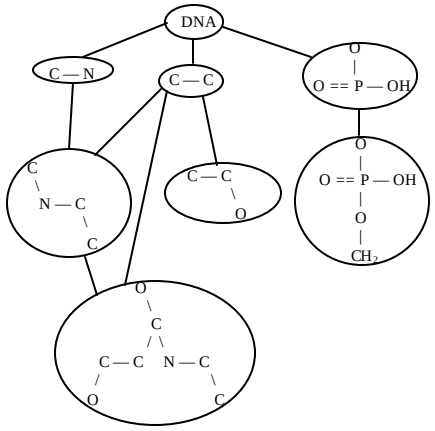
the ability to work with structured data–we present a task that involves describing a DNA sequence by clustering. To represent the DNA as a graph, atoms and small molecules are mapped to vertices, and bonds are represented by undirected edges. The edges are labeled according to the type of bond, single or double. A portion of the lattice generated is shown in the figure.

The lattice closely resembles a tree, with the exception that two nodes (bottom-left) have two parents. The lattice describes 71% of the DNA sequence. As the figure shows, smaller, more commonly occurring compounds are found first that compose the first level of the lattice. These account for more than 61% of the DNA. Subsequently identified clusters are based on these smaller clusters that are either combined with each other, or with other atoms or molecules to form a new cluster. The second level of the lattice extends the conceptual clustering description such that an additional 7% of the DNA is covered. Future work on Subdue will continue discovery of hierarchical clusterings in real-world domains, and both objective and expert-based comparisons to other clustering systems.



References

Fisher, D. H. Knowledge Acquisition Via Incremental Conceptual Clustering, *Machine Learning*. Kluwer, The Netherlands, 1987.

Holder, L. B. and D. J. Cook. Discovery of Inexact Concepts from Structural Data. In *IEEE Transactions on Knowledge and Data Engineering*, Volume 5, Number 6, 1993, pages 992-994.

Thompson, K. and P. Langley. Concept formation in structured domains. In Fisher, D.H., & Pazzani, M. (Eds.), *Concept Formation: Knowledge and Experience in Unsupervised Learning*, chap. 5. Morgan Kaufmann Publishers, Inc., 1991.