ISSN 2186-7437

NII Shonan Meeting Report

No. 2014–2

Towards the ground truth
Exact algorithms for bioinformatics research

Sebastian Böcker
Gunnar W. Klau
Hon Wai Leong

March 17–20, 2014

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Towards the ground truth
Exact algorithms for bioinformatics research

Organizers:
Sebastian Böcker[1], Gunnar W. Klau[2], Hon Wai Leong[3]

[1]Chair for Bioinformatics, Friedrich Schiller University Jena, Germany
[2]Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
[3]National University of Singapore

March 17–20, 2014

This report is joint work of the attendees of the seminar (see Section 4) and has been edited by the organizers.

Contents

1 Introduction

Today, bioinformatics has become an integral and indispensable part of life science research: Success stories include the assembly and deciphering of genomes, understanding the complexity of cellular processes by means of biological networks, recovering the "tree of life", and deciding on treatment plans for HIV or cancer patients. Applications range from fundamental questions such as the origin of life to multi-billion dollar decisions on novel drug leads and molecular modeling. None of these questions could be approached without massive support from bioinformatics.

Many of the core challenges in bioinformatics can be described as combinatorial optimization problems. Examples are the identification of genes and regulatory structures within genomes; discovering genomic or transcriptomic variations; mining biological networks for, say, protein-protein interactions; or, establishing the evolutionary history of organisms, to name just a few. Unfortunately, a large fraction—and arguably the majority—of these problems are NP-hard: Prominent problems are Multiple Sequence Alignment or Maximum Parsimony in phylogenetics, but there are many more—the query "bioinformatics NP-hard" yields over 12,000 hits in Google Scholar.

It is common practice in bioinformatics to approach these NP-hard problems using heuristics. Although the mathematical model provides only an imperfect approximation to the true goal, namely, to discover *nature's ground truth*, finding optimal solutions is indispensable to rigorously evaluate the quality of the model. Heuristics and approximation algorithms are useless for this purpose, for which *exact algorithms* are needed. Furthermore, good exact algorithms provide deep insight in the structure of the underlying combinatorial problem, which leads to a better understanding of what exactly makes the biological question hard to solve.

In particular, modern measurement techniques such as high-throughput sequencing provide such direct access to the biological ground truth, so that problem modeling can be focused to reverse engineer the biotechnology protocol. While the combinatorial modeling of, for example, assembly problems related to sequencing typically lead to NP-hard problems, the dramatic decrease in sequencing costs also enables multiplexing divide-and-conquer approaches such that inputs to each problem instance become smaller. In these scenarios, exact algorithms for hard problems can be feasible both from the computational and economical perspective.

Recently, there has been much progress on solving combinatorial problems in bioinformatics to provable optimality, despite their hardness. Different techniques have contributed to this progress: in particular, Integer Linear Programming, data reduction and kernelization, and fixed-parameter algorithms. In addition, Algorithm Engineering techniques, which exploit the fact that the structure in realistic problem instances often deviates from the worst case scenario, have contributed to the success of many exact approaches. In contrast, classical exponential-time algorithms such as exhaustive search or higher-dimensional Dynamic Programming have played a negligible role in bioinformatics research.

The aim of this workshop was to bring together researchers active in exact approaches for combinatorial bioinformatics problems. We wanted to tackle the difficult issues these problems pose, and to exchange ideas and theoretical frameworks that allow the design and implementation of algorithms and methods for their solution. Researchers in this workshop came from different areas of algorithmics, such as kernelization and Integer Linear Programming; assembling their views and ideas will foster the applicability of exact algorithms in bioinformatics. Through discussion and sharing knowledge, we promoted collaborations, contribute to the progress in this growing field, and make the field more visible for other scientists.

2 Seminar schedule

After a brief introduction, the participants decided to form working groups to discuss and assess the state of the art as well as work on particular problems and challenges for various topics. Working groups formed partially *ad hoc* and partially based on suggestion by previous working groups. The schedule of the seminar is shown below. Brief abstracts describing the conclusions of the individual working groups are reproduced in Section "Working groups" below.

- **Monday**
  - Transcript assembly and quantification
  - Comparative genomics and family-free gene assignment
- **Tuesday**
  - Somatic mutations and SNPs
  - The Maximum-Weight Connected Subgraph problem
  - Protein-protein interaction networks and dense subgraphs
- **Wednesday**
  - Superbubbles in genome assembly
  - The Maximal Common Induced Subgraph problem
  - Shortest Common Super-Sequence of p-Sequences
  - String Equations
  - Non-negative matrix factorization
- **Thursday**
  - Final discussion

3 Working groups

Detailed descriptions of the working groups were kindly provided by Alexandru Tomescu, Annelyse Thevenin, Fabio Vandin, Nadia Pisanti, Kunihiko Sadakane, Laurent Bulteau, Marco Pellegrini, Mohammed El-Kebir, and Sven Rahmann.

3.1 Transcript assembly and quantification

Transcriptome analysis has been essential in characterizing gene regulation and function, understanding development, disease, and disorders, including cancer. Depending on the individual, on the tissue the cell is in, or on various stimuli, a gene can produce multiple RNA transcripts, with different abundances, through a mechanism called alternative splicing. This mechanism is well understood: a gene transcribes preRNA, out of which some parts are removed to form the mature RNA transcript. However, reading the entire RNA transcripts and estimating their abundances is in practice a challenging problem. RNA-Seq is a recent high-throughput technology producing millions of short reads from the transcriptome, and it has allowed for breakthroughs in transcriptome analysis.

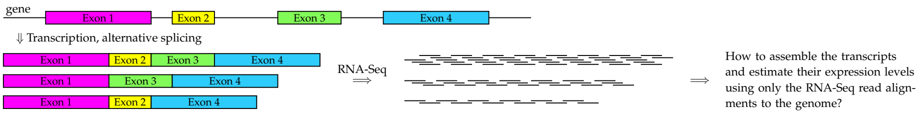
gene
Exon 1
Exon 2
Exon 3
Exon 4
⇓ Transcription, alternative splicing
RNA-Seq
⟹
⟹
How to assemble the transcripts and estimate their expression levels using only the RNA-Seq read alignments to the genome?

Figure 1: Alternative RNA splicing and the RNA-Seq multi-assembly problem.

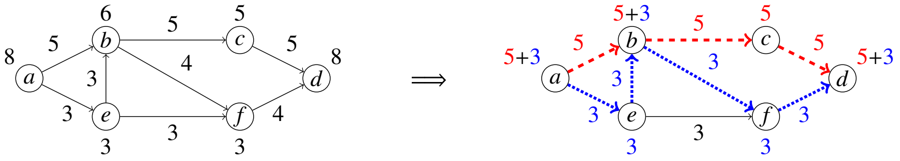

Figure 2: On the left, an input splicing graph whose nodes and edges are labeled with their observed coverages. The optimal paths are the ones depicted on the right, as red dashed and blue dotted, respectively, having abundances 5 and 3, respectively. Their cost is $(6-(5+3))^2+(4-3)^2+(4-3)^2+(3-0)^2=15$, from node b, and edges (b,f), (f,d), (e,f), respectively.

The multi-assembly of the RNA-Seq reads, and the quantification of the resulting transcripts, is usually tackled by first aligning the reads to the reference genome. The gene exons are identified from the read coverage of the gene, and exons that are consecutive in some transcript are identified by reads which span two exons. Then a 'splicing graph' is constructed with single exons as nodes, and consecutive ones as edges; moreover, nodes and edges are labeled with the observed read coverage. In this graph, transcripts correspond to paths.

One objective function for this problem (e.g. [1,2,3,4]) is to look for a number of paths and their corresponding abundances, such that the sum of squared differences between the observed coverage and the predicted coverage of each node and edge is minimized. If looking for a fixed (or bounded) number of paths, this problem is NP-hard. We discussed the following topics.

Connections with other problems

- This problem can be reduced to one having coverage values associated only to edges
- Similarities to a network flow problem, also raised in [3]: one could try employing different strategies for splitting a flow into paths (or design new ones), for example iteratively removing the path of maximum bottleneck (already done in [3]), or of longest length
- Possible formulation as Non-Negative Matrix Factorization problem, see below. However, it is not clear how to impose constraints that the matrix containing the abundance values of each node or edge actually corresponds to a collection of path. Exact algorithms for the NMF problem itself are not very well-known, so this could be a fruitful research direction for both problems

- Already some methods employ ILP formulations. We distributed some papers describing them for further study.

Heuristics

- The estimation of the number of paths, or of their abundance, can be done by looking at the weight of the in-coming edges to each node
- In practice, only a few transcripts (2-3) are highly abundant, so an optimization can be achieved by removing the edges with low coverage, and then looking for few paths in the resulting graph. All edges can then be added to the graph by fixing the already found paths (but not their abundances), and then looking for the remaining paths, together with all abundance levels
- The problem could be extended to use existing transcript annotation, in which case one can try to first explain this with some abundance associated to each transcript, and then try explaining the remaining graph

Practical issues

- There are some issues associated with the RNA-Seq technology not accounted by this model: transcripts have non-uniform coverage, possible start and end sites of the transcripts are hard to find, a low amount of reads from the preRNA is still present in the sample
- Sequencing technologies are able to produce longer reads, that can span multiple exons. These can impose additional constraints on the solution paths, and could provide a more accurate solution

Literature

1. J. Feng, W. Li, and T. Jiang, Inference of isoforms from short sequence reads, RECOMB 2010, Lecture Notes in Computer Science 6044 138-157
2. J. J. Li, C. Jiang, J. Brown, H. Huang, and P. Bickel, Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation, Proc. of the National Academy of Sciences, vol. 108, no. 50, pp. 19 86719 872, 2011.
3. A. I. Tomescu, A. Kuosmanen, R. Rizzi, V. Mäkinen, A novel min-cost flow method for estimating transcript expression with RNA-Seq, BMC Bioinformatics 14(S-5), S15, 2013 (Presented at RECOMB-Seq 2013)
4. A. I. Tomescu, A. Kuosmanen, R. Rizzi, V. Mäkinen, A Novel Combinatorial Method for Estimating Transcript Expression with RNA-Seq: Bounding the Number of Paths, WABI 2013, Lecture Notes in Computer Science 8126, 85-98, 2013

3.2 Comparative genomics and family-free gene assignment

Many methods in computational comparative genomics require gene family assignments as a prerequisite. While the biological concept of gene families is well established, their computational prediction remains unreliable. A new line of research is in which family assignments are not presumed. In this model of “family-free assignment” [1], we need specific data structures (bipartite ordered weighted graph) for which we looking for an optimal matching. Optimal means here the maximization of similarities (weight of saturated edges for example) and/or the minimization of distances (number of breakpoints for example). All relevant problems associated under this model are NP-hard problems. For some (adjacencies, no strict common intervals, DCJ) exact algorithms and heuristics are provided since a couple of years (for some, publications in progress). During this meeting we study the possibility of FTP algorithms for the detection of strict common intervals with for the parameter k the maximal size of the intervals.

Literature

1. M. Braga, C. Chauve, D. Doerr, K. Jahn, J. Stoye, A. Thévenin, R. Wittler. The potential of family-free genome comparison, Models and Algorithms for Genome Evolution conference (MAGE), chapter 13, pages 287-307, 2013

3.3 Somatic mutations and SNPs

The common topic of this working group was the analysis of single base variations in individual genomes: these are either *somatic* mutations, acquired during the lifetime of an individual and that play a crucial role in tumor development, or inherited *single nucleotide polymorphisms* (SNPs). Also, all topics share that members of the working group had a paper about it either RECOMB 2014 or ISMB 2014. We discussed two algorithmic problems related to somatic mutations, namely, reconstructing the subpopulations of tumor cells given a list of somatic mutations with their frequencies, and inferring the progression of somatic mutations from cross-sectional data. Finally, we discussed the problem of haplotype assembly given next-generation sequencing reads. The following paragraphs explain the respective problems and summarize our discussions.

Constructing subpopulations of tumor cells

Next-generation sequencing technologies have enabled the sequencing of many cancer genomes. Recent studies of tumor samples have shown that most tumors exhibit extensive intra-tumor heterogeneity, with multiple subpopulations of tumor cells containing different somatic mutations. We discuss combinatorial formulations of the problem of constructing the subpopulations of tumor cells , and possible solutions.

We discussed the problem, where (*one* list of frequencies is given and the solution presented in [1], which is an exact exponential-time algorithm. We discussed the extension, where multiple lists are given. Here, the goal is to find all trees with minimal number of internal nodes that explain the data. In this situation, it is possible that the problem is not resolvable, and we discussed how to find conditions for resolvability.

Literature

1. Iman Hajirasouliha, Ahmad Mahmoody and Benjamin Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. To appear in Bioinformatics, Proc. ISMB 2014.

Inferring mutation progression from cross-sectional data

Many methods have been proposed to identify the order of mutations in cancer using mutation data from a large number of cancer patients (i.e., cross-sectional data). Most approaches assume that the mutation order is at the level of single genes, while recent works have shown that the mutation order is better understood at the level of pathways (or sets of interacting genes). Current methods to reconstruct the order of mutations at the pathway level are limited to known, a priori defined pathways. Vandin et al. [1] recently introduced an exact algorithm to simultaneously reconstruct the pathways and their mutation order without restricting to known pathways and considering only mutation data. They consider a simple linear progression model for the mutation of pathways in cancer.

We first discussed the model presented in [1], its biological motivation and its relation to previous work, with particular emphasis on the Conjunctive Bayesian Network (CBN) model from Beerenwinkel et al. While CBN is a fairly general model, it does not capture the property of *exclusivity* among mutations in genes in the same cancer pathway, a property that is used in [1].

We then discussed the computational complexity of the combinatorial problem defined in [1], that is the identification of the linear pathway model that minimizes the number of "flips" (changes to the mutation data) to make the data satisfy the model. As proved in [1], this problem is NP-hard, but can be solved exactly for currently available datasets using an ILP formulation.

We went through different strategies to validate the reconstructed models: for example, when clinical data are available, one can test the association of the predicted stage in the linear progression with clinical variables (e.g., tumor stage, survival time); another strategy, used in [1], is to assess the enrichment for interacting genes among the genes in each stage of the reconstructed linear model.

Finally, we talked about some open problems, including i) the extension of the framework to progression models on pathways more complicated than the linear order (e.g., including accumulation of mutations, similar to the CBN model, ii) the design of exact algorithms to identify the best solution for more complicated models, and iii) methods to compare progression models of different complexity.

Literature

1. Raphael, B. and F. Vandin. Simultaneous Inference of Cancer Pathways and Tumor Progression from Cross-Sectional Mutation Data. Research in Computational Molecular Biology (RECOMB 2014). Lecture Notes in Computer Science, Volume 8394, pp. 250–264

Haplotype assembly

The human genome is diploid, that is each of its chromosomes comes in two copies. A full characterization requires to assign the single nucleotide polymorphisms (SNPs) to the two copies. The resulting haplotypes, lists of SNPs belonging to each copy, are crucial for downstream analyses in population genetics. Currently, statistical approaches constitute the state-of-the-art. With increasing read lengths of future generation sequencing, haplotype assembly, which addresses phasing directly from sequencing reads, will become competitive. We are not aware of any exact approach that can handle such kind of data. Recently, dynamic programming approaches have been presented (including Patterson et al., [1]) that address the (weighted) minimum error correction (MEC) problem. The approaches are are linear in the read length and practical up to a coverage of 20x.

- We discussed what makes the problem difficult. Intuitively there should be few flips only, because a SNP should only be called if there is clear evidence for it.
- We discussed the relation to the graph bipartization problem, for which previous work exists. We could formulate the MEC problem as a bipartization problem.
- We discussed whether we could apply data reduction in the graph representation of the problem and then continue to work in the matrix representation. Likely, this will not be possible as data reduction introduces gadgets.
- Solving the problem to optimality for higher coverages is an interesting research direction as well as FPT algorithms.
- We discussed applications for dividing into more than two partitions: this is the case for polyploid organisms, or when the input is given by a population which should be partitioned into a small number of haplotypes, for example, in viral quasispecies identification, when deep-sequencing is applied to strains of a virus.
- We discussed the problem how to distinguish rare SNPs from sequencing or mapping errors
- We also discussed the overall relevance of haplotype assembly.

Literature

1. Patterson, M., T. Marschall, N. Pisanti, L. van Iersel, L. Stougie, G. W. Klau, and A. Schönhuth. 2014. WhatsHap: Haplotype Assembly for Future-Generation Sequencing Reads. Research in Computational Molecular Biology (RECOMB 2014). Lecture Notes in Computer Science, Volume 8394, pp. 237–249

3.4 The Maximum-Weight Connected Subgraph problem

In the maximum-weight connected subgraph (MWCS) problem we are given a simple node-weighted graph and are asked to find a connected subgraph of maximum total weight. This problem is closely related to the prize-collecting Steiner tree problem and is a special case of the node-weighted Steiner tree problem. The main biological application of MWCS is to identify deregulated subnetwork modules by overlaying expression data with biological networks. During the session we discussed the following topics:

- *Data reduction:* We discussed several rules that aim at reducing the size of an input instance. Roughly the rules can be classified into exclusion and inclusion rules. The first category consists of conditions that, when met, exclude nodes from the optimal solution, whereas the latter category describes conditions for when adjacent nodes can be merged.

- *FPT on the number of nodes in the solution:* We spoke about an FPT with the number of solution nodes as a parameter. Combined with the data reduction, this FPT algorithm may be feasible in practice.

- *Characteristics of optimal solutions:* We discussed a sufficient condition that, when met, states that given two nodes u and v, node u is in the solution if and only if v is in the solution as well.

- *Combinatorial upper bound:* We also spoke of a way of obtaining upper bounds on the optimal solution. This may lead to stronger upper bounds than the previously known LP bound.

3.5 Protein-protein interaction networks and dense subgraphs

Protein-protein interaction (PPI) networks represent in graphical form our current systemic knowledge of the mutual interactions among proteins, as detected by high throughput experimental techniques. The interaction can be at the physical level, at the functional level, or represent a common co-expression. Complexes are agglomerations of proteins (usually through shared protein-protein interfaces) that cooperate towards producing a functional effect.

Many types of complexes often appear as dense subgraphs of a PPI network, and thus several models and algorithms have been proposed to predict potential complexes having the PPI network as the main input to the predictor. Most models and algorithms have a strong combinatorial flavor foundation, augmented with specific biological knowledge form protein annotation databases. A Recent new algorithmic result obtain good empirical results by modeling complexes as quasi-cliques, and by estimating the size and density of candidate quasi-cliques via an extension of Turan's theorem. However, as no algorithm or model seems able to cover all possible classes of complexes, thus it becomes important to be able to rank predicted complexes by measuring their features and estimating the probability that a complex with similar features appears by chance in a random PPI. If the PPI is modeled as an Erdos-Rényi random graph and complexes as complete subgraphs (cliques), then there is a very well defined size threshold that discriminates purely random complexes from significant ones.

In this workshop we have proposed that a specific theory is developed in the same spirit to determine thresholds to discriminate random vs significant quasi-cliques (and also, more ambitiously, complexes defined with the core-attachment model, or the conductance model). This is a challenging open problem, and there is widespread consensus among the workshop participants that is a valuable goal for computer scientists and has a potential for practical applications among biologists. However, in order to increase the impact of this research, some care should be placed in augmenting the Erdos-Rényi random graph model with suitable topological (non-uniform bias in the degree distributions) and biological information. For example the model should incorporate weights associated with edges that reflect the strength of an interaction, filters based on known functional annotations and localization, including the dynamic aspects of the interactions. This modeling effort should on one hand incorporate as much biological constraints as possible, while being amenable to effective mathematical derivation of the threshold functions in closed form. Other issue that need to be considered are the effect of experimental errors (false positive, and false negatives) and their impact on the robustness of the significance estimation. This aspect could be analyzed also with the help of simulations of induced errors in realistic PPI data sets.

Although complexes made of heterogeneous protein are easier to model in PPI networks, it is known that often large complexes are made of many copies of one (or few) protein types, thus also homogeneous complexes should be representable in the model with appropriate multiplicities. Proteins with an unusual number of interactions (hubs) may participate in multiple complexes active at different times. The special role of such proteins should be highlighted (since hubs in PPI networks also exhibit high betweenness value, this measure can be used to recognize hubs and give them proper weights).

3.6 Superbubbles in genome assembly

Superbubbles are subgraphs of a genome assembly graph, proposed in [Onodera, Sadakane, Shibuya WABI2013]. To detect sequencing errors in an assembly graph, tips and bubbles have been used. However these are too simple to detect complex errors. The superbubble is an extension of the bubble. By using it, we can detect more complex sequencing errors. Though the above paper has proposed an average-case linear time algorithm (*i.e.*, $O(n+m)$ for a graph with n vertices and m edges) for graphs with a reasonable model, the worst-case time complexity of the algorithm is quadratic (*i.e.*, $O(n(n+m))$). Therefore in this Shonan seminar we discussed a worst-case linear time algorithm for detecting superbubbles.

The definition of superbubbles is the following. Definition Let $G = (V, E)$ be a directed graph. If an ordered pair of distinct vertices (s, t) satisfies the following:

reachability t is reachable from s;

matching the set of vertices reachable from s without passing[1] through t is equal to the set of vertices from which t is reachable without passing through s;

acyclicity the subgraph induced by U is acyclic where U is the set of vertices in the above condition;

minimality no vertex in U other than t forms a pair with s that satisfies the conditions above,

then we say that the subgraph in the description of the acyclicity condition is a **superbubble** and s, t and $U \setminus \{s, t\}$ are this superbubble's **entrance**, **exit** and **interior** respectively. For any pair of vertices (s, t) that satisfies the above conditions, we denote the superbubble as $\langle s, t \rangle$.

[1] Passing through a vertex means that visiting and then leaving it, not just visiting or leaving alone.

In the seminar, we first discussed validity of the definition. For example, why is the minimality condition is not symmetric? Then we tried to improve the worst-case quadratic algorithm.

3.7 The Maximal Common Connected Subgraph problem

In the Maximum Common Connected Subgraph problem we are given two simple graphs and are asked to find the largest common (induced) subgraph that is connected. This problem has applications in computational chemistry as well as in computational biology. The current application we are studying is to identify maximum common fragments of two chemical structures. We discussed the following topics:

- *Applications in chemistry and biology:* We discussed applications of the problem formulation in chemistry (such as comparison of organic molecules) and biology (such as finding conserved protein complexes in more than one PPI network).

- *Finding maximum c-cliques in the product graph:* Maximim c-cliques in the product graph correspond to largest common *connected* induced subgraphs. We spoke about the definition of a c-clique and also how the Bron-Kerbosch algorithm can be adjusted to find c-cliques.

- *Solving the problem on the complement product graph:* We discussed how this can be done via reduction to the maximum independent set problem. Also the product graph exhibits structure that can be exploited in a divide-and-conquer scheme.

3.8 Shortest Common Super-Sequence of p-Sequences

In this problem, we are given a set of *p-sequences* (that is, strings where no letter may be duplicated) over an alphabet of size n, and a parameter k. The objective is to find a common super-sequence of length at most $n + k$. The problem of finding a shortest common super-sequence is well known in bioinformatics, since it aims at aggregating genomic or sequencing data with some dissimilarities. As an example, the input $\{abcde, bca, baec\}$ yields a solution with $k = 2$, namely string *abcadec*.

Interestingly, this problem generalizes Feedback Vertex Set (FVS), hence, aiming at an FPT algorithm, one could possibly use the *iterative compression* machinery which lead to an FPT algorithm for FVS.

3.9 String Equations

This problem was first defined as a slightly constrained version of Minimum Common String Partition (MCSP). In MCSP, the goal is to decompose two genomes represented as strings into a common multi-set of substrings (*blocks*). MCSP is intended as a first step or as an approximation for computing rearrangement distances. Indeed, a decomposition into blocks highlights the conserved regions between two genomes, hence a rearrangement distance can be more simply computed between the permutation of the blocks. MCSP is NP-hard but fixed-parameter tractable if parameterized by the number of blocks.

The string equation problem was thus intended as a variant of MCSP where the arrangement of the blocks is given in the input. For instance, decomposing the strings $X = abcab$, $Y = bcaab$ into the arrangement $X = X_1X_2X_3$, $Y = X_2X_1X_3$ yields the following solution: $X_1 = a$, $X_2 = bc$, $X_3 = ab$.

Surprisingly, it seems that the number of blocks does not yield a parameterized algorithm for this problem as it does for MCSP, hence we look at further restrictions. Considering several constraints on the "shape" of the equations (bounds on the number of blocks, maximum number of blocks per equations, etc.), we aim at better understanding the parameterized complexity of the problem.

3.10 Non-negative matrix factorization

Non-negative matrix factorization (NMF) is a technique to explain observed data as a weighted sum of simple (prototypical) parts. To be precise, let $Y \in \mathbb{R}^{m\times n}$ be a data matrix of n samples, each of which is an m-dimensional vector, such that $Y_{ij} \geq 0$ for all i, j. Assume that we find $A \in \mathbb{R}^{m\times k}$ and $X \in \mathbb{R}^{k\times n}$ for $k \ll \min\{m, n\}$ such that $Y = AX$ and all entries of A and X are non-negative as well. Then each column of Y (sample) is a non-negative linear combination of the columns of A (the k prototypical samples); the weights for j-th column of Y are given by the j-th column of X. Conversely, each row of Y (the values of a feature over time/samples) is a non-negative linear combination of the rows of X (the k prototypical feature behaviors), with the coefficients for the i-th row being given by the i-th row of A. Because of non-negativity, all effects are additive, and none cancel out.

In practice, we will not be able to achieve $Y = AX$ exactly, so in fact we are looking for a rank-k approximation of Y by minimizing $d(Y, AX)$ for an appropriate distance (error) function d. Typically the squared Frobenius norm $\|\cdot\|_\mathrm{F}^2$ is chosen because it is convenient and differentiable.

There are many variations of this basic problem. One possibility is to vary the error function d. In particular, one can add additional terms or constraints to the problem. In many applications, it is desired that X and/or A are sparse. To exemplify, let $s(A) := \sum_{i,c} |A_{ic}|$ be the ℓ_1-norm of A (interpreted as a vector in $\mathbb{R}^{mk}$). Then the optimization problem may be written as "minimize $d(Y, AX) + \lambda \cdot s(A)$" where $\lambda > 0$ is a parameter balancing the two objectives of fitting Y and obtaining a sparse A.

Challenges.

- As stated, the optimization problem is not convex in (A, X). However, it is convex in X when A is fixed and vice versa; it then is a least-squares problem. Therefore alternating least squares has become a popular method to "solve" NMF problems. However, usually both A and X are unknown, and one may find different local minima without the ability to make a statement on the global minimum.

- If (A, X) is one (say, locally) optimal solution to the problem, take an appropriate invertible $k\times k$ matrix Q such that $A' := AQ$ and $X' := Q^{-1}X$ are both non-negative. Then $A'X' = AX$ and so the solution (A', X') is indistinguishable from (A, X) in terms of objective function value.

- Even if we could characterize the level set of the global optimum exactly, it is not always clear that one of the contained decompositions (A, X) would be the "correct" one that best explains Y for the current application at hand. Frequently, there are constraints on A and Y that arise from the application but that are hard to formalize.

There are several heuristics to find a locally optimal solution, often with great success in practice. Many algorithms have been collected in the libNMF library. Several of these algorithms have been shown to work well with simulated data and seem to compute an intuitively appealing decomposition (even if not always the optimal one). However, it is worrying that NMF has become an important and frequently used tool (thanks to readily available heuristics), *while there is no exact algorithm* that will return the globally optimal solution (A, X) (or one exemplar of the level set). The NMF problem has been shown to be NP hard, but that does not mean that it is impossible to find practical exact algorithms for medium-sized problems. Furthermore, we are missing a *comprehensive theory about the uniqueness of the solution*, although progress has very recently been made from a geometric perspective [3]. There exist further results on uniqueness in particular cases, but no comprehensive theory yet.

Delineation of Cancer Types by Mutation Profiles. We now describe an NMF applications in bioinformatics [1]: We assume that we record m distinct mutation types in n tissue samples from cancer patients. A mutation type might be A(G → C)A, meaning that G mutates to C between A and A. The relative abundance of mutation type i in sample j is recorded as Y_{ij}, with $i = 1, \dots, m$ and $j = 1 \dots, n$. Hence each column of Y is a probability distribution. We assume that there exist $k \ll \min\{m, n\}$ distinct types of cancer, each of which has a specific mutation type profile, so the c-th column of A provides the profile of cancer type c. Each observed sample profile is to be written as a mixture of those profiles. Therefore X_{cj} is the coefficient of profile c in sample j, and we have the goal to write $Y \approx AX$.

Indeed, the mentioned article identifies 4 main cancer type profiles from 21 breast cancer samples and 96 mutation types by using the classical multiplicative update NMF algorithm after ad-hoc preprocessing. Their larger simulation studies show that the employed NMF algorithm was able to correctly reconstruct simulated noisy mixtures, but the lack of underlying theory is still discomforting.

Conclusion. There is now an urgent need to focus research effort on exact algorithms and uniqueness theory for NMF. Some concrete questions that could be addressed are:

- Identify a list of natural constraints from typical bioinformatics problems amenable to NMF analysis.

- How strong do additional constraints have to be in order to guarantee a unique globally optimal solution, up to re-scaling and permutations?

- Is adding ℓ_1 regularization terms to the objective be enough to guarantee uniqueness (together with a fixed scale for X)? If not, which other practically relevant constraints are required?

- Alternatively, what about ℓ_1 side constraints, such as $\sum_{i,c} |A_{i,c}| \leq T$ (given a fixed scale for X)?
- What about incorporating non-convex ℓ_0 (number of nonzero entries) regularization terms and/or constraints?

Literature

1. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton. Deciphering signatures of mutational processes operative in human cancer. Cell Reports, 3(1):246–259, 2013.
2. E. Fritzilas, M. Milanic, S. Rahmann, and Y. A. Rios-Solis. Structural identifiability in low-rank matrix factorization. Algorithmica, 53(3):313–332, 2010.
3. K. Huang, N. D. Sidiropoulos, and A. Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. IEEE Transactions on Signal Processing, 62(1):211–224, 2014.

4 Participants

- Sebastian Böcker, Friedrich Schiller University Jena
- Paola Bonizzoni, Universitá di Milano-Bicocca
- Laurent Bulteau, TU Berlin
- Francis Chin, The University of Hong Kong
- Mohammed El-Kebir, CWI Amsterdam
- Mike Fellows, Charles Darwin University
- Iman Hajirasouliha, Brown University
- Falk Hüffner, Technische Universität Berlin
- Gunnar W. Klau, CWI Amsterdam
- Mikko Koivisto, University of Helsinki
- Christian Komusiewicz, TU Berlin & Univ. Nantes
- Hon Wai Leong, National University of Singapore
- Matthias Müller-Hannemann, Martin-Luther University Halle-Wittenberg
- Marco Pellegrini, CNR
- Nadia Pisanti, University of Pisa
- Alberto Policriti, University of Udine and Applied Genomic Institute
- Sven Rahmann, University of Duisburg-Essen
- Frances Rosamond, Charles Darwin University
- Kunihiko Sadakane, NII

- Chuan Yi Tang, Providence University
- Annelyse Thévenin, Bielefeld University
- Alexandru Tomescu, University of Helsinki
- Fabio Vandin, Brown University
- Annegret Wagler, Université Blaise Pascal
- Tim White, FSU Jena
- Siu Ming Yiu, The University of Hong Kong