

Data bank struggles as protein imaging ups its game

Hybrid methods to solve structures of molecular machines create a storage headache.

BY EWEN CALLAWAY

Structural biology, the mapping of complex biological molecules such as proteins, is in the grip of a revolution. The field has long been dominated by X-ray crystallography, a technique made iconic by its role in decoding the DNA double helix in the 1950s. But the need to tackle more complex structures and to watch ‘molecular machines’ function in real time is fuelling a shift towards hybrid imaging methods that can create moving models.

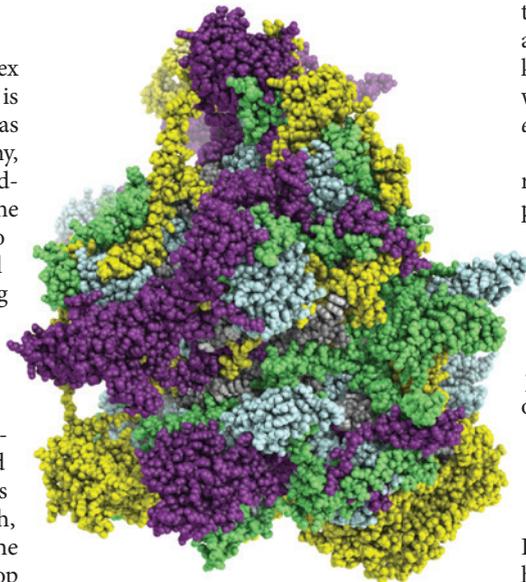
That is posing a challenge for the world’s official repository for protein structures: the Protein Data Bank (PDB), which relies almost exclusively on crystallography data and lacks the standards and software infrastructure to archive structures described by hybrid methods. This month, leaders of the four organizations around the world that host the data bank held a workshop in Hinxton, UK, to hatch a plan to ensure that hybrid models and their insights into fundamental biology and disease do not get lost.

Historically, structural biology has focused on generating three-dimensional (3D) descriptions of individual proteins. In many cases, this is a task perfect for crystallography, in which a molecule is bombarded with X-rays and the pattern of scattered radiation reveals the position of each atom. The technique underpins dozens of discoveries that led to Nobel prizes.

Archiving structures in the centralized, free PDB is crucial because it enables other researchers to use them to address questions never imagined by their discoverers. Most journals will publish structures only if they have been deposited in the PDB. This year, the database topped 100,000 registered structures, the vast majority of which were determined using X-ray crystallography (see *Nature* 509, 260; 2014).

But in the past decade or so, structural biology has moved on. Researchers now want to describe intricate cellular structures made up of dozens, or even hundreds, of proteins that move relative to each other do jobs such as recycling proteins or copying chromosomes. These molecular machines cannot be coaxed into the tidy, immobile crystals required for X-ray

“These days, being a crystallographer is not good enough.”



A subunit of a ribosome, a molecular machine.

crystallography. “These days, being a crystallographer is not good enough,” says Gerard Kleywegt, a structural biologist at the European Bioinformatics Institute in Hinxton, who heads the European annex of the PDB.

Hybrid methods take an ‘everything but the kitchen sink’ approach to structural biology, incorporating many different techniques. Some can offer a dynamic view of a molecular machine in motion; for example, fluorescence resonance energy transfer measures the distance and interactions between proteins. Others, such as cryo-electron microscopy, can deliver near-atomic detail of entire complexes without the need to crystallize them. Computer programs then integrate the various bits of information — including data from crystallography-friendly proteins inside the molecular machine — to produce a 3D model that best fits the data.

The scientific literature is now studied with products of the hybrid approach. In 2012, structural computational biologist Andrej Sali of the University of California, San Francisco, and his collaborators used hybrid methods to describe the structure of the 26S proteasome complex (K. Lasker *et al. Proc. Natl Acad. Sci. USA.* 109, 1380–1387; 2012), which recycles proteins and may malfunction in neurodegenerative diseases such as Alzheimer’s. The researchers have now used the model to identify potential drugs

that alter the proteasome’s activity. This year, another team published a hybrid model of the key HIV proteins that sneak the virus into a cell, which may help in vaccine design (M. Pancera *et al. Nature* <http://doi.org/wfz>; 2014).

The hybrid approach has also tackled the ribosome, which produces proteins; the nuclear pore complex, which provides a gateway between the genome in the nucleus and the rest of the cell; and the molecular syringes made by bacteria that inject proteins into cells. Models of many more molecular machines are expected. “We’re going to enter a period of exponential growth in the generation of these hybrid structures,” says Stephen Burley, a structural biologist at Rutgers University in Piscataway, New Jersey, who heads one of the two US annexes of the PDB.

At the PDB workshop, on 6–7 October, Kleywegt, Burley and three dozen others hashed out the challenges that these triumphs are creating for the PDB. Crystallography yields a standardized set of data files in which a structure and its level of precision are self-evident; by contrast, the underlying data for the hybrid models exist in a mishmash of formats such as X-ray diffraction patterns or electron-micrograph pictures. And going from raw data to a model involves more steps with hybrid methods than in crystallography; it also requires more assumptions, often leading to multiple possible ways of interpreting the results.

Most workshop attendees agreed that it will be crucial for structural-biology databases to capture not just the hybrid models’ raw data, but also how the models were put together, so that other researchers can verify and build on them. But there are many questions, such as how to store and distribute the data sets, which are much larger than crystallography files. The meeting ended with an agreement to seek funding for a new bank centred on molecular machines — and to come up with a name for it.

It is imperative to find a way to curate hybrid structures if structural biology is to realize its potential, says cell biologist Jan Ellenberg of the European Molecular Biology Laboratory in Heidelberg, Germany, who led one of the teams that modelled the nuclear pore complex. Ultimately, he says, “we want to have the molecular structure of an entire cell. That’s still science fiction at the moment — but it’s somewhere we can get to in 10, 20 years.” ■