Synthetic gene libraries: in search of the optimal diversity

Marc Ostermeier

Dept of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218, USA

Directed evolution has proven to be an effective method for evolving proteins with desired properties. A key step is the creation of suitably diverse gene libraries. Two new methods for creating such libraries make sole use of synthesized oligonucleotides and allow researchers to tailor the diversity of a library with greater precision and create libraries with greater diversity than was previously possible. Such increased diversity appears to accelerate directed evolution.

The DNA shuffling of variants of a single gene [1] or a family of genes [2] creates a library of homologously recombined genes. Combined with an efficacious screening or selection strategy, this *in vitro* molecular evolution strategy is an efficient (perhaps the most efficient) method of creating proteins with improved or novel properties. Notwithstanding the continued success of the original fragmentation-based DNA shuffling method [1], in which genes are fragmented using DNase I and reassembled using a primerless PCR reaction, limitations of the method have inspired several similar strategies [3].

The main limitations are a lack of 'high-resolution' crossovers (i.e. sites of recombination very near each other) and a dependence on homology for crossover generation. The former makes shuffling of short genes problematic and results in libraries in which polymorphisms near each other are linked [4]. The dependence on homology results in a bias towards crossovers in regions of highest homology, resulting in exchanges of blocks of sequences and impeded shuffling of genes with low homology. Although these issues have been addressed by other methods [5,6], no single method satisfies all concerns. Two improved recombination methods involving the shuffling of synthetic oligonucleotides partially address both limitations, particularly the lack of high-resolution crossovers [7,8]. These intriguing studies tentatively support previous findings [2,9,10] that show that increased diversity in a library accelerates directed evolution, a notion that is somewhat counter-intuitive considering that proteins are such highly coupled systems.

Synthetic shuffling

Previously, Ness and co-workers demonstrated the shuffling of 26 subtilisin genes using fragmentation-based DNA shuffling [11]. Recently, they revisited the shuffling of subtilisin using a new strategy called synthetic shuffling (Fig. 1) and compared it with fragmentation-based shuffling [7]. A series of degenerate oligonucleotides were designed to encode all polymorphisms in a central region of 15 subtilisin genes, to maximize sequence identity and to conform to the *Bacilus subtilis* codon usage table. The oligonucleotides were designed not to reflect the parental diversity, but to increase the frequency of rare polymorphisms. This overrepresentation of rare polymorphisms increases the diversity of the library by increasing the probability of finding rare combinations of amino acids present in only one or a few parents.

The 30 oligonucleotides were assembled using a primerless PCR gene-assembly method [12] and the resulting library was more diverse than the fragmentation-based library. The authors tested ~1500 active library members of the synthetic and fragmentation-shuffled libraries for activity and compared activity at pH 10 with that at pH 7, and the effect of heat treatment on activity at pH 10. The synthetic library had a lower average activity under all conditions. This is not unexpected because proteins are highly coupled systems and bringing two polymorphs together that are not found together naturally brings a risk of incompatibility.

The real measure of a library's quality, however, is whether it produces improved variants. The synthetic library had improved variants but comparing the two libraries at pH 10 and pH 7 did not clearly indicate the best method. The top ten sequences for each pH were predominantly from the fragmentation-based library but the best variant at pH 10 was from the synthetic library. This study, like their previous work [11], was limited because activity measurements were not normalized to protein expression levels or performed on purified protein. Thus, it is not clear to what extent the apparent diversity in activity could be attributed to variability in specific activity as opposed to variability in protein production levels resulting from properties of the variant itself or well-to-well variation in expression.

When activity at pH 10 after heat treatment was compared with the ratio of activities at pH 10 and pH 7, the synthetic library had variants that out-performed any variant from the fragmentation-based library. Owing to large changes in activity under different conditions, the activity profile of the variants cannot be attributed solely to expression levels. The superior performance of the synthetic library supports the hypothesis that increased diversity in a library accelerates directed evolution, much as was found in the comparison of DNA shuffling of single genes to DNA shuffling of a family of genes [2], in a study of

Corresponding author: Marc Ostermeier (oster@jhu.edu).

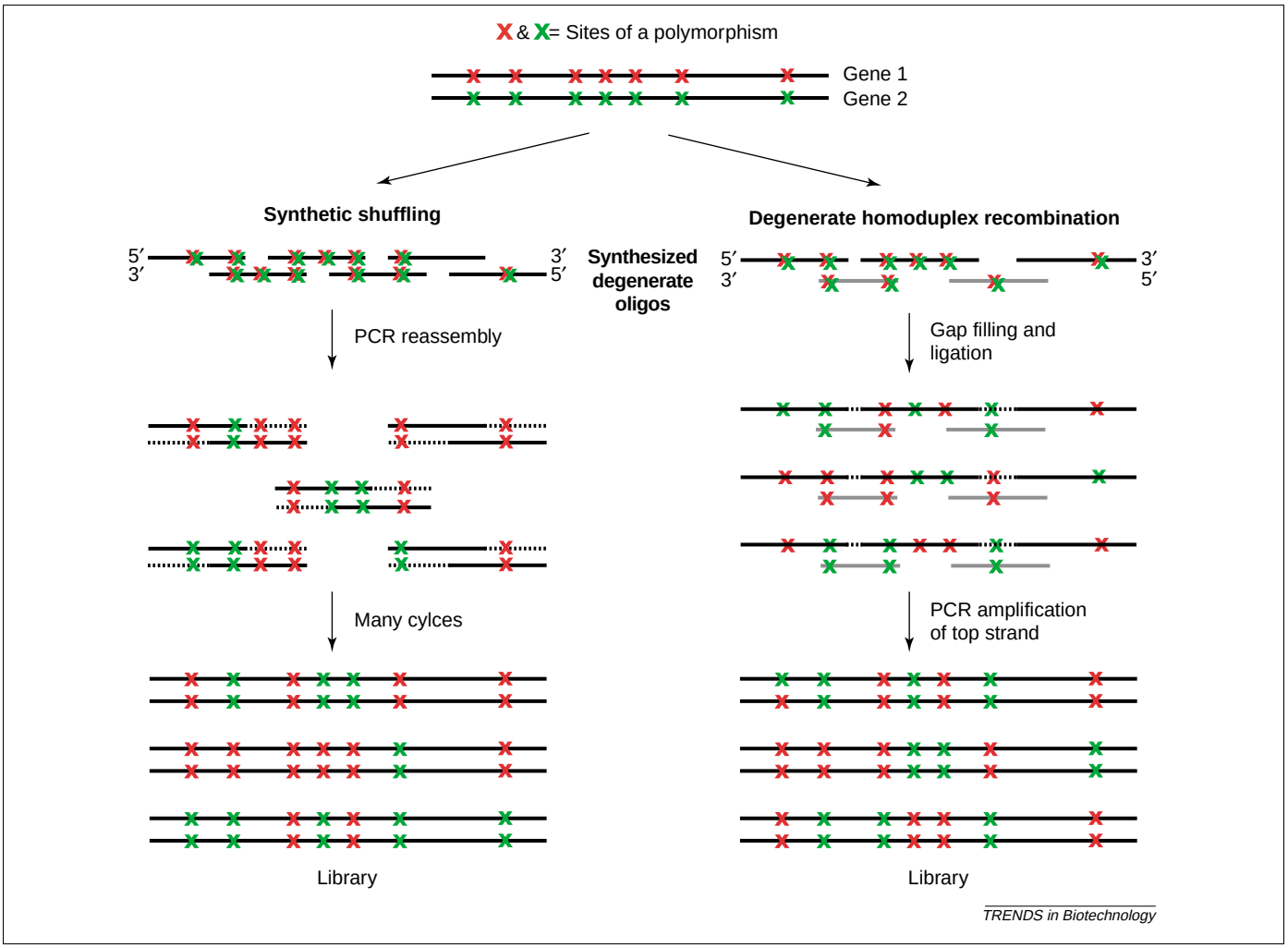

Fig. 1. Schematic representation of synthetic shuffling and degenerate homoduplex recombination. For simplicity, recombination of only two genes with only seven polymorphisms is shown. In synthetic shuffling, the gene is divided into equal length, overlapping, degenerate oligonucleotides that are reassembled in a PCR reaction. For degenerate homoduplex recombination, the oligonucleotides are designed to minimize divergence in the overlap region and the genes are reassembled on a series of scaffold oligonucleotides.

hypermutated antibody libraries [9] and in Monte Carlo's simulations of different shuffling algorithms [10].

Degenerate homoduplex recombination

Most methods for *in vitro* DNA shuffling differ in how the gene fragments are generated but reassemble the genes using a PCR reaction. Previously, Coco and co-workers [5] demonstrated a fundamentally different shuffling strategy, called RACHITT, which assembled recombined genes on a ssDNA scaffold using a series of enzymatic steps. The advantages of this strategy were an increased frequency of crossovers, an increase resolution of crossovers and an elimination of parental DNA from the library. However, the method had a complicated protocol.

Coco and co-workers [8,13] developed a similar but simpler method for the recombination of genes called degenerate homoduplex recombination (DHR), which divides the gene to be shuffled into a series of bridging 'partial scaffold' degenerate oligonucleotides (Fig. 1). Unlike synthetic shuffling, the bottom oligonucleotides are purely a scaffold; they cannot be ligated because they lack of a 5′ phosphate group and cannot be extended by a polymerase owing to 3′ amino modifications. These bottom-strand oligonucleotides form a scaffold on which the top strand is assembled by filling in the gaps with polymerase and ligating the ends together. Thus, DHR relies on a single hybridization and ligation of multiple oligos whereas synthetic shuffling relies on thermocycling and overlap extension. Another innovation of DHR is the selection of overlap regions between the oligonucleotides so that they have minimal divergence, especially near the 3′ end of the top strand. This minimizes biases in the creation of the library owing to some degenerate oligonucleotide pairs inefficiently annealing to each other. In addition, the slow ramping annealing step in DHR favors homoduplex formation, whereas the fast ramping used in synthetic shuffling would more favor heteroduplex formation and should be expected to result in more biased libraries.

A library from all mouse and human epidermal growth factors (EGFs) was created using DHR [8]. Sequencing of random members of the naïve library indicated a high crossover rate and distribution of crossovers that were statistically indistinguishable from complete randomization of all the parental polymorphisms. Because of this, the mouse-human library (3×10^6 transformants and a

degeneracy of 6.5×10^4) can be said to capture every permutation of the parental polymorphisms with a probability >99.99% and is the first complete gene-family library reported. A separately constructed five-species library was equally random with a crossover, on average, every 12.4 bases, the highest recombination density reported. These highly recombined libraries yielded proteins with improved agonist activity and improved binding the EGF receptor, improvements that were less successfully achieved by other approaches [14,15]. Because a fragmentation-based library could not be made for comparison for such a short gene, the significance of an increased diversity was not addressed directly. However, the fact that the improved variant differed from its nearest parent at nine positions strongly suggests that it could only have been made using a synthetic method.

Conclusions

The use of synthetic oligonucleotides in DNA shuffling is not new – synthetic oligos were spiked in some of the earliest fragmentation-based DNA shuffling experiments [16]. What is new is the sole use of synthetic oligos and their design so as to allow amino acids to recombine independently. Aside from a necessity for high-quality oligonucleotides (even minor (n − 1) products produced during oligonucleotide synthesis have a significant negative impact on the library for both methods), recombination methods using synthetic oligonucleotides offer significant advantages over non-synthetic methods. First, hybrid genes with crossovers at much high resolution can be created because the linkage between basepairs found in parental sequences has been destroyed. Second, crossovers in regions of lower homology can be created with a much higher frequency. Both types of crossovers are created by virtue of them being 'pre-made' crossovers present in the degeneracy of the oligonucleotides used in shuffling. Both result in a decrease in exchanges of blocks of sequences commonly seen in fragmentation-based DNA shuffling. In addition, starting with oligonucleotides allows the introduction of any desired 'artificial' diversity into the oligonuclotides not present in existing genes and reduces diversity on the DNA level to better facilitate recombination or to optimize codon-usage. Furthermore, the frequency of an amino acid at any position can be tailored to reflect its frequency among the parental genes at that position or to have an equal representation of rare and frequent amino acids at that position, which increases the diversity of the library.

If it is possible to create and evaluate all polymorphic combinations, as was achieved in the mouse-human EGF library using DHR, such a route will always be preferable. However, is increasing diversity for the sake of diversity always the best route? In most situations it is not possible to evaluate a comprehensive library because the degeneracy rapidly exceeds the library size that can be constructed and evaluated. For example, the number of polymorphic combinations of two, 300 amino acids proteins with 90% identity on the amino acid level exceeds 1 billion (2^{30}). If one can only screen a limited number of variants, what is the optimum degeneracy of the library?

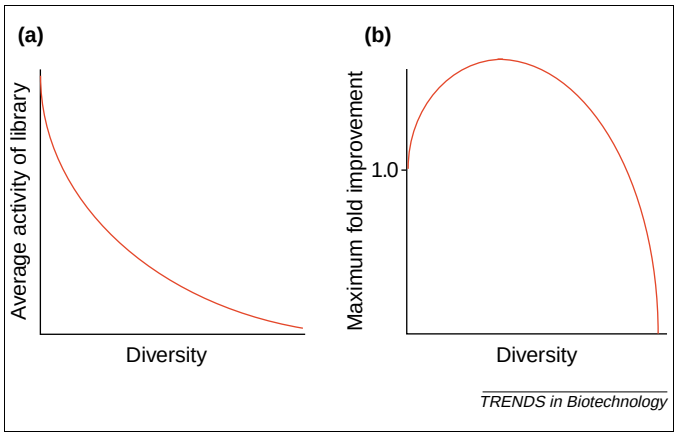

Fig. 2. Postulated relationships between diversity in a library and (a) average activity of library and (b) improvement of best variant. The shape of (b) is as suggested from a statistical mechanical model of the relationship between maximum improvement and average DNA mutation rate of a single gene [17].

How diverse should the library be? We expect that the average activity of the library will decrease with increasing diversity (Fig. 2a). However, we also know that diversity is necessary for creating improved variants and it is reasonable to expect that at some point too much diversity will be counter productive (Fig. 2b). A key question is where does the peak in Fig. 2b lie? How does the frequencies of improved variants and the best variant depend on library diversity? How diverse should the parental sequence be? The successful isolation of improved variants in the shuffling of synthetic oligonucleotides suggests that the diversity obtained by shuffling polymorphisms in an unlinked fashion might well be the preferred strategy. Given the limitations of experimental approaches, perhaps computational approaches have the greatest potential to address these questions in a rigorous manner [17].

As this article was going to press another study on the shuffling of synthetic oligonucleotides was published [18].

References

1 Stemmer, W.P.C. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* 370, 389–391
2 Crameri, A. *et al.* (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 391, 288–291
3 Stevenson, J.D. and Benkovic, S.J. (2002) Combinatorial approaches to engineering hybrid enzymes. *J. Chem. Soc. Perkin Trans. 2*, 1483–1493
4 Moore, G.L. *et al.* (2001) Predicting crossover generation in DNA shuffling. *Proc. Natl. Acad. Sci. U. S. A.* 98, 3226–3231
5 Coco, W.M. *et al.* (2001) DNA shuffling method for generating highly recombined genes and evolved enzymes. *Nat. Biotechnol.* 19, 354–359
6 Ostermeier, M. *et al.* (1999) A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.* 17, 1205–1209
7 Ness, J.E. *et al.* (2002) Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. *Nat. Biotechnol.* 20, 1251–1255
8 Coco, W.M. *et al.* (2002) Growth factor engineering by degenerate homoduplex gene family recombination. *Nat. Biotechnol.* 20, 1246–1250
9 Daugherty, P.S. *et al.* (2000) Quantitative analysis of the effect of the mutation frequency on the affinity maturation of single chain Fv antibodies. *Proc. Natl. Acad. Sci. U. S. A.* 97, 2029–2034
10 Bogorad, L.D. and Deem, M.W. (1999) A hierarchical approach to protein molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2591–2595

11 Ness, J.E. *et al.* (1999) DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* 17, 893–896
12 Stemmer, W.P. *et al.* (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 164, 49–53
13 Coco, W.M. (2001) Decision points and alternatives in DNA shuffling. *Gordon Research Conference in Applied and Environmental Microbiology* Connecticut College, New Haven, CT
14 Groenen, L.C. *et al.* (1994) Structure-function relationships for the EGF/TGF-alpha family of mitogens. *Growth Factors* 11, 235–257
15 Van Zoelen, E.J. *et al.* (2000) The EGF domain: requirements for binding to receptors of the ErbB family. *Vitam. Horm.* 59, 99–131
16 Stemmer, W.P.C. (1994) DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 91, 10747–10751
17 Voigt, C.A. *et al.* (2000) Rational evolutionary design: the theory of *in vitro* protein evolution. *Adv. Protein Chem.* 55, 79–160
18 Zha, D. *et al.* (2003) Assembly of designed oligonucleotides as an efficient method for gene recombination: a new tool in directed evolution. *ChemBioChem.* 4, 34–39

0167-7799/03/$ - see front matter
doi:10.1016/S0167-7799(03)00089-1

Transforming cyanobacteria into bioreporters of biological relevance

Till Bachmann

Institute of Technical Biochemistry, University of Stuttgart, Allmandring 31, D-70569 Stuttgart, Germany

Microbial bioreporters play an important role in environmental monitoring and ecotoxicology. Microorganisms that are genetically modified with reporter genes can be used in various formats to determine the bioavailability of chemicals and their effect on living organisms. Cyanobacteria are abundant in the photosynthetic biosphere and have considerable potential with regards to broadening bioreporter applications. Two recent studies described novel cyanobacterial reporters for the detection of environmental toxicants and iron availability.

Modern analytical methods, along with sophisticated instruments, allow the rapid and sensitive detection of virtually any chemical in a high-throughput format. However, these analytical techniques do not provide information about the potential biological harmfulness or the bioavailability of compounds under study. For this purpose, bioanalytical tools have been developed as a powerful supplement to chemical analysis. The technical concept of these devices usually allows for both the detection of individual analytes or of groups of chemicals that display common interactions with the biological material.

Wild-type cyanobacteria were among the first organisms to be used for developing whole-cell biosensors for environmental monitoring. They are phototrophic organisms and ecologically important and are therefore, ideally suited for the monitoring of compounds that inhibit photosynthetic activity, such as herbicides. Cyanobacteria have been used for the detection of phytotoxic pollutants based on amperometric sensors employing shuttled [1] or direct [2] electron transfer. Drawbacks of the sensors were their instability, lack of robustness and specific technical requirements.

To overcome these shortcomings, microbial cells can be genetically modified by introduction of a 'reporter gene' to connect the initial biological interaction of the tested chemical or physical event to an easily recordable output signal (e.g. light). Probably the most commonly used reporter proteins are β-galactosidase, green-fluorescent protein (GFP) and luciferase, from either bacteria (e.g. from *Vibrio fisheri, Vibrio harvei, Xenorhabdus luminescens*) or insects (e.g. *Photinus pyralis*). By convention, these systems are referred to as bioreporters. They follow two technical designs, rather non-specific reduction or specific induction of luminescence, which will be exemplified for cyanobacterial reporters in the following paragraphs. Microbial reporter sensors were reviewed by Daunert *et al.* [3] and Köhler *et al.* summarized their applications to environmental monitoring [4].

Detection of phytotoxicity

Recently, Shao and colleagues described a bioluminescent *Synechocystis* PCC 6803-derived reporter strain for monitoring the correlation between cyanobacterial activity and the presence of environmental toxicants [5]. *Synechocystis* PCC 6803 was selected for the experiments because of the availability of diverse tools for genetic alteration of the organism (the genome of *Synechocystis* PCC 6803 has been sequenced [6]; http://www.kazusa.or.jp/cyanobase/) and, more importantly, because this fresh water cyanobacterium is ecologically significant. When exposed to a sample, the decreasing luminescence of the reporter strain indicated the presence of photosynthetic inhibitors and other toxicants, such as heavy metals and volatile organic pollutants.

To create this reporter strain, Shao introduced a construct consisting of the constitutive *tac* promoter fused to *luc* (the gene encoding the firefly luciferase) and *luxAB* (encoding the bacterial luciferase) into the cyanobacterial cells by means of an integrative vector. The integration into the cyanobacterial chromosome ensured genetically stable bioreporters. Because the substrate of the bacterial enzyme proved to be inhibitory in their

Corresponding author: Till Bachmann (Till.Bachmann@po.uni-stuttgart.de).