Estimation of Super Pairwise Alignment (SPA) parameters on zika virus mutation using Artificial Bee Colony

by Mohammad Syaiful Pradana

Submission date: 29-Mar-2023 09:27PM (UTC-0700)
Submission ID: 2050676141
File name: MISEIC_-_Estimation_of_Super_Pairwise.pdf (1M)
Word count: 2565
Character count: 13395

Estimation of Super Pairwise Alignment (SPA) parameters on zika virus mutation using Artificial Bee Colony

Estimation of Super Pairwise Alignment (SPA) parameters on zika virus mutation using Artificial Bee Colony

D Rahmalia[1], A Rohmatullah[2], M S Pradana[3]

[1,2,3]Department of Mathematics, Universitas Islam Darul Ulum Lamongan

Email : [1]dinitarahmalia@gmail.com

Abstract. There are many diseases caused by viruses or bacteria. The structure of virus and bacteria always mutate to create new structures. Sequence alignment is important so that it can be used to research genetic diseases and epidemics. In this reseach, we take case study of zika virus. To see the similarity between original virus and the mutation virus, it is required the alignment process of two virus sequences. The method used for aligning two virus sequences is Super Pairwise Alignment (SPA). Due to the value of objective function depends on SPA parameters, in this research we will apply heuristic method, such as Artificial Bee Colony (ABC) algorithm to optimize SPA parameters minimizing penalty value or maximizing similarity value as objective function. ABC is the optimization method which is inspired by the behaviour of bee colony in which the advantages are there are three types of bees that will update optimal solution in approaching. From the ABC simulations, we can obtain optimal SPA parameters resulting minimum penalty value or maximum similarity value between two aligned zika virus protein sequences in approaching.

1. Introduction

In this time, there are many diseases caused by viruses or bacteria. Based on many researches, the existing of virus and bacteria are from virus and bacteria in the long time ago. They change their structures through mutation process. Mutations can lead to the growth and death of cells, and may lead to disease. Sequence alignment is important so that it can be used to research genetic diseases and epidemics. For example, it is possible to determine the origin, variance, and development of epidemics, and then finding the viruses and bacteria.

In this research, we take case study of zika virus. Zika virus can cause zika disease through *Aedes* mosquito bites with the symptons like dengue fever [1]. Based on previous research, the origin of zika virus is from South Africa and the virus mutate (change genetic sequence) and spread to other countries including Indonesia [2]. To see the similarity between original virus and the mutation virus, it is required the alignment process of two virus sequences. The method used for aligning two virus sequences is Super Pairwise Alignment (SPA) [3].

In sequence alignment using SPA, SPA parameters selection is applied by trial and error [4]. Due to the value of objective function depends on SPA parameters, in this research we will apply heuristic method, such as Artificial Bee Colony (ABC) algorithm to estimate and optimize SPA parameters minimizing penalty value or maximizing similarity value as objective function. In previous research, parameter estimations using heuristic method have been applied in optimization problem [4] and control problem [5],[6].

The Artificial Bee Colony (ABC) was introduced by Karaboga and Akay in 2009. ABC is the optimization method which is inspired by the behavior of bee colony. In the ABC algorithm, there are three types of bees, namely employed bees, onlooker bees, and scout bees in which the advantages are there are three types of bees that will update optimal solution in approaching [7],[8].

Simulations have been applied and we can obtain optimal SPA parameters resulting minimum penalty value or maximum similarity value between two aligned zika virus protein sequences in different countries in approaching.

2. Super Pairwise Algorithm (SPA)

Super Pairwise Alignment (SPA) is the method used for aligning two sequences and computing their penalty or similarity as objective function affected by parameters of SPA (n,θ,ϕ)

Let (A,B) are two sequences with sequence B is the mutation from sequence A. The parameters of SPA are (n,θ,ϕ) with $n>0$, $0.3\le\theta\le 0.4$, and $\theta<\phi\le 1$ [9].

1. Estimate the first mutation position $\hat{i}_1$ in T

   Initialize $i=j=0$ and calculate penalty value $w(A,B,i,j,n)$.

   a. If $(w(A,B,i,j,n)\ge\phi)$ then $\hat{i}_1=0$

   b. Otherwise, $i=j=k_1(n-\tau)$ with k_1 is the integer. Calculate $w(A,B,i,j,n)$. If $(w(A,B,i,j,n)<\phi)$, repeat the step 1(b) by updating $k\leftarrow k+1$ until $(w(A,B,i,j,n)\ge\phi)$, and keep the value of i and j as $\hat{i}_1$

2. Estimate $\hat{l}_1$ based on the estimation $\hat{i}_1$ of the first mutation position in T .

   $w(A,B,\hat{i}_1+l,\hat{i}_1,n)$, $w(A,B,\hat{i}_1,\hat{i}_1+l,n)$, $l=1,2,3,...,l_{\max}$

   a. If $(w(A,B,\hat{i}_1+l,\hat{i}_1,n)\le\theta)$, then $\hat{l}_1=-l$ and insert l virtual symbols (gaps) into sequence B following the position $\hat{i}_1$

   b. If $(w(A,B,\hat{i}_1,\hat{i}_1+l,n)\le\theta)$, then $\hat{l}_1=l$ and insert l virtual symbols (gaps) into sequence A following the position $\hat{i}_1$

   Based on these two steps, we can estimate the local mutation mode $T_1=\{(i_1,l_1)\}$, and its corresponding locally uniform alignment (C_1,D_1) . It is decomposed as follows:

   $C_1=(C_{1,1},A_{2,1})$ $D_1=(D_{1,1},B_{2,1})$

3. Repeat the above process. The process will terminate at some k_0 such that $C_{k_0}=(C_{1,k_0},A_{2,k_0})$ and $D_{k_0}=(D_{1,k_0},B_{2,k_0})$ have shifting mutations occurring in (A_{2,k_0},B_{2,k_0}) .

4. Calculate objective function :

   Penalty value between aligned sequence C_{k_0} and sequence D_{k_0} , $w(C_{k_0},D_{k_0})=\frac{1}{n}\sum_{j=1}^{n}w(c_j,d_j)$

$$w(c_j,d_j)=\begin{cases}0, & \text{if } c_j=d_j\\ 1, & \text{otherwise}\end{cases} \quad (1)$$

   Similarity value between aligned sequence C_{k_0} and sequence D_{k_0} , $s(C_{k_0},D_{k_0})=\frac{1}{n}\sum_{j=1}^{n}s(c_j,d_j)$

$$s(c_j,d_j)=\begin{cases}1, & \text{if } c_j=d_j\\ 0, & \text{otherwise}\end{cases} \quad (2)$$

   In this research, penalty value or similarity value optimization simulations are applied separately.

3. Artificial Bee Colony (ABC)

The Artificial Bee Colony (ABC) was introduced by Karaboga and Akay in 2009. ABC is the optimization method which is inspired by the behavior of bee colony. There are three groups in the ABC: employed bees, onlooker bees, and scouts. For every food source, there is only an employed bee. The position of food source represents possible solution of the optimization problem and the amount of nectar represents the fitness of solution [7],[8].

Given objective function $f : X \subseteq R^D \rightarrow R$ where D is the dimension of the search space.

In the initialization step, generate initial solutions $x_{ij}, i = 1..SN, j = 1..D$ randomly and evaluate the fitness $f(x_{ij}), i = 1..SN, j = 1..D$ [7]. In ABC applied to SPA optimization model, suppose $X = (n, \theta, \phi)$ as decision variable solution with objective function in equation (1) for minimizing penalty value or in equation (2) for maximizing similarity value.

Repeat

1. Employed Bees Step

   In the employed bees step, each employed bee generates solution as follows:

$$v_{ij} = \begin{cases} x_{ij} + \varphi_{ij}(x_{ij} - x_{kj}), i \neq k & \text{if } r_{ij} < MR \\ x_{ij} & \text{otherwise} \end{cases} \quad (3)$$

   with φ_{ij} is random number between $(-1,1)$, r_{ij} is random number between $(0,1)$, and MR is the modification rate.

   Evaluate $f(v_{ij}), i = 1..SN, j = 1..D$. If new solution v_{ij} is better than x_{ij}, replace x_{ij} with v_{ij}.

2. Onlooker Bees Step

   Onlooker bees select one of solutions based on probability $p_i = \frac{fitness_i}{\sum_{n=1}^{SN} fitness_n}$.

3. Scout Bees Step

   Determine the abandoned solution for the scout and replace with:

$$x_{ij} = x_j^{\min} + \tau_{ij}(x_j^{\max} - x_j^{\min}) \quad (4)$$

   with $\tau_{ij} \sim U(0,1)$ then eveluate $f(x_{ij}), i = 1..SN, j = 1..D$

4. Keep the best solution

4. Simulation Results

Based on previous research [2], the origin of zika virus is from South Africa and the virus mutate (change genetic sequence) and spread to other countries including Indonesia. Indonesia also derives zika virus from Cambodia. Data used in this research are zika virus protein data in Indonesia, Cambodia as its neighbour country, and South Africa as origin of zika virus. Data are obtained from National Center for Biotechnology Information (NCBI) accessed on May 2, 2019. The characteristics of zika virus protein data can be seen in Table 1.

Table 1. Zika virus protein data.

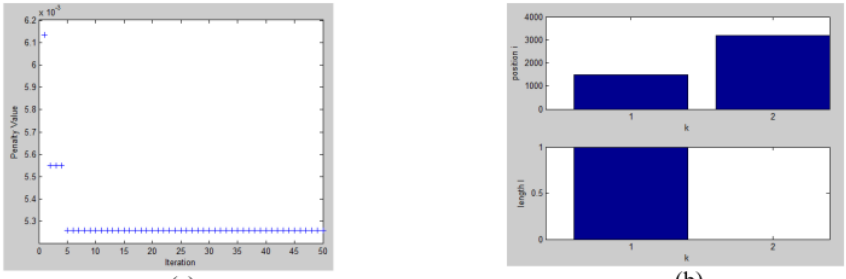
Access Code	Country	Length (bp)
ABI54480	South Africa	3429
AFD30972	Cambodia	3423
AMK49492	Indonesia	3423

There are two types of ABC simulations used : ABC simulation of penalty value minimization and ABC simulation of similarity value maximization. ABC simulation of penalty value minimization is to find SPA parameters (n,θ,ϕ) minimizing penalty value (including gaps) between two aligned zika virus protein sequences as objective function while ABC simulation of similarity value maximization is to find SPA parameters (n,θ,ϕ) maximizing similarity value (excluding gaps) between two aligned zika virus protein sequences as objective function. ABC parameters used in the simulations are : the number of bees : 10, modification rate : 0.8, maximum iteration : 50.

4.1. ABC Simulations of Penalty Value Minimization

Figure 1(a) shows ABC optimization process of penalty value minimization of aligned zika virus protein sequence in Cambodia and Indonesia. At the early iteration, the bees choose the position of food source as SPA parameters randomly. At the optimization process, we update nectar information as fitness function so that bees choose the SPA parameters resulting minimum objective function. Based on simulation, the optimal SPA parameters are $(n,\theta,\phi)=(246;0.3135;0.5395)$ with minimium penalty value is 0.005259 or 0.5259% and computation time is 36.57 seconds.

Figure 1(b) shows the positions i_k and lengths l_k of optimal SPA parameters. Optimal SPA parameters result iteration $k=2$. From the graphs, positions i_k and lengths l_k can be seen. If $l_k>0$, it occurs insertion, if $l_k<0$, it occurs deletion, and if $l_k=0$, there is no mutation shifting.

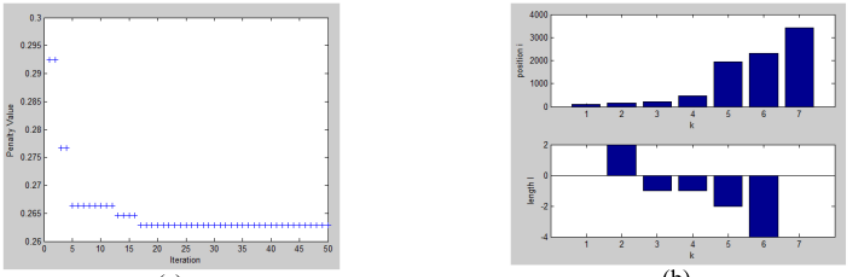


(a) (b)

Figure 1. Penalty value minimization of aligned zika virus protein sequences (Cambodia and Indonesia) (a) ABC optimization process (b) Positions and lengths of optimal SPA parameters

The similar procedure is applied to aligned zika virus protein sequence in South Africa and Indonesia. Based on figure 2(a), the optimal SPA parameters are $(n,\theta,\phi)=(51;0.3719;0.6819)$ with minimium penalty value is 0.262897 or 26.2897% and computation time is 45.02 seconds. Figure 2(b) shows the positions i_k and lengths l_k of optimal SPA parameters.

From the comparison, penalty value of aligned zika virus protein sequences in Cambodia is smaller than penalty value of aligned zika virus protein sequences in South Africa.



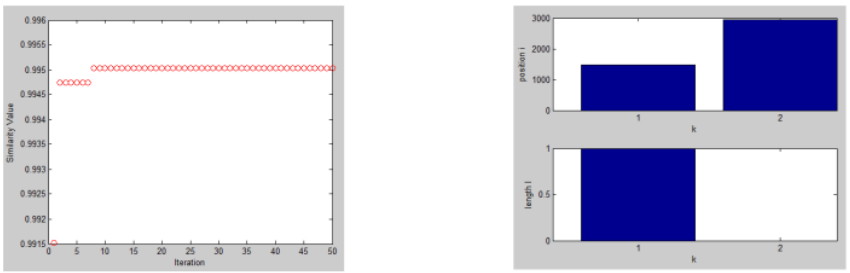
(a) (b)

Figure 2. Penalty value minimization of aligned zika virus protein sequences (South Africa and Indonesia) (a) ABC optimization process (b) Positions and lengths of optimal SPA parameters

4.2. ABC Simulations of Similarity Value Maximization

Figure 3(a) shows ABC optimization process of similarity value maximization of aligned zika virus protein sequence in Cambodia and Indonesia. At the early iteration, the bees choose the position of food source as SPA parameters randomly. At the optimization process, we update nectar information as fitness function so that bees choose the SPA parameters resulting maximum objective function. Based on simulation, the optimal SPA parameters are $(n,\theta,\phi)=(492;0.3682;0.6453)$ with maximum similarity value is 0.995032 or 99.5032% and computation time is 35.53 seconds.

Figure 3(b) shows the positions i_k and lengths l_k of optimal SPA parameters. Optimal SPA parameters result iteration $k=2$. From the graphs, positions i_k and lengths l_k can be seen. If $l_k>0$, it occurs insertion, if $l_k<0$, it occurs deletion, and if $l_k=0$, there is no mutation shifting.

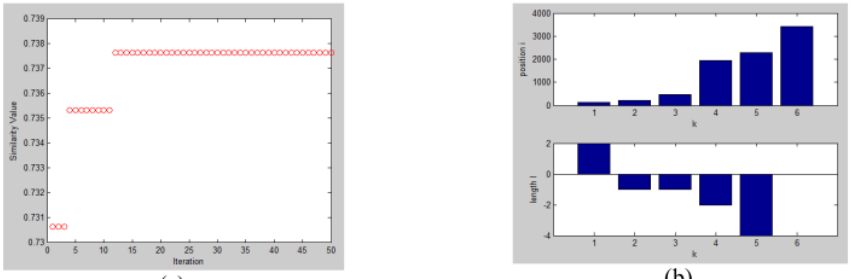


(a) (b)

Figure 3. Similarity value maximization of aligned zika virus protein sequences (Cambodia and Indonesia) (a) ABC optimization process (b) Positions and lengths of optimal SPA parameters

The similar procedure is applied to aligned zika virus protein sequence in South Africa and Indonesia. Based on figure 4(a), the optimal SPA parameters are $(n,\theta,\phi)=(67;0.3959;0.5850)$ with maximum similarity value is 0.737657 or 73.7657% and computation time is 43.90 seconds. Figure 4(b) shows the positions i_k and lengths l_k of optimal SPA parameters.

From the comparison, similarity value of aligned zika virus protein sequences in Cambodia is larger than similarity value of aligned zika virus protein sequences in South Africa.



(a) (b)

Figure 4. Similarity value maximization of aligned zika virus protein sequences (South Africa and Indonesia) (a) ABC optimization process (b) Positions and lengths of optimal SPA parameters

5. Conclusion

To see if the similarity between two sequences, it is required the alignment process of two sequences. The method used to align them is Super Pairwise Alignment (SPA). In SPA, the penalty value and the similarity value depend on SPA parameters. The SPA parameters can be optimized by heuristic method, such as Artificial Bee Colony (ABC) inspired from the behavior of bees. From the simulation with case study of zika virus protein, ABC can find optimal SPA parameters resulting minimum penalty value or maximum similarity value in approaching. The development of this research is for optimizing the similarity of multiple sequences.

Acknowlogdement

High appreciation to the Ministry of Research and High Education (Kemenristekdikti), Republic of Indonesia for the Penelitian Dosen Pemula fund for the research conducted in the year of 2019.

References

[1] Pradana M S and Amiroch S 2019 Protein Sequence Analysis of the Zika Virus and the Dengue Virus Using Smith Waterman Algorithm *AIP Conference Proceeding* **2084**
[2] Pradana M S and Amiroch S 2018 Zika Virus Mutation and the Spreading to Indonesia *International Journal of Computing Science and Applied Mathematics* **4** pp 15-18
[3] Isaev A 2004 *Introduction to Mathematical Methods in Bioinformatics* (Berlin : Springer)
[4] Shahab M L and Irawan M I 2017 Sequence Alignment Using Nature-Inspired Metaheuristic Algorithm *International Journal of Computing Science and Applied Mathematics* **3**
[5] Rahmalia D and Herlambang T 2018 Weight Optimization of Optimal Control Influenza Model Using Artificial Bee Colony *International Journal of Computing Science and Applied Mathematics* **4** pp 27-31
[6] Herlambang T, Rahmalia D, Yulianto T 2019 Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) for Optimizing PID Parameters on Autonomous Underwater Vehicle (AUV) Control System *Journal of Physics : Conference Series* **1211**
[7] Karaboga D and Basburk B A 2007 Powerful and Efficient Algorithm for Numerical Function Optimization : Artificial Bee Colony (ABC) Algorithm *Journal of Global Optimization* **39** pp 459-471
[8] Liao T, Aydin D, Stutzle T *Artificial Bee Colony for Continuous Optimization : Experimental Analysis and Improvements* (New York : Springer Science+Business Media)
[9] Shen S N and Tuszynski J A 2008 *Theory and Mathematical Model for Bioinformatics* (Berlin : Springer)

Estimation of Super Pairwise Alignment (SPA) parameters on zika virus mutation using Artificial Bee Colony

ORIGINALITY REPORT

15%	12%	9%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

2%

★ www.eri.u-tokyo.ac.jp

Internet Source

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Estimation of Super Pairwise Alignment (SPA) parameters on zika virus mutation using Artificial Bee Colony

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7