

# Exploring protein fitness landscapes by directed evolution

Philip A. Romero and Frances H. Arnold



**Darwin200**

Abstract | Directed evolution circumvents our profound ignorance of how a protein's sequence encodes its function by using iterative rounds of random mutation and artificial selection to discover new and useful proteins. Proteins can be tuned to adapt to new functions or environments by simple adaptive walks involving small numbers of mutations. Directed evolution studies have shown how rapidly some proteins can evolve under strong selection pressures and, because the entire 'fossil record' of evolutionary intermediates is available for detailed study, they have provided new insight into the relationship between sequence and function. Directed evolution has also shown how mutations that are functionally neutral can set the stage for further adaptation.

## Evolvability

A measure of the ability of a protein to adapt in response to mutation and selective pressure; for example, the frequency of beneficial mutations.

## Directed evolution

The application of iterative rounds of mutation and artificial selection or screening to alter the properties of biological molecules and systems

Millions of years of life's struggle for survival in different environments have resulted in proteins providing diverse, creative and efficient solutions to a wide range of problems, from extracting energy from the environment to repairing and replicating their own code. Good solutions to biological problems can also be good solutions to human problems — proteins are widely used in the food, chemicals, consumer products and medical fields. Not content with nature's protein repertoire, however, protein engineers are working to extend known protein function to new environments or tasks<sup>1–4</sup> and to create new functions altogether<sup>5–7</sup>.

Despite major advances, a molecular-level understanding of why one protein performs a certain task better than another remains elusive. This is perhaps not surprising when we remember that a protein often undergoes conformational changes during function and exists as a dynamic ensemble of conformers that are only slightly more stable than their unfolded and non-functional states and that might themselves be functionally diverse<sup>8</sup>. Mutations far away from active sites can influence protein function<sup>9,10</sup>. Engineering enzymatic activity is particularly difficult because very small changes in structure or chemical properties can have big effects on catalysis. Thus, predicting the amino acid sequence, or changes to an amino acid sequence, that would generate a specific behaviour remains a challenge, particularly for applications requiring high performance (such as an industrial enzyme or a therapeutic protein). Unfortunately, where function is concerned, details matter, and we just don't understand the details.

Evolution, however, had no difficulty generating these impressive molecules. Despite their complexity and finely tuned nature, proteins are remarkably evolvable: they can adapt under the pressure of selection by changing their behaviour, function and even fold. Protein engineers have learned to exploit this evolvability using directed evolution — the application of iterative rounds of mutation and artificial selection or screening — to generate new proteins. Hundreds of directed evolution experiments have revealed the ease with which proteins adapt to new challenges<sup>11</sup>. Notable recent examples include a recombinase evolved to remove proviral HIV from the host genome (providing a new strategy for treating retroviral infections)<sup>12</sup>, a cytochrome P450 fatty acid hydroxylase that was converted into a highly efficient propane hydroxylase (thereby proving that a cytochrome P450 is fully capable of hydroxylating small alkanes, even though most propane-using organisms use structurally and mechanistically unrelated enzymes)<sup>13</sup>, a more than 40 °C increase in the thermostability of lipase A (extending its application in biocatalysis to a whole new set of environments)<sup>14</sup> and a variant of green fluorescent protein that tolerates having all its leucine residues replaced with a non-natural amino acid, trifluoroleucine<sup>15</sup>. Roger Tsien won the Nobel Prize last year for his work on the fluorescent proteins that have transformed biological imaging<sup>16</sup>. Directed evolution had a key role by improving many features of fluorescent proteins, including emission and excitation properties, quantum yield, multimerization state and maturation rate<sup>4,17</sup>.

Directed evolution has become a common laboratory tool for altering and optimizing protein function (as well as the function of other biological molecules and

Division of Chemistry and Chemical Engineering, 210-41, California Institute of Technology, Pasadena, California, 91125, USA.  
e-mails: frances@cheme.caltech.edu; promero@caltech.edu  
doi:10.1038/nrm2805

**Box 1 | Directed evolution of other biological components and systems**

Evolution is unique because it works at all scales, from molecules to ecosystems — no other engineering design algorithm can make that claim. A simple algorithm of mutation and artificial selection has proved effective for everything from the selective breeding of plants and animals to discovering self-replicating nucleic acid sequences. Biological components and systems have shown a remarkable ability to adapt under the pressure of artificial selection with an evolvability that probably reflects their own history of natural selection<sup>100</sup>.

Functional nucleic acids have been evolved in the laboratory to achieve new and improved properties<sup>19</sup>. Because the phenotype and genotype are encoded in the same molecules these experiments involve *in vitro* selections, whereby pools of up to 10<sup>15</sup> sequences can be synthesized and evaluated outside of cells<sup>101</sup>. Hydrolysis of nucleic acid phosphodiester bonds and the binding of specified ligands are among the functions that have been discovered in this way<sup>102,103</sup>. Recently, a set of self-replicating RNA enzymes that catalyse their own synthesis in a self-sufficient manner was created<sup>104</sup>.

Directed evolution can also be applied to enzyme pathways and networks of interacting molecules such as genetic regulatory networks<sup>105,106</sup>. These systems are intimately tied to cellular function. Experimental selections for the desired behaviour can often be developed, allowing higher-throughput testing, particularly for the evolution of gene regulation<sup>107</sup>. However, the sequence space associated with these networks is enormous. It encompasses multiple protein coding sequences in addition to their regulatory regions. Mathematical models of how elements interact to generate desired functions can help focus the directed evolution search to components that are more likely to produce the targeted behaviour<sup>108</sup>. For example, an analysis of a mathematical model identified a particular ribosome binding site as having a key role in the target function of a circuit<sup>109</sup>. Experiments verified that mutations to this binding site were effective at altering the target function.

systems, including RNA, DNA regulatory elements, bio-synthetic pathways and genetic regulatory circuits<sup>18–20</sup>) (BOX 1). To understand the power, and the limitations, of directed evolution, it is helpful to view it as a biological optimization process. We therefore introduce the concept of evolution on a fitness landscape in protein sequence space and use this framework to explain directed evolution strategies. Data from laboratory evolution experiments have revealed important features of this fitness landscape and the types of trajectories that can traverse it efficiently. This landscape picture can help explain why decomposing a large functional hurdle into a series of smaller ones and exploiting protein modularity and structural information are useful strategies for dealing with the combinatorial explosion of possible paths in an evolutionary search. This also helps us to appreciate the power of recombination to generate functional sequences with numerous (mostly neutral) mutations, novel combinations of which can give rise to new protein behaviours and therefore new starting points for optimization of protein function.

There is little doubt that directed evolution is one of the most effective and reliable approaches to engineering useful new proteins. Perhaps less well appreciated, however, is how much our understanding of protein function and evolution has been enriched by data from these experiments. Directed evolution allows us to disconnect a protein from its natural context and observe how adaptation to different functional challenges can occur. These experiments can explore the boundaries between biological relevance (the ability of a protein to contribute to the reproductive fitness of an organism) and what is

physically possible (the ability of a protein to carry out a specific function *in vitro* or *in vivo*) in ways that studies on natural proteins alone cannot. Directed evolution can test alternative adaptive scenarios, explore the range of possible solutions to a given functional challenge, examine relationships between different protein properties (for example, trade-offs, in which improvements in one property are accompanied by losses of another) and provide biophysical explanations for evolutionary phenomena. Much has been discovered since these topics were first reviewed in the context of temperature adaptation<sup>21,22</sup>. In this Review, we revisit some of these early lessons and discuss new ones that have emerged.

**Protein fitness landscapes**

In his influential 1970 paper, John Maynard Smith eloquently described protein evolution as a walk from one functional protein to another in the space of all possible protein sequences<sup>23</sup>. He arranged all proteins of length  $L$  such that sequences differing by one amino acid mutation were neighbours. Although the distance between any two sequences is small (that is, it equals the number of mutations required to interconvert the sequences and is therefore  $\leq L$ ), this high-dimensional space contains an incomprehensibly large number of possible proteins. For even a small protein of 100 amino acids there are 20<sup>100</sup> ( $\sim 10^{130}$ ) possible sequences — more than the number of atoms in the universe. Searching in this space for billions of years for solutions to survival, nature has explored only an infinitesimal fraction of the possible proteins<sup>24</sup>. Furthermore, natural evolution keeps only sequences that are biologically relevant; others are discarded, even if they represent solutions to other interesting problems. There are so many proteins waiting to be discovered and we can only dream about the extent of their capabilities. Directed evolution is one way to extend protein function to new, non-natural tasks and convert dreams into actual proteins.

Each sequence in Maynard Smith's protein space can be assigned a 'fitness', which in natural evolution is a measure of the host organism's ability to reproduce in a given environment: fitter organisms reproduce faster and their genes spread throughout the population<sup>25</sup>. When artificial selection is imposed, fitness is defined by the experimenter. High-fitness sequences satisfy all of the criteria for a protein to function as desired, or at least to perform well in the assay used for screening, and might include the ability to recognize one substrate but not another, to be expressed at high levels in a particular host organism, to not aggregate and to have a long lifetime. Protein evolution can then be envisioned as a walk on this high-dimensional fitness landscape, in which regions of higher elevation represent desirable proteins, and iterations of mutation and artificial selection continuously discover new sequences further uphill, with higher fitnesses (FIG. 1a).

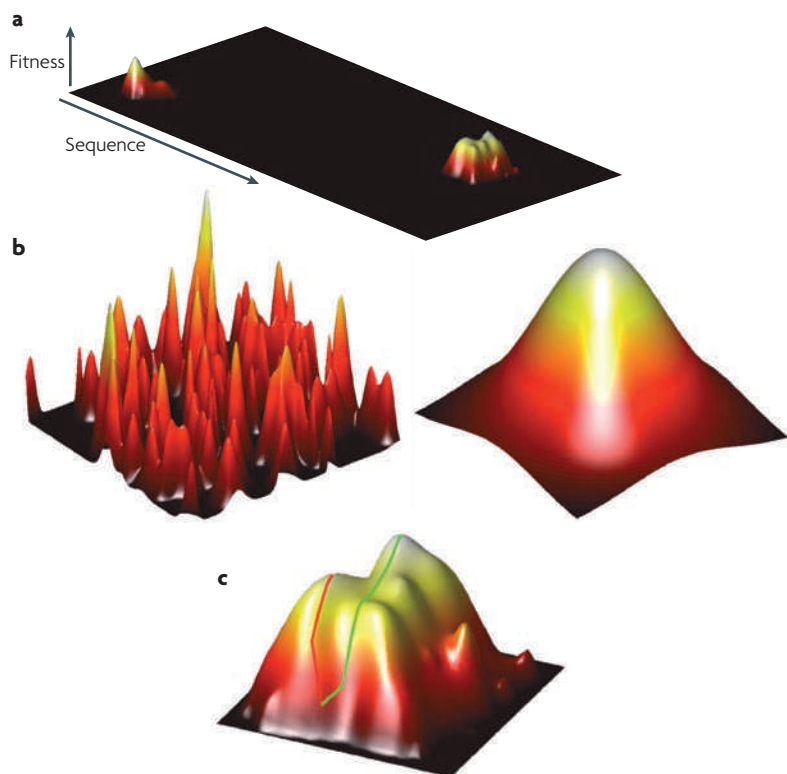
As with any optimization problem, the structure of the objective function (the fitness landscape) influences the effectiveness of a search strategy<sup>26</sup>. Possibilities range from smooth, single-peaked 'Fujiyama' landscapes to rugged, multi-peaked 'Badlands' landscapes<sup>27</sup> (FIG. 1b).

**Fitness landscape**

The mapping from genotype (target sequence) to phenotype (fitness; as measured in the experiment). Directed evolution is an optimization on the fitness landscape.

**Recombination**

A procedure whereby chimeric proteins are created by recombining sequence fragments from different (usually evolutionarily, and therefore structurally, related) parent proteins.



**Figure 1 | Protein fitness landscapes.** Directed protein evolution traverses a fitness landscape in sequence space. This fitness is the measure of how well a given protein performs a target function. **a** | The plot of fitness against sequence creates the landscape for evolution. The transition through black–red–orange–yellow represents increasing fitness. Although the details of this landscape are unknown, it is believed that most sequences do not function (black) and that the rare functional sequences encoding natural proteins are clustered near other functional sequences. However, this popular three-dimensional representation does a poor job of illustrating the numerous paths available to evolution and the numerous sequences in functional regions that do not encode functional proteins<sup>110</sup>. **b** | Similar to natural protein evolution, directed evolution moves along networks of functional proteins that differ by a single amino acid, because selection requires a continuous uphill walk and does not permit the fixation of non-functional sequences. Epistasis occurs when the effect of one mutation depends on the presence of another, which can create landscape ruggedness and local optima. Landscapes could range from the rugged ‘Badlands’ landscape (left panel), which is nearly impossible to climb by mutational steps, to the ‘Fujiyama’ landscape (right panel), in which any beneficial mutation brings the search closer to the optimum<sup>27</sup>. **c** | The presence of local optima might restrict some of the mutational paths uphill (red line). However, the large number of alternative routes leaves plenty of adaptive paths to a fitness optimum (green line).

**Protein sequence space**  
The space of all possible protein sequences arranged such that sequences that differ by single mutations are neighbours.

**Adaptive walk**  
An uphill trajectory on the fitness landscape, in which no deleterious mutations are accepted.

The rougher the landscape, the harder it is for evolution to climb. Local optima create traps that evolution cannot escape from unless a side-step or even a temporary decrease in fitness is permitted, or if multiple simultaneous mutations enable a jump to a new peak. The easiest landscape to climb is one that offers many smooth, uphill paths to the desired fitness (the Fujiyama landscape).

This terrestrial landscape analogy should be interpreted cautiously, however, because it cannot accurately represent the numerous possible paths that evolution can take to higher fitness (or the even larger number of possible downhill paths). Although it is easy to visualize being caught on a local optimum in a three-dimensional landscape, a local optimum in protein sequence space

(in which all possible mutations are deleterious) might be rare, unless stability has been compromised and few new mutations can be accepted. For example, the introduction of stabilizing mutations can increase a protein’s mutational robustness, opening new routes for further adaptation<sup>28,29</sup>.

The vast size of sequence space makes it impossible to characterize (or even model) more than a minute fraction of this fitness surface. Despite this, several important features have emerged from accumulated experimental studies. The first is the low overall density of functional sequences: the vast majority do not code for any functional protein, much less the desired protein<sup>30–32</sup>. Another important feature is the uneven distribution of functional sequences. Although representing a very small fraction of all possible sequences, functional sequences are often next to other functional sequences<sup>33–35</sup>. Maynard Smith recognized that this feature was a requirement for evolution by point mutation to be successful. Evolution can step one mutation at a time only if there is a continuous network of functional proteins, otherwise mutation would always lead to lower fitness and evolution would stop<sup>23</sup>. Proteins are in fact robust to mutation — a significant fraction of possible mutants retain their fold and function<sup>36,37</sup>.

Whereas natural evolution can discover new protein functions along circuitous paths that involve many neutral or even slightly deleterious mutations, directed evolution does not have that luxury. Because the possible evolutionary paths grow exponentially as mutations accumulate and there are too many ways to take neutral or deleterious steps that do not ultimately lead uphill, directed evolution is largely constrained to moving continuously uphill in an adaptive walk<sup>38</sup>. This is often not a severe limitation because many interesting proteins are accessible by short and simple adaptive walks. Although the resulting proteins, or even the mutations, might not be the same as those discovered by more convoluted paths to the same fitness level, they nonetheless provide valuable insights into protein function and routes of adaptation.

**Strategies for directed evolution**

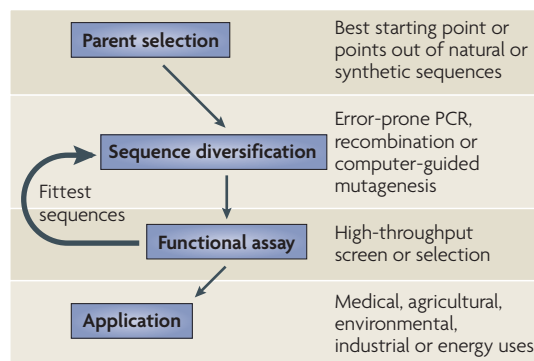
Before we describe some of the key lessons that directed evolution studies have taught us about protein function and evolution, we briefly discuss the experimental strategy. How the experiment is performed obviously influences the outcome and, therefore, the information that is extracted from it. Finding a sequence that performs a desired function in a vast space of possible sequences that is only sparsely populated with functional ones might seem like a daunting task. Inefficient searches of this space could take essentially forever and the task of the protein engineer is to choose a strategy that will reach the objective and do so quickly and easily. Starting with a functional protein, directed evolution uses repeated generations of mutation to create functional variation and selection of the fittest variants to direct the search to higher elevations on the fitness landscape. It involves four key steps (FIG. 2). First, identifying a good starting sequence; second, mutating this ‘parent’

to create a library of variants; third, identifying variants with improved function and last, repeating the process until the desired function is achieved. There are many options for the implementation of each step, the choice of which can greatly affect both the efficiency and the endpoint of an evolutionary search.

Directed evolution (and, indeed, natural evolution) relies on the ability of proteins to function over a wider range of environments or carry out a wider range of functions than might be biologically relevant at a given time and therefore selected for. This ability to tolerate a non-natural environment or to exhibit 'promiscuous' functions at some minimal level provides the jumping-off point for optimization towards that new goal. A good parent protein for directed evolution, therefore, exhibits enough of the desired function that small improvements (expected from a single mutation) can be reliably discerned in a high-throughput screen<sup>38</sup>. It is also easy to work with and sufficiently stable to accommodate multiple, potentially destabilizing, mutations if the target function is some other property. Some proteins are much more evolvable than others<sup>11,29,39,40</sup>. Possible molecular mechanisms that contribute to evolvability have been discussed, including the key role of the chemical mechanism in enzyme functional evolution<sup>41,42</sup> and the idea that evolvable proteins exist in multiple closely related but functionally diverse conformations, the distribution of which is easily altered by mutation<sup>8</sup>. These ideas, however, are still largely speculative, and little other than the ability to accept mutations<sup>29,43</sup> has been conclusively shown in laboratory evolution experiments to contribute directly to allowing one protein to adapt to a new challenge more readily than another protein. A good heuristic indicator of a protein family's evolvability is its natural functional diversity<sup>40,44</sup>. Proteins that have adapted to exhibit a range of functions across their family, for example members of an enzyme family that accepts a wide range of substrates (although individual enzymes in the family might be specific) are likely to be adaptable in the laboratory.

The next step is to create a library of variants. As screening is often the most difficult experimental step, the library is usually created to generate the highest probability of finding improved proteins given the screening capability. Because most mutations are deleterious and multiple mutations frequently inactivate proteins (see below), this usually involves a low mutation rate (one or two amino acid substitutions per gene). If screening is not difficult (for example, there is a good genetic selection), then the library can be constructed to generate the largest potential improvement. This might mean a slightly higher mutation rate<sup>45</sup>. In either case, mutations can be introduced randomly<sup>1</sup> or, if structural or mechanistic information is available, they can be made in a more directed manner<sup>46–48</sup> in an effort to increase the frequency of improved proteins and reduce the load in the next step.

Screening (with high-throughput functional assays) or selection (for example, a genetic selection in which hosts with improved proteins outcompete the others) is used to identify the library members improved in the



**Figure 2 | Overview of directed evolution.** The objective of directed evolution is to create a specific protein function through successive rounds of mutation and selection, starting from a parent protein with a related function. There are numerous options for implementing each step in the process, the choice of which can greatly affect the efficiency and success of the protein sequence optimization. A parent sequence (or sequences) is chosen based on its perceived proximity to the desired function and its evolvability. This parent sequence is then mutated to form a library of new sequences (error-prone PCR or other methods can be used to incorporate mutations randomly, recombination can be used to introduce mutations from other functional sequences or mutation sites can be chosen based on functional and/or structural information). These mutated sequences are evaluated for their ability to perform the desired function using a high-throughput screen or artificial selection. The fittest sequence (or sequences) is used as the parent for the next round of directed evolution, and this process is repeated until the engineering objective is met (usually after five to ten generations).

target property. A good screen or selection accurately assesses the target properties. The rule 'you get what you screen for' is always useful to remember — screening (or selecting) for something else is risky<sup>49</sup>. It is also important not to demand too much improvement in a single generation. The hurdle must be tuned to the screening capacity and should usually be no greater than the improvement that can be provided by a single mutation. If the desired function is beyond what a single mutation can accomplish, the problem can be broken down into a series of smaller ones that can be solved by the accumulation of single mutations, for example by gradually increasing the selection pressure or evolving against a series of intermediate challenges<sup>13</sup>. The process of mutation and selection is repeated until the fitness objective is met; the number of iterations required obviously depends on the starting fitness and the improvement that can be achieved in each round, but is often only five to ten generations.

**Mutational steps.** An evolutionary search relies on the presence of functional diversity in a population, which is the result of underlying genetic variation. At the molecular level, this genetic variation can take many forms; for example, point mutations, insertions, deletions, recombination and circular permutation<sup>50–52</sup>.

To search efficiently and minimize the screening load, the underlying genetic variation should be set to generate the highest probability of improvement. Statistically, random mutations tend to be quite harsh, usually decreasing activity and sometimes destroying it altogether. Typically, 30–50% of single amino acid mutations are strongly deleterious, 50–70% are neutral or slightly deleterious and 0.01–1% are beneficial<sup>11,29,37,53–56</sup>. If the fitness landscape is Fujiyama-like with many smooth uphill paths, only beneficial mutations need to accumulate (either in multiple rounds of mutagenesis and screening or by recombining beneficial mutations found in each round<sup>57,58</sup>) until the desired fitness is reached. In a single-peaked landscape, all beneficial mutations make a cumulative contribution to the desired function and all paths uphill eventually converge to the same optimal solution.

Of course, no real protein landscape consists of a single peak. Most mutations are deleterious and therefore most paths end downhill, with inactive proteins, rather than uphill at fitter sequences. Furthermore, epistatic interactions occur when the presence of one mutation affects the contribution of another mutation. Such epistatic interactions lead to curves in the fitness landscape and constrain evolutionary searches. Extreme forms of epistasis, in which mutations that are negative in one context become beneficial in another (so-called sign epistasis<sup>59</sup>), create local optima on the landscape that can frustrate evolutionary optimization. Epistatic interactions are a ubiquitous feature of protein fitness landscapes<sup>60,61</sup>. We argue, however, that they are not important for most optimizations by directed evolution, which instead follow one of many smooth paths that bypass the more rugged, epistatic routes on this high-dimensional surface<sup>62–64</sup>. Among the numerous mutational trajectories between a starting point and a solution, smooth uphill paths can often be found (FIG. 1c).

**Dealing with the combinatorial explosion.** Knowing of epistatic interactions and local fitness optima, some protein engineers worry about the need to make and find multiple mutations at one time. If multiple mutations are needed to climb the peak, the combinatorial explosion of mutational possibilities makes them especially challenging to find. For even a small protein of 100 amino acids, there are 1,900 single amino acid mutants and more than 1.5 million double mutants. The number of possible sequences increases exponentially with the number of mutations and a complete sampling of even just the double mutants is beyond the capacity of most screens.

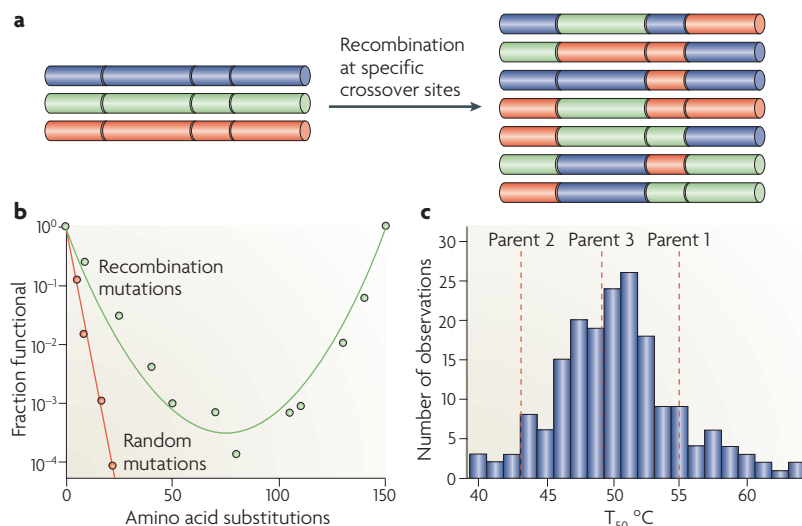
Higher-throughput screening approaches have been developed to enable sampling of more mutants and more combinations of mutations<sup>3,65,66</sup>. These screens can allow multiple paths to be explored simultaneously, increasing the probability of discovering good adaptive routes to higher fitness. However, higher-throughput screens or selections usually come at the cost of decreased accuracy, especially when a surrogate function that is more amenable to high-throughput measurement is substituted for the desired function. Furthermore, increasing

the mutation rate to capture rare synergistic mutations can make it more difficult to identify improved single-mutation variants because common deleterious mutations will tend to mask the rare beneficial ones. It is often better, therefore, to focus on sampling single mutants with a higher quality, lower-throughput screen rather than on increasing the throughput to capture multiple simultaneous mutations. Although a search through single adaptive steps cannot find mutations exhibiting negative epistasis, there are usually other, step-wise adaptive routes to the objective.

The high dimensionality of sequence space that makes finding simultaneous beneficial mutations so difficult can be reduced by taking advantage of structural, functional or phylogenetic information to focus mutations to those residues that are most likely to lead to the desired properties. For example, the modularity of protein structures permits the separate optimization of protein domains<sup>13,67</sup>. Phylogenetic analyses suggest that nature might separately optimize other, structurally non-obvious subunits, or 'sectors'<sup>68</sup>, which could prove to be appropriate targets for directed evolution. The search space can also be reduced by focusing mutations to specific residues in a domain; for example, in an active site or binding pocket in which functional changes might be more likely to occur<sup>11,46,69–71</sup>. This strategy only works, however, when the experimenter is able to select the right residue combinations for random mutagenesis, leaving out the possibility of finding surprising and informative solutions elsewhere. Numerous studies have shown, for example, that many activating mutations lie outside enzyme catalytic sites and exert their influence through mechanisms that might not be obvious from structural analysis<sup>9,10,72</sup>.

**Alternative search strategies.** Evolution by the accumulation of single mutations has proven to be very effective at optimizing a function or property that already exists or can be reached through a series of intermediate steps. Some functions, however, simply can not be reached through a series of small uphill steps and instead require longer jumps that include mutations that would be neutral or even deleterious when made individually. Examples of functions that might require multiple simultaneous mutations include the appearance of a new catalytic activity or an activity on a substrate for which the parent and its single mutants show no measurable activity.

Because most mutations are deleterious, the probability that a variant retains its fold and function declines exponentially with the number of random substitutions<sup>36,37</sup>, and random jumps in sequence space uncover mostly inactive proteins. Thus, new functions are extremely difficult to obtain without altering some aspect of the search. One approach is to create a new starting point — a parent protein with at least some minimal function — and improve that by directed evolution<sup>7</sup>. Where natural examples of a desired function are not practical or might not even exist, emerging protein design tools have identified functional sequences<sup>5</sup>. Expanding the sequence space by the incorporation of



**Figure 3 | Recombination of homologous sequences.** **a** | Recombination generates highly mutated sequence libraries. Multiple homologous parent sequences are divided into fragments, which can be chosen to minimize structural disruption<sup>73</sup>, and these fragments are recombined to form a combinatorial library of chimeric proteins. **b** | The mutations from homologous recombination (green) are much more conservative than random mutations (red). In  $\beta$ -lactamase, chimaeras with high levels of amino acid mutations (around 75) are  $10^{16}$  times more likely to fold than sequences with 75 random mutations<sup>56</sup>. **c** | Chimeric proteins contain new combinations of beneficial mutations. The histogram shows the distribution of thermostabilities ( $T_{50}$ ; the temperature at which 50% of the proteins are inactivated in 10 minutes) of 184 randomly selected chimeric cytochrome P450 enzymes made by structure-guided recombination. The thermostabilities of the three parents are marked by dashed red lines<sup>87</sup>. A significant fraction of chimaeras are more thermostable than any parent from which they are derived. Image in part **b** is modified, with permission, from REF. 56 © (2005) National Academy of Sciences, USA. Image in part **c** is modified, with permission, from *Nature Biotech.* REF. 87 © (2006) Macmillan Publishers Ltd. All rights reserved.

non-natural amino acids can also introduce a whole array of new functions and directed evolution can do the fine-tuning that might be needed to optimize these novel designs<sup>15</sup>. Another approach is to find more conservative ways to make multiple mutations; for example, using computational protein design tools to identify sets of mutations that are likely to be compatible with structure retention<sup>47</sup>.

An approach to making multiple mutations that is used extensively in nature is recombination. Naturally-occurring homologous proteins can be recombined to create genetic diversity within protein sequence libraries<sup>73–75</sup> (FIG. 3a). It has been shown that mutations made by recombination are much less disruptive and generate functional proteins with much higher frequency than random mutations<sup>56</sup> (FIG. 3b). Recombination methods based on DNA sequence hybridization direct crossovers to regions of high sequence identity and are generally limited to sequences that are very similar (with more than 70% identity)<sup>75</sup>, whereas various sequence-independent methods can recombine at random<sup>76,77</sup> or user-specified sites<sup>78,79</sup>. Recombining homologous proteins by choosing crossovers based on structural information allows the construction of libraries of chimeric proteins that simultaneously exhibit high levels of functionality and genetic diversity<sup>80</sup>. In all cases, the

chimeric proteins inherit the best (and worst) residues the parents have to offer, in new combinations that are not observed in nature.

Chimeric proteins can differ by tens or even hundreds of mutations from their parent sequences and still function. The conservative nature of recombination can be exploited to make whole families of novel enzymes. For example, in one set of more than 6,000 chimeric cytochrome P450 proteins with an average of 70 mutations from the closest parent, approximately half folded properly, and at least 75% of these folded P450 proteins displayed enzymatic activity<sup>80</sup>.

The new combinations of residues can give rise to novel properties<sup>81</sup>. Because many of the mutations made by recombination are neutral or nearly neutral, recombination is an efficient way to generate the neutral drifts (the accumulation of neutral mutations) that have been shown to lead to increases in promiscuous functions<sup>82,83</sup> and mutational robustness<sup>84,85</sup>. For example, members of the chimeric cytochrome P450 library exhibited higher enzymatic activity than any of the three parents across a panel of 11 non-native substrates that included substrates on which the parent enzymes showed no measurable activity<sup>86</sup>. Several P450 chimaeras were also more thermostable than the most thermostable parent enzyme, and dozens of thermostable chimaeras could be readily identified based on a small sampling of the library<sup>87</sup> (FIG. 3c). This approach was subsequently used to generate dozens of highly stable, highly active fungal cellobiohydrolase II enzymes that degrade cellulose into fermentable sugars (for biofuels applications, for example)<sup>79</sup>.

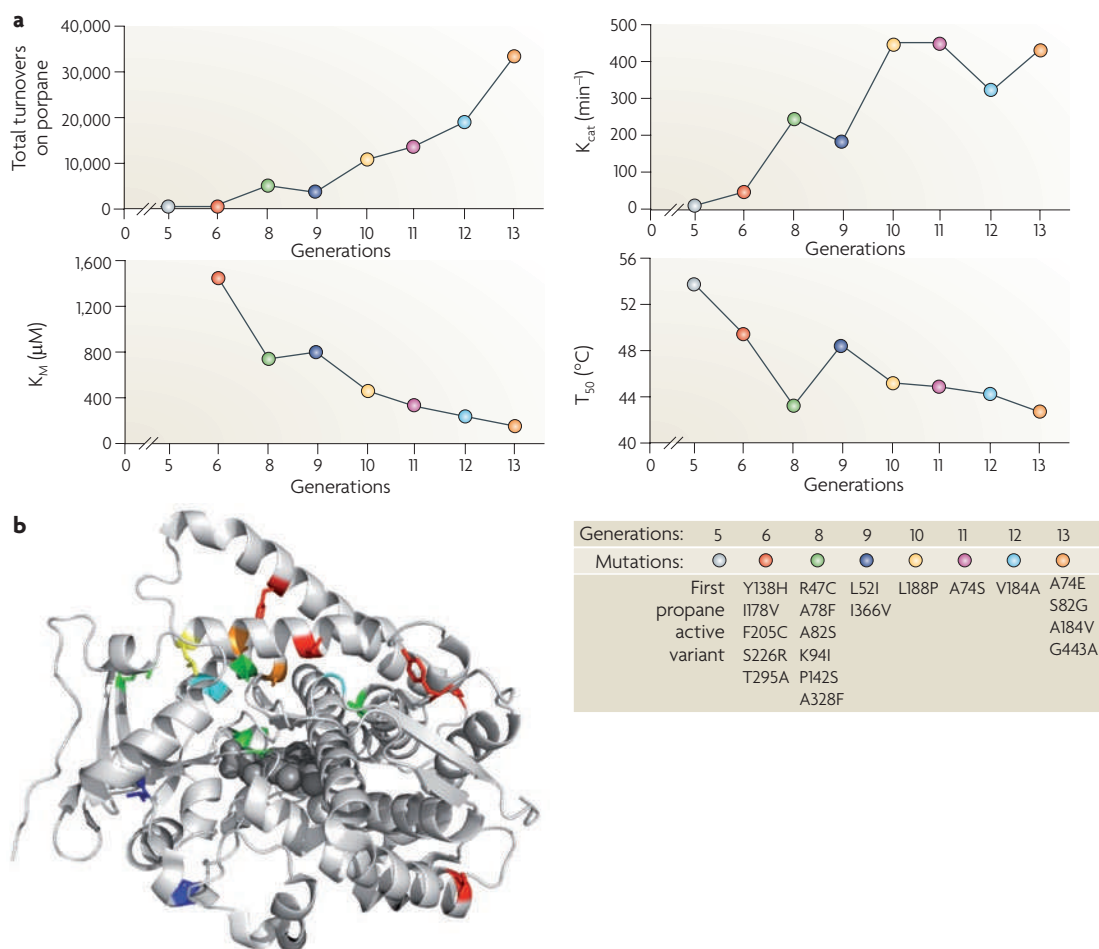
### Lessons from directed evolution

In addition to generating a plethora of novel proteins, directed evolution studies have elucidated available pathways and molecular mechanisms of adaptation, shown a key role for stability in epistasis and evolvability, identified important evolutionary trade-offs in protein properties and revealed the simultaneously conservative and exploratory nature of recombination, all of which have shed light on long-standing questions in protein chemistry and evolutionary biology. First and foremost, directed evolution experiments have shown time and again how rapidly proteins can adapt to exhibit new functions and properties. Protein behaviour can change dramatically on mutating a very small fraction of the protein sequence. Directed evolution also provides a detailed view of the adaptive process.

A directed evolution approach to studying sequence–function relationships circumvents several challenges associated with inferring mechanisms of adaptation using comparisons of evolutionarily related natural amino acid sequences<sup>21,22</sup>. Such studies are confounded by the numerous, mostly neutral mutations that accumulated during divergence of the sequences and the complex and largely unknown selection pressures under which the natural sequences evolved. By contrast, the sequences generated by directed evolution contain a small number of adaptive mutations that accumulated under well-defined selective pressures. Furthermore, performing the evolution

#### Neutral drift

The accumulation of mutations that have little or no effect on a particular protein function. These mutations, however, might affect other properties.



**Figure 4 | Directed evolution of a cytochrome P450 propane monooxygenase.** Cytochrome P450 BM3 from *Bacillus megaterium* catalyses the hydroxylation of long-chain fatty acids and has no measurable activity on propane. This enzyme was converted into a highly efficient and specific propane monooxygenase over 13 rounds of directed evolution<sup>13,111,112</sup>. The large change in substrate specificity was achieved using an incremental approach that first involved screening on an intermediate substrate. Because the native substrate contains a long alkyl chain and the target function was activity on a short alkane, an intermediate-length alkane (octane), towards which the parent enzyme had low but measurable activity, was chosen as the initial directed evolution target. Once high octane activity was achieved, the selective pressure was switched towards activity on propane. **a** | Selected kinetic and biophysical properties of evolutionary intermediates from later generations (with generation five being the first propane active variant)<sup>72</sup>. Total catalytic turnovers (moles of propanol produced per mole of P450),  $K_M$  and  $k_{\text{cat}}$  are reported for propane hydroxylation. Thermostability is shown as  $T_{50}$  (the temperature at which half of the enzyme inactivates after a 10 minute incubation). Variants were selected for total propane activity in all generations, except for generation nine, which was selected for  $T_{50}$ . The mutations acquired during each generation are listed. Even small numbers of mutations can be responsible for large functional changes. **b** | The crystal structure of the fifth generation P450 haem domain (Protein Data Bank identifier: 3CBD), with the locations of the mutations from subsequent generations colour-coded as in part **a**. Beneficial mutations are distributed over the haem domain and many are tens of Å from the catalytic iron. Image in part **a** is modified, with permission, from (REF. 72) © (2008) Elsevier.

in the laboratory permits access to the full ‘fossil record’ of evolutionary intermediates, the sequences, structures and functions of which can be analysed in an attempt to explain how new properties were acquired<sup>10,44,72,88</sup>. Fasan and co-workers analysed selected intermediates that arose during the directed evolution of a cytochrome P450 fatty acid hydroxylase into a highly efficient and highly specific propane monooxygenase<sup>13,72</sup> (FIG. 4). The gradual increase in activity on propane (as measured by total turnovers of propane to propanol — the property targeted during directed evolution) was accompanied by other interesting changes in the enzyme’s behaviour, the

most notable of which was the decrease in thermostability (as measured by  $T_{50}$ ; the temperature at which 50% of the proteins are inactivated in 10 minutes). Activating mutations came at the cost of thermostability, to the point that it became necessary to incorporate stabilizing mutations (generation nine in FIG. 4) before further increases in activity could appear. This apparent trade-off between functionally beneficial mutations and thermostability reflects the fact that most mutations are destabilizing and therefore most activating mutations are also destabilizing. Because evolution favours the most likely solutions over rarer ones, it favours marginal

stability in the absence of selection for higher stability. It also favours properties that are compatible with marginal stability<sup>32</sup>. Such trade-offs have also been shown to constrain the evolution of antibiotic resistance enzymes<sup>89</sup> and will be discussed further below.

The mutations that accumulated in the haem domain of the cytochrome P450 are depicted in FIG. 4b and are colour-coded according to the generation in which they appeared. Many of the mutations that conferred the increased activity on propane lie outside the substrate-binding pocket, where they influence substrate recognition through mechanisms that are difficult to discern from crystal structures or modelling. That the effects of the adaptive mutations are difficult to rationalize, much less predict, underscores how little we understand of how sequence determines protein structure and function. Directed evolution deals with the details of molecular interactions, and it is hoped that these details will eventually help protein design efforts<sup>7</sup>.

Directed evolution can explore alternative evolutionary scenarios; for example, to identify other possible solutions to the same functional challenge or to address whether multiple paths can lead to the same solution, as was done with a laboratory-evolved  $\beta$ -lactamase variant that contains 5 mutations responsible for a 100,000-fold increase in cefotaxime resistance<sup>63</sup>. In this study, the authors constructed variants with all 32 ( $2^5$ ) combinations of the adaptive mutations, representing all intermediate sequences along all 120 (5 factorial) possible mutational pathways. They were able to estimate the probability of each pathway based on the relative change in antibiotic resistance conferred to the bacteria by each mutation along each path. Whereas most of the possible paths were constrained by epistasis and were therefore highly unlikely, there were 18 different, simple uphill walks to the final solution.

**Empirical landscapes.** Even the earliest directed evolution experiments noted how rapidly proteins could adapt to new selective pressures<sup>1,58</sup>, indicating the ready availability of smooth uphill paths in the fitness landscapes. Stability, the ability to tolerate new environments and low-level side reactions or promiscuous functions usually respond well to directed evolution. One study used a well-controlled set of experiments to select for six different promiscuous activities starting from three different enzymes<sup>11</sup>. After 2 rounds of directed evolution, yielding just 1–4 mutations, the promiscuous enzyme activities ( $k_{\text{cat}}/K_M$ ) had increased by up to 150-fold over the activities of the parent enzymes. Interestingly, these newly evolved activities came at little cost to the native enzymatic activities, suggesting a particular robustness of the native functions to mutation and supporting a scenario for evolution of new activities that allows both the native and novel activities to be displayed in the same gene for some period of time<sup>8</sup>.

As well as demonstrating the availability of smooth uphill paths, directed evolution has provided insight into the molecular epistasis that curves the landscapes. Several studies have revealed a key link between stability and epistasis, where the effect of a mutation can be

conditional on the stability of the parent sequence<sup>36,43,90</sup> (FIG. 5). This was demonstrated, for example, in a study of cephalosporin antibiotic resistance mutations in  $\beta$ -lactamase, in which the fitness effects of several active site mutations were found to depend on the presence of a stabilizing M182T mutation<sup>89</sup> (FIG. 5a). These epistatic interactions are the result of catalytically beneficial but destabilizing mutations in the active site that cannot be tolerated unless the stabilizing M182T mutation is present. Without M182T, the active site mutations destabilize the enzyme to the point that total activity is compromised.

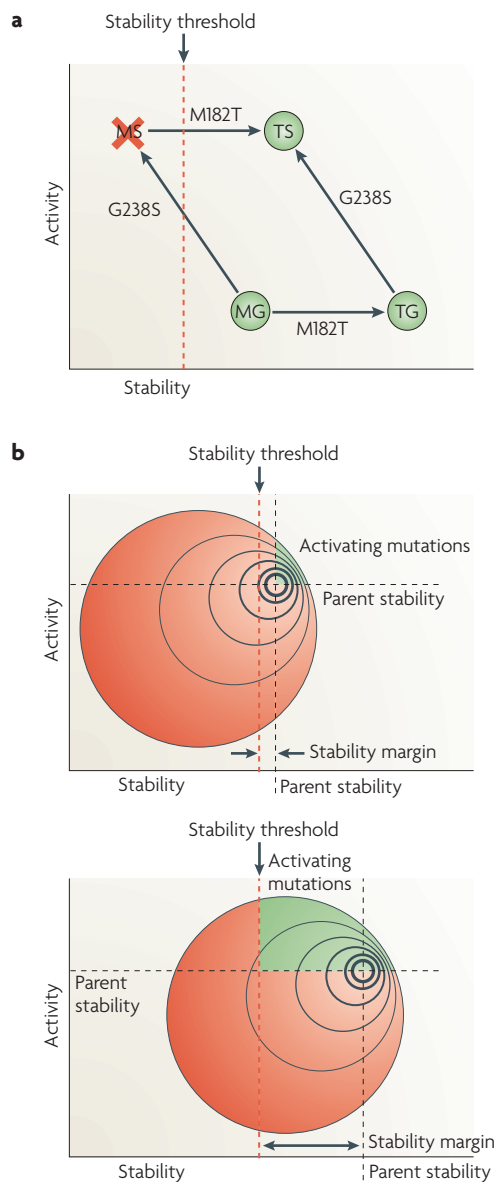
Many examples of stability-mediated epistasis are best explained in terms of a protein stability threshold, whereby stability is under selection only insofar as it allows a protein to fold and function<sup>36,43,91</sup> (FIG. 5). The consequences for evolution are profound: a protein with low stability cannot accept more than a small fraction of the possible mutations because most mutations are destabilizing. Thus, it can become trapped on a local optimum, unable to go further. As illustrated in FIG. 5b, proteins enjoying a larger margin above the minimal stability threshold can explore many more mutations and can therefore continue to adapt to other tasks, such as acquiring activity towards a new substrate or partner<sup>29</sup>. Stability-mediated epistasis is a mechanism whereby neutral mutations can shape the available adaptive pathways during natural evolution as well as in the laboratory. Experience has shown that when an evolutionary search in the laboratory seems to have exhausted all options for further uphill steps, the incorporation of stabilizing mutations is able to open up new adaptive routes<sup>13</sup>.

Despite being performed on different protein folds with selection for different protein functions, the repeated evaluation of thousands of random mutations has revealed the general features of protein fitness landscapes. In addition to the uphill paths that lie alongside numerous less favourable, epistatic routes there are an even larger number of side-steps in the protein fitness landscape. The high frequency of neutral mutations observed during evaluation of random mutant libraries suggests a myriad of sequences with essentially equivalent fitness. This is consistent with the existence of natural protein homologues that differ at several positions, the majority of which are functionally neutral. Even sequences that are highly optimized are probably just one of many potential solutions to a given functional challenge. Indeed, it is probably more accurate to imagine protein evolution occurring on neutral networks, rather than on fitness landscapes in which each neighbour has a different fitness<sup>28,62</sup>. This pervasive neutrality is exploited when families of functional proteins are constructed by recombination of homologous proteins<sup>79,80</sup>.

As discussed above in the context of stability-mediated epistasis, mutations that are neutral in one context might not be neutral in all and therefore can provide new opportunities for evolution. Directed evolution has shown an important role for stabilizing mutations (which can be functionally neutral or only

#### Neutral network

An interconnected network of functionally neutral sequences.



**Figure 5 | Stability threshold and epistasis.** Laboratory evolution studies have found many examples of mutational epistasis that are related to protein stability. The relationship between protein stability and epistasis is best explained in terms of a protein stability threshold, whereby stability is under selection only insofar as it allows a protein to fold and function<sup>36,43,91</sup>. **a** | Epistasis can arise as the result of the protein stability threshold. The G238S active-site mutation in this  $\beta$ -lactamase increases enzyme activity on cephalosporin antibiotics<sup>89</sup>. However, this mutation cannot be accepted into the wild-type sequence (MG) because the resulting protein (MS) is not sufficiently stable. Sequences with the beneficial G238S mutation can instead be reached by first finding the functionally neutral, but stabilizing, M182T mutation (sequence TG) and then incorporating the G238S mutation (sequence TS). **b** | Because most mutations are destabilizing, many of the single mutants of a protein close to the stability threshold (top panel) will be unstable and therefore inactive (red). This leaves few active mutants having beneficial mutations (green). A more stable protein (bottom panel) will be more tolerant to mutation, making more beneficial mutations available.

slightly deleterious) in adaptation. Laboratory evolution experiments have also shown that purposefully accumulated neutral mutations alter promiscuous activities and create new starting points for subsequent adaptive evolution<sup>82,83,92</sup>. Genetic drift and pre-existing diversity might have a similarly important role in natural adaptive evolution<sup>62</sup>.

**Directed evolution to understand natural evolution?**

An overall picture of the protein function landscape is therefore emerging from accumulating directed evolution data. This picture offers a description of the physical features that all proteins (synthetic or natural) must exhibit and the effects of mutations on these features. Extending the lessons learned from directed evolution to natural evolution, however, requires caution because these search processes operate under, for example, different time scales, population sizes, mutation rates and strength of selection. Furthermore, natural evolution works on a different fitness landscape and it is unclear how the protein fitness assayed during directed evolution is related to the organismal fitness that natural evolution optimizes. Differences reflect the consequences of interactions between the protein and the cellular environment and might include constraints related to metabolic burden, regulation, non-specific interactions and other factors.

The ability to disconnect a protein from its *in vivo* function is a valuable asset of directed evolution because it allows the exploration of physically possible proteins without the often-severe constraint of their being biologically relevant and contributing to organismal fitness. Thus, directed evolution can be used to identify which features of proteins are dictated by their physical properties versus those that are due to biological constraints or evolutionary origins and history. The laboratory evolution of the cytochrome P450 propane monooxygenase (FIG. 4), for example, showed the physical possibility, and indeed the ready availability, of such an enzyme, even though known organisms that live on small alkanes use mechanistically and evolutionarily unrelated enzymes for this transformation<sup>72</sup>. Another example is the generation of proteins with combinations of properties that are usually not found in natural proteins, such as high catalytic activity at low temperature and high stability at elevated temperature<sup>21,22</sup>. When properties seem to trade off like this, it might be tempting to infer that such trade-offs are dictated by physical requirements, such as the incompatibility between molecular rigidity that is needed for high stability and the flexibility that is required for catalytic activity<sup>93,94</sup>. If stability and enzymatic activity placed mutually exclusive demands on protein flexibility, then highly active, highly stable enzymes could not exist (a statement that protein engineers did not want to hear). Directed evolution, however, has little trouble finding enzymes that are both highly active and highly stable when the experiments select for both properties<sup>95</sup>. Clearly, such proteins are far rarer than highly active, marginally stable proteins and, without a good reason, natural sequences would not exhibit both features<sup>21,22,32,96</sup>.

## Conclusions

Despite the vast size of sequence space and the complex nature of protein function, the Darwinian algorithm of mutation and selection provides a powerful method to generate proteins with altered functions. This simple uphill walk on a fitness landscape in sequence space works because proteins are wonderfully evolvable and can adapt to new conditions or even take on new functions with only a few mutations.

In addition to providing useful proteins, directed evolution experiments have also taught us how proteins adapt and shed light on processes at work during natural evolution<sup>21,62,97</sup>. These experimental results allow us to look at sequence data in a functional context, providing a bridge between long-separated fields of evolutionary and molecular biology<sup>98</sup>. Directed evolution experiments have been used to address important evolutionary questions about the average effects of mutations,

mechanisms of functional divergence, evolvability and evolutionary constraints<sup>11,85,96,99</sup>.

With the growing number of applications for engineered proteins, directed evolution will continue to be an important strategy for making proteins that are well adapted to new environments and new functions. More advanced high-throughput screens and higher quality sequence libraries will make the searches easier and will enable evolution to solve increasingly complex problems. Advances in our understanding of proteins can be incorporated into library design, and the rapidly decreasing cost of DNA synthesis will relieve many sequence construction constraints. Directed evolution will help teach us how biological systems adapt to changing demands; it might also help us to address some of today's most challenging problems of providing effective treatments for disease or producing fuels and chemicals from renewable resources.

- Chen, K. & Arnold, F. H. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl Acad. Sci. USA* **90**, 5618–5622 (1993).  
**The first demonstration of directed evolution by successive rounds of mutagenesis and screening — a strategy now widely used to engineer enzymes.**
- Reetz, M. T. Combinatorial and evolution-based methods in the creation of enantioselective catalysts. *Angew. Chem. Int. Ed. Engl.* **40**, 284–310 (2001).
- Boder, E. T., Midelfort, K. S. & Wittrup, K. D. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc. Natl Acad. Sci. USA* **97**, 10701–10705 (2000).
- Campbell, R. E. *et al.* A monomeric red fluorescent protein. *Proc. Natl Acad. Sci. USA* **99**, 7877–7882 (2002).
- Jiang, L. *et al.* *De novo* computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
- Bolon, D. N. & Mayo, S. L. Enzyme-like proteins by computational design. *Proc. Natl Acad. Sci. USA* **98**, 14274–14279 (2001).
- Rothlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).  
**This study shows how computational design and directed evolution can be combined to create and improve new functions.**
- Tokuriki, N. & Tawfik, D. Protein dynamism and evolvability. *Science* **324**, 203 (2009).
- Shimotohno, A., Oue, S., Yano, T., Kuramitsu, S. & Kagamiyama, R. Demonstration of the importance and usefulness of manipulating non-active-site residues in protein design. *J. Biochem.* **129**, 943–948 (2001).
- Spiller, B., Gershenson, A., Arnold, F. & Stevens, R. A structural view of evolutionary divergence. *Proc. Natl Acad. Sci. USA* **96**, 12305–12310 (1999).
- Aharoni, A. *et al.* The 'evolvability' of promiscuous protein functions. *Nature Genetics* **37**, 73–76 (2005).  
**This work shows that enzymes with promiscuous activities that are improved in directed evolution tend to retain their native activities.**
- Sarkar, I., Hauber, I., Hauber, J. & Buchholz, F. HIV-1 proviral DNA excision using an evolved recombinase. *Science* **316**, 1912–1915 (2007).
- Fasan, R., Chen, M. M., Crook, N. C. & Arnold, F. H. Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting native-like catalytic properties. *Angew. Chem. Int. Ed. Engl.* **46**, 8414–8418 (2007).  
**An intermediate selective pressure (activity on octane) was used to direct the evolution of a P450 for high activity on propane — an activity which the original enzyme, a fatty acid hydroxylase, does not exhibit.**
- Reetz, M. T., D. Carballeira, J. & Vogel, A. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed. Engl.* **45**, 7745–7751 (2006).
- An alternative directed evolution strategy using structure information to focus mutations achieved a large increase in enzyme stability.**
- Yoo, T. H., Link, A. J. & Tirrell, D. A. Evolution of a fluorinated green fluorescent protein. *Proc. Natl Acad. Sci. USA* **104**, 13887–13890 (2007).
- Tsien, R. Constructing and exploiting the fluorescent protein paintbox (Nobel Lecture). *Angew. Chem. Int. Ed. Engl.* **48**, 5612–5626 (2009).
- Shaner, N. *et al.* Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature Biotech.* **22**, 1567–1572 (2004).
- Yokobayashi, Y., Weiss, R. & Arnold, F. H. Directed evolution of a genetic circuit. *Proc. Natl Acad. Sci. USA* **99**, 16587–16591 (2002).
- Beaudry, A. A. & Joyce, G. F. Directed evolution of an RNA enzyme. *Science* **257**, 635–641 (1992).
- Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proc. Natl Acad. Sci. USA* **102**, 12678–12683 (2005).
- Arnold, F. H., Wintrose, P. L., Miyazaki, K. & Gershenson, A. How enzymes adapt: lessons from directed evolution. *Trends Biochem. Sci.* **26**, 100–106 (2001).
- Wintrose, P. L. & Arnold, F. H. Temperature adaptation of enzymes: lessons from laboratory evolution. *Adv. Protein Chem.* **55**, 161–225 (2000).
- Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).  
**A beautiful description of protein evolution as a walk through sequence space.**
- Mandecki, W. The game of chess and searches in protein sequence space. *Trends Biotechnol.* **16**, 200–202 (1998).
- Wright, S. Evolution in mendelian populations. *Genetics* **16**, 0097–0159 (1931).
- Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comp.* **1**, 67–82 (1997).
- Kauffman, S. A. & Weinberger, E. D. The NK model of rugged fitness landscapes and its application to maturation of the immune-response. *J. Theor. Biol.* **141**, 211–245 (1989).
- Wagner, A. Robustness and evolvability: a paradox resolved. *Proc. Biol. Sci.* **275**, 91–100 (2008).
- Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874 (2006).  
**This work showed that excess stability provides increased mutational tolerance and allows greater room for adaptation in directed evolution.**
- Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
- Axe, D. D. Estimating the prevalence of protein sequences adopting functional enzyme folds. *J. Mol. Biol.* **341**, 1295–1315 (2004).
- Taverna, D. M. & Goldstein, R. A. Why are proteins marginally stable? *Proteins* **46**, 105–109 (2002).
- Govindarajan, S. & Goldstein, R. A. Evolution of model proteins on a foldability landscape. *Proteins* **29**, 461–466 (1997).
- Xia, Y. & Levitt, M. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins* **55**, 107–114 (2004).
- Taverna, D. M. & Goldstein, R. A. Why are proteins so robust to site mutations? *J. Mol. Biol.* **315**, 479–484 (2002).
- Bloom, J. D. *et al.* Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA* **102**, 606–611 (2005).
- Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101**, 9205–9210 (2004).
- Arnold, F. H. Directed evolution: creating biocatalysts for the future. *Chem. Eng. Sci.* **51**, 5091–5102 (1996).
- England, J. L. & Shakhnovich, E. I. Structural determinant of protein designability. *Phys. Rev. Lett.* **90**, 218101 (2003).
- O'Loughlin, T. L., Patrick, W. M. & Matsumura, I. Natural history as a predictor of protein evolvability. *Protein Eng. Des. Sel.* **19**, 439–442 (2006).
- Umeno, D., Tobias, A. V. & Arnold, F. H. Diversifying carotenoid biosynthetic pathways by directed evolution. *Microbiol. Mol. Biol. Rev.* **69**, 51–78 (2005).
- Glasner, M. E., Gert, J. A. & Babbitt, P. C. Evolution of enzyme superfamilies. *Curr. Opin. Chem. Biol.* **10**, 492–497 (2006).
- Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).  
**This study shows the role of protein stability in epistasis.**
- Claren, J., Malisi, C., Hocker, B. & Sterner, R. Establishing wild-type levels of catalytic activity on natural and artificial (β)<sub>3</sub>-barrel protein scaffolds. *Proc. Natl Acad. Sci. USA* **106**, 3704–3709 (2009).
- Drummond, D. A., Iversen, B. L., Georgiou, G. & Arnold, F. H. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *J. Mol. Biol.* **350**, 806–816 (2005).
- Reetz, M. T., Bocella, M., Carballeira, J. D., Zha, D. X. & Vogel, A. Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew. Chem. Int. Ed. Engl.* **44**, 4192–4196 (2005).
- Treynor, T. P., Vizcarra, C. L., Nedelcu, D. & Mayo, S. L. Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc. Natl Acad. Sci. USA* **104**, 48–53 (2007).
- Yoshikuni, Y., Ferrin, T. E. & Keasling, J. D. Designed divergent evolution of enzyme function. *Nature* **440**, 1078–1082 (2006).
- You, L. & Arnold, F. Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide. *Protein Eng.* **9**, 77–83 (1996).

50. Fujii, R., Kitaoka, M. & Hayashi, K. RAISE: a simple and novel method of generating random insertion and deletion mutations. *Nucl. Acids Res.* **34**, e30 (2006).
51. Qian, Z. & Lutz, S. Improving the catalytic activity of Candida antarctica lipase B by circular permutation. *J. Am. Chem. Soc.* **127**, 13466–13467 (2005).
52. Neylon, C. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucl. Acids Res.* **32**, 1448–1459 (2004).
53. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. Systematic mutation of bacteriophage-T4 lysozyme. *J. Mol. Biol.* **222**, 67–87 (1991).
54. Axe, D. D., Foster, N. W. & Fersht, A. R. A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry* **37**, 7157–7166 (1998).
55. Shafikhani, S., Siegel, R. A., Ferrari, E. & Schellenberger, V. Generation of large libraries of random mutants in *Bacillus subtilis* by PCR-based plasmid multimerization. *Biotechniques* **23**, 304–310 (1997).
56. Drummond, D. A., Silberg, J. J., Meyer, M. M., Wilke, C. O. & Arnold, F. H. On the conservative nature of intragenic recombination. *Proc. Natl Acad. Sci. USA* **102**, 5380–5385 (2005).
57. Moore, J. C., Jin, H.-M., Kuchner, O. & Arnold, F. H. Strategies for the *in vitro* evolution of protein function: Enzyme evolution by random recombination of improved sequences. *J. Mol. Biol.* **272**, 336–347 (1997).
58. Stemmer, W. P. Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* **370**, 389–391 (1994).
59. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).
60. Aita, T. *et al.* Surveying a local fitness landscape of a protein with epistatic sites for the study of directed evolution. *Biopolymers* **64**, 95–105 (2002).
61. Hayashi, Y. *et al.* Experimental rugged fitness landscape in protein sequence space. *PLoS ONE* **1**, e96 (2006).
62. Bloom, J. D. & Arnold, F. H. In the light of directed evolution: pathways of adaptive protein evolution. *Proc. Natl Acad. Sci. USA* **106**, 9995–10000 (2009).
63. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
- In this study, the authors construct every evolutionary intermediate between two sequences and evaluate the probability of all possible adaptive pathways.**
64. Reetz, M. T. & Sanchis, J. Constructing and analyzing the fitness landscape of an experimental evolutionary process. *ChemBiochem* **9**, 2260–2267 (2008).
65. Bernath, K., Magdassi, S. & Tawfik, D. S. Directed evolution of protein inhibitors of DNA-nucleases by *in vitro* compartmentalization (IVC) and nano-droplet delivery. *J. Mol. Biol.* **345**, 1015–1026 (2005).
66. Liu, L., Li, Y., Liotta, D. & Lutz, S. Directed evolution of an orthogonal nucleoside analog kinase via fluorescence-activated cell sorting. *Nucl. Acids Res.* **37**, 4472–4481 (2009).
67. Fischbach, M. A., Lai, J. R., Roche, E. D., Walsh, C. T. & Liu, D. R. Directed evolution can rapidly improve the activity of chimeric assembly-line enzymes. *Proc. Natl Acad. Sci. USA* **104**, 11951–11956 (2007).
68. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774–786 (2009).
69. Matsumura, I. & Ellington, A. D. *In vitro* evolution of  $\beta$ -glucuronidase into a  $\beta$ -galactosidase proceeds through non-specific intermediates. *J. Mol. Biol.* **305**, 331–339 (2001).
70. Park, S. *et al.* Focusing mutations into the *P. fluorescens* esterase binding site increases enantioselectivity more effectively than distant mutations. *Chem. Biol.* **12**, 45–54 (2005).
71. Paramesvaran, J., Hibbert, E. G., Russell, A. J. & Dalby, P. A. Distributions of enzyme residues yielding mutants with improved substrate specificities from two different directed evolution strategies. *Protein Eng. Des. Sel.* **22**, 401–411 (2009).
72. Fasan, R., Meharena, Y. T., Snow, C. D., Poulos, T. L. & Arnold, F. H. Evolutionary history of a specialized P450 propane monooxygenase. *J. Mol. Biol.* **385**, 1069–1080 (2008).
73. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. Protein building blocks preserved by recombination. *Nature Struct. Biol.* **9**, 553–558 (2002).
74. Hansson, L. O., Bolton-Grob, R., Massoud, T. & Mannervik, B. Evolution of differential substrate specificities in  $\mu$  class glutathione transferases probed by DNA shuffling. *J. Mol. Biol.* **287**, 265–276 (1999).
75. Cramer, A., Raillard, S., Bermudez, E. & Stemmer, W. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288–291 (1998).
76. Ostermeier, M., Shim, J. H. & Benkovic, S. J. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotech.* **17**, 1205–1209 (1999).
77. Lutz, S., Ostermeier, M., Moore, G. L., Maranas, C. D. & Benkovic, S. J. Creating multiple-crossover DNA libraries independent of sequence identity. *Proc. Natl Acad. Sci. USA* **98**, 11248–11253 (2001).
78. Hiraga, K. & Arnold, F. H. General method for sequence-independent site-directed chimeragenesis. *J. Mol. Biol.* **330**, 287–296 (2003).
79. Heinzelman, P. *et al.* A family of thermostable fungal cellulases created by structure-guided recombination. *Proc. Natl Acad. Sci. USA* **106**, 5610–5615 (2009).
80. Otey, C. R. *et al.* Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol.* **4**, e112 (2006).
- An example of the use of recombination to create thousands of chimeric enzymes with numerous mutations and new properties that are not exhibited by the parent enzymes.**
81. Campbell, R. K., Bergert, E. R., Wang, Y. H., Morris, J. C. & Moyle, W. R. Chimeric proteins can exceed the sum of their parts: implications for evolution and protein design. *Nature Biotech.* **15**, 439–443 (1997).
82. Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* **2**, 17 (2007).
83. Amitai, G., Gupta, R. D. & Tawfik, D. S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1**, 67–78 (2007).
84. Bershtein, S., Goldin, K. & Tawfik, D. S. Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.* **379**, 1029–1044 (2008).
85. Bloom, J. D. *et al.* Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.* **5**, 29 (2007).
86. Landwehr, M., Carbone, M., Otey, C. R., Li, Y. & Arnold, F. H. Diversification of catalytic function in a synthetic family of chimeric cytochrome P450s. *Chem. Biol.* **14**, 269–278 (2007).
87. Li, Y. *et al.* A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nature Biotech.* **25**, 1051–1056 (2007).
88. Counago, R., Chen, S. & Shamoo, Y. *In vivo* molecular evolution reveals biophysical origins of organismal fitness. *Mol. Cell* **22**, 441–449 (2006).
89. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J. Mol. Biol.* **320**, 85–95 (2002).
- An excellent example of stability-mediated epistasis.**
90. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).
91. Bloom, J. D., Arnold, F. H. & Wilke, C. O. Breaking proteins with mutations: threads and thresholds in evolution. *Mol. Syst. Biol.* **3**, 76 (2007).
92. Gupta, R. D. & Tawfik, D. S. Directed enzyme evolution via small and effective neutral drift libraries. *Nature Methods* **5**, 939–942 (2008).
93. Somero, G. N. Proteins and temperature. *Annu. Rev. Physiol.* **57**, 43–68 (1995).
94. Fields, P. A. Protein function at thermal extremes: balancing stability and flexibility. *Comp. Biochem. Physiol. A.* **129**, 417–431 (2001).
95. Giver, L., Gershenson, A., Freskgard, P. O. & Arnold, F. H. Directed evolution of a thermostable esterase. *Proc. Natl Acad. Sci. USA* **95**, 12809–12813 (1998).
96. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How protein stability and new functions trade off. *PLoS Comput. Biol.* **4**, e1000002 (2008).
97. Peisajovich, S. G. & Tawfik, D. S. Protein engineers turned evolutionists. *Nature Methods* **4**, 991–994 (2007).
98. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Rev. Genet.* **8**, 675–688 (2007).
99. Miller, S. P., Lunzer, M. & Dean, A. M. Direct demonstration of an adaptive constraint. *Science* **314**, 458–461 (2006).
100. Earl, D. J. & Deem, M. W. Evolvability is a selectable trait. *Proc. Natl Acad. Sci. USA* **101**, 11531–11536 (2004).
101. Ellington, A. D. & Szostak, J. W. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
102. Robertson, D. & Joyce, G. Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344**, 467–468 (1990).
103. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
104. Lincoln, T. & Joyce, G. Self-sustained replication of an RNA enzyme. *Science* **323**, 1229–1232 (2009).
105. Chatterjee, R. & Yuan, L. Directed evolution of metabolic pathways. *Trends Biotech.* **24**, 28–38 (2006).
106. Schmidt-Dannert, C. Directed evolution of single proteins, metabolic pathways, and viruses. *Biochemistry* **40**, 13125–13136 (2001).
107. Collins, C. H., Leadbetter, J. R. & Arnold, F. H. Dual selection enhances the signaling specificity of a variant of the quorum-sensing transcriptional activator LuxR. *Nature Biotech.* **24**, 708–712 (2006).
108. Haseltine, E. L. & Arnold, F. H. Synthetic gene circuits: design with directed evolution. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 1–19 (2007).
109. Feng, X. *et al.* Optimizing genetic circuits by global sensitivity analysis. *Biophys. J.* **87**, 2195–2202 (2004).
110. Gavrilits, S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12**, 307–312 (1997).
111. Glieder, A., Farinas, E. T. & Arnold, F. H. Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase. *Nature Biotech.* **20**, 1135–1139 (2002).
112. Peters, M. W., Meinhold, P., Glieder, A. & Arnold, F. H. Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. *J. Am. Chem. Soc.* **125**, 13442–13450 (2003).

## Acknowledgements

The authors acknowledge support from the U.S. Army Research Office, Department of Energy, National Science Foundation and the National Institutes of Health.

## DATABASES

PDB: <http://www.rcsb.org>  
3CBD

## FURTHER INFORMATION

Frances H. Arnold's homepage: <http://www.che.caltech.edu/groups/aha>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF