CHAPTER SIXTEEN

Efficient Sampling of SCHEMA Chimera Families to Identify Useful Sequence Elements

Pete Heinzelman*, Philip A. Romero†, Frances H. Arnold†,1
*Department of Chemical, Biological & Materials Engineering, University of Oklahoma, Norman, Oklahoma, USA
†Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA
1Corresponding author: e-mail address: frances@cheme.caltech.edu

Contents

Abstract

SCHEMA structure-guided recombination is an effective method for producing families of protein chimeras having high sequence diversity, functional diversity, and thermostabilities greater than any of the parent proteins from which the chimeras are made. A key feature of SCHEMA chimera families is their amenability to a "sample, model, and predict" operation that allows one to characterize members of a small chimera sample set and use those data to construct models that accurately predict the properties of every member of the family. In this chapter, we describe applications of this "sample, model, and predict" approach and outline methods for designing chimera sample sets that enable efficient construction of models to identify useful sequence elements. With these models we can also predict the sequences and properties of the most desirable chimeras.

Methods in Enzymology, Volume 523
ISSN 0076-6879
http://dx.doi.org/10.1016/B978-0-12-394292-0.00016-3

1. INTRODUCTION

SCHEMA structure-guided recombination is an effective method for producing large families of enzyme chimeras having high sequence and functional diversity. The chimeras are made by recombining a set of homologous parent proteins at crossover locations specifically chosen to minimize structural disruption. We have shown that members of these chimera families can have thermostabilities and maximum catalytic temperatures (T_{opt}) higher than those of any of the parent enzymes while retaining high catalytic activity (Heinzelman et al., 2009; Li et al., 2007; Smith et al., 2012). Additionally, for chimera families in which the residues that impact catalytic activity and substrate specificity are not highly conserved across the parent enzymes, it has proven possible to generate chimeras that are simultaneously thermostable and have substrate specificity profiles that are distinct from those of the parents (Li et al., 2007).

The ability to identify the sequences of the most desirable chimeras in a given family by using predictive modeling approaches contributes greatly to SCHEMA recombination's utility as a protein engineering tool. Such modeling allows one to design and construct a small sample set of chimera sequences (perhaps a few dozen), characterize their properties, and then use those data to predict the sequences of the chimera family members that have the most desirable property profiles. In this era of rapid and inexpensive gene synthesis, the construction of highly informative chimera sample sets has become accessible to virtually every laboratory. In this chapter, we describe some successful applications of this "sample, model, and predict" approach, whose main steps are illustrated and described in Fig. 16.1. We also outline methods for designing SCHEMA chimera family sample sets which with relatively moderate time and labor inputs can translate to the accurate prediction of dozens of useful new chimera sequences.

2. SCHEMA CHIMERA FAMILY DESIGN OVERVIEW

SCHEMA chimera families are constructed by recombining contiguous stretches of amino acids, or "blocks," taken from (structurally related) protein homologs, or "parents." In SCHEMA recombination, the crossover locations are chosen to maximize the number of chimeras that will be folded and functional. The design of SCHEMA chimera families uses the recombination as a shortest path problem (RASPP) algorithm to identify blocks

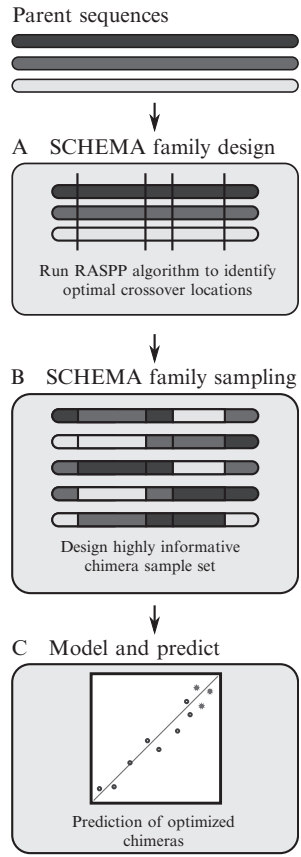

Figure 16.1 Schematic of "sample, model, and predict" approach for a SCHEMA chimera family. Once parent genes are chosen, the SCHEMA software is used to design a chimera family containing P^B unique sequences, where B is the number of blocks and P is the number of parent enzymes (Step A). A two-step active learning algorithm is then used to design a highly informative sample set of chimeras. Genes corresponding to these chimeras are synthesized, the chimeras are expressed in a recombinant host, and the properties of interest are measured experimentally (Step B). Linear regression is used to construct a model that can predict the sequences of the chimeras with the most desirable properties. Genes corresponding to these improved chimeras can then be synthesized, or constructed by standard cloning procedures, and the forecasted property improvements validated experimentally. This prediction and validation process is depicted in the bottom panel (Step C), where the abscissa denotes the predicted value for the chimera property of interest, for example, thermostability, while the experimentally measured property values appear on the ordinate. The lighter, starred points in this plot illustrate chimeras that are predicted and experimentally validated to be most desirable with respect to the chimera property that was modeled, for example, thermostability.

that minimize the number of amino acid side chain interactions, or contacts, that are disrupted when the blocks are swapped to generate new sequences (Endelman, Silberg, Wang, & Arnold, 2004). A given chimera is characterized by the number of parental contacts that have been broken, *E*. RASPP seeks to minimize the library average broken contacts $<E>$, which increases the number of enzyme chimera family members that are folded and functional (Meyer, Hochrein, & Arnold, 2006). The greater the sequence identity among the parents, the fewer the contacts that can be broken when they are recombined. Of course, the sequence diversity in the chimera family increases with the sequence diversity of the parents. Therefore, chimera sequence diversity and the fraction of chimeras in a library that are functional tend to trade off. The mutation level of a given chimera is measured by the number of amino acids that are different from the closest parent sequence, *m*. The library average mutation level $<m>$ is a measure of sequence diversity.

The SCHEMA algorithm source code, which is publicly available (http://www.che.caltech.edu/groups/fha/Software.htm), requires as inputs an alignment of the homologous parent sequences, a minimum desired block size, the number of blocks into which the parents are to be divided, and a set of crystal structure or structure model coordinates for one of the parent enzymes. A contact map, or listing of all the amino acid pairs with side chain heavy atoms lying within 4.5 Å of one another, is generated by the source code. Although the majority of SCHEMA families to date have been constructed by dividing three parents into eight blocks, there are no fundamental constraints on either the number of blocks or the number of parents; for example, active fungal cellobiohydrolase class I (CBH I) chimeras with improved thermostabilities have been obtained by recombining eight blocks from each of five parents (Heinzelman et al., 2010). However, the required sampling for model building and prediction increases with the number of blocks and parents (see Section 3.1).

Most SCHEMA enzyme families have been constructed from such parents having ~60–75% primary sequence identity; as many as half of the chimeras of such parents have been found to be functional (Otey et al., 2006). SCHEMA recombination can also be applied to parents with much lower sequence identities, although the fraction of sequences that encode functional enzymes is expected to decrease. More than 20% of the members of a beta-lactamase chimera family constructed from such parents sharing just 34–42% identity were found to be catalytically active (Meyer et al., 2006).

3. PREDICTION OF THERMOSTABLE CHIMERAS BY LINEAR REGRESSION MODELING

It has been demonstrated that the blocks comprising a chimera make linearly additive contributions to the chimera's thermostability (Heinzelman et al., 2009a; Li et al., 2007; Smith et al., 2012) as well as its temperature optimum for catalytic activity, T_{opt} (Smith et al., 2012). Thus, linear regression can be used to construct quantitative models that can accurately predict the thermostabilities of all of the members of the chimera family, where the chimeras with the greatest thermostability or the highest T_{opt} values are typically of particular interest. Constructing such a model requires thermostability measurements, for example, in the form of T_{50} values (the temperature at which 50% of enzyme activity is lost after incubation for a specified time interval), for an appropriate sample set of chimeras. As few as 35 chimera T_{50} measurements enabled construction of an accurate predictive linear regression model for a 6561-member cytochrome P450 chimera family (Li et al., 2007).

As shown in Eq. (16.1), linear thermostability models allow the T_{50} value for a given chimera to be expressed as the sum of the T_{50} for a reference parent sequence (T_{P1}) and contributions of each of its blocks ($B=1$–8 for eight-block recombination), which can be positive, negative, or zero, that are derived from the other parents ($P=2, 3$ for three-parent recombination).

$$T_{50} = T_{P1} + \sum_{B}\sum_{P} A_{BP} Q_{BP} \qquad [16.1]$$

A_{BP} is a dummy variable coded such that if a chimera contains block 1 from parent 2, $A_{12}=1$ and $A_{13}=0$. The regression coefficients Q_{BP} represent the thermostability contributions of the blocks A_{BP} relative to corresponding blocks from the reference parent, P1. Since the thermostability contributions of the blocks from P1 are accounted for in the T_{P1} term, no $A_{BP}Q_{BP}$ addition or subtraction is made to a chimera's predicted T_{50} value for block positions at which P1 appears. The predictive accuracy of these linear models can be determined using cross-validation.

Linear regression modeling was first applied in the context of cytochrome P450 chimeras to predict chimera thermostability (Li et al., 2007). T_{50} measurements were made for a sample set of 184 catalytically active chimeras chosen from a large pool of clones that featured all or most of the possible $3^8=6561$ unique sequences that can be constructed by

recombining eight blocks from three parents. Linear regression provided a model with good correlation between observed and predicted T_{50} values (10-fold cross-validated $R^2 = 0.73$). This model also provided high correlation ($R^2 = 0.90$) between observed and predicted T_{50} values for a set of 20 new P450 chimeras not included among the first 184 that were tested. Most importantly, the model was 100% accurate in predicting the sequences of chimeras with T_{50} values greater than those of any of the three parent enzymes, at least among the 11 tested. The predicted most stable chimera was in fact more stable than any other chimera tested, with a T_{50} value 9.5 °C higher than that of the most thermostable parent (Fig. 16.2).

In addition to giving rise to new thermostable P450 chimeras, SCHEMA recombination yielded new P450 enzymes with high sequence and functional diversity. The 184 sampled chimeras differed from each other by an average of 46 amino acid mutations and contained up to 99 mutations relative to the closest parent enzyme. This sequence diversity led to useful new activities. In particular, the thermostable P450 chimeras were able to hydroxylate drug compounds on which none of the parent enzymes were active (Landwehr et al., 2007; Sawayama et al., 2009).

Many fewer than 184 measurements were required to build useful predictive models of the P450 thermostability. In fact, linear regression models based on just 35 of the data points from the set of 184 measurements accurately predicted the T_{50} values of the P450 chimeras in the 20 chimera test set referenced above (Li et al., 2007). A considerable amount of cloning effort is required to construct a complete or near complete chimera family such as the one from which the randomly chosen P450 sample set chimeras were selected. Furthermore, at least half of these chimeras were expected to be unfolded and not functional. Thus, we have been considering alternative approaches to generating sample sets. Taking advantage of modern total gene synthesis capabilities we can rapidly and cost effectively obtain a small number of sample set chimera genes with sequences that have been designed to provide a high level of information content for use in linear regression modeling.

3.1. Chimera sample set design approaches

The construction and characterization of designed SCHEMA chimera family sample sets were first applied in the context of fungal cellobiohydrolase class II (CBH II) enzymes, processive glycosyl hydrolases that play a key role in industrial biomass-to-fuel conversion processes (Heinzelman et al., 2009).

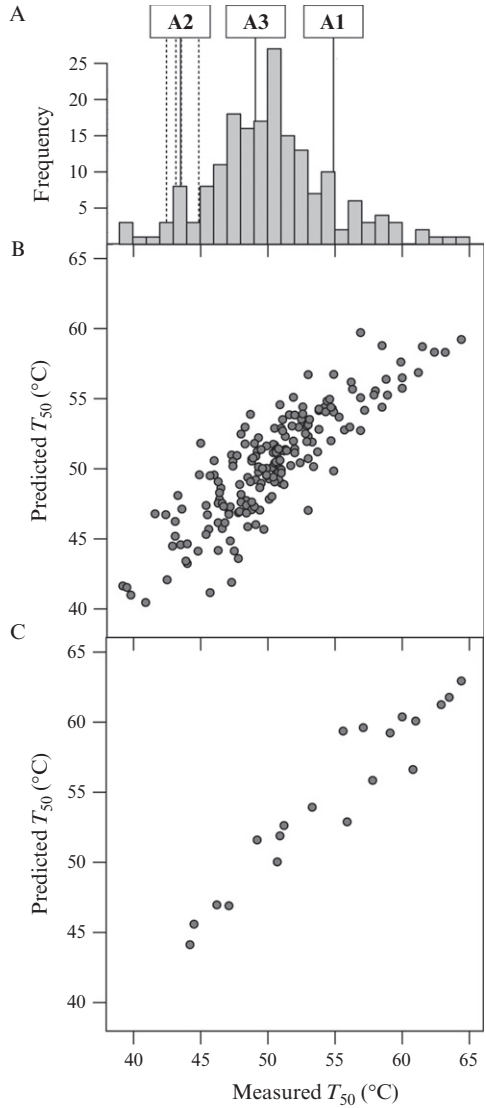

Figure 16.2 Sample, model, and predict allows accurate identification of chimeras that are more stable than the parent enzymes, illustrated here for a family of cytochrome P450 chimeras (Li et al., 2007). (A) Panel shows the distribution of T_{50} values for 184 chimeric cytochrome P450s, with T_{50}s for the three parent enzymes A1, A2, and A3 indicated (solid lines). Four experimental replicate measurements for parent A2 are shown to illustrate the high level of T_{50} measurement reproducibility that was achieved (dotted lines, standard deviation of 1.0 °C). (B) P450 chimera blocks make additive contributions to thermostability, allowing the use of linear regression to construct an accurate predictive model. Panel shows the good correlation between measured T_{50} values and T_{50} values predicted by an ordinary linear regression model ($R^2 = 0.73$). (C) Linear regression model from (B) accurately predicts T_{50} values for 20 new thermostable P450 chimeras ($R^2 = 0.90$), including the chimera predicted to be the most thermostable family member (top right-most point).

Gene synthesis company DNA 2.0 (Menlo Park, CA) provided a sample set of 48 synthetic CBH II chimera genes, which were codon-optimized for the expression host organism, *Saccharomyces cerevisiae*. To maximize the utility of the sample set chimera T_{50} dataset for constructing a linear regression model to predict chimera thermostability, the sample set was designed to provide equal representation to each of the three parents at each of the eight block positions. Specifically, 16 of the 48 chimeras featured block 1 (B1) from parent 1 (P1), 16 featured block 1 from parent 2 (P2), and 16 featured block 1 from parent 3 (P3). This equal representation was achieved by using each respective parent as a background in 16 of the 48 sample set chimeras and then substituting blocks from either one or both of the other two parents at three of the eight block positions.

Although this sample set design provided equal representation to each parent at each block position, it was not guided by any structural criteria or design algorithm. This approach was found to have a considerable drawback in that 25 of the 48 chimeras were not secreted by the *S. cerevisiae* expression host. This result is consistent with the expectation that about half of the 6561-member chimera family would be nonfunctional. Block 4 from parent 2 (B4P2) was particularly detrimental to CBH II chimera secretion; only 1 of the 16 sample set chimeras containing this block was secreted and active, the remaining 15 represented wasted gene synthesis effort. (We note here that all 23 of the CBH II chimeras that were secreted were also catalytically active.)

Measured thermostabilities of the 23 secreted chimeras showed that the sample set already contained CBH II chimeras that were more thermostable than any of the parents. This dataset, however, did not allow us to produce a quantitative linear model for predicting the thermostabilities of every member of the CBH II chimera family due to insufficient representation of some blocks in the sample set. Subsequent synthesis and characterization of an additional 41 CBH II chimeras, 31 of which were secreted (Heinzelman et al., 2009; Heinzelman et al., 2009a), led to the construction of an accurate (10-fold cross-validated $R^2 = 0.88$) linear regression model for predicting chimera T_{50} values. Although this was a highly desirable outcome, it appeared likely that the time, labor, and resources, specifically the number of synthesized genes required to build this model, could have been markedly reduced by choosing chimera sample set members that also had a high probability of being secreted. (We observed that more stable chimeras also tended

to be more highly secreted, consistent with what has been reported for heterologous expression of other proteins in *S. cerevisiae*; Shusta, Kieke, Parke, Kranz, & Wittrup, 1999.)

The aim of maximizing the fraction of synthesized chimera genes corresponding to catalytically active enzymes was pursued in the context of a SCHEMA arginase chimera family built by recombining eight blocks from this enzyme's two human isoforms, both of which are candidate cancer therapeutics (Romero et al., 2012). To minimize the number of synthesized chimera genes needed to construct an accurate model for predicting arginase stability in human serum, a two-step active learning algorithm was employed.

The first step was to find an "informative" set of chimeras for use in constructing a logistic regression model that could predict the probability that a given arginase chimera would be catalytically active. A key objective in designing this "informative" sample set was to achieve adequate representation of each block from all of the parent enzymes so that we could accurately assess each block's impact on chimera expression.

The second step of the active learning algorithm then used this predictive model to identify a set of chimeras that were both highly informative, that is, provided adequate representation of each block from all of the parents and were likely to be functional. Characterizing the chimeras in this second sample set provided data that were used to construct a model for predicting the chimeras that were most desirable with respect to the enzyme property of interest, which in the case of the arginases was stability in human serum at physiological temperatures. In the following sections, we detail the procedures for applying the two-step active learning algorithm to construct accurate regression models for predicting the properties of every member of a SCHEMA enzyme chimera family.

3.1.1 Members of chimera families are represented with a binary vector $\mathbf{x}$, which codes for the parent identity at every block position. The covariance matrix of any collection of chimeras is given by the dot product between all pairs of sequences

$$\Sigma_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j \qquad [16.2]$$

3.1.2 The first step of the active learning algorithm involves finding a set of sequences S that maximizes the mutual information between that candidate set and the other members of the chimera family ($L \backslash S$)

$$I(S; L \backslash S) = H(L \backslash S) - H(L \backslash S | S) \quad [16.3]$$

where $H(L \backslash S)$ is the entropy of the other members of the family, and $H(L \backslash S \mid S)$ is the entropy of the same sequences after the candidate set S has been observed. The mutual information for a given candidate sample set quantifies how much evaluating the properties of the members of the candidate sample set reduces the uncertainty (Shannon entropy) in predictions for the other members of the family.

Due to the fact that the Shannon entropy for a logistic regression model cannot be calculated in closed form, the logistic response is approximated with a Gaussian likelihood. This approximation allows the properties of collections of sequences and their relationships to be represented with a multivariate Gaussian distribution. Calculating the mutual information requires finding the covariance matrix of the sequences in $L \backslash S$

$$\Sigma_{L \backslash S} = X_{L \backslash S} {X_{L \backslash S}}^{\mathrm{T}} \quad [16.4]$$

where $X_{L \backslash S}$ is composed of all $\mathbf{x}$ in $L \backslash S$ ($X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^{\mathrm{T}}$) and T is the matrix transpose. The mutual information calculation also requires finding the covariance matrix of the sequences in $L \backslash S$ conditioned on S

$$\Sigma_{L \backslash S | S} = X_{L \backslash S} {X_{L \backslash S}}^{\mathrm{T}} - X_{L \backslash S} {X_S}^{\mathrm{T}} \left(X_S {X_S}^{\mathrm{T}} \right)^{-1} X_S {X_{S \backslash L}}^{\mathrm{T}} \quad [16.5]$$

With these covariance matrices, the Shannon entropy is given by

$$H = \frac{1}{2} \ln \left[(2\pi \mathrm{e})^k |\Sigma| \right] \quad [16.6]$$

where k is the dimensionality of the covariance matrix. The mutual information is the difference between the entropies before and after the sequences in S are observed (Eq. 16.3).

3.1.3 Having defined the mutual information objective function of Eq. (16.3), we now carry out an operation to find a set of sequences that maximizes it. Gaussian mutual information is a submodular set function (Krause & Guestrin, 2007) and therefore can be efficiently maximized using a greedy approximation algorithm that sequentially selects the most informative sequence (Nemhauser, Wolsey, & Fisher, 1978). To perform the greedy optimization, the mutual information of every sequence in the chimera family is evaluated and the sequence with the highest mutual information is used in the next iteration. Next, the mutual information of this sequence and all other sequences in the family is evaluated, and the set of two sequences with the highest mutual information is chosen. This greedy sequence selection process is repeated until the sample set has the desired number of sequences. For large chimera families, the speed of the greedy algorithm can be significantly increased by using "lazy" evaluations (Minoux, 1978). We typically use this algorithm to select a chimera sample set containing a number of sequences that is equal to the number of parameters in the logistic regression model. The MATLAB Toolbox for Submodular Function Optimization has useful functions for calculating Gaussian mutual information and maximizing submodular functions using accelerated greedy algorithms (Krause, 2010). We note here that greedy maximization of the objective function in Eq. (16.3) will not always find the global optimum but is guaranteed to find a solution near the optimum (Nemhauser et al., 1978). We also note that for every solution there are multiple, equivalent sample sets with the same level of mutual information.

3.1.4 After one of the maximally informative sample sets of chimera sequences has been identified, the corresponding genes are synthesized and expressed in a recombinant host. The chimera genes' functional status is then determined. This set of sequence/functional status data is then used to train a Bayesian logistic regression model that can predict the probability of forming a catalytically active enzyme (or otherwise functional protein) for all of the chimeras in the family. Chapter 4 of Bishop (2006) provides a detailed account of how to perform Bayesian logistic regression using the Laplace approximation, and specific methods we have used in applying Bayesian regression to the analysis of chimera families are further discussed in Section 3.4.

For the arginase SCHEMA chimera family, the first step of the active learning algorithm identified a set of eight (out of $2^8 = 256$ total chimeras in the family) arginase chimera sequences possessing maximum information content. As noted above, performing iterations of the greedy algorithm would have yielded additional sample sets with identical information content but differences in sequences for at least some of the chimeras contained therein. After identifying one of the maximally informative chimera sample sets, in particular, the sample set that was returned after the first time that the greedy algorithm was executed, codon-optimized genes encoding these eight arginase chimeras were synthesized and expressed in a recombinant *Escherichia coli* host. Three of the eight sample set chimeras were functional arginases. These data were sufficient to train a Bayesian logistic regression model for predicting the probability of function for each of the 256 chimeras in the family.

3.1.5 The second step of the active learning algorithm consists of identifying a set of highly informative and functional chimeras. The predictions from the logistic regression model can be used to calculate the expected value of the mutual information between the chosen set of sequences and the other members of the chimera family

$$E[I(S;L\backslash S)] = \sum_{A\in\mathcal{P}(S)} \left[I(A;L\backslash A) \prod_{\mathbf{c}\in A} p_{\mathbf{c}} \prod_{\mathbf{c}\in(S\backslash A)} 1-p_{\mathbf{c}} \right], \qquad [16.7]$$

where the sum is over all subsets A in the power set of S and $p_{\mathbf{c}}$ represents the predicted probability of being functional for chimera $\mathbf{c}$ from the catalytic activity logistic regression model.

3.1.6 Chimera sample sets with maximized expected mutual information can be identified by applying the expression appearing in Eq. (16.7). Maximizing this expected mutual information criterion identifies sets of chimera sequences that are both highly informative and likely to be functional. As the expected value of the mutual information is submodular, it can be efficiently maximized using a greedy algorithm, as described in Step 3.1.3 of Section 3.1. In performing this maximization, covariance matrices are conditioned (Eq. 16.5) on all functional sequences that were observed in the first step of the active learning algorithm so as to encourage the exploration of new regions of sequence space.

In the case of the arginase SCHEMA chimera family, a single execution of the greedy algorithm was performed to identify one of the possible sets of four additional chimeras that maximized the expected value of the mutual information. Genes corresponding to these four arginase chimeras were then synthesized and expressed in *E. coli*. All four of these new arginase chimeras were found to be highly active, validating the functional status logistic regression model's utility for predicting the sequences of catalytically active arginase chimeras.

3.2. Measurement of sample set chimera stability data

A number of techniques are available for obtaining chimera sample set stability data for regression model construction. Stability can be given in terms of free energies of (un)folding, as determined by standard methods, but this is not useful for the many proteins that do not unfold reversibly. A convenient thermostability data collection technique for enzymes is a T_{50} (temperature at which 50% activity is lost after a specified incubation time) measurement, which is useful for the many enzymes that undergo irreversible thermal denaturation.

3.2.1 In measuring chimera T_{50}s, one begins by specifying a thermal denaturation incubation time interval, which should be at least 10 min to allow for temperature equilibration. The sample set chimeras are incubated for the specified time interval across a range of temperatures, in aqueous buffered solution that does not contain substrate. Each thermal denaturation interval is halted by cooling the chimera samples in an ice water bath.

3.2.2 After completing the thermal denaturations, an appropriate substrate is added to all of the heat-treated samples as well as to an untreated reference sample, and the chimeras are assayed for activity. An appropriate enzymatic activity assay is performed to determine each heat-treated sample's residual activity, which is defined as the measured catalytic activity for a heat-treated sample divided by measured catalytic activity of the unheated reference sample.

3.2.3 The T_{50} value, defined as the temperature at which 50% of enzyme activity is lost after the specified incubation time, can then be determined by using nonlinear regression (we typically use the Microsoft Excel solver feature) to fit the residual activity-temperature data pairs. In order to ensure accurate T_{50} measurements, the thermal denaturation temperature range should be such that all of the chimeras

evaluated have residual activities $\leq$25% at the maximum temperature tested. Furthermore, we have observed that using a circulating water bath to make T_{50} measurements results in greater reproducibility (to within 1 °C) than is typically obtained by using a thermal cycler for the incubation step. Finally, one must be certain that enzyme denaturation is irreversible, where some disulfide-containing enzymes might require the addition of a reducing agent to prevent refolding after heat treatment (Heinzelman et al., 2010).

Although T_{50} measurement is convenient for obtaining sizable thermostability datasets for regression model construction in a short period of time, there may be cases in which it does not adequately break out differences in chimera stability or does not reflect the desired property well. Chimera stability can also be measured in terms of retention of catalytic activity over long time intervals. Such an approach was used to measure arginase chimera stability, where stability was quantified by area-under-curve (AUC) determination for activity as a function of time at moderate temperature (Romero et al., 2012). Thermal denaturation half-life ($t_{1/2}$) measurement (Heinzelman et al., 2009) and T_{opt} determination are additional ways (Smith et al., 2012) to describe chimera stability and obtain data for regression model construction.

3.3. Simple linear regression modeling of thermostability data

As noted above, the observation that the blocks which comprise a given enzyme chimera make additive contributions (Eq. 16.1) to that chimera's stability, e.g., T_{50} and/or T_{opt} value, enables the use of linear regression models to predict the sequences of the "best" chimera family members, at least with respect to that criterion. In this section, we describe the approaches that have been used to construct linear regression models for predicting chimera properties. This description features an emphasis on modeling chimera datasets using Bayesian linear regression, an approach that has proven particularly useful for modeling datasets much smaller than those obtained for the P450, CBH II, and family 48 glycosyl hydrolase chimera families, all of which contained at least 51 data points.

For the P450 and CBH II chimera families, simple least squares regression was used to obtain a linear model for predicting chimera T_{50} values (Heinzelman et al., 2009a; Li et al., 2007). The coefficient of determination (R^2 value) between observed and predicted T_{50} values for these respective models was confirmed using 10-fold cross-validation (Dietterich, 1998).

This approach was later improved upon in the context of modeling the thermostabilities of the members of the family 48 glycosyl hydrolase chimera family. In particular, the predictive accuracy of the model was improved by training the model on T_{50} data and E, the number of broken contacts in a given chimera (Smith et al., 2012). Including E as a parameter improved the correlation between observed and predicted T_{50}, increasing the model's R^2 value from 0.82 to 0.88.

3.4. Bayesian linear regression modeling of chimera sample set thermostability data

As noted above, the T_{50} and/or T_{opt} datasets for the P450, CBH II, and family 48 glycosyl hydrolase chimera families all contained 51 or more data points that could be used for linear regression analysis. In contrast, just seven data points (three expressed chimeras from the first sample set and all four from the second) were available for modeling arginase chimera stability. The relatively small size of this dataset motivated the construction of a Bayesian linear regression model, which was expected to improve the ability to accurately fit the limited number of arginase stability data points relative to the previously used linear least squares models (Bishop, 2006).

Ordinary least squares regression finds the set of model parameter values that minimizes the squared difference between experimentally measured and predicted thermostability values, for example, T_{50}s. This parameter set is known as the maximum likelihood estimate, since it is the parameter set that is most probable given the observed data. The maximum likelihood estimate approach is convenient to use as the model parameters can be found by using simple, well-known formulas. While this modeling approach is effective when relatively large datasets, that is, 30 or more data points, are available, it tends to "overfit" the data, or may even be unsolvable, when only small numbers of data points are available for model construction (Bernardo & Smith, 1994).

For cases in which the dataset to be modeled is of limited size, Bayesian linear regression can be used to include additional information. This additional information is specified by a prior probability distribution over the regression parameters.

3.4.1 In the case of modeling chimera stability, before any experimental measurements are made, the effect of block substitutions is assumed to be small and all blocks are taken as being equally likely to be stabilizing. This prior information is encoded with a zero mean,

isotropic Gaussian prior. With this prior, the Bayesian parameter estimates β can be found in closed form and are given by Eq. (16.8)

$$\beta = \left(X^{\mathrm{T}} X + \frac{\sigma_{\mathrm{b}}^2}{\sigma_{\mathrm{n}}^2} I \right)^{-1} X^{\mathrm{T}} y \qquad [16.8]$$

where X is the matrix that codes a chimera's block identities ($X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^{\mathrm{T}}$), I is the identity matrix, y is the vector of corresponding stability measurements, σ_{b}^2 is the prior block variance, and σ_{n}^2 is the variance of the measurement noise. We have found MATLAB to be useful for performing linear algebra calculations.

3.4.2 The values of the two variance hyperparameters, σ_{b}^2 and σ_{n}^2, that minimize the squared error are estimated by performing cross-validation on the chimera sequence/stability measurement dataset. To do this, a range of hyperparameters is scanned and for each combination the model is evaluated for its cross-validated mean squared error (MSE) of prediction. The hyperparameter combination that has the lowest MSE is then used for prediction.

3.4.3 If there are insufficient data to perform cross-validation, the Eq. (16.8) hyperparameters can be estimated directly from the data using the empirical Bayes method (Bernardo & Smith, 1994). In this case, the likelihood function is integrated over all possible regression coefficients (β in Eq. 16.8) to yield a marginalized likelihood function. This marginal likelihood is then maximized with respect to σ_{b}^2 and σ_{n}^2 using gradient descent, Newton's method, or other iterative approaches (Bishop, 2006). Finally, hyperparameters found by cross-validation or empirical Bayes are substituted into Eq. (16.8) to provide Bayesian parameter estimates that are used to construct the predictive chimera stability model.

In the case of modeling the arginase chimera sample set stability data, the Bayesian linear regression approach yielded a model that provided an excellent fit between observed and predicted arginase chimera log AUC values, returning a R^2 value of 0.96 despite being trained on only nine data points. This strong correlation between measured and predicted stability suggests that Bayesian linear regression will be valuable in reducing the numbers of synthetic chimera sample set genes and experimental measurements needed to construct accurate models for predicting the stabilities and possibly other properties of members of SCHEMA protein chimera families.

4. SUMMARY

The application of linear regression analysis to stability data obtained by characterizing small, designed sample sets of SCHEMA chimeras enables the efficient construction of predictive models that accurately identify the sequences of chimera family members whose stabilities are greater than those of the parent enzymes. This "sample, model, and predict" approach allows the sequences of hundreds of enzymes with improved properties and high sequence diversity to be identified and offers an avenue for improving enzymes that are not amenable to engineering using high-throughput screening methods.

A recently developed, two-step active learning algorithm that accounts for both the probability of a chimera being catalytically active and the information content of the chimera sample set markedly reduces the number of sample set genes needed for constructing robust predictive chimera property models. The ability of this approach to decrease the size of the chimera sample set needed to build accurate models will make it possible to model the properties of the members of very large SCHEMA chimera families without undue requirements for gene synthesis and data sampling. Furthermore, the Bayesian linear regression method used in the context of this learning algorithm has enabled the development of a robust predictive model based on data collected for an extremely small chimera sample set that would be extremely difficult to accurately characterize using linear regression. This improved ability to construct accurate regression models from very small datasets could be extrapolated to efficiently identifying the sequences of chimeras with improvements in properties other than (thermo)stability, such as specific catalytic activity or product inhibition, thus further increasing the utility of SCHEMA recombination in allowing the generation of large numbers of improved enzymes that feature high levels of sequence diversity.

ACKNOWLEDGMENTS

The authors acknowledge funding from the Institute of General Medical Sciences of the National Institutes of Health (ARRA grant 2R01-GM068664-05A1) for work on cytochrome P450s and the U.S. Army Research Office Institute for Collaborative Biotechnologies (grant W911NF-09-D-0001) for technology development and cellulase engineering. The contents of this chapter are solely the responsibility of the authors and do not necessarily represent the official views of the sponsors.

REFERENCES

Bernardo, J. M., & Smith, A. M. (1994). *Bayesian theory*. (1st ed.). New York: Wiley & Sons.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. (1st ed.). New York: Springer.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1923.

Endelman, J. B., Silberg, J. J., Wang, Z.-G., & Arnold, F. H. (2004). Site-directed protein recombination as a shortest-path problem. *Protein Engineering, Design & Selection*, *17*, 589–594.

Heinzelman, P., Komor, R., Kannan, A., Romero, P. A., Yu, X., Mohler, S., et al. (2010). Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Engineering, Design & Selection*, *23*, 871–880.

Heinzelman, P., Snow, C. D., Wu, I., Nguyen, C., Villalobos, A., Govindarajan, S., et al. (2009). A family of thermostable fungal cellulases created by structure-guided recombination. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 5610–5615.

Heinzelman, P., Snow, C. D., Smith, M. A., Yu, X., Kannan, A., Boulware, K., et al. (2009a). SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *Journal of Biological Chemistry*, *284*, 26229–26233.

Krause, A. (2010). SFO: A toolbox for submodular function optimization. *Journal of Machine Learning Research*, *11*, 1141–1144.

Krause, A., & Guestrin, C. (2007). Near-optimal observation selection using submodular functions. *Proceedings of 22nd conference on artificial intelligence (AAAI)*. Nectar Track 22, (pp. 1650–1654).

Li, Y., Drummond, D. A., Sawayama, A. M., Snow, C. D., Bloom, J. D., & Arnold, F. H. (2007). A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nature Biotechnology*, *25*, 1051–1056.

Landwehr, M., Carbone, M., Otey, C. R., Li, Y., & Arnold, F. H. (2007). Diversification of catalytic function in a synthetic family of chimeric cytochrome p450s. *Chemisty and Biology*, *14*, 269–278.

Meyer, M. M., Hochrein, L., & Arnold, F. H. (2006). Structure-guided SCHEMA recombination of distantly related beta-lactamases. *Protein Engineering, Design & Selection*, *19*, 563–570.

Minoux, M. (1978). Accelerated greedy algorithms for maximizing submodular set functions. *Proceedings of the 8th IFIP conference on optimization techniques* (pp. 234–243), New York: Springer.

Nemhauser, G. L., Wolsey, L. A., & Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, *14*, 265–294.

Otey, C. R., Landwehr, M., Endelman, J. B., Hiraga, K., Bloom, J. D., & Arnold, F. H. (2006). Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biology*, *4*, e112.

Romero, P., Stone, E., Lamb, C., Chantranupong, L., Krause, A., Miklos, A., et al. (2012). SCHEMA designed variants of human arginase I & II reveal sequence elements important to stability and catalysis. *ACS Synthetic Biology*, *1*, 221–228.

Sawayama, A. M., Chen, M. M., Kulanthaivel, P., Kuo, M. S., Hemmerle, H., & Arnold, F. H. (2009). A panel of cytochrome P450 BM3 variants to produce drug metabolites and diversify lead compounds. *Chemistry: A European Journal*, *15*, 11723–11729.

Shusta, E. V., Kieke, M. C., Parke, E., Kranz, D. M., & Wittrup, K. D. (1999). Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency. *Journal of Molecular Biology*, *292*, 949–956.

Smith, M. A., Rentmeister, A., Snow, C. D., Wu, T., Farrow, M. F., Mingardon, F., et al. (2012). A diverse set of family 48 bacterial glycoside hydrolase cellulases created by structure-guided recombination. *FEBS Journal*, *279*, 4453–4465.