Research Proposal:
Computational methods for reconstructing ancestral genomes

Ashok Rajaraman
Department of Mathematics, Simon Fraser University

1 Introduction

The broad theme of my research is the development and analysis of algorithms for problems that arise from the reconstruction of ancestral genomes, as well as classifying these problems according to their complexity and tractability. With biotechnology making major advances in the assembly and mapping of genomes from current species, there is a large amount of data to to study the functioning of genomes and their impact on complex biological systems. To understand the forces that shaped current genomes, though, we will need ancestral genomes. This provides us with valuable reference points from which we can study the evolution of molecular and genomic characters. For example, the evolution of the genome architecture in vertebrates is of interest to the study of long-range regulation and of the isochore structure of genomes [14, 24]. In bacteria, this knowledge may provide information to how pathogenicity evolves in microbial genomes, and diseases emerge in the population. A recent study of the causative agent of the Black Death, which has led to insights about the evolution of the disease in the human population, comes to mind as an example [4]. This begs the question what these ancestral genomes looked like, how they evolved into the current genomes, and how can we reconstruct them using the data we have. This project is aimed at laying the theoretical foundation for fully merging ancestral genomics to the fundamental field of comparative genomics, detailing a structured approach to the problem.

Ancestral reconstruction proceeds in a number of steps. The first step consists of defining ancestral markers on extant genomes. These may be genes or other families of genome segments that are inferred by comparative studies. The next step is to find out sets of markers that are conserved on extant genomes, which form the basis for ancestral local linkages. Here, the theory of common intervals, which was developed by Bergeron et al. for permutations [2], and by Schmidt and Stoye for sequences [26], is used to determine sets of markers that probably appeared together in the ancestor. Finally, these markers are ordered in a 'best possible' way into a number of sequences, representing large chromosomal segments of the ancestral genome, with the goal of obtaining full ancestral chromosomes. The term 'best possible' refers to an ordering that respects the sets that were generated. This gives rise to the main mathematical problem in the area, a variant of the Travelling Salesman Problem called the Consecutive Ones problem [3]. It is on variations of this problem that my primary research is based on and forms the first section of this proposal. Broadly speaking, the focus shall be on the tractability and fixed-parameter tractability of these problems, and the design of algorithms that tackle with the tractable cases. A second problem that arises from ancestral genome reconstruction is getting a linear arrangement of the genes that respects their mutual distances. Thus, the aim of the problem is to take a set of points and the mutual distances between the points, and embed it on a line while minimizing the errors between the original distances and the distances after embedding. This is called the distance problem, and forms the second section of this proposal.

Finally, the overarching aim of both sections of my research is to actually apply techniques developed while exploring these problems to real-world data. The final section addresses how to process the data, and in general, getting information from the data that can actually be used as input for the techniques developed.

2 Variants of the Consecutive Ones problem

An $m \times n$ binary matrix is said to have the *Consecutive Ones Property (C1P)* if there exists a permutation of the columns such that all the 1's in each row are consecutive. The original problem was solved by Booth and Leuker in their seminal work using the PQ-tree datastructure [3]. Furthermore, there are forbidden structure characterizations for matrices that do not have the C1P [27, 22], including structures that can be found in linear time.

Since not all matrices have the C1P, we often have to deal with relaxations of the property. For example, the circ-C1P(Ci1P) allows each row to have either all 1's consecutive, or all 0's consecutive [19], the gapped-C1P property allows gaps between consecutive series of 1's in each row, while restricting the size of the gaps by a constant [10, 21], while the x-C1P allows undefined entries in the matrix [16]. A different version of the C1P, called C1P with multiplicities (mC1P), asks for a *sequence* of the columns, instead of a permutation, such that the columns in each row occur as a substring somewhere in the sequence [8].

2.1 Computational complexity

Most C1P variants can be visualized as covering problems on hypergraphs, and can be shown to be NP-hard through reduction from variants of SAT. Other variants can be shown to be equivalent to the Travelling Salesman problem. This raises an interesting question: what relaxations are tractable, what conditions are these problems tractable under, and if a problem is not tractable, is it fixed-parameter tractable? Problems that are tractable are usually restricted to matrices in which each row entry only has two 1's. For example, the mC1P problem is tractable under this restriction.

Parameterized complexity As long as we are dealing with real-world data, there are also bounds on certain parameters that we can use. This opens possibilities for fixed-parameter tractability and parameterized complexity results for these problems. C1P problems usually translate to well-defined covering problems on graphs and hypergraphs. The presence of exact exponential algorithms for some graph covering problems [15] means that these problems may actually be tractable if the parameters are small enough. This is a comparatively new angle of approach to the problem, and still needs to be explored. Some results from parameterized complexity [6, 11] can possibly be extended to our cases, and point towards the directions that we might have to consider while dealing with them. In particular, these results are an encouraging sign that some C1P row deletion optimization problems, which usually turn out to be NP-hard, can be solved.

Non-deterministic algorithms and hardness of approximation Apart from parametric complexity, there is also the question of non-deterministic methods to attack C1P variants, as well as the existence of polynomial time approximation schemes. Since the tractability arguments made in fixed parameter tractability for edge deletion problems in graphs are often randomized [11], this is a possible path that the research may have have to focus on. Such algorithms may indeed be indispensable, since simplified versions of the problems have been proven not to have polynomial kernels unless there is a collapse in the polynomial hierarchy. Also, since there may be fixed parameter intractable problems as well, it would be good to have the option of efficient approximation algorithms.

2.2 Design of algorithms

Algorithms for the C1P are usually based on partition refinement techniques. The quest is to extend these to the variant problems, and to look at new techniques.

Spectral algorithms Spectral algorithms developed for the seriation problem are useful for the C1P problem [1], the main drawback being the non-linear computational time for the eigenvectors of a matrix. Furthermore, for cases that do not have the C1P, these algorithms still provide an output that seems to be related to the length of the gaps between non-consecutive 1's in rows [28]. The fact that the Laplacian spectra of graphs has been the subject of extensive research (survey of some results in [23]) means that there is a good starting point for using spectral algorithms for our variants. In particular, one software package for ancestral genome reconstruction already implements this and extends it to tackle the x-C1P problem [20].

There is still much to understand in this field, particularly about the exact behaviour of the spectra of the Laplacians of non-C1P matrices. The theory ties to the behaviour of the Laplacian spectra of weighted graphs. The gap minimization property is also interesting, and there is work to be done in that area as well.

Partition refinement Partition refinement has already been used to solve the classical C1P problem [22, 17], and has also pointed towards possible algorithms for the x-C1P variant [16]. Current research on partition refinement techniques have used them as heuristics for C1P variants. The goal would be to find such algorithms if they exist, to examine if the the heuristic algorithms currently being used provide good approximations for the actual answer, and extend partition refinement techniques to handle other such problems as well.

3 The Distances problem

The distances problem originates from the problem of estimating the order of markers on a chromosome given the pairwise distances between the markers. In other words, given the set of pairwise distances, we want to embed the markers on a line while respecting the distances. This is a well studied problem, generalizing to other metrics [13, 12, 5]. The challenge is to embed the points on a line while minimizing the errors in all pairwise distances, called the distortion. This problem is known to be NP-hard [25].

The spectral approach to this problem involves using the eigenvectors of the Laplacian matrix of the distances as embeddings on the real line. The function to minimize is non-trivial, and is not minimized by a spectral algorithm. However, there exist approximate algorithms for the same [13, 12], and the possibility of a polynomial time approximation scheme has not been ruled out. The possible theoretical realization of this, and the implementation of the known algorithms would be the main thrust areas here. As an example of how this might be useful in computational biology and ancestral reconstruction, such methods have already been applied to construct radiation hybrid maps [18].

4 Application

All the domains metioned before are forerunners to the actual application of these methods to real data. The generation of this data is taken as a-given, and it usually comes in the form of large common subsequences with few errors that are found on the genomes of the extant (currently still alive) species being examined. Genome data has to be processed, often manually, to a form that can actually be fed to software, and errors in the data are not uncommon.

There is already a freely distributed software package that implements known methods for reconstructing genomes [20]. This package is expected to be developed further to accommodate other methods that are uncovered. It is also expected to be used to actually reconstruct ancestral genomes. The current interest is in bacterial genomes, and the research group also has a strong record in the reconstruction of ancestral mammalian [9] and yeast genomes [7].

5 Summary of planned research

The breadth of my research ranges from the purely theoretical to highly applied. The main focus will be to develop the theoretical tools to deal with the ancestral reconstruction problem, in particular focussing on the complexity and tractability analysis of C1P variants. Parameterized complexity and non-deterministic algorithms are a new take in this field, while the distances project will prove to be useful in the long run.

The application of known methods is already underway, and as we develop more techniques, it is hoped that our estimate at ancestral genomes will get better. There is certainly no dearth of data sets, and the methodological issues that come with each are unique, often having to be dealt with on a case by case basis. An efficient ancestral genome reconstruction pipeline is the ultimate aim, one that can deal with the large variety of data that people in bioinformatics have to sift through. The completion of both parts of this project will be pivotal in the full integration of ancestral genomics in comparative genomics, a paradigm shift that can only follow from the theoretical research that forms the heart of my work.

References

[1] J.E. Atkins, E.G. Boman, B. Hendrickson, et al. A spectral algorithm for seriation and the consecutive ones problem. *SIAM J. Comput.*, 28(1):297–310, 1998.

[2] Anne Bergeron, Cedric Chauve, Fabien de Montgolfier, and Mathieu Raffinot. Computing common intervals of K permutations, with applications to modular decomposition of graphs. *SIAM J. Discrete Math.*, 22(3):1022–1039, 2008.

[3] Kellogg S. Booth and George S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms. *J. Comput. System Sci.*, 13(3):335–379, 1976. Working Papers presented at the ACM-SIGACT Symposium on the Theory of Computing (Albuquerque, N. M., 1975).

[4] K.I. Bos, V.J. Schuenemann, G.B. Golding, H.A. Burbano, N. Waglechner, B.K. Coombes, J.B. McPhee, S.N. DeWitte, M. Meyer, S. Schmedes, et al. A draft genome of yersinia pestis from victims of the black death. *Nature*, 478(7370):506–510, 2011.

[5] M. Bdoiu, K. Dhamdhere, A. Gupta, Y. Rabinovich, H. Räcke, R. Ravi, and A. Sidiropoulos. Approximation algorithms for low-distortion embeddings into low-dimensional spaces. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 119–128. Society for Industrial and Applied Mathematics, 2005.

[6] Leizhen Cai and Boting Yang. Parameterized complexity of even/odd subgraph problems. *J. Discrete Algorithms*, 9(3):231–240, 2011.

[7] C. Chauve, H. Gavranovic, A. Ouangraoua, and E. Tannier. Yeast ancestral genome reconstructions: the possibilities of computational methods ii. *Journal of Computational Biology*, 17(9):1097–1112, 2010.

[8] C. Chauve, J. Maňuch, M. Patterson, and R. Wittler. Tractability results for the consecutive-ones property with multiplicity. In *Combinatorial Pattern Matching*, pages 90–103. Springer, 2011.

[9] C. Chauve and E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS computational biology*, 4(11):e1000234, 2008.

[10] Cedric Chauve, Ján Maňuch, and Murray Patterson. On the gapped consecutive-ones property. In *European Conference on Combinatorics, Graph Theory and Applications (EuroComb 2009)*, volume 34 of *Electron. Notes Discrete Math.*, pages 121–125. Elsevier Sci. B. V., Amsterdam, 2009.

[11] M. Cygan, D. Marx, M. Pilipczuk, M. Pilipczuk, and I. Schlotter. Parameterized complexity of eulerian deletion problems. In *Graph-Theoretic Concepts in Computer Science*, pages 131–142. Springer, 2011.

[12] K. Dhamdhere. Approximating additive distortion of embeddings into line metrics. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 96–104, 2004.

[13] K. Dhamdhere, A. Gupta, and R. Ravi. Approximating average distortion for embeddings into line. In *Proceedings of the Symposium on Theoretical Aspects of Computer Science (STACS)*, 2004.

[14] M. Emmanuel, D. Ken, and B. Mathieu. Long-range regulation is a major driving force in maintaining genome integrity. *BMC Evolutionary Biology*, 9, 2009.

[15] F.V. Fomin and D. Kratsch. *Exact exponential algorithms*. Springer Verlag, 2010.

[16] H. Gavranović, C. Chauve, J. Salse, and E. Tannier. Mapping ancestral genomes with massive gene loss: A matrix sandwich problem. *Bioinformatics*, 27(13):i257–i265, 2011.

[17] Michel Habib, Ross McConnell, Christophe Paul, and Laurent Viennot. Lex-bfs and partition refinement, with applications to transitive orientation, interval graph recognition and consecutive ones testing. *Theoretical Computer Science*, 234(1–2):59–84, 2000.

[18] J. Håstad, L. Ivansson, and J. Lagergren. Fitting points on the real line and its application to rh mapping. *Journal of Algorithms*, 49(1):42–62, 2003.

[19] Wen-Lian Hsu and Ross M. McConnell. PC trees and circular-ones arrangements. *Theoret. Comput. Sci.*, 296(1):99–116, 2003. Computing and combinatorics (Guilin, 2001).

[20] Bradley R. Jones, Ashok Rajaraman, Eric Tannier, and Cedric Chauve. Anges: Reconstructing ancestral genomes maps. *Bioinformatics*, 2012.

[21] Ján Maňuch and Murray Patterson. The complexity of the gapped consecutive-ones property problem for matrices of bounded maximum degree. *J. Comput. Biol.*, 18(9):1243–1253, 2011.

[22] R.M. McConnell. A certifying algorithm for the consecutive-ones property. In *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 768–777. Society for Industrial and Applied Mathematics, 2004.

[23] Bojan Mohar. The Laplacian spectrum of graphs. In *Graph theory, combinatorics, and applications. Vol. 2 (Kalamazoo, MI, 1988)*, Wiley-Intersci. Publ., pages 871–898. Wiley, New York, 1991.

[24] J. Romiguier, V. Ranwez, E.J.P. Douzery, and N. Galtier. Contrasting gc-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome research*, 20(8):1001–1009, 2010.

[25] James B. Saxe. Embeddability of graphs into k-space is strongly np-hard. In *Allerton Conserence in Communication, Constrol and Computing*, pages 480–489. 1979.

[26] Thomas Schmidt and Jens Stoye. Quadratic time algorithms for finding common intervals in two and more sequences. In *In Proceedings of the 15th Annual Symposium on Combinatorial Pattern Matching, CPM 2004, volume 3109 of LNCS*, pages 347–358. Springer, 2004.

[27] A. Tucker. A structure theorem for the consecutive 1's property. *Journal of Combinatorial Theory, Series B*, 12(2):153–162, 1972.

[28] N. Vuokko. Consecutive ones property and spectral ordering. In *Proceedings of the 10th SIAM International Conference on Data Mining (SDM10)*, pages 350–360, 2010.