ISIT'2013 1569720459

Message Passing Algorithm for Inferring Consensus Sequence from Next-Generation Sequencing Data

Xiaohu Shen
ECE Department
The University of Texas at Austin
Austin, USA
xhshen@utexas.edu

Manohar Shamaiah
Broadcom Inc
Bangalore, India
manohar.shamaiah@gmail.com

Haris Vikalo
ECE Department
The University of Texas at Austin
Austin, USA
hvikalo@ece.utexas.edu

***Abstract*—In order to determine an individual's DNA sequence, sequencing platforms often employ shotgun sequencing where multiple identical copies of the DNA strand of interest are randomly fragmented and then the nucleotide content of the short fragments is determined. Assembly of the long DNA strand from short fragments is a computationally challenging task that has attracted significant amount of attention in recent years. We formulate reference-guided assembly as the inference problem on a bipartite graph and solve it using a message-passing algorithm. The message-passing algorithm does not need to rely on the quality score information which expresses reliability of the short reads. To assess the performance of the proposed methodology, we derive an expression for the probability of error of a genie-aided MAP consensus scheme. Simulation results on a *Neisseria meningitidis* data set demonstrate that the proposed message-passing algorithm performs close to the idealistic MAP consensus scheme.**

I. Introduction

Advancements in next-generation DNA sequencing technologies have enabled cheap and fast generation of massive amounts of sequencing data [1], [2]. Inferring the order of nucleotides in a long target DNA molecule typically involves use of shotgun sequencing technique where multiple copies of the target DNA strand are fragmented into short templates. Each template is then analyzed by a sequencing platform which provides reads that are used to assemble the desired long target sequence. Shotgun sequencing strategy is illustrated in Fig. 1.

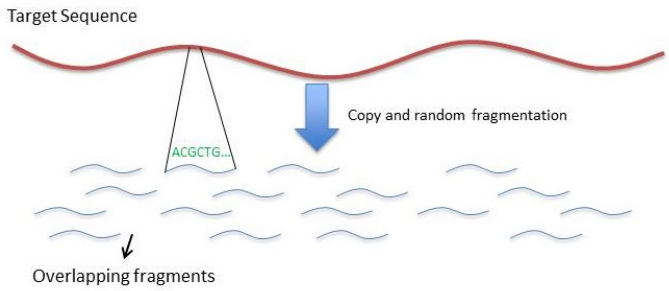

Fig. 1. *Multiple copies of a long target DNA molecule are fragmented into short templates, and the order of bases in the templates is determined using a sequencing platform.*

Next-generation sequencing methods rely on enzymatic synthesis of the complementary strands on templates to detect the order of nucleotides in the templates. Base calling algorithms attempt to infer the order of nucleotides in short templates from the acquired noisy signals. Measure of accuracy of the base caller's output is provided in form of quality scores. Many existing base calling algorithms rely on various heuristics to estimate quality scores, while more recent base calling methods employ Bayesian inference schemes to evaluate posteriori probabilities of the bases in the reads [3], [4]. State-of the-art target assembly methods typically rely on the base calling quality scores. Note, however, that since the quality scores are typically heuristically inferred, they are not necessarily a faithful measure of the accuracy of base calling results.

In re-sequencing tasks, individual genomes are sequenced with the goal of, for instance, studying genetic variations. However, even when the reference genome exists, assembly of a genome from short reads is computationally challenging. Note that when reconstructing the target DNA sequence using short-reads and a reference, the short reads are first mapped to a reference sequence (a DNA sequence highly similar to the target but not identical) using an alignment algorithm (e.g., [6], [7]). The target sequence is then determined by reaching consensus of the redundant information given by the overlapping reads. In practice, both the mapping and consensus steps are potentially erroneous. However, since each base in the target sequence is typically covered by a large number of short reads, accurate consensus using traditional techniques such as plurality voting is possible. To this end, there is a need for computationally efficient methods capable of fast correction of base calling errors.

In this paper, we formulate reference-guided assembly as the inference problem on a bipartite graph and solve it using a message-passing algorithm. The message-passing algorithm does not need to rely on the quality score information which expresses reliability of the short reads. This is a departure from the existing methods which typically employ quality scores in the assembly procedure. To assess the accuracy of the proposed technique, we derive an analytical expression for the probability of error of a genie-aided maximum a posteriori (MAP) consensus scheme. Simulation results on an *Neisseria Meningitidis* data set demonstrate that the proposed message-passing algorithm performs close to the idealistic MAP consensus scheme.

II. Graphical Model and the Message-Passing Consensus Algorithm

To graphically represent assembly of short reads generated by next-generation sequencing platforms, we introduce a bipartite graph representing the reads and bases in the target sequence that needs to be assembled. The sequence consists of four different kinds of nucleotides – A, C, G, T. We use unit vectors to represent them. In particular, nucleotide bases A, C, G and T are represented by $\mathbf{e}_A = [1,0,0,0]^T$, $\mathbf{e}_C = [0,1,0,0]^T$, $\mathbf{e}_G = [0,0,1,0]^T$, and $\mathbf{e}_T = [0,0,0,1]^T$, respectively. Assume the target sequence has length L, and denote the bases in the sequence as $b_{1:L}$. Then each symbol in the target sequence $b_i \in \{\mathbf{e}_A, \mathbf{e}_C, \mathbf{e}_G, \mathbf{e}_T\}$. For convenience, we will assume that all the short reads at our disposal are generated by the same sequencing platform and thus have identical read length l; note, however, that there is no loss of generality and that our scheme can combine reads generated by sequencing the same target on different platforms and of different read lengths. Let us denote the set of short reads by $\mathcal{R} = \{r_j\}$, $j = 1, 2, \ldots, n$. In general, the base calls in these reads are erroneous.

In reference-guided assembly, short reads are mapped onto the reference sequence using alignment algorithms. Each read is mapped to a specific position on the reference sequence, with the possibility of having several candidate positions (multiple mappings can be incorporated into the graphical model and resolved using the proposed scheme). Representing bases in the target sequence and reads by nodes and connecting read nodes with their corresponding target base nodes leads to a bipartite graph such as the one illustrated in Fig. 2. Note that if the alignment algorithm maps reads onto incorrect positions, those reads would provide unreliable information about the corresponding base nodes. Assuming reads of length l implies that each read node is connected to exactly l target base nodes. The bipartite graph $G(b_{1:L} \cup r_{1:n}, E)$ has L sequence base nodes and n read nodes. The edge (i,j) in the edge set E connecting b_i and r_j is associated with a unit vector $\mathbf{e}_{ij}$ indicating the type of base b_i provided by read r_j. Motivated by the iterative learning scheme and belief propagation methods [9], [10], we employ a message-passing algorithm to find consensus target sequence using overlapping reads.

The message passing algorithms rely on the exchange of messages between neighboring nodes in the graph [11]. Our algorithm operates on real-valued base messages $\{\mathbf{x}_{i\to j}\}_{(i,j)\in E}$ and read messages $\{y_{j\to i}\}_{(i,j)\in E}$. A base message $\mathbf{x}_{i\to j}$ is a 4×1 vector representing the likelihood of the base b_i being A, C, G, or T, while a read message $y_{j\to i}$ represents the reliability of read j. Read messages are initialized randomly from a Gaussian distribution, and the message update rules at iteration k are

$$\mathbf{x}_{i\to j}^{(k)} \leftarrow \sum_{j'\in\partial i\setminus j} \mathbf{e}_{ij'} y_{j'\to i}^{(k-1)}, \tag{1}$$

$$y_{j\to i}^{(k)} \leftarrow \frac{1}{l-1} \sum_{i'\in\partial j\setminus i} \mathbf{e}_{i'j}^T \mathbf{x}_{i'\to j}^{(k)}, \tag{2}$$

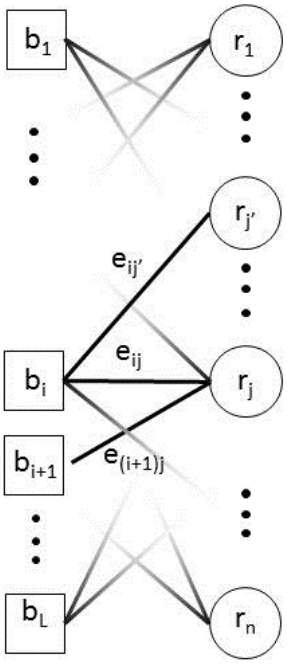

Fig. 2. *Graphical representation of the reference-guided DNA sequence consensus problem using short reads. Nodes b_i represent bases in the target DNA sequence and r_j represent reads. Each read node is connected to l base nodes.*

where ∂i and ∂j denote collection of the neighboring nodes of nodes i and j, respectively. Finally, we calculate a decision vector $\mathbf{x}_i$ as the weighted sum of the information provided by the reads weighted by each read's reliability.

$$\mathbf{x}_i = \sum_{j\in\partial i} \mathbf{e}_{ij} y_{j\to i}^{(k_{\max})}.$$

The symbol with the highest likelihood is chosen as the estimate. The estimate rule for the i^{th} base is

$$\hat{b}_i = \arg \max_{t\in\{A,C,G,T\}} \mathbf{x}_i^{\{t\}}, \tag{3}$$

where $\mathbf{x}_i^{\{t\}}$ denotes the likelihood corresponding to symbol $t \in \{A,C,G,T\}$ in vector $\mathbf{x}_i = \left[\mathbf{x}_i^{\{A\}}\ \mathbf{x}_i^{\{C\}}\ \mathbf{x}_i^{\{G\}}\ \mathbf{x}_i^{\{T\}}\right]^T$.

The algorithm is terminated when the change of reliability between subsequent iterations is small, i.e., $\sum |y_{j\to i}^{(k)} - y_{j\to i}^{(k-1)}| < \epsilon$. Note that the algorithm does not require exact knowledge of quality scores, and essentially infers reliability of individual reads.

III. Achievable performance

To evaluate performance of the proposed message passing scheme, in this section we analyze the probability of error of a genie-aided maximum a posteriori (MAP) estimator of the bases in the target sequence. In particular, we consider an idealized scenario where short reads are mapped to the reference genome with no errors and we assume that the MAP

estimator knows exact probabilities of mis-calling the bases in the short reads (i.e., has exact quality score information).

Let b_k denote the k^{th} base in the target sequence, and let $y_k^{(i)}$ denote the signal generated by sequencing b_k, $i = 1, 2, \ldots, c_k$, where c_k denotes the total number of reads covering b_k. Assume that the probability of erroneously calling b_k in the i^{th} read is $p_k^{(i)}$. Given the base calls of the reads covering b_k, $y_k^{(i)}$, the MAP estimate $\hat{b}_k$ is found as

$$\begin{aligned}\hat{b}_k &= \arg\max_x \prod_{i=1}^{c_k} P(y_k^{(i)}|b_k = x)P(b_k = x) \\ &= \arg\max_x \prod_{i=1}^{c_k} (1-p_k^{(i)})^{\delta(y_k^{(i)}=x)} (p_k^{(i)})^{(1-\delta(y_k^{(i)}=x))} \\ &= \arg\max_x \sum_{i=1}^{c_k} \delta(y_k^{(i)} = x) w_k^{(i)} + \log(P(b_k = x)), \quad (4)\end{aligned}$$

where $\delta(\cdot)$ denotes an indicator function taking value 1 if its argument is true and is 0 otherwise, and we introduced

$$w_k^{(i)} = \log\Big(\frac{1-p_k^{(i)}}{p_k^{(i)}}\Big).$$

In the absence of prior information $P(b_k = x)$, the MAP estimation of b_k in (4) is identical to the so-called weighted plurality voting [12].

For notational convenience, let us write the expression for the estimate in (4) as

$$\hat{b}_k = \arg\max_x W_k(x),$$

where $W_k(x) = \sum_{i:y_k^{(i)}=x} w_k^{(i)} + \log(P(b_k = x))$. [1]

To characterize the probability of error of the MAP consensus scheme, we rely on the so-called universal generating functions often used in reliability analysis of multi-state systems [14]. Consider n independent discrete random variables $X_1, \ldots, X_n$ with probability mass functions (pmf) represented by vectors $(\boldsymbol{x}_i, \boldsymbol{p}_i)$. In order to evaluate the pmf of an arbitrary function $f(X_1, \ldots, X_N)$, one has to find the vector $\boldsymbol{y}$ of all the possible values of $f(\cdot)$ and the vector $\boldsymbol{q}$ of the corresponding probabilities. The probability of the j^{th} combination of the realizations of the variables is $q_j = \prod_{i=1}^n p_{ij}$. If different combinations produce the same value of the function, then the probability that $f(\cdot)$ takes that value is equal to the sum of probabilities of the combinations resulting in it. We apply a variation of the above simple idea to represent the pmf of the base calling decisions using polynomial representation akin to z-transform. For brevity, we here derive the results and refer an interested reader to [14] for more background.

[1] Without a loss of generality, we will assume that ties where two different bases x and y lead to identical $W_k(x) = W_k(y)$ do not happen. The extension to this case is trivial but requires more cumbersome notation.

Consider the H-polynomial (HP) defined for each read position

$$\begin{aligned}H^i(z) &= \sum_{m=1}^4 s_m^i z^{v_m^i} \\ &= r_{x_1}^{(i)} z^{[w_i(x_1)\,0\,0\,0]} + r_{x_2}^{(i)} z^{[0\,w_i(x_2)\,0\,0]} \\ &+ r_{x_3}^{(i)} z^{[0\,0\,w_i(x_3)\,0]} + r_{x_3}^{(i)} z^{[0\,0\,0\,w_i(x_4)]}, \quad (5)\end{aligned}$$

where $r_{x_j}^i$ denotes the probability that the symbol from read i is x_j, and $w_i(x_j)$ is the associated weight. To obtain an H-polynomial of the consensus of two positions with their individual HP $H^1(z)$ and $H^2(z)$, the following composition operator can be used

$$\begin{aligned}H^{1,2}(z) &= \Omega\Big(H^1(z), H^2(z)\Big) \\ &= \Omega\Big(\sum_{m=1}^4 s_m^1 z^{v_m^1}, \sum_{m=1}^4 s_m^2 z^{v_m^2}\Big) \\ &= \sum_{m=1}^4 \sum_{n=1}^4 s_m^1 s_n^2 z^{v_m^1 + v_m^2} = \sum_m s_m^{\{1,2\}} z^{v_m^{\{1,2\}}} \quad (6)\end{aligned}$$

Note that some terms of HP $H^{1,2}(z)$ may involve the same vector. If so, in the last step in (6), probabilities corresponding to the same vector are summed up to obtain $s_m^{\{1,2\}}$. For an arbitrary subset λ, it is straightforward to obtain the HP for an extended subset $(\lambda \cup j)$ with an arbitrary $j \neq \lambda$ as

$$H^{\lambda \cup j}(z) = \Omega\Big(H^\lambda(z), H^j(z)\Big) = \sum_m s_m^{\lambda \cup j} z^{v_m^{\lambda \cup j}}. \quad (7)$$

Further reductions are obtained as follows. Consider the HP of an arbitrary weighted voting classifier (WVC) subsystem λ, $H^\lambda(z) = \sum_m s_m^\lambda z^{v_m^\lambda}$. Let $W_\Lambda = \sum_{j \in \Lambda} w_j$ be the total weight of all the votes belonging to the WVC, and let the total weight of the subsystem λ be given by W_λ. The weight not belonging to λ can be determined as

$$\sigma = \sum_{j \neq \lambda} w_j = W_\Lambda - W_\lambda.$$

If X is the maximal element of the vector v_m^λ and $v_m^\lambda(X) - v_m^\lambda(i) > \sigma$, then any element $v_m^\lambda(i) \neq X$ can be set to zero since this does not affect the probability of reliability even if all of the remaining votes is given to i. Similarly, vectors satisfying $v_m^\lambda(X) - v_m^\lambda(1) > \sigma, X \neq 1$ can be removed from further consideration. After these simplifications, the probability of correctly identifying the base is given by

$$P(\hat{b}_k = b_k) = r_k = \delta(H^\Lambda, \hat{b}_k) = \sum_{\hat{b}_k(V_m^\Lambda) = b_k} s_m^\Lambda, \quad (8)$$

where $\Lambda = \{1, 2, \ldots, c_k\}$.

Having computed the probability of error $P(\hat{b}_k \neq b_k)$ for a fixed coverage c_k (where the MAP estimator forms $\hat{b}_k$ by combining information from c_k reads that cover the k^{th} base), we can readily evaluate the probability of error for a random

coverage. In particular, the average consensus error probability P_{error} can be found by evaluating

$$P_{error} = \sum_{c_k} P(c = c_k) P(\overline{err}_c | c = c_k), \qquad (9)$$

where $\overline{err}_{c_k}$ denotes the consensus error averaged over different read positions given a fixed c_k. The probability distribution of c is often assumed to be Poisson [13], [15].

IV. Simulation Results

To demonstrate performance of the proposed algorithm, we simulated reference-guided sequence assembly of the genome of a strain of *Neisseria meningitidis*. The reference sequence is obtained from GenBank (http://www.ncbi.nlm.nih.gov/nuccore) database. The reference sequence is used to generate target sequences having 1% variation rate. We simulate short reads of length $l = 76$ (mimicking Illumina's Genome Analyzer II platform) from the target sequence by uniformly selecting read locations along the sequence. Sequencing errors are introduced into the reads according to the position-dependent base calling error profile found in [4]. The reads are then mapped to the reference sequence using an alignment algorithm based on Burrows-Wheeler transform [6]. Outcome of the mapping is used to form edges between the base and read nodes in the bipartite graph; the reads with multiple candidate mapping positions are replicated. Then the message passing algorithm from Section II is applied to obtain the consensus sequence. The parameter in the stopping criterion of the algorithm is set to $\epsilon = 0.01L$. As implied by Fig. 2, the graph has many short cycles yet in simulations the algorithm only needs ~ 20 iterations to converge. For a comparison, we also consider the plurality voting based consensus scheme often used in practice [13]. Notice that, in both message passing and plurality voting, we assume the error profiles of the reads (i.e., base calling error rates) are unknown. We also consider probability of error of the MAP consensus scheme in Section III which assumes perfect knowledge of the positions of reads along the target sequence and exact information about position-dependent base calling errors (both assumptions are unrealistic in practice). The three error rates are shown in Fig. 3 for various sequencing coverages (horizontal axis shows the average coverage). As can be seen from Fig. 3, proposed message-passing scheme outperforms plurality voting and is close to the aforementioned MAP consensus scheme.

V. Conclusion

We formulated reference-guided assembly as an inference problem on a bipartite graph and proposed a message-passing algorithm for finding consensus sequence. Unlike existing methods, the proposed algorithm finds consensus sequence without using reliability (i.e., quality scores) of the short reads. Moreover, we analyzed the probability of error of a genie-aided MAP consensus scheme in the idealized scenario where the base calling error rates and read mapping locations are known perfectly. Simulation results on a *Neisseria meningitidis* data set demonstrated that the proposed message-passing algorithm outperforms broadly used plurality voting method and is close to the genie-aided MAP consensus scheme.

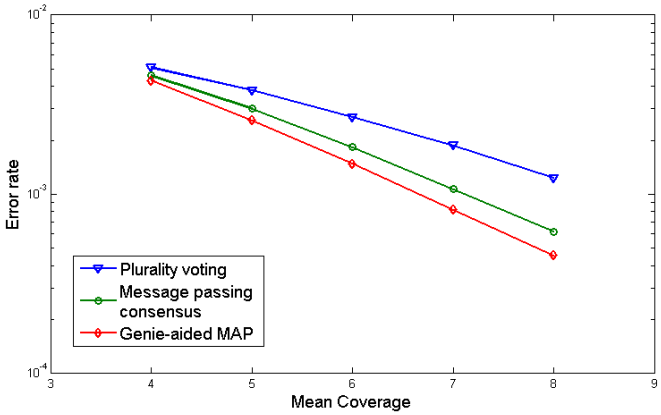

Fig. 3. *Error rates of the reconstructed target sequence and the probability of error of a genie-aided MAP consensus scheme.*

Acknowledgment

This work is funded by the National Institute of Health under grant 1R21HG006171-01. We thank Dr. Devavrat Shah for pointing out the reference [9] and useful discussions.

References

[1] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nat Biotechnology*, vol. 26, pp. 1135-1145, 2008.

[2] D. Bentley, "Whole-genome re-sequencing," *Curr Opin Genet Dev*, vol. 16, pp. 545-552, 2006.

[3] W. Kao, K. Stevens, and Y. Song, "BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing," *Genome Research*, vol. 19, pp. 1884-1895, 2009.

[4] X. Shen and H. Vikalo, "ParticleCall: A particle filter for base calling in next-generation sequencing systems," *BMC Bioinformatics*, vol. 13, July 2012.

[5] A. Aktmann, P. Weber, et al. "A beginners guide to SNP calling from high-throughput DNA-sequencing data," *Human Genetics*, pp. 1-14, 2012.

[6] B. Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, 2009.

[7] H. Li, R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, pp. 1754-1760, 2009.

[8] R. Li, Y. Li, et al. "SNP detection for massively parallel whole-genome resequencing," *Genome Research*, vol. 19: 1124-1132, 2009.

[9] D. Karger, S. Oh, D. Shah, "Iterative learning for reliable crowd-sourcing systems," *in Proceedings of NIPS*, 2011.

[10] J. Yedidia, W. Freeman, Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol 51, pp. 2282-2312, 2005.

[11] F. Kschischang, H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, 2001.

[12] X. Lin, S. Yacoub, J. Burns, and S. Simske, "Performance analysis of pattern classifier combination by plurality voting," *Pattern Recognition Letters*, vol. 24, pp. 1959-1969, 2002.

[13] W. C. Kao, A. H. Chan, and Y. S. Song, "ECHO: a reference-free short-read error correction algorithm," *Genome Research*, vol. 21, no. 7, pp. 1181-92, 2011.

[14] G. Levitin, *Universal Generating Function in Reliability Analysis and Optimization*, Springer-Verlag, 2005.

[15] G. A. Churchill and M. S. Waterman, "The accuracy of DNA sequences: Estimating sequence quality," *Genomics*, vol. 14, pp. 89-98, 1992.