Pacific Symposium on Biocomputing 12:145-156(2007)

BIOSPIDER: A WEB SERVER FOR AUTOMATING METABOLOME ANNOTATIONS

CRAIG KNOX, SAVITA SHRIVASTAVA, PAUL STOTHARD, ROMAN EISNER, DAVID S. WISHART

Department of Computing Science, University of Alberta, Edmonton, AB T6G-2E8 Canada

One of the growing challenges in life science research lies in finding useful, descriptive or quantitative data about newly reported biomolecules (genes, proteins, metabolites and drugs). An even greater challenge is finding information that connects these genes, proteins, drugs or metabolites to each other. Much of this information is scattered through hundreds of different databases, abstracts or books and almost none of it is particularly well integrated. While some efforts are being undertaken at the NCBI and EBI to integrate many different databases together, this still falls short of the goal of having some kind of human-readable synopsis that summarizes the state of knowledge about a given biomolecule – especially small molecules. To address this shortfall, we have developed BioSpider. BioSpider is essentially an automated report generator designed specifically to tabulate and summarize data on biomolecules – both large and small. Specifically, BioSpider allows users to type in almost any kind of biological or chemical identifier (protein/gene name, sequence, accession number, chemical name, brand name, SMILES string, InCHI string, CAS number, etc.) and it returns an in-depth synoptic report (~3-30 pages in length) about that biomolecule and any other biomolecule it may target. This summary includes physico-chemical parameters, images, models, data files, descriptions and predictions concerning the query molecule. BioSpider uses a web-crawler to scan through dozens of public databases and employs a variety of specially developed text mining tools and locally developed prediction tools to find, extract and assemble data for its reports. Because of its breadth, depth and comprehensiveness, we believe BioSpider will prove to be a particularly valuable tool for researchers in metabolomics. BioSpider is available at: **www.biospider.ca**

1. Introduction

Over the past decade we have experienced an explosion in the breadth and depth of information available, through the internet, on biomolecules. From protein databases such as the PDB [1] and Swiss-Prot [18] to small molecule databases such as PubChem (http://pubchem.ncbi.nlm.nih.gov/), KEGG [2], and ChEBI (http://www.ebi.ac.uk/chebi/), the internet is awash in valuable chemical and biological data. Unfortunately, despite the abundance of this data, there is still a need for new tools and databases to connect chemical data (small, biologically active molecules such as drugs and metabolites) to biological data (biologically active targets such as proteins, RNA and DNA), and vice versa. Without this linkage clinically important or pharmaceutically relevant information is often lost. To address

this issue we have developed an integrated cheminformatics/bioinformatics reporting system called BioSpider. Specifically, BioSpider is a web-based search tool that was created to scan the web and to automatically find, extract and assemble quantitative data about small molecules (drugs and metabolites) and their large molecule targets. BioSpider can be used as both a research tool or it can be used as a database annotation tool to assemble fully integrated drug, metabolites or protein databases.

So far as we are aware, BioSpider appears to be a unique application. It is essentially a hybrid of a web-based genome annotation tool, such as BASYS [3] and a text mining system such as MedMiner [4]. Text mining tools such as MedMiner, iHOP [5], MedGene [6] and LitMiner [7] exploit the information contained within the PubMed database. These web servers also support more sophisticated text and phrase searching, phrase selection and relevance filtering using specially built synonym lists and thesauruses. However, these text mining tools were designed specifically to extract information only from PubMed abstracts as opposed to other database resources. In other words MedMiner, MedGene and iHOP do not search, display, integrate or link to external molecular database information (i.e. GenBank, OMIM [8], PDB, SwissProt, PharmGKB [9], DrugBank [10], PubChem, etc.) or to other data on the web. This database or web-based information-extraction feature is what is unique about BioSpider.

2. Application Description

2.1. Functionality

Fundamentally, BioSpider is highly sophisticated web spider or web crawler. Spiders are software tools that browse the web in an automated manner and keep copies of the relevant information of the visited pages in their databases. However, BioSpider is more than just a web spider. It is also an interactive text mining tool that contains several predictive bioinformatic and cheminformatic programs, all of which are available through a simple and intuitive web interface. Typically a BioSpider session involves a user submitting a query about one or more biological molecules of interest through its web interface, waiting a few minutes and then viewing the results in a synoptic table. This hyperlinked table typically contains more than 80 data fields covering all aspects of the physico-chemical, biochemical, genetic and physiological information about the query compound. Users may query BioSpider with either small molecules (drugs or metabolites) or large molecules (human proteins). The queries can be in almost any form, including chemical names, CAS numbers, SMILES strings [11], INCHI identifiers, MOL files or Pubchem IDs (for small molecules), or protein names and/or Swiss-Prot IDs (for macromolecules). In extracting the data and assembling its tabular reports BioSpider employs several robust data-gathering techniques based on screen-scraping, text-

mining, and various modeling or predictive algorithms. If a BioSpider query is made for a small molecule, the program will perform a three-stage search involving: 1) Compound Annotation; 2) Target Protein/Enzyme Prediction and 3) Target Protein/Enzyme Annotation (see below for more details). If a BioSpider query is made for a large molecule (a protein), the program will perform a complete protein annotation. BioSpider always follows a defined search path (outlined in Figure 1, and explained in detail below), extracting a large variety of different data fields for both chemicals and proteins (shown in Table 1). In addition, BioSpider includes a built-in referencing application that maintains the source for each piece of data obtained. Thus, if BioSpider obtains the Pubchem ID for a compound using KEGG, a reference "Source: KEGG" is added to the reference table for the Pubchem ID.

Figure 1 - Simplified overview of a BioSpider search

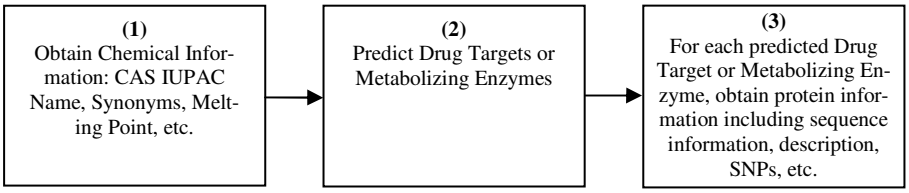


Table 1 - Summary of some of the fields obtained by BioSpider

Drug or Compound Information	Drug Target or Receptor Information
Generic Name	Name
Brand Names/Synonyms	Synonyms
IUPAC Name	Protein Sequence
Chemical Structure/Sequence	Number of Residues
Chemical Formula	Molecular Weight
PubChem/ChEBI/KEGG Links	pI
SwissProt/GenBank Links	Gene Ontology
FDA/MSDS/RxList Links	General Function
Molecular Weight	Specific Function
Melting Point	Pathways
Water Solubility	Reactions
pKa or pI	Pfam Domains
LogP or Hydrophobicity	Signal Sequences
NMR/Mass Spectra	Transmembrane Regions
MOL/SDF Text Files	Essentiality
Drug Indication	Genbank Protein ID
Drug Pharmacology	SwissProt ID
Drug Mechanism of Action	PDB ID
Drug Biotransformation/Absorption	Cellular Location
Drug Patient/Physician Information	DNA Sequence
Drug Toxicity	Chromosome Location

Step 1: Compound Annotation

Compound annotation involves extracting or calculating data about small molecule compounds (metabolites and drugs). This includes data such as common names, synonyms, chemical descriptions/applications, IUPAC names, chemical formulas, chemical taxonomies, molecular weights, solubilities, melting or boiling points, pKa, LogP's, state(s), MSD sheets, chemical structures (MOL, SDF and PDB files), chemical structure images (thumbnail and full-size PNG), SMILES strings, InCHI identifiers, MS and NMR spectra, and a variety of database links (PubChem, KEGG, ChEBI). The extraction of this data involves accessing, screen scraping and text mining ~30 well known databases (KEGG, PubChem), calling a number of predictive programs (for calculating MW, solubility) and running a number of file conversion scripts and figure generation routines via CORINA [12], Checkmol (http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html) and other in-house methods. The methods used to extract and generate these data are designed to be called independently but they are also "aware" of certain data dependencies. For instance, if a user only wanted an SDF file for a compound, they would simply call a single method: get_value('sdf_file'). There is no need to explicitly call methods that might contain the prerequisite information for getting an SDF file. Likewise, if BioSpider needs a Pubchem ID to grab an SDF file, it will obtain it automatically, and, consequently, if a Pubchem ID requires a KEGG ID, BioSpider will then jump ahead to try and get the KEGG ID automatically.

Step 2: Target/Enzyme Prediction

Target/enzyme prediction involves taking the small-molecule query and identifying the enzymes likely to be targeted or involved in the metabolism of that compound. This process involves looking for metabolite-protein or drug-protein associations from several well-known databases including SwissProt, PubMed, DrugBank and KEGG. The script begins by constructing a collection of query objects from the supplied compound information. Each query object contains the name and synonyms for a single compound, as well any similar but unwanted terms. For example, a query object for the small molecule compound "pyridoxal" would contain the term "pyridoxal phosphatase" as an unwanted term, since the latter name is for an enzyme. The list of unwanted or excluded terms for small molecule compounds is assembled from a list of the names and synonyms of all human proteins. These unwanted terms are identified automatically by testing for cases where one term represents a subset of another. Users can also include their own "exclusion" terms in BioSpider's advanced search interface.

The name and synonyms from a query object are then submitted using WWW agents or public APIs to a variety of abstract and protein sequence databases, including Swiss-Prot, PubMed, and KEGG. The name and synonyms are each sub-

mitted separately, rather than as a single query, since queries consisting of multiple synonyms typically produce many irrelevant results. The relevance of each of the returned records is measured by counting the number of occurrences of the compound name and synonyms, as well as the number of occurrences of the unwanted terms. Records containing only the desired terms are given a "good" rating, while those containing some unwanted terms are given a "questionable" rating. Records containing only unwanted terms are discarded. The records are then sorted based on their qualitative score. BioSpider supports both automated identification and semi-automated identification. For automated identification, only the highest scoring hits (no unwanted terms, hits to more than one database) are selected. In the semi-automated mode, the results are presented to a curator who must approve of the selection. To assist with the decision, each of the entries in the document is hyperlinked to the complete database record so that the curator can quickly assess the quality of the results. Note that metabolites and drugs often interact with more than one enzyme or protein target.

Step 3: Target/Enzyme Annotation

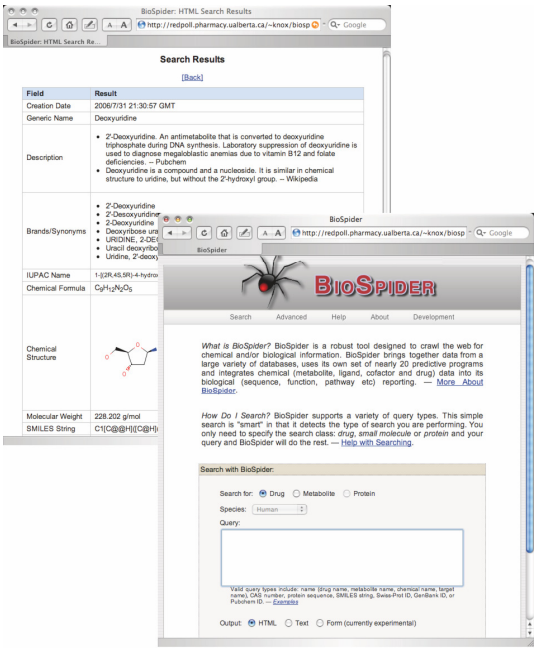
Target/Enzyme annotation involves extracting or calculating data about the proteins that were identified in Step 2. This includes data such as protein name, gene name, synonyms, protein sequence, gene sequence, GO classifications, general function, specific function, PFAM [13] sequences, secondary structure, molecular weight, subcellular location, gene locus, SNPs and a variety of database links (SwissProt, KEGG, GenBank). Approximately 30 annotation sub-fields are determined for each drug target and/or metabolizing enzyme. The BioSpider protein annotation program is based on previously published annotation tools developed in our lab including BacMap [14], BASYS and CCDB [15]. The Swiss-Prot and KEGG databases are searched initially to retrieve protein and gene names, protein synonyms, protein sequences, specific and general functions, signal peptides, transmembrane regions and subcellular locations. If any annotation field is not retrieved from the above-mentioned databases then either alternate databases are searched or internally developed/installed programs are used. For example, if transmembrane regions are not annotated in the Swiss-Prot entry, then a locally installed transmembrane prediction program called TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) is used to predict the transmembrane regions. This protein annotation tool also coordinates the updating of fields that are calculated from the contents of other fields, such as molecular weight and isoelectric point. The program also retrieves chromosome location, locus location and SNP information from GeneCards [16] on the basis of the gene name. BLAST searches are also performed against the PDB database to identify structural homologues. Depending upon the sequence similarity between the query protein sequence to a sequence represented in the PDB database, a program

called HOMODELLER (X. Dong, unpublished data) may generate a homology model for the protein sequence.

2.2. *Implementation*

The BioSpider backend is a fully objected-oriented Perl application, making it robust and portable. The frontend (website, shown in Figure 2) utilizes Perl CGI scripts which generate valid XHMTL and CSS. BioSpider uses a relational database (MySQL 5) to store data as it runs. As BioSpider identifies and extracts different pieces of information, it stores the data in the database. To facilitate this storage process, a module called a "DataBean" is used to store and retrieve the desired information from/to the database. This approach was chosen for 3 reasons: 1) it provides an "audit-trail" in terms of the results obtained, 2) it provides a complete search result history, enabling the easy addition of "saved-searches" to the website, and 3) it reduces memory load as the application is running. A screenshot of the BioSpider website is shown in Figure 2.

Figure 2 – A screen shot montage of BioSpider



3. Validation, Comparison and Limitations

Text mining and data extraction tools can be prone to a variety of problems, many of which may lead to nonsensical results. To avoid these problems BioSpider performs a number of self-validation or "sanity checks" on specific data extracted from the web. For example, when searching for compound synonym names, BioSpider will check that the PubChem substance page related to that synonym contains the original search name or original CAS number within the HTML for that page. This simple validation procedure can often remove bogus synonyms obtained from different websites. Other forms of such small-scale validation or sanity-checks includes a CAS number validation method, whereby the CAS number check-digit is used to validate the entire CAS number (CAS numbers use a checksum, whereby the checksum is calculated by taking the last digit times 1, the next digit times 2, the next digit times 3 etc., adding all these up and computing the sum modulo 10).

Since the majority of the information obtained by BioSpider is screen-scraped from several websites, it is also important to validate the accessibility of these websites as well as the HTML formatting. Since screen-scraping requires one to parse the HTML, BioSpider must assume the HTML from a given website follows a specific format. Unfortunately, this HTML formatting is not static, and changes over time as websites add new features, or alter the design layout. For this reason, BioSpider contains an HTML validator application, designed to detect changes in the HTML formatting for all the web-resources that BioSpider searches. To achieve this, an initial search was performed and saved using BioSpider for 10 pre-selected compounds, whereby the results from each of the fields were manually validated. This validation-application performs a search on these 10 pre-selected compounds weekly (as a cron job). The results of this weekly search are compared to the original results, and if there is any difference, a full report is generated and emailed to the BioSpider administrator.

The assessment of any text mining or report generating program is difficult. Typically one must assess these kinds of tools using three criteria: 1) accuracy; 2) completeness and 2) time savings. In terms of accuracy, the results produced are heavily dependent on the quality of the resources being accessed. Obviously if the reference data are flawed or contradictory, the results from a BioSpider search will be flawed or contradictory. To avoid these problems every effort has been made to use only high-accuracy, well curated databases as BioSpider's primary reference sources (KEGG, SwissProt, PubChem, DrugBank, Wikipedia, etc). As a result, perhaps the most common "detectable" errors made by BioSpider pertain to text parsing issues (with compound descriptions), but these appear to be relatively minor. The second most common error pertains to errors of omission (missing data that could be found by a human expert looking through the web or other references). In addition to these potential programmatic errors, the performance of BioSpider can be com-

promised by incorrect human input, such as a mis-spelled compound name, SMILES string or CAS number or the submission of an erroneous MOL or SDF file. It can also be compromised by errors or omissions in the databases and websites that it searches. Some consistency or quality control checks are employed by the program to look for nomenclature or physical property disagreements, but these may not always work. BioSpider will fail to produce results for newly discovered compounds as well as compounds that lack any substantive electronic or web-accessible annotation. During real world tests with up to 15 BioSpider users working simultaneously for 5-7 hours at a time, we typically find fewer than two or three errors being reported. This would translate to 1 error for every 15,000 annotation fields, depending on the type of query used. The number of errors returned is highest when searching using a name or synonym, as it is difficult to ascertain correctness. Errors are much less likely when using a search that permits a direct mapping between a compound and the source websites used by BioSpider. It is thus recommended that users search by structure (InChI, SDF/MOL, SMILES) or unique database ID (pubchem ID, KEGG ID) first, resorting to CAS number or name only when necessary. Despite this high level of accuracy, we strongly suggest that every BioSpider annotation should be looked over quickly to see if any non-sensical or inconsistent information has been collected in its annotation process. Usually these errors are quite obvious. In terms of errors of omission, typically a human expert can almost always find data for 1 or 2 fields that were not annotated by BioSpider – however this search may take 30 to 45 minutes of intensive manual searching or reading.

During the annotation of the HMDB and DrugBank, BioSpider was used to annotate thousands of metabolites, food additives and drugs. During this process, it was noted that BioSpider was able to obtain at least some information about query compounds 91% of the time. The cases where no information was returned from BioSpider often involved compounds whereby a simple web search for that compound would return no results. This again spotlights one of the limitations of the BioSpider approach – its performance is directly proportional to the "web-presence" of the query compound.

Perhaps the most important contribution for Biospider for annotation lies in the time savings it offers. Comparisons between BioSpider and skilled human annotators indicate that BioSpider can accelerate annotations by a factor of 40 to 50 X over what is done by skilled human annotators. In order to test this time-saving factor, 3 skilled volunteers were used. Each volunteer was given 3 compounds to annotate (2-Ketobutyric acid, Chenodeoxycholic acid disulfate and alpha-D-glucose) and the fields to fill-in for that compound. Each volunteer was asked to search for all associated enzymes, but only asked to annotate a single enzyme by hand. The data obtained by the volunteers were then compared to the results produced by BioSpider. These tests indicated that the time taken to annotate the chemical fields averages 40 minutes and 45 minutes for the biological fields, with a range between 22

and 64 minutes. The time taken by Biospider was typically 5 minutes. In other words, to fill out a complete set of BioSpider data on a given small molecule (say biotin) using manual typing and manual searches typically takes a skilled individual approximately 3 hours. Using BioSpider this can take as little as 2 minutes. Additionally, the quality of data gathered by BioSpider matched the human annotation for almost all of the fields. Indeed, it was often the case that the volunteer would give up on certain fields (pubchem substance IDs, OMIM IDs, etc.) long before completion.

In terms of real-world experience, BioSpider has been used in several projects, including DrugBank and HMDB (www.hmdb.ca). It has undergone full stress testing during several "annotation workshops" with up to 50 instances of BioSpider running concurrently. BioSpider has also been recently integrated into a LIMS system (MetaboLIMS – http://www.hmdb.ca/labm/). This allows users to produce a side-by-side comparison on the data obtained using BioSpider and the data collected manually by a team of expert curators. Overall, BioSpider has undergone hundreds of hours of real-life testing, making it stable and relatively bug-free.

4. Conclusion

BioSpider is a unique application, designed to fill in the gap between chemical (small-molecule) and biological (target/enzyme) information. It contains many advanced predictive algorithms and screen-scraping tools made interactively accessible via an easy-to-use web front-end. As mentioned previously, we have already reaped significant benefits from earlier versions of BioSpider in our efforts to prepare and validate a number of large chemical or metabolite databases such as DrugBank and HMDB. It is our hope that by offering the latest version of BioSpider to the public (and the metabolomics community in particular) its utility may be enjoyed by others as well.

5. Acknowledgments

The Human Metabolome Project is supported by Genome Alberta, in part through Genome Canada.

References

1. Sussman, JL, Lin, D, Jiang, J, Manning, NO, Prilusky, J, Ritter, O & Abola, EE. Protein data bank (PDB): a database of 3D structural information of biological macromolecules. *Acta Cryst.* 1998. D54:1078-1084.
2. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32(Database issue):D277-280.
3. Van Domselaar GH, Stothard P, Shrivastava S, Cruz JA, Guo A, Dong X, Lu P, Szafron D, Greiner R, Wishart DS. 2005. BASys: a web server for automated bacterial genome annotation. *Nucleic Acids. Res.* 1;33(Web Server issue):W455-9.
4. Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L. and Weinstein, J.N. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling**.** *Biotechniques* 1999. 27:1210-1217.
5. Hoffmann, R. and Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 2005. 21 Suppl 2:ii252-ii258.
6. Hu Y., Hines L.M., Weng H., Zuo D., Rivera M., Richardson A. and LaBaer J: Analysis of genomic and proteomic data using advanced literature mining. J *Proteome Res.* 2003. Jul-Aug;2(4):405-12.
7. Maier H., Dohr S., Grote K., O'Keeffe S., Werner T., Hrabe de Angelis M. and Schneider R: LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts. *Nucleic Acids Res.* 2005. Jul 1;33(Web Server issue):W779-82.
8. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33(Database issue):D514-517.
9. Hewett, M., Oliver, D.E., Rubin, D.L., Easton, K.L., Stuart, J.M., Altman, R.B. and Klein, T.E. 2002. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res.* 30:163-165.
10. Wishart, D.S., Knox, C., Guo, A., Shrivastava, S., Hassanali, M., Stothard, P. and Woolsey, J. 2006. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids. Res.* 34(Database issue):D668-672.
11. Weininger, D. 1988. SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* 28:31-38.
12. Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V: Chemical information in 3D-space. *J Chem Inf Comput Sci* **36**: 1030-1037, 1996.
13. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. 2004. The Pfam protein families database. *Nucleic Acids Res.* 32:D138–141.

14. Stothard P, Van Domselaar G, Shrivastava S, Guo A, O'Neill B, Cruz J, Ellison M, Wishart DS. BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucleic Acids Res.* 2005 Jan 1;33(Database issue):D317-20.
15. Sundararaj S, Guo A, Habibi-Nazhad B, Rouani M, Stothard P, Ellison M, Wishart DS. BacMap: an interactive picture atlas of annotated bacterial genomes. *Nucleic Acids Res.* 2005 Jan 1;33(Database issue):D317-20.
16. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics* 14:656-664.
17. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25:3389-3402.
18. Bairoch, A., Apweiler. R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. et al. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res*. 33(Database issue):D154-159.
19. Brooksbank, C., Cameron, G. and Thornton, J. 2005. The European Bioinformatics Institute's data resources: towards systems biology. *Nucleic Acids Res.* 33 (Database issue):D46-53.
20. Chen, X., Ji, Z.L. and Chen, Y.Z. 2002. TTD: Therapeutic Target Database. *Nucleic Acids Res.* 30:412-415.
21. Halgren, T.A., Murphy, R.B., Friesner, R.A., Beard, H.S., Frye, L.L., Pollard, W.T. and Banks, J.L. 2004. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* 47:1750-1709.
22. Hatfield, C.L., May, S.K. and Markoff, J.S. 1999. Quality of consumer drug information provided by four Web sites. *Am. J. Health Syst. Pharm.* 56:2308-2311.
23. Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. 2004. Recent improvements to the PROSITE database. *Nucleic Acids Res.* 32:D134–137.
24. Kramer, B., Rarey, M. and Lengauer, T. 1997. CASP2 experiences with docking flexible ligands using FlexX. *Proteins* Suppl 1:221-225
25. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305:567-580.
26. McGuffin, L.J., Bryson, K. and Jones, D.T. 2000. The PSIPRED protein structure prediction server. *Bioinformatics* 16:404–405.
27. Montgomerie, S., Sundararaj, S., Gallin, W.J. and Wishart, D.S. 2006. Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* 7:301-312.

28. Orth, A.P., Batalov, S., Perrone, M. and Chanda, S.K. 2004. The promise of genomics to identify novel therapeutic targets. *Expert Opin. Ther. Targets*. 8:587-596.
29. Sadowski, J. and Gasteiger J. 1993. From atoms to bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev*. 93: 2567-2581.
30. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. and Yaschenko, E. 2005. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33(Database issue):D39-45.
31. Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R.F., Sykes, B.D. and Wishart, D.S. 2003. VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res*. 31:3316–3319.