July 2004

Progress in the development and application of computational methods for probabilistic protein design

Sheldon Park
University of Pennsylvania, sheldonp@sas.upenn.edu

Hidetoshi Kono
Japan Atomic Energy Research Institute

Wei Wang
University of Pennsylvania, weiw@sas.upenn.edu

Eric T. Boder
University of Pennsylvania, boder@seas.upenn.edu

Jeffery G. Saven
University of Pennsylvania, saven@sas.upenn.edu

Postprint version. *Computers & Chemical Engineering* (in press)

Progress in the development and application of computational methods for probabilistic protein design

Abstract
Proteins exhibit a wide range of physical and chemical properties, including highly selective molecular recognition and catalysis, and are also key components in biological metabolic, catabolic, and signaling pathways. Given that proteins are well-structured and can now be rapidly synthesized, they are excellent targets for engineering of both molecular structure and biological function. Computational analysis of the protein design problem allows scientists to explore sequence space and systematically discover novel protein molecules. Nonetheless, the complexity of proteins, the subtlety of the determinants of folding, and the exponentially large number of possible sequences impede the search for peptide sequences compatible with a desired structure and function. Directed search algorithms, which identify directly a small number of sequences, have achieved some success in identifying sequences with desired structures and functions. Alternatively, one can adopt a probabilistic approach. Instead of a finite number of sequences, such calculations result in a probabilistic description of the sequence ensemble. In particular, by casting the formalism in the language of statistical mechanics, the site-specific amino acid probabilities of sequences compatible with a target structure may be readily identified. The computational probabilities are well suited for both de novo protein design of particular sequences as well as combinatorial, library-based protein engineering. The computed site-specific amino acid profile may be converted to a nucleotide base distribution to allow assembly of a partially randomized gene library. The ability to synthesize readily such degenerate oligonucleotide sequences according to the prescribed distribution is key to constructing a biased peptide library genuinely reflective of the computational design. Herein we illustrate how a standard DNA synthesizer can be used with only a slight modification to the synthesis protocol to generate a pool of degenerate DNA sequences, which encodes a predetermined amino acid distribution with high fidelity.

Keywords
proteins, peptide sequences, computational analysis

Comments
Postprint version. *Computers & Chemical Engineering* (in press)

Progress in the development and application of computational methods for probabilistic protein design

Sheldon Park[1], Hidetoshi Kono[2], Wei Wang[1], Eric T. Boder[3], and Jeffery G. Saven[1*]

1. Makineni Theoretical Laboratories, Department of Chemistry, University of Pennsylvania, 231 South 34th Street, Philadelphia, Pennsylvania 19104, USA.
2. Neutron Research 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.
3. Department of Chemical and Biomolecular Engineering, University of Pennsylvania, 220 South 33rd Street, Philadelphia, PA 19104, USA.

*Corresponding author:

Jeffery G. Saven

Department of Chemistry

University of Pennsylvania

231 South 34th Street

Philadelphia, PA 19104

Tel: 215-573-6062

Fax: 215-573-2112

Email: saven@sas.upenn.edu

Abstract

Proteins exhibit a wide range of physical and chemical properties, including highly selective molecular recognition and catalysis, and are also key components in biological metabolic, catabolic, and signaling pathways. Given that proteins are well-structured and can now be rapidly synthesized, they are excellent targets for engineering of both molecular structure and biological function. Computational analysis of the protein design problem allows scientists to explore sequence space and systematically discover novel protein molecules. Nonetheless, the complexity of proteins, the subtlety of the determinants of folding, and the exponentially large number of possible sequences impede the search for peptide sequences compatible with a desired structure and function. Directed search algorithms, which identify directly a small number of sequences, have achieved some success in identifying sequences with desired structures and functions. Alternatively, one can adopt a probabilistic approach. Instead of a finite number of sequences, such calculations result in a probabilistic description of the sequence ensemble. In particular, by casting the formalism in the language of statistical mechanics, the site-specific amino acid probabilities of sequences compatible with a target structure may be readily identified. The computational probabilities are well suited for both de novo protein design of particular sequences as well as combinatorial, library-based protein engineering. The computed site-specific amino acid profile may be converted to a nucleotide base distribution to allow assembly of a partially randomized gene library. The ability to synthesize readily such degenerate oligonucleotide sequences according to the prescribed distribution is key to constructing a biased peptide library genuinely reflective of the computational design. Herein we illustrate how a standard DNA

synthesizer can be used with only a slight modification to the synthesis protocol to generate a pool of degenerate DNA sequences, which encodes a predetermined amino acid distribution with high fidelity.

1. Introduction

Nature has been engineering nanoscale devices in the form of proteins for billions of years. In particular, biological proteins epitomize the elegance of biomolecules with their versatility, high fidelity synthesis, efficiency, and atomically well-defined complex structures. Proteins and other folding biomolecules perform a wide variety of functions within living organisms including selective catalysis, molecular recognition, cell signalling, and energy transduction. These machines have been "engineered" via evolution, but the potential exists to harness and extend the versatility of proteins and other folding chain molecules to provide new types of molecular structures and devices. Herein, we discuss recent methods for the design of particular proteins and---especially for cases where the determinants of structure and molecular function are not transparent---the probabilistic design of ensembles of proteins.

Despite their well-defined structures, proteins possess many special "processing" properties that provide versatile routes to structures that are ordered at nanometer length scales. Being linear heteropolymers, proteins are straightforward to synthesize from their constituent amino acids via solid phase peptide synthesis or the expression of an encoding gene in a suitable cell line. Proteins are not machined; no molding, templating, masking or other techniques are necessary for the protein to form a particular structure. Under appropriate conditions, the protein spontaneously folds to a well-defined three-dimensional shape, where the information necessary for this ordering is contained in the sequence of monomers, the amino acids. Upon folding, essentially a single structure is selected from among a huge ensemble of possible collapsed, unstructured conformations.

Stabilizing this folded state are a combination of effects: constraints on bond lengths and bond angles due to the covalent connectivity of the backbone and a myriad of noncovalent interactions, including van der Waals forces, hydrogen bonding, hydrophobic effects, and electrostatic interactions. The folded state of most proteins is a thermodynamic minimum and is robust with respect to thermal fluctuations. Exquisite control of structure is potentially achievable via the manipulation of sequence, which in turn extends the possibility of the *de novo* design of predetermined backbone structures, i.e., fold topologies. The search for sequences with desired structural and functional properties is also partially facilitated by the fact that there may be more than one solution. We have many natural examples of sequences folding to essentially the same overall structure, but these sequences may have little similarity with one another. Successful protein design has advanced dramatically with the use of computational methods to address the enormous range of monomer and conformational variability that is possible with these molecules.

1.1. Protein design

The design of proteins requires three main elements (Saven, 2001). One being a target structure or topology, which usually involves specifying the coordinates of atoms in the polypeptide backbone. Such structures may be extracted from the protein data bank (PDB)(Sussman et al., 1998) or created via molecular modeling (Summa et al., 1999). Secondly, an energy function is needed so as to quantify the compatibility between sequence and a particular structure. Lastly, constraints are typically imposed on the sequences. Such constraints can be used to pattern hydrophobic and hydrophilic amino acids appropriately(Kamtekar et al., 1993) or to specify the identities of amino

acids in the vicinity of a site in the structure to be used for binding or catalysis (Benson et al., 1998a; DeGrado et al., 1999).

Successes in protein design have come despite the fact that the noncovalent forces stabilizing proteins are some of the most difficult to quantify accurately. Most molecular potentials in use for protein design result (a) from fitting interatomic interactions to data derived from quantum chemical calculations or from spectroscopic experiments, resulting in molecular force fields such as CHARMM (Brooks et al., 1983) or AMBER (Weiner et al., 1984), or (b) from fitting the potentials to the structural preferences observed among atomically detailed structures in the PDB. Fortunately, current energy functions have proven practical, and the successful design of proteins is possible even without a complete, detailed quantitative understanding of all the forces involved in specifying their structures and stabilities. Via the design of targeted protein structures, including those having desired functions, our understanding of the forces and effects that specify the properties of the folded state can be further refined.

Designing proteins is nontrivial due to both the subtlety of the interactions that stabilize and specify the folded state and their conformational complexity. Proteins are large (tens to hundreds of amino acid residues), and many variables are required to specify the folded state: the amino acid sequence, the global (tertiary) structure of the backbone, and the side chain conformations. Each monomer (residue) has a pendant side chain, which may have multiple conformations even when the backbone structure is specified. In addition to structural complexity, there is also sequence complexity. Design involves identifying folding sequences from the enormous ensemble of possible sequences: for a modestly sized 100-residue protein, there are more than 10^{130} possible

sequences. Exhaustive searching of the m^N possible sequences is tractable only if the number of variable sites in the structure N is small and/or the number of degrees of freedom per monomer site m is greatly reduced, where the site degrees of freedom include both different amino acid identities and conformational states. If the different side-chain conformations (rotamer states (Dunbrack & Cohen, 1997; Ponder & Richards, 1987)) of each amino acid are considered, the complexity of the search is further compounded, since m may easily increase by an order of magnitude. As a result, complete enumeration of sequences is typically intractable for all but the smallest peptides. However, a number of search algorithms exist that identify suitable sequences without exhaustive enumeration. These searches are guided by the large degree of "consistency" seen in folded proteins (Go, 1983). On average, a folded protein is atomically well-packed with favorable van der Waals interactions, hydrophobic residues are sequestered from solvent, most hydrogen bonding interactions are satisfied, and the residues are in harmony with their local secondary structural preferences (α-helix or β-strand). However, it has been known since the first protein structure appeared that this global consistency is often complex and may have little simplifying symmetry. As mentioned, noncovalent interactions are some of the most difficult to accurately quantify, and estimating free energies associated with mutation or structural ordering remains a subtle and computationally demanding area of physical chemical research (Brooks, 2002; Shea & Brooks, 2001). Despite the predictive capabilities of molecular potentials, presently it is impractical to determine the relative stability changes of a large number of sequences using the type of detailed simulation methods commonly used to estimate such free energy differences. Nonetheless, molecular potentials derived from small molecules

and from the protein structure database do contain at least *partial* information about the interactions and forces known to be important for specifying and stabilizing protein structures, e.g., van der Waals interactions, steric (excluded volume) interactions, and hydrogen bonding. In some cases, the optimization of such potentials has lead to striking successes in protein design (Kraemer-Pecore et al., 2001). These potentials are necessarily approximate, however, and any sequence so designed may be sensitive to the particular potential and target structure used. Alternatively, the partial information contained in these potentials may be used in a probabilistic manner to yield directly the site-specific likelihoods of the amino acids. A probabilistic approach is also appropriate for characterizing the variability of sequences that fold to a common structure, where complete sequence enumeration is impractical.

Probabilistic approaches are particularly appropriate for de novo protein design in the context of combinatorial protein experiments, which create and rapidly assay many different sequences (Moffet & Hecht, 2001). Even though combinatorial methods can address large numbers of sequences (10^4-10^{12}), these numbers are still infinitesimal compared to the numbers of possible protein sequences, e.g., $20^{100} \approx 10^{130}$ for a 100-residue protein. Thus even with combinatorial methods, we still must focus on selected regions of sequence space. This pre-selection is most often done by identifying a few residues within the protein by inspection and allowing full or partial variability at these sites. Recently, computational methods have been developed that can keep track of a much wider range of sequence variability and provide quantitative methods for selectively winnowing the sequence search space. Here, we discuss directed and

probabilistic computational methods for sequence design with an emphasis on design applications of the probabilistic methods.

1.2. "Directed" methods of protein design

The sequence energy landscape may be explored in a directed manner so as to identify sequences having low energies when they take on the target structure. This we refer to as "directed protein design." Such sequences can then be realized experimentally, and their thermodynamic stability, structure, and functional properties may be determined. Early efforts in de novo design were guided by the local structural tendencies of the amino acids as inferred from known structures of natural proteins. Such tendencies include preferences for secondary structure(O'Neil & DeGrado, 1990) and exposure to solvent (hydrophobic effects) (Miller et al., 1987; Rose et al., 1985). Such inspection-based efforts identified polypeptides that were compact and had substantial secondary structure, but these early sequences did not necessarily fold to well-defined tertiary structures (Bryson et al., 1995). With their abilities to address quantitatively systems containing large numbers of degrees of freedom, computational methods have dramatically accelerated advances in protein design. Such methods most often cast the sequence search as an optimization process, wherein amino acid identity and side chain conformation are varied in order to optimize a scoring or energy function, which serves as a sequence-structure compatibility metric. Furthermore, in arriving at sequences where the target structure has a well-packed interior and favorable inter-atomic interactions, the search must include variation in the side-chain conformations (rotamer states) of each amino acid (see (Dunbrack, 2002)). Since the number of possible "states" for a residue can increase rapidly when the amino acid side chain conformers are taken

into account, it is often necessary to allow only a few residues to vary at a time while constraining the conformations of the remaining residues. Although complete enumeration is often not feasible, the sequence space can be sampled in a directed manner so that the search moves progressively toward optimal (or nearly optimal) sequences. Genetic algorithms and simulated annealing are two commonly used stochastic methods that search the sequence space in a partially random fashion (Desjarlais & Handel, 1995; Hellinga & Richards, 1994; Jones, 1994; Shakhnovich & Gutin, 1993). Such searches have sufficient "noise" or recombination to permit escape from local minima in the sequence-rotamer landscape while at the same time preferentially sampling low energy states as the calculation progresses. When applied to atomically detailed representations, the stochastic methods focus primarily on repacking the interior of a structure with hydrophobic residues (Hellinga & Richards, 1994). Both genetic algorithms and simulated annealing have been successfully applied to re-design a number of natural proteins: 434 Cro (Desjarlais & Handel, 1995), ubiquitin (Johnson et al., 1999), the B1 domain of protein G (Jiang et al., 2000), the WW domain (Kraemer-Pecore et al., 2001), and helical bundles (Bryson et al., 1998; Jiang et al., 1997). In many cases, these methods have aided in identifying experimentally viable sequences (Kraemer-Pecore et al., 2001; Walsh et al., 1999). However, since stochastic search methods do not always identify global optima (Voigt et al., 2000), other search methods have also been developed. For molecular potentials involving only site (residue-backbone) and pair (residue-residue) interaction energies, elimination and pruning methods such as "dead end elimination" can find the global optima (Gordon & Mayo, 1998; Gordon & Mayo, 1999; Looger & Hellinga, 2001; Pierce et al., 2000; Voigt et al., 2000). By

considering possible pair interactions between residues, these pruning methods successively remove from each site amino acid-rotamer states that cannot be part of the global optimum until no further states can be eliminated. Dead end elimination was applied to the full sequence design of a 28-residue zinc finger mimic (Dahiyat & Mayo, 1997), as well as a 51-residue homeodomain motif once the protein has been patterned with hydrophobic and polar sites (Marshall & Mayo, 2001). Residue subsets within portions of a variety of proteins have been redesigned (Malakauskas & Mayo, 1998; Shimaoka et al., 2000; Strop & Mayo, 1999). Functional properties such as metal or small molecule binding or catalysis may also be included as elements of the design process (Benson et al., 1998b; Bolon & Mayo, 2001; DeGrado et al., 1999; Looger et al., 2003). Directed protein design has been the subject of several recent reviews (Kraemer-Pecore et al., 2001; Saven, 2001; Street & Mayo, 1999).

Despite some striking successes, computational methods for the directed design of sequences have limitations. While stochastic methods can be applied to large proteins and permit many sites to be varied simultaneously, the computational times and resources used for such calculations can be extensive even for small proteins. Directed methods are necessarily sensitive to the energy or scoring function used, since they identify low energy states of a particular energy function. These energy functions are approximate, however, and inaccuracies in the energy function may not merit the search for global optima. In addition, many naturally occurring proteins are not optimized. In fact, most proteins are only marginally stable, e.g., $\Delta G^o < 10$ kcal/mol for folding (Gromiha et al., 2002), and mutants of natural proteins with increased stability are well known (Eriksson et al., 1992). This is not unexpected since nature selects for biological activity rather than

stability. Proteins need to be stable to be structured, but hyperstability is usually only a requirement of organisms such as hyperthermophiles. Thus it is important to develop methods complementary to those used for directed protein design, methods that reveal the properties of those sequences likely to fold to a particular structure but which may not be structurally "optimal." Such methods may be used in designing particular proteins and may also be applied to a new class of protein design studies, combinatorial experiments, where large numbers of proteins may be synthesized and subsequently assayed for desired structures and activities.

1.3. Probabilistic approaches to protein design

In the context of protein design, the use of site-specific amino acid probabilities rather than specific sequences is here referred to as *probabilistic* protein design. Probabilistic approaches, as opposed to directed approaches, are often used in science and engineering for cases where we have incomplete information about a system. For protein design, such a probabilistic approach is motivated by the complexity and uncertainty associated with describing folded proteins. Probabilistic design methods directly provide sequence information, particularly with regard to structurally important amino acids. The amino acid probabilities can guide the design of specific sequences and can also highlight sites likely to tolerate mutation with minimal impact on structure; such sites can be targets of variation upon multiple rounds of protein design and can mark regions likely to tolerate mutations that can confer biological activity or other desired properties.

The probabilistic methods described below may be used in several ways to guide protein design. In each case, sequences are generated in a manner consistent with the

calculated probabilities. Firstly, the most straightforward choice is a low energy consensus sequence, the sequence comprising the most probable amino acid at each position. Although the consensus sequence would not necessarily account directly for correlations between residue identities due to inter-residue interactions, such correlation may be better addressed by an iterative series of calculations, each time constraining an increasing number of amino acid degrees of freedom until a unique sequence is identified. Such an approach has been used in the design of a 114-residue four-helix metalloprotein (Calhoun et al., 2003). Secondly, the calculated probabilities may be used to guide a search algorithm. This approach has been applied to develop an efficient Monte Carlo (MC) based method which uses predetermined amino acid probabilities to bias the generation of trial sequences at each step of the Monte Carlo Markov trajectory (Zou & Saven, 2002). These methods are similar to well-known configurational bias MC methods but use a predetermined probability-profile to bias the sampling. Lastly, probabilistic methods may be used to quantitatively guide the design of combinatorial libraries of proteins, in which an ensemble of biased sequences is generated in a manner that best reproduces the calculated site-specific amino acid probabilities.

1.4. Combinatorial experiments

Combinatorial protein experiments can be used to investigate sequence-structure compatibility and to discover novel sequences folding to desired structures. In protein combinatorial design experiments, large numbers of sequences (libraries) are synthesized and screened for evidence of folding to predetermined structures or for a desired activity, e.g., small molecule binding or enzymatic activity. Depending upon how the sequence diversity is generated and assayed, experiments of this type can explore a large number of

sequences, up to 10^{12} (Keefe & Szostak, 2001). The protein diversity may be generated via solid phase peptide synthesis, but more commonly a library of partially random genes is expressed in phage (Hoess, 2001), bacteria (*E. coli*) (Kamtekar et al., 1993), or yeast (*S. cerevisiae*)(Boder & Wittrup, 1997). Such experiments can go "beyond the protein sequence database," since the diversity of the sequences is at the control of the researchers. Sequence features important to folding (and other biological properties) may be explored in a manner decoupled from the evolutionary pressures that determine the sequences of Nature's proteins. Combinatorial protein experiments have been used to identify helical proteins (Rojas et al., 1997; Roy et al., 1997a; Roy et al., 1997b), ubiquitin variants (Finucane et al., 1999), high affinity antibodies (Boder et al., 2000), self-assembled protein monolayers (Xu et al., 2001), proteins with amyloid-like properties (Xu et al., 2001), metal-binding peptides (Case & McLendon, 2000), and stable inter-helical oligomers (Arndt et al., 2000). Several excellent reviews of combinatorial experiments and methodology have appeared recently (Giver & Arnold, 1998; Hoess, 2001; Moffet & Hecht, 2001; Zhao & Arnold, 1997).

2. Methods for probabilistic protein design

For a given target structure, there are several methods that may be used to identify the site specific probabilities of the amino acids for sequences likely to fold to this structure. We give special emphasis here to statistical methods (section 2.3) that provide the amino acid probabilities directly.

2.1. Alignment of related sequences

The sequence variability of a protein structure can be examined using sequence and structural databases. Sequences known to have similar structures can be identified from the PDB or from a database of structural alignments (Holm & Sander, 1998). If the structure of a sequence is known, other proteins having sufficient sequence similarity (e.g., greater than 40% sequence identity) may be assumed to share the same overall structure. Multiple sequence alignments (MSA) of such similar sequences can used to estimate the site specific probabilities as simply the frequencies of each amino acid at each position in the alignment (Luthy et al., 1992). In case the number of sequences is insufficient to fully represent all amino acids at particular sites, pseudo-count and other methods may be used to "regularize" the frequencies (Durbin, 1998). Nonetheless, the probabilities from such a profile are likely to be heavily biased by known sequences in the database. This is especially a problem for sequences with a small number of other homologous sequences. At conserved sites, it may be difficult to resolve the determinants of amino acid identity, such as whether conservation arises from functional or structural constraints on the sequences. Finally,the database-derived probabilities are perforce inappropriate for engineering novel protein structures. Since there are many natural examples of sequences folding to similar structures but with little sequence similarity, it is desirable to obtain a broader understanding of the full range of sequence variability. Unencumbered with evolutionary constraints, more transferable computational methods permit direct determination of the amino acid probabilities using only a given backbone structure as a template.

2.2. Directed search methods to build profiles

Directed search methods can estimate the properties of an ensemble of sequences by repeatedly applying the search to build up a set of low energy sequences. A target structure is chosen by specifying the coordinates of the main chain (backbone) atoms. Several recent directed design studies have obtained sequences similar to the wild type sequence when a single protein structure was used (Koehl, 1999a; Koehl, 1999b; Kuhlman & Baker, 2000; Raha et al., 2000). For a given structure, multiple sequence search calculations may be run independently, resulting in a set of sequences whose alignment yields the site-specific probabilities. This approach was adopted by Desjarlais and coworkers, who used independent runs of their sequence prediction algorithm (SPA) for each member of an ensemble of closely related structures consistent with a particular fold. They were able to identify sequences consistent with the fold of a WW domain, a small beta-sheet protein (Kraemer-Pecore et al., 2001), some of which have been experimentally characterized. By designing sequences for each of 100 minor structural variants (1 Å root mean square deviation) of a particular fold using the SPA of Desjarlais, Larson et al have built computational profiles exhibiting much more diversity than those obtained using a single structure (Larson et al., 2002). Workers at Xencor, Inc. have recently used Monte Carlo sampling of sequences to mutate residues in the vicinity of the active site of β-lactamase (Hayes et al., 2002). Sequences with more than 1000-fold increases in resistance to an antibiotic were identified. Though obviously useful, these approaches to building profiles require repeated directed searches in order to build the site specific frequencies of the amino acids, making such calculations computationally demanding. Nonetheless, with the advent of faster processors and improved algorithms,

such calculations are not infeasible. Although stochastic and directed search methods may work well for optimization, it is not straightforward to impose constraints on the values of effective energies of the sequences in such methods. Given that empirical energies (e.g., a hydrophobicity score) may be well determined or bounded---but not optimized---for known folding sequences, it is of interest to develop methods than allow the facile implementation of such constraints.

2.3. Statistical theory of sequence ensembles

An alternative theory to identify amino acid probabilities for a given backbone structure has been developed (Kono & Saven, 2001; Zou & Saven, 2000). This entropy-based formalism applies statistical mechanical concepts to directly estimate the number of sequences and the site-specific probabilities. The theory addresses the whole space of available compositions and is not limited to the small fraction that is accessible to experiment or to computational enumeration and sampling. Using this approach, the features of suboptimal sequences may be readily examined. Large protein structures (more than 100 residues) can also be easily accommodated in the calculation. Here the "entropy" quantifies the variability of sequences consistent with the target structure. The number of possible sequences is reduced using concepts from thermodynamics—decreasing the energy or imposing constraints on the system reduces the entropy and hence diminishes the number of allowed sequences.

A key concept in this methodology is the notion of entropy maximization, which is also fundamental to statistical mechanics and information theory. There are an infinite number of possible sets of site-specific state probabilities, where "state" is defined by both monomer identity and side chain conformation. The most probable set of such

probabilities is determined by optimizing an effective entropy function, where the maximization is done subject to constraints. The method takes as input (a) a target structure determined by the coordinates of the backbone atoms, (b) energy functions that quantify sequence-structure compatibility, and (c) constraints on amino acid properties. For a target structure, the method estimates the probabilities of individual identity-rotamer states at each residue position. Both global considerations (e.g., the overall energy of the sequences in the target) and local features (e.g., the allowed amino acids at particular sites) can be specified via constraints. With judicious application of such constraints, the method provides a systematic means to reduce the size of sequence space to be searched to an experimentally feasible size.

Among sequences with desired properties as specified by constraint functions, let $w_i(\alpha, r_k(\alpha))$ denote the probability that amino acid α is present at residue position i, and its side chain is in a discrete rotamer state $r_k(\alpha)$ (Dunbrack & Cohen, 1997). The total sequence-conformational entropy S_c (here simply referred to as "conformational entropy") is written as

$$S_c = -\sum_{i,\alpha,k} w_i(\alpha, r_k(\alpha)) \ln w_i(\alpha, r_k(\alpha)) \qquad (1)$$

The sum extends over each sequence position i and all available amino acids α at each position. Furthermore, for each amino acid, the sum is taken over each of the k possible rotamer states $r_k(\alpha)$. Implicit in writing the entropy S_c in this manner is a factorization approximation, which would seem to imply that the probabilities are independent. Constraints on the sequences, however, will cause these probabilities to be coupled to one

another. The $w_i(\alpha,r_k(\alpha))$ are determined as those that maximize S_c subject to any constraints f_i, which are functions of the $w_i(\alpha,r_k(\alpha))$. In order to impose these constraints during maximization, a variational functional V is defined using the method of Lagrange multipliers:

$$V = S - \beta_1 f_1 - \beta_2 f_2 - \cdots \qquad (2)$$

To identify the state probabilities consistent with particular values of constraints, the m-th constraint function f_m is constrained to have a particular value $f_m^o = f_m(\{w_i(\alpha,r_k(\alpha))\})$. The constraint functions f_m may be used to specify such features as the overall energy of the structure or the patterning of residue properties. The set of equations that must be solved simultaneously to determine the probabilities and the Lagrange multipliers β_i then take the form:

$$\partial V / \partial w_i(\alpha,r_k(\alpha)) = 0 \qquad (3)$$

$$f_m(\{w_i(\alpha,r_k(\alpha))\}) = f_m^o \qquad (4)$$

This large set of on the order of 10^4 coupled, nonlinear equations is solved using root finding methods (Press et al., 1992). These equations may be solved for a series of constraint values to study how the amino acid profiles vary with the value of the constraint, e.g., how the probabilities change as the overall conformational energy is decreased. A typical calculation for a protein of 100 residues allowing all 20 amino acids at each position can be completed in less than 24 hours using a 1GHz Pentium IV processor. Typically, the site probabilities are sensitive to their initial conditions only for

very low effective energies and temperatures, well below those values where the probabilities to be used in protein design are determined (see Fig. 2 and discussion of effective heat capacity C_v in the Appendix).

Detailed descriptions of the energetic constraints have been discussed elsewhere (Calhoun et al., 2003; Kono & Saven, 2001) and are also presented in the Appendix. Here we briefly summarize how such effects are included in the calculations. A key energetic constraint on the sequences involves an effective "conformational energy" E_c, which quantifies van der Waals, electrostatic, and hydrogen bonding energies. The conformational energy consists of one-body energies which involve interactions of the side chains with the backbone as well as two-body energies involving interactions between side chains. E_c essentially accounts for intramolecular interactions within the protein. Effective reference energies for each amino acid may also be included in the conformational energy to crudely approximate the energetics of the unfolded ensemble of conformations. A second energy constraint involves an "environmental energy" E_{env}, which accounts for the solvation preferences of the amino side chains and hence quantifies intermolecular interactions involving protein and solvent. This environmental energy is a database-derived effective one-body potential and is a function of the local density of C_β atoms ρ about a particular amino acid side chain. Higher values of ρ appear more often within the interior of a protein than on the surface, and thus a local effective energy dependent on ρ may effectively parameterizes the propensities of amino acids to be buried within the typically hydrophobic interior of a folded protein or to be exposed to solvent. Conjugate to these effective energies, E_c and E_{env}, are effective temperatures, T_c and T_{env}, that arise as Lagrange multipliers during maximization subject

to constraints on the energies. We may choose to characterize the sequence energy landscape in terms of either effective energy or temperature.

By way of example the theory has been applied to a 57-residue protein, the SH3 domain. The conformational entropy decreases as the effective temperature T_c, or equivalently E_c, decreases (Fig. 1). At high energies (corresponding to high T_c), many high energy interactions between the possible amino acids are permitted, and one observes a broad distribution of sequence-rotamer states at each site. However, as energy is lowered, the number of probable amino acids and rotamer states per site also decreases on average. From a thermodynamic perspective, this is consistent with a positive temperature, since $1/T = \left(\frac{\partial S}{\partial E}\right)_{N,V}$. As shown in Fig. 1, C_v possesses a peak around 10 kcal/mol and reaches a valley around T_c = 2 kcal/mol, where the identities and conformations of residues in the interior become relatively well defined (Kono & Saven, 2001). Still, surface residues, while predominantly hydrophilic, may occupy a large number of rotamer states with comparable probability. This is consistent with the observation seen in structural databases that surface exposed residues are often less well defined structurally than interior residues. Thus, the heat capacity is a useful quantity to help determine at which "effective temperature" one should examine the amino acid probabilities. In addition, direct comparisons of the calculated profile shows good agreement with one obtained by sequence alignments using the HSSP database (Sander & Schneider, 1991). A few representative amino acid profiles from the buried region are shown in Fig. 2.

3. Gene libraries from site-specific probabilities

In section 1.3, we discussed how to make use of the site-specific amino acid probabilities. A specific sequence may be identified either as a consensus sequence or as the result of a directed search. This sequence can then be synthesized by solid phase peptide synthesis or expressed from a constructed gene coding for the sequence. In general, larger proteins (more than 50 residues) are most often created using gene expression. In order to apply the computed probabilistic sequence information to construct a combinatorial library, the protein profiles must first be reverse-translated into libraries of partially random gene sequences. As genes consist of nucleotide triplets, or codons, each of which is translated to a specific amino acid, non-uniform distributions of nucleotides are necessary to encode libraries of sequences with the amino acids at variable sites appearing with the predetermined probabilities. To obtain such libraries, we need computational methods to solve for the nucleotide frequencies consistent with the amino acid profiles, and methods to generate an ensemble of genes having predetermined nucleotide frequencies at selected positions.

3.1. Computational design of gene libraries

Pseudo-independent nucleotide probabilities at each position of a set of partially random genes can be calculated to best reproduce a protein library having the calculated set of site specific amino acid probabilities (Wang & Saven, 2002). The calculated gene library can then be constructed by standard DNA synthesis, as shown in the next section.

Let $P_1(n_1)$, $P_2(n_2)$, $P_3(n_3)$ be the probabilities of each of the four possible nucleotides (n_i = A, T, G, C) in the first, second and third position of a codon respectively.

If these are treated as independent, the probability that amino acid α will appear as encoded by the codon $n_1n_2n_3$ is $P(\alpha|n_1,n_2,n_3) = P_1(n_1)P_2(n_2)P_3(n_3)\delta(\alpha|n_1,n_2,n_3)$, where $\delta(\alpha|n_1,n_2,n_3) = 1$ only if n_1,n_2,n_3 is a codon for amino acid α and is zero otherwise. If the codons of amino acid α are equally likely (no codon bias), the probability of an amino acid α is the sum of codon probabilities corresponding to this amino acid.

$$P_{calc}(\alpha) = \sum_{n_1,n_2,n_3} P_1(n_1)P_2(n_2)P_3(n_3)\delta(\alpha \mid n_1,n_2,n_3) \qquad (5)$$

The next step is to identify an objective function that quantifies the deviation of the amino acid probabilities encoded by a given set of nucleotide probabilities from the desired set of site specific amino acid probabilities (Jensen, 1998; Wolf & Kim, 1999). To identify the nucleotide probabilities that not only best reproduce the desired amino acid frequencies but also prevent the occurrence of stop codons, a new objective function has been presented(Wang & Saven, 2002). The objective function comprises both a relative entropy term and a χ^2 function, which quantifies the difference between the desired and calculated amino acid probabilities in a least-squares manner. Relative entropies are commonly used to measure the 'distance' between two probability distributions and are strong indicators of when information in one distribution is not contained in the other (Durbin, 1998). The objective function takes the following form.

$$H = \sum_{\alpha=1}^{21} \{P_{calc}(\alpha)\ln\frac{P_{calc}(\alpha)+\varepsilon}{P_{des}(\alpha)+\varepsilon} + 0.5[P_{des}(\alpha)-P_{calc}(\alpha)]^2\} \qquad (6)$$

Here ε is introduced as an arbitrary small constant ($\varepsilon = 10^{-6}$) so as to avoid numerical instability in case $P_{des}(\alpha)$ vanishes. The stop codons are treated as the 21st amino acid for the purpose of this calculation. The objective function is optimized (minimized) subject to the usual constraints on the nucleotide probabilities: $0 \le P_i(n_i) \le 1$ and $\sum_{n_i} P_i(n_i) = 1$. This may be done by the Lagrange multiplier method or using computational packages available for constrained minimization (Wang & Saven, 2002). Codons optimized for a particular organism or expression system may also be included in objective function of this type (Wang & Saven, 2002).

As an example, this nucleotide design approach was applied to residue 54 of the SH3 domain discussed in the previous section, and the resulting nucleotide distributions are summarized in Fig. 3. The desired and encoded frequencies of the amino acids are displayed as open and filled bars, respectively, at the top of Fig. 3, while the computed nucleotide probabilities are shown at the bottom. The agreement between the two is excellent in this case, and in general, the calculated probabilities agree well with desired ones. In some cases, an exact match between the desired and calculated probability distribution cannot be achieved due to the partial degeneracy of codons to amino acids. This computational method, however, provides excellent yields of complete sequences, avoiding stop codons, which would otherwise render the protein incomplete. For several test proteins in which more than 50 residues were subject to selective randomization, the yield of complete sequences was 96% or greater, substantially better than other methods of determining nucleotide frequencies. High yields are particularly important when a large fraction (or all) of a gene is subject to combinatorial mutation, as introduction of premature stop codons can offset the advantages of having a large diversity library.

3.1. Synthesis of oligonucleotides subject to arbitrary nucleotide probabilities

While peptide libraries may be chemically synthesized if the peptide chain is short enough, there are times when the peptide chain is too long to be easily assembled. In such cases, the library can be assembled in a biological system (e.g. phage, bacteria, yeast, or mammalian cells) before being screened. This is accomplished by designing suitable DNA oligonucleotides (oligos), assembling the oligos to create the genes by polymerase chain reaction or primer extension, and expressing the gene products in vivo. In order to introduce targeted mutations in the library, degenerate oligos specifically tailored for the purpose must first be synthesized. However, the synthesis of all but the simplest degenerate oligos is technically difficult and often costly. For each degenerate nucleotide (base), various phosphoramidites need to be individually weighed out and be premixed. This process is time-consuming and error-prone. Once the mixture is prepared, it must be installed on the DNA synthesizer as a distinct base, occupying a dedicated port. As the number of such ports is typically limited, only a few can be fitted at any given time. For oligos containing more degenerate positions than there are ports, the synthesis must be halted regularly to allow bottles to be exchanged. This chain of events is labor-intensive, not amenable to parallel synthesis, and often results in a low overall yield due to the repeated stop-and-go cycles. Manual mixing of phosphoramidites also results in waste of reagents, as the leftover mix not used up by the end of the synthesis goes to waste, thus driving the cost of custom oligos higher.

Hence, it is preferable to develop an alternate means of synthesizing custom degenerate oligos that is reliable, efficient and cost-effective. The availability of such a

method would not only allow biased libraries to be constructed with high accuracy but would also reduce the cost of assembling them by keeping the prices of degenerate oligos down. In order to demonstrate the feasibility of automated synthesis of arbitrary degenerate oligos, we synthesized degenerate 7-base pair oligos of the sequence $ACX_1X_2X_3GT$, where X_i is a degenerate base. For the present study, the degenerate positions comprised the following arbitrary distributions of the bases:

Base	A	C	G	T
X_1	20%	20%	20%	40%
X_2	40%	30%	10%	20%
X_3	10%	20%	60%	10%

Table 1. Input nucleotide base probabilities

During the standard DNA synthesis, each base is coupled to the growing chain of oligonucleotides by adding a predetermined amount (to be determined by the scale of the synthesis) of activated phosphoramidite to a column containing resin. This is done in practice by periodically opening the mechanical valve controlling the outflow of phosphoramidite. The same principle can be applied to add bases nonuniformly at a variable position, where two or more bases are injected according to the prescription dictated by the calculated base frequencies. Their corresponding valves in the synthesizer are opened and closed for a specific number of times, each time allowing a fixed amount of phosphoramidite into the column. Coupling at each position was achieved with a total of ten such pulses. As such, opening a valve once raised the percentage of the corresponding base by 10%. For the example in Table 1, at position X_1, the A, C, G valves were each opened twice so that they are present in 20% of the final sequences, while the T valve was opened four times in accord with this base having 40%

probability. The phosphoramidite added first may have a greater chance of coupling to the growing oligonucleotide chain, and hence may be slightly overrepresented in the end. The synthesis of the simple oligonucleotide probes whether this would pose a major obstacle.

The resulting sixty-four sequences correspond to twenty distinct masses due to mass degeneracy. The synthesized oligos were purified by reverse phase HPLC on a C-8 column, run over an ion exchange column to exchange buffer to $(NH_4)_2SO_4$, lyophilized, and finally subjected to mass analysis by electrospray ionization (ESI) on Micromass Quattro II using 1:1 mix of acetonitrile:water as mobile phase. The oligos appeared as multiply charged species under the experimental condition with the minimum of charge two. For some adjacent peaks, the difference in mass to charge ratio (*m/z*) was less than one atomic mass unit. The complexity of the spectrum due to this near mass degeneracy as well as the isotopic abundances, made it difficult to resolve closely lying peaks. The peaks that were too close to be resolved separately, were considered as a single peak at the lower molecular weight with the combined intensity of both peaks. The intensities of the observed mass peaks were normalized to one to allow ready comparison with the predicted values. The observed peak intensities were then compared to the theoretical intensities. While there are many sources of uncertainties that could influence the ultimate sequence distribution, we wanted to estimate the sensitivities of the peak intensities to the uncertainties in the precise amounts of the phosphoramidites. To this end, we calculated the error in the expected peak intensities by assuming a 1% error in the probabilities of the bases at degenerate positions and propagating the error to the final

sequences. The expected peak intensities were further corrected to account for the molecular weight differences among phosphoramidites.

We calculated the individual base probabilities at each of the three degenerate positions from the observed peak intensities. As the intensity of the peak at mass m, I_m, is the sum of the probabilities of all sequences consistent with that mass, we have

$$I(m) = \sum_{s=1}^{64} \left(\prod_{i=1}^{3} P_{i,s(i)} \right) \delta \left(m, m_A + m_C + m_G + m_T + \sum_{k=1}^{3} m_{k,s(k)} \right) \qquad (7)$$

where s is a specific triplet sequence, $s(i)$ is the base at position i, $P_{i,s(i)}$ is the probability of the base $s(i)$ at position i, $m_{i,s(i)}$ is the mass of the base $s(i)$ at position i. The Kronecker delta $\delta(x, y)$ is 1 if x equals y, or 0 otherwise. Then the individual nucleotide probabilities are calculated via linear least squares fit subject to the constraints

$$P_{i,A} + P_{i,C} + P_{i,G} + P_{i,T} = 1 \qquad (8)$$

for i = 1, 2, and 3. The fitted base probabilities are as follows:

	A	C	G	T
X_1	18.7%	19.3%	19.0%	43.0%
X_2	38.4%	29.6%	9.8%	22.2%
X_3	9.5%	19.9%	59.4%	11.2%

Table 2 Fitted nucleotide base probabilities

Finally, the amino acid probabilities computed from the input and calculated nucleotide distributions plotted in Fig. 3 show that the desired amino acid profile is accurately mimicked by the designed library.

Given that no measures were taken to avoid the systematic error coming from sequential addition of phosphoramidites, the theoretical mass spectroscopy profile is

remarkably well reproduced experimentally. Even more remarkable is the consistency between the programmed and resulting nucleotide probabilities and the corresponding amino acid profiles. This study demonstrates that the synthesis of custom degenerate oligos can be fully automated by simple modifications to the software controlling standard oligonucleotide synthesizers. Furthermore, computation of the nucleotide probabilities and amino acid profile demonstrates that a peptide library with an arbitrary profile can be easily designed with high precision by using this approach. The gap between computational and experimental protein engineering may be bridged with an automated synthesis of degenerate oligos, which would help achieve routine construction of computationally designed polypeptide library.

4. Summary

Facile construction of large diversity libraries has led to continuous discovery of novel protein molecules with unexpected chemical properties. Equally impressive advances have also been made in computational protein engineering whereby polypeptide sequences that embody desired functions are beginning to be predicted in silico. No doubt further improvements will allow ever more complex protein systems to be examined computationally, leading to rapid identification of interesting de novo protein molecules.

This paper discusses two broad categories of computational protein engineering, i.e. direct protein design and probabilistic protein design. Whereas the former seeks to produce a list of protein sequences that satisfy functional and structure requirements, the latter attempts to describe these candidate sequences as a string of probabilities, each describing the likelihood of the amino acids at variable positions. The probabilistic

approach has the advantage of being readily configured to help design a biased peptide library. Given the subtlety of biological function and the uncertainties in the energy functions used to evaluate many competing sequences, the ability to combine a computational approach with a powerful experimental approach may be more productive in the long run. We have also presented how the probabilistic formalism may be applied to an exercisable laboratory procedure. Two further developments were required, one dealing with the reverse-translating the amino acid probabilities to nucleotide frequencies and the other addressing the physical synthesis of degenerate oligonucleotides satisfying the given nucleotide distribution. Together, they can convert a given amino acid profile to a pool of easily realized degenerate DNA sequences with the minimum infusion of incidental stop codons. These oligonucleotides can then be assembled to a functional gene and create a peptide library that is computationally biased in a site specific manner.

Proteomics is beginning to reveal the mysteries of how proteins encoded by a cell's genome interact as a complex system, controlling and modulating cellular activities and ensuring the ultimate survival of the cell and the organism. Many existing models of biochemical processes including cell signaling, response to environment, intracellular biosynthesis, reproduction, and gene expression are based upon those used to describe electric circuits and complex networks (Alm & Arkin, 2003). Just as we must understand how to build switches and transistors before we can build integrated circuits, we must first have a comprehensive understanding of the component protein molecules before we hope to understand and engineer biological systems as a whole. Here we have discussed recent progress in developing general methods for probabilistic protein design. The promise of such methods lies in their ability to accelerate the design of functional

molecules through synergistic collaboration with protein synthesis and library-based experiments, even when our understanding of the determininants of particular activities is less than complete.

5. Acknowledgments

We thank Adam Peritz for his help with the biased synthesis of oligonucleotide libraries. We acknowledge support from the NSF (CHE 99-84752 and DMR 00-79909) and the NIH (GM61267). JGS is a Cottrell Scholar of Research Corporation and an Arnold and Mabel Beckman Foundation Young Investigator.

6. Appendix

This appendix details both the various energy terms that are included as part of the energy constraints during the amino acid profile calculation and the determination of the site-specific amino acid probabilities.

1.1. Energy functions

In this statistical formalism, two energies--conformational energy E_c and environmental energy E_{env}--are considered and used as constraints in maximizing a conformational entropy S_c. The conformational energy E_c is calculated using an atomic based potential, the AMBER force field(Weiner et al., 1984). E_c includes van der Waals interactions, electrostatic interactions with a distant dependent dielectric ($4\varepsilon\ r_{ij}$), and a modified hydrogen bond term (Kono & Doi, 1996). For a particular sequence ($\alpha_1,...,\alpha_N$) where the conformational states of these amino acids are ($r_1(\alpha_1),\ldots,r_N(\alpha_N)$), E_c is:

$$E_c = \sum_i \varepsilon_i(\alpha, r_k(\alpha)) + \sum_{i,j>i} \varepsilon_{i,j}(\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha')) \qquad (9)$$

In the context of protein energy functions, the one-body term $\varepsilon_i(\alpha, r_k(\alpha))$ includes the interaction energies between backbone atoms and those of the amino acid side chains, as well as a reference energy (discussed below) of amino acid. The two-body term $\varepsilon_{i,j}(\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha'))$ is a sum over the inter-atomic interaction energies between two rotamers at two different sites in the structure. Fluctuations in E_c about its mean value due to variation of sequence is assumed to small for large numbers of sequences sharing common energetic properties. We may then write the conformational energy as a function of the site-specific probabilities $w_i(\alpha, r_k(\alpha))$.

$$E_c \approx \overline{E}_c = \sum_{i,\alpha,k} \varepsilon_i(\alpha, r_k(\alpha)) w_i(\alpha, r_k(\alpha)) + \sum_{\substack{i,j>i \\ \alpha,\alpha' \\ k,k'}} \varepsilon_{i,j}(\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha')) w_i(\alpha, r_k(\alpha)) w_j(\alpha', r_{k'}(\alpha')) \qquad (10)$$

Note that this equation results from the same implicit factorization assumption involving the site probabilities used to arrive at the expression for the entropy S_c (see Eq. 1). Other higher order approximations are also possible using more complex expressions for S_c and E_c that would yield the pair probabilities $w_{ij}(\alpha, r_k(\alpha), \alpha', r_k(\alpha'))$.

As another constraint, an environmental energy E_{env} is introduced to account for the hydrophobic effect in a way that is compatible with the statistical theory (Kono & Saven, 2001). This potential takes into account the surface exposure preferences of the amino acids. As for E_c, we can write E_{env} using amino acid probabilities as

$$E_{env} \approx \overline{E}_{env} = \sum_{i,\alpha,k} \varepsilon_{env}(\alpha, r_k(\alpha)) w_i(\alpha, r_k(\alpha)) \qquad (11)$$

where ε_{env} is a local environmental energy defined below. Note that this energy contains no two-body interactions and is dependent only upon the amino acid and rotamer state at each position. Since this effective energy is an approximate quantity derived from structures in the protein databank (PDB), it is usually implemented as a separate constraint rather than combined with the atom-based conformational energy E_c.

6.2. Solvation and hydrophobic energy

An important input to any protein design method is some means of quantifying the hydrophobic effect and other solvation properties. The hydrophobic effect in proteins is manifested in the tendency of apolar (uncharged, aprotic) amino acid side chains to be preferentially located within the interior of globular, soluble proteins. Explicit representation of solvent is impractical for calculations that examine a large number of sequences, and calculating solvent accessible surface areas alone---which often correlate well with hydrophobic tendencies---can be computationally prohibitive when many sequences must be considered. In considering solvation effects in a practical manner consistent with the formalism, we introduced an environmental energy that is a function of the density (ρ) of C_β atoms in the vicinity of each side chain(Kono & Saven, 2001). The position of the C_β atoms are determined by the backbone coordinates and hence invariant in the calculation. As hydrophobic residues tend to be located in buried interior regions of proteins, they are likely to have a higher C_β density than hydrophilic residues which tend to be located at the surface, exposed to solvent. Using 500 different globular proteins of known structure (training set), we derived effective potentials for the amino

acids using a standard equation for "statistical" potentials (Durbin, 1998; Miyazawa & Jernigan, 1985; Saven, 2003)

$$\varepsilon_{env}(\alpha,\rho) = -\beta^{-1}{}_{e} \ln \frac{p(\alpha,\rho)}{p(\alpha)p(\rho)} \qquad (12)$$

where $p(\alpha,\rho)$ is the fraction of times a local C_β density of ρ is observed for amino acid α; $p(\alpha)$ is the fraction of times amino acid α is observed in the training set; and $p(\rho)$ is the fraction of times a local density of ρ is observed, regardless of amino acid type (Kono & Saven, 2001). β_e is an effective inverse temperature. The density ρ is defined as the density of C_β atoms within the "free volume" within a sphere centered about a particular orientation of the side chain. We note that the local density is dependent upon the rotamer state of the amino acid, so $\varepsilon_{env}(\alpha,\rho(r_k(\alpha))) \equiv \varepsilon_{env}(\alpha,r_k(\alpha))$. This C_β density based potential exhibits good correlation with other amino acid hydrophobicity scales(Kono & Saven, 2001). For the sequence probability calculations, E_{env} is constrained to a value calculated using a known sequence consistent with the structure (if one is known), or a value representative of proteins of that particular size or chain length.

6.3. Reference energy

In computational protein design, the goal is to identify sequences whose energy when in the target structure is sufficiently below that of the ensemble of unfolded states. To address this issue about unfolded states, a reference energy $\gamma_{ref}(\alpha)$ for each amino acid is introduced into the energy E_c to mimic the effects of the denatured state (Raha et al.,

2000; Wernisch et al., 2000). The reference energy is calculated as a "free energy" of each amino acid, averaged over multiple rotamer configurations as well as backbone conformation (ϕ and ψ angles). The reference energies of an amino acid α may then be estimated using

$$\gamma_{ref}(\alpha,\beta_{ref}) = -\beta_{ref}^{-1}\ln\left(z_{ref}(\alpha,\beta_{ref})/z_{ref}(G,\beta_{ref})\right) \quad (13)$$

$$z_{ref}(\alpha,\beta_{ref}) = \sum_{\phi,\varphi,k}\exp\left(-\beta_{ref}\varepsilon_{ref}(\phi,\psi,r_k(\alpha))\right) \quad (14)$$

where ε_{ref} is the conformational energy of the model compound N-acetyl-N'-methylamide amino acid α, and the sum is over both rotamer and backbone degrees of freedom. Here $\beta_{ref}=1/k_BT$, where k_B is Boltzmann's constant and T is a temperature appropriate for the conformation sampling of side chain and backbone conformations (e.g., T = 300K). Finally, reference energies are calibrated with respect to that of glycine (G), which has no side chain and therefore is assumed to have zero reference energy. The energy constraint on the sequences involving inter-atomic interactions is then modified to

$$\begin{aligned} E_c \approx \overline{E}_c = & \sum_{i,\alpha,k}(\varepsilon_i(\alpha,r_k(\alpha)) - \gamma_{ref}(\alpha,\beta_{ref}))w_i(\alpha,r_k(\alpha)) \\ & + \sum_{\substack{i,j>i \\ \alpha,\alpha' \\ k,k'}}\varepsilon_{i,j}(\alpha,r_k(\alpha);\alpha',r_{k'}(\alpha'))w_i(\alpha,r_k(\alpha))w_j(\alpha',r_{k'}(\alpha')) \end{aligned} \quad (15)$$

6.4. Rotamer and identity probabilities

The theory maximizes the total conformational entropy S_c, yielding a probability $w_i(\alpha,r(\alpha))$ that a particular amino acid is present at site i with side chain conformation k. The conformational and environmental energies are constrained to take on particular values $E_c^{\,0}$ and $E_{env}^{\,0}$, while the probabilities are constrained to ensure normalization. As desired, additional constraints involving the patterning of amino acid residue identities may also be incorporated. The constrained values of E_c and E_{env} may be systematically varied so as to map out the sequence landscape as functions of these energies. Once the $w_i(\alpha,r(\alpha))$ have been determined, the amino acid probability $w_i(\alpha)$ can be easily obtained by summing over the rotamer state probabilities of amino acid α.

$$w_i(\alpha) = \sum_k w_i(\alpha, r_k(\alpha)) \qquad (16)$$

Using an analogy to statistical thermodynamics, the Lagrange multiplier that arises due to constraining the conformational energy, β_c, may be considered an effective inverse temperature $1/\beta_c = T_c$. Similarly, the "heat capacity" C_v of the system can be computed as :

$$C_v = \frac{\partial E_c}{\partial T_c} = \beta_c^2 \sum_i \left(\left\langle \varepsilon_i^2 \right\rangle - \left\langle \varepsilon_i \right\rangle^2 \right) \qquad (17)$$

$$\left\langle \varepsilon_i \right\rangle = \sum_{\alpha,k} \varepsilon_i^{loc}(\alpha, r_k(\alpha)) w_i(\alpha, r_k(\alpha)) \qquad (18)$$

$$\varepsilon_i^{loc}(\alpha, r_k(\alpha)) = \varepsilon_i(\alpha, r_k(\alpha)) - \gamma_{ref}(\alpha, \beta_c) + \sum_{j,\alpha',k'} \varepsilon_{i,j}(\alpha, r_k(\alpha); \alpha', r_{k'}(\alpha')) w_j(\alpha', r_{k'}(\alpha')) \quad (19)$$

where $\left\langle \varepsilon_i \right\rangle$ is so called a local mean field energy which denotes the average local field about a particular amino acid side *i*. The effective heat capacity C_v provides a quantitative measure of the fluctuations in the sequence-rotamer identities in response to changes in the constraint conditions during a calculation.

7. Figure captions

Fig. 1 (Top) Sequence-conformational entropy S_c of the SH3 domain of c-Crk (PDB code: 1CKA), plotted against effective temperature T_c, shows that entropy decreases monotonically as temperature is lowered. (Bottom) Effective heat capacities per residue C_v for all, buried and exposed residues are plotted against effective temperature. Temperatures are in arbitrary units determined by the molecular potential used, here in units of kcal/mol.

Fig. 2 A few representative amino acid probabilities from the buried region of the SH3 domain. Calculated (filled) and sequene alignment-based (open) profiles. F10, L18, F20 and L26 all have fractional solvent accessible surface areas less than 20 %.

Fig. 3 Probability distributions of amino acids (upper) and nucleotides (lower) for site 54 of a SH3 domain (PDB:1CKA). For the amino acid probabilities, the desired probability distribution is shown by open bars and that encoded by calculated gene library by filled bars. An oligonucleotide library with the frequencies of the nucleotides specified in the lower panel encodes for the site-specific amino acid probabilities in the upper panel.

Fig. 4 ESI mass spectroscopy of degenerate oligonucleotide $ACX_1X_2X_3GT$. "Expected" (black, triangles) are combined probabilities of all sequences resulting in the indicated molecular weight. "Observed" (gray, squares) are measured intensities normalized to one for all fifteen observed peaks. The error bars are based on assumed 1% uncertainty during synthesis.

Fig. 5 Input (filled) and designed (open) amino acid profiles. The input profile is computed from the original nucleotide probabilities at positions 1, 2 and 3. The designed profile is computed from the nucleotide probabilities calculated by fitting the mass spectroscopy peaks (see text for discussion). The amino acids are shown in the single letter notation. Asterisk (*) represents a stop codon.

8. Figures

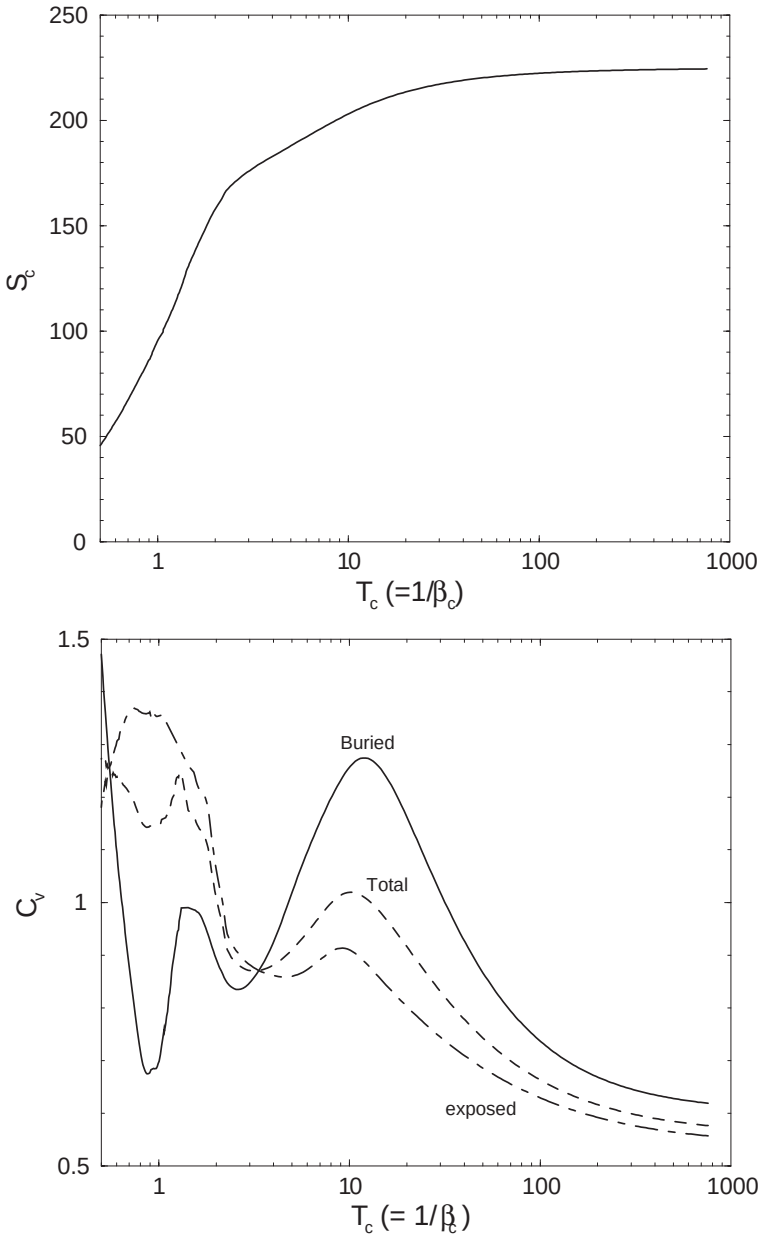

Fig. 1

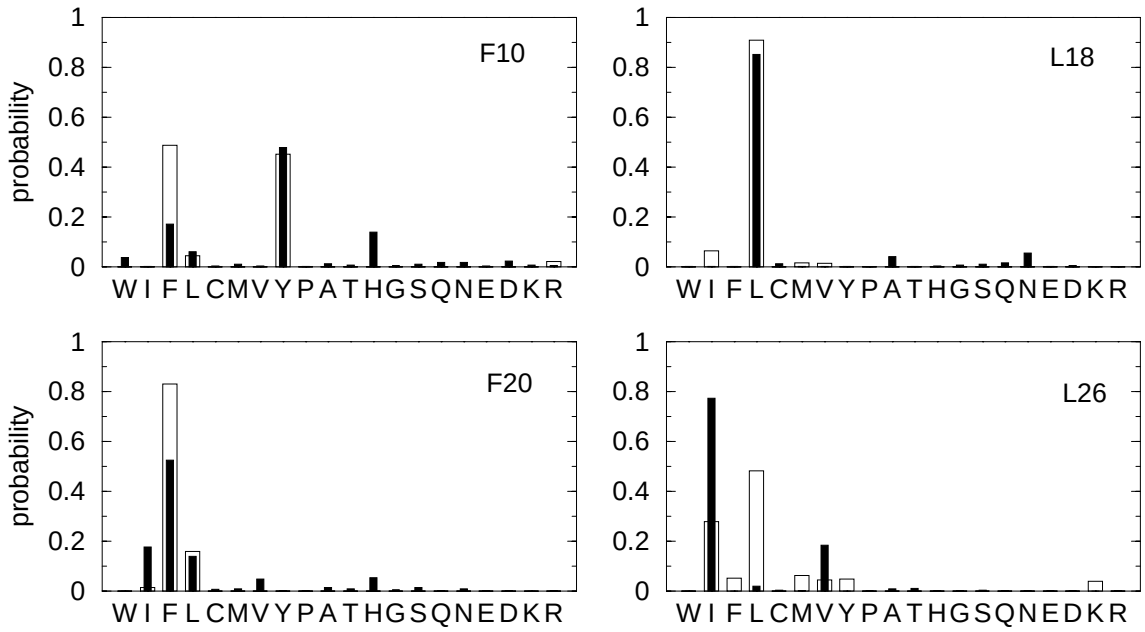
F10
L18
F20
L26
probability
probability
1
0.8
0.6
0.4
0.2
0
W I F L C M V Y P A T H G S Q N E D K R

Fig. 2

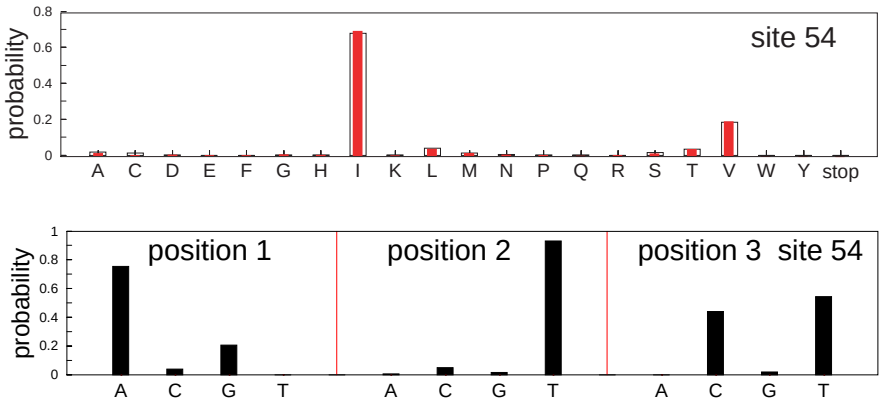
probability
0.8
0.6
0.4
0.2
0
site 54
A C D E F G H I K L M N P Q R S T V W Y stop
probability
1
0.8
0.6
0.4
0.2
0
position 1
position 2
position 3 site 54
A C G T
A C G T
A C G T

Fig. 3

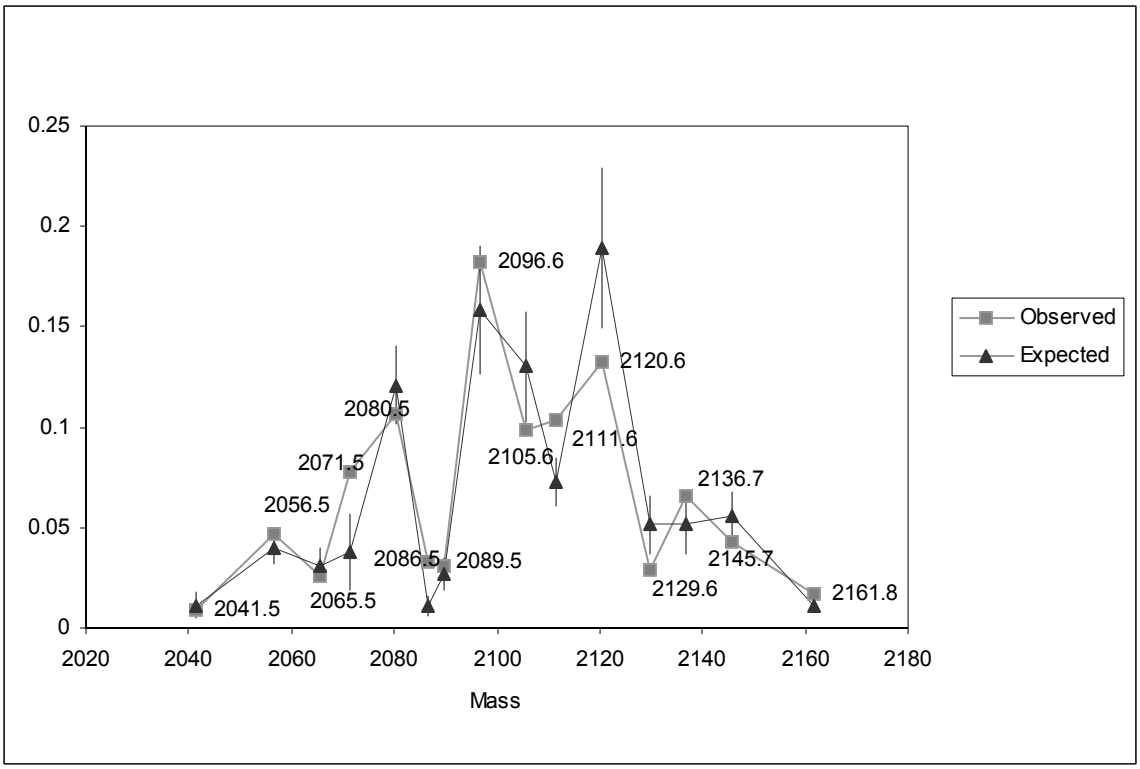
0.25
0.2
0.15
0.1
0.05
0
2020
2040
2060
2080
2100
2120
2140
2160
2180
Mass
Observed
Expected
2041.5
2056.5
2065.5
2071.5
2080.5
2086.5
2089.5
2096.6
2105.6
2111.6
2120.6
2129.6
2136.7
2145.7
2161.8

Fig. 4

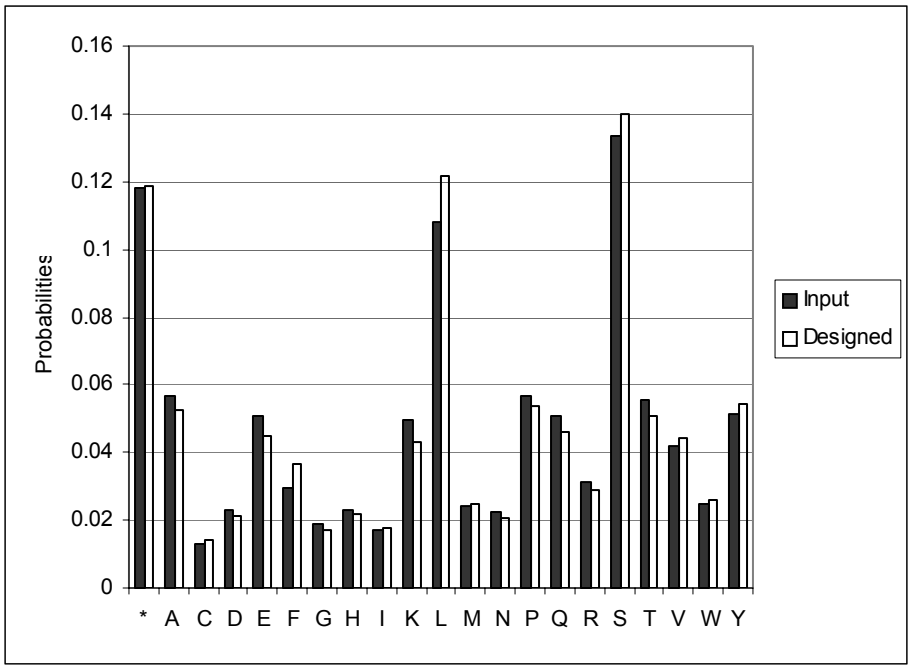
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Probabilities
* A C D E F G H I K L M N P Q R S T V W Y
Input
Designed

Fig. 5

9. References

Alm, E. & Arkin, A. P. (2003). Biological networks. *Curr Opin Struct Biol* 13, 193-202.

Arndt, K. M., Pelletier, J. N., Muller, K. M., Alber, T., Michnick, S. W. & Pluckthun, A. (2000). A heterodimeric coiled-coil peptide pair selected in vivo from a designed library-versus-library ensemble. *Journal of Molecular Biology* 295, 627-639.

Benson, D. E., Wisz, M. S. & Hellinga, H. W. (1998a). The development of new biotechnologies using metalloprotein design. *Current Opinion in Biotechnology* 9, 370-376.

Benson, D. E., Wisz, M. S., Liu, W. & Hellinga, H. W. (1998b). Construction of a novel redox protein by rational design: conversion of a disulfide bridge into a mononuclear iron-sulfur center. *Biochemistry* 37, 7070-6.

Boder, E. T., Midelfort, K. S. & Wittrup, K. D. (2000). Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc Natl Acad Sci U S A* 97, 10701-5.

Boder, E. T. & Wittrup, K. D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* 15, 553-7.

Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America* 98, 14274-14279.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* 4, 187-217.

Brooks, C. L., 3rd. (2002). Protein and peptide folding explored with molecular simulations. *Acc Chem Res* 35, 447-54.

Bryson, J. W., Betz, S. F., Lu, H. S., Suich, D. J., Zhou, H. X., O'Neil, K. T. & DeGrado, W. F. (1995). Protein design: a hierarchic approach. *Science* 270, 935-941.

Bryson, J. W., Desjarlais, J. R., Handel, T. M. & DeGrado, W. F. (1998). From coiled coils to small globular proteins: Design of a native-like three-helix bundle. *Protein Science* 7, 1404-1414.

Calhoun, J. R., Kono, H., Lahr, S., Wang, W., DeGrado, W. F. & Saven, J. G. (2003). Computational design and characterization of a monomeric helical dinuclear metalloprotein. *J Mol Biol* 334, 1101-15.

Case, M. A. & McLendon, G. L. (2000). A virtual library approach to investigate protein folding and internal packing. *Journal of the American Chemical Society* 122, 8089-8090.

Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: Fully automated sequence selection. *Science* 278, 82-87.

DeGrado, W. F., Summa, C. M., Pavone, V., Nastri, F. & Lombardi, A. (1999). De novo design and structural characterization of proteins and metalloproteins. *Annual Review of Biochemistry* 68, 779-819.

Desjarlais, J. R. & Handel, T. M. (1995). De-Novo Design of the Hydrophobic Cores of Proteins. *Protein Science* 4, 2006-2018.

Dunbrack, R. (2002). Rotamer libraries. *Current Opinion in Structural Biology* 12, 431-440.

Dunbrack, R. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain retainer preferences. *Protein Sci.* 6, 1661-1681.

Durbin, R. E., S.;Krogh, A.; Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge Univ. Press, Cambridge.

Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* 255, 178-183.

Finucane, M. D., Tuna, M., Lees, J. H. & Woolfson, D. N. (1999). Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry* 38, 11604-11612.

Giver, L. & Arnold, F. H. (1998). Combinatorial protein design by in vitro recombination. *Current Opinion in Chemical Biology* 2, 335-338.

Go, N. (1983). Theoretical studies of protein folding. *Annu Rev Biophys Bioeng* 12, 183-210.

Gordon, D. B. & Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *Journal of Computational Chemistry* 19, 1505-1514.

Gordon, D. B. & Mayo, S. L. (1999). Branch-and Terminate: a combinatorial optimization algorithm for protein design. *Structure with Folding & Design* 7, 1089-1098.

Gromiha, M. M., Uedaira, H., An, J., Selvaraj, S., Prabakaran, P. & Sarai, A. (2002). ProTherm, Thermodynamic Database for Proteins and Mutants: developments in version 3.0. *Nucleic Acids Res* 30, 301-2.

Hayes, R. J., Bentzien, J., Ary, M. L., Hwang, M. Y., Jacinto, J. M., Vielmetter, J., Kundu, A. & Dahiyat, B. I. (2002). Combining computational and experimental screening for rapid optimization of protein properties. *Proc Natl Acad Sci U S A* 99, 15926-31.

Hellinga, H. W. & Richards, F. M. (1994). Optimal Sequence Selection in Proteins of Known Structure by Simulated Evolution. *Proceedings of the National Academy of Sciences of the United States of America* 91, 5803-5807.

Hoess, R. H. (2001). Protein design and phage display. *Chemical Reviews* 101, 3205-3218.

Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucleic Acids Res* 26, 316-9.

Jensen, L. J. A., K. V.; Svendsen, A.;Kretzschmar, T. (1998). Scoring functions for computational algorithms applicable to the design of spiked oligonucleotides. *Nucleic Acids Res* 26, 697-702.

Jiang, X., Bishop, E. J. & Farid, R. S. (1997). A de novo designed protein with properties that characterize natural hyperthermophilic proteins. *Journal of the American Chemical Society* 119, 838-839.

Jiang, X., Farid, H., Pistor, E. & Farid, R. S. (2000). A new approach to the design of uniquely folded thermally stable proteins. *Protein Science* 9, 403-416.

Johnson, E. C., Lazar, G. A., Desjarlais, J. R. & Handel, T. M. (1999). Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. *Structure with Folding & Design* 7, 967-976.

Jones, D. T. (1994). De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.* 3, 567-574.

Kamtekar, S., Schiffer, J. M., Xiong, H. Y., Babik, J. M. & Hecht, M. H. (1993). Protein Design by Binary Patterning of Polar and Nonpolar Amino-Acids. *Science* 262, 1680-1685.

Keefe, A. D. & Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature* 410, 715-8.

Koehl, P. L., M. (1999a). De novo protein design. I. In search of stability and specificity. *Journal of Molecular Biology* 239, 1161-1181.

Koehl, P. L., M. (1999b). De Novo protein Design. II. Plasticity in sequence space. *Journal of Molecular Biology* 293, 1183-1193.

Kono, H. & Doi, J. (1996). A new method for side-chain conformation prediction using a hopfield network and reproduced rotamers. *Journal of Computational Chemistry* 17, 1667-1683.

Kono, H. & Saven, J. G. (2001). Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *Journal of Molecular Biology* 306, 607-628.

Kraemer-Pecore, C. M., Wollacott, A. M. & Desjarlais, J. R. (2001). Computational protein design. *Current Opinion in Chemical Biology* 5, 690-695.

Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* 97, 10383-10388.

Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. (2002). Thoroughly sampling sequence space:Large-scale protein design of structural ensembles. *Protein Sci.* 11, 2804-2813.

Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* 423, 185-190.

Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. *Journal of Molecular Biology* 307, 429-445.

Luthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with 3-dimensional profiles. *Nature* 356, 83-85.

Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure, and stability of a hyperthermophilic protein variant. *Nature Struct. Biol.* 5, 470-475.

Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *Journal of Molecular Biology* 305, 619-631.

Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J Mol Biol* 196, 641-56.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 218, 534-552.

Moffet, D. A. & Hecht, M. H. (2001). De novo proteins from combinatorial libraries. *Chemical Reviews* 101, 3191-3203.

O'Neil, K. T. & DeGrado, W. F. (1990). A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* 250, 646-51.

Pierce, N. A., Spriet, J. A., Desmet, J. & Mayo, S. L. (2000). Conformational splitting: A more powerful criterion for dead-end elimination. *Journal of Computational Chemistry* 21, 999-1009.

Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* 193, 775-791.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes*. 2nd edit, Cambridge University Press, Cambridge.

Raha, K., Wollacott, A. M., Italia, M. J. & Desjarlais, J. R. (2000). Prediction of amino acid sequence from structure. *Protein Science* 9, 1106-1119.

Rojas, N. R. L., Kamtekar, S., Simons, C. T., Mclean, J. E., Vogel, K. M., Spiro, T. G., Farid, R. S. & Hecht, M. H. (1997). De novo heme proteins from designed combinatorial libraries. *Protein Science* 6, 2512-2524.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science* 229, 834-8.

Roy, S., Helmer, K. J. & Hecht, M. H. (1997a). Detecting native-like properties in combinatorial libraries of de novo proteins. *Folding & Design* 2, 89-92.

Roy, S., Ratnaswamy, G., Boice, J. A., Fairman, R., McLendon, G. & Hecht, M. H. (1997b). A protein designed by binary patterning of polar and nonpolar amino acids

displays native-like properties. *Journal of the American Chemical Society* 119, 5302-5306.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56-68.

Saven, J. G. (2001). Designing Protein Energy Landscapes. *Chem. Rev.* 101, 3113-3130.

Saven, J. G. (2003). Connecting statistical and optimized potentials in protein folding via a generalized foldability criterion. *J. Chem. Phys.* 118, 6133-6136.

Shakhnovich, E. I. & Gutin, A. M. (1993). A new approach to the design of stable proteins. *Protein Engineering* 6, 793-800.

Shea, J. E. & Brooks, C. L., 3rd. (2001). From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu Rev Phys Chem* 52, 499-535.

Shimaoka, M., Shifman, J. M., Jing, H., Takagi, L., Mayo, S. L. & Springer, T. A. (2000). Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nature Structural Biology* 7, 674-678.

Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure with Folding & Design* 7, R105-R109.

Strop, P. & Mayo, S. L. (1999). Rubredoxin variant folds without iron. *Journal of the American Chemical Society* 121, 2341-2345.

Summa, C. M., Lombardi, A., Lewis, M. & DeGrado, V. F. (1999). Tertiary templates for the design of diiron proteins. *Current Opinion in Structural Biology* 9, 500-508.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural

information of biological macromolecules. *Acta Crystallogr D Biol Crystallogr* 54, 1078-84.

Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology* 299, 789-803.

Walsh, S. T. R., Cheng, H., Bryson, J. W., Roder, H. & DeGrado, W. F. (1999). Solution structure and dynamics of a denovo designed three-helix bundle protein. *Proc. Natl. Acad. Sci. USA* 96, 5486-5491.

Wang, W. & Saven, J. G. (2002). Designing gene libraries from protein profiles for combinatorial protein experiments. *Nucleic Acids Res* 30, e120.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *The Journal of the Americal Chemical Society* 106, 765-784.

Wernisch, L., Hery, S. & Wodak, S. J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. *J Mol Biol* 301, 713-36.

Wolf, E. & Kim, P. S. (1999). Combinatorial codons: A computer program to approximate amino acid probabilities with biased nucleotide usage. *Protein Sci.* 8, 680-688.

Xu, G. F., Wang, W. X., Groves, J. T. & Hecht, M. H. (2001). Self-assembled monolayers from a designed combinatorial library of de novo beta-sheet proteins. *Proceedings of the National Academy of Sciences of the United States of America* 98, 3652-3657.

Zhao, H. M. & Arnold, F. H. (1997). Combinatorial protein design: Strategies for screening protein libraries. *Current Opinion in Structural Biology* 7, 480-485.

Zou, J. & Saven, J. G. (2000). Statistical theory of combinatorial libraries of folding proteins: Energetic discrimination of a target structure. *Journal of Molecular Biology* 296, 281-294.

Zou, J. & Saven, J. G. (2002). Biasing Monte Carlo using self-consistently determined local fields to accelerate protein design and sequence sampling.