

Aufgabe 6 - Multiple Sequence Alignment

Silke Trißl

Wissensmanagement in der Bioinformatik



Multiples Sequenzalignment

- Bisher
 - Immer Vergleich zweier Strings
 - Alignment findet Teilstrings größtmöglicher Entsprechung
- Jetzt
 - Multiple Sequence Alignment (MSA)
 - Vergleich von $k > 2$ Strings

Definition

■ Definition

- Ein **multiple 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

- Das ist eine direkte Generalisierung des Alignment zweier Strings
- Es folgt, dass $l \leq \sum(|S_i|)$

Beispiel

```
S1: M---AIDE----NKQKALAAALGQIEKQFGKGSIMRLGEDR-SMDVETISTGSLSLDI
S2: MSDN-----KKQQALELALKQIEKQFGKGSIMKLGDG-ADHSIEAIPSGSIALDI
S3: M---AINTDTSGKQKALTMVLNQIERSFGKGAIMRLGDA-TRMRVETISTGALTLDL
S4: M-----DRQKALEAAVSQIERAFGKGSIMKLGKQVVEVETVSTRILGLDV
S5: M-----DE---NKKRALAAALGQIEKQFGKGAVMRMDHE-RQAIPAISTGSLGLDI
S6: MD-----KIEKSFGKGSIMKMGEVVEQVEVIPTGSIALNA
S7: M-----AL-----IE--FGKG--M--G-----L--
```

- Uns interessieren natürlich möglichst gute MSAs
 - Intuition
 - Möglichst wenig Spalten – wenig Leerzeichen
 - Möglichst „homogene“ Spalten – hohe Übereinstimmung
 - Exakte Score Functions
- MSAs erfassen das **Gemeinsame verschiedener Sequenzen**
- **Schwieriger zu fassen** als paarweise Alignments

Motivation

- Alignment sucht ähnliche Sequenzen
 - Da: ähnliche Sequenz – ähnliche Struktur – ähnliche Funktion
- MSA sucht „das Ähnliche“ in vielen Sequenzen
 - Argumentationsrichtung ist umgekehrt
 - Start mit Sequenzen, deren ähnliche Funktion/Struktur bekannt ist
 - MSA stellt fest, was das Gemeinsame dieser Sequenzen ist – Domänen, Motive, Signaturen, Profile, ...
 - Dieses **Gemeinsame ist offensichtlich biologisch relevant**
- Domänen oft nur schwach ausgeprägt
 - Schon geringe Ähnlichkeiten können zu gleicher Struktur führen
 - Alignment zweier Seqs wird keine signifikante Ähnlichkeit finden
 - MSA zeigt Bedeutung der (wenigen) Gemeinsamkeiten
 - Andererseits: schon kleine Änderungen können zu anderer Funktion führen (Mutationen – Erbkrankheiten)
- **Trennung des eventuell zufällig Gemeinsamen (Alignment) vom bedeutungsvoll Gemeinsamen (MSA)**

Konservierte Domänen

- Gedankengang
 - Gegeben: Proteine S_1, \dots, S_k mit ähnlicher Funktion
 - Z.B.: Können durch die Zellmembrane tunneln
 - Annahme: identischer evolutionärer Ursprung
 - Es gab einmal das „Mutterprotein“ S
 - S unterliegt Evolution
 - Mutation, Rekombination und Selektion
 - **Abschnitte in S_i , die trotz Evolution gleich blieben (konserviert sind), müssen wichtig sein**
 - Veränderung hat offensichtlich zum Funktionsverlust geführt
 - Andere Abschnitte dagegen sind nicht oder weniger wichtig
- Einige Annahmen ...
 - Gemeinsamer Ursprung
 - Man kennt nicht alle Nachfahren des Mutterproteins, sondern nur ein paar Gewinner
 - Sequenzen weiterer Organismen wichtig

MSA Zielfunktion

- **Zielfunktion** beim einfachen Alignment war klar
 - Möglichst wenig I,R,D
 - Eventuell mit Substitutionsmatrix und differenzierten Kosten
 - Eventuell mit spezieller Behandlung von Gaps
- Zielfunktion für MSA ist nicht so klar
 - Score einer Spalte mit 2 T, zwei G und einem Leerzeichen?
 - Angabe einer Substitutionsmatrix für k Sequenzen über Alphabet Σ würde $O(|\Sigma|^{k+1})$ Werte erfordern
 - Nicht machbar und biologisch nicht begründbar



MSA Überblick

- Weg über Substitutionsmatrizen nicht gangbar
- Verschiedene alternative Vorschläge für Zielfunktionen existieren
 - Maximiere die Summe aller paarweisen Alignments
 - Maximiere die Summe der Alignments jeder Sequenz zu einer Consensussequenz
 - Maximiere die Summe der Alignments folgend dem phylogenetischen Baum der Sequenzen

Beispiel

d/i = 1
r = 1
m = 0

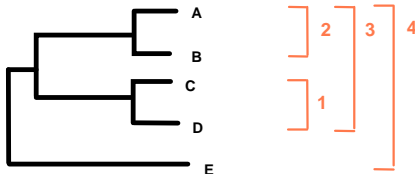
AAGAA_A } 4
AT_AATG } 5 } 14
CTG_G_G }

AAGAA_A } 4
_ATAATG } 5 } 7 } 16
C_TGG_G }

- Die Berechnung des SP-Scores für ein gegebenes M über k Sequenzen ist einfach
 - Komplexität $O(k^2 \cdot \max(|S_i|))$

ClustalW - Grobüberblick

A	1	-				
B	2	.17	-			
C	3	.59	.60	-		
D	4	.59	.59	.13	-	
E	5	.77	.77	.75	.75	-



1	PEEKSAVTALWGKVN--VDEVGG			
2	GEEKAAVLALWDKVN--EEEVGG			
3	PADKTNVKAAWGKVGAHAGEYGA			
4	AADKTNVKAAWSKVGGHAGEYGA			
5	EHEWQLVHLVWAKVEADVAGHGQ			

Berechne alle paarweisen Alignments



Konstruktion des Baums durch Clustern



Berechnung von MSAs entsprechend der Baumstruktur

ClustalW am EBI – www.ebi.ac.uk/clustalw

EMBL-EBI
European Bioinformatics Institute

SEQUENCE ANALYSIS

ClustalW Submission Form

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

[Download Software](#)

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
<input type="text"/>	<input type="text" value="Sequence"/>	<input type="text" value="interactive"/>	<input type="text" value="full"/>	<input type="text" value="single"/>
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="percent"/>	<input type="text" value="def"/>	<input type="text" value="def"/>
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
<input type="text" value="aln wnumbers"/>	<input type="text" value="aligned"/>	<input type="text" value="none"/>	<input type="text" value="off"/>	<input type="text" value="off"/>

Enter or Paste a set of Sequences in any supported format:

Upload a file:

Link zu Hilfe-Seiten

Sequenzeingabe
Copy & paste
File upload

Sequenzen – Fasta(Pearson) Format

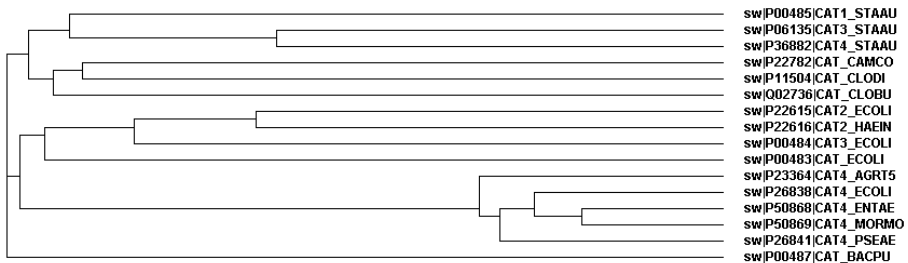
- Phenylalanine hydroxylase (PAH)
 - Das Enzym katalysiert die Reaktion von Phenylalanin zu Tyrosin
 - Sequenzen aus verschiedenen Organismen
- Death domain
 - Die death domain ist beteiligt an der Apoptose
 - Sequenzen aus verschiedenen Proteinen
- Chloramphenicol acetyltransferase
 - Das Enzyme ist verantwortlich für die Chloramphenicol (Antibiotikum) – Resistenz bei Bakterien
 - div. Sequenzen aus verschiedenen Organismen

Aufgabe

- Erstellt von allen drei gegebenen Sequenzsammlungen ein multiple sequence alignment
- Verändert die Reihenfolge der Sequenzen von PH4H.fasta
 - Was fällt auf?
- Unterteilt die Sequenzen von PH4H.fasta
 - Welche Unterteilung scheint aufgrund des Multiplen Alignments sinnvoll?

Clustering - graphisch

Cladogram



Cladogram von CAT

Nettes Tool – jalview

The screenshot displays the Jalview alignment editor interface. At the top, there is a menu bar with options: File, Edit, Font, View, Colour, Calculate, Align, Help. Below the menu is a sequence alignment of 17 protein sequences, each identified by an accession number and a protein name. The sequences are aligned over approximately 80 positions. A color-coded conservation plot is overlaid on the alignment, with a scale at the top ranging from 10 to 80. Below the alignment, a green bar chart represents the conservation plot for the sequence labeled 'Qual ity/ 1-229'. The plot shows varying levels of conservation across the sequence positions. At the bottom of the window, there is a status bar that reads 'done' and 'Java Applet Window'.

```
sw|P22615|CAT2_ECOLI/1-229-----MDFTKIDENFWRRRENFAFYRQQIKCQFSLTQKLDITAFRTAAETGVKIFVLMIVLISRAVWQIPDETMAKON-ELIV
sw|P22616|CAT2_HAEIM/1-229-----MDFTKIDENFWRRRENFAFYRQQIKCQFSLTQKLDITAFRTAAETGVKIFVPMIVLISRAVWQIPDETMAKON-ALIV
sw|P00484|CAT2_ECOLI/1-229-----MDVYTKIDENFWRRRENFAFYRHLFCQFSLTQKLDITLKKHDDSAVKFYPMIVLISRAVWQIPDETMAKOD-ELIV
sw|P00483|CAT2_ECOLI/1-229--HEKKITVYTTVVISQWRRKREHPEAFQSVQACTVQVQVLDITAFKTKKQKHKFYPAFIMHILARLWNAHEETMAKOD-ELAV
sw|P06135|CAT2_STAAN/1-229-----MTFMILYLENHWRRKVEPEYFM-QQTTVHKEIDDELFKMIGKGGVETVPSLIVAMEDVQVHKVYRFGISENKLGV
sw|P26882|CAT4_STAAN/1-229-----MTFMILYLENHWRRKVEPEYFM-QQTTVHKEIDDELFKMIGKGGVETVPSLIVAMEDVQVHKVYRFGISENKLGV
sw|P00485|CAT1_STAAN/1-229-----MDYENKIDLEHWRRKVEIPNMLN-QQTTVHKEIDDELFKMIGKGGVETVPAFIFLVKROISVTAIRGQVSDDELGV
sw|P22782|CAT2_CMICO/1-229-----HQYTKIDENWRRKVEYDRAVPEPCTVYRQKLDIS----KLEKQSEKLYPTLLVQNTIILHRELEPNAALDENQVGGV
sw|P11504|CAT2_CLOBU/1-229-----RVYFKIDENSRRKVEYDRAVPSQVCTVYRQVQVDS----QIKERQKLYPAALVYIMAVQVHSEPRNAALQDELGI
sw|IQ02736|CAT2_CLOBU/1-229-----MDYFMILDKHWRRKVEYDRAVPEPCTVYRQKLDIS----KLEKQSEKLYPTLLVQNTIILHRELEPNAALDENQVGGV
sw|P00487|CAT2_BACPU/1-229-----HFKQID-ENLVAEHPRAVMTLIRCSVSLVINDLIRGLHAILKEKGLKQVQVQVLLARAKQKIFESTVSDQVND-ELGV
sw|P50868|CAT4_ENTAE/1-229MTNYFESSFKGKLESEQVSNPNIKQGRYSYVSQVYRHSFDDCARYLLEDDEDDQK---LIIQSPCSIGSGAATIMAGQG-HEVD
sw|P50869|CAT4_NORMO/1-229MTNYFESSFKGKLESEQVSNPNIKQGRYSYVSQVYRHSFDDCARYLLEDDEDDQK---LIIQSPCSIGSGAATIMAGQG-HEVD
sw|P26838|CAT4_ECOLI/1-229MTNYFESSFKGKLESEQVSNPNIKQGRYSYVSQVYRHSFDDCARYLLEDDEDDQK---LIIQSPCSIGSGAATIMAGQG-HEVD
sw|P26841|CAT4_PSEAE/1-229MTNYFESSFKGKLESEQVSNPNIKQGRYSYVSQVYRHSFDDCARYLLEDDEDDQK---LVIQSPCSIGSGAATIMAGQG-HRAE
sw|P23264|CAT4_AGRYS/1-229MTNYFESSFKGKLESEQVSNPNIKQGRYSYVSQVYRHSFDDCARYLLED-ESADR---LVIQSPCSIGSGAATIMAGQG-HRME
```