

Unambiguous Chemical Database Searching with Chem-BLAST to Facilitate New Drug Development

NIST is developing intuitive and user friendly Web interfaces for organizing and retrieving chemical data, based on chemical structure, not chemical name. This will enable reliable exchange of chemical data over the Web, enabling new innovations in drug discovery and development.

T. N. Bhat and A. D. Nguyen (Div. 831)

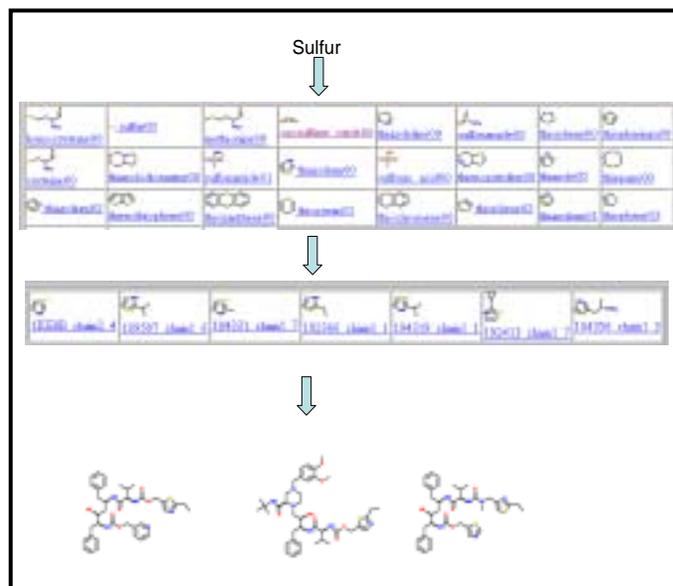
The emergence of a web-based data-rich era in a biochemical world is indisputable. In the past decade, databases have become an integral part of research and development in the biomedical sciences. Bioinformatics plays an essential role in deriving knowledge from complex biochemical data.

A common way of identifying or specifying chemical compounds is by name. IUPAC names are well-known examples. However, these names are difficult to generate, and machine searching for unique matches is slow. For this reason, chemical Web page developers have been relentlessly working on associating compounds with an ever growing list of so-called 'synonyms' to match against user queries. This approach is frustrating both for database providers and users of a Web page alike. A database provider can never be sure that he has included all the 'synonyms' of a compound, and a user can never be assured that a name he is familiar with has the same meaning in the context of the database he is planning to query. The result is often missed hits or an overwhelming number of hits.

NIST has developed a new technology called Chem-BLAST (Chemical Block Layered Alignment of Substructure Technique) that has proven to be highly successful to query therapeutic drugs that can be used to treat AIDS.

Chem-BLAST combines the usefulness of two-dimensional pictures of chemical structures with explicitly defined chemical structural components of interest to the user. Within an environment called a *resource description framework* (RDF), the user can choose the chemical components of the structure of interest from selections of components in the database. These selections are presented in a sequential fashion until the compound of interest is built up from its structural components. This approach is orderly, unambiguous, and efficient. In a typical setting, a

database with 10,000 compounds with four layers of RDF statements may allow a user to get a precise answer to his question in just four mouse clicks.



This figure illustrates a case where a user queries the database for drugs with sulfur containing a five member ring. The Web page displays all the drugs with such an element. In each mouse click the user builds the query from the structural elements presented to him and a total of three mouse clicks are used here.

NIST implemented Chem-BLAST in one of its most popular Web resources, the HIV Structural Database (<http://xpdb.nist.gov/hivpdb/hivpdb.html>).

HIVSDB is a collaborative effort between NIST, National Institute of Allergy and Infectious Diseases, National Cancer Institute, and University of Rutgers. Chem-BLAST and the Semantic Web concept have been embraced by the W3C, the International Body responsible for establishing standards for the future World Wide Web, which sees Chem-BLAST as a way to provide precise answers for complicated questions without compromising speed or convenience for either database developers or users.

Relevant Publication:

Prasanna MD, Vondrasek J, Wlodawer A, Rodriguez H, Bhat TN. **Chemical compound navigator: a web-based chem-BLAST, chemical taxonomy-based search engine for browsing compounds.** *Proteins* 2006;63(4):907-917