双序列比对

启发式算法

双序列比对的算法

- Dot Matrix，点阵法
- 动态规划算法：
  - Global: Needleman-Wunsch
  - Local: Smith-Waterman
- 启发式算法（基于Word or k-tuple）：FASTA, BLAST

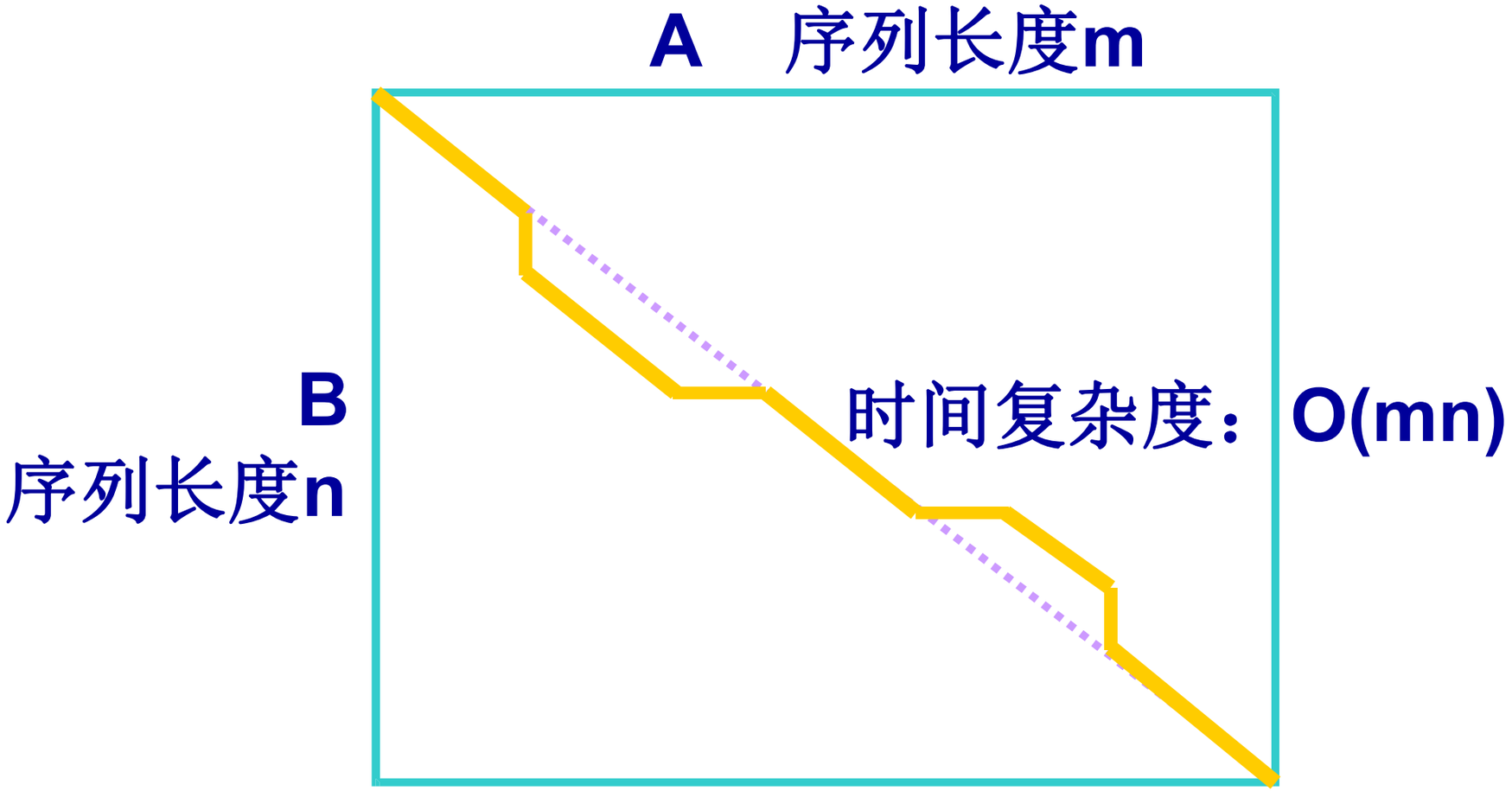
启发式算法Heuristic algorithms

- One of the major task of database mining is to search for homology of a query sequence against a large sequence database such as UniProt which involves millions of sequences. Although dynamic programming can provide accurate solution of alignment, it is too slow for large scale database searching.
- 当我们使用动态规划算法来搜索数据库时，由于数据库序列条目非常大，这时候直接使用动态规划算法，会非常慢。

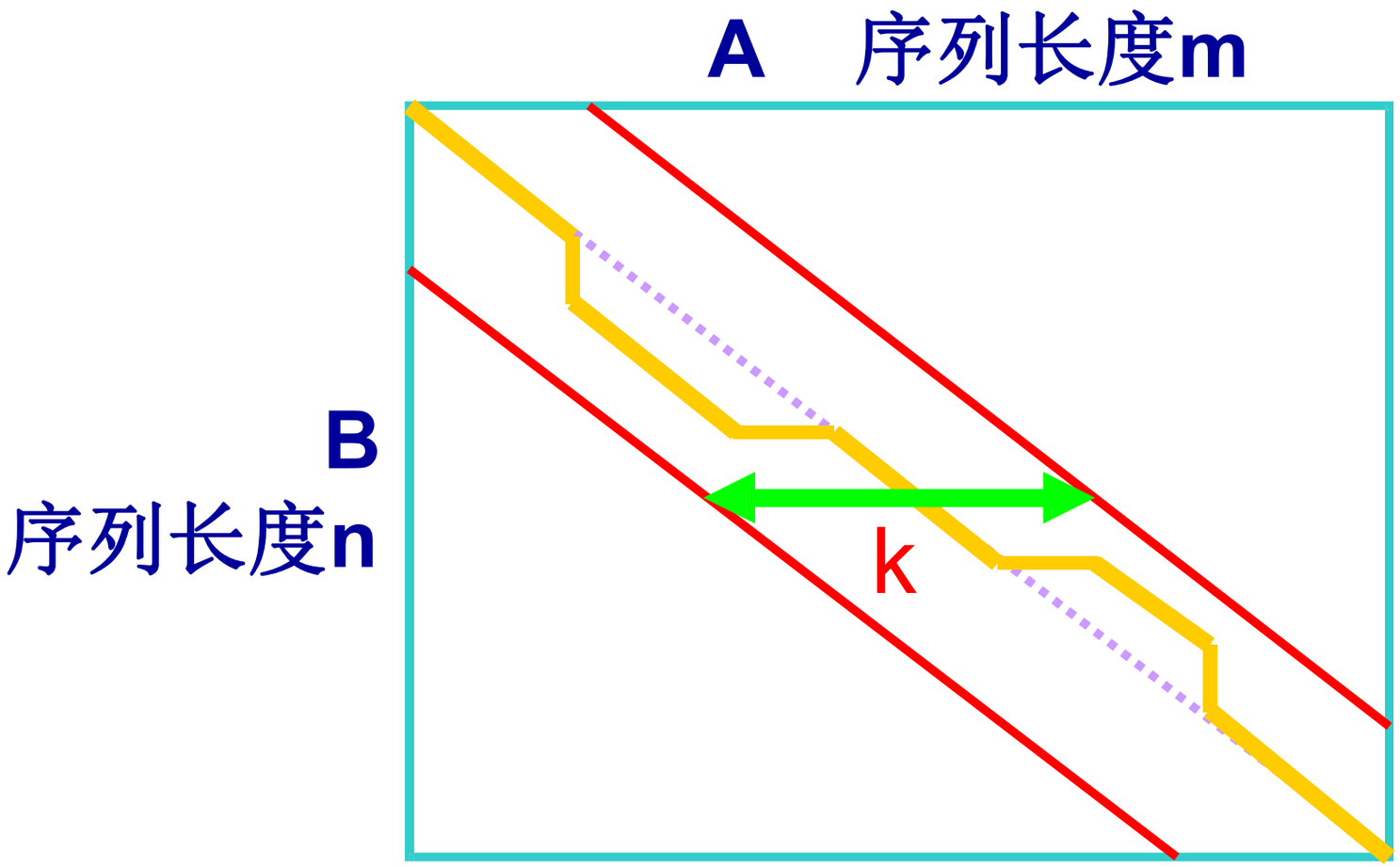
- Some heuristic algorithms, FASTA and BLAST, are designed to provide approximate alignment but with significantly increased speed (~50 times faster).
- 一些启发式算法，例如FASTA，BLAST 被提出。这些算法可以提供近似的比对结果，但速度显著提高，约提高50倍。

K-tup算法原理

- 对于两条序列A, B，若包含少量gap，则最优比对趋近对角线



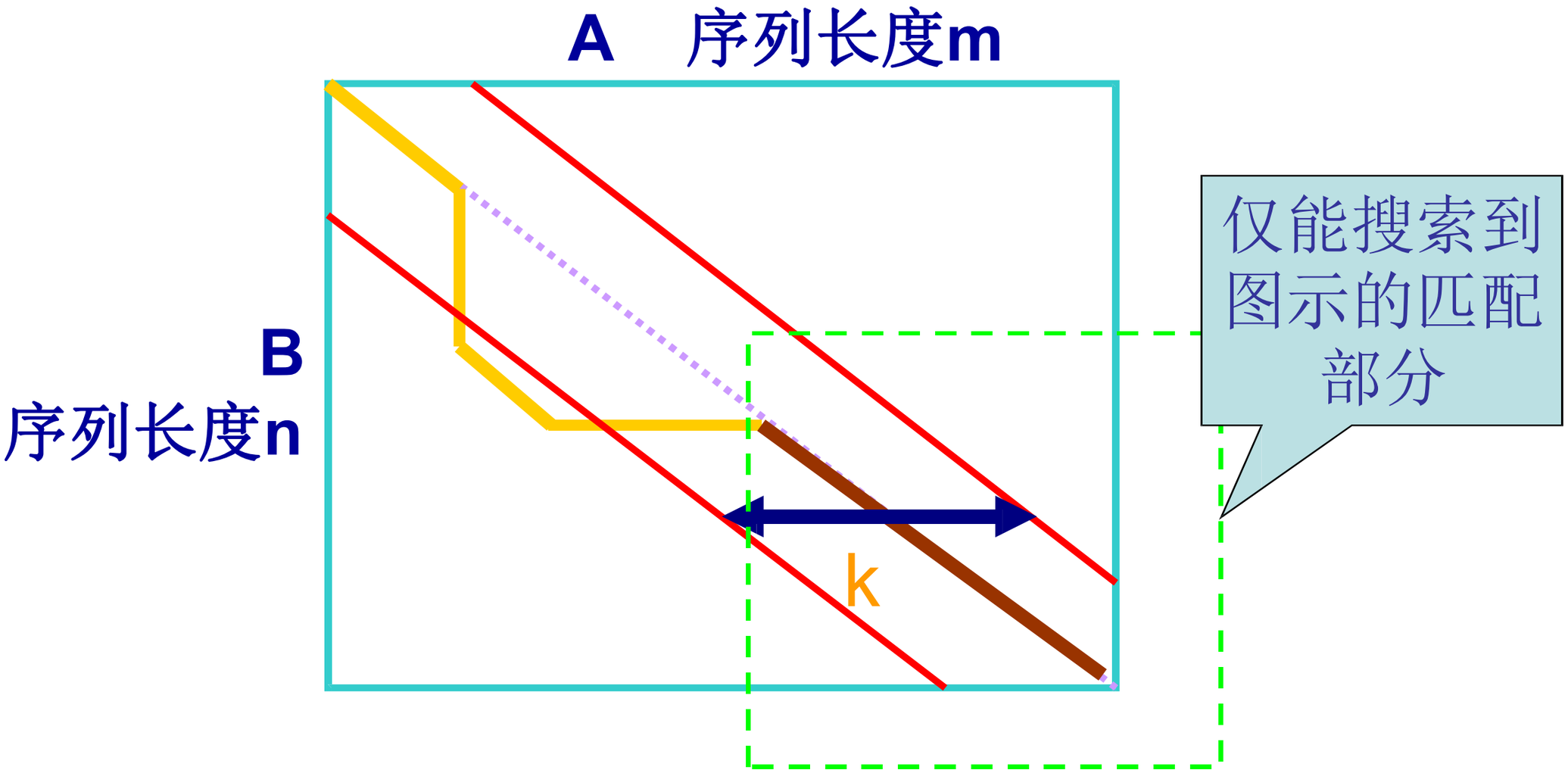
K-tup算法原理 (2)



令k为一常数，搜索限定区域内的最优比对

时间复杂度：O(kn)

K-tup算法原理：缺点



FASTA和BLAST

- 启发式算法, **heuristic algorithm**
- 不能保证搜索到最优的比对
- 具有很好的灵敏度，略为降低特异性
- 大大缩短序列比对的时间
- ***k-tup***算法：字符串匹配
- 应用：大的数据库搜索
- 时间复杂度：**< $O(n^2)$**

FASTA

- ***k-tup***: 蛋白质序列：**1~2 aa**；**DNA**序列：**4~6 nt**
- 以短序列构建索引，采用**hash**表存储方式
- 对于需要比较的两条序列，在**hash**表中查找所有完全匹配的片段；**FASTA**给每一个匹配给定一个**tup**值
- 产生**10**个最高分值片段，重新用**PAM250**打分；
- 将同一序列上的高分值区域连接在一起
- 采用**Needleman-Wunsch**或者**Smith-Waterman**算法对该高分值区域重新打分

FASTA 算法

FASTA: 索引表构建

- 以蛋白质序列为例
- ***k*=1**

氨基酸	*tup*分值
A	5
C	5
K	5
N	5
P	5
R	5
S	5
T	5

- 给定两条蛋白质序列：

Protein1: NCSPTA

Protein2: ACSPRK

氨基酸	位置1	位置2	*tup*分值
A	6	1	5
C	2	2	5
K	-	6	0
N	1	-	0
P	4	4	5
R	-	5	0
S	3	3	5
T	5	-	0

- 给定两条蛋白质序列：
  - Protein1: NCSPTA
  - Protein2: ACSPRK

氨基酸	位置1	位置2	*tup*分值
A	6	1	5
C	2	2	5
K	-	6	0
N	1	-	0
P	4	4	5
R	-	5	0
S	3	3	5
T	5	-	0

两条序列匹配结果

Protein1: NCSPTA

Protein2: ACSPRK

FASTA

1. FASTP
Lipman & Pearson, Science (1985) 227, 1435

2. FASTA
Pearson & Lipman, PNAS (1988) 85, 2444.

3. Lookup table
Dumas & Ninio, NAR (1982) 197.

FASTA

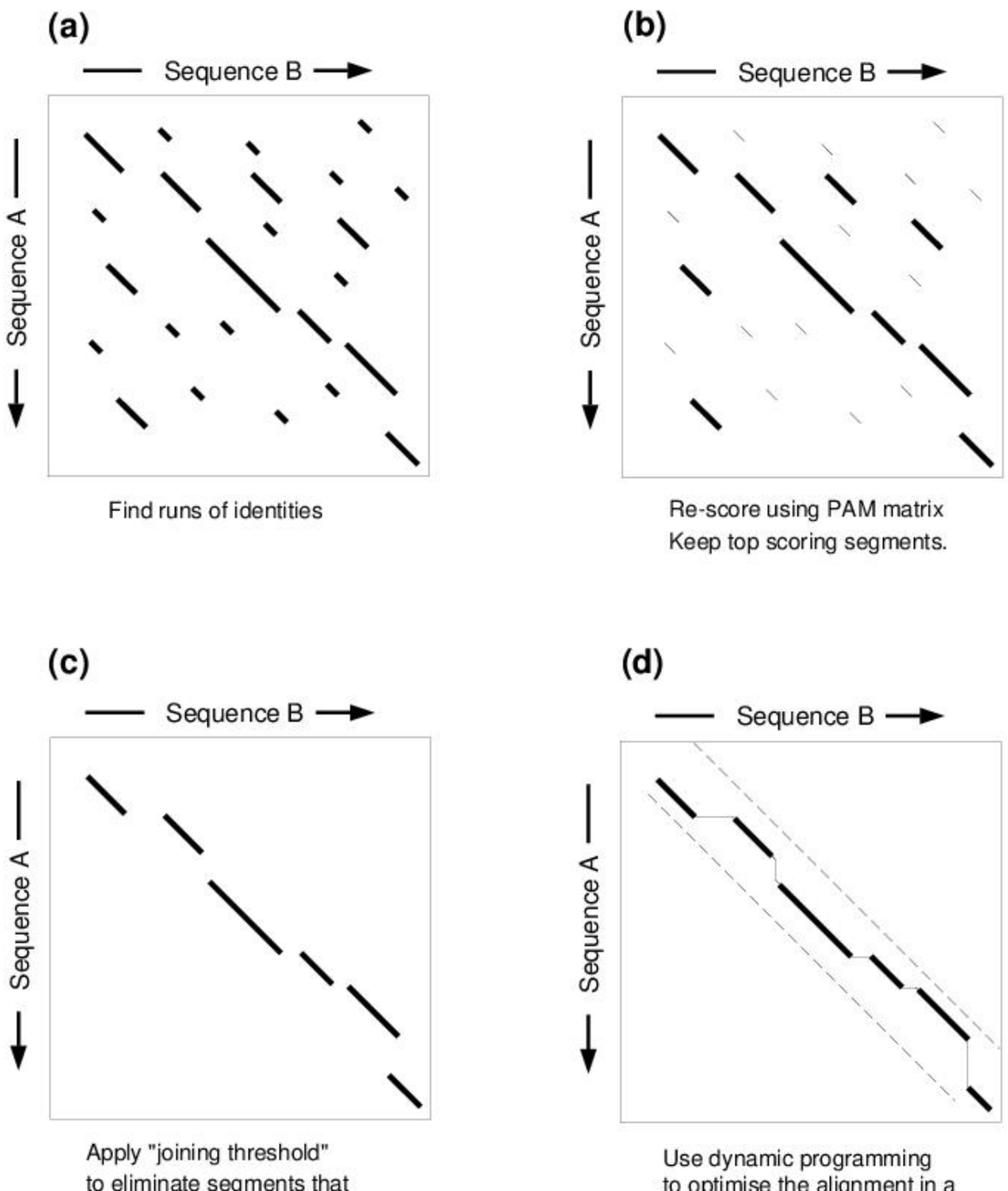
Four steps:

1. Identify common k-word (look-up table)

2. Score diagonals (PAM) to find 10 best diagonals

3. Join high scoring diagonals

4. Optimize alignment by DP

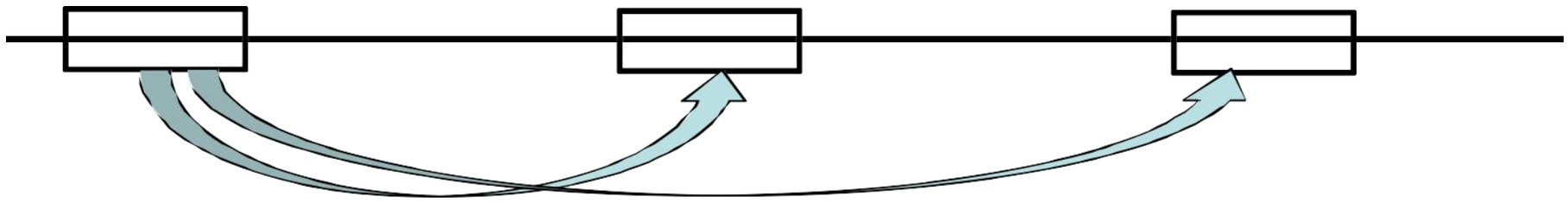


Dumas-Ninio look-up table

Original question:

For a sequences of length N, how to quickly find whether or not it contains repeated subsequences (length =k)?

Naïve methods: comparing every word with every other word of the sequence.



The time cost will increase with $O(N^2/2)$.
J P Dumas and J Ninio. Efficient algorithms for folding and comparing nucleic acid sequences. Nucleic Acids Res (1982) 10: 197-206.

Dumas-Ninio look-up table (example of k=2)

Question:

Given a sequence "TCGGATTCGTACGGTACGGATC", how to quickly find the locations of all the most frequently appeared words (length=2)?

1, Label the sequences by numbers

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
T	C	G	G	A	T	T	C	G	T	A	C	G	G	T	A	C	G	G	A	T	C
TC	CG	GG	GA	AT	TT	TC	CG	GT	TA	AC	CG	GG	GT	TA	AC	CG	GG	GA	AT	TC	

2, Map the word sequence to numerical sequence

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
GG	GC	AG	CG	GT	TT	AT	AA	GA	TA	TG	CA	CT	TC	AC	CC

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	pos.
T	C	G	G	A	T	T	C	G	T	A	C	G	G	T	A	C	G	G	A	T	C	
TC	CG	GG	GA	AT	TT	TC	CG	GT	TA	AC	CG	GG	GT	TA	AC	CG	GG	GA	AT	TC		
14	4	1	9	7	6	14	4	5	10	15	4	1	5	10	15	4	1	9	7	14		code

Dumas-Ninio look-up table (example of k=2)

3, Construct T/M-matrix to record the location of all words in one scan

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
code	T	C	G	G	A	T	T	C	G	T	A	C	G	G	T	A	C	G	G	A	T	C
	TC	CG	GG	GA	AT	TT	TC	CG	GT	TA	AC	CG	GG	GT	TA	AC	CG	GG	GA	AT	TC	
	14	4	1	9	7	6	14	4	5	10	15	4	1	5	10	15	4	1	9	7	14	

T-matrix

code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16					
pos.	3			2		6	5		4					1							
														7							

M-matrix

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
pos.							1														

Dumas-Ninio look-up table (example of k=2)

3, Construct T/M-matrix to record the location of all words in one scan

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
code	T	C	G	G	A	T	T	C	G	T	A	C	G	G	T	A	C	G	G	A	T	C
	TC	CG	GG	GA	AT	TT	TC	CG	GT	TA	AC	CG	GG	GT	TA	AC	CG	GG	GA	AT	TC	
	14	4	1	9	7	6	14	4	5	10	15	4	1	5	10	15	4	1	9	7	14	

T-matrix

code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
pos.	3			2		6	5		4					1		
				8										7		

M-matrix

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
pos.							1	2													

Dumas-Ninio look-up table (example of k=2)

3, Construct T/M-matrix to record the location of all words in one scan

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
code	T	C	G	G	A	T	T	C	G	T	A	C	G	G	T	A	C	G	G	A	T	C
	TC	CG	GG	GA	AT	TT	TC	CG	GT	TA	AC	CG	GG	GT	TA	AC	CG	GG	GA	AT	TC	
	14	4	1	9	7	6	14	4	5	10	15	4	1	5	10	15	4	1	9	7	14	

T-matrix

code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
pos.	3			2		6	5		4					1		
				8	9					10				7	11	
				12												

M-matrix

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
pos.							1	2				8									

Dumas-Ninio look-up table (example of k=2)

3, Construct T/M-matrix to record the location of all words in one scan

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
code	T	C	G	G	A	T	T	C	G	T	A	C	G	G	T	A	C	G	G	A	T	C
	TC	CG	GG	GA	AT	TT	TC	CG	GT	TA	AC	CG	GG	GT	TA	AC	CG	GG	GA	AT	TC	
	14	4	1	9	7	6	14	4	5	10	15	4	1	5	10	15	4	1	9	7	14	

T-matrix

code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
pos.	3			2		6	5		4					1		
				8	9					10				7	11	
	13			12												

M-matrix

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
pos.							1	2				8	3								

Dumas-Ninio look-up table (example of k=2)

3, Construct T/M-matrix to record the location of all words in one scan

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
code	T	C	G	G	A	T	T	C	G	T	A	C	G	G	T	A	C	G	G	A	T	C
	TC	CG	GG	GA	AT	TT	TC	CG	GT	TA	AC	CG	GG	GT	TA	AC	CG	GG	GA	AT	TC	
	14	4	1	9	7	6	14	4	5	10	15	4	1	5	10	15	4	1	9	7	14	

M(8)=2, indicates tat a dimer of rank 8 in the sequence occurred previously at position 2.

T-matrix

code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
pos.	3			2		6	5		4					1		
				8	9					10				7	11	
	13			12	14					15					16	
	18			17			20		19					21		

M-matrix

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
pos.							1	2				8	3	9	10	11	12	13	4	5	7

Dumas-Ninio look-up table (example of k=2)

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	T	C	G	G	A	T	T	C	G	T	A	C	G	G	T	A	C	G	G	A	T	C
	TC	CG	GG	GA	AT	TT	TC	CG	GT	TA	AC	CG	GG	GT	TA	AC	CG	GG	GA	AT	TC	
code	14	4	1	9	7	6	14	4	5	10	15	4	1	5	10	15	4	1	9	7	14	

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
GG	GC	AG	CG	GT	TT	AT	AA	GA	TA	TG	CA	CT	TC	AC	CC

T-matrix

code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
pos.	3			2		6	5		4					1		
				8	9					10				7	11	
	13			12	14					15					16	
	18			17			20		19					21		
pos.	18	0	0	17	14	6	20	0	19	15	0	0	0	21	16	0

M-matrix

pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
pos.							1	2				8	3	9	10	11	12	13	4	5	7

Using lookup table, we can quickly trace back identity and location of words.

'GG' appears at positions: 18, 13,3

'TC' appears at positions: 21, 7,1 etc

Dumas-Ninio look-up table

When we make an alignment, we only need to trace a limited number paths to find the matched words, i.e. from T to M

Advantage: fast $O(2N)$ vs. $O(N^2)$

Defect: only identical residue pairs can be aligned

BLAST

BLAST

Altschul et al, Basic Local Alignment Search Tool. J Mol Biol (1990) 215, 403.

Two steps:

Step 1. Search for exact matches of a small fixed length W between the query and sequences in the database.

For example, given the sequences AGTTATT and GCTTAAG and a word length **W** = 3, BLAST would identify the matching substring TTA that is common to both sequences. By default, **W** = 11 for nucleic acids.

```
···AGTTATT···
      |||
···GCTTAAG···
```

BLAST

Step 2. Try to extend the match in both directions, starting at the seed.

The ungapped alignment process extends the initial seed match of length **W** in each direction in an attempt to boost the alignment score. Insertions and deletions are not considered during this stage.

```
···AGTTATT···
   :|||:
···GCTTAAG···
```

The highest-scoring alignment will be returned.

BLAST

- **Word size: DNA, 11nt;** 蛋白质，**3aa**
- 蛋白质序列数据库，构建由**3aa**组成的分值表,采用**BLOSUM62**矩阵打分
- 待查询序列，打断成**3aa**的片段，在上述数据库中的分值表中进行查询
- 保留高于域值的结果，并进行两端的延伸，**HSP: high-scoring segment pair**
- **Nothing can be worse:** 牺牲灵敏度，提高计算速度

BLAST:索引表构建

- formatdb命令，将fasta格式的序列文件转换成blast能够识别的文件格式
- 构建索引表：

PQG

PQG	7+5+6=18
PEG	7+2+6=15
PWG	7-2+6=11
SQG	-1+5+6=10

BLAST: 序列匹配

- 两条蛋白质序列
- **Protein1：IVPQGRL**
- **Protein2：VAPEGKL**

- **Protein1：I V P Q G R L**
- **Protein2：V A P E G K L**

<Word>

7 2 6

3 0 2 4

两边延伸

HSP分值：3+0+15+2+4=24

BLAST

Different type of BLAST programs:

- 早期的**BLAST**版本：无空位罚分
- 新版本：**Gap Penalties: Existence: 11，Extension: 1**
  - Nucleotide-nucleotide BLAST (blastn)
  - Protein-protein BLAST (blastp)
  - Position-Specific Iterative BLAST (PSI-BLAST)
  - Pattern Hit Initiated BLAST (Phi-BLAST)

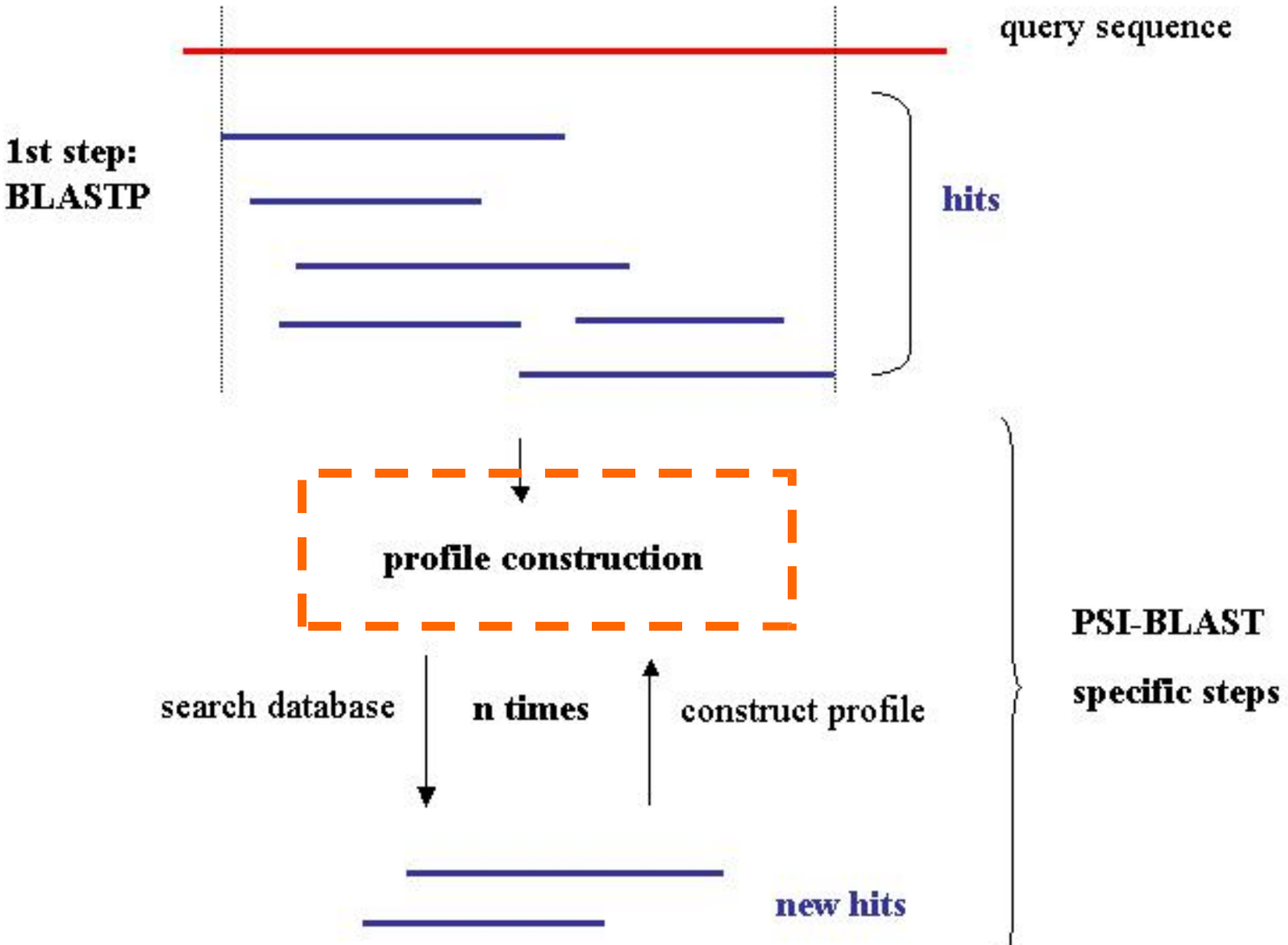
…

PSI-BLAST

Psi-BLAST：迭代搜索

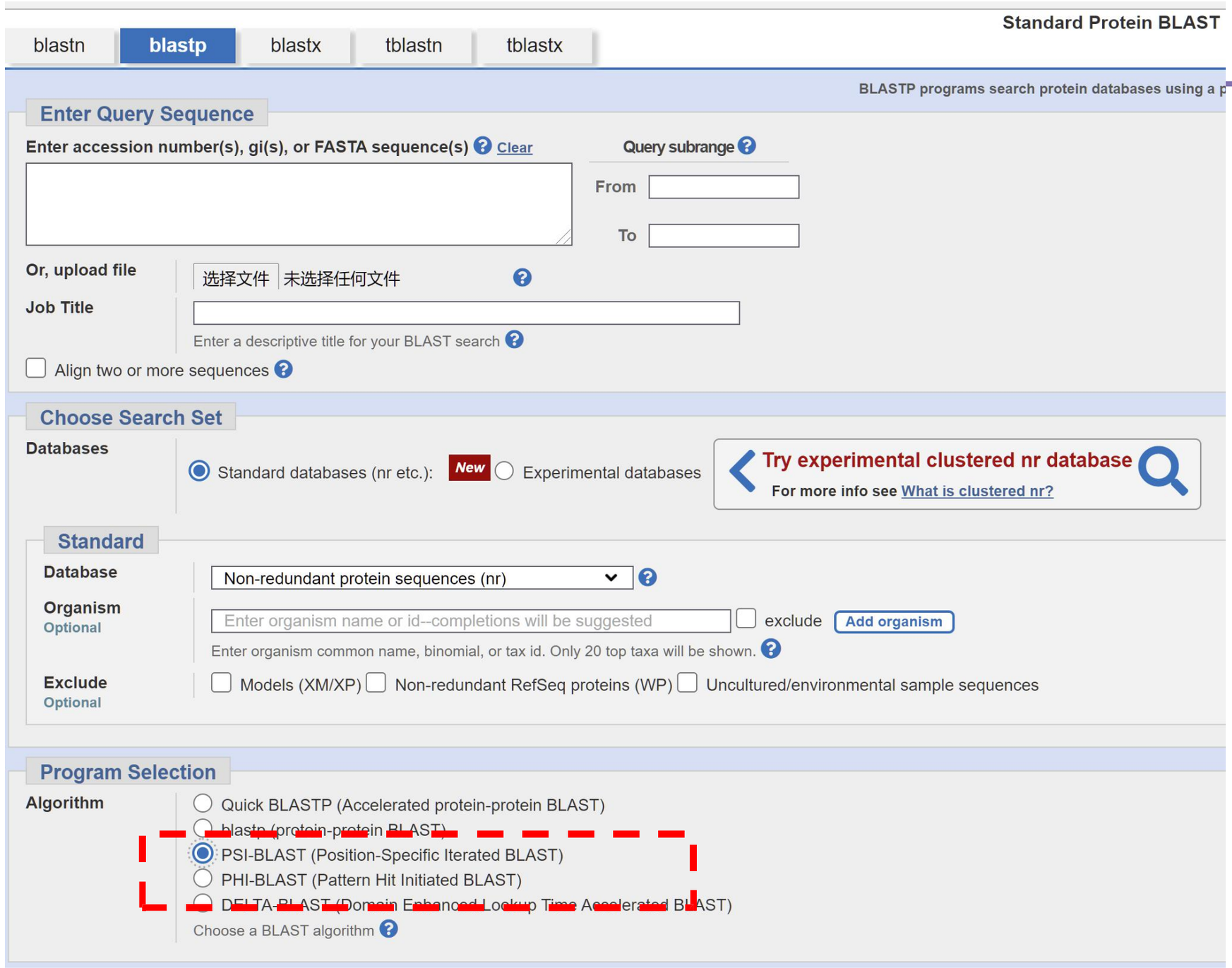
- 第一步，使用普通的blast算法进行搜索
- 第二步，将搜索得到的序列，包括输入的序列放在 一起，构建位点特异性的矩阵(Position Specific Matrix)
- 第三步，利用上面得到的矩阵谱(profile)，再次在数 据库中进行搜索
- 重复2，3步，直到不再有新的序列出现
- 优点：能够发现序列相似性非常低的同源序列
- 缺点：常常得到假阳性的结果

Psi-BLAST：迭代搜索 (2)



Psi-BLAST：例

- >NP_954673 ubiquitin-conjugating enzyme E2 Kua-UEV isoform 1 [Homo sapiens]
  MAGAEDWPGQQLELDEDEASCCRWGAQHAGAREL
  AALYSPGKRLQEWCSVILCFSLIAHNLVHLLLLARWE
  DTPLVILGVVAGALIADFLSGLVHWGADTWGSVELPI
  VGKAFIRPFREHHIDPTAITRHDFIETNGDNCLVTLLPL
  LNMAYKFRTHSPEALEQLYPWECFVFCLIIFGTFTNQI
  HKWSHTYFGLPRWVTLLQDWHVILPRKHHRIHHVSP
  HETYFCITTGVKVPRNFRLLEELEEGQKGVGDGTVS
  WGLEDDEDMTLTRWTGMIIGPPRTIYENRIYSLKIECG
  PKYPEAPPFVRFVTKINMNGVNSSNGVVDPRAISVLA
  KWQNSYSIKVVLQELRRLMMSKENMKLPQPPEGQC
  YSN



https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Run PSI-Blast iteration 2

Hit list size 500

Distance tree of results NEW

Sequences with E-value BETTER than threshold

Related Structures

Sequences producing significant alignments:			Score (Bits)	E Value	
NEW	☑	pdb|2C2V|C	Chain C, Crystal Structure Of The Chip-Ubc13-Uev1a...	305	2e-83	S
NEW	☑	pdb|2A4D|A	Chain A, Structure Of The Human Ubiquitin-Conjugat...	304	2e-83	S
NEW	☑	pdb|2HLW|A	Chain A, Solution Structure Of The Human Ubiquitin...	303	7e-83	S
NEW	☑	pdb|1J74|A	Chain A, Crystal Structure Of Mms2 >pdb|1J7D|A Cha...	282	1e-76	S
NEW	☑	pdb|1ZGU|A	Chain A, Solution Structure Of The Human Mms2-Ubiquit	276	8e-75	S
NEW	☑	pdb|2GMI|B	Chain B, Mms2UBC13~UBIQUITIN	150	4e-37	S
NEW	☑	pdb|1JAT|B	Chain B, Mms2UBC13 UBIQUITIN CONJUGATING ENZYME COMPL	150	5e-37	S
NEW	☑	pdb|2QOV|A	Chain A, Crystal Structure Of Ubiquitin Conjugatin...	142	1e-34	S
NEW	☑	pdb|2AWF|A	Chain A, Structure Of Human Ubiquitin-Conjugating Enz	51.6	3e-07	S
NEW	☑	pdb|2PWQ|A	Chain A, Crystal Structure Of A Putative Ubiquitin...	50.4	7e-07	S
NEW	☑	pdb|2E2C|A	Chain A, E2-C, An Ubiquitin Conjugating Enzyme Req...	49.3	2e-06	
NEW	☑	pdb|1Q34|A	Chain A, Crystal Structure Of Two [illegible] Enzyme...	48.9	2e-06	S
NEW	☑	pdb|2F47|A	Chain A, Toxoplasma Gondii Ubiquitin Conjugating E...	48.5	3e-06	S

显著性计算

Significance of alignment in BLAST: E-value

For any alignment, we can have an alignment score (S). The score itself does not tell how significant it is.

Definition: The ***E*-value** of an alignment with score S is the ***expected number*** of alignments to be found with score ≥S in two random sequences (of same lengths and letter compositions).

$$\textbf{E-value} = Kmne^{-\lambda S}$$

This eq is an approx. Exact solution is an open quiz

K and *λ* represent natural scales for the search space and the scoring system respectively. *m* and *n* are the sizes of the query and template sequences.

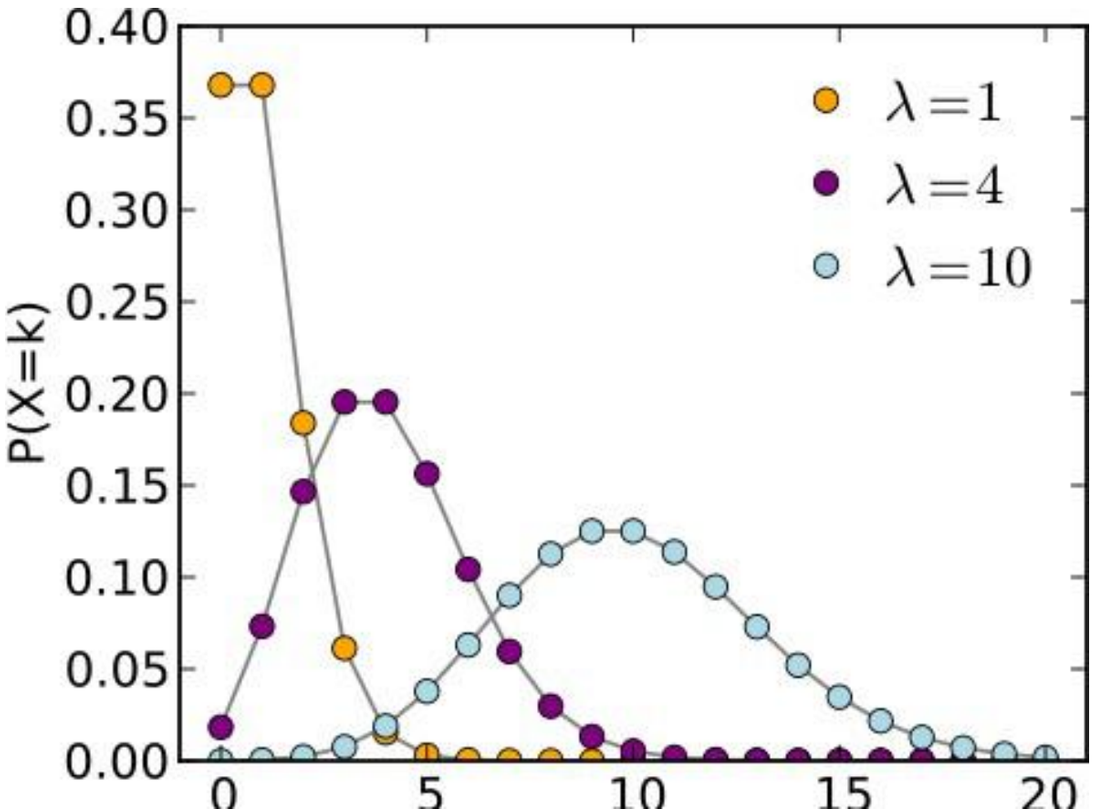
In general, the typical threshold for a good E-value from a BLAST search is 0.001 or lower. An alignment of low E-value means that that alignment is highly unique, and not due to error.

For a proof of the equation, see Karlin & Altschul, PNAS (1990), 87, 2264; PNAS (1993), 90, 5873.

Poisson distribution

If the ***expected number*** of an event to occur is λ, the probability that there are exactly *k* occurrences (*k* = 0, 1, 2, ...) is equal to

$$p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$



Significance of alignment in BLAST: P-value

Definition: The P-value of an alignment with score S is the *likelihood* that two ra
sequences will have (at least one) alignments with score ≥ S.

Relation between P-value and E-value

If E(S) is the expected number of alignment with score ≥S, the lik
of getting exactly *k* such (independent) alignments is

$$\textbf{P-value} = \frac{E(S)^k e^{-E(S)}}{k!} \leftarrow p(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Likelihood of getting 0 such alignment: $e^{-E(S)}$

- Likelihood of getting at least one such alignment: $1 - e^{-E(S)}$

显著性计算

以蛋白质CDC28_YEAST 为例在酵母中的序列比对结果，如何计算序列比对的显著性？

List of potentially matching sequences

Send selected sequences to Clustal W (multiple alignment) 提交查询内容 Select up to...

☐ Include query sequence

	Db	AC	Description	Score	E-value
☐	sp	P00546	CDC28_YEAST Cell division control protein 28 (EC 2.7.1...	607	e-175
☐	sp	P17157	PHO85_YEAST Cyclin-dependent protein kinase PHO85 (EC ...	309	3e-85
☐	sp	Q03957	CTK1_YEAST CTD kinase subunit alpha (EC 2.7.11.23) (CT...	212	5e-56
☐	sp	P23293	BUR1_YEAST Serine/threonine-protein kinase BUR1 (EC 2....	211	9e-56
☐	sp	P06242	KIN28_YEAST Serine/threonine-protein kinase KIN28 (EC ...	196	5e-51
☐	sp	P16892	FUS3_YEAST Mitogen-activated protein kinase FUS3 (EC 2...	182	8e-47
☐	sp	P32485	HOG1_YEAST Mitogen-activated protein kinase HOG1 (EC 2...	180	3e-46
☐	sp	P39073	SSN3_YEAST Meiotic mRNA stability protein kinase SSN3 ...	179	7e-46
☐	sp	Q00772	SLT2_YEAST Mitogen-activated protein kinase SLT2/MPK1 ...	163	3e-41
☐	sp	P14681	KSS1_YEAST Mitogen-activated protein kinase KSS1 (EC 2...	163	3e-41
☐	sp	P41808	SMK1_YEAST Sporulation-specific mitogen-activated prot...	155	6e-39
☐	sp	P38615	MDS1_YEAST Serine/threonine-protein kinase MDS1/RIM11 ...	146	5e-36
☐	sp	P36005	KKQ1_YEAST Probable serine/threonine-protein kinase YK...	145	8e-36
☐	sp	P50873	MRK1_YEAST Serine/threonine-protein kinase MRK1 (EC 2....	132	7e-32
☐	sp	P21965	MCK1_YEAST Protein kinase MCK1 (EC 2.7.11.1) (Meiosis ...	127	2e-30
☐	sp	P19454	CSK22_YEAST Casein kinase II subunit alpha' (EC 2.7.11...	124	2e-29
☐	sp	P15790	CSK21_YEAST Casein kinase II subunit alpha (EC 2.7.11....	121	1e-28
☐	sp	P32581	IME2_YEAST Meiosis induction protein kinase IME2/SME1 ...	117	2e-27
☐	sp	P06782	SNF1_YEAST Carbon catabolite-derepressing protein kina...	111	2e-25
☐	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

其中

- SNF1_YEAST的结果：
- Score: 111
- E-value: 2e-25

- 问题：
  - 如何计算Score？
  - 如何计算E-value？该值是何意义？

求近似值

- **S: bit**分值，有公式：

$$S = \frac{\lambda R - \ln(K)}{\ln(2)}$$

其中**R**，是**raw**分值，根据打分矩阵直接得到的分数

- $\boldsymbol{E(S) \approx Kmne^{-\lambda S} = mn2^{-S}}$;

因此，上例

```
sp  P06782      Carbon catabolite-derepressing protein kinase (EC 2.7.11.1) [SNF1]  633 AA
    SNF1_YEAST  [Saccharomyces cerevisiae (Baker's yeast)]                           align

 Score =  111 bits (277),  Expect = 2e-25
 Identities = 87/296 (29%), Positives = 144/296 (48%), Gaps = 44/296 (14%)
```

- R=277

上例 (2)

ungapped

```
Lambda     K      H
   0.323    0.142    0.434
```

```
Gapped
Lambda     K      H
   0.267   0.0410   0.140
```

- R=277
- λ= 0.267
- K=0.0410
- m=208
- n=2,657,097

$$S = \frac{\lambda R - \ln(K)}{\ln(2)} = \frac{0.267 * 277 - \ln(0.041)}{\ln(2)} = 111$$

```
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 2,033,710
Number of Sequences: 2415840
Number of extensions: 91335
Number of successful extensions: 543
Number of sequences better than 10.0: 100
Number of HSP's better than 10.0 without gapping: 118
Number of HSP's successfully gapped in prelim test: 7
Number of HSP's that attempted gapping in prelim test: 217
Number of HSP's gapped (non-prelim): 160
length of query: 298
length of database: 3,316,707
effective HSP length: 90
effective length of query: 208
effective length of database: 2,657,097
effective search space: 552676176
effective search space used: 552676176
```

$$E(S) = mn2^{-S} = 208 * 2657097 * 2^{-}$$
$$= 2e - 25$$

软件操作

Nucleotide BLAST三个program

https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/BLAST/nucleotide_blast.html

Nucleotide-Nucleotide BLAST (blastn)

Now that we have explored the program and database options, let's do a basic **blastn** search with the Jurassic Park sequence that you have copied/pasted into memory. If you haven't already copied the query sequence into memory, please do it now.

One more note before we do the search...

The nucleotide BLAST page provides a selection of three programs that vary in their sensitivity and speed: **megablast** (default), **discontiguous megablast**, and **blastn**.

For our sample search, use the traditional **blastn** program.

Footnote, for your **future reference**. Some of the differences between the algorithms are highlighted below.

Megablast	Retrieves **highly similar sequences** and is very **fast**. It efficiently find long alignments between very similar sequences -- it is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more. (word size* is 28 base pairs). learn more...
Discontiguous megablast	Retrieves **more dissimilar sequences** than megablast, but is more sensitive than blastn. It uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons -- the third base wobbling is taken into consideration by focusing on finding matches at the first and second codon positions while ignoring the mismatches in the third position. (word size* can be set only at 11 or 12 base pairs.) learn more...
Blastn	Retrieves **somewhat similar sequences**, so **can find more distantly related sequences**, but is **slower** than megablast and discontiguous megablast. (default word size* can range from 7 base pairs to 11 (default) base pairs) learn more...

* **Word Size** is discussed later in the module in the slide on how did BLAST work. It is mentioned here only so this slide can serve as a

同源序列搜索

- 同源序列通常具有相似的生物学功能
- 同源关系的分析：直系同源 or 旁系同源？
  - **直系同源（Orthologs）**是指来自于不同物种的由垂直家系，也就是物种形成，进化而来的基因，并且典型的保留与原始基因相同的功能。也就是说，随着进化分支，一个基因进入了不同的物种，并保留了原有功能。这时，不同物种中的这个基因就属于直系同源。
  - **旁系同源（Paralogs）**是指在同一物种中的来源于基因复制的基因，可能会进化出新的但与原功能相关的功能来。
- 直系同源序列的确定：相互最佳匹配
- 旁系同源序列的确定：BLAST，序列比对及数 据库搜索，至少存在一个共有的功能结构域
- 整体分析/蛋白质家族分析：系统发育树的构建

例：Bub1

- 芽殖酵母的**Bub1**:定位于动点，纺锤体检验点

UniProt

UniProtKB | Advanced | Search

BLAST Align Retrieve/ID mapping Peptide search Help Contact

From June 20, 2018 all traffic will be automatically redirected to HTTPS. More information or view this page using https

UniProtKB - P41695 (BUB1_YEAST)

Basket

Display

- Entry
- Publications
- Feature viewer
- Feature table

All None

- Function
- Names & Taxonomy
- Subcellular location
- Pathology & Biotech
- PTM / Processing
- Expression
- Interaction
- Structure

BLAST | Align | Format | Add to basket | History

Feedback | Help video | Other tutorials and videos

Protein: **Checkpoint serine/threonine-protein kinase BUB1**

Gene: **BUB1**

Organism: *Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)*

Status: Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level

Function

Involved in cell cycle checkpoint enforcement. The formation of a MAD1-BUB1-BUB3 complex seems to be required for the spindle checkpoint mechanism. Catalyzes the phosphorylation of BUB3 and its autophosphorylation. Associates with centromere (CEN) DNA via interaction with SKP1. The association with SKP1 is required for the mitotic delay induced by kinetochore tension defects, but not for the arrest induced by spindle depolymerization or kinetochore assembly defects.

1 Publication

Miscellaneous

Present with 414 molecules/cell in log phase SD medium. 1 Publication

Catalytic activity

ATP + a protein = ADP + a phosphoprotein.

Sites

获得FASTA序列

Display

- Entry
- Publications
- Feature viewer
- Feature table

All None

- [x] Function
- [x] Names & Taxonomy
- [x] Subcellular location
- [x] Pathology & Biotech
- [x] PTM / Processing
- [ ] Expression
- [x] Interaction
- [x] Structure
- [x] Family & Domains
- [x] Sequence
- [x] Similar proteins
- [ ] Cross-references
- [x] Entry information
- [x] Miscellaneous

▲Top

Sequence

fasta 第1条，共1条

Sequence status: Complete.

P41695-1 [UniParc] FASTA Add to basket

« Hide

Length: 1,021
Mass (Da): 117,868
Last modified: October 1, 1996 - v2
Checksum: 6D76FC980775D3F9

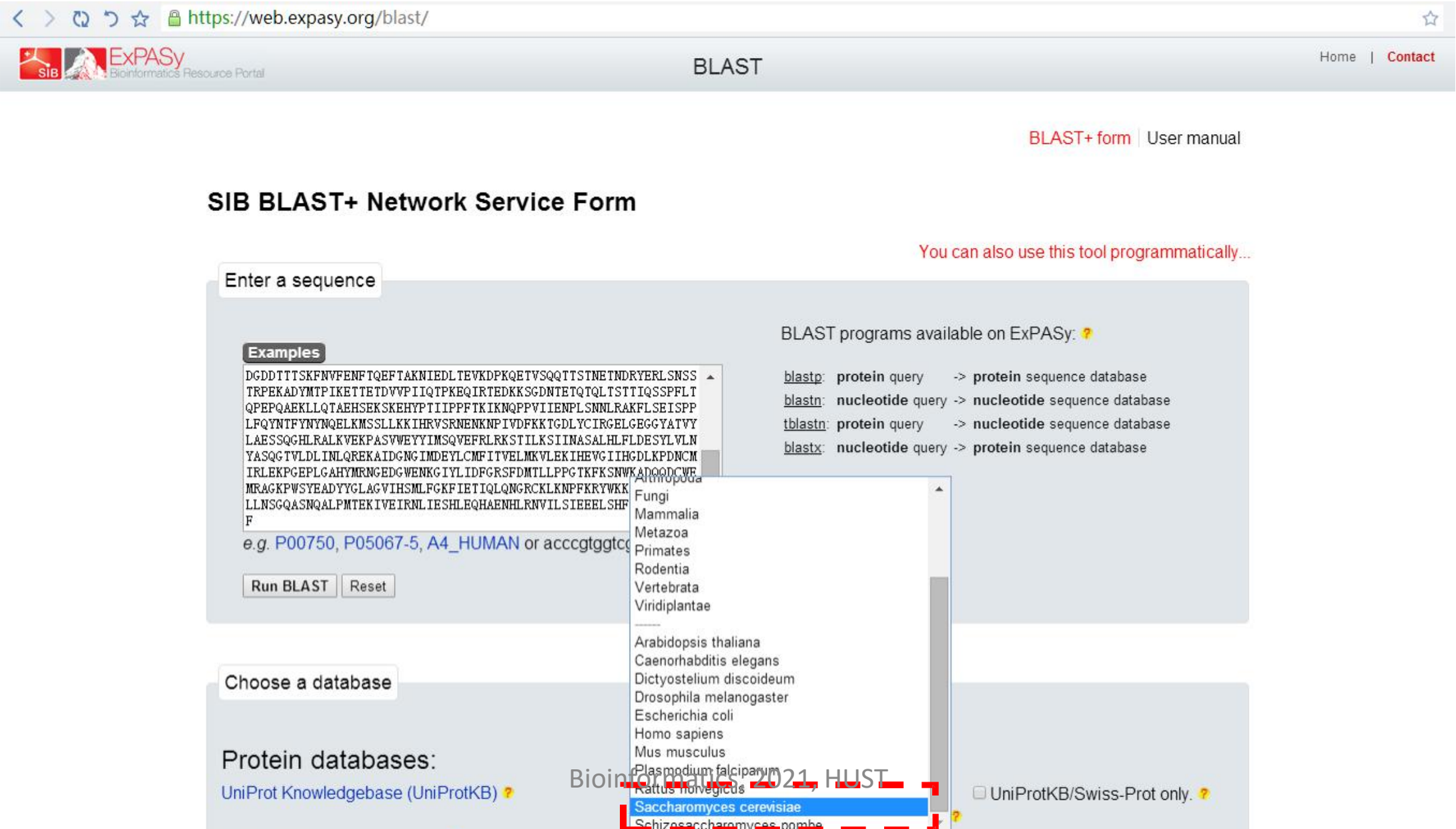
BLAST GO

```
        10         20         30         40         50
MNLDLGSTVR GYESDKDTFP QSKGVSSSQK EQHSQLNQTK IAYEQRLLND
        60         70         80         90        100
LEDMDDPLDL FLDYMIWIST SYIEVDSESG QEVLRSTMER CLIYIQDMET
       110        120        130        140        150
YRNDPRFLKI WIWYINLFLS NNFHESENTF KYMFNKGIGT KLSLFYEEFS
       160        170        180        190        200
KLLENAQFFL EAKVLLELGA ENNCRPYNRL LRSLSNYEDR LREMNIVENQ
       210        220        230        240        250
NSVPDSRERL KGRLIYRTAP FFIRKFLTSS LMTDDKENRA NLNSNVGVGK
       260        270        280        290        300
SAPNVYQDSI VVADFKSETE RLNLNSSKQP SNQRLKNGNK KTSIYADQKQ
       310        320        330        340        350
SNNPVYKLIN TPGRKPERIV FNFNLIYPEN DEEFNTEEIL AMIKGLYKVQ
       360        370        380        390        400
RRGKKHTEDY TSDKNRKKRK LDVLVERRQD LPSSQPPVVP KSTRIEVFKD
       410        420        430        440        450
DDNPSQSTHH KNTQVQVQTT TSILPLKPVV DGNLAHETPV KPSLTSNASR
       460        470        480        490        500
SPTVTAFSKD AINEVFSMFN QHYSTPGALL DGDDTTTSKF NVFENFTQEF
       510        520        530        540        550
TAKNIEDLTE VKDPKQETVS QQTTSTNETN DRYERLSNSS TRPEKADYMT
       560        570        580        590        600
PIKETTETDV VPIIQTPKEQ IRTEDKKSGD NTETQTQLTS TTIQSSPFLT
```

www.uniprot.org/uniprot/P41695.fasta

```
>sp|P41695|BUB1_YEAST Checkpoint serine/threonine-protein kinase BUB1
MNLDLGSTVRGYESDKDTFPQSKGVSSSQKEQHSQLNQTKIAYEQRLLNDLEDMDDPLDL
FLDYMIWISTSYIEVDSESGQEVLRSTMERCLIYIQDMETYRNDPRFLKIWIWYINLFLS
NNFHESENTFKYMFNKGIGTKLSLFYEEFSKLLENAQFFLEAKVLLELGAENNCRPYNRL
LRSLSNYEDRLREMNIVENQNSVPDSRERLKGRLIYRTAPFFIRKFLTSSLMTDDKENRA
NLNSNVGVGKSAPNVYQDSIVVADFKSETERLNLNSSKQPSNQRLKNGNKKTSIYADQKQ
SNNPVYKLINTPGRKPERIVFNFNLIYPENDEEFNTEEILAMIKGLYKVQRRGKKHTEDY
TSDKNRKKRKLDVLVERRQDLPSSQPPVVPKSTRIEVFKDDDNPSQSTHHKNTQVQVQTT
TSILPLKPVVDGNLAHETPVKPSLTSNASRSPTVTAFSKDAINEVFSMFNQHYSTPGALL
DGDDTTTSKFNVFENFTQEFTAKNIEDLTEVKDPKQETVSQQTTSTNETNDRYERLSNSS
TRPEKADYMTPIKETTETDVVPIIQTPKEQIRTEDKKSGDNTETQTQLTSTTIQSSPFLT
QPEPQAEKLLQTAEHSEKSKEHYPTIIPPFTKIKNQPPVIIENPLSNNLRAKFLSEISPP
LFQYNTFYNYNQELKMSSLLKKIHRVSRNENKNPIVDFKKTGDLYCIRGELGEGGYATVY
LAESSQGHLRALKVEKPASVWEYYIMSQVEFRLRKSTILKSIINASALHLFLDESYLVLN
YASQGTVLDLINLQREKAIDGNGIMDEYLCMFITVELMKVLEKIHEVGIIHGDLKPDNCM
IRLEKPGEPLGAHYMRNGEDGWENKGIYLIDFGRSFDMTLLPPGTKFKSNWKADQQDCWE
MRAGKPWSYEADYYGLAGVIHSMLFGKFIETIQLQNGRCKLKNPEKRYWKKEIWGVIEDL
```

酵母的同源序列：旁系同源序列



Mad3: 旁系同源序列

List of the matches

Clustal W (multiple alignment) 提交

Select up to...

Include query sequence

		Accession	Db	Description	Score	E-value
	1	P41695 (BUB1_YEAST)		Checkpoint serine/threonine-protein kinase ...	2086	0.0
	2	P47074 (MAD3_YEAST)		Spindle assembly checkpoint component MAD3 ...	184	1e-49
	3	Q03306 (PKH3_YEAST)		Serine/threonine-protein kinase PKH3 OS=Sac...	63.2	2e-10
	4	Q01389 (BCK1_YEAST)		Serine/threonine-protein kinase BCK1/SLK1/S...	62.0	4e-10
	5	Q12236 (PKH2_YEAST)		Serine/threonine-protein kinase PKH2 OS=Sac...	55.5	4e-08
	6	P53599 (SSK2_YEAST)		MAP kinase kinase kinase SSK2 OS=Saccharomy...	50.8	1e-06
	7	Q03407 (PKH1_YEAST)		Serine/threonine-protein kinase PKH1 OS=Sac...	50.1	1e-06

2. P47074 (MAD3_YEAST)
Spindle assembly checkpoint component MAD3 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) OX=559292 GN=MAD3 PE=1 SV=1 CRC64=36550249D81BB6D1
Length=515

```
 Score = 184 bits (466),  Expect = 1e-49, Method: Compositional matrix adjust.
 Identities = 126/353 (36%), Positives = 198/353 (56%), Gaps = 37/353 (10%)

Query  35   QLNQTKIAYEQRLLNDLEDMDDPLDLFLDYMIWISTSYIEVDSESGQEVLRSTMERCLIY  94
            ++NQ K ++EQRL+++L  + DP+ L+L+Y+ W++ +Y +    S Q  + + +ERCL +
Sbjct  55   EINQVKSSFEQRLIDELPALSDPITLYLEYIKWLNNAYPQ-GGNSKQSGMLTLLERCLSH  113

Query  95   IQDMETYRNDPRFLKIWIWYINLFLSNNFHESENTFKYMFNKGIGTKLSLFYEEFSKLLE  154
            ++D+E YRND RFLKIW WYI LF  N+F ES + F YM   GIG++L+ FYEEF+ LL
Sbjct  114  LKDLERYRNDVRFLKIWFWYIELFTRNSFMESRDIFMYMLRNGIGSELASFYEEFTNLLI  173
```

人类同源序列：直系同源序列

https://web.expasy.org/blast/

ExPASy Bioinformatics Resource Portal

BLAST

Home | Contact

BLAST+ form | User manual

SIB BLAST+ Network Service Form

You can also use this tool programmatically...

Enter a sequence

Examples

```
>sp|P41695|BUB1_YEAST Checkpoint serine/threonine-protein
kinase BUB1 OS=Saccharomyces cerevisiae (strain ATCC 204508
/ S288c) OX=559292 GN=BUB1 PE=1 SV=2
MNLDLGSTVRGYESDKDTFPQSKGVSSSQKEQHSQLNQTKIAYEQRLLNDLEDMDDPLDL
FLDYMIWISTSYIEVDSESGQEVLRSTMERCLIYIQDMETYRNDPRFLKIWIWYINLFLS
NNFHESENTFKYMFNKGIGTKLSLFYEEFSKLLENAQFFLEAKVLLELGAENNCRPYNRL
LRSLSNYEDRLREMNIVENQNSVPDSRERLKGRLIYRTAPFFIRKFLTSSLMTDDKENRA
NLNSNVGVGKSAPNVYQDSIVVADFKSETERLNLNSSKQPSNQRLKNGNKK
SNNPVYKLINTPGRKPERIVFNFNLIYPENDEEFNTEEILAMIKGLYKVQR
TSDKNRKKRKLDVLVERRQDLPSSQPPVVPKSTRIEVFKDDDNPSQSTHHK
```

e.g. P00750, P05067-5, A4_HUMAN or acccgtggtcg

Run BLAST | Reset

BLAST programs available on ExPASy:

blastp:	**protein** query	->	**protein** sequence database
blastn:	**nucleotide** query	->	**nucleotide** sequence database
tblastn:	**protein** query	->	**nucleotide** sequence database
blastx:	**nucleotide** query	->	**protein** sequence database

Arthropoda
Fungi
Mammalia
Metazoa
Primates
Rodentia
Vertebrata
Viridiplantae

Arabidopsis thaliana
Caenorhabditis elegans
Dictyostelium discoideum
Drosophila melanogaster
Escherichia coli
Homo sapiens
Mus musculus
Plasmodium falciparum
Rattus norvegicus
Saccharomyces cerevisiae

Choose a database

Protein databases:

UniProt Knowledgebase (UniProtKB)

人类Bub1？

List of the matches

Clustal W (multiple alignment) 提交

Select up to...

Include query sequence

		Accession	Db	Description	Score	E-value
	1	O43683-3 (BUB1_HUMAN)		Isoform 3 of Mitotic checkpoint serine/th...	174	7e-44
	2	O43683 (BUB1_HUMAN)		Mitotic checkpoint serine/threonine-protein...	174	7e-44
	3	O43683-2 (BUB1_HUMAN)		Isoform 2 of Mitotic checkpoint serine/th...	88.6	2e-17
	4	O60566 (BUB1B_HUMAN)		Mitotic checkpoint serine/threonine-protei...	68.2	4e-11
	5	O60566-3 (BUB1B_HUMAN)		Isoform 3 of Mitotic checkpoint serine/t...	59.3	2e-08
	6	O60566-2 (BUB1B_HUMAN)		Isoform 2 of Mitotic checkpoint serine/t...	57.8	6e-08
	7	O95835 (LATS1_HUMAN)		Serine/threonine-protein kinase LATS1 OS=H...	53.9	1e-06

在酵母中做比对

https://web.expasy.org/blast/

ExPASy Bioinformatics Resource Portal

BLAST

Home | Contact

BLAST+ form | User manual

SIB BLAST+ Network Service Form

You can also use this tool programmatically...

Enter a sequence

Examples

```
>sp|O43683|BUB1_HUMAN Mitotic checkpoint serine/threonine-
protein kinase BUB1 OS=Homo sapiens OX=9606 GN=BUB1 PE=1
SV=1
MDTPENVLQMLEAHMQSYKGNDPLGEWERYIQWVEENFPENKEYLITLLEHLMKEFLDKK
KYHNDPRFISYCLKFAEYNSDLHQFFEFLYNHGIGTLSSPLYIAWAGHLEAQGELQHASA
VLQRGIQNQAEPREFLQQQYRLFQTRLTETHLPAQARTSEPLHNVQVLNQMITSKSNPGN
NMACISKNQGSELSGVISSACDKESNMERRVITISKSEYSVHSSLASKVDVEQVVMYCKE
KLIRGESEFSFEELRAQKYNQRRKHEQWVNEDRHYMKRKEANAFEEQLLKQ
HQVVETSHEDLPASQERSEVNPARMGPSVGSQQELRAPCLPVTYQQTPVNM
VVPPLANAISAALVSPATSQSIAPPVPLKAQTVTDSMFAVASKDAGCVNKS
```

e.g. P00750, P05067-5, A4_HUMAN or acccgtggtcg

Run BLAST | Reset

BLAST programs available on ExPASy:

- blastp: **protein** query -> **protein** sequence database
- blastn: **nucleotide** query -> **nucleotide** sequence database
- tblastn: **protein** query -> **nucleotide** sequence database
- blastx: **nucleotide** query -> **protein** sequence database

Arthropoda
Fungi
Mammalia
Metazoa
Primates
Rodentia
Vertebrata
Viridiplantae

Arabidopsis thaliana
Caenorhabditis elegans
Dictyostelium discoideum
Drosophila melanogaster
Escherichia coli
Homo sapiens
Mus musculus
Plasmodium falciparum
Rattus norvegicus
Saccharomyces cerevisiae

Choose a database

Protein databases:

UniProt Knowledgebase (UniProtKB)

Best Hit!

List of the matches

Clustal W (multiple alignment) 提交

Select up to...

Include query sequence

		Accession	Db	Description	Score	E-value
	1	P41695 (BUB1_YEAST)		Checkpoint serine/threonine-protein kinase ...	174	7e-45
	2	P47074 (MAD3_YEAST)		Spindle assembly checkpoint component MAD3 ...	60.5	8e-10
	3	P14680 (YAK1_YEAST)		Dual specificity protein kinase YAK1 OS=Sac...	51.2	7e-07
	4	P06245 (KAPB_YEAST)		cAMP-dependent protein kinase type 2 OS=Sac...	45.8	3e-05
	5	P06244 (KAPA_YEAST)		cAMP-dependent protein kinase type 1 OS=Sac...	45.1	4e-05

Next class

- **Neighbor-joining method (for constructing phylogenetic tree)** N. Saitou and M. Nei. The neighbor-joining Method: A new method for reconstructing hylogenetic tree. Mol Biol Evol. (1987) 4: 406-425.
- **UPGMA**: https://en.wikipedia.org/wiki/UPGMA
- **PSI-BLAST (The most often-used algorithm for sequence-profile alignment)** S. F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. (1997) 25, 3389-3402
- **Hidden Markov Model (for multiple sequence alignment)** Haussler, D., Krogh, A., Mian, I. S., & Sjölander, K. (1993). Protein modeling using hidden Markov models: Analysis of globins. In: Proceedings of the Hawaii International Conference on System Sciences volume 1 pp. 792-802.
- **Sequence profile** Gribskov, Mclanchlan, Eisenberg. Profile analysis: Detection of distantly related proteins. PNAS (1987) 84, 4355-58
- **Henikoff weight** Steven Henikoff and Jorja G. Henikoff, Position-based sequence weights, Journal of Molecular Biology. Volume 243, Issue 4, 4 November 1994, Pages 574-578

Next class

- **Profile-profile alignments:**

Anna R. Panchenko. Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Research, 2003, Vol. 31, No. 2 683.

Edgar & Sjolander, A comparison of scoring functions for protein sequence profile alignment. Bioinformatics (2004) 20, 1301-8

G Wang, R. Dunbrack JR. Scoring profile-to-profile sequence alignments. Protein Sci. 13:1612-1626, 2004