J Mol Evol (2011) 72:193–203
DOI 10.1007/s00239-010-9415-2

Exploiting Models of Molecular Evolution to Efficiently Direct Protein Engineering

Megan F. Cole · Eric A. Gaucher

Received: 13 July 2010 / Accepted: 19 November 2010 / Published online: 4 December 2010

Abstract Directed evolution and protein engineering approaches used to generate novel or enhanced biomolecular function often use the evolutionary sequence diversity of protein homologs to rationally guide library design. To fully capture this sequence diversity, however, libraries containing millions of variants are often necessary. Screening libraries of this size is often undesirable due to inaccuracies of high-throughput assays, costs, and time constraints. The ability to effectively cull sequence diversity while still generating the functional diversity within a library thus holds considerable value. This is particularly relevant when high-throughput assays are not amenable to select/screen for certain biomolecular properties. Here, we summarize our recent attempts to develop an evolution-guided approach, Reconstructing Evolutionary Adaptive Paths (REAP), for directed evolution and protein engineering that exploits phylogenetic and sequence analyses to identify amino acid substitutions that are likely to alter or enhance function of a protein. To demonstrate the utility of this technique, we highlight our previous work with DNA polymerases in which a REAP-designed small library was used to identify a DNA polymerase capable of accepting non-standard nucleosides. We anticipate that the REAP approach will be used in the future to facilitate the engineering of biopolymers with expanded functions and will thus have a significant impact on the developing field of 'evolutionary synthetic biology'.

Keywords Directed evolution · Evolutionary models · Functional divergence · Protein engineering

M. F. Cole · E. A. Gaucher
School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA

E. A. Gaucher (✉)
School of Chemistry and Parker H. Petit Institute for Bioengineering and Biosciences, Georgia Institute of Technology, 310 Ferst Dr., Atlanta, GA 30332, USA
e-mail: eric.gaucher@biology.gatech.edu

Introduction

Protein engineering and directed evolution are powerful techniques for improving or modifying the activity, specificity and/or stability of proteins (Arnold and Georgiou 2003; Brakmann 2001; Crameri et al. 1998; Lutz and Patrick 2004; Ness et al. 2002). These approaches have been applied to a wide range of protein families for uses in technology development, therapeutics, agriculture, and chemistry. The general technique consists of fundamental steps that are repeated until a desired property emerges: (1) the introduction of sequence diversity to produce a library of variants from a parent protein; (2) a screen or selection that identifies the variant(s) with the desired phenotype; and, if necessary, (3) recombination between selected variants to produce new sequence combinations.

The success of these experiments depends on both the sequence/functional diversity sampled and the screening/selection assay. Researchers often design large libraries in order to capture as much functional diversity as possible. However, use of such large libraries requires high-throughput assays to select or screen for the desired functional variants. Ideally, these high-throughput assays efficiently capture protein variants with a preferred biomolecular function. In practice, however, high-throughput assays often capture variants whose behavior only serves as a proxy of a desired function (Ness et al. 2005). The importance of the assay's specificity for measuring a desired quality is evident

from the field's axiom of 'you get what you select for' (You and Arnold 1996). Therefore, more accurate low-throughput assays would be greatly preferred if library sizes could be reduced without sacrificing functional diversity.

Due to the necessity of low-throughput screening techniques to capture certain protein functions, many research groups have recently focused their attention on library quality instead of quantity in directed evolution and protein engineering experiments (Lehman and Unrau 2005; Liao et al. 2007; Lutz and Patrick 2004). Smaller pools of variants consisting of fewer and more-focused substitutions are displacing large libraries built from random or shuffled substitutions. The success of these small libraries ultimately depends on their ability to generate a sufficient amount of functional diversity within their reduced sequence space.

Here, we present an approach, termed 'Reconstructing Evolutionary Adaptive Paths' (REAP), that uses the evolutionary history of a protein and the functional diversity of extant homologs to guide the design of small libraries that capture meaningful sequence diversity. Previous work with molecular evolutionary models led us to conclude that understanding the evolutionary history of gene families can offer insight into the particular residues of a gene likely to alter function when working with reduced sequence space for small libraries (Gaucher et al. 2001, 2002a, b, 2003). This strategy can be used to identify residue substitutions that are likely to affect a particular functional property. Thus, a library targeting only these substitutions can be designed to capture functional diversity in a relatively small amount of sequence space.

The REAP approach is distinct from previous methods used to design libraries in that it is more explicit in its use of evolutionary information. Our approach relies on phylogenetic analysis of homologous sequences to detect signatures of functional divergence, and reconstruction of the individual mutations that occurred along these functionally divergent branches of the phylogeny. Here, we present the underlying principles of the REAP approach, an illustration of REAP compared to other common approaches, and demonstrate the power of the REAP approach by presenting a case where it was used to successfully engineer a DNA polymerase capable of utilizing non-standard nucleosides.

Theory

Signatures of Functional Divergence

Signatures of functional divergence and adaptive evolution can be identified using multiple models of molecular sequence evolution. For instance, we and others have developed a methodology that models site-specific rate shifts under a heterotachous framework (where mutation rates for a given residue are not necessarily constant across a phylogeny) such that the homotachous model (where the site-specific mutation rate remains constant across the phylogeny) can be treated as a special case (Gaucher et al. 2002b; Gu 2001; Gu and Vander Velden 2002; Knudsen and Miyamoto 2001; Lopez et al. 2002; Pupko and Galtier 2002; Wang et al. 2007). This provides an opportunity to statistically determine which of the two models better fit the data. Consider a phylogeny with at least two monophyletic clusters, generated by gene duplication or speciation event. It is proposed that a site has two states. In one state (S_0), a site has the same mutation rate in both monophyletic clusters; in the other state (S_1), a site has different rates between the two clusters. The prediction of functional divergence (θ) between two clusters is defined as the probability of a site being S_1, [i.e., $\theta = P(S_1)$], which is called the coefficient of evolutionary functional divergence (Gu 2001). With this approach, the homotachous model is a special case when $\theta = 0$. Conceptually, θ measures the degree of independence (i.e., lack of correlation) between the relative evolutionary rates at the sites in one protein subfamily/lineage versus those in another.

Two types of sequence change are associated with site-specific rate shifts. Both are based on the assumption that residues critical for function tend to be conserved over the course of evolution (Fig. 1). The first type of sequence change, type I functional divergence or heterotachy (also called covarion-like), involves a shift in the relative rate of evolution at a particular site: a site shifts from being relatively strongly conserved (functionally important) to being less conserved (functionally less important) or vice versa. Alternatively, in type II functional divergence, residues involved in the shift in function are highly conserved in both subfamilies but differ in the identity of their amino acids between the monophyletic subfamilies/lineages. Thus, analyses that combine measurements of θ with amino acid identities at sites across a protein can identify those sites associated with type I and type II functional divergence.

Additional models can also be exploited to identify episodes of adaptive evolution across a phylogeny, thereby identifying sites important for a variant library. In particular, the nonsynonymous-to-synonymous ratio is notable for detecting positive selection although the method can also identify sites predicted to be neutrally evolving (Benner and Gaucher 2001; Bielawski and Yang 2004; Gaucher et al. 2003; Wong et al. 2004). For sequences evolving under a neutral model of evolution, a comparison of the sequences will yield a nonsynonymous/synonymous ratio of ~ 1. Sequences under purifying/stabilizing selection will display a ratio less than 1, while sequences under diversifying/positive selection display a ratio greater than 1. Recent statistical advances now allow for the identification of either whole genes or specific sites within genes that have

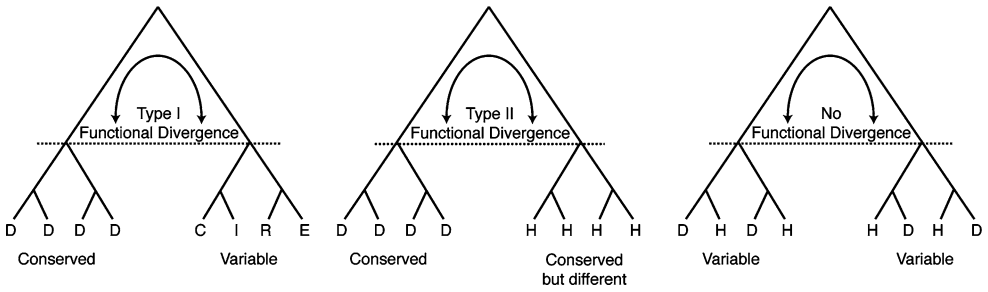

Fig. 1 Signatures of evolutionary functional divergence. Schematics of site-specific types I and II functional divergence between subfamilies of sequences. *Left*, type I functional divergence in which a specific site is occupied by a conserved aspartate residue (D) in one lineage but occupied by many residues at the homologous site in the other lineage. *Middle*, type II functional divergence in which a specific site is again occupied by a conserved aspartate residue (D) in one lineage while the homologous site in the other lineage is also conserved but occupied by a different residue (histidine, H). *Right*, no functional divergence associated with the replacements of aspartates and histidines. Functional inferences associated with these patterns require phylogenetic analysis otherwise they are indistinguishable from historical contingence (i.e., common ancestry)

undergone positive selection during their evolutionary histories (Bielawski and Yang 2004). Evolutionary analysis of nonsynonymous-to-synonymous ratios is therefore another powerful statistical technique to identify functionally important residues.

Once molecular evolutionary models have identified sites implicated in functional divergence, one must next identify the individual mutations that occurred during adaptive episodes at these sites through evolutionary history. This can be accomplished using ancestral sequence reconstruction (Gaucher et al. 2008; Yang et al. 1995). The most common approach for ancestral sequence reconstruction utilizes a model-based likelihood (Thornton 2004). The method follows standard Bayesian statistical theory: given the data at a site, the conditional probabilities of different ancestral states can be compared; and the reconstruction having the highest conditional probability is most often the accepted residue at an ancestral position. Although individual mutations occurring along phylogenetic branches implicated in functional divergence can also be determined by manual inspection, this can be difficult, especially when functional divergence is detected using the nonsynonymous-to-synonymous ratio. As such, the inference of ancestral character states using explicit models of molecular evolution is preferred. The inferred ancestral character states then serve as the sequence information used to design library variants.

Hypothetical Example of REAP and Traditional Library Design Methods

Identification of the various types of signatures of functional divergence left in the sequence record and inference of the amino acid replacements at these sites creates a powerful tool for variant library design. During the evolutionary divergence of a gene family, members of each lineage collect three types of mutations: (1) those that are responsible for the functional divergence of different lineages or homologs, (2) those that are due to neutral evolutionary forces along a branch, and (3) deleterious mutations either weeded out by natural selection or randomly fixed. While most library designs sample all three types of mutations when shuffling homologous sequences, the REAP approach specifically attempts to design variants that sample from only the first type of mutation, offering the tremendous advantage of culling out the much larger number of neutral or random mutations observed throughout evolution of a protein family. To illustrate how this affects the library size, sequence space, and functional space of a variant library, consider the hypothetical library design of a fluorescent protein family using two popular approaches and the REAP approach.

This hypothetical fluorescent protein family contains five homologous subfamilies of individual fluorescent spectra. Each subfamily contains five sequences and all five subfamilies share a common ancestor in the relationship of a polytomy (Fig. 2a). The evolution/engineering of fluorescent proteins with novel properties (i.e., a unique emission spectra) can be attempted using libraries designed by methods such as site-directed/random mutagenesis, DNA shuffling of homologous extant sequences, or the REAP approach.

When site-directed mutagenesis is employed, a parent protein sequence is identified and mutated to generate the variant library. This technique may not rely on evolutionary knowledge and may result in dense sampling in the immediate sequence space of the parent protein even when structural and/or biochemical information is available (Fig. 2a, right). While this approach is straightforward and works well when the desired function can be found in a sequence highly related to the parent protein, this approach generally samples a very limited area of sequence space and can thus entirely miss the functional sequence space of interest. This approach also has several limitations due to

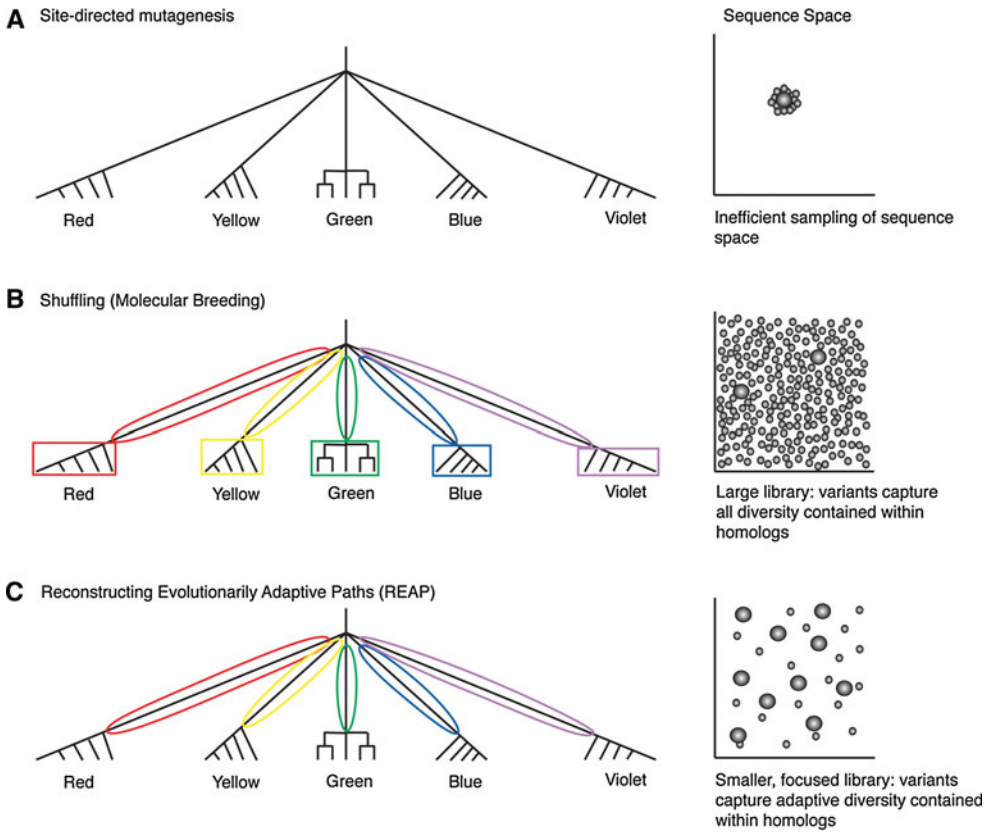

Fig. 2 The use of phylogenetics for directed evolution and protein engineering. This schematic shows how variation contained within homologous sequences is captured by different directed evolution approaches and how this relates to sampling of sequence/function space. Branch lengths are not to scale. **a** Site-directed mutagenesis approach randomly inserts mutations into the parent sequence. **b** Standard DNA shuffling approach builds libraries that incorporate homologous sequence information from all branches of a phylogeny because the approach uses only extant (modern) sequence information. **c** REAP approach builds libraries that incorporate homologous sequence information from only those branches of the phylogeny inferred to have undergone functional adaptation and divergence

the fact that proteins are metastable and the average (random) amino acid replacement is destabilizing rather than stabilizing, resulting in a large proportion of variant proteins being non-functional altogether (Taverna and Goldstein 2002). We expect the REAP approach, and others that consider extant sequence information, to circumvent this problem because they only considers DNA substitutions and/or amino acid replacements that have been accepted by natural selection. This assumes that the sequence background is relatively robust to change and that epistatic effects are minimal (Harms and Thornton 2010). However, as certain phenotypic shifts have been shown to require a particular ancestral background, it may sometimes be necessary to introduce additional residues ancestral to both functional branches in order to create a permissive environment for functional shifts (Bridgham et al. 2009). This is based on the observation that particular combinations of ancestral states can have destabilizing effects through epistatic interactions.

Another approach commonly used is the DNA shuffling technique whereby amino acids present in modern sequences are combinatorially shuffled or recombined to generate a library. As seen in Fig. 2b, patterns of amino acid residues that evolved either within a subfamily (branches bound by boxes) or along the branches that gave rise to the individual subfamilies (circled branches) are integrated during library design. Note that certain amino acid patterns observed in modern proteins arose within the subfamilies (boxed branches) and thus probably have little to offer in terms of generating novel biomolecular properties. These amino acid patterns arose mostly via neutral evolution assuming a lack of selective pressure to diversify within a given subfamily. Meaning, for example, that all proteins within the red family have equivalent emission spectra and the residues that differ between the five red-emitting proteins may not be useful when designing a library. The DNA shuffling approach will sample a large area of sequence space but may result in an intractably large library of variants to screen (Fig. 2b, right).

Unlike the above standard approaches, the REAP method is based on explicit models of molecular evolution that attempt to eliminate amino acid patterns predicted to

have minimal contributions to novel biomolecular functions. This is achieved by incorporating only the amino acid patterns that arose during the adaptive evolution of unique properties compared to the last common ancestor of the fluorescent proteins (Fig. 2c, circled branches), and neglecting the amino acid patterns that arose within a family. In doing so, this increases or at least maintains the unique behaviors captured using the standard DNA shuffling approach, while limiting the number of mutations by culling those that are inferred to have minimal impact on functional diversification.

The REAP Method

General Methodology

A general flowchart for the REAP approach that can be used to guide variant library design is presented in Fig. 3. The first step is to collect homologous sequences of a parent protein from databases such as NCBI or PFAM. A multiple sequence alignment is then created using software such as ClustalW (Larkin et al. 2007) or T-Coffee and manually inspected and refined as needed to obtain a trustworthy alignment. This alignment is used as input for a phylogenetic analysis to determine the relationships and evolutionary distances between the parent protein and its homologs. Software such as MrBayes (Huelsenbeck et al. 2001) can be used to construct a phylogenetic tree, which can be checked against existing knowledge of evolutionary relationships between the included species and adjusted if necessary.

The phylogenetic tree and multiple sequence alignment are then used as input into software programs such as DIVERGE (Gu and Vander Velden 2002) and Rate Shift Analysis Server (Knudsen and Miyamoto 2001) that use evolutionary models to describe the replacements of amino acids, rate heterogeneity among sites, etc. When these models detect functional divergence along branches of the phylogeny, ancestral sequence reconstruction, using programs such as PAML (Yang 2007), can be used to identify the specific residues that are changing along these branches. This list of residues can be further culled, as other directed evolution approaches often do, using protein

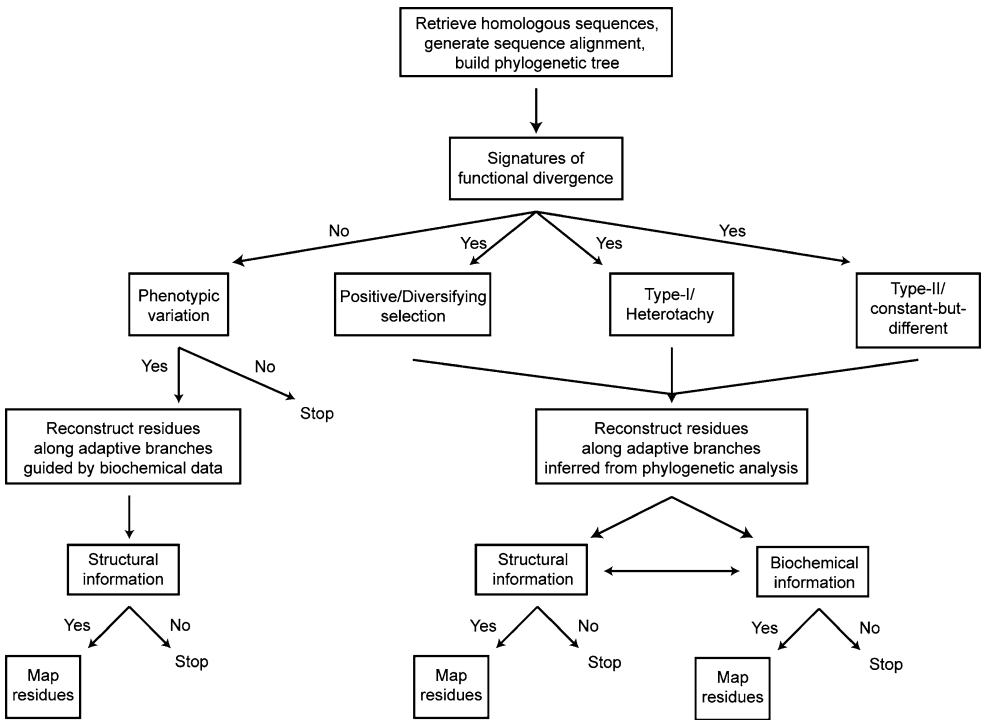

Fig. 3 Flowchart for the REAP approach. The implementation of the REAP approach begins by collecting and aligning sequences from a protein family. A phylogenetic tree must then be constructed to capture the evolutionary relationship and distance between homologs. From this information, molecular models are used to detect functional divergence along branches of the phylogeny. The computational reconstruction of ancestral states of the protein along these branches is then used to identify residues and amino acids associated with the functional divergence. This will result in a list of candidate residues/mutations that may affect the function of the protein. This list of candidates can be further reduced if needed by incorporating known structural or biochemical information about the protein. For example, residues may be selected based on their proximity to the protein's active site. The final candidate residue list is then used to design the variant library to screen for the desired function. In this manner, the REAP approach results in a small number of residues/mutations to vary in the sequence library

structural models or biochemical information to direct mutations to particular sites or domains of a protein (Perez-Jimenez et al. 2009; Yuen and Liu 2007). When signatures of functional divergence are not detected, but there is known phenotypic variation among homologs, ancestral sequence reconstruction can still be used to provide a list of substitutions observed in the evolutionary history of the protein but these substitutions cannot be culled according to their connection to functional divergence.

REAP Analysis of Family A DNA Polymerases

DNA polymerases are routinely used for basic research and biotechnology applications. Although it is common practice to use standard nucleoside triphosphates as substrates for these polymerases, there is a growing interest in identifying/evolving/engineering DNA polymerases capable of incorporating non-standard nucleosides (Henry and Romesberg 2005; Patel et al. 2001; Sismour et al. 2004). For instance, incorporating non-standard nucleosides would allow researchers to explore alternative sugar-ring structures and modified backbone linkages in DNA, expand the information capacity of DNA beyond the four standard nucleosides, and develop novel sequencing technology. One of the earliest and most influential developments along the latter line included the engineering of *Thermus aquaticus* DNA polymerase (Taq) to accept dideoxynucleoside triphosphates. These modified nucleosides serve as *non-reversible* terminators for Sanger-based cycle sequencing and revolutionized DNA sequencing. Next-generation sequencing technology, such as sequencing-by-synthesis (SbS), will require polymerases that incorporate *reversible* terminators. Nucleosides whose 3′ OH group, for instance, is replaced with an ONH_2 group will block extension until the O–N bond in the ONH_2 group is cleaved to restore the 3′ OH and allow extension to proceed. The REAP approach was applied to Taq polymerase and its homologous family members (Family A DNA polymerases) to identify sites involved in expanded substrate recognition in order to engineer polymerases capable of incorporating dNTP-ONH_2 reversible terminators.

A phylogenetic tree and multiple sequence alignment of Family A DNA polymerases from eukarya, archaea, bacteria, and viruses, were composed of 719 family A polymerase sequences available in the PFAM database at the time (PF00476) (Fig. 4) (Bateman et al. 2004). Type I and type II functional divergence was detected by feeding the alignment (with some taxa removed to reduce computational complexity) into DIVERGE and Rate Shift Analysis Server (http://www.daimi.au.dk/~compbio/rateshift/protein.html). The computational phylogenetic analysis of sequence information confirmed what was already known from the

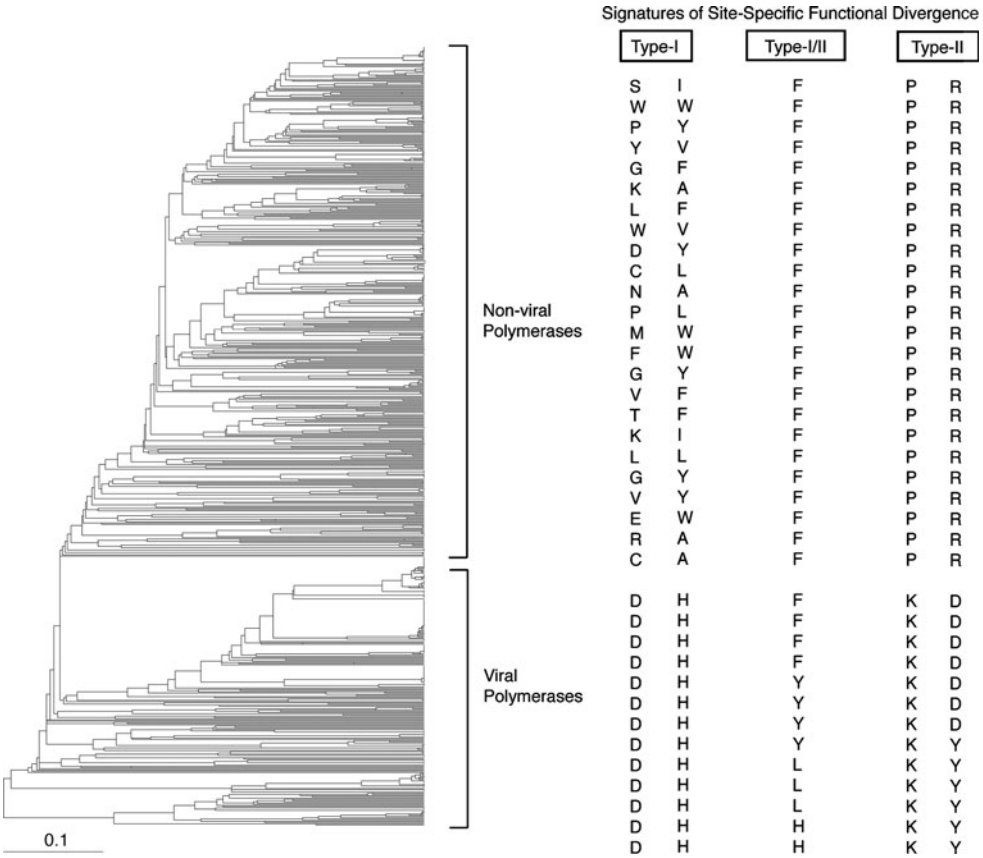

Fig. 4 Phylogeny of Family A DNA polymerases. The viral and non-viral clades used for REAP analysis are *highlighted*. *Scale bar* represents amino acid replacements/site/unit evolutionary time. Examples of patterns of types I and II functional divergence are also shown

literature: functional divergence of polymerase behaviors has occurred along branches of the phylogeny separating viral and non-viral polymerases (Horlacher et al. 1995; Leal et al. 2006; Sismour et al. 2004; Tabor and Richardson 1995). Based on the observation that viral polymerases are better able to accept modified nucleosides than non-viral polymerases, we reasoned that extracting specific sequence information (sites responsible for functional divergence) from viral polymerases and placing them within the genetic background of the non-viral Taq polymerase would generate evolution-guided engineered polymerases with modified substrate specificities.

Amino acid residues replaced along the branches separating viral and non-viral polymerases were inferred using PAML (version 4.1) by incorporating the WAG matrix with rate variation following a gamma distribution. These analyses identified numerous sites as potentially involved in functional divergence between viral and non-viral polymerases both inside and outside the active-site cleft of the polymerase structure. We elected to focus our analysis on sites within the active-site only since these are known to alter substrate specificity (Henry and Romesberg 2005). A total of 57 amino acid replacements distributed across 35 sites were predicted to have the potential to expand the substrate recognition of Taq polymerase (Fig. 5) and are listed in Table 1.

Identification of Polymerase Variants from the REAP-Designed Library

From the REAP analysis a library of 93 Taq polymerase variants was designed where each variant had 3–4 amino acid replacements and each replacement was present in six of the variants (Chen et al. 2010). Each variant was cloned, expressed and then assayed for the ability to incorporate the reversible terminator dNTP-ONH_2. Thirty variants (32%) were able to incorporate the reversible terminator to some degree and of these, eight of them were able to incorporate the reversible terminator with a threshold n + 1 extension efficiency of at least 50% in 2 min. Two of the eight variants were exceptional in their ability to incorporate the modified nucleoside and were thus used in further assays to demonstrate their utility for sequencing reactions using reversible terminators (full details provided in Chen et al. 2010).

Our previous study of DNA polymerases clearly demonstrates the power of the REAP method. Using the REAP approach, a small number of amino acid replacements were identified as potentially useful to expand protein function, in this case the ability of polymerase to accept a non-standard nucleoside substrate. A small library of variants was then generated that could be assayed in a low-throughput manner to accurately screen for the desired function. A high percentage of the engineered proteins had a detectable increase in the desired function and two variants had a level of function high enough for use in the desired application without further modifications. Thus, the REAP approach exploited evolutionary information and models of molecular evolution to efficiently design and identify a protein with new functionality.

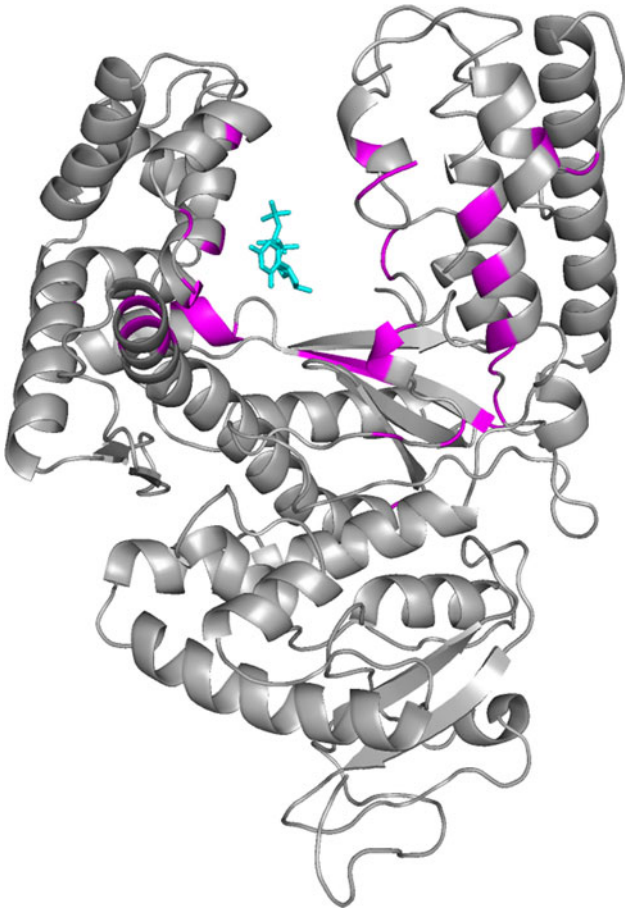

Fig. 5 Distribution of functional divergence sites mapped onto the structure of Taq polymerase. Locations of the 35 sites in Taq polymerase identified by the REAP analysis and mapped onto the polymerase structure and colored in magenta (PDB accession 5KTQ). The incoming nucleoside triphosphate substrate is shown in cyan

Discussion

Evolution is defined as change (mutation) in the heritable information (usually DNA) of biological systems from generation to generation. Mutations can be the result of either natural selection (i.e., adaptive) or genetic drift (neutral, or slightly deleterious), and occur randomly (Kimura 1991). In its purest form, experimental evolution of a protein in the laboratory would follow an analogous path; random changes followed by selection of variants for a particular property. The expanse of mutation space and limitations in selection schemes, however, render this

Table 1 Amino acid replacements identified by a REAP analysis of branches separating viral and non-viral lineages of Family A DNA polymerases

Site	Wild-type Taq	Engineered Taq variant
483	Asn	Arg
489	Gln	His
513	Ser	Ile
514	Thr	Val
520	Glu	Ile, Gly
536	Arg	Ile
540	Lys	Ile
544	Thr	Ala
545	Tyr	Glu
576	Ser	Glu, His
578	Asp	Phe, Thr
582	Gln	Ala
583	Asn	Gln, Ser
586	Val	Lys
587	Arg	Val
597	Ala	Thr, Cys
598	Phe	Val, Trp
600	Ala	Ser
604	Trp	Gly
608	Ala	Gly, Lys, Glu
609	Leu	Cys, Pro, Ser
610	Asp	Trp
614	Ile	Glu, Gly, Gln
615	Glu	Ile
616	Leu	Ala, Ile, Asp
625	Asp	Ser, Leu, Ala
660	Arg	Asp
667	Phe	Tyr, His, Leu
671	Tyr	Phe
673	Met	Gly, Ala
742	Glu	Pro, Arg
743	Ala	Ser, Arg
745	Glu	His, Val
746	Arg	Ala
777	Ala	His

Note: Sites are numbered according to wild-type Taq polymerse

approach impractical. To overcome these problems, researchers often focus their attention to particular positions and a subset of mutations to *direct* the evolution of a parent protein (Tobin et al. 2000; Van Regenmortel 2000).

The input for such direction varies considerably from in silico thermodynamic and steric structural considerations to in vitro mutagenesis experiments, and even activity profiles from initial rounds of directed evolution experiments (Fox et al. 2003, 2007; Korkegian et al. 2005; Saraf et al. 2004; Voigt et al. 2001). One of the most widely used approaches academically and commercially though involves shuffling of genetic differences between a parent protein and homologous members of the parent protein's evolutionary family (Crameri et al. 1996, 1998). The success of 'DNA shuffling' (Molecular Breeding) is threefold: (1) sampling of sequence space is restricted, (2) library variants can contain sufficient diversity to generate novel biomolecular functions, and (3) mutations contained within the set of homologs have already been subjected to evolutionary forces and are thus either adaptive, neutral or only slightly deleterious. This last point is noteworthy since it implies that none of the individual mutations extracted from the set of homologous sequences are deleterious enough to inactivate the protein. This assumption is, of course, incorrect if some of the homologs underwent pseudogenization or if particular mutations are context-dependent, beneficial or neutral in some sequence contexts but deleterious in others (Wang and Pollock 2005).

Although the shuffling approach greatly reduces the sequence space that is explored compared to randomly mutating a sequence, the approach scales exponentially with the number of homologs and the amount of sequence diversity contained within them. This restricts the amount of homologous sequence information that can be exploited to direct the evolution of a protein and thus decreases the chances of incorporating mutations that generate diversifying biomolecular functions. In addition, genome sequencing projects are rapidly identifying homologs for all gene families. This additional sequence information can be a burden for standard shuffling approaches, but it is favorable for approaches that exploit evolutionary information in designing libraries.

The ability to place large numbers of homologous sequences within an evolutionary framework provides an opportunity to determine whether conservation and variation are the result of functional divergence/constraint or common ancestry (Govindarajan et al. 2003; Lichtarge et al. 1996). Conserved sites are more likely to be associated with functional constraints if the evolutionary distance separating the sequences is long rather than short. Conversely, variable sites are more likely to be associated with functional divergence when they occur along short evolutionary paths (i.e., branches).

Several methods have been developed to identify functionally important sites within proteins based on evolutionary analysis of homologous sequences. For instance, ConSurf (Landau et al. 2005), uses multiple sequence alignments to score each residue for overall conservation amongst homologs. While this is a powerful way to identify sites implicated in the basic function of a protein, it is not able to identify sites associated with functional divergence in some protein subfamilies. One approach that can identify sites important for specific sub-family

functionalization is Evolutionary Trace (Lichtarge et al. 1996), which scores residues for conservation across all species and across consecutively smaller sub-classes. However, this method treats all mutations equally and does not use explicit evolutionary models to identify functionally relevant sites, thus missing certain types of sequence signatures of functional divergence. Two notable methods that attempt to harness the power of ancestral sequence reconstruction are Substitution Mapping (Skovgaard et al. 2006) and Ancestral Mutation (Yamashiro et al. 2010). Substitution mapping identifies sites of interest within a homolog that exhibit a desired functional quality by sequence comparison and then inserts these substitutions into a parent sequence. This approach relies on a priori knowledge of protein function and is thus not applicable to studies where the functions of various homologs have not been determined. The Ancestral Mutation method, on the other hand, does not rely on functional knowledge but rather on the fact that ancestral proteins often have increased thermostability. Thus, this method has been used to improve thermostability of enzymes by introducing ancestral residues into extant sequences.

REAP is unique from these other methods in that it uses explicit models of molecular evolution to identify sequence signatures of functional divergence within protein subfamilies. This evolution-guided strategy incorporates only those mutations inferred to be associated with new biomolecular properties during the evolution of a protein family, and to exclude the much larger set of mutations that do not lead to new functions (neutral or slightly deleterious mutations). For instance, assuming a background mutation rate of ca. 1×10^{-10} to 5×10^{-9} per base-pair per generation in *E. coli* (Lenski et al. 2003), neutral mutations will outnumber adaptive mutations 70–2000:1 per genome per generation (Perfeito et al. 2007). Eliminating the majority of these neutral mutations allows one to create libraries whose variants display a high level of functional diversity while restricting the overall sequence diversity, yielding a small library with a high proportion of active members which, in turn, permits the use of reliable low-throughput assays guaranteed to screen/select for variants exhibiting specific functions.

While the REAP methodology can be a powerful approach for protein engineering or directed evolution experiments it is not predicted to be ideal for all library designs. The approach requires numerous homologous sequences to generate an articulated phylogeny. Further, the phylogeny needs to represent a family of sequences with diverse behaviors guided by functional divergence, otherwise the extracted amino acid patterns may not generate novel function.

However, when there is sufficient homologous sequences and functional divergence, REAP may have substantial advantages over traditional library designs. The approach is intended to incorporate information from diverse family members to create a highly active and functional library. There is also no requirement for information regarding protein structure or the mutability of sites (mutagenesis experiments) to guide the library design. Equally important is that REAP libraries contain an order-of-magnitude fewer variants than most other types of libraries, allowing researchers to save time, money and to use low-throughput assays.

The general utility of the REAP approach will ultimately be determined by its ability to generate diverse biomolecules having novel functions. The approach has already been proven effective for the design of DNA polymerases capable of accepting non-standard nucleosides but further validation of the technique is needed for a wide array of protein designs. For now, we anticipate that the REAP approach will make substantial contributions to protein engineering and synthetic biology. For example, REAP could be used to generate protein variants capable of supporting unnatural amino acid incorporation during protein synthesis. The resulting biopolymers will then serve as the information (novel coding systems) and catalytic (novel side-chain chemistry) components of an expanded biology that we have termed "evolutionary synthetic biology" (Gaucher 2007).

Acknowledgments This work was supported by a National Institutes of Health grant to EAG. MFC was supported by an NIH NRSA award and in part by the Emory University Fellowship in Research and Science Teaching (FIRST) program's NIH/NIGMS IRACDA grant number K12 GM000680-11. This work was also supported by the National Aeronautics and Space Administration's Exobiology and Astrobiology Programs.

References

Arnold FH, Georgiou G (2003) Directed enzyme evolution: screening and selection methods. Humana Press, Totowa, New Jersey

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Res 32:D138–D141

Benner SA, Gaucher EA (2001) Evolution, language and analogy in functional genomics. Trends Genet 17:414–418

Bielawski JP, Yang Z (2004) A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J Mol Evol 59:121–132

Brakmann S (2001) Discovery of superior enzymes by directed molecular evolution. Chembiochem 2:865–871

Bridgham JT, Ortlund EA, Thornton JW (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature 461:515–519

Chen F, Gaucher EA, Leal NA, Hutter D, Havemann SA, Govindarajan S, Ortlund EA, Benner SA (2010) Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. Proc Natl Acad Sci USA 107:1948–1953

Crameri A, Whitehorn EA, Tate E, Stemmer WP (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. Nat Biotechnol 14:315–319

Crameri A, Raillard SA, Bermudez E, Stemmer WP (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 391:288–291

Fox R, Roy A, Govindarajan S, Minshull J, Gustafsson C, Jones JT, Emig R (2003) Optimizing the search algorithm for protein engineering by directed evolution. Protein Eng 16:589–597

Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, Chung LM, Ching C, Tam S, Muley S, Grate J, Gruber J, Whitman JC, Sheldon RA, Huisman GW (2007) Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol 25:338–344

Gaucher EA (2007) Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: realizing and extending the vision of Zukerkandl and Pauling. Oxford University Press, Oxford, pp 20–33

Gaucher EA, Miyamoto MM, Benner SA (2001) Function–structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc Natl Acad Sci USA 98:548–552

Gaucher EA, Das UK, Miyamoto MM, Benner SA (2002a) The crystal structure of eEF1A refines the functional predictions of an evolutionary analysis of rate changes among elongation factors. Mol Biol Evol 19:569–573

Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002b) Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27:315–321

Gaucher EA, Miyamoto MM, Benner SA (2003) Evolutionary, structural and biochemical evidence for a new interaction site of the leptin obesity protein. Genetics 163:1549–1553

Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451:704–707

Govindarajan S, Ness JE, Kim S, Mundorff EC, Minshull J, Gustafsson C (2003) Systematic variation of amino acid substitutions for stringent assessment of pairwise covariation. J Mol Biol 328:1061–1069

Gu X (2001) Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453–464

Gu X, Vander Velden K (2002) DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. Bioinformatics 18:500–501

Harms MJ, Thornton JW (2010) Analyzing protein structure and function using ancestral gene reconstruction. Curr Opin Struct Biol 20:360–366

Henry AA, Romesberg FE (2005) The evolution of DNA polymerases with novel activities. Curr Opin Biotechnol 16:370–377

Horlacher J, Hottiger M, Podust VN, Hubscher U, Benner SA (1995) Recognition by viral and cellular DNA polymerases of nucleosides bearing bases with nonstandard hydrogen bonding patterns. Proc Natl Acad Sci USA 92:6329–6333

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Evolution—Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314

Kimura M (1991) The neutral theory of molecular evolution: a review of recent evidence. Jpn J Genet 66:367–386

Knudsen B, Miyamoto MM (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci USA 98:14512–14517

Korkegian A, Black ME, Baker D, Stoddard BL (2005) Computational thermostabilization of an enzyme. Science 308:857–860

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33:W299–W302

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and clustal X version 2.0. Bioinformatics 23:2947–2948

Leal NA, Sukeda M, Benner SA (2006) Dynamic assembly of primers on nucleic acid templates. Nucleic Acids Res 34:4702–4710

Lehman N, Unrau PJ (2005) Recombination during in vitro evolution. J Mol Evol 61:245–252

Lenski RE, Winkworth CL, Riley MA (2003) Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. J Mol Evol 56:498–508

Liao J, Warmuth MK, Govindarajan S, Ness JE, Wang RP, Gustafsson C, Minshull J (2007) Engineering proteinase K using machine learning and synthetic genes. BMC Biotechnol 7:16

Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342–358

Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1–7

Lutz S, Patrick WM (2004) Novel methods for directed evolution of enzymes: quality, not quantity. Curr Opin Biotechnol 15:291–297

Ness JE, Kim S, Gottman A, Pak R, Krebber A, Borchert TV, Govindarajan S, Mundorff EC, Minshull J (2002) Synthetic shuffling expands functional protein diversity by allowing amino acids to recombine independently. Nat Biotechnol 20:1251–1255

Ness JE, Cox AJ, Govindarajan S, Gustafsson C, Gross RA, Minshull J (2005) Empirical biocatalyst engineering: escaping the tyranny of high throughput screening. American Chemical Society, Washington, DC

Patel PH, Kawate H, Adman E, Ashbach M, Loeb LA (2001) A single highly mutable catalytic site amino acid is critical for DNA polymerase fidelity. J Biol Chem 276:5044–5051

Perez-Jimenez R, Li JY, Kosuri P, Sanchez-Romero I, Wiita AP, Rodriguez-Larrea D, Chueca A, Holmgren A, Miranda-Vizuete A, Becker K, Cho SH, Beckwith J, Gelhaye E, Jacquot JP, Gaucher EA, Sanchez-Ruiz JM, Berne BJ, Fernandez JM (2009) Diversity of chemical mechanisms in thioredoxin catalysis revealed by single-molecule force spectroscopy (vol 16, pg 890, 2009). Nat Struct Mol Biol 16:1331

Perfeito L, Fernandes L, Mota C, Gordo I (2007) Adaptive mutations in bacteria: high rate and small effects. Science 317:813–815

Pupko T, Galtier N (2002) A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc Biol Sci 269:1313–1316

Saraf MC, Horswill AR, Benkovic SJ, Maranas CD (2004) FamClash: a method for ranking the activity of engineered enzymes. Proc Natl Acad Sci USA 101:4142–4147

Sismour AM, Lutz S, Park JH, Lutz MJ, Boyer PL, Hughes SH, Benner SA (2004) PCR amplification of DNA containing non-standard base pairs by variants of reverse transcriptase from Human Immunodeficiency Virus-1. Nucleic Acids Res 32:728–735

Skovgaard M, Kodra JT, Gram DX, Knudsen SM, Madsen D, Liberles DA (2006) Using evolutionary information and ancestral sequences to understand the sequence–function relationship in GLP-1 agonists. J Mol Biol 363:977–988

Tabor S, Richardson CC (1995) A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. Proc Natl Acad Sci USA 92:6339–6343

Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? Proteins 46:105–109

Thornton JW (2004) Resurrecting ancient genes: experimental analysis of extinct molecules. Nat Rev Genet 5:366–375

Tobin MB, Gustafsson C, Huisman GW (2000) Directed evolution: the 'rational' basis for 'irrational' design. Curr Opin Struct Biol 10:421–427

Van Regenmortel MH (2000) Are there two distinct research strategies for developing biologically active molecules: rational design and empirical selection? J Mol Recognit 13:1–4

Voigt CA, Mayo SL, Arnold FH, Wang ZG (2001) Computational method to reduce the search space for directed protein evolution. Proc Natl Acad Sci USA 98:3778–3783

Wang ZO, Pollock DD (2005) Context dependence and coevolution among amino acid residues in proteins. Methods Enzymol 395:779–790

Wang HC, Spencer M, Susko E, Roger AJ (2007) Testing for covarion-like evolution in protein sequences. Mol Biol Evol 24:294–305

Wong WS, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041–1051

Yamashiro K, Yokobori S, Koikeda S, Yamagishi A (2010) Improvement of *Bacillus circulans* beta-amylase activity attained using the ancestral mutation method. Protein Eng Des Sel 23:519–528

Yang ZH (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang ZH, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino-acid-sequences. Genetics 141:1641–1650

You L, Arnold FH (1996) Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide. Protein Eng 9:77–83

Yuen CM, Liu DR (2007) Dissecting protein structure and function using directed evolution. Nat Methods 4:995–997