**Übungsblatt 6**  
Sequenzmotive  
Deadline: 15.6.2022 um 20:00 MEZ  
Bioinformatik für Biochemiestudierende  
Dr. Florian Klimm  
Sommersemester 2022

**Aufgabe 6.1 Laplace Succession (10 %)**  
Die letzten sieben Vorlesungen haben stattgefunden. Nutzen Sie die Laplace Suc-  
cession um die Wahrscheinlichkeit, dass die nächste Worlesung auch stattfindet,  
zu bestimmen.

**Aufgabe 6.2 Sequenzmotive (20 %)**  
Sie haben die folgenden Motive gefunden

A picture containing text

Description automatically generated

Berechnen Sie das Profil dieser Motive. Berechnen Sie auch das Profil der Pseudo-  
counts und vergleichen Sie die Ergebnisse.

**COUNT**(Motifs)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| A | 3 | 2 | 0 | 1 | 2 |
| C | 0 | 1 | 0 | 2 | 1 |
| T | 0 | 1 | 4 | 0 | 0 |
| G | 1 | 0 | 0 | 1 | 1 |

PROFILE(Motifs)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| A | 3/4 | 1/2 | 0 | 1/4 | 1/2 |
| C | 0 | ¼ | 0 | 1/2 | 1/4 |
| T | 0 | ¼ | 1 | 0 | 0 |
| G | 1/4 | 0 | 0 | 1/4 | 1/4 |

**PSEUDOCOUNT**(Motifs)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| A | 4 | 3 | 1 | 2 | 3 |
| C | 1 | 2 | 1 | 3 | 2 |
| T | 1 | 2 | 5 | 1 | 1 |
| G | 2 | 1 | 1 | 2 | 2 |

PROFILE(Motifs)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| A | 1/2 | 3/8 | 1/8 | 1/4 | 3/8 |
| C | 1/8 | 1/4 | 1/8 | 3/8 | 1/4 |
| T | 1/8 | 1/4 | 5/8 | 1/8 | 1/8 |
| G | 1/4 | 1/8 | 1/8 | 1/4 | 1/4 |

Durch Einführen der Pseudocounts wird die Wahrscheinlichkeit seltener Events erhöht und Wahrscheinlichkeiten von Null werden eliminiert. Das ist vor allem in kleineren Datensets wie in diesem Beispiel wichtig, da die *sample size* nicht groß genug ist, um seltenere Vorkommnisse zu erfassen.

**Aufgabe 6.3 Profile (20 %)**  
Sie haben das folgende Profil gefunden

A picture containing table

Description automatically generated

Bestimmen Sie die Wahrscheinlichkeit der folgenden Strings  
1. ACTA

2. ACTG

🡪 die Wahrscheinlichkeit dieses Strings wäre nicht null, wenn das Profil auf Pseudocounts basieren würde

3. CCTAA (nicht sicher, ob ich das richtig verstanden habe…)

Theoretisch könnte man sich in der Profil-Tabelle eine sechste Spalte für ein fünftes Nukleotid im String dazudenken. Da wir keine Counts für diese Stelle haben, müssten wir die Wahrscheinlichkeiten für das Auftreten jedes Nukleotids an dieser fünften Stelle dann auf Null setzen.

Alternativ könnte das „fünfte“ Nukleotid auch vor der ersten Spalte Platz haben:

Würde wir anstatt von Counts mit Pseudocounts arbeiten, und würden somit alle Counts um 1 erhöhen, dann wäre auch die Wahrscheinlichkeit für das Auftreten von A als fünfte Base bzw. das Auftreten von C vor der ersten Stelle im Profil nicht mehr null sondern ¼. Dann würden wir für den String 3 eine Wahrscheinlichkeit erhalten, die nicht null ist. Da die Anzahl von Motiven in dieser Aufgabe nicht gegeben ist, können die Pseudocounts nicht ausgerechnet werden.

**Aufgabe 6.4 Sequence Logo (20 %)**  
Interpretieren Sie das folgende Sequence Logo.

Das abgebildete Sequenz-Logo enthält die Konsensus-Sequenz, an die das E. coli Ribosom während der Initiation der Translation rekrutiert wird. Die eigentliche mRNA-Sequenz ist hier in Form der korrespondierenden DNA-Sequenz ausgedrückt. Die Höhe der Buchstaben ist ein Maß für die „randomness“ – wenn der Buchstabe hoch ist, so kommt er in vielen Motiven an dieser Stelle vor und folglich ist die „randomness“ gering.

Die 27 Basen, die das Sequenz-Logo bilden, überspannen das Start Codon; alle Basen, welche mit einem negativen Vorzeichen gekennzeichnet sind, befinden sich *upstream* des Start Codons, solche mit positivem Vorzeichen *downstream*. In Prokaryoten wird die ribosome binding site (RBS) Shine-Dalgarno-Sequenz genannt und wird durch die Sequenz AGGAGG ausgemacht. Etwa 10 Basen vor dem Start Codon ist treten mehrere As und Gs mit geringer „randomness“ auf, sodass sich schließen lässt, dass es sich hier um die RBS halten kann. Das Sequenz-Logo zeigt außerdem, dass *E. coli* größtenteils das Start Codon ATG (AUG in mRNA) verwendet, aber dass auch GTG (GUG) und TTG (UUG) mögliche Start-Codons sind, die alternativ verwendet werden.

Chart, histogram

Description automatically generated

**Aufgabe 6.5 Programming (30 %)**

Schreiben Sie ein R Programm, das Münzwürfe simuliert und die folgenden Schritte  
ausführt  
• Input:  
– n Anzahl der Münzwürfe  
– p Wahrscheinlichkeit Kopf  
• Simuliere n Münzwürfe  
• Schätze aus diesen Simulierten Würfen die Wahrscheinlichkeit ˆp Kopf zu  
erhalten mit der Laplace Succession  
• Output ˆp

**Script (siehe R Dokument)**

**Wie gut ist die Schätzung ˆp und hängt dies von n und p ab?**

**Einfluss von n:**

Nach dem Gesetz der großen Zahlen wird sich bei größerem n die relative Häufigkeit des Zufallsergebnisses dem theoretischen Erwartungswert annähern. Sprich: bei einer geringen Anzahl an Münzwürfen weicht die relative Häufigkeit oft stärker vom Erwartungswert ab als bei einer großen Anzahl an Würfen. Dies kann man leicht erkennen, wenn man im R-Skript n variiert.

**Einfluss von p:**

Durch Variieren von p weg vom default Wert 0.5, was in der Praxis einer gezinkten Münze entspräche, ändern sich entsprechend auch die relativen Häufigkeiten der Ereignisse „heads“ und „tails“ und somit natürlich auch die Wahrscheinlichkeit .