D494–D496 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 20 October 2010*
doi:10.1093/nar/gkq902

TOPSAN: a dynamic web database for structural genomics

Kyle Ellrott[1,2], Christian M. Zmasek[3,4], Dana Weekes[1,4], S. Sri Krishna[2,4], Constantina Bakolitsa[1,4], Adam Godzik[1,2,3,4,*] and John Wooley[1,2,4,*]

[1]Joint Center for Structural Genomics, Bioinformatics Core, Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, [2]Center for Research in Biological Systems, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, [3]Joint Center for Molecular Modeling and [4]Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

Received August 16, 2010; Revised September 21, 2010; Accepted September 22, 2010

ABSTRACT

The Open Protein Structure Annotation Network (TOPSAN) is a web-based collaboration platform for exploring and annotating structures determined by structural genomics efforts. Characterization of those structures presents a challenge since the majority of the proteins themselves have not yet been characterized. Responding to this challenge, the TOPSAN platform facilitates collaborative annotation and investigation via a user-friendly web-based interface pre-populated with automatically generated information. Semantic web technologies expand and enrich TOPSAN's content through links to larger sets of related databases, and thus, enable data integration from disparate sources and data mining via conventional query languages. TOPSAN can be found at http://www.topsan.org.

INTRODUCTION

Over the past decade, structural genomics (SG) efforts in the USA alone have determined the structures of more than 3000 previously uncharacterized proteins at a sustained rate of over 500 novel structure depositions per year to the Protein Databank (PDB) (1). Through the discovery of numerous new folds and an even greater number of variants of known folds (2), SG structures provide key input for innovative research into protein evolution and function. One of the main challenges presented by such high-throughput research involves the timely annotation and integration of the resulting data to provide direct input into ongoing research within the greater biological community. Traditional mechanisms for publication are simply too slow to keep pace with the speed of structure determination. Thus, currently over 90% of SG deposited structures are not yet described in literature. The rate and volume of protein structures being produced requires novel mechanisms to ensure that the knowledge gained by these structures is disseminated in a timely manner.

Several new protein structure annotation platforms, using wiki-based methods, have been described (3–5). However, their content is largely static and derived from peer-reviewed publications, aspects that do not easily lend themselves to exploring new knowledge about structures. We developed The Open Protein Structure Annotation Network (TOPSAN) to serve both as an annotation and a communication platform with the goal of facilitating and accelerating research relevant to SG structures. TOPSAN integrates a wide range of information about SG proteins, from different high-throughput experiments to literature, evolutionary analysis and even functional predictions. Through the implementation of a semantic web layer in the current version, TOPSAN enables database-like searches through its entire content and thus promotes further integration between its content and mainstream biology.

THE DATABASE

Content and interface

TOPSAN currently contains annotations for over 7250 structures from SG efforts from around the world. Prominent among these are several hundred structures that represent the first experimentally characterized members of their respective families, as well as many proteins for which there is extensive interest within the research community. Annotations and collaborations developed via TOPSAN have led to peer-reviewed publications for several dozen proteins, with many more currently

*To whom correspondence should be addressed. Tel: 858 646 3168; Fax: 858 795 5249; Email: adam@sanfordburnham.org
Correspondence may also be addressed to John Wooley. Tel: 858 534 2494; Fax: 858 822 1452; Email: jwooley@ucsd.edu

in different stages of development. An overview of the database interface is given in Figure 1.

Implementation

Implementation of the TOPSAN platform has been described in detail elsewhere (6). In brief, TOPSAN was developed using MindTouch, an enterprise open source collaboration and integration platform, which provides tools and scripting capabilities that were used to develop a dynamic website. At the backend, data are collected from a variety of different sources, using multiple tools that have been integrated into the MindTouch platform. An application called TopsanApp is used to retrieve

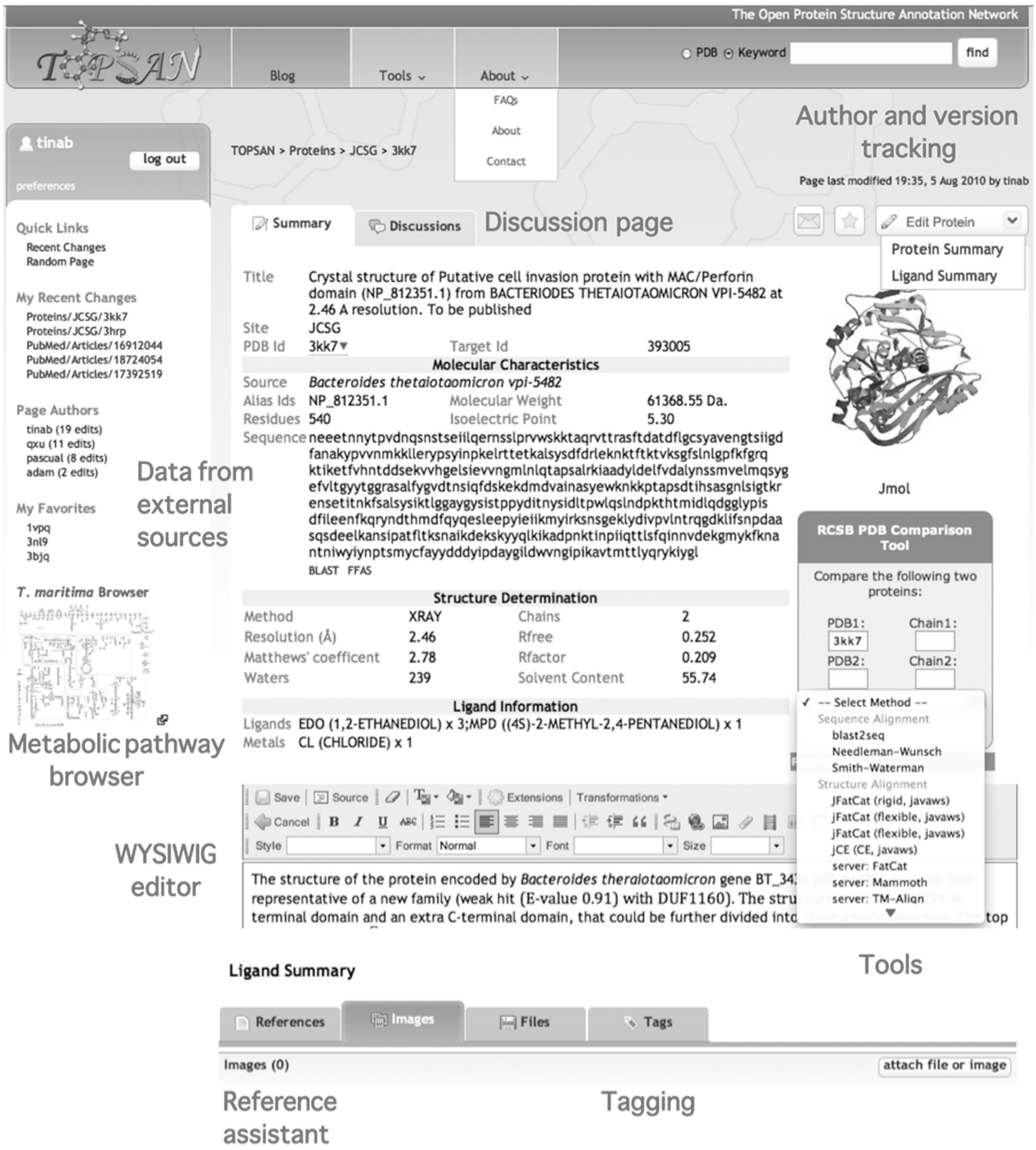

Figure 1. Screenshot of a TOPSAN entry [PDB id: 3kk7 (9)]. Automatically generated data (data from external sources) are combined with human input (WYSIWIG editor, tagging, discussion page) and analyses (tools). Authorship and version tracking enable accountability and quality control. Semantic web technologies enable easy import and export of this data. Details of the data specifications can be found on the website.

protein information from external resources and to create and store pages for specified proteins on the platform via an API. Data collected by TopsanApp are stored in a local MySQL database (termed topsanDB) that is used to generate individual protein pages with built-in functions for easy access and manipulation of information.

TOPSAN additionally utilizes a semantic web-based data import system that enables rapid integration of new data. Semantic web is an architectural layer built on top of existing web pages that consists of hidden embedded tags that employ standardized ontology, allowing searches normally associated with structured databases to be carried out across unstructured data collections (7,8). For our purposes, the semantic web provides a unified framework for integration of data automatically imported from external sources and human-curated annotations. In this environment, scripting calls made from the Dekiscript environment build requests to access and convert XML formatted data available on the web into a semantic web compatible format. Data from compatible sites that provide all records in a semantic web compatible format can be imported directly with no manual conversion. Thus, semantic web data can be automatically imported to TOPSAN from Pfam, UniProt, KEGG Pathway database and PDB. Other sources of data can be imported with a variety of modular, easily adapted plug-ins. Once imported, the data can be queried and utilized in the same web-based Dekiscript environment that was used for the import. Data export is handled by a variety of modular tools that make data available in formats including standard HTML, stripped-down XML, RDF/XML and RDFa. In addition to individual protein annotations, bulk compilations of the entire site are available as compressed files. TOPSAN also provides embeddable web interfaces to help other websites, such as Pfam, integrate TOPSAN annotations. Annotations can be viewed by anyone, but only registered users are enabled to contribute text. Accountability and ownership of ideas is preserved via time-stamped tracking of contributions.

Conclusions and future perspectives

TOPSAN explores an important nexus between human analysis and computational data mining, neither of which can independently handle the challenges of research in a high-throughput data generation era. Future TOPSAN improvements include developing statistical methods for determining the reliability of information extracted from databases and testing means for improving semantic functionality. Additionally, TOPSAN emphasizes user-friendly access to external resources and databases, which might otherwise be unknown or not easy to access for some users. Javascript-based widgets allow users to view and edit annotations while capturing and storing their input in a format that is compatible with the semantic web.

ACKNOWLEDGEMENTS

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

FUNDING

National Institutes of Health National Institute of General Medical Sciences Protein Structure Initiative (grant No. U54 GM074898). Funding for open access charge: University of California San Diego.

Conflict of interest statement. None declared.

REFERENCES

1. Norvell,J.C. and Berg,J.M. (2007) Update on the protein structure initiative. *Structure*, **15**, 1519–1522.
2. Levitt,M. (2007) Growth of novel protein structural data. *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.
3. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F. Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
4. Hodis,E., Prilusky,J., Martz,E., Silman,I., Moult,J. and Sussman,J.L. (2008) Proteopedia—a scientific 'wiki' bridging the rift between three-dimensional structure and function of biomacromolecules. *Genome Biol.*, **9**, R121.
5. Stehr,H., Duarte,J.M., Lappe,M., Bhak,J. and Bolser,D.M. (2010) PDBWiki: added value through community annotation of the Protein Data Bank. *Database*, **2010**. http://www.oxfordjournals.org/our_journals/databa/about.html, doi:10.1093/database/baq009.
6. Weekes,D., Krishna,S.S., Bakolitsa,C., Wilson,I.A., Godzik,A. and Wooley,J. (2010) TOPSAN: a collaborative annotation environment for structural genomics (accepted). *BMC Bioinformatics*, **11**, 426.
7. Neumann,E. (2005) A life science Semantic Web: are we there yet? *Sci. Signal. STKE*, **283**, pe22.
8. Antezana,E., Kuiper,M. and Mironov,V. (2009) Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief Bioinform.*, **10**, 392–407.
9. Xu,Q., Abdubek,P., Astakhova,T., Axelrod,H.L., Bakolitsa,C., Cai,X., Carlton,D., Chen,C., Chiu,H.-J., Clayton,T. *et al.* (2010) Structure of a membrane-attack complex/perforin (MACPF) family protein from the human gut symbiont *Bacteroides thetaiotaomicron* (accepted). *Acta Cryst. F*, **F66**, 1297–1305.