NCMHap: A novel method for haplotype reconstruction based on Neutrosophic c-means clustering

Fatemeh Zamani[1], Mohammad Hossein Olyaee[2], Alireza Khanteymoori[1,*]

[1]Department of Computer Engineering, University of Zanjan, Zanjan, Iran
[2]Faculty of Engineering, Department of Computer Engineering, University of Gonabad, Gonabad, Iran

Abstract

Background: Single individual haplotype (SIH) problem refers to reconstructing haplotypes of an individual based on several input fragments sequenced from a specified chromosome. Solving this problem is an important task in computational biology and has many applications in the pharmaceutical industry, clinical decision-making and genetic diseases.

Objective: It is known that solving the problem is NP-hard. Although several methods have been proposed to solve the problem, but it is found that most of them have low performances in dealing with noisy input fragments. Therefore, proposing a method which be accurate and scalable, is a challenging task.

Methods: In this paper, we introduced a method, named NCMHap, which utilizes the Neutrosophic c-means (NCM) clustering algorithm. The NCM algorithm can effectively detect the noise and outliers in the input data. In addition, it can reduce their effects in the clustering process.

Results: The proposed method has been evaluated by several benchmark datasets. Comparing with existing methods indicates that NCMHap is significantly superior in the most cases, particularly when the amount of noise increases, it outperforms the comparing methods.

Conclusion: The experimental results recommend the application of the proposed method on the datasets which involve the fragments with huge amount of gaps and noise.

Keywords Bioinformatics, Haplotype assembly, Minimum error correction, Neutrosophic c-means clustering.

1- Introduction

It has been revealed that the human genome shows some degrees of inter-individual and inter-population variations which make it an appropriate target to rigorous functional genomic analysis [1, 2]. Recent cost-effective next generation sequencing (NGS) technologies have provided a huge amount of genome sequences of individual human [3]. It has been discovered that more than 99% of human genomes are completely identical. Therefore, it turns out that the vast differences among people can potentially emerged from the less than 1% variations [4, 5]. Single nucleotide polymorphisms (SNPs) refer to the genetic variations which are more frequent. A sequence of SNPs which co-occur in a specific chromosome is named as haplotype. In diploid species like human, there are two copies of each chromosome. Since each haplotype is derived from a copy of specific chromosome, as a result, there are two copies of haplotypes.

Haplotype provide more attainable information than individual SNPs which can be remarkable

for investigating the relation between genetic variations and complex diseases [6], studying of human history [7], providing personalized medicine [8] and studying biological mechanisms [9].

In spite of the fact that obtaining the haplotypes is an important task, but direct experimental analysis of haplotypes is labor-intensive, expensive, and it is restricted to obtaining the local haplotypes. In practice, the human haplotypes are provided as sequencing reads (fragments). Assuming the importance of detecting genetic variations accompanied by limitations over molecular approaches, obtaining haplotype information from these numerous fragments may have profound effects in different aspects of medicine and molecular biology [10-13]. Availability of the fragments make it possible to assemble haplotypes in a process referred to as single individual haplotyping (SIH) [14] which is performed by *in silico* (computer aided) analysis using statistical and computational approaches.

For this purpose, the requested region of the specified chromosome is sequenced several times and a number of fragments are provided. Due to the limitations of sequencing methods, the fragments involve errors and gaps. It should be noted that the former derived from wrong determination of allele's measure; while, the latter is related to the low-confidence measures of allele positions. SIH attempts to assign each fragment to the right chromosome copy. Then, it detects and corrects the errors to reconstruct the desired haplotypes. In order to solve this problem, several models have been proposed which minimum SNP removal (MSR), minimum fragment removal (MFR), and minimum error correction (MEC) are the chief models. Among the existing models, MEC is more efficient and has been applied in several approaches [15-18]. The aim of this model is finding and correcting the errors by applying the minimum letter changes in the input fragments. It has been proved that all of the models are NP-Hard [14]. Most of the current methods construct a weighted graph such that each fragment corresponds with a vertex and the weight of each edge represents the amount of similarity between the connecting fragments. Based on the used model, each method transforms the built graph into a bipartite graph. For example in MEC model, this is performed by deleting the least number of conflicting edges. AROHap [18] and FCMHap [19] are two recently methods which have been addressed the problem according the MEC model. The first, through the use of asexual reproduction optimization (ARO) algorithm attempts to improve the fitness function which is designed based on MEC model. The second, by exploiting Fuzzy c-means (FCM) clustering algorithm tries to improve the initial haplotypes iteratively. It is worthwhile noting that the method divides the input fragments into two groups and the haplotypes are obtained as center of the clusters. However, some popular methods such as MCMC [20] and HapCUT [15] build the graph in a different way. These methods start with a set of arbitrary sequences as initial haplotypes and improve it step by step regarding the input fragments. They make a similar weighted graph in their distinctive model; but instead of fragments, SNPs are the vertices. Each pair of SNPs is connected if they are covered by at least one input fragments. The weight of each edge describes the amount of consistency with their corresponding positions in the current haplotypes. Albeit, this model efficiently describes the consistency of the current haplotype with the input fragments; but the existence of gaps and noise may lead to achieving inaccurate weights [21].

In this paper, we propose a fast and accurate method to solve haplotype reconstruction named NCMHap which involves two steps. First, a weighted fuzzy conflict graph is made such that each node corresponds with an input fragment and the weight of each edge represents the measurement of incompatibility between the corresponding input fragments. By removing the least of conflicting edges based on the MEC model and bi-partitioning the input fragments, an initial fragment clustering is obtained. Next, to decrease the effect of noise and outliers on the obtained clusters, Neutrosophic c-means (NCM) clustering method is applied. NCM by assigning a coefficient to each input fragment can reduce the noise effects on the clustering process. According to the experimental results, NCMHap can provide high throughput reconstructed haplotypes close to the optimal.

In the reminder of this paper, section 2 recalls the problem formulation. Section 3 provides a brief review on NCM algorithm. The details of the proposed method are described in section 4. Section 5 presents the experimental results. Finally, section 6 concludes this paper.

2- Problem formulation

As can be seen in Fig. 1, $X_{m\times n}$ is a SNP matrix where each row corresponds with an input fragment with length n. Since in most cases, there are two alleles at each SNP site, for simplicity, the major and minor alleles are represented by 0 and 1 respectively. It should be noted that if a SNP value cannot be determined with enough confidence, it is indicated by '-'.

Let f_i and f_j are two arbitrary input fragments. The Hamming distance (*HD*) can describe their similarity as below:

$$HD(f_i, f_j) = \sum_{k=1}^{n} D(f_{ik}, f_{jk}) \tag{1}$$

$$D(a,b) = \begin{cases} 1 & if\ a,b \neq' -' and\ a \neq b \\ 0 & else \end{cases} \tag{2}$$

Where f_i and f_j are compatible if $HD = 0$, else they are in conflict. In other words, when $HD(f_i.f_j)$ equals zero, it can be concluded that these fragments are originated from the same chromosome copy, otherwise the fragments belong to different chromosome copy or some of their positions are destroyed by noise. To solve the problem, the fragments of the SNP matrix must be divided into two clusters such that the elements of each cluster will be compatible by minimum number of letter flips i.e. MEC measure is minimized. Then, the center of each cluster equals with its corresponding haplotype. Fig. 1, demonstrates the haplotype reconstruction in the diploid genome, *X* is SNP matrix which divided into two parts and $H = \{h_1, h_2\}$ involves the reconstructed haplotypes of each clusters**.**

In order to evaluate the quality of the obtained haplotypes, reconstruction rate (RR) and MEC score are two useful measurements. Let $\widehat{H}$ and H contain the reconstructed haplotypes and the original haplotypes respectively. The RR describes the similarity between $\widehat{H}$ and H that it is computed as below.

$$RR_{(\hat{H}.H)} = 1 - \frac{min\left(HD(\hat{h}_1,h_1)+HD(\hat{h}_2,h_2).HD(\hat{h}_1,h_2)+HD(\hat{h}_2,h_1)\right)}{2n} \quad (3)$$

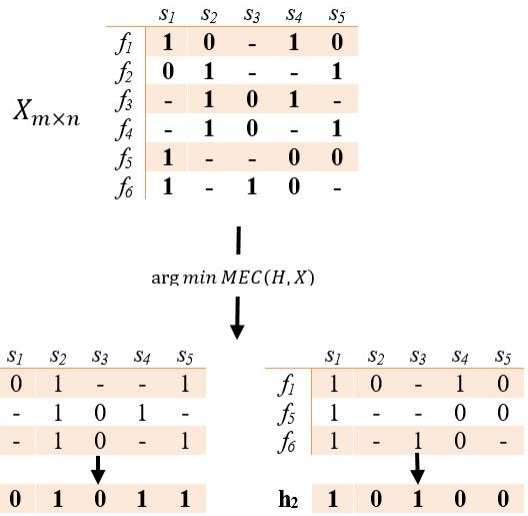

Fig. 1. An example of haplotype reconstruction using the MEC model [22].

3- Neutrosophic c-means (NCM) algorithm

As stated previously, fragment clustering is an important phase of the haplotype assembly. Also, huge amount of noise and gaps in the input fragments have made this phase as a challenging task. In order to perform this phase efficiently, we consider the Neutrosophic c-means (NCM) clustering algorithm. The algorithm computes the degrees belonging to the determinant and indeterminate clusters at the same time for each of the data points [23] [24]. Outlier and noise data are considered as Indeterminate clusters. Therefore, the NCM algorithm can detect outliers and noisy data. Also, by using some relevant functions, it can decrease the undesirable effects of noise and outliers on the clustering process. For this purpose, the NCM algorithm minimizes the objective function given in Eq. (4) through an iterative process, whereby the centers of the clusters are determined with the least error and the clustering accuracy is improved.

$$J(T.I.F.C) = \sum_{i=1}^{N}\sum_{j=1}^{C}\left(w_1 T_{ij}\right)^m \left\|x_i - c_j\right\|^2 + \sum_{i=1}^{N}(w_2 I_i)^m \left\|x_i - \bar{c}_{i\,max}\right\|^2 + \sum_{i=1}^{N}\delta^2 (w_3 F_i)^m \quad (4)$$

$$\bar{c}_{i\,max} = \frac{C_{pi}C_{qi}}{2} \quad (5)$$

$$p_i = \underset{j=1,2,\cdots,C}{\arg\max}\left(T_{ij}\right) \quad (6)$$

$$q_i = \underset{j\neq p_i \cap j=1,2,\cdots,C}{\arg\max}\left(T_{ij}\right) \quad (7)$$

In the above relations, T_{ij} is defined as the degree to determinant clusters, I_{i} is the degree to the boundary clusters, F_{i} is the degree belonging to the outlier data set, N number of data, C number

of clusters, w weighting factor, m is a constant, x_i is data point ,and δ is the number of objects that are considered as outliers. $\bar{C}_{\mathrm{i\,max}}$ is a constant that is calculated for each data point according Eq. (8). This parameter is used to precisely determine the value of function I_{i}, because the degree of indeterminacy of each data depends on the two largest definite clusters close to it, namely Eqs. (6) and (7). The cluster centers c_j and membershipT_{ij}, I_{i}, and F_{i} and are updated by Eqs. (8) respectively.

$$c_j = \frac{\sum_{i=1}^{N} (w_1 T_{ij})^m x_i}{\sum_{i=1}^{N} (w_1 T_{ij})^m} \tag{8}$$

$$T_{ij} = \frac{K}{w_1} \left(x_i - c_j\right)^{-(2/m-1)} \tag{9}$$

$$I_i = \frac{K}{w_2} (x_i - \bar{c}_{i\,max})^{-(2/m-1)} \tag{10}$$

$$F_i = \frac{K}{w_3} \delta^{-(2/m-1)} \tag{11}$$

4- Proposed method

As can be seen in Fig. 2, the proposed method involves two main steps. First, in order to provide an initial clustering of the input fragments, a weighted graph, called fuzzy conflict graph, is constructed based on the SNP matrix. In this graph, fragments are considered as vertices, and the weight of each edge is the normalized distance between corresponding fragments. This measure is given as follows:

$$\widehat{D}(f_i, f_j) = \frac{1}{S_{ij}} \sum_{k=1}^{n} \hat{d}(f_{ik}, f_{jk}) \tag{12}$$

In the above relations, f_i and f_j are two fragments of X, S_{ij} denotes the number of columns (SNPs) that are covered by either f_{ik} or f_{jk} in X. In fact, S_{ij} is a normalization factor that allows us to normalize the distance between the two fragments such that the resulting distance ranges from 0 to 1, and n represents the number of SNPs.

After constructing the graph, the edges with weight of 0.5 are removed because they do not provide sufficient information about the clustering of the connected fragments.

In the second phase, the initial clustering is given to the NCM algorithm. The centers of each cluster are considered as the primary centers in the NCM algorithm. Initial clustering can improve the convergence speed of the NCM algorithm. This algorithm determines the impact of fragments on clustering based on the three membership functions introduced and is able to reduce the impact of noise or outliers on the clustering process and consequently, the accuracy of clustering will be increased. Therefore, clustering is achieved by repeating the optimal objective function and the membership degree of the determinant and indeterminate clusters and the centers of the clusters in each iteration will be updated by Eqs. (8-11). The iteration is repeated until the difference between cluster centers at two successive iterations is greater than ε. Finally, the center of obtained clusters

construct the set of reconstructed haplotypes.

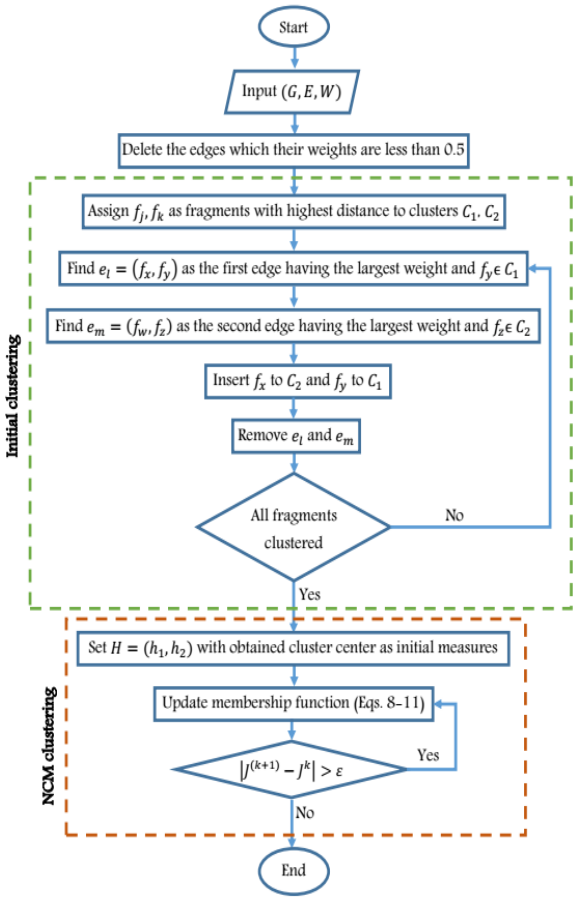

Fig. 2- Flowchart of the proposed method

5- Results

5.1- Setting the parameters

The proposed method was implemented in MATLAB and all experiments were completed on a Core i5 Intel with 2.7 GHz and 8G RAM. ε, δ, w_1, w_2, and w_3 are the main parameters which are set as 10^{-5}, 25, 0.7, 0.2, and 0.1, respectively. Similar to the pervious works [15, 18, 21, 25-28], RR measure is used to evaluate the quality of the obtained haplotypes.

5.2- Competitor methods

In this experiment, NCMhap is compared with a set of state-of-the-art and well-known methods. Some important notes about these competitors are described as follows:

- H-PoP [27] clusters the DNA reads into k groups such that the elements of each cluster have minimum distance with each other while are far from the reads of the other clusters. Moreover, it exploits the genotype information to improve the reconstructed haplotypes.
- SCGD [29] is a heuristic based method which models SIH as the low-rank matrix factorization problem and represents a modified of the gradient descent algorithm to solve the problem.
- Fast [26] is an iterative based method which models the similarities between the input

fragments with a weighed fuzzy conflict graph.

- FCM [19] uses Fuzzy C-means clustering method to divide the input fragments into two segments with minimum MEC measure.
- HG [21] exploits hypergraph model to describe the similarities between the input fragments more precisely.
- ARO [18] is a nature inspired method which utilizes Asexual Reproduction optimization method to cluster the input fragments with the best MEC score.
- ALT [28] is an iterative algorithm which formulates the haplotype assembly problem as a sparse tensor decomposition.
- HRCH [30] utilizes a chaotic viewpoint to reconstruct haplotypes. For this aim, the obtained haplotypes are mapped to some coordinate series by applying chaos game representation. Then, the positions with low confidences are improved by using a local projection.

5.3- Simulation results

In order to evaluate the performance of the proposed method, first, the experiments have been carried out on a widely used dataset named as Geraci's dataset [31] . It was provided by international Hapmap project which is based on 22 chromosomes of 269 different individuals.

The individuals have been nominated from Japan (JPT), China (HCB), Nigeria (YR) and Utah (CEU). Haplotype length (l), coverage (c), and error rate (e) are the main parameters which $l = \{100,350,700\}$, $c = \{3,5,8,10\}$ and $e = \{0.1,0.2,0.3\}$. It should be noted that for each combination of these parameters there are 100 instances.

Since the proposed method involves two steps, it can be desired to evaluate the influence of each step independently. For this purpose, the initial clustering, NCM algorithm, and NCMHap are separately executed on the Geraci's dataset. The obtained results for haplotypes with length 100, 350, and 700 are listed in Tables 1-3 respectively. It should be noted that the first two columns in these tables are the error rate e and the coverage c, respectively. In each table, The NCM column represents the results when it starts with a random initial guess for each cluster center.

It can be seen in the last column of Tables 1-3, the synergistic of these steps achieved the promising results which completely outperforms the other cases.

Table 1. The average reconstruction rate over 100 instances with length 100

e	c	Initial	NCM	NCMhap
0.1	3	0.657	0.817	0.916
	5	0.677	0.846	0.971
	8	0.676	0.946	0.983
	10	0.675	0.885	0.989
0.2	3	0.611	0.693	0.822
	5	0.620	0.730	0.907
	8	0.616	0.793	0.931
	10	0.633	0.826	0.936
0.3	3	0.554	0.581	0.684
	5	0.568	0.677	0.759

	8	0.565	0.653	0.816
	10	0.564	0.692	0.843

Table 2. The average reconstruction rate over 100 instances with length 350

e	c	Initial	NCM	NCMhap
0.1	3	0.639	0.688	**0.953**
	5	0.655	0.718	**0.982**
	8	0.664	0.805	0.989
	10	0.665	0.812	0.993
0.2	3	0.586	0.632	**0.856**
	5	0.600	0.678	**0.921**
	8	0.610	0.720	**0.939**
	10	0.619	0.703	**0.948**
0.3	3	0.527	0.580	**0.712**
	5	0.539	0.583	**0.803**
	8	0.540	0.590	**0.850**
	10	0.547	0.611	**0.870**

Table 3. The average reconstruction rate over 100 instances with length 700

e	c	Initial	NCM	NCMhap
0.1	3	0.635	0.704	**0.958**
	5	0.656	0.761	0.984
	8	0.647	0.745	**0.990**
	10	0.657	0.746	0.994
0.2	3	0.584	0.634	**0.865**
	5	0.607	0.624	**0.925**
	8	0.598	0.714	**0.938**
	10	0.599	0.708	**0.946**
0.3	3	0.520	0.545	**0.720**
	5	0.538	0.569	**0.808**
	8	0.535	0.590	**0.849**
	10	0.545	0.618	**0.958**

Table 4-6 demonstrate the RRs obtained from the run of the NCMHap as well as the benchmarking algorithms on Geraci's dataset for haplotypes with length 100, 350, and 700 respectively. In each table, the first two columns are error rate and coverage measure respectively. It should be noted that each element of these tables represents the average over 100 data samples. Also, the bold and gray values in the last column of each table represents the first and second best reconstruction rates, respectively.

By investigating the results of Table 4, it reveals that the proposed method can provide high quality results and completely comparable against the other approaches. Comparing the results demonstrates that the proposed method completely outperforms SCGD, FastHap, FCMHap ,and AROHap algorithms in all parameters.

As can be seen in Table 5, by increasing the length of fragments, the quality of the obtained haplotypes is efficiently improved. Particularly, when the amount of noise is increased, it can preserve the quality of reconstructed haplotypes against the other approaches and in most cases outperforms the benchmarking methods.

Table 4- Performance comparison of NCMHap and other methods on the Geraci's dataset [31] with haplotype block length *l = 100*.

e	C	SCGD	H-pop	Fast	FCM	HG	ARO	ALT	HRCH	NCMHap
0.1	3	0.918	0.921	0.823	0.882	0.941	0.844	0.944	0.957	0.916
	5	0.944	0.919	0.917	0.948	0.989	0.922	0.953	0.987	0.971
	8	0.948	0.900	0.955	0.971	0.994	0.945	0.945	0.991	0.983
	10	0.959	0.892	0.926	0.972	0.997	0.92	0.943	0.995	0.989
0.2	3	0.806	0.836	0.806	0.739	0.752	0.711	0.831	0.851	0.822
	5	0.825	0.865	0.834	0.772	0.899	0.736	0.865	0.926	0.907
	8	0.861	0.873	0.849	0.793	0.966	0.760	0.873	0.941	0.931
	10	0.886	0.878	0.899	0.835	0.981	0.788	0.878	0.956	0.936
0.3	3	0.671	0.717	0.578	0.629	0.621	0.627	0.694	0.695	0.684
	5	0.676	0.784	0.711	0.648	0.698	0.638	0.780	0.798	0.759
	8	0.740	0.835	0.700	0.664	0.79	0.649	0.841	0.861	0.816
	10	0.798	0.855	0.732	0.675	0.856	0.653	0.857	0.881	0.843

Table 5- Performance comparison of NCMHap and other methods on the Geraci's dataset [31] with haplotype block length *l = 350*.

e	C	SCGD	H-pop	Fast	FCM	HG	ARO	ALT	HRCH	NCMHap
0.1	3	0.941	0.921	0.872	0.873	0.939	0.844	0.943	0.939	**0.953**
	5	0.945	0.912	0.927	0.919	0.979	0.892	0.951	0.981	**0.982**
	8	0.950	0.896	0.977	0.934	0.988	0.908	0.930	0.991	0.989
	10	0.952	0.889	0.947	0.935	0.995	0.910	0.941	0.994	0.993
0.2	3	0.813	0.813	0.763	0.671	0.712	0.659	0.849	0.813	**0.856**
	5	0.817	0.860	0.811	0.719	0.905	0.691	0.896	0.897	**0.921**
	8	0.832	0.871	0.912	0.728	0.899	0.709	0.908	0.922	**0.939**
	10	0.838	0.873	0.923	0.733	0.907	0.719	0.913	0.937	**0.948**
0.3	3	0.637	0.629	0.575	0.597	0.602	0.595	0.664	0.640	**0.712**
	5	0.661	0.744	0.720	0.614	0.632	0.609	0.777	0.737	**0.803**
	8	0.690	0.830	0.790	0.626	0.675	0.628	0.838	0.788	**0.850**
	10	0.700	0.850	0.833	0.631	0.742	0.635	0.856	0.821	**0.870**

Table 6- Performance comparison of NCMHap and other methods on the Geraci's dataset [31] with haplotype block length *l = 700*

e	C	SCGD	H-pop	Fast	FCM	HG	ARO	ALT	HRCH	NCMHap
0.1	3	0.934	0.919	0.917	0.834	0.934	0.801	0.941	0.928	**0.958**
	5	0.951	0.923	0.872	0.881	0.990	0.862	0.951	0.972	0.984
	8	0.956	0.945	0.945	0.883	0.987	0.899	0.943	0.983	**0.990**
	10	0.973	0.951	0.983	0.996	0.997	0.912	0.942	0.992	0.994
0.2	3	0.796	0.811	0.703	0.652	0.677	0.644	0.852	0.797	**0.865**
	5	0.829	0.854	0.681	0.672	0.910	0.662	0.896	0.869	**0.925**
	8	0.832	0.868	0.916	0.686	0.884	0.695	0.905	0.885	**0.938**
	10	0.860	0.869	0.896	0.746	0.894	0.698	0.909	0.900	**0.946**
0.3	3	0.652	0.600	0.627	0.592	0.592	0.588	0.674	0.602	**0.720**
	5	0.659	0.733	0.682	0.599	0.621	0.598	0.735	0.699	**0.808**
	8	0.662	0.804	0.741	0.606	0.646	0.613	0.793	0.729	**0.849**
	10	0.714	0.844	0.805	0.606	0.696	0.618	0.829	0.759	**0.870**

Finally, as reported in Table 6, for input fragments with length 700, except one situation, NCMHap has achieved better reconstruction rates than any other algorithms.

Since the Neutrosophic c-means clustering is a developed form of Fuzzy c-means method and

moreover NCMHap like FastHap method uses weighted fuzzy conflict graph to model the similarity between the input fragments, its performance is compared against FCMhap and FastHap approaches when it deals with long block haplotypes and a huge amount of noise. Fig. 3 demonstrates the quality of obtained results for haplotypes with length 700 and error rate $e \geq 0.2$.

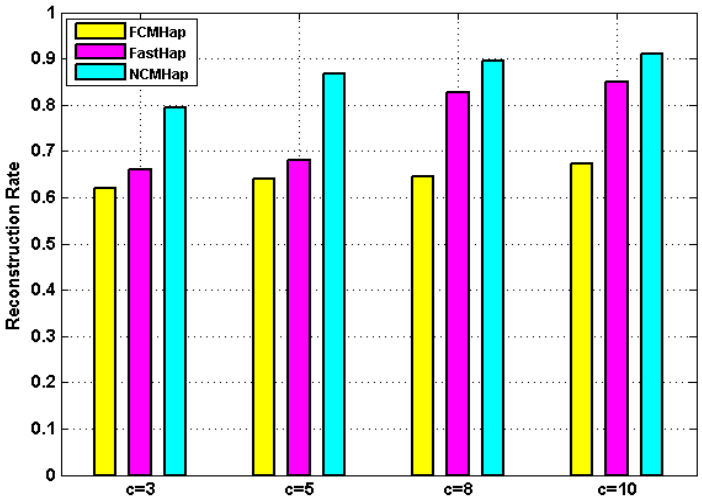

Fig. 3 Comparison the reconstruction rate of the proposed method against FastHap and FCMHap methods while $e \geq 0.2$

It is apparent the results of the proposed method are valuable against comparing methods in dealing with input fragments with high error rate.

5.4- Experimental results

For more investigation, we tested the performance of our method on a real dataset which involves data provided by 1000 genome project. This data belongs to an individual NA12878 which frequently is used to investigate the performance of the existing SIH methods. Moreover, the trio-phased variant calls from the GATK resource bundle[32] was used as the true haplotypes. The reconstruction rate of the proposed method is compared to those of H-PoP [27],SCGD[29],Fast[26],HG[21],ARO[18], ALT[28] ,and HRCH[30] in Table7. The obtained results demonstrate that our method achieves the highest and second highest RRs for most of the chromosomes.

Chr	H-pop	SCGD	FastHap	HGHap	AROHap	FCMHap	ALTHap	HRCH	NCMHap
1	0.957	0.925	0.919	0.937	0.935	0.913	0.974	0.954	0.972
2	0.956	0.926	0.922	0.929	0.943	0.908	0.953	0.943	**0.959**
3	0.912	0.919	0.923	0.928	0.94	0.913	0.933	0.944	**0.969**
4	0.970	0.927	0.933	0.923	0.949	0.923	0.969	0.96	0.961
5	0.966	0.939	0.914	0.932	0.942	0.912	0.972	0.952	0.957
6	0.952	0.930	0.938	0.935	0.948	0.929	0.949	0.958	**0.977**
7	0.924	0.935	0.921	0.925	0.951	0.904	0.970	0.954	**0.971**
8	0.947	0.907	0.906	0.906	0.934	0.903	0.962	0.949	0.950
9	0.910	0.971	0.940	0.901	0.966	0.937	0.971	0.921	0.956

10	0.945	0.926	0.923	0.940	0.945	0.913	0.968	0.954	0.956
11	0.915	0.932	0.931	0.939	0.942	0.923	0.933	0.963	**0.964**
12	0.903	0.923	0.923	0.945	0.935	0.908	0.921	0.954	**0.963**
13	0.941	0.970	0.941	0.930	0.935	0.925	0.970	0.946	0.965
14	0.971	0.911	0.934	0.917	0.934	0.932	0.903	0.949	0.970
15	0.974	0.991	0.917	0.920	0.937	0.905	0.972	0.951	0.959
16	0.935	0.930	0.932	0.932	0.946	0.924	0.967	0.962	**0.973**
17	0.911	0.967	0.944	0.931	0.951	0.920	0.975	0.963	0.973
18	0.976	0.903	0.926	0.924	0.949	0.919	0.910	0.954	0.973
19	0.978	0.972	0.930	0.949	0.942	0.923	0.976	0.960	0.968
20	0.950	0.968	0.931	0.945	0.946	0.922	0.973	0.957	0.971
21	0.970	0.943	0.919	0.933	0.941	0.915	0.974	0.960	0.960
22	0.983	0.941	0.926	0.951	0.941	0.914	0.973	0.964	0.976

Evaluating the obtained results on the both simulation and experimental datasets demonstrates that the proposed method can provide promising reconstructed haplotypes in dealing with low quality sequencing data. Moreover, in the worst case, NCMHap can solve the problem in the less than 3 minutes which this runtime is favorable against the existing approaches.

6- Conclusion

In this paper, we presented a method based on the Neutrosophic c-means (NCM) clustering algorithm for haplotype assembly problem. Time complexity and the handling high error rate datasets are the main challenges of the existing methods. Due to improve the NCM's convergence speed, the proposed method consists of two phases. First, the input fragments are divided into two partitions based on their similarities. Second, information of bi-partitioning is employed as initial centers by NCM clustering method. Applying the information in NCM can improve the speed of convergence and decrease number of iterations. Experimental results display that the proposed method provides high efficiency to reconstruct haplotypes with a high-error-rate.

As demonstrated in a series of recent publications (see, e.g., [21, 33-36]) in developing new prediction methods, user friendly and publicly accessible web-servers will significantly enhance their impacts [37], we shall make efforts in our future work to provide a web-server for the prediction method presented in this paper. Also, the source code of NCMHap is freely available at https://github.com/FatemehZamani/NCMHap.git.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Competing Interests

The authors declare no competing interests in relation to this study.

Funding

No funding.

Authors' contributions

A.R.K., M.H.O. and F.Z. designed the research, F.Z. and M.H.O. collected data, F.Z. and M.H.O. wrote and performed computer programs, A.R.K., M.H.O. and F.Z. analyzed and interpreted the results, F.Z. and M.H.O. wrote the first version of the manuscript, A.R.K. and M.H.O. revised and edited the manuscript.

References

1. Jorde LB, Wooding SP: Genetic variation, classification and'race'. *Nature genetics* 2004, 36(11s):S28.
2. Schneider JA, Pungliya MS, Choi JY, Jiang R, Sun XJ, Salisbury BA, Stephens JC: DNA variability of human genes. *Mechanisms of ageing and development* 2003, 124(1):17-25.
3. Snyder MW, Adey A, Kitzman JO, Shendure J: Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics* 2015, 16(6):344-358.
4. Hoehe MR, Köpke K, Wendel B, Rohde K, Flachmeier C, Kidd KK, Berrettini WH, Church GM: Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence. *Human molecular genetics* 2000, 9(19):2895-2908.
5. Terwilliger JD, Weiss KM: Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* 1998, 9(6):578-594.
6. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ: The importance of phase information for human genomics. *Nature Reviews Genetics* 2011, 12(3):215.
7. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y: A draft sequence of the Neandertal genome. *science* 2010, 328(5979):710-722.
8. Shastry BS: SNPs and haplotypes: genetic markers for disease and drug response. *International journal of molecular medicine* 2003, 11(3):379-382.
9. Adey A, Burton JN, Kitzman JO, Hiatt JB, Lewis AP, Martin BK, Qiu R, Lee C, Shendure J: The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* 2013, 500(7461):207.
10. Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB: Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nature genetics* 2001, 28(4):361.
11. Liu N, Zhang K, Zhao H: Haplotype-association analysis. *Advances in genetics* 2008, 60:335-405.
12. Ruano G, Kidd KK: Direct haplotyping of chromosomal segments from multiple heterozygotes via allele-specific PCR amplification. *Nucleic acids research* 1989, 17(20):8392.
13. Ruano G, Kidd KK, Stephens JC: Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proceedings of the National Academy of Sciences* 1990, 87(16):6296-6300.
14. Lancia G, Bafna V, Istrail S, Lippert R, Schwartz R: SNPs problems, complexity, and algorithms. In: *European symposium on algorithms: 2001*. Springer: 182-193.
15. Bansal V, Bafna V: HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* 2008, 24(16):i153-i159.
16. Qian W, Yang Y, Yang N, Li C: Particle swarm optimization for SNP haplotype reconstruction problem. *Applied mathematics and Computation* 2008, 196(1):266-272.
17. Wang T-C, Taheri J, Zomaya AY: Using genetic algorithm in reconstructing single individual haplotype with minimum error correction. *Journal of biomedical informatics* 2012, 45(5):922-930.

18. Olyaee M-H, Khanteymoori A: AROHap: An effective algorithm for single individual haplotype reconstruction based on asexual reproduction optimization. *Computational biology and chemistry* 2018, 72:1-10.
19. Olyaee MH, Khanteymoori A: Fuzzy c-means clustering for SNP haplotype reconstruction problem.
20. Bansal V, Halpern AL, Axelrod N, Bafna V: An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome research* 2008, 18(8):1336-1346.
21. Chen X, Peng Q, Han L, Zhong T, Xu T: An effective haplotype assembly algorithm based on hypergraph partitioning. *Journal of theoretical biology* 2014, 358:85-92.
22. Rhee J-K, Li H, Joung J-G, Hwang K-B, Zhang B-T, Shin S-Y: Survey of computational haplotype determination methods for single individual. *Genes & Genomics* 2016, 38(1):1-12.
23. Guo Y, Sengur A: NCM: Neutrosophic c-means clustering algorithm. *Pattern Recognition* 2015, 48(8):2710-2724.
24. Akbulut Y, Şengür A, Guo Y, Polat K: KNCM: Kernel neutrosophic c-means clustering. *Applied Soft Computing* 2017, 52:714-724.
25. Berger E, Yorukoglu D, Peng J, Berger B: Haptree: A novel bayesian framework for single individual polyplotyping using ngs data. *PLoS computational biology* 2014, 10(3):e1003502.
26. Mazrouee S, Wang W: FastHap: fast and accurate single individual haplotype reconstruction using fuzzy conflict graphs. *Bioinformatics* 2014, 30(17):i371-i378.
27. Xie M, Wu Q, Wang J, Jiang T: H-PoP and H-PoPG: Heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics* 2016, 32(24):3735-3744.
28. Hashemi A, Zhu B, Vikalo H: Sparse tensor decomposition for haplotype assembly of diploids and Polyploids. *BMC genomics* 2018, 19(4):191.
29. Cai C, Sanghavi S, Vikalo H: Structured low-rank matrix factorization for haplotype assembly. *IEEE Journal of Selected Topics in Signal Processing* 2016, 10(4):647-657.
30. Olyaee MH, Khanteymoori AR, Khalifeh K: A chaotic viewpoint-based approach to solve haplotype assembly using hypergraph model. *IEEE Access* 2020.
31. Geraci F: A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem. *Bioinformatics* 2010, 26(18):2217-2225.
32. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M: A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 2011, 43(5):491.
33. Liu Z, Xiao X, Qiu W-R, Chou K-C: iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Analytical biochemistry* 2015, 474:69-77.
34. Jia J, Liu Z, Xiao X, Liu B, Chou K-C: iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. *Journal of theoretical biology* 2015, 377:47-56.
35. Ding H, Deng E-Z, Yuan L-F, Liu L, Lin H, Chen W, Chou K-C: iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed research international* 2014, 2014.
36. Chen W, Feng P-M, Deng E-Z, Lin H, Chou K-C: iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry* 2014, 462:76-83.
37. Jia J, Liu Z, Xiao X, Liu B, Chou K-C: iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. *Molecules* 2016, 21(1):95.