Density-functional geometry optimization of the 150 000-atom photosystem-I trimer

Peter Canfield and Mats G. Dahlbom
School of Chemistry, The University of Sydney, New South Wales 2006, Australia

Noel S. Hush
School of Chemistry, The University of Sydney, New South Wales 2006, Australia and School of Molecular and Microbial Biosciences, The University of Sydney, New South Wales 2006, Australia

Jeffrey R. Reimers[a)]
School of Chemistry, The University of Sydney, New South Wales 2006, Australia

(Received 21 September 2005; accepted 10 November 2005; published online 9 January 2006)

We present a linear-scaling method based on the use of density-functional theory (DFT) for the system-wide optimization of x-ray structural coordinates and apply it to optimize the 150 000 atoms of the photosystem-I (PS-I) trimer. The method is based on repetitive applications of a multilevel ONIOM procedure using the PW91/6-31G(*d*) DFT calculations for the high level and PM3 for the lower level; this method treats *all* atoms in the structure equivalently, a structure in which the majority of the atoms can be considered as part of some internal "active site." To obtain a realistic single structure, some changes to the original protein model were necessary but these are kept to a minimum in order that the optimized structure most closely resembles the original x-ray one. Optimization has profound effects on the perceived electronic properties of the cofactors, with, e.g., optimization lowering the internal energy of the chlorophylls by on average 53 kcal mol^{-1} and eliminates the enormous 115 kcal mol^{-1} energy spread depicted by the original x-ray heavy-atom coordinates. A highly precise structure for PS-I results that is suitable for analysis of device function. Significant qualitative features of the structure are also improved such as correction of an error in the stereochemistry of one of the chlorophylls in the "special pair" of the reaction center, as well as the replacement of a water molecule with a metal cation in a critical region on the C_3 axis. The method also reveals other unusual features of the structure, leading both to suggestions concerning device functionality and possible mutations between gene sequencing and x-ray structure determination. The optimization scheme is thus shown to augment the molecular modeling schemes that are currently used to add medium-resolution structural information to the raw scattering data in order to obtain atomically resolved structures. System-wide optimization is now a feasible process and its use within protein x-ray data refinement should be considered. *© 2006 American Institute of Physics*. [DOI: 10.1063/1.2148956]

I. INTRODUCTION

Proteins and protein-cofactor complexes comprise the most diverse group of biologically functional molecules mediating, in one way or another, the great majority of biochemically relevant reactions. Because of this central role in biochemistry, scientists have long sought to model these chemical systems, but this effort is made difficult by their large size, the complex nature of the intramolecular forces, and the inherent difficulty in determining what shape these molecules take. The functions of proteins and complexes are diverse, providing biological systems with structural, motile, and catalytic features, most often in complex ways by interactions with other systems. These interactions occur at specific positions in the molecules and comprise the "active" regions in terms of function as opposed to simply providing a structural scaffold.

Photosystem-I (PS-I) is a particularly interesting protein complex found in all oxygenic photosynthetic organisms that is involved in the conversion of light energy into a chemical potential for use by the organism.[1] PS-I is also a particularly challenging system to model, as it is a member of the *trans*-membrane class of protein complexes and in *Synechococcus elongatus*[2] is composed of some 12 protein chains and a host of cofactors including 96 chlorophylls, four lipids, three 1,2dipalmitoylphosphatidylglycerole (LHG), one 1,2distearoylmonogalactosyldiglyceride (LMG), 22 β-carotenes, two phylloquinones, three Fe_4S_4 clusters, and 201 specifically identified water molecules. Additionally PS-I can exist as either a monomer or trimer or both depending on the organism and conditions.[3–5] The cofactors are situated throughout much of the PS-I complex with the chlorophylls, β-carotenes and quinones located in the *trans*-membrane region, and the Fe_4S_4 clusters in the stromal exomembrane region. Arguably each of these cofactors contributes to the overall function of PS-I through their absorp-

[a)] Author to whom correspondence should be addressed. Electronic mail: reimers@chem.usyd.edu.au tion of light, charge-transfer capabilities, and electronic couplings and thus they, and their immediate environment, must all be considered to be active regions of the functional complex.

Structural determination for large and complex systems such as PS-I constitutes a great achievement and, in general, makes use of NMR, x-ray crystallography, homology to other known proteins, and otherwise determined knowledge of the components. For x-ray crystallography, good quality single crystals of the native intact complexes must be carefully grown and heavy metal derivatives must be also obtained. From these crystals, several sets of diffraction data are usually produced and combined to generate electron-density maps for further interpretation. For PS-I, x-ray crystallography has yielded an enormous amount of detailed structural information. The representative uncertainty in the atomic coordinates provided can be estimated in various ways. The Luzzati and SigmaA values as reported by Jordan *et al.*[2] are 0.32 and 0.36 Å, respectively, while the diffraction-component precision index (DPI) of Cruickshank evaluates to 0.21 Å. This is sufficient, when combined with *molecular mechanics* models, to permit atomically resolved model structures of *medium* resolution to be determined, at least for all regions in which the participating amino-acid sequence and cofactor components are already known. It must be stressed that the coordinate sets published are molecular mechanics model *interpretations* of the x-ray experimental electron density. Many properties relating to the function of PS-I and proteins in general are sensitive to very *high*-resolution features of the structure,[6] however, features that cannot be reliably produced using the molecular mechanics models. Further, the molecular mechanics models, with their parameters optimized for proteins, perform particularly poorly for cofactors. Consequently, any electronic structure calculations done on cofactors *must* involve optimization of the cofactor at the very least, and preferably of the surrounding matrix. Until recently, systematic improvement of such high-resolution features has not been possible, but reliable linear-scaling quantum-chemical computational methods are becoming available for this purpose.[7–10]

X-ray analysis yields single static structures that are time and site averages of actual protein structures. As a result, disordered regions that may be present in the protein lead to either unassignable regions of electron density or to structures that represent an average of the actual instantaneous geometries. Such occurrences are conveniently flagged in published x-ray structures through the included isotropic B factor defined for each atom i from the Debye-Waller equation as

$$B_t = 8\pi^2\langle u_t\rangle^2, \tag{1}$$

where u_i is the uncertainty in the atomic coordinates averaged over zero-point and thermal motions as well as any structural variations and limitations set by factors such as the intrinsic resolution and the number of x-ray reflections collected; for PS-I, the published[2] uncertainties range from 0.4 Å for well-resolved features to 1 Å for poorly resolved ones. No single atomically resolved structure can properly represent structural variations within the crystal, and computationally this issue is usually addressed using molecular-dynamics simulations and ensemble averaging. The widespread success of x-ray structural analysis in interpreting protein function arises as the majority of the protein structure, including the most significant features, is usually well defined. However, standard coordinates from x-ray analyses in regions of structural flexibility are not usable in analyses of function as the deduced "average" structures contain unphysical bond lengths, etc., and are not realistic instantaneous configurations. Optimization of the x-ray coordinates enhances the structure by converting such regions into physically realistic instantaneous configurations that are useful in subsequent analyses of protein electronic structure and function.

In general, the x-ray structures of large biochemical systems are not directly suitable for studies of protein function as side chains, whole residues, counterions, solvent molecules, etc., are often missing while the hydrogen atoms are typically absent. Before optimization can commence, the original protein model depicted in the x-ray structure must be enhanced to overcome these limitations. Throughout, we take a minimalistic approach to changes in the protein model, producing the simplest physically realistic structure embodying all known functionality that is consistent with the original x-ray coordinates. In this process, e.g., missing key parts of cofactors are completed but missing residues are not added; also, some water molecules are added so as to stabilize surface ions, but a complete solvation shell is not added. As a consequence some minor structural anomalies persist in the final structure, but the final structure is overall as close as possible to the original x-ray structure.

A critical issue is, however, the presence and/or location of hydrogen atoms. While in most cases these may be added reliably using simple concepts of valence, a variety of feasible possibilities arises for rotatable groups such as the hydroxy groups found on serine (SER), threonine (THR) and tyrosine (TYR) residues, and rotatable molecules such as water. Further, the state of protonation of histidine (HIS) and other residues at physiological *p*H in *trans*-membrane proteins is often unclear and yet is typically critical to protein function. Also, key counterions such as Na^+, K^+, Mg^{2+}, and Ca^{2+} are difficult to properly identify, with Na^+ and Mg^{2+} being particularly difficult to distinguish from water, an isoelectronic molecule. Finally, the torsional conformations of acetyl groups, the imidazole ring of histidine, and the carboxamide group terminating the asparagine (ASP) and glutamine (GLN) side chains, etc., are not uniquely determined from the x-ray analysis and the actual orientations that are reported can be quite arbitrary.

The determination of molecular properties pertaining to protein function by quantum simulation, be they those associated with single or ensemble structures, should always involve some degree of geometry optimization. When quality high-resolution (i.e., <0.01 Å) information is essential, a procedure in common use[11] is that of a multilevel approach where, for example, an active site or small region of critical interest is modeled using a high level of theory, while the remainder of the system is treated using a computationally less demanding, lower level of theory such as molecular mechanics. Such a simple approach is not appropriate for PS-I as all of its many cofactors form "active sites" distributed throughout much of the complex, and for a complete description of the system to be obtained, almost the entire 150 000 atoms need to be considered for high-resolution optimization. We present a novel utilization of the ONIOM (Refs. 12–15) generalized multilevel method that allows for sequential optimization of the entire protein in an essentially *a priori* approach using quantum-mechanics methods that results in a linear-scaling algorithm. While more systemic linear-scaling semiempirical methods offer an alternate path for this procedure,[7–9] they are limited by intrinsic inadequacies in the methodologies and the need to treat a wide variety of atoms including metals and inorganic nanostructures. We use density-functional theory (DFT) to perform the quantum-chemical calculations, accessing the state of the art of high-resolution chemical structure optimization in large systems, with semiempirical methods applied to model weak interactions. The resulting model represents a single lowest-energy structural model at 0 K.

In Sec. II our minimal-change initial protein model is obtained starting with the recently published x-ray heavy-atom structure,[2] and in Sec. III we present our procedure for generating a reasonable guess at a starting geometry and the protonation state of the ionizable residues. In Sec. IV we describe our iterative piecewise procedure whereby the PS-I trimer is partitioned into computationally tractable fragments for geometry optimization at the DFT level using the PW91 functional[16] and a combination 6-31G(d)/6-31+G(d)
basis set, demonstrating that this level of optimization is practical for very large systems hitherto previously deemed inaccessible.

The major result sections are Secs. V and VI, with the final optimized coordinates supplied in full in supporting information. In Sec. V we describe key new qualitative features of the optimized structures involving modifications to the basic protein model. This includes a correction to the stereochemistry of $CL_2$1101 of the "special pair," proposes a metal cation in place of HOH143 on the C_3 axis, and the explicit solvation with water of various charged surface residues. In Sec. VI we discuss quantitative effects of optimization, providing a statistical analysis of the geometrical changes and energy distributions of the cofactors and amino-acid residues. Further, we show that PS-I presents an ideal system for demonstrating the applicability of our optimization method owing to the diversity and extent of its cofactors. In Sec. VII we identify problem residues and suggest alternative interpretations and implications for mechanism.

II. INITIAL PROTEIN MODEL

X-ray coordinates of PS-I by Jordan *et al.*[2] were used as published in the Protein Data Bank[17] as "1jb0.pdb" and converted into the HYPERCHEM format. HYPERCHEM (Ref. 18) was used to add hydrogens to the protein chains. Our own program then corrected chemically unrealistic hydrogen locations generated for the guanidino groups of the arginines (ARGs) and added hydrogens to the cofactors. During this process, the partitioning of the protein chains into the residues was preserved. However, the large cofactors such as carotenes, lipids, and chlorophylls that appear as single-residue molecules were each divided into a number of smaller "residues:" carotenes were split into two symmetrical fragments, chlorophylls had the phytyl chain and all other side chains split off as individual fragments, and the lipids were split into their fatty-acid end group as well as chemically intuitive components for the hydrophilic moiety. This division of the system into small units forms an essential part of the subsequent optimization procedure. In general, such subdivisions of molecules should strive to place the divisions between "residues" across non-delocalized bonds; this was not possible for β-carotene, however, but no adverse effects are anticipated in this case given the DFT level selection criteria as described in Sec. IV. All details of the partitioning are provided with the final optimized coordinates provided in supporting information.

The x-ray coordinates[2] of PS-I feature regions of the X chain containing residues that were absent from the known gene sequence of the protein. This region had been modeled in the x-ray coordinates as polyalanine, and we do not modify this. Also, the x-ray structure contains 91 missing residues; we do not add these but instead terminated intermittent chains with hydrogen atoms. In addition, 35 residues and a variety of cofactors are incompletely represented in the x-ray structure. We completed two missing proline (PRO) residues at 41-K and 57-K as for these the required configuration is unambiguous, but for all other residues the missing side chain was immediately terminated, essentially converting these into alanine (ALA) residues. All incomplete cofactor chains were immediately terminated with hydrogens, with the exceptions that at least the first carbon was retained from all phytyl chains, missing sp^2 centers were added, and the missing ester groups were added to chlorophyll (Chl) molecules $CL_1$1303 and $CL_1$1402 so as to complete the macrocyclic skeletons on all molecules.

III. ASSIGNMENT OF OXIDATION STATES, IONIZABLE RESIDUES, AND THE INITIAL HYDROGEN-BONDED NETWORK

Crystals for the x-ray structure were obtained at pH 6.4.[19] This is low enough to cause histidine to be protonated in aqueous solution, but the state of these residues *in vivo* in possibly hydrophobic pockets inside a *trans*-membrane protein can be quite variable. Initially, all histidine residues were assumed to be protonated except those ligated to magnesium, and similarly we assumed that all tyrosine residues are neutral and all protein-chain termini are ionized. As outlined in the Introduction, the OH groups of SER, THR, TYR, and water molecules are rotors as too is the SH group of cysteine (CYS), and these can take on a range of orientations. Similarly HIS, GLN, and ASN residues can adopt two orientations for their side chain terminal groups, as can deprotonated histidine residues. It is quite common for these rotatable groups, as well as the ionizable ones, to be involved in hydrogen bonding with other such groups, water molecules, and/or chain or cofactor fixed hydrogen-bond (H-bond) sites. Determining a realistic structure for these groups

TABLE I. Unusual protonation states of residues. [All residues by default are neutral except for ARG and LYS (protonated cations) and GLU and ASP (deprotonated anions.)]

Residue	Change	Reason
ASP113-B	Protonated	H-bonded to $CL_1$1206, hydrophobic region
HIS33-A	Protonated	On outside
HIS135-A	Protonated	On outside, increased H bonding
HIS633-A	Protonated	On outside
HIS33-B	Protonated	Not clear, in hydrophilic region
HIS121-B	Protonated	Near ASP367-B
HIS205-B	Protonated	Near ASP209- and ASP133-133
HIS241-B	Protonated	ASP367-B nearby
HIS368-B	Protonated	In hydrophilic region
HIS95-D	Protonated	Hydrophilic region, near ASP23-C
HIS63-E	Protonated	On outside
HIS50-F	Protonated	Near ASP46-F and GLU459-B
HIS3-J	Protonated	Near GLU26-A and LYS30-A

can thus be a complex multibody problem that requires thorough examination of the available configuration space and some means of estimating the relative energies of the various structures. Computational methods exist for protonation and rotor assignment including those implemented in the WHAT IF (Ref. 20) and REDUCE (Ref. 21) packages, but our needs are somewhat specialized in that we are considering a very large *trans*-membrane protein containing hydrophobic and hydrophilic regions for which we need an approximate *starting* structure for very many strongly coupled centers obtained without making any assumptions as to the nature of the hydrogen bonding. Our method is related to more comprehensive methods such as that of Nielsen and Vriend[22] for determining individual pK_a values, but considers only the electrostatic component of the interactions.

To proceed, we placed all rotatable groups at discrete positions on a sampling grid relative to the x-ray coordinates. In particular, the SH rotor of CYS, the OH rotors of SER and THR, along with those from the LHG and LMG cofactors were considered at 12 possible configurations, each separated by a 30° torsional motion, TYR OH at six locations, and GLN, ASN and each of the two isomers of neutral, deprotonated HIS (denoted HID and HIE) with two configurations. Water molecules (HOH) were considered at 72 different orientations obtained by pointing the dipole axis to one of the vertices of an icosahedron with sixfold rotations of the molecular plane about each axis direction. A total of 462 hydrogen-bond connectivity lists were then obtained containing all rotatable groups that could possibly interact with other such groups. The largest number of rotatable groups in a single hydrogen-bonded network was found to be 17, allowing for 9×10^{23} independent configurations of the product sampling grid. The vast majority of networks contain less than 10^7 independent configurations, and for these the energy of each configuration was determined and the lowest-energy configuration selected; for the seven networks with more than 10^7 configurations, a Monte Carlo sampling procedure was used to estimate the lowest-energy structure. The energies of the configurations were obtained considering only the electrostatic component of the total energy, obtained using standard AMBER (Ref. 23) charges for the residues and B3LYP/6-31G(d) charges obtained from the analysis of calculated molecular electrostatic potential.[24] These charges were evaluated using the GAUSSIAN 03 package.[25]

To provide a quantitative measure of the local environments around ionizable residues, the above calculations were repeated for each feasible ionization state. The prediction of enhanced environmental interactions for uncharged residues was taken as a clear indication that a particular residue is in fact not ionized; environments in which oppositely charged ions are located nearby, and those in which ions are surrounded by water molecules and/or other polar species clearly favor the ionized forms of the residues, however. In addition, for each ionizable residue, its local environment was quantified in terms of the density of nearby atoms (ionizable residues in the aqueous region of the system often have few resolved neighbors in the x-ray structure) as well as in terms of the local ratio of hydrophobic to hydrophilic groups. In most cases, these analyses led to clear assignments of the ionization state of the residues; the most ambiguous results occurred for only a small number of residues, and for these the ionization states were determined by visual inspection.

The result of this procedure is that the majority of the histidine residues were deprotonated; residues in ambiguous environments were kept protonated, however. Those that remained ionized, as well as other residues for which unusual ionization states were assigned, are listed in Table I along with a brief explanation.

The equilibrium oxidation states of the iron-sulfur complexes are not known with certainty. Calculations designed to investigate this question,[26] as well as parallels with other electron-transfer systems, suggest that the net charge on each $Fe_4S_4CYS_4$ unit is −2, indicating that the clusters are mixed-valence complexes containing two nominal Fe(II) centers and two nominal Fe(III) ones. We use this model and also assume that the clusters exist in singlet electronic states; while these details are critical to functional analysis, they do not affect the broader structural issues of interest herein.

Taking into account residues not explicitly included in the x-ray coordinates, the net charge of the PS-I monomer evaluates to −5. While this appears excessive, there is actually only one net charge per 10 000 atoms and so the charge density is quite low. This charge does not take into account the undetermined residues in the *X* chain, and the *K*-chain sequence is also only tentative and hence the actual charge may be slightly different from this value. Counterions may also be present that are not identified in the x-ray structure, either because their location is variable or because likely anions such as Na^+ and Mg^{2+} (the crystal was grown from $MgSO_4$ solution) are isoelectronic with water and may have been incorrectly assigned.

In a final preparatory step, all methyl groups were rotated using the same interconnected-network optimization scheme as was used to optimize the rotatable hydrogen-bonding residues. As a result, a single representative structure for the photosystem trimer was obtained in which there are no unphysically short intermolecular interactions other than those already identified as problem areas in the original x-ray structure.

The prepared structure was then subject to our optimization procedure, as detailed in the following sections. This optimization resulted in some minor changes to the original protein model: water HOH143 on the C_3 axis of the trimer was replaced with Na^+, modifying the net charge of the PS-I trimer to −14, and some 27 additional water molecules were added to the exomembrane protein surface to stabilize ionic states of the protein chains.

IV. OPTIMIZATION METHOD

The most common methods used for optimization of x-ray coordinates involves mixed quantum-mechanical and molecular-mechanical (QM/MM) methods but fully quantum-based molecular-orbital multilayer methods such as ONIOM (Refs. 12–15) are now available. Almost always these methods proceed by partitioning the system into a *single* active site that is treated at a high level of theory while the surroundings are treated using more efficient but less accurate methods. This approach is inadequate for PS-I, however, as the system contains not one but hundreds of regions that for problems of interest can be termed active sites. We use the generic ONIOM methodology for a single region but embed it in a higher-level approach that divides the *entire system* into computationally tractable and chemically sensible fragments, with each fragment being optimized while in the high-level region of *separate* ONIOM calculations. This process is well suited to modern massively parallel computer architectures as by grouping the fragments into noncontiguous subsets, thousands of ONIOM calculations may be launched in parallel. After each such subset is optimized, the protein coordinates are simultaneously updated. All subsets of fragments are optimized in turn, and the whole process is iterated until the geometry no longer changes. In practice there are many details to consider, however.

The computer time required to complete or iterative ONIOM calculations scales linearly with system size, a critical feature required for large systems. All linear-scaling methods for molecular electronic structure calculations rely in one way or another on molecular orbitals localizing spatially. For example, in implementations such as SIESTA (Refs. 27 and 28) this is achieved by the use of Wannier functions that go to zero value at a predetermined spatial extent. The semiempirical method in the MOZYME (Ref. 9) algorithm and the effective fragment potential[29] method of Ohta *et al.* both rely on localized molecular orbitals though these are approximate. Fortunately, the vast majority of organic systems are comprised of networks of low bond order or small delocalized systems within such networks that result in molecular-orbital (MO) descriptions that are naturally quite localized, making linear-scaling methods appropriate. Our approach has the advantage that the orbital localization is confined to the regions used in the high-level part of the ONIOM calculation and can thus be tailored to suit the chemical properties of the individual parts of the system. A similar approach has recently been introduced by Wada and Sakurai[10] and shown to be very reliable.

One of the most important technical aspects concerns the size of such fragments. The smaller the fragments that are optimized, the larger the number of iterations of the procedure that must be carried out, as this depends somewhat on the number of "fragment boundaries" in the system. Atoms at the boundaries are anchored by the stationary nonoptimizing neighboring atoms, and optimal treatment of the boundary regions is thus critical. Practical constraints limit the maximum size of fragments, however, due to the minimum $O(N^3)$ nature of quantum calculations, where N is the number of electrons included.

The ONIOM method was used as implemented in GAUSSIAN 03.[25] This method splits each calculation up into either two or three internal layers. All atoms to be optimized, as well as many of their surrounding ones, were included in the inner layer and treated with high-level methods. Typically 50–250 atoms were included in this layer, and it was treated using DFT with the PW91 (Ref. 16) density functional and typically the 6-31G(d) basis set. Only generalized-gradient-(GGA) type density functionals were considered for use due to their computational efficiency. PW91 was chosen after consideration of the structure of the hydrogen bonds and π-stacking interactions produced for a number of test cases. While DFT methods are known to produce high-quality optimized structures, hydrogen bonding and more importantly π-stacking interactions can be problematic. The computationally efficient 6-31G(d) basis set is well matched to this density functional for most structural properties but is inappropriate for the treatment of anions and lacks quantitative accuracy for hydrogen bonds. As very detailed analyses of hydrogen bonding require quantum treatment at a level well beyond that which is feasible in these calculations, we augmented the basis set to 6-31+G(d) for all anions but not for all hydrogen bonds. The structural mean error for this functional and basis set is expected to be of the order of 0.01 Å.[30] The second layer in the ONIOM calculation contained all residues for which an atom penetrated within 4 Å of any atom that is subject to optimization. Typically, this region contained 300–1300 atoms, and it was treated using the semiempirical quantum PM3 (Refs. 31 and 32) methodology. The iron-sulfur clusters could not be included in the second layer, however, as PM3 does not properly support Fe,

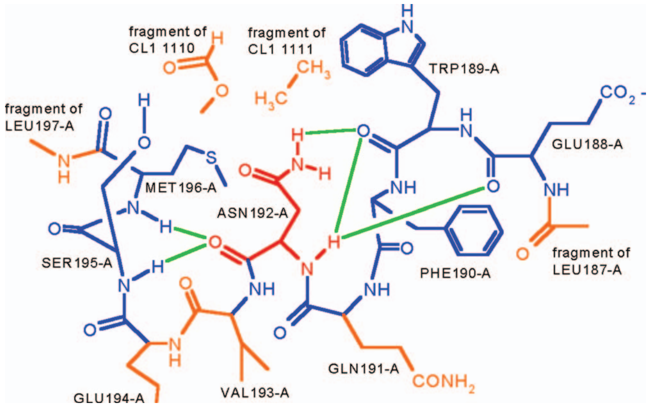

FIG. 1. (Color) Calculation constructed for the DFT optimization of ASN192-A (shown in red). Residues or part thereof included in the DFT level of the calculation but fixed in space are shown in blue. Residues or part thereof contained in the PM3 level of the calculation are shown in orange. H bonds involving ASN192-A are shown in green and determine the DFT/PM3 partitioning.

and hence these clusters were placed in an optional third layer of the ONIOM calculation that was treated using the universal force field (UFF).[33–35]

In practical terms, the key element in the construction of the optimization procedure is the method used to partition the atoms that are not being optimized between the inner (DFT) and outer (PM3, UFF) layers. All atoms whose presence significantly affects the optimized geometry need to be included in the inner layer, leaving the outer layers to account for exchange (steric) repulsion and polarization effects. The following rules were established.

- For amino-acid residues, delocalization across the peptide bonds requires at least two heavy atoms to be included from the neighboring residues in the protein backbone. The ONIOM procedure requires termination of the molecules used in each level, with the effect that the (nonproline) neighboring residues are treated internally as an acetyl group at the N terminus and as a methylamino group at the C terminus.
- For histidines, cysteines, water molecules, etc., that either form inter-residue covalent bonds such as disulfide linkages, or form ligands to Chl, metal, or Fe_4S_4 clusters, all interacting moieties are always kept together.
- All residues that hydrogen bond to the fragment being optimized are included in the inner layer.
- Where the fragment features an aromatic system then neighboring fragments that are deemed loosely π stacked to it (within 45° of parallel planes) are included in the inner layer. This feature was not included in the initial optimization cycles but was found to be essential.
- If two amino-acid residues in the same chain with numbers differing by two are included in the inner layer then the intervening residue is also included.

An example of the operation of these selection rules is provided in Fig. 1 concerning the optimization of residue ASN192-A. All optimized atoms are shown in red while the remaining atoms treated using DFT are shown in blue. This includes parts of the neighboring residues GLN191-A and VAL193-A, the residues to which ASN192-A forms hydrogen bonds GLU188-A, TRP189-A, SER195-A, and MET196-A, and the intermediary residue PHE190-A. Shown in orange are the atoms treated using only PM3, including the side chains of the neighboring residues GLN191-A and VAL193-A, as well as parts of the nearby residues LEU187-A, GLU194-A, LEU197-A, $CL_1$1110, and $CL_1$1111.

Poor convergence of the optimized geometry arises when strong interactions occur between two independently optimized fragment residues. To avoid this, groups of residues were marked for simultaneous optimization. Initially, all common ligands to a central metal atom were optimized together, although it was observed that this was not necessary for Chl ligation and, as these cofactors formed some of the largest subsystems for optimization, this process was not continued. The Chl molecules were internally divided into ten residues comprising the central macrocycle, the phytyl chain, and eight side chains. This dramatically reduced the size of the second layer in the ONIOM calculations, as often only a small side chain needed to be included in the neighborhood of a residue being optimized, not the entire chlorophyll. Note, however, that for optimization of the chlorophylls themselves, all residues except the phytyl chain were simultaneously optimized.

The region surrounding HOH143 forms a key link holding the three PS-I monomers together. For this, a very large region was simultaneously optimized comprising not only HOH143 on the C_3 axis but also 10 residues from *each* of the three monomers, these being per monomer HOH35, HOH71, HOH79, HOH125, HOH176, HOH177, HOH178, HOH180, SER42-L, and GLN119-L.

Finally, poorly convergent residues were identified during the optimizations and marked for special optimization where the inner (DFT) layer was extended to include extra neighboring residues. Many of these were acid-base pairs or linkages involving anomalously short hydrogen-bond lengths.

A. Fragment subsets and run sequence

When considering the strength of interactions between different fragment types, the fragments were divided into six distinct subsets on the basis of expected optimization efficiency and the chemical nature of the components. These subsets were

- the water–amino-acid cluster on the C_3 axis plus all manually defined cluster optimizations,
- all remaining odd numbered amino-acid residues,
- all remaining even numbered amino-acid residues,
- all remaining water molecules,
- the chlorophylls (without phytyl chains), Fe_4S_4 clusters, and one-half of each carotene, and

- the phytyl chains, phylloquinone, lipids, and the other half of each carotene.

In each cycle of the optimization procedure, all subsets were optimized at least once. However, as water molecules are discrete molecules that interact with other fragments through hydrogen bonding, small changes in the surrounding structure can result in large changes to their geometry. Given this and the fact that all water molecules in the structure are interacting with protein, the water subset was optimized immediately following each other subset in order to track these geometry changes. Also, as the amino-acid residues permeate the entire complex and generally interact strongly with other amino-acid residues through covalent and hydrogen bondings, the complementary odd and even subsets were always optimized successively and frequently during the overall iterative procedure. The cofactor subsets were less frequently optimized (usually once for every 2 or 3 odd/even amino-acid subset runs) as their geometries converged relatively quickly (they are generally isolated from each other by protein, have few hydrogen bonds, have ligands that are only relatively weakly bound, and interact with their environment often through poorly directional dispersive forces).

During the first two optimization cycles, the smaller 3 -21G(d) basis set was used and the heavy atoms were frozen at their x-ray coordinates. This facilitated the production of an intermediate optimized structure that retains as much as possible of the original crystallographic data while building within this constraint as realistically as possible a description of an all-atom structure. This is useful in interpreting the effects that the full coordinate relaxation had on the electronic structure of PS-I. In supporting information the full structure is provided after this extent of optimization.

V. RESULTS FROM THE OPTIMIZATION: MODIFICATIONS TO THE PROTEIN MODEL

As a result of the geometry optimization, various changes in the protein model appeared to be required in order to obtain a realistic structure of PS-I. In this section, these qualitative changes are described in detail.

A. Stereochemistry of the Chl-*a*′ $CL_2$1101

The x-ray structure clearly describes the presence of a Chl-*a*′ molecule named $CL_2$1101. Its structure is unusual, showing deformation of the ring plane and its surroundings, and this deformation has been considered in relation to the function of the molecule[36–39] one which forms half of the special pair associated with primary charge separation. Initial geometry optimization in the fixed field of the surrounding protein produced only the anticipated small changes to the structure of $CL_2$1101, a structure which appeared as a local minimum on the molecular potential-energy surface. At this point it was discovered that the stereochemistry of $CL_2$1101 was in fact that of a distorted Chl-*a*, possibly an artifact of an attempt at fitting a Chl-*a* structural template to the Chl-*a*′ electron density. After correcting the stereochemistry of $CL_2$1101 to that of Chl-*a*′ and subsequent reoptimization, including reoptimization of the surrounding protein, the energy of the molecule fell by 1400 kcal mol^{-1}. This removed all anomalies associated with the structure of this component of the special pair molecule.[36–39] Figure 2 illustrates the differences between the starting x-ray and optimized geometries for $CL_2$1101 with the unrealistic bond angle of 94° on the formally sp^3 stereocenter in the x-ray structure.

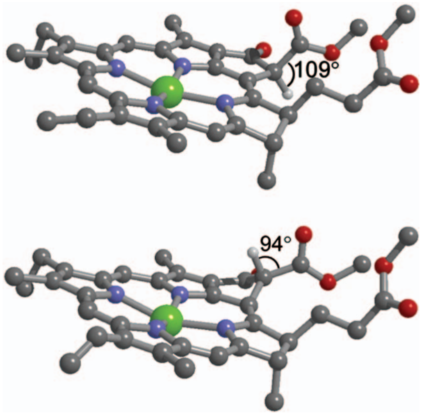

FIG. 2. (Color) $CL_2$1101 at x-ray and stereochemically corrected, optimized geometry. Phytyl chain and all hydrogens except epimeric hydrogen omitted for clarity.

B. Nature of intertrimer binding and HOH143

PS-I exists as a trimer with C_3 symmetry, the outline of which is shown in Fig. 3. There, the maximum lateral extent of each monomer is indicated, looking in the direction down the C_3 axis. A variety of weak interactions link one monomer to another, including six inter-residue hydrogen bonds (including $CL_1$1801—LYS159-B), one π-stacking interaction, and four hydrogen bonds involving water molecules; the major interaction, however, is associated with calcium ions CA 1001 (shown on the figure) in a binding pocket inside one monomer to which an extended chain from another is ligated. Note that the x-ray data are not in themselves sufficient to verify the presence of the calcium at the center of these clusters, but its presence is assumed based on chemical intuition and the observation of significant unaccounted electron density in the region. There is only one significant region of interaction near the C_3 axis that simultaneously links

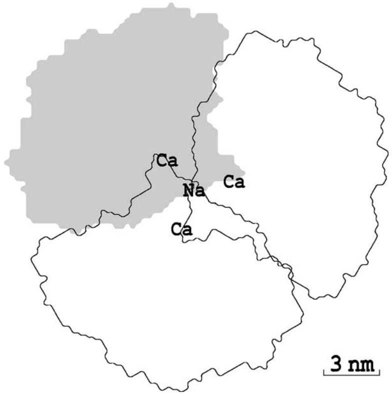

FIG. 3. View in the direction of the C_3 symmetry axis of the maximum lateral extent of the three PS-I monomer constituents (one shaded), indicating the locations of the calcium clefts and the proposed NA143 ion.

all three monomers, however, this is centered on HOH143 in the x-ray structure and is clearly apparent in Fig. 3. While this region could in principle play a key role in the trimerization process, the twofold symmetry of the water molecule acts to locally break the apparent threefold symmetry and this feature could actually disfavor trimerization.

As described previously, it was found necessary to simultaneously optimize a large hydrophilic region around HOH143 in order to obtain a structure that was chemically sensible. In this operation, many residues were allowed to break C_3 symmetry in order to obtain a representative configuration that featured a feasible hydrogen-bonding network. No satisfactory optimized structure was obtained, however, and optimization led to large changes in the location of HOH143 and of the surrounding structure. The optimization indicates that this structure is unstable and hence is not likely to contribute significantly to trimerization.

Two features of the original x-ray analysis suggest that the assignment of the observed electron density in the region to a water molecule may be incorrect. First, the nearby residues GLN119-L from each PS-I monomer are in very close contact. They appear on the anomalous heavy-atom separation list published with the x-ray structure,[12–15] the O–O separations between just 1.9 Å, a value not much larger than bond lengths in peroxide molecules. Optimization of the structure failed to satisfactorily relieve the large stresses associated with this configuration. Second, despite these residues being near the C_3 axis where their positional uncertainty as expressed by their B values [see Eq. (1)] could be expected to be quite small, they were actually moderately high at 30–40 for different atoms in the residue side chain. On the other hand, if HOH143 breaks the local symmetry then it must be in rapid equilibrium between a range of equivalent local structures and hence the experimental B factor should be quite large: it is reported as 20, however, much less than the values for the surrounding water molecules (27–69), indicating that its position is actually very well defined. The large B value for the neighboring oxygens can be interpreted not only as real uncertainty in their positions but also as an indication that the crystallographic refinement method was attempting to fit an incorrectly assigned atom.

The electron density that was assigned as HOH143 may also be assigned to say a Na^+ or Mg^{2+} ion, isoelectronic species to water. Sodium may be present as it is readily available *in situ*, while the reaction center's protein was crystallized from $MgSO_4$ solution and so magnesium could have either entered at this stage or been present *in situ*. It is also possible that a calcium ion also resides in this position, but this would have appeared more apparent in the x-ray structure due to its increased electron density. We considered only the simplest possibility that HOH143 should be reassigned as NA143. After making this substitution, the structure of the surrounding cluster was optimized in C_3 symmetry, forming a chemically sensible unit without major changes in the coordinates of the surrounding material. The resulting structure is shown in Fig. 4; note, however, that only the 157 optimized atoms are shown in this figure, not the 45 surrounding atoms in DFT calculation nor the remainder of the 1030 atoms included total. Also shown in the figure is a magnified view of the binding around the central metal, showing a stable pocket for the binding of species such as Na^+, Mg^{2+}, and Ca^{2+}.

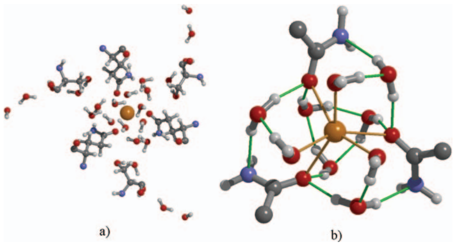

FIG. 4. (Color) Simultaneously optimized residues (*in situ*) associated with the central cluster about the C_3 axis showing the central Na^+ ion and the triplicate GLN119-L and SER42-L residues: (a) all residues and (b) detail of the coordination and H bonding about the central Na^+ ion, with coordinative bonds to the Na^+ ion in orange and H bonds in green.

Quantitatively, the internal cluster binding energy was calculated to be 70 kcal mol^{-1} more exothermic for Na^+ than for water. When provision is made for the electrostatic interaction of this cluster with the surrounding material (containing, for example, anions such as GLU115-L), at a (high-frequency) dielectric constant of $\epsilon=2$ the exothermicity increased to 103 kcal mol^{-1}. The water molecule or metal cation must be removed from the surrounding aqueous media, a process that is more endothermic for Na^+ than water by $\Delta G=80$ kcal mol^{-1} Hence, these calculations predict that Na^+ is favored at molecule 143 over water by 23 kcal mol^{-1} We thus replace HOH143 by NA143 in the protein model.

C. Surface ion stabilization

There were several instances of convergence problems for charged surface residues for which the surrounding material was not resolved in the original x-ray structure. Often, these groups optimized to enact proton exchange with neighboring residues, hence neutralizing nearby unsolvated charges. As this constitutes an unrealistic scenario, we solvated these charged groups with up to two water molecules each in order to stabilize the charges. This is a realistic procedure as surface water molecules cannot be directly seen in the x-ray experiment and solvation of the hydrophilic surface residues is expected. In all, 27 water molecules were added to the surface of the protein model. A minimalist approach was actually taken to the solvation of these outer residues as many more water molecules would be necessary in order to produce a quantitatively realistic structure. Our structure thus constitutes the smallest variation to the original x-ray structure that is consistent with the basic chemical properties of the protein. As a consequence, a total of nine residues remain with the proton on the expected side but with very short hydrogen-bond lengths in the region of 1.4–1.5 Å. This is the most significant remaining effects on the structure produced by the incomplete solvation, producing residues with large intramolecular strain energies as described in Sec. VII. While the quantum-chemical methods used are inadequate for the accurate description of real systems containing very short or shared hydrogen bonds, no compelling evidence is seen for the presence of any structure of this type in PS-I.

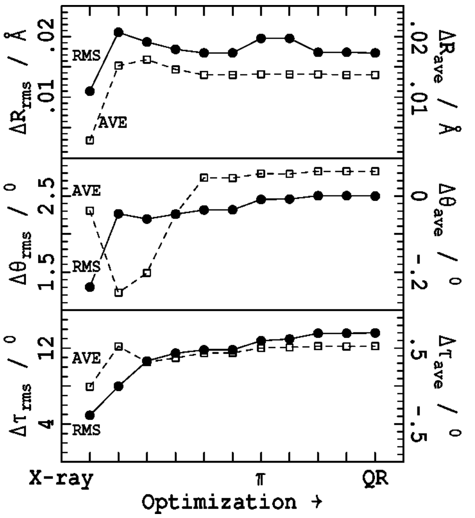

FIG. 5. Convergence of the root-mean-square (rms) and average (AVE) changes in heavy-atom bond lengths R, bond angles θ, and dihedrals τ, from the x ray to the optimized (OPT) structure, during the progress of the optimization. The point at which π-stacking interactions were included in the DFT level of the ONIOM calculation is indicated.

TABLE II. Changes $|\Delta x|$ and their standard deviation σ, both in Å, in the coordinates of the heavy atoms from the original x-ray structure to the optimized one.

PS-I components	Heavy atoms only		
	Mean $\|\Delta x\|$	σ	Max $\|\Delta x\|$
PsaA	0.116	0.145	1.955
PsaB	0.114	0.155	3.060
PsaC	0.126	0.140	0.845
PsaD	0.149	0.212	1.826
PsaE	0.137	0.192	1.933
PsaF	0.146	0.209	1.814
PsaI	0.131	0.197	1.419
PsaJ	0.172	0.262	1.982
PsaK	0.143	0.180	1.026
PsaL	0.122	0.162	1.422
PsaM	0.111	0.132	0.997
PsaX	0.175	0.194	1.019
Chlorophylls	0.151	0.145	1.588
β-carotenes	0.192	0.206	1.574
Waters	0.275	0.248	1.842
Other cofactors	0.146	0.150	1.138

VI. QUANTITATIVE RESULTS FROM THE OPTIMIZATION OF THE FINAL PROTEIN MODEL

In general, the optimization procedure led to the maintenance of qualitative structural elements, with only small changes in the positions of the heavy atoms. However, in all cases, the energies of the fragments decreased considerably, resulting in a significant release and equilibration of internal strain.

A. Convergence of the optimized geometry

Figure 5 shows the root-mean-square (rms) and average changes in the heavy-atom bond lengths R, bond angles θ, and dihedral angles τ from those in the starting geometry throughout the optimization procedure. These indicators, as well as a variety of others including the maximum deviations and changes in absolute Cartesian coordinates show rapid convergence with respect to cycling through the iterative ONIOM system-wide geometry optimization. The deviation in the trend midway during the optimization was due to the inclusion of π-stacking interactions within the DFT part rather than the PM3 part of the calculation. The final two configurations show no qualitative differences between hydrogen-bonding topologies, also indicating that the final structure is converged and stable.

B. Changes to atomic Cartesian coordinates

Table II shows the mean, standard deviation, and maximum changes to the Cartesian coordinates of the heavy atom from the modified starting geometry (x-ray heavy-atom coordinates with the orientation of the rotatable groups assigned and the stereochemistry of $CL_2$1101 corrected) to the final optimized (OPT) geometry. These are resolved into contributions from the different protein chains and cofactors. In all cases, the mean changes in nuclear positions are less than 0.3 Å with the standard deviations less than 0.3 Å. The maximum changes occur for surface residues whose surrounding atoms are not resolved in the x-ray structure; while these changes could have been minimized through the introduction of a complete hydrophilic surface hydration shell, or approach favored the production of a protein model that contains only the minimum essential changes to the original x-ray model. Also, the differences reported in the table between the protein chains and cofactors reflect largely the effects of the hydrophilic surface and are reasonably small at less than a factor of 2.

Shown in Fig. 6 is the probability distribution of the *difference* of the absolute optimization positional changes, $|\Delta x|$, compared with the expected mean vibrational motions, $\langle u^2\rangle^{1/2}$, as determined from the isotropic B factors using Eq. (1). This distribution is sharply distributed with a maximum at $|\Delta x|-\langle u^2\rangle^{1/2}=-0.55$ Å, indicating that the optimization procedure has moved most atoms by 0.55 Å less than the published uncertainty in its atomic coordinate. The net effect of the optimization is thus seen to be small compared with the experimental uncertainties and hence is physically realistic.

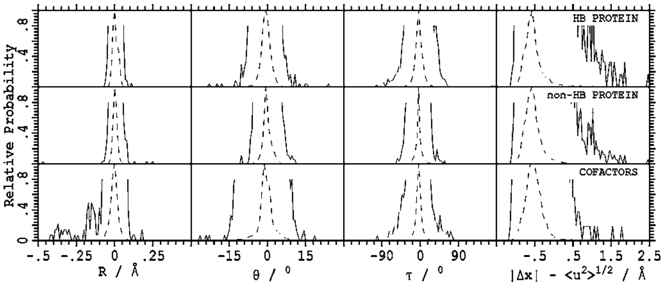

FIG. 6. Probability distributions for the changes on optimization of the heavy-atom bond lengths R, bond angles θ, and torsional angles τ as well as the norm of the Cartesian coordinate displacement $|\Delta x|$ less the experimental uncertainty in the atomic locations $\langle u^2\rangle^{1/2}$ [see Eq. (1)]. The results are partitioned into contributions from the cofactors, protein chain components involving N or O hydrogen bonding, and protein components that are free from hydrogen bonding. $\cdots\times 1$, $-\times 100$.

However, there are a relatively small but significant number of atoms that move considerably more than these uncertainties, up to $|\Delta x| - \langle u^2 \rangle^{1/2} = 2.5$ Å. An exhaustive analysis of these atoms with relatively large motions ($|\Delta x| - \langle u^2 \rangle^{1/2} > 0.8$) revealed that most were associated with the surface of the PS-I complex. Such motions are commonly associated with conformationally free side chains and thus do not appear to be important to the photophysical properties of PS-I for which this model was developed. All nonsurface atoms with large motions were found to be associated with strongly H-bonding groups, notably the terminal carboxamide groups of ASN and GLN residues. Individual visual inspection of these geometries revealed the optimized structures to be physically reasonable with much stronger intermolecular interactions than perceived in the original x-ray coordinates.

As reported earlier, anomalously large changes to the x-ray structure did eventuate when HOH143 was included, as specified in the original structure. This leads to its replacement by NA143 in protein model. This indicates that the optimization method can identify significant qualitative issues concerning the structure. The appearance of unexpected large structural variations provides an indicator for possible problems with the original assignment of the x-ray electron-density information.

C. Changes to bond lengths, bond angles, and torsional angles

While the average and rms changes to the heavy-atom bond lengths, bond angles, and torsional angles during the optimization are shown in Fig. 5, Fig. 6 shows the probability distributions of the changes in each property from the initial structure to the final optimized one. The analysis in Fig. 6 is presented in terms of the properties of the cofactors and properties of the protein in which hydrogen bonding to N or O is involved or not. The changes in bond length are sharply peaked around the average and rms values of ca. 0.015 Å, a value that is well within the accuracy of the molecular mechanics methods used to produce the original x-ray coordinates. However, the distribution contains long tails indicating that some bond lengths gave changes by enormous amounts up to 0.5 Å. All structures involving changes in excess of 0.1 Å have been examined, and in each case the initial structure did not depict a physically realistic single configuration of the system. The largest change was the disulfide bond between residues CYS8-F and CYS43-F that changed from 2.55 Å to the realistic value of 2.09 Å. Such a change may be indicative that several conformers exist in the protein crystal. Most other large changes are associated with the cofactors, particularly the Fe–S bond lengths and some Chl bond lengths, and it is readily apparent that the molecular mechanics methods used in the x-ray refinement are of substantially higher quality for treatment of amino-acid residues than they are for arbitrary molecules; this is understandable as these empirical methods were developed specifically to treat protein chains.

The changes to the bond and torsional angles show similar trends. They have central distributions that are much narrower and near zero for the nonhydrogen-bonded amino acids; these are broadened for the cofactors and especially the hydrogen-bonded protein. Broader distributions for the hydrogen-bonded amino acids are expected as the molecular mechanics methods used in the x-ray structure optimization do not use explicit hydrogen-bond information in determining the intermolecular interactions and hence the distortions that are required in order to optimize specific intermolecular interactions are not properly reproduced. Again, some instances of very large geometry changes are produced, however, and again manual inspection reveals that in most cases the original x-ray structure did not depict a physically realistic single configuration. In four cases the optimized structure rather than the x-ray one showed unexpected bond angles, but these residues all had extremely strong intermolecular interactions to drive the distortion. The most significant of these is ASP54-F for which the nominally tetrahedral C_α–C_β–C_γ angle optimized to 124° under the influence of a sterically hindered interaction with ARG56-F. Incomplete solvation of surface residues does introduce slight distortions in the optimized bond angles, but was a common occurrence for the bond torsions. The only heavy-atom torsional angle for a rotatable residue to change from the conformation used in the original structure was GLN718-A, producing the near-180° changes recorded in Fig. 6. All torsional angles that changed by more than 55° were examined manually, as well as all structures that moved significantly towards more eclipsed conformations. These changes were all induced by either incomplete hydrophilic surface solvation or strong intermolecular interactions; strong intermolecular interactions are features that are not properly represented in the molecular mechanics force field used for the original geometry optimization.

D. Changes to the Ramachandran plot indicating the protein secondary structure

A major triumph of x-ray protein crystallography has been the determination of the secondary structure of proteins, a feature that is quantified through Ramachandran plots of the torsional angles ϕ (C–N–C_α–C) and ψ (N–C_α–C–N) for each residue. Shown in Fig. 7 is the probability distribution for the change in the norm $(\Delta\phi^2 + \Delta\psi^2)^{1/2}$ of these angles [Fig. 7(a)], the Ramachandran plot indicating the density of different types of configurations in the structure [Fig. 7(b)], as well as vectors indicating the changes for all residues with $(\Delta\phi^2 + \Delta\psi^2)^{1/2} > 30°$ [Fig. 7(c)]. While the most likely change is only of order 5°, changes of up to 70° are found. Shown in the figure are colorings[40] indicating the relative likelihood of different (ϕ, ψ) pairs in protein structures, with regions of α-helix-type colored green, β-sheet-type colored blue, and near γ-turn colored cyan. Indeed, the majority of the points in the Ramachandran plot fall into the expected α-helix and β-sheet categories. Residues in regions colored yellow are also commonly found in protein structures, while those in brown-colored regions are rare and those in red-colored ones are very rare. A significant number of residues are found in rare to very rare regions with $\phi \sim 60°$, these being mainly associated with protein

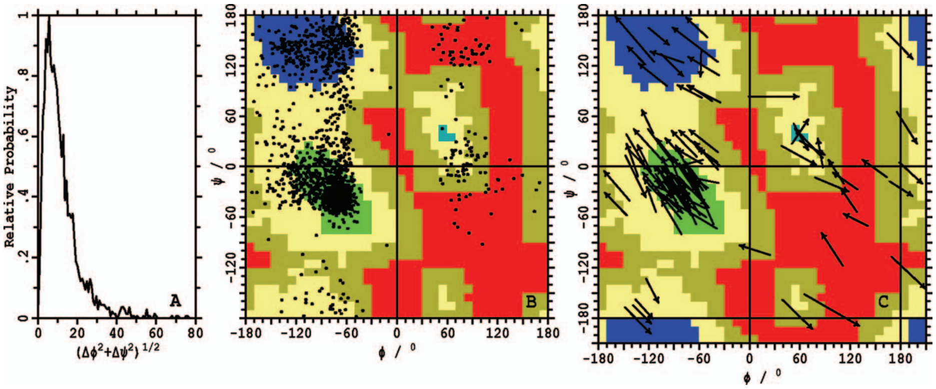

FIG. 7. (Color) Secondary structure of PS-I: (a) probability distribution for the norm of the changes on optimization of the C–N–C_α–C torsional angle ϕ and the N–C_α–C–N angle ψ, (b) Ramachandran plots indicating the distribution of protein residues with simultaneous angles ϕ and ψ, and (c) arrows indicating all large changes from the initial x-ray structure on optimization. The shaded regions indicate (Ref. 40) general regions in protein structures of high probability (green: α-helix-like, blue: β-sheet-like, and cyan: γ-turn-like), commonly occurrence (yellow), rare occurrence (brown), and very rare occurrence (red).

turns. The basic qualitative features of Fig. 7(b) for the optimized structure closely parallel those for the original x-ray data, and the secondary structure changes on optimization shown in Fig. 7(c) reveal that optimization takes only a few residues into or out of very rare regions. Manual inspection reveals that these changes are not centered at any particular region of the protein but rather are associated with incomplete residues, turn residues in the hydrophilic region, and glycine (GLY) and PRO residues whose nature is to give anomalous distributions on the Ramachandran plot due to reduced internal torsional constraints. Hence it is clear that the optimization procedure has maintained the essential nature of the secondary structure present in the original x-ray coordinates, a feature that the molecular mechanics methods used in its generation are optimized to produce.

E. Changes to the intramolecular energies of the residues and cofactors

The primary purpose of this work is to develop a structure for PS-I that is suitable for electronic structure analysis in order to ascertain the key chemical and photophysical properties responsible for its function. Such properties depend critically on small changes in intramolecular bond lengths and bond angles, properties for which large changes are recorded in Fig. 6. The simplest property is the intramolecular electronic energy of the components of the system; any changes found for this property will be magnified when reactivity or excited-state properties are considered.

Fortunately, the large number of similar components in PS-I allows for comparative statistics of intramolecular energies to be obtained and analyzed to isolate anomalous features. Energy distributions for some notable amino-acid residues, as well as the chlorophyll and carotene molecules, are shown in Fig. 8, while a complete catalog is provided in supporting information. Distributions are provided not only for the final optimized structure but also for the modified starting structures (x-ray coordinates for the heavy atoms after initial hydrogen atom placement and hydrogen atom optimization). These energies were calculated on the isolated fragments *in vacuo* using the PW91 density functional as per the geometry optimization calculations. In the case of the amino acids, only internal chain residues were considered, with each residue being terminated with hydrogen atoms. A summary of the results obtained for all species in provided in Table III, including the average energy decrease on optimization, $\Delta E_{\text{average}}$ and, for final structures, the largest variations found in the residue energies as well as the associated standard deviations σ. The normal distribution curves for the final optimized structures expected based on the calculated average residue energy and its standard deviation σ are also shown in this figure.

In Table III, the results are grouped into seven types, each type being typified by the chemical nature of the spe-

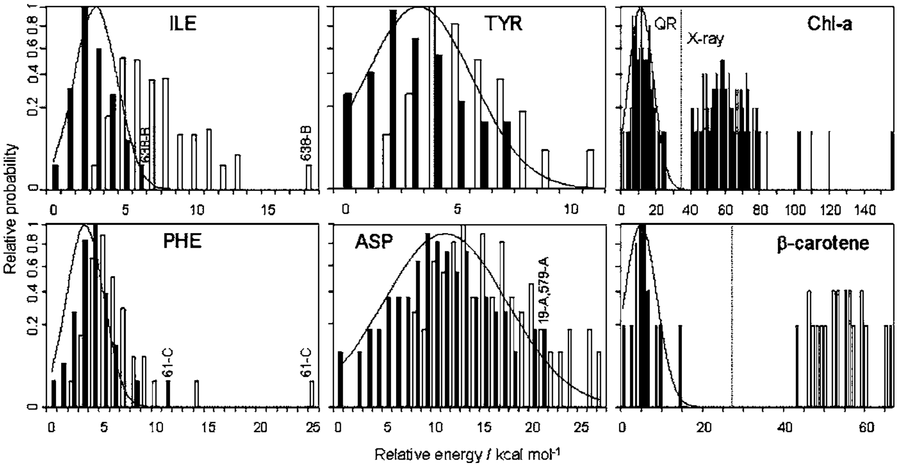

FIG. 8. Distributions of single point energies for a selection of amino acids and the chlorophyll and β-carotene cofactors. Energies are calculated at PW91/6-31G(d) level for isolated, terminated residues at their *in situ* geometry. ■ = fully optimized geometry and □ = heavy atoms at original x-ray coordinates. Superimposed are normal distribution curves obtained from the average energies and their standard deviations for the optimized data. Note that for the chlorophyll and β-carotene cofactors there are no overlaps of the distributions between the x-ray and optimized (QR) sets.

TABLE III. Comparative single point energy statistics between the modified (hydrogen added and optimized) x-ray and the optimized (OPT) structures for the amino acids, chlorophylls, and β-carotenes. The energies, in kcal mol^{-1}, are calculated at the PW91/6-31G(d) level.

			E range		σ	
R group	Residue	$\Delta E_{\text{average}}$	X ray	OPT	X ray	OPT
Neutral, non-H bonding	GLY	−1.19	26.30	8.27	3.01	1.22
	ALA	−0.78	21.10	8.04	2.17	1.13
	VAL	−3.32	17.73	7.75	3.20	1.42
	LEU	−5.75	25.20	9.55	4.56	1.40
	ILE	−4.41	14.05	6.80	2.17	1.07
	PRO	−2.66	11.27	6.54	1.73	1.44
	PHE	−1.77	22.16	10.25	2.25	1.24
-SMe	MET	−6.85	38.35	7.58	6.19	1.60
Neutral, H bonding, aromatic	TRP	−2.25	13.21	10.62	1.96	1.78
	TYR	−2.37	8.97	7.53	1.57	1.69
	HID	−5.22	5.88	5.33	1.25	1.29
	HIE	−4.83	8.51	5.95	1.91	1.34
Neutral, H bonding, nonaromatic	SER	−2.08	15.83	11.56	2.46	2.30
	THR	−3.40	39.67	12.39	4.37	2.30
	ASN	−2.52	20.23	9.27	2.97	2.08
	GLN	−3.19	15.08	14.79	2.83	2.36
Anionic, H bonding	ASP	−4.56	19.00	21.32	4.29	4.48
	GLU	−3.88	18.36	20.53	3.63	4.58
Cationic, H-bonding	HIS	−1.94	5.82	18.75	2.10	4.89
	LYS	−4.23	36.44	18.30	4.99	3.52
	ARG	−5.75	33.17	16.21	6.43	3.31
Cofactors	Chlorophyll	−52.57	114.98	24.41	16.96	4.98
	Carotene	−48.65	22.71	13.13	6.12	2.79

cies using descriptors such as charge state, hydrogen-bonding profile, and aromaticity. For the neutral, non-hydrogen-bonding residues GLY, ALA, valine (VAL), lencine (LEU), isolencine (ILE), PRO, and phenylalanine (PHE), the full optimization of the heavy-atom coordinates resulted in a decrease in the range of the residue energies by a factor of 2–3, indicating that a more internally consistent set of geometries has been obtained, consistent with the analysis of the bond-length changes given in Fig. 6. Also, the standard deviations of these distributions are in the range of 1–2 kcal mol^{-1}, of the order of that expected for a large thermal ensemble. Further, the distributions themselves appear Gaussian in shape, as indicated in Fig. 8 explicitly for ILE and PHE. The x-ray structures for the aromatic H-bonding residues tryptophan (TRP), TYR, HID, and HIE show significantly reduced energy spreads at the x-ray geometry than found for the neutral non-H-bonding residues, however, but the spreads become similar after optimization. This is indicative of the internally varying quality of the force field used in the original x-ray refinement.

For charged and strongly hydrogen-bonding residues, the distributions are significantly broadened compared with the aromatic and non-hydrogen-bonding residues. For example, the optimized distributions for the nonaromatic H-bonding residues SER, THR, ASN, and GLN increase in width to 9–15 kcal mol^{-1} from 5 to 10 kcal mol^{-1} while the standard deviations increase to 2.1–2.4 kcal mol^{-1} from 1.1 to 1.8 kcal mol^{-1}; the broadening is clearly seen in Fig. 8 for TYR. This arises as these intramolecular energy calculations do not take into account the effects of the environment in distorting the molecular structures: these effects increase in magnitude with the strength of the intermolecular interactions formed with the residues. For some of the ionic residues such as ASP, GLU, and especially HIS, the optimization actually resulted in an *increase* of the intramolecular energies compared with the original x-ray coordinates. Since hydrogen-bonding interactions are not realistically modeled with molecular mechanics, the x-ray coordinates for strongly hydrogen-bonding residues do not reflect the true variation in geometry and hence the spread of single point energies is underestimated.

The intramolecular energies for chlorophyll and carotene molecules, unlike those for amino acids, do not show any overlap in distributions between those for the starting geometry and those for the optimized structure, with the optimization lowering the energies by on average of 50 kcal mol^{-1}! This result highlights the relative treatment of various structures by the molecular mechanics model used in the x-ray structural determination; amino acids are modeled reasonably well but cofactors much less so. It is unexpected as the changes in bond lengths for the cofactors shown in Fig. 6 indicate that most bond lengths are realistic in the x-ray coordinates with only a few showing exceptional distortion. However, only one bond length per molecule needs be incorrectly represented in order for the electronic structure, chemical reactivity, and spectroscopic properties to be dramatically modified. A consequence of this is that x-ray coordinates for cofactors in protein complexes are much less reliable and should not be used unoptimized in electronic structure calculations. Previously, we have shown that incorrect qualitative conclusions have been drawn concerning the nature of the special pair in purple bacteria based on the use of raw x-ray heavy-atom coordinates;[6] subsequently, we show that the Q-band absorption energy profile for PS-I changes dramatically on quantum optimization, giving rise to a completely different calculated energy-transfer scenario.[41] Clearly there are some extremely highly distorted heavy-atom chlorophyll structures in the x-ray coordinates, with the original internal strain energy varying by 115 kcal mol^{-1} between cofactors, and these distortions have profound impacts on calculated molecular properties.

VII. OTHER PROBLEM RESIDUES IDENTIFIED BY THE OPTIMIZATION

Outliers in the single point energy distributions can highlight interesting and unusual fragments. These are apparent as residues whose energies are of very low probability according to the normal probability distribution. Inspection of energy outlying species during the optimization revealed that in the majority of cases these were improperly solvated surface residues. These poor solvation environments did not lead to undesired chemical reactions such as proton transfer and so additional solvation was not demanded, but the effects of the poor solvation can still be quantified. This effect is coupled to the anomalously short hydrogen-bond lengths described in Sec. V C. In other cases, outliers can be identified as originating from environments with unusually strong intermolecular interactions. For example, the outlier ASP579-A shown in Fig. 5 features five hydrogen-bond donors, including two arginines, bonding to the side chain carboxylate group; hence it is realistic that this environment could drive the unusual residue distortion. In this Section, we are concerned with those outliers for which no simple explanation is available, however.

A. Anomalies with PHE61-C

Figure 8 shows that PHE61-C is a clear outlier; given a normal energy distribution for the PHE residues, the probability a residue occurring with such a large strain energy is less than 1 in 10^7. This distorted geometry is not a result of the optimization, but rather, as shown in the figure, the optimization actually brings it much closer in energy to the other PHE residues. On visual inspection, optimized PHE61-C is seen to be appreciably distorted with a small angle on the β-carbon of 91° due to steric crowding about its phenyl side chain. We investigated a variety of alternate conformers for this residue but were unable to suggest a more appropriate structure. Given that the x-ray coordinates can only be deduced knowing the component amino-acid sequence as determined by gene sequencing, it is possible that genetic differences exist between the sequenced organism and the one studied by x-ray crystallography. Indeed, mutations of this type are known to have occurred, with additional residues being apparent in the PsaA chain in the x-ray studies. A single nucleotide substitution could have resulted in PHE replacing the sterically more appropriate and similarly hydrophobic residues; VAL, LEU, or ILE. Such possible assignments should be tested against the raw x-ray electron density.

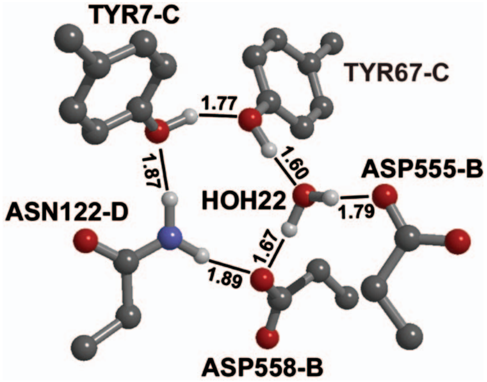

FIG. 9. (Color) Detail of the H-bonding interactions for the possible redox active residues TYR7-C and TYR67-C. H-bond lengths in Å, non-H-bonding hydrogens omitted for clarity.

B. Anomalies with ASP113-B

Another amino acid that may possibly be misassigned is ASP113-B. Our initial protonation-state determination procedure described in Sec. III indicated that this residue should be protonated, a relatively uncommon state for an aspartate residue. Qualitatively, this result is justified as the group appears to be H bonding to a chlorophyll located in the hydrophobic region of the protein through the proton in question, see Table I. While it is not completely unreasonable for this to be the case, the amino-acid ASN would also fulfill this role sterically and electronically without the presumed energy penalty involved in such a protonation process. As the gene codons for ASP and ASN differ by only one nucleotide, a mutation of this nature is feasible.

C. Anomalies with TYR7-C and TYR67-C

Two H-bonded residues, TYR7-C and TYR67-C, proved difficult for the DFT calculations due to convergence problems with the self-consistent-field electronic-structure optimization procedure in GAUSSIAN 03. Initially, these calculations established incorrect electronic structures (quinonoid forms) for these residues that appeared as very distant outliers in the tyrosine intramolecular energy distributions. These errors were manually corrected and do not appear in the final distributions. Figure 9 shows the final optimized structure of these residues and their immediate H-bonding environment; not shown is TYR136-D whose phenolic ring is π stacked with that of TYR67-C, providing another significant intermolecular interaction. During intermediate stages of the DFT optimization, unexpected proton-transfer reactions occurred from HOH22 to ASP558-B, leaving the phenolic species oxidized. Although a chemically sensible structure was eventually obtained, the nature of these problems, along with the proximity of this cluster of residues to the redox active Fe_4S_4 centers, is highly suggestive that these residues may be involved in the electron-transport pathway or related mechanisms.

D. Anomalies with TYR15-E

The environment of TYR15-E is also noteworthy. The initial procedure described in Sec. III for assignment of protonation states deprotonated this residue. This unusual action was predicted based upon the original x-ray heavy-atom coordinates that placed the tyrosine oxygen very close to ARG720-A. As a result of the optimization procedure, through consideration of DFT calculation stability and energy outlier analysis, this residue was reprotonated and returned to its expected state. Geometry optimization stabilized the usual structure of TYR15-E through small changes in the orientations of it and also ARG720-A. However, as TYR15-E is also surrounded by other cations such as LYS551-B and LYS42-E, it also appears to be in an activated chemical environment.

VIII. CONCLUSIONS

While the dramatic advances that have been made in x-ray crystallography have given rise to atomically resolved structures of important biological systems such as the PS-I trimer, the use of these structures in *quantitative* calculations of system properties requires additional optimization. This situation arises as the typical precisions that are obtained for such systems are of the order of 0.3 Å, orders of magnitude larger than that required for accurate electronic structure calculations. Historically, x-ray data have been reoptimized for this purpose in a pragmatic way using QM/MM methodologies and the like that concentrate on optimizing the properties of a small "active sites" *in situ* within the protein. Such approaches are limited in their applicability to systems such as PS-I, however, as for it a large fraction of the entire 150 000 atoms constitute multitudes of internal "active centers." We present a systematic procedure for the quantum optimization of entire protein systems starting from x-ray (or other) structural coordinates. This method embodies in its core not merely low-level quantum procedures such as semiempirical methods but rather state of the art density-functional methods with realistic basis sets.

The optimization scheme is shown to lead to improvements in understanding of the original x-ray coordinates in five areas.

(1) It allows for the generation of a realistic representation of the structure and function of the hydrogen atoms. These atoms are not seen in the x-ray data yet their locations are critical to function.

(2) The method allows for the identification of significant qualitative features that were previously misassigned; we demonstrate this for the functionally critical aspect of the structure of $CL_2$1101, one of the chlorophyll molecules that for the "special pair" in this system, and for the cluster based around what we identify as a metal cation rather than a water molecule in a key structural position on the C_3 axis.

(3) The optimization method allows missing atoms in key locations pertinent to system functionality and chemical integrity to be inserted. We demonstrate this for missing parts of cofactors and for missing solvent molecules only, taking the approach of producing the smallest set of changes to the original x-ray data that is required to establish these features. In an alternate implementation, the methods used could also be extended to produce realistic structures for the remaining incomplete parts of the x-ray structure.

(4) The method leads to a significant improvement in the precision to which the atomic structure is obtained (i.e., the differences between residues of the same type can be ascribed to chemical effects rather than random fluctuations). This effect is dramatic for the cofactors as for these the molecular mechanics methods used in the original x-ray fitting are not as advanced as those used for protein residues. Also, this effect is large for strongly interacting residues such as ions and ligands, and strong hydrogen bonders. Most important, as a result, the structures obtained become suitable for analysis in terms of system function. We will be presenting a reinterpretation of the energy transport networks in PS-I based on the optimized structure; previous calculations[42–45] have all been based on the x-ray heavy-atom coordinates, the errors in which are shown to completely randomize the perceived excited-state energies.

(5) The method allows for the identification of structural features that appear to be improbable. These may arise either from mutations that have made the system that was crystallized slightly different from that which had previously been sequenced, or from misinterpretations of the electron-density data. Unusual chemical features are also identified that are more plausible than these features, and they are identified as possible sites for active device functionality.

The molecular mechanics methods that are used in conjuction with x-ray analysis techniques to add medium-resolution information ensure that the majority of the structure is chemically sensible are computationally efficient and have been instrumental to the development of modern x-ray crystallography. However, computational approaches have now advanced to the stage where methods that offer even higher-resolution information, providing quantitative chemical accuracy in the description of a single realistic atomic structure, have now become feasible. Methods such as our optimization scheme as an augmentation to traditional molecular mechanics x-ray refinement should now be used, especially for the case of cofactors, in x-ray structural determinations, allowing for the generation of chemically realistic structures of arbitrary materials that are simultaneously consistent with the raw experimental scattering data.

IX. SUPPLEMENTARY DATA

Provided in EPAPS (Ref. 46) are the intramolecular energy distributions for all types of residues, the list of manually defined clusters that were optimized at the same time, and the coordinates for the initial hydrogen-only optimized structure as well as the complete final structure in both the ASCII formats of the Protein Data Bank (pdb) and HYPERCHEM (hin) formats.

ACKNOWLEDGMENTS

We thank the Australian Research Council for funding this research, the receipt of a Henry Bertie and Florence Mabel Gritton Research Scholarship, the Australian Centre for Advanced Computing and Communications (AC3) for providing most of the computational resources, and the Australian Partnership for Advanced Computing (APAC) for specialized computing.

[1] R. E. Blankenship, *Molecular Mechanisms of Photosynthesis*, 1st ed. (Blackwell Science, Oxford, 2002).
[2] P. Jordan, P. Fromme, H. T. Witt, O. Klukas, W. Saenger, and N. T. Krauss, Nature (London) **411**, 909 (2001).
[3] W. Westermann, O. Neuschaefer-Rube, E. Moerschel, and W. Wehrmeyer, J. Plant Physiol. **155**, 24 (1999).
[4] G. Tsiotis, W. Haase, A. Engel, and H. Michel, Eur. J. Biochem. **231**, 823 (1995).
[5] V. V. Shubin, V. L. Tsuprun, I. N. Bezsmertnaya, and N. V. Karapetyan, FEBS Lett. **334**, 79 (1993).
[6] J. R. Reimers, M. C. Hutter, J. M. Hughes, and N. S. Hush, Int. J. Quantum Chem. **80**, 1224 (2000).
[7] G. Galli, Phys. Status Solidi B **217**, 231 (2000).
[8] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
[9] J. J. P. Stewart, Int. J. Quantum Chem. **58**, 133 (1996).
[10] M. Wada and M. Sakurai, J. Comput. Chem. **26**, 160 (2005).
[11] A. Crespo, D. A. Scherlis, M. A. Marti, P. Ordejon, A. E. Roitberg, and D. A. Estrin, J. Phys. Chem. B **107**, 13728 (2003).
[12] S. Dapprich, I. Komaromi, K. S. Byun, K. Morokuma, and M. J. Frisch, J. Mol. Struct.: THEOCHEM **462**, 1 (1999).
[13] S. Humbel, S. Sieber, and K. Morokuma, J. Chem. Phys. **105**, 1959 (1996).
[14] T. Matsubara, S. Sieber, and K. Morokuma, Int. J. Quantum Chem. **60**, 1101 (1996).
[15] M. Svensson, S. Humbel, and K. Morokuma, J. Chem. Phys. **105**, 3654 (1996).
[16] J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).
[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[18] HYPERCHEM, Release 5.0, Hypercube Inc., Waterloo, Ontario, 1996.
[19] P. Fromme and H. T. Witt, Biochim. Biophys. Acta **1365**, 175 (1998).
[20] G. Vriend, J. Mol. Graphics **8**, 52 (1990).
[21] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, J. Mol. Biol. **285**, 1735 (1999).
[22] J. E. Nielsen and G. Vriend, Proteins **43**, 403 (2001).
[23] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case, J. Comput. Chem. **7**, 230 (1986).
[24] M. C. Hutter, J. M. Hughes, J. R. Reimers, and N. S. Hush, J. Phys. Chem. B **103**, 4906 (1999).
[25] M. J. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, GAUSSIAN 03, revision B.2, Gaussian Inc., Pittsburgh PA, 2003.
[26] R. A. Torres, T. Lovell, L. Noodleman, and D. A. Case, J. Am. Chem. Soc. **125**, 1923 (2003).
[27] P. Ordejon, D. Sanchez-Portal, A. Garcia, E. Artacho, J. Junquera, and J. M. Soler, RIKEN Rev. **29**, 42 (2000).
[28] J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejon, and D. Sanchez-Portal, J. Phys.: Condens. Matter **14**, 2745 (2002).
[29] K. Ohta, Y. Yoshioka, K. Morokuma, and K. Kitaura, Chem. Phys. Lett. **101**, 12 (1983).
[30] M. Swart and J. G. Snijders, Theor. Chem. Acc. **110**, 34 (2003).
[31] J. J. P. Stewart, J. Comput. Chem. **10**, 221 (1989).
[32] J. J. P. Stewart, J. Comput. Chem. **10**, 209 (1989).
[33] C. J. Casewit, K. S. Colwell, and A. K. Rappé, J. Am. Chem. Soc. **114**, 10035 (1992).
[34] A. K. Rappé, K. S. Colwell, and C. J. Casewit, Inorg. Chem. **32**, 3438 (1993).
[35] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard III, and W. M. Skiff, J. Am. Chem. Soc. **114**, 10024 (1992).
[36] S. Sinnecker, W. Koch, and W. Lubitz, J. Phys. Chem. B **106**, 5281 (2002).
[37] M. Pantelidou, P. R. Chitnis, and J. Breton, Biochemistry **43**, 8380 (2004).
[38] M. Platoa, N. Krauss, P. Fromme, and W. Lubitz, Chem. Phys. **294**, 483 (2003).
[39] N. Krauss, Curr. Opin. Chem. Biol. **7**, 540 (2003).
[40] A. L. Morris, M. W. MacArthur, E. G. Hutchinson, and J. M. Thornton, Proteins **12**, 345 (1992).
[41] M. Dahlbom, J. R. Reimers, and N. S. Hush, J. Phys. Chem. (to be published).
[42] M. K. Sener, D. Lu, T. Ritz, S. Park, P. Fromme, and K. Schulten, J. Phys. Chem. B **106**, 7948 (2002).
[43] M. Sener, S. Park, D. Lu, A. Damjanovic, T. Ritz, P. Fromme, and K. Schulten, J. Chem. Phys. **120**, 11183 (2004).
[44] M. Yang, A. Damjanovic, H. M. Vaswani, and G. R. Fleming, Biophys. J. **85**, 140 (2003).
[45] A. Damjanovic, H. M. Vaswani, P. Fromme, and G. R. Fleming, J. Phys. Chem. B **106**, 10251 (2002).
[46] See EPAPS Document No. E-JCPSA6-123-312548 for the optimized coordinates (both hydrogen only and all atom) in both the Protein Data Bank (pdb) and HYPERCHEM (hin) ASCII formats, as well as the full version of Fig. 8. This document can be reached via a direct link in the online article's HTML reference section or via the EPAPS homepage (http://www.aip.org/pubservs/epaps.hmtl).