An Evolution Based Approach to *de novo* Protein Design

Jeffrey R. Brender[1], David Shultis[1], Naureen Aslam Khattak[1], Yang Zhang[1,2]*

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA
[2]Department of Biological Chemistry, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA

*Correspondence should be addressed to
Yang Zhang
Department of Computational Medicine and Bioinformatics, University of Michigan,
100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA,
Phone: (734) 647-1549, Fax: (734) 615-6553,
Email: zhng@umich.edu

Running title: Evolutionary approach to protein design

Summary

EvoDesign is a computational algorithm that allows the rapid creation of new protein sequences that are compatible with specific protein structures. As such, it can be used to optimize protein stability, resculpt the protein surface to eliminate undesired protein-protein interactions, and optimize protein-protein binding. A major distinguishing feature of EvoDesign in comparison to other protein design programs is the use of evolutionary information in the design process to guide the sequence search towards native-like sequences known to adopt structurally similar folds as the target. The observed frequencies of amino acids in specific positions in the structure in the form of structural profiles collected from proteins with similar folds and complexes with similar interfaces can implicitly capture many subtle effects that are essential for correct folding and protein binding interactions. As a result of the inclusion of evolutionary information, the sequences designed by EvoDesign have native-like folding and binding properties not seen by other physics-based design methods. In this chapter, we describe how EvoDesign can be used to redesign proteins with a focus on the computational and experimental procedures that can be used to validate the designs.

Keywords: Protein design, Evolutionary profile, Protein structure modeling, Experimental protein validation, Recombinant expression, Circular dichroism, Nuclear magnetic resonance

1 Introduction

Computational protein design has expanded in recent years from the prediction of the effects of single site mutations to the complete redesign of entire proteins, including the alteration of protein-protein binding affinity and specificity [1-4], enzymatic activity [5,6], and even the creation of new folds [7] and functions [8] that are not seen in nature. On the theoretical side, protein design has been used to find sequence constraints necessary to generate specific folds or functions [9-11]. Through the use of these constraints, fundamental questions in protein evolution have been addressed by distinguishing what is physically possible from what is actually observed in evolution [10,12].

However, full protein redesign beyond the mutation of a few hot spot residues, called *de novo* design, is computationally difficult, which is reflected in the relatively low successful percentage of successful designs. Most algorithms for *de novo* protein design approach the problem as reverse *ab initio* protein folding, evaluating the energy of the sequence according to all-atom physical potentials. Several problems become apparent in the naïve application of this approach: (1) A very large number of sequences must be considered, which limits the force field only to approximate energy terms which can be rapidly calculated; (2) there is a mismatch between the low-resolution models generated in the sequence search and the all-atom physical potentials used for evaluation. To make the design simulation computationally tractable, the possible conformations of the side-chains of the protein are restricted to a limited set of discrete rotamer conformations. The small steric clashes that necessarily result from this approximation force the use of dampened potentials that may miss subtle interactions that exist in the native protein [13,14]; (3) the sequence search is considered only with the protein in isolation, not as the protein actually exists in the cellular context. This causes subtle problems in the real-life application of the designed proteins, particularly with respect to aggregation, as the highly hydrophobic sequences favored by folding energetics generally adopt highly compact sequences *in silico* but tend to aggregate in reality when actually expressed [15].

One approach to handle these challenges is to increase the accuracy of the design process by attempting to model physical reality at a higher resolution. In this spirit, design methodologies have been created that explicitly consider multiple conformations of the folded protein using ensemble techniques for multi-state design [16-18] or that explicitly consider the unfolded state during the design process [18]. Alternatively, other design methodologies have been created that recognize the inherent inaccuracy of the force fields and attempt to diminish the effects of known inaccuracies. One example is the use of soft-core potentials that lessen repulsive interactions, preventing strongly unfavorable interactions that can be alleviated by small backbone motions from overriding the other terms [19]. Another example of this approach is the inclusion of additional terms in the force field that consider factors relevant to real proteins that are missing in the simulation, for example, the explicit consideration of inappropriate hydrophobic surfaces to limit aggregation in the designed sequences [20,18]. The ongoing development of these methods has contributed greatly to the field and has led to some spectacular successes. However, complete *de novo* protein design is still a difficult process with routine application still in the future.

An alternative approach, based on hard-won knowledge from protein fold-recognition and structure prediction [21-24], is to recognize that evolution implicitly encodes information on protein folds and binding interactions that greatly exceeds our ability to describe it through reductionist, physics based methods. This evolution based method approach to protein design differs from the physics based methods in that most energy terms are not dependent on the full-atom representation of each tested

sequence, whose inaccuracy is a significant source of error. Instead, the sequence space search is constrained by the sequence and structural profiles collected from structurally analogous families, assisted by neural network predictions of local structural features, including secondary structure, backbone torsion angle and solvation [25,26].

2. Methods

2.1 EvoDesign: Evolution-based Method to Design Protein Folds and Interactions

The principle of EvoDesign follows the critical lessons learned from threading-based protein structure prediction methods, i.e. to use the reliable 'finger-print' of nature of multiple proteins from the same family in the form of structural profile information to guide the simulation to the sequence search. It first collects a set of proteins with similar folds to the target scaffold structure from the PDB library by the structural alignment program TM-align [27], using a TM-score cutoff value to define structural similarity (Fig. 1) [28]. In the second step, this set of structurally similar folds is used to create a position specific scoring matrix $M(p, a)$ for evaluating potential sequences [29,30].

To create the position specific scoring matrix, a multiple sequence alignment (MSA) is generated first according to the pair-wise structural alignments between the structural analogs identified in the first step and the target structure (Fig.1). An L×20 matrix (where L=length of the protein) is then created according to

$$M(p,a) = \sum_{x=1}^{20} w(p,x) \times B(a,x) \tag{1}$$

where x represents a particular amino acid, $B(a, x)$ is the BLOSUM62 substitution matrix [31] for amino acid x to amino acid a, and $w(p, x)$ is the frequency of the amino acid x appearing at position p in the MSA created by TM-align. The matrix $M(p, a)$ serves as a structural profile to guide the sequences towards native-like sequences known to adopt structurally similar folds as the target (Fig. 1).

While the structural profile as given by the position specific scoring matrix $M(p, a)$ is efficient in guiding the global fold, optimization on the profile alone can result in singularities (i.e. disjointed 'islands') in local sequences. To smoothen these singularities, back propagation neural network predictors are used to estimate the secondary structure (SS), solvent accessibility (SA), and torsion angles (φ/ψ) of the sequence. Unlike other predictors for these properties [32-34], these single-sequence based predictors do not require a computationally expensive PSI-BLAST search, which considerably speeds up prediction at little cost in accuracy [25].

The evolutionary potential in EvoDesign is defined as the maximum score of the optimal alignment path between the decoy and target structure obtained by Needleman-Wunsch dynamic programming, giving the energy function:

$$E_{evolution} = \max \sum \left[M(p,a) + w_1 \Delta SS(p) + w_2 \Delta SA(p) + w_3 \left(\Delta\varphi(p) + \Delta\psi(p) \right) \right] \tag{2}$$

where ΔSS, ΔSA, $\Delta\phi$ and $\Delta\psi$ are the difference in secondary structure, solvent accessibility and torsion angles between the target assignments and the predictions from the decoy sequences. The weighting factors (w_i) are decided by the relative accuracy of the single-sequence based predictions for each term on a training set [25].

A physics based potential can be used to predict potential favorable and unfavorable interactions among side-chains, such as steric interactions, which may be missed by the evolutionary based terms defined

above. While our computational benchmark results indicate the evolution-based energy function alone is sufficient to design protein sequences, adding a physics-based energy term from FoldX [35] improved the atomic packing of the local structures based on both computational structure prediction and experimental structure validations [25]. In this case, a full-atom representation of the sequence is needed which is created by SCWRL [36].

The final force field for single-chain protein design in EvoDesign is given by the weighted Z-scores of the evolution and physics based terms:

$$E = w_4 \frac{E_{evoluation} - \langle E_{evoluation} \rangle}{\delta E_{evoluation}} + w_5 \frac{E_{foldX} - \langle E_{foldX} \rangle}{\delta E_{foldX}} \quad (3)$$

where $\langle \ldots \rangle$ and δ indicate the average and standard deviation of the energy terms.

To actually generate the designed sequences, Monte Carlo searches are performed starting from 10 random sequences that are updated by random residue mutations (Fig.1). Due to the imprecision of the force field, the lowest energy states do not always correspond to the best sequence design. Instead of simply focusing on the lowest energy sequence, the sequences from all 10 runs are pooled and the sequence with the maximum number of neighbors is identified using the SPICKER clustering algorithm [37] where the pair-wise distance between sequences is measured by the sum of the BLOSUM62 substitution scores [38].

The above procedure finds sequences compatible with the target structure. To introduce new or altered functionality into the protein, the affinity of existing protein-protein interfaces can be improved by EvoDesign or new interfaces created though the optimization of non-native complexes created by docking. To modify interfaces, EvoDesign uses a multi-scale approach incorporating a variety of features at different levels of structural resolution (Fig. 2).

Similar to the design of protein folds with EvoDesign, a key feature of the binding potential is the mixture of physics based and evolutionary terms in the energy function [39]. For interface modification, the evolutionary terms are created from the structural alignment of similar interfaces from the non-redundant COTH structural library of dimeric proteins [40] by the IAlign program [41]. A series of interface similarity cutoffs has been used to define three separate interface structure profiles along with different metrics designed to assess the accuracy of the profiles relative to the other terms [39]. The interface profiles scores are then combined with physics based all-atom and residue level docking scores. Finally, sequence based scores based on phamacophore count differences between the native and designed sequences are calculated to complete the multi-scale approach. A random forest method trained to predict the experimental affinity changes ($\Delta\Delta G$) associated with single and multiple mutations at the interface is used for the final interface energy score. This energy score has a correlation to the experimental $\Delta\Delta G$ values equivalent or superior to the best state-of –the-art mutation prediction programs (Pearson's correlation coefficient =0.83 for the cross validated) but is fast enough to calculate the thousands of potential mutations necessary for protein design. The interface energy is then added to the regular EvoDesign scoring potential, using a user defined weighting function to balance fold stability and protein-protein affinity.

2.2 Using the EvoDesign Server Design Program

The EvoDesign program can be used as a server at http://zhanglab.ccmb.med.umich.edu/EvoDesign. The only input to the server is a PDB format file of the target structure, which can be either a full-atomic

or backbone only model. In either case, the backbone of the protein structure should be complete without breaks in the chain. Currently, the server is limited to design of one protein chain only.

There are three user-defined parameters to control the design simulation. The first parameter is the fold-similarity cutoff used for defining the structural profile (Eq. 1). By default, this is set to the relatively high value of a TM-score of 0.7, which is relaxed if less than 10 structural analogues are found in the PDB. This value can be adjusted to a higher or lower value; lower values incorporate more sequence and structural variability in constructing the profile while higher values incorporate less. The usual result is that higher cutoffs penalize deviations from the native sequence more strongly, which may or may not be desirable for the particular application. The second parameter controls whether the FoldX force field is used in the simulation or not. Inclusion of FoldX usually results in only a marginal improvement in the folding when validated by structure prediction (see next section) [25], most likely due to the fact that the side-chains are modeled by a different force field from SCWRL than that used for scoring. Including the FoldX in the simulation requires that the full atomic model of each sequence be constructed, which is the most computationally demanding step in the simulation. For this reason, the FoldX force field is turned off by default. The last parameter does not affect the design simulation but controls whether structure prediction is performed for each of the designed sequences through the creation of I-TASSER models (see **section 2.3.1**).

By default, the EvoDesign server operates without any residue restrictions on the design process. In many cases it desirable to freeze certain residues in the design process, such as those involved in disulfide bond formation or in ligand binding. Taken further, in other cases it is useful to redesign only the surface of the protein while keeping the inner core constant. An option is therefore provided to specify a set of residues (by residue number) which should be kept the same as in the input structure. It is also sometimes desirable to restrict the use of some residues completely or at certain positions. A prime example is cysteine residues on the surface, which can easily be oxidized to form intermolecular disulfide bonds that lead to a loss of activity through aggregation.

The output of the server is 10 sequences in decreasing order of cluster size from the clusters generated by the SPICKER algorithm. For each sequence, the sequence identity to the native sequence is calculated along with the predicted normalized relative error for the secondary structure, solvent accessibility, and torsion angles. Each property is calculated by a high accuracy predictor using PSI-BLAST profiles along with neural network predictors (PSSPred for secondary structure prediction [42], ANGLOR for torsion angle prediction [32], and the method of SOLVE for solvent accessibility [43], respectively). The normalized relative error (*NRE*) is reported for each prediction, which is defined by [25]

$$NRE = \frac{EDS - ETS}{ETS} \tag{4}$$

where *EDS* refers to 'error of designed sequence', i.e. the mismatch between the predicted structure feature from the designed sequence and the target structure. *ETS* refers to 'error of target sequence' which is defined similarly to *EDS* but for target sequence. The *NRE* defined thus counts for the uncertainty from the structure feature predictors. Finally, I-TASSER models of each of the designed sequences are provided if user selects the third option on I-TASSER modeling. The I-TASSER models represent a partial validation of the success of the design simulation as described below.

2.3 Computational Validation of Protein Designs

No computational design method is perfect, and validation remains an essential part of the design processes. Validating experimentally that the designed protein sequence successfully folds to the desired structure requires both successfully expressing the protein and successfully determining the structure. A full structure determination at the atomic level through either NMR spectroscopy or X-ray crystallography is a time-consuming and difficult task. Even simpler, less precise experimental methods for determining protein structure, such as comparing the secondary structure of the native and designed proteins through circular dichroism CD (see **section 2.4.7**) and recognition of the presence of folded tertiary structure through 1D NMR (see **section 2.4.8**), still require that the protein be successfully expressed. Compared to computational techniques, protein expression is relatively expensive, limited in throughput, and in some cases may be challenging to achieve. Before expression, it is therefore desirable to know which designed sequences are most likely to fold to the target structure. The first step is to visually confirm that the design sequences are compatible with the structure. Specifically, it is a good idea to look for buried charges without salt-bridges and buried side-chains without hydrogen bonding partners before proceeding. The EvoDesign program uses a fixed backbone approximation in its calculations. High energies from van der Waals clashes can usually be relieved by small changes in the backbone [44,45]. However, buried charges and missing hydrogen bonds are much harder to compensate for by small structural movements. Since, even one missed hydrogen bond or buried charge is enough to completely destabilize a structure, any designs possessing these features should be eliminated from consideration.

It is, however, not possible to tell reliably if a protein will fold correctly by simple visual analysis. Accurate structure prediction of designed sequences is therefore central to the EvoDesign methodology, as it allows a much larger number and variety of sequences to be tested for correct folding than can be experimentally checked. EvoDesign currently employs I-TASSER, which is a hierarchical approach to protein structure modeling that constructs protein 3D models by reassembling continuous fragments excised from the multiple threading templates [46-48,43]. I-TASSER has been extensively tested in both benchmarking [47,49,46] and blind tests [50-53]. In particular, the community-wide CASP (Critical Assessment of protein Structure Prediction) experiment is designed to benchmark the state-of-the-art of protein structure predictions every two years since 1994 [54-56]. I-TASSER was tested (as "Zhang-server") in the 7-11th CASP competitions in 2006-2015. Fig. 3 shows the histogram of the Z-score of the GDT-score, which measures the significance of the model predictions by each group of automated structure predictors compared to the average performance, in the latest 11th CASP competition. The data shows the advantage of the I-TASSER in comparison to other state of the art protein structure prediction methods, provided that the protein is already known to fold to a specific structure.

2.3.1 Estimating Structural Fidelity and Foldability of Designed Sequences Using I-TASSER

The I-TASSER based structure prediction of designed sequences in EvoDesign seeks to answer two related but distinct questions. First, does the designed sequence fold to any structure at all or is it only partially or completely unfolded when expressed? Second, given that the protein folds, does it fold to the correct structure? If a designed sequence is known to fold, there is considerable evidence from the benchmark and blind tests described above that I-TASSER could, with some confidence, tell if it will fold to its target structure. However, the ability of template based protein structure programs to determine whether or not a given sequence can fold correctly to any structure at all has been tested much less extensively (see **Note 1**).

In an early test, I-TASSER was shown to cleanly distinguish native sequences from random sequences with similar sequence identity and secondary structural propensity [38]. For a more stringent benchmark test, we recently tested 16 successfully designed sequences that are known to match their target structure and 29 unsuccessful sequences that were known to either fold to a different structure or were unable to fold at all in literature [25]. As shown in Figure 4, I-TASSER successfully captured the deviation of the structures of the designed sequences from the target structure. Furthermore, the confidence level (C-score) [57] of the I-TASSER prediction is roughly correlated with the chance of success of the design: a C-score below -1.5 indicates an almost certain failure and a C-score above 0 indicates a very strong possibility of success. I-TASSER prediction on designed sequences can therefore allow a winnowing out of poorly designed sequences without resorting to the lengthy procedure of expressing and experimentally determining the structures of designed proteins at each step.

2.4 Experimental Validation of Designed Sequences

True validation of the designed protein requires that protein be characterized experimentally for structurally fidelity and activity. The processes listed below have been employed in the EvoDesign studies [25,58], aiming to ensure that the designed proteins are thermodynamically stable, soluble, and adopt the desired fold. In all cases, the same tests should be performed with the wild type protein as well as a control.

2.4.1 Expression and Purification of Designed Proteins

Before a protein can be characterized experimentally, the pure protein must be generated in sufficient quantities for the experiments. This is done through a process called recombinant expression, which involves incorporating the DNA sequence of the designed protein into the genome of another organism and using that organism's protein production process to generate the target protein. Since there are many variations on the technique and the specifics of the process can vary with the protein being produced, a comprehensive description of the technique is not given here. Instead, key considerations are outlined in a basic manner for those unfamiliar with process. For further, more depth treatment readers are encouraged to consult several excellent reviews on this topic [59].

2.4.2 Choice of Host Cell

The first decision that must be made in setting up a recombinant protein expression system is the choice of the host cell whose protein synthesis machinery will produce the target protein. This choice is one of the most critical ones as the choice of the expression organism defines the scope of the project, the reagents and equipment needed, and the final outcome of the expression process [59]. Each protein expression has advantages and disadvantages. In most cases, bacterial expression systems are favored as they are low cost, easy to manipulate genetically, scale easily from small- to large-scale expression, and can easily incorporate isotopic labels for NMR studies. The main disadvantage of bacterial expression is that eukaryotic post-translational modifications such as glycosylation and phosphorylation are not performed. In the case that these post-translational modifications are essential, a eukaryotic host cell such as yeast or insect cells must usually be used and the process becomes considerably more complex.

Disulfide bond formation is also more difficult in bacteria, although this may be overcome in most cases by selecting a bacterial strain such as the Orgami cell line that have mutations in the thioredoxin reductase and glutathione reductase genes, which creates an oxidative environment that greatly enhances disulfide bond formation in the cytoplasm [60]. Expression can vary greatly for different bacterial strains. For this reason, different specialized strains of bacteria have been created to optimize the

expression of recombinant proteins. Most specialized bacterial strains for expression start with the BL21 genetic background which is deficient in the Ion and ompT proteases that can lead to improper cleavage of the protein product. Other bacterial strains attempt to minimize the difference in codon usage between the natural codon usage of the bacteria and the codon usage required to express the protein.

Recombinant expression of proteins can lead to a high demand for specific tRNAs that are normally produced in only small amounts by the bacteria. Depletion of these low abundance tRNAs can cause translation to stall on the ribosome, leading to premature release from the ribosome and the generation of truncated versions of the protein [61]. From our studies [25,62,58], we recommend for routine use the Rosetta 2 bacterial cell line which combines the protease mutations found in the BL21 strain along with additional modifications that allow the bacteria to generate low abundance tRNAs more efficiently and mutations that allow tunable expression through mutations in the Lac permease gene (see below). However, alternate strains may be considered in certain situations such as the Rosetta-gami strain, which adds the disulfide-bond promoting mutations of the Orgami strain to the Rosetta background.

2.4.3 Selection of Expression Vector

Once the host cell is selected, the next step is to create the vector that introduces the foreign DNA into host cell. This is typically a bacterial plasmid that contains several elements besides the DNA encoding the target protein. The first element is a gene for antibiotic resistance which provides a growth selection mechanism for discovery; only those bacteria which have incorporated the plasmid into their genome can grow in the presence of the antibiotic. The second is the promoter system, which ties the expression of the target protein to another protein whose expression is essential for the cell and whose expression can be readily induced at a specific time. Triggering expression at a specific time is essential as bacteria continue to grow during incubation and the time at which the protein are lysed determines the overall yield and final purity of the product. If the cell density is too low, the yield of expressed protein is naturally low. On the other hand, too high of cell density can also result in decreased yields and purity from loss of the plasmid from the bacteria [63], metabolism of the antibiotic within the medium, and death of the bacteria from lack of dissolved oxygen [64]. Typically this is done through the use of the Lac operon, in which protein expression can be induced at a specific time period during growth with the lactose analog isopropyl β-D-1-thiogalactopyranoside (IPTG).

2.4.4 Purification of Expressed Protein

Once expressed, the expressed protein still needs to be purified from the other proteins in the bacterial cell. Although this may be accomplished using the sequence of the designed protein without modification using multiple steps of column chromatography, it is easier to fuse the designed sequence to other protein domains to make purification easier. In many cases, the expressed protein is not soluble at the very high concentrations generated during expression. In this situation, the expressed protein accumulates in an insoluble form in the bacteria as particles known as inclusion bodies. The formation of inclusion bodies can make purification easier or more difficult. The inclusion bodies generally contain the expressed protein in highly pure form with only a small amount of the other proteins of the host cell mixed in, a clear advantage for the purification process. On the other hand, proteins within inclusion bodies must be first disaggregated and then refolded with urea, which may prove a difficult process [65]. If the stability of the protein is unknown, such as the case with designed proteins, it is often easier to try to purify already folded, soluble proteins.

To enhance the solubility of proteins during purification, a solubility tag such as the Mocr domain [66] can be fused to the target protein. This domain is usually fused N-terminal to the designed sequence. Since it is localized to the N-terminus, the Mocr domain is therefore synthesized first and folds into its native form before the translation of the designed sequence, stabilizing the designed domain's folding process. Moreover, the high negative charge on the Mocr domain increases the solubility during the purification process by preventing self-association by electrostatic repulsion. Along with the solubility tag, another sequence that specifically binds a particular column can be incorporated to assist purification. A common choice is the His tag, six consecutive histidine residues that strongly bind nickel (Ni) columns. A protease cleavage site is often placed between the Mocr domain with the His tag and the sequence of the designed protein so that the two domains can be separated. The expressed protein with the Mocr/His tag will bind the Ni column; most other bacterial proteins will not. The Mocr/His domain is then cleaved from the target sequence by the addition of a protease specific to the cleavage site and passed through the Ni column again. This time, the target protein does not bind the Ni column but all other nickel binding proteins will remain bound to the column. The end result of this process is a highly pure protein in soluble form.

2.4.5 Confirmation of Protein Solubility

In addition to adopting a stable folded conformation, many proteins must be soluble in water in order to perform their biological function. This requirement constrains the design process, as sequences that are optimized only for stability of the folded conformation may not be optimized for solubility. A key advantage of the EvoDesign method is that the structural profiles implicitly include all the constraints involved in determining the sequences that are compatible with a specific fold, not just those concerned with fold stability. As a result, sequences designed by EvoDesign are significantly more native-like in composition than those designed by physics only methods [25], which tend to overemphasize hydrophobic residues on the surface more than is found in native proteins [20,67,38]. Consequently, aggregation by the coalescence of exposed hydrophobic patches is a common source of failure in physics based design [20].

As aggregation generally makes a protein useless for most applications, the oligomeric state of the protein should be determined before proceeding at the highest concentration used for the other biophysical experiments. Typically, this is around 100 μM for a 100-residue domain. The limiting factor is usually sensitivity of the 1D NMR experiment for tertiary structure estimation and sensitivity of the urea denaturation experiment used for the determination of protein stability (see **Note 2**). An approximate concentration range may be established by measuring the signal to noise ratio at different concentrations of the native protein. The signal of both experiments is actually more sensitive to the total concentration by weight than the molar concentration. The 100 μM value may need to be adjusted upwards or downwards for proteins significantly shorter or longer than 100 residues.

The presence of aggregation is most readily determined quantitatively by dynamic light scattering, which measures the hydrodynamic radius of proteins in solution, or from a correctly calibrated analytical size exclusion column. In the absence of either instrument, aggregation may be measured semi-quantitatively by the absorbance at 400 nm. At this wavelength range, the protein does not absorb light and increases in absorbance are due to Raleigh scattering, which is proportional to the sixth power of the particle radius. A comparison to the corresponding absorbance at 400 nm of the native protein provides a qualitative estimate of the amount of aggregation in the sample (see **Note 3**).

2.4.6 Confirmation of Structural Fidelity

X-ray crystallography remains the gold standard for confirming whether a protein design has the desired structure. However, not all well-folded proteins crystallize and the expense of X-ray crystallography severely restricts the number of designs that can be studied. From a functional perspective, absolute structural fidelity is not necessary in many cases and small changes on the atomic scale are tolerated if the protein is stable, soluble, and functional. To test a larger number of sequences, faster low-resolution biophysical techniques can be used to eliminate obviously badly designed sequences [68,69].

2.4.7 Confirmation of Secondary Structure

Secondary structure is the most basic building block of protein structure. The existence of severely incorrect secondary structure in the designed protein therefore very strongly implicates a failed design. Since each secondary structure element (α-helix, β-sheet, and random coil) has a distinct circular dichroism (CD) spectra, the relative fractions of each in a protein can be estimated from a CD spectra by fitting to a reference set of proteins with known CD spectra and secondary structure [70]. The accuracy of this procedure is typically around ± 5%, with α-helical content determined more precisely than either random coil or beta sheet content. If available, infrared (IR) spectra can also be used in a similar manner to characterize the secondary structure, as it has been shown that IR and CD are largely complementary and a combination of the two techniques gives a more accurate picture of the secondary structure than either technique alone [71].

2.4.8 Confirmation of Existence of Tertiary Structure

The existence of tertiary structure has traditionally been defined in a qualitative way from the appearance of the 1D ^{1}H NMR spectra of the protein. A protein that is poorly folded, without extensive contacts within the protein core, has a distinctive 1D NMR spectra characterized by the lack of highly shielded peaks in the region of the spectra from -1 to 0.5 ppm and poor dispersion of the signal within the amide region (see Fig. 5) [72,73]. While this method is standard in the protein design field [68,69], it is subjective and qualitative. A more objective and quantitative method is to use the autocorrelation of a 1D ^{1}H [74] or unassigned 3D ^{15}N NOESY-HSQC NMR spectrum [75], which have been shown to accurately distinguish folded and unfolded proteins. A comparison of the binding of the dye SYPRO Orange, which binds to exposed hydrophobic surfaces, to the native sequence can provide an additional test for a misfolded protein structure [76].

2.4.9 Confirmation of Fold Stability

The free energy of folding can be measured using chemical denaturation with urea, with denaturation measured by the decrease in secondary structure as determined by CD [25]. As the concentration of urea is increased, the protein unfolds, in most cases by a two-step process without a significant population of partially unfolded intermediates. The first step of determining the stability is to measure the CD signal without denaturant (CD_{folded}), where it is assumed to be completely folded, and at a high concentration of denaturant, where it is assumed to be completely folded ($CD_{unfolded}$). If unfolding is a two-step process, the CD signal as a function of the urea concentration is [77]:

$$CD(urea) = f_{unfolded}(urea)CD_{unfolded} + f_{folded}(urea)CD_{folded} \quad (5)$$

where $f_{folded}(urea)$ and $f_{unfolded}(urea)$ refer to the fractions of folded and unfolded protein respectively, at a given urea concentration. Since the equilibrium constant can be calculated directly from fraction of folded and unfolded proteins, the Gibbs free energy of unfolding can be calculated for each urea concentration [77]:

$$K(urea) = \frac{f_{unfolded}(urea)}{1 - f_{unfolded}(urea)} \tag{6}$$

$$\Delta G(urea) = -RT \ln K(urea) = -RT \ln\left(\frac{f_{unfolded}(urea)}{1 - f_{unfolded}(urea)}\right) \tag{7}$$

The relevant free energy is the free energy of unfolding in the absence of denaturant, which can be obtained by linear extrapolation of the free energy to zero urea concentration.

3 Conclusions

Using an evolution-based approach, we have successfully designed, expressed, and experimentally characterized a number of single domain proteins [25,58]. In the first benchmark test, we used EvoDesign to redesign 87 globular proteins randomly collected from the PISCES server. I-TASSER was then used to test the fidelity of the predicted structure to the target. Although all homologous templates have been excluded from the I-TASSER template library, out of the 87 designed sequences, 80% were predicted to fold to structure with an RMSD of <2.0 Å to the target scaffold, and 42.5% were predicted to fold to an essentially identical structure with an RMSD<1.0 Å. This was a clear difference from designed sequences created using only the FoldX force field, for which only 54% of the predicted structures have an RMSD<2.0 Å to the target structure, and only 31% have an RMSD<1.0 Å.

In a separate test, we redesigned five globular proteins by EvoDesign and used the experimental validation procedures described in **Section 2.4** to confirm the success of the designs. All five proteins were successfully expressed using the expression system in **Section 2.4.3** and were soluble to at least 70 μM. Further, all five designed proteins have secondary structure consistent with the target protein (<12% difference). Three out of the five had a compact tertiary structure confirmed by NMR (**Section 2.4.8**, Figure 5), for an overall success rate of 60%. One of the three, the Phox homology domain of the cytokine-independent survival kinase (CISK-PX), could be crystallized and its structure compared to the native protein [78]. Despite having only 32% sequence identity, the structure of the designed protein showed a very close similarity to the target with a RMSD of 1.54 Å and a TM-score of 0.90 to the target template. The RMSD and TM-score between the I-TASSER model and the X-ray crystal structure of CISK-PX are 1.32 Å and 0.91, respectively. Most of the difference between the two structures was in a loop that is disordered in the original structure.

Finally, we have shown that EvoDesign can be used to create functional complexes for the X-linked inhibitor of apoptosis proteins (XIAP) with improved properties by designing a peptide-protein complex involved in apoptosis inhibition [58]. The XIAP protein inhibits apotosis by binding caspase-9, an activity which is in turn regulated by the second mitochondria-derived activator of caspases (SMAC). The designed XIAP protein by EvoDesign binds SMAC but does not possess affinity for caspase-9. As such, the designed protein can serve as a SMAC sink, altering the normal protein-protein interaction network involved in cell death. The circular dichroism and isothermal calorimetry data showed that the designed XIAP domain was more stable than WT-XIAP and bound the SMAC derived peptide with a K_d of 167± 67 nM, which compares favorably with the 80 ± 25 nM K_d found for WT-XIAP. Interestingly, a designed version of XIAP with native interface residues actually showed worse binding (K_d of 352±79 nM) and stability than the fully designed sequence, highlighting the efficiency of evolution-based full protein design.

Notes

1. The distinction between these two questions becomes clear when the nature of the benchmarks protein structure prediction is tested against is considered. Due to the experimental requirements of structure determination, the benchmark test largely consists of proteins that can be successfully expressed, successfully purified, and are stable for a prolonged period of time at high concentration. In addition, the protein also must be crystallized in the case of X-ray structures, which is a rather severe restriction for proteins with large unfolded regions as the disordered regions have poor crystal contacts which interferes with the crystallization process [79]. Even if the protein can be crystallized, the disordered regions will have poor electron density and will therefore not be resolved in the structure. Similarly, the structure of unfolded proteins is difficult to determine by NMR due to the lack of long-range NOE constraints and poor chemical shift dispersion [80]. These experimental constraints suggest that though the PDB library is largely complete with respect to the possible universe of monomeric folded domains [81,82], it is still biased towards compact folded structures, as proteins that are intrinsically unstable or unfolded are difficult to observe. The PDB library should therefore not be considered as completely representative of the conformational ensembles, folded or not, that all protein sequences can adopt.

2. The signal to noise ratio in an NMR experiment depends on a number of factors including the field strength of the NMR spectrometer (higher magnetic fields give higher resolution spectra and hence higher signal to noise ratios), the size of the protein (larger proteins give rise to broader signals), and other factors such as conformational exchange (transitions between conformations under certain timescales give rise broader signals). The signal to noise ratio in a CD spectrum also depends on a variety of factors, including the transparency of the buffer in the far UV region of the spectrum (180-260 nm), the pathlength of the cuvette, and the age of the xenon lamp used to acquire the spectrum. Of these factors, the transparency of the buffer usually has the most impact. A buffer strongly absorbing in the UV serves as an inner filter that attenuates the incoming light reaching the protein. Phosphate buffers are optimal for CD due to their transparency in the far UV region of the spectrum, although Tris buffers are nearly as good. Chloride ions absorb in this region and the proteins in NaCl solutions should be dialyzed against an equivalent of concentration of NaF. Finally, many additives used to stabilize proteins, such as glycerol, arginine, and Triton-X, absorb strongly in the UV and are incompatible with CD spectroscopy for this reason.

3. An alternative wavelength can be used if the protein possesses a cofactor such as FAD or FMN that absorbs in the visible light range.

Acknowledgement

The project is supported in part by the National Institute of General Medical Sciences (GM083107).

References

1. Karanicolas J, Kuhlman B (2009) Computational design of affinity and specificity at protein-protein interfaces. Current opinion in structural biology 19 (4):458-463.
2. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D (2004) Computational redesign of protein-protein interaction specificity. Nature structural & molecular biology 11 (4):371-379.

3. Shifman JM, Mayo SL (2003) Exploring the origins of binding specificity through the computational redesign of calmodulin. P Natl Acad Sci USA 100 (23):13274-13279.
4. Lopes A, Busch MSA, Simonson T (2010) Computational Design of Protein-Ligand Binding: Modifying the Specificity of Asparaginyl-tRNA Synthetase. J Comput Chem 31 (6):1273-1286.
5. Procko E, Hedman R, Hamilton K, Seetharaman J, Fleishman SJ, Su M, Aramini J, Kornhaber G, Hunt JF, Tong L, Montelione GT, Baker D (2013) Computational design of a protein-based enzyme inhibitor. Journal of molecular biology 425 (18):3563-3575.
6. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retro-aldol enzymes. Science 319 (5868):1387-1391.
7. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302 (5649):1364-1368.
8. Siegel JB, Smith AL, Poust S, Wargacki AJ, Bar-Even A, Louw C, Shen BW, Eiben CB, Tran HM, Noor E, Gallaher JL, Bale J, Yoshikuni Y, Gelb MH, Keasling JD, Stoddard BL, Lidstrom ME, Baker D (2015) Computational protein design enables a novel one-carbon assimilation pathway. Proc Natl Acad Sci U S A 112 (12):3704-3709.
9. Ollikainen N, Kortemme T (2013) Computational protein design quantifies structural constraints on amino acid covariation. Plos Comput Biol 9 (11):e1003313.
10. Fromer M, Linial M (2010) Exposing the co-adaptive potential of protein-protein interfaces through computational sequence design. Bioinformatics 26 (18):2266-2272.
11. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R (2012) The spatial architecture of protein function and adaptation. Nature 491 (7422):138-U163.
12. Schaefer C, Schlessinger A, Rost B (2010) Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. Bioinformatics 26 (5):625-631.
13. Ollikainen N, Smith CA, Fraser JS, Kortemme T (2013) Flexible Backbone Sampling Methods to Model and Design Protein Alternative Conformations. Methods in Protein Design 523:61-85.
14. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 79 (3):830-838.
15. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 424 (6950):805-808.
16. Smith CA, Kortemme T (2011) Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. Plos One 6 (7).
17. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS (1998) High-resolution protein design with backbone freedom. Science 282 (5393):1462-1467.
18. Pokala N, Handel TM (2005) Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J Mol Biol 347 (1):203-227.
19. Li Z, Yang Y, Zhan J, Dai L, Zhou Y (2013) Energy functions in de novo protein design: current challenges and future prospects. Annual review of biophysics 42:315-335.
20. Jacak R, Leaver-Fay A, Kuhlman B (2012) Computational protein design with explicit consideration of surface hydrophobic patches. Proteins 80 (3):825-838.
21. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253 (5016):164-170.
22. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21 (7):951-960.
23. Wu S, Zhang Y (2008) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72 (2):547-556.
24. Zhang Y (2008) Progress and challenges in protein structure prediction. Curr Opin Struct Biol 18 (3):342-348.
25. Mitra P, Shultis D, Brender JR, Czajka J, Marsh D, Gray F, Cierpicki T, Zhang Y (2013) An Evolution-Based Approach to De Novo Protein Design and Case Study on Mycobacterium tuberculosis. PLoS computational biology 9 (10):e1003298.

26. Mitra P, Shultis D, Zhang Y (2013) EvoDesign: de novo protein design based on structural and evolutionary profiles. Nucleic Acids Res 41 (W1):W273-W280.
27. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33 (7):2302-2309.
28. Xu J, Zhang Y (2010) How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 26 (7):889-895.
29. Gribskov M, Homyak M, Edenfield J, Eisenberg D (1988) Profile Scanning for 3-Dimensional Structural Patterns in Protein Sequences. Comput Appl Biosci 4 (1):61-66.
30. Gribskov M, Mclachlan AD, Eisenberg D (1987) Profile Analysis - Detection of Distantly Related Proteins. P Natl Acad Sci USA 84 (13):4355-4358.
31. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America 89 (22):10915-10919.
32. Wu ST, Zhang Y (2008) ANGLOR: A Composite Machine-Learning Algorithm for Protein Backbone Torsion Angle Prediction. Plos One 3 (10).
33. Chen HL, Zhou HX (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 33 (10):3193-3199.
34. Faraggi E, Zhang T, Yang YD, Kurgan L, Zhou YQ (2012) SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J Comput Chem 33 (3):259-267.
35. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. Nucleic acids research 33 (Web Server issue):W382-388.
36. Krivov GG, Shapovalov MV, Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77 (4):778-795.
37. Zhang Y, Skolnick J (2004) SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 25 (6):865-871.
38. Bazzoli A, Tettamanzi AGB, Zhang Y (2011) Computational Protein Design and Large-Scale Assessment by I-TASSER Structure Assembly Simulations. J Mol Biol 407 (5):764-776.
39. Brender JR, Zhang Y (2015) Recognizing Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. PLoS computational biology:in press.
40. Mukherjee S, Zhang Y (2011) Protein-protein complex structure predictions by multimeric threading and template recombination. Structure 19 (7):955-966.
41. Gao M, Skolnick J (2010) iAlign: a method for the structural comparison of protein-protein interfaces. Bioinformatics 26 (18):2259-2265.
42. Zhang Y (2012). http://zhanglab.ccmb.med.umich.edu/PSSpred.
43. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. Nature Methods 12 (1):7-8.
44. Davis IW, Arendall WB, Richardson DC, Richardson JS (2006) The backrub motion: How protein backbone shrugs when a sidechain dances. Structure 14 (2):265-274.
45. Smith CA, Kortemme T (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol 380 (4):742-756.
46. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5 (4):725-738.
47. Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 5:17.
48. Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69 (S8):108-117.
49. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.
50. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins 69 Suppl 8:38-56.
51. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A (2009) Evaluation of template-based models in CASP8 with standard measures. Proteins 77 Suppl 9:18-28.

52. Montelione GT (2012) Template based modeling assessment in CASP10. Paper presented at the 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Gaeta, Italy, December 9-12, 2012
53. Lee BK (2012) Template free modeling assessment in CASP10. Paper presented at the 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, Gaeta, Italy,
54. Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. Proteins 23 (3):ii-v.
55. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A (2009) Critical assessment of methods of protein structure prediction-Round VIII. Proteins-Structure Function and Bioinformatics 77:1-4.
56. Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Current Opinion in Structural Biology 15 (3):285-289.
57. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. Bmc Bioinformatics 9.
58. Shultis D, Mitra P, Aslam N, Gray F, Piper C, Chinnaswamy K, Stuckey J, Cierpicki T, Wang S, Lei M, Zhang Y (2015) Redesigning the fold and binding specificity of BIR3 domain of X-linked inhibitor of apoptosis proteins using evolutionary profiles.submitted.
59. Rosano GL, Ceccarelli EA (2014) Recombinant protein expression in Escherichia coli: advances and challenges. Front Microbiol 5:172.
60. Prinz WA, Aslund F, Holmgren A, Beckwith J (1997) The role of the thioredoxin and glutaredoxin pathways in reducing protein disulfide bonds in the Escherichia coli cytoplasm. J Biol Chem 272 (25):15661-15667.
61. Buchan JR, Stansfield I (2007) Halting a cellular production line: responses to ribosomal pausing during translation. Biol Cell 99 (9):475-487.
62. Shultis D, Czajka J, Marsh D, Gray F, Brender JR, Mitra P, Cierpicki T, Zhang Y Structural validation of computational protein designed through evolutionary methods. In preparation.
63. Baneyx F (1999) Recombinant protein expression in Escherichia coli. Curr Opin Biotechnol 10 (5):411-421.
64. Jana S, Deb JK (2005) Strategies for efficient production of heterologous proteins in Escherichia coli. Appl Microbiol Biotechnol 67 (3):289-298.
65. Burgess RR (2009) Refolding solubilized inclusion body proteins. Methods Enzymol 463:259-282.
66. DelProposto J, Majmudar CY, Smith JL, Brown WC (2009) Mocr: A novel fusion tag for enhancing solubility that is compatible with structural biology applications. Protein Expres Purif 63 (1):40-49.
67. Dantas G, Kuhlman B, Callender D, Wong M, Baker D (2003) A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. J Mol Biol 332 (2):449-460.
68. Koga N, Tatsumi-Koga R, Liu GH, Xiao R, Acton TB, Montelione GT, Baker D (2012) Principles for designing ideal protein structures. Nature 491 (7423):222-+.
69. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R (2005) Evolutionary information for specifying a protein fold. Nature 437 (7058):512-518.
70. Sreerama N, Woody RW (2000) Analysis of protein CD spectra: Comparison of CONTIN, SELCON3, and CDSSTR methods in CDPro software. Biophysical Journal 78 (1):334a-334a.
71. Oberg KA, Ruysschaert JM, Goormaghtigh E (2004) The optimization of protein secondary structure determination with infrared and circular dichroism spectra. Eur J Biochem 271 (14):2937-2948.
72. Rehm T, Huber R, Holak TA (2002) Application of NMR in structural proteomics: Screening for proteins amenable to structural analysis. Structure 10 (12):1613-1618.
73. Scheich C, Leitner D, Sievert V, Leidert M, Schlegel B, Simon B, Letunic I, Bussow K, Diehl A (2004) Fast identification of folded human protein domains expressed in E. coli suitable for structural analysis. Bmc Struct Biol 4:4.
74. Hoffmann B, Eichmuller C, Steinhauser O, Konrat R (2005) Rapid assessment of protein structural stability and fold validation via NMR. Nuclear Magnetic Resonance of Biological Macromolecules, Part C 394:142-+.
75. Schedlbauer A, Coudevylle N, Auer R, Kloiber K, Tollinger M, Konrat R (2009) Autocorrelation Analysis of NOESY Data Provides Residue Compactness for Folded and Unfolded Proteins. Journal of the American Chemical Society 131 (17):6038-+.
76. Niesen FH, Berglund H, Vedadi M (2007) The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat Protoc 2 (9):2212-2221.

77. Pace CN, Scholtz JM (1997) Measuring the conformational stability of a protein. In: Creighton TE (ed) Protein Structure: A Practical Approach. Oxford University Press, New York, pp 299–321.
78. Shultis D, Dodge G, Zhang Y (2015) Crystal structure of designed PX domain from cytokine-independent survival kinase and implications on evolution-based protein engineering.submitted.
79. Price WN, 2nd, Chen Y, Handelman SK, Neely H, Manor P, Karlin R, Nair R, Liu J, Baran M, Everett J, Tong SN, Forouhar F, Swaminathan SS, Acton T, Xiao R, Luft JR, Lauricella A, DeTitta GT, Rost B, Montelione GT, Hunt JF (2009) Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. Nat Biotechnol 27 (1):51-57.
80. O'Hare B, Benesi AJ, Showalter SA (2009) Incorporating 1H chemical shift determination into 13C-direct detected spectroscopy of intrinsically disordered proteins in solution. J Magn Reson 200 (2):354-358.
81. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J (2006) On the origin and highly likely completeness of single-domain protein structures. P Natl Acad Sci USA 103 (8):2605-2610.
82. Brylinski M, Gao M, Skolnick J (2011) Why not consider a spherical protein? Implications of backbone hydrogen bonding for protein structure and function. Phys Chem Chem Phys 13 (38):17044-17055.

Figures

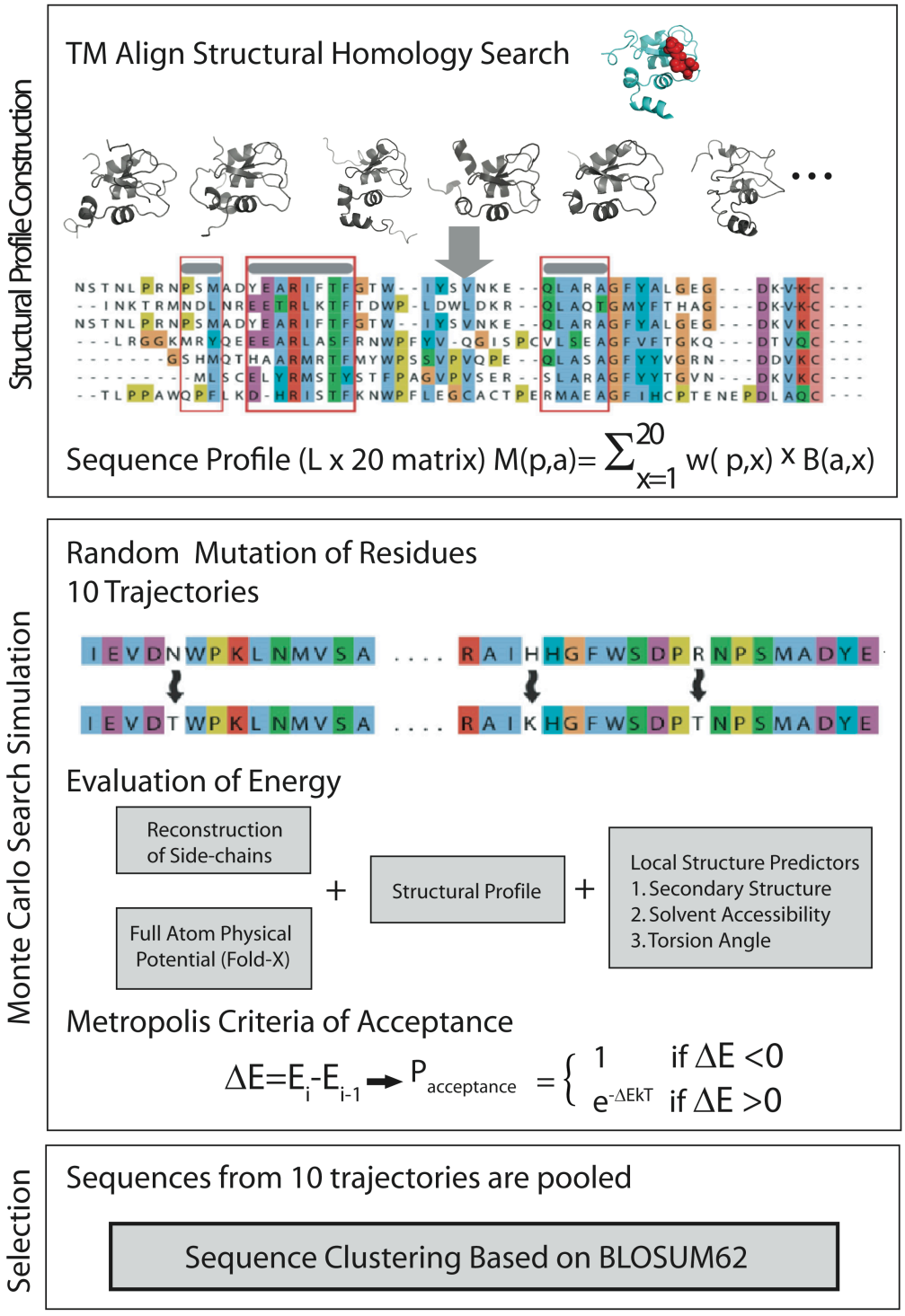

Figure 1. Overview of the EvoDesign method showing the construction of the structural profile, the Monte Carlo search in sequence space, and the final selection of the sequences by sequence clustering.

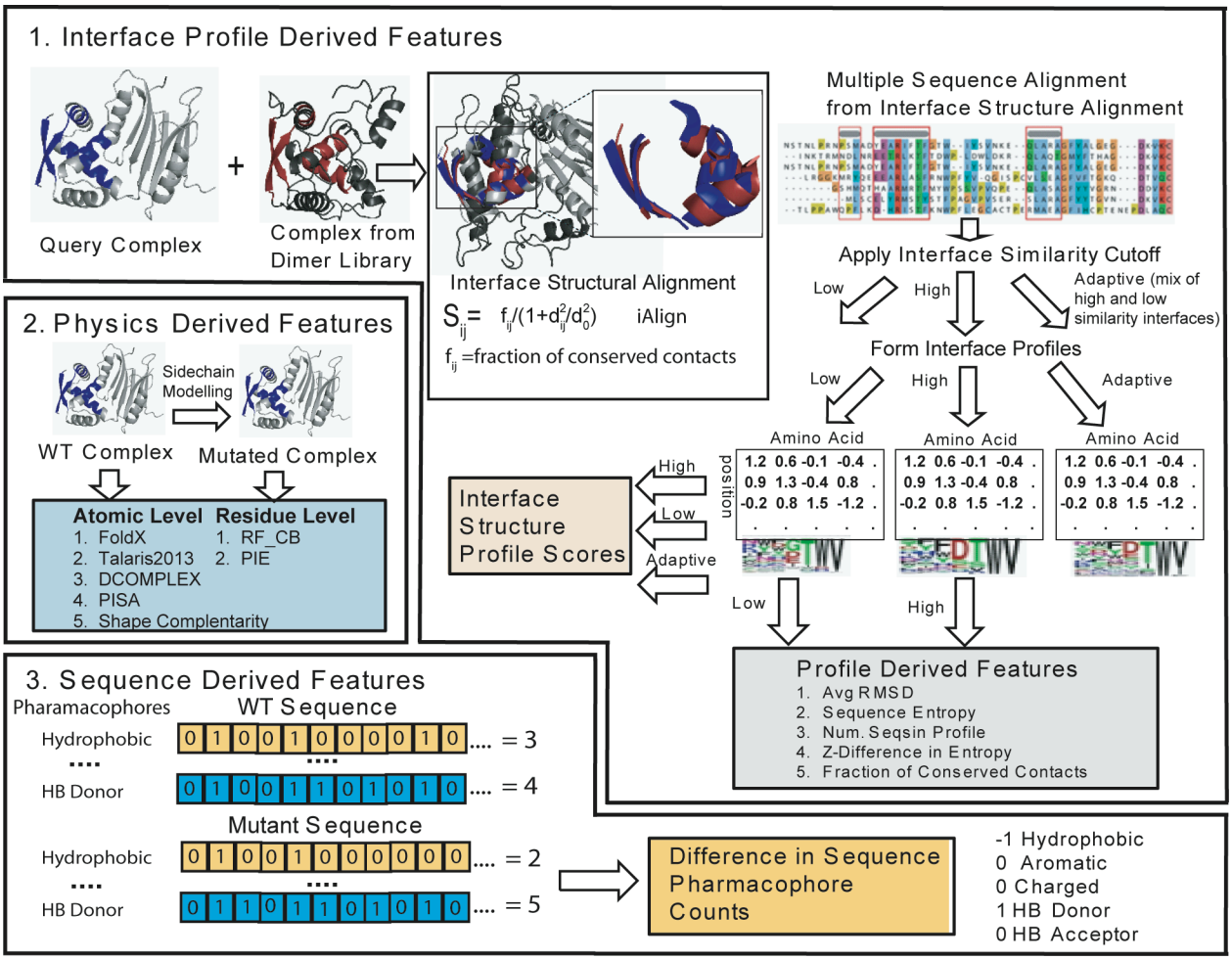

Figure 2. Multi-scale approach to predicting protein binding affinity using features derived from interface structural profiles, WT and mutant sequences, and physics based scoring of the structures of the wild type and mutant complexes. **1**) Interface profile scores derived by structural alignment of structurally similar interfaces using an interface similarity cutoff to define the aligned sequences that are used to build the profile. **2**) Physics based scores are formed at the residue or atomic level formed by modeling the mutant monomeric protein and complex and evaluating the difference in energy. **3**) Sequence features are formed by the difference between the WT and mutant sequences in the number of hydrophobic (V, I, L, M, F, W, or C), aromatic (Y, F, or W), charged (R, K, D, or E), hydrogen bond acceptors (D, E, N, H, Q, S, T, or Y), and hydrogen bond donating residues (R, K, W, N, Q, H, S, T, or Y) along with difference in amino acid volume calculated from the sequence.

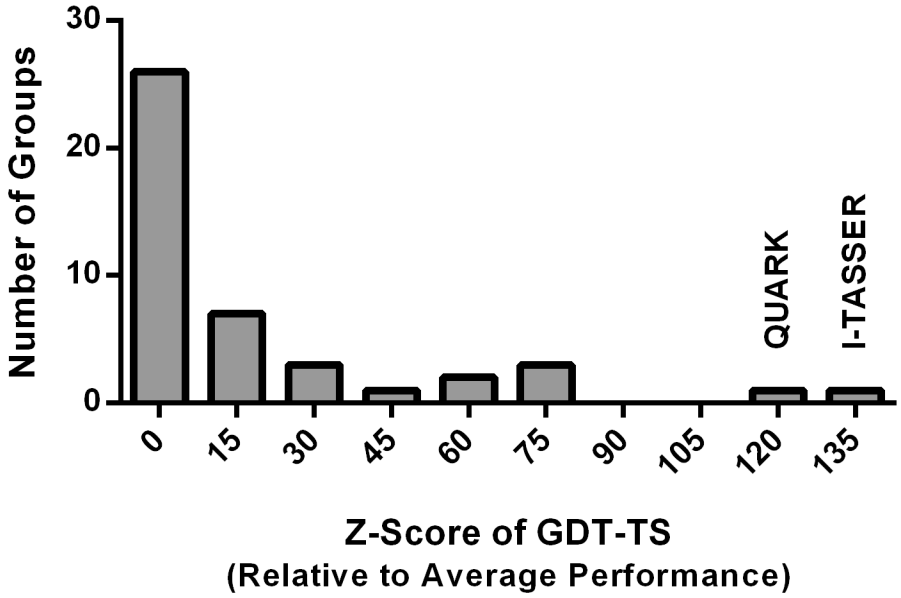

Figure 3. Histogram of the Z-scores of all automated protein structure predictors in the CASP11 experiment. The first bin contains groups that have Z-score below 0. Data are taken from official CASP webpage at URL http://www.predictioncenter.org/casp11/zscores_final.cgi?model_type=first&gr_type=server_only.

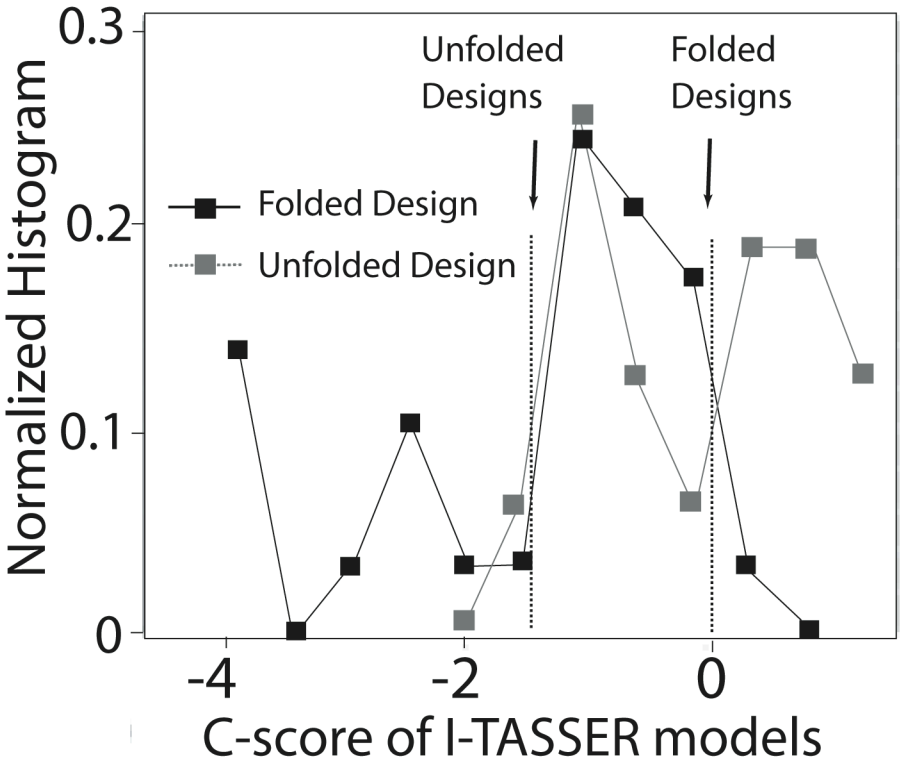

Figure 4. Divergence in the confidence score of the I-TASSER models for successfully and unsuccessfully designed sequences. Approximate cutoff values are indicated by the arrows. A C-score < -1.5 indicates a high probability that the design will be not folded correctly and a C-score >0 indicates a high probability that the design will fold to the target structure.

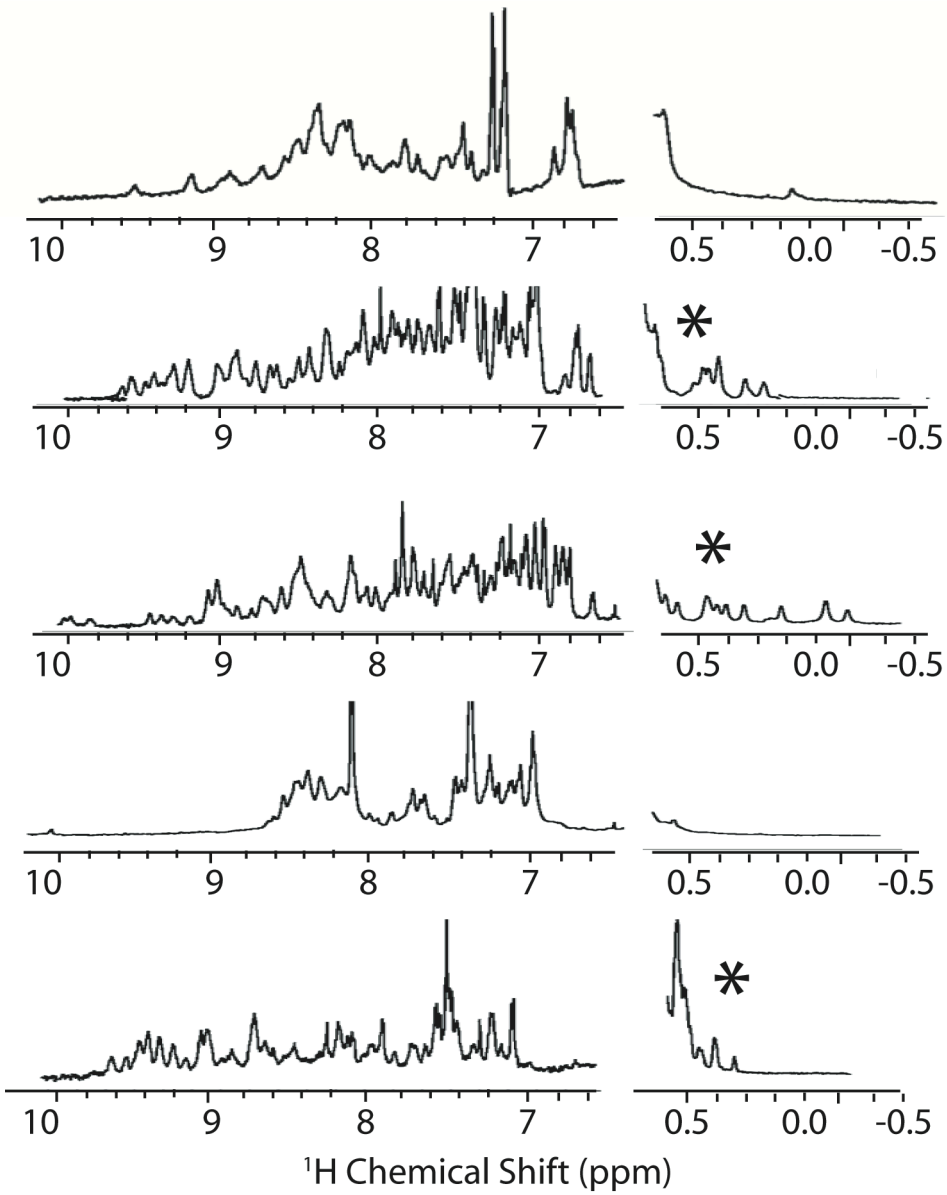

Figure 5. NMR spectra of folded (with asterisk) and unfolded designed proteins. The folded designs have a wider range of chemical shift values in the amide region of the spectrum (7-10 ppm) and have chemical shift values below 0.5 ppm indicating side-chains strongly shielded from solvent, as would be expected in a well-packed protein core.