

# A General Graphical Framework for Detecting Copy Number Variation

Xiao-Lin Yin<sup>1 2</sup>, Jing Li<sup>3</sup>

## 1 Introduction

Array comparative genomic hybridization (aCGH) allows identification of copy number variations (CNV) across genomes. The key challenge in analyzing CNV using aCGH data is the detection of segment boundaries of copy number changes and inference of the copy number state. We developed a novel statistical model based on the framework of conditional random fields (CRFs)[4] that can effectively combine data smoothing, segmentation and copy number state decoding into one framework. Our approach (termed CRF-CNV) provides great flexibilities in defining meaningful feature functions, therefore it can integrate local spatial information into the model. Experimental results demonstrate that CRF-CNV performs much better than two popular programs, CBS [2] and BHMM [1], in terms of breakpoint identifications as well as copy number assignments.

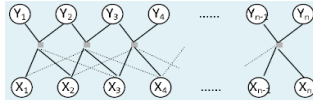


Figure 1: A linear chain conditional random field model for array CGH data.

## 2 Linear-Chain CRFs Model for aCGH Data

Let  $X = (X_1, \dots, X_n)$  denote the normalized  $\log_2$  ratio intensities along one chromosome for an individual, where  $X_i$  is the  $\log_2$  ratio for clone  $i$ . Let  $Y = (Y_1, \dots, Y_n)$  denote the corresponding hidden copy number state, where  $Y_i \in \{1, \dots, s\}$  and  $s$  is the total state number. The Linear-Chain CRFs Model (Figure 1) is defined as

$$P(Y|X) = \frac{1}{Z_\theta(X)} \exp\left\{ \sum_{i=1}^n \sum_{j=1}^s [\lambda_j f_j(Y_i, \tilde{X}_i(u)) + \mu_j g_j(Y_i, \tilde{X}_i(u))] + \sum_{j=1}^s \omega_j l_j(Y_1, \tilde{X}_1(u)) + \sum_{i=1}^{n-1} \sum_{j=1}^s \sum_{k=1}^s \nu_{jk} h_{jk}(Y_i, Y_{i+1}, \tilde{X}_{i,i+1}(u)) \right\}, \quad (1)$$

where  $Z_\theta(X)$  is the partition function,  $\theta = \{\lambda_j, \mu_j, \omega_j, \nu_{jk}\}$  are parameters.  $\tilde{X}_i(u)$  is defined as a neighbor set of  $X_i$  around clone  $i$  with dependence length  $u$ , i.e.,  $\tilde{X}_i(u) = \{X_{i-u}, \dots, X_i, \dots, X_{i+u}\}$ . Similarly,  $\tilde{X}_{i,i+1}(u) = \{X_{i-u}, \dots, X_i, \dots, X_{i+u+1}\}$ ,  $\tilde{X}_{i,i+1}^-(u) = \{X_{i-u}, \dots, X_i\}$  and  $\tilde{X}_{i,i+1}^+(u) = \{X_{i+1}, \dots, X_{i+u+1}\}$ . For notational simplification, we write  $\tilde{X}_i(u)$  as  $\tilde{X}_i$ , and *etc.* The emission feature functions  $f_j$  and  $g_j$  are defined as:

$$f_j(Y_i, \tilde{X}_i) = \begin{cases} \text{med } \tilde{X}_i & \text{if } Y_i = j \\ 0 & \text{otherwise,} \end{cases} \quad g_j(Y_i, \tilde{X}_i) = \begin{cases} (\text{med } \tilde{X}_i)^2 & \text{if } Y_i = j \\ 0 & \text{otherwise,} \end{cases}$$

$\text{med } \tilde{X}_i$  is the median value of  $\tilde{X}_i$ . The initial and the transition feature functions of  $l_j$  and  $h_{jk}$  are defined as:

$$l_j(Y_1, \tilde{X}_1) = \begin{cases} (a_{j+1} - a_j) / [(a_{j+1} - a_j) + 2(\text{med } \tilde{X}_1 - a_j)] & \text{if } Y_1 = j, \text{med } \tilde{X}_1 \geq a_j \\ (a_j - a_{j-1}) / [(a_j - a_{j-1}) + 2(a_j - \text{med } \tilde{X}_1)] & \text{if } Y_1 = j, \text{med } \tilde{X}_1 < a_j \\ 0 & \text{otherwise.} \end{cases}$$

<sup>1</sup>Case Western Reserve Univ., Dept. of Elect. Engn. & Comp. Sci., Cleveland, OH, USA. E-mail: xly@case.edu

<sup>2</sup>NE Normal Univ., Sch. Math. & Stat., Key Lab Appl. State MOE, Changchun, Jilin, China. E-mail: yinxl805@nenu.edu.cn

<sup>3</sup>Case Western Reserve Univ., Dept. of Elect. Engn. & Comp. Sci., Cleveland, OH, USA. E-mail: jingli@eecs.case.edu

$$h_{jk}(Y_i, Y_{i+1}, \tilde{X}_{i,i+1}) = \begin{cases} \frac{(a_{j+1}-a_j)/2+(a_{k+1}-a_k)/2}{(a_{j+1}-a_j)/2+(a_{k+1}-a_k)/2+\text{med } \tilde{X}_{i,i+1}^- - a_j + \text{med } \tilde{X}_{i,i+1}^+ - a_k} & \text{if } Y_i = j, \text{med } \tilde{X}_{i,i+1}^- \geq a_j, Y_{i+1} = k, \text{med } \tilde{X}_{i,i+1}^+ \geq a_k \\ \frac{(a_{j+1}-a_j)/2+(a_k-a_{k-1})/2}{(a_{j+1}-a_j)/2+(a_k-a_{k-1})/2+\text{med } \tilde{X}_{i,i+1}^- - a_j + a_k - \text{med } \tilde{X}_{i,i+1}^+} & \text{if } Y_i = j, \text{med } \tilde{X}_{i,i+1}^- \geq a_j, Y_{i+1} = k, \text{med } \tilde{X}_{i,i+1}^+ < a_k \\ \frac{(a_j-a_{j-1})/2+(a_{k+1}-a_k)/2}{(a_j-a_{j-1})/2+(a_{k+1}-a_k)/2+a_j - \text{med } \tilde{X}_{i,i+1}^- + \text{med } \tilde{X}_{i,i+1}^+ - a_k} & \text{if } Y_i = j, \text{med } \tilde{X}_{i,i+1}^- < a_j, Y_{i+1} = k, \text{med } \tilde{X}_{i,i+1}^+ \geq a_k \\ \frac{(a_j-a_{j-1})/2+(a_k-a_{k-1})/2}{(a_j-a_{j-1})/2+(a_k-a_{k-1})/2+a_j - \text{med } \tilde{X}_{i,i+1}^- + a_k - \text{med } \tilde{X}_{i,i+1}^+} & \text{if } Y_i = j, \text{med } \tilde{X}_{i,i+1}^- < a_j, Y_{i+1} = k, \text{med } \tilde{X}_{i,i+1}^+ < a_k \\ 0 & \text{otherwise,} \end{cases}$$

Here  $a_j$  denotes the mean  $\log_2$  ratio for clones with state  $j$ ,  $a_0$  and  $a_{s+1}$  denote the greatest lower bound and the least upper bound of all  $\log_2$  ratios, respectively. All types of our feature functions can capture the local spatial dependence over a set of adjacent clones thus provide more robust inference about copy number states.

For parameter estimations in model (1), we adopted the conjugate gradient(CG) method on a set of training data  $\mathcal{D} = \{(X^{(d)}, Y^{(d)}), d = 1, \dots, D\}$  and developed forward/backward algorithms within the CG framework.

### 3 An Real Example and Simulated Data

The Coriell data [3] is a ‘‘gold standard’’ real data. Table 1 shows the segment numbers of each sample from the Gold Standard, CRF-CNV, CBS and BHMM. The segment number detected by CRF-CNV is exactly the same as the Gold Standard for almost all samples (except for sample 9 and 10). It shows that both CBS and BHMM have generated many more segments comparing to the truth.

method \ sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	sum
Gold	5	3	5	3	5	3	5	5	5	5	3	5	2	3	3	60
CRF-CNV	5	3	5	3	5	3	5	5	3	3	3	5	2	3	3	56
CBS	17	42	7	6	9	5	5	6	5	5	13	7	17	3	9	156
BHMM	6	3	11	3	7	14	6	7	9	7	8	11	4	8	5	109

Table 1: Comparison of segment numbers return by three algorithms.

We pool all the breakpoints from all the samples and use the  $F$  measure (a combination of *precision* and *recall*) to compare the performance of the three algorithms. We use a match extent index  $D$  to allow some flexibility in defining matches of predicted breakpoints to those given by the gold standard. Table 2 shows  $F$  measure outcomes. Clearly, CRF-CNV achieved much better accuracy than the other two approaches while maintaining the same level of recall on the real data.

method \ math extent	0	1	2	3	4
CRF-CNV	0.638	0.914	0.948	0.967	0.967
CBS	0.333	0.500	0.519	0.519	0.519
BHMM	0.580	0.627	0.639	0.639	0.639

Table 2: The comparison of  $F$  measure with different match extent.

We also use a simulated dataset [5] to make comparisons. It shows that CRF-CNV performs much better than BHMM approach on simulated datasets in terms of breakpoint identifications as well as copy number assignments. Comparing to CBS approach, CRF-CNV has obtained comparable results on simulated data.

## References

- [1] Guha, S., Li, Y. and Neuberg, D. Bayesian hidden markov modeling of array cgh data, 2006.
- [2] Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. 2004. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–72.
- [3] Snijders, A. M., Nowak, N., Segreaves, R., and et al. 2001. Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics*, 29(3):263–4.
- [4] Sutton, C. and McCallum, A. 2007. An introduction to conditional random fields for relational learning. In Getoor, L. and Taskar, B., editors, *Introduction to Statistical Relational Learning*. MIT Press.
- [5] Willenbrock, H. and Fridlyand, J. 2005. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21(22):4084–91.