Published online December 7, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 21 ***6407–6413***
doi:10.1093/nar/gkh977

Predicting oligonucleotide-directed mutagenesis failures in protein engineering

Christopher D. Wassman[1], Phillip Y. Tam[2], Richard H. Lathrop[1,3] and Gregory A. Weiss[2,4,*]

[1]Department of Computer Science, [2]Department of Chemistry, [3]Department of Biomedical Engineering and [4]Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697-2025, USA

Received August 28, 2004; Revised November 3, 2004; Accepted November 15, 2004

ABSTRACT

Protein engineering uses oligonucleotide-directed mutagenesis to modify DNA sequences through a two-step process of hybridization and enzymatic synthesis. Inefficient reactions confound attempts to introduce mutations, especially for the construction of vast combinatorial protein libraries. This paper applied computational approaches to the problem of inefficient mutagenesis. Several results implicated oligonucleotide annealing to non-target sites, termed 'cross-hybridization', as a significant contributor to mutagenesis reaction failures. Test oligonucleotides demonstrated control over reaction outcomes. A novel cross-hybridization score, quickly computable for any plasmid and oligonucleotide mixture, directly correlated with yields of deleterious mutagenesis side products. Cross-hybridization was confirmed conclusively by partial incorporation of an oligonucleotide at a predicted cross-hybridization site, and by modification of putative template secondary structure to control cross-hybridization. Even in low concentrations, cross-hybridizing species in mixtures poisoned reactions. These results provide a basis for improved mutagenesis efficiencies and increased diversities of cognate protein libraries.

INTRODUCTION

In protein engineering, mutagenesis is used to alter and dissect protein function (1). Example applications include antibodies (2,3), enzymes (4–6), other receptor–ligand interactions (7,8), binding partner discovery (9) and epitope mapping (7,8,10). The technique requires the introduction of mutations into a gene encoding the protein target, resulting in a modified protein and a potentially altered function or activity. Perhaps due to an incomplete understanding of the relationship between protein sequence and structure/function, successful approaches to protein engineering often apply library-based mutagenesis, screens and selections.

Protein engineering uses many different molecular display scaffolds, including phage (11,12), yeast (13), mRNA (14), ribosome (15) and plasmid display (16,17). All link individual members of a protein library with encoding information. This linkage expedites identification of library members and can allow rapid molecular evolution through mutagenesis of the encoding information.

Mutations for protein engineering can be introduced into specific sites (rational protein design), randomly (stochastic search), or by a combination of the two (guided protein evolution). Stochastic techniques rely upon DNA copy errors introduced by PCR (18–21), propagation in a DNA repair-deficient strain of bacteria (13), recombination (22–24) or other techniques. Site-directed mutagenesis relies upon hybridization of a mutation-encoding oligonucleotide to a specific location in the gene encoding the protein of interest. This hybridization step is potentially problematic, and is the focus of this paper.

Site-directed combinatorial mutagenesis depends upon highly efficient chemical synthesis of degenerate oligonucleotide mixtures (also called degenerate sequences). This results in the construction and subsequent display of many different mutant proteins. In principle, increasing the diversity of protein libraries should increase the efficiency of protein engineering by exploring more protein function space. In practice, mutagenesis with diverse oligonucleotide mixtures presents technical challenges owing to many reasons, of which some are poorly understood.

Oligonucleotide-directed mutagenesis

The method for site-directed mutagenesis applied here, termed 'Kunkel mutagenesis', features hybridization of a mutation-encoding oligonucleotide to a target site on a uracil-doped template plasmid (25). The oligonucleotide consists of a variable region sandwiched between two annealing sequences that are complementary to the target site. Phosphorylated oligonucleotides are annealed to the template plasmid, followed by *in vitro* enzymatic synthesis of the complementary DNA strand. A successful reaction yields the desired product, covalently closed circular double-stranded DNA (ccc-dsDNA, termed '*good*', Figure 1). Two undesirable side products consume template DNA and degrade the reaction efficiency. Nicked DNA (termed '*bad*') results from failed phosphorylation or ligation reactions, and is often observed as a minor side product. Strand-displaced DNA (termed '*ugly*') results from aberrant DNA polymerization. During mutagenesis, *ugly* is always observed, and consumes a significant fraction of the

*To whom correspondence should be addressed. Tel: +1 949 824 5566; Fax: +1 949 824 9920; Email: gweiss@uci.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

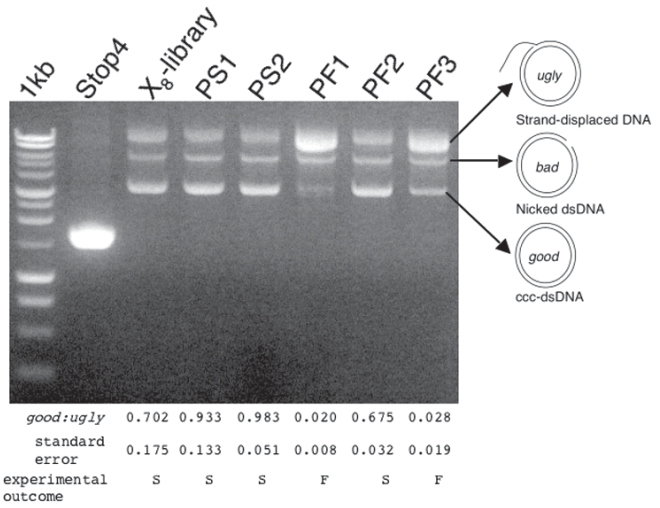

Figure 1. Mutagenesis with predicted successful and failure oligonucleotides. Mutagenesis reactions on template plasmid Stop4 used control (X_8-library), predicted successful (PS1 and PS2) and predicted failure (PF1, PF2 and PF3) oligonucleotides. Arrows indicate *good*, *bad* and *ugly* products, as described in the text. *Good:ugly* ratios are an average of two mutagenesis reactions with the indicated standard error. All depicted gels feature electrophoresis of approximately equal quantities of each mutagenesis reaction or template, with 1 kb indicating a 1 kilobase DNA ladder and the experimental outcome as described in Table 1. Each band indicates a different product, with consequent different levels of ethidium bromide intercalation. Therefore, integrating the intensities of the bands for comparison between reactions is potentially unreliable, and may not accurately represent the total amount of DNA.

starting materials. The mixture of *good*, *bad* and *ugly* is transformed into a bacteria host with intact uracil *N*-glycosylase, leading to uracil-targeted degradation of the template. The efficiency of a Kunkel mutagenesis reaction depends upon the fraction of completely filled-in and ligated reaction products, the *good*. In general, mutagenesis failures are associated with high levels of *ugly* (26). This side product could cause poor mutagenesis due to inefficient transformation of *Escherichia coli* (27), leading to reduced protein library diversity.

During synthesis of diverse phage-displayed peptide libraries, we observed many oligonucleotide mixtures that consistently yielded poor mutagenesis efficiencies and minimal amounts of *good*. Oligonucleotide mixtures yielding both low and high efficiencies had identical annealing sequences and mutagenesis conditions. Variable region length and sequence showed no discernable association with the success or failure of the reaction. These observations, which appeared to defy simple explanations, led us to explore possible mechanisms for different outcomes from similar oligonucleotide mixtures.

A hypothesis for mutagenesis reaction failures

Many possible factors could contribute to the generation of *ugly*, including DNA secondary structure, reaction conditions (salt concentration, annealing temperature, etc.), oligonucleotide purity and cross-hybridization. Of these, cross-hybridization, defined here as oligonucleotide annealing to incorrect locations, emerged as a significant predictor of experimentally observed *ugly* levels. The potential for cross-hybridization to cause deleterious effects has been described for DNA computing (28), DNA-based nanotechnology (29), PCR primer design (30), microarrays (31,32) and siRNAs (33). Cross-hybridization in combinatorial oligonucleotide-directed mutagenesis, to date, has not been studied extensively.

Perhaps this is due to the enormous theoretical diversities involved. Oligonucleotide mixtures for typical phage-displayed libraries range in theoretical diversities from $\sim 10^6$ to $\sim 10^{27}$ unique DNA sequences. Therefore, exhaustive analysis of every sequence in an oligonucleotide mixture is time-prohibitive.

The goal of this paper is to elucidate one important source of difficulty in protein engineering: problems associated with the oligonucleotide hybridization step of mutagenesis. This paper provides an (i) efficient algorithm to compute a cross-hybridization score for any oligonucleotide mixture to any template plasmid, and (ii) experimental verification that cross-hybridization can decrease the mutagenesis reaction efficiency.

MATERIALS AND METHODS

Materials

Reagents. The following materials were purchased commercially: *Taq* DNA polymerase and 10× PCR buffer from Continental Laboratory Products; M13-KO7 helper phage from Amersham–Pharmacia Life Science; *E.coli* XL1-Blue from Stratagene; exonuclease I and shrimp alkaline phosphatase from US Biochemicals; BigDye® terminator v3.1 sequencing reagents from Applied Biosystems; T4 polynucleotide kinase, T4 DNA ligase and T7 DNA polymerase (unmodified) from New England Biolabs; oligonucleotides from Sigma–Genosys; and 1 kb ladder from Promega. All other general reagents were of molecular biology grade or higher from commercial sources.

All buffers and media were prepared using water purified from a Millipore Nanopure™ system and sterilized as necessary, either by autoclave or by filtration through a 0.22 µm pore size filter (Corning). All enzymatic reactions were performed with irrigation water (Phoenix Pharmaceutical, Inc.).

Oligonucleotides. PS, PF and X_8 oligonucleotides were purified by PAGE. Lyophilized oligonucleotides were redissolved to a concentration of 330 ng/µl. Sequencing primers M13-F1 and M13-R were diluted to working concentrations of 0.8 pmol/µl. Variable or modified regions below are highlighted in bold. Degenerate bases in X_8 are named in accordance with IUBMB convention.

X_8
5′-GCTACAAATGCCTATGCA**NNSNNSNNSNNSNNSNNSS-NNS**GGTGGAGGATCCGGA-3′
PS1 (Predicted Successful 1)
5′-GCTACAAATGCCTATGCA**AAGAAGAAGCTCGAGAACT-TGATC**GGTGGAGGATCCGGA-3′
PS2 (Predicted Successful 2)
5′-GCTACAAATGCCTATGCA**ACCTAGTAGCTCGAGTAGA-AGATC**GGTGGAGGATCCGGA-3′
PF1 (Predicted Failure 1)
5′-GCTACAAATGCCTATGCA**CGGGTGCGCATGATCGTGC-TCCTG**GGTGGAGGATCCGGA-3′

PF2 (Predicted Failure 2)
5′-GCTACAAATGCCTATGCA**ATGGCCCACTACGTGAACCA-GATC**GGTGGAGGATCCGGA-3′
PF3 (Predicted Failure 3)
5′-GCTACAAATGCCTATGCA**GTGGACCGCTTGCTGCACAT-GATC**GGTGGAGGATCCGGA-3′
Stop4Δ-fix
5′-CAAAGGGCGAAAAACCGTCATGGCCCACTACGTGAACC-AAAGTTTTTTGGGGTCGAGGTG-3′
Stop4A-fix
5′-CAAAGGGCGAAAAACCGTC**TTTTTTTTTT**ATGGCCCAC-TACGTGAACCA**TTTTTTTTTTT**AAGTTTTTTGGGGTCGAG-GTG-3′
SAV-F1 5′-TGTAAAACGACGGCCAGTCGAGCACTTCAC-CAACAA-3′
SAV-R2 5′-CAGGAAACAGCTATGACGACAACAACCATC-GCCC-3′
3-fwd 5′-TGTAAAACGACGGCCAGTTGGTGCGGATATCT-CGGTAG-3′
3-rev 5′-CAGGAAACAGCTATGACTTGTTGGTGAAGTGC-TCGTG-3′
7-fwd 5′-TGTAAAACGACGGCCAGTAACTGTGAATGCGC-AAACC-3′
7-rev 5′-CAGGAAACAGCTATGACATAAAGCGGGCCATG-TTAAG-3′
8-fwd 5′-TGTAAAACGACGGCCAGTTGAGCATCCTCTCT-CGTTTCA-3′
8-rev 5′-CAGGAAACAGCTATGACCGCTTTCTTCCCTTC-CTTTC-3′
M13-F1 5′-TGTAAAACGACGGCCAGT-3′
M13-R 5′-CAGGAAACAGCTATGAC-3′

Computing the cross-hybridization score

The cross-hybridization score can be computed for any template plasmid and for any oligonucleotide mixture. It is the odds ratio of correct hybridization to cross-hybridization, as estimated by the following equation:

$$\text{Score} = \frac{\sum_{\text{i}_{\text{correct}}} f_i \cdot \text{e}^{-\Delta G_i/RT}}{\sum_{\text{j}_{\text{cross-hybridization}}} f_j \cdot \text{e}^{-\Delta G_j/RT}}. \qquad \mathbf{1}$$

Here, the numerator sums over correct hybridizations, the denominator sums over cross-hybridizations, ΔG is the free energy of a given hybridization site, f counts its occurrences in the mixture, R is the molar gas constant, T is the temperature and $\text{e}^{-\Delta G/RT}$ is the Boltzmann probability weight (likelihood). To compute the cross-hybridization score of Equation 1 requires (i) identifying hybridization sites, (ii) counting their occurrences in the mixture, and (iii) estimating their free energies.

The first step applies regular expression matching to identify initial seed hybridization sites. Small contiguous stretches of Watson–Crick base pairing occur in most significant cross-hybridizations. The oligonucleotide mixture, represented as a regular expression by a sequence of IUBMB degenerate base codes, is matched against the template plasmid to yield a list of all exact 8mer matches. The running time to find all 8mer matches is proportional to the product of the lengths of the template and the oligonucleotide.

The next step is to extend and count occurrences of N-mer matches. Longer exact N-mer matches are created from $(N-1)$-mer matches sharing an exact $N-2$ overlap at the same template site. This process continues until all possible longer exact matches have been constructed and added to the list. Each template site may be covered by several different N-mers of different lengths that match with different parts of the oligonucleotide degenerate sequence. Equation 1 sums up the final list of all matches. The count (f in Equation 1) of each N-mer match is the diversity of the degenerate bases not fixed by the N-mer, minus the counts of every longer match that contains it. This can be computed efficiently, because the relationship of superstring containment induces a directed acyclic graph over N-mers, which can be traversed easily. The time complexity of this step is bounded by the square of the number of 8mer matches found in the first step.

In the final step, free energies are estimated using nearest neighbor thermodynamic parameters (34). The running time of this step is proportional to the sum of the N-mer lengths in the final list.

Equation 1 is an approximate quantity, which neglects thermodynamic contributions from the oligonucleotide mixture bases not specified by the N-mer match, dangling ends, base mismatches, loops and other sources. In turn, the algorithm computes an approximate estimate of Equation 1 by examining only the strongest hybridizations.

Source code of the algorithms described here is freely available to users upon email request to cwassman@uci.edu. The algorithms were implemented on the Java™ 2 Platform, JDK 1.3.1.

Computational design of PS and PF oligonucleotides

Oligonucleotides predicted to have strong cross-hybridization (Predicted Failure, PF, oligonucleotides) were identified from the list of N-mer matches described above. Three N-mer matches, each with a low free energy, were chosen as PF1, PF2 and PF3. Unspecified bases were replaced randomly within the oligonucleotide mixture.

Oligonucleotides predicted to have very weak cross-hybridization (Predicted Success, PS, oligonucleotides) were chosen by template walking. A probe sequence, randomly chosen from X_8 by fixing degenerate bases, was compared stepwise with the template plasmid. Steps were 300 bases in length with an overlap of 150 bases, so step boundary end effects were avoided. Hybridization was tested at each step by predicting minimal energy structures between the probe and template (34–36). Bases in the probe sequence with the most tendency to cross-hybridize were identified by measuring the energy-weighted frequency of base pairing. Those bases most susceptible to cross-hybridization were changed in the probe sequence, and the walk was repeated. After quiescence or 200 mutations, the probe sequence with weakest cross-hybridization was recorded. This process was repeated 500 times. PS1 and PS2 were chosen from the 20 best probe sequences, in order to maximize sequence diversity.

Oligonucleotide mutagenesis

Mutagenesis reactions with PS, PF and X_8 oligonucleotides applied previously described protocols (25,27), using template Stop4 (pM1165a) (8) with analysis by 1% agarose gel electrophoresis. Mutagenesis reactions were conducted in

10× TM buffer (0.5 M Tris, pH 7.5 and 0.1 M $MgCl_2$) diluted to 1× as necessary. Mutagenic oligonucleotides (330 ng, 17.5 pmol) were phosphorylated at 37°C for 1 h in 10 µl reactions containing 1 mM ATP, 5 mM DTT and 5 U T4 polynucleotide kinase. Phosphorylated oligonucleotides (2 µl from the above solution) were annealed to Stop4 (1 µg, 0.49 pmol), for an oligonucleotide:template ratio of 7.16, in 25 µl reactions with an Eppendorf Mastercycler® thermocycler cycled at 85°C for 1 min, 50°C for 3 min and held at 4°C for 5 min (default temperature ramp time). The annealed oligonucleotide reaction was subjected to a complementary strand DNA synthesis in a reaction containing 0.34 mM ATP, 0.85 mM dNTPs, 5.10 mM DTT, 240 U T4 DNA ligase and 3 U T7 DNA polymerase, with overnight incubation at room temperature. Band intensity was quantified using Bio-Rad Quantity One 1-D software (version 4.5.0). Importantly, (*good*:*ugly*) product ratios were only compared among bands run on the same gel. The ratio form cancels many unknown common factors and mitigates the possibility that *good* and *ugly* might intercalate different levels of ethidium bromide. Comparisons of ratios across gels could encounter other experimental non-linearities.

DNA transformants into *E.coli* XL1-Blue were sequenced from individual colonies. PCR for sequencing used the SAV-F1/R2 (mutagenesis target site), 3-fwd/rev (PF3 cross-hybridization site), 7-fwd/rev (PF1 cross-hybridization site) or 8-fwd/rev (PF2 cross-hybridization site) primer pairs. PCR consisted of incubation at 94°C for 3 min; followed with 30 cycles of 30 s at 94°C, 30 s at 55°C and 30 s at 72°C; and finally incubation at 72°C for 7 min. PCR products were subjected to BigDye® terminator sequencing with the M13-F1 or M13-R sequencing primers, following the manufacturer's protocols.

Hairpin analysis and Stop4 mutations

Secondary structure at the PF2 predicted cross-hybridization site was analyzed with the mFold software package (34–36). All options were set to default values, with the exception of folding temperature (50°C) and Na^+ concentration (0.01 M). The PF2 cross-hybridization site of the Stop4 template, located in a non-coding region, was mutated with Stop4Δ-fix and Stop4A-fix to produce Stop4Δ and Stop4A, respectively (25,27). DNA sequencing, as described above, was used to confirm incorporation of the Stop4Δ and Stop4A mutations.

RESULTS

An efficient algorithm predicted cross-hybridization sites and calculated a cross-hybridization score, which correlated well with observed mutagenesis reaction efficiencies. Experimental evidence demonstrated the deleterious effects of cross-hybridization during mutagenesis for both single oligonucleotides and simple mixtures. A fragment of one oligonucleotide was incorporated at a cross-hybridization site. A putative 15 bp hairpin on the template plasmid blocked one predicted cross-hybridization, which was restored by hairpin destabilization.

Prediction and mutagenesis with oligonucleotide mixtures

Kunkel mutagenesis outcomes from 16 oligonucleotide mixtures were classified as Successful (S), Acceptable (A) or Failure (F). Categories were assigned based on the mutagenesis efficiency characterized by gel electrophoresis of the reaction products. Each mixture also received a cross-hybridization score calculated from Equation 1, which assessed cross-hybridization to the template plasmid. High scores were associated with successful mutagenesis reactions (Table 1). On a 2.2 GHz Pentium 4 Linux workstation, the mean time to calculate the cross-hybridization scores for the libraries in Table 1 was 7.2 min and the maximum was 17.2 min.

Table 1. Predicted and experimental outcomes with oligonucleotide mixtures

Oligonucleotide mixture	DNA insert size	Mixture diversity	Experimental outcome	Cross-hybridization score
$X_5CX_9CX_4$	60	1.24×10^{27}	F	0.24
$X_4CX_{10}CX_4$	60	1.24×10^{27}	F	1.21
CX_5CX_2	27	3.44×10^{10}	F	44.86
XCX_5C	24	1.07×10^{9}	F	77.10
$X_7CX_5CX_6$	60	1.24×10^{27}	F	91.90
CX_5C	21	3.36×10^{7}	F	1.10×10^{2}
CX_5CX	24	1.07×10^{9}	F	7.86×10^{2}
$X_6CX_7CX_5$	60	1.24×10^{27}	F	4.10×10^{3}
$X_7CX_4CX_7$	60	1.24×10^{27}	F	1.09×10^{4}
$X_2CX_5CX_2$	33	3.52×10^{13}	S	4.72×10^{11}
X_2CX_5C	27	3.44×10^{10}	A	9.22×10^{11}
X_8	24	1.10×10^{12}	S	2.18×10^{13}
$X_2CX_6CX_2$	36	1.13×10^{15}	S	9.16×10^{13}
$X_2CX_4CX_2$	30	1.10×10^{12}	F	9.39×10^{13}
$X_2CX_7CX_2$	39	3.60×10^{16}	F	2.30×10^{14}
$X_2CX_8CX_2$	42	1.15×10^{18}	A	2.39×10^{14}

Oligonucleotide mixture is a degenerate sequence describing the variable region, where X is a degenerate NNS codon encoding any of the 20 amino acids; X_n is X concatenated *n* times; C is variously TGC or TGT encoding cysteine; N is A, C, G or T; and S is C or G. DNA insert size is the number of nucleotides in the variable region. Mixture diversity is the number of distinct DNA sequences, calculated as 32 raised to the number of NNS codons. The experimental outcome is Successful (S), Acceptable (A) or Failure (F). Successful reactions yielded approximately equal amounts of *good* and *ugly* products. Failures were dominated by *ugly*. Acceptable reactions had intermediate ratios. Cross-hybridization scores were calculated from Equation 1.

Testing cross-hybridization effects on mutagenesis

The association between experimental outcomes and cross-hybridization scores in Table 1 implicated cross-hybridization as a factor in mutagenesis efficiency. To validate this hypothesis experimentally, several oligonucleotides were chosen from the high mutagenesis efficiency X_8 mixture (Table 1). Predicted Successful oligonucleotides, PS1 and PS2, should avoid cross-hybridization. Predicted Failure oligonucleotides, PF1, PF2 and PF3, should exhibit strong cross-hybridization.

Each PS and PF oligonucleotide was used for a separate site-directed Kunkel mutagenesis reaction. The outcome of each reaction was quantified by gel electrophoresis band intensities (Figure 1). DNA sequencing of four PF1 mutants revealed one clone with the 5′ portion of PF1 incorporated precisely at the predicted cross-hybridization site (Figure 2).

Unexpectedly efficient mutagenesis with the PF2 oligonucleotide

The PF2 oligonucleotide, chosen for its strong predicted cross-hybridization, unexpectedly yielded more *good* product during mutagenesis than either PF1 or PF3. Yields of *good* from

```
Stop4                    AGCGTTGGGTCCTGGCCACGGGTGCGCATGATCGTGCTCCTGTCGTTGAGGACCCGGC
                         ----------|||---|||||||||||||||||||||||||||||||||||||||
PF1 mutant               GCTACAAATGCCTATGCACGGGTGCGCATGATCGTGCTCCTGTCGTTGAGGACCCGGC
                         ||||||||||||||||||||||||||||||||||||||||--||-|||||-||||-
Oligonucleotide (PF1)    GCTACAAATGCCTATGCACGGGTGCGCATGATCGTGCTCCTGG-GTGGAGGATCCGGA
```

Figure 2. Sequence analysis of the PF1 cross-hybridization site. The Stop4 template sequence is from the predicted PF1 cross-hybridization site. The PF1 mutant is the cross-hybridization site sequence from one PF1 mutant. Oligonucleotide (PF1) is the PF1 sequence. Boxes or vertical bars indicate identical bases. The hyphen indicates a gap in the local alignment. Boldface indicates the oligonucleotide variable region.

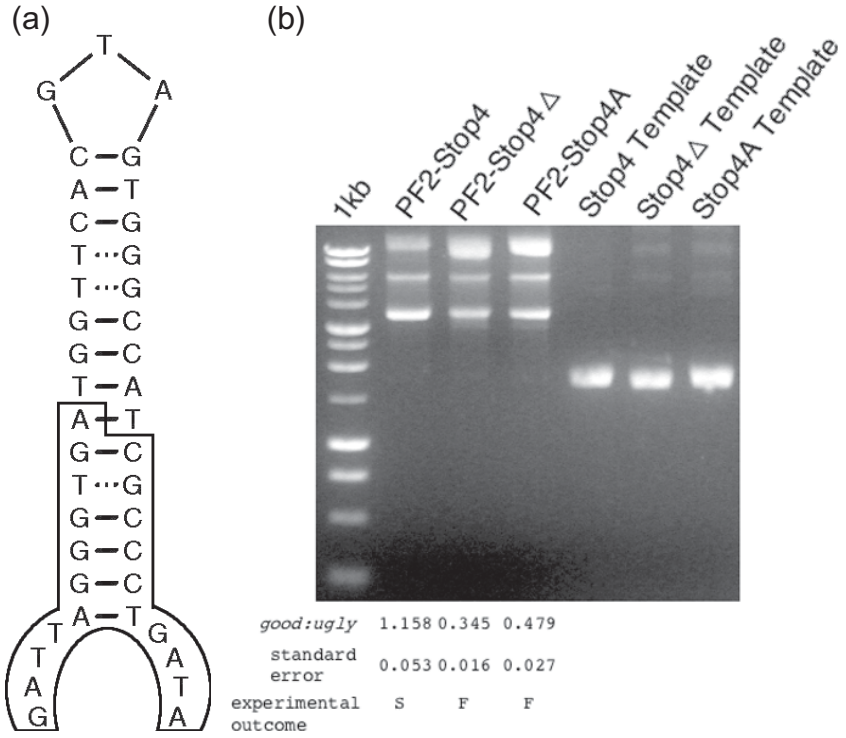

Figure 3. Mutagenesis with modified Stop4 templates. (**A**) A putative hairpin at the predicted PF2 cross-hybridization site. Horizontal lines indicate Watson–Crick base pairs. Dashed lines indicate G-T base pairs. Boxes enclose bases that were mutated from Stop4 by deletion (Stop4Δ) or replacement with adenosine (Stop4A). Unboxed bases are the predicted PF2 cross-hybridization site. (**B**) PF2 mutagenesis reactions on Stop4 templates. PF2-Stop4(Δ, A) shows mutagenesis reaction products. Stop4(Δ, A) template indicates the template starting material. *Good:ugly* ratios are an average of three mutagenesis reactions with the indicated standard error. Experimental outcomes are labeled as in Table 1.

PF2 were comparable to mutagenesis with X_8. Subsequent secondary structure analysis (34–36) revealed a putative hairpin (T_m = 60.2°C) at the PF2 cross-hybridization site (Figure 3A). Since this hairpin could interfere with PF2 cross-hybridization, site-directed mutagenesis of the template was used to reduce hairpin formation. Deletion of template nucleotides (Stop4Δ) or replacement with adenosines (Stop4A) was performed at bases required for hairpin formation, but not for cross-hybridization (Figure 3A). PF2 had lower mutagenesis efficiency with both Stop4Δ and Stop4A, as expected (Figure 3B).

Cross-hybridization in oligonucleotide mixtures

Complex reaction mixtures were modeled by mixing PS1 (correctly hybridizing) with PF1 (cross-hybridizing). With no PF1 added, the reaction was successful. The addition of 10% PF1 degraded the reaction substantially, and levels of PF1 ⩾25% resulted in reaction failure. Thus, even small fractions of cross-hybridizing oligonucleotides dominated the reaction outcome (Figure 4).

DISCUSSION

As demonstrated, mutagenesis failures can arise from cross-hybridizing members of oligonucleotide mixtures. First, a novel computational approach for estimating cross-hybridization correlated well with observed experimental outcomes across a wide range of existing libraries (Table 1). This led us to construct individual oligonucleotides predicted to yield mutagenesis success or failure, which behaved as expected (Figure 1). Several results provided direct experimental evidence that cross-hybridization affects mutagenesis: DNA

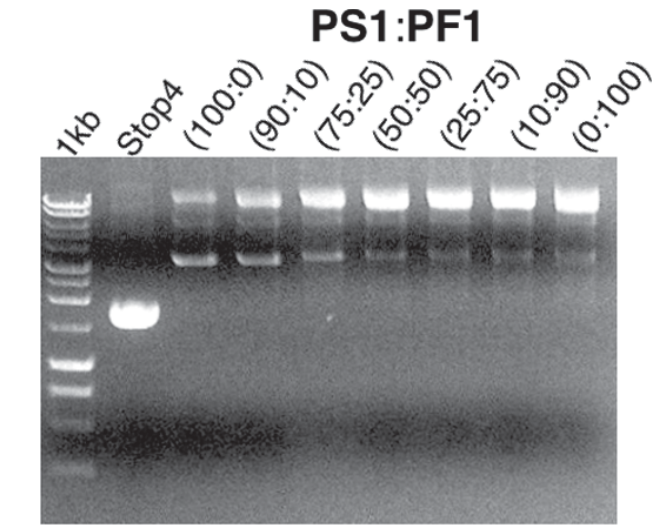

Figure 4. Mutagenesis with mixtures of oligonucleotides. Mutagenesis reactions with mixtures of PS1 (predicted successful) and PF1 (predicted failure) oligonucleotides. Stop4 is the template. Percentages of PS1:PF1 are indicated above each lane. The total quantity of oligonucleotides was 0.2 μg for all reactions. *Good:ugly* ratios are an average of two mutagenesis reactions with the indicated standard error. Experimental outcomes are labeled as in Table 1.

sequencing revealed partial oligonucleotide incorporation at a predicted cross-hybridization site (Figure 2), and putative template secondary structure blocked a cross-hybridization site (Figure 3). Dramatically, mutagenesis failed when cross-hybridizing species were present even in low concentrations relative to an excess of non-cross-hybridizing species (Figure 4).

Experimental and computational results suggest several mechanisms for mutagenesis failures mediated by cross-hybridization. Cross-hybridization could consume template plasmid, block DNA polymerase or incorrectly incorporate the mutagenic oligonucleotide. The resulting *ugly* would reduce the efficiency of competent cell transformation. The net effect of these mechanisms would limit protein library diversity.

Computational analysis

Computational prediction of nucleic acid hybridization appears in many contexts, but to our knowledge, not in degenerate oligonucleotide mixtures for combinatorial mutagenesis. Such mixtures can pose computationally intense problems due to their tremendous theoretical diversities.

Based on the computational analysis, the main contributor to cross-hybridization is the central, variable region of the oligonucleotides. Deleterious cross-hybridization events involve large numbers of oligonucleotides with low ΔG's. The resulting high melting temperatures are often greater than the annealing temperature of 50°C. Thus, cross-hybridization can out-compete annealing to the target site.

Equation 1 is not a partition function, which we presume to be intractable. A full partition function for an oligonucleotide mixture in a mutagenesis reaction would account for many interaction terms omitted here; for example, oligonucleotides from the mixture could bind to the same template plasmid at partially overlapping sites, some mutually exclusive and some not, resulting in different ΔG values for every combination of occupancy states. Instead, the odds ratio form of Equation 1 allows many unknown common factors to cancel from both numerator and denominator, including the partition function and most concentration-dependent effects. The result is an efficient predictive score, without requiring a partition function. The calculated cross-hybridization scores shown in Table 1 vary non-linearly, because the odds ratio is a non-linear transformation with numerator and denominator not linearly related.

This computational approach works particularly well for comparing cross-hybridization scores of degenerate oligonucleotide mixtures. However, cross-hybridization scores are not comparable between degenerate oligonucleotide mixtures and single oligonucleotides for several fundamental physical reasons. These include differences in the levels of oligonucleotide (i) dimers, (ii) hairpins and (iii) incomplete annealing at cross-hybridization sites. Dimeric and hairpin structures are inherent to degenerate mixtures, and are generally not present in designed single oligonucleotides. In degenerate mixtures, whenever one oligonucleotide anneals to a site, there are always a large number of related species that partially anneal to the same site, each with a potentially different binding mode and ΔG. This results in a spread distribution of binding modes and ΔG values at a given site. In contrast, a single oligonucleotide has only one binding mode and a single ΔG at any site, resulting in a point distribution and thus different behavior in Equation 1.

Mutagenesis with computational test oligonucleotides

Oligonucleotide mixtures contain both efficient and inefficient oligonucleotides. The PS and PF oligonucleotides were chosen from X_8 to examine the cross-hybridization scores at the high and low efficiency extremes of the mixture. These oligonucleotides had equal lengths, similar levels of secondary structure and identical annealing sequences. Their behavior in mutagenesis reactions is completely predicted by cross-hybridization effects.

The fragment of PF1 incorporated into the template plasmid of Figure 2 confirmed the predicted cross-hybridization. Incorporation of only a fragment of PF1 is puzzling. Oligonucleotides used in these experiments were PAGE-purified, so the original oligonucleotide was probably the correct length. Thus, this incorporation likely required a non-conventional mechanism, perhaps involving exonuclease activity by T7 DNA polymerase.

Modifying a putative hairpin at the PF2 cross-hybridization site determined the PF2 mutagenesis efficiency. The hairpin melts above the temperature used during PF2 annealing, and could block access to the site. Two approaches to reduce hairpin formation increased the levels of *ugly*. Manipulating access to a cross-hybridization site controlled the success or failure of mutagenesis.

Cross-hybridization poisoning in complex oligonucleotide mixtures was modeled by mixing PS1 and PF1. As expected, in a solution lacking cross-hybridizing members (100% PS1),

the reaction was successful. Some *ugly* was produced, as usual, so minor inefficiency in mutagenesis reactions is perhaps unavoidable with current protocols. Levels of PF1 above 10% resulted in failed mutagenesis reactions. Thus, low levels of cross-hybridizing oligonucleotides can poison mutagenesis reactions.

Outlook

These experiments, both computational and empirical, elucidate cross-hybridization problems in mutagenesis. If site-directed mutagenesis was better understood, one could engineer plasmids or alter mixture compositions to obtain mutagenesis reactions with improved success rates. Though plasmids for phage-displayed peptide libraries were used here, the ability to predict and control mutagenesis outcomes could lead to higher protein library diversities for many protein engineering techniques.

ACKNOWLEDGEMENTS

G.A.W. thanks Sachdev S. Sidhu (Genentech, Inc.) for many helpful discussions, including coining the nomenclature *good*, *bad* and *ugly*. We gratefully acknowledge support from a Young Investigator Award from the Arnold and Mabel Beckman Foundation (to G.A.W). This investigation was supported by National Institutes of Health, National Research Service Award 5 T15 LM00744 from the National Library of Medicine (to C.D.W. and P.Y.T.).

REFERENCES

1. Diaz,J.E., Howard,B.E., Neubauer,M.S., Olszewski,A. and Weiss,G.A. (2003) Exploring biochemistry and cellular biology with protein libraries. *Curr. Issues Mol. Biol.*, **5**, 129–145.
2. Swers,J.S., Kellogg,B.A. and Wittrup,K.D. (2004) Shuffled antibody libraries created by *in vivo* homologous recombination and yeast surface display. *Nucleic Acids Res.*, **32**, e36.
3. Shusta,E.V., Holler,P.D., Kieke,M.C., Kranz,D.M. and Wittrup,K.D. (2000) Directed evolution of a stable scaffold for T-cell receptor engineering. *Nat. Biotechnol.*, **18**, 754–759.
4. Cirino,P.C., Tang,Y., Takahashi,K., Tirrell,D.A. and Arnold,F.H. (2003) Global incorporation of norleucine in place of methionine in cytochrome P450 BM-3 heme domain increases peroxygenase activity. *Biotechnol. Bioeng.*, **83**, 729–734.
5. Peters,M.W., Meinhold,P., Glieder,A. and Arnold,F.H. (2003) Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. *J. Am. Chem. Soc.*, **125**, 13442–13450.
6. Petrounia,I.P. and Arnold,F.H. (2000) Designed evolution of enzymatic properties. *Curr. Opin. Biotechnol.*, **11**, 325–330.
7. Avrantinis,S.K., Stafford,R.L., Tian,X. and Weiss,G.A. (2002) Dissecting the streptavidin–biotin interaction by phage-displayed shotgun scanning. *Chembiochem*, **3**, 1229–1234.
8. Murase,K., Morrison,K.L., Tam,P.Y., Stafford,R.L., Jurnak,F. and Weiss,G.A. (2003) EF-Tu binding peptides identified, dissected, and affinity optimized by phage display. *Chem. Biol.*, **10**, 161–168.
9. Sidhu,S.S., Lowman,H.B., Cunningham,B.C. and Wells,J.A. (2000) Phage display for selection of novel binding peptides. *Meth. Enzymol.*, **328**, 333–363.
10. Weiss,G.A., Watanabe,C.K., Zhong,A., Goddard,A. and Sidhu,S.S. (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl Acad. Sci. USA*, **97**, 8950–8954.
11. Smith,G.P. and Petrenko,V.A. (1997) Phage display. *Chem. Rev.*, **97**, 391–410.
12. Sidhu,S.S. (2001) Engineering M13 for phage display. *Biomol. Eng.*, **18**, 57–63.
13. Boder,E.T. and Wittrup,K.D. (1997) Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.*, **15**, 553–557.
14. Roberts,R.W. and Szostak,J.W. (1997) RNA–peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl Acad. Sci. USA*, **94**, 12297–12302.
15. Mattheakis,L.C., Bhatt,R.R. and Dower,W.J. (1994) An *in vitro* polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl Acad. Sci. USA*, **91**, 9022–9026.
16. Cull,M.G., Miller,J.F. and Schatz,P.J. (1992) Screening for receptor ligands using large libraries of peptides linked to the C terminus of the lac repressor. *Proc. Natl Acad. Sci. USA*, **89**, 1865–1869.
17. van Huizen,R., Miller,K., Chen,D.M., Li,Y., Lai,Z.C., Raab,R.W., Stark,W.S., Shortridge,R.D. and Li,M. (1998) Two distantly positioned PDZ domains mediate multivalent INAD-phospholipase C interactions essential for G protein-coupled signaling. *EMBO J.*, **17**, 2285–2297.
18. Sawano,A. and Miyawaki,A. (2000) Directed evolution of green fluorescent protein by a new versatile PCR strategy for site-directed and semi-random mutagenesis. *Nucleic Acids Res.*, **28**, e78.
19. Fromant,M., Blanquet,S. and Plateau,P. (1995) Direct random mutagenesis of gene-sized DNA fragments using polymerase chain reaction. *Anal. Biochem.*, **224**, 347–353.
20. Henke,E. and Bornscheuer,U.T. (1999) Directed evolution of an esterase from *Pseudomonas fluorescens*. Random mutagenesis by error-prone PCR or a mutator strain and identification of mutants showing enhanced enantioselectivity by a resorufin-based fluorescence assay. *Biol. Chem.*, **380**, 1029–1033.
21. Leung,D.W., Chen,E. and Goeddel,D.V. (1989) A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique*, **1**, 11–15.
22. Voigt,C.A., Martinez,C., Wang,Z.G., Mayo,S.L. and Arnold,F.H. (2002) Protein building blocks preserved by recombination. *Nature Struct. Biol.*, **9**, 553–558.
23. Stemmer,W.P. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, **370**, 389–391.
24. Crameri,A., Raillard,S.A., Bermudez,E. and Stemmer,W.P. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, **391**, 288–291.
25. Kunkel,T.A. (1985) Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc. Natl Acad. Sci. USA*, **82**, 488–492.
26. Kunkel,T.A., Bebenek,K. and McClary,J. (1991) Efficient site-directed mutagenesis using uracil-containing DNA. *Meth. Enzymol.*, **204**, 125–139.
27. Sidhu,S.S. and Weiss,G.A. (2004) In Lowman,H.L. and Clackson,T. (eds), *Phage Display: A Practical Approach*. Oxford University Press, Oxford, UK, pp. 27–41.
28. Penchovsky,R. and Ackermann,J. (2003) DNA library design for molecular computation. *J. Comput. Biol.*, **10**, 215–229.
29. Shih,W.M., Quispe,J.D. and Joyce,G.F. (2004) A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature*, **427**, 618–621.
30. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
31. Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D. and Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
32. Rouillard,J.M., Zuker,M. and Gulari,E. (2003) OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res.*, **31**, 3057–3062.
33. Persengiev,S.P., Zhu,X. and Green,M.R. (2004) Nonspecific, concentration-dependent stimulation and repression of mammalian gene expression by small interfering RNAs (siRNAs). *RNA*, **10**, 12–18.
34. SantaLucia,J.,Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
35. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
36. Peyret,N. (2000) Prediction of nucleic acid hybridization: parameters and algorithms. PhD Dissertation. Wayne State University, Detroit, MI.