Chromosome instability for tumor progression inference

Claudia Cava[1]✉*, Italo Zoppis[1]*, Manuela Gariboldi[2,3], James F. Reid[2,3], Marco Antoniotti[1], Giancarlo Mauri[1]

[1]Department of Informatics, Systems and Communication, University of Milan Bicocca, Milan, Italy
[2]Department of Experimental Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[3]Molecular Genetics of Cancer, FIRC Institute of Molecular Oncology Foundation, Milan, Italy

* These authors have contributed equally to this work.

Motivation and Objectives

The development and progression of Colorectal Cancer (CRC) - as for most other solid cancers, is a multi-step process leading to the accumulation of chromosomal instability (CIN) that occurs over the lifetime of a tumor (Shen et al, 2007; Vogelstein et al, 1988; Fearon et al, 1990). CINs include DNA copy number alterations (CNAs), i.e., regions of aberrantly increased or decreased DNA. For this reason, it is a challenge to identify both the regions that have CNAs and the genes whose expression could be deregulated (i.e., increased or decreased) because of gain or loss of their copies.

In this paper we focus on the role of copy number alteration in assessing prognosis of patients with CRC. Specifically, we show that the inference of the CRC progression benefits from exploiting CNA information when a particular relational representation of patients is given. The employed framework outperforms standard approaches where patients are represented through a set of available attributes. Documentation and software are available at http://bimib.disco.unimib.it/people/claudia.cava/soft. The data set for this analysis was provided by IRCCS Istituto Nazionale dei Tumori Milano (INT) and deposited at NCBI Gene Expression Omnibus (GSE16125).

Methods

Tissue specimens from 53 consecutive sporadic CRCs were obtained from previously untreated patients lacking family history and high-frequency microsatellite instability (MSI) who underwent surgical resection at the "Fondazione IRCCS Istituto Nazionale dei Tumori" (INT) Milano, between 1998 and 2000. After surgery all patients continued to be treated in INT, where their clinical course was constantly recorded. Tumor specimens containing more than 70% neoplastic cells and their surrounding normal mucosa were selected by an experienced pathologist from cryopreserved tissue and used in a previous study of genetic features associated to colorectal carcinogenesis (Frattini et al, 2004). Microarray production was done following standard protocols by AROS Applied Biotechnology AS (Aarhus, Denmark). 51 DNA samples were hybridized to Affymetrix GeneChipVR Human Mapping 250K NspI (SNP arrays). Raw intensity CEL files of the SNP arrays were processed with CNAG program v.2.0 (Copy-Number Analysis for Affymetrix GeneChips; Santa Clara, CA (Nannya et al, 2005) to detect chromosomal CNAs . Some samples were excluded due to poor quality hybridizations and unknown stage tumor progression (Reid et al, 2009). Also, stage-I patients were excluded because of the lack of instances in the considered data set. The selected cohort can be finally summarized as follows: 10 type-II patients, 10 type-III patients and 23 type-IV patients.

In order to quantify relationships between patients expressing the CCR progression, we first define a dissimilarity function over both an "advanced-stage" patient group and a specific "representative" base group e.g., patients with the lowest stage ("prototype"), then we classify patients according to the induced representations. In other words, the considered dissimilarity values quantify, by construction, subject differences due to different CNA information belonging to each subject. While in a "standard" case-control classification subjects are discriminated on its own set of attribute values, the dissimilarity function D(fx, fy) is given through an estimation of the difference between the obtained CNA mean value frequency distributions fx and fy. In order to give a definition for D which can express dissimilarity between any pair of patients x and y (based on the CNA mean value frequency distribution fx and fy), we employ the symmetrised Kullback-Leibler (KL) divergence (Cover et al, 1991) between any pair of distribution fx and fy.

Table 1: a) Performances for the Standard Representation. b) Performances for the Dissimilarity Representation.

Test	Sensitivity	Specificfy	PPV	NPV	Accuracy
a)					
stage II vs stage III	90,00%	80,00%	81.81%	88.88%	85,00%
stage II vs stage IV	100,00%	30,00%	76.66%	100,00%	78.79%
stage III vs stage IV	95.65%	20,00%	73.33%	66.66%	72.72%
b)					
stage II vs stage III	100,00%	90,00%	90.91%	100,00%	95,00%
stage II vs stage IV	91.30%	80,00%	91.30%	80,00%	87.88%
stage III vs stage IV	91.30%	80,00%	91.30%	80,00%	87.88%

Results and Discussion

The first issue of our investigation was to check the capability, for a given standard approach, of distinguishing patient groups. For this, we considered the following case - control study: (i) stage II (as control group) vs stage III (as case group); (ii) stage III (as control group) vs stage IV (as case group); (iii) stage II (as control group) vs stage IV (as case group).

All our evaluations employ a class of algorithms widely used in the machine learning community (i.e., the Support Vector Machine (Cristianini et al, 2000) within a k-fold cross-validation process. For the "standard" case, SVMs are given (input) matrices where patients are represented through the sequence of chromosomes as attributes, and each i-th component of the sequence is given by the CNA mean value associated to the chromosome i. Moreover, all experiments are evaluated by standard indices which are broadly applied in this context to measure the precision and recall capability of an inference system; i.e., sensitivity, specificity, positive (PPV) and negative predictive values (NPV), see for example (Davis et al, 2006). Table 1a) shows the performances when the classifiers are applied to the standard representations as discussed above. The standard approach is not able to discriminate both stage III from stage-IV patients (20% specificity) and stage II from stage-IV patients (30% specificity). On this basis, we used CNAs information to represent patients through dissimilarities as reported above. Table 1b) reports the inference performance when the dissimilarity representation is applied. We obtained substantially better accuracies reporting higher values of performances (>=80%) for the whole set of the applied indices.

We showed that even a prediction analysis, concerning the progression of CRC, as characterized by the given staging classification system (Duke), benefits from exploiting CNA information when a specific representation of patients is considered. We point out that, in this work, the choice of a dissimilarity representation (i.e., the KL-divergence) has been addressed to obtain a function providing an estimation of the difference between the obtained CNA mean value frequency distributions for each pair of patients. More specific measures may be tested in future analysis.

Interesting questions on these arguments are reported in (Pekalska et al, 2005). Also the choice of a correct prototype set can be critical in this approach, and may change the results being investigated. This is another question which we are immediately interested for future analysis. Here we did not study the best possible prototype set, instead, the rationale for our choice was simply to employ a group of patients with a presumably low number of accumulated alterations. Numerical experiments indicate that the application of the applied representation for the considered data provide high precision and recall performances outperforming typical standard approaches where patients are represented through their set of available attributes. These results clearly suggest broader investigations either on different data sets or different CRC staging classification systems (Horton et al, 2005).

References

1. Cover T M, et al. (1991) Elements of information theory New York, NY, USA: Wiley-Interscience.
2. Cristianini N et al (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.

3. Davis J et al (2006) The relationship between precision-recall and roc curves ICML '06: Proceedings of the 23rd international conference on Machine learning. New York, NY, USA: ACM, 233-240.
4. Fearon, E. and Vogelstein, B. (1990). Genetic model for colorectal tumorigenesis. Cell, 61:759–767.
5. Frattini M, Balestra D, et al. (2004) Different Genetic Features Associated with Colon and Rectal Carcinogenesis Clin Cancer Res 10(12):4015-4021.
6. Horton J K et al (2005) Staging of colorectal cancer: past, present, and future. Clin Colorectal Cancer, 4(5):302-12
7. Nannya Y, Sanada K (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Research 65(14): 6071-6079.
8. Pekalska E et al (2005) The Dissimilarity Representation for Pattern Recognition: Foundations And ekalska E et al (2005) The Dissimilarity Representation for Pattern Recognition: Foundations And Applications Machine Perception and Artificial Intelligence. World Scientific Publishing Company.
9. Reid J F, Gariboldi M, et al. (2009) Integrative approach for prioritizing cancer genes in sporadic colon cancer Genes Chromosom. Cancer 48(11):953-962
10. Shen, L., Toyota, M., et al(2007). Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proceedings of the National Academy of Sciences, 104(47):18654-18659.
11. Vogelstein, B., Fearon, E.,et al. (1988). Genetic alterations during colorectal-tumor development. N. Engl. J. Med., 319:3526-3535.