EDITORIAL

Banking on protein structural data

Jeremy Berg
is professor of computational and systems biology at the University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. jberg@pitt.edu

In 1953, the proposed structure of DNA magnificently linked biological function and structure. By contrast, 4 years later, the first elucidation of the structure of a protein—myoglobin, by Kendrew and colleagues—revealed an inelegant shape, described disdainfully as a “visceral knot.” Additional complexity, as well as some general principles, was revealed as more protein structures were solved over the next decade. In 1971, scientists at Brookhaven National Laboratory launched the Protein Data Bank (PDB) as a repository to collect and make available the atomic coordinates of structures (seven at the time) to interested parties. The PDB now includes more than 180,000 structures, and this resource has fueled an incalculable number of advances, including the recent development of powerful structure prediction tools.

Biology takes place in three dimensions, yet most biological information is stored in one-dimensional sequences of DNA that encode the amino acid sequences of proteins. The transition from one to three dimensions is accomplished through the spontaneous folding of a sequence of amino acids into a folded protein structure. Comparing elucidated structures revealed that proteins that are at least 30% identical in amino acid sequence almost always have the same folded structure; evolutionarily, structure is much more conserved than sequence. Conversely, some short stretches of five or more amino acids can adopt completely different structures; structure is context dependent. Thus, the relationship between sequence and structure is not a simple one.

> “…this resource has fueled an incalculable number of advances…”

Predicting protein structures from sequences has been a grand challenge for decades. By 1994, fueled by the explosion of sequences, biophysicist John Moult and colleagues organized the first Critical Assessment of Structure Prediction (CASP) meeting. CASP is based on blinded assessments, which are common in clinical trials. Sequences of proteins whose structures had been determined but not publicly shared were made available to would-be predictors to develop and submit structural predictions for subsequent independent assessment. The first CASP meeting was somewhat depressing because the results revealed that predictors were doing substantially worse than they thought. CASP meetings have continued every 2 years and have driven the field forward through feedback and competition. The most recent CASP meeting, in November 2020, was shaken by results from the company DeepMind. Its AlphaFold program performed substantially better than other programs had in the past, producing many results that are of similar quality to that of experimental structures. The RoseTTAFold program, developed by the laboratory of structural biologist David Baker, builds on this laboratory’s previous work, combined with insights from the DeepMind success (see page 871). The results of both programs are sufficiently good that many are claiming that these represent relatively general (but certainly not perfect, and incomplete) solutions to the structure prediction problem. Notably, both groups have provided their computer code for their methods for others to use, test, and enhance.

These programs are based on deep-learning artificial intelligence methods. Such approaches depend on the availability of many thousands of questions with known answers to train the neural networks at their core. Thus, without the sequences with known structures from structural biologists from around the world shared in the PDB, these approaches would not have been feasible. The teams that developed these powerful programs deserve great credit for their accomplishments, but these stand on a foundation of the results from billions of dollars of public fund investments in structural biology and the sustained support of the PDB from around the world (now overseen by the Worldwide PDB).

Policies from funders, publishers, and the scientific community have led to requirements that reported structures be promptly deposited in the PDB. As someone who has interacted with the PDB as a consumer, a contributor, a policy-maker, and a funder, I have experienced the power and challenges of trying to optimize such a public resource. The cultural shifts, at the cutting (and often bleeding) edge of open science, were often controversial, but it is hard to argue that they have not both increased the impact of individual determined structures and accelerated scientific progress in many ways. The ever-growing PDB provides researchers with a universe of structures with which to compare their favorite structures. The new structure prediction tools expand this universe further and provide truly compelling evidence of the power of open science. Moreover, these tools bring truth to an old saying in structural biology circles, “The structure prediction problem has been solved; it’s hiding in the PDB.”

–Jeremy Berg

PHOTO: WENDIE BERG

10.1126/science.abl8151

Science

Banking on protein structural data

Jeremy Berg

Science **373** (6557), 835.
DOI: 10.1126/science.abl8151

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.