Università degli Studi di Verona

Dipartimento di Informatica

PhD course
"Computational methods for handling textual data"

- **Part I:** Introduction to information theoretic measures and compression - 4x2 hours *(Ferdinando Cicalese)*
- **Part II:** Data structures for strings: suffix trees, suffix arrays, Burrows-Wheeler-transform - 4x2 hours *(Zsuzsanna Lipták)*
- **Part III:** An application: InfoGenomics - 2x2 hours *(Giuditta Franco)*

Abstract:
In this course, we introduce basic computational methods for handling textual data. Textual data (strings, sequences) is ubiquitous in today's world: WWW (webpages) and biological data (genomic and other biological sequences) are only two examples of large amounts of textual data which need to be handled on a daily basis. This type of data is being produced at an ever increasing rate, and one of the major computational challenges now is to develop data structures which allow both storing the data efficiently, and at the same time extracting information from it (e.g. search, pattern matching).

In this course, directed at a general audience of computer science PhD students, we give an introduction to this topic. The course starts with an introduction to basic information theoretic measures, compression, dictionary based compression. In the second part, we introduce three data structures for strings which have been milestones in the area of string storage. These are suffix trees, suffix arrays, and the Burrows-Wheeler transform (BWT). All of these data structures have been or are being used in mainstream software for biological and other data. In this part, we will get an insight into of some of the major challenges in this research area. Finally, in the third part of the course, we give an introduction to an application of these concepts to genomic sequences.

Course times:
The course will take place from Mo 29 Feb - Wed 23 March 2016,
each week Mo+Wed+Fri (2 h per day), exact times to be agreed with students

Detailed contents of course: see overleaf

Part I: Introduction to information theoretic measures and compression *(Ferdinando Cicalese)*

1. Information Theoretic Measures
   - How to measure information
   - Entropy, Conditional Entropy, Mutual Information, Informational Divergence
   - Jensen Inequality; additivity properties
2. Compression:
   Shannon's lower bound; Shannon-Fano coding; Huffman coding;
   Lempel-Ziv; Gzip; Graph compression (tentative)
3. Computing with streams (tentative)

Course material:

- Cover T.M., Thomas J.A.: *Elements of Information Theory,* 2nd ed., Wiley (2006)
- V. Mäkinen, D. Belazzougui, F. Cunial, A.I. Tomescu: *Genome-Scale Algorithm Design*. Cambridge University Press, 2015.
- Course Lecture Notes
- Reading Assignments

Part II: Data structures for strings: suffix trees, suffix arrays, Burrows-Wheeler-transform *(Zsuzsanna Lipták)*

1. The suffix tree: definition, simple properties, WOTD construction algorithm
2. Some applications of suffix trees: text statistics, exact string matching, maximal repeats, shortest unique substring, maximal unique matches, palindromes
3. Suffix arrays: definition, properties, exact matching, backward search, construction from ST, construction via Prefix Doubling
4. Burrows-Wheeler-Transform: transformation, retransformation, exact matching, compression

Course material:

- course slides, exercises.
- V. Mäkinen, D. Belazzougui, F. Cunial, A.I. Tomescu: *Genome-Scale Algorithm Design*. Cambridge University Press, 2015.
- D. Gusfield: *Algorithms on Strings, Trees, and Sequences*. Cambridge U.P., 1997.
- M. Crochemore, C. Hancart and T. Lecroq. *Algorithms on Strings.* Cambridge University Press, 2007.
- B. Smyth: *Computing Patterns in Strings.* Addison Wesley, 2003.
- E. Ohlebusch: *Bioinformatics Algorithms.* Oldenbusch Verlag, 2013.

Part III: An application: InfoGenomics *(Giuditta Franco)*

An introduction will be given about an application of the above concepts within the context of a recent direction of research in computational genomics. Some theoretic information measures (such as empirical Shannon entropy) and some dictionary based statistical distributions have been defined and computed to analyze genomic sequences. A few intriguing results will be shown, for example in terms of genomic profiles and

features, all generated by a software developed at the University of Verona, IGtools, efficiently handling genomic data storage by specific string data structures. A few related open problems will conclude this part.

Course material:

- J. Bohlin, MW Van Passel, L. Snipen, AB Kristoffersen, D. Ussery, SP Hardy, Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. BMC Genomics 13:66, 2012.
- V. Bonnici, V. Manca, Infogenomics tools: a computational suite for informational analysis of genomes, Journal of Bioinformatics and Proteomics Review, 1(1):7-14, 2015.
- V. Bonnici, V. Manca, Recurrence distance distributions in computational genomics, American Journal of Bioinformatics and computational biology, 3(1): 5-23, 2015.
- A. Castellini, G. Franco, V. Manca: A dictionary based informational genome analysis, BMC GENOMICS, vol. 13 (1): 485-499, 2012.
- CH Chang, LC Hsiedh, TY Chen, HD Chen, L. Luo, HC Lee. Shannon information in complete genomes. J Bioinformatics Computational Biology 3:587-608, 2005.
- M Hackenberg, A. Rueda, P. Carpena, P. Bernaola-Galvan, G. Barturen, JL Oliver, Clustering of DNA words and biological function: a proof of principle. J Theoretical Biology 297:127-136, 2012.
- Course slides, lecture notes.