David Danko

Small Molecule Search: Independent Research with the Berger group at CSAIL

Introduction:

Small organic molecules are an important part of modern society. Every modern drug and many modern materials are based off of small organic molecules. Until recently every new drug has included the production of one ore more novel molecules which require a novel synthesis process and FDA approval.

Recently researchers have started to use the large existing database of molecules to search for compounds similar to drugs they want to synthesize. This allows for shorter drug development and approval times. Unfortunately current search methods for molecules are too slow to be practical.

This semester I plan to continue a project I began last semester with a post-doctoral researcher, Noah Daniels, and the Berger group at CSAIL to develop a faster multifold search for similarities between small molecules, using graph theory and some of the intrinsic properties of molecules.

Technique:

All publicly available techniques for searching databases of molecules to date work off of a maximum common subgraph technique. A query molecule has its maximum common subgraph (in this case a labeled and weighted graph) computed for each of a large set of target molecules. The molecules with the largest (strictly speaking a coefficient relating the size of the overlapping molecule to the sizes of the original molecules is used) of the common subgraphs with the original query are then selected as good matches. Finding a globally maximum common subgraph between two graphs is provably NP-Hard. MCS techniques differ in what method they use to find a locally maximum common subgraph, typically competing for faster algorithms though size of the locally optimal graph is also relevant.

Rather than develop a new MCS technique we have repurposed an existing MCS technique to find matching molecules by a two fold process. We plan to publish a paper detailing the exact method by which our technique works; unfortunately I am not able to go over many of the specifics of our technique until we have sent our paper for review. I will, however, provide an outline of our technique.

As a preprocessing step we find clusters within a database of molecules. Each of these clusters has a representative molecule that matches certain characteristic traits of the molecules in the cluster. The cluster representatives are then saved as a database with references to the molecules within the cluster. There is no requirement that the representative be a member of its own cluster.

Our preprocessing step allows for two distinct types of search, both of which we hope will be faster than existing techniques. The first technique is a straightforward conversion of existing MCS techniques. We find the cluster representatives that have the largest common subgraph with our query then we search for our queries largest common subgraph among the set of molecules represented by the cluster representatives we chose initially. The second search technique is to see if our query belongs to a cluster. If it does then we will search for our queries largest common subgraph among other members of that cluster.

The work I completed last semester allowed for a single level of clustering which improved upon existing techniques for molecule search by a reasonable factor. This semester I am building upon my previous work to develop a system which creates many layers of clustering: clusters, clusters of clusters, clusters of clusters of clusters, etc.

Multiple layers of clustering will allow us to build a tree which can be searched layer by layer. If a match between a query molecule and a cluster is found we will compare the query to the members of the clusters (which are themselves clusters) repeating the process until we have found actual molecules. Each cluster we reject likely contains a large number of molecules that we would otherwise have compared our query molecule to directly.

If we successfully develop a good clustering schema (we think we will) most of these molecules will be different than our query molecule; had we compared our query to them directly they would have been rejected. Almost inevitably some molecules that would have matched our query molecule will be rejected. The search technique we develop will have to strike a balance between reducing the number of comparisons we have to make by rejecting clusters and missing too many molecules that our query would otherwise have matched.

Once we have finished developing our algorithm we plan to release our implementation as a tool to the scientific community. A reasonable portion of my work will be polishing our research code to a high level. Similarly I will help to write the paper that will accompany the release of this tool.