ALGORITHM FOR DNA COPY NUMBER VARIATION DETECTION WITH READ DEPTH AND PARAMORPHISM INFORMATION

Rong Shen[1], *Kai Ying*[2], *Zhengdao Wang*[1], *Patrick S. Schnable*[3]

[1] Department of Electrical and Computer Engineering, Iowa State University, USA
[2] Center for Prostate Disease Research, Department of Defense, USA
[3] Center for Plant Genomics, Iowa State University, USA

ABSTRACT

Next-generation sequencing (NGS) has revolutionized the detection of structural variation in genome. Among NGS strategies, read depth is widely used and paramorphism information contained inside is generally ignored. We develop an algorithm that can fully exploit both read depth and paramorphism information. We embed mutation procedure in our system model for estimating prior likelihood of single nucleotide base. Hidden Markov model (HMM) is used to connect single base into segments and belief propagation algorithm is performed for the optimal solution of the HMM model. Simulations show promising results in detecting important types of structural variation. We have applied the algorithm on the maize B73 and MO17 genome data and compared the results with those obtained from arrayCGH method based micro-array data. Inconsistency between the two sets of data is discussed.

Index Terms— Copy number variation, graphical model, belief propagation, read depth, paramorphism

1. INTRODUCTION

Structural variation is defined as insertions, deletions and inversions in the sequence level or copy number variation (CNV) and presence/absence variation (PAV) in genomic level. Structural Variations (SV) are associated with the cause of disease as well as different traits between individuals [3, 4, 12]. Detecting structure variation in genomes has been under research quite long and great development has been achieved [1]. Among the SVs, CNV corresponds to relatively large regions of deletion or duplication more than the usual number of copies [5, 9]. Some CNVs are known to be associated with diseases [2, 11].

Compared to array-based methods, methods based on next-generation sequencing (NGS) require less labor and have less limitations in accuracy. Through *de novo* assembly, given long and accurate enough read sequence, all kinds of SV could be reconstructed [1]. However, *de novo* assembly is still under development to reduce its complexity and improve algorithm speed as well as reduce cost for large genome datasets.

Tuzun et al. [8] proposed a way of detecting accurate small SV segments less than 1 kbp using paired-end reading (PEM) strategy. Small length SVs as well as their boundaries could be estimated precisely through exploiting paramorphism information. Zollner et al. [12] applied Bayesian computations and expectation-maximization (EM) algorithm to a known CNV location and achieved accurate estimation of CNV carrier status and its boundaries. However, known location is necessary in this algorithm. Compared to them, read depth provides a wide detection range and statistical accuracy. Event-Wise Testing achieved fast algorithm speed via processing intervals of read depth and applied statistical significance test to the intervals with controlled significance level [10]. It achieved satisfactory results for CNV segments around 1000bp. A method called CNAseg applied Skellam distribution to read depth and employed Hidden Markov Model (HMM) on combining segments of read counts and found better estimation and precision besides lower significance level [4].

We are interested in detecting CNV based on NGS data in this paper. We will combine both read depth and paramorphism to achieve better CNV detection range and accuracy. Ideally, segments less or larger than 1 kbp can both be detected accurately. The main contributions are as follows:
1) We propose a novel model for the process of nucleotide copies, paramorphism, and the randomness in the sequencing process.
2) We evaluate likelihood of CNV considering mutation probability at each base pair location, taking into account both read depth and paramorphism information.
3) We proposes a simple HMM relating the copy number variation and presence/absence variation of neighboring base pairs. We also derive a belief propagation algorithm to estimate the copy numbers and possible presence/absence variation.

The algorithm has been applied to lab measured data comprising a reference genome and a sample genome. The results are compared with those obtained with arrayCGH.

2. SYSTEM MODEL

Our system model consists of two parts: 1) a single symbol model for a base pair that considers copy number variation, mutation, and the randomness in the sampling process; and 2) an HMM that incorporates the dependence among neighboring base pairs.

2.1. Single Base Pair Model

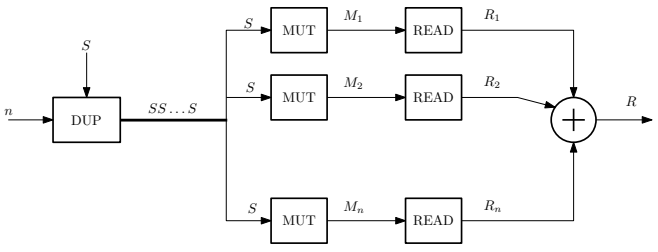

Fig. 1. Single Base Pair Model

Our proposed model is depicted in Figure 1. The model describes what happens to a single base pair in the sample species. A symbol S goes through three steps to produce the observed data: copying, mutation, and sampling (or reading). We are interested in using the model to estimate the likelihood of copy numbers from the known read depth and the read distribution, available in the PILEUP format data. There are 4 possibilities for S, namely $\{A, C, G, T\} := \mathcal{B}$.

Copy. For each symbol, assume its true copy number is n. The DUP block in the figure replicates the symbol S and produces n copies of S at its output.

Mutation. Each of the n copies can mutate to a different symbol with certain probability. This is represented as a MUT block in the figure. For simplicity, we assume the non-mutation probability for each copy of each symbol is the same as $(1-p)$. Let $M_1, M_2, \ldots, M_n$ be the n symbols after mutation. We denote the type of the mutated symbols as $\mathbf{n} = (n_A, n_C, n_G, n_T)$, where n_i is the number of symbols i in the sequence $(M_1, M_2, \ldots, M_n)$, for $i \in \{A, C, G, T\}$.

The mutation distribution $\mathbf{n}$ is a vector describing mutation distribution in the order of ATGC. For example, if $S = A$, and there is no mutation, then $\mathbf{n} = [n, 0, 0, 0]$. If A has one copy symbol mutated to G, then $\mathbf{n} = [(n-1), 0, 1, 0]$.

Given the known symbol S, and the copy number n, the mutation distribution is multinomial:

$$\Pr(\mathbf{n}|n, S) = \frac{n!}{\prod_{i\in\mathcal{B}} n_i!} \prod_{i\in\mathcal{B}} {P_i}^{n_i}, \tag{1}$$

where

$$P_i = \begin{cases} 1-p, & i = S, \\ p/3, & i \neq S. \end{cases} \tag{2}$$

Reading. Each mutated copy has a certain probability of being sampled by the sequencing procedure. The reading result has two parts of information: the read depth, and the read distribution.

We use Poisson distribution to model the read depth. Assuming the read depth number is k, the probability of getting a known k read depth given copy number n should be like this:

$$\Pr(k|n) = \frac{(n\lambda)^k e^{-n\lambda}}{k!} \tag{3}$$

where λ is the parameter of the Poisson distribution. Symbols in reference genome are assumed to have only one copy. If there is no copy procedure, sample symbol read depths are expected to have the same mean as the reference symbols. So we will choose the read depth in the reference genome times possible copy numbers as the parameter λ, the expected rate of increasing number for that symbol.

Reading Error. In the process of mapping symbols, there is a probability of error for each read symbol. However, in this work, for simplicity, we consider the probability of error to be the same for all, represented by ϵ. Suppose there are n_a counts of symbol A, n_G counts of G, n_C counts of C and n_T counts of T. For each symbol, the probability of detecting one A in the observed data is:

$$q_A = (1-\epsilon)n_a/n + (\epsilon/3)(n_g + n_c + n_t)/n,$$

where $n = n_A + n_C + n_G + n_T$. In general, for any symbol i, the probability of reading an i would be $q_i = (1-\epsilon)n_i/n + (\epsilon/3)\sum_{j\neq i} n_j/n$.

That is, given the mutation distribution $\mathbf{n}$, the distribution of the read results $\mathbf{k} := (k_A, k_C, k_G, k_T)$ should be as follows:

$$\Pr(\mathbf{k}|k, \mathbf{n}) = \frac{k!}{\prod_{i\in\mathcal{B}} k_i!} \prod_{i\in\mathcal{B}} {q_i}^{k_i} \tag{4}$$

where $\mathcal{B} = \{A, T, G, C\}$, and $k = \sum_{i\in\mathcal{B}} k_i$ is the read depth.

2.2. Likelihood Copy number

Using the chain rule of probability, we can write $\Pr(\mathbf{k}, k, \mathbf{n}|n, S)$ as

$$\begin{aligned} &\Pr(\mathbf{n}|n, S) \cdot \Pr(k|n, S, \mathbf{n}) \cdot \Pr(\mathbf{k}|n, S, \mathbf{n}, k) \\ &= \Pr(\mathbf{n}|n, S) \cdot \Pr(k|n) \cdot \Pr(\mathbf{k}|\mathbf{n}, k) \end{aligned} \tag{5}$$

where in (5), we have used

$$\Pr(k|n, S, \mathbf{n}) = \Pr(k|n) \tag{6}$$

which means that the read depth does not depend on what symbols are being read, and

$$\Pr(\mathbf{k}|n, S, \mathbf{n}, k) = \Pr(\mathbf{k}|\mathbf{n}, k) \tag{7}$$

which means that once the mutation distribution $\mathbf{n}$ is known, and for a given read depth k, $\mathbf{k}$ does not depend on the copy number n and the source symbol S.

We can then marginalize (average out) the mutation $\mathbf{n}$ to obtain

$$\Pr(\mathbf{k}, k|n, S) = \sum_{\mathbf{n}} \Pr(\mathbf{k}, k, \mathbf{n}|n, S), \tag{8}$$

where the summation is over $\mathbf{n}$ such that $\sum_{i\in\mathcal{B}} n_i = n$. Further marginalizing the total read depth k, which is trivial as $k = \sum_{i\in\mathcal{B}} k_i$ deterministically, we have

$$\Pr(\mathbf{k}|n, S) = \sum_{\mathbf{n}} \Pr(\mathbf{k}, k, \mathbf{n}|n, S) \tag{9}$$

2.3. Hidden Markov model

To model the dependency of the neighboring symbols' copy numbers, we assume a HMM and use it for combining long sequence of copy number estimates. HMM has been previously used to model single nucleotide polymorphisms (SNP) detection [4, 7].

Let n_{i-1} and n_i represent the copy number at two adjacent location $i-1$ and i. The variable δ_i could be viewed as the hidden state at location i. The mapping relationship between copy number and hidden variable is:

$$n_{i+1} = 1 + (n_i + \delta_i - 1) \mod m. \tag{10}$$

The m above stands for the number of states in the model.

The change δ_i has a probability distribution like follows:

$$\Pr(\delta = 0) = p, \quad \text{and} \quad \Pr(\delta \neq 0) = 1 - p,$$

where $\delta = 0$ represents no change in copy number, and $\delta \neq 0$ indicates a change. In the case where there is a change, we will assume that the probability $(1-p)$ is equally split among the cases where $n_{i+1} \neq n_i$. The probability p will be chosen according to desired gene CNV segment length. The Markov model corresponds to setting

$$\Pr(n_{i+1}|n_i, n_{i-1}, n_{i-2}, \ldots) = \Pr(n_{i+1}|n_i)$$

and

$$\Pr(n_{i+1} = n_i|n_i) = p, \ \Pr(n_{i+1} \neq n_i|n_i) = 1 - p.$$

3. ALGORITHM

The algorithm consists two steps. In the first step, each symbol is processed independently to obtain the likelihood for different copy numbers and presence/absence variations. In the second step, the

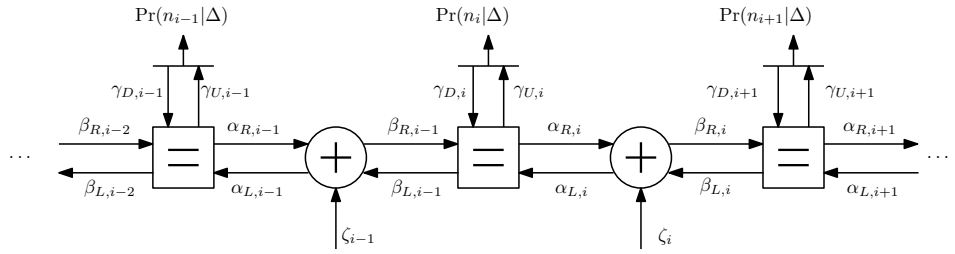

Fig. 2. Sum-Product Algorithm for the CNV problem

single-symbol information is jointly processed using the hidden Markov model through a belief propagation algorithm.

3.1. Single-Symbol Processing

The input is one input line in the PILEUP file data (the read results of the sequencing output for one base pair). And the output is the likelihood of various copy number possibilities (states).

In many genomes such as the maize genome, large copy number is common. However, the CNV that we do have interest is relative small numbers, usually less than 3. For this reason, and to reduce the computation complexity, we consider the cases where the copy number n is larger than 3 jointly.

For each symbol of the sample, we compute the likelihood of the copy number n being in state $i \in \mathcal{A} := \{0, 1^-, 1, 2, 3, 3^+\}$, where

- $n = 0$ means deletion: the segment was present in the reference but not present in the sample.
- $n = 1^-$ means copy number reduction: the copy number in the reference is larger than 1, and the copy number in the sample is smaller than that in the reference.
- $n = 1, 2, 3$ means respectively that the copy number is 1, 2, and 3.
- $n = 3^+$ means that the copy number is larger than 3.

Given the observation data, which is in the form of the read depth and read distribution, we can then compute the likelihood for each base pair. For simplicity and faster processing speed, we choose to ignore the mutation information when the read depth is large enough.

3.2. Belief propagation algorithm

Let N denotes the total length of a chromosome of interest, our observed data are $\Delta = (S_1, \ldots, S_N; \mathbf{k}_1 \ldots \mathbf{k}_N)$. Our goal is to estimate the copy numbers n_j for $j = 1, \ldots, N$. Specifically, we would like to compute the following distributions:

$$P(n_j|S_1, \ldots, S_N; \mathbf{k}_1 \ldots \mathbf{k}_N). \tag{11}$$

This problem in general has high complexity due to the need to marginalize all symbols but n_j in the joint posterior distribution of $(n_1, \ldots, n_N)$. For our problem, we have

$$\begin{aligned} & P(n_1, \ldots, n_N|S_1, \ldots, S_N; \mathbf{k}_1, \ldots, \mathbf{k}_N) && (12) \\ & \propto P(n_1, \ldots, n_N; \mathbf{k}_1, \ldots, \mathbf{k}_N|S_1, \ldots, S_N) \\ & = \prod_{i=1}^{N} P(n_i|n_{i-1}) \prod_{i=1}^{N} P(\mathbf{k}_i|S_i, n_i) && (13) \\ & = \prod_{i=1}^{N} P(n_i|n_{i-1}) P(\mathbf{k}_i|S_i, n_i) && (14) \end{aligned}$$

where in (13) we have used the assumption that the copy numbers $(n_1, \ldots, n_N)$ form a Markov chain, and the fact that given the copy number n_i and the symbol S_i the read $\mathbf{k}_i$ at location i is independent of symbols, copy numbers, and read results at other locations.

Due to the Markov structure, a low complexity algorithm for computing the posterior probabilities in (11) is possible through the belief propagation algorithm. We use factor graph [6] to represent the HMM constraints as a graphical model and use the sum-product algorithm to obtain the posterior distribution of the copy numbers.

There are two types of nodes in Fig 2. One type that has the "=" sign represents constraints of equal value passing through the node. Another node has + sign inside which represents constraints that one value equals to the summation of another two.

We computations to be performed in the message passing algorithm are as follows:

1. At the "Plus" nodes:
   (a) From left to right:
   $$\beta_{R,i}(n) = \sum_l \alpha_{R,i}(l)\zeta_i(n-l). \tag{15}$$
   (b) From right to left:
   $$\alpha_{L,i}(n) = \sum_l \beta_{R,i}(l)\zeta_i(l-n). \tag{16}$$
2. At the "Equal" nodes:
   (a) From left to right:
   $$\alpha_{R,i}(n) = \beta_{R,i-1}(n)\gamma_{D,i}(n) \tag{17}$$
   (b) From right to left:
   $$\beta_{L,i-1}(n) = \alpha_{L,i}(n)\gamma_{D,i}(n) \tag{18}$$
   (c) The message up:
   $$\gamma_{U,i-1}(n) = \alpha_{L,i}(n)\beta_{R,i-1}(n) \tag{19}$$

The output of the algorithm is the posterior probabilities in (11). Specifically, we have

$$P(n_j|S_1, \ldots, S_N; \mathbf{k}_1 \ldots \mathbf{k}_N) = \gamma_{D,i}(n)\gamma_{U,i}(n).$$

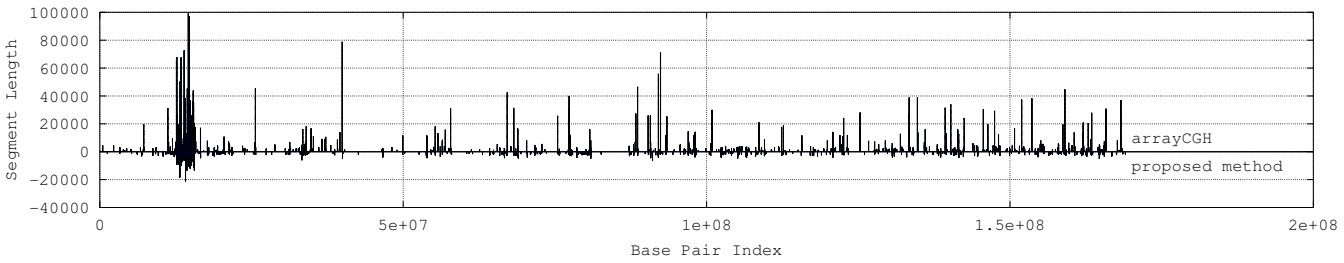

Fig. 3. PAV Segments

4. RESULTS

We used maize genome data for analysis. Two sets of data for two species are available. The reference genome is named B73 and the sample one named MO17. We process the genome from chromosome to chromosome. For each chromosome, we read the index, symbol, read depth and read depth distribution from the sample genome and the index, and retrieve the read depth from reference genome. We then transform the data from symbols to numbers. We use 1, 2, 3, and 4 in our algorithm to represent A T G C, respectively.

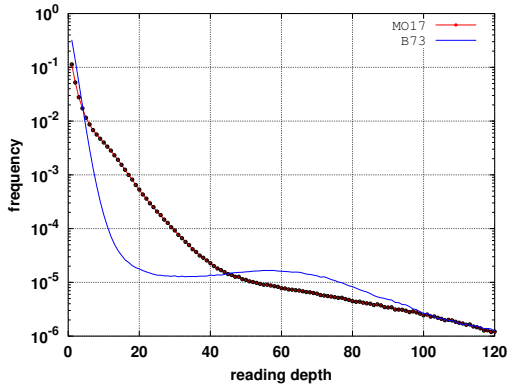

Fig. 4. read depth distribution in both the reference and the sample

We find the nucleotide indices that exist in both sample and reference genome and assign the reference read depth as the λ parameter for the corresponding symbol on the sample. The we take the union of the reference index and sample index. Thus if there is a total deletion, the corresponding value for λ is zero, which would be easy for the algorithm detection.

4.1. Distribution of the read depth

We obtained the distribution of the read depth in the two genomes. In Figure 4, the read depth distribution in both the reference and the sample is depicted. For the sample (MO17) and for $n = 1, 2, 3$, the distribution can be approximated as $(0.41)^n$. For $n \geq 4$, we averaged the ratio between two adjacent numbers and found it can be approximated as $0.0637 \times (0.8544)^{(n-4)}$. The likelihood function finally turns out to be

$$L(n \geq 4) = 0.2732 \times \frac{\lambda^k (k+1)}{(\lambda - \ln 0.8544)^k} [1 - \gamma^{(inc)}(4\lambda, k)] \quad (20)$$

The $\gamma^{(inc)}(\cdot)$ denotes the incomplete gamma integral from 1 to 4λ:

$$\gamma^{(inc)}(x, k) = \int_0^x t^{k-1} e^{-t} dt. \quad (21)$$

Use integral to replace the summation may lead to biased results. However, such bias would not affect the decision significantly.

4.2. Processing measured genomic data

For measured genomic data, we took chromosome number 6 in reference genome and its corresponding sample genome and ran our algorithm. Since there are no ground truth data on the structure variation in the chromosome, we compared our result with that of micro-array-based method arrayCGH.

We set the reading error probability to 10^{-3}. We only keep segments longer than 1kb in the output, by performing filtering on the message passing algorithm output. In Figure 3 we plotted the detected PAV result, as compared with that from the arrayCGH method. The x axis denotes the starting point of a structure variation segment, and the length in the y direction describes the length of the segment. For the arrayCGH result, we plot them upward, and for our result, we plot them downward. It can be seen that most of the locations are overlapped in the figure. However, the lengths of the segments tend to be longer in the arrayCGH case.

Except for the small segments of deletion, outliers would result in breakpoints. Fortunately, outliers usually affect a small area which lasts about 20 symbols. If we fill those regions as the CNV we want, it would be easy for the algorithm to detect CNV larger than 1kb. Almost all the CNV indicated from arrayCGH are included in our result. Besides, our result reports more possible regions of CNV.

5. CONCLUSIONS AND DISCUSSION

Using read depth and paramorphism information, we developed a method for detecting copy number variation between two different genomes. We exploited the paramorphism information to strengthen read depth power in detection of structural variation in genome. Also we applied belief propagation to solve the HMM and found the conditional single base copy number probability based on the prior information of other base pairs. Our results offered similar regions and reliable likelihood of the PAV and CNV region compared with those detected from arrayCGH method. Besides, our method provides accurate start and end locations for simulated region. Our future work will focus on improving the segment detection accuracy and reducing the overall algorithm complexity. It would also be useful to be able to treat diploid genomic data.

Acknowledgment: The research in this paper was supported in part by NSF Grant 1523374.

6. REFERENCES

[1] C. Alkan, B. Coe, and E. Eichler, "Genome structural variation discovery and genotyping," *Nature Reviews Genetics*, vol. 12, pp. 363–376, 2011.

[2] B. P. Coe, K. Witherspoon, J. A. Rosenfeld, et al., "Refining analyses of copy number variation identifies specific genes associated with developmental delay," *Nature genetics*, vol. 46, no. 10, pp. 1063–1071, 2014.

[3] E. Gonzalez, H. Kulkarni, H. Bolivar, et al., "The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility," *Science*, vol. 307, no. 5714, pp. 1434–1440, Mar. 2005.

[4] S. Ivakhno, T. R. T, A. C. A, et al., "CNAseg — a novel framework for identification of copy number changes in cancer from second-generation sequecing data," *Bioinformatics*, vol. 26, no. 24, pp. 3051–3058, 2010.

[5] N. Krumm, P. H. Sudmant, A. Ko, et al., "Copy number variation detection and genotyping from exome sequence data," *Genome research*, vol. 22, no. 8, pp. 1525–1532, 2012.

[6] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Processing Magazine*, vol. 21, no. 1, pp. 28–41, Jan. 2004.

[7] J. C. Marioni, N. P. Thorne, and S. Tavaré, "Biohmm: a heterogeneous hidden Markov model for segmenting array CGH data," *Bioinformatics*, vol. 22, no. 9, pp. 1144–1146, 2006.

[8] E. Tuzun, A. J. Sharp, J. A. Bailey, et al., "Fine-scale structural variation of the human genome," *Nature Genetics*, vol. 37, no. 7, pp. 727–732, May 2005.

[9] L. V. Wain and M. D. Tobin, "Copy number variation," in *Genetic Epidemiology*, pp. 167–183. Springer, 2011.

[10] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome Research*, vol. 19, no. 9, pp. 1586–1592, Sept. 2009.

[11] M. Zarrei, J. R. MacDonald, D. Merico, and S. W. Scherer, "A copy number variation map of the human genome," *Nature Reviews Genetics*, 2015.

[12] S. Zöllner, G. Su, W. Stewart, et al., "Bayesian EM algorithm for scoring polymorphic deletions from SNP data and application to a common CNV on 8q24," *Genetic Epidemiology*, vol. 33, no. 4, pp. 357–368, 2009.