1.1: Introduction

What is cheminformatics?

Modern cheminformatics evolved out of the "drug discovery" needs of the pharmaceutical industry. The term was originally coined as "chemoinformatics" by Frank Brown of the R.W. Johnson Pharmaceutical Research Institute in his 1998 manuscript, "Chemoinformatics: What is it and How does it Impact Drug Discovery". The term "chemoinformatics" still tends to be preferred in Europe today, we will use "cheminformatics" in this class, as that terminology aligns with the premier open-access journal of the field, the Journal of Cheminformatics.

Although much of the material in this class will be of value to students of pharmacology, we are taking a much broader perspective of the field and treating cheminformatic's skills as essential to one of the primary paradigms of science. The informatics, data representational and analytics skills learned in this class would be of value to a wide variety of tasks in the pursuit of knowledge in the chemical sciences, and this course is of value to students beyond those of the pharmaceutical sciences. In fact, we can go a step further and state that cheminformatics is changing the fundamental cognitive artifacts used to represent, manipulate and communicate chemical information, and in a world of instant access to interconnected digital data, a fundamental understanding of cheminformatics is an essential skill for tomorrow's practicing chemist.

Paradigms of Science

Cheminformatics can be considered to be a "fourth paradigm science" in the context of the 2009 Microsoft Research book published in honor of the late Jim Gray, "The Fourth Paradigm: Data-Intensive Scientific Discovery." In the forward of this book is a transcript based on Jim's last talk; "eScience: A Transformed Scientific Method", where he describes the four paradigms of science. The following paradigms of science are based on figure 1* of this transcript:

> **First Paradigm: Empirical Science** (thousands of years old)
> - the experimental chemist (scientist) making measurements and observations of the physical universe and generating empirical data.
>
> **Second Paradigm: Theoretical Science** (centuries old)
> - the theoretical chemist (scientist) defining complex mathematical relationships that underpin natural observations.
>
> **Third Paradigm: Computational Scienc**e (decades old)
> -the computational chemist (scientist) using computing machines to perform complex calculations to predict behavior and generate computational data.
>
> **Fourth Paradigm: Data Exploration** (emerging)
> -the eScientist using computing machines to discover complex relationships across datasets of both empirical and computational data.

The fourth paradigm actually depends on the other paradigms and requires the ability to acquire, manipulate and understand data. This class will introduce you to a variety of open source software programs and public compound databases, with a teaching focus on PubChem. But the skills will allow you to pursue data exploration with other data sets and resources. Since we are bringing in resources across the web, we will use a web annotation service, Hypothes.is, to connect those resources to the chapter discussions.

Hypothes.is Web Annotations

This is a collaboratively taught intercollegiate course and will use the Hypothes.is Web Annotation Service (https://web.hypothes.is/) to discuss the content in a webpage overlay. Students need to create an account at Hypothes.is, and then join the class discussion group through a link that their instructor will send them. Please review your syllabus before creating your account, as your instructor may include naming protocols for students in your class. You probably should also install the browser plugin when you create your account, as that will allow you to annotate pages external to the LibreText HyperLibrary.

There are two types of annotations faculty and students in this class will will make, those intrinsic to a page being discusses, and those extrinsic to the page.

Two types of Annotations

1. **Intrinsic Annotations:** You simply highlight content intrinsic to a page within the LibreText hyperlibrary, choose your class discussion group and annotate. After you save your annotation it will automatically appear in the overlay of that page. This is

possible because hypothes.is is integrated into LibreText, and note through the WYSIWYG editor you can format your annotations, add images, videos and the like.

2. **Extrinsic Annotations:** For content external to a page of the LibreText HyperLibrary you need to tag the annotation so it can be displayed on a page within the HyperLibrary. The table of contents of each chapter will be used for this purpose, and have a specific tag that you use for that chapter. When an annotation is made on any open access webpage on the web, and tagged with that chapter's tag, the annotation will appear at the bottom of that chapter's table of contents, and include a contextual link to the annotation. If the page being annotated is not part of the LibreText HyperLibrary you may need to install a browser plugin to be able to annotate it (which will be neccessary if hypothes.is is not integrated into the webpage). To install the plugin go to https://web.hypothes.is/start/.

Annotation Features

1. **Annotation Overlay:** Unlike web 2.0 comment features where people discuss an article at the bottom of a webpage, Hypothes.is uses an "overlay" on a web page that can be activated by clicking the arrow on the upper right corner of the webpage. If you click on an annotation in the overlay, the page scrolls down to the actual highlighted text. This is sort of like commenting on a piece of note-paper attached to a webpage, instead of commenting on the webpage itself, and you have to activate the overlay to see the annotation.
2. **Contextual Links:** A contextual link is a link to target text within a page. When you click a contextual link, it opens the page in a new browser, activates the overlay, and scrolls down to the targeted text. Technically, these are called direct links that combine a webpages URL with a selector that refers to specific text within the page, the target text of the contextual link.
3. **Groups:** Hypothes.is allows you to make comments and tags that can have either public or private group access. Only members of a group can see group annotations, and we will use a group that includes faculty and students from multiple campuses. Your instructor will provide you a link in your class syllabus that will allow you to join the group. Your instructor may also provide you with a unique username that "hides" your identity from everyone except your instructor, and provide instructions on how to create a Google email account for this class. This will not only allow your instructor to quickly identify students in your class, as compared to other classes that are involved with the course, but also allow you to create your own hypothes.is account that you can use outside of the class.
4. **Tags:** Learning how to tag annotations is an important ability, as it not only allows you to search and sort your tags, but also to connect tags to different web objects. So if you find 5 passages of text dealing with molecular fingerprints, you cold tag them, and then search them from your homepage (which automatically lists your tags), and have instant access to them. Your username is also a tag, and so if someone else had used the tag "fingerprint", you could find their target text, of by combining the tag with your username tag, filter the query to just the items you had tagged.
5. **Replies:** Within the overlay are discussions. As a student you will get an email if someone replies to a question you have, but also, through the overlay you can look at all the other questions dealing with chapter, and the discussions that evolved out of them. Our vision is that the overlay becomes a layer to the textbook where students across multiple campuses can learn from the questions and answer discussions of others.

Note on Annotations

It is interesting to note that annotations were part of Tim Berners-Lee's original 1989 proposal for the World Wide Web, and were integrated into the prerelease version of the 1994 NCSA graphical Mosaic Web Browser, and yet today, 30 years after Tim Bererns-Lee's original proposal, few faculty and students use them. In this class you will be expected to use web annotations to connect open-access cheminformatics resources across the web to the discussions of your class topics.

The W3C Web Annotation Working Group has an excellent interactive image describing Web Annotation Architecture that you might enjoy walking through.

What is PubChem?

A brief description of PubChem is warranted here, no more than a few small paragraphs. We should also state that there will be a chapter on other resources, but this class will focus on PubChem with respect to training students how to access data.

What is Programmatic Access?

A brief comparison of a GUI/webpage and an API. The goal here is to put the foundations down for training in programmatic access through PubChem, but the skills can be used with any database.

Non-Open Access Resources

This course is not attempting to provide a comprehensive coverage of contemporary cheminformatics resources, but to train students in the skill sets associated with data exploration in the chemical sciences. This is an intercollegiate course open to all schools, and although some will have access to content like ACS's SciFinder, STN and Elsevier's Reaxys, others, especially Primary Undergraduate Institutions, which often do not have the graduate level research needs to support those technologies, will not. None-the-less, the skills students learn in this class should assist them in utilizing those resources in their future endeavors.

In a similar vein, each chapter will provide a bibliography including suggested reading material, that will be delineated into open access and restricted access publications of the primary literature. But required reading assignment will be limited to open access publications. This is for for several reasons

1. Only open access content can be connected to the textbook discussions through Hypothes.is.
2. We can not expect our students to pay the exorbitant fees that publishers charge for access to single articles
3. This is an Open Education Resource (OER), which we expect others to use once the course is over, and to be of value they must have access to the content.

We regrettably recognize that in making the decision to limit this course to open access content that there will be a substantial amount of high-quality cheminformatics material that will not be available to our students. We believe that access to a quality education, which is one of the 17 United Nation's Sustainable Development Goals for 2030, is a fundamental human right, and that the content of this course needs to be available to all.

Contributors

Robert E. Belford, UA Little Rock

*This contextual link uses the https://web.hypothes.is/ annotation service to take you to the part of a PDF on the Fourth Paradigm that Microsoft Uploaded to the web. This is a public link and should be viewable to anyone on the web, but if you are in the class you will need to create an hypothes.is account and join the class group as outlined in your syllabus, as otherwise you will not be able to see or participate in the class discussions.