Mixture Modeling of DNA Copy Number Amplification Patterns in Cancer

Jarkko Tikka[1], Jaakko Hollmén[1], and Samuel Myllykangas[2]

[1] Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, FI-02015 TKK, Espoo, Finland
tikka@mail.cis.hut.fi, Jaakko.Hollmen@tkk.fi
[2] Department of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland Samuel.Myllykangas@helsinki.fi

Abstract. DNA copy number amplifications are hallmarks of many cancers. In this work we analyzed data of genome-wide DNA copy number amplifications collected from more than 4500 neoplasm cases. Based on the 0-1 representation of the data, we trained finite mixtures of multivariate Bernoulli distributions using the EM algorithm to describe the inherent structure in the data. The resulting component distributions of the mixtures of Bernoulli distributions yielded plausible and localized amplification patterns. Individual amplification patterns were tested for their role in cancer groups formed with known risk associations. Our detailed analysis of chromosome 1 showed that asbestos-exposure related and hormonal imbalance-associated cancers were clustered and specific chromosome bands, 1p34 and 1q42, were identified. These sites contain cancer genes, which might explain the condition-specific selection of these loci for amplification.

1 Introduction

Cancer is a complex disease, which results from changes in the genome. DNA copy number amplifications have an important role in cancer progression and amplified genes are appealing targets in clinical applications (therapy and prognosis). To probe further into the backgrounds of amplification patterns, we investigated a data collection of DNA copy number amplifications by probabilistic analysis. Mixture modeling was applied to reveal the natural structure in DNA copy number amplification patterns. Here, we demonstrate that analysis of chromosome-specific copy number data provides intuitive profiles and reveals specific sites of amplification. Based on molecular properties of cancers instead of clinical data, mixture modeling provides a practical setting for data mining of various factors related to manifestation, mechanisms, and specificity of DNA copy number amplifications in cancer. Specific DNA copy number amplification patterns for aetiological factors, namely asbestos-exposure and hormonal-imbalance, were identified in chromosome 1.

2 DNA Copy Number Amplifications in Cancer

DNA copy number amplifications are essential hallmarks of cancer. Amplifications are localized chromosomal aberrations, which result in copy number increase. Normal human genome contains two copies of the same chromosomal region and an amplification increases that number to at least five. A high-level amplification may contain hundreds of copies. The average size of an amplified region is 20 Mbp (million base pairs). Amplifications manifest as homogeneously staining regions (HSR) or extra-chromosomal DNA bodies, double minutes (DMs), and episomes. HSRs are ladder-like structures of inverted DNA stretch repeats in the chromosomes, and DMs and episomes are circular DNA molecules that have looped out from the chromosome. HSRs and DMs can be detected using microscopy techniques, whereas episomes can only be identified using molecular biology methods.

Comparative genomic hybridization (CGH) is a technique that can be used to screen DNA copy number changes throughout the entire genome [1]. In CGH, tumor DNA and reference DNA are differentially labeled using two different fluorochromes and hybridized on normal chromosome spreads. Chromosome regions with aberrant copy number are detected as changes in the intensity ratios of the two fluorochromes along the probe chromosomes. Microscope techniques that are used to detect hybridization intensities limit the resolution of conventional CGH to chromosome bands (roughly 400 specific chromosomal domains). By using specific staining methods, sub-banding can be visualized and the amount of bands increased to over 800.

We have previously profiled DNA copy number amplifications in human neoplasms [2]. The analysis was based on CGH data of amplifications (with resolution of 393 chromosome bands) that were collected from the literature between 1992 and 2002. The data collection covers 838 CGH publications with a total of 23 284 reviewed cases, 4590 cases with amplifications [3]. In our previous analysis [2], we estimated genome-wide, cancer-specific amplification profiles and formed clusters of neoplasms according to genome-wide similarities of DNA copy number amplifications. Furthermore, we defined amplification hot spots using the independent component analysis (ICA). Profiling analysis revealed that human neoplasms formed clusters based on the amplification frequency profiles. The specificity of amplifications based on cell lineage suggested that amplification patterns might be applied in classification of tumors, in addition to such clinical parameters as histology, aetiology, and pathogenesis[2]. The novel approach in this work is to model the general patterns of DNA copy number amplifications in cancers with a mixture of multivariate Bernoulli distributions. Performed one chromosome at a time, the modeling yields generative amplification patterns expressed as a probability distribution. The collection of DNA copy number amplifications in human neoplasms was obtained from http://www.helsinki.fi/cmg/cgh_data.html. The original classification of 73 human neoplasms [2] was refined to contain 95 specific neoplasm types. The neoplasms were classified using the WHO classification of tumors with respect to their aetiology regarding asbestos-exposure and hormonal-imbalance. Asbestos-

exposure and hormonal aetiology were chosen as examples, because these factors exhibit significant differences in chromosome 1 show case. The neoplasms were further filtered down to 82 types of cancer by discarding non-malignant, benign, and border line tumors. Asbestos is a generic name for mineral fibers that are known to cause asbestosis and several lung diseases, especially tumors of the lung. Asbestos-exposure has been shown to cause DNA double strand breaks [4] that may initiate the amplification pathway. Hormonal-imbalance and exogenous hormones have been associated with increased risk of cancer. For example, hormonal contraceptives and postmenopausal estrogen therapy have been reported to increase the risk for breast cancer [5].

3 Mixture of Multivariate Bernoulli Distributions

The objective of probabilistic modeling is to estimate an unknown probability distribution based on a finite number of observations. The estimated probability distribution describes the process that has generated the data. Finite mixtures of distributions are a flexible method for modeling complex distributions [6,7]. The advantage of a mixture model is that its components can represent different parts of the true distribution, which would be impossible to estimate by a single parametric distribution. In this work, we concentrated on the mixtures of multivariate Bernoulli distributions, since the presentation of our DNA copy number amplification data was binary. Finite mixtures of multivariate Bernoulli distributions have been shown to be non-identifiable [8], i.e. different parameter sets result in equivalent distributions. The non-identifiability problem has been examined further in [9], where the class of models is shown to be still useful in practice.

The amplification data can be presented as binary vectors $\boldsymbol{x} \in \{0,1\}^d$, in which digit 1 denotes an amplification and digit 0 a non-amplified chromosomal structure. In the probabilistic modeling of amplification, the probabilities of the outcomes of observation $\boldsymbol{x} = (x_1, \ldots, x_d)$ are modeled as $\theta_i = P(x_i = 1), \quad i = 1, \ldots, d$. The probability of the observed vector $\boldsymbol{x}$ is estimated using multivariate Bernoulli distribution

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}. \tag{1}$$

The finite mixture of multivariate Bernoulli distributions is defined as

$$p(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{j=1}^{J} \pi_j p(\boldsymbol{x}|\boldsymbol{\theta}_j) = \sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_i}(1-\theta_{ji})^{1-x_i}, \tag{2}$$

where π_j are mixture proportions with the properties $\pi_j \geq 0$ and $\sum_{j=1}^{J} \pi_j = 1$. In the case of J components and d dimensions, the parameters of the mixture model are $\boldsymbol{\Theta} = \{J, \{\pi_j, \boldsymbol{\theta}_j\}_{j=1}^{J}\}$.

3.1 Expectation-Maximization (EM) Algorithm

Let us consider the case that the number of mixture components J is fixed and we have N observations $\boldsymbol{x}_n, n = 1, \ldots, N$. The log-likelihood of the parameters $\{\pi_j, \boldsymbol{\theta}_j\}_{j=1}^J$ can then be written as

$$l = \sum_{n=1}^{N} \log \left[\sum_{j=1}^{J} \pi_j \prod_{i=1}^{d} \theta_{ji}^{x_{ni}} (1 - \theta_{ji})^{1 - x_{ni}} \right]. \tag{3}$$

The maximum likelihood estimates of the parameters are obtained by maximizing (3). The optimization can be carried out using the EM algorithm [10,11,12]. The detailed derivation of the EM algorithm for the finite mixture of multivariate Bernoulli distributions can be found from [7]. We show only the updating equations of the parameters. In the E-step, the posterior probabilities

$$p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \boldsymbol{\Theta}^k) = \frac{\pi_j^k p(\boldsymbol{x}_n|\boldsymbol{\theta}_j^k)}{\sum_{j'=1}^{J} \pi_{j'}^k p(\boldsymbol{x}_n|\boldsymbol{\theta}_{j'}^k)} \tag{4}$$

are calculated, where $\boldsymbol{\pi}^k = \{\pi_1^k, \ldots, \pi_J^k\}$ and $\boldsymbol{\Theta}^k = \{\boldsymbol{\theta}_1^k, \ldots, \boldsymbol{\theta}_J^k\}$ are parameter estimates at the iteration k. Equation (4) gives the posterior probability that observation $\boldsymbol{x}_n$ is generated by the component j. In the M-step, the values of the parameters are updated as follows

$$\pi_j^{k+1} = \frac{1}{N} \sum_{n=1}^{N} p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \boldsymbol{\Theta}^k), \qquad \boldsymbol{\theta}_j^{k+1} = \frac{1}{N\pi_j^{k+1}} \sum_{n=1}^{N} p(j|\boldsymbol{x}_n, \boldsymbol{\pi}^k, \boldsymbol{\Theta}^k) \boldsymbol{x}_n.$$

The parameters of the component distributions are initialized from the uniform distribution $\theta_{ji} \sim \mathcal{U}(0.25; 0.75)$ and the starting values of mixture proportions are $\pi_j = 1/J$. The iteration between the E- and M-steps gives the monotonically increasing series of the values for the log-likelihood. In this work, the iteration is stopped when the relative change in the log-likelihood is smaller than 10^{-4}.

4 Experiments: Modeling of DNA Amplification Patterns

The data collection of DNA amplifications contains 4590 observed cases. For our analysis purposes, we chose to work with one chromosome at the time. This made our density estimation more viable by greatly reducing the data dimension. Chromosome 1 is the largest of all chromosomes with 2202 genes and it is divided to 28 bands; therefore our data dimension was $d = 28$. Our data consisted of $N = 446$ cases, which have amplification in chromosome 1, i.e. cases with no amplifications were discarded. A fair amount of amplifications in chromosome 1 have been reported in cancers [2], but a single dominant driving gene has not been identified. The cancer genes database [13] assigns to chromosome 1 a total of 39 genes that have somehow been implicated in cancer.

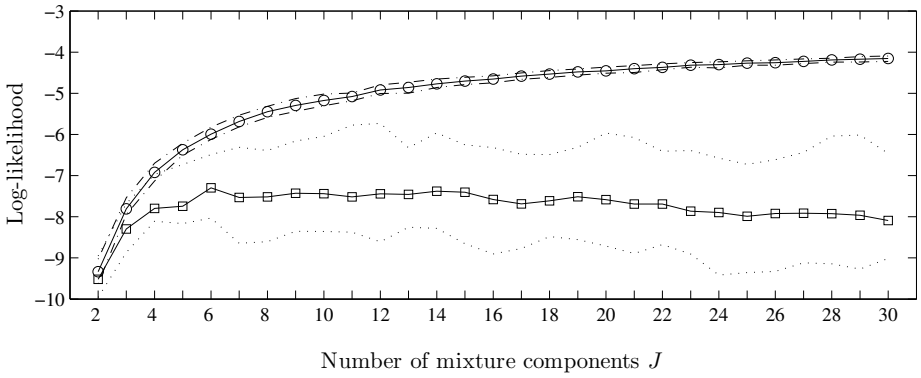

Fig. 1. The log-likelihoods for different training and validation sets for chromosome 1 as a function of the number of mixture components J. The mean of training log-likelihoods is marked with a solid line with circles, and a solid line with squares marks the mean of validation log-likelihoods. The interquartile range for the 50 training and validation runs is drawn with dash-dotted and dotted lines, respectively.

4.1 Model Selection on the Number of Component Distributions

In order to select a model with suitable complexity, the following cross-validation procedure was performed. The training of mixture models was repeated 10 times in a 5-fold cross-validation setting, varying the number of component distributions J from 2 to 30. Thus the setting produced 50 likelihood values for both the training sets and the validation sets for each model complexity. In all, 1450 mixture models were trained. Out of the 50 training and validation likelihoods for each complexity, the mean and the interquartile range were computed (Figure 1). The mean of the training likelihood is a smoothly increasing curve with increasing complexity, whereas the validation likelihood has more variance across the repeats. The model complexity can be chosen based on the mean of the validation likelihoods. For chromosome 1, six component distributions ($J = 6$) were chosen. The choice can also be motivated from the parsimony point of view: less complex models are preferred. In fact, additional components often produce identical component distributions within the mixture model. The final model was trained on all the data five times, out of which the best model was selected. This model was the basis for subsequent analysis.

4.2 Amplification Patterns in Chromosome 1

The values of parameters $\boldsymbol{\theta}_j$ of the jth component of the final mixture model are illustrated on the jth row in Figure 2. Darkness of the rectangles represents the value of the corresponding parameter θ_{ji}, so that darker rectangles indicate higher values. Figure 2 shows that the components of the mixture model clearly describe amplifications on different bands of chromosome 1. In four components

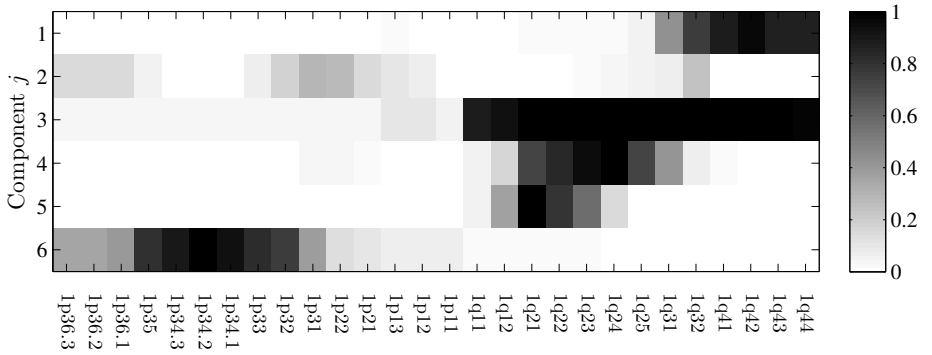

Fig. 2. Parameters $\theta_{ji}, j = 1, \ldots, 6,\ i = 1, \ldots, 28$ of the final mixture model. Each row shows the parameters of one component. The mixture proportions are $\pi_1 = 0.07, \pi_2 = 0.24, \pi_3 = 0.21, \pi_4 = 0.20, \pi_5 = 0.19$, and $\pi_6 = 0.09$. The names of the bands of chromosome 1 (corresponding to 28 variables) are shown under the x-axis.

(components 1 and 4-6), we can focus on the mode of the component distribution (the chromosomal band having the largest probability). Component 2 comprised rather low probabilities of amplification; component 3 embodies the amplification of the entire q-arm of chromosome 1 (1q11-1q44).

The data was divided into a risk group and a no-risk group according to asbestos-exposure related and hormonal-imbalance associated aetiologies. The hormonal risk and no hormonal risk groups consisted of 56 and 390 samples, respectively. The posterior probabilities were evaluated for each observation $\boldsymbol{x}_n$ according to Equation (4). The observations were classified to the components j according to the maximum posterior probability. The upper row of Figure 3 shows the proportions of observations coming from component j for the hormonal risk group (left) and for the no hormonal risk group (right). The lower row in Figure 3 shows the proportions for the asbestos risk group (left) and for the no asbestos risk group (right). The asbestos risk group and the no asbestos risk group include 37 and 409 samples, respectively.

To compare the differences in proportions of component j between the risk group and no-risk group, we performed a pooled proportions statistical test [14]. The null hypothesis was that the proportions are equal and the alternative hypothesis was that the proportions are unequal. In Figure 3, the components with statistical significant difference are marked by black bars. The difference is considered to be statistically significant if the p-value of the test is lower than 0.01. For instance, the difference in the proportions of the first component is statistically significant between the hormonal risk group and no hormonal risk group.

Component 1 was identified as enriched in cancers that were associated with hormonal-imbalance, and the empirical proportion of the asbestos-risk cases in component 6 in the asbestos risk group is significantly larger than that of the

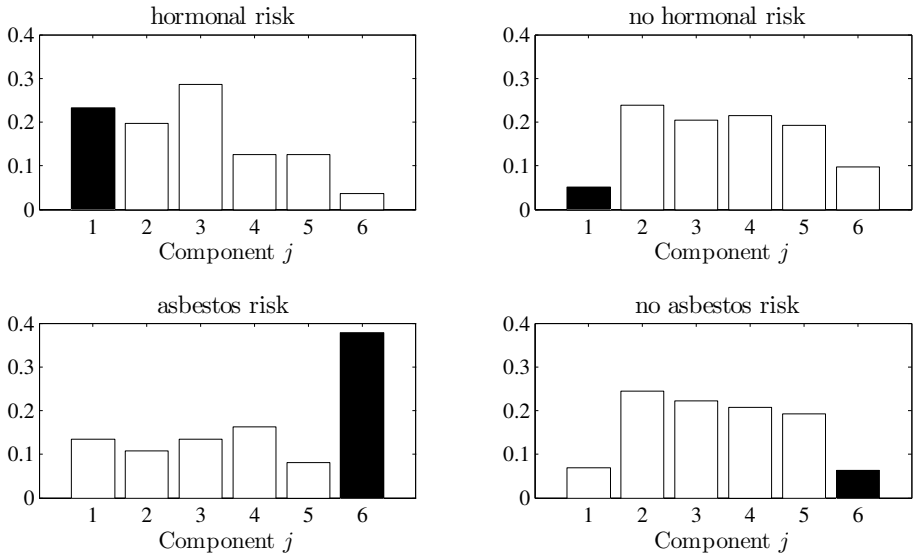

Fig. 3. The proportions of the observations coming from different components classified according to the posterior probabilities. Black bars denote that the difference of proportions in the jth component between the risk and no risk groups is statistically significant; the white bars indicate that the difference is not statistically significant.

non-risk group. The mode of component distribution 1 was 1q42, a region that contains the *FH* (fumarate hydratase) cancer gene. *FH* gene amplifications have not been reported in association with cancer, but heterozygous germ line mutations in the *FH* gene cause tumor predisposition syndrome known as hereditary leiomyomatosis and renal cell cancer (HLRCC). This syndrome is characterized by an increased risk of benign uterine leiomyomas, leiomyosarcoma, and renal cell carcinoma. At least leiomyosarcoma and leiomyoma are associated with hormonal-imbalance suggesting that the *FH* gene amplification could play a role in carcinogenesis induced by hormonal-imbalance. The mode of component distribution 6 was 1p34.2, containing the *MYCL1* oncogene. *MYCL1* has been identified to be amplified in small cell lung cancer and ovarian cancer [2]. Small cell lung cancer is related to asbestos-exposure, which might explain the finding.

5 Summary and Conclusions

We present a solution for probabilistic modeling of DNA copy number amplification data. Using the mixture modeling approach, we were able to provide plausible solutions for the DNA amplification patterns that are localized in the chromosome and can be interpreted with domain expertise. Furthermore, we investigated the role of the individual amplification patterns in their associa-

tion to cancers with known aetiological factors in a relation to a specific cancer. We found two amplification patterns to be statistically significant in cancers associated with hormonal-imbalance or asbestos-exposure. We believe that the amplification patterns will have many other uses as well as interpretations.

Acknowledgments

This work was supported by the Academy of Finland in the Research Program SYSBIO (Systems Biology and Bioinformatics), grant number 207469.

References

1. Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., Pinkel, D.: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science **258** (1992) 818–821
2. Myllykangas, S., Himberg, J., Böhling, T., Nagy, B., Hollmén, J., Knuutila, S.: DNA copy number amplification profiling of human neoplasms. Oncogene **25** (2006) 7324–7332
3. Knuutila, S., Autio, K., Aalto, Y.: Online access to CGH data of DNA sequence copy number changes. American Journal of Pathology **157** (2000) 689
4. Marczynski, B., Czuppon, A.B., Marek, W., Reichel, G., Baur, X.: Increased incidence of DNA double-strand breaks and anti-ds DNA antibodies in blood of workers occupationally exposed to asbestos. Human & Experimental Toxicology **13** (1994) 3–9
5. The International Agency for Research on Cancer. Post-menopausal oestrogen therapy in IARC monograph on the evaluation of carcinogenic risks to humans. In Hormonal contraception and post-menopausal hormonal therapy **72** (1999) 399–530
6. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics. John Wiley & Sons (2000)
7. Everitt, B.S., Hand, D.J.: Finite Mixture Distributions. Monographs on Applied Probability and Statistics. Chapman and Hall (1981)
8. Gyllenberg, M., Koski, T., Reilink, E., Verlaan, M.: Non-uniqueness in probabilistic numerical identification of bacteria. Journal of Applied Probability **31** (1994) 542–548
9. Carreira-Perpinan, M.A., Renals, S.: Practical identifiability of finite mixtures of multivariate Bernoulli distributions. Neural Computation **12** (2000) 141–152
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B **39** (1977) 1–38
11. Redner, R.A., Walker, H.F.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Review **26** (1984) 195–239
12. McLachlan, G.J.: The EM Algorithm and Extensions. Wiley & Sons (1996)
13. Higgins, M.E., Claremont, M., Major, J.E., Sander, C., Lash, A.E.: Cancergenes: a gene selection resource for cancer genome projects. Nucleic Acids Research **35** (2007) D721–726
14. Milton, J.S., Arnold, J.C.: Introduction to probability and statistics. McGraw-Hill second edition (1990)