Report from Dagstuhl Seminar 12291

Structure Discovery in Biology: Motifs, Networks & Phylogenies

Edited by

Alberto Apostolico[1], Andreas Dress[2], and Laxmi Parida[3]

1 **Georgia Institute of Technology, US,** axa@cc.gatech.edu
2 **Shanghai Institutes for Biological Sciences, CN,** andreas@picb.ac.cn
3 **IBM TJ Watson Research Center, US,** parida@us.ibm.com

Abstract

From 15.07.12 to 20.07.12, the Dagstuhl Seminar 12291 "Structure Discovery in Biology: Motifs, Networks & Phylogenies" was held in Schloss Dagstuhl – Leibniz Center for Informatics. The seminar was in part a follow-up to Dagstuhl Seminar 10231, held in June 2010, this time with a strong emphasis on large data. Both veterans and new participants took part in this edition. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar, as well as abstracts of seminar results and ideas, are put together in this report. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Seminar 15.–20. July, 2012 – www.dagstuhl.de/12291
1998 ACM Subject Classification algorithms, theory, pattern matching, text analysis, biology and genetics, health
Keywords and phrases mathematical biology, computational biology, algorithmic bioinformatics, pattern discovery, networks, phylogenetics, stringology
Digital Object Identifier 10.4230/DagRep.2.7.92
Edited in cooperation with Fabio Cunial

1 Executive Summary

Alberto Apostolico
Andreas Dress
Laxmi Parida

In biological systems, similarly to the tenet of modern architecture, form and function are solidly intertwined. Thus to gain complete understanding in various contexts, the curation and study of form turns out to be a mandatory first phase.

Biology is in the era of the "Omes": Genome, Proteome, Toponome, Transcriptome, Metabolome, Interactome, ORFeome, Recombinome, and so on. Each Ome refers to carefully gathered data in a specific domain. While biotechnology provides the data for most of the Omes (sequencing technology for genomes, mass spectrometry and toponome screening for proteomes and metabolomes, high throughput DNA microarray technology for transcriptomes, protein chips for interactomes), bioinformatics algorithms often help to process the raw data, and sometimes even produce the basic data such as the ORFeome and the recombinome.

The problem is: biological data are accumulating at a much faster rate than the resulting datasets can be understood. For example, the 1000-genomes project alone will produce more than 10^{12} raw nucleic acid bases to make sense of. Thus, databases in the terabytes, even petabytes (10^{15} bytes) range are the norm of the day. One of the issues today is that our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data with the ever advancing bio- and computing technologies. So, while the sheer size of data can be daunting, this provides a golden opportunity for testing (bioinformatic) structure-discovery primitives and methods.

Almost all of the repositories mentioned here are accompanied by intelligent sifting tools. In spite of the difficulties of structure discovery, supervised or unsupervised, there are reasons to believe that evolution endowed biological systems with some underlying principles of organization (based on optimization, redundancy, similarity, and so on) that appear to be present across the board. Correspondingly, using evolutionary thoughts as a "guiding light", it should be possible to identify a number of primitive characteristics of the various embodiments of form and structure (for instance, simply notions of maximality, irredundancy, etc.) and to build similarly unified discovery tools around them. Again, the forms may be organized as linear strings (say, as in the genome), graphs (say, as in the interactome), or even just conglomerates (say, as in the transcriptome). And the fact that even the rate of data accumulation increases continuously becomes rather a blessing in this context than a curse. It is therefore a worthwhile effort to try and identify these primitives. This seminar was intended to focus on combinatorial and algorithmic techniques of structure discovery relating to biological data that are at the core of understanding a coherent body of such data, small or large. The goal of the seminar was twofold: on one hand to identify concise characterizations of biological structure that span across multiple domains; on the other to develop combinatorial insight and algorithmic techniques to effectively unearth structure from data.

The seminar began with a town-hall, round-table style meeting where each participant shared with the others a glimpse of their work and questions that they were most excited about. This formed the basis of the program that was drawn up democratically. As the days progressed, the program evolved organically to make an optimal fit of lectures to the interest of the participants.

The first session was on population genomics, covered by Shuhua Xu and Laxmi Parida. The second was on methods on genomic sequences, covered by Rahul Siddharthan and Jonas Almeida. The next talks were on clinical medicine: an interesting perspective from a practicing physician, Walter Schubert, on treatment of chronic diseases, and Yupeng Cun spoke about prognostic biomarker discovery. Algorithms and problems in strings or genomic sequences were covered in an after-dinner session on Monday and in two sessions on Tuesday morning and late afternoon. The speakers were Sven Rahmann, Burkhard Morgenstern, Eduardo Corel, Fabio Cunial, Gilles Didier, Tobias Marschall, Matthias Gallé, Susana Vinga and Gabriel Valiente. The last speaker presented a system called "Tango" on metagenomics, and in a bizarre twist concluded the session and the day with a surprise live Argentine Tango dance performance with one of the organizers of the seminar. The early afternoon session was on metabolic networks, with lectures by Jörg Ackermann, Jun Yan and Qiang Li.

The Wednesday morning session was loosely on proteomics, with lectures by Alex Pothen, Benny Chor, Axel Mosig, Alex Grossmann, and Deok-Soo Kim. Coincidentally, three lecturers of this session shared very similar first names, leading to some gaffes and some light moments at the otherwise solemn meeting.

The Thursday sessions were on phylogenies and networks, with lectures by Mareike

Fischer, Mike Steel, Katharina T. Huber, Christoph Mayer, James A. Lake, Péter L. Erdös, Stefan Gruenewald and Peter F. Stadler. James A. Lake presented an interesting shift in paradigm, based in biology, called *cooperation and competition in phylogeny.* Péter L. Erdös gave a fascinating talk on the realization of degree sequences. Yet another session on strings was covered by Matteo Comin and Funda Ergun on Thursday. The day concluded with a lecture by Andreas Dress on pandemic modeling.

There were a few after-dinner sessions on big data, thanks to Jonas Almeida. An eclectic set of lectures were given on the last session on Friday, by Raffaele Giancarlo on clustering and by Concettina Guerra on network motifs. The meeting concluded with a fascinating lecture by Matthias Löwe on the combinatorics of graph sceneries. The impact of this on biology may not be immediately clear, but such is the intent of these far-reaching, outward-looking seminars.

2 Table of Contents

Executive Summary

Overview of Talks

3 Overview of Talks

3.1 Reduction techniques for network validation in systems biology

Jörg Ackermann (Universität Frankfurt am Main, DE)

The rapidly increasing amount of experimental biological data enables the development of large and complex, often genome-scale models of molecular systems. The simulation and analysis of these computer models of metabolism, signal transduction, and gene regulation are standard applications in systems biology, but size and complexity of the networks limit the feasibility of many methods. Reduction of networks provides a hierarchical view of complex networks, and gives insight knowledge into their coarse-grained structural properties. Although network reduction has been extensively studied in computer science, adaptation and exploration of these concepts are still lacking for the analysis of biochemical reaction systems. Using the Petri net formalism, we describe two local network structures, *common transition pairs* and *minimal transition invariants*. We apply these two structural elements for steps of network reduction. The reduction preserves the CTI-property (*covered by transition invariants*), which is an important feature for completeness of biological models. We demonstrate this concept for a selection of metabolic networks, including a benchmark network of *Saccharomyces cerevisiae* whose straightforward treatment is not yet feasible even on modern supercomputers.

3.2 Fractal decomposition of sequence representation for socializable genomics

Jonas Almeida (University of Alabama, Birmingham, US)

Joint work of Almeida, Jonas; Vinga, Susana
Main reference Almeida, J.S.; Grüneberg, A.; Maass, W.; Vinga, S. (2012). Fractal MapReduce decomposition of sequence alignment. Algorithms for Molecular Biology 7:12.

Universal Sequence Maps provide a generic numerical data structure to represent biological sequences. Recent work decomposing both the representation and the comparison of sequences raises the prospect of highly portable descriptions of human genomes. In this presentation we explore the analytical features of a participated route for computational genomics that uses the web's read-write feature of social networking infrastructure.

3.3 Metazoan conservation profiles reveal species-dependent functional enrichment patterns

Benny Chor (Tel Aviv University, IL)

Joint work of Witztum, Jonathan; Persi, Erez; Horn, David; Pasmanik-Chor, Metsada; Chor, Benny

The availability of a large number of annotated proteomes enables the systematic study of the relationships between protein conservation and functionality. In this work, we explore this question based solely on the presence (or absence) of protein homologues – the so called *conservation profile*. We study the proteomes of 18 metazoans: 11 vertebrates (including 7 mammals) and 7 invertebrates, and examine them from two distinct points of view: the human's (*Homo sapiens*) and the fly's (*Drosophila melanogaster*).

Two relevant protein groups in this context are the "universal proteins" – human/fly proteins having homologues in all 17 other species – and the "orphan proteins" – those with no homologues. But there are many additional complex patterns of conservation profiles (e.g. proteins having homologues in all vertebrates, but no invertebrate homologue), which are also of interest. In order to characterize the relations between such patterns and proteins functionality, and compare the two viewpoints, we employ Quantum Clustering (QC) and the Gorilla gene ontology tools.

We show many common enriched GO terms in universal proteins of human and fly, and lack thereof for non-universal proteins. Working solely with conservation profiles patterns, and clustering them with QC, we uncover interesting functional enrichments for resulting groups of proteins.

3.4 Whole-genome phylogeny based on non-overlapping patterns

Matteo Comin (University of Padova, IT)

With the progress of modern sequencing technologies, a number of complete genomes are now available. Traditional alignment tools cannot handle this massive amount of data, therefore the comparison of complete genomes can be carried out only with ad hoc, alignment-free methods.

In this talk we propose a distance function based on subword compositions called *Underlying Approach* (UA). We prove that the matching statistics, a popular concept in stringology that captures the statistics of common words between two strings, can be derived from a small set of "independent" subwords, namely the irredundant common subwords. We further refine this statistics by avoiding to count the same subword multiple times. This filter removes the overlaps, by discarding the subwords that occur in regions covered by other more significant subwords.

The UA builds a scoring function based on this set of patterns, called *underlying*. We prove that this set is by construction linear in the size of input, without overlaps, and it can be computed efficiently. Results show the validity of our method in the reconstruction of phylogenetic trees. The Underlying Approach outperforms the current state of the art methods. Moreover, we show that the accuracy of UA is achieved with a very small number of subwords, that in some cases carry meaningful biological information.

References

1 Sims, G.E.; Jun, S.-R.; Wu, G.A.; Kim, S.-H. (2009). Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. PNAS, 2009, 106(40): 17077–82.

2 Ulitsky, I.; Burstein, D.; Tuller, T.; Chor, B. (2006). The average common substring approach to phylogenomic reconstruction. J. Comput. Biol. 13(2): 336–50.

3 Comin, M.; Verzotto, D. (2012). Comparing, ranking and filtering motifs with character classes: application to biological sequences analysis. In "Biological knowledge discovery handbook: preprocessing, mining and postprocessing of biological data", M. Elloumi, A.Y. Zomaya (Eds.). Wiley.

4 Comin, M.; Verzotto, D. (2012). Whole-genome phylogeny by virtue of unic subwords. Proceedings of BIOKDD 2012, pp 190-195.

3.5 Rainbow graphs, alignments and motifs

Eduardo Corel (University of Evry, FR)

We present a graph-based approach to tackle the problem of integrating partial sequence similarity data into a multiple sequence alignment. The problem of finding shared similarities among the sequences is formulated as the clustering of a vertex-coloured graph (the incidence graph of the set of similarities) into *colourful* or *rainbow* components, where every colour appears at most once. We further present several combinations of algorithms to solve the NP-hard problem of finding colourful components by minimum edge-deletions. We show that including matching protein domains, or RNA secondary structure predictions, leads to improved multiple sequence alignments.

3.6 Faster variance computation for patterns with gaps

Fabio Cunial (Georgia Institute of Technology, US)

Main reference F. Cunial, "Faster variance computation for patterns with gaps," in Proc. of the Mediterranean Conference on Algorithms (MEDALG), 2012.

Determining whether a pattern is statistically overrepresented or underrepresented in a string is a fundamental primitive in computational biology and in large-scale text mining. We study ways to speed up the computation of the expectation and variance of the number of occurrences of a pattern with rigid gaps in a random string. Our contributions are twofold: first, we focus on patterns in which groups of characters from an alphabet Σ can occur at each position. We describe a way to compute the exact expectation and variance of the number of occurrences of a pattern w in a random string generated by a Markov chain in $O(|w|^2)$ time, improving a previous result that required $O(2^{|w|})$ time. We then consider the problem of computing expectation and variance of the *motifs* of a string s in an IID text. Motifs are rigid gapped patterns that occur at least twice in s, and in which at most one character from Σ occurs at each position. We study the case in which s is given offline, and an arbitrary motif w of s is queried online. We relate computational complexity to the structure of w and s, identifying sets of motifs that are amenable to $o(|w| \log |w|)$ time online computation after $O(|s|^3)$ preprocessing of s. Our algorithms lend themselves to efficient implementations.

3.7 Integrating prior knowledge into prognostic biomarker discovery

Yupeng Cun (Universität Bonn, DE)

Joint work of Cun, Yupeng; Frohlich, Holger

Stratification of patients according to their clinical prognosis is a desirable goal in cancer treatment in order to achieve a better personalized medicine. Reliable predictions on the basis of gene signatures could support medical doctors on selecting the right therapeutic strategy. However, during the last years, the low reproducibility of many published gene signatures has been criticized. It has been suggested that incorporation of network or pathway information into prognostic biomarker discovery could improve prediction performance. In the meanwhile, a large number of different approaches have been suggested for the same purpose.

First, we compared 14 published classification approaches (8 using network information) on six public breast cancer datasets with respect to prediction accuracy and gene selection stability [1]. A gene set enrichment analysis for the predictive biomarker signatures by each of these methods was done to show the association with disease related genes, pathways and known drug targets. We found that, on average, incorporation of pathway information or protein interaction data did not significantly enhance prediction performance, but increased greatly the interpretability of gene signatures. The results indicated that no single algorithm performs best with respect to all three categories in our study. Incorporating network of prior knowledge into gene selection methods in general did not significantly improve classification accuracy, but greatly increased the interpretability of gene signatures compared to classical algorithms.

Second, we present our newly developed algorithm, called fNet, which integrates protein-protein interaction network information into gene selection for prognostic biomarker discovery [2]. Our method is a simple filter-based approach, which focuses on central genes with large differences in their expression. Compared to several other competing methods, our algorithm reveals a significantly better prediction performance and higher signature stability. Moreover, obtained gene lists are highly enriched with known disease genes and drug targets. We extended our approach further by integrating information on candidate disease genes and targets of disease associated transcription factors. The first can additionally increase the association of gene lists to biological knowledge.

References

1 Cun, Yupeng; Fröhlich, Holger (2012). Prognostic gene signatures for patient stratification in breast cancer: accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. BMC Bioinformatics, 13:69 doi:10.1186/1471-2105-13-69.

2 Cun, Yupeng; Abnaof, Khalid; Fröhlich, Holger (2012). Integrating prior knowledge into prognostic biomarker discovery. Submitted.

3.8 Variable length decoding II

Gilles Didier (CNRS, Marseille, FR)

Let us consider a prefix code P (i.e. a set of words in which no word is prefix of another) in which each words is associated to a unique identifier. The classic way of coding with such a code is to transform a sequence over the alphabet of idents into a sequence over the prefix code alphabet by replacing each ident by its counterpart in the code (the prefix property making the reverse operation unambiguous). This is not what we are doing here: we use P to code sequences over the prefix code alphabet into a sequence of idents. The coding of a sequence s is the sequence t in which the ident at position i is the one corresponding to the (unique, thanks to the prefix property) words of the prefix code which occurs at the position i of s (we assume that such a word always exists). We have proved that this coding can be somehow reversed. Our result states that, being given the coding of sequence s by a prefix code P, there exists a sequence s' and a prefix code P' such that:

1. the coding of s' by P' is equal to that of s by P;
2. the words of P' and P have the same length;
3. if a pair (s'', P'') satisfies the two preceding assertions, then s'' can be obtained from s' by a letter-to-letter application.

The sequence s' is what we called the (variable length) local decoding of s.

From the third item above, the alphabet of the local decoding of s extends that of s (it has a greater diversity of symbols). Another, and most widely used in DNA sequence analysis, way of augmenting the alphabet of sequences is to consider k-mers (conversely, the sequence of k-mer of s can be seen as the coding – not the decoding – of s by a prefix code of constant length k). It turns out that the local decoding can be used in place of k- mers and that it somehow shows better properties.

In this talk, we first present variable length decoding, some of the ideas behind it, and a linear algorithm which computes it. We next apply the decoding to two questions where k-mers approaches are widely used: sequence comparison and sequence assembly. We show that local decoding approaches have good results with respect to k-mers, and try to explain why.

In both cases, a critical point in applying our method is the selection of a prefix code relevant with regard to the question. This is done in the following, heuristic way. We first define a score over a decoding which somehow predicts its relevance. Next, we consider a family of prefix codes parameterized by some quantity (something like a depth) and pick up the prefix code among this family which leads to the local decoding with the best score. Naturally, a better way would be to select the prefix code with the best code among the whole set of prefix codes. Because of some (or rather, a lack of) properties of the set of prefix codes and the associated local decodings, this problem is hard to solve in a feasible time.

3.9 Pandemics and the dynamics of quasispecies evolution: facts, models, and speculations

Andreas Dress (Shanghai Institutes for Biological Sciences, CN)

In my lecture, I argue that viral quasispecies dynamics offers the possibility of a "natural vaccination program", due to different degrees of virus virulence in highly heterogeneous viral populations which might explain why, so far, all pandemics have eventually petered out.

3.10 Graphical degree sequences and realizations

Péter L. Erdös (Hungarian Academy of Sciences, HU)

Joint work of Erdös, Péter L.; Kiraly, Zoltan; Miklos, Istvan
Main reference P.L. Erdös, Z. Kiraly, I. Miklos, "On the swap-distances of different realizations of a graphical degree sequence," arXiv:1205.2842v2 [math.CO], 2012.
URL http://arxiv.org/abs/1205.2842v2

One of the first graph theoretical problems which got serious attention (already in the fifties of the last century) was to decide whether a given integer sequence is equal to the degree sequence of a simple graph. One method to solve this problem is the greedy algorithm of Havel and Hakimi, which is based on the *swap* operation. Another, closely related question is to find a sequence of swap operations to transform one graphical realization into another one of the same degree sequence. This latter problem got particular emphasis in connection of fast mixing Markov chain approaches to sample uniformly all possible realizations of a given degree sequence.

Earlier there were only crude upper bounds on the shortest possible length of such swap sequences between two realizations. In this lecture we present formulae for these *swap-distance*s of any two realizations of simple undirected or directed degree sequences. The exact values in those formulae seem to be not computable efficiently. However the formulae provide sharp upper bounds on swap-distances.

3.11 Periodicity in data streams

Ayse Funda Ergun (Simon Fraser University, Burnaby, CA)

As our data sets grow in size, the need for techniques for processing them under limited resources becomes more critical. One model for processing large data sequences is that of *streaming computation*: the input is read sequentially, i.e., "streamed" in one long pass, and the computation is performed while using small (typically logarithmic in the size of the input) memory.

Streaming techniques are well studied in terms of statistical properties such as the frequency of elements in a stream, but not nearly as much in terms of the particular ordering of the input elements. In this talk, we discuss techniques for analyzing order-related trends of a stream, in particular, its self-similar properties. In this context, we show how to find the period of a stream by using polylogarithmic space if the stream is periodic. Surprisingly, we also show a linear space lower bound using communication theory techniques, i.e. we show that it takes linear space to show that a stream is aperiodic (has a very long period). We also show that one can approximate the distance to periodicity in small space.

3.12 Phylogenetically decisive taxon coverage

Mareike Fischer (Universität Greifswald, DE)

In a recent study, Sanderson and Steel defined and characterized *phylogenetically decisive sets* of taxon sets. A set is called phylogenetically decisive if, regardless of the trees chosen for each of its taxon sets, as long as these trees are compatible with one another, their supertree is always unique. It remained unclear whether deciding if a set of taxon sets is phylogenetically decisive can always be made in polynomial time or not. This question was one of the “Penny Ante” prize questions of the Annual New Zealand Phylogenetics Meeting 2012. In my talk, I explain phylogenetic decisiveness and demonstrate a new characterization for it, which then leads to a polynomial time algorithm both for the (simpler) rooted case as well as for the (more complicated) unrooted case.

3.13 On largest-maximal repeats

Matthias Gallé (Xerox Research Center Europe, Grenoble, FR)

Main reference M. Gallé, “Searching for compact hierarchical structures in DNA by means of the Smallest Grammar Problem,” Ph.D. Thesis, Université Rennes 1, 2011.
URL http://tel.archives-ouvertes.fr/tel-00595494/en/

Largest-maximal repeats (also known as *near-supermaximal repeats*) are a class of exact repeats. Stricter than maximal-repeats, they are less redundant and form a basis of all repeats (mirroring the definitions of irredundant/tiling motifs for rigid motifs). In this short talk I present what is known of them and argue that they could be more interesting for several applications that use maximal repeats.

3.14 A unifying framework for stability-based class discovery in microarray data

Raffaele Giancarlo (Università di Palermo, IT)

Joint work of Giancarlo, Raffaele; Utro, Filippo
Main reference R. Giancarlo, F. Utro, "Algorithmic paradigms for stability-based cluster validity and model selection – Statistical methods, with applications to microarray data analysis," Theoretical Computer Science, Vol. 428, pp. 58–79, 2012.
URL http://dx.doi.org/10.1016/j.tcs.2012.01.024

Clustering is one of the most well known activities in scientific investigation in many disciplines [1, 2]. In this beautiful area, one of the most difficult challenges is the *model selection* problem, i.e., the identification of the correct number of clusters in a dataset [4, 5, 6]. Among the few novel techniques for model selection proposed in the last decade, the stability-based methods are the most robust and best performing in terms of prediction, but the slowest in terms of time [3]. Unfortunately, such a fascinating area of statistics as model selection, with important practical applications, has received very little attention in terms of algorithmic design and engineering. Therefore, in order to partially fill this gap, we highlight: (a) the first general algorithmic paradigm for stability-based methods for model selection; (b) a novel algorithmic paradigm for the class of stability-based methods for cluster validity, i.e., methods assessing how statistically significant is a given clustering solution; (c) the main idea behind a paradigm that describes a very efficient speed-up for stability-based model selection methods.

References

1 Andreopoulos, B.; An, A.; Wang, X.; Schroeder, M. (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. Briefings in Bioinformatics 10(3), 297–314.

2 D'haeseleer, P. (2006). How does gene expression cluster work? Nature Biotechnology 23, 1499–1501.

3 Giancarlo, R.; Scaturro, D.; Utro, F. (2008). Computational cluster validation for microarray data analysis: experimental assessment of Clest, Consensus Clustering, Figure of Merit, Gap Statistics and Model Explorer. BMC Bioinformatics 9, 462.

4 Giancarlo, R.; Scaturro, D.; Utro, F. (2008). A tutorial on computational cluster analysis with applications to pattern discovery in microarray data. Mathematics in Computer Science 1, 655–672.

5 Giancarlo, R.; Scaturro, D.; Utro, F. (2009). Statistical indices for computational and data driven class discovery in microarray data. In: Chen, J.Y., Lonardi, S. (eds.) Biological Data Mining, pp. 295–335. CRC Press, San Francisco, USA.

6 Handl, J.; Knowles, J.; Kell, D. (2005). Computational cluster validation in Post-genomic data analysis. Bioinformatics 21(15), 3201–3212.

3.15 Comparing geometries of chains

Alex Grossmann (Laboratoire Statistique & Génome, Evry, FR)

Joint work of Grossmann, Alex; Lesieur, Claire; Feverati, Giovanni; Vuillon, Laurent

A chain is an ordered set of points in three-dimensional space. Let $N \geq 4$ be the number of points in a chain. The chain contains $N - 1$ pairs of consecutive points. Consider the distances between them. The chain contains $N - 2$ triplets of consecutive points. Consider the areas of the corresponding triangles. The chain contains $N - 3$ quadruplets of consecutive points. Consider the volumes of the corresponding tetrahedrons. We have so introduced $3N - 6$ numbers that are manifestly independent of the choice of the reference frame used to define the coordinates. So we have eliminated the six parameters of global translations and rotations, and are dealing with the intrinsic geometry of the chain. The talk illustrates the procedure with elementary examples, and by data from the Protein Data Bank. We also discuss the case where there are several chains in the same reference frame. The first step is the examination of of histograms of the data. This allows a partial decoupling of the geometry from the primary structure. The main problem is then understanding the histograms, which of course requires non-geometrical inputs.

3.16 On the quartet distance between phylogenetic trees

Stefan Gruenewald (Shanghai Institutes for Biological Sciences, CN)

Joint work of Gruenewald, Stefan; Huang, Libo

The quartet distance is one way to quantify how different two phylogenetic trees on the same taxa set are. It is defined to be the number of subsets of cardinality 4 of the taxa set for which the restrictions of the trees are different. Bandelt and Dress showed in 1986 that the maximum distance between two binary trees, when normalized by the number of all 4-sets, is monotone decreasing with n. They conjectured that the limit of this ratio is 2/3 (the 2/3-conjecture). In order to prove this conjecture, it seems to be helpful to look at a generalization for not necessarily binary trees. This allows us to compare trees with few splits (i.e. few interior edges) but many taxa.

A quartet is a split of a 4-set into to pairs. The quartet that splits a set $\{a, b, c, d\}$ into the pairs $\{a, b\}$ and $\{c, d\}$ is denoted by $ab|cd$. A phylogenetic tree whose taxa set contains $\{a, b, c, d\}$ displays the quartet $ab|cd$ if the paths between a and b and between c and d are vertex-disjoint. For two phylogenetic trees T, T' with identical taxa set of cardinality n, let $q(T, T')$ be the number of 4-sets for which each of T and T' displays one of the three possible quartets. Let $s(T, T')$ be the number of 4-sets for which both of T and T' display the same quartet. We conjecture that $s(T, T') \geq 1/3q(T, T') - o(n^4)$, and I give the easy proof for the case where both trees have only one interior edge. Clearly, the new conjecture implies the 2/3-conjecture, and it turns out that the converse is also true.

3.17 Conservation of complexes in protein-protein interaction networks

Concettina Guerra (Georgia Institute of Technology, US)

Joint work of Guerra, Concettina; Ciriello, Giovanni, Mina, Marco; Guzzi, Pietro Hiram; Cannataro, Mario
Main reference G. Ciriello, M. Mina, P.H. Guzzi, M. Cannataro, C. Guerra, "AlignNemo: A local network alignment method to integrate homology and topology," PLoS ONE 7(6):e38107, 2012.
URL http://dx.doi.org/10.1371/journal.pone.0038107

Comparative analysis of protein-protein interaction networks of different species is an important approach to understanding the mechanisms used by living organisms. One of the computational goals of network comparison (or alignment) is revealing sub-networks, or protein complexes, that are conserved throughout evolution.

In this talk, I analyze the behavior of algorithms for the alignment of protein-protein interaction networks, with respect to the architecture of protein complexes. Protein complexes in PPI networks of certain organisms, such as yeast, are thought to have a modular organization. For instance, according to one model proposed in the literature, complexes consist of core components and attachments. The core is defined as a small group of proteins that are functionally similar and have highly correlated transcriptional profiles. The core is surrounded by less strongly connected proteins, defined *attachments*, which allow diversification of potential functions.

I present the recently developed algorithm AlignNemo that identifies conserved complexes in a pair of PPI networks. The discovered conserved sub-networks have a general topology and need not correspond to specific interaction patterns, so that they more closely fit the models of functional complexes proposed in the literature. Based on reference datasets of protein complexes, AlignNemo shows better performance than other methods in terms of both precision and recall. The obtained solutions are biologically sound according to the semantic similarity measure as applied to Gene Ontology vocabularies.

3.18 Recognizing treelike k-dissimilarities

Katharina T. Huber (Univ. of East Anglia, Norwich, GB)

Joint work of Herrmann, Sven; Huber, Katharina T.; Moulton, Vincent; Spillner, Andreas
Main reference Herrmann, Sven; Huber, Katharina T.; Moulton, Vincent; Spillner, Andreas (2012). Recognizing treelike k-dissimilarities. Journal of Classification, (DOI: 10.1007/s00357-012-9115-2).
URL http://arxiv.org/abs/1206.1374

Many methods for constructing phylogenetic trees from distances essentially work by projecting an arbitrary pairwise dissimilarity onto some "nearby" tree metric. Even so, it is well-known that such methods can suffer from the fact that pairwise distance estimates involve some loss of information. As a potential solution to this problem, Pachter et al proposed using k-wise distance estimates, $k > 2$, to reconstruct trees. Their rationale was that k-wise estimates (as opposed to 2-wise estimates) are potentially more accurate since they can capture more information than pairwise distances, a point that was also made by Felsenstein.

In this talk we focus on the following question, which was originally posed by Pachter et al.: given an arbitrary k-dissimilarity, $k > 2$, how do we test whether this map comes from a tree?

3.19 Understanding molecular geometries via the beta-complexes

Deok-Soo Kim (Hanyang University, Seoul, KR)

Joint work of Kim, Deok-Soo; Cho, Youngsong; Sugihara, Kokichi; Ryu, Joonghyun; Kim, Donguk
Main reference D.-S. Kim, Y. Cho, K. Sugihara, J. Ryu, D. Kim, "Three-dimensional beta-shapes and beta-complexes via quasi-triangulation," Computer-Aided Design, Vol. 42, Issue 10, pp. 911–929, 2010.
URL http://dx.doi.org/10.1016/j.cad.2010.06.004

It has been generally agreed that structure is important in understanding biomolecular functions. Among others, geometry is one of the most important aspects of molecular structure. Despite of its importance, the theory for molecular geometry has not been sufficiently investigated. In this presentation, we present a unified geometric theory, the beta-complex theory, for biomolecules, and demonstrate how this theory can be used for solving important molecular structure problems.

The beta-complex is a generalization of the alpha-complex, which is a structure derived from the Voronoi diagram. For point sets, the Voronoi/Delaunay structures are useful for understanding the spatial structure of the point sets. Being a powerful computational tool, the generalization of the Voronoi/Delaunay structures has been made in various directions, including the Voronoi diagram of spherical balls (or also sometimes called spherical atoms). The Voronoi diagram of spherical atoms nicely defines the proximity among the atoms and its dual structure, called the quasi-triangulation, conveniently represents the topology structure of the Voronoi diagram.

This talk introduces the Voronoi diagram of spherical atoms and the quasi- triangulation in the three-dimensional space. Based on the quasi-triangulation, we define a new geometric structure called the *beta-complex* that concisely represents the proximity among all atoms. It turns out that the beta-complex can be used to precisely, efficiently, and easily solve many seemingly unrelated important geometry and topology problems for the atom set within a single framework. Among many potential application areas, structural molecular biology is the most immediate.

Application examples include the following: the most efficient/precise computation of van der Waals volume (and area), the volumes within an accessible/Connolly surface; an efficient docking simulation; the recognition of internal voids and their volume computation; the recognition of molecular tunnels; the comparison (or superposition) of the boundary structures of two proteins; shape reasoning such as measuring the sphericity of protein; the efficient computation of the optimal side-chain placement, etc. We anticipate that more important applications will be discovered. Several pieces of application software based on the Voronoi diagram and the beta-complex are freely available at the Voronoi Diagram Research Center (VDRC, `http://voronoi.hanyang.ac.kr`).

References

1 Kim, Deok-Soo; Cho, Youngsong; Sugihara, Kokichi; Ryu, Joonghyun; Kim, Donguk (2010). Three-dimensional beta-shapes and beta-complexes via quasi- triangulation. Computer-Aided Design, Vol. 42, Issue 10, pp. 911-929.

2 Kim, Deok-Soo; Cho, Youngsong; Sugihara, Kokichi (2010). Quasi-worlds and quasi- operators on quasi-triangulations. Computer-Aided Design, Vol. 42, Issue 10, pp. 874-888.

3 Cho, Youngsong; Kim, Jae-Kwan; Ryu, Joonghyun; Won, Chung-In; Kim, Chong-Min; Kim, Donguk; Kim, Deok-Soo (2012). BetaMol: a molecular modeling, analysis and visualization software based on the beta-complex and the quasi-triangulation. Journal of Advanced Mechanical Design, Systems, and Manufacturing, Vol.6, Issue 3, pp. 389-403.

3.20 Using genomes to track the evolution of life on Earth

James A. Lake (Univ. of California, Los Angeles, US)

Today evolutionary genomics is in a state of crisis because we mistakenly assumed that once complete genome became available, the complete tree of life on Earth could be easily reconstructed in considerable detail. Instead, all of us in the field agree that we cannot easily determine a single tree. Different genes have different histories. However, everyone seems to have different reasons for why they think that this happens. Here, I make the case that Darwinian tree-like evolution, and the "survival of the fittest" metaphor, give an incomplete view of evolution, and that we need to focus more upon both tree-like evolution and cooperation between organisms (endosymbioses and other types of gene sharing). Trees are easy to calculate from genomic data, but we must combine "survival of the fittest" and "cooperation", if we are to reconstruct the evolution of life on Earth. Methods to do this are vastly more complex and are just being developed. I describe some of the remarkable findings that are now being obtained using these new methods.

3.21 Systematic identification of novel gene members of mammalian metabolic pathways

Qiang Li (Shanghai Institutes for Biological Sciences, CN)

The metabolic functions of known enzymes in the metabolic pathways are among the best studied gene functions so far. However, how these enzymes are regulated and how they are linked to other metabolism-related genes such as metabolite transporters is still unclear. Using the fact that functionally related genes are often co-expressed, we develop an efficient computational method to predict novel genes participating in known metabolic pathways by screening genome-wide expression data. We identify the sets of enzymes associated with consecutive metabolic reactions that also show co-expression. Using these co-expressed consecutive enzymes as query sets or baits, we screen the entire mouse microarray datasets in the Gene Expression Omnibus (GEO) database for additional co-expressed genes. Using this method, we also gain insights into the physiological conditions that affect metabolic pathways. Our extended list of co-expressed metabolism-related genes facilitates the identification of their potential regulators using promoter analysis. We further validate that these novel genes also show spatial co-localizations with known enzymes in metabolic pathways by high-resolution in situ hybridization (ISH) data in E14.5 mouse embryos. Our prediction provides novel gene candidates with putative functional roles in metabolic pathways, which will be further investigated and validated by experiments.

3.22 Reconstruction of random scenery

Matthias Loewe (Universität Münster, DE)

A random scenery is a random coloring of the integer lattice Z^d. Usually this coloring is chosen to be IID. The problem of scenery reconstruction starts with a random scenery that cannot be directly observed. The only thing that can be seen are the observations along the path of one (infinite) realization of a random walk. The question now is: can one reconstruct the scenery when only given the observations? We survey results that answer this question in the affirmative (up to shift of the origin and rotation/reflection) under certain technical assumptions.

3.23 Speeding up exact motif discovery by bounding the expected clump size

Tobias Marschall (CWI, Amsterdam, NL)

Joint work of Marschall, Tobias; Rahmann, Sven
Main reference T. Marschall, S. Rahmann, "Speeding up exact motif discovery by bounding the expected clump size," in Proc. of 10th Workshop on Algorithms in Bioinformatics (WABI), pp. 337–349, 2010.
URL http://dx.doi.org/10.1007/978-3-642-15294-8_28

The overlapping structure of complex patterns, such as IUPAC motifs, significantly affects their statistical properties and should be taken into account in motif discovery algorithms. The contribution of this talk is twofold. On the one hand, we give surprisingly simple formulas for the expected size and weight of motif clumps (maximal overlapping sets of motif matches in a text). In contrast to previous results, we show that these expected values can be computed without matrix inversions. On the other hand, we show how these results can be algorithmically exploited to improve an exact motif discovery algorithm. First, the algorithm can be efficiently generalized to arbitrary finite-memory text models, whereas it was previously limited to IID texts. Second, we achieve a speed-up of up to a factor of 135. Our open-source (GPL) implementation is available at `http://mosdi.googlecode.com`.

3.24 Biases in phylogenetic reconstruction

Christoph Mayer (ZFMK, Bonn, DE)

Joint work of Mayer, Christoph; Wägele, Wolfgang; Kück, Patrick; Meid, Sandra; Richter, Stefan

We define a phylogenetic bias as every effect, which influenced the observable site patterns of a molecular data set, such that they cannot be explained by an evolution governed by a single stationary, homogeneous time-reversible Markov process with site rate heterogeneity constrained to invariant sites and gamma distributed rates. For nucleotide data sets this implies a deviation from an evolution governed by a single, time independent GTR+I+G substitution model. Such a bias can severely impede the reconstruction success of all model based tree reconstruction methods. If time reversibility is broken, a bias can be interpreted as a plesiomorphy effect, which is demonstrated using evolution scenarios based on time-dependent base frequencies as well as time-dependent heterogeneous site rates. Our results highlight the vulnerability of model-based tree reconstruction methods under realistic evolutionary scenarios.

3.25 Using heterogeneous sources of information for multiple sequence alignment

Burkhard Morgenstern (Universität Göttingen, DE)

Traditional methods for multiple sequence alignment are based on primary sequence information alone. Numerous algorithmic approaches have been proposed to calculate (near-)optimal alignments in the sense of some objective function defined in terms of sequence similarity. However, many more sources of information are nowadays available to find homologies between nucleic acid or protein sequences.

We use a recently published graph-theoretical approach to multiple protein alignment to combine primary-sequence similarity and other information, such as similarities to known protein domains, in order to obtain improved multiple protein alignments.

3.26 Co-localization and co-segmentation: algorithms and applications in image analysis

Axel Mosig (Ruhr-Universität Bochum, DE)

This talk introduces co-localization as a concept that occurs naturally in the analysis of bioimage data, either in the analysis of multi-label fluorescence microscopy or the segmentation of spectral images obtained from Raman or CARS microspectroscopes. As a universal tool for studying co-localization, the concept of co-segmentation, i.e. the simultaneous segmentation of two related images at the same time, is introduced. We propose algorithms that allow to compute co-segmentations between hierarchical images of different types. A main result are algorithms that allow to compute co-segmentations between fluorescence and spectral image, which has important applications in the annotation and registration of spectral images.

3.27 Combinatorial algorithms for flow cytometry

Alex Pothen (Purdue University, US)

Flow cytometry is a nearly 50-year old technology for studying properties of single cells via scattering and fluorescence induced by lasers, with applications in immunology and diagnosis of diseases. In recent years, flow ctyometry has become multispectral (thirty or more signals can be detected simultaneously), and high-throughput (millions of cells can be analyzed per minute at the single cell level). However, for analyzing the high dimensional, large-scale data generated by the new experimental methodologies, new algorithms from many areas of computer science, mathematics, and statistics are needed. We describe a few of the computational problems in this context.

We also describe FlowMatch, an algorithm for registering different cell types from patient samples using matchings in graphs and hierarchical template construction algorithms from multiple sequence alignment. These cell types are then followed across multiple time points and experimental conditions. We report results from flow cytometry data generated from leukemia, multiple sclerosis, and phiosphorylation shifts in T-cells. High throughput, multispectral flow cytometry coupled with new algorithmic advances enable systems biology discoveries at the single-cell level, leading to personalized medicine and new approaches to drug discovery.

3.28 The reverse complementarity relation is more complex than we thought

Sven Rahmann (Universität Duisburg-Essen, DE)

Joint work of D'Addario, Marianna; Kriege, Nils; Rahmann, Sven
Main reference M. D'Addario, N. Kriege, S. Rahmann, "Designing q-unique DNA sequences with integer linear programs and Euler tours in De Bruijn graphs," in Proc. of German Conference on Bioinformatics 2012. OpenAccess Series in Informatics (OASIcs), vol 26, pp. 82–92, 2012.
URL http://dx.doi.org/10.4230/OASIcs.GCB.2012.82

DNA nanoarchitectures require carefully designed oligonucleotides with certain non-hybridization guarantees, which can be formalized as the q-uniqueness property on the sequence level. We study the optimization problem of finding a longest q-unique DNA sequence. We first present a convenient formulation as an integer linear program on the underlying De Bruijn graph, that allows to flexibly incorporate a variety of constraints; solution times for practically relevant values of q are short. We then provide additional insights into the problem structure using the quotient graph of the De Bruijn graph with respect to the equivalence relation of reverse complementarity. Specifically, for odd q the quotient graph is Eulerian, and finding a longest q-unique sequence is equivalent to finding an Euler tour, hence solved in linear time (with respect to the output string length). For even q, self-complementary edges complicate the problem, and the graph has to be Eulerized by deleting a minimum number of edges. Two sub-cases arise, for one of which we present a complete solution, while the other one remains open.

3.29 Treatment of chronic diseases: is there a logic of failure?

Walter Schubert (Universität Magdeburg, DE)

The steadily increasing inefficiency in treating chronic diseases – in spite of elegant and logic molecular biological studies and helpful applied mathematics – has recently evoked urgent warnings in a report of the World Health Organization (WHO) 2011. The overall question asked by seriously worried scientists and editors of highly ranked journals is "Where are we going wrong?" (e.g. Nat. Rev. Clin. Oncol. 8, 189-190, 2011), whilst the disillusioned pharmaceutical industry closes discovery facilities and dismisses thousands of employees. Surprisingly, corresponding early warnings in a report entitled "The fruits of genomics" (Lehman Brothers and Mc Kinsey, 2001) were totally disavowed by the scientific community, whilst further promoting the likely cause of the problem, commonly known as large-scale expression profiling. This presentation provides insight into a logic of failure caused by a low content trap, which is a blind spot, but can have fatal impact on a patient's survival. Can mathematics help?

3.30 Evolution of eukaryotic centromeres, and the importance of correct sequence alignment

Rahul Siddharthan (The Institute of Mathematical Sciences, Chennai, IN)

Joint work of Siddharthan, Rahul; Jayaraman, Gayathri; Chatterjee, Gautam; Thattikota, Yogitha; Padmanabhan, Sreedevi; Jain, D; Raghavan, D.K.; Sanyal, Kaustuv

I discuss the alignment of non-coding DNA sequence, and show that the majority of alignment programs, written with proteins in mind, fare very poorly on non-coding DNA. I discuss an alternative approach using an evolutionary model, implemented in the program Sigma-2. I then discuss some recent work on the evolution of centromeres in Hemisacomycetous yeasts of the Candida clade, in collaboration with K. Sanyal (Bangalore). While the biology is very interesting, it also illustrates the importance of a careful approach to alignment.

References

1 Jayaraman, G.; Siddharthan, R. (2010). BMC Bioinformatics 11:464.
2 Padmanabhan, S.; Thakur, J.; Siddharthan, R.; Sanyal, K. (2008). Proc. Natl. Acad. Sci. USA, 105(50):19797-802.
3 Chatterjee, G.; Thattikota, Y.; Padmanabhan, S.; Jain, D.; Raghavan, V.K.; Siddharthan, R.; Sanyal, K. (2012). Submitted.

3.31 Computing a consensus of multilabeled trees

Andreas Spillner (Universität Greifswald, DE)

In this talk we consider two challenging problems that arise in the context of computing a consensus of a collection of multilabeled trees. Forming such a consensus is part of an approach to reconstruct the evolutionary history of a set of species for which events such as genome duplication and hybridization have occurred in the past. We outline exact algorithms that have an exponential run time in the worst case, and highlight the impact of several structural properties of the input on the performance of the algorithms. We conclude discussing some open problems and directions for future work.

3.32 Orthologs, co-graphs, gene trees, and species trees

Peter F. Stadler (Universität Leipzig, DE)

Orthology detection is an important problem in comparative and evolutionary genomics. Although most practical approaches start from inferred gene and/or species trees, orthology can be estimated at acceptable levels of accuracy without any phylogenetic information based on determining sets of closest relatives. We recently characterized orthology relations mathematically as co- graphs. This is equivalent to a (in general not fully resolved) gene tree together with an event labeling that identifies each interior node as either a speciation or a gene duplication event. This characterization opens a new avenue to improving computational methods for orthology detection without incorporating phylogenetic information for this purpose. Inferred orthology relations can then be used as partial information or constraints in the reconstruction of gene trees and their associated species trees. Indeed, given an event-labeled gene tree, the assignment of genes to species defines a partially resolved species tree, and hence establishes constraints on the possible phylogenetic relationships deriving directly from the orthology assignments. In this presentation an overview of the mathematical framework and a first glimpse at computational results is be presented.

References

1 Hellmuth, M.; Hernandez-Rosales, M.; Huber, K.T.; Moulton, V.; Stadler, P.F.; Wieseke N. (2012). Orthology relations, symbolic ultrametrics, and cographs. J. Math. Biol. 2012 DOI: 10.1007/s00285-012-0525-x.

2 Hernandez-Rosales, Maribel; Hellmuth, Marc; Wieseke, Nicolas; Huber, Katharina T.; Moulton, Vincent; Stadler, Peter F. (2012). From event-labeled gene trees to species trees. RECOMB RG Niteroi 2012 (accepted).

3.33 What can probability theory tell us about life's past and future?

Mike Steel (University of Canterbury, Christchurch, NZ)

Joint work of Steel, Mike; Gascuel, Olivier; Mooers, Arne; Li, Thomas; Stadler, Tanja

In a landmark 1925 paper, George Udny Yule FRS described a Markov process for explaining the observed distribution of species into genera [6]. In this model, each species can give rise to a new species according to a constant-rate pure-birth process. Eighty-five years later this Yule process and its extensions provide a basis for studying the shape of macroevolution [1]. In this talk, I highlight some results concerning Yule-type processes in evolutionary biology. First I show that the reliable estimation of ancestral information in the distant past depends on whether or not the ratio of speciation to mutation exceeds a critical ratio [2]. For a simple symmetric mutation model, this critical ratio turns out to be 4 (or 6 depending on the estimation method). Then I study the expected distribution of times between speciation events [4, 5] and its implications for how much evolutionary heritage might be lost under simple field of bullets models of extinction [3].

References

1 Aldous, D. (1995). *Probability distributions on cladograms.* In: D. Aldous, R. Pemantle (Eds.), Random Discrete Structures, IMA Volumes in Mathematics and its Applications 76, Springer, pp. 1–18.

2 Gascuel, O.; Steel, M. (2010). *Inferring ancestral sequences in taxon-rich phylogenies.* Mathematical Biosciences, 227: 125–135.

3 Mooers, A.; Gascuel, O.; Stadler, T.; Li, H.; Steel, M. (2012). *Branch lengths on Yule trees and the expected loss of phylogenetic diversity.* Systematic Biology. 61(2): 195–203.

4 Stadler, T.; Steel, M (2012). *Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models*, J. Theoretical Biology, 297, 33–40.

5 Steel, M.; Mooers, A. (2010). *Expected length of pendant and interior edges of a Yule tree.* Applied Mathematics Letters 23(11): 1315–1319.

6 Yule, G. U. (1925). *A mathematical theory of evolution. Based on the Conclusion sof Dr. J.C. Willis.* FRS Phil. Trans. Roy. Soc. 213, 21-87.

3.34 Sequence classification using reference taxonomies

Gabriel Valiente (TU of Catalonia, Barcelona, ES)

Joint work of Ribeca, Paolo; Valiente, Gabriel
Main reference P. Ribeca, G. Valiente, "Computational challenges of sequence classification in microbiomic data," Briefings in Bioinformatics 12(6):614-625, 2011.
URL http://dx.doi.org/10.1093/bib/bbr019

Next generation sequencing technologies have opened up an unprecedented opportunity for microbiology by enabling the culture-independent genetic study of complex microbial communities, which were so far largely unknown. The analysis of metagenomic data is challenging, since a sample may contain a mixture of many different microbial species, whose genome has not necessarily been sequenced beforehand. In this talk, we address the problem of analyzing metagenomic data for which databases of reference sequences are already known. We discuss both composition and alignment-based methods for the classification of sequence reads, and present recent results on the assignment of ambiguous sequence reads to microbial species at the best possible taxonomic rank.

3.35 Pattern matching through iterative function systems: bridging numerical and graph structures for biosequence analysis

Susana Vinga (Technical University, INESC-ID, Lisboa, PT)

Joint work of Vinga, Susana; Carvalho, Alexandra M; Francisco, Alexandre P; Russo, Luís MS; Almeida, Jonas S.
Main reference S. Vinga, A.M. Carvalho, A.P. Francisco, L.M.S. Russo, J.S. Almeida, "Pattern matching through Chaos Game Representation: bridging numerical and discrete data structures for biological sequence analysis," Algorithms for Molecular Biology, 7:10, 2012.
URL http://dx.doi.org/10.1186/1748-7188-7-10

Background Chaos Game Representation (CGR) is an iterated function that bijectively maps discrete sequences into a continuous domain. As a result, discrete sequences can be object of statistical and topological analyses otherwise reserved to numerical systems.

Characteristically, CGR coordinates of substrings sharing an L-long suffix will be located within 2^{-L} distance of each other. In the two decades since its original proposal, CGR has been generalized beyond its original focus on genomic sequences and has been successfully applied to a wide range of problems in bioinformatics. This presentation explores the possibility that it can be further extended to approach algorithms that rely on discrete, graph-based representations.

Results The exploratory analysis described here consists of selecting foundational string problems and refactoring them using CGR-based algorithms. We find that CGR can take the role of suffix trees and emulate sophisticated string algorithms, efficiently solving exact and approximate string matching problems such as finding all palindromes and tandem repeats, and matching with mismatches. The common feature of these problems is that they use longest common extension (LCE) queries as subtasks of their procedures, which we show to have a constant time solution with CGR. Additionally, we show that CGR can be used as a rolling hash function within the Rabin-Karp algorithm.

Conclusions The analysis of biological sequences relies on algorithmic foundations facing mounting challenges, both logistic (performance) and analytical (lack of unifying mathematical framework). CGR is found to provide the latter and to promise the former: graph-based data structures for sequence analysis operations are entailed by numerical-based data structures produced by CGR maps, providing a unifying analytical framework for a diversity of pattern matching problems.

3.36 Admixture, recombination, human population history, and local adaptation

Shuhua Xu (Shanghai Institutes for Biological Sciences, CN)

Recently available data on genome-wide high-density single nucleotide polymorphisms (SNPs), and the advent of whole-genome sequencing data for human populations, have demarcated a transition from single-locus based studies to genomics analysis of human population structure, history and local adaptation. Apart from the significant increase in the number of loci or markers, the accumulated recombination events in the genome are expected to provide additional information; in addition, it is now applicable to study human admixture at both population-level and individual-level. Here I report our recent research progress on human population history and local adaptation using recombination information in admixed genomes.

3.37 Inter-organ metabolic transport in mammals

Jun Yan (Shanghai Institutes for Biological Sciences, CN)

Complex organisms have evolved separate organs for specialized metabolic functions so that a metabolite is often synthesized in one organ but further catabolized in another. Membrane transporters, especially solute carrier (Slc) proteins, play important roles in shuttling metabolites in and out of the cells. Here we aim to reconstruct the network of inter-organ metabolic transport on the "-omic" scale. This is realized by systematically analyzing the organ- specific expression of enzymes and Slcs using microarray data and high- resolution in situ hybridization data. We provide convincing evidences that the entire metabolic network is segregated in different tissues and inter-organ transport of metabolites is facilitated by strategically located Slcs. Our study provides molecular correlates for the known inter-tissue metabolic transport systems as well as the unknown ones. We discover that there is a "metabolic code" of metabolite fluxes by combinatorial expression of enzymes and Slcs across tissues.

Participants

- Jörg Ackermann
Goethe-Universität Frankfurt am Main, DE
- Jonas Almeida
University of Alabama – Birmingham, US
- Alberto Apostolico
Georgia Inst. of Technology, US
- Benny Chor
Tel Aviv University, IL
- Matteo Comin
University of Padova, IT
- Eduardo Corel
University of Evry, FR
- Yupeng Cun
Universität Bonn, DE
- Fabio Cunial
Georgia Inst. of Technology, US
- Gilles Didier
CNRS – Marseille, FR
- Andreas Dress
Shanghai Institutes for Biological Sciences, CN
- Péter L. Erdös
Hungarian Acad. of Sciences, HU
- Ayse Funda Ergun
Simon Fraser University – Burnaby, CA
- Mareike Fischer
Universität Greifswald, DE
- Matthias Gallé
Xerox Research Center Europe – Grenoble, FR
- Raffaele Giancarlo
Universitá di Palermo, IT
- Alex Grossmann
Laboratoire Statistique & Génome – Evry, FR
- Stefan Gruenewald
Shanghai Institutes for Biological Sciences, CN
- Concettina Guerra
University of Padova, IT
- Katharina T. Huber
Univ. of East Anglia – Norwich, GB
- Deok-Soo Kim
Hanyang University – Seoul, KR
- Jack Koolen
POSTECH – Pohang, KR
- James A. Lake
Univ. of California – Los Angeles, US
- Qiang Li
Shanghai Institutes for Biological Sciences, CN
- Matthias Löwe
Universität Münster, DE
- Tobias Marschall
CWI – Amsterdam, NL
- Christoph Mayer
ZFMK – Bonn, DE
- Burkhard Morgenstern
Universität Göttingen, DE
- Axel Mosig
Ruhr-Universität Bochum, DE
- Laxmi Parida
IBM TJ Watson Research Center – Yorktown Heights, US
- Alex Pothen
Purdue University, US
- Sven Rahmann
Universität Duisburg-Essen, DE
- Walter Schubert
Universität Magdeburg, DE
- Rahul Siddharthan
The Institute of Mathematical Sciences – Chennai, IN
- Andreas Spillner
Universität Greifswald, DE
- Peter F. Stadler
Universität Leipzig, DE
- Mike Steel
University of Canterbury – Christchurch, NZ
- Gabriel Valiente
TU of Catalonia – Barcelona, ES
- Susana Vinga
Technical Univ. – Lisboa, PT
- Shuhua Xu
Shanghai Institutes for Biological Sciences, CN
- Jun Yan
Shanghai Institutes for Biological Sciences, CN