STRUCTURAL BIOLOGY

Mutagenesis-based protein structure determination

Genetics has played a key role in understanding the relationship between the DNA sequence encoding a protein and the protein's three-dimensional structure. Two new studies present similar analytical approaches to predict three-dimensional structure on the basis of genetic interaction data.

Melissa Chiasson and Douglas M. Fowler

Recent advances in multiplex assays have enabled measurement of the effects of hundreds of thousands of protein mutants[1,2]. New work from the Lehner[3] and Marks[4] groups leverages these large-scale mutagenesis datasets to infer protein structures by taking advantage of unexpected functional effects when alterations occur near each other in the folded structure of the protein.

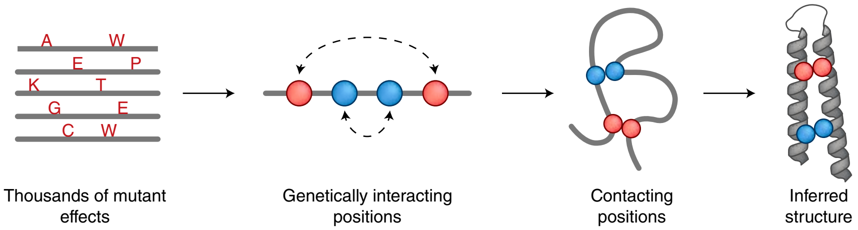
A
W
E
P
K
T
G
E
C
W
Thousands of mutant effects
Genetically interacting positions
Contacting positions
Inferred structure

Fig. 1 | Genetic interactions, determined from large-scale mutagenesis data, reveal protein structure.

Homologous sequences reveal structure

Resolving a protein's structure is necessary to understand how the protein functions and can also facilitate drug design or protein engineering. Experimental tools for determining structure, such as X-ray crystallography, nuclear magnetic resonance spectroscopy or cryo-electron microscopy, require purifying large quantities of the protein. This task is usually difficult and often impossible, particularly for proteins that form complexes or interact with membranes.

An alternative way of learning about a protein's structure is determining how changes in its sequence affect its function. In one approach, examination of homologous sequences can reveal the amino acids that are favored or disfavored at each position. These patterns of conservation can be analyzed between pairs of positions. Positions that are in contact in the folded structure of the protein tend to have co-constrained conservation patterns[5]. The identity of co-constrained and thus likely contacting pairs of positions can be used to produce an accurate model of a protein's structure[6–8]. This powerful approach is limited by the requirement for a large number of homologous sequences: approximately 70% of protein domains of unknown structure lack a sufficient number of sequences[4].

Experimental mutagenesis now also reveals structure

Building on prior work involving a smaller number of mutations[9], the Lehner and Marks groups used the functional effects of thousands of mutations, measured alone and in combination, to infer protein structure. The central idea of the approach depends on using the mutagenesis data to identify genetic interactions (Fig. 1). Two mutations interact genetically if they have an unexpectedly large effect when combined, relative to their individual effects[10]. Both groups primarily analyzed a remarkable dataset comprising the effects of all possible single mutants and nearly all 555,940 possible double mutants of the 56-amino-acid GB1 domain of *Streptococcus* protein G on binding to human IgG[11]. To begin, both groups took advantage of the distinct patterns of genetic interactions that occur in α-helices and β-sheets to predict the secondary structure for each position in the GB1 domain. Next, each group used the mutagenesis data to identify pairs of positions likely to be in contact. Here, the Marks group considered pairs of positions with the strongest positive genetic interactions to be in contact. The Lehner group used a more complex procedure that incorporated all the genetic interactions, both positive and negative, at each position to produce a composite genetic interaction score. These composite scores were then used to determine positions in contact. Both groups used standard algorithms to produce a structural model based on their proposed secondary-structure assignments and positions in contact. The results were remarkable: in each case, the mutagenesis-based GB1 structure was within a few angstroms of structures determined through traditional methods.

Importantly, both groups demonstrated that the approach is generalizable. For example, the authors analyzed less comprehensive mutagenesis data containing ~4% of all possible double mutants for the hYAP65 WW domain[12]. Despite lacking most double-mutant effects, the proposed mutagenesis-based structures agreed well with structures determined through traditional methods. Previously, the Lehner group measured the effects of ~33% of all possible double mutants in *trans* for the leucine-zipper domains of Fos and Jun, which interact in driving transcription[13]. The Lehner group showed that the genetic interactions derived from these data were enriched between contacting positions, and the Marks group used the data to dock Fos and Jun monomers and derive an accurate structure of the complex. Thus, both groups establish large-scale, experimentally derived genetic interactions as a viable and general way to identify contacting positions and thus propose protein structures.

Will the approach scale?

One important caveat is that both groups focused on the ~50-amino-acid GB1 domain and performed nearly exhaustive sampling of the double mutants. However, producing such comprehensive double-mutant data is presently difficult for small proteins (the GB1 dataset is the only one of its kind) and would be impossible for typical proteins with tens of millions of possible double mutants. Thus, both groups analyzed the performance of their approach by down-sampling the GB1 dataset. Contacting positions inferred from a random subset of 10% of the GB1 double-mutant data were more likely to be erroneous than contacting positions inferred by using all the data. These lower quality contacting positions could still be used to infer structure, albeit with lower accuracy. Interestingly, the Marks group showed that using just 5% of the data, provided that only deleterious mutations were sampled, yielded a performance equivalent to that of using the full dataset. Thus, mutagenesis-based structure determination can in principle be used on larger proteins, especially if deleterious mutations can be tested. However, more work is required to fully understand how much mutagenesis data will be required to solve large, complex protein structures.

The future

The Lehner and Marks groups elucidate how large-scale mutagenesis data, generated without purifying the protein, can be used to solve protein structures. Techniques for generating large-scale mutagenesis data are improving rapidly, and because these methods are being applied broadly, there is hope that the approach can be used to attack the myriad proteins that currently lack structures. To date, mutagenesis-based structure determination has been demonstrated by using only proteins whose structures were already known. Now, mutagenesis-based structures must be generated for one or more proteins whose structures are unknown. Ideally, the same proteins will be analyzed with traditional methods, and the results will be compared.

Finally, large-scale mutagenesis data can be gathered from a protein in its native context, thus suggesting that future analyses could go beyond static structure determination. For example, the Marks group reported that false-positive genetic interactions (for example, those between distant positions) tend to involve binding or active sites. After a structure has been determined, these strongly interacting, distant positions might reveal important aspects of the protein's function. Moreover, by varying the conditions under which the mutagenesis data are collected (for example, with or without ligand), understanding how a protein's structure changes in response to stimuli may be possible. Regardless, mutagenesis may be a promising tool for understanding protein structure.

Melissa Chiasson[1] and
Douglas M. Fowler[1,2,3*]
[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Department of Bioengineering, University of Washington, Seattle, WA, USA. [3]Genetic Networks Program, CIFAR, Toronto, Ontario, Canada.
*e-mail: dfowler@uw.edu

Published online: 28 June 2019
https://doi.org/10.1038/s41588-019-0461-4

References

1. Fowler, D. M. & Fields, S. *Nat. Methods* **11**, 801–807 (2014).
2. Gasperini, M., Starita, L. & Shendure, J. *Nat. Protoc.* **11**, 1782–1787 (2016).
3. Schmiedel, J. M. & Lehner, B. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0431-x (2019).
4. Rollins, N. J. et al. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0432-9 (2019).
5. Altschuh, D., Lesk, A. M., Bloomer, A. C. & Klug, A. *J. Mol. Biol.* **193**, 693–707 (1987).
6. Miller, C. S. & Eisenberg, D. *Bioinformatics* **24**, 1575–1582 (2008).
7. Marks, D. S., Hopf, T. A. & Sander, C. *Nat. Biotechnol.* **30**, 1072–1080 (2012).
8. Ovchinnikov, S. et al. *eLife* **4**, e09248 (2015).
9. Sahoo, A., Khare, S., Devanarayanan, S., Jain, P. & Varadarajan, R. *eLife* **4**, e09532 (2015).
10. Baryshnikova, A., Costanzo, M., Myers, C. L., Andrews, B. & Boone, C. *Annu. Rev. Genom. Hum. Genet.* **14**, 111–133 (2013).
11. Olson, C. A., Wu, N. C. & Sun, R. *Curr. Biol.* **24**, 2643–2651 (2014).
12. Araya, C. L. et al. *Proc. Natl Acad. Sci. USA* **109**, 16858–16863 (2012).
13. Diss, G. & Lehner, B. *eLife* **7**, e32472 (2018).

Competing interests

The authors declare no competing interests.

HUMAN DISEASE

Priority index for human genetics and drug discovery

Although human genetics can help identify new drug targets, the best way to prioritize genes as therapeutic targets is uncertain. A new study describes a framework to prioritize potential targets by integrating genome-wide association data with genomic features, disease ontologies and network connectivity.

Robert M. Plenge

Human genetics offers the potential to identify new therapeutic targets and prioritize decision-making throughout the process of drug discovery and development[1,2]. However, strategies to prioritize genes as therapeutic targets remain incomplete. In this issue, Fang et al.[3] describe a framework to prioritize potential targets for therapeutic intervention according to the integration of genome-wide association study (GWAS) data with genomic features, disease ontologies and network connectivity. Although their genetics-led drug-target-prioritization approach is focused on immunologically mediated traits, the framework should also be applicable to non-immunologically mediated diseases.

Two approaches

There are two general approaches to prioritize genes from human genetic studies as therapeutic targets (Fig. 1). The first is a gene-centric approach. One model leverages trait-associated alleles to estimate dose–response curves[1]. In that model, the trait-associated alleles could come from common-variant association studies, rare-variant association studies or studies of rare Mendelian phenotypes. For common diseases, examples of the 'allelic-series' model include *PCSK9* (coronary artery disease) and *TYK2* (immunologically mediated diseases), and for rare diseases,