PROTEINS: Structure, Function, and Bioinformatics 60:82–89 (2005)

Recognizing Protein Folds by Cluster Distance Geometry

Gordon M. Crippen*
College of Pharmacy, University of Michigan, Ann Arbor, Michigan

ABSTRACT **Cluster distance geometry is a recent generalization of distance geometry whereby protein structures can be described at even lower levels of detail than one point per residue. With improvements in the clustering technique, protein conformations can be summarized in terms of alternative contact patterns between clusters, where each cluster contains four sequentially adjacent amino acid residues. A very simple potential function involving 210 adjustable parameters can be determined that favors the native contacts of 31 small, monomeric proteins over their respective sets of nonnative contacts. This potential then favors the native contacts for 174 small, monomeric proteins that have low sequence identity with any of the training set. A broader search finds 698 small protein chains from the Protein Data Bank where the native contacts are preferred over all alternatives, even though they have low sequence identity with the training set. This amounts to a highly predictive method for *ab initio* protein folding at low spatial resolution. Proteins 2005;60:82–89.**

Key words: interresidue contacts; fold recognition potential; distance matrix; monomeric proteins; energy minimization; zero-filling; thermal denaturation

INTRODUCTION

Distance geometry is a way to deal with geometric problems in terms of distances between points, rather than using angles, coordinates, etc. In chemical applications the points are usually atoms, and although it is simple to calculate interatomic distances from atomic coordinates, it is also possible to go from limited information on interatomic distances to a sampling of configurations in terms of coordinates that are consistent with the given constraints.[1] Distance geometry can be generalized to treat disjoint sets of points rather than individual points, and the distances between pairs of points become the sums of squares of all distances between each pair of point sets. This lends itself to treating protein conformations where the polypeptide chain is broken up into blocks each containing a given number of sequentially adjacent residues, and the distance between two such blocks is the sum of the squares of the distances between the two sets of C^{α} atoms.[2] Breaking up all conformation space into regions defined in terms of upper and lower bounds on these generalized distances, it is possible to devise a simple contact energy function such that the statistical weight of the region containing the experimentally determined native conformation predominates over the that of the nonnative regions at some sufficiently low temperature.[3]

Qualitatively speaking, this problem has been addressed by many people over decades as what one might call the structure recognition problem: given the correct native conformation for some amino acid sequence and an assortment of incorrect or decoy conformations, devise an energy-like function that gives a lower value for the correct conformation over all alternatives, preferable for many different sequences. While there are many different ways to produce such functions (such as refs. 4–24) and many different ways to devise decoys (such as refs. 13, 16, 23, 25–28), a key assumption is that somewhere on the order of 10^5 decoys can adequately cover all possible conformations. However, suppose that it is sufficient to describe polypeptide conformations using two variables per residue, and there are 100 residues. Even sampling all corners of a 200 dimensional cube in such a conformation space would involve $2^{200} = 1.6 \times 10^{60}$ decoy structures, and surely this is inadequate. Indeed, we have shown that local optimization of the potential function with respect to conformation is easily able to locate nonnative structures having much superior function values, even for widely respected protein recognition potentials.[29]

The guiding principles in this work are that (1) conformation space needs to be divided into regions covering whole ranges of conformations rather than relying on a wide scattering of individual conformations, and (2) it may well be advantageous to represent protein structures at low resolution to reduce the dimensionality of the conformation space. This second point is the key motivating factor behind the current work. It is computationally infeasible to thoroughly explore all possible, self-avoiding, conformations of a polypeptide chain having $n = 100$ residues, even at a resolution of one point per residue. In terms of interpoint distances, the space to be searched has dimension $O(n^2)$. Although one may be able to solve the structure recognition problem for some sampling of thousands of nonnative structures, it remains a strong possibility that the result is a spurious fit to a woefully inadequate coverage of conformation space. On the other hand, if the

Grant sponsor: the Vahlteich Research Award Fund of the College of Pharmacy, University of Michigan.

*Correspondence to: Gordon M. Crippen, College of Pharmacy, University of Michigan, Ann Arbor, MI 48109-1065. E-mail: gcrippen@umich.edu

Received 25 October 2004; Accepted 24 January 2005

Published online 28 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20488

chain is grouped into blocks of b sequentially adjacent residues, the conformation space now has dimension $O(n^2/b^2)$, and the same finite sampling of nonnative structures is orders of magnitude closer to being adequate. Therefore, in this work we will not consider $b = 1$ models. Of course, there remains a tradeoff between a more feasible sampling problem and a cruder approximation to intramolecular interactions. At a ridiculous extreme, consider $b = n$. Now in the cluster distance geometry analysis, there is only one degree of conformational freedom, which corresponds to the radius of gyration of the polypeptide chain, and the search over conformation space is extremely easy. However, the single cluster of n residues is characterized only by amino acid content, rather than sequence, so the energy-like function is unable to discriminate between folding and nonfolding sequences having similar amino acid contents. Here we will explore models between the $b = 1$ and $b = n$ extremes, looking for enhanced conformational coverage while still retaining enough sequence discrimination to perform well at structure recognition.

METHODS

Sequence Partitioning

Here we are developing a model for protein folding at low spatial resolution. So far we have considered only soluble globular proteins consisting of a single polypeptide chain without substantial ligands or disulfide bridges, so that the stability of the native state can be attributed to intrachain interactions between residues and solvation of residues in an unspecified generic aqueous environment. Attention has been restricted to chain lengths $n \leq 128$ residues for the time being in order to keep the CPU and memory requirements within reason. There is no intrinsic reason why this limit could not be raised in the future. Instead of representing the chain by all its atoms, or all its nonhydrogen atoms, or even just the C^α atom of each residue, we group sequentially adjacent blocks of b residues together and represent each block as a cluster of b points located at the respective C^α atoms. Each point is labeled according to the type of the residue, considering only proteins having the standard 20 residue types.

Previously[3] we adopted the zero-filling approach often seen in signal processing, where shorter chains had imaginary noninteracting residues appended to the C-terminus to bring the total up to 128, and then all chains were segmented into blocks of eight residues. As a shorter example, suppose $b = 3$ and a given chain had seven residues denoted by 1234567. Zero-filling to nine residues gives 123456700 with two imaginary residues ("0") appended, so that it can be segmented evenly into 123|456|700. One problem is that the last segment has only one real residue while the others have three. The other is that perhaps a phase shift in segmenting would give more meaningful groups of residues. For instance, if residues 1 and 2 are disordered but 3 is the beginning of a helix, then having |345| as a block may be better. Here we do no zero filling but rather truncate residues as necessary from the N- and C-termini to have an even number of blocks. Thus, 1234567 becomes 123|456 or 234|567 or 345, and in subsequent calculations we consider equally all three alternative segmentations for all chains of seven residues. Such segmentation rather than zero-filling turns out to be essential to producing models that are successful at structure recognition.

In general, for chain length n and block size b, as long as $n \geq 2b - 1$, there will always be b different segmentations of the chain, of which $1 + n \bmod b$ of the segmentations will consist of $\lfloor n/b \rfloor$ blocks, and the rest will have $\lfloor n/b \rfloor - 1$ blocks. (The "floor" notation $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.) Any segmentation involves deleting fewer than b residues each from both ends. In this work we have used $n \leq 128$ and $b = 4$, 8, or 15.

Contact Energies

For a given conformation of a polypeptide chain, let d_{ij} be the distance between the C^α atoms of residues i and j. For a given segmentation of the chain, we summarize its conformation in terms of the b^2 squared distances between blocks I and J.

$$D_{IJ} = \sum_{i \in I} \sum_{j \in J} d_{ij}^2 \tag{1}$$

Note that intrablock distances D_{II} will tend to be smaller than interblock distances, especially when I and J are well separated in sequence. A simple sort of energy function would count all the residues of the two blocks to be in contact when the D_{IJ} fell below some cutoff, and otherwise they would not be interacting. If the cutoff is too low, nearly all blocks would not be in contact, whereas if the cutoff is too high, nearly all would be in contact. Using the nonredundant subset of polypeptide chains extracted from the Protein Data Bank[30] (PDB) and furnished with MOE[31] version 2004.03, we abstracted 1860 chains having $n \leq 128$. We will refer to these as the "chain set" (see Table I). Zero-filling each chain to the nearest multiple of the block size, there are many alternative interblock distance matrices for a given chain length. The cutoffs listed in Table II are those that produced the largest total number of contact/noncontact differences between comparable matrices. Of course the cutoffs depend strongly on b, and the offdiagonal values are much larger than the diagonal ones. Generally C_{II} is somewhat smaller than the observed upper bound on D_{II} for all-helical segments.[2] Note that these cutoffs are general properties of a large set of PDB structures without regard to the types of residues involved in contacts, so there is actually no contradiction between conveniently comparing zero-filled structures here and using multiple segmentations in all other aspects of this study.

A strictly binary cutoff for contact versus no contact can cause problems where small changes in conformation give large changes in contacts and hence energy. The usual solution is some sort of sigmoidal contact function $c(D_{IJ})$ that is 1 for short distances, 0 for long distances, and smoothly varying in between. Here we used a continuous, piecewise linear function

TABLE I. Sets of Proteins

Name	No. proteins	Composition
Chain set	1860	Derived from the MOE[31] nonredundant chain database of 7000 high-resolution PDB structures having less than 90% sequence identity. Selected those having $n \leq 128$, consisting of only the 20 standard amino acids where all residues have C^{α} coordinates.
Monomer set	262	Monomeric biological unit, no disulfide bridges or substantial ligands. $50 \leq n \leq 128$, compact radius of gyration PDB codes: 1a6s, 1aa3, 1adr, 1af8, 1ah9, 1aoy, 1aps, 1ark, 1auz, 1aw0, 1awj, 1b4r, 1b64, 1bax, 1blj, 1bo9, 1cb1, 1cdz, 1cpz, 1d1n, 1d6t, 1d8b, 1d8j, 1dcj, 1dgn, 1doq, 1dro, 1du6, 1e41, 1edi, 1egx, 1ehx, 1eik, 1eiw, 1eo1, 1eqk, 1ewi, 1f0z, 1f68, 1f6v, 1faf, 1fc1, 1fex, 1fho, 1fj7, 1fjc, 1fjd, 1fli, 1fna, 1fo5, 1fvq, 1fyj, 1g10, 1g2h, 1g6p, 1g7d, 1g7e, 1gb4, 1gh8, 1ghc, 1ghh, 1ghj, 1gjx, 1gl5, 1gxh, 1gxi, 1gyz, 1h5p, 1h67, 1h8c, 1h95, 1hce, 1hd0, 1hdj, 1hks, 1hqb, 1hqi, 1hs7, 1hsq, 1ig6, 1iio, 1ilo, 1ily, 1ipg, 1iqs, 1iqt, 1irz, 1itp, 1iv0, 1iyu, 1j0g, 1j2m, 1j3t, 1j7q, 1j8k, 1jei, 1jfw, 1jns, 1jrm, 1jt8, 1jw2, 1jxs, 1k5k, 1k85, 1k8b, 1ka5, 1kkg, 1klp, 1kmd, 1kom, 1ksr, 1kvi, 1kvn, 1l5i, 1l7b, 1l7y, 1lab, 1ll8, 1lq7, 1lwr, 1m5z, 1mc7, 1mg8, 1mjd, 1mp1, 1mvg, 1mwy, 1myo, 1n27, 1n87, 1n88, 1n91, 1neq, 1ngr, 1nho, 1nr3, 1nso, 1ny8, 1nz8, 1nz9, 1o1u, 1o78, 1oo3, 1plt, 1p68, 1p6r, 1p8g, 1p97, 1p9k, 1pav, 1pba, 1pc0, 1pfj, 1pls, 1pmr, 1pqx, 1pse, 1puz, 1pve, 1q02, 1q1o, 1q7x, 1qjo, 1qjt, 1qlc, 1qly, 1qp2, 1qqv, 1qzp, 1r4g, 1r57, 1rdu, 1rgw, 1rja, 1rlf, 1ryj, 1ryk, 1ryu, 1rzs, 1s79, 1s7a, 1sb6, 1sq8, 1sro, 1t0g, 1t1h, 1tac, 1tiv, 1tiz, 1tk7, 1tns, 1u2f, 1uaw, 1uc6, 1ucp, 1ucv, 1uep, 1ueq, 1uew, 1uez, 1ufm, 1ufn, 1ufw, 1ufx, 1ug0, 1ug1, 1ug7, 1ug8, 1ugv, 1uh6, 1uhc, 1uhf, 1uhp, 1uhr, 1uht, 1uhu, 1uhw, 1uhz, 1ujd, 1ujs, 1ujt, 1uju, 1ujv, 1ujy, 1ul7, 1um1, 1um7, 1uqv, 1uss, 1v2y, 1v31, 1v32, 1v38, 1v3f, 1v5j, 1v5k, 1v5l, 1v5p, 1v5s, 1v5t, 1v5u, 1v62, 1v6b, 1v89, 1vb7, 1vig, 1vye, 1wit, 1wot, 1yua, 2a3d, 2bby, 2bjx, 2cjn, 2fmr, 2fnb, 2mss, 2u1a, 2u2f, 2vik, 3crd, and 3hck.
Training set	211	monomer set except for 51 problematic proteins: 1ah9, 1aoy, 1aps, 1auz, 1b64, 1blj, 1bo9, 1cdz, 1cpz, 1du6, 1eqk, 1f0z, 1f6v, 1fex, 1fyj, 1gxh, 1hks, 1hqi, 1irz, 1iv0, 1iyu, 1jt8, 1jw2, 1k8b, 1ksr, 1l7b, 1lab, 1n87, 1ngr, 1nho, 1o1u, 1oo3, 1pc0, 1pqx, 1puz, 1qlc, 1qqv, 1ryj, 1rzs, 1sb6, 1sq8, 1t1h, 1tk7, 1uhz, 1ujs, 1ul7, 1yua, 2a3d, 2bjx, 2fmr, and 2u2f

TABLE II. Diagonal (C_{II}) and Offdiagonal (C_{IJ}) Contact Distance Cutoffs as a Function of Block Size (b)

b	C_{II} (Å^2)	C_{IJ} (Å^2)
4	364.3	5398.1
8	3553.4	22077.3
12	13197.2	51150.8
15	27028.4	86166.3

$$c(D_{IJ}) = \begin{cases} 1 & \text{for } D_{IJ} < (1-w)C_{IJ} \\ \dfrac{-D_{IJ}}{2wC_{IJ}} + \dfrac{w+1}{2w} & \text{otherwise} \\ 0 & \text{for } D_{IJ} > (1+w)C_{IJ} \end{cases} \quad (2)$$

where the relative half-width of the transition is $0 < w < 1$, and always $c(C_{IJ}) = 1/2$. In what follows we take $w = 0.2$ usually.

For the 20 standard amino acid types, there are $20 \times 21/2 = 210$ unordered residue pair types. Let $\boldsymbol{t}_{IJ}$ be a vector of 210 elements, where each element is the number of the corresponding type pairs found for one residue in block I and the other in block J. Thus, the sum of the elements of $\boldsymbol{t}_{IJ}$ is b^2 for $I \neq J$, but for intrablock interactions, we count only interactions between residue pairs i to $i + 4$ and beyond. In either case, we use the same vector of 210 adjustable parameters $\boldsymbol{a}$ to convert contact counts to an energy-like value

$$E = \sum_{s=1}^{b} \sum_{I \leq J} (\mathbf{a} \cdot \mathbf{t}_{s,IJ}) c(D_{s,IJ}) \quad (3)$$

summing over all segmentations s for chain length n of the given conformation of the protein. In the end, the energy is just a linear function of the adjustable parameters, although a nonlinear function of conformation. We find it essential to consider all segmentations, rather than choosing just one or adopting a zero-filling approach.

Given that the energy function has a myopic view of conformation that emphasizes close contacts but does not reflect how distant noninteracting blocks are, a matching measure of conformational similarity would be more appropriate than the usual RMSD (root-mean-square deviation in corresponding C^{α} coordinates after optimal rigid-body superposition). Here we have used the simple dissimilarity measure between two conformations A and B,

$$S(A, B) = k^{-1} \sum_{s} \sum_{l \leq J} |c(D_{A,s,IJ}) - c(D_{B,s,IJ})| \quad (4)$$

where k is the total number of terms in the double sum. In other words, $S(A, B)$ is the mean absolute difference in contact values for all intra- and interblock interactions. Customarily we take $S(A, B) < 0.1$ to mean that A and B are very similar conformations.

Using this similarity criterion, suppose we have a set of dissimilar conformations for each of a variety of chain lengths such that each protein P in some training set has a conformation P_{nat} similar to its PDB structure and several dissimilar conformations P_{non}. Adjusting the energy parameters amounts to solving the quadratic program: minimize $\|\mathbf{a}\|^2$ subject to $E(P_{\text{nat}}) + 1 \leq E(P_{\text{non}})$ for all nonnative conformations for all proteins in the training set. The

intent is to favor the native conformations of the proteins over their respective nonnative conformations by an arbitrary margin of one energy unit while keeping the magnitudes of the parameters as low as possible so that the natives are not favored by some fortuitous cancellation of large and opposing effects. The margin is truly arbitrary in that if $\boldsymbol{a}$ satisfies all the linear inequalities for a unit margin, then $k\boldsymbol{a}$ satisfies the inequalities for margin $k \geq 0$. A nonzero margin is needed to avoid solutions where $E(P_{\text{nat}}) \approx E(P_{\text{non}})$, that is, the energy function cannot make an unambiguous decision as to whether the native or some nonnative conformation of protein P is preferable. In the end, the differences in energy between native and nonnative structures were generally much greater than the margin, and the absolute values of the energies were typically at least on the order of 10.

In general, there may be no solution to a quadratic program[32] if the linear inequalities are mutually inconsistent, which is the same as saying there is no "feasible region." If a feasible region exists, it is some kind of (possibly unbounded) polyhedron in the space of the 210 adjustable energy parameters. If the objective function is positive definite, as it is in this case, then there is a unique solution to the quadratic program somewhere in the feasible region, and at the solution most of the inequalities will be "slack," that is, they are satisfied by more than the required margin. Anywhere between zero and 210 of the inequalities may be "active" at the solution. These active constraints in our problem correspond to having $E(P_{\text{nat}}) + 1 = E(P_{\text{non}})$ for some nonnative conformations of some of the training proteins. Exactly the same solution can be reached by deleting all the slack inequalities and using only the active constraints.

There are certainly specialized algorithms for efficiently solving quadratic programs, but in this case it was more expedient to convert it to a local nonlinear optimization of a function F that is the weighted sum of the objective function above plus quadratic penalty terms enforcing the inequalities ($v > 0$ being the weighting factor).

$$F(\mathbf{a}) = \|\mathbf{a}\|^2 + v \sum_{P} \sum_{\text{non}} (max\,[0, E(P_{\text{nat}}) + 1 - E(P_{\text{non}})])^2 \quad (5)$$

Then if the problem is infeasible, there will be violated inequalities at the minimum. Otherwise, there will be a unique local minimum corresponding to the solution of the quadratic program. Indeed, the same feasible optimal solution was always found from different random starting points. All this was programmed in MOE[31] in the SVL language.

Protein Structures

The advantage of dealing with cluster distance matrices is that there is a clear connection between two blocks being in contact and what types of residues are in contact, so that the energy function in Equation (3) is a simple function of the D_{IJ} elements and the sequence. Steric exclusion and basic chain connectivity translate into simple constraints[2] on the D_{IJ}, so that steric repulsion and virtual bond stretching terms are not required in the energy function. The disadvantage is that one can easily propose distance matrices corresponding to no three-dimensional structure whatsoever, or to structures incompatible with normal packing of protein secondary structural elements. Here, we restrict attention to matrices of interblock contacts, made up of elements $c(D_{IJ})$, that are calculated from sets of experimentally determined protein structures. For each chain length we retain only those matrices that correspond to some piece of a real structure and that otherwise differ from one another sufficiently, as measured by Equation (4) using the similarity cutoff $S(A, B) < 0.1$. In other words, these sets of contact matrices amount to the decoy sets used in other studies, except by the similarity cutoff used, each contact matrix covers a region of conformation space that can be rather large if it involves few contacts.

Of course, the coverage of conformation space depends on the set of protein structures surveyed to produce the set of significantly different contact matrices. The chain set described above consists of 1860 different structures assumed by real polypeptide chains, but they are taken out of context in the sense that they may be stabilized by substantial interactions with other polypeptide chains, polynucleotide chains, large prosthetic groups, and so on. A second set, called the "monomer set," was devised so as to include proteins that apparently fold as single polypeptide chains due to aqueous solvation and intrachain interactions without significant interactions with other peptide or nonpeptide moieties (see Table I). PDB has over 27,600 entries, but many of them have the same protein with different ligands or closely related proteins, etc. They list a subset where no two proteins have more than 90% sequence identity. Then PDB also lists a special database of biological unit files, explicitly describing the quaternary structure (monomer, dimer, etc.). From these we find 724 PDB entries having a single chain in the biological unit consisting of no more than 128 residues and having no nonpeptide atoms other than water molecules or sodium ions. Out of these, we find 262 proteins having at least 50 residues, coordinates for each C^{α} atom, no disulfide crosslinks, and are reasonably compact in that the radius of gyration is no more than 30% greater than the minimum value for that chain length.[6] In the case of NMR structures, simply the first model in the PDB entry is used. This "monomer set" is a rather comprehensive list of proteins for which the native state should be stabilized by effects included in the current model. These cover a variety of general fold types: 46 all α, 41 all β, 9 α/β, 57 $\alpha + \beta$, 4 small proteins, 4 peptides, 2 designed proteins, and 99 for which no SCOP classification[33] is yet available.

Covering sets of contact matrices are generated by first adding any contact matrix from the monomer set that differs from those already present for that chain length. Depending on the block size used, the 262 proteins in the monomer set are covered by 257 to 259 contact matrices, so only a few protein pairs have equivalent contact matrices even at the level of four or eight residue blocks. Next, novel contact matrices are added from the 1860 members of the chain set, which includes many less compact structures, typically resulting in a total of 1750 contact matrices of all

the various chain lengths observed. Hence, the chain set also has little redundancy in conformations. Finally, many more novel but realistic contact matrices can be generated from these matrices by deleting a block at the end of a chain. For a given chain length this continues until the number of distinct contact matrices exceeds a preset limit, such as 50 or 100. For short chains, fewer than the limit may be found because essentially all possible distinct contact matrices have been generated. For the longest chains, the limit may not be reached because there are fewer contact matrices available in the surveys of the chain and monomer sets. Generally the limit is reached for a wide range of intermediate lengths.

RESULTS

Key Parameters

The motivation behind this work is that viewing protein structure at low resolution, i.e., large b, greatly simplifies the dimensionality and combinatorial complexity of the protein conformation space. However, this must be balanced against the need to make meaningful distinctions between distinct folds. Surveying a set of 32 small, monomeric protein crystal structures having different folds (listed in ref. 3, Table I), we compared each protein chain of n residues with the first n residues of all the longer chains. For $b = 15$ there is an offdiagonal cutoff C_{IJ} such that all such pairs of structures produce distinguishable contact/noncontact matrices. For larger b there is no such cutoff. Hence, in Table II that is the largest value of b listed. Not surprisingly, smaller values discriminate better between native and nonnative conformations, so here we have tested $b = 8$ and 4, although other values may given even better results.

When adjusting the energy parameters, a key factor is the criterion for distinguishing native and nonnative contact matrices in the quadratic program. In Equation (4) we have used $S(A, B) < 0.1$ as the test for conformations A and B being essentially the same. Larger values of the cutoff, such as 0.15, tend to reduce the number of different contact matrices to be considered, but the resulting energy parameters have less predictive power (results not shown). The other parameter for distinguishability is the relative width w in Equation (2). A discontinuous contact function $c(D_{IJ})$ with $w = 0$ makes it very difficult to solve the quadratic program for all but trivial training sets. On the other hand, a very broad transition with $w = 0.9$ also makes it impossible to solve the quadratic program. In terms of predictive power, the best results are obtained when $w = 0.2$, and this value is used in all that follows.

Training and Prediction

Neither for block size 4 nor 8 is it possible to find a feasible set of energy parameters that satisfies all members of the monomer set. After minimizing F in Equation (5), some proteins had their respective native energies less than all nonnative energies by the required margin, while other proteins had violations of such inequalities. Successively eliminating the protein having an inequality violated by the greatest margin and readjusting the energy parameters, it is eventually possible to find a subset consisting of 211 proteins to use as a training set (see Table I), and all these can be satisfied simultaneously. With regard to sequence, the training set has little redundancy. By the construction of the whole monomer set, no pair of proteins has 90% or greater sequence identity after optimal alignment. Then in an alignment of all pairs of proteins in the training set, sequence identities range from 2.8 to 89.5%, and 99% of the pairs have less than 30% sequence identity, 89% of the pairs have less than 20% sequence identity, and all but the 20 most similar pairs have less than 47% sequence identity. It is not at all obvious what distinguishes the 211 proteins in the training set from the 262 − 211 = 51 problematic ones (see Table I). The monomer set as a whole consists largely of structures determined by NMR, often associated with DNA binding *in vivo*, but inspection of the problematic PDB entries and associated literature references reveals no evidence for substantial associated ligands, although often the ends of the chain are rather extended from the main protein globule.

TABLE III. Training and Prediction Results

b	No. structs	No. actives	Predicted monomers	Predicted chains
8	50	69	113	407
4	50	31	174	698
4	100	56	141	556

Table III summarizes the predictive power of these sorts of models. For a given block size, a set of contact matrices was found that are all different according to Equation (4), but every native conformation in the monomer and chain sets is included by the same criterion. In addition, truncated chains were also included so as to bring up the total number of alternative contact matrices for each chain length to either 50 or 100, as indicated in the table under number of structures. When the training set is used to adjust the energy parameters, most of the proteins contribute no active constraint at the solution to the quadratic program. The number of proteins that do is given in the table as the number of active proteins. Typically, the smaller the number of actives, the greater the predictive power. For example, for $b = 8$ and a maximum of 50 different contact matrices for each chain length n, only 69 of the 211 training set proteins were active at the solution. By the nature of quadratic programming, the same energy parameters would have been obtained if only these 69 active proteins had been used, instead of the full training set. A generous assessment of this outcome is to view the 262 −69 = 193 other members of the monomer set as a test set, remembering that the 211 member training set is a subset of the monomer set (see Table I). Then the correct native structure was recognized for 211 − 69 =142 proteins in the monomer set, which corresponds to a 142/193 = 75% success rate. A stricter interpretation is to discount any predictions where the sequence identity of the predicted protein is greater than 30% after optimal sequence alignment with any member of the active set. In

this case, 213 out of 262 monomer proteins were correctly predicted, but only 113 of these had low sequence identity with the active set, and this is listed in the table under predicted monomers. A more rigorous assessment is to view the chain set as a completely separate test set and exclude members of the chain set having sequence similarity to the active proteins. Thus, for the 1860 members of the chain set, 487 were correctly predicted, and 407 of those had less than 30% sequence identity to all members of the active set, as shown in the table under predicted chains.

Increasing the resolution to $b = 4$ dramatically decreases the size of the active set required to determine the energy parameters, and the numbers of sequentially dissimilar proteins that were correctly predicted rose for both the monomer and chain sets. The gross score is 226 out of the 262 monomer set were correctly predicted, of which 174 were sequentially dissimilar to all the actives, but that still leaves 36 members of the monomer set that are not in agreement with this model for reasons that have yet to be determined. On the other hand, there are 698 correct predictions of proteins in the chain set that are sequentially dissimilar to the active set, even though these chains may involve disulfide bridges or be associated with other polypeptide chains or large ligands that may play an important role in stabilizing the observed conformation. Of the 751 correctly predicted proteins in the chain set (without regard to sequence similarity to the active set), 132 had at least one disulfide bridge, several had seven, and one (1le6) had eight. Also, many of these correctly predicted chains are closely associated with other chains. For example, 1gmj.a is one of four largely helical chains forming a heterodimer of heterodimers in the crystal structure, each chain having a radius of gyration far greater than the minimum value for its chain length. It so happens that this conformation is viewed as the most stable of 50 alternatives for that sequence when one disregards its closely associated neighbors.

Raising the maximum number of alternative contact matrices to 100 for $b = 4$ presents a more challenging problem. Now there are 100 alternative contact matrices for chain lengths 10, 11, and 13–112, and the least coverage is 45, 43, 20, 14, 22, and 19 contact matrices for chain lengths 123–128, respectively. Parameters can still be determined (Table IV) on the basis of 56 active proteins, and only 42 of the 262 proteins in the monomer set are incorrectly predicted. Of the correct predictions, 141 have little sequence identity with the active set. Similarly for the 1860 members of the chain set, 648 are correctly predicted, and of these, 556 have little sequence identity with the active set (Table III). Although these interaction parameters work well, their values correspond only roughly with conventional wisdom about protein folding, in that hydrophobic-hydrophobic residue type pairs tend to have favorable negative values, and hydrophilic-hydrophilic interactions tend to have unfavorable positive values. The most unfavorable (positive) interaction in Table IV is between Gly and Met residues, while the most favorable is between Cys and Ala. It is not obvious this is due to some

TABLE IV. Residue–Residue Interaction Parameters Using $b = 4$ and at Most 100 Alternative Contact Matrices

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	0.20																			
S	0.33	−0.07																		
T	−0.76	−0.01	−0.01																	
P	0.17	0.09	0.10	−0.34																
A	−0.78	0.10	−0.08	−0.07	−0.09															
G	−0.13	0.09	0.01	−0.06	0.14	0.18														
N	0.34	−0.03	0.27	−0.05	0.01	0.04	−0.14													
D	−0.15	0.15	0.43	0.12	0.05	−0.15	0.11	0.08												
E	−0.04	−0.05	0.23	0.27	0.13	0.08	0.12	0.07	0.24											
Q	−0.28	0.04	0.09	−0.25	−0.11	0.12	−0.01	0.05	−0.15	−0.10										
H	0.08	0.14	−0.25	−0.01	−0.09	−0.01	−0.00	0.07	0.19	0.10	−0.05									
R	0.02	−0.03	0.02	0.10	−0.03	0.10	0.08	0.07	0.06	0.01	−0.06	0.08								
K	0.31	0.09	0.23	−0.01	−0.03	−0.16	0.09	0.02	−0.07	0.04	0.06	0.21	0.05							
M	−0.22	−0.36	0.13	0.26	−0.16	0.46	−0.39	0.23	0.24	0.36	−0.08	−0.14	−0.24	−0.35						
I	−0.06	0.22	−0.49	0.01	−0.15	−0.13	−0.17	0.12	−0.16	0.09	0.24	−0.12	0.21	−0.24	−0.31					
L	0.03	−0.13	−0.28	0.20	0.02	0.12	0.18	−0.23	−0.11	0.02	−0.11	−0.16	0.00	0.02	−0.27	−0.15				
V	0.02	0.01	−0.22	0.27	−0.11	0.18	−0.26	0.11	−0.11	0.07	0.12	−0.08	0.09	−0.00	−0.40	−0.09	−0.42			
F	0.04	0.08	0.08	0.05	0.05	0.11	0.00	−0.21	−0.01	0.09	−0.19	−0.04	−0.26	−0.15	−0.15	−0.25	−0.17	−0.24		
Y	−0.31	0.11	−0.01	−0.13	0.08	−0.03	0.24	−0.17	−0.20	−0.19	−0.02	0.05	−0.07	−0.64	−0.14	−0.13	−0.01	0.07	0.08	
W	−0.45	−0.08	−0.06	−0.11	−0.59	−0.01	−0.10	−0.16	−0.31	0.16	−0.14	0.17	−0.01	−0.40	0.20	−0.27	−0.44	−0.13	−0.05	−0.32

sort of overfitting, because these residue types are not that rare in the training set.

One might ask whether choosing for a given sequence the correct contact matrix out of 100 alternatives constitutes an accurate prediction of the tertiary structure. First note that these sets of contact matrices are not guaranteed to cover all possible energetically reasonable polypeptide conformations in three dimensions. They certainly include the conformations of a couple thousand chains in PDB, but further work will be required to establish a truly exhaustive set. As it stands, it is possible that the correct contact matrix is missing for some proteins. Second, consider a protein having a disordered lengthy end of the chain that is not in contact with the compact globular remainder. Two such conformations may agree well in the folded part while differing substantially in how the chain end extends out into the solution. By the similarity criterion used here, the contact matrices of the two conformations may be identical, yet the RMSD for the superposition of the entire chain may be large. Thus, it is possible to have very similar contact matrices but differ substantially in overall conformation in an energetically trivial way. Third, it is possible to make rather fine distinctions in conformation by contact matrices, such that choosing the right one actually amounts to a good prediction. For example, 1adr is a compact, all-α protein in the monomer set having 76 residues. Comparing with all contiguous 76 residue chain segments from the other 261 proteins in the monomer set, there are none having a similar contact matrix at the level of four residue blocks. However, residues 11–86 of protein 1dgn have a conformation that is visually clearly similar to 1adr, namely $\rho = 0.60$ in the universal scale[34] of 0 to 2, or equivalently 6.5 Å in RMSD. Yet because the two contact matrices are clearly different, this method would have to choose the native structure of 1adr over that of this piece of 1dgn.

CONCLUSIONS

It is possible to describe the conformations of polypeptide chains in terms of distances between several-residue segments such that a simple linear function of interresidue contacts can discriminate between the native conformation and numerous nonnative sets of contacts for many different proteins. Generally, the lower resolution models imply a more thorough coverage of conformation space, but the cruder approximation to interresidue interactions restricts the predictive power of the model. So far, attention has been restricted to fairly small, soluble, globular proteins that are stabilized by intrachain noncovalent interactions, rather than associations with other polypeptide chains, polynucleotides, or other large ligands. While this is a promising step, there remain two shortcomings. First, some proteins are not in agreement with the model for reasons that are not yet clear. Second, the sets of contacts considered for a given chain length are derived from experimental protein structures and thus embody general information about polypeptide flexibility, secondary structures, and overall packing, but the set does not necessarily cover all possible conformations. Thus, one cannot yet use this approach to develop a quantitative statistical mechanical model of protein folding.

ACKNOWLEDGMENTS

The author thanks Prof. Daniel Burns for helpful conversations and Yu Chen for searching PDB for monomeric proteins.

REFERENCES

1. Crippen GM, Havel TF. Distance geometry and molecular conformation. New York: Wiley; 1988.
2. Crippen GM. Cluster distance geometry of polypeptide chains. J Comput Chem 2004;25:1305–1312.
3. Crippen GM. Statistical mechanics of protein folding by cluster distance geometry. Biopolymers 2004;75:278–289.
4. Miyazawa S, Jernigan RL. Estimation of effective contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
5. Goldstein RA, Luthey-Schulten Z, Wolynes PG. Protein tertiary structure recognition using optimized hamiltonians with local interactions. Proc Natl Acad Sci USA 1992;89:9029–9033.
6. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
7. Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J Comput-Aided Mol Design 1993;7:473–501.
8. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682.
9. Kocher J-PA, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 1994;235:1598–1613.
10. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.
11. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264:1164–1179.
12. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. J Mol Biol 1996;256:623–644.
13. Park B, Levitt M. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258:367–392.
14. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
15. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci 1997;6:676–688.
16. Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:1101–1108.
17. Samudrala R, Moult J. An all-atom distance dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
18. Clementi C, Vendruscolo M, Maritan A, Domany E. Folding Lennard-Jones proteins by a contact potential. Proteins 1999;37:544–553.
19. Tobi D, Elber R. Distance-dependent pair potential for protein folding: results from linear optimization. Proteins 2000;41:40–46.
20. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins 2000;40:71–85.
21. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins 2000;38:134–148.
22. Dombkowski AA, Crippen GM. Disulfide recognition in an optimized threading potential. Protein Eng 2000;13:679–689.
23. Ohkubo YZ, Crippen GM. Potential energy function for continuous state model of globular proteins. J Comput Biol 2000;7:363–379.

24. Crippen GM. Constructing smooth potential functions for protein folding. J Mol Graph Mod 2001;19:87–93.
25. Wang Y, Zhang H, Li W, Scott RA. Discriminating compact nonnative structures from the native structure of globular proteins. Proc Natl Acad Sci USA 1995;92:709–713.
26. Huang ES, Subbiah S, Tsai J, Levitt M. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. J Mol Biol 1996;257:716–725.
27. Crippen GM, Ohkubo YZ. Statistical mechanics of protein folding by exhaustive enumeration. Proteins 1998;32:425–437.
28. Micheletti C, Seno F, Banavar JR, Maritan A. Learning effective amino acid interactions through interactive stochastic techniques. Proteins 2001;42:422–431.
29. Chhajer M, Crippen GM. Toward correct protein folding potentials. J Biol Phys 2004;30:171–185.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
31. Molecular Operating Environment (MOE). Version 2004.03. Chemical Computing Group, Inc., http://www.chemcomp.com.
32. Murty K. Linear and combinatorial programming. New York: Wiley; 1976. p 486–493.
33. Lo Conte L, Brenner SE, Hubbard TJP, Chothia C, Murzin A. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acid Res 2002;30:264–267.
34. Maiorov VN, Crippen GM. Size-independent comparison of protein three-dimensional structures. Proteins 1995;22:273–283.