Prediction and Evaluation of Side-Chain Conformations for Protein Backbone Structures

Peter S. Shenkin,[1] Hany Farid,[2] and Jacquelyn S. Fetrow[3]
[1]Department of Chemistry, Columbia University, New York; [2]Department of Computer Science, and [3]Department of Biological Sciences, University at Albany, SUNY, Albany, New York

ABSTRACT **A common approach to protein modeling is to propose a backbone structure based on homology or threading and then to attempt to build side chains onto this backbone. A fast algorithm using the simple criteria of atomic overlap and overall rotamer probability is proposed for this purpose. The method was first tested in the context of exhaustive searches of side chain configuration space in protein cores and was then applied to all side chains in 49 proteins of known structure, using simulated annealing to sample space. The latter procedure obtains the correct rotamer for 57% and the correct χ_1 value for 74% of the 6751 residues in the sample. When low-temperature Monte-Carlo simulations are initiated from the results of the simulated-annealing processes, consensus configurations are obtained which exhibit slightly more accurate predictions. The Monte-Carlo procedure also allows converged side chain entropies to be calculated for all residues. These prove to be accurate indicators of prediction reliability. For example, the correct rotamer is obtained for 79% and the correct χ_1 value is obtained for 84% of the half of the sample residues exhibiting the lowest entropies. Side chain entropy and predictability are nearly completely uncorrelated with solvent-accessible area. Some precedents for and implications of this observation are discussed.**

Key words: **rotamer prediction, side chain entropy, solvent accessibility, simulated annealing, Monte-Carlo**

INTRODUCTION

How an amino acid sequence encodes a protein's three-dimensional structure is a major unsolved problem in molecular biology. Currently, x-ray crystallography and nuclear magnetic resonance spectroscopy are the major methods of determining structure. These experiments are time-consuming and require elaborate equipment to carry out and extensive expertise to interpret. In addition, structure determination by crystallography cannot be performed on all proteins because not all proteins can be crystallized. Protein and DNA sequencing, on the other hand, are much less laborious procedures, which are carried out on a routine basis in thousands of laboratories equipped with the basic tools of molecular biology. Only about 3500 protein structures have been determined, while tens of thousands of sequences are known.

The utility of the information contained in protein- and nucleic-acid-sequence databases would be greatly enhanced if one could predict protein three-dimensional structure from sequence alone. Much work on structure prediction has involved prediction of protein secondary structure[1–6] and some work has done on packing such structures into a three-dimensional tertiary structure .[7,8] Other successful attempts to predict tertiary structure are based on model-building techniques that make use of observed homology.[9,10] This method is useful only when the sequence of unknown structure is found to be homologous to that of a protein of known structure. Loops and other regions exhibiting irregular secondary structure are usually the least conserved structurally, and can be modeled using a variety of techniques[11–21]; finally, side chains are added and the whole structure is subjected to energy minimization.[10]

Other methods of protein tertiary-structure prediction are based on the "inverse folding" concept,[22–25] or alignment of sequence-to-structure (reviewed in references 26 and 27). In this approach, the sequence of the unknown protein is "threaded" onto known backbone structures and a contact or other energy function is calculated for side chain–side chain and sometimes side chain–backbone interactions. Such an energy is determined for each "thread" or sequence-to-structure alignment. The backbone-threading combination is selected that gives the lowest energy. Like the homology-based methods, inverse folding methods give only backbone predictions, and do not build side chain or, in many cases, even loop backbone conformations.

Received May 31, 1996; revision accepted June 4, 1996.

Address reprint requests to Peter S. Shenkin, Department of Chemistry, Columbia University, 3000 Broadway, Mailcode 3153, New York, NY 10027.

Dr. Farid's present address is GRASP Laboratory, Department of Computer Science, University of Pennsylvania, Philadelphia, PA 19103–5544.

A recently published prediction algorithm uses a Monte Carlo strategy to iteratively search for local interactions in the protein backbone.[28] Local interactions are hierarchically accrued to obtain the backbone structure of small domains. Although successful at predicting backbone topologies, this folding algorithm does not build complete side chain configurations.

The focus of this paper is a fast, automated side chain conformation prediction algorithm that can be applied to backbone structures built using homology modeling or inverse folding or other computational methods. Existing algorithms for this purpose can be divided into three groups: knowledge-, database-, and rule-based methods[29–32]; local optimization of one or a small group of side chain structures at one time[33–38]; and global, but not necessarily exhaustive, optimization of all side chains.[38–44] In some of these methods, the search space can be limited to a rotamer library, as suggested by Ponder and Richards[45] in another context. These methods will be discussed and compared to the current method in the Discussion section of this paper.

We present a simple, automated, global optimization procedure for side chain prediction most similar to those of Holm and Sander[40] and Lee and Subbiah.[39] It begins with the demonstration by Ponder and Richards[45] that each side chain conformation can be represented by a selection from among a small number of discrete rotamers. We show that this is a reasonable approximation of nativelike structure in protein cores and in entire proteins, especially if the resulting structures will be further refined by energy minimization, as is often done in homology modeling and inverse folding problems. To compute the best set of side chain conformations, the scoring criteria of number of atomic overlaps and *side chain* probability from the rotamer library are used. We show that these simple criteria are adequate to determine side chain conformations in the hydrophobic cores of several proteins. To sample the conformation space for all side chains in a given protein, rotamer configurations are searched using a simulated annealing algorithm.[46] The "best" structure found by this search strategy is similar to the native structure in most positions, suggesting that our search strategy adequately explores conformational space and that our scoring criterion strongly selects for nativelike conformations. The algorithm can be applied to a complete protein containing 100 moving residues in about 2 CPU minutes on a MIPS R4400/150 processor (Silicon Graphics Iris Indigo).

Low-temperature Monte-Carlo simulations[47] initiated from the results of the simulated-annealing runs provide both a consensus configuration and an assessment of prediction accuracy for each residue. The low temperature guarantees that only low-energy configurations will be visited. From the fraction of times each rotamer is sampled, an entropy is calculated at each position. The side chain conformation prediction accuracy is highest at positions of low entropy (that is, at positions where the greatest consensus occurs), and becomes successively worse as the entropy increases. Thus, this method associates a reliability factor or confidence level with the prediction of each side chain. This allows the user to pay attention to only the most reliable predictions. Obtaining convergent entropies raises the simulation time for a protein containing 100 moving residues to about 15 minutes on a MIPS R4400/150 processor.

It is intuitive to expect surface residues to exhibit greater conformational freedom (greater entropy) than buried residues. However, this tendency is observed only very weakly: Surface residues are nearly as predictable, on average, as buried residues, and exhibit very similar entropy distributions. Some implications of this are discussed.

In addition to providing entropies of rotamer appearance, the low-temperature Monte-Carlo procedure gives a consensus configuration, which we take to be the most frequently visited rotamer at each position. The consensus configuration proves to be a slightly better predictor of the native side chain configuration, on average, than is the configuration resulting from any single simulated-annealing run.

This algorithm has been used in conjunction with backbone model-building techniques for fast and accurate side chain structure prediction.[48–50] The method presented here is global, rather than confined to use on a single side chain or a selected group of side chains, such as the buried residues. Its chief benefits are that it can be quickly applied to all residues of a protein, it provides an internally generated measure of its own prediction accuracy, and its efficacy has been demonstrated by evaluation over a large test-set: 49 proteins of diverse structure, comprising 6751 variable side chains. Aspects of this work were presented earlier.[51,52]

METHODS

Calculation of the Best Rotamer Configuration

The term "side chain configuration" is used here to mean a list of rotamers, one per side chain position. Given a rotamer library, the best rotamer configuration (BRC) of a protein or part of a protein of known structure is that configuration whose side chain atoms exhibit the smallest atomic root mean square (RMS) distance from the corresponding atoms of the native structure. The BRC is the best approximation that can be made to a known side chain configuration using the rotamers from the library.

Alanine and glycine do not have side chains, and therefore do not appear in the list of rotamers defining the BRC. In the rotamer library we use, the pyrrolidine ring of proline is regarded as fixed, so proline does not appear either. S—S bonded cysteines were also removed from the list, as were prosthetic groups and residues bonded to them. Finally, residues reported as ASX and GLX were excluded, as were residues whose numbers differed only by a suffix (such as VAL-62 and ASN-62A of elastase, PDB entry[53] 1hne), and only the first chains of multiple-subunit proteins were studied. Our programs are capable of dealing with these situations, but not all the programs used to calculate solvent-accessible surface handled these variations correctly. It was our initial intent to report results on the common subset of residues that could be handled by all the programs in use; however, certain of the above exceptions became apparent only during the subsequent analysis. Thus, surface-area analysis could not be applied to all residues for which we give simulation data. The simulation dataset includes 6751 residues; when we report surface areas, a subset of 6535 residues is utilized.

In determining the BRC, the following equation is used to compute the RMS interatomic displacement from the known native side chain conformation:

$$D_{\mathrm{rms,m}} = \min_{j=\mathrm{rotamers}} \left(\sum_{\substack{i=\mathrm{moving}\\ \mathrm{atoms}}} \frac{D^2_{i,\mathrm{nat-rot}_j}}{N_m} \right)^{1/2} \qquad (1)$$

where $D_{i,\mathrm{nat\text{-}rot}_j}$ is the distance between atom i in the native structure and the corresponding atom in rotamer j and N_m is the number of atoms in residue m. The summation is carried out only over "moving atoms": side chain atoms beyond Cβ; these are the ones whose positions change when χ angles are altered. Throughout this paper, we follow this convention in the calculation of RMS interatomic differences in side chain atom positions. Some, but not all, other researchers include Cβ in such calculations. This lowers the resulting RMS value for a residue by the factor

$$\sqrt{\frac{N_m}{N_m + 1}}.$$

This factor amounts to a 29% reduction in the worst case (serine, $N_m = 1$) and 5% in the best case (tryptophane, $N_m = 9$). For this reason, our RMS values may not be directly comparable to those of other authors.

For an entire protein, or for some subset of its residues, we have for the RMS interatomic displacement of the side- chain atoms:

$$D_{\mathrm{rms}} = \left(\frac{\sum_{m=\mathrm{residues}} N_m D^2_{\mathrm{rms},m}}{\sum_{m=\mathrm{residues}} N_m} \right)^{1/2} \qquad (2)$$

where m ranges over the residues in question, and N_m is the number of moving atoms in residue m.

In using Equation (1), allowance is made for number-order symmetry, as in χ_2 of phenylalanine, tyrosine, and aspartate and χ_3 of glutamate. If rotation of one of these torsions by 180° gives a lower value of $D_{\mathrm{RMS},m}$, then the rotated conformation is the one utilized in the calculation.

Since the BRC rotamers are, for a given rotamer library, the ones that best fit the native side chains, the BRC represents the best possible fit to the native structure which one can obtain using the rotamer library from which the BRC was derived. Thus, in a rotamer-based side chain prediction method, one must consider both how well the BRC fits the native structure and how well the method predicts the BRC.

Use of Rotamers to Model Side Chain Conformations

We employed a rotamer library based on that described by Ponder and Richards.[45] We extended the library, however, to encompass all χ angles, whereas Ponder and Richards reported significant populations only through χ_2 in most cases. The extension was performed by adding the missing rotamers based on the simplest of chemical considerations. For example, if three-way torsions were to be added, as for χ_3 and χ_4 of lysine, we added them with the three "generic" values of ±60° and 180°. The nine rotamers thus added to fill out a "parent" rotamer with χ_1 and χ_2 values given by Ponder and Richards were each given a probability of appearance equal to one-ninth that assigned to the parent by these authors.

We performed studies of both hydrophobic cores and whole proteins. Therefore, we assessed how well BRCs based on our rotamer library fit the native side chains both of hydrophobic cores and of whole proteins. The cores of nine proteins whose structures have been accurately determined by x-ray crystallography and whose coordinates have been deposited in Brookhaven database[53] were isolated using the Sculpt program (compliments of G. Rose). This program "peels away" consecutive layers of solvent-exposed atoms to ultimately reveal the innermost hydrophobic core. The cores were visualized using the modeling program InsightII (Biosym Technologies, San Diego, CA) and confirmed to be well-packed hydrophobic cores. The BRC and its fit to the native structure were determined using Equation (1).

For our study of whole proteins, 49 proteins (Table I) with resolution better than 2.0 Å and crystallographic R values less than 20% were selected from the Brookhaven Data Bank.[53] The BRC and its fit to the native structure for each side chain were calculated using Equation (1) and are listed in Table I.

The *B* and −log *P* Measures

The object is to explore the space of side chain rotamers and to select the BRC from among them. To this end, we evaluated the effectiveness of two

TABLE I. Proteins Studied and Summary of Results

			BRC		
PDB[a]		N_{at}[b]	RMS[c]	N_{res}[d]	N_{res2}[e]
1rdg	(*1*)	126	0.589	34	27
4pti	(*2*)	155	0.566	36	31
2ovo	(*3*)	129	0.699	39	26
5rxn	(*4*)	143	0.750	39	31
3ebx	(*5*)	153	0.668	45	30
1ctf	(*6*)	150	0.580	46	35
1r69	(*7*)	170	0.760	52	37
1hoe	(*8*)	184	0.775	53	32
2ci2	(*9*)	190	0.810	55	41
2utg	(*10*)	190	0.817	60	44
1ubq	(*11*)	221	0.865	65	51
5pcy	(*12*)	228	0.759	73	50
1ccr	(*13*)	274	0.641	76	58
4fd1	(*14*)	288	0.956	80	65
2cdv	(*15*)	267	0.716	81	57
2rhe	(*16*)	257	0.743	82	51
1paz	(*17*)	294	0.764	88	67
1bp2	(*18*)	320	0.765	91	73
2rsp	(*19*)	309	0.961	92	69
2mhr	(*20*)	346	0.680	93	73
1lz1	(*21*)	377	0.743	95	75
2lzt	(*22*)	355	0.717	95	72
7rsa	(*23*)	317	0.563	97	63
1mba	(*24*)	346	0.789	99	74
2aza	(*25*)	336	0.727	99	70
4hhb	(*26*)	351	0.785	105	71
4fxn	(*27*)	390	0.976	115	89
4tnc	(*28*)	448	1.040	127	109
2alp	(*29*)	418	0.728	132	75
2i1b	(*30*)	445	0.819	132	99
3lzm	(*31*)	493	0.755	135	107
1ppd	(*32*)	595	0.748	153	114
8dfr	(*33*)	562	0.783	155	121
1hne	(*34*)	551	0.817	163	110
1tld	(*35*)	510	0.653	164	103
3est	(*36*)	624	0.755	183	115
2cga	(*37*)	578	0.779	191	107
3bcl	(*38*)	618	0.843	194	101
1ca2	(*39*)	732	0.720	201	154
2cab	(*40*)	697	0.719	203	142
2apr	(*41*)	796	0.617	239	161
2app	(*42*)	764	0.603	245	146
4pep	(*43*)	797	0.729	254	162
1gd1	(*44*)	855	0.769	261	182
6ldh	(*45*)	901	0.974	276	196
2cpp	(*46*)	1142	0.743	318	247
2fb4	(*47*)	1054	0.756	331	189
2cts	(*48*)	1245	0.901	348	265
3grs	(*49*)	1194	0.742	361	248
Sum:		22885		6751	4715
Avg:		467.0	0.758	137.8	96.2
SD:		296.1	0.106	88.8	60.5
Min:		126	0.563	34	26
Max:		1245	1.040	361	265

(*continued*)

TABLE I. Proteins Studied and Summary of Results (*Continued*)

PDB[a]		S_1[f]	$S_{1\&2}$[g]	S_{all}[h]	M_1[i]	$M_{1\&2}$[j]	M_{all}[k]
1rdg	(*1*)	0.765	0.704	0.559	0.824	0.815	0.647
4pti	(*2*)	0.806	0.613	0.472	0.833	0.645	0.472
2ovo	(*3*)	0.641	0.500	0.462	0.641	0.538	0.538
5rxn	(*4*)	0.769	0.710	0.744	0.769	0.677	0.692
3ebx	(*5*)	0.733	0.667	0.578	0.756	0.700	0.622
1ctf	(*6*)	0.739	0.600	0.609	0.652	0.571	0.543
1r69	(*7*)	0.673	0.568	0.423	0.731	0.649	0.500
1hoe	(*8*)	0.660	0.625	0.604	0.642	0.594	0.566
2ci2	(*9*)	0.655	0.585	0.491	0.691	0.634	0.582
2utg	(*10*)	0.733	0.591	0.550	0.733	0.659	0.567
1ubq	(*11*)	0.615	0.529	0.477	0.677	0.549	0.554
5pcy	(*12*)	0.740	0.620	0.603	0.726	0.660	0.626
1ccr	(*13*)	0.711	0.569	0.553	0.684	0.638	0.539
4fd1	(*14*)	0.825	0.646	0.613	0.787	0.631	0.600
2cdv	(*15*)	0.630	0.439	0.383	0.642	0.491	0.395
2rhe	(*16*)	0.756	0.588	0.646	0.768	0.667	0.671
1paz	(*17*)	0.727	0.597	0.591	0.705	0.537	0.545
1bp2	(*18*)	0.692	0.507	0.473	0.681	0.466	0.429
2rsp	(*19*)	0.761	0.667	0.641	0.772	0.623	0.620
2mhr	(*20*)	0.699	0.493	0.516	0.753	0.603	0.613
1lz1	(*21*)	0.800	0.667	0.632	0.789	0.680	0.621
2lzt	(*22*)	0.811	0.681	0.589	0.779	0.708	0.600
7rsa	(*23*)	0.680	0.651	0.577	0.711	0.635	0.588
1mba	(*24*)	0.747	0.608	0.545	0.828	0.730	0.626
2aza	(*25*)	0.798	0.643	0.576	0.798	0.657	0.636
4hhb	(*26*)	0.724	0.549	0.562	0.743	0.662	0.619
4fxn	(*27*)	0.748	0.584	0.557	0.739	0.584	0.557
4tnc	(*28*)	0.646	0.550	0.496	0.677	0.578	0.520
2alp	(*29*)	0.795	0.720	0.697	0.780	0.760	0.689
2i1b	(*30*)	0.689	0.576	0.553	0.689	0.576	0.561
3lzm	(*31*)	0.741	0.607	0.519	0.748	0.617	0.526
1ppd	(*32*)	0.784	0.649	0.582	0.797	0.693	0.601
8dfr	(*33*)	0.755	0.620	0.555	0.781	0.653	0.606
1hne	(*34*)	0.712	0.645	0.601	0.736	0.618	0.577
1tld	(*35*)	0.750	0.612	0.579	0.774	0.689	0.634
3est	(*36*)	0.689	0.548	0.563	0.699	0.530	0.541
2cga	(*37*)	0.717	0.607	0.550	0.696	0.626	0.539
3bcl	(*38*)	0.691	0.584	0.521	0.711	0.614	0.567
1ca2	(*39*)	0.741	0.623	0.547	0.741	0.643	0.562
2cab	(*40*)	0.754	0.563	0.537	0.714	0.570	0.512
2apr	(*41*)	0.820	0.770	0.695	0.808	0.745	0.695
2app	(*42*)	0.784	0.747	0.698	0.796	0.740	0.698
4pep	(*43*)	0.732	0.654	0.606	0.732	0.698	0.634
1gd1	(*44*)	0.785	0.637	0.598	0.805	0.670	0.617
6ldh	(*45*)	0.728	0.561	0.565	0.714	0.520	0.554
2cpp	(*46*)	0.764	0.628	0.591	0.796	0.664	0.623
2fb4	(*47*)	0.719	0.646	0.577	0.734	0.677	0.613
2cts	(*48*)	0.710	0.566	0.514	0.698	0.577	0.529
3grs	(*49*)	0.740	0.593	0.562	0.740	0.625	0.590
Sum:							
Avg:		0.732	0.610	0.566	0.739	0.634	0.583
SD:		0.051	0.065	0.069	0.051	0.070	0.063
Min:		0.615	0.439	0.383	0.641	0.466	0.395
Max:		0.825	0.770	0.744	0.833	0.815	0.698

(*continued*)

TABLE I. Proteins Studied and Summary of Results (*Continued*)

PDB[a]		Avg res RMS[l]	Frac res <1.0 Å RMS[m]	Avg res RMS_{50}[n]	Frac res <1.0 Å RMS_{50}[o]	Total RMS[p]	Total RMS_{50}[q]
1rdg	(*1*)	1.061	0.618	0.376	0.941	1.392	0.575
4pti	(*2*)	1.420	0.528	0.987	0.778	2.044	1.601
2ovo	(*3*)	1.540	0.487	0.722	0.718	2.434	1.291
5rxn	(*4*)	1.102	0.692	0.367	0.974	1.604	0.515
3ebx	(*5*)	1.313	0.489	0.864	0.667	2.002	1.220
1ctf	(*6*)	1.275	0.565	0.594	0.870	1.860	1.135
1r69	(*7*)	1.275	0.565	0.594	0.870	1.860	1.135
1hoe	(*8*)	1.644	0.547	1.114	0.679	3.103	2.558
2ci2	(*9*)	1.449	0.491	1.040	0.655	2.147	2.028
2utg	(*10*)	1.465	0.533	1.320	0.567	2.065	1.991
1ubq	(*11*)	1.349	0.492	0.782	0.738	1.897	1.210
5pcy	(*12*)	1.220	0.548	0.635	0.822	1.679	0.993
1ccr	(*13*)	1.403	0.513	0.552	0.868	1.988	0.825
4fd1	(*14*)	1.357	0.525	0.870	0.750	1.941	1.295
2cdv	(*15*)	1.699	0.358	1.288	0.568	2.296	2.355
2rhe	(*16*)	1.196	0.646	0.915	0.780	2.347	2.259
1paz	(*17*)	1.424	0.523	0.718	0.818	2.115	1.517
1bp2	(*18*)	1.607	0.451	1.250	0.527	2.304	1.515
2rsp	(*19*)	1.419	0.554	0.733	0.826	2.390	1.413
2mhr	(*20*)	1.307	0.591	0.816	0.796	2.085	1.784
1lzl	(*21*)	1.297	0.589	0.532	0.884	2.036	1.018
2lzt	(*22*)	1.211	0.611	0.884	0.758	1.803	1.421
7rsa	(*23*)	1.220	0.588	0.681	0.825	2.011	1.670
1mba	(*24*)	1.177	0.545	0.734	0.727	1.606	1.067
2aza	(*25*)	1.229	0.576	0.749	0.788	1.689	1.068
4hhb	(*26*)	1.241	0.543	0.773	0.743	1.902	1.447
4fxn	(*27*)	1.470	0.461	1.022	0.609	2.134	1.546
4tnc	(*28*)	1.638	0.416	1.233	0.560	2.067	1.668
2alp	(*29*)	1.032	0.674	0.591	0.864	1.841	1.519
2i1b	(*30*)	1.391	0.492	1.009	0.667	1.942	1.481
3lzm	(*31*)	1.479	0.504	0.961	0.726	2.269	1.839
1ppd	(*32*)	1.287	0.562	0.830	0.784	2.142	1.986
8dfr	(*33*)	1.369	0.535	0.689	0.813	2.128	1.261
1hne	(*34*)	1.444	0.509	0.834	0.724	2.291	1.543
1tld	(*35*)	1.210	0.598	0.911	0.780	2.284	2.264
3est	(*36*)	1.405	0.503	0.994	0.699	2.249	2.066
2cga	(*37*)	1.444	0.524	1.144	0.712	2.306	2.359
3bcl	(*38*)	1.388	0.521	1.099	0.697	2.221	2.114
1ca2	(*39*)	1.415	0.517	0.844	0.746	2.035	1.316
2cab	(*40*)	1.389	0.488	0.842	0.719	2.019	1.444
2apr	(*41*)	0.998	0.661	0.692	0.812	1.589	1.269
2app	(*42*)	0.975	0.657	0.624	0.841	1.463	1.145
4pep	(*43*)	1.173	0.579	0.882	0.748	1.594	1.424
1gd1	(*44*)	1.150	0.556	0.804	0.743	1.733	1.505
6ldh	(*45*)	1.437	0.475	1.106	0.616	1.967	1.506
2cpp	(*46*)	1.261	0.575	0.768	0.767	1.898	1.183
2fb4	(*47*)	1.224	0.562	0.843	0.743	1.830	1.419
2cts	(*48*)	1.557	0.428	1.141	0.615	2.194	1.940
3grs	(*49*)	1.305	0.573	0.755	0.792	2.049	1.431
Sum:							
Avg:		1.338	0.540	0.860	0.744	2.029	1.538
SD:		0.167	0.066	0.226	0.098	0.301	0.463

(*continued*)

TABLE I. Proteins Studied and Summary of Results (*Continued*)

PDB[a]	Avg res RMS[l]	Frac res <1.0 Å RMS[m]	Avg res RMS_{50}[n]	Frac res <1.0 Å RMS_{50}[o]	Total RMS[p]	Total RMS_{50}[q]
Min:	0.975	0.358	0.367	0.527	1.392	0.515
Max:	1.699	0.692	1.320	0.974	3.103	2.558

[a] Protein Data Bank[53] entry code. Protein names as follows:
(*1*) Rubredoxin (*D. gigas*)
(*2*) Pancreatic trypsin inhibitor
(*3*) Ovomucoid (third domain)
(*4*) Rubredoxin (*C. pasteurianum*)
(*5*) Erabutoxin B
(*6*) 50S ribosomal protein (C-terminal domain)
(*7*) 434 Repressor (N-terminal domain)
(*8*) α-Amylase inhibitor
(*9*) Chymotrypsin inhibitor 2
(*10*) Uteroglobin
(*11*) Ubiquitin
(*12*) Plastocyanin
(*13*) Cytochrome c
(*14*) Ferredoxin
(*15*) Cytochrome c3
(*16*) Immunoglobulin (Bence-Jones)
(*17*) Pseudoazurin
(*18*) Phospholipase A2
(*19*) Protease (Rous sarcoma virus)
(*20*) Myohemerythrin
(*21*) Human lysozyme
(*22*) Lysozyme (egg white)
(*23*) Ribonuclease A
(*24*) Myoglobin
(*25*) Azurin
(*26*) Human hemoglobin
(*27*) Flavodoxin
(*28*) Troponin C
(*29*) α-Lytic protease
(*30*) Interleukin 1b
(*31*) T4 lysozyme
(*32*) Papain
(*33*) Dihydrofolate reductase
(*34*) Human neutrophil elastase
(*35*) β-Trypsin
(*36*) Elastase
(*37*) Chymotrypsinogen A
(*38*) Bacteriochlorophyll A protein
(*39*) Carbonic anhydrase
(*40*) Carbonic anhydrase form B
(*41*) Rhizopuspepsin acid proteinase
(*42*) Penicillopepsin acid proteinase
(*43*) Pig pepsin
(*44*) Glyceraldehyde-3-phosphate dehydrogenase
(*45*) Lactate dehydrogenase
(*46*) Cytochrome P450
(*47*) Immunoglobulin Fab
(*48*) Citrate synthase
(*49*) Glutathione reductase

[b] Number of moving atoms.

[c] RMS fit to BRC, computed over all side-chain atoms.

[d] Number of moving residues.

[e] Number of moving residues with two or more χ angles.

[f] Simulated annealing results, fraction χ_1 correct.

[g] Simulated annealing results, fraction χ_1 and χ_2 both correct.

[h] Simulated annealing results, fraction all χ (complete rotamer) correct.

[i] Monte-Carlo consensus results, fraction χ_1 correct.

[j] Monte-Carlo consensus results, fraction χ_1 and χ_2 both correct.

[k] Monte-Carlo consensus results, fraction all χ (complete rotamer) correct.

[l] Monte-Carlo consensus results, average RMS value of residues.

[m] Monte-Carlo consensus results, fraction of residues with RMS values less than 1 Å.

[n] Monte-Carlo consensus results, 50% lowest-entropy residues, average RMS value of residues.

[o] Monte-Carlo consensus results, 50% lowest-entropy residues, fraction of residues with RMS values less than 1 Å from BRC.

[p] Monte-Carlo consensus results, total RMS displacement of all side-chain atoms.

[q] Monte-Carlo consensus results, 50% lowest-entropy residues, total RMS displacement of all side-chain atoms.

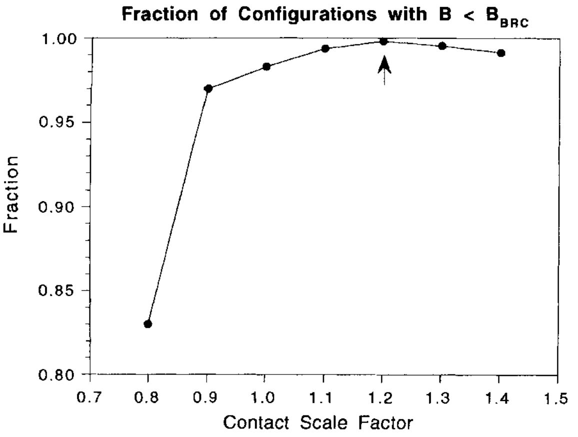

Fig. 1. Effect of varying contact factor on prediction accuracy of hydrophobic cores. The abscissa is the factor applied to the Ponder and Richards contact distances.[45] The ordinate is the fraction of total configurations in all cores exhibiting B values less than or equal to that of the BRC. The maximum at a contact-distance factor of 1.2 indicates that this value is most efficient in selecting for the BRC.

quantities that may be readily calculated for any such configuration. The first quantity, B, is the total number of bad interatomic contacts (bumps) exhibited by the configuration. A bad interatomic contact occurs whenever a moving atom in a side chain comes closer than some allowed interatomic distance to another atom in either the fixed part of the molecule or in another rotamer. The starting point for this list of distances was the table of allowable interatomic distances given by Ponder and Richards[45]; however, after some experimentation, it was determined (Fig. 1) that distances greater than these, by a factor of 1.2, give better results, and this factor was used throughout.

To facilitate the computation of B, two tables are built by the prediction program at the start of each simulation. The first table, B_f, the fixed-bump array, is a one-dimensional array with a single entry for each rotamer of each side chain. Each B_f element is filled with the number of bad interatomic contacts between the rotamer in question and the fixed part of the molecule. The second table, B_m, the moving-bump array, is a square array each of whose dimensions is equal to the total number of rotamers for all side chains. Each element of B_m contains the number of bad interatomic contacts that occur between the two rotamers corresponding to the column and row of the matrix. B_m elements for pairs of rotamers for the same side chain are not populated, nor are elements populated for interactions of rotamers from side chain pairs too far apart for their atoms to make contact. Thus, given a list of rotamers for some or all of the residues in a protein, the number of bad contacts in the configuration may be calculated rapidly, as follows:

$$B = \sum_i B_{f,i} + \sum_i \sum_{j>i} B_{m,ij} \qquad (3)$$

where indices i and j range over the rotamers in the configuration.

A secondary scoring criterion for a rotamer configuration is provided by the overall a priori probability of a configuration. This is derived from the frequencies of appearance of individual rotamers as given by Ponder and Richards[45] and modified as described earlier for extensions to the library. The frequencies are normalized to unity for each residue type in a preprocessing step. The a priori probability of appearance of a configuration, P, is equal to the product of the rotamer frequencies for the rotamers making up the configuration. In practice, multiplying hundreds of small probabilities gives a very small number indeed, so $-\log P$ (base 10) is used instead of P itself. The negative sign ensures that, as for the B scores, higher values correspond to less favorable configurations. $-\log P$ is calculated using the formula:

$$-\log P = - \sum_{j=\text{rotamers}} \log P_j \qquad (4)$$

where j ranges over the rotamers in the configuration and P_j is the probability of appearance of rotamer j. Thus, for any configuration of rotamers, we can easily determine two scores, B and $-\log P$, by table look-up.

Side Chain Prediction of Small Buried Cores

Side chain configurations for protein core regions were sampled exhaustively; that is, we generated every rotamer configuration, treating non-core atoms as fixed. For each configuration, B and $-\log P$ were calculated, and the results plotted on a two-dimensional histogram. We observed the fraction of residues exhibiting lower B and lower $-\log P$ than the BRC, and used this information to assess the relative efficacies of these criteria.

Side Chain Prediction of Entire Proteins

If a protein with 100 variable side chains is simulated with a rotamer library having about three rotamers per residue, there are $3^{100} \approx 10^{48}$ possible configurations. This makes exhaustive search intractable. Thus, simulated annealing[46] was employed to search configuration space for the rotamer combination with the smallest number of bumps and, within that group, the greatest overall probability of occurrence (lowest $-\log P$). Simulated annealing is a technique for identifying low-lying minima in a large discrete space; we used the implementation described in Press et al.,[54] which employs an adaptive cooling schedule.

The energy function (E) of a configuration used in the annealing is the number of bumps (B) plus a small constant times $-\log P$:

$$E = B - \epsilon \log P \qquad (5)$$

ϵ is chosen small enough to ensure that probability

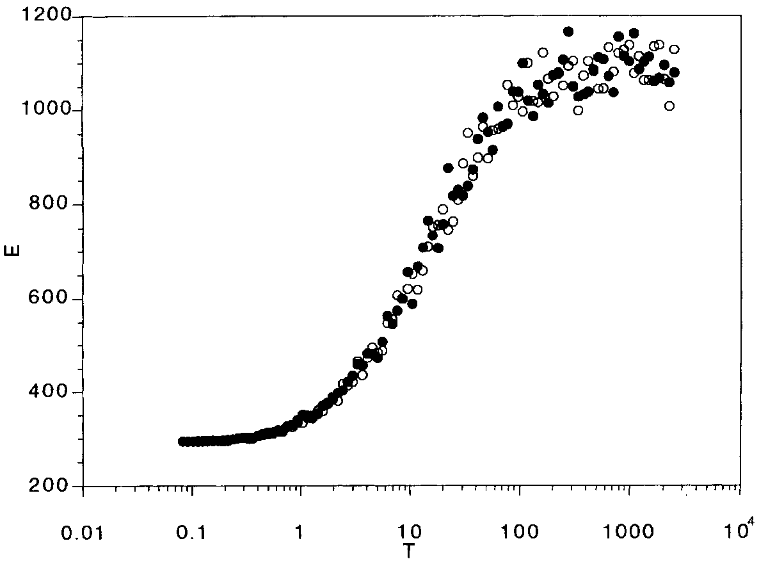

Fig. 2. Energy vs. log T for a simulated-annealing simulation. The data shown were obtained from simulations on Brookhaven PDB entry[53] 2fb4 (an immunoglobin Fab fragment). Squares and circles represent two simulations, which were initiated with different random configurations. The similarity between the shapes of the curves and the final energies obtained for duplicate runs is typical of the proteins studied.

is used only to break ties between configurations with equal numbers of bumps; that is, it is small enough so that ϵ log P can never be as large as unity, at least for nativelike configurations. This ensures that the $-\log P$ term will, at most, break ties between competing configurations exhibiting equal B values. In practice, ϵ is chosen to be the reciprocal of N_{mov}, the number of moving residues in the protein. This follows from the observation that $-\log P$ is always considerably smaller than N_{mov} for the BRC.

A random "move" is then made, which consists of selecting a single residue at random and selecting a new rotamer for it, also at random. After some experimentation, we found that predictions were slightly but consistently better if the rotamers were selected from a probability distribution taken from their a priori probabilities as given in the rotamer library rather than from a uniform distribution; therefore, this procedure was used. The energy of the new configuration (E') is then calculated. If the Metropolis criterion[47] is met, the new configuration is accepted; otherwise, the original configuration is kept. The Metropolis criterion can be stated: If $\Delta E = E' - E$ is negative, accept the new conformation; otherwise, accept it with probability $\exp(-\Delta E/T)$, where T is the "temperature" at the current point in the simulation.

T is a control parameter that dictates how likely the algorithm is to accept a configuration with a higher energy, in the hope of escaping local minima and eventually reaching a global energy minimum. If $T >> \Delta E$, then acceptance is nearly certain. Simulated annealing should be initiated at a temperature sufficiently high to allow all states to freely interconvert. In our exhaustive search of core conformations, we found that the highest B value observed was approximately six times the number of moving atoms in the core. Based on the assumption that this relationship would also hold for whole proteins, we used this value as the intitial value of T. Since the minimum value of E is theoretically zero, the starting value chosen will certainly well exceed the value of $|\Delta E|$ for nearly all configuration pairs. Thus, the initial value of T is appropriate. This is also borne out by Figure 2, of E vs. log T during the annealing process for one protein, which exhibits the characteristic sigmoidal shape of "good" annealing curves.

The configuration we report at the end of a simulated annealing run is not necessarily the last configuration found during the run. Rather, it is the lowest-energy configuration encountered at any time during the annealing process. This is sometimes colloquially termed a "pocket algorithm".

In practice, we repeat the annealing process four times for each protein, starting with a different random configuration each time. Nearly always, the four runs give different configurations with the same value of B, although the small $-\epsilon$ log P term may vary among them. This gives us confidence that in these situations we are finding configurations degenerate in globally minimal values of B. In some cases, there were small variations in $-\log P$ among these minimal B configurations; thus, convergence is not necessarily complete for the secondary scoring criterion.

Because finding the BRC for the entire protein would be the best possible result, the success of the algorithm was assessed by comparing predicted side chain conformations to those found in the BRC. The results reported are taken from the lowest-scoring simulated annealing run for each protein. We assess our success in predicting the correct rotamer for χ_1 only, χ_1 and χ_2 together, and all χ angles. For the χ_1 and all-χ assessments, all 6751 residues in the sample are included. The results we present for χ_1 and χ_2, however, include only those residues which have two or more χ angles. These constitute 4715 residues, about 70% of the total.

Low-Temperature Monte-Carlo Sampling and Entropy Analysis of Side Chain Conformations

Since the four annealing runs give different configurations, it would seem appropriate to look for consensus positions among these runs: residues that exhibit the same rotamer in all four reported configurations. Another observation, however, caused us to attempt additional simulations. The BRC itself always has a B score significantly greater than the minimal score found in simulated annealing. In fact, the BRC score is always greater than this minimal score by a value very close to the number of moving atoms in the side chains allowed to vary. We felt that if we could sample many configurations close to this energy value, we could perform a more thorough search for consensus than would be possible merely by examining the four low-energy configurations obtained from simulated annealing. This sampling can be accomplished by means of Monte Carlo[47] simulations at a constant low temperature. Starting with a configuration obtained from simulated annealing, and fixing the temperature at the empirically determined value of $T = 3$, the average energy of the structures sampled during the Monte-Carlo simulation was close to the BRC score for nine proteins randomly selected from among those studied. Thus, this temperature was utilized during the Monte-Carlo simulations.

During the constant-temperature sampling, a tally was kept of the number of times each rotamer was sampled for each residue. The statistical entropy was used to describe the results. For a given residue, m, suppose the rotamers 1, 2, ..., k are sampled with frequencies $f_1, f_2, ..., f_n$, where the f_j sum to unity. Then the entropy of this residue is given by

$$S_m = -\sum_{j=1}^{k} f_j \ln f_j \tag{6}$$

The effective number of rotamers sampled by residue m,k^*_m, is given by[55,56]

$$k^*_m = \exp(S_m) \tag{7}$$

k^* is monotonic in S and is equal to k, the total number of rotamers for the residue, when all the rotamers are sampled equally frequently ($S = \ln k$) and to unity when only a single rotamer is sampled ($S = 0$). Because of its straightforward interpretation, k^*, rather than S, is used in the discussion of the results.

Duplicate runs of varying lengths on nine of the 49 proteins studied demonstrated that k^* values for all residues are convergent to within 1% when the number of Monte Carlo steps reaches 10,000 times the number of moving residues; thus, all the Monte Carlo runs reported here were run for this number of steps. To avoid getting stuck in the basin of a single low-energy starting configuration, we divide the total number of Monte-Carlo steps among four runs, each initiated from a configuration obtained from a different simulated-annealing simulation. A common tally of rotamer composition was kept for all four Monte-Carlo runs. The Monte-Carlo consensus configuration is the list of the most often visited rotamer for each residue in the protein.

The results of the low-temperature Monte Carlo runs were analyzed using the same criteria utilized for the simulated annealing runs: fraction of all residues in the 49-protein sample correctly predicted at χ_1, at χ_1 and χ_2, and at all χ values. In addition, the analysis was performed as a function of k^*, in order to determine whether a high degree of consensus (low k^*) implies predictability. In addition, we report RMS interatomic displacement of moving atoms from their positions in the native proteins for the Monte-Carlo consensus results and for the half of the residues within each protein exhbiting the lowest k^* values.

A separate set of entropies was calculated on a χ angle-by-χ angle basis, rather than on a residue-by-residue basis, but this variability measure did not materially improve the predictability even of individual χ angles; therefore we report prediction accuracy only for the residue-by-residue analysis. On the other hand, χ_1 entropies have the advantage of being directly comparable for all residue types: each amino-acid type has only three generic values of χ_1: gauche(+), gauche(−), and trans. The χ_1 entropy of a rotamer is defined by Equation (6), where k is now equal to 3 and j ranges over the generic χ_1 values. The χ_1 entropies were used to investigate the extent to which conformational freedom is correlated with exposed surface area in a pooled sample of all residue types, taken together.

Correlation of Rotamer Entropies With Solvent Accessibilities

A modified version of SCULPT (compliments of George Rose) was used to determine fractional solvent accessibilities for the moving atoms of each protein studied. This program divides the solvent-exposed surface areas, as determined by the program of Lee and Richards,[57] by a precalculated "standard-state" maximum exposure to give the fractional degree of exposure for each atom. For each residue, the fractional exposure of the moving atoms, as a group, was calculated and tabulated.

From the overall appearance of scatter plots of k^* vs. fractional exposure (see Results), it was clear that no strong linear correlation is present between these observables. In order to assess whether subtle correlations might be present, we computed significance and strength using contingency-table analysis, as described by Press et al.[54] Both scales were

TABLE II. Protein Cores Studied

Protein	PDB[a]	Res[b]	RMS[c]	Residues[d]
Hemerythrin	1hmz	2.0	0.692	L28 T51 F55 L98
Glycosidase inhibitor	1hoe	2.0	0.376	S21 V33 K34 V35 V36 L70
Repressor (chain 3)	1lrd	2.5	0.458	L318 V336 M340 V347 F351 L357 L365
Immunoglobulin	2fb4	1.9	0.506	V117 I138 F141 Y142 V146 V148 Y174 V197 H199 V204
Lysozyme	3lzm	1.7	0.571	L7 D10 G11 L99 N101 M102 V149 F153 Y161
Hemoglobin (chain A)	4hhb	1.74	0.799	W14 Y24 L29 L66 M76 F98 L101 S102 L105 T108 L109 F128 V132 S133 L136
α-Chymotrypsin (chain A)	5cha	1.67	0.550	Q30 S45 V52 V53 T54 T138 L199 I212 V213 Y228
Cytochrome c	5cyt	1.5	0.273	L32 L94 L98
Dihydrofolate reductase	8dfr	1.7	0.538	N5 S6 I7 V50 I51 W113 V115 Y121

[a] Brookhaven Protein Data Bank entry.[53]
[b] Resolution of the x-ray structure in Ångstroms.
[c] RMS deviation between the BRC and the native structure.
[d] Residues in core.

divided into two portions. Except where otherwise indicated, "very buried" residues were defined as those whose moving atoms had solvent-accessible areas of 10% or less, and "very fixed" residues were defined as those with k^* values less than or equal to 1.9. The χ^2 test was used to assess the significance of deviation from the null hypothesis that degree of solvent exposure is independent of rotamer entropy for this twofold binning of each variable. Where correlation was significant, statistical entropy analysis was carried out to determine its strength.

These analyses were performed for the individual residue types valine and threonine, using rotamer entropies. These types were chosen because they both have a single χ angle and are topologically similar, and because one is polar and one is nonpolar. On the other hand, both are β-branched. To remove this bias, a similar analysis was performed for the entire sample, using χ_1 entropies. For this analysis, the fractional solvent-exposed surface utilized for each residue was that of those atoms moved by χ_1 and χ_1 only; these are the atoms in the side chain γ positions. This avoids a spurious area calculation when the end of a lysine side chain, for example, is exposed while the CG atom, which, alone, is affected only by χ_1, is buried.

Contingency-table analysis was also performed on the valine, threonine and χ_1 data sets in order to directly assess the extent to which fractional solvent exposure is correlated with χ-angle predictability. A "predicted" scale was defined, which was given a value of unity if the χ angle in question was correctly determined and a value of zero if it was not. The analysis was then used to determine the extent to which the property "correctly predicted" was correlated with the property "very buried."

RESULTS

Use of Rotamers to Model Side Chain Conformations

To investigate how well native side chains can be modeled using rotamers, the side chain RMS interatomic deviations between the native structure and the BRC were calculated as described in Equation 2 for each of nine small hydrophobic cores (Table II) and for each of forty-nine whole proteins (Table I). Histograms of these values for whole proteins are shown in Figure 3a on a protein-by-protein basis and Figure 3b on a residue-by residue basis. Except for a few residues, the extended Ponder and Richards rotamer library[45] can be used to reasonably accurately model side chain conformations in both hydrophobic cores and whole globular proteins. The BRC fits about 85% of the side chains studied and about 98% of the whole proteins studied (all but one, in fact) to better than 1 Å RMS interatomic displacement.

Side Chain Prediction of Small Buried Cores

Exhaustive search of every possible rotamer configuration is guaranteed to find the BRC if appropriate criteria for distinguishing the BRC can be determined. We now discuss how well the B and $-\log P$ measures described earlier perform this task.

The side chain configurations of small buried cores in nine proteins (Table II) were exhaustively generated and, for each configuration, B and $-\log P$ were computed. We investigated how many configurations exhibited better (lower) values of B, $-\log P$, and both for each core. The data in Table III demonstrate that B is a far more efficacious criterion than $-\log P$, since, in every case, the BRC ranks higher on a list of configurations sorted by B than on such a list sorted by $-\log P$. Furthermore, there is clearly an advantage to using both measures together in some way, since fewer configurations are "better" than the BRC in both criteria than in either criterion alone. For three of the nine cores studied—1hmz, 4hhb, and 5cyt—including hemoglobin, the largest core studied, no other configuration has lower B and $-\log P$ values than the BRC. For five of the remaining six cores, only two or three configurations rank lower than the BRC in both criteria. In

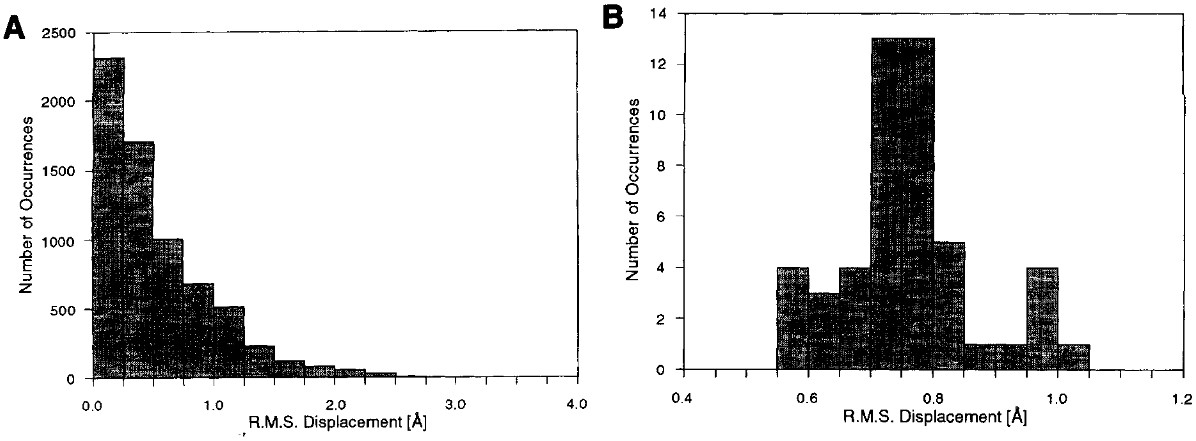

Fig. 3. Distribution of RMS atomic deviations of side-chain atoms between the native and BRC configurations. **a**: For the 49 whole proteins in the sample (mean = 0.758 Å, standard deviation = 0.106 Å). **b**: For the 6751 individual residues in the sample (mean = 0.529 Å, standard deviation = 0.464 Å).

TABLE III. Results of Exhaustive Search on Hydrophobic Cores

				N_{config} better in		
PBD[a]	N_{res}[b]	N_{at}[c]	N_{config}[d]	$-\log P$[e]	B[f]	Both[g]
1hmz	4	14	192	36	36	0
1hoe	6	14	20,412	1,818	7	2
1lrd	7	22	16,128	669	7	2
2fb4	10	38	466,560	1,710	11	1
3lzm	9	34	677,376	35,296	4	2
4hhb	15	55	891,814,000	2,372,752	57	0
5cha	10	30	1,749,600	634,582	540	392
5cyt	3	9	64	4	1	0
8dfr	8	30	97,200	2,421	5	1

[a] Brookhaven Protein Data Bank[53] entry. See Table I for names of proteins.
[b] Number of moving side chains.
[c] Number of moving atoms.
[d] Number of configurations of core residues.
[e] Number of configurations with lower values of $-\log P$ than the BRC.
[f] Number of configurations with lower values of B than the BRC.
[g] Number of configurations with both lower B and lower $-\log P$ than the BRC.

the worst case, 5cha, the BRC only 392 out of 1,749,600, or 0.03%, of all configurations rank better than the BRC in both B and $-\log P$.

A detailed analysis of the hydrophobic core of glycosidase inhibitor (1hoe) is presented in Table IV and Figure 4. This core has 20,412 rotamer configurations. A two-dimensional histogram of B and $-\log P$ for these configurations is shown in Figure 4. The bin containing the BRC, labeled $B = 12$, $-\log P = 4.064$, is marked with an asterisk. This figure illustrates the extent of the discrimination afforded by these criteria. As shown in Table IV, the two configurations that score better than the BRC differ from the BRC in only one of the six positions, lysine 34. Even for this residue, χ_1 and χ_2 are identical in all three configurations. For all cores studied, there is considerable rotamer consensus at some or most of the positions in those configurations that exhibit both low B and low $-\log P$ (Table V).

Side Chain Prediction of Entire Proteins

The key conclusion from these results on hydrophobic cores is that B is a more effective indicator of the "nativelikeness" of a protein side chain configuration than $-\log P$, but that $-\log P$ contributes additional information. This leads naturally to the choice of B as a primary and $-\log P$ as a secondary scoring criterion. This concept is embodied in the definition of the energy given in Equation (5).

This definition was used in attempts to predict the side chain conformations of all moving residues in the 49 proteins listed in Table I, using simulated annealing and low-temperature Monte-Carlo simulation to explore configuration space. The simulated-annealing results reported in this table pertain, in each case, to the single run of the four performed on each protein that gave the lowest-energy configuration, though, as mentioned earlier,

	-log(P): 2	3	4	5	6	7	8	B totals
B: 88	0	0	0	1	2	1	0	4
87	0	0	0	1	2	1	0	4
86	0	0	1	6	9	3	1	20
85	0	0	1	2	4	2	1	10
84	0	0	7	14	15	9	1	46
83	0	0	3	9	11	6	0	29
82	0	0	17	26	33	13	0	89
81	0	0	4	19	29	14	2	68
80	0	1	21	55	47	19	3	146
79	0	1	13	46	44	25	5	134
78	0	4	24	76	68	32	4	208
77	0	1	25	68	83	38	4	219
76	0	3	41	88	112	40	3	287
75	0	5	33	98	134	50	9	329
74	0	2	48	113	153	45	11	372
73	0	12	50	134	158	64	14	432
72	0	10	65	146	160	72	10	463
71	0	13	78	158	180	82	8	519
70	0	7	81	180	174	84	9	535
69	0	15	109	184	185	85	10	588
68	0	17	97	193	186	79	13	585
67	0	26	112	219	193	75	12	637
66	0	18	114	219	168	74	11	604
65	0	34	134	239	184	62	6	659
64	0	23	130	225	180	61	6	625
63	0	30	151	237	176	55	5	654
62	1	36	151	227	151	52	7	625
61	1	43	151	235	151	36	7	624
60	1	50	158	240	137	36	3	625
59	2	42	152	217	120	30	4	567
58	4	48	169	225	125	22	2	595
57	1	39	165	202	100	24	0	531
56	3	49	185	193	88	22	0	540
55	1	54	165	180	82	12	0	494
54	3	50	175	172	75	17	1	493
53	3	56	165	167	63	11	1	466
52	5	49	141	174	58	13	1	441
51	2	55	146	156	58	14	0	431
50	3	60	141	129	60	18	0	411
49	4	62	128	122	53	15	2	386
48	6	58	129	112	52	13	4	374
47	2	52	117	115	42	16	6	350
46	4	54	106	99	50	16	3	332
45	0	53	102	79	53	22	2	311
44	2	46	96	72	57	22	2	297
43	8	46	98	61	56	22	5	296
42	6	29	89	62	55	18	6	265
41	8	42	68	72	47	22	4	263
40	3	33	65	64	49	27	2	243
39	5	30	61	56	53	25	0	230
38	3	33	49	48	52	23	2	210
37	2	30	50	43	51	16	3	195
36	7	27	45	45	42	13	1	180
35	3	17	39	48	30	19	0	156
34	3	20	35	43	27	14	0	142
33	2	14	30	37	32	8	0	123
32	2	13	33	35	28	4	0	115
31	2	13	29	33	21	3	0	101
30	3	10	27	30	19	5	0	94
29	3	7	25	28	17	2	0	82
28	0	9	24	30	13	2	0	78
27	0	10	19	28	12	0	0	69
26	0	5	19	24	13	0	0	61
25	0	5	24	17	9	2	0	57
24	0	4	23	16	4	1	0	48
23	0	4	20	17	5	0	0	46
22	0	5	12	17	4	0	0	38
21	0	5	10	18	1	0	0	34
20	0	3	11	12	0	0	0	26
19	0	1	13	9	0	0	0	23
18	0	2	13	7	0	0	0	22
17	0	3	8	4	0	0	0	15
16	0	3	5	5	0	0	0	13
15	0	1	5	4	0	0	0	10
14	0	0	6	1	0	0	0	7
13	0	0	4	0	0	0	0	4
12	0	0	*3	0	0	0	0	3
11	0	0	3	0	0	0	0	3
10	0	0	1	0	0	0	0	1
-logP totals:	108	1602	5067	6786	4905	1728	216	

Fig. 4. Distribution of B and $-\log P$ for all configurations of the six residue core of glycosidase inhibitor (1hoe). The bin labeled $B = 12$, $-\log P = 4$, which contains the BRC, is marked with an asterisk. The totals shown for B and $-\log P$ are the one-dimensional distributions for these variables.

TABLE IV. 1hoe Core: Comparison of the BRC Configuration With Configurations Better in Both B and $-\log P^a$

Residue	BRC	Conf 1	Conf 2
Ser 21	−	−	−
Val 33	−	−	−
Lys 34	+ttt	+t+t	+tt−
Val 35	t	t	t
Val 36	t	t	t
Leu 70	−	−	−

[a] The symbols −, +, and t represent the gauche(−), gauche(+), and trans conformations, respectively.

TABLE V. Hydrophobic Cores: Comparison of Rotamer Configurations Better Than or Equal to the BRC in both B and $-\log P$

PBD[a]	N_{better}[b]	N_{res}[c]	Number constant[d]	Fraction constant[e]
1hoe	3	6	5	0.83
1lrd	3	7	6	0.85
2fb4	2	10	9	0.90
3lzm	3	9	6	0.67
5cha	393	11	4	0.36
8dfr	2	8	6	0.75

[a] Brookhaven Protein Data Bank entry.[53] See Table I for names of proteins.

[b] Number of configurations better than or equal to the BRC in both B and $-\log P$.

[c] Number of moving residues in the core.

[d] Number of residues having the same rotamer in all $N_{\text{better configurations}}$.

[e] Fraction of residues having the same rotamer in all $N_{\text{better configurations}}$.

multiple runs did tend to give nearly identical final energies. Predictions for whole proteins range from 38% to 74% correct on a rotamer basis (all χ angles predicted correctly), with a mean of 57%. χ_1 is predicted correctly an average of 73% of the time, and χ_1 and χ_2 together are predicted correctly on average 61% of the time, considering only residues that have two or more χ angles. Table VI gives pooled results for all the residues in the sample. The results are substantially the same as for whole proteins. Prediction results vary significantly among proteins (Table I), as others have found[31,40,42]; thus, testing a prediction method on only one or two proteins may not adequately indicate its overall performance.

Table VI also gives results for the pooled sample for the strategy of simply selecting the most likely rotamer for each residue. This strategy succeeds only 42.1% of the time.

Table VII summarizes the prediction accuracy by residue type. Looking at the results for all χ angles (i.e., prediction of the complete rotamer), we do most poorly for the charged residues ARG and LYS. The terminal residues for these amino acids tend to be exposed to solvent and to be poorly resolved in the x-ray structures; recall that the rotamer library had to be filled out for the last two χ angles of these types. On the other hand, χ_1 is predicted nearly as well for these as for the other residue types, and the prediction of χ_1 and χ_2, together, is also reasonably good. We also predict the overall rotamer poorly for ASN, GLN, and HIS, for reasons that can be easily understood. The terminal groups of these residues look symmetrical to our B criterion; thus, except for the contribution of the nonspecific $-\log P$ term, we can never get the terminal χ angle correct more than half the time. Again, we do reasonably well for χ_1 here, and, in the case of GLN, for χ_1 and χ_2.

Hydrophobic and aromatic residues are well predicted at all levels. This is not surprising, since B, the predominant term in our potential, measures only steric effects. The polar β-branched THR is also predicted well, indicating that, despite its polarity, its conformation is mitigated largely by steric factors. Among the χ_1 predictions, SER is most poorly predicted. This is in accord with the fact that, alone among the naturally occurring amino acids, all of the movable side chain atoms of serine (only the hydroxyl oxygen, in our simulation) are highly polar. This, together with its lack of β branching, is also consistent with the notion that, among the naturally occurring amino acid types, steric factors count the least in the conformational preference of serine.

Low-Temperature Monte-Carlo Sampling and Entropy Analysis of Side Chain Conformations

The configurations resulting from multiple simulated annealing processes usually give the same or nearly the same minimum energies, despite the fact that the configurations differ significantly from each other. To aid in the interpretation of the simulation results, we sought a method of determining a consensus configuration and of measuring the degree of consensus. The degree of consensus at a given residue position would then constitute a measure of confidence in the prediction of each side chain.

We observed that the BRC, which represents the native structure, generally exhibits a B value greater than the minimal value that simulated annealing provides. The B value exhibited by the BRC generally exceeds that of a configuration resulting from simulated-annealing by about the number of moving atoms in the side chains permitted to vary. Using the conditions described in the Methods section, the constant-temperature Monte-Carlo procedure samples configurations whose B values are less than or approximately equal to that of the BRC. This is demonstrated in Figure 5. Simulations were performed using these conditions, and, for each residue in each protein, the populations of the rotamers sampled were printed out, as were the values of k^*

TABLE VI. Comparison of Prediction Strategies for Whole Proteins

Strategy	χ_1 correct[a]	χ_1 and χ_2 correct[b]	All χ correct[c]
Most probable rotamer[d]	—[e]	—[e]	0.421
Simulated annealing[f]	0.737	0.614	0.573
Monte-Carlo, all residues[g]	0.743	0.634	0.588
Monte-Carlo, maximal $R-W$[h] (fraction of sample)	(0.981)	(0.621)	(0.629)
Monte-Carlo, 50% lowest k^*[i]	0.841	0.778	0.792

[a] Fraction of moving residues for which the generic χ_1 angle is correctly predicted.
[b] Fraction of moving residues possessing two or more χ angles for which χ_1 and χ_2 are correctly predicted.
[c] Fraction of moving residues for which the correct rotamer is predicted.
[d] The most probable rotamer is selected for each residue.
[e] Not determined.
[f] Results from the simulated annealing run that gave the lowest energy.
[g] Monte-Carlo consensus configuration, all residues.
[h] Monte-Carlo results for those residues whose k^* values are less than that at which $R-W$ is maximal. The following entries gives the fraction of residues this encompasses.
[i] Monte-Carlo results for the half of the residues exhibiting the lowest k^* values.

[Eq. (7)] for each residue derived from these population distributions.

The consensus configuration for a given protein then consists of the most frequently visited rotamer for each residue. The analysis presented here is based on this consensus configuration and on the k^* value for each residue. For example, the figures given for the fraction of times χ_1 was predicted correctly are based on the χ_1 values in the consensus configuration, not the pooled χ_1 frequencies for all rotamers visited by a given side chain. Experimentation with the latter method did not give substantial improvement over what is reported here.

Figure 6 shows how the success of prediction varies with k^*. Results are shown in Figure 6a for χ_1 predictions, Figure 6b for χ_1 and χ_2 predictions, and Figure 6c for the prediction of all χ angles. In each figure, the curve labeled $R+W$ gives the fraction of residues that have k^* values less than or equal to the x axis value. Curve R gives the fraction of residues which both are correctly predicted and also whose k^* values are less than or equal to the x axis value. Thus, the curve labeled $R/(R+W)$ represents the fraction of residues with k^* less than the x axis value which are correctly predicted. This curve starts near unity at $k^*=1$ and decreases as k^* increases. This shows that, indeed, the fewer the effective number of rotamers encountered by a given residue during the constant-temperature Monte Carlo simulation (i.e., the lower the k^* value), the greater the reliability of the prediction; consensus is correlated with predictability. The R and $R/(R+W)$ curves converge at the maximum k^* value because here $R+W$ is equal to unity.

The value of R or, equivalently, $R/(R+W)$, at maximum k^* gives the fractional prediction success for all residues, taken together. These values are tabulated for each protein studied in Table I and for the pooled sample in Table VI. χ_1 is predicted correctly 74% of the time, χ_1 and χ_2 are predicted correctly for 63% of residues that have two or more χ angles, and the correct rotamer is predicted correctly 59% of the time. These figures encompass all movable residues in the proteins studied.

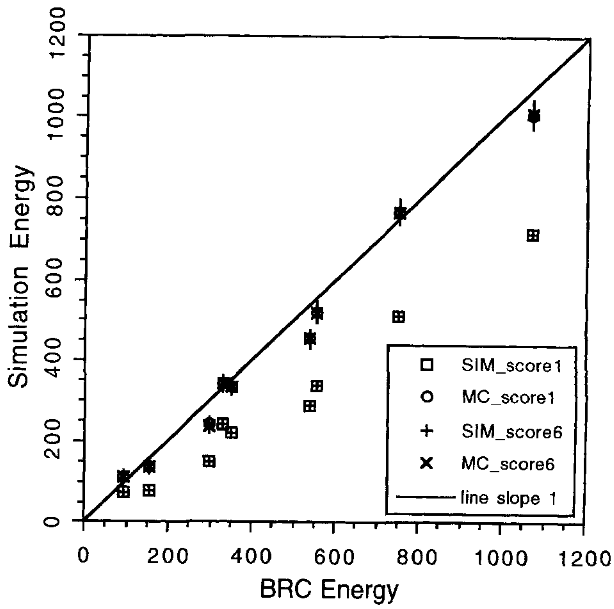

Fig. 5. Simulation energies vs. BRC energy. Duplicate simulated annealing runs (SIM_score1 and SIM_score6) and duplicate Monte-Carlo runs (MC_score1 and MC_score6) were performed on nine proteins randomly chosen from among the 49 studied. The coincidence of data points from duplicate runs demonstrates convergence from different starting configurations. The MC energies shown are average energies sampled during these runs, and the (small) error bars associated with these energies represent standard deviations of the energies sampled. The fact that the MC data points lie close to a straight line with a slope of unity and an intercept of zero indicates that the Monte-Carlo conditions utilized (T = 3) indeed samples configurations exhibiting energies close to that of the BRC.

The Monte-Carlo method provides the user with a choice: whether to pay attention only to side chains

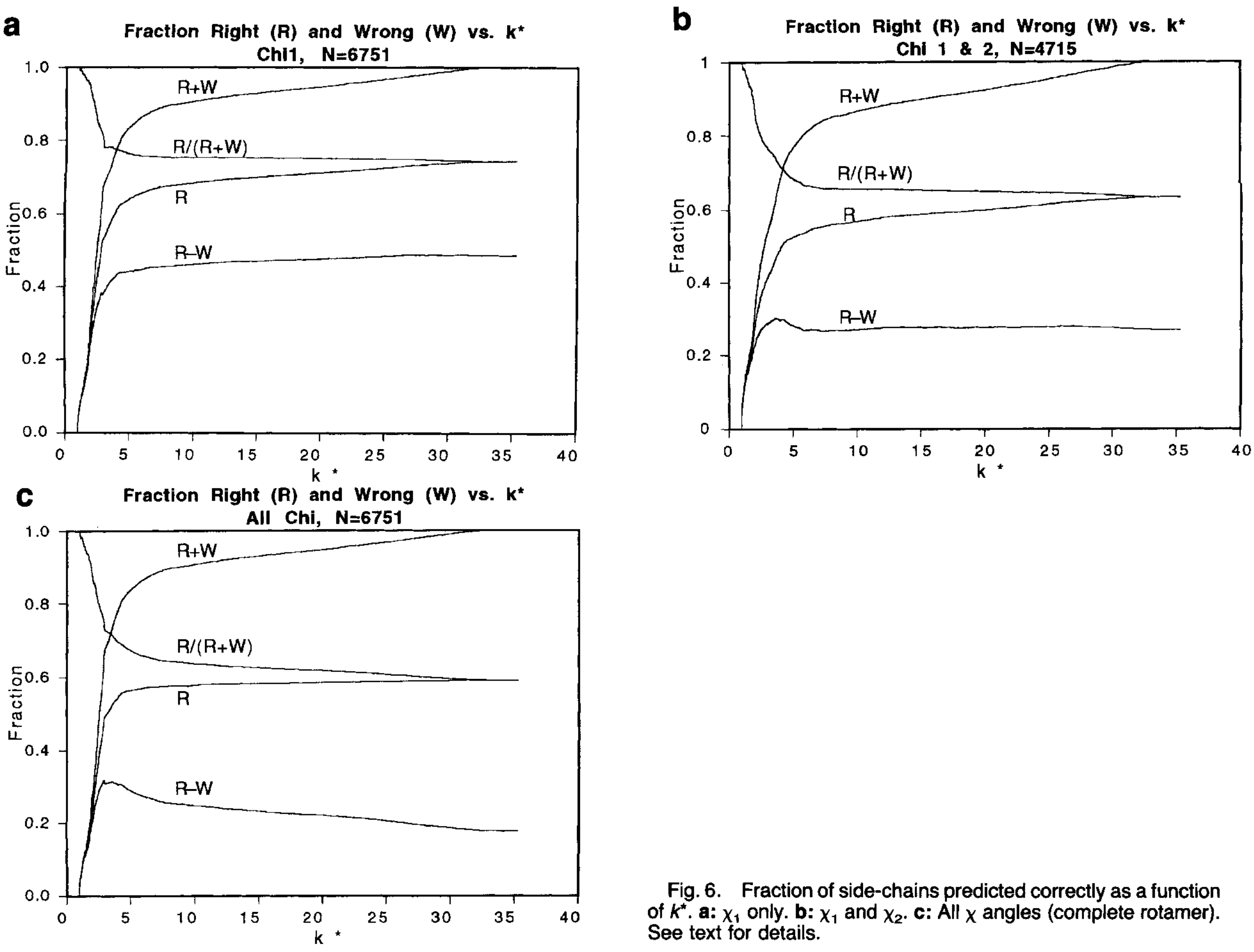

Fig. 6. Fraction of side-chains predicted correctly as a function of k^*. **a:** χ_1 only. **b:** χ_1 and χ_2. **c:** All χ angles (complete rotamer). See text for details.

predicted with a high degree of confidence (those with low k^* values), or, alternatively, to utilize predictions for all residues, regardless of k^*. Prediction accuracies are given in Table VI for several possible strategies involving entropy discrimination. One such strategy involves use of the curves labeled R − W in Figure 6a,b,c. $R - W$ constitutes one possible figure of merit that can be used when deciding which k^* cutoff to use for prediction; it penalizes an incorrect prediction to the same extent that it rewards a correct one. If this figure of merit is appropriate, then predictions should be ignored for k^* values greater than the x axis position where this curve is maximal. The $R + W$ curve at this cutoff gives the fraction of all the residues for which a prediction will be made, and the $R/(R + W)$ curve gives the fraction of residues correctly predicted at this k^* value.

Table VI shows that, using this procedure, predictions are very close to 75% correct, regardless of whether "correct" refers to prediction of χ_1, χ_1 and χ_2, or all χ angles; however, for these three situations, predictions are ventured for 98%, 62% and 63% of the residues, respectively. The "all-χ" prediction is more accurate at this level than the prediction of χ_1 and χ_2, only, because the all-χ dataset includes residues that have only a single χ angle, and these are predicted especially well, whereas the χ_1-and-χ_2 dataset excludes these residues.

Another possible strategy involves using a k^* cutoff value that encompasses some predetermined fraction of the side chains studied. For example, Table VI shows that if predictions are ventured only for the 50% of the residues exhibiting the lowest k^* values, then χ_1 is predicted correctly 84% of the time, χ_1 and χ_2 are predicted correctly 79% of the time, and the entire rotamer is predicted correctly 79% of the time.

Table VII shows how these results vary by residue type. They follow the same trends as the simulated annealing results.

Figure 6 and Table VII can be used to assess confidence in the prediction of a particular residue without any a priori knowledge of the protein structure—for example, without first examining whether the residue is buried or on the surface. Furthermore, the results shown below seem to indicate that excluding surface residues would remove from consideration many residues that are in fact predicted well without significantly improving the predictability of the remaining residues.

Correlation of Rotamer Entropies With Solvent Accessibilities

Figure 7a is a scatter plot of k^* vs. fractional solvent accessibility for all valine residues in the sample. There is certainly no linear correlation between these variables. χ^2 analysis[54] of the two-way contingency table shown in Table VIII does, however, demonstrate a correlation at a significance level of $P(\chi^2) = 99.3\%$; that is, the likelihood is only $Q(\chi^2) = 1 - P(\chi^2) = 0.7\%$ that a χ^2 value as large as that observed would be exhibited if correlation between the two categories were completely absent. Statistical entropy analysis,[54] however, indicates that $U(x,y)$, the strength of the correlation, is less than 1%; that is, less than 1% of the variation in either variable can be attributed to correlation with the other. χ^2 analysis for all threonine residues, performed the same way as described for valine, exhibits a significance of only $P(\chi^2) = 52\%$; exposed surface area and conformational flexibility thus appear independent here.

The use of 1.9 as the bifurcation value (see the Methods section) for k^* is based on the appearance of plots similar to those shown in Figure 6c for valine and threonine alone: both residues are close to 100% correctly predictable for k^* values less than 1.9. It might be argued, however, that a lower value of k^* should be used to define residues that are strongly fixed, in addition to being strongly predictable. Table IX lists the results of a similar analysis performed using a bifurcation value of 1.5 for k^*; the results are similar. The appearance of Figure 7a is consistent with a weak tendency for highly exposed residues to exhibit high k^* values. The results of a contingency-table analysis using bifurcation values of 0.75 and 2.0 for these two scales (Table IX) demonstrates that the effect is indeed weak, the two variables being correlated only to the extent of about 2%.

Similar calculations were performed for the combined χ_1 values of all residues in the sample (Figure 7b; Table IX). This procedure increases the number of data points, allowing for increased significance if an effect is indeed present. Figure 7b, like 7a, exhibits no apparent correlation. Here, the strength of the correlation (Table IX) is about 7% at most, still weak.

In Figure 7b, several data points appear which exhibit fractional solvent accessibilities considerably in excess of unity. Examination of these systems in detail revealed that the atoms in question have unrealistically long bond lengths. When these bond lengths are reduced to their generic values, the fractional solvent accessibilities of the atoms in question drop to near unity. The strong sensitivity of apparent solvent accessibility on bond length for exposed atoms is left as a subject of contemplation for others employing this measure.

The contingency-table analyses described above demonstrate that rotamer entropy is, at best, weakly correlated with solvent exposure. If this observation were an artifact of the simulation methodology employed, side chain predictions would not be expected to work as well for exposed as for buried residues, regardless of their entropy values. Figure 7c depicts histograms of fractional solvent exposure for correctly and incorrectly predicted valine residues in the data set. There appears to be only a weak tendency for the incorrectly predicted residues to exhibit the higher solvent exposures: both histograms are dominated by a high population of completely buried residues. Contingency-table analysis of these data and of the analogous data for threonine and for the pooled χ_1 sample shows that deeply buried residues are, at best, only slightly better predicted than exposed residues. The mutual effect of solvent exposure and predictability is less than 2% (Table IX).

Timings

Figure 8 is a plot of CPU times versus number of moving residues for the 49 whole proteins studied. The times were obtained on a MIPS R4400/150 processor (Silicon Graphics Iris Indigo). Two times are given for each protein: the simulated-annealing time, which is the time from the start of the run to the end of the first simulated-annealing procedure, and the Monte-Carlo time, which is the total time for the run, including four simulated-annealing processes and the constant-temperature Monte-Carlo steps associated with them. The figure shows that obtaining convergent entropies is about six times more expensive than a single simulated-annealing minimization.

If entropies are not desired, it is probably best to carry out two simulated-annealing minimizations, to check for the possibility that one might have converged to a high-energy minimum. This entire process would then take somewhat less than one third as much CPU time as the full Monte-Carlo procedure, since the setup time is small, but not insignificant: it constitutes about one quarter of the time through the end of the first simulated-annealing run. Each simulated-annealing minimization takes somewhat less time than its associated contant-temperature Monte-Carlo steps; put another way, the entire constant-temperature Monte-Carlo procedure takes more than four times as long as a single simulated-annealing minimization.

For a typical small protein containing about 100 moving residues (perhaps 130 residues in total), time through the first simulated-annealing run is about two CPU minutes; the full Monte-Carlo treatment takes about 12 minutes. For the largest sytems studied, which contain about 350 moving residues (about 450 residues in total), the corresponding

TABLE VII. Summary of Results by Residue Type

a. Overall Results

Res[a]	N_{res}[b]	S_1[c]	$S_{1\&2}$[d]	S_{all}[e]	M_1[f]	$M_{1\&2}$f2[g]	M_{all}[h]
Charged:							
ASP	478	0.653	0.653	0.653	0.657	0.657	0.657
GLU	430	0.644	0.426	0.426	0.656	0.451	0.451
ARG	307	0.648	0.466	0.114	0.671	0.537	0.143
LYS	532	0.701	0.509	0.124	0.677	0.536	0.147
Polar:							
SER	751	0.546	N/A	0.546	0.551	N/A	0.551
CYS	75	0.720	N/A	0.720	0.733	N/A	0.733
THR	531	0.708	N/A	0.708	0.702	N/A	0.702
ASN	423	0.676	0.336	0.336	0.662	0.383	0.383
GLN	325	0.708	0.517	0.271	0.723	0.529	0.289
Aromatic:							
HIS	165	0.752	0.406	0.406	0.764	0.382	0.382
PHE	316	0.873	0.826	0.826	0.883	0.823	0.823
TRP	140	0.864	0.636	0.636	0.857	0.650	0.650
TYR	292	0.887	0.815	0.815	0.901	0.849	0.849
Hydrophobic:							
VAL	679	0.814	N/A	0.814	0.831	N/A	0.831
LEU	681	0.860	0.838	0.838	0.881	0.859	0.859
MET	170	0.812	0.618	0.459	0.788	0.547	0.441
ILE	456	0.886	0.754	0.754	0.901	0.785	0.785
SUM:	6751						
AVG:	397.1	0.750	0.600	0.556	0.755	0.614	0.569
SD:	199.9	0.103	0.171	0.244	0.106	0.170	0.242
MIN:	75.0	0.546	0.336	0.114	0.551	0.382	0.143
MAX:	751.0	0.887	0.838	0.838	0.901	0.859	0.859

b. Entropy Dependence[i]

Type	M_1[f]	k^*_1[j]	F_1[k]	$M_{1\&2}$[g]	$k^*_{1\&2}$[l]	$F_{1\&2}$[m]	M_{all}[h]	k^*_{all}[n]	F_{all}[o]
ARG	0.600	20.55	1.000	0.600	11.85	0.720	0.600	2.43	0.036
	0.800	8.99	0.489	0.800	3.37	0.098	0.800	2.30	0.020
ASN	0.600	5.92	1.000	0.600	2.48	0.118	0.600	2.48	0.118
	0.800	3.59	0.496	0.800	1.71	0.002	0.800	1.71	0.002
ASP	0.600	2.97	1.000	0.600	2.97	1.000	0.600	2.97	1.000
	0.800	2.04	0.559	0.800	2.04	0.559	0.800	2.04	0.559
CYS	0.600	2.99	1.000	N/A	N/A	N/A	0.600	2.99	1.000
	0.800	2.43	0.547	N/A	N/A	N/A	0.800	2.43	0.547
GLN	0.600	8.17	1.000	0.600	6.43	0.738	0.600	3.65	0.188
	0.800	5.94	0.631	0.800	4.55	0.354	0.800	2.06	0.015
GLU	0.600	6.14	1.000	0.600	3.65	0.484	0.600	3.65	0.484
	0.800	3.59	0.442	0.800	2.57	0.165	0.800	2.57	0.165
HIS	0.600	4.75	1.000	0.600	1.68	0.061	0.600	1.68	0.061
	0.800	3.72	0.885	0.800	1.31	0.012	0.800	1.31	0.012
ILE	0.600	4.69	1.000	0.600	4.69	1.000	0.600	4.69	1.000
	0.800	4.69	1.000	0.800	4.38	0.967	0.800	4.38	0.967
LEU	0.600	3.09	1.000	0.600	3.09	1.000	0.600	3.09	1.000
	0.800	3.09	1.000	0.800	3.09	1.000	0.800	3.09	1.000
LYS	0.600	35.27	1.000	0.600	28.03	0.761	p	p	p
	0.800	23.91	0.517	0.800	15.58	0.188	p	p	p
MET	0.600	6.97	1.000	0.600	5.11	0.635	0.600	4.21	0.447
	0.800	6.56	0.947	0.800	2.03	0.065	0.800	2.03	0.065
PHE	0.600	3.71	1.000	0.600	3.71	1.000	0.600	3.71	1.000
	0.800	3.71	1.000	0.800	3.71	1.000	0.800	3.71	1.000
SER	0.600	2.95	0.806	N/A	N/A	N/A	0.600	2.95	0.806
	0.800	2.38	0.029	N/A	N/A	N/A	0.800	2.38	0.029
THR	0.600	2.99	1.000	N/A	N/A	N/A	0.600	2.99	1.000
	0.800	2.37	0.640	N/A	N/A	N/A	0.800	2.37	0.640
TRP	0.600	4.02	1.000	0.600	4.02	1.000	0.600	4.02	1.000
	0.800	4.02	1.000	0.800	1.57	0.629	0.800	1.57	0.629

(*continued*)

TABLE VII. Summary of Results by Residue Type (*Continued*)

b. Entropy Dependence[i]

Type	M_1[f]	k^*_1[j]	F_1[k]	$M_{1\&2}$[g]	$k^*_{1\&2}$[l]	$F_{1\&2}$[m]	M_{all}[h]	k^*_{all}[n]	F_{all}[o]
TYR	0.600	3.14	1.000	0.600	3.14	1.000	0.600	3.14	1.000
	0.800	3.14	1.000	0.800	3.14	1.000	0.800	3.14	1.000
VAL	0.600	3.00	1.000	N/A	N/A	N/A	0.600	3.00	1.000
	0.800	3.00	1.000	N/A	N/A	N/A	0.800	3.00	1.000

N/A, not applicable.

[a] Residue type.

[b] Number of occurrences.

[c] Fraction of moving residues for which χ_1 is predicted correctly, lowest-energy simulated-annealing run.

[d] Fraction of moving residues which have more than one χ angle for which χ_1 and χ_2 are predicted correctly, lowest-energy simulated-annealing run.

[e] Fraction of moving residues for which the complete rotamer is predicted correctly, lowest-energy simulated-annealing run.

[f] Fraction of moving residues for which χ_1 is predicted correctly, Monte-Carlo consensus.

[g] Fraction of moving residues which have more than one χ angle for which χ_1 and χ_2 are predicted correctly, Monte-Carlo consensus.

[h] Fraction of moving residues for which the complete rotamer is predicted correctly, Monte-Carlo consensus.

[i] For each residue type, we show the k^* cutoff values and the fraction of times we venture a prediction in the situations in which we predict 60% and 80% of the occurrences correctly.

[j] k^* cutoff value for prediction of χ_1, only.

[k] Fraction of occurrences with k^* below the χ_1 cutoff value.

[l] k^* cutoff value for prediction of χ_1 and χ_2.

[m] Fraction of occurrences with k^* below the χ_1 and χ_2 cutoff value.

[n] k^* cutoff value for prediction of all χ (complete rotamer).

[o] Fraction of occurrences with k^* below the all-χ cutoff value.

[p] Prediction accuracies as high as 60% were not achieved at any k^* cutoff value.

times are 20 CPU minutes and 2.5 CPU hours. Consideration of the algorithm indicates that CPU time should be $O(n^2)$ in the number of moving residues; the general appearance of Figure 8 is consistent with this notion.

DISCUSSION

Overview

We showed first that the side chains of hydrophobic cores and whole proteins in our sample can, for the most part, be well modeled using rotamers. For small buried cores, the RMS interatomic difference between the crystal structure and the BRC ranges from 0.2 to 0.8 Å (Table II). For complete proteins, RMS differences are only slightly greater (Table I, Figure 3a), although some individual residues are poorly fit (Figure 3b). Schrauber et al.[58] argue that rotamer libraries do not adequately cover side chain conformational space, since some nonrotameric conformations are found even in highly resolved structures. Although not an exact representation, the rotamer approximation is close enough to account for at least 80% of the residue conformations,[58] and would seem reasonable enough for the first step in homology model-building. Energy minimization[49] or more detailed modeling can subsequently be used to further refine the resulting structure.

Based upon studies of hydrophobic cores, we determined that two simple criteria—number of unfavorable van der Waals contacts (B) and configuration probability ($-\log P$)—are efficient indicators of the nativelike nature of a side chain configuration. These studies revealed that the B criterion is far more powerful than the $-\log P$ criterion if used alone; however, the latter adds power to the method if they are used together. For cores, few *configurations* score better than the BRC in both measures.

Discrimination in favor of the BRC could be improved by calculating B using minimum allowable interatomic contact radii 20% greater than those employed by Ponder and Richards[45] (Fig. 1). These

TABLE VIII. Contingency-Table Analysis of Correlation Between Fractional Solvent Exposure and Effective Number of Rotamers Sampled[a]

	fracarea ≤ 0.1		fracarea > 0.1	
N(total) = 655	N = 426		N = 229	
k^* ≤ 1.9	Found:	99	Found:	33
N = 132	Expected:	85.9	Expected:	46.1
k^* > 1.9	Found:	327	Found:	196
N = 523	Expected:	340	Expected:	183

[a] Two-way contingency-table analysis: x_{cut} = 0.1000, y_{cut} = 1.9000, n = 655. x = fracarea, y = k^*.

χ^2 analysis: df = 1; χ^2 = 7.2149. $Q(\chi^2)$ = 7.23*e*-03; $P(\chi^2)$ = 9.93*e*-01; Cramer's V = 1.050*e*-01.

Entropy analysis: H_x =0.64721; H_y = 0.50251, $H(x,y)$ = 1.14398. $H(x|y)$ = 0.64147, $H(y|x)$ = 0.49677. $U(x|y)$ = 8.871*e*-03, $U(y|x)$ = 1.143*e*-02, $U(x,y)$ = 9.988*e*-03.

P and Q are measures of correlation significance. Low Q (high P) means that it is unlikely that the underlying model—independence of x and y—would give a χ^2 as large as the observed value by chance. Thus, low Q (high P) implies significant correlation between x and y. Q and P range from 0 to 1.

$U(x|y)$, $U(y|x)$, and $U(x,y)$ are measures of correlation strength. Low $U(y|x)$ means that the value of y depends only weakly on x; similarly for $U(x|y)$. Low $U(x,y)$ indicates only a weak mutual influence of x and y on each other. The Us range from 0 to 1.

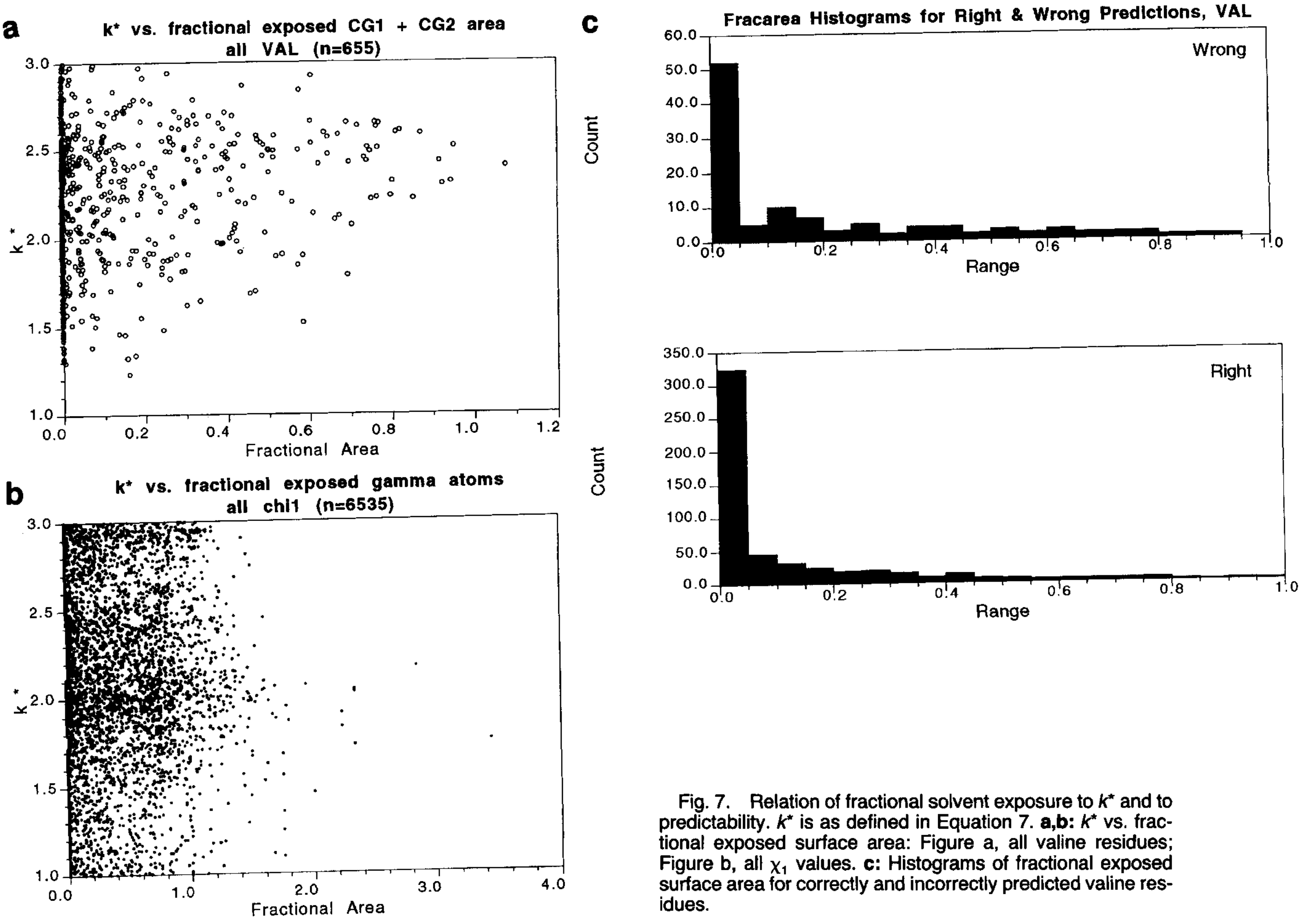

Fig. 7. Relation of fractional solvent exposure to k^* and to predictability. k^* is as defined in Equation 7. **a,b:** k^* vs. fractional exposed surface area: Figure a, all valine residues; Figure b, all χ_1 values. **c:** Histograms of fractional exposed surface area for correctly and incorrectly predicted valine residues.

authors employed an all-atom representation in their work, whereas we include heavy atoms only; thus, it is not surprising that the contact radii that work best for us are somewhat larger than those that work best for Ponder and Richards.

An energy function was then defined by the expression $E = B - \epsilon \log P$, with ϵ defined to be so small that no nativelike $-\log P$ value can exceed unity in absolute value. This ensures that for low-energy configurations, the $-\log P$ term will, at best, break ties between pairs of configurations exhibiting the same B value. This definition follows from

TABLE IX. Summary of Contingency-Table Analyses

Data[a]	x[b]	y[c]	x_{cut}[d]	y_{cut}[e]	$Q(\chi^2)$[f]	$U(x,y)$[f]
VAL	A	k^*	0.1	1.9	7.23e-03	0.010
VAL	A	k^*	0.1	1.5	7.54e-01	—[g]
VAL	A	k^*	0.75	2.0	1.57e-02	0.021
THR	A	k^*	0.1	1.9	4.80e-01	—[g]
THR	A	k^*	0.1	1.5	4.02e-01	—[g]
χ_1	A	k^*	0.1	1.9	6.09e-98	0.050
χ_1	A	k^*	0.1	1.5	2.06e-126	0.073
VAL	A	R	0.1	0.5	9.06e-04	0.015
THR	A	R	0.1	0.5	6.53e-01	—[g]
χ_1	A	R	0.1	0.5	2.40e-25	0.013

[a] Data set: all valines, all threonines, or pooled χ_1 data.

[b] A: fractional solvent-exposed surface.

[c] k^*: effective number of rotamers sampled (Eq. 7); R: 1 if χ is correctly predicted, 0 if incorrectly predicted.

[d] Bifurcation point for binning the x variable.

[e] Bifurcation point for binning the y variable.

[f] See footnote to Table VIII.

[g] $Q(\chi^2)$ is less than 10% (correlation is insignificant at the 90% level), so an assessment of correlation strength is not meaningful.

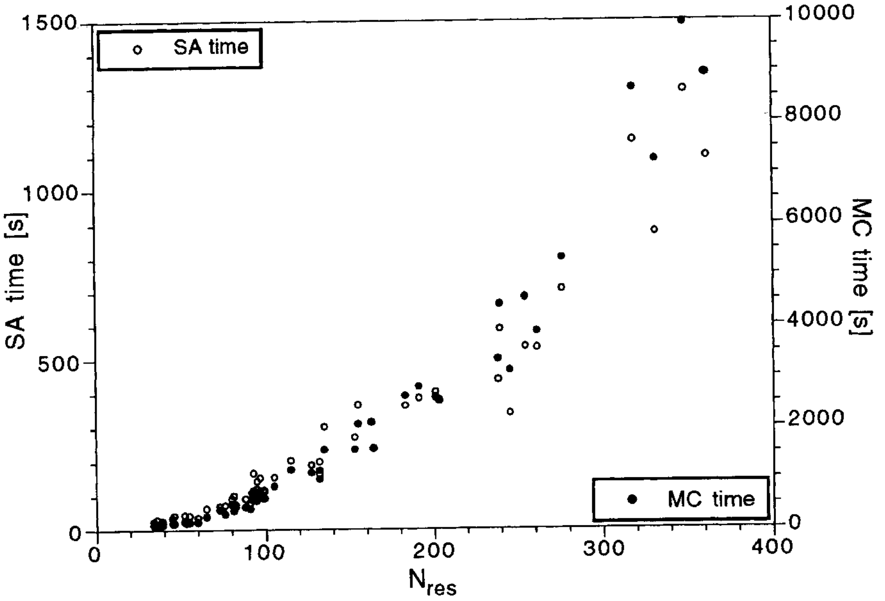

Fig. 8. CPU time vs. number of moving residues. The hollow symbols and the left-hand *y* axis represent the timings through the conclusion of a single simulated-annealing run. The filled symbols and the right-hand *y* axis represent the timings through the conclusion of the Monte-Carlo procedure.

the observation that *B*, taken alone, is far more efficient than $-\log P$ in selecting the BRC. This energy function was used in a simulated-annealing procedure to predict the rotamer configurations of 49 well-resolved whole proteins of diverse tertiary structure selected from the Brookhaven Protein Data Bank.[53] The results were assessed by comparison with the rotamers of BRC (the configuration which best fits the crystal structure). For all the residues in the sample (Table VI), 57% of the rotamers are predicted correctly. Results vary quite a bit from protein to protein (Table I), as observed by other workers (see below), but no obvious explanation could be found for this. This observation does, however, point out the need to test an algorithm of this nature on a variety of proteins of varying structure.

Using simulated annealing, predictions are more accurate for χ_1 and for χ_1 and χ_2, together, than for whole rotamers. In the entire sample, χ_1 is correctly predicted 74% of the time and χ_1 and χ_2 are correct 61% of the time (Table VI), considering, for the latter statistic, only those residues which have more than a single χ angle.

Whole rotamers are well predicted for hydrophobic and aromatic residues, and for threonine, which is β-branched (Table VII). Presumably, this is because the potential considers mainly steric factors, which dominate for these residues. This interpretation is consistent with the observation that χ_1 and χ_2 are reasonably well predicted for all residue types, with the exceptions noted in the next paragraph.

ARG and LYS, with positive charges at the termini, are well predicted in χ_1 and χ_2, but the whole rotamer is poorly predicted. This is presumably both because the potential does not "see" polar effects and also because these termini tend to be immersed in solvent and somewhat disordered. MET is also poorly predicted at the whole-rotamer level. This may not be fully ascribable to the modest polarity of its sulfur atom; simply the fact that it consists of a long, unbranched chain may render it less conformationally restricted than a branched residue would be.

ASN, GLN, and HIS are poorly predicted at the whole rotamer level, since their ends appear twofold symmetric to the dominant *B* criterion. This is perhaps the strongest limitation of the potential used here. We also observe that SER exhibits the least well predicted χ_1 value. This is consistent with the fact that, unique among the naturally occurring amino acids, both moving atoms on the serine side chain (the hydroxyl oxygen and hydrogen, though our simulation omits the hydrogen) are highly polar. We infer that the steric contribution to serine's χ_1 conformational preference is less than that for the other amino acid types.

Given that factors other than steric effects enter our potential only by means of the nonspecific $-\log$ *P* term, it is surprising that the method presented here yields results comparable to or better than those found using more sophisticated energetic measures (see below). This does, however, confirm the common doctrine that packing plays a large role in protein structure and stability. Our B measure, in fact, indexes only half of the packing problem: the avoidance of bad contacts. The other half—the avoidance of voids—is not taken into account; however, when studying whole proteins, we did observe that the BRC configurations always exhibited B values significantly greater than the minima obtained from simulated annealing. This may be an indication that the BRC is achieving "good" contacts not present in the lowest-energy configurations we obtain.

We were able to use this observation to advantage by finding an empirical, constant temperature which, when utilized in a constant-temperature Monte-Carlo simulation, caused the average energy of the configurations sampled to be quite close to that of the BRC (Fig. 5). Performing the Monte-Carlo steps affords two advantages over the simulated-annealing procedure described above. First, it gives rise to a consensus conguration whose prediction accuracy mildly exceeds that of the simulated-annealing method: χ_1 is predicted correctly 74% of the time, χ_1 and χ_2 63% of the time, and the whole rotamer 59% of the time (Table VI). The average residue in the average protein is predicted to about 1.3 Å RMS and 54% of the residues in the Monte-Carlo consensus configuration are predicted to better than 1.0 Å RMS with respect to the native structure (Table I). These values include both the deviation of the BRC from the native structure and the deviation of the predicted structure from the BRC.

Second, and more importantly, the Monte-Carlo procedure provides, for each residue, an entropy of rotamer appearance that is indicative of its prediction reliability, provided the simulation is carried out long enough for these entropies to converge. We indexed conformational consensus by k^*, an entropy-derived measure of the effective number of rotamers visited by a given residue. Figure 6 shows that the residues exhibiting low values of k^* (high consensus) are those most accurately predicted. For example, if prediction is restricted to the 50% of the sample exhibiting the lowest values of k^*, our prediction accuracies rise to 84% for χ_1, 78% for χ_1 and χ_2 taken together, and 79% for whole rotamers. For the 50% of each protein exhibiting the lowest k^* values (Table I), the average residue in the average protein is predicted to an atomic RMS displacement of 0.860 Å with respect to the native, and 74% of these residues are predicted to better than 1.0 Å RMS. We emphasize that the k^* values, which indicate which residues are most reliably predicted, arise naturally from the application of the algorithm itself. This, together with the computational efficiency of the method, allows the method to be used on whole proteins without first using separate procedures, such as definition of a core region, to restrict the search.

Rotamer entropies have previously been reported, but they have not been assessed as intrinsic measures of side chain predictability. Holm and Sander[59] interpreted nonzero side chain entropies (i.e., k^* values greater than unity) as evidence that their algorithm had not converged at residues exhibiting these values. Koehl and Delarue,[42] using a more physical potential function than ours, but perhaps a less physical notion of ensemble, framed their discussion in terms of the contribution of side chain configurational entropy to the free-energy of protein denaturation. By performing simulations of varying length, we demonstrated that entropy values are converged when the algorithm is carried out for sufficiently many steps; running longer will not lower the entropies further. The Monte-Carlo portion of the algorithm is the most time-consuming part of the method. If the user intends to ignore the entropic analysis, results nearly as good as the Monte-Carlo consensus results for the entire protein can be obtained merely from the simulated annealing runs at as little as one sixth the cost in computer time; however, even when reliable entropies are obtained, the computational cost of the procedure is modest.

Correlation of Rotamer Entropies With Solvent Accessibilities

We were surprised not to observe a strong correlation between rotamer entropy and solvent accessibility; on the whole, surface residues are nearly as immobile and are predicted nearly as accurately as buried residues. A given valine or threonine or χ_1 value is nearly as likely to be highly constrained—and hence to be predicted correctly—when it is on the surface as when it is in the interior. Koehl and Delarue[42] also observed that side chain entropy appears not to be correlated with solvent accessibility. We considered several possible explanations, some artifactual, for this observation. For example, we considered the possibility that some residues might be especially poorly fit by any rotamer from the library. For these residues, several rotamers , all making bad contacts, might occur in configurations having the same score. If this were the case, then we would expect to find high entropies at residues for which the BRC conformation fits the native conformation poorly. If buried residues were especially susceptible to this effect, then their conformational freedom would be exaggerated in our results. To test this hypothesis, we produced scatter plots (not shown) of k^* vs. the difference in B value between the BRC conformation and the native (nonrotamer) side chain, reasoning that where the BRC conformation is an especially poor approximation to the native side chain conformation, it ought to exhibit correspondingly more bumps. Such a plot, for either all residues of a given type or for only the deeply buried residues of a given type, exhibited no discernible correlation, thus ruling out this explanation.

It is also possible that a more realistic definition of the energy would improve the prediction of buried residues, only. This would mean that buried residues are, in fact, more constrained than surface residues, but by mechanisms not considered by our potential function. Against this possibility, we would point out that the prediction accuracies we report are respectable when compared with those achieved using more realistic potential functions (see below). Furthermore, if buried residues tend to be hydrophobic, then the protein interior ought to be better modeled than the surface by our method, which fails to consider polar effects in any detailed manner. Thus, the use of a more realistic potential function should, if anything, improve our predictions for surface residues more than for buried residues. This would strengthen, rather than weaken, the anomaly we observe.

Another possible explanation arises from the fact that the solvent-accessible area of a buried side chain is not necessarily a good measure of how much space there is around it. It may be that a similar measure, using a smaller solvent probe, or an entirely different method, such as one based on Voronoi polyhedra,[60] would show that some buried residues are surrounded by more free space than others. If so, the possibility remains that these residues would be the more mobile (higher-entropy) buried residues. If this were found to be the case, however, it would tend to rationalize, rather than refute, our observation that internal residues, on the whole, possess about as much conformational freedom as do surface residues.

Another consideration arises from the fact that we determine entropies only at the single-residue level. In the "independent residue" approximation, the overall configurational entropy is given by the sum of the entropies of the individual residues. There must, however, be correlations between the conformations of side chains close enough to interact. An example of such a correlation would be a situation in which residue 1 spends half its time in conformation A and the other half in conformation B, and likewise for residue 2, but in which rotamers 1A and 2A always occur together in low-energy configurations. Such correlations lower the entropy of the entire ensemble, but since our existing software does not keep track of pairwise or higher-order rotamer occurrences during the Monte-Carlo simulation, we cannot estimate this contribution. If such correlations were stronger in the interior of a protein than on the surface, this would constitute an additional restriction of the conformational freedom of the internal residues relative to those on the surface. Koehl and Delarue[42] tried to estimate this effect, and tentatively concluded that it was small. In principle, frequencies of rotamer pairs (in addition to those of individual rotamers) could be tallied during the Monte-Carlo phase of the simulation. This would allow the magnitude of this effect to be directly measured.

In a real protein in solution or in the crystal, the backbone is free to move to some extent. Backbone segments on the surface of a protein are believed to be more mobile than segments in the interior, both because of the relative absence of neighboring segments and also because loops and other regions of irregular secondary structure tend to occur on the surface. The extra mobility of backbone elements near the surface will contribute extra increments of fluctuation to the positions of surface side chain atoms, even if the side chains should be restricted to single rotamers. Beyond this, however, greater backbone flexibility near the surface could confer upon surface side chains an added tendency to populate multiple rotamer states. Our simulations would, however, be blind to this effect, since the procedure we have described holds all backbone atoms rigidly fixed.

Despite these many caveats, it appears likely that we are observing an effect that has a counterpart in real proteins. This hypothesis seems feasible, given the fact that our predictions are about as good for a surface valine or threonine, on average, as for a buried residue of the same type, and likewise for the pooled χ_1 data. If the surface residues that appear to our algorithm to be fixed were indeed mobile, we would not expect surface residues, on the whole, to be as predictable as buried residues; however, contingency-table analysis shows that they are indeed nearly as predictable. Note, however, that this does not imply that buried residues are, as a group, as bump-free as surface residues. When fractional solvent exposure is plotted against the *B* (bump) value of either the BRC conformation or the native conformation (not shown), the residues with the greatest *B* values do indeed exhibit the greatest tendency to be buried. However, when one moves away from the BRC, the increase in *B* for such residues is not significantly greater than that for surface residues; thus, selecting a non-BRC rotamer does not, on average, punish a buried residue more severely than it does a surface residue.

Solid-state dynamic NMR studies, though they have not been performed on large numbers of residues in large numbers of globular proteins, lend some support to the notion that conformational side chain mobility need not be correlated with exposed surface area. Keniry and colleagues[61] found, using ^{13}C NMR, that the terminal carbon in MET 55 of sperm-whale metmyoglobin (P.D.B.[53] entry 5mbn) exhibits dynamics that are "more solidlike" than those of the corresponding atom in MET 131. The Lee-Richards[57] ACCESS program indicates that for these two residues only the terminal carbon exhibits any solvent exposure at all. This atom is 10% exposed in MET 55 and 0.1% exposed in MET 131, based on a value of 74.8 Å^2 for standard-state accessibility. Thus, the more exposed methionine exhibits the more constrained motion. The average of the crystallographic temperature factors for the movable side chain atoms of these two residues follows the same trend.

This crystal structure was not a member of our test set; however, we later simulated it in order to compare our k^* results with the NMR results for these two residues. In order to do so, we first arbitrarily resolved the ambiguous atom-type designators in the P.D.B. file by converting all GLN AE1 and AE2 types to OE1 and NE2, respectively; similarly, all ASP AD1 and AD2 types were converted to OD1 and ND2, respectively. Our simulation of 5mbn gives 4.29 ± 0.04 and 5.83 ± 0.07 for k^* of MET 55 and MET 131, respectively. The error limits are standard deviations obtained from duplicate runs with different random seeds. The rotamer predicted from the Monte-Carlo consensus configuration was the BRC (native) rotamer in all simulations. These k^* values are in accord with both the NMR observations and the crystallographic temperature factors, and contrary to the expectation that the more buried residue should exhibit the lower mobility.

A single example such as this one can neither validate nor refute what amounts to a statistical statement; however, it does render feasible the notion that mobility need not correlate with solvent accessibility. In this regard, a recent NMR study[62] showed no correlation between residue surface accessibility and backbone amide-nitrogen mobility; on the other hand, another recent study[63] did show significant correlation between tryptophane side chain mobility and surface accessibility. We may perhaps look to future NMR studies for the further development of this story.

Comparison of Prediction Accuracy to That Reported for Other Methods

Several side chain prediction algorithms have been described in the literature. It is not always easy to compare published results, because different criteria are used to assess accuracy and different workers study different proteins. Some groups report results on only one or two structures; the data presented in Table I show, however, that results on single proteins can differ significantly. Often, only some of the side chains are modeled—for example, the core residues—whereas we present results on all movable residues. Different groups also use different criteria to judge success. One criterion sometimes used is to consider a χ angle to be correctly predicted if it falls within 30° or 40° of the dihedral angle found in the crystal structure; this criterion should map reasonably well onto our practice of considering a χ angle correctly predicted if it falls into the same generic class (gauche(−), gauche(+) or trans, for example) as the BRC value. The fraction of all dihedral (χ) angles correctly predicted has also been used. This can be an overly optimistic measure, because χ_2 may be counted as correctly predicted even if the χ_1 value of the same residue is not correctly predicted. The "correct" χ_2 value masks the fact that all side chain atoms of the residue are poorly placed. In the results we present, χ_2 is counted as correct only if the corresponding χ_1 is correct. Our own results are taken from the Monte-Carlo consensus results for whole proteins. For fraction of rotamers or χ angles predicted successfully, these are the values in the "M" columns of Table I when individual proteins are discussed and from the row labeled "Monte-Carlo Consensus" in Table VI when overall results are discussed.

In comparing RMS interatomic deviations from the native protein, bear in mind that not all authors follow our convention of excluding Cβ from the comparison. In addition, it is important to observe the distinction between average results for residues for a given protein and total results encompassing the same atoms. When the residue RMS deviations are averaged for a given protein, each residue has the same weight. In total figures, each atom (not each residue) has equal weight, so that larger residues have greater weight. Total values are generally greater than residue averages, since poorly predicted residues, such as arginine, lysine, and methionine, tend to have many atoms. Given these caveats, a summary of side chain prediction results and a comparison to ours is attempted here. The RMS interatomic displacement values we cite for our own work are taken from the third page of Table I and, as discussed earlier, are computed over side chain atoms, excluding Cβ.

The method presented here is most similar to the algorithms of Holm and Sander[40] and of Lee and Subbiah.[39] Holm and Sander developed a side chain construction algorithm in conjunction with their backbone-building algorithm[40] and applied it to homology model-built proteins with relatively good success.[59] Like us, they use a simulated-annealing search strategy to search conformations available in a rotamer library. They used the rotamer library of Tuffery et al.,[64] while we used the older Ponder and Richards library.[45] Calculations with a newer library similar to that of Tuffery (Fetrow and Libby, unpublished) gave little improvement over the results reported here. Holm and Sander applied a truncated 6–9 van der Waals potential to energy calculations. For each conformation, pairwise atomic interaction energies between atoms 6.0 Å or less apart were summed over the protein. In a set of 17 proteins, they computed an RMS difference between native and predicted structures of 1.56 Å for residues found in protein cores and 2.21 Å for all residues in entire proteins, with an average of 70.2% χ_1 values for all core residues. In our entire sample, the average overall RMS-deviation value was 2.03 Å and χ_1 was predicted correctly 74% of the time. This seems to indicate that our simple potential achieves results equal to or better than those obtained using a more physically realistic van der Waals potential. In application to homology modeling, Holm and Sander calculate entropies similar to the ones described in this paper. They interpret them, however, as measures of convergence of their algorithm. We showed that when simulations are run long enough that the entropies themselves converge to constant values, and when care is taken to sample only low-energy configurations during the determination of rotamer entropies, such values can be used to assess the reliability of prediction for each residue.

Lee and Subbiah[39] also used a simulated-annealing search to explore side chain conformations globally; however, they did not limit their search to conformations found in a rotamer library. In their method, χ angles are explored in 10° increments and the energy is calculated using a Lennard-Jones 6–12 potential. Because of the time required for the search, the method was applied only to a subset of residues in each of their test proteins. Prediction accuracies for whole proteins averaged 1.97 Å RMS interatomic deviation, including Cβ. This compares to our average overall value of 2.03 Å excluding Cβ. Since the difference between such values must be at least 5% (see the Methods section), our results appear comparable or a bit better. These researchers found arginine and lysine especially hard to predict, while hydrophobic residues were predicted with higher accuracy. These observations parallel our own. Because they explore a larger search space, this algorithm takes significantly more computer time than the one presented here. A medium-sized proteins (about 300 residues—about 230 moving residues) could not be searched in a reasonable time.

Another search strategy, termed self-consistent ensemble optimization, has recently been introduced by Lee.[43] This algorithm is not limited to rotamers, but searches side chain configurational space in 10° increments. Although the method was not applied to all side chains in one protein, it was successfully applied to the hydrophobic side chains in flavodoxin and to the hydrophobic core mutants of λ repressor. For 49 hydrophobic residues in flavodoxin, this algorithm yields a structure that exhibits a 0.99 Å RMS interatomic difference to the native structure, including Cβ. For the 57 residues comprising the lowest-entropy half of the side chains in this molecule, we found an overall RMS interatomic displacement of 1.02 Å. Again, given the minimum correction of 5% due to the inclusion of β atoms, our results appear comparable or perhaps a bit better. It should be mentioned that the 49 residues they quote results for, like our 57, are selected based on criteria internal to the algorithm.

The dead-end elimination algorithm was developed by Desmet et al.[41] to prune configuration space before beginning the search. These researchers reported a 71–72% prediction accuracy for χ_1 and χ_2 of all side chains in insulin and hemocyanin. Our overall χ_1 and χ_2 prediction accuracy was 63%; however, this figure ranges from 47% to 82% in individual proteins. Since insulin and hemocyanin are not in our database, specific results results cannot be directly compared. An improvement of this algorithm has recently been introduced[44]; however, it was not applied to any proteins, so its performance could not be evaluated.

Koehl and Delarue[42] used a self-consistent mean-field method to search conformation space using a standard rotamer library. They used a Lennard–Jones 6-12 potential to eliminate those combinations of rotamers that made bad van der Waals contact. This was followed by a general energy minimization of side chain conformations. For a set of 30 well-resolved proteins, this method yields an average of 72% correct predictions for χ^1 and 62% correct prediction for χ^1 and χ^2, roughly comparable to the 74% and 63%, which we present. For the seven proteins we studied in common with these workers, the average-residue RMS value was 1.2 Å, while ours was 1.3 Å, both groups excluding Cβ from the calculation. Thus, we appear to do a bit better than they using the χ value criterion and they appear to do a bit better than we do using the RMS criterion. Although these authors do not give timing data, energy minimization is likely to be more time-consuming than the procedure we describe.

Several groups have attempted to perform energy minimization on small clusters or cores of interacting residues.[33-37] Snow and Amzel[33] defined dependency sets for each residue and minimized each cluster in an attempt to predict changes caused by mutagenesis in an immunoglobulin structure. Schiffer and coworkers[35] start by placing side chains using a knowledge-based approach, but they then define clusters, which they call "molten zones," and minimize these zones independently. They also introduced a solvation term into their energy function, but they applied their method only to rat trypsin side chains on the structure of bovine trypsin. These papers involve building side chains onto a nonnative backbone, and are thus not directly comparable to our results.

Another energy-based rotamer search which includes a solvent term, and also allows the backbone to relax, was developed by Wilson and coworkers.[36] This improvement to the energy equation appears to improve prediction of accessible side chains, but results for only a few proteins were presented and results for individual amino acids were not presented. Using this algorithm, 88% of the buried residues, 67% of the exposed residues and 76% of all residues were predicted correctly at the rotamer level in α-lytic protease (2alp). For this protein, we achieve 69% correct prediction, overall. In r.m.s. interactomic displacement, these workers also do better than we do: for 2alp they achieve a 0.73 Å r.ms. displacement, on average, per residue and 1.31 Å RMS over all moving atoms; we achieve 1.03 and 1.83 Å RMS, respectively. Similarly, for 1ctf, they achieve 1.49 Å overall and we obtain a value of 1.86 Å. It is not clear from their paper whether they include Cβ in the calculation, but their results are certainly more accurate than ours. It is uncertain whether the solvation term or the ability to relax the backbone is responsible for more of the difference. Wilson and colleagues[36] find situations in which backbone relaxation is necessary if the correct rotamer is to be selected; however, such situations seem to be few.

Eisenmenger and associates[37] minimize each side chain individually in the context of backbone atoms, then combine individually minimized side chains in a global minimization protocol. Using a database of seven proteins, these researchers acheived an average of 74% (80%), 53% (65%), and 49% (57%) correct predictions for χ^1, χ^1 and χ^2, and all χ angles, respectively, where the numbers in parentheses were obtained using all atoms, backbone and side chain, in the minimization, while the other numbers were obtained using their so-called GAP method, which varies only backbone and Cβ atoms in the minimization. Our figures of 74%, 63%, and 59% are quite comparable with their better set of results. They report a 2.1 Å overall interatomic displacement for 4pti, compared with our 2.0 Å, and find a 1.5 Å displacement for buried residues of this protein, without saying what fraction of the total these comprise. For our lowest-entropy 50% of the residues, we observe a 1.6 Å RMS displacement. These results appear comparable; however, they do better than we for 1ubq, the other protein we both modeled in common: 1.4 Å and 0.5 Å RMS, overall and buried, compared to our 1.9 Å and 1.2 Å, overall and low-entropy half, respectively. It is interesting that relative results differ so greatly going from protein to protein. It is not known whether these authors included Cβ in the calculation.

Knowledge-based methods also produce comparable results. Reid and Thornton applied a knowledge-based method to homology modeling of flavodoxin[29], achieving an r.m.s side chain displacement of 2.41 Å. Our automatic method achieves a 2.1 Å value; however, this illustrates that hand-modeling is capable of reasonable results. In modeling rhizopuspepsin, Summers and Karplus[30] first place homologous atoms based on the protein of known structure, use a rigid rotor van der Waals approximation to place remaining atoms, then apply an energy minimization routine to relax atomic overlaps. Dunbrack and Karplus[31] developed a backbone-dependent side chain rotamer library and used this library to initially place side chains on the protein backbone. Side chains were then iteratively minimized and reoriented to eliminate van der Waals repulsions. For several proteins, we and they studied either different mutants or different chains; results can be directly compared only for rhizopuspepsin (2apr), pancreatic trypsin inhibitor (we studied 4pti; they studied 5pti), and ribonuclease (7rsa). We predict χ_1 correctly 81%, 83%, and 71% of the time, respectively, for these systems. Dunbrack and Karplus predict these values correctly 54%, 65%, and 56% of the time using a backbone-independent rotamer library and 82%, 85%, and 79% of the time using a backbone-dependent rotamer library. Our predictions, which use a backbone-independent rotamer library, are more successful, than theirs, when they use a similar library, and not quite as good as theirs when they use a backbone-dependent library. Their criterion for correctness was obtaining a χ angle within 40° of the native value, which, as mentioned earlier, should map reasonably well to our criterion. This comparison seems to bear out the assertion that, despite its simplicity, the simple energetic criterion and search strategy proposed here achieve results at least as good as more complicated and time-consuming methods. From the table they present of frequency and average RMS interatomic displacement of the various residue types, an average residue RMS interatomic displacement can be computed for the 635 nonproline residues in their sample. This value, 1.43 Å, is not quite as good as our comparable value of 1.34 Å. It is not clear whether they included Cβ in the computation. In any event, the data of Dunbrack and Karplus also suggest that a backbone-dependent rotamer library would improve the accuracy of any rotamer-based side chain modeling method, including ours.

Laughton[32] uses a database of side chain contacts to build side chains from comparable side chains found in the database. The complete structure is then energy-minimized. The proteins they, as well as we, modeled are 1ctf, 7rsa, 1lzl, and 4fxn. The average over these four proteins of the overall r.m.s. interatomic displacement of the side chain atoms is 1.95 Å; for the same proteins, our average value is 2.01 Å; however, they included Cβ in the computation, so, given the 5% minimum correction for this factor, our results would appear to be comparable or perhaps a bit better. For core residues, they obtain an average value over these proteins of 1.20 Å, whereas for our 50% low-entropy residues we obtain an average of 1.34 Å; however, aside from the Cβ correction, it is not clear what fraction of the total their core residues comprise. We appear to achieve accuracies similar to those of this group, without going through an energy-minimization step.

Tanimura and coworkers[65] compare several methods of side chain prediction using rotamer libraries of several sizes (101, 263, and 624 rotamers, total). We compare our results, using a rotamer library of 189 rotamers, total, with their results using the medium-size library. The dead-end elimination method[41] gives the best overall results with this library (and could not be carried out on large proteins with the large library). Their overall RMS interatomic displacement, over all side chain atoms in the 11 proteins they studied, is 1.76 Å using this method, and is 1.10 Å for core residues, excluding Cβ. Both χ_1 and χ_2 are predicted correctly 68% and 80% of the time for these two groups. We find overall interatomic displacements of 2.03 Å and 1.54 Å RMS, respectively, for all residues and for the 50% lowest-entropy residues, averaged over single proteins. We obtain both χ_1 and χ_2 correctly 63% of the time and 78% of the time, respectively, for these two groups. Since there is no clearly observable trend in accuracy as proteins increase in size (Table I), the overall RMS displacement results presented by these workers is probably comparable to our average over proteins. The results of Tanimura and coworkers are thus more accurate than ours, especially in terms of rms displacement. This cannot be due to a better fit of the rotamer library to the native side chains, since both their library and ours exhibit a BRC rms interatomic displacement of 0.76 Å. Their smaller library of 101 residues, which exhibits a BRC rms interatomic displacement of 0.88 Å, gives, using the dead-end method, results close to ours. It is uncertain whether the higher accuracy obtained by these workers is a due to the use of a full molecular mechanics forcefield or to the method used to place the rotamers. It should also be mentioned that Tanimura and colleagues used an all-hydrogen model of the protein, rather than a heavy-atom-only model, as we did. This probably contributes to the accuracy as well.

Vasquez[66] experimented with a variety of methods, all starting with a reference rotamer library about twice as large as ours, giving an RMS interatomic displacement in the BRC of 0.58 Å. He uses two basic algorithms (termed the heat-bath method and the mean-field method) to select rotamers, and follows this with a stage of energetic refinement in side chain torsion space. Throughout, he uses Lennard-Jones-like energy functions. The two basic algorithms perform essentially equally well, and are not improved materially by means of an intermediate "customization" step, in which the rotamer library for each residue is refined based on its local environment; thus, this procedure was discarded. His overall results over a test set of 30 proteins is an RMS interatomic displacement of 1.78 Å, excluding Cβ (1.53 Å including Cβ), compared to our value of 2.03 Å. For cores comprising 45% of the residues, his results improve to 1.14 Å RMS, compared to our value of 1.54 Å for the 50% low-entropy residues. His results are also more accurate than ours for χ angle predictions: χ_1 alone is predicted correctly about 81% of the time, compared to our 74%, and both χ_1 and χ_2 are predicted correctly 71% of the time, compared to our 63%. It is difficult to unravel the several factors which might be responsible for the greater accuracy of Vasquez's results, compared to ours. The larger rotamer library presumably contributes, and Vasquez states that postrefinement is also important. The heat-bath or mean-field methods are probably not an improvement over the simulated annealing method used here. Our observation that simulated annealing nearly always succeeds in finding the same minimal-B score indicates that, as both Vasquez and Eisenmember and colleagues[37] concluded, the combinatorial problem is not performance-limiting for side chain prediction. Interestingly, Vasquez observed, as we did, that the lowest-energy structures found by his algorithm always had lower energies than the BRC.

With the exception of the work of Tanimura[65] and Vasquez[66], whose results are comparable and more accurate than ours, the results presented here appear to be similar in accuracy to those presented by other workers. The factors responsible for the accuracy obtained by Tanimura and Vasquez appear to be distinct. Tanimura's improved accuracy is probably due to the use of a full molecular-mechanics energy function and/or the inclusion of explicit hydrogens; it cannot be due to his use of an intrinsically better rotamer library, since the BRC does not fit the native structure any better for his library than for ours. Vasquez uses neither explicit hydrogens nor a full molecular-mechanics energy function, but he does employ a larger and intrinsically better-fitting rotamer library than other workers (except for the large-library experiments of Tanimura, which we have not discussed). He also performs a postplacement refinement of the side chain χ angles. Though the work of Vasquez, like ours and that of most other workers, achieves success without considering polar interactions, underscoring the importance of packing in the determination of side chain conformation, careful examination of results with polar groups (especially ASN, GLN, and HIS) make it clear that using a more realistic energy function must improve such predictions. Be this as it may, the entropy-based selectivity criteria discussed here would probably continue be useful in conjunction with bettter rotamer libraries, force-fields and placement algorithms.

Limitations of and Possible Improvements to the Algorithm

The very use of a rotamer library is a simplification which allows a large region of configuration space to be explored in a finite amount of time, at some sacrifice of accuracy in atomic position. In 1987, Ponder and Richards[45] presented evidence that side chains largely fall into a limited number of rotamer categories. This library was examined and extended by Tuffery and colleagues[64] and Thornton and coworkers presented evidence that side chain conformations become more tightly clustered with higher structure resolution.[67] However, Schrauber and colleagues[58] present convincing evidence that 5% to 30% (depending on amino acid type) of all amino acid conformations do not fall into clear rotamer categories, even in very high-quality crystal structures, calling into question the rotamer-library approximation for conformational searching. So, can side chain conformations be adequately searched using a rotamer library? Data presented here and elsewhere suggest that, for the vast majority of side chains, a rotamer library is sufficient for an initial side chain conformational search, as long as the model protein is subjected to energy minimization or dynamics following the conformational search to relieve further atomic overlaps and strain imposed by the limited rotamer libraries.[31,42,48,49] Rotamer library searching is not adequate for drug or ligand design, where atomic resolution is critical. Further techniques for thoroughly searching conformational space in critical parts of a protein remain to be further developed; however, techniques such as those presented here can be used for initial side chain placement before further refinement[49], as is also apparent from the work of Vasquez.[66]

A second limitation of the present work is the extremely simple energy function used to distinguish among alternate conformations. Koehl and Delarue[42] mention briefly that they experimented with a simple on/off atomic overlap, but that this gave results somewhat worse than those obtained using a Lennard-Jones potential. They also mention that by increasing the atomic radii, they achieved worse results, whereas by decreasing the radii they obtained slightly better results using this simplified potential. On the surface, these results seem in direct contrast to ours, in which a simplified on/off potential was applied and results comparable to those obtained with other methods were achieved using atomic radii somewhat greater than normal. However, our results, like theirs, suggest that a large fraction of side chain packing interactions can be ruled out based on steric (nonpolar) factors. We obtain results comparable to most results found using more complicated potentials, demonstrating that the simple model described here captures much of the physical reality mitigating side chain conformation. On the other hand, the excellent results of Tanimura[65] may, at least in part, be due to his use of additional force-field terms.

A major benefit of our method is computational efficiency coupled with reasonably accurate prediction results of all side chains in the protein of interest. We would like to improve the prediction results without significantly slowing the computation time. Several such improvements can be imagined. First, a rotamer library based on the most recent structures from the Brookhaven database[53] could be used. Preliminary results suggest, however, that this change makes little difference in prediction accuracies (Farid and Shenkin, unpublished). However, it has been shown[31,68] that side chain dihedral preferences correlate with main chain conformations. Comparison of our data with those of Dunbrack and Karplus[31] strongly suggests that incorporating secondary structure-specific rotamers into the library would yield significant improvement (see above). It is worth recalling that our algorithm keeps the backbone fixed; thus, the program could itself determine the nature of the secondary structure in the vicinity of each residue and select the appropriate rotamer set accordingly.

Currently, atomic overlaps or bumps are treated in a simple manner, and changes here might result in better prediction accuracies. For instance, moderate and severe atomic overlaps are now counted equally; furthermore, distances between hydrogen-bonded atoms can be smaller than those that are not hydrogen bonded. Thus, it might be more realistic to weight large atomic overlaps more heavily than small ones or to allow oxygen and nitrogen atoms to approach each other more closely than other atoms. It might be worth weighting backbone-side chain bumps differently from side chain-side chain overlaps. However, comparison of our data to that of other researchers suggests that most of these changes are only likely to result in small, if any, visible difference in the prediction accuracies. Certainly, however, addition of terms to account for polar effects can only improve prediction accuracy for glutamine, asparagine and histidine residues. Even in the context of a pure packing model, incorporation of explicit hydrogens might improve results.

The work of Wilson and colleagues,36, Schiffer and colleagues,35, and Snow and Amzel[33] suggests that solvent effects can improve the prediction of surface residues. Side chain conformations could first be positioned using the algorithm described here, then the conformations of those residues that are solvent accessible could be further searched with a solvation term included in the energy calculation. However, the utility of predicting charged or polar surface residues, which may not be well determined in crystal structures and may adopt multiple conformations in solution, is questionable.

Applications of the Method

The question of side chain entropy and its relation to local environment deserves further investigation. As mentioned earlier, tallying the pairwise as well as individual rotamer occurrences during the low-temperature Monte-Carlo process would allow the effect of correlation to be directly assessed. Detailed comparison with multiply occupied crystallographic positions and with dynamic NMR studies are in principle possible; also, if correlated side chain conformation is a feature of real proteins, taking correlation into account in the prediction algorithm ought to improve the overall prediction accuracy.

The best test of a side chain conformational search algorithm is its application to real homology modeling and inverse folding problems, because small shifts in backbone conformation may cause drastic changes in the distribution of side chain conformations. On the other hand, because of its speed, the method described here can be easily be applied to a variety of problems. This algorithm is currently being applied to real modeling problems: to a homology model of galactose repressor sugar-binding domain,[48] to a model of fasciclin III neural adhesion protein whose sequence-to-structure alignment was determined by the inverse folding algorithm of Bryant and Lawrence,[50] to a homology model of two human papillomavirus E2 transcriptional regulatory proteins (Altman, Brenowitz, and Fetrow, unpublished), to modeling of side chains in protein loops, and to side chain dressing of de novo-generated backbone structures in fundamental studies of protein folding.[49] Comparison of predicted models to experimental results, such as actual crystal structures, will provide better tests of the accuracy of this approach.

ACKNOWLEDGMENTS

This work was initiated with the help of Jonathan Greene, now at CombiChem, Inc., and his contribution is gratefully acknowledged. The authors would like to thank Ann McDermott, George Berg, Brian Kell, Laurie Castonguay, and Bill Rennie for stimulating discussions during the course of this work. This work was supported by NIH grant GM44336 to PSS, when at Barnard College, and by a University at Albany Faculty Research Award to JSF. Support to P.S.S. at Columbia University was provided, in part, by NIH grant P41-RR06892.

REFERENCES

1. Chou, P.Y., Fasman, G.D. Prediction of protein conformation. Biochem. 13:222–245, 1974.
2. Garnier, J., Osguthorpe, D.J., Robson, B. Analysis of accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1987.
3. Qian, N., Sejnowski, T.J. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol 202:865–884, 1988.
4. Holley, L.H. Karplus, M. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. USA 86:152–156, 1989.
5. Kneller, D.G., Cohen, F.E., Langridge, R. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214:171–182, 1990.
6. Kneller, D.G., Cohen, F.E., Langridge, R. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214:171–182, 1990.
7. Cohen, F.E., Richmond, T.J., Richards, F.M. Protein folding: Evaluation of some simple rules for assembly of helices into tertiary structures with myoglobin as an example. J. Mol. Biol. 132:275–288, 1979.
8. Hurle, M.R., Matthews, C.R., Cohen, F.E., Kuntz, I.D., Toumadge, A., Johnson Jr., W.C. Prediction of the tertiary structure of the α-subunit of tryptophan synthase. Proteins 2:210–224, 1987.
9. Greer, J. Comparative model-building of the mammalian serine proteases. J. Mol. Biol. 153:1027–1042, 1981.
10. Greer, J. Comparative modeling of homologous proteins. Methods Enzymol. 202:239–252, 1991.
11. Shenkin, P.S., Yarmush, D.L., Fine, R.M., Wang, H., Levinthal, C. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 26:2053–2085, 1987.
12. Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L., Levinthal, C. Predicting antibody hypervariable loop conformations. II. Minimization and molecular dynamics studies of MCPC603 from randomly generated loop conformations. Proteins 1:342–362, 1986.
13. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5:819–822, 1986.
14. Bruccoleri, R.E., Haber, E., Novotny, J. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. Nature 335:564–568, 1988.
15. Chothia, C., Lesk, A.M., Tramontano, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., Colman, P.M., Spinelli, S., Alzari, P.M., Poljak, R.J. Conformations of immunoglobulin hypervariable regions. Nature 342:877–883, 1989.
16. Martin, A.C.R., Cheetham, J.C., Rees, A.R. Modeling antibody hypervariable loops, a combined algorithm. Proc. Natl. Acad. Sci. USA 86:9268–9272, 1989.
17. Higo, J., Collura, V., Garnier, J. Development of an extended simulated annealing method: Application to the modeling of complementary determining regions of immunoglobulins. Biopolymers 32:33–43, 1992.
18. Palmer, K.A., Scheraga, H.A. Standard-geometry chains fitted to x-ray derived structures: validation of the rigid-geometry approximation. II. Systematic searches for short loops in proteins: Applications to bovine pancreatic ribonuclease A and human lysozyme. J. Comp. Chem. 13:329–350, 1992.
19. Zheng, Q., Rosenfeld, R., Vajda, S., DeLisi, C. Determining protein loop conformation using scaling-relaxations techniques. Protein Sci. 2:1242–1248, 1993.
20. Collura, V., Higo, J., Garnier, J. Modeling of protein loops by simulated annealing. Protein Sci. 2:1502–1510, 1993.
21. Smith, K.C., Honig, B. Evaluation of the conformational free energies of loops in proteins. Proteins 18:119–132, 1994.
22. Bowie, J.U., Luthy, R., Eisenberg, D. A method to identify proteins sequences that fold into a known three-dimensional structure. Science 253:164–170, 1991.
23. Blundell, T.L., Sibanda, B.L., Sternberg, M.J., Thronton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326:347–352, 1987.
24. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through folding motif. Proteins 1993.
25. Sippl, M.J., Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271, 1992.
26. Fetrow, J.S., Bryant, S.H. New programs for protein tertiary structure prediction. Bio/Technology 11:479–484, 1993.
27. Wodak, S.J., Rooman, M.J. Generating and testing protein folds. Curr. Opin. Struct. Biol. 3:247–259, 1993.
28. Srinivasan, R., Rose, G.D. LINUS: A hierarchic procedure to predict the fold of a protein. Proteins 22:81–99, 1995.
29. Reid, L.S., Thornton, J.M. Rebuilding flavodoxin from Ca coordinates: A test study. Proteins 5:170–182, 1989.
30. Summers, N.L., Karplus, M. Construction of side chains in homology modeling: Application to the C-terminal lobe of Rhizopuspepsin. J. Mol. Biol. 210:785–811, 1989.
31. Dunbrack Jr., R.L., Karplus, M. Backbone-dependent rotamer library for proteins: Application to side chain prediction. J. Mol. Biol. 230:543–574, 1993.
32. Laughton, C.A. Prediction of protein side chain conformations from local three-dimensional homology relationships. J. Mol. Biol. 235:1088–1097, 1994.
33. Snow, M.E., Amzel, L.M. Calculating three-dimensional changes in protein structure due to amino acid substitutions: The variable region of immunoglobulins. Proteins 1:267–279, 1986.
34. Bruccoleri, R.E., Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.
35. Schiffer, C.A., Caldwell, J.W., Kollman, P.A., Stroud, R.M. Prediction of homologous protein structures based on con-

formational searches and energetics. Proteins 8:30–43, 1990.

36. Wilson, C., Gregoret, L.M., Agard, D.A. Modeling side chain conformation for homologous proteins using an energy-based rotamer search. J. Mol. Biol. 229:996–1006, 1993.

37. Eisenmenger, F., Argos, P., Abagyan, R. A method to configure protein side chains from main-chain trace in homology modelling. J. Mol. Biol. 231:849–860, 1993.

38. Tuffery, P., Etchebest, C., Hazout, S., Lavery, R. A critical comparison of search algorithms applied to the optimization of protein side chain conformations. J. Comp. Chem. 14:790–798, 1993.

39. Lee, C., Subbiah, S. Prediction of protein side chain conformation by packing optimization. J. Mol. Biol. 217:373–388, 1991.

40. Holm, L., Sander, C. Database algorithm for generating protein backbone and side chain coordinates from a Ca trace: Application to model building and detection of coordinate errors. J. Mol. Biol. 218:183–194, 1991.

41. Desmet, J., DeMaeyer, J., Hazes, B., Lasters, I. The deadend elimination theorem and its use in protein side chain positioning. Nature 356:539–542, 1992.

42. Koehl, P., Delarue, M. Application of a self-consistent mean field theory to predict protein side chains conformation and estimate their conformational entropy. J. Mol. Biol. 239:249–275, 1994.

43. Lee, C. Predicting protein mutant energetics by self-consistent ensemble optimization. J. Mol. Biol. 236:918–939, 1994.

44. Goldstein, R.F. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys. J. 66:1335–1340, 1994.

45. 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. 193:775–791, 1987.

46. Kirkpatrick, S., Gelatt, C.D., Becchi, M.P. Optimization by simulated annealing. Science 220:671–680, 1983.

47. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.J. Equation of state calculations by fast computing machines. Chem. Phys. 21:1087–1089, 1953.

48. Hsieh, M., Hensley, P., Brenowitz, M., Fetrow J.S. A molecular model of the inducer binding domain of the galactose repressor of *Escherichia coli.* J. Biol. Chem. 269: 13825–13835, 1994.

49. Monge, A., Lathrop, E.J.P., Gunn, J.R., Shenkin, P.S., Friesner, R.A. Computer modeling of folding: Conformational and energetic analysis of reduced and detailed protein models. J. Mol. Biol. 247:995–1012, 1995.

50. Castonguay, L.A., Bryant, S.H., Snow, P.M., Fetrow, J.S. A proposed structural model of domain 1 of fasciclin III neural cell adhesion protein based on an inverse folding algorithm. Prot. Sci. 4:472–483, 1995.

51. Greene, J. Shenkin, P.S. Can Rotamer Libraries be Used to Predict Side-chain conformations in Globular Proteins? Presented at the Symposium on "The Process of Protein Folding," American Association for the Advancement of Science, San Francisco, CA January, 1989.

52. Farid, H., Shenkin, P.S., Greene, J., Fetrow, J. Prediction of side chain conformations in protein cores and loops from rotamer libraries. Biophys. J. 61:A350, 1992.

53. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J.C. Protein data bank. 1987. In: Allen, F.H., Bergerhoff, T., Sievers, R. (eds.):"Crystallographic Databases: Information Content, Software Systems, Scientific Applications." Bonn: International Union of Crystallography, 1987:107–132.

54. Press, W.H., Flannery, B.R., Teukolsky, S.A., Vetterling, W.T. "Numerical Recipes." Cambridge: Cambridge University Press, 1986.

55. Swanson, R.M. Entropy measures amount of choice. J. Chem. Educ. 67:206–208, 1990.

56. Shenkin, P.S., Erman, B. Mastrandrea, L.D. Information-theoretical entropy as a measure of sequence variability. Proteins 11:297–313, 1991.

57. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55: 379–400, 1971.

58. Schrauber, H., Eisenhaber, F., Argos, P. Rotamers: To be or not to be? An analysis of amino acid side chain conformations in globular proteins. J. Mol. Biol. 230:592–612, 1993.

59. Holm, L., Sander, C. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. Proteins 14:213–223, 1992.

60. Richards, F.M. The interpretation of protein structures: Total volume, group volume distributions, and packing density. J. Mol. Biol. 82:1–14, 1974.

61. Keniry, M.A., Rothgeb, T.M., Smith, R.L., Gutowsky, H.S., Oldfield, E. Nuclear magnetic resonance studies of amino acids and proteins: Side-chain mobility of methionine in the crysalline amino acid and in crystalline sperm whale (*Physeter catodon*) myoglobin. Biochemistry 22:1917–1926, 1983.

62. Ishima, R., Nagayama, K. Protein backbone dynamics revealed by quasi spectral density function analysis of amide N-15 nuclei. Biochemistry 34:3162–3173, 1995.

63. Mandel, A.M., Akke, M. Palmer, A.G. III. Backbone dynamics of *Escherichia coli* ribonuclease HI: Correlations with structure and function in an active enzyme. J. Mol. Biol. 246:144–163, 1995.

64. Tuffery, P., Etchebest, C., Hazout, S., Lavery, R. A new approach to the rapid determination of side-chain conformations. J. Biomol. Struct. Dynam. 8:1267–1289, 1986.

65. Tanimura, R., Kidera, A. Nakamura, H. Determinants of protein side-chain packing. Prot Sci 3:2358–2365, 1994.

66. Vasquez, M. An evaluation of discrete and continuum search techniques for conformational analysis of side chains in proteins. Biopolymers 36:53–70, 1995.

67. Morris, A.L., MacArthur, M.W., Hutchinson, E.G., Thornton, J.M. Stereochemical quality of protein structure coordinates. Proteins 12:345–364, 1992.

68. Janin, J., Wodak, S., Levitt. M., Maigret, B. Conformation of amino acid side chains in proteins. J. Mol. Biol. 125: 357–386, 1978.