量化基因體中的重複序列

QUANTIFYING REPETITIVE SEQUENCE IN GENOMES

指導教授：賀保羅
專題成員：李哲宇

目錄

- 動機.....3
- 專題系統流程.....4
- Approach 1：Plain Markov Model.....6
- Approach 2：Hidden Markov Model.....9
- Approach 3：Autocorrelation.....11
- 結論.....18

動機

- DNA定序：分析鹼基A、T、C、G的排列方式。可被應用於研究基因組及其對應編碼的蛋白質，並快速的推動分子生物學與現代醫學。
- 其中重複的基因片段與漢亭頓氏症、小腦萎縮症等等基因疾病有關聯。研究這些高重複性的序列能夠對診斷基因疾病有著莫大的幫助。
- 本專題的目的便是運用Markov Model、HMM、Autocorrelation等等序列、訊號處理演算法，將人類基因體中重複出現的DNA序列片段數據化、圖表化。讓我們能更清楚瞭解人類基因體中的DNA重複片段。

專題系統流程

一. 建立輸入資料。本專題所使用的輸入資料為美國國家生物技術資訊中心NCBI網站提供的人類基因體FA檔如下圖。

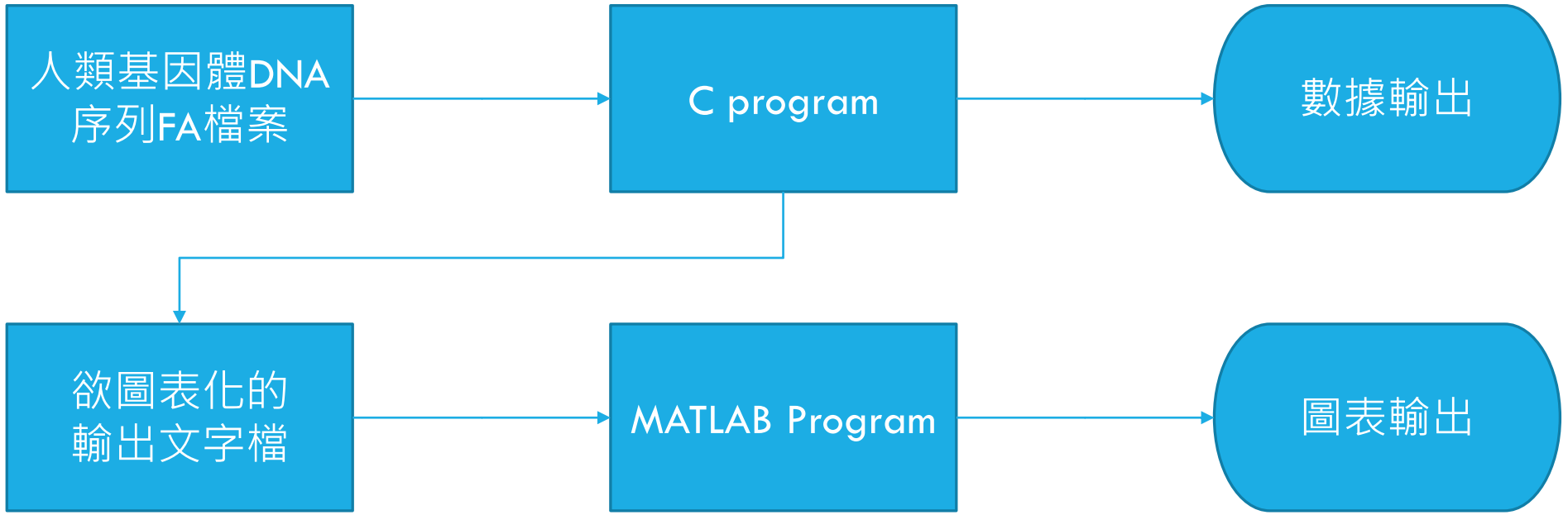
二. 由C語言程式讀取輸入，經指定演算法（後續介紹）計算後得出數據結論、或者欲做成圖表的數據文字檔（TXT檔）。

三. 由MATLAB程式讀取上述TXT檔作為輸入，將該數據轉換成折線圖並輸出。

```
>ref|NC_000006.12|:100001-1100000 Homo sapiens chromosome 6, GRCh38.p13 Primary Assembly
TTGGTACCATTCCTTCTGAAACTATTCCAAACAACAGAAAAAGAGAGAATCCTCCCTAACTCATTTTATG
AGGCCAGAATAATTCTGGTACCAAAATTTGGCAGAGACACACACACAAAAAAAGAAAATTTCAAGCCAAT
ATCCCTGATGAACATCGATGCAAAAATCCTCAATAAAATACTGGCAAACCAAATCCAGCAGCACATCAAA
AGCTTGTCCACCACAATCAAGTCGGCTTCATCCCTGGGATACAAGGCTAGTTCAACATACGCAAATCAAT
```

專題系統流程

系統架構圖：



APPROACH 1：PLAIN MARKOV MODEL

- 使用order 0至2的Markov Model，根據輸入資料手動調整Model參數，使得這些Model產生出輸入序列的機率愈大愈好。
- Order 0 = 4 state Model
- Order 1 = 16 state Model
- Order 2 = 64 state Model
- 右圖為各state轉換至A、T、C、G state之個別機率。

```
double prob0[BASE] = {0.29, 0.29, 0.21, 0.21};
double prob1[BASE][BASE] = {
    {0.30, 0.25, 0.25, 0.20},  // A + A/T/G/C
    {0.20, 0.32, 0.26, 0.22},  // T + A/T/G/C
    {0.30, 0.25, 0.25, 0.20},  // G + A/T/G/C
    {0.35, 0.32, 0.10, 0.23}   // C + A/T/G/C
};
double prob2[BASE][BASE][BASE] = {
    {{0.38, 0.24, 0.22, 0.16},  // AA + A/T/G/C
     {0.25, 0.31, 0.25, 0.19},  // AT + A/T/G/C
     {0.30, 0.22, 0.26, 0.22},  // AG + A/T/G/C
     {0.40, 0.30, 0.07, 0.23}}, // AC + A/T/G/C
    {{0.32, 0.30, 0.20, 0.18},  // TA + A/T/G/C
     {0.20, 0.38, 0.20, 0.22},  // TT + A/T/G/C
     {0.26, 0.27, 0.26, 0.21},  // TG + A/T/G/C
     {0.32, 0.36, 0.05, 0.27}}, // TC + A/T/G/C
    {{0.33, 0.21, 0.29, 0.17},  // GA + A/T/G/C
     {0.20, 0.27, 0.33, 0.20},  // GT + A/T/G/C
     {0.30, 0.21, 0.26, 0.23},  // GG + A/T/G/C
     {0.33, 0.32, 0.07, 0.28}}, // GC + A/T/G/C
    {{0.24, 0.24, 0.29, 0.23},  // CA + A/T/G/C
     {0.16, 0.27, 0.32, 0.25},  // CT + A/T/G/C
     {0.21, 0.27, 0.28, 0.24},  // CG + A/T/G/C
     {0.34, 0.33, 0.07, 0.26}}  // CC + A/T/G/C
};
```

APPROACH 1：PLAIN MARKOV MODEL

- 將人類基因體第六對染色體（ NC_000006.12 Homo sapiens chromosome 6，GRCh38.p14 ）作為主要輸入。
- 以下為第100001至1100000個鹼基序列結果：

```
level 0 model log base 2 probability: -1988107.761833
level 1 model log base 2 probability: -1949355.207585
level 2 model log base 2 probability: -1932983.620669
```

- 以下為第1100001至2100000個鹼基序列結果：

```
level 0 model log base 2 probability: -1987564.798109
level 1 model log base 2 probability: -1950023.124025
level 2 model log base 2 probability: -1933193.840351
```

APPROACH 1：PLAIN MARKOV MODEL

- 在手動調整Model參數的過程中能觀察到一些事情：A與T的出現率高於G與C；C之後出現G的機率比其他鹼基低許多。
- 比較難以觀察到短片段中重複性高的序列。

```
double prob0[BASE] = {0.29, 0.29, 0.21, 0.21};
double prob1[BASE][BASE] = {
    {0.30, 0.25, 0.25, 0.20},  // A + A/T/G/C
    {0.20, 0.32, 0.26, 0.22},  // T + A/T/G/C
    {0.30, 0.25, 0.25, 0.20},  // G + A/T/G/C
    {0.35, 0.32, 0.10, 0.23}   // C + A/T/G/C
};
double prob2[BASE][BASE][BASE] = {
    {{0.38, 0.24, 0.22, 0.16},  // AA + A/T/G/C
     {0.25, 0.31, 0.25, 0.19},  // AT + A/T/G/C
     {0.30, 0.22, 0.26, 0.22},  // AG + A/T/G/C
     {0.40, 0.30, 0.07, 0.23}}, // AC + A/T/G/C
    {{0.32, 0.30, 0.20, 0.18},  // TA + A/T/G/C
     {0.20, 0.38, 0.20, 0.22},  // TT + A/T/G/C
     {0.26, 0.27, 0.26, 0.21},  // TG + A/T/G/C
     {0.32, 0.36, 0.05, 0.27}}, // TC + A/T/G/C
    {{0.33, 0.21, 0.29, 0.17},  // GA + A/T/G/C
     {0.20, 0.27, 0.33, 0.20},  // GT + A/T/G/C
     {0.30, 0.21, 0.26, 0.23},  // GG + A/T/G/C
     {0.33, 0.32, 0.07, 0.28}}, // GC + A/T/G/C
    {{0.24, 0.24, 0.29, 0.23},  // CA + A/T/G/C
     {0.16, 0.27, 0.32, 0.25},  // CT + A/T/G/C
     {0.21, 0.27, 0.28, 0.24},  // CG + A/T/G/C
     {0.34, 0.33, 0.07, 0.26}}  // CC + A/T/G/C
};
```

APPROACH 2：HIDDEN MARKOV MODEL

- 模型構思：以order 1 Markov Model為基礎，考慮可能出現ATATAT...等等重複序列，額外加入兩個state。
- 從initial state開始後，有小機率進入上述兩個state，兩個state有高機率會互相走訪，並有高機率產生A與T鹼基，以此模擬ATATAT...此類重複序列的出現。
- 右圖為initial state機率、轉換至各state機率、及各state產生A、T、G、C之機率。

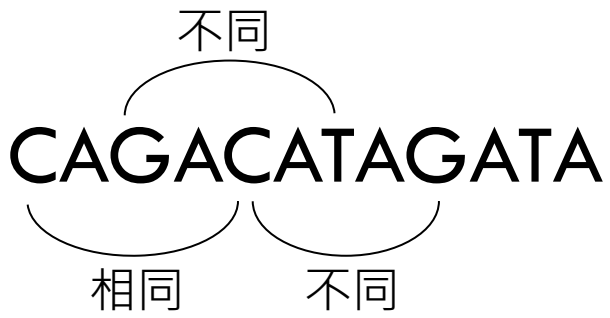
```
//set parameter of hidden Markov model
double init[STATE] = {0.29, 0.29, 0.20, 0.20, 0.01, 0.01};
double state_trans[STATE][STATE] = {
    {0.29, 0.24, 0.24, 0.18, 0.04, 0.01},
    {0.18, 0.31, 0.25, 0.21, 0.01, 0.04},
    {0.30, 0.24, 0.24, 0.20, 0.01, 0.01},
    {0.35, 0.32, 0.10, 0.21, 0.01, 0.01},
    {0.16, 0.01, 0.01, 0.01, 0.01, 0.80},
    {0.01, 0.16, 0.01, 0.01, 0.80, 0.01}
};
double bases[STATE][BASE] = {
    {0.80, 0.07, 0.07, 0.06},
    {0.07, 0.80, 0.06, 0.07},
    {0.06, 0.07, 0.80, 0.07},
    {0.07, 0.06, 0.07, 0.80},
    {0.02, 0.95, 0.02, 0.01},
    {0.94, 0.02, 0.02, 0.02}
};
```

APPROACH 2：HIDDEN MARKOV MODEL

- 以第100001至1100000個鹼基序列做輸入：經Forward Algorithm計算出log2機率為-196.546934
- 以第1100001至2100000個鹼基序列做輸入：經Forward Algorithm計算出log2機率為-196.435012
- 比 order 0 model 高，但比 order 1 model 低。
- 模型需自定義的參數過多，實作較為困難。

APPROACH 3：AUTOCORRELATION

- 定義：將輸入序列分成固定大小的Window，在指定Window size（20~100）中，計算第i個鹼基及第$i+k$個鹼基相同的機率（$1 \leq k \leq 5$）。
- 以下為範例：設$k = 4, wsize = 12$

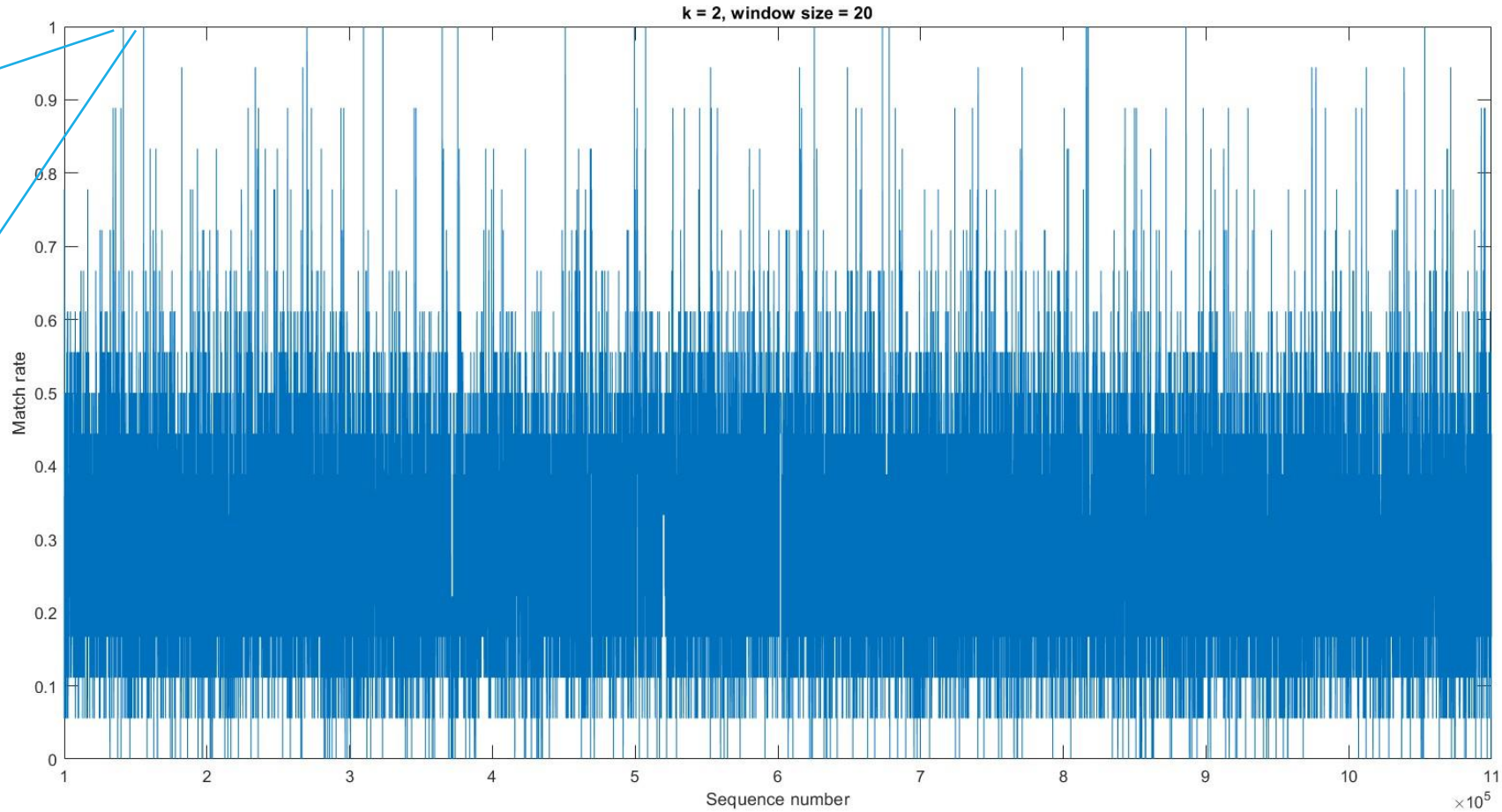


此Window的$match\ rate = 6 \div 8 = 0.75$

APPROACH 3 : AUTOCORRELATION

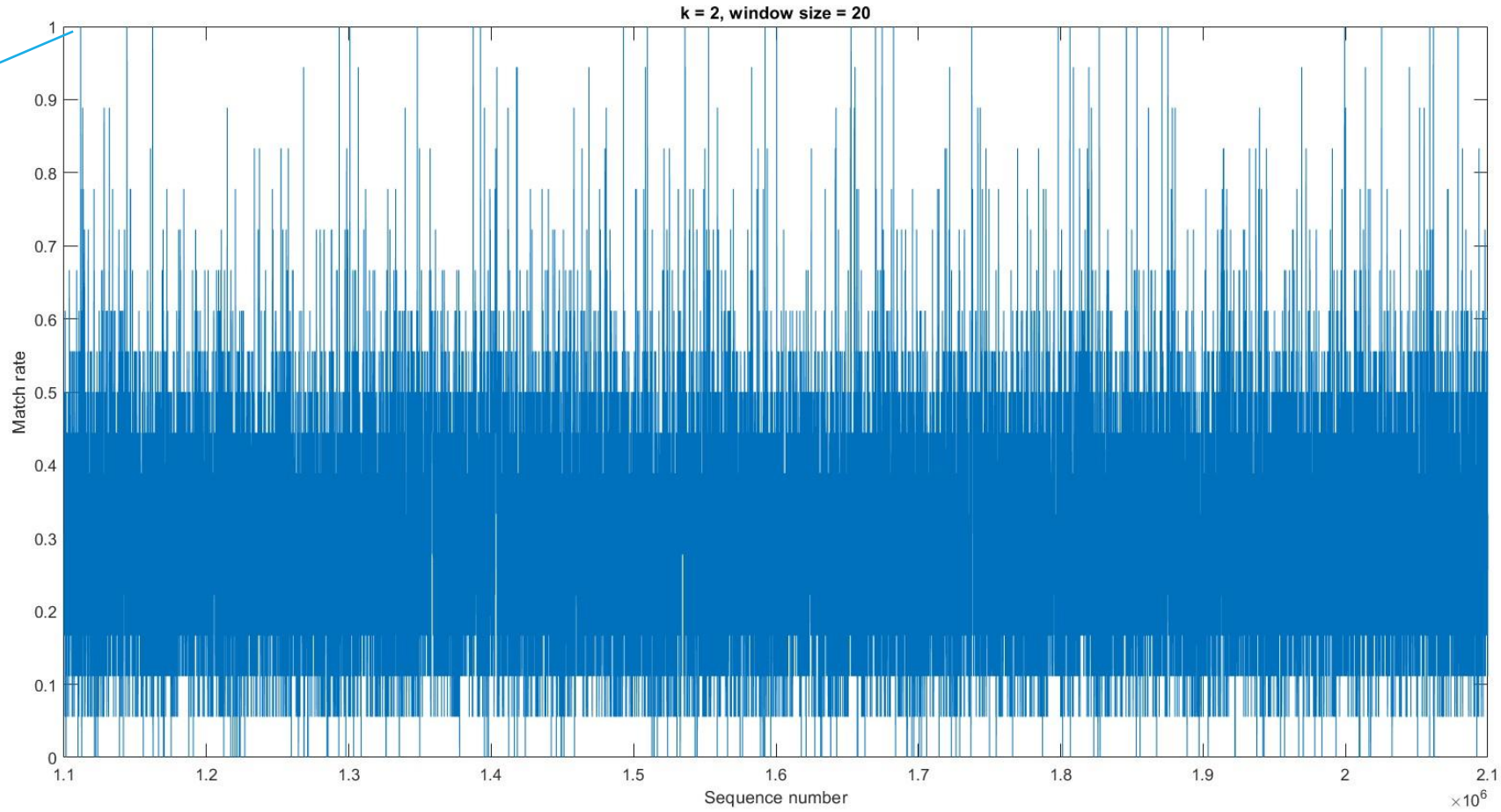
序列：
AAAAAAAAAA
AAAAAAAAAA
Match rate
= 1

序列：
TTTTTTTTTTTTT
TTTTTT
Match rate
= 1

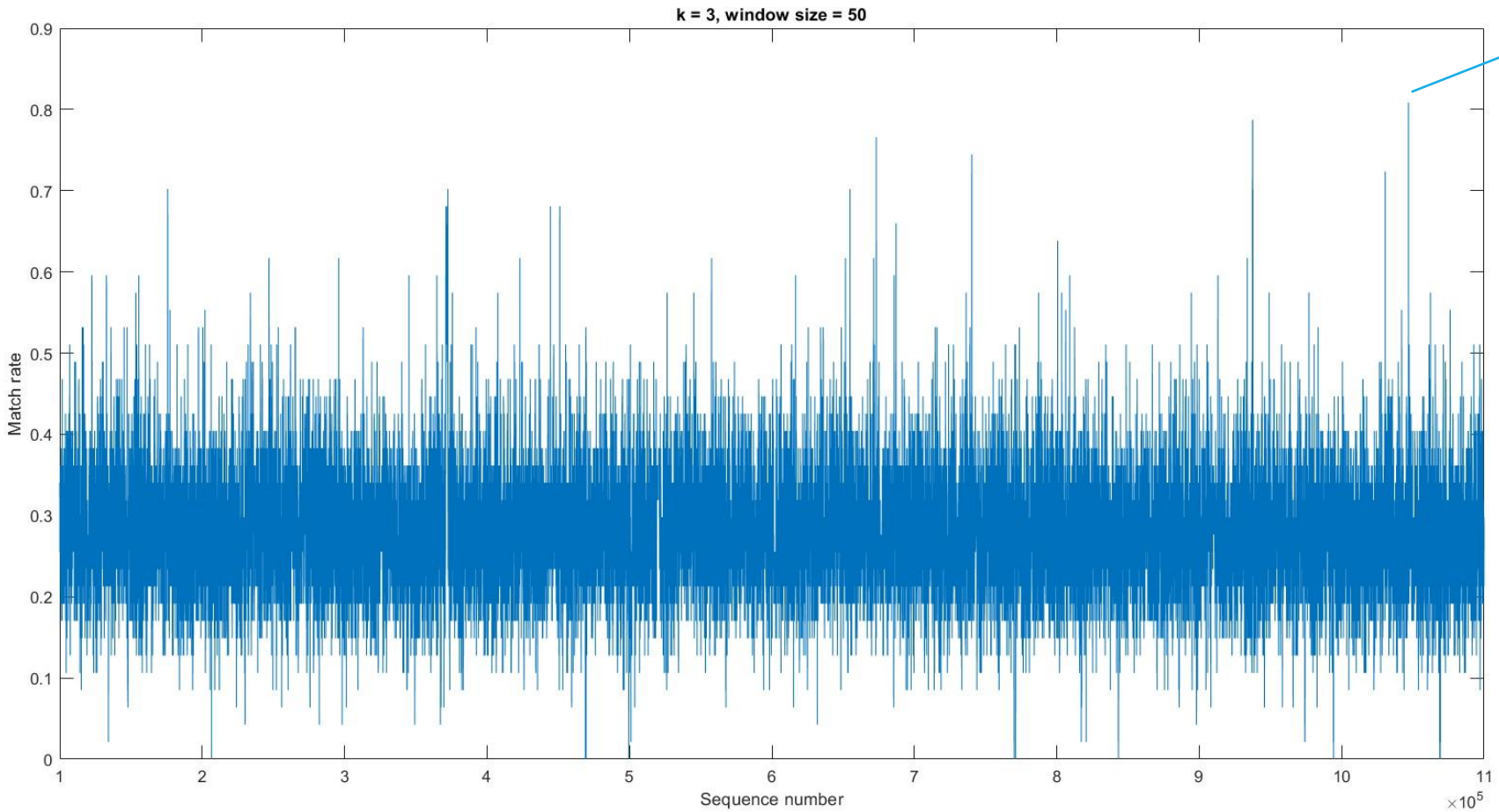


APPROACH 3 : AUTOCORRELATION

序列：
GTGTGTGTGT
GTGTGTGTGT
Match rate
= 1



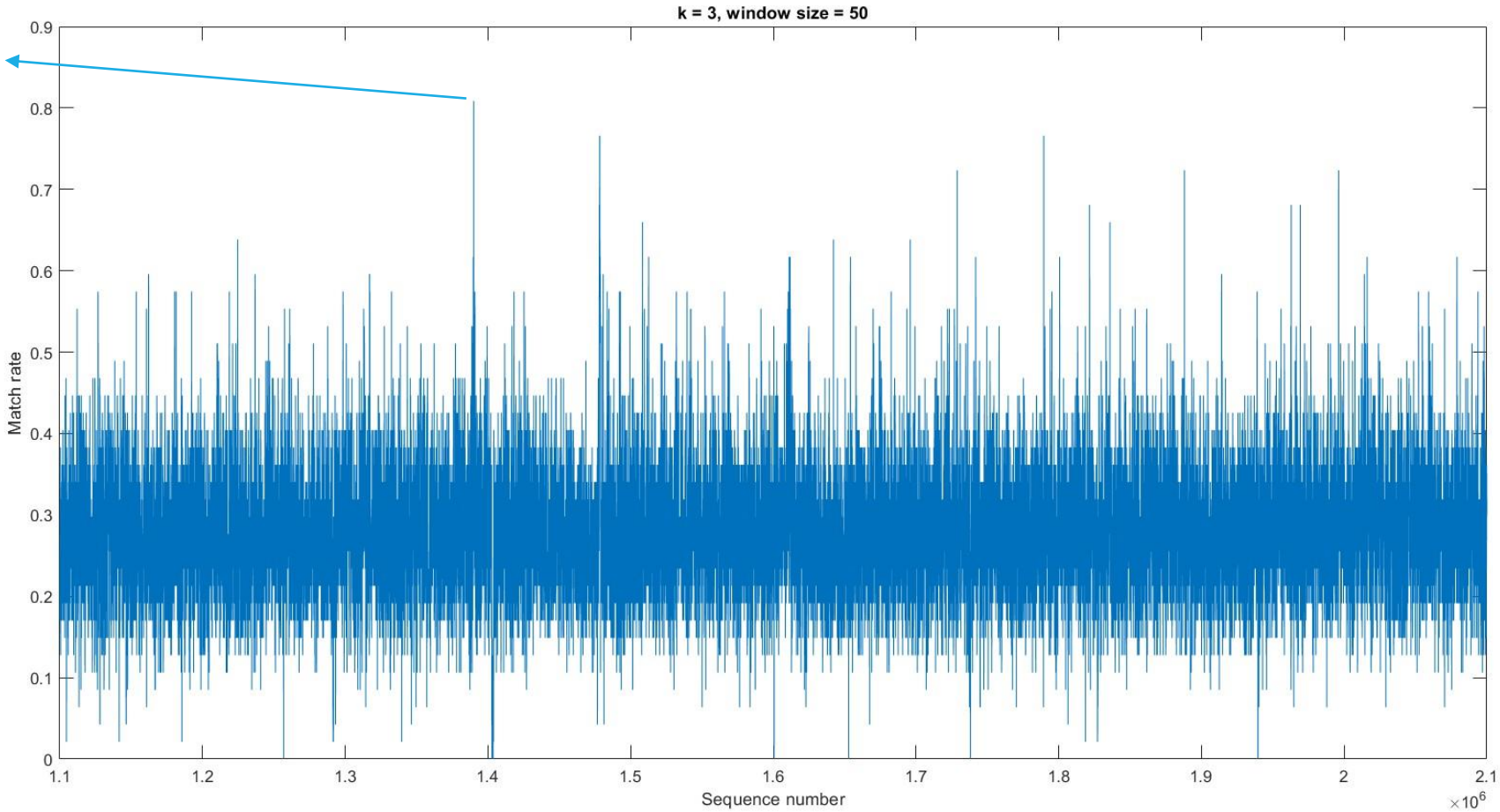
APPROACH 3 : AUTOCORRELATION



序列：
CTCAACAACAA
CAACAAAAAA
AAAAAAAAAA
AAAAAAAGAA
AGAAAGAA
Match rate
= 0.808511

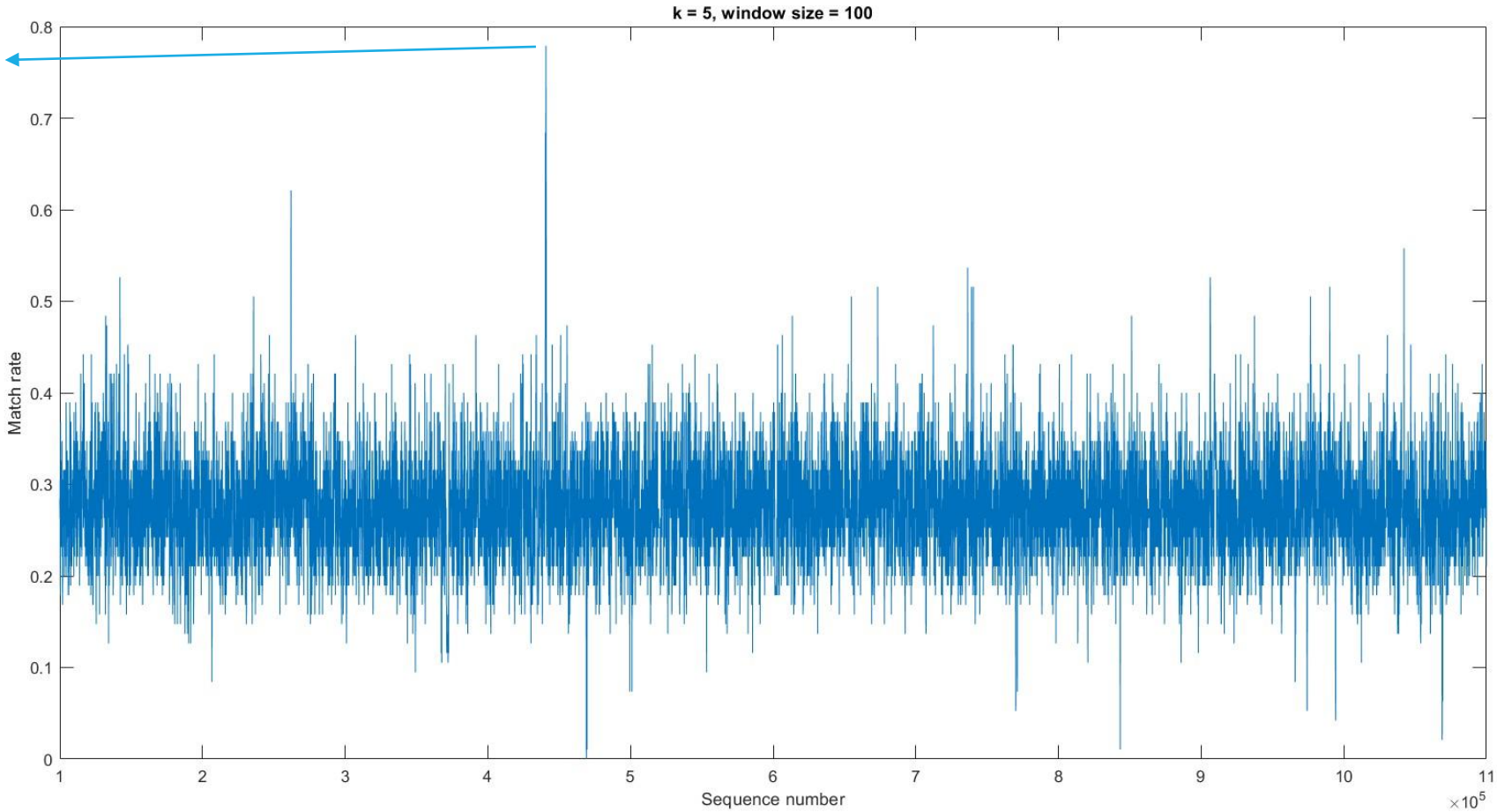
APPROACH 3 : AUTOCORRELATION

序列：
CCGCCGCCG
CCGCCGCCG
CCCCGGAGA
CCACCTCCTCC
TCCTCGTCGTC
G
Match rate
=0.808511



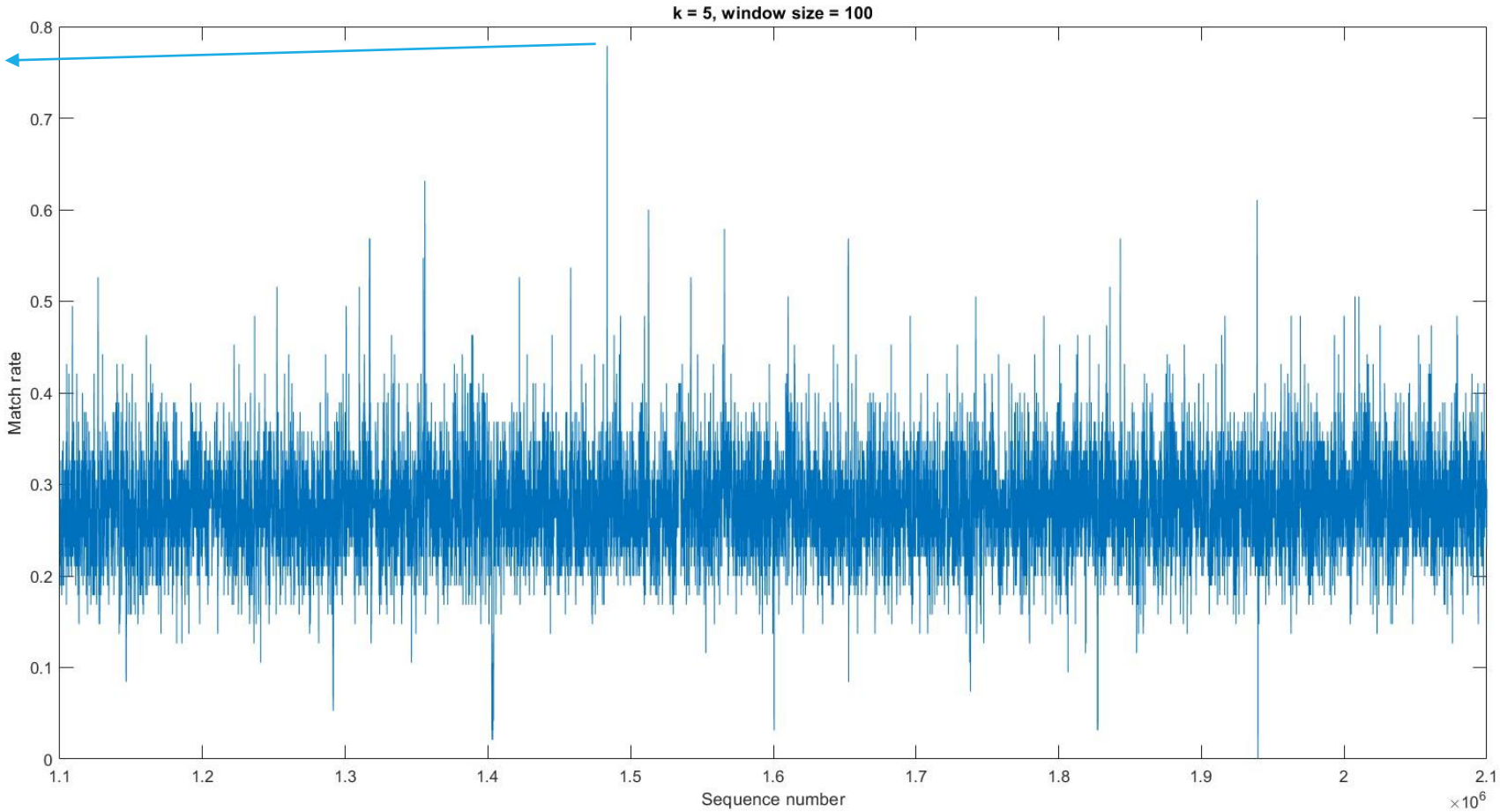
APPROACH 3 : AUTOCORRELATION

序列：
CCCACCCATC
CATCCCATCCA
TCCTATCCATC
CCATCCCATCC
CATCCCATCCC
ATCCCATCCCA
TCCCATCCCAT
CCCATCCCTTC
CTACCCTGTCT
CA
Match rate
=0.778947



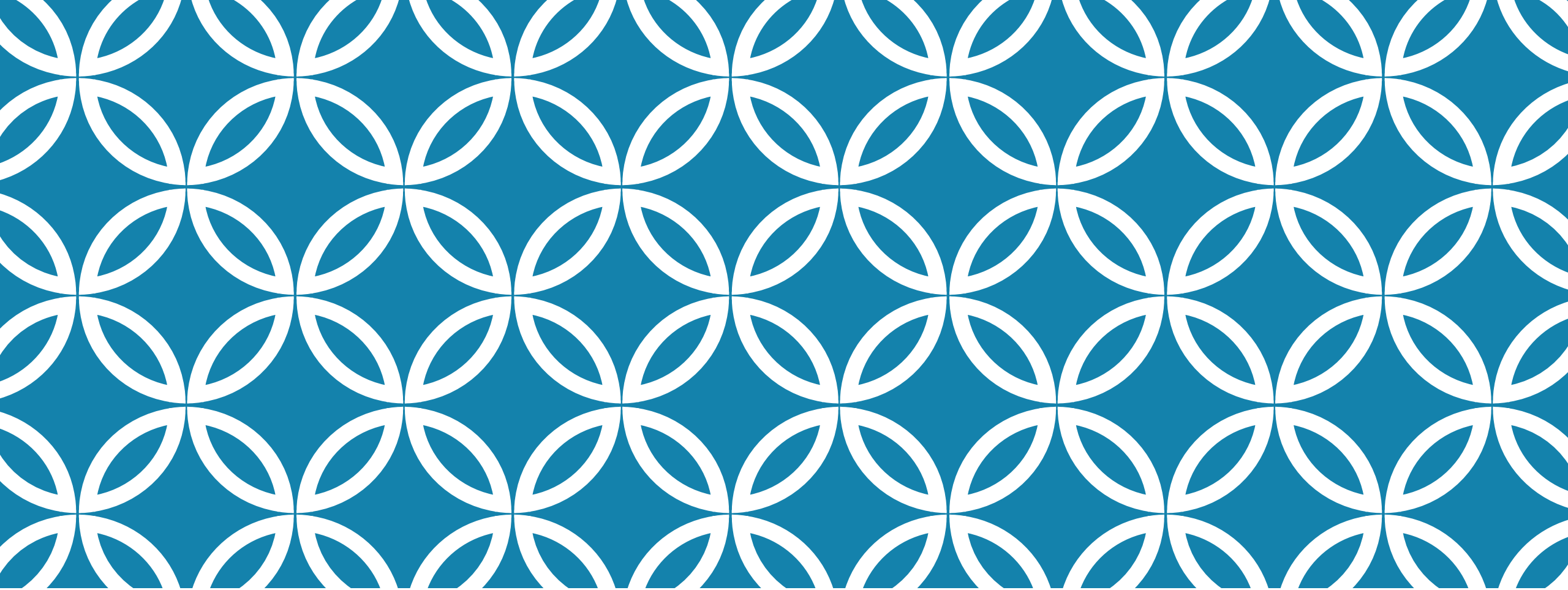
APPROACH 3 : AUTOCORRELATION

序列：
GAAAGAAAG
AAAGAAAGA
AAGAAAGAA
AGAAAGAAA
GAAAAAAGA
AAAGAAAAG
AGAAAAGAA
AAGAAAAGA
AAAGAAAAG
AAAAGAAAA
GAAAAGAAA
A
Match rate
=0.778947



結論

- 4種鹼基並非平均的分散在DNA序列中，反而時常出現有規則性、重複性高的短DNA片段。
- Autocorrelation window size愈高，match rate高的window數目愈少，但仍會出現match rate比預期中高的片段。
- Autocorrelation能夠有效率、並能容忍誤差的找出重複序列。
- 較長的ALU Element沒辦法透過這些方法找出。



THE END

報告者：李哲宇