Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biomolecular Engineering 22 (2005) 105–112
Biomolecular Engineering
www.elsevier.com/locate/geneanabioeng

Review

Strategies and computational tools for improving randomized protein libraries

Wayne M. Patrick [a,*], Andrew E. Firth [b]

[a] *Center for Fundamental and Applied Molecular Evolution, Emory University, 1510 Clifton Road, Atlanta GA 30322, USA*
[b] *Department of Biochemistry, University of Otago, P.O. Box 56, Dunedin, New Zealand*

Received 2 May 2005; received in revised form 20 June 2005; accepted 21 June 2005

Abstract

In the last decade, directed evolution has become a routine approach for engineering proteins with novel or altered properties. Concurrently, a trend away from purely 'blind' randomization strategies and towards more 'semi-rational' approaches has also become apparent. In this review, we discuss ways in which structural information and predictive computational tools are playing an increasingly important role in guiding the design of randomized libraries: web servers such as ConSurf-HSSP and SCHEMA allow the prediction of sites to target for producing functional variants, while algorithms such as GLUE, PEDEL and DRIVeR are useful for estimating library completeness and diversity. In addition, we review recent methodological developments that facilitate the construction of unbiased libraries, which are inherently more diverse than biased libraries and therefore more likely to yield improved variants.

Keywords: Directed evolution; Library; Molecular diversity; Random mutagenesis; Redundancy

Contents

* Corresponding author. Tel.: +1 404 727 6669; fax: +1 404 727 3452.
E-mail address: wpatric@emory.edu (W.M. Patrick).

1389-0344/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.bioeng.2005.06.001

1. Introduction: is blind better?

It has now been 7 years since Frances Arnold surveyed a series of advances and exhorted that ''blind is better'' [1] in the complicated world of protein engineering. In the intervening years we have witnessed a plethora of papers describing the directed evolution of an enormous variety of proteins, the development of new techniques for generating molecular diversity, and the arrival of commercially available kits for random mutagenesis (reviewed in [2–4]). Each has served to emphasize the power of mimicking Darwinian evolution in vitro to alter the properties and functions of proteins, and to address fundamental questions in biochemistry and evolutionary biology.

The amount of molecular diversity that it is possible for nature to generate is staggering: an oft-quoted example is that there are more potential 100-residue protein sequences ($20^{100} \approx 10^{130}$) than there are atoms in the observable Universe ($\sim 10^{80}$). The enormity of these numbers ensures that any laboratory-generated randomized library (i.e. $<10^{15}$ variants [5]) represents a minute proportion of the possible sequence-, structure- or function-space. It also highlights the need to target the diversity that we can produce experimentally, in such a manner that the probability of finding a clone displaying a desired property is increased. Consequently, a trend towards making blind 'irrational' design more rational has become apparent, even as methods such as error-prone PCR (epPCR) [6], oligonucleotide-directed randomization [7] and DNA shuffling [8,9] have become routine.

One factor accounting for this trend has been the rapid expansion of the RCSB Protein Data Bank (PDB). As the structures of target proteins or their homologues are elucidated, new computational tools that 'use evolution to guide evolution' are being developed (for an excellent review, see [10]). These algorithms suggest that reliable in silico predictors will soon be available to guide the laboratory researcher; some of these will be discussed in Section 2.

In Section 3, we will discuss the statistical limitations of biased libraries and review recent methodological advances aimed at removing bias and hence increasing library diversity. Finally, in Section 4, we will summarize user-friendly tools for guiding library design and statistically analysing library composition.

2. Structure-guided evolution

2.1. Targeting randomization

With the advent of structural genomics initiatives and the concomitant growth of the PDB, it is becoming increasingly likely that a protein engineer will know the structure of their target protein and/or its sequence homologues. Correspondingly, however, the fraction of PDB entries that lack full functional annotation is also predicted to grow rapidly [11]. A number of groups have therefore investigated approaches for the automated identification of functionally important residues (reviewed in [12]).

Of more specific relevance to purveyors of directed evolution, Wiederstein and Sippl have reported the use of knowledge-based potentials to detect the regions of proteins that are most likely to tolerate randomization without severely compromising stability [13]. Another appealing application is the ConSurf-HSSP web server [14]. On specifying a PDB code, the evolutionary conservation at each amino acid position of a target protein is assessed using the pre-calculated multiple sequence alignments of the Homology-derived Secondary Structure of Proteins (HSSP) Database [15]. The result is projected onto the protein's structure, generating an interactive, three-dimensional map that emphasizes conserved and variable regions (Fig. 1A).

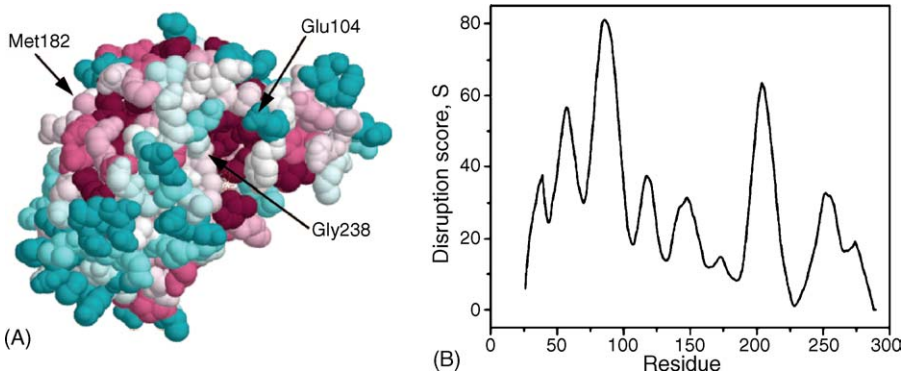

Fig. 1. Structural information to guide the directed evolution of TEM-1 β-lactamase. (A) ConSurf-HSSP mapping of evolutionary conservation, projected onto the Van der Waals surface of PDB entry 1BTL [57]. More conserved residues are in darker shades of red, more variable residues are in darker shades of blue and a sulfate ion in the active site is partly visible below Gly238. Three residues mutated in an improved variant (TEM-52) are indicated. Note that they range from being conserved but distant from the active site (Met182), to highly variable and adjacent to the active site (Glu104). (B) SCHEMA disruption profile for hybrids generated by recombining TEM-1 and PSE-4 (PDB code 1G68 [58]). The disruption score, *S*, was calculated using a window size of 14 and a distance cut-off for interaction of 4.5 Å [21]. Residues are numbered according to the TEM-1 sequence.

While tools such as ConSurf-HSSP provide insights into the structure and evolution of target proteins, there is little consensus on how best to use this information. Mapping functionally important regions tends to identify those parts of the protein that are most likely to abrogate parental function when mutated. This may be a desired outcome—for example, removing the conserved catalytic nucleophile of a glycosidase can destroy its hydrolysis activity but impart glycosynthase activity [16]. In contrast, many directed evolution experiments have identified mutations in non-conserved regions far from the active site that, nevertheless, affect properties such as activity and thermostability [17]. This dichotomy is illustrated in TEM-1 β-lactamase, where a combination of three mutations (E104K/M182T/G238S) that improve activity has been identified by directed evolution and in clinical isolate TEM-52 [18]. As shown in Fig. 1A, two of these mutations are in evolutionarily variable residues that line the active site, while the third, M182T, is distant but at a conserved position.

Kazlauskas and co-workers have suggested that the majority of variants in an epPCR library are statistically likely to contain distant mutations, such as M182T, simply because most residues in a protein lie far from the active site. They demonstrated that targeting randomization to the substrate binding site of the *Pseudomonas fluorescens* lipase yielded a greater proportion of variants with improved enantioselectivity than whole-gene epPCR [19]. However, an attempt to generalize this finding across a range of enzymes and properties painted a more confusing picture [17], especially as the analysis was dominated by site-directed mutagenesis experiments in which inactive variants (presumably) went unreported.

2.2. Guiding recombination

The most dramatic examples of directed evolution have all utilized in vitro recombination. This has spurred the development of a number of computational models for assessing the relative fitness of recombined variants, and therefore for predicting optimal crossover locations. The first – and most user-friendly – of these to be described was the SCHEMA algorithm [20], which scans a target structure and sums interactions within a user-defined primary sequence window. It then determines the number of these interactions that are disrupted in creating a hybrid protein, yielding a profile in which minima are hypothesized to correspond to ideal crossover sites (Fig. 1B). SCHEMA has been used to guide recombination of the β-lactamases TEM-1 and PSE-4 [21,22] and the haem domains of the cytochrome P450 enzymes CYP102A1 and CYP102A2 [23]. A correlation between minimizing the SCHEMA disruption score and maximizing chimera function was observed for both model systems. However, when crossovers were directed towards both SCHEMA minima and maxima, it was noteworthy that 13 of the 23 functional β-lactamase hybrids incorporated crossovers at maxima (i.e. the peaks at residues 84 and 204 in Fig. 1B). This led the authors to suggest that profile minima are not always a good guide for predicting optimal crossover sites [21], which may reflect the inability of the algorithm to account for interactions between residues that are distal in the primary sequence of the protein.

Maranas and co-workers have extended the SCHEMA concept to assess both favourable and unfavourable residue interactions in hybrid proteins (instead of scoring all contacts between residues from different parents as disruptive) and have incorporated data from multiple sequence alignments to recognise co-varying but distant residues [24–26]. While the resulting algorithms are correspondingly more CPU-intensive than SCHEMA, they have shown promise in predicting preferred crossover sites and, in the case of FamClash [26], ranking the activities of the resulting hybrids. However, Govindarajan et al. have also provided a caveat on over-emphasizing the importance of co-varying residues: when they recombined 100 amino acid pairs that appeared to co-vary in 15 subtilisin enzymes, only 7 were found to reflect true functional constraints [27].

2.3. Outlook

A key advantage of the directed evolution approach is that there is no absolute prerequisite for structural information. However, there is also no doubt that having a high-resolution structure of the target greatly facilitates a successful engineering strategy. An excellent recent example was provided by Acharya et al., who used structural considerations to rationalize mutations generated by epPCR, to discard those that appeared deleterious, and ultimately to construct a lipase variant with a 300-fold increase in its half-life at elevated temperatures [28]. To obtain such a marked improvement in thermostability from two rounds of asexual epPCR and a relatively low-throughput screen (2000 colonies per round) illustrates the benefits of a structure-guided, semi-rational approach to engineering.

The computational tools for guiding directed evolution are at a nascent stage of development—while extremely promising, they are not yet able to provide the average laboratory researcher with definite and straightforward predictors of regions to target for randomization or recombination. The current algorithms are also primarily geared towards predicting the retention of parental function; strategies for altering a given function or property in a specified manner are clearly going to prove much more difficult to predict. However, the gap between computation and experimentation is closing rapidly, and iterating in silico and in vitro approaches to evolution will undoubtedly become a prevalent and powerful strategy.

3. Methods for reducing redundancy

A rare, improved variant is most likely to be found in a maximally diverse (and therefore minimally biased) library.

Ideally, every gene in any given library of DNA sequences should encode a unique protein. Unfortunately, however, the unbalanced mutation spectra of error-prone polymerases, biases in the locations of crossovers during fragment reassembly, and the inherent degeneracy of the genetic code itself all act to reduce the amount of useful diversity found in randomized libraries. This realization has led a number of groups to propose methods for removing unwanted redundancy. We review some of these below.

3.1. Oligonucleotide-directed randomization

Conventional methods for codon randomization employ NNN-, NNB-, NNK- or NNS-containing oligonucleotides (N: A/C/G/T; B: C/G/T; K: G/T; S: G/C), as these combinations each encode all 20 amino acids. Incorporating the full degeneracy of the genetic code into a library by using NNN codons has severe limitations (Table 1). A significant fraction of the library will contain premature termination codons, especially if multiple codons are randomized. Further, the most common protein variants (containing combinations of Arg, Leu and Ser, each of which are encoded by six codons) will be vastly over-represented compared to the rarest (with Met or Trp at each randomized position). While this is likely to be deleterious, and to complicate downstream screening or selection [29], amino acids with more synonymous codons do tend to be better tolerated in amino acid substitutions [30], so it is also possible that a bias towards variants containing Arg, Leu and Ser, etc. may result in more functional variants than otherwise.

A superior strategy for improving library quality and diversity is to use reduced codon sets: NNB codons have the smallest probability of encoding a stop codon (one in 48), while NNK and NNS codons minimize the over-representation of the commonest variants (Table 1). It is noteworthy that the amino acid distributions resulting from the use of NNK and NNS codons are identical; however, codon usage preferences in *Escherichia coli* and especially in *S. cerevisiae* suggest that NNK is generally a better option for libraries maintained in these two microorganisms.

Recently, Hine and co-workers have described a method for constructing maximally diverse libraries in which the degeneracy of the genetic code is completely negated [29,31]. In their maximum efficiency (MAX) randomization protocol, sets of 20 primers containing the preferred codon for each amino acid in *E. coli* are hybridized to a template strand containing fully randomized codons (NNN or NNK) at the desired positions. The template strand is replaced by PCR, resulting in a library containing no premature stop codons and an equal representation of each amino acid (Table 1). While the current protocol requires the synthesis of 20 oligonucleotides per randomized position, the availability of premixed pools of trinucleotide phosphoramidites (from Glen Research) suggests an economical, one-pot method for synthesizing all 20 MAX oligonucleotides simultaneously. Despite not enabling the randomization of multiple adjacent codons, MAX therefore represents an optimal approach to oligonucleotide-directed randomization.

3.2. Error-prone PCR

Error-prone PCR is commonly used to generate diversity at any position in a target gene, and the original protocol [6] has been modified extensively to afford control over the nucleotide substitution rate. However, library diversity can be adversely affected by biases in the mutation spectra produced by various polymerases under error-prone conditions [32]. We suggested previously that polymerases such as Strategene's Mutazyme® – which in combination with *Taq* DNA polymerase displays a near-uniform mutational spectrum – would provide a simple method for offsetting this bias [33]; this approach has now been validated experimentally [34,35]. Site-directed mutants of the *Pyrococcus furiosus* DNA polymerase have also been investigated for their utility in epPCR [36].

As with oligonucleotide-directed randomization, a more difficult source of bias to address in epPCR libraries results from the degeneracy of the genetic code. Although multiple substitutions in the same codon are observed, their occurrence is infrequent and this effects which amino acid substitutions are likely to be accessed [3]. Indeed, on average a single base substitution in a sense codon allows only 5.7 alternative amino acids, rising to 15.7 amino acids when two bases are substituted [37] (although, on the plus side, amino acids accessible by a single base substitution are, on average, slightly more likely to be tolerated than those accessible by two or three mutations [38]). Moreover, attempts to incorporate non-synonymous mutation rates into

Table 1
The effects of randomized codon selection on library redundancy

Codon	P(stop) per codon	P(stop), 5 codons[a]	P(Ser)/P(Trp)[b]	$P([Ser]_5)/P([Trp]_5)$[c]
NNN	0.047	0.213	6	7776
NNB	0.021	0.100	5	3125
NNK/NNS	0.031	0.147	3	243
MAX	0	0	1	1

[a] Probability of encoding at least one stop codon when five codons are randomized.
[b] Ratio of the most common amino acid variant (e.g. Ser) to the rarest amino acid variant (e.g. Trp).
[c] Ratio of the most common amino acid variant to the rarest variant in a library when five codons are randomized.

calculations of library diversity are made inaccurate by the sequence dependence of these mutation rates. This would appear to be an unavoidable outcome of the epPCR process, although in an interesting recent result, Deem and co-workers showed that the choice of codon in the template sequence will affect the probabilities of synonymous, conservative and non-conservative mutations [39]. The implication, then, is that site-directed mutagenesis of the template prior to error-prone amplification may afford some control over the mutational spectrum obtained on randomization, although this approach remains untested experimentally.

A third source of bias that has been proposed for epPCR libraries results from the uneven representation of those random variants generated in the early cycles of the amplification process compared to those generated late [3]. We note that this source of bias is usually limited by the inefficiency of the subsequent library sub-cloning and/or transformation steps: a 'typical' epPCR might consist of amplifying $\sim 10^9$ template DNA molecules (e.g. 1 ng of a 1000 bp gene) to yield $\sim 10^{15}$ molecules (1 mg of PCR product), yet only a tiny fraction of these sequences will be incorporated into the final, cloned library (which will almost always consist of $<10^9$ variants). Provided the final library size is less than the number of template molecules in the epPCR, any 'amplification bias' is rendered insignificant.

A more controllable source of redundancy in epPCR libraries stems from over-sampling those variants that differ from the parental sequence in only a few (e.g. <3) positions. Most libraries are constructed using low rates of mutation in order to minimize the accumulation of deleterious mutations. However, it has been shown that significant fractions of such libraries will contain the unmutated parent sequence and multiple copies of all possible one- and two-substitution variants, at the expense of clones with greater numbers of mutations [33,40]. A viable method for increasing the diversity represented in any given epPCR library is therefore to increase the average mutation rate. The algorithm Program for Estimating Diversity in Error-prone PCR Libraries (PEDEL, vide infra) [33,41] enables an optimal error rate to be estimated such that redundancy is minimized but that those sequences with one to two mutations remain well represented.

3.3. In vitro recombination

The protocols for in vitro recombination – DNA shuffling and the staggered extension process (StEP PCR) [42] – are those that are most prone to yield severely biased libraries. Extensive in silico modelling of DNA shuffling has demonstrated that crossovers accumulate in regions of sequence identity and that there are intrinsic trade-offs between annealing temperature, template concentration and the efficiency of reassembling full-length products [43–45]. In analogy to in vivo recombination, tightly linked polymorphisms are unlikely to be recombined in daughter sequences and this can result in a failure to sample all possible diversity, even in large libraries [33]. The related methods 'degenerate homoduplex recombination' [46], 'synthetic shuffling' [47] and 'assembly of designed oligonucleotides' [48] therefore represent significant advances. Each involves the use of degenerate, overlapping oligonucleotides encoding all parental diversity as substrates in the reassembly reaction. The result is unbiased libraries in which all shuffled daughter sequences are equally likely, independent of linkage in the parents.

The proportion of 'junk' products is even greater in homology-independent recombination protocols such as incremental truncation [49,50], sequence homology-independent protein recombination [51] and SCRATCHY [52]. Useful library diversity is reduced significantly by the presence of non-parental sized hybrids and out-of-frame crossovers. This necessitates the inclusion of size selection steps in each protocol, and has led to the construction of a plasmid vector specifically for enriching libraries by removing frame-shifted variants [53].

4. The statistics of library construction and analysis

Computational algorithms and new experimental methods provide the laboratory researcher with tools to target randomization and to construct diverse libraries. While these are undoubtedly useful, there remains something of a disconnect between 'designing the perfect library on paper' and 'constructing the perfect library in the lab'. This is accentuated by the realization that – short of sequencing thousands or millions of variants – it is impossible to provide a complete description of the diversity contained in most randomized libraries. In this section, we review statistical approaches to assess library completeness and diversity.

4.1. Estimating library completeness

As discussed above, a maximally diverse library is one in which every possible variant is represented once and only once. In reality, however, the statistics of sampling a large pool of randomized variants ensure that this is not the case, even when the underlying method of library construction is unbiased. We have previously described a suite of programs – GLUE, PEDEL and Diversity Resulting from In Vitro Recombination (DRIVeR) – for estimating various library statistics [33]. We have also recently constructed easy-to-use web interfaces for these programs [41]. The programs may be used to guide experimental design (e.g. to determine what library size is required to sample a given amount of diversity, or to optimize the mutation rate to maximize diversity), as well as to interpret results (e.g. by estimating how many unique sequences are represented in a given library).

GLUE is applicable to any methodology where all possible variants are equally likely (such as oligonucleotide-directed randomization, MAX randomization and synthetic

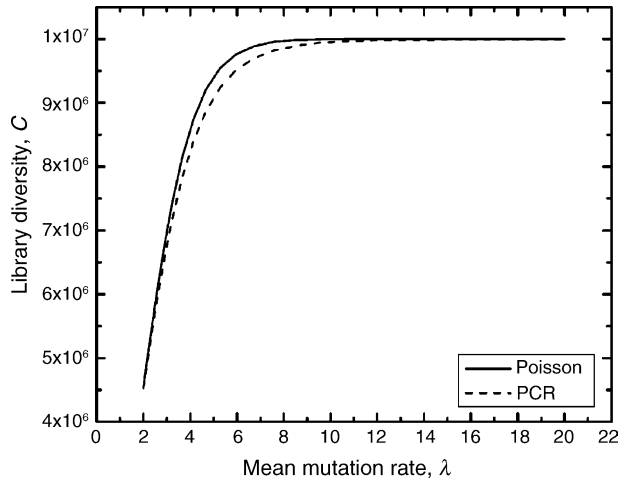

Fig. 2. Use of PEDEL to estimate how the total number of distinct variants, C, in an epPCR library varies with mean mutation rate, λ. In this hypothetical case, the parental gene is assumed to be 600 bp in length and the library size has been fixed at 10^7 clones. For most values of λ, the PCR distribution [40,54] (dashed line, calculated assuming 30 PCR cycles and a PCR efficiency of 0.5) is closely approximated by a Poisson distribution (solid line). In this example, it can be seen that increasing λ to >5 mutations per daughter sequence will ensure that almost every daughter variant is unique, leading to a maximally diverse library.

shuffling). On entering the total number of possible variants, the user can calculate the expected number of unique variants in a given library, or the library size required to sample a given fraction of all variants.

PEDEL can be used to calculate the expected number of distinct sequences present in an epPCR library, given the library size, mean mutation rate and parent sequence length. The web interface may be used to produce plots of the expected diversity of the library as a function of library size and of mutation rate (Fig. 2). The program can also produce statistics and plots giving a breakdown of the library into those sequences with exactly 0, 1, 2, 3, etc. mutations. An underlying assumption in the original implementation of PEDEL was that the number of mutations per daughter sequence follows a Poisson distribution [33,41]. Recently, however, Arnold and co-workers have revisited the work of Sun [54] and provided experimental evidence that the mutational distribution in epPCR is not Poisson [40]. The effect on estimated library diversity is minimal (<5%) under typical experimental conditions (Fig. 2), although the compositions of individual sub-libraries can be altered more significantly. The PEDEL web interface has been updated to accommodate this non-Poisson 'PCR distribution'.

DRIVeR can be used to estimate the number of distinct daughter sequences in a library constructed by recombining two highly-homologous parent sequences, given the library size, the parent sequence length, the mean crossover rate and the positions of the nucleotides (or amino acids) that vary between parents. The user can also produce plots of the library diversity as a function of crossover rate or library size.

Even in cases where the underlying assumptions are not justified (such as when *Taq* alone is used to construct a biased epPCR library), GLUE, PEDEL and DRIVeR provide an upper estimate of the proportion of all possible variants that will be represented in a given library. This in turn allows the user to make an informed judgement on whether their library is large enough to contain sufficient diversity.

Once a library has been constructed and a sample of members that survive screening or selection have been sequenced, it becomes possible to estimate the rate of false positives. An excellent example was provided by Doyle and co-workers, who sequenced 12 of 300 positive clones in a selection for novel ligand–receptor pairs [55]. While all 12 were full-length with designed mutations, they stopped short of assuming that their entire set of 300 clones was free from false positives and instead used binomial statistics to put a lower bound on the useful diversity in the selected pool (in this case, that at least 78% of the clones were true positives). This sort of careful analysis represents a more accurate description of experimental results, and should prove useful in rationalizing experimental outcomes.

4.2. How big is big enough?

A related problem to that of library completeness (i.e. 'How many distinct variants does my library contain?') is to estimate the probability that a library is 100% complete (i.e. 'What is the probability that I have *every* distinct variant?'). GLUE [33,41] can be used to calculate this, and therefore to provide a target library size to aim for in the laboratory, in cases where all variants are equally probable. More recently, an identical result has been derived independently [56], albeit in a rearranged and less readily calculable form.

Finally, it has been noted that propagating a library by sub-cloning or by transforming a new host strain introduces an additional way to lose diversity through under-sampling [56]. In effect, a second round of library completeness calculations must be carried out, and one or both of the original and propagated libraries will need to be increased in size to retain diversity.

5. Conclusions

It seems that 'blind' is no longer the best way to do directed evolution. Instead, an abundance of structural information and increasingly powerful computational algorithms now enable the laboratory researcher to approach library construction and analysis in a targeted, semi-rational manner. Web-based tools such as ConSurf-HSSP and SCHEMA serve to guide library design, albeit with some caveats. Techniques such as MAX randomization for saturation mutagenesis, dual-polymerase approaches to epPCR and the synthetic shuffling methods for in vitro recombination, allow the construction of minimally biased, and therefore maximally diverse, libraries. The programs

Table 2
Web servers and/or source code for applications described in the text

Application	URL	Comments	Reference
ConSurf-HSSP	http://consurf-hssp.tau.ac.il/	Displays evolutionary conservation of each amino acid. Currently, only compatible with Internet Explorer	[14]
SCHEMA	http://www.mayo.caltech.edu/~pcs/initialschemapage.html	Convenient web interface for predicting crossover locations	[20]
FamClash	http://fenske.che.psu.edu/faculty/cmaranas/	Source code available on request	[26]
GLUE/PEDEL/DRIVeR	http://guinevere.otago.ac.nz/stats.html	Web interfaces for estimating library completeness and diversity	[41]

GLUE, PEDEL and DRIVeR estimate the number of distinct sequence variants in a library, or provide a library size to aim for to encode a specified proportion of all possible variants. Taken together, these strategies and tools increase the probability of finding a rare, improved variant and hint that directed evolution methodologies will provide many more success stories in the future.

Supplementary information

Links to the web-based algorithms and applications discussed above are summarized in Table 2, and are also compiled at the following website: http://www.famecenter.emory.edu/relatedLinks.htm.

Acknowledgements

The authors thank Monica Gerth for her critical reading of this manuscript. A.E.F. acknowledges funding from the New Zealand Foundation for Research, Science and Technology, grant number UOOX0304. The publication costs of this article were generously met by Dr. Ichiro Matsumura and the NIH/NIAID (1 R21AI054602-01).

References

[1] Arnold FH. When blind is better: protein design by evolution. Nat Biotechnol 1998;16:617–8.
[2] Lutz S, Patrick WM. Novel methods for directed evolution of enzymes: quality, not quantity. Curr Opin Biotechnol 2004;15:291–7.
[3] Neylon C. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. Nucleic Acids Res 2004;32:1448–59.
[4] Williams GJ, Nelson AS, Berry A. Directed evolution of enzymes for biocatalysis and the life sciences. Cell Mol Life Sci 2004;61:3034–46.
[5] Roberts RW, Ja WW. In vitro selection of nucleic acids and proteins: what are we learning? Curr Opin Struct Biol 1999;9:521–9.
[6] Cadwell RC, Joyce GF. Randomization of genes by PCR mutagenesis. PCR Methods Appl 1992;2:28–33.
[7] Hermes JD, Parekh SM, Blacklow SC, Koster H, Knowles JR. A reliable method for random mutagenesis: the generation of mutant libraries using spiked oligodeoxyribonucleotide primers. Gene 1989; 84:143–51.
[8] Stemmer WP. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. Proc Natl Acad Sci USA 1994;91:10747–51.
[9] Stemmer WP. Rapid evolution of a protein in vitro by DNA shuffling. Nature 1994;370:389–91.
[10] Moore GL, Maranas CD. Computational challenges in combinatorial library design for protein engineering. AIChE J 2004;50:262–72.
[11] Brenner SE. A tour of structural genomics. Nat Rev Genet 2001; 2:801–9.
[12] Minshull J, Ness JE, Gustafsson C, Govindarajan S. Predicting enzyme function from protein sequence. Curr Opin Chem Biol 2005;9:202–9.
[13] Wiederstein M, Sippl MJ. Protein sequence randomization: efficient estimation of protein stability using knowledge-based potentials. J Mol Biol 2005;345:1199–212.
[14] Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. Proteins 2005;58:610–7.
[15] Holm L, Sander C. Protein folds and families: sequence and structure alignments. Nucleic Acids Res 1999;27:244–7.
[16] Williams SJ, Withers SG. Glycosynthases: mutant glycosidases for glycoside synthesis. Aust J Chem 2002;55:3–12.
[17] Morley KL, Kazlauskas RJ. Improving enzyme properties: when are closer mutations better? Trends Biotechnol 2005;23:231–7.
[18] Orencia MC, Yoon JS, Ness JE, Stemmer WP, Stevens RC. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. Nat Struct Biol 2001;8:238–42.
[19] Park S, Morley KL, Horsman GP, Holmquist M, Hult K, Kazlauskas RJ. Focusing mutations into the *P. fluorescens* esterase binding site increases enantioselectivity more effectively than distant mutations. Chem Biol 2005;12:45–54.
[20] Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. Protein building blocks preserved by recombination. Nat Struct Biol 2002; 9:553–8.
[21] Meyer MM, Silberg JJ, Voigt CA, Endelman JB, Mayo SL, Wang ZG, et al. Library analysis of SCHEMA-guided protein recombination. Protein Sci 2003;12:1686–93.
[22] Hiraga K, Arnold FH. General method for sequence-independent site-directed chimeragenesis. J Mol Biol 2003;330:287–96.
[23] Otey CR, Silberg JJ, Voigt CA, Endelman JB, Bandara G, Arnold FH. Functional evolution and structural conservation in chimeric cytochromes p450: calibrating a structure-guided approach. Chem Biol 2004;11:309–18.
[24] Moore GL, Maranas CD. Identifying residue–residue clashes in protein hybrids by using a second-order mean-field approach. Proc Natl Acad Sci USA 2003;100:5091–6.
[25] Saraf MC, Moore GL, Maranas CD. Using multiple sequence correlation analysis to characterize functionally important protein regions. Protein Eng 2003;16:397–406.
[26] Saraf MC, Horswill AR, Benkovic SJ, Maranas CD. FamClash: a method for ranking the activity of engineered enzymes. Proc Natl Acad Sci USA 2004;101:4142–7.
[27] Govindarajan S, Ness JE, Kim S, Mundorff EC, Minshull J, Gustafsson C. Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. J Mol Biol 2003;328:1061–9.
[28] Acharya P, Rajakumara E, Sankaranarayanan R, Rao NM. Structural basis of selection and thermostability of laboratory evolved *Bacillus subtilis* lipase. J Mol Biol 2004;341:1271–81.

[29] Hughes MD, Nagel DA, Santos AF, Sutherland AJ, Hine AV. Removing the redundancy from randomised gene libraries. J Mol Biol 2003; 331:973–9.
[30] Dufton MJ. The significance of redundancy in the genetic code. J Theor Biol 1983;102:521–6.
[31] Hughes MD, Zhang ZR, Sutherland AJ, Santos AF, Hine AV. Discovery of active proteins directly from combinatorial randomized protein libraries without display, purification or sequencing: identification of novel zinc finger proteins. Nucleic Acids Res 2005;33:e32.
[32] Shafikhani S, Siegel RA, Ferrari E, Schellenberger V. Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. Biotechniques 1997;23:304–10.
[33] Patrick WM, Firth AE, Blackburn JM. User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. Protein Eng 2003;16:451–7.
[34] Rowe LA, Geddie ML, Alexander OB, Matsumura I. A comparison of directed evolution approaches using the β-glucuronidase model system. J Mol Biol 2003;332:851–60.
[35] Vanhercke T, Ampe C, Tirry L, Denolf P. Reducing mutational bias in random protein libraries. Anal Biochem 2005;339:9–14.
[36] Biles BD, Connolly BA. Low-fidelity *Pyrococcus furiosus* DNA polymerase mutants useful in error-prone PCR. Nucleic Acids Res 2004;32:e176.
[37] Hermes JD, Blacklow SC, Knowles JR. Searching sequence space by definably random mutagenesis: improving the catalytic potency of an enzyme. Proc Natl Acad Sci USA 1990;87:696–700.
[38] Wong JT. Role of minimization of chemical distances between amino acids in the evolution of the genetic code. Proc Natl Acad Sci USA 1980;77:1083–6.
[39] Tan T, Bogarad LD, Deem MW. Modulation of base-specific mutation and recombination rates enables functional adaptation within the context of the genetic code. J Mol Evol 2004;59:385–99.
[40] Drummond DA, Iverson BL, Georgiou G, Arnold FH. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. J Mol Biol 2005;350:806–16.
[41] Firth AE, Patrick WM. Statistics of protein library construction. Bioinformatics 2005;21:3314–5.
[42] Zhao H, Giver L, Shao Z, Affholter JA, Arnold FH. Molecular evolution by staggered extension process (StEP) in vitro recombination. Nat Biotechnol 1998;16:258–61.
[43] Moore GL, Maranas CD, Lutz S, Benkovic SJ. Predicting crossover generation in DNA shuffling. Proc Natl Acad Sci USA 2001;98: 3226–31.
[44] Moore GL, Maranas CD. Predicting out-of-sequence reassembly in DNA shuffling. J Theor Biol 2002;219:9–17.
[45] Maheshri N, Schaffer DV. Computational and experimental analysis of DNA shuffling. Proc Natl Acad Sci USA 2003;100:3071–6.
[46] Coco WM, Encell LP, Levinson WE, Crist MJ, Loomis AK, Licato LL, et al. Growth factor engineering by degenerate homoduplex gene family recombination. Nat Biotechnol 2002;20:1246–50.
[47] Ness JE, Kim S, Gottman A, Pak R, Krebber A, Borchert TV, et al. Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. Nat Biotechnol 2002; 20:1251–5.
[48] Zha D, Eipper A, Reetz MT. Assembly of designed oligonucleotides as an efficient method for gene recombination: a new tool in directed evolution. ChemBioChem 2003;4:34–9.
[49] Ostermeier M, Shim JH, Benkovic SJ. A combinatorial approach to hybrid enzymes independent of DNA homology. Nat Biotechnol 1999;17:1205–9.
[50] Lutz S, Ostermeier M, Benkovic SJ. Rapid generation of incremental truncation libraries for protein engineering using α-phosphothioate nucleotides. Nucleic Acids Res 2001;29:e16.
[51] Sieber V, Martinez CA, Arnold FH. Libraries of hybrid proteins from distantly related sequences. Nat Biotechnol 2001;19:456–60.
[52] Lutz S, Ostermeier M, Moore GL, Maranas CD, Benkovic SJ. Creating multiple-crossover DNA libraries independent of sequence identity. Proc Natl Acad Sci USA 2001;98:11248–53.
[53] Gerth ML, Patrick WM, Lutz S. A second-generation system for unbiased reading frame selection. Protein Eng Des Sel 2004;17: 595–602.
[54] Sun F. The polymerase chain reaction and branching processes. J Comput Biol 1995;2:63–86.
[55] Schwimmer LJ, Rohatgi P, Azizi B, Seley KL, Doyle DF. Creation and discovery of ligand–receptor pairs for transcriptional control with small molecules. Proc Natl Acad Sci USA 2004;101:14707–12.
[56] Bosley AD, Ostermeier M. Mathematical expressions useful in the construction, description and evaluation of protein libraries. Biomol Eng 2005;22:57–61.
[57] Jelsch C, Mourey L, Masson JM, Samama JP. Crystal structure of *Escherichia coli* TEM1 β-lactamase at 1.8 Å resolution. Proteins 1993;16:364–83.
[58] Lim D, Sanschagrin F, Passmore L, De Castro L, Levesque RC, Strynadka NC. Insights into the molecular basis for the carbenicillinase activity of PSE-4 β-lactamase from crystallographic and kinetic studies. Biochemistry 2001;40:395–402.