

Analytical Approaches in Comparative Genomics

Traditional biology has changed dramatically in the past two decades to the degree that advancements in the 21st century will be dominated by research at the convergence of biological, physical, and information sciences. Driven by large scientific endeavors such as the human genome project, it has become very much an information science, closely tied to tools and methods of the mathematical sciences (e.g. modeling, algorithms, etc.). In parallel, high-throughput sequencing, expression profiling and proteomics technologies are generating exponentially growing amounts of data. This interplay between the mathematical and biological sciences, and the necessity to process and annotate this new data gave rise to a new, interdisciplinary field, called *bioinformatics*.

My Ph.D. was focused on mathematical and algorithmic solutions to problems in bioinformatics, in particular in the field of evolutionary tree reconstruction (phylogenetics). The problems I have investigated are characterized by a scarcity of accurate, fast methods. This is partly due to the following reasons:

- Lack of understanding of the biological factors influencing the evolutionary process.
- Almost all problems arise in this field are computationally hard.

My research has sought to address these two issues. I have done so by pursuing a number of directions:

- Understand existing mathematical models and inspect how well they represent the actual biological processes.
- Introduce new efficient algorithmic procedures to address established problems.
- Develop new models for evolutionary processes using large biological data sets.

For my research I use tools from fields such as combinatorial optimization, statistics and probability, mathematics, and information theory. During my postdoc and in Israel I have established myself as an independent researcher setting his own agenda and focus. My research is characterized by ample collaborations with researchers from broad disciplines under the general framework of a systematic analysis of evolutionary processes aiming at finding biologically significant patterns. Below I detail on my future research directions.

Phylogenetics: Phylogenetics, the reconstruction of the evolutionary history of a group of species, is increasingly integrated into modern biological areas such as preventive medicine and epidemiology. Understanding the mechanisms underlying the evolution of pathogen species is crucial for controlling important human, animal and plant diseases. Maximum Likelihood (ML) is currently considered as the most accurate phylogenetics method. Part of my Ph.D. work involved analytical solutions to ML phylogenetic reconstruction [1, 3, 5, 4, 2]. The novelty of my approach was the representation of the ML problem as a constraint optimization problem and using algebraic geometry tools for obtaining the solution. The work appeared in RECOMB, the world's most important bioinformatics conference and have since sparked a wave of mathematical interest in the field, in particular at UC Berkeley where I later took on a position as a postdoctoral research fellow. In Berkeley I used semi-definite programming (SDP) approach to solve a different problem in phylogenetics - constructing a big tree from a set of overlapping small trees. In a series of papers [13, 16, 14], the SDP approach was established as a successful strategy in particular for coping with conflicting inputs. Inferring reliable small trees (quartets) under a stochastic process of evolution is a subject of another recent project I pursue with researchers from the Technion [7]. The reliability of the quartets with the speed

of the SDP approach makes the combination of the two a promising future research direction I intend to pursue in this project through my collaboration with Profs. Warnow and Rao from UT Austin and UC Berkeley.

Horizontal Gene Transfer: Horizontal gene transfer (HGT), the passage of genetic material between genetically distant organisms, is a significant factor in microbial evolution. HGT plays a major role in the emergence of novel human diseases, as well as promoting the spread of antibiotic resistance in bacteria species. HGT puts in question the validity of a single common (*tree like*) evolutionary history suggesting an alternative evolutionary *network*. Trivial problems on trees turn to be computationally hard on a network setting. Current methods for HGT analysis are mostly based on intuition or weak signals. As W. Ford Doolittle, possibly the greatest authority on HGT, wrote in his inaugural article for the National Academy of Sciences -"In the near future, even more sophisticated methods should be available, because mathematical research into phylogenetic network reconstruction is presently very active" [6]. It is not a co-incidence that Doolittle chose to cite my ML [8] work in this paragraph by its pioneering role.

My research of HGT proceeds along seemingly unrelated tracks. In the first, the *phylogenetic* track, I have formulated several rigorous frameworks to detect and analyze HGT [8, 9, 10, 11, 12, 15]. These works were the first to model realistically HGT. Incongruities in the data set at hand are explained by introducing the least number HGT events, represented by horizontal edges along the species phylogeny. I formulated both combinatorial and statistical models for the HGT phenomenon. On a second, the *sequence based* track, the goal is to identify statistically significant horizontally transferred elements (HTE's) between genomes. I devised an algorithm operating under a stochastic model of evolution to identify unexpectedly highly similar regions. These regions can result from various evolutionary processes such as conservation, translocation, HGT. Using probabilistic and graph theoretical tools, we can distinguish between these events. The work is currently under review.

My future plan in this line of research is to extract more information from remnants of evolutionary events. These may provide us better ideas with respect to major evolutionary processes operating at both intra- and inter-genome levels. Below, I detail on few future topics: **(A) Structural Properties of Phylogeny Networks:** Our algorithms have provided good experimental as well as biological results. However, we still lack deep understanding of structural properties of these networks. Obtaining such result will provide us with better algorithmic solutions such as approximation and/or fixed parameter algorithms. These are of prime importance as they can be useful in realistic cases where the number of HGT events is small. **(B) Algorithmic Accuracy:** In a recent paper [15] I formulated the network by the accurate hidden Markov model (HMM) where the network trees are the states of the HMM and state transitions signify HGT events. I showed that mutation rate accelerates after HGT, a process called *amelioration*. This data is unique and important. Similar improvements in algorithm accuracies will provide further results as this. **(C) HT in Eukaryotes:** HGT in bacteria is straightforward, lending its name: Horizontal **Gene** Transfer. The recent finding of HT in eukaryotes may reveal the transfer of not necessarily coding, or even functional DNA. A major advantage of our approach over other techniques, is being alignment free, naturally fit for HT analysis in eukaryotes where coding DNA comprises a small fraction. Discoveries here may have tremendous impact on topical areas such as functionality.

A grant proposal of \$120,000 for the Israel Science Foundation on this direction was submitted aiming at employing two research students for three years.

Structural Patterns of Non-Coding Eukaryotic Genomes: Revealing the interplay between the functional and structural genome organization is one of the main objectives of modern genomics. The complexity of the problem derives from the fact that in higher eukaryotes the proportion of coding DNA is small. The current comparative genomics relies on aligned sequences, hence mainly limited to coding sequences. Recent discoveries that large conserved blocks of non-coding DNA may have dramatic phenotypic impacts indicate that gene-based analysis is not sufficient for understanding functional organization, regulation, and evolution of eukaryotic genomes. Absence of a systematic approach to characterize the organization of non-coding genome reduces also the promises of detecting meaningful patterns in genomic distribution of functionally related genes in eukaryotes.

In a recent endeavor I adapted a linguistic-like approach of genome comparisons at the *above-gene-level*, referred to as *compositional spectra analysis* (CSA). A genomic segment is characterized by scoring the occurrences of fixed length words of a fixed *vocabulary*, defining the segment's *spectrum*. Spectrum similarities implies segment similarities where similar segments are clustered together. The approach is alignment independent. Moreover, it relaxes the premise that similarity implies homology. This allows for the first time a whole genome analysis and extension and verification of cardinal questions previously studied only at the gene-level. Among my future directions in this area: **(A)** Classification of major patterns of non-genic sequence organization within eukaryotic genomes and their distribution within and between chromosomes. This might serve for defining a *compositional orthogonal basis* in the space of major compositional patterns of across different genomic segments. **(B)** The sought for *synteny* is fundamental in genome evolution research. Defining compositional basis at the genome level will enable extension and revision of parallel syntenies among eukaryotes established earlier on the basis of conserved orders of structural genes. **(C)** Looking for shared compositional patterns (signatures) of gene-rich regions carrying clusters of functionally related genes. Results in this direction can be extremely helpful in the field of gene finding, or regulatory networks.

The aforementioned goals also pose major technical challenges I intend to tackle. The current application of CSA is rather intuitively based, requiring analytical, well established handling. Formulating it in a statistical framework will allow likelihood analysis of the segment clustering penalized by the model complexity as measured by information theoretical standards (BIC, AIC). Also the vocabulary selection process done today is based on intuitions rather than on combinatorial rigor. As the current approach of CSA allows some level of mismatches, this task correlates to the field of coding theory and the specific vocabulary sought is denoted as *covering code*. As there are known results in coding theory with respect to this task, I have already taken preliminary steps in this direction.

References

- [1] A. Khetan B. Chor and S. Snir. Maximum likelihood on four taxa phylogenetic trees: Analytic solutions.
- [2] B. Chor, M. Hendy, and S. Snir. Maximum likelihood jukes-cantor triplets: Analytic solutions. *Molecular Biology and Evolution*, 23(3):626–632, 2006.
- [3] B. Chor, A. Khetan, and S. Snir. Maximum likelihood on molecular clock comb: Analytic solutions. *Journal of Computational Biology*, 13(3):819–837, 2006.

- [4] B. Chor and S. Snir. Molecular clock fork phylogenies: Closed form analytic maximum likelihood solutions. *Systematic Biology*, 53(6):963–967, 2004.
- [5] B. Chor and S. Snir. Molecular clock forks: Symbolic mathematical analysis. *Mathematical Biosciences*, 208(2):347–58, 2007.
- [6] W. Ford Doolittle and Eric Baptiste. Inaugural Article: Pattern pluralism and the Tree of Life hypothesis. *Proceedings of the National Academy of Sciences*, 104(7):2043–2049, 2007.
- [7] I. Gronau, S. Moran, and S. Snir. Fast and reliable reconstruction of phylogenetic trees with very short branches. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 379–388, 2008.
- [8] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Maximum likelihood of phylogenetic networks. *Bioinformatics*, 22(21):2604–11, 2006. **Authors by alphabetical order.**
- [9] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics*, 23(2):e123–8, 2007. **Authors by alphabetical order.**
- [10] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. *Molecular Biological & Evolution*, 24(1):324–37, 2007. **Authors by alphabetical order.**
- [11] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. A new linear-time heuristic algorithm for computing the parsimony score of phylogenetic networks: Theoretical bounds and empirical performance. In *ISBRA*, pages 61–72, 2007. **Authors by alphabetical order.**
- [12] G. Jin, L. Nakhleh, S. Snir, and T. Tuller. Parsimony score of phylogenetic networks: Hardness results and a linear-time heuristic. 2008. **Alphabetical order. All authors contributed equally.**
- [13] S. Snir and S. Rao. Using max cut to enhance rooted trees consistency. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(4):323–33, 2006.
- [14] S. Snir and S. Rao. Quartets maxcut: A divide and conquer quartets algorithm. *accepted IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2008.
- [15] S. Snir and T. Tuller. Novel phylogenetic network inference by combining maximum likelihood and hidden markov models. In *WABI*, pages 354–368, 2008. **Alphabetical order. All authors contributed equally.**
- [16] S. Snir, T. Warnow, and S. Rao. Short quartet puzzling: A new quartet-based phylogeny reconstruction algorithm. *Journal of Computational Biology (JCB)*, 2007. accepted.