

Computational approaches to Molecules, Reactions and Catalysis

1.12.11

Background. In describing the behavior of a set of molecules, know which reactions can take place and which molecules within the set would accelerate the reactions are clearly central. Typically, in systems biochemistry, metabolic control analysis and kinetic analysis, these are assumed known. However, in reality they are not known and answering them is highly relevant as having a set of molecules and wanting to predict the changes in concentrations is central to both systems chemistry, origin of life research and beyond – the equations of systems biology are often taken from “established biological knowledge” that is bound to contain many errors.

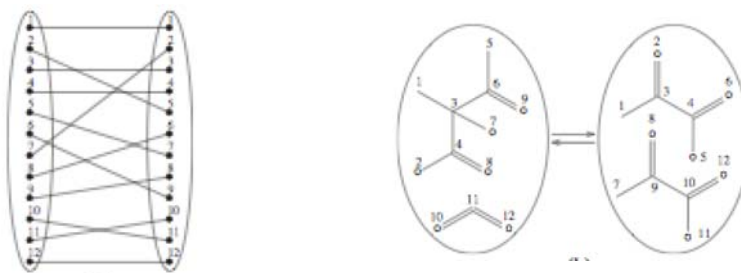
The motivation for this problems comes from “systems chemistry” (Hunt and Otto, 2011) and formal models of origin of life [Steel, 2000], where it would be a major advantage if observed molecules, reactions and catalysis could be computed as these in real systems will consist of multiply interacting molecules that is not pure or controlled by an experimenter.

The three problems – molecule enumeration, reaction prediction and catalysis prediction – are very different. The first has a more than hundred year history, the second decades, while that last is only started to be studied.

Much of modern combinatorics was initiated by attempts to *count molecules*. Cayley counted saturated hydrocarbons [alkanes] in mid-19th century, which happens also to be equivalent to counting the shapes of phylogenies in biology. Polya extended the ability to count molecules considerably by the use of group theory to be able to take account of symmetries. In the last decades molecule enumeration has been used to screen virtual drug targets computationally (Blum, van Deursen and Reymond, 2011). Molecule enumeration views a molecule as a node labeled graph. Viewing a molecule as a 3D entity will introduce chirality of certain atoms, when a molecule is different from its mirror image. Additionally, a large set of molecule graphs are hard to realize as physical structures due to steric collision, making them energetically unfavourable. Since, a molecule with a sub-molecule that energetically unfavourable, is itself energetically unfavourable, this lends itself to a branch-and-bound traversal of the set of molecules.

Many problems remains open in this line of research and numbers become very large as a function of number of atoms, so imposing restrictions on the set of relevant molecules is important to become a realistic tool. Andronico et al. (2011) discussed a variety of methods to predict the 3D structure from the molecular graph, where using databases of known small molecule structures to extract constraints on fragments of the molecules seemed to be the most most promising approach.

Most *reactions* are either fission/fusion of molecules or swapping of groups. This is so for the simple reason that more that two molecules colliding in space is unlikely and we will restrict ourselves to these case. Higher order reactions with more molecules reacting simultaneously can be explained as a string of reactions only involving pairs of molecules. To investigate if $A+B \rightarrow C$ or $A+B \rightarrow C+D$ is possible suggests a natural approach: Are **A** and **B** sub-graphs of **C** [or **C** and **D**] such that they don't overlap and that they cover **C** [or **C** and **D**]?



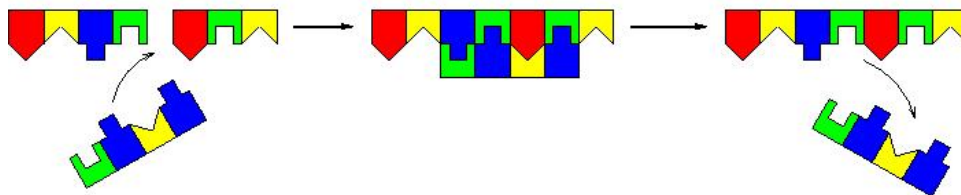
To the left which atoms in A, B could map to which atoms in C,D. The two covering disjoint sub-sets from A,B is found that maps into two covering disjoint sub-sets in C,D that preserves sub-graphs of the sub-sets (From Rover and Viari)

Kayala et al. (2011) discuss machine learning approaches to predict reactions incorporating various levels of chemical realism in the form of molecular orbital information.

Catalysis is harder to predict as it doesn't lend itself as easily to be reduced to a simple computational problem such as sub-graph isomorphism with constraints. However, there are special cases where catalysis is well established and that is in the case of ligation and hydrolysis of RNA/DNA. $A+B \rightarrow C$

[concatenation of **A** and **B**] is catalysis by a molecule **M** if **M** is complementary to the end of **A** and beginning of **B**.

Abstracting a little in the hope of applying it generally, the double complementarity of **M** to **A** and **B** can be viewed as double docking. Docking [Brooijmans and Kuntz, 2003] is when two molecules “lock” into each other. Docking is hard to predict as it involves real molecules that should be represented in full and described by quantum mechanics. This has led to a series of simplified approaches. These approaches have not been reliable in being able to predict docking with high confidence, but it still represent progress relative to uniform random guessing. Docking comes in many flavours dependent on the problem at hand and the computational resources available. Complementary is always a goal, but the shapes can be refined with charge and flexibility. One approach to docking would be that **M** should dock to **A** and to **B**, so that **A** and **B** was close, but not physically overlapping.



Two polymers of length 4 and 3, respectively are brought together by a polymer of length 4 that is complementary to the end of the first polymer and the beginning of the next polymer.

Possible Contents of report/presentation:

- Introduction/Motivation
- History of Molecule Combinatorics, Reaction and Catalysis Prediction
- Molecule Combinatorics
- Reaction Prediction
- Catalysis Prediction
- Key Technical Challenges
- Possible computational research on the topic

Comment: Since molecule counting and reaction prediction have an established literature, covering these can be done by studying the key papers. Since catalysis prediction is new this part should probably be done by suggesting an algorithm for doing this. Since there are three students and three sub-projects, this suggests a natural division of labour.

References:

- Arita (2000) Metabolic reconstruction using shortest path Simulation *Practice and Theory* 8.1-2,109-12
- Arita (2000) Graph Modeling of Metabolism JS Art Intel. 15.4??
- Gil Benko □,† Christoph Flamm,* and Peter F. Stadler (2003) “A Graph-Based Toy Model of Chemistry” *J. Chem. Inf. Comput. Sci.* 43, 1085-1093
- Boyer (2003) “Ab Initio reconstruction of metabolic pathways” *Bioinformatics* 19.2.26-34
- Brooijmans N and I Kuntz (2003) “Molecular Recognition and Docking Algorithms” *Annu.Rev.Biophys.Biomol.Struct.* 32.335-73.
- JChen, Baldi. (2009) “No Electron Left-Behind: a Rule-Based Expert System to Predict Chemical Reactions and Reaction Mechanisms”. *J. of Chem Inf Mod.* 49, 9, 2034-2043
- Crabtree and Metha (2009) “Automated Reaction Mapping” *ACM J. Experimental Algorithms* 13.1.1-29
- Diettrich, Ziegler and Banzhaf (2001) “Artificial Chemistries—A Review” *Artificial Life* 7: 225–275
- Halperin et al. (2002) “Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions” *Proteins* 47:409–443
- Hunt and Otto (2011) “Dynamic combinatorial libraries: new opportunities in systems chemistry” *Chem. Commun.*, 2011, 47, 847–858
- Kayala et al. (2011) “Learning to Predict Chemical Reactions” *Journal of Chemical Information and Modeling*. 2011
- Kerber, Laue, Meringer,§ and Rucker (2007) “Molecules in Silico: A Graph Description of Chemical Reactions” *J. Chem. Inf. Model.* 47, 805-817
- Kuntz, ID, Blaney, JM; Oatley, SJ; Langridge, R; Ferrin, TE (1982). “A geometric approach to macromolecule-ligand interactions”. *Journal of molecular biology* 161 (2): 269–88
- Rosello and Valiente (2005) “Chemical Graphs, Chemical Reaction Graphs, and Chemical Graph Transformation” *Electronic Notes in Theoretical Computer Science* 127. 157–166
- Steel (2000) The emergence of a self-catalysing structure in abstract origin-of-life models. *Applied Mathematics Letters* 3, pp.91-95.
- Lorenz C. Blum • Ruud van Deursen • Jean-Louis Reymond (2011) “Visualisation and subsets of the chemical universe database GDB-13 for virtual screening” *J Comput Aided Mol Des* (2011) 25:637–647
- Fink, Bruggesser, and Reymond *Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons* *Angew. Chem. Int. Ed.* 2005, 44, 1504–1508
- Tobias Fink and Jean-Louis Reymond (2007) *Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery* *J. Chem. Inf. Model.*, 47, 342-353
- Van Deursen and JL Reymond (2007) *Chemical Space Travel* *Cem Med Chem* 2.636-40
- Reymond, van Deursen, Blum and Ruidtjgkeit (2010) *Chemical space as a source for new drugs* *Med. Chem. Commun.*, 2010, 1, 30–38
- Trinajstić (1992) *Chemical Graph Theory* CRC Press

Acknowledgements: JH has received advice on literature from Tack Kuntz, Rene Thomsen, Martin Simonsen, Charlie Carter, Sijbren Otto, Wim Hordijk, Pierre Baldi and probably more. JH is surprised that not more work has been done on catalysis prediction as there clearly are a series of unexplored approaches.