are also seen in the unliganded subunit.

What is the benefit of using two subunits when FAcD can use one? To address this question, the authors compared *B*-factor values, which typically reflect protein motions associated with the crystallographic data; the more mobile a part of the protein, the higher the *B*-factor values. Kim *et al*. found an increase in the *B*-factor values on the unliganded subunit during catalysis, indicating increased motions. The increased motions coincide with a reduction in water molecules observed in the unliganded subunit (see the figure). Thus, the unliganded subunit pays the entropic cost incurred by the catalytic subunit upon ligand binding.

Kim *et al.*'s study provides remarkable progress toward understanding the full range of functional mechanisms used by enzymes to achieve their catalytic power. The key to success was the use of several biophysical approaches that collectively illuminate the process. It remains to be shown whether other enzymes use similar mechanisms to carry out catalysis, to what extent the involved residues contribute to the process, and whether amino acid substitutions can alter catalytic rates by means of changes in protein motions. Answers to these questions will require a similarly integrated approach and will yield fascinating insights in the coming years. ■

> ***"The freeze-trapped structures show that the two subunits work in concert during the whole [catalytic] process."***

REFERENCES

1. S. J. Benkovic, S. Hammes-Schiffer, *Science* **301**, 1196 (2003).
2. H. Frauenfelder, S. G. Sligar, P. G. Wolynes, *Science* **254**, 1598 (1991).
3. K. Henzler-Wildman, D. Kern, *Nature* **450**, 964 (2007).
4. A. G. Palmer 3rd, *Acc. Chem. Res.* **48**, 457 (2015).
5. T. H. Kim, P. Mehrabi, Z. Ren, A. Sljoka, C. Ing, A. Bezginov, L. Ye, R. Pomès, R.S. Prosser, E.F. Pai, *Science* **355**, eaag2355 (2017).
6. A. J. Baldwin, L. E. Kay, *Nat. Chem. Biol.* **5**, 808 (2009).
7. S. R. Tzeng, C. G. Kalodimos, *Nat. Chem. Biol.* **9**, 462 (2013).
8. D. D. Boehr, D. McElheny, H. J. Dyson, P. E. Wright, *Science* **313**, 1638 (2006).
9. S. R. Tzeng, C. G. Kalodimos, *Nature* **462**, 368 (2009).
10. S. R. Tzeng, C. G. Kalodimos, *Nature* **488**, 236 (2012).
11. K. K. Frederick, M. S. Marlow, K. G. Valentine, A. J. Wand, *Nature* **448**, 325 (2007).
12. N. Popovych, S. Sun, R. H. Ebright, C. G. Kalodimos, *Nat. Struct. Mol. Biol.* **13**, 831 (2006).

10.1126/science.aal4632

PROTEIN STRUCTURE

Big-data approaches to protein structure prediction

Metagenomics sequence data give protein structure prediction a boost

By **Johannes Söding**

A protein's structure determines its function. Experimental protein structure determination is cumbersome and costly, which has driven the search for methods that can predict protein structure from sequence information (*1*). About half of the known proteins are amenable to comparative modeling; that is, an evolutionarily related protein of known structure can be used as a template for modeling the unknown structure. For the remaining proteins, no satisfactory solution had been found. On page 294 of this issue, Ovchinnikov *et al.* (*2*) used recently developed methodology for predicting intra-protein amino acid contacts in combination with protein sequences from metagenomics of microbial DNA to compute reliable models for 622 protein families, and discovered more than 100 new folds along the way. The fast-paced growth of metagenomics data should enable reliable structure prediction of many more protein families.

This advance builds on progress in two areas: prediction of contacts between pairs of amino acids and template-free (de novo) protein structure prediction. Most de novo methods rely on assembling proteins from short peptide fragments (*3*, *4*). These fragments are sampled from the database of known protein structures based on the similarity of their sequences. Thousands of trial structures are assembled and assessed by their expected stability according to detailed statistical and heuristic energy functions derived from known structures. Although many striking successes of predicting complex and large proteins were achieved, only a fraction of models for proteins larger than 100 amino acids usually had the correct fold. Worse, in contrast to comparative modeling, it was hard to tell when a model would be reliable.

Progress in de novo prediction was slow until it was discovered that contacts between amino acids can be reliably predicted from large multiple sequence alignments. The basic idea was old: When the side chains of two amino acid residues are in contact, the selection pressure to maintain this favorable interaction can lead to compensatory mutations. A correlation signature develops between the amino acids of the two corresponding columns in a large multiple sequence alignment of the protein family (see the figure).

However, the contact predictions suffered from high false-positive rates—too many pairs had sizable correlations without being in contact. A breakthrough occurred when it was found that the prime source of false positives was from correlations arising through indirect chains of interactions. Similar problems had been solved in physics and statistics, and such methods could be applied to distinguish correlations from direct statistical couplings between the amino acids (*5*, *6*).

In the past 5 years, these methods have been improved and applied to predict the structures of many proteins (*7*) and even protein complexes (*8*). A limitation to their more large-scale application is the requirement of the multiple sequence alignments containing hundreds to thousands of sequences. Out of 5211 protein families in the Pfam database (*9*) for which no template is found, only ~400 have alignments large enough to yield reliable models.

A solution lies in metagenomics, in which microbial DNA is extracted and sequenced directly from environmental samples. The pool of species sequenced is much more diverse than those that can be cultured in the lab, which greatly enhances contact prediction. By carefully including these lower-quality sequences in their sequence searches to build the alignments, Ovchinnikov *et al.* brought the number of Pfam families with sufficient diversity up to 1300. Crucially, they have developed and convincingly validated a quality score that allows them to identify those 614 out of the 1300 models that are very likely to be correct. The quality of their models is impressively highlighted by the recently published structures for six protein families that had been modeled by Ovchinnikov *et al.* (*10*). All six models, most of which have several hundred amino acids and complex folds, are highly similar to the experimental structures.

Quantitative and Computational Biology, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany. Email: soeding@mpibpc.mpg.de

Much of protein space is still in the dark: Pfam covers only 47% of residues in representative proteomes (*9*), and the present study still leaves 4600 Pfam families without structural information. Nonetheless, the rapid growth of metagenomics data means that more and more protein families will become amenable to the new approach (see the figure). Ovchinnikov *et al.* used a database of 2 billion metagenomic protein sequences, but a much larger number of sequences is probably contained in metagenomics data sets. One challenge will be to improve sequence searching, clustering (*11*), and assembly algorithms to extract more and higher-quality protein sequences. One limitation of using metagenomics sequences is the preponderance of prokaryotic sequences. Metatranscriptomic experiments targeted more specifically at eukaryotes might address this.

Recent results from this year's blind Critical Assessment of Techniques for Structure Prediction (CASP12) (*12*) show that a few other methods perform almost as well or as well as the Rosetta server of Ovchinnikov *et al.* In any case, apart from the new quality score developed in (*2*), a huge advantage of Rosetta is the free availability of its source code and a collaborative and open community of researchers, both of which allow the field to profit on a large scale from the progress described here.

Homo-oligomers still present particular difficulties because intra- and interchain contacts are hard to distinguish. Predicting functional sites as well as ligand and cofactor binding sites would also be very useful. Rigorous statistical methods that would make use of the types of coupled amino acids should further lower the required alignment diversity and thereby make the approach applicable to smaller protein families. ■

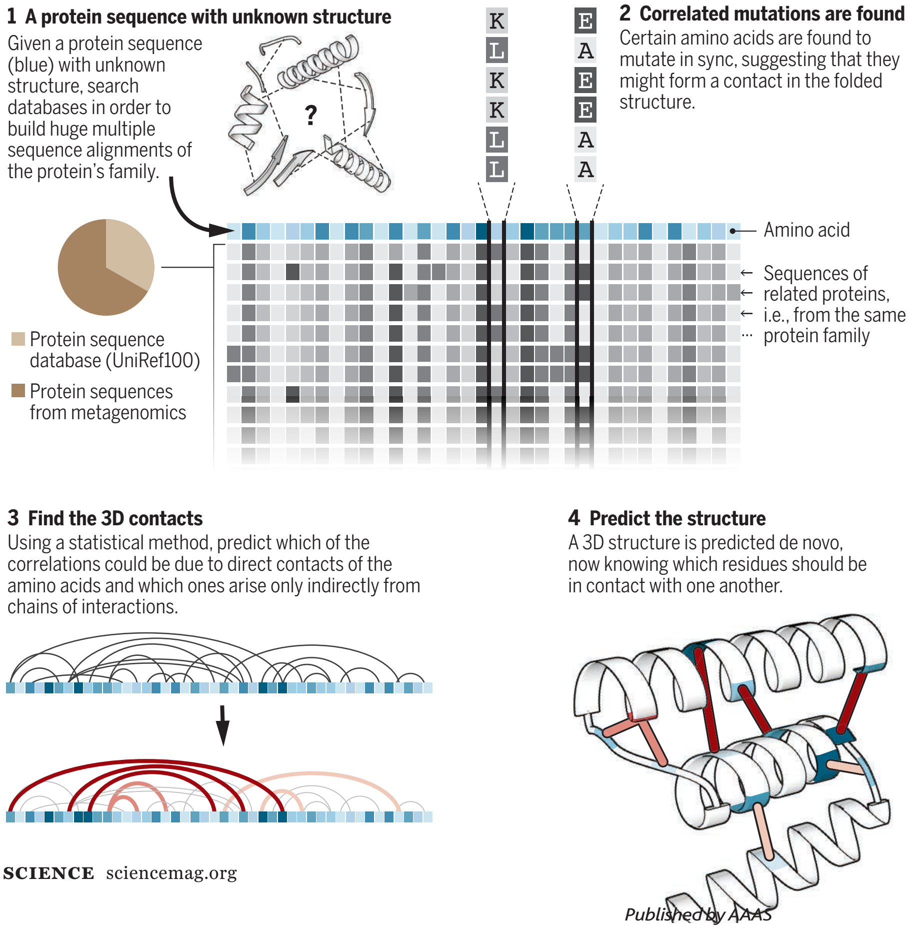
REFERENCES AND NOTES

1. K. A. Dill, J. L. MacCallum, *Science* **338**, 1042 (2012).
2. S. Ovchinnikov *et al.*, *Science* **355**, 294 (2017).
3. L. N. Kinch *et al.*, *Proteins* **84**, 51 (2016).
4. W. Zhang *et al.*, *Proteins* **84**, 76 (2016).
5. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67 (2009).
6. D. S. Marks, *PLOS ONE* **6**, e28766 (2011).
7. T. Nugent, D. T. Jones, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1540 (2012).
8. T. A. Hopf *et al.*, *eLife* **3**, e03430 (2014).
9. R. Finn *et al.*, *Nucleic Acids Res.* **44**, D279 (2016).
10. S. Ovchinnikov *et al.*, *eLife* **4**, e09248 (2015).
11. M. Steinegger, J. Söding, *bioRxiv* 10.1101/079681 (2016).
12. http://predictioncenter.org/casp12/zscores_final.cgi
13. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; E, Glu; K, Lys; and L, Leu.

10.1126/science.aal4512

Structures from sequences

Protein structures are reliably predicted from nothing more than large multiple sequence alignments (*13*).



1 A protein sequence with unknown structure
Given a protein sequence (blue) with unknown structure, search databases in order to build huge multiple sequence alignments of the protein's family.

2 Correlated mutations are found
Certain amino acids are found to mutate in sync, suggesting that they might form a contact in the folded structure.

3 Find the 3D contacts
Using a statistical method, predict which of the correlations could be due to direct contacts of the amino acids and which ones arise only indirectly from chains of interactions.

4 Predict the structure
A 3D structure is predicted de novo, now knowing which residues should be in contact with one another.

?
K
L
K
K
L
L
E
A
E
E
A
A
Amino acid
Sequences of related proteins, i.e., from the same protein family
Protein sequence database (UniRef100)
Protein sequences from metagenomics

GRAPHIC: V. ALTOUNIAN/*SCIENCE*

Chromosomal chaos silences immune surveillance

A high level of chromosomal structural abnormality can suppress the immune response to tumor cells

By **Maurizio Zanetti**

Not all cancers, and not all individuals with the same cancer type, respond equally to immunotherapy—the use of antibodies to block so-called immune checkpoints in T cells—thereby unleashing immune responses against tumor cells. This can be partially explained by nonsynonymous mutations, which can create neoantigen epitopes that induce T cell responses against cancer cells (*1*). However, such mutations scattered throughout the genome may or may not activate the immune system, and if they do, their effect wanes over time. Is there a role for other genomic abnormalities of cancer cells in immune surveillance beyond the generation of neoantigens? On page 261 of this issue, Davoli *et al.* (*2*) propose that structural abnormalities in chromosomes, including variation in the number of chromosome copies (aneuploidy), adversely affect immune cell action against the tumor.

In human cells, aneuploidy is present at some level in 90% of solid tumors and 50% of blood cancers (*3*). Somatic copy-number alterations (SCNAs) arise when sections of the genome are repeated. One-quarter of the genome of a typical cancer cell is affected by whole-arm or whole-chromosome SCNAs, and 10% is affected by focal SCNAs (*3*).

Davoli *et al.* examined data from 5255 tumor and normal samples from 12 cancer types from The Cancer Genome Atlas, distinguishing three types of SCNA (focal, arm, and chromosome). They found that high-level aneuploidy (i.e., higher than 70th percentile SCNA score, which included all

The Laboratory of Immunology, Department of Medicine and Moores Cancer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0815, USA. Email: mzanetti@ucsd.edu

Big-data approaches to protein structure prediction
Johannes Söding (January 19, 2017)
Science **355** (6322), 248-249. [doi: 10.1126/science.aal4512]

Editor's Summary

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2016 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.