Protein Engineering, Design & Selection vol. 23 no. 11 pp. 871–880, 2010
Published online September 16, 2010 doi:10.1093/protein/gzq063

Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination

Pete Heinzelman, Russell Komor, Arvind Kanaan, Philip Romero, Xinlin Yu, Shannon Mohler, Christopher Snow and Frances Arnold[1]

Division of Chemistry and Chemical Engineering, California Institute of Technology 210-41, Pasadena, CA 91125, USA

[1]To whom correspondence should be addressed.
E-mail: frances@cheme.caltech.edu

Received August 18, 2010; revised August 18, 2010; accepted August 24, 2010

Edited by Stephen Withers

We describe an efficient SCHEMA recombination-based approach for screening homologous enzymes to identify stabilizing amino acid sequence blocks. This approach has been used to generate active, thermostable cellobiohydrolase class I (CBH I) enzymes from the 390 625 possible chimeras that can be made by swapping eight blocks from five fungal homologs. Constructing and characterizing the parent enzymes and just 32 'monomeras' containing a single block from a homologous enzyme allowed stability contributions to be assigned to 36 of the 40 blocks from which the CBH I chimeras can be assembled. Sixteen of 16 predicted thermostable chimeras, with an average of 37 mutations relative to the closest parent, are more thermostable than the most stable parent CBH I, from the thermophilic fungus *Talaromyces emersonii*. Whereas none of the parent CBH Is were active >65°C, stable CBH I chimeras hydrolyzed solid cellulose at 70°C. In addition to providing a collection of diverse, thermostable CBH Is that can complement previously described stable CBH II chimeras (Heinzelman *et al.*, *Proc. Natl Acad. Sci. USA* 2009;106:5610–5615) in formulating application-specific cellulase mixtures, the results show the utility of SCHEMA recombination for screening large swaths of natural enzyme sequence space for desirable amino acid blocks.

Keywords: biofuel/cellulase/directed evolution/protein recombination/protein thermostability/CBH I/Cel7A

Introduction

The high cost of the fungal cellulase mixtures that are commonly employed in biomass-to-biofuel conversion processes is one of the major limitations to achieving economically viable production of transportation fuel from inedible plant matter. The operating costs of cellulase treatments can be reduced by improving the thermostability of these enzyme mixtures (Viikari *et al.*, 2007). Cellulase operating lifetime increases with thermostability, allowing thermostable cellulases to hydrolyze more cellulose per unit enzyme than their less stable counterparts throughout the course of biomass degradation. Thermostable cellulases can also operate at higher temperatures and derive a benefit from higher specific activities. This increased hydrolysis reduces the enzyme loading needed to convert cellulosic biomass into fermentable sugars. In addition to stability, properties such as specific activity, pH dependence, product inhibition and productive versus non-productive adsorption on solid substrate surfaces all contribute to the overall performance of a cellulase mixture (Gusakov *et al.*, 2007).

Class I cellobiohydrolases (CBH Is or family 7 glycosyl hydrolases) are the principal components of industrial cellulase mixtures and account for ~60 wt% of the cellulases secreted by the prevalent commercial cellulase production host, the filamentous fungus *Hypocrea jecorina* (*Trichoderma reesei*) (Le Crom *et al.*, 2009). As such, CBH Is have been the subject of multiple enzyme engineering efforts aimed primarily at improving CBH I thermostability. Both high throughput screening (HTS) of CBH I random mutant libraries (Voutilainen *et al.*, 2009) and rational disulfide bond engineering (Voutilainen *et al.*, 2010) have been employed to create stable CBH I variants. The applicability of HTS is limited to CBH Is that are expressed by a suitable heterologous host at levels sufficient to enable library characterization. The applicability of disulfide bond engineering is limited to CBH Is for which a crystal structure exists. Neither of these approaches generate the CBH I gene sequence diversity that could lead to improvements in the suite of enzyme properties enumerated above. Here, we describe a method for engineering cellulases (and other proteins) that reliably improves thermostability while simultaneously retaining function and providing a high level of sequence diversity.

SCHEMA structure-guided recombination allows multiple related enzymes, with identities as low as 30% (Meyer *et al.*, 2006) and perhaps even lower, to be recombined to yield enzyme chimeras that exhibit improved thermostability (Li *et al.*, 2007; Heinzelman *et al.*, 2009a,b) as well as diversity with respect to properties such as substrate specificity (Li *et al.*, 2007) and pH dependence (Heinzelman *et al.*, 2009a). The ability to accurately predict thermostable sequences from a small sample of the chimeras (Meyer *et al.*, 2006) eliminates the need for HTS in stability engineering. The requirements for applying SCHEMA are minimal: the sequences of the parent enzymes and a structure for either a parent enzyme or homolog suffice for identifying suitable block boundaries for recombination.

SCHEMA uses protein structure data to define boundaries of amino acid 'blocks' which minimize $<E>$, the family average number of side-chain contacts that are broken when the blocks are swapped. An additional important parameter is $<m>$, the family average number of mutations. As $<m>$ reflects the diversity of the sequences in a chimera family, all

else being equal, one seeks to identify crossovers that generate high $<m>$ values. The RASPP (recombination as shortest path problem) algorithm (Endelman *et al.*, 2004) can be used to identify the block boundaries that minimize $<E>$ relative to $<m>$, thus maximizing the fraction of chimera family members that are folded and functional for a given level of sequence diversity. The computational tools required to implement SCHEMA and RASPP and instructions for applying them can be found at http://www.che.caltech.edu/groups/fha/schema-tools/schema-overview.html.

To date, SCHEMA chimera families have been constructed from pools of 24 blocks, where each of three parent genes has been divided into eight blocks. This design was initially chosen (Meyer *et al.*, 2006) based on the desire to balance chimera sequence diversity, the number of sequences that must be sampled to predict the most thermostable chimeras in the family and the fraction of chimeras that are catalytically active. Sequence diversity can be increased by recombining more than three parents and/or dividing the recombination parents into more than eight blocks, with a concomitant increase in the labor required for representative sampling of chimeras from such expanded recombination pools.

We recently used SCHEMA recombination to create a family of thermostable fungal CBH II (family 6 glycosyl hydrolase) chimeras (Heinzelman *et al.*, 2009a). This family featured the three-parent, eight-block construction. Because recombination makes mutations outside the highly conserved cellulase active site, properly folded enzyme chimeras usually retain cellulase activity. Recombination is nonetheless disruptive to the folded structure, and SCHEMA serves to increase the fraction of properly folded chimeras. The 6561-member CBH II SCHEMA family contains more than 3000 catalytically active enzymes, and a linear regression model for CBH II thermostability predicted that more than one-third of these are more thermostable than any of the three fungal enzymes from which they were assembled (Heinzelman *et al.*, 2009a,b). Stability increases also translated to improved cellulase performance: seven of the eight thermostable chimeras studied hydrolyzed more solid cellulose than the parent CBH IIs in long-time solid cellulose hydrolysis assays (Heinzelman *et al.*, 2009b). Some of the stable CBH II chimeras exhibited broadened pH-activity profiles. Further investigations allowed us to identify a single Cys–Ser mutation that markedly increased the thermostability of the parent CBH IIs that did not already contain it as well as a wild-type CBH II not included in the recombination parent set (Heinzelman *et al.*, 2009b).

For the CBH II family, a sample set of 48 chimeras was constructed by total gene synthesis. However, fungal cellulases are secreted at low levels by *S. cerevisiae*, and only 23 of the sample set chimeras were secreted at levels sufficient for the thermostability measurement. These 23 measurements nonetheless allowed us to make qualitative predictions that accurately identified the sequences of chimeras more thermostable than the parent CBH IIs. Constructing and measuring the stabilities of an additional 31 predicted thermostable CBH II chimeras allowed development of a quantitative stability model (Heinzelman *et al.*, 2009b). This model showed that the blocks made additive contributions to chimera thermostability and also indicated that the qualitative stability predictions had not overlooked any exceptionally thermostable CBH II chimera sequences.

For the current work, we used the CBH II sample set expression and modeling results to guide the design of a CBH I chimera sample set corresponding to an eight-block, five-parent family containing more than 390 000 unique sequences. In order to predict the most stable members of this chimera family while still sampling only a limited set of chimeric genes (~30–40), we used past experience to simplify the sample set design and maximize the number of sample genes expected to be secreted in functional form. In particular, we hypothesized that SCHEMA blocks would make additive or at least cumulative contributions to chimera stability. We further assumed that using a highly expressed parent as the background into which single blocks from homologous parents are substituted would increase the probability that the sample sequence will be secreted and functional. Thus, we constructed a set of CBH I 'monomeras', chimeras that contain a single-block substitution, in the background of a well-expressed parent enzyme. We show that this is an efficient approach for rapidly screening homologous enzymes for stabilizing blocks of sequence. The task of predicting the most stable chimeras is reduced to making stability measurements for the parent enzyme and 32 monomeras made in that background. Diverse thermostable chimeras can then be assembled from stabilizing and neutral blocks.

Materials and methods

Parent and chimeric genes encoding CBH I enzymes were cloned into yeast expression vector YEp352/PGK91-1-αss and transformed into expression strain YDR483W as described (Heinzelman *et al.*, 2009a). Parent CBH I genes featured native codon usage and were synthesized by DNA 2.0 (Menlo Park, CA, USA). Five milliliter synthetic dextrose casamino acids (SDCAA) media starter cultures were grown overnight at 30°C with shaking at 225 rpm, expanded into 40 ml of yeast peptone dextrose (YPD) medium and incubated for 48 h. Culture supernatants were brought to 1 mM phenylmethylsulfonyl fluoride and 0.02% NaN_3.

Total yeast-secreted CBH I activity toward the soluble substrate 4-methylumbelliferyl lactopyranoside (MUL) was determined by adding 125 μl of culture supernatant to 25 μl of 1.8 mM MUL (Sigma) dissolved in 750 mM sodium acetate, pH 4.8, incubating at 45°C for 30 min and quenching with 150 μl of 1 M Na_2CO_3. MUL hydrolysis rates were determined by using a microplate reader to measure sample fluorescence with excitation at 365 nm and emission at 445 nm and comparing values to a standard curve prepared with 4-methylumbelliferone (Sigma).

The T_{50} value is defined as the temperature at which a 10-min incubation in the absence of substrate causes loss of one-half of the activity, measured after reaction with MUL substrate, relative to a 100% activity reference sample that does not undergo the incubation. For T_{50} assays, culture supernatants were diluted using a supernatant from a negative control YPD yeast culture not containing secreted cellulase so that approximately equivalent MUL hydrolysis rates of 1.6×10^{-5} mol/l/s were obtained for samples not incubated for thermal denaturation. These diluted samples were adjusted to 1 mM DTT and 125 mM sodium acetate, pH 4.8. Aliquots of 125 μl were incubated for 10 min in a water bath across a range of temperatures bracketing the T_{50} value.

Water bath temperatures were measured using two different alcohol thermometers and observed to be consistent within 0.1°C. After cooling, 25 μl of 1.8 mM MUL in 50 mM sodium acetate, pH 4.8, was added to the incubated sample and an unheated sample, and these were incubated in a 45°C water bath for 90 min. MUL hydrolysis was determined as above, and the T_{50} value was calculated by linear interpolation of data using Microsoft Excel. A representative T_{50} data set is provided in SI 7.

To determine total yeast-secreted CBH I activity toward solid cellulose, 500 μl of yeast culture supernatant was incubated with 500 μl of 120 mg/ml Lattice NT microcrystalline cellulose (FMC) in 50 mM sodium acetate, pH 4.8, for 1 h at 4°C in a thermal block with shaking at 1000 rpm. Samples were centrifuged at 3000 rcf for 3 min and washed with 1 ml of ice-cold 50-mM sodium acetate, pH 4.8, containing 1 mg/ml of BSA. Solid cellulose with bound CBH I was resuspended in 1 ml of same buffer, incubated with shaking at 37°C for 90 min and the amount of reducing sugar in the reaction supernatant was determined by the Nelson–Somogyi assay as described (Heinzelman *et al.*, 2009a).

Ni^{2+} affinity-isolated CBH I sample preparation, protein concentration measurement and SDS-PAGE analyses were performed as described for CBH II (Heinzelman *et al.*, 2009a). Post-Ni^{2+} isolation CBH I yield estimates range from 500 μg/l culture for the poorly secreted *Thermoascus aurantiacus* parent CBH I to between 5 and 10 mg/l for the *Talaromyces emersonii* parent CBH I and most highly secreted CBH I chimeras. CBH I solid cellulose temperature activity profiles were obtained by assuming that all protein in the affinity-isolated CBH I samples was fully active CBH I and adding 4 μg to 270 μl of 50 mM sodium acetate, pH 4.8, containing 60 mg/ml Lattice NT cellulose. After incubation for 16 h in a water bath at the temperature of interest, supernatant reducing sugar was determined by the Nelson–Somogyi assay as above.

For *T. emersonii* CBH I circular dichroism (CD) and half-life $t_{1/2}$ thermostability comparison experiments, metal affinity-isolated CBH I samples were treated as 100% fully active CBH I, and enzyme, substrate and buffer conditions identical to those described (Voutilainen *et al.*, 2010) were used. Half-life assay samples in which CBH I was supplied by addition of culture supernatants received supernatant containing MUL-hydrolyzing CBH I activity approximately equal to that added to the assays performed with affinity-isolated CBH I. CBH I deglycosylation was performed using PNGaseF (New England Biolabs) as per the manufacturer's instructions with a CBH I concentration of 100 μg/ml. CBH I secretion culturing of the described (Gusakov *et al.*, 2007) hyperglycosylating yeast strain was performed as above with the exception that overnight starter cultures were grown in synthetic dropout-Uracil media prior to expansion into YPD.

Results

Parent fungal CBH I enzymes

Four of the five CBH I recombination parents, from the filamentous fungi *Chaetomium thermophilum* (P1), *T. aurantiacus* (P2), *H. jecorina* (P3) and *Acremonium thermophilum* (P4), were chosen on the basis of their having been overexpressed from the popular industrial cellulase secretion host *T. reesei* (teleomorph *H. jecorina*; Voutilainen *et al.*, 2008), which is important for industrial applications. The fifth CBH I (P5), from the thermophilic fungus *T. emersonii*, was included by virtue of its reported high thermostability (Voutilainen *et al.*, 2010). To eliminate the possibility of generating unpaired Cys residues upon recombination, residues G4 and A72 in the *T. emersonii* and *T. aurantiacus* CBH Is were changed to Cys, so that each parent CBH I catalytic domain contained 10 disulfide bonds. A sequence alignment of the five parent catalytic domains appears in SI 1, and the catalytic domain pairwise sequence identities are shown in SI 2. The *T. emersonii* and *T. aurantiacus* CBH Is, which do not contain carbohydrate binding modules (CBMs), were appended with the C-terminal linker and CBM from the *H. jecorina* CBH I, mimicking a construction previously used for heterologous expression of the *T. aurantiacus* CBH I (Voutilainen *et al.*, 2008). The *C. thermophilum*, *H. jecorina* and *A. thermophilum* parent genes featured their respective wild-type linkers and CBMs. The sequences for all of the CBH I parents are provided in SI 3.

As shown in SI 4, the *T. emersonii* CBH I showed much higher expression than the other four parents in an SDS-PAGE gel. *Talaromyces emersonii* yeast secretion culture supernatant also contained more than three times the activity [$(2.3 \pm 0.3) \times 10^{-4}$ mol/l/s under conditions described in the 'Materials and Methods' section] toward the fluorescent, soluble CBH I substrate MUL than supernatant for the second most highly expressed parent, from *C. thermophilum*. Accurate CBH I thermostability measurements, in the form of 10-min T_{50} values, required a total MUL hydrolysis rate of $\geq 1.6 \times 10^{-5}$ mol/l/s. Neither the *H. jecorina* (P3) nor the *A. thermophilum* (P4) parents reached this threshold. We classified CBH Is with supernatant activity values below this level as 'not secreted'. The *T. emersonii* parent had a T_{50} (62.9 ± 0.3°C) greater than those of the *C. thermophilum* (59.9 ± 0.3°C) and *T. aurantiacus* (62.2 ± 0.4°C) parents. The relatively high stability and secretion of *T. emersonii* CBH I motivated our choosing it as the background for screening sequence blocks from other enzymes.

SCHEMA chimera family design

The *T. emersonii* CBH I crystal structure (pdb 1Q9H; Grassick *et al.*, 2004) was used to prepare the contact map used by SCHEMA to evaluate disruption upon recombination, which is needed by RASPP for choosing the block boundaries that minimize library average disruption $<E>$. As crystal structures for neither a CBH I linker nor CBM are available, SCHEMA recombination was applied only to the CBH I catalytic domain. CBH I chimeras therefore contain the linker and CBM corresponding to the parent represented at block 8. An analysis of the five-parent, eight-block family designs returned by the RASPP algorithm led us to choose the block boundaries depicted in Fig. 1. The $5^8 = 390\,625$ chimeras in this family have $<E> = 20.3$ and $<m> = 66.0$, providing a desirable balance between a large number of mutations and a low number of broken contacts.

Sample chimeras for stability analysis

Fungal CBH Is are poorly secreted from the *Saccharomyces cerevisiae* host. To maximize the fraction of sample set chimeras that provide useful data, we have implemented a block screening strategy in which 32 blocks from four parents are

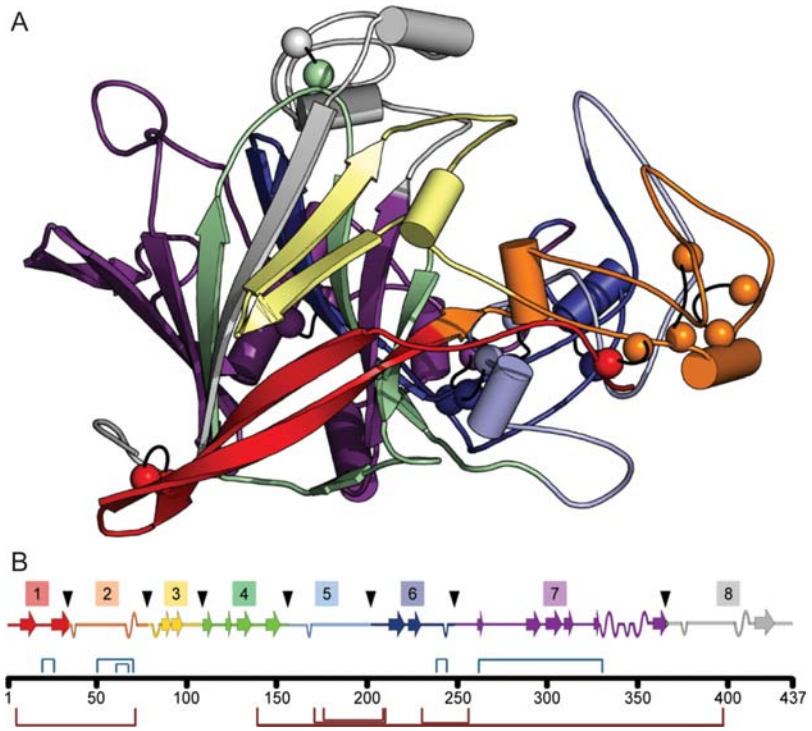

Fig. 1. (**A**) CBH I catalytic domain three-dimensional structure. Disulfide-bonded Cys residues are denoted by spheres connected by black lines. (**B**) CBH I recombination block divisions and secondary structure diagram. Interblock disulfide bonds are denoted by maroon lines, intrablock disulfides by light blue lines and block divisions by black arrows. Residue numbering for *T. emersonii* CBH I.

substituted, one at a time, into the background of a CBH I that is secreted at relatively high levels (parent 5). The 32-member CBH I 'monomera' sample set has $<E> = 5.9$ and $<m> = 15.6$. These are considerably lower than the average values of the 390 625 sequences in the family and are therefore expected to have a high likelihood of retaining fold and cellulase function.

The 32 monomeras were prepared by total gene synthesis. As shown in Fig. 2, 28 monomeras (88%) were secreted in functional form from *S. cerevisiae*. However, none of the monomeras containing substitutions at block 7, the largest block, were secreted. Substitutions at block 4 were also highly detrimental to monomera secretion. On the other hand, several of the monomeras with substitutions at blocks 2 and 5 were more highly secreted than the *T. emersonii* background parent. Although we previously observed an inverse relationship between *E* and secretion for CBH II chimeras (Heinzelman *et al.*, 2009a), *E* is not predictive of CBH I monomera secretion (SI 5). Our treating total secretion culture supernatant CBH I supernatant activity toward MUL as a proxy for CBH I secretion is based on our observation that Ni-NTA affinity-isolated, C-terminally His_6-tagged CBH I parents and chimeras have similar specific activities toward MUL (SI 6). We did not attempt to determine whether the non-secreted or poorly secreted monomeras were expressed but retained within the host cells.

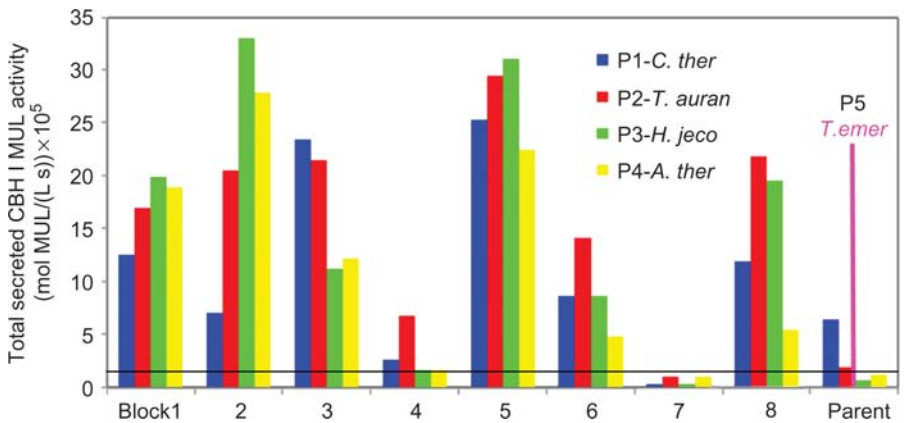

Fig. 2. Total secreted CBH I MUL-hydrolyzing activity for parent CBH Is and 32 monomeras. Monomeras contain single-block substitutions from parents 1–4 into parent 5 (from *T. emersonii*). Total secreted CBH I MUL-hydrolyzing activity for *T. emersonii* CBH I denoted by a pink bar. Yeast secretion culture supernatants were incubated with 300 μM soluble, fluorescent MUL substrate for 30 min at 45°C. Mean of single activity measurements for three independent *T. emersonii* secretion cultures is 2.3×10^{-4} mol MUL/(l s), standard deviation is 3.0×10^{-5} mol MUL/(l s). All other values represent single cultures and measurements. Black line at the bottom of figure denotes the threshold activity value of 1.6×10^{-5} mol MUL/(l s) for T_{50} measurement.

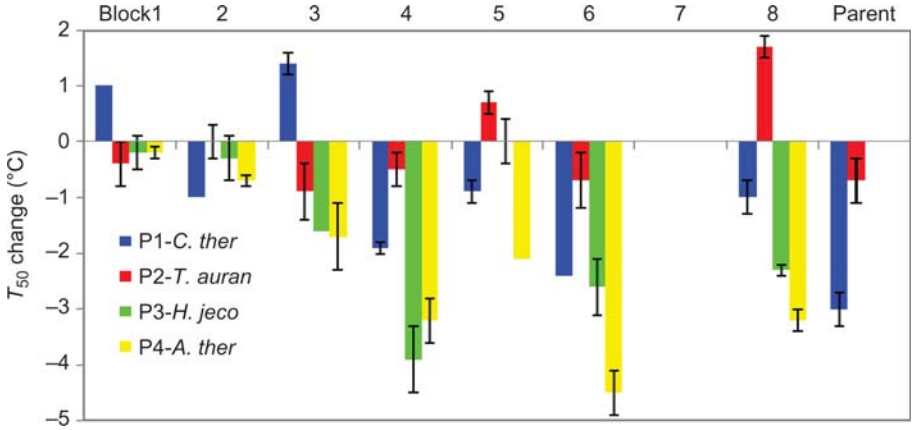

Fig. 3. Changes in the T_{50} values (°C) relative to *T. emersonii* ($T_{50} = 62.9 \pm 0.3$°C) parent for 28 CBH I monomeras. Monomeras contain single-block substitutions from parents 1–4 into parent 5 (from *T. emersonii*). Error bars for monomeras represent the extreme values of two duplicate measurements. Error bars for parents represent the standard deviations for between 3 and 8 replicates. T_{50} values for respective *C. thermophilum* and *T. aurantiacus* parent CBH Is are 59.9 ± 0.3°C and 62.2 ± 0.4°C. *Hypocrea jecorina* and *A. thermophilum* parent CBH Is were not secreted.

Fig. 3 summarizes the block stability contributions and shows that four-block substitutions resulted in CBH I chimeras with increased T_{50} values; the stabilizing blocks B1P1, B3P1, B5P2 and B8P2 increase the T_{50} by between ~0.7°C and ~1.6°C. Although no stabilizing blocks were obtained from the two parents that were not secreted (P3 and P4), these parents did provide five neutral blocks, B1P3, B2P3, B5P3, B1P4 and B2P4. Assuming that block neutrality is independent of chimera background, these blocks can be used to increase chimera sequence diversity without reducing thermostability.

Of the 10 disulfide bonds in CBH I, five involve Cys residues originating from different blocks. For example, Cys135 (parent 5 numbering) of block 4 forms a disulfide bond with Cys401 of block 8, and Cys253 of block 7 is paired with Cys227 of block 6 (Fig. 1). We explored whether the recombination of disulfide-bonded Cys pairs was responsible for the detrimental effect on secretion and/or stability that results from the block 4 and 7 substitutions. We tested this by substituting the 4–8 and 6–7 block pairs from parents 1 and 2 into parent 5. Conserving the disulfides in this way, however, resulted in expression levels that fall between those of the monomeras containing the respective single-block substitutions or are not secreted at all (Table I). T_{50} values for the chimeras with substitutions at blocks 4 and 8 fall between those for the respective block 4 and 8 monomeras. These results show that C135–C401 and C227–C253 disulfide bonds containing Cys residues in blocks taken from different parents do not reduce secretion or stability relative to these blocks coming from the same parent.

The lack of secretion for block 7-substituted monomeras prevented the assignment of stability contributions to blocks at this position. Only one monomera in which B7P5 was substituted into the other four parents was secreted, where moving block 7 into parent 2, which has the highest identity (81%) to parent 5, increased expression more than 5-fold and increased the T_{50} of parent 2 by 1.5°C.

Thermostable CBH I chimera design and characterization

We then sought to design a set of diverse, thermostable chimeras that would also be secreted at relatively high levels. To achieve high stability, all 16 members of this set include the two most stabilizing blocks, B3P1 and B8P5. Similarly, as both B5P3 and B5P5 were observed to have significant and similar stabilizing effects, all of the designed chimeras contain one of these two blocks. As blocks 4 and 7 from parents other than *T. emersonii* parent 5 were found to either eliminate or markedly reduce secretion, all 16 designed chimeras feature both B4P5 and B7P5. Finally, to obtain high sequence diversity without sacrificing thermostability and/or secretion level, we incorporated a collection of 11 blocks, B1P1, B1P2, B1P3, B1P4, B1P5, B2P2, B2P3, B2P4, B2P5, B6P2 and B6P5, that were expected to be either beneficial or neutral with respect to the chimera stability and secretion level.

The chimeras are thus comprised of 17 of the 40 available CBH I blocks and contain an average of 37 mutations relative to the closest parent (of 441 total residues). They differ from each other by 21 mutations on average and give representation to all five parent CBH Is. As shown in Fig. 4, all 16 of these predicted-stable CBH I chimeras in fact have T_{50} values that are significantly greater than that of the most stable CBH I parent (from *T. emersonii*). Eight of the 16 thermostable chimeras have T_{50} values that are 2 or more

Table I. Total yeast-secreted MUL activity and T_{50} values for disulfide-paired CBH I chimeras and underlying monomeras

Chimera	Total secreted activity (mol MUL/(l s) × 10^5)	T_{50} (°C)
55 515 555	2.6	61.0 ± 0.1
55 555 551	11.8	61.9 ± 0.2
55 515 551	6.3	58.2 ± 0.3
55 525 555	6.7	62.4 ± 0.2
55 555 552	21.8	64.6 ± 0.2
55 525 552	11.0	63.1 ± 0.1
55 555 155	8.6	60.5 ± 0.0
55 555 515	0.3	NS
55 555 115	0.1	NS
55 555 255	14.1	63.2 ± 0.5
55 555 525	1.0	NS
55 555 225	0.2	NS

T_{50} values represent extremes of two duplicate measurements, MUL activity values for a single measurement of a single culture, 300 μM MUL, 30-min incubation at 45°C. NS indicates insufficient secretion for T_{50} measurement.

CBH I sequence	CBH I parent represented at each block position	T_{50} (°C)	Secreted activity
11111111		59.9±0.5	7.5
22222222		62.2±0.4	1.9
33333333		NS	0.6
44444444		NS	1.1
55555555		62.9±0.3	23.0
34152252		64.0±0.1	22.6
55153552		64.3±0.0	33.2
32153252		64.3±0.2	10.2
55155552		64.4±0.7	21.9
22153252		64.4±0.2	12.8
52152552		64.5±0.0	34.3
12153252		64.7±0.2	6.4
45153252		64.8±0.2	25.3
12153552		64.9±0.3	10.9
25152252		65.0±0.1	22.2
13152552		65.0±0.0	34.7
12152252		65.3±0.1	10.2
55153252		65.3±0.2	20.0
55552252		65.6±0.7	18.5
55152552		65.7±0.1	29.4
55152252		66.3±1.0	19.6

Fig. 4. T_{50} values, total yeast-secreted activity (mol MUL/(l s) $\times 10^5$) and block sequences for parent CBH Is. T_{50} error bars for monomeras represent the extreme values of two duplicate measurements, and error bars for parents represent the standard deviations for between 3 and 8 replicates. Total secreted activity values [mol MUL/(l s)] are a single measurement for a single culture, with the exception of parent 5, *T. emersonii*, which has mean and standard deviation total yeast-secreted activity of $(2.3 \pm 0.3) \times 10^{-4}$ mol MUL/(l s) for single measurements of three independent cultures. Secretion levels for parent 3 (*H. jecorina*) and parent 4 (*A. thermophilum*) are below the threshold for the T_{50} measurement.

degrees above *T. emersonii*, with the most thermostable chimera, 55 152 552, having a T_{50} that is higher by 3.4°C. As shown in Fig. 4, all but one of the 16 stable chimeras are secreted at levels equal to or greater than that for the second most highly secreted parent, from *C. thermophilum*, and 8 chimeras were secreted at levels equal to or greater than that for the most highly secreted parent, from *T. emersonii*.

As our attempts to substitute B7P5 into the backgrounds of the four other parents were successful only for parent 2, the parent most identical to parent 5, we substituted B7P2 for B7P5 in the background of five thermostable chimeras. As shown in SI 8, this substitution either markedly reduced or abrogated secretion in all five cases and decreased secreted chimera T_{50} values by an average of 2.3 ± 0.8°C.

We next explored whether smaller stretches of amino acids, or sub-blocks, lying within block 7 could be swapped in chimeric CBH Is and whether these blocks could make positive thermostability contributions. We continued in the mode of interchanging sequence between the two most identical sequences, parents 2 and 5. We selected six sub-blocks within B7P2, chosen on the basis of cloning convenience and a relatively equal distribution of the 32 mutations separating B7P2 and B7P5. As shown in SI 9, the six sub-blocks feature between 2 and 7 mutations. SI 10 shows that three of the six sub-blocks (C, D and E) either increase or do not reduce secretion when substituted into parent 5. Sub-block C, which contains six mutations, was found to increase the T_{50} of *T. emersonii* CBH I by ~1.0°C.

As shown in Table II, where B7P5 containing sub-block C from parent 2 is denoted by '*5' at position 7, this sub-block improved the thermostability of all five chimeras into which it was substituted, with an average T_{50} increase of 1.5 ± 0.4°C. Furthermore, the B7P*5 chimeras are all secreted at higher levels than the corresponding B7P5 chimeras.

Table II. Total yeast-secreted MUL activity (mol MUL/(l s) $\times 10^5$) and T_{50} values for B7P5 chimeras and corresponding B7P*5 substituted chimeras

B7P5 chimera	Total secreted activity	T_{50} (°C)	B7P*5 chimera	Total secreted activity	T_{50} (°C)
55 153 552	33.2	64.3 ± 0.0	551 535*52	42.2	65.7 ± 0.2
12 153 252	6.4	64.7 ± 0.2	121 532*52	10.6	66.0 ± 0.0
25 152 252	22.2	65.0 ± 0.1	251 522*52	28.7	66.8 ± 0.1
12 152 252	10.2	65.3 ± 0.1	121 522*52	17.7	66.9 ± 0.1
55 152 252	19.6	66.3 ± 1.0	551 522*52	34.0	66.9 ± 0.1

T_{50} values represent extremes of two duplicate measurements, MUL activity values for a single measurement of a single culture, 300 μM MUL, 30-min incubation at 45°C.

Cellulose hydrolysis using thermostable CBH I chimeras

We sought to determine whether an increase in T_{50}, which is measured after thermal denaturation in the absence of substrate, corresponds to an increase in the maximum CBH I solid cellulose hydrolysis temperature. To this end, we built yeast secretion constructs for the three secreted CBH I parents and five thermostable B7P*5 chimeras in which the CBH I N-terminus was appended with a His_6 tag to allow purification by Ni-NTA affinity chromatography from the components in the yeast culture medium. As shown in SI 11, although CBH I bands appear at the anticipated molecular weight of ~60 kDa in the SDS-PAGE, there are also unexpected bands at ~20 kDa. Although these samples are not sufficiently homogeneous to permit CBH I specific activity measurements, the removal of background protein and carbohydrates from the medium allows valid comparison of maximum solid cellulose hydrolysis temperatures.

As shown in Fig. 5, higher T_{50} values are indicative of a greater ability to hydrolyze solid cellulose at elevated temperatures over a 16-h interval. Whereas none of the parent enzymes were active at temperatures >65°C, all five of the tested thermostable chimeras, which contain an average of 42 mutations and differ from each other by an average of 16 mutations, retained some hydrolytic activity at 70°C. The five tested thermostable chimeras all have between 30% and 50% lower specific activity, however, than the *T. emersonii* parent at 50°C (assuming that all of the protein loaded into each reaction is active CBH I).

The Ni-NTA affinity-isolated CBH I samples are also useful for evaluating whether CBH I specific activities toward the soluble MUL substrate, measured at 45°C, are retained upon recombination. As shown in SI 6, the estimated specific activities of the five thermostable His_6-tagged chimeras, based on the assumption that the affinity-isolated CBH I samples are 100% pure, lie within 4×10^{-5} mol MUL/(l s μg CBH I) of the mean value of 2.8×10^{-4} mol MUL/(l s μg CBH I). These specific activities fall between the respective values of $(4.3 \pm 0.1) \times 10^{-4}$, $(2.3 \pm 0.2) \times 10^{-4}$ and $(4.3 \pm 0.1) \times 10^{-4}$ mol MUL/(l s μg CBH I) measured for parents 1, 2 and 5. Thus, the thermostable chimeras have not increased in stability at the cost of their specific activities toward the soluble MUL substrate.

The total cellulase activity secreted from yeast is relevant in consolidated bioprocess applications, where recombinant strains of *S. cerevisiae* convert cellulosic biomass to fermentable simple sugars and ferment the simple sugars to biofuel in a single process step (Den Haan *et al.*, 2007). We measured the total solid cellulose hydrolysis activities of the five CBH I parents and a set of six stable chimeras with total secreted activities toward the soluble MUL substrate that are equal to or greater than that of *T. emersonii* CBH I. As shown in Fig. 6, all six of the CBH I chimeras also exhibit more total solid cellulose hydrolysis activity than any of the parents.

Discussion

The total number of chimeras that can be made by swapping sequence blocks is p^b, where p is the number of parents and b is the number of blocks into which each parent is divided. Including more parent enzymes in the construction of a SCHEMA recombination family generates many more potential unique chimeras and enables inclusion of more potentially beneficial mutations. Whereas 6561 chimeras can be made by recombination of three parents and eight blocks, adding two more parent sequences increases the family size to more than 390 000. The number of mutations explored by recombination depends on the parent sequence identities.

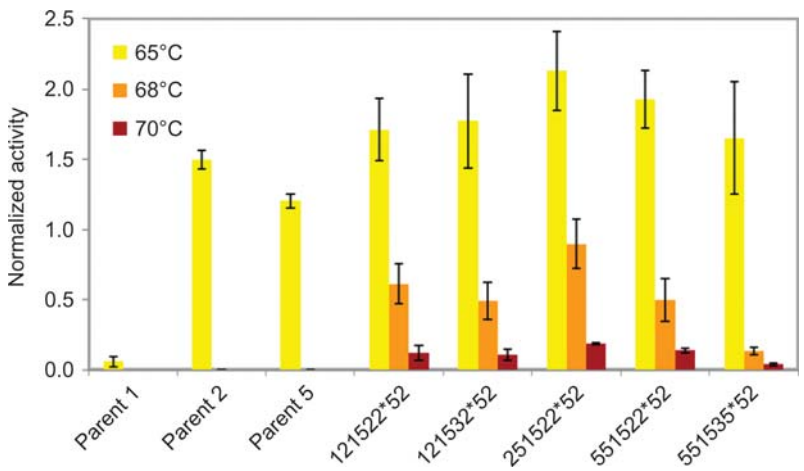

Fig. 5. Normalized high-temperature solid cellulose hydrolysis activity for affinity-isolated CBH I parents and chimeras. Values presented are activity at given temperature relative to activity at 50°C. Reactions were carried out for 16 h in 50 mM sodium acetate, pH 4.8, with 60 mg/ml solid cellulose and 14.6 μg/ml affinity-isolated CBH I. Error bars denote the standard deviations for three replicates. *5 denotes block 7 from parent 5 with stabilizing sub-block C insertion.

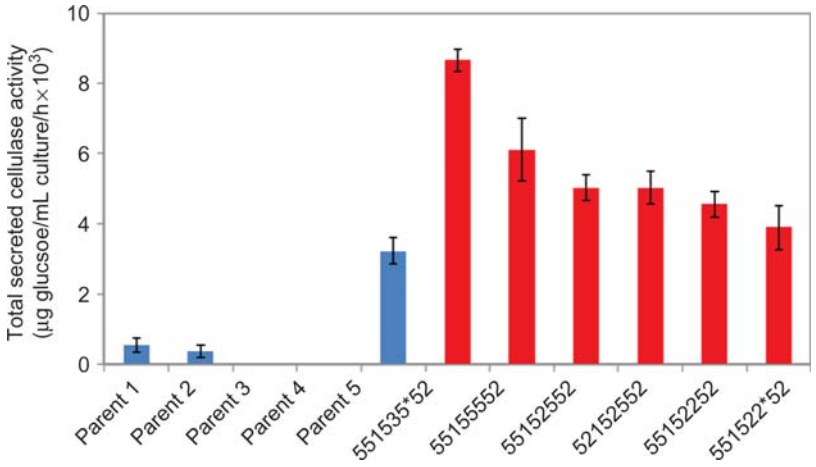

Fig. 6. Total yeast-secreted activity toward solid cellulose for CBH I parents and chimeras. Yeast culture supernatant was incubated with microcrystalline cellulose in 50 mM sodium acetate, pH 4.8, for 1 h at 4°C to bind CBH I. Cellulose was subsequently washed and hydrolysis allowed to proceed for 90 min at 37°C. Error bars represent the standard deviations for three replicates. *5 denotes block 7 from parent 5 with stabilizing sub-block C insertion.

For the CBH Is, relative to the *T. emersonii* background parent (P5), parent 1 (*C. thermophilum*) contains 151 mutations, parent 2 (*T. aurantiacus*) adds 43 unique mutations, parent 3 (*H. jecorina*) brings 100 more unique mutations and parent 4 (*A. thermophilum*) increases the mutation count by 52, bringing the total number of mutations that can be searched by recombination to 336.

The drawback to working with a larger chimera family is that more chimeras must be characterized in order to build a predictive stability model. It can be costly if a significant proportion of the sample chimeras do not express in functional form. This work demonstrates that the desirable sequences can be identified efficiently with a monomera screening approach, in which individual block substitutions are made in the background of a stable, well-expressed parent. Relative to a chimera sample set chosen to test interactions among blocks, i.e. the importance of the background sequence, this strategy reduces the number of non-productive sequences that are constructed.

Stability measurements made for the background parent and 28 secreted members of a 32-member CBH I monomera set allowed stability contributions to be estimated for 36 of the 40 blocks comprising the five-parent, eight-block CBH I chimera family. Assuming no non-linear stability effects among blocks and that B7P5 is the most stabilizing block at position 7 in all chimera backgrounds, these measurements allow prediction of the most stable of $5^8 = 390\,625$ CBH I chimera sequences. This represents an increase in screening efficiency relative to the prior CBH II recombination work (Heinzelman *et al.*, 2009a), but rests on the assumption that the blocks contribute additively to overall stability and does not test the linear model.

This work also demonstrates the robustness of SCHEMA recombination for creating active chimeras from parent enzymes that feature a large number of disulfide bonds. SCHEMA seeks to define block boundaries so that interactions among blocks are similar to those that occur in the parent enzymes. Block boundaries, however, are defined without regard for disulfide bonds. As such, the presence of 10 disulfide bonds, 5 of which link Cys residues lying in different blocks, poses a new test of SCHEMA's ability to generate chimera family designs that lead to a large fraction of active members. As shown by 28 of the 32 monomeras and 16 of the 16 predicted stable chimeras being secreted as active cellulases, SCHEMA recombination can generate a large fraction of active chimeras even when the protein is crosslinked by a large number of disulfide bonds. These results suggest that SCHEMA recombination conserves the appropriate position and orientation of Cys residues for disulfide formation.

Linear block stability contributions that allow quantitative prediction of chimera thermostability (Heinzelman *et al.*, 2009a,b) stand alongside the high sequence diversity and large fraction of active members as useful features of SCHEMA chimera families. Although we have shown that the monomera screening approach allows thermostable chimera sequences to be predicted with high accuracy, the monomera sample set and predicted thermostable chimera thermostability measurement set do not provide the degree of block representation needed to determine whether block stability contributions are indeed linear. This will be addressed by future construction and evaluation of CBH I chimeras containing a broader representation of blocks.

Block 7 is the largest block, with 116 residues that comprise 27% of the CBH I catalytic domain. The inability to make substitutions at this position markedly reduces the total number of mutations encompassed by the monomera sample set screen. In particular, 119 of the 336 total unique mutations in the 32 monomera sample set are contained within block 7. High E values do not necessarily predict the resistance of block 7 to recombination (e.g. the 55 555 525 chimera is not secreted, despite having an E value of 2). This observation motivated us to construct and test sub-blocks of block 7. That we identified a sub-block that increases the stability of not only the corresponding monomera but also all five of the stable chimeras into which it is substituted shows that subdividing a recombination block can generate further stability improvements.

Given the demonstrated utility of SCHEMA and the monomera block screening approach for creating new thermostable enzymes, it is instructive to compare and contrast this strategy with other methods for improving enzyme thermostability. Consensus mutagenesis (Lehmann *et al.*, 2002; Amin *et al.*, 2004) is possibly the most broadly used enzyme

thermostabilization strategy that does not employ HTS. Consensus mutagenesis is based on aligning a large, i.e. dozens or hundreds, number of related enzyme sequences and identifying residues that appear with high frequency at a given position as being potentially stabilizing. Changing the residue identity from a low frequency to a higher frequency amino acid at a given position is then predicted to improve the thermostability of the enzyme into which such a substitution is made.

Despite the successful use of consensus mutagenesis to predict single-residue substitutions that improve enzyme thermostability (Lehmann *et al.*, 2002; Amin *et al.*, 2004), the need for a large number of phylogenetically diverse sequences to ensure prediction accuracy is a considerable limitation (Jäckel *et al.*, 2010). Successful applications of consensus mutagenesis (Lehmann *et al.*, 2002; Amin *et al.*, 2004) have incorporated dozens, if not hundreds, of enzyme homolog sequences. While the CAZy database (www.cazy.org) contains more than 40 CBH I or CBH I-related gene sequences that could be used in applying consensus mutagenesis to CBH I stabilization, there are many enzyme classes for which such a large set of known sequences is not available. Furthermore, even when many sequences are available, the ability to make accurate predictions of stabilizing residues is limited by the fact that the enzymes have evolved from common ancestors (Jäckel *et al.*, 2010). Evolution from a small starting pool biases residue frequencies in the full homolog set toward amino acids appearing in the parental sequences, which confounds any stabilizing role that amino acid might have. Given that SCHEMA recombination requires only the sequences of the parent enzymes and a crystal structure for either parent enzyme or homolog, the monomera block screening approach we have described can be a useful alternative to consensus mutagenesis for improving stability.

This demonstration of enzyme stabilization by SCHEMA recombination has been made in the context of industrially relevant fungal CBH Is, which are the principal components of cellulase mixtures used in large-scale biomass conversion processes. These enzymes are notoriously difficult to express in a heterologous host (Godbole *et al.*, 1999) and few protein engineering efforts have led to improved enzymes, despite their industrial importance. The most thermostable CBH I described to date is a variant of *T. emersonii*, secreted from a recombinant *S. cerevisiae* host, that contains three additional, rationally designed disulfide bonds, G4C-A72C, N54C-P191C and T243C-A375C (Voutilainen *et al.*, 2010). The single G4C-A72C engineered disulfide *T. emersonii* catalytic domain used as a SCHEMA recombination parent was also described. The respective T_m values of the single- and triple-disulfide-bond variants are reported to be 80°C and 84°C, as measured by CD, and their half-lives at 70°C are reported to be 270 and 320 min in the absence of substrate (Voutilainen *et al.*, 2010). These numbers for the G4C-A72C mutant do not align with our observed T_{50} value of 62.9 ± 0.3°C for the *T. emersonii* CBH I parent in yeast secretion culture supernatant, however, and also imply thermostability much greater than what we observe in solid substrate hydrolysis assays, where the *T. emersonii* parent is inactive at temperatures >65°C.

In an attempt to resolve these differences, we performed CD measurements on Ni-NTA affinity isolated *T. emersonii* CBH I and also made $t_{1/2}$ measurements at 70°C using both affinity-isolated CBH I, which contains an N-terminal His_6 tag, and CBH I without the His_6 tag, from both SDCAA and YPD media yeast secretion cultures that were diluted into the assay buffer described in the previous report (50 mM sodium acetate, pH 5.0; Voutilainen *et al.*, 2010). Both our CD measurements (T_m ~65°C) and measured half-lives (<3 min at 70°C) are consistent with the T_{50} we report here, but not with the very high thermostability reported elsewhere (Voutilainen *et al.*, 2010). In further efforts to identify the root of this disagreement, we also made new constructs designed to more closely mimic the vector described in the literature. We made separate constructs in which the N-terminal Ala-Ser residues, a product of the N-terminal NheI cloning site in the expression vector, have been removed, in which the *H. jecorina* linker and CBM are not appended to the *T. emersonii* C-terminus, and in which the *T. emersonii* catalytic domain is preceded by the native signal peptide, as described (Voutilainen *et al.*, 2010). None of these modifications change the T_{50} value relative to our original *T. emersonii* secretion construct (data not shown).

We also explored whether glycosylation can account for this thermostability difference. Treating metal-affinity isolated CBH I with PNGase F to remove N-linked glycosylation had no impact on $t_{1/2}$ as measured in the described assay buffer (data not shown; Voutilainen *et al.*, 2010). This finding is consistent with the report that deglycosylation changes do not affect $t_{1/2}$ or T_m as determined by CD (Voutilainen *et al.*, 2010). Unlike the glycosylation-deficient Δ*KRE2* yeast strain we use as a secretion host, the described *S. cerevisiae* strain (Voutilainen *et al.*, 2010) is expected to hyperglycosylate N-linked sites. We therefore transformed our original and the three modified *T. emersonii* secretion constructs into this strain. Our finding that T_{50} values are independent of secretion host strain (data not shown) indicates that the stability difference cannot be attributed to differences in N-linked glycosylation. We did not perform experiments to determine whether O-linked glycosylation plays a role in the observed thermostability difference. The absence of O-linked glycosylation from the *T. emersonii* CBH I catalytic domain (Grassick *et al.*, 2004), however, coupled with the above observed difference in thermostability for the *T. emersonii* CBH I catalytic domains without the flexible linker and CBM indicates that O-linked glycosylation likely cannot explain the stability difference.

This five-parent SCHEMA recombination has generated a set of thermostable CBH I chimeras that are a key addition to the previously described thermostable CBH II chimeras (Heinzelman *et al.*, 2009a,b) in the assembly of an inventory of thermostable fungal cellulases from which application-specific mixtures can be formulated. Additionally, this work shows that the monomera screening strategy makes tractable the prediction of desirable chimera sequences within large families, thus increasing the utility of SCHEMA for exploring large swaths of enzyme sequence space. Furthermore, the observed improvements in chimera properties and the high fraction of active recombined enzymes shows that SCHEMA recombination can be applied to enzymes that contain extensive post-translational modifications. As such, these results are relevant not only to enzyme engineering in the context of industrial biomass conversion processes but also for engineering other proteins for which high sequence diversity is

desirable and/or whose properties are not easily improved by mutagenesis and HTS.

Funding

This work was supported by grants from the Army-Industry Institute for Collaborative Biotechnologies and the Caltech Innovation Institute.

References

Amin,N., Liu,A.D., Ramer,S., Aehle,W., Meijer,D., Metin,M., Wong,S., Gualfetti,P. and Schellenberger,V. (2004) *Protein Eng. Des. Sel.*, **17**, 787–793.

Den Haan,R., Rose,S.H., Lynd,L.R. and van Zyl,W.H. (2007) *Metab. Eng.*, **9**, 87–94.

Endelman,J.B., Silberg,J.J., Wang,Z.G. and Arnold,F.H. (2004) *Protein Eng. Des. Sel.*, **17**, 589–594.

Godbole,S., Decker,S.R., Nieves,R.A., Adney,W.S., Vinzant,T.B., Baker,J.O., Thomas,S.R. and Himmel,M.E. (1999) *Biotechnol. Prog.*, **15**, 828–833.

Grassick,A., Murray,P.G., Thompson,R., Collins,C.M., Byrnes,L., Birrane,G., Higgins,T.M. and Tuohy,M.G. (2004) *Eur. J. Biochem.*, **271**, 4495–4506.

Gusakov,A.V., Salanovich,T.N., Antonov,A.I., Ustinov,B.B., Okunev,O.N., Burlingame,R., Emalfarb,M., Baez,M. and Sinitsyn,A.P. (2007) *Biotechnol. Bioeng.*, **97**, 1028–1038.

Heinzelman,P., Snow,C.D., Smith,M.A., Yu,X., Kanaan,A., Boulware,K., Villalobos,A., Govindarajan,S., Minshull,J. and Arnold,F.H. (2009a) *J. Biol. Chem.*, **284**, 26229–26233.

Heinzelman,P., Snow,C.D., Wu,I., Nguyen,C., Villalobos,A., Govindarajan,S., Minshull,J. and Arnold,F.H. (2009b) *Proc. Natl Acad. Sci. USA*, **106**, 5610–5615.

Jäckel,C., Bloom,J.D., Kast,P., Arnold,F.H. and Hilvert,D. (2010) *J. Mol. Biol.*, **399**, 541–546.

Le Crom,S., Schackwitz,W., Pennacchio,L., *et al*. (2009) *Proc. Natl Acad. Sci. USA*, **106**, 16151–16156.

Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P. and Wyss,M. (2002) *Protein Eng.*, **15**, 403–411.

Li,Y., Drummond,D.A., Sawayama,A.M., Snow,C.D., Bloom,J.D. and Arnold,F.H. (2007) *Nat. Biotechnol.*, **25**, 1051–1056.

Meyer,M.M., Hochrein,L. and Arnold,F.H. (2006) *Protein Eng. Des. Sel.*, **19**, 563–570.

Viikari,L., Alapuranen,M., Puranen,T., Vehmaanpera,J. and Siika-aho,M. (2007) *Adv. Biocheml. Eng. Biotechnol.*, **108**, 121–145.

Voutilainen,S.P., Puranen,T., Siika-Aho,M., *et al*. (2008) *Biotechnol. Bioeng.*, **101**, 515–528.

Voutilainen,S.P., Boer,H., Alapuranen,M., Jänis,J., Vehmaanperä,J. and Koivula,A. (2009) *Appl. Microbiol. Biotechnol.*, **83**, 261–272.

Voutilainen,S.P., Murray,P.G., Tuohy,M.G. and Koivula,A. (2010) *Protein Eng. Des. Sel.*, **23**, 69–79.