Computational prediction of CpG islands

  - BMC Bioinformatics 7: 446

  - Algorithms for Molecular Biology 6:2
DNA methylation

- Causes C to turn into T
- Accounts for low occurrence of CpG dinucleotides in vertebrates
  - Expectation is 6.25% randomly
  - Actually 1% of total sequence (Bird 1986)
CpG islands

In the human genome, the observed CpG frequency is around 5 times lower than expected by the GC content, most likely due to high mutation rates of methylated CpGs (between 70 and 80% of all CpGs in the human genome are methylated).

→“regions of DNA with a high G + C content and a high frequency of CpG dinucleotides relative to the bulk genome” -- Gardiner-Garden and Frommer (1987)

Properties:

• between 50-60% of all genes have a CpG island in its promoter regions

• Almost all house keeping genes have a CpG island

• CpG islands are usually hypomethylated but aberrant methylation of these regions plays an important role in epigenetic formation of many cancer types.

• CpG islands are important “attributes” in the in silico detection of promoters and to predict methylation probabilities.
CpG islands & Genes

5’ end

CpGi

Gene

Promoter CpG islands

Gene

CpG islands in body

Gene

3’ end CpG islands
Traditional finders

Sliding Windows and thresholds:

High Parameter space:
Island criterion: Observed/expected CpG ratio, G+C content, minimum island length and minimum number of CpGs (as G and CpG poor and C-rich regions will meet the O/E criterion)
Technical Parameters: window length, step size and aggregation distance

Shortcomings:
1) High Parameter space → fine tuning in various ways to eliminate spurious Alus or restrict the prediction to putative promoter islands
2) Statistically properties of CGIs are predetermined to some extent by the prediction algorithm.

From Takai and Jones (2002)
Main statistical properties of CpG islands are biased towards the lower thresholds chosen in the prediction process.

→ These variables have been used as attributes in data mining approaches to classify CGIs in promoter/no-promoter or methylation prone/non-methylated islands which may be methodologically doubtfull.

Gardiner and Frommer (1987): Stretches of DNA with a high G+C content, and a frequency of CpG dinucleotides close to the expected value, appear as CpG clusters within the CpG-depleted bulk DNA, and are now generally known as CpG islands.

→ Try to detect these clusters without using %G+C, O/E and length in the prediction process.
A distance-based algorithm for CpG-island prediction.

BMC Bioinformatics 7: 446

Read the DNA sequence

CpG \rightarrow 1; \text{other} \rightarrow 0

Binary sequence:

001000010100000010100011000010001010100011

Determine the distance ($d$) of each CpG to the next CpG downstream in the DNA sequence:

10,5,5,3,1,8,23,34,21,12,2,5,8,6,9,...N-1

Let be $d_m$ a distance threshold \(\rightarrow\) If $d_i \leq d_m \rightarrow\) Cluster seed

For example, for $d_m = 5$:

10,\underline{5},\underline{5},3,1,8,23,34,21,12,\underline{2},5,8,6,9,...N-1

List of CpG clusters with coordinates, length and number of CpGs

Assign a $P$-value to each CpG cluster

Statistically significant cluster \(\equiv\) CpG island

Calculate statistical sequence properties: G+C content, O/E ratio, CpG density, intra-clustering of CpGs, overlap with Alus, PhastCons etc.
Distribution of distances among CpGs

If CpGs are distributed totally at random along the chromosome sequence, the distances between neighboring CpG dinucleotides should follow the geometric distribution:

\[ P(d) = (1 - p)^{d-1} p \]

where \( P(d) \) represents the probability of finding a distance \( d \) between neighboring CpGs and \( p \) corresponds to the probability of CpGs in the sequence.
Set the distance threshold

Distances between within island CpGs and bulk CpGs should be strikingly different

→ Calculate observed and expected (geometric distribution) PDFs (probability density function)

→ Short distances are overrepresented (more likely than under random expectation) = inner island distances

→ The observed PDF subtend the expected fairly good at its median distance (in all chromosomes)

→ The median distance separates the inner CpG island distances from the rest

→ This suggest the use of the median distance as the distance threshold
Computing P-values

What is the probability to find a cluster with X CpGs and length N in a random distribution?

Negative Binomial Distribution:
Probability to get r failures (non-CpGs) when the number of successes (CpGs) is fixed in advance

\[ P \text{-value} \equiv \text{Cumulative Density Function at point } n_f \text{ (number of failures = N-2*X)} \]
Benchmark test

Compare predictions to five commonly used algorithms by means of a prepared set of test sequences (400 experimental CpG islands embedded into a random background):

• CpGcluster shows moderate Sensitivity (Sn) but reaches highest values for Specificity (Sp) and Correlation (between Sn and Sp)
• CpGcluster “hits” more islands than the other finder (at least the “core” of the island gets predicted)

→ High specificity and correlation together with highest hit percentage seems to indicate an advantage of CpGcluster

CpGcluster parameters: Median distance and P-value 1E-5

<table>
<thead>
<tr>
<th>Program</th>
<th>Sn ± SD</th>
<th>Sp ± SD</th>
<th>CC ± SD</th>
<th>Hit* [%] ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newcpgreport</td>
<td>0.545 ± 0.002</td>
<td>0.973 ± 0.002</td>
<td>0.725 ± 0.005</td>
<td>87.000 ± 0.540</td>
</tr>
<tr>
<td>CpGProD</td>
<td>0.918 ± 0.003</td>
<td>0.657 ± 0.003</td>
<td>0.772 ± 0.006</td>
<td>94.675 ± 0.808</td>
</tr>
<tr>
<td>CpGIS</td>
<td>0.832 ± 0.003</td>
<td>0.756 ± 0.007</td>
<td>0.789 ± 0.013</td>
<td>86.675 ± 1.528</td>
</tr>
<tr>
<td>CpGIE</td>
<td>0.910 ± 0.002</td>
<td>0.667 ± 0.003</td>
<td>0.775 ± 0.006</td>
<td>94.650 ± 0.810</td>
</tr>
<tr>
<td>CpGED</td>
<td>0.819 ± 0.013</td>
<td>0.584 ± 0.004</td>
<td>0.685 ± 0.005</td>
<td>84.075 ± 1.191</td>
</tr>
<tr>
<td>CpGcluster (d_i = median, or 44 bp)</td>
<td>0.655 ± 0.003</td>
<td><strong>0.976 ± 0.005</strong></td>
<td>0.797 ± 0.009</td>
<td><strong>95.475 ± 0.870</strong></td>
</tr>
<tr>
<td>CpGcluster (d_i = 75th percentile, or 94 bp)</td>
<td>0.866 ± 0.006</td>
<td>0.832 ± 0.009</td>
<td>0.846 ± 0.006</td>
<td>95.050 ± 0.643</td>
</tr>
</tbody>
</table>
### Statistical properties compared to traditional finders

- Higher mean G+C content, O/E ratios and CpG densities
- Lower mean and maximum length
- Lower overlap with spurious Alu elements
- Higher overlap with conserved phylogenetic elements (PhastCons)
- Detects short CpG islands in tissue specific genes which are normally missed by other finders

<table>
<thead>
<tr>
<th></th>
<th>CpGcluster</th>
<th>CpGproD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome length (without N-runs, bp)</td>
<td>2.85E+09</td>
<td>2.85E+09</td>
</tr>
<tr>
<td>Total number of CpGs</td>
<td>28,073,991</td>
<td>28,073,991</td>
</tr>
<tr>
<td>CpG-dinucleotides in CpG-islands (%)</td>
<td>4,489,575 (15.99)</td>
<td>4,323,799 (15.40)</td>
</tr>
<tr>
<td>Number of islands predicted</td>
<td>197,727</td>
<td>76,793</td>
</tr>
<tr>
<td>Island coverage (%)</td>
<td>1.90</td>
<td>2.81</td>
</tr>
<tr>
<td>Island length (bp):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>273.5 ± 246.7</td>
<td>1043.8 ± 761.7</td>
</tr>
<tr>
<td>Minimum</td>
<td>8</td>
<td>500</td>
</tr>
<tr>
<td>Maximum</td>
<td>7,774</td>
<td>42,276</td>
</tr>
<tr>
<td>Average island GC-content (%)</td>
<td>63.76 ± 7.51</td>
<td>54.58 ± 6.12</td>
</tr>
<tr>
<td>Average CpG O/E ratio</td>
<td>0.855 ± 0.265</td>
<td>0.636 ± 0.089</td>
</tr>
<tr>
<td>Average CpG-density</td>
<td>0.087 ± 0.041</td>
<td>0.047 ± 0.016</td>
</tr>
</tbody>
</table>

### Overlap with TSS of MAGE genes

<table>
<thead>
<tr>
<th>Program</th>
<th>#CGI</th>
<th>Average length ± SD</th>
<th>% of overlap with Alus</th>
<th>% of overlap with PhastCons</th>
</tr>
</thead>
<tbody>
<tr>
<td>newcpgreport</td>
<td>2</td>
<td>271.0 ± 18.4</td>
<td>19.49</td>
<td>23.73</td>
</tr>
<tr>
<td>CpGPromD</td>
<td>3</td>
<td>1,314.3 ± 525.1</td>
<td>23.40</td>
<td>13.31</td>
</tr>
<tr>
<td>CpGIS</td>
<td>3</td>
<td>800.0 ± 243.3</td>
<td>10.52</td>
<td>20.59</td>
</tr>
<tr>
<td>CpGIE</td>
<td>3</td>
<td>1,093.0 ± 476.1</td>
<td>23.99</td>
<td>14.00</td>
</tr>
<tr>
<td>CpGED</td>
<td>2</td>
<td>730.5 ± 320.3</td>
<td>15.32</td>
<td>15.82</td>
</tr>
<tr>
<td>CpGcluster</td>
<td>8</td>
<td>258.3 ± 100.8</td>
<td>6.79</td>
<td>28.53</td>
</tr>
</tbody>
</table>
http://bioinfo2.ugr.es/CpGcluster/

http://bioinfo2.ugr.es/wordCluster/wordCluster.php
http://bioinfo2.ugr.es/CpGislands/

CpG Island database

CpG islands predicted by

CpG cluster
gatcta

Computer prediction of CpG islands