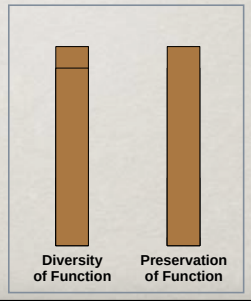
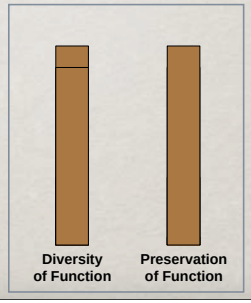
Protein Library Design

- Goal: discover new/improved protein function by screening combinatorial libraries of proteins that contain a high fraction of functional molecules displaying a wide range of functional diversity



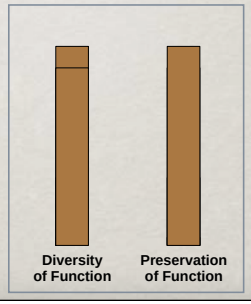
Protein Library Design

- Problem: "preservation of function" and "diversity of function" are anti-correlated when using common methods like epPCR and multi-parent gene recombination for library construction



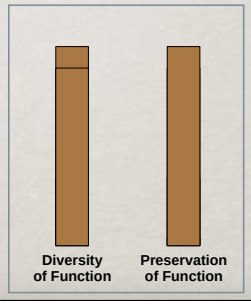
Protein Library Design

- Solution: develop structure based computational method that allows for high mutation rate (*diversity of function*) with maintenance of stably folded structure (*preservation of function*)



Protein Library Design

- Fold before Function: hypothesis is that stability (and the implied potential energy functions) can be used as a surrogate for explicitly computing function



DBIS Design Algorithm

(Diversity Benefit applied to Interacting Sets of amino acids)

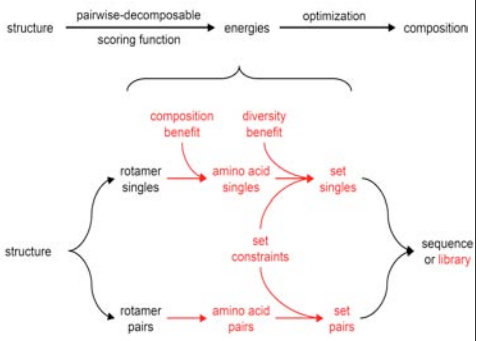
composition benefit: to enhance or force appearance of certain amino acids at certain positions -- here forces wild type sequence

diversity benefit: user controlled parameter; at fixed library size, controls mutation rate (higher benefit gives more mutated positions)

set constraints: defines clustering of amino acids -- here use of degenerate codons; set singles and pairs values calculated using an aggregation function

DBIS Design Algorithm

(Diversity Benefit applied to Interacting Sets of amino acids)

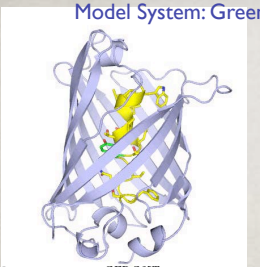


composition benefit: to enhance or force appearance of certain amino acids at certain positions -- here forces wild type sequence

diversity benefit: user controlled parameter; at fixed library size, controls mutation rate (higher benefit gives more mutated positions)

set constraints: defines clustering of amino acids -- here use of degenerate codons; set singles and pairs values calculated using an aggregation function

Model System: Green Fluorescent Protein



GFP-S65T

- 15 core positions (contiguous stretch from position 57 to 72 excluding residue 66)
- Fluorescence emission spectra recorded *in vivo* using a 96-well fluorescence plate reader
- Libraries evaluated by preservation of function (integrated fluorescence intensity) and diversity of function (extremes and dispersion of fluorescence maxima) relative to GFP-S65T

- $DBIS^{ORBIT}$: library size of 2^9 (2-fold degenerate codons at 9 positions in design region)
- $DBIS^{ORBIT}4^4$: library size of 4^4 (4-fold degenerate codons at 4 positions)
- $SCMF^{ORBIT}$: library size of 32^2 (32-fold degenerate codons at 2 positions); structure-based method of Voigt, et al. (PNAS, 2001) with saturation mutagenesis at the two positions with highest computed site entropy;
- RANDOM: library size of 2^9 (random amino acid selection at the 9 positions identified by $DBIS^{ORBIT}$)
- epPCR: error prone PCR method directed at entire gene
- All computationally designed libraries forced to contain wild-type protein sequence

Approaches for generating libraries of proteins can be broadly classified into two categories: those that rely on randomly generated sequence diversity and those that rationally attempt to predict which sequences will give a desired property. The first method includes commonly used techniques such as error-prone PCR and DNA shuffling. The latter includes “expert” design, simple sequence heuristics (e.g., hydrophobic/polar patterning) and computational design methods. While the utility of both of these approaches has been demonstrated repeatedly, a systematic comparison of many random and rational methods has not been carried out.

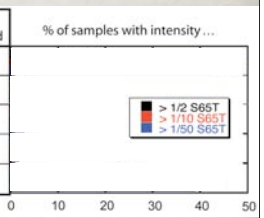
Since the combination of high-throughput methods with rational techniques may be imperative for achieving the ambitious goals of protein engineers, it is important to understand the benefits and drawbacks of different engineering

Designed Libraries

	$DBIS^{ORBIT}$	$DBIS^{ORBIT}4^4$	$SCMF^{ORBIT}32^2$	Random
57	W	W	W	W
58	PA	PAST	all	PQ
59	TS	T	T	TN
60	L	L	L	L
61	VL	VALS	V	VD
62	TA	TAGS	T	TN
63	T	T	T	T
64	F	F	F	F
65	TA	T	T	TK
67	G	G	G	G
68	VA	V	V	VM
69	QL	QELV	Q	QE
70	C	C	all	C
71	FL	F	F	FY
72	SA	S	S	SI

Preservation of Function

library	theoretical library size	average # mutations	# sampled
$DBIS^{ORBIT}$	512	4.5	1512
$DBIS^{ORBIT}4^4$	256	3.0	1344
$SCMF^{ORBIT}32^2$	1024	1.9	1260
Random	512	4.5	1509
epPCR	$\sim 10^6$	2.5	1510



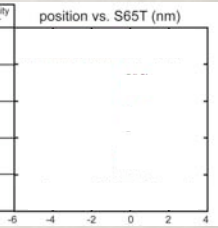
- Preservation of function increases with average mutation level among libraries design with structure-based methods in contrast to general expectations -- better to spread conservative mutations among many positions instead of saturating a few positions when constrained for library size

A functional variant was defined by having an emission spectrum that had 1/2, 1/10 or 1/20 of the integrated intensity of the spectrum for cells expressing GFP-S65T.

The 2^9 libraries designed using ORBIT had the highest fraction of functional variants. The MSA-based libraries as well as the DBISORBIT 4^4 library preformed similarly. The random and SCMF libraries performed the worst. An additional library was constructed by carrying out error-prone PCR on the entire GFP-S65T gene. The rate of mutagenesis was varied to obtain a library with approximately the same fraction of functional variants as the DBISORBIT library.

Diversity of Function at 50% Minimum Intensity

library	# w/ intensity > 1/2 S65T
DBIS^{ORBIT}	178
DBIS^{ORBIT} 4^4	12
SCMF 32^2	10
Random	12
epPCR	327



- Diversity of function increases with preservation of function among designed libraries
- Supports an approach to library design where protein stability is modeled as a surrogate for protein function (i.e., the protein must be stably folded in order to be functional)

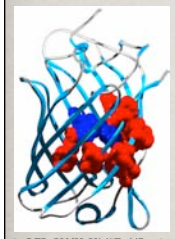
For the libraries designed using ORBIT or the MSA, there is greater diversity of function in libraries that have higher retention of function. The most extreme function is observed in the epPCR library. However, the overall distribution of function in the epPCR library is narrow, indicating that functional variants with perturbed properties are observed with a lower frequency than functional variants with peak position equivalent to that of GFP-S65T.

Conclusions

- ☑ **Preservation of Function: enhanced by using a novel structure-based computational method (DBIS)**
- ☑ **Diversity of function: greater diversity in designed libraries that better preserve function**
- ☑ **Fold before Function: computed stability is a good surrogate for function in library design**

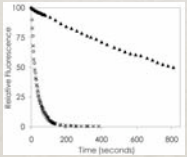
Treynor et al., *PNAS*, **104** 48, 2007

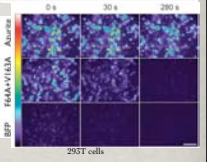
Enhanced BFP via DBIS (and some luck)



Position	61	64	145	148	150	163	165	167	203	205	220	224
wtBFP	Val	Phe	Phe	His	Val	Val	Phe	Ile	Thr	Ser	Leu	Val
	GTC	TTT	TTT	CAC	GTT	GTT	TTC	ATT	ACT	TCT	TTG	GTT
Design	DYA	TTK	KBC	CAC	DTA	RYG	TWC	ATK	RSC	KCA	MTG	RBG
	Ala	Leu	Ala	His	Ile	Ala	Phe	Ile	Ala	Ala	Leu	Ala
	Ile	Phe	Cys		Leu	Met	Tyr	Met	Gly	Ser	Met	Arg
	Leu		Gly		Val	Thr			Ser			Gly
	Ser		Phe			Val			Thr			Met
	Thr		Ser									Thr
	Val		Val									Val

IUPAC nomenclature for degenerate codons is as follows: B(CGT), D(AGT), K(GT), R(AG), S(CG), W(AT), Y(CT), M(AC)





- BFP (Y66H+Y145F; 447 nm):
  - poor quantum yield
  - pH sensitivity
  - rapid photobleaching
- Design/Screening:
  - 12 design positions
  - 3×10^6 variants (10^{15})
  - Screened by FACS (10^7)

- Azurite: BFP+F64L+V150I+V224R
- Enhanced quantum yield from 0.34 to 0.55
- Reduced pH sensitivity
- 40-fold improvement in photobleaching half-life

Mena et al., *Nature Biotech.*, 24 1569, 2006

Acknowledgments

- DBIS Design Method and Experimental Evaluation
  - Thomas Treynor
  - Christina Vizcarra
  - Daniel Nedelcu
- Enhanced BFP (U.C. Santa Barbara)
  - Marco Mena
  - Patrick Daugherty
- Funding
  - DARPA
  - Parsons Foundation
  - Weston Havens Foundation