Student Conference on Mathematical Foundations in Bioinformatics

King's College London

20th July 2016


#MatBio2016
Message from Chairs

The students of the Algorithms & Bioinformatics research group would like to welcome you to the first International Student Conference on Mathematical Foundations in Bioinformatics, sponsored by the London Mathematical Society and the Department of Informatics at King’s College London.

The aim of this Conference is to enable communication between early stage pioneers in the field and showcase applications of combinatorics, statistics and general mathematical principles to solving problems in Bioinformatics. It is expected this will inspire professional networking and future advancements in the careers of the postgraduates involved; and it will also promote the sharing and development of methods and approaches to problem solving.

We hope today will be an enjoyable learning experience and encourage future collaborations.

Ritu Kundu, Ahmad Retha and Fatima Vayani
YOUR CONFERENCE PRESENTATION

HOW YOU PLANNED IT:

- Introduce Yourself
- Describe Outline of Talk
- Motivation
- Methodology and Experiment Design
- Results
- Conclusions
- Applause
- Engaging Q&A

15 MINUTES

HOW IT GOES:

- Previous Speaker Runs Late and Eats into Your Time.
- Technical Difficulties Connecting Your Laptop.
- Forget Introducing Yourself.
- Spend Way Too Much Time Describing Your Outline.
- Annoying Audience Member Interrupts with Self-Aggrandizing Question.
- Realize You Only Have 3 Minutes Left.
- Power Through the Rest of Your 30 Slides.
- Awkward Silence Q&A.

15 MINUTES

PhDcomics 21/01/2013, © Jorge Cham
# Programme

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location*</th>
<th>Speaker</th>
<th>Talk Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30 - 09:00</td>
<td>Registration</td>
<td>K3.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:00 - 09:10</td>
<td>Welcome</td>
<td>K3.11</td>
<td>Prof. Peter McBurney</td>
<td></td>
</tr>
<tr>
<td>09.10 - 09.50</td>
<td>Invited talk</td>
<td>K3.11</td>
<td>Prof. Rajeev Raman</td>
<td>Succinct Dynamic Tries</td>
</tr>
<tr>
<td>09.50 - 10.15</td>
<td>Student talk</td>
<td>K3.11</td>
<td>Vladimir Gligorijevic</td>
<td>Revealing modular organization in multi-layer biological networks by non-negative matrix factorization</td>
</tr>
<tr>
<td>10.15 - 10.55</td>
<td>Invited talk</td>
<td>K3.11</td>
<td>Prof. Gregory Kucherov</td>
<td>Algorithms and data structures for large-scale metagenomic analysis</td>
</tr>
<tr>
<td>10.55 - 11.30</td>
<td>Morning Coffee Break</td>
<td>Terrace Cafe*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Session 1 - Chaired by Dr. Solon Pissis

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location*</th>
<th>Speaker</th>
<th>Talk Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.30 - 12.10</td>
<td>Invited talk</td>
<td>K3.11</td>
<td>Prof. Hélène Touzet</td>
<td>Some algorithms for RNA analysis</td>
</tr>
<tr>
<td>12.10 - 12.35</td>
<td>Student talk</td>
<td>K3.11</td>
<td>Jamie Alnasir</td>
<td>Transcriptomics: Leveraging a MapReduce algorithm and Python for gene-expression analysis on Apache Spark</td>
</tr>
<tr>
<td>12.35 - 13.00</td>
<td>Student talk</td>
<td>K3.11</td>
<td>Alice Heliou</td>
<td>Combinatorial RNA Design</td>
</tr>
<tr>
<td>13.00 - 14.50</td>
<td>Lunch</td>
<td>Fire and Stone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td>Location*</td>
<td>Speaker</td>
<td>Talk Title</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------</td>
<td>-----------</td>
<td>----------------</td>
<td>----------------------------------------------------------------</td>
</tr>
<tr>
<td>14.50 - 15.30</td>
<td>Invited talk</td>
<td>K3.11</td>
<td>Dr. Simon Puglisi</td>
<td>Indexing for Pan-genomics</td>
</tr>
<tr>
<td>15.30 - 15.55</td>
<td>Student talk</td>
<td>K3.11</td>
<td>Nicola Prezza</td>
<td>A Randomized In-Place LCE Data Structure</td>
</tr>
<tr>
<td>15.55 - 16.20</td>
<td>Student talk</td>
<td>K3.11</td>
<td>Jia Gao</td>
<td>Optimal Computation of Avoided Words</td>
</tr>
<tr>
<td>16.20 - 16.40</td>
<td>Afternoon tea break</td>
<td></td>
<td></td>
<td>* Terrace Cafe*</td>
</tr>
</tbody>
</table>

**Session 4 - Chaired by Fatima Vayani**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location*</th>
<th>Speaker</th>
<th>Talk Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.40 - 17.20</td>
<td>Invited talk</td>
<td>K3.11</td>
<td>Dr. Nadia Pisanti</td>
<td>Dynamic Programming Approaches to Haplotype Assembly</td>
</tr>
<tr>
<td>17.20 - 17.45</td>
<td>Student talk</td>
<td>K3.11</td>
<td>Sofia Teixeira</td>
<td>SpliceTAPyR – Efficient method for transcriptome alignment</td>
</tr>
<tr>
<td>17.45 - 18.10</td>
<td>Student talk</td>
<td>K3.11</td>
<td>Maryam Abdollahyan</td>
<td>Transcription Factor Binding Site-based Alignment of Conserved Non-coding Sequences</td>
</tr>
<tr>
<td>18.10 - 18.20</td>
<td>Closing remarks &amp; Award announcement</td>
<td>K3.11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location*</th>
<th>Speaker</th>
<th>Talk Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.20 - 19.30</td>
<td>Drinks reception</td>
<td></td>
<td></td>
<td>* Terrace Cafe*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.45: Dinner</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L‘ulivo</td>
</tr>
</tbody>
</table>

* K3.11 - King’s College, Strand Campus, King’s Building, 3rd floor
Terrace Cafe - King’s College, Strand Campus, Macadam Building, Ground floor
Abstracts

Invited Speakers

**Succinct Dynamic Tries**
**Prof. Rajeev Raman** – University of Leicester

With the proliferation of large string data, we need efficient data structures such as tries for storing large collections of strings, and process them in main memory. Succinct data structures are one answer to this problem: they store data in almost the information-theoretical minimum space but perform operations on them quickly. Succinct data structures are typically static: this work will discuss some theoretical and practical advances in succinct dynamic tries (joint work with Andreas Poyias and Simon Puglisi).

**Prof. Gregory Kucherov** – Université Paris-Est Marne-la-Vallée
**Algorithms and data structures for large-scale metagenomic analysis**

The ever-increasing body of DNA sequencing data is turning bioinformatics into a big data science. Current sequence analysis projects routinely use hundreds of gigabytes of sequence data, and terabyte-scale projects are becoming a commonplace. In this talk, we focus on metagenomics – an approach to study microbial populations through massive sequencing all DNA contained in an environmental sample. I will address computational problems arising in this context, and algorithmic techniques and data structures designed to solve them.

**Prof. Hélène Touzet** – Université Lille 1
**Some algorithms for RNA analysis**

Ribonucleic acid (RNA) is one of the three major macromolecules, along with DNA and proteins, that are essential for all forms of life. RNAs are now recognized to play an active role in cells by catalyzing biological reactions and controlling gene expression, and have attracted a lot of interest in molecular biology and in computational biology recently.

Like DNA, RNA is made up of a chain of nucleotides. But a main difference is that RNAs are single stranded and fold into a three-dimensional structure. Due to this spatial structure, RNAs are modeled by complex combinatorial models, such as ordered trees or graphs. In this talk, I will present several basic problems raised by the analysis of RNA (alignment, structure inference) and introduce a universal framework to design dynamic programming algorithms that is especially well-suited for RNA analysis.
Dr. Simon Puglisi – University of Helsinki

Indexing for Pan-genomics

Advances in DNA sequencing mean databases of thousands of human genomes will soon be commonplace. This talk will introduce a simple technique for reducing the size of conventional indexes on such highly repetitive data. Given upper bounds on pattern lengths and edit distances, we preprocess the input data with Lempel-Ziv parsing to obtain a filtered version of it - or kernel - for which we store a conventional index. Later, given a query, we find all matches in the kernel, then use their positions and the structure of the Lempel-Ziv parse to find all matches in the original data. Our experiments show this also significantly reduces both index size and query times, and that the technique scales comfortably to terabytes of data.

Dr. Nadia Pisandi – University of Pisa/INRIA Rhone Alpes

Dynamic Programming Approaches to Haplotype Assembly

The human genome is diploid, which requires to assign heterozygous single nucleotide polymorphisms (SNPs) to the two copies of the genome. The resulting haplotypes, lists of SNPs belonging to each copy, are crucial for downstream analyses in population genetics. Currently, statistical approaches, which are oblivious to direct read information, constitute the state-of-the-art. Haplotype assembly, which addresses phasing directly from sequencing reads, suffers from the fact that sequencing reads of the current generation are too short to serve the purposes of genome-wide phasing.

While future-technology sequencing reads will contain sufficient amounts of SNPs per read for phasing, they are also likely to suffer from higher sequencing error rates. I will describe WhatsHap, the first approach that yields provably optimal solutions to the weighted minimum error correction problem in runtime linear in the number of SNPs. WhatsHap is a fixed parameter tractable (FPT) approach with coverage as the parameter. We demonstrate that WhatsHap can handle datasets of coverage up to 15x, and that 15x are generally enough for reliably phasing long reads, even at significantly elevated sequencing error rates.

I will then show several theoretical results on the optimization problem that lead to HapCol, a fixed parameter algorithm and tool with the number of errors as the parameter. HapCol can handle coverage higher than WhatsHap while being more sensible to the error rate.
Revealing modular organization in multi-layer biological networks by non-negative matrix factorization

Recent technological advances in high-throughput biology have resulted in the accumulation of biological data representing different and complementary views of molecular interactions inside the cell. Such data, also called omics data, enabled construction of complex biological networks with different layers of complexity representing various types of interactions between the same set of molecules. One of the main challenges is identification of functional modules, i.e., sets of molecules, sharing common cellular function, that are densely connected within themselves in the network but sparsely connected with the rest of the network. Standard network methods were shown to be limited in extracting these modules from heterogeneous networks having multi-type interactions. Thus, we propose a novel method capable of collective mining (integration) of various types of biological networks. Our method is based on non-negative matrix factorization, a dimensionality reduction technique; First, it creates a low-dimensional non-negative feature matrices representing individual networks. Then, it performs collective integration of the feature matrices and creates a common feature matrix from which the functional modules are extracted. We apply our method on four large-scale biological networks, including protein-protein interaction, gene co-expression, metabolic and protein complex networks. We identify novel modules that could not be identified by single-network analysis and demonstrate the superior performance of our method over the state-of-the-art methods.

Transcriptomics: Leveraging a MapReduce algorithm and Python for gene-expression analysis on Apache Spark

RNA-Seq (RNA Sequencing) is a next-generation, high-throughput sequencing technology employed in the field of Transcriptomics. A common task in transcriptomics is identifying those transcripts whose expression abundance is altered by experimental conditions which differ between sets of samples and which typically employs complex computational methods in the quantification of expression levels for observed transcripts. RNA-Seq enables researchers in the biomedical and basic science fields to study various aspects of the transcriptome from alternative splicing isoforms, post-transcriptional modifications to mutations and gene expression.

RNA-Seq and other Next-generation sequencing techniques are prone to biases that may be introduced in a number of the steps of a typical sequencing workflow as well as downstream computational methods. This issue has been covered in detail in the authors' paper entitled “Investigation into the annotation of protocol sequencing steps in the sequence read archive”.

Following on from this we have devised a distributed MapReduce algorithm (Hercules) to analyse transcriptomics data, in particular we are interested in examining non-uniform gene expression.

Data generated from transcriptomic studies tends to be large and complex - Bigdata. Delaney characterised such datasets as possessing volume, velocity and variety. To study biases in the data we therefore utilise distributed computing and employ bigdata analytics tools routinely used in industry (Apache Spark) to address the challenge.

The Hercules algorithm therefore applies a distributed programming paradigm (MapReduce) which is central to its implementation. To address part of the bias issue we have devised a methodology for quantifying and quality assessing non-uniform coverage of aligned transcriptomics reads within exons of a transcriptome given genome annotation and aligned reads data.

We have applied Hercules to the analysis of the highly annotated Drosophila Melanogaster transcriptome. We investigated how the GC content of particular motifs, median GC content of a given exon and the random hexamer primer effects described by Hansen et al influence uniformity of read coverage. This is crucial to the measurement of gene expression. We will show our preliminary results and will discuss how we have developed the analyses and the techniques and technologies employed.

**Alice Heliou** – LIX, Ecole Polytechnique

**Combinatorial RNA Design**

We consider the Combinatorial RNA Design problem, a minimal instance of RNA design where one must produce an RNA sequence that adopts a given secondary structure as its minimal free-energy structure. We consider two free-energy models where the contributions of base pairs are additive and independent: the purely combinatorial Watson-Crick model, which only allows equally-contributing A-U and C-G base pairs, and the real-valued Nussinov-Jacobson model, which associates arbitrary energies to A-U, C-G and G-U base pairs. We first provide a complete characterization of designable structures using restricted alphabets and, in the four-letter alphabet, provide a complete characterization for designable structures without unpaired bases. When unpaired bases are allowed, we characterize extensive classes of (non-)designable structures. Finally, we consider a structure-approximating relaxation of the design, and provide a \(\Theta(n)\) algorithm which, given a structure \(S\) that avoids two trivially non-designable motifs, transforms \(S\) into a designable structure constructively by adding at most one base-pair to each of its stems.
Nicola Prezza – University of Udine
A Randomized In-Place LCE Data Structure

Let $T \in \Sigma^n$ be a text of length $n$ over an alphabet $\Sigma = \{0, \ldots, \sigma - 1\}$. The longest common extension (LCE) problem asks to find the length of the longest common prefix between any two suffixes of $T$. Efficient solutions to this problem find numerous applications in bioinformatics, being LCE structures at the basis of several suffix sorting, indexing, and text compression algorithms. In this talk I will present a randomized data structure of size $n \log_2(\sigma)$ bits (the same size of the plain text) based on Rabin-Karp fingerprinting. The structure can be built in optimal $O(n)$ time and supports optimal time extraction of any text substring and $O(\log(n))$-time LCE queries w.h.p. (with high probability). Our approach exploits particular properties of Mersenne prime numbers—i.e. primes of the form $q = 2^p - 1$—in order to reach the information-theoretic optimal space usage and speed-up operations modulo $q$. I will conclude by presenting a C++ implementation that is very fast and space-efficient in practice.

Jia Gao – King’s College London
Optimal Computation of Avoided Words

The deviation of the observed frequency of a word $w$ from its expected frequency in a given sequence $x$ is used to determine whether or not the word is avoided. This concept is particularly useful in DNA linguistic analysis. The value of the standard deviation of $w$, denoted by $\text{std}(w)$, effectively characterises the extent of a word by its edge contrast in the context in which it occurs. A word $w$ of length $k > 2$ is a $\rho$-avoided word in $x$ if $\text{std}(w) \leq \rho$, for a given threshold $\rho < 0$. Notice that such a word may be completely absent from $x$. Hence computing all such words naively can be a very time-consuming procedure, in particular for large $k$. In this article, we propose an $O(n)$-time and $O(n)$-space algorithm to compute all $\rho$-avoided words of length $k$ in a given sequence $x$ of length $n$ over a fixed-sized alphabet. We also present a time-optimal $O(\sigma n)$-time algorithm to compute all $\rho$-avoided words (of any length) in a sequence of length $n$ over an integer alphabet of size $\sigma$. We provide a tight asymptotic upper bound for the number of $\rho$-avoided words over an integer alphabet and the expected length of the longest one. We make available an implementation of our algorithm. Experimental results, using both real and synthetic data, show the efficiency of our implementation.
Sofia Teixeira – INESC-ID
SpliceTAPyR – Efficient method for transcriptome alignment

Background: The RNA-Seq protocol for sequencing the messenger RNA in a cell generates millions of short sequence fragments in a single run. These fragments, or reads, can be used to measure levels of gene expression and to identify novel splice variants of genes. One of the critical steps in an RNA-Seq experiment is mapping NGS reads to the reference transcriptome. Currently, there are several tools developed to make the direct alignment of these fragments in a reference genome.

However, due to the splicing phenomenon of eukaryotic cells, that regards a transcript that may contain parts of more than one exon, these tools may fail, since the case in which a read spans to more than one exon is not covered. Because RNA-Seq reads are short, the task is challenging. In recent years, tools for RNA-Seq alignment have emerged, but most of those tools find junctions by mapping reads to the reference in two phases. In the first phase, the pipeline maps all reads to the reference genome using a tool like Bowtie. All reads that do not map to the genome are set aside as initially unmapped reads. Then, they use heuristics based approaches, or even annotations, to decide where to align them. This work addresses the problem of transcriptome alignment and it presents an efficient computational solution for it, named SpliceTAPyR, which relies on heuristics to identify splice junctions and on compressed full-text indexing methods and succinct data structures for an efficient mapping. SpliceTAPyR extends the TAPyR tool for sequence alignment and is able to align RNA-Seq reads in only one phase.

Methods: SpliceTAPyR uses an implementation of the FM-Index optimized for the DNA alphabet. The FM-Index is a compressed index based on the Burrows-Wheeler transform requiring only $O(n \log \alpha)$ bits of memory space. It then employs a multiple seed heuristic to anchor the best candidate alignment. Contrary to other seed-based alignment tools, TAPyR’s strategy adds more flexibility by dispensing the need of determining the number and length of the seeds. These dynamic seeds correspond to all the left maximal exact matches between the read and the genome, due to the backward nature of the FM-Index’s matching procedure. The heuristic relies on some assumptions that can be reasonably expected to hold true for re-sequencing projects based on pyrosequencing data, namely, that the optimal alignments are mostly composed of relatively large chunks of exact matches interspersed by small, possibly gapped, divergent regions. Banded dynamic programming is then used in the gaps between these multiple anchored seeds to finish up the alignment of the candidate read positions, considering user specified error constraints. Taking advantage of all the characteristics of this computational model, SpliceTAPyR deals with transcriptome alignment by bringing into play a new module that is executed when necessary, reporting the existence of splice evidence, if there is any. We rely on the seeds already efficiently computed by TAPyR and on a greedy strategy to efficiently identify splice signals such as: "GT-AG", "GC-AG" and "AT-AC". This information about splice evidence in the inner tips of pairs of seeds is then used to guide the correct alignment of the transcriptomic reads.
Results: SpliceTAPyR provides many advantages that are crucial to a good performance. The first great advantage, with a great impact in the running time, is the fact that this is the first approach to analyze the set of reads just once, unlike the other tools for transcriptome alignment that have two phases and re-process the reads that were not aligned. Other major aspect is that, like some of the most competitive tools in this area, TAPyR does not require any splice junctions model, any annotation or any junctions library. This allows to discover new splice sites without any influence. Comparing with MapSplice, one of the most competitive tools for transcriptome alignment, we could align more reads and while using only a fraction of the time. For instance, in a standard human dataset, SpliceTAPyR aligned 98.92% of the reads in 12 seconds while MapSplice aligns 98.89% in more than 2 minutes.

Maryam Abdollahyan – Queen Mary, University of London

Transcription Factor Binding Site-based Alignment of Conserved Non-coding Sequences

The identification and functional characterization of regulatory modules in the human genome is a challenging task. Regulatory modules act through the sequence-specific binding of transcription factors and previous studies have demonstrated that co-occurrence of transcription factor binding sites (TFBSs) in close proximity can be a good indicator of regulatory activity. In this study, we analysed the co-occurrence of TFBSs within a set of highly conserved non-coding elements (CNEs) that are associated with the regulation of early vertebrate development. From a computational point of view, analysis of the co-occurrence of TFBSs is complicated by the fact that TFBSs overlap. This rules out the use of classic alignment algorithms (that cannot handle alternative motifs in sequences) or k-mer-based approaches (that count the occurrences of motifs and would enumerate all alternative motifs indiscriminately). Our approach is fundamentally different in that we wrote each CNE as a sequence of symbols, each representing a TFBS identified within that element. We then constructed a graph representation of the CNEs which accounts for the ambiguity due to the overlapping of TFBSs and used a dynamic programming approach to find the optimal alignment between these graphs. We then computed the relative enrichment of short sequences of TFBSs in the alignments of CNEs compared to a background distribution. Our results identify a number of enriched TFBS alignments within CNEs, including a regulatory signature that has been functionally validated in this set of CNEs previously and is associated with hindbrain enhancer activity.