Sequence-Based Data Mining

Jaroslaw Pillardy

Computational Biology Service Unit
Cornell University

Sequence analysis: what for?

- Finding coding regions (gene finding)
- Finding regulatory regions
- Analyzing mutation rates
- Determine properties of a sequence (repeats, low complexity regions)
- Functionally annotate genes
- Associate ESTs with genes
- Make cross-species comparison
- Build a model for a protein in order to understand its function, mutations etc
- And many more …

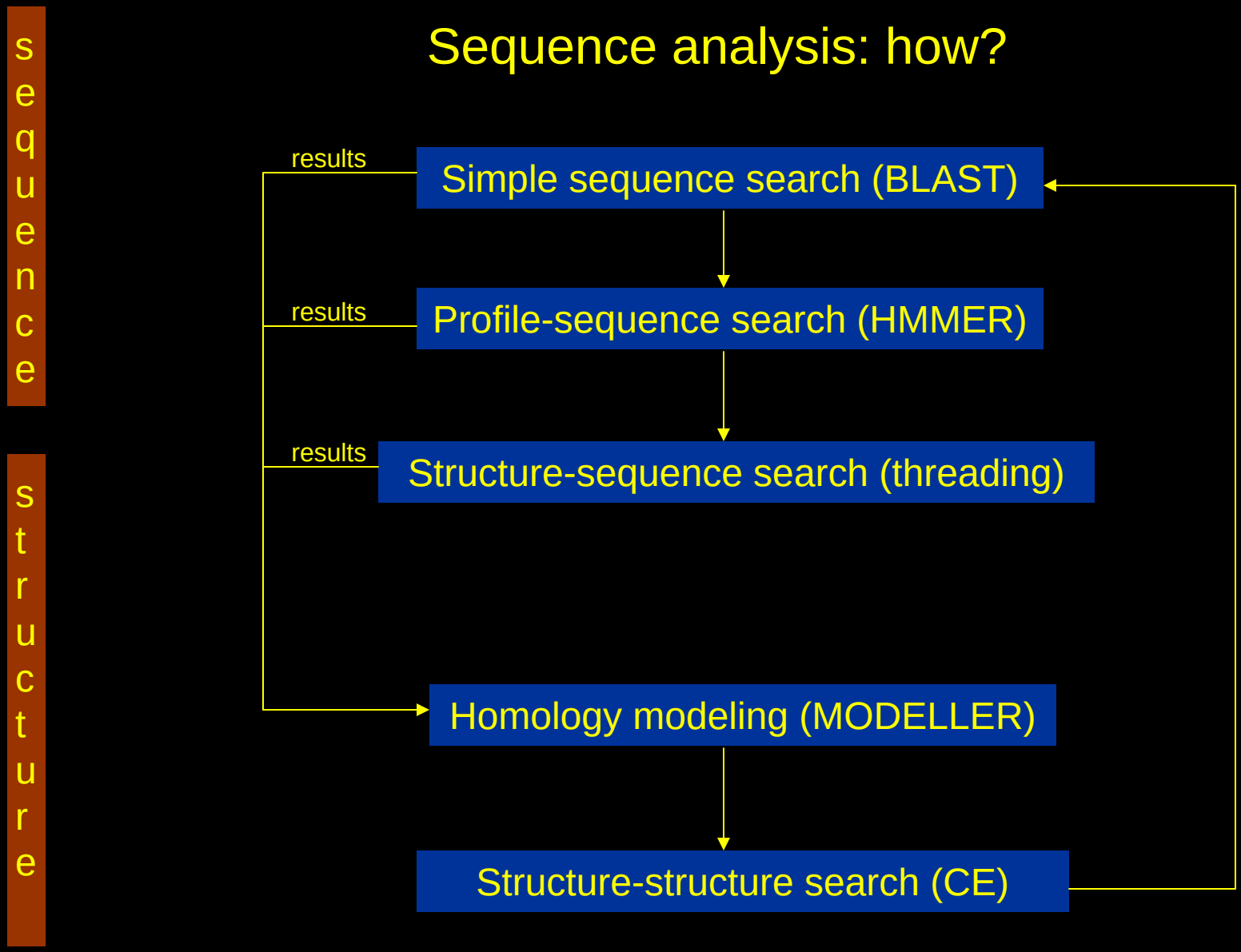
Sequence analysis: an example of a problem

Quiz:

A human geneticist identified a new gene that would significantly increase the risk of colon cancer when mutated. By using BLASTP, she found that this protein exists in a few vertebrate and invertebrate species with very low homology, but she was not able to find any good BLAST hits in *Drosophila melanogaster*.

Before making the conclusion that this gene does not exist in fly, what other approaches would you take?

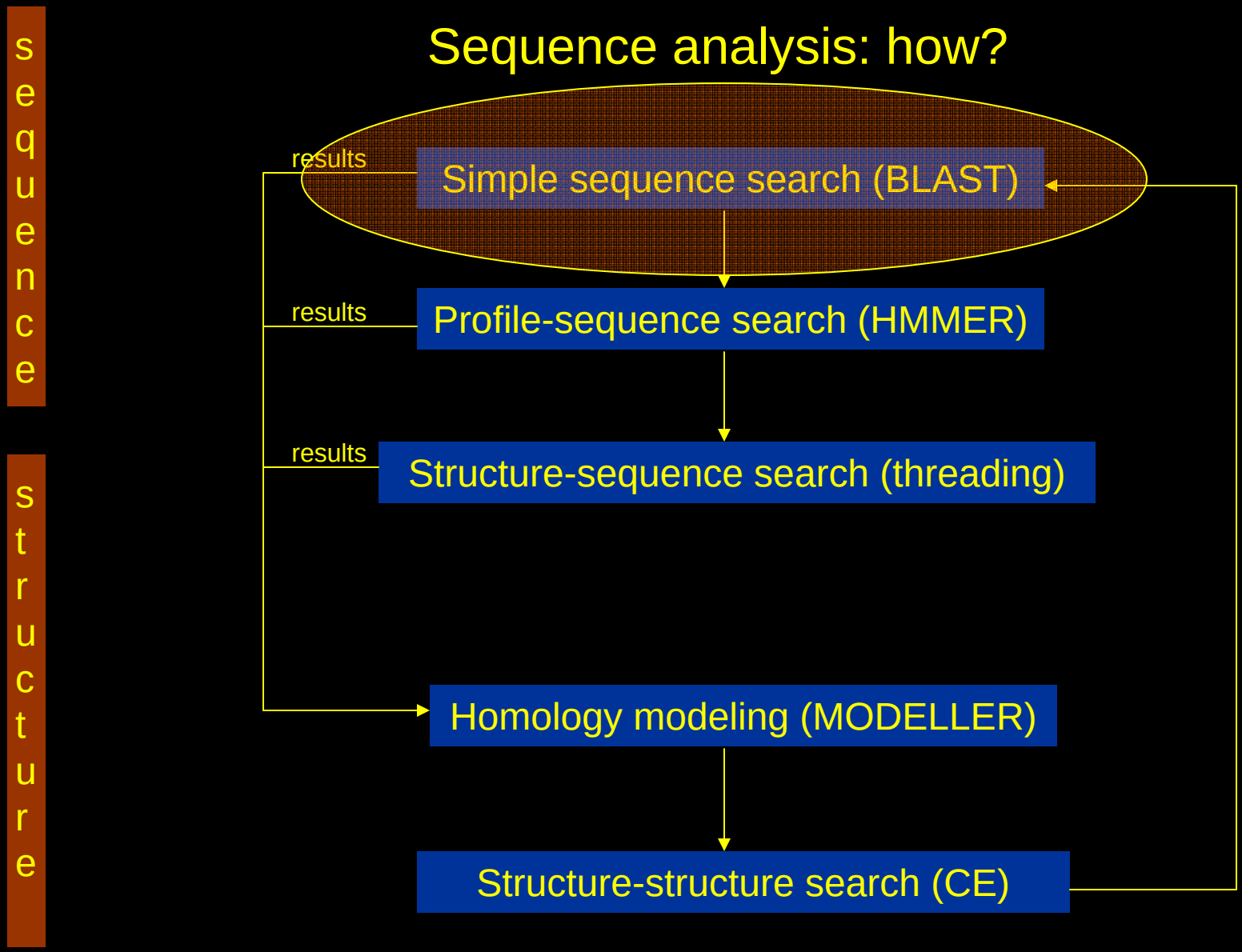
Sequence analysis: how?



Searching for similar proteins in a Database

	Simple sequence search	Profile-sequence search	Structure-sequence search
Sensitivity:	Least sensitive →		Most sensitive
Speed:	Seconds →	Minutes →	Hours
DB size:	4×10^6 →	4×10^6 →	4×10^4 (PDB)

Sequence analysis: how?



Simple sequence search

- Sequence similarity search looks like *syntactic* problem: comparing strings using alphabets

- Sequence homology is based of common ancestor and is *semantic* in nature
  - *orthologs* similar genes in different species, usually with same function
  - *paralogs* similar genes created by duplication, may be in same species, may not have the same function

- High sequence similarity does not imply homology, it is only a base for further investigation

- Physics can be reintroduced to sequence similarity search via scoring matrices

Scoring alignments

Scoring Matrices

- Relative entropy: $H = \Sigma\ q_{ij}c_{ij}$
- Shows information content per pair
- Matrices with larger entropy values are more sensitive to less divergent sequences
- Matrices with smaller entropy values are more sensitive to distantly related sequences

	a_1	a_2	a_3	a_4
a_1	c_{11}	c_{21}	c_{31}	c_{41}
a_2	c_{12}	c_{22}	c_{32}	c_{42}
a_3	c_{13}	c_{23}	c_{33}	c_{43}
a_4	c_{14}	c_{24}	c_{34}	c_{44}

- Relative entropy can be used to compare matrices
- Scores can be related to biology: negative=dissimilarity, zero=indifference, positive=similar

Scoring DNA alignments

Identity Matrix

```
AATTGGCTAGCTAA
|    || |||||||
...AAAAATGCAAAATGCGGGTAGCTTATTCTAGAAGATT...
```

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

Relative entropy: 1.0

Matches: 10
Mismatches: 4
Score: 10 x 1 + 4 x 0 = 10
Max score: 14
Expected score: 3.5
Minimum score: 0
Score: 71%

Scoring DNA alignments

BLAST Matrix

```
AATTGGCTAGCTAA
|   || |||||||
...AAAAATGCAAAATGCGGGTAGCTTATTCTAGAAGATT...
```

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

Relative entropy: -1.0

Matches: 10
Mismatches: 4
Score: 10 x 5 + 4 x (-4) = 36
Max score: 70
Expected score: -24.5
Minimum score: -56
Score: 73%

Scoring DNA alignments

Transition-Transversion Matrix

```
AATTGGCTAGCTAA
|   :|| |||||||
...AAAAATGCAAAATGCGGGTAGCTTATTCTAGAAGATT...
```

	A	T	C	G
A	1	-5	-5	-1
T	-5	1	-1	-5
C	-5	-1	1	-5
G	-1	-5	-5	1

Relative entropy: -4.5

Matches: 10 (1)
Mismatches: 3
Score: 10 x 1 + 3 x (-5)
+ 1 x (-1) = -6
Max score: 14
Expected score: -35
Minimum score: -70
Score: 42%

Scoring protein alignments

- 20 letter sequences, more possibilities
- Scoring may be based on physical properties of amino acids (polarity, size, hydrophobicity etc)
- Scoring may based on genetic code: minimum number of nucleotides substitutions necessary to convert
- Hard to put the above into a consistent scoring table
- Most popular matrices (PAM, BLOSUM) are based on observed substitution rates

```
ADCFDGGFAA
|  ||  ||  ||
AECFCGGEAA
```

Score = 4 + 2 + 9 + 6 -3 + 6 + 6 -3 + 4+ 4 = 35

	A	C	D	E	F	G	H →
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	[illegible]
G	0	-3	-1	-2	-3	[illegible]	
H ↓	-2	-3	-1	0			

BLOSUM 62

Scoring protein alignments : PAM

Deriving **P**oint **A**ccepted **M**utation matrix

- Dataset of families of very closely related proteins (identity >= 85%)
- Phylogenetic tree was constructed for each family
- Substitution frequency F_{ij} was computed
- Relative mutability m_i was computed for each amino acid (ratio of occurring mutation to all possible ones)
- Mutation probability $M_{ij} = m_j F_{ij} / \Sigma_l F_{ij}$
- $c_{ij} = \log(M_{ij}/f_i)$ – log odds matrix, f_j is frequency of occurrence

Scoring protein alignments : PAM

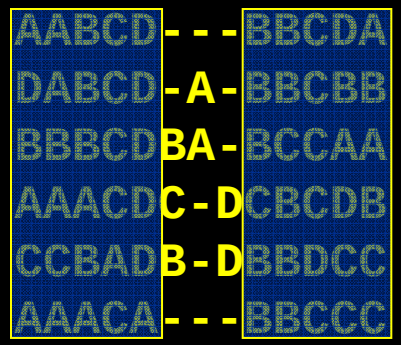
Using **P**oint **A**ccepted **M**utation matrix

- Matrix normalization to PAM-1 unit: 1 substitution over 100 residues
  "what is the probability of substitution of a residue during the time when 1% of residues mutated"
- Multiplication of PAM-1 unit produces substitution rates for multiple units
- PAM-1 is good for very closely related sequences, PAM-250 for intermediate and PAM-1000 for very distant

Scoring protein alignments : BLOSUM

BLOck **SU**bstitution **M**atrix

- Based on comparisons of Blocks of sequences derived from the Blocks database (derived from Prosite)
- The Blocks database contains multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins
- BLOSUM matrices are categorized by sequence identity above which blocks were clustered (i.e. BLOSUM62 is derived from blocks clustered at 62% sequence identity)
- Focused on highly conserved regions



Scoring protein alignments : BLOSUM vs. PAM

Matrix	Entropy	Expected score
BLOSUM30	0.1424	-0.1074
BLOSUM35	0.2111	-0.1550
BLOSUM40	0.2851	-0.2090
BLOSUM45	0.3795	-0.2789
BLOSUM50	0.4808	-0.3573
BLOSUM55	0.5637	-0.4179
BLOSUM60	0.6603	-0.4917
BLOSUM62	0.6979	-0.5209
BLOSUM65	0.7576	-0.5675
BLOSUM70	0.8391	-0.6313
BLOSUM75	0.9077	-0.6845
BLOSUM80	0.9868	-0.7442
BLOSUM85	1.0805	-0.8153
BLOSUM90	1.1806	-0.8887

Matrix	Entropy	Expected score
PAM-10	3.430	-8.270
PAM-20	2.950	-6.180
PAM-30	2.570	-5.060
PAM-40	2.260	-4.270
PAM-50	2.000	-3.700
PAM-60	1.790	-3.210
PAM-70	1.600	-2.770
PAM-80	1.440	-2.550
PAM-90	1.300	-2.260
PAM-100	1.180	-1.990
PAM-120	0.979	-1.640
PAM-140	0.820	-1.350
PAM-160	0.694	-1.140
PAM-180	0.591	-1.510
PAM-200	0.507	-1.230
PAM-250	0.354	-0.844
PAM-300	0.254	-0.835
PAM-350	0.186	-0.701

Scoring protein alignments : BLOSUM vs. PAM

Equivalent PAM and BLOSUM matrices based on relative entropy

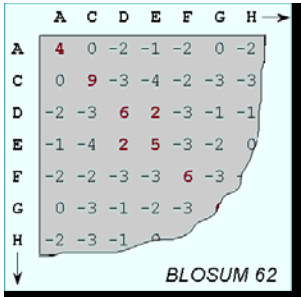
PAM100 <==> Blosum90
PAM120 <==> Blosum80
PAM160 <==> Blosum60
PAM200 <==> Blosum52
PAM250 <==> Blosum45

- PAM matrices have lower expected scores for the BLOSUM matrices with the same entropy
- BLOSUM matrices “generally perform better” than PAM matrices

Simple sequence search : scoring gaps

```
AATCTATA        AATCTATA        AATCTATA
AAG-AT-A        AA-G-ATA        AA--GATA
```

- Gap should correspond to insertion/deletion (indel) even in evolution
- Multiple (block) nucleotide indels are common as single nucleotide indels
- It is then more probable that fewer indel events occurred, i.e. gaps should be grouped
- Gaps are scored negatively (penalty)
- Two scores for gaps: origination and continuation
- Origination score > continuation score



	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	[illegible]
G	0	-3	-1	-2	-3	[illegible]	[illegible]
H	-2	-3	-1	0	[illegible]	[illegible]	[illegible]

BLOSUM 62

Substitution Matrix and Gap Cost

Query Length	Substitution Matrix	Gap cost
<35	PAM-30	(9,1)
35-50	PAM-70	(10, 1)
50-85	BLOSUM-80	(10, 1)
>85	BLOSUM-62	(11, 1)

Simple sequence search - alignment

- Direct enumeration impossible: 100 vs. 95 with 5 gaps = ~55 million choices
- Optimal solution comes from Dynamic Programming: extending solution to *n* based on all optimal solutions for *n-1* problems (*Needleman-Wunsh*)
- Solution is a path in the Dynamic Programming score table

		A	C	T	C	G
	0	-1	-2	-3	-4	-5
A	-1					
C	-2					
A	-3					
G	-4					
T	-5					
A	-6					
G	-7					

- Initiate table with gap penalties (1,1)
- Fill table top-left to low-right
- Fill element with maximum value of

  = take left cell add gap penalty

  = take upper cell add gap penalty

  = take diagonal cell add score

Simple sequence search - alignment

- This alignment uses identity scoring table with (1,1) gaps
- Aligns full sequences: global alignment

ACAGTAG

AC--TCG

		A	C	T	C	G
	0	-1	-2	-3	-4	-5
A	-1					
C	-2					
A	-3					
G	-4					
T	-5					
A	-6					
G	-7					

		A	C	T	C	G
	0	-1	-2	-3	-4	-5
A	-1	1	0	-1	-2	-3
C	-2	0	2	1	0	-1
A	-3	-1	1	2	1	0
G	-4	-2	0	1	2	2
T	-5	-3	-1	1	1	2
A	-6	-4	-2	0	1	1
G	-7	-5	-3	-1	0	2

		A	C	T	C	G
	0	-1	-2	-3	-4	-5
A	-1	**1**	0	-1	-2	-3
C	-2	0	**2**	1	0	-1
A	-3	-1	**1**	2	1	0
G	-4	-2	**0**	1	2	2
T	-5	-3	-1	**1**	1	2
A	-6	-4	-2	0	**1**	1
G	-7	-5	-3	-1	0	**2**

Simple sequence search - alignment

- Global alignment is not useful when searching databases
- Semiglobal alignment: terminal gaps allowed
- Achieved by initializing gaps to zero in the first step and allowing no gap penalties in the last row/column

		A	C	G	T	C
	0	-1	-2	-3	-4	-5
A	-1	1	0	-1	-2	-3
A	-2	0	0	-1	-2	-3
C	-3	-1	1	0	-1	-1
A	-4	-2	0	0	-1	-2
C	-5	-3	-1	-1	-1	0
G	-6	-4	-2	0	-1	-1
G	-7	-5	-3	-1	-1	-2
T	-8	-6	-4	-2	0	-1
G	-9	-7	-5	-3	-1	-1
T	-10	-8	-6	-4	-2	-2
C	-11	-9	-7	-5	-3	-1
T	-12	-10	-8	-6	-4	-2

```
AACACGGTGTCT
-A-C-G-TC---

AACACGGTGTCT
---ACG-TC---
```

		A	C	G	T	C
	0	0	0	0	0	0
A	0	1	0	-1	-1	0
A	0	1	0	-1	-2	0
C	0	0	2	1	0	0
A	0	1	1	1	0	0
C	0	0	2	1	0	1
G	0	-1	1	3	2	1
G	0	-1	0	2	2	1
T	0	-1	-1	1	3	2
G	0	-1	-2	0	2	2
T	0	-1	-2	-1	1	2
C	0	-1	0	-1	0	2
T	0	0	0	0	0	2

Simple sequence search - alignment

- Local alignment: best subsequence matching
- Dynamic programming algorithm for local alignment: *Smith-Waterman*
- Starts like semiglobal alignment with fourth option for filling table:

  = place 0 in the cell when maximum possible value is negative
- Start with the cell with maximum score

		G	C	G	A	T	A	T	A
	0	0	0	0	0	0	0	0	0
A	**0**	-1	-1	-1	1	0	1	0	1
A	0	**-1**	-2	-2	0	0	1	0	1
C	0	-1	**0**	**-1**	-1	-1	0	0	1
C	0	-1	0	-1	**-2**	-2	-1	-1	1
T	0	-1	-1	-1	-2	**-1**	-2	0	1
A	0	-1	-2	-2	0	-1	**0**	-1	1
T	0	-1	-2	-3	-1	1	0	**1**	1
A	0	-1	-2	-3	-2	0	2	1	**2**
G	0	1	0	-1	-2	-1	1	1	**2**
C	0	0	2	1	0	-1	0	0	**2**
T	0	0	1	1	1	1	1	1	**2**

```
AAC-CTATAGCT
-GCGATATA---
```

```
AACCTATAGCT
GCGATATA
```

		G	C	G	A	T	A	T	A
	0	0	0	0	0	0	0	0	0
A	0	0	0	0	1	0	1	0	1
A	0	0	0	0	1	0	1	0	1
C	0	0	1	0	0	0	0	0	1
C	0	0	1	0	0	0	0	0	1
T	0	0	0	0	0	**1**	0	1	1
A	0	0	0	0	1	0	**2**	1	2
T	0	0	0	0	0	2	1	**3**	2
A	0	0	0	0	1	1	3	2	**4**
G	0	1	0	1	0	0	2	2	4
C	0	0	2	1	0	0	1	1	4
T	0	0	1	1	1	1	1	2	4

The BLAST Search Algorithm

query word (W = 3)

```
Query:  GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL
```

neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

neighborhood score threshold (T = 13)

```
Query:  325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
            +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct:  290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
```

High-scoring Segment Pair (HSP)

	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	[illegible]
G	0	-3	-1	-2	-3	[illegible]	
H	-2	-3	-1	0			

BLOSUM 62

FASTA search algorithm

- Breaks up query sequence into words (like BLAST)
- Using lookup tables with words finds areas of identity
- Areas of identity are joint to form larger pieces
- Full Smith-Waterman algorithm is used to align these pieces

- FASTA is slower than BLAST, but produces optimal alignment for pieces

Bit Score and E-value

Bit Score: $S' = (\lambda S - \ln K)/\ln 2$

Expect Value: $E = mn\, 2^{-S'}$

E=0.01 -> 1% chance that the match is due to a random match
E value depends on database size
E value: expected number of HSPs with score S or higher

P value: probability of finding zero HSPs with score S or higher

$$P = 1 - \exp(-E)$$

Programs and Database selection

1. nucleotide sequence: blastn
Query: nucleotide sequence
Database: nucleotide sequence database

e.g. nt htg est

Programs and Database selection

2. protein sequence: blastp
Query: protein sequence
Database: protein sequence database

e.g. nr

Programs and Database selection

3. translated blast search:

blastx

nucleotide sequence -> protein database

tblastn

protein sequence -> nucleotide database

tblastx

nucleotide sequence->nucleotide

Programs and Database selection

Protein sequence alignment is more sensitive than nucleotide sequence alignment !

Filtering the low complexity and repetitive sequences

1. **Low complexity: DUST and SEG programs**

2. **Repetitive sequences: RepeatMasker**

(DNA sequences: "NNNNNNNN")
(Protein sequences: "XXXXXXXXXX")

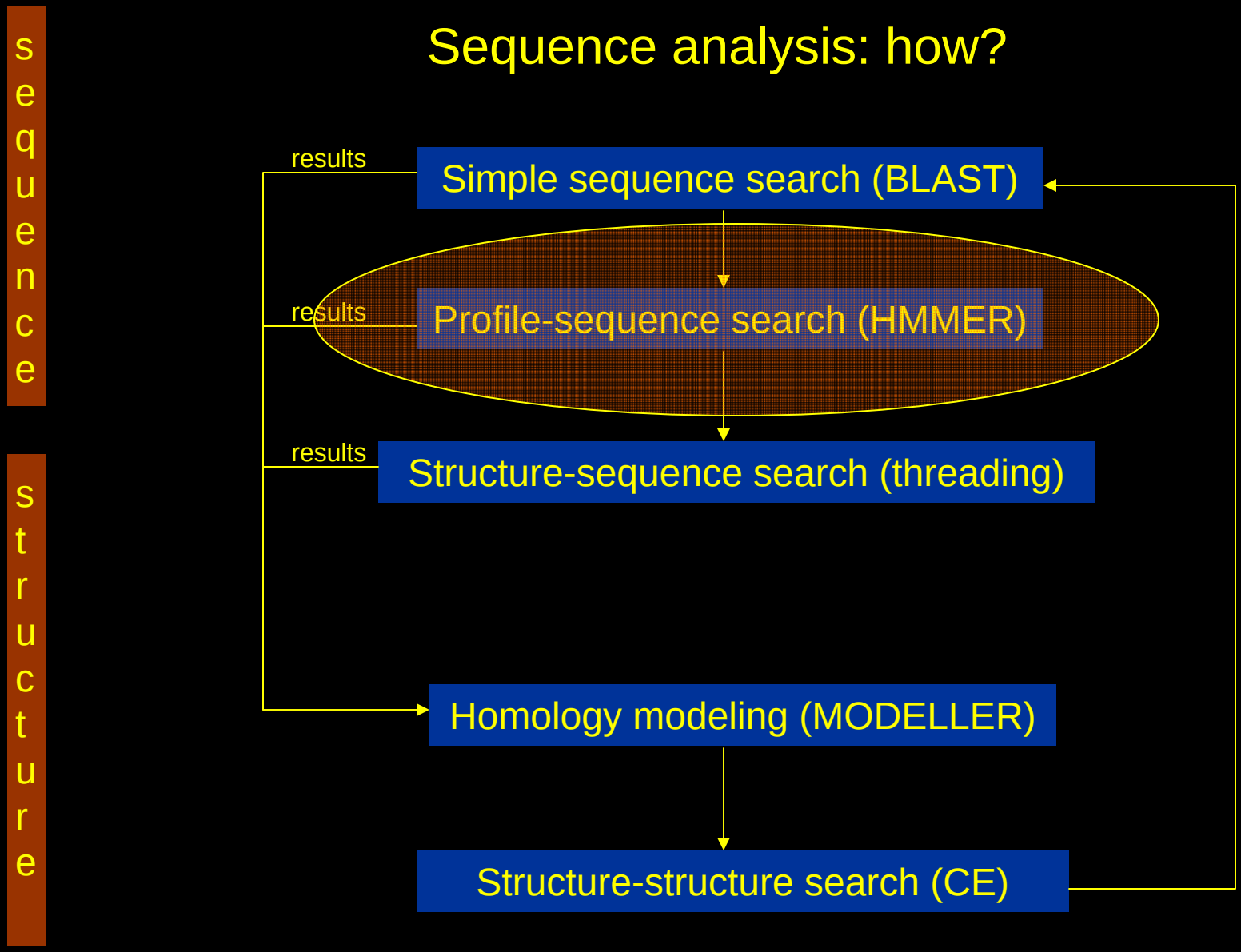
BLAST Servers

1. NCBI http://www.ncbi.nlm.nih.gov/BLAST/

2. Batch Blast http://cbsuapps.tc.cornell.edu/cbsu/blast_s.aspx

Input files: Fasta format sequence files
Output files:
1. standard
2. -m 8 format
3. CBSU parsed format
4. CBSU parsed format 2

Sequence analysis: how?



Scoring system of BLAST

```
Query:   ACCGGEFFGACD
         || |||| ||
Target:  ACGGGCFCGAGG
Score:   49366462631
```

	A	C	D	E	F	G	H →
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	[illegible]
G	0	-3	-1	-2	-3	[illegible]	
H ↓	-2	-3	-1	0			

BLOSUM 62

Sequence alignment of domain X

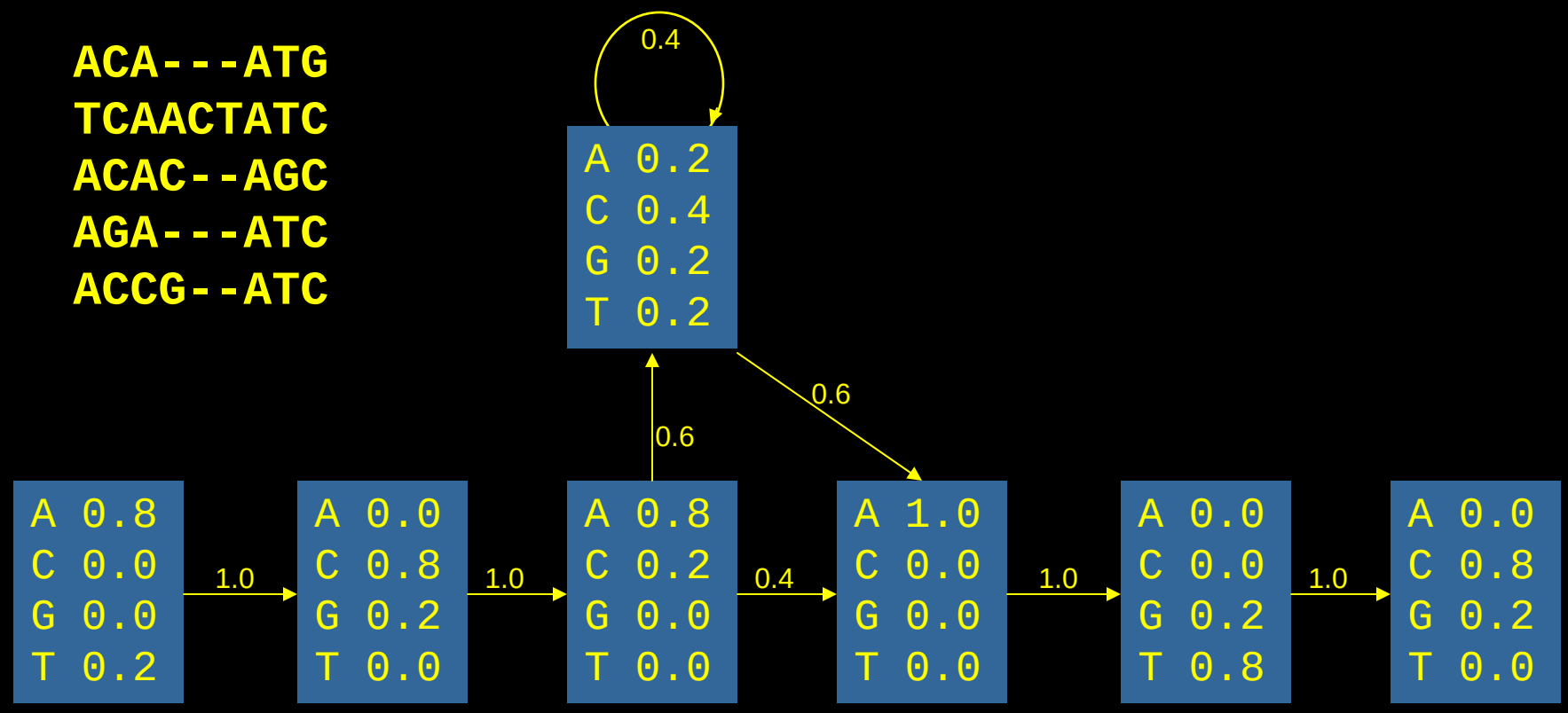
```
ACHGGEFFGAC
ACCGGCFCGAG
ACACCEFFCAC
ACACTCFFGAC
ACLGPEFFGAC
```

	A	C	G	H	S
1	1.0	0.0	0.0	0.0	0.0
2	0.0	1.0	0.0	0.0	0.0
3	0.4	0.2	0.0	0.2	0.0
4	0.0	0.4	0.4	0.0	0.0
..	..	..	..	..	..

	A	C	G	H	S
1	100	-100	-100	-100	-100
2	-100	100	-100	-100	-100
3	50	10	-50	10	-50
4	-60	60	60	-60	-60
..	..	..	..	..	..

What is Hidden Markov Model?

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

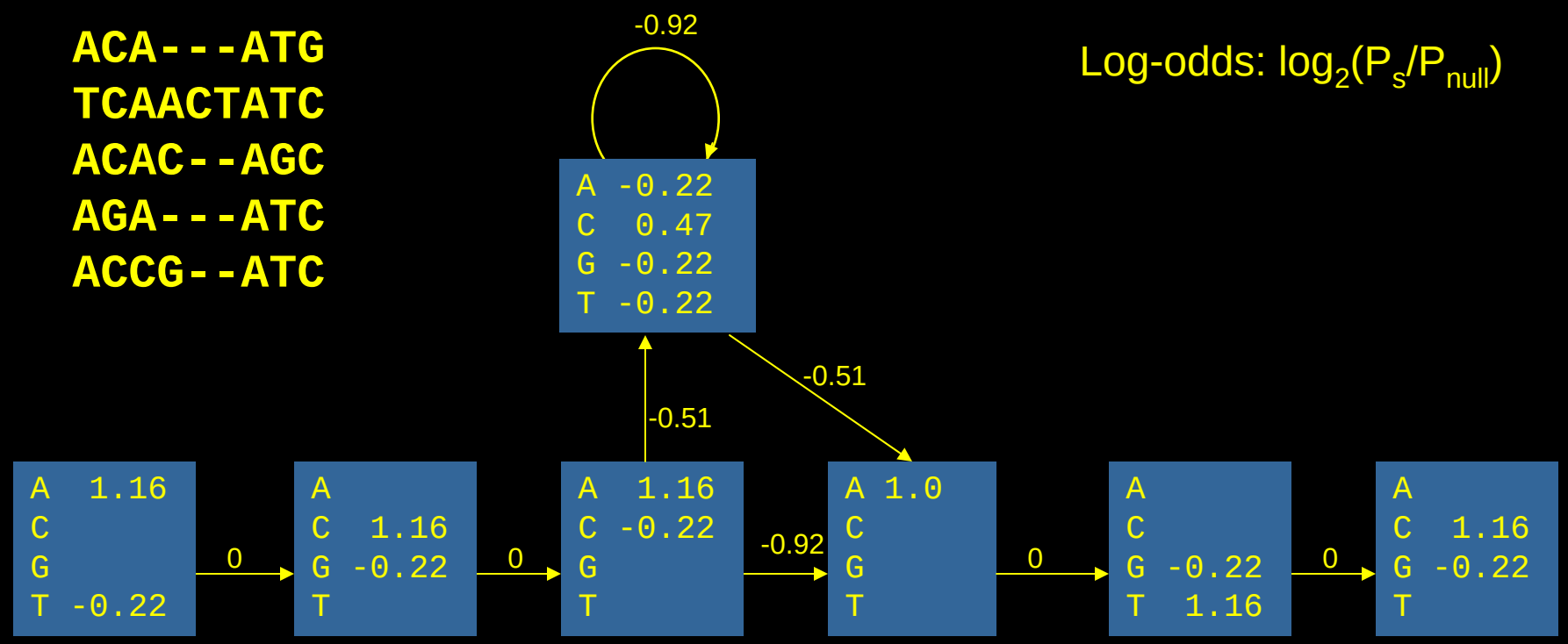


$P(ACACATC)=0.8\times1.0\times0.8\times1.0 \ldots \times0.8=4.7\times10^{-2}$

What is Hidden Markov Model?

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

Log-odds: $\log_2(P_s/P_{null})$



Log-odds(ACACATC)=1.16+0+1.16+0 … +1.16=6.64

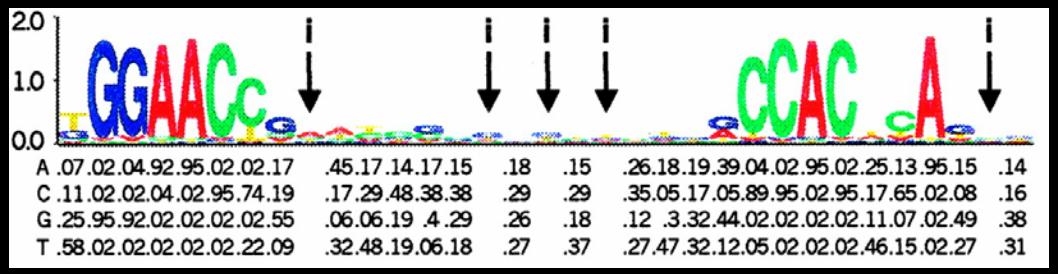
What is Hidden Markov Model?

```
ACA---ATG
TCAACTATC
ACAC--AGC
AGA---ATC
ACCG--ATC
```

	Sequence	P %	Log odds
Consensus	ACAC--ATC	4.7	6.7
	ACA---ATG	3.7	4.9
	TCAACTATC	0.0075	3.0
	ACAC--AGC	1.2	5.3
	AGA---ATC	3.3	4.9
	ACCG--ATC	0.59	4.6
Bad sequence	TGCT--AGG	0.0023	-0.97

```
40ptsH      TTTTGTGGCCTGCTTCAAACTT
41ptsH      TTTTATGATTTGGTTCAATTCT
42rhaS      AATTGTGAACATCATCACGTTC
43rot       TTTTGTGATCTGTTTAAATGTT
```

Alignment



Model

GTGTGATCAGAGTGATTGTGTCAGTGTGTAGCGCTCTGTT
TCGTGTGTTTGTGTTCATTTATTGTGTTGT GGCTTCTCATT
GCCCCTTTGGTTCTGTTCTTAAACCTTCATCTTCGCTTAGT
AAAGTTAGATTCCACCGA TCCGTTTCTGTTA AAGAAAAAG
TGATCAACAAACTTCAAGAAAATCTAAATGTGCAGTAATTT
GAAATTTATGCTTATTGTGT

Search for matches

HMM model table

```
HMMER2.0  [2.3.2]
NAME  AA_kinase
ACC   PF00696.17
DESC  Amino acid kinase family
LENG  318
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   hmmbuild -F HMM_ls.ann SEED.ann
COM   hmmcalibrate --seed 0 HMM_ls.ann
NSEQ  108
DATE  Tue Feb 21 02:42:42 2006
CKSUM 7209
GA    -40.0 -40.0
TC    -39.2 -39.2
NC    -40.5 -40.5
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT      -4  -8455
NULE     595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201    384  -1998   -644
EVD   -134.910873   0.147785
HMM         A      C      D      E      F      G      H      I      K      L      M      N      P      Q      R      S      T      V      W      Y
          m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
           -18      *  -6337
     1   -442  -4997   -726    -30  -5318   -606  -3157  -2580   2335  -2272   2946   -718  -4591    898    966   -637  -2003  -4619  -5180   -542     1
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11485 -12527   -894  -1115   -701  -1378    -18      *
     2  -3924  -3758  -6216  -2272     84  -2772  -4334    150    958    254   2151  -5094   -246  -4823   2414  -4548   1095   -766   1517  -1497     2
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
     3  -1251    -80  -6262  -2333   -382  -5472   -800   2750  -2121     29   -310  -5115  -5522  -1594   -675  -1992     -4   1889     -8  -3869     3
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
     4  -2336  -5399  -8620  -8305  -5890  -8502  -8549   1204  -8290    133   -519  -8157  -8176  -8175  -8444  -7901  -5913   3484  -7736  -7303     4
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
     5  -2325   -602  -6272  -5636    441  -5474    328   2498  -5231    835    606  -5120  -5524   1214  -5031  -2377  -3867   1824  -4211  -1401     5
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
     6    858  -5563  -1423  -5664  -7743  -5588  -5944  -7511   3703  -7583   -760  -5461  -6266  -5679  -5766  -1043  -2248  -6499  -7688  -7349     6
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
     7  -6142    229  -8586  -8089   2523  -8180  -6968   1713  -7853   1859  -3664  -7805  -7776  -7152  -7599  -7422  -6068   1024  -6130   1322     7
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
     8  -5620  -6116  -8303  -8671  -8641   3699  -7887  -8683  -8615  -8808  -7964  -7205  -7115  -8115  -8210    549  -6093  -7417  -8370  -8783     8
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
     9  -4637  -5247  -7901  -8257  -7980   3590  -7294  -7815  -7975  -8072  -7095  -6356  -6342  -7394  -7582    543    132  -6473  -8186  -8186     9
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
    10    746  -5180  -3527    790  -5520  -2312    301  -5269  -2960  -5219  -4298   1891  -4788  -1285  -3470   2074   1689  -1565  -5391  -4708    10
     -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394     45     96    359    117   -369   -294   -249
     -     -1 -11609 -12651   -894  -1115   -701  -1378      *      *
```

PSI-BLAST

Position-Specific Iterative BLAST

BLAST search

↓

Align the sequences of the blast targets

↓

Construct profile from the blast targets

↓

Modify substitution matrix to fit profile

↓

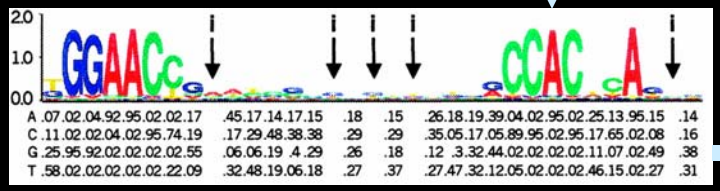
Search the database with the new scoring

PSI-BLAST uses position-dependent substitution matrix instead of probabilities (HMM)

```
40ptsH      TTTTGTGGCCTGCTTCAAACTT
41ptsH      TTTTATGATTTGGTTCAATTCT
42rhaS      AATTGTGAACATCATCACGTTC
43rot       TTTTGTGATCTGTTTAAATGTT
```

Sequence alignment

hmmbuild



Model

hmmsearch

More sequence motifs that fit this model

Programs:	Databases:
HMMER	PFAM http://pfam.wustl.edu/
SAM	SMART http://smart.embl-heidelberg.de/
PSI-BLAST	COG http://www.ncbi.nlm.nih.gov/COG/
	Superfamily http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/

Web based programs:

PFAM: http://pfam.wustl.edu/hmmsearch
An HMM library based on the Swissprot 48.9 and SP-TrEMBL 31.9 protein sequence databases. 8296 protein families in current version.

SMART: http://smart.embl-heidelberg.de/
More than 500 extensively annotated domain families

InterProScan: http://www.ebi.ac.uk/interpro/scan.html
Combines many HMM and other methods

The input and output:

MLYQLSKATTRIRLKRQKAVPQHRWLWSLAFLAAFTLKVSERANKNMAKTHNSGDVRCADLAI
SIPNNPGLDDGASYRLDYSPPFGYPEPNTTIASREIGDEIQFSRALPGTKYNFWLYYTNFTHHD
WLTWTVTITTAPDPPSNLSVQVRSGKNAIILWSPPTQGSYTAFKIKVLGLSEASSSYNRTFQVN
DNTFQHSVKELTPGATYQVQAYTIYDGKESVAYTSRNFTTKPNTPGKFIVWFRNETTLLVLWQ
PPYPAGIYTHYKVSIEPPDANDSVLYVEKEGEPPGPAQAAFKGLVPGRAYNISVQTMSEDEISL
PTTAQYRTVPLRPLNVTFDRDFITSNSFRVLWEAPKGISEFDKYQVSVATTRRQSTVPRSNEPV
AFFDFRDIAEPGKTFNVIVKTVSGKVTSWPATGDVTLRPLPVRNLRSINDDKTNTMIITWEADPA
STQDEYRIVYHELETFNGDTSTLTTDRTRFTLESLLPGRNYSL

	Model	Seq-from	Seq-to	HMM-from	HMM-to	Score	E-value	Alignment	Description
!!	fn3	139	221	1	84	58.1	1.2e-14	glocal	Fibronectin type III domain
!!	fn3	233	317	1	84	59.4	5.1e-15	glocal	Fibronectin type III domain
!!	fn3	328	410	1	84	36.3	4.4e-08	glocal	Fibronectin type III domain
!!	fn3	421	501	1	84	58.4	9.8e-15	glocal	Fibronectin type III domain
!!	fn3	512	591	1	84	27.0	3e-05	glocal	Fibronectin type III domain
!!	fn3	599	677	1	84	78.9	6.9e-21	glocal	Fibronectin type III domain
!!	fn3	689	778	1	84	40.8	2e-09	glocal	Fibronectin type III domain
!!	fn3	789	869	1	84	14.8	0.0063	glocal	Fibronectin type III domain
!!	fn3	880	955	1	84	67.6	1.7e-17	glocal	Fibronectin type III domain
!!	fn3	974	1060	1	84	58.4	1e-14	glocal	Fibronectin type III domain
!!	Y_phosphatase	1312	1542	1	274	393.6	1.3e-115	glocal	Protein-tyrosine phosphatase

Evaluating the significance of a hit:

1. E-value: <= 0.1
(10% chance that you would've seen a hit this good in a search of random sequences)

2. Raw score >= GA (the scores used as cutoffs in constructing Pfam, you may consider TC and NC as well)

3. Raw score > $\log_2$(number of sequences in the database)
(20 for the nr)