

FIRST SET OF TOPICS

1. **Indexing of compressed biological sequences**
Indicative Bibliography
 - [Francisco Claude](#), [Antonio Fariña](#), [Miguel A. Martínez-Prieto](#), Gonzalo Navarro: Compressed q-Gram Indexing for Highly Repetitive Biological Sequences. [BIBE 2010](#): 86-91
 - [Veli Mäkinen](#), Gonzalo Navarro, [Jouni Sirén](#), [Niko Välimäki](#): Storage and Retrieval of Individual Genomes. [RECOMB 2009](#): 121-137

2. **Biological sequence databases**
Indicative Bibliography
 - H. Berman, K. Henrick, H. Nakamura, J. Markley, The worldwide Protein Data Bank (wwPDB) ensuring a single uniform archive of PDB data, *Nucleic Acids Research*, 2007.
 - T. Kulikova, R. Akhtar, et al. EMBL Nucleotide Sequence Database in 2006, *Nucleic Acids Research*, 2007.

3. **Biological Sequence Categorization Algorithms**
Indicative Bibliography
 - G. Ratch, C.S. Ong, Introduction to Machine Learning for Predictive Sequence Analysis
 - C. Leslie, E. Eskin, J. Weston, and W.S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4), 2003.

4. **Presentation of algorithms for finding repetitions .**
Indicative Bibliography
 - Gerth Stølting Brodal, R.B. Lyngsew, C. Pedersen, J. Stoye, Finding Maximal Pairs with Bounded Gaps. In *Journal of Discrete Algorithms, Special Issue of Matching Patterns*, volume 1(1), pages 77-104, 2000.
 - G.S. Brodal, C.S. Pedersen, Finding Maximal Quasiperiodicities in Strings, In *Proc. 11th Annual Symposium on Combinatorial Pattern Matching*, volume 1848 of *Lecture Notes in Computer Science*, pages 397-411. Springer Verlag, Berlin, 2000.

5. **Presentation of algorithms for evolutionary trees**
Indicative Bibliography
 - G.S. Brodal, Rolf Fagerberg, Christian N. S. Pedersen, and Anna Ostlin, The Complexity of Constructing Evolutionary Trees Using Experiments In *Proc. 28th International Colloquium on Automata, Languages, and Programming*, volume 2076 of *Lecture Notes in Computer Science*, pages 140-151. Springer Verlag, Berlin, 2001.
 - G.S. Brodal, R. Fagerberg, C. S. Pedersen, Computing the Quartet Distance Between Evolutionary Trees in Time $O(n \log n)$, In *Algorithmica*, Special Issue on ISAAC 2001, volume 38(2), pages 377-385, 2003.

- Phylogenetic Tree Construction
http://linneus20.ethz.ch:8080/5_6.html
- Chapter 17. Dan Gusfield. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge 1997

6. **Presentation of algorithms for Dynamic String Matching** **Indicative Bibliography**

- P. Ferragina, Incremental Text Indexing: a new Data Structure, In Proc. Od 2nd ESA, LNCS 855, pages 495-507, 1994.
- P. Ferragina and R. Grossi, Optimal On-Line Search and Sublinear Time Update in String Matching, In Proc. of IEEE FOCS, pages 604-612, 1995.
- S. Alstrup, G.S. Brodal, T. Rauhe, Pattern Matching in Dynamic Texts, In Proc. of Symposium on Discrete Algorithms, pages 819-828, 2000.
- A. Amir, M. Farach, Z Galil, R. Giancarlo, K. Park, Dynamic Dictionary Matching, Journal of Computer and System Sciences, 49:208-222, 1994.
- R.M. Idury and A.A. Schafer, Dynamic Dictionary Matching with Failure Functions, TCS, 131:295-310, 1994.

7. **DNA sequence compression** **Indicative Bibliography**

- Behshad Behzadi, Fabrice Le Fessant, DNA Compression Challenge Revisited: A Dynamic Programming Approach, Combinatorial Pattern Matching: 16th Annual Symposium, CPM 2005, June 19-22, 2005.
- G. Manzini and M. Rastero. A simple and fast DNA Compressor. Software Practise and Experience, 2004

8. **XML and Bioinformatics** **Indicative Bibliography**

- F. Achard, G. Vaysseix, and Emmanuel Barillot, XML bioinformatica and data integration, Bioinformatics review, 17, 115-125.
- L. Wang, J.P. Riethoven, A. Robinson, XEMBL: distributing EMBL data in XML format, Bioinformatics, Applications Note, 18, 1147-1148

SECOND SET OF TOPICS

Analysis of Gene Expression Maps:

L. An, H. Ling, Z. Obradovic, D.J. Smith, and V. Megalooikonomou, "Learning pairwise gene functional similarity by multiplex gene expression maps", *BMC Bioinformatics*, 13(Suppl 3):S1, 2012.

L. An, H. Xie, M.H. Chin, Z. Obradovic, D.J. Smith and V. Megalooikonomou, «Analysis of multiplex gene expression maps obtained by voxelation», *BMC Bioinformatics*, Vol. 10 (Suppl 4): S10, pp. 1-15, 2009.

Protein Signatures:

C. S. Carvalho, D. Vlachakis, G. Tsiliki, V. Megalooikonomou, S. Kossida. "Protein signatures using electrostatic molecular surfaces in harmonic space", *PeerJ*, 1:e185, <http://dx.doi.org/10.7717/peerj.185>

D. Vlachakis, C. Feidakis, V. Megalooikonomou, S. Kossida, "A two-dimensional visualization tool for amino acid domain sequences", *Theoretical Biology and Medical Modelling*, vol. 10, No.14, 2013.

D. Vlachakis, G. Tsiliki, D. Tsagkrasoulis, C. S. Carvalho, V. Megalooikonomou, S. Kossida, "Speeding up the drug discovery process: structural similarity searches using molecular surfaces", *EMBNet journal, Bioinformatics in Action*, Vol 18, No. 1, 2012.

Drug Design:

D. Vlachakis, P. Fakourelis, V. Megalooikonomou, C. Makris, S. Kossida, "DrugOn: a fully integrated pharmacophore modeling and structure optimization toolkit", *PeerJ* 3:e725; DOI 10.7717/peerj.725, 2015.

D. Vlachakis, D. Tsagrasoulis, V. Megalooikonomou, S. Kossida, "Introducing Drugster: a comprehensive and fully integrated drug design, lead and structure optimization toolkit", *Bioinformatics* 29 (1): 126-128, 2013.

Data Mining in Bioinformatics:

A. Amidi, S. Amidi, D. Vlachakis, V. Megalooikonomou, N. Paragios, E. I. Zacharaki, "EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation", *PeerJ* 6:e4750; DOI 10.7717/peerj.4750, 2018.

C. Salis, E. Papakonstantinou, K. Pierouli, A. Mitsis, L. Basdeki, V. Megalooikonomou, D. Vlachakis, M. Hagidimitriou, "A genomic data mining pipeline for 15 species of the genus *Olea*", *EMBNet Journal*, DOI: <https://doi.org/10.14806/ej.24.0.922>, 2019.

THIRD SET OF TOPICS

One chapter (and relevant papers) from Srinivas Aluru, Handbook of Computational Molecular Biology, CRC Press 2005.

(<https://www.taylorfrancis.com/books/mono/10.1201/9781420036275/handbook-computational-molecular-biology-srinivas-aluru>)

✓ **Sequence alignment**

- 12.1. B. Jackson, S. Aluru, Pairwise sequence alignment
- 12.2. A. Mironov, M. Gelfand, Spliced alignment and similarity based gene recognition
- 12.3. O. Gotoh, S. Yamada, T. Yada, Multiple sequence alignment
- 12.4. D. Fenandez-Baca, B. Venkatachalaim, Parametric sequence alignment

✓ **String data structures**

- 12.5. S. Aluru, Pang-Ko, Lookup tables suffix trees and suffix arrays
- 12.6. P. Ko, S. Aluru, Suffix tree applications in Computational Biology
- 12.7. M. Abouelhoba, S. Kurtz, E. Ohlebusch, Enhanced suffix tree and applications

✓ **Genome assembly and EST clustering**

- 12.8. X. Huang, Computational methods for genome assembly
- 12.9. R. Agarwala, Assembling the human genome
- 12.10. V. Vereramachandni, Comparative methods for sequence assembly
- 12.11. S. Bhandarkar, J. Huang, J. Arnold, Information theoretic approach to genome reconstruction
- 12.12. A. Kalyanaraman, S. Aluru, Expressed sequence tags clustering and applications
- 12.13. S. Emrich, A. Kalyanaraman, S. Aluru, Algorithms for large scale clustering and assembly of biological sequence data

✓ **Genome scale computational methods**

- 12.14. M. Brudno, I Dubchak, Comparison of long genomic sequence Algorithms and applications
- 12.15. E. Ohlebusch, M. Abouelhoba, Chaining algorithms and applications in comparative genomics
- 12.16. M. Gelfand, Computational analysis of alternative splicing
- 12.17. A. Schaffer, Human genetic linkage analysis
- 12.18. D. Gusfield, S. Orzack, Haplotype inference

✓ **Phylogenetics**

- 12.19. C. Linder, T. Warnow, Phylogenetic reconstruction
- 12.20. O. Eulenstein, Consensus trees and supertrees
- 12.21. T. Warnow, Large scale phylogenetic analysis
- 12.22. D. Bader, M. Yan, High performance phylogeny reconstruction

✓ **Microarrays and gene expression analysis**

- 12.23. C. Ball, G. Sherlock, Microarray data: annotation retrieval, storage and communication
- 12.24. H. Chou, Computational methods for microarray design
- 12.25. P. Baldi, G. Hatfield, L. Fu, Clustering algorithms for gene expression analysis
- 12.26. A. Tanay, R. Sharan, R. Shamir, Biclustering algorithms: a survey
- 12.27. V. Filkov, Identifying gene regulatory networks from gene expression data
- 12.28. H. Samad, M. Rammish, Modeling and analysis of gene networks using feedback control analysis

✓ **Computational Structural Biology**

- 12.29. M. Singh, Predicting protein structure and supersecondary structure

- 12.30. W. Hart, A. Newman, Protein structure prediction with lattice models
- 12.31. G. Lin, X. Tu, X. Wan, Protein structure determination via NMR spectral data
- 12.32. C. Bajaj, Z. Yu, Geometric and signal processing of reconstructed 3D maps of molecular complexes
- 12.33. D. Xu, O. Buzlevski, X. Fend Wan, In search of remote homologs,
- 12.34. L. Kale, K. Schulten, R. Skeel, G. Matyna, M. Tuckerman, J. Phillips, S. Kumar, G. Zheng, Biomolecular modeling using parallel supercomputers
- ✓ **Bioinformatic databases and data mining**
- 12.35 Kahveci, A. Singh, Index structures for approximate matching in sequence databases
- 12.36 S. Rajaeskaran, Algorithms for motif search
- 12.37 M. Zaki, K. Sequeira, Data mining in computational biology

FOURTH SET OF TOPICS

One chapter (and relevant papers) from Data Analytics in Bioinformatics: A Machine Learning Perspective Editor(s): Rabinarayan Satpathy, Tanupriya Choudhury, Suneeta Satpathy, Sachi Nandan Mohanty, Xiaobo Zhang, 2021, Wiley (<https://onlinelibrary.wiley.com/doi/book/10.1002/9781119785620>)

- ✓ Nancy Anurag Parasa, Jaya Vinay Namgiri, Sachi Nandan Mohanty, Jatindra Kumar Dash, Introduction to Unsupervised Learning in Bioinformatics (Pages: 35-49)
- ✓ Vrs Jhalia, Tripti Swarnkar, A Critical Review on the Application of Artificial Neural Network in Bioinformatics (Pages: 51-76)

- ✓ Saurav Roy, Ratula Ray, Satya Ranjan Dash, Mrunmay Kumar Giri, Plant Disease Detection Using Machine Learning Tools With an Overview on Dimensionality Reduction (Pages: 109-144)
- ✓ S. Mahapatra, T. Swarnkar, Gene Selection Using Integrative Analysis of Multi-Level Omics Data: A Systematic Review (Pages: 145-171)
- ✓ Sudhansu Shekhar Patra, Om Praksah Jena, Gaurav Kumar, Sreyashi Pramanik, Chinmaya Misra, Kamakhya Narain Singh, Random Forest Algorithm in Imbalance Genomics Classification (Pages: 173-190)
- ✓ Shubham Raj, Swati Singh, Avinash Kumar, Sobhangi Sarkar, Chittaranjan Pradhan, Feature Selection and Random Forest Classification for Breast Cancer Disease (Pages: 191-210)
- ✓ Swati Sucharita, Barnali Sahu, Tripti Swarnkar, A Comprehensive Study on the Application of Grey Wolf Optimization for Microarray Data (Pages: 211-248)
- ✓ V. Aakash, S. Sridevi, G. Ananthi, S. Rajaram, Forecasting of Novel Corona Virus Disease (Covid-19) Using LSTM and XG Boosting Algorithms (Pages: 293-311)
- ✓ P. Poongodi, E. Udayakumar, K. Srihari, Nandan Mohanty Sachi, An Innovative Machine Learning Approach to Diagnose Cancer at an Early Stage (Pages: 313-337)

- ✓ G. Aparna, G. Anitha Mary, G. Sumana, Performance Analysis of Signal Processing Techniques in Bioinformatics for Medical Applications Using Machine Learning Concepts (Pages: 391-429)

- ✓ Yuvaraj Natarajan, Srihari Kannan, Sachi Nandan Mohanty, Survey of Various Statistical Numerical and Machine Learning Ontological Models on Infectious Disease Ontology (Pages: 431-442)

- ✓ Ritesh Choudhary, T. Gopalakrishnan, D. Ruby, A. Gayathri, Vishnu Srinivasa Murthy, Rishabh Shekhar, An Efficient Model for Predicting Liver Disease Using Machine Learning (Pages: 443-457)

FIFTH SET OF TOPICS

One chapter (and relevant papers) from Handbook of Statistical Bioinformatics, Henry Horng-Shing Lu, Bernhard Schölkopf, Hongyu Zhao Springer, 2011 (<https://link.springer.com/book/10.1007/978-3-642-16345-6>)

✓ **Sequence Analysis**

Accuracy Assessment of Consensus Sequence from Shotgun Sequencing Lei M. Li

Statistical and Computational Studies on Alternative Splicing Liang Chen

Statistical Learning and Modeling of TF-DNA Binding Bo Jiang, Jun S. Liu

Computational Promoter Prediction in a Vertebrate Genome Michael Q. Zhang

Discovering Influential Variables: A General Computer Intensive Method for Common Genetic Disorders Tian Zheng, Herman Chernoff, Inchi Hu, Iuliana Ionita-Laza, Shaw-Hwa Lo

STORMSeq: A Method for Ranking Regulatory Sequences by Integrating Experimental Datasets with Diverse Computational Predictions Jim C. Huang, Brendan J. Frey

Mixture Tree Construction and Its Applications Grace S. C. Chen, Mingze Li, Michael Rosenberg, Bruce Lindsay

✓ **Expression Data Analysis**

Experimental Designs and ANOVA for Microarray Data Richard E. Kennedy, Xiangqin Cui

The MicroArray Quality Control (MAQC) Project and Cross-Platform Analysis of Microarray Data Zhining Wen, Zhenqiang Su, Jie Liu, Baitang Ning, Lei Guo, Weida Tong et al.

A Survey of Classification Techniques for Microarray Data Analysis Wai-Ki Yip, Samir B. Amin, Cheng Li

Statistical Analysis of Single Nucleotide Polymorphism Microarrays in Cancer Studies Pierre Neuvial, Henrik Bengtsson, Terence P. Speed

Computational Analysis of ChIP-chip Data Hongkai Ji

eQTL Mapping for Functional Classes of *Saccharomyces cerevisiae* Genes with Multivariate Sparse Partial Least Squares Regression Dongjun Chung, Sündüz Keleş

Statistical Analysis of Time Course Microarray Data Lingyan Ruan, Ming Yuan

✓ **Systems Biology**

Kernel Methods in Bioinformatics Karsten M. Borgwardt

Graph Classification Methods in Chemoinformatics Koji Tsuda

Hidden Markov Random Field Models for Network-Based Analysis of Genomic Data
Hongzhe Li

Review of Weighted Gene Coexpression Network Analysis Tova Fuller, Peter
Langfelder, Angela Presson, Steve Horvath

Liquid Association and Related Ideas in Quantifying Changes in Correlation Ker-
Chau Li

Boolean Networks Tung-Hung Chueh, Henry Hornng-Shing Lu

Protein Interaction Networks: Protein Domain Interaction and Protein Function
Prediction Yanjun Qi, William Stafford Noble

Reverse Engineering of Gene Regulation Networks with an Application to the
DREAM4 in silico Network Challenge Hyonho Chun, Jia Kang, Xianghua Zhang,
Minghua Deng, Haisu Ma, Hongyu Zhao

Inferring Signaling and Gene Regulatory Network from Genetic and Genomic
Information Zhidong Tu, Jun Zhu, Fengzhu Sun

Computational Drug Target Pathway Discovery: A Bayesian Network Approach
Seiya Imoto, Yoshinori Tamada, Hiromitsu Araki, Satoru Miyano

Cancer Systems Biology

Elana J. Fertig, Ludmila V. Danilova, Michael F. Ochs

Comparative Genomics Xuhua Xia

Robust Control of Immune Systems Under Noises: Stochastic Game Approach
Bor-Sen Chen, Chia-Hung Chang, Yung-Jen Chuang