Ion Motion Optimization Algorithm Applied to CpG Island Prediction

Cheng-Hong Yang and Ying-Jui Hao
Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
Email: chyang@cc.kuas.edu.tw, xuite0711@gmail.com

Li-Yeh Chuang
Department of Chemical Engineering & Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan
Email: chuang@isu.edu.tw

Abstract—In a normal cell, CpG islands are usually unmethylated. When cell is changed by diseases, CpG islands are transformed into hypermethylation. Therefore, correctly predicting the locations of CpG islands provides important information in the biomedical. In this study, the ion motion optimization (IMO) algorithm was implemented to predict CpG islands called CpGIMO. A total of six contigs, chromosome 21 and chromosome 22 were tested to compare the efficiency of IMO with other methods for CpG island identification. IMO algorithm resolves local optima entrapment through the force of anion and cation. The results showed that the performance of IMO algorithm is better than the other methods.

Index Terms—methylation, CpG islands, ion motion optimization, chromosome, prediction

I. INTRODUCTION

Many diseases associated with hypermethylation in CpG islands have been confirmed. CpG islands are short sequences which are rich in Cytosine (C) and Guanine (G) in DNA. Tykocinski and Max proposed that CpG island regions in the genome contain the restriction enzyme HpaII in 1984 [1]. Gardiner-Garden and Frommer (GGF) proposed the definition of CpG islands in 1987, and the features are: (i) the length of the region over 200 bp, (ii) GC content above 50%, and (iii) the observed/expected (O/E) ratio is higher than 0.6 [2]. Takai and Jones proposed the problem of Alu (Arthrobacter luteus) in 2002, and defined strict conditions of CpG islands including the length at least 500 bps, GC content more than 55%, and the O/E ratio of 0.65.

Many methods have been proposed to predict CpG islands such as CpGIS [3], CpGProD [4], CpGcluster [5], CpGPPlot [6] and CPSORL [7]. CpGPPlot predicting CpG islands by sliding window approach is proposed by Rice, P. et al in 2000. It defines ten windows which are over than 200 bps, and calculates the expectation of window. The formula of expectation is calculated as the number of ‘C’s in the window multiplied by the number of ‘G’s in the window divided by the window length.

So far, there are many methods to solve the optimization problems such as differential evolution (DE) [8], particle swarm optimization (PSO) [10], and Genetic Algorithm (GA) [11]-[13]. These methods have been applied to many different fields, for example, function optimization [14,] [15], parameter optimization [16], and traveling salesman problem (TSP) [17] et al. Ion motion optimization (IMO) algorithm is similar to particle swarm optimization. IMO is published by Javidy and Hatamlou in 2015 [18]. The main characteristic of IMO is that it has two types of charged particles (ions) called anion (ions with negative charge) and cation (ions with positive charge). The ions have two phases which are liquid and solid. In liquid phase, ions can move to each other freely by the force. The opposite ions have attraction forces and the similar ions have repulsion forces in liquid phase [19]. In solid phase, ions have ceased and gathered at an optimal point. However, it could be a local optimal, so the authors proposed a mechanism which is producing repulsion forces to crack the solid. The advantages of IMO have the least number of tuning parameters, low computational complexity, and fast convergence.

In the study, we propose IMO algorithm for CpG island prediction and use some sequences including 6 contigs, chromosome 21 and chromosome 22 to compare IMO with other methods. The results show that CpGIMO obtained a higher performance coefficient and a higher correlation coefficient in all selected experimental contigs than the methods of CpGPSORL, CpGIS, CpGProd, CpGcluster and CpGPPlot.

II. METHOD

A. Ion Motion Optimization (IMO) Algorithm

In order to improve the balance between convergence rate and quality of the final solution, the researchers proposed the method of Ion Motion Optimization (IMO). The inspiration of IMO algorithm is from the properties of ions in nature. Charged ions have two types: the ion with negative charge called anion and the ion with positive charge called cation. Charged ions move continuously by the force to find an optimal point.
The processes of IMO are introduced, including (1) Initialization, (2) Fitness evaluation, (3) Update the best ion, (4) Calculate the force, and (5) Update the position. The pseudo code of IMO is shown below.

1) Randomly initialize the position of ions
In the study, we first initialize the ions into two types which are anions and cations randomly. The position of ions is $x_i = \{x_{i1}, x_{i2}, \ldots, x_{id}\}$, and where $i$ represents the ion index and $d$ indicates the dimension index.

2) Calculate ions of the fitness
The fitness is the position of an ion calculating the length of CpG islands plus GC content and Observed/expected by the formula 1-4.

   a) Length of CpG islands formula:
   $\text{length}_{CpG} = \frac{\text{length}_{\text{min}} + \text{length}_{\text{max}}}{\text{length}_{\text{max}}} - \frac{\text{length}_{\text{min}}}{\text{length}_{\text{max}}}$

   b) GC content formula:
   $GC(P) = \frac{\#C + \#G}{\#A + \#T + \#C + \#G}$

   c) Observed/expected (O/E) ratio formula:
   $\frac{\text{Obs}_{CpG}/\text{Exp}_{CpG}(P)}{\text{CpG}_{\text{length}}(P) \times \text{CpG}_{\text{length}}(P)}$

   The fitness value of the $i$th ion is calculated as follows:
   $\text{fitness} = \text{CpG}_{\text{length}}(P_i) + GC(P_i) + \frac{\text{Obs}_{CpG}/\text{Exp}_{CpG}(P_i)}{\text{CpG}_{\text{length}}(P_i)}$

3) Update the best ion:
The best ion is the highest fitness in all ions at once iteration. The best cation is called “Cbest”, and the best anion is called “Abest”.

4) Calculate the force
The opposite ions have attraction forces and the similar ions have repulsion forces in liquid phase [19]. In Fig. 1, dotted line means the repulsion and real line means the attraction. Where $i$ is the ion index and $d$ means the dimension index, $AD$ is the distance between an anion and the best cation, $CD$ is the distance between an anion and the best cation, $AF$ is the attraction force of anions, $CF$ is the attraction force of cations.

   a) Distance between an anion and the best cation:
   $AD_{i,j} = |A_{i,j} - C_{\text{best}}|$ (7)

   b) Distance between a cation and the best anion:
   $CD_{i,j} = |D_{i,j} - A_{\text{best}}|$ (8)

5) Update the position
The update position is done as Eq. 9-10.

The pseudo code of the IMO
1. Begin;
   2. Randomly initialize anions and cations
   3. While (the stopping criterion is not met);
      4. Evaluate fitness of anions and cations;
      5. For $n = 1$ to number of ions;
         6. Find $A_{\text{best}}$ and $C_{\text{best}}$
         7. Find $A_{\text{worst}}$ and $C_{\text{worst}}$
         8. For $d = 1$ to number of dimension of ions
            9. Calculate the force of anions and cations by Eq. 5-8
            10. Update the position of anions and cations by Eq. 9-10
      11. next $n$
   12. next $d$
   13. next generation until stopping criterion
   14. End;

III. RESULTS

A. Parameter Settings
In IMO algorithm, only two parameters need to be set: the number of iterations and the population size. In this study, the parameters for the population size is 300 (each of anions and cations are 150) [20], the iteration number is 100 [7].

B. Data Sets
In the study, we used contigs NT_113952.1, NT_113953.1, NT_113954.1, NT_113955.2, and NT_113958.2 in chromosome 21 and NT_113958.2 in chromosome 22 to test all methods. In long sequence test, we utilized the entire sequence of chromosome 21 and 22. The verified CpG islands were used to collate the CpG islands prediction. The data sources including contigs, chromosomes and verified CpG islands were obtained from NCBI (http://www.ncbi.nlm.nih.gov).

### TABLE I. CONTINGENCY TABLE

<table>
<thead>
<tr>
<th>Predicted Condition</th>
<th>True condition</th>
<th>$p$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p'$ True position</td>
<td>$n'$ False position</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n'$ False negative</td>
<td>True negative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
C. CpG Island Prediction in Six Contigs

We propose IMO algorithm to predict CpG islands, and compare the results with other methods. All methods use the Gardiner-Garden and Formmer (GGF) definition to detect the CpG islands. The definition is according to the properties: (i) the length of the region over 200 bps, (ii) GC content above 50%, and (iii) the observed/expected (O/E) ratio is higher than 0.6 [2]. First, the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) were determined (Table I). Second, the accuracy (ACC), sensitivity (SN), specificity (SP), performance coefficient (PC) and correlation coefficient (CC) were calculated through the formulas 11-15. Table II shows the results of all methods predicted in six contigs and the performance of CpGIMO was higher than other methods in a part of contigs. In Table II, we saw that the CC of IMO method was the highest in the contigs of NT_113952.1 (85.54%), NT_113953.1 (74.58%), NT_11394.1 (74.58%), NT_113955.2 (88.7%), NT_113958.2 (81.6%) and NT_028395.3 (74.2%).

\[
ACC = \frac{TP + TN}{TP + FP + TN + FN}
\]

\[
SN = \frac{TP}{TP + FN}
\]

\[
SP = \frac{TN}{TN + FP}
\]

\[
PC = \frac{TP}{TP + FP + FN}
\]

\[
CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}}
\]

D. CpG Island Prediction in Chromosomes 21 and 22

Chromosomes 21 and 22 have been widely used in many literatures; therefore, we use the data to identify the results of each method. Table III shows the results of long sequences in chromosomes 21 and 22 including the length of CpG islands, the coverage of CpG islands, average length, minimum length, maximum length, number of islands predicted, GC content, and O/E ratio. The average of GC content is higher than 50% and O/E ratio is higher than 0.6, therefore, it conforms to the GGF standard.

The length of true CpG islands in Chromosome 21 is 1719555 bps and the length of true CpG islands in chromosome 22 is 3114716 bps. The prediction length of CpGPlot, CpGcluster and CpGProD has significant gaps for true CpG islands. CpGcluster is merged the distance of CpG sites. In addition, Table III shows the minimum length in CpGcluster is 8 bps. It could make the prediction accuracy glissade in long sequence.

E. CpG Island Prediction in the ROC Space

In Fig. 2, the true positive rate (TPR) and false positive rate (FPR) were calculated through the formulas 16 and 17 to create the distribution of xy [21]. TPR is defined to y-axis, and FPR is defined to x-axis in the receiver operating characteristic (ROC) space. When FPR is as possible as low and TPR is as possible as high, the accurate rate is higher. In other words, the point is closer to the upper left corner coordinates; the accuracy of the result is higher. Fig. 2 shows that CpGcluster is at the bottom left corner coordinates in all contigs. The TPR of CpGIMO is higher than other methods, and the FPR of CpGIMO is more left than other methods. Therefore, CpGIMO provided higher accuracy than other methods.

\[
FPR = \frac{FP}{FP + TN}
\]

\[
TPR = \frac{TP}{TP + FN}
\]
IV. DISCUSSION

Currently, many famous methods use sliding window approach to predict CpG islands such as CpGIS, CpGPlot, and CpGProD. They use fixed-length to predict CpG islands, but the fixed-length window could influence the prediction results. Hackenberg proposed the window size have a great influence in the CpG island prediction. Therefore, we proposed IMO algorithm, and the definition is based on Gardiner-Garden and Frommer (GGF). We set a 200 to 2000 bp window to predict CpG islands randomly, and the results (Table II) show that the detection efficiency is higher than other methods.

CpGcluster merges CpG sites each other by the percent of distance. Thus, it is suitable for the short sequence prediction, but not for the long sequence prediction. The SP of CpGPlot is higher than other methods in all contigs. Because the length of CpG island prediction in CpGPlot is shorter than the length of true CpG islands, it has minimal FN, and gets a higher SP. However, it has maximum FP, and gets a lower SN that causes it not grow in all performances.

CPSORL is based on particle swarm optimization and sliding window approach, and combines with complementary and Reinforcement Learning (RL) method. It updates the velocity by the formula wherein \( r_1 \) and \( r_2 \) are random number \([0, 1]\). It could be unstable for the updating particle position, and caused the local optima. Combing with many methods also leads to increase operating complexity. In order to reduce operating speed, we use IMO algorithm to predict CpG islands. In fact, CpGIMO indeed enhanced the computing effectiveness, and improved CpG island detection.

<table>
<thead>
<tr>
<th>Contig</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CpGPlot</td>
</tr>
<tr>
<td>NT_113952.1</td>
<td></td>
</tr>
<tr>
<td>SN</td>
<td>56.43</td>
</tr>
<tr>
<td>SP</td>
<td>100.0</td>
</tr>
<tr>
<td>ACC</td>
<td>98.09</td>
</tr>
<tr>
<td>PC</td>
<td>56.42</td>
</tr>
<tr>
<td>CC</td>
<td>74.38</td>
</tr>
<tr>
<td>NT_113955.2</td>
<td></td>
</tr>
<tr>
<td>SN</td>
<td>47.19</td>
</tr>
<tr>
<td>SP</td>
<td>100.0</td>
</tr>
<tr>
<td>ACC</td>
<td>98.08</td>
</tr>
<tr>
<td>PC</td>
<td>47.14</td>
</tr>
<tr>
<td>CC</td>
<td>67.94</td>
</tr>
<tr>
<td>NT_113958.2</td>
<td></td>
</tr>
<tr>
<td>SN</td>
<td>51.29</td>
</tr>
<tr>
<td>SP</td>
<td>99.99</td>
</tr>
<tr>
<td>ACC</td>
<td>96.90</td>
</tr>
<tr>
<td>PC</td>
<td>51.24</td>
</tr>
<tr>
<td>CC</td>
<td>70.38</td>
</tr>
<tr>
<td>NT_113953.1</td>
<td></td>
</tr>
<tr>
<td>SN</td>
<td>22.80</td>
</tr>
<tr>
<td>SP</td>
<td>100.0</td>
</tr>
<tr>
<td>ACC</td>
<td>97.76</td>
</tr>
<tr>
<td>PC</td>
<td>22.80</td>
</tr>
<tr>
<td>CC</td>
<td>47.21</td>
</tr>
</tbody>
</table>
and crossover, or hybridizing with other algorithms. In the future work, it might be possible to improve the performance of the proposed algorithm using chaotic maps, evolutionary operators such as mutation and crossover, or hybridizing with other algorithms.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Island length (bp)</th>
<th>Island coverage (%)</th>
<th>Number of islands predicted</th>
<th>Total length of CpG islands</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT_028395.3</td>
<td>31.24</td>
<td>99.94</td>
<td>98.72</td>
<td>97.62</td>
</tr>
<tr>
<td>ACC</td>
<td>97.47</td>
<td>96.90</td>
<td>97.00</td>
<td>96.83</td>
</tr>
<tr>
<td>PC</td>
<td>31.24</td>
<td>26.19</td>
<td>38.94</td>
<td>47.05</td>
</tr>
<tr>
<td>CC</td>
<td>55.17</td>
<td>43.81</td>
<td>54.68</td>
<td>63.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Island length (bp)</th>
<th>Island coverage (%)</th>
<th>Number of islands predicted</th>
<th>Total length of CpG islands</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT_113954.1</td>
<td>27.11</td>
<td>94.89</td>
<td>54.18</td>
<td>76.68</td>
</tr>
<tr>
<td>SP</td>
<td>100.0</td>
<td>99.47</td>
<td>99.45</td>
<td>98.93</td>
</tr>
<tr>
<td>ACC</td>
<td>97.98</td>
<td>97.53</td>
<td>98.19</td>
<td>98.14</td>
</tr>
<tr>
<td>PC</td>
<td>27.10</td>
<td>39.26</td>
<td>45.36</td>
<td>59.36</td>
</tr>
<tr>
<td>CC</td>
<td>51.51</td>
<td>57.21</td>
<td>62.26</td>
<td>73.57</td>
</tr>
</tbody>
</table>

TABLE III. COMPARISON OF THE NUMBER OF CpG ISLANDS IDENTIFIED IN THE HUMAN GENOME WITH DIFFERENT METHODS.

V. CONCLUSION

In the study, we proposed IMO algorithm to predict CpG islands. IMO algorithm searches the fitness to find the best solution through the force of anion and cation. It avoids the local optima highly by changing phases continuously until the satisfaction of an end criterion.

The advantages of IMO algorithm include the least number of tuning parameters, low computational complexity, and fast convergence. Among all the test sequences, CpGIMO obtained better prediction results for CpG islands than the other methods except in NT_028395.3. In the future work, it might be possible to improve the performance of the proposed algorithm using chaotic maps, evolutionary operators such as mutation and crossover, or hybridizing with other algorithms.

ACKNOWLEDGMENT

This study was partly supported by the National Science Council of Taiwan for Grant NSC 103-2221-E-151-029-MY3, 102-2221-E-151-024-MY3.

REFERENCES


