FACHHOCHSCHULE KIEL
UNIVERSITY OF APPLIED SCIENCES
Fachbereich: Informatik und Elektrotechnik *Institut für Kommunikationstechnik*

Laborübungen im Fach:
Bioinformatische Datenanalyse

Aufgabe Nr. 1: Dotplot-Visualisierung

Laborleiter : M. Eng. Christian Rohrandt

Gruppe: _____

Name: _____ Matr.-nr.: _____

Name: _____ Matr.-nr.: _____

Name: _____ Matr.-nr.: _____

Übungstag: Vortestat:

Abgabetag: Testat:

Rücksprache:

Allgemeine Informationen:

1. **Schlüssel Laborberichte:** EIN Laborbericht pro Team (nicht pro Person).
2. **Deadline:** Die Laborberichte sind *immer spätestens 2 Wochen nach* dem jeweiligen *Labortermin* fällig.
3. Format und Abgabe: Die Laborberichte können abgegeben werden entweder
    a. in ausgedruckter Form: Bitte einwerfen in die Metallbox am Ende des Ganges im 2. Stock in Gebäude C13, Grenzstr. 5, 24149 Kiel
    b. in pdf-Format: Bitte schicken Sie Ihre Berichte per email an: christian.rohrandt@fh-kiel.de
    c. andere Formate: NUR nach vorheriger Rücksprache mit dem Laborleiter.
4. **Inhalt der Laborberichte:** Jeder Laborbericht mus detaillierte Informationen zu jeder spezifischen Aufgabe enthalten, insbesondere:
    a. Den Python-Code aller Module inclusive der Kommentare.
        i. Stellen Sie sicher, dass Sie STETS Python-Code HINREICHEND kommentieren, *d.h. so, dass ein Außenstehender, der sich mit Python auskennt, Ihren Python-code verstehen kann, ohne vorher mit Ihnen Rücksprache zu halten.*
        ii. Stellen Sie sicher, dass Sie mit Ihren Code Kommentaren immer erklären, WARUM Sie so codiert haben, wie sie es tun, (nicht was Sie codiert haben.)

    NOTE: Hinreichend dokumentierter Python-Code (s.o.!) ist als Laborbericht ausreichend. Falls Ihr Python-Code keine oder nur wenige Kommentare enthält, ist eine zusätzliche textuelle Erläuterung, wie Sie Ihr Design angelegt und zu Ihrem Ergebnis gekommen sind, zwingend erforderlich.

Ziele dieses Labors:

Sie sollen:

1. den Umgang mit den gängigen Dateiformaten (FASTA, FASTQ und SAM/BAM) der Bioinformatik vertiefen.
2. Die grundsätzlichen Zusammenhänge von Sequenz-Alignment kennenlernen.
3. Eine Methode anwenden, die eine einfache Visualisierung von Ähnlichkeiten zweier Sequenzen ermöglicht.

1) Dotplot-Visualisierung

Gegeben sind zwei FASTA-Dateien (gi7305369.fasta & gi12643549.fasta). In diesen Dateien sind zwei verschiedene Protein-Sequenzen gespeichert, die jedoch Ähnlichkeiten aufweisen.

a) Lesen Sie beide Sequenzen ein und erzeugen Sie eine Dotplot-Visualisierung.
b) Erweitern Sie ihre Dotplot-Routine so, dass verschiedene Fensterlängen und Übereinstimmungszahlen verwendet werden können.
c) Stellen Sie die Dotplot-Visualisierung für die Fensterlängen 2, 3, 5 und 15 dar und vergleichen Sie die Ergebnisse miteinander.

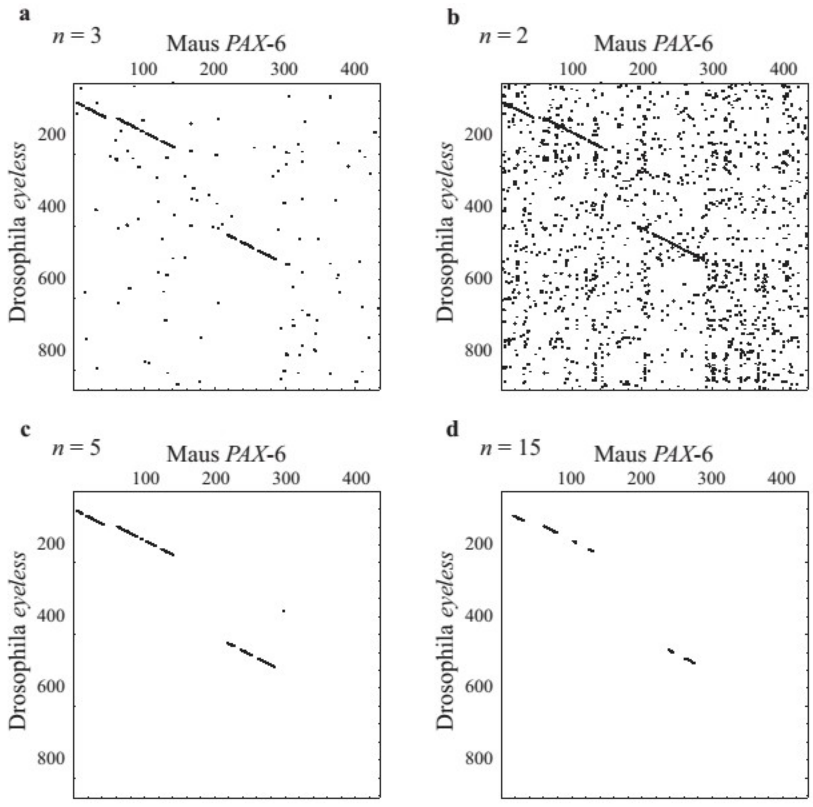

Abb. 3.7 Vergleich zweier realer Proteinsequenzen (*eyeless* von *D. melanogaster* und *PAX-6* von *M. musculus*) im Dotplot für verschiedene Werte der Fenstergröße *n*, nämlich **a**: $n = 3$, **b**: $n = 2$, **c**: $n = 5$ und **d**: $n = 15$. (In Anlehnung an: Lesk 2014)

Abbildung 1: [1]M.-T. Hütt and M. Dehnert, Methoden der Bioinformatik. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

Erläuterungen:

Fensterlänge: Als Fensterlänge wird verstanden, dass nicht nur einzelne exakte Übereinstimmungen als Dot im Dotplot eingetragen werden , sondern, dass zusammenhängende Worte der Fensterlänge eingetragen werden. Dadurch wird die Übersicht in dem Dotplot erhöht, die Genauigkeit jedoch verringert.

Übereinstimmungszahl: Die Übereinstimmungszahl kann als Erweiterung der Fensterlänge verwendet werden um die Genauigkeit etwas zu erhöhen. Dabei wird ein Fenster als Übereinstimmung eingezeichnet, wenn mindestens so viele einzelne Übereinstimmungen innerhalb der Fensterlänge vorkommen, wie die Übereinstimmungszahl angibt.