Probabilistic approach to the design of symmetric protein quaternary structures

Xiaoran Fu¹, Hidetoshi Kono² and Jeffery G. Saven¹,³

¹Makineni Theoretical Laboratories, Department of Chemistry, University of Pennsylvania, 231 South 34th Street, Philadelphia, PA 19104, USA and
²Neutron Science Center and Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, 8-1, Umemidai, Kizu-cho, Souraku-gun, Kyoto 619-0215, Japan
³To whom correspondence should be addressed.
E-mail: saven@sas.upenn.edu

Probabilistic methods have been developed that estimate the site-specific probabilities of the amino acids in sequences likely to fold to a particular target structure, and such information can be used to guide the de novo design of proteins and to probe sequence variability. An extension of these methods for the design of symmetric homo-oligomeric quaternary structures is presented. The theory is in excellent agreement with the results of studies on exactly solvable lattice models. Application to an atomically detailed representation of proteins verifies the utility of a symmetry assumption, which greatly simplifies and accelerates the calculations. The method can be applied to a wide variety of symmetric and periodic protein structures.

Keywords: computational protein design/protein oligomers/p53 tetramerization domain/quaternary structure

Introduction

De novo protein design identifies amino acid sequences with predetermined folded structures and functions. Such a design probes the determinants of protein stability and function and may potentially lead to novel enzymes, therapeutics and biomaterials. The design of entire sequences is complicated by the size of proteins, their many conformational degrees of freedom, the subtlety of the interactions that stabilize the folded state, and the exponentially large number of possible sequences (Saven, 2001). Despite this complexity, computationally aided design has led to successful realization of a variety of proteins (DeGrado et al., 1999; Kraemer-Pecore et al., 2001) and the redesign of natural proteins to confer novel functionalities (Bryson et al., 1995; Hellinga, 1999; Looger et al., 2003). Herein, we present an efficient methodology for estimating the site-specific amino acid probabilities of oligomeric protein structures, where the target structure for design comprises both quaternary and tertiary structure.

Most efforts in computational design apply ‘directed’ methods. Here ‘directed protein design’ refers to the search for a sequence (or a small set of sequences) consistent with a predetermined backbone structure. The search is directed toward those sequences having low energy (or a favorable score) when they assume the target structure. Pioneering efforts in design identified proteins having substantial ordering but not necessarily well defined tertiary structures (Bryson et al., 1995). Since exponentially large numbers of sequences are possible even for small proteins, computational methods have accelerated successful design. Exhaustive searching of all \( m^N \) possible sequences is usually intractable, where \( N \) is the number of variable residues and \( m \) is the number of residue degrees of freedom, e.g. the number of allowed amino acids and side chain conformations. Genetic algorithms (Jones, 1994) and simulated annealing methods (Shakhnovich and Gutin, 1993; Hellinga and Richards, 1994) search sequence space in a partially random fashion, on average progressing toward lower energy sequences. Such stochastic methods have been used to redesign a variety of natural proteins (Desjarlais and Handel, 1995, 1999; Desjarlais and Clarke, 1998; Johnson et al., 1999; Kraemer-Pecore et al., 2001) and novel helical proteins (Bryson et al., 1998). Elimination methods such as ‘dead end elimination’ provide better estimates of global optima and have also been used to automate both the redesign of sub-regions within natural proteins (Malakauskas and Mayo, 1998; Strop and Mayo, 1999; Bolon and Mayo, 2001; Looger et al., 2003), as well as full sequence design (Dahiya and Mayo, 1997). Directed approaches are sensitive to the energy functions used, which may be problematic given that such energy functions are necessarily approximate. Uncertainties in the energy function may not merit the search for global optima. For cases where information about a problem is incomplete, probabilistic rather than directed approaches to design may be appropriate.

Statistical methods complementary to directed protein design have been developed (Zou and Saven, 2000; Kono and Saven, 2001). Such methods reveal the features of sequences likely to fold to a particular structure but which may not be thermodynamically ‘optimal’. Such probabilistic methods provide site-specific information about the range of allowed amino acid substitutions. Within this methodology, much of the formalism of statistical thermodynamics is recast so as to reveal the properties of sequences likely to fold to a target three-dimensional structure. The site-specific probabilities of the amino acids are determined by maximizing an effective entropy function, subject to constraints on the sequences. Such constraints can be physically based, such as the energy of sequences after they acquire the target backbone structure, or functionally based, such as the patterning of amino acids at predetermined positions to confer metal binding. The theory takes as input (i) a given target structure, (ii) energy functions for quantifying sequence–structure compatibility and (iii) a set of constraints on the sequences. For some forms of the constraints, the approach reduces to a form of heterogeneous mean field theory (Koehl and Delarue, 1996). The theory yields estimates of the number of sequences and, most importantly, the site-specific probabilities of the amino acids and their side chain conformations. The computation time of the calculations scale is \( (Nm)^2 \). With statistical methods, in a much shorter time larger numbers \( N \) of variable residues can be examined using a
larger diversity of residue states \( m \) than with other computational methods. Sequences are not explicitly sampled; the calculations yield the amino acid probabilities directly, which are useful in protein design for identifying allowed amino acids at each position in a target structure. The probabilities may also be used both to identify sites tolerant of mutations, where functional properties may be engineered, and to guide the construction of a combinatorial library of protein sequences.

Most computational efforts in protein design have focused on the design of tertiary structures, but some of the earliest design targets were oligomeric structures. The DeGrado group has crafted a variety of dimeric, trimeric and tetrameric helical bundles (Bryson et al., 1995; DeGrado et al., 1999). Directed computational methods have been used in the design of coiled coils, where the symmetry of the structure facilitates the design (Harbury et al., 1998; Keating et al., 2001). Just as in nature, quaternary structure provides a facile route to large, well structured protein systems. It is of interest to extend the probabilistic methods for designing protein and protein libraries to include such an intermolecular structure, which is presented herein. For symmetric arrangements of protomers, a symmetry assumption greatly reduces the computational overhead so that arbitrary oligomers and multidimensional arrays may be included as design targets.

Theory

In this section, we briefly review statistical theory that yields estimates of the site-specific amino acid probabilities for a given backbone structure. Molecular energy functions are a key component of the formalism for quantifying sequence–structure compatibility, and methods for incorporating inter-and intra-molecular energies into the statistical formalism are presented. For polymeric or oligomeric systems, i.e. homomorphic systems with quaternary structure, a simplifying symmetry assumption greatly improves the efficiency of the calculations.

Overview of statistical theory

A statistical, entropy-based formalism has been developed to identify the features of sequences that are likely to fold to a given backbone structure (Saven and Wolynes, 1997; Zou and Saven, 2000; Kono and Saven, 2001). Generally, the method takes as input a given template backbone structure and a function for characterizing sequence–structure compatibility. The calculation yields the site-specific probabilities of the amino acids. Since the method does not involve the explicit generation of sequences, information about exponentially large numbers of sequences may be obtained. This is particularly useful given the large numbers of possible protein sequences \( 20^N \) for an \( N \)-residue protein) and the wide natural diversity of sequences sharing common folds. The generality of the method allows a number of requirements to be prescribed upon sequences as constraints. Such constraints can include both the patterning of residues, e.g. so as to confer solubility, or global energetic constraints, so as to identify the properties of sequences likely to have a particular target structure as their folded state.

The fundamentals of the method have been presented in previous work (Saven and Wolynes, 1997; Zou and Saven, 2000; Kono and Saven, 2001), and are briefly reviewed here. The most probable set of site-specific amino acid probabilities at each position is determined by maximizing an effective entropy function subject to imposed constraints using the method of Lagrange multipliers (McQuarrie, 1976), thus specifying a variational functional \( V \):

\[
V = S - \lambda_1 f_1 - \lambda_2 f_2 - \cdots
\]

(1)

The sequence entropy \( S \) quantifies the number of sequences likely to fold to a particular structure. The \( f_k \) are functions that impose the constraint conditions and the \( \lambda_k \) are their corresponding Lagrange multipliers. The variational function \( V \) and the constraint function \( f_k \) depend upon the site-specific probabilities \( w_i(\alpha, r(\alpha)) \). Each residue position (site) in the structure is labeled by the index \( i (i = 1, ..., N) \), where \( N \) is the total number of residues. Here \( w_i(\alpha, r(\alpha)) \) is the probability that residue position \( i \) is in a particular ‘state’, where the state of the residue is specified by both the identity of the amino acid \( \alpha \) and the conformation of its side chain \( r(\alpha) \). Generally, \( \alpha \) labels any one of the possible amino acids (natural or non-natural), and \( r(\alpha) \) labels the members of a discrete set of side chain conformations for each amino acid, so-called rotamer states (Dunbrack, 2002). Here \( S \) is written solely in terms of the one-body probabilities \( w_i(\alpha, r(\alpha)) \):

\[
S = - \sum_i \sum_{\alpha, r(\alpha)} w_i(\alpha, r(\alpha)) \ln w_i(\alpha, r(\alpha))
\]

(2)

In identifying the state probabilities consistent with particular values of the constraints, the \( k \)-th constraint function \( f_k \) is constrained to have a particular value \( f_k^\alpha \):

\[
f_k^\alpha = f_k(w_i(\alpha, r(\alpha)))
\]

(3)

For example, such constraints may be used to specify the normalization of the probabilities, to pattern residue types, or to specify the energy sequences acquire when in the folded structure. As a result, the properties of sequences sharing these common constraints can be examined. The site-specific probabilities \( w_i(\alpha, r(\alpha)) \) are determined as those that maximize \( V \) subject to given values of \( f_1^\alpha, f_2^\alpha, \ldots \) (see Equation 3). It is important to note that although the form of \( S \) in Equation 2 would seem to imply that the \( w_i(\alpha, r(\alpha)) \) are independent, the constraint conditions in Equation 3 will cause them to be coupled to one another. The probability of an amino acid at a particular position may be obtained by summing over the probabilities of its rotamers: \( w_i(\alpha) = \sum_{r(\alpha)} w_i(\alpha, r(\alpha)) \). This formalism may be applied to different representations of protein structure. In many cases, it is useful to simplify the representation of each residue so as to eliminate explicit side chain degrees of freedom, thus yielding effective united-residue representations and energy functions often used in protein science (Miyazawa and Jernigan, 1996; Liwo et al., 1998). For such representations, each amino acid has effectively only one conformational state: \( w_i(\alpha) = \sum_{r(\alpha)} w_i(\alpha, r(\alpha)) = w_i(\alpha, r(\alpha)) \), and for such cases we may omit the variable indicating the rotamer state.

Constraints and protein sequence energetics

The theory may accommodate a wide variety of constraints of the form in Equation 3. Each residue site \( i \) must be occupied, i.e. the probabilities \( w_i(\alpha, r(\alpha)) \) are normalized, which leads to:
\[
\sum_{\alpha} \sum_{r(\alpha)} w_i(\alpha, r(\alpha)) = 1 \tag{4}
\]

The hydrophobic effect may be included as an effective ‘one-residue’ energy that is dependent upon the local density of β-carbons (Kono and Saven, 2001). The sum of these solvation or environment scores may also be constrained by specifying the average ‘environmental energy’ as summed over the positions in a particular target structure:

\[
E_{env} = \sum_i \sum_{\alpha, r(\alpha)} e_i^m(\alpha, r(\alpha)) w_i(\alpha, r(\alpha)) \tag{5}
\]

Other multi-body representations of the hydrophobic effect and effective pair interactions between residues may also be incorporated in the theory (Miyazawa and Jernigan, 1985).

For realistic, atom-based representations of proteins, design algorithms that focus primarily on optimizing inter-atomic interactions within the folded state have had substantial success (Hellinga and Richards, 1994; Dahiyat et al., 1997; Kortemme et al., 1998; Baker and DeGrado, 1999; Hellinga, 1999; Strop and Mayo, 1999; Bolon and Mayo, 2001; Looger et al., 2003).

Such methods use atom based potentials to account for both covalent and non-covalent interactions, e.g. van der Waals forces, hydrogen bonds, and electrostatic interactions. The statistical theory of sequence ensembles may be formulated so as to include such atom-based descriptions of both intra- and inter-molecular interactions.

For a protomer structure in an oligomeric complex, the intramolecular energy \( E_i \) of a single chain depends upon the set of amino acid identities \( \{\alpha_1, ..., \alpha_N\} \) and the rotameric state of each of these amino acids \( r(\alpha) \). \( E_i \) may be written as a sum of effective one-residue and two-residue interactions:

\[
E_i = \sum_i \gamma_i^{(1)}(\alpha_i, r(\alpha_i)) + \sum_{i<j} \gamma_{ij}^{(2)}(\alpha_i, r(\alpha_i); \alpha_j, r(\alpha_j)) \tag{6}
\]

The indices \( i \) and \( j \) refer to residue positions in the structure, and the second term sums only over unique interactions between pairs of residues. The ‘one-residue’ energy \( \gamma_i^{(1)}(\alpha, r(\alpha)) \) is the energy associated with locating the amino acid \( \alpha \) with conformation \( r(\alpha) \) at site \( i \) within a single protomer structure. This energy is determined by side-chain backbone interactions or intrinsic structural tendencies of the amino acids, such as preferences for solvent exposure or secondary structure. Similarly, the pair energy \( \gamma_{ij}^{(2)}(\alpha_i, r(\alpha_i); \alpha_j, r(\alpha_j)) \) is the sequence dependent interaction energy between residue \( i \) and residue \( j \), each of which is a member of the same protein chain. Such interaction energies may be inferred from a database using reduced descriptions of amino acids (Miyazawa and Jernigan, 1996; Onuchic et al., 1997) or, using an atomically detailed model, as a sum over inter-atomic interactions of the identity-rotamer states at sites \( i \) and \( j \), as specified by \( (\alpha_i, r(\alpha_i)) \) and \( (\alpha_j, r(\alpha_j)) \).

Similarly, the intermolecular association energy \( E_a \) is that due to the intermolecular interactions among sites in a complex of \( M \) protein chains, where chains are labeled by the indices \( m \) and \( n \). The association energy of the complex \( E_a \) may be written in terms of the identity-rotamer states of each site in the complex:

\[
E_a = \sum_{m,i} \epsilon_i^{(1)}(\alpha_i^m, r(\alpha_i^m)) + \sum_{m<n,i,j} \epsilon_{ij}^{(2)}(\alpha_i^m, r(\alpha_i^m); \alpha_j^n, r(\alpha_j^n)) \tag{7}
\]

where \( \epsilon_i^{(1)}(\alpha_i^m, r(\alpha_i^m)) \) is the additional effective energy associated with locating the amino acid \( \alpha_i^m \) with conformation \( r(\alpha_i^m) \) at site \( i \) on chain \( m \) in the complex. For example, a residue that is exposed to solvent in the folded structure of an isolated protomer may become sequestered from solvent or interact with the backbone of an adjacent protomer upon complex formation; \( \epsilon_i^{(1)}(\alpha_i^m, r(\alpha_i^m)) \) quantifies the effective change in energy upon association. For given values of the identities and rotamers at sites \( i \) and \( j \) on two different protomers, the pair interaction \( \epsilon_{ij}^{(2)}(\alpha_i^m, r(\alpha_i^m); \alpha_j^n, r(\alpha_j^n)) \) is the sequence-dependent intermolecular interaction energy between residue \( i \) on chain \( m \) and residue \( j \) on chain \( n \).

Equation 7 may be applied to any given complex, including hetero-dimers and other proteins having asymmetric quaternary structure. For symmetric complexes comprising identical chains, however, symmetry implies that equivalent positions on each chain have the same identity and conformational state. This symmetry assumption may be expressed as follows:

1. The amino acid identities at equivalent sites on each chain are the same: \( \alpha_i^m = \alpha_i \) for each protomer \( m = 1, ..., M \).
2. Side chains at equivalent sites on different chains take on equivalent rotamer states: \( r(\alpha_i^m) = r(\alpha_i) \) for each \( m = 1, ..., M \).

This assumption is introduced to expedite the design process, though it is perhaps overly stringent, particularly with regard to side chain conformation. For example, distant, exposed residues of a protein need not always have precisely the same conformation in solution. For systems with only one state per site, e.g. the lattice model discussed in the Applications and results (section ‘Lattice model’), this symmetry assumption is exact.

Within this symmetry assumption, Equation 7 may be written in a form analogous to Equation 6, involving effective interactions between sites on a single chain:

\[
E_a = \sum_i \tilde{\epsilon}_i^{(1)}(\alpha_i, r(\alpha_i)) + \sum_{i<j} \tilde{\epsilon}_{ij}^{(2)}(\alpha_i, r(\alpha_i); \alpha_j, r(\alpha_j)) \tag{8}
\]

where

\[
\tilde{\epsilon}_i^{(1)}(\alpha_i, r(\alpha_i)) = \sum_{m=1}^M \epsilon_i^{(1)}(\alpha_i^m, r(\alpha_i^m)) \tag{9}
\]

\[
\tilde{\epsilon}_{ij}^{(2)}(\alpha_i, r(\alpha_i); \alpha_j, r(\alpha_j)) = \sum_{m<n} \epsilon_{ij}^{(2)}(\alpha_i^m, r(\alpha_i^m); \alpha_j^n, r(\alpha_j^n)) \tag{10}
\]

The effective interactions include contributions from both the one-residue energy at each site in the chain (\( \tilde{\epsilon}_i^{(1)}(\alpha_i, r(\alpha_i)) \)) and the pair energy \( \tilde{\epsilon}_{ij}^{(2)}(\alpha_i, r(\alpha_i); \alpha_j, r(\alpha_j)) \) between sites in the chain.
We note that $E_a$ now has a form exactly analogous to $E_t$. Each residue ‘sees’ other residues on the same chain and ‘images’ of.

For a given set of constraint conditions, the values of $E_t$ and $E_a$ are assumed to be well represented by their sequence averages (Zou and Saven, 2000; Kono and Saven, 2001). With this approximation and the factorization implicit in Equation 2, $E_t$ and $E_a$ may be expressed as functions of the site probabilities $w_i(\alpha, r(\alpha))$:

$$E_t \approx \sum_i \sum_{\alpha, r(\alpha)} \gamma_i^{(1)}(\alpha, r(\alpha)) w_i(\alpha, r(\alpha)) + \sum_{i<j} \sum_{\alpha, r(\alpha)} \sum_{\alpha', r(\alpha')} \gamma_{ij}^{(2)}(\alpha, r(\alpha); \alpha', r(\alpha')) w_i(\alpha, r(\alpha)) w_j(\alpha', r(\alpha'))$$

$$w_i(\alpha, r(\alpha)) w_j(\alpha', r(\alpha'))$$ (12)

$$E_a \approx \sum_i \sum_{\alpha, r(\alpha)} (\epsilon_i^{(1)}(\alpha, r(\alpha)) + \epsilon_i^{(2)}(\alpha, r(\alpha))) w_i(\alpha, r(\alpha)) + \sum_{i<j} \sum_{\alpha, r(\alpha)} \sum_{\alpha', r(\alpha')} \epsilon_{ij}^{(2)}(\alpha, r(\alpha); \alpha', r(\alpha')) w_i(\alpha, r(\alpha)) w_j(\alpha', r(\alpha'))$$

Note that we sum over amino acids at each site and no longer consider just a single sequence as would be indicated by the set $\{\alpha_i\}$. The $E_t$ and $E_a$ are now each functions of the site-specific probabilities $w_i(\alpha, r(\alpha))$ and may now appear as constraints of the form suggested in Equation 3. With the symmetry assumption, the number of variables decreases dramatically relative to a calculation that treats each site $i$ in the complex separately, i.e. a calculation where $N$ is the total number of residues in the complex rather than in a single protomer. For symmetric structures, only the $w_i(\alpha, r(\alpha))$ of residues on a single chain need be determined. There are typically $N \times m$ such independent variables, where $m$ is the number of allowed identity rotamer states at each site and $N$ is the number of residues per protomer.

Other energetic constraints may be imposed on the sequences as well. For simplified representations of the amino acids, in designing protein sequences it is often important to account for non-target structures that a sequence may also acquire. For a foldable sequence, the target state should be energetically removed from these other possible collapsed structures. There are many ways to achieve this (Saven, 2001), but perhaps the simplest is to optimize a stability gap $\Delta f = E_f - \langle E(\alpha) \rangle$, where $\langle E(\alpha) \rangle$ is an average over the an appropriate ensemble of unfolded structures. Similarly, we may also define a ‘binding stability gap’ $\Delta b = E_b - \langle E(\beta) \rangle$, where $\langle E(\beta) \rangle$ represents an energy average over configurations of the complex other than the target quaternary structure. Herein, these alternate configurations involve different relative orientations of the individually folded protomers. The nature of the configurational averagings $\langle .. \rangle$ depends on the particular model and may be impractical for atomically detailed models due to both the large number of unfolded conformations and the large numbers of identity-rotamer states. For contact type energy functions, however, this averaging is straightforward (Zou and Saven, 2000), and an example is presented in the next section.

Applications and results

The statistical theory of homo-oligomeric protein complexes is applied to two systems: a lattice model and a realistic atom-based representation of a protein. The lattice model may be solved exactly via explicit enumeration and is used both to test the accuracy of the theory and to exemplify how sequence space may be resolved in terms of the folding and association energetics. An atomically detailed representation of the protein is used to test and illustrate the symmetry assumption and show how the method may be applied to realistic representations of protein structures.

Lattice model

The model consists of a 27 residue self-avoiding polymer whose monomers occupy sites on a cubic lattice (Lau and Dill, 1990; Shakhnovich and Gutin, 1990; Leopold et al., 1992). Two types of monomers, hydrophobic (H) and polar (P), are used to construct sequences. Each of these residue types has a single conformational state: $w_i(\alpha, r, (\alpha)) = w_i(\alpha)$, where $\alpha = H$ or $P$. The $2^7$ possible sequences may be exhaustively enumerated. As a result, this model may be ‘solved’ exactly, and the results may be compared with theoretical results. A simplified potential developed for HP type monomers (Li et al., 1996; Zou and Saven, 2000) is used that contains only two-body interactions, those interactions make the calculations non-trivial. The two body interactions are non-zero only if the amino acids are nearest neighbors on the lattice but are not bonded to one another. Let $r_{ij}$ be the distance between residues $i$ and $j$ and $r_0$ is the distance between nearest neighbors on the lattice. Then the contact parameter $\sigma_{ij}$ is non-zero only if $r_{ij} \leq r_0$, for which $\sigma_{ij} = 1$. The contact variable for intermolecular interactions appears as $\sigma_{ij}^{mn}$, which is unity only if residues $i$
The same energy function is used for both folding and for inter-molecular association: $\gamma^{(2)}(\alpha; \alpha') = \varepsilon^{(2)}(\alpha; \alpha')$. The contact energies are chosen as (Li et al., 1996; Zou and Saven, 2000): $\gamma^{(2)}(H_i, H_j) = -3\varepsilon$, $\gamma^{(2)}(P_i, H_j) = \gamma^{(2)}(P_i, P_j) = -\varepsilon$, and $\gamma^{(2)}(P_i, P_j) = 0$. Equating $\gamma^{(2)}(\alpha, \alpha') = \varepsilon^{(2)}(\alpha, \alpha')$ corresponds to treating inter-residue interactions in the same fashion for both intra-molecular (folding) and inter-molecular (specific oligomerization) organization. As necessary, distinct potentials for intra-molecular folding and for inter-molecular association may be used.

The energetic constraints in this lattice model study are the folding stability gap ($\Delta_f$) and binding stability gap ($\Delta_b$). For such folding criteria, ensembles of unfolded and mis-associated states are necessary. The choice of folded structure is arbitrary, and a structure that is the conformational ground state for a large number of sequences is chosen in this study (Li et al., 1996). The remaining 103,345 compact, cubic structures of the 27-mer are chosen as the ensemble of unfolded states in the calculation of $\langle E_{f,\alpha} \rangle$. The target structures of the complex are those depicted in Figure 1. As an ensemble of mis-associated states for the lattice proteins, we consider arrangements of the oligomer for which each chain takes on the target folded structure and the interface between two protomers has nine contacts, i.e. the faces of the protomers are in registry with one another (Figure 1). This loosely mimics the expectation that the mis-associated states of a particular complex most likely to compete with the target structure are those involving large numbers of residues in inter-molecular contact. For the dimer, there are 84 unique configurations of the two protomers.

Calculations were performed for the homo-dimer depicted in Figure 1. Using this structure, the site-specific amino acid probabilities were determined using entropy optimization for prescribed values of $\Delta_f$ and $\Delta_b$. The entropy surface $S(\Delta_f, \Delta_b)$ is presented in Figure 2, where $S$ has been normalized such that $\int d\Delta_f d\Delta_b \exp(S(\Delta_f, \Delta_b)) = 2^{27}$ (see Zou and Saven, 2000). The exact results in Figure 2 are determined using $S = \ln \Omega(\Delta_f, \Delta_b)$, where $\Omega$ is the number of sequences having common values of $\Delta_f$ and $\Delta_b$. In comparing the entropy surfaces, we note that there is excellent agreement of the theory with the exact results. Interestingly, those sequences that are optimal with regard to $\Delta_f$ are not those that are optimal with respect to $\Delta_b$ and vice versa. Sequences having low values of $\Delta_f$ are those for which the target structure is well removed from other competing folds. Such sequences have largely exposed hydrophilic groups. For this model energy function, such sequences have unfavorable values of the binding gap energy, $\Delta_b = 0$.

The polar probabilities of each residue $w_i(P)$ as obtained by theoretical calculation and enumeration are presented in Figure 3. There is excellent agreement. Although only probabilities calculated at $\Delta_f = -14\varepsilon$ and $\Delta_b = -5\varepsilon$ are shown for this illustration, excellent agreement is also obtained at other values of these energies (data not shown). For the sequences considered, $\Delta_f < 0$ and $\Delta_b < 0$. Within the context of this model, such sequences are likely to form stable isolated structures and/or stable complexes, i.e. sequences for which both the folded state of each chain and the oligomer structure are of lower energy than competing collapsed and mis-associated states.
Atomically detailed model

For realistic representations of proteins, an all-atom model is used that includes side chain conformational degrees of freedom. Side chains assume discrete conformations, rotamers (Ponder and Richards, 1987; Dunbrack and Cohen, 1997; Tuffery et al., 1997). Here we use the backbone-dependent rotamer library of Dunbrack and coworkers (Dunbrack and Cohen, 1997). The Amber force field (Weiner et al., 1984) is used to calculate non-bonded interactions. This same potential is used to quantify both the intra- and inter-molecular energies. To address the hydrophobic effect, the 'environmental energy' (Equation 5) for the complex is constrained to the value of the wild-type sequence (Kono and Saven, 2001). The site-specific probabilities are determined for multiple (constrained) values of \( E = E_f + E_a \), and conjugate to this energy is an effective inverse temperature \( \beta \). In what follows, the probabilities are presented at \( \beta = 0.5 \text{ mol/kcal} \), which is also the effective temperature at which (unfolded) reference energies of the amino acids are determined, as described in Calhoun et al. (2003).

We select the tetramerization domain of p53 tumor suppressor (PDB code: 1C26) as the target structure (Figure 4). This transcription factor is involved in a number of important physiological roles, including regulation of the cell cycle, apoptosis, DNA repair and angiogenesis, and mutations of p53 have been linked to cancers (Vogelstein et al., 2000). The domain considered here is a tetramer with four identical 32 residue chains. In this study, all 20 amino acid residues are allowed at each position. Thus, a total of \( 20^{32} = 4 \times 10^{41} \) sequences are possible for a single chain. Taking side chain conformations into account, there are a total of 320 states per residue position. In the absence of the symmetry assumption, there are 128 variable positions and \( 320^{128} \) possible states. The imposition of the symmetry constraint reduces this complexity to 32 variable positions and \( 320^{32} \) possible states; the number of independent site probabilities to be determined is reduced from \( 128^3 \times 320 \) to \( 32^3 \times 320 \).

Calculations are performed both with and without the symmetry assumption, permitting full sequence variability at each site in the tetramer. Without using symmetry assumption, the four chains have identical site-specific probabilities at equivalent positions. This is expected, since the structures are identical. In terms of both memory usage and computer time, the efficiency of the calculation scales as \( M^2 \) for a symmetrical structure with \( M \) subunits. For the p53 tetramer \( M^2 = 16 \), imposition of the symmetry assumption reduces the computation time by a factor of 10. For larger oligomers, we would expect even greater savings in computer time and memory usage; for infinite symmetries, e.g. crystals, the assumption makes calculations on such systems feasible. Comparing the results with and without symmetry assumption, we find excellent agreement (Figure 5).

In many cases, the site-specific probabilities are in agreement with the wild-type amino acid residues. For the well structured, stable tetrameric structure, it is not unreasonable to expect that wild-type residues are among the more probable at many sites. This finding is also in harmony with previous design studies that have identified sequences having consider-

![Fig. 5. Amino acid probabilities at selected sites for the p53 tetramerization domain. Results shown are those calculated without (open bars) and with (black bars) the symmetry assumption. Residues numbers and wild-type amino acid are indicated on each panel.](image)
able wild-type character (Kuhlman and Baker, 2000; Raha et al., 2000). The wild-type amino acid is the most probable at 31.3% (10/32) of the 32 positions (E326, L330, I332, G334, E336, E346, A347, E349, Q354 and G356, where the letters denote the wild-type amino acids). For example, see site L330 in Figure 5. At eight of the remaining positions (G325, R333, E339, F341, E343, L344, L348 and A353), the wild-type amino acids are among the five most probable (for example, see sites 333 and 341 in Figure 5). Sites Y327, T329, Q331, R335, R342, D352 and A355 tolerate many different amino acids, including the wild-type. Interestingly, these 25 sites also include interfacial positions where protomers come into contact, such as L330, I332, F341, L344 and A347. Nonetheless, there are seven positions that have calculated probabilities significantly different from the wild-type amino acids. Although for residues M340, N345 and L350 the wild-type identities are of much lower probability than the calculated most probable residues, the most probable amino acids maintain hydrophobic properties and suggest the following mutations: M340L, N345Y and L350V. At sites F328 and F338, the wild-type residues assume side chain conformations in which the side chains are in van der Waals contact and this arrangement cannot be recovered with the rotamer library used (Figure 4). This does, however, suggest mutations that may reduce side chain ‘strain’: F328H and F338T. In the wild-type, two intermolecular salt bridges R337–D352 and K351–E343 are formed (Figure 4), of which R337 and K351 are buried in the complex and are replaced by hydrophobic amino acids, R337V and K351V. The other two residues involved in the salt bridges, D352 and E343, are partially exposed, and the corresponding most probable residues turn out to be a neutral, hydrophobic amino acid Q (D352Q and D343Q) in both cases. The seven suggested mutations may increase the stability of this p53 tetramer, but this may come at the expense of function, since no information about the binding properties other than oligomerization was included in the calculations.

Summary

Herein we have provided a formalism for the probabilistic design of oligomeric protein structures. The theory quantitatively recovers the exact results obtained from a simple lattice model of proteins. The application to an atomically detailed representation of the protein p53 tetramerization domain reveals the utility and accuracy of the symmetry assumption and also recovers many of the sequence properties observed in the wild-type structure. The probabilities obtained from the theory may be used either iteratively in protein design, where increasing numbers of residues are specified with each round of the calculation, or to guide a biased search for low energy sequences (Zou and Saven, 2003). These probabilities may also be used to specify the composition of a protein combinatorial library. Such computational methodologies will be useful in designing large, oligomeric protein complexes or multidimensional arrays of proteins for biomaterials applications.

Acknowledgements

We acknowledge support from the NSF (CHE 99-84752 and DMR 00-79909). J.G.S. is a Cottrell Scholar of Research Corporation and an Arnold and Mabel Beckman Foundation Young Investigator.

References


Received August 20, 2003; revised October 24, 2003; accepted October 28, 2003