Large-Scale Annotation of Small-Molecule Libraries Using Public Databases

Yingyao Zhou,*,† Bin Zhou,† Kaisheng Chen,† S. Frank Yan,† Frederick J. King,†,‡
Shumei Jiang,† and Elizabeth A. Winzeler†

Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego,
California 92121, and Developmental and Molecular Pathways, Novartis Institutes for BioMedical Research,
250 Massachusetts Avenue, Cambridge, Massachusetts 02139

Received March 9, 2007

While many large publicly accessible databases provide excellent annotation for biological macromolecules, the same is not true for small chemical compounds. Commercial data sources also fail to encompass an annotation interface for large numbers of compounds and tend to be cost prohibitive to be widely available to biomedical researchers. Therefore, using annotation information for the selection of lead compounds from a modern day high-throughput screening (HTS) campaign presently occurs only under a very limited scale. The recent rapid expansion of the NIH PubChem database provides an opportunity to link existing biological databases with compound catalogs and provides relevant information that potentially could improve the information garnered from large-scale screening efforts. Using the 2.5 million compound collection at the Genomics Institute of the Novartis Research Foundation (GNF) as a model, we determined that ∼4% of the library contained compounds with potential annotation in such databases as PubChem and the World Drug Index (WDI) as well as related databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) and ChemIDplus. Furthermore, the exact structure match analysis showed 32% of GNF compounds can be linked to third party databases via PubChem. We also showed annotations such as MeSH (medical subject headings) terms can be applied to in-house HTS databases in identifying signature biological inhibition profiles of interest as well as expediting the assay validation process. The automated annotation of thousands of screening hits in batch is becoming feasible and has the potential to play an essential role in the hit-to-lead decision making process.

INTRODUCTION

Large amounts of knowledge about small molecules can be found in patents, the scientific literature, and marketed and investigational drug databases.[1,2] Other sources include cocrystal structures from databases such as PDB[3,4] and large, proprietary, activity databases accumulated within pharmaceutical companies.[5] However, automatic annotation of small-molecule libraries remains an unsolved problem for the cheminformatics community mainly due to the complexity of how small molecules are presented. The vast majority of chemical structures cannot be reliably represented by a few keywords in a manner similar to what is common for macromolecules such as biologically relevant natural product molecules. Therefore, knowledge about small molecules is difficult to retrieve by text indexing services.

Using the anticancer drug Gleevec as an example, a Google Scholar search (http://www.scholar.google.com) does not yield useful results if search terms are based on its IUPAC name, InChI code for “Imatinib”, or its Chemical Abstracts Service (CAS) registration number “152459-95-5”. In contrast, search terms such as “Imatinib”, “Gleevec”, “Glivec”, or “STI571” do lead to instructive information. However, these edifying synonyms often are not assigned to a compound until it enters the late stages of clinical development.

Initiatives are underway to automatically retrieve chemical information from text sources such as the 8 million pages of the U.S. Patent Library.[6] However, these tools are currently at the beta-testing phase and unavailable to the research community. Commercial products such as SciFinder (http://www.cas.org/SCIFINDER/) and IDdb3 (http://www.iddb3.com) often rely on manual curation, which results in high costs that impedes access by a large number of biomedical researchers. Despite the concerns about the business model of commercial proprietary registration systems such as CAS,[7] there are few alternative annotation services so far.

Modern drug discovery programs that incorporate high-throughput screening (HTS) technologies typically identify hundreds to several thousand primary hits from a single screening campaign. The application of filters to the initial screening results based upon factors such as compound potency, selectivity, physicochemical properties, and the existence of assigned intellectual property provides an opportunity to reduce the effort spent on a significant number of compounds at an early stage that presumably would be triaged eventually. However, the development of the appopriate informatic tools is encumbered by the lack of annotation systems that provide application programming interfaces (APIs) for subscribers to interrogate a large number of chemical structures within a single request. Therefore, annotation searches such as patent searches typically are restricted to only a handful of selected structures; no

* Corresponding author e-mail: yzhou@gnf.org.
† Genomics Institute of the Novartis Research Foundation.
‡ Novartis Institutes for BioMedical Research.

10.1021/ci700092v CCC: $37.00
Published on Web 07/03/2007

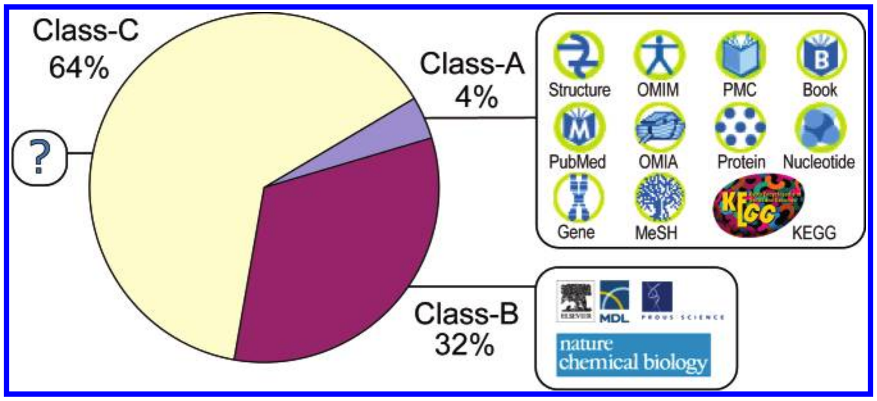

Figure 1. Classification of all available PubChem and WDI structures divided into three classes based on their levels of annotations. Class-A comprises 4% of the structures that have annotations in databases such as Entrez PubMed, Structure, Protein, Nucleotide, Gene, OMIM, and KEGG, etc. Class-B comprises 32% of the structures that are contributed by content providers such as DiscoveryGate, Prous, Nature Chemical Biology, etc. Class-C comprises the remaining 64% of the structures that have no readily available annotations.

information typically is obtained for the majority of the hits. Providing annotation information to a much broader scale of screening hits can be of great interest in several aspects: annotated hits may assist in biological validation of the assay; the mechanism of action of false negatives derived from the current screening assay protocol could be identified; known patented scaffolds could be avoided from further analyses; desirable scaffolds for scaffold hopping or usage repositioning could be prioritized; etc. Therefore, the difficulties inherent in retrieving compound annotations for an extensive compound list result in this information being underutilized in the hit-to-lead decision making process.

PubChem (http://pubchem.ncbi.nlm.nih.gov), a critical part of the Molecular Libraries Initiative of the NIH Roadmap (Austin 2004), is a new member of the NCBI Entrez data warehouse.[8] As of September 2006, PubChem covered 8.0 million unique structures from 48 contributing sources. We separated these structures into three classifications (Figure 1): (class-A) 337 875 unique structures in PubChem that are linked to annotation data sources such as PubMed, PMC, MeSH, MMDB, KEGG, ChemIDplus, MICAD, etc.; (class-B) 2.5 million structures contributed from content providers such as MDL/Elsevier DiscoveryGate (http://www.discoverygate.com); and (class-C) the remaining 5 million structures where little or no annotation is recorded.

In addition to the PubChem database, the World Drug Index (WDI) adds an additional 37 168 structures to the class-A category. Across multiple sources, we identified approximately 3 million class-A or -B structures that are at the disposal of the cheminformatics community to construct such an annotation database. Random sampling was used for the three structure classes to estimate the correlation between PubChem annotations and our current understanding of small molecules as represented by the CAS database. Statistics obtained from these analyses were then applied to estimate the scope of CAS annotations. Data presented here support the argument that PubChem currently plays a complementary rather than redundant role to the CAS database.[9]

In order to annotate a compound collection, one straightforward approach would be to hyperlink each in-house structure to its PubChem match and redirect scientists to the PubChem Web site for further information. This does not solve the above-mentioned library annotation problem for several main reasons. First, scientists have to interrogate each structure individually, which remains time-consuming. Second, annotations often do not reside on the PubChem summary page, and it takes several additional steps to locate the pertinent information. Third, many interesting bioinformatics annotations, such as compound−gene ontology associations, are not present in PubChem. Fourth, annotations for structurally similar compounds or similar annotations for structurally distinct compounds are not pooled and analyzed together.

Here, we propose an integrated cheminformatics and bioinformatics pipeline, which is capable of automatically creating a comprehensive local compound annotation database. We further illustrate how such an annotation database can provide significant and substantial information to scientists for their screening compound collection at the very early stages of the drug discovery process. A Web service has also been established for public users to submit compounds in batch and directly retrieve relevant compound knowledge from the annotation database.

MATERIALS AND METHODS

PubChem-Centered Annotation Pipeline. PubChem compound and substance databases were downloaded from the NCBI (ftp://ftp.ncbi.nlm.nih.gov/pubchem/) in September 2006. At that time, the database was comprised of 8 000 122 unique structures from 12 798 100 substances. Substance SDF files provide links to nucleotide, protein, MMDB, PubMed databases, etc. Using the NCBI programming interface (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html), we retrieved links from compounds to PMC, MeSH, and MICAD data sources. Nucleotide and Protein objects were further cross-linked to the corresponding Gene objects and then be linked to OMIM and OMIA databases containing disease phenotypes data as well as the GO database containing potential functional and pathway annotations. Summary annotation information for all NCBI database objects can be downloaded in XML format using the NCBI programming interface.

We next downloaded the Kyoto Encyclopedia of Genes and Genomes (KEGG) compound and drug databases, which contain 12 135 compounds and 2465 drugs and a total of 12 832 links to PubChem substance identifiers. KEGG is useful in terms of providing metabolic pathway annotations for compounds and potential target annotations for drugs.

Table 1. List of Data Sources That Are Used in the Annotation Integration Pipeline

acronym	description	URL
PubChem	Include NCBI Compound, Substance and BioAssay databases; disseminate chemical and biological data as part of the NIH's Molecular Libraries Roadmap Initiative	http://pubchem.ncbi.nlm.nih.gov
Nucleotide	NCBI Nucleotide database	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
Protein	NCBI Protein database	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
MMDB	NCBI Molecular Modeling Database	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure
PubMed	NCBI PubMed database, biomedical papers	http://www.ncbi.nlm.nih.gov/entrez
PMC	PubMed Central, full text access	http://www.pubmedcentral.nih.gov/
MeSH	Medical Subject Headings provided by the National Library of Medicine, including Pharmacological Action annotation	http://www.nlm.nih.gov/mesh
ChemIDplus	Nomenclature and toxicity link	http://sis.nlm.nih.gov/chemical.html
MICAD	Molecular Imaging & Contrast Agent	http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/micad/home.html
Gene	Entrez Gene database, replacing LocusLink	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
OMIM	Online Mendelian Inheritance in Man	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=omim
OMIA	Online Mendelian Inheritance in Animals	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=omia
GO	Gene Ontology	http://www.geneontolgy.org
KEGG	Kyoto Encyclopedia of Genes and Genomes (free access for academic institute)	http://www.genome.jp/kegg/
WDI	World Drug Index, drug and developmental compounds (commercial)	http://scientific.thomson.com/products/wdi/

In addition, we also retrieved the mechanism of action keywords from the licensed World Drug Index database, which contains 62 582 therapeutics.

Statistical Analysis of CAS Annotations for PubChem Structures. For each of the 300 PubChem structures randomly selected from the three compound classes (100 compounds per class), we carried out a manual exact structure search using SciFinder. We recorded information such as whether a structure was found, whether it had references, whether the reference list contained patent data, and whether there were reaction and/or synthesis schema available. According to the binomial distribution, if *n* out of the 100 structures for a specific compound class are found to have a certain type of annotation in CAS, we can infer the probability of a randomly selected structure from the same category to carry the same type of annotation in CAS to be *n*% with a standard error of the mean of $\sqrt{n(100-n)/100}$ %.

RESULTS AND DISCUSSION

PubChem-Centered Annotation Integration. PubChem is the main database enabled for this study. Additional databases such as KEGG and WDI were also used to increase the potential sources for content (see Materials and Methods). Although databases such as DrugBank (http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/drugcard_field_expl.cgi) have not been included in PubChem, they are most likely a subset of PubChem and WDI; therefore, these satellite databases were not included in our pilot study. A methodology for potentially increasing the scope of annotations is exemplified by the SMID database (http://smid.blueprint.org). This collection was developed by applying protein sequences retrieved from ligand−protein structure databases to homology BLAST searches, resulting in an expansion of the list of potential compound−protein interaction pairs. However, this approach is not applied in our study in order to preserve the quality of our annotations and avoid the complication of false positives introduced via homology extrapolation. A summary of the databases used in our integration pipeline and their associated URLs is shown in Table 1. Starting from PubChem, we retrieved data and derived links to all the data sources described above and then followed the links and downloaded summary annotation information from each corresponding data source. Chemical structures and their annotations were integrated using various public bioinformatics programming utilities such as NCBI eUtils as well as in-house program routines. The resultant annotation database contained 2 372 195 annotations for 337 875 PubChem structures and is schematically represented in Figure 2. PubChem provides annotations (links) for these compounds from a large variety of data sources: 82 199 links to PubMed, 99 316 links to MeSH, 1 597 991 links to PMC, 56 638 links to MMDB, 276 056 links to ChemIDplus (which further contains links to the TOXicology Data NETwork−TOXNET, http://www.nlm.nih.gov/pubs/factsheets/toxnetfs.html), 125 links to MICAD, 3151 links to NCBI Nucleotide, and 172 583 links to NCBI Protein databases. One can further derive 1703 links to the NCBI Gene database, 158 links to OMIM, 5 links to OMIA, and 1175 to the GO databases. Together with the WDI entries, the resultant knowledgebase contains annotations for about 400 456 unique class-A structures (Figure 1). Further analyses show as many as 51% of these structures are indexed by ChemIDplus, 11−15% of them have links to MeSH, WDI, and PubMed, and 2−5% of them have links to PMC, KEGG, and MMDB (Figure 3). Except for the links to commercial knowledgebases that need to be established by scientists manually, the pipeline described above can be executed periodically to keep the resultant annotation database current.

Class-B and Class-C Compounds. For the remaining PubChem structures, direct public annotation is not available at this point (class-B or -C; Figure 1). However, these compounds could still be interrogated if their structures were found in catalogs of content providers. For example, if a structure is contributed by MDL xpharm, one would expect additional pharmacological data to be available. Similarly, if a structure is present in a source such as Prous or Nature Chemical Biology, one would expect research literature offered at the Web sites of the respective publishers. An entry

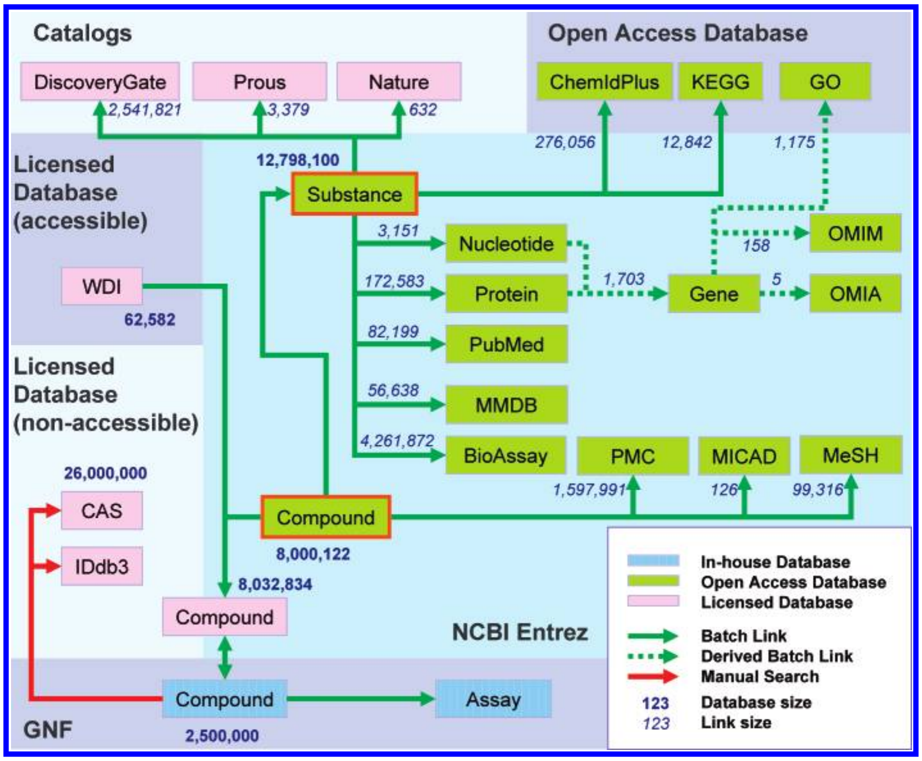

Figure 2. Schematic illustration of the GNF compound annotation pipeline. 12 798 100 PubChem substances and 8 000 122 PubChem compounds are linked to various open-access databases and commercial catalogs. These links can be either directly retrieved from PubChem or identified by additional bioinformatics tools in batch mode. Eight million PubChem structures were then merged with 62 582 WDI structures, which resulted in a collection of 8 032 834 unique structures. This composite library was then mapped to 32% of the 2.5 million GNF in-house compound collections via exact and similar structure searches. Links between GNF structures and commercial knowledgebases, such as CAS and IDdb3, have to be established manually by scientists due to the lack of programming interfaces.

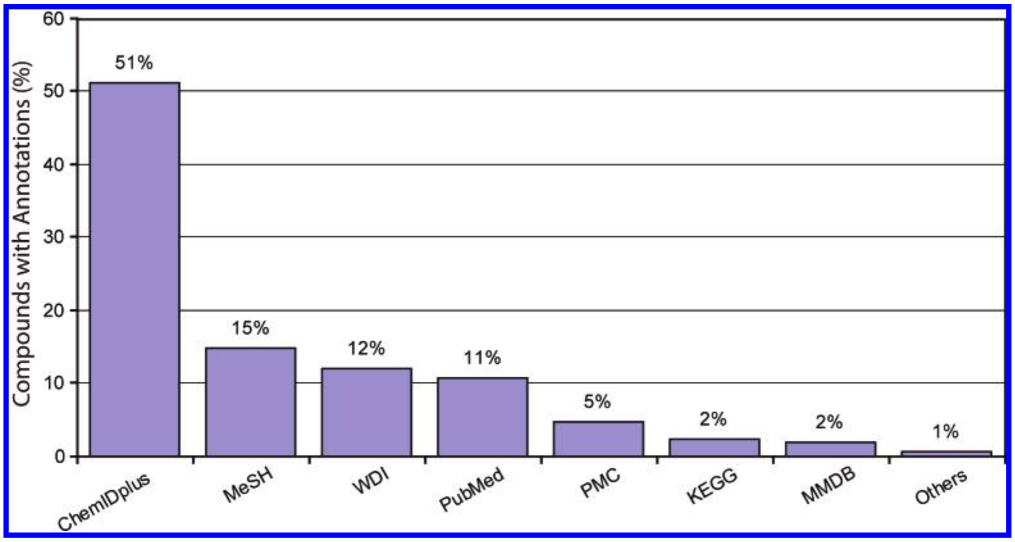

Figure 3. Percentages of the class-A structures that carry various annotations. ChemIDplus, MeSH, WDI, and PubMed are the four annotation categories that cover more than 90% of the class-A structures.

in NMRShiftDB, NIST, or NIST Chemistry WebBook implies that experimental analytical chemistry data are available.

Readers may refer to the PubChem Web page for more updated contributor information in terms of the type of data they can provide (http://pubchem.ncbi.nlm.nih.gov/sources/sources.cgi?mode=substanceinfo). Therefore, a simple binary flag designating whether a compound is expected to be found in various external databases can be an important time-saving tool to identify compounds with existing information. The nature of this information, however, can be diverse, such as intellectual property documentation, reaction and synthesis information, and/or simply the compound suppliers. We classified 2.5 million PubChem structures originated from DiscoveryGate, publishers, and other content providers as class-B structures.

The remaining 5 million PubChem structures likely are from chemical vendors or various screening centers such as ChemDB and the NIH Molecular Libraries Screening Center Network (MLSCN). Because these compounds do not have direct biological annotations (biological measurements provided by PubChem BioAssay database are not considered in this study), we expect that these class-C compounds are

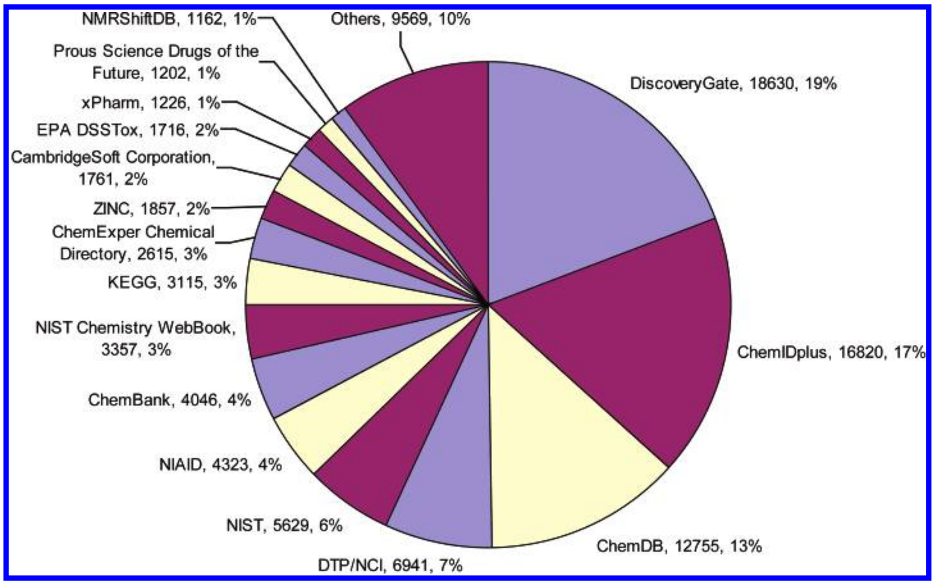

Figure 4. Source distribution of WDI structures according to PubChem substance contributors. Notice that DiscoveryGate, ChemIDplus, and ChemDB contribute about 50% of the WDI structures deposited in PubChem, while the remaining 50% of the compounds are presented in over a dozen additional sources.

less well characterized compared to class-A and class-B structures.

WDI and PubChem Are Complementary. We compared the annotations derived from the open-access PubChem (including KEGG) database to the WDI database. Our analysis showed as many as 37 168 structures (60%) in the WDI database still are not covered by PubChem; therefore, publicly available data sources are not redundant with the information present within the WDI at this time. For the 40% of WDI structures that overlap with PubChem, Figure 4 illustrates that approximately 50% of them also are covered predominantly by DiscoveryGate, ChemIDplus, and ChemDB and to a lesser extent by a dozen additional sources. It should be noted that for the WDI entries that overlap, PubChem generally provides much more extensive annotation obtained from a variety of sources compared to the simple keyword annotation present in the WDI database. Clearly, commercial databases such as WDI and public counterparts such as PubChem currently play complementary roles, and their integration should greatly benefit users of each database.

Statistical Analysis of PubChem Annotation Quality. We classified PubChem structures into the three categories described above based on their level of annotation available in the public domain. We assumed that the CAS database was the "golden standard" for representing all of the current knowledge of small molecules largely because of its extensive comprehensiveness compared to other sources. If the knowledge captured by PubChem is representative of the current annotation information that exists for these compounds, we would expect a high likelihood that class-A structures would also be annotated in the CAS database. Furthermore, we would predict a significant percentage of the CAS database to contain information for class-B structures, whereas annotation for class-C structures should be relatively infrequent. In order to examine whether these assumptions are true, we randomly selected 100 PubChem structures from each of the three annotation classes. Next, we performed exact structure searches on the 300 structures using SciFinder. The types of CAS annotations were recorded if a match was identified. Our statistical analysis is shown in Figure 5 (also see Materials and Methods). The results validated our hypothesis that the extent of available PubChem annotations, represented by the three annotation classes, correlated with the probability of having CAS annotation. That is, class-A structures have a 99% chance to be found in the CAS database, along with a 95% and 65% probability to have associated literature references and patent documents, respectively. However, this does not imply that the associated information regarding a particular compound for the two knowledgebases is redundant. Specifically, as previously discussed, CAS provides chemical, commercial, and patent information to chemists, while PubChem integrates medical information for medical researchers (http://www.arl.org/info/frn/gov/ACS/backgroundfaqpb.pdf). Therefore, an integration of the annotations from both databases would create a more complete description of the biological and pharmacological properties for the compounds of interest. Unfortunately, as many as 36% of the class-C structures found in the PubChem database currently are not present in the CAS database, indicating that CAS has not taken advantage of this public resource.

Only 20% of the structures in the DiscoveryGate catalogs have literature annotation in the CAS database. However, manual searching of the 100 class-B structures through DiscoveryGate indicated that the vast majority of these compounds (93 ± 3%) only have associated "trivial" annotations such as compound identifiers, in silico physicochemical properties, computational 3D structures, etc. In contrast, the remaining 7% of the structures have experimental data with the corresponding citation information. All of these seven structures also are found in the CAS database, where more extensive annotations are present. Due to the fact that the majority of DiscoveryGate compounds do not provide biological and pharmacological data (which is provided under a separate MDL product called xPharm), we estimated that only 175 000 class-B DiscoveryGate structures have nontrivial annotations (and this is mostly synthesis

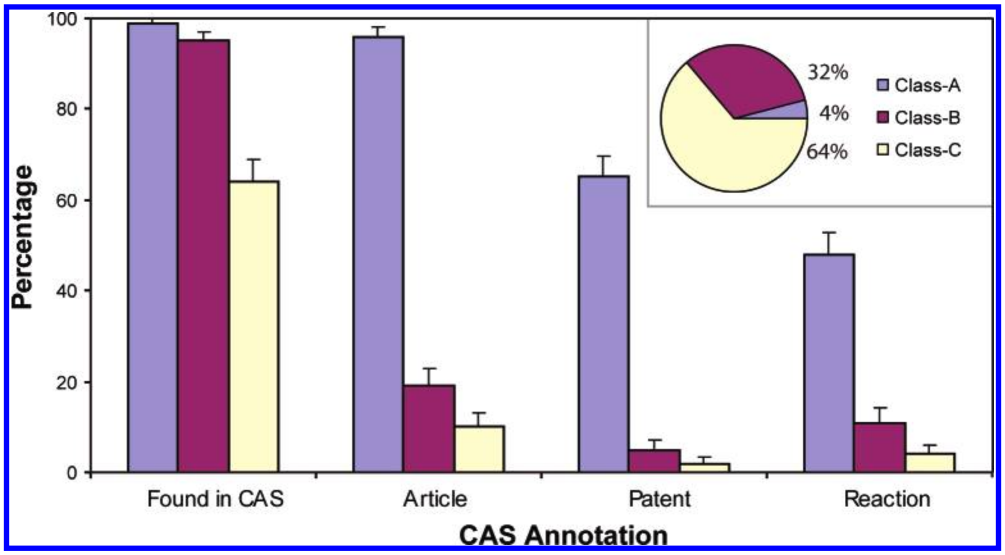

Figure 5. The probability of finding articles, patents, and reaction information in the CAS database by SciFinder searches using structures randomly selected from class-A, class-B, and class-C. Notice that the probability of locating valuable annotation information from the CAS database decreases as the amount of annotation available in PubChem decreases for these three classes. Statistics for class-B and class-C structures are similar largely because most of the DiscoveryGate compounds do not have appropriate annotations. The fact that 36% of class-C PubChem structures are not found in the CAS database indicates that CAS has not yet taken advantage of all of the available data in the public domain.

information). This observation also suggests that the annotation levels of class-B and class-C structures are fairly comparable in our CAS search analyses.

We observed that the availability of compounds with annotation in our resultant database roughly parallels the availability of annotations in the CAS database. By applying our compiled annotation database to a list of several thousand structures, one would be able to distinguish and prioritize structures with a relatively high chance of CAS annotation (i.e., class-A) versus those with a lower chance of CAS annotation (i.e., class-B and class-C). This type of batch annotation tool would be valuable in addressing such issues as increasing competition and intellectual property space.

Annotation by a CAS Registration Number. It would be preferable if all 26 million structures in the CAS system were readily accessible via PubChem or other informatic services, in order to prescan in-house compound collections quickly and identify those that have patent, reaction, and/or other literature data in the CAS database. However, the lack of a programming interface capable of searching thousands of HTS hit candidates against the CAS catalog limits the ability to use this database for hit-to-lead analyses.

There are currently 394 508 PubChem structures that have CAS registration numbers. Among them, 121 509 PubChem structures have no direct annotation other than CAS registration numbers; these are not designated as class-A compounds since the type of annotation available in the CAS database is not clear. However, the data are included in our resultant database, because the CAS registration numbers enable us to search these structures in SciFinder in a much more convenient manner than using chemical structures one at a time. For these compounds, a batch search using SciFinder in a semiautomated fashion is possible.

Comparative Analysis of Annotation Sources. Based on our estimations, approximately 175 000 class-B compounds within DiscoveryGate have nontrivial annotations, and about 500 000 class-C compounds may have nontrivial annotations in the CAS database. Without automated access to the CAS database, it is difficult to determine how many structures of the 26 million in the CAS catalog actually provide useful annotations. This is a relevant question, since simply finding a structure in the CAS catalog and obtaining a CAS registration number for a structure provide little help for understanding the biological and pharmacological characteristics of the compound. Indeed, we have seen that as many as 80% of the DiscoveryGate compounds that are present in the CAS database have no annotation (Figure 5). We attempted to explore this issue based upon statistical interpolations.

PubChem consists of 4% class-A, 32% class-B, and 64% class-C structures. The polling analysis shows these classes have a 96%, 19%, and 10% chance of finding literature data in the CAS database, respectively (Figure 5). Therefore, a structure randomly selected from the PubChem collection, would have a 16% probability of containing literature annotations (4% $\times$ 96% + 32% $\times$ 19% + 64% $\times$ 10%). Similarly, we estimated the probability of obtaining patent and reaction information for a randomly selected PubChem structure to be 5% and 8%, respectively. Since the structures in PubChem were not collected with any obviously biased filters, we hypothesize that the chemical space covered by PubChem and the 26 million structures collected by the CAS database to be similar. It is then straightforward to extrapolate that the CAS database contains literature annotations for 4.2 million structures, patent annotations for 1.3 million structures, and reaction data for 2.1 million structures. Considering there are only 0.4 million class-A PubChem compounds, the knowledgebase provided by PubChem is less than 10% of the size of the CAS database. Although these numbers are only estimations, it qualitatively demonstrates that the two knowledgebases, PubChem and CAS, presently are complementary.

Large-Scale Annotation of GNF Compounds. The key value of the annotation pipeline described above at GNF has been from the integration of external annotation databases with the existing infrastructure, i.e., cross-linking the annotated structures to in-house compound collections. Applying exact structure search analyses (ignoring stereochemistry) to the entire 8 million plus PubChem structures, we found annotation for 747 711 compounds out of the 2.5 million compounds situated at GNF (30%), among which 28 924 (0.12%) were classified as class-A compounds. For these class-A structures, we carried out a structure similarity expansion using a Tanimoto similarity threshold of 0.95 to increase the number of matched structures to 103 624 compounds (4%). ChemAxon Java libraries were used in our in-house program for fingerprint calculation and related structure searching algorithms (http://www.chemaxon.com). A conservative similarity threshold of 0.95, instead of the commonly used threshold of 0.85 for SAR analysis, was applied in order to increase the likelihood that the associated annotation would remain relevant for these related compounds.[10]

Taking into account additional annotation sources, as many as 32% of GNF compounds are hyperlinked to the annotations obtained using the automation pipeline described above as the result of this study. As PubChem is evolving into the official structural depository for publishers such as Prous and Nature Chemical Biology as well as new initiatives such as the Chemistry Central (http://www.chemistrycentral.com/) in open-access publications, there is a strong reason to believe that the extent of annotation available for compound libraries will expand in the future. GNF (along with the pharmaceutical industry in general) will likely benefit more from these open-access databases.

Annotation-Driven HTS Data Analysis. Yan et al. previously proposed an ontology-based pattern identification (OPI) algorithm, which first clusters compounds into structurally similar families and then identifies statistically significant biological selectivity patterns shared by a majority of the members of a compound family (i.e., structure−profile relationship).[5,11] This approach successfully identified compound scaffolds of general cytotoxicity, differential tumor cytotoxicity, and potential report gene assay artifacts as well as inhibitors for specific protein target families. Our annotation database provides an alternative way to group compounds based upon their mechanisms of action. In this way, we will be able to associate a biological activity profile directly to a putative mechanism of action.

Among the 33 107 substances with biological activities identified from the 74 HTS assays previously described by Yan et al.,[5] we identified 356 structures that belong to 67 MeSH pharmacological action categories. We then applied the OPI algorithm to all of these 67 MeSH terms individually to search for statistically significant signature patterns characterizing the corresponding pharmacological action. Our results show that 30 out of 39 structures associated with the MeSH term, “82000893 (anti-inflammatory agents)”, share a unique biological profile (Figure 6A). Since the *p*-value of such a coinhibition pattern is $10^{-29.8}$ and only 5 out of the remaining 317 structures share this pattern (false positives), this would suggest that this pattern of activity is indeed a biological fingerprint for anti-inflammatory agents. In this example all 30 structures are steroids. The nine outliers include nonsteroidal anti-inflammatory agents, such as CID4674, CID2466, and CID2396, along with a stereoisomer to CID343585 (Figure 6B).

The MeSH annotation is extremely useful when structurally distinct compounds apparently share the same biological activity profile. Indeed, 4 out of 6 structures associated with MeSH 82000972 (antineoplastic agents, phytogenic) show a unique profile with a *p*-value of $10^{-7.6}$ (Figure 6C). These four compounds are structurally very different comparing CID10607 and CID13342. A similar finding was also observed for CID83979 and CID446541 under MeSH term 82000903 (antibiotics, antineoplastic) with a *p*-value of $10^{-8.6}$ (Figure 6D). It is interesting to point out that even though the difference between the two profiles representing MeSH 82000972 and MeSH 82000903 appear subtle, the OPI algorithm is capable of distinguishing their dissimilarities.

A biological “fingerprint” also was found for adrenergic beta-agonists (MeSH 82000318), having a *p*-value of $10^{-13.6}$ (data not shown). Therefore, the annotation database obtained was able to guide in-house data mining efforts using signature patterns that assign a potential mechanism of action annotations to compounds. It can also serve as a discovery tool to find a novel scaffold that bears little structural similarity to a known scaffold in a drug discovery project (i.e., scaffold-hopping). Considering the instrumental role the gene ontology database has been playing for system biology studies, it is reasonable to expect a compound mechanism of an action database will lead to more interesting cheminformatics discoveries in the near future.

Annotation-Driven Antimalarial Assay Validation. “Biological validation” can be ascribed to an HTS assay when the screen successfully identifies compounds that have a described mechanism of action consistent with the intended goals of the screen. For example, researchers at GNF recently performed a 1.6 million compound screen designed to provide lead compounds for the development of novel antimalarials. We first searched the WDI database and found only three “antimalarials” available in our compound collection, namely, quinacrine, chloroquine, and primaquine. We then searched the annotation database with the MeSH heading “Antimalarials” (MeSH ID: 82000962) and found 141 PubChem structures, among which 46 unique structures were available at GNF.

These 49 antimalarials were tested at 1 μM concentration in the antimalarial cell-based assay. Twenty-six structures showed more than 50% inhibition of parasite invasion, a value that would be designated as a “hit”. The hit rate of 51 ± 7% is highly statistically significant (*p*-value of 10^{-10}), considering the same compounds had an extremely low hit rate (<5%) in 125 previously performed cell-based screens at GNF. This suggested that the selected compounds lacked general cellular cytotoxicity that often leads to a “false positive” signal in cell-based assays. The specificity of the activity associated with this set of compounds provided validation for the screening format and confidence that the unannotated hit compounds were of biological importance in the search for novel antimalarials.

Even though almost half of the previously described antimalarials did not score as hits, further inspection of the annotation linked to these compounds provided an explanation. For example, many of the compounds lacked sufficient potency at the concentration used in the assay to be expected to score as a positive. Other examples include compounds only effective in combination therapy or in their metabolized forms. Therefore, additional examinations are required to

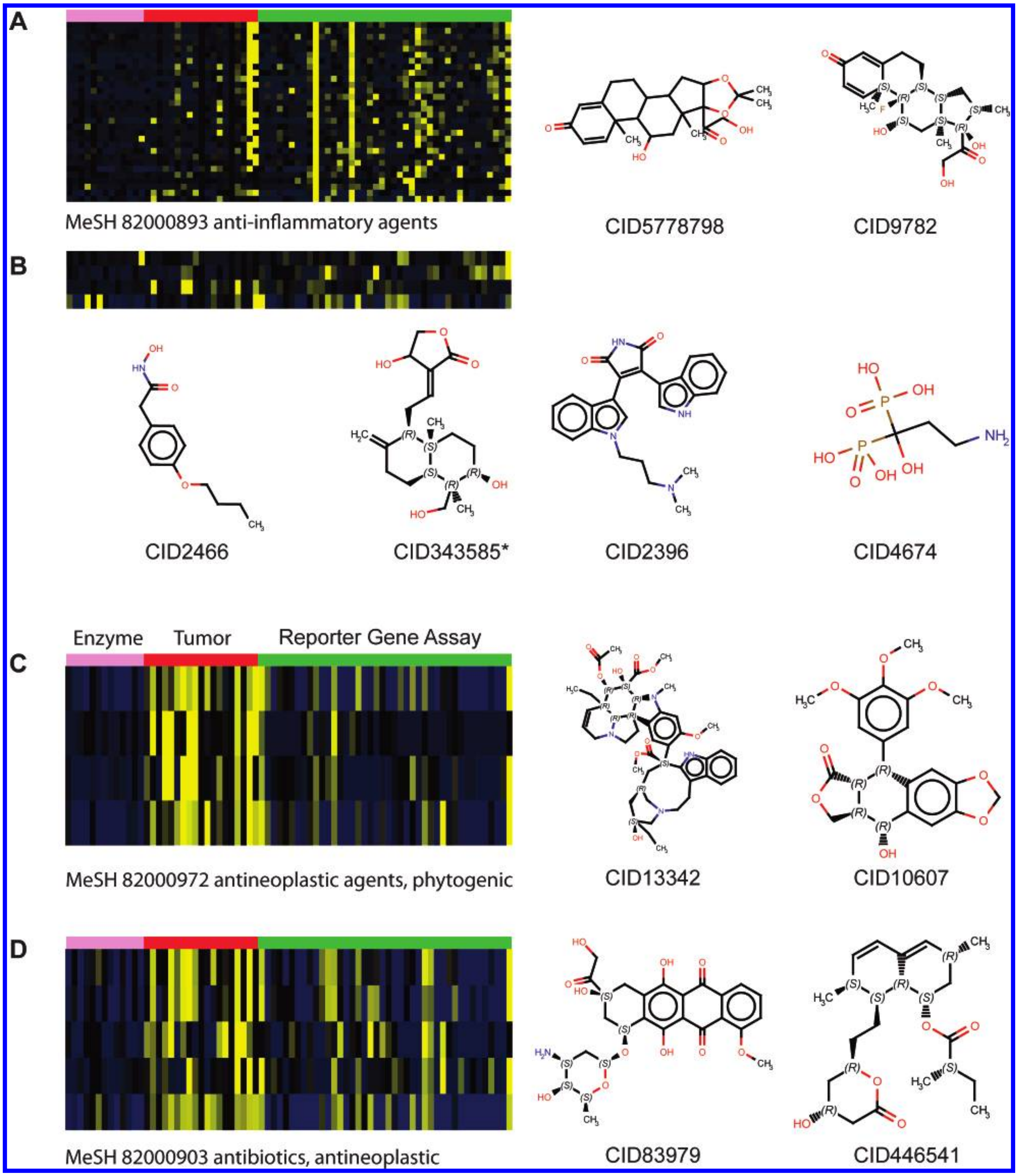
A
MeSH 82000893 anti-inflammatory agents
CID5778798
CID9782
B
CID2466
CID343585*
CID2396
CID4674
Enzyme
Tumor
Reporter Gene Assay
C
MeSH 82000972 antineoplastic agents, phytogenic
CID13342
CID10607
D
MeSH 82000903 antibiotics, antineoplastic
CID83979
CID446541

Figure 6. MeSH ontology-driven analysis of GNF HTS databases discovers signature biological profiles for several pharmacological actions. (A) 30 out of 39 anti-inflammatory agents are steroids, and they all share a unique HTS profile. (B) The HTS profiles for nonsteroid anti-inflammatory agents are rather different, which indicates that in-house HTS data can be used to provide granular resolution to the MeSH classifications. (C) Structurally distinct compounds (two are shown as examples) can be found to share a common unique HTS signature pattern for phytogenic antineoplastic agents. This signature pattern would not have been discovered, if the compounds had been grouped based on chemical structural similarity rather than MeSH annotation. The signature pattern is meaningful considering these anticancer agents tend to show antiproliferative activities in GNF tumor cell viability assays but are largely inactive in enzymatic assays and reporter gene assays. Detailed methods have been described in a previous study.[5] (D) Unique HTS signature pattern for antibiotics, antineoplastic agents based on MeSH ontology (two structures are shown as examples). Notice that the small difference between profiles in (C) and (D) is sufficient to distinguish the two related pharmacological actions.

identify the true negatives from these compounds. Nevertheless, being able to provide valuable chemical structures for assay validation in a straightforward manner offered significant time-savings that are otherwise not available.

Analysis of Bioassay Annotations. Biological, pharmacological, and measured physicochemical properties are also valuable forms of annotation that can offer insights into a compound's efficacy, selectivity, toxicity, and ADME properties. Our analysis shows at this early stage that PubChem provides 4.3 million records of experimental measurements for about 378 371 substances; however, most of them are from the legacy NCI60 data sets (http://dtp.nci.nih.gov/branches/btb/ivclsp.html). Currently, there are only 65 854 compounds in the compound collection owned by MLSCN that are being studied by the ten NIH small-molecule screening centers. Our study shows 32 452 structures out of this collection already exist in the 2.5 million GNF compound library; therefore, the additive value of the bioassay results from PubChem is not significant at its current stage. We expect both the NIH screening collection and PubChem BioAssay database to grow quickly as the centers start to operate at full capacity. As the amount of biological activity measurements grows, the utility associated with the integration of PubChem with an in-house drug discovery database will expand.

GNF Public Compound Annotation Web Tools. A Web site has been set up (http://carrier.gnf.org/publications/PubChem) that allows the general research community to apply the search tools described above. Upon user submission of a list of compound structures, the Web service carries out structure similarity searches using a ChemAxon chemical cartridge and creates a spreadsheet report for all of the annotations identified. The regularly updated GNF chemical compound knowledgebase can also be obtained, which can provide an important addition to an organization's existing infrastructure as illustrated in this study.

CONCLUSIONS

PubChem provides a valuable research tool that has linked small molecules to the complex biomedical annotation network that has been described by the biomedical community over the past decade. Its role in serving as the central catalog repository for chemical vendors and content providers not only enables clients to better evaluate the add-on value of commercial compound libraries but also encourages more convenient entry points into commercial databases for their customers. We hope more content providers will contribute their catalogs to PubChem and enables licensees to prescan their in-house chemical collections. This would save biomedical researchers from the laborious effort associated with manually searching commercial databases. Our analyses also demonstrate that commercial parties such as CAS could benefit from the incorporation of additional open-access chemical and biological data found in PubChem.

MeSH-driven HTS data analysis and antimalarial assay validation provide two examples made feasible by our annotation database. It is our opinion that similar applications would be rather laborious to carry out otherwise. Therefore, the informatics pipeline described here serves as a starting point for the community involvement. Besides CAS, several companies are working on solutions to better annotate chemical-related literatures, especially making the rich patent repository more structure-search-friendly. The particular implementation presented here is only as validated as the underlying data source and therefore is open for debate. The fact that small compound annotation is a relatively less studied field, compared to that of biological macromolecules, provides opportunities for creative solutions. Therefore, we expect any such implementation should be continuously reevaluated and polished as new results are made available.

We demonstrated that the PubChem database is already capable of providing an extensive level of annotation information that, even at its early stages of development, should not be neglected by biomedical researchers. The bioassay data generated by the NIH Roadmap initiative will likely grow significantly in the near future, and PubChem will play an increasingly important role in disseminating these data sets in a standard format that provides facile data integration. Our study also suggests that the PubChem-centered annotation database already covers a significant portion of compound collections found in a typical pharmaceutical company. The hit-to-lead decision making process in drug discovery programs can greatly benefit from such an automated batch annotation service.

ACKNOWLEDGMENT

We would like to thank James Hasse and the PubChem team for providing support during this study. We also want to thank ChemAxon, Inc. for generously providing us the licenses for the structure drawing and searching tools on our public compound annotation Web service.

REFERENCES AND NOTES

(1) Baker, D. B.; Horiszny, J. W.; Metanomski, W. V. History of abstracting at Chemical Abstracts Service. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 193−201.
(2) Cheeseman, E. N. IDdb3: pharmaceutical intelligence for the third millennium. *Online* **2002**, *26*, 44−49.
(3) Huang, R.; Wallqvist, A.; Covell, D. G. Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen. *Genomics* **2006**, *87*, 315−328.
(4) Feldman, H. J.; Snyder, K. A.; Ticoll, A.; Pintilie, G.; Hogue, C. W. A complete small molecule data set from the protein data bank. *FEBS Lett.* **2006**, *580*, 1649−1653.
(5) Yan, S. F.; King, F. J.; He, Y.; Caldwell, J. S.; Zhou, Y. Learning from the data: mining of large high-throughput screening databases. *J. Chem. Inf. Model.* **2006**, *46*, 2381−2395.
(6) Richard, A. M.; Gold, L. S.; Nicklaus, M. C. Chemical structure indexing of toxicity data on the internet: moving toward a flat world. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 314−325.
(7) Murray-Rust, P.; Mitchell, J. B.; Rzepa, H. S. Communication and re-use of chemical information in bioscience. *BMC Bioinformatics* **2005**, *6*, 180.
(8) Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; DiCuccio, M.; Edgar, R.; Federhen, S.; Geer, L. Y.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Ostell, J.; Miller, V.; Pruitt, K. D.; Schuler, G. D.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L.; Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2007**, *35*, D5−D12.
(9) Kaiser, J. Science resources. Chemists want NIH to curtail database. *Science* **2005**, *308*, 774.
(10) Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469−474.
(11) Zhou, Y.; Young, J. A.; Santrosyan, A.; Chen, K.; Yan, S. F.; Winzeler, E. A. *In silico* gene function prediction using ontology-based pattern identification. *Bioinformatics* **2005**, *21*, 1237−1245.

CI700092V