Connecting TOPSAN to Computational Analysis

Christian M Zmasek[2,5], Kyle Ellrott[3], Dana Weekes[1,2], Constantina Bakolitsa[1,2], John Wooley[3], Adam Godzik[1,2,3,4]

[1] Joint Center for Structural Genomics, http://www.jcsg.org/
[2] Sanford-Burnham Medical Research Institute, La Jolla, California, USA
[3] University of California, San Diego, La Jolla, California, USA
[4] Joint Center for Molecular Modeling, http://jcmm.burnham.org/
[5] czmasek@burnham.org

Project Site: http://www.topsan.org
Software: http://www.topsan.org/Tools
Open Source Licenses: Creative Commons Attribution 3.0 License (data)
GNU General Public License (software)

The National Institute of Health's Protein Structure Initiative (PSI) has produced over four thousand protein structures. As a member of that initiative, the Joint Center of Structural Genomics (JCSG) uses TOPSAN to organize the annotations of these proteins. TOPSAN, the wiki based protein annotation system, sits at an important nexus between computationally generated protein annotations and human expertise. The information in TOPSAN is both the product of and a source to biological computational analysis. Unlike Protopedia, the goal of which is to catalog established literature about proteins in Wiki form, TOPSAN's primary purpose is to promote rapid theoretical discussion about protein structures.

TOPSAN has successfully provided a link from the library of protein information generated by the JCSG program to human curators. On the other hand, it is also important to close the loop and bring that human expertise to feed back into the database of information that JCSG is producing. To this end, we are attempting to draw more links from protein descriptions texts written in TOPSAN to the organized databases that are used to annotate proteins. One example of this is the effort to map TOPSAN proteins to standard ontologies, such as the Gene Onotology (GO) classifications. GO terms have been used to link protein families into large hierarchical groups related by function.

GO annotation will be done manually by the curators. In order to facilitate this effort we are beginning to provide automatic suggestions that the curators can review and approve. The first stage of this is to use the existing Pfam to GO mappings to take care of trivial cases. In situations where there are no existing mappings, we use automatic text scanning to generate suggestions for the user. Using published literature, GO terms will be suggested using standard indexing/searching strategies.

There is also a need to connect the work being done by the annotators to other sources of data. By linking annotations to semantic web compliant databases, TOPSAN becomes connected to a larger set of databases. This is done via a simple in-line notation embedded in the text of the wiki article, that is then automatically exported to RDFa and other semantic web compliant technologies.

More importantly, semantic web technologies will also allow TOPSAN to be interrogated computationally. A piece of software could, for example, easily extract a list of all proteins from a particular organism and/or of a given function. The program could use described protein interaction links to extract all available information about a pathway and then perform further analysis on the data delivered by TOPSAN.

The wiki platform provides a system that allows for rapid updates to data, while the semantic web enables a system for the representation of data in a way that is both easy to extract and malleable enough to allow a variety of new data sources to be integrated.