ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

Proteins from an Unevolved Library of *de novo* Designed Sequences Bind a Range of Small Molecules

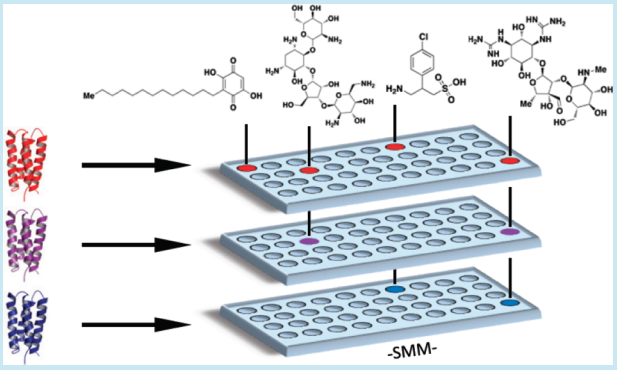
Izhack Cherny,[†] Maria Korolev,[†] Angela N. Koehler,[‡] and Michael H. Hecht*[,†]

[†]Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States

[‡]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, United States

Supporting Information

ABSTRACT: The availability of large collections of *de novo* designed proteins presents new opportunities to harness novel macromolecules for synthetic biological functions. Many of these new functions will require binding to small molecules. Is the ability to bind small molecules a property that arises only in response to biological selection or computational design? Or alternatively, is small molecule binding a property of folded proteins that occurs readily amidst collections of unevolved sequences? These questions can be addressed by assessing the binding potential of *de novo* proteins that are designed to fold into stable structures, but are "naïve" in the sense that they (i) share no significant sequence similarity with natural proteins and (ii) were neither selected nor designed to bind small molecules. We chose three naïve proteins from a library of sequences designed to fold into 4-helix bundles and screened for binding to 10,000 compounds displayed on small molecule microarrays. Several binders were identified, and binding was characterized by a series of biophysical assays. Surprisingly, despite the similarity of the three *de novo* proteins to one another, they exhibit selective ligand binding. These findings demonstrate the potential of novel proteins for molecular recognition and have significant implications for a range of applications in synthetic biology.



KEYWORDS: *protein design, binary code, small molecule microarray, four helix bundle, molecular evolution*

Recent advances enabling the design and construction of large collections of novel proteins present new opportunities to devise functional macromolecules for applications in synthetic biology. For many of these applications, proper function will depend on the ability of a novel protein to recognize and bind to a small molecule (SM). SMs can contribute to protein function by acting as substrates, cofactors, or allosteric regulators. Moreover, SM binding can enable functions that range from enzyme catalysis to gene regulation, and more than 30% of natural proteins require SMs to fold and/or function.[1] For some natural proteins, SM binding is tight and specific, while for others it is weak and/or promiscuous. Although tight and specific binding is unlikely to arise naturally without eons of selective pressure, it is not clear whether moderate binding affinity and/or specificity might be found among collections of unevolved proteins designed in the laboratory.

These broad questions can now be addressed by direct experimentation using purified proteins isolated from libraries of *de novo* designed sequences. To probe the SM binding potential of such sequences, we assessed the binding of three *de novo* designed α-helical proteins to diverse compounds displayed on small molecule microarrays.[2] The sequences of these three *de novo* proteins are not related to known natural sequences, and they were neither selected nor designed to bind SMs. We considered three possible outcomes:

(i) SM binding is difficult to achieve without selection (or rational design): If this is true, then specific (non-promiscuous) interactions between SMs and unevolved *de novo* proteins would occur only very rarely.

(ii) The binding of SMs by naïve proteins is inherently promiscuous: This model would predict that three *de novo* proteins sharing sequence and structural similarity would bind weakly and non-specifically to the same SMs.

(iii) Specific (non-promiscuous) binding of SMs to folded protein structures is *not* a rare occurrence: If this is correct, then *de novo* proteins that were designed to fold but *not* explicitly designed to recognize SMs would nonetheless bind SMs with reasonable affinities and specificities.

To distinguish between these possibilities, we probed the SM binding capabilities of three novel proteins chosen from a combinatorial library of *de novo* sequences. In order to focus our studies on the binding capabilities of *folded* proteins, rather than unfolded, aggregated, or insoluble sequences, we used

Received: November 7, 2011
Published: March 23, 2012

   dx.doi.org/10.1021/sb200018e | *ACS Synth. Biol.* 2012, 1, 130–138

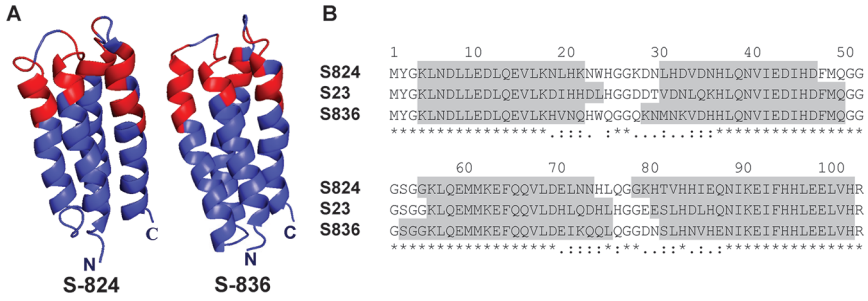


Figure 1. *De novo* proteins probed for small molecule binding. (A) Solution structures of proteins S824 and S836 (PDB codes 1P68, 2JUA). Constant residues among these proteins are shown in blue. (B) Amino acid sequences of proteins S824, S23, and S836. Residues that are identical (*) and highly similar (:) are indicated. Residues are colored according to secondary structure (α-helices in gray background).

proteins from a collection of sequences designed to fold into soluble 3-dimensional structures, rather than from libraries of random sequences, which would rarely yield well-folded structures.[3,4] Specifically, our three proteins were drawn from a library of sequences designed to fold into 4-helix bundles. As described previously, these sequences were designed using the binary code for protein design, which posits that stably folded proteins can be encoded by specifying the sequence pattern of polar and nonpolar residues (the binary pattern) to coincide with the exposed and buried parts of a structure, respectively.[5] Accordingly, a 4-helix bundle is designed to contain four stretches of the following pattern of polar (○) and nonpolar (●) residues: ○●○○●●○○●○○●●○. This pattern is consistent with the α-helical repeat of 3.6 residues/turn in an amphiphilic α-helix. Indeed, this binary pattern (or shorter segments of this pattern) is found frequently among amphiphilic α-helices of natural proteins.[6]

The three artificial proteins chosen for the current studies were S824, S836, and S23 from a second-generation library described previously.[7] These sequences were chosen for four reasons: (i) They have been structurally and thermodynamically characterized, and the 3-dimensional structures of both S824 and S836 were determined by NMR (Figure 1A).[8,9] (ii) The sequences of the three *de novo* proteins are very similar to one another (Figure 1B),[7] allowing us to ascribe differences in binding to small changes in sequence. (iii) Despite their sequence similarities, these three proteins cover a range of structural stability: S824 is extremely well-ordered; S836 has a well-defined structure but is more dynamic than S824; S23 is considerably more dynamic and resembles a molten globule.[7] (iv) None of these *de novo* sequences share significant similarity with known natural sequences.

Because these proteins were obtained from libraries of *de novo* sequences, they are not products of biological selection for SM binding. Indeed, they are not biased by any requirement to provide life-sustaining functions. Their only requirement is that they be non-toxic to *E. coli*, and readily expressed and purified. Therefore, these proteins are well suited for a first assessment of the binding potential of sequences that are folded, but unbiased by evolutionary history.

The three *de novo* proteins were screened for binding to 10,000 different compounds displayed on microarrays. Hits were confirmed, and binding constants were assessed by spectroscopic assays. The results show that these unevolved 4-helix bundle proteins indeed recognize and bind a range of SMs. Furthermore, despite their sequence and structural similarities, proteins S824, S836, and S23 display some level of selectivity: they distinguish between different compounds and bind targets with different affinities. Our results show that *de novo* proteins that were designed to fold but *not* explicitly designed to recognize SMs can nonetheless bind organic compounds with reasonable affinities and specificities. These results indicate that specific (non-promiscuous) binding between SMs and *de novo* folded proteins is not rare and can be found by screening relatively modest-sized libraries. These findings have implications for the early evolution of protein function and provide a foundation for engineering novel proteins for applications in synthetic biology.[10]

RESULTS AND DISCUSSION

Collections of *de novo* designed proteins present new opportunities for synthetic biology. Because many synthetic biological applications will require novel proteins that bind small molecules, it is important to assess whether SM binding is (i) a relatively common feature of folded proteins that can be found by medium throughput screening or (ii) a rare feature that must either be computationally designed or selected by evolution (in nature or *in vitro*). As a first step toward assessing the potential of unevolved (naïve) proteins to bind SMs, we probed the abilities of three *de novo* proteins from a combinatorial library of 4-helix bundles to bind to a diverse collection of organic compounds displayed on microarrays.

Small-Molecule Microarrays Enable High-Throughput Screening for Binding to *de novo* Proteins. We used small-molecule microarrays (SMMs) to identify compounds that bound each of our three *de novo* proteins. SMMs are glass slides on which libraries of SMs are covalently immobilized in an array of microscopic spots. A single slide typically contains nearly 10,000 compounds, allowing rapid screening of diverse libraries of SMs. SMM slides can be probed with a fluorophore- or epitope-tagged protein, and compounds that bind the protein are detected by automated fluorescence read-out.[11] SMMs have been used previously to identify compounds that bind a range of natural proteins, including calmodulin, transcriptional regulators, Alzheimer's Aβ peptide, and histone deacetylases.[12−15] For the current study, we used SMMs displaying a range of compounds, including bioactive molecules, natural products, and molecules originating from diversity-oriented syntheses.[11,16]

To detect binding to a spot on a SMM, we used variants of proteins S824, S836, and S23 that were tagged at their C-termini with the octapeptide FLAG-tag. Proteins that bound to spots on the array were visualized by probing with an anti-FLAG antibody followed by a fluorescently labeled secondary antibody. Each screen was performed in triplicate, and signal over background (DMSO controls) scores were calculated as composite Z values,[2,17,18] which are available at the ChemBank

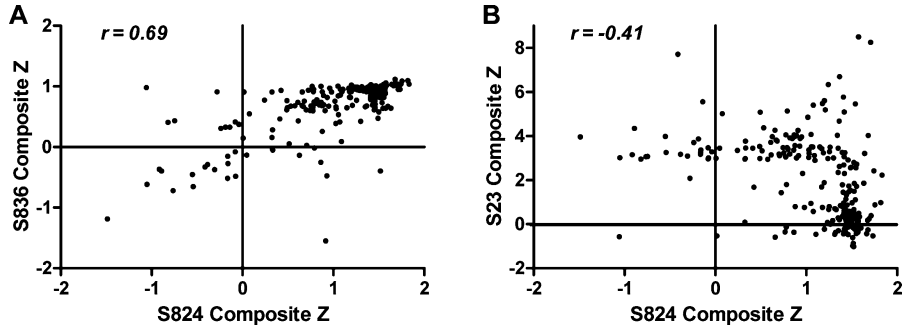


Figure 2. Comparison of the distribution of 233 compounds comprising the top 100 ranked putative binders of S824, S836, and S23 according to their composite *Z* scores. (A) S836 vs S824. (B) S23 vs S824. The correlation coefficient (*r*) between data sets is indicated. S836 and S824 appear to share many of their putative binding compounds, while protein S23 has a distinctly different profile.

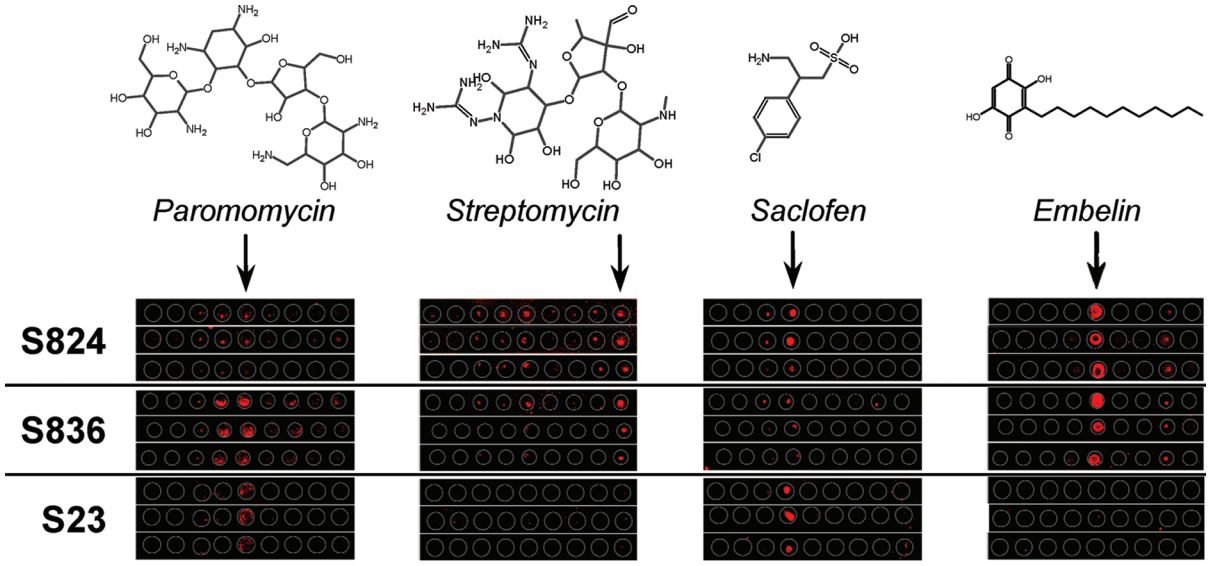


Figure 3. SMM binding assays. Arrays probed with proteins S824, S836, or S23, followed by fluorescently labeled antibody detection (Alexa-647). Arrows mark the position of the indicated compound.

database at http://chembank.broadinstitute.org/. These composite *Z* scores do not merely report signal intensity; they are normalized relative to the background across all spots on a given array and to the reproducibility of the signal between replicated plates of the same experiment. Detailed explanation of composite *Z* score calculation can be found elsewhere.[18] Compounds that were shown previously to be promiscuous binders to a range of natural proteins (according to ChemBank database) or exhibited binding in the control experiments (see Methods section) were disqualified prior to ranking. Thus, the considered SMs include only those compounds that are not promiscuous binders.

Because composite *Z* scores do not necessarily correspond to solution affinities, we focused on the top assay positives for each protein, rather than on those that scored above an arbitrary cutoff for composite *Z* score. The composite *Z* scores of the top 20 assay positives are shown in Supplementary Table S1, and their structures are presented in Supplementary Figure S1. Interestingly, among these top 20 hits, proteins S824 and S836 recognize many of the same compounds, and more than 50% of these shared hits possess aliphatic chains (Supplementary Figure S1).

A broader comparison of the binding preferences of the three *de novo* proteins is shown in Figure 2, which plots the top 100 ranked molecules for each protein (a total of 233 potential binders) against one another and presents the correlation coefficient (*r*) between data sets. These correlations, consistent with the top 20 binding preferences shown in Supplementary Table S1, show that proteins S824 and S836 recognize many of the same compounds (r = 0.69), while the molten globule structure of protein S23 seems to have a more distinctive binding profile (r = −0.41).

Despite overall similarities in the binding profiles of proteins S824 and S836, there are distinct differences in the compounds they recognize. For example, saclofen, RELII062-R1, and NP-005403 are highly ranked for S824 but do not come close to the top rankings for S836. Protein S23 differs from the other two proteins in several respects: it has a considerably less ordered (more molten globule-like) structure than S824 or S836[7] and favors a different group of compounds (Figure 2 and Supplementary Figure S1).

Estimation of Binding Affinities. SMM assays provide lists of putative binders, rather than quantitative measures of affinity. Therefore actual binding must be confirmed using biophysical assays. To provide such confirmation, we chose four commercially available compounds and measured the binding of these compounds to all three of our *de novo* proteins. To ensure that recognition was mediated by the *de novo* sequence

dx.doi.org/10.1021/sb200018e | *ACS Synth. Biol.* 2012, 1, 130–138

without contributions from the FLAG tag, these assays were performed with untagged proteins.

The structures of paromomycin, streptomycin, saclofen, and embelin are shown in Figure 3. To address the possibility that that these compounds might be generic (sticky) binders, we analyzed the ChemBank database for binding to natural proteins. (Because data in ChemBank are based on unvalidated SMM studies, we defined binding permissively, by including all compounds with composite Z scores $\geq$1.0 and within the first 500 SMM "hits.") This analysis showed that paromomycin, streptomycin, saclofen, and embelin bind very few natural proteins. Indeed, only single natural proteins, among several hundred assayed, were suggested to bind saclofen, streptomycin, or embelin. Only six natural proteins were suggested to bind paromomycin. Thus, the four compounds studied here are *not* promiscuous binders.

Results of the SMM assay for the four compounds are shown in Figure 3. The dissociation constants (K_D's) were estimated using a range of biophysical methods, as summarized in Table 1.

Table 1. Estimated Binding Constants (μM) of Four Small Molecules Binding to Three *de novo* Proteins

compound	method	S824	S836	S23
streptomycin	SPR	40 ± 5	5 ± 1	no binding observed
	PFGSE NMR		76 ± 32	650 ± 420
paromomycin	SPR	50 ± 4	8 ± 0.5	10 ± 1
	PFGSE NMR		22 ± 20	15 ± 14
embelin	CD	20 ± 7 to 60 ± 30	760 ± 330	no binding observed
	PFGSE NMR	N/A		1000 ± 520
saclofen	PFGSE NMR	41 ± 15		42 ± 17

Paromomycin is an antibiotic that inhibits protein synthesis by binding the 16S RNA of the bacterial ribosome with a K_D estimated to be ~0.1 μM.[19,20] The SMM screen identified paromomycin among the top 20 binders to protein S23. It also scored relatively high (ranked 26th) for protein S836 and near the top 20 for S824. To confirm binding and estimate affinities, we used pulse field gradient spin echo (PFGSE) NMR.[21] This method is well-suited to study the affinities between macromolecules and SMs, and we have used it previously to study binding to novel proteins.[22] PFGSE NMR enables estimation of binding affinities from the mole fraction of the protein/ligand complex, which is estimated from the observed diffusion coefficients of the ligand in the free and bound states. PFGSE experiments monitoring the interaction of paromomycin with S836 and S23 yielded K_D values in the range of 15−20 μM (Figure 4a). To confirm this binding by an orthogonal biophysical method, we used surface plasmon resonance (SPR) with paromomycin immobilized on a sensor chip. Immobilization was accomplished using EDC/NHS chemistry to capture the compound with heterogeneous display via different amines. Dissociation constants of 50, 8, and 10 μM, were estimated for S824, S836, and S23, respectively (Figure 4b and Supplementary Figure S2A), verifying that all 3 *de novo* proteins bind paromomycin.

Streptomycin, also an antibiotic, inhibits protein synthesis by binding the S12 protein of the 30S subunit of the bacterial ribosome with a K_D in the range of 1 μM.[23−25] Streptomycin was among the top 20 assay positives for protein S836, a mediocre binder for S824, and non-binder to protein S23 (Supplementary Table S1). PFGSE NMR confirmed the differential affinity for streptomycin, with moderate binding (estimated K_D, 70 μM) for S836 and weak binding (estimated K_D, 650 μM) for S23 (Figure 4A). As shown in Figure 4B and Supplementary Figure S2B, SPR verified the differential binding, with estimated K_D's of 5 μM for S836, 40 μM for S824, and no binding for S23. The difference between the K_D values for S836 estimated from SPR and PFGSE likely stem from (i) the intrinsic insensitivity of the PFGSE NMR technique and (ii) the fact that PFGSE NMR is done in solution, while SPR requires ligand immobilization. Although the two techniques yielded slightly different values for K_D, both methods confirm the *de novo* designed proteins S824 and S836 bind streptomycin in the low to mid micromolar range.

Saclofen is a competitive antagonist of the $GABA_B$ receptor, which binds with a K_D of ~10 μM.[26] In the SMM screen, it ranked among the top 20 hits for proteins S23 and S824 but did not bind S836. PFGSE NMR estimated K_D's of approximately 40 μM for both S23 and S824 binding to saclofen (Figure 4A). SPR indicated extremely weak binding, suggesting this method is not suitable for saclofen, perhaps because of its small size and/or because functional groups essential for binding were compromised by immobilization to the chip. We also attempted isothermal titration calorimetry (ITC) to assess saclofen binding; however, the very low enthalpy of this interaction precluded an accurate determination of the K_D. Thus, in the case of saclofen, we were able to estimate binding affinities to S23 and S824 by PFGSE NMR; however, we were unable to confirm these K_D's by an orthogonal technique.

Embelin is a natural product isolated from plants. It displays a variety of biological functions, including antioxidative and pro-apoptotic activities.[27,28] The natural binding partner for embelin is not known; however, it was found to bind the BIR3 domain of the XIAP (X-linked inhibitor of apoptosis) protein with a K_D estimated in the lower micromolar range ($\leq$1 μM).[29] Due to its long alkyl chain, embelin is poorly soluble in aqueous solution (<250 μM at neutral pH). In the SMM screen, embelin was among the top-20 assay positives for S824, a mediocre binder for S836, and below background for S23. Initial indications that embelin also binds S824 in solution came from observations that (i) in the presence of S824, embelin remained soluble (Figure 4C), and (ii) embelin coeluted with S824 from a gel filtration column. Binding of embelin to the *de novo* proteins was probed using PFGSE NMR (Figure 4A). As suggested by the SMM screen, S23 bound weakly, with a K_D estimated around 1 mM. In contrast, the S824−embelin interaction was too strong to measure accurately by PFGSE NMR; when the concentration of S824 was equal to or greater than that of embelin, the ^{1}H peaks corresponding to the alkyl chain of free embelin (0.74 ppm, 1.14−1.16 ppm and 2.1 ppm) were absent (Supplementary Figure S3), indicating the SM was bound with a relatively long residence time. Therefore we estimated the affinity using CD spectroscopy. Monitoring the CD signal of embelin at 334 nm (Supplementary Figure S4A) following titration into a solution of protein S824 indicates that binding occurs with equimolar stoichiometry (Supplementary Figure S4B) and an apparent K_D of 60 ± 30 μM (Figure 4D). Binding was confirmed in a reciprocal experiment, with the titration of protein S824 into embelin yielding an estimated K_D of 20 ± 7 μM (Supplementary Figure S4C). The dissociation

dx.doi.org/10.1021/sb200018e | *ACS Synth. Biol.* 2012, 1, 130−138

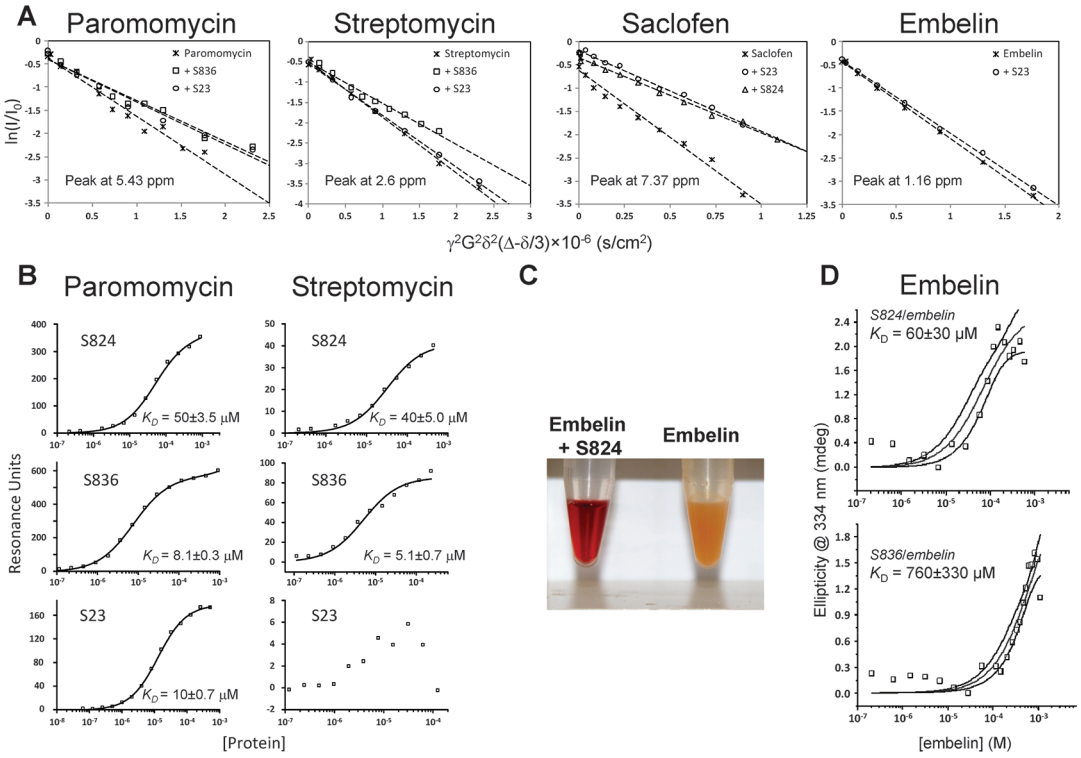


Figure 4. Biophysical characteristics of binding to *de novo* proteins. (A) ^{1}H PFGSE NMR diffusion decay plots for the SMs either alone or in the presence of the indicated proteins. Proton resonance amplitude intensities (I) at indicated δ peaks are shown. Dashed lines show the linear regression of the diffusion decay. (B) Binding isotherms for paromomycin (left) and streptomycin (right) derived from concentration-dependent SPR. Compounds immobilized on the chip surface were titrated with proteins at concentrations ranging from micromolar to millimolar in successive analyte injections at 25 °C. Instrument response value at steady state, as derived from the flat part of sensorgrams, is shown in arbitrary units after subtraction of response in a parallel mock-immobilized channel. (C) Protein S824 enhances the solubility of embelin. Embelin (2 mM) suspended in buffer (right) and in a solution of 2 mM S824 (left). (D) Binding of embelin to each protein was monitored by the CD signal at 334 nm following titration of embelin into 260 μM protein. All association curves show the best fit for a 1:1 model with the indicated dissociation constants, while side curves indicate the 95% confidence range according to Origin 7 software (OriginLab).

constant for protein S836 appeared at least 10-fold weaker than for S824 (Figure 4D), and as expected from the SMM and NMR experiments, binding to S23 was not detectable by CD (Supplementary Figure S4D).

To confirm that binding of embelin to S824 is both specific and non-covalent, we performed several additional experiments. First, we examined the possibility of non-specific binding at multiple sites in the protein. If embelin bound multiple sites, one would expect the ^{1}H NMR peaks of the SM to remain invisible at embelin/S824 ratios greater than 1. This was not the case (Supplementary Figure S3). Equimolar stoichiometry was also confirmed by CD, as noted above (Supplementary Figure S4B). Second, we assayed for formation of a covalent adduct[30] using both mass spectrometry and organic/aqueous extraction. Mass spectrometry of a S824-embelin sample confirmed the mass of the apoprotein with no trace of a covalent adduct (Supplementary Figure S5A). Next, we showed that extraction with butanone removed the embelin and left apo-S824 in the aqueous phase (Supplementary Figure S5B). Taken together, our results indicate that protein S824 binds embelin in a non-covalent one-to-one complex with an affinity in the mid micromolar range.

Binding Embelin Induces a Conformational Change in Protein S824. Because the binding of embelin was relatively specific for protein S824, we were especially interested in the structural changes that accompany this binding. ^{1}H−^{15}N HSQC NMR spectroscopy showed that binding to embelin reduced the dispersion of cross-peaks (Figure 5A) and accelerated H-D exchange (Supplementary Figure S6). The previously assigned peaks in the HSQC spectrum of S824[31] enabled identification of side chains that interact with embelin (Supplementary Figure S7). Most of the affected residues occur in the hydrophobic core of the protein, the second loop, and the chain termini (Figure 5B). (Additional nonpolar α-helical residues beyond those shown may contribute to binding; however, these resonances were difficult to reassign because of decreased resolution in the crowded region of the spectrum.) Embelin binding also seems to affect residue Asp-32. However, this residue seems too remote to contact embelin. Presumably, embelin binding causes overall changes in the structure of S824, which may be transmitted to this residue.

We were surprised to note that although protein S824 binds embelin with higher affinity than proteins S23 or S836, most of the residues involved in binding are conserved among all three sequences. Thus, the impact of changing the protein sequence is subtle. To probe the complementary effect, i.e., the impact of changing the small molecule, we measured binding to the following analogues of embelin:

- 2,5-Dihydroxy-1,4-benzoquinone (DHBQ), which contains the quinone ring by itself, did not bind S824.
- 2,5-Dihydroxy-3-heptyl-1,4-benzoquinone (DHHBQ), a derivative with a shorter alkyl chain, bound S824 with affinity ∼100-fold weaker than that of embelin (Supplementary Figure S8A).

dx.doi.org/10.1021/sb200018e | *ACS Synth. Biol.* 2012, 1, 130−138

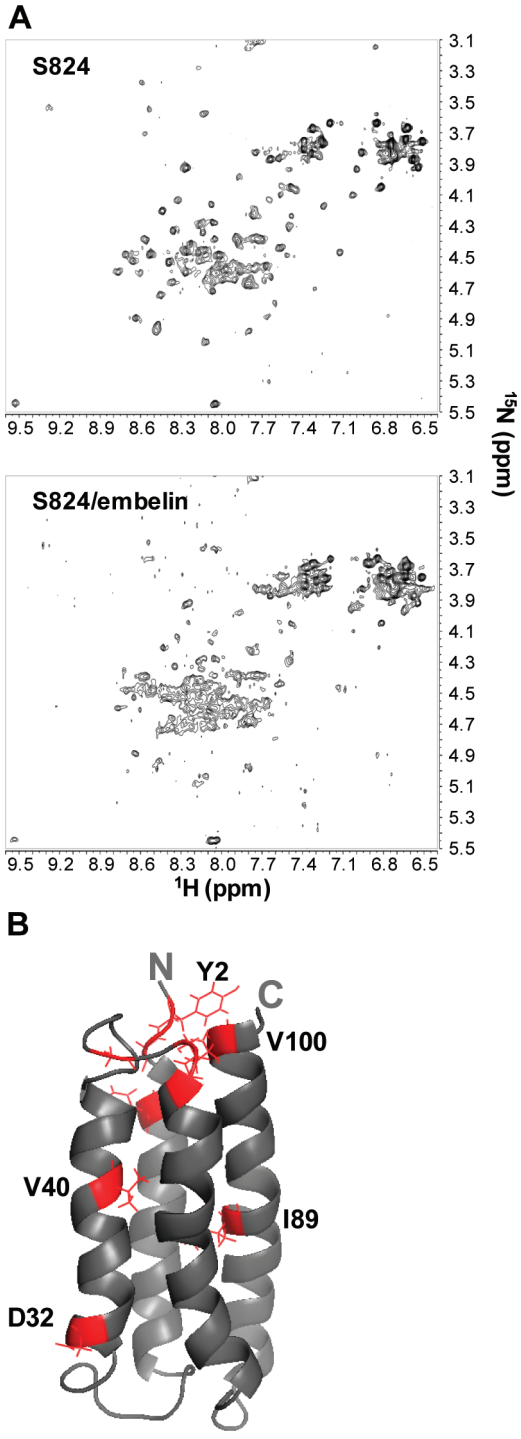


Figure 5. ^{1}H−^{15}N HSQC NMR. (A) Spectra of free S824 (0.7 mM) and S824 with embelin (1 mM). Addition of the small molecule produces a spectrum with reduced chemical shift dispersion. The loss of dispersion is particularly apparent in the crowded region between 4.4 and 4.8 ppm, corresponding to α-helical residues.[31] (B) Structure of S824 with residues affected by embelin marked in red (based on changes in chemical shifts). Residues are Y2, G3, N6, D32, V40, G52, G54, G55, I89, and V100. All but D32 are "conserved" residues. Additional changes corresponding to buried helical residues could not be assigned with confidence due to the loss of dispersion.

- 2,5-Dimetoxy-3-undecyl-1,4-benzoquinone (DMUBQ), an analogue with a modified ring, bound with a somewhat weaker affinity (K_D = 180 μM, Supplementary Figure S8B).

These results show that the nonpolar residues in the core of the protein and the alkyl chain of the SM are important for binding; however, structural features of the quinone are less important. These conclusions were further supported by the ^{1}H NMR spectra of S824/embelin complex, which showed that binding to S824 caused the 23 aliphatic protons of embelin to populate new molecular environments (Supplementary Figure S3). Taken together, these data suggest that binding is mediated primarily by insertion of the alkyl chain into the core of the protein, with the quinone acting as a placeholder on the surface of the protein. Such interactions presumably disrupt the well-packed interior of the protein and stabilize the more dynamic and less stable structure that is observed upon binding.

Similar *de novo* Proteins Have Different Binding Specificities. The three proteins examined in this study have sequences that are ∼70% identical. Not surprisingly, these proteins bind many of the same SMs. An unexpected result, however, was the finding that some compounds distinguish between the three proteins by binding with dramatically different affinities. This selectivity may be related to differences in the structural flexibility of the three proteins. As noted above, previous NMR studies showed that proteins S824 and S836 form well-ordered structures, whereas S23 is more like a dynamic molten globule. The importance of protein flexibility in dictating the binding profile is highlighted by Figure 2, which shows that the two well-ordered proteins share many SM hits in common, whereas the S23 molten globule favors a different set of SMs. Further evidence that similar sequences do not automatically lead to similar binding profiles is provided by our observation that although embelin binds S824 with an affinity much stronger than that of S23 or S836, the residues responsible for this binding occur in conserved regions of the protein sequence (Figure 5).

Structural plasticity is likely to facilitate greater promiscuity in ligand binding,[32,33] as flexible structures sample more conformations that may accommodate a wider range of binding partners. However, such binding often comes at the expense of affinity, as malleable binding sites are not preorganized and must pay an entropic cost to form a complementary binding site. The trade-off between specificity and affinity (or stability) has been described previously in several systems.[34−36] These considerations suggest that the molten globule, S23, would bind to a wider range of SMs than the more rigid proteins, S824 and S836. Indeed, S23 may bind a significant fraction of the compounds bound by S824 and S836, but with affinities that are below the detection sensitivity of the SMM screen. This type of flexible recognition with relatively low affinity may explain why the residues in S824 that contact embelin (Figure 5) are conserved among all 3 proteins.

Implications for Protein Evolution in Natural Biology and for Protein Design in Synthetic Biology. The overall agreement between the SMM screens and the biophysical measurements demonstrate that although the *de novo* proteins were neither evolved nor designed to bind SMs, they indeed recognize a range of compounds with moderate affinities and some level of specificity. These findings have implications both for the evolution of functional proteins in nature and for the design of proteins *de novo*.

In nature, functional proteins can be divided into two classes: (i) those that require only the chemical moieties provided by

dx.doi.org/10.1021/sb200018e | *ACS Synth. Biol.* 2012, 1, 130−138

the amino acid side chains and polypeptide backbone and (ii) those that require bound cofactors for activity. Both classes evolved from pools of sequences that served as the feedstock of molecular evolution; however, the differing requirements of the two classes may have led to different evolutionary trajectories. The first class of functional proteins, the purists, underwent selective pressure to fold into structures that place the side chain and backbone atoms of the active site in the precise orientations required for catalysis. The serine proteases represent a classic example of this first type of enzyme.[37] The second class of functional proteins underwent selective pressure to bind SMs that enable function. The cytochromes, which bind the heme cofactor, comprise a diverse collection of this second class of enzymes.[38]

By analogy to the evolution of functional proteins in nature, recent work on the design of functional proteins *de novo* can also be divided into two classes: those that function with[39−46] and those that function without[40,47−49] bound cofactors. While some level of success has been achieved for both classes, there are more examples of *de novo* proteins that rely on cofactors for activity. This is not surprising since using preorganized activity modules to impart function is likely to be easier than designing proteins with active sites that are preorganized into fixed locations by precise positioning of each side chain and backbone atom. Similarly, it seems likely that the early evolution of natural systems may have favored proteins with cofactor-based activities.[50,51]

If SM binding is an initial step toward protein function, then the occurrence of functional proteins early in evolution would have depended upon how frequently SM binding occurred in a population of unselected sequences. To probe the frequency of SM binding in random sequence space, Keefe and Szostak performed a pioneering study in which they screened 6×10^{12} random 80-residue sequences for ATP binding.[52] From this vast collection of sequences, they found 4 that bound ATP. While those results suggested that binding to specific SMs is difficult to achieve and occurs very rarely in unselected sequence space, two factors must be considered: (i) their library was constructed randomly, and therefore the vast majority of sequences would not be expected to fold into stable protein-like structures; and (ii) screening was performed only against one target, ATP.

In contrast to the studies by Keefe and Szostak, our experiments focus on a family of sequences that are predisposed to fold into stable structures. Our results show that non-promiscuous SM binding, with a degree of selectivity, occurs readily in a collection of unevolved folded sequences. Depending on one's perspective, this finding may or may not have been expected *a priori*; however, it could not have been probed experimentally without the availability of collections of *de novo* designed folded proteins, such as those described here. This initial study focused on one simple family of folds, the 4-helix bundle. For this family of structures, our findings support the premise that binding to small molecules occurs relatively frequently among unevolved collections of folded proteins. Because proteins that bind SMs can provide a feedstock for the optimization of specificity and function, their occurrence among unselected libraries of folded sequences suggests a mechanism for the rapid evolution of biological functions.

METHODS

Protein Expression and Purification. The plasmid for expressing epitope tagged proteins was based on the pT7-FLAG-4 vector (Sigma), which allows C-terminal tagging with the octapeptide, DYKDDDDK. Untagged proteins were expressed using a modified pCA24N vector (p3GLAR). Proteins were expressed in *E. coli* BL21 star (DE3) (Invitrogen) as described (Supporting Information).

SMM Screen. SMMs were prepared, screened, and analyzed as described.[11] Data and procedures are summarized in Supporting Information. A complete list of screened SMs and SM binding profiles can be found in the ChemBank database.[18] All screens were performed in triplicate. Detection of interactions with the FLAG-tagged proteins was performed using a monoclonal mouse anti-FLAG M2 antibody (Sigma) and AlexaFluor 647 labeled goat anti-mouse antibody (Invitrogen). Slides were scanned for fluorescence at 532 and 635 nm using a GenePix 4200A slide scanner. To avoid false positives, control analyses were also performed with FLAG peptide (instead of protein) as well as with goat anti-mouse (secondary) antibody alone or together with anti-FLAG (primary) antibody.

The signal-to-noise ratios for screens involving proteins S824 and S836 were generally lower than for S23 (maximum composite Z scores of 1.9 and 1.1, respectively). Because of the different normalization between experiments, using composite Z scores to make comparisons across different proteins is not valid.[18,53] Nonetheless, the composite Z value is a good indication for binding likelihood within each experiment: the higher the score, the greater the likelihood that binding is real. Accordingly, a negative or zero composite Z should give a good indication for no or very weak binding.

Calculation of Correlation Coefficients. The extent of a linear relationship between composite Z scores of two SMM data sets was calculated according to Pearson product moment correlation coefficient, r:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

where x and y are the composite Z scores, and $\bar{x}$ and $\bar{y}$ represent the mean average of the respective data sets. The range of the coefficient is $-1 \leq r \leq +1$, where $r = +1$ reports perfect correlation, $r = 0$ no correlation, and $r = -1$ inverse correlation.

Circular Dichroism (CD) Spectroscopy. Spectra were acquired using an AVIV 62DS spectropolarimeter equipped with temperature-controlled sample holder and a 1-mm path length cuvette. All experiments were performed in PBS, pH 7.3. Ellipticity at each wavelength was averaged for 5 s. For titration analyses, 1−2 μL aliquots of concentrated embelin stock (20 mM in DMSO) were added to 300 μL of sample. Solubilization of the SM was facilitated by bath sonication for 1−2 min.

Surface Plasmon Resonance (SPR). Affinities were evaluated using a BIAcore 3000 (BIAcore Inc.). Approximately 840 and 200 resonance units (RU) of paromomycin or streptomycin were immobilized onto a research grade sensor chip CM5 using amine coupling kit (BIAcore) as described by the manufacturer. Double dilutions of S824, S836, and S23 (at indicated concentrations in PBS, pH 7.3) were passed over the chip at a flow rate of 30 μL/min. The chip surface was regenerated after each run with 10 mM NaOH (40 μL/min for 7 s) and re-equilibrated in PBS buffer. Sensogram data were analyzed using the BIAevaluation 4.1. Steady-state equilibrium binding constants were calculated from the final RU values after binding saturation (reaching plateau) using Origin 7.0 (OriginLab) with the 1:1 Langmuir binding model.

Heteronuclear Single Quantum Coherence (HSQC) NMR. Spectra were acquired at 22 °C on a Bruker AVANCE-II 500 MHz spectrometer equipped with a TCI ($^{1}H/^{13}C/^{15}N$) cryoprobe. Typical samples for the $^{1}H-^{15}N$ HSQC experiments[54] contained 0.6−0.8 mM ^{15}N-labeled protein in 0.6 mL of buffer (10 mM phosphate, pH 7.0, 50 mM NaCl and 10% (v/v) $^{2}H_2O$ (Cambridge Isotope Laboratories)). Powdered SMs were first dissolved in water or DMSO (embelin) and then diluted in the protein sample. All samples containing embelin were further centrifuged (10 min at 13,000 rpm) to remove insoluble material. Water signals were suppressed using excitation sculpting.[55] Reassignment of protein peaks from pH 4.0 (determined previously[31]) to pH 7.0 was accomplished manually by acquiring spectra every 0.5 pH unit between 4.0 and 7.0. Measurement parameters are detailed in Supporting Information.

PFGSE NMR and Estimation of K_D. Spectra were acquired at 22 °C using the same spectrometer as above equipped with *z*-field gradient of max 53 G/cm. A stimulated echo pulse sequence was used.[21] The gradient pulse duration (δ) was set to 3 ms, and the magnetic field gradient (G) was ramped from 0.53 to 42.4 G/cm. The diffusion time (Δ) was set to 200 ms. To maximize the signal-to-noise ratio, 128 or 512 scans were run over at least 9 gradient steps, depending on the sample. Calculations of diffusion coefficients and estimation of K_D are explained in Supporting Information.

Mass Spectrometry. The S824/embelin sample (0.1 mM S824, 0.2 mM embelin in H_2O) was analyzed on an Agilent 6220 TOF LC/MS system. Sample was diluted 1:1 in 90% CH_3CN, 10% H_2O, 0.1% formic acid.

Isothermal Titration Calorimetry. ITC was performed using a MCS ITC instrument (Microcal). Concentrations were 2.3 mM saclofen and 0.9 mM S824. Fifty aliquots (2−4 μL each) of saclofen solution were automatically injected into the sample cell containing protein S824. Titrations were performed at 30 °C.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary methods, supporting figures and table. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 609-258-2901. Fax: 609-258-6746. E-mail: hecht@princeton.edu.

Author Contributions

M.H.H. and I.C. planned the project and experimental design; A.N.K. designed, printed, and analyzed the SMM screens; I.C. performed most of the experiments; M.K. participated in ^{15}N protein labeling and NMR analyses. All of the coauthors wrote or edited the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Supported by grants MCB-0817651 and MCB-1050510 from the NSF and contract N01-CO-12400 from the NCI Initiative for Chemical Genetics. We gratefully acknowledge support from the Machiah Foundation to I.C. (Fellowship 20070117). We thank Olivia M. McPherson for help with SMM experiments and analysis, Istvan Pelczer for help with NMR, and Jannette Carey and Rebecca E. Strawn for help with CD, SPR, and ITC.

■ REFERENCES

(1) Gray, H. B. (2003) Biological inorganic chemistry at the beginning of the 21st century. *Proc. Natl. Acad. Sci. U.S.A. 100*, 3563−3568.

(2) Vegas, A. J., Fuller, J. H., and Koehler, A. N. (2008) Small-molecule microarrays as tools in ligand discovery. *Chem. Soc. Rev. 37*, 1385−1394.

(3) Mandecki, W. (1990) A method for construction of long randomized open reading frames and polypeptides. *Protein Eng. 3*, 221−226.

(4) Tanaka, J., Doi, N., Takashima, H., and Yanagawa, H. (2010) Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci. 19*, 786−795.

(5) Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M., and Hecht, M. H. (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science 262*, 1680−1685.

(6) West, M. W., and Hecht, M. H. (1995) Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci. 4*, 2032−2039.

(7) Wei, Y., Liu, T., Sazinsky, S. L., Moffet, D. A., Pelczer, I., and Hecht, M. H. (2003) Stably folded de novo proteins from a designed combinatorial library. *Protein Sci. 12*, 92−102.

(8) Wei, Y., Kim, S., Fela, D., Baum, J., and Hecht, M. H. (2003) Solution structure of a de novo protein from a designed combinatorial library. *Proc. Natl. Acad. Sci. U.S.A. 100*, 13270−13273.

(9) Go, A., Kim, S., Baum, J., and Hecht, M. H. (2008) Structure and dynamics of de novo proteins from a designed superfamily of 4-helix bundles. *Protein Sci. 17*, 821−832.

(10) Nobeli, I., Favia, A. D., and Thornton, J. M. (2009) Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol. 27*, 157−167.

(11) Bradner, J. E., McPherson, O. M., and Koehler, A. N. (2006) A method for the covalent capture and screening of diverse small molecules in a microarray format. *Nat. Protoc. 1*, 2344−2352.

(12) Barnes-Seeman, D., Park, S. B., Koehler, A. N., and Schreiber, S. L. (2003) Expanding the functionay of small-molecule microarrays: discovery of novel calmodulin ligands. *Angew. Chem., Int. Ed. 42*, 2376−2379.

(13) Koehler, A. N., Shamji, A. F., and Schreiber, S. L. (2003) Discovery of an inhibitor of a transcription factor using small molecule microarrays and diversity-oriented synthesis. *J. Am. Chem. Soc. 125*, 8420−8421.

(14) Chen, J., Armstrong, A. H., Koehler, A. N., and Hecht, M. H. (2010) Small molecule microarrays enable the discovery of compounds that bind the Alzheimer's Abeta peptide and reduce its cytotoxicity. *J. Am. Chem. Soc. 132*, 17015−17022.

(15) Vegas, A. J., Bradner, J. E., Tang, W., McPherson, O. M., Greenberg, E. F., Koehler, A. N., and Schreiber, S. L. (2007) Fluorous-based small-molecule microarrays for the discovery of histone deacetylase inhibitors. *Angew. Chem., Int. Ed. 46*, 7960−7964.

(16) Duffner, J. L., Clemons, P. A., and Koehler, A. N. (2007) A pipeline for ligand discovery using small-molecule microarrays. *Curr. Opin. Chem. Biol. 11*, 74−82.

(17) Clemons, P. A., Bodycombe, N. E., Carrinski, H. A., Wilson, J. A., Shamji, A. F., Wagner, B. K., Koehler, A. N., and Schreiber, S. L. (2010) Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U.S.A. 107*, 18787−18792.

(18) Seiler, K. P., George, G. A., Happ, M. P., Bodycombe, N. E., Carrinski, H. A., Norton, S., Brudz, S., Sullivan, J. P., Muhlich, J., Serrano, M., Ferraiolo, P., Tolliday, N. J., Schreiber, S. L., and Clemons, P. A. (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res. 36*, D351−359.

(19) Fourmy, D., Recht, M. I., Blanchard, S. C., and Puglisi, J. D. (1996) Structure of the A site of *Escherichia coli* 16S ribosomal RNA complexed with an aminoglycoside antibiotic. *Science 274*, 1367−1371.

(20) Recht, M. I., Douthwaite, S., Dahlquist, K. D., and Puglisi, J. D. (1999) Effect of mutations in the A site of 16 S rRNA on aminoglycoside antibiotic-ribosome interaction. *J. Mol. Biol. 286*, 33−43.
(21) Momot, K. I., and Kuchel, P. W. (2006) PFG NMR diffusion experiments for complex systems. *Concepts Magn. Reson., Part A 28A*, 249−269.
(22) Das, A., Wei, Y., Pelczer, I., and Hecht, M. H. (2011) Binding of small molecules to cavity forming mutants of a de novo designed protein. *Protein Sci. 20*, 702−711.
(23) Birge, E. A., and Kurland, C. G. (1969) Altered ribosomal protein in streptomycin-dependent Escherichia coli. *Science 166*, 1282−1284.
(24) Chang, F. N., and Flaks, J. G. (1972) Binding of dihydrostreptomycin to Escherichia coli ribosomes: characteristics and equilibrium of the reaction. *Antimicrob. Agents Chemother. 2*, 294−307.
(25) Llano-Sotelo, B., Hickerson, R. P., Lancaster, L., Noller, H. F., and Mankin, A. S. (2009) Fluorescently labeled ribosomes as a tool for analyzing antibiotic binding. *RNA 15*, 1597−1604.
(26) Drew, C. A., Johnston, G. A., Kerr, D. I., and Ong, J. (1990) Inhibition of baclofen binding to rat cerebellar membranes by phaclofen, saclofen, 3-aminopropylphosphonic acid and related GABAB receptor antagonists. *Neurosci. Lett. 113*, 107−110.
(27) Singh, D., Singh, R., Singh, P., and Gupta, R. S. (2009) Effects of embelin on lipid peroxidation and free radical scavenging activity against liver damage in rats. *Basic Clin. Pharmacol. 105*, 243−248.
(28) Ahn, K. S., Sethi, G., and Aggarwal, B. B. (2007) Embelin, an inhibitor of X chromosome-linked inhibitor-of-apoptosis protein, blocks nuclear factor-kappaB (NF-kappaB) signaling pathway leading to suppression of NF-kappaB-regulated antiapoptotic and metastatic gene products. *Mol. Pharmacol. 71*, 209−219.
(29) Nikolovska-Coleska, Z., Xu, L., Hu, Z. J., Tomita, Y., Li, P., Roller, P. P., Wang, R. X., Fang, X. L., Guo, R. B., Zhang, M. C., Lippman, M. E., Yang, D. J., and Wang, S. M. (2004) Discovery of embelin as a cell-permeable, small-molecular weight inhibitor of XLAP through structure-based computational screening of a traditional herbal medicine three-dimensional structure database. *J. Med. Chem. 47*, 2430−2440.
(30) Fisher, A. A., Labenski, M. T., Malladi, S., Gokhale, V., Bowen, M. E., Milleron, R. S., Bratton, S. B., Monks, T. J., and Lau, S. S. (2007) Quinone electrophiles selectively adduct "electrophile binding motifs" within cytochrome *c*. *Biochemistry 46*, 11090−11100.
(31) Wei, Y., Fela, D., Kim, S., Hecht, M., and Baum, J. (2003) 1H, 13C and 15N resonance assignments of S-824, a de novo four-helix bundle from a designed combinatorial library. *J Biomol. NMR 27*, 395−396.
(32) Tapley, T. L., Korner, J. L., Barge, M. T., Hupfeld, J., Schauerte, J. A., Gafni, A., Jakob, U., and Bardwell, J. C. (2009) Structural plasticity of an acid-activated chaperone allows promiscuous substrate binding. *Proc. Natl. Acad. Sci. U.S.A. 106*, 5557−5562.
(33) Fink, A. L. (2005) Natively unfolded proteins. *Curr. Opin. Struct. Biol. 15*, 35−41.
(34) Prajapati, R. S., Indu, S., and Varadarajan, R. (2007) Identification and thermodynamic characterization of molten globule states of periplasmic binding proteins. *Biochemistry 46*, 10339−10352.
(35) Vamvaca, K., Jelesarov, I., and Hilvert, D. (2008) Kinetics and thermodynamics of ligand binding to a molten globular enzyme and its native counterpart. *J. Mol. Biol. 382*, 971−977.
(36) Bolon, D. N., Grant, R. A., Baker, T. A., and Sauer, R. T. (2005) Specificity versus stability in computational protein design. *Proc. Natl. Acad. Sci. U.S.A. 102*, 12724−12729.
(37) Campbell, M. K., Farrell, S. O. (2006) *Biochemistry*, 5th ed., Thomson-Brooks/Cole, United States.
(38) Bernhardt, R. (2006) Cytochromes P450 as versatile biocatalysts. *J. Biotechnol. 124*, 128−145.
(39) Moffet, D. A., Certain, L. K., Smith, A. J., Kessel, A. J., Beckwith, K. A., and Hecht, M. H. (2000) Peroxidase activity in heme proteins derived from a designed combinatorial library. *J. Am. Chem. Soc. 122*, 7612−7613.
(40) Patel, S. C., Bradley, L. H., Jinadasa, S. P., and Hecht, M. H. (2009) Cofactor binding and enzymatic activity in an unevolved superfamily of de novo designed 4-helix bundle proteins. *Protein Sci. 18*, 1388−1400.
(41) Faiella, M., Andreozzi, C., de Rosales, R. T. M., Pavone, V., Maglio, O., Nastri, F., DeGrado, W. F., and Lombardi, A. (2009) An artificial di-iron oxo-protein with phenol oxidase activity. *Nat. Chem. Biol. 5*, 882−884.
(42) Koder, R. L., Anderson, J. L. R., Solomon, L. A., Reddy, K. S., Moser, C. C., and Dutton, P. L. (2009) Design and engineering of an O-2 transport protein. *Nature 458*, 305−U364.
(43) Yeung, N., Lin, Y. W., Gao, Y. G., Zhao, X., Russell, B. S., Lei, L. Y., Miner, K. D., Robinson, H., and Lu, Y. (2009) Rational design of a structural and functional nitric oxide reductase. *Nature 462*, 1079−U1144.
(44) Simmons, C. R., Stomel, J. M., McConnell, M. D., Smith, D. A., Watkins, J. L., Allen, J. P., and Chaput, J. C. (2009) A synthetic protein selected for ligand binding affinity mediates ATP hydrolysis. *ACS Chem. Biol. 4*, 649−658.
(45) Fry, H. C., Lehmann, A., Saven, J. G., DeGrado, W. F., and Therien, M. J. (2010) Computational design and elaboration of a de novo heterotetrameric alpha-helical protein that selectively binds an emissive abiological (porphinato)zinc chromophore. *J. Am. Chem. Soc. 132*, 3997−4005.
(46) Bender, G. M., Lehmann, A., Zou, H., Cheng, H., Fry, H. C., Engel, D., Therien, M. J., Blasie, J. K., Roder, H., Saven, J. G., and DeGrado, W. F. (2007) De novo design of a single-chain diphenylporphyrin metalloprotein. *J. Am. Chem. Soc. 129*, 10732−10740.
(47) Wei, Y., and Hecht, M. H. (2004) Enzyme-like proteins from an unselected library of designed amino acid sequences. *Protein Eng., Des. Sel 17*, 67−75.
(48) Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F. 3rd, Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008) De novo computational design of retro-aldol enzymes. *Science 319*, 1387−1391.
(49) Rothlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., and Baker, D. (2008) Kemp elimination catalysts by computational enzyme design. *Nature 453*, 190−195.
(50) Ma, B.-G., Chen, L., Ji, H.-F., Chen, Z.-H., Yang, F.-R., Wang, L., Qu, G., Jiang, Y.-Y., Ji, C., and Zhang, H.-Y. (2008) Characters of very ancient proteins. *Biochem. Biophys. Res. Commun. 366*, 607−611.
(51) Caetano-Anolles, G., Kim, K. M., and Caetano-Anolles, D. (2012) The phylogenomic roots of modern biochemistry: Origins of proteins, cofactors and protein biosynthesis. *J. Mol. Evol. 74*, 1−34.
(52) Keefe, A. D., and Szostak, J. W. (2001) Functional proteins from a random-sequence library. *Nature 410*, 715−718.
(53) Bradner, J. E., McPherson, O. M., Mazitschek, R., Barnes-Seeman, D., Shen, J. P., Dhaliwal, J., Stevenson, K. E., Duffner, J. L., Park, S. B., Neuberg, D. S., Nghiem, P., Schreiber, S. L., and Koehler, A. N. (2006) A robust small-molecule microarray platform for screening cell lysates. *Chem Biol 13*, 493−504.
(54) Mori, S., Abeygunawardana, C., Johnson, M. O., and van Zijl, P. C. (1995) Improved sensitivity of HSQC spectra of exchanging protons at short interscan delays using a new fast HSQC (FHSQC) detection scheme that avoids water saturation. *J. Magn. Reson. B 108*, 94−98.
(55) Hwang, T. L., and Shaka, A. J. (1995) Water suppression that works−Excitation sculpting using arbitrary wave-forms and pulsed-field gradients. *J. Magn. Reson. Ser. A 112*, 275−279.