ACS Synthetic Biology

Letter

pubs.acs.org/synthbio

Mapping of Amino Acid Substitutions Conferring Herbicide Resistance in Wheat Glutathione Transferase

Sridhar Govindarajan,[†] Bengt Mannervik,[‡] Joshua A. Silverman,[§] Kathy Wright,[†] Drew Regitsky,[§] Usama Hegazy,[∥] Thomas J. Purcell,[†] Mark Welch,[†] Jeremy Minshull,[†] and Claes Gustafsson*,[†]

[†]DNA2.0, Inc., 1140 O'Brien Drive, Suite A, Menlo Park, California 94025, United States

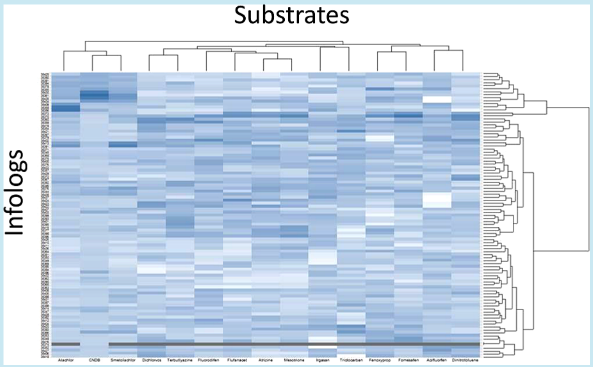
[‡]Department of Neurochemistry, Stockholm University, SE-10691 Stockholm, Sweden

[§]Calysta, 1140 O'Brien Drive, Suite B, Menlo Park, California 94025, United States

[∥]National Research Centre, Dokki, Cairo 12311, Egypt

S *Supporting Information*

ABSTRACT: We have used design of experiments (DOE) and systematic variance to efficiently explore glutathione transferase substrate specificities caused by amino acid substitutions. Amino acid substitutions selected using phylogenetic analysis were synthetically combined using a DOE design to create an information-rich set of gene variants, termed infologs. We used machine learning to identify and quantify protein sequence−function relationships against 14 different substrates. The resulting models were quantitative and predictive, serving as a guide for engineering of glutathione transferase activity toward a diverse set of herbicides. Predictive quantitative models like those presented here have broad applicability for bioengineering.



KEYWORDS: *synthetic biology, machine learning, design of experiment, enzymes, sequence space, optimization, bioengineering, herbicide resistance*

Glutathione transferase (GST, EC 2.5.1.18) catalyzes the conjugation of glutathione (GSH) to the electrophilic center of a variety of endogenous and xenobiotic substrates for the purpose of subsequent detoxification. Mammals and plants have numerous diverse GSTs allowing for the conjugation of a broad range of diverse substrates. Members of the GST superfamily are broadly distributed in amino acid sequence homology, and the preferred substrate(s) for a large fraction of the GST sequences deposited in public databases are unknown.[1] The GST enzymes are often expressed at high concentrations, constituting up to 10% cytosolic protein in mammalian liver[2] and up to 2% of total foliage protein in cereal crops.[3] In plants, the family of soluble GSTs comprises at least six classes of proteins based on phylogenetic relationships: theta, zeta, phi and tau, lambda, and dehydroascorbate reductase,[4] of which the phi and tau classes, unique to plants, are the most numerous. Previous plant genomics approaches and laboratory derived directed evolution experiments have identified and characterized a number of native and synthetically evolved GST proteins displaying widely varying expression profiles and substrate specificities.[5,6] Detoxification of several herbicides, including dimethenamid,[7] fenoxaprop-ethyl,[8] and flupyrsulfuronmethyl,[9] has been shown to be catalyzed by native wheat GSTs.

Potential non-GMO transgenic trait engineering relies on altering few nucleotides in the genome without the aid of markers or nonendogenous genetic sequences.[10] Building a sequence−function map of natural wheat GST will allow for the identification of the minimal number of GST substitutions that result in the most change of the trait to be affected. Such maps could be useful for many different herbicide resistance traits.

In this paper, we have mapped the relationship between GST sequence and substrate specificity to illustrate this approach as a general model for navigating sequence−function space. Protein engineering is a search for variables (typically amino acid substitutions) and their preferred combinations that produce desired functional properties. For most enzyme engineering applications, these properties can be orthogonal or correlated. For example, one may wish to improve catalytic activity, while maintaining high stability and substrate specificity, and minimizing any deleterious interactions with the production host or application system. Structure−function space thus has a high dimensionality in both the dependent and independent variables. Efficient navigation of this space requires optimized search methodologies.

With the 20 naturally occurring amino acids as possibilities for each residue in a protein, sequence space is hyperdimensional and vast (20^N where N is the number of residues in

Received: April 23, 2014
Published: June 6, 2014

```
                               N
              I       Y A      S
            VRV  V A ALLS  R   H            VF R      Y    RRV
MHHHHHHAGGDDLKLLGAWPSPFVTRVKLALALKGLSYEDVEEDLYKKSELLLKSNPVHKKIPV

                                                      LK
 L       D LV        W      I                F     L  MQ   K
LIHNGAPVCESMIILQYIDEVFASTGPSLLPADPYERAIARFWVAYVDDKLVAPWRQWLRGKTE

     E                      DP
     A     V        F       KC               L  FL  A
EEKSEGKKQAFAAVGVLEGALRECSKGGGFFGGDGVGLVDVALGGVLSWMKVTEALSGDKIFDA

                               Y
 R  R        V A  V   I     V  LV  K
AKTPLLAAWVERFIELDAAKAALPDVGRLLEFAKAREAAAAASK*
```

Figure 1. TaGSTU4−4 sequence. The protein sequence of TaGSTU4−4 (wtGST) depicted above. The amino acid substitutions explored within the infolog set are shown above the corresponding wild type amino acid.

the protein). Previous work has shown that the majority of this protein sequence space is nonfunctional for any given property, in most cases failing even to fold into a defined structure.[11] Naturally occurring proteins with fitness for one or more functional properties occupy extremely small regions of this space, much like very rare islands within the total available sequence space. For desirable activities, functional islands are far too rare to be found by a random search. Despite recent advances current rational understanding of protein folding and structure−function relationship is inadequate to permit reliable *de novo* design of catalytic function in proteins.[12] In contrast, directed evolution methods starting from existing natural protein islands in sequence space have been proven reasonably successful.[13] Even the best directed evolution methods have significant room for improvement: random sampling of very large functionally sparse sequence space necessitates high-throughput functional screening, which in turn limits the types of activity measurements that can be made.[14] Furthermore, directed evolution methods generally keep only the sequence information in the handful of "best" clones that appear in the course of the selection process, discarding the information from the majority of the variants tested.

A few protein sequence−function mapping studies have attempted to thoroughly explore a narrowly defined region of total sequence space. Noel and co-workers analyzed the impacts of all 9 amino acid substitutions separating two sesquiterpene synthases in the catalytic landscape. They created a library of 2^9 (= 512) gene variants where >80% of all possible combinations were tested for product formation, and the relative contribution of each substitution and combination of substitutions were quantified resulting in a measure of the mutational accessibility between two terpene synthases.[15] Similarly, Keasling and co-workers performed site saturation mutagenesis at 19 residues close to the active site of a sesquiterpene synthase and combined the preferred amino acid substitutions allowing for only 2500 gene variants to be screened and identifying 7 novel substrate specificities.[16] The amino acid substitution interdependence can be captured as linearly additive as in the above examples or by using Bayesian statistics as was used for modeling chimeric cytochrome P450s.[17]

We have broadened this approach by using design of experiments (DOE) strategies to minimize the number of individual sequences that must be tested to map a sequence−function landscape. Briefly, we used phylogenetic, structural and experimental information to computationally identify amino acid substitutions likely to affect GST-mediated herbicide resistance. The top scoring 59 amino acid substitutions were systematically introduced into the natural wheat GST protein encoding gene to generate a small set of systematically varied gene variant "infologs", so that covariation between amino acid substitutions of the encoded proteins was minimized and each dimension was sampled uniformly. Infologs were tested for their activity on 14 different substrates, and machine learning was used to estimate the relative weights of the independent and combinatorial functional contributions of the amino acid substitutions.

We have previously argued that the scheme of natural evolution is based on a molecular quasi-species consisting of an ensemble of functionally related variants from which enhanced functions can emerge by stochastic mutagenesis.[18,19] The design strategy based on infologs described here is a rational primary-structure guided approach by which a corresponding ensemble of mutants is specifically synthesized for optimal and quantitative information content and directed evolution.

We show that a set of 95 GST infologs is sufficient for the identification of amino acid substitutions that increase enzymatic activity against 6 herbicides. We also find activities against 8 herbicides toward which the natural wheat GST has no detectable activity. This information set allowed us to construct a crude map of the GST sequence−function space in 14 functional dimensions. We also show that the weights of the amino acid substitutions from sequence−activity models are predictive, enabling navigation of enzyme specificity through megadimensional herbicide activity space.

As a model GST protein for this study, the tau-class GST enzyme TaGSTU4−4 from wheat (*Triticum aestivum L.*) was chosen. The sequence of wheat TaGSTU4−4 is presented in Figure 1. TaGSTU4−4 is a homodimer present in wheat that catalyzes the conjugation of reduced GSH to several different substrates *in vitro*.[7,20,21] The TaGSTU4−4 protein has previously been purified from etiolated shoots of diploid wheat, showing that the native protein can detoxify dimethenamid, a common herbicide, and CDNB (1-chloro-2,4-dininitrobenzene), the standard chromogenic GST substrate.[7] The corresponding TaGSTU4−4 gene has been cloned and the recombinant protein crystallized at 2.2 Å resolution in complex with *S*-hexylglutathione, a substrate analogue inhibitor.[20] The recombinant TaGSTU4−4 protein and engineered versions thereof was shown to catalyze the glutathionylation of eight related compounds.

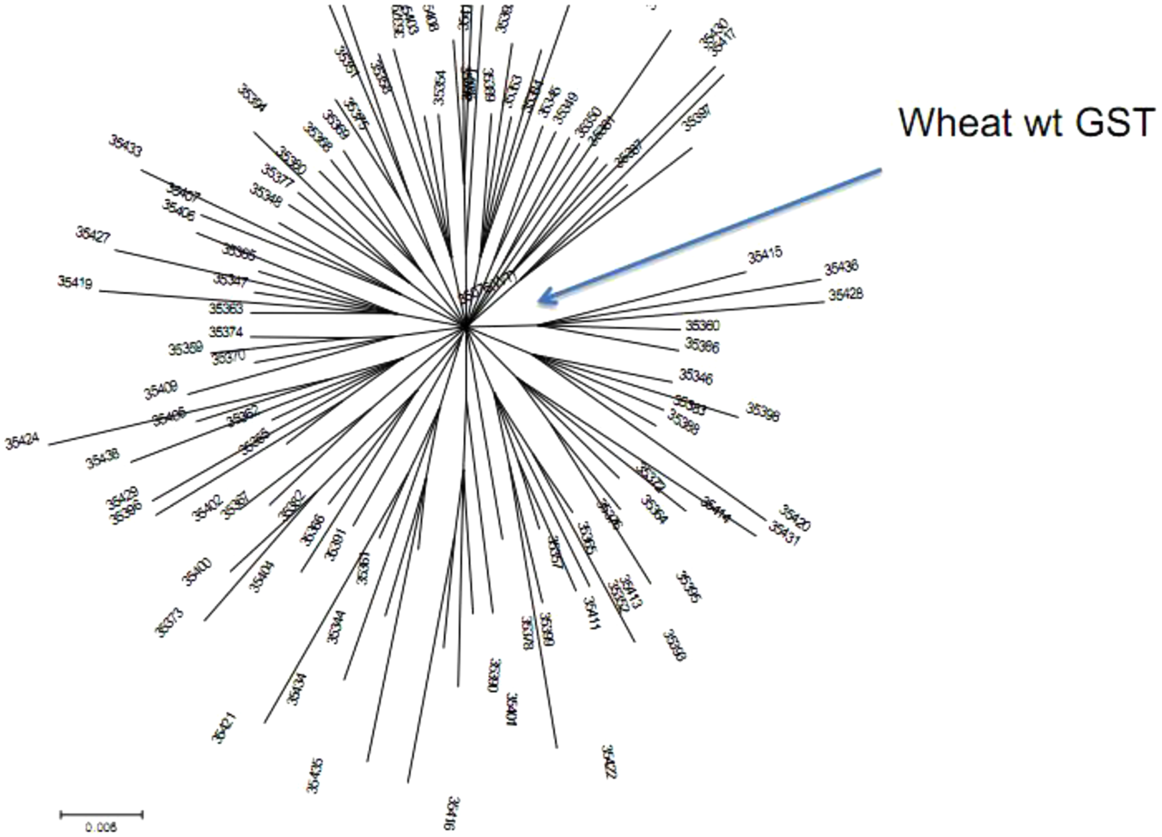


Figure 2. Global sequence distribution of GST infologs. Unrooted tree generated by neighbor-joining method, based on the pairwise Hamming distance (number of nucleotide difference) between each infolog. The tree encodes 95 GST infologs surrounding the wtGST center point. Each infolog is equidistant from every other infolog having the same number substitutions.

We added an N-terminal His-tag to the amino acid sequence of TaGSTU4−4, and the resulting protein sequence was converted to a nucleotide sequence using the GeneGPS technology[22] designed to recode genes to express high level of soluble protein in *Escherichia coli*. The synthetic construct was denoted wtGST, synthesized and cloned in expression vector pJ401 (DNA2.0) under the control of a modified T5 promoter. The wtGST was expressed in *E. coli* strain BL21, purified by Ni-NTA chromatography, and shown to glutathionylate CDNB with specific activity of ∼600 nkat mg^{-1} protein, very similar to that previously reported,[20] illustrating that the recoding and addition of N-terminal His-tag did not significantly alter the functional properties of the protein.

Variations among homologues of extant proteins have arisen by natural evolution from neutral or adaptive changes within functional islands of the sequence space. Unlike random mutations, naturally existing amino acid variations generally retain the overall protein fold while allowing for the exploration of improved functionality.[23] Individual amino acid variations observed in nature are a preferred source of amino acid substitutions as they are less likely than random amino acid substitutions to introduce detrimental changes in the protein. In addition, due to the nature of molecular evolution, there is a low likelihood for simultaneous mutations to be fixed in the population. The effects of natural amino acid substitutions are also predominantly independent,[24] making naturally existing amino acid substitutions better suited to statistical modeling of sequence−function than substitutions generated by random or saturation mutagenesis, rational design using structure information, or by any other random means.

We accordingly used amino acid substitutions present in existing GST homologues to explore the sequence−function correlation of GST substrate specificity. The wtGST protein sequence was used as a BLAST query[25] against Genbank release 174 to extract 180 homologue sequences with >40% identity to wtGST. The 180 sequences were aligned using ClustalW[26] and all amino acid diversity present at every position in the GST protein alignment was captured. A total of ∼1700 amino acid substitutions were identified and rank ordered using automated variable selection and characterization tools as described previously.[27,28] In short, the ∼1700 substitutions distributed across the length of the gene were quantified for 10 properties such as scoring in a global Dayhoff PAM matrix as well as a Dayhoff PAM matrix built for this family of proteins, sequence conservation, convergent evolution rate, rate of synonymous codon mutations, surface *versus* core location in protein structure, proximity to active site and information content. Each of the ∼1700 substitutions was evaluated for each property with scoring ranging from 0 (*e.g.*, most distant from active site) to 1 (closest to active site). The scores were mean centered and normalized for each property. The corresponding value was averaged across all properties for each substitution and used to rank-order the substitutions. The 59 highest ranked substitutions are mapped on the wtGST gene sequence in Figure 1 and listed in Table S1 (Supporting Information). We predominantly identify amino acid substitutions that are similar (L to I/V, and K to R, *etc.*) suggesting minor perturbation of the enzyme instead of drastic alterations. These 59 amino acid substitutions define a 59-dimensional qualitative sequence space centered on the wtGST sequence,

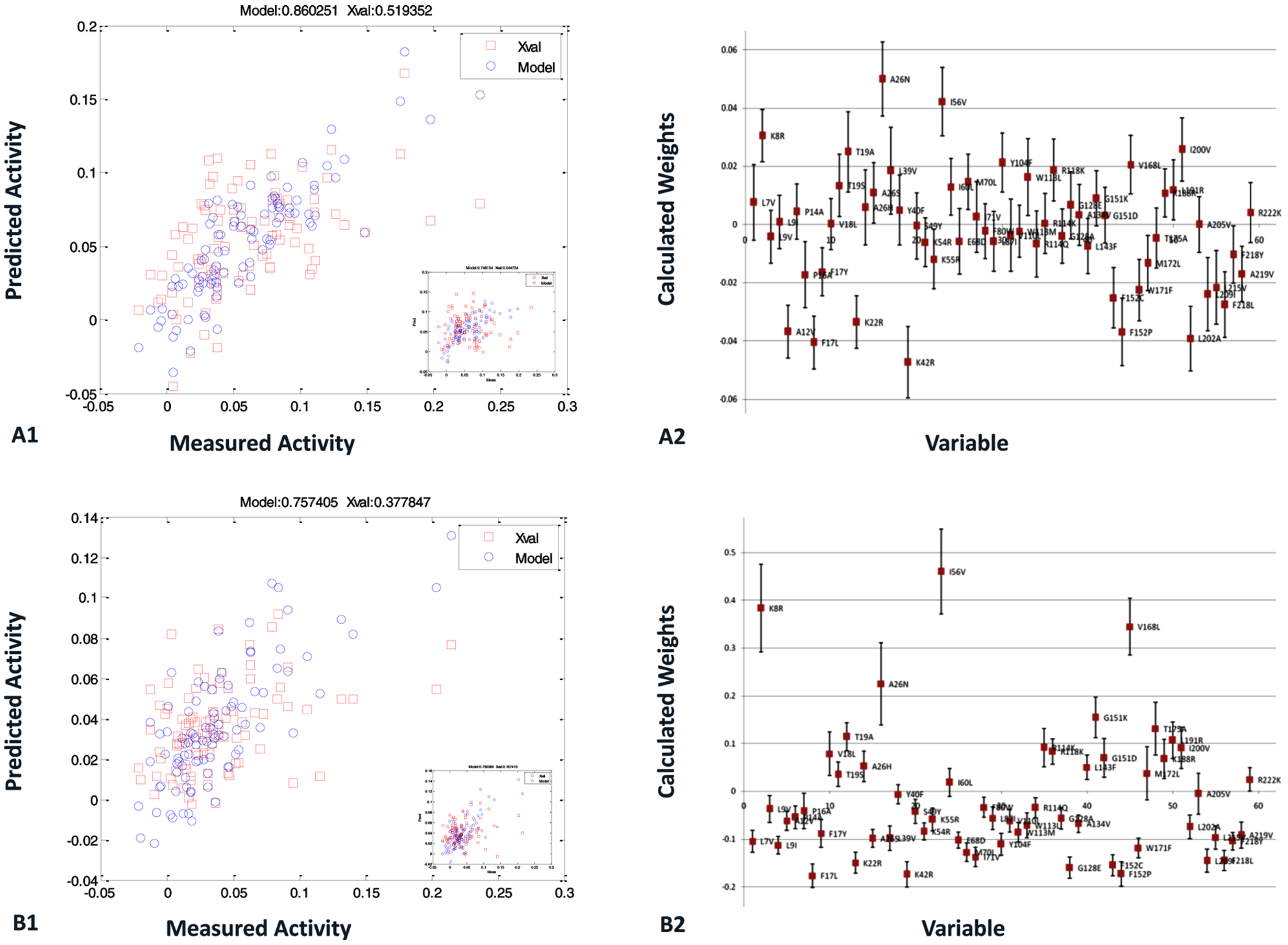


Figure 3. Predictive models for substrates *S*-metolachlor and alachlor. The measured activity of each infolog is used to generate a model (A1, B1) assigning a weight (A2, B2) to each individual variable (substitution) describing its effect on activity. A1 and A2 is the model and weight assignment respectively for *S*-metolachlor. B1 and B2 is the model and weight assignment respectively for alachlor. The distribution bar for each variable indicates the Gaussian distribution of the calculated weight within the 1000 bootstraps of subsampling as described in the text. Instick graph in A1 and B1 shows model if sample order is randomized (cross validation of 0.04 and 0.17 respectively).

where each varying position is a variable with a value of 0 (native residue) or 1 (substitution).

Following the identification of 59 amino acid substitutions, we designed synthetic GST sequences using DOE methods[29] to systematically sample the 59 amino acid substitutions within the backbone of the functional parent enzyme and assess their contribution to GST substrate specificity. These synthetic gene variants are referred to as infologs and defined as genes and/or proteins related by synthetic ancestry designed to approach perfect diversity distribution. This type of unbiased systematic sampling allows for maximal exploration of sequence space with a minimal number of test samples.

A total of 95 infologs were designed to systematically incorporate 3 (48 infologs), 4 (24 infologs), or 5 (23 infologs) identified substitutions from Table S1 (Supporting Information), with each substitution represented in 6 of the 95 infologs. Each amino acid substitution was encoded by a codon identified by the GeneGPS algorithm to ensure that consistent and high heterologous protein expression level was retained. Each infolog was designed to be as distant and as uncorrelated as possible from all other infologs in the 59-dimensional amino acid substitution space. This design provides an unbiased and uniform distribution of the sampled substitutions as seen in Figure 2. By measuring the functional effect of each substitution 6 times in 6 different contexts, the contributions from individual substitutions can be assessed, along with their dependency on the context, without a significant increase in the total number of test samples required.

Chemical gene synthesis enabled quick and efficient synthesis of the full infolog set. Proteins were expressed and purified in parallel, and GST activity assays were adapted for 96-well plates. Most GST infologs expressed >50 μg of protein per mL as quantified by PAGE. The 95 purified infologs and the positive control wtGST were diluted to 25 μg/mL and dispensed in 96-well plates.

Fourteen xenobiotic compounds were selected based on commercial relevance and the presence of an electrophilic center reasonably likely to be amenable to GST-mediated detoxification (Figure S1, Supporting Information). Commonly used herbicides *S*-metholachlor, alachlor, fenoxaprop and flufenacet all target fatty acid biosynthesis in grass. Herbicides fomesafen, acifluorfen and fluorodifen are inhibitors of the chlorophyll precursor synthase protoporphyrinogen oxidase. Atrazine and terbuthylazine are herbicides inhibiting photosystem II and mesotrione inhibits pigment formation. Dichlorvos is an insecticide inhibiting acetylcholinesterase and the active ingredients triclosan and triclocarban are commonly used antibacterial/antifungal compounds used in soap and other consumer products. The final xenobiotic molecule in the set is dinitrotoluene, primarily known as a precursor to the explosive material trinitrotoluene (TNT), but mainly produced as a precursor to commodity chemical toluene diisocyanate. All of the listed xenobiotic compounds are toxic and environmental hazards to various degrees.

The catalytic activity of each infolog against every substrate was quantified using a standard GST assay measuring the consumption of GSH (thiol donor) in the sample relative to a negative control.[30] The canonical GST substrate CDNB was used as positive control for GST activity.

The starting point wtGST had detectable activity against 6 of the 14 xenobiotics, with catalytic activities ranging from level of detection (5 mmol product/mol enzyme min^{-1}) to 150 mmol product/mol enzyme min^{-1} measured against preferred substrate *S*-metolachlor as detailed in Table S2 (Supporting Information). Assessment of protein-dependent GSH conjugation from the full infolog set against *S*-metolachlor identified 4 out of 95 infologs with improved activity relative to wtGST. Further, infologs with measurable catalytic GSH conjugation activity were identified for all 15 tested substrates (including against the eight substrates where wtGST showed no detectable activity). Improvement of catalytic activity over the starting point wtGST was identified for all 15 substrates tested. The success rate among the 95 infologs is striking. Wide functional diversity was also observed in specific activity and substrate specificity across the set as can be seen in the sequence function heat map in Figure S2 (Supporting Information). In addition, every one of the 95 infologs sampled had functional activity in at least one substrate dimension. Thus, within the small and defined sequence space of wtGST and its 95 infologs representing 59 systematically varied amino acid substitutions, a vast functional space could be accessed.

Sequence−function correlation models were built essentially as described.[27] Briefly, machine learning algorithms were used to build linear models of the data set by calculating a 59-dimensional weight vector w (one dimension per variable), where the activity Y_j of a variant X_j is estimated as $Y_j = (\sum_{j=1..59} w_j x_{i,j})$. The weight w_j is associated with the *j*-th substitution, providing a quantitative estimate of the relative effect of the *j*-th substitution on GST activity. $x_{i,j}$ represents the presence or absence of substitution *j* in variant *i* (takes a value from (0,1)). Bootstrapping techniques were used to create 1000 data sets containing a training subset (the set of all *x*,*y* pairs used for a cycle of machine learning) and test subsets by randomly splitting 20% of infolog sequences (x_i) into the test set and the rest into the training set. We interpret the weight distribution as an indirect measure of variable epistasis. Machine learning algorithms were used to select values for w_j that resulted in the highest correlation between measured and predicted activities for test subsets.

Predictive models could be built for 10 of the 15 substrates. For the remaining five substrates for which we identified novel activity, the number of infologs with detectable activity was too low and the sampling too sparse to allow for statistically significant models to be built. The data and the models were validated *in silico* by data permutation testing where Y_i data were randomized relative to X_i and models were constructed for the permuted data. Several randomizations were performed and the Wilcoxon signed rank test was used to assess the significance of the nonpermuted model.[29] Cross validation will only quantify the accuracy of the model within the range of training set (*i.e.*, predict activity from gene variants performing better than the worst gene in the training set and worse than the best gene in the training set). Accuracy of the model predictions may not necessarily extend to the tangent of the function, and the accuracy of the model will with certainty degrade with increasing distance from the training set.

Models for substrates alachlor (R^2 cross validation of 0.38) and *S*-metolachlor (R^2 cross validation of 0.52) with high activities and models with robust statistical significance are presented in Figure 3. The amino acid substitution weights denoted in the right panel of Figure 3 illustrate well the functional distribution of the substitutions for two relatively similar compounds. The substitution weights were subsequently used to calculate the predicted activity of each infolog against each substrate. Comparing each amino acid substitution weight for the two related substrates alachlor and *S*-metolachlor as is done in Figure S3 (Supporting Information) revealed substitutions with minimal effect in either functional direction (*e.g.*, Y40F), substitutions that improve both activities (*e.g.*, I56 V), substitutions that decrease both activities (*e.g.*, F17L) and substitutions that are positive in one functional dimension and negative in the other (*e.g.*, P16A). Mapping of the identified substitutions on the TaGSTU4−4 structure does not suggest any obvious rationale for choice of amino acid or location (Figure S5, Supporting Information)

The GST enzyme family is abundant and within it are a broad range of substrate specificities.[5,31] We have used amino acid substitutions found within the GST family to design GST infologs to systematically explore determinants of sequence−function correlation in a GST sequence. The 95 infologs were designed to capture and quantify the relative functional contribution of 59 amino acid substitutions. The number of substitutions per infolog was controlled to maximize the sampling of single substitutions and combinations while limiting the number of variants with unmeasurable activity resulting from deleterious substitutions and combinations. Similarly, choosing phylogenetically validated amino acid substitutions also biased the infolog set for functionality. The infolog design enabled a search space of 2.2×10^{16} proteins (38 positions that are available in 2 alternate amino acids, 9 positions available in 3 alternate amino acids and one position that is available in 4 alternate amino acids = $2^{38} \times 3^9 \times 4^1$). The search space size is still very small compared to the total available TaGSTU4−4 space, although we believe it provides us access to the part of the GST space that is robust, relatively additive, and can be measured with low throughput high quality assay technology.

Figure 3 illustrates the ability of machine learning models to predict the relationship of protein sequence to activity toward substrates *S*-metolachlor and alachlor. Activity for both substrates is well predicted. Similar models can be built for 10 of the 14 substrates. The relative variable weight contribution (Figure 3, right panels) identifies the variables that contribute to either *S*-metolachlor activity and/or alachlor activity. The relative weights for all 59 amino acid substitutions against *S*-metolachlor and alachlor substrates can be displayed as a biplot using principal component analysis (Figure S4, Supporting Information) where the relative weight of each substitution is denoted for two of the 10 functional dimensions for which statistically relevant models could be built. Thus, the infolog analysis provides a direct tool for simultaneous navigation in the different sequence−function dimensions of

GST. The process can be iterated for further improvement in one or many functional dimensions with increasing resolution.[32,33]

The results demonstrate that sequence space, if searched using conservative amino acid substitution selection and explored systematically, is locally rich in function and functional diversity. The infolog approach here using only 95 sequences centered on starting point wtGST provides functional correlation weights for each of the 59 amino acid substitutions tested in 15 dimensions of relevant functional space. Thus, a infolog library affords tractable information, which is more conducive to rational directed evolution than a stochastically derived quasispecies.[6] The technology can be iterated for protein engineering[27,32,33] in accordance with the current shift toward small, functionally rich libraries.[13] The infolog approach is also well suited to a variety of protein engineering applications, including using protein domains as variables,[34,35] protein expression optimization[22,36] and structure−function analysis.[37]

METHODS

Molecular Biology. The GST infologs with an N-terminal hexaHis tag were synthesized by DNA2.0 and incorporated into vector pJ401 using Electra cloning.[38] Vector pJ401 (DNA2.0) is a high copy vector (pUC origin of replication) characterized by kanamycin resistance and an IPTG inducible T5 promoter. The GST infologs were transformed into *E. coli* strain BL21. Single cell colonies were subsequently grown overnight in LB and antibiotic. An aliquot of the culture was induced by adding 1 mM IPTG and grown for a further 4 h before harvesting cells, lysis and isolation of His-tagged proteins using standard Ni-NTA chromatography. With the exception of GST infologs encoding the K8R substitution, all infologs expressed at similar levels and all were soluble.

Biochemistry. All xenobiotic compounds were purchased from Riedel-de Haen or Fluka. Positive control substrate CDNB was purchased from Sigma-Aldrich. The substrates were dissolved in H_2O or DMSO. The enzymes were dialyzed overnight with 20 mM sodium phosphate, 1 mM EDTA, and 0.2 mM DTT, pH 7. The enzymes were mixed with 2 mM glutathione and herbicide in 0.1 M sodium phosphate and 1 mM EDTA, pH 7.5 at 24 °C for suitable time period, 10 μL of this mixture was mixed with 240 μL of 0.1 M sodium phosphate, 1 mM EDTA and 0.2 mM DTNB (5,5′-dithiobis(2-nitrobenzoic acid), Ellman's reagent), pH 7.5 and absorbance at wavelength 412 nm was measured.[30]

Computational Modeling. Computational modeling was performed using PLS Toolbox (eigenvector Research Inc., Wenatchee, WA) implemented in a MATLAB environment (The MathWorks Inc., Natick, MA),

ASSOCIATED CONTENT

Supporting Information

This material is available free of charge *via* the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: cgustafsson@dna20.com.

Notes

The authors declare the following competing financial interest(s): SG, JM, and CG are inventors of US patents 8005620, 8412461, and 8635029 related in part to material presented here. All authors except BM and UH are employed and hold shares in DNA2.0 and/or Calysta Energy. BM and UH declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the DNA2.0 production team for synthesizing all genes described. The work was supported by a grant to BM from the Swedish Research Council.

REFERENCES

(1) Atkinson, H. J., and Babbitt, P. C. (2009) Glutathione transferases are structural and functional outliers in the thioredoxin fold. *Biochemistry 48*, 11108−11116.

(2) Guthenberg, C., Morgenstern, R., DePierre, J. W., and Mannervik, B. (1980) Induction of glutathione S-transferases A, B and C in rat liver cytosol by trans-stilbene oxide. *Biochim. Biophys. Acta 631*, 1−10.

(3) Edwards, R., Dixon, D. P., and Walbot, V. (2000) Plant glutathione S-transferases: enzymes with multiple functions in sickness and in health. *Trends Plant Sci. 5*, 193−198.

(4) Cummins, I., Dixon, D. P., Freitag-Pohl, S., Skipsey, M., and Edwards, R. (2011) Multiple roles for plant glutathione transferases in xenobiotic detoxification. *Drug Metab. Rev. 43*, 266−280.

(5) McGonigle, B., Keeler, S. J., Lau, S. M., Koeppe, M. K., and O'Keefe, D. P. (2000) A genomics approach to the comprehensive analysis of the glutathione S-transferase gene family in soybean and maize. *Plant Physiol. 124*, 1105−1120.

(6) Kurtovic, S., and Mannervik, B. (2009) Identification of emerging quasi-species in directed enzyme evolution. *Biochemistry 48*, 9330−9339.

(7) Riechers, D. E., Irzyk, G. P., Jones, S. S., and Fuerst, E. P. (1997) Partial characterization of glutathione S-transferases from wheat (*Triticum* spp.) and purification of a safener-induced glutathione S-transferase from *Triticum tauschii*. *Plant Physiol. 114*, 1461−1470.

(8) Cummins, I., Cole, D. J., and Edwards, R. (1997) Purification of multiple glutathione transferases involved in herbicide detoxification from wheat (*Triticum aestivum* L.) treated with the safener fenchlorazole-ethyl. *Pestic. Biochem. Physiol. 59*, 35−49.

(9) Koeppe, M. K., Barefoot, A. C., Cotterman, C. D., Zimmerman, W. T., and Leep, D. C. (1997) Basis of selectivity of the herbicide flupyrsulfuron-methyl in wheat. *Pestic. Biochem. Physiol. 59*, 105−117.

(10) Waltz, E. (2012) Tiptoeing around transgenics. *Nat. Biotechnol. 30*, 215−217.

(11) Keefe, A. D., and Szostak, J. W. (2001) Functional proteins from a random-sequence library. *Nature 410*, 715−718.

(12) Baker, D. (2010) An exciting but challenging road ahead for computational enzyme design. *Protein Sci. 19*, 1817−1819.

(13) Lutz, S. (2010) Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin Biotechnol. 21*, 734−743.

(14) Levay-Young, B., Olesiuk, M., Gustafsson, C., and Minshull, J. (2013) Library format for bioengineering: Maximizing screening efficiency through good design. *Genet. Eng. Biotechnol. 33*, 18−19.

(15) O'Maille, P. E., Malone, A., Dellas, N., Andes Hess, B., Jr., Smentek, L., Sheehan, I., Greenhagen, B. T., Chappell, J., Manning, G., and Noel, J. P. (2008) Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nat. Chem. Biol. 4*, 617−623.

(16) Yoshikuni, Y., Ferrin, T. E., and Keasling, J. D. (2006) Designed divergent evolution of enzyme function. *Nature 440*, 1078−1082.

(17) Romero, P. A., Krause, A., and Arnold, F. H. (2013) Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U. S. A. 110*, E193−201.

(18) Emren, L. O., Kurtovic, S., Runarsdottir, A., Larsson, A. K., and Mannervik, B. (2006) Functionally diverging molecular quasi-species evolve by crossing two enzymes. *Proc. Natl. Acad. Sci. U. S. A. 103*, 10866−10870.

(19) Runarsdottir, A., and Mannervik, B. (2010) A novel quasi-species of glutathione transferase with high activity towards naturally occurring isothiocyanates evolves from promiscuous low-activity variants. *J. Mol. Biol. 401*, 451−464.

(20) Thom, R., Cummins, I., Dixon, D. P., Edwards, R., Cole, D. J., and Lapthorn, A. J. (2002) Structure of a tau class glutathione S-transferase from wheat active in herbicide detoxification. *Biochemistry 41*, 7008−7020.

(21) Dixon, D. P., McEwen, A. G., Lapthorn, A. J., and Edwards, R. (2003) Forced evolution of a herbicide detoxifying glutathione transferase. *J. Biol. Chem. 278*, 23930−23935.

(22) Gustafsson, C., Minshull, J., Govindarajan, S., Ness, J., Villalobos, A., and Welch, M. (2012) Engineering genes for predictable protein expression. *Protein Expression Purif. 83*, 37−46.

(23) Govindarajan, S., and Goldstein, R. A. (1997) Evolution of model proteins on a foldability landscape. *Proteins 29*, 461−466.

(24) Govindarajan, S., Ness, J. E., Kim, S., Mundorff, E. C., Minshull, J., and Gustafsson, C. (2003) Systematic variation of amino Acid substitutions for stringent assessment of pairwise covariation. *J. Mol. Biol. 328*, 1061−1069.

(25) Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol. 215*, 403−410.

(26) Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics 23*, 2947−2948.

(27) Liao, J., Warmuth, M. K., Govindarajan, S., Ness, J. E., Wang, R. P., Gustafsson, C., and Minshull, J. (2007) Engineering proteinase K using machine learning and synthetic genes. *BMC Biotechnol. 7*, 16.

(28) Minshull, J., Ness, J. E., Gustafsson, C., and Govindarajan, S. (2005) Predicting enzyme function from protein sequence. *Curr. Opin. Chem. Biol. 9*, 202−209.

(29) Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2009) *Multivariate Data Analysis*, 7th ed., Prentice-Hall, Upper Saddle River, NJ.

(30) Mannervik, B., and Jemth, P. (2001) Measurement of glutathione transferases. *Current Protocols in Toxicology*, Chapter 6, Unit 6.4, Wiley, New York.

(31) Zhang, W., Dourado, D. F., Fernandes, P. A., Ramos, M. J., and Mannervik, B. (2012) Multidimensional epistasis and fitness landscapes in enzyme evolution. *Biochem. J. 445*, 39−46.

(32) Ehren, J., Govindarajan, S., Morón, B., Minshull, J., and Khosla, C. (2008) Protein engineering of improved prolyl endopeptidases for celiac sprue therapy. *Protein Eng., Des. Sel. 21*, 699−707.

(33) Midelfort, K. S., Kumar, R., Han, S., Karmilowicz, M. J., McConnell, K., Gehlhaar, D. K., Mistry, A., Chang, J. S., Anderson, M., Villalobos, A., Minshull, J., Govindarajan, S., and Wong, J. W. (2013) Redesigning and characterizing the substrate specificity and activity of *Vibrio fluvialis* aminotransferase for the synthesis of imagabalin. *Protein Eng., Des. Sel. 26*, 25−33.

(34) 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.

(35) Heinzelman, P., Snow, C. D., Smith, M. A., Yu, X., Kannan, A., Boulware, K., Villalobos, A., Govindarajan, S., Minshull, J., and Arnold, F. H. (2009) SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *J. Biol. Chem. 284*, 26229−26233.

(36) Welch, M., Govindarajan, S., Ness, J. E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C. (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One 4*, e7002.

(37) Chen, F., Gaucher, E. A., Leal, N. A., Hutter, D., Havemann, S. A., Govindarajan, S., Ortlund, E. A., and Benner, S. A. (2010) Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. *Proc. Natl. Acad. Sci. U. S. A. 107*, 1948−1953.

(38) Whitman, L., Gore, M., Ness, J., Theodorou, E., Gustafsson, C., and Minshull, J. (2013) Rapid, scarless cloning of gene fragments using the electra vector system. *Genet. Eng. Biotechnol. 33*, 42.