

A database of interactions between protein domains and small molecules has been created using data from the Protein Data Bank (PDB). Using a set of fully annotated protein-small molecule interactions from the PDB as a starting point, interactions of the Small Molecule Interaction Database (SMID) were generated by identifying protein domains that bind to small molecules, using NCBI's Reverse Position Specific BLAST (RPS-BLAST) algorithm. Record redundancy was addressed by clustering interactions that involved the same small molecule, same domain, and had a 50% or greater overlap in binding site residues. Through the user-friendly SMID interface, at <http://smid.blueprint.org/>, users may view the interactions using the Cn3D or RasMol programs. The database may be queried with a protein database identifier, small molecule name, PDB ID or SMID ID. The SMID interface also includes the SMID-BLAST tool, which provides accurate transitive annotation of small-molecule binding sites for proteins not found in the PDB. Given a protein sequence, SMID-BLAST identifies domains using RPS-BLAST and then produces a listing of potential small molecule ligands and binding sites on the query sequence based on existing SMID records. A ligand score is calculated for all SMID-BLAST hits in order to assign a level of confidence. SMID-BLAST is also available as a command line tool at <ftp://ftp.blueprint.org/pub/SMID/tool/>. Several examples of the use of this tool include: to provide insight into potential protein drug targets related to human disease; to propose functions of novel unannotated proteins; and to investigate the evolutionary relationship of protein families.

The SMID-BLAST command line tool is written in PHP (www.php.net) and requires a UNIX or Windows machine with PHP 4.3.0 or newer on it. A MySQL server (www.mysql.com; we suggest 4.1.7 or higher) is required as well, though it need not be on the same machine, to house the SMID data used by SMID-BLAST. All other required support tools are downloaded automatically by the software the first time it is run, and are kept up to date automatically as well, as new versions of SMID data, BLAST executables, or RPS-BLAST databases become available. The user may choose to install and maintain these manually if desired. The program takes a FASTA format file of protein sequences as its input. Output of small molecule hits is provided in ASN.1 format, readable by the Sequin program (<http://www.ncbi.nlm.nih.gov/projects/Sequin/>) as well as GenPept format. Further analysis and annotation can then be carried out using the Sequin program, or any GenPept parser. The tool is available at the above location, free of charge for academic use. Commercial use requires a license.