Automated design of specificity in molecular recognition

James J. Havranek[1] and Pehr B. Harbury[2]

Published online 2 December 2002; doi:10.1038/nsb877

Specific protein–protein interactions are crucial in signaling networks and for the assembly of multi-protein complexes, and represent a challenging goal for protein design. Optimizing interaction specificity requires both positive design, the stabilization of a desired interaction, and negative design, the destabilization of undesired interactions. Currently, no automated protein-design algorithms use explicit negative design to guide a sequence search. We describe a multi-state framework for engineering specificity that selects sequences maximizing the transfer free energy of a protein from a target conformation to a set of undesired competitor conformations . To test the multi-state framework, we engineered coiled-coil interfaces that direct the formation of either homodimers or heterodimers. The algorithm identified three specificity motifs that have not been observed in naturally occurring coiled coils. In all cases, experimental results confirm the predicted specificities.

Computational protein design provides a rigorous test of our understanding of proteins. In effect, a design algorithm translates our hypotheses about protein structure and function into amino acid sequences. Experimental analysis of these sequences reports on the validity of the hypotheses. Recent design efforts have resulted in the realization of a novel backbone fold[1], the redesign of a folding pathway[2] and the design of a zinc finger domain that does not require metal binding for stability[3].

Most design studies follow the 'inverted-folding' strategy in which an optimal sequence for a preexisting backbone is selected by the design algorithm[4]. Protein structure is represented by a fixed backbone and a rotamer-based description of side chain conformation[5]. Amino acid sequences are selected that minimize a potential energy function when computationally modeled in the target conformation. We refer to this procedure as 'single-state' design. The potential energy functions used in protein design include empirically weighted contributions derived from molecular mechanics potentials, secondary structure propensities, structural database statistics and surface-area scaled terms that depend on hydrophobic/polar (H/P) character[6,7]. Because they combine a diverse set of energetic and statistical considerations, we refer to these as 'hybrid' potential energy functions. This general approach has led to numerous impressive results from several groups[2,3,8–13].

These successes suggest that the automated stabilization of fixed structures may be considered a solved problem. However,

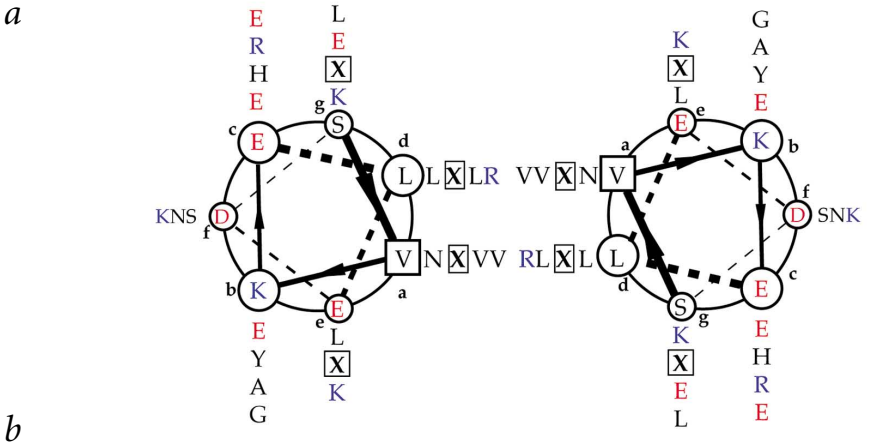

Fig. 1 A GCN4-derived scaffold for coiled-coil design. ***a***, A helical wheel diagram of the pCAP sequence. Positions allowed to vary in the design calculation are denoted by 'X'. These positions form the interface of the central heptad. The pCAP sequence is identical to an N-terminally capped variant of GCN4 (ref. 60) with the asparagine at position 16 shifted by one heptad level to position 9. ***b***, Constructs used for the experimental characterization of designed sequences. 'H_6' denotes a $(His)_6$-tag; 'SH', a (Gly-Gly-Cys) linker; and 'PKA', a protein kinase A-tag.

[1]Biophysics Program and [2]Department of Biochemistry, Stanford University, Stanford, California 94305, USA.

Correspondence should be addressed to P.B.H. *e-mail: harbury@cmgm.stanford.edu*

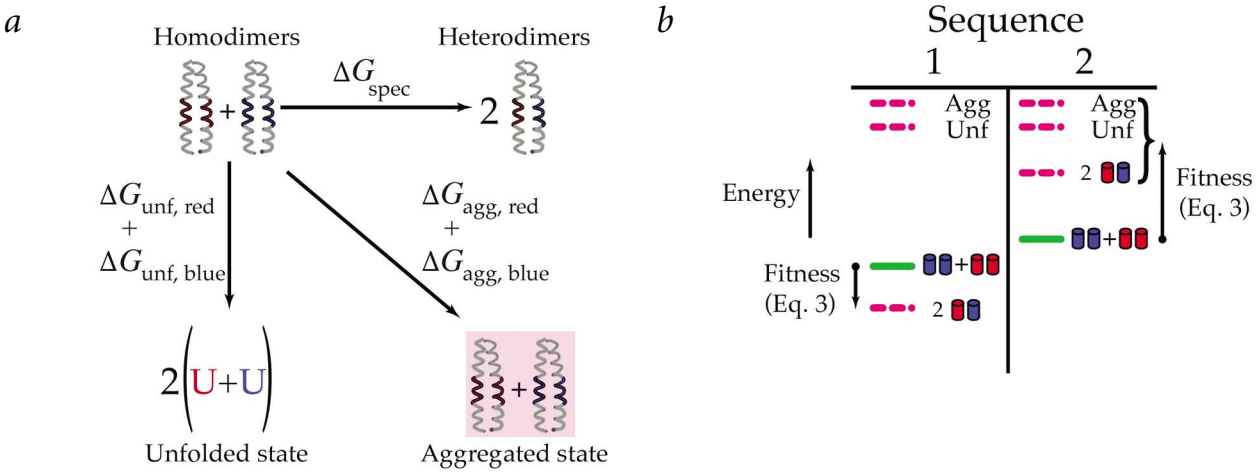

Fig. 2 Positive and negative design states. ***a***, Schematic representation of competing states included in the design calculations targeting the homodimer conformation. Competing states are included to enforce homospecificity (upper right), solubility (lower right) and stability (lower left). The aggregated state is modeled as the target conformation embedded in a medium with a dielectric constant of 65. Each sequence considered by the genetic algorithm is subjected to conformational optimization in each of the four states. The fitness score for a given sequence is the transfer free energy from the target state (homodimer) to the ensemble of competing states (heterodimers, aggregated state and unfolded state). ΔG_{spec} and ΔG_{unf} are transfer free energies between states and can be measured experimentally. ***b***, Energy diagram for two protein sequences in different conformational or associative states. A solid green line indicates the energy of the target state, the coiled-coil homodimers. The dashed magenta lines indicate the energies of competing states, including the unfolded protein, aggregated protein and the coiled-coil heterodimers. Sequence 1 minimizes the energy of the target state and would be incorrectly selected by the single-state design algorithm. The single-state design algorithm cannot recognize that the heterodimers are more stable than the homodimers because stabilities are computed only for the target state. The multi-state algorithm would correctly select sequence 2, because its transfer free energy from the target state to the ensemble of competing states is more positive than for sequence 1 (Eq. 3).

single-state approaches do not explicitly address discrimination between multiple states, which is a central feature of molecular specificity. Examples include proteins that selectively bind one small molecule without binding chemically related compounds, allosteric proteins that change conformation in the presence of a regulatory ligand and enzymes capable of binding transition states more tightly than ground states. To maximize specificity for a target state, the design algorithm must both stabilize the desired physical result (positive design) and destabilize undesired conformations, arrangements or states (negative design).

We present a general method for the automated design of specificity in molecular recognition. Following previous work[1,14,15], we represent each design requirement as a separate state that the protein can adopt. The algorithm achieves specificity by selecting sequences calculated to have an energetic preference for the target state over the negative design states. We refer to this procedure as 'multi-state' design. Using a coiled-coil model system for molecular recognition, we show that the use of multiple states in our calculations is necessary. The multi-state algorithm discovers new specificity motifs unreported in naturally occurring coiled coils.

Design of specific coiled-coil interfaces

We chose a dimeric coiled coil as our design scaffold because it represents the simplest protein–protein interface. Coiled coils have a characteristic heptad repeat (**a**–**g**, Fig. 1*a*). Positions **a** and **d** are typically occupied by hydrophobic residues, positions **e** and **g** by charged residues and positions **b**, **c** and **f** by polar residues. We redesigned positions **a**, **d**, **e** and **g** in the central heptad of the prototypical and well-studied homodimeric coiled coil GCN4 (ref. 16). Eight residues (four per helix) were varied, generating two distinct sequences. All non-proline amino acids were considered at the designed positions, allowing for a total of 8×10^9 possible sequence outcomes.

A design intended to select two coiled-coil sequences that preferentially associate into homodimers and do not cross-hybridize with each other is illustrated (Fig. 2). Four states were modeled. The first state is defined as the folded homodimer conformation, which is the target state. The second state is the folded heterodimer conformation, which is included as a competing state to select against sequences that cross-hybridize. The third state is the unfolded state of the polypeptides, which is included as a competitor to select against sequences that are unstable. The fourth state is the aggregated state, which is included as a competitor to select against sequences with poor water solubility.

Free energies were evaluated for candidate sequences in each of the four states. The fitness of a sequence was defined as its computed transfer free energy from the target state to the ensemble of competing states (Fig. 2*b*). Single-state design algorithms select sequences with the lowest computed energy in the target state (sequence 1, Fig. 2*b*). The multi-state design algorithm selects sequences that maximize the fitness (sequence 2, Fig. 2*b*), ensuring specificity towards the target state.

A genetic algorithm was used to evolve a population of sequences that maximize the transfer free energy from the target state to the ensemble of competing states. To distinguish different classes of solutions within the population, we clustered the 100 sequences of highest fitness into four groups using BLASTClust[17]. The sequence with the largest transfer free energy from each cluster is reported (Table 1; Fig. 3).

Identifying specific pair interactions

The designed sequences incorporate both previously identified and new amino acid motifs. To identify the pairwise interactions in these motifs that are responsible for the computed specificity, computational double-mutant cycles were performed. In the cases of sequences iv and vi (Fig. 3), specificity was achieved by patterning charged residues on the protein surface, which occurs naturally in coiled coils[18] and has been used in protein engineer-

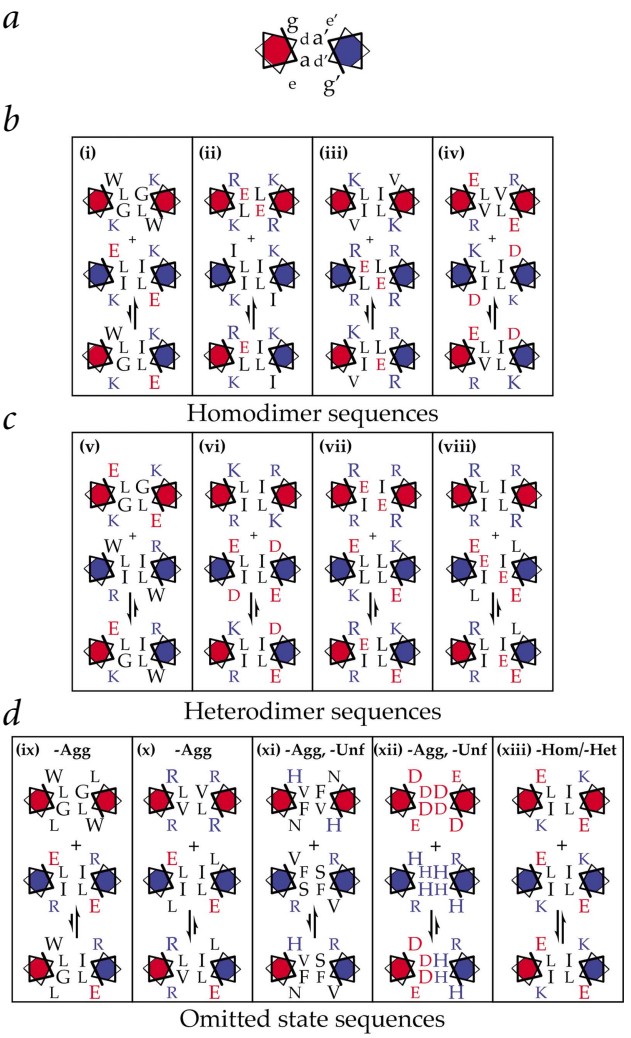

Fig. 3 Results of design calculations. ***a***, Positions of designed residues in the central heptad of pCAP (Fig. 1). ***b***, Designed homodimer sequences. The arrangement of residues for both the homodimer and heterodimer species is shown for each sequence pair. The target state of the design is indicated by the direction of the equilibrium arrow. Single letter abbreviations are used for the amino acids. Basic and acidic residues are shown in blue and red, respectively. ***c***, Designed heterodimer sequences. The arrangement of residues for both the homodimer and heterodimer species is shown for each sequence pair. ***d***, Omitted state sequences. Sequences result from calculations omitting one or more competing states (Table 1). The competing states omitted from each calculation are indicated at the top of the panel.

ing studies[19]. Several novel sequence patterns also emerge. Volume complementarity between a Trp side chain and a Gly side chain confers specificity in sequences i and v (Fig. 4*a*). Poor packing between a Leu side chain at heptad position **a** against β-branched side chains at positions **g′** and **a′** accounts for the homospecificity of sequence ii. In sequences iii and vii, a Glu side chain at heptad position **d** favors a basic amino acid at position **e′** over a hydrophobic alternative (Fig. 4*b*). Because position **d** of the heptad repeat is located in the hydrophobic core of the coiled coil, these sequences contain buried polar residues computationally engineered to confer specificity.

Multiple design goals require multiple states

To test whether multiple states are required to achieve our design criteria, we performed a second set of calculations in which one or more competing states were omitted (sequences ix–xiii, Fig. 3*d*; Table 1). The calculations with limited sets of competitors demonstrate that the neglect of any state yields inferior results relative to the results obtained with the full set of competitors. The omission of the aggregated state gives rise to sequences with fewer charged residues (compare sequences ix and x with i and vi). Although it is not clear whether the aggregated state is required for the success of our design, the loss of polar residues at surface positions is generally undesirable. Designs lacking both the aggregated and unfolded states lead to sequence pairs predicted to be specific but also unstable (sequences xi and xii). When

Table 1 Calculated (using the OPLS-UA potential) and observed thermodynamic quantities for designed sequence pairs

Sequence pair		States[1]				$\Delta G_{Fitness}$	$\Delta G_{spec,calc}$	$\Delta G_{spec,obs}$	$\Delta G_{unf,calc}$	$\Delta G_{unf,obs}$	$\Delta G_{agg,calc}$
	(A / B)	Hom	Het	Unf	Agg	(kcal mol^{-1})[2]	(kcal mol^{-1})[3]	(kcal mol^{-1})	(kcal mol^{-1})[4]	(kcal mol^{-1})	(kcal mol^{-1})[5]
i	WGLK / EILK	*	C	C	C	+4.3	+9.7	+0.9	+1.8 / +2.9	−5.4 / +1.6	+4.8 / +5.1
ii	RLEK / IILK	*	C	C	C	+3.8	+7.1	+3.0	+1.4 / +2.1	−3.8 / +1.9	+4.6 / +4.6
iii	KILV / RLER	*	C	C	C	+3.8	+7.6	+1.6	+2.7 / +1.6	−0.5 / −3.6	+4.3 / +4.5
iv	EVLR / KILD	*	C	C	C	+3.7	+4.8	+2.7	+3.1 / +1.1	−0.9 / −1.9	+4.9 / +4.6
v	EGLK / WILR	C	*	C	C	+4.7	−12.1	−1.7	+3.0	−0.6	+4.8
vi	KILR / EILD	C	*	C	C	+4.7	−8.7	−2.5	+2.9	−0.9	+4.8
vii	RIER / ELLK	C	*	C	C	+4.5	−5.6	−2.5	+2.9	+1.7	+4.7
viii	RILR / EIEL	C	*	C	C	+4.4	−12.2	−1.7	+2.3	−1.6	+4.6
ix	WGLL / EILR	*	C	C	–	+5.7	+12.8		+3.4		+4.9 / +3.9
x	RVLR / EILL	C	*	C	–	+6.4	−6.8		+3.9		+4.4
xi	HFVN / VSFR	*	C	–	–	+50.2	+50.2		−16.7		+4.6 / +4.1
xii	DDDE / HHHR	C	*	–	–	+109.3	−109.3		−10.4		+4.9
xiii	EILK / EILK	*	–	C	C	+5.7	0.0		+2.9		+5.1 / +5.1
	EILK / EILK	–	*	C	C	+5.7	0.0		+2.9		+5.1

[1]The target state for each calculation is denoted with an asterisk, competing states with a 'C', and states omitted from the calculation with a minus sign. The abbreviations for the states are Hom, homodimers; Het, heterodimers; Unf, the unfolded state; and Agg, the aggregated state.
[2]Transfer free energy from target state to ensemble of competitors.
[3]ΔG_{spec} is defined as the free energy change when the two homodimers are rearranged to form the two heterodimers.
[4]$\Delta\Delta G_{unf}$ is defined as the free energy difference between the unfolded and target states, subtracted by the same value for the pCAP (KVLE / KVLE) sequence. For homodimer species, $\Delta\Delta G_{unf}$ is reported for both sequences (A / B).
[5]ΔG_{agg} is defined as the free energy difference between the aggregated and target states. For homodimer species, ΔG_{agg} is reported for both sequences (A / B).

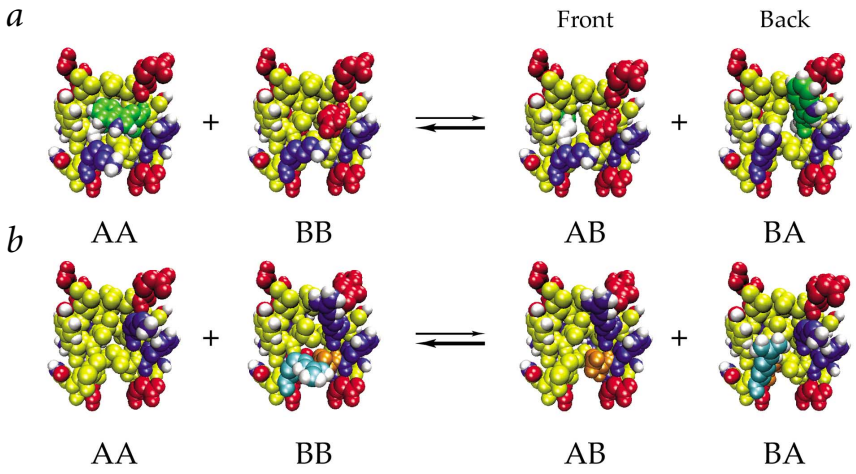

Fig. 4 Predicted specificity motifs. ***a***, Sequence pair i (Table 1). In the AA homodimer, a Trp residue (green) occupies a space created by a Gly residue on the opposite helix (white). In the AB heterodimer, a Glu residue is opposite the Gly residue. The Glu residue extends into solvent, leaving a cavity in the hydrophobic core of the coiled coil. The Trp residue is placed opposite an Ile residue and cannot pack into the core. ***b***, Sequence pair iii. In the BB homodimer, a Glu residue at the heptad **d** position (orange) is in close contact with an Arg residue at the opposing **e′** position (cyan). In the AB heterodimer, the Glu is opposite an uncharged residue (Val). Figures were generated with VMD[61].

the homodimer competitor is omitted from the heterodimer design, all association specificity is lost (sequence xiii). Omission of the heterodimer competitor from the homodimer design results in the same loss of specificity (sequence xiii). We conclude that both positive and negative design states must be considered to achieve specificity in our calculations.

Experimental validation of designed sequences

To test the energetic predictions, sequence pairs i–viii were expressed, purified and characterized experimentally. We first determined whether the target species form parallel dimeric structures by measuring whether the apparent melting temperatures (T_m) of C-terminally disulfide-bonded–target coiled coils vary with peptide concentration[18]. All of the T_ms were observed to be concentration independent. These data rule out the possibility that the disulfide-bonded coiled coils form higher order oligomers or adopt antiparallel conformations. For six of the species, the dimer oligomerization state was confirmed independently by analytical ultracentrifugation.

A disulfide-exchange assay was used to measure directly the equilibrium between the homodimer and heterodimer states of the designed coiled coils (Fig. 5*a*,*b*). In each instance, specificity for the desired association state was achieved (Table 1; Fig. 5*c*). Two sets of predictions are shown. The first set is computed with the OPLS-UA potential energy function, which was used in the design calculation[20]. In addition, we report specificities calculated identically, using the CHARMM19 potential energy function[21], which was applied after the design to evaluate the selected sequences. Both sets of predictions correlate with the measured values (the square of the correlation coefficient (R^2) is 0.7 for both OPLS-UA and CHARMM19).

To test our predictions of unfolding free energies, stabilities were measured for sequences i–viii by urea denaturation. All measurements were taken in 5 mM phosphate buffer, consistent with the low salt environment used for the design calculation. For homodimer species, melts of unmodified coiled coils were performed. For heterodimer species, disulfide-bonded coiled coils were studied to prevent the formation of a mixed population of dimers. The data were fit assuming a two-state bimolecular (homodimers) or unimolecular (heterodimers) folding reaction (Fig. 6*a*,*b*). Stabilities were extracted from the data and referenced to that of the pCAP peptide, the parental sequence for the design calculation. We compared the stabilities predicted using the OPLS-UA[20] and CHARMM19 potential energy func-

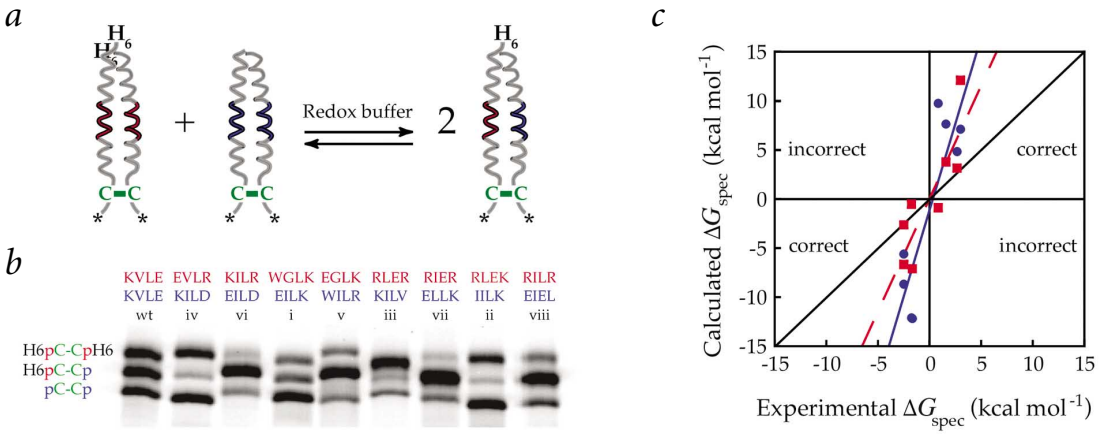

Fig. 5 Specificity of designed coiled coils. ***a***, One member of each sequence pair was expressed in *E. coli* with an N-terminal His_6-tag (red), whereas the other was expressed without the tag (blue). The proteins were allowed to exchange helix partners in the presence of redox reagents, which facilitated the breaking and reforming of disulfide bonds, until equilibrium was reached. The exchange reaction was then quenched, radioactively labeled and analyzed by SDS-PAGE. ***b***, Autoradiograph of electrophoretically separated exchange reaction. The top band is composed of His_6-tagged homodimers; the middle band, heterodimers; and the bottom band, untagged homodimers. Sequence pairs are colored corresponding to whether they were expressed with (red) or without (blue) a His_6-tag. ***c***, Specificities calculated using the OPLS-UA potential energy function[20] (blue circles; slope = 3.5 and R^2 = 0.7) or the CHARMM19 potential energy function[21] (red squares; slope = 2.3 and R^2 = 0.7) plotted against the measured values. Lines of best fit are shown in solid blue (OPLS-UA) and dashed red (CHARMM19). The diagonal is shown as a solid black line. The quadrants of the graph are labeled to indicate where the computation correctly predicts measured specificity.

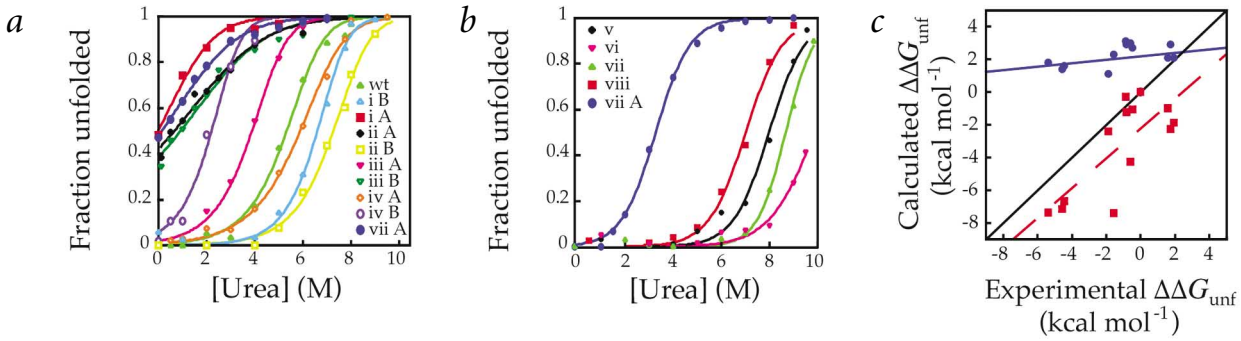

Fig. 6 Stability of designed coiled coils. ***a***, Urea denaturation of designed homodimer coiled coils. Sequences are identified by lower case roman numerals (Table 1). The pCAP sequence is denoted 'wt'. Curves were fit using a two-state bimolecular model. ***b***, Urea denaturation of disulfide-bonded dimers of designed heterodimer coiled coils. Curves were fit using a two-state unimolecular model. Sequences are identified by lower case roman numerals (Table 1). ***c***, The calculated stability for each of the designed species (eight homodimers, four heterodimers and the scaffold) is plotted against the observed value. Blue circles denote values calculated using the OPLS-UA potential energy function[20], and red squares are those using the CHARMM19 potential energy function[21]. Lines of best fit are shown in solid blue (OPLS-UA; slope = 0.1, R^2 = 0.1) and dashed red (CHARMM19; slope = 0.9, R^2 = 0.6). The diagonal is shown as a solid black line.

tions[21] with the experimental stabilities. All target species are predicted to be more stable than the pCAP peptide by the OPLS-UA potential energy function. In contrast, stabilities calculated with the CHARMM19 potential are in closer agreement with the observed values.

Discussion

Our design algorithm differs from previous efforts[2,3,8,9] in two ways. First, we select sequences that maximize the transfer free energy of a protein from a target state to an ensemble of explicitly represented competitors, rather than optimizing the computed potential energy for a single target state. As a result, sequence optimization is necessarily distinct from structural optimization, and structural optimization must be performed separately for each state[22]. The computed specificities are used to guide the subsequent sequence search. Second, we evaluate conformational free energies using a standard molecular mechanics potential energy function with a continuum solvent model (MM/CS) rather than a hybrid energy function. This allows us to directly compare predicted and observed free energies. Although MM/CS potential energy functions are still under development, they are more thoroughly parameterized and tested than hybrid energy functions. Because they can account for the energetics of small molecules, nucleic acids and proteins, they are expected to be more general than hybrid energy functions. Finally, the use of a standard molecular mechanics potential energy function allows for its modular substitution with improved potentials, as advances (such as polarizable potential energy functions) emerge from other fields of computational chemistry.

Specificity in protein design

Specificity is, by definition, a multi-state property[23]. The probability that a designed protein will adopt a target conformation or state is given by:

$$P_{\text{target}} = (e^{-A_{\text{target}}/RT}) / (\sum_{\text{target, competitors}} e^{-A/RT}) \qquad (1)$$

where A_{target} is the free energy of the target conformation, and A is the free energy of each conformation in the sum in the denominator.

Single-state design is predicated on maximizing the numerator of Eq. 1, neglecting the effects of sequence variation on the denominator (the partition sum). This approach has been used successfully to engineer specificity[15,24]. However, if any competitor conformation structurally resembles the target conformation, optimization for the target will be correlated with optimization for the competitor, and the single-state strategy will likely break down[14,25,26].

Failure of single-state design is observed in lattice models of proteins, resulting in heteropolymer sequences that fold into multiple conformations[27]. To address this deficiency, a new generation of lattice-design algorithms selects sequences that directly optimize P_{target} in Eq. 1 rather than target stability[28–30]. Although optimizing P_{target} may seem computationally prohibitive, given the large number of states that could contribute to the denominator in Eq. 1, it has been noted[31,32] that the partition sum is dominated by a small number of low-energy conformations that are structurally similar to the target. The partition sum can thus be approximated by modeling this subset of near-native conformations.

The differences between the single-state and multi-state strategies are highlighted by the manner in which they achieve the 'hydrophobic in/polar out' pattern observed in naturally occurring proteins. This pattern reflects the realities that buried charges can be destabilizing and that an excess of hydrophobic residues at the surface can lead to aggregation. In single-state design algorithms, the unfolded and aggregated states are not modeled. Consequently, the selection of charged residues at buried positions is discouraged by penalizing the burial of polar surface area, by excluding polar residues from buried positions[7,12] or by constraining amino acid composition[33]. Likewise, hydrophobic residues are often excluded from consideration at surface positions to prevent aggregation. Although sequence constraints may be expedient for enforcing 'hydrophobic in/polar out' patterning, protein function often depends on exceptions to this rule[34–40]. In the multi-state approach, the patterning of hydrophobic and polar residues arises as a natural consequence of simultaneous competition against the unfolded and aggregated states. Thus, polar residues are not excluded from the cores of proteins; their selection is based on an energetic balance between the requirements for stability and specificity.

The multi-state approach only offers an advantage for designing against undesired competitors states that are known and can be modeled. With respect to unknown competitors, the single-state and multi-state approaches are equivalent. One must

hope that specificity against unknown competitors will arise fortuitously as a consequence of sequence optimization for the target state. To assess the magnitude of such fortuitous specificity, we measured ΔG_{spec} for seven pairs of homodimer sequences that were not deliberately designed to disfavor cross-hybridization with each other (iA/iiiB, iA/ivB, iiA/iiiB, iiA/ivB, iiiA/ivB, ivA/iiB and ivA/iiiB; Fig. 3)). The ΔG_{spec} values for these pairs range from –0.1 to 1.1 kcal mol^{-1}, averaging 0.6 kcal mol^{-1} (data not shown). Seven of the eight engineered sequence pairs (sequences i–viii, Fig. 3; Table 1) show values of ΔG_{spec} exceeding in magnitude the largest value of ΔG_{spec} that arises fortuitously.

Assessment of the physical model

Comparison of experimentally measured free energies with predicted values reports on the accuracy of the physical model used for design, which includes the side chain rotamer library, the backbone representation and the potential energy function. The results demonstrate that MM/CS energy potentials are capable of conferring functional specificity on designed proteins. However, the quantitative agreement between energetic predictions and measured values leaves room for improvement.

As observed by others, we find that the accuracy of energetic estimates strongly depends on the number of rotamers used to model side chain conformations[41]. The library we used included 1,064 rotamers to represent the 19 non-proline amino acids. Optimizing the side chain coordinates in the presence of the fixed backbone after rotamer placement was necessary to achieve energy values that could be compared across all states[9,42].

Our physical model does not consider backbone movements of the protein, which could mitigate unfavorable interactions in the negative design states. This limitation probably contributes to the overestimation of the specificities by both the OPLS-UA[20] and CHARMM19 potential energy functions[21] (Fig. 5*c*). Incorporating backbone flexibility in the multi-state framework should be possible by including several fixed backbone conformations for each of the competing states.

All potential energy functions contain errors. The design process likely inflates the cumulative error in the force field used for the design calculation. Presumably, the genetic algorithm selects sequences for which errors in the OPLS-UA potential energy function are correlated and preferentially stabilize the target state. In contrast, the designed sequences are expected to sample errors that are present in the CHARMM19 potential energy function randomly, yielding a more accurate assessment of their stabilities. The differences in the OPLS-UA and CHARMM19 stability predictions derive from the bonded and Lennard-Jones terms in the potential energy function (data not shown). The results suggest that cross-validation by independent potential energy functions could be used to identify designed sequences that likely contain accumulated errors before time-consuming experimental efforts are initiated.

Conclusion

We have presented a general method for incorporating specificity into protein design. Each positive and negative design requirement is embodied as a separate state in our algorithm. We have verified experimentally that this multi-state framework produces functionally specific protein–protein recognition. The use of a molecular mechanics potential energy function with a continuum solvent model allows for comparison of the predicted and observed free energies. The results suggest that the use of several potential energy functions may help to minimize the effects of errors present in these functions.

Our framework for multiple competing states is applicable beyond the simple protein–protein interactions that we have considered here. For example, larger sets of orthogonal coiled coils could be designed to direct complex self-assembly processes. Competing states in which a protein is bound to 'decoy' ligands could be used to direct the design of specific small-molecules (for example, see ref. 14). Finally, an explicit framework for the stabilization of the transition state of a reaction relative to its ground states should be possible[43].

Methods

Rotamer library and optimization. Backbone coordinates for a symmetric idealized coiled coil were generated from a mathematical model using parameters optimal for Val and Leu residues at heptad positions **a** and **d**[44] (Fig. 1). The most commonly occurring rotamers for an α-helix were taken from the backbone-dependent library of Dunbrack and Karplus[45]. Sufficient rotamers for each amino acid were extracted to account for 95% of all observed conformations. Sulfhydryl and hydroxyl hydrogens were added with dihedral angles of –60°, 60° and 180°. Rotamers were built onto the backbone structure and energy minimized using either the OPLS-UA[20] or CHARMM19 (ref. 21) geometric and van der Waals potential energy terms and a 20° square-well dihedral restraint[20]. Additional rotamers were then introduced, offset from the minimized values by 1.3 s.d. in the χ_1 dihedral angle[45] for lysine, methionine, glutamine, glutamate and arginine, and in χ_1 and χ_2 for all other amino acids (20° for the hydrogen dihedrals above). Atom positions for additional rotamers were energy minimized, with their dihedral angles held fixed. Rotamer probabilities were optimized following Koehl and Delarue[46]. Because the mean-field algorithm does not guarantee convergence to the global minimum[47], each sequence in Table 1 was repacked 50× with different random initial rotamer probabilities. The results agreed to within 0.01 kcal mol^{-1}, suggesting that the repacking algorithm identifies the globally optimal conformation for these sequences.

Energy function. The potential energy of the system is approximated in a pairwise factorable form[46] and decomposed into the following contributions:

$$U^{total} = U^{Geom} + U^{LJ} + U^{MTK\,/\,\gamma} \quad (2)$$

U^{Geom} consists of the bonded energy terms from the OPLS-UA[20] or CHARMM19 (ref. 21) force field. $U^{MTK\,/\,\gamma}$ is identical to the FDPB / γ solvation energy[48], with the electrostatic energies calculated from PARSE parameters[48] using the modified Tanford-Kirkwood algorithm[49]. The solvent and protein dielectric constants used were 80 and 4, respectively. Pairwise surface areas were calculated similarly to Street and Mayo[50]. Separate scaling factors were stored for backbone and side chain atoms at each position. The scaling factors were selected so that the pairwise-calculated buried surface would equal the exact buried surface areas for all residues in the GCN4 structure (PDB entry 2ZTA)[16]. For sequences i–viii (Table 1), the average difference between pairwise computed and exact surface area was 68 Å^2. U^{LJ} is the Lennard-Jones potential energy. For one-body energies, the Lennard-Jones function with OPLS-UA or CHARMM19 parameters was used. For interactions between rotamers (two-body energies), we used a fuzzy Lennard-Jones function. Lennard-Jones interaction energies were calculated with radii scaled[51] by 0.9, and negative (favorable) interaction energies were set to zero. Surface area buried between side chain rotamers was assigned an energy density of –16 cal mol^{-1} Å^{-2}; this value was derived from two constants taken from the literature. First, the experimentally determined surface-area energy density for transfer of acetyl-X-amide analogs of non-polar side chains from water to octanol[52,53] is 21 cal mol^{-1} Å^{-2}. Second, the FDPB / γ solvation model assigns a surface-area energy density for transfer of hydrocarbons from water to vacuum[48] of 5 cal mol^{-1} Å^{-2}. The difference between these values, –16 cal mol^{-1} Å^{-2}, is the surface-area energy density for transfer from vacuum to octanol, the appropriate value for the fuzzy Lennard-Jones function.

Design targets and competitors. The eight designed positions in each calculation are located in two distinct polypeptides, A and B (Fig. 1). Free energies for the coiled-coil states were calculated using the coiled-coil backbone template subject to mean-field repacking of side chains. The free energy of the unfolded state was calculated in two steps according to the following scheme: D↔2M↔2U, where D is the folded dimer, M is a monomer α-helix and U is the unfolded polypeptide monomer. Free energies for monomer helices were calculated using an isolated helix backbone subject to mean-field repacking of side chains. The free energy for unfolding of the monomer helices was computed using the AGADIR parameters[54]. We added a sequence-independent constant to the energy of the unfolded state so that the stability of the pCAP sequence would evaluate to 3 kcal mol^{-1}, its measured stability at 1 μM concentration. The aggregated state was modeled as the target conformation embedded in a medium with a dielectric constant of 65 rather than 80. This value yields an energy gap between the native and aggregated states of the pCAP sequence comparable to its stability at 1 μM concentration. Inclusion of the aggregated state as a competitor guarantees that the designed sequences will have an unfavorable transfer free energy to a solvent of lower dielectric constant. The free energies of states involving heterodimers were decreased by $RT\ln(2)$ to account for the entropy of mixing.

Each state in the design consisted of two copies of the A and B polypeptides in different environments and arrangements. The homo- and heterodimer states consisted of the appropriate arrangements of the A and B sequences evaluated in the folded conformation. Three unfolded-state competitors were considered, corresponding to the unfolding of AA, BB or one copy of AB, with the other polypeptides remaining folded. Designs targeting the homodimer state included two aggregated state competitors, and those targeting the heterodimer state included one. These competitors involved the transfer of single coiled coils to the lower dielectric environment, analogous to the unfolded-state competitors.

Fitness function. The fitness of a given sequence is defined as the transfer free energy of that sequence from the target state to the ensemble of competing states (Fig. 2*b*).

$$\text{Fitness} \equiv -RT\ln(\Sigma_{\text{competitors}} e^{-A_c / RT}) - A_{\text{target}} \quad (3)$$

where A_c is the free energy of the competing states, A_{target} is the free energy of the target state and RT is evaluated at room temperature.

Genetic algorithm. An initially random population of 4,800 discrete sequences was propagated for 30 generations. Three rules dictated the composition of subsequent generations once fitness scores were evaluated. First, the most-fit sequence of each generation was automatically propagated. Uniform crossover recombination was then used to generate 99% of the remaining sequences[55]. Finally, mutation of single sequences at 20% probability per site was used to generate the remainder of the population. Sequences chosen for the recombination and mutation processes were selected randomly, biased by fitness scores such that P_{s*} (the probability of selecting sequence s*) was

$$P_{s*} = (e^{-(F_s^* / \sigma)}) / (\Sigma_s e^{-(F_s / \sigma)}) \quad (4)$$

where F_s is the fitness of member s, σ is the standard deviation in fitness for the current generation and the sum in the denominator extends over the entire population. We performed each design calculation three times with different random initial sequence populations. These calculations identified the same best sequence, suggesting that the genetic algorithm finds the global optimum.

Cloning and expression. The pCAP and pCAP-SH-PKA constructs (Fig. 1*b*) were appended to the *TrpLE′* leader sequence[56]. All constructs were cloned into the pET24a vector (Novagen) using standard molecular biology techniques. Mutations were introduced using the method of Kunkel[57] and verified by DNA sequencing. The pCAP and pCAP-SH-PKA peptides were purified from inclusion bodies and cleaved from the *TrpLE′* leader sequence with cyanogen bromide. Peptides in the pH6-CAP-SH-PKA construct were purified by nickel-NTA affinity chromatography directly from cell lysates. Final purification of all peptides was performed by reversed-phase HPLC. Peptide identities were confirmed by electrospray mass spectrometry. Protein concentrations were determined using the method of Edelhoch[58].

Measurement of specificity. Redox exchange reactions were performed at 10 μM peptide concentration in 5 mM Tris-HCl, pH 9.0, 50 μM β-mercaptoethanol and 100 μM 2-hydroxyethyl disulfide[18]. The reactions were equilibrated overnight and quenched by the addition of iodoacetamide to 10 mM for 1 h. The peptides were labeled by incubation with 5 U protein kinase A (Sigma) at 37 °C for 3 h at final concentrations of 5 μM peptide, 5 mM Tris-HCl, pH 9.0, 0.005% (v/v) Triton X-100, 40 μM ATP and 10 μCi [γ-P^{33}]ATP. The labeled mixture was analyzed by SDS-PAGE using the Tris-tricine system[59] in the absence of reducing agents. Gel bands were quantitated on a Phosphorimager (Molecular Dynamics). ΔG^{X}_{spec} is defined as $-RT\ln(K^X / K^{pCAP})$, where $K^X \equiv [A^XB^X]^2 / ([A^XA^X][B^XB^X])$ and K^{pCAP} denotes the equilibrium constant for the pCAP sequence. Results from exchange reactions initiated from the pure homodimer and pure heterodimer forms of the protein agreed to within 0.1 kcal mol^{-1} in all cases, indicating that equilibrium had been reached[18].

Measurement of stability. $\Delta\Delta G_{unf}$ is equal to the difference in stability (ΔG_{unf}) between each designed peptide and the original pCAP peptide. ΔG_{unf} was determined by urea denaturation in 5 mM potassium phosphate, pH 7.1, monitored by circular dichroism spectroscopy at 222 nm and 4 °C on an Aviv DS-62A spectropolarimeter at 5 μM peptide concentration. Data were converted to the fraction of dimer unfolded. Where a folded baseline was unavailable, data were collected at low concentrations of urea in the presence of 20% (v/v) trifluoroethanol (TFE), a potent helix-inducing solvent. A folded baseline was then extracted from this data. For well-folded species, 20% TFE did not affect the *y*-intercept or slope of the baseline (data not shown). Where an unfolded baseline was unavailable, a value of zero was assumed. A two-state bimolecular model was used to fit the homodimer data, and a two-state unimolecular model was used to fit the disulfide-bonded heterodimer data. Stabilities were measured for sequence vii A in both the disulfide-bonded and unmodified forms to serve as a calibration between the data sets.

Acknowledgments

We thank F.E. Boas, J.A. Silverman, D.R. Halpin, S.J. Wrenn and R.L. Baldwin for stimulating conversations and criticism during the course of this work and for comments on the manuscript. We also acknowledge helpful suggestions from the anonymous referees. This research was funded by a Searle scholar grant to P.B.H. from the Chicago Community Trust.

Competing interests statement

The authors declare that they have no competing financial interests.

Received10 July, 2002; accepted 8 November, 2002.

1. Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T. & Kim, P.S. High-resolution protein design with backbone freedom. *Science* **282**, 1462–1467 (1998).
2. Nauli, S., Kuhlman, B. & Baker, D. Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* **8**, 602–605 (2001).
3. Dahiyat, B.I. & Mayo, S.L. *De novo* protein design: fully automated sequence selection. *Science* **278**, 82–87 (1997).
4. Pabo, C. Molecular technology. Designing proteins and peptides. *Nature* **301**, 200 (1983).
5. Ponder, J.W. & Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791 (1987).
6. Kuhlman, B. & Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA* **97**, 10383–10388 (2000).
7. Gordon, D.B., Marshall, S.A. & Mayo, S.L. Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509–513 (1999).
8. Hellinga, H.W. & Richards, F.M. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. USA* **91**, 5803–5807 (1994).
9. Desjarlais, J.R. & Handel, T.M. *De novo* design of the hydrophobic cores of proteins. *Protein Sci.* **4**, 2006–2018 (1995).
10. Lazar, G.A., Desjarlais, J.R. & Handel, T.M. *De novo* design of the hydrophobic core of ubiquitin. *Protein Sci.* **6**, 1167–1178 (1997).
11. Malakauskas, S.M. & Mayo, S.L. Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* **5**, 470–475 (1998).
12. Marshall, S.A. & Mayo, S.L. Achieving stability and conformational specificity in designed proteins *via* binary patterning. *J. Mol. Biol.* **305**, 619–631 (2001).
13. Marvin, J.S. & Hellinga, H.W. Conversion of a maltose receptor into a zinc biosensor by computational design. *Proc. Natl. Acad. Sci. USA* **98**, 4955–4960 (2001).
14. Wilson, C., Mace, J.E. & Agard, D.A. Computational method for the design of enzymes with altered substrate specificity. *J. Mol. Biol.* **220**, 495–506 (1991).
15. Shimaoka, M. *et al.* Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.* **7**, 674–678 (2000).
16. O'Shea, E.K., Klemm, J.D., Kim, P.S. & Alber, T. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* **254**, 539–544 (1991).
17. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
18. O'Shea, E.K., Rutkowski, R., Stafford, W.F. III & Kim, P.S. Preferential heterodimer formation by isolated leucine zippers from Fos and Jun. *Science* **245**, 646–648 (1989).
19. O'Shea, E.K., Lumb, K.J. & Kim, P.S. Peptide Velcro: design of a heterodimeric coiled coil. *Curr. Biol.* **3**, 658–667 (1993).
20. Jorgensen, W.L. & Tiradorives, J. The OPLS potential functions for proteins: energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **110**, 1666–1671 (1988).
21. Brooks, B.R. *et al.* CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1983).
22. Koehl, P. & Levitt, M. *De novo* protein design. II. Plasticity in sequence space. *J. Mol. Biol.* **293**, 1183–1193 (1999).
23. Janin, J. Quantifying biological specificity: the statistical mechanics of molecular recognition. *Proteins* **25**, 438–445 (1996).
24. Ghirlanda, G., Lear, J.D., Lombardi, A. & DeGrado, W.F. From synthetic coiled coils to functional proteins: automated design of a receptor for the calmodulin-binding domain of calcineurin. *J. Mol. Biol.* **281**, 379–391 (1998).
25. Hellinga, H.W. Rational protein design: combining theory and experiment. *Proc. Natl. Acad. Sci. USA* **94**, 10015–10017 (1997).
26. Hellinga, H.W. Construction of a blue copper analogue through iterative rational protein design cycles demonstrates principles of molecular recognition in metal center formation. *J. Am. Chem. Soc.* **120**, 10055–10066 (1998).
27. Yue, K. *et al.* A test of lattice protein-folding algorithms. *Proc. Natl. Acad. Sci. USA* **92**, 325–329 (1995).
28. Deutsch, J.M. & Kurosky, T. New algorithm for protein design. *Phys. Rev. Lett.* **76**, 323–326 (1996).
29. Irback, A., Peterson, C., Potthast, F. & Sandelin, E. Monte Carlo procedure for protein design. *Phys. Rev. E* **58**, R5249–R5252 (1998).
30. Dima, R.I., Banavar, J.R., Cieplak, M. & Maritan, A. Statistical mechanics of protein-like heteropolymers. *Proc. Natl. Acad. Sci. USA* **96**, 4904–4907 (1999).
31. Banavar, J.R. *et al.* Structure-based design of model proteins. *Proteins* **31**, 10–20 (1998).
32. Rossi, A., Maritan, A. & Micheletti, C. A novel iterative strategy for protein design. *J. Chem. Phys.* **112**, 2050–2055 (2000).
33. Koehl, P. & Levitt, M. *De novo* protein design. I. In search of stability and specificity. *J. Mol. Biol.* **293**, 1161–1181 (1999).
34. Jones, S. & Thornton, J.M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **93**, 13–20 (1996).
35. Buckle, A.M., Schreiber, G. & Fersht, A.R. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-Å resolution. *Biochemistry* **33**, 8878–8889 (1994).
36. Shoichet, B.K., Baase, W.A., Kuroki, R. & Matthews, B.W. A relationship between protein stability and protein function. *Proc. Natl. Acad. Sci. USA* **92**, 452–456 (1995).
37. Derewenda, U. *et al.* An unusual buried polar cluster in a family of fungal lipases. *Nat. Struct. Biol.* **1**, 36–47 (1994).
38. Warshel, A. & Aqvist, J. Electrostatic energy and macromolecular function. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 267–298 (1991).
39. Lumb, K.J. & Kim, P.S. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry* **34**, 8642–8648 (1995).
40. Bolon, D.N. & Mayo, S.L. Polar residues in the protein core of *Escherichia coli* thioredoxin are important for fold specificity. *Biochemistry* **40**, 10047–10053 (2001).
41. Desjarlais, J.R. & Handel, T.M. Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305–318 (1999).
42. Keating, A.E., Malashkevich, V.N., Tidor, B. & Kim, P.S. Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils. *Proc. Natl. Acad. Sci. USA* **98**, 14825–14830 (2001).
43. Jencks, W.P. *Catalysis in Chemistry and Enzymology* (Dover, New York; 1987).
44. Harbury, P.B., Tidor, B. & Kim, P.S. Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc. Natl. Acad. Sci. USA* **92**, 8408–8412 (1995).
45. Dunbrack, R.L. & Karplus, M. Backbone-dependent rotamer library for proteins. *J. Mol. Biol.* **230**, 543–574 (1993).
46. Koehl, P. & Delarue, M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249–275 (1994).
47. Mendes, J., Soares, C.M. & Carrondo, M.A. Improvement of side-chain modeling in proteins with the self-consistent mean field theory method based on an analysis of the factors influencing prediction. *Biopolymers* **50**, 111–131 (1999).
48. Sitkoff, D., Sharp, K.A. & Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978–1988 (1994).
49. Havranek, J.J. & Harbury, P.B. Tanford-Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. USA* **96**, 11145–11150 (1999).
50. Street, A.G. & Mayo, S.L. Pairwise calculation of protein solvent-accessible surface areas. *Fold. Des.* **3**, 253–258 (1998).
51. Dahiyat, B.I. & Mayo, S.L. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **94**, 10172–10177 (1997).
52. Fauchère, J.-L. & Pliska, V. Hydrophobic parameters-π of amino-acid side-chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem.* **18**, 369–375 (1983).
53. Wimley, W.C., Creamer, T.P. & White, S.H. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* **35**, 5109–5124 (1996).
54. Lacroix, E., Viguera, A.R. & Serrano, L. Elucidating the folding problem of α-helices: local motifs; long-range electrostatics; ionic-strength dependence and prediction of NMR parameters. *J. Mol. Biol.* **284**, 173–191 (1998).
55. Mitchell, M. *An Introduction to Genetic Algorithms* (MIT Press, Cambridge, Massachusetts; 1996).
56. Kleid, D.G. *et al.* Cloned viral protein vaccine for foot-and-mouth disease: responses in cattle and swine. *Science* **214**, 1125–1129 (1991).
57. Kunkel, T.A., Bebenek, K. & McClary, J. Efficient site-directed mutagenesis using uracil-containing DNA. *Methods Enzymol.* **204**, 125–139 (1991).
58. Edelhoch, H. Spectroscopic determination of tryptophan and tyrosine in proteins. *Biochemistry* **6**, 1948–1954 (1967).
59. Schagger, H. & von Jagow, G. Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal. Biochem.* **166**, 368–379 (1987).
60. Lu, M. *et al.* Helix capping in the GCN4 leucine zipper. *J. Mol. Biol.* **288**, 743–752 (1999).
61. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).