

CYANOS: A Data Management System for Natural Product Drug Discovery Efforts Using Cultured Microorganisms

George E. Chlipala, Aleksej Kronic, Shunyan Mo, Megan Sturdy, and Jimmy Orjala*

Department of Medicinal Chemistry and Pharmacognosy, University of Illinois at Chicago, 833 South Wood Street, Chicago, Illinois, United States

Received July 20, 2010

A software package, termed “CYANOS”, has been developed to facilitate the data management and mining for natural product drug discovery efforts. This system allows for the management of data associated with field collections, culture conditions, harvests, extractions, chemical separations, and biological evaluation. This software utilizes a MySQL database for data storage, which allows for reporting and data mining via third party tools. In addition, a Web-based interface was constructed to allow for multiuser access from a variety of desktop platforms. The code for this system is freely available and has been released under the Illinois Open Source license.

Chemoinformatics has been recognized as an important part of drug discovery efforts. A review by Brown¹ defined chemoinformatics as the process “to transform data into information and information into knowledge”. This knowledge is then used to improve decision making in a drug discovery effort. In particular, the two decisions that Brown emphasized were “what to test next” and “what to make next”. Typically, a chemoinformatic approach for a drug discovery effort will utilize the chemical structure data of the compounds in the screening library and bioinformatics of a target to produce the “information” and “knowledge” as defined by Brown.² For synthetic libraries, these chemical data are easily obtained since the structure of the compounds is defined when the library is created.

This is rarely the case for natural product drug discovery efforts, where the structural information of an active constituent is not known until after multiple rounds of bioassay guided fractionation. In addition, natural product samples are typically complex mixtures and not a single chemical entity. Techniques have been developed to short cut the process of bioassay guided fractionation and thereby decrease the time from initial extraction to when the structures are determined of the active components.^{3–6} While these techniques may speed the identification of active components, they often increase the amount of data generated for each sample analyzed. Thus, it is important to effectively manage the data generated, in order to improve decision-making through data mining and aid prioritization during the natural product drug discovery effort.

The use of simple files and spreadsheets, along with a directory structure standard, can simplify the data management of the various pieces of information. However, the process of searching and data mining becomes increasingly difficult, especially if the data are stored in a variety of file formats, e.g., Microsoft Excel, PDF, and plain text. It was with these challenges in mind that we began the development

of a data management system that would store and maintain data as well as allow for searching and data mining to improve data access and report generation for natural product drug discovery efforts.

Requirements. We established four requirements for the design of this data management system:

1. It must allow multiple researchers to work together and combine their data in real time.
2. The data should be easily accessible for generating reports and data mining.
3. The data schema and program functions should facilitate the dereplication of natural products.
4. The system should provide programming interfaces and a development library to allow further customization for specific deployments of the application.

In this article, we describe the logical design and give a high-level view of the implementation of this data system. Further details of the implementation and developer information can be found at <http://www.uic.edu/labs/orjala/cyanos>. The logical design of CYANOS included various entities and associated attributes, listed below, that are important in natural product drug discovery efforts. Figure 1 depicts the entity relationship diagram (ERD) for these objects.

LOGICAL DESIGN

Sample Information. At the heart of the data is the sample that is evaluated biologically and chemically. In a natural product drug discovery effort, there are three main classes of samples: crude extracts, fractions, and pure compounds. From a hierarchical standpoint, extracts are the children of harvested biomass and are the parent of fractions, via a chemical separation. Fractions in turn can be the parent of further fractions as well as purified compounds. In the case of all three classes, daughter samples can be created for storage libraries or sample submission for biological evaluation. A single sample class was defined that would contain common information for all samples, e.g., location, label, date, and amount. Specific attributes and relationships were then used to define the sample as a crude extract, fraction,

* Corresponding author. E-mail: orjala@uic.edu. Telephone: 1-312-996-5583.

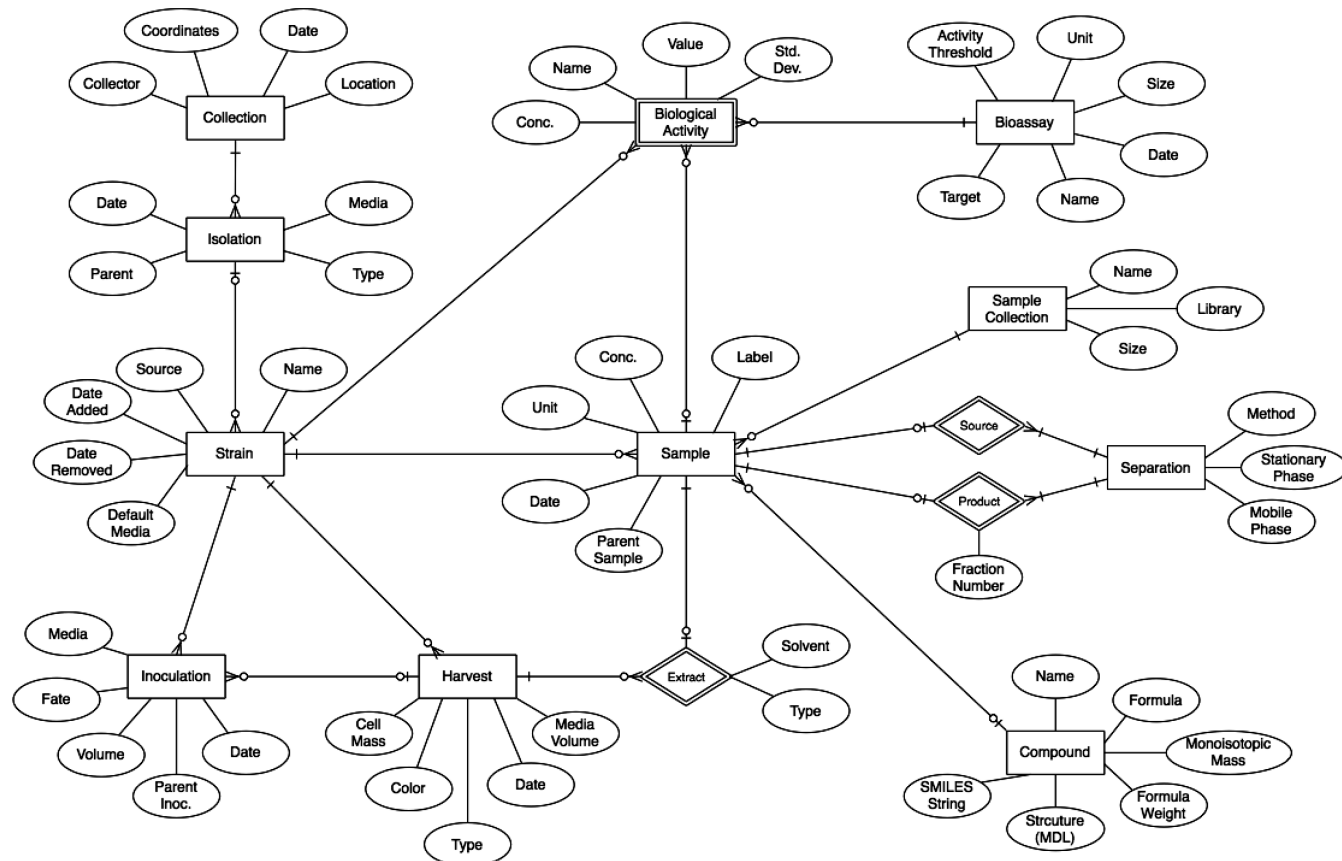


Figure 1. Overview of the CYANOS data objects and their relationships.

or pure compound. The source material and resulting fractions of a chemical separation were also linked to the separation record, which stored data relating to the chromatographic parameters. All samples had transaction histories, which linked daughter samples and allowed a view of the full history of any particular sample. Sample collections, either physical or logical, were defined to help organize samples, and these sample collections were then grouped into sample libraries.

Compound Information. In order to facilitate dereplication of compounds previously isolated, a data class was defined for compound information. A compound record would contain the key chemical attributes, i.e., formula weight, monoisotopic mass, molecular formula, SMILES string, and the 3D structure in MDL format. When a particular compound was purified from a sample, the relevant sample record was linked to the associated compound record. Since sample records were linked to other records in the database, e.g., strain and bioassay, one could retrieve relevant biological information for a particular compound.

Strain Information. For our project in particular, we were working with cultured cyanobacteria as the source for natural products. Thus, it was important to capture culture information, which included strain attributes (strain), culture conditions (inoculation), and harvest information (harvest). In addition, it was important to maintain all information related to the collection of biological material from the field as well as the strain isolation information for strains unique to our culture collection. This was implemented by creating separate records for collections and isolations, then creating relationships from collections to isolations, allowing iterative isolation records, and finally linking the isolation record to a

strain. Since biological diversity often leads to chemical diversity, it was important to maintain taxonomic information in the data management system to improve searching and reporting functions of the software. Taxonomic data, i.e., kingdom, phylum, class, order, and family, were separated from strain information but linked via the genus.

Biological Activity Data. Ultimately, biological activity is the driving force for any natural products drug discovery project. A single class was defined to store assay data. The assay record contained important attributes, e.g., date, target, and activity threshold, and separate records would store the data for samples evaluated, which included parent sample/strain, concentration, and activity value. The determination of active vs inactive was performed on demand using the stored attributes of the assay.

Occasionally, the data reported from an assay is indefinite, e.g., $IC_{50} < 3 \mu\text{g/mL}$. From a computing standpoint, a standard primitive, e.g., a floating point number or integer, could not properly store this value. To accommodate indefinite values, the sign, i.e., less than (<), greater than (>), or none, of a value would be stored along with the numerical value for any activity data point.

External Data. Aside from the basic attributes of each record, e.g., date, weight, source, and activity, it was also important to be able to link records to external data files, such as HPLC-UV chromatograms, mass, and NMR spectra. For samples, extracts, fractions, and compounds, one can link data files for HPLC chromatograms, MS, and NMR data. Separation records could be linked to chromatographic data, while strains could be linked to external URLs and photos. Assay records could be linked to raw data files and reports. For any of the external data files, any binary or text format

Table 1. Number of Strains Added to the Collection during the Specified Period, Grouped by Culture Source and Taxonomic Order

order	source					total
	CCALA	CCMP	SAG	UIC	UTEX	
Chroococcales	1	5	5	18	12	41
Pleurocapsales	—	1	5	2	4	12
Oscillatoriales	—	13	18	77	40	148
Nostocales	6	7	32	54	47	146
Stigonematales	2	—	7	—	6	15
total	9	26	67	151	109	362

was acceptable, and thus one could link proprietary, raw data, e.g., a Bruker FID or Agilent Chemstation datafile, or a standard representation, such as a PDF report or PNG/BMP graphic.

Project Management. A project class was defined to allow researchers to organize data based upon the associated project. The stored attributes of the project class was the project code, a short name or label, and a brief description of the project. The project would link the associated data via project codes stored in records of field collections, isolations, strains, inoculations, harvest, samples, separations, and bioassays.

REPORTING AND DATA MINING

The ultimate value of storing data in a relational database is the ability to rapidly access select data elements for report generation and data mining. A variety of third party tools, e.g., Crystal Reports and Microsoft Excel, are available to generate reports from and data mine using an SQL-based database. For simple reporting, we found that Microsoft Excel (via Pivot Tables) and OpenOffice.org (via DataPilot) worked well to generate query-based reports, which could be regenerated on a regular basis. Example reports from our database are shown in Tables 1 and 2. These simple reports were useful for creating status reports as well as tracking strains from the time they are added to the collection through cultivation, harvest, extraction, and biological evaluation.

Reports generated from the database are valuable for tracking the progress of a project, however greater value arises from data mining. Data mining can be defined as the process of elucidating relationships in a large data set, which has been indicated as an important part of chemoinformatics for drug discovery.⁷ A key relationship for natural product drug discovery efforts is that of biological diversity (taxonomy) and biological activity. The report query for Table 2 can be modified, as shown below, to report activity data and thus calculate hit rates for various taxonomic groups

Table 3. Example Data Mining Report for the Bioassay Hit Rate of Extracts, Grouped by Taxonomic Order

order	assay					
	20S proteasome	brine shrimp	H460	HT-29	MCF7	SF268
Chroococcales	34%	0%	5%	0%	0%	5%
Pleurocapsales	0%	0%	0%	0%	0%	0%
Oscillatoriales	21%	3%	4%	0%	2%	2%
Nostocales	21%	8%	11%	6%	9%	11%
Stigonematales	50%	18%	11%	22%	11%	11%
total	23%	5%	8%	3%	6%	7%

(Table 3). This query utilizes a custom SQL function, named ACTIVE, to determine in the stored value satisfies the activity threshold stored with the assay information. This function returns the number “1” for values that satisfy the activity threshold and “0” for values that do not. A report using DataPilot (OpenOffice.org) or Pivot Tables (Microsoft Excel) could use a count of culture IDs (s.culture_id) to determine the number of strains evaluated and the sum of the active field to reveal the number that produced an active extract.

```
SELECT s.culture_id, a.target, t.ord, MAX
(ACTIVE(ad.activity + ad.sign,
a.active_level, a.active_op)) AS "Active"
FROM assay_info a
JOIN assay ad ON (ad.assay_id = a.assay_id)
JOIN species s ON (ad.culture_id =
s.culture_id)
LEFT OUTER JOIN taxonomic t ON (s.genus =
t.genus)
GROUP BY s.culture_id, a.target;
```

The results from our data revealed that only the orders Nostocales and Stigonematales had active members in each assay. In the case of the whole cell assays; i.e., H460, HT-29, MCF7, and SF268, the hit rate for either of these two orders was at least twice that of any other taxonomic order. These results would indicate that members of Nostocales and Stigonematales, which comprise of all the heterocyte-forming genera, are a good source of biologically active extracts for these targets. Based upon these results, strain isolation efforts could be modified to ensure that any Nostocalean and Stigonematalean cyanobacteria found in field collections would become part of the culture collection.

This query is not limited to the taxonomic order of the strain. It would also be possible to report and group the activity data by other taxonomic groups, e.g., family or genus.

Table 2. Number of Unique Strains Assayed during the Same Time Period, Grouped by Assay Target and Culture Source Type^a

order	assay					
	20S proteasome	brine shrimp	H460	HT-29	MCF7	SF268
Chroococcales	24 (79%)	23 (83%)	19 (79%)	23 (78%)	19 (79%)	19 (79%)
Pleurocapsales	6 (50%)	6 (50%)	4 (75%)	3 (0%)	4 (75%)	4 (75%)
Oscillatoriales	84 (81%)	74 (77%)	50 (90%)	69 (83%)	50 (90%)	50 (90%)
Nostocales	97 (95%)	82 (95%)	74 (95%)	76 (93%)	74 (95%)	74 (95%)
Stigonematales	12 (100%)	11 (100%)	9 (100%)	8 (100%)	9 (100%)	9 (100%)
total	223 (87%)	196 (86%)	156 (91%)	179 (86%)	156 (91%)	156 (91%)

^a Each source type is further separated by taxonomic order. Number denotes total number of extracts evaluated with percent freshwater strains in parentheses.

Table 4. Example Data Mining Report for the Bioassay Hit Rate of Extracts, Grouped by Source^a

order	assay					
	20S proteasome	brine shrimp	H460	HT-29	MCF7	SF268
non-UIC	27% (168)	6% (150)	9% (97)	3% (135)	5% (97)	7% (97)
UIC	15% (84)	4% (70)	5% (63)	4% (76)	6% (63)	6% (63)
total	23% (252)	5% (220)	8% (160)	3% (211)	6% (160)	7% (160)

^a The number in parentheses denotes the number evaluated for each category.

Furthermore, it would be valuable to group activity data by culture medium or culture source. In the case of our data, we were interested to determine if strains acquired from outside collections had similar hit rate as those strains recently isolated in house. We had hypothesized that strains acquired from other collections would have a lower hit rate. We based this hypothesis on the premise that many of these strains have been cultured from numerous years and that the lack of competition or other selective pressures has resulted in a reduction of secondary metabolite production. The previous query was modified, as follows, and when the results were grouped by source, it was possible to calculate the hit rate for each source, i.e., UIC or non-UIC (Table 4).

```
SELECT s.culture_id, a.target, t.ord, MAX
(ACTIVE(ad.activity + ad.sign,
a.active_level, a.active_op)) AS "Active",
IF(s.culture_source LIKE 'UIC%', 'UIC',
'Non-UIC') AS 'Source'
```

```
FROM assay_info a
```

```
JOIN assay ad ON (ad.assay_id = a.assay_id)
```

```
JOIN species s ON (ad.culture_id =
s.culture_id)
```

```
LEFT OUTER JOIN taxonomic t ON (s.genus =
t.genus)
```

```
GROUP BY s.culture_id, a.target;
```

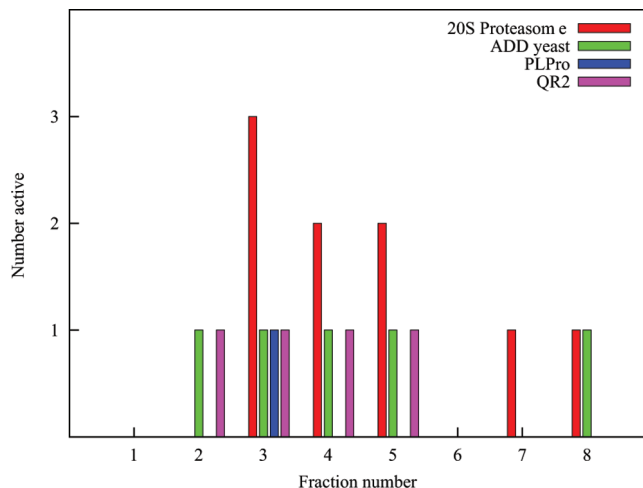
The results revealed that the strains acquired from other collections (non-UIC) displayed a hit rate similar to and in many cases greater than the hit rate displayed by strains recently isolated from the field (UIC). These results were contrary to our hypothesis and indicate that strains acquired from other collections can be as valuable as strains that are recently isolated from the field. Based upon these results, we have decided to continue to supplement our culture collection with strains acquired from other collections.

The data mining of biological activity is not limited to basic strain attributes, e.g., taxonomic order or source, but can also link chemical information. In particular, initial fraction data can be added to the database, and the query can be modified, as shown below, to relate the bioassay hit rates with fractionation values, e.g., fraction number. For this example, separation data tagged as "PREFRACTIONATION" were performed using a standardized protocol.

```
SELECT d.culture_id, f.fraction_number,
a.target
```

```
FROM assay d
```

```
JOIN assay_info a ON (d.assay_id =
a.assay_id)
```

**Figure 2.** Example data mining report of the bioassay hit rate of first round fractions grouped by target and fraction number.

```
JOIN separation_product f ON (f.sample_id
= d.sample_id)
```

```
WHERE active(d.activity + d.sign,
a.active_level, a.active_op) = 1
```

```
AND f.tag = 'PREFRACTIONATION';
```

The results of this query (Figure 2) displayed a small trend of activity for fractions 3–5. The report also showed activity associated with fractions 7–8, further analysis of these fractions revealed interference due to pigments. Based upon these results, we have modified our screening protocol to include an extra cleanup step for fractions 7–8 and thereby reducing the number of false positives due to interference. This evaluation is preliminary, since the number of active fractions is low ($N = 28$). As more assay data are generated and added to the database, the query can be re-evaluated easily. In addition, it would be possible to organize the data by other fields, so for example, one could evaluate and compare the fraction activity trends for strains in different taxonomic orders.

Results from the first data mining query (Table 3) revealed that extracts from cyanobacteria of the orders Nostocales and Stigonematales displayed the highest hit rate. Given this activity trend, we decided to evaluate our collection efforts for overall efficiency, i.e., the number of strains produced per collection as well as taxonomic diversity. In particular, we were interested in determining which months would be the best to collect cyanobacteria in the orders Nostocales and Stigonematales. Two queries were created for this data mining exercise. The first query was used to report the total number of collections for each month of a year, and the second was used to determine taxonomic information of strains isolated from those collections.

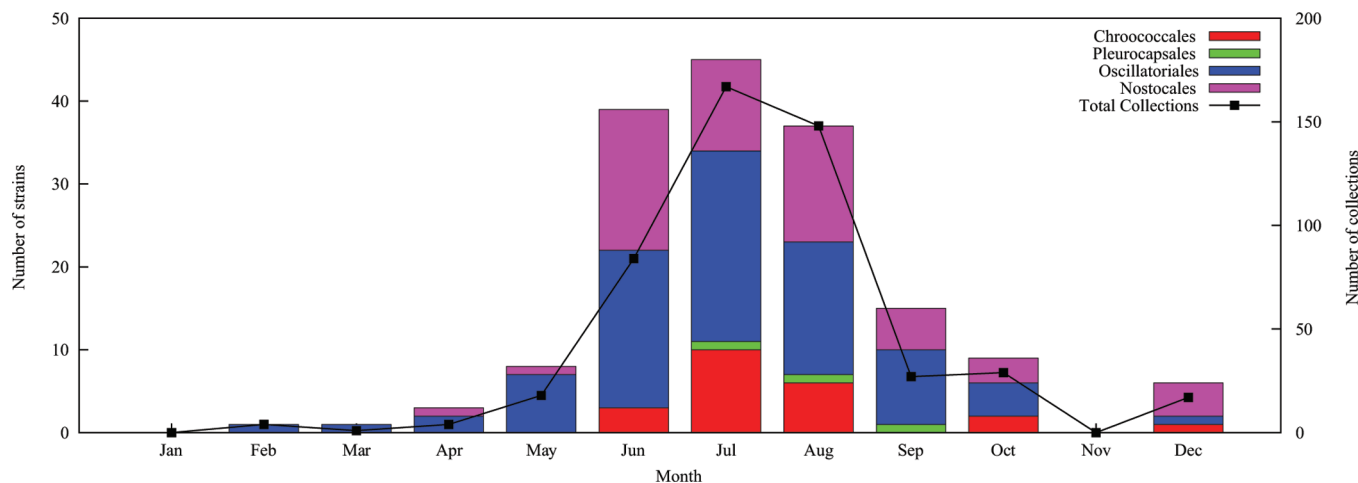


Figure 3. Example data mining report for strains isolated from field collections grouped by taxonomic order and month of collection. The black line displays the total number of collections for each month.

```
SELECT c.collection_id, MONTH(c.date)
FROM collection c
GROUP BY MONTH(c.date)
ORDER BY MONTH(c.date);
SELECT s.culture_id, MONTH(c.date), t.ord
FROM species s
JOIN isolation i ON (s.culture_source =
i.isolation_id)
JOIN collection c ON (c.collection_id =
i.collection_id)
JOIN taxonomic t ON (t.genus = s.genus);
```

The results from this query (Figure 3) revealed that the month of July was the most productive, in terms of total number of collections, however the month of June had a higher percentage of strains per collection (46% vs 27%). In addition, collections from the month of June produced higher percentage of strains in the order Nostocales (44% vs 24%) than any of the summer months, i.e., June, July, and August. Based upon some initial observations of material collected, we had hypothesized that early summer was dominated by green algae, e.g. *Spirogyra*, and that cyanobacteria would be more prevalent in the late summer. However, these results (Figure 3) have shown that early summer is just as important, if not more so, than the rest of the summer collection season. Based upon these results, we will now schedule our collection trips in the Midwest for both early and late summer.

Data mining is not limited to the examples shown here. The current database schema allows one to add chemical data, e.g., MDL data files and SMILES strings, and link these entries to sample records, which would then be linked to other records, e.g., strain and bioassay. The following example query would report compounds isolated from strains along with the taxonomic order and culture source.

```
SELECT s.culture_id, t.ord, IF(s.culture_
source LIKE 'UIC%', 'UIC', 'Non-UIC') AS
'Source', ch.name
FROM sample sa
```

```
JOIN species s ON (sa.culture_id =
s.culture_id)
JOIN compound ch ON (ch.compound_id =
sa.compound_id)
LEFT OUTER JOIN taxonomic t ON (s.genus =
t.genus);
```

The chemical data stored within the database is not limited to simple reporting of the data. Chemoinformatic functions and queries can be used to create data mining queries that directly utilize the stored chemical data. The following example uses the myChem⁸ package to allow matching of stored chemical data using a SMARTS query.⁹ The SMARTS query in the example would match compounds with an aldehyde moiety. Software packages, such as Marvin and iBabel, are available that can generate SMARTS queries using a drawn structure.^{10,11}

```
SELECT s.culture_id, t.ord, IF(s.culture_
source LIKE 'UIC%', 'UIC', 'Non-UIC') AS
'Source', ch.name
FROM sample sa
JOIN species s ON (sa.culture_id =
s.culture_id)
JOIN compound ch ON (ch.compound_id =
sa.compound_id)
LEFT OUTER JOIN taxonomic t ON (s.genus =
t.genus)
WHERE MATCH_SUBSTRUCT('[CX3H1](=O)',
ch.smiles) = 1;
```

Reporting Geospatial Information. In the previous data mining examples, results were used to generate report tables and categorical plots. The inclusion of position data, i.e., latitude and longitude, in collection records allows one also to visualize trends as a function of geographic location. The query for Figure 3 can be modified, as shown below, to report taxonomic diversity with associated geographic information. The results of this query could be imported into a geographic information system (GIS), such as GRASS or Quantum GIS.^{12,13} Once the data is accessible to a GIS, the results of the query could be plotted on a map, which would give one

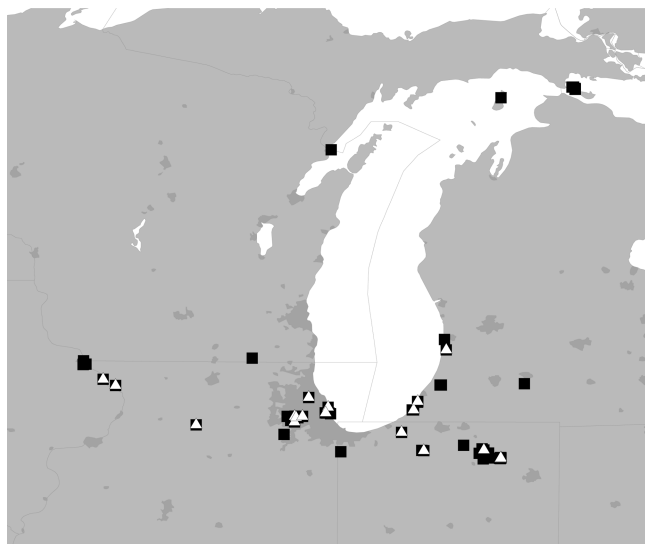


Figure 4. Map of collections sites in the Great Lakes region. All collection sites are marked with by a black square (■), and collections yielding cyanobacteria of the order Nostocales are overlaid with a white triangle (△). The map was generated using the GRASS GIS package¹² with Natural Earth vector map data.¹⁴

a view of taxonomic diversity in relationship to geographic location (Figure 4).

```
SELECT s.culture_id, MONTH(c.date), t.ord,
       c.latitude, c.longitude
```

```
FROM species s
```

```
JOIN isolation i ON (s.culture_source =
                    i.isolation_id)
```

```
JOIN collection c ON (c.collection_id =
                     i.collection_id)
```

```
JOIN taxonomic t ON (t.genus = s.genus);
```

The query can be further modified, as seen below, to report strains that are considered active in a biological assay and the associated assay target. The results of this query would then allow one to visualize biological activity as a function of the geographic location of the original collection.

```
SELECT DISTINCT s.culture_id, MONTH
(c.date), t.ord, c.latitude, c.longitude,
a.target
```

```
FROM species s
```

```
JOIN isolation i ON (s.culture_source =
                    i.isolation_id)
```

```
JOIN collection c ON (c.collection_id =
                     i.collection_id)
```

```
JOIN taxonomic t ON (t.genus = s.genus)
```

```
JOIN assay a ON (a.culture_id =
                 s.culture_id)
```

```
JOIN assay_info ai ON (a.assay_id =
                      ai.assay_id)
```

```
WHERE active(a.activity + a.sign,
ai.active_level, ai.active_op) = 1;
```

The query can also be modified to link chemical data. A simple query, as shown below, could return the name of the compounds isolated with associated geographic data.

```
SELECT DISTINCT s.culture_id, ch.name,
               c.latitude, c.longitude
```

```
FROM species s
```

```
JOIN isolation i ON (s.culture_source =
                    i.isolation_id)
```

```
JOIN collection c ON (c.collection_id =
                     i.collection_id)
```

```
JOIN sample sa ON (sa.culture_id =
                  s.culture_id)
```

```
JOIN compound ch ON (ch.compound_id =
                    sa.compound_id);
```

The results from the queries listed previously can also be used to visualize trends based upon taxonomic order and month of the collection as well as the geographic coordinates. All of the queries can be re-evaluated on a regular basis to provide updated results and thus simplifying the process of generating reports.

Dereplication. The ability to quickly identify known compounds is a very important part of natural product research. To aid this process, we developed two search engines, one to search using MS data and another utilizing ¹H NMR data. The MS search form (Figure 5) allows queries based upon an *m/z* value, a specified deviation (in Da or ppm), and a list of possible adducts (including a molecular ion, "M"). The MS search engine will then build a query that utilizes the stored monoisotopic mass values to find compounds that would satisfy the specified search parameters. This query would calculate the mass of selected adducts, thus eliminating the need to store values for all possible adducts.

While MS data can be very useful in identifying a compound, variability in both the ionization of compounds and accuracy of different spectrometers can obscure the true identity of the compounds of interest. On the other hand, ¹H NMR data can provide a clearer picture of the compounds in a sample and the structural features of those molecules. While it is possible to catalog and search NMR spectra directly,¹⁵ a simpler method is to utilize ¹H NMR data to determine the number of key moieties, e.g., singlet methyls and aromatic protons, and search against the chemical structure data rather than the NMR spectral data.¹⁶ This method of NMR-based dereplication has been successfully utilized by Lang et al. to dereplicate compounds using the Anti-Marin database.¹⁷ Following their example, the NMR search engine of CYANOS was designed to find compounds using key chemical moieties (Figure 5). For each moiety, a specific quantity or range can be specified, e.g. 1–3 or 2+. From an implementation standpoint, CYANOS utilizes the Mychem extension for MySQL,⁸ which allows the search engine to utilize SMARTS patterns to find compounds based upon SMILES strings stored in the database.

The CYANOS dereplication search engine was useful in identifying a known alkaloid from an extract of a *Hapalosiphon welwitschii*. LC-MS data revealed a compound with a major ion of *m/z* 305.2004 in positive-ion mode. The mass spectrum also revealed an ion of *m/z* 327.1833. The mass

☒ **Mass Spec**

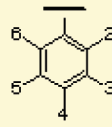
Positive Mode ± ppm

Adducts
M ☐ H ☐ Na ☐ NH₄ ☐ K ☐ Custom:

☒ **NMR Data**

Methyl
Triplet (-CH₂-CH₃) Doublet (-CH-CH₃) Singlet (-C-CH₃) -O-CH₃ -N-CH₃

sp² Carbons
-C=CH₂ -CH=CH -CH=CH₂

Aromatic Rings
Aromatic Protons Substitutions ☐ unsubstituted
☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6


Special
-C-CHO

Figure 5. Screen shot of CYANOS dereplication search forms.

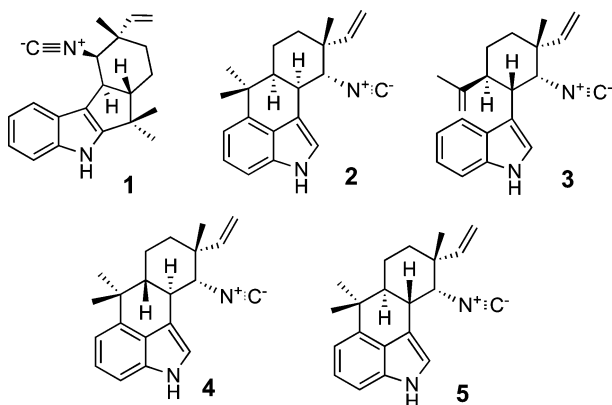


Figure 6. Structures of compounds from the dereplication search results.

difference of 22 indicated that the first ion, m/z 305.2004, corresponded to the $[M + H]^+$ adduct, and the second ion was the $[M + Na]^+$ adduct. A search of either of these two ions, returned a list of five compounds (1–5) (Figure 6). The 1H NMR spectrum indicated the presence of a 1,2-disubstituted aromatic ring and adding these parameters to the search narrowed the list to two compounds (1 and 3). The 1H NMR spectrum also revealed only four aromatic signals and adding this parameter to the search reduced the list to a single compound (1, fischerindole U isonitrile).¹⁸

Ultimately, both the MS and NMR moiety search engines in CYANOS are only as useful as the data stored in the database. To aid the creation of compound data, CYANOS utilizes the Chemistry Development Toolkit (CDK)¹⁹ to generate the molecular formula, formula weight, monois-

topic mass, and SMILES string of a compound using chemical structure data from an uploaded MDL file. This ability to extract and store chemical information from a MDL file has simplified the process to add compound data to CYANOS, and subsequently, we have employed it to build a focused database of known compounds from cyanobacteria. Also, it is important to note that CYANOS and its dereplication search engine were not designed to replace other chemical databases, e.g., SciFinder and Anti-Marin,^{17,20} but rather to provide a simple search engine for unpublished or proprietary compounds that would not be found in commercial or third party databases.

IMPLEMENTATION DETAILS

The system was implemented using a third party relational database management system (RDBMS). This provided the advantage of the possible integration of various existing third party tools that could aid the administration, management, searching, and data mining of the information stored in the database. A variety of RDBMS are available from different software companies, e.g., IBM DB2, Oracle, Microsoft SQL Server, and MySQL, with each providing robust storage, indexing, and searching functions. We chose to start development with MySQL v5.0, since the RDBMS and development libraries were freely available.²¹

A Web-based interface was developed to provide the multiuser access, as specified by requirement 1. We utilized Apache Tomcat v5.5²² as the Web application server. For software development, we utilized version 1.5 of the Java Development Kit (JDK)²³ from Sun Microsystems, Santa

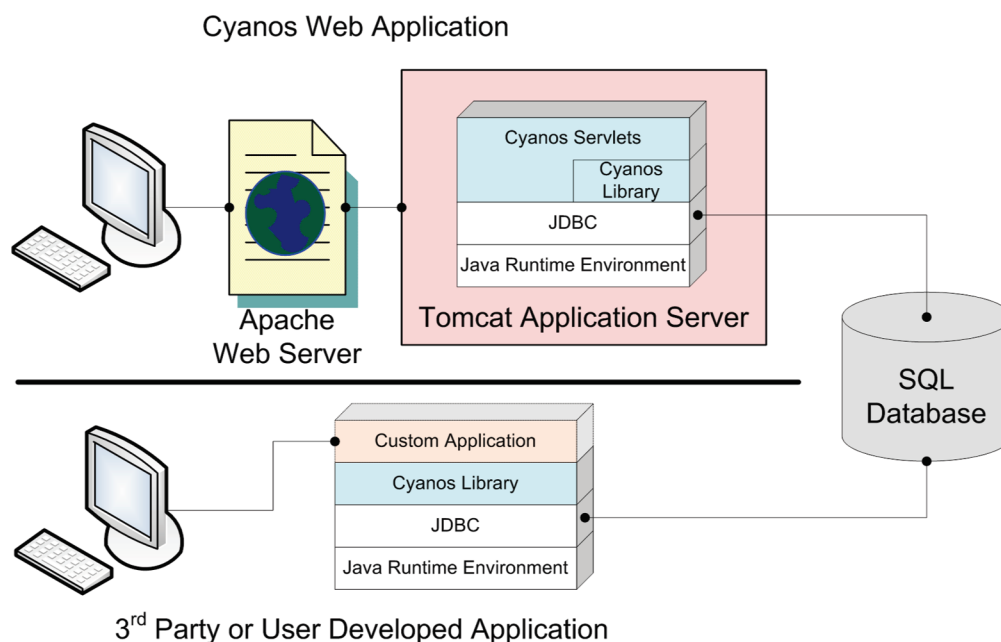


Figure 7. Detailed view of the implementation of CYANOS showing the currently developed Web-based application (top) and a hypothetical custom application (bottom).

Clara, CA, and the Eclipse SDK v3.2.²⁴ In addition, the software for the data management objects were separated from the Web interface code, in essence providing a software library that could be used to develop or integrate other applications (Figure 7). The schema SQL code, Web application, and development library for CYANOS have been released under the Illinois Open Source License.²⁵

Record Identifiers. From an entity relationship standpoint, many of the objects, e.g., inoculation, sample, separation, and harvest, could be weak entities (Figure 1). A weak entity is defined as an object where the identifier (primary key) is a combination of the parent object's key and a secondary key. For example, an inoculation could be identified by the strain ID and date of the inoculation. While this system of identification is valid, it can make data management and cross referencing difficult. Also, if the keys are not chosen carefully, then it is possible to have multiple instances of the same class share the same key, e.g., multiple inoculations of one strain on the same date. To avoid many of these pitfalls, the main objects in CYANOS, i.e., strain, field collection, isolation, bioassay, inoculation, sample, separation, and harvest, were defined as strong entities in which each entity has their own independent, unique key.

In the CYANOS database strains, field collections, isolations, and bioassays have a manually entered, alphanumeric identifier with a maximum length of 32 characters. For inoculations, harvests, samples, and separations, the MySQL database server automatically generated the identifier in the form of a serial number for each new record created. Each of the "serial number" classes, i.e., inoculations, harvests, samples, and separations, maintained their own sequence of serial numbers, thus a number itself was not unique to a particular class. The serial number field was defined as an unsigned 64-bit integer and allows numbers from 0 to $2^{64} - 1$ (approximately 1.84×10^{19}).

Data Entry and Collaboration. The developed Web application allowed for data entry through the use of online Web forms or the upload of Microsoft Excel XML or

OpenOffice.org spreadsheets. The online forms were well suited to low-throughput tasks, such as creating a strain, making a harvest, or creating an extract, while spreadsheet uploads allowed for the quick entry of results from biological evaluations or the creation of sample libraries and separations. Given that the Web site was accessible to anyone with a suitable Web browser, tasks were efficiently delegated to members of the team. For example, the culture specialist in our research group was responsible for adding new strains, inoculating growths, and harvesting the material. Using this system the culture specialist was able to enter relevant information into the database, and other members of the group were able to view this data and associate these records with extracts, chemical separations, and biological activity data.

Online Access and Security. The Web-based interface made the database instantly accessible by any computer via a Web browser. This "instant access" was crucial in promoting the sharing of data among members of the project, however this could also allow unauthorized users access to our data. The application code and the associated database do not, in themselves, require access to the Internet. If remote access is not required, then one could setup the system as standalone or place the system on a network with no access to the Internet. In the case of other, less secure networks, e.g., the Internet, the Web application is compatible with existing network firewalls and standard network encryption technologies, e.g., SSL and VPN. Our installation of CYANOS, utilizes both a firewall to control traffic and SSL to protect data transmission between Web browsers and the server. In addition to these basic network security procedures, we have implemented a simple user/password authentication and authorization system. A set of basic roles is defined to restrict access to various parts of the data, e.g., culture vs bioassay vs sample, with permission bits to refine access for read, write, create, and delete actions. In addition, we setup a project specific authorization system to allow one the ability to restrict roles and permissions based upon the associated

project of a specific record. Additional details of the authorization system can be found in the product documentation on our Web site (<http://www.uic.edu/labs/orjala/cyanos>).

System Requirements. CYANOS can be installed on any system that supports the required services, i.e., Apache Tomcat v5.5 with Java v1.5 and MySQL v5.0.^{21–23} There are no strict hardware requirements for CYANOS, other than the ability to run the previously mentioned services. The instance of CYANOS used for this publication was installed on a computer system with dual 800 MHz Intel Pentium III processors and 512 MB of RAM. The operating system for this machine was Fedora Core 6, although we have successfully installed CYANOS on Windows XP and Mac OS X systems. The CYANOS Web application binary requires 13.8 MB of disk space, and our database currently utilizes approximately 10 MB of disk space, which contains 235 assay records with 9966 associated activity data points, 4556 sample records, 373 separation records, 472 strain records, 844 collection records, 1359 isolation records, 2991 inoculation records, and 730 harvest records as well as user and project records. This system hosted both the database and Web-interface portions of CYANOS and was more than sufficient for a group of nine users. This system also provided file and print share services for our group, thus CYANOS can be installed on an existing network server.

CYANOS has additional software dependencies, which include JavaMail, CDK v1.0.4,¹⁹ Apache Commons Fileupload v1.2,²⁶ Apache Lucene v3.0.0,²⁷ and Jericho HTML Parser v3.1.²⁸ In addition, the NMR dereplication search engine requires Mychem extension for MySQL.⁸ All distribution packages include Apache Lucene and Jericho HTML Parser in the CYANOS Web application binary, thus it is not necessary to install these requirements separately. Other distribution packages are available that also include CDK and Apache Commons Fileupload in the CYANOS Web application binary. It is important to note that all of these software packages have separate licenses, and even though the components may be included in the CYANOS application binary, their use is still covered by their respective license.

CONCLUSIONS

Chemoinformatics has become an important part of drug discovery efforts. Typically, these informatic systems require detailed knowledge of the compounds to be screened. For many natural product drug discovery efforts this would create an “informatic gap” since these efforts often culminate with the identification of compounds, rather than begin with this vital chemical information. We found that CYANOS could fill that gap by allowing us to manage data from field collection, through isolation, strain deposition, growth, harvest, extraction, and multiple rounds of fractionation with associated biological evaluation to the ultimate purification of a compound. It is important to note that once the compound has been purified, a traditional chemioinformatic system could be utilized to manage the drug development effort.

The multiuser nature of the CYANOS system also allowed for real-time data collaboration. This allowed the delegation of various tasks but at the same time allowed each member to record the data in a central repository so that it would be accessible to other members of the team. Thus, when a strain was assigned to a member of the team for isolation and

structure elucidation work, that member would have quick access to the data regarding the source, previous growths, initial extracts, and biological evaluations. We also found that CYANOS would allow members of the team to easily keep track of their work, e.g., number of crude extracts screened during a time period and to be able to detect any potential gaps, e.g., crude extracts that have yet to be evaluated in the various assays. The ability to unify data from various members of the team has been a great asset, which has allowed CYANOS to show its value as a platform for reporting and data mining. This has allowed researchers in our group to effectively manage their data so that they could more readily transform their data into information, which has led to improved decision making.

ACKNOWLEDGMENT

We thank D.L. Lantvit and Dr. S.M. Swanson from the University of Illinois at Chicago (UIC) for 20S proteasome data; H.Y. Kim from UIC for brine shrimp toxicity data; and Drs. V. Grum-Tokars, S.D. Pegan, and A.D. Mesecar from UIC for the QR2, PLpro, 3CLPro, and ADD yeast data. We also thank Y. Nakanishi from the Research Triangle Institute for providing MCF7, SF268, and H460 cytotoxicity data and Dr. H. Chai of The Ohio State University for providing HT-29 cytotoxicity data. This research was supported by NIH grants R01 GM0758556 and P01 CA125066.

Supporting Information Available: Diagram of CYANOS MySQL schema and SQL queries for reports in Tables 1 and 2. This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Brown, F. K. Chemoinformatics: What is it and How does it Impact Drug Discovery. In *Annual Reports in Medicinal Chemistry*; Bristol, J. A., Ed.; Academic Press: London, U.K. 1998; Vol. 33, pp 375–384.
- (2) Hrib, N. J.; Peet, N. P. Chemoinformatics: Are We Exploiting this New Science?: ‘We Need to Make Chemoinformatics Tools More Accessible to the Bench Chemist’. *Drug Discovery Today* **2000**, *11*, 483–485.
- (3) Clarkson, C.; Staerk, D.; Hansen, S. H.; Jaroszewski, J. W. Hyphenation of Solid-phase Extraction with Liquid Chromatography and Nuclear Magnetic Resonance: Application of HPLC-DAD-SPE-NMR to Identification of Constituents of *Kanahia laniflora*. *Anal. Chem.* **2005**, *11*, 3547–3553.
- (4) Manly, C. J. Managing Laboratory Automation: Integration and Informatics in Drug Discovery. *J. Automat. Meth. Manag. Chem.* **2000**, *6*, 169–170.
- (5) Ng, J.; Bandeira, N.; Liu, W. T.; Ghassemian, M.; Simmons, T. L.; Gerwick, W. H.; Linington, R.; Dorrestein, P. C.; Pevzner, P. A. Dereplication and De Novo Sequencing of Nonribosomal Peptides. *Nat. Methods* **2009**, *8*, 596–599.
- (6) Bugni, T. S.; Richards, B.; Bhoite, L.; Cimbor, D.; Harper, M. K.; Ireland, C. M. Marine Natural Product Libraries for High-Throughput Screening and Rapid Drug Discovery. *J. Nat. Prod.* **2008**, *6*, 1095–1098.
- (7) Claus, B. L.; Underwood, D. J. Discovery Informatics: Its Evolving Role in Drug Discovery. *Drug Discovery Today* **2002**, *18*, 957–966.
- (8) Pansanel, J.; De Luca, A.; Gruening, B. *Mychem3*, version 0.6.0; SourceForge.net: Mountain View, CA, 2008.
- (9) SMARTS - A Language for Describing Molecular Patterns; Daylight Chemical Information Systems, Inc.: Laguna Niguel, CA; <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>. Accessed November 2, 2010.
- (10) *MarvinSketch*, version 5.3.8; ChemAxon: Budapest, Hungary, 2010.
- (11) *iBabel*, version 2.2; SourceForge.net: Mountain View, CA, 2006.
- (12) GRASS Development Team; *Geographic Resources Analysis Support System (GRASS GIS) Software*, version 6.4.0; Open Source Geospatial Foundation: Vancouver, BC, Canada, 2008.

- (13) Quantum GIS Development Team. Quantum GIS Geographic Information System; Open Source Geospatial Foundation: Vancouver, BC, Canada, 2009.
- (14) Natural Earth; North American Cartographic Information Society: Milwaukee, WI; <http://www.naturalearthdata.com/>. Accessed on May 13, 2010.
- (15) Steinbeck, C.; Kuhn, S. NMRShiftDB -- Compound Identification and Structure Elucidation Support Through a Free Community-built Web Database. *Phytochemistry* **2004**, *19*, 2711–2717.
- (16) Bradshaw, J.; Butina, D.; Dunn, A. J.; Green, R. H.; Hajek, M.; Jones, M. M.; Lindon, J. C.; Sidebottom, P. J. A Rapid and Facile Method for the Dereplication of Purified Natural Products. *J. Nat. Prod.* **2001**, *12*, 1541–1544.
- (17) Lang, G.; Mayhudin, N. A.; Mitova, M. I.; Sun, L.; van der Sar, S.; Blunt, J. W.; Cole, A. L. J.; Ellis, G.; Laatsch, H.; Munro, M. H. G. Evolving Trends in the Dereplication of Natural Product Extracts: New Methodology for Rapid, Small-scale Investigation of Natural Product Extracts. *J. Nat. Prod.* **2008**, *9*, 1595–1599.
- (18) Stratmann, K.; Moore, R. E.; Bonjouklian, R.; Deeter, J. B.; Patterson, G. M. L.; Shaffer, S.; Smith, C. D.; Smitka, T. A. Welwitindolinones, Unusual Alkaloids from the Blue-Green Algae *Hapalosiphon welwitschii* and *Westiella intricata*. Relationship to Fischerindoles and Hapalinodoles. *J. Am. Chem. Soc.* **1994**, *116*, 9935–9942.
- (19) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an Open-source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *2*, 493–500.
- (20) SciFinder, Web version; Chemical Abstracts Service: Columbus, OH, 2010.
- (21) MySQL Server Community edition, version 5.0; MySQL AB: Uppsala, Sweden 2005.
- (22) *Apache Tomcat*, version 5.5; Apache Software Foundation: Forest Hill, MD, 2004.
- (23) *Java Development Kit*, version 1.5; Sun Microsystems: Santa Clara, CA, 2004.
- (24) *Eclipse SDK*, version 3.2; Eclipse Foundation: Portland, OR, 2006.
- (25) University of Illinois, Open Source License; University of Illinois: Chicago, IL; http://otm.illinois.edu/uiuc_openSource. Accessed on October 25, 2010.
- (26) *Apache Commons Fileupload*, version 1.2; Apache Software Foundation: Forest Hill, MD, 2007.
- (27) *Apache Lucene*, version 3.0.0; Apache Software Foundation: Forest Hill, MD, 2009.
- (28) Jericho, M. *Jericho HTML Parser*, version 3.1; SourceForge.net: Mountain View, CA, 2009.

CI100280A