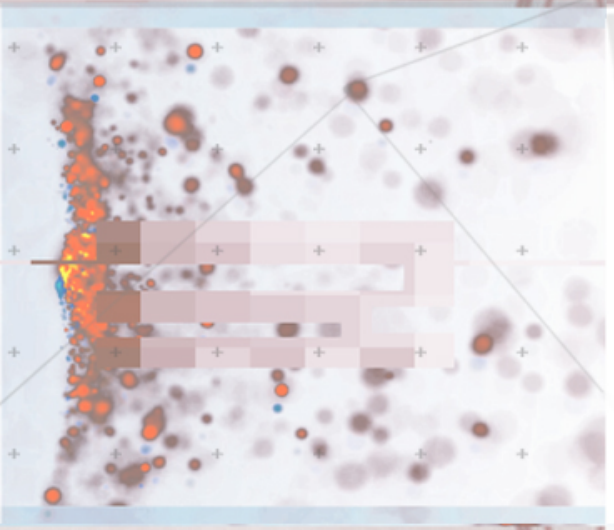
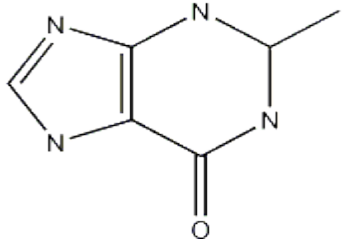
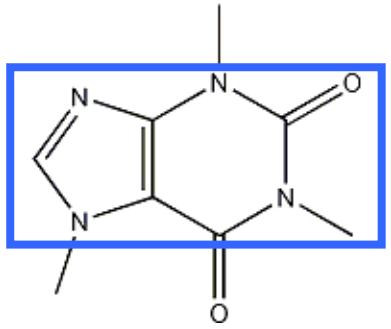


Graph Pattern Mining Application II: Graph Similarity Search

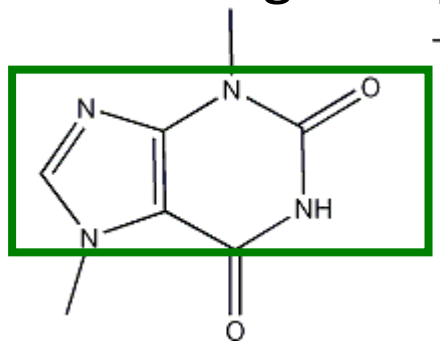
Application II: Support Substructure Similarity Search

- Find graphs in a graph DB containing substructures similar to a given query graph
- Ex. Data: A chemical compound DB
  - A query graph q:

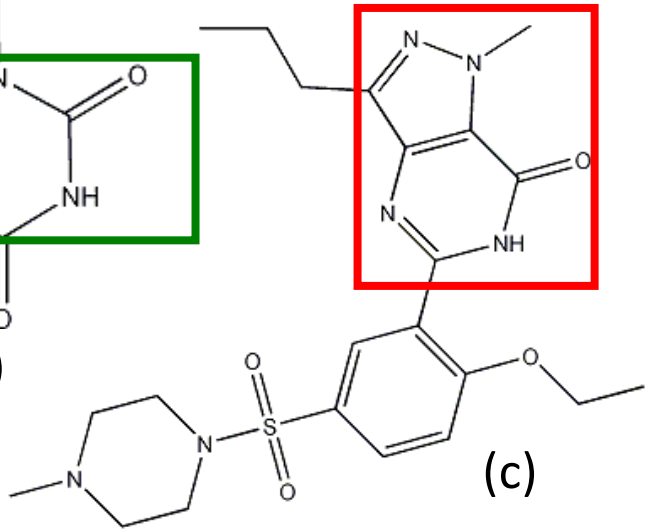




(a)



(b)

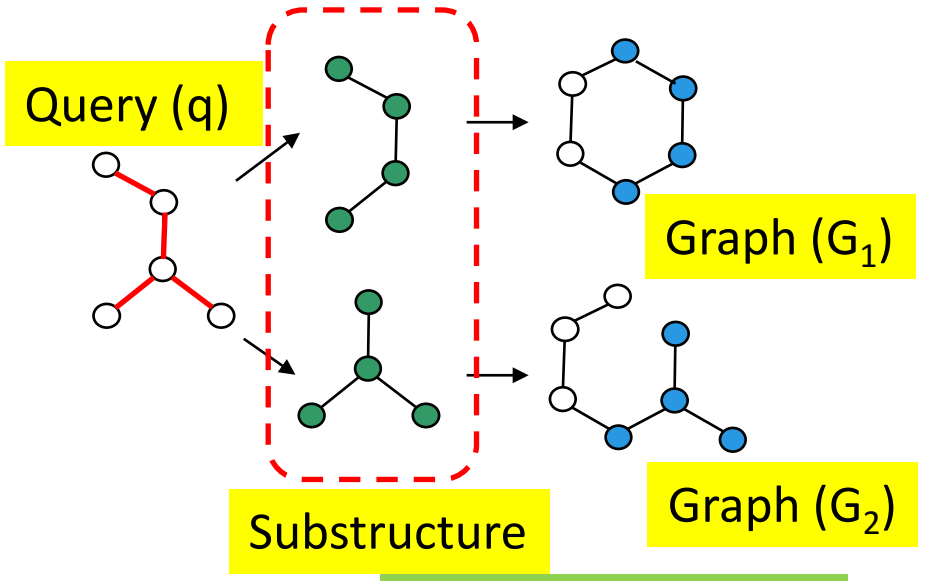


(c)

- How to do similarity search efficiently?
  - No indexing? – Sequential scan + computing subgraph similarity – too costly!
  - Build graph indices to support approximate search?
    - Need an explosive number of subgraphs to cover all the *similar* subgraphs!
- An elegant solution (Yan, Yu, & Han, SIGMOD'05):
  - Keep the graph index structure, but select features in the query space

Feature-Based Similarity Search

- Decompose a query graph into a set of features
- Feature-based similarity measure
  - Each graph is represented as a feature vector $X = \{x_1, x_2, \ldots, x_n\}$
  - Similarity is defined by the distance of their corresponding vectors
- If graph G contains the major part of a query graph q, G should share a number of common features with q
  - Given a relaxation ratio, one can calculate the maximal number of features that can be missed!



Assume: Query graph has 5 features

Relaxation threshold: Can miss at most 2 features

Then: G_1, G_2, G_3 are pruned

Graphs in database

features	G_1	G_2	G_3	G_4	G_5
f_1	0	1	0	1	1
f_2	0	1	0	0	1
f_3	1	0	1	1	1
f_4	1	0	0	0	1
f_5	0	0	1	1	0

A feature-graph matrix