Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Classification of cancers based on copy number variation landscapes

Ning Zhang [a,1], Meng Wang [b,1], Peiwei Zhang [b], Tao Huang [b,*]

[a] *Department of Biomedical Engineering, Tianjin Key Lab of Biomedical Engineering Measurement, Tianjin University, Tianjin, PR China*
[b] *Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, PR China*

ARTICLE INFO

Article history:
Received 26 February 2016
Received in revised form 25 May 2016
Accepted 1 June 2016
Available online xxxx

Keywords:
DNA copy number variants
Cancer type
Sequential minimal optimization
Feature selection
IFS

ABSTRACT

Genomic alterations in DNA can cause human cancer. DNA copy number variants (CNV), as one of the types of DNA mutations, have been considered to be associated with various human cancers. CNVs vary in size from 1 bp up to one complete chromosome arm. In order to understand the difference between different human cancers on CNVs, in this study, we developed a method to computationally classify six human cancer types by using only CNV level values. The CNVs of 23,082 genes were used as features to construct the classifier. Then the features are carefully selected by mRMR (minimum Redundancy Maximum Relevance Feature Selection) and IFS (Incremental Feature Selection) methods. An accuracy of over 0.75 was reached by using only the CNVs of 19 genes based on Dagging method in 10-fold cross validation. It was indicated that these 19 genes may play important roles in differentiating cancer types. We also analyzed the biological functions of several top genes within the 19 gene list. The statistical results and biological analysis of these genes from this work might further help understand different human cancer types and provide guidance for related validation experiments. This article is part of a Special Issue entitled "System Genetics" Guest Editor: Dr. Yudong Cai and Dr. Tao Huang.

1. Introduction

Copy number variants (CNV) are duplications and deletions of chromosomal segments, from small deletions or insertions (1–50 bp) to huge chromosomal aberrations (>1 Mbp) [1]. It is known that up to 12% of human genome is subject to it [2]. Up to now, CNVs have been widely reported in human populations. The widespread changes in gene copy numbers among human individuals are associated with their fast formation rates and cellular stress. The following introduces current knowledge about the elusive landscape of CNVs in its mutation pattern, distribution and correlation with human population and diseases [1], in order to illustrate that CNVs could be used to classify varying types of cancers.

The mapping of the CNV landscape among general population is seldom done. One study based on SNP microarray data from 2500 individuals with no known disease showed that a CNV that is at least 100 kb is approximately harbored by 65% to 80% of individuals [3]. A normal misconception is that CNVs are mainly inherited, and copy number variants in individuals during their life time could be dismissed. Admittedly, fixed CNVs account for large parts of total CNVs, and are crucial factors to distinguish us with other primate lineages. However, both somatical and meiotical changes in organs and tissues are also significant based on experiments on identical twins [4]. As for the frequencies of copy number variants, 10^{-6}–$5 * 10^{-5}$ is suggested by the analysis of sperm cells [5], and over 10^{-6} is estimated in blood experiments [6]. Compared to point mutations, at least several orders of magnitude higher are suggested in variants in copy numbers.

Currently, researches have shown that many human diseases involve copy number variants (CNV) that could alter the diploid status of particular locus of the genome. For example, neurodevelopmental diseases, such as intellectual disability, autism, and schizophrenia, researches have shown that CNV at least account for approximately 15% of these diseases [1]. Reasons for the relevance between CNVs and neurodevelopmental diseases could be the perturbation of gene pathways involving in neuron development [7]. Several neurodevelopmental relevant genes, such as A2BP1, IMMP2L, and AUTS2, are reported with mutational CNVs [7].

Cancer formation and progression are also associated with change in copy number [8]. Based on the oncogenetic tree analysis of CNVs in breast cancer, Li et al. found that the genetic alteration of ErbB2 occurs early in breast cancer and the CNVs of AKT2, PTEN, CCND1, RAS, and PIK3CA are late events [9]. The is partly because of the relevance between cellular stress and CNVs, since copy number change is easily happened due to the fact that stress, such as hypoxia, might switch repair of broken replication from homologous recombination to non-homologous repair [2]. Ample evidence has proven that individuals with some CNVs might show cancer prone [10]. Research has revealed about 140 tumor driver genes, and a common tumor might contain two to eight driver gene mutations [11].

[*] Corresponding author.
E-mail addresses: zhni@tju.edu.cn (N. Zhang), mengwang@sibs.ac.cn (M. Wang), zhangpeiwei@sibs.ac.cn (P. Zhang), tohuangtao@126.com (T. Huang).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.bbagen.2016.06.003
0304-4165/© 2016 Published by Elsevier B.V.

Despite the rapid findings of de novo CNVs, the associations between CNVs and different types of cancer still remain elusive and difficult to be illuminated. Previous studies have shown that the CNV pattern between lung adenocarcinoma and squamous cell carcinoma are very different [12] and there are rich mutations and copy number changes in various cancers [13]. Here, we developed a computational method, which successfully uses the information of gene CNV levels to classify six types of cancers — breast adenocarcinoma (BRCA), colon and rectal carcinoma (COAD/READ), glioblastoma multiforme (GBM), kidney renal clear cell carcinoma (KIRC), ovarian serous carcinoma (OV), and uterine corpus endometrial carcinoma (UCEC). A Dagging method in 10-fold cross validation is used, and 19 genes are finally chosen to differentiate cancer types with the accuracy 0.751. Experiment evidences on some of these 19 genes are given to further support our findings. We concluded that CNVs of these 19 genes play crucial roles in the distinguishing these six cancer types.

2. Data and method

2.1. Dataset

The data of copy number variants (CNV) in different types of cancers were downloaded from the cBioPortal for Cancer Genomics (http://cbio.mskcc.org/cancergenomics/pancan_tcga/, Release 2/4/2013) [13–15]. The CNV values were discretized into 5 different values in the database, with “−2” denoting a deep loss (possibly a homozygous deletion), “−1” presenting a shallow loss (possibly heterozygous deletion), “0” for diploid, “1” for a low-level gain, and “2” for a high-level amplification. Each sample has the CNVs of 24,174 genes.

11 cancer types were provided in the cBioPortal database. However, since the sample number of several cancer types were below 400, in this study, we only downloaded six cancer types of them with samples more than 400. The six cancer types and the sample numbers were listed in Table 1. Totally there were 3480 samples in the six cancer types.

2.2. Feature selection

2.2.1. The mRMR method

The 24,174 gene CNV level values were regarded as features for the cancer type classification, constructing one 24,174-dimensional vector for one sample. The mRMR (minimal-redundancy-maximal-relevance) criterion [17] was used to rank the importance of the 24,174 features. The method rank features according to both their relevance to the problem and the feature redundancies. If a ranked feature had a smaller index, it would have a better trade-off between the maximum relevance and minimum redundancy.

We define mutual information (*MI*) as:

$$I(x,y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{1}$$

where x, y are two vectors, $p(x,y)$ is the joint probabilistic density, $p(x)$ and $p(y)$ are the marginal probabilistic densities. *MI* is used to quantify both the relevance and the redundancy.

Suppose Ω denotes the entire space containing all the aforementioned 10,092 feature components, Ωs denotes the already-selected feature set containing mm features, and Ω_t denotes the to-be-selected feature set containing nn features. The relevance D between the feature f in Ω_t and the target c can be calculated by

$$D = I(f,c) \tag{2}$$

The redundancy R between the feature f in Ω_t and all the features in Ω_s can be calculated by

$$R = \frac{1}{mm}\sum_{f_i\in\Omega_s} I(f,fi) \tag{3}$$

To get the feature f_j in Ω_t with the maximum relevance and the minimum redundancy, let us combine Eq. (5) with Eq. (6), as formulated by

$$\max_{f_j\in\Omega_t}\left[I\left(f_j,c\right)-\frac{1}{mm}\sum_{f_i\in\Omega_s} I(f_j,f_i)\right](j=1,2,\ldots,nn) \tag{4}$$

The evaluation will continue for 24,174 rounds in this study because our feature set contains 24,174 ($=mm + nn$) features. After these evaluations, a feature set S can be obtained:

$$S=\left\{f'_1,f'_2,\ldots,f'_h,\ldots,f'_N\right\} \tag{5}$$

where the subscript index of each feature in S indicates at which round the feature is selected. If a feature has been selected earlier, it will have a smaller index and could be a better feature.

However, to reduce the computational time, we only compute and retrieve the top 200 features in the ranked feature list in this study, not ranking all the 24,174 features.

2.2.2. Incremental feature selection (IFS)

To determine the optimal feature set, the Incremental Feature Selection (IFS) [18–22] method was applied. Features in the ranked list by mRMR were added one by one from higher to lower rank. When 1 more feature had been added, a new feature set was constructed. For each of the feature set, a classifier was constructed and examined. Totally, 200 classifiers were constructed since we only examine the top 200 features in the ranked mRMR feature table. The performances of all the 200 classifiers were evaluated. And finally we chose the one classifier as the final one that yielded the first stable high performance at changing point of the IFS curve. The corresponding feature set that the final classifier used was regarded as the selected feature set.

2.3. Classification algorithms

The classification algorithms used in this study was Dagging. The Dagging algorithm [23] is a meta classifier. It creates a number of disjointed and stratified folds from the data. Then, it feeds each chunk of data into a base single learning algorithm. Each base algorithm provides an output for a given sample. The final classification is made based on the most votes obtained using the base classifiers. We employed Weka 3.6.4 [24] to construct the cancer type classification models with default parameters.

Table 1
The number of samples in the 6 cancer types in our dataset.

Type index	Cancer type	Samples
1	BRCA (Breast invasive carcinoma)	847
2	COAD/READ (Colon adenocarcinoma/Rectum adenocarcinoma)	575
3	GBM (Glioblastoma multiforme)	563
4	KIRC (Kidney renal clear cell carcinoma)	490
5	OV (Ovarian serous cystadenocarcinoma)	562
6	UCEC (Uterine corpus endometrioid carcinoma)	443
Total		3480