tissues or organs, and many tumors.

Nevertheless, bioluminescent quantum dot technology has the potential to greatly improve near-infrared fluorescence detection in living tissue. And it stimulates one to think about how the interconversion of energy from one form to another might solve major problems in the field of *in vivo* imaging.

1. So, M.-K., Xu, C., Loening, A.M., Gambhir, S.S. & Rao, J. *Nat. Biotechnol.* **24**, 339–343 (2006).
2. Xu, Y., Piston, D.W. & Johnson, C.H. *Proc. Natl. Acad. Sci. USA* **96**, 151–156 (1999).
3. Sevick-Muraca, E.M., Houston, J.P. & Gurfinkel, M. *Curr. Opin. Chem. Biol.* **6**, 642–650 (2002).
4. Ntziachristos, V., Bremer, C. & Weissleder, R. *Eur. Radiol.* **13**, 195–208 (2003).
5. Frangioni, J.V. *Curr. Opin. Chem. Biol.* **7**, 626–634 (2003).
6. De Grand, A.M. *et al. J. Biomed. Optics* in press (2006).
7. Michalet, X. *et al. Science* **307**, 538–544 (2005).
8. Lim, Y.T. *et al. Mol. Imaging* **2**, 50–64 (2003).

Fancy footwork in the sequence space shuffle

Frances H Arnold

Recent reports on directed evolution broaden the scope of evolutionary enzyme engineering.

If we were able to explore all possible proteins, we would likely find fantastical molecules that might solve any number of human problems. The riches of sequence space include cures for cancer and solutions to the energy crisis, along with countless other valuable molecules. Unfortunately, such a comprehensive exploration is not even remotely possible, as sequence space is Vastly—for "Very-much-more-than-astronomically"—large[1]. Just a single copy of each 300-amino acid sequence, for example, would fill dozens of universes. That nature has explored only the tiniest fraction of this space over the entire history of life on Earth should leave protein engineers excited about their own opportunities for discovery.

Yet excitement over the untold riches of sequence space must be tempered by the recognition that the great majority of those sequences don't code for anything interesting; most don't even fold. Estimates for the density of functional proteins in sequence space range anywhere from 1 in 10^{12} to 1 in 10^{77}. No matter how you slice it, proteins are rare. Useful ones are even more rare. This might lead one to believe that discovering new proteins by mutation and selection is highly unlikely and to discount evolution as an algorithm for discovery. So how do laboratory evolutionists discover new proteins on the timescale of a PhD thesis or, worse, a commercial deadline? They do it by taking the right kinds of steps and starting from the right places.

Three recent reports describe new twists on this theme. Writing in the *Journal of the American Chemical Society*, Qian and Lutz[2] show how circular permutation might complement other, more tried-and-true steps in the search for better enzymes. In a report in *Nature Genetics*, Peisajovich *et al.*[3] provide a laboratory demonstration of how this permutation step might happen in nature. Finally, writing in *Science*, Park *et al.*[4] demonstrate how some 'rational design,' with inspiration from studying evolutionarily related proteins, can help find a good place to do the sequence space shuffle.

Evolution works because functional proteins are not evenly distributed in sequence space. Functional proteins are surrounded by other functional proteins that share the same overall structure. Even though most random amino acid substitutions are deleterious, many are not. Sometimes, a single substitution can improve a protein; accumulating such beneficial mutations over iterative rounds of mutagenesis and selection is an effective evolutionary strategy. Random mutation is only one search mechanism that explores sequence space efficiently. Recombination also accesses functional proteins with high probability and can make much larger jumps in sequence space than random mutation[5]. Laboratory evolutionists have also used less-natural search operations: saturation mutagenesis and random mutagenesis targeted to key portions of a protein (for example, the active site) are widely believed to provide advantages over more random approaches, especially when detailed structural information is available.

Qian and Lutz demonstrate that circularization and random opening should be included on a list of preferred search steps by making permutations on a lipase from *Candida antarctica* that is widely used in chemical synthesis. Postulating that the enzyme might hydrolyze bulkier substrates more efficiently if it had greater flexibility in its active site, these authors set out to determine whether opening up new C- and N-termini might provide that flexibility, especially if the new ends appeared near the active site. Working at the gene level, they connected the termini with a flexible linker, circularized the construct and proceeded to make random cuts. Screening for the genes that produced lipase activity in *Pichia pastoris* yielded functional permutants. Not only did they identify 63 new ways to start and stop the *C. antarctica* lipase, they also found some variants with significantly higher (10–60 fold) k_{cat} values on lipase substrates (*p*-nitrophenol butyrate and 6,8-difluoro-4-methylumbelliferyl octanoate).

The net result of permutation is (presumably) a protein of the same overall structure, and with most of the amino acids in the same places in the structure. However, the sequences and topologies might be completely different if the polypeptide chain starts and ends at very different positions in the structure (**Fig. 1**). What role this strategy plays in the evolution of new functional proteins still remains to be determined, but it could wreak havoc for patent attorneys!

Various natural protein families bear the marks of having undergone permutation, leading to rearrangement of functional modules and diversification of their topologies. The circularization strategy used by Qian and Lutz to obtain their permutants is not likely to happen *in vivo*. However, Peisajovich *et al.* replicated what is considered the most likely natural equivalent of circular permutation—gene duplication and in-frame fusion followed by degradation from the 5′ and 3′ ends to generate new N- and C-termini. They tested this set of steps on a gene for a DNA methyltransferase (M.*Hae*III) and demonstrated that not only could this mechanism produce active permuted methyltransferases, it could do so through a series of functional intermediates in which only the N- or the C-terminus was degraded. These intermediates, which contained some wholly or partially duplicated modules, folded and functioned, albeit at a reduced level compared with the unmodified protein. Because all of the steps used for this laboratory demonstration have natural counterparts, it is likely that similar events can and

Frances H. Arnold is in the Division of Chemistry and Chemical Engineering, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA.
e-mail: frances@cheme.caltech.edu

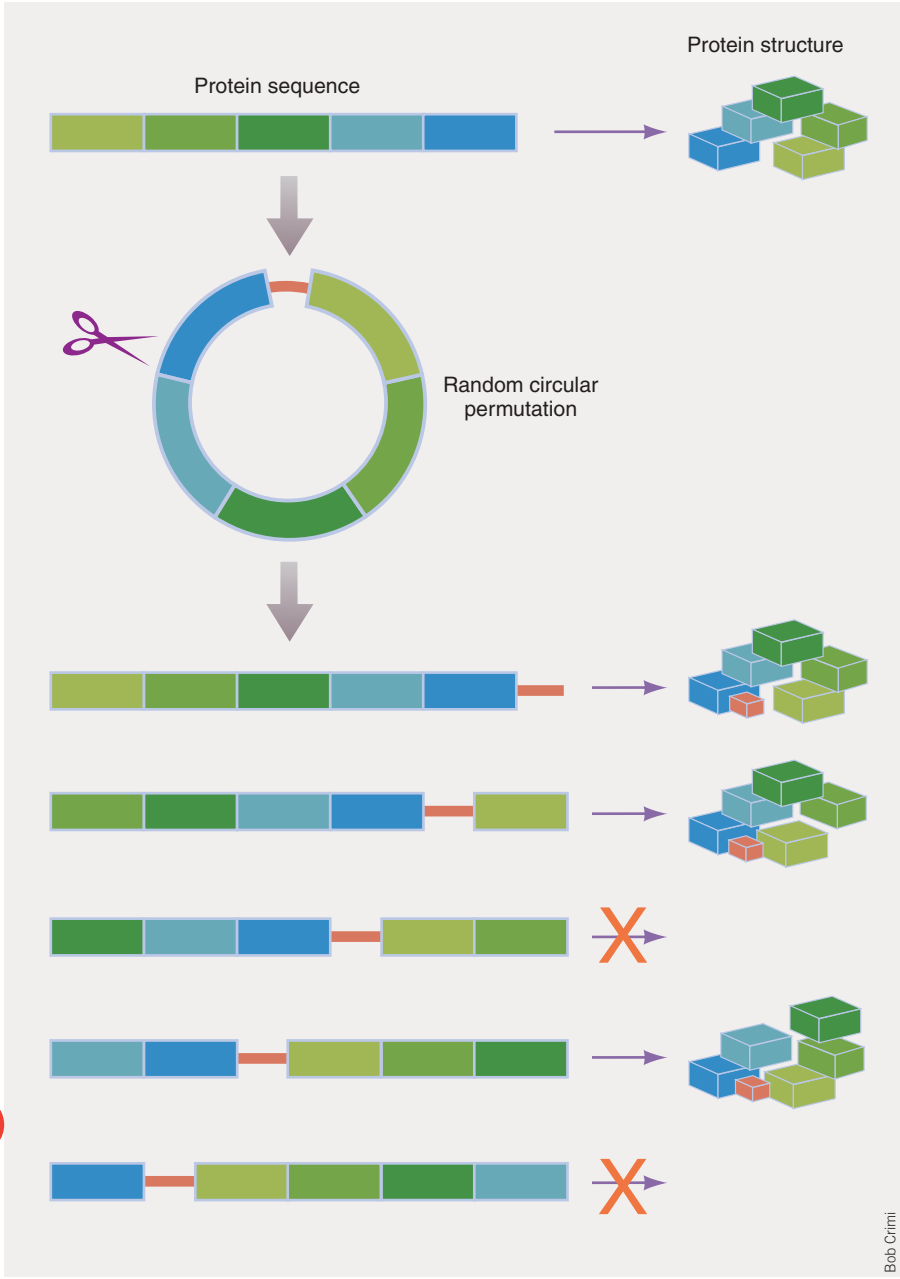

Figure 1 Random circular permutation provides access to some of the riches of sequence space. Working at the gene level, a short peptide linker (orange) is added to connect the native termini and the DNA construct is cleaved randomly[2]. Changing the termini of a protein sequence results in the ordered reorganization of its amino acid building blocks (the primary structure). Although some permutated proteins will no longer be able to fold properly (marked X), others retain the protein's global structure. Termini relocation may result in changes in conformation and flexibility, particularly near the site of peptide cleavage, and may affect enzyme activity. Simply aligning the permuted sequences might mask the fact that circular permutation, while changing the primary structure, does not change the identity or position of the amino acid residues in the three-dimensional structure. Figure provided by the Stefan Lutz laboratory.

do occur in nature. In fact, the existence of a new class of methyltransferases predicted on the basis of the laboratory results was validated through searching sequence databases.

Whereas Qian and Lutz found 63 unique functional solutions to permuting the lipase, Peisajovich *et al.* found far fewer functional permutants of M.*Hae*III. This is despite the fact that methyltransferases have clearly undergone such gene rearrangements in the past, and there is no evidence that lipases have. Clearly, guidelines for which proteins are likely to accept such an operation, and especially which ones are likely to benefit from it by developing new or improved function[6,7], must still be determined.

Understanding how functional proteins are distributed in sequence space is fundamental to the success of directed evolution. Another key factor, often ignored, however, is the starting point. Yes, the protein scraped from the bottom of your shoe or collected from your refrigerator is one of those rare sequences that encodes a functional protein. But it is not necessarily a good starting point for obtaining the protein of your dreams. Natural evolution can take twists and turns that the graduate student or industrial biotechnologist does not have the luxury of taking. Thus it makes sense to use all the shortcuts you can to breed new molecules. A champion racehorse is more likely to be born of fleet parents, or at least ones with the requisite physiognomy. A new functional protein is likewise more likely to appear when the laboratory evolutionist makes a discriminating choice of parent(s), thereby starting his search in a promising ballpark.

Recent indications are that good starting points might be accessible by rational design, using powerful computational approaches[8] or inspiration derived from a related protein, particularly if that relative already exhibits the targeted function[4]. Mixing rational design with a little randomization, Park *et al.*[4] converted glyoxalase II, which hydrolyzes the thioester bond of *S*-D-lactoylglutathione, into a metallo-β-lactamase, which catalyzes a similar hydrolysis reaction, but on cefotaxime, a very different substrate. The two naturally occurring enzymes that hydrolyze these substrates share the same overall fold and ancestry, but exhibit low sequence identity. Rationally introduced changes to glyoxalase II included altering the metal binding site to accommodate Zn and duplicate the coordination pattern observed in the β-lactamase. In addition, the C-terminal glutathione-binding domain was removed and new loop regions based on metallo-β-lactamase family templates were grafted on, each containing variable amino acids.

Bacteria transformed with a diverse gene soup of all these changes and seasoned with a sprinkling of random mutations were selected using cefotaxime. Positives were further evolved with multiple rounds of DNA shuffling and selection until the evolved enzyme conferred resistance to 4 μg/ml of the antibiotic. The resulting protein displays only 59% identity to glyoxalase II and contains mutations throughout. Although competent enough to confer antibiotic resistance at a low level, the evolved enzyme is significantly less active on cefotaxime than its role model. And, unlike

natural β-lactamases, it does not hydrolyze the other β-lactam antibiotics tested.

One way to view the work of Park *et al.* is as a beautiful demonstration of how evolution can rescue a less-than-perfect (but good-as-you-can-get) rational design. The other way to look at it is that rational design narrowed the vast sequence space down to the infinitesimally small (compared to sequence space) ballpark actually searchable in a real experiment. What remains to be seen, however, is whether the ballparks targeted by human design[4,8] also contain enzymes as good as the ones that nature makes, in addition to the relatively mediocre versions discovered so far. In other words, are all mediocre enzymes surrounded by good ones? Answering this will require more exploration.

1. Dennett, D.C. *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (Simon & Schuster Inc., New York, NY; 1995) p. 109.
2. Qian, Z. & Lutz, S. *J. Am. Chem. Soc.* **127**, 13466–13467 (2005).
3. Peisajovich, S.G., Rockah, L. & Tawfik, D.S. *Nat. Genet.* **38**, 168–174 (2006).
4. Park, H.-S. *et al. Science* **311**, 535–538 (2006).
5. Drummond, D.A. *et al. Proc. Natl. Acad. Sci. USA* **102**, 5380–5385 (2005).
6. Baird, G.S., Zacharias, D.A. & Tsien, R.Y. *Proc. Natl. Acad. Sci. USA* **96**, 11241–11246 (1999).
7. Guntas, G., Mansell, T.J., Kim, J.R. & Ostermeier, M. *Proc. Natl. Acad. Sci. USA* **102**, 11224–11229 (2005).
8. Dwyer, M.A., Looger, L.L. & Hellinga, H.W. *Science* **304**, 1967–1971 (2004).

Know the enemy

Ee Chee Ren

The sequencing of 169 avian influenza virus genomes provides a much-needed boost to preparations for a pandemic.

As health authorities across the world gear up efforts to ward off the spread of highly pathogenic avian influenza virus (AIV)[1], they are hampered by a critical lack of basic knowledge about the enemy. In a recent issue of *Science*, Obenauer *et al.*[2] provide vital new information on AIV in the form of 169 complete genome sequences and 2,196 partial sequences. These genetic data will facilitate understanding of the specific components of AIV that cause damage to the human body and, more importantly, provide the basis for developing effective vaccines and therapies.

The influenza virus has eight segments of negative-strand RNA (PB1, PB2, PA, HA, NP, NA, M and NS) totaling about 13.6 kb. There are 16 hemagglutinin (H1–H16) and 9 neuraminidase (N1–N9) serologically defined variants, which provide the basis for virus typing. As Obenauer *et al.* note, the existing public sequence data are understandably skewed towards these genes, and the relative scarcity of sequence information about the rest of the genome has turned out to be grossly insufficient for scientists to understand how the virus as a whole undergoes genetic drift and shift.

In their study, Obenauer *et al.* tapped the rich repository of ~7,000 AIV samples available at the St. Jude Children's Research Hospital in Memphis, Tennessee, selecting 336 random samples drawn from ducks, gulls, shorebirds and poultry. One of the first things they noticed was the existence of previously undetected subgroups of AIV. Although this is not surprising considering the wide range of samples tested, it provides a more complete picture of what is 'out there.'

Next, the group attempted to combine their new data with all other known AIV sequences. This analysis, although not computationally complex, was hard to visualize. As a result of the many sporadic mutations that have accumulated in the viral RNA genome over time, it was difficult to discern the principle mechanisms of AIV evolution. To remove the genetic noise and clutter, the group applied two main rules: (i) visualize the different genome fragments at the amino acid level, and (ii) describe the virus segments according to the most frequently occurring residues. By stripping away the consensus 'nonformative' regions and joining what remained in a contiguous stretch, they generated a unique protein sequence signature, termed a 'proteotype.' This treatment was applied to all eight gene fragments, and within each gene, each unique proteotype was assigned a clade and proteotype number (**Fig. 1**).

The power and utility of this approach becomes apparent when all 169 complete AIV genomes are displayed according to their proteotypes and one can immediately appreciate the combinatorial shuffling of gene segments that results in new AIV compositions. The proteotype patterns showed that the HA and NA genes, which define AIV's surface characteristics, have undergone reassortment much more frequently than have the six core genes (PB1, PB2, PA, NP, M and NS). This finding suggests that AIV is constantly testing different ways of cloaking itself to inject a relatively preserved set of replication machinery into host cells. In addition, the authors saw cosegregation among the six core genes, which implies that certain genes work better with preferred partners and/or that compensatory mutations confer a positive selection and fitness advantage.

Obenauer *et al.*'s collection of AIV sequence data also revealed that a short stretch of four amino acid residues at the C terminus of the NS1 protein forms a binding site that interacts with PDZ domains. As host cellular proteins containing PDZ domains are often involved in signaling pathways, this finding suggests that NS1 may interfere with key cellular processes such as membrane protein trafficking, cell morphology and neuronal signaling. Using NMR and protein arrays, the authors were able to demonstrate direct binding of avian influenza NS1 to 30 out of 123 tested human PDZ domains.

Intriguingly, human samples from the highly pathogenic infections of 1997 and 2003 carried the avian NS1 'ESEV' motif, in contrast to the low-pathogenicity infections of 1957 and 1968, which carried the human NS1 'RSEV' motif. It is too early to know whether this single change between an acidic residue (E, glutamate) and a basic residue (R, arginine) accounts for these extremes of pathogenicity. Nevertheless, the highly pathogenic 1918 flu strain[3], which was largely avian in composition, carried the NS1 'KSEV' motif. Although these data hint at the NS1 protein having a role as a virulence-determinant factor, its cellular targets must be defined before its true significance can be understood. It is most likely, however, that the genetic correlates of avian influenza pathogenicity lie in the sum of the viral gene sets and are further modulated by interactions with the host.

As described in companion studies published in *Nature*[4] and *PLoS Biology*[5], large-scale sequencing of human influenza A viruses has so far yielded 209 complete human influenza genomes consisting predominantly of H3N2 isolates. These two large data sets of influenza genomes should provide an unprecedented resource for public access. What's lacking,

Ee Chee Ren is at the Genome Institute Singapore, 60 Biopolis St., Singapore 138672.
e-mail: renec@gis.a-star.edu.sg