ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

SCHEMA-Designed Variants of Human Arginase I and II Reveal Sequence Elements Important to Stability and Catalysis

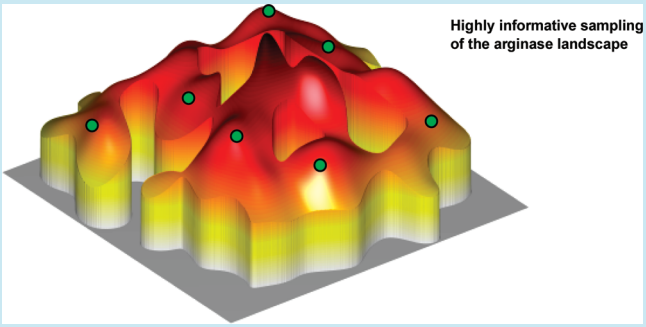
Philip A. Romero,[†,#] Everett Stone,[∥,#] Candice Lamb,[⊥] Lynne Chantranupong,[¶] Andreas Krause,[‡,§] Aleksandr E. Miklos,[△] Randall A. Hughes,[△] Blake Fechtel,[⊥] Andrew D. Ellington,[△,○] Frances H. Arnold,*[,†] and George Georgiou*[,∥,⊥,¶,△,○]

Divisions of [†]Chemistry and Chemical Engineering and [‡]Engineering and Applied Science, California Institute of Technology, Pasadena, California 91125, United States

[§]Department of Computer Science, Swiss Federal Institute of Technology, Zurich, Switzerland

Departments of [∥]Biomedical Engineering, [⊥]Chemical Engineering, [¶]Molecular Genetics and Microbiology, and [△]Chemistry and Biochemistry and [○]Institute for Cell and Molecular Biology, University of Texas, Austin, Texas 78712, United States

ABSTRACT: Arginases catalyze the divalent cation-dependent hydrolysis of L-arginine to urea and L-ornithine. There is significant interest in using arginase as a therapeutic antineogenic agent against L-arginine auxotrophic tumors and in enzyme replacement therapy for treating hyperargininemia. Both therapeutic applications require enzymes with sufficient stability under physiological conditions. To explore sequence elements that contribute to arginase stability we used SCHEMA-guided recombination to design a library of chimeric enzymes composed of sequence fragments from the two human isozymes Arginase I and II. We then developed a novel *active learning algorithm* that selects sequences from this library that are both highly informative and functional. Using high-throughput gene synthesis and our two-step active learning algorithm, we were able to rapidly create a small but highly informative set of seven enzymatically active chimeras that had an average variant distance of 40 mutations from the closest parent arginase. Within this set of sequences, linear regression was used to identify the sequence elements that contribute to the long-term stability of human arginase under physiological conditions. This approach revealed a striking correlation between the isoelectric point and the long-term stability of the enzyme to deactivation under physiological conditions.



KEYWORDS: *enzyme engineering, arginase, homologous recombination, SCHEMA library design, active learning, protein stability*

Humans produce two arginase isozymes (EC 3.5.3.1) that catalyze the hydrolysis of L-arginine (L-Arg) to urea and L-ornithine (L-Orn). The Arginase I (hArgI) gene is located on chromosome 6 (6q.23), is highly expressed in the cytosol of hepatocytes, and functions in nitrogen removal as the final step of the urea cycle. The Arginase II (hArg II) gene is found on chromosome 14 (14q.24.1). Arginase II is localized in the mitochondria in tissues such as kidney, brain, and skeletal muscle, where it is thought to provide a supply of L-ornithine (L-Orn) for L-proline and polyamine biosynthesis.[1] The two enzymes share 61% amino acid sequence identity and adopt a homotrimeric structure composed of an α/β fold consisting of a parallel eight-stranded β-sheet surrounded by several helices. These enzymes contain a dinuclear metal cluster that generates a hydroxide for nucleophilic attack on the guanidinium carbon of L-arginine.[2,3] In eukaryotes and the vast majority of prokaryotes, the native metal cofactor in arginase is believed to be Mn^{2+}.

There is significant interest in applying arginases as cancer chemotherapeutic agents. A number of high morbidity tumors such as hepatocellular carcinomas (HCCs), melanomas, renal cell, and prostate carcinomas[4−6] are deficient in the urea cycle enzyme argininosuccinate synthase (ASS) and thus are sensitive to L-arginine (L-Arg) depletion. Non-malignant cells typically enter into quiescence (G_0) when deprived of L-Arg and remain viable for several weeks. However, ASS-deficient tumor cells experience cell cycle defects that lead to the reinitiation of DNA synthesis even though protein synthesis is inhibited, in turn resulting in major imbalances that lead to rapid cell death.[7,8] The selective toxicity of L-Arg depletion for HCC, melanoma, and other urea-cycle enzyme-deficient cancer cells has been extensively demonstrated *in vitro*, in xenograft animal models, and in clinical trials.[4,5,7,9]

Additionally, rare autosomal recessive mutations in hArgI can cause hyperargininemia, which results in hyperammonemia, spasticity, seizures, and failure to thrive.[10] Dietary management in combination with oral phenylbutyrate is often successful in

Received: March 7, 2012
Published: March 30, 2012

controlling hyperammonemia, but the underlying hyperargininemia can persist, which can result in L-arginine-associated neurotoxicity.[11] Red blood cell replacement, which provides supplemental hArgI within red blood cells, has shown promise in treating hyperargininemia as evidenced by reduced serum L-Arg levels and improved clinical outcomes.[12,13]

To function as a therapeutic agent, arginase must efficiently degrade L-Arg to very low levels (<5 μM) under physiological conditions (~100 μM L-Arg, 37 °C, and pH 7.4). Unfortunately, hArgI and hArgII display low enzymatic activity at physiological pH and are rapidly inactivated in serum, with half-lives of only a few hours. Arnold and co-workers have demonstrated the utility of SCHEMA-guided recombination for generating libraries of chimeric proteins between low-homology sequences.[14,15] In an effort to understand the sequence determinants of arginase that are important for long-term stability, we designed a SCHEMA-guided recombination library composed of sequence fragments from the human arginases hArgI and hArgII (Figure 1). By coupling this SCHEMA library with a novel active learning algorithm we efficiently identified a diverse set of enzymatically active chimeric arginases. These chimeras highlighted an important correlation between isoelectric point and long-term stability, providing a key insight into how these enzymes might be further optimized for stability.

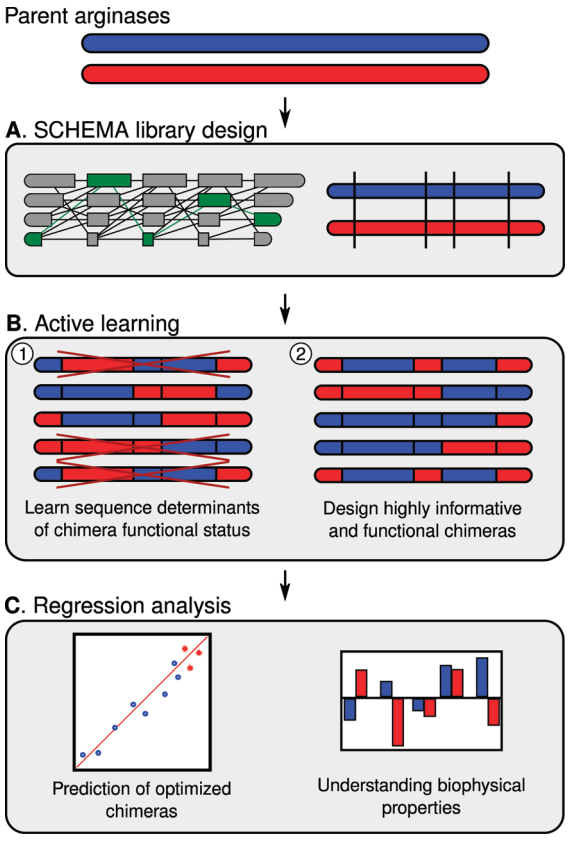


Figure 1. Overview of method. (A) Starting with the two parent arginases, we used SCHEMA (structure-guided recombination) to identify optimal recombination sites. (B) Next, a two-step active learning algorithm was used to identify a highly informative subset of this SCHEMA library. The first step of this algorithm efficiently learns which sequence elements contribute to loss of function. The second step then uses this information to design a set of chimeras that are highly informative and functional. (C) With experimental data on chimeric arginases, regression analysis can be used to make predictions across the entire library or to understand how each sequence element contributes to arginase properties.

RESULTS AND DISCUSSION

SCHEMA Library Design. When homologous proteins are recombined, new interactions between structural fragments are often deleterious to protein function. The presence of these interactions within a chimeric protein can be estimated from the SCHEMA disruption, which counts interactions that are not observed in the parents.[16] A chimera's SCHEMA disruption is calculated from the parent sequences and a residue−residue contact map representation of the protein structure. Large combinatorial libraries of chimeric proteins can be designed using the Recombination as a Shortest-Path Problem (RASPP) algorithm, which identifies the library that minimizes the average SCHEMA disruption with constraints on the number and size of sequence fragments.[17]

hArgI and hArgII share 61% amino acid sequence identity (64% nucleotide identity) and were chosen as parents for a SCHEMA recombination library. The trimeric structure of hArgII (PDB ID: 1PQ3) was used to prepare the contact map, which included both intra- and intersubunit contacts. The RASPP algorithm was used to design a library of chimeric sequences having seven recombination sites (eight sequence blocks).

The chimera blocks chosen for the arginase recombination library are illustrated in Figure 2. Within each monomer's central β-sheet, seven of eight strands came from different blocks, while the trimer interface was formed from blocks 5, 6, and 8. Substrate recognition in arginase is achieved by several loops that flank the active site and numerous water-mediated hydrogen bonds.[18] Within the chimera library, each of these "specificity" loops was located in different blocks. We believed these design choices should provide multiple opportunities for identifying more functional catalysts, especially since the residues that coordinate the catalytic binuclear manganese cluster were conserved within the library, while the surrounding, second-shell residues came from different parental combinations of blocks 3, 4, 7, and 8.

The sequences within the designed chimera library were diverse: on average, chimeras differed from one another by 60 mutations (as few as 6 and as many as 120). These chimeras were also novel: the average mutational distance between a chimera and a parent arginase was 40.3 mutations. Nonetheless, on the basis of results from earlier studies[14,15] and the average SCHEMA disruption score for the designed library ($\langle E \rangle = 16$), it was predicted that approximately half of the chimeras would be functional arginases.

Rational Generation of an Informative Set of Chimeras. While the SCHEMA algorithm limits the protein sequence space that must be explored in order to identify functional variants, the problem of deciding which proteins to construct and assay is still a challenging one. For example, in the current library there were 256 (2^8) possible chimeric arginases, and synthesis and characterization of all these possibilities would have been daunting. Systematically chosen chimera sets are more effective than randomly chosen ones,[19] but the criterion for selection are very much open to discussion. Selecting chimeras that equalize the representation of each parent at each block position will not generate a maximally informative set of proteins,[15] primarily because of the significant proportion of nonfunctional sequences that provide

dx.doi.org/10.1021/sb300014t | ACS Synth. Biol. 2012, 1, 221−228

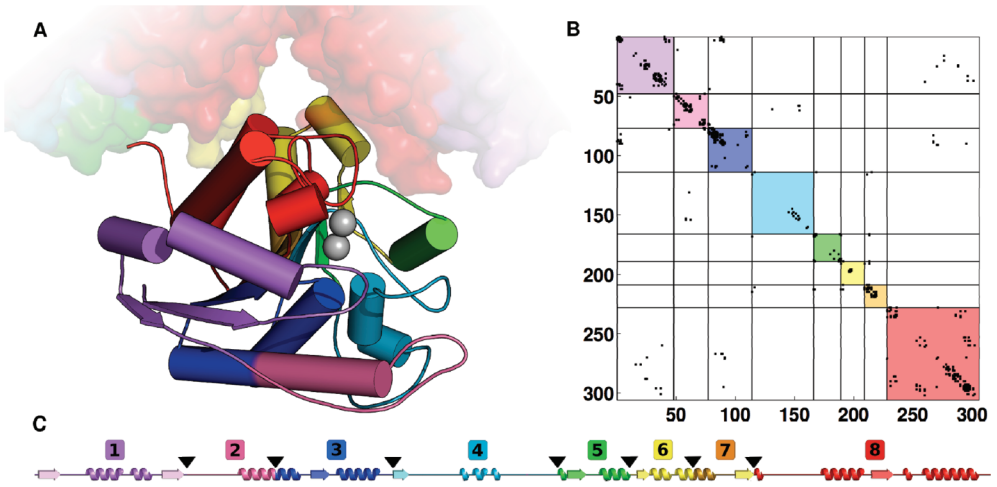


Figure 2. Arginase chimera library block boundaries. (A) Arginase three-dimensional structure with blocks represented by different colors. The trimer interface is shown as a transparent surface. (B) Contact map displaying residue−residue contacts that could be broken upon recombination. The colored squares correspond to the block divisions of the library. (C) Secondary structure diagram showing the chimera library block divisions.

no information about functional properties. Such nonfunctional sequences can be avoided by making single block perturbations, which better avoid major, disruptive interactions. However, these one-factor-at-a-time designs closely resemble the wild-type parents and thus limit the sequence and functional diversity of the data sets produced.[20] To balance these considerations, we developed a two-step active learning algorithm that efficiently identifies an informative set of functional chimeras by first training a model that can predict if a chimera will form a functional protein and then using this functional status classifier to guide an experimental design.

The first step of the algorithm involved finding an informative set of chimeras for a logistic regression classifier that models the probability that a chimera will form a functional protein. Here, we quantify the "informativeness" of a set of chimeras as the mutual information between that set and the remainder of the library (see Methods). Intuitively, this mutual information measures how much observing a given set of chimeras reduces the uncertainty (Shannon entropy) of prediction for the remainder of the library. On the basis of these criteria, we initially chose to study a set of eight arginase chimeras that maximized this mutual information criterion. The genes encoding these eight chimeras (Table 1, SCHEMA A−H) were synthesized and expressed (see Methods). As expected, approximately half (3/8) of the sequences produced functional arginases. With the functional status of these sequences now defined, it proved possible to train a Bayesian logistic regression model to predict the probability of functioning for all chimeras within the library.

The second step of the algorithm then consisted of finding a highly informative set of functional chimeric arginases. We used the predictions from the logistic regression model to select sequences that maximized the *expected value* of the mutual information between the chosen set and the remainder of the library (see Methods). This criterion should have simultaneously identified sequences that were both informative and had a high probability of being functional. A set of four additional chimeras was chosen that maximized the expected value of the mutual information. Significantly, when these gene sequences were synthesized and expressed, they were all found to encode functional enzymes that hydrolyzed L-Arg at significant rates (Table 1, SCHEMA I−L).

Table 1. Chimeric Arginase Data[a]

name	chimera blocks	*AUC*	AUC_{Mn}	T_m (°C)	k_{cat}/K_m (mM^{-1} s^{-1})
hArgI	11111111	2350	5326	81.0	130 ± 20
hArgII	22222222	4042		80.6	114 ± 18
SCHEMA A[b]	11112122				
SCHEMA B	12122211	3664	2927	81.2	19 ± 7
SCHEMA C	11221221	3872	5838	68.1	53 ± 10
SCHEMA D[b]	12211212				
SCHEMA E[b]	21121212				
SCHEMA F[b]	21212211				
SCHEMA G[b]	22111221				
SCHEMA H	22221111	3089	4109	68.5	31 ± 7
SCHEMA I	21122121	1654		67.5	27 ± 11
SCHEMA J	11222112	4710	6188	70.7	42 ± 8
SCHEMA K	22121121	1311	2655	78.9	27 ± 10
SCHEMA L	21222111	2828		70.5	19 ± 7
SCHEMA M	12122122	4005		71.2	45 ± 11
SCHEMA N	12222112	3901		70.6	36 ± 5
SCHEMA O	11122222	1026	3108	82.5	138 ± 19
SCHEMA P	21121111	1417		74.3	39 ± 10

[a]SCHEMA A−L are the designed set of highly informative sequences, SCHEMA M and N are the sequences used to validate the regression model, and SCHEMA O and P are two additional chimeras that were generated during this study. [b]No protein expression detected.

Overall, the active learning algorithm efficiently identified a highly informative set of nine functional arginases (two parents and seven chimeras). Within this set of chimeric sequences, each parent at each block was typically observed multiple times, and 103 of the 112 possible sequence block pairs were observed. Some blocks (such as block 4 parent 1) were underrepresented, presumably because they contributed to loss of function and were therefore avoided in the second step of the sequence selection algorithm.

Regression Model for Long-Term Stability. We used the highly informative set of chimeras to explore sequence−function relationships within the arginase library. In particular, the temporal inactivation of all nine enzymes within the designed set of sequences was measured (see Methods). Because the chimeras displayed either exponential, sigmodial, or biphasic decay of activity, for ease of comparison we derived

dx.doi.org/10.1021/sb300014t | *ACS Synth. Biol.* 2012, 1, 221−228

each chimera's normalized area under the inactivation curve (*AUC*), which provides a measure of a chimera's overall kinetic stability (Table 1). A Bayesian linear regression model was used to correlate sequence fragments with the experimentally measured *AUC* values (see Methods). This model resulted in an excellent fit (r = 0.98, Figure 3A), and the block regression parameters are given in Table 2.

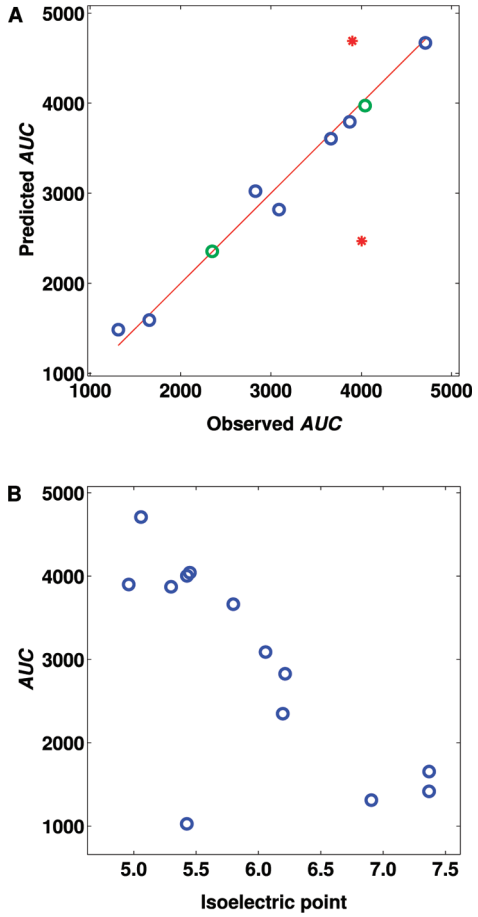


Figure 3. Arginase long-term stability. (A) Bayesian linear regression model for *AUC*. Green and blue circles correspond to the parents and chimeras (respectively) within the initial data set (r = 0.98 and $p = 9 \times 10^{-7}$). Red stars represent the model's predictions on the validation set. (B) Correlation between isoelectric point and *AUC* for all chimeras tested (r = −0.74 and p = 0.004).

Table 2. Regression Model Parameters[a]

parameter name	log *AUC*
reference	7.76
B1P2	−0.23
B2P2	0.00
B3P2	**0.39**
B4P2	0.02
B5P2	0.07
B6P2	0.32
B7P2	−0.26
B8P2	0.20

[a]The parameters specify how substituting parent 2 for parent 1 at a given block changes the logarithm of the *AUC*. The most significant substitution occurs at block 3, which is highlighted in bold.

To validate the linear regression model we designed two additional chimeric arginases (SCHEMA M and N) that were predicted to have enhanced long-term stability. These sequences were synthesized and characterized. The regression model showed good predictive ability (Figure 3A), and both sequences were more stable than 80% of the other chimeric arginases.

From the regression analysis, the most stabilizing sequence element was found to be block 3, where substituting hArgI for hArgII is estimated to increase the *AUC* by almost 50%. Closer inspection of the amino acid sequences for this important chimera block revealed an abundance of charged residues. Consistent with this observation, we found the estimated isoelectric point[21,22] of the chimeras to show a striking negative correlation (r = −0.74, p = 0.004) with the *AUC*, Figure 3B. Thus, chimeras with greatest net negative charge under the assay conditions (pH 7.4 and 37 °C) were the most stable, while those closer to their isoelectric point exhibited faster inactivation.

Metal Dependence of Stability. To test if metal binding affected thermal unfolding, the melting temperature (T_m) for all sequences was measured in the presence and absence of a chelator (see Methods; Table 1). The melting temperatures showed no significant correlation with long-term stability (r = −0.30, p = 0.31). However, as expected the addition of EDTA resulted in lower thermodynamic stability for all enzymes, with an average decrease in T_m values of 15 ± 5 °C, in close agreement with results of similar chelation experiments with beef liver and *Saccharomyces cerevisiae* arginases.[23,24] This highlights that bound Mn^{2+} stabilizes the correctly folded state under thermal equilibrium conditions. To determine if metal chelation is also a key factor in long-term kinetic stability at thermodynamically stable conditions, we measured stability in the presence of excess manganese (500 μM $MnCl_2$) (AUC_Mn in Table 1) at 37 °C, far below the average T_m of 74 ± 6 °C. In accord with the denaturation data, excess manganese was shown to increase long-term stability while maintaining the overall trend in long-term stability as a function of isoelectric point. These findings indicate that the enzymes may be stabilized by excess charge as a function of p*I* and that metal stabilizes the correctly folded form of the enzyme. As activity is dependent on active-site bound metal, it is clear that excess manganese will drive the equilibrium toward the metal-bound (active), folded state and delay irreversible inactivation. A possible mechanism of inactivation is depicted in Figure 4. Here, arginase is irreversibly inactivated by loss of metal followed by protein unfolding/aggregation.

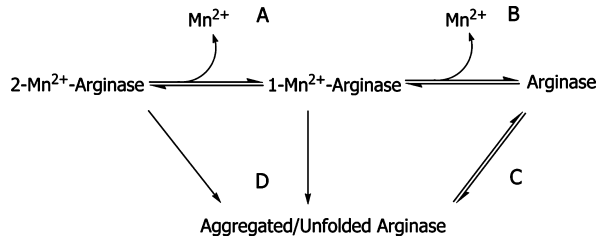


Figure 4. Schematic of potential arginase inactivation mechanisms. (A) Loss of first equivalent of bound metal and decrease of some activity. (B) Loss of second equivalent of bound metal and loss of all activity. (C) Equilibrium between folded and unfolded states. (D) Probably irreversible precipitation/aggregation.

dx.doi.org/10.1021/sb300014t | *ACS Synth. Biol.* 2012, 1, 221−228

For all chimeric arginases, we performed Michaelis–Menten kinetic measurements (see Methods) and calculated the resulting catalytic efficiencies (k_{cat}/K_M) (Table 1). Intriguingly, the fold stabilization upon addition of 500 μM $MnCl_2$ (AUC_{Mn}/AUC) displayed a linear relationship with catalytic efficiency (r = 0.85, p = 0.02), Figure 5. A similar trend has been observed within a set of Cu^{2+} complexes.[25] In that study, the authors found the stability of a Cu^{2+} complex to be inversely related to its rate of glycine methyl ester hydrolysis, indicating that more stable complexes lower the Lewis acidity of the Cu^{2+} ion. Likewise, arginases that bind Mn^{2+} more tightly (i.e., that are not as dependent on an excess of Mn^{2+} for long-term activity) may have reduced Lewis acidity for coordinating substrate or water ligands, and therefore diminished catalytic efficiency.

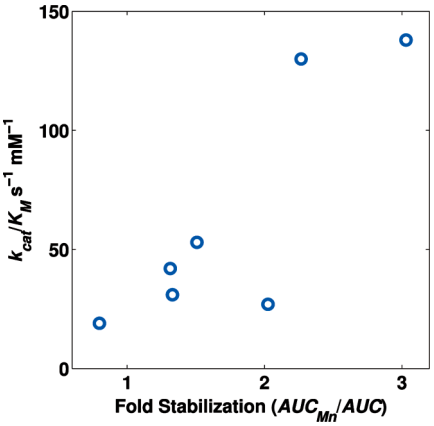


Figure 5. Correlation between fold change in long-term stability (AUC_{Mn}/AUC) and catalytic efficiency (k_{cat}/K_M), r = 0.85 and p = 0.02.

Summary and Conclusions. The combination of structure-guided SCHEMA recombination and an efficient active learning procedure was used to generate a highly informative set of catalytically active chimeric arginases. Site-directed recombination libraries between low homology parental genes provide unique data sets for probing sequence–function relationships, offering distinct advantages over sets of point mutants or naturally existing proteins. The effects of point mutations are frequently too small to resolve experimentally, while the large numbers of neutral mutations in naturally existing proteins make it difficult to pinpoint the basis of functional differences. In contrast, libraries of chimeric proteins contain an intermediate level of sequence diversity, and mutational changes are observed in multiple sequence backgrounds.

The resulting set of chimeric human arginases displays measurable variation, and the sequence basis of this variation can be efficiently identified using linear regression. The high level of sequence diversity within the hArg chimeras translates into extraordinary functional diversity, as evidenced by the fact that many of the measured properties were outside the range displayed by the two parents (Table 1). For example, recombination of hArgI and hArgII (pI = 6.8 and 5.7, respectively) generated a set of functional chimeras with isoelectric points ranging from 5.5 to 7.5. A linear regression model helped identify a strong negative relationship between a chimeric arginase's isoelectric point and its long-term stability (r = −0.74, p = 0.004). Since the long-term stability experiments were performed at physiological pH (7.4), chimeras with the greatest net charge (low pI) displayed the greatest stability. Similar relationships between a protein's net charge and its stability have been observed previously; for example, a large survey across multiple protein families found many proteins to be less stable near their isoelectric point.[26] Similarly, engineered ribonuclease variants show decreased solubility and increased aggregation near their isoelectric point.[27,28]

The relationships ferreted out in this study have practical consequences for protein engineering. Arginase inactivation is strongly linked to the loss of the metal center of arginase, as activity as well as structural/thermal stability are metal-dependent. Given the mechanism of inactivation depicted in Figure 4, it might be further hypothesized that stability issues could be resolved by (a) engineering proteins with increased numbers of negative surface charges;[29] (b) increasing the concentration of metal (likely therapeutically intractable); or (c) introducing a different metal that might lead to improved binding and hence stability.

With respect to the latter hypothesis, it is noteworthy that we have recently reported that Co^{2+}-substituted hArgI (Co-hArgI) displays a dramatically reduced K_M for L-Arg relative to the native Mn^{2+}-containing enzyme and has a 12-fold increase in k_{cat}/K_M. More importantly, Co-hArgI is significantly more stable in serum, with an inactivation half-life of more than 30 h.[30] The improved pharmacological properties of Co-hArgI have been shown to mediate potent tumor cytoxicity against numerous cancer cell lines *in vitro* and to lead to the inhibition of hepatocellular and pancreatic carcinomas in the mouse xenograft model.[30,31]

In the case of arginase replacement therapy to treat hyperargininemia, it would be preferred to reduce elevated serum L-Arg levels that range from 600 to 900 μM[32] to normal reference values of 50–150 μM[33] rather than completely eliminate the amino acid from the bloodstream. The ideal enzyme for this application would have exceptional long-term stability but not necessarily the increased efficiency that the Co-substituted enzyme shows. The SCHEMA J variant (blocks 11222112) identified in this study has a stable linear decay rate of only 1% per hour and thus may hold promise for therapeutic purposes. A simple kinetic model based on substrate hydrolysis rates, inactivation rates, and L-Arg replenishment estimates suggests that a single dose of SCHEMA J could maintain L-Arg levels in hyperargininemia patients within the normal range for 5 days longer than a single dose of the more active but less stable Co^{2+}-loaded hArgI.[34,35]

Such a treatment option is especially interesting for a number of reasons. As the SCHEMA J variant comprises two human arginases, only the three chimeric junctions represent potential new T-cell epitopes. Using software from the Immune Epitope Database Analysis Resource,[36–38] we analyzed each of these sequence junctions for any significant changes in predicted epitope binding relative to the parent sequences for the eight most common HLA alleles (see Methods). Calculations for the HLA-DRB1*15:01 allele for the second junction suggested a 3- and 3.5-fold increase in binding affinity relative to hArgI and hArgII, respectively; all other junctions and alleles did not show a significant change relative to the parental sequences, suggesting that SCHEMA J is not likely to be highly immunogenic. Moreover, since hArgI has been under investigation as an anti-neoplastic agent, its serum retention time has already been pharmacologically optimized via

PEGylation, resulting in dose-dependent L-Arg depletion in rats for up to days at a time,[39] and thus methods for further extending the lifetime of the chimera may already exist.

Overall, the ability to design enzymes that are customized to specific reaction conditions is of significant interest to biomedical science. SCHEMA recombination coupled with an active learning algorithm provided a diverse and efficient sampling of the protein fitness landscape, revealing features that could not be observed by traditional biochemical methods. These data sets therefore provide a unique opportunity to explore the relationships between protein sequence and protein function, quickly yielding fundamental principles that can be used to engineer highly optimized protein sequences.

METHODS

Active Learning Algorithm. The active learning algorithm consists of a two-step experimental design. The first step involves finding an informative set of chimeras for a logistic regression functional status model. Here, we would like to find the set of sequences that maximize the mutual information between the chosen set of chimeras S and the remainder of the library $L\backslash S$, which is given by

$$I(S; L\backslash S) = H(L\backslash S) - H(L\backslash S|S)$$

where $H(L\backslash S)$ is the Shannon entropy of library L excluding the chimeras in subset S, and $H(L\backslash S|S)$ is the entropy of the same sequences after the chimeras in S have been observed. We approximate the intractable entropy of the Bayesian logistic regression model by replacing the logistic response with a Gaussian likelihood. With this approximation, the properties of collections of sequences and their relationships can be represented with a multivariate Gaussian distribution, and their Shannon entropy can be calculated from the determinant of the covariance matrix. Gaussian mutual information is a submodular set function[40] and therefore can be efficiently maximized using a greedy approximation algorithm.[41] We used a greedy algorithm to find a set of sequences S with maximized mutual information. The functional status of the resulting sequences was then used to train a Bayesian logistic regression model that can predict the probability of functioning for all chimeras in the library.

The second step of the algorithm consists of finding a highly informative set of functional chimeric arginases. Here, we want to find the set of chimeras S that maximize the expected value of the mutual information

$$\mathrm{E}[I(S; L\backslash S)] = \sum_{A\in\mathcal{P}(S)} \left[I(A; L\backslash A) \prod_{c\in A} p_c \prod_{c\in(S\backslash A)} 1 - p_c \right]$$

where the sum is over all subsets A in the power set of S, and p_c is the predicted probability of being functional for chimera c from the logistic regression model. This objective is chosen to simultaneously find sequences that are informative and have a high probability of being functional, similar to the most informative positive (MIP) active learning algorithm.[42] Since submodular functions are closed under positive linear combinations, the expected value of the Gaussian mutual information is also submodular, and therefore greedy maximization provides strong performance guarantees. The covariance between sequences was calculated using the chimera-block coding scheme described in the Regression Analysis section (below). All experimental designs were performed with the Submodular Function Optimization Matlab Toolbox.[43]

Gene Synthesis and Cloning. Genes encoding the SCHEMA designed arginase chimeras were synthesized from oligonucleotides as described previously.[44] In brief, long DNA oligonucleotides (99 bases) were synthesized in-house and assembled into two 560-base pair fragments using inside-out PCR. These primary fragments were combined without purification in a secondary overlap-extension reaction that formed the final desired 1086-base pair product. Custom software directed the assembly schemes and the efficient reuse of oligonucleotides across multiple related sequences; 32-base pair overlaps were designed between adjacent oligonucleotides and a 35-base pair overlap was designed between the two primary fragments. Genes were synthesized with an N-terminal 6x His tag followed by a tobacco etch virus protease cleavage site and NcoI and *Eco*RI restriction sites as described previously.[30] These genes were cloned into a pET28a expression vector, and the sequences were verified using DNA sequencing.

Two variants (SCHEMA O and SCHEMA P) were not designed by the algorithm but were chosen from preliminary experiments on the basis of regions of sequence homology. These chimeras were constructed by overlap extension PCR and are included in this study as they contain SCHEMA identified blocks from hArgI and hArgII.

Expression and Purification. *E.coli* cells expressing arginase variants were grown at 37 °C in minimal media to an OD_{600} of 0.8−1. Cells were collected by centrifugation, resuspended in fresh minimal media containing 0.5 mM IPTG and 100 μM $MnSO_4$, and incubated for an additional 8−12 h at 37 °C with shaking. After protein expression, cells were collected by centrifugation, lysed using a French pressure cell, and centrifuged at 14,000*g* for 20 min at 4 °C. The clarified cell lysate was applied to a nickel IMAC column and washed with 10−20 column volumes of IMAC buffer, and the purified arginases were eluted with IMAC elution buffer (50 mM $NaPO_4$, 250 mM imidazole, 300 mM NaCl, pH 8). The purified arginases were buffer exchanged several times into PBS, 10% glycerol, pH 7.4 using a 10,000 MWCO centrifugal filter device (Amicon). Aliquots of purified arginase variants were then flash frozen in liquid nitrogen and stored at −80 °C.

Enzyme Kinetics. Michaelis−Menten kinetics for L-Arg hydrolysis were determined in 100 mM HEPES buffer at 37 °C, pH 7.4 as previously described.[30]

Long-Term Stability. The long-term stability of the arginase chimeras was measured in 100 mM HEPES buffer, pH 7.4 at 37 °C, with or without 500 μM $MnCl_2$. Proteins were diluted to 2 μM with 100 mM HEPES, pH 7.4 and placed at 37 °C. Aliquots of 30−50 μL were taken at different time points (typically t = 0, 0.5, 3, 24, 48, and 72 h). The activity at each time point was immediately measured using 1 mM L-Arg, as described previously.[30] The data were plotted as percent activity as a function of time, and the area under this inactivation curve (AUC) was calculated using Kaleidagraph. The data was also fit to various models to calculate the rates of decay of activity over time: (i) for biphasic decay: % Act = $(((100\% - amp\%)e^{-kt}) + amp\%)/(1 + e^{-hs(T_{0.5}-t)})$ where t = time, amp = amplitude of the first decay, k = the rate of exponential decay, hs= hill slope, and $T_{0.5}$ = the half-life of the sigmoidal decay; for sigmoidal decay: % Act = $100\%/(1 + e^{(-hs(T_{0.5}-t))})$, and finally a single exponential decay model was used for some enzymes as described in the results section.

dx.doi.org/10.1021/sb300014t | *ACS Synth. Biol.* 2012, 1, 221−228

Thermal Stability. Arginase variants (20−40 μM) in PBS, pH 7.4 with or without EDTA (10 mM final concentration) were incubated in 96-well low-profile PCR plates (Fisher Scientific, Rockford, IL) on ice for 30 min. SYPRO orange dye (Life Technologies, Grand Island NY) was added into each well immediately before the plate was placed in an RT (real-time)-PCR machine (LightCycler 480, Roche, Mannheim Germany). The temperature dependence of protein unfolding between 20 and 95 °C was measured in at least duplicate experiments. T_m values were derived from the monophasic melting curves curves. To determine the circular dichroic spectra, a 6 μM sample of hArgII in a 100 mM phosphate buffer, pH 7.4 was analyzed on a Jasco J-815 CD spectropolarimeter. The change in molar ellipticity at 222 nm (θ_{222}) was monitored from 25 to 90 °C. The fraction of denatured protein at each temperature was calculated by the ratio of $[\theta_{222}]/[\theta_{222}]_d$ where $[\theta_{222}]_d$ is the molar ellipticity of the completely unfolded protein. The resulting data were fit to a modified logistic equation to determine the thermal transition midpoint.

Regression Analysis. For regression models, the independent variable corresponded to chimera sequences and is represented with a binary vector x, where x_i indicates the parent identity at block i. Because of our limited data, we used Bayesian parameter estimation, which outperforms maximum likelihood estimation for small data sets.

A chimera's binary functional status was modeled with a Bayesian logistic regression model, which contains a Bernoulli likelihood function and a zero-mean, isotropic Gaussian prior on coefficients.[45] The resulting posterior distribution was approximated using Laplace's method and prior variance was estimated from the data by maximizing the marginal likelihood function. Using Newton's method, we found the maximum a posteriori (MAP) estimates for each chimera block's contribution to functionality. The probability that a chimera is functional was estimated by applying the MAP parameter estimates to the logistic model.

The logarithm of a chimera's long-term stability (*AUC*) was modeled with a Bayesian linear regression model, which consists of a Gaussian likelihood function with a zero-mean, isotropic Gaussian prior on coefficients.[45] The measurement noise and prior variance were estimated from the data by maximizing the marginal likelihood function. With these hyperparameters, MAP estimates for each chimera block's contribution to long-term stability were found in closed-form.

Immunogenicity Calculations. We used software from the Immune Epitope Database (IEDB) (consensus method for MHC(II) binding)[46] to evaluate peptides spanning 15 residues on either side of the hArgI and hArgII junctions of the SCHEMA J variant (blocks 11222112) to compare with the corresponding sequences from the hArgI and hArgII parents. Using the predicted binding constants for the 8 most common HLA alleles as reported previously[47] we then calculated the ratio of the predicted binding values for each (hArgI/SCHEMA J and hArgII/SCHEMA J) peptide for each HLA allele to assess any significant changes relative to both parents.

AUTHOR INFORMATION

Corresponding Author

*E-mail: frances@cheme.caltech.edu; gg@che.utexas.edu.

Author Contributions

[#]These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This project was supported by grants (#HF0032 and F-1654) from TI3D/Welch Foundation and National Institutes of Health (CA 139059). In addition, this work was supported by the National Security Science and Engineering Faculty Fellowship (FA9550-10-1-0169), and L.C. was supported by a fellowship from the Arnold & Mabel Beckman Foundation. The authors also acknowledge the National Institutes of Health, ARRA (grant R01-GM068664 to FHA) for funding SCHEMA library design, and the U.S. Army Research Office, Institute for Collaborative Biotechnologies (grant W911NF-09-D-0001 to FHA) for funding the regression analysis work. These contents are solely the responsibility of the authors and do not necessarily represent the official views of the sponsors.

ABBREVIATIONS

L-Arg, L-arginine; L-Orn, L-ornithine; hArgI, human Arginase I; hArgII, human Arginase II; T_m, Melting temperature; p*I*, isoelectric point; *AUC*, area under the curve

REFERENCES

(1) López, V., Alarcón, R., Orellana, M. S., Enríquez, P., Uribe, E., Martínez, J., and Carvajal, N. (2005) Insights into the interaction of human arginase II with substrate and manganese ions by site-directed mutagenesis and kinetic studies. Alteration of substrate specificity by replacement of Asn149 with Asp. *FEBS J. 272*, 4540−4548.

(2) Cama, E., Emig, F. A., Ash, D. E., and Christianson, D. W. (2003) Structural and functional importance of first-shell metal ligands in the binuclear manganese cluster of arginase I. *Biochemistry 42*, 7748−7758.

(3) Dowling, D. P., Di Costanzo, L., Gennadios, H. A., and Christianson, D. W. (2008) Evolution of the arginase fold and functional diversity. *Cell. Mol. Life Sci. 65*, 2039−2055.

(4) Ensor, C. M., Holtsberg, F. W., Bomalaski, J. S., and Clark, M. A. (2002) Pegylated arginine deiminase (ADI-SS PEG20,000 mw) inhibits human melanomas and hepatocellular carcinomas in vitro and in vivo. *Cancer Res. 62*, 5443−5450.

(5) Feun, L. G., Marini, A., Landy, H., Markoe, A., Heros, D., Robles, C., Herrera, C., and Savaraj, N. (2007) Clinical trial of CPT-11 and VM-26/VP-16 for patients with recurrent malignant brain tumors. *J. Neuro-Oncology 82*, 177−181.

(6) Yoon, C.-Y., Shim, Y.-J., Kim, E.-H., Lee, J.-H., Won, N.-H., Kim, J.-H., Park, I.-S., Yoon, D.-K., and Min, B.-H. (2007) Renal cell carcinoma does not express argininosuccinate synthetase and is highly sensitive to arginine deprivation via arginine deiminase. *Int. J. Cancer 120*, 897−905.

(7) Shen, L.-J., Beloussow, K., and Shen, W.-C. (2006) Modulation of arginine metabolic pathways as the potential anti-tumor mechanism of recombinant arginine deiminase. *Cancer Letters 231*, 30−35.

(8) Scott, L., Lamb, J., Smith, S., and Wheatley, D. N. (2000) Single amino acid (arginine) deprivation: rapid and selective death of cultured transformed and malignant cells. *Br. J. Cancer 83*, 800−810.

(9) Ascierto, P. A., Scala, S., Castello, G., Daponte, A., Simeone, E., Ottaiano, A., Beneduce, G., De Rosa, V., Izzo, F., Melucci, M. T., Ensor, C. M., Prestayko, A. W., Holtsberg, F. W., Bomalaski, J. S., Clark, M. A., Savaraj, N., Feun, L. G., and Logan, T. F. (2005) Pegylated arginine deiminase treatment of patients with metastatic melanoma: results from phase I and II studies. *J. Clin. Oncol. 23*, 7660−7668.

(10) Jain-Ghaia, S., Sreenath Nagamanic, S. C., Blasera, S., Siriwardenaa, K., and Feigenbaum, A. (2011) Arginase I deficiency: Severe infantile presentation with hyperammonemia: More common than reported? *Mol. Genet. Metab. 104*, 107−111.

(11) Segawa, Y., Matsufuji, M., Itokazu, N., Utsunomiya, H., Watanabe, Y., Yoshino, M., and Takashima, S. (2011) A long-term survival case of arginase deficiency with severe multicystic white matter and compound mutations. *Brain Dev. 33*, 45−48.
(12) Sakiyama, T., Nakabayashi, H., Shimizu, H., Kondo, W., Kodama, S., and Kitagawa, T. (1984) A successful trial of enzyme replacement therapy in a case of argininemia. *Tohoku J. Exp. Med. 142*, 239−248.
(13) Mizutani, N., Hatakawa, C., Maehara, M., and Watanbe, K. (1987) Enzyme replacement therapy in a patient with hyperargininemia. *Tohoku J. Exp. Med. 151*, 301−307.
(14) Otey, C. R., Landwehr, M., Endelman, J. B., Hiraga, K., Bloom, J. D., and Arnold, F. H. (2006) Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol. 4*, e112.
(15) Heinzelman, P., Snow, C. D., Wu, I., Nguyen, C., Villalobos, A., Govindarajan, S., Minshull, J., and Arnold, F. H. (2009) A family of thermostable fungal cellulases created by structure-guided recombination. *Proc. Natl. Acad. Sci. U.S.A. 106*, 5610−5615.
(16) Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L., and Arnold, F. H. (2002) Protein building blocks preserved by recombination. *Nat. Struct. Biol. 9*, 553−558.
(17) Endelman, J. B., Silberg, J. J., Wang, Z.-G., and Arnold, F. H. (2004) Site-directed protein recombination as a shortest-path problem. *Protein Eng., Des. Sel. 17*, 589−594.
(18) Shishova, E. Y., Di Costanzo, L., Emig, F. A., Ash, D. E., and Christianson, D. W. (2009) Probing the specificity determinants of amino acid recognition by arginase. *Biochemistry 48*, 121−131.
(19) Li, Y., Drummond, D. A., Sawayama, A. M., Snow, C. D., Bloom, J. D., and Arnold, F. H. (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol. 25*, 1051−1056.
(20) Heinzelman, P., Komor, R., Kannan, A., Romero, P. A., Yu, X., Mohler, S., Snow, C. D., and Arnold, F. H. (2010) Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Eng., Des. Sel. 23*, 871−880.
(21) Nelson, D. L., Cox, M. M. (2005) *Lehninger Principles of Biochemistry*, Vol. *1*, 4th ed., W. H. Freeman, New York.
(22) Subramaniam, S. (1998) The biology workbench - A seamless database and analysis environment for the biologist. *Proteins 32*, 1−2.
(23) ROSSI, V., GRANDI, C., DALZOPPO, D., and FONTANA, A. (1983) Spectroscopic study on the structure and stability of beef liver arginase. *Int. J. Pept. Protein Res. 22*, 239−250.
(24) Green, S., Ginsburg, A., Lewis, M., and Hensley, P. (1991) Roles of metal ions in the maintenance of the tertiary and quaternary structure of arginase from Saccharomyces cerevisiae. *J. Biol. Chem. 266*, 21474.
(25) Nakon, R., Rechani, P. R., and Angelici, R. J. (1974) Copper(II) complex catalysis of amino acid ester hydrolysis. A correlation with complex stability. *J. Am. Chem. Soc. 96*, 2117−2120.
(26) Alexov, E. (2003) Numerical calculations of the pH of maximal protein stability. *Eur. J. Biochem. 271*, 173−185.
(27) Shaw, K. L., Grimsley, G. R., Yakovlev, G. I., Makarov, A. A., and Pace, C. N. (2001) The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. *Protein Sci. 10*, 1206−1215.
(28) Schmittschmitt, J. P., and Scholtz, J. M. (2003) The role of protein stability, solubility, and net charge in amyloid fibril formation. *Protein Sci. 12*, 2374−2378.
(29) Lawrence, M. S., Phillips, K. J., and Liu, D. R. (2007) Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc. 129*, 10110−10112.
(30) Stone, E. M., Glazer, E. S., Chantranupong, L., Cherukuri, P., Breece, R. M., Tierney, D. L., Curley, S. A., Iverson, B. L., and Georgiou, G. (2010) Replacing Mn(2+) with Co(2+) in human arginase I enhances cytotoxicity toward l-arginine auxotrophic cancer cell lines. *ACS Chem. Biol. 5*, 333−342.
(31) Glazer, E. S., Stone, E. M., Zhu, C., Massey, K. L., Hamir, A. N., and Curley, S. A. (2011) Bioengineered human arginase I with enhanced activity and stability controls hepatocellular and pancreatic carcinoma xenografts. *Transl. Oncol. 4*, 138−146.
(32) Crombez, E. A., and Cederbaum, S. D. (2005) Hyperargininemia due to liver arginase deficiency. *Mol. Genet. Metab. 84*, 243−251.
(33) Rodriguez, P. C., and Ochoa, A. C. (2008) Arginine regulation by myeloid derived suppressor cells and tolerance in cancer: mechanisms and therapeutic perspectives. *Immunol. Rev. 222*, 180−191.
(34) Stone, E., Glazer, E. S., Chantranupong, L., Cherukuri, P., Breece, R. M., Tierney, D. L., Curley, S. A., Iverson, B. L., and Georgiou, G. (2010) Replacing Mn2+ with Co2+ in human Arginase I enhances cytotoxicity towards L-arginine auxotrophic cancer cell lines. *ACS Chem. Biol. 5*, 333−342.
(35) Stone, E., Chantranupong, L., Gonzalez, C., O'Neal, J., Rani, M., VanDenBerg, C., and Georgiou, G. (2011) Strategies for optimizing the serum persistence of engineered human Arginase I for cancer therapy. *J. Controlled Release 158*, 171−179.
(36) Bui, H. H., Sidney, J., Peters, B., Sathiamurthy, M., Sinichi, A., Purton, K. A., Mothé, B. R., Chisari, F. V., Watkins, D. I., and Sette, A. (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. *Immunogenetics 57*, 304−314.
(37) Nielsen, M., Lundegaard, C., and Lund, O. (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinf. 8*, 238.
(38) Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., Braxenthaler, M., Gallazzi, F., Protti, M. P., and Sinigaglia, F. (1999) Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol. 17*, 555−561.
(39) Cheng, P. N.-M., Lam, T.-L., Lam, W.-M., Tsui, S.-M., Cheng, A. W.-M., Lo, W.-H., and Leung, Y.-C. (2007) Pegylated recombinant human arginase (rhArg-peg5,000mw) inhibits the in vitro and in vivo proliferation of human hepatocellular carcinoma through arginine depletion. *Cancer Res. 67*, 309−317.
(40) Krause, A., and Guestrin, C. (2007) Near-optimal observation selection using submodular functions, in *Proceedings of the 22nd National Conference on Artificial Intelligence*, pp 1650−1654, AAAI Press, Vancouver.
(41) Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978) An analysis of approximations for maximizing submodular set functions—I. *Math. Programming 14*, 265−294.
(42) Danziger, S. A., Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G. W., Kaiser, P., and Lathrop, R. H. (2009) Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning. *PLoS Comput. Biol. 5*, e1000498.
(43) Krause, A. (2010) SFO: A toolbox for submodular function optimization. *J. Mach. Learn. Res. 11*, 1141−1144.
(44) Cox, J. C., Lape, J., Sayed, M. A., and Hellinga, H. W. (2007) Protein fabrication automation. *Protein Sci. 16*, 379−390.
(45) Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*, 1st ed., Springer, New York.
(46) Wang, P., Sidney, J., Dow, C., Mothé, B., Sette, A., and Peters, B. (2008) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput. Biol. 4*, e1000048.
(47) Cantor, J. R., Yoo, T. H., Dixit, A., Iverson, B. L., Forsthuber, T. G., and Georgiou, G. (2011) Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proc. Natl. Acad. Sci. U.S.A. 108*, 1272−1277.