Protokoll

Algorithmen der Bioinformatik I

Konrad Buchmann

23.12.2021

Programmiersprachen und Bibliotheken

Für die Lösung der Aufgaben wurde Python 3.9.9 verwendet. Außerdem die folgenden Bibliotheken:

- NumPy 1.21.5
- Matplotlib 3.5.1

Die Ausführung des Programms ist in der `README.md` beschrieben.

Ergebnisse

Anzahl der Startcodons

Mit der Regular Expression `(?=(ATG|GTG|TTG))` lässt sich die Anzahl der Startcodons im Datensatz finden. Insgesamt gibt es 7031 Startcodons, wobei auch überlappende Codons gezählt wurden.

Training der PWM

Für das schätzen der PWM wurden zunächst alle Sequenzen als Trainingsmenge verwendet.

Zuerst werden alle möglichen Startcodons gesucht. Nun werden alle potenziellen TIS, bestehend $L = 30$ Basen vor den Startcodons, betrachtet. Für diese werden die relativen Häufigkeiten der einzelnen Basen bestimmt, und in eine $4 \times L$ Matrix eingetragen. Die Zeilen der Matrix entsprechen je einer Base.

Mit der Gleichverteilung als Hintergrundverteilung kann man über die log-likelihood Methode die PWM bestimmen.

Die ermittelte PWM ist in Abb. 1 dargestellt. Subjektiv lässt sich bereits ein deutliches Muster erkennen, welches darauf hindeutet, dass das Schätzen erfolgreich war.

Schätzen des Detektionsschwellwerts

Zuerst wird allen Genstart Kandidaten mit der PWM ein score zugewiesen. Nun kann man die bekannt positiven Kandidaten, welche im gegebenen Datensatz immer am Index 100 auftreten, von den bekannt negativen trennen. Um eine

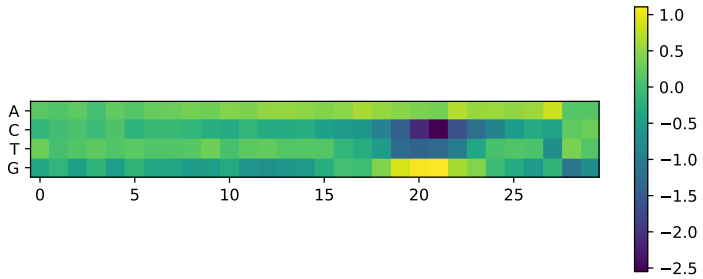

Figure 1: PWM

Sensitivität von p zu erhalten, kann man nun den Score-Threshold mit dem p-Quantil der Scores der echt positiven TIS schätzen. Der so ermittelte Threshold führt zumindest für die Trainingsdaten zu einer Sensitivität von genau p. Konkret wird zunächst $p = 0.5$ gewählt, womit sich ein Threshold von 3.038 ergibt.

Anzahl der falsch positiven Kandidaten

Wendet man den zuvor geschätzten Threshold an, erhält man 360 echt positive und 139 falsch positive Kandidaten.

Unterteilung in Trainings- und Validierungsdaten

Nun werden die Daten in eine 400 Sequenzen große Trainingsmenge und eine Validierungsmenge, bestehend aus den verbleibenden Sequenzen, unterteilt. Für die Trainingsdaten ergibt sich mit dem 50%-Kriterium ein Threshold von 2.93. Damit erhält man auf den Validierungsdaten 159 echt positive und 55 falsch positive Kandidaten.

ROC-Kurve

Wenn man durch mögliche Thresholds iteriert, in diesem Fall wurde naiv das Intervall $[-100, 100]$ gewählt, und anschließend die True Positive Rate gegen die False Positive Rate aufträgt, erhält man mit den Validierungsdaten eine ROC-Kurve. Diese ist in Abb. 2 dargestellt.

Der berechnete AUC-Wert beträgt 0.74. Dieser erscheint mir mit Blick auf die ROC-Kurve etwas unpassend. Eventuell ist die Reihenfolge der Datenpunkte in den Arrays `tprs` und `fprs` (`line 155`) nicht ganz linear, was dazu führen könnte, dass `numpy.trapz()` (`line 174`) einen falschen Wert berechnet.

Verschiebung des TIS-Fensters um eine Base nach rechts

Bezieht man die erste Base des Startcodons mit ein, verschlechtert sich die Detektionsgenauigkeit, und damit die AUC deutlich auf 0.63. Eine denkbare Erklärung wäre, dass die verschiedenen Varianten der Startcodons einen geringen Einfluss auf die Wahrscheinlichkeit eines Genstarts an dieser Stelle haben.

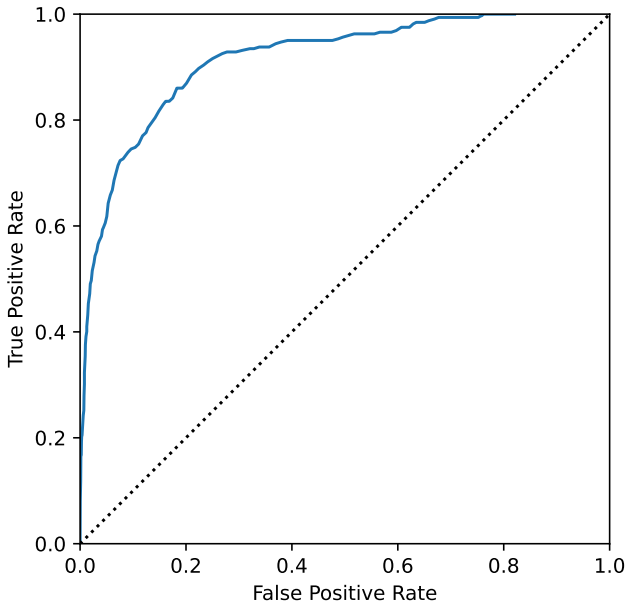

Figure 2: ROC-Curve

Da die Varianten aber mit unterschiedlichen Häufigkeiten auftreten, kommt es während des PWM-Trainings dazu, dass häufig auftretende Startcodons bevorzugt werden, auch wenn dies eigentlich kein Indikator für einen Genstart ist.

Variation der Pseudocounts

Variiert man die Pseudocounts r ergeben sich folgende AOC-Werte:

r	AOC
0.1	0.735
0.5	0.735
1.0	0.735
2.0	0.735
10.0	0.734

Offenbar haben verschiedene Pseudocounts keinen signifikanten Einfluss auf die Genauigkeit des Modells. In den betrachteten PWM-Fenstern gibt es scheinbar genug Variation, sodass niedrige Positionsfrequenzen nicht auftreten. Dies würde sich ändern wenn man die sehr ähnlichen Bereiche der Trainingsdaten, also konkret das Startcodon selbst, in das PWM-Fenster mit aufnimmt.