Comprehensive Engineering of Biological Systems

Mark Welch, PhD
Director, Research
DNA2.0, Inc.

PepTalk
January 19, 2015
DNA2.0 – Picks and Shovels for Bioengineering

• Engineering Toolbox
  – Synthetic genes, libraries, vectors, markers, software

• Bioengineering Solutions
  – Genes, proteins, vectors, pathways, genomes
Whole-System Optimization

- Amino acid sequence
- Codon choice
- Promoters, Enhancers, Rep ori, RBS strength
- Chaperonins, Proteases, Redox, Sec pathways, etc.
- Temperature, Carbon source, Dissolved O₂, pH

Protein + Gene + Vector + Strain

DNA Sequence Space + Environmental Space

Product Output

Fermentation
Bioengineering Variable Space

System

Protein
- Substitutions
- Domain exchange
- Fusions
- Localization signals
- Homologs/Analogs
  - Specific activity
  - Condition dep.
  - Specificity
  - Stability
  - Solubility
  - Aggregation

Gene
- Codon usage (Global, Local)
- RNA structure
- Regulatory elements
  - Expression level
  - Expression control
  - RNA stability

Vector
- Promoters
- Enhancers
- Terminators
- mRNA processing
- ORI
- RBS, IRES
- Nuclear export
- Integration elements
- Size
  - Expression level
  - Expression control
  - Protein solubility
  - Viral packaging
  - Integration
  - Stability

Host
- Parent strain
- Knock-out, -down
- Gene augmentation
- Pathways
  - Robustness
  - Biomass
  - Condition tolerance
  - Pathway flux
  - Protein folding
  - Protein stability

Output
Bioengineering GPS

- **Expertise**
- **Constraints**
  - (Time, Resource, Assay)
- **Design tools**
  - (Software)
- **Project Goals**
- **Select Variables**
- **Design Experiment**
- **Construct**
- **Learn**
- **Test**
- **Products**

- **Gene synthesis**
- **Molecular toolbox**
- **Machine-learning technologies**

**GPS**
Protein Engineering
Variables = Amino Acid Substitutions

Challenge:
• Protein variation space is vast
• 16 positions = \(20^{16} \approx 10^{21}\)
  \(\approx\) Total grains of sand on the planet
  \(\approx\) 100kg DNA to sample
• Most is junk
• Which positions/substitutions?

Solution:
➢ Systematic sampling of substitutions
➢ Learning based search

→ ProteinGPS
ProteinGPS
Sampling Substitutions

Variable Selection

Test Proteins
Experimental Design
~60 subs per 96 variants

Parent Protein

Test Proteins "Infologs"
Assay

DNA 2.0
ProteinGPS
Machine Learning Identifies Substitution Impacts
Industrial Partner X, Protein X

$R^2 = 0.95$
$R^2(CV) = 0.78$
p < 0.001

Partial Least Squares regression model for soluble expression
Multiple Functional Criteria
Industrial Partner X, Protein X

Model Weight

Substitutions

- Activity
- Thermostability
- Soluble Expression

DNA 2.0
Multi-parameter Protein Engineering

Goals exceeded in <400 test variants
Efficient Improvement

employed in industrial bioprocess

Goal reached in ~300 test variants
Gene ORF Engineering
Variables = Synonymous Codon Choices

Challenge:
• Coding space is astronomical
• \( \sim 10^{100} \) ways to code a 30kDa protein
  > Total particles in universe
• Key variables not accessible by randomization

Solution:
• Control over sequence variables
• Learning based search

\( \rightarrow \) GeneGPS

Dr. MOT design software
• Flexible constraints and weights
• Gene design in the cloud
Controlling Gene Variables

Sequence Diversity

Randomized Library

Codon Usage Diversity

GeneGPS

Infolog

Set

DNA 2.0
Custom Gene Design Algorithms

>50 host organisms interrogated to date

Baculovirus/Sf9

- Predicted Expression
- Measured Expression
- $R^2 = 0.68$
- $R^2 (CV) = 0.52$
- $p = 0.04$

S. cerevisiae

- $R^2 = 0.86$
- $R^2 (CV) = 0.71$
- Initial Genes
- Hybrids

P. pastoris

- $R^2 = 0.81$
- $R^2 (CV) = 0.57$

Clostridium sp.

- $R^2 = 0.71$
- $R^2 (CV) = 0.64$

Corn (Ind. partner)

- $R^2 = 0.63$
- $R^2 (CV) = 0.53$

Mammalian Cells (Ind. & DNA2.0)

- $R^2 = 0.74$
- $R^2 (CV) = 0.65$
- $p < 0.001$

>300 test genes analyzed in HEK293 and CHO-S

Other partners:

- DNA 2.0
- Pfenex
- Perelman School of Medicine
- National Institutes of Health
Vector Engineering
Variables = Functional Elements

Other: Secretion leader panels, transposases, ProteinPaintbox

- Sample element diversity
- Learning based search

→ VectorGPS
VectorGPS
HEK and CHO Show Distinct Preferences

- Systematically sampled mammalian vector elements
  - Promoters, enhancers, introns, polyA, replication origins
- Transient expression of Dasher GFP tested in CHO-K1 and HEK293
VectorGPS

- Current data includes a wide range of 100s of elements
- HEK293, CHO-S, CHO-K1 (Transient and stable)
Strain Engineering
Variables = Host Genes

Solution:
- Reliable genome editing
- Learning based search

→ StrainGPS

Endless possibilities
- Attenuation
- Deletion
- Augmentation
- Gene replacement

Toolbox
- Nickase Ninja - CRISPR vectors
- Guide RNA design tool
- Transposases
- Chaperonin panels
- More
gRNA Design for CRISPR

Matches for USP53 Chromosome: 4

Select one or more gRNAs:

<table>
<thead>
<tr>
<th>Position</th>
<th>gRNA</th>
<th>Distance between gRNA</th>
<th>Score</th>
<th>Overlapping Genes</th>
<th>Splice variants targeted</th>
</tr>
</thead>
<tbody>
<tr>
<td>120,181,066</td>
<td>AGCTCTUAATTTTCTGATTCAC...ACCCTATCCACGATGGATAGGAAAA</td>
<td>6</td>
<td>100</td>
<td></td>
<td>7/7</td>
</tr>
<tr>
<td>120,181,024</td>
<td>TCTGCTTCTCTTCTGCATACG...TCTGTTACTACTTTCTAACACACATTT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120,181,735</td>
<td>TCTGCTTCTCTTCTGCATACG...TCTGTTACTACTTTCTAACACACATTT</td>
<td>4</td>
<td>81.47</td>
<td></td>
<td>7/7</td>
</tr>
<tr>
<td>120,181,697</td>
<td>ACGCTAACGAAATTACACGATCTCTCTAATGTTACCCAGATGGATAGGAAAA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>120,180,992</td>
<td>AGCTCTUAATTTTCTGATTCAC...ACCCTATCCACGATGGATAGGAAAA</td>
<td>7</td>
<td>80.62</td>
<td></td>
<td>7/7</td>
</tr>
<tr>
<td>120,180,951</td>
<td>TCTGCTTCTCTTCTGCATACG...TCTGTTACTACTTTCTAACACACATTT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Whole-System Optimization

- Amino acid sequence
- Codon choice
- Promoters, Enhancers, Rep ori, RBS strength
- Chaperonins, Proteases, Redox, Sec pathways, etc.
- Temperature, Carbon source, Dissolved $O_2$, pH

DNA Sequence Space

Protein + Gene + Vector + Strain

Fermentation

Environmental Space

Product Output
Atlas Plasmid Mapper

Demos at booth 305 9:30am, 3:30pm daily
See us at booth #305

**DNA 2.0**

Genome editing tools and services
Gene design and synthesis
Protein/pathway engineering
Universal scar-free cloning
IP-free fluorescent/colored proteins

DNA construct management
GeneDesigner software
Custom library synthesis
IP-free expression vectors
Genome synthesis