Ordering copy number alteration data to analyze colorectal cancer progression

Iuliana M. Bocicor[1]✉, Giulio Caravagna[2], Alex Graudenzi[2], Claudia Cava[2], Giancarlo Mauri[2], Marco Antoniotti[2]

[1]Department of Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania
[2]Department of Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy

Motivation and Objectives

Cancer is a very complex disease and understanding its dynamics and evolution is one of the challenges of modern biosciences. As most available data on cancer is static, extracting dynamic information about its progression from "static" biological data would have a major significance.

We are approaching the Temporal Ordering Reconstruction (TOR) problem, that is the sorting of a collection of multi-dimensional biological data to reflect an accurate temporal progression of the target disease.

The most general form of the TOR problem has been studied from many points of view. Firstly, the TOR problem, as defined above has been tackled mostly in two works, which use **gene expression** data as the "raw'" data in the samples (Gupta and Bar-Joseph, 2008; Magwene et al., 2003). Secondly, another series of works start by analyzing comparative genomic hybridization data to build a plausible tree of possible gene mutation events and continue towards a use of Bayesian models to assess pathways variations in a disease (Desper et al., 1999; Pathare et al., 2009; Gerstung et al., 2011; Beerenwinkel et al., 2005).

Our work is more focused on a specific approach to the TOR problem, previously proposed by Gupta and Bar-Joseph (Gupta and Bar-Joseph, 2008), which has been shown to work for gene expression data and we develop a methodology which enables us to apply this technique on a **Copy Number Alterations (CNAs)** data set. We also aim to provide a building block in an analysis pipeline that can be used to look at temporal reconstruction problems that assume an already (partially) ordered dataset (Ramakrishnan, 2010; Antoniotti, 2010).

Methods

The technique presented by Gupta and Bar-Joseph (Gupta and Bar-Joseph, 2008) is based on the reduction of the sorting problem to the **Travelling Salesman Problem** (TSP), under two biologically realistic assumptions over the gene expression data set. As we can assume that the CNAs data also fulfils these two assumptions, we develop a methodology which enables us to apply the technique on a CNAs data set.

In order to capture distinct aspects of the complex CNAs phenomenon, we define several chromosome-related measures and certain filters targeting significant portions of chromosomes. We also aim to identify which of these measures performs best regarding tumour progression or whether chromosomal gains (amplifications) or losses (deletions), considered separately, could influence the outcome.

As chromosome measures, we introduce the following notions: value, intensity, number and the averaged analogous: average of the values and average of the intensities, all these referring to alterations, deletions and amplifications. Furthermore, we propose two filtering methods to be applied on the initial data set, which could lead us towards obtaining more accurate orderings:

- recurrent CNAs - we consider those CNAs that belong to regions of the chromosomes that have suffered alterations in a higher number of different samples;
- recurrent CNAs, as well as CNAs belonging to regions that include at least one of the genes known to be involved in tumor progression (**cancer driver genes**).

In order to build the TSP instance, we consider the cities to be represented by the 22-dimensional samples (each dimension corresponding to one chromosome, not considering the gender-linked chromosome) and a distance matrix is used to define distances between any two samples. Two types of metrics are used: the L_1 **distance** and the **Euclidean distance**.

Figure 1 briefly illustrates our methodology, highlighting the most important steps that were

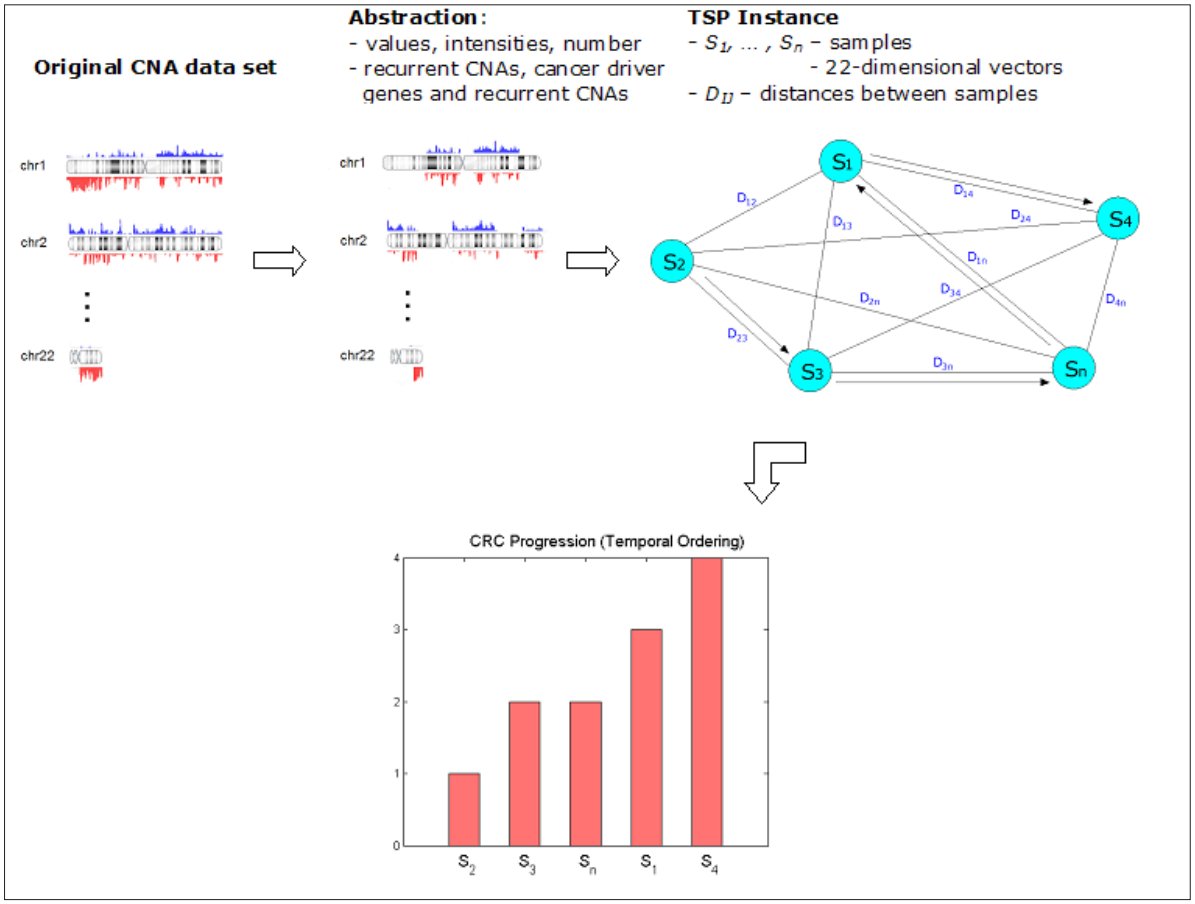

Figure 1: Representation of the proposed methodology. Starting from the input data set, different subsets are defined, using abstraction mechanisms. A TSP instance is built for each new data set and finally, the solution to the TSP represents the temporal ordering of the given samples.

used to determine a temporal ordering for a set of biological samples.

We tested the algorithm on a CNAs data set (Reid et al., 2009), consisting of 44 samples, in different stages of colorectal cancer (CRC). Three types of tests were made, one for the initial input data set and two for the subsets obtained by applying the above mentioned filters, therefore obtaining several different orderings. As a validation criterion, we used the survival time of each patient, after being diagnosed. We defined the "ideal ordering'" as the one in which the first sample has the maximum, while the last one has the minimum overall survival time. Using the **Squared Deviation Distance (SDD)**, the distance from each obtained solution to the ideal one was computed. Therefore, the orderings having smaller SDDs (with regard to the ideal ordering) were considered to be more accurate.

Results and Discussion

Results show that the test in which recurrent CNAs are used in conjunction with CNAs belonging to cancer driver genes produces the highest similarities with respect to the ideal ordering, i.e., the lowest values of the SDD, in terms of minimum value. Therefore, the more filters we apply on the input data set, the closer the minimum obtained orderings are, with respect to the ideal one. This clearly outlines the importance of combining biological knowledge with mathematical techniques to achieve significant results.

The best result was obtained for the test that takes into account the CRC driver genes and recurrent CNAs, with the chromosome measure average of values of alterations and for the L_1 distance. The samples in the first half of this ordering belong to patients who have (on average) significantly higher survival times than those in the last half. Although in our data set the CRC histologi-

cal stages are not always directly correlated to the survival time, we observe that the best ordering is also compatible with the CRC stages, to a certain degree.

Concerning the other chromosome measures, we have noticed that, on average, in the case of values and intensities, amplifications and deletions, considered separately, induce a better ordering than all alterations; in the case of the number, deletions seemed to be more relevant, when considering the L_1 distance, while for the Euclidean distance, all alterations inferred orders with lower SDDs. For the averaged values and intensities, all alterations have proven to be more important than either gains or losses, when considering the L_1 distance. On average, the orderings obtained using the L_1 distance are more accurate, compared to those using the Euclidean distance.

We have presented a particular solution for the temporal ordering reconstruction problem. We have built our approach on a previously proposed solution (Gupta and Bar-Joseph, 2008), by adapting it to chromosomal CNA data and we tested it on a CRC data set. To the best of our knowledge, our work is the first to adapt the TSP approach to the TOR problem, in conjunction with CNA data.

Acknowledgements

The authors would like to thank Manuela Gariboldi for providing the colorectal cancer data set. We also acknowledge Regione Lombardia (project RetroNet, grant 12-4-5148000-40; U.A 053) and NEDD for financial support of this work. The work was also possible with the financial support of the Sectoral Operational Programme for Human Resources Development 2007-2013, co-financed by the European Social Fund, under the project number POSDRU/107/1.5/S/76841 with the title "Modern Doctoral Studies: Internationalization and Interdisciplinarity".

References

1. Antoniotti M, Carreras M, Farinaccio A, Mauri G, Merico D et al. (2010) An Application of Kernel Methods to Gene Cluster Temporal Meta-Analysis. Comput Oper Res 37(8): 1361-1368. doi: 10.1016/j.cor.2009.03.011
2. Beerenwinkel N, Rahnenfuhrer J, Daumer M, Hoffman D, Kaiser R et al. (2005) Learning multiple evolutionary pathways from cross-sectional data. J Comput Biol 12: 584-598. doi: 10.1089/cmb.2005.12.584
3. Desper L, Jiang F, Kallioniemi OP, Moch H, Papadimitriou CH et al. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. J Comput Biol 6(1): 37-51.
4. Gerstung M, Eriksson N, Lin J, Volgestein B, Beerenwinkel N. (2011) The temporal order of genetic and pathway alterations in tumorigenesis. PLoS ONE 6(11): 1-9. doi: 10.1371/journal.pone.0027136
5. Gupta A and Bar-Joseph Z. (2008) Extracting dynamics from static cancer expression data. IEEE/ACM Trans Comput Biol Bioinform 5:172-182. doi: 10.1109/TCBB.2007.70233
6. Magwene PM, Lizardi P, Kim J. (2003) Reconstructing the temporal ordering of biological samples using microarray data. Bioinformatics 19(7):842-850. doi: 10.1093/bioinformatics/btg081
7. Pathare S, Schaffer AA, Beerenwinkel N, Mahimkar M. (2009) Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression. Int J Cancer 124(12): 2864-2871. doi: 10.1002/ijc.24267
8. Ramakrishnan N, Tadepalli S, Watson LT, Helm RF, Antoniotti M et al. (2010) Reverse engineering dynamic temporal models of biological processes and their relationships. PNAS 107: 12511-12516. doi: 10.1073/pnas.1006283107
9. Reid JF, Gariboldi M, et al. (2009) Integrative approach for prioritizing cancer genes in sporadic colon cancer. Genes, Chromosomes and Cancer 48: 953-962. doi: 10.1002/gcc.20697