J Mol Model (2009) 15:765–805
DOI 10.1007/s00894-008-0420-y

ORIGINAL PAPER

Application of the PM6 method to modeling proteins

James J. P. Stewart

Received: 18 July 2008 / Accepted: 14 October 2008 / Published online: 10 December 2008

Abstract The applicability of the newly developed PM6 method for modeling proteins is investigated. In order to allow the geometries of such large systems to be optimized rapidly, three modifications were made to the conventional semiempirical procedure: the matrix algebra method for solving the self-consistent field (SCF) equations was replaced with a localized molecular orbital method (MOZYME), Baker's Eigenfollowing technique for geometry optimization was replaced with the L-BFGS function minimizer, and some of the integrals used in the NDDO set of approximations were replaced with point-charge and polarization functions. The resulting method was used in the unconstrained geometry optimization of 45 proteins ranging in size from a simple nonapeptide of 244 atoms to an importin consisting of 14,566 atoms. For most systems, PM6 gave structures in good agreement with the reported X-ray structures. Some derived properties, such as pKa and bulk elastic modulus, were also calculated. The applicability of PM6 to model transition states was investigated by simulating a hypothetical reaction step in the chymotrypsin-catalyzed hydrolysis of a peptide bond. A proposed technique for generating accurate protein geometries, starting with X-ray structures, was examined.

This work was funded by the National Institutes of Health Grant No. 1 R43 GM083178-01

Electronic supplementary material The online version of this article (doi:10.1007/s00894-008-0420-y) contains supplementary material, which is available to authorized users.

J. J. P. Stewart (✉)
Stewart Computational Chemistry,
15210 Paddington Circle,
Colorado Springs, CO 80921, USA
e-mail: MrMOPAC@OpenMOPAC.net

Keywords PM6 · Hydrogen bonding · Metalloenzymes · MOZYME · Proteins · Salt bridge · Young's modulus

Introduction

Semiempirical methods, such as MNDO [1, 2], AM1 [3], and PM3 [4, 5], have been used for a long time for modeling small organic and inorganic systems. They have not, however, enjoyed much success when used for modeling proteins, primarily due to the poor accuracy in reproducing the geometries of large organic systems such as oligopeptides, and to the large computational effort involved.

The recently developed semiempirical method PM6 has been shown to reproduce the heats of formation and geometries of small molecules [6], simple organic and inorganic crystals [7], and a hormone—the nonapeptide oxytocin [7]—with good accuracy. Because of these encouraging results, determining the applicability of PM6 to modeling larger biochemical molecules, particularly proteins, was of obvious interest.

Despite their large size, proteins can be regarded as simple organic compounds, being composed mainly or even entirely of residues of the 20 common amino acids. Some, for example, cytochrome-P450 and hemoglobin, are more complicated, in that they contain organometallic structures, e.g., the metalloporphyrin heme ring system, while others, such as the structural protein collagen, contain modified residues. But with the exception of photoactive sites of the type that occur in chlorophyll and in retinal-containing proteins, most of the subtle electronic phenomena frequently encountered in transition metal chemistry is absent. At the primary structural level, therefore, proteins can be described in terms of simple covalent bonds and

weak hydrogen bonds. As such, the primary structure is easily and accurately reproduced by NDDO [8] methods, in particular PM6. It is only when secondary, tertiary, and quaternary structures are involved that the wide range of properties of proteins becomes apparent, among the more important of which is the ability to catalyze reactions.

Several successful approaches have been developed to reduce the computational effort required for modeling proteins using semiempirical quantum chemical methods. One such technique is the divide-and-conquer method [9], in which a large molecule is divided into fragments, each of which is relatively easy to manipulate using quantum chemistry methods. This approach has been applied extensively to biochemical systems [10–13].

Another, hybrid, approach has been to model the interesting parts of a protein using quantum mechanics (QM), and to model the rest of the system using molecular mechanics (MM). This approach, QM/MM, although complicated because of problems associated with the boundary between the QM and MM regions, has enjoyed considerable success [14].

In this work, the entire protein is modeled using semiempirical QM methods in which the self-consistent field (SCF) equations are solved using the localized molecular orbital method MOZYME[15] instead of the more conventional matrix algebra methods. Minor changes to the basic MOZYME approach were made in an attempt to reduce the computational requirements. These changes were incorporated into the program MOPAC2007[16]. All calculations were performed using 3.6 GHz PCs, each of which had either 1 or 2 Gb of RAM.

Theory

Solving the SCF equations

A consequence of the way that semiempirical methods developed is that the SCF equations are usually solved using matrix algebra techniques, and because the computational effort required for such operations scales as the third power of the number of atomic orbitals involved, conventional methods of solving the SCF equations are impractical when applied to large systems.

One way to accelerate the solution of the SCF equations for a large system is to divide it into smaller pieces and then solve the SCF equations for each piece. Thus, if a large system is divided into n approximately equal fragments, then the CPU time required to solve the SCF equations for each of the n fragments would be $1/n^3$ of that for the entire system. The increase in speed as a result of using this divide-and-conquer method would therefore be approximately n^2. The divide-and-conquer [17] approach has been used successfully in modeling biological systems [9], including proteins [13] of the type discussed here.

Obviously, the greatest increase in speed is obtained when the system is divided into the maximum number of fragments. If this concept is taken to its logical limit, the smallest fragment of an electronic structure of a molecule is the individual chemical bond or lone pair. This is the principle behind the localized molecular orbital (LMO) method, MOZYME[15].

The MOZYME technique begins with generating the Lewis structure for the system. By their nature, Lewis structures are a better approximation to the correct electronic structure than is the starting point in conventional SCF methods. Given a starting Lewis structure consisting of monatomic (lone pairs) and diatomic (bonds) LMOs, an improved electronic structure can then be obtained by annihilating the energy terms connecting the nearby occupied and virtual LMOs by performing a series of 2×2 Euler rotations. This results in a lower energy, and, by implication, an improvement in the electronic structure. When this annihilation process is performed repeatedly, the energy of the system is reduced to a minimum, and therefore, by implication, the SCF equations are solved. As with the SCF procedure in conventional methods this is necessarily an iterative procedure but, because LMOs are used, the computational effort required scales almost linearly with the size of the system, instead of the N^3 dependency of conventional methods.

Use of point charges

A further, large, increase in computational efficiency can be obtained by replacing the NDDO set of approximations with simpler terms for pairs of atoms that are separated by large distances. The NDDO approximation is only necessary for pairs of atoms that have a finite electronic interaction, that is, for atom pairs whose common density matrix terms are significantly non-zero. Without significant loss of precision, the NDDO approximation for all other atom pairs can be replaced by either a point charge or a point charge plus polarization term, to represent the effect of lone pairs. To allow the transition from exact NDDO to point charge plus polarization to be varied, a parameter, CUTOFF = *n.nn*, was added to the set of keywords used by MOPAC.

The effect of different values of the transition distance was investigated by optimizing the structure of a medium-sized protein, chymotrypsin, using various values of CUTOFF. The results are shown in Table 1. Only small geometric changes, in the order of 0.01 Å, were observed when CUTOFF was in the range 8–18 Å, indicating that any value of CUTOFF in that range would yield geometries essentially indistinguishable from the exact PM6 structure.

Table 1 Effect on geometry optimization of chymotrypsin of varying CUTOFF. Times are averages over six cycles. Times for cycles of less than 5 Å varied too much to be useful. *RMS* Root mean square

Cutoff (Å)	Average time per cycle (s)	Gradient norm of "exact" PM6 geometry (kcal mol^{-1} $Å^{-1}$)	RMS difference from "exact" PM6 geometry (Å)	ΔH_f of optimized geometry (kcal mol^{-1})	ΔH_f of "exact" PM6 geometry (kcal mol^{-1})
4	–	175.1	0.092	−25,680.7	−25,673.4
4.5	–	157.6	0.073	−26,014.3	−25,990.5
5	256	143.7	0.021	−26,261.5	−26,251.5
6	260	116.0	0.019	−26,127.2	−26,112.1
8	478	83.4	0.014	−26,183.3	−26,174.4
10	604	67.9	0.007	−26,205.1	−26,200.5
12	747	58.1	0.004	−26,192.6	−26,189.2
14	886	48.0	0.004	−26,194.9	−26,191.7
16	940	14.2	0.004	−26,197.0	−26,194.1
18	937	14.3	0.000	−26,193.4	−26,193.4

The advantage of a small value for the transition distance becomes apparent when the value of CUTOFF is decreased: the computational effort decreases rapidly, so that, at CUTOFF=8 Å, an optimization cycle requires only about half the computational effort of CUTOFF=18 Å. When the value of CUTOFF drops below 6 Å, the computational effort decreases even more dramatically, but in that domain the value of the results becomes questionable, due to the rapid increase in distortions from the exact PM6 structure.

Changes in forces acting on the fully optimized PM6 structure as a result of varying the value of CUTOFF were calculated using the structure obtained from CUTOFF= 18 Å as the reference PM6 geometry. When single-point calculations were performed on the reference PM6 structure, the gradient norm increased rapidly as the transition distance decreased, reflecting the close relationship of forces acting on the geometry (Table 1, column 3), and the degree of distortion of that geometry (Table 1, column 4). These geometric changes were also reflected in the heat of formation, where, as expected, the change in ΔH_f on going from the reference PM6 structure to the fully optimized structure increased as CUTOFF decreased. Interestingly, the changes in ΔH_f for the optimized geometry and for the reference geometry did not follow any obvious pattern as the value of the CUTOFF decreased.

Geometry optimization

Internal coordinates are normally used when the geometries of small molecules are being optimized, because of their increased efficiency compared to using Cartesian coordinates. This is true even for large systems, provided that the topology of the system does not include any large rings. In this context, hydrogen bonds, salt bridges, and other weak interactions can be regarded as bonding interactions, and therefore by implication they contribute to the topology. As proteins invariably contain many such weak interactions, their topologies also invariably contain many large rings. If internal coordinates were used in defining systems that contain large rings, then any small change in internal coordinate angles would result in large changes in interatomic separations involving bonded atoms, atoms often far away from those used in defining the angle. This would result in large fluctuations in the heat of formation and a consequent failure of the optimization procedure. To avoid this problem, Cartesian coordinates were used in all optimizations reported here.

Two conventional geometry optimization methods have been developed for use in semiempirical packages such as MOPAC. Both, however, were found to be unsuitable for optimizing the geometries of large systems. The more efficient method, Baker's Eigenfollowing (EF) technique [18], involves matrix operations which, while very rapid for small numbers of geometric parameters, soon become impractical when applied to large systems because the number of arithmetic operations required by such operations scales as the third power of the number of parameters. To a lesser degree, the same problem precludes the use of the older BFGS [19–22] procedure. In addition, since both methods require the construction and manipulation of matrices whose size is proportional to the square of the number of parameters, as the size of system increases, the memory requirements of these methods eventually becomes prohibitive.

The problem of optimizing large numbers of parameters to minimize the value of a function occurs often in computational methods, and has been a focus of interest to the Optimization Technology Center. This group developed the L-BFGS method [23, 24], a limited-memory quasi-Newton code for unconstrained function optimiza-

tion. The L-BFGS optimization technique is, as its name suggests, a modification of the BFGS method that is specifically designed for use with systems involving large numbers of parameters. Because it is based on the BFGS method, and because it does not make full use of the partial Hessian constructed in the EF procedure, the L-BFGS method is less efficient than EF for optimizing the structures of systems of less than about 2,000 parameters. However, for systems with over 2,000 variables, the L-BFGS method was found to be significantly more efficient, and as a result was made the default for large systems.

Geometry optimization was typically performed in three stages. First, the positions of all hydrogen atoms were optimized. In those cases where the hydrogen atoms were not reported in the starting X-ray structure, hydrogen atoms were added, but obviously their initial positions were only estimates. Where hydrogen atoms were present, the hydrogen bond lengths were usually too short by 0.05–0.15 Å. Errors of this type became immediately apparent when the positions of the hydrogen atoms were optimized, the heat of formation of the resulting structure being invariably much less than that of the starting geometry, typically by hundreds of kcal mol^{-1}. CUTOFF was set to 5 Å during this process. In the second stage, the positions of all atoms were optimized, again using a CUTOFF of 5 Å. This operation was terminated when the gradient dropped to 20 kcal mol^{-1} $Å^{-1}$. Finally, CUTOFF was set to 9 Å, and the geometry re-optimized. For small proteins, optimization was terminated when the gradient dropped below 1 kcal mol^{-1} $Å^{-1}$. For many of the larger proteins this criterion was too severe, and, instead, optimization was terminated when the heat of formation did not decrease significantly over 20 cycles of optimization. With the exception of the largest proteins, this typically corresponded to a gradient of 10 kcal mol^{-1} $Å^{-1}$.

Preconditioning

Starting geometries of all the systems reported in this work were obtained from the Protein Data Bank (PDB) [25]. Because of the way X-ray structures are determined, modifications had to be made to all the structures obtained from the PDB before they could be used in meaningful quantum mechanical calculations. The most common preconditioning operations were:

Where positional or structural disorder existed, the disorder was resolved to yield a single structure. Disorder of this type occurs naturally, but because semiempirical simulations require a well-defined system, only one of the various structures reported could be used. Serendipitously, disorder of this type invariably occurred only where its presence was not important. That is, in all cases where disorder exists, all candidate structures were equally suitable, and the choice could be made arbitrarily without significantly affecting any important sites in the system.

Where atoms in a residue or even entire residues were missing, valency requirements were satisfied by adding hydrogen atoms. As with positional and structural disorder, when groups or residues were missing, the absences again invariably occurred in sections of the chain far away from any interesting features of the protein such as active sites.

The main structural deficiency, arising from the way X-ray structures are generated, was the frequent and complete absence of any hydrogen atoms. These were added as needed. Several residues, such as Asp and Asn, contain ionizable groups; all such groups were represented by their initially neutral forms, this being regarded as the most easily definable ionization state for the entire system.

Where the positions of hetero molecules, such as water, sulfate, phosphate, ethanol, or other small organic species, were given or indicated, such hetero molecules were used in the model.

Preconditioning was completed with a preliminary calculation to optimize the positions of all hydrogen atoms.

Applications

The suitability of PM6 for modeling proteins was investigated by modeling several properties of proteins obtained from the PDB. Because the focus of this work was to determine the applicability of PM6 to modeling proteins, systems were selected that illustrated specific properties. Many of the proteins examined here have been the subject of intense and extensive study because of some important biological property, such as the actinic response of bacteriorhodopsin. However, for the purposes of this work, such properties were considered to be of secondary importance.

Before any computational method can be used for modeling proteins, the ability of the method to accurately reproduce known structures must be determined. This will be demonstrated for PM6, after which the applicability of PM6 to the study of other properties, including both chemical, such as the ability to catalyze reactions and prediction of pKa, and physical, such as biomechanical behavior, can then be examined.

Geometries

Proteins can be very large molecules. Even the smallest of the systems used in this work is much larger than the largest molecule that was used during the development and testing of PM6. Because of this, analyses of the type used

previously in reporting PM6 geometric results were considered to be inappropriate, and, instead, the analysis of the accuracy of prediction of protein geometries given here will be split into the various levels of structural complexity normally found in proteins. In order of increasing complexity, these are: primary, which is the structures of individual amino acid residues and their order in the polypeptide chain; secondary, which deals with common local structures, such as α helices, β sheets, and turns, in which the individual structural motifs are held together by hydrogen bonds; tertiary, which deals with packing motifs involving secondary structures; and, finally, quaternary, which is the packing together of entire protein subunits, i.e., two or more chains.

Primary structure

Of the four levels of complexity found in proteins, the primary structure is the only one that can be related to the geometric analyses of the type commonly used in reporting the accuracy of prediction of semiempirical methods, i.e., a useful description of the primary structure that can be obtained by reference to bond-lengths and angles. An analysis of these for PM6 has already been reported [6], and therefore will not be elaborated further here; instead, a more useful, and simpler, measure of the predictive power of a method will be used, namely the root mean square (RMS) difference between the optimized and reference structures. The use of X-ray structures of proteins as reference data is less than ideal in determining the accuracy of prediction of primary structures, in that X-ray structures of proteins typically have resolutions many times larger than the average error in predicted geometries of individual residues. A more appropriate source of reference data would be to use the results of high-level calculations. For this purpose, the optimized geometry predicted by the B3LYP[26] method using the 6–31G(d) basis set was selected. Table 2 shows the RMS differences for the 20 isolated amino acids commonly found in proteins. The average difference is 0.225 Å, with most of this being attributable to rotation about the relatively flexible bonds to the carboxylic acid and amino groups.

The peptide linkage can be represented by N-methyl acetamide, where PM6 predicts the C–N distance to be 1.396 Å, slightly longer than the B3LYP value of 1.367 Å.

Secondary structure

Unlike the primary structure, there is a wealth of accurate X-ray data for secondary structures. Three structural elements occur frequently in secondary structures: the α helix, the β sheet, and turns. As these involve very different structural features, they will be treated separately.

Table 2 Root mean square differences in structures of the 20 amino acids calculated using B3LYP and PM6

Amino acid	RMS error (Å)
Glycine	0.050
Alanine	0.094
Valine	0.160
Leucine	0.237
Isoleucine	0.239
Serine	0.283
Threonine	0.292
Aspartic acid	0.303
Asparagine	0.385
Lysine	0.194
Glutamic acid	0.256
Glutamine	0.190
Arginine	0.425
Histidine	0.157
Phenylalanine	0.220
Cysteine	0.228
Tryptophan	0.161
Tyrosine	0.219
Methionine	0.134
Proline	0.271
Average	0.225

Alpha helix

Alpha helices consist of a right-handed helical arrangement of the polypeptide backbone in which the pitch, that is, one complete turn of the helix, involves 3.6 residues and results in a translation of about 5.4 Å along the axis. Helices are stabilized by the N–H of the amide group on residue *n* forming a hydrogen bond with the C=O of the amide group on residue *n* −4. A good example of helix structure is provided by the halobacteria protein bacteriorhodopsin (bR). Bacteriorhodopsin is a trans-membrane protein consisting of seven α-helices in the center of which is an extended conjugated π system, retinal, that forms a Schiff base with one of the residues, Lys216.

Several structures for bR are given in the PDB, although all the entries report significant positional disorder, with the result that, in all cases, one or more residues are missing. One of the more accurate structures is 1C3W [27]. After preconditioning, the structure of 1C3W was optimized using PM6, and the predicted and native geometries compared. All seven helices exhibited similar distortions (Fig. 1), so only the structure of helix "A" will be described, it being representative of the other helices. In this helix, the RMS error for the 26 residues involved was 0.73 Å. Structures within the helix were reproduced with good accuracy, thus the hydrogen bond distances increased only slightly (Fig. 2), resulting in an increase in the average N–hydrogen bonded–O distance from 2.80 Å to 3.00 Å,

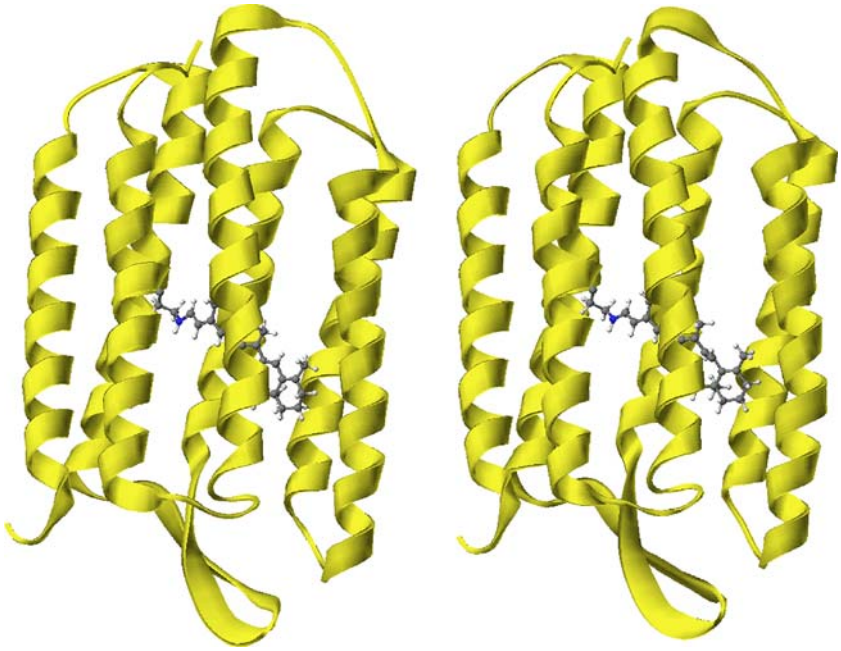

Fig. 1 Comparison of X-ray structure of bacteriorhodopsin (bR) 1C3W with PM6 structure, showing α helices and retinal group. *Left* X-ray structure, *right* PM6

and an increase in the pitch by 0.1 Å from the 5.4 Å in the native bR structure to 5.5 Å. The value predicted by PM6 for the number of residues in one turn of the α helix matched the value from 1C3W within the limits of computational uncertainty. This high accuracy can be attributed to the presence of intra-chain hydrogen bonding, in that any significant change in the number of residues per turn would have required breaking of the original hydrogen bonds.

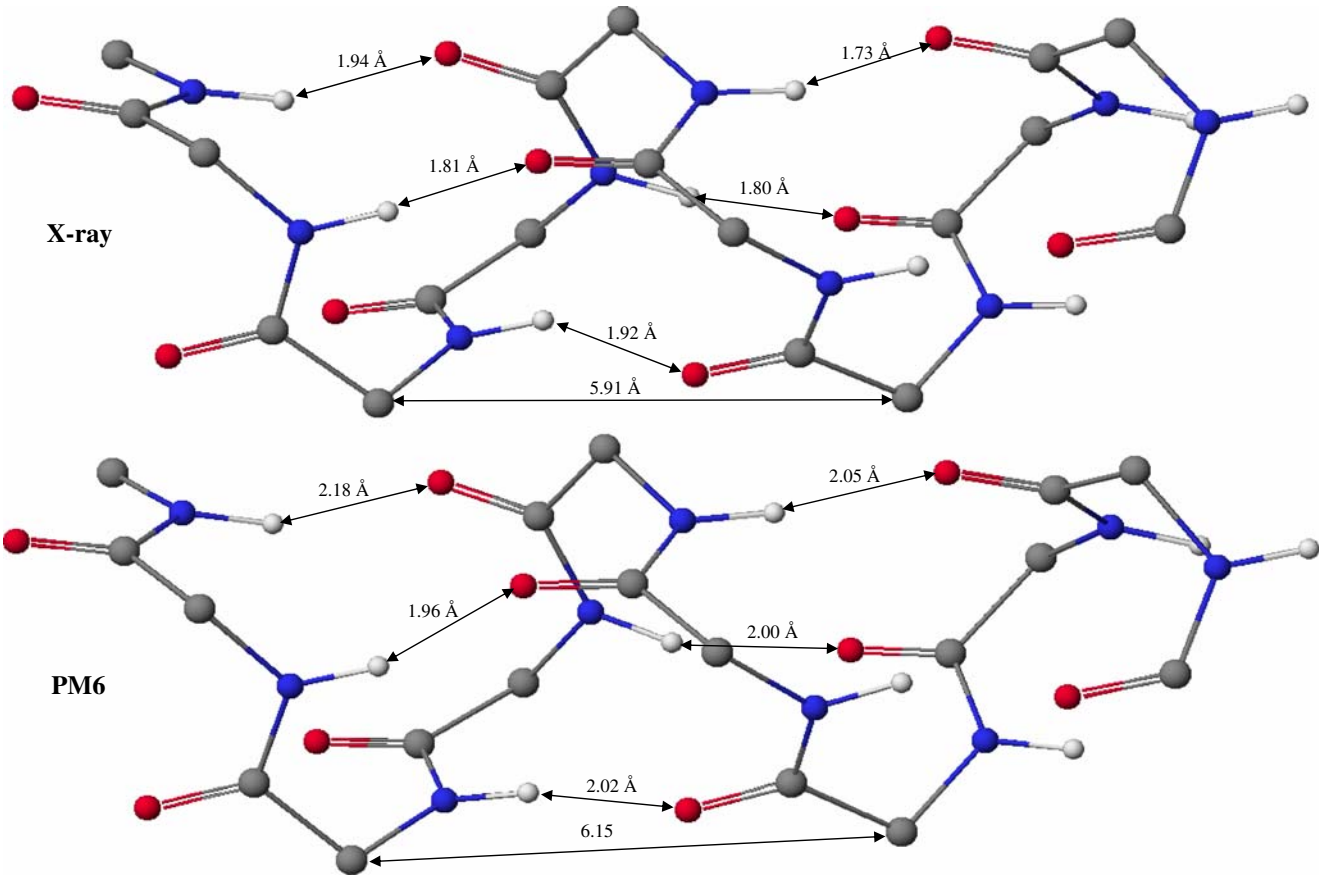

Fig. 2 Detail of α helix in bR

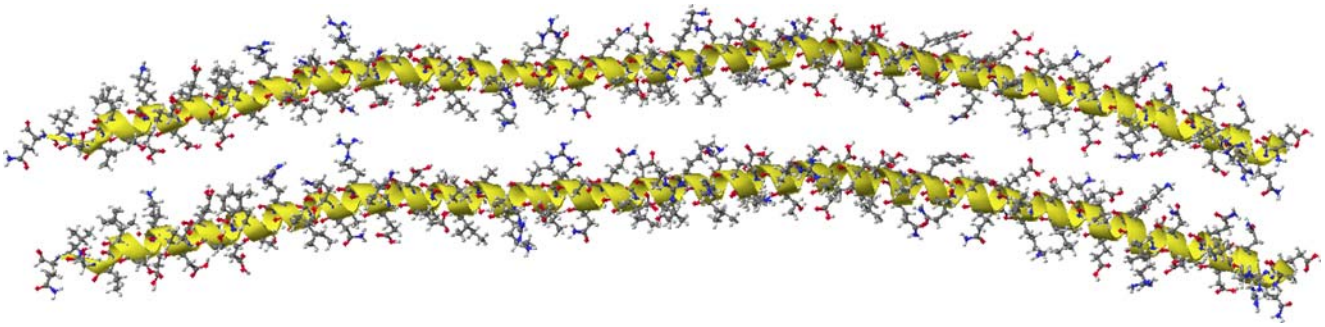

Fig. 3 Comparison of one chain in native (*top*) and PM6 structures of 2V66 showing extended α helix. Note distortion of PM6 structure at N terminus (*left end*) of the chain

An example of a longer α helix structure is provided by 2V66, a tetrameric coiled-coil structural protein consisting of four identical chains, each of which forms an α helix with 30 complete turns. In the native structure, the distance between the N and C termini of each chain spans a range from 164.0 to 164.7 Å, the spread of values presumably being a consequence of the quaternary structure. The optimized PM6 structure differed only slightly from that of the native structure, as illustrated for one of the chains in Fig. 3. PM6 predicts the termini to be separated by distances ranging from 161.4 to 164.3 Å, i.e., the α helices are predicted to be very slightly compressed, in contrast to those in bR where they are expanded. Most of the difference arises from distortions of the residues at the exposed N termini, shown at the left end of the molecule in Fig. 3. These distortions are likely due to inadequate modeling of the local environment; in the crystal, the N termini would presumably contribute to the formation of salt bridges, possibly with other chains.

A useful way to represent protein secondary structures is the Ramachandran plot, where the backbone torsion angles for individual residues are represented by points on a graph spanning the domains from −180° to 180°. In a Ramachandran plot for an α helix, the ϕ angles typically span a range from −80 to −70°, and the ψ angles range from −70° to −30°. The Ramachandran plot for the optimized PM6 geometry of 2V66 (Fig. 4) shows good agreement with the X-ray structure. Points on the plot far away from the α helix region arise from residues near the ends of the α helices, where most of the largest differences in torsion angles also occur, the largest of these being about 97°. Another useful method of comparison is to plot individual

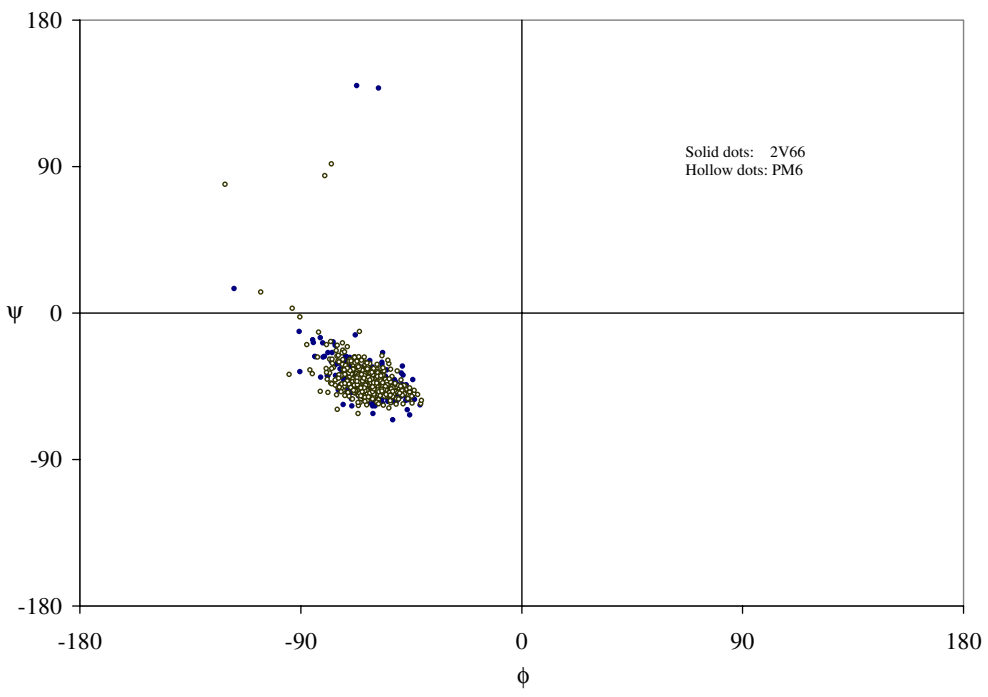

Fig. 4 Ramachandran plot for α helix protein 2V66 and PM6

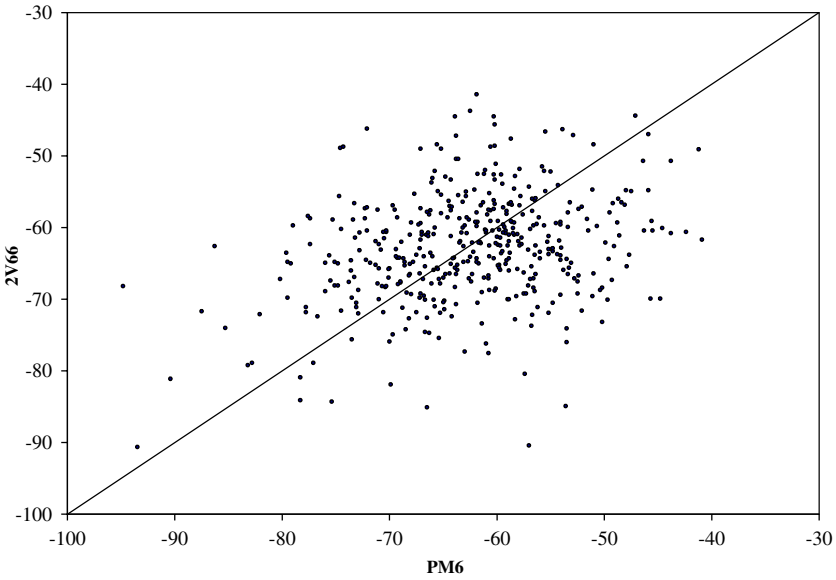

Fig. 5 Comparison of ϕ angles for 2V66 and PM6

ϕ and ψ angles for the PM6 and X-ray structures (Figs. 5 and 6, respectively). These graphs reinforce the result obtained from the Ramachandran plot, the median unsigned error being 6.3°.

All other structural proteins containing large α helices that were examined, such as the mid-section of the contractile protein tropomyosin, 2B9C, where 39 complete turns occur, and the cytoskeletal protein spectrin, 2SPC (Fig. 7), exhibited similar behavior.

β antiparallel sheet

Another important secondary structure in proteins is the β sheet, in which individual β strands form hydrogen bonds with adjacent β strands, resulting in either parallel or antiparallel β sheets, depending on the relative direction of the strands. A classic example of the stronger of these two arrangements, the β antiparallel sheet, is provided by silk. Silk is a strong fibrous protein produced by arachnids and some insects. Because of its wide availability, one of the most studied forms of silk is Silk I from the mulberry silkworm, *Bombyx mori*. A putative structure for the crystalline domain of Silk I is represented in the PDB by a simulation, entry 1SLK [28], where, for computational simplicity, the naturally occurring structure was replaced by the model high polymer, poly(L-Ala-Gly). In that work, the authors modeled the crystalline structure using a large discrete cluster consisting of 15 heptapeptides. They postulated a novel structure that was consistent with reported X-ray diffraction patterns, in which the packing

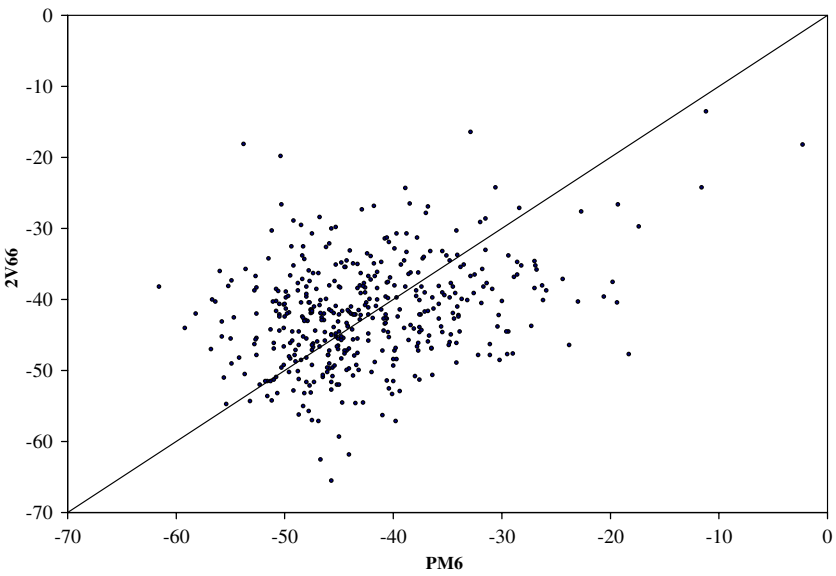

Fig. 6 Comparison of ψ angles for 2V66 and PM6

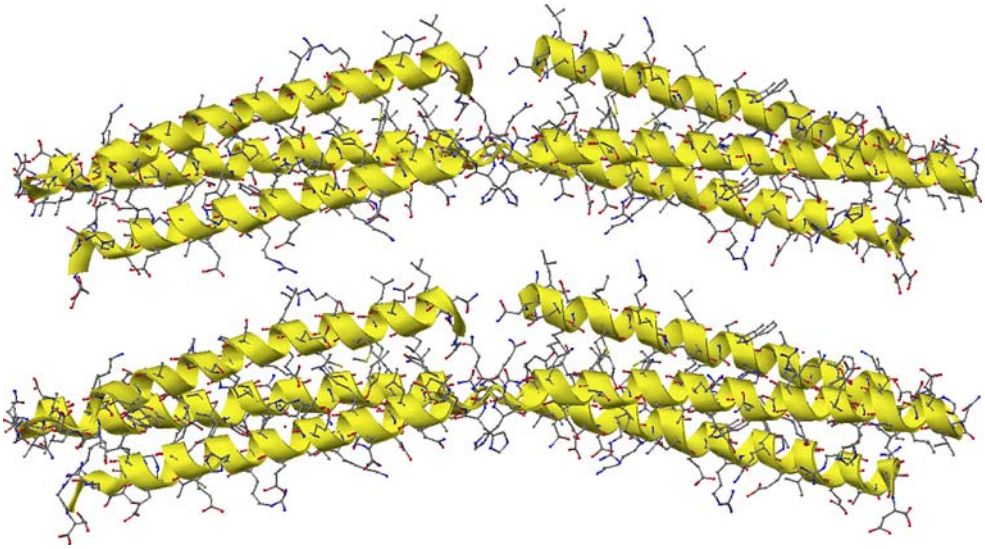

Fig. 7 X-ray and PM6 structures of spectrin (2SPC), showing extended α helices. *Top* X-ray structure, *bottom* PM6

of the antiparallel β-sheets gave rise to a highly symmetric structure. The unit cell was reported to be orthorhombic, with dimensions a=8.94 Å, b (fiber axis) = 6.46 Å, and c= 11.26 Å. Subsequent X-ray work [29] on *B. mori* Silk I, that is, using the naturally occurring polymer, confirmed this structure, and gave the unit cell dimensions as a= 9.38 Å, b=9.49 Å, and c (fiber axis) = 6.98 Å.

Using PM6, and employing periodic boundary conditions (PBC), a solid state geometry optimization of poly(L-Ala-Gly) was carried out, starting with a data set constructed from the quasi-crystalline central section of the large cluster used in [28], i.e., the starting topology was that of 1SLK. The resulting optimized unit cell dimensions were a=8.78 Å, b (fiber axis) = 6.30 Å, and c=9.97 Å, and the internal structure of the unit cell (Fig. 8), although similar to that in 1SLK, had some subtle differences. In 1SLK, each β strand formed hydrogen bonds with the adjacent strands, but in the PM6 optimized structure, in addition to the normal hydrogen bonds connecting the antiparallel β sheets, there were hydrogen bonds connecting adjacent sheets. Each amide hydrogen on a glycine residue formed two hydrogen bonds, one to a glycine residue in an adjacent

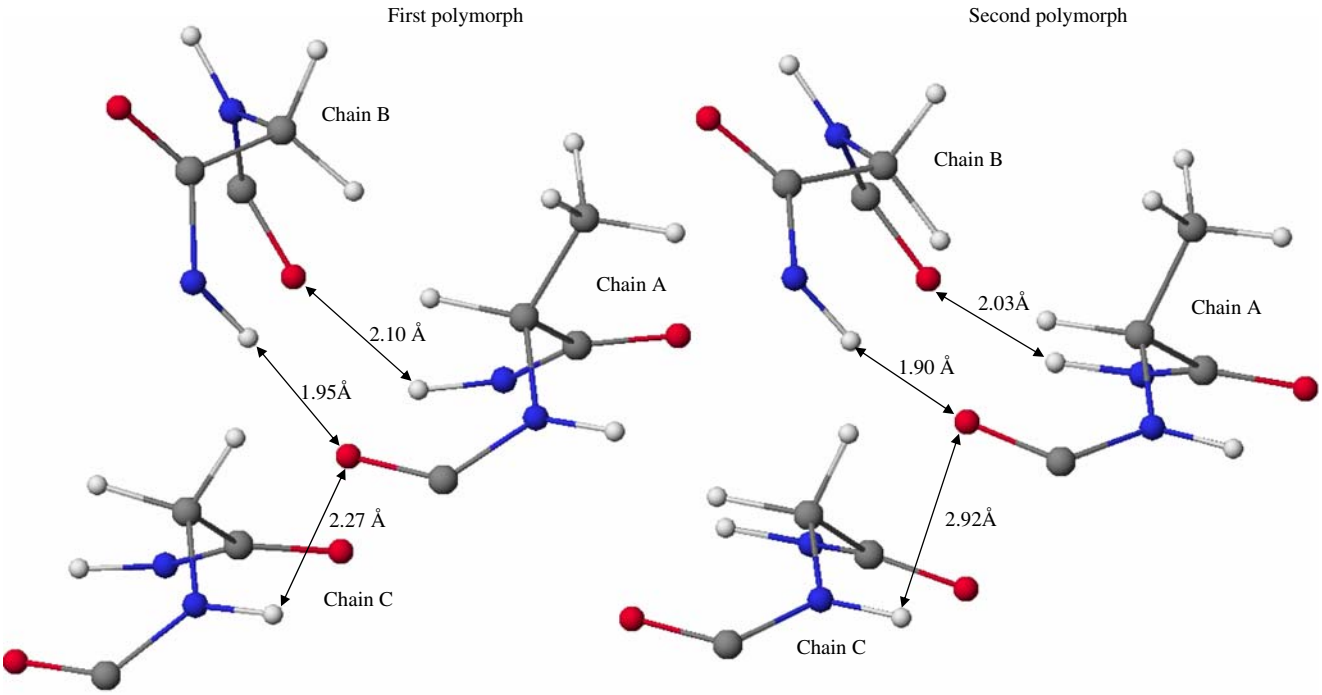

Fig. 8 Hydrogen bonds in poly(L-Gly-Ala). Chains A and B are in the same antiparallel β sheet, chain C is in an adjacent sheet

parallel strand, and one to an alanine residue in an adjacent antiparallel strand, and each alanine amide hydrogen formed one hydrogen bond with a glycine residue in an adjacent antiparallel strand. PM6 also predicted that a more conventional polymorph of poly(L-Ala-Gly) should exist. In this form, the inter-sheet hydrogen bonds were absent, the backbone chains were less angular, and the unit cell had an increased length in the direction of the fiber axis, resulting in the unit cell dimensions: a=9.10 Å, b (fiber axis) = 6.85 Å, and c=9.05 Å. To verify that these were indeed distinct polymorphs, both structures were optimized further, until the gradient dropped below 1.0 kcal mol^{-1} $Å^{-1}$. At that point, any attempt to increase the size of the translation vector corresponding to the fiber axis for the first polymorph, or to decrease the size of the translation vector corresponding to the fiber axis for the second polymorph, resulted in an increase in ΔH_f, confirming that the two polymorphs were separated by an activation barrier, and that the two polymorphs were in fact in different potential wells.

In addition to extended planar surfaces, antiparallel β sheets can form cylinders. A good example is the green fluorescent protein (GFP), 1EMA [30], found in the jellyfish *Aequorea victoria*. GFP consists of an 11-stranded beta barrel, with the photoactive site being composed of a section of chain oriented along the axis of the barrel. Minor defects, consisting of several missing residues and some missing atoms, were reported in the X-ray structure, but these were unlikely to influence the secondary structure. Geometry optimization of GFP resulted in only a small motion of the atoms, so that the optimized and reference structures were similar (Fig. 9). Torsion angles were also reproduced with good accuracy, albeit not as good as for the α helix, as shown in the Ramachandran plot (Fig. 10). As expected for a β antiparallel sheet, most of the reference and PM6 points are in the top left quadrant, but because the structure of GFP involves many turns, there is a significant fraction of points scattered throughout the plot.

A more representative example of the β antiparallel sheet is provided by the *Borrelia burgdorferi* outer surface protein 2G8C [31], in which the residues between Gln96 and Ile150 form a typical β sheet consisting of nine strands, and exhibiting the characteristic twist of β sheets. After preconditioning, the geometry of this highly hydrated system—each protein is associated with 388 water molecules—was optimized using PM6. A comparison of the native 2G8C structure and the fully optimized PM6 structure is shown in Fig. 11, and details of the β sheet are given in Fig. 12. For simplicity, all side chains in the figure have been truncated to the C_β atom, and only the hydrogen bonds between amide hydrogen atoms and carbonyl oxygen atoms shown. The β sheet twist is conserved in the optimized PM6 structure. As with GFP, the Ramachandran plot for 2G8C (Fig. 13) shows the characteristic clustering in the top left quadrant, and slightly more distinct clustering due to the short section of α helix.

β parallel sheets

Possibly because parallel β sheets have less favorable hydrogen bonds, they occur less frequently than the antiparallel form, and, when they do occur, it is usually in assemblies of several parallel strands. A common form of parallel β sheet is the beta-helix, in which the polypeptide

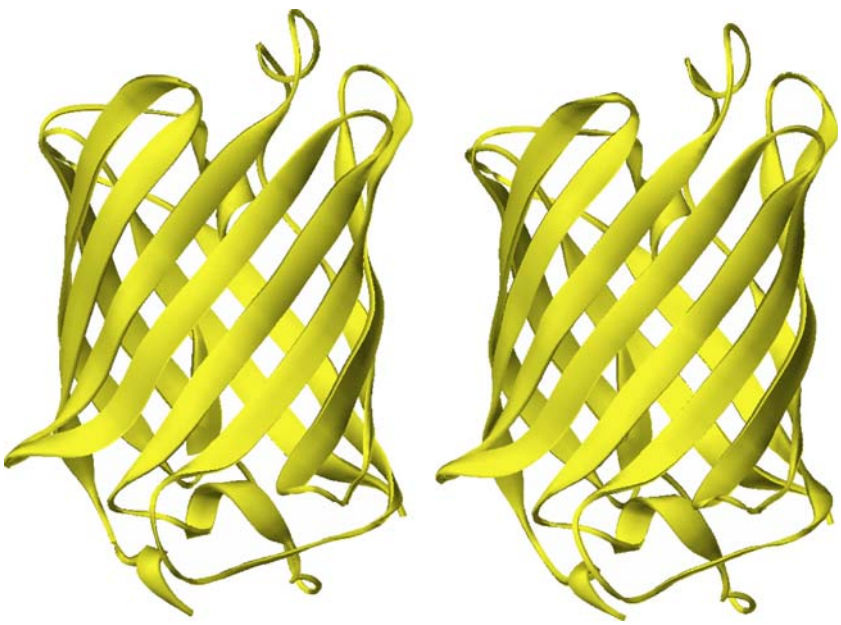

Fig. 9 Green fluorescent protein (GFP) 1EMA. *Left* X-ray structure, *right* PM6

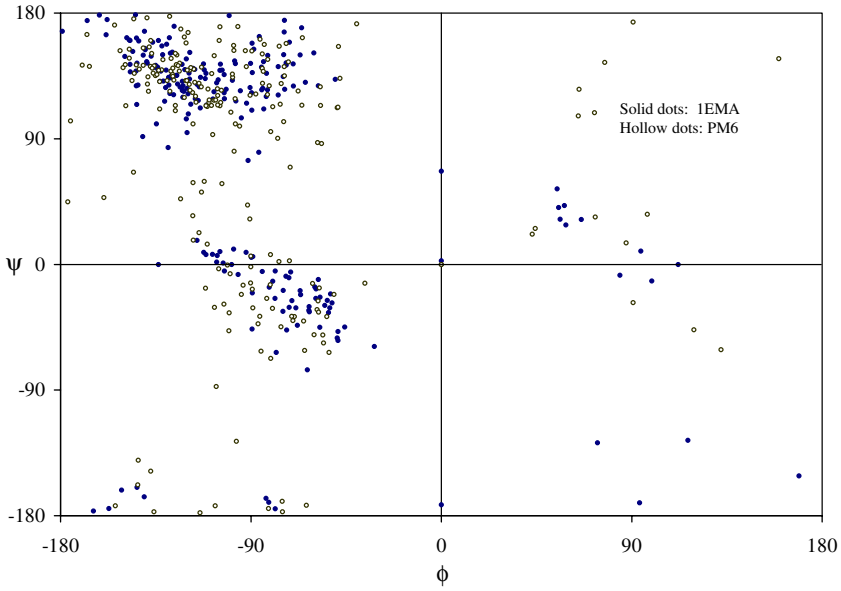

Fig. 10 Ramachandran plot for GFP 1EMA, an antiparallel beta barrel

backbone forms a large helix. These helices can be right- or left-handed. An example of a right-handed parallel beta-helix containing a relatively large region of parallel β sheet is provided by the mealworm thermal hysteresis protein YL-1, 1EZG [32]. After preconditioning, the geometry of 1EZG was optimized and the β sheet regions compared (Fig. 14).

Left-handed parallel beta-helices are a relatively rare secondary conformation; one example is UDP-N-acetylglucosamine acyltransferase from *Escherichia coli*, 1LXA [33]. X-ray and PM6 structures for this enzyme are shown in Fig. 15. Despite the fact that many unusual crossover connections have been reported [33] between the parallel β sheets in this structure, a superficial inspection revealed no obvious differences in the hydrogen bonding of the X-ray and PM6 structures, both being essentially similar to the right-handed beta-helix in 1EZG.

β turn

Another common motif in proteins is the β turn, in which a sharp bend occurs in the polypeptide chain. An example of a typical β turn is provided by the connection between the α helices "B" and "C" in bR, and consists of the residues Pro70, Phe71, Gly72, Gly73, and Glu74. From an examination of six X-ray structures for bR, there appears to be no single definitive geometry for the β turn, with even the locations of the hydrogen bonds changing on going from one PDB entry to another. Each of these structures, upon optimization, yielded a structure for the β turn that, with one exception, was significantly different from both the X-ray structure and the other PM6 structures. The exception was that the PM6 β turns resulting from optimization of the 1BRX and 1C3W structures were similar, with the average backbone torsion angles for this pair differing by only 5°. This lack of a definitive reference structure makes any meaningful comparison of calculated and reported geometries distinctly more difficult. For simplicity, therefore, the geometry of the β turn in the optimized PM6 structure for bR was compared with the X-ray structure 1C3W, this being representative of the other structures. The results are shown in Fig. 16. Examination of the turn shows that, where 1C3W had one strong hydrogen bond between the carbonyl of Phe71 and the hydrogen attached to nitrogen in Glu74, PM6 predicted two weaker

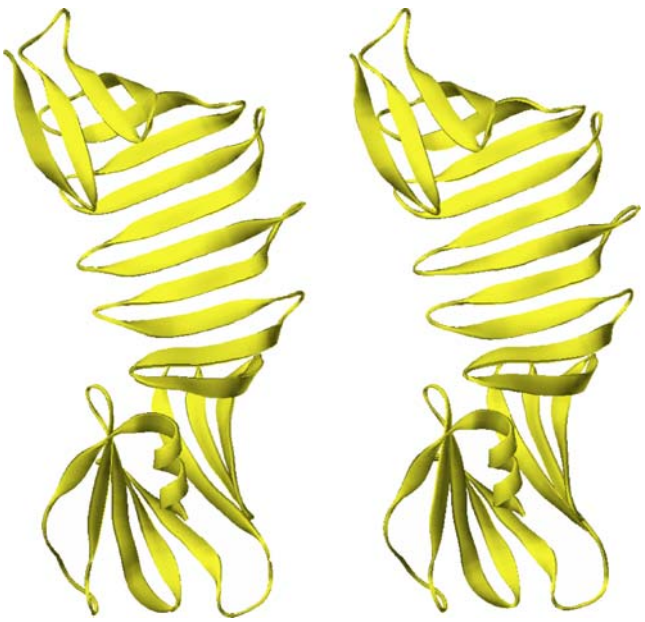

Fig. 11 Comparison of secondary structure of outer surface protein 2G8C with PM6. *Left* X-ray structure, *right* PM6

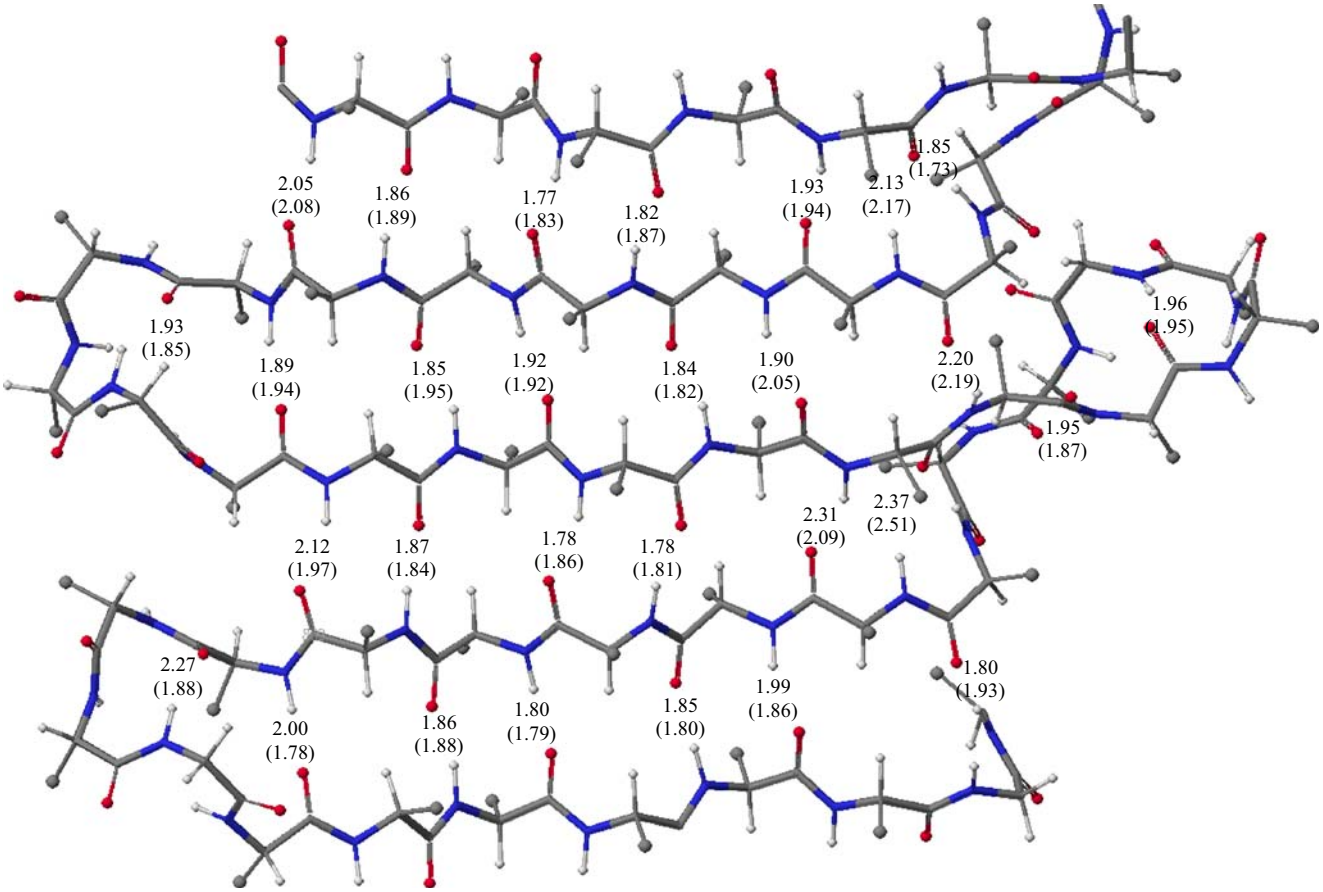

Fig. 12 Hydrogen bonding between amide and carboxyl groups in antiparallel β sheet in outer surface protein 2G8C. Distances in Ångstroms

hydrogen bonds, the second one being between the carbonyl of Pro70 and the hydrogen attached to nitrogen in Gly73. Despite the fact that the two structures have distinctly different geometries, all bond lengths and angles (but not dihedrals) were similar, and the distances between the termini of the β turns differed by only 0.09 Å, suggesting that PM6 can reproduce the secondary features of the β turn, but not necessarily all of the fine detail.

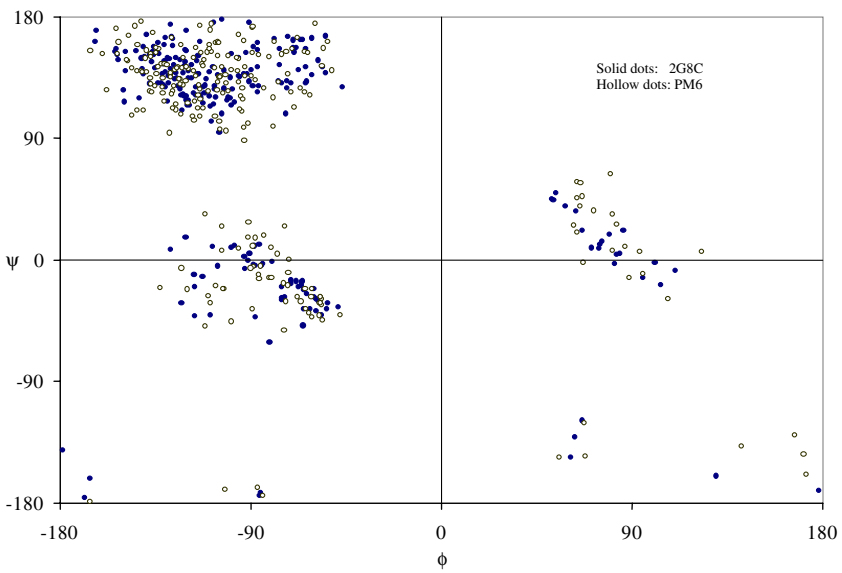

Fig. 13 Ramachandran plot for antiparallel β-sheet protein 2G8C

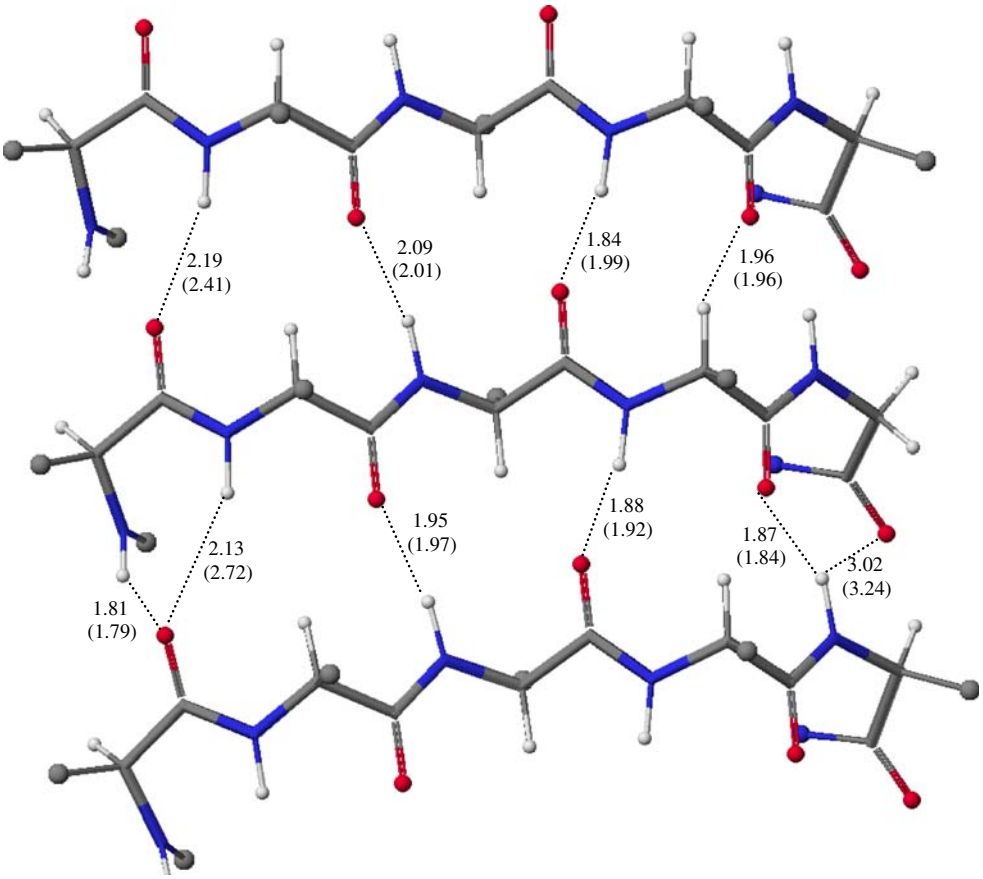

Fig. 14 PM6 hydrogen bonding distances in right-handed parallel β-sheet in thermal hysteresis protein isoform YL-1, 1EZG

Salt bridges

Protein structures are often stabilized by salt bridges, which, although much weaker than covalent bonds, are significantly stronger than hydrogen bonds, principally due to the electrostatic interaction arising from the large, almost unit, charges on the ionized residues.

Some salt bridges have been identified in X-ray structures, such as the two salt bridges reported in the high-resolution structure of crambin (1CBN [34]) between Arg10 and the carboxylate terminus Asn46, and between Arg17 and Glu23. Positional disorder was indicated in 1CBN by the letters "A" and "B". By default, set "A" was used. Crambin is a small globular protein found in Abyssinian cabbage seeds, contains only 46 residues, and, although it has no obvious biochemical activity, is a popular test of modeling methods. In part, this can be attributed to its unusual rigidity arising from three disulfide bridges and

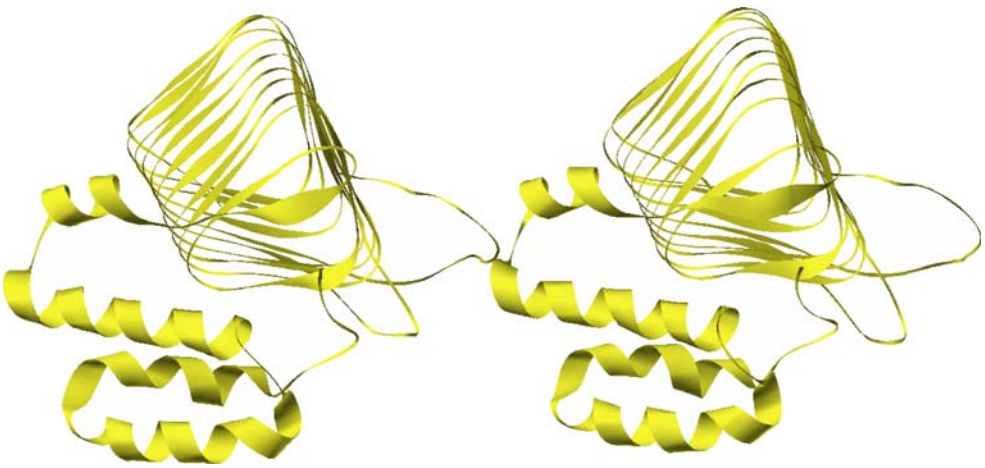

Fig. 15 Comparison of left-handed parallel beta helix UDP-N-acetylglucosamine acyltransferase 1LXA with PM6. *Left* X-ray structure, *right* PM6

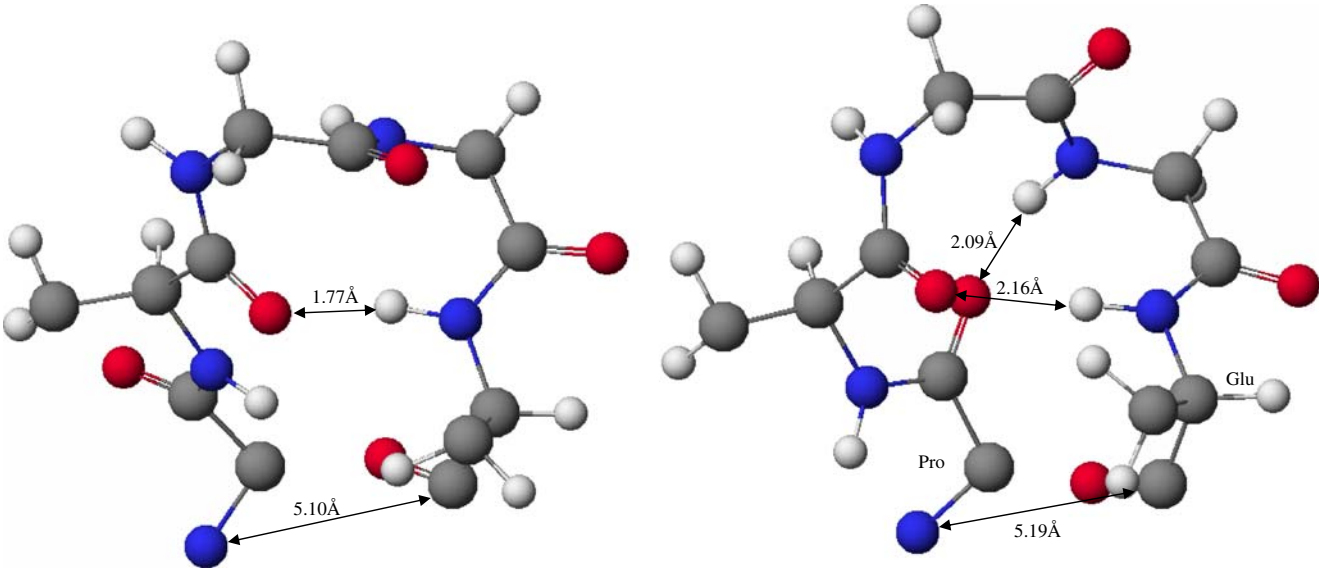

Fig. 16 Hydrogen bonds in β turn between a helices B and C in bR

to the presence of two distinct α helices, a set of features normally only found in larger proteins. Optimization of the X-ray structure of 1CBN using PM6 proceeded smoothly: the salt bridges were preserved in the optimized structure, and the RMS difference between the PM6 and X-ray structures was ~0.9 Å. When the salt bridges were neutralized and the structure re-optimized, the salt bridges reformed. That is, during the optimization procedure, the protons on the two carboxylate groups spontaneously migrated to the nearby arginine residues, the implication being that crambin with salt bridges is considerably more stable than crambin in the neutral form.

Crambin is unusual in that the locations of the hydrogen atoms were reported. In general, only the positions of the heavy atoms are given, so that, for most proteins, no inferences can be made regarding the locations of salt bridges. However, as with crambin, when the geometry of the neutral form of a protein is optimized, salt bridges might form spontaneously, strongly suggesting the existence of salt bridges in the natural form. An example of such a phenomenon is provided by barnase—a ribonuclease produced by *Bacillus amyloliquefaciens*. PDB entry 1RNB [35] reports the structure of the complex formed between barnase and the dinucleoside monophosphate d(GpC). One residue, Ala1, and several atoms in Gln2 were missing, but as these were distant from the sites of interest their absence is unlikely to affect the analysis significantly. Barnase is interesting in that two pairs of ionizable residues, Asp93 and Arg69, and Asp75 and Arg83, are in close proximity and are therefore candidates for the formation of salt bridges.

In accordance with the standard preconditioning protocol, the initial geometry used in the optimization had all potentially ionized sites represented by their neutral form. However, during the optimization procedure, a proton that was initially on Asp93 spontaneously migrated to Arg69 to form a salt bridge. The optimized distance between C_γ of Asp93 and the C_ζ of Arg69 was 3.97 Å; in 1RNB, the distance is 4.22 Å, i.e., the optimized PM6 distance was 0.25 Å shorter than that found experimentally. The other putative salt bridge, between Asp75 and Arg83, was not present in the optimized PM6 structure, but when the proton was moved from the aspartic acid to the arginine and the optimization re-run, the resulting heat of formation dropped by 17.5 kcal mol^{-1}. This strongly suggested that the salt bridge was indeed the stable form, and that a possible potential energy barrier to proton migration existed between the neutral and Zwitterionic forms. Additional evidence for the salt bridge is provided by the X-ray structure: in 1RNB the distance between C_ζ of Arg83 to C_γ of Asp75 is 4.21 Å. In the PM6 salt-bridge form, the equivalent distance was 4.00 Å, and in the neutral form the distance increased to 5.08 Å.

Reliability of prediction of secondary structures

One consequence of the fact that all geometry optimizations reported here start with the preconditioned native structure is that, of necessity, the structures predicted by PM6 calculations are biased towards the reference structure. An estimate of the degree to which this compromises the validity of the inferences being presented here was obtained by examining the effect of optimizing the ten reference structures in 1CAG [36], a nonapeptide from *Drosophila melanogaster*, all of which represent the same system.

This fragmentary system contained 121 atoms, and had the sequence PFCNAFTGC, with the two Cys residues

being connected by a disulfide bond. No obvious hydrogen bonds were present. All ten NMR structures reported for 1CAG were used and, as the locations of the hydrogen atoms were given, the only preconditioning required was the removal of H_8 from Pro1 and the addition of a hydroxyl group to Cys9, to make a terminal carboxylate group. All optimizations proceeded rapidly, each taking only about 13 CPU minutes. The RMS difference between the PM6 optimized structure and that reported in 1CAG averaged about 1.1 Å, a value much larger than expected for a biochemical of this size. Examination of the optimized structures showed that each of them corresponded to a different minimum on the potential energy surface (PES). That is, there were potential energy barriers between all the conformers.

Comparison of two of the structures, those corresponding to the highest and the lowest energy minima, showed that the higher energy conformer differed from the lower energy conformer by two distinct conformational differences, both of which involved a moiety of form $–CH_2R$ with the "R" group being rotated by ~120° in one conformer relative to the other. The higher energy structure was 18 kcal mol^{-1} above the lower energy structure, and the RMS difference in geometries was 1.81 Å. When the "R" groups of the higher energy conformer were rotated to match the lower energy conformer, and the structure re-optimized, the heat of formation dropped to within 1 kcal mol^{-1} of the lower energy conformer. Despite this, the geometries were still very different, with the RMS difference dropping only slightly, to 1.78 Å. When the higher energy system was continuously distorted so as to reduce the RMS difference between it and the lowest energy structure, and the geometry re-optimized at each point, the heat of formation did not rise. Instead, it decreased monotonically, indicating that no barrier existed between these structures. The obvious inference was that the two structures were in the same minimum, and that the PES near the bottom of that minimum was extremely flat. This condition—a very flat PES in the vicinity of the minimum—was also initially anticipated to occur with the larger proteins investigated. However, with few exceptions, this was not the case; subsequent work showed that 1CAG was unusual among the systems investigated in that all the other systems had better-defined minima. One possible reason for this phenomenon is that, because the non-apeptides in 1CAG were not complete biological molecules, they did not have the requisite hydrogen bonds and other structures that would normally confer rigidity to naturally occurring proteins. To investigate this conjecture, two naturally occurring proteins, bR and crambin, both of which have multiple structures in the PDB, were examined.

Each of the published structures of bR were optimized using PM6, and the geometries of the non-hydrogen atoms in the common sequence of residues, Pro8 to Val151 and Ala168 to Ile224, compared with those of the starting structures. By using only the sections of chain common to all the published structures, those regions where the structure is likely to be very flexible would automatically be excluded. The RMS differences are presented in Table 3 along with the differences between the various X-ray structures and between the various optimized geometries.

Surprisingly, the errors did not decrease. Instead, in all cases the RMS between the optimized PM6 structures was larger than that between the equivalent reference structure. This unexpected result prompted further examination of the reference data. The two most similar structures were 1BRX [37] and 1C3W, with the RMS difference being only 0.78 Å, and, since the distance between them on the PES was a minimum, they constitute the pair of structures that would be the least likely to have a barrier between them.

Table 3 Root mean square differences for X-ray and PM6 structures of bacteriorhodopsin (bR), heavy atoms for residues 8:151 and 168:224 only. PM6 structures were obtained by starting the optimization from the appropriate X-ray structure

		X-ray						PM6					
		1AT9	2AT9	1BRR	1BRX	1C3W	1AP9	1AT9	2AT9	1BRR	1BRX	1C3W	1AP9
X-ray	1AT9	0.00											
	2AT9	1.69	0.00										
	1BRR	1.57	1.31	0.00									
	1BRX	1.57	1.34	0.95	0.00								
	1C3W	1.61	1.34	0.81	0.78	0.00							
	1AP9	2.59	2.49	2.39	2.41	2.43	0.00						
PM6	1AT9	1.15	1.91	1.75	1.74	1.73	2.69	0.00					
	2AT9	1.70	1.67	1.73	1.73	1.74	2.70	1.82	0.00				
	1BRR	1.75	1.53	0.94	1.22	1.05	2.55	1.81	1.79	0.00			
	1BRX	1.69	1.47	1.12	0.79	0.97	2.48	1.78	1.73	1.20	0.00		
	1C3W	1.71	1.48	1.08	1.06	0.75	2.51	1.79	1.77	1.16	1.17	0.00	
	1AP9	2.61	2.57	2.41	2.43	2.43	1.13	2.66	2.69	2.52	2.46	2.46	0.00

Nevertheless, on examining the X-ray structures, the N–C–C–N torsion angle, i.e., the ψ angle, for Leu62, was very different in the two structures, being −8.9° in 1C3W and 149.6° in 1BRX, i.e., cis in 1C3W and trans in 1BRX. Conversion between cis and trans does not occur readily because of the presence of large steric barriers, so these two structures obviously represented points on the sides of different minima. As all other pairs of structures are separated by even greater distances, the likelihood is that each such pair is in a different minimum, and, by extension, all six bR are in different minima. The failure of the energy minimization procedure to show that any pair of structures were related by continuous deformation is thus rationalized.

In the middle of bR is a highly conserved biochemical structure, an extended (> 18 Å long) conjugated π system, retinal, that forms a Schiff base with Lys216. This structure is central to the actinic behavior of bR. Comparison of the 1C3W and 1BRX optimized structures for retinal gave an RMS difference of 0.31 Å, that is, they were almost identical. This was the expected result, given that the retinal group is critical to the functioning of bR and therefore its geometry would be expected to be highly conserved. Retinal is normally represented by the ionized form, bR(+), with the ionized Lewis structure site being the nitrogen atom. Interestingly, the results of a PM6 calculation indicate that there was almost no charge on the nitrogen, and instead that the charge was delocalized over the extended π system. This delocalization can be rationalized in that there are 11 valid Lewis structures for the Schiff base, each of which puts the positive charge on 1 of the 12 atoms in the conjugated system—no Lewis structure can put the charge on the penultimate carbon, that is, the first carbon of the cyclohexene. As all of these are equally valid, there is no a priori reason to select the nitrogen atom.

The issue of bias arising from the initial choice of geometry was investigated further by examining two high-accuracy structures, 1CBN and 1EJG [38], of crambin. Like 1CBN, 1EJG had positional disorder, and again set "A" was used. An unexpectedly large RMS difference of 0.83 Å was found when the two X-ray structures were compared. Most of this difference was traced to different sequences of atoms in the two files, and after re-sequencing the RMS difference dropped to 0.24 Å. Further examination of the two PDB files showed that the structures corresponded to different conformers: where the structure derived from 1CBN had a methyl group on Ile7, the equivalent methyl group in the 1EJG PM6 structure was rotated about C_β–C_γ by about 120°. This positional disorder was noted in both PDB files. Additionally, and not reported in the PDB files, the structures of Asn46 were different, in that the side-chain O and NH_2 groups were interchanged.

Optimization of both structures resulted in the side chain of Asn46 in 1CBN rotating by almost half a circle. The resulting ΔH_f and optimized geometry were similar to those of the optimized PM6 structure of 1EJG. When the torsion angle of the Asn46 side chain in 1CBN was constrained at the X-ray value and the geometry again optimized, the resulting ΔH_f was several kcal mol^{-1} higher. Based on these results, the obvious deduction can be made that the orientation of Asn46 in 1EJG is correct, and that the orientation in 1CBN is wrong. Additionally, the effect of bias in favor of the original PDB structure, although present, is seen to be quite small—the side chain of Asn46 in 1CBN spontaneously rotated to match 1EJG.

Tertiary structure

In single-chain proteins, the tertiary structure is the three-dimensional geometry of the entire polypeptide. For multi-chain proteins, tertiary structure refers to the structure of each chain. Two main types are of interest here: (1) globular proteins, composed of α helices, β sheets, and turns, with the overall structure being stabilized by weak intra-chain interactions, such as salt bridges, hydrogen bonds and π-bonding, and (2) long, thin, rigid proteins, composed mainly of α helix, and usually consisting of two or more chains.

Because tertiary structure is a property of entire polypeptide chains, a useful measure of the ability of a method to reproduce protein tertiary structure is the RMS difference between the calculated and reference geometries. The RMS differences for a set of proteins is presented in Table 4.

For most proteins, the RMS errors in the PM6 structure are in the order of an Ångstrom, suggesting that PM6 can be used for accurately modeling protein structures, and by implication can be used for investigating protein properties, such as catalysis by enzymes. No obvious dependence of RMS error on size of system was noted. This was unexpected, in that small systematic errors in bond-lengths were expected to exist, and, if present, these would increase the RMS error in proportion to the size of the system.

Quaternary structure

In proteins that consist of two or more chains, the arrangement in the protein of the various sub-units gives rise to the quaternary structure. The structures of several such oligomeric proteins were optimized and the resulting geometries compared with the reference X-ray structure. Despite their size, the RMS differences (Table 5) were comparable with those of simple proteins. Some specific oligomeric proteins of interest are:

Potassium selectivity filter

The potassium channel homotetrameric membrane protein 1JVM [39] forms a filter for migration of potassium ions. In

Table 4 Root mean square errors for PM6 optimized structures of simple proteins. *PDB* Protein data bank

PDB entry	Name	Formula	No. of residues	RMS error (Å)	No. of atoms
1V46	Cardioactive peptide	$(C_{42}H_{56}N_{10}O_{12}S_2)$	9	1.11	244
1A3J	Collagen-like peptide	$(C_8H_{12}N_3O_3)_7$ $(H_2O)_{40}$	21	0.41	371
3ZNF	Zinc finger	$(C_{157}H_{244}N_{48}O_{41}S_3Zn)$	30	2.00	493
1CBN	Crambin	$(C_{202}H_{315}N_{55}O_{64}S_6)$	46	0.91	642
1EJG	Crambin	$(C_{202}H_{315}N_{55}O_{64}S_6)$	46	1.27	642
1EF4	Zinc bundle	$(C_{280}H_{440}N_{81}O_{81}S_6Zn)$	55	2.37	889
2B97	Hydrophobin HFBII	$(C_{305}H_{495}N_{81}O_{92}S_8)$	70	1.15	981
2J7J	Transcription factor IIIA	$(C_{446}H_{672}N_{130}O_{120}S_8Zn_3)(SO_4)(H_2O)_{254}$	85	1.43	1,765
1CAG	Collagen-like peptide	$(C_{359}H_{516}N_{88}O_{118})$ $(Ac)_6$ $(H_2O)_{85}$	88	0.89	1,384
1RNB	Barnase	$(C_{569}$ H_{870} N_{159} O_{194} PS) $(H_2O)_{94}$	109	1.16	2,066
1A62	*Escherichia coli* rho	$(C_{624}H_{998}N_{170}O_{191}Se_3)$ $(H_2O)_{114}$	125	0.82	2,328
135L	Turkey egg white lysozyme	$(C_{611}H_{947}N_{191}O_{182}S_{10})$ $(H_2O)_{114}$	129	1.02	1,183
193L	Hen egg white lysozyme	$(C_{613}H_{951}N_{193}O_{185}S_{10})$ $(H_2O)_{142}$ (NaCl)	129	0.88	2,380
1CPQ	Bacteria cytochrome *c*	$(C_{605}H_{936}N_{160}O_{192}S_6Fe)$ $(H_2O)_{115}$	129	0.80	2,245
1CUO	Bacteria azurin Iso-2	$(C_{605}H_{934}N_{163}O_{192}S_8Cu)$ $(H_2O)_{81}$	129	0.81	2,146
1C7K	Zinc endoprotease	$(C_{619}H_{913}N_{185}O_{208}S_3ZnCa)$ $(H_2O)_{116}$	132	1.15	2,278
1M6C	Pig myoglobin	$(C_{763}H_{1213}N_{209}O_{227}S_3)$ $(C_{34}H_{34}N_4O_4Fe\cdots CO)$ (H_2O_2) $(H_2O)_{272}$	153	1.07	3,312
2F3Y	Calmodulin	$(C_{707}H_{1098}N_{186}O_{251}S_9)$ $(C_{119}H_{179}N_{29}O_{25})(Mg_2Ca_4)(H_2O)_{330}$	164	0.93	3,599
1EZG	Mealworm thermal hysteresis protein isoform YL-1	$(C_{309}H_{479}N_{99}O_{129}S_{16})_2$	164	1.14	2,064
1RCF	Anabaena-7120 flavodoxin	$(C_{856}H_{1298}N_{218}O_{291}PS)(H_2O)_{295}(SO_4)$	169	1.08	3,550
1AMM	Bovine γ-B crystallin	$(C_{926}H_{1396}N_{266}O_{281}S_{14})$ $(H_2O)_{379}$	174	1.09	4,020
2SPC	*Drosophila melanogaster* spectrin	$(C_{1062}H_{1698}N_{320}O_{339}S_6)$ $(H_2O)_{155}$	212	0.87	3,890
1EMA	*E. coli* green fluorescent protein (GFP)	$(C_{1134}H_{1727}N_{293}O_{338}S_2Se_4)$ $(H_2O)_{95}$	225	0.83	3,783
1AT9	Halobacteria bacteriorhodopsin	$(C_{1214}H_{1869}N_{272}O_{303}S_9)$	230	1.26	3,667
1AFQ	Bovine γ-chymotrypsin	$(C_{1096}H_{1730}N_{295}O_{345}S_{12})$ $(C_{22}H_{28}N_3O_2F)_2$ (SO_4) $(H_2O)_{124}$	243	0.90	3,968
2G8C	*E. coli* outer surface protein	$(C_{1112}H_{1887}N_{299}O_{411}S)$ $(C_3N_2H_4)$ $(C_8H_{18}O_5)$ $(C_5H_{12}O_3)$ $(H_2O)_{388}$	246	0.99	4,934
2YPI	Yeast triose phosphate isomerase	$(C_{1196}H_{1895}N_{320}O_{365}S_2)$	247	1.44	3,778
1LXA	UDP N-acetylglucose O-acetyltransferase	$(C_{1237}H_{1972}N_{360}O_{368}S_9)$	262	1.04	3,946
2B9C	α-Tropomyosin	$(C_{1341}H_{2250}N_{384}O_{467}S_6)(H_2O)_{345}$	278	1.42	5,483
4BCL	Bacteriochlorophyll (Fenna-Mathews-Olsen protein)	$(C_{1722}H_{2675}N_{481}O_{514}S_6)(C_{55}H_{74}N_4O_6Mg)_7$ $(H_2O)_{119}$	350	0.99	6,735
1UW8	Oxalate decarboxylase	$(C_{1916}H_{2934}N_{504}O_{587}S_8Mn_2)$ $(C_4H_{11}NO_3)(H_2O)_{387}$	378	1.18	7,131
2V3N	Bovine transcobalamin-2	$(C_{2008}H_{3228}N_{572}O_{586}S_{13})$ $(C_{62}H_{96}N_{13}O_{14}PCo\cdots CO)$ $(Cl)_3$ $(H_2O)_{174}$	405	0.59	7,121
1DT6	Rabbit cytochrome-P450 2C5	$(C_{2309}H_{3631}N_{603}O_{655}S_{22}Fe)$ $(C_{34}H_{34}N_4O_4)$ $(SO_4)_3$	460	0.70	7,312
1WYI	Human triose phosphate isomerase	$(C_{2362}H_{3782}N_{648}O_{723}S_{14})$ $(H_2O)_{449}$	496	0.90	8,876
1DMS	Dimethylsulfoxide (DMSO) reductase	$(C_{3744}H_{5737}N_{998}O_{1120}S_{28})$ (MoO_2) $(C_{20}H_{24}N_{10}O_{13}P_2S_2)_2$ $(H_2O)_{355}$	766	0.71	12,837

it, four residues from each chain, Thr75, Val76, Gly77, and Tyr78, form part of a channel or tube through which alkali ions can migrate. Because a large fraction of each chain consists of α helix, almost all of the hydrogen bonds are intra-chain, and there are very few available for inter-chain hydrogen bonding. The reason for the stability of the tetrameric structure in therefore of interest.

Inspection of the optimized PM6 structure showed that, in addition to the stabilization provided by the metal cations interacting with the hydroxyl of Thr75 and the dipoles of the peptide carbonyls, salt bridges had formed between Arg89 of one strand and Asp80 and Glu71 of the adjacent strand, as shown in Fig. 17. Unlike most other protein salt bridges, the salt bridges in 1JVM involve three residues. In a section of α helix, Glu71 shares a bridging hydrogen atom with Asp80 located in a turn in the same chain, the two-residue assembly having a net unit negative charge. That assembly then forms a salt bridge with an ionized

Table 5 Root mean square errors for PM6 optimized structures of oligomeric proteins

PDB entry	Name	Formula	No. of residues	No. of chains	RMS error (Å)	No. of atoms
1SLK	Silk poly(L-Ala-Gly)	$((C_5H_8N_2O_2)_{32})_{\infty}$	2	∞	N/A	544
1A3J	Collagen-like peptide	$((C_{84}H_{125}N_{21}O_{21})(H_2O)_{40})_{\infty}$	7	∞	0.41	371
1CAG	Collagen-like peptide	$(C_{359}H_{516}N_{88}O_{118})\ (Ac)_6\ (H_2O)_{85}$	88	3	0.89	1,384
1SN9	Tetrameric β-β-α-mini-protein	$(C_{472}H_{740}N_{124}O_{132})\ (H_2O)_{118}$	92	4	1.03	1,822
1EZG	Mealworm thermal hysteresis protein isoform YL-1	$(C_{309}H_{479}N_{99}O_{129}S_{16})_2$	164	2	1.14	2,064
1JVM	Voltage-gated potassium channel	$K_3((C_{462}H_{719}N_{115}O_{120}S_2)\ (C_{470}H_{732}N_{118}O_{123}S_2))_2\ (C_8H_{20}N)(H_2O)$	390	4	1.23	5,749
2V66	*E. coli* nuclear distribution protein nudE-line 1	$(C_{567}H_{919}N_{167}O_{191})_4$	444	4	1.03	7,745
1WYI	Triose phosphate isomerase	$(C_{1181}H_{1896}N_{324}O_{364}S_7)_2(H_2O)_{449}$	496	2	0.98	8,876
2AAI	Ricin	$(C_{2683}H_{4175}N_{733}O_{838}S_{17})\ (C_{12}H_{22}O_{11})_2\ (H_2O)_{123}$	529	2	1.18	8,905
1GZX	Human hemoglobin	$(C_{2818}H_{4390}N_{764}O_{795}S_{12})\ (C_{34}H_{32}N_4O_4Fe\cdots O_2)_4\ (H_2O)_{200}$	574	4	1.89	9,687
1QGK	Importin	$(C_{4288}H_{6853}N_{1145}O_{1330}S_{47})\ (C_{225}H_{396}N_{84}O_{64}S_2)\ (H_2O)_{44}$	921	2	0.87	14,566

Arg89 at the end of an α helix in an adjacent strand. Each of the three residues in this assembly—a proton shared by two carboxylate groups electrostatically interacting with a cationic residue—contribute strongly to the stability of the system.

Ricin

The superpoison ricin is a heterodimeric protein composed of a 32 kDa N-glycoside hydrolase ribosome-inactivating enzyme, the "A" chain, joined by a disulfide bond to a 34 kDa lectin, the "B" chain. At 529 residues, ricin is one of the larger proteins; nonetheless, like most of these proteins, it is a relatively simple system, in that, with the exception of two residues on chain B, Asn95 and Asn135, which are glycosylated, it is composed only of residues of the 20 common amino acids. From a computational perspective, one of the interesting features of ricin is the nature of the interface between the two chains. The most suitable entry in the PDB for studying this system was 2AAI [40], which contained both the A and B chains, residues of 14 saccharide moieties, and 123 water molecules.

As a result of its size, geometry optimization was slower than for most other proteins. Each geometry optimization cycle required only about 15 CPU minutes; however, an unusually large number of optimization cycles were required in order to reduce the gradient to 15 kcal mol^{-1} $Å^{-1}$, and this resulted in the entire optimization requiring almost 10 CPU days. To verify that the minimum had, in fact, been reached, the optimization was continued, and the gradient reduced further to 9 kcal mol^{-1} $Å^{-1}$. This required an additional 3 CPU days of effort, and resulted in the ΔH_f decreasing by 26.8 kcal mol^{-1} to −51071.5 kcal mol^{-1}, that is, a drop of 0.003 kcal $atom^{-1}$. The two geometries differed by a RMS of 0.15 Å, and the RMS error between the calculated and X-ray geometries increased from 1.18 Å to 1.23 Å. Based on these results, the decision was made that further work was not justified.

The optimized PM6 structure agreed well with the X-ray structures (Fig. 18) lending support to the validity of the description of the interface between the two chains. At the interface (Fig. 19), in addition to the disulfide bridge, PM6 predicted the existence of a salt bridge between His40 of chain A and Asp94 of chain B. PM6 predicted that 11 other

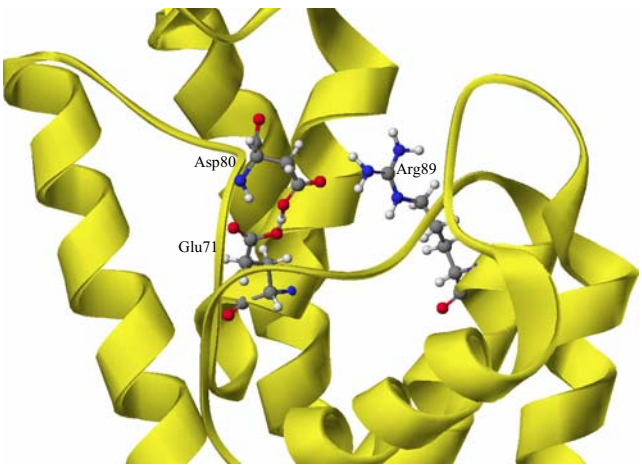

Fig. 17 Salt bridge in potassium channel membrane protein 1JVM. Glu71 and Asp80 have a net charge of −1

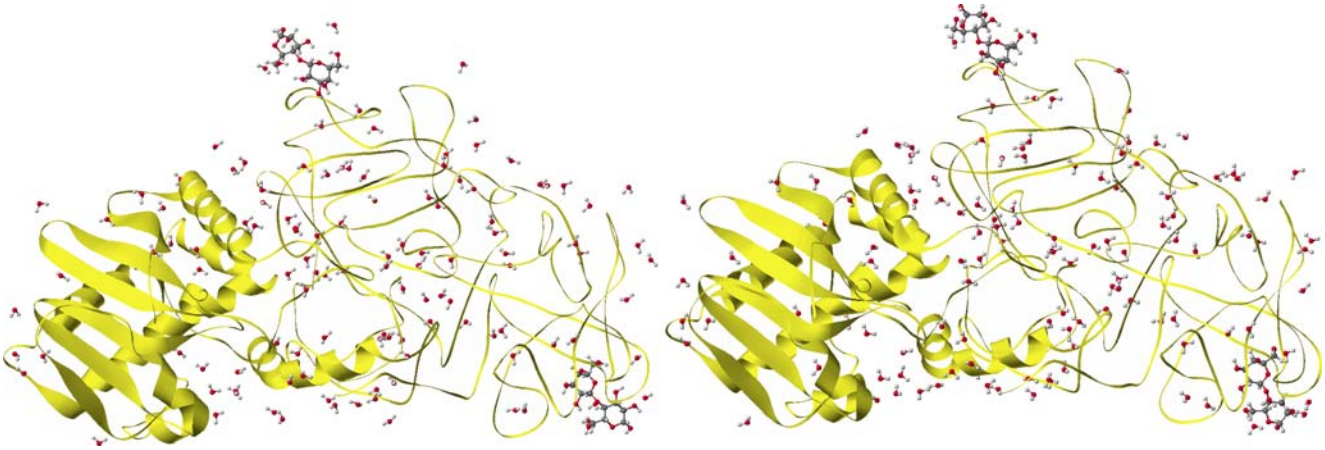

Fig. 18 X-ray and PM6 structures of ricin, 2AAI. *Left* X-ray structure, *right* PM6

salt bridges should also exist, 3 in chain A and 8 in chain B. Examples of these bridges are, in chain A, between Asp37 and Arg39, and in chain B between Lys219 and Asp221. The large number of salt bridges prompted speculation as to the existence of more inter-chain interactions. To investigate this, other possible salt bridges between the chains were artificially constructed by moving a proton from one chain to the other, but, in each case, optimization of the structure resulted in either the proton migrating back or in an increase in heat of formation, indicating that only one salt bridge connects the two chains.

Hemoglobin

Human oxy-hemoglobin was one of the larger proteins studied. It consists of a pair of two α and two β sub-units, each of which contains a heme unit, to give the tetramer $\alpha_2\beta_2$. A consequence of the quaternary structure is the

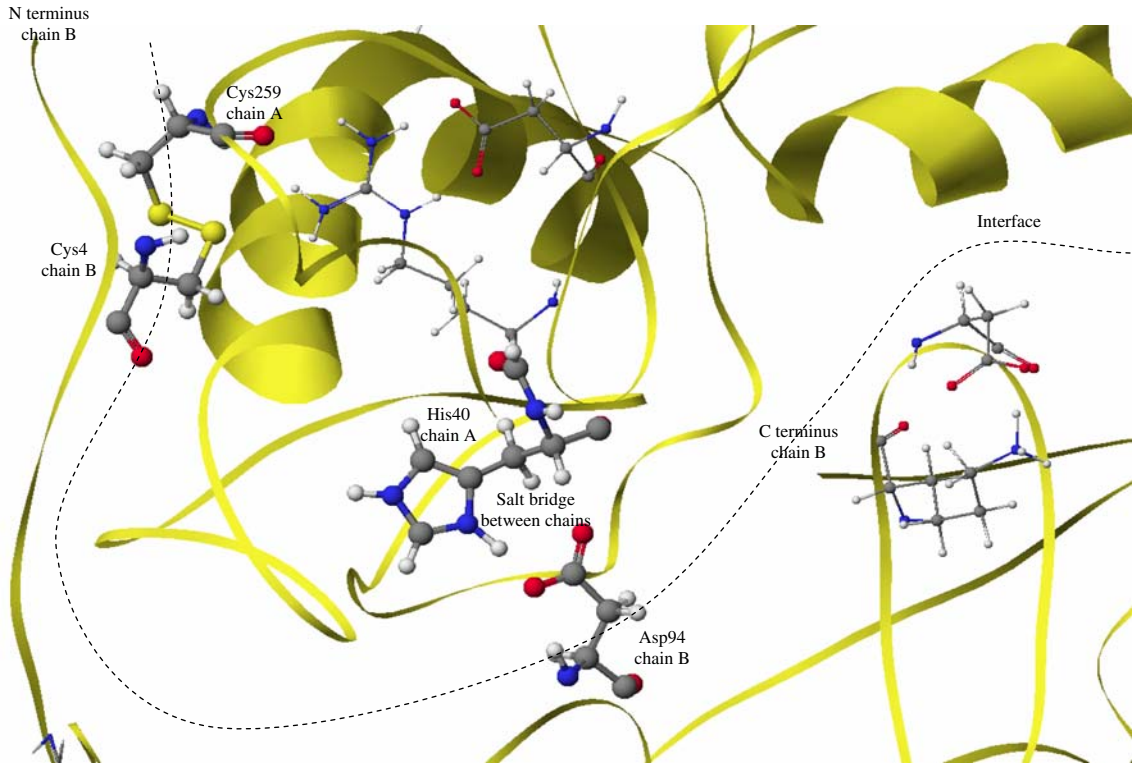

Fig. 19 Interface between chains A and B in ricin showing disulfide bridge and inter-chain salt bridge. Also shown are salt bridges in chain A (Asp37–Arg39) and chain B (Lys219–Asp221)

allosteric behavior of hemoglobin: as each heme becomes oxygenated, the remaining heme rings increase their affinity for oxygen.

The X-ray structure of hemoglobin 1GZX [41] included 205 oxygen atoms representing water molecules; these, together with the four sub-units were used in the simulation. The resulting optimized structure and original X-ray structure are shown in Fig. 20.

An attempt was made to model the allosteric properties of hemoglobin, but the results were inconclusive.

Metalloproteins

In addition to the common main-group elements, many proteins also contain metal atoms. These metals can be either covalently or ionically bound to side chains, to non-protein moieties, such as porphyrin ring systems, or they might be ionized and as a result would be highly mobile. Biochemical activity, such as enzymatic catalysis or charge transfer across cell membranes, often depends on the immediate environment of the metal atom. Modeling metalloproteins using earlier semiempirical methods has not been very successful, so the degree to which PM6 can reproduce the environment of the various metal atoms is of obvious interest.

Magnesium

The photosynthetic system of the pigment bacteriochlorophyll-A, 4BCL [42], contains seven chlorophyll molecules, and thus seven magnesium atoms. This protein is unusual in that a large fraction of its surface is composed of β antiparallel sheet, underneath which are the chlorophyll molecules in close proximity, i.e., they form a compact cluster.

Minor problems were encountered during preconditioning, in that the Lewis structure generated by the MOZYME procedure for the chlorophyll molecules was different to the conventional structure, but as the computed structure was nevertheless a valid Lewis structure, and as there was no overriding reason to use the conventional one, the computed Lewis structure was used to start the SCF procedure. This resulted in no complications. Preconditioning was completed with the optimization of the positions of the added hydrogen atoms.

The optimized and X-ray structures are shown in Fig. 21, and a detail of one of the chlorophyll molecules is given in Fig. 22. Other than a minor rotation of an ester group (top left of each structure), the environment of the magnesium atom was reproduced with good accuracy.

Potassium

The metal atom in potassium-containing proteins is invariably almost completely ionized, and therefore does not form covalent bonds. Instead, it is free to migrate within cavities in the protein. One such structure consists of potassium ions in a pore or channel in a transmembrane protein; such proteins are usually involved in regulating the electrical potential across the cell wall.

An example of such a potassium channel membrane protein is the 1JVM protein mentioned earlier; its four chains form a channel in the center of the protein that contains three metal ions, a tetrabutylammonium ion, and a

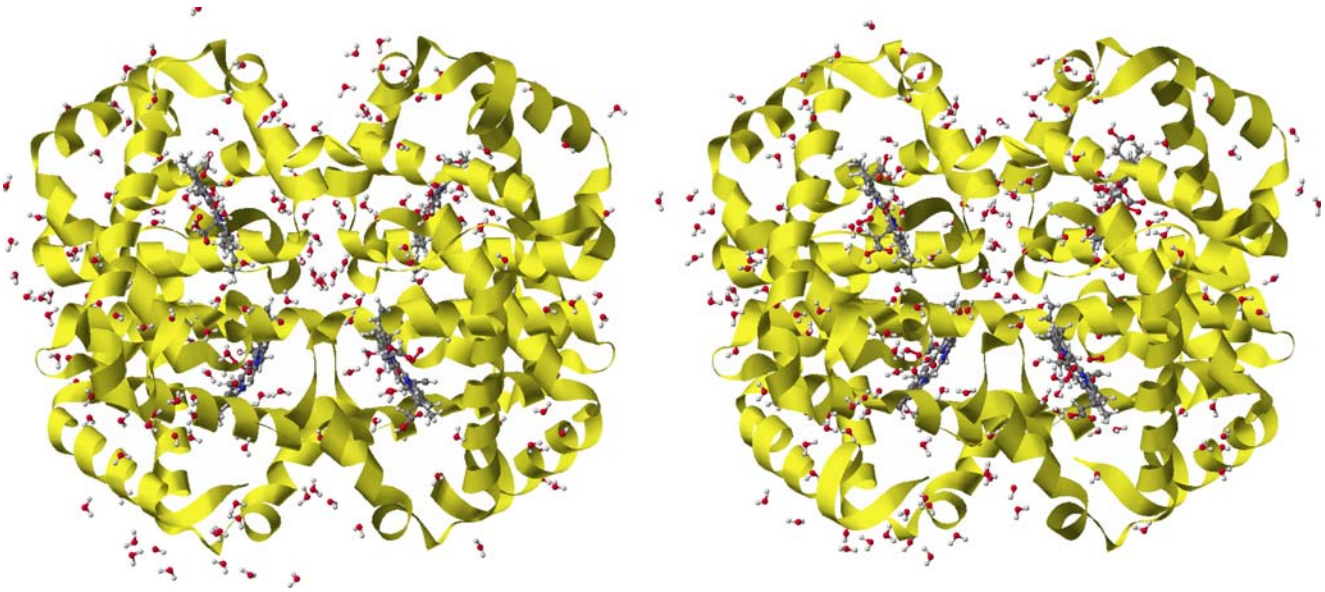

Fig. 20 X-ray and PM6 structures of hemoglobin, 1GZX. *Left* X-ray structure, *right* PM6

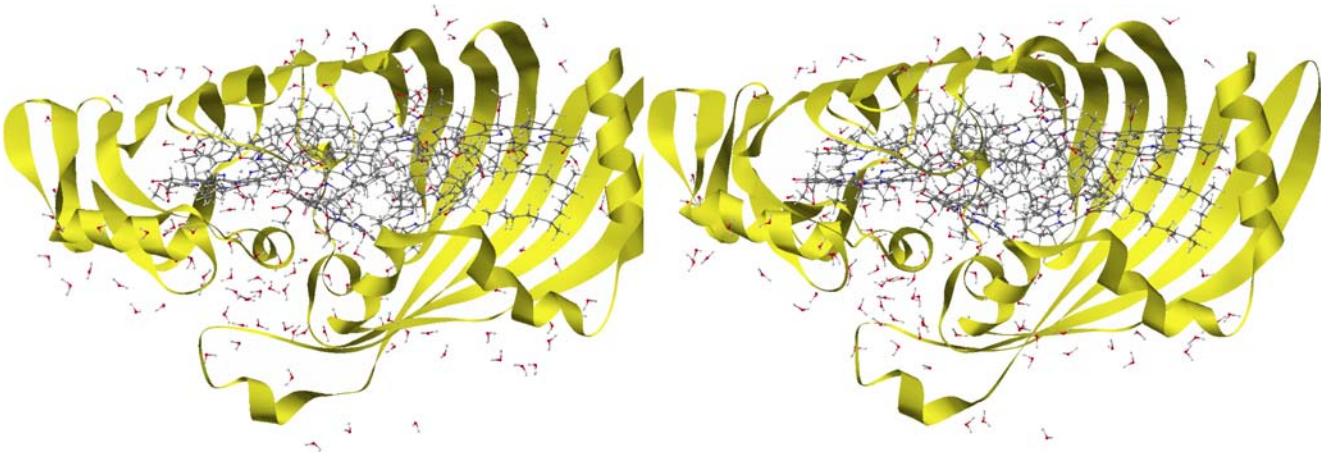

Fig. 21 X-ray and PM6 structures of bacteriochlorophyll, 4BCL. *Left* X-ray structure, *right* PM6

water molecule. In the PDB file, all three K^+ ions were replaced by Rb^+ ions, that is, by ions of similar type and size, but as the purpose of this study is to investigate systems that can occur in vivo, and because naturally occurring proteins do not normally contain significant amounts of Rb^+, before any work was done, the Rb^+ ions were replaced once again by K^+ ions.

The protein was preconditioned and given a net charge of +4, reflecting the charges on the three K^+ ions and on the included tetrabutylammonium ion. Despite the large net charge, geometry optimization proceeded smoothly, terminating in a structure very similar to that in the PDB file, the RMS difference being 1.23 Å, as shown in Fig. 23. In 1JVM, the alkali metal ions are in an approximately square antiprism environment coordinated by eight oxygen atoms. This structure was reproduced with good accuracy, as is shown in Fig. 24. In agreement with the X-ray structure, the central cation coordinates with Val76 and Gly77 in each of the four chains. The center of the tetrabutylammonium ion and the water molecule had both drifted about 0.3 Å off-axis, suggesting a possible fault in the PM6 method.

Calcium

Calcium is a highly electropositive metal that invariably exists in biochemical systems as the formal dipositive ion coordinated by oxygen atoms. An example of a calciferous protein is calmodulin, 2F3Y [43], which contains four calcium ions, each of which is bound to a helix-turn-helix assembly or EF-hand in the chain. Each EF-hand consists of about 12 residues, 4 of which bind to the calcium ion using side-chain oxygen atoms, and 1 binding through the peptide carbonyl. A representative EF-hand is DKDGDGTITTKE, in which the calcium-binding residues are Asp20, Asp22,

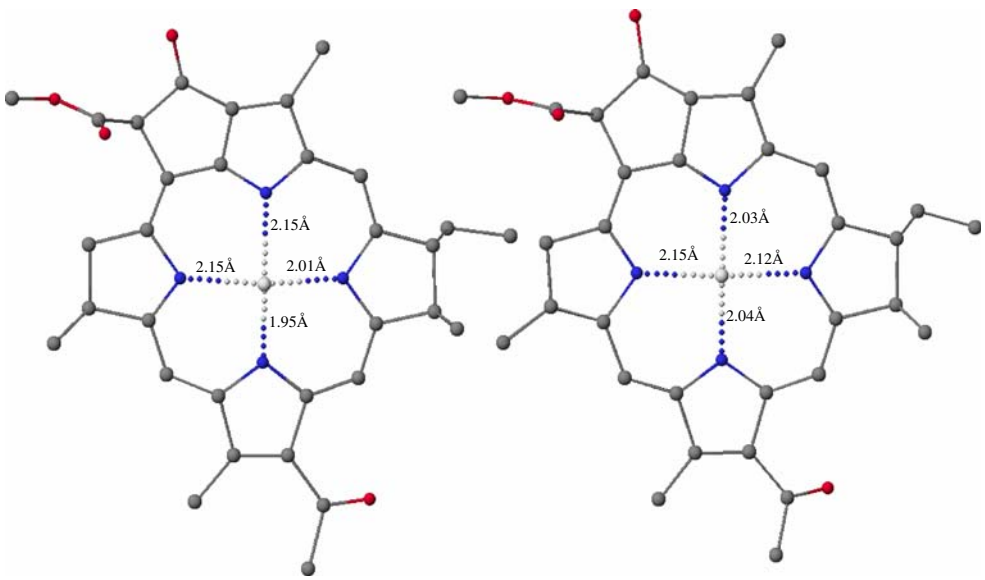

Fig. 22 Detail of one of the chlorophyll molecules in bacteriochlorophyll, 4BCL, showing the environment of magnesium

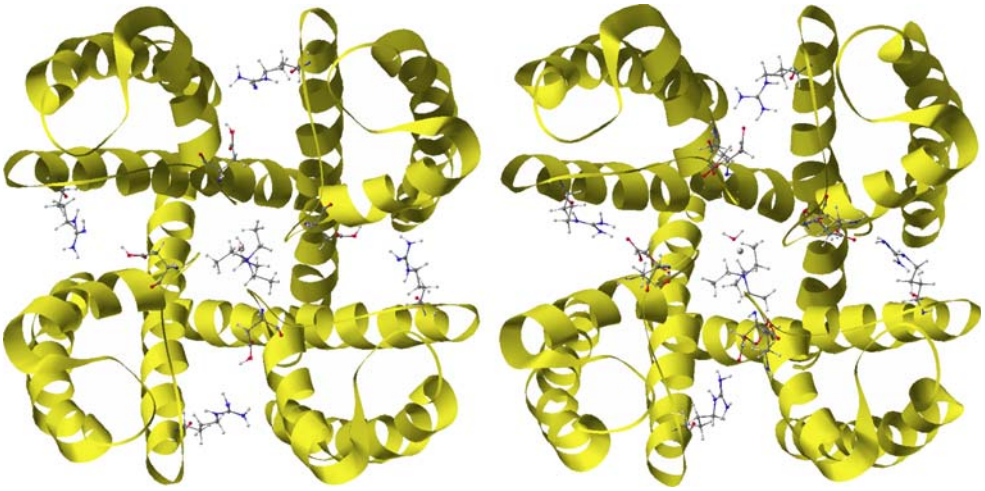

Fig. 23 X-ray and PM6 structures of potassium channel membrane protein 1JVM. *Left* X-ray structure, *right* PM6

Asp24, Thr26, and Glu31. In this EF-hand, calcium also coordinates to a water molecule. The EF hand is reproduced with low accuracy by PM6 (Fig. 25), one residue, Asp20, being only weakly bound in the PM6 structure. That the coordination complex formed by calcium is not purely ionic is indicated by the reduced charge on calcium, +1.2, suggesting that a substantial amount of covalent bonding exists between the calcium and the oxygen atoms.

Calcium also occurs in the endoprotease 1C7K [44] found in *Streptomyces caespitosus*, where it binds to the backbone through Asp76 and Thr78. The environment of calcium in this protein is reproduced with reasonable accuracy, with the calcium to Thr78 side chain oxygen distance being 2.50 Å, only slightly larger than the X-ray value of 2.45 Å, but the optimized Asp78 carboxylate oxygen to calcium distance was 2.26 Å, substantially smaller than the reported 2.53 Å. Calcium also bonds to four water molecules, with PM6 predicting the Ca–O distance to be 2.41 Å, in good agreement with the reported 2.35–2.51 Å. Although the calcium ion is near to only two backbone residues, the carbonyl oxygen on Thr113 forms a hydrogen bond with one of the water molecules in the first coordination shell of calcium, so the calcium atom could be regarded as also bonding to a third residue.

Manganese

Oxalate decarboxylase, 1UW8 [45], which catalyzes the reaction of oxalate to give a formate ion and carbon dioxide, contains two manganese binding sites, the second of which is postulated[46] to be the active site. In this site, Mn binds to His273, His275, Glu280, and His319. After preconditioning, the structure of 1UW8 was optimized, as shown in Fig. 26. PM6 predicts the Mn–O distance with good accuracy (2.02 Å vs the X-ray 2.09 Å) but underestimates the Mn–N distance, averaging 1.82 Å vs the X-ray 2.23 Å.

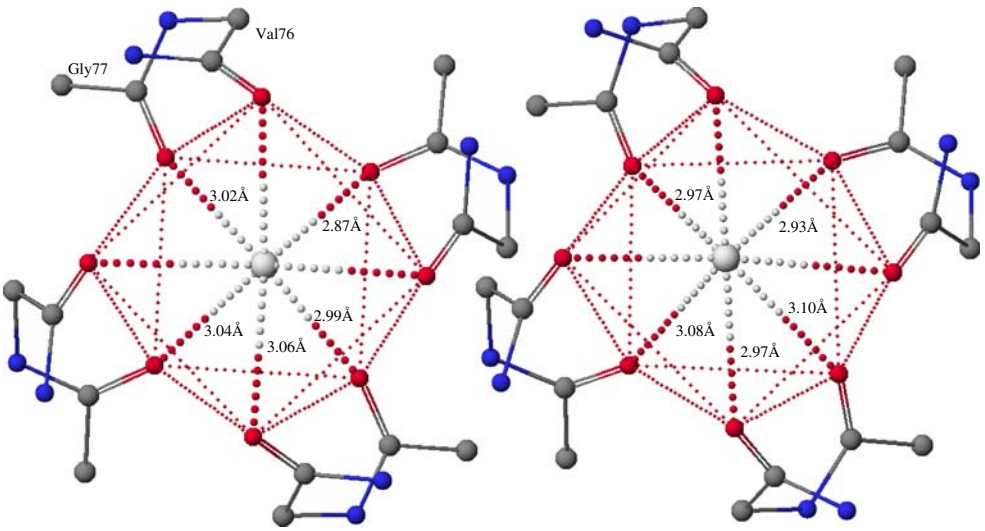

Fig. 24 Environment of middle potassium ion in potassium channel filter protein 1JVM. *Left* X-ray structure, *right* PM6

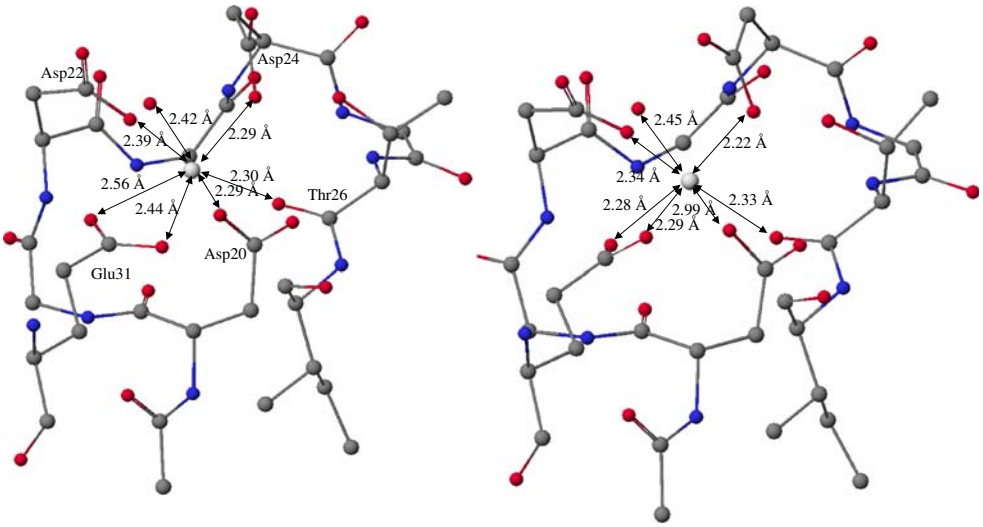

Fig. 25 X-ray and PM6 structures of an EF hand in calmodulin, 2F3Y, binding calcium ion. *Left* X-ray structure, *right* PM6

Iron

Iron is one of the most important metals in the biochemistry of the animal kingdom; iron-containing proteins are almost ubiquitous in oxygen-breathing creatures. In its most common form, iron is the central atom in a porphyrin heterocyclic ring system, the heme molecule. A wide range of proteins containing heme molecules are known, and the applicability of PM6 to model such systems is of obvious interest. A potential problem in modeling heme-containing proteins was anticipated as a consequence of the restriction of the current MOZYME technique to closed-shell systems. Iron atoms in the heme system have formal oxidation states Fe(II) or Fe(III), and therefore are almost certainly open shell. Nevertheless, structures of various proteins containing heme were optimized without problems.

The smallest heme-containing system modeled was cytochrome *c*, 1CPQ [47]. In this protein, the rim of the heme ring is covalently bound to residues Cys118 and Cys121, and, in addition to bonding to the four nitrogen atoms of the porphyrin ring, the iron atom is also coordinated to a nitrogen on the heterocyclic ring of His122. These three bonds to the heme system effectively hold it rigidly within the protein. When the structure of 1CPQ was optimized, the changes in geometry of the entire protein, and of the heme system in particular, were quite small, with the RMS error for the entire heme system plus nearby atoms being 0.40 Å; errors for individual N–Fe distances are shown in Table 6. In all heme systems studied, the iron–heme nitrogen distances were similar to those given here.

The environment of the iron atom in the slightly larger protein carboxy-myoglobin, 1 M6C [48], is similar to that of cytochrome *c*, except that in addition to the five nitrogen atoms coordinating to the iron, the sixth coordination site is occupied by a carbonyl ligand. In the optimized PM6 structure, the iron remained octahedrally coordinated, and the Fe–CO bond length, at 1.75 Å, was essentially unchanged from that in the X-ray structure, 1.78 Å.

In the X-ray structure, residues Glu18 and Lys77 are far apart in the backbone, but because of the folding of the chain they are relatively near in space, with an oxygen atom of the carboxylate on Gly18 being only 2.56 Å from the amine nitrogen on Lys77. This very small interatomic separation strongly suggests a bridging hydrogen atom instead of a salt bridge. This conjecture was reinforced

Fig. 26 X-ray and PM6 structures of oxalate decarboxylase, 1UW8. *Left* X-ray structure, *right* PM6

Table 6 Bond lengths to iron in heme in cytochrome *c*, 1CPQ

	X-ray	PM6
Ring N_1-Fe	1.98	2.01
Ring N_2-Fe	1.96	2.00
Ring N_3-Fe	1.99	1.98
Ring N_4-Fe	1.98	1.98
His_{122} N-Fe	2.01	1.99

when, during preconditioning, a hydrogen atom positioned itself approximately midway between the oxygen and nitrogen atom; after global optimization the relative positions of the atoms in the O–H–N complex (Fig. 27) did not change significantly. The presence of such a structure implies a highly stabilizing interaction between the two residues, so the two residues can be regarded as being connected, the net effect being a strong connection or link between two well-separated sections of the backbone chain.

Cytochrome-P450 forms a very large and diverse set of hemoproteins. A typical example of such a system is the rabbit cytochrome-P450, 1DT6 [49]. It is structurally similar to cytochrome *c* and myoglobin, the main difference being that, instead of a histidine nitrogen, the fifth coordination site consists of a sulfur atom from Cys432 forming an axial thiolate bond to the iron atom. At 1.82 Å, PM6 significantly underestimates the Fe–S distance, the X-ray structural value being 2.46 Å. This large error might be attributable to the closed-shell method used, in that the Restricted Hartree-Fock (RHF) calculation requires a covalent single bond to exist between the sulfur and the iron atoms, but this conjecture cannot be tested until either a configuration interaction or unrestricted Hartee-Fock method is available.

With four heme molecules, per-oxy-hemoglobin was the largest iron-containing protein studied. In 1GZX each iron atom is octahedrally coordinated: four bonds to the nitrogen atoms of the porphyrin ring as usual, with one bond to a nitrogen in the imidazole ring of His87, and one bond to an oxygen atom of molecular oxygen. As with the other heme systems, PM6 reproduces the Fe–N of the porphyrin ring with good accuracy, but it underestimates the Fe–N distance, at 1.98 Å compared to the X-ray 2.26 Å. Conversely, the iron–molecular oxygen interaction is slightly over-estimated at 1.71 Å compared to the X-ray 1.61 Å.

Cobalt

Transcobalamin is a relatively large protein that contains the cobalt-containing molecule cobalamin (vitamin B12). This molecule is similar to heme in that it consists of a porphyrin ring, with cobalt in the center instead of iron. However, unlike heme, where the ring system is almost planar, the ring system in cobalamin is highly buckled. In the transcobalamin complex 2V3N [50], the fifth coordination site of cobalt is occupied by a nitrogen atom of dimethylbenzimidazole, and the final, sixth, site is occupied by a carbonitrile ion. The cobalamin molecule is tightly bound inside the protein, despite the fact that there are no covalent bonds connecting it to the polypeptide. The geometry of 2V3N was first preconditioned then optimized. Only relatively small changes resulted from the optimization, as shown in Fig. 28. Most of the environment of the cobalt atom was reproduced with good accuracy (Fig. 29), the exception being the fifth coordinate site, where the cobalt–nitrogen distance was underestimated by 0.15 Å.

Copper

As its name suggests, azurin iso-2, 1CUO [51], is a blue protein, consisting of 129 residues and a copper atom that forms strong bonds with three residues, His46, His117, and Cys112, and longer, and consequently weaker, bonds to Gly45 and Met121. Because it contains Cu(II), a d^9 system, azurin is unambiguously an open-shell system; therefore, like the heme-containing proteins, it was not initially expected to be readily modeled using a closed-shell system. But, as with heme, when the geometry of 1CUO was optimized using the MOZYME technique, the bond lengths of the strong metal bonds were accurately reproduced (Fig. 30). Unfortunately, some of the weaker bonds were essentially destroyed: PM6 incorrectly predicted that the Gly45 Cu–O distance would increase from 3.34 Å to 4.05 Å and the Cu–S bond to Met121 would increase from 2.99 Å to 5.45 Å.

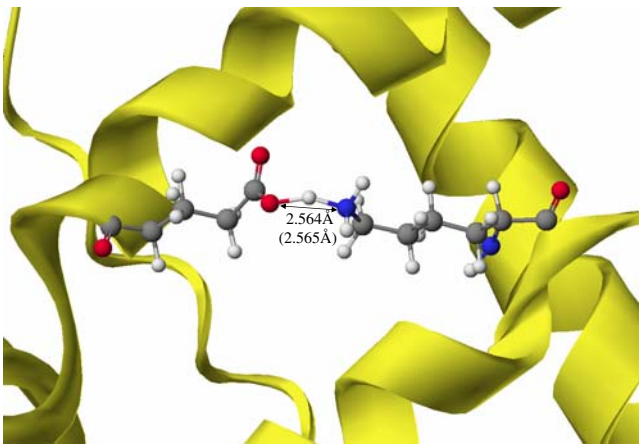

Fig. 27 Example of bridging hydrogen bond in myoglobin, 1M6C

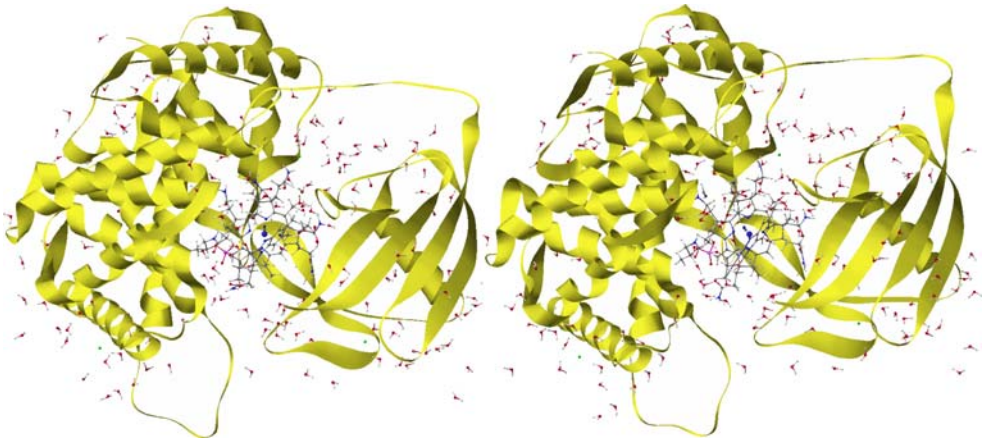

Fig. 28 Comparison of PM6 and X-ray structures of transcobalamin, 2V3N. *Left* X-ray structure, *right* PM6

Zinc

Zinc-containing proteins frequently involve zinc tetrahedrally coordinated to nitrogen and sulfur atoms. A typical structural motif is the zinc finger, in which a zinc atom binds to two cysteine residues in an antiparallel β sheet, and to two histidine residues in an α helix. A recently published structure of transcription factor IIIA, 2J7J [52] from *E. coli*, contains three zinc fingers in a single protein chain, involving residues Cys4, Cys9, His22 and His26; Cys34, Cys39, His52, and His56; and Cys61, Cys67, His80, and His85. Although the RMS error of the optimized PM6 structure was larger than that for similar proteins, the structures of the zinc fingers were reproduced with good accuracy (Fig. 31). PM6 predicted the Zn–S distance to range from 2.23 Å to 2.28 Å, averaging 2.25 Å, in exact agreement with the X-ray average, after one X-ray Zn–S distance of 2.93 Å was excluded on the grounds that it was unrealistically long. The predicted Zn–N distances ranged from 1.97 to 1.99 Å, somewhat shorter than the X-ray average of 2.13 Å, again after exclusion of one unrealistically short X-ray Zn–N distance of 1.53 Å.

Two zinc fingers, from *Methanobacterium thermoautotrophicum*,1EF4 [36], and from the human enhancer binding protein, 3ZNF [53], had unusually large RMS errors, 2.35 Å and 2.00 Å, respectively. These large errors can be attributed in part to the large positive charges on the proteins, and in part to the lack of weak bonds connecting distant parts of the polypeptide chain. In order to allow a realistic calculation to be performed, protons were removed from ionized nitrogen atoms until the systems were neutral. Despite the large RMS distortion, the immediate environment of the zinc atom was accurately reproduced, PM6 predicting the Zn–S distances in 1EF4 to range from 2.27 to 2.48 Å versus the reported NMR values of 2.22 to 2.32 Å, while in 3ZNF the Zn–S distances were predicted to be 2.24 Å versus the NMR structural value of 2.30 Å, and the Zn–N distances averaged 1.98 Å versus the NMR 2.00 Å.

Selenium

Although selenium is not a true metal, it is convenient to consider it here. Selenium-containing proteins are relatively rare, with one of the smaller proteins being the transcriptional

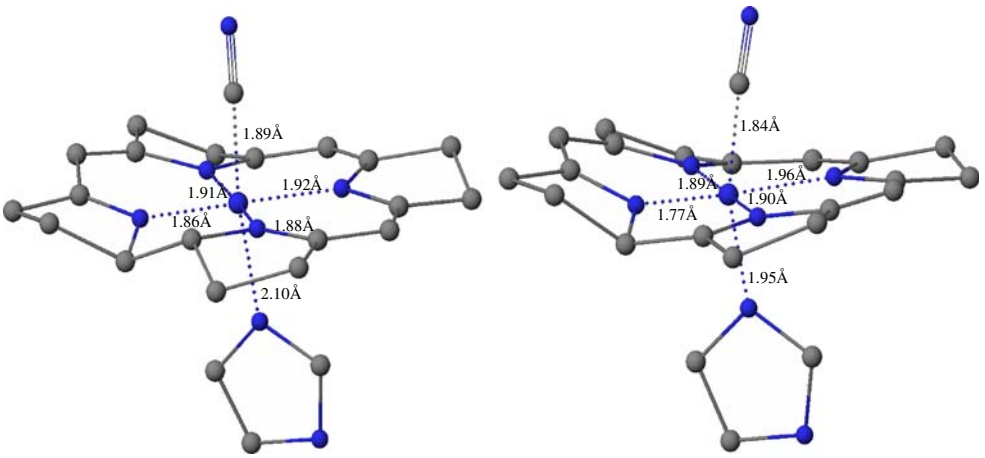

Fig. 29 Detail of transcobalamin showing environment of cobalt atom. *Left* X-ray structure, *right* PM6

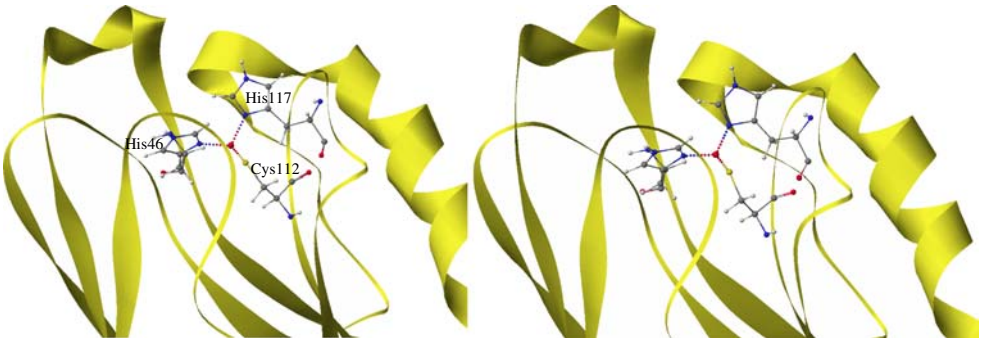

Fig. 30 Comparison of PM6 and X-ray environment of copper atom in azurin iso-2, 1CUO. *Left* X-ray structure, *right* PM6

terminator protein rho, 1A62 [54]. Rho contains three selenomethionine, $–CH_2–CH_2–Se–CH_3$, groups, in which selenium forms normal covalent single bonds with two adjacent aliphatic carbon atoms: that is, selenium is in a common chemical environment. After preconditioning, geometry optimization proceeded without problem, and resulted in a RMS difference of the X-ray and PM6 structures of 0.82 Å.

A comparison of the selenomethionine geometries showed a wide range of Se–C distances in the X-ray structure, from 1.78 to 2.06 Å, where PM6 gave a narrower range, from 1.95 to 1.97 Å, close to the 1.93±0.02 Å expected for organo-selenium compounds. The C–Se–C angles were similar, averaging 100.2° for the X-ray structure versus 98.1° for the PM6 structure.

Molybdenum

Molybdenum is the only second-row transition metal that is required by most living organisms [55]. An example of a molybdenum-containing molecule is the enzyme cofactor molybdopterin, in which it exists as Mo^{IV} or Mo^{VI}, depending on the nature of its ligands. In molybdopterin, the molybdenum atom forms strong bonds to two oxygen atoms, to form the very stable MoO_2 moiety, and weaker bonds to two sulfur atoms. An example of a molybdopterin-containing protein is the enzyme dimethyl sulfoxide (DMSO) reductase, 1DMS [56], from *Rhodobacter capsulatus*.

In the PDB entry for 1DMS, positions of heavy atoms in 766 of the 781residues were given, and all of these were used in the PM6 calculation. The stated formal oxidation state of molybdenum in 1DMS is four, therefore during preconditioning hydrogen atoms were added to both sulfur atoms, so that they would form dative rather than simple covalent bonds to the molybdenum, and thus not contribute to the formal oxidation state. In the native structure, an ionizable oxygen atom on Ser147 is positioned near to the molybdenum atom. As the Mo–O distance is only 1.96 Å, the assumption was made that this oxygen atom was ionized, i.e., unprotonated.

Because 1DMS is larger than most enzymes, the memory requirements were also unusually large. As a result, in order to allow the geometry optimization to be run, it was necessary to reduce the value of CUTOFF from 9 Å to 8 Å, this resulting in a significant reduction in the size of arrays used. With this one change, optimization of 1DMS proceeded without difficulty.

A comparison of the native and PM6 optimized geometries (Fig. 32) shows that both the overall structure and the immediate environment of the molybdenum were reproduced with good accuracy. This accuracy is easily sufficient to allow the various steps in the enzyme catalyzed reduction of DMSO to dimethyl sulfite to be modeled, but because the system is so large, at over 12,800 atoms, such work should be regarded as only marginally practical using current readily available computers,.

Reactions

Enzymes catalyze reactions by lowering the activation barrier, that is, the heat of formation of the transition state relative to that of the reactants. Although this is a very

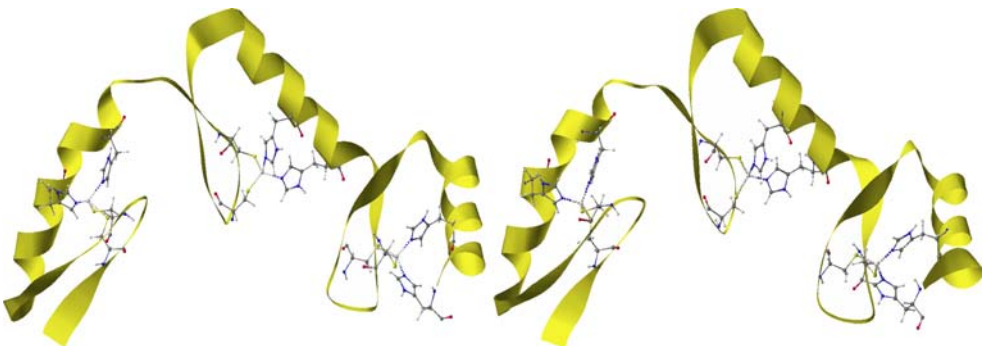

Fig. 31 Comparison of PM6 and X-ray structures of PDB 2J7J, showing the three zinc fingers. *Left* X-ray structure, *right* PM6

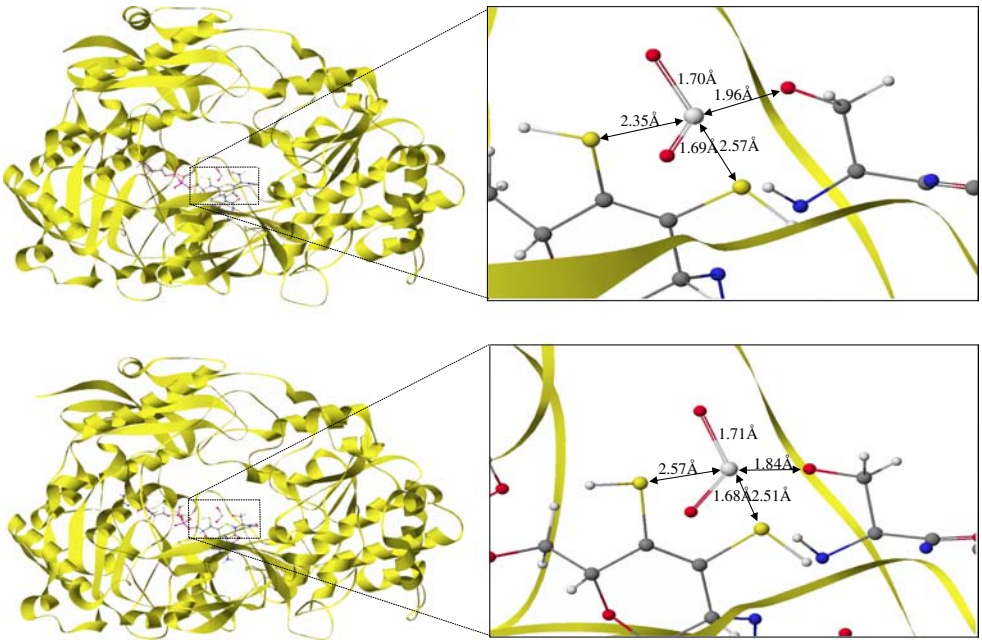

Fig. 32 Comparison of X-ray and PM6 structures of dimethyl sulfoxide (DMSO) reductase, 1DMS, showing environment of molybdenum atom (*insets*).*Top* X-ray structure, *bottom* PM6

important field of biochemistry, computational chemistry tools have hitherto enjoyed only limited success, partly because of limitations in available computational chemistry modeling tools. Reactions cannot be modeled by molecular mechanics methods because such processes involve electronic changes; conversely, while ab initio methods can correctly represent such phenomena, they are impractical because of the prohibitive amount of calculation involved. Although earlier semiempirical methods were much faster, and could model the electronic processes involved, the computational effort was still large, and, even if the calculations were done, the accuracy of the results was insufficient to allow much confidence to be placed in them.

In this work, the ability of PM6 to model structural features of proteins has been demonstrated, and, by using the MOZYME localized molecular orbital method, the computational effort has been significantly reduced, so determining the applicability of PM6 to the study of enzyme reaction mechanisms is appropriate.

A suitable, simple, test case for this work is the formation of a tetrahedral intermediate in the active site of chymotrypsin. Chymotrypsin is a proteolytic enzyme that catalyzes the hydrolysis of peptide linkages, in particular those involving aromatic residues. The generally accepted reaction mechanism involves a triad of residues, His57, Asp102, and Ser195, in which a positive charge, a proton, is shuttled from the serine via the histidine to the aspartate, with the result that the energy required for ionizing the serine is reduced. In turn, this directly lowers the activation barrier for hydrolysis, thus giving rise to the catalytic activity. For the mechanism of the catalytic triad to work, the three residues involved must be positioned precisely in that, if they were too far apart, the barrier to proton migration would be insurmountable.

The objective of this investigation is to determine the validity and ease of modeling enzyme reactions in general, so the precise reaction to be modeled is of little importance. Because of this, and because an X-ray structure of chymotrypsin containing a docked substrate, albeit an inhibitor, was available, the reaction chosen for study was the formation of the tetrahedral intermediate resulting from a hypothetical reaction of D-leucyl-L-phenylalanyl-P-fluorobenzylamide with the active site of bovine γ-chymotrypsin. The starting point for this reaction is represented in the PDB by 1AFQ [57], and consists of 243 of the residues of chymotrypsin, a molecule of the substrate D-leucyl-L-phenylalanyl-P-fluorobenzylamide docked in the active site, a second molecule of substrate loosely associated with residues Cys122, Leu123, and Pro124, 124 water molecules, and a sulfate ion, for a total of 3,967 atoms. Preconditioning consisted of neutralizing all ionizable sites, except Asp102, and the addition of hydrogen atoms as necessary. Optimization of the structure of the enzyme plus substrate was then performed. The X-ray and optimized PM6 structures are shown in Fig. 33.

Examination of both the preconditioned and optimized structures shows that the Ser195 hydroxyl oxygen is much nearer than the ionizable nitrogen on His57 to the peptide nitrogen of the substrate. Based on this, hydrogen migration from the serine hydroxyl to the substrate peptide nitrogen would appear to occur more readily than the conventional mechanism in which the hydrogen first migrates to His57 and only later moves on to the peptide bond being hydrolyzed. Therefore, instead of the conventional first

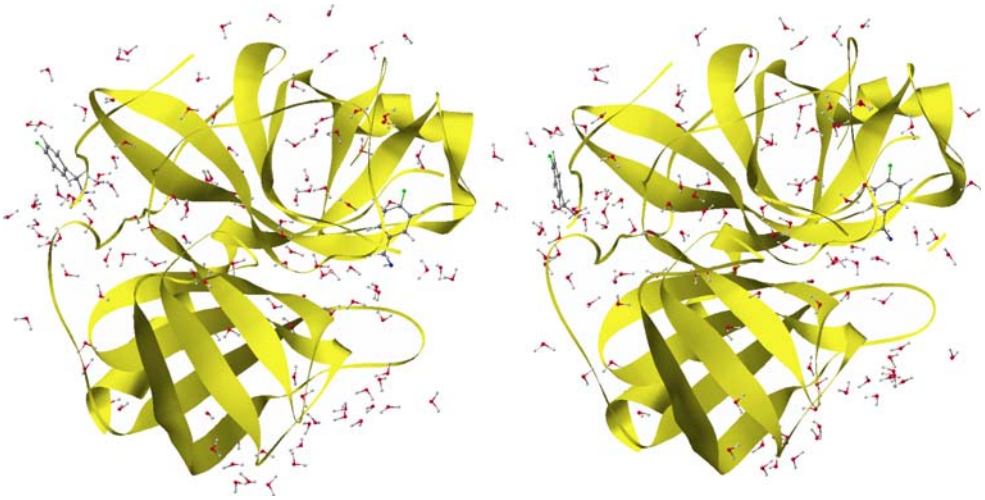

Fig. 33 X-ray and PM6 structures of chymotrypsin, 1AFQ. *Left* X-ray structure, *right* PM6

step in the charge relay mechanism being used, the reaction investigated here was that shown in Fig. 34, the migration of a proton from the hydroxyl to the amide nitrogen and the formation of a tetrahedrally coordinated carbon.

An approximate structure of the tetrahedral intermediate was constructed by moving the hydroxyl hydrogen atom to the region of the amide nitrogen, and reducing the oxygen–carbon separation to that of a C–O single bond. The expected product from optimization of the geometry of the intermediate was the ester and amine, but the actual structure obtained as a result of energy minimization was the tetrahedral intermediate, i.e., the tetrahedral intermediate was a stationary point on the PES. Geometry optimization of much simpler systems containing the same structural feature resulted in decomposition to ester plus amine, from which it can be inferred that the tetrahedral intermediate was being stabilized by its environment. PM6 has a known fault in that it over-stabilizes Zwitterions, which reduces the likelihood of the tetrahedral intermediate being a minimum on the PES. However, for the purpose of this work, the presence of a stationary point on the PES immediately after the transition state was serendipitous in that it yielded well-defined points on both sides of the transition state.

The prospect of attempting to locate even an approximate structure for the transition state appeared to be a formidable task; conventional methods, such as the SADDLE technique [58], synchronous transit [59], assuming a reaction path, etc, would all require an excessive computational effort. In an attempt to simplify this task, a modified synchronous transit method was used.

In the synchronous transit method, the assumption is made that the geometry of a transition state is intermediate between those of the reactant(s) and product(s). If, as is the case here, the reactants and products are very different, in that all the atoms involved in the system are moving, the fact that the assumption might be true still does not provide a reliable guide as to how to locate the transition state. However, by modifying the geometry in such a manner as to reduce the difference in the geometries, while at the same time ensuring that they are on opposite sides of the transition state, the size of the domain in which the transition state is located can be reduced. This result can be achieved by a constrained optimization of the type described below for improving the geometries of X-ray structures. That is, the reactant geometry is re-optimized using a penalty function whose value depends on the difference between the current geometry and that of the product. The same procedure can then be performed on the product geometry. The resulting geometries are then used in a second, and sometimes a third, optimization, culminating in two geometries that are near to the points of inflection on the reaction PES. At that point the two heats of formation are similar, and a good approximation to the transition state can then be obtained by averaging the two geometries. A good approximation to the transition state for the chymotrypsin reaction was obtained using this method.

Conventional procedures are unsuitable for refining transition states of large systems. Thus, in one of the more efficient methods, Baker's Eigenfollowing technique [18], the gradient norm is minimized using a quasi-Newton minimizer that requires evaluation of the associated Hessian. In the current chymotrypsin system this would entail evaluation of gradients for 11,901 separate geometries.

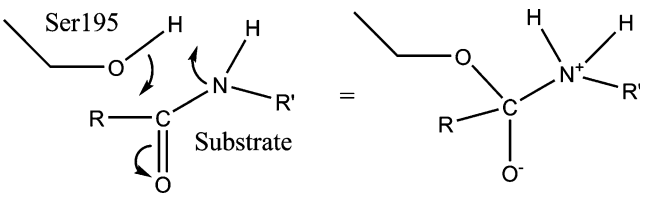

Fig. 34 Formation of a tetrahedral intermediate in chymotrypsin

Fortunately, the normal mode representing motion through the transition state invariably involves only a very few atoms. This, together with the necessary and sufficient requirements that a transition state is a stationary point on the PES in which there exists exactly one normal mode with a negative force constant, allows a rapid and reliable procedure to be developed for refining transition states. This procedure is iterative, and consists of repeated pairs of operations. In the first of these, the gradient norm for the set of atoms that are central to the reaction, the "core atoms," is minimized using Bartels' non-linear least squares (NLLSQ) method [60]. During this operation, the geometry of the rest of the system is frozen. In the second step, the positions of the core atoms are frozen, and the rest of the geometry optimized. By repeated execution of this pair of operations, the geometry of the transition state is rapidly and reliably obtained, resulting in the middle structure shown in Fig. 35.

As with the other operations, validation of the transition state required modification of the standard procedure. Instead of constructing the full Hessian, advantage is once again made of one of the necessary and sufficient conditions that characterize transition states, specifically, the existence of a single negative force constant. Normal modes of vibration can be represented by eigenvectors of the mass-weighted Hessian. A consequence of this is that, for transition states, the associated eigenvalue of the normal mode representing the reaction coordinate has the most negative value possible. Any modification of the eigenvector would, of necessity, cause its eigenvalue to increase. An implication of this property, unique to the eigenvector with the lowest eigenvalue, is that the reaction coordinate normal mode has intensity on only a few atoms—the atoms directly involved in the reaction plus a small number of adjacent atoms. Although it is intuitively obvious that the reaction normal mode would be highly localized, the simplest mathematical description of this involves an apparently counterintuitive eigenvalue equation: given the Hessian matrix, F, and the negative force constant, ε_1, each eigenvector coefficient, $\Psi_{\lambda 1}$ for motion of a given atom must satisfy Eq. 1.

$$\Psi_{\lambda 1} = \varepsilon_1^{-1} \sum_{\sigma} F_{\lambda\sigma} \Psi_{\sigma 1} \tag{1}$$

As the distance between any two atoms increases, the absolute value of $F_{\lambda\sigma}$ decreases, and rapidly converges to zero. Because all $F_{\sigma\sigma}$ far from the transition site are necessarily large and positive, and because the transition mode eigenvalue is large and negative, it follows that the eigenvector can have a significant intensity only on those atoms that are important to the reaction, and that coefficients for all other atoms must be small, decreasing in value rapidly as the interatomic distance increases. Therefore, instead of using the entire system in the construction of the Hessian, only the core atoms are needed. Using this reduced Hessian, the normal mode representing the reaction coordinate is then readily constructed, and its eigenvalue evaluated. By use of this technique, most of the computationally intensive operations are thus avoided.

There is no obvious choice of the number of atoms to be considered as core atoms, so the normal modes of vibration for the transition state were evaluated using three different sets of atoms: one set of four atoms, consisting of the carbon and nitrogen atoms of the peptide bond being hydrolyzed plus the hydroxyl of Ser195; a second set, of nine atoms, consisting of the first set plus adjacent atoms; and a third set of 18 atoms consisting of the second set plus adjacent atoms. For these three sets, the lowest vibrational frequencies were, in order, *i*925, *i*934 and *i*940 cm^{-1}, and the next lowest vibrational frequency in each set was real and over 100 cm^{-1}. These results provided unequivocal justification for the steps used in this process. The narrow spread of imaginary frequencies implied that even the smallest set of core atoms was large enough to verify that the system was at a transition state. The second and higher vibrational modes, although positive, were significantly different in the various calculations. Again, this was expected, in that, although the character of the transition

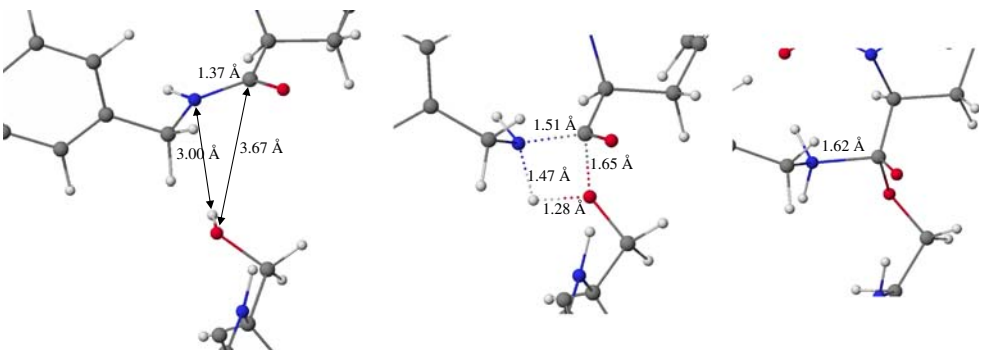

Fig. 35 Reactant, transition state and tetrahedral intermediate in chymotrypsin. *Left* Reactant, *middle* transition state, *right* tetrahedral intermediate

mode should be conserved between sets of core atoms, this requirement did not extend to the other modes.

A useful operation to confirm that the transition state has indeed been reached is to generate the intrinsic reaction coordinate (IRC) [61]: that is, the minimum energy path in mass-weighted coordinate space connecting reactants and products. The starting point for this operation is the putative transition state geometry, at which point any small perturbation in the direction of the reaction normal mode would cause it to move downhill in energy to either the reactants or the products, depending on the phase of the mode. Using the normal coordinate for the reaction from the previous step, the IRC path or trajectory for the migration of the hydrogen atom to form the tetrahedral intermediate was generated in two steps, first from the transition state to the reactants, then from the transition state to the products. Each trajectory was terminated when the forces acting on the atoms had dropped below a pre-set limit. The steps were then joined together to form the entire reaction coordinate (Fig. 36). Computational artifacts generated in the SCF calculation resulted in small spikes in the low-energy regions.

As expected, the product side of the IRC terminated in the region of the tetrahedral intermediate. Exhaustive optimization of the intermediate gave a ΔH_f of −26,165.3 kcal mol^{-1}, only 0.6 kcal mol^{-1}, below the termination of the IRC, at −26,164.7 kcal mol^{-1}. Conversely, and initially unanticipated, the geometry of the starting structure on the reaction side of the IRC was distinctly different to that expected from the X-ray structure. Where the IRC had the ΔH_f of the reactants equal to −26,168.5 kcal mol^{-1}, the ΔH_f of the optimized X-ray structure was considerably more negative, at −26,190.1 kcal mol^{-1}. The geometries were also very different: although the optimized X-ray structure to transition state geometry represented a motion of over 800 Å, in the IRC the transition state to reactant distance was only 55 Å.

A conjecture to explain the large difference between the reactant geometry from the IRC and that from the optimized X-ray structure can be derived based on a consideration of the nature of the two structures. First, there is overwhelming evidence for the existence of myriad minima on the PES of real proteins, and since the X-ray structure presumably corresponded to the lowest energy conformation of the docked substrate in the active site of chymotrypsin, optimization of the X-ray geometry would naturally result in a very low energy structure. This need not be true when the starting geometry is a transition state. In order to reach the transition state, various geometric changes had to be made, and because the energies involved are relatively large, these motions almost certainly involved large conformational changes, amounting to a RMS motion of 0.35 Å per atom. On descending from the transition state in the direction of the reactant, the IRC would be expected to terminate at the first minimum, a minimum that in this case was 22 kcal mol^{-1} above the optimized X-ray minimum. This conjecture implied that further examination of the PES between the IRC reactant minimum and the optimized X-ray minimum should reveal many small intermediate minima separated by transition states corresponding

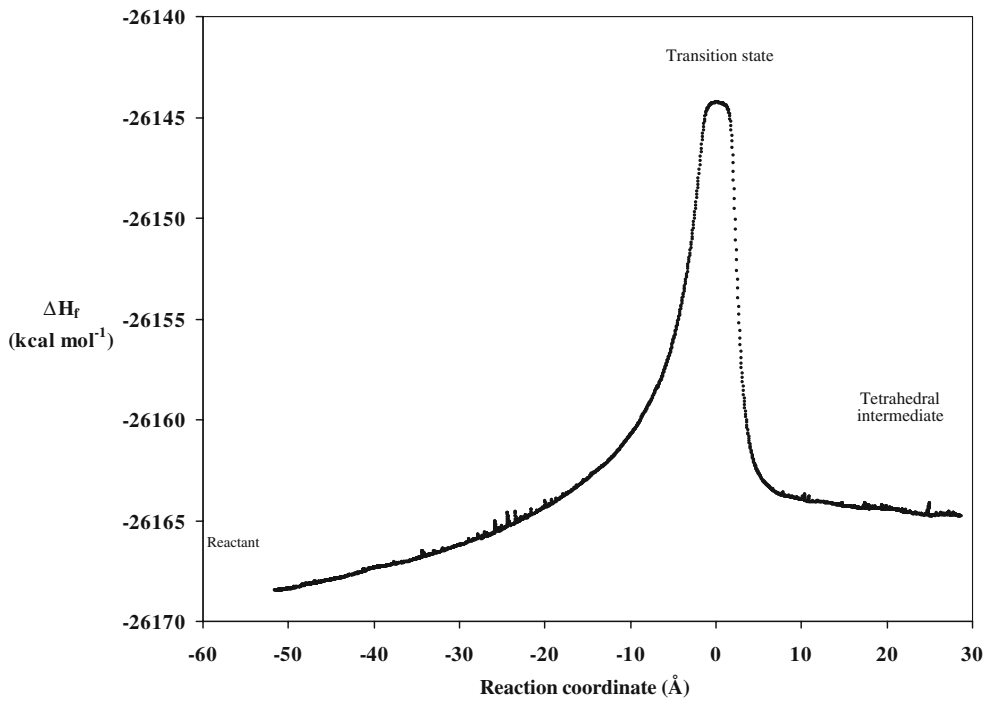

Fig. 36 Intrinsic reaction coordinate for the formation of the tetrahedral intermediate in chymotrypsin. The small spikes in the low energy regions are computational artifacts

to conformational changes, but not any that involve high energy transition states of the type encountered in covalent bond making or bond breaking.

This conjecture was tested by selecting the terminal geometry from the IRC and performing an exhaustive optimization. Although the terminal geometry was optimized within the criteria of the IRC calculation, when the criterion was tightened, the geometry continued to change, accompanied by a steady decrease in heat of formation, until, after 688 further cycles of optimization, it converged to a geometry similar to that resulting from optimization of the X-ray structure. As a result of the monotonic decrease in energy on going from the terminus of the IRC calculation to the fully optimized geometry, the conclusion can be made that the conjecture was incorrect—the hypothetical small local minima that were postulated to impede motion from the transition state to the true minimum did not in fact exist, and that the failure of the IRC to converge on the true minimum can be attributed simply to the different termination criteria.

Mechanical properties: Young's modulus

Some proteins have distinct biomechanical properties, often forming long elastic fibers that have great tensile strength. The behavior of such systems when stretched is therefore of interest. As no work has thus far been reported on the applicability of PM6 to modeling mechanical properties, a preliminary examination of a simple and well-characterized high polymer is warranted.

Description of method and application to polyethylene

An estimate of the accuracy of PM6 for prediction of elastic moduli was obtained by modeling the stretching of a polyethylene molecule, using the cluster method for elastic moduli [62]. Given that the density, ρ, of polyethylene is 0.96 g cm^{-3}, the repeat distance, l, for the C_2H_4 unit calculated using PM6 is 2.537 Å, and the molecular weight of the C_2H_4 unit is 28, then the cross-sectional area, A, of a polyethylene molecule can be determined using Eq. 2, in which N is Avogadro's number, 6.022×10^{23}.

$$A = \frac{MWt}{N \times l \times 10^{-8} \times \rho} = 19.09 \text{ Å}^2 \quad (2)$$

Using the cluster technique [7, 63], elongation of a short section of polyethylene chain consisting of $C_{16}H_{32}$ yielded a stress-strain curve in which the increase in heat of formation, $\Delta\Delta H_f$, in kcal mol^{-1}, resulting from strain, Δx, could be represented by a quadratic equation as $\Delta\Delta H_f = 17.580478\cdot\Delta x^2 + 3.204278\cdot\Delta x - 0.27136$, with an R^2 of 0.999937. The effect of anharmonicity was estimated by fitting the stress-strain curve using a cubic expression; this yielded a $\Delta\Delta H_f$ of $-1.1595312\cdot\Delta x^3 + 20.9721068\cdot\Delta x^2 + 0.5921442\cdot\Delta x + 0.1260261$, and an R^2 of 0.9999986. Both equations give similar stretching curves, the difference at a 10% elongation amounting to 0.4 kcal mol^{-1}, at which point the stress, $\Delta\Delta H_f$, was over 70 kcal mol^{-1}.

The force constant, K, can be estimated from the quadratic coefficient, C, using Eq. 3.

$$K = \frac{2 \times C}{N} \quad (3)$$

After converting to MKS units (factor=$4.184\cdot10^{23}$), the quadratic equation yielded K=24.42 N m^{-1}; the cubic equation gave a similar result: K=29.13 N m^{-1}.

Young's modulus, E, can be estimated from the quadratic term, K, the cross-sectional area of the polymer, A, and the translational repeat distance, l, of $C_{16}H_{32}$ of 20.296 Å, using Eq. 4.

$$E = \frac{Kl}{A} \quad (4)$$

This gives a modulus of 259.8×10^9 N m^{-2}, or 259.8 GPa, for the quadratic and 309.9 GPa for the cubic form, in good agreement with an experimental value [64] of 288 GPa, and ab initio results of 334 GPa [65] and 360.2 GPa [66].

Silk

The most important characteristic of silk protein is its ability to form strong elastic fibers, a property that can be directly attributed to its antiparallel β sheet structure. Based on X-ray analyses, values of Young's modulus for the crystalline domains in silk range from 20 to 28 GPa [67], depending on the degree of crystallinity, and calculations [68] give 13 or 16 GPa, depending on the method of analysis.

Young's modulus for a model silk system, crystalline poly(L-Ala-Gly), was calculated using an orthorhombic cluster of 32 units of (L-Ala-Gly). This cluster had dimensions a=13.7 Å, b=18.19 Å, and c=18.09 Å, with the "a" translation vector corresponded to the fiber axis, and a cross-sectional area of 329.06 $Å^2$. Starting with the equilibrium system, the value of the "a" vector was steadily increased, representing increasing strain in the crystal, which, in turn caused the ΔH_f, i.e., the stress, to rise, as shown in the region of Fig. 37 where the strain is 0 to 1 Å.

Unlike polyethylene, the stress-strain curve for solid poly(L-Ala-Gly) was only very approximately parabolic. At low strains, the force constant was low: K=~70 N m^{-1}, and E=~29 Gpa. In this domain, the increased values of the translation vector is achieved by changes in torsion angles. When the strain increased, the modulus also increased, with

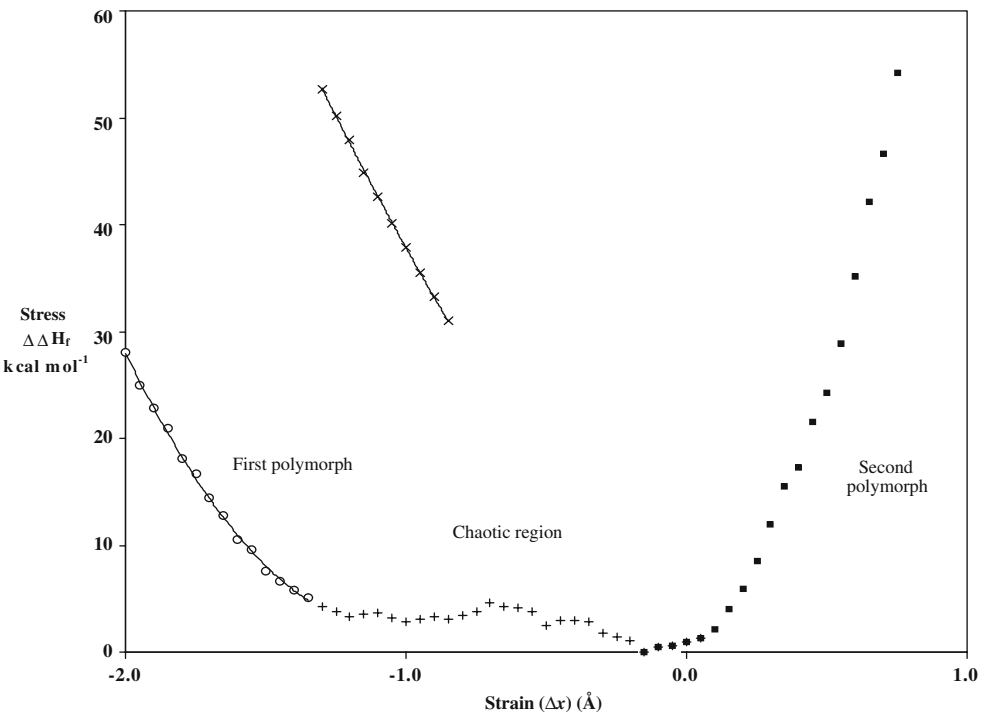

Fig. 37 Stress-strain behavior of crystalline poly(L-Gly-Ala)

E eventually rising to ~60 GPa. This increase can be attributed to the transition from changes in torsion angles to changes in bond-angles and bond-lengths, so that the nature of the strains become similar to those for polyethylene.

An interesting result was obtained when the crystal was compressed. If silk behaved like a classical solid, then, for small deformations, the effect of compression would be essentially the same as that of tension, i.e., the potential well should be approximately parabolic. Instead, with increased compression, the ΔH_f, varied erratically until, at a compression of about 1.3 Å, a new parabolic surface appeared, one unrelated to that obtained by stretching. The obvious interpretation is that as the uniaxial compression of the crystal is steadily increased, the internal structure changes from one polymorph to the other, with these polymorphs being those described earlier. After all the inter-sheet hydrogen bonds formed, any further compression resulted in stress which manifested itself as an increased ΔH_f.

A metastable structure was also found in the compression region. The structure of this polymorph was similar to that found in the tension region, where inter-sheet hydrogen bonds did not form, and its stress-strain curve was apparently a continuation of the tension curve.

Collagen

Collagen, found in ligaments and tendons, is another structural protein that has great tensile strength. In contrast to silk, which forms β sheets, collagen consists of fibrils, each of which is built from three molecules of tropocollagen: long α helix chains that contain a repeating motif of three residues, which, in the case of the synthetic collagen-like polypeptide 1A3J [69], are Pro-Pro-Gly. The three tropocollagen helices twist around each other to form a coiled-coil superpolymer. Because of steric requirements, the Gly residues in each molecule are positioned in the helix nearest to the center of the triple strand, with the two Pro residues pointing out away from the axis, as shown in Fig. 38. In the X-ray analysis of 1A3J [69], the authors assumed that the structure could be represented by a hydrated infinite high polymer in which the repeat unit was $(Pro\text{-}Pro\text{-}Gly)_7 \cdot (H_2O)_{40}$. This assumption allowed the

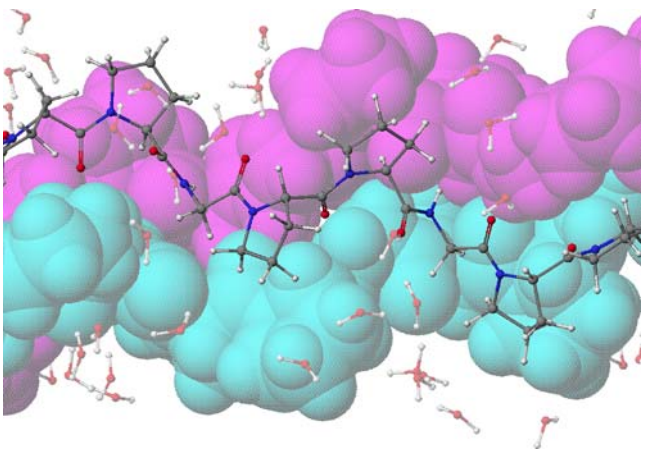

Fig. 38 Section of collagen, 1A3J, showing triple strand

repeat or translation distance to be determined; the value reported was 20.42 Å.

The structure of 1A3J was optimized using PM6 using PBC to simulate the infinite polymer. Because of the large size of the unit cell, only one unit cell was needed, so the system used to represent collagen was correspondingly small; only 21 residues, that is, only one-third of the repeat distance of a single chain was used. The result of translating one of the tropocollagen sections was to place it at the end of the adjacent tropocollagen section, i.e., the end of each of the three chains joined on to the other end of one of the other two chains, thus preserving the three-fold symmetry. Optimization proceeded rapidly and smoothly, and yielded a translation distance of 20.39 Å, in excellent agreement with that reported.

An estimate of the Young's modulus for collagen was obtained by stretching the collagen polymer, which involved increasing the translation distance in several steps and allowing the geometry to relax between each step. If the one-third unit cell used in the geometry optimization was used, then increasing the translational distance would result in one chain being pulled apart from the two other chains. To avoid this catastrophe the stretching operation was performed using the entire unit cell, i.e., $(Pro\text{-}Pro\text{-}Gly)_{21}\cdot(H_2O)_{120}$. In other words, one complete repeat unit for each of the three chains was used in the modulus calculation.

The stress-strain curve for collagen is given in Fig. 39. In contrast to polyethylene, where all the deformation arose from the C–C–C angle increasing and from the C–C stretch, the variation with strain in heat of formation for collagen exhibited excursions of up to several kcal mol^{-1}, reflecting the much greater complexity of the polymer. In particular, the water molecules changed orientation as the strain increased, giving rise to hydrogen bonds being made and broken. To avoid the effect of these excursions unduly affecting the results, only the quadratic fit was used. The reported size of the orthorhombic unit cell for collagen (27.01×26.42×20.42 Å, with 20.42 being along the axis) allowed the cross-sectional area of the fiber to be determined: 713.6 $Å^2$. This represents four parallel fibers; therefore, the area of one fiber to be used in estimating Young's modulus was one-quarter of that: 178.4 $Å^2$. The translation distance corresponding to zero strain, obtained from the stress-strain curve, was 62.5 Å. This was slightly greater than that expected from the first optimized PM6 structure, 61.2 Å, and likely reflected the changes arising from the different orientations of the surrounding water molecules.

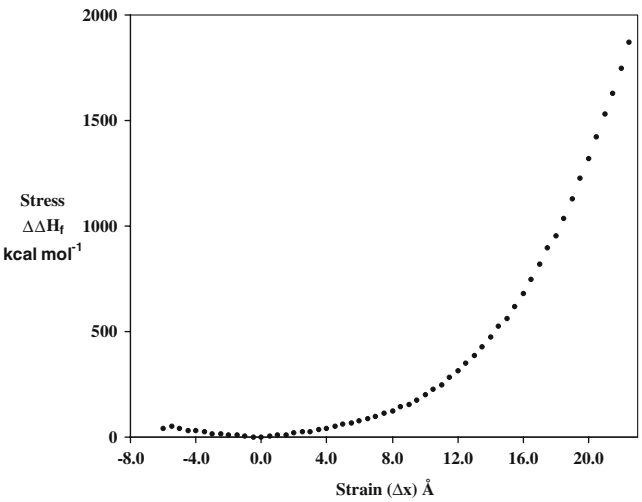

Fig. 39 Stress-strain curve for collagen, 1A3J

For strains up to 10%, the heat of formation increased in proportion to the square of the strain, as expected, giving $\Delta\Delta H_f = 1.27\cdot\Delta x^2 + 3.69\cdot\Delta x + 0.48$, with an R^2 of 0.9987. This implied a force constant of 1.76 N m^{-1}, and a Young's modulus of 6.18 Gpa, in good agreement with other theoretical studies, which give values of 4.8 GPa[70] and ≈7 GPa[71].

Above a 10% strain, collagen became significantly more stiff, with the result that, for a 36% strain, the best quadratic fit of the stress-strain curve was $\Delta\Delta H_f = 5.28\cdot\Delta x^2 - 44.64\cdot\Delta x + 100.22$ with an R^2 of 0.9931.

Constrained optimization

Optimized PM6 structures for proteins are, in general, in good agreement with the starting X-ray structures, albeit thus far PM6 has not been shown to be able to predict such structures de novo. Indeed, there is strong evidence that the myriad local minima on the PES would militate against PM6 being a suitable computational method for making such predictions. Nevertheless, since PM6 predicts very short-scale geometric quantities, such as the primary structure of proteins, e.g., bond lengths and angles, with good accuracy, and since X-ray analyses are ideal for generating secondary, tertiary, and quaternary structures, a combination of the two methods would most likely be of much higher accuracy than either method in isolation. This idea, to refine protein crystal structures using semiempirical quantum mechanically derived energy constraints, was first proposed by Yu, Yennawar, and Merz, in 2005 [72], and subsequently critically assessed in 2006 [73].

That primary structures predicted by X-ray analyses are of limited accuracy can readily be demonstrated by observing the precipitous decrease in heat of formation of the first few cycles of a PM6 geometry optimization. For the larger proteins, changes in calculated heats of formation are often in the order of several thousand kcal mol^{-1}. During these early cycles of geometry optimization, errors

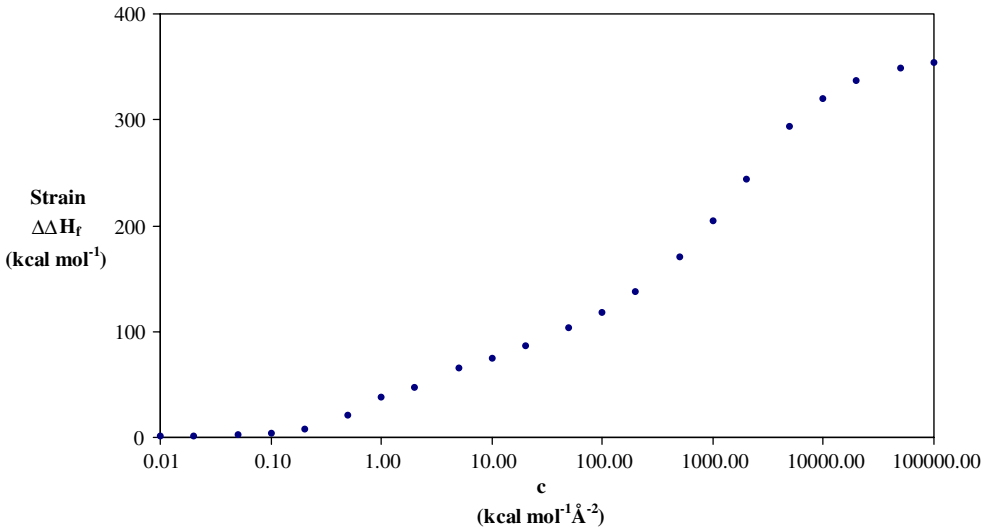

Fig. 40 Dependency of strain induced in crambin as a result of constrained optimization

in X-ray bond lengths and angles—geometric quantities with large force constants—are corrected, and, since large scale motions are not occurring, the RMS error of the optimizing structure remains small.

The rapid drop in energy is then followed by a slow decrease in energy, frequently lasting many hundreds of cycles. During this phase, large changes occur in the secondary, tertiary, and quaternary structures of the protein. As these geometric changes are associated with small or very small force constants, the resulting structure becomes less and less accurate. If a small force constant constraint, of the type proposed by Yu et.al. [72], were to be added to the overall system, then errors in the secondary and higher order structures could be minimized, and the optimized geometry resulting from the combined method, PM6 plus X-ray structure, would be of unprecedented accuracy.

Crambin was selected to test this idea, and a set of geometry optimizations performed, in which the constraint used was represented by the addition to the calculated heat of formation of a potential equal to a constant, c, times the square of the sum of the displacements of the non-hydrogen atoms from their initial positions, as shown in Eq. 5.

$$\Delta H_f' = \Delta H_f + c \times \sum_i \left(x_i - x_i^0\right)^2 \quad (5)$$

For any given value of c, the system is subjected to a stress arising from the constraint. When c is small, the strain imposed on the system is small (Fig. 40), and the RMS difference (Fig. 41), between the optimized structure and the X-ray structure, is large: i.e., the geometry optimizes to the ideal PM6 structure. Conversely, when c is very large, the geometry is highly strained, and is, for all practical purposes, constrained to the X-ray structure, with only the positions of the hydrogen atoms being optimized.

The region of interest is where the RMS difference is small, and, at the same time, the strain is also relatively small; this occurs when c is in the domain of 1 to 10. When c=10, the strain, at 75 kcal mol^{-1} is 80% less than that in the X-ray structure (354 kcal mol^{-1}), and the RMS error is 0.07 Å. When c is reduced to 1, the strain decreases to

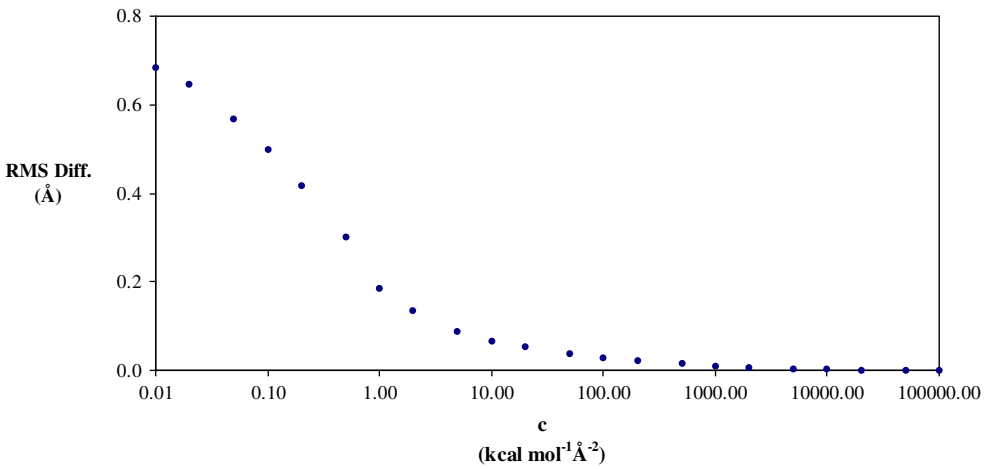

Fig. 41 Root mean square distortion from X-ray geometry of crambin resulting from constrained optimization

38 kcal mol^{-1}, i.e., 90% less than that in the X-ray structure, but this is offset by the RMS error increasing to 0.19 Å.

From these results, it is apparent that an improved, i.e., nearer to the true structure, geometry can be obtained by using a constrained optimization. It is not obvious whether it is preferable to use a large value of *c* and only allow about 80% of the strain from the X-ray structure to be relieved, or to use a smaller value, thus relieving more strain, but, at the same time, increasing the risk that significant errors due to faults in PM6 might be introduced; such a decision would depend on the purpose for which the resulting geometry would be used.

In all the earlier geometry optimizations reported here, the cutoff for NDDO to point charge plus polarization terms was assigned a large value in order to more precisely reproduce the PM6 method. When a constrained optimization is done, the resulting optimized geometry cannot be described as a PM6 structure, and therefore there is no a priori reason to try to reproduce the PM6 method. On the other hand, a large increase in computational efficiency can be obtained by reducing the cutoff. A cursory test of reducing the cutoff to 6 Å resulted in a negligible distortion of the geometry, but was accompanied by a large reduction in computational effort. The implication of this is that by using a small value for CUTOFF, the method described here provides a simple, very rapid, and general procedure for improving X-ray structures of proteins.

pKa

One of the commonest reactions in proteins occurs when the pH of the surrounding medium is changed, resulting in the gain or loss of a proton from a hydroxyl group to form either the neutral moiety or the anion. Sites where hydroxyl groups occur are Asp, Glu, Cys, Ser, Tyr, Thr and at the carboxylic acid terminus. In recognition of the importance of this process, an attempt was made to predict the pKa for the hydroxyl groups for a set of organic compounds, using the optimized hydroxyl bond lengths, R_{O-H}, and calculated charges, Q_H, on the ionizable hydrogen atom, and using an expression of the type shown in Eq. 6.

$$pKa = aR_{O-H} + bQ_H + c \tag{6}$$

The resulting values were of low accuracy, the average unsigned error (AUE) being about 0.6 pKa units. While disappointing, this was not surprising insofar as, although PM6 has been shown to reproduce bond lengths with good accuracy, partial charges are not an observable and therefore no attempt had been made during the development of PM6 to reproduce the charges on specific atoms. In an attempt to correct this, the NDDO parameters that determine partial charge were reoptimized to reproduce the pKa of a set of simple organic compounds, again using Eq. 6. During this process the geometries were frozen at the optimized PM6 structures. This resulted in a significant increase in accuracy. A further improvement was obtained when the reference compounds were solvated using Klamt's COSMO [74] method, which resulted in the AUE dropping to 0.31 pKa units for a set of 109 compounds. The final optimized values of the parameters for use in Eq. 6 were a=−288.0531, b=28.6889, and c=89.1172. The optimized parameters for generating the charges are given in Table 7.

The combined method, Eq. 6 with O–H distances calculated using PM6 and charges derived using the parameters in Table 7, was used to predict the pKa for all ionizable hydroxyl sites in various proteins. The results are presented in Table 8.

Some pKa values predicted using PM6 were unexpected in that they were large and negative and implied an impossibly strong acid. Examination of the environment of the proton involved revealed several interesting structures. In the simplest of these, a hydroxyl proton had formed a strong hydrogen bond to a water molecule, the

Table 7 Parameters used in generating charges for pKa calculation

	Hydrogen	Carbon	Nitrogen	Oxygen
U_{ss} [eV]	−9.355540	−49.939003	−55.430747	−89.947265
U_{pp} [eV]		−43.823335	−49.824730	−70.724996
β_s [eV]	−2.678981	−12.644614	−21.761245	−66.749718
β_p [eV]		−9.461422	−16.933587	−21.779505
ζ_s [$bohr^{-1}$]	1.245857	1.641208	1.486068	4.399104
ζ_p [$bohr^{-1}$]		1.581098	2.042007	2.161493
g_{ss} [eV]	14.596444	16.761654	7.114914	16.387563
g_{sp} [eV]		12.392540	7.224266	16.029363
g_{pp} [eV]		11.267926	14.810368	16.609899
g_{p2} [eV]		11.015834	11.450082	11.033971
h_{sp} [eV]		0.420002	4.587341	4.792060

Table 8 PM6 predicted values of pKa for ionizable hydrogen atoms in proteins

Bacteriorhodopsin		Barnase		Chymotrypsin	
ASP 85	−4.03	ASP 86	−11.36	TYR 228	4.02
GLU 166	−3.35	ASP 12	−3.43	TYR 94	4.69
GLU 9	−2.98	ASP 23	−1.07	SER 217	5.19
TYR 43	−2.84	ASP 101	−0.58	THR 151	5.60
TYR 83	−0.95	ARG 110	0.54	THR 232	6.07
ASP 115	1.88	ASP 8	0.70	SER 214	6.53
ASP 104	1.90	ASP 44	1.18	THR 117	6.80
ASP 102	2.13	GLU 60	3.11	SER 32	6.95
GLU 204	2.44	SER 50	3.47	THR 104	7.86
ASP 212	2.81	GLU 73	3.57	SER 113	8.23
GLU 74	3.61	GLU 29	3.69	THR 166	9.46
ASP 36	4.10	ASP 75	3.73	SER 96	9.83
ASP 96	4.19	TYR 97	4.48	THR 219	9.95
GLU 194	4.37	TYR 17	4.78	THR 222	10.08
TYR 57	5.32	SER 92	4.89	TYR 146	10.50
TYR 26	7.49	TYR 103	7.11	SER 186	10.55
THR 205	7.52	ASP 54	7.47	SER 119	10.57
ASP 38	7.75	TYR 13	7.85	SER 189	10.88
SER 226	7.89	THR 6	8.85	THR 144	10.97
TYR 131	8.38	TYR 78	9.46	TYR 171	11.06
TYR 133	8.86	THR 105	10.12	THR 54	11.12
SER 162	9.12	THR 26	10.52	THR 62	11.22
THR 107	9.78	SER 85	10.61	SER 75	11.24
TYR 147	9.87	SER 28	10.72	THR 139	11.36
TYR 64	9.92	SER 57	11.83	SER 127	11.38
TYR 150	9.97	THR 107	12.34	THR 110	11.42
SER 35	10.11	THR 100	12.49	THR 241	11.50
PHE 156	10.73	THR 79	12.71	SER 77	11.77
THR 170	11.02	TYR 24	14.09	THR 224	11.93
THR 90	11.62	SER 91	14.26	SER 221	12.15
THR 67	11.86	SER 80	14.27	SER 195	12.34
THR 128	12.32	THR 99	14.28	SER 190	12.46
SER 132	12.69	SER 67	14.56	SER 76	12.46
TYR 79	12.96	THR 70	14.77	SER 109	12.61
THR 121	13.56	SER 38	15.57	THR 208	12.82
SER 169	14.01	THR 16	15.60	SER 223	12.98
THR 46	14.38	TYR 90	15.64	THR 61	13.22
SER 59	14.71			SER 45	13.43
THR 5	14.77			THR 37	13.77
THR 89	15.08			SER 159	13.84
TYR 185	15.51			SER 115	13.98
THR 55	15.56			SER 63	14.10
THR 178	15.65			SER 92	14.42
SER 214	16.12			SER 218	14.88
SER 141	17.20			THR 174	15.19
SER 183	17.51			THR 98	15.77
THR 142	17.94			SER 164	15.85
THR 17	18.47			SER 125	16.07
THR 47	18.86			THR 135	16.85
SER 193	18.86			THR 134	17.18
THR 24	19.75			THR 138	18.20
				SER 26	19.23

resulting structure being intermediate between the neutral system and a hydronium ion in close proximity to an oxygen anionic site. A somewhat more complicated system involved a hydroxyl proton strongly hydrogen bonding to the ionizable nitrogen of a neutral Arg residue: this formed a nascent salt bridge. Still another structure involved two ions, a cation, invariably Arg(+), and two to four water molecules having a net charge of −1, e.g., $[H_7O_4]^-$, positioned near to the hydroxyl. The close proximity of the large poly-water anion to the hydroxyl unit resulted in an unusually strong hydrogen bond to the hydroxyl proton.

Discussion

Related semiempirical work

The primary structure of hen egg white lysozyme, represented by 193L [75], has been optimized [76] with AM1 [3] using a divide-and-conquer method in which the fundamental unit was the residue. In their work, the authors did not optimize the secondary or tertiary structure "because traditional semiempirical Hamiltonians (AM1 and PM3) have serious shortcomings in the description of peptide backbones." This assertion was investigated here by performing a global optimization of 193L using AM1 and PM3. Somewhat surprisingly, these optimizations yielded structures that were compatible with PM6, the RMS error for AM1 being 0.97, for PM3 0.68, and for PM6 0.88 Å.

193L is interesting in that its size, 129 residues, corresponds to the maximum in the size distribution of proteins in the PDB, that is, 193L represents a typical PDB entry. As such, and because it had already been modeled using AM1, the decision was made to optimize the structure of 193L using PM6. The starting structure was the PDB entry. This was complete in that all heavy atoms in all residues were located and, in addition, the PDB file also contained 142 water molecules, and sodium and chloride ions. All these moieties were used in the calculation. This meant that the only preconditioning necessary was the addition of hydrogen atoms to satisfy valency requirements; as usual, all residues were represented by their neutral forms. Optimization proceeded without complication. After geometry optimization was complete, the active site in egg white lysozyme was examined. This active site catalyzes the hydrolysis of a polysaccharide, and, in 193L, is composed of the residues Glu35, Asp52, Trp62 and Trp63. No significant distortions were observed. That is, the active site was accurately reproduced, the RMS error for the four residues involved being 0.53 Å. An exact comparison of the computational effort required for the PM6 calculation with the AM1 work reported by Wada and Sakurai was not possible, but the general impression was that the PM6 calculation using MOZYME ran significantly faster than the combined MOZYME and divide-and-conquer method used in the AM1 calculation.

Geometries

Given that the starting point for all optimizations were geometries derived from files in the PDB, the influence of the correct answer, i.e., the initial X-ray or NMR structure, must be considered. In all systems examined, there was no ambiguity: the starting structure had a large and obvious influence on the final optimized structure. This is a natural consequence of the presumed existence of a very large number of local minima in even the smallest system studied, and it is undoubtedly the presence of these minima that determines the final structure. No attempt was made to locate the global minimum—this operation is well known to be extraordinarily difficult, and, even if it were possible to locate the global minimum on the PM6 PES, there is no reason to believe that this would be the true minimum. Therefore, all structures reported here should be regarded as being derived from the reference data, and should not be considered as de novo structures. What can be addressed, however, is the level of accuracy of prediction of some of the structures that exist in proteins. These are treated in the following sections, in order of increasing complexity.

Primary and secondary structure

Because of its high accuracy, PM6 would be useful in detecting large errors in original X-ray structures. After preconditioning a structure from the PDB, the gradients or forces acting on the heavy atoms can be calculated. Consistent with the premise that the PDB structure is of good accuracy, most of these should be small. Any large gradients would then be indicative of significant differences between the PDB and optimized PM6 structure. In a survey of several proteins, most large differences were found to occur in arginine residues and in residues containing carboxylate, i.e., Asp and Glu. Carboxylate X-ray structures have C–O distances in a single range, from 1.24 to 1.26 Å, but PM6 predicts three distinct sets of the C–O distances: 1.20–1.23 Å, 1.25–1.27 Å, and 1.34–1.39 Å, corresponding to the C = O and C–O–H of the neutral carboxylic acid, and the $C–O^{1/2-}$ of an ionized carboxylate, respectively. These differences arise from the fact that an ionizable hydrogen atom in a carboxylate group can be in one of two different locations, and presumably in proteins both locations are fractionally occupied, but in the quantum chemical calculation any ambiguity of this type must be resolved. When appropriate averaging was done, the differences in observed and predicted C–O bond-lengths decreased considerably. A similar condition exists for arginine, and, when averaging

was done, the large difference in C–N bond lengths again vanished. After discounting atoms with large forces attributable to the resolution of fractional populations or ambiguities, the remaining forces can be interpreted as indicators of possible errors in the PDB structures. Several types of such putative errors were found. Representative examples of these are:

Some covalent bonds are of unexpected lengths. Thus, in the X-ray structure of 1CBN, the C_γ–N_δ distance in Asn_{46} is reported to be 1.25 Å, much less than the expected 1.33–1.38 Å for a bond of this type. A PM6 calculation showed that the forces acting on these two atoms were very large, over 200 kcal mol^{-1} $Å^{-1}$, given that the median force on an atom in 1CBN was 4 kcal mol^{-1} $Å^{-1}$. Two other Asn residues are present in 1CBN, and for both of these the C_γ–N_δ distances, 1.312 and 1.347 Å, were nearer to the values expected. Many errors of this type were found in X-ray structures, and their occurrence was apparently unpredictable.

Some geometric quantities had systematic errors. In the PDB structure for ricin, 2AAI, for example, the reported angles for C_γ–C_δ–N_ε in tryptophan residues averaged about 104°. This is significantly less than the typical 109–111° angles found in other proteins, less than that in crystalline tryptophan from CSD entry LIHLIX (110.6°), less than that found in the structure of ricin predicted by PM6 (109.6°), and less than that predicted for tryptophan by B3LYP (110.2°), and by PM6 (109.3°).

Some non-bonded distances were unexpectedly short. In the X-ray structure of hemoglobin, 1GZX, an oxygen atom on Phe188 was positioned only 2.55 Å from a nitrogen atom of Asn200. During preconditioning, hydrogen atoms were added and after their positions were optimized one hydrogen atom was positioned between the oxygen and the nitrogen atoms, implying that the hydrogen bond distance to the carbonyl oxygen was only ~1.5 Å. This would represent a highly unusual, extremely short, and therefore strong, hydrogen bond. On optimizing the structure, the N–O distance increased to the expected 2.9 Å.

Some groups were incorrectly assigned. In crambin, the side-chain, $–CH_2–CONH_2$, of residue Asn46 has the locations of the oxygen and amine groups exchanged in PDB entries 1CBN and 1EJG. Obviously only one of these structures can be correct. Errors of this type can occur in X-ray structures when the number of electrons in two groups are similar. Only minor motion of Asn46 occurred when the structure of 1EJG was optimized, but a large motion occurred when 1CBN was optimized, with the side chain rotating by almost half a circle, essentially converting to the orientation in 1EJG, which strongly suggests that the orientation of the side chain in 1EJG was correct.

Some differences between X-ray and PM6 structures can be attributed to the lack of environmental effects in the computational model. These are most obvious in surface residues, where solvent and other external effects would be largest. An example of this is provided by the surfactant protein hydrophobin HFBII found in *Trichoderma reesei*, for which a very high resolution structure has been reported, 2B97 [77]. Hydrophobin is a globular protein strongly stabilized by four sets of disulfide bridges. Its surfactant behavior arises from the presence of an unusually large fraction of hydrophobic residues on its surface, and as a result the protein is, as its name suggests, highly hydrophobic.

Geometry optimization resulted in only minor motion of the heavy atoms of the hydrophobic residues, the RMS difference being 0.82 Å, but in a much larger motion of the ionized residues, the RMS error for these being 1.43 Å. All the ionized sites lie on the surface of the protein, and, in both the crystal and in vivo, these sites would be either involved in salt bridges or be solvated. As the PM6 calculation was performed using only the isolated molecule, and therefore solvation effects were not included, the large distortions of the ionized sites can be rationalized.

Tertiary structure

Most of the proteins considered in this work are globular, with the secondary structures cross-linking using disulfide bonds, salt bridges, and bridging and normal hydrogen bonds. As bonds of this type are strong relative to the other interactions that determine the shape of the backbone, tertiary structures are reproduced with an accuracy similar to that of the secondary structure. Only when the structures become very large, as in hemoglobin, do the RMS errors approach 2 Å. In some oligopeptides such as the nonapeptides in 1V46[78], stabilizing inter- and intra-chain bonds are absent; in those cases the PES is very flat, and the structures predicted by PM6 are severely in error.

The agreement between the predicted and reported structures for the zinc finger proteins 1EF4 and 3ZNF were unexpectedly poor, the RMS difference being very large. These systems, and the nonapeptides in 1V46, were among the few structures examined that were derived from NMR analyses rather than from X-ray. While there is no obvious reason for these large differences, the speculation can be made that since the NMR structures were derived from solvent studies and the PM6 calculations modeled the isolated, gas-phase system, the distortions are due to the neglect of solvent effects.

Quaternary structure

Several proteins with quaternary structure were examined. Among these were ricin, with two sub-units connected by a disulfide bond, and per-oxy-hemoglobin, with four sub-units joined by salt bridges and hydrogen bonds. As with the tertiary structures, the quaternary structure is deter-

mined mainly by the same types of inter-chain interactions as those found in tertiary structures. No problems specific to quaternary structures were identified in the PM6 optimizations or optimized geometries.

Reactions

As a result of various modifications, transition states for closed-shell biochemical reactions can now be modeled easily and rapidly. Thus for the hypothetical reaction described above, involving formation of a tetrahedral intermediate in chymotrypsin, the transition state was obtained using a straightforward procedure. The first step in this procedure consists of evaluating stationary points on the PM6 PES corresponding to the reactant and product. There are several ways of preparing these starting points. Using molecular mechanics methods, a substrate can be docked into the active site of an enzyme, and, after preprocessing if necessary, the resulting geometry optimized using PM6. Alternatively, the starting point could be an X-ray structure of the docked substrate. Again, after preprocessing as necessary, geometry optimization would yield the stationary point. Generating the stationary point corresponding to the product requires a knowledge of the reaction, or purported reaction, and a likely first step would consist of modifying the stationary point corresponding to the reactant, followed by energy minimization.

Once both stationary points are available, the systems would then be moved in the direction of the transition state by re-optimizing each geometry after adding to the Hamiltonian a perturbative potential corresponding to the distance to the other geometry. This process is iterative: first, the stationary point with the lower energy is reoptimized in the field of the other geometry, then the geometries are swapped around and the second geometry is optimized in the field of the first, perturbed, geometry. At each stage, the geometry with the lower energy is moved in the direction of the geometry with the higher energy. The process is terminated when the distance between the two geometries is small, e.g., less than 2 Å. In the case of chymotrypsin, this required three complete iterations.

An approximation to the transition state is then readily obtained by averaging the structures of the two strained geometries: this is similar to that used in the synchronous transit method. Given the approximate transition state, refining it to obtain the stationary point, i.e., the minimization of the gradient norm, using traditional methods such as EF [18], would require evaluation of the Hessian. For large systems such as enzymes, construction of the entire Hessian would be extremely computationally intensive, but, by using the unique property of the reaction eigenvector, that its associated eigenvalue is irreducibly negative, the process can be simplified. Because it is irreducibly negative, the reaction eigenvector has significant intensity on only a few atoms, and if only these atoms are used in the gradient minimization, location of the stationary point can be performed rapidly and efficiently. Of course, such an operation necessarily introduces perturbative forces in the rest of the system, so, as with the previous process, a tandem or two-step iterative procedure is necessary: in the first step, the gradient norm of the atoms likely to contribute significantly to the reaction eigenvector is minimized, and in the second step, the heat of formation of the rest of the system is minimized. The stationary point for the whole system is then rapidly achieved; three iterations were sufficient to refine the transition state for the chymotrypsin system. A further increase in computational efficiency was obtained by eliminating the need to construct the small Hessian used in the gradient minimization step. This was achieved by using Bartels' non-linear least squares method for obtaining the Chebyshev solution, which, when used in gradient minimization for refining transition state stationary points, turned out to be much more efficient than EF.

Validation of the transition state was performed using two methods. First, the vibrational frequency of the normal mode corresponding to reaction was determined. As with earlier steps, only those atoms that were likely to contribute to the transition eigenvector were included in the Hessian. That a transition state stationary point existed was confirmed by the presence of one, and only one, large imaginary frequency. A second, definitive, test was to map out the IRC. Starting with the stationary point, the geometry was perturbed using the reaction eigenvector from the normal mode calculation. The IRC was then followed until a new stationary point was reached. On reversing the phase of the normal mode, and repeating the IRC calculation, a different stationary point was reached. Examination of these two new points showed that they corresponded to reactant and product.

Implications of size

Both the computational effort required for solving the SCF equations and the number of steps required for geometric operations increases with the increasing size of the protein. Because of this, computational experiments involving systems of 9,000 or more atoms, while possible, involve a significant computational effort, often in the order of CPU weeks. As a result, the technique described here should not be regarded as being practical for routine work on systems of more than 9,000 atoms. On the other hand, operations involving relatively small systems—up to 5,000 atoms—can be performed facilely, each taking only about 1 CPU day of computational effort. Thus, optimizing the geometry of a substrate docked in an active site of an enzyme, optimizing the product structure, locating the transition

state, refining it, and characterizing it would each require about 1 CPU day for systems of a few thousand atoms. Generating the plot of the IRC would involve more effort, typically in the order of 5–10 CPU days.

As most of the proteins in the PDB are smaller than 5,000 atoms, this implies that the bulk of the systems in the PDB are amenable to modeling with the techniques described here using readily available desktop computers.

Conclusion

The PM6 method has been used successfully for modeling a large number of properties of proteins, including metalloproteins, ranging from generating optimized geometries of enzymes, including the primary, secondary, tertiary, quaternary, and active site structures, to comparison of various candidate structures, to predicting Young's modulus for the stretching of silk and collagen, and, by implication, any regular polymeric system. When the course of a PM6 optimization is biased using experimentally determined geometries of the type found in the PDB, the resulting geometry is likely to be more accurate than either the experimentally derived geometry or that predicted by PM6 alone.

Some active sites were successfully modeled, and, in the case of chymotrypsin, one of the simplest reaction steps in the charge relay mechanism was mapped and verified by determining that the force constant for reaction was negative, and by mapping the intrinsic reaction coordinate. In principle, no problems are anticipated in modeling other reaction mechanisms.

All salt bridges reported to exist in the various enzymes examined were reproduced, but as PM6 is known to favor the formation of salt bridges over the neutral equivalent, the ability of PM6 to reproduce known salt bridges should be tempered with concern that some neutral systems might incorrectly be predicted by PM6 to exist as salt bridges. All calculations reported here involved neutral or singly ionized proteins. A significant lowering of energy would likely occur if solvated higher ionized species were considered. This would be an obvious field of application of the method described here.

Acknowledgments Support for this work was provided by the National Institutes of Health grant No.1R43GM083178-01. The author also gratefully recognizes the generous contribution of Fujitsu for giving permission to use the MOZYME method and for providing the relevant source code.

References

1. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:4907–4917
2. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:4899–4907
3. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) J Am Chem Soc 107:3902–3909
4. Stewart JJP (1989) J Comp Chem 10:209–220
5. Stewart JJP (1989) J Comp Chem 10:221–264
6. Stewart JJP (2007) J Mol Modeling 13:1173–1213
7. Stewart JJP (2008) J Mol Modeling 14:499–535
8. Pople JA, Santry DP, Segal GA (1965) J Chem Phys 43:S129–S135
9. Lee T-S, Lewis JP, Yang W (1998) Comp Mater Sci 12:259–277
10. Dixon SL, K. M. Merz J (1996) J Chem Phys 104:6643–6649
11. Dixon SL, Merz MM (1997) J Chem Phys 107:879–893
12. Vaart Avd, Gogonea V, Dixon SL, Jr. KMM (2000) J Comp Chem 21:1494–1504
13. Vaart Avd, Suárez D, K. M. Merz J (2000) J Chem Phys 113:10512–10523
14. Lin H, Truhlar D (2007) Theor Chem Acc 117:185–199
15. Stewart JJP (1996) Int J Quant Chem 58:133–146
16. Stewart JJP (2007) MOPAC2007. Stewart Computational Chemistry, Colorado Springs, CO
17. Yang W, Lee T-S (1995) J Chem Phys 103:5674–5678
18. Baker J (1986) J Comp Chem 7:385
19. Broyden CG (1970) J Inst Math Appl 6:222–231
20. Fletcher R (1970) Comput J 13:317–322
21. Goldfarb D (1970) Math Comput 24:23–26
22. Shanno DF (1970) Math Comput 24:647–656
23. Nocedal J (1980) Math Comput 35:773–782
24. Liu DC, Nocedal J (1989) Math Program B 45:503–528
25. http://www.pdb.org/; 98 (2007) Research Collaboratory for Structural Bioinformatics, The San Diego Supercomputer Center, San Diego, CA
26. Becke AD (1993) J Chem Phys 37:5648–5652
27. Luecke H, Schobert B, Richter H-T, Cartailler J-P, Lanyi JK (1999) J Mol Biol 291:899–911
28. Fossey SA, Némethy G, Gibson KD, Scheraga HA (1991) Biopolymers 31:1529–1541
29. Takahashi Y, Gehoh M, Yuzuriha K (1999) Int J Biol Macromol 24:127–138
30. Ormö M, Cubitt AB, Kallio K, Gross LA, Tsien RY, Remington SJ (1996) Science 273:1392–1393
31. Makabe K, Tereshko V, Gawlak G, Yan S, Koide S (2006) Protein Sci 15:1907–1914
32. Liou Y-C, Tocilj A, Davies PL, Jia Z (2000) Nature 406:322–324
33. Raetz CR, Roderick SL (1995) Science 270:997–1000
34. Teeter MM, Roe SM, Heo NH (1993) J Mol Biol 230:292–311
35. Martin C, Richard V, Salem M, Hartley RW, Mauguen Y (1982) Nature 297:162–164
36. Bella J, Eaton M, Brodsky B, Berman HM (1994) Science 266:75–81
37. Luecke H, Richter H, Lanyi JK (1998) Science 280:1934–1937
38. Jelsch C, Teeter MM, Lamzin V, Pichon-Pesme V, Blessing RH, Lecomte C (2000) Proc Natl Acad Sci USA 97:3171–3176
39. Morais JH, Cabral1 YZRM (2001) Nature 414:37–42
40. Rutenber E, Katzin BJ, Ernst S, Collins EJ, Mlsna D, Ready MP, Robertus JD (1991) Proteins 10:240–250
41. Paoli M, Liddington R, Tame J, Wilkinson A, Dodson G (1996) J Mol Biol 256:775–792
42. Tronrud DE, Matthews BW (1993) Refinement of the structure of a water-soluble antenna complex from green photosynthetic bacteria by incorporation of the chemically determined amino acid sequence. In: Deisenhofer J, Norris JR (eds) The photosynthetic reaction center 1. Academic, San Diego

43. Fallon J, Halling D, Hamilton S, Quiocho F (2005) Structure 13:1881–1886
44. Kurisu G, Kai Y, Harada S (2000) J Inorg Biochem 82:225–228
45. Just VJ, Stevenson CEM, Bowater L, Tanner A, Lawson DM, Bornemann S (2004) J Biol Chem 279:19867–19874
46. Anand R, Dorrestein PC, Kinsland C, Begley TP, Ealick SE (2002) Biochemistry 41:7659–7669
47. Tahirov TH, Misaki S, Meyer TE, Cusanovich MA, Higuchi Y, Yasuoka N (1996) J Mol Biol 259:467–479
48. Murshudov GN, Krzywda S, Brzozowski AM, Jaskolski M, Scott EE, Klizas SA, Gibson QH, Olson JS, Wilkinson AJ (1998) Biochemistry 37:15896–15907
49. Williams PA, Cosme J, Sridhar V, Johnson EF, McRee DE (2000) Mol Cell 5:121–131
50. Wuerges J, Geremia S, Fedosov SN, Randaccio L (2007) IUBMB Life 59:722–729
51. Inoue T, Suzuki S, Nishio N, Yamaguchi K, Kataoka K, Tobari J, Yong X, Hamanaka S, Matsumura H, Kai Y (2003) J Mol Biol 333:117–124
52. Lu D, Klug A (2007) Proteins 67:508–512
53. Omichinski JG, Sakaguchi K, Clore GM, Gronenborn AM, Appella E (1992) J Protein Chemistry 11:408–409
54. Briercheck DM, Wood TC, Allison TJ, Richardson JP, Rule GS (1998) Nat Struct Biol 5:393–399
55. Hille R (2002) Trends Biochem Sci 27:360–367
56. Schneider F, Lowe J, Huber R, Schindelin H, Kisker C, Knablein J (1996) J Mol Biol 263:53–69
57. Yennawar NH, Yennawar HP, Farber GK (1994) Biochemistry 33:7326–7336
58. Dewar MJS, Healy EF, Stewart JJP (1984) J Chem Soc, Faraday (II) 80:227–233
59. Peng C, Schlegel HB (1993) Isr J Chem 33:449–454
60. Bartels RH, Golub GH (1968) Commun ACM 11:401–406
61. Fukui K (1981) Accounts Chem Res 14:363–368
62. Klei HE, Stewart JJP (1986) Int J Quant Chem Symposium 20:529–540
63. Perkins PG, Stewart JJP (1980) J Chem Soc, Faraday (II) 76:520–533
64. Avitabile G, Napolitano R, Pirozzi B, Rouse KD, Thomas MW, Willis BTM (1975) J Polym Sci 13:351–355
65. Hageman JCL, Meier RJ, Heinemann M, Groot RAd (1997) Macromolecules 30:5953–5957
66. Barrera GD, Parker SF, Ramirez-Cuesta AJ, Mitchell PCH (2006) Macromolecules 39:2683–2690
67. Sinsawat A, Putthanarat S, Magoshi Y, Pachter R, Eby RK (2002) Polymer 43:1323–1330
68. Sinsawat A, Putthanarat S, Magoshic Y, Pachter R, Eby RK (2003) Polymer 44:909–910
69. Kramer RZ, Vitagliano L, Bella J, Berisio R, Mazzarella L, Brodsky B, Zagari A, Berman HM (1998) J Mol Biol 280:623–638
70. Lorenzo AC, Caffarena ER (2005) J Biomech 38:1527–1533
71. Buehler MJ (2006) Proc Natl Acad Sci USA 103:12285–12290
72. Yu N, Yennawar HP, K. M. Merz J (2005) Acta Cryst 61:322–332
73. Yu N, Li X, Cui G, Havik SA, K. M. Merz J (2006) Protein Sci 15:2773–2784
74. Klamt A, Schüürmann G (1993) J Chem Soc Perkin Trans 2:799–805
75. Vaney MC, Maignan S, Riès-Kautt M, Ducruix A (1996) Acta Cryst 52:505–517
76. Wada M, Sakurai M (2005) J Comp Chem 26:160–168
77. Hakanpää J, Paananen A, Askolin S, Nakari-Setälä T, Parkkinen T, Penttilä M, Linder MB, Rouvinen J (2004) J Biol Chem 279:534–539
78. Nagata K, Tanokura M (2005) Pept Sci 2004:441–445