Algorithms and Tools in Bioinformatics

Algorithms: Sequence Alignment
(adapted from Prof. Stephan Winkler)

Julia Vetter
julia.vetter@fh-hagenberg.at

SS2024

(4) Heuristic Methods

BLAST and FASTA

Heuristics

- heuristic methods
  - based on (simplifying) rules of thumb
  - do not necessarily produce an exact (optimal) result
  - but are fast and based on reasonable assumptions

- exact sequence comparison
  - dynamic programming: optimal alignment (relative to model / evaluation scheme)
  - runtime and space complexity: $O(n^2)$
  - too slow for DB search

- heuristic sequence comparison
  - FAST, BLAST...
  - linear runtime and memory requirements, i.e. $O(n)$
  - at least about 10-100 times faster than Smith-Waterman

Heuristic DB Search

- rule of thumb: almost all homologous sequences contain short partial sequences with a high degree of similarity
- goal: find those DB sequences with very highly rated, short local alignments (highly conserved sequence sections) – and find them fast!
- Calculate as few cells of the alignment matrix as possible
- Collect all of these high scoring segments
- Extend these sections into longer alignments

- The best known and most frequently used program for searching sequence databases is BLAST
- Altschul, Gish, Miller, Myers and Lipman [1990]: Journal of Molecular Biology
- Gapped BLAST and PSI-BLAST: Altschul, Madden, Schäffer, Zhang, Zhang, Miller and Lipman [1997]: Nucleic Acids Research
- BLAST is also based on the idea of hot spot search.

BLAST - Definition

- Let S be the database (a long sequence) and P the pattern.

- A segment pair consists of two equally long subsequences of S and P:

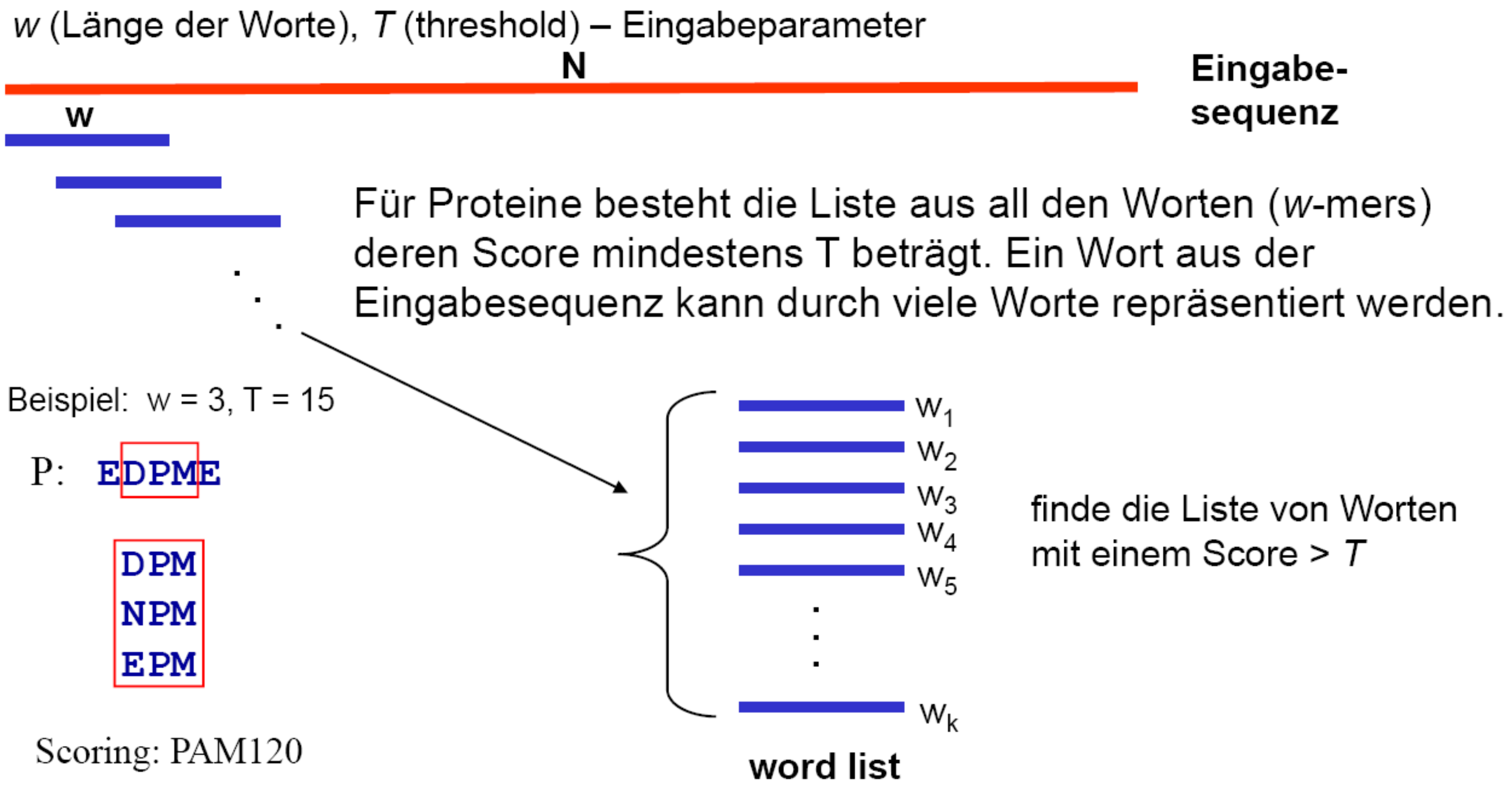
$$s_i s_{i+1} \dots s_{i+l}$$
$$p_j p_{j+1} \dots p_{j+l}$$

- A pair of segments that is maximal with respect to a given scoring function, i.e., the value of the pair does not increase as it is lengthened or truncated, is referred to as a maximal pair of segments.

- All segment pairs with a fixed predetermined length k are called word pairs.

BLAST

(1) Find all words of length k in the alphabet whose similarity to any word of length k is greater than a bound.



w (Länge der Worte), *T* (threshold) – Eingabeparameter

Eingabe-sequenz

Für Proteine besteht die Liste aus all den Worten (*w*-mers) deren Score mindestens T beträgt. Ein Wort aus der Eingabesequenz kann durch viele Worte repräsentiert werden.

Beispiel: w = 3, T = 15

P: EDPME

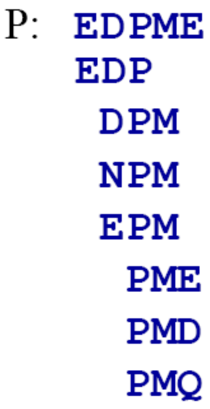
DPM
NPM
EPM

Scoring: PAM120

finde die Liste von Worten mit einem Score > *T*

word list

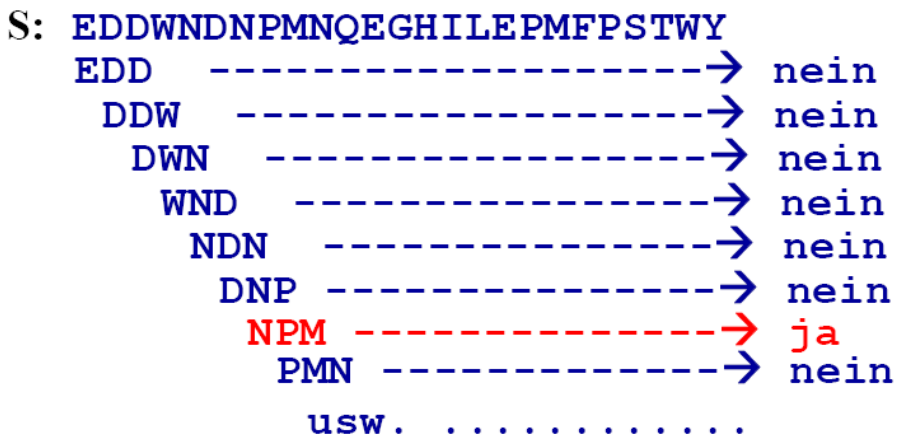
BLAST



```
P:  EDPME
    EDP
     DPM
     NPM
     EPM
      PME
      PMD
      PMQ
```

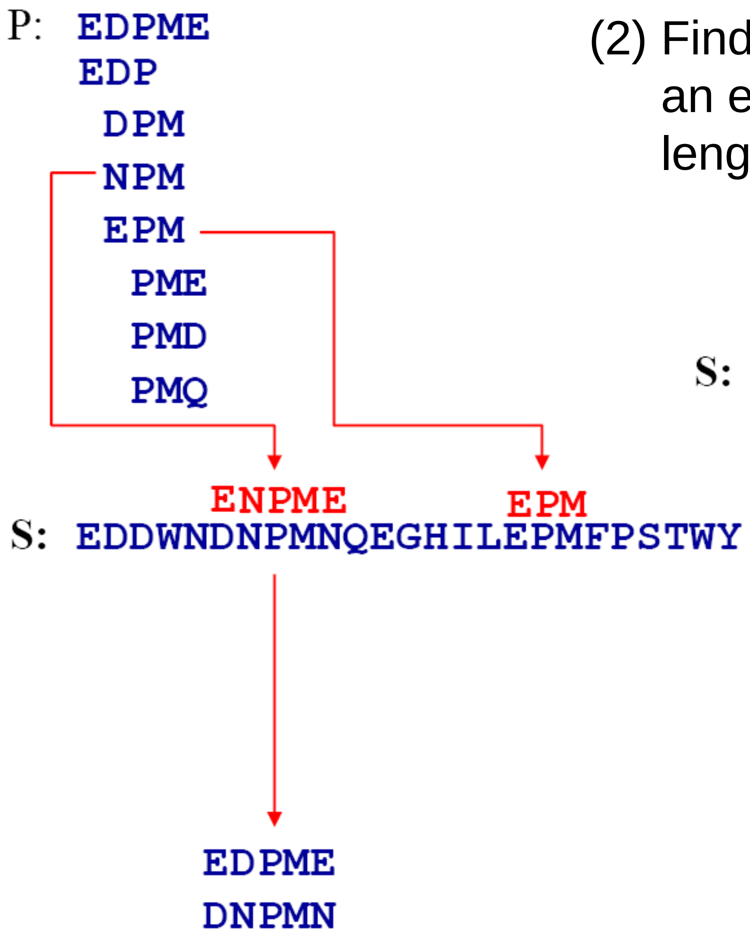
(2) How can one determine all occurrences of the word list in S? For each word of length k in S, test whether it belongs to P's word list using an efficient data structure.

BLAST uses a deterministic finite automaton for this (see Mealy [1955], Hopcroft & Ullman [1979]).



```
S:  EDDWNDNPMNQEGHILEPMFPSTWY
    EDD  ---------------------> nein
     DDW  --------------------> nein
      DWN  -------------------> nein
       WND  ------------------> nein
        NDN  -----------------> nein
         DNP -----------------> nein
          NPM ----------------> ja
           PMN ---------------> nein
             usw. ............
```

BLAST



(2) Find all occurrences of the word set in S. Using an efficient data structure, test for each word of length k in S whether it belongs to P's word list.

```
S: EDDWNDNPMNQEGHILEPMFPSTWY
   EDD  ---------------------→ nein
    DDW  --------------------→ nein
     DWN  -------------------→ nein
      WND  ------------------→ nein
       NDN  -----------------→ nein
        DNP -----------------→ nein
         NPM ----------------→ ja
          PMN ---------------→ nein
           usw. ............
```

BLAST

P: EDPME

EDP

DPM

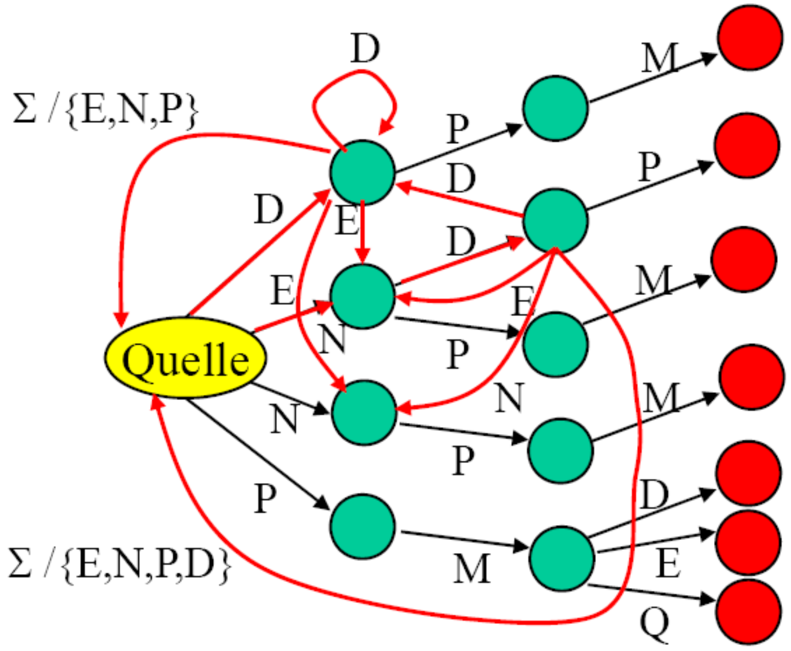
NPM

EPM

PME

PMD

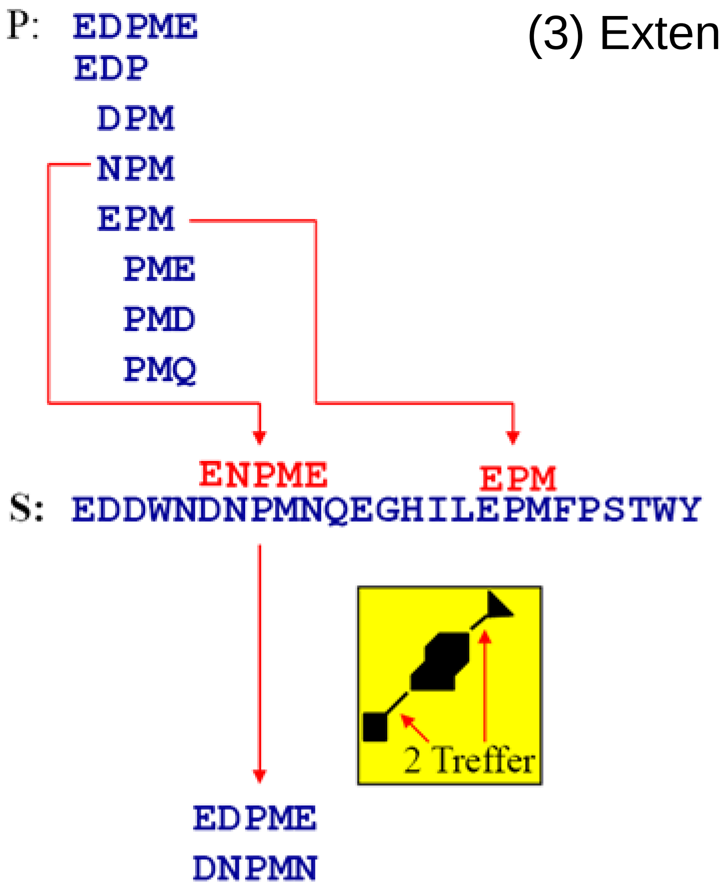
PMQ



Mealy Automat
(nicht vollständig)

BLAST

(3) Extend local hits to maximal segment pairs



P: EDPME

EDP

DPM

NPM

EPM

PME

PMD

PMQ

ENPME EPM

S: EDDWNDNPMNQEGHILEPMFPSTWY

EDPME

DNPMN

P values, E values

- p value (probability) – A. M. Lesk

– $P \leq 10^{-100}$	exact match
– P between 10^{-100} and 10^{-50}	almost identical sequences
– P between 10^{-50} and 10^{-10}	closely related sequences, homology sure
– P between 10^{-10} and 10^{-1}	distant relatives
– $P > 10^{-1}$	similarity probably not significant

- e value (expectancy)

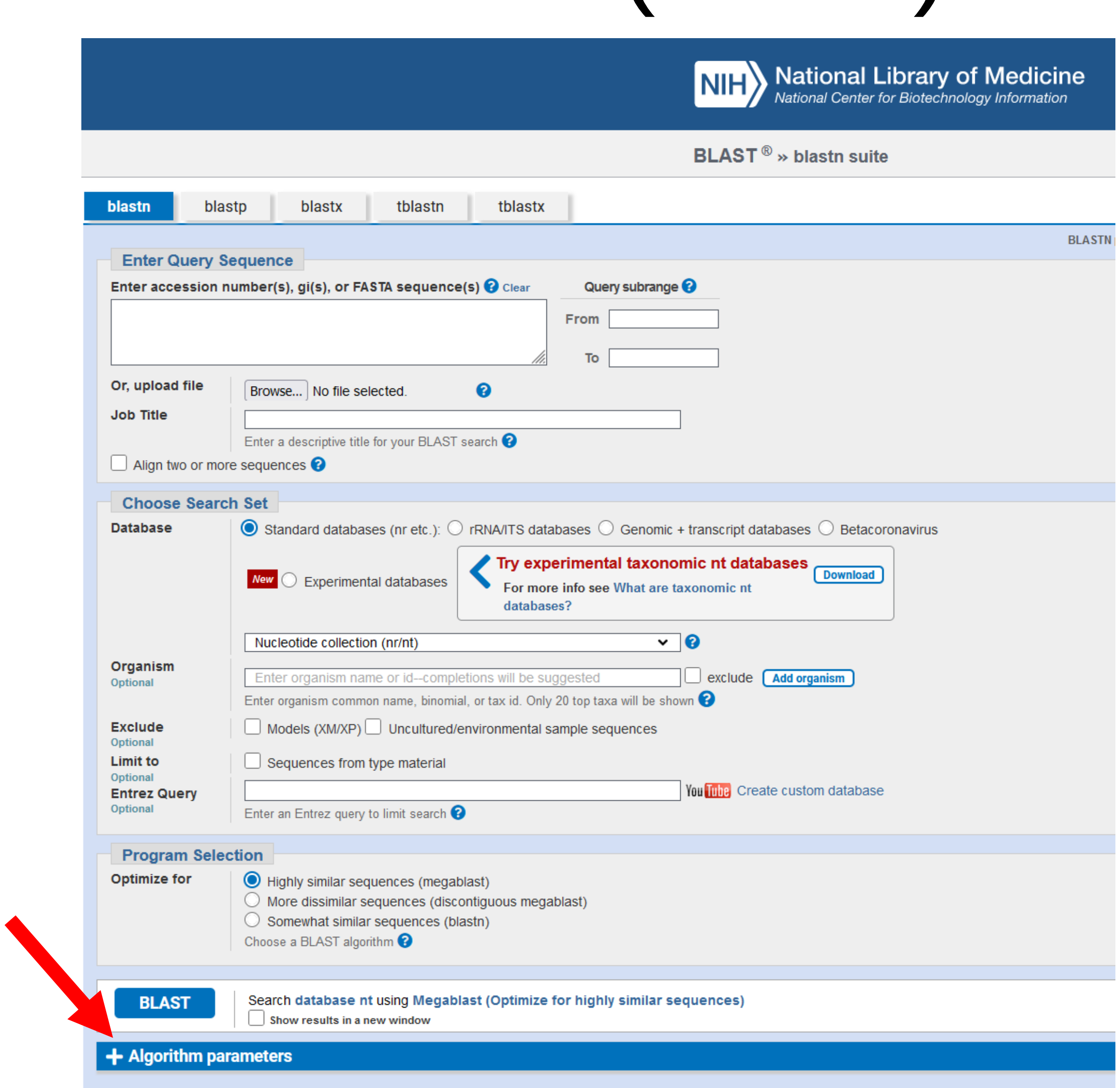
– E ≤ 0,02	sequences probably homologous
– E between 0,02 und 1	Homology cannot be ruled out
– E ≥ 1	good match might probably be random hit

- http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

BLAST Programmsuite

Program	**Database**	**Query**	**Typical uses**
BLASTN	Nucleotide	Nucleotide	Mapping oligonucleotides, cDNAs and PCR products to a genome, screening repetitive elements; cross-species sequence exploration; annotating genomic DNA; clustering sequencing reads
BLASTP	Protein	Protein	Identifying common regions between proteins; collecting related proteins for phylogenetic analyses
BLASTX	Protein	Nucleotide	Finding protein-coding genes in genomic DNA; determining translated into if a cDNA corresponds to a known protein protein
TBLASTN	Nucleotide translated into protein	Protein	Identifying transcripts, potentially from multiple organisms, similar to a given protein; mapping a protein to genomic DNA
TBLAST	Nucleotide translated into protein	Nucleotide translated into protein	Cross-species gene prediction at the genome or transcript level; searching for genes missed by traditional methods protein or not yet in protein database

BLAST (NCBI)



NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastn suite

blastn | blastp | blastx | tblastn | tblastx

BLASTN

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

Query subrange ?

From

To

Or, upload file Browse... No file selected. ?

Job Title

Enter a descriptive title for your BLAST search ?

□ Align two or more sequences ?

Choose Search Set

Database ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

New ○ Experimental databases

Try experimental taxonomic nt databases Download
For more info see What are taxonomic nt databases?

Nucleotide collection (nr/nt) ?

Organism
Optional

Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown ?

Exclude
Optional
□ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to
Optional
□ Sequences from type material

Entrez Query
Optional
YouTube Create custom database

Enter an Entrez query to limit search ?

Program Selection

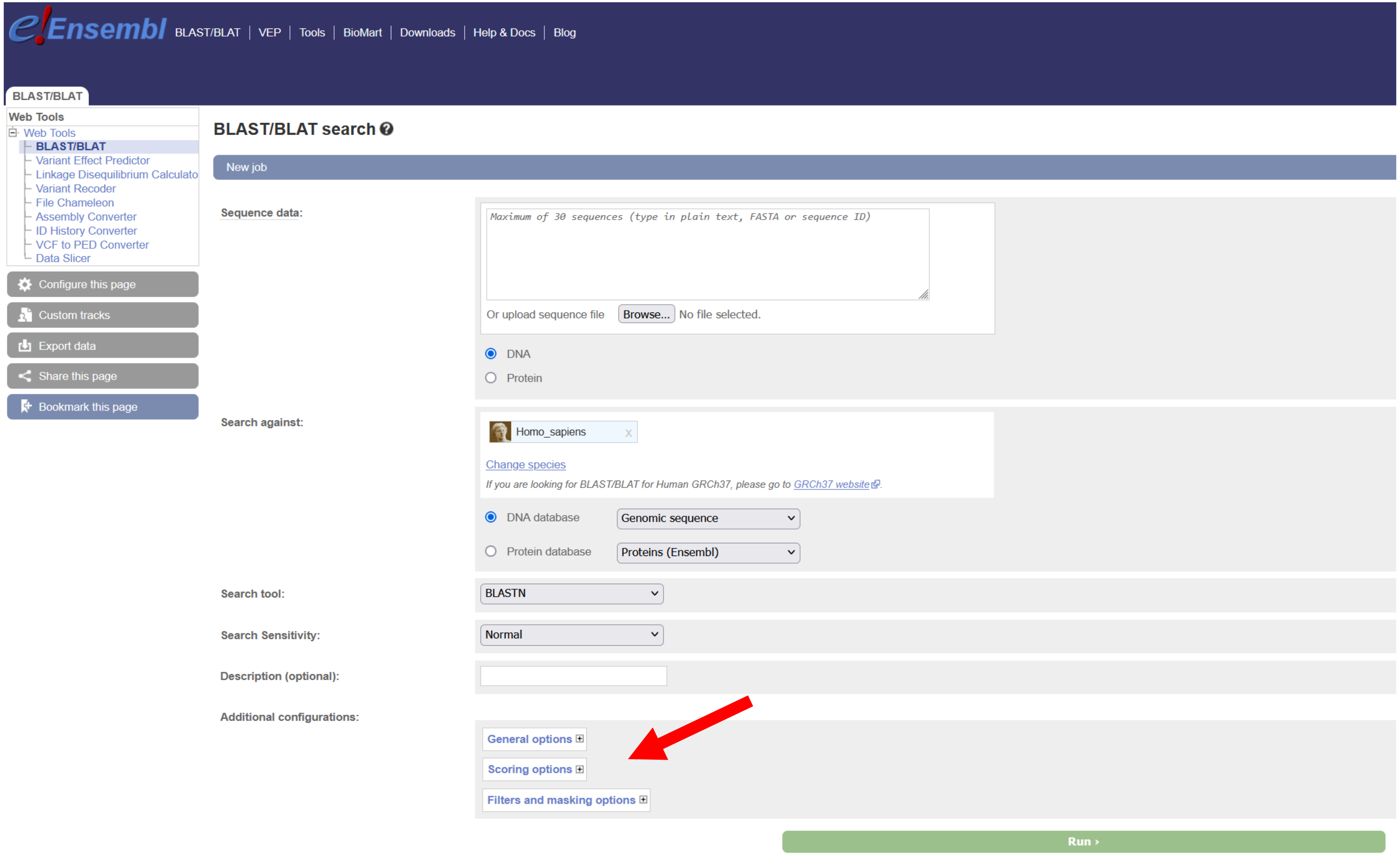
Optimize for
● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm ?

BLAST Search database nt using Megablast (Optimize for highly similar sequences)
□ Show results in a new window

+ Algorithm parameters

BLAST (EMBL-EBI Ensembl)



BLAST/BLAT search

Sequence data: Maximum of 30 sequences (type in plain text, FASTA or sequence ID)

Or upload sequence file No file selected.

● DNA

○ Protein

Search against: Homo_sapiens x

Change species

If you are looking for BLAST/BLAT for Human GRCh37, please go to GRCh37 website.

● DNA database: Genomic sequence

○ Protein database: Proteins (Ensembl)

Search tool: BLASTN

Search Sensitivity: Normal

Description (optional):

Additional configurations:

- General options ⊞
- Scoring options ⊞
- Filters and masking options ⊞

FASTA

- FASTA (Fast All)
  - by Pearson & Lipman (1985/88), Department of Biochemistry, University of Virginia

- 4 phases
  1. simple index search (indices = short exact match sequences)
  2. 'rough' evaluation of locally optimal sections
  3. connect sections to larger regions
  4. calculation of a local optimal narrow stripe alignment around the best regions

FASTA Phase 1: Index Search

- Separation into (overlapping) “words” of fixed length
  - word length is called ktup parameter; ktup for k-tuple: protein 1-3, DNA 4-6
- Example: Sequence R K T U R K (word length 2)
  - 1st word R K
  - 2nd word K T
  - 3rd word T U etc.
- all positions of a word in table (lookup-table)
  - (Example: Word RK at sequence positions 1 and 5, ...)
- Compare identical words (hot spots) from query and DB sequences quickly using hashing or lookup tables (sorted array of all ktup)
- initial score between query and DB sequence: number of hot spots within a narrow region

FASTA Phase 1: Index Search

- Query Sequenz: **FLWRTWS**
- DB-Sequenz: **SWKTWT**

Aminosäure	F	L	W	R	T	S
Index	1	2	3,6	4	5	7

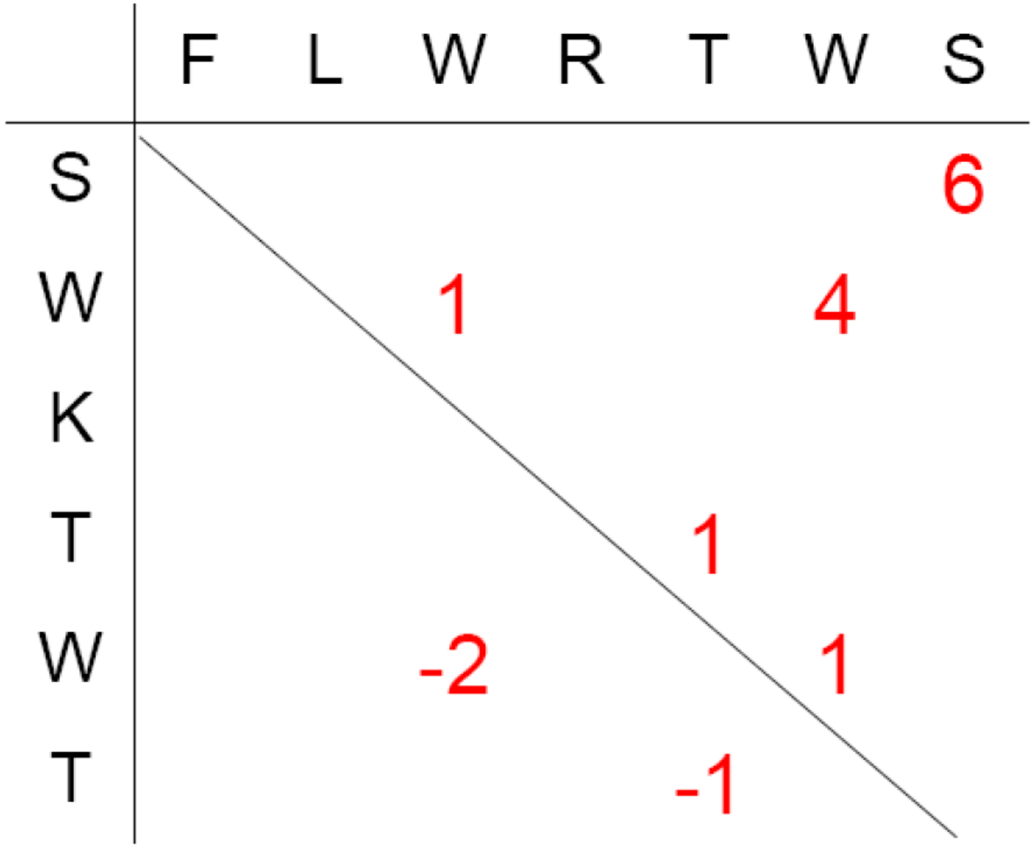
Aminosäure	S	W	K	T	W	T
Position	1	2	3	4	5	6

Hot-spots und deren relative Lage (Query zu DB-Seq.):

A.säure, Position	S, 1	W, 2		K, 3	T, 4	W, 5		T, 6
Abstand	7-1=6	3-2=1	6-2=4	-	5-4=1	3-5=-2	6-5=1	5-6=-1

→ Tabelle ‚entspricht' Dotplot

FASTA Phase 1: Index Search (Diagonals on a Dotplot)

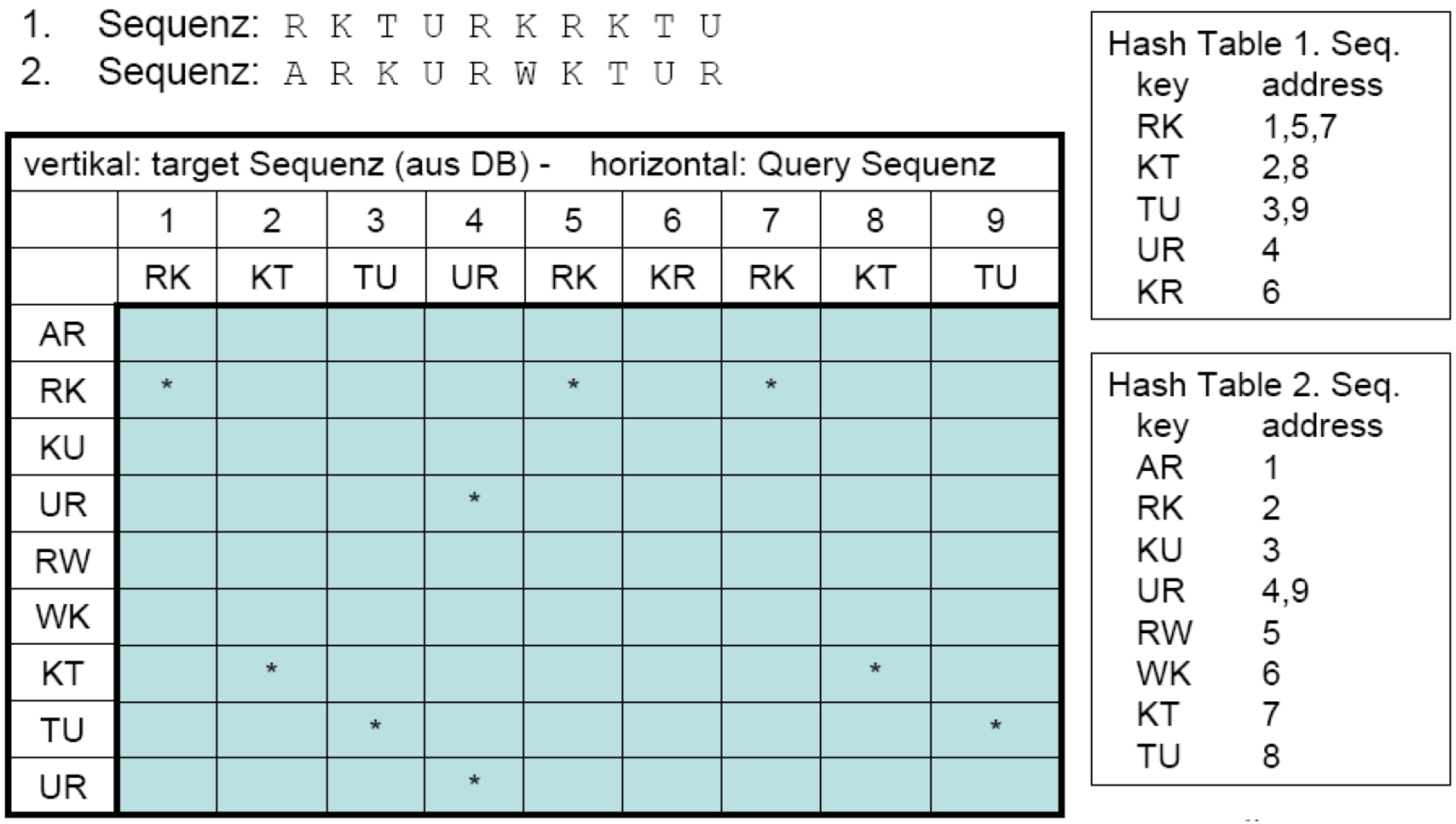


	F	L	W	R	T	W	S
S							6
W			1			4	
K							
T					1		
W			-2			1	
T					-1		

hot spots mit
gleicher Differenz
der Positionen:
auf einer Diagonalen

Differenz nennt man
Offset

FASTA Phase 1: Index Search
Location of all k-tuple matches



1. Sequenz: R K T U R K R K T U
2. Sequenz: A R K U R W K T U R

vertikal: target Sequenz (aus DB) -   horizontal: Query Sequenz

	1	2	3	4	5	6	7	8	9
	RK	KT	TU	UR	RK	KR	RK	KT	TU
AR									
RK	*				*		*		
KU									
UR				*					
RW									
WK									
KT		*						*	
TU			*						*
UR				*					

Hash Table 1. Seq.

key	address
RK	1,5,7
KT	2,8
TU	3,9
UR	4
KR	6

Hash Table 2. Seq.

key	address
AR	1
RK	2
KU	3
UR	4,9
RW	5
WK	6
KT	7
TU	8

FASTA Phase 1: Index Search Diagonals

- Sort hot spots by diagonals

- diagonal sequence = consecutive hot spots

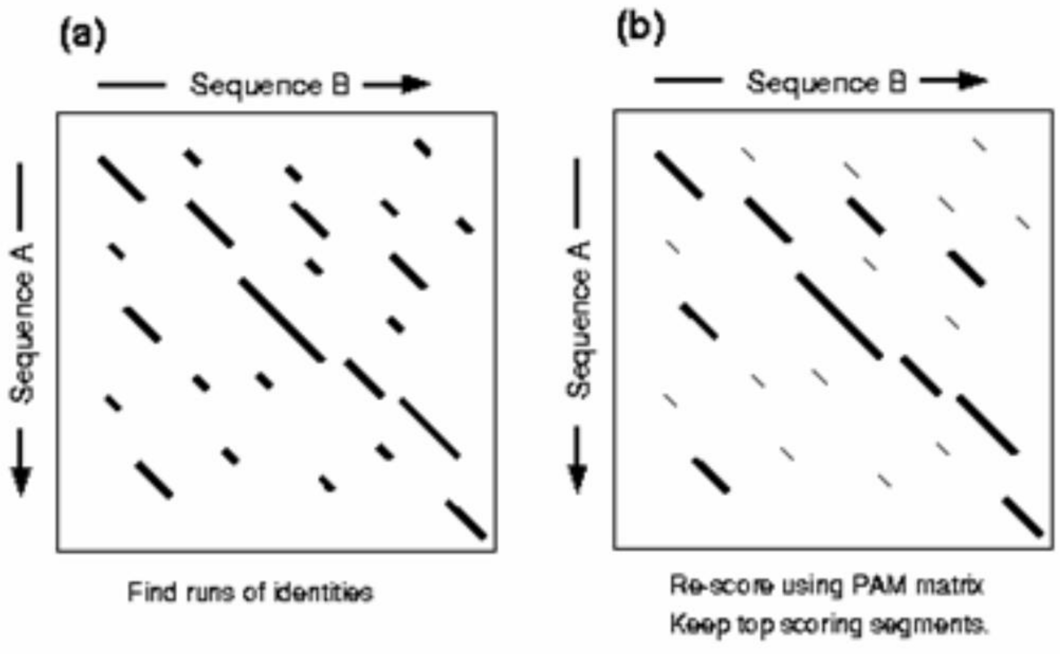
- Evaluation of a diagonal sequence: Sum of positive score according to the number of hot spots and negative score: number and length of 'inter-spot' areas; the longer these areas, the higher the score

- Determine ten best diagonal sequences
  (gap-free sections of potentially high-scoring alignments)

- Complexity: O(#hot-spots) << O(n*m)

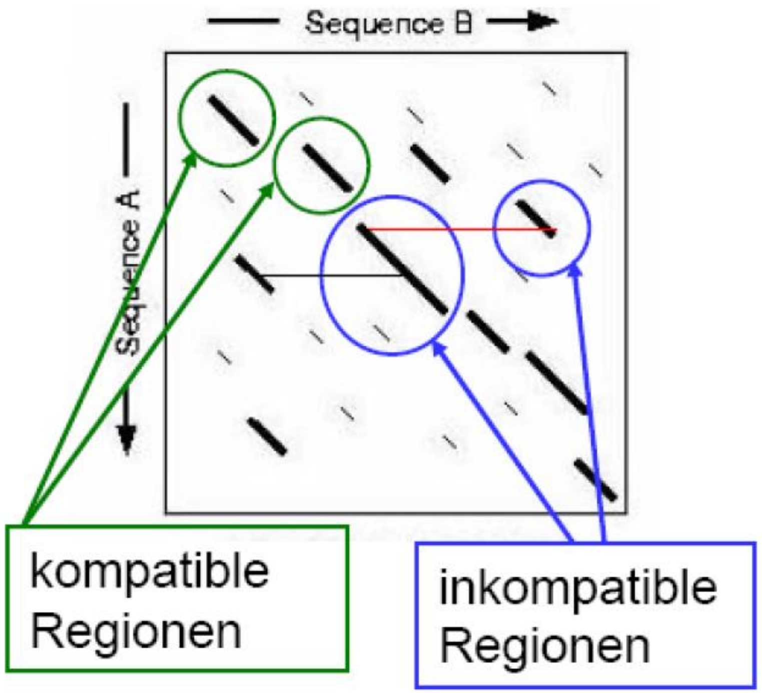
FASTA Phase 2: init1 Score

- Re-evaluation of all diagonal sequences
- all matches and mismatches according to PAM or BLOSUM
- associated sections of the diagonals are called initial regions
- init1 Score: best so received re-evaluation



FASTA Phase 3: init*n* Score

- only consider initial regions with score > cutoff (parameter)
- Sequence of regions is compatible if all related parts of a sequence do not overlap
- distance between regions < join (parameter, e.g. 36)
- Scoring a compatible episode
  - positive: sum of the scores of the initial regions
  - negative: relative position (distance of the regions) of the initial regions according to a gap penalty
- initn Score: maximum score of a compatible sequence



FASTA Phase 4: opt Score

- If initn score is sufficiently large, calculation of opt score (optimal local alignment score)
- Restriction to narrow, diagonal stripes

- fixed width around init1 region
- Perform Smith-Waterman inside this strip

- Determining the stripe width (parameter)
- Heuristic: the stronger the identities, the less likely an optimal alignment path is far away from the init1 diagonal (i.e. contains a lot of gaps)
  - ktup=2: 16 diagonals
  - ktup=1: 32 diagonals

- opt score (formerly initn score) as the basis for ranking the DB sequences (ranking)

FASTA Program Suite

- FASTA
  - Query Protein vs Protein DB
  - Query DNA vs DNA DB
- FASTX
  - Query DNA vs Protein DB
- TFASTX
  - Query Protein vs DNA DB
- FASTS
  - Query Protein (MALDI analyses) vs Protein DB

FASTA Weaknesses

1. Example: Two protein sequences:
   - ABABABABAB
   - ACACACACAC
   - 50% identity, but with ktup=2 no hot-spots

2. The narrow band of e.g. 32 residues in phase 4 can be too narrow: two proteins can be identical except for a gap of length >32 in the middle of one of the sequences. In phase 4, only half of the identity would be found.

3. FASTA only considers perfect matches but not conserved substitutions in proteins. As a result, sequences that are functionally homologous but have little identity cannot be found.