

Übungen 1: paarweise Sequenzalignments, BLAST

**1. Dynamisches Alignment**

Führe ein globales Alignment nach Needleman & Wunsch durch. Der Wert für jedes Feld setzt sich aus dem größten Score der folgenden 3 Möglichkeiten zusammen:

- Match Score: Wert der Diagonalzelle links oben + Wert des Alignments (gleich +1; ungleich -1)
- Horizontal Gap Score: Wert der linken Zelle + gap score (-1)
- Vertical Gap Score: Wert der oberen Zelle + gap score (-1)

Der Pfeil für das Trace-Back zeigt in die Richtung, woher der beste Score kam. Falls dieser nicht eindeutig ist, ist die Diagonale zu bevorzugen bzw. es gibt mehrere Möglichkeiten.

		D	P	F	M	C	D	C	M	V	I
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
		←	←	←	←	←	←	←	←	←	←
<b>Q</b>	↑ -1	0+(-1)=-1 -1+(-1)=-2 -1+(-1)=-2	-1+(-1)=-2 -1+(-1)=-2 -2+(-1)=-3								
<b>F</b>	↑ -2	-1+(-1)=-2 -2+(-1)=-3 -1+(-1)=-2									
<b>M</b>	↑ -3										
<b>K</b>	↑ -4										
<b>D</b>	↑ -5										
<b>C</b>	↑ -6										
<b>M</b>	↑ -7										

Globales Alignment:

**2. Einfache Suche mit Blast**

Gehe zu <http://www.ncbi.nih.gov/blast> → Protein → Protein-protein BLAST (blastp).

Erklärungen zu den Feldern und Optionen von Blast gibt es im Anhang und unter <http://www.ncbi.nih.gov/blast/blastcgihelp.shtml> (Hyperlinks anklicken).

- (a) Um eine Sequenz einzusetzen, finde in **SRS** (<http://srs.ebi.ac.uk>) die Proteinsequenz mit der Accession-Nummer P00042. Kopiere die Sequenz im Fasta-Format und füge sie in das Feld *Search* ein. (Tipp: benutze bei SRS den Button *Apply Display Options* mit der passenden Option *FastaSeqs*.)
- (b) Wähle nun eine Datenbank aus. Hier **swissprot**.
- (c) Mit einem Klick auf **BLAST!** wird die Suche abgeschickt.
- (d) Als nächstes kommt ein Fenster, wo das Format der Ausgabe noch einmal geändert werden kann. Belasse alles bei default und klicke auf **Format!**.
- (e) Mache Dich mit der Ausgabe vertraut.

### 3. Erweiterte Suchen und Vergleiche

- (a) Gehe wieder zurück zur Such-Seite. Verändere Expect, Description, Alignments und Alignment views, um die jeweilige Funktion dieser Optionen genau zu verstehen.
- (b) Nun benötigen wir die Protein-Sequenz des Proteins MJ0577 aus dem Organismus *Methanococcus jannaschii* im FASTA-Format. Finde in swissprot die 10 homologsten Proteine und lasse Dir ihre Sequenzen anzeigen (*Get selected sequences*).
- (c) Führe eine Blast-Suche mit dem BlastP-Tool bei SRS (*Entry Options – Launch analysis tool*) durch mit denselben Parametern, wie sie bei NCBI als default verwendet werden, und ebenfalls gegen SwissProt. Vergleiche die Ergebnisse mit denen aus Aufgabe 2b. Lasse auch hier die Sequenzen anzeigen (Tipp: link).
- (d) Wie viele Treffer findest Du in der **nr**-Datenbank, wenn Du einen E-Value von 0.1 als oberste Schranke verwendest?

### 4. PSI-BLAST

- (a) Gehe (in einem neuen Tab oder Fenster) zu <http://www.ncbi.nih.gov/blast> → Protein → Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST). Benutze die Protein-Sequenz des Proteins MJ0577 aus dem Organismus *Methanococcus jannaschii* im FASTA-Format (aus der vorherigen Aufgabe). Wähle die nr-Datenbank und *Expect* 0.1.
- (b) Aktiviere die Option *Format for PSI-BLAST*, damit BLAST in mehrere Iterationen abläuft.
- (c) Verwende *with inclusion threshold* 0.001 für den E-value der Sequenzen, die zur Erstellung der PSSM verwendet werden.
- (d) Führe nun die erste Suche durch. Die Ergebnisse stammen aus der 1. Iteration, d.h. dieses Ergebnis sollte dem Ergebnis aus Aufgabe 2d entsprechen. Wo sind aber trotzdem Unterschiede?
- (e) Führe eine weitere Iteration durch. (Wenn Du *Layout – Two windows* belassen hast, mußt Du nach Klicken von *Run PSI-Blast iteration 2* im separaten Format-Fenster noch einmal auf *Format!* drücken, um das Ergebnis der zweiten Iteration angezeigt zu bekommen. *One Window* bewirkt, daß sich Format- und Ergebnisseiten abwechselnd im selben Fenster öffnen.)  
Jetzt wirst Du sehen, daß sich das Ergebnis verändert. Was ist zu beobachten? Was geschieht bei einer weiteren Iteration? (Tipp: Beachte vor allem die E-values.)

### Anhang: Erklärungen

- *Search*: Definiert den Bereich für die Sequenz, die im FASTA-Format eingefügt werden sollte.
- *Set subsequence*: Hier kann man den Bereich eingrenzen, wenn nicht die ganze Sequenz, sondern nur ein Teil davon betrachtet werden soll.
- *Choose database*: **nr** ist eine nicht-redundante Ansammlung aus vielen Datenbanken (GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF) und stellt die größte Proteinsequenz-Datenbank im BLAST-Modul dar. swissprot entspricht UniprotKB/Swiss-Prot.
- *Do CD-Search*: Die Suchsequenz wird gegen PSSMs der *Conserved domain Database* ausgerichtet, die auch Pfam enthält, um konservierte Domänen zu finden.

### Options for advanced blasting

- *Limit by entrez query or select from*: Mit dem Pulldown-Menu kann die Suche auf bestimmte Organismen(gruppen) eingeschränkt werden. Es können Beschränkungen für die Art der Datenbankeinträge erstellt werden. Mit NOT lassen sich unerwünschte explizit ausschließen.
- *Compositional adjustments – Composition-based statistics*: Dabei wird die Zusammensetzung der Query (Suchsequenz) und der Datenbank betrachtet, um genauere E-Values zu erhalten.
- *Choose filter*:
  - *Low-complexity*: Filter, um Regionen mit eingeschränkter Sequenzzusammensetzung in der Query zu maskieren, d.h. vom Alignment auszuschließen. Solche z. B. Prolin-reiche Regionen sind zwar statistisch, aber nicht biologisch signifikant. In Proteinen wird jede betroffene Aminosäure durch ein X ersetzt, in DNA die Nukleotide durch Ns.

- *Mask for lookup table only*: Hier werden die *low complexity*-Regionen nur von der Generierung der w-mers ausgeschlossen, bei den Alignments aber mit betrachtet.
- *Mask Lower Case*: Es gibt Programme (z.B. RepeatMasker), die bestimmte Bereiche von Sequenzen als „uninteressant“ maskieren, indem sie sie mit Kleinbuchstaben kennzeichnen. Diese können mit der Option *Mask Lower Case* explizit ausgeschlossen werden.
- *Expect*: Gibt die Schranke für die E-values der paarweisen Sequenzalignments an, die im Alignment ausgegeben werden. Mit dem default-Wert von 10 würde man erwarten, daß maximal 10 weitere Sequenzen aus der Datenbank denselben Score rein zufällig erreichen. Je geringer der E-Value, desto signifikanter ist das Ergebnis. Als sinnvoll für das Ausschließen unsignifikanter Treffer erweisen sich Werte von 0.02 und kleiner. Damit reduziert sich auch die Ausgabeliste des Alignments.
- *Word Size*: Mit 2 statt 3 würde eine größere Liste für die w-mers entstehen und entsprechend mehr Treffer in der Datenbank (HSPs) gefunden. Die Rechenzeit erhöht sich, bringt aber genauere Ergebnisse. Erfahrungsmäßig ist der default von 3 ein guter Kompromiß aus Genauigkeit und Rechenintensität.
- *Matrix*: Die Wahl der Aminosäure-Austauschmatrix ist für den Raw-Score und schließlich das gesamte Alignment von entscheidender Bedeutung. Generell gilt bei enger Verwandtschaft: niedrige PAM und hohe BLOSUM; entfernte Verwandtschaft: hohe PAM und niedrige BLOSUM. Die BLOSUM62 ist ein gutes Mittelmaß. *Gap costs*: Hier wird die Bestrafung zum Öffnen einer Lücke (*Existence*) und deren Erweiterung (*Extension*) festgelegt.
- *PSSM: Position Specific Score Matrix*: Statt Suchsequenz und Austauschmatrix kann auch eine positionsspezifische Scoringmatrix verwendet werden, wie sie von Psi-Blast erstellt wird. Damit wird das Alignment besser und spezifischer gewertet und es können auch entfernt verwandte Sequenzen gefunden werden.
- *Other advanced*: Hier kann man alle zuvor genannten Optionen als kurze Parameter formulieren und andere Werte als den vorgegebenen verwenden.
- *PHI pattern*: PHI-BLAST führte eine kombinierte Suche mit regulären Ausdrücken (ähnlich PROSITE-Motive) und einem lokalen Alignment um so erhaltene Treffer durch.

## Format

- *Show*: Hier kann man sich alternativ zum Alignment ab der 2. Iteration von Psi-Blast die PSSM anzeigen lassen, speichern und für eine neue Suche in einer anderen Datenbank in das obige Feld PSSM einfügen.
- *Masking Character/Color*: falls maskierte Residuen vorliegen, können sie im Alignment besonders hervorgehoben werden.
- *Descriptions*: Anzahl der angezeigten Sequenzen mit den zugehörigen Scores und E-values.
- *Alignments*: Anzahl der angezeigten Alignments.
- *Alignment views*: Es gibt verschiedene Arten der Alignmentdarstellung. *Pairwise* ist die übliche, dabei bezeichnet + ähnliche, der 1-letter-Code identische Aminosäuren. (Bei DNA-Sequenzen werden identische Nukleotide durch | bezeichnet.) Leerzeichen treten bei Gaps (-) und unpassenden Residuen auf. Bei *query-anchored with identities* werden identische Positionen mit . und abweichende mit 1-letter-Code bezeichnet.
- *Format for PSI-BLAST*: Aktiviert das iterative Alignment für das PSI-Blast
- *with inclusion threshold*: Gibt den Schwellwert für die E-values der alignierten Sequenzen wieder, die zur Erstellung der PSSM verwendet werden sollen. Je niedriger dieser Wert, desto weniger Sequenzen werden für die PSSM verwendet und desto spezifischer die Matrize. Dies ist aber nicht immer sinnvoll, weil gerade stark divergierte Sequenzen wichtige Informationen zur Evolution der Sequenz besitzen und schließlich in der Suche nach homologen Sequenzen hilfreich sein könnten. Auf der anderen Seite „korrumpieren“ unpassende Sequenzen die PSSM und führen zum Auffinden falsch positiver Treffer.