

Survival of the Fittest Molecule

Biochemists harness a novel form of evolution to sculpt new compounds for the fight against dengue fever, cancer and other modern plagues

Willem Stemmer and Brett Holland

Over billions of years, life has evolved into a spectacular diversity of forms—more than a million species presently exist. For each, the source of its uniqueness is the particular constellation of proteins found within its cells. Yet in the midst of this diversity, the similarities between living things are profound. For example, although the fruit fly genome encodes about 14,000 different proteins, and humans have two to three times that number, many proteins are still recognizably similar in sequence and task, reflecting their common ancestry. In fact, when scientists have put human disease genes into flies, they often cause the same symptoms in the insects as they do in people. Furthermore, addition of a normal human gene can sometimes compensate for the deletion of the same gene from the fly.

The differences that do exist between equivalent genes in flies and humans are the result of 900 million or so years of DNA mutations—the amount of time that has passed since the divergence of arthropods (including fruit flies) and chordates (including human beings). Some of these mutations also changed the sequence of amino acids

in the proteins encoded by these genes. Out of these modified proteins, a small fraction worked in a superior or novel way and gave some advantage to their bearers; ultimately, organisms with the mutation left more descendants than those without it. In other words, mutation provides the raw material for natural selection.

Evolution also works over shorter timescales. Anatomically modern humans are probably less than 170,000 years old (see “We Are All Africans,” page 496), yet the genetic diversity among living *Homo sapiens* runs to several million mutations, also known as *polymorphisms*, which form the basis for a vast number of combinations, each with a potentially different set of properties. Because preserving a low mutation rate is important for complex organisms, the principal source of functional genetic diversity is recombination between sister chromosomes, a process that creates new combinations from existing point mutations. These rearrangements are not necessarily beneficial—many are neutral or harmful—but the process is endlessly iterative, creating in each generation novel combinations of mostly old mutations. Indeed, recombination followed by natural selection is the foremost mechanism of organic evolution.

Now biomolecular engineers seeking to craft proteins that will perform certain functions have co-opted this evolutionary strategy. Pharmacologists and industrial biochemists are patently interested in tailoring proteins to individual needs, and developing the tools to create these molecular thoroughbreds is crucial. In this article, we discuss the importance of custom-built proteins and describe some of our own breakthroughs that have dramatically

increased the speed and efficiency of their creation.

Biotech Promise

Proteins make and maintain life, so they play an understandably central role in biomedical science. Many human maladies come from disease genes that cause specific protein changes, including hemophilia, muscular dystrophy and sickle cell anemia. While some of these altered-protein diseases remain difficult to correct, others can be treated. In the case of classic hemophilia, replacing the blood-clotting factor VIII protein that is missing or reduced in people with the disease can lessen the chronic bleeding problems. The original source for this purified protein was plasma from human blood, but in 1984 the normal version of the factor VIII gene was inserted into a microbial host. These transformed bacteria were designed to produce massive quantities of the human factor VIII protein, and in 1990 a large clinical trial demonstrated the success of the new drug. The recombinant protein was safer as well as more efficient to produce: Bioengineered factor VIII freed patients from the risks associated with blood borne diseases such as HIV and hepatitis.

For all its promise, the biotechnology revolution is still relatively young. The first therapeutic use of genetically engineered human proteins was only 25 years ago, after recombinant insulin was produced in bacteria. Such therapies have exploded since then, aided by the increasing sophistication of tools for manipulating DNA. The main advance was the brilliantly simple use of a combination of two well-studied classes of bacterial enzymes: restriction endonucleases and ligases. These enzymes cut

Willem “Pim” Stemmer invented DNA breeding and co-founded Maxygen Incorporated in 1997 to develop the technology. Currently, Maxygen and its subsidiaries, Codexis and Verdia, are focused on applying this technique to pharmaceuticals, vaccines, chemicals and agriculture, and Pim has co-founded a new company called Avidia Research Institute to focus on antibody-like molecules. Brett Holland is an assistant professor of biology at California State University in Sacramento, where he teaches evolution and genetics. His research interests include sexual selection and sexually antagonistic coevolution. Dr. Stemmer’s address is Avidia, 2450 Bayshore Drive, Mountain View, CA 94043. Internet: pim.stemmer@avidia.org

© 2003 Sigma Xi, The Scientific Research Society. Reproduction with permission only. Contact perms@amsoci.org.

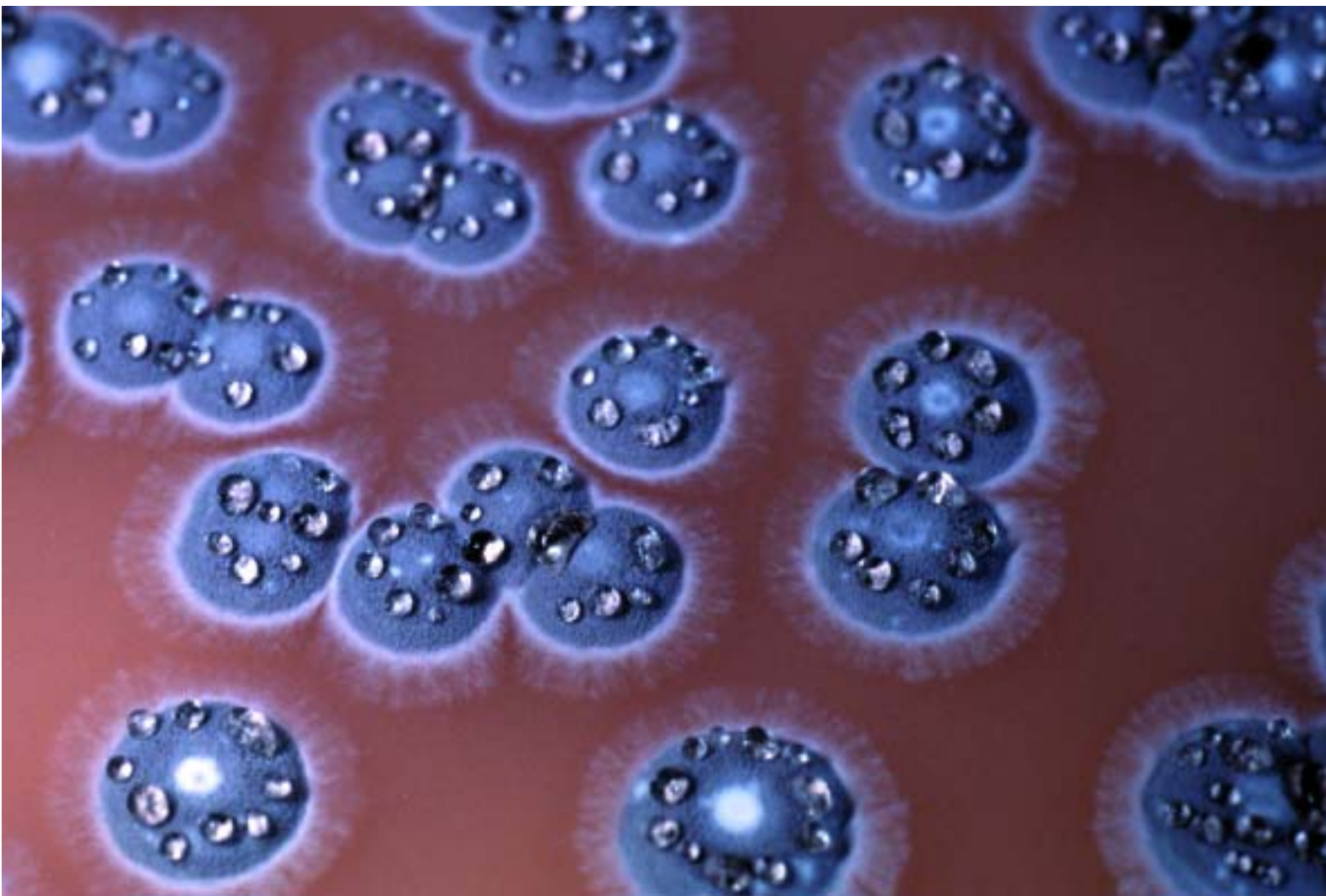


Figure 1. A strain of the bacteria *Streptomyces ambofaciens* produces the antibiotic polyketide spiramycin, seen as beads of clear liquid on the surface of each blue colony. In this article, the authors discuss their development of strategies to optimize the function of proteins and entire bacterial genomes by mimicking *in vitro* the natural process of DNA recombination—an approach that they term “DNA breeding.” (Photograph courtesy of Tobias Kieser, John Innes Centre, Norwich, U.K.)

and rejoin, respectively, pieces of DNA in a sequence-specific, origin-independent manner. This allowed molecular biologists to create hybrid DNA molecules from different species, meaning that genes could be copied, or cloned, endlessly by putting them into fast-dividing bacteria. Likewise, the proteins encoded by DNA could be made in quantities that were impossible to get from the original sources.

Modern molecular technologies now provide unprecedented access to the enormous diversity of natural proteins. However, these proteins tend to be very dependent on the cellular context in which they evolved. When they are removed from the original cellular milieu and produced and assayed in a different environment, the proteins are apt to perform poorly or not at all.

Improving Nature’s Design

In light of this inherent limitation, an important goal of biotechnology is to re-optimize molecules to preserve or enhance their function. For example, some industrial enzymes need to be modified for expression at extremely high levels, yet retain their function under harsh physical conditions. In the pharmaceutical industry some human proteins may require a different receptor-binding property or circulatory half-life to survive therapeutic administration. Another example comes from our own lab, where we are optimizing a viral vaccine to induce a strong and broadly protective immune response. This is a challenge because viruses have evolved specifically to evade the immune systems of their hosts.

At present two different but complementary strategies are being pursued

for the optimization and redesign of proteins. They are generally known as *rational design* and *directed evolution*.

Rational design, also known as computer modeling, attempts to modify or create molecules for specific applications by predicting which amino acid sequence will produce a protein with the desired properties. Using a combination of x-ray crystallography, three-dimensional structure determination and computer-simulated amino acid interactions, biochemists can often predict which amino acid substitutions are the best candidates to elicit the desired change in a protein.

Unfortunately, the task of accurately modeling protein function is Herculean—there are a staggering number of interdependent variables that influence protein function. Cells go through

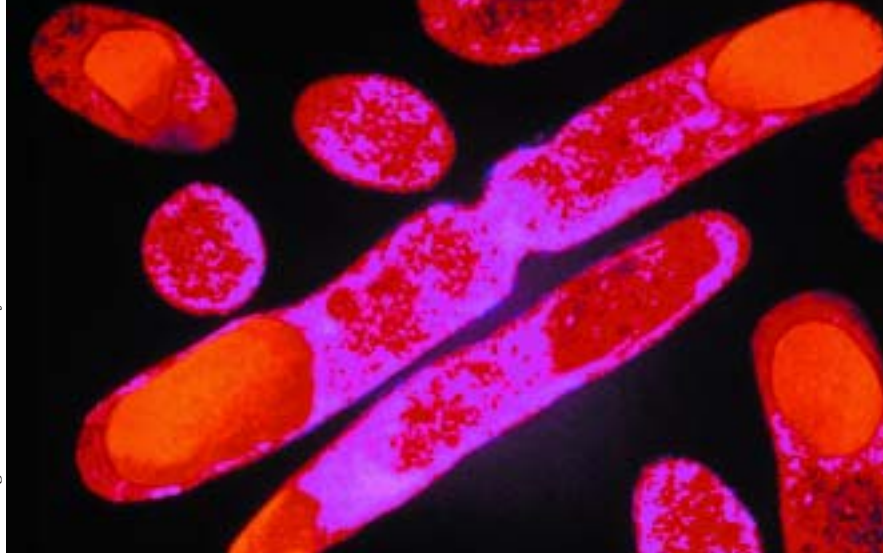


Figure 2. Insulin was the first recombinant protein produced by genetically engineered bacteria for therapeutic use in human beings. This pseudocolored transmission electron micrograph shows *Escherichia coli* bacteria carrying the human insulin gene; sites of protein expression are orange.

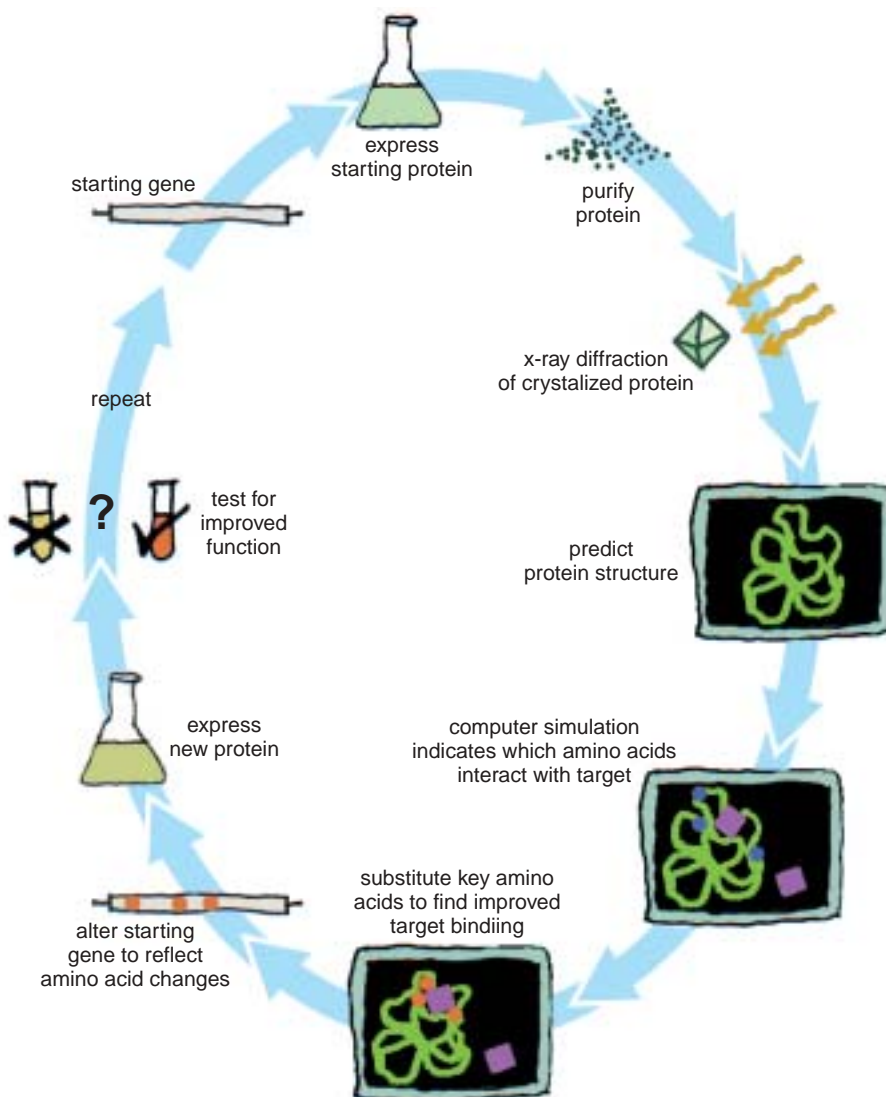


Figure 3. Rational protein design is one strategy for tailoring proteins to specific needs. To optimize a known protein, the starting gene product is purified and crystallized, and the three-dimensional structure of the protein is deduced using x-ray diffraction. A computer model predicts the forces between amino acids and identifies key sites that interact with a target molecule. Based on an understanding of the physical principles that govern these interactions, specific amino acid changes can be tested that may improve protein function.

many steps between DNA and an active protein—including but not limited to RNA and peptide synthesis, post-translational modification, subcellular targeting and intermolecular binding—and each step is regulated by multiple mechanisms. Protein folding and stability alone are sensitive to dozens, if not hundreds, of internal and external factors, and the consideration of any additional properties, such as activity in the presence of organic solvents, complicates matters further. Even without calculating such interactions between molecules, modeling the forces exerted by amino acids within the protein is an enormous undertaking—an average-size polypeptide has 300 amino acids, and each interaction is influenced by changes in the solution composition.

With all of these variables, computer models are understandably constrained by their requirement for massive computational power. The steady rise of microprocessor speed suggests that this approach may be increasingly fruitful in the future, but it will remain limited by the detail with which we can describe a protein's interactions with thousands of other molecules in the cell.

Directed evolution is an alternative to rational design that does not rely so heavily on future technology. Over the past decade, a much smaller number of scientists, including us, have followed this alternate approach with increasing success. Rather than trying to modify existing proteins or design new proteins by computer simulation of physical principles, we harness natural selection at the molecular level and direct the evolution of proteins that are customized to meet specifications set by medicine, agriculture and industry.

Evolution In a Tube

The most powerful form of directed evolution, called *DNA breeding*, is a modern derivative of classical breeding, which is familiar to anyone who has supervised the reproduction of plants or animals. The strategies are the same: Select promising parents, breed them to create a diverse pool of genetic variants and select those offspring that have the best combination of desirable traits. Our task resembles that of early man who domesticated the dog 14,000 years ago. Starting with the wolf *Canis lupus*—a magnificent animal but a poor domestic compan-

ion—prehistoric humans selectively bred those individuals with favorable traits and shaped hundreds of dog breeds in a relatively short time. Some dogs were even bred to perform highly specialized roles, such as sheepherding in border collies—a prime example of why this technique is powerful. Herding behavior is so far removed from the wolf's original behavioral genetic program that it could never have been designed rationally, even with the most sophisticated models and the most powerful computers imaginable. By contrast, directed evolution does not require *a priori* knowledge of how a system works. Its great advantage lies in the ability to modify a complex property without knowing every detail of its mechanism.

DNA breeding makes use of established biochemical techniques to apply strong selective pressure to molecules rather than whole animals. And because the evolution takes place in test tubes rather than in kennels, the entire process is notably faster.

The basic advance of DNA breeding is the recombination of diverse genetic material into novel and potentially more productive forms. This strategy can be applied at many levels, with the first requirement being some reservoir of genetic variation. In the simplest form of DNA breeding, which uses only a single gene, the functional diversity that is normally provided by a natural population needs to be generated in the laboratory. To do this, thousands of copies of the gene are randomly mutated and a small number of improved variants are identified by protein expression and screening. These selected clones are chopped into fragments of random length with a restriction enzyme, reassembled into full-length sequences and amplified using the *polymerase chain reaction*, or PCR, creating a much larger number of new combinations. This fragmentation-religation process creates novel combinations of the original set of mutations, while preserving the order of the pieces. The exponential nature of PCR generates an enormous number of new sequences in only a few hours.

Generating molecular diversity is far easier than evaluating the performance of the new molecules. Hundreds of unique progeny produced by PCR must be analyzed to find the small fraction that can perform the desired

function. The method used to screen these candidates depends on the function of the protein. If binding a specific molecular target, like a receptor or ligand is critical, then the target can be immobilized on a solid support. A solution containing the pool of new molecules is passed over the target, and molecules that bind it are retained as nonbinding proteins wash away. To cull the strongest-binding proteins, the immobilized target is washed several times with increasingly stringent solutions until only a small number of molecules remain. Other screening methods are used if some catalytic ability is sought: Reaction rates can be evaluated through chemically linked indicator molecules that change color as the reaction proceeds. With this technique, automated optical assays can screen thousands of reactions per hour for the color intensity within each sample.

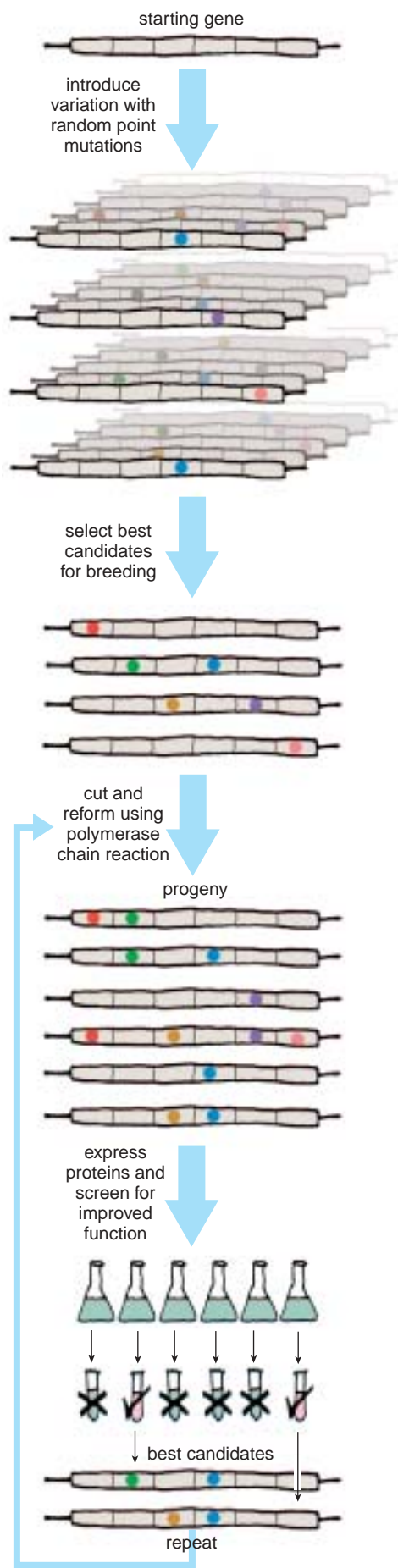
Once the population of mutated compounds has been winnowed to a scant number of contenders, positive mutations and a few remaining negative mutations from the selected candidates are refashioned into new combinations. This generates an entirely new but closely related pool of molecules to analyze for the targeted trait. Thus, the iterative process of diversity generation and selection is repeated until we achieve the desired quality or combination of qualities.

DNA Family Breeding

Directed evolution through random mutation of a single starting sequence can result in many-fold improvements in activity and substrate specificity within a few generations. However, this strategy has the disadvantage of having to examine an extremely high number of candidates to find the rare improvements that are needed as the input for DNA breeding.

A more potent variation of the DNA breeding strategy goes by the term *multi-gene shuffling* or *DNA family breeding*; it refers to the recombination of multiple equivalent genes from related species rather than random mu-

Figure 4. Single-gene breeding starts with the generation of sequence diversity through random mutagenesis. After an initial screen to select promising mutants from among thousands of clones, the best candidates can be “bred” by cutting and recombining the parental sequences to yield “progeny.” A second screen identifies new molecules with even further improvements in function.



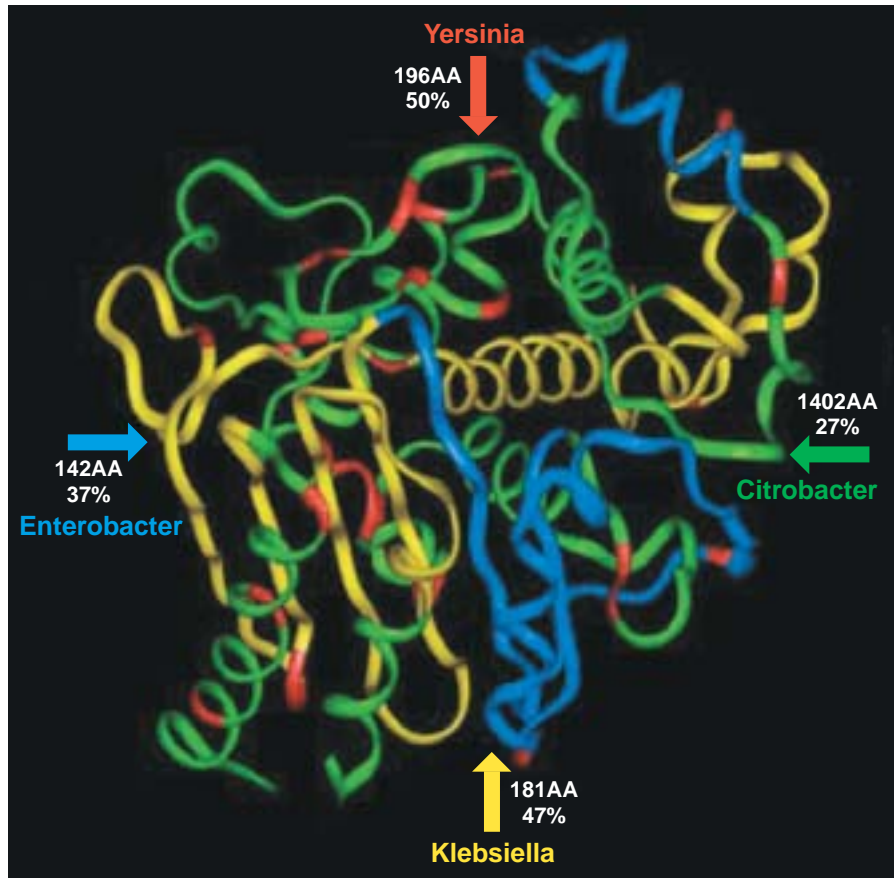
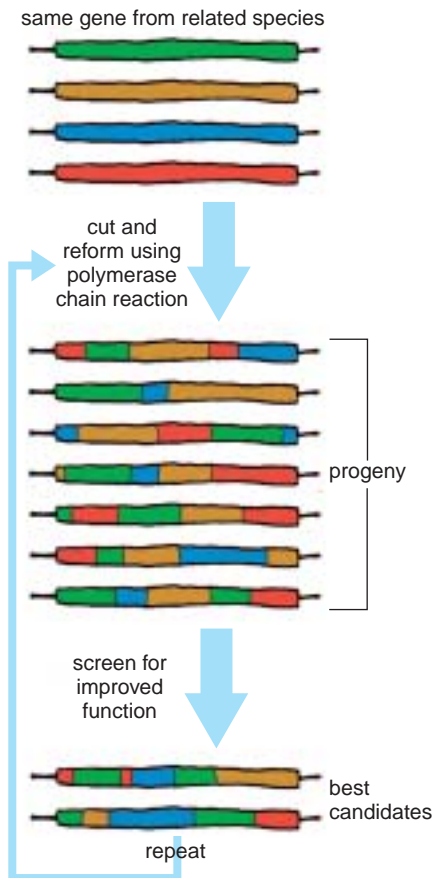


Figure 5. In DNA family breeding, the initial sequence diversity is provided by recombining the same gene from different species (left). Similar to single-gene breeding, the parental DNA is shuffled into new arrangements while preserving the order of the pieces. An example of this technique is the recombinant beta-lactamase enzyme shown above, which is a mosaic made from portions of the same gene from four distant related microbes. This particular shuffled protein is 270 times more efficient than the best parental sequence. (Courtesy of the authors.)

tagenesis of a single gene. This approach takes advantage of the fact that while extensive novelty is provided by the cross-species interchange, most of the deleterious mutations have long ago been removed by natural selection. The reassortment of old, proven mutations yields a higher frequency of functional progeny sequences, and because multi-gene shuffling starts with more than one parental sequence, it accesses a broader range of progenitor combinations. These attributes make the process more efficient, minimizing loss-of-function mutations so that fewer progeny molecules need to be screened to discover superior performers. In our lab, we observed that this variation of DNA breeding yields striking improvements in complex properties even with just a few hundred progeny. Because screening is the most laborious step in directed evolution, DNA family breeding is a major advance.

This kind of directed evolution is not only faster—it is also remarkably powerful. This can be seen in some of

our earlier work with the technique, which we presented in 1998. In that study, we compared random mutagenesis of single genes to multi-gene shuffling of an identical set of related sequences. We chose to focus on bacterial genes that encode the beta-lactamase enzyme, which are clinically relevant because these proteins inactivate the antibiotic penicillin. The starting genes came from four distantly related microbes and differed in sequence by 58 to 69 percent. We found that the best clone generated by random mutagenesis showed an eight-fold increase in beta-lactamase catalytic activity. However, when the same four genes were shuffled, the best clone was 33 times better than the random mutation champion—in other words, a 270-fold increase in the rate of catalysis.

Widespread Shuffling

We can also broaden the technology by applying the concept of DNA breeding to entire genomes rather than single genes; the advantages of such *genome*

shuffling are proportional to the size of the genome. One example of this technique used *Streptomyces* bacteria, which are valuable as natural sources of the antibiotic tylosin. Understandably, drug makers would like to find strains that show higher tylosin expression to increase production efficiency. Using a classic approach based on random point mutation and screening, a team of scientists developed *Streptomyces* variants with a six-fold increase in tylosin production. This achievement required 20 years of mutation and selection, and the cumulative screening of more than 1,000,000 mutants. Our lab performed a similar search using genome shuffling, and we found the same improvement—but it was achieved in a single year after screening only 24,000 mutants.

This technique is applied more frequently to mimic protein evolution in prokaryotes, which have continuous genes. We have developed other methods specifically suited to evolve eukaryotic proteins, which are encoded by many short stretches of DNA sepa-

rated by long noncoding spans. We can mimic eukaryotic protein evolution by focusing on those parts of the genome that actually encode proteins.

To understand how the technique works, it's helpful to look at how eukaryotic proteins are made. A single protein often resembles a string with many beads, in which each bead performs a specific task, such as binding a target. These structurally and functionally autonomous regions, or domains, work in combination to execute the overall function of the protein. One or more DNA segments called *exons*, which make up only about 1 percent of the human genome, encode each domain. Between the exons are large spans of intervening, noncoding *introns*, which get distinguished from exons during RNA splicing. Tracts of so-called junk DNA, mostly short repeated sequences that do not encode any protein information, separate the

genes; these regions make up 75 percent of the human genome.

Exons can move around the genome using a variety of mechanisms. This form of mutation based on mobile, functional modules is termed *exon shuffling*. It is predisposed to be far more useful, in frequency and magnitude, than random changes in sequence because exons or groups of exons frequently encode autonomous functional domains.

Last year, we published a method for *in vitro* exon shuffling, once again mimicking a natural evolutionary process, by PCR-based recombination of multiple exons. This technique promises to be especially useful in the creation of new therapeutic proteins because the starting variation can come entirely from human genes, rather than from nonhuman species or random mutations. We expect this will minimize the risks of triggering an immune

response to the therapeutic protein, because the constituent exons will be familiar to human immune systems on the prowl for foreign invaders.

Breeding Molecular Medicines

Technological advances are important, but ultimately they are only tools used to accomplish larger goals. Today, virtually any human protein, such as a cytokine, growth factor, antibody or enzyme, can be cloned, expressed, purified and rendered administrable as a potential therapeutic. The completion of the human genome project has enabled scientists to identify hundreds of new proteins with the potential to treat disease. However, because ideal drug properties are generally different from those of the native protein, most of these pharmaceutical candidates will require optimization to meet specific therapeutic goals. DNA breeding is ideally suited for this purpose. It can

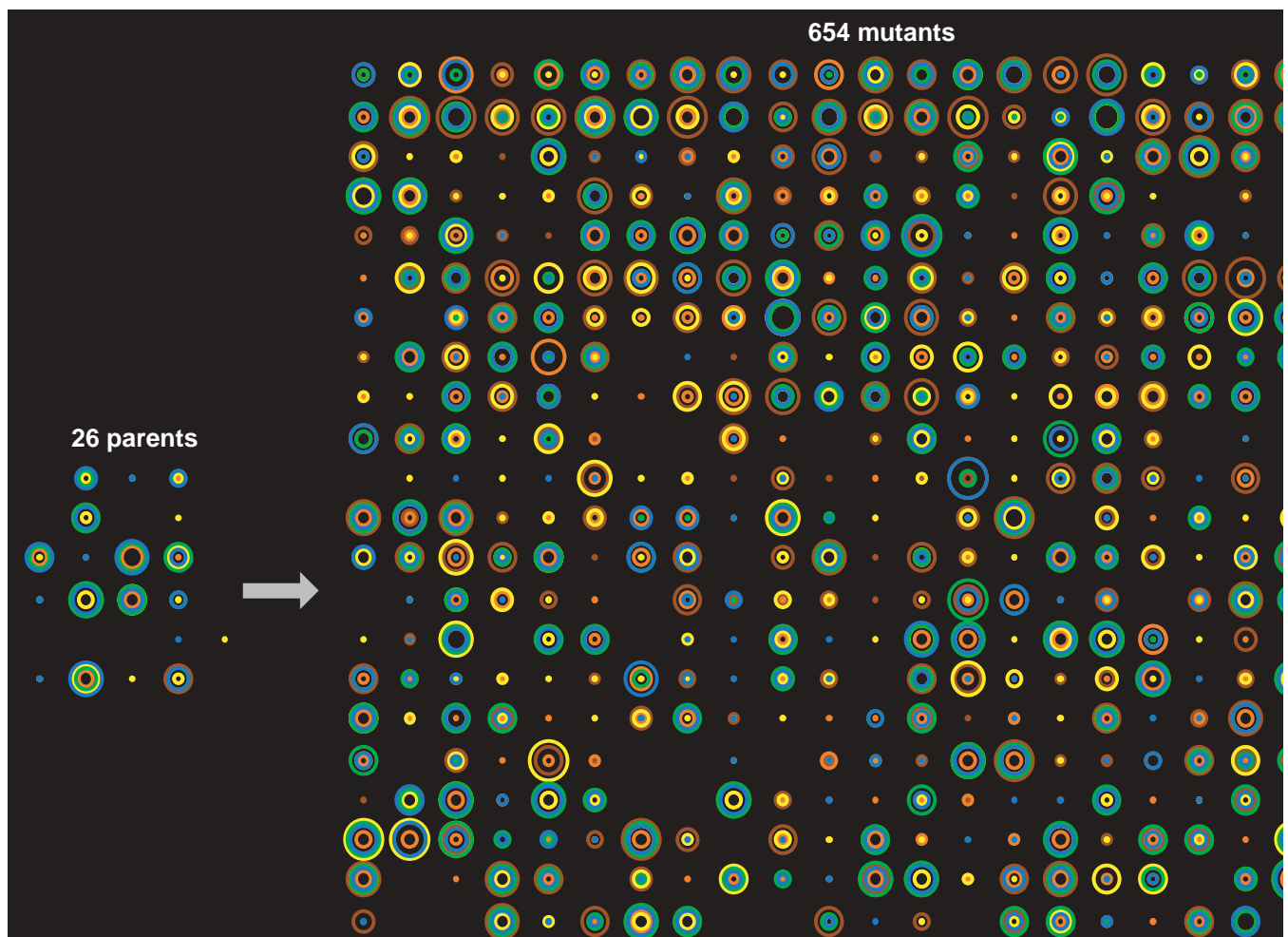


Figure 6. DNA family breeding of the subtilisin gene illustrates the potential for rapid improvement in protein function. Subtilisin is a protein-degrading enzyme used in laundry detergent and the most highly engineered enzyme in existence. Twenty-six subtilisin genes from various *Bacillus* microbes were shuffled to yield 654 progeny that were screened for enzyme activity at five commercially relevant conditions. Each clone is shown as a set of concentric circles, with the size of the circle indicating the activity of the enzyme and each assay condition represented by a different color. Several progeny showed simultaneous improvement in multiple properties over the best parental sequences.

© 2003 Sigma Xi, The Scientific Research Society. Reproduction

with permission only. Contact perms@amsci.org.

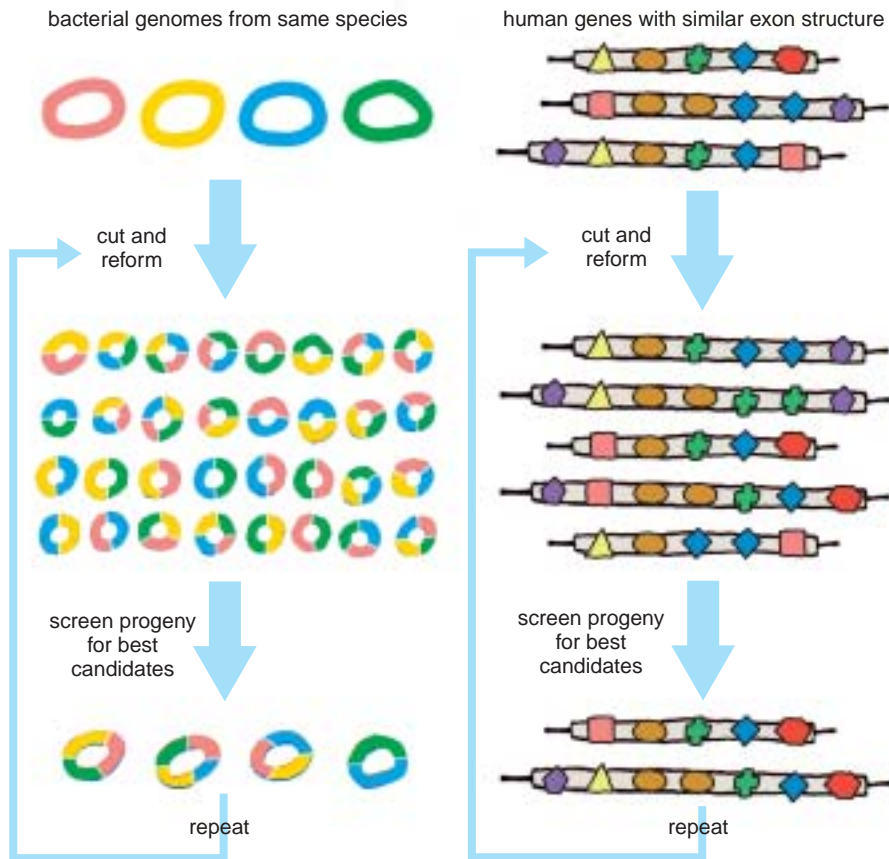


Figure 7. Genome shuffling (*left*) and exon shuffling (*right*) are two other applications of the DNA breeding technique. Genome shuffling is modeled after the natural evolution of prokaryotic organisms, whereas exon shuffling mimics eukaryotic protein evolution. In both cases the progeny are more likely to have useful properties because the breeding step starts with “proven diversity”—natural selection having already removed most deleterious point mutants.

enhance desirable biological activities such as binding activity, receptor specificity, circulation half-life, expression level and stability, and reduce undesirable side effects such as immune-system rejection and toxicity.

One example of the potential for directed evolution to improve existing drug therapies comes from our work with interferon alpha, variants of which are encoded by more than 20 genes in humans. Interferons are a group of cytokine proteins that have a broad spectrum of anticancer and antiviral activities, and human interferon alpha is already a billion-dollar product used to treat these conditions. We sought to optimize interferons to combat specific types of human cancer or viruses using DNA breeding. With a starting pool of 20 interferon alpha genes, we screened 1,700 shuffled clones with a total of only 68 antiviral assays and found several progeny proteins with improved activity on mouse cells. The best of these showed specific activity 135,000 times greater than the

existing drug. After the second round of shuffling and screening, we isolated clones with a 285,000-fold increase in potency compared with the product now available. DNA sequencing showed that the three best progeny each consisted of segments from up to five parental interferons without any amino acid point mutations—an important advantage because the lack of point mutations decreases the likelihood of an immune response, a common problem with traditional mutants. In studies using live mice, animals that were given the novel interferons were fully protected against viral infection, whereas mice given the most potent native interferon enjoyed only partial protection.

Another project focuses on our aim to fine tune immune responses by modifying the T-cell surface protein CD80. This protein is crucial to the rapid immune response that targets foreign antigens, such as pathogens, while ignoring native proteins. The decision to attack or remain quiescent is

mediated by the interaction of CD80 with two receptors with very different effects. Binding to one of them increases the immune response, which we sought to bolster in order to improve the response to vaccines and to fight cancer and infection. In contrast, when CD80 binds to the other receptor the immune response is reduced, which would be useful in the treatment of autoimmune diseases and to create tolerance for organ transplants. We strove to increase the selectivity of these binding events by shuffling CD80 genes from seven species—human, orangutan, rhesus monkey, baboon, cat, cow and rabbit—and screening the progeny for preferential binding to one receptor but not the other. The strategy was successful, and the two distinct types of CD80 variants have behaved as expected in a variety of *in vitro* assays. One of these new molecules is being evaluated further in monkeys.

A third area of research in our lab is the development of a vaccine for dengue fever, a viral disease transmitted by mosquitoes. Dengue was originally a tropical disease, but it is rapidly spreading throughout the world and has entered the United States. While surviving a bout of the disease does confer immunity, there are four viral variants or *serotypes*, and they each elicit a distinct immune response. Survivors that are reinfected with a different serotype are at increased risk of developing an often-fatal disease called dengue hemorrhagic fever, so dengue vaccines must protect against all four serotypes simultaneously. Our approach to this dilemma was to create a single-protein vaccine: Genes encoding the antigens from all four variants were shuffled, and the resulting protein progeny was tested for binding to antibodies that recognize each of the four. The best vaccine candidates were complex combinations of the key sites from all four parental serotypes. These proteins were evaluated in mice for their vaccine potential, and several induced antibody responses that cross-neutralized all four dengue serotypes—completely preventing viral infection. Candidate vaccines are now progressing to trials in primates.

We haven't yet reached the end of the potential offered by directed evolution. The application of the technique to vaccine development is especially promising: By shuffling sequences from multiple viral subtypes into a sin-

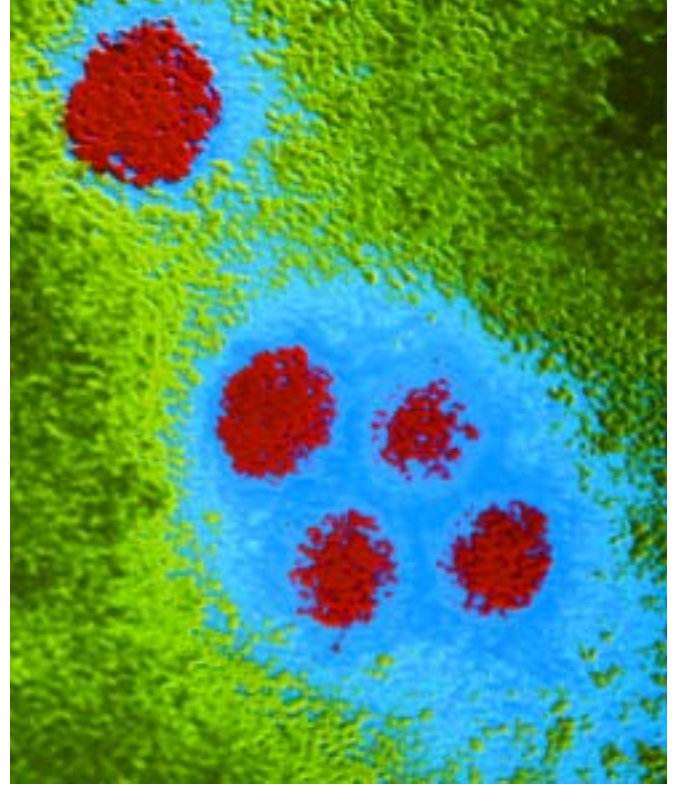


Figure 8. Dengue fever is a viral disease that is carried by the mosquito *Aedes aegypti* (left), a species that feeds during the day and prefers human beings. Dengue and dengue hemorrhagic fever can be caused by any of four distinct viral varieties, or serotypes, and this makes the effort to develop an effective vaccine especially suited to the application of DNA breeding. Shuffling the viral antigens into a single protein inoculant appears promising: Early results in mice suggest that several of the vaccine candidates prevent infection from all four serotypes. At right, a pseudocolored transmission electron micrograph illustrates dengue virus particles (red circles) within a host cell.

gle protein, we might be able to conquer influenza, human papilloma, herpes simplex 1 and 2, foot and mouth, and even bacterial infections such as those caused by the ulcer-inducing *Helicobacter pylori*. This approach may even help reverse the traditional advantage conferred by the overwhelming number of pathogens that assault us. With a world of microbes to select from, evolution exerts exceptional pressure to generate substrains and varieties that threaten to evade immunization efforts and antimicrobial drugs. Yet maybe an immune system primed with a shuffled, multistrain inoculant would be prepared to meet the summed diversity of an entire slice of the pathogenic spectrum.

As we demonstrated in 2001, laboratory-directed evolution can be used to rapidly predict the future course of natural evolution, as observed in antibiotic-resistant pathogens from clinical samples. Of course, this advantage is temporary—the shuffling technique cannot anticipate all future infectious agents. Like the Red Queen from Lewis Carroll's *Alice in Wonderland*, medical technology and the human immune system must constantly run in order to stay in place. But for the first time we have the ability to affect the treadmill itself, evolving our medical and im-

mune system defenses at a faster rate than that of its challengers.

DNA breeding is explicitly modeled on natural evolution—both are blindly opportunistic, and this is an indisputable strength of both processes. Virtually any combination of sequences may be used in a shuffling reaction, and as true, bottom-up, rational protein engineering continues to progress, it too will be incorporated into directed-evolution methodologies to increase further the fraction of useful progeny. We are optimistic that directed evolution will exponentially catalyze development of the next generation of useful proteins to flow from biotechnology's fertile cornucopia.

Bibliography

- Chang, C.-C., T. T. Chen, B. W. Cox, G. N. Dawes, J. Punnonen, W. P. C. Stemmer and P. A. Patten. 1999. Rapid evolution of a cytokine using molecular breeding. *Nature Biotechnology* 17:793–797.
- Cramer, A., S.-A. Raillard, E. Bermudez and W. P. C. Stemmer. 1998. DNA shuffling of genes from diverse species accelerates directed evolution. *Nature* 391:288–291.
- Joo, H., Z. Lin and F. H. Arnold. 1999. Laboratory evolution of peroxide-mediated cytochrome P450 hydroxylation. *Nature* 399:670–673.
- Kolkman, J. A., and W. Stemmer. 2001. Directed evolution of proteins by exon shuffling. *Nature Biotechnology* 19:423–428.

- Orencia, C., J. S. Yoon, J. E. Ness, W. P. C. Stemmer and R. C. Stevens. 2001. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nature Structural Biology* 8:238–242.
- Schmidt-Dannert, C., D. Umeno and F. H. Arnold. 2000. Molecular breeding of carotenoid biosynthetic pathways. *Nature Biotechnology* 18:750–753.
- Stemmer, W. P. C. 1994. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* 370:389–391.
- Zhang, J., G. Dawes and W. P. C. Stemmer. 1997. Evolution of an effective fucosidase from a galactosidase by DNA shuffling and screening. *Proceedings of the National Academy of Science of the United States of America* 94:4504–4509.
- Zhang, Y.-X., K. Perry, V. A. Vinci, K. Powell, W. P. C. Stemmer and S. B. DelCardayré. 2002. Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 415:644–646.

For relevant Web links, consult this issue of
American Scientist Online:

<http://www.americanscientist.org/>

IssueTOC/issue/406