

Finding focus	623
Screening sequences <i>in silico</i>	624
In chance we trust	626
Box 1: Easy screening	625

# Protein engineering: navigating between chance and reason

Monya Baker

Researchers use large libraries, focused libraries and rational design to engineer useful proteins.

Protein engineering has been used to create antibody drugs, very bright fluorescent proteins and enzymes that cut down on toxic waste. Genentech's drug for macular degeneration, Lucentis, is an antibody fragment created through site-specific mutagenesis and successive screening in phage to bind its target 100-fold more tightly than its precursor<sup>1</sup>. More recently, researchers at Codexis and Merck crafted an enzyme to replace an organo-metallic catalyst in the synthesis of sitagliptin, a diabetes drug with more than a billion dollars in annual sales. The researchers began with an enzyme that could create the appropriate bonds but had no activity for the desired substrate; about a dozen rounds of mutagenesis and modeling eventually produced an enzyme with high activity<sup>2</sup>. The enzyme had yields about 10% higher and produced 20% less waste compared with the original rhodium-based

catalyst. The companies won a Presidential Green Chemistry Award for the achievement, and Merck is now working with the US Food and Drug Administration to switch to a manufacturing route using the engineered enzyme.

Success in designing useful proteins is largely a matter of trying over and over again. But protein engineers often find themselves having to choose between trying to make bigger libraries and focusing libraries on sequences most likely to be effective. "Everything depends on how frequently successful solutions appear in the libraries. At the outset of any experiment, you don't really have an idea," explains Donald Hilvert, a protein engineer and laboratory head at the Swiss Federal Institute of Technology in Zurich (ETH Zurich).

Strategies for protein engineering are manifold. Many rely on display technolo-

gies (Table 1), which link protein to gene by putting proteins on the outer surface of viral particles or cells, or by physically linking mRNA to protein in cell-free systems. In a process called directed evolution, protein libraries are screened for desired activity; the best of the lot are collected and modified in hopes of further improvements, generating another library for another screen. Deciding how to find the best versions depends on the particular application and personal preference, but options include screening large random libraries, screening focused libraries as well as using rational design and computer tools.

## Finding focus

It is impossible to screen every possible protein sequence: even for proteins containing a mere 100 amino acids, the options are immense. Twenty possible amino acids at

**Table 1** | Display technologies

Technology (typical number of sequences screened per library)	Description	Strengths or weaknesses
Bacterial display (10 <sup>8</sup> –10 <sup>9</sup> )	Proteins are displayed on the surface or cell envelope of <i>Escherichia coli</i>	Selects proteins that can be made in cells Flow cytometry allows multiparameter, quantitative screening
mRNA display (10 <sup>15</sup> )	mRNA-protein fusions are synthesized through a puromycin linker; reverse-transcription PCR allows amplification after rounds of selections	Large libraries Can screen proteins that would be toxic to cells Works best with small proteins Stringent conditions required
Phage display (10 <sup>11</sup> )	DNA libraries encoding displayed proteins and required phage genes are put into bacteria, which produce the library attached to the phage surface	Robust and quick Smaller libraries than cell-free systems
Ribosome display (10 <sup>15</sup> )	DNA libraries encode the displayed proteins as a fusion to a sequence that tethers both mRNA and protein on stalled ribosomes; reverse-transcription PCR allows amplification after rounds of selections	Large libraries Can screen proteins that would be toxic to cells Requires stringent conditions and stable proteins
Yeast display (10 <sup>8</sup> –10 <sup>10</sup> )	Gene libraries code for the target protein fused to a yeast surface protein	Flow cytometry allows multiparameter, quantitative screening Selects proteins that can be made in eukaryotic cells

each of 100 positions works out to over  $10^{130}$  combinations<sup>3</sup>. To put that in perspective, the number of atoms in the observable universe is around  $10^{80}$ . Though many groups are working on ways to make screening more efficient (**Box 1**), the largest protein libraries screened have had around  $10^{15}$  members.

“Larger libraries and greater diversity are better, but you can’t cover everywhere,” says Gregory Weiss, who has adapted phage display to study membrane proteins at the University of California, Irvine. Generally fewer than 20 residues, if that, will be varied in an initial protein sequence of about 200 amino acids, he says. “That doesn’t give you a lot of room to work in. You have to choose your 10% very carefully.”

Several long-standing techniques can help researchers decide which sections of proteins to vary: alanine scanning creates proteins in which one position at a time is converted to alanine, one of the simplest amino acids. This technique can reveal which residues are particularly important for a protein’s function but does not neces-



Donald Hilvert at ETH Zurich says engineering enzymes to perform the same chemistry on new substrates is becoming straightforward. Redesigning enzymes for new reactions is more difficult.

sarily indicate how to improve it. Saturation mutagenesis samples all 20 amino acids at a particular residue, but creates too many variants to test many residues. Intermediate approaches can be used to interrogate more residues by sampling subsets of amino acids, such as whether they are hydrophobic versus hydrophilic or contain bulky versus tiny side chains.

Iterative saturation mutagenesis is a conceptual framework to systematically explore amino acids at several sites of interest (each site typically contains one to three amino acids). This approach was formalized by Manfred Reetz and colleagues at the Max Planck Institute for Coal Research, who have used it to improve an enzyme’s stability and also to select enzymes that favor one enantiomer over another<sup>4</sup>. The process requires multiple rounds of screening; hits from earlier libraries are used as starting points for subsequent rounds that explore other sites, hunting for ‘couples’ of amino



David Baker at the University of Washington, Seattle designs new enzymes and protein-protein interactions computationally, then tests them in the real world<sup>6</sup>.

acids with synergistic effects. A freely available program called the combinatorial active-site saturation test, more commonly known as CASTER, can help researchers plan which sites to explore and how to explore them efficiently.

Gene-synthesis companies can manufacture libraries by randomizing codons at particular locations, but this still makes for more possibilities than can be screened, explains Markus Enzelberger, vice president of research and development at MorphoSys. For antibodies, screening becomes considerably harder when researchers move past binding affinity to properties such as stability, aggregation and specificity. Even if researchers limit random variation to, say, all 20 amino acids at six particular residues, that would produce  $10^{10}$  DNA sequences, and because different sequences of nucleotides can code for the same amino acid, many of these are redundant. Cutting out redundant clones and stop codons reduces the six-residue library to  $10^7$ , but ordering  $10^7$  distinct sequences from a gene-synthesis company is not feasible. The library size could be reduced to a few dozen by varying one amino acid at a time, but then effects of combining variation at different positions would be lost, says Enzelberger. To explore the effects of combinations, his company recently acquired the company Sloning BioTechnology, which developed a technology that relies on a clever combination of double-stranded hairpin DNA and restriction enzymes to transfer desired codons into specific regions of a gene. “That makes the difference between being able to screen all the variants and not,” says Enzelberger, who says his company will make the proprietary technology available to other researchers under certain conditions.

Protein engineering strategies may shift depending on the project, says Herren Wu, vice president of research and development at MedImmune. “If you know what you want to do, and you have a lot of structure and function information, you can use a more

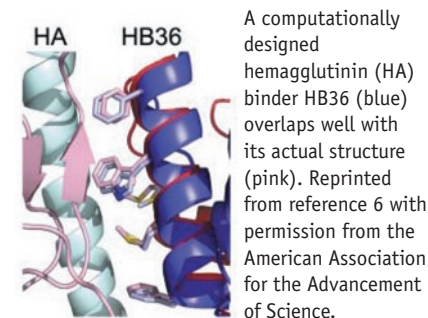
focused library or do rational design. If you know less, you really need a big library.” To find an antibody that binds to a particular target, researchers at his company sometimes begin with genes from individuals whose immune cells make antibodies to a pathogen. But because such starting points are often unavailable, most projects begin with phage display, a robust technique capable of displaying many sorts of antibodies. If an antibody with promising affinity is identified, discovery teams may then turn to ribosome display for optimization. Ribosome display can screen much larger libraries, he says, but is currently not as useful for *de novo* screening because not all antibodies can be displayed stably and efficiently.

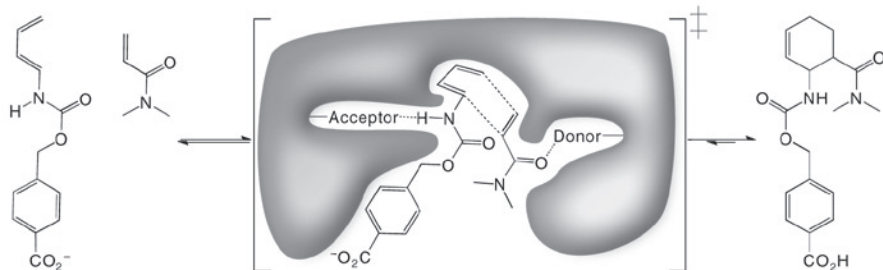
Different goals require different approaches, agrees Sachdev Sidhu at the University of Toronto, who is using phage display to design new protein-protein interactions. To find naive interactions without natural precedents, large simple libraries work better, but once an interaction is actually identified, the strategy shifts to optimizing a specific region of the protein. “Then you need a very precise interaction in an existing interface,” says Sidhu, so focused, structure-based and computational approaches can make an existing interaction stronger.

### Screening sequences *in silico*

Compared to wet-lab approaches, computational design can be used to screen many more variants more quickly. Designing an enzyme starts with making models of how an enzyme could facilitate catalysis of a given substrate. “There are a number of ways that a reaction can be catalyzed, and we come up with constellations of amino acid residues in ideal active sites,” explains David Baker, a computational biologist at the University of Washington, Seattle.

For each constellation, *in silico* screening initially identifies appropriate positions in protein scaffolds where the active site can be placed. After a few rounds of *in silico* opti-





To design an enzyme computationally, researchers model ideal active sites. Reprinted from reference 5 with permission from the American Association for the Advancement of Science.

mization, the most promising sequences are made into actual proteins and tested. Most of these do not work, but the approach has created functional enzymes for a handful of specific reactions. Last year, Baker's group described two enzymes with very different scaffolds that each catalyze the Diels-Alder reaction, creating two carbon bonds and four stereoisomers from appropriate substrates<sup>5</sup>.

Still, rationally designed enzymes usually are not very good compared to their natural counterparts. Nature has created enzymes that boost rates of reactions by up to  $10^{14}$  fold, says Baker, but the best computationally designed enzymes enhance reaction

rates by only  $10^6$ , even after a few rounds of directed evolution.

There are many possibilities why billions of years of evolution have outperformed the last few years of computational engineering. For one thing, whereas natural enzymes often have six or more amino acids that carry out an enzyme's chemistry, designed active sites usually start with three or four amino acids; additional residues are highly constrained by the protein scaffold. More importantly, the actual active site may not have the structure that was computationally predicted. The amino-acid changes necessary to give the enzyme its activity could

shift the protein's structure to produce an active site with a suboptimal conformation.

"Computation allows you to explore sequence spaces that are not accessible by experiment," says Hilvert. But before asking whether a new configuration works as predicted, researchers have to answer another question, he says. "How do I fit all of that into a scaffold that hasn't seen this kind of apparatus before?"

A broader choice of scaffolds could boost the quality of computationally designed proteins, says Baker. In fact, some options have been inspired by online volunteers playing the computer game Foldit, which Baker and coworkers developed to see whether game players could solve problems in predicting protein structures.

Another application of computational design is in creating protein-protein interactions. Baker and colleagues recently reported a small protein that binds a small, well-conserved region on influenza hemagglutinin, the protein which the virus uses to gain access to host cells<sup>6</sup>. The predicted structure of the designed protein matched the crystallographically observed structure almost perfectly,

## BOX 1 EASY SCREENING

The number of stars in the observable universe has been estimated at upward of  $10^{22}$ . Even that number is tiny compared to that of potential protein sequences. With so many possibilities, protein engineers are hunting for more efficient ways to explore sequence space. Whereas cell-free systems can be used to churn through greater numbers of proteins (including proteins toxic to yeast and bacteria), cell-based systems offer a ready means of production, particularly for large, complex proteins. Researchers including George Georgiou at the University of Texas, Austin have pioneered techniques in which cells producing proteins with desirable properties can be easily sorted by flow cytometry using fluorescence-activated cell sorting.

Still, getting display technologies to select for properties beside binding is often impractical. This year, researchers led by David Liu at Harvard University described ways to use phage or yeast to screen for a wide range of catalytic and binding functions. Instead of using phage to display proteins for potential binders, phage-assisted continuing evolution depends on proteins' function inside an *Escherichia coli* cell<sup>7</sup>. Phage are genetically altered so that they cannot reproduce except in the presence of a protein with desired functions; a mutagenesis plasmid in *E. coli* generates variation around the relevant genes and allows dozens of rounds of mutagenesis and selection to occur in a single day without manual intervention.

In separate work, researchers in Liu's lab adapted yeast display to select for enzymes that join two molecules<sup>8</sup>. In addition to displaying the enzyme of interest on the cell surface, yeast also display another peptide, which serves as a hook to which an enzyme's substrate can be attached. Next, the second substrate is added followed by a fluorescent label, which allows cells encoding enzymes that join the substrates to be purified. In an initial demonstration of the system, they selected for yeast encoding enzymes with substrate-joining activity and used flow cytometry to enrich this population by a factor of 6,000. Eight rounds of screening produced an enzyme with a 140-fold increase in activity over its predecessor. Although this work involved the enzyme sortase A, says Liu, the technique should work for many proteins that catalyze bond formation.

Other work has sped up the pace of screening itself. Researchers led by Andrew Griffiths at the University of Strasbourg and David Weitz at Harvard University combined yeast display with microfluidics, creating picoliter-sized drops, most containing only a single yeast cell<sup>9</sup>. In a proof-of-principle study, the researchers used a total reagent volume of less than 150 microliters to screen  $\sim 10^8$  variations of horseradish peroxidase. The enzyme generates a fluorescent product, and the brightest drops can be sorted easily. Griffiths and Weitz estimate that the technique increases speed of screening by 1,000-fold and decreases cost a million fold.

It is too early to know whether these techniques will be robust enough to work in many labs, but if these and similar technologies spread beyond the labs in which they were invented, the size of libraries screened could expand astronomically. Protein engineering would, in a sense, be reaching for the stars.

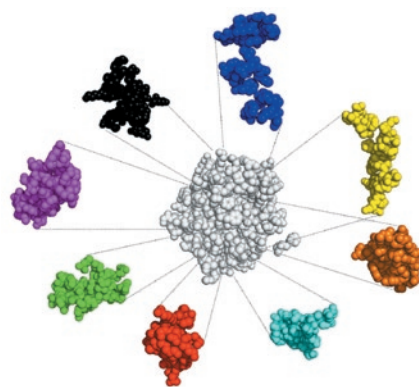
even down to the side chains on individual amino acids. Nonetheless, says Baker, researchers have much to learn before they can reliably design proteins that act as expected.

### In chance we trust

Right now, protein engineers might do better to trust in chance than design, says Frances Arnold at the California Institute of Technology, whose directed evolution techniques using random mutagenesis are used in laboratories around the world. Researchers who design libraries for beneficial mutations rarely compare libraries to randomly generated ones, she says. When researchers in her lab analyze beneficial mutations in enzymes, results show a surprising lack of patterns. “You get mutations all over the enzyme. You couldn’t explain them, let alone predict them,” she says. “What we’ve learned is that we know less than we thought we did. People learned



“There are lots of ways to evolve a protein,” says Frances Arnold at California Institute of Technology. “If one doesn’t work, you can try another.”



M. Smith, Arnold lab, California Institute of Technology

Chimeric protein libraries, made by combining parts of related proteins, create diverse libraries with a high proportion of functioning proteins.

the hard way that you couldn’t just go in and make a mutation where you thought and expect it to improve an enzyme.”

Error-prone PCR makes on average one variation per gene, so Arnold increases variation by recombining genes encoding pieces of homologous proteins from different species. This produces highly varied libraries, whose members may differ from each other by sometimes hundreds of residues. At the same time, the number of sequences that could code for a stable protein is high, explains Arnold. “Recombination is a great way to make

many mutations simultaneously. Because these have already been vetted by nature, you know that they can make a folded protein.”

Given the number of sequences to explore, even a rough guide to productive sequence space can make a big difference. “If your choice of building blocks is a sensible one, then your chance of getting a hit is higher,” says Hilvert. Nonetheless, he cautions that finding more hits does not guarantee finding the best possible hit. “You do make a sacrifice because you’re not looking at everything.”

1. Chen, Y. *et al. J. Mol. Biol.* **293**, 865–881 (1999).
2. Savile, C.K. *et al. Science* **329**, 305–309 (2010).
3. Jäckel, C. & Hilvert, D. *Curr Opin Biotechnol.* **21**, 753–759 (2010).
4. Reetz, M.T. *Angew. Chem. Int. Edn. Engl.* **50**, 138–174 (2011).
5. Siegel, J.B. *et al. Science* **329**, 309–313 (2010).
6. Fleishman, S.J. *et al. Science* **332**, 816–821 (2011).
7. Esvelt, K.M., Carlson, J.C. & Liu, D.R. *Nature* **472**, 499–503 (2011).
8. Chen, I., Dorr, B. & Liu, D.R. *Proc. Natl. Acad. Sci. USA* published online 22 June 2011 (doi:10.1073/pnas.1101046108).
9. Agresti, J.J. *et al. Proc. Natl. Acad. Sci. USA* **107**, 404–409 (2010).

Monya Baker is technology editor for *Nature* and *Nature Methods* (m.baker@us.nature.com).