Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Chemical Biology

Utilizing natural diversity to evolve protein function: applications towards thermostability

Megan F Cole[1] and Eric A Gaucher[1,2]

Protein evolution relies on designing a library of sequences that capture meaningful functional diversity in a limited number of protein variants. Several approaches take advantage of the sequence space already explored through natural selection by incorporating sequence diversity available from modern genomes (and their ancestors) when designing these libraries. The success of these approaches is, partly, owing to the fact that modern sequence diversity has already been subjected to evolutionary selective forces and thus the diversity has already been deemed 'fit to survive'. Five of these approaches will be discussed in this review to highlight how protein engineers can use evolutionary sequence history/diversity of homologous proteins in unique ways to design protein libraries.

Addresses
[1] School of Biology, Georgia Institute of Technology, Department of Biology, Atlanta, GA, 30332, United States
[2] School of Chemistry and Parker H. Petit Institute for Bioengineering and Biosciences, Georgia Institute of Technology, Atlanta, GA, 30332, United States

Corresponding author: Gaucher, Eric A (eric.gaucher@biology.gatech.edu)

Current Opinion in Chemical Biology 2011, **15**:399–406

This review comes from a themed issue on
Molecular Diversity
Edited by T. Ashton Cropp and Dewey McCafferty

Available online 4 April 2011

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2011.03.005

Introduction

The engineering of proteins to improve or modify enzyme properties is often necessary or desirable for many industrial and biomedical applications. The most common approach used by researchers to obtain desired enzymatic functions involves generating a library of sequence variants, screening or selecting for the best variants, and repeating this process as many times as necessary [1]. The success of this approach, termed protein engineering/evolution, relies on the size and quality of the variant library used. Complete coverage of sequence space is often infeasible for even relatively short protein sequences and large library sizes can restrict the screening/selection assays amenable to a study. Library design is thus a crucial component affecting the success of protein evolution studies.

One approach to library design is to exploit the diversity that millions or billions of years of natural selection provides us. Nature samples a large amount of sequence space and natural selection typically selects against unfit variants during the course of a protein's evolution across time and phyla. This selection can generate modified behaviors or properties *within* a homologous gene family either through gene duplication and diversification or through adaptation of orthologous sequence among species. Scientists can exploit this by designing variants that sample across modern species' sequence diversity and backwards in time through the evolutionary history of a gene family to capture functional diversity while excluding the many deleterious mutations already weeded out by natural selection.

Design strategies guided by evolutionary sequence information range from recombining modern sequences to entirely resurrecting ancestral sequences. The composition of a variant library depends on where along the continuum of sequence history and diversity an approach lies. One metric for assessing libraries is 'functional information content' [2], which measures the number of sequences that pass some threshold for functionality in relation to the total relevant sequence space. Therefore, libraries that contain a high number of variants exhibiting a desired property, in a small overall library size, have high functional information content. This review will discuss a number of successful approaches that exploit sequence history and diversity in unique ways to design variant libraries, will highlight their strengths for certain applications such as thermostability, and discuss their relative functional information contents (Table 1).

DNA shuffling

The field of protein evolution was revolutionized with the invention of DNA shuffling (Figure 1) [3]. In this approach, a pool of homologous sequences are randomly fragmented and recombined via PCR. Generally a pool of modern protein sequences are used to create a variant library composed of novel combinations of these extant sequences. Improved approaches, such as synthetic shuffling, have since been introduced to eliminate recombination biases and allow for independent reshuffling of amino acids [4].

DNA shuffling has been used with great success in many applications [5,6]. One study that highlights the power of the unbiased nature of this approach used DNA shuffling on the industrially relevant enzyme Chitinase and found that the amino acid replacements that generated improved

Table 1

Summary of approaches

Approach	Methodology	Use of sequence information	Functional information content	References
DNA Shuffling	Variants created by recombination among extant sequences	Relies only on modern sequence diversity	Low	[3–8,39]
Consensus Sequence	Most common amino acid at sites is used across whole protein or a select number of residues are incorporated into extant sequence	Relies only on modern sequence diversity	Moderate	[9–12]
Ancestral Sequence Resurrection (ASR)	Computationally infer and experimentally resurrect ancestral sequences	Uses wide range of both sequence history and diversity	Moderate	[13–20]
Ancestral Mutation Method (AMM)	Replace ancestral residues into an extant protein sequence	Uses a single modern sequence and a subset of its ancestral residues	High	[27,28,29,30••]
Reconstructing Evolutionary Adaptive Paths (REAP)	Replace ancestral residues associated with functional divergence into a modern protein sequence	Uses a single modern sequence and a subset of its ancestral residues	High	[31,33,35]

activity lay outside the substrate binding pocket and the catalytic region [7]. These residues probably would not have been predicted *de novo* to be important for activity. DNA shuffling was also notably used to engineer plants resistant to the pesticide glyphosate by evolving a glyphosate acetyltransferase [8]. Through eleven rounds of shuffling, enzyme activity was increased by four orders of magnitude, enough to confer strong resistance.

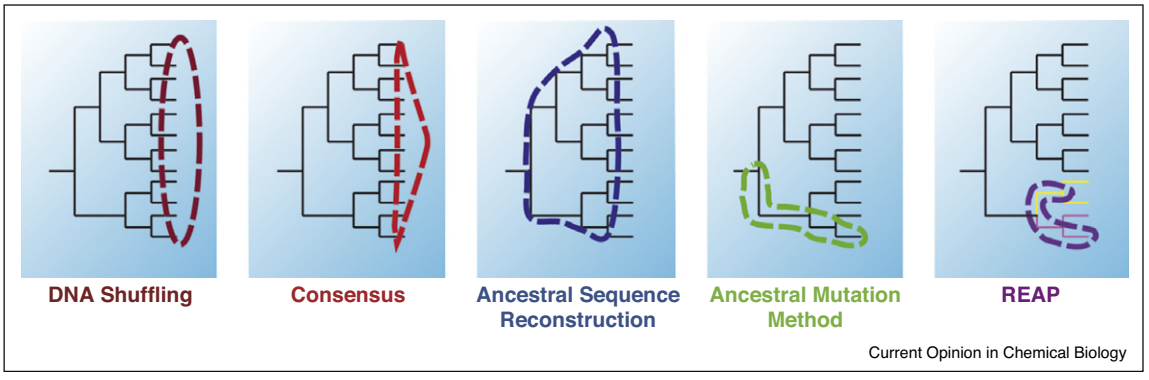
DNA shuffling is a powerful approach that fully exploits the known sequence and functional diversity of a protein family. A major drawback, however, is that it generates very large libraries and thus may not be amenable to all studies depending on resources and assays. It is thus important to be able to cull the sequence diversity contained within extant sequences. Four such approaches are presented below.

Consensus sequence

One way to cull extant sequence information is to generate a consensus sequence (Figure 1). This approach first determines the most commonly used amino acid found in extant organisms at each position in a sequence alignment and then incorporates these residues across the length of the sequence when engineering a consensus protein. Consensus amino acids can also be used to replace individual residues in an extant protein sequence to create a library of consensus variants.

This approach has proven useful for engineering various proteins with increased stabilities [9,10•]. For instance,

Figure 1



Approaches and their use of evolutionary sequence space.Theoretical phylogenetic trees are depicted for each of the five approaches with circled regions representing the sequence space sampled by each when designing variants. DNA shuffling uses recombination between extant sequences. Consensus sequence uses extant sequences to design a single consensus variant. ASR resurrects ancestral sequences at internal nodes of the phylogeny. AMM replaces some residues in an extant sequence with ancestral amino acids. REAP replaces residues in an extant sequence with ancestral amino acids associated with functional divergence between two subfamilies in a phylogeny (depicted by yellow and purple branches).

Lehmann *et al.* used the consensus approach to improve the thermostability of fungal phytase since no naturally occurring phytases had been identified with the thermostability required for use in animal foodstocks [11]. Thirteen fungal sequences were used to design a consensus protein that had both normal catalytic capabilities and a 15–22 °C increase in thermostability. Another study used a library of consensus variants to generate variants of the β-lactamase enzyme having improved thermal and proteolytic stabilities [12].

The strength of the consensus approach relies both on its utility to improve thermostability (a common target of protein engineering studies) as well as its straightforward implementation. Libraries as small as a single variant can result in successful identification of improved proteins. Another benefit to this approach is that it can engineer thermophilic enzymes derived solely from mesophilic sequences. The drawback to the approach, however, is that a consensus sequence is a heuristic of a true ancestral sequence thus it is neither able to account for the context-dependent nature of amino acids in proteins (epistasis) nor has a consensus sequence been subjected to natural selection at any point in evolutionary history.

Ancestral sequence reconstruction

Reconstructing ancestral sequences has the potential to overcome some of the limitations associated with a consensus sequence (Figure 1). Ancestral Sequence Reconstruction (ASR), first suggested by Pauling and Zuckerkandl in 1963 [13], consists of aligning extant protein sequences, building a phylogenetic tree to determine evolutionary relationships and then inferring ancestral amino acid states at nodes in the phylogeny using Parsimony or Bayesian probability [14]. The inferred sequences of any of the ancestral proteins at some or all nodes in the phylogeny can then be synthesized in the laboratory to form a library of variants.

The application of ASR for protein engineering studies is new compared to DNA shuffling and consensus, but several groups have already used ancestral sequence reconstruction to explore the evolution and diversification of protein properties during the course of evolutionary history [15•,16]. One such study resurrected variants of the Elongation Factor Tu (EF-Tu) protein family and found a striking trend between thermostability and geologic time [17•]. The most ancient ancestral proteins were the most thermostable. Other studies using ASR include the resurrection of the ancestral glucocorticoid and mineralcorticoid receptors to gain insights into the nature of protein evolution and diversification of ligand binding [18], the resurrection of ancient fluorescent proteins to understand phenotypic diversification and functional divergence [19], and a study of the functional evolution of a protein involved in type-2 diabetes, glucagon-like peptide-1 [20].

Ancestral sequence reconstruction offers the benefit of small library sizes but is limited by the properties of ancestral proteins. This approach may be most useful for designing more thermostable and/or acid-tolerant proteins since it is hypothesized that ancient life lived in hotter and more acidic environments [21–23]. Additionally, ancestral sequences may prove to be a better 'parent' sequence for use in protein evolution libraries as they often have increased stability and more biochemical promiscuity [24••,25]. However, ancestral sequence reconstruction also has two significant limitations. First, computational inferences are limited by the available extant sequences and their correct alignments. Thus, an ancestor dating back to the last common ancestor of a domain of life can only be inferred when homologous sequences exist in a sufficient number of extant organisms from *multiple* domains of life. Another limitation is the difficulty in handling insertions, deletions, or sequence regions that cannot be correctly aligned. However, both of these issues will become less of a concern as more protein sequences continue to be added to databases and as our understanding of the evolution of insertion and deletion events increases [26].

Ancestral mutation method

An approach related to ancestral sequence reconstruction is the ancestral mutation method (AMM) (Figure 1), which introduces ancestral residues into a modern protein sequence. This approach thus places ancestral residues within the framework of a protein in which all other residues are provided by a modern sequence. Potential benefits of this approach are that it can combine useful properties of modern and ancestral sequences and that it allows the researcher control over the library size. However, the crux of ancestral mutation method is in selecting which ancestral residues to introduce into a library.

This approach has been applied to a handful of proteins to improve thermostability [27]. In one application, ancestral residues were placed into a modern 3-isopropylmalate dehydrogenase and of the variants tested, five had increased thermostability over the wild-type protein [28]. The method was also used on Glycyl-tRNA synthetase from *Thermus thermophilus* and the majority of the variants assayed displayed higher thermostability and even higher activity than the wild-type protein [29]. In another study, this approach was applied to a mesophilic enzyme, β-amylase, to identify a variant with higher thermostability and activity [30••].

This approach will probably prove useful where characteristics of both modern and ancestral proteins are desired. For example, if a particular extant enzyme has a unique substrate specificity but is not particularly thermostable, AMM would be a useful approach to introduce higher thermostability into this protein while maintaining its enzymatic characteristics. However, for this approach

to be broadly used it will be necessary to define good strategies for identifying which specific residues should be selected for replacement.

Reconstructing Evolutionary Adaptive Paths

One approach for selecting a subset of modern residues to replace with ancestral residues in an extant sequence is Reconstructing Evolutionary Adaptive Paths (REAP) (Figure 1) [31]. This approach uses signatures of functional divergence found within the phylogenetic sequence history of a protein family to identify a specific set of sites that are likely to alter a protein's function [32–34]. An example of such a sequence signature is where an amino acid residue differs either in its identity or in its conservation between two subfamilies of a phylogeny. The underlying assumption of the REAP approach is that when such a difference is present between two subfamilies in a tree, that the branch connecting the two subfamilies has experienced functional divergence and thus the residues changing along the branch may be responsible for functional divergence of the homologous proteins between the two subfamilies.

REAP has already demonstrated utility in a study of family A DNA polymerases [35••]. This work exploited the observation that viral polymerases have a higher substrate promiscuity than non-viral polymerases to engineer a *Taq* polymerase variant capable of accepting a non-standard nucleoside. A library containing only 93 variants based on the evolutionary REAP analysis was used to identify a polymerase capable of incorporating a useful non-standard nucleoside with high fidelity.

The strength of the REAP approach lies in its ability to cull the large number of potential ancestral residues to a subset that are most probably responsible for functional divergence between homologous proteins. Such a quality makes the REAP approach valuable because properties from two functionally diverse subfamilies in a phylogeny can be combined into a single protein. REAP is likely to also prove valuable for gene duplication events where researchers want to identify the residues responsible for functional innovation after the divergence of two paralogous protein families resulting from either subfunctionalization or neofunctionalization.

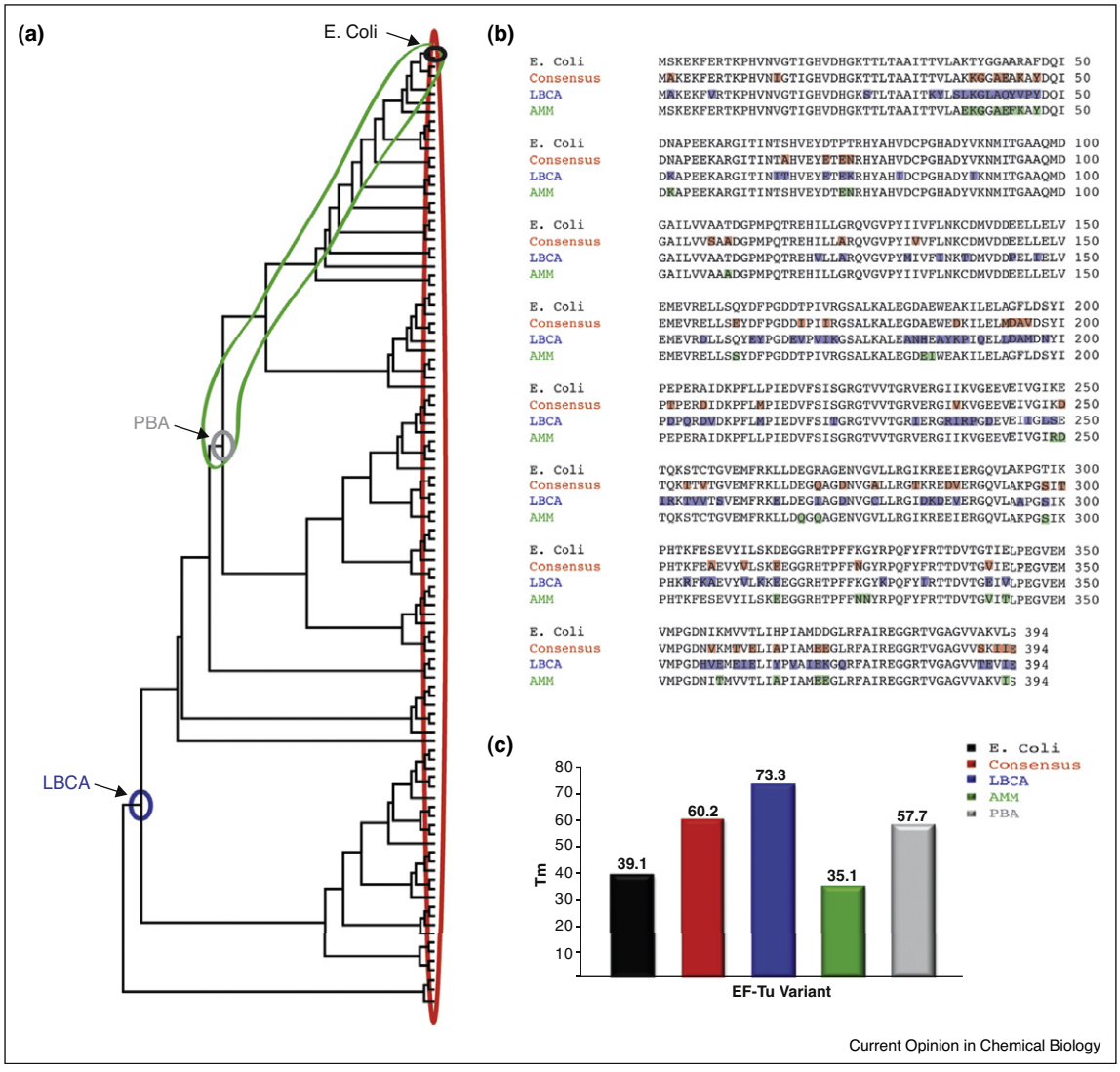
Applications towards thermostability

Enhanced thermostability is one of the most common properties desired as output from a protein engineering study. Many industrial and biomedical applications require enzymes with activity at high temperatures for greater chemical control or overall stability for long-term usage or storage. As such, there has been great interest in isolating proteins from thermophilic or hyperthermophilic organisms over the years, and many such enzymes are either currently used for these applications or served as the parent protein for a protein engineering/evolution study [36]. However, there remains a need for effective strategies to design or evolve increased thermostability since not all protein families are present in thermophiles and since thermophilic enzymes often do not have optimal characteristics such as a particular substrate specificity or activity required for industry or biomedicine.

Enzyme stability is a determinant of protein fitness. For many globular proteins from microscopic organisms, thermostability is dictated by environmental temperature [37]. Thus, an organism living in a hot environment contains proteins that are thermostable. Conversely, an organism living in a cooler environment contains proteins that are unstable at high temperatures (mesostable). Extant sequence diversity and evolutionary history of a protein can thus reveal valuable information about amino acid patterns contributing to a protein's stability. The approaches described in this review each exploit the sequence diversity provided by natural selection to design variant libraries for protein engineering. Here we can examine experimental applications for some of the approaches discussed above towards engineering thermostability in Elongation Factor Tu (EF-Tu) proteins.

EF-Tu is an essential protein necessary for translation and whose thermostability is linearly correlated to the optimal growth temperature of its host organism [17•,37]. Bacterial EF-Tu sequences were previously collected and used to make a phylogenetic tree and subsequently reconstruct ancestral sequences (Figure 2a). We computationally generated and experimentally synthesized four EF-Tu variants; the extant *E. coli* sequence, the bacterial consensus EF-Tu, an ASR of the Last Bacterial Common Ancestor (LBCA), and an *E. coli* sequence with some ancestral mutations, AMM (Figure 2b). The AMM variant was designed using residues from the inferred proteobacteria ancestor (PBA) EF-Tu sequence, which has a melting temperature of 57 °C. The goal then is to convert a 39 °C protein (*E. coli*) into a 57 °C protein. Of the 69 total residues different between the PBA and *E. coli* EF-Tu sequences, we picked the top 30 predicted to enhance thermostability from physicochemical properties and evolutionary parameters we and others have found to be correlated among thermostable proteins (solvent accessibility, interaction energies, hydrophobicity and site-specific evolutionary conservation) [38]. The DNA shuffling technique was not explored owing to the large number of variants required for this approach, but DNA shuffling has been successfully used previously to improve protein thermostability [39]. REAP was also not explored since this approach requires a functional shift between at least two subfamilies in a phylogeny (e.g. EF-Tu sequences with high melting temperatures in one subfamily of the tree and low melting temperatures in another subfamily) and there is currently not enough diversity of sequence data among thermophilic branches

Figure 2



Thermostability of EF-Tu variants.
(a) A phylogenetic tree of the EF-Tu family of proteins. Black circle denotes the *E. coli* node, red indicates nodes used to construct the consensus sequence, green indicates the ancestral nodes leading back to a proteobacteria ancestor (PBA) circled in gray, and blue indicates the node for the last bacterial common ancestor (LBCA). **(b)** A multiple sequence alignment of EF-Tu variant sequences. Colored boxes indicate residues differing between the variant and extant *E. coli* sequences. **(c)** The protein variants were cloned, purified and their melting temperatures (in degrees Celcius) were measured by circular dichroism.

to support this approach. However, the REAP approach was able to improve thermostability of a DNA Polymerase (data not shown [35••]).

The four EF-Tu variants were purified and assayed for their melting temperatures using circular dichroism (Figure 2c) as previously described [17•]. The LBCA variant had the highest melting temperature, which is more than 30 degrees higher than the extant *E. coli* protein (73 °C vs. 39 °C). The consensus EF-Tu variant also shows enhanced thermostability (60 °C) but not as enhanced as the LBCA. The sequence identities between *E. coli* versus the consensus and LBCA variants are 86% and 72% respectively, and between the consensus and LBCA is 76%. The observation that the AMM variant did not display increased thermostability (35 °C) demonstrates the difficulty in identifying sites to replace using the AMM approach despite the fact that we incorporated 30 of the 69 amino acid differences between the *E. coli* and ancestral proteins.

In addition to the various melting temperature profiles, it is equally important to note that each approach yielded a folded protein. This observation alone demonstrates the

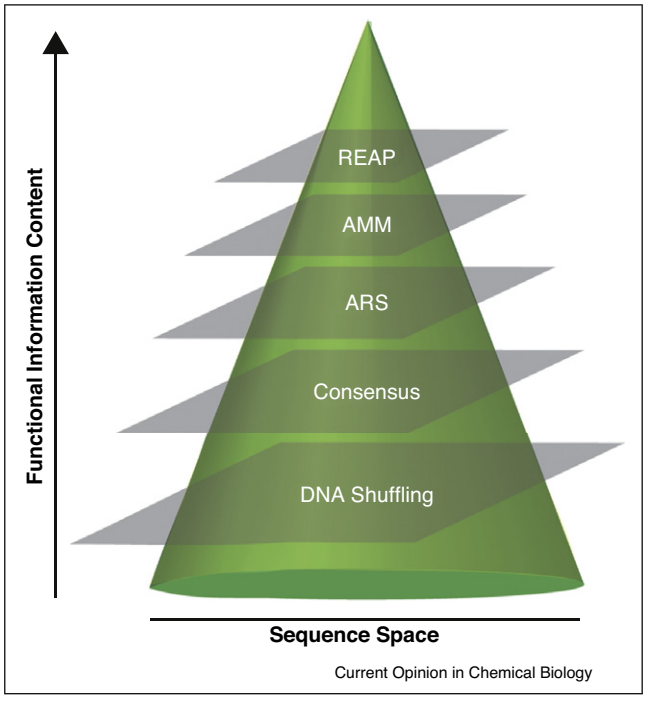
utility of using evolutionary sequence history and diversity to guide protein design. The ability to constrain the design of variants to properly folded protein space is a benefit that eliminates much of the need to assay large numbers of variants in protein engineering studies.

Conclusions

Each of the approaches described in this review takes advantage of some of the sequence space already explored over evolutionary time and screened via natural selection. In so doing, these approaches increase the functional information content of their libraries by having a higher ratio of functional sequences to library size as compared to random mutagenesis. The approaches do, however, display a continuum of functional information content (Figure 3).

DNA shuffling produces the largest libraries in terms of functional sequence but also the largest in terms of sequence space. This requires researchers to sample a large number of variants to identify desired protein properties. As such, DNA shuffling has low functional information content. In comparison, a library of consensus variants contains far fewer sequences than a shuffled library yet also contains substantial functional diversity and thus may contain improved functional information content than a shuffled library. ASR further reduces both sequence space and non-functional space while still providing functional diversity since these sequences presumably have already experienced some form of natural selection in their evolutionary history. Functional information content can be still further enhanced by AMM and REAP especially when an ancestral protein had properties different than its descendent proteins, allowing AMM and REAP variants to contain beneficial properties from both ancestral and modern proteins.

Figure 3



Exploiting evolutionary sequence information can increase functional information content of variant libraries.
Diagram depicting how each of the approaches may relate to the functional information content of their variant library designs. Approaches such as REAP and AMM that can target a small set of mutations to design proteins with desired properties have the potential to have very high functional information content while approaches such as DNA shuffling that require large library sizes generally have lower functional information content.

In determining which approach will yield the maximum functional information content for a *specific* protein engineering study, however, one must consider many factors such as sequence availability, the number of variants desired in a library, the type of selection/screen to be employed, and the selective forces that governed the evolution of a protein family, among other things. It is important to note that an exhaustive DNA shuffling library encompasses the libraries of each of the other four approaches. Thus, with unlimited resources and time, DNA shuffling would always be the preferred technique when exploiting modern sequence information for protein engineering studies. However, given that most studies are limited in how many variants can be created and screened by the allotted timeframe and budget, it is often desirable to use one of the approaches with smaller library size if possible.

It is also important to note that there are other successful approaches to protein evolution that leverage additional information beyond the sequence record, such as structural knowledge, to improve the functional information content of libraries [40]. It is also necessary in some cases to sample residues not found in nature in order to identify a desired property. Nonetheless, we anticipate that approaches that exploit sequence information would still prove useful by serving as desirable parent (or seed) sequences that could be randomly or selectively mutated to screen residues not found in the sequence record in attempt to generate a variant with the desired property.

It is clear that there will continue to be a need for protein engineering, especially as fields such as synthetic biology continue to develop. The approaches for variant design described here demonstrate the great power in exploiting what natural selection provides us in the sequence record. While these approaches each utilize this information in different ways, there is still a lot of room to develop additional methods that cull the entirety of sequence space to create libraries having optimal functional information content. As more sequence data become available each year the power of these techniques will be realized as the need for additional ways to cull ever-expanding

sequence databases becomes imperative. We anticipate that these and other approaches that make use of this sequence information will continue to aid the specifically engineering of new proteins and the emerging field of 'evolutionary synthetic biology' in general [33].

Acknowledgements

We thank Dr. Michael Gromiha, James Kratzer, Dr. Eric Ortlund, Dr. Manish Chandra Pathak, and Ryan Randall for assistance with our computational and experimental analyses. This work was funded partly by NASA Astrobiology at Georgia Tech, NASA Exobiology grant NNX08AO12G to E.A.G., NIH NRSA grant F32GM095182-01 to M.F.C. and the Emory University Fellowship in Research and Science Teaching (FIRST) program's NIH/NIGMS/IRACDA grant GM000680-11.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Lutz S, Bornscheuer U: *The Protein Engineering Handbook*. Weinheim: Wiley-VCH; 2009.

2. Szostak JW: **Functional information: molecular messages**. *Nature* 2003, **423**:689.

3. Stemmer WP: **Rapid evolution of a protein in vitro by DNA shuffling**. *Nature* 1994, **370**:389-391.

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

5. Sclimenti CR, Thyagarajan B, Calos MP: **Directed evolution of a recombinase for improved genomic integration at a native human sequence**. *Nucleic Acids Res* 2001, **29**:5044-5051.

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

7. Fan Y, Fang W, Xiao Y, Yang X, Zhang Y, Bidochka MJ, Pei Y: **Directed evolution for increased chitinase activity**. *Appl Microbiol Biotechnol* 2007, **76**:135-139.

8. Castle LA, Siehl DL, Gorton R, Patten PA, Chen YH, Bertain S, Cho HJ, Duck N, Wong J, Liu D *et al.*: **Discovery and directed evolution of a glyphosate tolerance gene**. *Science* 2004, **304**:1151-1154.

9. Steipe B, Schiller B, Pluckthun A, Steinbacher S: **Sequence statistics reliably predict stabilizing mutations in a protein domain**. *J Mol Biol* 1994, **240**:188-192.

10. • Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon AP, Wyss M: **The consensus concept for thermostability engineering of proteins: further proof of concept**. *Protein Eng* 2002, **15**:403-411.
This article introduces the consensus approach and applies the approach to increase the thermostability of Phytase proteins to nearly 40 °C greater than any parent protein.

11. Lehmann M, Kostrewa D, Wyss M, Brugger R, D'Arcy A, Pasamontes L, van Loon AP: **From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase**. *Protein Eng* 2000, **13**:49-57.

12. Amin N, Liu AD, Ramer S, Aehle W, Meijer D, Metin M, Wong S, Gualfetti P, Schellenberger V: **Construction of stabilized proteins by combinatorial consensus mutagenesis**. *Protein Eng Des Sel* 2004, **17**:787-793.

13. Pauling L, Zuckerkandl E: **Chemical paleogenetics, molecular restoration studies of extinct forms of life**. *Acta Chemica Scandinavica* 1963, **17**:9-16.

14. Liberles D: *Ancestral Sequence Reconstruction*. USA: Oxford University Press; 2007.

15. • Bridgham JT, Ortlund EA, Thornton JW: **An epistatic ratchet constrains the direction of glucocorticoid receptor evolution**. *Nature* 2009, **461**:515-519.
This article shows how ASR can be used to dissect the evolutionary paths responsible for biochemical and functional adaptation of receptors towards new substrates.

16. Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, Aris JP, Benner SA: **Resurrecting ancestral alcohol dehydrogenases from yeast**. *Nat Genet* 2005, **37**:630-635.

17. • Gaucher EA, Govindarajan S, Ganesh OK: **Palaeotemperature trend for Precambrian life inferred from resurrected proteins**. *Nature* 2008, **451**:704-707.
This article applies ASR to ancient proteins inferred to be on the order of hundreds of millions to billions of years old. These ancient proteins display a near linear increase in thermostability the older a protein is in geologic time.

18. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW: **Crystal structure of an ancient protein: evolution by conformational epistasis**. *Science* 2007, **317**:1544-1548.

19. Ugalde JA, Chang BS, Matz MV: **Evolution of coral pigments recreated**. *Science* 2004, **305**:1433.

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

21. Di Giulio M: **The universal ancestor was a thermophile or a hyperthermophile: tests and further evidence**. *J Theor Biol* 2003, **221**:425-436.

22. Woese CR: **Bacterial evolution**. *Microbiol Rev* 1987, **51**:221-271.

23. Perez-Jimenez R, Ingles-Prieto A, Zhao Z, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Holmgren A, Sanchez-Ruiz JM, Gaucher EA, et al.: **Single-molecule paleoenzymology probes the chemistry of resurrected enzymes**. *Nat Struct Mol Biol*, in press, 2011.

24. •• Bershtein S, Goldin K, Tawfik DS: **Intense neutral drifts yield robust and evolvable consensus proteins**. *J Mol Biol* 2008, **379**:1029-1044.
This article demonstrates that ASR and Consensus generate variants that are more tolerant of deleterious mutations owing to increased stability of ancient and consensus residues. The authors show that these variants are more likely to acquire new functions and thus promote such variants as 'parent' sequences in protein engineering studies.

25. Tokuriki N, Tawfik DS: **Stability effects of mutations and protein evolvability**. *Curr Opin Struct Biol* 2009, **19**:596-604.

26. Blanchette M, Diallo AB, Green ED, Miller W, Haussler D: **Computational reconstruction of ancestral DNA sequences**. *Methods Mol Biol* 2008, **422**:171-184.

27. Watanabe K, Ohkuri T, Yokobori S, Yamagishi A: **Designing thermostable proteins: ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree**. *J Mol Biol* 2006, **355**:664-674.

28. Miyazaki J, Nakaya S, Suzuki T, Tamakoshi M, Oshima T, Yamagishi A: **Ancestral residues stabilizing 3-isopropylmalate dehydrogenase of an extreme thermophile: experimental evidence supporting the thermophilic common ancestor hypothesis**. *J Biochem* 2001, **129**:777-782.

29. Shimizu H, Yokobori S, Ohkuri T, Yokogawa T, Nishikawa K, Yamagishi A: **Extremely thermophilic translation system in the common ancestor commonote: ancestral mutants of Glycyl-tRNA synthetase from the extreme thermophile *Thermus thermophilus***. *J Mol Biol* 2007, **369**:1060-1069.

30. •• Yamashiro K, Yokobori S, Koikeda S, Yamagishi A: **Improvement of Bacillus circulans beta-amylase activity attained using the ancestral mutation method**. *Protein Eng Des Sel* 2010, **23**: 519-528.
This article presents the AMM approach as a combination of ASR and Consensus. The authors use AMM to improve the thermostability of β-amylase.

31. Gaucher EA: **Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: realizing (and extending) the vision of Zukerkandl and Pauling**. In *Ancestral Sequence Reconstruction.* Edited by Liberles DA. Oxford University Press; 2007:20-33.

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

33. Cole MF, Gaucher EA: **Exploiting models of molecular evolution to efficiently direct protein engineering**. *J Mol Evol* 2010.

34. Benner SA, Gaucher EA: **Evolution, language and analogy in functional genomics**. *Trends Genet* 2001, **17**:414-418.

35. •• Chen F, Gaucher EA, Leal NA, Hutter D, Havemann SA, Govindarajan S, Ortlund EA, Benner SA: **Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection**. *Proc Natl Acad Sci USA* 2010, **107**:1948-1953.
This article presents the REAP approach as a novel way to exploit natural sequence diversity using an evolutionary phylogenetic context. REAP was applied to DNA polymerases to generate variants having novel substrate binding properties.

36. Arnold FH, Georgiou G: *Directed Enzyme Evolution: Screening and Selection Methods*. Totowa, NJ: Humana Press; 2003.

37. Gromiha MM, Oobatake M, Sarai A: **Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins**. *Biophys Chem* 1999, **82**:51-67.

38. Gromiha MM, Sarai A: **Thermodynamic database for proteins: features and applications**. *Methods Mol Biol* 2010, **609**: 97-112.

39. Park YM, Phi QT, Song BH, Ghim SY: **Thermostability of chimeric cytidine deaminase variants produced by DNA shuffling**. *J Microbiol Biotechnol* 2009, **19**:1536-1541.

40. Heinzelman P, Snow CD, Smith MA, Yu X, Kannan A, Boulware K, Villalobos A, Govindarajan S, Minshull J, Arnold FH: **SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability**. *J Biol Chem* 2009, **284**:26229-26233.