Label Propagation Algorithm for Haplotype Assembly Problem

Yuying Zhao, Jinshan Li

College of Science, Beijing Forestry University, Beijing 100083, China,

zhyuying@bjfu.edu.cn

Keywords: SNP, haplotype assembly problem, label propagation algorithm

Abstract. This template explains Haplotype assembly problem is one of the most important problems in molecular biology and life sciences. Computationally, the key work for this problem is to divide the aligned sequence fragments into two disjoint clusters and then to reconstruct the corresponding haplotypes from every cluster respectively. In this work we formulate the haplotype assembly problem as network community detection problem, and then it is solved by the advanced modularity-specialized label propagation algorithm. The simulation results show the well performance of the algorithm for the haplotype assembly problem.

1. Introduction

The human genome is composed of 23 pairs of homologous chromosomes, and in each pair one chromosome is inherited from mother and the other from father. Within the chromosomes, the genetic information for humans is encoded as DNA sequences of four nucleotide bases, A, C, G and T. Human genome sequences consist of 3234.84 Mega base pairs. However, 99.5% of any two individuals' genome sequences are identical in humans. Mutation in DNA sequences is the principle factor that is responsible for the phenotypic differences among human beings, and single nucleotide polymorphisms (SNPs) are the most common mutations. A string of SNPs on a single chromosome is called haplotype. Haplotype information is essential for understanding genetic causes of various diseases and for advancement of personalized medicine [1].

However, haplotypes obtained by experimental methods is very time-consuming, labor-intensive, and expensive. Alternative computational methods for obtaining haplotypes from genotype data or SNP fragments have attracted many researchers' attention recently. The haplotype assembly problem is based on DNA fragments with SNPs and methodology of sequence assembly. High-throughput sequencing systems produce a redundant library of contiguous DNA sequence fragments from a pair of chromosomes [2]. The short sequence reads allow acquisition of the knowledge about distant SNPs on the same chromosome. Through overlapping information, all of the sequence reads can be divided into two clusters and then assembled to two haplotypes. Given a set of aligned sequence fragments, the haplotype assembly problem aims to reconstruct the haplotype sequences for each chromosome copy [3-5]. If there are no sequencing errors, the haplotype assembly problem can be resolved easily. While the sequence reads contain errors, various simplified formulations of the haplotype assembly problem have been considered, such as minimum fragment removal, minimum SNP removal, minimum error correction, and longest haplotype reconstruction etc. For these formulations of large scale, exact algorithms and various heuristic algorithms have been designed by many researchers [3-12].

In this paper, we formulate the haplotype assembly problem as network community detection problem and design a label propagation algorithm for it. The paper is organized as follows. Section 2 give some formulations of the haplotype assembly problem and the advanced modularity-specialized label propagation algorithm is designed for it. Simulation results are reported in section 3. The last section gives a brief conclusion and discussion.

© 2015. The authors - Published by Atlantis Press

2. Problem Formulations and Algorithm

2.1 Problem Formulations

In order to describe the haplotype assembly problem, some formulations must be developed first. It is assumed that we have exactly two chromosomes and two possible alleles at each SNP site. For the sake of convenience, we will denote one of them as 0, while the other one as 1. It is also assumed that the sequence fragments input to the haplotype assembly problem have been aligned and we only consider the SNP sites of these sequence fragments. That is to say, the input to the haplotype assembly problem can be denoted as a matrix M over alphabet {0, 1, -}, whose rows correspond to aligned sequence fragments and columns correspond to SNPs. In the matrix, "-" refers to the lack of information about a SNP site of some fragment reads. For convenient, we denote the i-th row of M as Mi. The output of the haplotype assembly problem is a pair of haplotypes reconstructed from the fragments matrix M. With this notation, a haplotype sequence H of one chromosome can be represented by a binary string, where we denote H = (h_1, h_2, …, h_n), and n is the length of the haplotype. The sequence fragments matrix M can be formulated as a network N = (V, E), where V is the set of vertexes and E is the set of edges. In the network, every vertex represents a sequence fragment, and an edge is connected between two vertexes if the public part of them is enough.

According to the relationships between the sequence fragments, the haplotype assembly problem can be formulated as a network community detection problem. In order to formulate the haplotype assembly problem or the weighted haplotype assembly problem as a network community detection problem, we will make some preparations as follows.

Definition1. For two values $x, y \in \{0, 1, -\}$,

$$d(x,y) = \begin{cases} 1, & if\ x \neq -, y \neq -, and \quad x \neq y \\ 0, & otherwise \end{cases} \tag{1}$$

$$c(x,y) = \begin{cases} 1, & if\ x \neq -, y \neq - \\ 0, & otherwise \end{cases} \tag{2}$$

Definition2. For two sequence fragments of length n f_1=(f_{11},f_{12},f_{13},…,f_{1n}) and f_2=(f_{21},f_{22},f_{23},..,f_{2n}), the distance between them can be denoted as $d(f_1, f_2)$, and the number of common SNP sites can be denoted as $c(f_1, f_2)$.

$$d(f_1, f_2) = \sum_{i=1}^{n} d(f_{1i}, f_{2i}) \tag{3}$$

$$c(f_1, f_2) = \sum_{i=1}^{n} c(f_{1i}, f_{2i}) \tag{4}$$

Definition3. For a cluster of sequence fragments C= {f_1, f_2,…, f_k}, the number of 0 at the i-th SNP site of all fragments in cluster C is denoted as $N_i^0(C)$, and the number of 1 at the i-th site of all fragments in cluster C is denoted as $N_i^1(C)$. A haplotype H = (h_1, h_2, ..., h_n) can be assembled from the fragments of cluster C by the following formulation.

$$h_i = \begin{cases} 1, & if N_i^1(C) > N_i^0(C) \\ 0, & if N_i^1(C) < N_i^0(C), i = 1,2,\ldots,n \\ -, & if N_i^1(C) = N_i^0(C) \end{cases} \tag{5}$$

2.2 Algorithm

A network can be designed according to the sequence fragments matrix M, and the haplotype assembly problem can be solved by detecting correct communities in the network. Experiments on networks show the advanced modularity-specialized label propagation algorithm can detect communities with high modularity quickly [13].

In order to solve the haplotype assembly problem using the advanced modularity-specialized label propagation algorithm, the problem must be formulated as a network community detection problem. The network N = (V, E) can be designed according to matrix M. In the network, every vertex represents a sequence fragment, and an edge is connected between two nodes if the public part of

them is enough. Because the sequencing error is often smaller than 5% in high-throughput sequencing systems, an edge is connected between two vertexes if $d(f_1, f_2) / c(f_1, f_2)$ is less than 15%. The advanced modularity-specialized label propagation algorithm can detect the communities of this designed network in the next step. At last, two haplotypes can be reconstructed. The advanced modularity-specialized label propagation algorithm for haplotype assembly problem can be described as follows:

Input: The sequence fragments matrix M. Every row of M is represented by a vertex. An edge is connected between two vertexes if $d(f_1, f_2) / c(f_1, f_2)$ is less than 15%.

Step1. Initially every vertex is assigned with a unique label, indicating the cluster it belongs to.

Step2. The label of every vertex is updated to a new one. This step is performed iteratively until the label of every node does not change. In the end of this step, the nodes bearing the same labels belong to the same community.

Step3. Compute the haplotypes of each community according to formula (5). If the distance between the haplotypes of two clusters is the shortest, merge them until the number of clusters is two.

Otuput: Two haplotypes assembled from the obtained two clusters.

3. Results and Discussions

To assess its performance, we tested the label propagation algorithm to a set of benchmark datasets of haplotype assembly problem. The main purpose here is to evaluate how accurately our algorithm can reconstruct two haplotypes from input sequence fragments. All the tests have been done on a Windows-xp (32 bits) desktop PC with 2.0 GHz CPU and 2GB RAM. The sequence fragments matrix M of every instance consists of 100 rows, each of which is generated by randomly copying one of the two seed haplotypes. The missing data, or the gaps, in every row are also produced randomly at missing rate Rm. The SNP error in a SNP fragment is simulated by turning 0 to 1 or vice versa at some error rate Re.

Some of the instances of the haplotype assembly problem, the seed haplotypes come from public Daly's data set. Each reconstruction rate shown in table 1 is the average over 147 instances under the same parameter setting. From table 1, it can be seen that the output haplotypes are very close to the seed haplotypes when the error rate and missing rate are relatively low. When the error rate and the missing rate are high, the reconstruction becomes a little low.

Table 1 The Reconstruction Rate on Real Seed Haplotypes

Rm \ Re	0.02	0.04	0.06	0.08	0.10
0.1	1.000	1.000	1.000	0.995	0.985
0.3	1.000	1.000	0.997	0.994	0.982
0.6	0.998	0.997	0.973	0.974	0.971
0.9	0.994	0.949	0.936	0.923	0.835

Some of the instances of the haplotype assembly problem, the seed haplotypes with SNP sites n = 100 are generated randomly. From table 2, we can see the reconstruction rate of the random seed haplotypes is a little higher than that of the real seed haplotypes. From the table, it can be seen that with the increase of missing rate and error rate, the reconstruction rate of the haplotypes declines.

Table 2 The Reconstruction Rate on Random Seed Haplotypes

Rm \ Re	0.02	0.04	0.06	0.08	0.10
0.1	1.000	1.000	1.000	0.997	0.987
0.3	1.000	1.000	0.996	0.995	0.975
0.6	0.997	0.986	0.969	0.957	0.946
0.9	0.966	0.937	0.932	0.925	0.843

4. Conclusion

Haplotype assembly problem is a difficult task and it has attracted attention of many researchers. In this paper, the haplotype assembly problem is formulated as a network community detection problem and a label propagation algorithm is proposed to solve it. To my opinion, this is the first time the haplotype assembly problem is formulated as a network community detection problem. From the simulation results, it can be seen that the label propagation algorithm performs well in many cases.

Acknowledgements

This research is financially supported by the Fundamental Research Funds for the Central Universities (No. TD 2014-03).

References

[1] A. Chakravarti. It's raining SNPs, hallelujah? Nature Genetics. vol. 19(1998) No. 3, p. 216–217.

[2] A. Edwards, H. Voss, P. Rice, A. Civitello, J. Stegemann, C. Schwager, J. Zimmermann, H. Erfle, C. T. Caskey, and W. Ansorge. Automated DNA sequencing of human HPRT locus. Genomics. Vol.6 (1990), p. 593-608.

[3] G. Lancia, V. Bafna, S. Istrail, et al. SNPs problems, complexity, and algorithms. Proceedings of the 9th Annual European Symposium on Algorithms (LNCS 2161), 2001, p. 182-193.

[4] R. Lippert, R. Schwartz, G. Lancia, et al. Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. Briefings in Bioinformatics. Vol. 3(2002), p.23-31.

[5] H. J. Greenberg, W.E. Hart, G. Lancia. Opportunities for combinatorial optimization in computational biology. Informs Journal on Computing. Vol. 14(2004), p. 211-231.

[6] R. Rizzi, V. Bafna, S. Istrail, and G. Lancia.Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem.Workshop on Algorithms in Bioinformatics (LNCS 2452), 2002, p. 29-43.

[7] R.S. Wang, L.Y. Wu, Z.P. Li, X.S. Zhang. Haplotype reconstruction from SNP fragments by minimum error correction. Bioinformatics. Vol. 21(2005), p. 2456-2462.

[8] S. Levy, G. Sutton, et al. The diploid genome sequence of an individual human. PLoS Biology. Vol. 5(2007), p. 2113-2144.

[9] M.H. Moeinzadeh, E. Asgarian. Information fusion and multiple classifiers for haplotype assembly problem from SNP fragments and related genotype. Journal of Bioinformatics and Sequence Analysis. Vol. 3 (2011), p. 63-69.

[10] Chen ZZ, Deng F, Wang L. Exact algorithms for haplotype assembly from whole-genome sequence data. Vol. 29 (2013), p.1938-1945.

[11] Fei Deng, Wenjuan Cui, Lusheng Wang. A highly accurate heuristic algorithm for the haplotype assembly problem. BMC Genomics. Vol. 14(2013), p. 52-62.

[12] Hongbo Si, Haris Vikalo, and Sriram Vishwanath. Haplotype assembly: an information theoretic view. CoRR abs/1404.0097, 2014.

[13] X. Liu, T. Murat. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. [EB/OL], http://arxiv.org/pdf/0910.1154.pdf, 2010.