

ISMB 2006 TUTORIALS

CHEMOINFORMATICS

Pierre Baldi, Jonathan Chen, S. Joshua Swamidass
Institute for Genomics and Bioinformatics
School of Information and Computer Sciences
University of California, Irvine

Tutorial Outline:

- 1 Introduction**
- 2 Chemical Data and Databases**
- 3 Molecular Models, Representations, and Annotations**
- 4 Chemical Similarity and Searching**
- 5 Chemical Reactions**
- 6 Predictive Methods/Machine Learning**
- 7 Molecular Docking**
- 8 Applications: Drug Screening/Design**
- 9 Conclusions and Discussion**

Notes Outline:

- 1 Introduction**
- 2 Molecular Models, Representations, and Annotations**
- 3 Chemical Similarity and Searching**
- 4 Chemical Reactions**
- 5 Molecular Docking**
- 6 Applications: Drug Screening/Design—A Case Study**
- 7 References**
- 8 Web Resources**

1 Introduction

Definition

Chemoinformatics is a relatively young term, patterned after bioinformatics, still with alternative spellings (“cheminformatics”) and declinations (“chemical informatics”). Like bioinformatics, the boundaries of chemoinformatics are not well defined and may vary depending on people, context, etc. The narrowest definitions tend to emphasize drug discovery applications. For instance, in a recent book, chemoinformatics is defined as “the set of computer algorithms and tools to store and analyse chemical data in the context of drug discovery and design projects”. In a similar vein (Brown 1998), chemoinformatics is defined as “the mixing of information resources to transform data into information and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization”. While the emergence and expansion of chemoinformatics is indeed largely driven by the vast quantity of data associated with, or generated by, drug discovery projects (e.g. HTS, combinatorial chemistry), it is probably counterproductive to use a narrow definition, and futile to try to precisely carve the boundaries of chemoinformatics as a scientific discipline. In our view, it is wiser to use more general and broadly encompassing definitions such as: “chemoinformatics encompasses the design, creation, organisation, management, retrieval, analysis, dissemination, visualization and use of chemical information”, or “the application of informatics methods to solve chemical problems”, or “the intersection of the computational and chemical sciences”. In this broader sense that goes well beyond drug discovery, computational chemistry, quantum mechanical simulations, retrosynthesis, reaction discovery, molecular docking, compounds databases, reaction databases are all examples of topics that fall within the scope of chemoinformatics.

Historical Perspective and Comparisons with Bioinformatics

From an historical perspective, it is also informative to draw analogies between chemoinformatics and bioinformatics. In spite of its central role between physics and biology, chemistry has remained in a backward state of informatics development compared to its two close relatives. Computers, public databases, and large collaborative projects have become the pervasive hallmark of research in physics and biology. The Human Genome Project, for instance, required collaboration

among dozens if not hundreds of scientists across the world. And the resulting human DNA sequence, as well as a wealth of other biological information, are available for anyone to download from public repositories on the Web such as GenBank, Swissprot, the PDB, and PubMed. Virtually every biologist today uses publicly available tools, such as BLAST, to search sequence databases and analyze high-throughput data. Similar observations can be made in physics with large collaborative efforts in, for instance astronomy or high-energy physics. The Web itself was born at CERN, a European consortium with over half a century of history, and the world largest particle physics laboratory. In stark contrast, large collaborative efforts and public databases and software are comparatively absent from chemical research.

This is not to say that chemists do not use computers or databases at all. Of course they do and chemoinformatics has a long tradition (Gasteiger 2006), but these uses have remained limited and somewhat peripheral to the chemical sciences. Suffice it to say that to this date there is no publicly available repository of all known molecules publicly available and downloadable over the Internet, and no large-scale collaborative effort to annotate any significant portion of chemical space. The equivalent of BLAST for chemistry remains to be created.

The underdeveloped state of chemical informatics is even more surprising when one realizes that chemists were among the first to understand the importance of annotated repositories. The Beilstein system was created more than two centuries ago. However, most of these repositories have not kept pace with the explosion of chemical information, the computer/Internet revolution, and movements toward openness in other sciences.

This unfortunate state of affairs and the overall conservatism of the chemistry community is unlikely to result from some intrinsic properties of chemistry as a science. Rather, it is likely to be the product of complex historical and sociological factors, that may include: (1) the origins of chemistry in, for instance, secretive alchemy; (2) the early but large-scale industrial and commercial applications of chemistry; in contrast with more recent applications of biology to biotechnology; (3) related to (2) is the parallel development of modern computer and genomic sciences, as opposed to the early start of chemistry. Finally, in modern times, the American Chemical Society has certainly played a role in the current state of affairs (Marris 2005a and b, Kaiser 2005a) by controlling and profiting from the

dissemination of chemical information through journal and database ownership and commercialization.

Development of new informatics methods and algorithms to search chemical space requires having access to large corpus of data in order to compute statistical properties and detect patterns that can then be used to develop search algorithms and other modern datamining methods. In many ways, the state of chemoinformatics today recalls the state of bioinformatics a few decades ago, before the advent of Genbank and BLAST. The lessons learnt from bioinformatics' exponential development over the last few decades strongly suggest that two ingredients are essential to develop the chemistry cyberinfrastructure: (1) large public data repositories; and (2) the tools to search them efficiently.

Data

Although the methods to be developed apply to other areas of chemistry, in this tutorial we will focus on organic chemistry and small molecules for several reasons. Small molecules, containing at most a few dozen atoms and the associated chemical reactions, are very important for a variety of purposes in biology, chemistry, and other areas. For instance, small molecules occur ubiquitously as metabolites during biochemical reactions, and their study is important for understanding biological systems (Camilli 2006). Small molecules are routinely used as building blocks in chemical synthesis to build more complex molecules (Schreiber 2000, Agrafiotis 2002), including polymers. Natural and man-made polymers, from DNA/RNA, to proteins, to silk and nylon, are made of small molecular building blocks. In addition, most drugs consist of small molecules capable of selectively interacting with specific proteins (Lipinski 2004, Jonsdottir 2005). More broadly, identifying molecules that can selectively interact with and modify the behaviour of particular proteins is fundamental not only for drug design, but also for chemical genomics (Schreiber 2003, Stockwell 2004, Dobson 2004). Being able to selectively perturb molecular pathways is key to systems biology (Ideker 2001) and our ability to reverse engineer, model, and understand these pathways. Finally, huge arrays of new small molecules can be produced in a relatively short time (Houghten 2000, Schreiber 2000).

In addition to their scientific and technological appeal, small molecules offer also technical advantages from an informatics standpoint. The space of small molecules is vast and largely unexplored. The current estimates for the total number of small

molecules are in the range of 10^{60} (Bohacek 1996). In contrast, only a few million molecules are found in the best current databases. Computers are bound to become an essential tool for exploring such chemical space (Baldi 2005). Finally, as we shall see in this tutorial, small molecules have simple compact representations that are suitable for developing fast search methods.

Over the past three years, a few groups have developed large, downloadable, publicly accessible repositories of compounds including, UCSF's ZINC (Irwin et al. 2005), NIH PubChem (<http://pubchem.ncbi.nlm.nih.gov>), Harvard's ChemBank (Strausberg et al. 2003), and UCI's ChemDB (Chen et al. 2005 (<http://cdb.ics.uci.edu>)). Aggregation and organization of datasets of chemical information allows for massive in silico processing that would be impractical or even impossible in a traditional experimental setting. In parallel with databases of compounds, it is important to develop also databases of chemical reactions. Here again the main databases (e.g.~Beilstein) are commercial, expensive, and of limited use for developing large-scale methods, for instance in reaction discovery and retrosynthesis. Needless to say, even with a small library of reactions, as reactions are applied to a database of compounds, the number of new compounds generated grows *exponentially*, raising important algorithmic challenges both from a database and a datamining/datasearching standpoint.

Similarity, Search, and Prediction

The central notion to developing search methods is the notion of *similarity* between molecules. Similarity is central not only for searching current databases, but also for searching virtual compounds, and discovering new reactions and retrosynthetic pathways. Similarity between molecules can be defined in many ways and based on different representations, ranging from SMILES strings, to 2D graph of bonds, to molecular surfaces, and 3 D structures. Creating efficient search tools for small molecules is far from hopeless, particularly because to a first degree of approximation, by breaking cycles in the 2D graph of bonds, molecules can be viewed as small trees. The trees are small because both the number of vertices (atoms) in a small molecule is relatively small, and the branching factor for organic molecules is small. Efficient techniques for storing and rapidly searching such data structures exist and can be further developed. Different kinds of similarities may be appropriate in different situations, and may have different computational costs. Efficient search requires combining multiple filters, with different resolutions and speeds.

Computational methods in chemistry can be organized along a spectrum ranging from Schrodinger equation, to molecular dynamics, to statistical machine learning methods. Quantum mechanical methods, or even molecular dynamics methods, are computationally intensive and do not scale well to very large datasets. These methods are best applied to specific questions on focused small datasets. Statistical and machine learning methods are more likely to yield successful approaches for rapidly sifting through large datasets of chemical information. Similarity is also essential, in the form of statistical machine learning kernels, for developing methods that can predict the chemical, physical, and biological properties of molecules from training examples. This is not too surprising since, given an annotated training set of molecules (e.g. toxic/non-toxic), the properties of a new molecule ought to be inferred from its similarities to the molecules in the training set. Good kernels can be derived from different molecular representations (1D, 2D, 3D, etc). Spectral kernels in particular, counting the number of occurrences of each possible substructure, lead to efficient molecular ``fingerprints" and similarity measures that are useful both in database searches and statistical machine learning applications (Ralaivola et al. 2005, Swamidass et al. 2005).

In short, the notion of chemical similarity is complex and central to chemoinformatics. Understanding, modelling, and measuring chemical similarity are central computational tasks from which many applications can be derived. Thus in logical order this tutorial is organized around: (1) molecular compound and reaction data and representations; (2) similarity measures, search, and prediction; (3) applications, including molecular docking and drug discovery/screening.

2 Molecular Models, Representations, and Annotations

Communicating Chemical Data

Like any scientific discipline, studying chemistry requires the ability to catalogue and communicate large amounts of data. Unlike most disciplines however for which such data is restricted primarily to text and numbers, chemistry has the additional special challenge of modelling and representing molecules in a consistent manner, amenable to communication (Gasteiger 2003). The standard valence-bond model of chemistry with respective 2D depictions is the most commonly used and understood representation with which non-informatics inclined chemists would naturally communicate. This model accounts for discrete atoms with lines drawn between them to represent bonds, which themselves are abstractions of shared electron pairs. While these 2D sketch representations of molecules are convenient and intuitive for chemists, communication and processing, particularly informatics processing within a computer, requires the meaning in these graphical depictions to be codified into a reproducible representation.

2D Graph of Atoms and Bonds

The graphical depiction maps very well to a labelled graph model with atoms mapping to labelled nodes and bonds mapping to labelled edges. Note that it is common shorthand when depicting organic structures (such as that shown in the respective slide) to assume unlabeled nodes as carbon atoms and that hydrogen atoms implicitly populate all atoms based on standard valence rules, which indicate the expected number of connections (bonds) each type of atom has. For example, the standard valences for common organic elements include 4 for carbon, 3 for nitrogen, 2 for oxygen and 1 for all halides (fluorine, chlorine, bromine and iodine).

2D Data Formats

Assuming a molecule is modelled as such a labelled graph, there are two common ways to encode these into a format fit for computer processing. The first is a graph adjacency matrix, or bond matrix for molecules (Dugundji 1973). Such a matrix M has one row and column for each atom in the molecule with each element M_{ij} equal to the bond order between the i th and j th atoms or zero if the atoms are not bonded. The respective slide shows the graphical depiction of acetamide with sequence numbers labelling each atom, along with a respective bond matrix. Note that this is more specifically an example of a bond-electron matrix because the diagonal elements further specify the number of free electrons at the atom (2 pairs for

oxygen, 1 pair for nitrogen). This confers useful matrix properties, in particular that the sum over a column or row equals the number of valence electrons for the respective atom (assuming all hydrogens are also explicitly accounted for in the matrix).

The adjacency matrix format is relatively simple and powerful, with some chemically meaningful mathematical properties, but in general, it is a very sparse matrix (mostly zeros) with size proportional to the square of the number of atoms. A more compact representation would be a connection table that simply lists all of the atoms and then only the bonds that exist, referencing the atoms by an index position. The size of such tables will only grow linearly with respect to the number of atoms and bonds. A couple of the most commonly accepted molecular file formats, SDF and Mol2, use just such a representation. While these are very useful and the most widely used formats for computer storage and transmission of molecular data, they would be too complex to expect a human to systematically read and write.

1D Line Notations

Line notations, which can describe a molecule's complete constitution and connectivity with a single line of text, are desirable to facilitate rapid communication of molecular structures, especially in this age over the Internet. Furthermore, a notation that is human readable and writeable would greatly facilitate human interaction with any chemistry information system.

A few of the most important such chemical line notations include nomenclature systems to assign names to molecular structures (systematic and common), SMILES strings that were originally proprietary but have since become the de facto standard for much chemical communication, as well as the more recently developed InChi standard (<http://www.iupac.org/inchi>) officially supported by IUPAC (though not yet as widely accepted). The accompanying slide shows a molecular structure and the respective line notations for each of the mentioned schemes.

IUPAC Nomenclature

The International Union of Pure and Applied Chemistry (IUPAC) is an organization of chemists that developed a systematic naming scheme for molecules. The accompanying slide depicts the IUPAC standard names for a series of incrementally more complex molecules from propane to 2-amino-3-hydroxy-propanoic acid. This is a fairly well established system that should produce unambiguous chemical

names, though not necessarily unique names due to inconsistent application by different parties. The more significant problem with the system however is simply the extensive and complex nature of all of the naming rules, which makes both the composition and the actual length of names for larger complex molecules quite unwieldy. Alternatively, many structures or substructures are known by common / trivial names such as the slide example of 2-amino-3-hydroxy-propanoic acid which is much more commonly known as the amino acid serine. Such common names are extensively used in communicating chemical and biological information, but obviously these do not lend themselves to systematic name translation methods, particularly for novel compounds.

SMILES Basics

SMILES is a chemical line notation in widespread use for communicating structure information with chemical informatics services and databases with a relatively simple set of construction rules.

1. Each atom is represented by it's atomic symbol
2. Bonds are represented by special characters, distinct by bond order
 - a. Single bond: - (dash, implicit, need not be specified)
 - b. Double bond: = (equals)
 - c. Triple bond: # (hash)
3. Parentheses indicate a structure branch
4. Matching numerical annotations indicate atoms connected in a cycle
5. Hydrogens are implicitly assumed based on standard valence rules, though they may be specified for non-standard cases such as charged species

The accompanying slides include several examples demonstrating these simple rules. Additional extensions to the SMILES grammar allows for the specification of additional properties such as formal charges, stereochemistry, aromaticity, composite molecules and reactions.

Canonical Representations

Most of the representations discussed so far produce an unambiguous encoding of the molecular structure. That is, given the encoding, we can reliably reproduce the molecule that it came from. However, these have also been non-unique encodings, meaning a single structure can produce many different but equivalent encodings. A unique encoding with a one-to-one mapping between structure and encoding would be much more desirable when we wish to address questions such as verifying the

uniqueness of a molecule amongst a pool of molecules or performing a rapid database lookup for a molecule record.

Coming up with such a unique encoding, such as the so-called "canonical SMILES" (D Weininger 1989) essentially comes down to finding a unique and consistent manner to sequentially order the atoms of a molecule. For N atoms, there are $N!$ such orders (~3.6 million for 10 atoms) out of which a single one must be selected consistently. One important algorithm for accomplishing this effect is the Morgan graph algorithm that iteratively labels nodes (atoms) based on their connectivity and the connectivity of their neighbors to establish an extended connectivity (EC) value for each node. The nodes can then be sequentially numbered, essentially based on their EC ranking, with tie-breaking by atom and bond distinctions. The algorithm works very well in general, though it can be broken by some confounding structures with high structural symmetry. For practical purposes, the point is that a one-to-one mapping between molecular structure and computational representation exists.

Stereochemistry / Isomers

In our discussions of molecular structure thus far, we have really only considered the topological connectivity of the molecules, specifying which atoms are bonded to which. In actuality however, atoms will have topographical spatial relationships with respect to each other. This brings us to the issue of stereochemistry regarding molecules that have identical connectivity but are not super-imposable in real space due to distinct spatial configurations, conferring a "handedness" to molecules.

This primarily occurs in two instances for organic chemicals. Atoms with at least 4 distinct connections (typically carbon with a tetrahedral geometry) are not super-imposable with its mirror image. Double bonds that have distinct components on both ends are not super-imposable on the equivalent structure with constituents on one side of the bond flipped. Note that this is because double bonds resist rotation and will normally maintain a fixed planar configuration.

Note that it is very common for chemicals that are otherwise identical except for stereochemical configuration to have completely distinct biochemical effects, generally because biological receptor structures themselves have non-symmetric spatial configurations. To fully specify a molecule's configuration then, we must also label stereospecific atoms and double bonds such as in the isomeric SMILES strings found in the accompanying slide.

3D Atomic Coordinates

Beyond even stereochemical configuration, a more complete molecular representation would indicate the complete 3D spatial coordinates of the atoms. Unfortunately, such complete 3D structures are only known for a small fraction of the millions of known molecules with about 300,000 available in the Cambridge Structure Database (<http://www.ccdc.cam.ac.uk/>). Such coordinates are essential for more advanced analysis of physical, chemical and biological properties of chemicals however, so many structure prediction packages have been developed to fill this gap such as CORINA (J. Sadowski 1994).

4D Conformers

With respect to bond lengths and angles, molecules are fairly rigid structures. However, with respect to torsion angles around single bonds, molecules are quite free to rotate. Different conformations of such torsion angles for a single molecule specify different conformers of that molecule. Specifying any single rigid 3D structure for a molecule is thus misleading as it discounts the flexibility of the molecule, only accounting for a single conformer. A more complete representation would account for all conformations, but this would be an unmanageable number, so a more common approach is to sample several low energy conformations for a single molecule.

Molecular Surfaces

When considering intermolecular interactions, one final representation discussed here, the molecular surface, can be especially important. For intermolecular interactions, the "interior" of a molecule is relatively unimportant since the solvent and other molecules can never "see" the interior. Instead, only the solvent-accessible molecular surface (Richards 1977) and physicochemical properties there should be primarily relevant.

Of course, molecules and atoms have no real hard surface in the macroscopic sense. In the microscopic sense, at best they have electron probability density isocontours or Van der Waals radii where atomic attraction forces are overtaken by repulsion forces. The accompanying slide demonstrates conceptually how a solvent-accessible molecular surface can be constructed. To begin with, the Van der Waals radii for each of the three atoms are traced out in red. Probe spheres in blue representing the standard radius of the solvent (usually a water molecule) are used to trace out the surface of the atom Van der Waals radii. This will mostly

correspond to the atomic radii themselves where the probe sphere can contact the atoms, but at certain concave areas of the structure, the probe sphere cannot completely contact as shown. Tracing out the border of the probe sphere in these regions yields a completed and smooth molecular surface, all of whose points are accessible by the solvent.

Valence Model Limitations

All of the models discussed thus far have been based on the valence-bond model of chemistry with atoms connected by one or more bonds. However, bonds are only models for shared electrons across molecular orbitals and as a result valence-bond based models have inherent difficulty modelling certain concepts such as aromaticity, resonance and tautomers. The accompanying slide illustrates several pairs of molecules or atoms that are chemically equivalent, despite the fact that the valence-bond model suggests they are distinct. The Representation Architecture for Molecular Structures by Electron Systems (RAMSES) (S. Bauerschmidt 1997) is at least one computational representation that has been developed to address many of these shortcomings by more directly modelling the molecular orbital systems. Unfortunately, these have yet to see widespread use and acceptance, probably in large part due to the entrenchment of the valence-bond model in chemical communication in general, let alone chemical informatics.

3 Chemical Similarity and Searching

The Similarity Problem

Assessing similarity between chemicals is a fundamental operation in chemical informatics. Good measures of similarity allow us to construct meaningful database indexes, predict properties of molecules, cluster groups of related compounds, and even de-noise screening datasets (Klon, Glick et al. 2004; Camastra and Verri 2005; Swamidass, Chen et al. 2005).

Similar chemicals have similar properties. Chemicals similar to estrogen are more likely to bind estrogen receptor (ER) than other chemicals. We can imagine computing similarity between chemicals along several different dimensions with varied importance for given applications. We could, for example, measure similarity in terms of the size, the shape, the polar surface area, or atom composition. Different similarities will have higher correlation with different properties. For example, compounds with similar polar surface areas will have similar logP, and compounds with similar shapes will tend to bind similar protein pockets.

So the question now becomes: what are fast ways to compute meaningful similarity between chemicals?

The Historical Progression

Database searching naturally introduces basic algorithms in computing chemical similarity. The earliest similarity measures were directed at just this application (Daylight Chemical Information Systems 1992).

One of the earliest ways chemical similarity was the size of Maximum Common Substructure (MCS) of the atom-bond graphs of the two chemicals. MCS reduces to the subgraph isomorphism problem which is known to be NP-complete in the general case. On chemical graphs, this algorithm normally works in polynomial time: tolerable for small datasets. As databases grew in size, a new method, structural keys, was used to pre-filter a database before running MCS.

Structural keys are bitmaps, vectors of ones and zeros. Each bit position in the key corresponds with a predetermined structure. If the structure is in the chemical its key's bit corresponding to that structure will be set to one. For example, if the first bit corresponds to a benzene ring, this first bit of the corresponding key will be set to one. We can now do fast, linear-time comparisons between molecules by assessing the similarity between these fixed-length, pre-computed keys. Subset searches on well constructed keys correspond well with exact substructure queries. Structural keys have the useful property that the key of molecule A is a subset of the key of molecule B if and only if A is a subset of B.

Structural keys require us to choose before knowing the application a set of structures to look for in chemicals. This is a problem. In some applications certain structures are important while in others they are irrelevant. We manually choose the structures in the key with our goals in sight. Our key, however, has limited utility for other applications. How do we create a general structural key which can be used for many different applications?

Fingerprints were designed be more broadly useful than structural keys. Rather than only searching for structures in a predetermined list, chemical fingerprints are constructed by enumerating *all* substructures of a certain size in a given molecule. We once again set a bit to one for each substructure observed in a molecule. However, rather than finding the bits position from a lookup table as we did for structural keys, we compute the corresponding bit's position by calculating the hash value of a canonical representation of the structural key. While improving the generality of structural keys, fingerprints still have the useful the property that the key of molecule A is a subset of the key of molecule B if and only if A is a subset of B. Fingerprints are the current standard in large molecular database searches.

Most fingerprint systems sequentially scan all fingerprints in a database to answer queries. Newer methods are being developed which can prune these scans using bounds on similarity (Swamidass and Baldi 2006). Additionally, Locality Sensitive Hashing (LSH) may be the next advance, allowing for $O(\log n)$ complexity searches of large datasets (Dutta, Guha et al. 2006).

Venn Similarity

How do we compute similarity between fingerprints? There are many different formula which can be used to compute similarity between two fixed length bitmaps, Euclidian distance, hamming distance, cosine angle, and more. Two measures of similarity for comparing chemical fingerprints arose early on and tend to produce the most useful measures of similarity.

These two standard formulas, Tanimoto and Tversky, can best be rationalized with a Venn diagram. The area in common corresponds with the number of features found both in chemical A and B, and the area not in common corresponds to the features observed in A alone and B alone, then Tanimoto similarity is the percentage overlap between the two Venn circles. It is computed as the area in common over the total area covered: i.e. the percentage overlap.

Superstructure and Substructure Searches

A molecule is a substructure of another if it is exactly contained inside the other. A superstructure is the opposite. If molecule B is a substructure of A, then A is a superstructure of B. Tversky similarity reduces to Tanimoto similarity if α and β are set equal to one. Choosing α and β correctly allows us to penalize the mismatched bits asymmetrically, allowing us to search for substructures or superstructures of our query.

2D Graph Substructures

How are *all* substructures enumerated? This is an implementation detail. Most systems use something like a depth-first search to enumerate all paths of a particular length. We can of course, imagine other types of substructures which maybe suited for certain applications.

There are polynomial time algorithms for enumerating all paths if the connectivity of a graph is bounded. In the case of chemicals, the branching factor is relatively low so we can safely apply exhaustive, exact, algorithms. Even though the space of all possible paths is quite large and difficult to count, the number of paths in a single molecule is manageable. Fingerprints are all pre-computed and stored in the database, so once the index is created, comparisons are linear and rapid.

Mapping Structures to Bits

For fingerprints, how do we map substructures to particular bit positions? The algorithm is simple: 1) find a canonical representation of the substructure, 2) compute a good hash value of this representation, and 3) compute the bit position by calculating the modulus of the hash value and the length of the fingerprint. So, to construct a fingerprint, we set the bits corresponding with every substructure we enumerate.

An obvious concern with this algorithm is that sometimes different substructures will be mapped to the same bit position. We refer to this conflict as a clash. If two substructures we observe clash, we set their bit position equal to one ($1+1=1$). This amounts to a sort of lossy compression of data. Each bit position corresponds to a family of unrelated structures.

The Fingerprint Approximation

So fingerprints can be thought of as a compressed representation of a very long structural key. Within certain limits, Tanimoto similarity computed between fingerprints approximates similarity computed on this longer structural key. What

makes this approximation fall apart? The more clashes the more error. The higher the density of ones in a fingerprint, the more error in computing its similarity with other fingerprints.

2D MinMax

Fingerprints as described do not consider the frequency/counts of substructures. So no distinction is made between molecules with one benzene ring vs. two benzene rings other than a unique path which might cross over both rings in one molecule but not the other. It has been shown that measures that appropriately consider the counts of substructures correspond more directly with molecular properties. MinMax is one of the best performing measures. MinMax is a generalization of Tanimoto similarity which incorporates information about the counts of substructures (Ralaivola, Swamidass et al. 2005).

Fingerprint Similarity is a Spectral Similarity/Kernel

Computing similarity between objects is a fundamental operation in Machine Learning as well as Chemical Informatics. Kernel methods have been established as a powerful method of solving classification, regression, visualization and clustering problems (Camastra and Verri 2005; Swamidass, Chen et al. 2005).

It is important to note that Tanimoto, Tversky (when $\alpha=\beta$) and MinMax similarity computed between either fingerprints or chemicals are examples of spectral kernels. Spectral kernels are a type of similarity computed by 1) enumerating all substructures of an object and 2) comparing these enumerations. Tanimoto and MinMax similarity are Mercer kernels, therefore they can be used as the core of any general kernel methods in order to solve chemical problems.

Normal Distribution of Fingerprints

What are the statistical properties of fingerprints? What is their distribution? Lessons from sequence analysis have shown us that these questions can help us design more efficient powerful algorithms.

Using a set of 50K random chemicals and plotting the distribution of the number of bits sets to one, we can see that fingerprint bit counts are distributed approximately normally.

Pruning Search Space Using Bounds

We can bound the similarity of a given database molecule using a simple formula. This bound is dependent on bit count of the query and the bit count of the database

fingerprint. If we are only interested in molecules of a 0.9 or greater similarity to our query, we can prune most of the database by using the bounds formula. This can dramatically accelerate database searches, on average as much as an 8x improvement in speed.

Speedup from Pruning

This speedup depends on the similarity threshold we choose and the bit count of the query. The speedup can range from 100-fold to 2-fold speedup. It will never be worse than a sequential scan.

Aggregate Queries

We can imagine situations where we would like to search a database use a group of chemicals as the query. For example, we may want to search for all molecules which bind ER by querying by all known binders of ER. There are many ways to think of constructing this sort of aggregate query. This is an active area of research which will hopefully lead to new sorts of more accurate searches.

4 Chemical Reactions

Basic Principles

Reactions represent the dynamic nature of chemistry whereby different compounds can interconvert between one another, perhaps yielding energy for biological pathways or for constructing an industrial polymer under regulated conditions. At a minimum, specifying a chemical reaction requires identification of the chemical structures of the reactants and subsequent products. Supplementary information such as any catalysts used in the reaction, solvent and temperature conditions, etc. are not strictly necessary to understand the chemical structural changes that occur in the reaction, but can be very useful for building and understanding reaction knowledge bases and for practical application.

Given the reactants and products for a reaction, the reaction center is the specific substructure of atoms and bonds that are actually rearranged. More generally, it refers to the functional groups of the reactants and how they are rearranged to form the products.

Needed Information

For chemical informatics, a complete reaction specification must include more than just the reactant and product structures. A complete specification must also include a mapping between the reactant and product atoms, at least at the reaction center. Without this, ambiguous mechanistic pathways for transforming the reactants into products could be inferred such as the slide example either showing the hydroxyl group directly substituting the bromide or indirectly by adding to the double-bonded carbon on the other end.

Furthermore, a complete and correct reaction specification must respect conservation of mass in the universe by fully specifying a stoichiometrically balanced reaction equation. That is, over the course of a reaction, no atoms or electrons can be created or destroyed. Unfortunately, "trivial" reactants and products such as a water molecule in condensation reactions are often neglected in reaction specifications, making it much harder to systematically process them.

Note that for practical chemistry, even more information is necessary including reaction catalysts, solvent and temperature conditions, yield, rate and other factors that are necessary to reproduce the reaction in a laboratory.

Reaction Databases

While small molecule databases are becoming more common and available today, databases cataloguing reaction information are still generally in a poor state. Most repositories consist only of thousands of records, perhaps millions, but in general the data has poor consistency. In particular, the data is often incomplete with respect to balanced stoichiometry and reaction conditions (Gasteiger 2006). Well-developed, publicly available reaction databases have not been identified by this group, though some privately licensed ones are referenced in the literature such as the CASREACT and ChemInform RX systems.

To search through reaction databases, at least for the structural component of the data, we can reuse many of the same search techniques for simple chemicals. Searching for reactions by reactant or product structure is fundamentally no different than a simple chemical search. Alternatively, one can search based on just those atoms and bonds which change over the course of a reaction to focus in on reaction centers and thus find reactions of similar class.

Virtual Chemical Space

Once a collection of reaction profiles is known and available in a computational representation, this offers us the power to address such problems as exploring virtual chemical space. Searching for chemicals in a database similar to a query molecule has already been well established, but consider the target structure molecule in the accompanying slide. No structure in the UCI ChemDB (Jonathan Chen 2005) is found to be directly similar to it. If this were theoretically a very important compound however, we could instead search for it in virtual chemical space that is just one reaction away from the directly available chemical space represented by the database of available chemicals. We can accomplish this by applying the retro form of one of our reactions (Diels-Alder in this case) to produce a pair of precursor molecules. Searching for similar chemicals to each precursor independently does yield several similar results. Reapplying the normal forward version of the reaction to each pair of similar results yields theoretical compounds that are not directly available in the database, but should be indirectly accessible by applying one reaction to pairs of readily available compounds

Knowledge Based Reactions

Exploring a virtual space of chemicals is one example of the power and utility of having these reactions. The most common way of working with reactions

computationally is what we refer to as “knowledge based reactions.” These explicitly specify what functional groups can react with one another and precisely how to rearrange the atoms and bonds to form the product. This specific and discrete representation can be convenient for many uses, but has its limitations.

Knowledge Based Limitations

Even if reaction databases were better developed and curated, a knowledge based method requires manual pre-specification of many different reaction profiles to achieve any level of generality. The accompanying slide illustrates 3 example reactions, all of which would require a separate reaction profile specification for the computer to understand how to process them.

Reaction Discovery

One reaction research area then is to discover reaction profiles by more general principles, with the virtue that this would not be limited to existing knowledge bases. Furthermore, if doing so allowed us to discover wholly new and novel reaction schemes that chemists haven’t already discovered, this would already be inherently useful as a chemist’s tool and could even suggest leads for targeting biologically relevant functional groups. For example, the post-translational modification of nitro-tyrosine is a known marker for diseases such as coronary artery disease, so if we could find a reaction scheme that uniquely reacts with the nitro-tyrosine functionality, that could be used to probe that disease system, determine if it is a causative agent, and maybe even offer a therapeutic lead if the reaction product alters the disease process.

This problem of predicting how two arbitrary chemicals will react is essentially solvable with quantum chemical methods, but this is too demanding computationally to be done on a large scale. Systems developed to predict chemical reactions using more approximate theoretical concepts such as partial charge and frontier molecular orbitals include CAMEO (Julia Schmidt Burnier 1984) and EROS (Robert Hollering 2000).

We present here a simplified approach to reaction discovery, touching on several concepts used in such systems. To discover reaction profiles from more general principles, consider that the very simple reaction profile in the accompanying slide involving any four atoms where the bonds just exchange positions already accounts for all of the reactions in the previous slides, and in fact about 50% of organic

reactions (Johann Gasteiger 2003). The accompanying slide includes the concrete example of an amide bond formation. Here the carbon, chlorine, nitrogen and hydrogen are the four atoms A, B, C and D. Removing the 2 original bonds and exchanging them for bonds in the other direction yields the expected amide product. This simple approach has already covered 50% of all reactions, but of course this general profile has issues.

Generic Reaction Profile Issues

Allowing any four atoms to exchange bonds in this generic pattern yields many unreasonable products like those shown in the accompanying slide. Furthermore, there are many reactions with more sophisticated profiles not covered by this scheme such as the Diels-Alder and azide + alkyne aromatic cyclization reactions, and others where the reaction involves more than 4 atoms. As is, these will not be covered unless we manually specify more knowledge-based profiles, perhaps involving 6 atoms.

Reaction Favorability Scoring

To address the first issue of unreasonable reaction predictions, we need some scoring system to suggest reaction favourability. One such mechanism based on thermodynamic favourability is illustrated in the accompanying slide by estimating the change in enthalpy of proposed reactions, and only taking those with energetically favourable changes. A simple additive method to do so is to simply look up bond-dissociation energies for all of the bonds in the reactants and products and assess which side of the reaction is more stable in that respect. For additional robustness, this scoring method can offer stability bonuses and penalties for aromatic compounds and compounds with ring strain. Other schemes that consider additional effects beyond thermodynamics such as reaction kinetics provide greater accuracy, but quantitative data is much less available.

Pseudo-Mechanistic Reactions

To address the issue of modelling reactions with reaction centers more sophisticated than 4 atoms, we can try a more formal “pseudo-mechanistic” approach. The main addition is introducing the concept of intermediates into the reaction predictor. For example in the accompanying slide, using the same basic 4 atom reaction profile, instead of directly exchanging the bonds, we first model the shifting of the bond electrons to the attached atoms by just applying formal positive

and negative charges. In that case, closing the intermediates to produce the product is just a matter of matching + and – charges and closing the bonds.

Thus far, we added an extra intermediate step but still have the same 4 atom bond rearrangement profile. The difference is that we can now allow these intermediates to rearrange themselves based on general electron-shifting rules before reclosing into products. The accompanying slide shows an enol with its OH bond opened to create an intermediate. Rather than allowing it to immediately react with another intermediate to create a product, we can apply basic electron shifting rules to yield an equivalent intermediate since it now has a negative charge (representative of a lone pair of electrons) adjacent to an atom with a π orbital double bond. Putting all of these pieces together, many known reactions can be discovered by basic principles, including the Diels-Alder and azide+alkyne aromatic cyclization reactions depicted in the slides.

Chemical Synthesis

As a simplified concept, chemical reagents applied to appropriate chemical reactants will result in a reaction. The accompanying slide illustrates a simple chemical synthesis pathway, which is simply a chain (or tree) of several reactions applied to starting reactants to reach a final product. An important reaction based research problem is to reverse-engineer these synthesis pathways.

Synthesis Design Problem

A standard setup for chemical synthesis design problems is to be given a desired target molecule (e.g., a natural product drug), a collection of readily available starting reactants (e.g., a chemical vendor catalog), and known reagents that can perform reactions on those reactants (i.e., a reaction database). The goal then is to find a proper combination and sequence of reaction reagents to apply to the reactants to generate the product.

Performing an exhaustive search to divine the synthesis pathway by recursively applying all known reactions to all available starting material reactants would be intractable. The starting material pool itself could consist of millions of chemicals. Alternatively, a retro-synthetic approach (E.J. Corey 1985) starts from the product and computationally applies retro-reactions (transforms) to generate precursors until it can trace a path back to available starting materials. Existing packages such as LHASA, SECS and SYNCHEM apply this basic methodology (Todd 2004), with the

former two calling upon human interaction to guide the pathway search. Other packages like SST and CHIRON attempt the forward direction of search in a sense, though only in terms of looking for abstract structural pattern correlations between the starting materials and products, not tracing out a specific reaction pathway. Other packages like IGOR, EROS and SYNGEN use more formal methods to model the reactivity of molecules, lending themselves well to extensions like CAMEO and WODCA to predict whole new reaction schemes in manners similar to that discussed earlier.

Retro-Synthesis Example

The accompanying slide illustrates the framework for the retro-synthetic search strategy. Given a target molecule, we apply known reactions in reverse to produce several possible precursors. If one of these precursors is found amongst the available starting materials, the search is complete. If not, we can recursively search for a retro-synthesis pathway for the best precursors. This is comparable to a search space with a branching factor of P where P is the number of possible precursors generated for each target product. For large numbers of known reactions and large, complex target products, this branching factor can be quite large, necessitating heuristic measures to guide the search. For example, one could pursue only those precursor branches where the precursor has greater similarity to compounds in the starting material reactant pool.

5 Molecular Docking

The Docking Problem

In this section we consider Virtual Ligand Screening (VLS): the task of identifying chemicals, i.e. ligands, in a database which fit into a protein's binding pocket. There are other types of docking, including protein-protein and protein database screening, but we focus here on the task of screening a large database of chemicals for activity by geometrically fitting them into a given protein cavity.

Small molecule-protein docking is distinct from molecular modelling. Molecular modelling attempts to model feasible atomic trajectories as accurately as possible. Current methods require hours, days, or weeks to simulate a single ligand's trajectory into a binding pocket. Docking, on the other hand, can be rapidly calculated in minutes for each ligand. This allows tens of thousands of chemicals to be docked in just a few days on a cluster. This speed is achieved by simplifying the molecular dynamics formulation. Instead of searching for feasible trajectories for the whole system in dynamic motion, the ligand is allowed only to rotate bonds and the protein is kept rigid. Instead of simulating a trajectory, docking algorithms search for a single low energy geometric configuration, or pose.

There are two critical modules in any docking program: a search algorithm and a scoring function. The search algorithm heuristically searches for the best scored poses. The scoring function quickly computes how well a pose fits a ligand into the protein's cavity. These two modules work together to rapidly find low-energy poses.

Challenges

Screening databases by docking is a challenging task for a number of reasons. The search algorithm must robustly find good minima in a high dimension, variable size space full of local minima, singularities and sharp curvature. The scoring algorithm must robustly and rapidly screen out false negatives and correlate well with experimental binding affinity.

If we are screening a database for chemicals with activity in a biological system, we face additional problems. Experimental binding affinity (pK_d) does not always correspond with biological activity. For example, but estrogen receptor antagonists and agonists bind in the same pocket with high affinity. Each class causes different structural shifts in the protein resulting in different biological activities.

Docking is designed to answer one question: which chemicals bind this protein? If we find chemicals which bind our protein, we still have more questions to answer: what is the absorption rate, the distribution, the metabolism, and excretion of the chemical? These questions are collectively referred to as ADME concerns and are left unresolved by docking studies.

Despite these challenges docking can be useful as a first pass screen for more expensive, time-consuming experiments. Limiting experiments to compounds with higher chance of binding a protein can reduce the time and cost of discovering new drugs.

Scoring Function

Predicting binding affinity from a single pose is difficult. Many phenomenon are simultaneously at play as a ligand binds a protein: electrostatics, hydrophobic interactions, desolvation, loss of entropy as bonds are frozen in place, hydrogen bonding, protein flexibility, alternate binding positions, cavity accessibility, inclusion of precisely oriented active site waters, etc.

Docking programs do not model binding affinity from first principles. Their scoring functions use simplified approximations which can be rapidly computed from a single pose. For example, most programs use grid-based optimizations which allow for linear computation of energy at the cost of holding the protein rigidly in place.

Scoring functions must also screen out false positives with high efficiency. Even a low false positive rate can rapidly allow a large number of false positives to quickly overwhelm the small number of expected binders.

Search Algorithms

There are a large number of search algorithms used by docking programs to find good binding configurations. Most of them are tuned to the specific details of molecular energy functions. For example, many programs treat ligand flexibility using 'incremental construction,' to first dock in one rigid segment of a molecule and then grow out the rest(Ewing, Makino et al. 2001). Other methods include simulated annealing and other forms of Monte Carlo simulation(Bursulaya, Totrov et al. 2003).

Docking Programs

One of the oldest, and most affordable, docking programs is DOCK from Tack Kuntz's lab in UCSF (Ewing, Makino et al. 2001). It is open source software which can be run on large clusters of computers for very fast results.

There are number of other programs, some of which seem produce more accurate results than DOCK. However, most of them cost quite a bit for yearly single-processor licensing and are prohibitively expensive for cluster computing. Some of the best are ICM (Bursulaya, Totrov et al. 2003) and Glide (Halgren, Murphy et al. 2004). Using DOCK on a cluster to rank a large database of compounds and then rerunning the top few results on a slower, more accurate setup has been a workable strategy for screening large databases on reasonable budgets. Confirming the results of one docking program with the results of another controls some types of systematic error (Clark, Strizhev et al. 2002).

Cluster Based Computing

Docking is trivially parallelized by dividing input files across a large number computational nodes. Licensing costs are the typical barrier. For DOCK licensing is not per-processor, so we have automated large database runs using a combination of bash and python scripting. PVM and MPI are parallel computing standards which can be used to parallelize code, however they prove to be much more complex to use and less robust to equipment and software failures.

Visualization

Viewing docking results is surprisingly challenging. Common molecular viewers like SwissPDB Viewer, RasMol, and VMD are not designed to flip through a large number of molecules listed in one file. There are two good options: VIDA and Chimera. VIDA is free for academics from Open Eyes and Chimera from UCSF is free for everyone. Both these viewers have the ability to read a concatenated file of molecules and scroll through them one by one. This allows researchers to manually assess the quality of particular poses using familiar visualization tools.

6 Applications: Drug Screening/Design-A Case Study

Tuberculosis (TB)

TB is still a real threat. Multi-drug resistant (MDR) TB is difficult to treat and has a high morbidity and mortality rate. TB typically infects the lungs but can also cause serious infections in bones and even the digestive track. Treatment includes a six-month course of special antibiotics with undesirable side-effects. Additional drugs targeting TB could reduce side-effects of treatment, shorten treatment time, and provide physicians with additional therapeutic options for MDR TB. This results of this work have been recently published (Lin, Melgar et al. 2006).

The Cell Wall: Key to Pathogen Survival

TB is difficult to treat because of its mycolic acid cell wall. This cell wall is both a shield against most standard antibiotics and the target of effective drugs. Its cell wall is especially waxy and densely packed with a number of fatty acids unique to this family. Targeting the biosynthesis of these fatty acids kills TB and can cure the disease.

AccD5

AccD5 is an enzyme necessary for the synthesis of the TB cell wall. It is part of a family of Acyl-CoA-ases which elongate fatty acids. This particular enzyme is sufficiently different from human enzymes that it could be used as a drug target.

AccD5 Protein Structures

One of our collaborators crystallized and solved three different isozymes in the AccD family from TB. AccD forms a large halo-shaped hexamer. This structure was solved using x-ray diffraction experiments. Each of these enzymes were characterized by *in vitro* experiments showing activity and different specificities for different substrates.

Structure-Based Drug Design

We used an iterative method to prioritize compounds for experimental assays. For the first pass we docked a representative set of molecules from the ChemDB (Chen, Swamidass et al. 2005) using both ICM and DOCK into the active pocket of AccD5. The top few were assayed. We then searched the ChemDB for compounds similar with our confirmed positives and prioritized them based on further docking studies. The top few from this iteration were assayed again.

Two Strategies

We used a combination of two strategies to search ChemDB, similarity searches and docking simulations. Similarity searches produces compounds which look very much like known binders. Docking yields much more diverse results. Combining the two strategies heuristically explores the database and biases computation toward experimental information about known binders.

Identified Inhibitor

From these two studies we identified two inhibitors of AccD5. One of which has about 5 micro molar inhibition constant and kills about 50% of TB cells in culture at about 50 micro molar concentration. Ideally, further iterations of searching will find compounds with nano-molar inhibition constants and with favorable ADME. This example shows how a combination of docking and similarity searching can be used to find novel inhibitors of important protein targets.

7 REFERENCES

Agrafiotis,D.K., Lobanov,V.S. and Salemme,F.R. (2002) Combinatorial informatics in the post-genomics era. *Nature Reviews Drug Discovery*, **1**, 337–346.

F. R. Bach and M. I. Jordan. Kernel Independent Component Analysis. *Journal of Machine Learning Research*, **3**:1–48, 2002.

P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach*. MIT Press, Cambridge, MA, 2001. Second edition.

Baldi,P. and Pollastri,G. (2003) The principled design of large-scale recursive neural network architectures–DAG-RNNs and the protein structure prediction problem. *Journal of Machine Learning Research*, **4**, 575–602.

S. Bauerschmidt, J. G. (1997). "RAMSES: Representation Architecture for Molecular Structures by Electron Systems." Journal of Chemical Information and Computer Sciences **37**: 705-714

Bergstrom,C., Norinder,U., Luthman,K. and Artursson,P. (2003a) Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Model.*, **43** (4), 1177– 1185.

Bergstrom,C.A.S., Norinder,U., Luthman,K. and Artursson,P. (2003b) Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.*, **43** (4).

Berman,H.M.,Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N.,Weissig, H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucl. Acids Res.*, **28**, 235–242.

Brown, F. K. (1998) Chemoinformatics: What is it and How Does it Impact Drug Discovery? *Annual Reports in Medicinal Chemistry* **33**, 375-384.

Budavari,S.,E. (1996) The merck index, 12th ed. *Merck and Co, Inc Whitehouse Station*, Faraday Trans.

Julia Schmidt Burnier, W. L. J. (1984). "General Treatment of Periselectivity." Journal of Organic Chemistry **49**: 3001-3020.

Bursulaya, B. D., M. Totrov, et al. (2003). "Comparative study of several algorithms for flexible ligand docking." J Comput Aided Mol Des **17**(11): 755-63.

Camasta, F. and A. Verri (2005). "A novel kernel method for clustering." IEEE Trans Pattern Anal Mach Intell **27**(5): 801-805.

J. Chen, S. J. Swamidass, Y.Dou, J. Bruand, and P. Baldi. (2005). "ChemDB: A Public Database of Sm all Molecules and Related Chemoinformatics Resources." Bioinformatics, **21**(22): 4133-9.

Chang,C.C. and Lin,C.J. (2001) LIBSVM: a library for support vector machines. . available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Cherqaoui,D. and Villemin,D. (1994) Use of neural network to determine the boiling point of alkanes. *J. Chem. Soc. Faraday Trans.*, **90**, 97–102.

Clardy,J. and Walsh,C. (2004) Lessons from natural molecules. *Nature*, **432**, 829–837.

Clark, R. D., A. Strizhev, et al. (2002). "Consensus scoring for ligand/protein interactions." *J Mol Graph Model* **20**(4): 281-95.

Collins,M. and Duffy,N. (2002) Convolution Kernels for Natural Language. In *Adv. in Neural Information Processing Systems* 14.

E. J. Corey, A. K. L., Stewart D. Rubenstein (1985). "Computer-Assisted Analysis in Organic Synthesis." *Science* **228**(4698): 408-418.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK.

Daylight Chemical Information Systems, I. (1992). Daylight Theory Manual.

Debnath,A.K., Lopez de Compadre,R.L., Debnath,G., Shusterman, A.J. and Hansch,C. (1991) Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, **34**, 786–797.

Delaney,J.S. (2004) ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput.Sci.*, **44** (3), 1000–1005.

Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.
Y. Dou, P. Baisnee, G. Pollastri, Y. Pecout, J. Nowick, and P. Baldi. ICBS: a database of interactions between protein chains mediated by beta-sheet formation. *Bioinformatics*, 20(16):2767–2777, 2004.

Dumais,S., Platt,J., Heckerman,D. and Sahami,M. (1998) Inductive learning algorithms and representations for text categorization. In *Proc. of the 7th Int. Conf. on Information and Knowledge Management (Bethesda, MD)*.

Dutta, D., R. Guha, et al. (2006). "Scalable partitioning and exploration of chemical spaces using geometric hashing." *J Chem Inf Model* **46**(1): 321-33.

J. Dugundji, I. U. (1973). "(Bond-Electron Matrix)." *Topics in Current Chemistry* **39**: 19-64.

Ewing, T. J., S. Makino, et al. (2001). "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases." *J Comput Aided Mol Des* **15**(5): 411-28.

Fligner,M.A., Verducci,J.S. and Blower,P.E. (2002) A Modification of the Jaccard/Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics*, **44**, (2), 110–119.

- Flower,D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. of Chemical Information and Computer Science*, **38**, 378–386.
- P. Frasconi, M. Gori, and A. Sperduti. A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, **9**(5):768–786, 1998.
- Y. Freund and R. E. Schapire. Large Margin Classification Using the Perceptron Algorithm. *Machine Learning*, **37**(3):277–296, 1999.
- B. J. Frey. *Graphical Models for Machine Learning and Digital Communication*. MIT Press, 1998.
- Frimurer,T.M., Bywater,R., Naerum,L., Lauritsen,L.N. and Brunak, S. (2000) Improving the odds in discriminating “drug-like” from “non drug-like” compounds. *Journal of Chemical Information and Computer Sciences*, **40**, 1315–1324.
- Gärtner,T., Flach,P.A. and Wrobel,S. (2003) On Graph Kernels: Hardness Results and Efficient Alternatives. In *Proc. of the 16th Annual Conf. on Computational Learning Theory and 7th Kernel Workshop* pp. 129–143 Springer Verlag, NY, NY.
- Gasteiger, J. (2006). "Chemoinformatics: a new field with a long tradition." *Anal Bioanal Chem*(384): 57-64.
- J. Gasteiger, T. E. and Engel, T. (Editors) (2003). *Chemoinformatics: A Textbook*. Wiley
- Gasteiger,J., Sadowski,J., Schuur,J., Selzer,P., Steinhauer,L. and Steinhauer,V. (1996) Chemical information in 3D-space. *Journal of Chemical Information and Computer Sciences*, **36**, 1030–1037.
- Gower,J.C. and Legendre,P. (1986) Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, **3**, 5–48.
- Gower,J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.
- Hadjipavlou-Litina,D. and Hansch,C. (1994) Quantitative structureactivity relationship of the benzodiazepines. A review and reevaluation. *Chemical Reviews*, **94**, 1483–1505.
- Halgren, T. A., R. B. Murphy, et al. (2004). "Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening." *J Med Chem* **47**(7): 1750-9.
- Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143** (1), 29–36.
- Hann, M and Green, R. (1999). Chemoinformatics—A New Name for an Old Problem? *Current Opinoin in Chemistry and Biology*, **3**, 379-383.
- Heckerman,D. (1998) A tutorial on learning with Bayesian networks. In *Learning in Graphical Models*, (Jordan,M., ed.),. Kluwer Dordrecht.
- Helma,C., King,R.D., Kramer,S. and Srinivasan,A. (2001) The predictive toxicology challenge 2000-2001. *Bioinformatics*, **17** (1), 107–108.

R. Herbrich. *Learning Kernel Classifiers, Theory and Algorithms*. MIT Press, 2002.

Robert Hollering, J. G., Larissa Steinhauer, Klaus-Peter Schulz, Achim Herwig (2000). "Simulation of Organic Reactions: From the Degradation of Chemicals to Combinatorial Synthesis." Journal of Chemical Information and Computer Sciences **40**: 482-494

Hohmann,A.G., Suplita,R.L., Bolton,N.M., Neely,M.H., Fegley,D., Mangieri,R., Frey,J.K., Walker,J.M., Holmes,P.V., Crystal,J.D., Duranti,A., Tontini,A., Mor,M., Tarzia,G. and Piomelli,D. (2005) An endocannabinoid mechanism for stress-induced analgesia. *Nature*, **435**, 1108–1112.

Hou,T.J. and Xu,X.J. (2003) ADME evaluation in drug discovery. 2. Prediction of partition coefficient by atom-additive approach based on atom-weighted solvent accessible areas. *Journal of Chemical Information and Computer Sciences*, **43**, 1058–1067. Erratum in Issue 44(4) of the same journal, 1516-1516, (2004).

Houghten,R.A. (2000) Parallel array and mixture-based synthetic combinatorial chemistry: tools for the next millenium. *Annual Review of Pharmacology and Toxicology*, **40**, 273–282.

Huuskonen,J. (2000) Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *Journal of Chemical Information and Computer Sciences*, **40** (3), 773–777.

T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science*, **292**:929–934, 2001.

J. J. Irwin and B. K. Shoichet. ZINC—a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Computer Sciences*, 45:177–182, 2005.

Jaakkola,T., Diekhaus,M. and Haussler,D. (1999) Using the Fisher kernel method to detect remote protein homologies. *Intelligent Systems for Molecular Biology*, 149–158.

James,C.A., Weininger,D. and Delany,J. (2004) *Daylight Theory Manual*. Available at <http://www.daylight.com/dayhtml/doc/theory/theory.toc.html>.

Jonsdottir,S.O., Jorgensen,F.S. and Brunak,S. (2005) Prediction methods and databases within chemoinformatics: Emphasis on drugs and drug candidates. *Bioinformatics*, **21**: 2145-2160.

Karthikeyan,M., Glen,R. and Bender,A. (2005) General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.*, **45** (3), 581–590.

Kashima,H., Tsuda,K. and Inokuchi,A. (2003) Marginalized Kernels between Labeled Graphs. In *Proc. of the 20th International Conference on Machine Learning (Washington, DC)* pp. 321–328.

Kaiser,J. (2005a) Chemists want NIH to curtail database. *Science*, **308**, 774.

Kaiser,J. (2005b) House approves 05% raise for NIH, comments on database. *Science*, **308**, 1729.

Kazius,J., McGuire,R. and Bursi,R. (2005) Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, **48** (1), 312–320.

King,R.D., Muggleton,S.H., Srinivasan,A. and Sternberg,M.J.E. (1996) Structure-Activity Relationships Derived by Machine Learning: The Use of Atoms and their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. *Proc. of the National Academy of Sciences*, **93** (1), 438–442.

R. D. King, A. Srinivasan, and M.J. E. Sternberg. Relating chemical activity to structure: an examination of ILP successes. *New Generation Computing*, **13**:411–433, 1995.

Klon, A. E., M. Glick, et al. (2004). "Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results." *J Med Chem* **47**(11): 2743-9.

Koza,J. (1994) Evolution of a computer program for classifying protein segments as transmembrane domains using genetic programming. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, (Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D., eds),. AAAI Press Menlo Park, CA pp. 244–252.

Kramer,S. and De Raedt,L. (2001) Feature construction with version spaces for biochemical application. In *Proc. of the 18th Int. Conf. on Machine Learning* pp. 258–265. Lauritzen,S.L. (1996) *Graphical Models*. Oxford University Press, Oxford, UK.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, UK, 1996.

A. R. Leach and V. J. Gillet (2005). *An Introduction to Chemoinformatics*. Springer.

Leslie,C., Eskin,E., Cohen,A., Weston,J. and Noble,W. (2004) Mismatch String Kernels for Discriminative Protein Classification. *Bioinformatics*, **20** (4), 467–476.

Leslie,C., Eskin,E., Cohen,A., Weston,J. and Noble,W.S. (2003) Mismatch string kernels for SVM protein classification. In *Advances in Neural Information Processing Systems*, (S. Becker,S.T. and Obermayer,K., eds), vol. **15**,. MIT Press Cambridge, MA pp. 1417–1424.

Leslie,C., Eskin,E. and Noble,W.S. (2002) The spectrum kernel: a string kernel for svm protein classification. In *Proc. of the Pacific Symposium on Biocomputing, 2002* pp. 564–575.

T. Lin, M. Melgar, S. J. Swamidass, J. Purdon, T. Tseng, G. Gago, D. Kurth, P. Baldi, H. Gramajo, and S. Tsai. Structure-Based Inhibitor Design of AccD5, an Essential acyl-CoA Carboxylase Carboxyltransferase Domain of *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences USA*, **103**, 9, 3072-3077, (2006).

Lipinski,C. and Hopkins,A. (2004) Navigating chemical space for biology and medicine. *Nature*, **432**, 855–861.

Lipinski,C.A., Lombardo,E., Dominy,B.W. and Feeney,P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, **23** (3), 3–25.

Lodhi,H., Saunders,C., Shawe-Taylor,J., Cristianini,N. and Watkins, C. (2000) Text Classification using String Kernels. *Journal of Machine Learning Research*, **2**, 419–444.

Mahé,P., Ueda,N., Akutsu,T., Perret,J.L. and Vert,J.P. (2004) Extension of Marginalized Graph Kernels. In *Proc. of the 21st Int. Conf. on Machine Learning*, New York, NY, USA.

Marris,E. (2005) Chemistry society goes head to head with NIH in fight over public database. *Nature*, **435** (7043), 718–719.

Micheli,A., Sperduti,A., Starita,A. and Biancucci,A.M. (2003) A novel approach to QSPR/QSAR based on neural networks for structures. In *Soft Computing Approaches in Chemistry*, (Cartwright, H. and Sztandera,L.M., eds),. Springer Verlag Heidelberg, Germany pp. 265–296.

Micheli,A., Sperduti,A., Starita,A. and Biancucci,A.M. (2001a) Analysis of the internal representations developed by neural networks for structures applied to quantitative structure-activity relationship studies of benzodiazepines. *J. Chem. Inf. Comput. Sci.*, **41**, 202–218.

Micheli,A., Sperduti,A., Starita,A. and Bianucci,A.M. (2001b) Analysis of the internal representations developed by neural networks for structures applied to quantitative structure-activity relationship studies of benzodiazepines. *J. Chem. Inf. Comput. Sci.*, **41** (1).

Mihalic,Z., Nikolic,S. and Trinajstic,N. (1992) Comparative study of molecular descriptors derived from the distance matrix. *J. Chem. Inf. Comput. Sci.*, **32** (1), 28–37.

Muggleton,S. (1992), vol. 38 of APIC Series,. Academic Press, London. Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

K.-R. Müller, G. Rätsch S. Mika, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Neural Networks*, **12**(2):181–201, 2001.

Newman,D.J., Cragg,G.M. and Snader,K.M. (2002) Natural products as a source of new drugs over the period 1981-2002. *Journal of Natural Products*, **66**, 1002–1037. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA., 1988.

PHYSPROP (1994). Physical/Chemical Property Database (PHYSPROP). **10**

L. Ralaivola, J. S. Swamidass, H. Saigo, and P. Baldi. Graph Kernels for Chemical Informatics. *Neural Networks*, special issue on Neural Networks and Kernel Methods for Structured Domains, **18**, 8, 1093-1110, (2005).

J. W. Raymond and P. Willett. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of Computer-Aided Molecular Design*, **16**:59–71, 2001.

Richards, F. M. (1977). "Areas, volumes, packing and protein structure." *Annual Review Biophysics and Bioengineering* **6**: 151-176.

Rouvray, D. (1992). *Journal of Chemical Information and Computer Sciences*, **32** (6), 580–586.

Russo, E. (2002) Chemistry Plans a Structural Overhaul. *Nature*, **419**, 4-7.

Sadowski, J., Gasteiger, J. and Klebe, G. (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *Journal of Chemical Information and Computer Sciences*, **34**, 1000–1008.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjölander, K., Underwood, R.C. and Haussler, D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucl. Acids Res.*, **22**, 5112–5120.

Salton, G. (1991) Developments in automatic text retrieval. *Science*, **253**, 974–980.

Schölkopf, B. and Smola, A.J. (2002) *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. MIT University Press.

Schreiber, S.L. (2003) The small-molecule approach to biology: chemical genetics and diversity-oriented organic synthesis make possible the systematic exploration of biology. *Chemical and Engineering News*, **81**, 51–61.

Schreiber, S.L. (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science*, **287**, 1964–1969.

Shoichet, B.K. (2004) Virtual screening of chemical libraries. *Nature*, **432**, 862–865.

A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, **8**(3):714–735, 1997.

M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *Journal of Medicinal Chemistry*, **44**(7):1035–1042, 2001.

Stockwell, B.R. (2004) Exploring biology with small organic molecules. *Nature*, **432**, 846–854.

R. L. Strauseberg and S. L. Schreiber. From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science*, **300**(5617):294–295, 2003.

Swamidass, S. J. and P. F. Baldi (2006). A One to Two Order Speedup of Chemical Fingerprint Searches. *unpublished*, University of California, Irvine.

Swamidass, S.J., Chen, J., Bruand, J., Phung, P., Ralaivola, L. and Baldi, P. (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity, and anti-cancer activity. *Bioinformatics*, **21** (Supplement 1), i359–368.

Todd, M. H. (2004). "Computer-Aided Organic Synthesis." Chemical Society Reviews(34): 247-266.

Townsend,J., Adams,S.A., Waudby,C.A., de Souza,V.K., Goodman, J.M. and Murray-Rust,P. (2004) Chemical documents: machine understanding and automated information extraction. *Organic and Biomolecular Chemistry*, **22**, 3294–3300.

Tversky,A. (1977) Features of similarity. *Psychological Review*, **84** (4), 327–352.

Ukkonen,E. (1995) On–line construction of suffix trees. *Algorithmica*,**14** (3), 249–260.

C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 1978.

V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, inc., 1998.

Veber,D., Johnson,S.R., Cheng,H., Smith,B.R., Ward,K.W. and Kopple,K.D. (2002) Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, **45**, 2615–2623.

Vert,J.P. (2002) A tree kernel to analyze phylogenetic profiles. *Bioinformatics*, **18**, 276–284.

Viswanadhan,V.N., Ghose,A.K., Revankar,G.R. and Robins,R.K. (1989) Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *Journal of Chemical Information and Computer Sciences*, **29** (3), 163–172.

Vishwanathan,S.V.N. and Smola,A.J. (2003) Fast Kernels for Strings and Tree Matching. In *Adv. in Neural Information Processing Systems* vol. 15,

Voigt,J.H., Bienfait,B., Wang,S. and Nicklaus,M.C. (2001) Comparison of the nci open database with seven large chemical structural databases. *Journal of Chemical Information and Computer Sciences*, **41** (3), 702–712.

Wang,R., Fu,Y. and Lai,L.A. (1997) New atom-additive method for calculating partition coefficient. *Journal of Chemical Information and Computer Sciences*, **37**, 615–621.

Weiner,P. (1973) Linear Pattern Matching Algorithms. In *Proc. of the 14th IEEE Ann. Symp. on Switching and Automata Theory* pp. 1–11.

D. Weininger, A. Weininger., JL Weininger (1989). "SMILES 2: Algorithm for generation of unique SMILES notation." Journal of Chemical Information and Computer Sciences **29**: 97-101.

L. Xue, F. L. Stahura, and J. Bajorath. Similarity search profiling reveals effects of fingerprint scaling in virtual screening. *Journal of Chemical Information and Computer Science*, **44**:2032–2039, 2004.

Yalkowsky, S.H. and Dannelfelser, R.M. (1990). The ARIZONA dATABASE of Aqueous Solubility.

Yang, Y. and Pedersen, J.O. (1997) A comparative study on feature selection in text categorization. In *Proc. of the 14th Int. Conf. on Machine Learning (Nashville, TN)* pp. 412–420.

8 WEB RESOURCES

The following list provides a few pointers and is not meant to be comprehensive in any way.

DATABASES, DATASETS, SEARCH:

Cambridge Crystallographic Data Centre

<http://www.ccdc.cam.ac.uk/>

DrugBank

http://redpoll.pharmacy.ualberta.ca/drugbank/cat_browse.htm

eMolecules (formerly Chmoogle)

<http://www.emolecules.com/>

ChemBank

<http://chembank.broad.harvard.edu/>

ChemDB and other datasets

<http://cdb.ics.uci.edu>

IUPAC InChi Website

<http://www.iupac.org/inchi>

Ligand Info (PDB chemical info)

<http://ligand.info/>

MSD Ligand Chemistry (PDB chemical Info)

<http://www.ebi.ac.uk/msd-srv/chempdb/cgi-bin/cgi.pl>

NCI Data

<http://dtp.nci.nih.gov/webdata.html>

PubChem

<http://pubchem.ncbi.nlm.nih.gov/>

Standard Datasets <http://www.cheminformatics.org/datasets/index.shtml>

<http://www.cheminformatics.org/>

ZINC

<http://blaster.docking.org/zinc/>

TOOLKITS:

Chemistry Development Kit

<http://www.chemistry-development-kit.org/>

Frowns

<http://frowns.sourceforge.net/>

Jmol

jmol.sourceforge.net/

OEChem

<http://www.eyesopen.com/products/toolkits/oechem.html>

OpenBabel

<http://openbabel.sourceforge.net/>

VISUALIZATION:

Chimera

<http://www.cgl.ucsf.edu/chimera/>

VIDA

<http://www.eyesopen.com/products/applications/vida.html>

MISCELLANEOUS:

<http://www.chemaxon.com/>

<http://www2.chemie.uni-erlangen.de/index.html>

<http://www.daylight.com/>

<http://www.daylight.com/dayhtml/doc/theory/index.html>

<http://www.eyesopen.com/>

<http://www.tripos.com/>

