JCTC Journal of Chemical Theory and Computation

A Statistical Framework for Hierarchical Methods in Molecular Simulation and Design

David F. Green*

Department of Applied Mathematics and Statistics and Graduate Program in Biochemistry and Structural Biology, Stony Brook University, Stony Brook, New York 11794-3600

Received August 24, 2009

Abstract: A statistical framework for performance analysis in hierarchical methods is described, with a focus on applications in molecular design. A theory is derived from statistical principles, describing the relationships between the results of each hierarchical level by a functional correlation and an error model for how values are distributed around the correlation curve. Two key measures are then defined for evaluating a hierarchical approach—completeness and excess cost—conceptually similar to the sensitivity and specificity of dichotomous prediction methods. We demonstrate the use of this method using a simple model problem in conformational search, refining the results of an *in vacuo* search of glucose conformations with a continuum solvent model. Second, we show the usefulness of this approach when structural hierarchies are used to efficiently make use of large rotamer libraries with the Dead-end Elimination and A* algorithms for protein design. The framework described is applicable not only to the specific examples given but to any problem in molecular simulation or design that involves a hierarchical approach.

1. Introduction

The ability to efficiently and robustly design molecules with particular characteristics is an objective targeted by many disciplines. Pharmaceutical development is largely a problem of designing a molecule that interacts specifically with a given protein target while maintaining additional characteristics (such as solubility in water, the ability to diffuse across plasma membranes, and stability to the varied environments of the body). Many applications in biotechnology similarly involve the development of proteins (most often through the modification of an existing sequence) that interact with specific targets or that catalyze particular reactions. Additional applications of molecular design abound, including materials engineering, nanotechnology, and catalyst development. While differing in the details of the design goal, all of these applications share the basic necessity of searching extremely large spaces of chemical and structural variation for individual molecular structures with the desired properties. In many cases, accurate computational models are available for the evaluation of individual molecules, but the immense size of the search spaces involved requires tradeoffs be made between computational efficiency and predictive accuracy.

Numerous biomedical and biotechnological applications have demonstrated a need for the ability to design new or modified proteins with particular stability and interaction properties.[1–3] It has been recognized that, by phrasing an inverse problem to the traditional protein-folding problem, great progress can be made in this area of practical protein design.[4,5] Typically in these approaches, a target backbone structure is chosen, and then a set of amino acid side-chain arrangements is selected to stabilize the target structure.[6] However, even given a fixed sequence length and backbone structure, the space of possible sequences to search is still immense. For a sequence of N residues, there are 20^N possible sequences—with as few as 10 variable residues, this is a space of over 10^{12} possibilities. When structural variability of each sequence is included (for example, by treating each residue as a set of commonly observed rotameric states), this complexity can easily grow to upward of 10^{26} (400 choices at 10 positions), and for larger designs, the search space can be greater than 10^{100}.

* Author phone: (631) 632-9344, fax: (631) 632-8490, e-mail: david.green@stonybrook.edu.

10.1021/ct9004504
Published on Web 04/09/2010

Ligand docking—a key method for virtual high-throughput screening of pharmaceutical lead compounds—also involves a search over a huge structural space.[7,8] The size of the chemical space is essentially unbounded (even when "drug-like" restrictions are placed in molecular selection), and libraries of candidate molecules may easily contain hundreds of thousands of molecules. The docking process, which must be done on each molecule in a library, requires searching over six global degrees of freedom (three translational and three rotational) as well as any internal flexibility in the molecule and protein.[9,10] Again, even small systems can easily involve 10^{12} or more individual states.

Algorithmic advances, coupled with the rise in available computing power, have made these search problems tractable.[6,9] However, simplified descriptions of structural energetics are generally required for a number of reasons. In some cases, the large number of energetic evaluations required demands simplifications purely for computational tractability; for other methods, the algorithm itself requires an energetic description with particular properties (such as being decomposable into a sum of pairwise interactions between atoms or groups of atoms).

In parallel with developments in algorithms for searching large spaces of chemical and structural diversity, methodological advances have shown a remarkable ability to reproduce important experimental values, such as binding free energies and the effects of mutation on protein stability. However, these methods can be costly—free energy perturbation simulations using explicit solvent models being a perfect example.[11–13] Approximations can be made for efficiency, but with costs in accuracy. The frequently used Generalized-Born (GB) solvation model, for example, is orders of magnitude faster than explicit solvent simulation but suffers from well-characterized inaccuracies.[14–17]

Thus, while the power of computational approaches has developed strongly, there remains a fundamental trade-off that must often be made between the accuracy of the model used and the ability to sample an adequate space to solve the problem at hand. One solution to this dilemma is the use of a hierarchy of models. An inexpensive (but relatively inaccurate) model may be used for an initial search, and top ranking solutions from this search may be passed on to a more expensive, but more accurate, treatment. This procedure may be repeated, with successively increasing accuracy and expense in the models used. For most applications, a final level of the hierarchy is represented by experimental testing and validation.

Conceptually, the hierarchical approach is simple and has been applied in numerous applications.[18–21] However, there remains one important issue with a hierarchical approach—how do the cutoffs chosen at each level of the hierarchy affect the final results? Here, we present a statistical framework to help in answering this question. First, we describe the underlying statistics that describe the transfer of distributions with varying cutoffs. We use this framework to define two key descriptors of the efficacy of a hierarchical procedure: the completeness of the final set of results and the excess work done in calculations ultimately excluded from this set. The applications of this method is then outlined using a simple problem involving a conformational search with and without consideration of the solvent. Finally, these measures are used to consider the performance of hierarchical methods for an example application in protein design.

2. Theory

Here, we outline the fundamental statistical theory that can be used to characterize the performance of hierarchical methods. In this context, two levels of a hierarchy may be considered to be two energetic models for the same system; the levels may differ in the level of structural detail or in the Hamiltonian used. For example, levels may involve coarse-grained and fully atomistic descriptions of molecular structure, molecular mechanical and quantum mechanical Hamiltonians, or various treatments of the solvent. At each level, a "state" denotes an entity that can be associated with a single energetic value; these may be true structural microstates (such as a single conformation of a molecule) or an ensemble of microstates described by a free energy. A key requirement, however, is that the set of states at one level of the hierarchy be uniquely mapped to equivalent states at the next level.

We begin with a detailed outline of the theory; this section is purposefully constructed to be very general in scope and is thus somewhat abstract in its presentation. However, it may help the reader to consider that the ultimate goal in applying these methods will be to derive system-dependent functional relationships between hierarchical levels, and to use these to gain insight into the real-world performance of these approaches. Such applications are discussed in more detail in the Results and Discussion section. It should also be noted that this section is written using a formalism of continuous probability distributions; in many applications, including those presented as examples, the distributions will involve a finite number of discrete states. While a discrete variation of this theory can be easily derived (essentially by replacing all integrals by the appropriate sums), it is not clear that there is a significant motivation to do so.

Relating an Ensemble between Two Models. Consider an ensemble of states in a reference model, $\{\mathrm{E}^0\}$, and in a target model, $\{E^1\}$, where the distribution of states in the reference model is given by $P(E^0)$. If the energies in model 1 are correlated with those in model 0, then for all states with a given energy in model 0 ($\{E_i^0\}$), the energies of those states in model 1 ($\{E_i^1\}$) will distributed according to some error model, $P(E_i^1)$, about an expectation energy of $\bar{E}_i^1$. The expected energy in model 1 should be related to the energy in model 0 by $\bar{E}_i^1 = f(E_i^0)$, where $f(x)$ is a monotonically varying function. Similarly, with no loss of generality, we may assume an arbitrary error model, written as $P(E_i^1) = g(E_i^1, \bar{E}_i^1, \vec{\mu}_i)$, where $g(x, \bar{x}, \vec{\mu})$ is some general probability distribution, described by one or more parameters, $\vec{\mu}$. These parameters may include the standard deviation, higher-order moments of the distribution, or other descriptors. Since the expected energy varies with the reference energy, as may the parameters of the error model, we write:

$$P(E^1|E^0) = g(E^1, f(E^0), \vec{\mu}(E^0)) \quad (1)$$

The probability of a given (E^1, E^0) pair is then $P(E^1|E^0)P(E^0)$ or

$$P(E^1, E^0) = g(E^1, f(E^0), \vec{\mu}(E^0))P(E^0) \tag{2}$$

and the probability of any E^1 over the full distribution of $\{E^0\}$ is then given by

$$P(E^1) = \int_{-\infty}^{\infty} g(E^1, f(E^0), \vec{\mu}(E^0))P(E^0)\, dE^0 \tag{3}$$

requiring knowledge only of the reference distribution, $P(E^0)$, and the forms of $f(x)$, $g(x)$, and $\vec{\mu}(x)$. $P(E^0)$ may be considered to be either a normalized probability distribution or an unnormalized distribution of states. This choice does not affect the generality of the discussion but does fix the interpretation of $P(E^i)$ for all levels, i.

Multiple Levels of Propagation. Using the distribution $\{E^1\}$ as a reference for a target distribution $\{E^2\}$ and an analogous logic gives

$$\begin{aligned} P(E^2) &= \int_{-\infty}^{\infty} P(E^1)\, g_1(E^2, f_1(E^1), \vec{\mu}_1(E^1))\, dE^1 \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} P(E^0)\, g_0(E^1, f_0(E^0), \vec{\mu}_0(E^0))\, dE^0 \right) \times \\ &\quad g_1(E^2, f_1(E^1), \vec{\mu}_1(E^1))\, dE^1 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(E^0)\, g_0(E^1, f_0(E^0), \vec{\mu}_0(E^0)) \times \\ &\quad g_1(E^2, f_1(E^1), \vec{\mu}_1(E^1))\, dE^0\, dE^1 \end{aligned} \tag{4}$$

This may be extended to any number of levels of propagation:

$$\begin{aligned} P(E^N) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(E^0) \times \\ &\quad \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dE^0\, dE^1 \ldots dE^{N-1} \\ &= \int_V P(E^0) \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dV \end{aligned} \tag{5}$$

where the integral $\int_V dV$ is taken from $(-\infty,\infty)$ over all dimensions E^i with i in the range $[0, N-1]$.

Defining Ensemble Completeness. Equation 5 defines the propagation of entire ensembles from one hierarchical level through another. In practice, however, only a sampling of each distribution will be passed on from one level to the next. One ensemble of particular interest is that of the low-energy states at the highest level of the hierarchy, $\{E^N | E^N < E^N_{\text{cut}}\}$. From the full distribution, $P(E^N)$, the size of this ensemble is given by

$$\begin{aligned} D(E^N_{\text{cut}}) &= \int_{-\infty}^{E^N_{\text{cut}}} P(E^N)\, dE^N \\ &= \int_{-\infty}^{E^N_{\text{cut}}} \int_V P(E^0) \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dV\, dE^N \end{aligned} \tag{6}$$

Now, consider what happens if each level of the hierarchy is truncated at some maximal value, E^i_{cut}. In this case, the propagation of the integrals is not done over the full space of $(-\infty,\infty)$ for each dimension. Rather, for each dimension i, the integral is taken over the range $(-\infty, E^i_{\text{cut}})$:

$$\begin{aligned} P'(E^N) &= \int_{-\infty}^{E^{N-1}_{\text{cut}}} \int_{-\infty}^{E^{N-2}_{\text{cut}}} \cdots \int_{-\infty}^{E^0_{\text{cut}}} P(E^0) \times \\ &\quad \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dE^0\, dE^1 \ldots dE^{N-1} \\ &= \int_{V'} P(E^0) \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dV \end{aligned} \tag{7}$$

where V' represents the reduced space due to truncation of each distribution. Note that $P'(E^N)$ is not normalized, and thus its integral over all energies is less than that of $P(E^N)$. Given this distribution, the total size of the ensemble of interest is given by:

$$\begin{aligned} D'(E^N_{\text{cut}}) &= \int_{-\infty}^{E^N_{\text{cut}}} P'(E^N)\, dE^N \\ &= \int_{-\infty}^{E^N_{\text{cut}}} \int_{V'} P(E^0) \times \\ &\quad \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dV\, dE^N \end{aligned} \tag{8}$$

Given the size of the complete ensemble and that of the ensemble obtained through subsequent levels of truncation, we may define the ensemble completeness to be

$$C(E^N_{\text{cut}}) = \frac{D'(E^N_{\text{cut}})}{D(E^N_{\text{cut}})} \tag{9}$$

$C(E^N_{\text{cut}})$ gives the fraction of the complete low energy ensemble that is propagated through the truncated levels of the hierarchy. The completeness is analogous to the sensitivity (the true positive rate) of a dichotomous prediction method.

While $C(E^N_{\text{cut}})$ describes the completeness of the final ensemble, a second expression can be defined to describe the "excess work" required to obtain that level of completeness. At a given level in the hierarchy, the total size of the distribution carried through is given by

$$\begin{aligned} D''(E^N) &= \int_{-\infty}^{\infty} P'(E^N)\, dE^N \\ &= \int_{-\infty}^{\infty} \int_{V'} P(E^0) \times \\ &\quad \prod_{i=0}^{N-1} g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dV\, dE^N \end{aligned} \tag{10}$$

However, only $D'(E^N_{\text{cut}})$ of this is valuable—either for carrying on to the next level of the hierarchy or for inclusion in the final ensemble. Thus, we define the "excess work" to be

$$X(E^N_{\text{cut}}) = \frac{D''(E^N) - D'(E^N_{\text{cut}})}{D'(E^N_{\text{cut}})} \tag{11}$$

That is, the excess work is the relative amount of time spent evaluating "kept" and "discarded" states. $X(E^N_{\text{cut}})$ may be arbitrarily large but will equal 0.0 if no false positives are carried along and will equal 1.0 if equal numbers of false and true positives are found. The excess work is related to the false-positive rate, or specificity, of a dichotomous

prediction method but is scaled relative to the true-positive rate to account for cost. Note that this definition of excess work does not directly relate to a computational cost estimate, as there is no explicit consideration of the relative expense of each step. An ideal search method, applied directly to the final energetic ensemble, would require only evaluation of the energies of the lowest energy states; the excess work describes how much extra effort must be put into evaluation of energies at the higher hierarchical level, compared to this ideal. Note that this measure does not consider the cost of the search at the lower level.

Propagation in Normal Error Models. Consider the case where $\bar{E}^1$ is linearly correlated with E^0 ($\bar{E}^1 = m_1E^0 + b_1$), and the error distribution of $\{E_i^1\}$ about $\bar{E}_i^1$ is a normal distribution with constant variance ($\vec{\mu}_1 = \{\sigma_1\}$, independent of E^0). Thus, eq 1 becomes

$$P(E^1|E^0) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{\frac{-(E^1 - (m_1E^0 + b_1))^2}{2(\sigma_1)^2}} \tag{12}$$

If $\{E^0\}$ is normally distributed about E^0 with a standard deviation of σ_0, then eq 2 becomes:

$$P(E^1, E^0) = \frac{1}{(2\pi)\sigma_0\sigma_1} e^{\frac{-(E^1 - (m_1E^0 + b_1))^2}{2(\sigma_1)^2} + \frac{-(E^0 - \bar{E}^0)^2}{2(\sigma_0)^2}} \tag{13}$$

The zero points of the energy distributions are arbitrary, and thus we may make the simplifying assumption of $b_1 = 0$ and $E^0 = 0.0$. Furthermore, a correlation with slope m and variance σ is equivalent to a correlation with unit slope and variance $\sigma' = \sigma/m$, allowing the further simplification of $m_1 = 1.0$. This gives

$$P(E^1, E^0) = \frac{1}{(2\pi)\sigma_0\sigma_1'} e^{\frac{-(E^1 - E^0)^2}{2(\sigma_1')^2} + \frac{-(E^0)^2}{2(\sigma_0)^2}} \tag{14}$$

which integrates (eq 3) to

$$P(E^1) = \frac{1}{\sqrt{2\pi}\sqrt{(\sigma_0)^2 + (\sigma_1')^2}} e^{\frac{-(E^1)^2}{2((\sigma_0)^2 + (\sigma_1')^2)}} \tag{15}$$

$P(\mathrm{E}^1)$ is a normal distribution, with variance $(\sigma_0)^2 + (\sigma_1')^2$. This can be extended simply through multiple levels to give $P(E^N)$ as a normal distribution with variance $\sum_{i=0}^{N}(\sigma_i')^2$. Truncating the integral at E_{cut}^0 gives

$$P'(E^1) = \frac{1}{2\sqrt{2\pi}\sigma_{01}'} e^{\frac{-(E^1)^2}{2(\sigma_{01}')^2}} \left[1 + \mathrm{erf}\left(\frac{E_{\mathrm{cut}}^0\sigma_{01}'}{\sqrt{2}\sigma_0\sigma_1'} - \frac{\sigma_0E^1}{\sigma_{01}'}\right)\right] \tag{16}$$

where $\sigma_{01}' = [(\sigma_0)^2 + (\sigma 1)2]1/2$. The general form of $\int e{-}x2\mathrm{erf}(ax + b)\, dx$ can not be analytically determined, and thus propagation of the truncated set, as well as determination of D' and D'', must be done numerically.

Propagation of a Sampled Low-Energy Distribution. The above treatment was based on obtaining *all* members of the ensemble below a given E^{cut}. While a number of methods are designed to give this set deterministically, other methods yield a set that is enriched in low-energy states but is not guaranteed to give all states within any energy cutoff. The same definitions of completeness and excess cost can be applied to these methods. However, the description of how the ensembles of states are passed through the hierarchy is different. Consider a sampling algorithm that, given some uniform distribution over a variable x, yields a distribution $Q(x)$. The sampled distribution of $P'(E^i)$ will then be given by $P''(E^i) = P'(E^i)\, Q(E^i)$. The distribution obtained by propagating this distribution up a level of the hierarchy is given by integrating (over all energies) the product of this distribution with the correlation function:

$$\begin{aligned} P'(E^{i+1}) &= \int_{-\infty}^{\infty} P''(E^i) g_i(E^{i+1}, f_1(E^i), \vec{\mu}_i(E^i))\, dE^i \\ &= \int_{-\infty}^{\infty} P'(E^i) Q(E^i) \times \\ &\quad g_i(E^{i+1}, f_1(E^i), \vec{\mu}_i(E^i))\, dE^i \end{aligned} \tag{17}$$

This may be extended to any number of levels of propagation in various ways. First, the sampled distribution may be passed on using another sampling-based algorithm (possibly the same one), in which case:

$$\begin{aligned} P'(E^N) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(E^0) \times \\ &\quad \prod_{i=0}^{N-1} Q(E^i) g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dE^0\, dE^1 \ldots dE^{N-1} \\ &= \int_V P(E^0) \prod_{i=0}^{N-1} Q(E^i) g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dV \end{aligned} \tag{18}$$

where the integral $\int_V dV$ is taken from $(-\infty, \infty)$ over all dimensions E^i with i in the range $[0, N-1]$. The infinite domain of integration indicates that the entire sampled distribution is used as the input for each subsequent step. However, an alternative is to pass on only the lowest-energy states from the sampled distribution. In this case, the result is analogous to eq 7:

$$P'(E^{i+1}) = \int_{-\infty}^{E_{\mathrm{cut}}^i} P'(E^i)\, Q(E^i) g_i(E^{i+1}, f_1(E^i), \vec{\mu}_i(E^i))\, dE^i \tag{19}$$

$$\begin{aligned} P'(E^N) &= \int_{-\infty}^{E_{\mathrm{cut}}^{N-1}} \int_{-\infty}^{E_{\mathrm{cut}}^{N-2}} \cdots \int_{-\infty}^{E_{\mathrm{cut}}^0} P(E^0) \prod_{i=0}^{N-1} Q(E^i) \times \\ &\quad g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dE^0\, dE^1 \ldots dE^{N-1} \\ &= \int_{V'} P(E^0) \prod_{i=0}^{N-1} Q(E^i)\, g_i(E^{i+1}, f_i(E^i), \vec{\mu}_i(E^i))\, dV' \end{aligned} \tag{20}$$

where V' represents the reduced space due to truncation of each distribution. The total sizes of the ensembles of interest are still given by $D'(E_{\mathrm{cut}}^N) = \int_{-\infty}^{E_{\mathrm{cut}}N} P'(E^N)\, dE^N$ and $D''(E^N) = \int_{-\infty}^{\infty} P'(E^N)\, dE^N$. It should be noted that the results given earlier for the truncation of an enumerated system are simply a special case of eqs 18 and 20 when $Q(E^i) = H(E^i - E_{\mathrm{cut}}^i)$, the Heaviside step function.

3. Methods

Conformational Analysis of Glucose. All calculations were done using the CHARMM computer program[22] with parameters from the Carbohydrate Solution Force Field

(CSFF).[23] D-β-glucopyranose was built in a standard geometry ($^4C_1$), and conformational states were generated by rotation about each of the hydroxyl dihedral angles (C_i-O_i, $i = 1-4, 6$) as well as about the exocyclic C_5-C_6 dihedral. An exhaustive enumeration of states was performed, sampling each dihedral at intervals of 60°, for a total of roughly 46 000 states. Energies were evaluated with no cutoffs, both in a vacuum (with a uniform dielectric constant of 1) and with the GBSW implementation of the Generalized-Born implicit solvent model.[24] GBSW calculations used an internal dielectric constant of 1 and an external dielectric constant of 80, with the dielectric boundary defined by a set of radii that have been optimized for this use.[25] The scaling coefficients were set to standard values of $a_0 = 1.2045$ and $a_1 = 0.1866$, the molecular surface was used, and a smoothing length of 0.2 Å was applied.

Protein Design. All calculations were done starting with the minimized average structure from the NMR structure of Calmodulin bound to the M13 peptide from rabbit skeletal muscle myosin light chain kinase (Protein Data Bank ID 2BBM).[26] Hydrogen atoms attached to carbons were removed for consistency with the PARAM19 parameter set used in the calculations. The positions of hydrogen atoms attached to heteroatoms were reoptimized using the HBUILD facility of the CHARMM computer program.

Sequences compatible with a low-energy complex structure were selected using a discrete structural search. The Dunbrack and Karplus rotamer library[27] was used, augmented by rotamers at $\pm 10°$ of χ_1 and χ_2 for each rotamer. The selected set of positions consisted only of basic and acidic residues: lysine and arginine were varied to Asp, Glu, Asn, and Gln, and aspartate and glutamate were varied to Lys, Arg, His, Asn, and Gln. The three protonation states of histidine were each considered as individual choices.

Energies for the initial search were calculated using the CHARMM computer program[22] with the PARAM19 polar-hydrogen force field.[22] A distance-dependent dielectric of $4r$ was used for electrostatic interactions. All energies were calculated relative to isolated model compounds of the variable side chains. Software written by Tidor and colleagues was used for the search.

Different levels of structural detail were considered at various stages in the design. A rotameric structure refers to a specific choice of an amino acid conformer at each position in the protein. In many cases, similar rotamers make similar interactions—to prevent this from complicating the search, the fleximer model of Mendes et al. was used.[28] Here, all "sub-rotamers" derived from enhanced sampling of a Dunbrack and Karplus rotamer were grouped into a single entity denoted as a (DK-)fleximer. The energies of interaction between fleximers were taken as weighted averages of the interaction energies between the component rotamers. The same approach was used to group all rotamers of a single amino acid into an entity we term a sequence-mer,[29] or all rotamers of a given amino acid with the same χ_1 and χ_2 into a $\chi_{1,2}$-fleximer.

The structural search involved a hierarchical process over these levels. The Dead-End Elimination (DEE) and A* algorithms[30–32] were first used over the space of fleximers (or sequence-mers) to find all fleximeric states with an energy within a given cutoff of the global minimum. DEE and A* were then used to find the lowest rotameric state for each low-energy fleximer; at this stage this problem is one of side-chain placement for a single sequence, not of sequence design. A maximum of 10 unique DK-fleximers were processed for each sequence, and the final result was a single, minimum-energy rotameric state for each low-energy sequence.

4. Results and Discussion

Theoretical Framework. Hierarchical approaches to molecular design are not new; rather, it has long been recognized that the vast space of chemical and conformational variations necessitates the use of approximate models for computational tractability, but that more rigorous calculations are required to achieve reasonable predictive capability with respect to experimental results. While hierarchical filtering procedures have been used in many applications, these have generally been constructed in an *ad hoc* manner—it is often not clear whether false negatives arise due to inadequate sampling at any given stage, or whether significantly more computational expense was applied than necessary. A mathematical framework for assessing the effectiveness of different hierarchical approaches would allow for a more rigorous consideration of these and other issues.

The Theory section described precisely such a framework, based on a statistical description of state distributions and correlations between hierarchical levels. In particular, a number of key descriptors were outlined. First, it was shown how a distribution of selected, low-energy states at one level of a hierarchy is transformed on moving up the chain, as well as how to generate an estimate of the expected density of low-energy states at the highest hierarchical level. Two scalar quantities of performance assessment were also defined: the completeness of the final solution and the excess work required to achieve the final result. Completeness provides a quantitative metric for the success of a hierarchical solution, describing the fraction of low-energy solutions found relative to those expected. Excess work provides a measure of efficiency, describing the relative amount of effort spent on discarded solutions relative to those kept in the final solution. For a given application, one of these may be more important than the other, or a balance of the two may be sought. These metrics provide a quantitative basis on which to address this balance.

Conformational Analysis of Glucose. As an example of how these methods may be applied, we consider a conformational analysis of D-β-glucopyranose (Figure 1a). This molecule overwhelmingly prefers the $^4C_1$ ring conformation, which places all substituents in an equatorial arrangement, and thus the primary degrees of conformational flexibility are the rotations of the exocyclic hydroxyls. This is a six-dimensional space of finite volume (360° for each angle), and thus an exhaustive evaluation of conformational states is feasible (for a moderate sampling of each dihedral). While the number of states may not be beyond enumeration, the computational cost of the evaluation of each state must also be considered, and an accurate model of the conformational free energy surface must take into account the solvent; for

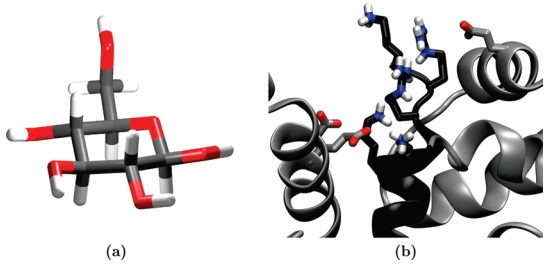

Figure 1. Model systems for application of statistical methods. (a) The minimum energy conformation of glucose found (in a Generalized-Born implicit solvent model) is displayed. (b) The residues of the CaM·M13 complex chosen for variation are displayed, using the minimized average solution structure. Calmodulin is shown in gray and M13 in black. Figures generated with VMD.[36]

most biological molecules, the environment of interest is that of an aqueous solution of moderate ionic strength. Among the most commonly used models for the inclusion of solvent effects are those based on continuum electrostatics: the Poisson−Boltzmann model and the Generalized-Born approximation.[33–35] As the computation of free energies in an implicit solvent model is significantly more costly than the corresponding calculation in a vacuum, one strategy for reducing the computation cost may be to screen the full space with a vacuum energy model and then refine the lowest energy states with a continuum model. However, the question arises as to how many low energy (vacuum) states need be considered in order to obtain an accurate description of the minimum-energy solvated states. This is precisely the question our technique aims to answer.

All six degrees of freedom were uniformly sampled at 60° intervals, giving a total of 46 656 distinct conformational states. The energy of each state was then evaluated with the CHARMM all-atom force field in a vacuum. These data follow a nearly perfect normal distribution ($R^2 = 0.9991$ for a nonlinear least-squares fit, see Supporting Information Figure S1) with a mean of 81.09 kcal/mol and a standard deviation of 6.42. All states with energies in the lowest 20 kcal/mol (8483 total, 18% of all states) were then selected for subsequent evaluation with a Generalized-Born (GB) solvent model;[24] the correlations of these two data sets are shown in Figure 2a. The two energies are correlated (although not strongly in a linear sense, $R^2 = 0.48$) and give a linear best fit with a slope of 0.62. However, for the application of the statistical analysis, we need to obtain an error model for the GB energies as a function of the vacuum energy.

The vacuum energies were divided into 1 kcal/mol bins, and the GB energies of all states in each bin were fit to a normal distribution (see Figure 2b and Supporting Information Figure S2). Several observations can be made that clearly demonstrate the applicability of the statistical model. First, in all cases, the fit to a normal error model was reasonable ($R^2 > 0.7$), and the fit was excellent ($R^2 > 0.96$) in every bin containing at least 200 states. Second, the mean GB energy in each bin is highly correlated with the vacuum energy, with an R^2 of over 0.99. Finally, for bins with a significant population, the standard deviation of GB energies in the bin is roughly constant, with a mean value of 1.97. These results motivate the use of a quite simple error model: (1) the expectation value of the GB energy varies linearly with the vacuum energy; (2) the distribution of GB energies around the expectation value follows a normal distribution of constant variance.

While in this case the full underlying distribution of vacuum energies is known, in many applications, it may not be. Thus, we additionally considered how well the full distribution may be fit using only those data within the lowest 20 kcal/mol. As clearly shown in Figure 2c and d, the fit is excellent, giving a mean of 81.00 (compared to the true mean of 81.09) and a standard deviation of 6.28 (true value, 6.42). Thus, strictly using the ensemble of low vacuum energies is reasonable.

Given the vacuum energy distribution and the error model for how GB and vacuum energies correlate, 10 000 model distributions were generated. As can be seen in Figure 2e, the distribution of model data matches the observed data (where known) very well. The model distributions were then used to estimate how well various low-GB-energy ensembles are captured; the lowest 2, 4, 6, 8, and 10 kcal/mol ensembles (relative to the lowest observed GB energy) were all considered. For each of these ensembles, the completeness and excess cost were computed as a function of the vacuum-energy cutoff used (see Figure 2f). With the actual cutoff applied to the data (20 kcal/mol), the 10 kcal/mol ensemble of GB states is found to have a completeness of roughly 90%, and the lower-energy ensembles are all near 100% complete.

In all these cases, however, the full 20 kcal/mol of low-vacuum-energy states had to be evaluated in the GB model, and many of these may not have been in the final low-energy ensemble; this is measured by the excess cost. For the 90% complete 10 kcal/mol ensemble, the excess cost is roughly 1, meaning half of the states that were evaluated were not part of the final low energy ensemble. On the other hand, the excess cost for the 2 kcal/mol ensemble is above 100; less than 1% of the evaluated states were part of the final set. This suggests that a cutoff of 20 kcal/mol in vacuum energy is inefficient if only the very lowest GB energies are desired; using a 15 kcal/mol cutoff would still give a near-perfect completeness, but with an excess cost 10-fold less.

To consider in more detail what this analysis provides, consider the middle of the ensembles considered (red curves); these are all states with GB energies within 6.0 kcal/mol of the global minimum and thus includes all states that would be populated more than 0.01% at room temperature. When only the lowest 6 kcal/mol of vacuum energies are considered, the completeness of this ensemble is only 7.6%; increasing this to the lowest 10 kcal/mol of vacuum energies increases the completeness to 39%, and thus many states are still missed. However, a 15 kcal/mol cutoff in vacuum energy gives a completeness of 91%, and the 20 kcal/mol cutoff gives a completeness of 99.89%.

To obtain this most complete ensemble required an excess cost of 16; for each state that was evaluated and "kept" in the low-energy ensemble, 16 were "discarded" as being outside the desired range. In other words, using the hierarchi-

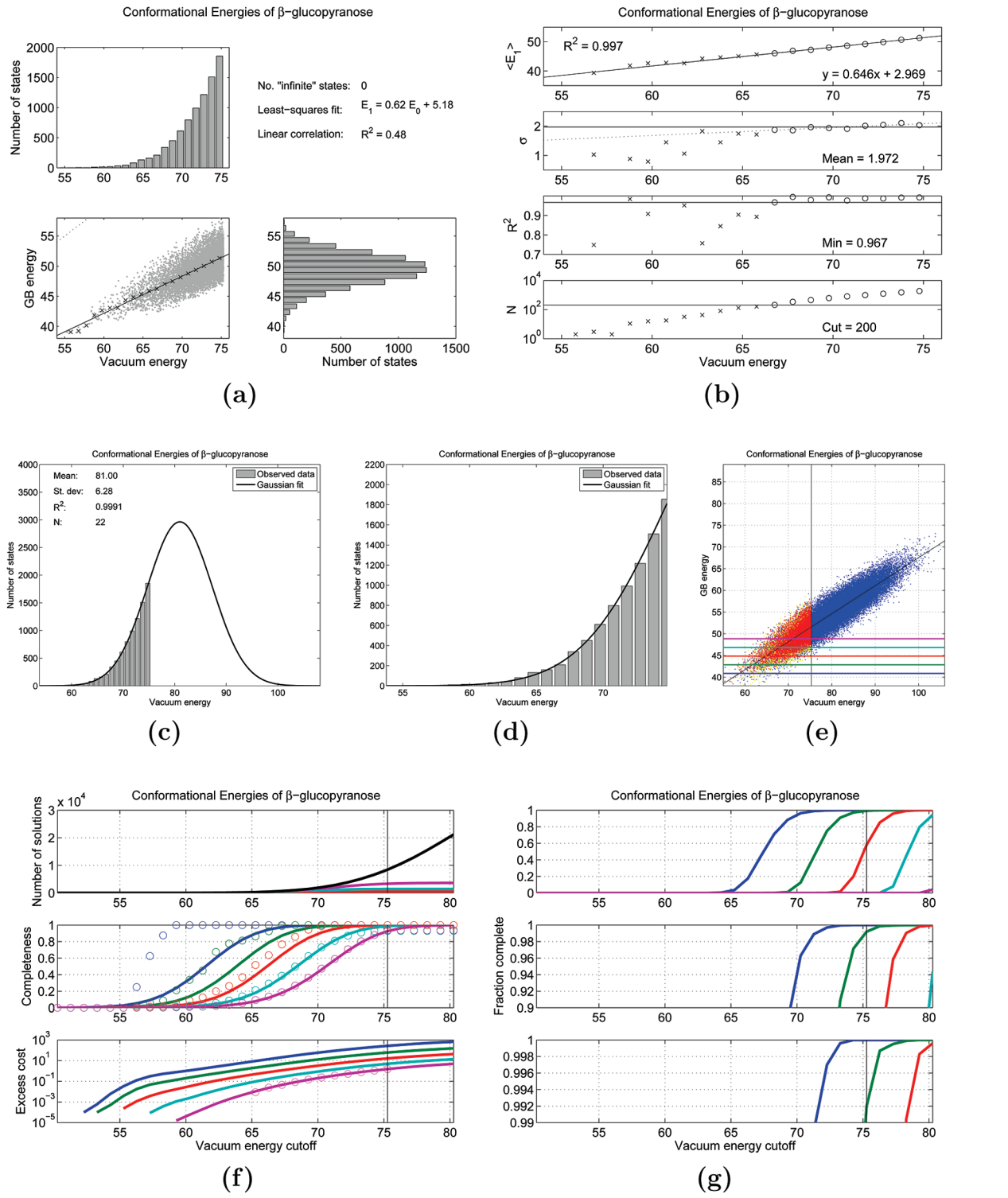

Figure 2. Conformational energetics of β-glucopyranose. Details of a conformational analysis of glucose are shown. (a) Correlations between Generalized-Born and vacuum energies for the lowest 20 kcal/mol of vacuum energy states. (b) Details of normal distributions fit to the GB energies of states having vacuum energies within 1.0 kcal/mol bins. $\langle E_1 \rangle$ is the mean GB energy, σ is the standard deviation, R^2 is the proportion of the variance described by the fit, and N is the number of data points in each bin. Bins of at least 200 points are indicated by open circles; these bins were used to evaluate the linear best fit equation of the mean, the mean standard deviation, and the minimum R^2 of the fits. (c, d) The distribution of low-vacuum energy states, fit to a normal distribution; d shows the same data with focused axes. (e) Model data, generated from the fit parameters of b and c. Blue points indicate model data and red, the actual data (yellow points are a model of GB energies from actual vacuum energies). The horizontal lines denote cutoffs in GB energy of 2, 4, 6, 8, and 10 kcal/mol, from the lowest actual value. (f, g) Metrics of performance for each GB-energy cutoff (colors match those of the cutoffs marked in e. In f, the total number of solutions, completeness, and excess cost from both model (lines) and actual (circles) data are shown. In g, the fraction of cases in which 100% completeness was achieved (out of 10 000 model calculations) is shown; each panel shows an increasingly focused *y*-axis range.

cal approach required evaluation of 16 times more states than were actually found. For the 91% complete ensemble, the excess cost drops to 3.6; 25% of the evaluated states are, in fact, part of the final solution.

Now, what does a 99.89% completeness really mean? In this case, we have 393 states in the low-GB-energy ensemble, and thus a 99.89% completeness suggests that we expect to have missed roughly "half a state". A complete enumeration of the space confirms that the ensemble is, in fact, 100% complete. The confidence in whether we found all states was assessed by considering the fraction of the 10 000 model distributions that led to a *perfectly* complete set of low GB energy states at a given vacuum energy cutoff (see Figure 2g); for a 20 kcal/mol vacuum cutoff, this value is 58.8% for the 6 kcal/mol GB ensemble. This can be interpreted as the confidence value that the ensemble is truly complete and thus provides one of the most important results of the statistical analysis. If a given level of confidence in the completeness of the solution is desired, the analysis also gives this: for a 96% confidence level, a vacuum energy cutoff of 22 kcal/mol should be used, and a solution that is complete to over 99.9% confidence can be found with a 25 kcal/mol cutoff.

A brief discussion of computational costs and the meaning of excess work is appropriate. As noted above, the excess work is not a direct measure of computational expense, but rather a measure of how many states selected at one level of the hierarchy (in this case, the vacuum energy) were not included in the final set of solutions (low-GB-energy states). The actual costs involve two primary contributions: (1) the cost of the first-level search and (2) the cost of re-evaluation in the second level. In an ideally performing system, the second step would only involve those states that are, in fact, low energy; the excess cost describes, in relative terms, how much more effort must be expended in the second step than this ideal bound.

In this case, the full conformational search over vacuum energies took 5.59 min, for a total of 0.0072 s per state, on a single 3.40 GHz Intel Xeon processor; due to software overhead, this is reduced to 0.0048 s per state when a large number of states are considered, for example by finer sampling. The GB calculations are slightly less than twice as costly, taking 0.013 s per state, and thus to perform the complete grid search requires 9.93 min. To evaluate the lowest 20 kcal/mol of vacuum energy states with GB, however, only requires 1.81 min, thus making the net time for the hierarchical search 7.40 min, or 75% of the exhaustive search time using GB. To achieve a 96% or 99.9% confidence in the completeness would require 8.32 or 10.05 min, respectively (84% or 101% of the exhaustive search cost).

Of course, these results suggest that there is minimal motivation for the use of a hierarchical method. Part of this arises from the relatively small cost differential of the two methods; in this small system, computing a GB energy is only fractionally more expensive than the corresponding vacuum energy. However, in more typical applications involving large biological macromolecules, GB is roughly 4-fold more costly. With this cost difference, the hierarchical approach would give 99.9% confidence in a complete low-energy GB ensemble at 70% of the cost of an exhaustive search; 96% confidence would be attained with 52% of the cost.

It should be noted that the model distributions give notably divergent results from the actual data for the very lowest GB-energy ensembles. There are only eight states in the lowest 2 kcal/mol, and only 80 in the lowest 4 kcal/mol. For samples of this small size, deviations must be expected. Additionally, the error model was fit primarily with data from a slightly higher energy range, where the density of states is larger; while this was done to reduce errors in the model from inadequate sampling, it could affect the accuracy of the error model in the lowest-energy regime. A comparison of how the two energies correlate across the full spectrum of energies (Supporting Information Figure S3) shows additional deviations from the model at high vacuum energies; as these states do not contribute to the low-GB-energy ensembles, these differences do not impact the analysis. The statistical analysis is most accurate for those data directly used in the derivation of the model.

Applications to Protein Design. A significant motivating force behind the development of this framework was for direct application to protein design. Thus, it is informative to consider a problem in this application space. We have recently described initial progress toward the development of variant Calmodulin−M13 peptide complexes with altered specificity.[21] In that work, we applied a hierarchical technique to the protein design problem at a number of focused sites. The same system is used here as an example with which we may evaluate the theory developed here.

Eight residues (five basic residues on M13 and three acidic groups on CaM) at a surface-exposed site at one end of the CaM−M13 binding site were varied to evaluate the viability of charge-reversal mutants at these positions (Figure 1b shows the design site). Each positive group was allowed to vary to the acidic amino acids and the amides, and each negative group was allowed to vary to the basic amino acids (including His) and the amides. As the three protonation states of histidine were considered individually, this corresponds to 1.6×10^6 possible sequences; with structural flexibility considered, there were 1.6×10^{26} individual structures under consideration. A significant number of rotameric states led to easily detected clashes with the fixed portion of the protein. After removal of these, the total structural search space was 8.8×10^{23}.

This space must be then be searched for low energy structures; the Dead-End Elimination (DEE) and A* algorithms may be used to enumerate the lowest-energy states in a guaranteed manner.[30,32] Rather than a single global minimum sequence and structure, we aim to find all sequences within a given cutoff of the global minimum, for several reasons. First, the DEE/A* approach requires a pairwise decomposable (in terms of individual side-chain positions) energy function. Thus, approximations to fundamentally non-pairwise energetic components (such as solvation free energies) must be made. Finding a number of low-energy states allows these to be reranked with more accurate energy functions, and to thereby obtain a better

Table 1. Number of States Found in Initial Search

E_0^{cut}	0	1	2	3	4	5	10	15	20
fine states	1	28	337	2352					
fine sequences	1	2	7	18					
coarse states	1	10	73	309	1095	3402	336 142	11 928 271	221 817 700
coarse sequences	1	4	21	51	99	188	2080	10 353	

estimate of the "true" lowest-energy sequences. Second, the design algorithm works on a single target energy function, while in reality several energies may need to be simultaneously optimized. For example, in optimizing the binding free energy between a pair of proteins, it is important to additionally maintain the stability of each individual structure. This can be addressed by optimizing on a single term that is a prerequisite for satisfying all others; in this case, optimizing the total complex energy (the sum of folding and binding free energies) satisfies the requirement. Given an ensemble of sequences with low complex energies, some will have higher affinity, and some higher stability; as we are particularly interested in high-affinity complexes, the low-energy set can be subsequently screened for this criterion. Finally, current models for protein energetics are still not ideal, and thus for experimental testing, a set of possible variants is desired.

A Fine Rotamer Library Makes Enumeration Infeasible. When DEE/A* is used to enumerate low-energy states of the full space, a problem becomes readily apparent: due to the large density of states, many with similar energies, it is infeasible to enumerate states beyond 3 kcal/mol above the global minimum (all computations beyond this level required beyond 8 GB of internal memory and many days of computational expense). While over 2000 structural states are found in this range, these states correspond to only 18 distinct sequences (see Table 1). This is a result of two features: the relative solvent exposure of the site and the size of the rotamer library used in the search. An augmented version of the Dunbrack−Karplus library was used in this search, with χ_1 and χ_2 sampled at $\pm 10°$ around each standard rotamer. Since the design site is fairly exposed, it is reasonable to consider using the unaugmented library (the finer sampling is often needed in buried sites to allow reasonable packings to be found). Using this coarser library allows a much more extensive sampling of sequences (see Table 1). For example, 2080 sequences can be found within 10 kcal/mol of the global minimum, from a total of over 300 000 structural states.

However, when the results of the two calculations are compared, there is little correlation. Of the 18 sequences within 3 kcal/mol of the global minimum with the fine library, only four are within the same cutoff with the coarse library. While all 18 are found within 10 kcal/mol of the coarse global minimum, the sequence ranking third with the fine library ranks 595 with the coarse library. In terms of energies, the top fine library sequences are roughly 20 kcal/mol more favorable than those from the coarse library, and essentially no correlation between the two values is seen.

A Fleximer Model Makes the Search Tractable. These issues have been recognized previously, and the "fleximer" model of Mendes et al. is an elegant solution.[28] Briefly, in this approach, pairwise energies are computed for all rotameric states in the fine library, but the discretization of the coarse library is used in the search. When computing energies for the search, the interactions of any parent rotamer at a given position is given by the Boltzmann-weighted average of the interactions of all substates of that rotamer. The results of such a search do not correspond to any single structural state, and thus a second step is required, in which the minimum energy state for a given set of fleximers is found. As this involves a search over only roughly nine choices at each position, this evaluation is very fast. This approach is very successful; using the fleximer model, the true, fine-library global minimum is ranked third, with a difference in energy of only 0.2 kcal/mol from the fleximer global minimum (see Table 2). Of the 18 top fine-library sequences, 11 are found within the same 3 kcal/mol of the fleximer global minimum, and 17 are found within the top 5 kcal/mol.

Table 2. Rank of Full Search Global Minimum in Initial Search

	rotamer	DK-fleximer	$\chi_{1,2}$-fleximer	sequence-mer
rank	1	3	272	148
ΔE	0.0	0.20	7.88	9.71

The search with the fleximer model is much more efficient than that with the fine library, and thus it is feasible to enumerate as high as 20 kcal/mol from the global minimum (in which range there are over 40 million fleximer states, and 11 000 distinct sequences). When collapsed into a single rotameric state, the energy obtained for a given fleximer is identical to that found in the fine-library search. However, since the search is performed on the fleximer energy, it is possible that true low-energy sequences are not found in the fleximer-based search. This limitation depends directly on how many fleximer states are enumerated, and how many true low-energy states are desired. For example, we have seen that 5 kcal/mol of fleximer states are required to find 17 of the 18 sequences within 3 kcal/mol of the fine-rotamer minimum. How far must the fleximer energetic landscape be explored to find all sequences within a given cutoff in fine-library energy? Again, this question can be directly addressed with the statistical framework presently here.

Defining Correlations between Rotamer and Fleximer Energies. In order to address this question, it is necessary to know both the degree of correlation between the fleximer and rotamer energies and the distribution of energetic states in the fleximer model. As can be seen in Table 3, the number of states increases exponentially with increasing distance from the global minimum. This is consistent with an ensemble of states that is normally distributed: the extremes of a normal distribution are essentially exponential, and only 0.7% of the total sequence space is sampled in the lowest

Table 3. Number of States Found in Initial Search

	fine rotamer		DK fleximer		$\chi_{1,2}$-fleximer		sequence-mer
E_0^{cut}	states	seq.	states	seq.	states	seq.	sequences
0	1	1	1	1	1	1	1
1	28	2	7	3	2	1	1
2	337	7	47	15	3	1	3
3	2352	18	156	32	7	4	11
4			481	45	19	9	17
5			1316	69	35	11	25
10			77466	632	975	145	161
15			2205651	3136	13591	788	743
20			41998715	11030	129939	3198	2576
25					971173	9585	7441
30					5773982	24501	18310
40					116352240		78564
50							229332
75							1019658
100							1538405
150							1600000
total	1.49×10^{26}		3.58×10^{18}		1.42×10^{13}		1.60×10^{6}

20 kcal/mol. As was done in the conformational analysis of glucose, the fleximer energies were grouped into bins, and the distributions of rotamer energies within each bin were characterized (Figure 3a and b). While the distribution of rotamer energies in each bin deviates somewhat from normality, the bulk of each distribution is, in fact, well fit by a normal curve—the R^2 of the fit was greater than 90% in all cases, except in the lowest-energy bins, where there were few data. Samples of the fits to individual bins may be found in the Supporting Information (Figure S4). The deviations are most notable at low energies, where the observed number of states is less than would be expected from a normal curve, and at high energies, where a greater number of states is observed. This is expected, as there is a lower bound on the rotamer energy for a given fleximer energy, but no upper bound. While it may be possible to define an alternate distribution that is a slightly better fit for the tails, we have chosen to use the normal distribution for further analysis.

The details of a linear fit to the mean are given in Table 4 and Figure 3b. The means are essentially perfectly correlated ($R^2 = 0.999$), and the best-fit line is very near the $y = x$ line, with a slope of 0.99 and an intercept of −0.17. The standard deviation in each bin is on average 1.49 kcal/mol. This fit was done with all sequences within 20 kcal/mol of the global minimum, including data from all bins (1 kcal/mol in width) with a population of at least 100; this provides 11 points for the fitting. However, it is clear that the fit also matches well those data from low-energy bins with lower sampling. Eleven-thousand sequences were considered in order to reach this result, with 42 million individual structures enumerated. Thus, a natural question is whether fewer data would suffice. Table 4 additionally shows the results of fitting only data within 10 or 15 kcal/mol of the global fleximer minimum; in these cases, all bins with at least 50 points were included, so as provide a reasonable number of bins for fitting. The results are very similar—the slope of the best-fit line ranges from 0.97 to 0.99, and the average standard deviation is between 1.5 and 1.6 kcal/mol. Thus, it is clear that very similar results are obtained even when only the very lowest-energy fleximer states are considered; the 10 kcal/mol ensemble contains only 0.04% of the total sequence space, with less than 80 000 structures, while the 20 kcal/mol ensemble contains 0.7% of the sequence space (and 42 million structures).

Model Distributions Capture Observed Behaviors in the Hierarchy. In addition to the correlation between energetic levels, the theoretical framework outlined above requires a knowledge of the distribution of states in the lowest (fleximer) level. The DEE/A* methodology gives a rigorous enumeration of the lowest-energy states, and we additionally have prior knowledge of the total number of sequences in the space. Thus, it is possible to perform a nonlinear least-squares fit of a normal distribution to the available data. The fit (see Supporting Information Figure S5) is excellent in the region where there are data, with an R^2 of 98.8%, although these data only occupy the region from −4.2 to −2.5 standard deviations from the mean. Thus, this should be considered an estimate, and it should be expected that there will be significant deviations between this estimate and the true distribution.

This estimated fleximer distribution was then combined with the observed correlation and error model for transferring between fleximer and rotamer energies to provide an estimate of the complete distribution of minimum rotamer energies. This was then used to compute expected completeness and excess-cost curves, as a function of fleximer-energy cutoff, for different low-rotamer-energy ensembles (see Figure 3c and d). Given the 20 kcal/mol cutoff that was used, near 100% completeness is expected for ensembles up to 15 kcal/mol of the global minimum in rotamer energy, roughly 80% completeness is expected for rotamer energies within the same 20 kcal/mol cutoff, and about 35% of sequences in the lowest 25 kcal/mol of rotamer energies are expected to have been found. Comparing to the observed data, the agreement with the completeness estimates is remarkable. For the lowest energy ensembles (5, 10, and 15 kcal/mol), the 100% completeness is strongly supported by the observation that the ensemble is converged with increasing fleximer energy. For the higher-energy rotamer ensembles (20 and 25 kcal/mol), the number of observed states matches very closely the number of states predicted by the model; the expected total of number of states can be used with the

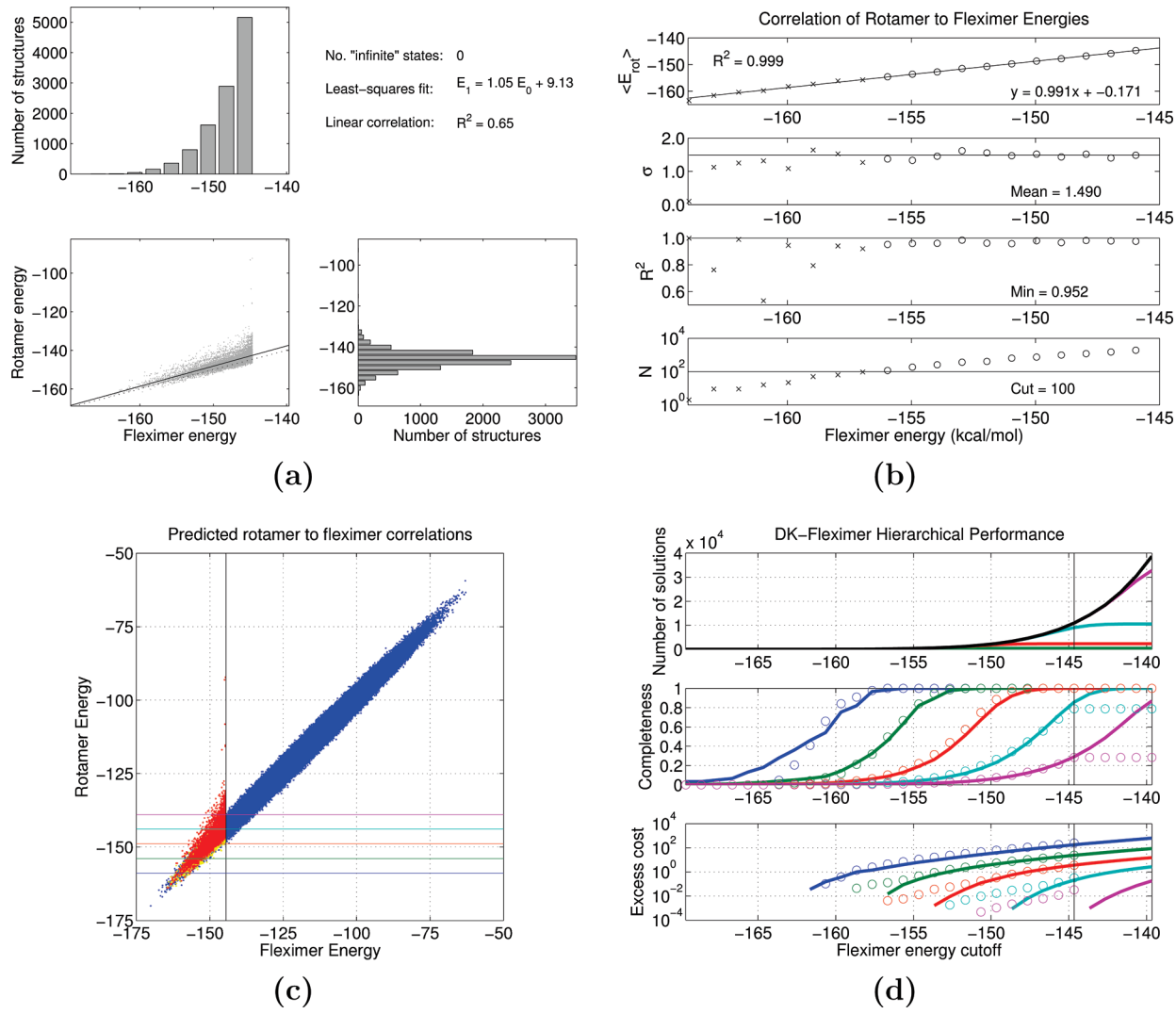

Figure 3. Correlations of rotamer to fleximer energies. (a) The distribution of fleximer energies within 20 kcal/mol of the global minimum are shown, along with the distribution of minimum rotamer energies for the same ensemble and the correlation between the two. A linear least-squares fit gives a slope near unity, with a modest correlation ($R^2 = 0.65$). (b) The results of fitting a Gaussian to the distribution of rotamer energies within 1.0 kcal/mol bins of fleximer energies are shown. The mean ($\langle E_{rot} \rangle$) shows strong linear correlation. The standard deviation (σ) is uniform, with a value of 1.5 kcal/mol in the nearly all bins, and the fit to a normal distribution is excellent ($R^2 > 0.9$) in all cases with 100 data points (N) or higher, shown as open circles. (c) The correlation of rotamer to fleximer energies simulated using the distribution of fleximer states and correlations of rotamer to fleximer energies computed from low energy states (blue points) are shown, along with the observed data (red points). Yellow points indicate the simulated distribution of rotamer energies given the actual (low-energy) fleximer energies. (d) The computed performance metrics are shown for the simulated data (solid lines), and for the observed data (open circles). Colors correspond to different rotamer energy cutoffs (from 5 to 25 kcal/mol, in increments of 5 kcal/mol), indicated by horizontal lines of the same color in the right panel. The black line in the number of solutions indicates the total number of solutions at the fleximer level.

Table 4. Statistics of Rotamer to Fleximer Fit

	DK fleximer			χ_1, χ_2 fleximer			sequence-mer		
E_0^{cut}	slope	intercept	$\langle\sigma\rangle$	slope	intercept	$\langle\sigma\rangle$	slope	intercept	$\langle\sigma\rangle$
30[a]				0.83	−20.28	3.73	0.66	−41.70	4.05
20[a]	0.99	−0.17	1.49						
30[b]				0.79	−26.14	3.74	0.66	−42.80	3.99
25[b]				0.79	−25.34	3.88	0.65	−44.09	3.86
20[b]	0.98	−1.93	1.48	0.84	−17.38	4.01	0.65	−43.04	3.72
15[b]	0.99	−0.58	1.51	1.07	19.25	4.42	0.62	−48.63	3.59
10[b]	0.97	−2.70	1.57	0.62	−49.67	4.68			

[a] Fit included all bins with at least 100 values. [b] Fit included all bins with at least 50 values. In all cases, bins were of 1 kcal/mol width.

number of observed states to compute a completeness measure that agrees with prediction.

Excess-cost estimates suggest that 80% complete ensembles can be obtained at very little excess cost (10%),

and 100% complete ensembles can be realized with an excess cost of roughly 1 (equal number of kept and discarded states). Additionally, it can be seen that a 100% complete ensemble of the top 10 kcal/mol (rotamer energy) should be attainable with enumerating only up to 15 kcal/mol in rotamer energy, and that the top 5 kcal/mol of rotamer energy sequences can be fully determined with a fleximer cutoff of 10 kcal/mol. Considering the observed data, there is good agreement in the regime of near-complete sampling (completeness greater than about 75%), but the observed excess work is significantly larger than the model would predict for less-complete sampling. This is not surprising, as the true distribution of rotamer energies for a given fleximer energy has a longer positive tail than the model normal distribution. Thus, there are a larger number of discarded states than expected by the model, which leads to higher excess cost. This is most dramatic when there are few low-energy states (low completeness), and less apparent when there are many states.

Statistical Guarantees for a Hierarchical Approach. We thus have a very important result—while with a direct search using a fine library only the top 3 kcal/mol could be enumerated (giving 18 sequences), using the fleximer model allows complete enumeration of 15 kcal/mol of low-energy states (2388 sequences). While the completeness of this ensemble is not algorithmically guaranteed, as is the case when a space is directly searched with DEE/A*, a *statistical guarantee* has been provided; that is, we can rigorously define a confidence value that all solutions have been found. Averaging over 10 000 model distributions, the completeness of the 15 kcal/mol ensemble was found to be 99.992%. Given the size of this ensemble, this leads to perfect completeness in 83% of the cases, and in an additional 15%, a single sequence was not found. The ensemble may thus be described as perfectly complete with greater than 80% confidence, and there is greater than 98% confidence that no more than a single sequence has been omitted.

Improving Efficiency with Alternative Hierarchies. While the above approach allowed for complete sampling of the top 15 kcal of energies, a great deal of computational expense was involved. In particular, 42 million fleximer states had to be enumerated, and about 100 000 fleximer states expanded to a unique rotameric state (up to 10 per sequence). This involved roughly one week of computation on a single AMD Opteron 250 (2.4 GHz) processor. Could this expense be reduced by creating coarser-grained models for the initial search? To test this, the fleximer model was applied to alternate groupings of rotameric states. In the first, all rotamers with the same χ_1 and χ_2 angles (from the Dunbrack−Karplus library) were grouped in a single fleximer; the finer samplings of ±10° were included in the fleximer defined by the parent angles. In the second, all rotamers of a given amino acid were grouped into a single "sequence-mer". Thus, whereas the Dunbrack−Karplus-based fleximer sampled 282 fleximer states at each acidic position, and 174 at each basic position, the $\chi_{1,2}$-fleximer model samples 54 and 39 per acidic or basic position, respectively, and the sequence-mer model samples eight and five states per position. This leads to a dramatic reduction in the overall size of the search space—from 3.6×10^{18} for the Dunbrack−Karplus fleximer to 1.4×10^{13} for the $\chi_{1,2}$-fleximer and 1.6×10^{6} for the sequence-mer.

For each fleximer model, all low-energy states were enumerated with DEE/A*, and up to 10 states per sequence were expanded into single rotamer structures, as discussed above (see Table 3). For the $\chi_{1,2}$-fleximer model, there were 130 000 states corresponding to 3200 sequences in the top 20 kcal/mol of fleximer energies; for the sequence-mer model, there were 2600 sequences in the same range. This is, of course, much smaller than the number of sequences found in the top 20 kcal/mol of the initial (Dunbrack−Karplus, DK) fleximer model, and thus larger cutoffs were considered. In the top 30 kcal/mol, 24 500 sequences (almost 6 million states) were found for the $\chi_{1,2}$ model, and 18 300 sequences were found with the sequence-mer model.

As above, these data were binned according to fleximer energy, and the relationships between rotamer and fleximer energy were determined (see Table 4 and Figure 4). As for the DK-fleximer model, the mean rotamer energy is linearly correlated with both coarser models, with R^2 values of greater than 99%. However, the slope of the correlation is below unity in both cases—0.83 for the $\chi_{1,2}$-fleximer and 0.66 for the sequence-mer. The distributions of the rotamer energies around the mean are also fit well by a Gaussian, and the standard deviation is constant across the observed range. Not surprisingly, however, the distributions are much broader than was the case for the Dunbrack−Karplus fleximer (3.7 kcal/mol for $\chi_{1,2}$, 4.0 kcal/mol for sequence-mer). While these seem quite similar, it is important to note that, if the data in each set were scaled to give a slope of unity in the correlation of the means, the standard deviation about the mean would scale by the reciprocal of the original slope. Thus, the normalized standard deviation for the $\chi_{1,2}$ distribution is 4.5 kcal/mol, and that for the sequence-mer distribution is 6.1 kcal/mol. This compares with 1.5 kcal/mol for the Dunbrack−Karplus fleximer.

These linear correlations were then combined with best-Gaussian fits to the original fleximer energy distribution (Table 5) to give the expected completeness and excess-cost for varying energetic cutoffs (see Figure 5). The completeness curves for the $\chi_{1,2}$-fleximer model transition more sharply than those for the sequence-mer model; this is expected, as the slope of the transition depends on the accuracy of the correlation between models. However, at the highest cutoff in fleximer-energy considered (30 kcal/mol), both models give similar overall completeness. Curiously, the completeness for the 20 kcal ensemble in rotamer energy shows roughly the same degree of completeness (80%) with both these models (and a 30 kcal/mol fleximer cutoff) as with the DK-fleximer model using a 20 kcal/mol cutoff. This is coincidental, but allows for an interesting observation to be made concerning the ensembles of slightly higher and lower rotamer energy. The 15 kcal/mol (rotamer energy) ensemble was 99.992% complete when searching with the DK-fleximer model, and the completeness of the 25 kcal/mol ensemble was 35%. The completeness of the 15 kcal/mol ensemble is somewhat reduced when the search is performed with the coarser fleximer models (97.0% for the $\chi_{1,2}$-fleximer and 97.7% for the sequence-mer). This is a

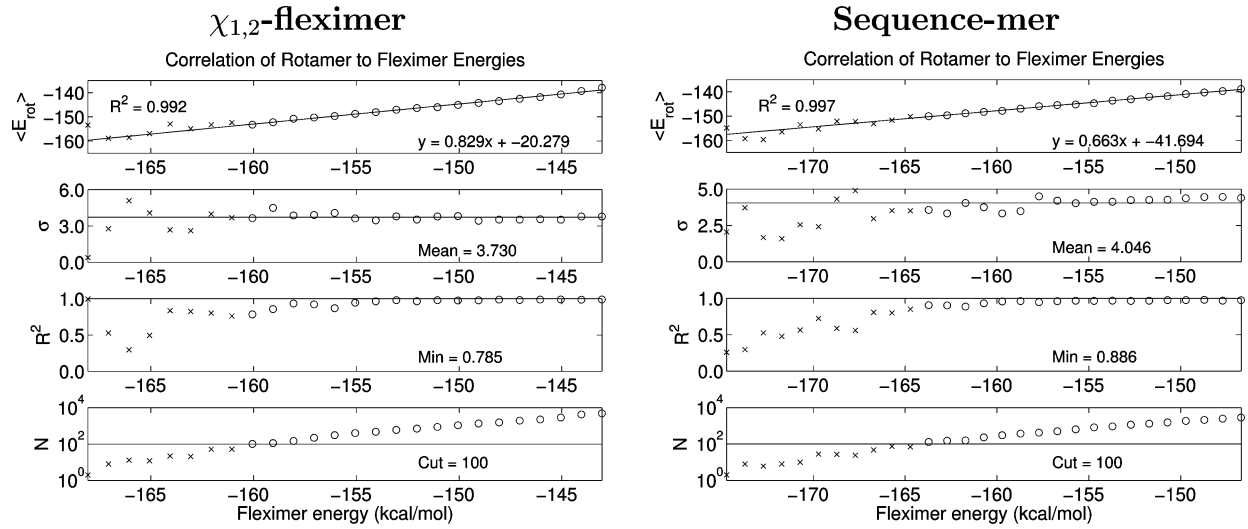

Figure 4. Correlations of rotamer to $\chi_{1,2}$-fleximer and sequence-mer energies. The results of fitting a Gaussian to the distribution of rotamer energies within 1.0 kcal/mol bins of $\chi_{1,2}$-fleximer (left) and sequence-mer (right) energies are shown. In both cases, the mean ($\langle E_{rot} \rangle$) shows strong linear correlation, the standard deviation (σ) is uniform, and the fit to a normal distribution (R^2) is very good in all cases with 100 data points (N) or higher, shown as open circles.

Table 5. Statistics of Gaussian Fit to Low-Energy Sequences

	mean	σ	R^2	bins
DK-fleximer	−119.58	10.16	0.9876	21
$\chi_{1,2}$-fleximer	−122.59	9.23	0.9529	27
sequence-mer	−110.78	15.61	0.9992	31

result of the broader transition to completeness and thus is not unexpected. However, the broadness of the transition also contributes to a *higher* completeness of the less-fully sampled ensembles—the 25 kcal/mol (rotamer energy) ensemble is estimated to be 44% complete when the $\chi_{1,2}$-fleximer model is used, and more than 52% complete when the sequence-mer approximation is used.

The more coarsely sampled models naturally have a larger excess-cost; achieving near 100% completeness (for example, in the 15 kcal/mol ensemble) requires evaluation of 10 times as many states as are found in the final ensemble. While this may initially seem like a large drawback compared with the DK-fleximer model (which only required evaluation of double the number of kept states), it is important to additionally consider the computational cost of each step. For all cases, the full pairwise energy matrix for the fine library must be computed; this took roughly 50 min on a single AMD Opteron 250 (2.4 GHz) processor. To enumerate the top 20 kcal/mol of states in the DK-fleximer model (41 million states and 11 030 sequences) took 6.5 days on a single CPU, while to enumerate the top 30 kcal/mol of the $\chi_{1,2}$-fleximer model (5.8 million states, 24 501 sequences) took just under 1 h, and to enumerate the 18 310 sequences in the top 30 kcal/mol of the sequence-mer model took less than 1 min. These dramatic differences in the initial search are somewhat offset by the need to do additional calculations to find the true low-energy structures corresponding to each fleximeric state; this process is more costly for fleximers containing more members. For the DK-fleximer model, the cost of this stage was negligible, roughly 1 h, while it took roughly 1.5 days for the $\chi_{1,2}$-fleximer model and 2 days for the sequence-mer model. However, when all steps are considered, both coarser fleximer models (requiring about 1.5−2 days total time) are significantly more cost-efficient than the DK-fleximer model (requiring roughly 1 week total CPU time).

While slightly more cost-effective overall than the sequence-mer model, the $\chi_{1,2}$-fleximer model gives a somewhat poorer best-Gaussian fit to the original distribution ($R^2 = 95.2\%$) than either the DK-fleximer ($R^2 = 98.8\%$) or the sequence-mer model ($R^2 = 99.9\%$) and noticeably underestimates the true density of states at low fleximer energies (Supporting Information Figure S5). The reason for this is not entirely clear, and work on additional systems will be needed to determine whether this is a broader issue.

The Underlying Sequence-mer Distribution Is Normally Distributed. The above analysis was based on an assumption that the underlying (lowest level) energy distribution can be reasonably estimated by a normal distribution fit to the lowest energy states. In most cases, the vast number of states precludes directly assessing this. However, as our model system consists of only 1.6 million distinct sequences, we can, in fact, enumerate the sequence-mer distribution. Figure 6 shows this full distribution, along with a Gaussian fit to all the data, and a Gaussian fit to only the lowest 30 kcal/mol of sequences (2% of the total).

Considering these data, we may make two observations. First, it is notable that a normal distribution models the full distribution of energies nearly perfectly, and thus the initial assumption is validated in this case. Second, while the low-energy fit is not perfect, the estimated distribution matches the actual data remarkably well; the fit mean is shifted to slightly lower energy, and the fit variance is slight smaller. Thus, the number of moderate-energy states will be overestimated and the number of high energy sequences underestimated. As the most significant deviations are at high energies, which contribute little to the low-energy states at higher hierarchical levels, the errors from this approach will

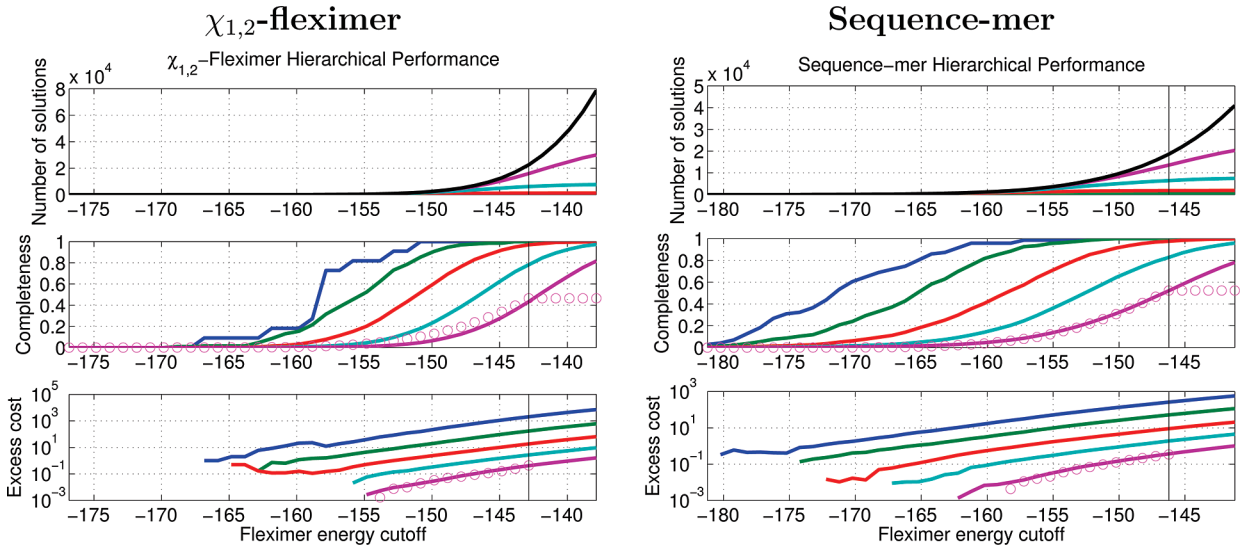

Figure 5. Performance metrics of rotamer to $\chi_{1,2}$-fleximer and sequence-mer hierarchies. The computed performance metrics for $\chi_{1,2}$-fleximer (left) and sequence-mer (right) are displayed. Solid lines of different colors correspond to simulated data for different rotamer energy cutoffs (from 5 to 25 kcal/mol, in increments of 5 kcal/mol, as in Figure 3). Open circles correspond to the actual data for highest rotamer-energy cutoff (25 kcal/mol). The black line in the number of solutions indicates the total number of solutions at the fleximer level.

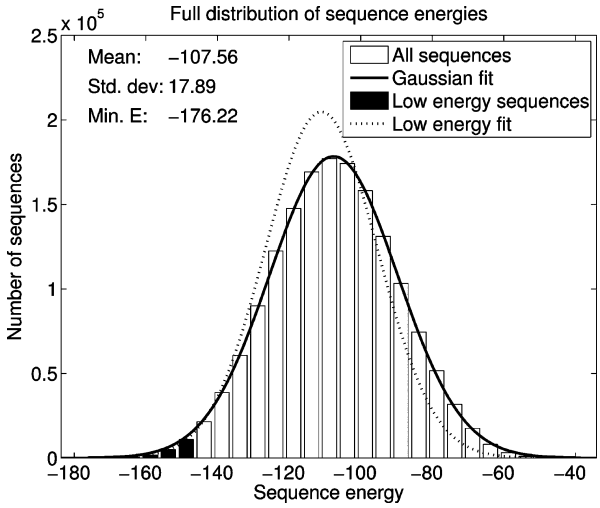

Figure 6. Complete histogram of sequence-level energies. The distribution of the enumerated set of sequence energies for the design site are shown. The dark curve is a best-fit normal distribution to all the given data. The dotted curve is the best-fit Gaussian curve, using only the lowest 30 kcal/mol of sequences (denoted by black bars).

be further reduced. As a result, reasonable results may be obtained using a predicted sequence-mer distribution.

Applications of the Statistical Framework. The previous sections detailed applications for the theory developed here in conformational search and in protein design. However, the methods may be applied with equal ease to any of a large number of applications in molecular simulation and design. The statistical framework provides a number of key benefits. One of the most significant of these is the ability to assess completeness, which can be used to provide a level of confidence that the global minimum has been found. It may also be used to answer a challenging problem: if no satisfactory solution is found, is it the result of incomplete sampling or due to the true absence of a satisfactory solution? These questions are important ones in protein design, in ligand docking, and in other areas of molecular design.

An additional application is for problems where an ensemble of states is necessary. Important problems in protein evolution can be addressed by determining the set of all sequences compatible with a given protein structure; protein-design methods can be applied to this problem, but completeness measures are essential. Conformational search methods are also often used to generate ensembles of states from which ensemble-averaged properties, such as free energy and entropy, may be computed. The challenge in these approaches, though, is that these properties are only accurate with adequate sampling of the low-energy regions of phase space; completeness provides a direct means of evaluating such sampling.

It should be noted that there is an important distinction that can separate the commonly-used search algorithms. Some methods are designed to enumerate all low-energy structures (these include exhaustive search approaches and tree-search-based methods such as Dead-End Elimination and A*), while other methods produce a *sampling* of the low-energy states (genetic algorithms, Monte Carlo, and simulated annealing are among these). While in an ideal situation, the sampling produced by the latter methods will be complete (or nearly so), this cannot be guaranteed with finite resources; the first class of methods can provide guarantees that *all* low-energy states (of the set considered) have been found. The statistical theory described here is compatible with both types of algorithm so long as a functional description of the expected sampling produced by the search algorithm can be given—DEE/A* gives a sampling distribution described by a step-function, while Monte Carlo gives a sampling distribution equivalent to the Boltzmann distribution at a given temperature.

Possible Limitations. The utility of the approach outlined here is fundamentally limited by the accuracy to which the

underlying (lowest level) probability distribution and the correlations between hierarchical levels can be estimated. In general, these will not be known *a priori* and thus, as discussed above, must be derived *a posteriori* from sampled data. If there is a systematic error which leads to a dramatic difference in energy for only a subset of states, an error model derived solely from states that are low energy in the reference model may not be representative of the full set of states. For example, consider two models in which a particular class of states (e.g., molecules with a net charge of $-1e$) are destabilized in the lower-level model but stabilized in the higher-level model, relative to all other states. The low-energy states from the first model may not include any molecules of this class, and thus this bias would be excluded from the error model. As a result, the completeness could be estimated to be very high, even though a significant number of lower-energy states (at the higher-level) were missed.

A related limitation is the need to be able to sample enough of the space in the lower-level model to derive a reasonable error model. In a system where the density of states very close to the global minimum is large, methods which aim to enumerate low-energy states may not be able to sample a wide enough range of energies for a reasonable model to be obtained. This should be less of an issue for sampling methods which include some higher energy states (such as Monte Carlo), although cases that remain problematic in this regime could still be constructed.

These caveats are important to be aware of, and care should be taken to evaluate how accurate the error models are expected to be. It should also be noted that it is possible to decouple the derivation of the error model from the search. For example, a random (or less-biased) search over states in the lower-level model would give a broad sampling of states that may be used to derive an error model across a wide energy range; the model could then be used to evaluate the performance of more targeted search strategies. Although there will always be *some* possibility of error, the use of a statistical model explicitly considers this; terms like the completeness are estimated to a certain level of confidence, rather than given as precise predictions.

5. Conclusion

We have outlined a statistical framework for performance analysis in hierarchical methods, with a particular focus on applications in molecular design. The theory is derived from fundamental statistical principles, presuming that the relationship between the results of each hierarchical level may be described by some functional correlation (linear or not), and an error model for how values are distributed around the correlation curve. Two example problems—one in conformational search and one in protein design—clearly show the usefulness of this approach; measures of completeness of the final ensemble can be computed, providing a level of confidence that adequate sampling of low-energy states has been achieved.

The framework we have described here is applicable not only to specific examples presented here but to any problem in molecular design that involves a hierarchical approach. Perhaps the most common of these is that of protein−ligand docking and virtual high-throughput screening, and we look forward to seeing this framework applied to these problems.

Acknowledgment. D.F.G. would like to thank Bruce Tidor for making the protein-design software available, as well as for helpful discussions on an early version of this manuscript. D.F.G. also thanks the State University of New York, the College of Engineering and Applied Sciences and the Department of Applied Mathematics and Statistics for financial support.

Supporting Information Available: Figures demonstrating the various fits in more detail are presented. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Szymkowski, D. E. Creating the next generation of protein therapeutics through rational drug design. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 590.

(2) Rosenberg, M.; Goldblum, A. Computational protein design: A novel path to future protein drugs. *Curr. Pharm. Des.* **2006**, *12*, 3973.

(3) Razeghifard, R.; Wallance, B. B.; Pace, R. J.; Wydrzynski, T. Creating functional artificial proteins. *Curr. Protein Pept. Sci.* **2007**, *8*, 3.

(4) Drexler, K. E. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 5275.

(5) Pabo, C. O. Molecular technology: Designing proteins and peptides. *Nature* **1981**, *301*, 200.

(6) Park, S.; Yang, X.; Saven, J. G. Advances in computational protein design. *Curr. Opin. Struct. Biol.* **2004**, *14*, 487.

(7) Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10*, 787.

(8) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: Current status and future challenges. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 15.

(9) Halperin, I.; Ma, B. Y.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 409.

(10) Bonvin, A. M. Flexible protein-protein docking. *Curr. Opin. Struct. Biol.* **2006**, *16*, 194.

(11) McCarrick, M. A.; Kollman, P. A. Predicting relative binding affinities of non-peptide HIV protease inhibitors with free energy perturbations calculations. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 109.

(12) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem. B* **2003**, *107*, 9535.

(13) Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. Direct calculation of the binding free energies of FKBP ligands. *J. Chem. Phys.* **2005**, *123*, 084108 .

(14) Mardis, K. L.; Luo, R.; Gilson, M. G. Interpreting trends in the binding of cyclic ureas to HIV-1 protease. *J. Mol. Biol.* **2001**, *309*, 507.

(15) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032.

(16) Gohlke, H.; Case, D. A. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem.* **2004**, *25*, 238.

(17) Jaramillo, A.; Wodak, S. J. Computational protein design is a challenge for implicit solvation models. *Biophys. J.* **2005**, *88*, 156.

(18) Given, J. A.; Gilson, M. K. A hierarchical method for generating low-energy conformers of a protein-ligand complex. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 475.

(19) Grüneberg, S.; Stubbs, M. T.; Klebe, G. Successful virtual screening for novel inhibitors of human carbonic anhydrase: Strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45*, 3588.

(20) Floriano, W. B.; Vaidehi, N.; Zamanakos, G.; Goddard, W. A., III. HierVLS Hierarchical docking protocol for virtual ligand screening of large-molecule databases. *J. Med. Chem.* **2004**, *47*, 56.

(21) Green, D. F.; Dennis, A. T.; Fam, P. S.; Tidor, B.; Jasanoff, A. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. *Biochemistry* **2006**, *45*, 12547.

(22) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187.

(23) Kuttel, M.; Brady, J. W.; Naidoo, K. J. Carbohydrate solution simulations: Producing a force field with experimentally consistent primary alcohol rotational frequencies and populations. *J. Comput. Chem.* **2002**, *23*, 1236.

(24) Im, W.; Lee, M. S.; Brooks, C. L., III. Generalized born model with a simple smoothing function. *J. Comput. Chem.* **2003**, *24*, 1691.

(25) Green, D. F. Optimized parameters for continuum solvation calculations with carbohydrates. *J. Phys. Chem. B* **2008**, *112*, 5238.

(26) Ikura, M.; Clore, G. M.; Gronenborn, A. M.; Zhu, G.; Klee, C. B.; Ad, B. Solution structure of a Calmodulin-target peptide complex by multidimensional NMR. *Science* **1992**, *256*, 632.

(27) Dunbrack, R. L., Jr; Karplus, M. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.* **1993**, *230*, 543.

(28) Mendes, J.; Baptista, A. M.; Arménia Carrondo, M.; Soares, C. M. Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 530.

(29) Hanf, K. J. M. *Protein design with hierarchical treatment of solvation and electrostatics*, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, September 2002.

(30) Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **1992**, *356*, 539.

(31) Gordon, D. B.; Mayo, S. L. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.* **1998**, *19*, 1505.

(32) Leach, A. R.; Lemon, A. P. Exploring the conformational space of protein side chains using dead-end elimination and the A * algorithm. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 227.

(33) Warwicker, J.; Watson, H. C. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.* **1982**, *157*, 671.

(34) Gilson, M. K.; Honig, B. H. Calculation of electrostatic potentials in an enzyme active site. *Nature* **1987**, *330*, 84.

(35) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(36) Humphrey, W.; Dalke, A.; Schulten, K. VMD − Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33.

CT9004504