Journal of Engineering and Applied Sciences 8 (2): 49-57, 2013
ISSN: 1816-949X
© Medwell Journals, 2013

DNA Sequence Design Using Artificial Immune Systems

Mohd Zakree Ahmad Nazri, M. Daman Huri, Azuraliza Abu Bakar,
Salwani Abdullah, Masri Ayob Dan and Tri Basuki Kurniawan
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
43600 Bangi, Selangor, Malaysia

Abstract: The fundamental principle in the fields of DNA computing and DNA nanotechnology is based on the complementary pairing of the Deoxyribonucleic Acid (DNA). In this field of research, it is essential to obtain good DNA sequences in order to obtain accurate DNA-based computational information. In this process, however there are four constraints involved, namely $H_{measure}$, similarity, continuity and hairpin. In addition, two other constraints also play a role to maintain the uniformity in the sequence of the GC-content and the melting temperature (T_m) that would arise. Therefore, a DNA sequence design tool is needed to facilitate the design process with the ability to monitor and completely satisfy the specified constraints. In this study, a biologically-inspired DNA sequence design algorithm is presented and it allows generated sets of DNA that satisfy the several thumb rules in the DNA sequence design. The algorithm is based on the Negative Selection Algorithm (NSA). NSA is a common technique inspired by the negative selection process that occurs during the maturation of the T cells in the thymus. The proposed algorithm is able to prevent risks of fraying strands of the DNA and to limit cross hybridizations. In addition, it is able to design unique sequences. Furthermore, the NSA based algorithm can prevent the formation of self-complimentary and hairpin structures of certain lengths and only allows minimum interaction with neighbouring sequences. In this study, the results are compared to an Ant Colony Optimization (ACO) based on the DNA sequence design tool. The analysis shows that the NSA based algorithm performs better than ACO in generating the DNA sequences that satisfy the given constraints.

Key words: Component, artificial immune system, clonalg, DNA sequence design, optimization, Malaysia

INTRODUCTION

The inception of the Human Genome Project in 1989 encompasses the goal of sequencing and identifying all 3 billion chemical units in the human genetic instruction set. From there on researchers believe that the significant scientific and technological breakthrough in the 21st century would be related to the processing and interpretation of the vast information that was currently being revealed from sequencing the genomes. Protein and Deoxyribonucleic Acid (DNA) are examples of the genomic sequences and self-assembly and self-complimentary are the two known unique properties of the DNA. These features enable the DNA to save an enormous amount of data and perform massive parallel reactions that have been exploited by scientists to open new avenues for further advancement in many fields like biotechnology, nanotechnology and even computer science, especially in the field of DNA computing.

Instead of using the traditional silicon-based computer technologies, DNA computing or bio-molecular computing uses the DNA itself for computing. The computation uses specific biochemical reactions between the different DNA strands as found by the Watson-Crick complementary based pairing which allows DNA computing to have an advantageous property such as having a vast memory capacity with massive parallelism. However, a successful computation depends on the quality of the DNA sequences used and a good DNA sequence design is therefore, the essence of achieving high computation accuracy. But however, a DNA sequence design is not a trivial task. Adleman (1994) who demonstrated the possibility of using DNA to compute and solve complex problems expresses his doubts on an all-purpose library of sequences that can effectively cater for the requirements of all laboratory experiments. His concern is due to the differences in the experimental requirements. Kashiwamura *et al.* (2003) stressed the need of a systematic method for designing the DNA sequences because a design of the DNA sequence is only reliant on the protocol of biological experiments. Furthermore, Khalid states that the *in vitro* reactions may lead to

Corresponding Author: Mohd Zakree Ahmad Nazri, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia incorrect computing solutions because of the biochemical complexity of the experiment which fails to repeat identical results for the same problem even when using the same algorithm. Thus, many researchers focus on improving the reliability and efficiency of the DNA computing (Kobayashi and Kondo, 2002). The necessity in solving these challenges is to design a good hybridization between a sequence and its base pairing complement in order to retrieve the information stored in the sequences and to operate the computation processes. The aim is to have a stable duplex where the complement and the two sequences do not complement one another but are two important design requirements. Thus, various heuristics-based approaches have been developed to meet these requirements in order to produce good DNA sequences.

For the last 2 decades, computer scientists turn to nature and biology in their quest for effective and efficient solutions for solving complex problems. Natural biological systems, such as the genetic algorithm has intrinsically posed great features and delivered many great concepts. Adaptability and robustness of biological systems are among the leading attributes that motivate computer scientists to use them as one of the computing algorithms. The borrowing of nature or biological principles and processes, such as the human immune system has been applied in many domains. The first known heuristic approach in designing the DNA sequence was proposed by Parsons and Johnson (1995) and followed by Deaton *et al.* (1996a). Both study research on the genetic algorithm to design DNA sequences. Deaton *et al.* (1996a, b) uses the Hamming distance for measuring the similarity between the DNA strands in order to generate a unique DNA strand. As Fang *et al.* (2005) used genetic algorithm to solve the DNA fragment assembly problem. Genetic algorithm is applied again in designing DNA sequences but has been enhanced with a tuning function by using a different heuristic algorithm as proposed by Kikuchi and Chakraborty (2006). Alba and Luque (2008) hybridized the genetic algorithm with PALS (Parallel Adaptive Learning Search) while in the same year Nebro *et al.* (2008) developed a grid-based genetic algorithm for a DNA fragment assembly problem. Other than the genetic algorithm, researchers have applied other meta-heuristic algorithms, such as the Ant Colony system by Meksangsouoy and Chaiyaratna and Kurniawan (2009) proposed the population ant colony optimization and the particle swarm optimization by Ravi and Sanjay (2011).

However based on a survey by Indumathy and Maheswari, the Artificial Immune System (AIS) principles and processes have yet to be applied in designing a DNA sequence and therefore, leave ample opportunities for exploration. Previous research has shown that the Artificial Immune System (AIS) has attributes that would fulfill such task. There are a number of motivating factors to use the immune system as inspiration for designing the DNA sequence which includes recognition, diversity, memory, self-regulation and learning (Dasgupta, 1998). In this study, a new DNA sequence design is proposed and which is inspired by the Negative Selection Algorithm (NSA) and CLONALG (De Castro and von Zuben, 2002).

MATERIALS AND METHODS

DNA sequence design: James Watson and Francis Crick discovered the chemical structure of the DNA in 1953. DNA and Ribonucleic Acid (RNA) are the most common nucleic acids. Nucleic acid is a macromolecule composed of chains of monomeric nucleotide molecules that carry genetic information or form structures within cells. The DNA in particular which can be found in all cells and viruses is a polymer that is strung together by a series of monomers. Monomers construct the building blocks of nucleic acids which are known as nucleotides. Each nucleotide contains a sugar (deoxyribose), a phosphate group and one of the four bases: Adenine (A), Thymine (T), Guanine (G) or Cytosine (C). It is the combination of these four bases that determines the precise function and coding capacity of the DNA. By using the same 4 alphabets, strings of these alphabets have mathematically represented the digital DNA.

A DNA is double stranded and is held together by the hydrogen bonds between the base pairs. This is termed as a duplex or double stranded DNA. Based on the Watson-Crick DNA base pairing, A forms a base pair with T and G forms a base pair with C. A sequence of DNA can be read from 5-end (the ribose end) of one sequence and the 3-end (the phosphate end) of another sequence. This complementary base is critical for various DNA testing techniques and the basic principles of the DNA chemistry. When put together in a unique way, the string of A, C, T or G can serve as a template for messenger Ribonucleic Acid (mRNA) which in turn codes for the proteins. These proteins finally form the structure and function for each and every process inside the cells and inside an organism. As a result, sequences of DNA coding for proteins enables the process of constructing and maintaining the cell which is finally responsible for all genetic processes. However, perfect hybridization between a sequence and its base-pairing complement is important to retrieve the information stored in the sequences and in order to operate the computation processes.

Objectives and constraints: The DNA sequence design is a multi-objective problem. However in this study, the problem has been converted into a single objective problem using the weighted sum method. The same weighted sum method is used by Tribasuki and Khalid to scale down a set of objectives into a single objective by pre-multiplying each objective with a user-supplied weight. However, setting the value of the weights is not easy as it depends on the importance of each objective in the context of the problem and a scaling factor. The multi-objective problem is converted into a single objective problem using Eq. 1 as follows:

$$\min f_{DNA} = \sum_{i} \omega_i f_i \qquad (1)$$

Equation 1 is subjected to T_m and $GC_{content}$ constraints where f_i is the objective function for each i_{ϵ} {$H_{measure}$, similarity, hairpin, continuity} and ω_i is the weight for each f_i. Here, ω typically set by the decision maker such that $\Sigma_{i=1}^{k} \omega_i = 1$ and $\omega > 0$. If all the weight is committed or set to 1 then all objectives are treated equally (Arora, 2004). The basic notations which will be used comprehensively in this study are displayed in Table 1. Table 2 shows the additional notations that are used to formulate Eq. 2-4. Furthermore for a given sequence $x \epsilon \Lambda^*$, the numbers of non-blank nucleotides are defined as Eq. 5 and 6. While a move of sequence x by i bases are formulated as Eq. 7 and 8.

Design criteria: DNA computing would only succeed when the hybridization between a DNA sequence and its base-pairing complement is perfect. This is important so that information stored in the DNA molecules can be retrieved easily and for this reason the design criteria is very important in the DNA design process. In this study, four objective functions namely $H_{measure}$, similarity, hairpin and continuity and two other constraints which are the GC-content and the melting temperature are employed as the design criteria. The selection and the calculations of all the design criteria applied in this research are based on Shin *et al.* (2005). The objective functions and constraints are described as follows:

$H_{measure}$: This function calculates the number of complementary nucleotides in an attempt to prevent the occurrence of a hybrid cross between two sequences whereas the system shifting of one sequence against the other on $H_{measure}$ is basically the minimum of the hamming distance that was introduced by Garzon *et al.* (1997). This method has an advantage and is considered as $H_{measure}$ which is then used to calculate numerous nucleotides that are complementary. This strategy is to avoid cross breeding between two sequences that belong in the shift position. The equation of $H_{measure}$ is shown in Eq. 9. Where Σ_i and Σ_j are anti-parallel to each other. $H_{measure}$ (x-y) is divided into two measurements which are h_{con} and h_{dis}. Each section has a function, the function of h_{con} which calculates the overall complement in Eq. 11 and h_{dis} for the penalty or limits the continuous complement as in Eq. 12.

Table 1: Basic definition

Notation	Description
Λ	{A,C,G,T}
$x, y \epsilon \Lambda$	x, y = {A, C, G, T}
$\|x\|$	Length of x
x_i $(1 \le i \le \|x\|)$	ith nucleotide from 5'end of sequence x
Σ	A set of n sequences with the same length l
Σ_i	ith member of Σ
$\bar{a}$	Complementary base of a
l	Length of sequence
n	No. of sequences

Table 2: Notation

Equations	No.
$bp(a,b) = \begin{cases} 1 & a = \bar{b} \\ 0 & \text{otherwise} \end{cases}$	2
$T(i,j) = \begin{cases} i & i > j \\ 0 & \text{otherwise} \end{cases}$	3
$eq(a,b) = \begin{cases} 1 & a = b \\ 0 & \text{otherwise} \end{cases}$	4
$Length(x) = \sum_{i=1}^{\|x\|} n(x_i)$	5
$n(a) = \begin{cases} 1 & a \in A \\ 0 & \text{otherwise} \end{cases}$	6
$Shift(x,i) = \begin{cases} (-)^i x_1 \ldots x_{1-1} & i \ge 0 \\ x_{i+1} \ldots x_1 (-)^i & i < 0 \end{cases}$	7
$rev(x) = x_1 \ldots x_i \quad \text{for} 1 \le i \le$	8

Similarity: The function of the similarity is to calculate the nucleotides which are similar in the two parallel strands of the DNA where calculations are used to maintain the uniqueness of every strand of DNA that is generated. The formulation of similarity is shown in Table 3, Eq. 14 where Σ_i and Σ_j are parallel to each other. Similarity (x, y) is also divided into a two functional measurements which are S_{dis} and S_{con}. S_{dis} they are measurements for the overall complementary show in Eq. 16 and S_{con} is the penalty measurement for the continuous complementary area, Eq. 17. S_{dis} real value between 0 and 1 and S_{con} is an integer between 1 and l. Both values presented are set by the user.

Table 3: Equation function

Equations	No.	Equations	No.
$f_{Hmeasure}(\Sigma)=\sum_{i=1}^{n}\sum_{j=1}^{n}H_{measure}(\Sigma_i,\Sigma_j)$	9	$S_{con}(x,y)=\sum_{i=1}^{1}T(ceq(x,y,i),S_{con})$	17
$h_{measure}(x,y)=\max_{\|i\|<1-1}\begin{pmatrix}h_{dis}(x,shift(rev(y),i)+\\ h_{con}(x,shift(rev(y),i)\end{pmatrix}$	10	$ceq(x,y,i)=\begin{cases}c & \text{if } \exists c, \text{s.t. } eq(x_i,y_i)=0, eq(x_{i+j},y_{i+j}) \\ & \text{for } 1\le j\le c, eq(x_{i+c+1},y_{i+c+1}) \quad =0 \\ 0 & \text{otherwise}\end{cases}$	18
$h_{dis}(x,y)=T\left(\sum_{i-1}^{1}bp(x_i,y_i),\ h_{dis}\times length(y)\right)$	11	$f_{hairpin}(\Sigma)=\sum_{i=1}^{n}hairpin(\Sigma_i)$	19
$h_{dis}(x,y)=T\left(\sum_{i=1}^{1}bp(x_i,y_i),\ h_{dis}\times length(y)\right)$	12	$hairpin(x)=\sum_{p=P_{min}}^{(1-R_{min})/2}\sum_{r=R_{min}}^{1-2p}\sum_{i=1}^{1-2p-r}T\begin{pmatrix}\sum_{j=1}^{pinlen(p,r,i)} & bp(x_{p+i+j},x_{p+i+r+j}) \\ & pinlen(p,r,i)/2\end{pmatrix}$	20
$cbp(x,y,i)=\begin{cases}c \text{ if } \exists c, \text{s.t. } bp(x_i,y)=0, \\ bp(x_{i+j},y_{i+j})=1 \text{ for } 1\le j\le c, \\ \quad bp(x_{i+c+1},y_{i+c+1})=0 \\ \quad 0 \quad \text{otherwise}\end{cases}$	13	$pinlen(p,r,i)=\min(p+i,1-r-i-p)$	21
$f_{similarity}(\Sigma)=\sum_{i=1}^{n}\sum_{j=1,j\ne i}^{n}similarity\ (\Sigma_i,\Sigma_j)$	14	$f_{continuity}(\Sigma)=\sum_{i=1}^{n}continuity(\Sigma_i)$	22
$similarity(x,y)=\max_{\|i\|<1-1}\begin{pmatrix}S_{dis}(x,shift(y,i)+\\ S_{con}(x,shift(y,i)\end{pmatrix}$	15	$continuity(x)=\sum_{1\le i\le 1}\left(\sum_{a\in A}T(c(a,i),t)^2\right)$	23
$S_{dis}(x,y)=T\left(\sum_{i=1}^{1}eq(x_i,y_i),S_{dis}\times length(y)\right)$	16	$c(a,i)=\begin{cases}n & \text{if } \exists n, \text{s.t.} eq(a_i,a_{i+j})=1 \\ \text{for} & 1\le j<n, eq(a_i,a_{i+n})=0 \\ 0 & \text{otherwise}\end{cases}$	24
$GC_{content}=(yG+zG)/(wA+xT+yG+zG)$	25	$T_m(x)=\frac{\Delta H}{\Delta S+R\ln C_T}+16.6\log(Na^+)$	26

Hairpin: This function is used to calculate the probability of the occurrence of double structures of a DNA strand. For example, a hairpin can be seen in Fig. 1 which shows the probability of the formation of ring structures based on certain values of p (pair) and r (ring) to 20-mer DNA sequence, this formula is based on Kurniawan (2009) as shown in Eq. 19 and equation for the pinlen shown in Eq. 19.

Continuity: To calculate whether any base (A, T, C, G) is located in a continuous sequence in a DNA strand. This happens when the objective is to get the calculation of the amount base which is continuous from the set sequences (Fig. 2). Kurniawan (2009), defines it as shown in Eq. 22 where every i, x represents Σ_i. Other than the four functions of these objectives, there are two constraints where each function of the constraints maintains uniformity for every strand of the DNA sequence.

Gc$_{content}$: The percentage of G and C in a sequence is very important, since it can influence the chemical attributes that are existing in a DNA sequence (Brenneman and Condon, 2002). The formulation of GC$_{content}$ is shown in Eq. 25.

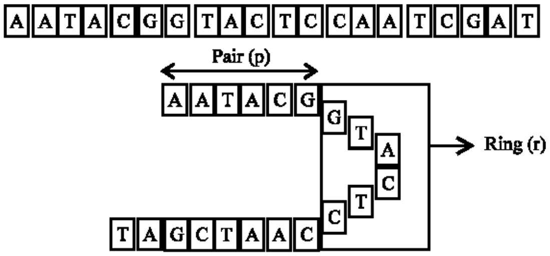

Fig. 1: Formation of sequence DNA being composed into pair (p) and ring (r) = 6

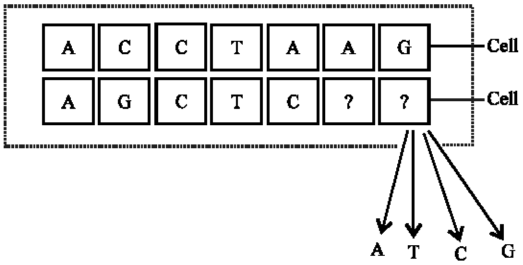

Fig. 2: The A, T, C, G elements of DNA representing cells (antigen)

Melting temperature (T_m): The melting temperature (Tm) is also an important factor in the experimental design of the DNA sequence. This is the temperature at that point where half of the two strands are split and separated to form one strand while the melting temperature exceeds a certain given threshold. In this study, the molar free enthalpy of DNA duplex formation is then calculated. This calculation is based on the nearest-neighbour model Santa Lucia Unified (Lucia, 1998) with parameters taken from the same study. Equation 26 depicts the formulation where ΔH and ΔS are enthalpy and entropy changes of the annealing reaction as shown in Table 4. CT is the total oligonucleotide strand concentration. For non-self-complementary molecules, CT is replaced by CT/4. Na+ is the salt concentration.

Artificial immune systems: The aim of this study is to provide an introduction to Artificial Immune System (AIS) and its relation to the proposed algorithm termed as NS-Cl. AIS can be defined as a computing paradigm inspired from the theory of the immune system with regard to the functions, principles and mechanisms of the immune system (De Castro and Timmis, 2002a). The immune system has a number of basic mechanisms. Each mechanism has different working principles, such as the clonal selection, the negative selection and the danger theory. The proposed method is examined in this study from the results of the tailor made NSA with evolutionary properties, i.e., cloning and mutations. The proposed algorithm in this study is based on the Negative Selection Algorithm (NSA) but the Clonal Optimization method (CLONALG) is employed to optimize the detectors in the NSA.

Negative selection: Forrest developed the NSA which is inspired by the negative selection process. Theoretically, natural immune systems are capable of differentiating any foreign cell (non-self) or molecule from the body's own cell. During the production of T lymphocyte cells (T cells), there are receptors that are made for the censoring process as the T cells have to be matured in the thymus before they are released into the circulatory system. This censoring process which occurs in the thymus is called a negative selection. During the negative selection process, the T cells that react against the self-proteins are destroyed. Only the unrecognized or unique T cells that do not bind to self-proteins are allowed to leave the thymus and such successful T cells then will then enter into the circulatory system. These successful T cells have an interesting feature where they can only be activated by foreign cells.

Table 4: ΔH and ΔS of Santa Lucia Unified

	Unified	
Sequence	ΔH_x	ΔS_x
AA/TT	-7.9	-22.2
AT/AT	-7.2	-20.4
AG/CT	-7.8	-21.0
AC/GT	-8.4	-22.4
TA/TA	-7.2	-21.3
TG/CA	-8.5	-22.7
GA/TC	-8.2	-22.2
GG/CC	-8.0	-19.9
GC/GC	-9.8	-24.4
CG/CG	-10.6	-27.2
G-C base pair on the end	0.1	-2.8
A-T base pair on the end	2.3	4.1
Equilibrium rectification	0.0	-1.4

NSA has an important component of memory that contains three sets of memory which are P, M and C. Set P is the set of self-contained and the goal of the NSA is to provide a set of patterns where P should be protected. Set M or Detector Set (M) is responsible for identifying all the elements that produce a set containing non-self-elements. NSA begins by generating candidate element (C) randomly. Each element in C will be compared with the elements in P. If there are similarities where the P element can be identified by the elements in C then C may be removed. If the element C does not identify any of the elements of P then these elements are kept in the Detector Set M. After generating a set of M sensors, the next step in the APN, state in full is to monitor the system for the presence of cells or non-self-antigens. Set P contains the cells that are needed to be protected, set P can be formed from set P or from new cells or where the sets of all members of the set P are new. Set P and M are used to identify foreign cells in the system.

Clonal selection: Inspired by the natural clonal selection theory which is proposed by Burnet in 1959, CLONALG (Clonal Selection Algorithm) is designed by De Castro and von Zuben (2000). CLONALG has been successfully applied to deal with numerous complex computational problems (Wang *et al*., 2004). The natural clonal selection theory explains how an immune response is mounted when a non-self-antigenic pattern is recognized by the B cells. This theory is based on the response of the immune system to antigen stimulation. Clonal selection occurs in T and B cells. When the antibody is bound to the antigen, the antibody will be activated and differentiated into plasma or memory cells (De Castro and Timmis, 2002b). One of the important principles in clonal selection is that only cells that are capable of recognizing an antigen will grow in numbers. CLONALG generates an N population of antibodies and each set releases random solutions for the process optimization. Over several iterations, some of the best results are chosen where mutation is doubled and it is trying to become a population of the candidate solution as a whole. The new antibodies are then evaluated and the best antibodies will be added to the native population and the number of antibodies that have the lowest value will be replaced with an antibody derived from a new population which has a better value (De Castro and von Zuben, 2002).

As researchers know, NSA has no evolutionary properties, such as cloning that can be used for optimization. There are two well-known principles where the AIS which has been applied for optimization and which is the immune network theory (De Castro and Timmis, 2002a) and clonal selection principle (De Castro and von Zuben, 2002). In this study, the focus is on clonal selection. The immune network has a lot of potential and will be studied in future planned projects. Cloning and mutation are two important features of CLONALG that promotes diversification. The clonal selection theory states that antibodies that are able to recognize the intruding antigens will be selected to proliferate by cloning. Besides undergoing the cloning process, the antibodies will be hyper mutated and the ones with better affinity will be selected while the random antibodies will be generated to enhance the variety of the population. This is where the bone marrow will be responsible to produce the antibodies. This ability for adaptation is known as selection and affinity maturation by hyper mutation or clonal selection (Garrett, 2004).

The NSCL algorithm: In order to understand the NSCL, it is important to understand the difference between the affinity used in AIS and the affinity introduced in NSCL. The affinity used by NSCL is not the same as there are two types of affinity measurements in NSCL which are the local (single) and the global (a set of DNA strands) affinity. In this study, the proposed algorithm described is based on the modified Negative Selection Algorithm (NSA) and the Clonal Selection Algorithm (CLONALG). A detailed and thorough discussion on the original CLONALG algorithm can be found in De Castro and von Zuben (2002). An overview of the NSCL algorithm can be summarized as follows:

```
Set the iteration parameters r, n and p
Initialization: Generate random population of T cell in P that is in
compliance with the constraints
Move population P into C as Selfdata
Detector Generation
Cycle: While iteration <r, DO
T cell presentation:
    For each T cell (DNA strand) in P, DO
    Determine its local affinity in P
    Move the best T cell in P into M as the new detector
Clonal expansion: Clone T cell with the best affinity in M. The clone size
is the pre-determined n strands and must be in compliance with the
constraints
Affinity maturation: Mutate all cloned T cell in M
Meta-dinamics:
Randomly create a new population of T-cells and insert into M
    For each T cell in M, DO
        Determine its affinity
        Compare with low quality elements in C
        If a T cell of M has a better Global affinity then a C's Selfdata
            Eliminate the C's Selfdata and move the better T cell from
            M to C
            Calculate the global affinity of C
            Update P with T cell in C
Clear P and C
r ++
```

The objective of the NSCL is to find a good set of DNA strands that satisfy the given constraints. NSCL is designed to obtain a set of DNA sequences where each sequence is unique or cannot be hybridized with other sequences in the set. In this research ions, namely $H_{measure}$ and similarity are chosen to estimate the uniqueness of each DNA sequence. Another two additional functions, the hairpin and the continuity are used to prevent the secondary structure of a DNA sequence. $GC_{content}$ and melting temperature are used as the constraints where the ranges for these constraints are set by user's preference. The formulations for all objectives and constraints are displayed in Table 4. There are three parameters that need to be predefined which are r, n and p. After setting the parameters, the first step is the initialization which is to generate a set of T cells (DNA strands) randomly but must satisfy the given constraints before a strand can be accepted as part of the first set P of T cells which in NSA is termed as selfdata.

After the initialization process, the next phase is a NSA process known as Detector generation. An interesting aspect of this phase is that it is responsible for managing a population of DNA strands (i.e., immature T cells) that do not have any similarity with other cells in order to find a unique DNA sequence. This problem is known as self-nonself discrimination. The detector generation phase is a cyclic process and the number of iterations (r) of the detector generation process has to be determined before the NSCL starts operating. While the iteration is less than r, the affinity for each T cell in P will be determined before it is presented to each selfdata in C. If the T cell pattern matches a selfdata in C and any detector of M, the T cell will be removed or deleted and if it is not similar, the T cell will be moved into memory, M as a new detector for further processes. The next phase, clonal expansion is the evolutionary process where the best T cell is cloned. The size of clone is a predetermined parameter n. After the cloning task, the affinity maturation will take place where all the cloned T cells will be mutated to diversify the DNA strands sequence. The mutation task is still guided by the rules and has to satisfy the given constraints.

The next phase is detector application and suppressions. In this phase, the NSCL will measure all affinity of the T cells and then compare each T cell with each element in C. If any elements of M are similar to a selfdata in C, it will be deleted while a unique detector will not be removed from M. However, if the affinity of the detector is better than any of the selfdata in C, the better detector (T cell) will replace the selfdata position in C. When the phase of detector application and suppressions is completed, the global affinity of C will be determined and the memory of P will be cleared for a new generation of T cells. The detector generation will repeat the same process until a pre-specified number of iterations are reached.

RESULTS AND DISCUSSION

The sequence generated by generate-and-test algorithm is performed on a computer equipped with 1.86 GHz processor and 2 GHz RAM. The programme is developed using the Visual Basic net. The results obtained are compared with the sequence generated by Kurniawan (2009). All the variables for the generation of the DNA sequences are listed in Table 5.

In this experiment, the parameters value for $H_{measure}$ and similarity (continuous parameter) and (h_{con} and S_{con}) values were fixed at 6 while discontinuous parameter (h_{dis} and S_{dis}) is 0.17, continuity is 2. The hairpin formation was set at least 6 base-pairings and a 6 base loop. The Melting temperature was calculated using the Nearest-Neighbour (NN) method with 1 M salt concentration and 10 nM DNA concentration is set between 30-80°. The parameter settings for NS-CL are as follows: Number of selection is 1 sequence; cloning size is set to 30.

The comparison of the DNA strands generated by PACO and NS-CL are shown in Table 6. The comparison of the results is diplayed in Table 7 and when visually compared in Fig. 3, they clearly show that the sequences designed by NS-CL have higher $H_{measure}$ average value than the sequences generated by PACO. However, the sequences in this experiment show lower similarity average value than the sequences designed by PACO. The sequences designed by PACO and NS-CL have similar value of hairpin that equals to zero but however, the average value of continuity for NS-CL is larger than PACO which is 18.

Based on the results, the sequences obtained have a lower average value at 67.3 than the sequences produced by PACO at 75.3. As an overall, it can be concluded that NS-CL performs better than PACO.

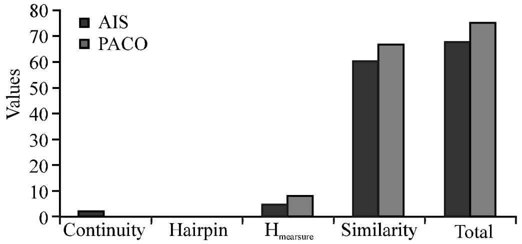

Fig. 3: Comparison of results

Table 5: Parameter settings for generation of DNA sequences

Constraints (allowable range)	Values
No. of sequences (1-10)	7
Length of sequences (8-20)	20
Continuity (1-3)	t = 2
Hairpin	Rmin = 6, Pmin = 6
H-measure	h_{con} = 6, h_{dis} = 0.17%
Similarity	S_{con} = 6, S_{dis} = 0.17%
GC percentage	Min = 20%, Max = 80%
Melting temperature method	Nearest-Neighbour (NN) Min = 30°, Max = 80°
Na +	1 M
C_t	10 nM

Table 6: DNA strands comparison between PACO and AIS method

AIS proposed method		PACO method	
Sequences	Total	Sequences	Total
TTCTCTATTCTTCTTGTTCT	58	GCAGAACACACACCACCAAC	72
TCTGTGTCTCCTTCTTCCTA	62	CACACACACACACACACGAA	72
CCTTCGTTCCTTCCTGCGTC	71	ACACCACAACACCACATAGC	77
TCTCTTCACACTCCTCTTCT	65	CAAGAGAACAACAACCAAGC	78
TCCGCTCCCTTCCGTCCTTC	69	CCACCACCACCACCACTACA	75
ATCTTGTCCTCTTCTCTCTT	72	CACACAAGACACCACAACAG	79
CTTCCATCCTTCACCTCTTT	74	TACAAGACACACAAGACACA	74
Total	471	Total	527
Average	67.3	Average	75.3

Table 7: comparison of each fitness function

Fitness function	PACO method	AIS proposed method
$H_{measure}$	61	32
Similarity	466	421
Continuity	0	18
Hairpin	0	0
Total	527	471
Average	75.3	67.3

CONCLUSION

In this study, the proposed algorithm for the DNA sequence design problems have been described in which each base (A, T, C, G) in the DNA strand represents a structure of a T cell that needs to be matured in the Thymus. NS-CL is designed based on the Negative Selection Algorithm (NSA) and Clonal Selection Algorithm (CLONALG). The main objective of this integration is that the NSA does not have the evolutionary properties that can be used to mature the T cell inside the thymus. Cloning and mutation helps NS-CL to diversify the search for better DNA strands created randomly in the initialization process of NS-CL. The results of the experiment are compared with PACO which is an initial experiment of the proposed algorithm. The studies conducted show that the proposed NS-CL algorithm could be used to design a set of DNA sequences for DNA computing. The sequences generated based on this algorithm are better than the sequences designed by PACO. However, these results have not been compared to other advanced algorithms as this study is just an attempt to explore ideas. Hence for future studies, the NS-CL would be enhanced in order to be compared with other advanced optimization algorithms used in the DNA sequence design, such as the Particle Swarm Optimization.

ACKNOWLEDGEMENT

This research is wholly financially supported by Universiti Kebangsaan Malaysia GGPM-2012-001.

REFERENCES

Adleman, L.M., 1994. Molecular computation of solutions to combinatorial problems. Sci., 266: 1021-1024.

Alba, E. and G. Luque, 2008. A hybrid genetic algorithm for the DNA fragmentassembly problem. Recent Adv. Evol. Comput. Comb. Optimiz., 153: 101-112.

Arora, J.S., 2004. Introduction to Optimum Design. 2nd Edn., Academic Press, New York.

Brenneman, A. and A. Condon, 2002. Strand design for biomolecular computation. Theor. Comput. Sci., 287: 39-58.

Dasgupta, D., 1998. Artificial Immune Systems and Their Applications. 1st Edn., Springer, Heidelberg.

De Castro, L.N. and F.J. von Zuben, 2000. The clonal selection algorithm with engineering applications. Proceedings of the Conference on Genetic and Evolutionary Computation, Workshop on Artificial Immune Systems and their Applications, July 2000, Las Vegas, USA., pp: 36-37.

De Castro, L.N. and F.J. von Zuben, 2002. Learning and optimization using clonal selection principle. IEEE Trans. Evolutionary Comput., 6: 239-251.

De Castro, L.N. and J. Timmis, 2002a. An artificial immune network for multimodal function optimization. Proceedings of the 2002 Congress on Evolutionary Computation, Volume: 1, May 12-17, 2002, Honolulu, HI., pp: 699-704.

De Castro, L.N. and J. Timmis, 2002b. Artificial Immune Systems: A New Computational Intelligence Approach. Springer-Verlag, New York.

Deaton, R., R.C. Murphy, M. Garzon, D.R. Franceschetti and S.E. Stevens Jr., 1996a. Genetic search for reliable encodings for DNA-based computation. Proceedings of the 1st Conference on Genetic Programming, July 28-31, 1996, Stanford University, USA., pp: 159-171.

Deaton, R., R.C. Murphy, M. Garzon, D.T. Franceschetti and S.E. Jr. Stevens, 1996b. Good encodings for DNA-based solutions to combinatorial problems. Proceedings of the 2nd Annual Meeting on DNA Based Computers, June 10-12, American Mathematical Society, Princeton University, pp: 159-171.

Fang, S.C., Y. Wang and J. Zhong, 2005. A genetic algorithm approach to solving DNA fragment assembly problem. J. Comput. Theor. Nanosci., 2: 499-505.

Garrett, S.M., 2004. Parameter-free, adaptive clonal selection. Proceedings of the Congress on Evolutionary Computation, Volume 1, June 19-23, 2004, Portland, OR., USA., pp: 1052-1058.

Garzon, M., P. Neathery, R. Deaton, R.C. Murphy, D.R. Franceschetti and S.E. Stevens Jr., 1997. A new metric for DNA computing. Proceedings of the 2nd Annual Conference on Genetic Programming, July 13-16, 1997, Stanford University, CA, USA., pp: 472-478.

Kashiwamura, S., A. Kameda, M. Yamamoto and A. Ohuchi, 2003. Two-step search for DNA sequence design. Proceedings of the International Technical Conference on Circuits/Systems, Computers and Communications, July 7-9, 2003, Phoenix Park, Korea, pp: 1815-1818.

Kikuchi, S. and G. Chakraborty, 2006. Heuristically tuned GA to solve genome fragmentassembly problem. Evolutionary Computation, IEEE pp: 1491-1498.

Kobayashi, S. and T. Kondo, 2002. On template method for DNA sequence design. Proceeding of 8th International Meeting on DNA Based Computers, June 10-13, 2002, Springer-Verlag, Berlin, pp: 205-214.

Kurniawan, T.B., 2009. A population based ant colony optimization approach for DNA sequence design. Proceedings of the 3rd Asia International Conference on Modelling and Simulation, May 25-29, 2009, Bandung, Bali, Indonesia, pp: 246-251.

Lucia, Jr. J.S., 1998. A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of National Academy of Science, Feb. 17-17, USA., pp: 1460-1465.

Nebro, A.J., G. Luque, F. Luna and E. Alba, 2008. DNA Fragment assembly using a grid-based geneticalgorithm. Comput. Operat. Res., 35: 2776-2790.

Parsons, R. and M.E. Johnson, 1995. DNA sequence assembly and genetic algorithms-newresults and puzzling insights. Proc. Int. Conf. Intell. Syst. Mol. Biol., 3: 277-284.

Ravi, V. and Sanjay, 2011. DNA sequence assembly using particle swarm optimization. Int. J. Comput. Appl., 28: 33-38.

Shin, S.Y., I.H. Lee, D. Kim and B.T. Zhang, 2005. Multiobjective evolutionary optimization of DNA sequences for reliable DNA computing. IEEE Trans. Evolut. Comput., 9: 143-158.

Wang, X., X.Z. Gao and S.J. Ovaska, 2004. Artificial immune optimization methods and applications -A survey. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, October 2004, The Hague, The Netherlands, pp: 3415-3420.