Molecular Breeding of Genes, Pathways and Genomes by DNA Shuffling

Willem P. C. Stemmer

Maxygen, Inc., 515 Galveston Drive, Redwood City, CA 940663, USA

Abstract Existing methods for optimization of sequences by random mutagenesis generate libraries with a small number of mostly deleterious mutations, resulting in libraries containing a large fraction of non-functional clones that explore only a small part of sequence space. Large numbers of clones need to be screened to find the rare mutants with improvements. Library display formats are useful to screen very large libraries but impose screening limitations that limit the value of this approach for most commercial applications. By contrast, in both classical breeding and in DNA shuffling, natural diversity is permutated by homologous recombination, generating libraries of very high quality, from which improved clones can be identified with a small number of complex screens. Given that this small number of screens can be performed under the conditions of actual use of the product, commercially relevant improvements can be reliably obtained.

Keywords: molecular breeding, DNA shuffling, single gene shuffling, family shuffling, whole genome shuffling

INTRODUCTION

All of the beautifully complex biological structures and sequences that Nature has provided us with have been designed by a single process, that of natural evolution. In fact, all truly complex adaptive systems were generated by an evolutionary process. However, molecular biologists mostly use a fundamentally different approach to engineering these same molecules, which is rational design by molecular modeling. Although powerful, molecular modeling operates only at the level of protein structure and cannot predict the effect of protein mutations at the level of DNA, mRNA, protein folding or protein interactions with other molecules in the environment.

One logical approach for further adapting natural sequences is to apply the same proven, evolutionary tools that created these sequences in the first place. An applied and directed variant of natural evolution, classical breeding, has already shown how sequences obtained by natural evolution can be further improved. The green revolution clearly demonstrated that complex, multigenic traits of whole organisms can be optimized without any sequence or structural information regarding the underlying genes. Breeding is a slow but proven method for optimizing whole, typically eukaryotic genomes. Breeding is a practical and conceptually simple recursive process that involves mostly homologous recombination of DNA between related genomes followed by screening of the progeny for clones with improved properties. If whole genomes can be readily improved by classical breeding, then variations of this approach should be useful for the improvement of single genes, pathways and microbial genomes, which are the typical targets of molecular biologists.

'Molecular breeding' is conceptually similar to classical breeding in that it aims for rapid functional improvement of sequences by recursive cycles of sexual evolution, each cycle comprising diversity generation (by homologous recombination and other forms of mutagenesis), screening for the best individuals, and amplification of the winners to go on to the next cycle. Molecular breeding comprises a diversity of technical formats for diversity generation that can be applied to a much wider range of targets, including single genes, pathways, episomes, viruses and bacterial and eukaryotic genomes. Molecular breeding techniques also allow much better control over all aspects of the breeding reaction (such as number of parents, level of homology, cross-species, crossover number, level of point mutation) than classical breeding. The progeny can be expressed and screened in cell-free systems, in bacteria, in eukaryotic tissue culture cells, or even in whole organisms, as the application demands. This allows the selection pressure to be focused on commercially relevant properties. Molecular breeding typically has a generation time of a few days.

COMPARISON WITH EXISTING TECHNIQUES

Already there is a variety of useful sequence improvement approaches. Classical breeding works by recom-

* **Corresponding author**
Tel: +1-650-298-5300, Fax: +1-650-364-2715
e-mail: pim.stemmer@maxygen.com

Classical Breeding	Molecular Breeding
• Cycle time = years	• Cycle time = days
• Whole genome	• Genes, pathways, genomes
• Breed within species	• Breed across species
• Two parents	• One to multiple parents
• Limited control	• Multi-level control
• Complex selection pressure	• Focused selection pressure
• Whole plants and animals	• Microbes, cells, whole organisms
Slow breeding of whole eukaryotic genomes	Rapid breeding of genes, pathways, episomes, viruses, partial & whole genomes

Fig. 1. Comparison of classical breeding with molecular breeding.

bination of natural, proven diversity, rather than by random point mutagenesis, but is limited to whole eukaryotic genomes. Classical strain improvement uses random point mutagenesis to improve whole microbial genomes. In both approaches, even though the target is large, recombination and natural diversity are not used. Existing approaches to improve single proteins include: (1) specific modeling-directed point mutations; (2) the generation of libraries of mutants by synthetic mutagenesis using oligonucleotide cassette libraries (including random, biased, saturation and codon-based oligonucleotide mutagenesis), which introduces partially random mutations in a specific, generally single location of a gene and (3) random point mutagenesis (including error-prone PCR, base-analog, mutator strains and UV or chemical mutagenesis), which can introduce (partially) random mutations throughout a gene.

Molecular breeding has the greatest advantage over these alternative approaches when the target is complex, such as a multi-component protein with complex interactions, when structural information is not yet available, or when the principal target does not involve a protein structure [1].

Since the potential diversity of most libraries, which is the number of sequences this library could potentially contain (potential sequence space), is generally much larger than the actual library size (the number of different clones actually generated in a single library), significant additional improvement can be obtained by performing additional cycles of evolution in a recursive fashion, where the output of one cycle becomes the input for the next.

A key question has been how to go from cycle to cycle. For theoretical and practical reasons, it is important that the diversity generation can be applied to a pool of selected clones blindly, without sequencing all of the selected clones. Some of the diversity generation formats (such as the popular synthetic oligo mutagenesis formats) require sequencing of the pool of selected clones and resynthesis of oligonucleotides.

Regardless of the source of the initial diversity, recombination of the best sequences is generally regarded as the preferred approach to go from one evolutionary cycle to the next. This is equally true in other fields of evolutionary engineering, such as genetic algorithms and artificial life. While recombination plays a fundamental role in the evolutionary design process, there are many variations of recombination that one can choose from and these details can affect the outcome.

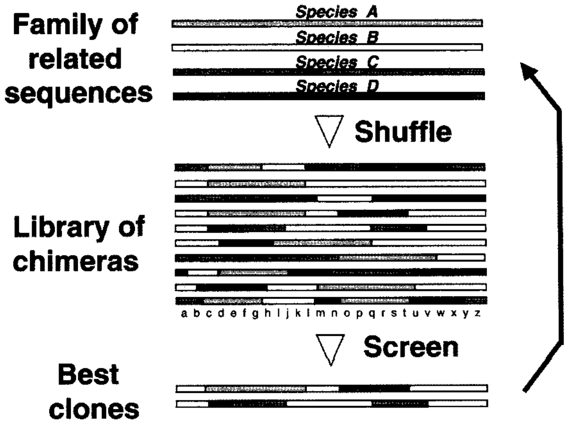

Fig. 2. Homologous recombination formats.

DNA SHUFFLING FORMATS

DNA shuffling involves the recombination of multiple DNA sequences to create a library of chimeras. Important variables for constructing shuffled libraries are: (1) the length of the shuffled DNA, (2) the DNA homology, (3) the quality of the sequence diversity (natural, proven diversity versus random point mutagenesis), (4) the crossover number and (5) the number and molar ratio of the parents.

HOMOLOGOUS RECOMBINANT FORMATS

Maxygen Inc. has developed many formats for homologous recombination, both *in vitro* and *in vivo*. A widely practiced format for DNA shuffling involves the random fragmentation of a pool of related DNA sequences [2]. After denaturation of the fragments, homologous sequences from different templates hybridize and prime each other and the resulting crossovers are locked in by polymerase extension. Multiple cycles of this fragment reassembly reaction result in a library of full-length, homologously recombined chimeric sequences that contain a variable number of crossovers. After insertion of this library of DNA sequences into an expression vector and transfer into a host cell for expression, the clones are screened for new or improved properties.

NON-HOMOLOGOUS RECOMBINANT FORMATS

A variety of formats for recombination that do not require DNA sequence homology has also been developed mostly for *in vitro* use but also for use *in vivo*. Several of these formats are useful to recombine genes in 'homologous' locations in a protein (based on protein sequence alignment) but without relying on DNA sequence homology to create the crossover. These formats are used to recombine related genes that can be aligned but of which the sequences are too different to allow recombination based on DNA sequence homology.

Another class of non-homologous recombination formats is aimed at the shuffling of proven building blocks, such as the protein domains and exons found in eukaryotic proteins. In contrast to most other shuffling methods, exon shuffling (and other block permutation approaches) typically creates length variations, such as exon insertions and deletions, in addition to substitutions.

The key to making high-quality libraries of recombinants by non-homologous recombination, such as exon shuffling, is to perform conservative substitutions. However, there are many different ways to be conservative other than substitution based on the similarity of the DNA or protein sequence. For example, one can substitute on the basis of similarity in structure (*i.e.* shuffling structurally related P450s), similarity in function (*i.e.* swapping exons encoding structurally unrelated serine proteases), similarity in immunological response (*i.e.* replacing one human exon with another human exon in a human pharmaceutical protein, to avoid raising an immune response), or similarity in exon splice frame to avoid a frameshift.

These libraries are then screened or selected to identify the clones with the best combinations of the permutated sequence diversity. As in classical breeding, a pool of the best clones (rather than the best single clone) is used as the input for the next round of breeding.

BETA-LACTAMASE

After developing the very first format (fragmentation) for DNA shuffling, we wanted to compare it with existing methods for the evolution of a gene. We picked a known test system, using the widely used TEM-1 β-lactamase and evolved it in *E. coli* for increased resistance to B-lactam antibiotic called cefotaxime. Starting with a single gene, three recursive cycles of shuffling (under conditions that introduce random point mutations) and selection on plates with increasing levels of the drug yielded a clone with a 16,000-fold increased drug resistance [3]. This gene was found to have many amino acid and silent mutations and we were interested in understanding which of these caused the improved activity.

To flush out those mutations that do not contribute to the mutant phenotype, the mutant was backcrossed by breeding with the parental sequence. Whereas classical backcrossing is limited to a 1:1 molar ratio and needs to be performed multiple times, with DNA shuffling we can use any molar ratio that we desire. Using a 40:1 molar ratio of parental to mutant DNA followed by selection for resistance, we obtained a further improved mutant with a 32,000-fold increased drug resistance [3]. All four of the silent mutations had been replaced by wildtype sequence, showing that the backcross had efficiently removed non-contributing mutations. Backcrossing is particularly useful for reducing the number of amino acid mutations in proteins, to reduce the immunogenicity of pharmaceutical proteins or to focus on the key mutations to understand the mechanism of the improvement.

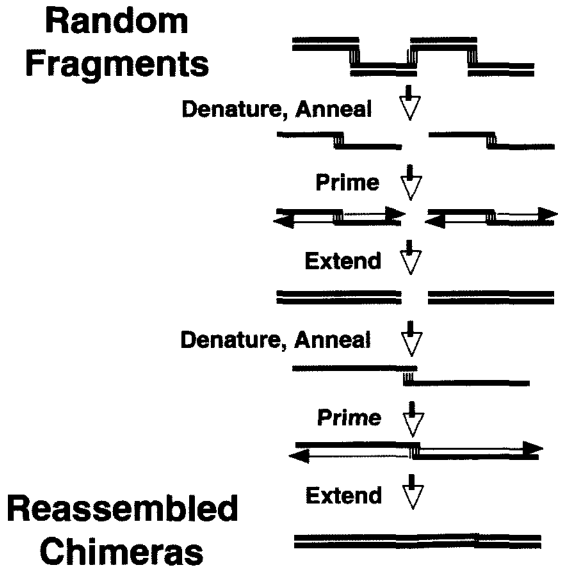

Fig. 3. Non-homologous recombination formats.

The optimization of this same enzyme for this same drug by rational design had been previously published. Modeling of the known structure of this B-lactamase identified three active site loops that were randomized using oligonucleotide cassette mutagenesis. The most resistant clone that was found had a 16-fold increased drug resistance. Thus, without using molecular models and in a three-week time frame, DNA shuffling yielded a 2,000-fold greater drug resistance than the standard approach of modeling and cassette library mutagenesis.

In comparison, we also evolved the same enzyme-drug combination by random point mutagenesis, using error-prone PCR. Surprisingly, the best mutant obtained after three cycles of point mutagenesis only showed a 16-fold increase in drug resistance. The shuffled mutant contained two well known mutations (E104K and G238S), which together yield a 500-fold increase in resistance. These two mutations were also found by cassette mutagenesis but only in separate clones because the libraries were made in three different areas and never combined, resulting in only a four to eightfold increased resistance. Given that two mutations can give a 500-fold increase, it is surprising that three cycles of error-prone PCR only yielded a 16-fold improvement. The probable explanation is that the ratio of deleterious mutations to useful mutations is so high that the frequency of mutants containing several positive mutations but no deleterious mutations is exceedingly low. The initial diversity in this example was created by random point mutagenesis, whereas shuffling creates new combinations of the mutations in the selected clones and replaces deleterious mutations with wildtype sequence.

GREEN FLUORESCENT PROTEIN

Green fluorescent protein (GFP), is a widely used marker protein obtained from jellyfish, and has quantum efficiency of 70-80%, our construct was expressed at an unusually high level of 75% of total protein in *E. coli*. One would *a priori* expect that shuffling could not improve the fluorescence signal in such a system.

We performed three cycles of visual screening of 10,000 clones per cycle for increased fluorescence [4]. The best mutant we obtained had a 45-fold improved fluorescence signal over the standard, native GFP gene. Sequencing of the evolved GFP gene showed that it contained three substitutions of hydrophobic amino acids with hydrophilic residues. Fractionation of *E. coli* cells showed that most of the native GFP was in inclusion bodies and that only a small amount of soluble GFP was fluorescent. By contrast, most of the evolved GFP remained soluble and was fluorescent. Thus, the three mutations allowed GFP to avoid an aggregation pathway and instead allowed the protein to stay soluble and fold properly, this proper folding is required for the autocatalytic activation of the fluorescent chromophore.

Traditionally, the property that limits the activity of GF would be determined, for example promoter activity, mRNA stability, mRNA translation, protein stability, specific activity or one of many other factors. Even if this analysis had shown protein folding to be the limiting factor, it would have been nearly impossible to solve such a folding problem using rational means because the modeling of protein folding is still primitive.

In summary, DNA shuffling was able to solve a complex protein folding problem without having to determine what the performance-limiting property of the system was. The 45-fold improvement in signal was maintained on transfer of the GFP genes to mammalian and plant cells.

FUCOSIDASE

In another example, the substrate specificity of an enzyme was changed; β-galactosidase, the largest *E. coli* protein was changed into a fucosidase. Using the chromogenic substrates nitrophenylgalactose and nitrophenylfucose 10,000 clones per cycle were screened. The best clone of the seventh cycle resulted in a 66-fold improved in fucosidase activity and a 1000-fold in fucosidase specificity over the native β-galactosidase construct [5]. Plated on X-fucose, the starting construct formed white colonies whereas the evolved construct formed blue colonies, showing a full phenotypic switch.

MULTI-GENE PATHWAYS

As a starting point, molecular breeding simply requires a DNA fragment that encodes a screenable phenotype. The number of genes on this fragment, their complex interactions, or even the DNA sequence of this fragment do not have to be known. As in classical breeding, no sequence information is required at any point.

We have demonstrated the ability to evolve multigene pathways in three examples, an arsenate detoxification operon, an atrazine degradation pathway and a mercury detoxification pathway.

ARSENATE DETOXIFICATION OPERON

Starting with a 2.3 kb operon from *Staphylococcus aureus*, encoding three genes of an arsenate resistance pathway, the whole plasmid was shuffled in order to access any possible mechanism for improving resistance Growth was selected for on increasing concentrations of arsenate and following three cycles of shuffling the best mutant that was obtained was 40-fold more resistant [6]. This mutant grew at up to 500 mM arsenate, close to the solubility limit for arsenate. Shuffling and selection reproducibly caused the operon to integrate into the chromosome, contributing to the arsenate resistant phenotype. A control plasmid, without shuffling but with three cycles of selection, showed no increase in arsenate resistance and the plasmid did not integrate. Sequencing of the operon after retrieval from the chromosome showed that it contained 13 base mutations, including three amino acid mutations in the membrane pump. We had expected that the improvement would be obtained by mutations in the reductase; although no mutations were found in the reductase, the activity of the reductase was increased approximately tenfold, but this must have been as a result of mutations in flanking sequences.

This example shows that even if the rate-limiting target gene is known, the mutations that alleviate this limitation might lie outside of that gene, and thus it is advisable to shuffle a sequence larger than just the gene itself. The results suggest that shuffling activated the insertion sequences that were known to flank the operon and that the activated sequences mediated insertion at a favorable location on the chromosome. This chromosomal insertion library strategy by shuffling of IS sequences could be a useful tool for the expression optimization of chromosomal constructs.

SINGLE SEQUENCE SHUFFLING

In single gene shuffling, the source of diversity is generally random point mutagenesis. When one makes a library of random point mutants of a single gene, in general all the active gene products generally differ by only one, two or three amino acid mutations from the starting sequence. A higher level of mutagenesis results

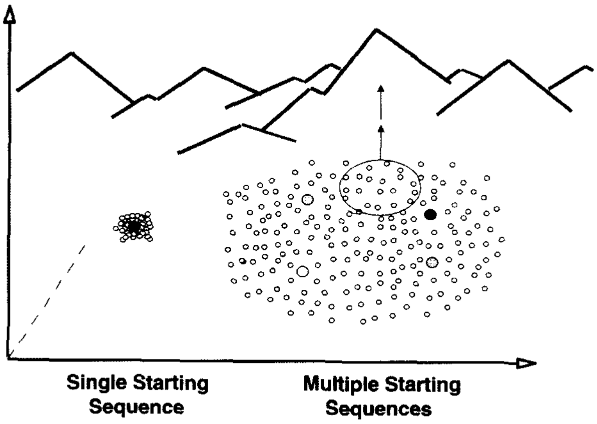

Fig. 4. Single starting sequence and multiple starting sequence.

in a higher level of inactive clones. Random point mutagenesis libraries, therefore, over-sample a small sequence space directly surrounding the starting sequence.

By contrast, initially we would prefer to sparsely sample a much larger part of sequence space. Screening of this 'sparse library' would then lead to areas of sequence space representing high function. In additional cycles of breeding of the pool of selected sequences, these areas are sampled at increasingly higher density, resulting in (hill-climbing) toward the peaks representing the highest local optima.

FAMILY SHUFFLING

Such sparse libraries can be obtained by pool-wise shuffling of multiple related sequences, typically homologs obtained from natural diversity. This approach, called "family shuffling", involves mixing of a family of homologous sequences obtained from nature, typically the same gene from related species or related genes from a single species. The library of chimeric sequences is screened and a pool of the best clones is used as the input for the next shuffling cycle. Because sequence blocks carrying multiple mutations are exchanged, 'neighboring' clones generally differ by many mutations, causing a sparse distribution of the library in sequence space. This sparse distribution is ideal in that it allows sampling of a very large sequence space by screening a relatively small number of mutants. Higher density sampling is performed in subsequent cycles but only in areas of proven high functionality.

CEPHALOSPORINASE FAMILY

Our first example of family shuffling used four 1.6 kb cephalosporinase genes obtained from four different genera of bacteria – *Citrobacter, Klebsiella, Enterobacter* and *Yersinia* [7]. The DNA homologies were low, ranging from 58% to 82%. Each of these four genes was shuffled separately and 50,000 clones from each of the four libraries were plated on moxalactam. For all four single gene libraries, the best mutant clones showed and increase in moxalactam resistance of approximately 8-fold over the parental gene.

As a direct process comparison, all four genes were family-shuffled together as a single pool to create a library of chimeric genes and plated 50,000 clones on moxalactam. The best clone obtained with this process was 270-fold improved over the most resistant parents and 540-fold improved over the least resistant parents. *Therefore, a single step of family shuffling resulted in a* 50-fold better solution than the single gene shuffling method.

Characterization of the best clone obtained by family shuffling demonstrated that it contained eight segments from three of the four parents. The seven crossovers all occurred in areas of sequence homology. Whereas the best clones in this example contained up to 33 point mutations, in subsequent family shuffling examples reduced the rate of random mutations using proof-reading polymerases.

The structure of one of the four parental cephalosporinases is known and this was used to obtain a model of the chimeric enzyme. After energy minimization, the α-chain backbone was nearly identical to that of the known structure, despite the presence of 142 amino acid side chain differences. The shuffled enzyme was substantially different from any of its parents: it differs at 196 aa from the *Yersinia* parent (50% of residues), at 102 aa from the *Citrobacter* parent, at 181 aa from the *Klebsiella* parent and at 142 aa from the *Enterobacter* parent. Thus, in a single step of family shuffling we obtained a sharply improved enzyme by making a huge leap in sequence space, which clearly could not have been possible using any other protein engineering methods.

NATURAL DIVERSITY VERSUS RANDOM MUTATIONS

Why was the cephalosporinase result achievable by family shuffling and why would a similar result not be achievable by random point mutagenesis? The explanation lies in the quality of natural diversity. Compared with random mutations, natural diversity is far more conservative – functionally, structurally and perhaps also immunologically and evolutionarily. Because of the conservative nature of natural mutations, the average shuffled clone in these libraries can be as active as the average of the parents. Because large blocks of sequences carrying multiple natural mutations are exchanged, sequence space is sampled sparsely, so that the same library size covers a much greater sequence space. Therefore, family shuffling mediates rapid optimization of sequences by creating complex combinations of proven, functional steps obtained from natural diversity.

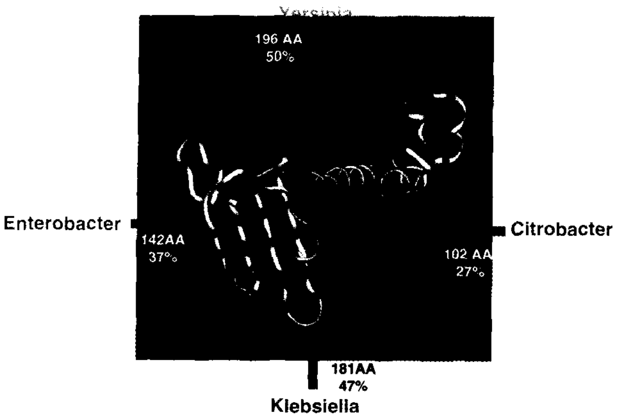

Fig. 5. Family shuffling of cephalosporinase from four different genera of bacteria.

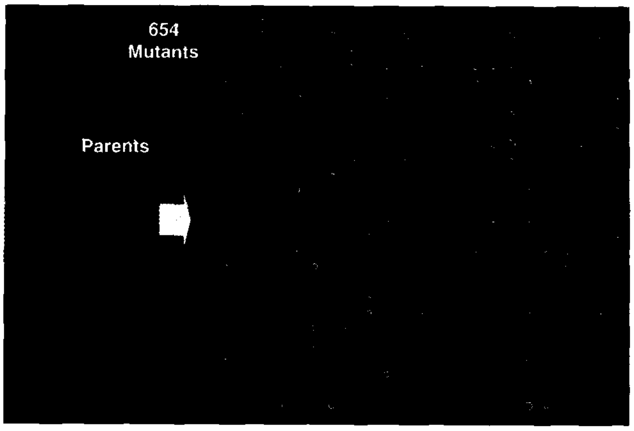

Fig. 6. Shuffling of subtilisin genes.

FINDING HOMOLOGS

Family shuffling requires multiple sequence homologs and it is now almost always possible to obtain or synthesize multiple homologous sequences from the assay databases now available. When homologs are not known, they can easily be found by screening for hybridization to the known gene or by PCR using multiple sets of primers based on the known sequence. Even if only two gene homologs are available, for example two genes that differ at only 25 amino acid positions, the shuffling of these genes should still create a library of up to 2^{25} or 30 million different chimeras. Homologs do not have to be active to contribute useful sequences to the breeding reaction.

MIMICKING OTHER CLASSICAL BREEDING TECHNIQUES

Classical breeding comprises a broad variety of breeding concepts as well as tools for thorough mathematical analysis. We developed simple molecular mimicks of classical techniques such as inbreeding, outbreeding, crossbreeding, founder effect and backcrossing. Inbreeding refers to the repeated self-fertilization of a population without adding additional diversity, which is the format that we used for most of the examples described here. In contrast, out-breeding routinely adds new diversity, by breeding with parental genes that were not involved in the earlier crosses. We routinely exploit the founder effect by constructing multiple libraries in parallel from different combinations of parental genes, as a small variation in parental composition can dramatically affect the outcome. Then a small number of clones can be screened from each of these libraries to determine which library is best, and then a much larger number of clones are screened to obtain the best individual sequences. In backcrossing, improved mutant gene is breed with a molar excess of the parental gene to flush out those mutations that least contribute to the mutant phenotype. The goal of backcrossing is to obtain a sharp reduction in the number of mutations with only a minor decrease in performance.

SUBTILISINS

Subtilisins are a family of serine proteases with sales of approximately US$ 500M, mostly for laundry detergent. It is the most highly engineered protein with >130 solved crystal structures and nearly every residue is covered by a patent. Thus, even a two-fold improvement can be commercially significant.

In collaboration with Novo Nordisk, which sells more than half of the world's industrial enzymes, 26 subtilisin genes were shuffled (including the gene for Savinase, Novo's leading commercial protease) as a single pool. The pairwise DNA homologies ranged from 56% to 99%. A library of 654 active mutants was obtained and this small library was screened for protease activity under five different conditions: pH 5, pH 7.5, pH 10, thermostability and solvent stability [9].

For each of the five screening conditions, 4-12% of the clones were improved over the best of the 26 parents for that assay condition. Furthermore, the library of chimeras showed a greatly increased diversity of combinations of enzymatic properties. For example, 77 out of the 654 clones were better than the best parent in terms of activity at pH 10. In addition, some of these clones were three-fold more thermostable and others about 1.5-fold more solvent-stable. Several clones were improved in all three properties over the best parental enzyme.

The best clones for each of the five screening conditions were all complex chimeras and differed from savinase, the leading commercial protease, by 21 to 32 amino acid positions, suggesting that these results could not have been obtained from known structures by rational design.

The results demonstrate that family shuffling can create a high quality 'single-pot' library from which enzymes with a wide range of desirable properties can be obtained.

THYMIDINE KINASE

The most active thymidine kinase (tk), that of Herpes Simplex Virus I, has been used in several clinical trials as a 'suicide' gene for killing cancer cells. To increase its rate of phosphorylation of AZT, as phosphorylated AZT is the active compound, HSVI tk was shuffled with the HSVII tk, despite its 400-fold lower activity (the DNA homology is 78%). After four cycles of robotic screening of 11,500 clones per cycle, the best chimera showed a 32-fold improved in activity over the best parent, HSVI tk. The specificity for AZT over thymidine was improved 44-fold. The best mutant was a complex chimera formed by ten crossovers between the two genes and contained five non-parental amino acid mutations [10]. This chimera differs from HSVI tk at 22 amino acids and from HSVII tk at 86 amino acid residues.

MURINE LEUKEMIA VIRUS

The performance of viral vectors continues to be a critical issue in gene therapy and a diversity of techniques for virus optimization have been pursued [15]. The utility of shuffling for virus improvement has been demonstrated in two examples. In the first example, a retrovirus with a new host specificity was evolved. The six envelope genes were shuffled and reinserted into the Moloney virus and a library of 5 × 10^6 clones was incubated with a co-culture of CHO K1 cells and a semi-permissive cell line that was included to keep viral titers high. After five passages, this selection for viral replication yielded multiple isolates that were able to efficiently infect hamster CHO K1 cells [13]. Controls with the pool of six parental constructs in Moloney did not yield any CHO K1 infectious activity. The envelope sequence of the best chimeric clone consisted of four segments from three parental viruses. The results suggest a potential role for glycosylation in the interaction of the envelope and the cellular receptor.

A second example of virus breeding was in a collaboration with Novartis and aimed at solving a critical problem in the manufacture of retroviruses for gene therapy. The virus purification and concentration process results in a sharp reduction in viral yield due to a lack of mechanical stability. A library of 5 × 10^6 replication-competent shuffled viruses, was subjected to the purification with amplification after each cycle. The best viral clone that was isolated exhibited no loss in titer under conditions that reduced the titers of the parental viruses by 30-100-fold [14].

INTERFERON ALPHA

hIFN α-2a is used in anti-viral and anti-cancer therapy and has a market size of approximately $1B. For unknown reasons, humans have more than 20 genes for IFN-alphas, with DNA homologies ranging from 85% to 95%. The human IFN-α gene family has been shuffled to find sequences with greater potency. Multiple sets of ambiguous primers against the known IFN genes were used to clone the whole repertoire of hIFN-α genes from pooled human genomes. These PCR products were pooled and shuffled directly, without subcloning and sequencing. The quality of the library was confirmed by sequencing and the chimeric proteins were expressed and directly screened for activity in an anti-viral assay using lymphocytic choriomeningitis virus, growing on murine cells [12]. Since we assay for the activity of human interferons on the mouse interferon receptor, the starting activity is low. Half the shuffled clones were found to be as active as the parental interferons, demonstrating the unusually high quality of shuffled libraries. Because the anti-viral assay is labor intensive, pools of 12 and 96 randomly picked clones were assayed and the most active pools were deconvoluted.

Only 68 assays were needed to find a chimera that had 135,000-fold higher specific activity than hIFNα-2a. The best clone following a second round of shuffling had a 285,000-fold higher specific activity than hIFNα-2a, and 185-fold more active than hIFNα-1, the most potent human interferon. Despite the fact that the human sequences are only about 65% identical to the murine sequences, the shuffled human genes are 3-fold more active than the most active mouse IFN, mIFN-α-4. The best chimeras all contained segments from multiple IFN genes, but no new point mutations. The ability to improve proteins by simply permutating diversity that is naturally present and well tolerated in humans is important as it should significantly decrease the risk of immunogenicity of engineered proteins.

This work also demonstrates that screening of a surprisingly small number of clones (only 68 assays) from a family-shuffled library can lead to powerful enhancements in specific activity.

FEWER ASSAYS MEANS MORE APPLICATIONS

In single genes only about 0.1% of random point mutations result in an improved phenotype. Because of the deleterious nature of most random point mutations, random mutagenesis methods generate libraries containing a majority of clones that are inactive or have reduced activity. Because of the low library quality, large numbers of mutants need to be screened to find the rare improved clones. Library display formats such as phage display, ribosome display and cell surface display are capable of screening very large numbers of clones and are therefore often required to screen libraries generated by random mutagenesis methods. The general trend in the display field is towards larger and larger libraries.

A major drawback of these display formats, however,

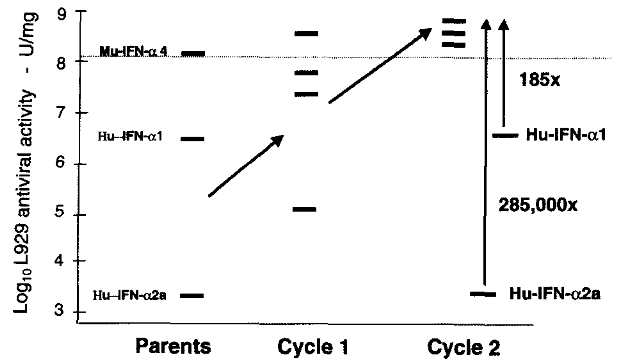

Fig. 7. Shuffling of interferon α gene family.

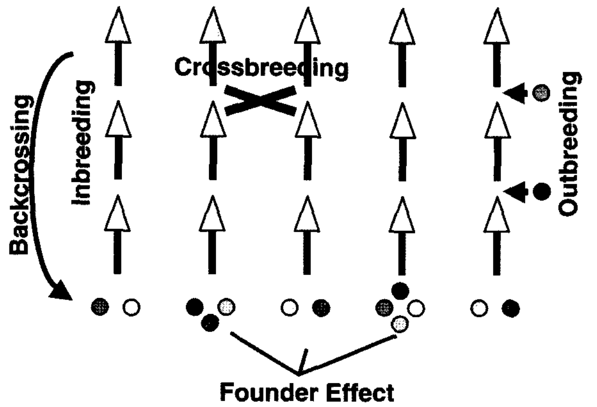

Fig. 8. Dog breeding example.

is their incompatibility with commercially relevant sequences, including eukaryotic proteins, large proteins, membrane proteins, co-factor requiring enzymes, glycosylated proteins, multi-protein complexes and non-coding sequences, because most of the display formats are bacterial. Display formats are good at screening for binders, but are generally not good at screening for commercially important properties such as catalysis, transfer properties, or function inside a cell or whole organism. The use of surrogate binding assays instead of *screening for the property of interest directly* frequently leads to clones that perform well in the surrogate assay but are not improved in the property of commercial interest. In addition, strong sequence biases are imposed by each of the display formats themselves.

By contrast, many commercial applications depend on the ability to access a small number of high quality, complex screens rather than high-throughput, low quality screens. Examples are fermentation yield, immunogenicity of vaccines, gene therapy vectors or pharmaceutical proteins or yield drag assays in whole transgenic plants. The higher the quality of the library, the smaller the number of assays required to obtain improvements and the larger the number of accessible commercial applications.

In some cases the number of assays can be reduced by screening pools of clones. High quality libraries could be screened directly in whole transgenic plants or animals. This is particularly useful when the desired property can only be screened for in whole organisms, such as the yield of a transgenic plant.

The quality of the libraries is therefore critical for broad commercial success. Obtaining improvements with a small number of clones requires that most of the clones in the shuffled libraries are active, which is achieved by conservative family shuffling of closely related genes. Conservative shuffling requires limiting the target size, the degree of sequence divergence, the quality of the diversity, the number of parents and the number of crossovers. The following example suggests that if these variables are carefully controlled we can expect to be able to improve whole prokaryotic and eukaryotic genomes with a surprisingly small number of assays.

DOG BREEDING

To understand the number of assays needed to improve sequences, we consider an extreme example, which is classical breeding of dogs. In dog breeding one shuffles whole genomes of 10^9 basepairs, which can differ by millions of mutations are shuffled. Despite the hundreds of random crossovers that are created during meiosis, the resulting puppies are phenotypically diverse but healthy, which suggests high library quality. *In five generations of outbreeding with ten puppies per* litter one can obtain almost any phenotype in 'dog-space'. Thus, a total of 50 assays is sufficient to breed whole eukaryotic genomes resulting in dramatic phenotypic changes. The ability to obtain significant improvements with a small number of assays allows the screening to be performed directly in whole transgenic animals and plants. This is useful when the desired property can only be screened for in whole organisms, such as the yield of a transgenic plant.

By focusing all of the diversity and assay capacity on a single gene rather than on 30,000 genes, a much more efficient optimization of that specific single gene would be expected.

The dog breeding example shows that it is possible to obtain significant improvements in targets of any length with a surprisingly small number of assays. The critical variable in both the whole genome and the single gene extremes is the conservative nature of libraries created by permutation of natural, proven diversity. In practice, for targets of a few kilobases selected sequences with a proportionally lower degree of homology (*i.e.* 70% to 90% versus >99.9%) should be used, a higher crossover density should be applied and more parents should be included. Therefore, for single genes we generally breed across the species barrier.

WHOLE GENOME SHUFFLING

We have also developed formats for shuffling contiguous pathways, distributed pathways (for example targeting ten specific genes in distributed locations, without altering the rest of the genome) and whole microbial genomes have also been developed. The dog breeding example shows that, in principle, it will be possible to extend whole genome shuffling to eukaryotic genomes. Some of the shuffling formats for whole microbial genomes can also be applied to optimize whole microbial communities. The preferred technical format depends on the specific application and what is known about the target(s).

Microbial whole genome shuffling is particularly promising for strains that were developed by classical strain improvement. Classical strain improvement involves multiple cycles of point mutagenesis by UV, radiation or chemical treatment of the whole microbial genome, followed by screening for improved mutants. Because of the very high ratio of deleterious mutations to positive mutations, approximately 10,000 mutants have been screened to find one with improvements in the property of interest. The improved strain typically accumulates many mutations that have deleterious effects in properties that are not directly related to the property of interest and become sensitive to growth conditions, 'brittle' and of reduced viability.

Whole genome shuffling by recursive protoplast fusion was applied to *Streptomyces fradiae* to improve tylosin production [16]. Only two rounds of genome shuffling and screening yielded strains with tylosin titres ~10-fold higher than the parental strain and identical to modern production strains. While the production strain required 21 rounds of classical strain improvement by random point mutagenesis and screening of more than a million clones over 20 years, our microbial breeding approach required only a year and 25,000 clones.

Cross-breeding of strains that were independently generated by classical strain improvement allows the positive mutations from each of the parental strains to be combined and permutated. Simultaneously, the deleterious mutations of one parent strain can be replaced with the wildtype sequence of the other parent strain.

REFERENCES

[1] Stemmer, W. P. C. (1995) Searching sequence space. *Bio/Technology* 13: 549-553.

[2] Stemmer, W. P. C. (1994) DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA* 91: 10747-10751.

[3] Stemmer, W. P. C. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* 370: 389-391.

[4] Crameri, A., E. Whitehorn, E. Tate, and W. P. C. Stemmer (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nature Biotechnol.* 14: 315-319.

[5] Zhang, J., G. Dawes, and W. P. C. Stemmer (1997) Evolution of an effective fucosidase from a galactosidase by DNA shuffling and screening. *Proc. Natl. Acad. Sci. USA* 94: 4504-4509.

[6] Crameri, A., G. Dawes, E. Rodriguez, S. Silver, and W. P. C. Stemmer (1997) Molecular evolution of an arsenate detoxification pathway by DNA shuffling. *Nature Biotechnol.* 15: 436-438.

[7] Crameri, A., S.-A. Raillard, E. Bermudez, and W. P. C. Stemmer (1998) DNA shuffling of genes from diverse species accelerates directed evolution. *Nature* 391: 288-291.

[8] Stemmer, W.P.C., A. Crameri, K. D. Ha, T. M. Brennan, and H. L. Heyneker (1995) Single-step PCR assembly of a gene and a whole plasmid from large numbers of oligonucleotides. *Gene* 164: 49-53.

[9] Christians, F. C., L. Scapozza, A. Crameri, G. Folkers, and W. P. C. Stemmer (1999) Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nature Biotechnol.* 17: 259-264.

[10] Ness, J., M. Welch, L. Giver, M. Bueno, J. Cherry, T. Borchert, W. P. C. Stemmer, and J. Minshull (1999) Creation of a functionally diverse enzyme library by DNA family shuffling. *Nature Biotechnol.* 17: 893-896.

[11] Chang, C.-C., T. T. Chen, B. W. Cox, G. N. Dawes, W. P. C. Stemmer, J. Punnonen, and P. A. Patten (1999) Rapid evolution of a cytokine using molecular breeding. *Nature Biotechnol.* 17: 793-797.

[12] Soong, N.-W., L. Nomura, K. Pekrun, M. Reed, L. Sheppard, G. Dawes, and W. P. C. Stemmer (2000) Molecular breeding of viruses. *Nature Genetics* 25: 436-439.

[13] Powell, S. K., M. A. Kaloss, A. Pinkstaff, R. McKee, I. Burimski, M. Pensiero, E. Otto, W. P. C. Stemmer, and N.-W. Soong (2000) Breeding of retroviruses by DNA shuffling for improved stability and processing yields. *Nature Biotechnol.* 18: 1279-1282.

[14] Stemmer, W. P. C and N.-W. Soong (1999) Molecular breeding of viruses for targeting and other clinical properties. *Tumor Targeting* 4: 59-62.

[15] Zhang, Y.-X., K. Perry, V. A. Vinci, K. Powell, W. P. C. Stemmer, and S. B. del Cardayre (2002) Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 415: 644-646.

[16] Raillard, S., A. Krebber, Y. Chen, J. E. Ness, E. Bermudez, R. Trinidad, R. Fullem, C. Davis, M. Welch, J. Seffernick, L. P. Wackett, W. P. C. Stemmer, and J. Minshull (2001) Novel enzyme activities and functional plasticity revealed by recombining highly homologous enzymes. *Chemistry Biol.* 125: 1-9.

[Received February 7, 2002; accepted May 18, 2002]