

# Aufgabe 3 - Dotplot

Silke Trißl

Wissensmanagement in der Bioinformatik



## Dotplot

### ■ Definition:

Ein *Dotplot* zweier Strings  $A$ ,  $B$  ist eine Matrix  $M$ :

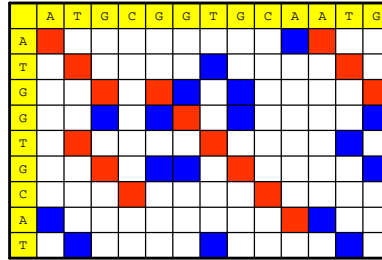
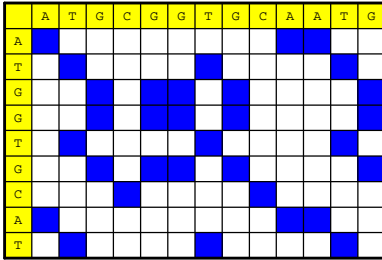
- Die Spalten entsprechen den Zeichen von  $A$
- Die Zeilen entsprechen den Zeichen von  $B$
- $M[a,b]=1$  gdw.  $A[a] = B[b]$

### ■ Beispiel

	A	T	G	C	G	G	T	G	C	A	A	T	G
A	1									1	1		
T		1					1						1
G			1		1	1		1					1
G					1	1		1					
T		1										1	
G			1		1	1		1					1
C				1					1				
A	1									1	1		
T		1					1						1

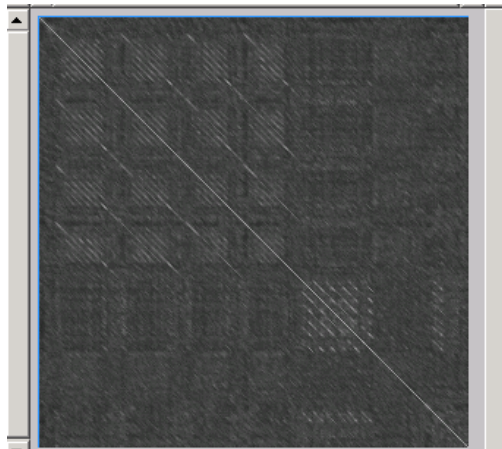
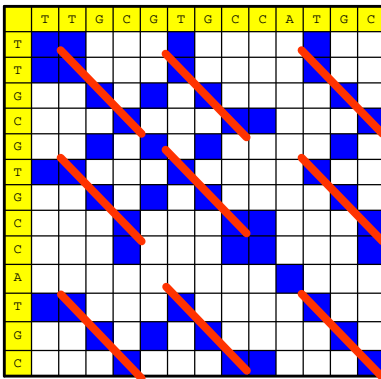
# Dotplot und gleiche Teilstrings

- Wie erkennt man gleiche Teilstrings im Dotplot?



- Diagonalen von links-oben nach rechts-unten
  - Größter gemeinsamer Teilstring – längste Diagonale
  - Visuell bei kurzen Strings möglich

# Repetitive Sequenzen



- Dotplot mit  $A=B$ 
  - Zitat (Genbank, P24014):  
[SIMILARITY] CONTAINS 7 EGF-LIKE DOMAINS.  
[SIMILARITY] Contains 24 leucine-rich (LRR) repeats.

# Abstandsmaße

---

- Approximatives Stringmatching sucht Ähnlichkeiten
  - Welcher Substring von T ist am ähnlichsten zu P ?
  - Welcher String  $T_1, \dots, T_n$  ist am ähnlichsten zu T ?
- Voraussetzung dafür
  - Was heißt ähnlich?
  - Was heißt „am ähnlichsten“?
- Quantifizierung des Abstandes zweier Strings
  - In der Regel eine sehr schwierige Aufgabe
  - Ähnlichkeit ist abhängig vom Gegenstand und Aufgabe
    - Wann sind sich Gesichter ähnlich - Haarfarbe zählt weniger als Augenfarbe ?
    - Wann sind sich Texte ähnlich – gleiche Wörter oder gleicher Inhalt?
  - Wir tun im folgenden so, als ob es einfach wäre
  - Und kommen auf das „Schwere“ später zurück

# Aufgabe 2

---

- Erstellt eine Anwendung, die
  - die Ähnlichkeit von von 2 Strings zeigt
- möglich:
  - Dotplot
  - Dotplot mit Sliding Windows

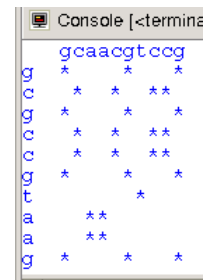
# Eingabe

- Einfach
  - 2 Strings innerhalb des Programms definieren
- Erweitert
  - die Strings aus einer Datei einlesen

# Ausgabe



- nur in Konsole
  - Ausgabe der beiden (kurzen und im Programm definierten) Strings
  - Zeichen, bei Gleichheit der Zeichen
- Als Java-Applett im Internet
  - <http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html>





# Beispielsequenzen

---

- Repeated Domain
    - SLIT\_DROME (gegen sich selbst)
  - Conserved Domains
    - AD08\_HUMAN gegen ADAM\_CROAD
  - Intron / Exon
    - AF404777 gegen AF404777mRNA
  - Reverse Complements
    - BSGTABX (gegen sich selbst)
  - Low Complexity Regions
    - SERA\_PLAFG (gegen sich selbst)
-