Protein Engineering, Design & Selection, 2017, vol. 30 no. 8, pp. 543–549
doi: 10.1093/protein/gzx045
Advance Access Publication Date: 30 August 2017
Original Article

Original Article

Exploring sequence-function space of a poplar glutathione transferase using designed information-rich gene variants

Yaman Musdal[1], Sridhar Govindarajan[2], and Bengt Mannervik[1,*]

[1]Department of Neurochemistry, Arrhenius Laboratories, Stockholm University, Svante Arrhenius väg 16B, SE-10691 Stockholm, Sweden, and [2]ATUM, 37950 Central Ct, Newark, CA 94560, USA

[*]To whom correspondence should be addressed. E-mail: bengt.mannervik@neurochem.su.se
Edited by Alan Fersht

Received 23 July 2017; Revised 0 0; Editorial Decision 25 July 2017; Accepted 15 August 2017

Abstract

Exploring the vicinity around a locus of a protein in sequence space may identify homologs with enhanced properties, which could become valuable in biotechnical and other applications. A rational approach to this pursuit is the use of 'infologs', i.e. synthetic sequences with specific substitutions capturing maximal sequence information derived from the evolutionary history of the protein family. Ninety-five such infolog genes of poplar glutathione transferase were synthesized and expressed in *Escherichia coli*, and the catalytic activities of the proteins determined with alternative substrates. Sequence–activity relationships derived from the infologs were used to design a second set of 47 infologs in which 90% of the members exceeded wild-type properties. Two mutants, C2 (V55I/E95D/D108E/A160V) and G5 (F13L/C70A/G122E), were further functionally characterized. The activities of the infologs with the alternative substrates 1-chloro-2,4-dinitrobenzene and phenethyl isothiocyanate, subject to different chemistries, were positively correlated, indicating that the examined mutations were affecting the overall catalytic competence without major shift in substrate discrimination. By contrast, the enhanced protein expressivity observed in many of the mutants were not similarly correlated with the activities. In conclusion, small libraries of well-defined infologs can be used to systematically explore sequence space to optimize proteins in multidimensional functional space.

Key words: alternative substrates, directed protein evolution, enhanced activities, glutathione transferase, infologs

Introduction

Enzymes have naturally evolved to acquire valuable functions including high catalytic efficiencies in living systems. Crystal structures and computational studies have provided us with knowledge about the reaction mechanisms and have identified critical active-site residues that influence the functional properties. This knowledge has enabled many research groups to use protein engineering methods to tailor active sites of enzymes for altered efficiency, substrate specificity or other desired features (Norrgård *et al.*, 2006; Laurino *et al.*, 2016). However, currently used methods are limited in efficacy due to lack of solved protein structures and incomplete knowledge of the intricacies of enzyme catalysis. Computational predictions of 3D protein structures are less accurate, even though major advances are being made (Huang *et al.*, 2016; Tiwari, 2016). In many instances protein engineering therefore requires strategies based on empirical screening of small or large libraries of designed or random mutants.

Infologs are designed variants of a given gene where substitutions are systematically incorporated to achieve a high information content that enables modern machine-learning tools to deconvolute sequence–activity relationships. The use of infologs is the basis of our rational approach to protein engineering in which a matrix of well-defined amino-acid substitutions are used to map the targeted fitness landscape. By this approach libraries of <100 mutants are characterized for functional properties of interest and serve as a

basis for machine-learning tools to design a new generation of infologs based on empirical probing of the fitness landscape. The initial set of infologs synthesized are designed to have the same (or a similar) number of substitutions (~3), and thereby probe regions at the same (or approximately the same) Hamming distance from the reference locus in sequence space. Substitutions in the mutants are selected from a pool of substitutions and each set of infologs is designed such that several variants contain the same substitution albeit in presence of different additional mutations. Thus, the functional consequences of individual substitutions can be modeled and quantitatively evaluated. In an early application to the engineering of proteinase K based on a training set of 24 variants followed by two sequential libraries, the enzymatic activity was increased 20-fold by a mere 95 mutants in total (Liao *et al.*, 2007). In another investigation an infolog library based on 59 amino-acid substitutions in a tau class glutathione transferase (GST) from wheat afforded increased activity against most of a number of herbicides tested (Govindarajan *et al.*, 2015).

A recent study of tau class GSTU45 from poplar (*Populus trichocarpa*), which was of interest as an enzyme inducible by 2,4,6-trinitrotoluene (TNT), demonstrated that the enzyme had minimal activity with 16 alternative substrates, including TNT, and only modest activity with 1-chloro-2,4-dinitrobenzene (CDNB), the standard GST substrate (Musdal and Mannervik, 2015). Considering its structural similarity to homologous GSTs displaying higher activity it was of interest to find out if the functions of GSTU45 could be enhanced, in support of the evolutionary principle that modest mutagenesis could yield proteins with novel and useful properties. In the present investigation GSTU45 was therefore subjected to protein engineering based on the infolog approach. The enzyme activity was primarily monitored by CDNB and phenethyl isothiocyanate (PEITC), the latter substrate being a natural product serving as an effective substrate for many GSTs (Zhang *et al.*, 2015). In addition, the catalytic function with TNT, structurally similar to CDNB, was tested. Two of the second-generation mutants with enhanced functions were characterized and found to have elevated activities with additional substrates.

Materials and Methods

Expression and purification of poplar GSTU45 variants

Gene sequences encoding wild-type poplar GSTU45 and variants thereof were codon-optimized for high-level expression in *Escherichia coli* BL21 by ATUM (Newark, CA) using the codon optimization algorithm described by Gustafsson *et al.* (2012). The genes were designed to encode N-terminal His_6 to enable efficient purification, and inserted behind the inducible T5 promoter in the pJ401 expression vector (Kan^r, high copy number) by ATUM (Newark, CA). All constructs were expressed in 150 ml LB culture medium at 30°C and purified using pre-packed Ni-Sepharose His GraviTrap affinity columns (GE Healthcare) as described before (Kolm *et al.*, 1995a; Musdal and Mannervik, 2015). Purified enzymes were dialyzed against 10 mM Tris–HCl pH 7.8, 1 mM EDTA, 0.2 mM dithiothreitol and mixed with glycerol to a final concentration of 20% (v/v). Proteins were aliquoted and stored with marginal decrease of enzymatic activity at 5°C as well as at −80°C (data not shown).

Measurement of GST activities

Enzymatic activities were measured with the standard GST substrate CDNB, the naturally occurring substrates PEITC and benzyl isothiocyanate (BITC), as well as with cumene hydroperoxide (CuOOH) using a Shimadzu UV-2501 PC spectrophotometer. Absorbance change was monitored in 1-ml quartz cuvettes using 0.1 M sodium phosphate buffer pH 6.5 (pH 7.0 for CuOOH) containing 1 mM EDTA for 1 min at 30°C. Activities were based on initial rates corrected for the non-enzymatic reactions and using the molar extinction coefficients of substrates at selected wavelengths. Assay conditions with minor modifications (Musdal and Mannervik, 2015) were as previously described: for CDNB (Habig *et al.*, 1974); for PEITC and BITC (Kolm *et al.*, 1995); for TNT (Mazari and Mannervik, 2016); and for CuOOH (Lawrence and Burk, 1976).

Determination of steady-state kinetic parameters

Steady-state kinetic measurements of GSTs with CDNB, PEITC and CuOOH as varied substrates were made under the same conditions as used for specific activity determinations. Measurements were made using at least six different concentrations of substrate and a fixed concentration of GSH. In the experiments CDNB was used in the range of 0.025–2 mM with 5 mM GSH, PEITC was used in the range of 0.0125–1 mM with 1 mM GSH, and CuOOH was used in the range of 0.025–6 mM with 1 mM GSH. In all assays the content of the solvents ethanol (5%, v/v) and acetonitrile (0.5%, v/v) used to dissolve substrates had no effect on enzyme activity. All measurements were performed at least in triplicate.

Kinetic parameters V_{max} and K_m were obtained by nonlinear regression analysis of the initial rates using GraphPad Prism software. The k_{cat} values were calculated from V_{max} values based on the mass of the dimer.

Sequence alignment and structure modeling

The crystal structure 4J2F available in Protein Data Bank (PDB) was used for modeling of the C2 and G5 mutant proteins. The amino acid sequences were aligned with Clustal omega software for comparison with GSTU45 (Sievers and Higgins, 2014). Chimera software was used for modeling the protein structures (Pettersen *et al.*, 2004). Modeled structures of C2 and G5 were matched with the previously published model of GSTU45 (Musdal and Mannervik, 2015) and possible interactions with S-(p-nitrobenzyl)glutathione ligand in the mutant active sites were determined.

Results

GSTU45 infolog set

Recently two members of the extensive GSTome of poplar (*P. trichocarpa*), GSTU45 and GSTU16, were heterologously expressed and characterized. They had previously been found to be inducible by the explosive 2,4,6-trinitrotoluene (TNT) and were of interest for their possible contribution to the biotransformation of this toxic compound (Brentner *et al.*, 2008). However, neither of the enzymes showed noteworthy activity with TNT, and GSTU45 was also comparatively inefficient with a variety of alternative substrates, whereas GSTU16 displayed good activity with BITC and PEITC (Musdal and Mannervik, 2015).

Starting with wild-type GSTU45, we obtained the multiple sequence alignment and a phylogenetic tree of all homologs found in the Genbank non-redundant protein database using the BLAST algorithm. The alignment was used to enumerate all possible changes that can be made to GSTU45 that are seen from the alignments. These changes were then scored based on the pattern of convergence and divergence on the tree and ranked for adaptability score (Liao *et al.*, 2007; Ehren *et al.*, 2008). The top 57 were chosen

to be included in the GSTU45 infolog set, which was constructed to consist of independently designed synthetic DNA sequences derived from the GSTU45 gene (see Supplementary Table 1). The substitutions were systematically incorporated to maximize information content and devised for optimal diversity distribution to maximize search efficiency (Govindarajan *et al.*, 2015). This primary set of GSTU45 infologs comprised the wild-type enzyme (219 amino acids preceded by N-terminal MHHHHHH) and 95 infologs, each infolog containing three specified amino acid substitutions from the top 57, distributed across 45 sites in the sequence. In eight positions two and in two positions three alternative substitutions were introduced. Accordingly, the total number of positions explored in sequence space was $>3 \times 10^{15}$ ($2^{35} \times 3^{8} \times 4^{2}$). The substitutions were made in combinations such that the mutant matrix contained each mutation evenly distributed and in distinct but orthologous sequence contexts in order to maximize the information content. By this fractional factorial design the experimental data obtained could be evaluated quantitatively and the weight of each substitution to every measured experimental data point calculated.

Synthesis, purification and activity measurement of GSTU45 infologs

DNA encoding each member of the designed infologs was synthesized with an N-terminal His_6-tag for heterologous expression in *E. coli*, and the expressed protein was purified via IMAC (Porath *et al.*, 1975) using Ni-Sepharose. SDS-PAGE demonstrated the purity of the proteins and the expected molecular mass of the GST subunits of approximately 25 kDa (Fig. 1).

Each of the 95 purified mutant proteins was tested for activity with three alternative substrates (Fig. 2). CDNB and PEITC were two of the favored substrates of the wild-type GSTU45, and TNT was of interest from the phytoremediation perspective. The majority (>90%) of the GSTU45 variants demonstrated measureable activities with CDNB and PEITC, but no statistically significant activity was found with TNT as substrate. Determination of the protein concentration of the purified enzyme variants allowed calculation of the specific activities. With CDNB 35% and with PEITC 53% of the infologs scored higher specific activity than wild-type GSTU45. The specific activities were individually modeled as a function of the substitutions by linear regression. The results provide relative weights of each substitution for activity with CDNB as well as with PEITC (Fig. 3 and Supplementary Table 2). An obvious positive correlation between the weights for the activities with CDNB and PEITC was noted, such that favorable mutations for one substrate were also favorable for the other substrate.

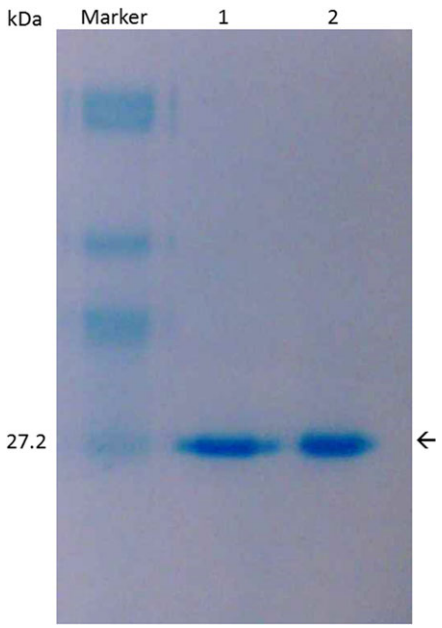

Fig. 1 Representative SDS-PAGE analysis of purified poplar GST45 variants stained with Coomassie Brilliant Blue. M_r markers (lane 1 left), C2 (lane 2) and G5 (lane 3) are shown with an arrow indicating 25 kDa

Second-generation infolog library

All the mutations shown with positive weights for both substrates in Fig. 3 were used in different combinations with other substitutions to construct a new library of GSTU45 mutants. The design of the library was based on the principles and machine-learning algorithms earlier described (Govindarajan *et al.*, 2015). A total of 47 GSTU45 variants displayed 25 selected amino-acid substitutions. The number of mutations in a given variant ranged from one, as in mutant M69I or mutant E122G, to five as in, e.g. mutant K29R/M69I/C110I/N123E/V160A.

The members of the second infolog set were expressed and purified to homogeneity as described earlier in yields ranging from 0.4 to 10 mg/ml.

The activities of the proteins were measured with the GST substrates CDNB and PEITC. With the exception of three proteins, all the infologs showed increased specific activity with both substrates compared to wild-type GSTU45. Among the infologs showing the most prominent specific activities and also being expressed at a high level were mutants C2 and G5, which were obtained in stock solutions of C2 (4.1 mg/ml) and G5 (4.9 mg/ml), with a total of more than 20 mg of each protein. The only differences from wild-type GSTU45 were that C2 has four substitutions (V55I/E95D/D108E/A160V) and G5 has three substitutions (F13L/C70A/G122E) away from wild-type (Fig. 4). These two infologs were chosen for further analysis by means of kinetic measurements and modeling studies.

Molecular properties and homology modeling of GST variants

The subunit molecular masses and isoelectric points of the homodimeric mutants C2 and G5, respectively, were calculated (http://web.expasy.org/compute_pi/) as 25 336 and 25 378 and 6.03 and 6.36. Both mutant proteins have 226 amino acid residues per subunit (including the His_6-tag) like the wild-type GSTU45.

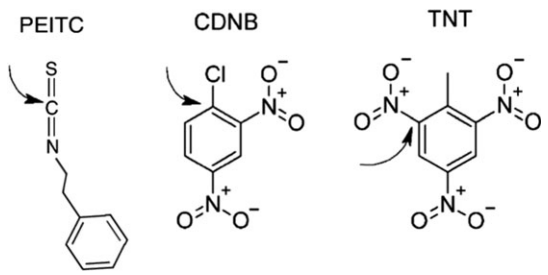

Fig. 2 Structures of alternative GST substrates used. PEITC, phenethyl isothiocyanate; CDNB, 1-chloro-2,4-dinitrobenzene; TNT, 2,4,6-trinitrotoluene; arrows mark the sites of attack by the second substrate glutathione

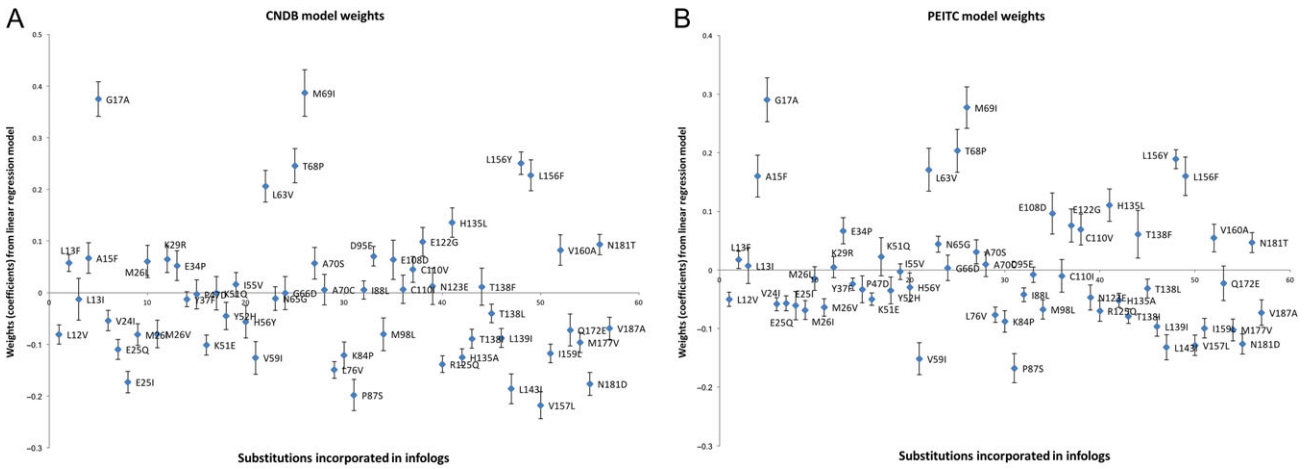

Fig. 3 Substitution weights calculated for CDNB (A) and for PEITC (B) based on the primary infolog library. Diamonds indicate mean values with standard deviations obtained using a machine-learning algorithm (Liao *et al.*, 2007). See Supplementary Table 2 for the data behind the figures

```
GSTU45    MHHHHHHAEVKLLGAWGSPF SRRVEMALKLKGVEYEYIDE DLANKSPLLLKYNPIHKKVP
C2        MHHHHHHAEVKLLGAWGSPF SRRVEMALKLKGVEYEYIDE DLANKSPLLLKYNPVHKKVP
G5        MHHHHHHAEVKLHGAWGSPF SRRVEMALKLKGVEYEYIDE DLANKSPLLLKYNPIHKKVP
          ***********:******* ******************** **************:*****

GSTU45    VLLHNGKTMAESLVILEYID ETWKSNPILPEDPYDKAMAR FWAKFIDEKCMPAIWQIMLS
C2        VLLHNGKTMAESLVILEYID ETWKSNPILPEDPYEKAMAR FWAKFIDDKCMPAIWQIMLS
G5        VLLHNGKTMCESLVILEYID ETWKSNPILPEDPYDKAMAR FWAKFIDEKCMPAIWQIMLS
          *********.********** **************:***** *******:************

GSTU45    KENEREKAIEEAIQHLKTLE NELKDKKFFGGETIGLVDIV ANFIGFWLGAAQEATGMELV
C2        KENEREKAIEEAIQHLKTLE NELKDKKFFGGETIGLVDIA ANFIGFWLGAAQEATGMELV
G5        KGNEREKAIEEAIQHLKTLE NELKDKKFFGGETIGLVDIV ANFIGFWLGAAQEATGMELV
          * ****************** ******************** ********************

GSTU45    NKERFPVLCKWIDEYANCSV VKENLPPRDKLIAFLRPRLS ASSWKY
C2        NKERFPVLCKWIDEYANCSV VKENLPPRDKLIAFLRPRLS ASSWKY
G5        NKERFPVLCKWIDEYANCSV VKENLPPRDKLIAFLRPRLS ASSWKY
          ********************* ******************** ******
```

Fig. 4 Sequence alignment of poplar GST45 with the C2 and G5 mutants selected from the second-generation infolog library. Mutated residues are indicated without asterisk (*). Residues in the H-site, based on the position of S-(p-nitrobenzyl)glutathione in the homologous GSTU45 structure are indicated in light shading.

Crystal structures of tau class GSTs in the PDB consistently show homodimeric proteins with a highly conserved glutathione-binding site (G-site) and a more variable site for electrophilic substrates (H-site). Blasting the amino acid sequences of C2 and G5 gave the highest score with a tau class GST from *Ricinus communis* (PDB designation 4J2F) with sequence identities of 54.0 and 53.1%, respectively. Homology models of C2 and G5 were obtained on the basis of the 4J2F crystal structure (Fig. 5). The 4J2F structure lacks an active-site ligand, but the previously published model of GSTU45 (Musdal and Mannervik, 2015) shows the ligand S-(p-nitrobenzyl) glutathione occupying both the G-site and the H-site. Matching the GSTU45 structure to the models using Chimera indicated the likely binding mode of this ligand in the modeled C2 and G5 structures.

The modeled subunit structures show that all the mutations in C2 and G5 are located remote from the G- and H-sites. Figure 5C shows the ligand in the active site cavity in a surface representation of a subunit of the C2 model. The functional enzyme is a dimer, and the novel residues Val55 and Asp108 marked in Fig. 5C face the neighboring subunit. However, based on comparison with the crystallographic dimeric structure of the homologous 4J2F, the mutated residues do not contact the other subunit and are distant from the active site of the neighboring subunit. The other mutated residues Glu95 and Ala160 are further away from the subunit interface as well as from the active site.

In mutant G5 two residues, Phe13 and Cys70, rest on the β-sheet buried below helix α2. Their side chains are pointing towards this helix, which is known to be relatively flexible in GSTs (Hitchens *et al.*, 2001). The third residue, Gly122, is located in the N-terminus of helix α5 in another mobile region at the surface of the protein molecule. None of the mutated residues in G5 are close the subunit interface.

Substrate specificities and steady-state kinetics of the C2 and G5 mutants

Five substrates were used to compare the specific activities of the two selected variants with the wild-type GSTU45: CDNB, undergoing aromatic substitution by glutathione; PEITC and BITC, occurring in cruciferous plants, forming thiocarbamates with glutathione; and CuOOH, an oxidant being reduced by glutathione. In addition,

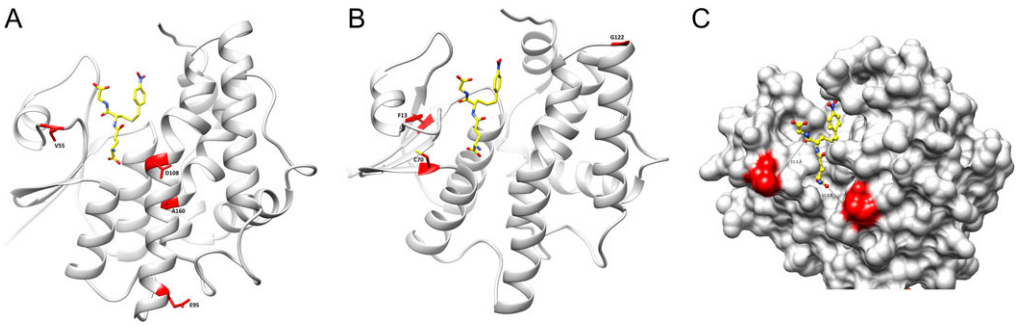

Fig. 5 Homology models of the subunits of the C2 (A) and G5 (B) mutants of poplar GSTU45. The substitutions Val55Ile/Glu95Asp/Asp108Glu/Ala160Val in (A) and Phe13Leu/Cys70Ala/Gly122Glu in (B) are marked and labeled. The active-site ligand S-(p-nitrobenzyl)glutathione is rendered (in the on-line version) in yellow with oxygen and nitrogen atoms in red and blue, respectively. Panel (C) shows the surface facing the neighboring subunit of the C2 homodimer with the two mutated residues marked

TNT, an environmental pollutant, was tested without demonstrating noteworthy activity.

Table I shows a comparison of specific activities of the two mutants and wild-type GSTU45. Mutant G5 is the most active with CDNB (13-fold higher than the wild-type activity), BITC (150-fold), and CuOOH (12-fold). By contrast, mutant C2 shows the highest increase of activity with PEITC (15-fold). C2, like G5, demonstrated a remarkable elevation of the activity with BITC (100-fold). Overall, the specific activities of C2 and G5 were similar and in a ratio of ≤ 2.5. It is also noteworthy that the observed expressivity of C2 and G5 in *E. coli* is 6–7-fold higher than that noted with the wild-type protein. This could be an effect of increased expression, increased solubility and/or increased stability.

Steady-state kinetic measurements were performed using the substrates CDNB, PEITC and CuOOH with a fixed concentration of GSH (Table II and Supplementary Fig. 1). The k_{cat} values largely reflect the elevated specific activities, as expected. However, the K_m values for CDNB are higher than for the wild-type GSTU45, whereas the K_m values of C2 and G5 for PEITC and CuOOH are significantly lower in comparison with those for wild-type GSTU45.

The majority of the mutants (77%) were expressed at levels above the GSTU45 wild-type level, but the substitutions that promoted enhanced expression were generally not the same as those enhancing CDNB or PEITC activity. Models were calculated for each of the three variables to obtain weights of the individual substitutions analyzed in the second infolog set. The data form a rectangular 25 × 3 matrix containing information about the effects of the substitutions. A Gabriel biplot can be used to illustrate the relationships between the rows and columns in matrices (Gabriel, 1971), and Fig. 6 shows a biplot of weights for the two substrates as well as for expressivity. The weights are depicted as vectors as are the two activities and the expression yield. Thus, the biplot demonstrates the influence of the 25 substitutions on the activities and on the expression in the infolog library. Clearly, the CDNB and PEITC activities are similarly affected by the mutations, whereas the expressivity is influenced differently.

Table I. Specific activities of poplar GSTU45 and mutants C2 and G5 with alternative substrates. Values are means ± SD obtained from at least three different measurements

Substrates	Specific activities (μmol/min per mg)		
	GSTU45[a]	C2	G5
CDNB	3.32 ± 0.20	17.1 ± 0.1	43.1 ± 2.2
PEITC	0.43 ± 0.04	6.4 ± 0.2	3.5 ± 0.3
BITC	0.074 ± 0.04	7.6 ± 1.1	10.8 ± 0.8
CuOOH	0.03 ± 0.002	0.29 ± 0.03	0.36 ± 0.02
TNT	$<0.1 \times 10^{-4}$	$<0.1 \times 10^{-4}$	$<0.1 \times 10^{-4}$

[a]Data from Musdal and Mannervik (2015).

Discussion

A myriad of diverse chemical reactions take place in biological organisms, and more or less all of them are enabled by the evolution of enzymes. It could therefore be proposed that new designed protein catalysts could be created for a wealth of novel chemical transformations that do not occur in Nature (Arnold, 2015; Obexer *et al.*, 2017). Such designer enzymes have value in both basic and applied research. However, the problem is how the protein polypeptide chains should be synthesized from amino acids in order to obtain a properly folded structure with the desired functions. Sequence space is essentially unlimited and catalytically active polypeptides are outstandingly rare. However, the sequences of naturally occurring enzymes help to identify areas in the multidimensional space where proteins can fold into biologically active structures. Searches of subspaces around such areas have repeatedly proven successful in enhancing protein functions and creating new activities. In protein engineering different approaches have been used ranging from introducing random mutations to the parent structure (Copp *et al.*, 2014) to recombination of related sequences by DNA shuffling (Stemmer, 1994) or other methods (Acevedo-Rocha *et al.*, 2014; Obexer *et al.*, 2016). When the 3D structure of the parent protein is known, mechanistic considerations and computational approaches can guide site-specific mutations, but the predictive power is still limited by incomplete understanding of transition states and reaction trajectories.

In general, it would appear that the creation and screening of mutants not too distant from the parental sequence is a worthwhile general approach to protein engineering. However, the construction of libraries of such mutants involves considerations of the optimal number and variability of the members. Random empirical screening of large mutant libraries may be tedious and time-consuming and require expenditure of valuable materials. A rational approach based on analysis of homologous sequences, synthetic genes and machine learning from small informative libraries has proven effective in optimizing the engineering of several enzymes *Tritirachium album* proteinase K (Liao *et al.*, 2007; Ehren *et al.*, 2008; Midlefort *et al.*, 2013;

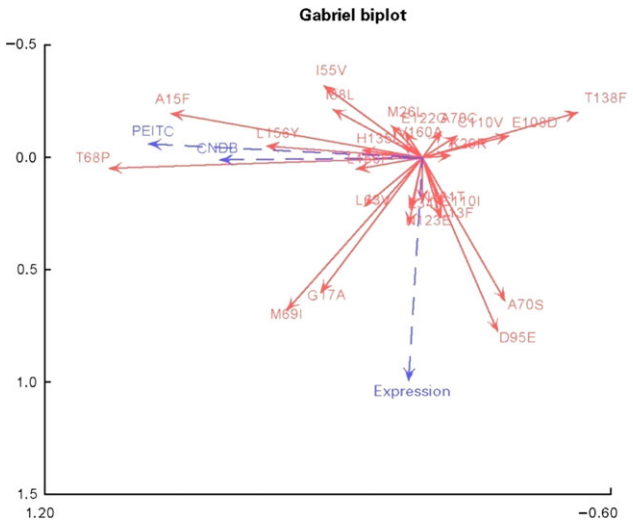

Fig. 6 Gabriel biplot demonstrating the interdependence of the 25 amino acid substitutions in the second-generation GSTU45 infolog library and their effects (dashed arrows) on catalytic activity with CDNB and PEITC as well as the expressivity of the protein. A linear PLS (mean-centered) model was fitted to each of the three dependent variables and the resulting 25 × 3 matrix of weights was represented as vectors, which in the graph are projected in two dimensions following singular value decomposition (Gabriel,1971). The axes are unit-less. The 2D representation accounted for 94% of the variance in the analysis

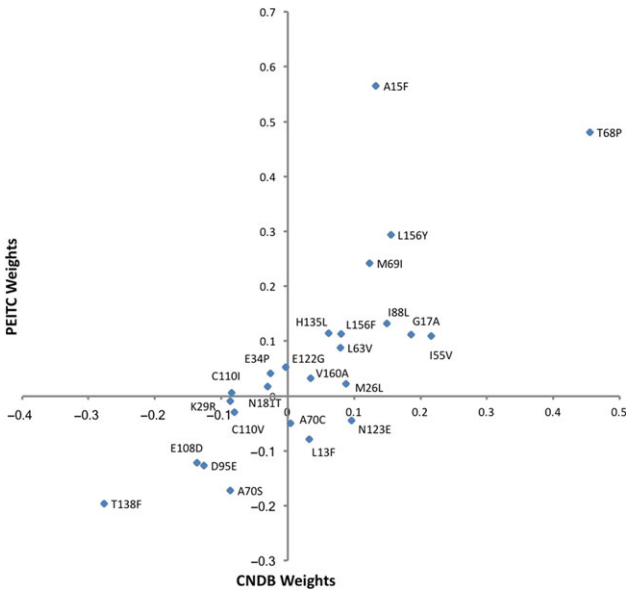

Fig. 7 Substitution weights for CDNB (*x*-axis) plotted against substitution weights for PEITC (*y*-axis) calculated from assays of the second-generation infolog library. The numbering of the substitutions includes the N-terminal His_6-sequence

Govindarajan *et al.*, 2015). By this approach the synthetic variants of the parental gene were designed to contain substitutions of amino acids identified in corresponding positions of homologous primary structures. The mutants are information-rich with respect to the desired function and are called infologs (Govindarajan *et al.*, 2015). The functional outcome of the selected substitutions provide quantitative information in each position, which allows design and execution of new rounds of synthesis and testing of improved variants. In the present investigation the infolog approach was adopted to enhance the catalytic activity of a GST from *P. trichocarpa*.

GSTU45 from *P. trichocarpa* was previously found to have very modest catalytic activities with a variety of conventional GST substrates (Musdal and Mannervik, 2015). Particular interest was directed to its potential activity with TNT in relation to phytoremediation (Tzafestas *et al.*, 2017). The present investigation was undertaken to find out to what extent the feeble activities of GSTU45 could be enhanced by protein engineering based on the infolog approach.

In the first mutant library each of the 95 infologs contained three defined point mutations. The properties of the mutants consequently sampled the information content at equal Hamming distances of three units from the wild-type in sequence space. The 45 positions mutated (some with alternative replacements) were based on substitutions identified in homologous amino acid sequences in GenBank and were distributed over the native primary structure consisting of 219 residues. In total 57 replacements were created including a few positions in which two or three alternative substitutions were introduced (e.g. Cys110Ile and Cys110Val). Each of the 95 mutant proteins was purified and tested for activity with CDNB, TNT and PEITC. With CDNB 35% and with PEITC 53% of the mutants scored higher activity than the wild-type GSTU45 activity. No reliable enzyme activity with TNT was observed.

Based on the results derived from the first infolog library, 25 amino acid substitutions were selected to construct a second set of 47 mutants to be compared with wild-type GSTU45. The expressed mutant proteins were again assayed with CDNB and PEITC, and the expression levels were also determined. With the exception of three for CDNB and two for PEITC all the mutants demonstrated significantly higher activity than the wild-type enzyme. As in the case of the first library, none of the mutants showed reliable activity with TNT. Modeling of the data from the second library again demonstrated that overall the substitutions promoting CDNB activity similarly promoted PEITC activity (Fig. 7). However, in one case (Ala15Phe) the positive effect (weight ± SD, 0.565 ± 0.077) preferentially affected the activity with PEITC.

Notably, the activities with two additional substrates, BITC and CuOOH, were enhanced in a similar way as those with CDNB and PEITC, as demonstrated with the two mutants C2 and G5. The most dramatic increase was noted with BITC, with activities elevated 100-fold in C2 and 150-fold in G5, but from a 6-fold lower wild-type value than the related isothiocyanate substrate PEITC (Table II). All in all, the substrates represent three types of reactions: double-bond addition (BITC and PEITC), hydroperoxide reduction (CuOOH) and aromatic substitution (CDNB). None of the different reaction types is selectively enhanced, with the exception of A15F, showing that the chosen mutations promote a general increase of activity.

The substitutions that enhance activity are all located far away from the bound substrates and do not introduce any catalytic functionalities. Their effects consequently are indirect and possibly involve parameters such as conformation and flexibility of the protein. It is noteworthy that the enhanced activities in most GST variants are not accompanied by significantly altered substrate selectivity. In particular, activity with TNT did not rise above the current level of detection in our assay. This outcome could be explained by the fact that the infologs in our design did not target the active-site residues, but other residues known from homologous GST proteins to be mutable. Substantial changes in substrate selectivities obtained by mutations of substrate-contacting amino acids (Nilsson *et al.*, 2000; Pettersson *et al.*, 2002; Ivarsson *et al.*, 2003; Norrgård *et al.*, 2006; Blikstad *et al.*, 2008; Shokeer and Mannervik, 2010) suggest the prospect of

Table II. Steady-state-kinetic parameters of GSTU45 and mutants C2 and G5.

h	GSTU45			C2			G5		
	K_m (mM)	k_{cat} (s^{-1})	k_{cat}/K_m $(mM\ s)^{-1}$	K_m (mM)	k_{cat} (s^{-1})	k_{cat}/K_m $(mM\ s)^{-1}$	K_m (mM)	k_{cat} (s^{-1})	k_{cat}/K_m $(mM\ s)^{-1}$
CDNB	0.29 ± 0.10	5.30 ± 0.53	18.3 ± 6.8	1.55 ± 0.21	51.3 ± 3.8	33.1 ± 5.2	0.39 ± 0.03	66.6 ± 1.7	170 ± 15
PEITC	0.036 ± 0.006	0.14 ± 0.01	3.75 ± 0.66	0.01 ± 0.001	4.97 ± 0.10	522 ± 59	0.013 ± 0.001	2.59 ± 0.06	194 ± 21
CuOOH	4.85 ± 1.20	0.27 ± 0.05	0.06 ± 0.02	3.42 ± 0.23	1.10 ± 0.04	0.32 ± 0.02	1.95 ± 0.16	1.0 ± 0.03	0.51 ± 0.04

Initial rates were determined with varying concentration of substrates at a fixed concentration of GSH and the experimental data were analyzed using nonlinear regression. The experimental data are shown in Supplementary Fig. 1.

altering the substrate selectivity also of GSTU45. Incisive application of the infolog approach to the H-site residues may be similarly effective for this purpose as in boosting existing activities as shown in the present investigation.

In summary, we demonstrate that the design of a very limited number of informative-rich gene variants enable rapid and efficient engineering of proteins with increased functional activities as well as identifying residues critical for the orthologous beneficial property of increased expression yield.

Supplementary Data

Supplementary data are available at *Protein Engineering, Design & Selection* online.

Acknowledgments

We thank Dr Birgitta Sjödin for valuable help and advice, Dr Claes Gustafsson for editing the article and Johanna Mannervik, BSc, for graphical assistance.

Funding

This work was supported by grants from the Swedish Research Council, Carl Tryggers Stiftelse and ATUM.

Authors' contributions

Y.M. performed the experiments, S.G. designed GST infologs and performed model building and machine learning, Y.M. and B.M. designed and analyzed the experiments, B.M. directed the entire project.

References

Acevedo-Rocha,C.G., Hoebenreich,S. and Reetz,M.T. (2014) *Methods Mol. Biol.*, **1179**, 103–128.

Arnold,F.H. (2015) *Q. Rev. Biophys.*, **48**, 404–410.

Blikstad,C., Shokeer,A., Kurtovic,S. and Mannervik,B. (2008) *Biochim. Biophys. Acta*, **1780**, 1458–1463.

Brentner,L.B., Mukherji,S.T., Merchie,K.M., Yoon,J.M., Schnoor,J.L. and Van Aken,B. (2008) *Chemosphere*, **73**, 657–662.

Copp,J.N., Hanson-Manful,P., Ackerley,D.F. and Patrick,W.M. (2014) *Methods Mol. Biol.*, **1179**, 3–22.

Ehren,J., Govindarajan,S., Moron,B., Minshull,J. and Khosla,C. (2008) *Protein Eng. Des. Sel.*, **21**, 699–707.

Gabriel,K.R. (1971) *Biometrika*, **58**, 453–467.

Govindarajan,S., Mannervik,B., Silverman,J.A. *et al* (2015) *ACS Synth. Biol.*, **4**, 221–227.

Gustafsson,C., Minshull,J., Govindarajan,S., Ness,J., Villalobos,A. and Welch,M. (2012) *Protein Expr. Purif.*, **83**, 37–46.

Habig,W.H., Pabst,M.J. and Jakoby,W.B. (1974) *J. Biol. Chem.*, **249**, 7130–7139.

Hitchens,T.K., Mannervik,B. and Rule,G.S. (2001) *Biochemistry*, **40**, 11660–11669.

Huang,P.S., Boyken,S.E. and Baker,D. (2016) *Nature*, **537**, 320–327.

Ivarsson,Y., Mackey,A.J., Edalat,M., Pearson,W.R. and Mannervik,B. (2003) *J. Biol. Chem.*, **278**, 8733–8738.

Kolm,R.H., Stenberg,G., Widersten,M. and Mannervik,B. (1995a) *Protein Expr. Purif.*, **6**, 265–271.

Kolm,R.H., Danielson,U.H., Zhang,Y., Talalay,P. and Mannervik,B. (1995b) *Biochem. J.*, **311**, 453–459.

Laurino,P., Rockah-Shmuel,L. and Tawfik,D.S. (2016) *Adv. Exp. Med. Biol.*, **945**, 491–509.

Lawrence,R.A. and Burk,R.F. (1976) *Biochem. Biophys. Res. Commun.*, **71**, 952–958.

Liao,J., Warmuth,M.K., Govindarajan,S., Ness,J.E., Wang,R.P., Gustafsson, C. and Minshull,J. (2007) *BMC Biotechnol.*, **7**, 16.

Mazari,A.M. and Mannervik,B. (2016) *Biochem. Biophys. Rep.*, **5**, 141–145.

Midelfort,K.S., Kumar,R., Han,S. *et al* (2013) *Protein Eng. Des. Sel.*, **26**, 25–33.

Musdal,Y. and Mannervik,B. (2015) *Biochim. Biophys. Acta*, **1850**, 1877–1883.

Nilsson,L.O., Gustafsson,A. and Mannervik,B. (2000) *Proc. Natl. Acad. Sci. USA*, **97**, 9408–9412.

Norrgård,M.A., Ivarsson,Y., Tars,K. and Mannervik,B. (2006) *Proc. Natl. Acad. Sci. USA*, **103**, 4876–4881.

Obexer,R., Godina,A., Garrabou,X., Mittl,P.R., Baker,D., Griffiths,A.D. and Hilvert,D. (2017) *Nat. Chem.*, **9**, 50–56.

Obexer,R., Pott,M., Zeymer,C., Griffiths,A.D. and Hilvert,D. (2016) *Protein Eng. Des. Sel.*, **29**, 355–366.

Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) *J. Comput. Chem.*, **25**, 1605–1612.

Pettersson,P.L., Johansson,A.S. and Mannervik,B. (2002) *J. Biol. Chem.*, **277**, 30019–30022.

Porath,J., Carlsson,J., Olsson,I. and Belfrage,G. (1975) *Nature*, **258**, 598–599.

Shokeer,A. and Mannervik,B. (2010) *J. Biol. Chem.*, **285**, 5639–5645.

Sievers,F. and Higgins,D.G. (2014) *Methods Mol. Biol.*, **1079**, 105–116.

Stemmer,W.P.C. (1994) *Nature*, **370**, 389–391.

Tiwari,V. (2016) *Front. Chem.*, **4**, 39.

Tzafestas,K., Razalan,M.M., Gyulev,I., Mazari,A.M., Mannervik,B., Rylott, E.L. and Bruce,N.C. (2017) *New. Phytol.*, **214**, 294–303.

Zhang,W., Dourado,D.F. and Mannervik,B. (2015) *Biochim. Biophys. Acta*, **1850**, 742–749.