Pattern Matching and Pattern Discovery in Scientific, Program, and Document Databases*

Jason T. L. Wang† Kaizhong Zhang‡ Dennis Shasha§

1 Overview

Over the past several years we have created or borrowed algorithms for combinatorial pattern matching and pattern discovery on sequences [2] and trees.

In *matching* problems, given a pattern, a set of data objects and a distance metric, we find the distance between the pattern and one or more data objects. In *discovery* problems by contrast, given a set of objects, a metric, and a distance, we seek a pattern that matches many of those objects within the given distance. (So, discovery is a lot like data mining.) Our toolkit performs both matching and discovery with current targeted applications in molecular biology and document comparison.

2 Prototype

Our demonstration shows:

- How to find approximately common regular expression motifs in a set of protein sequences (obtained from the Cold Spring Harbor Laboratory) and DNA sequences (obtained from the Whitehead Institute of MIT).
- How to classify a given protein (we invite conference participants to bring in a sequence) into a family in the SWISS-PROT database. The output displays not only the family, but relevant literature published to describe this family.
- How to find approximately common substructures in a set of RNA secondary structures (trees) obtained from the National Cancer Institute pertaining to the sabin strain, human rhinovirus and coxsackievirus.
- How to align two LaTex and SGML documents [1] according to their hierarchical structures and similarly for a set of programs.

The main applications (and users) of our toolkit so far have been in biology and medicine, but the potential uses extend to any application where string, tree, or (eventually) graph comparison and discovery is an important operation.

*This work was supported, in part, by NSF under Grants IRI-8901699, CCR-9103953, IRI-9224601 and IRI-9224602, by ONR under Grants N00014-90-J-1110, N00014-91-J-1472 and N00014-92-J-1719, by the Natural Sciences and Engineering Research Council of Canada under Grant OGP0046373, by NJIT under Grant SBR-421280, and by a grant from the AT&T Foundation.

†Department of Computer and Information Science, New Jersey Institute of Technology, University Heights, Newark, New Jersey 07102 (jason@vienna.njit.edu).

‡Department of Computer Science, The University of Western Ontario, London, Ontario, Canada N6A 5B7 (kzhang@csd.uwo.ca).

§Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012 (shasha@cs.nyu.edu).

Acknowledgements

The project is a group research effort. We thank the following individuals for their help, advice and contributions to the various aspects of the project: Nat Goodman (Whitehead Institute of MIT), Jim Kaminski (Schering-Plough Research), Tom Marr (Cold Spring Harbor Lab), Steve Rozen (Whitehead Institute of MIT), Bruce Shapiro (National Cancer Institute), and our students Chia-Yo Chang, Gung-Wei Chirn, Karpjoo Jeong and Karen Pysniak.

References

[1] V. Christophides, S. Abiteboul, S. Cluet, and M. Scholl. From structured documents to novel query facilities. In *SIGMOD*, pages 313–324, Minnesota, May 1994.

[2] J. T. L. Wang, G.-W. Chirn, T. G. Marr, B. A. Shapiro, D. Shasha, and K. Zhang. Combinatorial pattern discovery for scientific data: Some preliminary results. In *SIGMOD*, pages 115–125, Minnesota, May 1994.