An Information Theoretic Approach to Macromolecular Modeling: II. Force Fields

Tiba Aynechi* and Irwin D. Kuntz†
*Graduate Group in Biophysics, and †Department of Pharmaceutical Chemistry, University of California, San Francisco, California

ABSTRACT In this article, we explore the information content of molecular force-field calculations. We make use of exhaustive lattice models of molecular conformations and reduced alphabet sequences to determine the relative resolving power of pairwise interaction-based force fields. We find that sequence-specific interactions that operate over longer distances offer greater amounts of information than nearest-neighbor or non-sequence-specific interactions. In a companion article in this issue, we explored the information content of sequence alignment procedures and the calculation of gap penalties. Both articles have implications for protein and nucleic-acid computations.

INTRODUCTION

Predicting the three-dimensional structure of macromolecules from their sequences remains a fundamental problem of great interest in biology and a most challenging theoretical puzzle. To date, most attention has been focused on proteins. Current protein-folding hypotheses assume that the native structure is the conformational set of lowest free energy when compared to all other accessible conformational ensembles. To test this assumption, the free energy of all conformations must be calculated, an intractable undertaking at present, although progress is being made through massively parallel simulations (1,2). An important aspect of this problem is developing a force field, typically based on atomic interactions, which, in principle, can be evaluated for all conformers. Force fields are developed from experimental geometries and energies of small molecules, often with the help of high-level quantum mechanical calculations (3). The accuracy and validity of force fields is tested by comparison against experiments; however, only in the simplest cases can conformational space be explored exhaustively (4). Our aim here is to present a framework for understanding the discriminatory power of various force fields using information theory. We emphasize, however, that information theory, alone, cannot be used to verify the physical correctness of force-field terms.

In a companion article (5) we use information theory to quantify the information gained during sequence-alignment procedures and show how to extract gap penalty values based on analytical formulations and gap distributions rather than empirical optimization. In this work, we wish to extend our methods toward the understanding of force fields.

Determining the lowest energy conformation from among all possible three-dimensional structures of a given sequence requires an energy function capable of discriminating among native and decoy conformations. Many force fields and scoring functions have been developed for this purpose (6–11). Some are physics-based, concentrating on pairwise or higher-order atomic interactions, while others, such as potentials of mean force, incorporate properties of the ensemble. Alternatively, one can define an empirical score function using training data and then apply these functions to sequences/structures to test their efficacy.

However, there is no agreed measure for evaluating the quality of force fields. It has been shown that most statistically derived potentials are inadequate in the representation of real-world proteins (12,13). In addition, because the structures in the Protein Databank (14) represent only a fraction of the protein conformational space, the question of whether parameter-based methods, or those relying on potentials of mean force, will apply to a new structure that lies outside the training data, remains unanswered. Physics-based methods avoid this particular extrapolation problem, but, as noted above, can fall short due to the computational costs of conformational sampling and the complexity of comprehensive atomic interaction models (15–17). However, the impressive agreement between experimental kinetic data and simulation results for several small proteins using molecular mechanics force fields and explicit solvent models is quite encouraging (2,18). The utility of simplified models depends on the level of resolution required (19). Thus, efficient use and development of future and current methodologies require an in-depth understanding of their behavior and dependencies. As previously demonstrated (5,20), exhaustive two-dimensional lattice models (21) and information theory (22) allow us to move from an empirical regime toward an analytical formulation. We can then quantitatively measure the discriminating power of various scoring functions and force fields, develop metrics for performance analysis, and draw inferences that cross over to real proteins.

In this article, we will examine various force fields for their ability to distinguish among conformations given a set of all possible conformations and all possible sequences.

Submitted January 13, 2005, and accepted for publication August 9, 2005.

Address reprint requests to Dr. Irwin D. Kuntz, Dept. of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA 94143-0446. Tel.: 415-476-1937; Fax: 415-502-1411; E-mail: kuntz@cgl.ucsf.edu.

0006-3495/05/11/3008/09 $2.00

doi: 10.1529/biophysj.105.059618

Combining all conformers and all sequences into one ensemble is the broadest formulation of the problem of macromolecular structure and evolution. It allows us to contemplate such issues as the designability of structures and detailed evolutionary models. Such a formulation is a departure from the traditional focus of an ensemble consisting of all the conformers accessible to a single sequence.

METHODS

Overview

Our basic approach is to write out all possible occurrences of the set of interest (conformations, sequences, etc.) and then ask: What are the informational consequences of performing an operation that combines or clusters some of the objects? In particular, we will determine the energies of a set of conformations using various force fields and cluster the conformations based on the degeneracy of their energies. The information content of the original set of W objects is $\log_2 W$. An *object* in this context will be a specific conformer with a specific sequence. Any clustering will reduce the effective number of objects and hence reduce the information content of the transformed set. Of course, it is not feasible to write out all possible protein or nucleic-acid sequences or all possible macromolecular conformations. Instead, we will make use of walks on two-dimensional lattices and simplified alphabets (21,23,24). However, we are also interested in obtaining results of practical application whenever possible. We must also add the caution that information theory is analogous to statistical mechanics: it provides formal consequences of initial assumptions, but the truth of those assumptions has to be established from additional context. For example, the information content of an electrostatic model in which like charges attract is indistinguishable from the information content of a theory in which like charges repel, but only one of these models describes our experiential universe.

Entropy and information

As before (5), the information content of an ensemble recovered by a constraint set M in bits is defined to be the Shannon entropy (25),

$$I^{\mathrm{M}} = -\Sigma[p_{\mathrm{k}} \log_2(p_{\mathrm{k}})], \tag{1}$$

in which p_k is the population of cluster k expressed as a fraction of the ensemble, summed over all clusters.

The theoretical information content of the ensemble of size, W, is defined as

$$I^{\mathrm{S}} = \log_2(W), \tag{2}$$

in which I^{S} is referred to as the *source* information (25).

There is extensive literature on the relationship between Shannon entropy and thermodynamic entropy (26). In essence, a Shannon entropy can be calculated for clustering operations, whether or not they correspond to physical processes that yield thermodynamic entropies. We will return to these differences, below.

Protein models

We employ the two-dimensional lattice models of Chan and Dill (21) and focus our discussion on proteinlike conformation space. We use the Cartesian (through-space) distance, d, between beads i, j defined as

$$d_{\mathrm{i,j}} = ((x_{\mathrm{i}} - x_{\mathrm{j}})^2 + (y_{\mathrm{i}} - y_{\mathrm{j}})^2)^{1/2}. \tag{3}$$

Chains of beads, each bead representing one residue, are arranged in self-avoiding walks according to the following rules. The elementary step, the distance between consecutive beads, $d_{\mathrm{i,i+1}}$, is fixed at unit length. The move set is limited to a single step with diagonal moves disallowed. Beads cannot overlap (the excluded volume constraint). This set of walks is the same as the exhaustive ensembles of Chan and Dill (21) and Irback and Troein (27) that count all conformations not related by translation, rigid rotation, or reflection. The N-terminus to C-terminus directionality of proteins is preserved in these ensembles.

We will explore two types of conformational sets: *exhaustive*, which contains all conformers allowed by the rules (Table 1), above, and *compact*, the set in which all vertices of an $i \times j = N$ two-dimensional lattice must be occupied (Table 2). All conformations of up to length 26 have been enumerated. We also generate semicompact structures by fitting the N-mer to the next smallest perfect-square lattice (Table 2). Compact lattices were obtained in an efficient manner by modifying the generation program to terminate whenever the i or j limits were exceeded.

Simplified potentials

HP interaction model

The sequence space of proteins grows exponentially as 20^{N} if the natural amino acids are used. To exhaustively explore sequence space, the HP model (28) of residues is used to generate all possible sequences of length N (Table 1). The residues are typed as H (nonpolar) and P (polar). For every sequence, all possible conformations are generated on the two-dimensional lattice and the interaction energy is calculated for each structure according to an interaction rule set. For example, residues can be said to interact if their geometric distance, d_{ij}, within the lattice is one unit length or less, and they do not occupy adjacent positions in the chain. The energy of a structure is lowered by an amount ε if the interacting residues are both of the type H. In the original Dill formulation, HP and PP interactions do not lower the energy of the conformations. The interaction energy, e_{HH}, is

$$e_{\mathrm{HH}} = \sum_{1}^{\mathrm{k}} \varepsilon, \tag{4}$$

where k is the number of interacting HH pairs. The partitioning power of the interaction energy will be determined in bits using Shannon entropy (Eq. 1; also see Appendix for sample calculations). In our initial implementation, we assume that all energy differences are resolvable. In effect, we are measuring the diversity of the energy landscape, that is, the distribution of conformer-sequence pairs that have specific energies. As noted above, we will not get thermodynamic entropies from this characterization because we are assuming the reversible interconversion of all conformers and all sequences in a high temperature limit. We will also obtain the temperature-dependence of the

TABLE 1 HP sequences and self-avoiding two-dimensional walks on a square lattice

Chain length (N)	Number of HP sequences generated	Number of fully enumerated structures	Number of structures generated	Total pairwise (W)	I^{Source} (bits)
4	16	5	5	80	6.322
5	32	13	13	416	8.700
6	64	36	36	2304	11.170
7	128	98	98	12,544	13.615
8	256	272	272	69,632	16.087
10	1024	2034	2034	2,082,816	20.990
12	4096	15,037	15,037	61,591,552	25.876
16	10,000*	802,075	10,000*	10,000†	13.288
25	10,000*	5,768,299,665	10,000*	10,000†	13.288

*Stochastically generated.

†Random sequence/structure pairs.

TABLE 2 Compact and semicompact Hamilton walks

Chain length (N)	Lattice dimensions	Number of HP sequences	Number of compact structure	Total pairwise (W)	I^{Source} (bits)
9	3 × 3	512	5	2560	11.322
12	3 × 4	4096	31	126,976	16.954
16	4 × 4	65,536	69	4,521,984	22.109
24	5 × 5	10,000*	5398	53,980,000	25.686
25	5 × 5	10,000*	1081	10,810,000	23.366
26	6 × 6	10,000*	10,000*	10,000	13.288

*Stochastically generated.

information content when sequences are treated as nonconvertible, a more physically apt assumption (see below).

Solvation model

Interactions between solvent and protein can be difficult to model with any realism. We employ a simple procedure that again uses all possible conformations of N-mer chains on the two-dimensional lattice (Table 1). Residues are classified as buried, exposed, or partially buried. On a two-dimensional lattice, each vertex has a coordination number $z = 4$. A buried residue is one whose three surrounding vertices (other than its immediate predecessor in the chain) in the lattice are occupied by other residues. An exposed residue is surrounded by three unoccupied vertices and a partially buried residue has one or two of the vertices filled. The same sequence/structure pairs used above are used here as well. However, in addition to the HH contact score, a solvation score is also added to each residue's energy contribution. For every filled coordination site, a contribution $\boldsymbol{\xi}$ is added. A fully buried residue would incur an additional term equaling $3\boldsymbol{\xi}$, whereas an exposed residue would have no added term. Again, we will assume that each energy level is fully resolvable, and we count the number of objects (sequence-conformer pairs) in each energy level.

Electrostatic interactions

Electrostatic interactions within molecules play an important role in defining their properties (29). Pairwise attractive and repulsive interactions between atoms are governed by Coulomb's law whose energy, e, depends on charge and Cartesian distance, $e \sim q_i q_j / d_{ij}$. In our simplified model, we allow every residue to have either a unit positive (+) or negative (−) charge. All possible ± sequences of length N are subsequently generated and for each one the Coulomb energy, E_c, of all possible two-dimensional conformations is calculated. We assume the residues to be spherical point charges laid on a two-dimensional lattice yielding the familiar $1/r$ dependence on separation. Alternative two-dimensional representations use disk charges, yielding ln d dependence (K. A. Dill, private communication; Dill and Bromberg (30)). Thus, we will take the Coulomb energy of each conformation as

$$e_{ij} = \frac{q_i q_j}{d_{ij}}$$
$$E = \sum_{i=1}^{N}\sum_{j>i}^{N} e_{ij}. \tag{5}$$

We again use Shannon's equations for entropy to quantify the partitioning power of electrostatic interactions. For a more realistic model, we also define sequences where the sum of ± residues is limited to a representative percentage of the total residues.

Gō-type potentials

Gō-type potentials were proposed by Gō and Taketomi (31,32) to elucidate the effects of long-and short-range interactions during protein folding. In their work, interactions that involve bond lengths and angles are called short-range and often associated with secondary structure formation, whereas long-range interactions are defined, as among residues, nearest-neighbors in space (but far in sequences). Protein structures were stabilized by specific long- and short-range interactions of varying weights during Monte Carlo simulations. It was concluded that native-state stability is achieved through long-range (in space) interactions, whereas folding rates were affected by the short-range interactions. We investigate whether knowledge of long-range interactions, i.e., large space separations, as defined by the Gō potential, alone is sufficient for discriminating among various conformations when averaged over the ensemble of structures. All possible W compact and semicompact conformations of N-mer chains were generated (Table 2). We suppose residues i and j to be interacting if $j > i + 1$ in the chain and they occupy nearest-neighbor vertices on the lattice. The energy of the interacting units is assumed to be identical, with a value of $-\varepsilon$. We define the potential for every structure in the ensemble as the number of interacting pairs in units of ε. The energies of all other conformations are evaluated based on the potential for a target structure. Conformations are scored as follows: Given a target structure with a set of interacting pairs, S_T, the energy of the conformation is lowered by ε for every interacting pair present in both the target structure and the conformer being evaluated. Thus, the total score of any structure is

$$E = \sum_{k} \varepsilon_k, \tag{6}$$

where k equals the total number of shared pairs. All conformations are scored as above, averaged over W target structures. The resolving power is determined by the partitioning effect of the resulting scores on the conformer set.

Temperature dependencies

We carry out two protocols for calculating the explicit temperature-dependence of the information content of the force fields. Such calculations require an assumption relating the energy spacing to kT, where k is the Boltzmann constant and T is the absolute temperature. We assume that the energy is expressed directly in units of kT. We then use Boltzmann statistics to calculate the population of each conformer for each sequence and calculate the entropy of mixing of these conformers. We assume that the conformers can interconvert freely and that the sequences have no thermal pathway for exchange. Calculations of Shannon entropy are carried out as follows: For every conformer, the sequence/conformer pair with the lowest energy (under a particular force field) is defined as the ground state, to calculate the relative energy levels, E_i. We use the Boltzmann probability function (Eq. 7) to obtain probabilities, p_{E_i}, for each energy level. In the first protocol, each conformer is considered to be distinct and we use the energy for each sequence/conformer pair, i (w total pairs), to calculate entropy (Eq. 8). This formulation yields a number proportional to the average conformational entropy, log W, as a function of temperature. Alternatively, we can determine Shannon entropy by clustering the degenerate populations, k, in each distinct energy level, n (Eq. 9; also see Appendix for sample calculations).

$$p_{E_i} = \frac{e^{\frac{-E_i}{kT}}}{\sum_{i}^{w} e^{\frac{-E_i}{kT}}}, \tag{7}$$

$$I_T = \sum_{1}^{w} p_{E_i} \log_2 p_{E_i}, \tag{8}$$

$$p_{E_{i,n}} = \sum_{1}^{k} p_{E_{i,k}}$$
$$I_T = \sum_{i}^{n} p_{E_{i,n}} \log{}_2 p_{E_{i,n}}. \tag{9}$$

RESULTS

We begin by presenting our measures of Shannon information for simplified force fields applied to sets of two-dimensional conformations threaded onto sets of sequences. Using Shannon's entropy for the energy distributions, we measure each force field's classifying power: its ability to distinguish among the full set of conformer-sequence pairs. To do so, exhaustive and stochastic sets of conformers are generated as described in Methods, above. Our models assume that all distances and sequences are known exactly and are free from errors. In general, the most informative force fields are those whose energy functions produce the least degenerate set of values for a given set of conformer/sequence pairs. Our results indicate that force fields with terms that include long-range interatomic distances yield much more information than force fields that make use of pairwise contact potentials only.

HP interactions

Self-avoiding two-dimensional conformations of N bead chains were enumerated (Table 1) and threaded with all possible HP sequences (see Methods). Each conformer/sequence pair was subsequently scored according to Eq. 4 and the Shannon entropy was calculated using Eq. 1. Fig. 1 shows that there is a steady increase in the amount of information retrieved, I^{M}, as N becomes larger (Eq. 10).

The information increase for the fully enumerated sets is approximated by

$$I(HP_{\mathrm{N}}) = 0.61 \times \log_2 N - 0.99. \qquad (10)$$

With increasing N, the conformation space becomes exponentially larger, creating more energy states. However, the rise in information with increasing N is much slower than the rise in the bits of information required to fully classify the ensemble, referred to as I^{S} (see Fig. 1, *inset*).

The properties of the fully enumerated ensembles are dominated by the extended structures, analogous to the denatured state of proteins. To better resemble native and molten globule states (21), we also study subsets consisting of compact and semicompact conformers only. Compact conformers are generated as perfect-square Hamilton walks where every lattice site is occupied. For semicompact structures, the lattice is restricted to the smallest square that fits a chain of length N (Table 2). The percentage of information recovered is higher in the compact and semicompact structures, ~10–15% compared to ~5% for the fully enumerated population (Fig. 2). The HP potential quantifies the number of HH contacts. Since the beads in compact and semicompact geometries (depending on the tightness of the lattice fit) are limited to rectangles, there is a higher occurrence of HH contacts leading to somewhat finer partitioning of the ensemble by the potential.

Solvent interactions

Protein interactions of interest to biochemistry do not occur in a vacuum, but in a matrix of interactions with some form of solvent that influences their energetics. Atom-solvent interactions have been modeled both explicitly (33,34) and in more simplified models (continuum models) (35). Although it has been shown that solvation terms improve theoretical calculations (34), there has never been a quantitative analysis of their contribution.

To explore the information value of solvent interactions, we assume that each conformer is immersed in a uniform solvent. We determine the burial state of each residue within a

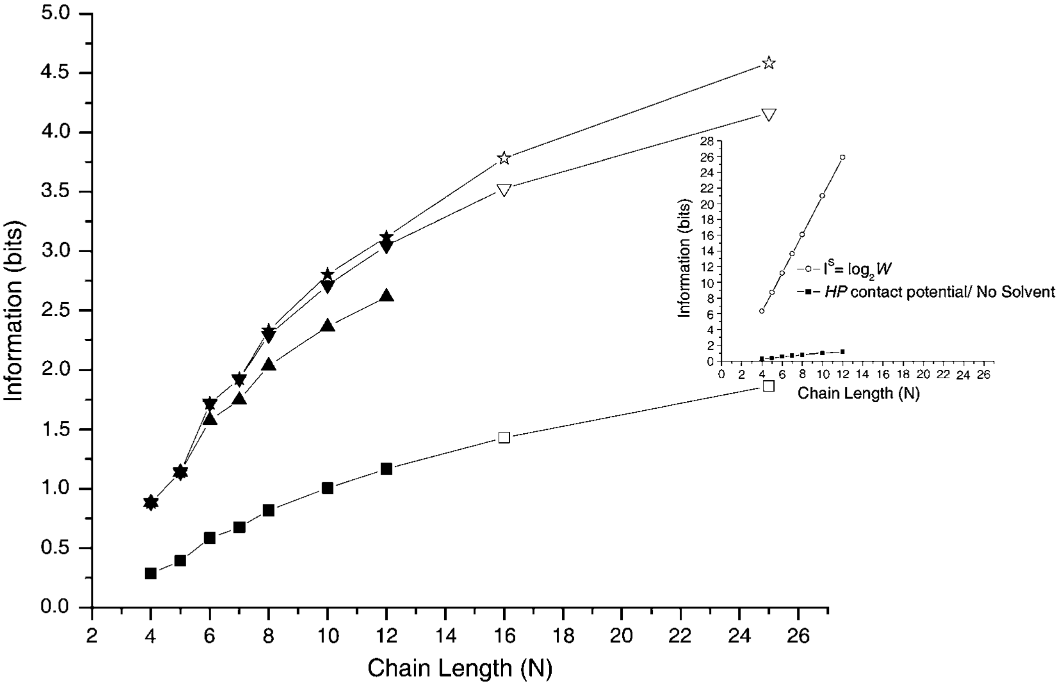

FIGURE 1 Information of HP and solvent contacts. (*Square symbols*, HP contacts; *up-triangles*, solvent with weight scale 0.5; *down-triangles*, solvent with weight scale 1.0; and *stars*, solvent with weight scale 0.2.) Solid symbols, exhaustive set; open symbols, stochastic sample. (*Inset*) Open circle, I^{S}.

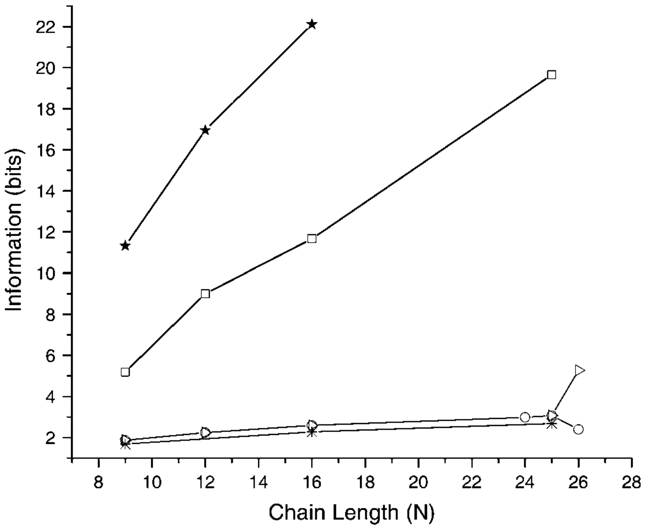

FIGURE 2 Information of force fields for compact structures. (*Stars*, I^S; *diamonds*, Coulomb interactions; *right triangles*, HP + solvent with weight 0.2; *circles*, HP contacts; and *asterisks*, Gō-type potential.)

conformer by counting the number of empty lattice vertices around it. For two-dimensional lattice walks, the coordination number is $z = 4$, resulting in, at most, three additional solvent/bead interactions per residue.

For the fully enumerated sets, there is a ~2.5-fold increase in the amount of recovered information (Fig. 1). We also experiment with varying the weight scale for the solvent score. The amount of information is larger for weights that increase the score separation among conformers of varying solvation states. The separation of the solvation curves (Fig. 1) for varying weights becomes larger with increasing N because the more diverse conformer sets reduce the scoring degeneracies.

As expected, the compact structures show only marginal rises in information when the energy score includes a solvation score (Fig. 2). By design, the compact structures minimize the number of exposed residues by maximally filling all lattice vertices. The only exposed residues are those placed on the four lattice edges. We observe a consistent but small rise in information with increasing N. More information is recovered in the semicompact structures (modeling packing defects) which may be physically relevant to protein globular states. The smaller the ratio of N to the lattice dimensions, the larger the enrichment.

Electrostatic interactions

Biological processes are often governed by long-range electrostatic interactions (29). These distance-dependent energies are modeled by Eq. 5. Although the HP force fields mimic short-distance interactions, the distance-dependent function can be used to illustrate the discrimination power of long-range (large-space separations) pairwise interactions.

For a set of two-dimensional conformers, we assign every lattice point as either a positive or negative charge. All possible ± combinations are explored for a given set of conformers. The energy of each lattice and subsequent entropy per set are evaluated according to Eqs. 5 and 1, respectively. For exhaustive sets of fully enumerated conformers we observe that, on average, close to 80% of the maximum information is retrieved. Furthermore, electrostatic screening among residues, modeled by a $1/r^2$ potential, offers nearly the same amount of information (Fig. 3). It is also worth mentioning that the information connected with various terms in common force fields is not additive. Instead it is only as informative as its most discriminating descriptor. For our simple models, where both a Coulomb energy function and a solvent potential function are used, there is no significant additional information supplied by the latter term (Fig. 3). For compact structures, the amount of recovered information is slightly >50% of I^s (Fig. 2). Comparison of the performance of the pairwise energy function on compact versus the fully enumerated set seems to indicate that extended structures are better described by long-range electrostatic terms than compact structures. Because the Coulombic energy term is a sum over all residue pairs in the lattice, this function will fail to discriminate among compact conformers with the same sequence elements placed in different chain positions. In effect, the constraint for compactness reduces the effective number of conformers to one, and highlights the energy differences among sequences.

One could argue, however, that the relative increase in information from the Coulomb energy is an artifact of our highly charged model. Our experiments using partially charged sequences (20% random charge on 10-mer sequences) show

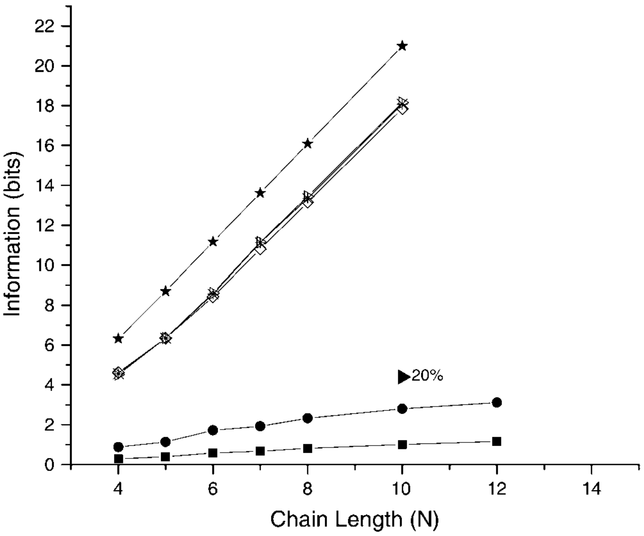

FIGURE 3 Information from Coulombic pairwise interactions. (*Stars*, I^S; *right-triangles*, Coulomb energy + solvent; *asterisks*, $1/r$ Coulomb energy; *diamonds*, $1/r^2$ dielectric screening; *squares*, HP contacts; *circle*, HP + solvent; and *solid right triangle*, Coulomb with 20% randomly charged sequences.)

that a single pairwise Coulomb term per sequence enriches the information retrieved by four-times the HP model, and double compared to the solvent model (Fig. 3, Table 3). This demonstrates the richness of long-range spatial distances from an information point of view.

Gō-type potential

The force fields, above, utilize two different descriptors for each conformer/sequence pair: 1), A geometric descriptor, i.e., Cartesian coordinates; and 2), a bead descriptor, i.e., residue type, charge. Consequently, two pairs of residues occupying the same chain position and lattice points can yield different energy contributions. In dealing with a Gō-type potential we only consider geometric descriptors, in particular pairs of interacting residues, not adjacent in chain position but occupying neighboring lattice vertices (Eq. 6). For every target conformer (see Methods) we consider all possible interactable pairs, the equivalent of nonspecific interactions. Although all compact conformations adhere to the square shape, the chain trace on the lattice is different, and as a result most conformers share few similar contacts. As a result, partitions can be either highly populated or sparse; to be part of a partition, each of the members must exhibit the same degree of dissimilarity to the target structure as others (resemblance to the other cluster members is neither necessary nor required). Fig. 2 shows that for compact conformers, Gō-type potentials offer only a small amount of information, similar to the HP contact and solvent potentials. This is consistent with the absence of unique clusters containing similar structures. Given a target structure, the Gō potential is capable of identifying a similar structure or structures, based on the relative energies. However, it cannot effectively describe an ensemble by differentiating among its members.

Temperature dependencies

We explored the formal temperature-dependence of the information content by calculating the populations of energy levels as a function of kT/E, where E is the energy difference. We then clustered the conformer populations at each temperature under the two protocols described in Methods. The first protocol parallels the calculation of conformational entropy and looks at the population of each conformer as a function of temperature. It leads to the expected result that, at high temperatures, the entropy is just $R\ln W$, where W is the number of conformers for the system under study. Under this protocol, the conformational entropy and the Shannon information content are independent of the force field at sufficiently high temperatures (Fig. 4). On the other hand, the details of the information content at intermediate temperatures reflect the distributions of energy levels and consequently offer an alternative method for characterizing force fields.

The second protocol measures the distribution of the system by energy level. These results are more complicated. The high-temperature results are related to those given earlier in the article, with the difference being that the earlier values are based on interconversion of sequences while the temperature-dependent results in this section do not allow interconversion. Depending on the energy levels, the Shannon information can actually go through a maximum as a function of temperature (Figs. 5 and 6). These results are quite sensitive to the details of the force fields and parallel the general trends seen above. The simulation reported in Figs. 5 and 6 for the Coulombic force field corresponds to Coulombic energies scaled for a dielectric constant of 80 and a bead-to-bead separation of 4 Å to represent protein spacing of charges in an aqueous environment. Decreasing the dielectric constant or the distance would reduce the information content at a fixed temperature, but would not alter the high-temperature limits.

TABLE 3 Average information per residue for various force fields (*kT* independent)

FF type	*I* (bits)/residue Compact	 Exhaustive
I^S	1.403	2.099
Gō-type	0.107	—
HP	0.123	0.101
HP + Solvent (0.2)	0.123	0.280
HP + Solvent (1.0)	0.123	0.271
Coulomb 20% (1/*r*)	—	0.440
Coulomb 100% (1/*r*)	0.786	1.807
Coulomb 100% (1/*r*) + Solv	—	1.815

DISCUSSION

Score matrices and force fields, used in sequence and structure alignments respectively, are valued for their

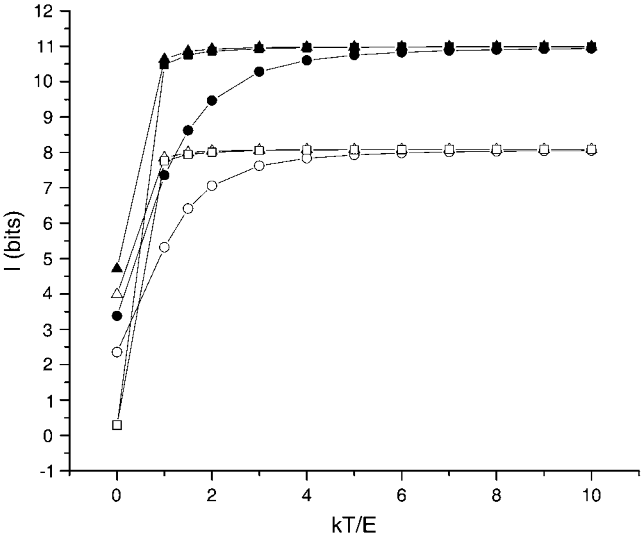

FIGURE 4 The *kT* dependence of information, based on conformer populations averaged over all sequences. (*Squares*, Coulomb; *circles*, HP + solvent; and *up-triangles*, HP.) Solid symbols, 10-mers; open symbols, 8-mers.

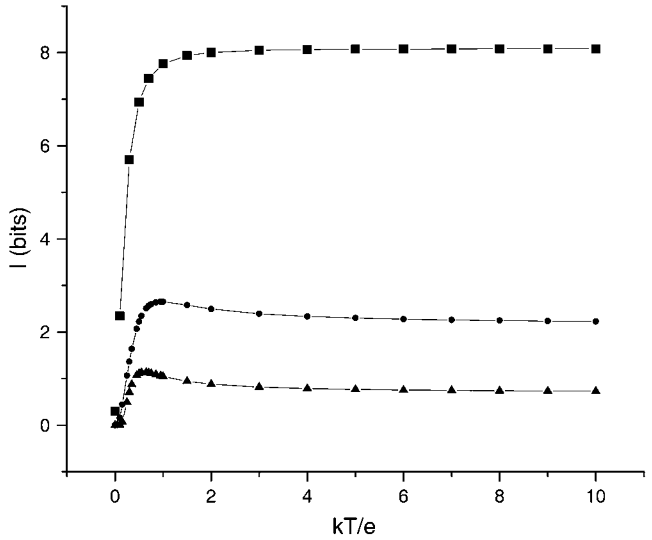

FIGURE 5 Information of force fields for 8-mer exhaustive conformers with kT dependence using energy levels averaged over all sequences. (*Squares*, Coulomb; *circles*, HP + solvent; and *up-triangles*, HP.)

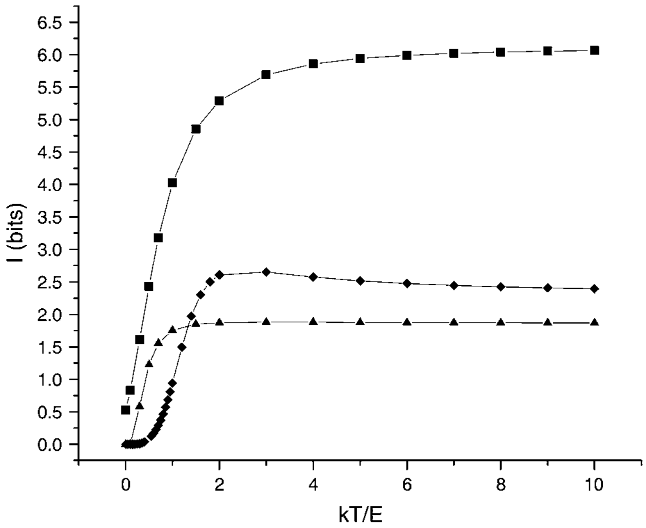

FIGURE 6 Information of force fields for 16-mer compact conformations with kT dependence using energy levels averaged over all sequences. (*Squares*, Coulomb; *diamonds*, Gō; and *up-triangles*, HP.)

discriminating power when faced with choices among similar sequences or conformers. Their resolution is determined by both sensitivity and selectivity thresholds. By using information theory we are able to quantify the resolving power of several basic force fields in bits. Although our metric cannot comment on the correctness of the physical assumptions, such measures could indicate whether there is enough information in the force field to serve a particular purpose. Depending on the particular task at hand, one might ask whether a force field can discriminate well between open and closed conformations or various compact states. Our data show that contact-based interactions (i.e., solvent/solute, HP, and Gō-type potentials) have much lower resolving power than interactions over larger distances, such as Coulomb forces, even when the number of parameters is small. Further, distance-dependent potentials retain this advantage as the number of parameters (e.g., specific amino-acid interactions) increases.

Another issue is how orthogonal are the terms included in force fields. For example, we see that adding solvation terms and/or distance-dependent terms greatly expands the information content compared to simple square-well representations. We have certainly overestimated the importance of Coulombic contributions in our simple model by treating every residue as ionic. However, the principle (given our results from the partially charged model) remains clear that any distance-dependence expands the resolving power of the force field (Fig. 3).

It is important to note that the information content of a force field, per se, is not a direct measure of the utility of a force field for a particular task. Force fields such as the HP and Gō models were developed to simplify the description of complex systems by reducing the degrees of freedom. Their usefulness is judged on how well they can illuminate specific features of these systems. Our interest is to provide a way to quantitate the information being used in each setting.

TABLE 4 Sample calculations of information content for the temperature-dependent HP potentials

			HH						Eq. 8			Eq. 9
			energy		$e^\wedge$	$kT = 1\ \Sigma e^\wedge$		pE_i	$-\Sigma pE_i$		$pE_{i,n}$	$-\Sigma pE_{i,n}$
i	*n*	*Seq*	(*kT*)	E_i	$(-E_i/kT)$	$(-E_i/kT)$	pE_i	$\log pE_i$	$\log pE_i$	$pE_{i,n}$	$\log pE_{i,n}$	$\log pE_{i,n}$
1	1	HHHH	−1	0	1.000	2.472	0.405	−0.528	2.164	0.405	−0.528	0.974
2	2	HHHH	0	1	0.368		0.149	−0.409		0.595	−0.445	
3		HHHH	0	1	0.368		0.149	−0.409				
4		HHHH	0	1	0.368		0.149	−0.409				
5		HHHH	0	1	0.368		0.149	−0.409				
6	**1**	**HHHP**	**0**	**0**	**1.000**	**5**	**0.200**	**−0.464**	**2.322**	**1.000**	**0.000**	**0.000**
7		**HHHP**	**0**	**0**	**1.000**		**0.200**	**−0.464**				
8		**HHHP**	**0**	**0**	**1.000**		**0.200**	**−0.464**				
9		**HHHP**	**0**	**0**	**1.000**		**0.200**	**−0.464**				
10		**HHHP**	**0**	**0**	**1.000**		**0.200**	**−0.464**				
						$I = [4 * (2.164) + 12 * (2.322)]/16$:			2.282537057	$I = [4 * (-0.973) + 12 * 0]/16$:		0.243

We return to the question addressed in the final section of the previous article: how does the information content of force fields compare to the information available from sequence alignment or from a large number of structural measurements? Although we cannot answer this question in analytical detail, we can use a back-of-the-envelope estimate for two-dimensional lattices based on the idea that the information content of a force field is the $\log_2$ (number of energy levels). For an HP or Gō force field on a compact 10×10 two-dimensional lattice, the global-minimum ground state has a energy of -81 units. Assuming that every energy level up to zero is obtainable with alternate sequences, the full conformer-sequence space would yield only 6.3 bits for the protein or 0.06 bits/bead. We obtain a value of ~0.1 bits/bead based on a 5×5 model (in agreement with the ground state of -16). Our calculations also show a distance-dependent force field for a two-dimensional compact structure yields <1 bit/bead (Table 3). In the previous article (5), we estimated the information to be ~3.5 and 2.5 bits/residue for sequence and structural alignments, respectively. Thus, it appears that force fields have much less information content than alignment procedures. This striking result is actually just a reflection of the well-known multiple minima problem that has been a major source of difficulty in the protein-folding field. That is, a force field provides enough information to refine a structure locally but needs to be augmented by extensive sampling or modeling to find the global minimum from a large number of minima with similar energies.

Conclusions on the relative resolving power of force fields, the information content of various interactions, and the additivity of information appear extendible to real proteins. The move from simple to exact models will shift the reference states but not the general trends. Such quantitative assessments are critical for improving the effectiveness of current force fields and score functions used in various alignment protocols.

APPENDIX

Example: information content of an HP potential

The information content of an HP potential, for a given ensemble is calculated by partitioning the ensemble based on the ε_{HH} distribution. The fraction of the ensemble having a particular value for ε_{HH} defines the value for $p_{\mathbf{k}}$. The indexing length for k is determined by the number of energy states in the ensemble.

For example, for a chain of length $N = 4$, there are five distinct conformers and 16 possible HP sequences, resulting in 80 sequence/conformer pairs in the ensemble. We define the interacting pairs as above (see HP Interaction Model, above). From the five conformers, only the fully bent conformer is capable of having an HH contact, and of the 16 possible sequences, only four have an H in both the first and last positions, resulting in two partitions in the ensemble—namely, one with an energy of 1 and probability $p_k = 4/80$, and another with energy 0 and probability 76/80. The information is determined as follows: The number of conformer/sequence pairs in the ensemble, W, is 80. Since the only allowed energies are 0 and 1, $k = 2$, $p_0 = 76/80$, and $p_1 = 4/80$.

Using Shannon's equation,

$$I(HP) = -[0.05(\log_2(0.05)) + 0.95(\log_2(0.95))] = 0.29\,\text{bits}.$$

The information content of other force fields is measured in a similar manner, by substituting the appropriate scoring function to determine the energies of the conformers.

Example: information content of an HP potential—temperature-dependence

For a 4-mer HP model, there are five conformers and 16 different sequences resulting in 80 sequence-conformer pairs. The 4-mer sequence/conformer pairs result in two types of subpopulations. The first type has two populated energy levels, while the other has only a single populated energy level (separated by data in *boldface type* in Table 4). All sequences fall in one of two categories, and so for brevity we only show one of each and multiply the results by the corresponding number of sequences in each subtype (four for type *1* and 12 for type *2*). Table 4 shows HP contact energies calculated according to Eq. 4 of the text and the Boltzmann distributions determined by Eqs. 8 and 9. For every sequence, the sequence/conformer pair with the lowest energy is said to be the *ground state*. Calculations are carried out as described in Methods. The resulting average entropies calculated according to Eqs. 8 and 9 are shown in the bottom row.

We are grateful for helpful discussions with Scott Pegg and Kevin Masukawa.

This work was supported by a grant from the National Science Foundation (NSF No. CHE-0118481), R. Kip Guy, Principal Investigator; a National Institutes of Health grant (No. RR019864), C. Pancerella, Principal Investigator; and a University of California President's Dissertation Year Fellowship (to T.A.).

REFERENCES

1. Zagrovic, B., C. D. Snow, M. R. Shirts, and V. S. Pande. 2002. Simulation of folding of a small α-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* 323:927–937.
2. Snow, C. D., E. J. Sorin, Y. M. Rhee, and V. S. Pande. 2005. How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* 34:43–69.
3. Kollman, P. 1993. Free energy calculations—applications to chemical and biochemical phenomena. *Chem. Rev.* 93:2395–2417 [Review].
4. Kuntz, I. D., and D. A. Agard. 2003. Assessment of the role of computations in structural biology. *Adv. Protein Chem.* 66:1–25.
5. Aynechi, T., and I. D. Kuntz. 2005. An information theoretic approach to molecular modeling: I. Sequence alignments. 89:2998–3007.
6. Halgren, T. A. 1995. Potential energy functions. *Curr. Opin. Struct. Biol.* 5:205–210.
7. Sippl, M. J. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5:229–235.
8. Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. A new approach to protein fold recognition. *Nature.* 358:86–89.
9. Park, B., and M. Levitt. 1996. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.* 258: 367–392.
10. Tenette, C., J. C. Smith, and C. M. Topham. 1996. Force field development and conformational search strategy in the simulation of biomolecular recognition processes. *Biochem. Soc. Trans.* 24:268–274.
11. Bahar, I., and R. L. Jernigan. 1997. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* 266:195–214.
12. Park, B. H., E. S. Huang, and M. Levitt. 1997. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. *J. Mol. Biol.* 266:831–846.

13. Thomas, P. D., and K. A. Dill. 1996. Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* 257:457–469.

14. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

15. Feldman, H. J., and C. W. Hogue. 2002. Probabilistic sampling of protein conformations: new hope for brute force? *Proteins.* 46:8–23.

16. Sullivan, D. C., and I. D. Kuntz. 2004. Distributions in protein conformation space: implications for structure prediction and entropy. *Biophys. J.* 87:113–120.

17. Habeck, M., M. Nilges, and W. Rieping. 2005. Replica-exchange Monte Carlo scheme for Bayesian data analysis. *Phys. Rev. Lett.* 94:018105.

18. Shirts, M. R., and V. S. Pande. 2005. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* 122:134508.

19. Huber, T., and A. E. Torda. 1998. Protein fold recognition without Boltzmann statistics or explicit physical basis. *Protein Sci.* 7:142–149.

20. Sullivan, D. C., T. Aynechi, V. A. Voelz, and I. D. Kuntz. 2003. Information content of molecular structures. *Biophys. J.* 85:174–190.

21. Chan, H. S., and K. A. Dill. 1989. Compact polymers. *Macromolecules.* 22:4559–4573.

22. Young, J. F. 1971. Information Theory. Butterworth & Company, Bristol, UK.

23. Solis, A. D., and S. Rackovsky. 2002. Optimally informative backbone structural propensities in proteins. *Proteins.* 48:463–486.

24. Dill, K. A., S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. 1995. Principles of protein folding—a perspective from simple exact models. *Protein Sci.* 4:561–602.

25. Shannon, C. E. 1948. A mathematical theory of communication. *Bell Sys. Tech. J.* 27:379–423,632–656.

26. Schneider, T. D. 2000. Evolution of biological information. *Nucleic Acids Res.* 28:2794–2799.

27. Irback, A., and C. Troein. 2002. Enumerating designing sequences in the HP model. *J. Biol. Phys.* 28:1–15.

28. Lau, K. F., and K. A. Dill. 1989. A lattice statistical-mechanics model of the conformational and sequence-spaces of proteins. *Macromolecules.* 22:3986–3997.

29. Honig, B., and A. Nicholls. 1995. Classical electrostatics in biology and chemistry. *Science.* 268:1144–1149.

30. Dill, K. A., and S. Bromberg. 2003. Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology. Garland Science, New York.

31. Gō, N., and H. Taketomi. 1978. Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. USA.* 75:559–563.

32. Taketomi, H., Y. Ueda, and N. Gō. 1975. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* 7:445–459.

33. Levy, R. M., and E. Gallicchio. 1998. Computer simulations with explicit solvent: recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.* 49:531–567.

34. Kollman, P. A., I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham 3rd. 2000. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* 33:889–897.

35. Bashford, D., and D. A. Case. 2000. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* 51:129–152.