Taking Structure Searches to the Next Dimension

ARTICLE *in* STRUCTURE · JULY 2014

Impact Factor: 5.62 · DOI: 10.1016/j.str.2014.06.007 · Source: PubMed

2 AUTHORS:

Andrea Schafferhans

Technische Universität München

17 PUBLICATIONS **218** CITATIONS

SEE PROFILE

Burkhard Rost

Technische Universität München

279 PUBLICATIONS **25,286** CITATIONS

SEE PROFILE

Taking Structure Searches to the Next Dimension

Andrea Schafferhans[1,*] and Burkhard Rost[1]
[1]Department of Informatics, Bioinformatics-I12, TUM, Boltzmannstrasse 3, 85748 Garching/Munich, Germany
*Correspondence: andrea.schafferhans@rostlab.org
http://dx.doi.org/10.1016/j.str.2014.06.007

Structure comparisons are now the first step when a new experimental high-resolution protein structure has been determined. In this issue of *Structure*, Wiederstein and colleagues describe their latest tool for comparing structures, which gives us the unprecedented power to discover crucial structural connections between whole complexes of proteins in the full structural database in real time.

Comparing protein domains is state-of-the-art in structure comparison. It is widely thought that protein structure determines protein function (Redfern et al., 2008). However, one early structural genomics discovery by the Protein Structure Initiative (PSI) at the NIH/National Institute of General Medical Sciences was that structure comparisons could boost the impact of each experimental structure (Bertonati et al., 2009). Details about protein function become apparent through highlighting differences and similarities in protein 3D structures. Several databases offer precomputed structure comparisons. For instance, CATH (Sillitoe et al., 2013), SCOP (now SCOP2; Andreeva et al., 2014), VAST+ (Madej et al., 2014), or COPS (Suhrer et al., 2009) group proteins into families related in structure. The typical unit for classifications of those resources and their underlying programs are domains, that is, the compact constituents of proteins assumed to fold independently and to represent the "building blocks" relevant for function and evolution. In fact, many resources have been dedicated to improving the split of large protein complexes into their domain constituents (Sillitoe et al., 2013), and one of the major differences between the proteins of known structure and sequences of unknown structure is in their domain composition. Proteins with one single domain make up 20% of all known proteins, but over 70% of all structures have a single domain (Liu and Rost, 2004; Sillitoe et al., 2013). Although there are many good reasons to first split and then compare, such structural comparisons of the building blocks completely miss important relations that become visible only through the assembly of domains into large units or molecular machines through which proteins act.

Comparing protein complexes is more complex. With an increasing number of experimental high resolution structures of complexes, the time has come for a tool that breaks through and compares complexes in their full size and glory directly. Switching from *comparing constituents* to *matching machines*, methods need to relax stringency with careful tuning. TopMatch (Sippl and Wiederstein, 2012) realized this task by building less stringent composite global alignments from more stringent local alignments. Empirical ranking circumvented the explosion of possible combinations and ascertained acceptable computing times. A genially simple and powerful similarity metric made this possible (Sippl, 2008; Sippl and Wiederstein, 2008). The resulting metric score for similarity implicitly leaped from measuring 3D similarity to estimating 3D distance.

TopSearch efficiently reduces complexity to gain speed. In this issue of *Structure*, Wiederstein et al. (2014) introduce an essential improvement of their methodology. Queried by coordinate sets for entire complexes, the newly enhanced tool, TopSearch, quickly unravels macromolecular relations in the entire Protein Data Bank (PDB). The construction and maintenance of a continuously updated local resource with a representative subset for all biological assemblies in the PDB combined with the similarity metric introduced ascertain speed and precision. The efficient matching considers all biologically relevant alternative assemblies in all known 3D structures. TopSearch also maintains precomputed links among all related macromolecular complexes in the PDB. Visualization of structural matches is extremely fast and can easily handle complexes as large as entire ribosomes. Figure 1 shows

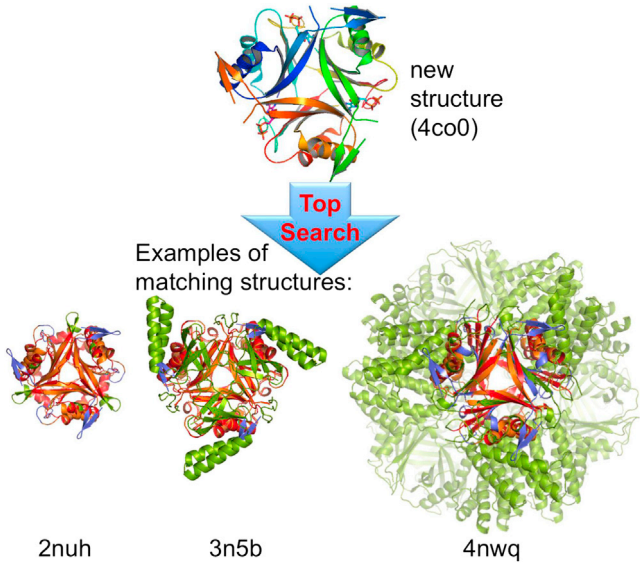

Figure 1. TopSearch Reveals Structural Connections between Whole Complexes of Proteins in the Full Structural Database
The figure shows an example query structure, trimeric nitrogen signal transduction PII protein GlnZ from *Azospirillum brasilense*, released in June 2014 (PDB code 4CO0), and some of its structural matches: a protein of unknown function with 10% sequence identity (PDB code 2NUH); nitrogen regulatory protein P-II from Nostoc in complex with an interacting protein (PDB code 3N5B); and a designed hetero 24-mer that shares a similar interaction pattern (PDB code 4NWQ).

an example of a recently released trimeic PDB structure and some of the related complexes found by TopSearch.

Unraveling relations between complexes gives new insights. Two teasers demonstrate the power of the new way to trigger discoveries in the wealth of high-resolution data about protein structures (Wiederstein et al., 2014). The first teaser presents a protein of yet unknown function (PDB code 2GJV), the structure of which was determined by the PSI. This protein could assemble into two alternative biological assemblies. TopSearch shows one of these to be significantly similar to secretory proteins, thereby addressing two questions: first, as to which assembly is most biologically relevant, and second, as to what the possible function is. By using a bacterial DNA clamp to search against the structure database, Wiederstein et al. (2014) can explore the structural variations and constants found in this universal class of proteins. The same overall ring structure is dimeric in bacteria, trimeric in eukaryotes, and tetrameric in a viral structure.

TopSearch opens new horizons. The TopSearch resource could also be the beginning of new studies that comprehensively analyze in detail how the interfaces can vary between related complexes as a function of decreasing sequence relation and decreasing functional similarity. This may, in turn, lead to new insights that aid the effort of unraveling interaction networks. Predictions of biological assemblies might best begin with the TopSearch resources from now on. TopSearch could also benefit the study of evolutionary relations; comparing the 3D structures of domains allows much better inferences of evolutionary relations than sequences alone do. In analogy, we expect that matching entire machines (instead of isolated domains) will bring about new insights into protein evolution, structure, and function.

REFERENCES

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A.G. (2014). Nucleic Acids Res. *42*, D310–D314.

Bertonati, C., Punta, M., Fischer, M., Yachdav, G., Forouhar, F., Zhou, W., Kuzin, A.P., Seetharaman, J., Abashidze, M., Ramelot, T.A., et al. (2009). Proteins *75*, 760–773.

Liu, J., and Rost, B. (2004). Proteins *55*, 678–688.

Madej, T., Lanczycki, C.J., Zhang, D., Thiessen, P.A., Geer, R.C., Marchler-Bauer, A., and Bryant, S.H. (2014). Nucleic Acids Res. *42*, D297–D303.

Redfern, O.C., Dessailly, B., and Orengo, C.A. (2008). Curr. Opin. Struct. Biol. *18*, 394–402.

Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R., et al. (2013). Nucleic Acids Res. *41*, D490–D498.

Sippl, M.J. (2008). Bioinformatics *24*, 872–873.

Sippl, M.J., and Wiederstein, M. (2008). Bioinformatics *24*, 426–427.

Sippl, M.J., and Wiederstein, M. (2012). Structure *20*, 718–728.

Suhrer, S.J., Wiederstein, M., Gruber, M., and Sippl, M.J. (2009). Nucleic Acids Res. *37*, W539–W544.

Wiederstein, M., Gruber, M., Frank, K., Melo, F., and Sippl, M.J. (2014). Structure *22*, this issue, 1063–1070.