# SURVEYING THE BIND

*Using various computational methods, molecular modelers are delving into the cracks and crevices of proteins to derive drug targets.*

BY RANDALL C. WILLIS

As the various genomic and proteomic databases fill with sequences, the pressure is increasing to turn this data into information and eventually into druggable targets. Efforts are being made on a variety of fronts to clone, express, purify, and determine the structure of many of these proteins, but this can be an arduous and expensive prospect.

Even if you are lucky enough to have a protein that purifies in high yield, folds well, and either is stable in an NMR tube or crystallizes well, there are no guarantees that you can determine its structure with a bound ligand or that its structure will exhibit obvious ligand-binding sites.

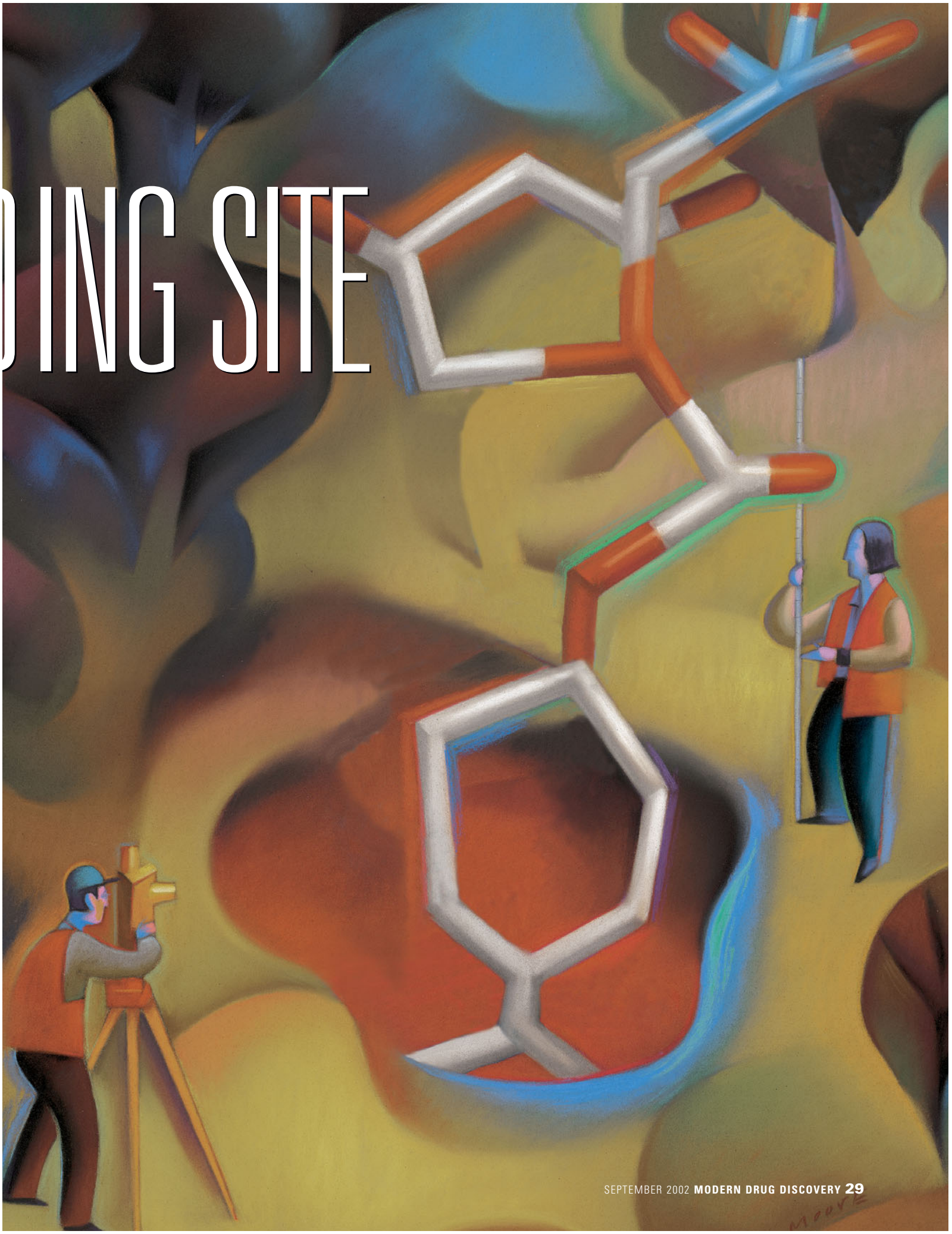
There are two basic cases to examine when looking for ligand-binding sites. In the first, the protein structure has been determined but the binding site remains unknown. In the second and much more complicated case, the researcher might be looking for an alternative site, other than the known ligand-binding site, that offers a new target on the same protein. In this case, the researcher is looking to take advantage of allosteric protein structural changes that do not necessarily block the ligand from binding but might alter how the protein performs its natural function in some other way.

So if you want to use a protein's structure for virtual drug screening, what do you do if you don't know which of its many cracks and crevices might be important to its function? Basically, how do you find its ligand-binding sites?

## HOMOLOGY-BASED METHODS

In many cases, the genomic and protein databases available already contain all the information that you need to locate the potential ligand-binding sites on your protein sequence or structure. "If you're looking for something that is the binding site of a natural ligand, it's reasonably straightforward to find those through the information that you get from protein sequence alignment," says David Lowis, senior director of applied science at Tripos, Inc. (St. Louis, *www.tripos.com*). An example of such a method is the ICM system developed by Ruben Abagyan, researcher

ILLUSTRATION: LARRY MOORE



# ING SITE

# Advertisers

at Scripps Research Institute and co-founder of MolSoft LLC (La Jolla, CA, *www.molsoft.com*), which uses various alignment, prediction, and optimization algorithms to determine a test protein's structure.

In addition, "evolutionary tracing leverages the fact that the key amino acids in the binding site are retained throughout evolution and in multiple species," says Osman Guner, director of lead identification and optimization at Accelrys (San Diego, *www.accelrys.com*). "When you compare the sequences and identify certain blocks that are maintained, these might represent binding or active sites."

An example of this phenomenon is the serine–histidine–aspartic acid triad that comprises the catalytic site of serine proteases.

On the other hand, if the structure of the protein of interest is unknown, it becomes more difficult to align it with known structures. Instead, it might be necessary to model the structure of the protein onto a known protein using a computational method such as threading, in which the amino acid string of the test protein is run along the polypeptide chain of a homologous protein whose structure is known (see "Going for fold in Asilomar", *Modern Drug Discovery*, November/December 2000, pp 40–46). In theory, this will provide the researcher with a model structure in which the ligand-binding site has already been determined.

The Augmented Homology Modeling component of the ProMax database from Structural Bioinformatics (San Diego, *www.strubix.com*) performs such a function by using the crystal structure of one member of a protein family to model the structure of any of the other members. Unfortunately, inserted or deleted sequences might adversely affect how well the test and template proteins fit together.

Another problem in homology modeling arises from the fact that two proteins with the same function are not necessarily sequentially homologous. In fact, there are numerous examples in which 3-D structure has been evolutionarily conserved over the amino acid sequence. But this situation need not stop the protein modeler from identifying a protein's ligand-binding site, because several algorithms and databases have been developed in which proteins are aligned on the basis of structure, not sequence (Table 1). For example, the FSSP (fold classification based on structure–structure alignment of proteins) database includes the structural neighbors of all proteins contained in the Protein Databank (PDB). After being unable to identify sequence homologues of barley endochitinase, a protein involved in plant defense, Holm and Sander used the FSSP database to identify a subclass of lysozymes that were structural homologues (*1*). As the number of protein structures available increases, the ability to determine ligand-binding sites by homology or similarity modeling can only improve. But for now, there are still limits on how effective these methods can be. Furthermore, says Lowis, "I think where the biggest challenge comes is when you're looking for a secondary or auxiliary binding site that may not be that well conserved."

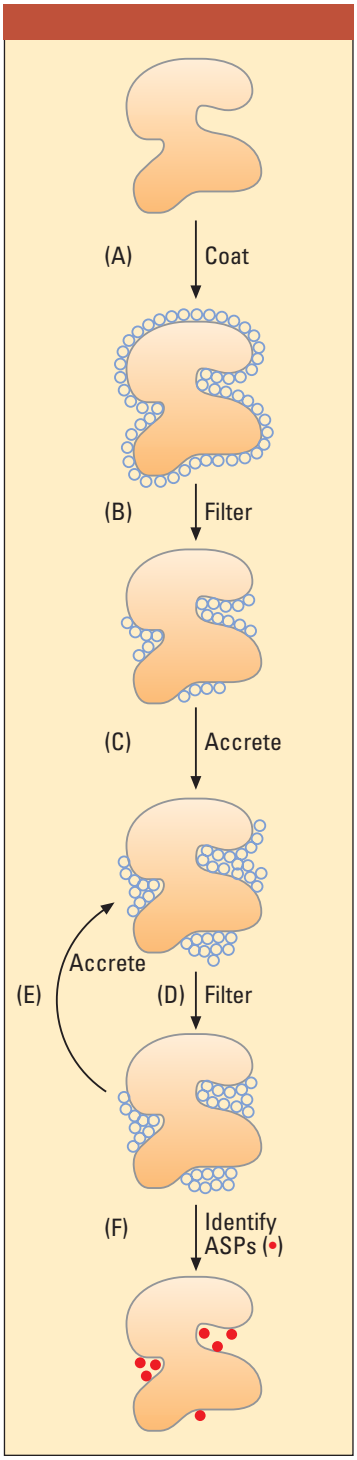



**Figure 1. A flood of information.** Researchers developed the putative active site with spheres (PASS) method to determine where ligands might bind on a protein. (A) The protein's surface is coated with spheres. (B) Spheres insufficiently buried are removed (filtered). (C) New spheres are accreted. (D) These new spheres are filtered. (E) The process is repeated. (F) Probe weights are calculated and active site points (ASPs) are identified. (Adapted from reference 2.)

"But the nice thing about those kind of sites," continues Lowis, "is that they're probably much more specific to that [particular] protein for when you're looking to generate drugs with fewer side effects." Thus, computational chemists have developed other modeling methods to find these sites.

## SOLVATION-BASED METHODS

One way of identifying cavities on protein surfaces is by "coating" the surface with model spheres, removing the excess, and then determining where the greatest concentration of spheres remains. Depending on the algorithm, these spheres can represent merely geometric shapes or they can mimic water molecules.

One such method is the putative active site with spheres (PASS) algorithm developed by Patrick Brady, Jr., and Pieter Stouten, then at DuPont Pharmaceuticals (Wilmington, DE) (*2*). In this case, a set of spheres is layered over the surface of a protein such that each sphere is in contact with three unique protein atoms (Figure 1). The least buried (most solvent-accessible) spheres are then filtered out. This burial count is calculated by counting the number of protein atoms within 8 Å of an individual sphere, and spheres that fall below a preset threshold are eliminated. A second layer of spheres is then added to the protein, and these too are filtered for solvent accessibility. This process is then repeated iteratively until no further spheres can be added.

Cavities filled with spheres are then assigned active site points (ASPs)—potential binding sites for ligand atoms—which are central probes in regions that contain many deeply buried spheres. Each of the ASPs is assigned a probe weight that is proportional to the number of spheres in its vicinity and their burial counts. Using this method, Brady and Stouten successfully identified the binding sites on 29 of 32 proteins with

**Table 1**

**Databases for homology-based modeling of ligand-binding sites**

| Database | Comparison | Ref. |
|---|---|---|
| CATH | Protein function | 5 |
| CAVBASE | Protein cavities | 1 |
| Entrez Structure | Protein sequence and structure | 6 |
| FSSP | Protein structure | 7 |
| ProMax | Protein structure | 8 |
| RELIBASE | Protein–ligand complexes | 9 |

X-ray crystal structures that they pulled from the PDB.

Other algorithms use a similar method to identify protein cavities (Table 2).

"SiteID has two different algorithms that it can use to identify possible binding sites," says Tripos's Lowis. "The first is a solvation-type algorithm where we effectively place the protein in a fine grid and place probes that are designed to simulate a water molecule at each grid point. We then look for those grid points that are surrounded by a significant number of protein atoms." (The second algorithm, based on physical and structural parameters, is described below.)

## GRID-BASED AND OTHER METHODS

Rather than use virtual water molecules to delineate potential binding sites, other algorithms rely on grid-based methods to determine where to start docking their ligands.

"Envision placing a protein into a three-dimensional grid and deciding which grid points belong to the protein (red) and which are available (green)," says Guner, describing how the Active Site Finding tool of Insight II, shortly to be incorporated into Accelrys's Discovery Studio platform, works. "Then envision a big eraser. You start erasing the green points and approach the protein until you reach a red point. Depending on the size of the eraser, there is going to be a moment when the eraser won't fit into a particular cavity. It will continue erasing outside of the domain, and this cavity is then identified as a potential binding site. There are typically multiple cavities in a protein, and there are typically multiple binding sites and possibly multiple active sites."

The size of the virtual eraser can be changed in the program such that larger erasers will leave more cavities while smaller erasers will elucidate fewer. Fundamentally, the eraser size parallels the known ligand sizes.

The program LigSite, developed by Manfred Hendlich and colleagues at the University of Marburg (Germany), similarly embeds the protein in a regularly spaced grid, and lattice intersections falling within a protein atom's van der Waals sphere are discarded (*3*). The remaining lattice points are then scored by how deeply buried they are within surface depressions, which is based on scanning along the *x*, *y*, and *z* axes and the four diagonals for areas that are enclosed by protein atoms. Adjacent lattice points with high degrees of burial indicate continuous cavities.

Still other programs, like the second mode of SiteID, move away from grids and solvation altogether, relying instead on physical and structural parameters. "It makes a spreadsheet of the protein residues," says Lowis, "where each row represents the protein atoms and the columns include information about those atoms such as their [solvent] exposure, distance from the center of the protein, lipophilicity, hydrophilicity, et cetera. We then use clustering algorithms to look for [for example] hydrophobic exposed residues. As soon as you identify those groups, you can visualize those residues on the protein structure."

## PSEUDORECEPTORS

In rare cases, the only information available to a researcher is the biological information about which ligands affect an unknown protein and to what degree. But even here, there is enough information to begin modeling a binding site on a ligand receptor, even in the total absence of information about the protein. Some people would

**Table 2**

**Binding site modeling programs**

| Type | Program | Website |
|---|---|---|
| Solvation-based | DynaPharm | www.strubix.com/dphOSP.htm |
| | MOLCAD | www.tripos.com |
| | Putative active site with spheres (PASS) | www.delanet.com/~bradygp/pass/ |
| | SiteID | www.tripos.com |
| Grid-based | Cerius$^2$-LigandFit | www.accelrys.com |
| | LigSite | Reference 3 |
| | Site Finder | www.chemcomp.com/feature/sitefind.htm |
| Pseudoreceptor | Cerius$^2$-Receptor | www.accelrys.com/cerius2/receptor_ms.html |
| | Genetically evolved receptor models (GERM) | www.finchcms.edu/biochem/Walters/germ.html |
| | Hypothetical active site lattice (HASL) | www.eslc.vabiotech.com/hasl/ |
| | Pseudoatomic receptor model (PARM) | Reference 10 |
| | Quasi-atomistic receptor surface models | www.biograf.ch/software.html |

even argue that the non-binding-site atoms of a protein simply use up valuable computing space and can be ignored. To that end, these researchers focus their efforts on the computation of virtual pseudo- or minireceptors.

A wide variety of programs can calculate pseudoreceptors, but they all rely on three implicit assumptions:

- all of the test ligands bind to a common site on the protein,
- biological activity is proportional to ligand–protein affinity, and
- all ligands bind in a low-energy conformation (although not necessarily the lowest energy).

An example of such an algorithm is the genetically evolved receptor models (GERM) method developed by Eric Walters and his colleagues at Finch University of Health Sciences (Chicago) (*4*).

In GERM, several ligands are modeled in their low-energy conformations and are superimposed to determine which chemical groups are most important for binding (e.g., all of the ligands carry an aromatic side group in the same general location). Ideally, the compounds encompass a broad range of structural types and biological activities. A shell of 45–60 atoms is then created around the superimposed ligands, and these atoms are classed according to their chemical nature (e.g., a hydrogen bond donor pseudoreceptor atom is placed near a hydrogen bond acceptor ligand atom). Because the atoms are placed randomly, they can occupy several positions, so a series of shells is created. The atoms of these shells are used to create a "gene" that consists of the types of atom (aliphatic H, carbonyl C, hydroxyl O, etc.) and location in space (Figure 2). Thus, each shell has a corresponding gene.

Each of these genes is then passed through a genetic algorithm in which pairs of parent (initial) genes are allowed to rearrange components and to mutate. Each offspring gene is then tested for its ability to bind the ligand group. If it binds better than any of the parent genes, it is retained, and the low parent is removed from the algorithm. No duplicates of existing genes are allowed, and the algorithm proceeds until no new genes survive the screening process.

At the end of the evolutionary process, you are left with one pseudoreceptor that best fits the binding characteristics of the known ligands. This pseudoreceptor can then be used as a template for virtual high-throughput screening for new ligands with better binding properties.

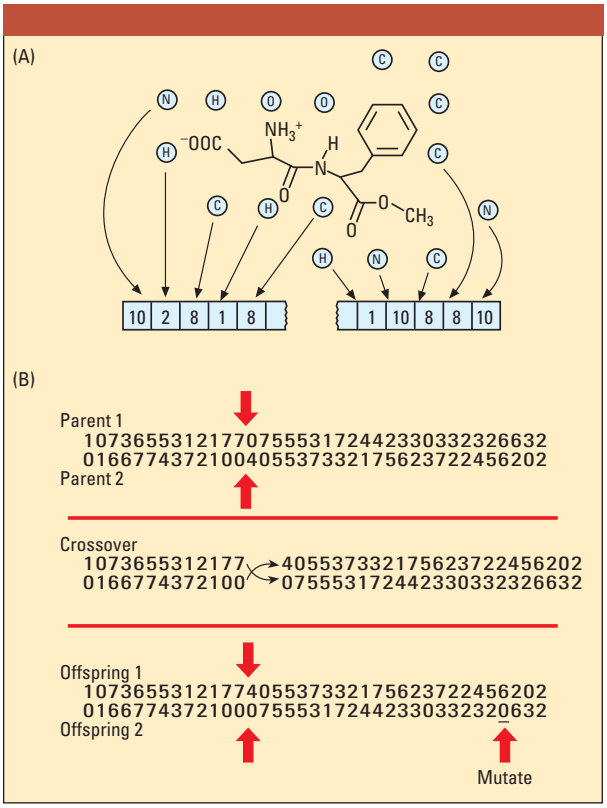



**Figure 2. GERM in action.** To create a pseudoreceptor using the genetically engineered receptor models (GERM) algorithm, a series of shells composed of different atoms (A) is created around ligand molecules that perturb a common protein. (B) The location and identity of the shell atoms are then coded to form parent genes (top) that can recombine or mutate (middle) in a genetic algorithm to form new genes (bottom), the shells of which might better fit the ligands. (Adapted from reference 4.)

## CORRECT OR SPECIOUS?

Ultimately, however, the problem remains that just because something might be a ligand-binding site does not mean that it is.

"If you're in a situation where you know something binds but you don't know where," says Lowis, "then you start thinking about using these kinds of site-identification algorithms in conjunction with a good docking program that's going to predict binding affinities well. The idea being that you might have four or five pockets that you're interested in and you've got to be able to prioritize those with some kind of good docking tool."

These docking tools help the researcher determine how well a particular ligand will bind within a given protein cavity and, like the programs just described, work on the basis of various ligand and protein characteristics (see "2001: A dock odyssey", *Modern Drug Discovery*, September 2001, pp 26–32).

"The quality of modeling work depends on the quality of the data that you put into it," Lowis adds. "As soon as you're starting to guess at things like binding sites, I would definitely try to run that in parallel with a ligand-based approach and look for a consensus between the two methods."

Although having the structure of a target protein with a variety of bound ligands remains the ideal starting point for virtual high-throughput screening, the various modeling techniques available and in development offer hope to researchers that their projects do not have to stop until better data become available.

## REFERENCES


(1) Sotriffer, C.; Klebe, G. *Il Farmaco* **2002,** *57,* 243–251.
(2) Brady, Jr., G. P.; Stouten, P. F. W. *J. Comput.-Aided Mol. Des.* **2000,** *14,* 383–401.
(3) Hendlich, M.; Rippman, F.; Barnickel, G. *J. Mol. Graphics* **1997,** *15,* 359–363.
(4) Walters, D. E.; Hinds, R. M. *J. Med. Chem.* **1994,** *37,* 2527–2536.
(5) www.biochem.ucl.ac.uk/bsm/cath_new.
(6) www.ncbi.nlm.nih.gov/entrez/structure.html.
(7) www2.ebi.ac.uk/dali/fssp/fssp.html.
(8) www.strubix.com/PromOSP.htm.
(9) http://relibase.ccdc.cam.ac.uk.
(10) Chen, H.; Zhou, J.; Xie, G. *J. Chem. Inf. Comput. Sci.* **1998,** *38,* 243–250.



**Randall C. Willis** is a senior associate editor of *Modern Drug Discovery*. Send your comments or questions about this article to mdd@acs.org or to the Editorial Office address on page 3.■