De novo design of biocatalysts
Daniel N Bolon*, Christopher A Voigt* and Stephen L Mayo[†]

The challenging field of *de novo* enzyme design is beginning to produce exciting results. The application of powerful computational methods to functional protein design has recently succeeded at engineering target activities. In addition, efforts in directed evolution continue to expand the transformations that can be accomplished by existing enzymes. The engineering of completely novel catalytic activity requires traversing inactive sequence space in a fitness landscape, a feat that is better suited to computational design. Optimizing activity, which can include subtle alterations in backbone conformation and protein motion, is better suited to directed evolution, which is highly effective at scaling fitness landscapes towards maxima. Improved rational design efforts coupled with directed evolution should dramatically improve the scope of *de novo* enzyme design.

Addresses
*Biochemistry and Molecular Biophysics Option, California Institute of Technology, mail code 147–75, Pasadena, California 91125, USA
[†]Correspondence: Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, mail code 147-75, Pasadena, California 91125, USA; e-mail: steve@mayo.caltech.edu

Abbreviations
FACS   fluorescence-activated cell sorter
k_{cat} catalytic rate constant
K_M Michaelis–Menten equilibrium constant
PRAI   phosphoribosyl-anthranilate isomerase
TSA   transition-state analogue

Introduction
There are many properties of natural enzymes that make them appealing for chemical production. Natural enzymes are highly efficient catalysts with typical apparent second-order rate constants (k_{cat}/K_M) from 10^6 to 10^8 $M^{-1}sec^{-1}$ [1]. Naturally occurring enzymes are able to effectively catalyze difficult chemical transformations under mild conditions. Rate accelerations over background (k_{cat}/k_{uncat}) can reach 10^{17} [1]. By binding substrates in well-defined active sites with catalytic side chains geometrically positioned, enzymes are highly selective catalysts. The enantioselectivity of enzymes is increasingly being used in the production of chiral pharmaceuticals [2]. In addition, the general selectivity of enzymes results in few by-products, which, combined with their energy-efficient operation under mild conditions, makes enzymes environmentally friendly [3,4]. Because of the high selectivity of enzymes, relevant natural enzymes do not exist for many industrially important transformations. For this reason, there is great practical interest in the design of enzymes with novel activities.

This review highlights the strengths of rational design, catalytic antibodies, and directed evolution for the design of novel enzymes. The strengths of different techniques are highly complementary, suggesting that combined approaches may have significant advantages over approaches that rely on a single method [5••,6,7••].
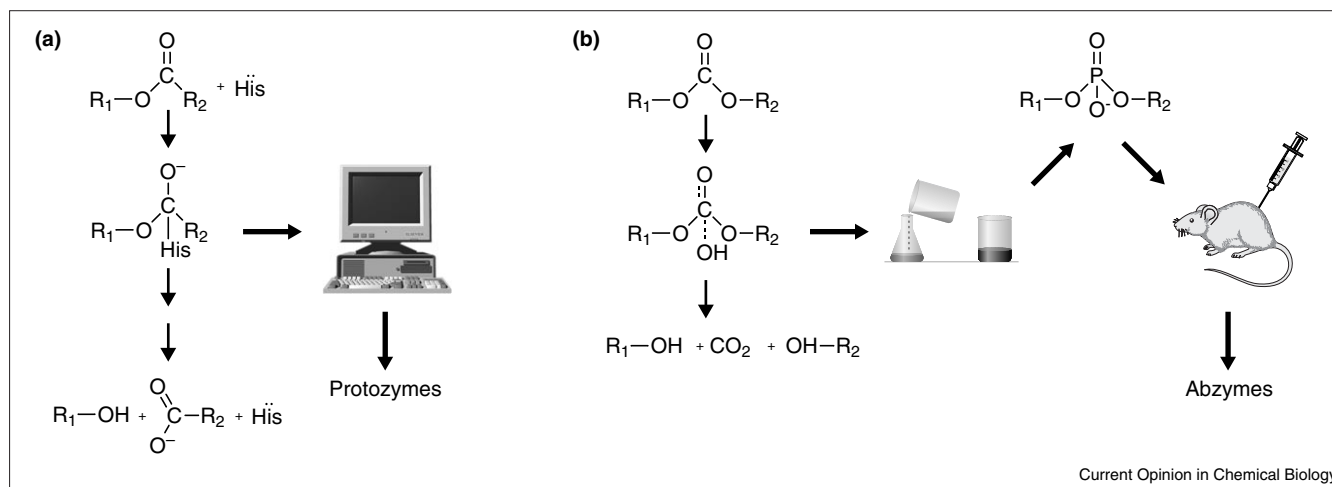
Rational computational design
Computational protein design [8] starts with the coordinates of a protein main chain and uses a force field to identify sequences and geometries of amino acids that are optimal for stabilizing the backbone geometry. Even for small proteins, the number of possible sequences far exceeds that which can be exhaustively searched. The development of powerful search algorithms to find optimal solutions has provided a major stimulus to the field [9]. Studies correlating predicted and experimental stability have since been used to iteratively improve an empirical force field for protein design calculations [10]. The combination of predictive force fields and fast search algorithms is now being applied to functional protein design [11•].

A novel active site for activated ester hydrolysis was computationally designed in the scaffold of the inert protein thioredoxin [12••]. In a method analogous to the design of catalytic antibodies, the high-energy state of a reaction was modeled with a side chain geometrically oriented for catalysis (Figure 1a). Numerous low-energy geometries were modeled to generate a library of high-energy-state rotamers for the design. To identify a novel active site, optimal sequences were computed with the high-energy-state rotamers at different positions in the protein. This active-site scan explicitly limited the search to a relevant phase space where the catalytic residue could be positioned to properly interact with the substrate. Ranking of different active sites from the scan on the basis of substrate recognition successfully predicted 'protozymes' that were experimentally shown to catalyze the reaction by the designed method of histidine-mediated nucleophilic catalysis. Although the catalytic activity of the designed protein was very modest, on the same order as the initial catalytic antibodies, the generalness of the approach promises interesting future results, including the ability to design proteins to catalyze reactions that are inaccessible by natural enzymes.

The diverse and powerful chemistry of metals makes design of metalloproteins a promising approach to enzyme design [13]. Early pioneering work included the development of a computational method to identify sites in proteins capable of using side chains to complex metal atoms [14]. By modeling iron with one primary coordination sphere ligated with dioxygen, a number of metalloproteins were designed in the thioredoxin fold [15••]. These proteins were experimentally shown to bind iron and catalyzed a variety of oxygen chemistries.

Figure 1

The similar procedures for computationally designed protozymes and traditional TSA-induced catalytic antibodies are illustrated. **(a)** Computational enzyme design models a reaction with the involvement of at least one catalytic amino acid side chain. A high-energy state of the reaction is then computationally modeled and used to identify active sites capable of binding substrate and orienting catalytic side chains. **(b)** Catalytic antibody design identifies a transition state in a reaction pathway. TSAs are then synthesized and used to elicit antibodies with binding sites that are complementary to the transition state.

The engineering of a completely new function into an inert protein scaffold is likely to require multiple coupled mutations where individual mutations may have no effect. In terms of a fitness landscape, engineering a novel function is likely to require traversing inactive sequence space (Figure 2). Computational design benefits from the ability to search a larger space than possible by purely experimental methods and is well suited to designing novel function. Computational design is not currently an effective tool for increasing fitness, which may occur through subtle changes in backbone structure and protein dynamics. However, current research is actively exploring these issues. The use of a flexible main chain was recently used to successfully design a left-handed coiled coil [16]. The backbone geometry and sequence of a β-turn was recently designed in order to increase the folding rate and improve thermodynamic stability [17•]. Backbone flexibility has also been explored in the design of protein cores [18,19].

Rational homology-driven design
The rational redesign of enzymes uses structural and/or sequence homology to graft desired properties of one enzyme onto another. The goals of homology-driven experiments include engineering binding sites to fit different substrates as well as construction of new catalytic residues to alter mechanism and function [5••]. Rational design generally focuses on changing residues that contact substrate. The efficiency of most homology-designed enzymes has been poor relative to that of natural enzymes; however, some experiments have resulted in strikingly efficient enzymes.

The engineering of a catalytic Ser–His–Asp triad into a peptidyl-prolyl isomerase resulted in a remarkably efficient proline-specific endopeptidase [20]. Based on structural alignment of 4-chlorobenzoyl-CoA dehalogenase and crotonase, engineering of two catalytically active acid/base residues as well as six positioning residues resulted in efficient crotonase activity in the dehalogenase [21•]. Both of these designs started with significant catalytic machinery in place to polarize a carbonyl group within the substrate, and successfully included additional catalytic residues in order to support a related reaction.

Homology-driven design has greatly improved our understanding of enzyme mechanisms and substrate specificity. This information together with more accurate and detailed physical models should greatly benefit future enzyme designs.

Catalytic antibodies
The use of transition-state analogues (TSAs) to elicit catalytic antibodies (abzymes) effectively uses the diversity of the immune system to identify proteins with substrate affinity and catalytic rate enhancement. The method includes synthesis of a TSA, immunization, isolation of monoclonal antibodies via hybridomas, and screening for catalytically effective antibodies (Figure 1b). The technique of immunization with a TSA has been used to identify antibodies that successfully catalyze a number of different reactions [22•]. However, the catalytic efficiency of the resulting antibodies has been low relative to natural enzymes [23].

In a significant advance, catalytic antibodies that utilize a nucleophilic mechanism were selected by reactive immunization and resulted in efficient catalysts [22•]. Rather than a TSA, a mechanism-based inhibitor was used to

Figure 2

Fitness landscapes, illustrating a path to go from high fitness for Function A to high fitness for a novel Function B. The first jump to attain an initial level of Function B requires crossing an inactive region of sequence space. The extremely large sequence space that can be searched by computational design allows it to effectively jump over this inactive space. Directed evolution should then be an effective tool to iteratively improve Function B to a high level of activity.

Current Opinion in Chemical Biology
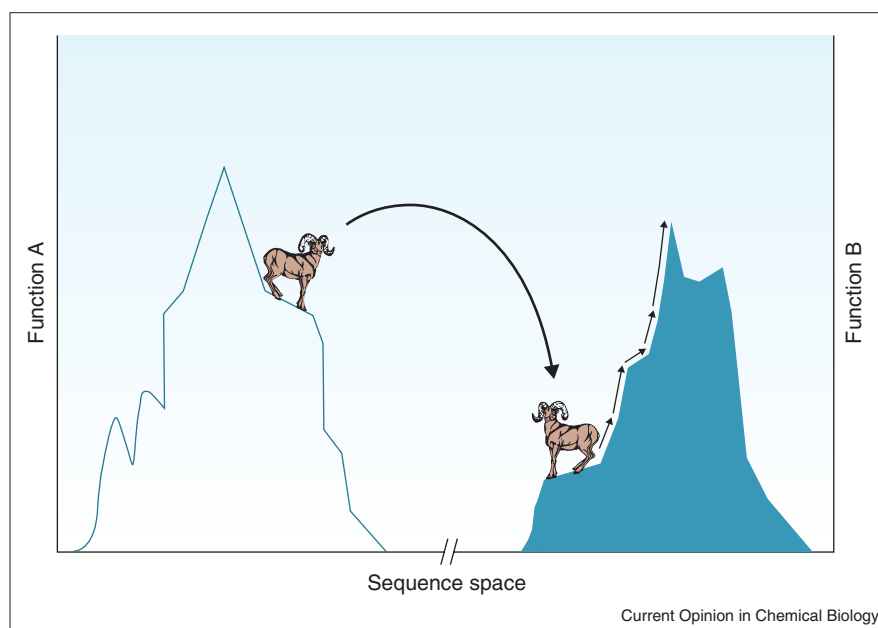
elicit the immune response. Antibodies that formed stable covalent attachments to the suicide inhibitor were effectively selected. This method was employed in the selection of an efficient abzyme with a nucleophilic lysine for aldol condensations [24]. The efficiency of this aldolase demonstrates the effectiveness of covalent catalysis. The ability to select for powerful catalytic groups and active sites with high transition-state specificity could theoretically yield more efficient catalysts.

Directed evolution

Directed evolution has emerged as a popular method for protein engineering [4,25]. Mimicking natural evolution, an initial parent gene is chosen and a diverse library of offspring genes is created through mutagenesis or recombination. A screen or selection is applied to the library and the mutants that exhibit the greatest improvement in the desired properties are chosen to become the parents to the next generation. This iterative search has generated large improvements in properties such as activity, specificity (including enantioselectivity [26]) and stability, and has been used to evolve systems of enzymes, such as found in metabolic pathways [27••]. Directed evolution often discovers these improvements by making a few amino acid substitutions that collectively have an important functional effect. For this reason, the method generally requires a starting protein with some activity towards the desired reaction [28].

One of the critical steps in directed evolution is the creation of a screen or selection that is rapid enough to process millions of mutants while accurately measuring a desired property [29]. Screening by monitoring formation of the

exact product of interest is desirable, but limits throughput. For this reason, the development of ingenious high-throughput methods is an active area of research. The use of fluorogenic substrates for high-throughput screening is a productive field that has been recently reviewed [30]. Other methods include the linking of substrate to phage particles, after which, product binding can be used to enrich for phage displaying active enzymes [31–34]. The *in vivo* three-hybrid system is based on substrate competition [35]. In this system, the substrate of the reaction competes with a synthetic dimerizer and inactivates a transcription factor. Conversion of substrate to product shifts the system in favor of dimerization and can be followed by the expression of a reporter gene.

Most selection schemes indirectly screen for catalysis, often through binding events. One of the difficulties of this is that the mutants that are discovered tend to bind more tightly to the substrates, but are not necessarily better catalysts. To overcome this limitation, Iverson and co-workers [36•,37] coupled catalytic turnover with fluorescence, so that high throughput could be achieved using FACS (fluorescence-activated cell sorter). A fluorophore (F) and a quencher (Q) were tethered to a peptide substrate. Upon action of a protease, the substrate was cleaved, separating the F and Q moieties and disrupting the intramolecular quenching. Fluorescence is thus linked to a turnover event.

Because screening is often the limiting step, there have been several studies to optimize the mutation or recombination rate with respect to the number of mutants that can be screened [38••]. Several groups have developed models

that use mutagenesis data to expose the ruggedness of the fitness landscape, where non-additive effects between mutations are more prevalent as the ruggedness increases [39,40•]. Using these models, it has been found that the optimal mutation rate decreases as the ruggedness increases [38••,41] and that the ruggedness affects the optimal ratio of parent genes for recombination [42].

Mutagenizing restricted regions of the gene can also optimize the search. A difficulty in doing this is deciding where the diversity should be targeted, as improvements can be propagated by subtle changes distant from the active site [43•,44,45]. One approach has been to highly mutagenize the active site. The majority of the mutants created by this technique have drastically reduced fitnesses, thus requiring extremely large selection capabilities to discover improvements [46,47]. Another successful approach has been to saturate residues for which directed evolution previously discovered improvements [48]. Overall, there is a need for algorithms that can aid in targeting regions of the protein structure that are likely to demonstrate improvement.

Several algorithms have been proposed that focus the diversity towards regions that are likely to preserve the structure, under the assumption that maintaining the stability of the structure is a prerequisite for discovering improvements. Towards this end, there is promise in merging computational and experimental combinatorial methods. Inverse folding algorithms can be used to predict the amino acid substitutions that are consistent with the parental structure. Mean-field techniques have been used to accelerate the calculation [49•,50•]. It is useful to condense the sequence information with an entropy measure to identify those residues that are the most likely to retain the structure. The residues that least perturb the structure can then be mutagenized and screened for improvements in the desired properties.

Conclusions

The challenging field of *de novo* enzyme design has begun to yield proteins with impressive catalytic efficiency. However, the current methods are not sufficient to design efficient enzymes for many reactions. Most design efforts to date have focused on using a single method, a drawback given the potential of combining complementary methods. Computational design is well suited to optimizing direct interactions including catalytic machinery, but is not well suited for identifying distant mutations that may subtly alter the protein structure to enhance catalysis. Directed evolution effectively optimizes activity, but is not as well suited to introducing completely novel activity. A few recent efforts have taken advantage of the complementary traits of the different methods. Rational design was used to engineer a low level of phosphoribosyl-anthranilate isomerase (PRAI) activity into indole-3 glycerol phosphate synthase, allowing selection in bacteria [7••]. Directed evolution was then used to generate a protein with highly efficient PRAI activity. Starting with an RNA-binding antibody, structurally guided mutagenesis of a single catalytic histidine yielded moderate ribonuclease activity [51]. The combined use of computational design with directed evolution and antibody techniques should continue to result in improved *de novo* biocatalyst designs.

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. Radzicka A, Wolfenden R: **A proficient enzyme.** *Science* 1995, **267**:90-93.

2. McCoy M: **Biocatalysis grows for drug synthesis.** *Chem Eng News* 1999, **77**:10-14.

3. Schmid A, Dordick JS, Hauer B, Kiener A, Wubbolts M, Witholt B: **Industrial biocatalysis today and tomorrow.** *Nature* 2001, **409**:258-268.

4. Arnold FH: **Combinatorial and computational challenges for biocatalyst design.** *Nature* 2001, **409**:253-257.

5. Cedrone F, Menez A, Quemeneur E: **Tailoring new enzyme functions**
•• **by rational redesign.** *Curr Opin Struct Biol* 2000, **10**:405-410.
An interesting review of site-directed mutagenesis with insight into the benefits of combining rational and combinatorial design techniques.

6. Petsko GA: **Enzyme evolution. Design by necessity.** *Nature* 2000, **403**:606-607.

7. Altamirano MM, Blackburn JM, Aguayo C, Fersht AR: **Directed**
•• **evolution of new catalytic activity using the alpha/beta-barrel scaffold.** *Nature* 2000, **403**:617-622.
Rational design was first used to introduce a low level of new activity into an existing enzyme. Directed evolution was then used to develop highly efficient activity.

8. Street AG, Mayo SL: **Computational protein design.** *Structure Fold Des* 1999, **7**:R105-R109.

9. Desjarlais JR, Clarke ND: **Computer search algorithms in protein modification and design.** *Curr Opin Struct Biol* 1998, **8**:471-475.

10. Gordon DB, Marshall SA, Mayo SL: **Energy functions for protein design.** *Curr Opin Struct Biol* 1999, **9**:509-513.

11. Shimaoka M, Shifman JM, Jing H, Takagi J, Mayo SL, Springer TA:
•   **Computational design of an integrin I domain stabilized in the open high affinity conformation.** *Nat Struct Biol* 2000, **7**:674-678.
Computational design was used to optimize the core of an integrin domain in two different conformations. The activity of the predicted proteins confirmed that one conformation is active and the other inactive.

12. Bolon DN, Mayo SL: **Enzyme-like proteins by computational**
•• **protein design.** *Proc Natl Acad Sci USA* 2001, **98**:14274-14279.
Computational protein design was used to predict active sites for activated ester hydrolysis in the inert protein scaffold of thioredoxin. Experimentally, these 'protozymes' demonstrated enzyme-like properties including catalytic rate enhancement and substrate saturation.

13. Benson DE, Wisz MS, Hellinga HW: **The development of new biotechnologies using metalloprotein design.** *Curr Opin Biotechnol* 1998, **9**:370-376.

14. Hellinga HW, Richards FM: **Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry.** *J Mol Biol* 1991, **222**:763-785.

15. Benson DE, Wisz MS, Hellinga HW: **Rational design of nascent**
•• **metalloenzymes.** *Proc Natl Acad Sci USA* 2000, **97**:6292-6297.
Computational design was used to identify a number of sites in thioredoxin that could bind iron. By modeling one of the primary coordination spheres of

the metal complexed with oxygen, proteins with a variety of experimentally determined chemistries were generated.

16. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS: **High-resolution protein design with backbone freedom.** *Science* 1998, **282**:1462-1467.

17. Nauli S, Kuhlman B, Baker D: **Computer-based redesign of a**
• **protein folding pathway.** *Nat Struct Biol* 2001, **8**:602-605.
Structural analysis indicated strain in a β-turn in protein G. Modeling of a new backbone conformation and optimization of the local sequence resulted in a protein that folded dramatically faster and had increased thermodynamic stability.

18. Su A, Mayo SL: **Coupling backbone flexibility and amino acid sequence selection in protein design.** *Protein Sci* 1997, **6**:1701-1707.

19. Desjarlais JR, Handel TM: **Side-chain and backbone flexibility in protein core design.** *J Mol Biol* 1999, **290**:305-318.

20. Quemeneur E, Moutiez M, Charbonnier JB, Menez A: **Engineering cyclophilin into a proline-specific endopeptidase.** *Nature* 1998, **391**:301-304.

21. Xiang H, Luo L, Taylor KL, Dunaway-Mariano D: **Interchange of**
• **catalytic activity within the 2-enoyl-coenzyme A hydratase/ isomerase superfamily based on a common active site template.** *Biochemistry* 1999, **38**:7638-7652.
Rational design of substrate-contacting residues based on structural alignment resulted in the engineering of efficient crotonase activity into 4-chlorobenzoyl-CoA dehalogenase.

22. Hilvert D: **Critical analysis of antibody catalysis.** *Annu Rev Biochem*
• 2000, **69**:751-793.
A thorough review of catalytic antibodies with analysis of future challenges in the field.

23. Stewart JD, Benkovic SJ: **Transition-state stabilization as a measure of the efficiency of antibody catalysis.** *Nature* 1995, **375**:388-391.

24. Wagner J, Lerner RA, Barbas CF III: **Efficient aldolase catalytic antibodies that use the enamine mechanism of natural enzymes.** *Science* 1995, **270**:1797-1800.

25. Petrounia IP, Arnold FH: **Designed evolution of enzymatic properties.** *Curr Opin Biotechnol* 2000, **11**:325-330.

26. May O, Nguyen PT, Arnold FH: **Inverting enantioselectivity by directed evolution of hydantoinase for improved production of L-methionine.** *Nat Biotechnol* 2000, **18**:317-320.

27. Schmidt-Dannert C, Umeno D, Arnold FH: **Molecular breeding of**
•• **carotenoid biosynthetic pathways.** *Nat Biotechnol* 2000, **18**:750-753.
By mutagenizing the enzymes participating in a secondary metabolic pathway, a library of cells was created that produced a diverse and colorful array of carotenoids.

28. Shao Z, Arnold FH: **Engineering new functions and altering existing functions.** *Curr Opin Struct Biol* 1996, **6**:513-518.

29. Zhao H, Arnold FH: **Combinatorial protein design: strategies for screening protein libraries.** *Curr Opin Struct Biol* 1997, **7**:480-485.

30. Wahler D, Reymond JL: **Novel methods for biocatalyst screening.** *Curr Opin Chem Biol* 2001, **5**:152-158.

31. Pedersen H, Holder S, Sutherlin DP, Schwitter U, King DS, Schultz PG: **A method for directed evolution and functional cloning of enzymes.** *Proc Natl Acad Sci USA* 1998, **95**:10523-10528.

32. Demartis S, Huber A, Viti F, Lozzi L, Giovannoni L, Neri P, Winter G, Neri D: **A strategy for the isolation of catalytic activities from repertoires of enzymes displayed on phage.** *J Mol Biol* 1999, **286**:617-633.

33. Jestin J-L, Kristenson P, Winter G: **A method for the selection of catalytic activity using phage display and proximity coupling.** *Angew Chem Int Ed Engl* 1999, **38**:1124-1127.

34. Atwell S, Wells JA: **Selection for improved subtiligases by phage display.** *Proc Natl Acad Sci USA* 1999, **96**:9497-9502.

35. Firestine SM, Salinas F, Nixon AE, Baker SJ, Benkovic SJ: **Using an AraC-based three-hybrid system to detect biocatalysts *in vivo*.** *Nat Biotechnol* 2000, **18**:544-547.

36. Olsen MJ, Stephens D, Griffiths D, Daugherty P, Georgiou G,
• Iverson BL: **Function-based isolation of novel enzymes from a large library.** *Nat Biotechnol* 2000, **18**:1071-1074.
A substrate cleavage event is linked to fluorescence so FACS can be used to screen for a turnover event.

37. Daugherty PS, Iverson BL, Georgiou G: **Flow cytometric screening of cell-based libraries.** *J Immunol Methods* 2000, **243**:211-227.

38. Voigt CA, Kauffman S, Wang ZG: **Rational evolutionary design: the**
•• **theory of *in vitro* protein evolution.** *Adv Protein Chem* 2000, **55**:79-160.
A review of techniques in theoretical biophysics, genetic algorithms, and population genetics as applied to the goal of accelerating *in vitro* evolution.

39. Aita T, Husimi Y: **Fitness spectrum among random mutants on Mt. Fuji-type fitness landscape.** *J Theor Biol* 1996, **182**:469-485.

40. Aita T, Uchiyama H, Inaoka T, Nakajima M, Kokubo T, Husimi Y:
• **Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin.** *Biopolymers* 2000, **54**:64-79.
A mathematical model of the fitness landscape is introduced that includes terms for additive and non-additive contributions from mutations. Data sets generated by mutagenizing two enzymes are fit to the model and used to give insight into the ruggedness of those landscapes.

41. Matsuura T, Yomo T, Trakulnaleamsai S, Ohashi Y, Yamamoto K, Urabe I: **Nonadditivity of mutational effects on the properties of catalase I and its application to efficient directed evolution.** *Protein Eng* 1998, **11**:789-795.

42. Aita T, Husimi Y: **Theory of evolutionary molecular engineering through simultaneous accumulation of advantageous mutations.** *J Theor Biol* 2000, **207**:543-556.

43. Spiller B, Gershenson A, Arnold FH, Stevens RC: **A structural view**
• **of evolutionary divergence.** *Proc Natl Acad Sci USA* 1999, **96**:12305-12310.
The long-range effects of mutations are demonstrated by comparing the structures of a wild type and evolved enzymes.

44. Yano T, Oue S, Kagamiyama H: **Directed evolution of an aspartate aminotransferase with new substrate specificities.** *Proc Natl Acad Sci USA* 1998, **95**:5511-5515.

45. Oue S, Okamoto A, Yano T, Kagamiyama H: **Redesigning the substrate specificity of an enzyme by cumulative effects of the mutations of non-active site residues.** *J Biol Chem* 1999, **274**:2344-2349.

46. Skandalis A, Loeb LA: **Enzymatic properties of rat DNA polymerase beta mutants obtained by randomized mutagenesis.** *Nucleic Acids Res* 2001, **29**:2418-2426.

47. Patel PH, Loeb LA: **Multiple amino acid substitutions allow DNA polymerases to synthesize RNA.** *J Biol Chem* 2000, **275**:40266-40272.

48. Miyazaki K, Arnold FH: **Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function.** *J Mol Evol* 1999, **49**:716-720.

49. Voigt CA, Mayo SL, Arnold FH, Wang ZG: **Computational method to**
• **reduce the search space for directed protein evolution.** *Proc Natl Acad Sci USA* 2001, **98**:3778-3783.
By applying mean-field theory to a structural model, the residues that least affect the structural integrity are determined. It is demonstrated that beneficial mutations discovered by directed evolution tend to occur at these residues.

50. Kono H, Saven JG: **Statistical theory for protein combinatorial**
• **libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure.** *J Mol Biol* 2001, **306**:607-628.
The energetic contributions between amino acids that lead to structural motifs are determined using mean-field theory.

51. Fletcher MC, Kuderova A, Cygler M, Lee JS: **Creation of a ribonuclease abzyme through site-directed mutagenesis.** *Nat Biotechnol* 1998, **16**:1065-1067.