Searching in Sequence Databases

- Keyword Searching → Database entries,
  that contain a keyword specified.
  (Medline, SRS, ….)

- Sequence Searching → Similar sequences,
  and their alignment with the
  query sequence.
  (FastA, Blast, ….)

HUSAR

Search strategies:

Searching in Sequence Databases

- The comparison of a **sequence** against a database ….

  In order to find similar sequences

- The comparison of a **pattern,** a **profile**, or a **HMM** against a database ….

  In order to find **distantly related** sequences.

Search algorithms:

Searching in Sequence Databases

- Sensitive searching using **rigorous** searching algorithms:

  - Smith - Waterman search, accurate comparison ⟶ compute time !
    ⟶ special hardware (Bioccelerator)

- Extremely fast searching using **heuristic** searching algorithms:

  - Approximations in comparison, insertions and deletions are widely ignored;
    ⟶ reduced compute time!

HUSAR

Heuristic Methods

- FastA *(Pearson and Lipman)*
- Blast / Blast2 *(Altschul)*

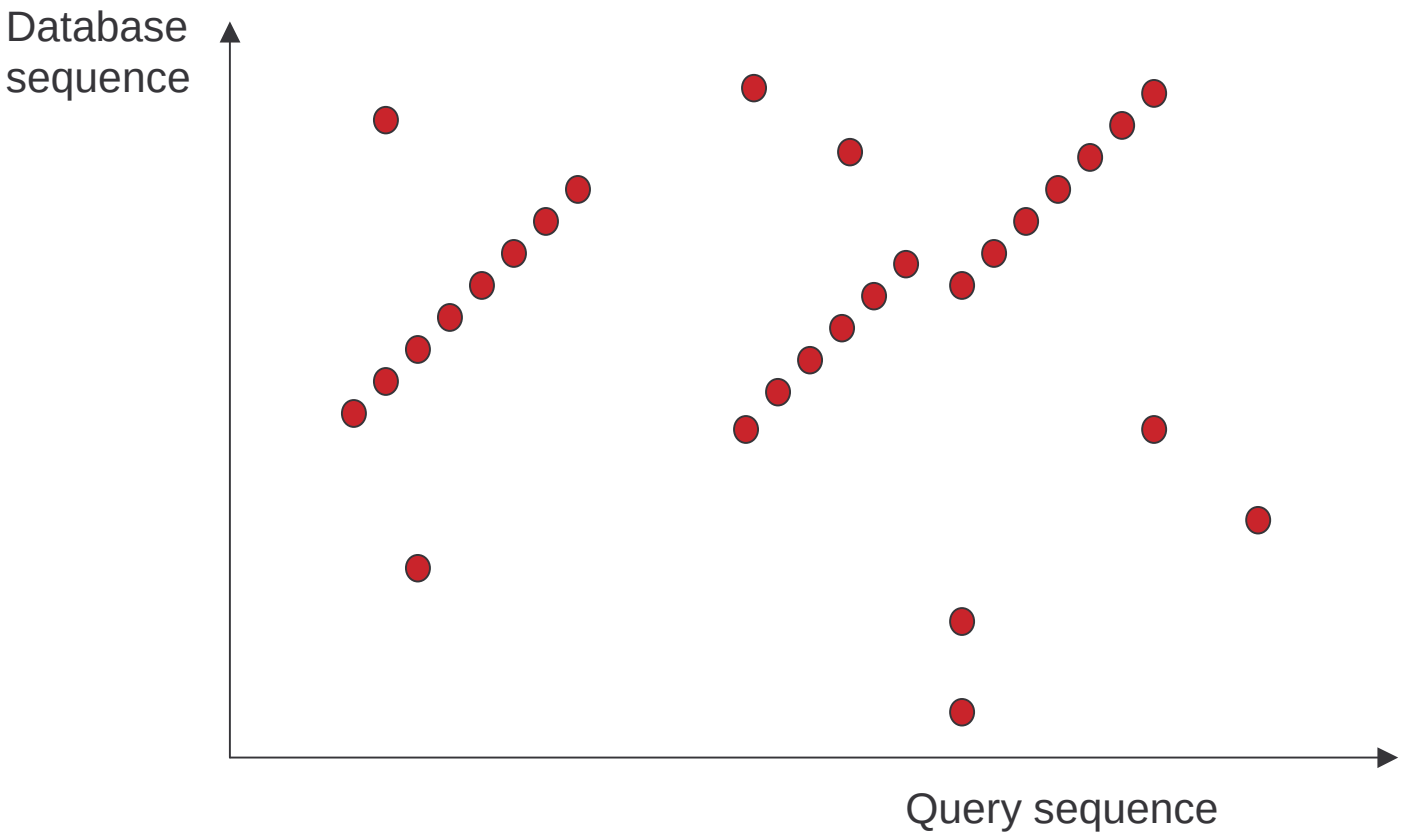
FastA *(Pearson and Lipman)*

1. Search for short identities of a fixed length (word size).
2. The score of each diagonal is determined.
3. The ten highest scoring regions (diagonals) are rescored using scoring tables (initial regions).
   - Highest score is called *init1*.
4. Adjacent initial diagonals are connected.
   - Highest score is called *initn*.
5. For sequences with an initn score greater than a given threshold an *opt-score*, a *z-score*, and an *E() value* are calculated.
6. Sequences with an E() value smaller than a given threshold are listed

Example:

Step 1

Search for short identities of a fixed length



Word size:

DNA: 6

Protein: 2

HUSAR

FastA *(Pearson and Lipman)*

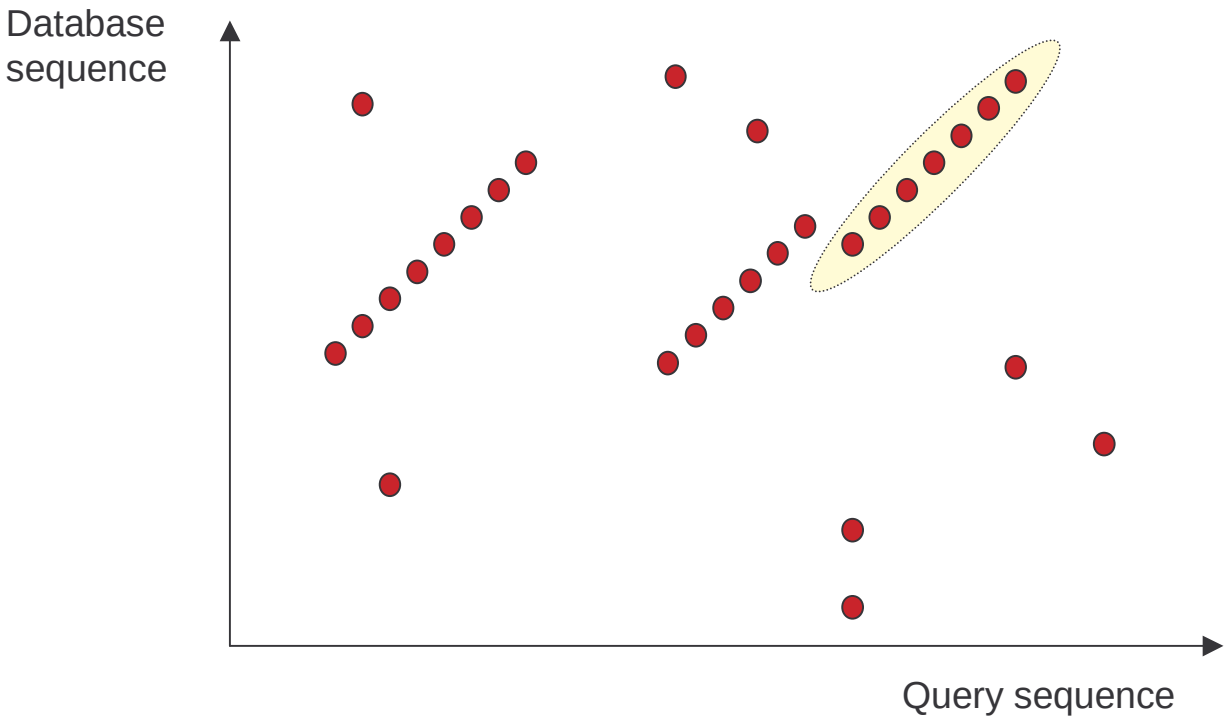
1. Search for short identities of a fixed length (word size).
2. The score of each diagonal is determined.
3. The ten highest scoring regions (diagonals) are rescored using scoring tables (initial regions).
   - Highest score is called *init1*.
4. Adjacent initial diagonals are connected.
   - Highest score is called *initn*.
5. For sequences with an initn score greater than a given threshold an *opt-score*, a *z-score*, and an *E() value* are calculated.
6. Sequences with an E() value smaller than a given threshold are listed

HUSAR

Example:

Step 2

Scoring of diagonals



DNA:

Match: 5

Mismatch: - 4

Protein:

Scoring Table

Score = 60

HUSAR

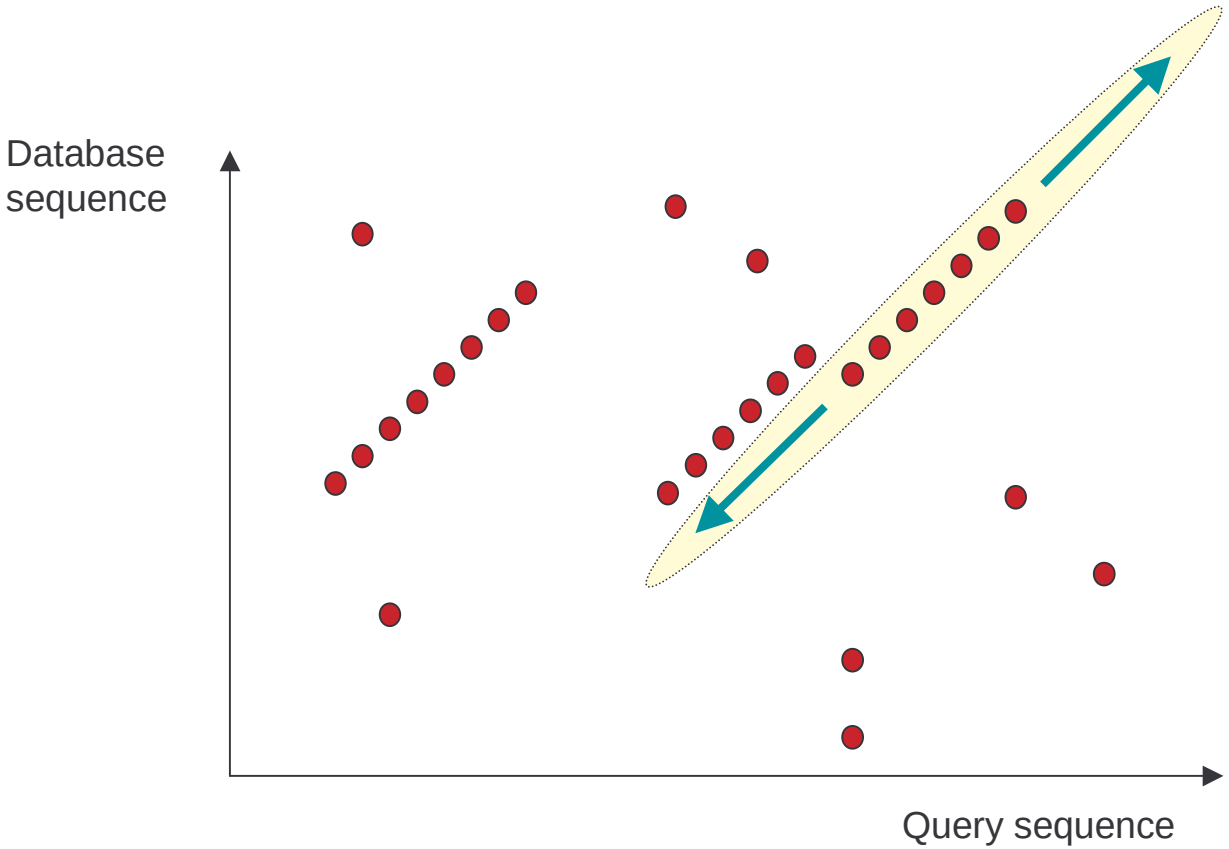
FastA *(Pearson and Lipman)*

1. Search for short identities of a fixed length (word size).
2. The score of each diagonal is determined.
3. The ten highest scoring regions (diagonals) are rescored using scoring tables (initial regions).
   - Highest score is called *init1*.
4. Adjacent initial diagonals are connected.
   - Highest score is called *initn*.
5. For sequences with an initn score greater than a given threshold an *opt-score*, a *z-score*, and an *E() value* are calculated.
6. Sequences with an E() value smaller than a given threshold are listed

Example:

Step 3

Scoring of diagonals



DNA:

Match: 5

Mismatch: - 4

Protein:

Scoring Table

Score > 60 (**INIT1**)

HUSAR

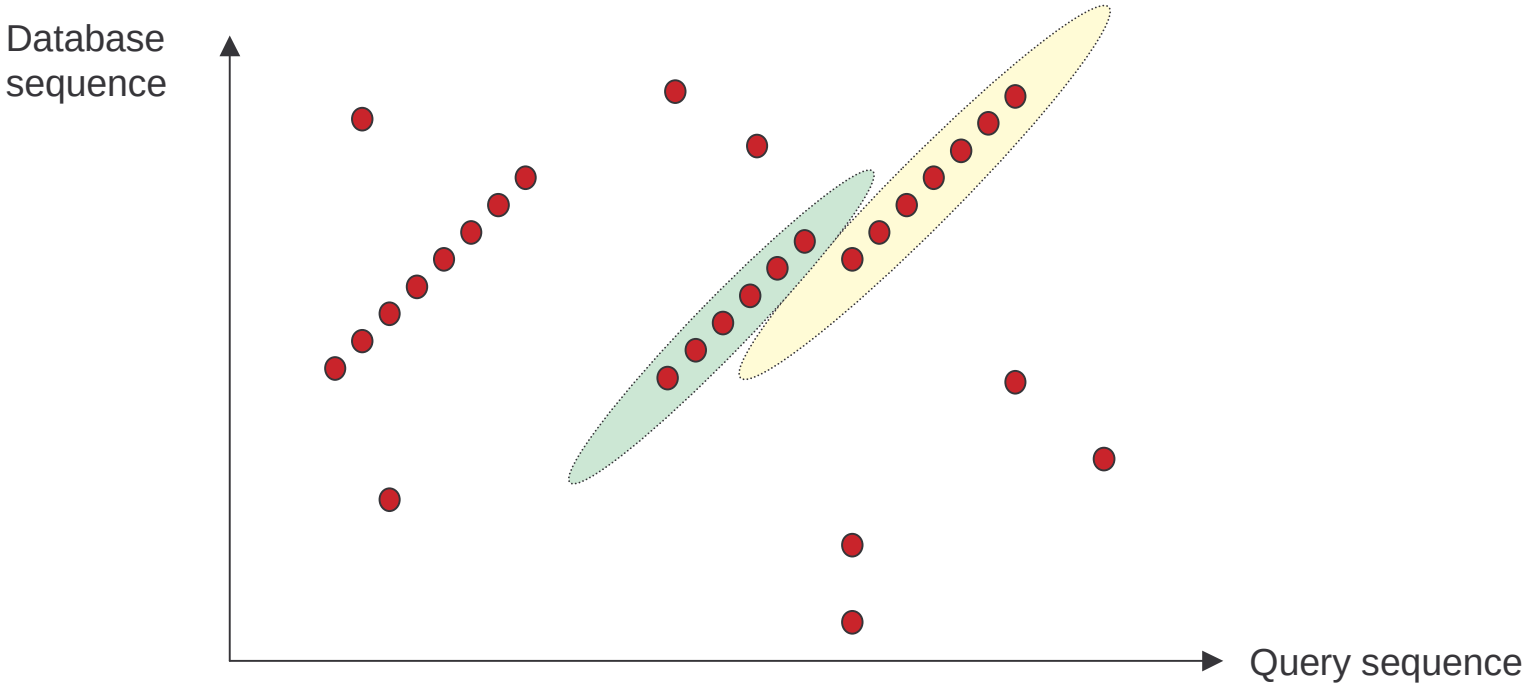
FastA *(Pearson and Lipman)*

1. Search for short identities of a fixed length (word size).
2. The score of each diagonal is determined.
3. The ten highest scoring regions (diagonals) are rescored using scoring tables (initial regions).
   - Highest score is called *init1*.
4. Adjacent initial diagonals are connected.
   - Highest score is called *initn*.
5. For sequences with an initn score greater than a given threshold an *opt-score*, a *z-score*, and an *E() value* are calculated.
6. Sequences with an E() value smaller than a given threshold are listed

Example:

Step 4

Connection of neighbouring diagonals



INITN = green score + yellow score - “joining penalty”

HUSAR

FastA *(Pearson and Lipman)*

1. Search for short identities of a fixed length (word size).
2. The score of each diagonal is determined.
3. The ten highest scoring regions (diagonals) are rescored using scoring tables (initial regions).
   - Highest score is called *init1*.
4. Adjacent initial diagonals are connected.
   - Highest score is called *initn*.
5. For sequences with an initn score greater than a given threshold an *opt-score*, a *z-score*, and an *E() value* are calculated.
6. Sequences with an E() value smaller than a given threshold are listed

Step 5

Score calculation

Opt-score:	Smith-Waterman score
Z-score:	normalization with respect to data base sequence length
E() value	expectation value of a score

FastA *(Pearson and Lipman)*

1. Search for short identities of a fixed length (word size).
2. The score of each diagonal is determined.
3. The ten highest scoring regions (diagonals) are rescored using scoring tables (initial regions).
   - Highest score is called *init1*.
4. Adjacent initial diagonals are connected.
   - Highest score is called *initn*.
5. For sequences with an initn score greater than a given threshold an *opt-score*, a *z-score*, and an *E() value* are calculated.
6. Sequences with an E() value smaller than a given threshold are listed

Example:

FastA

FastA output: part I

```
 Results sorted and z-values calculated from opt score
 1770 scores saved that exceeded 107
 4614416 optimizations performed
 Joining threshold: 47, optimization threshold: 32, opt. width: 16

The best scores are:                      init1 initn  opt  z-sc E(5219455)

EMORG:CHPHET01    Begin: 1  End:  162
! M37322 P.hybrida chloroplast rpS19   810   810   810   614.0   5e-25
EMORG:CHPHETIR    Begin: 31  End:  183  Strand: -
! M35955 P.hybrida chloroplast rps19'  410   410   699   531.8  1.7e-20
EMORG:SNCPJLB    Begin: 2  End:  150
! Z71250 S.nigrum chloroplast JLB reg  457   457   659   499.2  6.8e-19
EMORG:NPCPJLB    Begin: 2  End:  151
! Z71235 N.palmeri chloroplast JLB re  642   642   659   501.5   7e-19
EMORG:NBCPJLB    Begin: 2  End:  158
! Z71226 N.bigelovii chloroplast JLB   472   472   644   485.5  2.7e-18
EMORG:STCPJLB    Begin: 2  End:  149
! Z71248 S.tuberosum chloroplast JLB   452   452   641   485.4  3.7e-17
```

Example:

FastA

FastA output: part II

```
SCORES      Init1:   472  Initn:   472  Opt:   644 z-score: 485.5 E(): 2.7e-18
  91.7% identity in 157 bp overlap

                      10        20        30        40          50       59
chphet01.em_ TCTTCGTCGCCATAGTAAATAGGAGAGGAAATCGAATTTAATTCTTCG-TTTTACAAAAA
                      |||||||||||||||| |||| ||||| ||||||||| ||||||||||
NBCPJLB               GTAGTAAATAGGAGAGAAAATTGAATTAAATTCTTCGTTTTTACAAAAA
                                10        20        30        40

            60              70        80        90       100       110
chphet01.em_ CTT------ACAAAAAAAAAAAAAAATAGGAGTAAGCTTGTGACACGTTCACTAAAAAAAA
             |||      | ||||||||||||||||||||||||||||||||||||||||||||||||
NBCPJLB      CTTTACAAAAAAAAAAAAAAAAAAATAGGAGTAAGCTTGTGACACGTTCACTAAAAAAAA
            50        60        70        80        90       100

                 120       130       140       150       160
chphet01.em_ ACCCCTTTGTAGCCAATCATTTATTAAATAAAATTGATAAGCTTAACAC
             | |||||||||||||||||||||||| ||||||||||||||||||
NBCPJLB      ATCCCTTTGTAGCCAATCATTTATTAAAAAAAATTGATAAGCTTAACACAAAAGCAGAAA
           110       120       130       140       150       160
```

FastA programs:

FastA — searches for similarity between a query sequence and any group of sequences (DNA and Protein).

TFastA — compares a peptide sequence against a set of nucleotid sequences.

FastX — compares a nucleotide sequence against a protein database taking frameshifts into account.

TFastX — compares a peptide sequence against a nucleotide sequence database taking frameshifts into account.

Blast

(“Basic Local Alignment Search Tool”) (*S. Altschul*)

1. Search for short regions of a given length that score at or above a certain threshold (initial hits).
2. Initial hits are extended until score drops off by a certain amount from its maximum.
   - The region with the maximum score is called “ high scoring segment pair “ (HSP).
3. All HSPs with a score as high or higher than a cutoff-score are displayed. This cutoff score corresponds to a predefined expectation value.

Blast - essential parameters -

	Word size	Threshold
DNA	11	-
Protein	3	11 (blosum62)

Example:

Step1

BlastP

Initial hit

A	W	T	P	S
H	W	T	C	S
-2	11	5	-5	4

14

Word size: 3

Threshold: 11

Extension; 22

BLOSUM 62

HUSAR

Blast

(“Basic Local Alignment Search Tool”) (*S. Altschul*)

1. Search for short regions of a given length that score at or above a certain threshold (initial hits).
2. Initial hits are extended until score drops off by a certain amount from its maximum.
   - The region with the maximum score is called “ high scoring segment pair “ (HSP).
3. All HSPs with a score as high or higher than a cutoff-score are displayed. This cutoff score corresponds to a predefined expectation value.

Example:

Step 2

BlastP

Initial hit

A	W	T	P	S
H	W	T	C	S
-2	11	5	-5	4

14

Word size: 3

Threshold: 11

Extension; 22

BLOSUM 62

HUSAR

Example:

Step 2

BlastP

Initial hit

A	W	T	P	S
H	W	T	C	S
-2	11	5	-5	4

14

9

Word size: 3

Threshold: 11

Extension: 22

BLOSUM 62

HUSAR

Example:

Step 2

BlastP

Initial hit

A	W	T	P	S
H	W	T	C	S
-2	11	5	-5	4

14

9

13

Word size: 3

Threshold: 11

Extension: 22

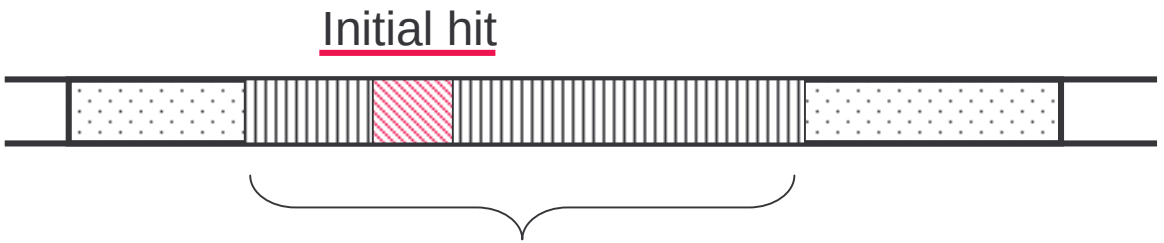
BLOSUM 62

HUSAR

Example:

Step 2

BlastP



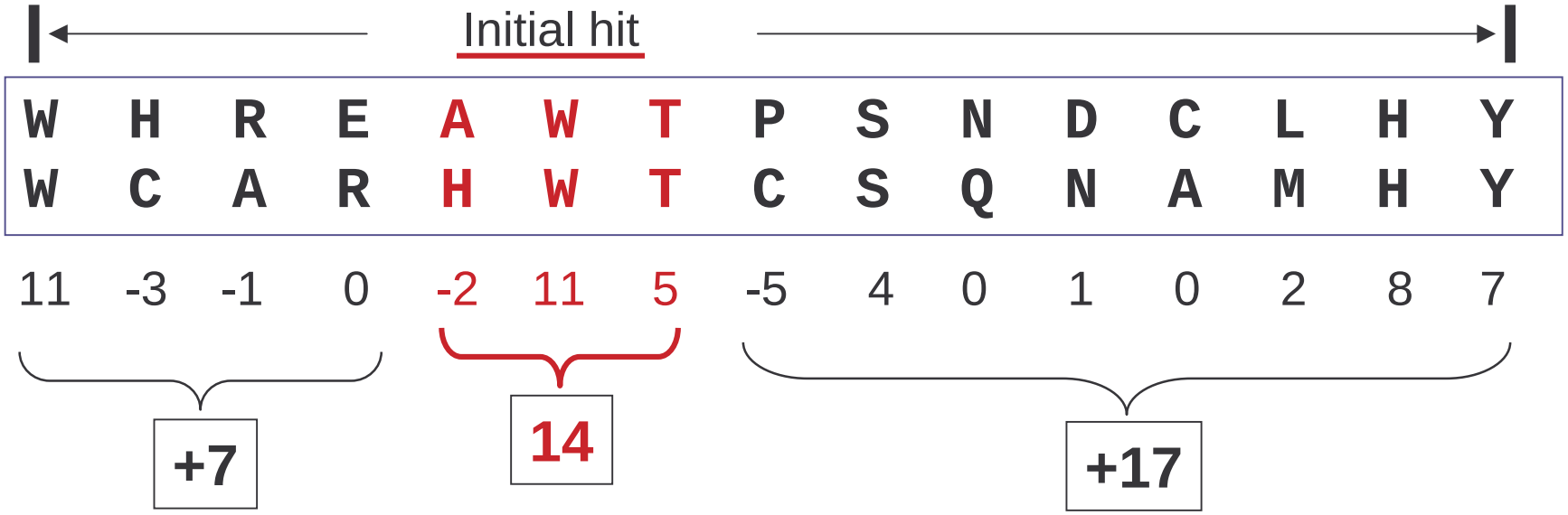
Stop extension, if score of total region less than maximum score minus extension parameter

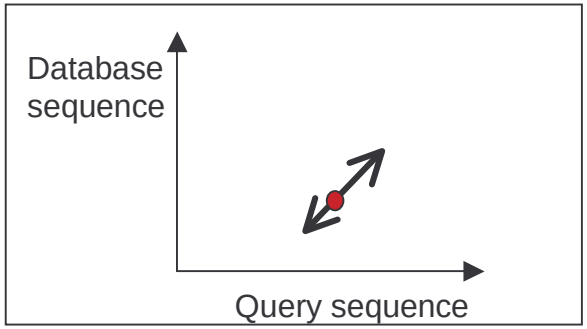
HUSAR

Example:

Step1 +2

Evaluation of final HSP





Score(HSP) = 38

HUSAR

Blast

(“Basic Local Alignment Search Tool”) (*S. Altschul*)

1. Search for short regions of a given length that score at or above a certain threshold (initial hits).
2. Initial hits are extended until score drops off by a certain amount from its maximum.
   - The region with the maximum score is called “ high scoring segment pair “ (HSP).
3. All HSPs with a score as high or higher than a cutoff-score are displayed. This cutoff score corresponds to a predefined expectation value.

Blast - statistics -

There is a mathematical theory that allows us to compute the probability of occurrence of a certain score when searching a random database of a given size.

→ One can compute an expectation value, which tells us how often a certain score would occur simply by chance.

Blast uses an expectation value of **10** as a threshold.

→ Blast displays all HSPs scoring as high or higher as the score corresponding to this expectation value.

HUSAR

Example:

BlastN output: part I

```
                                                                    Smallest Sum
                                                          High      Probability
Sequences producing High-scoring Segment Pairs:           Score     P(N)         N

>>>empri:HSCHIKER    M37818 Human keratin (psi-K-alpha) p...  1115  4.8e-252     5
>>>empri:HSKERUV     X05803 Human radiated keratinocyte m...   718  6.0e-106     2
>>>empri:HSEPKER     J00124 Homo sapiens 50 kDa type I ep...   687  2.4e-105     4
>>>empri:HSKERELP    X62571 H.sapiens mRNA for keratin-re...   718  3.1e-105     2
>>>emrod:MMKTEPIA    M13806 Mouse keratin (epidermal) typ...   691  5.3e-101     2
>>>emrod:MMKTEPIC    M13805 Mouse type I epidermal kerati...   700  1.1e-96      2
>>>empri:HSCYTOK17   Z19574 H.sapiens gene for cytokerati...   651  1.8e-90      3
>>>gb_pr_only:S79867 S79867 type I keratin 16 [human, epi...   709  1.5e-87      2
>>>empri:S72493      S72493 keratin=keratin 16 homolog [h...   693  1.3e-86      2
>>>empri:HSKERP2     M22928 Human keratin pseudogene, exo...   624  6.4e-86      3
                              ...
>>>emvrl:HESH1ULTR   L24958 Pseudorabies virus long termi...   125  0.99994      1
>>>emvrl:HESH1SEQ    M81222 Herpesvirus-pseudorabies viru...   125  0.99995      1
>>>emvrl:HEHSSLLT    M57505 Pseudorabies virus ORF1, ORF2...   125  0.99995      1
```

Example:

BlastN output: part II

```
 Score = 1115 (308.1 bits), Expect = 4.8e-252, Sum P(5) = 4.8e-252
 Identities = 223/223 (100%), Positives = 223/223 (100%), Strand = Plus / Plus

Query:   276 AGAAAGCATCGCTGGAGGGCAGCCTGGTGGAGACGGAGGTGTGTTACAGGACCCAGCTGG 335
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1325 AGAAAGCATCGCTGGAGGGCAGCCTGGTGGAGACGGAGGTGTGTTACAGGACCCAGCTGG 1384

                ...

Query:   456 AGATCGCCACCTACAGCCGCTTGCTAGAGGTTGAGGACGCCCG 498
             |||||||||||||||||||||||||||||||||||||||||||
Sbjct:  1505 AGATCGCCACCTACAGCCGCTTGCTAGAGGTTGAGGACGCCCG 1547


Score =  810 (223.8 bits), Expect = 4.8e-252, Sum P(5) = 4.8e-252
 Identities = 162/162 (100%), Positives = 162/162 (100%), Strand = Plus / Plus

Query:     1 GAAATGAACGCCCTTTGAGGTCAGGTGGACGAGGATGTCAGTGTGAAGATGGACACTGTG 60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:   195 GAAATGAACGCCCTTTGAGGTCAGGTGGACGAGGATGTCAGTGTGAAGATGGACACTGTG 254

                ...
```

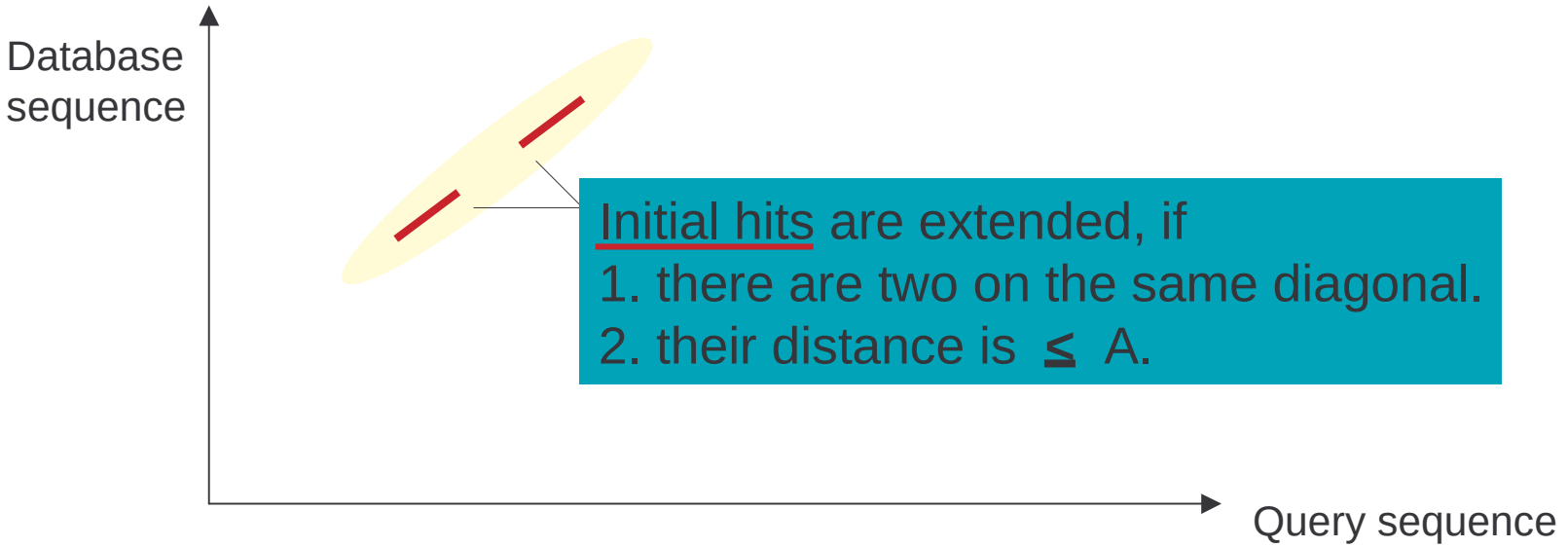
Blast2

Advantages of Blast2 generation:

- Speed
- Gapped alignments

Blast2 - extension -

Replacement of single-hit by two-hit method:



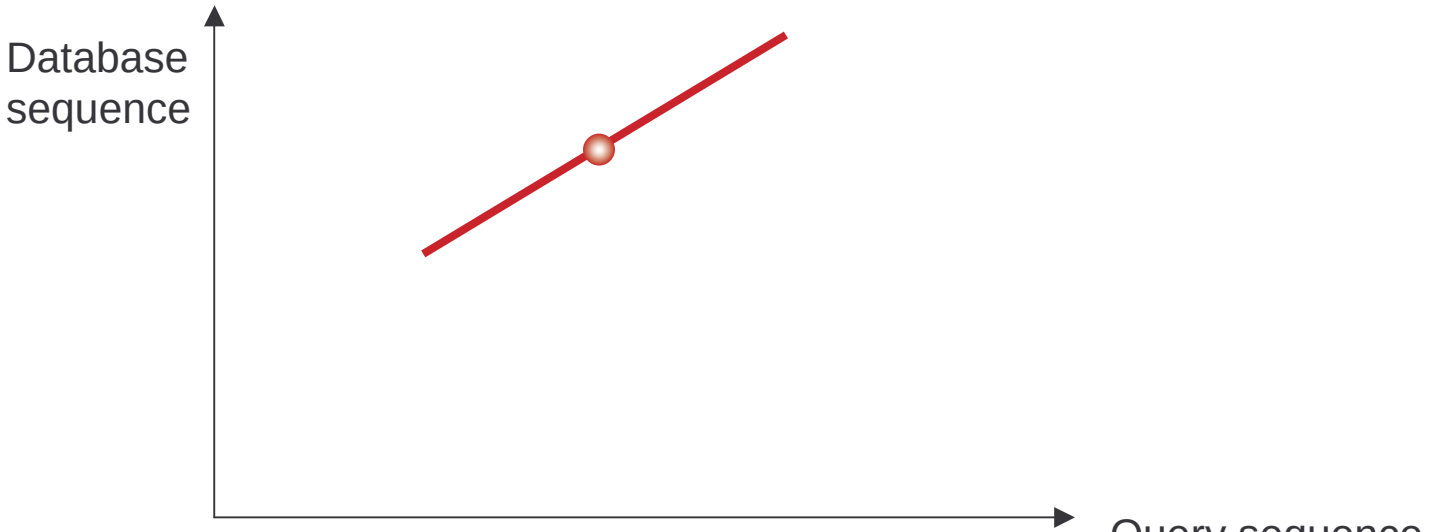
→ Reduces number of extensions.

→ Threshold for initial hits has to be reduced to achieve same level of sensitivity.

HUSAR

Blast2 - extension -

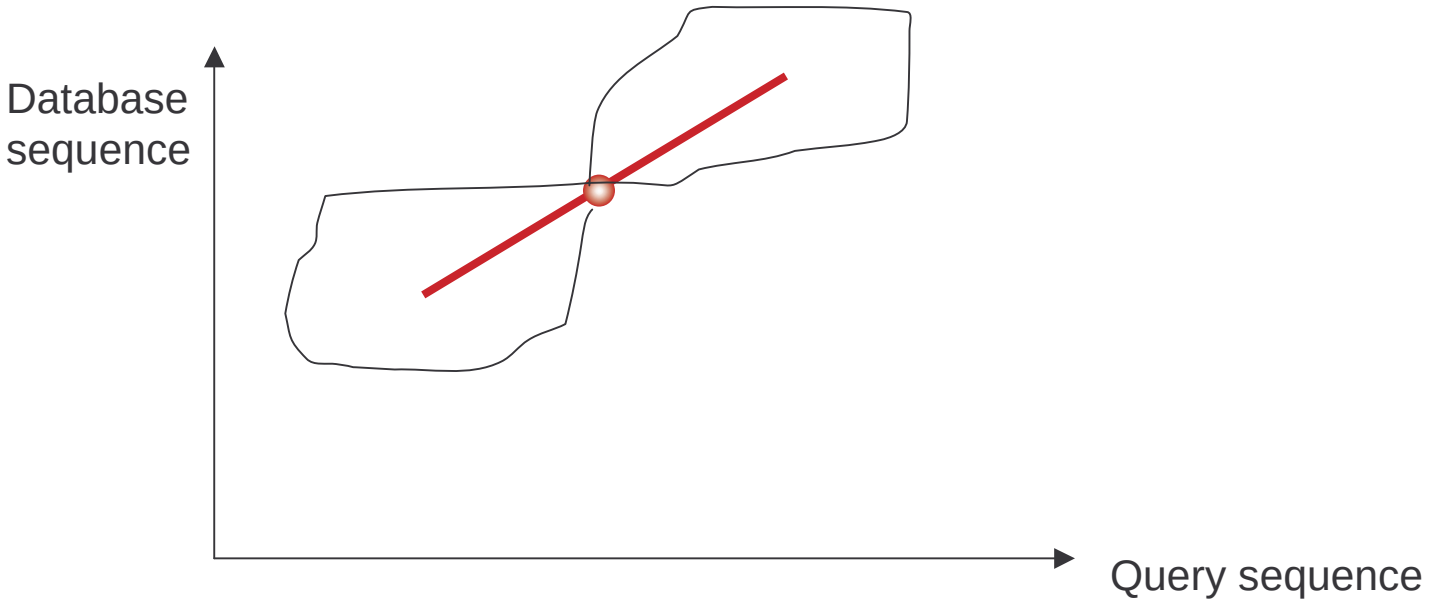
Introduction of gapped alignments



HUSAR

Blast2 - extension -

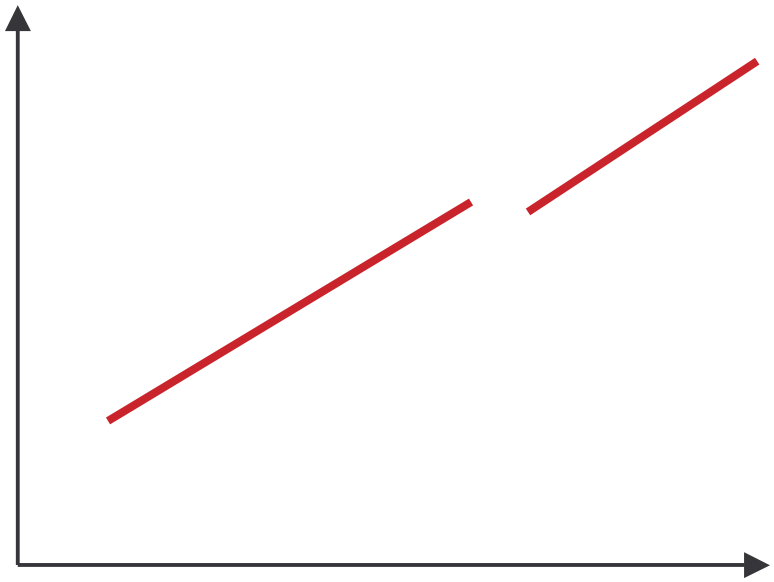
Introduction of gapped alignments



HUSAR

Blast2 - extension -

Due to gapped alignments, only one HSP out of a group needs to be found!



→ Threshold of initial hits can be increased.

HUSAR

Blast2 - statistics -

Due to gapped alignments, no statistical treatment possible for arbitrary substitution scores.

Programs use pre-computed statistical parameters for predefined substitution scores.

→ Only limited number of substitution matrices available.

Gap penalties are only slightly changeable.

Blast available in HUSAR:

BlastN(2)	- fast comparison of nucleotide sequence to DNA sequence database
TBlastN(2)	- fast comparison of peptide sequence to DNA sequence database
BlastP(2)	- fast comparison of peptide sequence to protein sequence database
BlastX(2)	- comparison of nucleotide sequence to protein sequence database
TBlastX(2)	- compares six-frame translation of nucleotide query against six-frame translation of nucleotide sequence database

HUSAR

Gaps in Blast2

- **BlastN2, BlastP2** - fully gapped alignments
- **TBlastN2, BlastX2** - in-frame gapped alignments
- **TBlastX2** - no gaps

Comparison between FASTA and BLAST

- **Nucleotide sequences: Fasta much more sensitive**
  BlastN(2) much faster
- **Blast locates multiple hits between query and ONE database sequence**
- **Both algorithms cannot deal with short query sequences very well**

Smith-Waterman style searches:

Available programs in HUSAR:

SW_Bic	Protein query against protein database
	DNA query against DNA database
TSWN	Protein query against DNA database
TSWP	DNA query against protein database
Frame*	Protein query against DNA database or DNA query against protein database recognizing frame shifts

Example:

BlastP // BlastP2

BlastP output: part I

Query sequence: Rat adenylat cyclase:

```
                                                                    Sum
                                                          High  Probability
Sequences producing High-scoring Segment Pairs:          Score  P(N)      N

>>>swissprot 25A1_RAT  Q05961 rattus norvegicus (rat). (...   1915  1.8e-263  1
>>>swissprot:25A1_MOUSE P11928 mus musculus (mouse). (2'-...  1583  8.6e-222  2
>>>swissprot:25A2_HUMAN P04820 homo sapiens (human). (2'-...   782  9.2e-179  3
>>>swissprot:25A1_HUMAN P00973 homo sapiens (human). (2'-...   782  6.0e-173  2
>>>swissprot:25A2_MOUSE P29080 mus musculus (mouse). (2'-...   782  6.0e-173  2
>>>swissprot 25A6_HUMAN 9728 homo sapiens (human). 69/7...    386  1.8e-118  4
>>>swissprot:[illegible] 0081 mus musculus (mouse). (2'-...   785  5.7e-106  1
>>>swissprot:TR14_HUMAN Q15646 homo sapiens (human). thyr...  161  1.5e-16   2
>>>swissprot:YKH2_YEAST P36085 saccharomyces cerevisiae (...    66  0.71      1
```

Example:

BlastP // BlastP2

BlastP2 output: part I

Query sequence: Rat adenylat synthetase:

```
Searching.................................................done

                                                                 Score     E
Sequences producing significant alignments:                      (bits)  Value
>>>swissprot:25A1_RAT Q05961 rattus norvegicus (rat). (2'-5')ol...   750  0.0
>>>swissprot:25A1_MOUSE P11928 mus musculus (mouse). (2'-5')oli...   629  e-180
>>>swissprot:25A2_MOUSE P29080 mus musculus (mouse). (2'-5')oli...   495  e-140
>>>swissprot:25A2_HUMAN P04820 homo sapiens (human). (2'-5')oli...   495  e-140
>>>swissprot:25A1_HUMAN P00973 homo sapiens (human). (2'-5')oli...   495  e-140
>>>swissprot:25A3_MOUSE P29081 mus musculus (mouse). (2'-5')oli...   492  e-139
>>>swissprot:25A6_HUMAN 9728 homo sapiens (human). 69/71 kd (...   350  2e-96
>>>swissprot:IRI4_HUMAN Q15646 homo sapiens (human). thyroid re...    77  4e-14

>>>swissprot:RN14_YEAST P25298 saccharomyces cerevisiae (baker'...    32  1.5
```

Example:

BlastP // BlastP2

BlastP output: part II

25A6_HUMAN

```
 Score = 386 (178.8 bits), Expect = 1.8e-118, Sum P(4) = 1.8e-118
 Identities = 68/121 (56%), Positives = 94/121 (77%)

Query:   228 LPPQYALELLTVYAWERGNGITEFNTAQGFRTILELVTKYQQLRIYWTKYYDFQHPDVSK 287
             LPP+YALELLT+YAWE+G+G+ +F+TA+GFRT+LELVT+YQQL I+W   Y+F+   V K

Sbjct:   562 LPPKYALELLTIYAWEQGSGVPDFDTAEGFRTVLELVTQYQQLGIFWKVNYNFEDETVRK 621

                                    . . .

Score = 260 (120.4 bits), Expect = 1.8e-118, Sum P(4) = 1.8e-118
 Identities = 54/118 (45%), Positives = 73/118 (61%)

Query:    58 KVVKGGSSGKGTTLKGKSDADLVVFLNNFTSFEDQLNRRGEFIKEIKKQLYEVQREKHFR 117
             ++V+GGS+ KGT LK  SDADLVVF N+  S+  Q N R + +KEI +QL    REK
Sbjct:   389 QIVRGGSTAKGTALKTGSDADLVVFHNSLKSYTSQKNERHKIVKEIHEQLKAFWREKEEE 448

  Score = 222  ...
  Score = 174  ...
  Score = 142  ...
  Score = 121
  Score = 98
  Score = 69
  Score = 47
  Score = 35
  Score = 35   ...
```

BlastP reports 11 HSPs

Example:

BlastP // BlastP2

BlastP2 output: part II

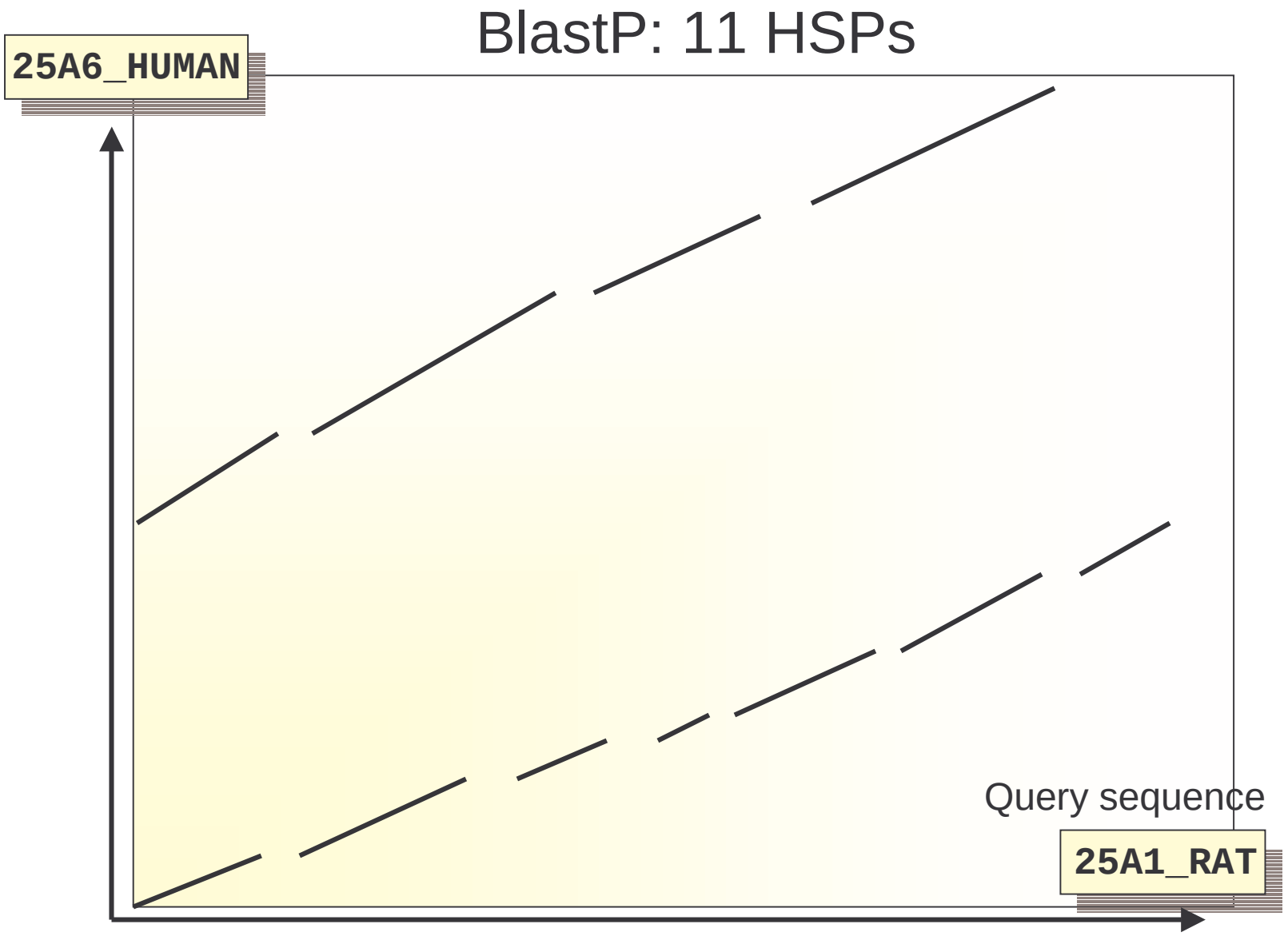
25A6_HUMAN

```
 Score =  350 bits (888), Expect = 2e-96
 Identities = 178/348 (51%), Positives = 237/348 (67%), Gaps = 8/348 (2%)

Query: 5   LRSTPSWKLDKFIEVYLLPNTSFRDDVKSAINVLCDFLKERCFRDTVHPVRVSKVVKGGS 64
           L +TP   LDKFI+ +L PN  F + + SA+N++  FLKE CFR +   +++   V+GGS
Sbjct: 339 LFTTPGHLLDKFIKEFLQPNKCFLEQIDSAVNIIRTFLKENCFRQSTAKIQI---VRGGS 395

                          . . .

Query: 301 LDPADPTGNVAGGNQEGWRRLASEARLWLQYPCFMNRGGSPVSSWEVP 348
           LDP +PTG+V GG++  W  L  EA++ L  PCF +  G+P+  W+VP
Sbjct: 635 LDPGEPTGDVGGGDRWCWHLLDKEAKVRLSSPCFKDGTGNPIPPWKVP 682

 Score =  214 bits (540), Expect = 2e-55
 Identities = 136/349 (38%), Positives = 187/349 (52%), Gaps = 21/349 (6%)

Query: 2   EQELRSTPSWKLDKFIEVYLLPNTSFRDDVKSAINVLCDFLKERCFRDTVHPVRVSKVVK 61
           E +L S P+ KL  FI+ YL P    +  +N +CD      C      P+ V  V
Sbjct: 4   ESQLSSVPAQKLGWFIQEYLKPYEECQTLI

                          . . .

Query: 299 VILDPADPTGNVAGGNQEGWRRLASEARLWLQYPCFMNRGGSPVSSWEV 347
           VILDP DPT NV+ G++  W+ L  EA+ WL  P   N    P  SW V

Sbjct: 289 VILDPVDPTNNVS-GDKICWQWLKKEAQTWLTSPNLDNE--LPAPSWNV 334
```

BlastP2 reports 2 HSPs

Example:

BlastP // BlastP2



Example:

BlastN // BlastN2 // FastA // SW

BlastN output

Query sequence: cDNA HUMAN Keratin against HONEST

emhum1: HSCHIKER M37818

Sequences producing High-scoring Segment Pairs:			High Score	Smallest Sum Probability P(N)	N
>>>emhum1:HSCHIKER	M3	ps...	1115	8.5e-252	5
>>>emhum1:HSKERUV	X05803	Human radiated keratinocyte mR...	718	1.0e-105	2
>>>emhum1:HSEPKER	J00124	Homo sapiens 50 kDa type I epi...	687	4.2e-105	4
>>>emhum1:HSKERELP	X62571	H.sapiens mRNA for keratin-rel...	718	5.4e-105	2
>>>emrod:MMKTEPIA	M13806	Mouse keratin (epidermal) type...	691	9.3e-101	2
>>>emrod:MMKTEPIC	M13805	Mouse type I epidermal keratin...	700	2.0e-96	2
>>>emhum1:HSCYTOK17	Z19574	H.sapiens gene for cytokeratin...	651	3.2e-90	3
>>>emhum2:S79867	S79867	type I keratin 16 [human, epid...	709	2.7e-87	2
>>>emhum2:S72493	S72493	keratin=keratin 16 homolog [hu...	693	2.3e-86	2
>>>emhum1:HSKERP2	M22928	Human keratin pseudogene, exon...	624	1.1e-85	3
>>>emhum1:HSKERC15	X07696	Human mRNA for cytokeratin 15....	560	3.3e-72	2
>>>emhum2:HSTYPIKER	X90763	H.sapiens mRNA for type I kera...	527	2.1e-56	2
>>>emhum1:HS1K14	M28646	Homo sapiens epidermal keratin...	745	3.2e-52	1
>>>emhum2:HSPK161	M37817	Human keratin psi-K#16-1 pseud...	480	6.2e-49	2
>>>emhum1:HSKERI6	M21755	Human pot. pseudo-keratin K16 ...	651	6.6e-46	1
>>>emhum1:HSKER16A6	M28437	Human keratin type 16 gene, ex...	642	3.8e-45	1

Example:

BlastN // BlastN2 // FastA // SW

BlastN2 output

Query sequence: cDNA HUMAN Keratin against HONEST

emhum1: HSCHIKER M37818

```
Sequences producing significant alignments:                          Score     E
                                                                     (bits)  Value
>>>emhum1:HSCHIKER M37818 [illegible]udogen...                         442  e-122
>>>emhum1:HS1K14 M28646 Homo sapiens epidermal keratin, 3' end....     103  5e-20
>>>emhum1:HSEPKER J00124 Homo sapiens 50 kDa type I epidermal k...     103  5e-20
>>>emmam:BTKERVII X01461 Bovine mRNA 3'-region for epidermal ke...      96  1e-17
>>>emrod:MMKTEPIB J02644 Mouse type I epidermal keratin mRNA, 3...      96  1e-17
>>>emrod:MMKTEPIA M13806 Mouse keratin (epidermal) type I mRNA,...      88  3e-15
>>>emhum1:HSKERI6 M21755 Human pot. pseudo-keratin K16 type I, ...      84  4e-14
>>>emhum1:HSCYTOK17 Z19574 H.sapiens gene for cytokeratin 17. 6/93      78  3e-12
>>>emhum1:HSKERELP X62571 H.sapiens mRNA for keratin-related pr...      78  3e-12
>>>emhum1:HSKERUV X05803 Human radiated keratinocyte mRNA 266 (...      78  3e-12
>>>emrod:MMKTEPIC M13805 Mouse type I epidermal keratin mRNA, c...      76  1e-11
>>>emhum1:HSKER16A6 M28437 Human keratin type 16 gene, exon 6. ...      76  1e-11
>>>emhum2:S79867 S79867 type I keratin 16 [human, epidermal ker...      76  1e-11
>>>emhum2:HSPK161 M37817 Human keratin psi-K#16-1 pseudogene, 3
```

Example:

BlastN // BlastN2 // FastA // SW

FastA output

Query sequence: cDNA HUMAN Keratin against HONEST

emhum1: HSCHIKER M37818

```
The best scores are:                                      init1 initn  opt..
emhum1:hschiker  M37818 Human keratin (psi-K-alpha) pseud... 892  1924   912
emhum1:hsepker  J00124 Homo sapiens 50 kDa type I epiderm... 559  1101   564
emrod:mmktepia  M13806 Mouse keratin (epidermal) type I m... 559  1082  1109
emhum1:hskerelp  X62571 H.sapiens mRNA for keratin-relate... 521  1015  1116
emhum1:hskeruv  X05803 Human radiated keratinocyte mRNA 2... 521  1015  1116
emrod:mmkt                                                .. 503  1005  1054
emhum1:hsk                                                .. 447   993   511
emhum1:hscytok17  Z19574 H.sapiens gene for cytokeratin 1... 461   977   543
emhum2:s79867  S79867 type I keratin 16 [human, epidermal... 503   890   976
emhum2:s72493  S72493 keratin=keratin 16 homolog [human, ... 496   883   965
emhum1:hscytk  X52426 H.sapiens mRNA for cytokeratin 13. ... 499   753   776
emhum1:hsker13c  X14640 Human mRNA for keratin 13. 1/94      492   746   769
emhum1:hskerc15  X07696 Human mRNA for cytokeratin 15. 9/93  391   735   832
emhum1:hs1k14  M28646 Homo sapiens epidermal keratin, 3' ... 601   641   647
```

Fasta detects only the longest exon!

Example:

BlastN // BlastN2 // FastA // SW

SW output

Query sequence: cDNA HUMAN Keratin against HONEST

```
Sequence              Strd Orig ZScore    EScore    Len  ! Documentation
emhum1:HSKERUV         +  280.40    1079.43  0.000000e2147483647  956  ! X05803 Human
radiated keratinocyte mRNA 266 (keratin-related protein). 7/95

emhum1:HSKERELP        +  280.40    1076.40  0.000000e2147483647  1512  ! X62571
H.sapiens mRNA for keratin-related protein. 11/92

em_ro:MMKTEPIA         +  278.80    1071.56  0.000000e2147483647  1224  ! M13806
Mouse keratin (epidermal) type I mRNA, clone p52, partial  4/90

em_ro:MMKTEPIC         +  265.20    1021.34  0.000000e2147483647  801  ! M13805 Mouse
type I epidermal keratin mRNA, clone pkSCC-50, partial cds. 4/90

emhum2:S79867          +  248.40     952.05  0.000000e2147483647  1422  ! S79867 type
I keratin 16 [e                                                        . 2

emhum2:S72493          +  245.80     944.40  0.000000e2147483647  976  ! S72493
keratin [human, tracheobronchial epithelial cells, mRNA P

emhum1:HSCHIKER        +  2                  483647  9697  ! M37818
Human                                        e, exons 4,5,6,7,8,
```

SW detects only the longest exon!

emhum1: HSCHIKER M37818

Example:

BlastX // BlastX2 // Framesearch

Framesearch output

Genbank: atts0012 EST against Swissprot

atts0012.gb_est5
G3PC_ARATH

```
                Quality:  343.0              Length:    240
                  Ratio:  4.397                Gaps:      2
            Percent Similarity: 100.000   Percent Identity: 97.436

              .         .         .         .         .
  3 GAAATCAAGAAGGCCATCAAGGAGGAATCTGAAGGCAAAATGAAGGGAAT 52
    |||||||||||||||||||||||||||||||||||||:::||||||||
261 GluIleLysLysAlaIleLysGluGluSerGluGlyLysLeuLysGlyIl 277   . . .
              .         .         .         .         .
103 ACAGGTCAAGCATTTTCGATGCCAAGGCTGGATTGCATTGCATTGAGCGA 152
    |||||||||||||||||||||||||||||||    |||||||||||||||
295 snArgSerSerIlePheAspAlaLysAlaGly....IleAlaLeuSerAs 309
              .         .         .         .         .
153 CAAGTTTGTGAAGTTGGTGTCATGGTACGACAACGAATGGGGTTACACAG 202
    ||||||||||||||||||||||||||||||||||||||||||||||||  ||
310 pLysPheValLysLeuValSerTrpTyrAspAsnGluTrpGlyTyr..Se 325

              .         .         .         .

203 TTCTCGTGTCGTTGACCTTATCGTTCACATGTCAAAGGCC 242
    |||||||||||||||||||||||||||||||||||||||

326 rSerArgValValAspLeuIleValHisMetSerLysAla 338
```

Example:

BlastX // BlastX2 // Framesearch

BlastX output

Genbank: atts0012 EST against Swissprot

G3PC_ARATH

```
 Plus Strand HSPs:

 Score = 215 (98.9 bits), Expect = 1.6e-44, Sum P(3) = 1.6e-44
 Identities = 42/45 (93%), Positives = 45/45 (100%), Frame = +3

Query:     3 EIKKAIKEESEGKMKGILGYSEDDVVSTDFVGDNRSSIFDAKAGL 137
             EIKKAIKEESEGK+KGILGY+EDDVVSTDFVGDNRSSIFDAKAG+
Sbjct:   261 EIKKAIKEESEGKLKGILGYTEDDVVSTDFVGDNRSSIFDAKAGI 305

 Score = 114 (52.4 bits), Expect = 1.6e-44, Sum P(3) = 1.6e-44
 Identities = 20/21 (95%), Positives = 21/21 (100%), Frame = +1

Query:   139 IALSDKFVKLVSWYDNEWGYT 201
             IALSDKFVKLVSWYDNEWGY+
Sbjct:   305 IALSDKFVKLVSWYDNEWGYS 325

 Score = 67 (30.8 bits), Expect = 1.6e-44, Sum P(3) = 1.6e-44
 Identities = 14/15 (93%), Positives = 15/15 (100%), Frame = +3

Query:   198 HSSRVVDLIVHMSKA 242
             +SSRVVDLIVHMSKA

Sbjct:   324 YSSRVVDLIVHMSKA 338
```

Example:

BlastX // BlastX2 // Framesearch

BlastX2 output

Genbank: atts0012 EST against Swissprot

G3PC_ARATH

```
>>>>swissprot [illegible]858 arabidopsis thaliana (mouse-ear
           cress). glyceraldehyde 3-phosphate dehydrogenase,
           cytosolic (ec 1.2.1.12). 5/92
           Length = 338

 Score =  104 bits (257), Expect = 4e-23
 Identities = 59/80 (73%), Positives = 65/80 (80%)

Query: 3   EIKKAIKEESEGKMKGILGYSEDDVVSTDFVGDNRSSIFDAKAGLHCIERQVCEVGVMVR 182
           EIKKAIKEESEGK+KGILGY+EDDVVSTDFVGDNRSSIFDAKAG+   ++ V  V
Sbjct: 261 EIKKAIKEESEGKLKGILGYTEDDVVSTDFVGDNRSSIFDAKAGIALSDKFVKLVS-WYD 319

Query: 183 QRMGLHSSRVVDLIVHMSKA 242
              G +SSRVVDLIVHMSKA
Sbjct: 320 NEWG-YSSRVVDLIVHMSKA 338
```

Search strategies:

Searching in Sequence Databases

- The comparison of a **sequence** against a database ….

  In order to find similar sequences

- The comparison of a **pattern,** a **profile**, or a **HMM** against a database ….

  In order to find **distantly related** sequences.

Pattern searching

Pattern language

Fixed symbols	A
Alternatives	(A,B,C)
Exclusions	~(D,E,F)
Repetitions	(A,B){n1,n2}

Example:

Pattern searching

Pattern: G(D,E)~PX{0,4}(D,E){1,2}W

GDSTRLDDW

GEAEW

HUSAR

Pattern searching

Programs available in HUSAR

FindPatterns

- Finds DNA or protein patterns in a database of the same kind. (*GCG*)

TFindPatterns

- Finds a protein pattern in a dynamically translated nucleic acid sequence.

HUSAR

Profile Searching

- Suited for finding and aligning distantly related proteins.
- Input is not a single sequence, but a profile generated from a multiple alignment of similar sequences.
- The profile reflects the likelihoods of all amino acids occurring at every position in the alignment.
- Position specific gap weights.

Profile Searching

Programs available in GCG / HUSAR

Profile analysis package: *M. Gribskov*

ProfileMake - creates a profile from a multiple alignment.

ProfileSearch - compares a profile against a database using Smith-Waterman style search.

TProfile Search - compares a protein profile against a nucleic acid database.

ProfileSegments - displays the alignments of a ProfileSearch result.

ProfileGap - compares a profile against a sequence.

HUSAR

Hidden Markov Models

- Suited for finding and aligning distantly related proteins.
- A Hidden Markov Model is a more general description of a multiple alignment than a profile.
- Input is not a single sequence, but a Hidden Markov Model generated from a multiple alignment of similar sequences.

Hidden Markov Models

Programs available in HUSAR

HMMBUILD	- generate a Hidden Markov Model (HMM)
HMMSEARCH	- Smith-Waterman search using a HMM against a sequence database
HMMSCAN	- scan a sequence against a database of HMMs

PSIBlast

- Iterative Blast2 program starting with a single protein sequence.
- Automatically generates profile from search results.
- Blast2 algorithm extended to work with profiles.

Example:

Detecting weak protein homologies

A Comparison of different methods:

Input: Human hemoglobin alpha (BlastP, FastA, SW, PSIBlast)

Profile of the alignment of 7 hemoglobin and myoglobin sequences (ProfileSearch)

HMM of the alignment of 7 hemoglobin and myoglobin sequences (HMMSearch)

Database: containing 14 globin sequences (total 10.000 sequences)

Example:

BlastP

HBAC_ANGAN	P80726	HEMOGLOBIN	ALPHA, CATHODIC CHA...	255	1.6e-46	2
HBAA_ANGAN	P80945	HEMOGLOBIN	ALPHA, ANODIC CHAIN...	218	6.4e-41	2
HBBC_ANGAN	P80727	HEMOGLOBIN	BETA, CATHODIC CHAI...	188	5.6e-37	3
HBBA_ANGAN	P80946	HEMOGLOBIN	BETA, ANODIC CHAIN....	116	7.1e-17	2
MYG_HORSE	P02188	MYOGLOBIN.	2/97	72	7.8e-09	2
PTB_PIG	Q29099	POLYPYRIMIDINE TRACT-BINDING P...		59	0.45	1
DLDH_MYCPN	P75393	LIPOAMIDE DEHYDROGENASE COMPON...		56	0.79	1
XERC_SALTY	P55888	INTEGRASE/RECOMBINASE XERC. 2/97		39	0.87	3
SYI_CAEEL	Q21926	PROBABLE ISOLEUCYL-TRNA SYNTHE...		49	0.993	2
LEUA_MICAE	P94907	2-ISOPROPYLMALATE SYNTHASE (EC...		40	0.994	3
TPIS_RHIET	P96985	TRIOSEPHOSPHATE ISOMERASE (EC ...		43	0.995	2
HJA3_METJA	Q58655	PROBABLE ARCHAEAL HISTONE 3. 2/97		49	0.999	1
CMTG_PSEPU	Q51983	4-HYDROXY-2-OXOVALERATE ALDOLA...		45	0.999	2

HUSAR

Example:

FastA

```
hbac_angan   P80726 HEMOGLOBIN ALPHA, CATHODIC CHAIN... 263   370   381
hbaa_angan   P80945 HEMOGLOBIN ALPHA, ANODIC CHAIN. ... 226   331   351
hbbc_angan   P80727 HEMOGLOBIN BETA, CATHODIC CHAIN.... 187   231   300
hbba_angan   P80946 HEMOGLOBIN BETA, ANODIC CHAIN. 2/97 120   149   238
syl_syny3   P73274 LEUCYL-TRNA SYNTHETASE (EC 6.1.1....  42    66    57
ymb3_yeast   Q04228 HYPOTHETICAL 66.8 KD PROTEIN IN ...  40    66    42
rdh1_bovin   Q27979 11-CIS RETINOL DEHYDROGENASE (EC...  54    65    65
amd_mouse   P97467 PEPTIDYL-GLYCINE ALPHA-AMIDATING ...  36    64    41
faf_mouse   P70398 PROBABLE UBIQUITIN CARBOXYL-TERMI...  45    63    45
ycby_haein   P44524 HYPOTHETICAL PROTEIN HI0116/15. ...  49    61    56
myg_horse   P02188 MYOGLOBIN. 2/97                       40    61   172
ptga_brela   Q45298 PTS SYSTEM, GLUCOSE-SPECIFIC IIA...  49    60    69
e6ap_human   Q05086 ONCOGENIC PROTEIN-ASSOCIATED PRO...  44    60    44
mauf_thive   Q56463 METHYLAMINE UTILIZATION PROTEIN ...  48    59    48
```

Example:

SW

```
HBAC_ANGA P80726 HEMOGLOBIN ALPHA, CATHODIC CHAIN. 2/97       .0000   378.0
HBAA_ANGA P80945 HEMOGLOBIN ALPHA, ANODIC CHAIN. 2/97         .0000   348.0
HBBC_ANGA P80727 HEMOGLOBIN BETA, CATHODIC CHAIN. 2/97        .0000   296.0
HBBA_ANGA P80946 HEMOGLOBIN BETA, ANODIC CHAIN. 2/97          .0000   231.0
MYG_HORSE P02188 MYOGLOBIN. 2/97                              .0000   172.0
Y168_HAEI P43958 HYPOTHETICAL PROTEIN HI0168/69. 2/97         .0001    80.0
GLBI_CHIT Q23763 GLOBIN CTT-VIIB-8 PRECURSOR. 2/97            .0004    73.0
HMPA_ERWC Q47266 FLAVOHEMOPROTEIN (HAEMOGLOBIN-LIKE PROTEIN) .0006    75.0
GLBZ_CHIT Q23761 GLOBIN CTT-Z PRECURSOR (HBZ). 2/97           .0007    70.0
MAOX_MYCT P71880 PUTATIVE MALATE OXIDOREDUCTASE (NAD) (EC 1. .0011    75.0
YCHM_ECOL P40877 HYPOTHETICAL 58.4 KD PROTEIN IN PTH-PRSA IN .0014    73.0
GLBK_CHIT Q23762 GLOBIN CTT-VIIB-10 PRECURSOR. 2/97           .0014    67.0
FCPC_MACP Q40299 FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN .0018    67.0
                      . . .
```

Position 25:

```
GLB1_CHIT P02221 GLOBIN CTT-I/CTT-IA PRECURSOR (ERYTHROCRUOR .0037    62.0
```

Position 35:

```
GLB6_CHIT P02224 GLOBIN CTT-VI PRECURSOR. 2/97                .0047    61.0
```

Example:

PSIBlast

```
Results from round 1

                                                                  Score     E
Sequences producing significant alignments:                      (bits)  Value

>>>swnew:HBAC_ANGAN P80726 HEMOGLOBIN CATHODIC, ALPHA CHAIN. 1...   141  2e-35
>>>swnew:HBAA_ANGAN P80945 HEMOGLOBIN ANODIC, ALPHA CHAIN. 10/...   125  2e-30
>>>swnew:HBBC_ANGAN P80727 HEMOGLOBIN CATHODIC, BETA CHAIN. 11...   102  1e-23
>>>swnew:HBBA_ANGAN P80946 HEMOGLOBIN ANODIC, BETA CHAIN. 11/1997    83  7e-18
>>>swnew:MYG_HORSE  P02188 MYOGLOBIN. 10/2000                       48  4e-07

Results from round 2

>>>swnew:PYRC_AERPE Q9yfi5 DIHYDROOROTASE (EC 3.5.2.3) (DHOASE...    24  5.7
>>>swnew:GLB_PAREP P80721 GLOBIN-3 (MYOGLOBIN). 7/1998              24  5.7
>>>swnew:Y373_HUMAN O15078 HYPOTHETICAL PROTEIN KIAA0373. 10/2000   24  7.5
```

HUSAR

Example:

ProfileSearch

```
SWNEW:HBBC_ANGAN      +   33.39  51.16    146 ! P80727 HEMOGLOBIN BETA, CATHODIC
SWNEW:HBBA_ANGAN      +   29.72  46.88    147 ! P80946 HEMOGLOBIN BETA, ANODIC C
SWNEW:HBAC_ANGAN      +   28.88  45.16    142 ! P80726 HEMOGLOBIN ALPHA, CATHODI
SWNEW:HBAA_ANGAN      +   27.36  43.34    142 ! P80945 HEMOGLOBIN ALPHA, ANODIC
SWNEW:MYG_HORSE       +   27.02  44.37    153 ! P02188 MYOGLOBIN. 2/97
SWNEW:GLBI_CHITH      +   11.95  26.38    161 ! Q23763 GLOBIN CTT-VIIB-8 PRECURS
SWNEW:GLB1_CHITH      +   11.66  25.81    158 ! P02221 GLOBIN CTT-I/CTT-IA PRECU
SWNEW:GLBK_CHITH      +   11.65  26.01    161 ! Q23762 GLOBIN CTT-VIIB-10 PRECUR
SWNEW:GLBZ_CHITH      +   11.35  25.75    163 ! Q23761 GLOBIN CTT-Z PRECURSOR (H
SWNEW:GLB_PAREP       +   10.99  24.20    147 ! P80721 GLOBIN (MYOGLOBIN). 2/97
SWNEW:GLB_ISOHY       +   10.98  24.26    148 ! P80722 GLOBIN (MYOGLOBIN). 2/97
SWNEW:GLBW_CHITH      +    8.27  21.63    159 ! Q23760 GLOBIN CTT-W PRECURSOR (H
SWNEW:GLB6_CHITH      +    8.20  21.72    162 ! P02224 GLOBIN CTT-VI PRECURSOR.
SWNEW:Y395_METJA      +    5.36  17.94    158 ! Q57838 HYPOTHETICAL PROTEIN MJ03
SWNEW:YM34_YEAST      +    4.86  16.93    150 ! Q03818 HYPOTHETICAL 17.2 KD PROT
SWNEW:YA74_METJA      +    4.71  14.66    112 ! Q58474 HYPOTHETICAL PROTEIN MJ10
SWNEW:Y267_MYCPN      +    4.48  14.53    114 ! P75397 HYPOTHETICAL PROTEIN MG26
                           . . .
```

Position 50: SWNEW: HMPA_ERWCH (flavohemoprotein)

HUSAR

Example:

HMMSearch

172.07	1	153	10	165	MYG_HORSE
135.21	4	145	12	158	HBBC_ANGAN
113.30	4	147	12	159	HBBA_ANGAN
111.99	1	142	10	159	HBAC_ANGAN
96.50	1	142	10	159	HBAA_ANGAN
33.70	8	150	1	147	GLBI_CHITH
31.65	8	150	1	147	GLBK_CHITH
26.41	8	157	1	154	GLBZ_CHITH
23.70	7	154	1	154	GLB1_CHITH
16.57	7	149	1	147	GLB6_CHITH
12.42	26	147	32	157	GLB_ISOHY
11.24	31	90	30	90	GLBW_CHITH
10.06	22	145	28	156	GLB_PAREP
8.16	75	126	114	165	SMF1_YEAST
6.51	347	434	76	165	AEFA_ECOLI
3.64	2	83	84	165	CAPP_AMAHP
3.60	72	96	1	25	PHZA_PSEAR
3.36	30	142	49	165	HMPA_ERWCH
2.58	36	73	1	39	ATPN_CAEEL
2.06	91	166	89	165	YFBM_ECOLI

Benchmarks

Protein query against protein database (Swissprot 38+ 85661 entries)

18.05.00

Seq.Length	BlastP	BlastP2	FastA	SSEARCH	SW(Biocc)
250	20.6	8.3	49	379	22.6
500	37.2	12.6	67	787	35.1
1000	80	16.0	88	1672	59.8
2000	194	26.6	116	3499	105
4000	308	62.1	195	8628	321

[time in seconds]

Benchmarks

DNA query against nucleic acid database (NR 610 000 entries)

18.05.00

Seq.Length	BlastN	BlastN2	FastA	SSearch	SW(Biocc)
200	40	39	1314	32970	1448
1000	48	62	2247	-	5254
5000	130	112	2748	-	24465
25000	473	459	10866	-	120562

[time in seconds]

HUSAR

Benchmarks

ProfileSearch:
- 28 sequences of length 720 against Swissprot (35)

FrameSearch:
- EST query (286 bp) against Swissprot (35)

	CVX	Biocc.
ProfileSearch	8380	97
FrameSearch	6300	140

[time in seconds]

Sequence Masking

DNA

DUST	(masking of locally biased regions, BlastN2)
XBlast	(masking of repetitive elements)
RepeatMasker	(“ “)

Protein

SEG	(masking of locally biased regions)
XNU	(“ “)

HUSAR

Example:

Sequence Masking

BlastN result against HONEST with masking

Sequences producing High-scoring Segment Pairs:		High Score	Smallest Sum Probability P(N)	N
>>>emhum1:HS367641	U36764 Human TGF-beta receptor interac...	470	1.6e-53	3
>>>emhum2:HSU39067	U39067 Human translation initiation fa...	461	1.1e-52	3
>>>emhum1:AC004161	Ac004161 Homo sapiens BAC clone RG208K...	249	1.4e-23	3
>>>empln:AT36765	U36765 Arabidopsis thaliana TGF-beta r...	189	2.3e-05	1
>>>eminv:DMU90930	U90930 Drosophila melanogaster TRIP-1 ...	189	2.3e-05	1
>>>emfun:SPD187	D89187 Schizosaccharomyces pombe mRNA,...	150	0.045	1
>>>emfun:SPSUM1	Y09529 S.pombe mRNA for SUM1 protein. ...	150	0.046	1
>>>emnew:SPSUM1	Y09529 S.pombe mRNA for SUM1 protein. ...	150	0.046	1
>>>emfun:SPAC4D7	Z98602 S.pombe chromosome I cosmid c4D...	150	0.049	1

Example:

Sequence Masking

BlastN result against HONEST without masking

```
WARNING:  Descriptions of 30,108 database sequences
          were not reported due to the limiting value
          of parameter -LIST=500.
```

DNA versus protein sequence comparison

	E(DNA)	E(protein)
MUSGLUTA	0	0
MAMGLUTRA	10^{-11}	0
MMGLUT	1.1	10^{-7}
HUMGSTB	14	10^{-6}

Distantly related sequences (Pearson,CABIOS 13, 4 (1997)

- Always compare proteins if the genes encode proteins
- Matches that are >50 % identical in a 20-40 amino acid region frequently occur by chance
- Most sequences with E<0.01 are homologous but distantly
- related sequences do not share significant homology
- Use “transitive” method!