

Pattern Discovery in Biosequences

Tutorial for ISMB 2002

Stefano Lonardi

Department of Computer Science
University of California
Riverside, CA 92521

Tutor

Stefano Lonardi is Assistant Professor at University of California, Riverside, CA. He belongs to the Computational Biology group, and he is also a faculty of the Genetics program. During the Fall of the year 2001 he started a new seminar graduate course on “Pattern Discovery in Biosequences”. The proposed tutorial draws from the experience gained during that course.

Stefano received his *Ph.D.* in the summer of 2001 from the Department of Computer Sciences, Purdue University, West Lafayette, IN. His thesis, supervised by Prof. A. Apostolico, is entitled “Global Detectors of Unusual Words: Design, Implementation, and Applications to Pattern Discovery in Biosequences”.

He also received the *Dottorato di Ricerca* degree in Computer and Electrical Engineering from the Department of Electrical and Computer Engineering, University of Padua, Italy. During the summer of 1999, he was intern at Celera Genomics, Department of Informatics Research, Rockville, MD, working under the supervision of E.W. Myers. He published papers in several journals, like *Science*, *Journal of Computational Biology*, *Proceedings of the IEEE*, among others.

Length

Half day

Abstract

Characterizing and finding regularities in strings are fundamental problems in many areas of Science: molecular biology, analysis of stochastic processes, data compression, machine learning, speech recognition, coding, automata, formal language theory, etc. Among the commonly studied regularities in texts, the most prominent role is played by frequent occurrences of the same word.

Patterns that occur unexpectedly often or rarely in genetic sequences have been variously linked to biological meanings and functions. The underlying probabilistic and statistical models have been studied extensively (see, e.g., [Reinert, Schbath, Waterman, JCB 2000] for a review). With increasing availability of whole genomes, exhaustive statistical tables

and global detectors of unusual patterns on a scale of millions, even billions of bases become conceivable. It is natural to ask how large such tables may grow with increasing length of the input sequence, and how fast they can be computed. These problems need to be regarded not only from the conventional perspective of asymptotic space and time complexities, but also in terms of the volumes of data produced and ultimately, of practical accessibility and usefulness.

The tutorial is intended to describe recent algorithms that are aimed to attack this problem. For clarity of exposition, we first classify the methods based upon the type of patterns they are designed to find: deterministic, rigid or profiles. Deterministic patterns are simply words over the alphabet while rigid patterns allow substitutions but their length cannot change. Matrix profiles are matrices in which each position is associated with a probability distribution over the symbols of the alphabet.

In the class of deterministic patterns, I plan to cover VERBUMUCULUS [Apostolico, Bock, Lonardi, Xu, JCB 2000][Apostolico, Bock, Lonardi, RECOMB 2002]. Among the methods for rigid patterns, I plan to describe TEIRESIAS [Rigoustos, Floratos, Bioinformatics 98], WINNOWER [Pevzner, Sze, ISMB 00], PROJECTION [Buhler, Tompa, RECOMB 01], and WEEDER [Pavesi, Mauri, Pesole, ISMB 01]. I close the tutorial by covering two statistical methods to discover matrix profiles, MEME [Bailey, Elkan, Machine Learning, 95] and GIBBS sampler [Lawrence, Altschul, Boguski, Liu, Neuwald, Wootton, Science 93].

This selection may seem somewhat arbitrary and leave out some important algorithms. For example: SPEXS and the related generalization attempts described in [Brazma et al., JCB 98][Brazma et al., Genome Research 98], PRATT by Jonassen [cabios 97], SPLASH [Califano, Bioinformatics 2000], YEBIS [Yada et. al, Bioinformatics 98], CONSENSUS [Hertz, Stormo, Bioinformatics 99].

This particular choice is an attempt to cover the most *recent* techniques to attack the problem of pattern discovery. I am describing two more established techniques (MEME and GIBBS) in order to cover matrix profiles.

Intended Audience

Computer scientists, biologists, statisticians with interests in the analysis of biological sequences (DNA or proteins). The tutorial aims to be introductory and requires only basic knowledge on probability, statistics and molecular biology.

Outline

- Why “pattern discovery”? Gene regulation and Promoters
- Basic concepts
- Problem definition
- Classification of patterns
- Discovering Deterministic patterns

– VERBUMUCULUS

- Complexity results
- Discovering Rigid patterns
 - TEIRESIAS
 - WINNOWER
 - PROJECTION
 - WEEDER
- Discovering Profiles
 - GIBBS sampling
 - MEME

References

- [1] APOSTOLICO, A., BOCK, M. E., LONARDI, S., AND XU, X. Efficient detection of unusual words. *J. Comput. Bio.* 7, 1/2 (Jan. 2000), 71–94.
- [2] BAILEY, T. L., AND ELKAN, C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* 21, 1/2 (1995), 51–80.
- [3] LAWRENCE, C. E., ALTSCHUL, S. F., BOGUSKI, M. S., LIU, J. S., NEUWALD, A. F., AND WOOTTON, J. C. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262 (Oct. 1993), 208–214.
- [4] PAVESI, G., MAURI, G., AND PESOLE, G. An algorithm for finding signals of unknown length in dna sequences. In *Proc. of the International Conference on Intelligent Systems for Molecular Biology* (2001), AAAI press, Menlo Park, CA, pp. S207–S214.
- [5] PEVZNER, P. A., AND SZE, S.-H. Combinatorial approaches to finding subtle signals in DNA sequences. In *Proc. of the International Conference on Intelligent Systems for Molecular Biology* (2000), AAAI press, Menlo Park, CA, pp. 269–278.
- [6] RIGOUTSOS, I., AND FLORATOS, A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* 14, 1 (1998), 55–67.
- [7] TOMPA, M., AND BUHLER, J. Finding motifs using random projections. In *Annual International Conference on Computational Molecular Biology* (Montreal, Canada, Apr. 2001), pp. 67–74.