Lecture 2
Sequence Searching: Pattern Matching and Blast

- Sequence format and notation
- Sequence comparison models
- Sequence searching phases
- BLAST: 3 step heuristic
- Hit statistics: P-value and E-value
- BLAST result interpretation
- BLAST flavours

Sequence Searching

Sequence

- a string of characters that represents the chain of building blocks in a biopolymer
- building blocks are amino acids (proteins) and nucleotides (DNA, RNA)
- the side chains define the sequence

Sequence Searching

Sequence databank searching is the process of extracting homologues of one or several query sequence(s) from a sequence database.

Sequence Format Conventions

- A sequence is composed of a name (often including an accession number) and the residue string.

- A sequence databank is a formatted (and often sorted) list of sequences (here FASTA format).

>1epi epidermal growth factor (Mus musculus)
NSYPGCPSSYDGYCLNGGVCMHIESLDSYTCNCVIGYSGDRCQTRDLRWWELR
>1ixa EGF-like module coagulation factor (Homo sapiens)
VDGDQCESNPCLNGGSCKDDINSYECWCPFGFEGKNCEL

- Proteins are written from the N-terminus to the C-terminus.

- Nucleotide sequences are defined within a 'reading frame'.

Sequence Notation - Positions and Chains

Mutations

- Y35G-BPTI (bovine pancreatic trypsin inhibitor)
  mutation from Tyr to Gly at position 35
- K(B29)P-insulin
  mutation from Lys to Pro at position 29 in chain B
- Des(B27-B30)-insulin-B26-carboxamide
  residues 27 to 30 deleted in chain B and C-terminus amidated

Chain notation

Chains are denoted A, B, C, D,... in successive order.

Sequence Searching and Sequence Alignment

• Sequence searching and sequence alignment are different techniques.

• Sequence searching uses sequence matching (like alignment), but sequence alignment works on a pre-defined sequence set.

• In sequence searching one tries to extract homologous sequences, whereas in sequence alignment one assumes homology and tries to correctly identify similarities between homologues.

Sequence Searching Problem

given

required

Sequence Searching Problem

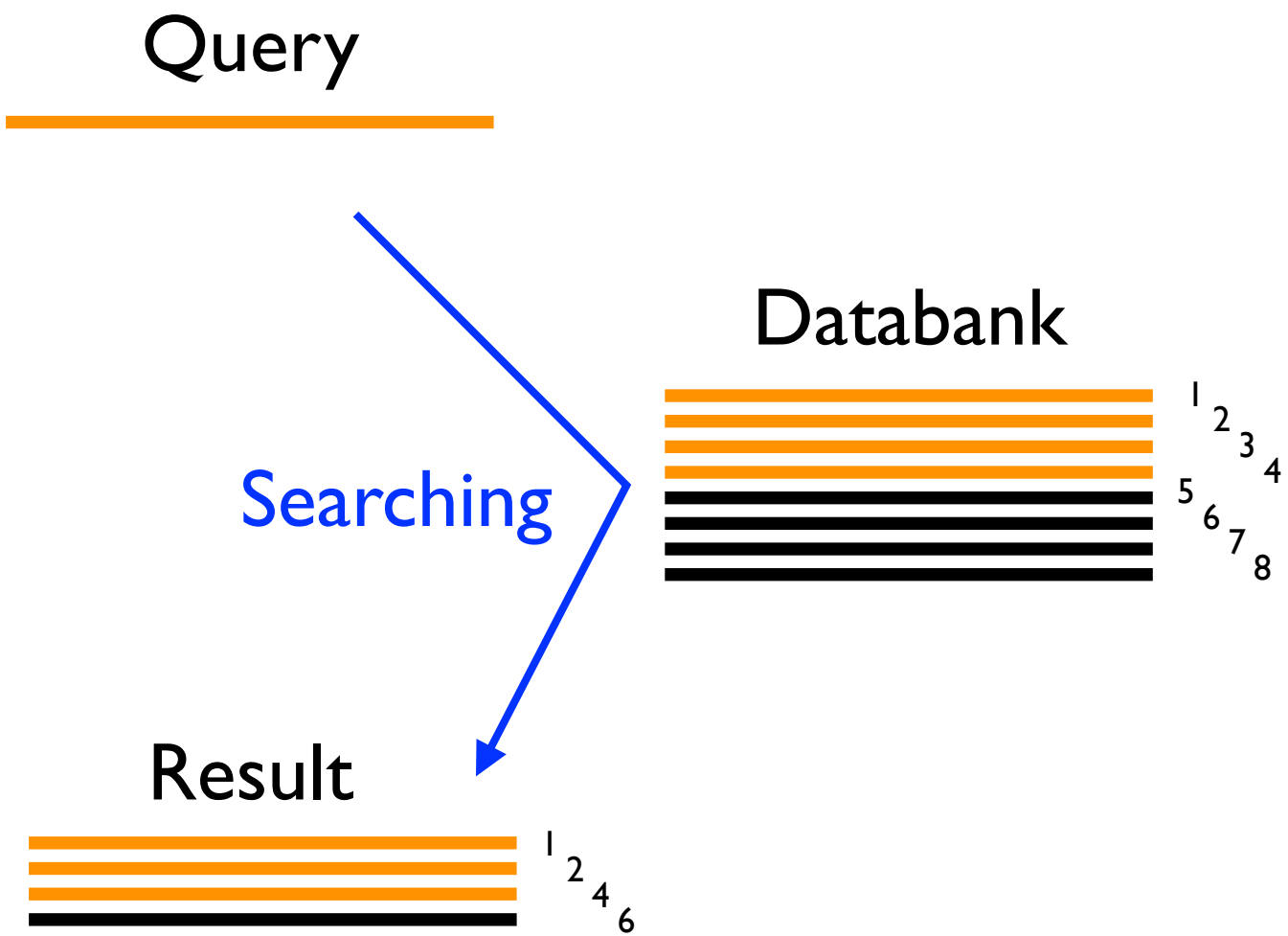
given

- Query: sequence to search with (assume 300 aa)
- Database (search space): very many sequences (assume > 100000)
- Goal: find sequences homologous to query

required

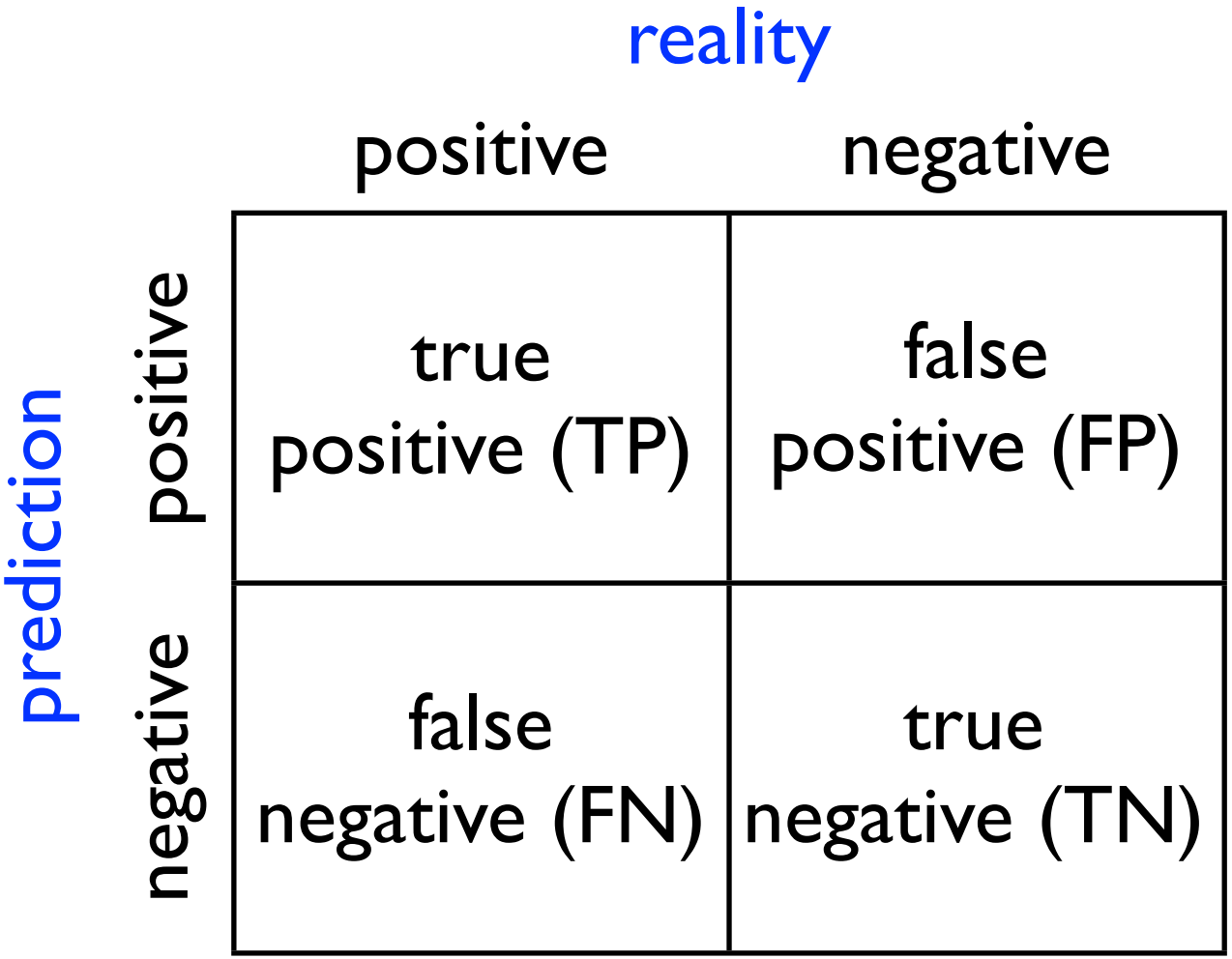
- a fast tool
- primarily a filter: most sequences will be unrelated to the query
- fine-tune the alignment later

Sequence Databank Searching

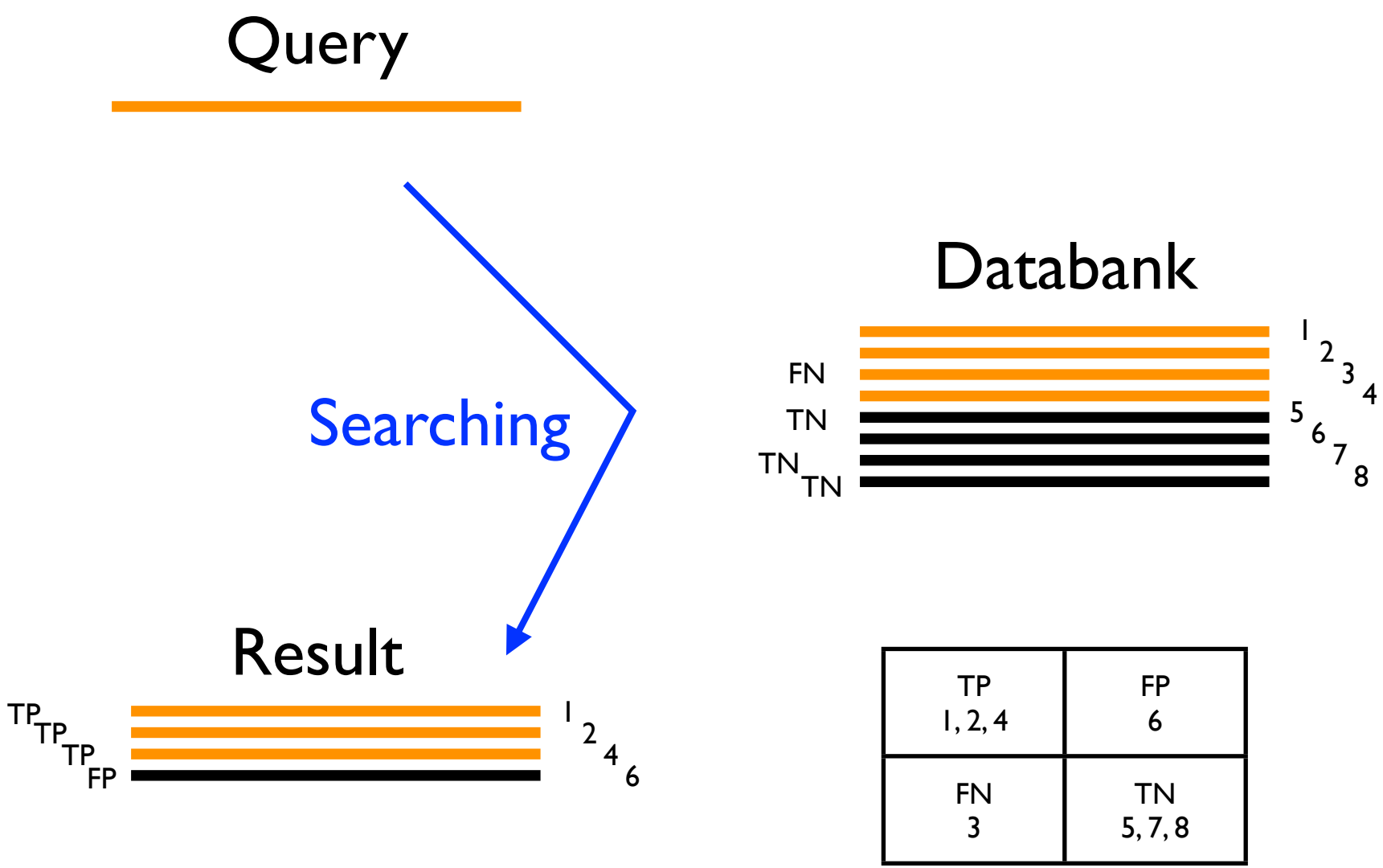


Prediction : Contingency Table

Sequence searching is a prediction about the homology relation between query and DB sequence.



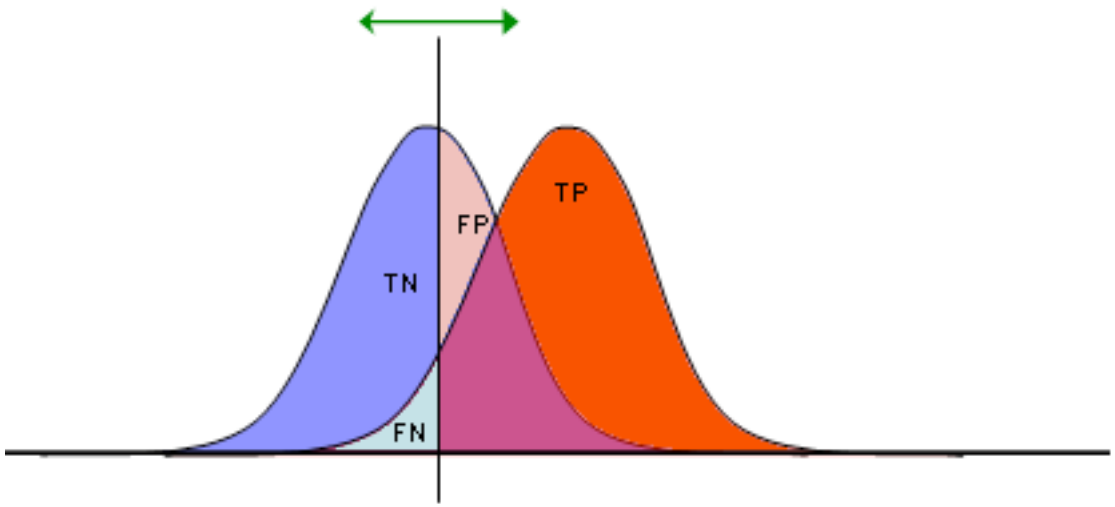
Sequence Databank Searching



TP 1, 2, 4	FP 6
FN 3	TN 5, 7, 8

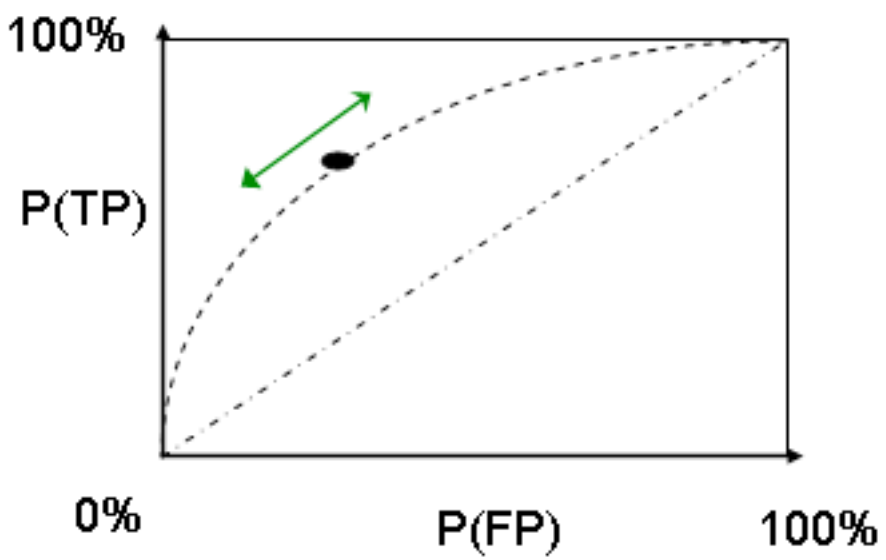
Inference / Prediction Statistics

Random and Target Distributions



Contingency Table

TP	FP
FN	TN
1	1



ROC or benchmark curve

Models Underlying Sequence Comparison

Sequence comparison (searching and alignment) is predominantly based on three models:

1. Scoring Model PAM or BLOSUM matrix + gap penalties

Information content of matched amino acid pair is defined as

Score(XZ) = log [P(XZ) / P(X)P(Z)]

with P(XZ): observed probability of matched XY in trusted alignments

and P(X),P(Z): probabilities of X and Z in a random sequence

2. Alignment Model

local or global alignment of homologues

=> alignment scores of biologically meaningful alignments

3. Random Model

local or global alignment of un-related (=random) sequence pair

=> alignment scores of random alignments

Scoring Model: Substitution Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	2																					
R	-2	6																				
N	0	0	2																			
D	0	-1	2	4																		
C	-2	-4	-4	-5	12																	
Q	0	1	1	2	-5	4																
E	0	-1	1	3	-5	2	4															
G	1	-3	0	1	-3	-1	0	5														
H	-1	2	2	1	-3	3	1	-2	6													
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5												
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6											
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5										
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6									
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9								
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6							
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2						
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3					
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17				
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10			
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4		
B	0	-1	2	3	-4	1	2	0	1	-2	-3	1	-2	-5	-1	0	0	-5	-3	-2	2	
Z	0	0	1	3	-5	3	3	-1	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3

Scoring Model: PAM Matrices

- Derive permissible mutations from trusted alignments.
- Step 1: Construct phylogenetic tree from protein alignments.
- Step 2: Fill substitution matrix with observed transition probabilities.
- A 'PAM' is an evolutionary distance:
  1 PAM = 1 accepted mutation per 100 amino acids
  250 PAM = 2.5 accepted mutations per amino acid
- There are modern versions of the PAM matrix:
  Gonnet matrices or Jones-Taylor matrices.

Scoring Model: BLOSUM Matrices

- Mostly used family of amino acid substitution matrices is the BLOSUM series (BLOSUM50, BLOSUM62).
- The BLOSUM matrices are derived from the BLOCKS database of multiple alignments (Henikoff & Henikoff, 1992).
- BLOSUM50 is derived from BLOCKS (core) alignment regions with >=50% sequence identity, Blosum62 from those >= 62%, etc.
- Note that there is NO underlying Markov model of amino acid substitution; BLOSUM is based on observed pairs of aligned residues. Higher numbers mean shorter evolutionary distance (opposite to PAM!).

Alignment Model

- The alignment model is the result of a meaningful sequence alignment, i.e. a correct alignment of two homologous sequences. In the ideal case this is identical to the alignment derived from structure superpositioning.

- Such meaningful alignments are called ‘reference alignments’ or ’trusted alignments’.

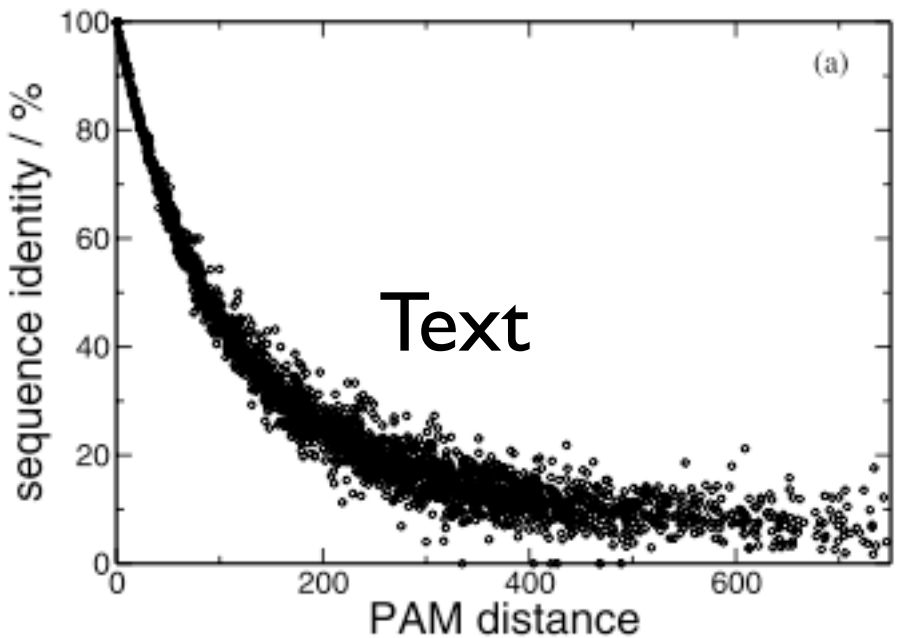
- Alignment models include local, global, probabilistic or other forms matching residue pairs.

- Generally we want to maximise P(XZ).

Random Model

- The random model is the result of an alignment of random sequences.
- The probability of finding a matched pair is just a chance event, which is given by the product of the probability to find the individual amino acids in the sequences (joint probability of independent events): P(X) * P(Z).
- The random model is important, because amino acids have different frequencies of occurrence.
- Is that the only random model you can think of?

Evolutionary Distance and Sequence Identity



Distant sequences are difficult to match,
because the amount of identical or similar residues is low
(signal / noise ratio is low).

Sequence Searching Phases

Matching Phase

The query sequence is matched (partially) with the databank sequence.

Scoring Phase

The matched residues pairs are scored the scores are summed up.

Selection phase

Based on a statistical criterion, database sequences with score above a user-defined cutoff score are returned as hits. Each sequence in the hitlist is a potential homologue.

BLAST: Basic Local Alignment Search Tool

BLAST is a program designed to compare a query sequence with every sequence in a database and to report the most similar sequences.

Basic idea:

- High scoring segments have well conserved (almost identical) parts.
- After well conserved parts have been identified, extend them to the real alignment.

BLAST: History

Smith and Waterman, 1981

Exact Local Dynamic Programming

Lipman and Pearson, 1985

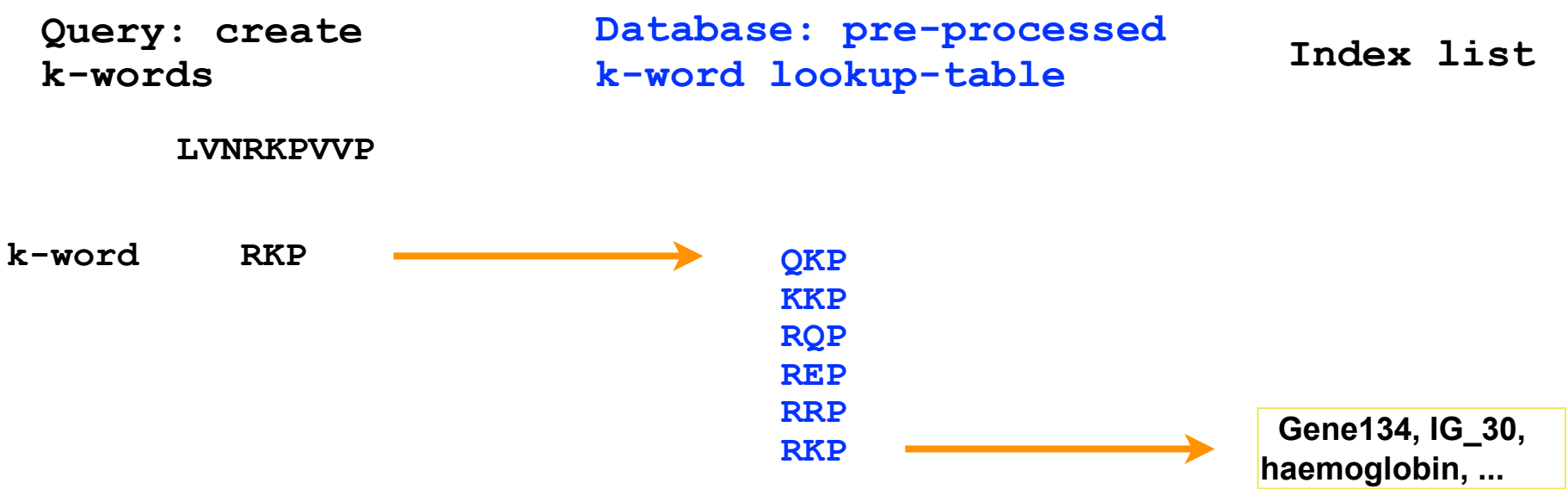
FASTA finds similar words (k-tup) on the same diagonal.

Altschul et al., 1990

BLAST: The most widely cited tool in Biology

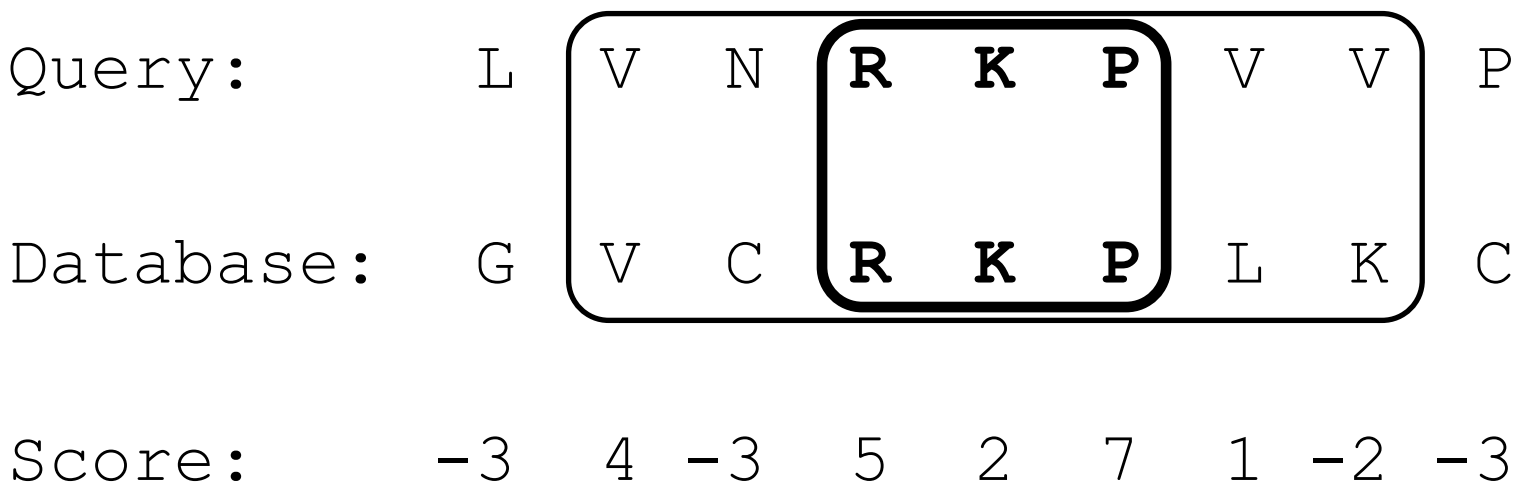
BLAST: k-word search

Considers only sequences with at least two k-word matches above a certain score.



BLAST: match extension

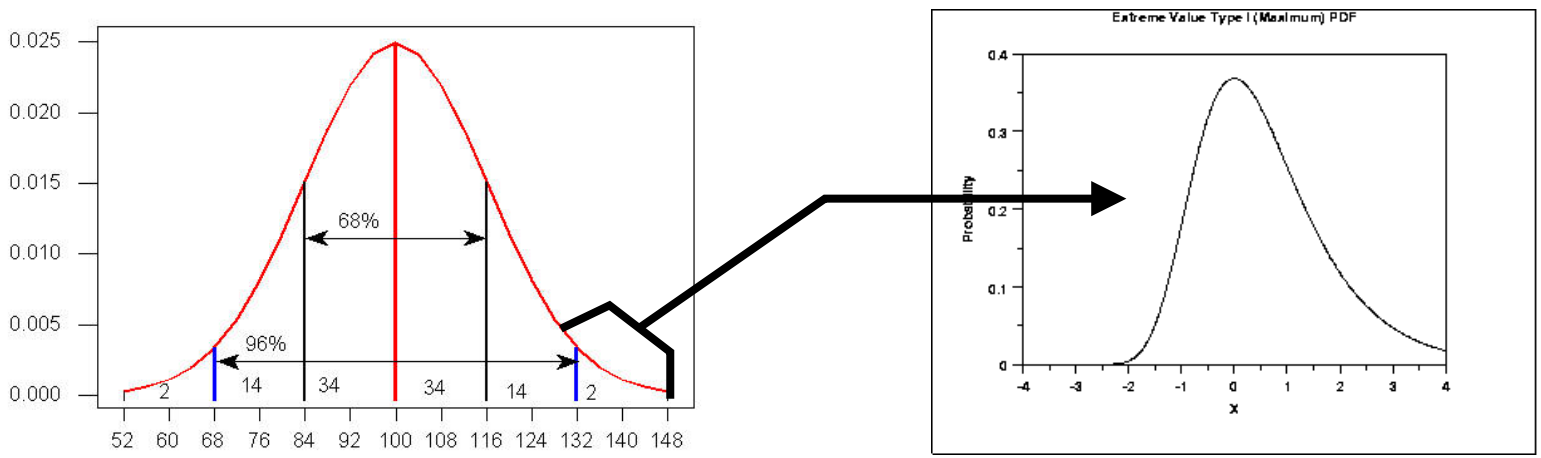
Elongate k-word matches in both directions as far as possible.



Extension terminates if score drops below value X (ungapped local alignments).

BLAST: score evaluation

The BLAST search yields a match (alignment) score for each query/hit sequence pair.



normal distribution

extreme value distribution

Sampling from the extreme end of a normal distribution yields an extreme value distribution. BLAST scores are extreme value distributed.

BLAST: Score Evaluation

P-value:
Probability of finding at least a match score x.

$$P(S \geq x) = Kmne^{-\lambda x}$$

K is calibrated with the database composition.
Lambda is calibrated with the substitution matrix.
mn is the search space, i.e. length 'm' of query sequence and length 'n' of the database sequence.

BLAST: Score Threshold

The likelihood of random sequences to yield a score greater than T increases linearly with the logarithm of the search space mn.

This gives a formula for accepting hits:

$$S > T + \log(mn) - \lambda$$

BLAST: E-value

E-value: P-value adjusted by database size.
Number of database hits expected by chance

E-values are easier to compare than P-values, because they represent a sequence count rather than a probability.

P31383 Vs YEAST

sp P31383 2AAA_YEAST Protein phosphatase PP2A regulatory subunit...	1170	0.0
sp P33892 GCN1_YEAST Translational activator GCN1 [GCN1] [Saccha...	47	1e-05
sp P53978 EF3B_YEAST Elongation factor 3B (EF-3B) [YEF3B] [Sacch...	46	2e-05
sp P22219 VP15_YEAST Protein kinase VPS15 (EC 2.7.1.-) [VPS15] [...	40	0.001
sp P32337 IMB3_YEAST Importin beta-3 subunit (Karyopherin beta-3...	39	0.002
sp P49955 S3B1_YEAST U2 snRNP component HSH155 [HSH155] [Sacchar...	38	0.005
sp Q06708 YL86_YEAST Hypothetical 99.8 kDa protein in IKI3-RPS29...	38	0.006
sp P35194 YBA4_YEAST Hypothetical 287.5 kDa protein in PDR3-HTA2...	37	0.010
sp P32917 STE5_YEAST STE5 protein [STE5] [Saccharomyces cerevisi...	33	0.15
sp P32074 COPG_YEAST Coatomer gamma subunit (Gamma-coat protein)...	32	0.25
tr Q12150 Chromosome XII COSMID 9449 [CSF1] [Saccharomyces cerev...	32	0.32

P31383 Vs YEAST

sp	P31383	2AAA_YEAST Protein phosphatase PP2A regulatory subunit...	1170	0.0
sp	P33892	GCN1_YEAST Translational activator GCN1 [GCN1] [Saccha...	47	1e-05
sp	P53978	EF3B_YEAST Elongation factor 3B (EF-3B) [YEF3B] [Sacch...	46	2e-05
sp	P22219	VP15_YEAST Protein kinase VPS15 (EC 2.7.1.-) [VPS15] [...	40	0.001
sp	P32337	IMB3_YEAST Importin beta-3 subunit (Karyopherin beta-3...	39	0.002
sp	P49955	S3B1_YEAST U2 snRNP component HSH155 [HSH155] [Sacchar...	38	0.005
sp	Q06708	YL86_YEAST Hypothetical 99.8 kDa protein in IKI3-RPS29...	38	0.006
sp	P35194	YBA4_YEAST Hypothetical 287.5 kDa protein in PDR3-HTA2...	37	0.010
sp	P32917	STE5_YEAST STE5 protein [STE5] [Saccharomyces cerevisi...	33	0.15
sp	P32074	COPG_YEAST Coatomer gamma subunit (Gamma-coat protein)...	32	0.25
tr	Q12150	Chromosome XII COSMID 9449 [CSF1] [Saccharomyces cerev...	32	0.32

P31383 Vs UniProt

sp	Q10178	S3B1_SCHPO U2 snRNP component prp10 [PRP10] [Schizosac...	50	2e-04
tr	Q9FMF9	Nuclear protein-like [Arabidopsis thaliana (Mouse-ear ...	49	2e-04
tr	O17873	Hypothetical protein F46C5.6 [F46C5.6] [Caenorhabditis...	49	3e-04
tr	Q7PLL6	CG17514-PA.3 [CG17514] [Drosophila melanogaster (Fruit...	49	3e-04
tr	Q86JC5	Hypothetical protein [Dictyostelium discoideum (Slime ...	49	3e-04
tr	Q7RM47	Hypothetical protein [PY02341] [Plasmodium yoelii yoelii]	48	4e-04
tr	O42900	Protein kinase with calcium binding domain [SPBC119.07...	48	6e-04
sp	P33892	GCN1_YEAST Translational activator GCN1 [GCN1] [Saccha...	47	0.001
sp	P53978	EF3B_YEAST Elongation factor 3B (EF-3B) [YEF3B] [Sacch...	46	0.002
tr	Q8C0Y0	Hypothetical heat repeat containing protein [8430415E0...	46	0.002
tr	Q9CRR0	8430415E04Rik protein (Fragment) [8430415E04RIK] [Mus ...	46	0.002
tr	Q7SDG6	Hypothetical protein [NCU03042.1] [Neurospora crassa]	45	0.003
tr	Q7QEW2	AgCP13260 (Fragment) [AGCG49048] [Anopheles gambiae st...	45	0.004
tr	Q9HCF0	Hypothetical protein KIAA1622 (Fragment) [KIAA1622] [H...	45	0.004
tr	Q8VD65	Hypothetical protein [PIK3R4] [Mus musculus (Mouse)]	45	0.005
tr	Q8C948	Phosphatidylinositol 3 kinase [C730038E05RIK] [Mus mus...	45	0.005
tr	Q99570	Adaptor protein [P150] [Homo sapiens (Human)]	45	0.005
sp	Q10105	YAQ5_SCHPO Putative translational activator C18G6.05c ...	44	0.006
tr	Q7XJN7	At2g40730 protein [AT2G40730] [Arabidopsis thaliana (M...	44	0.006
tr	Q8LQE7	Kinase-like protein [OJ1529_G03.14] [Oryza sativa (jap...	44	0.006
tr	Q21909	T08A11.2 protein [T08A11.2] [Caenorhabditis elegans]	44	0.006
tr	Q9CXV4	13 days embryo head cDNA, RIKEN full-length enriched l...	44	0.008
tr	Q7QFZ0	AgCP13790 (Fragment) [AGCG51792] [Anopheles gambiae st...	44	0.008
sp	O57683	S3B1_XENLA Splicing factor 3B subunit 1 (Spliceosome a...	44	0.011
sp	Q99NB9	S3B1_MOUSE Splicing factor 3B subunit 1 (Spliceosome a...	44	0.011
sp	O75533	S3B1_HUMAN Splicing factor 3B subunit 1 (Spliceosome a...	44	0.011
tr	Q9VPR5	CG2807 protein [CG2807] [Drosophila melanogaster (Frui...	44	0.011
tr	Q8SQL2	Protein phosphatase PP2-A regulatory subunit A [ECU09_...	43	0.014
tr	O59191	Hypothetical protein PH1522 [PH1522] [Pyrococcus horik...	43	0.014
tr	Q8II56	PF16 protein, putative [PF11_0318] [Plasmodium falcipa...	42	0.031
tr	Q9V0Z1	Hypothetical protein PYRAB06470 [PYRAB06470] [Pyrococc...	42	0.031
tr	Q7T160	SI:dZ146N9.1 (Novel protein similar to human general c...	42	0.041
tr	Q7PY15	EbiP8519 (Fragment) [EBIG8519] [Anopheles gambiae str....	42	0.041
tr	Q86Y56	Hypothetical protein FLJ20397 [Homo sapiens (Human)]	42	0.041
sp	P22219	VP15_YEAST Protein kinase VPS15 (EC 2.7.1.-) [VPS15] [...	40	0.091

Psi-BLAST

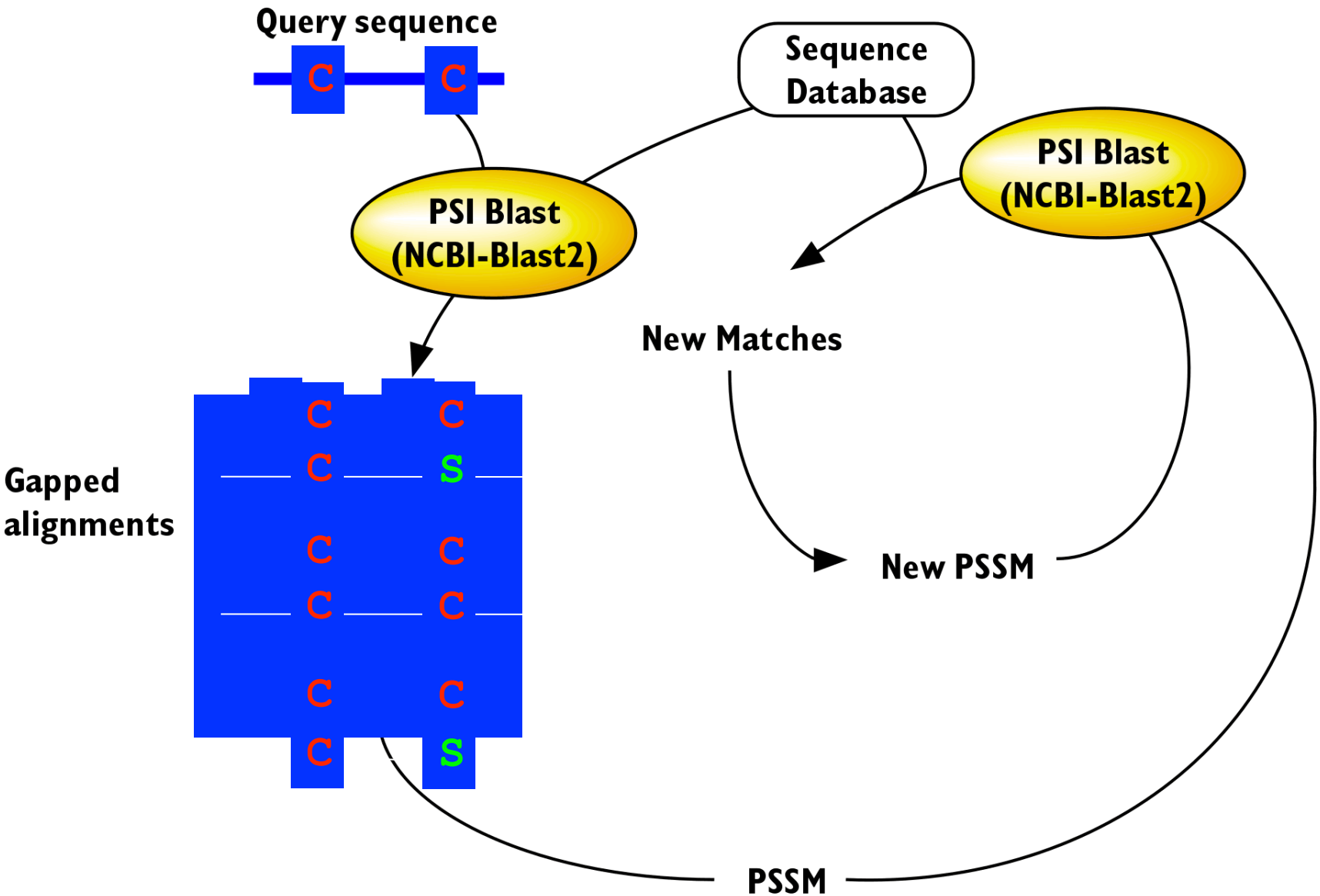
Position Specific Iterated version of BLAST.

Psi-BLAST performs a gapped BLAST database search.

Psi-BLAST program uses the information from any significant alignments returned to construct a position-specific score matrix, which replaces the query sequence for the next round of database searching.

Psi-BLAST may be iterated until no new significant alignments are found.

Psi-BLAST Iteration



BLAST: E-value

	M	Y	**C**	E	Q	U	E	N	**C**	E	S	.	.
A	0	2	-1	0	0	0	0	-1	0	-1	3		
S	-1	-1	-1	0	-1	0	0	0	**5**	-1	-1		
C	-1	-1	**10**	1	-1	0	0	5	**5**	4	-1		
.													
.													
Y	-1	6	-1	-1	-1	0	-1	-1	-1	-1	-1		
V	-1	1	-1	-1	-1	0	-1	-1	-1	1	-1		


NCBI
BLAST
PubMed Entrez BLAST OMIM Taxonomy Structure
Info
FAQs
News
References
NCBI Contributors
Education
Program selection guide
Tutorial
URL API guide
Download
Databases
Documentation
Executables
Source code
Support
Helpdesk
Mailing list
NEW 12 May 2004 BLAST 2.2.9 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)
PHI- and PSI-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 (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Chicken, cow, pig, dog, sheep, cat NEW
Environmental samples
Human, mouse, rat
Fugu rubripes, zebrafish
Insects, nematodes, plants, fungi, malaria
Microbial genomes, other eukaryotic genomes
Special
Search for gene expression data (GEO BLAST)
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
Meta
Retrieve results by RID
Get this page with javascript-free links
Disclaimer
Privacy statement
Accessibility
Valid XHTML 1.0, CSS.

Address http://www.ncbi.nlm.nih.gov/BLAST/

NCBI

BLAST

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

NEW **12 May 2004** BLAST 2.2.9 has been released. Read more...

Info

- FAQs
- News
- References
- NCBI Contributors

Education

- Program selection guide
- Tutorial
- URL API guide

Download

- Databases
- Documentation
- Executables
- Source code

Support

- Helpdesk
- Mailing list

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)
- PHI- and PSI-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 (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Chicken, cow, pig, dog, sheep, cat NEW
- Environmental samples
- Human, mouse, rat
- Fugu rubripes, zebrafish
- Insects, nematodes, plants, fungi, malaria
- Microbial genomes, other eukaryotic genomes

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)

Meta

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

Low Complexity Regions

Some genome sequences contain low-complexity regions. These can give false-positive hits.

Example:

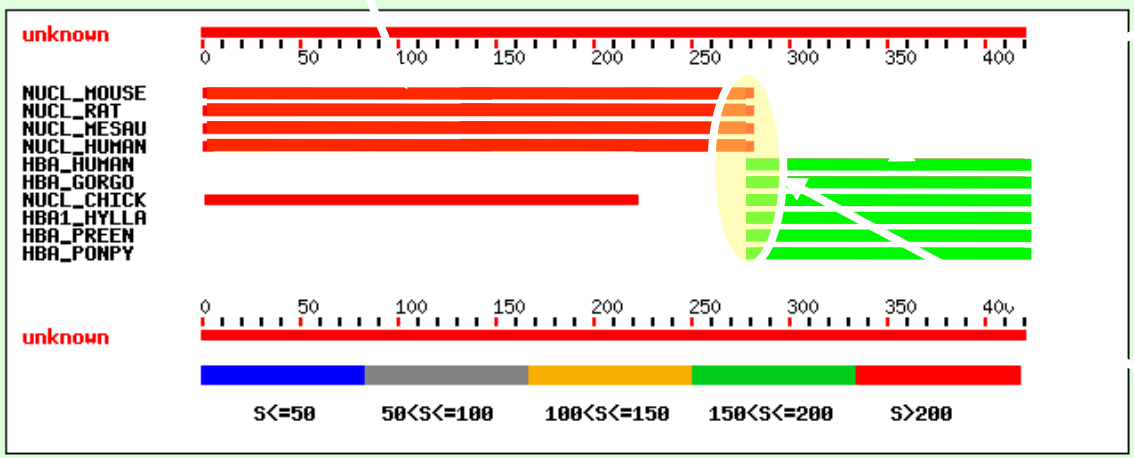
```
HSGDLPERTCPPCPPPCPPCPPPPCPPPCPCPPCPPPPLWQPSSERTD
         |-  low-complexity region  -|
```

Most sequence searching programs use filters to recognise and skip such low-complexity regions. If such regions are by chance included in the hit, the output looks like:

```
HSGDLPERTXXXXXXXXXXXXXXXXXXXXXXXXXXXXXLWQPSSERTD
         |-  low-complexity region  -|
```

Domain 1

Domain 2



No Overlap

Sequences producing significant alignments:	Score (bits)	E Value
sp|P09405|NUCL_MOUSE|89505EE39C89F832 (NCL..)Nucleolin (Protein ...	425	e-119
sp|P13383|NUCL_RAT|68774A214E550F90 (NCL..)Nucleolin (Protein C2...	407	e-113
sp|P08199|NUCL_MESAU|79DDCF724CED7DB4 (NCL)Nucleolin (Protein C2...	397	e-110
sp|P19338|NUCL_HUMAN|85A2F2CA22EA03DB (NCL)Nucleolin (Protein C2...	371	e-102
sp|P01922|HBA_HUMAN|34D13618E62A33C1 (HBA1..)Hemoglobin alpha ch...	285	2e-76
sp|P01923|HBA_GORGO|25D13618E72A3306 (HBA)Hemoglobin alpha chain...	283	4e-76
sp|P15771|NUCL_CHICK|7996C504BE9459A1 Nucleolin (Protein C23).[G...	283	5e-76
sp|Q9TS35|HBA1_HYLLA|25D13618E36A7706 (HBA1)Hemoglobin alpha-1 c...	281	2e-75
sp|P01924|HBA_PREEN|3771361D402A35C7 (HBA)Hemoglobin alpha chain...	280	5e-75
sp|P06635|HBA_PONPY|37DE74049545CE88 (HBA)Hemoglobin alpha chain...	279	7e-75

Tricky Problems

- Repeats
- Multi-domain proteins
- Low-complexity regions
- Redundancy
- Very short queries
- Very distant sequences
- Un-annotated sequences

Post Processing of Search Results

- Match identification numbers and key words
- Grep sequences out of the database
- Compare results of different searches (on different databases)
- Filter sequences on specified criteria
- Align results using a multiple alignment
- Recognise family patterns or generate a family profile

Typical Method Combinations

- Sequence search -> multiple alignment -> phylogenetic tree
- Sequence search -> sequence alignment -> homology modelling
- Sequence search -> sequence alignment -> mutation analysis
- Sequence search -> multiple alignment -> functional annotation

Learning Outcomes

- Principles of Sequence Searching
- Substitution Matrix and Scoring
- BLAST heuristics: k-word matching and local alignment
- Extreme value distribution
- True/false positive/negative hit
- P-value and E-value
- Interpretation of hit list
- BLAST for nucleotides and proteins
- PHI- and PSI-BLAST