Algorithmische Bioinformatik

Multiple Sequence Alignment

Sum-of-pairs Score

Center-Star Score

Ulf Leser

Wissensmanagement in der Bioinformatik

Ziel dieser Vorlesung

- Aufgabenstellung „Multiples Sequenzalignment“ verstehen
- Einen elegante (aber nutzlosen) Algorithmus kennenlernen
- Dadurch das Problem der MSA-Zielfunktion verstehen

Inhalt dieser Vorlesung

- Multiples Sequenzalignment
- Sum-Of-Pair Zielfunktion
- Center-Star Zielfunktion

Definition

- Bisher: Alignment zweier Strings
- Jetzt: Alignment von k>2 Strings
- Definition
  Ein multiples Sequenzalignment (MSA) von k Strings S_i, $1 \leq i \leq k$, ist eine Tabelle mit k Zeilen und l Spalten, so dass
  - *In Zeile i steht String S_i, mit beliebig eingefügten Leerzeichen*
  - *Jedes Zeichen jedes S_i steht in exakt einer Spalte*
  - *In keiner Spalte stehen nur Leerzeichen*
- Bemerkungen
  - Es folgt, dass $l = |MSA| \leq \Sigma(|S_i|)$

Beispiel

```
S1: M---AIDE----NKQKALAAALGQIEKQFGKGSIMRLGEDR-SMDVETISTGSLSLDI
S2: MSDN--------KKQQALELALKQIEKQFGKGSIMKLGDG-ADHSIEAIPSGSIALDI
S3: M----AINTDTSGKQKALTMVLNQIERSFGKGAIMRLGDA-TRMRVETISTGALTLDL
S4: M-----------DRQKALEAAVSQIERAFGKGSIMKLGGKDQVVETEVVSTRILGLDV
S5: M------DE---NKKRALAAALGQIEKQFGKGAVMRMGDHE-RQAIPAISTGSLGLDI
S6: MD----------------------KIEKSFGKGSIMKMGEE-VVEQVEVIPTGSIALNA
S7: M---------------AL------IE--FGKG--M--G----------------L--
```

- Uns interessieren „möglichst gute“ MSAs
  - Möglichst wenig Spalten – wenig Leerzeichen
  - Möglichst homogene Spalten – hohe Übereinstimmung

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *
```

Motivation

- Alignment als Maß für Ähnlichkeitssuche sucht ähnliche Sequenzen in einer DB
  - Grund: Ähnliche Sequenz – ähnliche Struktur – ähnliche Funktion
- MSA sucht „das Ähnliche“ in vielen Sequenzen
  - Das, was viele Sequenzen ähnlich zueinander macht
  - Man startet mit vielen Sequenzen, bei denen man ähnliche Funktion / Struktur vermutet
  - MSA stellt fest, was das Gemeinsame dieser Sequenzen ist – Domänen, Motive, Signaturen, Profile, ...
  - These: Dieses Gemeinsame ist biologisch relevant

Konservierte Domänen

- Gedankengang
  - Gegeben: Proteine $S_1,\ldots,S_k$ mit ähnlicher Funktion
  - Annahme: Identischer evolutionärer Ursprung S
  - S unterliegt Evolution und generiert $S_1,\ldots,S_k$
  - Abschnitte in S_i, die trotz Evolution gleich blieben (konserviert sind), müssen für die gemeinsame Funktion wichtig sein
- Alignmengt – Multiples Alignment
  - Gemeinsamkeiten zwischen zwei Sequenzen oft per Zufall
  - Gemeinsamkeiten zwischen vielen Sequenzen eher kein Zufall

```
AAC__GTG_AT_T_GAC__
AACGAGTGC_TTTACACGT
```

```
AAC__GTG_AT_T_GAC__
AACGAGTGC_TTTACACGT
GCCG__TGC_TA__GTTG_
TTC_AGTGGACGTG__GTA
G____GTGCA__TGACC__
```

Blöcke, Domänen, Sites

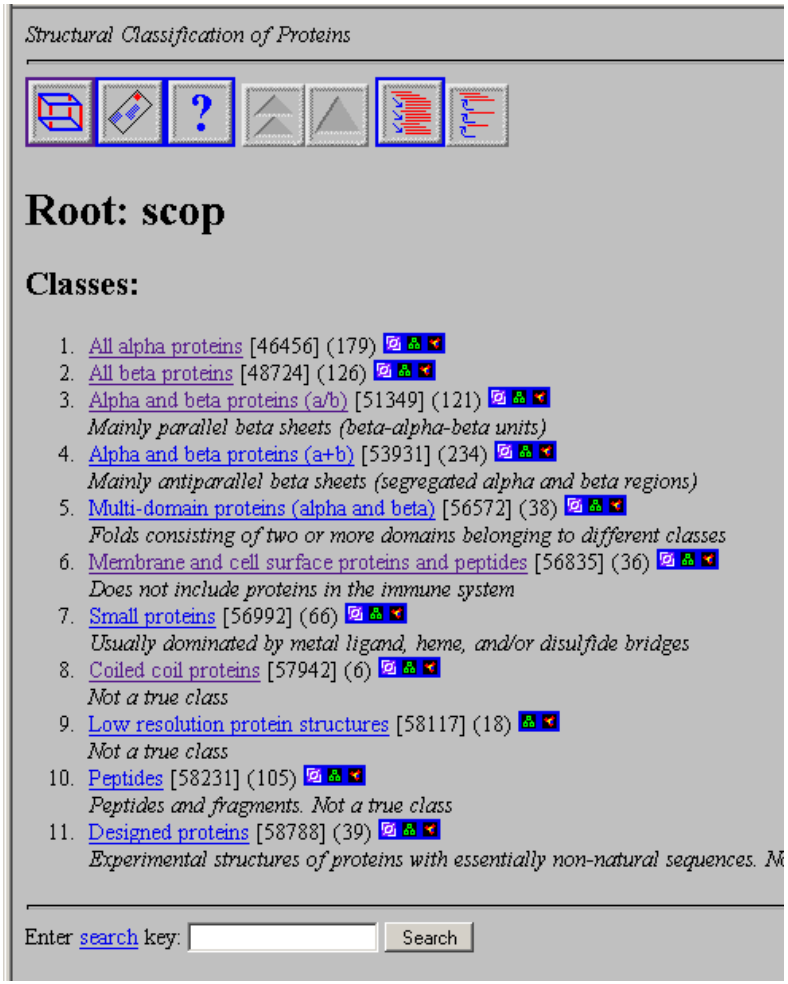
- Proteine
  - Bindungsstellen für andere Proteine / Liganden / Moleküle
  - Bindungsstellen an DNA
  - Signale zur Phosphorylierung / Dephosphorylierung
  - Signale zum Transport des Proteins
  - Signale zum Abbau des Proteins
  - ...
- DNA
  - Bindungsstellen für Transcriptionsfaktoren
  - Start- und Stoppcodons
  - Signal für differenzielles Splicen
  - ...

Proteinfamilien

- Proteine werden in Familien, Superfamilien, ... unterteilt
- Diverse Klassifikationen vorhanden (CATH, SCOP, ...)
- Idee: X00.000 Proteine zerfallen in X.000 Klassen ähnlicher Funktion?, Struktur?, Substruktur?
- Finden von Proteinfamilien
  - Starte mit Menge von Proteinen ähnlicher Funktion
  - Finde das Gemeinsame durch MSA
  - Suche nur mit konservierten Blöcken nach weiteren Vertretern
  - Modifiziere Familie entsprechend
  - Iteriere, bis Zufriedenheit eintritt

Beispiel: SCOP

- Structural Classification of Proteins
- Hierarchische Anordnung
  - *Fold: Major structural similarity*
    - All Alpha, All Beta, Membrane proteins, ...
  - *Superfamiliy: Probable common evolutionary origin*
    - Nucleotide-binding domain, Neurotransmitter-gated ion-channel transmembrane pore, ...
  - *Family: Clear evolutionarily relationship*
    - Globins, Death Domain, 4 families of Immunoglobulin ...
  - Protein
  - Spezies

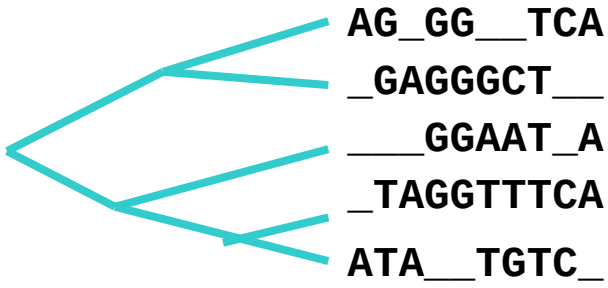


Inhalt dieser Vorlesung

- Multiples Sequenzalignment
- Sum-Of-Pair Zielfunktion
- Center-Star Zielfunktion

MSA Zielfunktion

- Zielfunktion beim einfachen Alignment war klar
  - Möglichst wenig evolutionäre Ereignisse – I, R, D
  - Eventuell mit Substitutionsmatrix / Gapmodellen
- Zielfunktion für MSA ist nicht klar
  - Score einer Spalte mit 2 T, zwei G und einem Leerzeichen?
  - Angabe einer Substitutionsmatrix für k Sequenzen über Alphabet Σ würde $O((|\Sigma|+1)^k)$ Werte erfordern
  - Motivation mit evolutionären Ereignissen funktioniert nicht mehr



MSA Überblick

- Es gibt diverse Vorschläge für Zielfunktionen

- Maximiere die Summe aller paarweisen Alignments
- Maximiere die Summe der Alignments jeder Sequenz zu einer Konsensussequenz (Center-Star)
- Maximiere die Summe der Alignments folgend dem phylogenetischen Baum der Sequenzen
  - Allerdings kennen wir die Sequenzen an der Wurzel und an den inneren Knoten nicht

Formal

- Definition
  - *Gegeben ein MSA M für Sequenzen $S_1,\ldots,S_k$. Das durch M induziertes Alignment für zwei Sequenzen S_i und S_j ist durch folgende Vorschrift definiert:*
    - *Entferne aus M alle Zeilen außer i und j*
    - *Entferne alle Spalten, die in i und j je ein Leerzeichen enthalten*
  - *Gegeben ein MSA M für Sequenzen $S_1,\ldots,S_k$. Der Sum-Of-Pairs Score für M (SP-Score) ist die Summe aller Alignmentscores der durch M induzierten paarweisen Alignments*
  - *Das SP-Alignment Problem für Sequenzen $S_1,\ldots,S_k$ sucht das MSA M mit minimalem SP-Score*

Beispiel

```
d/i = 1
r   = 1
m   = 0
```

```
AAGAA_A
AT_AATG
CTG_G_G
```

4
5
5
} 14

```
AAGAA_A
_ATAATG
C_TGG_G
```

4
5
6
} 15

- Die Berechnung des SP-Scores für ein gegebenes MSA über k Sequenzen ist also einfach
  - Komplexität?

Beispiel

```
d/i = 1
r   = 1
m   = 0
```

```
AAGAA_A
AT_AATG
CTG_G_G
```

4
5
5
} 14

```
AAGAA_A
_ATAATG
C_TGG_G
```

4
5
6
} 15

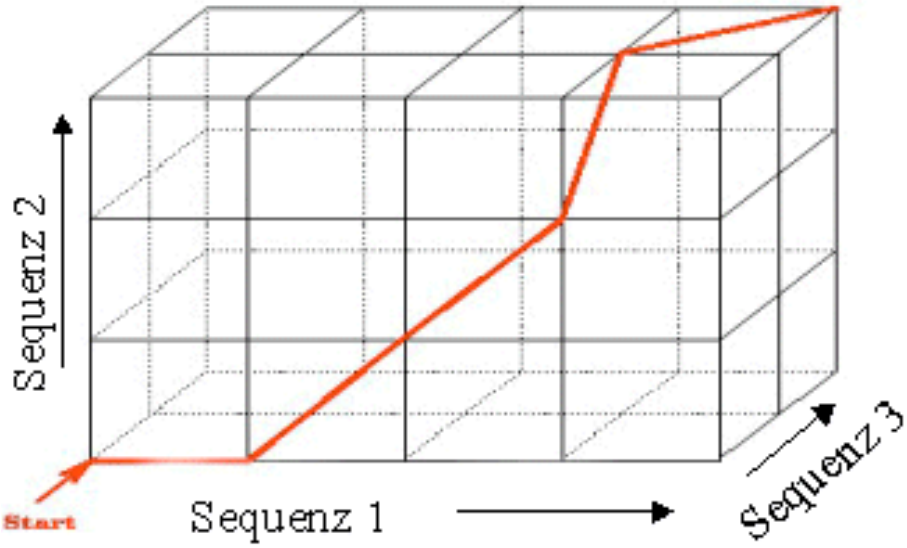
- Die Berechnung des SP-Scores für ein gegebenes MSA über k Sequenzen der Länge l ist also einfach
  - Komplexität $O(k^2 * l)$
- Aber wie findet man das MSA mit minimalem SP-Score?

DP in k Dimensionen

k=2: 2-dimensionale Matrix

		0	1	2	3	4	5	6	7
			w	r	i	t	e	r	s
0		0	1	2	3	4	5	6	7
1	v	1	1	2	3	4	5	6	7
2	i	2	2	2	2	3	4	5	7
3	n	3	3	3	3	3	4	5	6
4	t	4	4	4	4	3	4	5	6
5	n	5	5	5	5	4	4	5	6
6	e	6	6	6	6	5	4	5	6
7	r	7	7	6	7	6	5	4	5

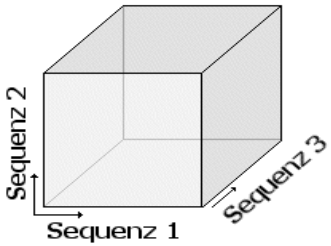
k=3: 3-dimensionale Matrix



- Was kostet ein Schritt?
- Im SoP sind die Kosten einfach zu finden – Paare vergleichen

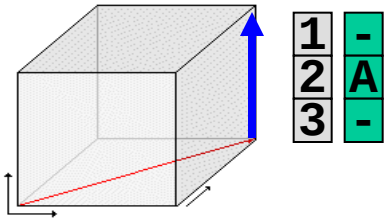
Erinnerung

- Grundidee der DP für zwei Sequenzen S_1, S_2
  - Berechnung des Alignment d(i,j) von $S_1[1..i]$ und $S_2[1..j]$ für steigende Werte (i, j) bis $i=|S_1|$ und $j=|S_2|$
  - Berechnung von d(i,j) aus d(i-1,j-1), d(i,j-1), d(i-1,j)
    - Man verlängert d(i-1,j-1) um Match oder Mismatch
    - … oder man verlängert d(i,j-1) um ein Insert
    - … oder man verlängert d(i-1,j) um eine Deletion
  - Statische Initialisierung der Werte d(i,0) und d(0,j)
- Wir betrachten im Folgenden nur den Fall k=3



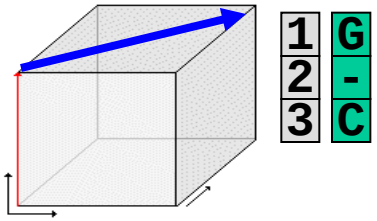
Analogie

d(i,j-1,k)

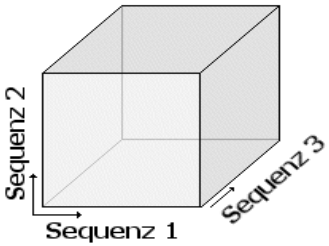


- SP-Alignment von d(i,j-1,k) ist bekannt
- Wir erweitern zu d(i,j,k)
- Dazu alignieren wir bei dieser Option $S_2[j]$ zweimal mit Leerzeichen (Inserts)

d(i-1,j,k-1)

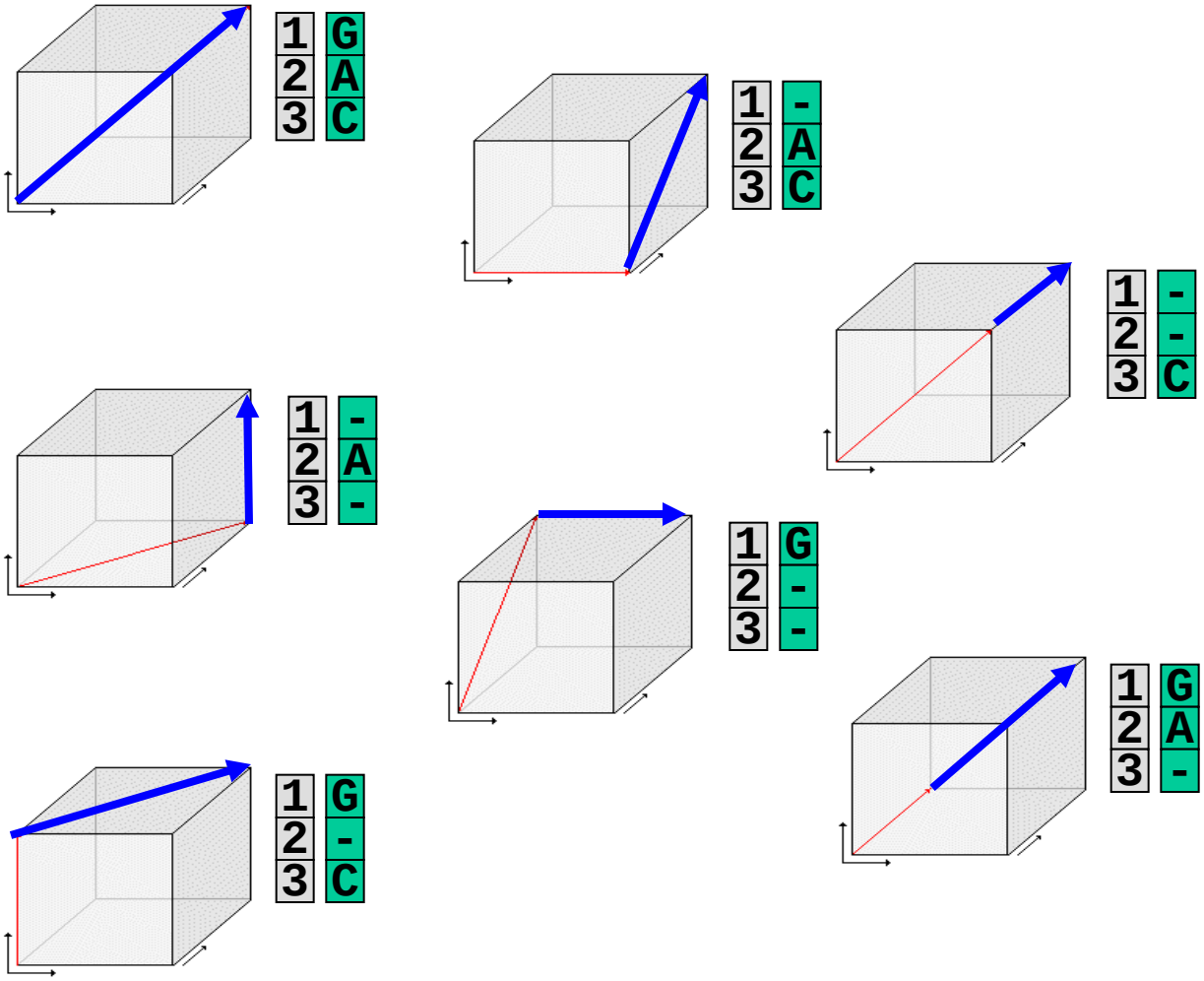


- SP-Alignment von d(i-1, j,k-1) ist bekannt
- Wir erweitern zu d(i,j,k)
- Dazu alignieren wir hier ein Leerzeichen mit $S_1[i-1]$ und mit $S_3[k-1]$



Mögliche Schritte zu d(i,j,k)

- d(i-1,j-1,k-1)
- d(i,j-1,k-1)
- d(i,j,k-1)
- d(i,j-1,k)
- d(i-1,j,k)
- d(i-1,j-1,k)
- d(i-1,j,k-1)



Formal

- Einfaches Kostenmodell (I/D/R=1, M=0)
- Theorem
  - *Gegeben Sequenzen S_1, S_2, S_3.*
    - *Sei d(i,j,k) der Score des SP-optimalen Alignments der Strings $S_1[1..i]$, $S_2[1..j]$, $S_3[1..k]$*
    - *Sei $c_{ij} = 0$, wenn $S_1[i] = S_2[j]$, sonst 1*
    - *Sei $c_{ik} = 0$, wenn $S_1[i] = S_3[k]$, sonst 1*
    - *Sei $c_{jk} = 0$, wenn $S_2[j] = S_3[k]$, sonst 1*
  - *Dann berechnet sich d(i,j,k) als:*

$$d(i, j, k) = \min \begin{cases} d(i-1, j-1, k-1) + c_{ij} + c_{ik} + c_{jk} \\ d(i-1, j-1, k) + c_{ij} + 2 \\ d(i-1, j, k-1) + c_{ik} + 2 \\ d(i, j-1, k-1) + c_{jk} + 2 \\ d(i-1, j, k) + 2 \\ d(i, j-1, k) + 2 \\ d(i, j, k-1) + 2 \end{cases}$$

Randbedingungen

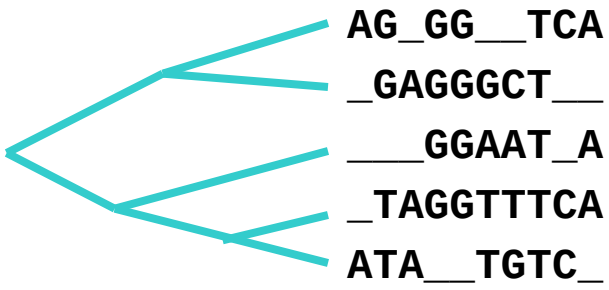
- Theorem Fortsetzung
  - *...*
  - *mit Initialisierung*
    - *Sei $D_{a,b}(i,j)$ der optimale Alignmentscore von $S_a[1..i]$ mit $S_b[1..j]$*
    - $D(0, 0, 0) = 0$
    - $D(i, j, 0) = D_{1,2}(i, j) + (i+j)$
    - $D(i, 0, k) = D_{1,3}(i, k) + (i+k)$
    - $D(0, j, k) = D_{2,3}(j, k) + (j+k)$
- Bemerkung
  - Alignment eines Leerzeichen mit einem Leerzeichen ist im induzierten paarweisen Alignment nicht enthalten

Komplexität

- Für drei Sequenzen der Länge n
  - Würfel hat n^3 Zellen
  - Für jede Zelle sind 7 Berechnungen notwendig
  - Zusammen $O(7*n^3)$
- Allgemeiner Fall: k Sequenzen der Länge n
  - Hyperwürfel hat n^k Zellen
  - Für jede Zelle sind 2^k-1 Vorgänger zu beachten
    - Alle Ecken eines k-dimensionalen Würfels minus eins (Das ist die Ecke die gerade berechnet wird)
  - Zusammen $O(2^k * n^k)$
    - Eigentlich: $O(2^k*n^k*k)$ wg Berechnung des Scores jeder Spalte
    - Häufigkeit aller Zeichen in Spalte in O(k) zählen und Score in $O(|\Sigma|^2)$ ausrechnen
- Tatsächlich: *Das SP-Alignment Problem ist NP-vollständig*

MSA unlösbar?

- SP-Score für mehr als eine Handvoll Sequenzen nennenswerter Länge nicht berechenbar
- Aber: SP berechnet gar nicht die minimale „Menge“ an Evolution



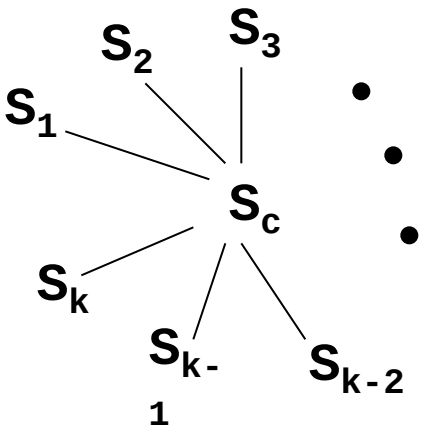
- Andere Zielfunktionen: Center-Star, MSA entlang des phylogenetischen Baums
- Praxis: Heuristiken (iterative, lokal-Greedy, ...)

Inhalt dieser Vorlesung

- Multiples Sequenzalignment
- Sum-Of-Pair Zielfunktion
- Center-Star Zielfunktion

MSA mit Konsensussequenz

- Minimiere Summe der Alignmentscores aller Sequenzen $S_1,\ldots,S_k$ mit einer Konsensussequenz S_c
  - S_c kann eine der S_i sein, muss aber nicht
  - Konstruktion von S_c z.B. durch Untereinanderschreiben der S_i ohne Gaps und Wahl des häufigsten Buchstaben (Consensus)
    - Funktioniert nur, wenn alle S_i ungefähr gleich lang – typisch bei MSA
  - MSA wird aus dem „Star“ abgeleitet
- Bei geeigneter Wahl von S_c gilt
  - Der SP Score des berechneten MSA ist höchstens doppelt so hoch wie der SP-optimale
  - Center-Star approximiert Sum-of-Pairs also bis auf Faktor 2



Center-Star Verfahren

- Gegeben k Sequenzen der Länge n
- Wähle als Konsensus S_c die Sequenz, die den kleinsten durchschnittlichen Abstand zu allen Sequenzen hat
- Beginne: $M=S_c$
- Iteriere
  - Wähle eine noch nicht alignierte Sequenz S beliebig
  - Aligniere M und S
    - Genaue Methode führen wir nicht aus (Gusfield, p. 347-)
    - Im Kern aligniert man S mit S_c und fügt dadurch entstehende neue Leerzeichen in S_c auch in M ein
  - Bis alle Sequenzen in M enthalten sind
- Progressiv: Sukzessives Hinzufügen von einzelnen Sequenzen zu einem wachsenden MSA

Beispiel

1. ATGGC
2. AGCC
3. TGCGAT
4. GCATG
5. TGCCTA
6. CAACTA

	S1	S2	S3	S4	S5	S6
S1	0	2	4	4	4	5
S2	2	0	4	4	3	4
S3	4	4	0	3	3	5
S4	4	4	3	0	3	4
S5	4	3	3	3	0	3
S6	5	4	5	4	3	0
Durchschnitt	3,8	3,4	3,8	3,6	3,2	4,2

Quelle: Martin Filip, Proseminar, 2005

Beispiel 2

- Kern des MSA: S_5= **TGCCTA**
- Wähle Sequenz: S_3= **TGCGAT**
  - Alignment:

```
TGCC_TA
TGCGAT_
```

- Wähle Sequenz: S_2= **AGCC**
  - Alignment:

```
TGCC_TA
TGCGAT_
AGCC___
```

- Wähle Sequenz: S_1= **ATGGC**
  - Alignment:

```
_TGCC_TA
_TGCGAT_
_AGCC___
ATGGC___
```

1. ATGGC
2. AGCC
3. TGCGAT
4. GCATG
5. TGCCTA
6. CAACTA

Beispiel 3

- Wähle Sequenz: S4= **GCATG**
  - Alignment:

```
_TGCC_TA
_TGCGAT_
_AGCC____
ATGGC____
___GC_ATG
```

- Wähle Sequenz: S6= **CAACTA**
  - Alignment:

```
_TGCC_TA
_TGCGAT_
_AGCC____
ATGGC____
___GC_ATG
CAAC__TA
```

Approximationsgüte

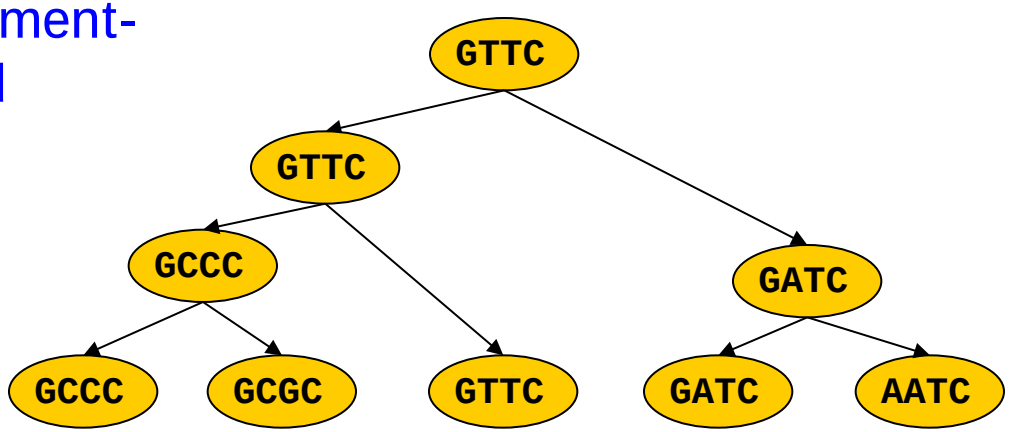
- Wenn für Zeichen i,j,k gilt: $d(i,j)+d(j,k) \geq d(i,k)$
  - Dreiecksungleichung
  - Gilt nicht für viele relevante Substitutionsmatrizen
- Theorem
  - *Gegeben k Sequenzen, $k \geq 3$. Sei d der SP Score des durch den Center-Star Algorithmus berechneten MSA. Sei d* der optimale Sum-of-Pairs Score. Dann gilt*

$$\frac{d}{d^*} \leq 2 - \frac{1}{k} < 2$$

- Beweis
  - Siehe Gusfield

MSA mit phylogenetischen Bäumen

- These: Sequenzen sind aus einer Ursequenz entstanden
- Natürliche Zielfunktion
  - Suche den Baum T so, dass die Summe aller Alignmentscores von benachbarten Sequenzen in T minimiert wird
  - Aus T kann man ein MSA ableiten (später)
- Leider
  - Wir kennen die inneren Sequenzen nicht
  - Das phylogenetische Alignment-problem ist MAX-SNP-hard
  - Siehe Maximum Parsimony



Suche mit MSA

- Erinnerung: Erzeugung von Proteinfamilien
  - Starte mit Proteinen gleicher/ähnlicher Funktion
  - Finde das Gemeinsame durch MSA
  - Suche mit dem MSA nach weiteren Vertretern
  - Modifiziere Familie entsprechend
  - Iteriere, bis Zufriedenheit eintritt
- Wie sucht man mit einem MSA?
  - Wir müssen entscheiden, wie gut eine (neue) Sequenz S zu einem gegebenen MSA M passt
  - Verschiedene Möglichkeiten: Profile, RegExp, Profile-HMM
  - Nächste Stunde

Selbsttest

- Welche biologische Frage versucht man mit MSA zu beantworten?
- Was ist eine Protein-Domäne? Beispiele? Was wäre eine Entsprechung bei Genen?
- Warum ist SP nicht das sinnvollste MSA-Verfahren?
- Wie könnte ein lokales MSA-Problem definiert sein? Wann könnte so ein Maß sinnvoll sein?
- Erklären Sie den Algorithmus, um das SP-optimale MSA für k Sequenzen zu finden