X-RAY CRYSTALLOGRAPHY

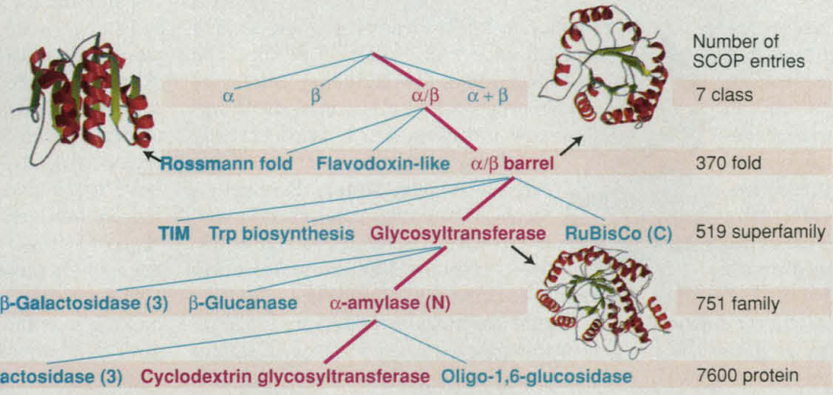
Taking a Structured Approach To Understanding Proteins

Like many x-ray crystallographers, Sung-Hou Kim of the University of California, Berkeley, has built his reputation by solving the structure of whatever protein seemed most interesting at the time. But over the past few years, the flood of gene sequences pouring into the public databases has convinced him that he needs to change his ways. Kim understands that knowing the structures of the proteins encoded by these genes would give a big boost to biomedical researchers seeking to understand their functions. Yet, he complains, "there are too many genes," making it "too difficult" to reasonably determine them all.

So to tackle this daunting task, he has taken a new approach in his research and with about 70 other researchers has begun formulating a structural biology undertaking that may one day rival the Human Genome Project. Meeting 24 and 25 January at Argonne National Laboratory outside Chicago, they agreed that the goal should be to determine the structures of a core set of molecules that are representative of all types of protein structure. That targeted approach, they said, would yield almost as good a picture of the universe of proteins—and be much more practical—than attempting to get the three-dimensional (3D) structure of every protein that turns up through gene sequencing, which in the case of the human genome could be 100,000.

Once in hand, these representative structures will help improve the computer programs that predict the structure and function of other proteins directly from the DNA sequence of their genes. "We will be playing with a complete deck," says George Rose, a biophysicist at Johns Hopkins University in Baltimore who develops these kinds of predictive models. Having a complete catalog of fundamental protein structures should also speed up the work of x-ray crystallographers and nuclear magnetic resonance researchers when a computer prediction of a new protein's structure isn't sufficient—for example, when the protein is a potential drug target.

Indicating the interest in the idea, the workshop organizers had more than twice as many applicants as they could handle, forcing them to turn many away. Representatives from the Department of Energy and the National Institutes of Health, two agencies that might help foot the bill for a protein structure initiative, also received the idea warmly. "I am convinced that an effort of this type is important and will give enormous benefit," says Marvin Cassman, director of the National Institute of General Medical Sciences in Bethesda, Maryland.

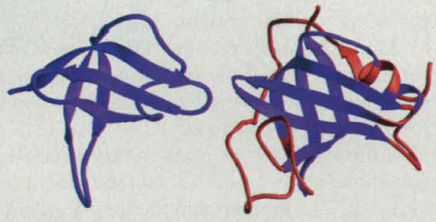


Protein clan. Hierarchical scheme reveals structural kinships among proteins.

SOURCE: D. S. BRENNER/STANFORD UNIVERSITY

He points out, however, that before the agencies can support the initiative, the researchers will have to resolve many uncertainties, including which set of proteins should be studied. The number could range from a modest 1000 or so to close to 10,000. At $150,000 per structure, the total bill for the larger number could reach $1.5 billion. Meeting participants noted, however, that more efficient techniques for determining structures could reduce the cost.

Which proteins researchers pick to study would depend on which approach to classifying proteins they rely on. Currently, genome researchers group genes—and presumably their proteins—into functional classes based on similarities in DNA sequences. But that approach has its limitations, as sometimes genes that look very different yield proteins with very similar 3D structures and functions. "With the programs out there, it's easy to cluster proteins," says Argonne computational biologist Paul Bash. "But it's quite another matter to produce meaningful clusters."



Close cousins. These look-alike proteins have genes with very dissimilar sequences.

G.T. MONTELIONE AND W. FENG/RUTGERS UNIVERSITY

Consequently, several groups have taken a different tack—classifying proteins based on how their strings of amino acids are arranged in spirals, turns, and zigzag sheets. Because of the close connection between a protein's structure and the way it operates, this type of classification "is a far more powerful way of recognizing evolutionary relationships," says Steven Brenner, a computational biologist at Stanford.

Alexey Murzin at the Medical Research Council Centre for Protein Engineering in Cambridge, United Kingdom, was one of the first to classify proteins by their structures, beginning in the 1970s with existing ones and keeping tabs on new ones ever since. His approach first groups into a single family proteins whose genes have closely related sequences. Families whose proteins have very similar 3D structures, but not necessarily similar genes, are then grouped into superfamilies, which are in turn grouped into "folds," a technical term that refers to particular arrangements of key 3D components, such as helices or pleated sheets. And finally, the folds are combined into classes.

Brenner's team turned Murzin's work into a computerized database, called SCOP for Structural Classification of Proteins, in 1994. A researcher seeking to study the protein product of a new gene can compare its sequence to those in the database and instantly know what other proteins it's likely to resemble. In addition, "you can see all the known folds," explains Brenner. "Before, you couldn't begin to figure this all out," because there was no comprehensive organization of these structures. Thus far, SCOP has grouped the more than 7600 proteins in the protein database into 751 families, 519 superfamilies, 370 folds, and seven classes.

Structural biologist Christine Orengo of University College in London is developing a second classification scheme, which she has turned into a computerized database she calls CATH for Class, Architecture, Topology, Homology, some of the categories in her scheme. Like SCOP, CATH groups proteins in families based on their sequences, then into superfamilies according to similarities in their 3D structures, and finally into ever broader categories defined in much the same way as the

SCOP categories but having different names.

These databases are already proving to structural biologists that they need to choose new targets for their efforts. In 1994, for example, only 10% of the proteins submitted to the protein database represented new families—with no sequence similarity to other proteins—and only a third of those had a new fold. Based on the number of new folds found over time, researchers predict that the universe of proteins may contain 1000 or more folds. It could take decades to find them all, Brenner points out, unless crystallographers change their tactics.

For this reason, Kim and others have begun to streamline fold discovery by trying to pick proteins that should have novel folds and working on several of them at once. As they go through the difficult task of making and crystallizing the proteins, they eliminate those candidates that might bog down the work because they are hard to synthesize in sufficient quantities or won't crystallize properly. In this way, "[we] get the structure information we need from the one that's the easiest to get it from," says Tom Terwilliger, a structural biologist at Los Alamos National Laboratory in New Mexico.

One way to efficiently find proteins that have novel folds, says Kim, is to choose genes from organisms that don't have many proteins. "The idea is just to pick a small, self-replicating organism that presumably has a smaller number of genes but a large fraction of the three-dimensional folds," he explains. For example, using genome data from The Institute for Genomic Research (TIGR) in Rockville, Maryland, Kim's team hopes to nail down the core structures of many of the 1700 proteins coded for in the completely sequenced genome of *Methanococcus jannaschii* (*Science*, 23 August 1996, pp. 1043, 1058).

As a test, Kim had TIGR send him 10 *Methanococcus* genes that look like known genes and 10 with no recognizable similarity to previously discovered genes. He and his colleagues put those genes in bacteria and eliminated from the study any that did not yield heat-stable proteins that could be purified readily. So far, they have purified five proteins and solved the structures of three. While none of them turned out to contain a new fold, Kim is confident this strategy is a good one for identifying those proteins that do.

The meeting participants did not reach a consensus about what to do next, but they did agree to meet twice more, once in April and again in October, to come up with a more concrete plan. A few are worried that most structural biologists are too independent to sign on to a project in which their goals are so well-defined. But Kim is adamant. "We don't have any choice," he says. "What else can we do if we're trying to get the function of as many gene [products] as possible?"

–Elizabeth Pennisi

CANADIAN SOCIAL SCIENCE

Database Funding Left Out in the Cold

OTTAWA—Last month's killer ice storm, which caused extensive damage in eastern Canada and Maine and left hundreds of thousands without power for up to a month, was a reminder of modern society's vulnerability to the elements. To a group of Canadian social science researchers, it also presented a rare opportunity to build a research database on how people coped with the once-in-a-lifetime disaster. Unfortunately, the researchers soon confronted another cold reality: No Canadian funding agency was prepared even to review a proposal to fund such a venture.

This snub, social scientists contend, is the latest piece of evidence that their field doesn't receive the same respect—or financial support—accorded the natural and biomedical sciences. There is no government program to fund infrastructure projects in the social sciences, and leading practitioners in the field say a new $600 million program to fund infrastructure needs at Canadian universities will leave most social science projects out in the cold. "It's the task of the social scientist and the humanist to set the larger context. And for any government to exclude [their needs] is myopic in the extreme," says University of Victoria historian Eric Sager, who spent 2 years cobbling together funding for a $700,000 project to digitize a randomized 5% sample of the 1901 Canadian population census.

The proposed disaster database would compile and analyze the mountains of documents generated by this winter's storm. Potential studies include the dynamics of disaster response, the adequacy of public emergency preparedness programs, society's reliance on technology, the impact of severe individual stress on community relations, and the role of the national reserves in maintaining civil order during natural disasters. The results, say proponents like University of Montreal sociologist Paul Bernard, would greatly expand knowledge of how systems respond to crises.

The government's Social Sciences and Humanities Research Council (SSHRC) is desperately trying to broker a funding package among public agencies and departments that would allow the social scientists to begin amassing such a disaster database. Its own coffers are bare after taking a $20 million cut in the last 3 years, and "there is no mechanism or program" to create, maintain, and update databases or other forms of social sciences research infrastructure, says Chad Gaffield, president of the Humanities and Social Sciences Federation of Canada and a historian at the University of Ottawa.

AP PHOTO/T. RIVIERES / *LE NOUVELLISTE* VIA CP

Nature 1, Quebec 0. Transmission tower was no match for ice storm.

The much-ballyhooed Canada Foundation for Innovation (CFI), a 5-year infrastructure program announced last year (*Science*, 28 February 1997, p. 1256), won't help those seeking to establish the disaster database, because its first awards won't be made before fall. In any case, it's not yet clear what kinds of social science infrastructure projects would be eligible for CFI support.

Initially, the CFI said databases weren't eligible. After objections from the community, CFI officials drew a murky line between the creation and maintenance of databases. The former will be eligible, the latter will not. But that's not the only bone of contention: Annual additions to existing databases and digitization projects and virtual libraries are also nonstarters. The line between an eligible database and an ineligible virtual library is, however, "subject to interpretation," says CFI spokesperson Janet Halliwell.

University of Calgary academic vice president Ron Bond says that the definitional squabble indicates that the infrastructure deck is stacked against the social sciences. However, acting CFI President Denis Gagnon says the fund's long-term intention is to serve all disciplines. And he says he'll recommend "drastic measures" should the agency discover after a few competitions that the social sciences have been shortchanged or that database proposals are being routinely rejected.

Social scientists aren't sure how to respond to the controversy. Some, like Univer-