Beyond the Natural Proteome: Nondegenerate Saturation Mutagenesis - Methodologies and Advantages

M.M. Ferreira Amaral*, L. Frigotto†, A.V. Hine*,1

*School of Life and Health Sciences, Aston University, Aston Triangle, Birmingham, B4 7ET, UK.

†Isogenica Ltd., The Mansion, Chesterford Research Park, Great Chesterford, Essex, CB10 1XL, UK.

1Corresponding author, email a.v.hine@aston.ac.uk

CONTENTS

ABSTRACT

Beyond the natural proteome, high-throughput mutagenesis offers the protein engineer an opportunity to “tweak” the wild-type activity of a protein to create a recombinant protein with required attributes. Of the various approaches available, saturation mutagenesis is one of the core techniques employed by protein engineers and in recent times, nondegenerate saturation mutagenesis is emerging as the approach of choice. This review compares the current methodologies available for conducting nondegenerate saturation mutagenesis with traditional, degenerate saturation and briefly outlines the options available for screening the resulting libraries, to discover a novel protein with the required activity and/or specificity.

KEYWORDS

Gene shuffling, molecular evolution, directed evolution, synthetic biology, saturation, mutagenesis, nondegenerate, diversity; protein engineering; antibody engineering

1. INTRODUCTION

Natural proteomes provide a plethora of proteins that scientists employ in widespread applications. But despite nature's bounty, there are many occasions where the natural proteins don't quite have the activity needed. Perhaps an enzyme's affinity for a key substrate is poor, or that enzyme lacks the necessary stability. Maybe a transcription factor that binds a particular DNA sequence is needed and the required specificity cannot be found naturally. More frequently, the CDR region of an antibody or antibody fragment will need to be changed to create a key therapeutic. This is where protein engineering comes to the fore, to create those novel proteins. Protein engineering itself relies on two key components: mutagenesis of the encoding gene and screening of the resulting proteins created.

The earliest methods of mutagenesis were conceptually crude. The entire genome of an organism was targeted with a mutagen (chemical or physical), high kill rates were accepted and any surviving organisms were screened for a phenotype of interest. Notwithstanding the lack of knowledge regarding causative mutation(s), the study of such mutants led to information that was key to elucidating many biochemical pathways. However, by the late 1970's alternative, more refined approaches had begun to emerge. For example, Shortle and Nathans (1978) were able to target such random mutagenesis to short, specific regions of DNA within a viral chromosome. Around this time, the first reports of successful site-directed mutagenesis were also published. Michael Smith and co-workers had already established the importance of comparing mutant with wild type sequences in localizing genetic function (Smith, Brown, Air, Barrell, Coulson, Hutchison & Sanger, 1977) and in 1978, Smith and co-workers published the first example of site-directed mutagenesis mediated by oligonucleotides (Hutchison, Phillips, Edgell, Gillam, Jahnke & Smith, 1978). These two approaches: site-directed mutagenesis and targeted random mutagenesis would serve as forerunners for the high-throughput mutagenesis strategies of today.

Site directed mutagenesis developed rapidly from 1978 onwards, with protocols to improve mutagenesis efficiency such as Kunkel mutagenesis (using *Escherichia coli dut*$^{-}$ *ung*$^{-}$ strains)

coming to the fore in the 1980's (Kunkel, 1985). However, even such "rapid" protocols were relatively lengthy and by the late 1980's favor had returned to simple primer-extension-based methods, as originally described by Michael Smith and co-workers. Smith shared the Nobel Prize for Chemistry in 1993 "For his fundamental contributions to the establishment of oligonucleotide-based, site-directed mutagenesis and its development for protein studies", with Kary B. Mullis, the inventor of PCR (Smith, 1994). Site directed mutagenesis and PCR were subsequently combined to create megaprimer mutagenesis and its various modifications (Ke and Madison, 1997 and references therein). Megaprimer mutagenesis is a rapid process that generates a PCR amplicon for incorporation into the original gene via cassette mutagenesis (Wells, Vasser & Powers, 1985). Meanwhile, a commercial approach, the Stratagene QuikChange® Site-Directed Mutagenesis Kit (Agilent Technologies, cat # 200518) and its improvements (Liu & Naismith, 2008) were yet faster since they require no further cloning steps.

Thus, by the early 1990s, introduction of a single point mutation at a single location was a routine and rapid laboratory procedure. The next major advancement would be to introduce multiple mutations in one step, to generate a whole library of variations based on a single gene. Now, methodologies diversified, depending on the knowledge of the original gene and the likely number of mutations required. To introduce random mutations at random locations within an amplicon, error-prone PCR was first employed (Leung, Chen, & Goeddel, 1989; Cadwell, & Joyce, 1991). Thereafter, error-prone PCR was itself employed within gene shuffling (Stemmer, 1994; Crameri, Raillard, Bermudez, & Stemmer, 1998), which effectively "breeds" homologous genes together on a laboratory timescale. Such techniques are immensely powerful, albeit that by their nature, they can never cover all of the theoretical sequence space. Nonetheless, gene shuffling offers an elegant solution when engineering a protein to create a desired activity, particularly when predicting the necessary locations of mutations would be impossible (eg. Campbell, Tour, Palmer, Steinbach, Baird, Zacharias & Tsien, 2002). By contrast, saturation mutagenesis offers changes to protein residues at one

or more defined locations simultaneously, also producing diverse variants but potentially within the theoretical sequence space, depending on the number of codons targeted and the methodology employed. In effect, saturation mutagenesis is conceptually a simple extension of site-directed mutagenesis, but in a high-throughput format. Instead of a single location, saturation mutagenesis targets multiple codons, making multiple substitutions in each, in a single experiment. Unsurprisingly therefore, saturation mutagenesis is a key tool in a protein engineer's arsenal. However, the techniques used to generate such mutations can be far from straightforward and are the subject of the current review.

2. ADVANTAGES OF NONDEGENERATE SATURATION

By its very nature, the genetic code is degenerate. A codon (3 bases) encodes a single amino acid. Each position within that codon can be any one of the four bases A, C, G and T. Thus, there are a total of 64 codons (4^3 combinations of the four bases, where order is important). Of those 64 codons, 3 are termination codons, leaving the remaining 61 codons to encode the twenty naturally-occurring amino acids. However, as any student of biology knows, those 61 codons are not distributed evenly. Specifically, some amino acids are encoded by 6 codons (Leu, Ser, Arg), some by 4 codons (Ala, Gly, Pro, Thr, Val), one by 3 codons (Ile), several by 2 codons (Cys, Asp, Glu, Phe, His, Lys, Asn, Gln, Tyr) whilst two amino acids are encoded by just 1 codon each (Met, Trp). Thus a conventional degenerate saturation codon (NNN, where N = a mixture of A, C, G & T and NNN a mixture that collectively contains all four bases at each position of the codon) is actually a mixture of all 64 possible sequences, that necessarily encodes the various amino acids disproportionately. More degenerate codons equates to yet more sequences / disproportionality (bias). Thus a piece of DNA containing three such codons would in practice be a mixture of 262144 (64^3) different DNA sequences.

Aside from the huge numbers involved, the disproportionality / bias encoded by conventional saturation has major impact when screening the encoded library for the “best” protein(s), particularly where ligand-based screening technologies are involved (see section 4.1). Thus scientists have attempted to reduce both sheer numbers and concomitantly the encoded bias either by reducing redundancy or else by eliminating it altogether.

Initial approaches to reduce redundancy involved using simple *limited* codon redundancy such as NNK or NNS saturation codons (K=T/G, S=G/C) and indeed, this became the “norm” for many years. NNK / NNS reduces from 64 to 32 codons to encode the 20 amino acids (+1 termination codon), so reducing numbers and decreasing the bias but not yet eliminating it. More recently, the 22c-trick has been described, which reduces the number of codons to just 22 per saturated position (Kille, Acevedo-Rocha, Parra, Zhang, Opperman,

Reetz & Acevedo, 2013), resulting in 2 codons for Val and Leu, 1 codon for all other amino acids and 0 termination codons. The consequences of these methods in comparison with fully nondegenerate saturation (exactly 20 codons encoding 20 amino acids) are compared in Figure 1.

As can be seen from Figure 1, conventional methods of degenerate saturation are largely inadequate. Diversity is the gold-standard when creating gene libraries. Within the context of protein engineering, diversity is a measure of the percentage of unique species within a library. Fig 1a demonstrates that diversity is poor even when just one codon is saturated via conventional degenerate codons and even drops off rapidly with the 22c-trick. By contrast, it is maintained at a theoretical level of 100% via nondegenerate approaches, no matter how many codons are saturated. Bias is an alternative way of addressing this problem. Fig 1b explains why diversity is so poor in conventionally-constructed saturation libraries. Whilst the numbers in Fig 1b reflect a worst-case scenario (i.e. the ratio of the most common codon combinations to the rarest codon combinations), it is clear that there is no equality of representation between different gene sequences, with some being very populous in comparison with others. Finally, only nondegenerate methods and the 22c trick prevent encoded truncation with the gene library (Fig 1c). Truncation can be a problem because truncated, non-functional proteins can be prone to aggregation, which leads to protein precipitation.

By contrast, non-degenerate methods allow the user to include all 20 codons in approximately equal ratio. Some nondegenerate methodologies go still further and permit the user to choose exactly which amino acids are (and are *not*) encoded at a specified codon and further still, in which relative proportions (either equal or alternative, defined ratios) - see section 3.2. In these ways, along with the removal of termination codons, nondegenerate methods permit the size of the library to be minimized whilst concomitantly maximizing the number of encoded variants for screening. The combination of those attributes has very positive ramifications since modern, directed evolution puts pressure on

creating high-quality libraries with reduced number of variants to boost the efficiency of screening experiments (Tang, Gao, Zhu, Wang, Zhou, & Jiang, 2012) – see section 4.

3. SATURATION MUTAGENESIS METHODS

3.1 Core techniques in saturation mutagenesis

Whether saturation mutagenesis is degenerate, near nondegenerate (the 22c-trick, Kille et al., 2013) or wholly nondegenerate, all methodologies rely on the same core techniques, as described in this section.

3.1.1 Cassette mutagenesis

A DNA “cassette” is simply a double-stranded piece of DNA that can be spliced into a gene of interest, so replacing the original gene sequence. Splicing usually occurs via conveniently-located restriction sites that are present in the original gene and are incorporated at either end of the cassette. In its first iteration, saturation mutagenesis was achieved via cassette mutagenesis using synthetic oligonucleotide cassettes containing degenerate bases at defined locations (Wells et al., 1985). DNA cassettes may vary in length from tens to hundreds of base pairs and can be simple synthetic DNA (containing degenerate bases as required), PCR amplicons (made from primers with degenerate bases as required) or be created by nondegenerate methodologies such as MAX randomization (section 3.2.2), ProxiMAX randomization (section 3.2.3) or Slonomics® (section 3.2.4).

3.1.2 Simple primer-extension mutagenesis

The original method of site-directed mutagenesis was based on primer extension (Hutchison et al., 1978; Smith, 1994). In essence, a primer containing a centrally-located mutation was annealed to a single-standed circular DNA template and extended around the template to create a heteroduplex. Because the parental strand of wild type DNA remained intact, the efficiency of this mutagenesis was low. However, the methodology has been developed to high efficiency over the decades. One of the most favored strategies is now QuikChange®

Mutagenesis (Agilent Technologies). Here, double-stranded plasmid template is amplified by PCR with a pair of complementary primers that each contain the required mutation(s) in a central location. After extension around each strand of the entire plasmid, the parental DNA is destroyed by *Dpn*I digestion (requires Dam methylation) and the two mutated strands are annealed together (to create a plasmid with two staggered nicks) and transformed into bacteria without further modification. In terms of saturation mutagenesis, the pair of primers need not contain a single mutation, but may instead contain a degenerate codon or indeed take the form of a DNA cassette as described in section 3.1.1. To expand the technology further, the QuikChange® Multi Site-Directed Mutagenesis Kit has been developed to target up to five sites simultaneously (Hogrefe, Cline, Youngblood & Allen, 2002). Target residues have to be at least 5 codons apart and so automatically exclude targeting contiguous sites. Although quick and simple, QuikChange® has many limitations for example, complementary primers are prone to self-priming or give low efficiency with more than one mutated position (Liu & Naismith, 2008). The protocol of QuikChange® has been the subject of modifications by many research laboratories.

3.1.3 Overlap extension mutagenesis

The use of overlap extension PCR in mutagenesis was first described in 1988 (Higuchi, Krummell & Saiki, 1988). In essence, PCR fragments (created from primers that contain mutations) which have complementary 3' regions can be used to prime each other and so join those fragments together. This technique is used extensively in both degenerate and nondegenerate saturation mutagenesis, with fragments created as originally described by Higuchi et al (in which the overlaps contain the mutations), or for joining cassettes together in which the mutations are contained internally.

3.2 Nondegenerate saturation methodologies

3.2.1. TRIM technology: trinucleotide phosphoramidites

The original approach to nondegenerate saturation mutagenesis, TRIM technology employs trinucleotide phosphoramidites (Virnekäs, Ge, Plückthun, Schneider, Wellnhofer & Moroney, 1994). Essentially, rather than adding one base at a time as in oligonucleotide DNA synthesis, three bases are added to a growing oligonucleotide in a single reaction. Thus to saturate a given codon, a pre-defined mixture of trinucleotide phosphoramidites is added to the growing DNA strand. However, single bases of DNA have differing coupling efficiencies during oligonucleotide synthesis meaning that a biased mixture of phosphoramidites is required to effect equimolar addition (Ho, Britton, Stone, Behrens, Leffet, Hobbs, et al., 1996) and this effect is amplified when coupling trinucleotides. To mitigate the problem, it is probably wise to use a commercial source of DNA produced using TRIM technology, such as Gene Art™ (Thermofisher) rather than undertaking synthesis with trinucleotide phosphoramidites in one's own laboratory. The resulting DNA can be used in any saturation protocol of choice.

3.2.2 MAX randomization

MAX randomization (Hughes, Nagel, Santos, Sutherland & Hine, 2003) was one of the first techniques to be published that achieves nondegenerate saturation without the need for any specialized chemicals, reagents or equipment. In essence, MAX randomisation relies on "selectional hybridisation" in which a series of short selection oligonucleotides hybridise with a complementary template oligonucleotide that is conventionally-saturated at the relevant codons. After the selection oligonucleotides have been ligated, asymmetric PCR ensures that only the selection strand is amplified (Figure 2). Thus MAX randomization generates a cassette which is typically used in cassette mutagenesis, but could be employed as a double-stranded primer in QuikChange® mutagenesis, or be joined with other sequences via

overlap PCR. Note that although MAX randomization can be used to mutate multiple codons, only a maximum of two contiguous (neighbouring) codons can be saturated, owing to the addressing function of the conserved part of the selection oligonucleotides. Again, MAX randomization yields a double-stranded DNA cassette that can be used in any mutagenesis protocol of choice.

3.2.3 ProxiMAX randomization

Like other nondegenerate techniques, ProxiMAX is a nondegenerate saturation technology that uses one codon only per amino acid (Ashraf, Frigotto, Smith, Patel, Hughes, Poole et al., 2013). In common with MAX randomization, ProxiMAX does not require any specialized reagents, but rather relies on conventional oligonucleotides, a Type IIS restriction enzyme and blunt-ended ligation. The process involves cycles of ligation, amplification and digestion with the consequence that one codon is added to the end of a growing DNA fragment, per cycle (Figure 3). However, that “codon” is really a user-defined mixture of up to 20 individual, double-stranded oligonucleotides. These oligonucleotides are largely-conserved sequences that can be fully or partially double-stranded, or be self-complementary hairpins, but each terminates with a unique triplet encoding just one amino acid (typically chosen for *max*imal expression in the organism of choice – hence the names MAX/ProxiMAX randomization). In practice, the mixture is often selected to comprise less than 20 such MAX codons, for example omitting codons for cysteine and methionine or selecting just the polar residues. As such, ProxiMAX randomization allows the user to define exactly which amino acids are encoded at each individual saturated position, which can either be contiguous or separated by regions of conserved sequence. The relative proportions of codons at each saturated position may also be user-defined. ProxiMAX can be achieved manually with good results (Poole, 2015) or via automation, which gives excellent compliance with library design (Ashraf et al., 2013). Automation also permits the addition of hexameric, two-codon units rather than one codon per cycle, though this modification requires automation, owing to the sheer

number (400) of oligonucleotides (codon donors) involved (Frigotto, Smith, Brankin, Sedani, Cooper, Kanwar et al., 2015). The use of two codons or hexamer nucleotides, rather than one per cycle, boosts efficiency and enhances performance of high-throughput mutagenesis reducing synthesis time. The commercial development of ProxiMAX, Colibra™ (Isogenica Ltd.) involves comprehensive quality control tests, via next generation sequencing, which allow for careful monitoring of library fragments manufacture. ProxiMAX can be used to saturate multiple contiguous codons and the resulting DNA cassettes are typically linked together by ligation or used in cassette mutagenesis, but could also be used in overlap extension PCR or as primers in QuikChange® mutagenesis.

3.2.4 Slonomics® / SlonoMax™

Originally published as an automated gene synthesis technology (Van den Brulle, Fischer, Langmann, Horn, Waldmann, Arnold et al., 2008; Schatz, O'Connell, Schwer & Waldmann, 2010), Slonomics® is also a cycle-based process, though one that avoids amplification and involves sticky-ended ligation of hairpin oligonucleotide building blocks typically with three-base single-stranded overhangs, called splinkers and anchors. The process requires a total of 64 splinkers and 4096 anchors. Initially, a selected splinker is ligated to a selected anchor. The ligated product is immobilized via a biotin moiety contained within the anchor. Washing removes any unligated splinker and the resulting immobilized product is then digested with a Type IIS endonuclease that leaves a three-base, single-stranded overhang. The immobilized sequences are then discarded, so removing digested (and any unligated) anchor. Meanwhile the supernatant, containing the extended splinker, enters the next cycle. In essence, each cycle transfers six bases of DNA (a staggered three bases on each strand) from the anchor to the splinker (Figure 4). The process is repeated to generate up to six codons (18 bp) in what is termed an “elongation fragment” and several elongation fragments can be combined via digestion with two further Type IIS restriction enzymes that each leave unique sticky ends, followed by multiple ligations. Slonomics® has been adapted to make combinatorial

libraries by using mixtures of splinkers and anchors in a commercial process called SlonoMax™ (Waldmann, 2006; Waldmann, 2013).

3.2.5. DC-analyzer and MDC-analyzer

DC-analyzer stands for Degenerate Codon Analyzer and is a computational approach to designing a saturated gene library (Tang et al., 2012). Specifically, DC-analyzer is downloadable software that designs combinatorial degenerate primers. As a consequence of DC-analyzer's programming, termination codons, codon degeneracy and rare codons of *E. coli* are eliminated from the design. In practice, DC-analyzer selects one codon for each amino acid and outputs a series of mainly limited degeneracy primers that the user must synthesize and then employ in combination within a PCR reaction or choice, to perform saturation mutagenesis. As such, DC-analyzer eliminates bias and so increases the diversity of the encoded library. DC-analyzer is used to design "small-intelligent libraries", as an alternative to NNS randomization. Owing to the numbers of primers involved, DC-analyzer typically targets small numbers of codons (1 and 2 codons were demonstrated by Tang et al., 2012). To target higher numbers of codons, Multi-Site Degenerate Codon Analyzer (MDC-analyzer) was developed (Tang, Wang, Ru, Sun, Huang & Gao, 2014).

MDC-analyzer allows for randomization of more than two codons that may be contiguous (three contiguous codons were demonstrated), and rather than producing full saturation, aims to reduce library size by designing DNA sequences to encode selected subsets of amino acids at saturated positions, rather than all twenty amino acids. Using prediction programs in conjugation with mutagenesis technology gives a so-called 'rational random' approach. Once the required subsets of amino acids have been chosen, using applications such as ConSurf-HSSP (Glaser, Rosenberg, Kessel, Pupko & Ben-Tal, 2005) and HotSpot Wizard (Pavelka, Chovancova & Damborsky, 2009) which analyze the favourable properties of amino acids at key regions of a protein, MDC designs multiple degenerate oligonucleotides to encode the required amino acids without termination codons or *E. coli*

rare codons. By accepting that some non-required amino acids will also be encoded, MDC-Analyzer drastically reduces the number of oligonucleotides required to perform the mutagenesis, in comparison with DC-Analyzer (Tang et al., 2014).

3.3 Near nondegenerate saturation: the 22c-trick

Conceptually similar to DC-Analyzer, the 22c-trick (Kille et al, 2013) uses a combination of PCR primers that contain pre-defined, limited degeneracy at the selected codons. Specifically, saturated codons are generated during a PCR reaction via a mixture of three primers: one contains codon NDT (A/C/G/T; A/G/T; T) another VHG (A/C/G; A/C/T; G) and the final, codon TGG. As described in section 2, when used in combination, these primers collectively encode valine and leucine twice, the other 18 amino acids once each and no termination codons. Note that this process cannot saturate multiple contiguous codons. Moreover, optimization of annealing temperature is essential to achieve good saturation (Kille et al., 2013), since different primers will necessarily have differing annealing temperatures and diversity would be affected quite severely if higher numbers of codons were targeted using this methodology (Figure 1).

4. SCREENING THE LIBRARY: METHODS, LIBRARY SIZE AND THE IMPORTANCE OF DIVERSITY.

In extending the natural proteome via protein engineering, library production is only the first part of a complex process. Once the gene library has been expressed, the resulting protein library may contain from hundreds up to ten trillion (10^{13}) different components, depending on the library design and methodology used in construction. Without a suitable screening procedure such libraries would be useless – they can hardly be examined one component at a time. In fact, the majority of libraries fall somewhere in between the two size limits and thus screening several million to billion components is non-trivial. Libraries with sizes greater than 10^{10} cannot generally be screened *in vivo* since transformation efficiency will limit the yield of clones. Thus the largest libraries are usually screened by using *in vitro* methodologies. Meanwhile, the type of protein library is also important in selecting a screening methodology. Ligand-binding libraries such as antibodies, transcription factors etc. tend to be screened via biopanning approaches (immobilized ligand bound by solution-based protein – section 4.1). In contrast, unless seeking an essentially irreversible enzyme inhibitor, enzyme libraries tend to require a different approach, since binding an immobilized substrate (ligand) would normally be followed by processing and release of the substrate, so preventing the immobilisation of functional proteins that forms a key part of the biopanning process.

4.1 Methods for screening ligands

When biopanning a ligand-binding protein library, the ligand of choice is immobilized and the protein library added in solution. After binding, the majority of the library (non-binding protein) is washed away, bound proteins are eluted, these populations are amplified and the process is repeated, typically 4-5 times. Thereafter, individual species are isolated, sequenced and their characteristics investigated.

Thus biopanning is dependent on mass action, which in turn requires that the library components are all present in approximately equal concentrations for accurate discovery of the “best” proteins. Thus good diversity is essential. If all library components are unique, i.e.100% diversity, and expressed similarly (which is why rare codons should be avoided during saturation), all protein components will have approximately equal concentrations. In contrast, a library with low diversity will have many different genes encoding one protein and just one gene encoding another, which will lead to biased concentrations of proteins (see Figure 1) and thus delivery of the populous proteins as the best “hits”, regardless of their suitability.

The various options for biopanning strategies are described in this section.

4.1.1 Phage display

George Smith first proposed that filamentous phage can serve as carriers of foreign DNA and so provide a means to screen library variants such as those of antibodies (Scott and Smith, 1990; Smith, 1985).

This method of screening libraries relies on infecting bacterial cells with viruses called bacteriophage. Phage chromosomes have capacity to take in foreign DNA fragments and both replicate and express them from within host bacteria, usually *E. coli.* Since this foreign DNA is spliced into the phage coat protein genes, upon its expression the protein will be exposed on the outer surface of the phage particle. (Smith and Petrenko, 1997). Specifically, the library of variants is cloned into phage vectors which then transfect bacterial cells. One cell will carry one variant only and expressed proteins of that variant will be exposed on the surface of phage which are then screened by biopanning. After each round of selection, the eluted phage are amplified by passage through *E. coli*. In addition to antibody fragments, phage display has been successful in screening libraries to find insulin and IRF-1 receptor agonists and antagonists (Dedova, Fletcher, Liu, Wang, Blume & Brissette et al., 2004);

small peptides that mimic erythropoietin (Wrighton, Farrell, Chang, Kashyap, Barbone & Mulcahy et al., 1996; McConnell, Dinh, Le, Brown, Becherer & Blumeyer et al., 1998) and even an inhibitor for the enzyme pancreatic lipase (Lunder, Bratkovic, Kreft & Strukelj, 2005). Though useful for many applications, phage display is limited by transformation efficiency and cannot therefore be used to screen libraries of greater than 10^9 to 10^{10} components (Odegrip, Coomber, Eldridge, Hederer, Kuhlman & Ullman et al., 2004).

4.1.2 Ribosome display

Ribosome display is an *in vitro* method of selection invented by Plückthun (Hanes & Plückthun, 1997). It was the first *in vitro* selection method, inspired by work of Mattheakis and co-workers (Mattheakis, Bhatt & Dower, 1994) who demonstrated affinity selection using polysomes which enable the critical link between genotype and phenotype that is essential for any biopanning application. As an *in vitro* method ribosome display is not limited by transformation efficiency and can therefore be used to screen very large libraries.

The *E. coli* S30 system is a coupled transcription/translation system that forms a key part of ribosome display and was also used by Mattheakis and co-workers. They displayed a library of short, 13-mer opioid peptides on *E. coli* S30 followed by affinity screening against immobilized antibodies. Because the expressed proteins were coupled with mRNA, selected hits could be reverse transcribed into cDNA and then sequenced to determine the nucleotide sequence encoding the displayed peptide (Mattheakis et al., 1994).

Ribosome display also uses the *E. coli* S30 system to create whole, correctly-folded proteins that remain coupled with the S30 complex and mRNA (Hanes & Plückthun, 1997). In the first instance, PCR is employed to both amplify the library and couple it with a T7 promoter and ribosome-binding site. Following transcription to RNA, the *E. coli* S30 coupled transcription/translation system translates the mRNA *in vitro*. This system also includes various factors than enable correct folding of the translated proteins and stabilize produced

ribosome-mRNA-protein complexes. These complexes are then screened by biopanning. After each round of selection, isolated library members are eluted from the ribosome complex and the mRNA is reverse transcribed into cDNA ready for another round of biopanning or for sequencing of individual isolates. (Hanes & Plückthun, 1997).

Ribosome display is often used to improve protein affinity or stability and can be used in combination with other selection pressure mechanisms (Buchanan, Ferraro, Rust, Sridharan, Franks & Dean et al., 2012). The first exemplification of this cell-free system was used to screen single-chain fragments (scFv) of an antibody (Hanes & Plückthun, 1997). Subsequently, Minter and co-workers have used ribosome display in conjunction with three stability selective pressure factors for two therapeutic proteins that gave problems during the drug development phase (Buchanan et al., 2012). Specifically, the tendency of erythropoietin to aggregate was reduced and a 1000-fold improvement in the soluble expression of granulocyte colony-stimulating factor was gained.

In a conceptually similar approach to ribosome display, mRNA may be covalently bound to its encoded protein (Xiao, Bao & Zhao, 2014). Because this modified approach and ribosome display more generally are performed wholly *in vitro*, neither method is limited by transformation efficiency.

4.1.3 CIS display

CIS display is an *in vitro* screening technology that uses the RepA family of proteins of the R1 plasmid as key components. RepA is a bacterial replication initiator protein which has a special feature of high-fidelity cis-activity: it binds only to the DNA from which it was expressed. The first part of the method involves construction of DNA consisting of: N-terminal promoter, library, RepA gene (*repA*), CIS element and C-terminal *ori*. Once these elements are joined in that order, *in vitro* transcription begins at the promoter and ends when it reaches the CIS element. Simultaneous, *in vitro* translation produces RepA protein which

binds to its encoding DNA at the CIS region and *ori* sequence. As a result, the nascent polypeptide (from the library) is fused with RepA. Ultimately, through its cis activity, RepA protein provides the crucial physical linkage between genotype (DNA) and phenotype (protein; Figure 5). After each round of biopanning, DNA from the eluted complex is either amplified by PCR ready for the next round of biopanning, or sequenced to identify the interacting library component (Odegrip et al., 2004; Mathonet, Ioannou, Betley & Ullman, 2011).

CIS display has been used in antibody research (Odegrip et al., 2004), peptide maturation, ligand discovery including therapeutic peptides (Mathonet et al., 2011), and engineering of small WW scaffolds (small β-sheet motifs; Patel, Mathonet, Jaulent & Ullman, 2013). CIS display technology has also been used to identify 12-mer peptides resistant to thrombin, chymotrypsin and plasma proteases (Eldridge, Cooley, Odegrip, McGregor, FitzGerald & Ullman, 2009).

4.2 Methods for screening enzymes

Enzymes are often engineered to effect a change in substrate specificity, solvent tolerance or stability. Unless an irreversible enzyme inhibitor is sought (see section 4.1) there is much more variety in screening methodologies, since screening is necessarily based on phenotypic properties. Examples of two key strategies are given herein (sections 4.2.1. and 4.2.2).

4.2.1 Double selection: positive and negative

Functional assays are performed in *in vivo* systems and rely on production of phenotypes which are different to the native molecule and thus, recognized as mutant. Double selection is a method that has been used with different selection markers – one positive and one negative (Liu & Schultz, 1999; Pastrnak & Schultz, 2001). A well-established positive marker

is based on antibiotic resistance such as the β-lactamase gene while negative selection can be based on a toxic gene such as barnase. In further studies β-lactamase was replaced with chloramphenicol acetyl transferase (CAT), since chloramphenicol has proved to have stronger selection pressure than other antibiotics owing to its bacteriostatic nature (Pastrnak, Magliery & Schultz, 2000).

Schultz and co-workers use this approach to study the aminoacyl-tRNA synthetase (AARS) family of enzymes, specifically to engineer these enzymes to aminoacylate suppressor tRNA (specific to the amber stop codon) with various unnatural amino acids. During positive selection, variants of an AARS engineered in the amino acid binding pocket are selected based on aminoacylation of the suppressor tRNA either with a natural amino acid (AA) or the chosen unnatural amino acid (UAA). Thus, any enzyme that effectively aminoacylates the suppresssor tRNA (with any amino acid) will permit read-through of the amber stop codon so that antibiotic resistance can be expressed. Subsequently, during negative selection, the UAA is omitted. This time, read-through of the stop codon will permit expression of a toxic gene (barnase or uracil phosphoribosyltransferase). Negative selection is carried out in the absence of the UAA, so that any clones that express AARS enzymes that aminoacylate with native amino acids will be killed, so leaving behind enzymes that aminoacylate only with the required UAA (Melançon & Schultz, 2009).

4.2.2 FACS screening

Fluorescence-activated cell sorting (FACS) is a cell-display and activity-based selection screening procedure that employs flow cytometry. It is an ultrahigh-throughput technique, capable of screening up to 10^8 mutants per day (Yang & Withers, 2009). FACS is also characterized by high sensitivity. FACS is suitable for screening enzymatic activity where production of fluorescence is feasible. As with any another screening, FACS relies on linking the genotype with phenotype. For comparison, an example used in engineering AARSs is again described. Here, amber stop codons were engineered into genes for both T7 RNA

polymerase and GFP under the control of a T7 promoter. Using genes for an orthogonal AARS/suppressor tRNA pair (i.e the suppressor tRNA is not recognised by host AARSs and the orthogonal AARS does not recognise the host tRNA), the amino acid binding pocket of the orthogonal AARS was engineered to bind UAAs. In this instance, two positive selections (rather than a positive then a negative selection) are employed. The first again involves chloramphenicol resistance as described previously (section 4.2.1), whilst the second relies on UAA aminoacylation of the suppressor tRNA to read though both the T7 RNA polymerase and the GFP genes, to generate a fluorescent signal detectable by FACS (Santoro, Wang, Herberich, King & Schultz, 2002).

Alternative methods for FACS-based screening of gene/protein libraries have been reviewed recently (Xiao et al., 2014).

5. CONCLUSIONS

The transition from low- to high-throughput protein engineering has enabled the creation of novel proteins in a myriad of applications and saturation mutagenesis plays an invaluable role within that sphere. However, the protocol selected to perform that saturation depends on several factors including the design of the library and the importance placed on the quality versus the complexity of library synthesis, as follows:

i. how many codons are being targeted
ii. whether full or partial saturation is required
iii. whether or not controlled ratios of codons are required
iv. importance of the quality of the retrieved product, versus
v. the effort / expense involved in library synthesis

The conflicting and synergistic features of these properties are examined herein.

In the preceding sections of this review, it becomes fairly clear that degenerate saturation is of very limited or indeed *no* practical utility when saturating three or more codons in a library that is to be screened by biopanning (Figure 1), although its simplicity may be more attractive when screening enzyme libraries (section 4.2), where life/death rather than mass action (section 4.1) determines whether or not a protein is selected. In contrast, for small numbers of saturated codons (~1-3 codons), the near nondegenerate 22c-trick or the fully non-degenrate DC-Anaylzer / MDC-Analyzer used in conjunction with overlap PCR or QuikChange® mutagenesis (section 3.1) are attractive options owing to their simplicity of use and lack of relative expense. Indeed, the financial benefit of employing these methodologies in comparison with NNK/NNS has been examined recently (Acevedo-Rocha, Reetz & Nov, 2015). However, three saturated codons is the maximum exemplified for each of these techniques as described by the inventing authors (Kille et al., 2013; Tang et al., 2012; Tang et al., 2014), presumably because the number of primers becomes

unmanageable for higher numbers or saturated positions and also because those primers are specific to individual saturation experiments.

For higher numbers of saturated positions, a more complex strategy of nondegenerate saturation is required. ProxiMAX randomization has recently been used to target 24 consecutive codons (Frigotto et al., 2015), whilst the maximum number of residues that can be targeted by SlonoMAX™ has not been defined by the inventors (Van den Brulle et al., 2008), but is surely equal to that of ProxiMAX. Both of these techniques are expensive in terms of DNA synthesis, but the oligonucleotides involved may be consistently re-used, because everything except the saturated codons is removed by Type IIS restriction digestion, once the saturation protocol has been completed. Moreover, both offer the ability to define not only the precise residues encoded at each saturated position, but also the relative ratios of each codons at a saturated location, meaning that natural antibody libraries (for example) can be mimicked. In contrast, neither the 22c-trick (Kille et al., 2013) nor MDC-Analyzer (Tang et al., 2014) offer that possibility, even with low numbers of saturated positions. Meanwhile DC-Analyzer (Tang et al., 2012) does allow the user to specify which amino acids are encoded, but not their relative proportions, because of the defined-degenerate primer synthesis involved.

Thus in the end, the selected method of saturation will come down to a decision about where the funds and effort are to be spent. Up-front investment in library synthesis is economically advantageous even when considering small libraries (Acevedo-Rocha, 2015). This effect can only be amplified when screening large ($>10^9$ protein libraries). Here, even the expense of having a gene library synthesized commercially must pale into insignificance in comparison with screening costs of examining so many components via a biopanning strategy. More serious still is the possibility of taking forward a sub-optimal candidate protein that was identified because the library from which it was identified contained poor diversity. The relative expense of such a mistake is hard to calculate.

ACKNOWLEDGEMENTS

This work was supported by BBSRC Grant BB/L015633/1. The authors gratefully acknowledge Dr Andrew J Sutherland (Aston University) for critical reading of the manuscript.

REFERENCES

Ashraf, M., Frigotto, L., Smith, M.E., Patel, S., Hughes, M.D., Poole, A.J., Husam R.M., Hebaishi, H.R.M., Ullman, C.G. & Hine, A.V. (2013). ProxiMAX randomization: a new technology for nondegenerate saturation mutagenesis of contiguous codons. *Biochemical Society Transactions,* **41**, 1189–1194.

Acevedo-Rocha, C.G., Reetz, M.T & Nov, Y. (2015). Economical analysis of saturation mutagenesis experiments. *Scientific Reports* **5**, 10654.

Buchanan, A., Ferraro, F., Rust, S., Sridharan, S., Franks, R., Dean, G., McCourt, M., Jermutus, L. & Minter, R. (2012). Improved drug-like properties of therapeutic proteins by directed evolution. *Protein Engineering, Design & Selection*, **25**, 631–638.

Cadwell, R. & Joyce, G., (1991). Randomization of genes by PCR mutagenesis*. PCR Methods and Applications*, **2**, 28–33.

Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A. & Tsien, R.Y. (2002). A monomeric red fluorescent protein. *Proceedings of the National Academy of Sciences USA*, **99**, 7877–7882.

Crameri, A., Raillard, S.A., Bermudez, E. & Stemmer W.P.C. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, **391**(6664), 288-291.

Dedova, O., Fletcher, P., Liu, H., Wang, P., Blume, A., Brissette, R., Hsiao, K., Lennick, M., Pillutla, R. & Goldstein, N. (2004). US Patent No. US2004023887. Washington, DC: U.S. Patent and Trademark Office.

Eldridge, B., Cooley, R.N., Odegrip, R., McGregor, D.P., FitzGerald, K.J. & Ullman, C.G. (2009). An in vitro selection strategy for conferring protease resistance to ligand binding peptides. *Protein Engineering, Design & Selection*, **22**, 691-698.

Frigotto, L., Smith, M.E., Brankin, C., Sedani, A., Cooper, S.E., Kanwar, N., Evans, D., Svobodova, S., Baar, C., Glanville, J., Ullman, C.G. & Hine, A.V. (2015). Codon-precise, synthetic, antibody fragment libraries built using automated hexamer codon additions and validated through next generation sequencing. *Antibodies*, **4**, 88-102.

Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T. & Ben-Tal, N. (2005). The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins*, **58**, 610-617.

Hanes, J. & Plückthun, A. (1997). *In vitro* selection and evolution of functional proteins by using ribosome display. *Proceedings of the National Academy of Sciences USA*, **94**, 4937-4942.

Higuchi, R., Krummell, B. & Saiki, R.K. (1988). A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. *Nucleic Acids Res.* **16**, 7351-7367.

Ho, S.P., Britton, D.H.O., Stone, B.A., Behrens, D.L., Leffet, L.M., Hobbs, F.W., Miller, J.A. & Trainor, G.L., 1996. Potent antisense oligonucleotides to the human multidrug resistance-1 mRNA are rationally selected by mapping RNA-accessible sites with oligonucleotide libraries. *Nucleic Acids Res.* **24**, 1901–1907.

Hogrefe, H.H., Cline, J., Youngblood, G.L. & Allen, R.M. (2002). Creating randomized amino acid libraries with the QuikChange® multi site-directed mutagenesis kit. *Biotechniques,* **33**, 1158-1160.

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.

Hutchison, C.A. III, Phillips, S., Edgell, M.H., Gillam, S., Jahnke, P. & Smith, M. (1978). Mutagenesis at a specific position in a DNA sequence. *The Journal of Biological Chemistry*, **253**(18), 6551-60.

Ke, S.-H & Madison, E.L. (1997). Rapid and efficient site-directed mutagenesis by single-tube 'megaprimer' PCR method. *Nucleic Acids Research*, **25**, 3371–3372.

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

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

Kunkel, T.A. (1985) Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences USA*, 82, 488-492.

Leung, D.W., Chen, E. & Goeddel, D.V. (1989). A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique*, **1**, 11-15.

Liu, D.R. & Schultz, P.G. (1999). Progress toward the evolution of an organism with an expanded genetic code. *Proceedings of the National Academy of Sciences USA*, **96**, 4780-4785.

Liu, H. & Naismith, J.H. (2008). An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnologies*, **8**, 91.

Lunder, M., Bratkovic, T., Kreft, S. & Strukelj, B. (2005). Peptide inhibitor of pancreatic lipase selected by phage display using different elution strategies. *Journal of Lipid Research*, **46**, 1512–1516.

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

Mathonet, P., Ioannou, A., Betley, J. & Ullman, C. (2011). CIS display, a DNA-based in vitro selection technology for therapeutic peptides. *Chimica Oggi – Chemistry Today*, **29**, 10-12.

Mattheakis, L.C., Bhatt, R.R. & Dower, W.J. (1994). An *in vitro* polysome display system for identifying ligands from very large peptide libraries. *Proceedings of the National Academy of Sciences USA*, **91**, 9022-9026.

Melançon, C.E.III. & Schultz, P.G. (2009). One plasmid selection system for the rapid evolution of aminoacyl-tRNA synthetases. *Bioorganic & Medicinal Chemistry Letters*, **19**, 3845–3847.

McConnell, S.J., Dinh, T., Le, M.-H., Brown, S.J., Becherer, K., Blumeyer, K., Kautzer, C., Axelrod, F. & Spinella, D.G. (1998). Isolation of erythropoietin receptor agonist peptides using evolved phage libraries. *Biological Chemistry*, **379**, 1279-1286.

Odegrip, R., Coomber, D., Eldridge, B., Hederer, R., Kuhlman, P.A., Ullman, C., FitzGerald, K. & McGregor, D. (2004). CIS display: In vitro selection of peptides from libraries of protein–DNA complexes. *Proceedings of the National Academy of Sciences USA.*, 101, 2806-2810.

Patel, S., Mathonet, P., Jaulent, A.M. & Ullman, C.G. (2013). Selection of a high-affinity WW domain against the extracellular region of VEGF receptor isoform-2 from a combinatorial library using CIS display. *Protein Engineering, Design & Selection*, **26**, 307–315.

Pastrnak, M., Magliery, T.J. & Schultz, P.G. (2000). A new orthogonal suppressor trna/aminoacyl-tRNA synthetase pair for evolving an organism with an expanded genetic code. *Helvetica Chimica Acta*, **83**, 2277-2286.

Pastrnak, M. & Schultz, P.G. (2001). Phage selection for site-specific incorporation of unnatural amino acids into proteins in vivo. *Bioorganic & Medicinal Chemistry*, **9**, 2373–2379.

Pavelka, A., Chovancova, E. & Damborsky, J. (2009). HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Research*, 37, W376-W383.

Poole, A.J. The integration of ProxiMAX randomisation with CIS display for the production of novel peptides. (2015). PhD thesis, Aston University.

Santoro, S.W., Wang, L., Herberich, B., King, D.S. & Schultz, P.G. (2002). An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nature Biotechnology*, **20**, 1044-1048.

Schatz, O., O'Connell, T., Schwer, H. & Waldmann, T. (2010). US Patent No. US 7695906. Washington, DC: U.S. Patent and Trademark Office.

Scott, J.K. & Smith, G.P. (1990). Searching for peptide ligands with an epitope library. *Science*, **249**, 386-390.

Shortle, D. & Nathans, D. (1978). Local mutagenesis: a method for generating viral mutants with base substitutions in preselected regions of the viral genome. *Proceedings of the National Academy of Sciences USA*, **75**(5), 2170–2174.

Smith, M., Brown, N.L., Air, G.M., Barrell, B.G., Coulson, A.R., Hutchison, C.A. III & Sanger, F. (1977). DNA sequence at the C termini of the overlapping genes *A* and *B* in bacteriophage ΦX174. *Nature*, **265**, 702-705.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, **228**, 1315-1317.

Smith, M. (1994). Synthetic DNA and Biology. *Bioscience Reports*, 14(2), 52-66.

Smith, G.P. & Petrenko, V.A. (1997). Phage display. *Chemical Reviews*, **97**, 391-410.

Stemmer, W.P.C. (1994). Rapid evolution of a protein in vitro by DNA shuffling. *Nature*, 370(6488), 389-91.

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

Tang, L., Wang, X., Ru, B., Sun, H., Huang, J. & Gao, H. (2014). MDC-Analyzer: A novel degenerate primer design tool for the construction of intelligent mutagenesis libraries with contiguous sites. *BioTechniques,* **56**, 301-310.

Van den Brulle, J., Fischer, M., Langmann, T., Horn, G., Waldmann, T., Arnold, S., Fuhrmann, M., Schatz, O., O'Connell, T., O'Connell, D., Auckenthaler, A. & Schwer, H., (2008). A novel solid phase technology for high-throughput gene synthesis. *Biotechniques*, **45**, 340-343.

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

Waldmann, T. (2006, September 28). Sloning announces the development of SlonoMax™ gene variant libraries of previously unmatched quality – creating new possibilities for the directed evolution of proteins. Retrieved July 13, 2016, from http://www.businesswire.com/news/home/20060928005357/en/Sloning-Announces-Development-SlonoMax-TM-Gene-Variant

Waldmann, T. (2013, April 1). Achieving a new quality level in tailoring genetic diversity. Retrieved July 13, 2016, from http://www.scientistlive.com/content/19548

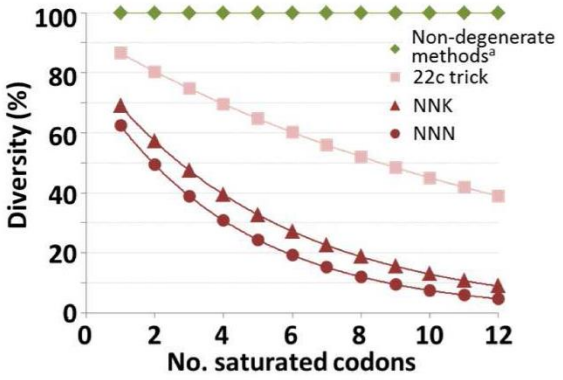
Wells, J.A., Vasser, M. & Powers, D.B. (1985). Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites. *Gene*, **34**, 315-323.

Wrighton, N.C., Farrell, F.X., Chang, R., Kashyap, A.K., Barbone, F.P., Mulcahy, L.S., Johnson, D.L., Barrett, R.W., Jolliffe, L.K & Dower, W.J. (1996). Small peptides as potent mimetics of the protein hormone erythropoietin. *Science*, **273**, 458-463.

Xiao, H., Bao, Z. & Zhao, H. (2014). High throughput screening and selection methods for directed enzyme evolution. *Industrial & Engineering Chemistry Research*, **54**, 4011-4020.

Yang, G. & Withers, S.G. (2009). Ultrahigh-throughput FACS-based screening for directed enzyme evolution. *ChemBioChem*, **10**, 2704-2715.

a) Diversity



b) Encoded bias

No. saturated codons	Ratio most common : rarest codon combinations[b] NNN	NNK / NNS	22c-trick	Non-degenerate methods[a]
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%		
6	25%	17%		
9	35%	25%		0%
12	44%	32%		

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 nondegenerate methods[a] may be created via various methodologies as described in section 3. **a)** Diversity was calculated using the formula $d=1/(N\sum_k p_k^2)$ (Makowski & Soares, 2003) and is in agreement for a 12-mer peptide saturated with codon NNN (Krumpe, Schumacher, McMahon, Makowski. & Mori, 2007). **b)** 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). **c)** Truncation is calculated as the percentage of sequences that contain 1 or more termination codons within the saturated region. *Reproduced in part from Ashraf, M., Frigotto, L., Smith, M.E., Patel, S., Hughes, M.D., Poole, A.J., et al. (2013). ProxiMAX randomization: a new technology for nondegenerate saturation mutagenesis of contiguous codons.* Biochemical Society Transactions, *41, 1189–1194 under the Creative Commons Attribution License (CC-BY).*

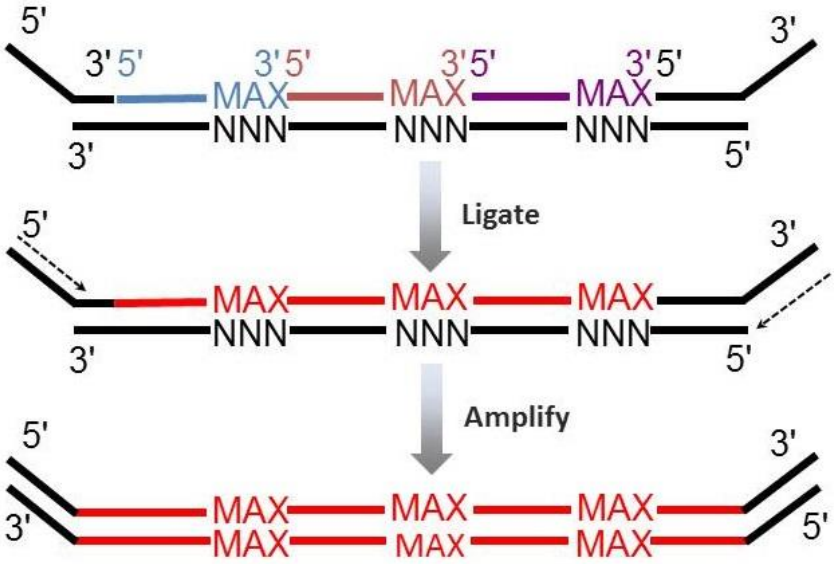


Figure 2: Overview of the MAX Randomization technique (Hughes et al., 2003)
A single template oligonucleotide is synthesized that is fully-degenerate at the designated, saturated codons. Meanwhile, a set of up to 20 small selection oligonucleotides are synthesised individually, for each saturated position. 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, where a MAX codon is the favoured codon for a single amino acid in the organism of interest. 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 randomization cassette.

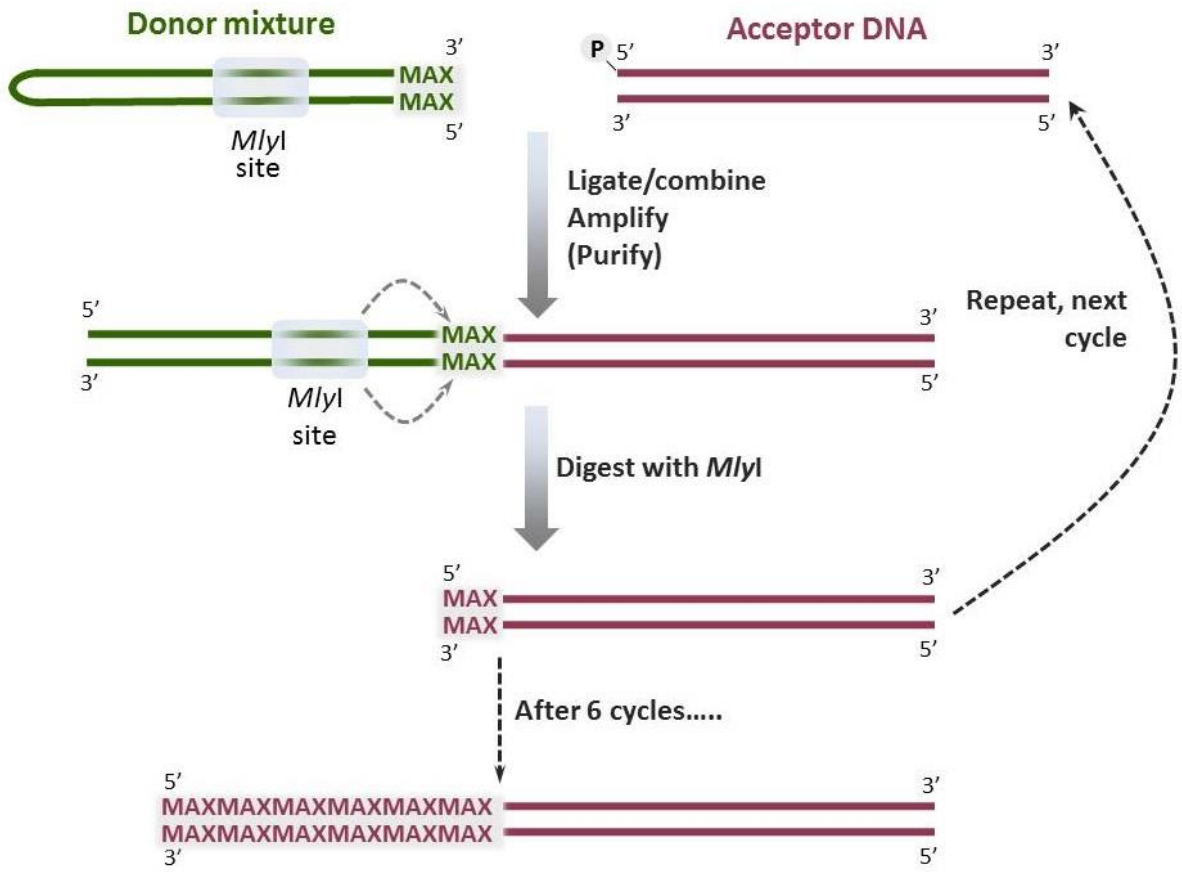


Figure 3: Overview of the ProxiMAX randomisation process.

Double-stranded DNA donors, carrying the required 'MAX' codons at their termini, are ligated 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, fully double-stranded DNA or hairpin oligonucleotides (as shown). After ligation, the products are amplified by PCR. Depending on whether the process is performed with automation or manually, the donor oligonucleotides can either be combined before or after ligation, with the automated process substantially reducing the number of steps involved and permitting the use of hexanucleotide donors as required (Frigotto et al., 2015). The amplified, purified product is then digested with *Mly*I and the process repeated, using the digestion product from round 1 as the acceptor for the next round of ligation. Different sets of donor oligonucleotides (up to 20 independently-synthesized, double-stranded oligonucleotides) are cycled to prevent potential carry-over from one round of addition to the next. Adapted from *Adapted from Ashraf, M., Frigotto, L., Smith, M.E., Patel, S., Hughes, M.D., Poole, A.J., et al. (2013). ProxiMAX randomization: a new technology for nondegenerate saturation mutagenesis of contiguous codons.* Biochemical Society Transactions, *41, 1189–1194.*

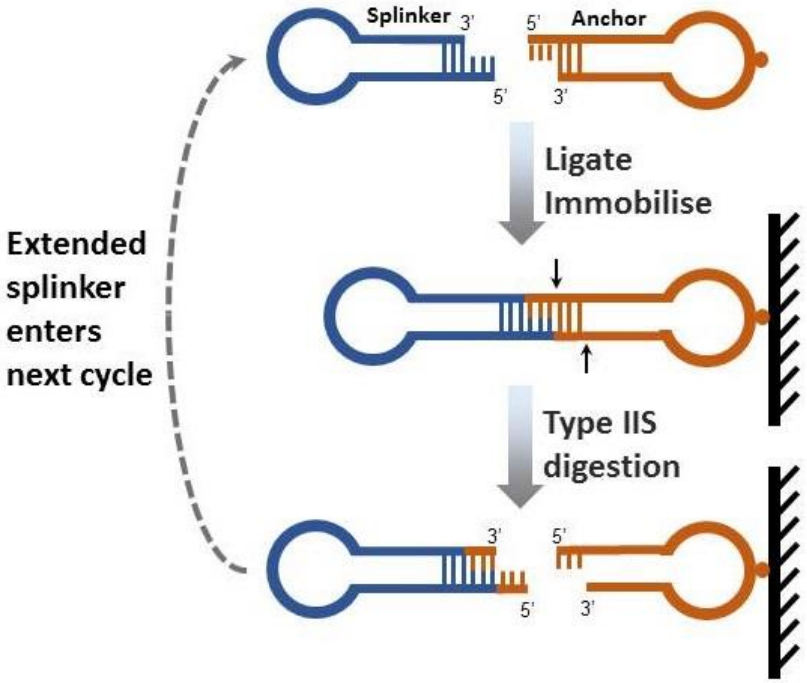


Figure 4: Overview of the Slonomics® process (Van den Brulle et al., 2008; Waldmann 2006; Waldmann, 2013).

Hairpin splinker oligonucleotides are joined, via sticky-ended ligation, to a mixture of hairpin anchor oligonucleotides. The ligated product is then immobilised via a biotin moiety contained within the anchors, washed and digested with Type IIS restriction enzyme *Eam*1104I, which generates a three-base sticky ended overhang. The resulting extended splinker, which is now free in solution, enters the next round of addition. Up to six rounds of addition are performed to generate an “elongation block” and several elongation blocks may be joined together via further Type IIS digestion and subsequent ligation (Van den Brulle et al., 2008). No PCR amplification is involved in the process.

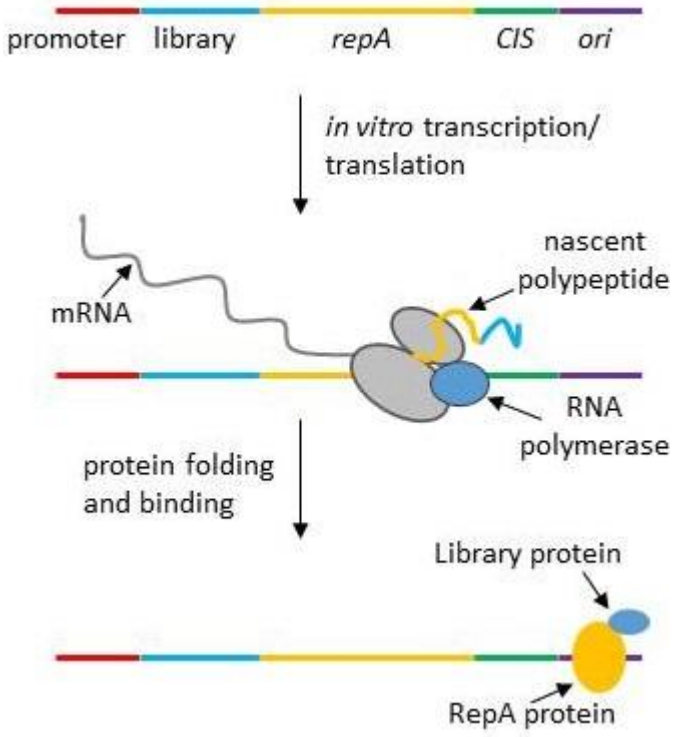


Figure 5: Overview of CIS display.

Adapted from Odegrip et al., 2004. Double-standed DNA is generated consisting of (in 5' ➔ 3' order) a promoter, the saturated library fused in-frame to *repA*, the *CIS* element and the *ori* sequence. The DNA is then subject to coupled *in vitro* transcription/translation using an *E. coli* S-30 extract. Transcription pauses when RNA polymerase reaches the *CIS* element. Meanwhile translation of the newly-produced mRNA produces a library component/RepA fusion protein that binds to the *ori* sequence, so linking the protein to its encoding gene. The resulting complex is screened via biopanning. *Adapted from Odegrip, R., Coomber, D., Eldridge, B., Hederer, R., Kuhlman, P.A., Ullman, C. et al. (2004). CIS display: In vitro selection of peptides from libraries of protein–DNA complexes.* Proceedings of the National Academy of Sciences USA.*, 101, 2806-2810.*