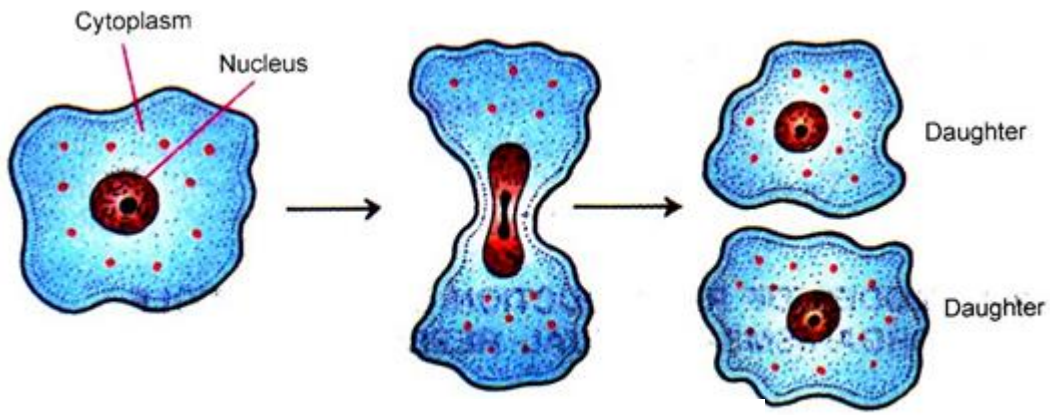


Asexual reproduction optimization(ARO)

Weaknesses of GA

- Several parameters should be tuned by user
  - Size of population
  - Pc
  - Pm
- Convergence speed

Asexual reproduction optimization

- ARO is individual-based method
- Suppose $X \in R^n$ as an individual i.e. $X = (x_1, x_2, \ldots, x_n)$
- Each x_i is considered as a chromosome with the below structure:

- The length of each chromosome is *l*

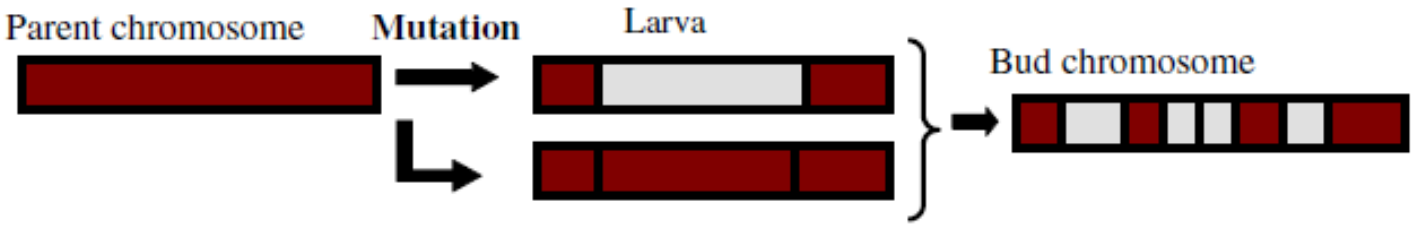
- To start the algorithm, an individual is randomly initiated
- Next, the individual reproduces an offspring labeled **bud** by a particular operator
- The parent and its offspring compete to survive according to a fitness function.
  - If the bud wins the competition, its parent will be discarded. Therefore, the bud is replaced with its parent and it becomes the new parent.
  - If the parent is better, then, the bud will be thrown away.
- The algorithm repeats steps until the stopping criteria are satisfied.

Pseudo code of ARO.

```
Begin
    t = 1;
    P = Initialize (L,U); % Parent Initialization between lower and upper bound
    Fitness_P = fit(P); % Fitness of P is calculated
    While stopping conditions are not met % Stopping Criteria
        Bud(t) = Reproduce(P); % P reproduces a Bud
        Fitness_Bud(t) = fit(Bud(t)); % Fitness of Bud(t) is calculated
        If Fitness_Bud(t) is better than Fitness_P
          P = Bud(t); % Bud(t) is replaced with P
        Else
          clear Bud(t); % Bud(t) is discarded
        end
        t = t + 1;
    End
end
```

- It is obvious that the choice of an appropriate reproduction operator is very crucial
- While ARO only applies one operator, most evolutionary algorithms use the number of operators to explore the search space and to exploit available information according to the traditional control theory.
- In order to reproduce, a substring which has gbits g ~ Uniform[1,L] in each chromosome is randomly chosen.
- Afterward bits of the substring mutate such that in any selected gene, 1 is replaced by 0 and vise versa.

- This substring named larva is a mutated form of its parent



- in the earlier version of ARO, merging of the genes were done with the probability of 0.5
- But in the new versions, to control the exploration, this probability if computed as:

$$p = \frac{1}{1 + Ln(g)}.$$

- It is obvious that when g increases, p decreases and vice versa.

ARO strength and weakness points

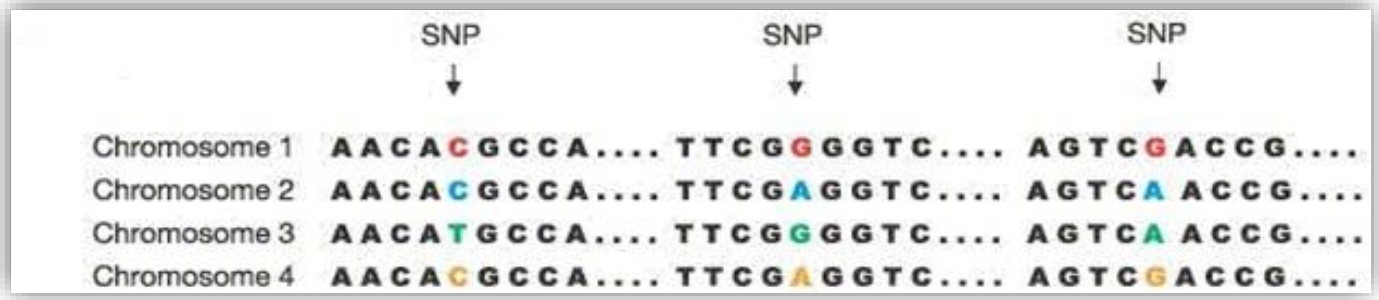
ARO is an individual-based algorithm; hence despite population-based algorithms taking a lot of energy (i.e. time) to evolve, ARO consumes a little energy resulting a remarkable fast convergence time. This property of ARO make it very appropriate for real time applications.

ARO does not require any parameter tuning.

Furthermore, any selection mechanism is not necessary in ARO.

HapMap Project

- Detecting specific DNA sequence variants that determine complex traits
- The Project is a collaboration among scientists in Japan, the U.K., Canada, China, Nigeria, and the U.S.
- The Project officially started with a meeting on October 27-29, 2002



Importance

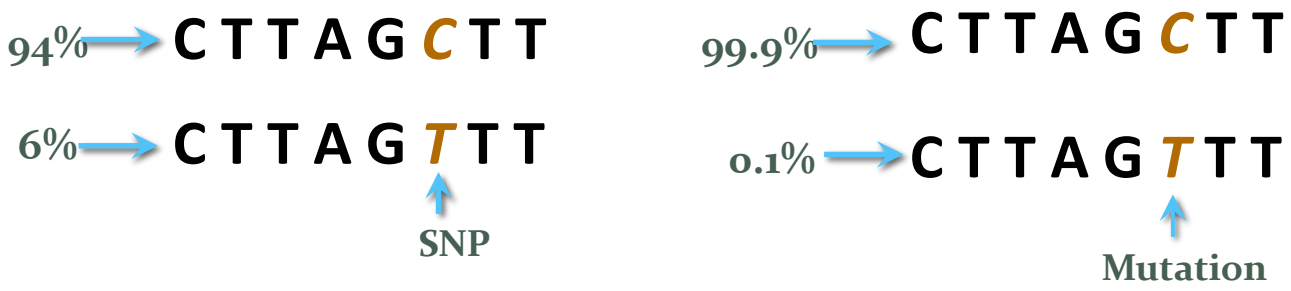
The genetic variations in DNA sequences (e.g., insertions, deletions, and mutations) have a major impact on genetic diseases and phenotypic differences.

All humans share 99% the same DNA sequence!

Single Nucleotide Polymorphism

A **Single Nucleotide Polymorphisms (SNP),** is a genetic variation when a single nucleotide (i.e., A, T, C, or G) is altered and kept through heredity.

- **SNP: Single DNA base variation found >1%**
- **Mutation: Single DNA base variation found <1%**



Allele

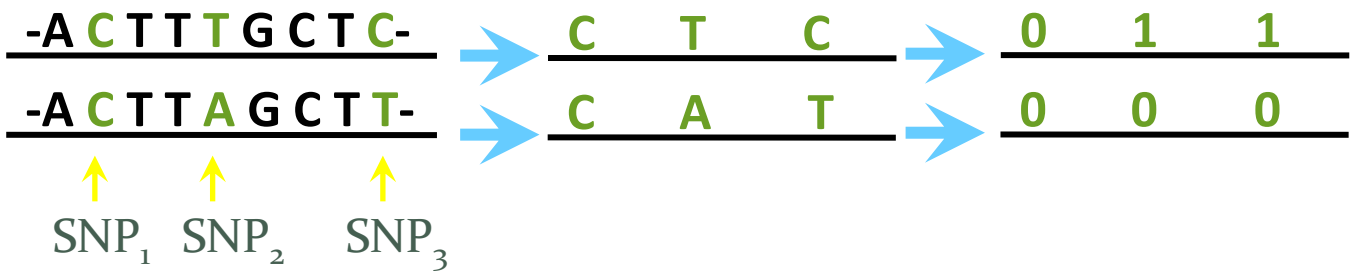
- Each of two or more alternative forms of a base that arise by mutation and are found at the same place on a chromosome
- The nucleotide on a SNP locus is called:
  - a major allele (if allele frequency > 50%), or
  - a minor allele (if allele frequency < 50%).

94% → **A C T T A G C T T** ← T: Major allele
6% → **A C T T A G C T C** ← C: Minor allele

Haplotypes

A **haplotype** is a set of linked SNPs on the same chromosome.

- **A haplotype can be simply considered as a binary string since each SNP is binary.**

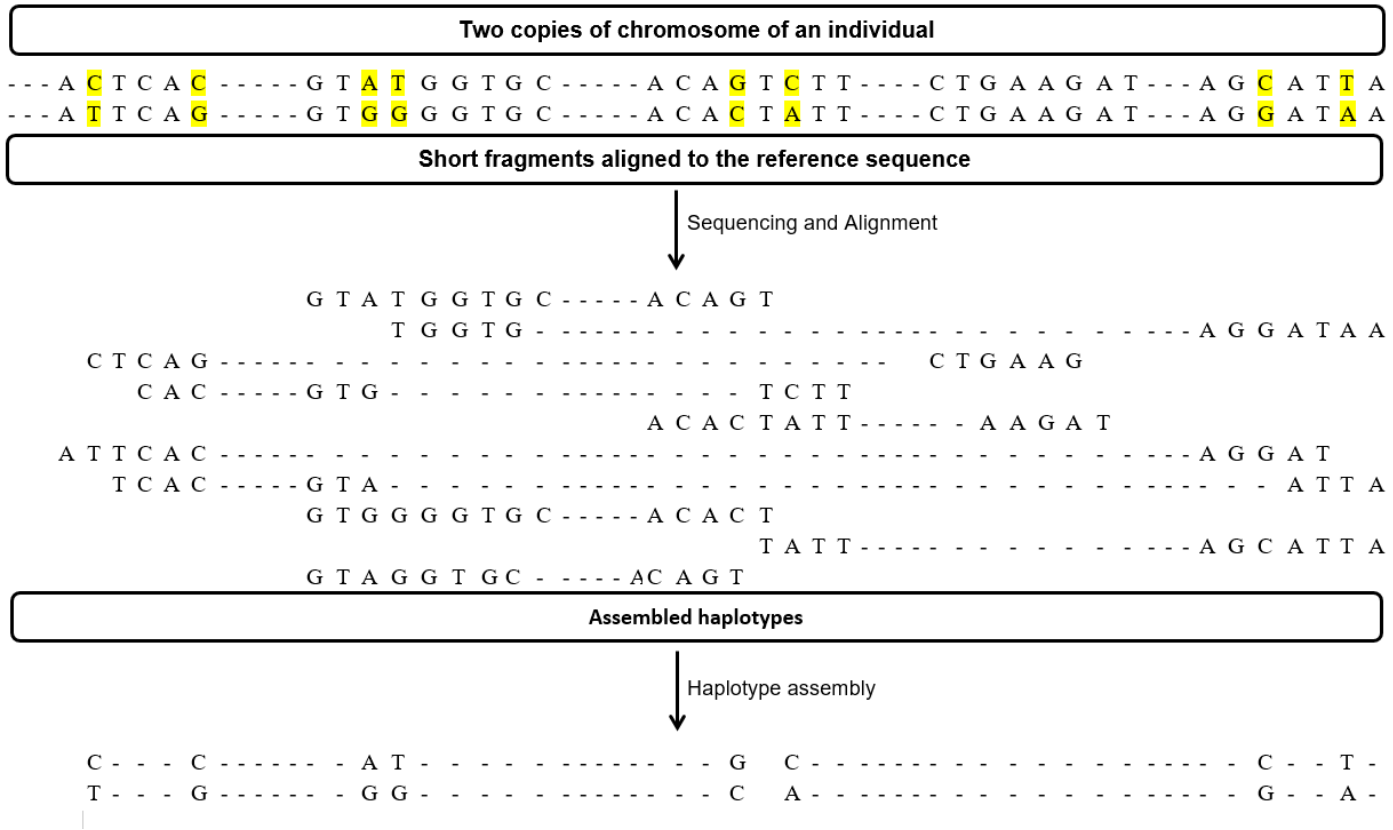


Haplotype Reconstruction

- **Experimental methods**
  - Expensive
  - Time-consuming
  - Low throughput

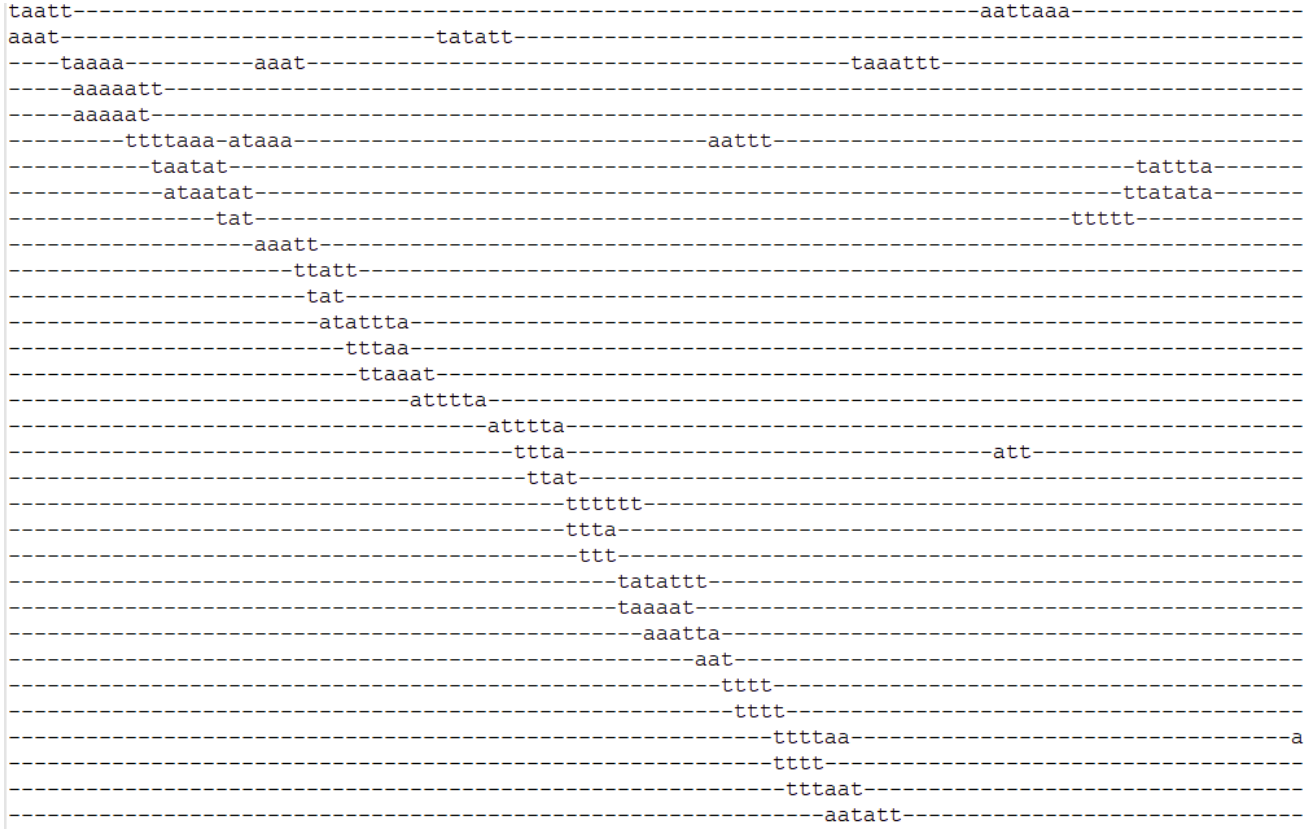
Haplotype assembly

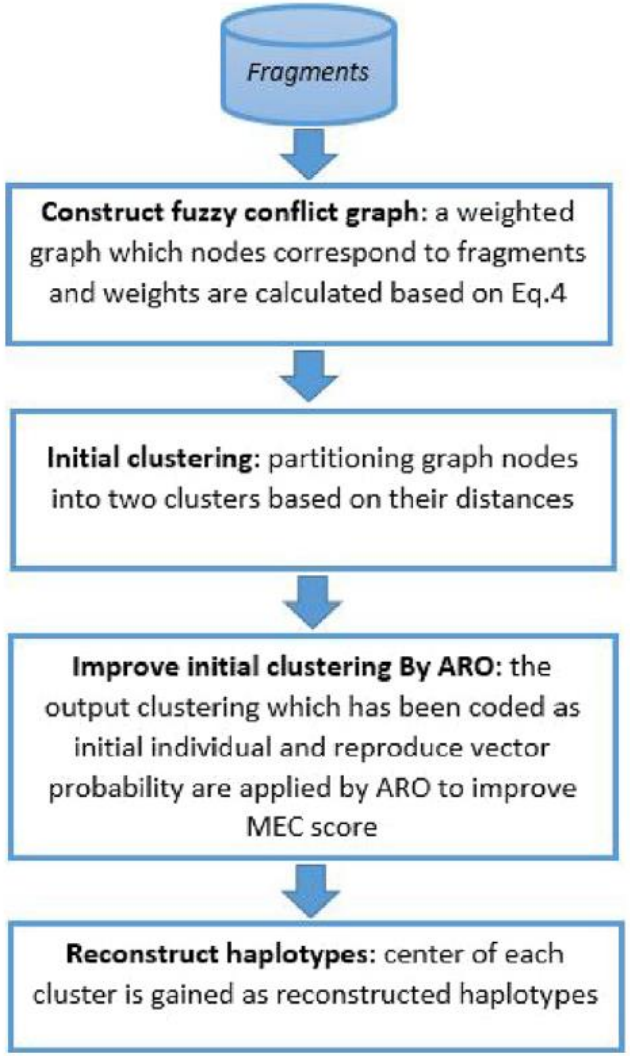
- **Problem statement**

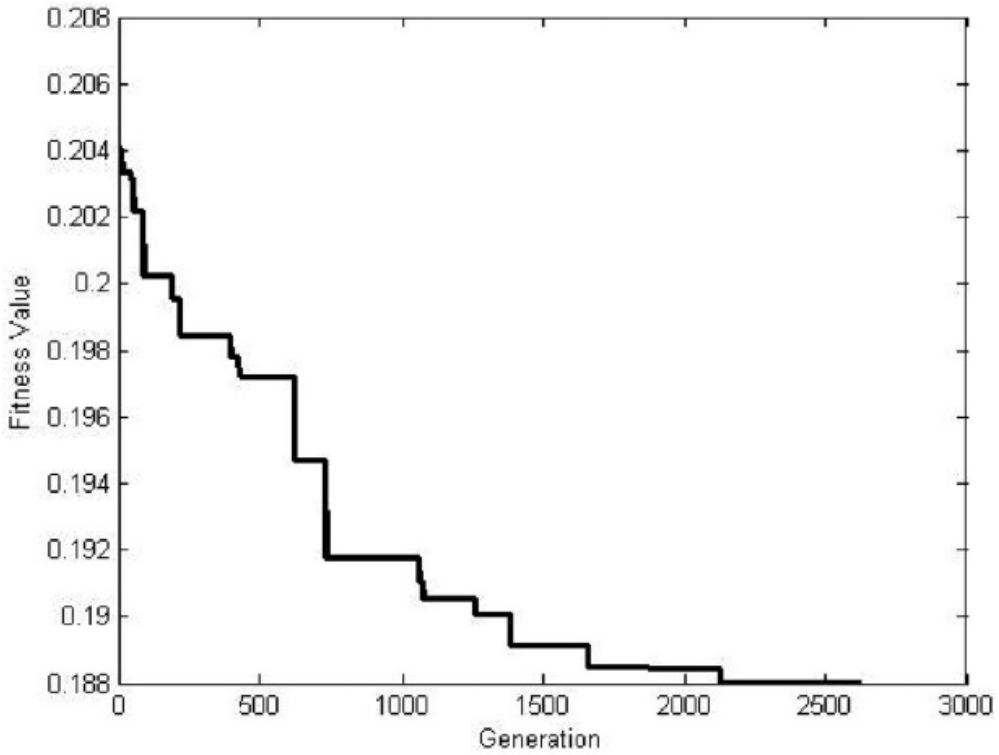


Haplotype assembly

- Data and type of errors
  - Missing information
  - Sequencing errors




Fragments
Construct fuzzy conflict graph: a weighted graph which nodes correspond to fragments and weights are calculated based on Eq.4
Initial clustering: partitioning graph nodes into two clusters based on their distances
Improve initial clustering By ARO: the output clustering which has been coded as initial individual and reproduce vector probability are applied by ARO to improve MEC score
Reconstruct haplotypes: center of each cluster is gained as reconstructed haplotypes


0.208
0.206
0.204
0.202
0.2
0.198
0.196
0.194
0.192
0.19
0.188
Fitness Value
0
500
1000
1500
2000
2500
3000
Generation

e	C	Baseline	SPH	Fast	2d	Cut	MLF	SHR	DGS	GAHap	Fasthap	FCMHap	HGHap	AROHap
0%	3	1.000	0.999	0.999	0.990	1.000	0.973	0.816	1.000	0.996	0.916	1.000	0.999	**1.000**
	5	1.000	1.000	0.999	0.997	1.000	0.992	0.861	1.000	1.000	0.953	1.000	1.000	**1.000**
	8	1.000	1.000	1.000	1.000	1.000	0.997	0.912	1.000	1.000	0.956	1.000	1.000	**1.000**
	10	1.000	1.000	1.000	1.000	1.000	0.998	0.944	1.000	1.000	1.000	1.000	1.000	**1.000**
10%	3	0.971	0.895	0.913	0.911	0.928	0.889	0.696	0.930	0.922	0.823	0.882	0.941	**0.957**
	5	0.992	0.967	0.964	0.951	0.920	0.969	0.738	0.985	0.983	0.917	0.948	0.989	0.972
	8	0.997	0.989	0.993	0.983	0.901	0.985	0.758	0.989	0.989	0.955	0.971	0.994	0.989
	10	0.999	0.990	0.998	0.988	0.892	0.995	0.762	0.997	0.993	0.926	0.972	0.997	0.977
20%	3	0.898	0.623	0.715	0.738	0.782	0.725	0.615	0.725	0.824	0.806	0.739	0.752	**0.858**
	5	0.944	0.799	0.797	0.793	0.838	0.836	0.655	0.813	0.888	0.834	0.772	0.899	**0.919**
	8	0.967	0.852	0.881	0.873	0.864	0.918	0.681	0.878	0.937	0.849	0.793	0.966	0.924
	10	0.980	0.865	0.915	0.894	0.871	0.938	0.699	0.917	0.954	0.899	0.835	0.981	0.956
30%	3	0.788	0.480	0.617	0.623	0.602	0.618	0.557	0.611	0.869	0.578	0.629	0.621	0.694
	5	0.840	0.637	0.639	0.640	0.629	0.653	0.599	0.647	0.791	0.711	0.648	0.698	0.773
	8	0.878	0.667	0.661	0.675	0.673	0.697	0.632	0.663	0.859	0.700	0.664	0.790	0.783
	10	0.903	0.676	0.675	0.678	0.709	0.715	0.632	0.688	0.875	0.732	0.675	0.856	0.823

Computational Biology and Chemistry 72 (2018) 1–10

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

AROHap: An effective algorithm for single individual haplotype reconstruction based on asexual reproduction optimization

Mohammad-H Olyaee, Alireza Khanteymoori*

Department of Computer Engineering, University of Zanjan, Zanjan, Iran

ARTICLE INFO

Article history:
Received 15 October 2016
Received in revised form 22 November 2017
Accepted 10 December 2017
Available online 14 December 2017

Keywords:
Bioinformatics
Haplotype reconstruction
Minimum error correction
Asexual reproduction optimization

ABSTRACT

In this paper, a method for single individual haplotype (SIH) reconstruction using Asexual reproduction optimization (ARO) is proposed. Haplotypes, as a set of genetic variations in each chromosome, contain vital information such as the relationship between human genome and diseases. Finding haplotypes in diploid organisms is a challenging task. Experimental methods are expensive and require special equipment. In SIH problem, we encounter with several fragments and each fragment covers some parts of desired haplotype. The main goal is bi-partitioning of the fragments with minimum error correction (MEC). This problem is addressed as NP-hard and several attempts have been made in order to solve it using heuristic methods. The current method, AROHap, has two main phases. In the first phase, most of the fragments are clustered based on a practical metric distance. In the second phase, ARO algorithm as a fast convergence bio-inspired method is used to improve the initial bi-partitioning of the fragments in the previous step. AROHap is implemented with several benchmark datasets. The experimental results demonstrate that satisfactory results were obtained, proving that AROHap can be used for SIH reconstruction problem.