MAX Randomisation:

Designed, non-degenerate saturation mutagenesis of armadillo repeat proteins.

Anupama Chembath[1], Mohammed Ashraf[1], Ben Wagstaffe[1], Yvonne Stark[2], Andreas Plückthun[2] and Anna V. Hine[1]
[1]School of Life & Health Sciences, Aston University, Birmingham, B4 7ET, UK.
[2]Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.

Abstract: Non-degenerate saturation mutagenesis is critical to library composition both in terms of library size and amino acid representation. Unlike conventional methodologies, non-degeneracy permits equal representation of all encoded amino acids and effectively eliminates termination codons from saturated positions. We have previously described both 'MAX' randomisation, which saturates physically separated codons and 'ProxiMAX' randomisation, which saturates contiguous codons. Both allow the additional advantage of encoding only required amino acids without reference to codon sequence, but their methodologies are quite different. Whilst MAX randomisation employs a manual process of selectional hybridisation between individual oligonucleotides and a conventionally-randomised template, ProxiMAX relies on saturation cycling; repeated cycles of blunt-ended ligation, Type IIS restriction and PCR amplification. MAX randomisation is thus typically employed in the research laboratory to engineer active sites of enzymes (or α-helices within other proteins), whilst ProxiMAX is now a chiefly commercial, automated process employed in antibody engineering. Here we present the application of MAX randomisation to engineer libraries of Armadillo Repeat Proteins (ArmRPs), α-helical proteins that selectively bind extended peptides. We have utilised the MAX randomisation technique to engineer ArmRPs for the generation of gene libraries encoding multiple repeat modules, saturating seven positions and encoding between 4 and 18 amino acids within each location, achieving excellent correlation between the design and observed specifications. We also present early developments to extend MAX randomisation into the realm of multiple contiguous codons.

Why non-degeneracy matters

Degeneracy limits diversity. Diversity is the percentage of unique species within a library and as such, is a measure of library quality. The importance of non-degeneracy increases with the number of codons saturated (**Figure 1**).

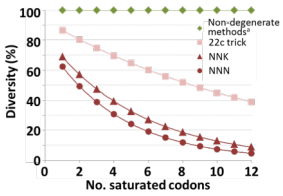

Figure 1: Theoretical diversity within saturated libraries. Green coloration indicates ideal performance; pale pink coloration indicates tolerable performance and deep pink coloration indicates unacceptable performance. Diversity was calculated using the formula d=1/(N∑kpk2) and is in agreement for a 12-mer peptide saturated with codon NNN (Makowski & Soares, 2003; Krumpe *et al.*, 2007). 22c trick, Kille *et al.*, 2013; [a]MAX & ProxiMAX Randomisation, Hughes et al., 2003, Ashraf et al., 2013).

Degeneracy encodes bias. Typical screening protocols rely on mass action, which necessitates approximately equivalent concentrations of all library components. Thus, if a protein is over/under represented in the encoded library **(Table 1)**, screening will mis-represent its activity. False negatives / positives result.

	Ratio most common : rarest codon combinations			
No. saturated codons	NNN	NNK / NNS	22c-trick	Non-degenerate methods
3	216:1	27:1	8:1	1:1
6	4.7×10^4:1	729:1	64:1	
9	1.0×10^7:1	2.0×10^4:1	512:1	
12	2.2×10^9:1	5.3×10^5:1	4096:1	

Table 1: Encoded bias within saturated libraries. Ratios represent the theoretical relative concentrations of each individual gene combining any of the most common codons (Leu/Arg/Ser, NNN/NNK; or Leu/Val, 22c trick) versus each individual gene containing any combination of the rarest codons (Met/Trp, NNN; Cys/Asp/Glu/Phe/His/Ile/Lys/Met/Asn/Gln/Trp/Tyr, NNK; or 18 codons (omitting Leu/Val), 22c trick[3]).

Degeneracy encodes truncation. Truncation (Table 2) is problematic because truncated, non-functional proteins are prone to aggregation, leading to precipitation.

No. saturated codons	NNN	NNK / NNS	22c-trick	Non-degenerate methods[a]
3	13%	9%	0%	
6	25%	17%		
9	35%	25%		
12	44%	32%		

Table 2: Encoded truncation within saturated libraries. Truncation is calculated as the percentage of sequences that contain 1 or more termination codons within the saturated region.

What are MAX codons? MAX codons are a **non-degenerate** mixture of **selected** codons (exactly one codon per amino acid). MAX uses a mixture of codons that encode all 20 amino acids, or any lesser set of amino acids selected. The DNA sequence of the codons is unimportant and can be selected for maximal expression in any organism.

Figure 2: Exemplar MAX codons. Codons chosen to encode all amino **acids except Cys and Met, selected according to** *E. coli* codon preference, for MAXimal expression. Note that the NNN would be required as a conventional degenerate codon to include all in this selection.

1st Base \ 2nd Base	A	C	G	T	3rd Base
A	Lys	Thr	Arg	Ile	A
	Asn	Thr	Ser	Ile	C
	Lys	Thr	Arg	Met	G
	Asn	Thr	Ser	**Ile**	T
C	Gln	Pro	Arg	Leu	A
	His	Pro	**Arg**	Leu	C
	Gln	**Pro**	Arg	**Leu**	G
	His	Pro	Arg	Leu	T
G	**Glu**	Ala	Gly	Val	A
	Asp	Ala	**Gly**	Val	C
	Glu	**Ala**	Gly	**Val**	G
	Asp	Ala	Gly	Val	T
T	*	Ser	*	Leu	A
	Tyr	Ser	Cys	Phe	C
	*	Ser	**Trp**	Leu	G
	Tyr	Ser	Cys	**Phe**	T

MAX Randomisation

While ProxiMAX randomisation saturates contiguous codons (e.g. CDR loops in antibodies), MAX randomisation saturates separated codons, such as those encoding residues on the surface of α-helices. PRe-ART, employs MAX randomisation to saturate peptide-binding residues within Armadillo Repeat Proteins (ArmRPs). ArmRPs are naturally-occurring eukaryotic proteins characterised by a right-handed superhelix formed by 4–12 stacked tandem armadillo repeat motifs. Each repeat consists of approximately 42 amino acids folded into three α-helices in a triangular assembly. ArmRPs are abundant in eukaryotes, involved in signalling, nucleocytoplasmic transport and cell adhesion. They bind peptides of various lengths remarkably consistently, in extended conformations (**Figure 3**).

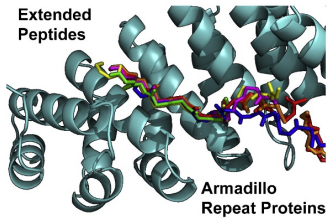

Figure 3: Peptides in extended conformation bound by a designed Armadillo Repeat Protein (Varadamsetty et al., 2012).

MAX methodology: MAX randomisation works via a process of selectional hybridisation followed by ligation and asymmetric amplification (**Figure 3**). Because MAX codons are non-degenerate, MAX randomisation can add all 20 chosen codons (or any required subset) at each saturated position.

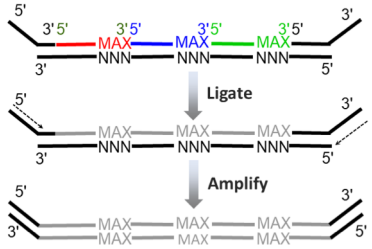

Figure 4: Schematic representation of the MAX randomisation process. A single template oligonucleotide is synthesized to be fully-degenerate at the designated, saturated codons. Meanwhile, a set of up to 20 small selection oligonucleotides are synthesised individually, for each saturated position. Each selection oligonucleotide consists of a short (typically in the order of 6bp) addressing region that is fully-complementary to the template and one MAX codon. The selection oligonucleotides are mixed as required and alongside two terminal oligonucleotides, are hybridised with the template and ligated together. The ligated strand is then selectively amplified with primers complementary to the terminal oligonucleotides, to generate a randomisation cassette (Hughes et al., 2003).

Exemplar MAX library design:

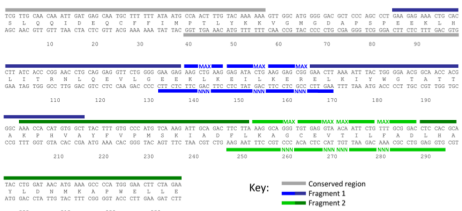

Figure 5: Exemplar MAX library design. A MAX library will typically be designed in multiple fragments that are synthesised as described in Figure 4. After asymmetric amplification, the fragments are assembled together by overlap PCR (or alternative methodologies). Saturation of 7 codons is illustrated. **NB.** Illustrated DNA / peptide sequences are arbitrary and do not reflect ArmRP gene or protein sequences.

ArmRP library construction

An ArmRP library was designed in three sections to contain one fully-conserved region, one containing 3 saturated codons (Fragment 1) and a third containing 4 saturated codons (Fragment 2), similar in concept to the design illustrated in Figure 5. The three sections were synthesised independently (Figure 6A). Since the completed library is relatively small, with a theoretical diversity of 1.1 x 10^8 components, individual fragments were combined by overlap PCR (Figure 6B).

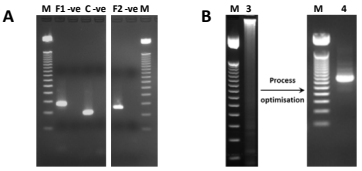

Figure 6: ArmRP library construction. A) synthesis of the conserved region by overlap PCR, and Fragments 1 and 2 by MAX randomisation. **B)** Optimisation of overlap PCR to combine the three sections. Required process optimisations included annealing temperature, fragment dilutions and PCR cycles. **Lanes:** M, 50 bp ladder; -ve, negative controls; C, conserved region (99 bp); F1, F2, Fragments 1 (132 bp) & 2 (135 bp) respectively; 3 & 4, overlap PCR before and after process optimisation respectively. The predicted size of the final PCR fragment is 336 bp.

Quality control: The resulting library was examined both by Sanger sequencing and by Illumina NGS, revealing good correlation between library design and content (Figure 7).

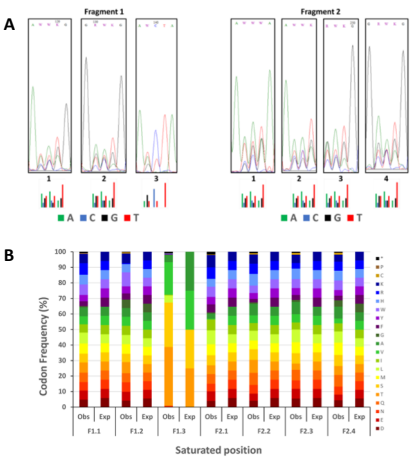

Figure 7: Quality control data. A) Chromatograms of saturated codons and neighbouring, conserved bases within the completed library. Histograms represent the designed distribution of A,C,G & T at each of the three bases within each saturated codon. **B)** NGS data showing observed and expected (designed) distribution of codons within the completed library.

ParaMAX: applying MAX to contiguous codons: One disadvantage of MAX randomisation is its limitation to two contiguous codons (resulting from addressing regions in selection oligos; see centre of Fragment 2, Figure 5). However, akin to Golden Gate cloning and ProxiMAX randomisation, judicious use of Type IIS restriction enzymes can allow groups of two contiguous saturated codons created by MAX randomisation to be linked together seamlessly, though unlike ProxiMAX, the randomisation oligos cannot be re-used – they are unique to each codon.

References

- Ashraf, M, Frigotto, L, Smith, ME, Patel, S, Hughes, MD, Poole, AJ, Hebaishi, HRM, Ullman, CG & Hine, AV (2013) *Biochem. Soc. Trans.* **41**, 1189–1194.
- Hughes, MD, Nagel, DA, Santos, AF, Sutherland, AJ & Hine, AV (2003) *J. Mol. Biol.* **331**, 967-972.
- Kille, S, Acevedo-Rocha, CG, Parra, LP, Zhang, Z-G, Opperman, DJ, Reetz, MT & Acevedo, JP (2013) *ACS Synth. Biol.* **2**, 83–92.
- Krumpe, LRH, Schumacher, KM, McMahon, JB, Makowski, L & Mori, T (2007) *BMC Biotechnology* **7**, 65.
- Makowski, L & Soares, A (2003) *Bioinformatics* **19**, 483-489.
- Varadamsetty, G, Tremmel, D, Hansen, S, Parmeggiani, F & Plückthun, A. (2012). *J. Mol Biol* **424**, 68-87.