

Aufgabe 5 - BLAST

Silke Trißl

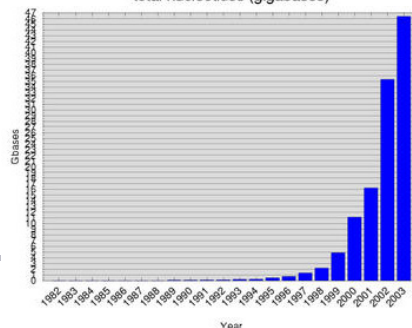
Wissensmanagement in der Bioinformatik



Heuristische Alignierung

- Annotation neuer Sequenzen basiert auf Suche nach homologen Sequenzen in Sequenzdatenbanken
- Datenmenge wächst exponentiell – selbst lineare Algorithmen sind zu langsam
- Gesucht sind Verfahren
 - Sublinearer Laufzeit
 - Möglichst geringer Verlust an Ergebnissen

EMBL Database Growth
total nucleotides (gigabases)



Suche in Datenbanken

■ Gedankenkette

- Gegeben Sequenz S und Datenbank mit Sequenzen S_1, \dots, S_n
- Gesucht: Welche Sequenzen in DB **sind homolog** zu S?
 - Kann nicht beantwortet werden
- Annäherungen
 - Welche Sequenzen in DB sind sehr ähnlich zu S?
 - Operationalisierung fehlt
 - Welche Sequenzen haben einen hohen Alignment-Score zu S?
 - Berechnung Alignmentsscore dauert zu lange
 - Welche Sequenzen haben **wahrscheinlich einen hohen Alignmentsscore** zu S?
 - BLAST

■ Wie viele Hits verlieren wir dabei?

Sensitivität und Spezifität

		Reality	
		+	-
Prediction	+	TruePositive (TP)	FalsePositive (FP)
	-	FalseNegative (FN)	TrueNegative (TN)

- Spezifität = $TP / (TP + FP)$ (Precision)
 - Wie viele der Treffen des Verfahrens sind wirklich welche?
- Sensitivität = $TP / (TP + FN)$ (Recall)
 - Wie viele der echten Treffer findet das Verfahren?
- Immer **eine Balance**
 - Ergebnismenge klein: SP=hoch, SE=klein
 - Ergebnismenge groß: SP=niedrig, SE=hoch

BLAST

- Altschul, Gish, Miller, Myers, Lipman: „Basic Local Alignment Search Tool“, K Mol Bio, 1990.
- **Heuristische Suche**
 - Sehr schnell, findet aber nicht alle optimalen Alignments
 - Berechnet statistische Signifikanz der Treffer
- ****Die** Erfolgsgeschichte der Bioinformatik**
 - Für Biologen tw. äquivalent zu „Bioinformatik“
 - Eingesetzt auf NCBI/EBI Server – von der ganzen Welt benutzt
 - Software frei erhältlich – auch lokale Installation möglich
- Weiter- / Parallelentwicklung von FASTA
 - FASTA kaum noch in Benutzung
- Diverse Weiterentwicklungen
 - Gapped-BLAST und PSI-BLAST (1997)

BLAST Parameter

- **Zunächst**
 - Suche in DNA Sequenzen
 - Keine Beachtung von Gaps
- **Gegeben**
 - Suchsequenz P, Datenbank $DB = \{S_1, \dots, S_n\}$
 - **Minimale Länge w** der „Seeds“
 - Substitutionsmatrix M
 - **Minimaler initialer Schwellwert t**
 - **Minimaler Gesamtschwellwert c**
 - Wird berechnet in Abhängigkeit von t, M, |DB|, |P|
 - Erwünschte Anzahl Treffer v

Prinzip

- Schritt 1
 - Bestimme alle Teilwörter P_1, \dots, P_m der Länge w in P
- Schritt 2
 - Suche nach Volltreffern von P_1, \dots, P_m in DB mit Score über t
 - Diese heißen „Hits“ oder „Seeds“
 - Manche P_i sind von vorneherein statistisch so wenig signifikant, dass sie von der Suche ausgeschlossen werden
- Schritt 3
 - Gegeben ein Hit zwischen DB-Sequenz S_j und Wort P_i
 - Verlängere Bereich um Hit P_i in P und in S_j
 - Alignment erst nach links, dann nach rechts wachsen lassen
 - Solange, bis
 - Sequenz P oder S_j ist zu Ende
 - Alignmentwert fällt unter Threshold c
 - Alignmentwert fällt „signifikant“ unter bisherige beste Treffer
 - „Signifikant“ heuristisch bestimmt, abhängig von c und v
 - Ergibt „Maximal Segment Pairs“ – die besten v sind das Ergebnis

Beispiel

$W=5, t=5$

$P=ACGTGATA$

$S=GATTGACGTGACTGCAAGTGATACTATAT$

Schritt 1
Teilwörter



$P_1=ACGTG$

$P_2=CGTGA$

$P_3=GTGAT$

$P_4=TGATA$

GATTG**ACGTG**ACTGCAAGTGATACTATAT
 GATTG**ACGTG**ACTGCAAGTGATACTATAT
 GATTGACGT**GATA**CTGCAAGTGATACTATAT

Schritt 2
Hitsuche

Schritt 3
Verlängerung

GATTG**ACGTG**ACTGCAAGTGATACTATAT

ACGTGATA

5

ACGTGATA

5+1=6

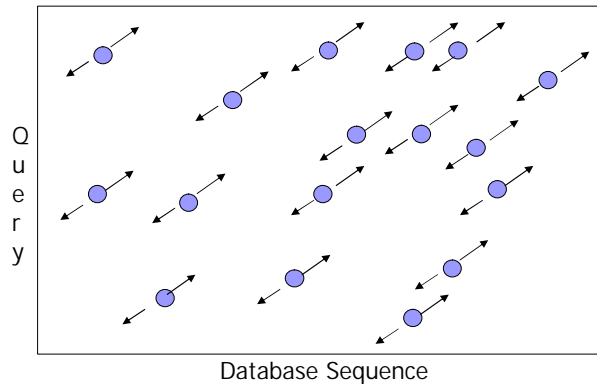
ACGTGATA

6-3=3

...

...

Veranschaulichung



- Original BLAST – kein Zusammenfügen mehrerer Maximal Segment Pairs zu einem (kompletten) Alignment
- Gapped - Blast

Bemerkungen

- BLAST ist eine **Exklusionsmethode**
 - Bestimme und suche nach Seeds = minimale Alignments, die im Kern jedes optimalen Alignments stecken müssen
 - Erweitere (nur) diese zu vollen Alignments
 - Weitere Algorithmen nach diesem Muster bekannt: Myer 's sublineares Alignment, Baeza-Yates&Perleberger
- Heuristik
 - BLAST **kann nicht alle optimalen Alignments** finden
 - Keine Inserts/Gaps, Schwellwerte w und t
 - Kleinere w/t Werte – höhere Sensitivität, aber auch geringere Performance
 - Parameterwerte und Algorithmen sind **über 10 Jahre getuned** worden

Defaulteinstellungen

- Typische Einstellungen, oft übernommen
- Proteine
 - BLOSUM62 Matrix
 - $W=4$, $t=14$
- DNA
 - Einheitsmatrix (Match +1, Mismatch -3)
 - $w=12$, $t=12$

BLAST 2

- Altschul, Madden, Schaffer, Zhang, Zhang, Miller, Lipman: „Gapped BLAST and PSI-BLAST: a new generation of protein database search programs“, NAR, 1997
- Drei Verbesserungen
 - **Performance verbessern** durch andere Extensionsstrategie
 - Sequenzdatenbanken wachsen schneller als CPU Geschwindigkeit
 - **Extensionen fressen >90% der gesamten Laufzeit**
 - Folge: Mehr verlangen als ein Wort, bevor man die Extension beginnt
 - BLAST2: Extension erfolgt nur, wenn **zwei nicht-überlappende Hits auf einer Diagonale** mit Abstand höchstens a gefunden wurden
 - **Gaps beachten** durch Verknüpfung von MSPs mit dyn. Prog.
 - Denn: Mehrere kurze Alignments mit Gaps werden vom alten BLAST übersehen, wenn keines der Kurzen signifikant ist
 - Durch Verlangen von zwei Hits kann w reduziert werden
 - **Motivsuche verbessern** durch Iteration mit Anpassung der Substitutionsmatrix (PSI-BLAST)
 - Denn: Während der Suche kristallisiert sich das konservierte Motiv oft erst heraus

NCBI BLAST

PubMed Entrez BLAST OMIM Taxonomy Structure

NEW 10 February 2004 BLAST 2.2.8 has been released. [Read more...](#)

Nucleotide

- Discontiguous megablast
- Megablast
- Nucleotide-nucleotide BLAST (blastn)**
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)**
- PPI- and PPI-BLAST**
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Search by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)**
- Protein query vs. translated database (tblastx)**
- Translated query vs. translated database (tblastx)**

Genomes

- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Special

- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobulin BLAST (IgbLst)

Meta

- Retrieve results by RID
- Get this page with javascript-free links

[Disclaimer](#)
[Privacy statement](#)
[Accessibility](#)
Valid [XHTML 1.0](#), [CSS](#)

BLAST Varianten

- BLAST gibt es in verschiedenen Varianten
 - **Blastn** : DNA-Anfrage/DNA-Datensammlung
 - DNA -> DNA
 - **Blastp** : Protein-Anfrage/Protein-Datensammlung
 - Protein -> Protein
 - **Blastx** : translatierte DNA-Anfrage/Protein-Datensammlung
 - DNA -> Protein
 - **Tblastn** : Protein-Anfrage/translatierte DNA-Daten
 - Protein -> Protein
 - **Tblastx**: translatierte DNA-Anfrage/translatierte DNA-Daten
 - DNA -> Protein
- Bei DNA/Proteinsuche immer Übersetzung in **alle sechs Reading Frames**

blastp - Standardeingabe

NCBI
Nucleotide Protein

protein-protein **BLAST**

- Sequenzen in
 - Fasta-Format >Sequenzname SEQUENZ
 - nur Sequenz
 - Accession P24021

Search

Set subsequence From: To:

Choose database nr

Do CD-Search

Now: **BLAST!** or Reset query Reset all

Sequenzen

- Nucleotide – Nucleotide: blastn
 - AQ008298.fasta
- Protein – Protein: blastp
 - Q7RNI7.fasta
- Translated Nucleotide – Protein: blastx
 - BI670787.fasta
- Protein – translated NucleotideDB: tblastn
 - Q9DGL7.fasta
- Aufgabe:
 - Finden Sie heraus, um welche Proteine / Gene es sich handeln könnte

Verfügbare Datenbanken

- Proteine

- nr
 - all non-redundant GenBank CDS translations+RefSeq Proteins+SwissProt+PIR+PRF
- swissprot
 - major release of SwissProt (manually annotated Protein-DB)
- pat
 - Proteins from patients
- pdb
 - Sequences derived from the 3-dimensional structure DB (PDB)
- month
 - same as nr, but just sequences from the last 30 days

Verfügbare Datenbanken

- DNA

- nr
 - All GenBank+RefSeq Nucleotides+EMBL+DDBJ+PDB sequences
- EST
 - Sequences from EST division of GenBank, EMBL, DDBJ
- gss
 - Genome Survey Sequence
- Chromosome
 - completely sequenced chromosomes and genomes
- month
 - same as nr, but just sequences from the last 30 days

■ ... see blast Dokumentation

blastp - Optionen

Options for advanced blasting

[Limit by entrez query](#) or select from:

[Composition-based statistics](#)

[Choose filter](#) Low complexity Mask

[Expect](#) **Expected E-value
E-values < Wert
werden ausgegeben**

[Word Size](#) **Word size**

[Matrix](#) Gap Costs


[PSSM](#)

[Other advanced](#)

[PHI pattern](#) **Erweiterte Optionen
-W wordsize[Integer]**

Silke Trißl: Bioinformatik für Biophysiker, WS 2006/07

blastp - Warteschlange

 **formatting BLAST**

Nucleotide Protein Translations Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = uniprot|Q7RJW0|Q7RJW0 Similar to S. cerevisiae BDF1. (260 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is

or

The results are estimated to be ready in 21 seconds but may be done sooner.

Zeigt Motive von Pfam Smart ...

Silke Trißl: Bioinformatik für Biophysiker, WS 2006/07

blastp - Output



results of **BLAST**

BLASTP 2.2.8 [Jan-05-2004]

Reference:

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res.* 25:3389-3402.

RID: 1076489669-10464-134267037774.BLASTQ3

Query= uniprot|Q7RJW0|Q7RJW0 Similar to S. cerevisiae BDF1.
(260 letters)

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF
1,642,191 sequences; 538,623,868 total letters

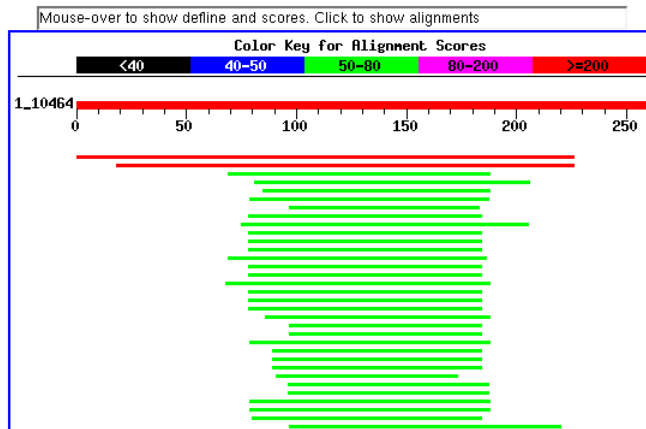
If you have any problems or questions with the results of this search
please refer to the [BLAST FAQs](#)

[Taxonomy reports](#)

Ergebnisse
aufgeschlüsselt nach
Organismen / Species

blastp - Output - 2

Distribution of 101 Blast Hits on the Query Sequence



blastp – Output – 3

Sequences producing significant alignments:

		Score (bits)	E Value
gi 23491055 gb FAA22686.1	similar to <i>S. cerevisiae</i> BDF1 [P...	395	e-109
gi 23508131 ref NP_700091.1	Hypothetical protein (Bromodi...	2	5e-75
gi 19114532 ref NP_055114.1	bromodomain-containing protein ...	69	3e-13
gi 40744729 gb EAA66622.1	...	75	1e-12
gi 7493489 pir IT400000	...	73	4e-12
gi 33146482 dbj BAC11111	...	73	5e-12
gi 31204491 ref XP_005555555.1	...	72	8e-12
gi 27804346 gb AA022222	Hom...	72	9e-12
gi 19173489 ref NP_055114.1	R [...]	71	2e-11
gi 33879612 gb AAH33333	...	69	8e-11
gi 37674287 ref NP_055114.1	...	69	9e-11
gi 3184498 gb AAC27777	...	69	9e-11
gi 7657218 ref NP_055114.1	bromodomain-containing protein ...	69	1e-10
gi 34862361 ref XP_343176.1	similar to bromodomain-contain...	69	1e-10
gi 15242727 ref NP_201138.1	DNA-binding protein family (Ar...	69	1e-10
gi 38014413 gb AAH60452.1	LOC398944 protein [<i>Xenopus laevis</i>]	68	1e-10
gi 31206311 ref XP_312107.1	ENSANGP00000016848 [<i>Anopheles</i> ...	68	1e-10
gi 19931486 gb AA02191.1	cell proliferation related protei...	68	1e-10
gi 31560788 ref NP_065254.2	bromodomain containing 4 isofo...	68	1e-10
gi 11358472 pir IT48600	kinase-like protein - <i>Arabidopsis t...</i>	68	1e-10
gi 19718731 ref NP_490597.1	bromodomain-containing protein...	68	2e-10
gi 18417335 ref NP_568297.1	DNA-binding protein family (Ar...	68	2e-10

E-Value
je kleiner, desto besser;
hängt von
Querylänge
Datenbankgröße
Alignmentscore

blastp – Output – 4

[gi|19114532|ref|NP_593620.1](#) protein with 2 bromodomains, putative involvement with sporulation [Schizosaccharomyces pombe]
[gi|34098718|sp|Q9HGR4|YK82_SCHP0](#) Hypothetical bromodomain protein C631.02
[gi|19955810|emb|CAC05484.1](#) SPAC631.02 [Schizosaccharomyces pombe]
 Length = 727

Score = 77.0 bits (188), Expect = 3e-13
 Identities = 44/121 (36%), Positives = 65/121 (53%), Gaps = 1/121 (0%)

Query: 70 PKNGDVKIKNEI-SLNNDPFNKQWYSLANQIIQSLSKYEGGHIFERLVDTKKQNCPPDYDV 128
 P + V K EN + P K + + + + L + F VD KQN PDY +
 Sbjct: 212 PHDETIVKKEENDKQYPPMIKEQHKYIHAMLRQLRRGRDSIPFRAPVDPVKQNPDPYPTI 271

Query: 129 IKNPMSFSCIKAKLKKGGYSSPQEFINDVQLVFYNCVSYNTPGTIVAITGKNIEAYFNQ 188
 IKNP+ ++ K G YSS Q FI+D+ L+F NC +YN + V + GKN+ +A F Q
 Sbjct: 272 IKNPIDLGTMQKFFSSGVYSSAQHFIDDMMLMFSNCFLYNGTESPVGVMGKNLQATPERQ 331

Query: 189 L 189
 L
 Sbjct: 332 L 332

Score = 59.3 bits (142), Expect = 7e-08
 Identities = 29/75 (38%), Positives = 45/75 (60%)

Query: 111 FERLVDTKKQNCPPDYDVVIKNPMSFSCIKAKLKKGGYSSPQEFINDVQLVFYNCVSYNTP 170
 F + V + CPDY+ VIK+PM ++ KL +Y+S + F D+ L+F NC +N+
 Sbjct: 416 FYKPVNPTACGCPDYFVKVIKHPMDLGTMQKLNHNEYASMKAFEDMVLMPKNCYKFNFA 475

Query: 171 GTIVAITGKNIEAYF 185
 GT V + GK +E+ F
 Sbjct: 476 GTPVHLMGKKLESIF 490



Aufgabe 5

- Verwendet BLAST, um für die gegebenen Sequenzen herauszufinden, um welche Proteine bzw. Gene es sich handeln könnte.

- Variiert dabei die Parameter