ProxiMAX randomization: a new technology for non-degenerate saturation mutagenesis of contiguous codons

Mohammed Ashraf*[1], Laura Frigotto†, Matthew E. Smith†, Seema Patel†, Marcus D. Hughes*[1], Andrew J. Poole*, Husam R.M. Hebaishi*, Christopher G. Ullman† and Anna V. Hine*[1,2]

*School of Life and Health Sciences, Aston University, Aston Triangle, Birmingham B4 7ET, U.K., and †Isogenica Ltd, The Mansion, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, U.K.

Abstract

Back in 2003, we published 'MAX' randomization, a process of non-degenerate saturation mutagenesis using exactly 20 codons (one for each amino acid) or else any required subset of those 20 codons. 'MAX' randomization saturates codons located in isolated positions within a protein, as might be required in enzyme engineering, or else on one face of an α-helix, as in zinc-finger engineering. Since that time, we have been asked for an equivalent process that can saturate multiple contiguous codons in a non-degenerate manner. We have now developed 'ProxiMAX' randomization, which does just that: generating DNA cassettes for saturation mutagenesis without degeneracy or bias. Offering an alternative to trinucleotide phosphoramidite chemistry, ProxiMAX randomization uses nothing more sophisticated than unmodified oligonucleotides and standard molecular biology reagents. Thus it requires no specialized chemistry, reagents or equipment, and simply relies on a process of saturation cycling comprising ligation, amplification and digestion for each cycle. The process can encode both unbiased representation of selected amino acids or else encode them in predefined ratios. Each saturated position can be defined independently of the others. We demonstrate accurate saturation of up to 11 contiguous codons. As such, ProxiMAX randomization is particularly relevant to antibody engineering.

Background

Saturation mutagenesis (replacement of wild-type codons with codons for all 20 amino acids) is a core technique within the protein engineer's repertoire. Its importance in engineering non-native ligand-binding domains is undisputed. Thus saturation mutagenesis has played a vital role in creating synthetic zinc-finger-based transcription factors [1,2], antibodies and antibody-derived scaffolds [3,4] for many years, and, more recently, has proved valuable in engineering modified enzymes [5].

Conventional saturation mutagenesis employs degenerate codons NNN, NNK or NNS. Although technically undemanding, degeneracy leads to significant problems that have provoked a drive for non-degenerate alternatives. Traditionally, non-degeneracy was achieved by using trinucleotide phosphoramidites which add whole codons (rather than single bases) during oligonucleotide synthesis [6]. In 2003, we described 'MAX' randomization [7], which uses standard oligonucleotides and is useful for enzyme engineering, but requires separation between randomized codons and thus cannot saturate more than two contiguous codons. Recently, simpler alternatives of 'small-intelligent libraries' designed by the program DC-analyzer [8] and the '22c trick' [9] have been described, which are optimal methodologies to saturate small numbers of codons effectively and efficiently, irrespective of location; although, owing to multiplex PCR primers, these approaches cannot saturate larger numbers of codons (see below). In the present paper, we describe ProxiMAX randomization, which offers all the advantages of Slonomics^{TM} [10,11] (an automated, non-degenerate, enzyme-based process), but can be performed in a standard molecular biology laboratory. ProxiMAX likewise combines the benefits of non-degeneracy with the ability to saturate larger numbers of contiguous codons, a key requirement for antibody engineering.

Saturation mutagenesis: comparison of techniques

The advantages and disadvantages of the various approaches to saturation mutagenesis are compared in Figure 1. The most critical consequences of degeneracy are the loss of diversity/functionality [12,13] and inherent encoded bias [7]. Diversity is a measure of the percentage of unique species within a library. Even the '22c trick' [9] (NDT/VHG/TGG degeneracy) leads to >60% loss of diversity over 12

Key words: antibody engineering, gene library, MAX randomization, overlap PCR, protein engineering, saturation mutagenesis.

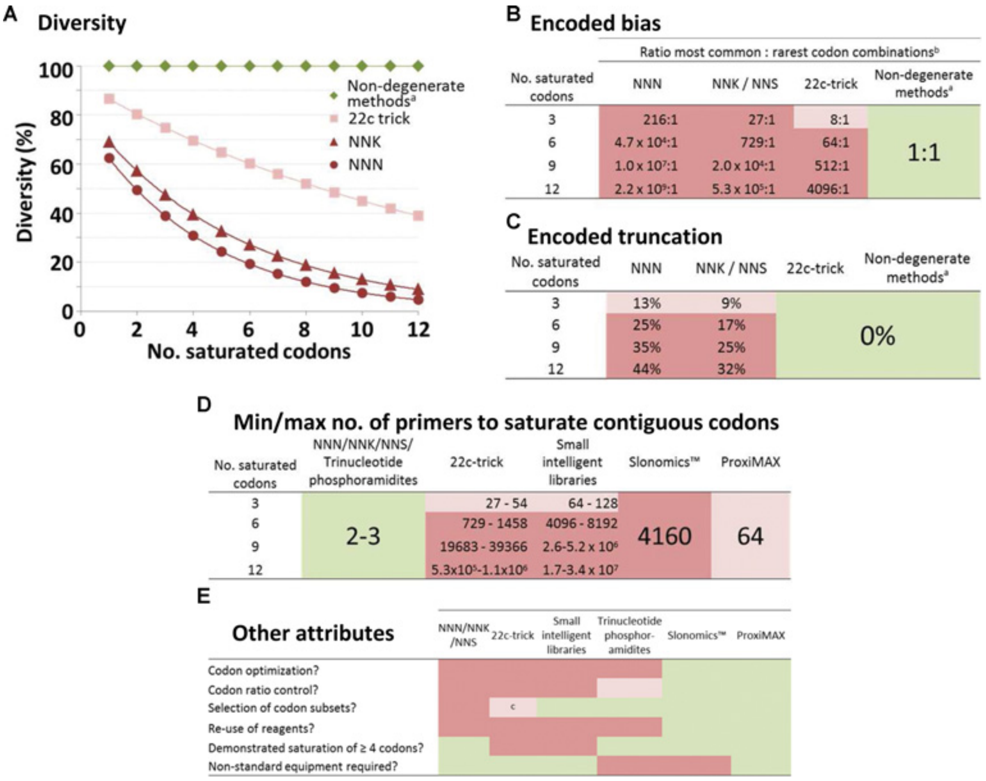
Abbreviation used: CDR3, complementarity-determining region 3.

[1] ProxiMAX randomization is protected by patent numbers EP 1907548, EP 2236612 and US 8357638 B2 filed by Mohammed Ashraf, Marcus Hughes and Anna Hine, and is operated commercially under the trade name Colibra™ (Isogenica).

[2] To whom correspondence should be addressed (email a.v.hine@aston.ac.uk).

Figure 1 | Comparison of performance of common saturation mutagenesis techniques

Green coloration indicates ideal performance, pale pink coloration indicates tolerable performance, and deep pink coloration indicates unacceptable performance, where non-degenerate methods ([a]) include 'small-intelligent libraries' [8], Slonomics^{TM} [10,11] and ProxiMAX. (**A**) Diversity was calculated using the formula $d = 1/(N\Sigma_k p_k{}^2)$ [12] and is in agreement for a 12-mer peptide saturated with codon NNN [13]. (**B**) Ratios represent the theoretical relative concentrations of each individual gene combining any of the most common codons (leucine/arginine/serine, NNN/NNK; or leucine/valine, 22c trick) compared with each individual gene containing any combination of the rarest codons ([b]) [methionine/tryptophan, NNN; cysteine/aspartate/glutamate/phenylalanine/histidine/isoleucine/lysine/methionine/asparagine/glutamine/tryptophan/tyrosine, NNK; or 18 codons (omitting leucine/valine), 22c trick]. (**C**) Truncation is calculated as the percentage of sequences that contain one or more termination codons within the saturated region. (**D**) NNN/NNK/trinucleotide oligonucleotides may be used as a DNA cassette, or as primers in PCR-based mutagenesis; Slonomics^{TM} and ProxiMAX require a fixed number of oligonucleotides and the numbers of primers for the 22c trick [9] and small-intelligent libraries [8] were calculated using the formulae in the respective publications for saturating consecutive codons. (**E**) [c]As dictated by degeneracy limitations.



B Encoded bias

No. saturated codons	NNN	NNK / NNS	22c-trick	Non-degenerate methods[a]
	Ratio most common : rarest codon combinations[b]			
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	

C Encoded truncation

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%		

D Min/max no. of primers to saturate contiguous codons

No. saturated codons	NNN/NNK/NNS/ Trinucleotide phosphoramidites	22c-trick	Small intelligent libraries	Slonomics™	ProxiMAX
3	2-3	27 - 54	64 - 128	4160	64
6		729 - 1458	4096 - 8192		
9		19683 - 39366	2.6-5.2 x 10^6		
12		5.3×10^5-1.1×10^6	1.7-3.4 x 10^7		

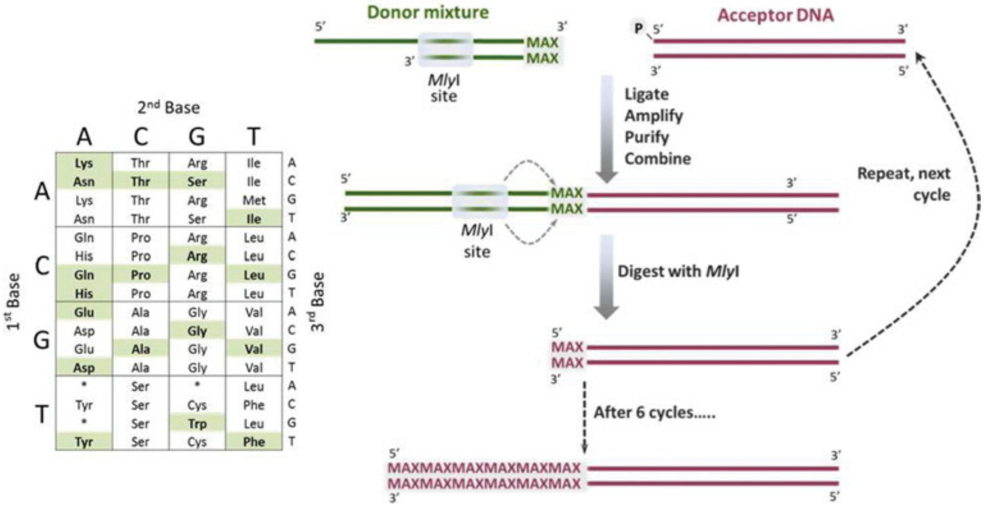
E Other attributes

	NNN/NNK /NNS	22c-trick	Small intelligent libraries	Trinucleotide phosphor-amidites	Slonomics™	ProxiMAX
Codon optimization?	🟥	🟥	🟥	🟥	🟩	🟩
Codon ratio control?	🟥	🟥	🟥	▨	🟩	🟩
Selection of codon subsets?	🟥	▨ c	🟩	🟩	🟩	🟩
Re-use of reagents?	🟥	🟥	🟥	🟥	🟩	🟩
Demonstrated saturation of ≥ 4 codons?	🟩	🟥	🟥	🟩	🟩	🟩
Non-standard equipment required?	🟩	🟩	🟩	🟥	🟥	🟩

saturated codons (Figure 1A). It might be argued that excess screening capacity (e.g. with ribosome or CIS display [14]) diminishes this issue, but in libraries with more than three randomized positions, the use of degenerate codons will probably have a severe impact on the quality of the output. Since the magnitude of protein–ligand interactions is a function of both affinity and concentration, all display screens are based on a pretext of equivalent concentration of library members. Figure 1(B) demonstrates that the concentration differences between common and rare codon combinations are unworkable beyond three degenerate saturated codons, regardless of methodology. Library diversity is further restricted by NNK and NNN saturations which randomly introduce termination codons (Figure 1C) that may lead to non-functional proteins, possibly leading to aggregation. Practical issues must also influence method choice. Notwithstanding objections of bias, the number of primers required to saturate more than three consecutive codons using either 'small-intelligent libraries' or the '22c trick' are impractical to handle manually (Figure 1D). Neither do degenerate methods (including the '22c trick') allow the potential to exclusively eliminate cysteine, which is usually undesired in protein and peptide libraries. Finally, Figure 1(E) compares other desirable attributes of saturation techniques, including the ability to select codons to suit the organism of choice, including codon optimization, ratio-control, subset-selection, etc. Thus we propose that ProxiMAX is the first technology to offer all desirable attributes in a manual setting and, as such, will be an invaluable addition to the protein engineer's toolbox.

Figure 2 | ProxiMAX methodology

Double-stranded DNA donors, carrying the required 'MAX' codons at their termini, are ligated individually on to a double-stranded DNA acceptor sequence (phosphorylated at the required 5′ end only). The donors can take the form of partially double-stranded DNA (as shown), fully double-stranded DNA or hairpin oligonucleotides. After ligation, the products are amplified, purified, quantified and then combined in the required ratios. The combined product is digested with MlyI. The process is then repeated, using the digestion product from cycle 1 as the acceptor for the next round of ligation. Different sets of donors are cycled to prevent potential carry-over from one cycle to the next. The inset of the genetic code shows that any codons may be selected as 'MAX' codons, regardless of their sequence.



1st 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

ProxiMAX randomization: the concept

ProxiMAX randomization is based on iterative cycles of blunt-ended ligation, amplification and digestion with the Type IIS restriction endonuclease MlyI. More specifically, oligonucleotide donors (containing the required codons at their termini) are ligated on to a conserved acceptor sequence. Following ligation, the PCR amplification step provides ample ligated product, while concomitantly diluting non-ligated contaminants. Subsequent digestion removes all but the final codon of the donor sequences to generate a new acceptor, which enters the next cycle (Figure 2). By alternating different sets of oligonucleotide donors for each cycle, codon additions can be restricted to a specific saturation cycle.

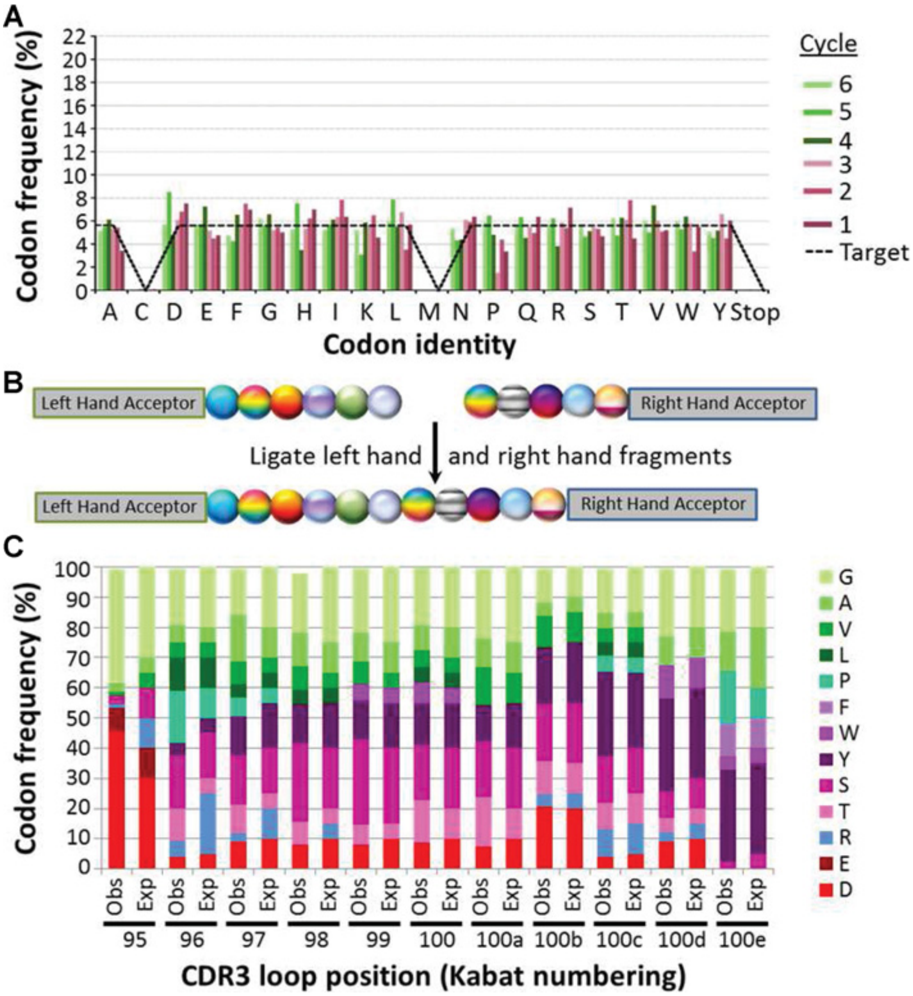
ProxiMAX development: T4 DNA ligase exhibits some sequence preference in blunt-end ligations

To examine process feasibility, preliminary experiments used a mixture of all 20 donors combined in nominally equal quantities (based solely on manufacturer's data). Results demonstrated that the process worked, but, unsurprisingly, there was bias in the resulting small-scale library (Supplementary Figure S1 at http://www.biochemsoctrans.org/bst/041/bst0411189add.htm). Bias might result from variant oligonucleotide quality, inaccurate estimations of oligonucleotide concentration, sequence preferences of T4 ligase, amplification bias during PCR or even sequence preference of the restriction enzyme MlyI. To examine and/or exclude these factors, ligations and amplifications were performed in individual parallel reactions. Only after purification and quantification were the individual amplicons combined. Analysis at this midpoint in each of six cycles of saturation (i.e. after combination of individual reactions and before MlyI digestion) demonstrated that equimolar mixes (5% of each codon) had been generated successfully during each cycle of addition (Supplementary Figure S2A at http://www.biochemsoctrans.org/bst/041/bst0411189add.htm). However, these data offered no insight as to whether subsequent ligations would proceed in an equimolar fashion when those mixtures acted as acceptors. Therefore, after six cycles of saturation mutagenesis, composition of the final assembled library was assessed by Roche 454 pyrosequencing, which demonstrated that, whereas most codons were introduced into the library at the expected ratio, a few were noticeably over- and under-represented. In particular, the histidine codon 5′-CAT-3′ was favoured, whereas codons for lysine and threonine were under-represented (Supplementary Figure 2B).

By eliminating donor concentration and/or quality and PCR preference as potential sources of bias, these data suggest that T4 DNA ligase exhibits some sequence preference in blunt-ended ligations, when presented with a mixture of 5′ and 3′ ends, so that, in this substrate mixture, the enzyme has a different reactivity rate for different codons. To examine this possibility, data from the preceding 6-mer sequence analysis were studied for expected and observed ratios with respect

Figure 3 | Library construction with controlled codon ratios

(**A**) DNA sequence analysis of a library encoded by six cycles of codon addition, optimized to provide unbiased representation. Six cycles of ProxiMAX randomization were undertaken as described in Supplementary Figure S2 (at http://www.biochemsoctrans.org/bst/041/bst0411189add.htm), using right-handed hairpins (Supplementary Table S3 at http://www.biochemsoctrans.org/bst/041/bst0411189add.htm) as donors and an amplicon of pUC19 as the acceptor, with optimized mixtures of 18 codons adjusted to reflect the sequence bias illustrated in Supplementary Figure S2(B). The resulting library was analysed by DNA sequencing, using a MiSeq DNA sequencer according to the manufacturer's instructions. Data represent the analysis of 286684 sequences of the correct length, which represented 79.9% of the entire library. Bars represent the frequency of each codon from each cycle of saturation mutagenesis. The broken line depicts the target representation for each codon. Further analysis of the library can be found in Supplementary Table S2 (at http://www.biochemsoctrans.org/bst/041/bst0411189add.htm). (**B**) Schematic representation of the procedure for assembling a representative 11-mer CDR3 antibody library. Left-hand (LH, six cycles) and right-hand (RH, five cycles) of codon additions were performed in parallel as described in Supplementary Figure S2(B) with adjusted ratios of codons to compensate for sequence preference as determined in Supplementary Figure S2(B). After the final MlyI digestions, the two assemblies were joined together to generate a region of 11 contiguous saturated codons. (**C**) DNA sequence analysis of the representative 11-mer CDR3 antibody library. The library was assembled as described in (**B**) and sequenced using a MiSeq sequencer according to the manufacturer's instructions. Data represent the analysis of 461274 sequences of the correct length, which comprised 74.1% of the entire library. Single codon deletions comprised an additional 7.4% of that library, whereas double and triple codon deletions comprised 2.0% and 1.2% respectively. Expected (designed) frequencies of each codon are compared with observed frequencies for locations 95–100e within the CDR3 loop. Further analysis of the library can be found in Supplementary Table S2.



to the last base in the phosphate donor/acceptor junction. A χ^2 statistic with one degree of freedom was calculated for the 16 possible combinations of 3′–5′ preference at the ligation junction and where $P < 0.0001$, this was noted (Supplementary Table S1 at http://www.biochemsoctrans.org/bst/041/bst0411189add.htm). Although there were no discernible rules for the junction preference directly at the point of ligation, there are certainly preferences for 3′-T and 5′-C, which supports the observed over-representation for histidine (CAT), but does not fully explain the reason for histidine being more preferred than arginine (CGT).

Successful application of ProxiMAX randomization

To compensate for these apparent sequence preferences, the six cycles of addition were repeated, with the concentration of each donor sequence being adjusted during mixing, to counteract the over- or under-representation shown in Supplementary Figure S2(B). Thus concentrations of donors carrying over-represented codons were reduced, whereas those carrying under-represented codons were increased. Cysteine and methionine codons were also excluded as these are deemed to have liabilities in peptide libraries by being susceptible to oxidation. It is clear from Figure 3(A) that T4 DNA ligase sequence preference can be compensated for by adjustment of donor concentrations.

Having demonstrated successful equimolar representation, we next sought to create a more relevant library from a protein engineering perspective. Saturation of 11 consecutive codons within the loop of a CDR3 (complementarity-determining region 3) domain of an antibody V_H (variable heavy) chain domain was undertaken. Each position in the loop required a separate mixture of codons so that favourable protein-binding amino acids were included at the correct positions and those that would encode structural or manufacturing liabilities were excluded. The frequencies of the desired amino acids were rounded to the nearest 5% in order to simplify mixing of the codon oligonucleotides and analysis of the final library.

For this longer loop, both right-hand and left-hand acceptors were employed to build two library components in parallel (Figure 3B). The 11-mer CDR3 loop was then assembled by ligating the two fragments. The loop was sequenced, and remarkable compliance was achieved between design specifications (expected) and the actual library generated (observed; Figure 3C). One notable anomaly was the arginine codon which was poorly represented in the left-hand acceptor ligations. This is apparent in Figure 2(C), where the lack of under-representation of arginine at a particular addition has led to a compensating over-representation of other amino acids, for example aspartate at position 95. Despite this, the majority of codons are remarkably close to their desired frequency.

Conclusions

Although more involved than conventional degenerate saturation mutagenesis, we believe that the benefits of ProxiMAX randomization easily outweigh the increased outlay on reagents and of time. One to two cycles of ProxiMAX randomization can be achieved in a day and thus our exemplar 11-mer library can readily be prepared in 2 weeks. The bottleneck of directed evolution is library screening [9] and thus the smaller and more accurate the initial library, the greater the efficiency of sampling in selection or screening and the better the likely outcome. As stated by Neugebauer and co-workers (Morphosys AG) “The most critical parameter for all *in vitro*-based approaches is the quality of the antibody library” [15]. Thus although the requirement to perform individual ligations and PCRs may initially seem unwieldy, performing 20 parallel reactions from a single mastermix is not a significant undertaking. The phrase ‘if you want a good result, preparation is nine-tenths of the work’ was never more apposite.

Acknowledgements

We gratefully acknowledge Dr C. Baar for assistance with data collection and analysis, Dr A.J. Sutherland and Dr D.A. Nagel (Aston University) for helpful discussions and Professor A.R. Pitt (Aston University) and Professor J.D. Sutherland (Laboratory of Molecular Biology, Cambridge, U.K.) for a critical reading of drafts before submission.

Funding

This work was supported by an Aston University departmental studentship (to M.A.), the Biotechnology and Biological Sciences Research Council [grant number BB/D525756/1 to A.V.H. (M.D.H.)], an Aston Research Centre for Healthy Aging (ARCHA) studentship (to H.R.M.H.) and Technology Strategy Board [grant number 720032 (to Isogenica)].

References

1 Choo, Y. and Klug, A. (1994) Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proc. Natl. Acad. Sci. U.S.A. **91**, 11163–11167

2 Jamieson, A.C., Kim, S.H. and Wells, J.A. (1994) *In vitro* selection of zinc fingers with altered DNA-binding specificity. Biochemistry **33**, 5689–5695

3 Chen, G., Dubrawsky, I., Mendez, P., Georgiou, G. and Iverson, B.L. (1999) *In vitro* scanning saturation mutagenesis of all the specificity determining residues in an antibody binding site. Protein Eng. **12**, 349–356

4 Caravella, J. and Lugovskoy, A. (2010) Design of next-generation protein therapeutics. Curr. Opin. Chem. Biol. **14**, 520–528

5 Reetz, M.T., Prasad, S., Carballeira, J.D., Gumulya, Y. and Bocola, M. J. (2010) Iterative saturation mutagenesis accelerates laboratory evolution of enzyme stereoselectivity: rigorous comparison with traditional methods. Am. Chem. Soc. **132**, 9144–9152

6 Virnekäs, B., Ge, L.M., Pluckthun, A., Schneider, K.C., Wellnhofer, G. and Moroney, S.E. (1994) Trinucleotide phosphoramidites-ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. Nucleic Acids Res. **22**, 5600–5607

7 Hughes, M.D., Nagel, D.A., Santos, A.F., Sutherland, A.J. and Hine, A.V. (2003) Removing the redundancy from randomised gene libraries. J. Mol. Biol. **331**, 967–972

8 Tang, L., Gao, H., Zhu, X., Wang, X., Zhou, M. and Jiang, R. (2012) Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers. BioTechniques **52**, 149–158

9 Kille, S., Acevedo-Rocha, C.G., Parra, L.P., Zhang, Z.-G., Opperman, D.J., Reetz, M.T. and Acevedo, J.P. (2013) Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. ACS Synth. Biol. **2**, 83–92

10 Van den Brulle, J., Fischer, M., Langmann, T., Horn, G., Waldmann, T., Arnold, S., Fuhrmann, M., Schatz, O., O'Connell, T., O'Connell, D. et al. (2008) A novel solid phase technology for high-throughput gene synthesis. BioTechniques **45**, 340–343

11 Zhai, W., Glanville, J., Fuhrmann, M., Mei, L., Ni, I., Sundar, P.D., Van Blarcom, T., Abdiche, Y., Lindquist, K., Strohner, R. et al. (2011) Synthetic antibodies designed on natural sequence landscapes. J. Mol. Biol. **412**, 55–71

12 Makowski, L. and Soares, A. (2003) Estimating the diversity of peptide populations from limited sequence data. Bioinformatics **19**, 483–489

13 Krumpe, L.R.H., Schumacher, K.M., McMahon, J.B., Makowski, L. and Mori, T. (2007) Trinucleotide cassettes increase diversity of T7 phage-displayed peptide library. BMC Biotechnol. **7**, 65

14 Odegrip, R., Coomber, D., Eldridge, B., Hederer, R., Kuhlman, P.A., Ullman, C., FitzGerald, K. and McGregor, D. (2004) CIS display: *in vitro* selection of peptides from libraries of protein–DNA complexes. Proc. Natl. Acad. Sci. U.S.A. **101**, 2806–2810

15 Ponsel, D., Neugebauer, J., Kathrin Ladetzki-Baehs, K. and Tissot, K. (2011) High affinity, developability and functional size: the holy grail of combinatorial antibody library generation. Molecules **16**, 3675–3700

Received 26 June 2013
doi:10.1042/BST20130123

SUPPLEMENTARY ONLINE DATA

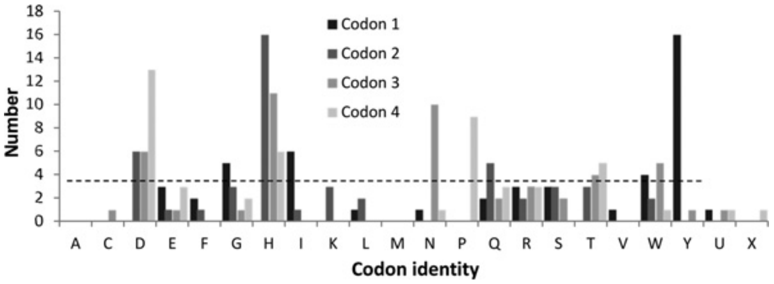
ProxiMAX randomization: a new technology for non-degenerate saturation mutagenesis of contiguous codons

Mohammed Ashraf*[1], Laura Frigotto†, Matthew E. Smith†, Seema Patel†, Marcus D. Hughes*[1], Andrew J. Poole*, Husam R.M. Hebaishi*, Christopher G. Ullman† and Anna V. Hine*[1,2]

*School of Life and Health Sciences, Aston University, Aston Triangle, Birmingham B4 7ET, U.K., and †Isogenica Ltd, The Mansion, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, U.K.

Figure S1 | **Preliminary assessment of ProxiMAX randomization**

Individual pairs of complementary oligonucleotides (Table S3; MAX_{RH1}–MAX_{RH4} and Rev_{RH}) were hybridized together, at 10 μM final concentration (using the manufacturer's stated concentrations for each oligonucleotide) and then diluted to a final concentration of 1 μM. These hybridized stocks were then combined to generate equimolar mixtures (R1 pairs for cycle 1 etc.). Then 10 pmol of the R1 mixture was ligated to 3.3 pmol of test acceptor, in a 20 μl reaction volume. The ligation was diluted 1 in 1000 and duplicate 100 μl PCR amplifications were performed, using a high-fidelity polymerase (Pfu). The two PCRs were combined and processed using a QIAquick® PCR purification kit (Qiagen). The resulting DNA was resuspended in 30 μl of water and 25 μl of the cleaned product digested using 10 units of fast-digest MlyI (Fermentas) in a 50 μl reaction volume, which was then heat-inactivated. The process was then repeated using the next set of oligonucleotide donors and primer (R2 and primer 2 for cycle 2, etc.), except that, at the ligation stage, 5 μl of MlyI-restricted DNA from the previous cycle (estimated to be ~3.3 pmol of DNA) replaced the previous acceptor. Two further cycles of saturation (cycles 3 and 4) were performed, using new sets of oligonucleotide donors as primers as indicated in Table S3. The products of the fourth round of saturation cycling (before MlyI digestion) were inserted into SmaI- and alkaline phosphatase-treated pUC19 (Fermentas). The histogram represents analysis of inserts contained within 48 resulting clones, where U represents unreadable codons and X represents non-MAX codons. Ideal 2.4% representation of each codon at each location is indicated by the broken line (but note that only integer values of cloned codons can be achieved experimentally).



[1] ProxiMAX randomization is protected by patent numbers EP 1907548, EP 2236612 and US 8357638 B2 filed by Mohammed Ashraf, Marcus Hughes and Anna Hine, and is operated commercially under the trade name Colibra™ (Isogenica).

[2] To whom correspondence should be addressed (email a.v.hine@aston.ac.uk).

Figure S2 | **Sources of bias in ProxiMAX randomization**

The protocol described in Figure S1 was modified to identify and/or eliminate potential sources of library bias. To exclude oligonucleotide concentration differences during annealing, right-handed hairpin donors were employed (Table S3). To exclude potential T4 DNA ligase preference for donor sequences and/or amplification bias, ligations and amplifications were performed in 20 individual parallel reactions. Individual PCR products were electrophoresed through pre-cast 2% agarose gels, gel-extracted and quantified by UV absorbance. They were then combined in equimolar amounts and the combined product was digested with MlyI. The digested product was then used as the acceptor in the subsequent saturation cycle. Six cycles of saturation were performed using MAX$_{RH}$ hairpins R1–R3 in sequential succession. (**A**) DNA sequence analysis of codon mixing. To determine whether the equimolar mixing had been achieved, mixtures were sampled midway through each cycle (i.e. after equimolar combination and before MlyI digestion), and analysed by Roche 454 pyrosequencing to assess the identity and proportion of the latest codon addition (each donor should be present at 5% representation). Bars represent the frequency of each codon during each saturation cycle. (**B**) DNA sequence analysis of a library encoding a 6-mer random peptide mixture, made from equimolar mixtures of codons at each position. To determine the influence of acceptor sequence on T4 DNA ligase, the completed library (after six cycles of randomization) was also analysed by Roche 454 pyrosequencing. Data represent the analysis of 9074 sequences of the correct length, which represented 85.1% of the entire library. Bars represent the frequency of each codon in each saturated position. The broken line depicts the designed representation of 5% for each codon.

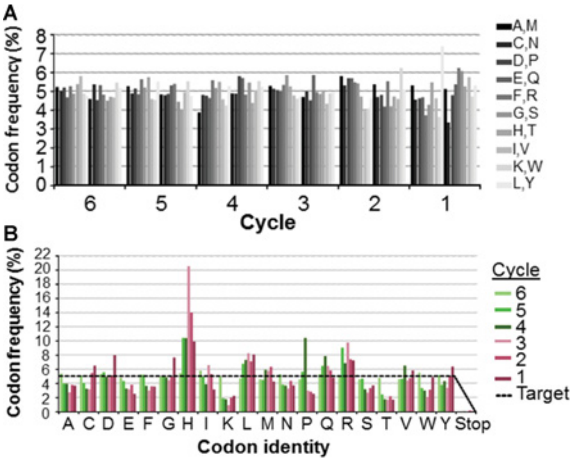


Table S1 | **Examination of potential sequence preference during blunt-ended ligations**

All unique sequences (8734 of 8874, 98.4%) within the 6-mer library generated using equimolar mixtures of the 20 donors (Figure S2B) were analysed for the identity of nucleotides immediately adjacent to the six ligation points. The observed frequencies (upper value) are noted against the expected frequencies (lower value; rounded to the nearest integer) for the 3′–5′ ligation interfaces. Expected frequencies are calculated from the identities of the 20 selected codons (Supplementary Table 3). The χ^2 statistic with one degree of freedom was calculated for the 16 possible combinations of 3′–5′ preference at the five ligation junctions. Values in bold italic are those that have a statistically significant difference between observed and expected frequencies, where P < 0.0001.

	5′ position at junction (phosphate donor)			
3′ position at junction	C	G	T	A
C	**57**	116	94	88
	109	109	109	109
G	**511**	**1199**	**642**	**643**
	873	**873**	**873**	**873**
T	**445**	**1322**	829	**391**
	764	**764**	764	**764**
A	**517**	**821**	462	**599**
	437	**437**	437	**437**

Table S2 | **Further library analyses**

Uncorrected MiSeq data from the libraries illustrated in Figure 2 of the main paper were analysed. Correct length refers to sequences that have both correct flanking sequences and the correct insert. Errors and diversity are expressed as percentages of sequences that are the correct length. Error percentages represent an average across all randomized positions.

		6-Mer peptide library	**CDR3 library (11 codons)**
Length	Correct length (%)	79.9	74.1
	Single base additions/deletions (%)	8.0	7.4
	Other incorrect length (%)	12.0	18.5
Errors	Specified codons (%)	99.5	99.3
	Non-specified codons (%)	0.4	0.7
	Termination codons (%)	0.0	0.1
Diversity	Unique sequences (one occurrence) (%)	99.4	99.9
	Two occurrences (%)	0.6	0.1
	Three or more occurrences (%)	0.0	0.0
	Total library sequences	286684	461274

Table S3 | **Oligonucleotide sequences**

'MAX' within an oligonucleotide indicates a mixture of independently synthesized oligonucleotides, each containing one of the 'MAX' codons [alanine = GCT; cysteine = TGC; aspartate = GAC (right-hand) or GAT (left-hand); glutamate = GAA; phenylalanine = TTC; glycine = GGC; histidine = CAT; isoleucine = ATC; lysine = AAG; leucine = CTG; methionine = ATG; asparagine = AAC; proline = CCG; glutamine = CAG; arginine = CGT; serine = TCT; threonine = ACC; valine = GTG; tryptophan = TGG; tyrosine = TAC), whereas 'MAX$_{(RC)}$' represents the reverse complement of each MAX codon. Rev$_{RH}$ may be used as a minimalist reverse complement to the conserved part of all MAX$_{RH}$ primers (Figure 1 of the main paper and Figure S1) or, alternatively, fully complementary or hairpin oligonucleotides may be substituted in which a (T)$_4$ sequence joins the two strands (as demonstrated above, in the MAX$_{LH1–3}$ sequences). Acceptor sequences are fully user-defined. Those employed in Figure S1 (test acceptor) and the CDR3 randomization experiment are listed. The RH acceptor consists of a mixture of three oligonucleotides in which codons 1, 2 and 3 are TTT, ATC and CAT; TTT, TAC and CCT; and ATG, GTC and CAC respectively. All acceptor sequences were hybridized to fully complementary oligonucleotides before use. 6-Mer peptide randomizations employed a pUC19 amplicon (primers 5′-phosphateGTAAGATCCTTGAGAGTTTTCGCC-3′ and 5′-CGGTTAGCTCCTTCGGTCCTC -3′) as the acceptor. Underlined regions correspond to PCR primer sequences (where dotted underlines signify the reverse complement as the primer).

Oligonucleotide	**Sequence**
MAX$_{RH1}$	5′- GTGCTACGATGTCATTGCGAGTCACGTAMAX-3′
MAX$_{RH2}$	5′- AGGTAGATCAGTGACACGGAGTCACGTAMAX-3′
MAX$_{RH3}$	5′- ACAGAACAGGCACTTAGGGAGTCACGTAMAX-3′
MAX$_{RH4}$	5′-GATCTCACAGTCAGATGGGAGTCACGTAMAX-3′
Rev$_{RH}$	5′-MAX$_{(RC)}$TACGTGACTC-3′
MAX$_{LH1}$	5′-MAXTACGTGACTCCCATCTGACTGTGAGATCTTTTGATCTCACAGTCAGATGGGAGTCACGTAMAX$_{(RC)}$-3′
MAX$_{LH2}$	5′-MAXTACGTGACTCCGAGATGCTACTGTCGAATTTTTTCGACAGTAGCATCTCGGAGTCACGTAMAX$_{(RC)}$-3′
MAX$_{LH3}$	5′-MAXTACGTGACTCGCTCAAGTTACGGTCGATTTTTATCGACCGTAACTTGAGCGAGTCACGTAMAX$_{(RC)}$-3′
Test acceptor	5′-phosphate-ATCCGATCCTAGTGTATTCTTCGCTGTTCTCTGACTGAT-3′
RH acceptor	5′-phosphate$_{Codon1}$GAC$_{Codon2}$TGGGGCCAGGGCAC$_{Codon3}$GGTCACCGTCTCCTCAGGTGGAGGCGGTTCAGG CGGAGGTGGCAGCGGCGGTGGCGGATCGCAGTCTGTGCTGATTGAGCCTGCCTCCGTGTCTGCACTTCG-3′
LH acceptor	5′-CGAAGACCGACAGGGGTAAGACCATCAGTAGTAGGAAAGGTCTCGGCCGTTTCACCATCTCCCGTGACA ATGCCAAGAACTCACTGTATCTGCAAATGAACAGCCTGCGTGCCGAGGACACGGCCGTATATTACTGTGCGAGA -3′

Received 26 June 2013
doi:10.1042/BST20130123