

Chemical informatics for libraries

Presentation by Jeffery Loo (www.jeffloo.com)

Speaking notes (by slide number)

1

(title slide)

2

Good morning. Today we are here to discuss chemical informatics and its opportunities for libraries.

My name is Jeffery Loo. I am a medical librarian in the Associate Fellowship program at the National Library of Medicine. Currently, I am completing my second year at the Welch Medical Library of the Johns Hopkins University.

I'm very interested in chemical informatics because I have a background in chemistry with my undergraduate degree in this field.

3

"What is chemical informatics?" you may ask.

Very generally, it is the use of information science and technology to solve chemical problems.

4

In this presentation, my goals are to:

- Explain how chemical informatics manages and advances scientific developments
- Explore how library and information professionals can participate in this field.
- Outline the skill sets and training for chemical informatics
- Explore library service models for developing chemical informatics services.

5

Let's begin by looking at the importance of chemical informatics for managing and advancing scientific developments.

6

Certain developments are leading to rapid growth in chemical data and information, particularly in 3 areas:

- Chemistry
- Life and health sciences
- Drug discovery

I'm going to explore these developments and look at what chemical informatics can do.

7

Let's look at chemistry.

Traditionally, there have been 3 fundamental domains behind the field:

1. the design and synthesis of chemicals
2. finding out the structure of a chemical
3. finding the relationship between the structure of a chemical and the chemical's properties and activities

Currently, there are more than 41 million chemical compounds known. Already, there is a lot of information in these 3 domains of chemical knowledge.

8

This knowledge base is growing.

More new chemicals are being synthesized with combinatorial chemistry. This is a technique that can produce large numbers of structurally related compounds very quickly. Materials science and drug discovery relies on this technique.

More data is collected and generated through computational chemistry. In this field, computers are used to perform chemical studies and activities, such as synthesis design, chemical reaction prediction, and molecular modeling. Computational chemistry can explore the atomic and molecular levels impossible to reach by experimentation alone.

9

Chemical informatics help chemists manage the growing base of chemicals and data.

It facilitates information storage, focusing on data collection, organization and representation of chemicals and reactions, and organizing the data into databases.

In retrieval, it deals with the search for information on chemical structures and reactions.

Analyzing data can advance chemical knowledge. Chemical informatics provide tools for making calculations and analyses, employing machine learning, statistics, data mining, and other techniques.

This is a very general look at what chemical informatics does. Further detail is available in Gasteiger and Engel's Chemoinformatics: A Textbook, and in the Handbook of Chemoinformatics.

10

Now, let's look at another area of development.

Chemistry is increasingly important in the life and health sciences. As a result, life scientists need and are generating more chemical information.

To demonstrate this, I'm going to look at developments in chemical biology, small molecule chemistry, and high throughput screening.

11

Chemical biology is the application of chemistry to study molecular events in biological systems. It is an interdisciplinary field and encompasses biochemistry, biological chemistry, organic chemistry and more.

By helping scientists understand molecular mechanisms and pathways, chemical biology could be important for medicine and drug development.

Recently, Nature has developed a new journal, Nature Chemical Biology. It suggests that chemical biology will be important to expanding both chemistry and biology in new directions, forging collaborations between these two fields.

12

Another development is small molecule chemistry.

Small molecules are a class of chemicals that can affect protein function and physiology very nicely. For example, they may activate a biological function or inhibit it. They affect proteins in a reversible and controllable manner.

Because of their ability to perturb biological systems, scientists use small molecules to study biological mechanisms and pathways.

There are lots of small molecules available, and some estimate that there are between 10^{30} and 10^{200} possible small molecule compounds. This variety can be developed through techniques such as combinatorial chemistry.

13

Scientists are interested in bioactivity experiments that reveal the effects small molecules have on biological systems. With so many small molecules possible, this experimental work can be very time consuming.

Fortunately, robots are being used to perform experiments on small molecules. Thousands of samples can be run in a day. This use of robotic technology is known as high throughput screening.

The robot can take a combinatorial library of small molecules, test them on biological assays, and then record the results.

This is a video of the Kalypsys suite of ultra-high throughput screening technologies. It can evaluate the biological activity of more than 1 million chemical compounds in a single day.

14

There is an infrastructure of information resources to support chemical biology, small molecule chemistry and high throughput screening.

The ChemBank Initiative for Chemical Genetics includes a database of small molecules, their properties and their biological effects. This project is supported by the National Cancer Institute and is a collaboration based at Harvard Medical School.

At the NIH, the roadmap plan includes the Chemical Genomics Center. Also, there are plans to develop a network of screening centers to experiment small molecules on biological assays that scientists submit to these centers. The results are then publicly shared. This is known as the Molecular Libraries Screening Centers Network.

In the library world, the NCBI at the National Library of Medicine is working on the PubChem project. I think this is a terrific project and I'm going to spend a moment to explain.

15

The PubChem project consists of 3 databases.

Two of them are PubChem Compound and PubChem Substance. They both contain records on chemical compounds and their properties. These databases are different. The records in PubChem Compound have been validated and converted to a standardized form.

The third database is PubChem Bioassay. This includes information on bioactivity studies on various small molecules and the results.

Here is a record in PubChem compound.

You see the chemical structure. There are links to other resources as well. To chemical properties, MeSH terms, PubMed searches, toxicology information and related chemical structures. Here is a link to bioactivity studies.

It takes you to the PubChem Bioassay database. Here experiments are listed with their activity outcome. You can view and download the experimental data.

16

Let's look at another development driving chemical informatics: drug discovery.

Chemistry is important to key steps in the discovery and design of drugs.

In target identification, researchers try to understand the biological processes behind a disease. Scientists may use small molecules for elucidating disease mechanisms and elements.

In lead finding, researchers test chemicals on a biological system. If there is a desired biological effect, the chemical may have the potential to become a drug. This testing can be an iterative process: making and testing thousands of chemicals. Some scientists use computers to quickly identify leads by performing virtual screening experiments.

Then there is lead optimization. In this step, the scientist refines the chemical structure of a drug lead to create the optimal drug.

17

Chemical information is important for drug researchers. I won't go into all the details, but chemical informatics provides methods and tools that help the drug researcher to collect and organize chemical data, to analyze this data, and to make experimental predictions. Chemical informatics also has an important role in facilitating experimental work. As you can see here, there are many different facets within these broad categories.

18

Let me review what I've discussed so far.

Chemical informatics is the use of information science and technology to solve chemical problems.

There are developments that are generating more chemical data and increasing the need for chemical information.

I list some of the developments here. Notice that chemical information is very important for a lot of areas in the life and health sciences.

These developments are resulting in more:

- More compounds synthesized
- More methods to consider for chemical syntheses
- More experimental data generated
- Greater chemical information needs

Chemical informatics is a solution for managing this growth to increase our chemical knowledge.

19

Now I'm going to explore how library professionals can participate in chemical informatics.

20

The obvious role is for librarians to expand reference services and collection development for chemical informatics.

But, I want to focus on three specialized areas for participation:

- data acquisition and sharing
- information management, focusing on information storage and retrieval.
- information use, particularly with data analysis and information resource awareness.

I'm going to explore some of the challenges in each area, and then offer solutions we library professionals can provide.

21

First of all, there is a challenge to improve how chemical data is captured and shared.

A lot of chemical information is manually extracted from the primary literature and then sold back to the community.

Many chemists rely on these manually curated secondary publications. Some of these publications have had little change over the past 120 years, and are incomplete in time, coverage and coverage of information types. For example, more than 99% of all newly synthesized compounds reported in papers have IR spectra, but only a tiny fraction of this is available in electronic form.

Compare this situation to the biosciences or crystallography. Authors and publishers seem to value author-based deposition of data that is later aggregated in communally accessible data banks. Scientists are then freely able to apply data- and text-mining tools for research.

(paraphrased from Murraray-Rust, Mitchell and Rzepa. Chemistry in bioinformatics. BMC Bioinformatics. 2005 Jun 7;6(1):141.)

We information professionals can perform an institution-wide study of what is being done with chemical data.

From this awareness, we can develop software tools or methods to manage data acquisition and sharing. Generally, we can share our expertise in management and organization.

To facilitate data sharing, librarians can educate scientists on the value of sharing their data through communally accessible data banks. We need to outline how this will advance research, facilitate knowledge discovery, and promote a lab's research. We can show how sharing can be as simple as uploading data onto personal websites or institutional repositories. In future, scientists can make deposits into open data banks such as ChemBank or PubChem.

22

For information management, principal challenges are:

Efficient information storage and retrieval of data.

The integration of chemical data and information with other disciplines is increasingly important, particularly with the biosciences. However, a lot of chemical information is available in a dispersed manner in the primary literature or in disciplinary tools.

It is important to integrate disparate chemical resources such as chemical indexes, records, abstracts, toxicological information, data, and spectra, and more. These unified resources can then be integrated into resources of other disciplines.

This integration will support the development of “smart instruments.” These instruments are capable of collecting and analyzing experimental data as they are generated, and be able to incorporate information from existing resources. This helps scientists to make intelligent decisions during the course of an experiment.

When integrating information resources, it is important to consider the open movements, including Open Data, Open Source, Open Access, and Open Standards. Very generally, these opens promote a transparency and a freedom to share and use information and information tools.

23

Librarians and information professionals can meet these challenges.

Education is an important role. We can teach scientists about:

- the trends towards integrated and open systems in scientific information
- to adopt common standards and languages that facilitate data management and sharing
- to publish their articles in open access publications.

24

We can facilitate with the development of lab informatics, such as electronic notebooks or lab workflow management software, with the important goal of integrating information resources into daily lab processes. Librarians can consult on the customization of software tools to incorporate library and information resources.

25

Another role is the development of integrated and customizable search systems. If I were a scientist studying a chemical on a biological system I may need to search several different databases for my background research. It would be efficient to have a one-stop solution that lets users search across several databases at once.

This can be done with metasearch software solutions, such as MetaLib, that let you search across multiple databases at once.

Ideally, resources should be linked at the record level. For instance, the PubChem database can link individual chemical records to medical literature citations. These links may facilitate knowledge discovery.

At the least, libraries can develop subject guides/portals that list chemical resources useful for different disciplines.

26

Another role is working with digital repositories to archive, preserve, and provide access to chemical data and information.

I quickly looked at the use of DSpace by chemistry departments at MIT and other prominent institutions. I saw that it was effectively being used to store text files such as journal articles, reports, and theses. However, there was little use of digital repositories for storing chemical data.

Currently, there are some interesting DSpace projects for the archiving of chemical data.

One of them is SPECTRa: Submission, Preservation and Exposure of Chemistry Teaching and Research Data.

This is a project lead by the University of Cambridge in collaboration with Imperial College London.

Initially, the project will study researchers' needs and scope their data-handling requirements. It will then develop automated tools so that high-volume data can be identified, extracted, and ingested to repositories, where it will be preserved and accessible for use to support research and teaching.

Another project is the WorldWideMolecularMatrix. This is a collection of information on small molecules. This is a project at the University of Cambridge at the Unilever Centre for Molecular Informatics. This project is trying to develop open electronic collections of chemical information. This collection will contain the calculated properties of over 200,000 molecules provided by the National Cancer Institute.

Library professionals can educate faculty about the benefits of depositing their digital content into repositories. We can also help faculty use digital repositories. For example, some university libraries have dedicated FTE's who visit faculty, office by office, to help them deposit copies of their articles. The St. Andrews University Library in Scotland asks faculty to send in their articles as email attachments and library staff will then deposit them into the repository.

27

Let's now look at chemical information use.

Some challenges include:

- awareness among scientists of chemical information resources and tools
- data analysis for extracting information and generating knowledge
- knowledge of computational methods for chemistry

Library professionals have a role to promote chemical information resources. We can work closely with scientists to learn what tools and resources they find useful, and communicating these tips to others

We can promote software tools for data analysis. The library could purchase and provide networked access to software packages. Or, the library could create a subject guide of chemical software tools.

Also, library professionals can work closely with scientists in applying chemical literature to research problems. A consultant could be in the lab working closely with the research team to provide advanced and specialized user services.

It would be helpful to look into medical librarianship, and their model for the informationist – this is a specialized information professional working closely with a clinical team.

28

We've looked at opportunities for information professionals, now let's look at the training to prepare to work in chemical informatics.

29

There are three principal Master's programs in chemical informatics:

- Indiana University
- The University of Sheffield
- The University of Manchester

However, many chemistry departments already have course modules in chemical informatics.

At the University of Massachusetts, Lowell, there is a Master's program in Computer Science with an option for Cheminformatics.

30

In general, students take a variety of courses in chemistry, library and information science, and computer science and informatics.

I list here the possible courses in these areas.

The programs also permit a variety of elective courses including the life sciences, intellectual property, e-business and e-commerce, healthcare information and electronic publishing.

Topics of dissertations include: databases, molecular modeling, chemical publications and services, information searching and retrieval, and computational chemistry.

As you can see in the training, there are many different areas of specialization in chemical informatics.

31

How can libraries provide services for chemical informatics? I would like to share two model programs at medical libraries for bioinformatics services. I believe these programs are quite innovative. I'm looking at bioinformatics principally because it is a relatively well established informatics field in libraries.

32

One model is the bioinformatics program at the University of Washington Health Sciences Libraries.

In 1995, the library recruited a Ph.D. scientist who had a background in molecular and cellular biology for its bioinformatics consultation position.

First, the library assessed the information needs of molecular biology researchers. This led to the development of a three-pronged program involving: consultation, education, and resource development.

The consultation service provides in-depth assistance with bioinformatics data analysis and help in planning experimental strategies. The library believes that the consultant needs a strong background in molecular biology or specialized training in this area, as only 20% of the questions received could be expected to be answered by a librarian without these qualifications.

Also, there are library-based and graduate level bioinformatics training. The library's education program includes a bioinformatics series. And, the bioinformatics consultant has been teaching a graduate level bioinformatics course in collaboration with medical faculty. In addition, the consultant also regularly presents lectures about biological information resources to graduate classes.

For resource development, the library has been providing networked access to biological information resources and to bioinformatics software. I think that it is important for libraries to consider nonbibliographic biological information resources as part of collection development. The library has also developed a website geared towards biological researchers, with news, and links to useful resources. Finally, the library has a communication and outreach program for scientists working with bioinformatics.

33

The Eskind Biomedical Library is a leader in training librarians to provide highly specialized services.

The library developed a curriculum designed to increase bioinformatics competencies. Professional library staff participated in a 12-week training course consisting of 5 distinct modules in molecular biology, genetic analysis, biotechnology, research literature, and databases.

Some staff later went on to pursue in-depth individual training. This program increased the number of library staff with advanced molecular biology expertise.

With these specialized staff skills, Eskind Library is able to provide strong outreach services. They have successfully moved information specialists into clinical and research settings outside the library walls, making librarians vital additions to multidisciplinary teams.

For instance, they developed RICS, the Research Informatics Consult Service. The purpose of the program is to provide proactive, targeted information services delivered to health sciences researchers at the point of need. RICS services include training, grant assistance, access to electronic resources, database searching and information filtering.

They have similar consultation service programs in clinical informatics, evidence-based medicine, and consumer health.

The library attributes their success to the library's culture of lifelong learning in daily professional practice, and to the strong support of library leadership.

34

To summarize this presentation, let me highlight the key points.

Chemical data and information needs are growing.

Several key scientific developments are behind this trend, particularly the increasing importance of chemistry to the life sciences.

Chemical informatics provide tools for managing this growth.

Library professionals have an important role working closely with scientists in data acquisition and sharing, information management and information use.

Training and skill sets for chemical informatics centers on 3 areas: chemistry, library and information science, and computer science and informatics.

Look at library service models for bioinformatics when developing services.

35

Thank you very much for your attention. May I take your questions?