BIOINFORMATICS CENTRE
UNIVERSITY OF COPENHAGEN

BLAST, gapped BLAST and PSI-BLAST

Overview

- Quick repetition of pairwise alignment
- Statistical fundament of BLAST
- The basic BLAST algorithm
- Gapped BLAST
- PSI-BLAST

Pairwise alignment

- Global alignment
  - Needleman-Wunsch
- Local Alignment
  - Smith-Waterman
- Dynamic programming, fill out matrix
- Find optimal solution
- Time complexity: *O(mn)*

BLAST

- Heuristic method
- Fast search through large databases
- Good but not necessarily best solution
- Skip explicit search of the entire matrix
- Extensions:
  - Faster
  - Include gaps
  - Use position-specific scoring matrices

Statistical fundament

Recall random walks and extreme value distributions

Assumption: Independet background distribution of amino acids, P_i

s_{ij} denote the score of aligning AAs *i* and *j*

The **expected score** must be negative

$$\sum_{i,j} P_i P_j s_{ij} < 0$$

(cf. random walks – else drift to infinity)

Statistical fundament

Given P_i and s_{ij},derive parameters *λ* and *K*

Let *S* be the nominal score of a sequence pair. The **normalized score** in *bits* is:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

The number *E* of **chance occurrences** of pairs giving *S'* is approximated by:

$$E = \frac{mn}{2^{S'}} \quad \text{or} \quad S' = \lg\left(\frac{mn}{E}\right)$$

Statistical fundament

The **target frequency** of aligned pairs:

$$q_{ij} = P_i P_j e^{\lambda_u s_{ij}} \quad \text{or} \quad s_{ij} = \frac{\ln(q_{ij} / P_i P_j)}{\lambda_u}$$

Note: Makes it possible to scale scores to fit desired frequencies q_{ij}

All this holds only when not using gaps

Expected to hold for gaps as well when costs are sufficiently large

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool

Brief repetition of basic BLAST:

Fast search for local ungapped alignments

W: Word size – find *W*-mers in target/query

T: Threshold – focus on pairs scoring >*T*

X: Drop-off – stop extending when loss >*X*

S: Score – the final score of segment pair

BLAST

Look for high scoring words of length *W*

Compile list *L* of all *W-mers* that score >*T* with some word in query sequence

Scan database for words in *L*

When some word found: *Extend alignment*

When score drops more than *X* below hitherto best score stop extension

Report all words with large score *S*

BLAST

If *W* too large: Too many words in *L*

- or too few

If *T* too large: Too restrictive search

- or too many extensions

Choose cut-off for relevant hits

High-scoring Sequence Pairs (*HSPs*)

It turns out that **>90%** computation time is in extending hits!

The Two-Hit Method

The goal: Faster algorithm

Reduce number of extensions

Observation:

- HSP much longer than W
- often contains more than one word-pair

Idea: Focus on two or more words on same diagonal

The Two-Hit Method

How to do it:

- For each hit, remember diagonal position
  - If overlapping: Ignore
  - If distance to previous hit < *A*: Extend
- Must lower *T* to get same sensitivity
  - Many more single hits
  - Only a few are extended due to diagonal constraint

The Two-Hit Method

Evaluation:

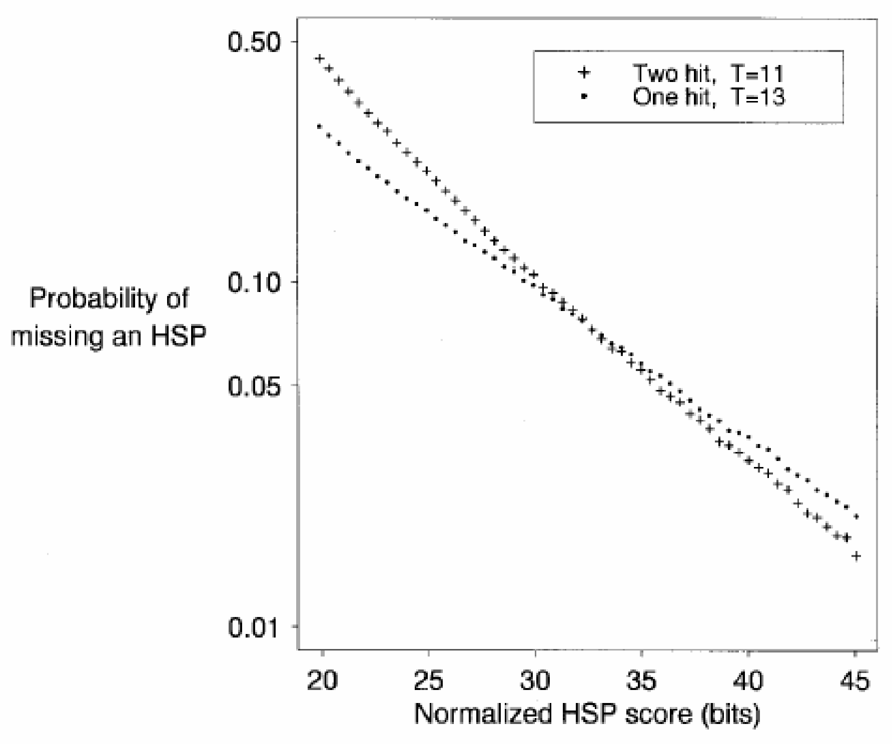
Generate 100,000 model HSPs

Check for hits > *T*

W=3, T=13

W=3, T=11, A=40

For S≥33, two-hits more sensitive



The Two-Hit Method

Test on real data:

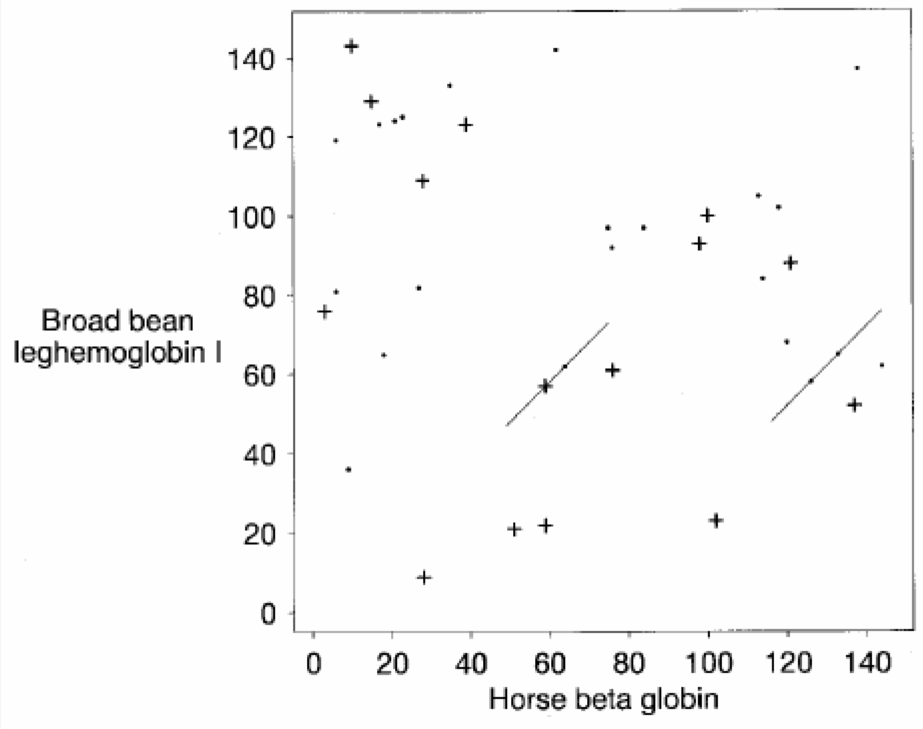
15 hits with T≥13 (+)

Additional:

22 hits with T≥11 (•)

One-hit extends all 15

Two-hit extends 2 pairs



More hits, fewer extensions

Gapped BLAST

How to let BLAST find gapped alignments?

Already implicit ’gapped’ alignment:

When more HSPs in same sequence → assess combined result

If one HSP is missed the combined result might be missed, too

Lower T needed – large execution time

Gapped BLAST

New idea: Introduce a moderate score S_g

If *HSP* exceeds S_g start gapped extension

Choose S_g to trigger ~ 1 extension per 50 sequences in database ($S_g \approx 22$ bits)

Costly operation but few of them

Gapped extension based on a single HSP – we may tolerate missing more HSPs

Raise T

Gapped BLAST

New algorithm:

Two-hit method: Two words of score ≥T trigger ungapped extension

If *HSP* scores ≥ S_g, start gapped extension

Report final alignment if significant (low E-value)

Gapped BLAST

How to construct gapped local alignment?

Standard way:

Limit search to a banded matrix

- Gapped extension may stray outside band

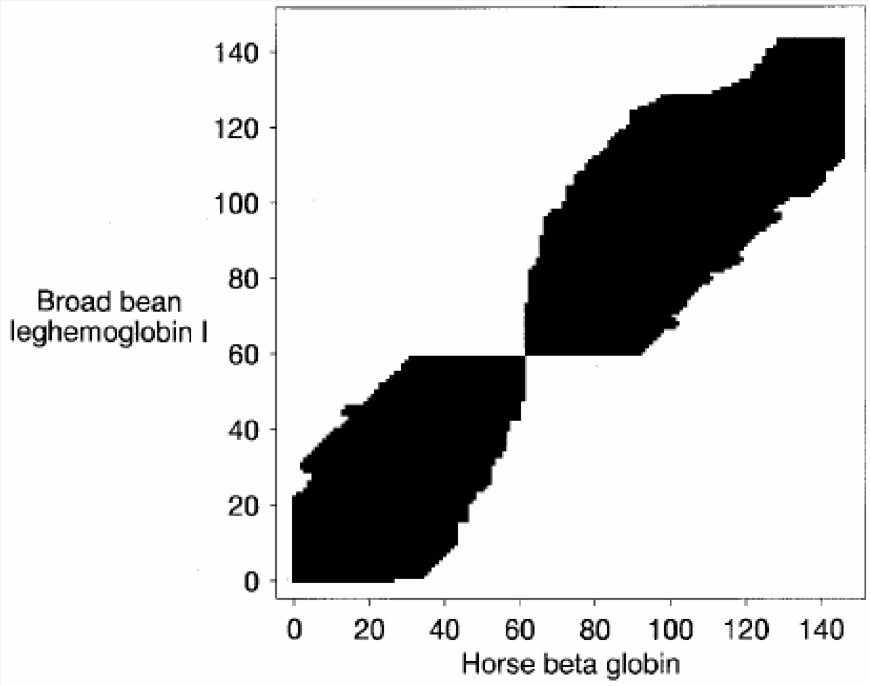
Instead use standard BLAST procedure:

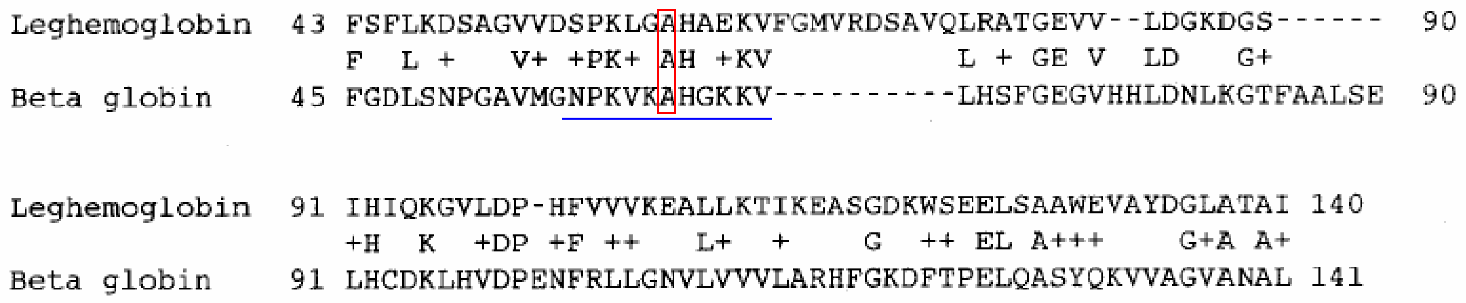
Look in cells where the score drops no more than X_g

But how to begin…

Gapped BLAST

- Use central aligned pair as seed
- Heuristic: Find length-11 segment with highest score. Use central pair
- Extend forward and backward





```
Leghemoglobin  43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
                  F  L +   V+ +PK+ AH +KV          L + GE V  LD   G+
Beta globin    45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90


Leghemoglobin  91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                  +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin    91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

Gapped BLAST

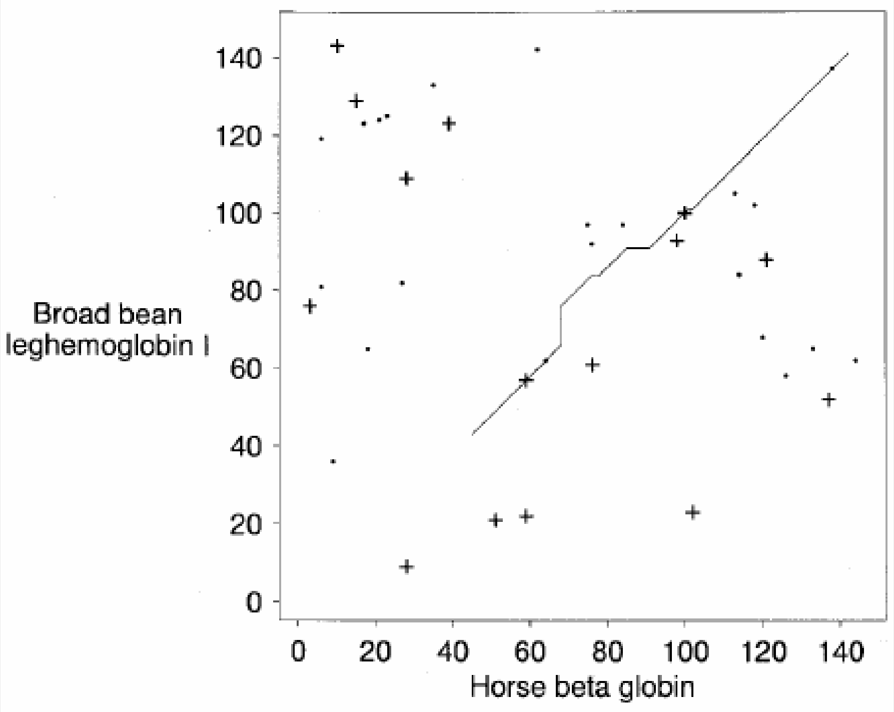
This result is not found by standard BLAST

Combined result of first and last HSP gives E-value 31

Gapped BLAST:

S=32.4 bits

E=0.54

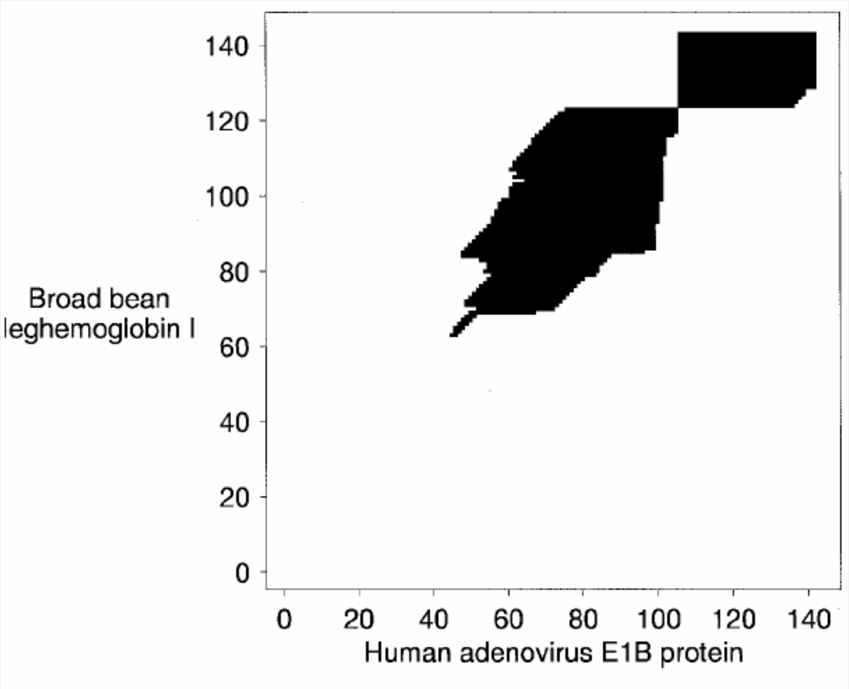


Gapped BLAST

What about spurios hits? Does that give extra work?

No: Score decays fast

Small part of matrix explored



Gapped BLAST

New version faster. Relative times:

	BLAST	Gapped BLAST
Overhead	8 (8%)	8 (24%)
Extend?	-	12 (37%)
Ungapped	92 (92%)	5 (15%)
Gapped	-	8 (24%)
	100	33

Gapped BLAST

What about the parameters λ and K?

Cannot be estimated during execution since BLAST looks at only some sequences

No theory covers gapped alignments

Use estimations made in advance

Drawback: Cannot use arbitrary scoring systems

PSI-BLAST

Position **S**pecific **I**terated BLAST

Use sequence information to build position-specific scoring matrices

Readers of X-Men will know that psy blasts are something else entirely …

PSI-BLAST

More sensitive procedure

Each iteration a little slower

Issues:

i) Architecture of score matrix

ii) Construction of multiple alignment

iii) Sequence weights

iv) Target frequencies and scores

v) Applying BLAST to scoring matrices

PSI-BLAST: Architecture

Automated generation is difficult

Boundaries, many motifs, subsets…

1) Length of query determines dimensions
2) No position-specific gap cost
   - No theory for deriving gap costs from M
   - Estimate statistical significance

So they build a L×20 scoring matrix

PSI-BLAST: Constructing M

Collect BLAST output with *E<0.01*

Remove similar sequences

- Sequences identical to query segments
- Only one copy of sequences >98% similarity

Local alignment → varying number of sequences per column

No true multiple alignment methods

PSI-BLAST Constructing M

Reduce *M* to M_C:

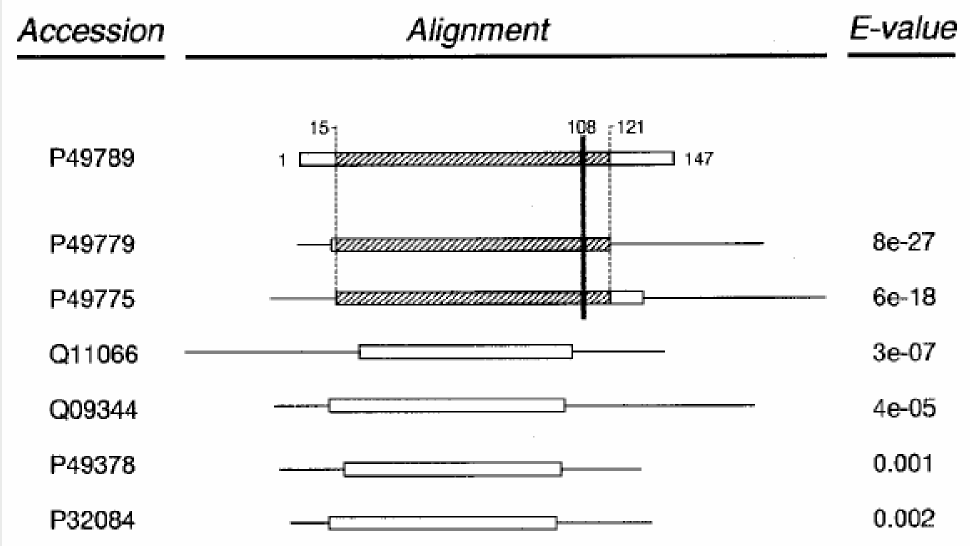
Treat columns independently

For each column *C*: Let *R* be the set of sequences with a residue in *C*

The columns of M_C:

Columns from *M* with all sequences in *R*

Now: Characters in all positions



PSI-BLAST: Weights

Weighting needed to avoid bias

Many methods, roughly same results

- Voronoi, maximum entropy, *position based*

Information content

N_C: Number of independent observations

Simple estimate: Mean number of different residues in each column

PSI-BLAST: Target frequencies

Many methods for creating scoring matrices

Good theoretical foundation:

$$\log(Q_i/P_i)$$

P_i: Background. How to estimate Q_i?

Pseudocount frequencies g_i for column C

f_j: Observed frequency, q_{ij}: Implicit target (7)

$$g_i = \sum_j \frac{f_j}{P_j} q_{ij}$$

PSI-BLAST: Target frequencies

Weight observed and pseudocount freq’s

$\alpha = N_C - 1$ $\beta = 10$ (empirical)

Now Q_i is given as:

$$Q_i = \frac{\alpha f_i + \beta g_i}{\alpha + \beta}$$

This makes it possible to build a matrix

But how to use it with BLAST…

PSI-BLAST: Application

Minor modifications to

- find words in query matrix
- find hits
- extend hits (gapped and ungapped)

But what about the parameters T and X_g?

Test whether the scale λ_u of the matrix corresponds to the scale of s_{ij}

If similar: Probably the same scale λ_g

PSI-BLAST: Application

Test the hypothesis:

Construct matrix by BLASTing length-567 influenza A virus hemagglutinin precursor

Compare to 10,000 random sequences

Plot local alignment score versus count

Fit best extreme value distribution

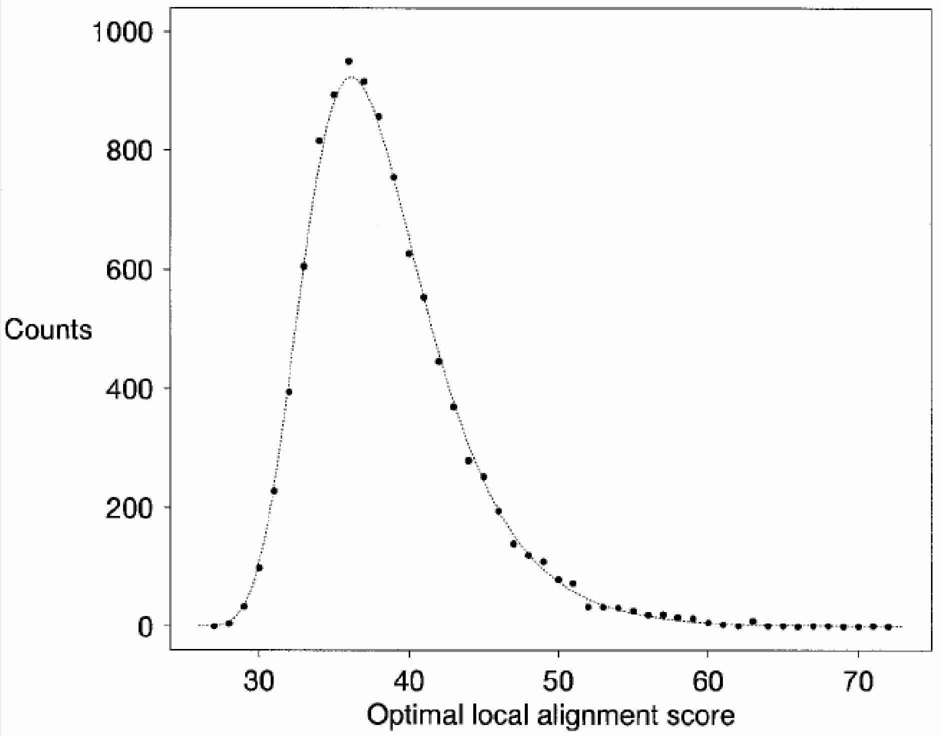
$$\lambda_g=0.251 \qquad K_g=0.031$$

PSI-BLAST: Application

Good fit to data

Supported by other experiments

Generally: Less than 2% deviation when using precomputed parameters



Testing PSI-BLAST

Create scoring matrix for 11 families

Compare to shuffled SWISS-PROT

Record:

- Lowest E-value
- No. of sequences with *E≤1* and *E≤10*

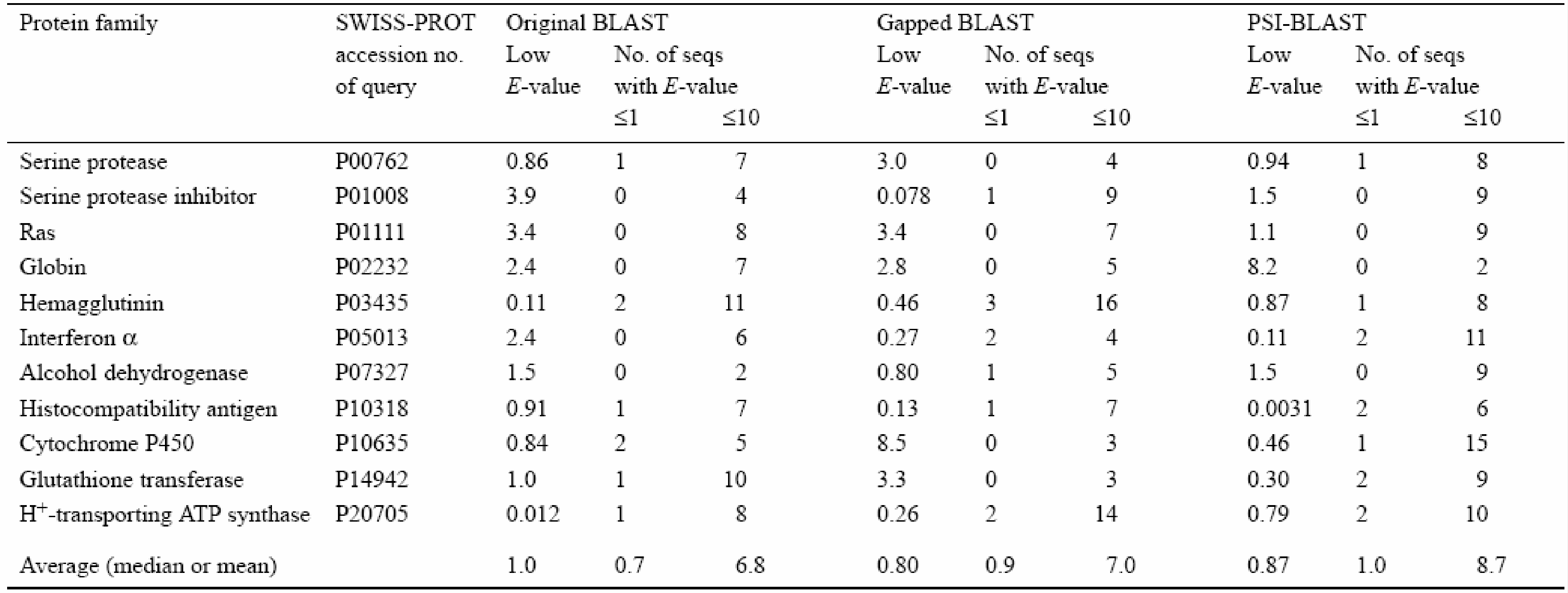
BLAST, gapped BLAST and PSI-BLAST

Testing PSI-BLAST

Within uncertainties of the theory

PSI-BLAST can automate the procedure

Beware of including used sequences



Protein family	SWISS-PROT	Original BLAST			Gapped BLAST			PSI-BLAST		
	accession no.	Low	No. of seqs		Low	No. of seqs		Low	No. of seqs	
	of query	*E*-value	with *E*-value		*E*-value	with *E*-value		*E*-value	with *E*-value	
			≤1	≤10		≤1	≤10		≤1	≤10
Serine protease	P00762	0.86	1	7	3.0	0	4	0.94	1	8
Serine protease inhibitor	P01008	3.9	0	4	0.078	1	9	1.5	0	9
Ras	P01111	3.4	0	8	3.4	0	7	1.1	0	9
Globin	P02232	2.4	0	7	2.8	0	5	8.2	0	2
Hemagglutinin	P03435	0.11	2	11	0.46	3	16	0.87	1	8
Interferon α	P05013	2.4	0	6	0.27	2	4	0.11	2	11
Alcohol dehydrogenase	P07327	1.5	0	2	0.80	1	5	1.5	0	9
Histocompatibility antigen	P10318	0.91	1	7	0.13	1	7	0.0031	2	6
Cytochrome P450	P10635	0.84	2	5	8.5	0	3	0.46	1	15
Glutathione transferase	P14942	1.0	1	10	3.3	0	3	0.30	2	9
H^+-transporting ATP synthase	P20705	0.012	1	8	0.26	2	14	0.79	2	10
Average (median or mean)		1.0	0.7	6.8	0.80	0.9	7.0	0.87	1.0	8.7

Testing PSI-BLAST

Compare sensitivity and speed of

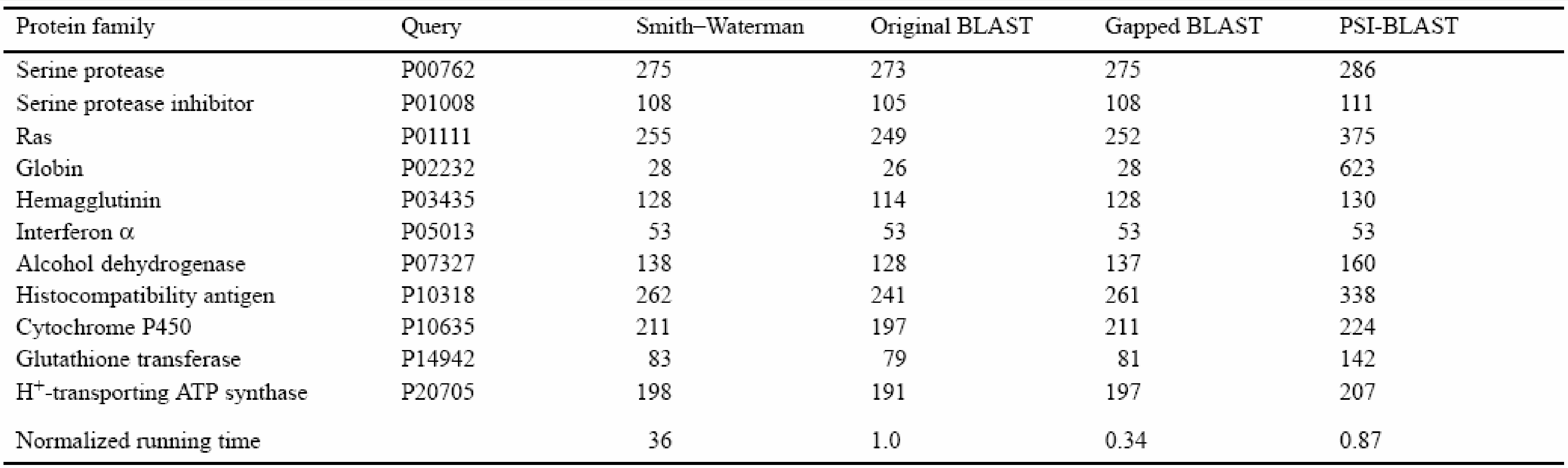
- Smith-Waterman
- Original BLAST
- Gapped BLAST
- PSI-BLAST (1 iteration)

Testing PSI-BLAST

All but one are true homologs

PSI-BLAST is faster and more sensitive

Other BLAST algorithms good as well



Protein family	Query	Smith–Waterman	Original BLAST	Gapped BLAST	PSI-BLAST
Serine protease	P00762	275	273	275	286
Serine protease inhibitor	P01008	108	105	108	111
Ras	P01111	255	249	252	375
Globin	P02232	28	26	28	623
Hemagglutinin	P03435	128	114	128	130
Interferon α	P05013	53	53	53	53
Alcohol dehydrogenase	P07327	138	128	137	160
Histocompatibility antigen	P10318	262	241	261	338
Cytochrome P450	P10635	211	197	211	224
Glutathione transferase	P14942	83	79	81	142
H^+-transporting ATP synthase	P20705	198	191	197	207
Normalized running time		36	1.0	0.34	0.87

Conclusions

- The two-hit method improves speed
- Gapped BLAST is fast
- PSI-BLAST finds weak homologs fast
- The theory can be extended

Future work

- Gap costs: Generalized affine gap cost
- Input scoring matrices to PSI-BLAST
  - Problems with parameters
- More refined multiple alignment
  - Use most significant hits
  - Rescore and realign sequences
  - Iterate