

Reprogrammed: This is the structure of the endonuclease I-MsoI computationally redesigned to target a new DNA site. The redesigned enzyme displays altered site specificity with a level of target discrimination comparable to that of wild type (see ref. 8). Such methods are currently being applied to more ambitious targets, such as disease-gene hot spots.

PROTEINS BY DESIGN

New functional proteins are being built on advances in modeling and structure prediction

By David Baker

Imagine having the power to create a brand new protein – a biosensor for any small molecule, say, or a novel enzyme – on demand. It's not pure fantasy. Computational structural biology is poised to put this power into our hands.

Along with a team of research groups around the world, we have begun designing novel proteins and folds from scratch, computing amino acid sequences that will fold to create enzymatic activities never before seen in nature. The possibilities are limited only by our imaginations: Picture an endonuclease designed to thwart malaria, molecular sensors for bioterror agents, or a vaccine that HIV is less likely to evolve around.

The mechanics of these engineering feats are closely related, perhaps not surprisingly, to their logical inverse: structure prediction. Scientists have for years tried to develop methods for predicting a protein's structure simply from its amino acid sequence. Imagine that in the time it takes to sequence the genome of an

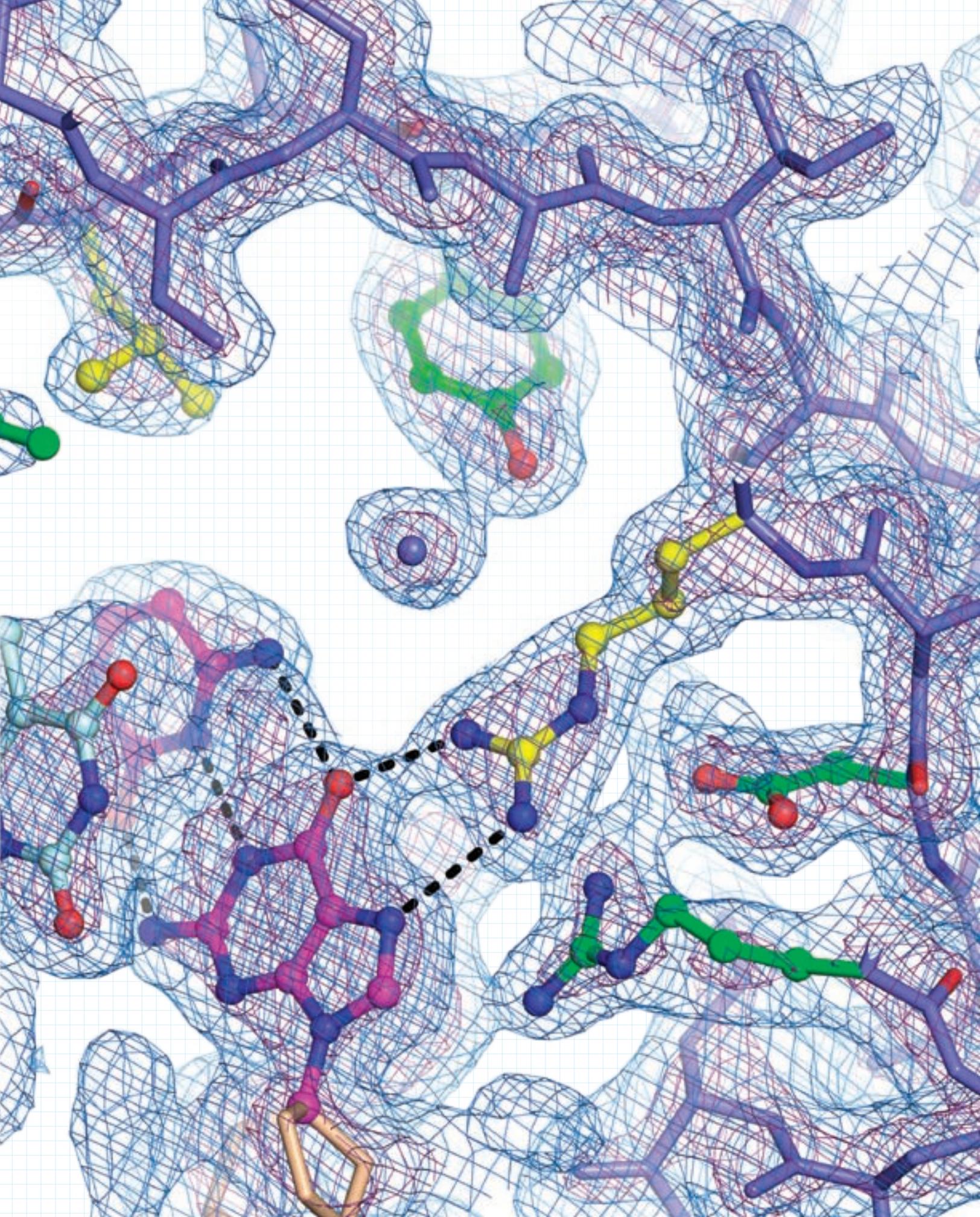
organism, scientists could also characterize the structure of each of its proteins. This would unveil biomolecular interactions, structural homologies, functional roles, and potential drug targets that might never be found from gene sequence alone.

Although there is still a long way to go, with improvements to algorithms and increases in computing power, exciting progress is being made in both prediction and design.

STRUCTURE PREDICTION

Now-classic experiments conducted with RNAase in the early 1970s demonstrated that all the information necessary to fold a protein resides in its amino acid sequence. This suggested that protein-structure prediction could be fairly straightforward. But going from sequence to structure has proven phenomenally difficult – biology's version of predicting the weather – at least in part





because even a relatively small protein can assume a vast number of possible conformations. According to “Levinthal’s Paradox,” which assumes each amino acid has three rotational degrees of freedom (and that’s an underestimate), a 100-residue sequence could adopt at least 3^{100} possible conformations.

To attack this problem, we have developed a computer program called ROSETTA, which has at its core a method for computing the energy of a given protein conformation. Eliminating unlikely structures that have, for instance, hydrophobic residues exposed to solvent, the program intelligently samples the total protein-folding landscape, testing perhaps a million or so possible conformations for the lowest energy structure.

To benchmark our progress, we have since 1998 been entering ROSETTA predictions in a worldwide structure-prediction experiment called CASP (Critical Assessment of techniques for protein Structure Prediction; www.predictioncenter.org). Instituted more than a decade ago, CASP is a sort of structural biology proving ground, a community experiment/competition in which participants are asked to predict the shape of proteins whose structures have been elucidated but not yet published. The predictions are then compared to the correct structures, and a meeting is held to discuss the results and identify the most promising methods and the most important problems remaining to be solved.

No algorithm has yet produced a precisely correct structure, but ROSETTA has performed very well in these tests, and our predictions (and those of other groups) get closer every year. A highlight of the most recent event, CASP6, for instance, was the prediction by Phil Bradley in my group of a 76-residue protein to within 1.5 Angstroms of the correct structure. Phil followed up on this achievement by showing we could predict structures with this accuracy for a number of small proteins.¹

DE NOVO DESIGN

Protein design is essentially the inverse of prediction; here, we are asking for the

amino acid sequence that will fold in such a way as to create a protein structure that carries out a desired function.

The field has made exciting progress designing proteins with new structures and functions. In 1998, Stephen Mayo and coworkers at the California Institute of Technology computed a novel sequence that folded into a naturally occurring zinc finger structure. In 2003, Brian Kuhlman, now a professor at the University of North Carolina, Chapel Hill, and Gautam Dantas in my group went a step further, using ROSETTA to design an exceptionally stable protein called Top7, which has a sequence and structure unrelated to any known protein. The 93-amino acid protein was found to be monomeric and folded, and its X-ray structure lined up remarkably well with our prediction, demonstrating that modern protein-design methodology can design brand-new proteins with atomic-level accuracy.²

While the creation of completely new structures is exciting, current efforts in the protein-design field have primarily taken aim at giving existing proteins functions found elsewhere in protein chemistry. Mayo, along with Daniel Bolon, for instance, used the catalytically inactive *Escherichia coli* protein, thioredoxin, as a scaffold for a novel enzyme capable of catalyzing the histidine-mediated hydrolysis of p-nitrophenyl acetate.³ Bill DeGrado’s group at the University of Pennsylvania engineered a metalloenzyme site into a designed four-helix bundle protein, while Homme Hellinga of Duke University Medical Center in Durham, NC, and colleagues have used members of the *E. coli* periplasmic binding protein superfamily as scaffolds upon which to design new biosensors for, among other molecules, trinitrotoluene (TNT) and glucose.⁴

Remarkably, Hellinga and coworkers succeeded in coupling their new designed biosensors to cellular signaling pathways to generate bacteria that turn blue when exposed to TNT; this work may lead to new detection methods for the land mines plaguing much of the world. More recently, Hellinga’s team successfully converted an otherwise inactive ribose-binding protein

into an extremely active catalyst of the triose phosphate isomerase reaction.⁵

The next step is the important but formidable challenge of creating enzymes to catalyze chemical reactions not performed by naturally occurring proteins (see “Designing a New Catalyst”). Our group is heading up a worldwide team of researchers representing the wide range of expertise that will be required for success. The team includes the design groups of Mayo, Hellinga, Kuhlman, and Jens Meiler; the computational chemistry expertise of William Jorgensen’s group at Yale; Ken Houk’s quantum chemistry group at the University of California, Los Angeles, which brings the ability to accurately compute structures of active sites optimal for stabilizing reaction transition states; the molecular evolution and catalytic antibody expertise of Don Hilvert’s group in Switzerland; and other groups with expertise ranging from computer science and physical chemistry to biochemistry and molecular biology. Although a tremendous challenge, with this stellar team, brought together and funded by the Defense Advanced Research Projects Agency (DARPA), I am optimistic we will see some real breakthroughs.

INTERACTION DESIGN

Beyond de novo design of catalysts, reengineering the interface of macromolecular interactions, whether between proteins or between a protein and a nucleic acid, is an important goal. Macromolecular interactions play critical regulatory and functional roles in the cell, and redesigning these could lead to the development of new drug compounds, research tools, and/or diagnostics. As with de novo protein design, we have made exciting progress in this area.

Lukasz Joachimiak, a graduate student in my group, and Tanja Kortemme, now a professor at the University of California, San Francisco, used as a test bed the interaction between colicin E7 (a nonspecific DNAase) and its inhibitor, immunity protein Im7. We first identified contacts on one partner in the pair that would destabi-

lize the complex, and then identified compensatory changes we could make in the second partner to restore the interaction. In this way, we developed new protein-protein pairs that interact with each other with subnanomolar affinities, but which do not interact with their cognate wild-type partners.⁶

We are also working to design new protein-DNA interactions. In 2002, for example, working with the groups of Ray Monnat at the University of Washington and Barry Stoddard at the Fred Hutchinson Cancer Research Center in

WHILE THE CREATION OF COMPLETELY NEW STRUCTURES IS EXCITING, CURRENT EFFORTS IN THE PROTEIN-DESIGN FIELD HAVE LARGELY AIMED AT GIVING EXISTING PROTEINS FUNCTIONS FOUND ELSEWHERE.

Seattle, we developed a novel endonuclease, *E-DreI*, by fusing domains from two separate homing enzymes, *I-DmoI* and *I-CreI*.⁷ The new enzyme has a DNA-binding specificity that is a hybrid of the two parent enzymes.

Recently, Jim Havranek in my group extended the ROSETTA design methodology to the reengineering of protein-DNA interaction specificity, and graduate student Justin Ashworth computationally redesigned the DNA-binding interface of the *I-MsoI* homing endonuclease to cleave a new DNA sequence.⁸ Justin’s experimental characterization of the redesigned enzyme showed that it efficiently cleaves the new site, but not the original site, and the high-resolution crystal structure confirmed the accuracy of the design (see image, p. 27).

We are now using this computational design approach to try to create therapeutically useful endonucleases. Designed enzymes could be introduced into mutant cells or organisms, together with a wild-type copy of the mutant gene, to drive gene therapy, for instance. Following cleavage of the disease gene by the novel ►

enzyme, the wild-type sequence would be used to drive DNA repair, thereby fixing the genetic defect. In collaboration with a team of researchers worldwide we are also seeking to design new endonucleases that would inactivate the genes in mosquitoes required for the malaria parasite to survive and propagate; such engineered enzymes could play roles in malaria eradication programs.

OTHER APPLICATIONS

Protein design methodology can be useful in unanticipated ways. A number of significant human diseases, including Alzheimer and Parkinson, are associ-

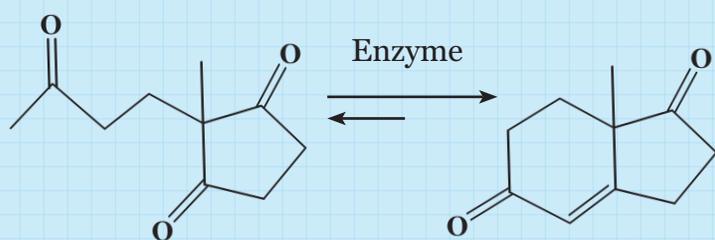
ated with proteins that misfold to form amyloid fibrils. David Eisenberg's group at UCLA made a major breakthrough last year in the understanding of how this occurs when they reported the first high-resolution structure of an amyloid-forming peptide.⁹ The study revealed a set of interactions that seem very likely to be general to most, if not all, amyloid structures. John Karanicolas in my group has been collaborating with Eisenberg's group to try to predict the portions of proteins responsible for amyloid fiber formation. The design methodology in ROSETTA was used to identify sequences compatible with a generalized model of their amyloid structure.¹⁰

DESIGNING A NEW CATALYST

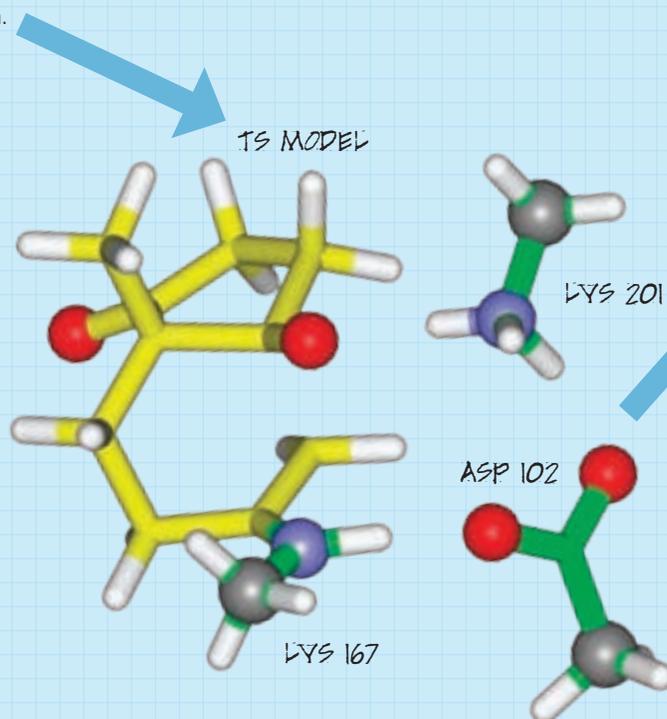
In a recent demonstration of computational protein-design principles, we approached an intramolecular aldol reaction whose substrate is not catalyzed by naturally occurring enzymes. The aldol reaction constitutes one of the most powerful tools for the formation of carbon-carbon bonds both in nature and the lab. Our goal is to design a non-natural aldolase able to catalyze reactions of non-natural substrates.

TARGET REACTION:

Hajos-Parrish-Eder-Sauer-Wiechert intramolecular aldol reaction.



SITE DESCRIPTION: The proposed catalytic residues in the active site are based on three key residues - Lysine 167, Aspartate 102, and Lysine 201 - found in the native deoxyribose-phosphate aldolase (DERA). The transition state (TS) model (yellow) and optimal functional group positions of the catalytic residues (green) are calculated by quantum mechanical (QM) methods.



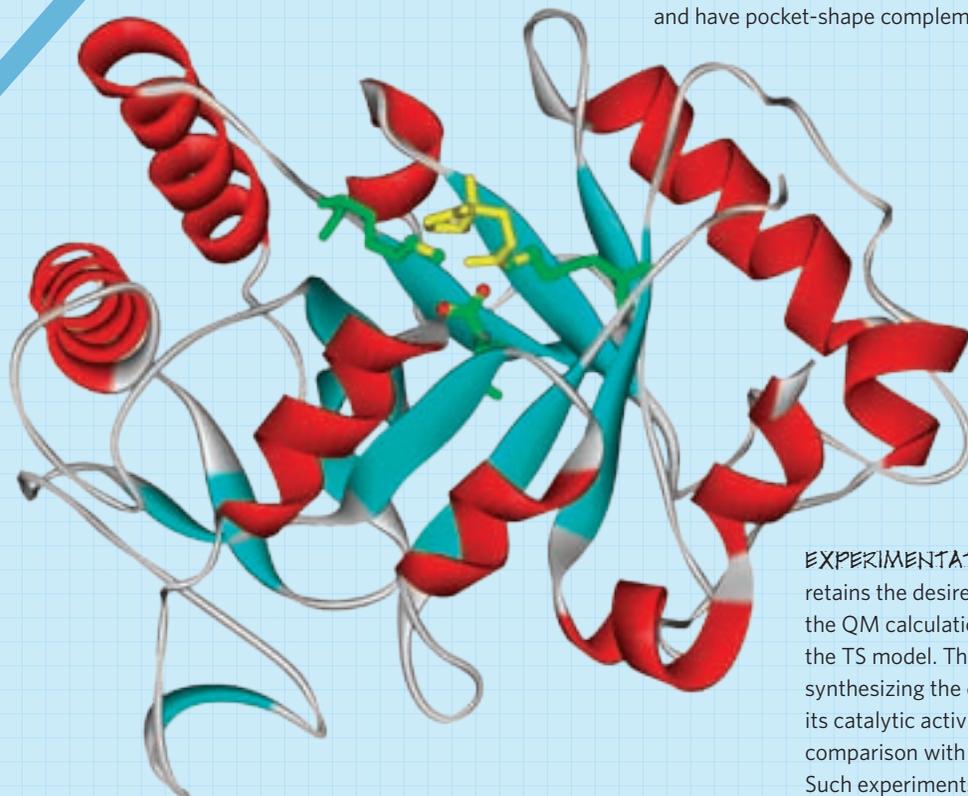
Protein design also has commercial implications. Xencor, a protein-engineering firm in Monrovia, Calif., for instance, has used its Protein Design Automation technology to develop XPro-1595, a dominant-negative inhibitor of tumor necrosis factor alpha, which is expected to enter clinical trials in 2006. Researchers at Sangamo Biosciences recently announced their success in repairing genes in vivo using a designed, zinc-finger-based enzyme. Biosensors, like Hellinga's novel designs, could be used to build chips, or even to regulate an implantable artificial pancreas. And then there are the nearly limitless options in organic chemistry for enzymes capable of accelerating complicated synthesis reac-

tions, such as that of the antimalarial drug, artemisinin.

Yet we remain constrained by computational power. The calculations required to probe and refine a protein's structure are intensive, as they must account for all the various rotational and conformation degrees of freedom. It took 15 processor-days (3.2 GHz CPUs) to model the docking interface between two proteins for a community experiment similar to CASP, the Critical Assessment of Prediction of Interactions (CAPRI, <http://capri.ebi.ac.uk>). The same calculations would have taken a year on decade-old hardware. It took approximately 150 CPU-days to make the folding predictions for our 2005 *Science* paper! ▶

ROSETTA

DESIGN: Rosetta algorithms search for new active sites in a library of protein scaffolds and design the residues surrounding these potential active sites to further stabilize the TS model. The protein should fold in a way that will place the desired catalytic groups in the active site and have pocket-shape complementary to the TS model.



EXPERIMENTATION: This designed protein retains the desired catalytic geometry from the QM calculation and the pocket to bind the TS model. The next steps would include synthesizing the designed protein, evaluating its catalytic activity, and crystallizing it for comparison with the predicted structure. Such experiments will further test our understanding of catalytic sites.

To enable the searching of the huge conformational and sequence spaces associated with protein-structure prediction and design, David Kim in my group has developed a distributed computing version of ROSETTA, called rosetta@home, which harnesses the collective computing power of tens of thousands of computers around the world.

At last count rosetta@home had some 67,000 users worldwide, yielding a combined 29 teraflops of computational power – good enough for eighth place on the November 2005 list of the world's top 500 supercomputers. The resulting web of computing power is greatly increasing the rate at which we can improve the ROSETTA structure-prediction methodology, as we can test new ideas much more quickly. Currently rosetta@home users are field-testing our new methods on the just-released prediction targets for the ongoing CASP7 structure-prediction experiment, which closes August 4.

When CASP7 is completed, we will also direct our users' computational power to important design problems, such as the development of a vaccine for HIV. The reason people have difficulty fighting off HIV is that the major viral coat protein, gp120, has many highly variable loops. Immune responses to the virus tend to be aimed at these variable regions and hence are relatively ineffective against the virus. There are, however, a few key regions of gp120 that cannot change because they are critical to virus infectivity. Bill Schief in my group, in collaboration with the groups of Peter Kwong, Rich Wyatt, and Gary Nabel at the National Institutes of Health, Leo Stamatatos at the Seattle Biomedical Research Institute, Roland Strong at the Fred Hutchinson Cancer Research Center, and Dennis Burton at Scripps Research Institute, is now designing a series of novel protein vaccines designed to mimic these Achilles-heel regions of the virus. Our collaborators will be testing whether antibodies made against these potential vaccines can neutralize virus infectivity.

These examples are just the beginning. I would guess the potential for protein design is nearly as vast as the diversity of biology itself. With the power to create unheard of catalysts, improved biomolecular interactions, and genuinely useful new proteins, I consider myself privileged to be part of a field limited merely by imagination and computing power. And I am even more privileged to have had the opportunity to work with the wonderful students and postdoctoral fellows who have come through my research group, many of whom continue at their respective universities to develop the ROSETTA prediction and design methodology and apply it to problems I never would have dreamed of. ■

David Baker is a professor of biochemistry and a Howard Hughes Medical Institute investigator at the University of Washington, Seattle. You can sign up for the rosetta@home project from his lab Web page, www.bakerlab.org
dbaker@the-scientist.com

REFERENCES

1. P. Bradley et al., "Toward high-resolution de novo structure prediction for small proteins," *Science*, 309:1868–71, 2005.
2. B. Kuhlman et al., "Design of a novel globular protein fold with atomic level accuracy," *Science*, 302:1364–8, 2003.
3. D.N. Bolon, S.L. Mayo, "Enzyme-like proteins by computational design," *Proc Natl Acad Sci*, 98:14274–9, 2001.
4. L.L. Looger et al., "Computational design of receptor and sensor proteins with novel functions," *Nature*, 423:185–90, 2003.
5. M.A. Dwyer et al., "Computational design of a biologically active enzyme," *Science*, 304:1967–71, 2004.
6. T. Kortemme et al., "Computational redesign of protein-protein interaction specificity," *Nat Struct Mol Biol*, 11:371–9, 2004.
7. B.S. Chevalier et al., "Design, activity, and structure of a highly specific artificial endonuclease," *Mol Cell*, 10:895–905, 2002.
8. J. Ashworth et al., "Computational redesign of endonuclease DNA binding and cleavage specificity," *Nature*, in press, 2006.
9. R. Nelson, "Structure of the cross- β spine of amyloid-like fibrils," *Nature*, 435:773–8, 2005.
10. M.J. Thompson et al., "The 3D profile method for identifying fibril-forming segments of proteins," *Proc Natl Acad Sci*, 103:4074–8, March 14, 2006.