Identifying Copy Number Variations based on Next Generation Sequencing Data by a Mixture of Poisson Model

Karin Schwarzbauer, Günter Klambauer, Andreas Mayr, and Sepp Hochreiter
Institute of Bioinformatics, Johannes Kepler University, Linz, Austria

Motivation: Next generation sequencing (NGS) technologies have profoundly impacted biological research and are becoming more and more popular due to cost effectiveness and their speed. NGS can be utilized to identify DNA structural variants, namely copy number variations (CNVs) which showed association with diseases like HIV, diabetes II, or cancer. There have been first approaches to detect CNVs in NGS data, where most of them detect a CNV by a significant difference of read counts within neighboring windows at the chromosome. However these methods suffer from systematical variations of the underlying read count distributions along the chromosome due to biological and technical noise. In contrast to these global methods, we locally model the read count distribution characteristics by a mixture of Poissons which allows to incorporate a linear dependence between copy numbers and read counts. Model selection is performed in a Bayesian framework by maximizing the posterior through an EM algorithm. We define a CNV call which indicates a deviation of the Poisson mixture parameters from the null hypothesis represented by the prior which is a model for constant copy number across the samples. A CNV call requires sufficient information in the data to push the model away from the null hypothesis given by the prior.

Results: We test our approach on the HapMap cohort where we rediscovered previously found CNVs which validates our approach. It is then tested on the tumor genome data set where we are able to considerably increase the detection while reducing the false discoveries.

The Model

- α_i : percentage of samples with copy number i
- λ: Poisson parameter for copy number 2
- For copy number i the Poisson parameter is (i λ) / 2

$$p_{\mathrm{mix}}(x \mid \boldsymbol{\alpha}, \lambda) = \sum_{i=0}^{n} \alpha_i \; p_{\mathrm{Poisson}}(x; \tfrac{i}{2}\lambda)$$

$$p_{\mathrm{Poisson}}(x; \lambda) = \tfrac{1}{x!} e^{-\lambda} \lambda^x$$

Dirichlet Prior on α

MOTIVATION

- predominantly locations with copy number 2 for all samples appear → null hypothesis
- Posterior can deviate from prior only by a high information content in the data
- CNV call possible as deviation from the prior (constant copy number across samples)

→Dirichlet prior $D(\boldsymbol{\alpha}^1; \boldsymbol{\gamma}) = b(\boldsymbol{\gamma}) \prod_{i=0}^{n} (\alpha_i)^{\gamma_i - 1}$

$$\boldsymbol{\alpha}^1 = (\alpha_1, \ldots, \alpha_n)$$

$$\alpha_i^p = E(\alpha_i) = \frac{\gamma_i}{\gamma_s}$$

$$\text{with } \gamma_s = \sum_{i=0}^{n} \gamma_i, \quad \gamma_2 \gg \gamma_i \text{ for i} \neq 2$$

- Thus α^p_2 is large and α^p_i is small for i ≠ 2.

Model Selection: EM Algorithm

The posterior for the i-th mixture component for sample x_k is

$$\hat{\alpha}_{ik} = \frac{\alpha_i \, (x; \frac{i}{2}\lambda)}{p(x; \lambda)}$$

The EM algorithm minimizes an upper bound on the negative log-posterior.

Resulting EM update rules:

$$\alpha_i^{\mathrm{new}} = \frac{\hat{\alpha}_i + \gamma_i - 1}{1 + \gamma_s - n} \qquad \lambda^{\mathrm{new}} = \frac{\frac{1}{N} \sum_{k=1}^{N} x_k}{\sum_{i=0}^{n} \hat{\alpha}_i \, \frac{i}{2}}$$

$$\hat{\alpha}_i = \frac{1}{N} \sum_{k=1}^{N} \hat{\alpha}_{ik}$$

CNV region calls

Information Gain of the max. posterior vs. max. prior

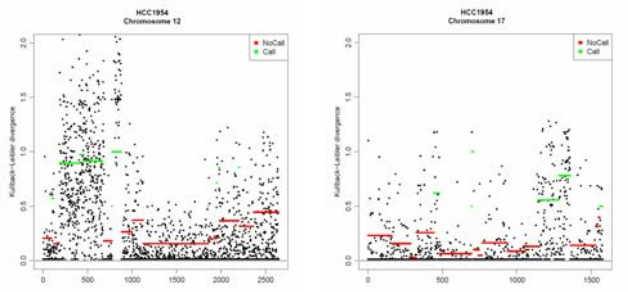
- Max. prior (ε →0) $\boldsymbol{\alpha}^p = (\epsilon^2, \epsilon, \alpha_3, \epsilon, \epsilon^2, \epsilon^3, \ldots, \epsilon^n) \quad \alpha_3 = 1 - \sum_{i \neq 2} \alpha_i^p$
- Measured via Kullback-Leibler divergence $KL(\boldsymbol{\alpha}^p \| \boldsymbol{\alpha}^{new}) = \sum_{i=0}^{n} \alpha_i^p log \frac{\alpha_i^p}{\alpha_i^{new}}$
- Alternative measures:
  - Mutual information $I(\boldsymbol{\alpha}^{new}, \boldsymbol{\alpha}^p) = \sum_{i=0}^{n} \alpha_i^p log \alpha_i^p - \sum_{i=0}^{n} \alpha_i^{new} log \alpha_i^{new}$
  - Entropy of alpha $H(\boldsymbol{\alpha}) = -\sum_{i=0}^{n} p(\alpha_i) log p(\alpha_i)$

CNV calls allow for FDR control
Detection power is increased

Experiments and Results

Experiment on Tumor Genomes

- Breast ductal carcinoma data set from Chiang et al. [2]
- Detected segment on tumor vs. normal
- False positives on "normal" cell lines



HCC1954	SegSeq	Chiang Calls*	Mixture of Poissons
Detected segments	691	194	314
False Positives	29	7	1

HCC1143	SegSeq	Chiang Calls*	Mixture of Poissons
Detected segments	407	126	205
False Positives	29	0	0

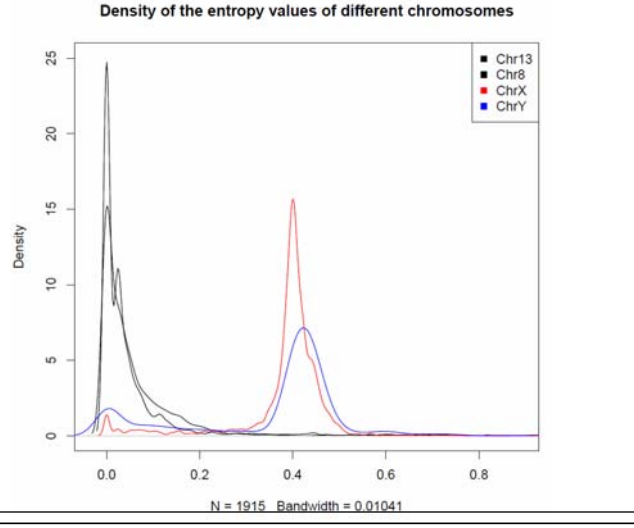
*Chiang calls with ratio y < 0.5 or y >1.5

Conclusion:

- Mixture of Poisson model reduces starting segments by a half without using labels
- False discovery rate is decreased
- Detection power is increased two fold
- Loss of only 5 of 194 called segments on HCC1954 / 5 of 126 on HCC1143

Experiment on HapMap Data

- NGS data from 46 HapMap samples
- Read counts for fixed windows of 50 kb for each sample
- Normalization: lane effects, multiple mappable reads
- A mixture of Poisson model was fitted for each 50kb window on the human genome.
- CNV region were called using a threshold on the Kullback Leibler divergence.
- Significant overlap with previously known CNV regions from [1]



References

[1] Conrad et al. Origins and functional impact of copy number variations in the human genome. *Nature* 464, 704-712, 2010.
[2] Chiang et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* 6(1),99-103, 2009.