Haplotype Assembly Using Rank Minimization

S. Majidian*[a], M. H. Kahaei[a]
a School of Electerical Engineering, Iran University of Science & Technology, Narmak, Tehran, Iran
s_majidian@elec.iust.ac.ir, kahaei@iust.ac.ir

Abstract: For many years, DNA sequencing was done using Sanger's method introduced in 1975. Nowdays, Next Generation Sequencing (NGS) becomes popular due to its speed and resolution [1]. In diploids, there are two homologous copies of each somatic chromosome. Homologous pairs mean that each chromosome consists of similar nucleotides sequences except a little difference. The frequent variation of a single nucleotide is named as the Single Nucleotide Polymorphism (SNP). The sequence of SNPs of each chromosome is called haplotype. The haplotype of an individual is used in drug-design and genome wide association studies [2]. Unfortunately, NGS does not provide haplotype information. Haplotype can be found using high-cost experiments or computational methods [3]. Using the NGS data for Haplotyping is known as haplotype assembly which is a NP-Hard problem. A new method for haplotype assembly is SDhaP [4] which is superior to ReFHap [2] and HapCut [5]. The core idea of SDhaP is correlation clustering. In this paper, a novel modeling for the haplotype data and NGS reads is presented based on [6]. This model results in a constraint on the rank of the read matrix. Then, the problem of haplotype assembly converted to a rank minimization optimization. Since the rank functional which is from vector space of matrices to the positive integer set, is non-convex, it can be relaxed to the nuclear norm, *i.e.* the sum of matrix singular values. This relaxation has been well mentioned as the power of convex relaxation [7-8]. Simulations on the data addressed in [9] shows that the proposed approach improves the resolution of haplotype assembly in terms of reconstruction rate compared to the new method, SDhaP [4].

Keywords: Haplotype Assembly; Next Generation Sequencing (NGS); Rank minimization.

References

[1] E. Hayden, "Technology: the $1,000 genome." Nature 507.7492 (2014), 294-295.

[2] J. Duitama , *et. al.* "Fosmid-based whole genome haplotyping of a hapmap trio child: evaluation of single individual haplotyping techniques." Nucleic Acids Res. 2011;40(5):2041–53.

[3] M. Snyder, *et. al.* "Haplotype-resolved genome sequencing: experimental methods and applications." Nature Reviews. Genetics 16.6 (2015): 344.

[4] S. Das, and H. Vikalo. "SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming." BMC genomics 16.1 (2015): 260.

[5] V. Bansal, V. Bafna "Hapcut: an efficient and accurate algorithm for the haplotype assembly problem." Bioinformatics.24,16 (2008)153–9.

[6] C. Cai, S. Sanghavi, and H. Vikalo, "Structured low-rank matrix factorization for haplotype assembly." IEEE Journal of Selected Topics in Signal Processing, 10.4 (2016) 647-657.

[7] B. Recht, M. Fazel, and P. Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." SIAM review 52.3 (2010): 471-501.

[8] E. Candès, and T. Tao. "The power of convex relaxation: Near-optimal matrix completion." IEEE Transactions on Information Theory 56.5 (2010): 2053-2080.

[9] F. Geraci, "A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem." Bioinformatics 26.18 (2010): 2217-2225.