CHARITÉ / HUMBOLDT UNIVERSITÄT ZU BERLIN
INSTITUTE FOR THEORETICAL BIOLOGY

www.sys-bio.net/teaching
Prof. Hanspeter Herzel — h.herzel@biologie.hu-berlin.de
Prof. Nils Blüthgen — nils.bluethgen@charite.de
Dr. Manuela Benary — manuela.benary@biologie.hu-berlin.de
Dr. Christoph Schmal — christoph.schmal@charite.de

BIOINFORMATIK SS 2017 – ÜBUNGSBLATT 8

Gib deine Lösung bitte am 03.07.2017 in der Vorlesung ab. Alternativ kannst du die Lösung auch per E-Mail an `christoph.schmal@charite.de` schicken.

1. Poisson-Approximation

Für große Stichproben ($N \to \infty$) und kleine Wahrscheinlichkeiten ($p \ll 1$) lässt sich die Binomialverteilung

$$W_{\text{Bin}}(k) = \binom{N}{k} p^k (1-p)^{N-k}$$

durch eine Poisson-Verteilung

$$W_{\text{Poi}}(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

approximieren, wobei $\lambda = N\,p$ gelte. Berechne die Wahrscheinlichkeiten $W_{\text{Bin}}(k)$ und $W_{\text{Poi}}(k)$ für $N = 10$, $p = 0.2$ sowie $k = 1, 2, ..., N$. Stelle die Ergebnisse tabellarisch dar und zeichne die zugehörigen Wahrscheinlichkeitsverteilungen.

2. Positional weight matrix

Transkriptionsfaktorbindestellen für einen Transkriptionsfaktor sind normalerweise leicht variabel in der Sequenz. Mögliche Bindestellen werden *aligned*, was die Berechnung einer *Position Weight Matrix (PWM)* erlaubt. Mit *PWMs* kann unter anderem das Auftreten von neuen Bindestellen und deren Bindungsenergie für den Transkriptionsfaktor vorhergesagt werden. Hier ein Beispiel für ein Bindestellen-Alignment:

	alignment position						
site	1	2	3	4	5	6	7
1	T	T	C	T	T	C	T
2	C	T	A	T	A	A	C
3	G	C	G	G	A	G	T
4	G	T	G	A	A	T	C
5	G	T	G	G	A	C	T
6	G	C	G	T	G	C	T
7	C	T	G	G	A	G	T
8	G	T	G	T	A	A	T
9	G	A	C	C	A	A	T
10	G	G	C	A	A	A	T
11	T	T	G	A	T	C	A
12	G	T	G	A	A	T	A
13	G	C	A	T	T	G	T
14	T	A	G	A	T	G	T
15	A	G	G	C	A	T	A

Die Tabelle kann auch als Text-Datei heruntergeladen werden unter:
`https://itb.biologie.hu-berlin.de/~schmal/teaching/alignment_table.txt`

1. Berechne aus dem Alignment die *Position Count Matrix (PCM)*.
2. Berechne mit Hilfe der *PCM* die *Position Weight Matrix (PWM)* mit den Einträgen W_{ij} bei einem G+C-Gehalt von 40 %.

   Hinweis: *Vor* der Berechnung, addiere bitte eine Eins (*pseudo-count*) auf jede Position der *PCM* und aktualisiere N dementsprechend. Warum muss dies gemacht werden (Stichwort $\log_2 0$)?
3. Was bedeutet ein positiver Matrix-Eintrag W_{ij} für einen Buchstaben i an der Position j? Wie könnte demnach das Gewicht einer neuen Bindestelle biologisch interpretiert werden?
4. Wie könnte eine mögliche Konsensussequenz aussehen.
5. Berechne die Gewichte folgender mutmaßlicher Bindestellen:

   (a) `GTGGATT`

   (b) `AATGAGG`

   (c) `AGTGGAG`

   Welchen cut-off-Wert würdest du vorschlagen und warum? Welche der drei Sequenzen könnte(n) demzufolge als mögliche Bindestelle(n) gelten?
6. Schlage eine hypothetische Bindestelle mit einem hohen Score vor.
7. Erzeuge ein Sequenz-Logo für die *PWM*. Benutze dabei WebLogo:
   `http://weblogo.threeplusone.com/`