

Introduction to Chemoinformatics

Jeffery Loo
NLM Associate Fellow
jloo1@jhmi.edu

What is chemoinformatics?

Chemoinformatics is the use of information technology to:

- manage chemical information, and
- solve chemical problems

Chemoinformatics deals with the chemical information for **drug discovery**.

The need for chemoinformatics

Recent chemical developments for drug discovery are generating a lot of chemical data. These developments are combinatorial chemistry and high-throughput screening. Some scientists have described this situation as a chemical information explosion. This has created a demand to effectively collect, organize, and apply the chemical information.

What is drug discovery?

Drug discovery is a research process:

- The goal is to make a chemical drug that causes a desired biological effect.
- First, researchers test a variety of chemicals on biological systems to see if they generate a desired biological response. This testing phase identifies potential chemical candidates for medical drugs, known as **drug leads**.
- Researchers then optimize these drug leads to maximize their beneficial properties and minimize their side effects. Optimization may involve adding, removing, or replacing chemical components or by changing the shape of the chemical.
- After optimization, a drug has been developed, which then undergoes further testing and study.

(Scientific note: this drug discovery methodology is for the case when a protein receptor structure is unknown and the ligand structure is also unknown)

Drug discovery can be a very slow process:

1. Identifying and testing drug leads can be inefficient. A drug lead may belong to a family of similar chemicals that may include thousands of variants. It can be tedious to test all the variants, and it can be extremely difficult to identify the worthwhile drug leads in a family of chemicals.

2. It can take a long time to test a chemical for a biological response. The researcher needs to prepare a control sample, prepare the biological assay, prepare the reagents, monitor the reaction, record observations, report results, and more.

Two developments address these limitations and speed up drug discovery. They are:

- combinatorial chemistry
- high throughput screening

Combinatorial chemistry: a driving force for chemoinformatics

Combinatorial chemistry is a process that generates large numbers of related chemical compounds. This can help the scientist find drug lead candidates more quickly.

In the combinatorial chemistry process, you combine variant molecules to form a chemical compound. (Molecules are the 'ingredients' of a chemical compound.) Take the example of the chemical compound A–B that is formed by the two different groups of molecules A and B.

Suppose there are 2 variants for each group of molecules: A1, A2, B1, and B2.

From these **4** molecules, **4 possible A–B** chemical compounds can be made:

A1–B1	A2–B1
A1–B2	A2–B2

Now, suppose there are 3 variants for each group of molecules: A1, A2, A3, B1, B2, and B3.

From these **6** molecules, **9 possible A–B** chemical compounds can be made:

A1–B1	A2–B1	A3–B1
A1–B2	A2–B2	A3–B2
A1–B3	A2–B3	A3–B3

Here lies the power of combinatorial chemistry. By using 2 more molecule variants, we have more than doubled the chemical compounds made. This simply demonstrates that adding a few more "molecular ingredients" can dramatically increase the possible number of chemical compounds made.

Combinatorial chemistry speeds up drug discovery in two ways:

1. It can generate several million structurally related chemical molecules. These collections of chemicals are known as combinatorial libraries or chemical libraries. By increasing the chemicals available for biological testing, the chances of finding a drug lead may be higher.
2. If you know the biological effects of different molecules, you can combine them to make chemicals with particular designed biological effects. The result may be better drug leads.

High-throughput screening: another driving force for chemoinformatics

High-throughput screening (HTS) speeds up drug discovery by automating the drug lead testing process.

In HTS:

- There is a robot machine that tests large numbers of chemicals for biological effects.
- A scientist prepares many biological samples and then loads them into a machine along with a combinatorial library (a collection of many related chemicals).
- The machine will automatically add the chemical to the sample, measure for a response, and then record the data.

Chemical data explosion

Combinatorial chemistry and high-throughput screening are data dependent and data rich technologies. When making combinatorial libraries of chemical compounds, you need information on the molecular components, their biological effects, and information on how to prepare the compound. There is also data for managing and storing the libraries. In high throughput screening, the test results need to be captured, stored, and then analyzed.

Chemoinformatics tasks

Chemoinformatics collects, manages, analyzes and disseminates the chemical information needed for drug discovery. Some of the tasks in chemoinformatics research are:

- analysis of HTS data
- similarity search chemicals
- design of combinatorial libraries
- design of focused libraries
- comparison of the similarity/diversity of libraries
- virtual screening
- docking
- *de novo* design
- pharmacophore perception
- prediction of affinities, physicochemical properties and pharmacokinetic properties
- establishment of QSAR models which can be interpreted and guide the further development of a new drug

The following book chapter provides more detail:

Terfloth L. Drug Design. In: Gasteiger J, Engel T, eds. *Chemoinformatics: A Textbook*. Weinheim, Germany: Wiley-VCH; 2003:597-622.

Prevalence of chemoinformatics

In addition to private sector investments by the pharmaceutical industry, here are some prominent public and academic supporters of chemoinformatics:

National Institutes of Health

The NIH Roadmap, the National Institutes of Health's ambitious plan to advance medical research for the 21st century, has identified chemoinformatics as one of the "New Pathways to Discovery". It is an important component to developing NIH research in Molecular Libraries and Imaging. (<http://nihroadmap.nih.gov/>)

National Library of Medicine

The National Center for Biotechnology Information, a division of the NLM, has developed PubChem (<http://pubchem.ncbi.nlm.nih.gov>). The PubChem project is developing a database of chemical structures, properties, and activities that will be integrated with other databases and literature. There will also be links to data generated from HTS screening centers. PubChem is in line with NIH Roadmap plan to advance medical research.

Universities with dedicated chemoinformatics programs

Several universities have developed MSc programs devoted to chemoinformatics:

- University of Indiana
<http://www.informatics.indiana.edu/academics/chem.asp>
- University of Sheffield
http://www.shef.ac.uk/is/courses/pg_mscci.html
- University of Manchester
<http://www.manchester.ac.uk/degreeprogrammes/postgraduate/taught/1107.htm>

Johns Hopkins University School of Medicine

At JHMI, there is the JHU ChemCORE Facility, which is an integrated robotics and chemical repository unit. Services include:

- high throughput screening against ChemCORE compound libraries
- assay development for screening
- compound library duplicate plates

Visit their website at http://www.molecularinteraction.org/chemcore_first%20page.htm

Electronic resources for chemoinformatics

There are many electronic resources available for chemoinformatics. On top of the traditional chemistry information resources, some good resources are:

- PubChem
<http://pubchem.ncbi.nlm.nih.gov/>
- ChemBank
<http://chembank.med.harvard.edu/>

For a detailed list of resources, see:

Engel T. Databases and data sources in chemistry. In: Gasteiger J, Engel T, eds. *Chemoinformatics: A Textbook*. Weinheim, Germany: Wiley-VCH; 2003:227-290.

References

Broach JR, Thorner J. High-throughput screening for drug discovery. *Nature*. 1996;384(Supp 7 Nov 1996):14-16.

Brown FK. Chemoinformatics: What is it and how does it impact drug discovery? *Annu Rep Med Chem*. 1998;33:375-384.

Gasteiger J, Engel T, editors. *Chemoinformatics: A Textbook*. Weinheim, Germany: Wiley-VCH; 2003. 649p.

Plunkett MJ, Ellman JA. Combinatorial chemistry and new drugs. *Sci Am*. 1997 April:68-73.

Russo E. Chemistry plans a structural overhaul. *Nature*. 2002;419(6903):4-7.