e-letter for COMPLEAT database

<http://stke.sciencemag.org/about/help/letters>

<http://www.sciencemag.org/site/misc/editor.xhtml>

<http://www.sciencemag.org/site/collections/online/eletters/guidelines.xhtml>

The COMPLEAT (Complex Enrichment Analysis Tool) database introduced in the article (Vinayagam et al., 2013) is the first of it’ s kind. It contains the protein complexes of three organisms, human, fruit fly and yeast, and provides tools for annotation of high-throughput data sets (e.g RNAi screens, proteomics, gene expression) as well as annotation results visualization. The incorporated complexes of COMPLEAT are derived from other databases with manually curated data and individual experiments as well as bioinformatics prediction tools. The tools used for the prediction of protein complexes are CFinder and NetworkBlast (Kalaev, Smoot, Ideker, & Sharan, 2008).

In this e-letter, we want to indicate that the tool NetworkBlast predicted protein complexes with sizes that range from 3 to 15 proteins for yeast and fruit fly (for human these data are not provided). This is in contrary with the complexes mentioned in the literature as well as predicted with CFinder. In the following section we summarize the sizes of complexes, in terms of number of proteins, and the respective resource.

Fruit fly

NetworkBlast: 2893 complexes, size range: 3 to 15 proteins

CFinder: 389 complexes, size range: 3 to 189 proteins

Literature: 2045 complexes, size range: 1 to 175 proteins

Total: 5327 complexes (different from the article numbers)

Yeast

NetworkBlast: 2863 complexes, size range: 3 to 15 proteins

CFinder: 223 complexes, size range: 3 to 189 proteins

Literature: 1044 complexes, size range: 1 to 175 proteins

Total: 4130 complexes (different from the article numbers)

Human

No such data provided from COMPLEAT

The fact that around half of the protein complexes are predicted from NetworkBlast creates a bias towards small protein complexes. This is noticeable as a “break” in the distribution of complexes sizes. It can be easily confirmed that is caused by the complexes from NetworkBlast because the distributions of the sizes of complexes from the literature and CFinder follow a power-law-like distribution with no such “break”.

Further investigation is needed to determine if the upper limit of protein complex size is an inherent bias of NetworkBlast or is caused by the its implementation in the COMPLEAT database.

In either case, we believe that the users of the COMPLEAT database need to be informed of the fact that there is a bias towards small complexes that is caused by the predicted complexes of NetworkBlast.

**Names and affiliations**

**References**

Kalaev, M., Smoot, M., Ideker, T., & Sharan, R. (2008). NetworkBLAST: Comparative analysis of protein networks. *Bioinformatics*, *24*(4), 594–596. https://doi.org/10.1093/bioinformatics/btm630

Vinayagam, A., Hu, Y., Kulkarni, M., Roesel, C., Sopko, R., Mohr, S. E., & Perrimon, N. (2013). Protein complex-based analysis framework for high-throughput data sets. *Sci Signal*, *6*(264), rs5. https://doi.org/10.1126/scisignal.2003629

**Conflict of interest:**

None declared