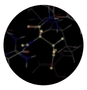
Assignment of Polar States for Protein Amino Acid Residues Using an Interaction Cluster Decomposition Algorithm and its Application to High Resolution Protein Structure Modeling

Matthew P Jacobson
University of California, San Francisco
157 PUBLICATIONS **5,970** CITATIONS



Kai Zhu
Schrodinger
17 PUBLICATIONS **424** CITATIONS

Suwen Zhao
ShanghaiTech University
18 PUBLICATIONS **423** CITATIONS

PROTEINS: Structure, Function, and Bioinformatics 66:824–837 (2007)

Assignment of Polar States for Protein Amino Acid Residues Using an Interaction Cluster Decomposition Algorithm and its Application to High Resolution Protein Structure Modeling

Xin Li,[1] Matthew P. Jacobson,[2] Kai Zhu,[1] Suwen Zhao,[1] and Richard A. Friesner[1*]
[1]*Department of Chemistry, Columbia University, New York, NY*
[2]*Department of Pharmaceutical Chemistry, University of California, San Francisco, CA*

ABSTRACT **We have developed a new method (Independent Cluster Decomposition Algorithm, ICDA) for creating all-atom models of proteins given the heavy-atom coordinates, provided by X-ray crystallography, and the pH. In our method the ionization states of titratable residues, the crystallographic mis-assignment of amide orientations in Asn/Gln, and the orientations of OH/SH groups are addressed under the unified framework of polar states assignment. To address the large number of combinatorial possibilities for the polar hydrogen states of the protein, we have devised a novel algorithm to decompose the system into independent interacting clusters, based on the observation of the crucial interdependence between the short range hydrogen bonding network and polar residue states, thus significantly reducing the computational complexity of the problem and making our algorithm tractable using relatively modest computational resources. We utilize an all atom protein force field (OPLS) and a Generalized Born continuum solvation model, in contrast to the various empirical force fields adopted in most previous studies. We have compared our prediction results with a few well-documented methods in the literature (WHATIF, REDUCE). In addition, as a preliminary attempt to couple our polar state assignment method with real structure predictions, we further validate our method using single side chain prediction, which has been demonstrated to be an effective way of validating structure prediction methods without incurring sampling problems. Comparisons of single side chain prediction results after the application of our polar state prediction method with previous results with default polar state assignments indicate a significant improvement in the single side chain predictions for polar residues. Proteins 2007;66:824–837. © 2006 Wiley-Liss, Inc.**

Key words: protonation state; mis-assignment; hydrogen bond; all-atom force field; side chain prediction; generalized born model

INTRODUCTION

Macromolecular structures determined by X-ray crystallography generally cannot specify the positions of hydrogen atoms, other than a small handful of structures solved at unusually high resolution. This is a significant limitation of crystal structures because ionization states of titratable residues and the positions of polar hydrogens (i.e, hydrogens on the —OH group of carboxylic acid/C-terminal end groups and NH_3^+ groups on Lysine side chains/N-terminal end groups) are often important for understanding the structural properties, dynamic behaviors, and ligand binding of macromolecules. Another related problem is that the amide groups in Gln/Asn and the imidazole ring in His are sometimes oriented incorrectly because of the ambiguity caused by the similar electron density of oxygen and nitrogen atoms. The His side chain is particularly complicated because it has three possible protonation states as well as ambiguity in the orientation of the ring, resulting in a total of six alternative states. One of the common characteristics possessed by both protonation states of titratable groups and Asn/Gln/His side chain orientations is that they both strongly depend on, and in turn significantly impact, the hydrogen bonding patterns in the vicinities of these residues.

Several previous studies have addressed the problem of protonation state prediction, mainly through the calculation of effective pKa's of titratable groups.[1–8] For example, Bashford et al.[2] designed a Monte Carlo stochastic algorithm for effective pKa calculations and applied their method to the calculation the titration curve of the titratable residues in several lysozymes with different unit cell shapes. Baptista et al.[1] have designed an extended molecular dynamics algorithm by explicitly taking pKa into consideration as a dynamic variable in the evolution of the whole dynamic system. A recent extension of this work by Lee et al.[4] generalized the idea of λ-dynamics

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: NIH; Grant number: GM-52018; Grant sponsor: HHMI; Grant number: 5300246; Grant sponsor: Sandler Program in Basic Sciences.

*Correspondence to: Richard A. Friesner, Department of Chemistry, Columbia University, 3000 Broadway, New York, NY 10027.
E-mail: rich@chem.columbia.edu

Received 8 February 2006; Revised 31 May 2006; Accepted 13 June 2006

Published online 11 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21125

into a continuous titration variable and used an implicit model as a macroscopic description of solvent.

With respect to the Asn/Gln/His side chain χ_2 misassignment problem, Word et al.[9] have performed a very detailed study based on an analysis of molecular contacts including all explicit hydrogen atoms, and incorporated their method into the REDUCE software package. Their method, although extremely fast, addresses only side chain orientation, and uses a very simplified scoring function only involving hydrogen bond and overlap volume.

The approach most closely related to that in the present paper is the method developed by Hooft et al.[10] which explicitly treats ionization states and crystallographic misassignments in a unified framework. Their method has been incorporated into the well-known WHATIF molecular modeling program.[11] The WHATIF method uses a coarse-grained empirical hydrogen bond force field derived from small molecule crystal structures, in contrast to the physics-based, all-atom energy function employed here. A second key difference is that the WHATIF method uses knowledge of explicit waters identified in the crystal structures, which makes it appropriate for structure verification but not for other modeling tasks, where this information would be unavailable. We have instead used an implicit solvation model (see below for details) to treat water, which allows us to employ our method in protein structure modeling tasks such as homology modeling. In addition, explicit water generally account for only a portion of the water in a protein crystal, and the implicit solvent model accounts for the effects of dielectric screening from bulk water.

In this work, we have developed a general algorithm for assigning positions for all polar hydrogens in a protein, given the pH. The protonation states of all titratable residues are considered: Arg, Asp, Cys, Glu, His, Lys, and Tyr (Cys residues forming disulfide bonds are not considered, although this can be incorporated into the methodology fairly easily). The algorithm performs placement of polar hydrogens for the protonated residues (including Ser, Thr, Tyr, and the NH_3^+ group of Lys), as well as the ambiguous χ_2 angles for Asn/Gln/His. The union of all of the side chains considered will be referred to as "polar residues" and the union of all conformations and protonation states considered will be referred to as "polar states." The energy function is the OPLS all atom force field[12–14] with SGB/NP implicit solvation model,[15,16] which has proved to be a very effective model for all atom macromolecular modeling, as demonstrated in our previous works.[17–20]

The problem of optimizing the positions of polar hydrogens (including protonation states, orientations of OH/SH groups, and side chain orientations of Asn/Gln/His) is a very hard computational problem in principle. This is due to the exponential scaling of the state space $O(a^N)$, where N is the number of polar residues, and where the number of polar states a ≥ 2. Because of the current computational bottleneck in calculating the energy functions, especially nonbonded pairwise interactions, which scale as $O(N^2)$, it is arguably a very challenging task to incorporate the effect of polar states into the computational modeling of proteins and other macromolecules explicitly. Our method partitions all polar residue in a given protein (including all symmetry copies) into disjoint "independent" clusters. For this reason, we refer to this method as the Independent Cluster Decomposition Algorithm (ICDA). The partitioning scheme is based on the local hydrogen-bonding network, and reduces the computational complexity from $O(a^N)$ to $O(\frac{N}{t}a^t)$, where t is the average size of the cluster (which is [≪]N for all test cases we have examined to date).

A critical issue in polar state assignment is the validation of assignments made by the algorithm under study; the "right" answer cannot simply be read off from raw X-ray crystallographic data, unless one is considering ultra high resolution structures in which hydrogen positions can be observed unambiguously. We argue later that single side chain prediction (SSP)-prediction of each side chain, one at a time, keeping the remainder of the protein fixed-provides an appropriate metric for evaluating the correctness of polar state assignments. If the assignment is correct, and the energy model is accurate, prediction of every side chain in the crystal environment should reproduce the native structure, in particular the hydrogen bond patterns of the side chain as seen in the crystal structure. Although some side chain predictions will in fact exhibit errors because of problems with the energy function, one can calculate the fraction of successful SSP predictions for each residue type, and associate a higher success rate with a superior assignment methodology. This argument is elaborated in some detail in what follows. Finally, we discuss the limitation of the current method and provide perspectives for future improvement.

Although the paper is primarily focused on presenting our new methodology and the results generated by it, we have also benchmarked performance against existing methods so as to calibrate whether or not any improvement has been achieved. We compare our predicted results with those obtained in previous work, including the hydrogen placement method of Hooft et al.[10] and the Asn/Gln/His flipping detection method developed by Word et al.[9] as implemented in REDUCE. In addition to explicit comparisons for specific residues and proteins, we evaluate SSP accuracy for all three methods and present statistics summarizing effectiveness as assessed by this metric.

METHODS

Overview

Hydrogen bond networks play a vitally important role in determining the relative stability of alternative polar states. In many cases, the local hydrogen bond network can be used to unambiguously assign protonation states or χ_2 orientations of side chains, even without a detailed calculation. For example, in cases of two carboxylic acid oxygen atoms being very close in space (distance: <2.8Å), it is almost always the case that one of them must be protonated so as to relax the strong electrostatic repulsion that otherwise would have been incurred. Also, for this particular problem long-range interactions do not appear to play as important a role as the local hydrogen bond network. We can therefore, as a first order approximation (which turns out to work very well in practice as will be shown below), assume that the polar state of a polar residue depends only upon its neighboring polar residues, based on a certain distance cutoff, and is independent of all the other polar residue states (since the structure of the whole protein, except for the polar hydrogens, are held fixed, the polar state of a particular residue is completely determined once the polar states of all the other polar residues are fixed). We may therefore optimize the whole set of polar states by optimizing each cluster separately and finally combining the results to yield the final prediction.

Such simple distance based cluster decomposition of polar state space is in no way meant to completely replace all atom energy calculations, as has been the case for some alternative approaches,[9,10] but merely serves as an efficient pre-processing step to reduce the potentially huge sampling of states, which is hardly tractable for current computers. If we were to evaluate all $O(a^N)$ states in a "brute force" approach (where $a \geq 2$), we would have to evaluate the energy of the whole system many times. However, by partitioning the set of polar residues into K clusters each of which has roughly t members, and independently optimizing the polar states of each cluster, we only need to perform $O(\frac{N}{t}a^t)$ energy calculations. Typically we have $t[\ll]N$ (in most cases the maximum cluster size t is less than 4 using the default cutoff, see Fig. 1). This level of effort renders the problem tractable using realistic energy and solvation models, including geometry optimization of the hydrogen bonded network, a feature that has been missing from many prior methods for pKa determination employing, for example solutions to the Poisson-Boltzmann equation to treat solvation effects. Our experience has been that it is difficult, if not impossible, in many cases to compare the energies of alternative hydrogen bonding patterns without optimization of the total energy, including solvation.

In addition, by eliminating clearly incompatible partial polar state assignments early in the calculation, a further reduction in sampling space can be attained. In our method we use a tree based search approach so that once we detect any incompatible "partial assignment" we immediately prune an entire branch of the search tree and avoid any further energy evaluations. This can improve efficiency significantly in optimizing large clusters.

The final assignment of the whole protein is obtained by combining the best assignment of each cluster. A full energy minimization is then applied to all the polar hydrogens to yield the final structure with the predicted polar states assigned to the set of polar residues.

Assessing the accuracy of ICDA and related methods is far from trivial. In the present paper, we employ single side chain prediction (SSP) as a measure of the accuracy of polar state assignment by various algorithms (both ours and those of other groups). The definition of a single side chain prediction is straightforward: all of the protein except the side chain of the target residue is fixed, complete phase space sampling of that residue is carried out, and the resulting lowest predicted free energy conformation is compared with the experimental structure. A crucial point is that, for comparison with experimental crystallographic data to be meaningful, SSP *must* be carried out in the crystal environment if there is any sort of intermolecular interaction in the native structure.

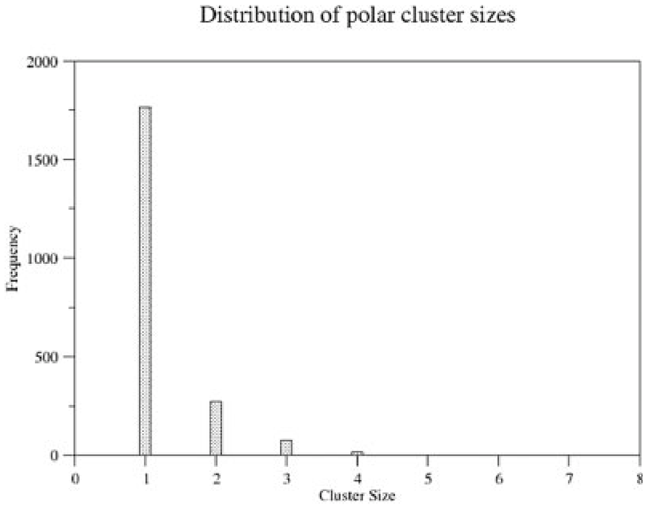


Fig. 1. Distribution of polar cluster sizes for the 30 test protein structures. The maximum cluster size is 8, with only one occurrence in our test. The majority of clusters have sizes varying from 1 to 3.

A number of variables can be examined to assess the accuracy of the prediction, including χ_1 and χ_2 accuracy, RMSD, and hydrogen bonding pattern. In the present paper we rely principally on RMSD (more accurate than χ_1/χ_2, less labor intensive than examination and categorization of hydrogen bonding patterns) but similar results would be obtained with any of the relevant measures.

Of course, some errors in SSP are due to problems with the force field and the solvation model, and these will not be improved by superior polar state assignment. However, it is a reasonable hypothesis that incorrect polar state assignment will in many if not most cases lead to erroneous single side chain prediction. Obvious cases are when mismatches of donors and acceptors result from such incorrect assignments; the example discussed in the introduction, of two carbonyl oxygens being placed 2.8 Å from each other with no intervening hydrogen, is a canonical example. In this case, one would expect that if SSP is carried out for either of the carboxylate-bearing side chains that compose the structure, the native conformation, forcing the two negatively charged atoms to approach each other so closely, would not be formed, leading to a substantial deviation in RMSD. Thus, the expectation is that effective polar state assignment will significantly reduce the number of errors in SSP results, when compared, for example with SSP calculations starting with "default" assignments—e.g., "normal" protonation states for titratable residues and His/Asn/Gln conformational assignments from the PDB file. Similarly, two methods for polar state assignment can be compared by carrying out SSP for structures prepared with each method, and determining which method yields a smaller fraction of errors.

Ultimately, the polar state assignment process is connected to the energy model, and SSP data provides a self-consistent approach to assessing the entire machinery. SSP involves a small enough number of degrees of freedom that sampling errors can be entirely eliminated, i.e. it is possible to search the entire phase space rigorously with a relatively modest expenditure of computation time. A truly accurate and reliable energy model/polar state assignment methodology will yield SSP results with few or no errors when compared with experimental data for a wide range of high resolution test cases in the PDB. At present we (and, to our knowledge, other groups working on this problem) are far from this ideal state of affairs; nevertheless, it is possible to assess improvement in SSP prediction as a result of addressing a particular aspect of the problem, such as polar state assignment, and that is what we have chosen to do in what follows.

Formal Specification of the Algorithm

Our aim is to assign a unique set of polar states to the N polar residues in a protein, such that the overall assignment is optimal energetically, i.e., the set of optimal assignment $S_{\text{opt}} = \{s_1, s_2, \ldots, s_N\}$ satisfies the following criteria:

(a) The assignment is "pairwise compatible," i.e., for each pair of predicted states: $s_i, s_j, i \neq j.$, we have that

$$\text{compatible}(s_i, s_j) = \text{true}$$

We call any assignment S satisfying condition (a) a compatible assignment.

(b) S_{opt} has the lowest energy over the space of all compatible assignments, i.e.,

$$E(S_{\text{opt}}) = \min_{S \text{ is compatible}} E(S)$$

where E includes both the force field energies (including solvation contribution) as well as a pH dependent term (with corrections), which will be explained below.

Our algorithm proceeds as follows:

Step 1: Identify all polar residues for which "polar states" have to be assigned

As described in the introduction, we include the ionizable residues (His, Glu, Asp, Lys, Arg, Cys and Tyr), the side chains with ambiguous χ_2 angles (His, Asn, Gln), and side chains with OH/SH groups (Tyr, Cys, Ser, Thr). Here we do not address the assignment of end groups, because they are generally disordered and rarely of interest.

Step 2: Partition the set of all polar residues into independent clusters

This step raises the issue of how to define the "independent clusters." Here, on the basis of extensive observation of the data from the native structures, we choose a simple distance criterion to partition the polar residues into clusters. Given a pair of polar residues, if there is at least one pair of nonhydrogen side chain atoms (interaction between backbone and side chain polar atoms will be considered in subsequent steps) from the two residues respectively lies within a specified distance cutoff D_c, we will call such a pair of residues a pair of neighbors. A cluster is thus defined as a maximal collection of such neighboring polar residues, i.e., it cannot be expanded by adding another polar residues into the cluster. Note that in considering the neighbor relations of residue i and j, all the symmetry copies of i and j in the crystal are also explicitly taken into consideration, to account for the crystal packing effects, which plays a non-negligible role in determining the structural properties especially for surface polar residues[20]. The "self-interactions," i.e., interactions between the two symmetry copies of the same residue (e.g., a carboxylic acid dimer in "resonance") are considered separately during the compatibility checking step.

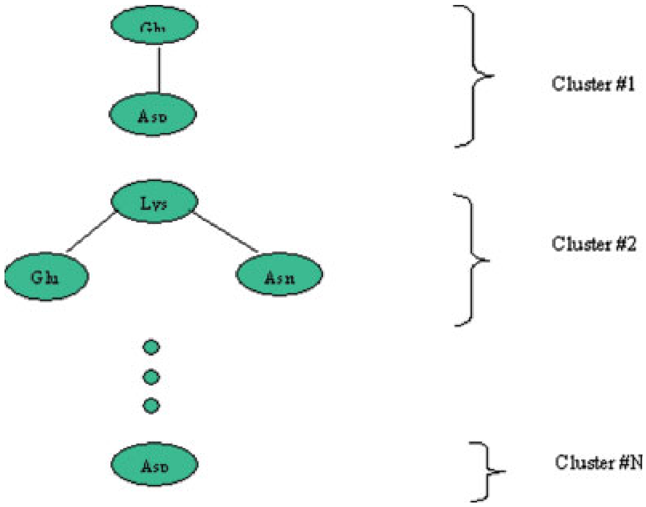


Fig. 2. Illustration of the idea of independent cluster decomposition. All polar residues are partitioned into disjoint clusters according to the hydrogen bonding equivalence relations, and each cluster is optimized independently, i.e., the polar state of any particular polar residue only depends on those polar residues belonging to the same cluster as it. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The partition of polar residues can be implemented using a standard graph theory algorithm. More specifically, we construct an undirected graph G = (V, E) for the given protein molecule such that the set of vertices of the graph represent the polar residues. An edge is assigned between vertex i and j if and only if residue i and j are neighbors, as defined earlier. V and E are the set of vertices and edges respectively. The partitions will then be found with a standard Depth First Search (DFS) traversal algorithm.[21] The independent cluster decomposition scheme is depicted in Figure 2.

The cutoff distance D_c is an adjustable parameter in our implementation of the algorithm. It determines the size of a cluster. Large clusters will definitely result in more extensive search in subsequent sampling steps, thereby demanding more computations, while too many small clusters would be unlikely to recover the hydrogen bond interactions as most of the interacting clusters would be treated as independent. Here we set $D_c = 3.1$ Å, which is a more or less standard hydrogen bond distance cutoff. In fact, this cutoff works quite well in practice in terms of the balance between sensitivity of inter-residue interactions and computational efficiency, as demonstrated in numerous tests that we have performed (data not shown). In future work the use of a single cutoff value can likely be improved, by defining the cutoff to be specific to each pair of partners in the hydrogen bonding interaction.

Step 3: Discretization of polar hydrogen positions

Each of the polar residues is then assigned a fixed set of states, on the basis of the alternative polar forms it can exist in. For most titratable residues as well as Asn/Gln there are just 2 states, corresponding to the protonated/deprotonated forms, or the two side chain flipping states. For His there are 6 states (3 tautomeric forms and 2 conformational flips of the ring). However, for protonated states with rotatable polar hydrogens (i.e., hydrogens on the hydroxyl group of Asp, Glu, and Tyr; the SH group of Cys; and the Lys amine group), we must also specify the angular position of the polar hydrogens in order to uniquely determine the "polar state" of the residue. Here we expand the protonated states via discretization of the polar hydrogen positions, such that the original protonated state will be split into a set of "child states," each corresponding to a polar hydrogen position. The way we discretize the polar hydrogen positions here is to assign a set of hydrogen "rotamer" states such that each state corresponds to an alternative hydrogen bond based on proximity of neighboring heavy atoms. Finally, there is a state corresponding to the "solvated state," in which no hydrogen bond is made to any other protein atoms and contacts only water. Each of these initial child states is then split into two states by perturbing the rotation angle around the bond axis by $\pm 15^\circ$, to allow for the possibility of minor deviation from the perfect hydrogen bond directions. Thus, for example, for a polar hydrogen which can potentially make 4 alternative hydrogen bonds with its 4 different neighbor heavy atoms, we will end up expanding the protonated state into 9 child states; adding in the deprotonated state, there will be 10 polar states altogether to sample.

Step 4: Check the compatibility between pairs of polar states for each cluster

For each cluster, the compatibility table for a pair (i, j) of interacting residues is N times M if there are N possible states of residue i and M possible states of residue j. Each entry is either a 1 or a 0. A 1 indicates that the two states are compatible; a 0 indicates that they are incompatible. In the computational representation, each entry of the compatibility table only needs to be represented by a single bit.

The compatibility test is based on whether there is any "violation of physical sense." Basically, for a given state, if two atoms carrying like-charges are in close contact (e.g., an O—O or N—N pair within a distance less than 2.8 Å), without a hydrogen atom lying in between, which would indicate a clear sign of incompatibility.

Step 5: Independent prediction of polar states for each independent cluster

This turns out to be the most computationally intensive part of the algorithm as a large number of energy evaluations are needed, which is far from a trivial calculation since we are using an all atom force field plus implicit solvation. In addition to the decomposition into independent clusters as explained earlier, which makes it only necessary to enumerate the energy states for each independent cluster separately, we further reduce the computation time via an efficient tree pruning approach. Basically, for a given cluster, the state of each member residue is evaluated in a tree-based process. Each polar residue would then correspond to a node of the tree, with each alternative state assignment as a separate branch. The tree grows in a top-down manner. Once a leaf is reached, a new set of polar state assignments is obtained by following the unique path from the root of the tree (which is the first residue to be evaluated) down to this leaf node, and the energy of the system is evaluated with the given assignment. The pruning occurs if, during any stage of the assignment, an incompatibility of the assignment of the current residue with a previous assigned residue is detected. At this point, the whole branch from that node down to the leaf is pruned, thereby eliminating the need to evaluate the energy of all state combinations containing that incompatible assignment. Finally, states surviving the pruning process are evaluated via local minimization of the total energy model (force field plus solvation), and the states are then ranked according to energy. The minimization minimally perturbs the hydrogen atom positions but can significantly change the energy. Note that, although the polar side chains are clustered according to their local interactions, the energy evaluations (including the Coulomb electrostatic term and the solvation free energies) are performed over the entire protein, without any distance cutoff.

At the end of this step, we end up with a list of all (polar) residue state assignments, ranked by energy from lowest to highest. We then take the one with the lowest energy as the optimal assignment. Alternatively, one could also retain more than one top ranked structures for subsequent calculations, or use a clustering approach for selecting representative low energy assignments. In the present paper, however, we only consider the lowest energy assignment for each independent cluster. In practice, the energy differences among the states can range widely. For example, for the first 6 Glu side chains in protein 1wer, the energy differences between the best and second best state are 26.00, 17.27, 0.38, 0.22, 2.12, and 13.94 kcal/mol.

Step 6: Combine the best assignment of each cluster and get the overall optimal assignment of the whole set of polar residues

After getting the optimal assignment of each cluster as in the previous step, we finally obtain the optimal assignment for all polar residues in the protein molecule. We then assign each residue its optimal polar states, and relax the whole structure by a complete energy minimization of all polar hydrogens.

Energy Function

All energy evaluations are performed using an all-atom energy model, based upon the OPLS-AA force field and the Surface Generalized Born implicit solvent model. The force field and solvation model have been extensively discussed and tested in previous publications (including their effectiveness in loop and side chain prediction),[17–20] and we shall not repeat this discussion here. Compared with many previous works addressing polar state assignment, our force field model is highly detailed. For a particular atom type in a given titratable residue type, the force field parameters such as the atomic radii and partial charges are customized for each alternative ionization state, thereby taking into explicit consideration the partial charges/vdW radii relaxation during the proton association/dissociation process. Detailed parameter optimization of this type is often not performed in empirical methods, which is likely to lead to a lower degree of accuracy and robustness.

Comparing Free Energies of Different Protonation States

An energy correction term, assigned to each protonation state of each titratable residue, is required in order to appropriately compare the total energies obtained from the various protonation states. One approach to computing a term of this type is to use a thermodynamic cycle, in which one incorporates the gas phase deprotonation energy of the species (which can be computed via quantum chemistry, for example) in question and the solvation free energy of the proton. However, neither of these quantities is known with high precision, so we have adopted an alternative approach, which employs an experimental reference state with a known pKa. Specifically, we compute free energy difference between the protonated/unprotonated forms for the blocked amino acid monomer (acetylating the N-terminus and amidating the C-terminus); similar approaches have been adopted by others, e.g. Ref. 4

The correction term is simply taken as the force field energy difference between the protonated state and unprotonated state of the minimized capped amino acid monomer (including SGB solvation term). Such a correction term is calculated once and all the values are stored. In all subsequent calculations, the stored correction terms will be added to the energy difference between the nonstandard and standard ionization forms of each titratable residue. There is a distinct correction term for each titratable residue (with the exception of His, which has two correction terms, corresponding to the two alternative protonation site N_δ and N_ε). For flipped states of Asn/Gln/His, there are no correction terms. Furthermore, we have developed a set of empirical adjusting terms for correction terms, motivated by the consideration that residues assume different conformations in real proteins than the conformations of the corresponding monomers we use to find the correction terms. These adjusting terms are optional, however, including them improves our results significantly.

TABLE I. Model pKa Values and Free Energy Correction Terms for all Titratable Residues

Residue	Model pKa	Free energy correction term (kcal/mol)
Arg	12.0	37.48
Asp	3.9	–23.27
Cys	8.5	60.74
Glu	4.3	–35.65
His	6.4	28.42 (HIP*); 6.58 (HIE*)
Lys	11.1	–23.95
Tyr	10.0	86.33

*The default protonation state for His is HID, which is protonated on N_δ only; HIP signifies protonation on both N_δ and N_ε HIE signifies protonation on N_ε only.

Therefore, if we take the correction term as well as the usual pH dependent term into consideration, the energy difference between the two ionization forms for a particular titratable residue, assuming the polar states and conformations of other parts of the protein being identical, would be

$$\Delta U = U^P - U^U = U^P_{FF} - U^U_{FF} + 2.303kT(\text{pH} - \text{pKa}^\text{m}) \\ + U^\text{corr},$$

where pKa$^\text{m}$ is the pKa for the model compound, which is readily available from experiment; U^P_{FF} and U^U_{FF} are the force field energies for the protonated and unprotonated form respectively (see previous section). The model pKa's and correction terms used are listed in Table I. These values are dominated by differences in the solvation free energies between the protonation states, as computed using the SGB implicit solvent model. However, a second major contributor is the Coulombic electrostatic interactions with the side chain, especially "1–4" interactions (i.e., atoms separated by 3 bonds) involving the titratable proton.

Crystal Packing

In previous work,[20] we have found that crystal packing forces can affect structural details of proteins, especially the conformations of polar side chains on the surfaces of proteins. To remove any uncertainty about effects of neglecting crystal packing, and to provide a fair comparison with experimental crystal structures, we perform all predictions in the crystal environment. That is, crystal unit cells are explicitly reconstructed using the dimensions and space group reported in the Protein Data Bank files. We do not attempt to employ explicit lattice summation techniques (e.g., Ewald summation), but instead define the simulation system to consist of one asymmetric unit (which may contain more than one protein chain) and all atoms from other, surrounding asymmetric units that are within 30 Å. Every copy of the asymmetric unit is identical at every stage of the calculation; that is, space group symmetry is rigorously enforced.

COMPUTATIONAL IMPLEMENTATION

The method is implemented in Fortran 90 and all computational tests are performed on our Linux cluster with 32 nodes. All nodes are 1.4 GHz Pentium III processors. The typical running time for a protein with 100 polar residues is about 20 min. Recently, a parallel version with MPI interface has also been developed which results in a 2–3 times speed-up of the CPU time.

RESULTS AND DISCUSSION

Overview

For this work, we performed two sets of computational experiments to validate of our polar state assignment algorithm (ICDA). We first present and discuss a few residue-by-residue comparisons with WHATIF and REDUCE, using data from the literature as well as data that we have generated by running both programs on test cases selected from the PDB. As a second part of the test, we have applied WHATIF, REDUCE, and ICDA to an extensive set of high-resolution proteins that have been used in the single side chain prediction test, and demonstrated that the correctly predicted polar state assignments emerging from ICDA result in statistically meaningful improvement of the single side chain prediction results for polar residues, as compared to a simple default assignment approach and also to results obtained from structures assigned using either WHATIF or REDUCE. We also present some preliminary results using ICDA to improve loop structure prediction. Here we just aim to offer some anecdotal evidence of effectiveness of ICDA, instead of providing a comprehensive investigation. This task will be performed in a subsequent publication.

Comparison With Previous Work

We first compare ICDA with the hydrogen placement algorithm of Hooft et al.[10] as implemented in the WHATIF software package, using test cases identified by the authors of that package. In Ref. 10 prediction results are presented and analyzed for the protonation states of the 2 pairs of carboxylic acids in the protein penicillopepsion (1APT): Glu16/Asp115 and Asp33/Asp213, as well as a single carboxylic acid Glu45. For the first 4 carboxylic acids, none of them were predicted to be protonated, although the author strongly believed that at least one in each pair should. The author attributed this failure to the inappropriate value of penalty terms they assigned for protonation, or the inappropriate treatment of the special hydrogen bond pattern. Using ICDA, we found that both Glu16 and Asp115 are predicted to be protonated: Glu16 on OE2 and Asp115 on OD1. The protonation of Glu16 on OE2 yields a better hydrogen bond pattern in correspondence to the spatial vicinity of the two oxygens Glu16:OE2 and Asp115:OD1, which are separated by a distance of only 2.94 Å. The OE1 atom on Asp115 is protonated because of the presence of a side chain-backbone hydrogen bond pair, the backbone oxygen being one of its own crystal copies. For the second pair, Asp213 is predicted to protonate on OD1, while Asp33 is left unprotonated, resulting in a good hydrogen bond arrangement between the spatially close oxygens Asp33:OD2 and Asp213:OD1 (distance = 2.92 Å). For Glu45, both our method and that of Hooft et al. predicted the protonation of OE2, due to the presence of its close hydrogen bond neighbor Asn84:OD1 (distance = 2.53 Å).

TABLE II. General Information for the Proteins in the Test Set

PDB ID	pH	Number of residues
1a2y	6.5	352
1a3c	5.1	166
1akz	7.9	223
1awd	8.0	94
1awq	8.4	170
1b2p	4.7	238
1bkr	6.6	108
1brt	8.5	277
1btk	8.5	329
1c52	8.1	131
1cvl	6.4	316
1dhn	6.5	121
1edg	6.0	380
1f94	8.5	63
1ig5	6.4	75
1ixh	4.5	321
1jse	4.2	129
1kpf	6.5	111
1nox	6.0	200
1qto	5.7	122
1rcd	5.5	171
1rhs	7.6	293
1u9a	7.5	160
1wer	6.5	324
2a0b	4.1	118
2fcb	5.3	173
2ilk	6.5	155
2pth	7.5	193
3lzt	4.6	129
3vub	4.5	101

For each protein, PDB ID, crystallization pH, as well as the total number of residues are indicated. In the case of proteins consisting of multiple chains, the number of residues includes the residues in all chains.

To examine assignments for Asn/Gln/His residues, we select 4 protein structures and compare our predicted states for all Asn/Gln/His with those obtained using WHATIF and REDUCE. The 4 proteins being tested are: 1b2p, 1c44, 1c52, and 1awq. The comparison results are presented in Supplementary Materials. It can be seen that in many (although not all) cases, the ICDA, WHATIF and REDUCE assignments are in agreement. It should be noted that there is no attempt being made to address the issue of polar states of other residues (e.g., carboxylic acids) in REDUCE, so the correlation effect among different polar residues tend to be under-estimated. The question of which results are correct in cases where there is disagreement is,

TABLE III. Comparison of Single Side Chain Prediction Quality Using Default Assignment and ICDA

Residue type	Correct–Correct (%)	Incorrect–Correct (%)	Correct–Incorrect (%)	Incorrect–Incorrect (%)
Asn	62.46	21.50	5.12	10.92
Gln	46.90	17.70	3.98	31.42
His	41.06	37.75	5.96	15.23
Asp	78.22	5.28	1.98	14.52
Glu	50.92	11.66	3.37	34.05
Ser	74.92	3.93	4.83	16.31
Thr	86.29	1.25	0.93	11.53
Lys	45.32	5.44	2.72	46.53
Arg	46.92	2.69	1.92	48.46
Cys	88.68	0.00	5.66	5.66
Tyr	96.46	0.51	1.01	2.02
Overall	64.05	9.20	3.15	23.59

For each polar residue type, four percentages (with respect to the total number of that particular residue in the test set) are reported: (1) Percentage of side chains that are correctly predicted for both the default and ICDA assignment (Correct–Correct); (2) Percentage of side chains that are correctly predicted for ICDA assignment while incorrectly predicted for the default assignment, or the percentage of improvement (Incorrect–Correct); (3) Percentage of side chains that are incorrectly predicted for ICDA assignment while correctly predicted for the default assignment, or the percentage of degradation (Correct-Incorrect); (4) Percentage of side chains that are incorrectly predicted for both the default and ICDA assignment (Incorrect–Incorrect). The advantage of the ICDA assignment over the default assignment can be seen from the contrast between the high percentage of improvement and the low percentage of degradation.

as noted above, a highly nontrivial one; the single side chain prediction statistics, discussed in the next subsection, represent our approach to answering this question.

Single Side Chain Predictions With Explicit Polar State Assignment

For the single side chain prediction test, we have selected a test suite composed of 30 globular protein crystal structures from the PDB database. These proteins are selected such that all of them have resolution <2 Å and do not have any serious heavy atom steric clashes. These structures vary widely in their size and crystallization pH values. Table II lists the PDB IDs and sizes (number of residues) along with their respective crystallization pH values as reported in the PDB file header.

The 30-protein test suite, in addition to serving as a test set for single side chain prediction, is also meant to be a comprehensive suite for studying our ICDA polar state assignment algorithm, from which a lot of characteristics of the algorithms could be extracted. Figure 1 gives the distribution of the independent cluster sizes (see Methods section). It can be clearly seen that the majority of clusters have less than 3 polar residue members, and in fact sizes >5 are rarely seen, demonstrating that most residues are either fully "independent" or connected in relatively small interacting groups.

As explained in Methods, the single side chain prediction is used here as a benchmark of our polar state prediction method. We restrict our attention to the set of polar residues, as defined to be the limited subset of the 11 residue types (titratable plus Asn/Gln) described previously. Based on experience, we found that the RMSD cutoff of 1.5 Å is a suitable threshold in distinguishing between the success/failure of single side chain prediction, and serves as a more accurate measure of side chain prediction correctness than commonly used criteria based on accuracy of the χ_1 or χ_{1+2} angles,[22,23] in consideration of the fact that many longer side chains have more than 2 side chain χ angles, such as Lys and Arg. In this work, both the RMSD and χ_1/χ_{1+2} criteria have been used. Basically, RMSD <1Å is usually reasonable, while an RMSD >2Å is normally thought of as problematic. Table III and IV summarize the effect of polar state assignment on the single side chain prediction accuracies.

One simple analysis is to determine how many residues' SSP results were improved, left invariant, or made worse by the ICDA assignments, as compared to the default assignment approach discussed above. Statistics addressing this issue are presented in Table III. We define four possible outcomes of carrying out default and ICDA predictions: correct (RMSDs less than 1.5 Å for both predictions); default correct, and ICDA incorrect; default incorrect, and ICDA correct; and both predictions incorrect. As our conclusions are intended to be qualitative, we do not attempt a more quantitative assessment of the alterations in structure induced by ICDA calculations. The first and fourth of these categories indicate that there has been little change in SSP accuracy as a result of ICDA assignment; the fourth category presumably predominantly represents errors due to intrinsic problems in the force field or solvation model. The second category is most likely reflective of an error in the ICDA model, whereas the third category indicates a successful revision of the structure proximate to the side chain in question. For all residues, successes outnumber failures, typically by a substantial margin. However, the significant number of cases falling into categories 2 and 4 imply that there is more work to do in improving both the assignment and energy models. Nevertheless, this data demonstrates that a reasonable start has been made.

In terms of individual residue types, we found that our algorithm is most effective in fixing the flipping of Asn/Gln amides and the protonation states of certain carboxylic acids and histidines. Achieving high absolute accuracy

TABLE IV. Accuracies of Single Side Chain Prediction for all Test Proteins, Before and After Polar States Assignment Using ICDA

		Without ICDA assignment				With ICDA assignment			
Target	Total polar residue #	Correct RMSD #	Correct RMSD %	Correct χ_1 %	Correct $\chi_{1,2}$ %	Correct RMSD #	Correct RMSD %	Correct χ_1 %	Correct $\chi_{1,2}$ %
1a2y	176	120	68.2	80.7	71.6	135	76.7	83.5	73.9
1a3c	71	44	62.0	76.1	60.6	48	67.6	84.5	69.0
1akz	113	77	68.1	82.3	68.1	82	72.6	83.2	76.1
1awd	50	39	78.0	84.0	76.0	41	82.0	88.0	80.0
1awq	81	57	70.4	93.8	79.0	61	75.3	96.3	81.5
1b2p	110	68	61.8	82.7	66.4	78	70.9	81.8	69.1
1bkr	61	40	65.6	77.0	65.6	43	70.5	78.7	68.9
1brt	130	103	79.2	90.0	78.5	107	82.3	91.5	80.8
1btk	176	113	64.2	80.1	66.5	117	66.5	80.1	66.5
1c52	55	31	56.4	78.2	61.8	39	70.9	78.2	69.1
1cvl	151	113	74.8	86.8	78.8	124	82.1	89.4	84.1
1dhn	64	44	68.8	82.8	71.9	47	73.4	82.8	73.4
1edg	200	158	79.0	88.5	78.0	163	81.5	88.0	79.5
1f94	32	23	71.9	81.3	71.9	25	78.1	87.5	78.1
1ig5	41	21	51.2	78.0	58.5	24	58.5	80.5	63.4
1ixh	151	104	68.9	80.8	72.2	116	76.8	82.8	76.2
1jse	58	42	72.4	87.9	70.7	46	79.3	89.7	75.9
1kpf	52	38	73.1	82.7	69.2	41	78.8	84.6	80.8
1nox	79	57	72.2	91.1	75.9	57	72.2	88.6	75.9
1qto	55	34	61.8	72.7	63.6	39	70.9	78.2	70.9
1rcd	101	63	62.4	73.3	65.3	75	74.3	82.2	74.3
1rhs	141	99	70.2	85.1	77.3	102	72.3	83.0	74.5
1u9a	79	51	64.6	83.5	60.8	56	70.9	88.6	65.8
1wer	175	89	50.9	66.9	56.0	100	57.1	68.0	57.7
2a0b	50	25	50.0	70.0	54.0	31	62.0	80.0	62.0
2fcb	98	63	64.3	76.5	68.4	75	76.5	83.7	77.6
2ilk	77	44	57.1	79.2	71.4	49	63.6	80.5	70.1
2pth	75	50	66.7	76.0	60.0	56	74.7	81.3	73.3
3lzt	47	39	83.0	89.4	80.9	39	83.0	91.5	83.0
3vub	44	28	63.6	79.5	63.6	30	68.2	84.1	75.0
Overall	2793	1877	67.2	81.5	69.6	2046	73.3	83.7	73.5

For each single side chain prediction test target, the number of correctly predicted residues based on the RMSD < 1.5 Å criterion, the percent of correctly predicted residues based on the RMSD < 1.5 Å criterion as well as the commonly used $|\Delta\chi_1| < 40^\circ$ and $|\Delta\chi_{1,2}| < 40^\circ$criteria are listed.

with the longer side chains in these categories (Gln and Glu) appears to be quite challenging, presumably because longer side chains have greater opportunity for finding incorrect structures exhibiting energy errors; nevertheless, the assignment protocol does lead to significant improvements in both cases. However, the ICDA assignment has little effect on the long flexible basic groups such as Lys and Arg, whose side chain prediction errors are primarily attributable to energy errors, rather than problems with protonation states. This can be understood by the fact that the pKa's of Lys and Arg are usually very high (~10), it is very rare to observe deprotonated states for these residues, as also demonstrated by various other studies.

Based on the above observation, we shall now focus our attention on the carboxylic acid/histidine protonation states assignment and Asn/Gln amide mis-assignment.

Carboxylic Acid Protonation States

The protonation state of a carboxylic acid residue (Asp/Glu) can often be determined by simple hydrogen bond network analysis. For example, the close proximity of two carboxylic oxygens often serves as an indication that at least one of them must be protonated. To determine exactly which one is to be protonated can be determined by our all atom energy calculations of the alternative protonation patterns. In the case of the close proximity of a carboxylic acid and a backbone oxygen (either from the same or a different residue), it is usually the case that the carboxylic acid is protonated on that oxygen. The algorithm succeeds in a reasonable number of cases, but as indicated above, there are still a nontrivial fraction of side chains where energy errors are a serious problem.

Histidine Protonation States

Histidine is probably the most complicated among all the polar residues considered in this work, because it has two alternative protonation sites (N_δ and N_ε) as well as the possibility of side chain χ_2 flipping, resulting in a total of 6 polar states, making the prediction of histidine side chain conformation a challenging task.

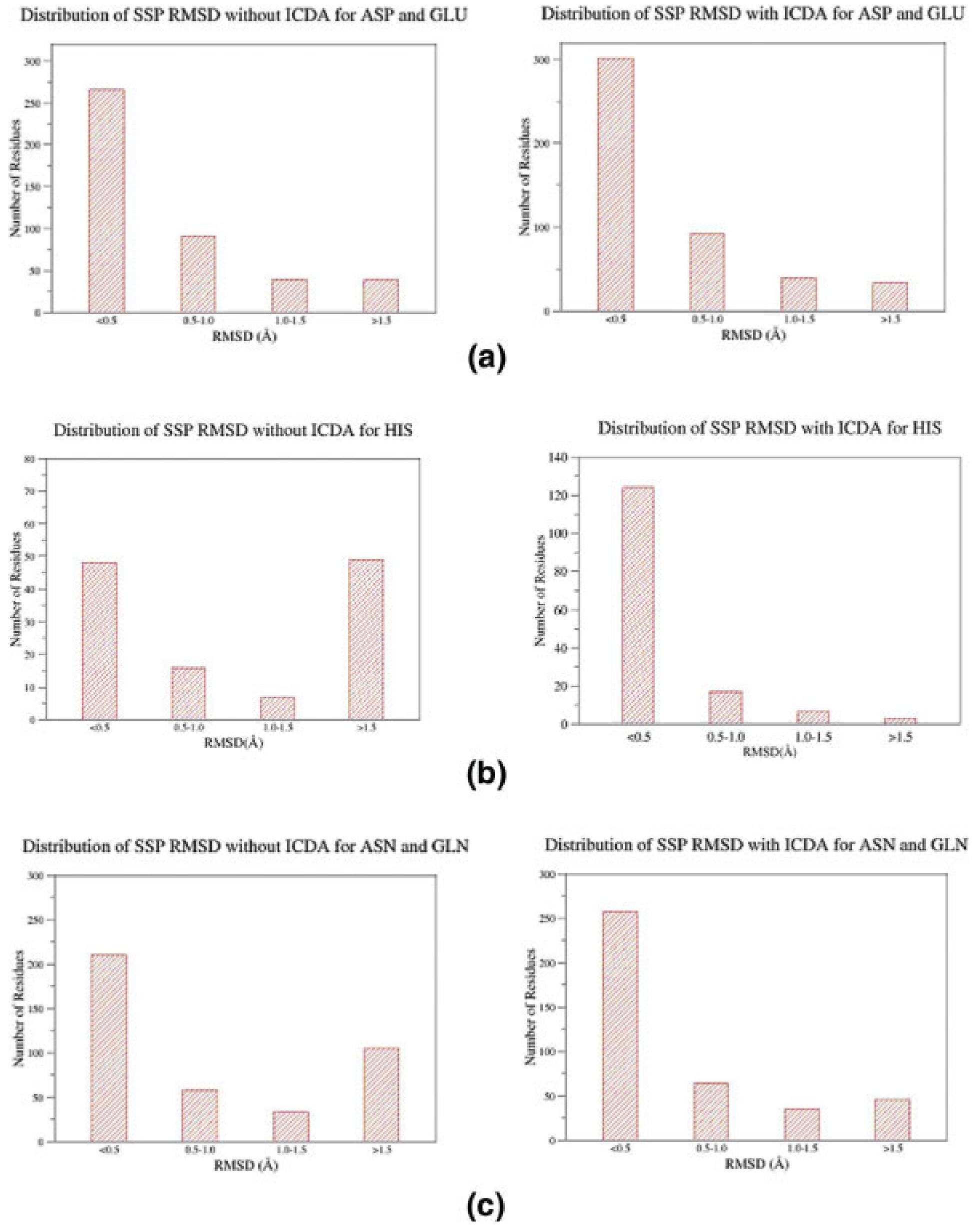


Fig. 3. Distribution of single side chain prediction (SSP) accuracies (in terms of RMSD for side chain atoms) before and after application of ICDA polar states assignment procedure for the subsets of: (**a**) carboxylic acids (Glu, Asp); (**b**) histidines; (**c**) Asn/Gln.

As shown in Figure 3, the single side chain prediction results improves substantially after the ICDA assignment. Table V gives an example in which two His residues form a cluster with a Glu. Default assignment yields very poor SSP prediction results, while after the correction of polar states via ICDA, the SSP results significantly improved, as reflected in the RMSDs.

Asn/Gln Amide Group Flipping Detection

Our algorithm successfully detects and corrects a large portion of the Asn/Gln amide group mis-assignments, which lead to significant improvement in the SSP prediction accuracy as shown in Figure 3. These potential mis-assignments are identified either based on the incompatible hydrogen bond pattern as explained previously or by the all atom energy calculations. These two different levels of methodologies complement each other and prove to be very powerful in the successful identification of many potential side chain flippings of Asn/Gln residues. For example, the three residues: Gln56, Asn61 and Asn114, which belong to different clusters in the protein 1qto, are all identified by WHATIF to be potential amide mis-assignments. This also leads to the poor single side chain prediction results for these 3 residues. For the latter two residues, the mis-assignments are identified by the obvious incompatible

TABLE V. Effect of Protonation State Prediction Using ICDA for Cluster 34 in Protein 2fcb, which consists of 3 Polar Residues: Glu70, His85, and His155

	Default protonation state	RMSD of SSP with default protonation state	ICDA predicted protonation state	RMSD of SSP with predicted protonation state
Glu70	Glu	2.71	Glu	0.16
His85	Hid	4.16	Hie	0.23
His155	Hid	2.21	Hip	0.13

SSP is an acronym for Single Side chain Prediction. For this cluster, a total of 180 sets of compatible states are found. The optimal solution and the default protonation state as assigned based on the modal pKa values of amino acids are both listed. The optimal assignment yields an all atom energy of –10,356 kcal/mol, when compared with the next and third lowest energy assignment, whose energies are –10,354.08 kcal/mol and –10,342 kcal/mol, respectively.

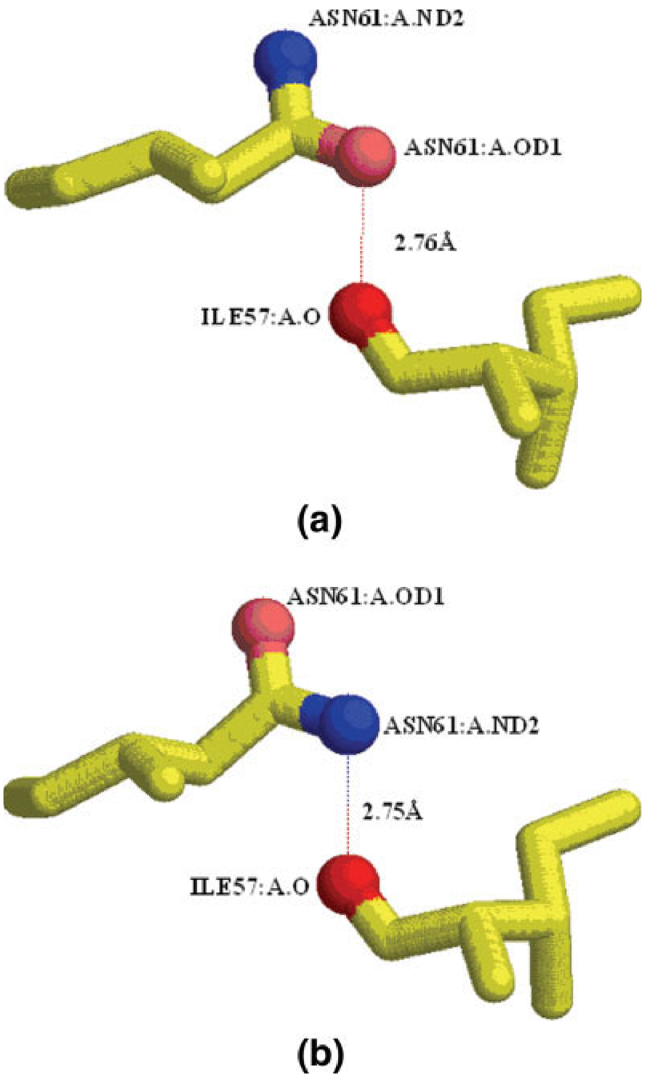


Fig. 4. Illustration of the misassignment of the amide groups of Asn61 of 1qto, and the correction of it by our ICDA assignment algorithm: (**a**) the default (incorrect) assignment and (**b**) the assignment predicted by ICDA. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

TABLE VI. Comparison of Single Side Chain Prediction Accuracies for the Three Asn/Gln Residues in Protein 1qto

	RMSD of SSP with default amide positions	RMSD of SSP with predicted amide positions
Gln56	1.58	0.11
Asn61	1.83	0.12
Asn114	1.83	0.18

TABLE VII. Comparison of SSP Accuracies (RMSD measure) Between ICDA, WHATIF, and REDUCE

Residue	Total	SSP accuracy (%)			
type	count	Default	ICDA	WHATIF	REDUCE
Asn	293	67.6	84.0	72.3	78.5
Gln	226	50.9	64.6	57.3	53.8
His	151	47.0	78.8	43.7	45.7
Asp	303	80.2	83.5	77.3	75.2
Glu	326	54.3	62.6	56.0	48.3
Ser	331	79.8	78.9	76.4	74.0
Thr	321	87.2	87.5	86.9	86.6
Lys	331	48.0	50.8	48.3	47.4
Arg	260	48.9	49.6	48.2	51.0
Cys	53	94.3	88.7	92.4	92.5
Tyr	198	97.5	97.0	97.5	97.5
Overall	2793	67.2	73.3	67.5	66.6

Here the prediction accuracies are broken into residue types. For each (polar) residue, we report the total number of occurrences of the residue type, as well as the SSP prediction accuracy (%) for default, ICDA, WHATIF, and REDUCE.

hydrogen bond patterns without any energy calculations: for Asn61 the amide oxygen is in close contact with the backbone oxygen of Ile57, with an interatomic distance of 2.77 Å. For Asn114, a similar scenario arises: the distance between the amide oxygen and the backbone oxygen of Thr12 is 2.80 Å. For both of these cases, the O—O distances are too close, and therefore the amide oxygen must be swapped with the nitrogen atom on the same group. Figure 4 illustrates the Asn61 case. For Gln56, no obvious hydrogen bond incompatibility is detected, but the energy calculation of the two states indicates that the original assignment is less energetically favorable than the alternative, flipped state (−6684.3 kcal/mol vs. −6695.3 kcal/mol). The significant improvement of the single side chain predictions, as shown in Table VI, also suggests the correctness of the assignments.

Table VII compares the single side chain prediction accuracies after the crystal structures are corrected by three assignment schemes: ICDA, REDUCE and WHATIF. The

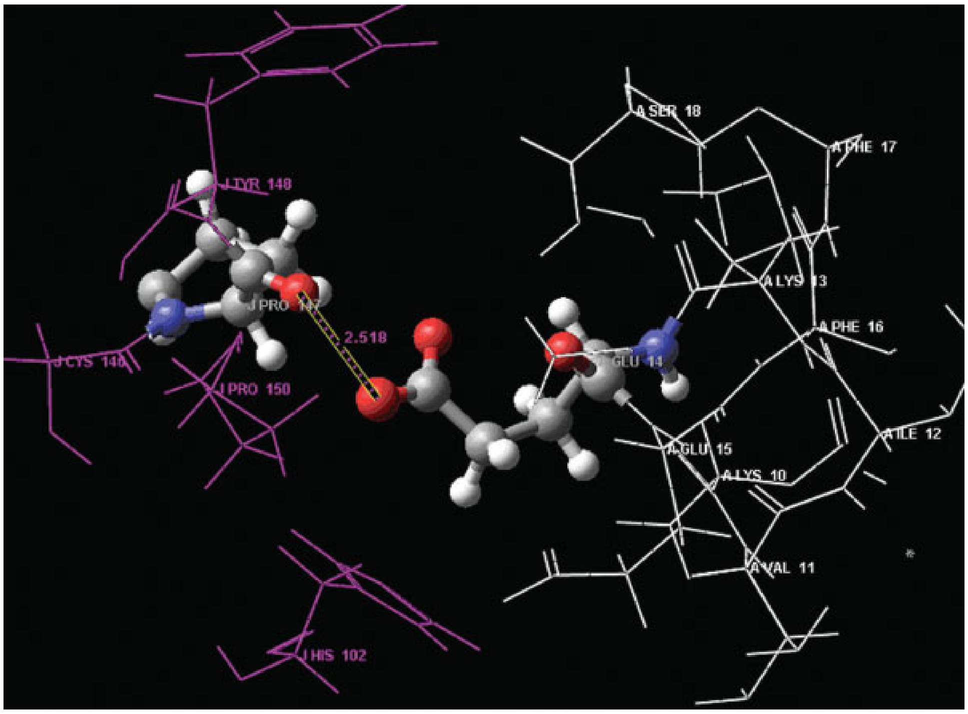


Fig. 5. The close hydrogen bond contact between Glu14:OE2 in asymmetric unit and Pro147:O in crystal copies in the 1a8l crystal structure, illustrating the impact of crystal packing environment on the protonation states. The protein and its crystal copies are colored by white and purple respectively, while the two hydrogen bond partner residues are shown with ball and stick model. To relax the strong electrostatic repulsion with the backbone oxygen of J: Pro147 and A: Glu14 must be protonated on OE2. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

ICDA assignments perform significantly better in this test than these alternatives, with a particularly dramatic effect on histidine prediction (the most challenging case due to the interplay of protonation states and tautomerism). REDUCE and WHATIF do not show significant improvements in the SSP test relative to the default assignments for many of the side chains.

It should be noted that in a small fraction of cases, ICDA assignment converts side chains for which the default polar state assignment gave reasonable SSP results into a case where the results are degraded. These cases in general are due to difficulties with the energy model, which can lead to inversion of the ranking of the protonated and deprotonated states in unfavorable cases. The elimination of errors of this type would obviously be desirable, and is a priority for future work.

Effect of Crystal Packing

Crystal packing effects have been demonstrated to play an important role in the structural details of protein conformation. In the current study, in order to probe the effect of crystal packing on the polar residue states, we reran the whole process (polar states prediction + single side chain optimization) on all the test cases without crystal packing and compared with the results with the crystal packing turned on. The effect of crystal packing can clearly be significant as has been pointed out by us in a previous publication.[20] In many cases, especially those with direct interactions between residues in neighboring copies of the asymmetric unit, the predicted polar states as well as single sidechain conformation are dramatically different.

A notable example is Glu14 in 1a8l, which is not included in the 30-protein test suite due to certain steric clashes present in the native structures. Without considering the crystal environment, the side chain of Glu14 appears not to form any hydrogen bond with any other residues and is predicted to be deprotonated. However, considering the crystal environment, the OE2 atom is detected to form a strong hydrogen bond with the backbone oxygen of PRO147, with an O-O distance of 2.52 Å, and therefore is predicted to be protonated on the OE2. Single side chain prediction test confirms the correctness of the latter assignment. When ignoring crystal packing and with Glu14 taken to be deprotonated the predicted RMSD is 3.0 Å. When considering the crystal packing environment in the SSP test while leaving Glu14 deprotonated, the results are a bit worse: the RMSD is 3.1 Å, possibly due to the repulsion of the backbone oxygen atom of PRO147. Only when we take crystal packing effects into consideration in the ICDA procedure do we get the correct answer for this case: the RMSD in the SSP test is now 0.2 Å, with the hydrogen bond network accurately reproduced, as shown in Figure 5.

TABLE VIII. Long Loop Prediction After Using ICDA

		Default assignment		ICDA assignment	
PDB	Residues	Egap (kcal/mol)	RMSD (Å)	Egap (kcal/mol)	RMSD (Å)
1edt	93–103	74.5	5.5	4.5	0.3
1eur	87–97	−7.4	4.6	−13.2	1.7
1hnj	191–203	−47.2	8.3	−43.8	3.1

The RMSD is calculated on the loop backbone atoms after superimposing the rest protein body. The Egap means the energy difference between predicted structure and native structure.

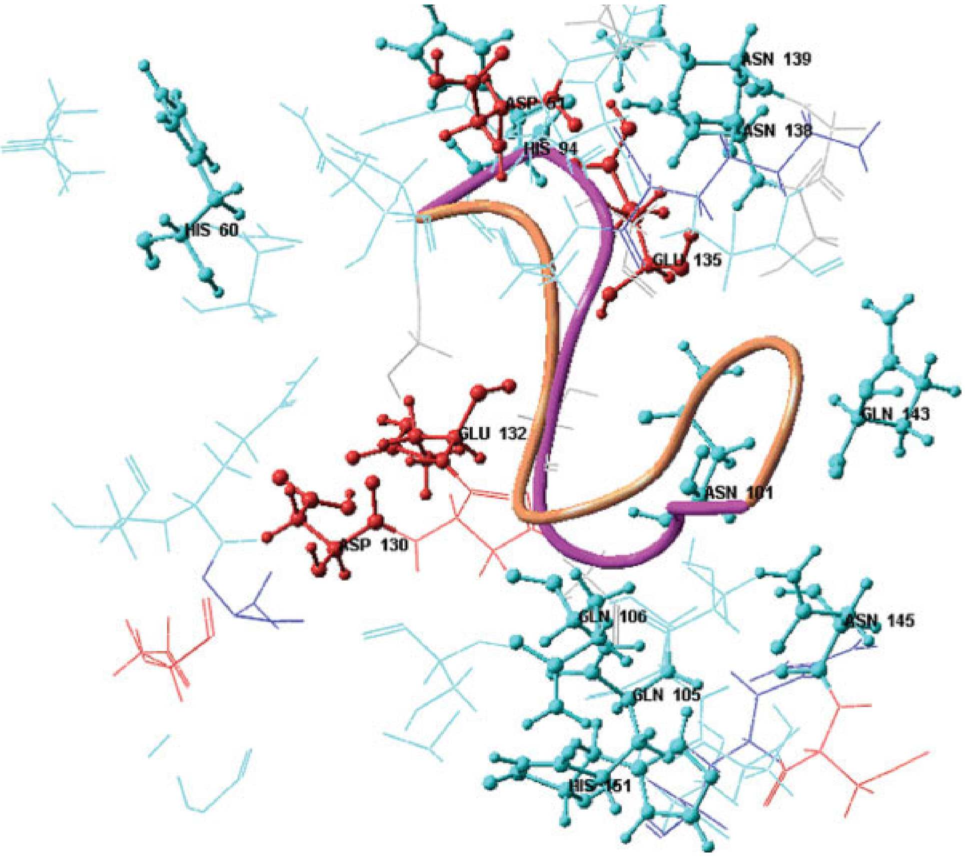


Fig. 6. Protonation state assignments affect prediction for 1edt. All the residues within 10 Å of the loop (residue 93–103) are shown, and colored by their polar properties, i.e. green for hydrophobic (here omitted for clarity), cyan for polar, blue for positive charged, red for negative charged, and gray for Gly. The residues whose protonation states are changed by assignment algorithm are represented by ball and stick. The purple ribbon is the correctly predicted loop geometry, and the orange one is the incorrectly predicted loop without ICDA protonation assignment. The native structure is omitted because it is almost identical to the correctly predicted structure. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Effects of ICDA on Loop Prediction

The assignment of protonation states can have a significant effect on the accuracy of loop prediction results. As we discussed in a previous study,[19] a number of loop prediction failures can be attributed to the misassignment of protonation states and therefore, we tried to filter out such cases in our test set by restricting the crystallization pH to be near neutral (6.5–7.5) in order to focus on the development of the methodology for loop prediction. This test set filtering reduces errors related to incorrect protonation state assignments, but does not eliminate such errors completely. In our recent loop prediction experiments[17] on a set of long loop targets ranging from 11 to 13 residues, we encountered a number of situations in which protonation state mis-assignments were responsible for failure of the loop prediction. Here we investigate whether using ICDA assignment of the crystal structures could eliminate these errors.

We have applied ICDA to a majority of the substantive prediction failure cases (all cases with prediction global backbone RMSD greater than 4 Å), and found that the ICDA algorithm gave alternative protonation states in the target loop region or neighboring amino acids for several cases. Using the structures generated by the ICDA algorithm for the loop prediction improves the results substantially, as shown in Table VIII. One of these cases, 1edt (residues 93–103), is illustrated in Figure 6. Within 10 Å of the target loop, there are a total of 14 residues whose protonation states are changed or χ_2 angles are flipped based on the ICDA assignment algorithm. Also, His 94 in this loop region is predicted to be protonated on both nitrogen atoms. Using the new structure yields a 0.3 Å RMSD prediction while the standard assignment results in a 5.9 Å RMSD prediction. In contrast to the single side chain prediction results, no single hydrogen bond or salt bridge appears to be responsible for the previous failure or current success; however, it is reasonable to assume that the ICDA hydrogen atom assignments provides a more correct description of the electrostatic environment.

CONCLUSIONS

We have developed a new method for predicting protonation states, hydrogen atom positions, and side chain orientations of His/Asn/Gln, all of which are ambiguous in a large majority of protein crystal structures. Two novel aspects of our methodology (ICDA) are an independent cluster decomposition strategy to reduce the exponential search space of polar residue states, and the use of an all-atom physical chemistry based energy function plus a Generalized Born implicit solvent model.

In addition to examining anecdotal cases, the method has been quantitatively evaluated by assessing improvements in single side chain prediction, comparing with a default assignment strategy as well as two competitive methods, REDUCE and WHATIF. The results suggest that ICDA represents a significant advance in the ability to assign polar states, although it does not yet represent a complete solution to the problem. We are investigating several possible improvements to the current algorithm. For example, there are quite a few heuristics in the choice of parameters, e.g., the cutoff distance for hydrogen bond interaction. The neglect of long-range interactions, although a good approximation to the first order, is still a potential error in the prediction results. Another major source of errors comes from systematic errors in the energy functions, particularly the SGB solvation model. This is particularly severe for basic residue groups such as Lys and Arg, which have long flexible side chains and possess many alternative hydrogen bond patterns.

ACKNOWLEDGMENTS

XL thank David L. Pincus and Raj Chakrabarti in Columbia University for many stimulating discussions and suggestions.

REFERENCES

1. Baptista A, Teixeira V, Soares C. Constant-pH molecular dynamics using stochastic titration. J Chem Phys 2002;117:4184–4200.
2. Bashford D, Karplus M. pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry 1990;29:10219–10225.
3. Juffer A, Argos P, Vogel H. Calculating acid-dissociation constants of proteins using the boundary element method. J Phys Chem B 1997;101:7664–7673.
4. Lee M, Freddie R, Brooks C. Constant-pH molecular dynamics using continuous titration coordinates. Proteins: Struct Funct Genet 2004;56:738–752.
5. Rabenstein B, Ullmann G, Knapp E. Calculation of protonation patterns with structural relaxation and molecular ensembles—application to the photosynthetic reaction center. Eur Biophy J 1998; 27:626–637.
6. Sandberg L, Edholm O. A fast and simple method to calculate protonation states in proteins. Proteins: Struct Funct Genet 1999; 36:474–483.
7. Yang A, Honig B. On the pH dependence of protein stability. J Mol Biol 1993;231:459–474.
8. Yang A, Honig B. Structural origins of pH and ionic strength effects on protein stability: acid denaturation of sperm whale apomyoglobin. J Mol Biol 1994;237:602–614.
9. Word J, Lovell S, Richardson J, Richardson D. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 1999;285:1735–1747.
10. Hooft R, Sander C, Vriend G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins: Struct Funct Genet 1996;26:363–376.
11. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8:52–56.
12. Jacobson M, Kaminski G, Friesner R, Rapp C. Force field validation using protein side chain prediction. J Phys Chem B 2002; 106:11673–11680.
13. Kaminski G, Friesner R, Tirado-Rives J. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 2001;105:6474–6487.
14. Jorgensen W, Maxwell D, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Sci 1996;118:11225–11236.
15. Ghosh A, Rapp CS, Friesner RA. Generalized born model based on a surface integral formulation. J Phys Chem B 1998;102:10983–10990.
16. Gallicchio E, Zhang L, Levy R. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. J Comput Chem 2002;23:517–529.
17. Zhu K, Pincus D, Zhao S, Friesner R. Long loop prediction using the protein local optimization program. Proteins: Struct Funct Genet 2006;65:438–452.
18. Li X, Jacobson M, Friesner R. High resolution prediction of protein helix positions and orientations. Proteins: Struct Funct Genet 2004;55:368–382.
19. Jacobson M, Pincus D, Rapp C, Day T, Honig B, Shaw D, Friesner R. A hierachical approach to all-atom protein loop prediction. Proteins: Struct Funct Genet 2004;55:351–367.
20. Jacobson M, Friesner R, Xiang Z, Honig B. On the role of crystal packing forces in determining protein side chain conformations. J Mol Biol 2002;320:597–608.
21. Cormen T, Leiserson C, Rivest R, Stein C. Introduction to algorithms. Cambridge, MA: MIT Press; 2001.
22. Huang E, Koehl P, Levitt M, Pappu R, Ponder J. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. Proteins: Struct Funct Genet 2004; 33:204–217.
23. Canutescu A, Shelenkov A, Dunbrack R. A Graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001–2014.