This introduction has two purposes; to introduce you to cheminformatics, and to introduce you to the course.

**Part I: The Introduction to Cheminformatics.**

In this page we are posting an introduction to cheminformatics from the perspective of an in silico Medicinal Chemist, Nathan Brown, who has also shared his recent text; "In Silico Medicinal Chemistry", which students who have logged in can access from the bottom of this page. Please note, you only need to click on the file name. If you click the radio button and "save", you will delete the file. Although the modules will follow the initial chapters of Dr. Brown's text, the course will focus on public chemical compound databases, and how chemicals, and chemical data are represented on computer. That is, there is a lot more to the field of cheminformatics than what this course will attempt to cover, and this introduction is to help us see where the course material fits in the larger field of cheminformatics.

**Part II: The Introduction to this Course and the Participants.**

This is an intercollegiate course where students, faculty and non-academic professionals have an chance to interact through the course website, and we thought a good way to learn how to use the course website would be for everyone to introduce themselves in a short comment. We would appreciate if students could indicate their school, and tell us a little about themselves and why they are taking this course, and we ask students to only use first names, not last names. We also encourage students and anyone who is new to the course to look at the videos on the WebTutorials page, especially the first video on Logging in and Discussing Modules.

**Part I: Introduction to Cheminformatics*  

By Nathan Brown

Please note, this is being developed like a blog or a wiki, and the text of this introduction is currently being written. The advent of the widespread availability of electronic computers, primarily since the 1970s, has led to huge advances in many scientific disciplines. The field of chemistry itself has benefitted greatly from this availability, but the development of many new methods, algorithms, and data sources was necessary to realise the compute power now available to the chemist. The interface science of Cheminformatics has the objective of applying computer science approaches in the representation, analysis, design, and modelling of chemical structures and associated metadata, such as biological activity endpoints and physicochemical properties. The field of Cheminformatics not only draws on expertise in computer
science and chemistry, but also mathematics, statistics, biology, physics, and biochemistry. In this introduction to the course on Cheminformatics, we will introduce some of the overarching concepts in the field and introduce some of the open access resources that may be applied in understanding the data types and methods that are widely available to the community.

Representing Chemical Structures in the Computer

The representation of chemical structures in the computer has a history going back some centuries, to the advent of atomistic theory in the mid-19th century, and even further to the development of the mathematical discipline of graph theory in the early-18th century. The famous mathematician, Leonhard Euler, used an abstraction of a real-world problem in the early-18th century to understand whether it is possible to devise a walk around the town of Königsberg in Prussia (present day Kaliningrad, Russia), while crossing each and every one of the seven bridges connecting the mainland to the island in the centre of town across the river Pregel, once and only once. In analysing this problem, it led Euler to devise an abstraction of the real-world problem - that could be easily represented on a geographic map - into one that pared back the details to only those that were important. The salient details required to solve this problem were, namely: the land masses of Königsberg and bridges, or connections, between them. It was not at all necessary to know the shape, topography, elevation, or any other details of the land masses including the routes internal to each landmass, other than that they existed. Similarly, it was only necessary to know that two land masses were or were not connected to another landmass, and how many connections there may be between them.

In devising this abstraction of a real-world problem, Euler was to make a significant impact on mathematics, essentially formalising a new sub-discipline called graph theory. Euler’s work here led to the development of a field of endeavour that is today applied widely, not only in chemistry, but also in social network analysis and biochemical pathways, amongst others. But what of Euler’s initial problem? Euler demonstrated that for a walk to be possible between the land masses (or nodes or vertices), across the multiple bridges (or edges or arcs), it was reliant on the number of connections to each of those nodes. The number of connections to each node is called the degree of the node. Euler showed with his abstraction that for such a walk to be possible, then only zero or two nodes may be permitted to have an odd degree. Therefore, given that all nodes in the Koenigsberg representation have an odd degree (3, 3, 3, and 5, respectively) then a walk fulfilling the defined restrictions would be impossible. The solution is called the Eulerian walk or path in Euler’s honour and led to the formalisation of the fields of graph theory, topology, network analysis, and combinatorics.

Molecular Similarity

One of the key and enduring concepts in Cheminformatics is that of molecular similarity. Quantifying the similarity of molecules has a wide range of applications, many of which will be covered later in this introduction, but the fundamental aspect that underpins all of these applications is the similar property principle. The similar property principle suggests that often, if two chemical structures are similar, they will also exhibit a number of similarities in their properties. However, although this heuristic holds true in many examples, it is also observed that highly similar chemical structures have significantly differences in properties, particularly in biological activity, a phenomenon known as Activity Cliffs. In some instances this can give rise to terminology such as the Magic Methyl, where a single carbon atom may bestow or remove activity, although the effect of this kind of alteration can typically be rationalised by some property change, such as a clash with a protein binding site, or a forced conformational change that is advantageous or detrimental.
Given the subjective nature of some aspects of molecular similarity, particularly when simply compared visually, it is often important to generate objective measures of molecular similarity based on the actual chemical structures, similarity of molecular descriptors, or similarity in some measured or predicted property. The comparisons made according to structure only often rely on graph theoretic algorithms to calculate molecular graph similarity, but also can be a shape and electronic similarity, such as that generated in pharmacophoric descriptor generation tools like ROCS (Rapid Overlay of Chemical Structures) from OpenEye Scientific Software. There exist many molecular descriptors in the literature that are used to rapidly generate molecular similarities, which can be simply classified into property descriptors, topology descriptors (those generated from molecular connectivity alone), and topographical descriptors (those that are generated from the geometric shapes of molecular structures).

**Molecular Property Descriptors**

The first class of molecular descriptors to be covered here are the property descriptors, or modelled properties that indicate some reliable prediction of a physicochemical property, such as molecular weight or the octanol-water partition coefficient (ClogP). The descriptors tend to convolute any different properties into these simple scalar descriptors, but can be highly effective in certain circumstances and are widely appreciated for their interpretability in interactive systems. One such set of property descriptors that has gained wide acceptance is the Lipinski rule-of-five, which has been suggested as an heuristic for indicating the oral absorption of a potential drug based on marketed orally-dosed drugs. The rule-of-five applies four calculated properties and defined cut-offs, each of which is a multiple of five. The four properties and their cut-off ranges are: Molecular Weight (MW or MWt) less than 500 daltons; predicted octanol-water partition coefficient (ClogP) less than five; fewer than five hydrogen bond donors (HBD = total number of nitrogen-hydrogen and oxygen-hydrogen bonds); and fewer than ten hydrogen bond acceptors (HBA = total number of all nitrogen and oxygen atoms). As indicated the Lipinski rule-of-five is an heuristic, albeit useful, and is often applied, somewhat crudely, as a drug-likeness descriptor in curating screening collections.

**Topological Descriptors**

The second class of molecular descriptor to be discussed in this introduction is the class of topological descriptors. Topological descriptors are those calculated from the molecular structure, typically using only the atomic connectivity data and eschewing any geometric data - although exceptions do exist. Two types of molecular descriptor are often used, molecular indices and molecular fingerprints. Molecular indices are single real-valued descriptors that summarise some characteristics of the molecular structure under consideration. One of the older topological indices is the Wiener index, developed by Harry Wiener in 1947. The Wiener index is calculated as the sum of distances between all carbon atoms. Another popular index is the Randic index, developed in 1975 by Milan Randic, and focusses on the atom connectivities, or node degrees.

The second class of topological descriptor to be considered here is the molecular fingerprint. A molecular fingerprint is often a long, contiguous array of bits, but also sometime integers and real-valued descriptors, which can be compared to each other using a similarity coefficient. As with many molecular descriptors, a large number of molecular fingerprints have been defined. The fingerprint was originally designed as a rapid screen-out descriptor prior to the more computationally intensive substructure search being performed in chemical information retrieval systems. The substructure of interest was encoded into a fingerprint to be compared a database of pre-calculated fingerprints for
chemical structures of interest. If, when using the substructure query fingerprint as a bit-mask, a given database fingerprint has precisely the same bits set, then there is a high probability, depending on the molecular fingerprint being used, that the substructure is contained within that database structure. If a match is identified then the database fingerprint is passed to the more computationally intensive substructure searching algorithm using graph theory. More recently, however, molecular fingerprints have been applied to a variety of pressing challenges in Cheminformatics, including cluster analysis, predictive modelling and similarity searching, more of which later in this introduction.

Molecular fingerprints can be subdivided into two different classes: knowledge-based fingerprints, and information-based fingerprints. Knowledge-based fingerprints use dictionaries of molecular substructures with a corresponding bit in the fingerprint assigned to each substructure unambiguously if it is present in the structure under consideration. Typically, even if the substructure appears multiple times, it will only be counted once in the fingerprint. Dictionaries of substructures tend to be relatively small, a few hundred, and can suffer from brittleness when considering new and unusual chemistry that may not have been considered when the dictionary was compiled - a little like Samuel Johnson not know what an Aardvark is. This brittleness can be overcome by applying information-based fingerprints, which do not suffer from such brittleness.

The information-based molecular fingerprint takes the chemical structure under investigation and transforms that structure into a fingerprint representation using one of a variety of algorithms. One of the most famous information-based molecular fingerprints is the Daylight Fingerprint designed and implemented by Daylight Chemical Information Systems. Here, the chemical structure is examined by iterating over each individual atom and enumerating all possible atom-bond-atom paths up to a specific length, typically seven bond length paths in these fingerprints. Each path, of all lengths from zero (just the atom itself), up to and including length seven, are then passed to a hashing algorithm that converts that string path into a number that is in a high range of something in the order of $(-2^{32}$ to $2^{32})$. The resulting number is then ‘folded’ into the length of the fingerprint and the corresponding bit at that index is set to one. One challenge in this approach in encoding the fingerprint is the probability of bit collisions, where two different paths encode to the same fingerprint index. The effect of bit collisions can be somewhat overcome by passing the original hashed value as a seed to a pseudo random number generator (RNG), and the first few values taken from the RNG and those values folded into the fingerprint and set to one. Fingerprints are often quite long, 1024 or 2048 bits is not uncommon, and these lengths offer an effective balance between calculation speed of the molecular similarity and the information capacity to appropriately describe the molecular structures.

Contributors

• Nathan Brown, Institute of Cancer Research

Adapted from Spring 2017 Cheminformatics OLCC