CHEMICAL BIOLOGY

From structure to function

By docking potential substrates into the active site of an enzyme of known structure, researchers accurately predict the catalytic activity of the enzyme.

In recent years, structural genomics initiatives have had great success in generating large numbers of structures of so-far uncharacterized proteins. But, as Brian Shoichet at the University of California at San Francisco puts it, "even if you know what a protein looks like, this doesn't necessarily mean you know what it does. We decided to take the next step and ask [whether] we can broadly predict function of an enzyme if we know the structure." In collaboration with Frank Raushel at Texas A&M University and Steve Almo at Albert Einstein College of Medicine, Shoichet and his group have tackled this question for Tm0936, a protein from the microorganism *Thermotoga maritima* (Hermann *et al*, 2007).

The researchers took a computational docking approach, in which a database of potential substrates is docked into the active site of the enzyme. The degree of structural complementation, or fit, of a given substrate could then predict what the activity of the enzyme is. This approach has been used before for the design of inhibitors, but the challenges for substrate prediction are much greater. "Inhibitor design is hard enough," Shoichet says, "but in that case all you have to do is get a molecule that the enzyme recognizes. It doesn't have to do the next step, which is to turn the molecule over, to do the catalytic reaction on it. And capturing that is no joke."

To address this challenge, the researchers docked not the ground-state of substrates but rather a high-energy intermediate state that would be recognized by the enzyme, if catalysis were possible with that substrate. For the 4,207-metabolite database that they screened, this involved computational conversion of all of the relevant functional groups to their high-energy states, and resulted in a final set of about 22,500 forms
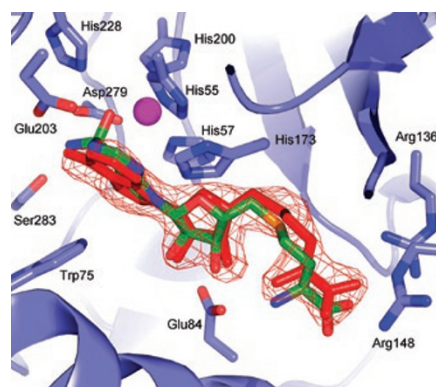

Figure 1 | Comparing predicted with actual structure. The high-energy form of SAH (green) in the active site of Tm0936, as predicted by docking, was superimposed on the crystal structure of the enzyme-substrate complex (substrate in red). Reprinted from *Nature*.

of the test molecules. As the enzyme under study, Tm0936, is broadly classifiable as belonging to the amidohydrolase superfamily, the researchers could limit the possible forms of the test substrates by conversion only of the appropriate functional groups.

Beginning with a good database is critical, and creating one is not easy. "If you're talking about 4,000 molecules with 20,000 states, that's 20,000 problems," Shoichet says. "Each molecule has its own questions—'What's the protonation state?', 'What's the isomerization state?', 'Do you have to test all the possible stereocenters?', 'Which ones are likely to be the most biologically relevant?'" All the more laudable, therefore, is that the results of generating the test structures and docking them into the active site of Tm0936 were so informative.

When the substrates were ranked in terms of fit, the researchers found adenosine and several adenosine analogs among the highest ranked molecules; nine of the ten top-scoring hits belonged to this category. They tested four of these molecules experimentally for catalytic turnover by Tm0936, and verified that three of them, 5-methyl-

thioadenosine, *S*-adenosyl-L-homocysteine (SAH) and adenosine itself, were bona fide substrates of the enzyme. Notably, the researchers determined the X-ray crystal structure of Tm0936 with the product of the SAH deamination reaction, and found that the docking prediction and the actual structure of the complex had close congruence (**Fig. 1**). Thus, Tm0936 was identified as an adenosine deaminase.

In another recent related study, substrate docking was applied to predict function of a member of the enolase superfamily. In that case, a homology model of the test enzyme, rather than a crystal structure, was used as the template for docking (Song *et al.*, 2007). Subsequent experimental studies led to identification of the enzyme as an *N*-succinyl arginine/lysine racemase. The amidohydrolase and enolase superfamilies contain about 6,000 and 1,000 proteins, respectively. The stage now seems set for function prediction for more of these.

There are certainly challenges ahead, however. The right substrate for a particular enzyme may simply not be in the database being tested. In addition, problems could arise for enzymes that undergo substantial conformational change during catalysis. Although retrospective studies indicate that docking high-energy intermediates makes this sort of predictive approach less susceptible to conformation-based problems, it does not eliminate them completely.

But the structural genomics pipelines continue to flow, the raw material is at hand, and the extent to which the docking approach is successful at going from form to function will become clear as it is applied to more and more protein structures.

Natalie de Souza

RESEARCH PAPERS
Hermann, J.C. *et al*. Structure-based activity prediction for an enzyme of unknown function. *Nature* **448**, 775–779 (2007).
Song, L. *et al*. Prediction and assignment of function for a divergent *N*-succinylamino acid racemase. *Nat. Chem. Biol.* **3**, 486–491 (2007).