

WS2018/2019

„Genomforschung und Sequenzanalyse - Einführung in Methoden der Bioinformatik-“

Thomas Hankeln



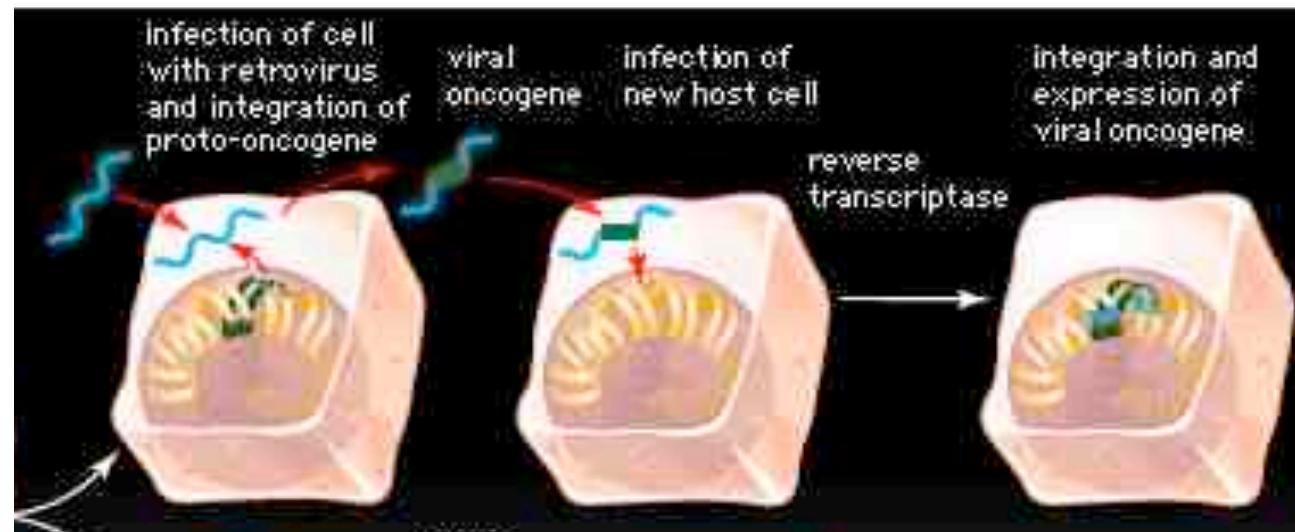
Alignments & Datenbanksuchen

break-thru...



Doolittle et al. 1983, Waterfield et al. 1983 > DB-Suche

„.... das virale Oncogen v-sis ist eine modifizierte Form des zellulären Gens für den platelet-derived growth factor (PDGF)!!“



Spezielle Such-Algorithmen erforderlich...

- „optimale“ Algorithmen wie N-W oder S-W sind viel zu aufwändig für das Durchsuchen großer Datenbanken
- „**Heuristische**“ Methoden des Sequenzvergleichs ermöglichen schnelle Alignments, jedoch mit geringer Gefahr, eine noch besser passende Sequenz zu übersehen.

Heuristik ([altgr.](#) εύρισκω *heurísko* „ich finde“; von εύρισκειν *heurískein* ‚auffinden‘, ‚entdecken‘) bezeichnet die Kunst, mit begrenztem Wissen ([unvollständigen Informationen](#)) und wenig Zeit dennoch zu wahrscheinlichen Aussagen oder praktikablen Lösungen zu kommen.^[1]

Allgemeine Strategie heuristischer Methoden

- Suchsequenz in kurze Abschnitte („words“ bzw. „ k -tuple“) aufbrechen (Wilbur und Lipman, 1983).
- zunächst sehr schnell nach „word hits“ in der DB suchen
- hat man mehrere „word hits“ in einem DB-Eintrag?
Dann dort genauer schauen...

Zwei Programmfamilien für die DB-Suche

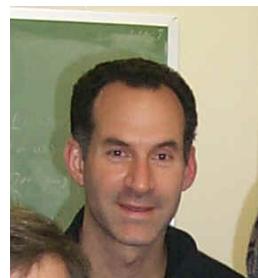
- **FASTA (Lipman und Pearson 1983)**
- **BLAST (Altschul et al. 1991, 1997)**

„Basic Local Alignment Search Tool“

Beide Tools machen lokale Alignments!



Stephen Altschul



David Lipman



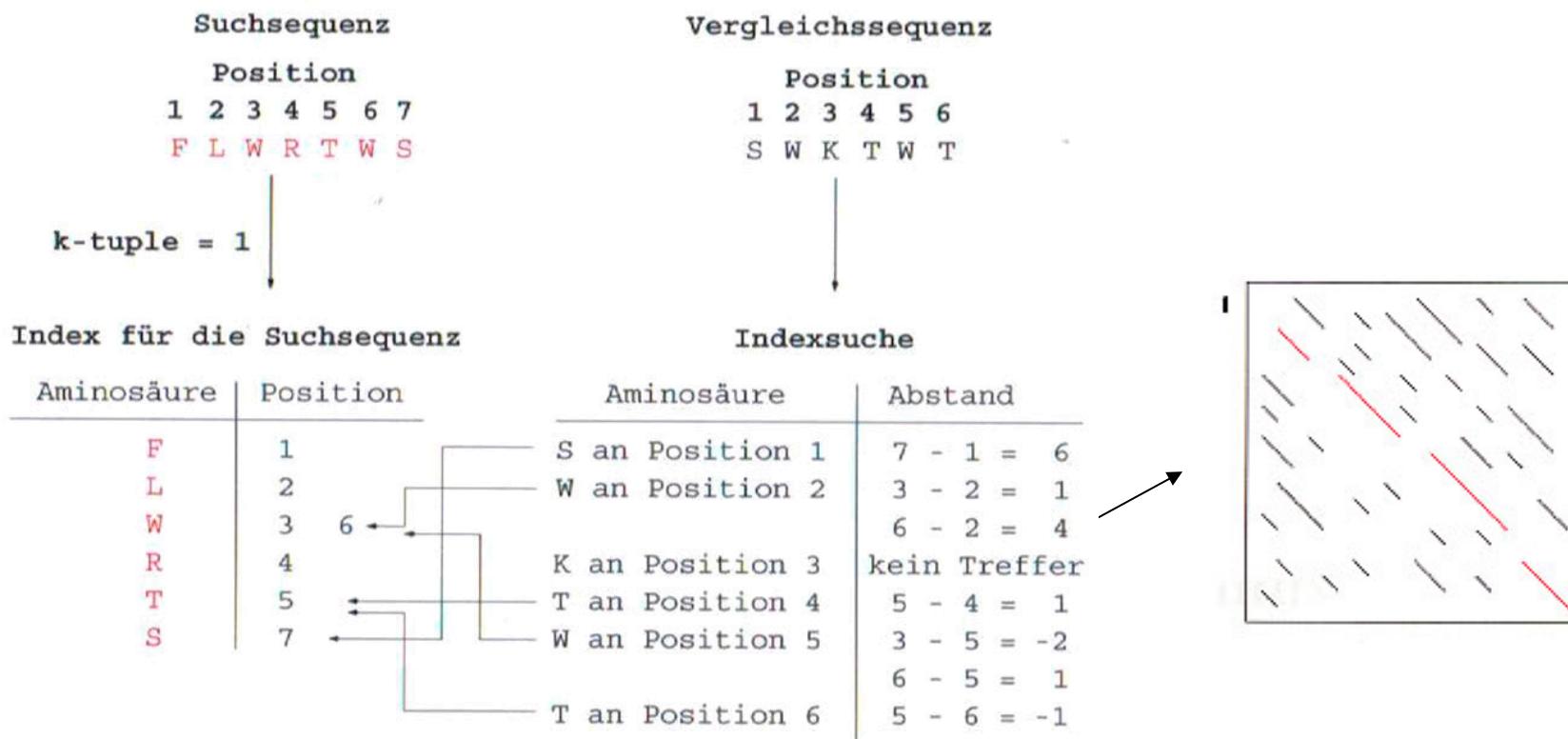
Bill Pearson

FASTA

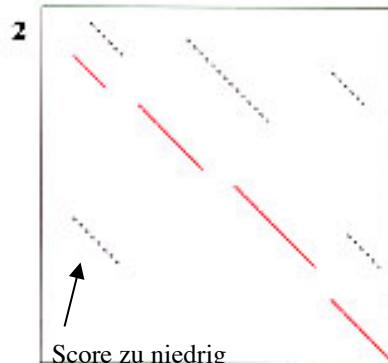
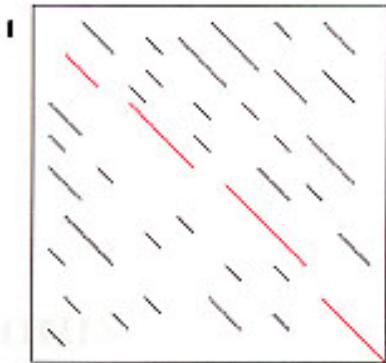
1. Erzeugung eines **Index** (lookup table) der Suchsequenz:

> Länge der Index-Einträge = **k-tuple**

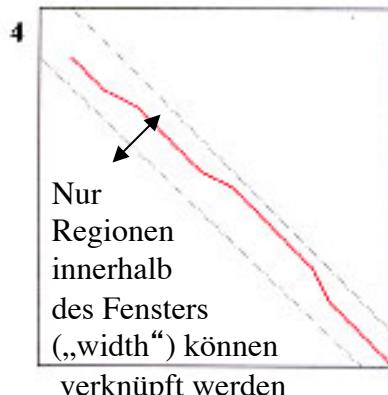
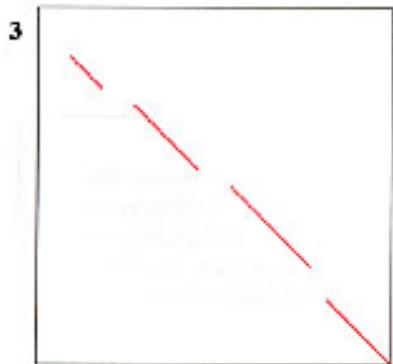
> mit Index wird nach ident. Posit. in Vergleichssequenz gesucht



FASTA



2. Verlängerung der ersten Matches ohne gaps; Berechnung des **init1**-Scores



3. Verknüpfung unter Einführen von gaps (**initn**-Score = **init1** - joining penalty)

4. Verknüpfung der **init1**-Regionen mit höchstem Score (> „opt“) nach der sensitiveren S-W-Methode (nach Normalisierung auf Länge wird **Z-Score** angegeben)



FASTA

- sensitiv, aber vergleichsweise etwas langsam
- *default:* k-tup (DNA) = 6, k-tup (Protein) = 2
- höheren k-tup
 - > mehr speed, weniger noise, weniger Sensitivität
- niedrigeren k-tup > höhere Sensitivität für entfernte Matches
- **größter Nachteil:** nur ein einziges optimales lokales Alignment wird gezeigt
 - > Nachbearbeitung (z.B. mit LALIGN) erforderlich



http://www.ebi.ac.uk/fasta33/nucleotide.html

Phylogeny p... and system Phylemon LEO UCSC HighWire Press molgen News Google NCBI HomePage Ele...

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

EBI > Tools > Similarity & Homology > Fasta

Fasta - Nucleotide Similarity Search

Provides sequence similarity searching against nucleotide and protein databases using the Fasta programs. Fasta can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity searching against complete [proteome](#) or [genome](#) databases using the [Fasta programs](#).

[Download Software](#)

PROGRAM	DATABASES	RESULTS	SEARCH TITLE	YOUR EMAIL
fasta3	Nucleic Acid EMBL Release EMBL Updates EMBL Coding Sequence EMBL Environmental	email	Sequence	
MATRIX	GAP OPEN DNA STRAND	GAP EXTEND HISTOGRAM	KTUP MOLECULE TYPE	EXPECTATION UPPER VALUE LOWER VALUE
none	-14	-4	6	10.0
both	no	DNA		default
SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	STATISTICAL ESTIMATES
50	50	START-END	START-END	none Regress

Enter or Paste a **DNA/RNA** Sequence in any format: [Help](#)

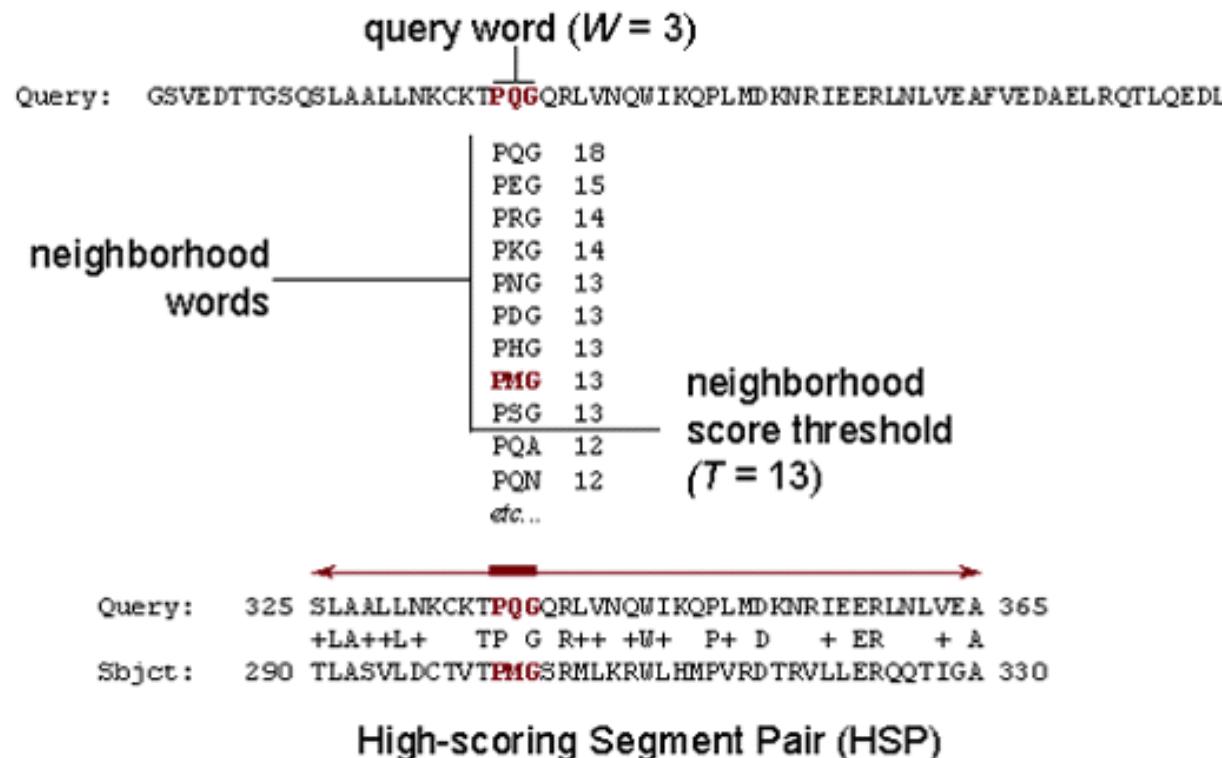
Upload a file: Keine Datei ausgewählt

Verschiedene Typen von FASTA-Suchen stehen auf EBI-Seite zur Verfügung...

BLAST

Altschul et al. 1990, 1997
> 74000 Zitate

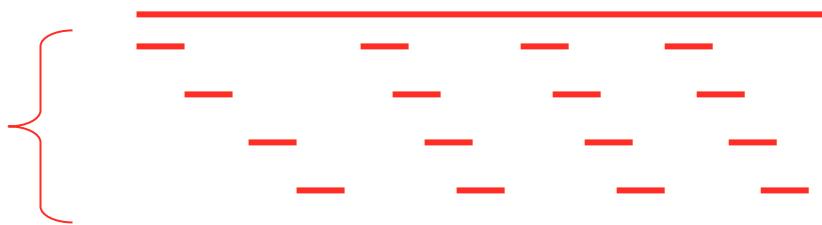
- schneller als FASTA!
- liefert mehrere lokale Alignments
- berücksichtigt Ähnlichkeiten!



(word size $W = 11$ bei DNA)

BLAST

Index-
Einträge
der Länge w



Suchsequenz



erster Hit

Datenbanksequenz



—
Fensterlänge A

Gibt es 2. Hit?



← → ← →

HSPs

zwei lokale Alignments,
Verknüpfung über Lücken falls möglich erlaubt

BLAST

1. Suchsequenz wird in ‚words‘ der Länge w „zerbrochen“
2. mit Index dieser ‚words‘ wird DB durchsucht
3. ein „word hit“ liegt vor, wenn das ‚word‘ exakt oder in ähnlicher Form* (threshold-Score $> T$) erkannt wird
 - > word size kann hoch bleiben (speed) ohne Sensitivitätsverlust
 - > erhöhe T : weniger ‚background words‘, schneller
 - > erniedrige T : entfernte Verwandschaften zu finden
4. ausgehend von ‚word hit‘ wird lokales optimales alignment verlängert, bis Score S durch mismatches stark abfällt (= HSP, high-scoring segment pair)
 - > dabei können kleine Lücken toleriert werden

*das kann FASTA nicht!

BLAST bewertet die Signifikanz eines Alignments !!

Score E-Wert

3.	dbj BAA29916 (AP000003)	170aa long hypothetical protein [P...	107	6e-23
4.	sp Q57951 Y531_METJA	HYPOTHETICAL PROTEIN MJ0531 >gi 212801...	91	4e-18
5.	gi 2622094	(AE000872) conserved protein [Methanobacterium t...	85	4e-16
6.	gi 2621993	(AE000865) conserved protein [Methanobacterium t...	81	4e-15
7.	gi 2621194	(AE000803) conserved protein [Methanobacterium t...	80	7e-15

$$E = k m n e^{-\lambda S}$$

k, Konstante

λ , Konstante für Normalisierung des HSP-scores

m, Nukleotidanzahl in Suchsequenz

n, Nukleotidanzahl in Datenbank

S = score des HSP-matches

Der **E (Expect)-Wert** gibt die Zahl der Treffer an, die in einer Datenbank der verwendeten Größe zufällig erwartet werden können.

(Je kleiner der Wert, desto höher die Signifikanz des betrachteten Treffers)

Wann habe ich einen guten Treffer?

Faustregel:

- DNA: $< e -6$, $> 60\%$ Sequenzidentität
- Protein: $< e -3$, $> 25\%$ Identität

Grenzfall „Neuroglobin“:

Sequences producing significant alignments:	Score (bits)	E Value
dbj AU036042.1 AU036042 AU036042 Sugano mouse brain mnccb Mu...	41	0.003
gb BE648697.1 BE648697 UI-M-BG1-aid-e-09-0-UI.r1 NIH_BMAP_M...	37	0.045
gb AW548186.1 AW548186 L0032E08-3 Mouse E12.5 Female Mesone...	32	0.89
gb AW546198.1 AW546198 L0005A02-3 Mouse E12.5 Female Mesone...	32	0.89
gb AW548428.1 AW548428 L0036F07-3 Mouse E12.5 Female Mesone...	32	1.1
emb AL362383.1 AL362383 AL362383 ICRFp 522 and 523 Mus musc...	32	1.3

[dbj|AU036042.1|AU036042](#) AU036042 Sugano mouse brain mnccb Mus musculus cDNA clone MNCb-7114.
Length = 740 Score = 40.8 bits (126), Expect = 0.003
Identities = 33/154 (21%), Positives = 63/154 (40%), Gaps = 5/154 (3%)
Frame = +3

Query: 1 MNSDEVQLIKKTWEIPVATPTDSGAAILTQFFNRFPSNLEKFPFRDVPL---EELSGNAR 57
M E +LI+++W + +P + G + + F PS L F + E+ +
Sbjct:156 MERPESELIROQSWRVVSRSPLEHGTVLFARLFALEPSLLPLFQYNGRQFSSPEDCLSSPE 335

BLAST : Endecke die Möglichkeiten...

blastn DNA-Sequenz ÷ DNA-DB
 > für nahe Verwandschaft; beide Stränge verglichen

blastp As-Sequenz ÷ Protein-DB
 > für entfernte Verwandschaft (default: BLOSUM62)



Ich habe die DNA-Sequenz aus einer exotischen Spezies neu entschlüsselt.

Ich will wissen, ob diese DNA-Sequenz ein bekanntes Protein kodiert, und welches Protein aus welcher anderen Spezies am Ähnlichsten ist...

Was muss BLAST idealerweise können, um das zu beantworten?

BLAST : Endecke die Möglichkeiten...

- blastx**
 - DNA-Seq > in 6 Leserahmen translatiert
 - ÷ Protein-DB
 - > findet mögliche Proteine in einer nicht-charakterisierten („anonymen“) DNA-Sequenz (z.B. EST)!

- tblastn**
 - As-Seq gegen DNA-DB (6-frame translatiert!)
 - > findet nicht-annotierte Genregionen in DNA-DB-Sequenzen

- tblastx**
 - 6-frame-Translation einer DNA-Seq ÷
 - 6-frame-Translation einer DNA-DB
 - > Analyse von ESTs auf Proteinebene zur Detektion entfernter Verwandschaft
 - > kann nicht mit nr-DB benutzt werden (zu aufwändig)

Meistbesucht news LEO Wikipedia Google UniMail NCBI NTV SZ süd-D ZEIT UCSC molgen Aktuelle Nachrichten Uni Mainz fileahre Inbox - Outlook

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST®

Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

BLAST+ 2.7.1 released
A new version (2.7.1) of the BLAST+ executables is now available.
Mon, 23 Oct 2017 08:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Human Mouse Rat Microbes

Standalone and API BLAST

[Download BLAST](#)
Get BLAST databases and executables

[Use BLAST API](#)
Call BLAST from your application

[Use BLAST in the cloud](#)
Start an instance at a cloud provider

Viele spezialisierte BLAST-Optionen weiter unten auf der Seite... 18

BLAST-Suche (1)

Copy/paste

DB wählen!

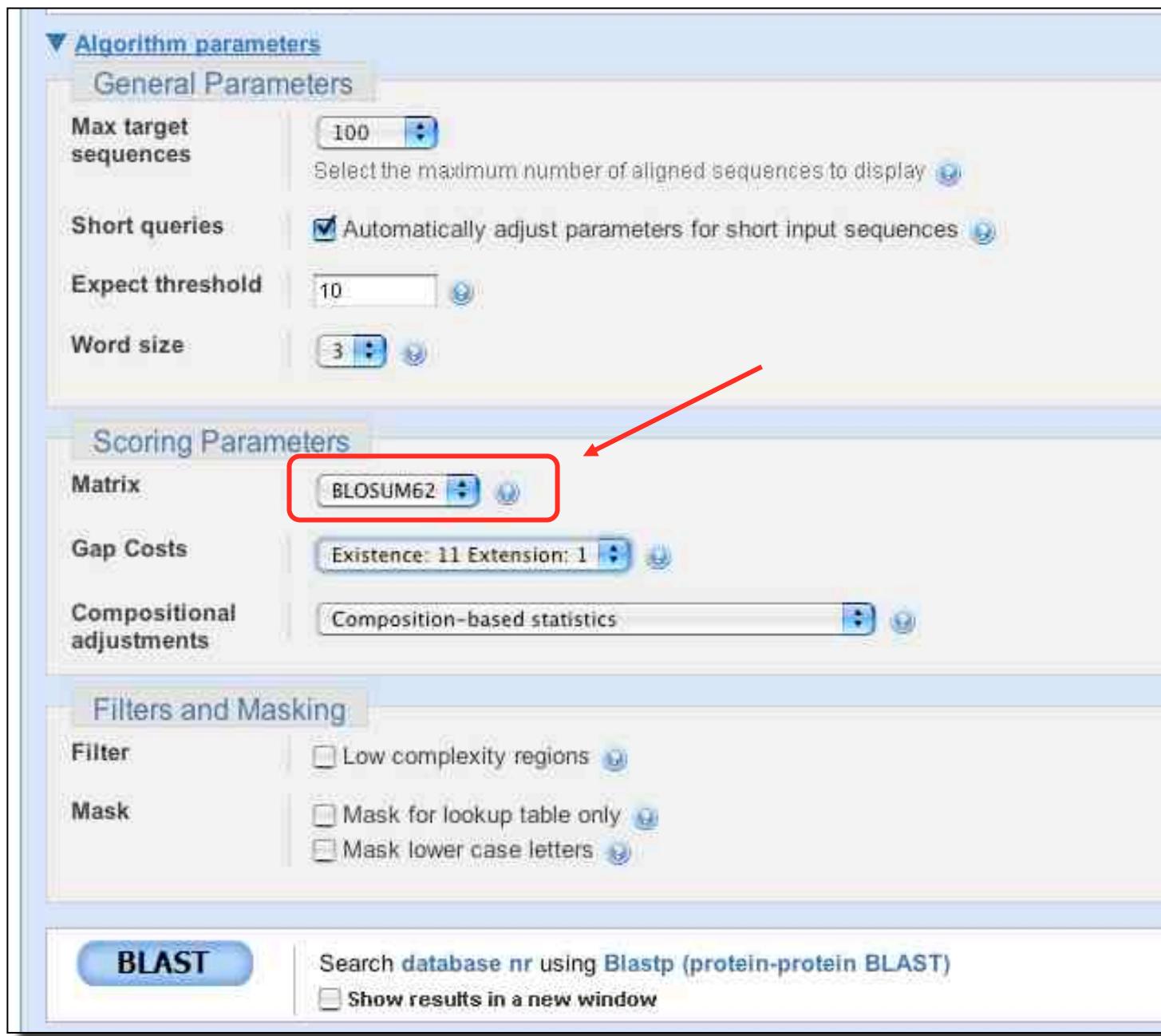
Algorithmus wählen

BLAST-Algorithmen auf Nt-Ebene

- Megablast: längere Word size, daher schneller für gut passende matches, aber weniger sensitiv als BlastN:
für Suchen mit >80 % Identität
- *discontiguous*
Megablast: „unterbrochene“ word hits erlaubt;
ignoriert mismatches der 3. Kodonposition
in kodierenden Sequenzen;
sensitiver als BlastN für entfernte Suchen
- BlastN: „gut für den Rest...“

BLAST-Suche (2)

evtl. die Parameter verändern



▼ Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Scoring Parameters

Matrix: BLOSUM62 (highlighted with a red box)

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Composition-based statistics

Filters and Masking

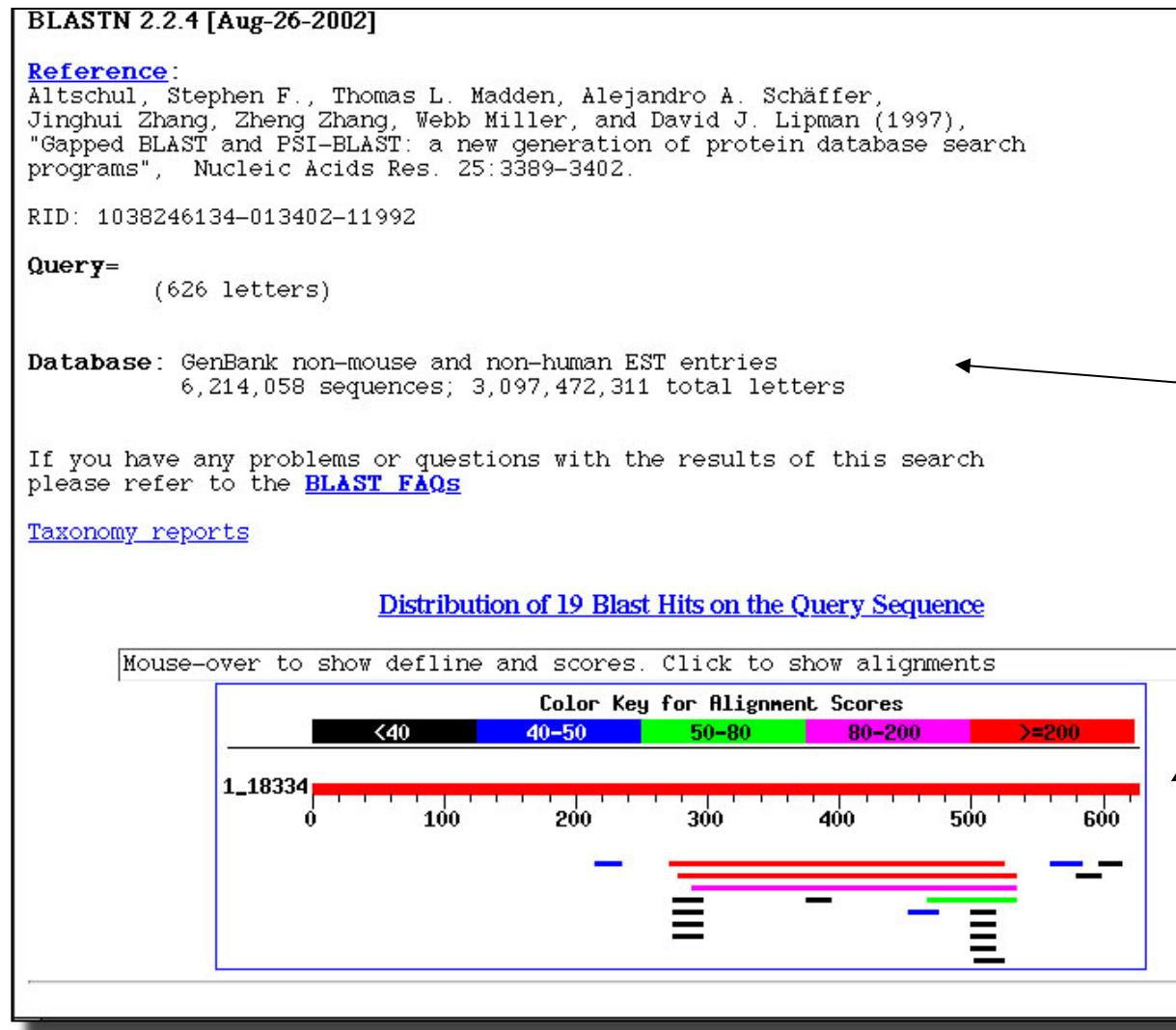
Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask lower case letters

BLAST

Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window

BLAST-Suche (3) ..das Ergebnis



Zur Erinnerung...

Suchsequenz
("Query")

Matches mit
unterschiedlicher
Qualität

BLAST- Suche (4) ..das Ergebnis

Anschauen:

			Score (bits)	E Value
Sequences producing significant alignments:				
gi 11089284 gb BF198326.1 BF198326	248106	MARC 2PIG Sus scrofa cDNA 5'	234	5e-59
gi 6943903 gb AW416021.1 AW416021	50699	MARC 2PIG Sus scrofa cDNA 5'	234	5e-59
gi 8327284 gb BE032275.1 BE032275	131150	MARC 1PIG Sus scrofa cDNA 5'	174	4e-41
gi 6952978 gb AW425031.1 AW425031	50990	MARC 2PIG Sus scrofa cDNA 5'	56	3e-05
gi 9922596 dbj AV663566.1 AV663566	AV663566	Bos taurus brain cDNA 5'	48	0.006
gi 11397297 gb BF409322.1 BF409322	UI-R-CA1-ble-f-11-0-UI.s...		40	1.5
gi 9297954 gb BE356397.1 BE356397	DG1_124_B03.g1_A002	Dark	40	1.5
gi 24558097 gb CA339999.1 CA339999	NISC_ly10c01.y1	NCI_CGAP...	38	5.9
gi 23181414 emb AL916116.1 AL916116	AL916116	PJR-Z1+Z2 Dani...	38	5.9
gi 18235816 dbj AV940019.1 AV940019	AV940019	K. Sato unpubl...	38	5.9
gi 18234096 dbj AV938299.1 AV938299	AV938299	K. Sato unpubl...	38	5.9
gi 18211618 dbj AV915841.1 AV915841	AV915841	K. Sato unpubl...	38	5.9
gi 17995218 gb BM286192.1 BM286192	526142	MARC 3BOV Bos tau...	38	5.9
gi 15206276 gb BI429044.1 BI429044	fr71g03.y1	zebrafish adu...	38	5.9
gi 14261240 gb BG884148.1 BG884148	fp72e05.y1	Zebrafish adu...	38	5.9
gi 13418848 dbj AT003990.1 AT003990	AT003990	POSLM01 Pleuro...	38	5.9
gi 13091106 gb BF629137.2 BF629137	HVSMEb0010A15f	Hordeum v...	38	5.9
gi 12354741 gb BF937421.1 BF937421	fm64f06.y1	Zebrafish adu...	38	5.9
gi 7131799 emb AJ398145.1 AJ398145	AJ398145	dkfz426 Gallus...	38	5.9
Alignments				
<input type="button" value="Get selected sequences"/>	<input type="button" value="Select all"/>	<input type="button" value="Deselect all"/>		
<input type="checkbox"/> >gi 11089284 gb BF198326.1 BF198326	248106	MARC 2PIG Sus scrofa cDNA 5'		
	Length = 539			
Score = 234 bits (118), Expect = 5e-59				
Identities = 223/258 (86%)				
Strand = Plus / Plus				
Query: 278 gtgatgctagtgattgatgctgcagtgaccaacgtggaggacctgtttcattggaggag	337			
Sbjct: 1 gtgatgctgtgattgatgctgcagtgactaacgtggaggacctgtttcgctggaggag	60			
Query: 338 tacctgaccagcttggcgaggaagcatggcagtggagtggactcagtccttcgt	397			
Sbjct: 61 taccttgcggcctggcgaggaagcacggcagtgggtgtgaagctcagtccttcgt	120			
USW...				

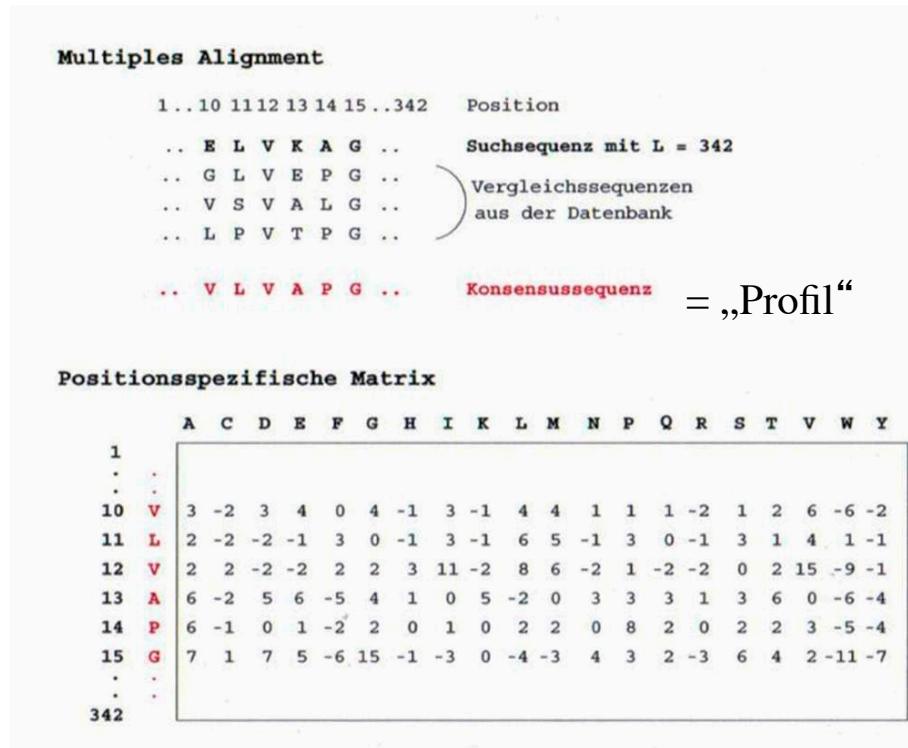
...das erste Alignment (Query = Suchsequenz)



PSI-BLAST

Position-specific iterated BLAST

- speziell für die Suche sehr entfernt verwandter Proteine, die durch BLASTP nicht gefunden werden



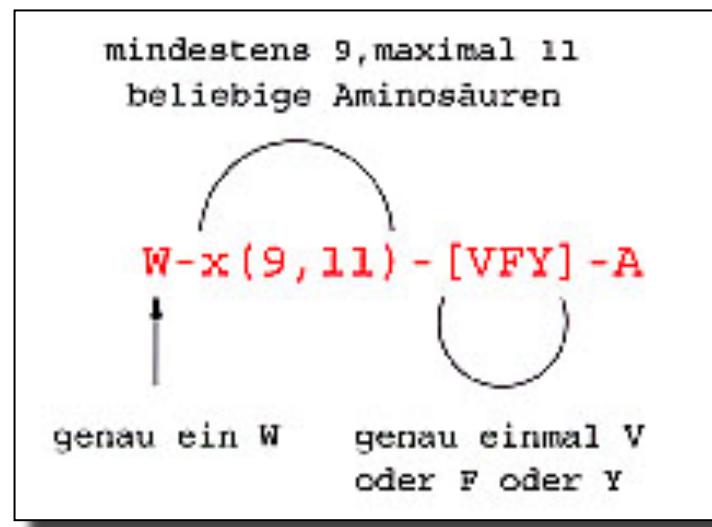
1. Erste Suche = einfacher BLAST
 2. Matches untereinander schreiben,
> Konsensussequenz errechnen
„Profil“)
 3. „Positionsspezifische“ Substitutions-
Matrix errechnen
 4. BLAST mit dem „Profil“ und der
PSSM mehrfach wiederholen
- Gezieltere Suche nach verwandten
Proteinen wegen Positionsinformation



PHI-BLAST

Pattern-hit initiated BLAST

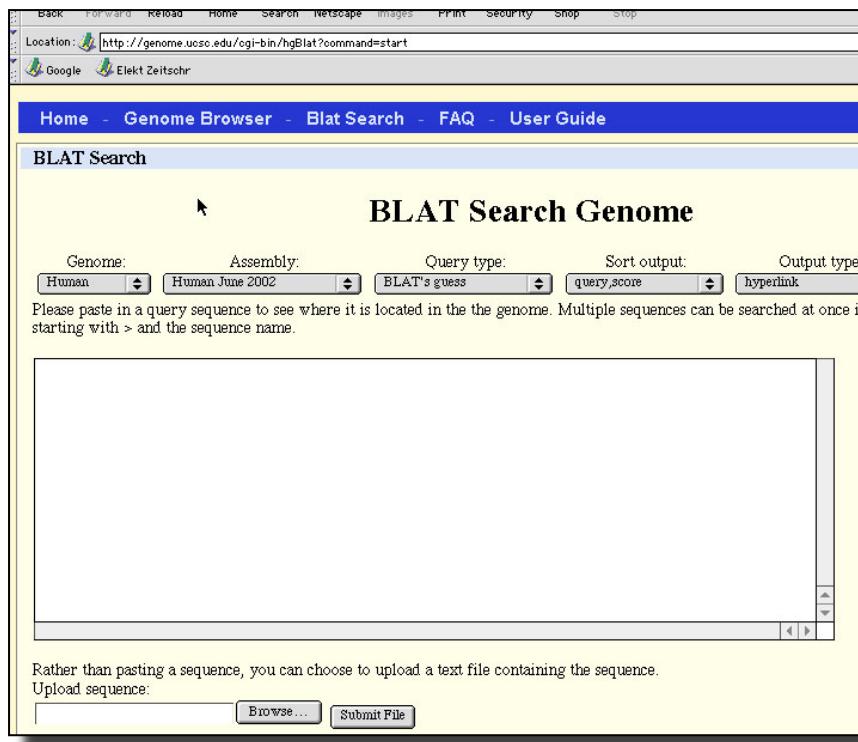
- sucht Sequenzmuster („Signatur“), das typisch für Proteindomäne ist
- Suche über „qualitatives“ Sequenzmotif (PSI-Blast über quantitatives Motiv)



- Muster zusammen mit Suchsequenz gegen DB laufen lassen
- Treffer = Proteine mit Ähnlichkeit zur Suchsequenz und das Motiv enthaltend

Ultraschnelle DB-Suche über BLAT

„BLAST-like alignment tool“



The screenshot shows a web browser window with the URL <http://genome.ucsc.edu/cgi-bin/hgBlat?command=start> in the address bar. The page title is "BLAT Search Genome". It features a search form with dropdown menus for "Genome" (Human), "Assembly" (Human June 2002), "Query type" (BLAT's guess), "Sort output" (query_score), and "Output type" (hyperlink). Below the form is a text input area with placeholder text: "Please paste in a query sequence to see where it is located in the genome. Multiple sequences can be searched at once by starting with > and the sequence name." At the bottom, there is a note: "Rather than pasting a sequence, you can choose to upload a text file containing the sequence." followed by "Upload sequence:" and two buttons: "Browse..." and "Submit File".

<http://genome.ucsc.edu/cgi-bin/hgBlat?command=start>

- DNA-BLAT findet 40 Bp (>95% id) bzw. perfekte matches von >33Bp
- Protein-BLAT findet 20 aa (<80%id)
- Index (DNA) enthält alle nicht-überlappenden 11-mere des Genoms (1 Gb RAM)!!!
- Index wird gebraucht um passende Regionen im Genom schnell zu identifizieren, die dann für genaueren Vergleich „hochgeladen“ werden



Inbox - Outlook

ONLINE ANALYSIS TOOLS
(INTERNET RESOURCES for MOLECULAR BIOLOGISTS)

Analysis of nucleotide and protein sequence data was initially restricted to those with access to complicated mainframe or expensive desktop computer programs (for example PC/GENE, Lasergene, MacVector, Accelrys etc.). The availability of online tools permits even the novice molecular biologist the opportunity to derive a considerable amount of useful information from nucleotide or protein sequence data. For those with no experience I have provided three [sequences](#): (a) a DNA sequence, (b) a protein sequence, and (c) four protein sequences presented in [FASTA format](#). Prior to trying out a Web Site select the sequence and copy to clipboard. Each of the items in blue text is hyperlinked to a site on the Web. Each of these Web Sites has a box into which you can "Paste" your sequence. Click on the button labeled "Search," "Run" or "Submit." If in doubt use the default setting that the sites provide, but for the more adventuresome some of the sites offer the chance of modifying the search strategy.

Bioinformatic tutorials

- [ONLINE RESOURCES](#) (tutorials and glossaries) - Needs work

Carbohydrates

- [TERTIARY STRUCTURE PREDICTIONS OF SACCHARIDES](#)

DNA sequence analysis

CMS Molecular Biology

<http://www.sdsc.edu/~ginai/cmsmbr/>

UNIMAIL JGU Anmelden Google Phylogeny p... and system ilias 78 molgen UCSC LEO

CMS MBR CMS Molecular Biology Resource Portal to Data Analysis Tools & Databases

Science means simply the aggregate of all the recipes that are always successful. The rest is literature. Author: P. Valery

POPULAR TOOLS

- Conferences
- Academic Programs
- Societies
- Workshops & Training
- Institutes & Depts
- Techniques-Protocols
- Project Information
- Contributions
- Contact Information
- Site Map
- Mirror Sites**
- Europe
- ABI/U Aix-Marseille (France)
- U Bielefeld (Germany)
- IFOM (Italy)
- IBBA (Italy)
- ILLUSIO (Poland)
- Asia
- Asahi Glass Co.

PROTEIN ANALYSIS

- [Sequence Search & Analysis Tools](#)
- [Sequence Databases](#)
- [Structure Prediction & Databases](#)
- [ID Based Upon Physical & Chemical Data](#)
- [Physico-Chemical Features Analyses](#)
- [Enzyme Info and Structure Databases](#)
- [Immunology Databases & Info Servers](#)
- [Protein Family Resources](#)
- [Techniques & Protocols](#)

GENETICS-PHYLOGENY

- [Phylogenetic Analyses](#)
- [General Phylogeny Resources](#)

DNA ANALYSIS

- [Sequence Homology & Structure Analyses](#)
- [Physico-Chemical Properties & Analyses](#)
- [Sequence & Structure Databases](#)
- [Organism-Specific Genome Databases](#)
- [DNA Restriction Enzyme Databases](#)
- [Gene Family Resources](#)
- [Techniques & Protocols](#)

BIOCHEMISTRY

- [Metabolism & BioCompounds](#)

ExPASy: SIB Bioinformatics Resource Portal - Categories

UNIMAIL JGU Anmelden Google Phylogeny p... and system ilias 78 molgen UCSC LEO News ElektrZeitschr

ExPASy Bioinformatics Resource Portal

Query all databases

Visual Guidance

Categories

proteomics

genomics

- sequence alignment
- similarity search
- characterisation/annotation
- structural bioinformatics
- systems biology
- phylogeny/evolution
- population genetics
- transcriptomics
- biophysics

Databases

- EPD • collection of eukaryotic promoters • [\[more\]](#)
- OrthoDB • Hierarchical catalog of eukaryotic orthologs • [\[more\]](#)
- SwissRegulon • annotations of regulatory sites • [\[more\]](#)
- smirnaDB • miRNA expression profiles analysis • [\[more\]](#)
- OMA • orthology inference among complete genomes. • [\[more\]](#)
- arrayMap • Curated array data repository for cancer genomics • [\[more\]](#)
- CLIPZ • binding sites of RNA-binding proteins • [\[more\]](#)
- EIMMo • miRNA target predictions • [\[more\]](#)
- GPSDB • gene and protein synonyms • [\[more\]](#)
- ImmunoDB • insect immune-related genes and gene families • [\[more\]](#)

Tools

- EPD • collection of eukaryotic promoters
- smirnaDB • miRNA expression profiles analysis
- OMA • orthology inference among complete genomes
- ALF • simulation of gene expression
- Alignment tools
- arrayMap • Curated array data repository for cancer genomics
- Association Viewer
- BayeScan • identify
- BLAST - NCBI • Bio
- BLAST - PBIL • BLA

Inbox - Outlook

old.nsu.ru

old.nsu.ru/education/i4biol/mirrors/abim.html

This is a local (NSU) copy of the page from [L'Atelier BioInformatique de Marseille \(ABIM\)](#).

- [Search a sequence database](#)
- [Nucleic acids sequences](#)
- [Patterns \(proteins\)](#)
- [Prediction on protein sequences](#)
- [Alignments - Phylogeny](#)
- [Analysis tools package](#)