Description of CLOSE/CLOSE-R models[*]

Xuefeng Wang[†]

1 Introduction

This document will briefly describe models and algorithms used in package CLOSE and its companion R package CLOSE-R. CLOSE is a suite of tools for CNA and LOH analysis (as well as potential CLOnality analysis) with SEquencing data. Current pipeline majorly facilitates the analysis on paired tumor and normal samples. CLOSE includes functions for both segment-level and read-level processing. Here we will mainly discuss segment-level models. At the segment level, the major data input are LRR and BAF values (or mean/median values) from each pre-defined segments or windows.

2 Segmental LRR and BAF models

LRR is also called Log R or Log 2 ratio. It is a term originated from array data analysis, where "R" represents variables for probe intensity. In sequencing data, it can naturally be used to represent Read Depth ratio, i.e. LRR $= \log_2(R_T/R_N)$, where R_T and R_N are mapped read counts of tumor and matched normal samples. The probability of observing n reads in a segment/bin is usually modeled through a Poisson distribution (and their ratio through a binomial). We can approximate the distribution of segmental LRR (mean) value by a Gaussian distribution. Let r_k denotes the sLRR value of the kth segment, we have

$$r_k|\mu_k, \sigma^2 \sim \mathcal{N}(\mu_k^{(r)}, \sigma^2)$$

, where μ_k reflects copy number changes while σ is estimated globally (often set at 0.18 in array data, see table below). To allow for more flexible dispersion control, a non-standardized Student's t distribution (Favero et al. 2015) can also be applied: $r_k \sim t(\mu_k, \nu, \sigma^2)$, where the degrees of freedom parameter ν can be estimated adaptively or set at a fixed value (e.g., ν=5 was used in Favero–at the depth ratio scale).

Given true local copy numbers in a pair of tumor and normal tissues n_T and n_N in a segment, the expected value of log2 depth ratio μ_k should be given by $\log_2(n_T/n_N)$. In reality, however, this value is affected by many factors inherent to tumor genomes and sequencing data. Let ϕ and ρ

[*] Supplementary materials for Wang et al. *Bioinformatics* 2015 --:--

[†] xuefeng.wang@stonybrook.edu; xuefeng.wang@yale.edu

denote the tumor purity and ploidy, respectively. The expected relative read ratio can be calculated as follows:

$$\mu_k^{(r)} = \log_2 \left[\frac{2(n_T/n_N)\phi + 2(1-\phi)}{\rho\phi + 2(1-\phi)} \times \frac{1}{\lambda} \right] = \log_2 \left[(n_C\phi + 2(1-\phi)) \times \frac{1}{2\lambda^*} \right] \tag{1}$$

,where $2(n_T/n_N)\phi+2(1-\phi)$ reflects the average copy number in the mixed sample and $\rho\phi+2(1-\phi)$ is the average ploidy. λ is a parameter that accounts for coverage difference between tumor and normal samples and $\lambda = 1$ if read counts are normalized. For simplicity, in (1) we defined a new parameter λ^* (λ^*=1 when $\rho = 2$ and $\lambda = 1$). n_C will be defined in the following.

BAF (B-allele frequency) is also a term adapted from array data analysis where BAF=intensity of B/ (intensity of A+B). Given a SNP site (after variant calling) with A and B allele, there are different genotypes under different copy number status {A, B, AB, AA, BB, AAB, ABB, AABB, AAAB, $\cdots$}. Therefore, BAF values associated with these genotypes are {0, 1, 1/2, 0, 1, 1/3, 2/3, 1/2, 1/4, $\cdots$}, respectively. In real data analysis, it is recommended to add a small (positive/negative) value, e.g. 0.05, to the theoretical BAFs that are exact (0/1), in order to compensate sequence mapping bias and other global errors. The deviation of BAFs from 0.5 is thus an indicator of loss of heterozygosity (LOH) event: AB(in the normal tissue)→other genotypes in tumor tissue. With paired tumor-normal samples, we only consider BAFs that are the heterozygous (SNPs) in the normal sample because homozygous will be non-informative for BAF estimation. Similarly, we can reasonably model the distribution of segmental BAF (sBAF) values by a Gaussian distribution. Let b_k denotes the sBAF mean value of the kth segment, we have $b_k \sim \mathcal{N}(\mu_k^{(b)}, \sigma^2)$, where the expected value of $b_k$is given by $\frac{n_B}{n_C}$ (where n_C=$n_A + n_B = 2$ if genotype AB). With mixed tumor and normal cells and purity ϕ (allelic imbalance not affected by ploidy ρ), it becomes

$$\mu_k^{(b)} = \frac{n_B\phi + (1-\phi)}{n_C\phi + 2(1-\phi)} \tag{2}$$

Because BAF signal is symmetric about 0.5 and much more sparse than LRR especially with WES data, a more desirable alternative is to calculate segmental LAF (lesser allele frequency):

$$sLAF = \begin{cases} sBAF & \forall sBAF < 0.5 \\ 1 - sBAF & \forall sBAF > 0.5 \end{cases} \tag{3}$$

This transformation makes the LOH modelling more efficient since usually we do not need to distinguish between certain genotype pairs such as AAB and ABB. As a reference for the subsequent parameterization of joint likelihood model, we summarize the genotype-CNA relationship (modified from the model implemented in Illumina GenomStudio) in the following table. Note that the standard deviations (SD) in this table are derived from the array data (but we have found it works surprisingly well as an approximation in many sequencing data).

Genotypes	Copy number	LRR mean	LRR SD	LAF mean	LAF SD
DD	0	-5	2	NA	NA
A or B	1	-0.45	0.18	0	0.03
AA or BB	2	0	0.18	0	0.03
AB	2	0	0.18	0.5	0.03
AAA or BBB	3	0.3	0.18	0	0.03
AAB or ABB	3	0.3	0.18	1/3	0.03
AAAA	4	0.75	0.18	0	0.03
AAAB or ABBB	4	0.75	0.18	0.25	0.03
BBBB	4	0.75	0.18	0	0.03

3 Joint likelihood model

The total likelihood is then the product of the likelihood (probability density) for each pre-defined segment, i.e.,

$$L(\Theta; \mathbf{r}, \mathbf{b}) = \prod L_s = \prod f(r_k, b_k|\Theta) \tag{4}$$

,where $\Theta = \{\mathbf{x}, \phi, \rho\}$ indicates the unknown parameter set including local copy number of the segment and global parameters purity and ploidy. It is justifiable to assume that, conditional on the underlying true copy number x_k, sLRR and sBAF are independent. Therefore, we have

$$f(r_k, b_k|\Theta) = f(r_k|\Theta) \times f(b_k|\Theta) \tag{5}$$

If we assume Gaussian density and ignore purity and ploidy parameter, the likelihood of genotype AB (based on previous genotype-CNA table) is simply

$$L^{(AB)} = \frac{1}{0.18\sqrt{2\pi}} exp - \frac{(r_k - 0)^2}{2(0.18^2)} \times \frac{1}{0.03\sqrt{2\pi}} exp - \frac{(b_k - 0.5)^2}{2(0.03)}$$

We may also construct composite likelihoods based on major CN status by pooling individual genotype likelihood, e.g.,

$$\begin{array}{cc}
L^{(0)} = L^{(DD)} & \text{Homozygous deletion} \\
L^{(1)} = L^{(A)} + L^{(B)} & \text{Hemizygous deletion} \\
L^{(2)} = L^{(AA)} + L^{(AB)} + L^{(BB)} & \text{Dizygous normal} \\
L^{(3)} = L^{(AAA)} + L^{(BBB)} + L^{(AAB)} + L^{(ABB)} & \text{Trizygous gain} \\
L^{(4)} = L^{(AAAA)} + L^{(AAAB)} + L^{(ABBB)} + L^{(BBBB)} & \text{Tetrazygous}
\end{array}$$

For each segment (with fixed ϕ and ρ) , CN status is determined by the largest composite likelihood value calculated above. To consider all segments, the CNA status and global parameters should be estimated globally by maximum the likelihood given in (4),

$$\Theta = \underset{\Theta}{\text{argmax}} \prod L_s \tag{6}$$

The MLE or MAP (if priors are specified) problem can be solved by EM algorithm (by treating copy number information $\mathbf{x}$ as latent/missing variables). MAP method is more robust to model violations and can be achieved at a substantially lower computational cost compared to MCMC estimates.

4 Global purity and ploidy

Global parameters (ϕ and ρ) can be estimated by solving:

$$\underset{\rho,\phi,x}{\operatorname{argmax}} \prod f(r_k, b_k|\phi, \rho, \tilde{x_i})p(\tilde{x_i}) \tag{7}$$

The solution to (7) can be found by EM algorithm or simply through a grid search over the parameter space $\{(\phi, \rho) : 0 < \phi < 1; \rho \in \{1, 2, 3, 4...\})\}$. A Dirichlet prior on copy numbers: $p(x_i) = \frac{1}{B(\alpha)} \prod x_i^{\alpha_i - 1}$and set hyperparameters to impose the prior knowledge such as copy number 2 is more common than others. Estimating all free parameters jointly in (6) and (7) may cause model overfitting especially when sequencing depth is low or when there are not enough large number of segments due to limited CNA vents. As an alternative to full likelihood model, we use a modified "canonical point" method based on the originally described method of Li et al. (2014). It includes three major steps (1) calculate all the possible canonical LRR-LAF points values under each pair of (ϕ, ρ) (2) assign each segment to the nearest canonical points; and (3) the final ϕ, ρ estimates are determined by the smallest total distance over all grid pairs. We weight the distance metric in step 2 to reflect segment length and to emphasize points with LAF values smaller than 0.25. Segments with smaller LAFs are less prone to global bias such as mapping error and are often act as critical anchor points in purity and ploidy pattern recognition. It can be shown that the canonical point method is essentially an approximation to the likelihood method.

To ensure robust estimation, we need to employ few important pre-processing steps: (1) center LRR values around the global segmental median; (2) remove all segments with LRR < -2 (homozygous deletion regions do not contribute to purity estimation); and (3) exclude all "normal" regions (segments with LRR close to 0 and LAF close to the global mode).

5 ASCN models

Once the global purity and ploidy are estimated, the ASCN (allele specific copy number) estimates can be obtained by solving two simultaneous equations (1) and (2). Here we derive an alternative formula to calculate ASCN based on segmental LRR (r) and LAF (l):

$$\begin{cases} \hat{n}_C = 2^r \times \rho + 2 \times (2^r - 1) \times \frac{1-\phi}{\phi} \\ \hat{n}_B = 2^r \times \rho \times l + (2^r \times l - 1) \times \frac{1-\phi}{\phi} \text{ or } n_C \times l + (2 \times l - 1) \times \frac{1-\phi}{\phi} \end{cases} \tag{8}$$

Equations (8) are the key functions used recursively in CLOSE. First, with fixed ρ andϕ, they can act as a mapping function to transform the input LRR-LAF values to more convenient ASCN space; therefore, the local (absolute) copy number calling (of each segment or each DP cluster that will be introduced below) can be performed by searching the nearest ASCN canonical points. Second, by relaxing ϕ, we can get canonical lines that enables the calculations of local aneuploid mixture ratio for each segment or cluster, which can be used as an effective surrogate for sub-clonal profiling. The following figure (generated by the demo function in CLOSE-R) illustrates the parameter trajectory of different copy number status along with purity.

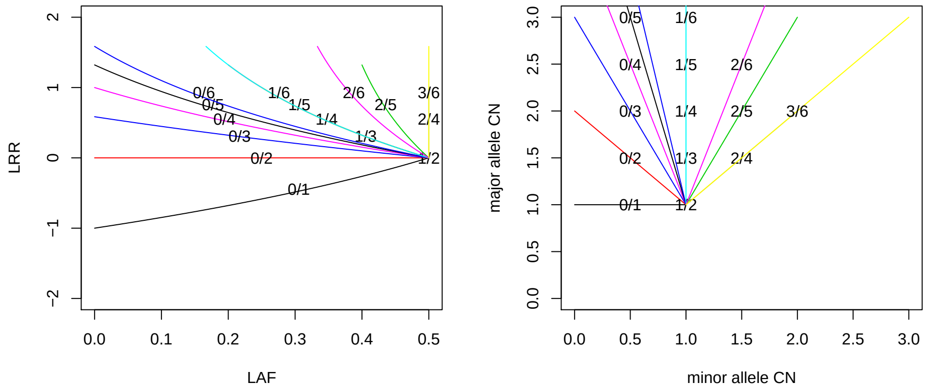

Figure 1: Parameter trajectory plots (canonical lines) with $\rho = 2$ and varying ϕ. Canonical points tend to shrink towards the normal status "1/2" with decreasing purity. The straight ASCN canonical lines (right panel) are apparently more convenient than the curved LRR-LAF lines in calculating distances. This figure also illustrates the non-identifiability problem of ploidy and purity. For genotypes sharing the same line (such as 2/4 and 3/6), the copy number status is chosen based on more criteria such that it is more consistent with global ploidy and purity estimates. Parameter plots of other ploidy can be generated by the demo function in CLOSE-R, e.g.$Clines.demo(la = 3)$ for triploid samples.

Based on lines and coordinates shown in Figure 1, we can (1) determine local copy number status based on the line of shortest distance from the point (2) estimate the corresponding local purity or mixture ratio indicated by the projection of the point to the line. As a complementary approach to segment-specific calling, a bivariate clustering analysis can be done first by partitioning all segments into clusters of different copy number status. In CLOSE-R, we use a distance-based Chinese Restaurant Process (Teh et al. 2006, Blei and Frazier 2011) to perform clustering analysis. The original CRP algorithm is as follows: (1) start with an empty group and the first data point is assigned to this group (2) When a new data point z_i comes in we make a stochastic decision.

$$z_i|z_{1:i-1},\alpha = \begin{cases} z_k & \text{with probability} \propto n_k \\ \text{new group} & \text{with probability} \propto \alpha \end{cases}$$

where n_k is the number of data points in group k. α is the parameter controls dispersion. In the distance based CRP the assignment to a previous group is proportional to the distance of the new datapoint to each previously assigned groups, i.e., $p(z_i|z_{1:i-1}, D, \alpha) \propto f(d_{ij})$. The major advantage of this approach over k-means clustering is that one does not need to specify a fixed number of clusters. The copy number status of segmental clusters is then determined by comparing the distances from the centroid of each cluster to all canonical lines as shown in this figure. This additional data-driven step ensures more reliable results, while providing better visualization of the global ASCN profiling.

References

1. Blei and Frazier *J. Mach. Learn. Res.* 12:2383-2410 (2011)
2. Carter et al. *Nat Biotechnol.* 30: 413-21 (2012).
3. Li and Li. *Genome Biol.* 15:473 (2014).
4. Li and Xie. *Bioinformatics* 30:2121-9 (2014).
5. Favero et al. *Ann Oncol.* 26:64-70 (2015).
6. Teh et al. *JASA.* 101:1566-1581 (2006)