Statistically Based Reduced Representation of Amino Acid Side Chains‡

Jan K. Rainey† and M. Cynthia Goh*

Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada

Received August 12, 2003

Preferred conformations of amino acid side chains have been well established through statistically obtained rotamer libraries. Typically, these provide bond torsion angles allowing a side chain to be traced atom by atom. In cases where it is desirable to reduce the complexity of a protein representation or prediction, fixing all side-chain atoms may prove unwieldy. Therefore, we introduce a general parametrization to allow positions of representative atoms (in the present study, these are terminal atoms) to be predicted directly given backbone atom coordinates. Using a large, culled data set of amino acid residues from high-resolution protein crystal structures, anywhere from 1 to 7 preferred conformations were observed for each terminal atom of the non-glycine residues. Side-chain length from the backbone C^{α} is one of the parameters determined for each conformation, which should itself be useful. Prediction of terminal atoms was then carried out for a second, nonredundant set of protein structures to validate the data set. Using four simple probabilistic approaches, the Monte Carlo style prediction of terminal atom locations given only backbone coordinates produced an average root mean-square deviation (RMSD) of ∼3 Å from the experimentally determined terminal atom positions. With prediction using conditional probabilities based on the side-chain χ_1 rotamer, this average RMSD was improved to 1.74 Å. The observed terminal atom conformations therefore provide reasonable and potentially highly accurate representations of side-chain conformation, offering a viable alternative to existing all-atom rotamers for any case where reduction in protein model complexity, or in the amount of data to be handled, is desired. One application of this representation with strong potential is the prediction of charge density in proteins. This would likely be especially valuable on protein surfaces, where side chains are much less likely to be fixed in single rotamers. Prediction of ensembles of structures provides a method to determine the probability density of charge and atom location; such a prediction is demonstrated graphically.

INTRODUCTION

Many biophysical techniques are not suitable for routine atomic level determination of protein structure. Such methods are often more suited to proteins or supramolecular complexes which cannot be determined to high-resolution. For example, a feature such as topography in a scanning probe microscopy image may be correlated to protein sequence without the determination of an entire structure.[2] For many investigations into protein structure and function, the prediction or experimental determination of the precise positions of all atoms in amino acid side chains is unnecessary or unwieldy, if even possible. From a physicochemical perspective, only two properties of side chains need to be represented: chemical character (i.e., charge, dipole, aromaticity, aliphaticity) and extent of steric bulk. To allow the efficient and accurate reduced representation of amino acid side chains, we employ the statistical principles employed in rotamer library compilation to produce a new set of parameters allowing both reduced representation and prediction of amino acid conformations. The applicability of this representation is demonstrated through a Monte Carlo styled prediction of side-chain positions in a large set of known protein structures.

Reduced, or minimal, modeling of proteins is by no means a new concept. Various representations have been developed over the past 3 decades or so. In the most minimal case, the entire protein may be represented in a schematic manner such as an ellipsoid.[3,4] Beyond a simple ellipsoidal representation, surface features, or patches, may be taken into account to better represent chemical heterogeneity and predict potential interactions.[5] This style of minimal model is most effective if a tertiary structure is already known. For protein folding simulations, reduced models typically consist either of a polymer physics style approximation[6] or of an approximated backbone along with side chains represented by their centroid following the initial work of Levitt and Warshel.[7−10] Other approaches include the following: multiple virtual atoms[11−13] or centroids;[14] single, mobile bead representations;[15] and, multiple spheres with defined chemical characteristics.[16,17]

Dihedral angle preferences for both the backbone and side-chain atoms, following from Ramachandran and co-workers,[18,19] have been calculated, experimentally measured, and extensively tabulated. Side-chain dihedral angle preferences are now typically classed into specific clusters, termed rotamers, following the 1987 work of Ponder and Richards.[20] Several rotamer libraries have been compiled since this

* Corresponding author phone: (416)978-6254; fax: (416)978-4526; e-mail: cgoh@chem.utoronto.ca.

† Current address: Protein Engineering Network of Centres of Excellence (PENCE), University of Alberta, Edmonton, Alberta T6G 2S2, Canada.

‡ Abbreviations used: amino acid 3-letter codes and dihedral angles ϕ, ψ, ω, and χ_1 as standard.[1] ASA − accessible surface area; $\mathbf{A^T}$ − terminal atom (see Table 1); μ, σ, f*- mean, standard deviation, and mode, respectively; PDB − protein data bank; RMSD − root-mean square deviation. Probabilistic methods for fixing $\mathbf{A^T}$: MP − most probable; PC− probability based on counts; PSS − probability based on secondary structure; PD − probability based on backbone dihedrals; CHI − probability based on χ_1 rotamer.

10.1021/ci034177z CCC: $27.50
Published on Web 02/21/2004

original work,[21,22] with various characteristics of rotamer preference taken under consideration. While rotamer libraries have proven highly useful for such applications as X-ray and NMR structure validation,[23,24] homology modeling,[25,26] and protein design and engineering,[27] determination of the entire side-chain structure must typically be carried out.

Given the existence of distinct rotamer conformations, we would like to bypass the intermediate atoms of a side chain and predict the location of the terminal atoms purely based upon the backbone, providing both surface topography and chemical functionality. This approach is most akin to the coarse-grained representation introduced by Keskin and Bahar,[13] where side-chain virtual bonds and atoms were determined by statistical fitting of a data set comprised of 302 proteins. We believe that there is a great deal of value to be gained by predicting terminal atoms instead of more nebulous centroidal, or virtual atom, positions closer to the backbone. Since terminal atoms locate the end of the side chain, a definite steric boundary is provided, despite the drastic reduction in atoms required. Furthermore, side-chain chemistry tends to take place at or near the terminal (as extensively covered in ref 28), making these atoms often of crucial importance in reactions or protein−protein interactions. It should be noted that many previous reduced representations are rather difficult to adapt to more general use, often requiring a great deal of mathematical expertise, or having been designed with a very specific application in mind. It is our goal to make the generally applicable statistical representations herein also readily accessible. While we have focused on terminal atoms, the framework and methods we have developed are equally well suited to determination and parametrization of any side-chain atom.

Common methodology for rotamer library preparation is to select a culled set of high-resolution protein structures from the PDB (Protein Data Bank[29,30] http://www.rcsb.org/pdb). Since the work of Ponder and Richards, culling has been based on a cutoff for acceptable structure resolution (e.g. only structures below 1.7−2.0 Å) and maximum allowable sequence identity (e.g. no more than 25−50% identity) within the library. The rapid growth in the number of structures deposited within the PDB has led to notable increases in rotamer library resolution (e.g. refs 21, 31, and 32) and the introduction of less frequently observed rotamers. With such large numbers of high-resolution crystallographic examples of each type of amino acid now available, culling may be carried out with great zeal without sacrificing statistically devastating numbers of samples. In the recent library of Lovell et al.,[31] the traditional culling techniques were expanded upon. The veracity of the reported coordinates for each residue were taken into account, as initially detailed by Word et al.[33] Namely, those residues with high-temperature factors, alternate structures, overlapping van der Waals radii, or at the N- or C-termini were not included within the library. We have followed this improvement in culling procedures in order to produce our library of parameters for reduced residue representation.

One potential weakness of reliance on all-atom rotamer libraries and methodology is the manner in which side chains are typically fixed in a single rotameric conformation. On a protein surface, there is no reason to expect that a side chain will assume a single conformation; rather, it should statistically sample all favorable rotamer states with probabilities influenced by neighboring residues and by intermolecular interactions with binding partners, solvent, or ions. This dichotomy between rotameric flexibility and rigidity is demonstrated very clearly by Daley and Sykes.[34] The ability to efficiently carry out rigorous probabilistically based prediction of topology and surface charge density is likely one of the most appealing features of the reduced representation parameters introduced herein. A Monte Carlo style prediction is used to demonstrate this capability.

We feel that a set of parameters combining both the chemical and conformational nature of the side chain while minimizing the amount of information to be manipulated will prove extremely useful for modeling and representation as well as for computational and experimental investigations of protein structures and interactions. Extensive optimized prediction methods are not developed herein; instead, prediction of side chains is carried out in a rudimentary, randomized manner in order to validate the parameter set. The goal in the present work is to provide these parameters and to demonstrate their effectiveness in representing terminal atom location.

METHODS

I. Extraction of Conformational Parameters. Definition of Side-Chain Parameters. Considering any amino acid, a minimum set of necessary parameters required to describe its side chain would be the location and type of its terminal atom or atoms, denoted $\mathbf{A^T}$. The atoms we have examined are listed in Table 1. As with any general coordinate in three-dimensions, three parameters must be provided in order to fix $\mathbf{A^T}$ in space. The three parameters we have chosen are illustrated in Figure 1: the distance from C^α to $\mathbf{A^T}$, r($C^\alpha \rightarrow \mathbf{A^T}$), and two angles relative to the plane containing N, C^α, and C′ for the residue in question: the angle from the plane to the vector $C^\alpha \rightarrow \mathbf{A^T}$, ζ_{above}, and the angle from the projection of this vector onto the plane to the vector $C^\alpha \rightarrow$ N, ζ_{in}. A more "minimal" definition of a plane, requiring definition only of the C_α backbone, for residue number n could be envisioned as that containing $C^\alpha(n-1)$, $C^\alpha(n)$, and $C^\alpha(n+1)$. For the present, however, we will discuss only the first of these plane definitions.

The sign convention we have chosen for these angles is based upon the stereochemistry of an L-amino acid. The cross product between the vectors $C^\alpha \rightarrow$ N and $C^\alpha \rightarrow$ C′ produces a vector lying in the direction of the C^β of an L-amino acid. We have chosen a positive ζ_{above} to mean that $\mathbf{A^T}$ lies on the same side of the plane as C^β. If the plane containing N, C^α, and C′ is halved along vector $C^\alpha \rightarrow$ N, the projection of C^β falls in the half of the plane opposite that containing C′. We have defined a ζ_{in} of less than 180° to fall in the half-plane not containing C′—i.e. a clockwise rotation of ζ_{in} away from vector $C^\alpha \rightarrow$ N in the plane will reach the projection of $C^\alpha \rightarrow \mathbf{A^T}$. These angular definition conventions are illustrated in Figure 2.

Selection, Parsing, and Culling of PDB Entries. A culled set of polypeptide chains from PDB entry files at a resolution of 1.8 Å or better and a maximum of 40% pairwise sequence identity was selected from the Oct 4, 2001 revision of the CulledPDB compiled by R. L. Dunbrack (http://www.fccc.edu/research/labs/dunbrack/pisces) based on the

Table 1. Side-Chain Atoms (Non-Hydrogen) Used as Representative $\mathbf{A^T}$ for Each Amino Acid Residue[c]

residue	in culled data set: number	in culled data set: percent	number of side-chain atoms: total	number of side-chain atoms: retained	$\mathbf{A^T}$ atoms retained
Gly	13699	8.88	0	0	
Ala	15245	9.85	1	1	C^{β}
Pro[a]	7387	4.77	3	1	C^{γ}
Pro (NT *cis*)[a]	396	0.26	3	1	C^{γ}
Cys[b]	1539	0.99	2	1	S^{γ}
Cystine[b]	933	0.60	2	1	S^{γ}
Ser	9109	5.88	2	1	O^{γ}
Thr	9692	6.26	3	2	$O^{\gamma 1}$, $C^{\gamma 2}$
Val	12048	7.78	3	2	$C^{\gamma 1}$, $C^{\gamma 2}$
Asn	7041	4.55	4	2	$O^{\delta 1}$, $N^{\delta 2}$
Asp	8732	5.64	4	2	$O^{\delta 1}$, $O^{\delta 2}$
Ile	9058	5.85	4	1	$C^{\delta 1}$
Leu	14246	9.20	4	2	$C^{\delta 1}$, $C^{\delta 2}$
Gln	4924	3.18	5	2	$O^{\epsilon 1}$, $N^{\epsilon 2}$
Glu	6910	4.46	5	2	$O^{\epsilon 1}$, $O^{\epsilon 2}$
Met	2998	1.94	4	1	C^{ϵ}
Arg	5846	3.78	7	2	$N^{\eta 1}$, $N^{\eta 2}$
Lys	6153	3.98	5	1	N^{ζ}
His	3643	2.35	6	2	$N^{\delta 1}$, $N^{\epsilon 2}$
Phe	6681	4.32	7	1	C^{ζ}
Trp	2533	1.64	10	2	$N^{\epsilon 1}$, $C^{\eta 1}$
Tyr	5931	3.83	8	1	O^{η}

[a] Pro have trans ($|\omega| > 135°$) peptide bonds at N- and C-terminal; Pro (NT cis) have cis ($|\omega| < 45°$) peptide bond at N-terminal but trans peptide bond at C-terminal. [b] Cys are free cysteine and cystine are cross-linked residues as determined by DSSP.[35] [c] The four backbone atoms (N, C^{α}, C′, and O) are included in the data set, although O is not needed for the backbone dependent parametrization developed. The numbers of each residue are retained, and the percentage of the total in the culled data set of 154 736 residues are also given. Residues are grouped by the length/complexity of the side chain up to those terminating at an ϵ-substituent and by the primary chemical character for aromatic and basic residues.

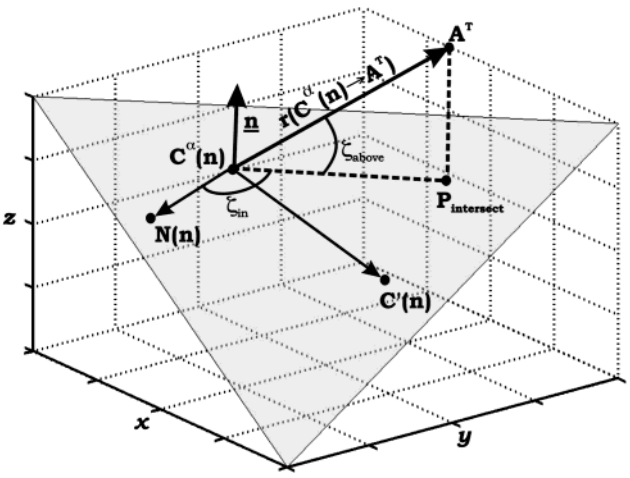

Figure 1. Definition of parameters determined for residue number n, where $\mathbf{A^T}$ represents an atom of interest in its side chain.

methodology developed by Hobohm, Sander, and co-workers.[36,37] This provided an initial data set of 792 polypeptide chains from 749 PDB entries with a total of 183 206 residues. The complete list of polypeptide chains employed, including the number of residues in the chain and resolution of the structure, is supplied in the Supporting Information.

Prior to usage, each PDB file was analyzed by DSSP[35] and PROCHECK.[38] All parsing, culling, and processing of PDB, DSSP, and PROCHECK derived data was carried out

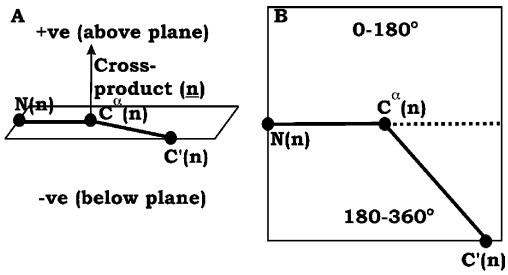

Figure 2. Conventions employed in definition of (A) ζ_{above} and (B) ζ_{in}. Note that **n** is given by the cross-product ($C^{\alpha} \rightarrow N$) × ($C^{\alpha} \rightarrow C'$).

using IDL 5.5 (Interactive Data Language, Research Systems, Inc., Boulder, CO) on an SGI Octane (Mountain View, CA) workstation. Source code is freely available upon request. The PROCHECK modified PDB entry files containing the IUPAC−IUP Commission preferred atom labels, as recently clarified by Markley et al.,[39] were then parsed for the following data concerning each residue: PDB entry code and chain; Cartesian coordinates for all backbone atoms and side-chain atoms of interest; the maximum B-factor of all non-H atoms; the presence of any atoms with occupancy less than 1.0; the existence of alternate specified conformations; and, N- or C-terminal character. The corresponding DSSP output files were parsed to obtain each residue's accessible surface area (ASA; as defined by Lee and Richards[40]) and, where applicable, to differentiate between cysteine and cystine residues. Parsing of PROCHECK output files provided the secondary structure assignment (expanded due to relaxed constraints over the DSSP derived assignment) and the region of the Ramachandran plot following the classification of Morris et al.[41] Exact backbone (ϕ, ψ, and ω for both N- and C-terminal directions) and side-chain (χ_1) dihedral angles were also calculated for each residue. Culling beyond the initial selection of the PDB chains employed was then carried out residue-by-residue. Any residues with an incomplete set of defined non-H atoms or with extra atoms due to co- or posttranslational modification (such modified residues are either flagged by a MODRES field or contain an extra non-H atom in the ATOM fields) were removed in order to prevent skewing of the statistics by such nonstandard conformations. Residues at the C- or N-terminal of a chain were also discounted, since these are likely not representative of the standard mid-peptide conformation.

Following the work of the J. and D. Richardson group,[31,33] three further culling parameters were employed. Any residue with an atom with reported occupancy less than 1.0 was removed; any residue with more than one reported set of coordinates (i.e. with alternate structures provided) was removed; and, any residue with an atom with a large (>40) temperature factor (B) or with an unspecified (0.0) temperature factor was disqualified. Note that any residue with the mean-square displacement of the atom ($\langle u^2 \rangle$) specified instead of B in the B-factor column of the PDB entry was first converted using $B = 8\pi^2\langle u^2 \rangle$, which assumes an isotropic and harmonic vibration of the atom within the crystal lattice.[42] Each of these three culling factors was shown by to be effective in reducing the number of residues involved in a steric clash.[33] Any such clashing residues are probably misreported or poorly determined in the initial protein structure.

A final culling factor was implemented: any nonproline residue with a cis ω-dihedral angle was excluded. While cis peptide bonds are becoming more acceptable in protein structures (recently reviewed in ref 21), the more prevalent *trans* conformation was deemed most relevant for our statistical analysis. In the case of proline residues, cis peptide bonded residues were considered as a separate data set from the trans peptide bonded residues. Our data set of culled proline residues contains 5.05% cis peptide bonds, which is on the lower end of the 5.0−6.5% range reported in previous studies.[21,43] This value is 2 orders of magnitude greater than the 0.047% occurrence of cis peptide bonds in the culled nonproline residues after all other filters have been applied, which is near the upper end of the 0.03−0.05% occurrence rates shown in the past.[21]

In combination, the seven culling procedures beyond initial chain selection provided a final data set of 154736 residues. (Data concerning the number of residues flagged for culling by each culling procedure are provided in the Supporting Information.) This represents a culling rate of 15.5%. Interestingly, ~36% of the residues culled were flagged for removal by more than one of these parameters. The number of each residue retained and its percentage composition within the total culled data set is provided in Table 1.

Calculation of $\mathbf{A^T}$ Conformational Parameters. Calculation of the three parameters for $\mathbf{A^T}$ illustrated in Figure 1 was performed as follows. Note that vectors between two atoms A_1 and A_2 are written as $(A_1 \rightarrow A_2)$ during the subsequent discussion. The distance $r(C^\alpha \rightarrow \mathbf{A^T})$ is simply the norm of the vector $C^\alpha \rightarrow \mathbf{A^T}$:

$$r(C^\alpha \rightarrow \mathbf{A^T}) = ||C^\alpha \rightarrow \mathbf{A^T}||$$

To calculate the angles ζ_{above} and ζ_{in} for each residue, all coordinates for that residue were first normalized to an origin centered at C^α. The equation of the plane containing N, C^α, and C′, in form $Ax + By + Cz = 0$, is given by

$$(\mathbf{C}^\alpha(n) \rightarrow \mathbf{N}(n)) \times (\mathbf{C}^\alpha(n) \rightarrow \mathbf{C}'(n)) = \mathbf{n}$$

where the cross-product provides the normal to the plane, **n**, and the components of **n** are $n_x = A$; $n_y = B$; and, $n_z = C$.

The location of $P_{intersect}$, as defined in Figure 1, was then determined as follows. Starting from $\mathbf{A^T}$, a displacement opposite in direction to **n** of the shortest distance from $\mathbf{A^T}$ to the plane

$$r(\mathbf{A^T} \cdot P_{intersect}) = \frac{\mathbf{n} \cdot \mathbf{A^T}}{||\mathbf{n}||}$$

will provide $P_{intersect}$.

The magnitude of ζ_{above} is the three-dimensional angle between $P_{intersect}$ and $\mathbf{A^T}$. Its sign is given by the angle between $\mathbf{A^T}$ and **n**: if this angle is greater than 90°, $\mathbf{A^T}$ must lie below the plane, and the sign should be negative. The value of ζ_{in} is the three-dimensional angle between $P_{intersect}$ and $(C^\alpha \rightarrow N)$. If the cross product $P_{intersect} \times (C^\alpha \rightarrow N)$ is parallel to **n**, then $P_{intersect}$ falls in the positive half-plane defined in Figure 2B, if antiparallel, the negative half-plane.

Analysis of Preferential Side-Chain Conformers. Initial graphical examination of statistical data was performed on an octane using MATLAB 6.0 (The MathWorks, Natick, MA); we have since implemented comparable graphing with IDL. Regions of dense population in $(r(C^\alpha \rightarrow \mathbf{A^T}), \zeta_{ab}, \zeta_{in})$ space were selected visually using a variety of 2-D plots of ζ_{ab} vs ζ_{in},1-D histograms of each of the three parameters and 3-D plots in (r, ζ, ζ). In cases where multiple regions overlap in one of the parameters, an arbitrary cutoff was employed to separate regions. Fitting of multiple Gaussian or Laplacian distributions to such a histogram would not be immediately constructive, since only the parameter in question is represented on a given histogram, and analysis of further probabilistic information would not be obtained with such a fitting. (Exact cutoffs employed for regions are available upon request.) Unfortunately, standard deviations may be misrepresented in cases where regions overlap (less than 20% of regions) in that such statistical variables may appear narrower than they actually are due to data tails being cut off and causing artificially low σ. Conversely, distant outliers are often still within a given region since cutoffs are only imposed to separate regions of relatively high population leading to an increased σ. Use of modal values minimizes the impact of arbitrary region cutoffs. To determine modes, the data were divided into bins of width $\sigma/15$ over the range $\mu \pm (4\sigma + \sigma/30)$, from which the bin containing the highest frequency of counts was considered the mode. The bin width $\sigma/15$ was an arbitrary decision; the factor of $\sigma/30$ in the histogram range allows the possibility that the mode may be exactly equal to μ. Ellipsoid cutoffs calculated as $\mu \pm 2.33\sigma$ in each of r, ζ_{above}, and ζ_{in} for each cutoff region were used to count a variety of parameters used below in a probabilistic manner. This overall process will likely be improved through use of an automated and less arbitrary region selection algorithm. Text files containing all $\mathbf{A^T}$ cluster parameter data for each atom analyzed may be downloaded at www.pence.ca/~jrainey or www.chem.utoronto.ca/staff/MCG.

II. Predictive Validation of Parameters. All validation analysis was carried out using IDL 5.5 on an SGI Octane IRIX workstation or an Intel Pentium III 533 MHz Windows NT 4.0 workstation. Source code is freely available upon request.

Outline of Probabilistic Validation. A second large data set of high-resolution structures was used to validate the use of the reduced side-chain representations. Backbone coordinates for each member of this data set were used as a scaffold to build a set of statistically predicted reduced representation structures. A variety of strategies for choice of the $\mathbf{A^T}$ conformational probabilities were tested (for convenience 2−3 letter codes are used):

MP − "Most Probable" − using the absolutely preferred conformation (i.e. that with the highest number of residues in its $\mu \pm 2.33\sigma$ ellipse);

PC − "Probability based on Counts" − using the probability of finding the given atom in each of the $\mu \pm 2.33\sigma$ ellipses;

PSS − "Probability based on Secondary Structure" − using the relaxed DSSP[35] style PROCHECK[38] derived secondary structure assignments along with the conditional probability of finding an $\mathbf{A^T}$ in that secondary structure element for each of the $\mu \pm 2.33\sigma$ ellipses;

PD − "Probability based on backbone Dihedrals" − using the PROCHECK derived secondary regions of the Ramachandran plot along with the conditional probability of

finding $\mathbf{A^T}$ in that region of (ϕ, ψ) space for each of the $\mu \pm 2.33\sigma$ ellipses; and,

CHI − calculating the χ_1 dihedral angle and using the conditional probabilities in the $\mu \pm 2.33\sigma$ ellipse corresponding to the given χ_1 rotamer.

The mode, mean, and distribution of normalized ASA and distance from protein center (mode for ASA and distance from center calculated using a histogram of bin width 0.05 over the range $0 \rightarrow 1$) were also determined for each region. No significant correlations were noted between these factors and the regional populations.

The RMSD (root-mean-square deviation) of the predicted $\mathbf{A^T}$ conformation to the experimentally determined atom location is then determined for each prediction made with each of the five probabilistic methods as described in detail below. Backbone dependent rotamer predictions by SCWRL,[25] both in the initial, most likely rotamer position and the energy minimized form are used as a benchmark. Each of the five methods is considered as a whole as well as for the specific deviations observed for each residue.

PDB Structure Data Set Used for Validation. Using the SearchFields interface at the PDB Web site (date of search—September 25, 2002), a set of structures was chosen with a release date after Nov 2, 2001 to avoid duplication with the culled PDB list used in reduced representation statistics, a minimum chain length of 50 residues (with nucleic acid and carbohydrate structures excluded), resolution between 0.5 and 1.8 Å, and maximum 50% sequence identity allowed (SearchFields uses a Hobohm and Sander type algorithm[36,37] as implemented in cluster analysis by Li et al.[44]) This provided a set of 723 polypeptide chains. For each PDB entry, specific chains for analysis were then selected as either the longest chain in the file or as the first chain, if all chains were equal in length. This gives an initially culled data set of 374 chains, one for each of the 374 PDB files employed.

As with the initial parameter determination, the PROCHECK modified PDB files (".new" files) were employed, and nonrepresentative residues were culled. This provided a final test data set with 372 polypeptide chains and 116 785 non-glycine residues. PROCHECK derived secondary structure and Ramachandran regions were collected using the same criteria as above, and backbone and χ_1 dihedral angles were calculated.

Deriving $\mathbf{A^T}$ Coordinates. The calculation of Cartesian coordinates for an $\mathbf{A^T}$ given a set of (r, ζ, ζ) parameters and a set of coordinates for N, C^α, and C′ requires multiple steps. (This nontrivial process is provided as a detailed derivation in the Supporting Information.) This procedure is readily automated, allowing an $\mathbf{A^T}$ to be very efficiently fixed in space.

Probabilistic Parameter Choice. Output files generated during parameter calculation were parsed in order to produce each of the parameter and probability values. For simplicity in this analysis, the conditional probabilities of finding a second $\mathbf{A^T}$ for a residue given the location of the first are not taken into account—each side-chain atom is considered as an independent entity. Inclusion of these additional constraints should strengthen the predictive ability. If method MP was selected, the ($r(\mathbf{C^\alpha} \rightarrow \mathbf{A^T})$, ζ_{above}, ζ_{in}) parameters coinciding with the maximum number of residues in any $\mu \pm 2.33\sigma$ ellipse were assigned to each residue. If any other method was selected, the conditional probability, P(R | A), of finding the representative side-chain atom in a region R was calculated as

$$P(R \mid A) = \frac{N(R \mid A)}{\sum_R N(R \mid A)}$$

where N(R | A) is the number of residues in the $\mu \pm 2.33\,\sigma$ ellipse for region R given that condition A is true. (Note that all outliers are excluded from the sum in the denominator, leading to a total probability of 1 in all instances.) The condition A depends on the method being employed. For method PC, this simplifies to the residue being within the ellipse for the given region and could be written as P(R) instead of P(R | A). For method PSS, separate P(R | A) calculations were carried out for each secondary structure assignment of PROCHECK. Methods PD and CHI follow suit, with the appropriate conditions. A two-dimensional array, **PrA,** containing [R × A] elements in dimensions of R columns and A rows is then defined for each amino acid. Each element in this array is given as

$$\mathbf{PrA_{ij}} = \sum_i P(R_i \mid A)$$

such that each row of the array will sum to a total probability of one representing one of the conditions A_j.

For each chain, the following general procedure could then be employed. The numbers of each residue found in the PDB chain, and not culled, were counted. A vector of random numbers, **Rn**, was then generated such that $0 \leq \mathbf{Rn_k} \leq 1$ with each element k representing each instance of a given residue. This vector was simply extended by a factor of N in order to easily and quickly produce values for an ensemble of N predicted structures. Each element $\mathbf{Rn_k}$ was associated with its given condition A_j, such as secondary structure type. The ($r(\mathbf{C^\alpha} \rightarrow \mathbf{A^T})$, ζ_{above}, ζ_{in}) parameters predicted for each atom number k of the given residue are then given by the region corresponding to

$$i = \max(\mathbf{Rn_k} \leq \mathbf{PRA_{ij}})$$

where only the row corresponding to the appropriate condition A_j in array **PrA** is considered in this inequality and the max function gives the maximum column element, i, holding with the inequality. Parameters for each atom may then be rapidly assigned for large numbers of residues. The method in the Supporting Information may then be used to assign the Cartesian coordinates for each $\mathbf{A^T}$ assignment using the associated backbone coordinates. A PDB format file may then be written for external analysis or graphical exploration, or the RMSD of the prediction versus the experimentally determined coordinates may be calculated directly for either entire predicted structures or for a given $\mathbf{A^T}$.

III. Prediction of Surface Charge Density and Topology. The backbone coordinates for pepsin (PDB entry 5PEP[45]) were used to produce an ensemble of 1015 predictions of each $\mathbf{A^T}$ for all residues with method PC. A script was developed in Tcl/Tk (a powerful freeware scripting language available at www.tcl.tk) to carry out this prediction using an algorithm similar to that discussed in section II; this script is freely available under the GNU

general public license from www.pence.ca/~jrainey or www.chem.utoronto.ca/staff/MCG. The Cartesian space of the protein was divided into cubic volume elements with dimension 0.5 × 0.5 × 0.5 Å using the *x*, *y*, and *z* axes defined in the original PDB file and the numbers of atoms, positive and negative charges were counted in each volume element using a second Tcl/Tk script, again freely available. Charges were counted as the following: +1 for N^{ζ} of Lys, +0.5 for $N^{\eta 1}$ and $N^{\eta 2}$ of Arg; −0.5 for $O^{\delta 1}$ and $O^{\delta 2}$ of Asp and $O^{\epsilon 1}$ and $O^{\epsilon 2}$ of Glu; residue 23 was considered as a cis Pro. The freeware open source program rotater by Craig Kloeden (modified slightly from version 5.0b2 for Apple Macintosh OS X available from http://casr.adelaide.edu.au/rotater/ to provide additional coloring capabilities) was used to visualize the predicted atom and charge density using an input file generated automatically by the same Tcl/Tk script used to count atom and charge elements. (A volume element output file could equally easily be generated for calculation purposes instead of graphing, if an application beyond simple visualization is desired.) Briefly, each volume element is represented as a cube with a gray scale intensity depending on the number of atoms contained within it and normalized to the largest count observed; if the element contains positive or negative charge, it is colored blue or red respectively, again with hue intensity determined by the count of charges; finally, this is superimposed on a C^{α} trace colored in green, with residues connected in order from N- to C-terminal.

RESULTS AND DISCUSSION

Preferred Side-Chain Conformations. The initial culled PDB data set for $\mathbf{A^T}$ calculation contained 792 polypeptide chains from 749 PDB entries, with 183 206 entries. The further culling provided a final data set of 154 736 residues, with $\mathbf{A^T}$ data sets ranging in size from 396 proline residues with N-terminal cis and C-terminal trans peptide bonds, 933 cystine residues, 1539 cysteine residues, and 2533 tryptophan residues up to 14 246 leucines and 15 245 alanines. In examining the three-dimensional distributions of $\mathbf{A^T}$ groups, anywhere from 1 to 7 clusters in (r, ζ, ζ) space containing ~5% or more of a given side-chain atom may be readily located. $\mathbf{A^T}$ groups found in 1−3 clusters are summarized in Table 2; the remaining $\mathbf{A^T}$ group conformations have 4−7 clusters, given in the Supporting Information. Note that while only the $\mathbf{A^T}$ groups listed in Table 1 are analyzed herein, all substituents from $\mathbf{A}^{\beta}$ out to $\mathbf{A}^{\eta}$ are represented, indicating the ability to calculate a reduced representation at any position in the side chain. As may be expected, the Tyr C^{ζ}, for example, displays preferences extremely similar to those for Phe C^{ζ} (Rainey and Goh, unpublished). Preferences of each cluster with respect to secondary structure, (ϕ, ψ) dihedral angle, accessible surface area, and distance from the protein center were calculated; these will be made freely available in text file format (www.pence.ca/~jrainey or www.chem.utoronto.ca/staff/MCG), since their envisioned utility is in an automated probabilistic use in computation (as used herein for validation), rather than explicit consideration of each preference.

As one would expect, the side-chain length increases directly with the number of bonds between $\mathbf{A^T}$ and C^{ξ} (Figure 3). Notably, as the number of bonds to $\mathbf{A^T}$ increases, the spread in the r($C^{\alpha} \rightarrow \mathbf{A^T}$) values observed also increases. Although anywhere from 1 and 7 conformational clusters are observed for the various $\mathbf{A^T}$ groups examined, the populated regions of (ζ_{above}, ζ_{in}) space are noticeably clustered. This is demonstrated in Figure 4, where all observed angular conformation pairs are plotted for each of the tabulated $\mathbf{A^T}$ groups. Such clustering in the reduced space defined herein is not immediately intuitive simply based upon all-atom rotamers but seems illustrative of the potential power of a reduced approach. A brief overview of characteristics observed for each class of side chain will now be given prior to discussion of the predictive validation.

Alanine. Alanine C^{β} atoms cluster tightly in (r, ξ, ξ) space. An extremely tight r($C^{\alpha} \rightarrow C^{\beta}$) distribution is seen, with μ 1.526 Å, f^* 1.523 Å, and σ 0.015 Å. This C−C distance is extremely close to the 1.521 Å reported by Engh and Huber in 1991.[46] The standard deviation of our result (0.015), however, is less than half that reported by Engh and Huber (0.033), implying improved statistics due to the much larger high resolution data set now available. The angular distribution relative to the backbone plane in Figure 1 also demonstrates a very tightly clustered preferential conformation.

Proline. The fact that proline can only assume g^+ and g^- χ_1 rotamers is directly mirrored by the observed (r, ζ, ζ) clustering pattern (Table 2). Interestingly, the proportion observed in each rotamer is practically equal for Pro residues with both N- and C-terminal trans peptide bonds, while the g^+ rotamer is strongly preferred by the set of Pro residues following an N-terminal cis peptide bond.

γ-Substituent Terminated Residues. As would be anticipated, each of the four residues with a terminal γ-substituent (cysteine and the disulfide linked cystine, serine, threonine and valine) displays three clusters in (r, ζ, ζ) space (Table 2). Each corresponds to one of the three χ_1 rotamers (g^-, t, or g^+). These correspond to the regions in (ζ_{above}, ζ_{in}) labeled 1, 2, and 3, respectively. Note that for region 3 (g^+) the distributions in ζ_{in} have extremely large σ. This is due to the fact that the $\mathbf{A^T}$ is nearly perpendicular to the backbone plane given in Figure 1 and that ζ_{in} therefore has an extremely minimal effect on the position of the residue.

The observed side-chain lengths vary depending upon the type of γ-substituent atom. In order of increasing length, these are O^{γ}(~2.3−2.4 Å), C^{γ}(~2.5 Å), and S^{γ}(~2.8 Å). These lengths are practically unaffected by branching, as can be observed by comparing the Ser O^{γ} lengths to Thr $O^{\gamma 1}$, or by the nature of the branch, in comparison of either Val C^{γ} with Thr $C^{\gamma 2}$. Stereochemistry dictates that locating one atom in a branched side chain must unambiguously provide the location of the second atom; this is upheld in the conditional statistics observed for the Thr and Val substituents (given in the Supporting Information.)

δ-Substituent Terminated Residues. As would be anticipated from the addition of at least one bond and atom to each side chain, the four δ-terminated residues (asparagine, aspartate/aspartic acid, isoleucine, and leucine) tend to display more complicated distributions than the previous classes. All four residues have branched side chains: Asn, Asp, and Leu branch at the γ-substituent, Ile at the β-substituent. The conformational preferences for each δ-substituent (Table 2, Figures 3 and 4, and Supporting Information) all overlap with each other in various ways. The conditional proportions of one $\mathbf{A^T}$ being located given

Table 2. Statistically Preferred Conformations of $\mathbf{A^T}$ for Given Residues with 1−3 Observed Preferred Conformations[a]

Ala-C$^{\beta}$	P	0.913		
N = 15 245	*f**	[1.523, 52.59, 124.48]		
	σ	(0.015, 2.38, 2.70)		
Pro-C$^{\gamma}$	P	0.458{0.965, 0, 0.000, 0.000}	0.444{0.000, 0.966, 0.000}	
N = 7387	*f**	[2.377, 37.04, 64.27]	[2.378, 67.07, 30.52]	
	σ	(0.029, 4.89, 3.10)	(0.032, 4.63, 10.69)	
Pro-C$^{\gamma}$ (*cis*)	P	0.119{0.951, 0.000, 0.000}	0.816{0.000, 0.976, 0.000}	
N = 396	*f**	[2.396, 45.67, 60.83]	[2.368, 69.80, 23.62]	
	σ	(0.032, 5.15, 3.02)	(0.036, 3.88, 11.99)	
Cys-S$^{\gamma}$	P	0.488{0.919, 0.000, 0.000}	0.264{0.000, 0.000, 0.925}	0.169{0.000, 0.925, 0.000}
N = 1539	*f**	[2.805, 24.36, 95.01]	[2.786, 26.74, 158.94]	[2.829, 83.70, 183.32]
	σ	(0.043, 5.03, 5.28)	(0.038, 4.12, 4.92)	(0.039, 4.27, 64.12)
Cystine-S$^{\gamma}$	P	0.627{0.926, 0.000, 0.000}	0.192{0.000, 0.000, 0.913}	0.099{0.000, 0.902, 0.000}
N = 933	*f**	[2.791, 28.52, 91.41]	[2.776, 29.60, 155.76]	[2.817, 87.74, 80.41]
	σ	(0.041, 4.97, 5.81)	(0.042, 5.02, 7.49)	(0.042, 5.08, 75.32)
Ser-O$^{\gamma}$	P	0.267{0.932, 0.000, 0.000}	0.217{0.000, 0.000, 0.928}	0.448{0.000, 0.902, 0.000}
N = 9101	*f**	[2.427, 26.92, 91.99]	[2.432, 31.76, 158.85]	[2.434, 83.55, 176.74]
	σ	(0.036, 4.44, 5.30)	(0.037, 4.80, 6.32)	(0.036, 3.63, 45.27)
Thr-O$^{\gamma 1}$	P	0.404{0.928, 0.000, 0.000}	0.074{0.000, 0.000, 0.957}	0.454{0.000, 0.937, 0.000}
N = 9692	*f**	[2.424, 28.87, 91.70]	[2.438, 27.05, 153.75]	[2.432, 83.87, 188.25]
	σ	(0.032, 3.74, 4.09)	(0.036, 5.04, 5.96)	(0.032, 3.36, 55.43)
Thr-C$^{\gamma 2}$	P	0.451{0.000, 0.929, 0.003}	0.401{0.922, 0.000, 0.000}	0.073{0.001, 0.000, 0.944}
N = 9692	*f**	[2.524, 28.83, 92.80]	[2.521, 29.34, 160.45]	[2.547, 82.31, 162.93]
	σ	(0.033, 4.45, 4.26)	(0.033, 3.44, 4.13)	(0.035, 4.36, 37.18)
Val-C$^{\gamma 1}$	P	0.178{0.957, 0.000, 0.00}	0.690{0.000, 0.000, 0.933}	0.066{0.000, 0.937, 0.000}
N = 12 048	*f**	[2.515, 27.98, 93.44]	[2.525, 29.78, 161.91]	[2.539, 83.20, 157.06]
	σ	(0.030, 4.49, 4.77)	(0.028, 3.57, 3.79)	(0.029, 3.95, 37.04)
Val-C$^{\gamma 2}$	P	0.690{0.000, 0.000, 0.933}	0.068{0.000, 0.975, 0.000}	0.175{0.942, 0.000, 0.000}
N = 12 048	*f**	[2.517, 27.91, 92.42]	[2.532, 25.29, 156.60]	[2.530, 84.08, 147.63]
	σ	(0.029, 4.07, 3.89)	(0.031, 6.35, 7.21)	(0.031, 3.54, 40.99)
His-N$^{\epsilon 2}$	P	0.515{0.935, 0.000, 0.000}	0.300{0.000, 0.000, 0.935}	0.120{0.000, 0.952, 0.000}
N = 3643	*f**	[4.424, 15.49, 81.23]	[4.538, 15.19, 166.98]	[4.520, 78.42, −79.63]
	σ	(0.096, 7.54, 8.84)	(0.92, 7.35, 9.94)	(0.074, 5.43, 51.45)
Phe-C$^{\zeta}$	P	0.504{0.925, 0.000, 0.000}	0.311{0.000, 0.000, 0.920}	0.120{0.000, 0.952, 0.000}
N = 6681	*f**	[5.115, 12.76, 87.24]	[5.132, 17.17, 168.62]	[5.144, 79.01, −40.57]
	σ	(0.066, 5.84, 7.92)	(0.065, 5.18, 7.39)	(0.059, 4.70, 41.97)
Trp-N$^{\epsilon 1}$	P	0.462{0.930, 0.000, 0.000}	0.326{0.000, 0.000, 0.940}	0.145{0.000, 0.953, 0.000}
N = 2533	*f**	[4.520, 10.90, 79.62]	[4.516, 12.48, 160.95]	[4.574, 78.12, −13.88]
	σ	(0.114, 8.01, 9.38)	(0.101, 8.10, 9.52)	(0.081, 5.20, 55.12)
Tyr-O$^{\eta}$	P	0.500{0.928, 0.000, 0.000}	0.313{0.000, 0.000, 0.920}	0.112{0.000, 0.937, 0.000}
N = 5931	*f**	[6.438, 9.46, 79.68]	[6.446, 13.81, 174.60]	[6.480, 73.47, −69.54]
	σ	(0.083, 6.36, 8.62)	(0.085, 5.94, 8.29)	(0.80, 4.89, 36.82)
Leu-C$^{\delta 1}$	P	0.571{0.860, 0.000, 0.000}	0.258{0.001, 0.000, 0.828}	0.039{0.045, 0.000, 0.004}
N = 14 246	*f**	[3.901, 40.14, 103.00]	[3.108, 4.52, 150.49]	[3.105, 1.32, 76.78]
	σ	(0.040, 3.97, 5.37)	(0.079, 3.84, 7.90)	(0.107, 7.45, 13.36)
Leu-C$^{\delta 2}$	P	0.574{0.864, 0.000, 0.000}	0.249{0.000, 0.000, 0.803}	0.049{0.056, 0.011, 0.013}
N = 14 246	*f**	[3.101, 1.47, 108.12]	[3.908, 42.84, 149.24]	[3.903, 36.56, 116.63]
	σ	(0.086, 4.06, 8.10)	(0.039, 3.71, 5.70)	(0.091, 5.33, 9.47)
Gln-O$^{\epsilon 1}$	P	0.465{0.487, 0.434, 0.432}	0.273{0.294, 0.000, 0.295}	0.073{0.118, 0.000, 0.001}
N = 4924	*f**	[4.313, 50.80, 124.46]	[3.396, -5.84, 127.73]	[4.861, 27.06, 101.43]
	σ	(0.131, 10.14, 19.84)	(0.339, 7.66, 19.06)	(0.075, 11.43, 9.67)
Glu-O$^{\epsilon 1}$	P	0.599{0.579, 0.540, 0.652}	0.216{0.255, 0.000, 0.208}	0.080{0.104, 0.222, 0.000}
N = 6910	*f**	[4.309, 52.99, 128.29]	[3.177, -8.01, 123.73]	[3.407, 41.18, 42.61]
	σ	(0.111, 8.60, 16.93)	(0.224, 6.96, 11.48)	(0.158, 8.22, 10.87)

[a] P − proportion of total $\mathbf{A^T}$ groups observed within ellipse (or ellipsoid) enclosed by $\mu \pm 2.33\sigma$; χ_1 rotamer P for ellipse given in order {g^-, g^+, t}. *f** modes [r(C$^{\alpha}$ → $\mathbf{A^T}$), ζ_{ab}, ζ_{in}] and σ standard deviations (r(C$^{\alpha}$ → $\mathbf{A^T}$), ζ_{ab}, ζ_{in}) given respectively.

the location of the other $\mathbf{A^T}$ are given for Asn, Asp, and Leu in the Supporting Information.

Unlike the γ-substituents, these conformational preferences cannot be clearly categorized into simple, superimposable groups. However, several observations may be made. Two conserved clusters falling within region 1 in Figure 4 are, without exception, the most populated regions for all substituents. The conditional probabilities of each branched side chain demonstrate that a given δ-substituent in one of these positions corresponds exactly to the other δ-substituent being located in the other of these positions. Compared with the previous, shorter side chains, more variability is seen in the distance from C$^{\alpha}$ to each of the δ-substituents. Interestingly, however, the conditional probabilities show that one substituent will be significantly further from the C$^{\alpha}$ than the other substituent. Finally, Leu is surprisingly simple, with only three regions demonstrating significant numbers of residues for either of its C$^{\delta}$ substituents, with ∼86% of Leu residues accounted for in these populated regions.

ϵ-Substituent Terminated Residues. Even with a further increase in side-chain length, glutamine, glutamate/glutamic acid, and methionine display a similar complexity in side-chain conformational preference to the δ-substituent terminated residues. Notably, the ϵ-substituent terminated residues no longer display a significant population in region 3 of Figure 4 (the broadly distributed conformations with $\zeta_{ab} >$ 60°). Instead, all of the residues are fairly concentrated within several regions (Table 2, Figures 3 and 4, and Supporting

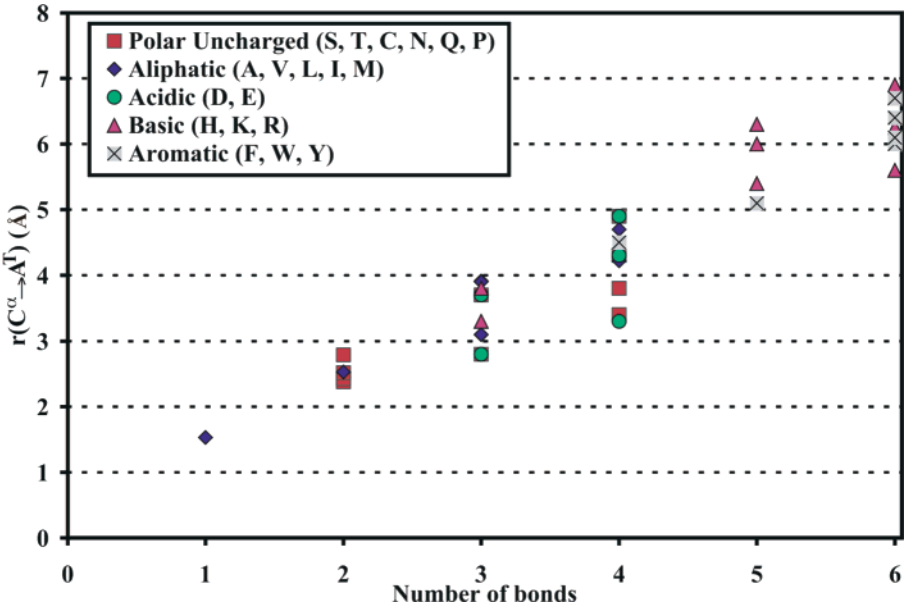

Figure 3. Observed length of side chains to $\mathbf{A^T}$ over given numbers of bonds.

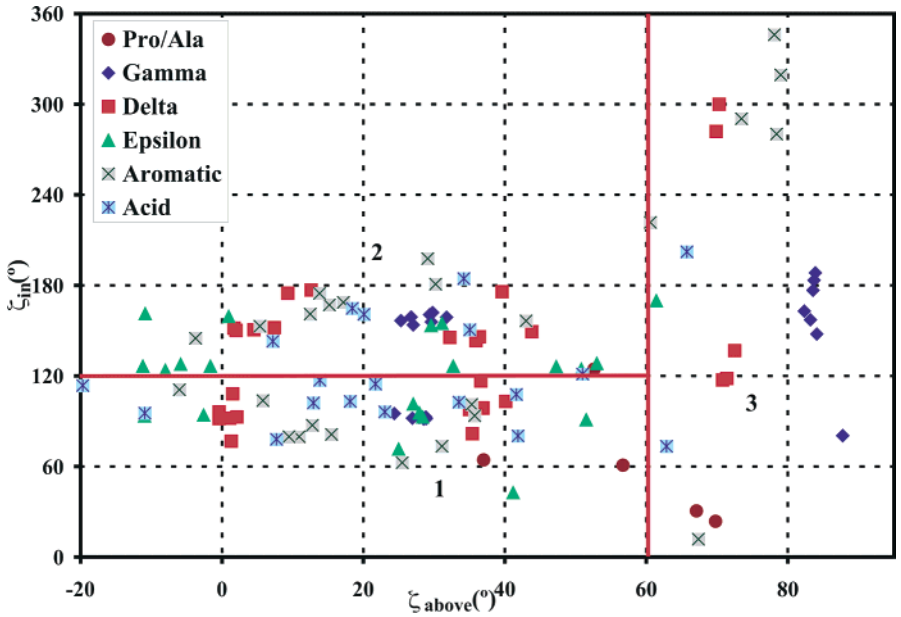

Figure 4. Observed pairs of angular conformations for given $\mathbf{A^T}$ types. Regions are numbered 1−3 for convenience.

Information.) Amazingly, 80−90% of the $O^{\epsilon 1}$ substituents of both Gln and Glu are highly concentrated within three regions of conformational space. These display very similar correlations with their ϵ2-substituents, with statistics provided in the Supporting Information. Unfortunately, only four regions of major population in conformational space were discernible for the C^{ϵ} atom of Met, leaving almost 35% of Met C^{ϵ} atoms out of the statistical calculations. A more elaborate selection scheme is likely needed to categorize further Met C^{ϵ} conformational preferences.

Aromatic Residues. Each of the four aromatic residues, histidine, phenylalanine, tryptophan, and tyrosine, has an $\mathbf{A^T}$ displaying highly similar conformational preference clustered into each one of regions 1−3 of Figure 4. As with the γ-substituents, this corresponds directly to the χ_1 rotamer. For His and Tyr, where two $\mathbf{A^T}$ groups were used (Table 1), only one of the $\mathbf{A^T}$ choices fits into this common category. The two ring N atoms of His were chosen for analysis, since these will share a positive charge in the protonated, basic state. It is the more distal $N^{\epsilon 2}$ atom which fits into the common conformational class. For Trp, $N^{\epsilon 1}$ was chosen as the representative of the five-membered ring, while $C^{\eta 2}$ was chosen as a representative limit of steric bulk for the six-membered ring. The $N^{\epsilon 1}$ conformations fall into the common class. For Phe, the C^{ζ} atom is considered; for Tyr, the O^{η} atom (the C^{ζ} shows the conformations indistinguishable from Phe C^{ζ}.) Note that, unlike the γ-substituents, each of the aromatic $\mathbf{A^T}$ groups has the same order of preference of the three conformations.

Despite the drastically different configurations of the His $N^{\delta 1}$ and Trp $C^{\eta 2}$ atoms, these aromatic substituents show an interesting number of common conformational preferences. Statistical values for the seven most populated regions of each of these $\mathbf{A^T}$ atoms are given in the Supporting Information. Conditional probabilities of finding each of these atoms given the location of either three conformation

Table 3. Average Terminal Atom RMSD between Predicted and Experimentally Determined Coordinates for the Data Set of 372 PDB Chains Containing 116 785 Predictable Non-Glycine Residues[c]

$\mathbf{A^T}$ prediction method[a]	number of structures	number of $\mathbf{A^T}$'s	RMSD $\mu(\sigma)$ − (Å)	best prediction (N)[b]	worst prediction (N)[b]
MP	372	117 685	2.71 (0.35)	1.73 (63)	7.15 (94)
PC	9300	2 919 625	3.04 (0.32)	1.64 (88)	9.98 (326)
PSS	9300	2 919 625	3.00 (0.32)	1.56 (63)	8.44 (94)
PD	9300	2 919 625	3.00 (0.33)	1.67 (63)	9.96 (94)
CHI	9300	2 919 625	1.74 (0.34)	0.56 (56)	9.61 (94)
SCWRL − most likely	369	113 019	2.37 (0.26)	1.53 (108)	3.00 (85)
SCWRL − minimized	360	108 294	1.94 (0.26)	1.33 (101)	2.80 (134)

[a] Methods acronyms refer to those introduced in text − 25 predictions were carried out for each polypeptide with methods PC, PSS, PD, and CHI. The SCWRL "most likely" prediction is that contained in the A-file output by SCWRL; "minimized" is the final output. [b] N − number of $\mathbf{A^T}$ predictions in best or worst predicted structure of entire ensemble of structures predicted by given method. [c] The predictions carried out using the $\mathbf{A^T}$ conformation library and conditional probabilities given herein are presented alongside the RMSD provided with the Dunbrack and Cohen backbone dependent rotamers, as predicted by SCWRL.[24] For comparison, the RMSD values of the best and worst predictions provided by each method are also given.

N^{ϵ} substituent are again given in the Supporting Information. Although the relative populations in each region differ, there is a striking correspondence between the region in which the ϵ-substituent is found and the region in which the His $N^{\delta 1}$ or Trp $C^{\eta 2}$ atom is found.

Basic Residues. The two basic residues, Lys and Arg, may be logically grouped together both by functionality and by the length of their side chains. Both of these residues show a variety of conformational preferences, provided in the Supporting Information and visible in Figures 3 and 4. The Lys residue displays 7 populated conformations for the N^{ζ} atom. Arg displays 5 conformations for the $N^{\eta 1}$ atom and a further 7 conformations for the $N^{\eta 2}$ atom. Conditional probabilities for the two N^{η} atoms of Arg are given in the Supporting Information.

Validation of Representation via Prediction. It should be reiterated that the goal of this work is not to develop a highly sophisticated and optimized predictive methodology—instead, at this stage we opted to carry out a Monte Carlo styled randomized test. We feel that the best way to demonstrate the applicability of the reduced $\mathbf{A^T}$ representations is to generate large numbers of predictions and compare these to the predictive capabilities using the now mature all-atom backbone dependent rotamers. In generating 25 probabilistic based structures with each method for each of the 372 PDB chains used as a test case, a number of predictions should be in good agreement with the experimental structure if the reduced representation is indeed a valid predictive tool. Presumably, with the appropriate algorithm, the same agreement with experiment could be consistently arrived at. We hope that these predictions serve not only to demonstrate the potential reduction of computing time and speed of prediction available using these representations but also to show that the reduced representations are indeed highly capable of providing representative locations for substituents that may be several Ångstroms from a C^{α} using only 3 parameters.

Overall Success of Crude Predictive Methods. Each of the five probabilistic side-chain placement methods was tested using a data set of 372 polypeptide chains (culled from 374 PDB entries), with 116 785 non-Gly residues in the final culled data set. Using the MP method described above, only a single prediction will ever be generated for a given backbone. For the remaining 4 methods, the results reported here are for ensembles of 25 predictions carried out for each structure by each method. Running on a Pentium III 533 MHz Win NT workstation with 128 MB RAM, this particular ensemble of 37 572 protein structures containing ∼1.18 billion predicted atoms took less than 4 h to generate and compare with the parent experimental data.

Table 3 provides RMSD values based on the ensemble of predicted structures for each of the five predictive methods used herein. Notably, the MP method provides a very reasonable RMSD with mean of 2.71 Å over the 372 predicted polypeptide chains, with a total of 116 785 $\mathbf{A^T}$ predictions. The methods PC, PSS, and PD are practically indistinguishable with RMSDs of 3.00−3.04 Å for each set of 372 × 25 predictions. Finally, the CHI method provides an excellent RMSD of only 1.74 Å for the ensemble of 25 predictions for each of the 372 PDB structures. In comparison, using backbone dependent rotamers, the SCWRL predicted structures provide an average RMSD of 2.37 Å using the most likely rotamer for a given backbone configuration and 1.94 Å after the lengthy energy minimization process. Table 3 also demonstrates the RMSD of both the best and worst predicted structures produced in an ensemble of predictions. (One structure, 1K3I, had only 6 residues that were not culled—it is excluded from the "best" prediction category.) A great deal of variability in the agreement of the predictions to the experimental structures is quite apparent for the five methods tested herein. The best predictions from each method are very promising, while the worst predictions are rather far from the experimental and are drastically worse than the SCWRL worst cases.

We tested the possibility of improving side-chain prediction using probabilities dependent upon observed secondary structure or general region of Ramachandran plot. Clearly (Table 3) this did not serve to generally improve prediction. Prediction of the same set of structures using the most likely conformation given the region of Ramachandran plot (i.e. method MP for the *conditional probabilities* of PD) did not improve prediction beyond MP except in the case of four $\mathbf{A^T}$ groups: Cys S^{γ}, Asp $O^{\delta 1}$ and $O^{\delta 2}$, and Met C^{ϵ}. A more elaborate backbone-dependent statistical analysis, along the lines of those developed by Dunbrack with Karplus[47] or Cohen,[32] is probably required to improve these conditional predictions.

At first glance, it is not intuitive that the inclusion of less likely conformations with the probabilistic method PC worsens the mean RMSD for structural prediction over just choosing the most likely conformation (method MP). This is probably due to some combination of unlikely conforma-

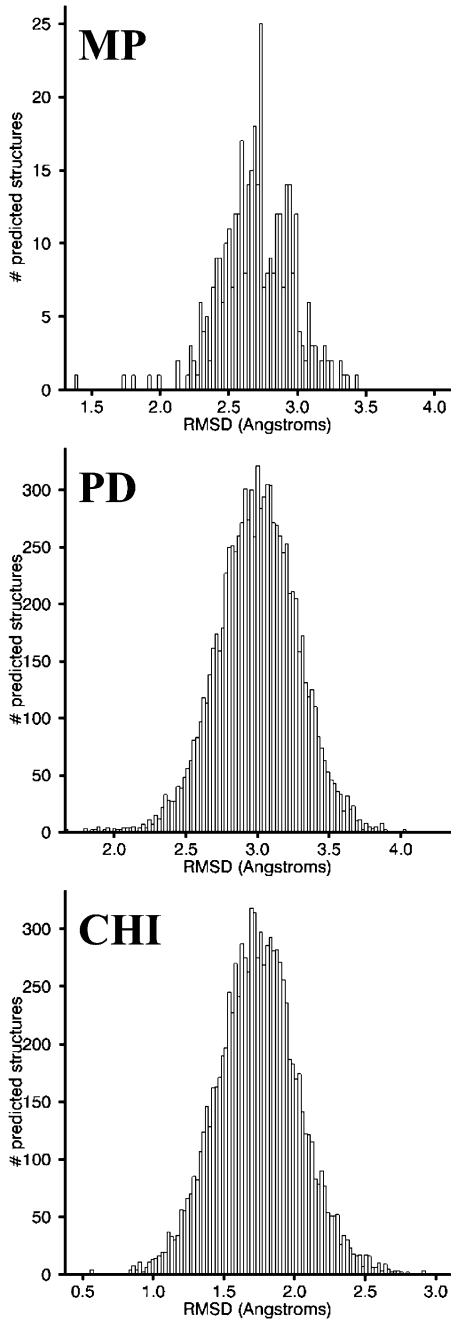

Figure 5. Distributions of RMSD of predicted vs experimentally determined terminal atom positions for entire polypeptide over a set of 372 PDB structures with terminal atoms predicted using PDB backbone coordinates: in the most probable position (MP) and with conditional probabilities for ensembles of 25 predictions for each polypeptide based on backbone dihedral angles (PD) or χ_1 dihedral angles (CHI).

tions being randomly chosen along with the fact that our random selection method, once MP is no longer enforced, takes no account of conditional probabilities for branched substituents of a side chain that may not, therefore, be positioned correctly with relation to each other for a given predicted residue. Note that Bower et al. employ the most likely (backbone dependent all atom) rotamers as their starting point in SCWRL[25]—analogously, our results indicate that the most likely $\mathbf{A^T}$ is probably the best starting point for structural exploration with our reduced representation. Frequency distributions of RMSDs are shown in Figure 5 for the methods MP, PD, and CHI. With both MP and PD, a number of excellent predictions are seen in the range below 2.5 Å RMSD. Furthermore, while the μ RMSD observed for MP is better than that of PD and the other probabilistic methods, Figure 5 demonstrates that the inclusion of less likely conformations do in a large number of cases (note the scale difference) lead to a better overall prediction than simply using the most likely conformation. A direct comparison with all-atom based rotamer libraries is difficult, since these studies usually only compare predictions to a few, select protein structures. The ensemble of SCWRL predictions presented herein is not significantly better than the lower end of this range, however. Therefore, with improved statistics, prediction at a comparable level of accuracy to all-atom rotamers should be achievable.

The deviations observed for prediction of each $\mathbf{A^T}$ group are given in Tables 4 and 5. As would be anticipated, an increase in side-chain length or number of preferred conformations leads to an increase in the deviation between prediction and experiment. Despite the lack of satisfactory backbone dependent statistics, many of the MP predictions are as good as or, in some rare cases, better than those made by SCWRL using backbone dependent most likely all atom rotamers. Conversely, the mean deviations obtained using the random (non-CHI) probabilistic methods are in all cases but one or two worse than those of SCWRL. Subsequent development of a more elaborate backbone dependent set of statistics may therefore be anticipated to provide $\mathbf{A^T}$ predictions on par with all atom rotamers.

Predicting Statistical Surface Charge Density and Topology. As discussed in the Introduction, surface exposed residues of a protein may be anticipated to assume an ensemble of conformations, rather than a single rotameric state (e.g., ref 34). The reduced representation introduced here is one excellent way in which one can predict large statistically relevant ensembles of locations of important sites such as charge carrying atoms. This is possibly the most powerful use of the prediction method we introduce here. To graphically demonstrate this capability, we have carried out over 1000 predictions of $\mathbf{A^T}$ position for each residue of porcine pepsin (PDB entry 5PEP). The density of atoms found in 0.5 Å cubed volume elements, with the Asp $O^{\delta 1}$ and $O^{\delta 2}$, Glu $O^{\epsilon 1}$ and $O^{\epsilon 2}$, Lys N^{ζ}, and Arg $N^{\eta 1}$ and $N^{\eta 2}$ atoms counted as acidic and basic charged atoms as appropriate, is plotted in Figure 6 (b). In comparison to the crystal structure itself (Figure 6(a)), the reduced representation is much more suited to the determination of probability density of charge, especially on the protein surface, as is evident from smearing of each charged $\mathbf{A^T}$ over a wide range of volume in Figure 6(b) as compared to the single point seen for each in (a).

The ensemble of 1015 predictions illustrated in Figure 6(b) indicates that a much wider volume may be sampled by a given charged side chain than is initially indicated by the solved X-ray structure. Since the X-ray structure only represents that rotamer conformation which is favorable within the context of the crystal packing of the particular protein crystal being solved, the reduced representation prediction should be far more indicative of a realistic surface charge density where many rotamers are sampled than a calculation based upon a single side-chain conformation. All-atom prediction approaches are also typically most suited to

Table 4. Average RMSD (Top Row for Each $\mathbf{A^T}$) and Mean (μ) and Standard Deviation (σ) (Bottom Row for Each $\mathbf{A^T}$) between Predicted and Experimentally Determined $\mathbf{A^T}$ Position for the Test Set of 372 Polypeptide Structures from the PDB[a]

		RMSD (top) & $\mu(\sigma)$ (bottom) from experiment (Å) − methods					SCWRL	
residue ($\mathbf{A^T}$)	N	MP	PC	PSS	PD	CHI	likely	min
Ala (C^{β})	7690	0.133	0.0805					
		0.048 (0.124)	0.081 (0.071)					
Pro (C^{γ})	4013	1.01	1.00	0.96	0.96	0.30	0.57	0.55
		0.78 (0.64)	0.78 (0.62)	0.73 (0.61)	0.74(0.61)	0.20 (0.21)	0.38 (0.42)	0.37 (0.41)
Pro (C^{γ} − cis)	246	0.55	0.66	0.66	0.66	0.23	0.46	0.45
		0.33 (0.44)	0.43 (0.51)	0.44 (0.51)	0.43 (0.51)	0.17 (0.16)	0.31 (0.34)	O.30 (0.33)
Cys (S^{γ})	801	1.99	2.21	2.13	2.11	0.32	1.55	1.12
		1.51 (1.30)	1.82 (1.25)	1.72 (1.27)	1.68 (1.28)	0.27 (0.18)	1.02 (1.17)	0.63 (0.92)
Cystine (S^{γ})	431	1.57	1.89	1.87	1.86	0.24	0.90	0.87
		1.08 (1.14)	1.43 (1.21)	1.42 (1.21)	1.42 (1.21)	0.34 (0.26)	0.52 (0.73)	0.51 (0.71)
Ser (O^{γ})	4741	1.65	1.90	1.88	1.87	0.28	1.36	1.31
		1.26 (1.06)	1.54 (1.15)	1.51 (1.10)	1.49 (1.14)	0.22 (0.17)	0.92 (1.00)	0.87 (0.98)
Thr ($O^{\gamma 1}$)	5095	1.75	1.81	1.75	1.80	0.29	0.87	0.77
		1.36 (1.10)	1.45 (1.08)	1.37 (1.09)	1.44 (1.08)	0.23 (0.18)	0.48 (0.73)	0.41 (0.65)
Thr ($C^{\gamma 2}$)	5095	1.73	1.84	1.78	1.83	0.26	0.91	0.80
		1.32 (1.12)	1.46 (1.12)	1.37 (1.13)	1.44 (1.12)	0.19 (0.17)	0.49 (0.77)	0.42 (0.69)
Val ($C^{\gamma 1}$)	6415	1.22	1.56	1.53	1.55	0.22	0.84	0.71
		0.72 (0.99)	1.08 (1.12)	1.06 (1.12)	1.07 (1.12)	0.17 (0.14)	0.42 (0.73)	0.34 (0.62)
Val ($C^{\gamma 2}$)	6415	1.23	1.56	1.54	1.56	0.23	0.83	0.70
		0.74 (0.98)	1.10 (1.12)	1.06 (1.11)	1.09 (1.11)	0.19 (0.14)	0.43 (0.71)	0.36 (0.61)
Asn ($O^{\delta 1}$)	3565	2.62	3.00	2.99	2.86	1.31	2.33	2.05
		2.08 (1.58)	2.58 (1.53)	2.56 (1.54)	2.41 (1.54)	1.05 (0.78)	1.98 (1.23)	1.73 (1.11)
Asn ($N^{\delta 2}$)	3565	2.19	2.49	2.48	2.44	1.43	2.48	2.22
		1.88 (1.12)	2.23 (1.13)	2.21 (1.13)	2.15 (1.15)	1.22 (0.75)	2.12 (1.29)	1.87 (1.19)
Asp ($O^{\delta 1}$)	4786	2.67	3.02	2.97	2.86	0.73	2.46	2.27
		2.03 (1.74)	2.50 (1.70)	2.44 (1.71)	2.15 (1.74)	0.56 (0.47)	2.30 (0.87)	2.09 (0.88)
Asp ($O^{\delta 2}$)	4786	1.95	2.14	2.11	2.00	0.82	2.66	2.40
		1.58 (1.13)	1.84 (1.08)	1.79 (1.11)	1.65 (1.13)	0.66 (0.49)	2.47 (0.97)	2.21 (0.94)
Ile ($C^{\delta 1}$)	4752	1.67	2.14	2.12	2.13	1.33	1.46	1.18
		1.21 (1.15)	1.70 (1.30)	1.66 (1.31)	1.69 (1.30)	0.93 (0.97)	0.97 (1.09)	0.74 (0.92)
Leu ($C^{\delta 1}$)	7730	1.95	2.31	2.27	2.30	1.08	1.75	1.17
		1.36 (1.40)	1.76 (1.48)	1.73 (1.48)	1.76 (1.49)	0.62 (0.88)	1.16 (1.30)	0.70 (0.93)
Leu ($C^{\delta 2}$)	7730	2.04	2.28	2.24	2.27	0.95	1.73	1.16
		1.48 (1.40)	1.78 (1.43)	1.74 (1.43)	1.77 (1.43)	0.64 (0.70)	1.19 (1.26)	0.73 (0.89)

[a] Predictions are compared for the five $\mathbf{A^T}$ conformational preference methods given here and for the most likely backbone dependent rotamer or the energy minimized prediction from SCWRL.[24] Note that methods PC, PSS, PD, and CHI are ensembles of 25 x N predictions. See Table 5 for ϵ-substituent terminated, aromatic, and basic residues.

finding a single structure such as that in Figure 6(a), rather than to generating a large ensemble of possible structures such as those that may in fact be representative of a protein in solution.[34] Furthermore, the idea of a more plastic topology arising from statistical sampling of all likely side-chain conformations, instead of a single rotamer, should provide a more reliable representation of protein surface topology.

Other Proposed Applications. In some instances the side-chain length alone could be extremely useful as a first approximation in modeling, while in others, the preferred conformation including the length would be desired. In the static instance, if a protein backbone is either predicted or experimentally determined, a reasonable prediction of the locations of chemically relevant side-chain moieties may then be readily carried out using the reduced parameter set. In dynamic use, folding simulations may be envisioned in which side-chain conformations sample the known stable forms given by the parameter set herein. Finally, in examination of experimental data, deviations from the conformations herein may be used to pinpoint interesting regions of a protein structure or as a guide to assignment of terminal atoms in carboxylic acids or amide groups.

Simulations requiring side-chain configurations with a reduction in degrees of freedom could very readily be carried out with the parameters given herein. Our data are highly complementary to the various reduced representations and coarse-grained approaches reviewed in the Introduction but provides a rather different starting point for such studies. The CHI prediction method looks promising for high-throughput structural studies. If a protein backbone is determined experimentally along with even some χ_1 rotamer values (to a fairly rough accuracy), the statistical parameters given here could provide a very satisfactory prediction of the terminal and, functional, side-chain atom locations. This could be combined with appropriate high-throughput NMR methods, for example, to screen a large set of proteins for a desired structural character. Of course, using a similar conditional probability with all-atom rotamers would likewise increase their performance significantly. Therefore, it would be up to the user to weigh the advantages of decreased computational time vs atomic detail.

The power of such a reduced representation should be very notable in combinatorial applications. In particular, the searching of large sets of primary sequence dependent topological and chemical configuration for a desired structural and chemical motif could be readily carried out prior to protein engineering. For example, if a particular topography is defined, and a polypeptide framework laid out (e.g., a set of amphipathic α-helices or of parallel β-sheets), the surface exposed portion could be combinatorially scanned

Table 5. Average RMSD (Top Row for Each $\mathbf{A}^{\mathbf{T}}$) and Mean (μ) and Standard Deviation (σ) (Bottom Row for Each $\mathbf{A}^{\mathbf{T}}$) between Predicted and Experimentally Determined $\mathbf{A}^{\mathbf{T}}$ Position for the Test Set of 372 Polypeptide Structures from the PDB[a]

		RMSD (top) & $\mu(\sigma)$ (bottom) from experiment (Å) − methods					SCWRL	
residue ($\mathbf{A}^{\mathbf{T}}$)	N	MP	PC	PSS	PD	CHI	likely	min
Gln($O^{\epsilon 1}$)	2791	2.65	2.93	2.92	2.92	2.81	2.68	2.41
		2.27 (1.37)	2.56 (1.42)	2.55 (1.42)	2.55 (1.43)	2.45 (1.38)	2.29 (1.39)	1.99 (1.36)
Gln($N^{\epsilon 2}$)	2791	3.26	3.59	3.56	3.57	2.74	3.02	2.66
		2.88 (1.53)	3.24 (1.53)	3.20 (1.53)	3.22 (1.54)	2.45 (1.25)	2.62 (1.50)	2.21 (1.48)
Glu ($O^{\epsilon 1}$)	4067	2.50	2.90	2.86	2.88	2.81	2.43	2.24
		2.01 (1.48)	2.43 (1.59)	2.38 (1.59)	2.41 (1.58)	2.34 (1.56)	1.94 (1.47)	1.72 (1.43)
Glu ($O^{\epsilon 2}$)	4067	3.21	3.69	3.64	3.67	2.40	3.00	2.68
		2.72 (1.71)	3.23 (1.79)	3.18 (1.78)	3.20 (1.78)	2.00 (1.33)	2.49 (1.67)	2.13 (1.63)
Met (C^{ϵ})	1448	3.83	3.56	3.56	3.46	3.19	3.14	2.23
		3.07 (2.30)	2.99 (1.68)	2.95 (1.70)	2.97 (1.78)	2.68 (1.70)	2.74 (1.52)	1.72 (1.42)
His ($N^{\delta 1}$)	2066	2.70	3.08	3.05	3.06	1.62	2.48	2.06
		2.29 (1.43)	2.73 (1.42)	2.69 (1.42)	2.71 (1.42)	1.40 (0.83)	2.14 (1.24)	1.76 (1.07)
His ($N^{\epsilon 2}$)	2066	3.97	4.44	4.40	4.41	0.96	3.24	2.29
		3.07 (2.51)	3.68 (2.49)	3.64 (2.50)	3.63 (2.50)	0.80 (0.52)	2.30 (2.27)	1.51 (1.71)
Phe (C^{ζ})	3561	4.51	5.04	4.88	4.94	0.83	3.30	2.01
		3.41 (2.95)	4.08 (2.94)	3.88 (2.96)	3.96 (2.95)	0.70 (0.45)	2.10 (2.55)	1.12 (1.67)
Trp ($N^{\epsilon 1}$)	1424	4.25	4.59	4.47	4.55	1.11	3.31	2.26
		3.39 (2.56)	3.89 (2.43)	3.74 (2.45)	3.84 (2.44)	0.94 (0.59)	2.41 (2.28)	1.45 (1.73)
Trp ($C^{\eta 2}$)	1424	5.81	6.45	6.34	6.38	4.10	5.09	3.84
		4.83 (3.22)	5.78 (2.84)	5.65 (2.91)	5.69 (2.88)	3.47 (2.19)	4.12 (2.99)	2.76 (2.67)
Tyr (O^{η})	3126	6.12	6.89	6.72	6.81	1.26	4.48	2.66
		4.61 (4.01)	5.63 (3.98)	5.36 (4.01)	5.53 (3.99)	1.06 (0.69)	2.91 (3.41)	1.54 (2.17)
Arg ($N^{\eta 1}$)	3340	4.01	4.54	4.54	4.52	4.18	4.64	4.10
		3.57 (1.84)	4.08 (2.00)	4.07 (2.00)	4.05 (2.01)	3.74 (1.85)	4.29 (1.76)	3.73 (1.70)
Arg ($N^{\eta 2}$)	3340	4.99	5.52	5.52	5.52	5.03	4.39	3.94
		4.44 (2.29)	5.00 (2.36)	4.99 (2.35)	5.00 (2.35)	4.53 (2.24)	4.12 (1.51)	3.64 (1.50)
Lys (N^{ζ})	3418	3.74	4.23	4.16	4.19	3.19	3.56	3.05
		3.25 (1.85)	3.77 (1.87)	3.73 (1.87)	3.75 (1.86)	2.78 (1.56)	3.05 (1.83)	2.51 (1.73)

[a] Predictions are compared for the five $\mathbf{A}^{\mathbf{T}}$ conformational preference methods given here and for the most likely backbone dependent rotamer or the energy minimized prediction from SCWRL.[24] Note that methods PC, PSS, PD, and CHI are ensembles of 25 x N predictions. See Table 4 for Ala, γ-, and δ-substituent terminated residues.

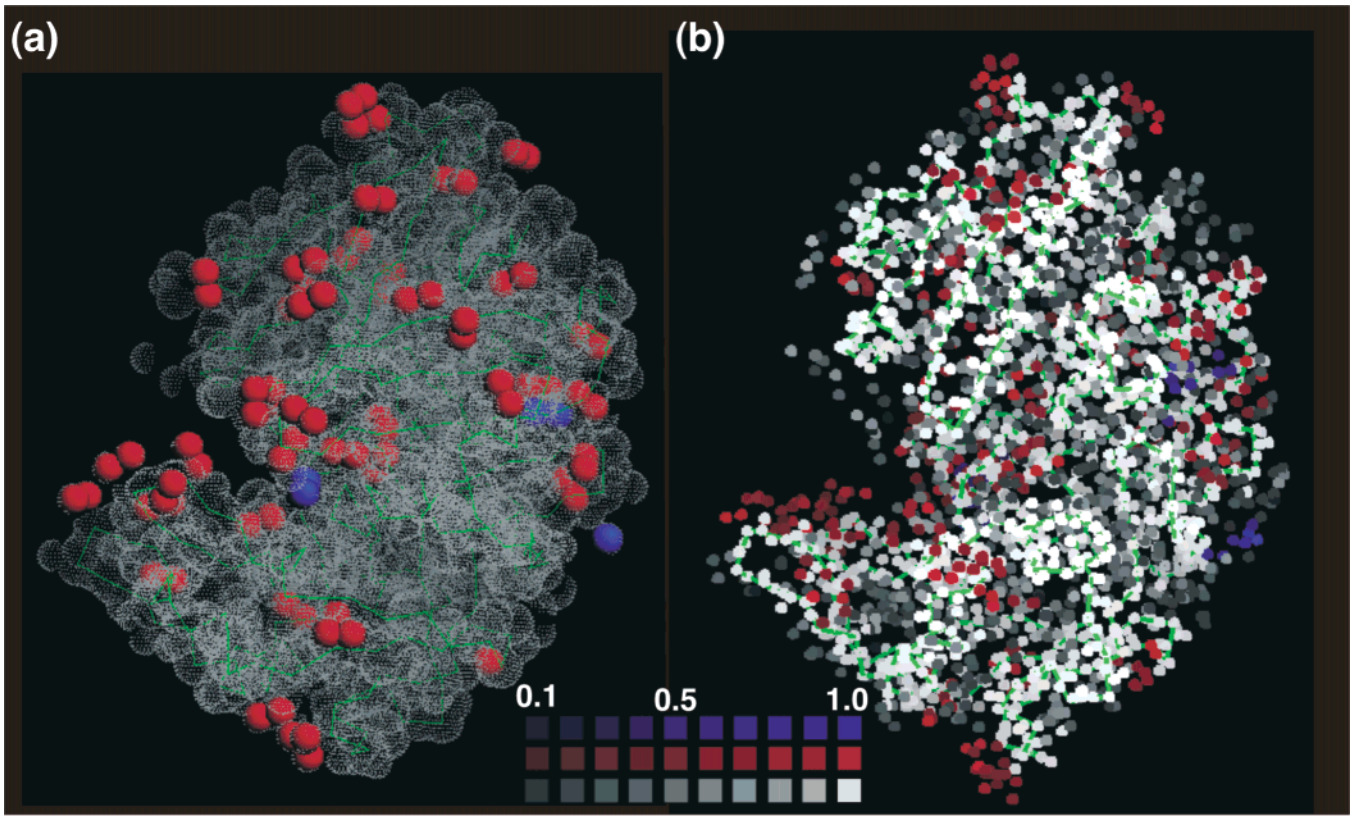

Figure 6. (a) Crystal structure of porcine pepsin (PDB entry 5PEP); backbone C^{α} trace in green, all $\mathbf{A}^{\mathbf{T}}$ atoms (Table 1) shown with gray electrostatic shell dots, Asp $O^{\delta 1}$ & $O^{\delta 2}$ Glu $O^{\epsilon 1}$ & $O^{\epsilon 2}$ colored red and Lys N^{ζ} Arg $N^{\eta 1}$ & $N^{\eta 2}$ colored blue and spacefilled (drawn in RasMol,[48] www.openrasmol.org). (b) Normalized numbers of $\mathbf{A}^{\mathbf{T}}$ predictions in cubic volume elements with side 0.5 Å; backbone C^{α} trace in green, atom density shaded in gray and acidic and basic atom density shaded in red and blue respectively (see Methods section III for details.) Scale bars correspond to normalized density of atoms in volume elements of (b).

through all combinations of appropriate residues in order to engineer the given topological and chemical features.

As mentioned in the Introduction, an experimental technique which may readily benefit from the availability of fast, accurate topographical calculation is scanning probe microscopy (SPM). Approaches such as that of Todd et al. rely upon prediction of topography and correlation to observed Ångstrom to nanometer scale measurements of protein structure.[2] The reduced representation herein provides a rapid way to calculate topographical information for systems which may not otherwise be suitable for molecular modeling. Further, the ability to describe topology in a probabilistic manner (e.g. Figure 6) would allow an SPM image to be analyzed in a much different context than that provided by a single structure. Since SPM cannot generally reach single side-chain resolution, searching an ensemble of possible topologies convoluted to the resolution observable with a given SPM probe may provide a more fruitful match to real topographic data.

CONCLUSIONS

Using a large, culled data set of high-resolution protein structures, we have determined conformational specifications in a generalized parameter representation for the terminal atoms of each of the non-Gly residues. A geometric framework has also been developed to allow the incorporation of these parametrized conformations into any polypeptide backbone of interest. These preferences range in simplicity from Ala (with a single preferred conformation) and Pro (with two preferred conformations) to the Trp $C^{\eta 2}$, Arg $N^{\eta 2}$, and Lys N^{ζ} atoms (with seven preferred conformations in clusters containing at least 5% of the $\mathbf{A^T}$'s). These conformations alone should prove valuable as a readily accessible physical representation for steric extent or the location of charged moieties of side chains relative to a peptide backbone.

The representative nature of these conformational preferences was tested by comparing a large ensemble of predicted structures made using the statistical side-chain representations to a nonredundant set of input protein structures. Despite the relative simplicity of the $\mathbf{A^T}$ representation, the population or backbone based conformational preferences are shown to provide side-chain predictions with mean RMSD of 2.73−3.04 Å over the entire predicted structure comparing the experimental versus the predicted $\mathbf{A^T}$ location. Even with the crude probabilistic positioning methods herein, the distribution of RMSD values shows success in terminal atom positioning that is equal to or better than that provided by conventional backbone dependent rotamers (using SCWRL without energy minimization) for approximately 10% of the ensemble of predictions. With more elaborate probabilistic selection procedures and minimization algorithms, the predictive capability of this reduced representation should be routinely equal to conventional rotamer libraries. This would allow highly efficient computation of predicted protein conformation. In the case where a χ_1 rotamer could be experimentally determined, protein side chains could be readily predicted to a very high accuracy using only N, C^{α}, and C′ atom locations—the mean RMSD of $\mathbf{A^T}$ prediction using this method was 1.74 Å.

To demonstrate the applicability of the reduced representation, a volume density plot is shown alongside its corresponding high-resolution crystal structure. Atom locations on the protein surface are generally observed to smear over space, rather than localize to a single Cartesian coordinate. This corresponds more closely to experimental observation of proteins in solution and provides a starting point for a more representative manner of carrying out surface charge density and topology calculations for proteins. In contrast to a single local minima structure that may be produced with a typical all-atom approach, the reduced representation allows for massive ensembles of probable structures to be efficiently generated for computational use as desired.

ACKNOWLEDGMENT

This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). J.K.R. is grateful for an NSERC Postgraduate Scholarship and to the Government of Ontario for an Ontario Graduate Scholarship; since moving to University of Alberta, J.K.R. has been supported by PENCE and by an Alberta Heritage Foundation for Medical Research Fellowship; J.K.R. is grateful to PENCE for the use of computational resources and to Darren Anderson at U of T and Steffen Graether at PENCE for helpful discussions.

Supporting Information Available: Numbers of residues culled by each selection procedure during the culling process, a full derivation of the manner by which $\mathbf{A^T}$ may be fixed in Cartesian coordinates given backbone coordinates and reduced representation statistics, statistical data for all $\mathbf{A^T}$ groups with 4−7 conformations, conditional probabilities of finding branched substituents, and a detailed list of PDB files used in compiling reduced representation statistics. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Creighton, T. *Proteins − Structures and Molecular Properties*, 2nd ed.; W. H. Freeman and Company: New York, 1993.

(2) Todd, B. A.; Rammohan, J.; Eppell, S. J. Connecting nanoscale images of proteins with their genetic sequences. *Biophys. J.* **2003**, *84*, 3892−3991.

(3) Taylor, W. R.; Thornton, J. M.; Turnell, W. G. An ellipsoidal approximation of protein shape. *J. Mol. Graphics* **1983**, *1*, 30−38.

(4) Prabhakaran, M.; Ponnuswamy, P. K. Shape and surface-features of globular-proteins. *Macromolecules* **1982**, *15*, 314−320.

(5) Thornton, J. M.; Edwards, M. S.; Taylor, W. R.; Barlow, D. J. Location of 'continuous' antigenic determinants in the protruding regions of proteins. *EMBO J.* **1986**, *5*, 409−413.

(6) Dill, K. A. Polymer principles and protein folding. *Protein Sci.* **1999**, *8*, 1166−1180.

(7) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253*, 694−698.

(8) Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **1976**, *104*, 59−107.

(9) Wilson, C.; Doniach, S. A computer model to dynamically simulate protein folding: studies with crambin. *Proteins: Struct., Funct., Genet.* **1989**, *6*, 193−209.

(10) Sun, S. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci.* **1993**, *2*, 762−785.

(11) Head-Gordon, T.; Brooks, C. L., III Virtual rigid body dynamics. *Biopolymers* **1991**, *31*, 77−100.

(12) Herzyk, P.; Hubbard, R. E. A reduced representation of proteins for use in restraint satisfaction calculations. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 310−324.

(13) Keskin, O.; Bahar, I. Packing of side chains in low-resolution models for proteins. *Folding Des.* **1998**, *3*, 469−479.

(14) Wallqvist, A.; Ullner, M. A simplified amino acid potential for use in structure predictions of proteins. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 267−280.

(15) Smith, A. V.; Hall, C. K. α-Helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 344−360.
(16) Gibbs, N.; Clarke, A. R.; Sessions, R. B. Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 186−202.
(17) Zacharias, M. Protein−protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* **2003**, *12*, 1271−1282.
(18) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharn, V. Sterechemistry of polypeptide chain configurations. *J. Mol. Biol.* **1963**, *7*, 95−99.
(19) Chandrasekaran, R.; Ramachandran, G. N. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformation in proteins. *Int. J. Protein Res.* **1970**, *2*, 223−233.
(20) Ponder, J. W.; Richards, F. M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **1987**, *193*, 775−791.
(21) Chakrabarti, P.; Pal, D. The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.* **2001**, *76*, 1−102.
(22) Dunbrack, R. L., Jr. Rotamer Libraries in the 21st Century. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431.
(23) Philippopoulos, M.; Lim, C. Exploring the dynamic information content of a protein NMR structure: Comparison of a molecular dynamics simulation with the NMR and X-ray structures of *Escherichia coli* ribonuclease HI. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 87−110.
(24) Kleywegt, G. J. Validation of protein crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 249−265.
(25) Bower, M. J.; Cohen, F. E.; Dunbrack, R. L., Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J. Mol. Biol.* **1997**, *267*, 1268−1282.
(26) Al-Lazikani, B.; Jung, J.; Xiang, Z. X.; Honig, B. Protein structure prediction. *Curr. Opin. Chem. Biol.* **2001**, *5*, 51−56.
(27) Pokala, N.; Handel, T. M. Review: protein design−where we were, where we are, where we're going. *J. Struct. Biol.* **2001**, *134*, 269−281.
(28) Hermanson, G. *Bioconjugate Techniques*; Academic Press: New York, 1996.
(29) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(30) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(31) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The penultimate rotamer library. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 389−408.
(32) Dunbrack, R. L., Jr.; Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **1997**, *6*, 1661−1681.
(33) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **1999**, *285*, 1711−1733.
(34) Daley, M. E.; Sykes, B. D. The role of side chain conformational flexibility in surface recognition by *Tenebrio molitor* antifreeze protein. *Protein Sci.* **2003**, *12*, 1323−1331.
(35) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(36) Hobohm, U.; Scharf, M.; Schneider, R.; Sander, C. Selection of representative protein data sets. *Protein Sci.* **1992**, *1*, 409−417.
(37) Hobohm, U.; Sander, C. Enlarged representative set of protein structures. *Protein Sci.* **1994**, *3*, 522−524.
(38) Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M. PROCHECK − a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **1993**, *26*, 283−291.
(39) Markley, J. L.; Bax, A.; Arata, Y.; Hilbers, C. W.; Kaptein, R.; Sykes, B. D.; Wright, P. E.; Wuthrich, K. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J. Mol. Biol.* **1998**, *280*, 933−952.
(40) Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.
(41) Morris, A. L.; Macarthur, M. W.; Hutchinson, E. G.; Thornton, J. M. Stereochemical Quality of Protein-Structure Coordinates. *Proteins: Struct., Funct., Genet.* **1992**, *12*, 345−364.
(42) Drenth, J. *Principles of Protein X-ray Crystallography*, 2nd ed.; Springer-Verlag: New York, 1994.
(43) Jabs, A.; Weiss, M. S.; Hilgenfeld, R. Nonproline cis peptide bonds in proteins. *J. Mol. Biol.* **1999**, *286*, 291−304.
(44) Li, W.; Jaroszewski, L.; Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **2001**, *17*, 282−283.
(45) Cooper, J. B.; Khan, G.; Taylor, G.; Tickle, I. J.; Blundell, T. L. X-ray analyses of aspartic proteinases. II. Three-dimensional structure of the hexagonal crystal form of porcine pepsin at 2.3 A resolution. *J. Mol. Biol.* **1990**, *214*, 199−222.
(46) Engh, R. A.; Huber, R. Accurate bond and angle parameters for X-ray protein-structure refinement. *Acta Crystallogr. A* **1991**, *47*, 392−400.
(47) Dunbrack, R. L., Jr.; Karplus, M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **1993**, *230*, 543−574.
(48) Sayle, R. A.; Milner-White, E. J. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **1995**, *20*, 374.

CI034177Z