Protein engineering: rational design and directed evolution

楊安綏 An-Suei Yang, Ph.D.
genomics research center
academia sinica
The fate of cellulose in our environment.

Bayer et al., Current Opinion in Biotechnology 2007, 18:237–245
Engineering potent cellulolytic microbes for the production of desired end products. Genes encoding for cellulases and/or designer cellulosome components (i.e. chimaeric scaffoldin and desired dockerin-containing hybrid enzymes) can be cloned into a desired bacterial or fungal host cell, and the secreted proteins can be overexpressed for the degradation of cellulosic biomass in an industrial reactor (in vitro assembly). Alternatively, the genes can be cloned into a suitable bacterial, fungal or yeast host, and the transformed cell with either de novo or improved cellulose-degrading capacity can be grown directly on cellulosic biomass to produce a desired and product.
Principles of designer cellulosome action. (a) Architecture of native versus designer cellulosomes. The cohesion–dockerin interaction is of uniform specificity in the native complex; the specificity is divergent in designer cellulosomes to facilitate controlled incorporation of enzyme components. (b) Enhanced synergism of binary designer cellulosomes, by combined targeting and enzyme proximity effects [36]. Targeting of enzymes to the substrate through the cellulose-binding module (CBM) results in enhanced synergistic action (red trace), compared to the free (Figure 2 Legend Continued) enzymes (blue trace). Integration of the enzymes into a single complex generates an additional enhancement of synergy (cyan trace). (c) Enhanced cellulase-hemicellulase synergy of a tertiary designer cellulosome on a crude cellulose substrate [37**]. Incorporation of an additional cellulase to a chimaeric complex containing two processive enzymes (C. cellulolyticum Cel48F and Cel9G) augments the activity of the complex on hatched straw. The different colored bars (pale blue, green, gray, red, yellow and violet) represent different C. cellulolyticum cellulases (from left to right, Cel5A, Cel8C, Cel9E, Cel9M, Cel9G and Cel48F). The adjacent bar in each case represents the activity of the three free enzymes alone (without the scaffoldin). By including a xylanase (C. thermocellum Xyn10Z) in the complex (dark blue bar), a dramatic increase in activity on hatched straw is observed.
Rational design

A cross-disciplinary area blending structural biology, computation, and bioinformatics
Computational redesign of endonuclease DNA binding and cleavage specificity

Figure S6. Structure of the re-designed I-MsoI – DNA interface, showing the regions that were re-designed to recognize the previously untolerated basepairs -6G:C and +6C:G (magenta). Re-designed residues K28->L and T83->R are shown in yellow, and additional residues that were allowed to move and mutate are shown in green. Water molecules are light blue, metal ions are gray. This image was generated from the crystal structure of the engineered I-MsoI-K28L-T83R protein bound to its non-native cognate DNA (rcsb: 2FLD).

Ashworth et al., Nature. 2006 Jun 1;441(7093):656-9
Energy calculation:

• van der Waals (Lennard-Jones) packing energy
• orientation-dependent hydrogen bonding potential
• Born-based electrostatic solvation energy
• PDB derived sidechain torsional potential
• Amino acid residue reference energy in unbound and unfolded state

Conformational search:

• Monte-Carlo search algorithm
• Torsional angle-based energy optimization
Van der Waals interactions

Lennard–Jones potential of the form:

\[ E(d) = \frac{C_n}{d^n} - \frac{C_6}{d^6} \quad (n > 6) \]

**Figure 4.2**
Representative profile of the energy of the van der Waals interaction as a function of the distance \( d \) between the centers of the two atoms. The individual attractive and repulsive components are indicated by the dashed lines, the net interaction by the solid line. The optimal interaction between the two atoms occurs where the energy is at a minimum. The sum of the van der Waals radii of the two atoms is given by the distance at which the energy increases sharply. The interaction energy was calculated using the Lennard–Jones 6,12 potential (Eq. 4.6) with \( C_{12} = 2.75 \times 10^6 \) Å\(^6\) kcal/mol and \( C_{6} = 1425 \) Å\(^6\) kcal/mol for the interaction between two carbon atoms (M. Levitt, *J. Mol. Biol.* 82:393–420, 1974).

**Table 4.1** Van der Waals Radii of Atoms Found in Proteins

<table>
<thead>
<tr>
<th>Atom</th>
<th>Observed range (Å)</th>
<th>Radius when singly bonded (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrogen</td>
<td>1.0–1.54</td>
<td>1.17</td>
</tr>
<tr>
<td>Oxygen</td>
<td>1.4–1.7</td>
<td>1.40</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>1.55–1.60</td>
<td>1.55</td>
</tr>
<tr>
<td>Carbon</td>
<td>1.70–1.78</td>
<td>1.75</td>
</tr>
<tr>
<td>Sulfur</td>
<td>1.75–1.80</td>
<td>1.80</td>
</tr>
</tbody>
</table>

Hydrogen bonds

$H...O = 1.8 - 2.0 \text{Å}$

**FIGURE 4.3**
Linearity of $\text{N} - \text{H} \cdots \text{O}$ hydrogen bonds observed in crystal structures of small molecules. The degree of linearity is measured by the angle $\phi$ (top), which would have a value of $180^\circ$ for a perfectly linear hydrogen bond. The histogram of observed values of intermolecular hydrogen bonds (left) is affected by a geometric factor in which the various ranges of the value of $\phi$ include different volumes of three-dimensional space. Correcting for this gives the histogram on the right, which illustrates the marked tendency of hydrogen bonds to be linear. (Adapted from R. Taylor and O. Kennard, *Acc. Chem. Res.* 17:320–326, 1984.)

**FIGURE 4.4**
The geometries of $\text{C} = \text{O} \cdots \text{H} \cdots \text{N}$ hydrogen bonds observed in crystal structures of small molecules. The definitions of the angles $\phi$ and $\theta$ are illustrated at the top, and the relative frequencies of their observed values in intermolecular hydrogen bonds (R. Taylor et al., *J. Amer. Chem. Soc.* 105:5761–5766, 1983) are given by the contours. The angle $\phi$ measures departures from linearity of the $\text{C} = \text{O}$ bond and the $\text{H}$ atom; the most frequently observed values are in the region of $50^\circ$–$60^\circ$. The angle $\theta$ measures the extent to which the $\text{H}$ atom lies out of the plane defined by the $\text{R}$, $\text{C}$, and $\text{O}$ atoms; the most commonly observed values are in the region of $0^\circ$–$7^\circ$. The lone-pair electrons of the oxygen atom are believed to project at angles of $\phi = 60^\circ$, $\theta = 0^\circ$. The spherical polar coordinate system used here gives a bias toward small values of $\theta$ that could be corrected by plotting $\sin \theta$. 
PDB derived sidechain torsional potential

\[ \Delta G = \log(P_1/P_0) \]
\[ \Delta G_{\text{GB}} = -\frac{1}{2} \sum_{ij} f_{\text{GB}}^i(r_{ij}, R_i, R_j) \left( 1 - \frac{e^{-k_r^i}}{\epsilon_w} \right) \]

\[ k_\text{[Å}^{-1}] \approx 0.316 \sqrt{[\text{salt}]} \]

\[ f_{\text{GB}} = [r_{ij}^2 + R_i R_j \exp(-r_{ij}^2/4R_i R_j)]^{1/2} \]

$r_{ij}$ = dist between i and j
$q_i$ = charge of i
$R_j$ = Born radius of j
$\epsilon_w$ = dielectric constant of water (80)

**Generalized Born model:**
**Example for a positively charged particle from vacuum to water**

\[ \Delta G_1 = \text{ASA} \times \text{const} \]

ASA = accessible surface area

\[ \Delta G_2 = -\frac{e^2}{2R} \]

In pure water
Amino acid residue reference energy in unbound and unfolded state

\[ \Delta \Delta G = \Delta G_{\text{unfold}} + \Delta G_{\text{fold}} \]

\[ \Delta G = 0 \]
Metropolis Monte Carlo algorithm:

1. Choose the initial position, calculate energy $E_1$

2. Pick the random displacement, calculate the new location and energy $E_2$

3. Decide whether to accept the move:
   - if $E_2 > E_1$, calculate draw a random number $r$ from 0 to 1.
   - if $r < \exp(-\frac{(E_2 - E_1)}{RT})$ or $E_2 < E_1$, accept the new position, update it and the energy, otherwise, stay at the same place, add the current values to the running averages.

4. go back to step 2
Computational redesign of endonuclease DNA binding and cleavage specificity

Figure S3. WebLogo\cite{1,2} representation of the predicted sequence specificity of the wildtype I-MsoI endonuclease, expressed in terms of relative Boltzmann probabilities as in equation (9) in Morozov, et al.\cite{3} with a temperature of 0.5 mol/kcal. The wildtype recognition site of the endonuclease is shown at the bottom.

Ashworth et al., Nature. 2006 Jun 1;441(7093):656-9
Computational redesign of endonuclease DNA binding and cleavage specificity

Figure 1 | Comparison of the predicted interactions in cognate and non-cognate binding complexes, illustrating the designed specificity switch. a, Wild-type I-MsoI, −6G·C (wild type). A water molecule present in the original structure is shown. b, Wild-type I-MsoI, −6G·C. c, I-MsoI-K28L/T83R, −6G·C. d, I-MsoI-K28L/T83R, −6G·C. In parts c and d, the van der Waals surfaces of Leu 28 and +6C are shown in grey. Figures were generated using the molecular graphics program PyMOL (Delano Scientific). WT, wild type; DES, designed; blue strands, protein backbone; beige spheres and sticks, DNA backbone; other spheres, constant nucleotides; dashed lines, hydrogen bonds.

Ashworth et al., Nature. 2006 Jun 1;441(7093):656-9
Computational redesign of endonuclease DNA binding and cleavage specificity

**Figure 2 | Switch in nuclease cleavage specificity.** Equimolar amounts of linearized plasmid DNAs containing wild-type (WT) or designed (DES) I-MsoI cleavage sites were digested by serial dilutions of wild-type or designed I-MsoI endonuclease, and analysed by gel electrophoresis. The switch in sequence specificity is defined as (wild type vs. DES/wild type vs. WT) × (designed vs. WT/design vs. DES), where quantities in parentheses indicate the lowest enzyme concentration at which significant cleavage of the site is observed. Here, the wild-type enzyme favours the WT site by \( >2^7 \)-fold, the designed enzyme favours the DES site by \( \sim 2^5 \)-fold, and hence the specificity switch is greater than \( 2^7 \times 2^5 \) (\( >4,000 \)-fold).
Computational redesign of endonuclease DNA binding and cleavage specificity

Figure 3 | Crystal structure of the designed enzyme-DNA complex. Left, $F_o-F_c$ electron-density map of the redesigned region calculated from a refinement model lacking the redesigned side chains and bases (cyan). The computational design model (grey) fits well into the unassigned density (blue mesh, ±2σ). Right, superposition of the design model (salmon) and the refined crystal structure (cyan) confirms the accuracy of the design. A new coordinated water molecule (red sphere) is also apparent.

Ashworth et al., Nature. 2006 Jun 1;441(7093):656-9
Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design

L:N52Y
\(\Delta \Delta G_p = -1.0, \quad \Delta \Delta G_m = -0.1\text{ kcal/mol}\)

L:S28Q
\(\Delta \Delta G_p = -0.7\text{ kcal/mol}\)

H:T50V
\(\Delta \Delta G_p = -1.6, \quad \Delta \Delta G_m = -0.3\text{ kcal/mol}\)

H:K64E
\(\Delta \Delta G_p = -0.2\text{ kcal/mol}\)

Figure 1. Visualization of the mutated positions (yellow and orange) on the antibody–antigen interface. The view looks down through the antigen onto the complementary determining region (CDR) loops of the antibody. Only residues with an atom within 5 Å of the opposite side of the interface and mutation positions are shown. The antigen residues are colored purple, the light chain is green, and the heavy chain is blue. Positions of beneficial mutations are shown in orange. Expressed mutants at each position are noted to the right of the position number. All structural figures have been made using PyMOL (DeLano Scientific).

Clark et al 2006
protein sci. 15, 949
Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design

Conformational energy calculation:
- Van der Waals interaction
- Explicit hydrogen bond
- Chemical bond torsional energy
- Finite-difference Poisson-Boltzmann model for electrostatic interaction and solvation energy

\[ \nabla \left[ \varepsilon(\mathbf{r}) \nabla \Psi(\mathbf{r}) \right] = -4\pi \rho^{f}(\mathbf{r}) - 4\pi \sum_{i} c_{i} z_{i} q_{i} \lambda(\mathbf{r}) e^{-\frac{z_{i} q_{i} \Psi(\mathbf{r})}{kT}} \]

Side chain conformational search:
- Dead-end elimination algorithm

Flexible backbone modeling:
- Molecular dynamics simulation

Clark et al 2006
protein sci. 15, 949
Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design

L:N52Y  
H:K64E  
H:T50V  
L:S28Q

Clark et al 2006  
protein sci. 15, 949

Figure 2. Changes in binding for the mutants. (A) Comparison of calculated with experimental $\Delta \Delta G$ of binding. All mutants with competition ELISA EC50s near wild type were reevaluated using the solution-based KinExA assay (inset). (B) Examples of the solution-phase (KinExA)

\[ \Delta \Delta G = -RT \ln \left( \frac{K_d \text{wt}}{K_d \text{mt}} \right) \]
Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design

Figure 3. Visualization of the quadruple mutant crystal structure (dark gray) near side chain repacking mutants. (A) Comparison to the predicted structure (light gray) in the vicinity of the S28Q mutation in the light chain. The predicted structure forms a similar stacking-like interaction between Tyr264 on antigen and the glutamine. The electron density (2Fo−Fo, σ = 1.1) indicates that the tyrosine has swung inward toward the bulk of the antigen. (B) Comparison of the wild-type (light gray) and quadruple mutant crystal (dark gray) structures in the vicinity of the N52Y mutation in the light chain.

Figure 5. Comparison of the wild-type (light gray) and quadruple mutant (dark gray) crystal structures in the vicinity of the H:T50V mutation. The wild-type threonine hydrogen bonds with the tryptophan. When substituted with a valine, the local environment rearranges to eliminate an unsatisfied hydrogen bond.
55 fXa-inhibitors Complex structure from Protein Data Bank – models for protein-substrate interaction
Pharmacophore models derived from known ligand-enzyme complexes

Yang Lab
Acyl-enzyme transition state complex model (substrate=GPQGR)
Acyl-enzyme transition state complex model (substrate=GIEGR)
Acyl-enzyme transition state complex model (substrate=GYRGR)
Acyl-enzyme transition state complex model (substrate=GPSGR)

Yang Lab
Acyl-enzyme transition state complex model (substrate=GLLQR)
Acyl-enzyme transition state complex model (substrate=GAQFR)
Acyl-enzyme transition state complex model (substrate=GPRAR)
Protein-substrate recognition can complicate enzyme engineering

GIEGRSGG (nature substrate sequence)

Yang Lab
Combinatorial biology:

Powerful tools for directed evolution
Selection of the preferred experimental approach for enzyme engineering based on the availability of experimental tools and prior knowledge of structure and function. Rational design, semi-rational design or whole-gene randomization each refer to multiple methodologies, as outlined in the text. The enzyme engineering approach that may have the greatest potential for success is in upper case letters, while alternative approaches are in lower case letters.
High throughput screening for protein expression and assay

Figure 1
Diagram of the robotic workcell (shown without biocontainment hood). Position #1A: Track 1; #1B: Track 2; #2A: StackLink active stacker (Track 1); #2B: StackLink active stacker (Track 2); #3: 4-axis Hudson Plate Crane EX; #4: BioRad VersArray colony picker/arrayer; #5: PCR thermal cycler with motorized heated lid; #6: UV/VIS plate reader using KC4 software; #7: ABgene 300 plate sealer using foil tape; #8: Brandel RS3000 plate sealer using porous tape; #9: liquid handler with centrifuge; #10: Hudson microl0 filler; #11: Hudson track-based sterile plate aspirator; #12: automated incubator; #13: passive stackers; #14: computer and monitor; #15: barcode reader; #S1-S6: StopLink plate positions.

Hughes et al., Proteome Science 2006, 4:10
High throughput screening for protein expression and assay:
high-throughput screening of mutants of cellulase F, an endoglucanase from the anaerobic fungus *Orpinomyces* PC-2.

Hughes et al., *Proteome Science* 2006, 4:10
DNA libraries
keystone for directed evolution
Error-prone polymerase chain reaction

**PCR**
- Taq
- dCTP, dTTP
- dGTP, dATP
- Mg$^{2+}$

**Error prone PCR**
- Taq
- dCTP, dTTP $\uparrow$
- dGTP, dATP $\downarrow$
- Mg$^{2+}$ $\uparrow$
- Mn$^{2+}$

Additional information:
- 7 mM MgCl$_2$
- 50 mM KCl
- 20 mM Tris (pH 8.4)
- 0.5 mM MnCl$_2$
- 1 mM dTTP
- 1 mM dCTP
- 0.2 mM dGTP
- 0.2 mM dATP
- 100 ng Template
- 2 µM Primer s
- 2 µM Primer as
- 1 µl Taq
- add H$_2$O to 100 µl

Temperature and duration:
- 95°C 3 min
- 94°C 45 s
- 58°C 45 s
- 72°C 3 min
- 72°C 10 min
- 4°C
Single gene DNA shuffling

Fig. 1. (A) A pool of homologous genes with different point mutations is fragmented with DNase I. (B) For simplicity, all mutations shown are considered beneficial and additive. (C) Reassembly of the random fragments into full-length genes results in frequent template switching and recombination. A recombinant gene containing the four crossovers (thick lines) can be selected from the library of recombinants based on its improved function. (D) Selected pool of improved recombinants provides the starting point for another round of mutation and recombination. The recombination process alone causes a low level of point mutations but, if desired, additional mutations could be introduced by error-prone PCR or UV mutagenesis of the pool of genes.

Stemmer, PNAS 1994, 91, 10747
Single gene DNA shuffling

**Fig. 2.** Reassembly of a 1.0-kb gene from 10- to 50-bp random fragments. (a) A 1-kb DNA fragment encoding lacZα was amplified by PCR. After digestion of the gene with 0.15 unit of DNase I for 15 min at 20°C (b), fragments of 10–50 bases were purified from an agarose gel (c). (d) Purified fragments were reassembled into a full-length gene at a high fragment concentration (30 ng/μl) in the absence of primers. The average size of the PCR products increases gradually via the priming of one product on another. The heterogeneous appearance of the product is largely due to the partially single-stranded nature of the product. (e) After addition of primers and additional cycles of PCR, a single PCR product of the correct size is typically obtained. Cloning of this product into a plasmid yielded ~84% light- to dark-blue colonies, reflecting mutations that occurred during the reassembly process.
Single gene DNA shuffling: β-lactamase model system

FIG. 2. Three successive rounds of DNA shuffling were done and the cells were grown on increasing cefotaxime levels. The MIC of cefotaxime (Sigma) for E. coli XL1-blue (Stratagene, San Diego) carrying wild-type p182Sfi is 0.02 μg ml⁻¹. A mutant with a 16,000-fold increased resistance to cefotaxime was obtained (MIC = 320 μg ml⁻¹). This mutant was backcrossed twice, by shuffling with a 40-fold excess of wild-type DNA. The backcrossed mutant was 32,000-fold more resistant than the wild type (MIC = 640 μg ml⁻¹). After selection, the plasmid of selected clones was transferred back into wild-type XL-1 blue cells to ensure that none of the measured drug resistance was due to chromosomal mutations. DNA sequencing showed that both mutants had 9 single-base-pair mutations.
Gene family shuffling

Figure 2: a, Comparison of single sequence shuffling versus sequence family shuffling. b, Sequence of chimeraic mutant A obtained by family shuffling. The segments derived from Enterobacter are shown in blue, those from Klebsiella are shown in yellow, and those from Citrobacter are shown in green. The grey segments are where the crossovers have taken place. Because of DNA homology in the grey segments, the exact location of the crossover cannot be determined more exactly. The amino-acid point mutations are shown with underlined red letters. The numbers at the beginning and end of each segment are the numbers from the Genbank protein files of the wild-type enzymes and differ from those used for the Enterobacter cloacae enzyme.
Gene family shuffling

Figure 3 Computer model of evolved mutant A obtained by a single cycle of family shuffling. The 142 amino-acid mutations were introduced into the Enterobacter cloacae sequence, whose structure is known [4], followed by energy minimization. The predicted structure of the α-chain backbone of the chimaeric enzyme is within an r.m.s. deviation of 0.768 Å from the known native structure. The segments derived from Enterobacter are shown in blue, those from Klebsiella are shown in yellow, and those from Citrobacter are shown in green. The 33 amino-acid point mutations are shown in red.

Figure 4 Searching sequence space by family shuffling versus by single sequence shuffling. Single sequence shuffling yields clones with a few point mutations and the library members are typically 97–99% identical. Family shuffling causes sequence block exchange which yields chimaeras that have greater sequence divergence. At equal library size, the increased sequence diversity of the chimaeric library results in sparse sampling of a much greater area of sequence space, allowing more promising areas to be found and subsequently explored at increased sampling density.

Krameri et al, Nature 1998 (391), 288
Staggered extension process (StPE) in vitro recombination

Figure 1. StPE recombination. Only one primer and single strands from two parent genes (templates) are shown. (A) Denatured template genes are primed with one defined primer. (B) Short fragments are produced by brief polymerase-catalyzed primer extension. (C) Through another cycle of StEP, fragments randomly prime the templates (template switching) and extend further. (D) This process is repeated until full-length genes are produced. (E) Full-length genes are purified and (optionally) amplified in a PCR reaction with external primers.

Figure 2. Agarose electrophoresis gel showing the progress of recombining two thermostable subtilisin E genes RC1 and RC2 by StEP. Lane 1: AmpliSize DNA Size standards (Bio-Rad, Hercules, CA), from top to bottom: 2000, 1500, 1000, 700, 500, 400, 300, 200, 100, and 50 bp; lane 2: after 20 cycles; lane 3: after 40 cycles; lane 4: after 60 cycles; lane 5: after 80 cycles; lane 6: 1 kb ladder from Boehringer Mannheim.

Figure 3. Results of screening 368 variants from the recombined library for activity after incubation at 65°C for 10 min (initial activity, A) and 40 min (residual activity, A'). The ratio A/A' (thermostability index) was used to estimate thermostability. Data from active variants are sorted and plotted in descending order.

In vitro heteroduplex formation and in vivo repair

Figure 3. (a) Heteroduplex recombination using heteroduplexes prepared directly from parent plasmids. Circular parental plasmids are linearized on either side of the target gene using restriction endonucleases with unique cutting sites. Linear plasmids are mixed, denatured and annealed. Possible products of this reaction are parent homoduplexes (linear) and recombinant heteroduplexes (circular). Only circular plasmids efficiently transform bacterial cells. (b) Heteroduplex recombination using heteroduplexes formed by insert hybridization and ligation. Target genes are amplified in a PCR reaction, mixed and annealed together. After digestion with appropriate restriction endonucleases, the annealing products are ligated into a vector. Asymmetric synthesis should be used to suppress parent homoduplexes.

Volkov et al., Nucleic Acids Res. 1999 Sep 15;27(18):e18
In vitro heteroduplex formation and in vivo repair

Figure 1. To the extent that regions of non-identity are repaired independently, in vivo repair of a heteroduplex gives rise to a library of parent and recombined (bold) sequences.
Combinatorial libraries enhanced by recombination in yeast

Figure 1. Principle of the library construction. Plasmidic DNA was subjected to DNase I digestion (see Materials and Methods) and fragments were separated on a 1% agarose gel. (A) Lane 1, DNA ladder; lanes 2–4 correspond to DNase I treated p1A1/V60 and p1A2/V60 respectively. Lanes 2 and 5 correspond to fragmentation with 0.0112 U, lanes 3 and 6 to 0.0056 U and lanes 4 and 7 to 0.0028 U of DNase I per µg of DNA. (B) Reassembly reaction. Lane 1, DNA ladder; lanes 2–4 correspond to reassembly reactions between fragmented p1A1/V60 and p1A2/V60 mixing the reactions from lanes 2 and 5, 3 and 6, 4 and 7 respectively. (C) Amplification reaction. Lane 1, DNA ladder; lanes 2–4 correspond to the amplification with full-length plasmid pYeDP60, p1A1/V60 and p1A2/V60, respectively; lanes 5–7 correspond to the amplification with previously reassembled DNA as a matrix (lanes B2, B3 and B4 respectively). The band presented in (C) lane 6 was purified and used as such to cotransform Scerevisiae with linearized pYeDP60.

Combinatorial libraries enhanced by recombination in yeast

Naphthalene + cytochromes P450 -> phenolic metabolite -> diazo-BLUE-B -> red

Random chimeragenesis on transient templates

Figure 1. Random chimeragenesis on transient templates (RACHTT). A uracil-containing Rhodococcus erythropolis IGTS8 dstrC gene was made single-stranded to serve as a transient scaffold template (black bar). The molecule mediates the ordering of Nocardia asteroleuca A3H1 top-strand fragments (white bars) and serves as a template for the filling of gaps. Unequilibrated 5' and 3' termini or "flaps" are expected on some annealed fragments because of overlaps with adjacent fragments, or because of mismatches with respect to the scaffold, and are trimmed using nuclease. Filling gaps between hybridized fragments allows inclusion of template sequences as mosaic insertions into the parent chimeric strand. After flap trimming, gap filling, and ligation, the heteroduplexed parental bottom-strand template is rendered nonamplifiable by uracil-DNA-glycosylase treatment and is replaced by homoduplex chimeric bottom strand during PCR.

More crossovers

Figure 2. RACHTT generates highly mosaic shuffled proteins. In the top panel are bars representing the entire protein sequence for the two parental DBTMOs (solid white and solid black). Vertical lines in the parental sequences indicate the position of each of the 38 residues that differ between the parents, including a one-residue deletion (−). Of 22 randomly chosen unselected clones, 0 are shown. The improved clones were selected or screened as described in the text. The F7 clone combined the best rate and affinity phenotype. Spontaneous mutations are indicated by caret (>) above the bars. The locations of recombinational crossovers in regions of 0 to 5 amino acid residue identities are also identified above the bars (○ indicates recombination between adjacent residues). While bars connect A3H1 donated residues that are not interrupted by unique IGTS8 sequences, and will generally correspond to contiguous blocks of sequence donated by one or more A3H1 fragments. The locations of crossovers are parsimoniously depicted at the midpoint of regions of identity between parental differences. The minimum average number of crossovers per protein by this analysis is 0.3 for the unselected clones, but analysis of phenotypically silent differences at the DNA level reveals that 14.0 is the true minimum (see text).

Random chimeragenesis on transient templates

Figure 6. Evolution of substrate range. (A) A high level of phenotypic diversity is generated from the chimeric library. The *Rhodococcus*-borne chimeric *dszC* library was plated to rich medium, grown three days, subjected to indole vapors 3 h, and incubated five additional days. The amount of color development in plates containing *dszC* from strain IGTS8 (top left) or from A3H1 (top right) was reliably uniform, with the former producing little indigo and the latter somewhat more. Note the wide variability among the unselected library transformants (bottom). This “filling in” of phenotype space between and beyond both parents is what is expected from a diverse chimeric library. (B) A strain determined to be over 20-fold improved for indole oxidation in spectrophotometric assays was treated as above, but subjected to indole vapors and grown for only two days (lower plate) along with otherwise isogenic strains containing the parental A3H1 and IGTS8 dszC genes (upper left and right plates, respectively).

Compartmentalized self-replication

Water in oil emulsion: oil phase - 4.5% Span 80, 0.4% Tween 80, 0.05% Triton X-100 Mineral oil.

Ghadessy et al., Proc Natl Acad Sci U S A. 2001 Apr 10;98(8):4552-7
Problem with DNA shuffling:

DNA shuffling relies on long overlap in homologous sequence fragment for full length progeny gene extension - blocks of parent gene segments tend to be conserved in progeny genes.

Improvement?
Sequence homology-independent protein recombination

Figure 1. SHIPREC procedure. A gene dimer that consists of (from 5' to 3') the gene of protein 1 (1A2), a linker sequence containing useful restriction sites, and the gene of protein 2 (BM3) is constructed. (1) This dimer is fragmented (e.g., by digestion with DNase I in the presence of MnCl₂) and treated to produce blunt ends (e.g., using S1 nuclease or T4 polymerase). (2) Fragments of the length of a single gene (plus the length of the linker sequence) are separated from the pool (e.g., by agarose gel electrophoresis). (3) Single-gene length fragments are circularized by intramolecular blunt-end ligation. (4) Linearization of the circular DNA pieces by restriction digestion in the linker region creates a library of chimeric genes with members having an N-terminal part from protein 2 and a C-terminal part from protein 1, the crossovers being distributed over the entire length of the gene. (5) The chimeric genes are cloned into an expression vector directly or after amplification by PCR using one terminal primer from each of the two parents. The numbering shown in the genes represents percentage of gene length.

Incremental truncation for the creation of hybrid enzymes: for crossover between sequences of low homology

Sequence-independent site-directed chimeragenesis

Figure 1. A schematic of sequence-independent site-directed chimeragenesis (SISDC), illustrating recombination between two parents at three sites (four elements). (a) The nucleotide sequences of the parent genes (A and B) are aligned, and consensus sequences at each targeted site (I, II and III) are determined. (b) Marker tags (I–III) are inserted into targeted sites (I–III). The marker tags contain consensus 5 bp sequence (S), a type IIIb endonuclease (Bae I) recognition sequence (blue), two variable regions (X and Y; pink for site I, cyan for site II and green for site III), a constant region (yellow) and a Sma I recognition site (underline). (c) Inserted tag regions are removed by treatment with endonuclease and specific sticky ends are produced. (d) Mixed fragments are ligated on the fragments from different parents that ligate with each other in the right order. After this step, digestion by Sma I can eliminate any untreated tag remaining in the final, chimeric library.

**Figure 1.** Oligonucleotide design strategy for synthetic shuffling of subtilisins. Diversity from 15 subtilisin parents was encoded in a total of 30 synthetic oligonucleotides. Relative positions of 16 backbone oligos (F1–F8 and R1–R8) and 14 spiking oligos (S1–S14) are shown; arrows indicate the 5’ to 3’ direction. Positions where oligonucleotides differ from one another are indicated by X above the oligonucleotide.
Figure 2. Comparative phylogenies of synthetic shuffled and fragmentation-shuffled library members. Active variants were randomly picked and sequenced. The amino acid sequences of the diversified region were distributed on a phylogenetic tree together with the sequences of the parental genes using MEGA 2.1 Tree Explorer by the neighbor-joining method. The depth of the branches indicates the divergence between the sequences. (A) Fragmentation-based shuffled library. (B) Synthetic shuffled library. Also included are four variants shown in Fig. 3.

Ness et al., 2002Nat. Biotechnol. 20:1251-1255
Library Construction-Oligonucleotide directed mutagenesis

ssDNA
Template phagemid

T4 Ligase
T7 Polymerase

CCC-dsDNA
phagemid

Mutant

E.coli

Phage library

★ : Randomized primer

Yang Lab
Primer for constructing template

Av6

..TAATAAGAATTC..
Stop: EcoR1 site..

Yang Lab
**VL-CDR2, VL-CDR3, VH-CDR2, VH-CDR3 library construction**

<table>
<thead>
<tr>
<th>L2</th>
<th>L3</th>
<th>H2</th>
<th>H3</th>
<th>L2</th>
<th>L3</th>
<th>H2</th>
<th>H3</th>
<th>V1</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3397</td>
<td>133</td>
<td>251</td>
<td>144</td>
<td>1329</td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3530</td>
<td>251</td>
<td>144</td>
<td>1329</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3397</td>
<td>384</td>
<td>144</td>
<td>1329</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3397</td>
<td>547</td>
<td>395</td>
<td>1329</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3397</td>
<td>547</td>
<td>251</td>
<td>1473</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3397</td>
<td>547</td>
<td>1724</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3397</td>
<td>384</td>
<td>1473</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3397</td>
<td>528</td>
<td>1329</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3530</td>
<td>251</td>
<td>1473</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3530</td>
<td>395</td>
<td>1329</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3781</td>
<td>144</td>
<td>1329</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3397</td>
<td>1857</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3530</td>
<td>1724</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3781</td>
<td>1473</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>v</td>
<td>v</td>
<td>v</td>
<td>3925</td>
<td>1329</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5254</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Size map after EcoRI contained primer insertion by QuickChange® II Site-Directed Mutagenesis Kit

Yang Lab

insertion check by restriction enzyme cutting

↑ : sequencing to check all 4 primer are inserted
**VL-CDR2, VL-CDR3, VH-CDR2, VH-CDR3 library construction**

---

**Kunkle protocol for library construction**

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primer kinase phosphorylation</td>
<td></td>
</tr>
<tr>
<td>6 $\mu$L 100 $\mu$M library primer</td>
<td></td>
</tr>
<tr>
<td>10 $\mu$L 10X T4 phosphorylate kinase buffer</td>
<td></td>
</tr>
<tr>
<td>10 $\mu$L 10 mM ATP</td>
<td></td>
</tr>
<tr>
<td>5 $\mu$L 100 mM DTT</td>
<td></td>
</tr>
<tr>
<td>5 $\mu$L T4 kinase</td>
<td></td>
</tr>
<tr>
<td>ddH$_2$O supplement to total 100 $\mu$L</td>
<td></td>
</tr>
<tr>
<td>37°C, 2 h</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>primer annealing</td>
<td></td>
</tr>
<tr>
<td>4 $\mu$g template ssDNA</td>
<td></td>
</tr>
<tr>
<td>16 $\mu$L of each phosphorylated primer</td>
<td></td>
</tr>
<tr>
<td>5 $\mu$L 10XTM buffer</td>
<td></td>
</tr>
<tr>
<td>ddH$_2$O supplement to total 50 $\mu$L</td>
<td></td>
</tr>
<tr>
<td>90°C, 3 min</td>
<td></td>
</tr>
<tr>
<td>90°C-0.5°C/30sec X80</td>
<td></td>
</tr>
<tr>
<td>50°C-0.5°C/30sec X60</td>
<td></td>
</tr>
<tr>
<td>20°C, $\infty$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>extension</td>
<td></td>
</tr>
<tr>
<td>10 $\mu$L 10 mM dNTP</td>
<td></td>
</tr>
<tr>
<td>6 $\mu$L 10 mM ATP</td>
<td></td>
</tr>
<tr>
<td>4 $\mu$L 100 mM DTT</td>
<td></td>
</tr>
<tr>
<td>1.5 $\mu$L T7 DNA polymerase</td>
<td></td>
</tr>
<tr>
<td>0.4 $\mu$L T4 DNA ligase</td>
<td></td>
</tr>
<tr>
<td>20°C, O/N</td>
<td></td>
</tr>
</tbody>
</table>

---

**Yang Lab**

---

![Image of gel electrophoresis](image.png)

- **M**: 1Kb marker
- **1**: template ssDNA
- **2**: V4 library dsDNA
VL-CDR2, VL-CDR3, VH-CDR2, VH-CDR3 library construction

Insertion ratio check by restriction enzyme cutting

Before helper phage rescue:
1 / 5 insert (20%)

After helper phage rescue:
7 / 15 insert (46%)

Sequencing result:
all 8 inserts are V4 inserted

Complexity from 20ug of ssDNA=\~10^8

✓: all sites are successfully inserted

Yang Lab
VL-CDR2, VL-CDR3, VH-CDR2, VH-CDR3 library construction

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>N</td>
<td>K</td>
<td>N</td>
<td>N</td>
<td>K</td>
<td>N</td>
<td>N</td>
<td>K</td>
<td>N</td>
<td>N</td>
<td>K</td>
<td>N</td>
<td>N</td>
<td>K</td>
</tr>
<tr>
<td>L2</td>
<td>A%</td>
<td>22</td>
<td>17</td>
<td>0</td>
<td>32</td>
<td>14</td>
<td>0</td>
<td>9</td>
<td>35</td>
<td>0</td>
<td>23</td>
<td>30</td>
<td>9</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>T%</td>
<td>43</td>
<td>30</td>
<td>48</td>
<td>23</td>
<td>27</td>
<td>65</td>
<td>30</td>
<td>22</td>
<td>52</td>
<td>23</td>
<td>22</td>
<td>48</td>
<td>26</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>C%</td>
<td>17</td>
<td>39</td>
<td>0</td>
<td>36</td>
<td>41</td>
<td>4*</td>
<td>35</td>
<td>26</td>
<td>4*</td>
<td>41</td>
<td>39</td>
<td>9*</td>
<td>30</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>G%</td>
<td>17</td>
<td>13</td>
<td>52</td>
<td>9</td>
<td>18</td>
<td>30</td>
<td>26</td>
<td>17</td>
<td>43</td>
<td>14</td>
<td>9</td>
<td>35</td>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td>total%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>L3</td>
<td>A%</td>
<td>17</td>
<td>26</td>
<td>0</td>
<td>30</td>
<td>26</td>
<td>0</td>
<td>35</td>
<td>27</td>
<td>0</td>
<td>30</td>
<td>26</td>
<td>0</td>
<td>26</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>T%</td>
<td>35</td>
<td>13</td>
<td>39</td>
<td>26</td>
<td>35</td>
<td>65</td>
<td>22</td>
<td>41</td>
<td>65</td>
<td>22</td>
<td>30</td>
<td>70</td>
<td>26</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>C%</td>
<td>30</td>
<td>39</td>
<td>0</td>
<td>35</td>
<td>30</td>
<td>0</td>
<td>13</td>
<td>23</td>
<td>0</td>
<td>30</td>
<td>22</td>
<td>0</td>
<td>22</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>G%</td>
<td>17</td>
<td>22</td>
<td>61</td>
<td>9</td>
<td>9</td>
<td>35</td>
<td>30</td>
<td>9</td>
<td>35</td>
<td>17</td>
<td>22</td>
<td>30</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>total%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>H2</td>
<td>A%</td>
<td>17</td>
<td>13</td>
<td>4*</td>
<td>22</td>
<td>4</td>
<td>0</td>
<td>17</td>
<td>26</td>
<td>0</td>
<td>22</td>
<td>39</td>
<td>0</td>
<td>26</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>T%</td>
<td>39</td>
<td>30</td>
<td>65</td>
<td>30</td>
<td>57</td>
<td>52</td>
<td>43</td>
<td>35</td>
<td>57</td>
<td>26</td>
<td>22</td>
<td>70</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>C%</td>
<td>39</td>
<td>35</td>
<td>0</td>
<td>30</td>
<td>22</td>
<td>0</td>
<td>26</td>
<td>9</td>
<td>0</td>
<td>26</td>
<td>22</td>
<td>4*</td>
<td>30</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>G%</td>
<td>4</td>
<td>22</td>
<td>30</td>
<td>17</td>
<td>17</td>
<td>48</td>
<td>13</td>
<td>30</td>
<td>43</td>
<td>26</td>
<td>17</td>
<td>26</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>total%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>H3</td>
<td>A%</td>
<td>13</td>
<td>26</td>
<td>0</td>
<td>30</td>
<td>39</td>
<td>0</td>
<td>22</td>
<td>26</td>
<td>0</td>
<td>17</td>
<td>26</td>
<td>0</td>
<td>39</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>T%</td>
<td>43</td>
<td>35</td>
<td>65</td>
<td>13</td>
<td>17</td>
<td>43</td>
<td>43</td>
<td>26</td>
<td>57</td>
<td>26</td>
<td>22</td>
<td>52</td>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>C%</td>
<td>22</td>
<td>26</td>
<td>0</td>
<td>43</td>
<td>26</td>
<td>0</td>
<td>30</td>
<td>39</td>
<td>0</td>
<td>35</td>
<td>30</td>
<td>4*</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>G%</td>
<td>22</td>
<td>13</td>
<td>35</td>
<td>13</td>
<td>17</td>
<td>57</td>
<td>4</td>
<td>9</td>
<td>43</td>
<td>22</td>
<td>22</td>
<td>43</td>
<td>26</td>
<td>17</td>
</tr>
<tr>
<td>total%</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

*: some value are influenced by sequencing quality
Yang Lab
Future readings