



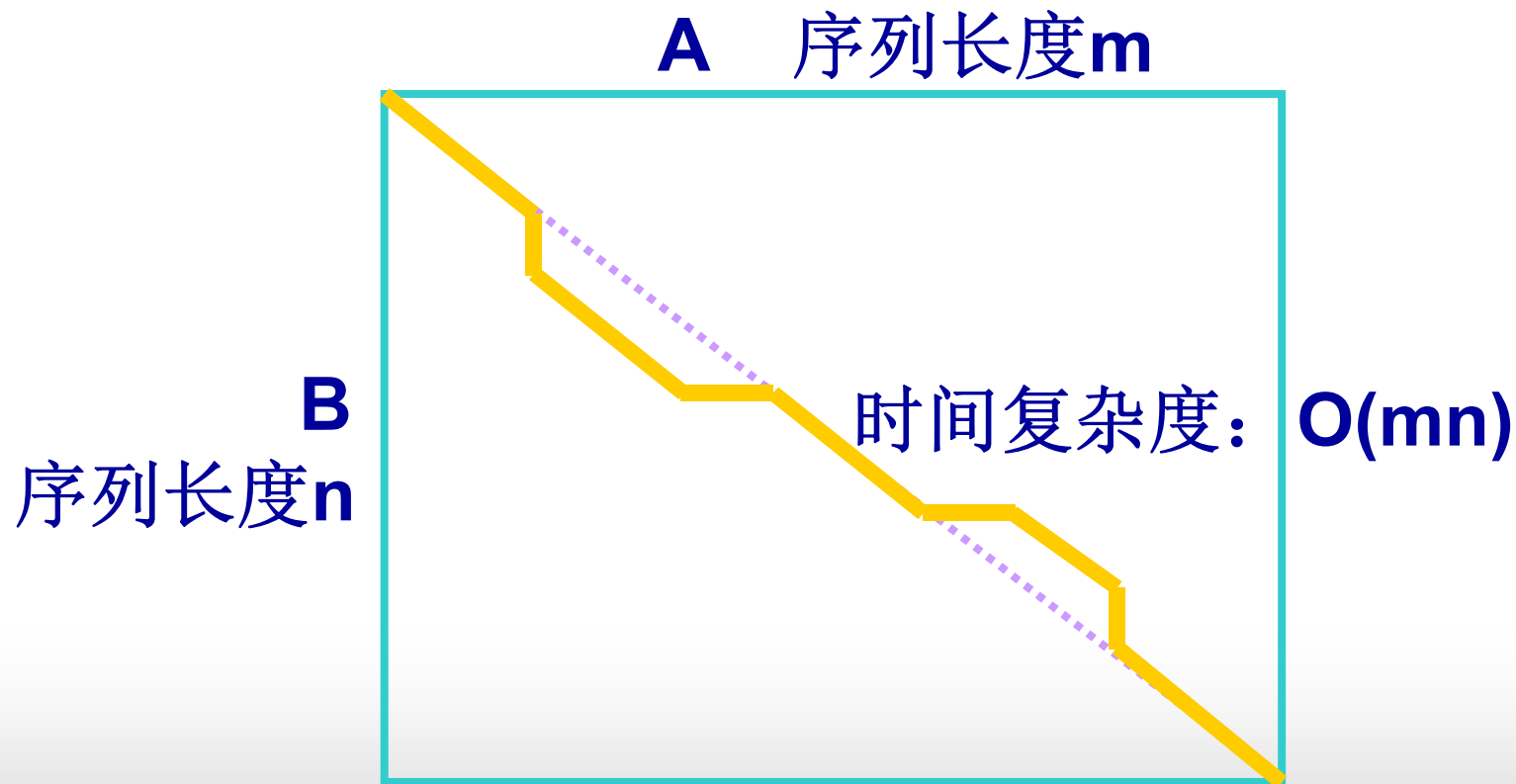
生物信息学

第六章 双序列比对 (2)

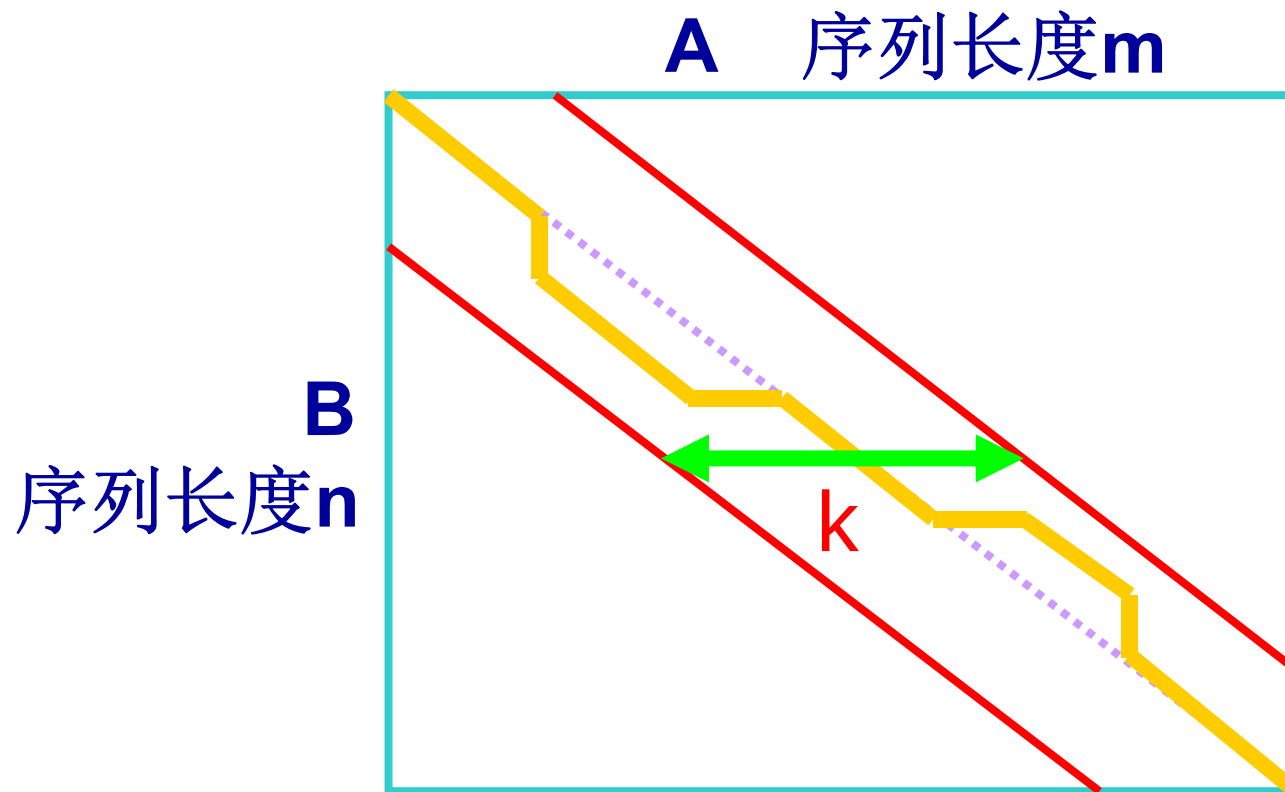
*K-tup*算法原理



- 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)

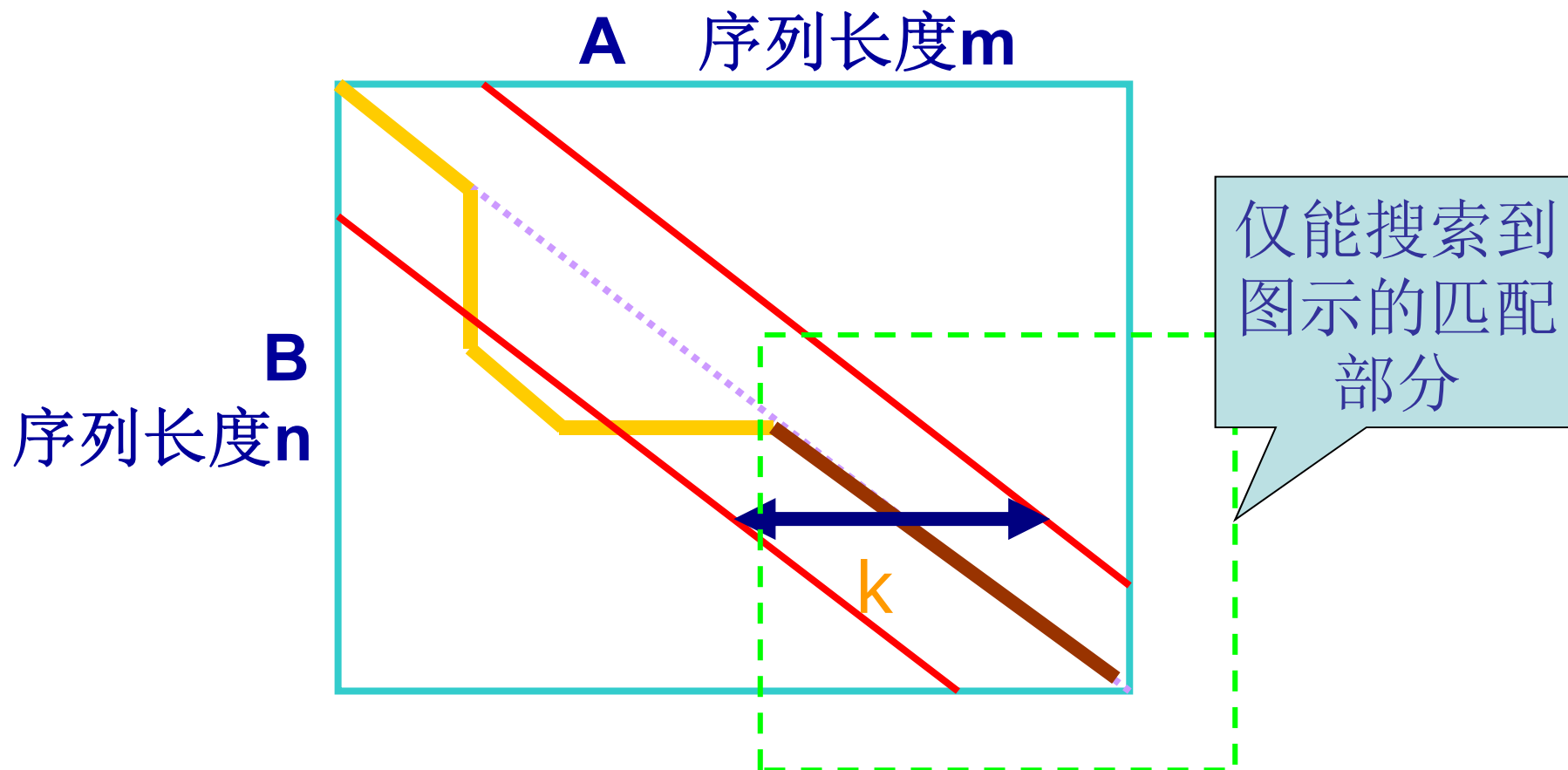


令 k 为一常数，
搜索限定区域
内的最优比对

时间复杂度: $O(kn)$



K -tup算法原理：缺点



关于“新冠病毒不是源自实验室”的科普 2020-02-06

讲正事儿之前，我们先按惯例闲话几句。“时近年节返故乡，过汉江，出黄冈。暂离江城，高堂迎道旁。黍米鸡粟忙遍尝，举杯酒，庆吉祥！”这是某粗通“简单的生物 ...

(8302)次阅读 | (21)个评论

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: 索引表构建



□ 以蛋白质序列为例

□ $k=1$

氨基酸	<i>tup</i> 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	<i>tup</i> 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	<i>tup</i> 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: NCSP**TA**

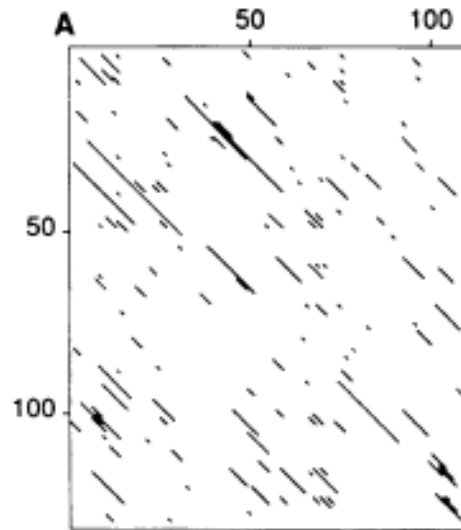


Protein2: ACSPRK

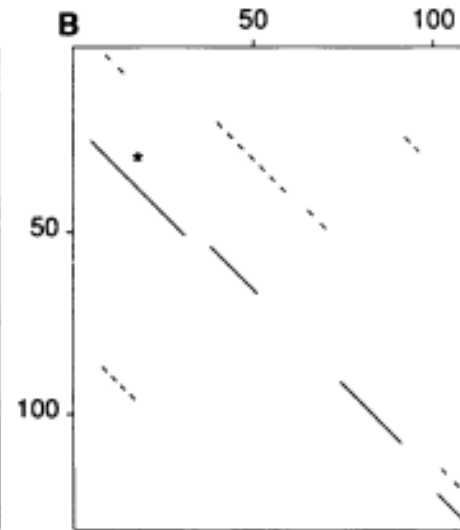
FASTA: 序列比对的过程



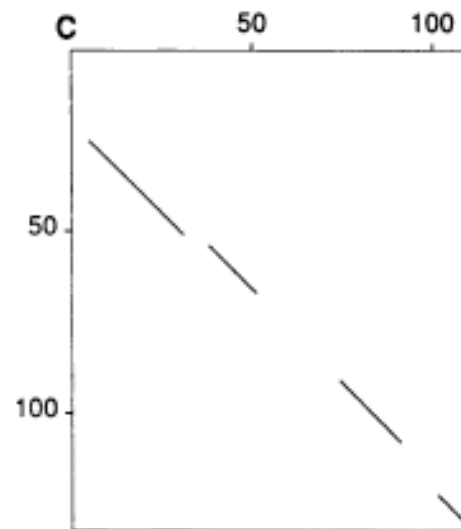
A. 找出完全匹配的区域



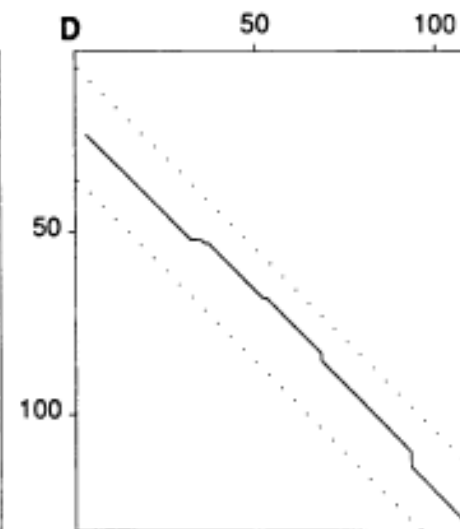
B. 保留分数最高的10个匹配



C. 将较好的匹配连接起来



D. 重新用动态规划算法计算最优比对



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**

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 useful reference after the course.

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: 发展与改进



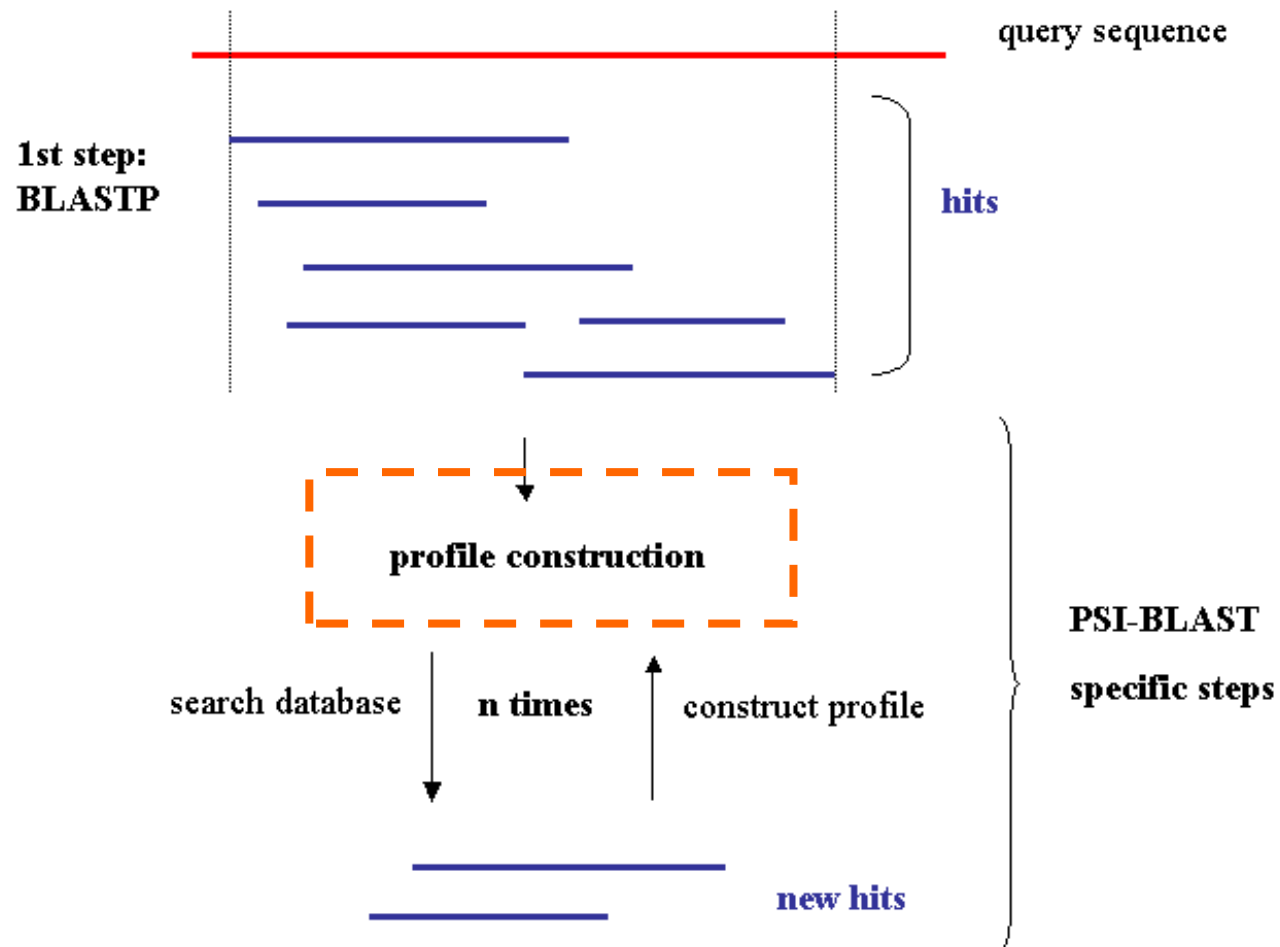
- ❑ 早期的BLAST版本：无空位罚分
- ❑ 新版本：Gap Penalties: Existence: 11, Extension: 1
- ❑ Psi-BLAST: 构建位点特异性矩阵
- ❑ Phi-BLAST: 包含特定模体的序列相似性搜索

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
LNMAYKFRTHSPEALEQLYPWECFVFCLIFGTFTNQI
HKWSHTYFGLPRWVTLLQDWHVILPRKHHRIHHVSP
HETYFCITTGVKVPRNFRLLLEELEEGQKGVGDGTVS
WGLEDDMTLTRWTGMIIGPPRTIYENRIYSLKIECG
PKYPEAPPFVRFTKINMNGVNSSNGVVDPRASVLA
KWQNSYSIKVVLQELRRLMMSKENMKLPQPPEGQC
YSN



Enter Query Sequence

Enter accession number, gi, or FASTA sequence ?

[Clear](#)

Query subrange ?

```
LVTLPLLNMMAYKFRTHSPEALEQLYPWECFVFCIIIFGTFTNQIHKWSHTYFGLPRWVTLQDWHVILP  
RKHHRIHHVSPHETYPFCITTTGVKVPFRNFRLLLEELEGQKGVGDGTVSUGLEDDMTLTRWTGMIIGPPR  
TIYENRIYSLKIECGPKYPEAPPFVRFVTKINMNGVNSSNGVVDPRATSVLAKWQNSYSIKVVLQELRRL  
MMSKENMKLPQPPEGQCYSN
```

From

To

Or, upload file

浏览...

Job Title

Enter a descriptive title for your BLAST search ?

Choose Search Set

Database

Protein Data Bank proteins(pdb) ?

Organism

Optional

Enter organism name or id-completions will be suggested

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

Entrez Query

Optional

Enter an Entrez query to limit search ?

Program Selection

Algorithm

☐ blastp (protein-protein BLAST)

☒ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm ?

BLAST

Search database **pdb** using **PSI-BLAST (Position-Specific Iterated BLAST)**



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 Structures Of Two Ubc (E2) Enzyme...	48.9	2e-06	S
NEW	☒	pdb 2F47 A	Chain A, Toxoplasma Gondii Ubiquitin Conjugating E	48.5	3e-06	S

打分矩阵及其含义



- ❑ Dayhoff: PAM系列矩阵
- ❑ Henikoff: BLOSUM系列矩阵
- ❑ 常用氨基酸打分矩阵: **BLOSUM62**

PAM系列矩阵



- ❑ Margaret Dayhoff, 1978
- ❑ Accepted point mutation (PAM): 可接受的点突变，氨基酸的改变不显著影响蛋白质的功能
- ❑ 进化模型：
 - ✿ 基本假设：中性进化，Kimura, 1968
 - ✿ 进化的对成性： $A \rightarrow B = B \rightarrow A$
 - ✿ 扩展性：通过对较短时间内氨基酸替代关系的计算来计算较长时间的氨基酸替代关系

PAM1矩阵的构建



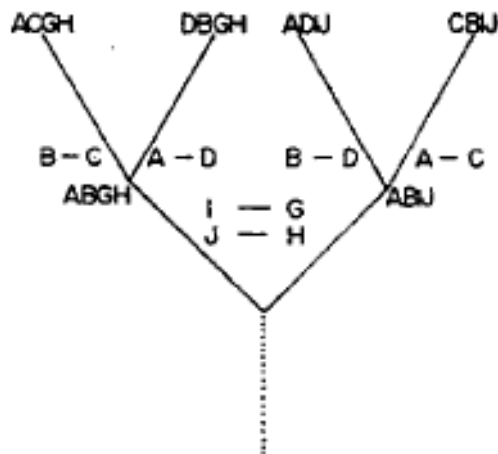
- ❑ 两个蛋白质序列的~1%氨基酸发生变化
- ❑ 定义进化时间以氨基酸的变异比例为准，而不是时间：各个蛋白质家族进化的速度并不相等
- ❑ $PAM2 = PAM1 * PAM1$
- ❑ $PAM3 = (PAM1)^3$
- ❑ $PAM250 = (PAM1)^{250}$

使用的序列数据集



- 34个蛋白质超家族
- 71个蛋白质分组
- 15,720个突变
- 序列相似性 > 85%

分子进化树



氨基酸替代矩阵

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

15,720

Step 2: 计算氨基酸出现概率



□ 蛋白质序列数据库中20种氨基酸的出现频率

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

$$P_F = 0.04$$

$$P_Y = 0.03$$

Step 3: 计算相对突变能力



□ 相对突变能力: **relative mutability, m_a**

$$m_a = \frac{f_a}{100 f P_a}$$

关于氨基酸a的突变数

100个位点的突变权值*所有的突变数*氨基酸a的概率



相对突变能力的计算

- 例如，对于A，突变数3644，则
 - ✿ $m_A = 3644 / (100 * 0.087 * 15720) = 0.0266$
- 对于R，突变数为1191，则
 - ✿ $m_R = 1191 / (100 * 0.041 * 15720) = 0.0173$
- 对于F，突变数为682，则
 - ✿ $m_F = 682 / (100 * 0.04 * 15720) = 0.0108$
- 对于Y，突变数512，则
 - ✿ $m_Y = 512 / (100 * 0.03 * 15720) = 0.0108$
- 令 $m_A = 100$ ，则
- $m_R = \sim 65$ ， $m_F = \sim 41$ ， $m_Y = \sim 41$

20种氨基酸的相对突变能力



Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Ala: 主观设定为100

总和: 1,468

Step 4: ~1%氨基酸突变



- M_{aa} 反映 a 没有突变成其他氨基酸的能力
- $M_{aa} = 1 - \sum M_{ab}$
- 问题:
- $\sum M_{ab}$ 可能会 $\gg 1\%$, 则
- 使用 λ , 使得 $\lambda \sum M_{ab}$ 约为 1% , M_{aa} 约为 99% , 则
- PAM1 矩阵:
- $M_{aa} = 1 - \lambda \sum M_{ab} \approx 99\%$

Step 5: 计算a->b的相对概率



□ P_a = 氨基酸a出现的概率

□ f_{ab} = a->b替代的总数

□ f_a = 所有与a相关的氨基酸替代数

□ $= \sum f_{ab}$ (a不等于b)

□ $f = \sum f_a$

□ 定义：相对概率 M'_{ab} ,

□ $M'_{ab} = \Pr(a \rightarrow b) = f_{ab}/f_a$

F->Y:

$$M'_{FY} = 260/682$$

Y->F:

$$M'_{YF} = 260/512$$

Step 6: λ 的计算

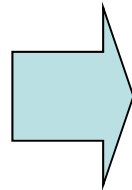


□ 20种氨基酸，突变总数1%

✿ 200,000个氨基酸，突变总数2,000

✿ $\lambda = 2000/1468 = \sim 1.36$

Asn	134	His	66	Asn	182	His	90
Ser	120	Arg	65	Ser	163	Arg	88
Asp	106	Lys	56	Asp	144	Lys	76
Glu	102	Pro	56	Glu	139	Pro	76
Ala	100	Gly	49	Ala	136	Gly	67
Thr	97	Tyr	41	Thr	132	Tyr	56
Ile	96	Phe	41	Ile	131	Phe	56
Met	94	Leu	40	Met	128	Leu	54
Gln	93	Cys	20	Gln	126	Cys	27
Val	74	Trp	18	Val	101	Trp	24



□ 连接概率 $M_{ab} = M'_{ab} * m_a$

2025,

$M_{FY} = 260/682 * 56 = 21$

$M_{YF} = 260/512 * 56 = 28$

ORIGINAL AMINO ACID

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
REPLACEMENT AMINO ACID	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	12	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1										0	0	1
E Glu	10	0	7	56	0	55	9865	4										0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	2	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947							3	4	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1													9946						
P Pro	13													1	9926	12	4	0	0	2
S Ser	26	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13						2	3	3	57	11	1	17	1	3	2	10	0	2	9901

$\sim 99\% * 20 * 10000$

$$1,954/20=97.7$$

PAM1矩阵，乘以10000

加和为10000

例：DNA打分矩阵PAM1的构建



□ 假设DNA四种碱基概率为：

✿ $P_A = P_T = 0.3, P_C = P_G = 0.2$

□ 若干序列比较发现1000个碱基替代

✿ $f = 1000$

✿ $f_A = 550, f_T = 550, f_C = 350, f_G = 550$

□ 相对突变能力：

✿ $m_A = 0.0183, m_T = 0.0183,$
 $m_C = 0.0175, m_G = 0.0275$

A				
C	100			
G	200	150		
T	250	100	200	
	A	C	G	T

$$m_a = \frac{f_a}{100 f P_a}$$

例：DNA打分矩阵PAM1的构建 (2)



□ PAM1, 总数40000个碱基, 400个突变

✿ $\lambda(m_A + m_T + m_C + m_G) = 400$

✿ $\lambda m_A = \lambda m_T = 90, \lambda m_C = 85, \lambda m_G = 135$

PAM1矩阵

	A	C	G	T
A	9910			
C	?	9915		
G			9865	
T				9910

例：DNA打分矩阵PAM1的构建 (3)



□ $\lambda m_A = \lambda m_T = 90$, $\lambda m_C = 85$, $\lambda m_G = 135$

✿ $\lambda m_{AC} = 16$, $\lambda m_{AG} = 33$, $\lambda m_{AT} = 41$

✿ $\lambda m_{CA} = 24$, $\lambda m_{CG} = 37$, $\lambda m_{CT} = 24$

✿ $\lambda m_{GA} = 49$, $\lambda m_{GC} = 37$, $\lambda m_{GT} = 49$

✿ $\lambda m_{TA} = 41$, $\lambda m_{TC} = 16$, $\lambda m_{TG} = 33$

PAM1矩阵

	A	C	G	T
A	9910	24	49	41
C	16	9915	37	16
G	33	37	9865	33
T	41	24	49	9910

A				
C	100			
G	200	150		
T	250	100	200	
	A	C	G	T

PAM2矩阵



- 基本假设：每个氨基酸的突变的概率独立于前次突变。因此， $PAM2 = PAM1 * PAM1$

	aa1	aa2	aa3
aa1	a	b	c
aa2	d	e	f
aa3	g	h	i

X

	aa1	aa2	aa3
aa1	a	b	c
aa2	d	e	f
aa3	g	h	i

	aa1	aa2	aa3
aa1	A	B	C
aa2	D	E	F
aa3	G	H	I

$$A = a^2 + bd + cg + \dots$$

$$B = ab + be + ch + \dots$$

$$C = ac + bf + ci + \dots$$

$$D = da + ed + fg + \dots$$

ORIGINAL AMINO ACID

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
REPLACEMENT AMINO ACID	A Ala	13	6	9	9	5	8	9	12	5	8	6	7	7	4	11	11	11	2	4	9
	R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
	N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
	D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
	E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
	G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
	H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
	I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
	L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
	K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
	M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
	F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
	P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
	S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
	T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
	W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
	Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
	V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

PAM250矩阵，乘以100

PAM250矩阵（前景）



- ❑ PAM250: 250%的期望的突变
- ❑ 蛋白质序列仍然有15-30%左右的相似性，例如：
 - ✿ F->F: 32%

Expected % similarity

Other PAM matrices:	PAM 120 – 40%	}	Use for similar sequences
	PAM 80 – 50%		
	PAM 60 – 60%		
PAM250 – 15-30% similarity.			

当前使用的PAM250



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0	-8
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1	-8
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	-8
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1	-8
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3	-8
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	-8
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	-8
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	-8
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1	-8
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-8
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	-8
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	-8
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	-8
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	-8
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1	-8
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0	-8
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	-8
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	-8
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2	-8
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	-8
B	0	-1	2	3	-4	1	3	0								0	0	-5	-3	-2	3	2	-1	-8
Z	0	0	1	3	-5	3	3	0								0	-1	-6	-4	-2	2	3	-1	-8
X	0	-1	0	-1	-3	-1	-1	-1								0	0	-4	-2	-1	-1	-1	-1	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

对称分布！

Recall: 替代矩阵的模型



- 考虑长度为 n 的序列 x 和长度为 m 的序列 y
- 令 x_i 为 x 序列中的第 i 位； y_j 为 y 序列中的第 j 位
- 对于不相关或者随机的模型 R ，假设 x_i 出现的频率为 q_{xi} ， y_j 出现的频率为 q_{yj} ，则两条序列匹配的概率为：

$$P(x, y | R) = \prod_i q_{xi} \prod_j q_{yj}$$

替代矩阵的模型 (2)



- 对于备择假设/匹配模型M，两个字符匹配的概率为连接概率 p_{ab} ，因此：

$$P(x, y | M) = \prod_i P_{xiyi}$$

- 两个似然性值之间的比值称为几率值(odds ratio):

$$\frac{P(x, y | M)}{P(x, y | R)} = \frac{\prod_i p_{xiyi}}{\prod_i q_{xi} \prod_i q_{yi}} = \prod_i \frac{p_{xiyi}}{q_{xi} q_{yi}}$$

替代矩阵的模型 (3)



- 连乘->连加；取对数，求对数几率值(log-odds ratio):

$$S = \sum_i s(x_i, y_i)$$

- 并且

$$s(a, b) = \log\left(\frac{P_{ab}}{q_a q_b}\right)$$

PAM系列矩阵的构建思想



- 对于上面的公式，这里，
- $S(a,b)=[S(a,b|t)*S(b,a|t)]^{1/2}$; 时间单位：序列相似性的%
- $S(a,b|t)=\log[P(a|b,t)/q_a]$
- $P_{ab}=[P(a|b,t)*P(b|a,t)]=(M_{ab}*M_{ba})^{1/2}$
- 因此，
- $S(a,b) = 1/2\log(M_{ab}*M_{ba})/q_aq_b$

PAM250的计算



- 计算相关比率值 R_{ab} : 通过进化形成的关联的两条蛋白质序列上的两个氨基酸有关联的几率 vs. 随机的两条蛋白质序列上, 两个氨基酸有关联的几率
- M_{ab} = 相关的蛋白质序列上a替换成b的概率
- P_a^{ran} = 氨基酸随机出现的概率, 因此,
- $P_a^{\text{ran}} = f_a$;
- $R_{ab} = M_{ab}/f_a$

因此



- 根据 $R_{ab} = M_{ab} / f_a$
- 考虑 F $\rightarrow$ Y 的替代
- $M_{FY} = 0.15$
- $f_F = 0.04$
- 得 $R_{FY} = 0.15 / 0.04 = 3.75$
- $\text{Log}_{10}(R_{FY}) = \log_{10}(3.75) = 0.57$
- $10^{0.57} = 5.7$
- 考虑 Y $\rightarrow$ F 的替代
- $M_{YF} = 0.20$
- $f_Y = 0.03$
- 得 $R_{YF} = 0.20 / 0.03 = 6.7$
- $\text{Log}_{10}(R_{YF}) = \log_{10}(6.7) = 0.83$
- $10^{0.83} = 8.3$

因此，平均值为 $(5.7 + 8.3) / 2 = 7$



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0	-8
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1	-8
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	-8
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1	-8
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3	-8
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	-8
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	-8
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	-8
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1	-8
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-8
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	-8
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	-8
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	-8
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	-8
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1	-8
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0	-8
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	-8
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	-8
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2	-8
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	-8
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1	-8
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1	-8
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

打分矩阵的使用



- ❑ PAM250: ~15-30%的序列相似性
- ❑ PAM120: ~40%的序列相似性
- ❑ PAM80: ~50%
- ❑ PAM60: ~60%
- ❑ 如何选择最合适的矩阵?
- ❑ 遍历尝试...

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

PAM矩阵的问题及改进



□ PAM系列矩阵存在的问题：

- ✿ 氨基酸的打分矩阵，不关心核酸
- ✿ 进化模型的构建需要系统发育树的分析，因此，成为一个循环论证的问题：序列比对->矩阵构建->打分，进行新的序列比对
- ✿ 数据集很小

□ 打分矩阵的改进

- ✿ 选用大量的序列数据，构建矩阵（JTT矩阵）
- ✿ BLOSUM系列矩阵
- ✿ 核酸的打分矩阵

BLOSUM系列矩阵



- ❑ **BLOCK**: 蛋白质家族保守的一段氨基酸，无gap，一般几个~上百个氨基酸
- ❑ **Prosite**家族：至少有一个BLOCK存在于该家族的所有蛋白质序列中
- ❑ **BLOSUM62**: 序列的平均相似性为62%的BLOCK构建的打分矩阵
- ❑ 最被广泛使用的氨基酸打分矩阵系列

BLOSUM系列矩阵



- 1992, Steven Henikoff
- 利用"Block" (模块) 计算矩阵
 - ✿ ~500组相似的蛋白质
 - ✿ ~2,000个Blocks

EcORF708	EEVI-----	AAL E E G E F D L A I G L S V F H H I V H L H G I D E V K R L L S R L A D V T Q A V I L E L A V K E E
DmPka-C2	-----	T N L E N Y I T R A V L G N G S F G T V M L V R E K S G K N Y Y A A K M M S K E D L V R L K
DmCG12069	-----	T G L D D Y E I K A T L G S G S F G K V Q L V R E R E S G V Y Y A S K Q L S K D Q I V K T K
ScTPK2	-----	Y T L H D F Q I M R T L G T G S F G R V H L V R S V H N G R Y Y A I K V L K K Q Q V V K M K
ScTPK1	-----	Y S L Q D F Q I L R T L G T G S F G R V H L I R S R H N G R Y Y A M K V L K K E I V V R L K
ScTPK3	-----	Y S L S D F Q I L R T L G T G S F G R V H L I R S N H N G R F Y A L K T L K K H T I V K L K
Cekin-1	BLOCK →	A C L D D F D R I K T L G T G S F G R V M L V K H K Q S G N Y Y A M K I L D K Q K V V K L K
DmPka-C1	-----	A A L D D F E R I K T L G T G S F G R V M I V Q H K P T K D Y Y A M K I L D K Q K V V K L K
HsPKACg	-----	A S S D Q F E R L R T L G M G S F G R V M L V R H Q E T G G H Y A M K I L N K Q K V V K M K
HsPKACa	-----	A H L D Q F E R I K T L G T G S F G R V M L V K H K E T G N H Y A M K I L D K Q K V V K L K
HsPKACb	-----	A G L E D F E R K K T L G T G S F G R V M L V K H K A T E Q Y Y A M K I L D K Q K V V K L K
DmPKA-C3	-----	D G N E T D D E E D D D E S E S S S V Q T A K G V R K Y H L D D Y Q I I K T V G T G T F G R V C L C R D R I S E K Y C A M K I L A M T E V I R L K
HsPRKX	-----	Y S L Q D F D T L A T V G T G T F G R V H L V K E K T A K H F F A L K V M S I P D V I R L K
HsPRKY	-----	Y R L Q D C D A L V T M G T G T F G R V H L V K E K T A K H F F A L K V M S I P D V I R R K

计算方法



□ 例：给定一系列数据，9个A，1个S

✿ $f_{AA} = (8+1)*8/2=36; f_{AS}/f_{SA}=9; f_{SS}=0$

□ $q_{ij} = f_{ij} / \sum_{i=1}^{20} \sum_{j=1}^i f_{ij}$

✿ $q_{AA}=36/45=0.8, q_{AS}=9/45=0.2$

□ $p_i = q_{ii} + \sum_{j \neq i} q_{ij}/2$

✿ $p_A=(36+9/2)/45=0.9; p_S=0.1$

✿ $e_{AS}=p_A p_S + p_S p_A = 0.9*0.1*2=0.18$

□ $s_{ij} = \log_2(q_{ij}/e_{ij})$

✿ $S_{as}=\log_2(0.2/0.18)=0.152$

Relative entropy

$$H = \sum_{i=1}^{20} \sum_{j=1}^i q_{ij} \times s_{ij}$$

Expected Score

$$E = \sum_{i=1}^{20} \sum_{j=1}^i p_i \times p_j \times s_{ij}$$

多列数据



- 两列数据：16个A， 4个S
- $p_A=16/20=0.8$; $p_S=0.2$
- $e_{AS}=p_Ap_S+p_Ap_S=0.8*0.2*2=0.32$
- $q_{AS}=(9+21)/90 =0.333$
- $S_{AS}=\log_2(0.333/0.32)=0.059$

AA
AA
AA
AA
AA
AA
AA
AS
AS
SS

AA
AA
AF
FA
FA
AA
AA
AF
AS
SS

- 两列数据：13个A， 3个S
- $p_A=13/20=0.65$; $p_S=0.15$
- $e_{AS}=p_Ap_S+p_Ap_S=0.65*0.15*2=0.195$
- $q_{AS}=(7+12)/90 =0.211$
- $S_{AS}=\log_2(0.195/0.211)=-0.11$

BLOSUM62



```
# Matrix made by matblas from blosum62.iij
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks_5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
  A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
G  0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  2  2  2  2  7  1  2  2  1  1  1
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2 -1  2  2  2  2  2  1  2  2  1  1  1
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -1  2  2  2  2  2  1  2  2  1  1  1
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1  2  2  2  2  2  1  2  2  1  1  1
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1  2  2  2  2  2  1  2  2  1  1  1
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  2  2  2  2  2  1  2  2  1  1  1
```

Lod ratios are multiplied by a scaling factor of 2 and then rounded to the nearest integer value

BLOSUM vs. PAM



□ BLOSUM系列比PAM系列好

Matrix aligned	Program	Residue positions missed*	
		All positions	Side chains
	MSA	12	6
PAM 120	MULTALIN	31	22
PAM 160	MULTALIN	30	22
PAM 250	MULTALIN	30	22
+6/-1	MULTALIN	34	26
BLOSUM 45	MULTALIN	9	5
BLOSUM 62	MULTALIN	6	4
BLOSUM 80	MULTALIN	9	6

序列比对的显著性检验



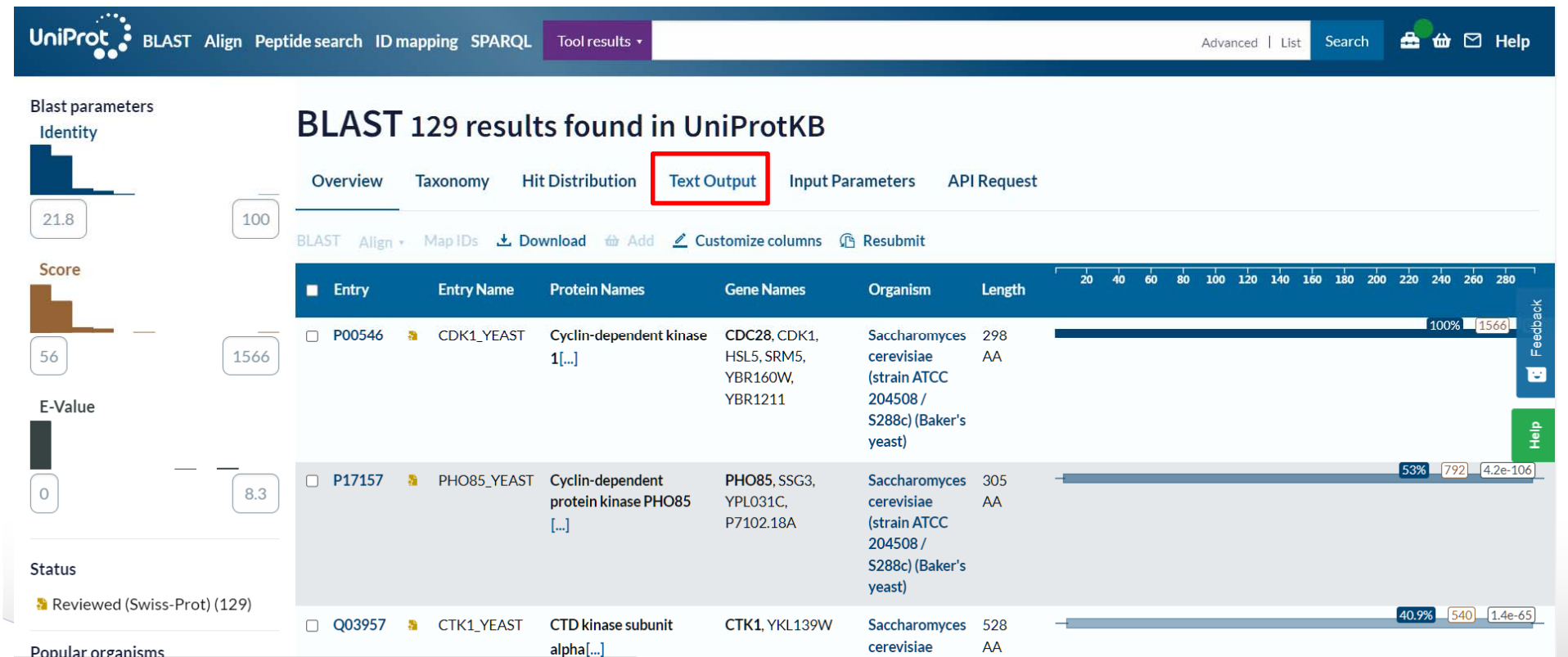
- 两条序列在进化上显著相关：
- Sander & Schneider, 1991, 序列相似性的显著性于匹配的序列长度 k 有一定关系：
 - ✿ (1) 当匹配序列 $k < 10$ 时，无相似性
 - ✿ (2) $10 < k < 80$ 时，域值为 $290.15k^{-0.562}\%$
 - ✿ (3) $k > 80$ 时，域值为24.8%
- Brenner, 1998,
 - ✿ (1) $k \sim 150$, $> 25\%$
 - ✿ (2) $k \sim 70$, $> 40\%$
 - ✿ (3) 10%~20%时，无相似性

显著性计算



□ 如何计算序列比对的显著性？

酵母CDK1在酵母中的序列比对结果



比对结果



SP:P21965MCK1_YEAST Protein kinase MCK1 OS=Saccharomyces cerevisi...	127	9e-35
SP:P19454CSK22_YEAST Casein kinase II subunit alpha' OS=Saccharom...	124	7e-34
SP:P15790CSK21_YEAST Casein kinase II subunit alpha OS=Saccharomy...	121	2e-32
SP:P32581IME2_YEAST Meiosis induction protein kinase IME2/SME1 OS...	117	6e-30
SP:P06782SNF1_YEAST Carbon catabolite-derepressing protein kinase...	111	7e-28
SP:Q12222YGK3_YEAST Glycogen synthase kinase-3 homolog YGK3 OS=Sa...	108	2e-27
SP:P22216RAD53_YEAST Serine/threonine-protein kinase RAD53 OS=Sac...	104	2e-25
SP:Q01389BCK1_YEAST Serine/threonine-protein kinase BCK1/SLK1/SSP...	104	2e-25
SP:P27636CDC15_YEAST Cell division control protein 15 OS=Saccharo...	100	3e-24

□ SNF1_YEAST的结果:

⚙ Score: 111

⚙ E-value: 7e-28

□ 问题:

⚙ 如何计算Score?

⚙ 如何计算E-value? 该值是何意义?

贝叶斯方法：模型比较



- 考虑两个模型：两条序列无关的概率为 $P(R|x,y)$ 以及两条序列相关的概率为 $P(M|x,y)$
- 两个模型的前向概率为： $P(M)$ 和 $P(R)$ ，存在：
 - ✿ $P(R)=1-P(M)$
- 因此：

$$\begin{aligned} P(M | x, y) &= \frac{P(x, y | M)P(M)}{P(x, y)} = \frac{P(x, y | M)P(M)}{P(x, y | M)P(M) + P(x, y | R)P(R)} \\ &= \frac{P(x, y | M)P(M)}{1 + \frac{P(x, y | R)P(R)}{P(x, y | M)P(M)}} \end{aligned}$$

贝叶斯方法：模型比较(2)



$$\text{令 } S' = S + \log\left(\frac{P(M)}{P(R)}\right) \text{ 则 } S = \log\left(\frac{P(x, y | M)}{P(x, y | R)}\right)$$

得 $P(M|x,y)=\sigma(S')$, 其中,

$$\sigma(x) = \frac{e^x}{1 + e^x} \quad \text{Sigmoid方程}$$

求近似值



□ $E(S) \approx K m n e^{-\lambda S} = m n 2^{-S};$

$$\sum_{a,b} q_a q_b e^{\lambda S(a,b)} = 1$$

□ S : bit分值, 有公式:

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

□ R , raw分值, 根据打分矩阵直接得到的分数

SNF1



>SP:P06782 SNF1_YEAST Carbon catabolite-derepressing protein kinase OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) OX=559292
GN=SNF1 PE=1 SV=1
Length=633

Score = 111 bits (277), Expect = 7e-28

Identities = 87/296 (29%), Positives = 144/296 (49%), Gaps = 44/296 (15%)

Query	5	LANYKRLEKVGEGTYGVVYKALDLRPGQGQRVVAL-KKIRLESEDEGVPSTAIRESLLK	63
		+ NY+ ++ +GEG++G V A GQ + + KK+ +S+ +G REIS L+	
Sbjct	52	IGNYQIVKTLGEGSFGKVKLAYHTTTGQKVALKIINKKVLAKSDMQG---RIEREISYLR	108
Query	64	ELKDDNIVRLYDIVHSDAHKLYLVFEFLDLKRYMEGIPKDQPLGADIVKKFMMQLCKG	123
		L+ +I++LYD++ S ++ +V E+ +L Y+ + +D+ + ++F Q+	
Sbjct	109	LLRHPHIKLYDVIKSK-DEIIMVIEYAGNELFDYI--VQRDK-MSEQEARRFFQQIISA	164
Query	124	IAYCHSHRILHRDLKPQNLINKDGNLKLDFGLARAF--GVPLRAYTHEIVTLWYRAPE	181
		+ YCH H+I+HRDLKP+NLL+++ N+K+ DFGL+ G L+ + Y APE	
Sbjct	165	VEYCHRHKIVHRDLKPENLLLDEHLNVKIADFGLSNIMTDGNFLKT---SCGSPNYAAPE	221
Query	182	VLLGGKQYSTGVDTSIGCI-FAEMCNRKPIFSGDSEIDQIFKIFRVLGTPNEAIWPDIV	240
		V+ G VD WS G I + +C R P D I +FK I +	
Sbjct	222	VISGKLYAGPEVDVWSCGVILYVMLCRRLPF--DDESIPVLFK-----NISNGVY	269

因此，上例



```
>SP:P06782 SNF1_YEAST Carbon catabolite-derepressing protein kinase OS=Saccharomyces  
cerevisiae (strain ATCC 204508 / S288c) OX=559292  
GN=SNF1 PE=1 SV=1  
Length=633
```

```
Score = 111 bits (277), Expect = 7e-28
```

```
Identities = 87/296 (29%), Positives = 144/296 (49%), Gaps = 44/296 (15%)
```

□ **R=277**

上例 (2)



Lambda	K	H
0.323	0.142	0.434

Gapped

Lambda	K	H
0.267	0.0410	0.140

Effective search space used: 590525104

Database: uniprotkb_swissprot

Posted date: Dec 14, 2022 02:19 PM

Number of letters in database: 205,548,017

Number of sequences in database: 568,744

- ❑ $\lambda = 0.267$
- ❑ $K = 0.0410$
- ❑ $m = 298$
- ❑ $n = 590,525,104$

上例 (3)



$$\begin{aligned} S &= \frac{0.267 * 277 - \ln(0.041)}{\ln(2)} \\ &= \frac{73.959 - (-3.194)}{0.693} \approx 111 \end{aligned}$$

$$\begin{aligned} E &= 298 * 590525104 * 2^{-111} \\ &= 6.78e - 23 = 7e - 23 \end{aligned}$$

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？
- ❑ 直系同源序列的确定：Reciprocal Best Hits
- ❑ 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

BLAST Align Peptide search ID mapping SPARQL UniProtKB

Advanced | List Search

Help

Function

Names & Taxonomy

Subcellular Location

Phenotypes & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

P41695 · BUB1_YEAST

Protein ⁱ	Checkpoint serine/threonine-protein kinase BUB1	Amino acids	1021
Gene ⁱ	BUB1	Protein existence ⁱ	Evidence at protein level
Status ⁱ	UniProtKB reviewed (Swiss-Prot)	Annotation score ⁱ	5/5
Organism ⁱ	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)		

Entry Feature viewer Publications External links History

BLAST Download Add Add a publication Entry feedback

Functionⁱ

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

获得FASTA序列



Function

Names & Taxonomy

Subcellular Location

Phenotypes & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

P41695 · BUB1_YEAST

Proteinⁱ Checkpoint serine/threonine-protein kinase BUB1

Amino acids 1021

Geneⁱ BUB1

Protein existenceⁱ Evidence at protein level

Statusⁱ UniProtKB reviewed (Swiss-Prot)

Annotation scoreⁱ 5/5

Organismⁱ Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)

Entry Feature viewer Publications External links History

BLAST Download Add Add a publication Entry feedback

Text

FASTA (canonical)

FASTA (canonical & isoform)
JSON

Misc
XML

Preser
RDF/XML

Cata
GFF

```
>sp|P41695|BUB1_YEAST Checkpoint serine/threonine-protein kinase BUB1
MNLDLGSTVRGYESDKDTFPQSKGVSSSQKEQHSQLNQTKIAYEQRLNLDLDDPLDL
FLDYMIWISTSYIEVDSESGQEVLRSTMERCLYIQDMETRYNDPRFLKIWIWYINFLS
NNFHESNTFKYMFNKGIGTKLSLFYEEF SKLENAQFFLEAKVLLLEGAENNCRPYNRL
LRSLSNYEDRLREMNIVENQNSVPDSRERLKGRLIYRTAPFFIRKFLTSSLMTDDKENRA
NLNSNVGVGKSAPNVYQDSIVVADF KSETERLNLNSSKQPSNQRLKNGNKKTSIYADQKQ
SNNPVYKLINTPGRKPERIVFNFLIYPENDEEFNTEEILAMIKGLYKVQRRGKKHTEDY
TSDKNRKKRKL DVL VERRQDLPSSQPPVVPKSTRIEVF KDDDNPSQSSTHHKNTQVQVQT
TSILPLKPVVDGNLAHETPVKPSLT SNASRSPTVTAF SKDAINEVF SMFNQHYSTPGALL
DGDDTTTSKFNVFENFTQEF TAKNIEDLTEVKDPKQETVSQQTSTNETNDRYERLSNSS
TRPEKADYMTPIKETTETDVVPIIQTPKEQIRTEDKKSGDNTETQTQTSTTIQSSPFLT
QPEPQAEKLLQTAEHSEKSEHYPTIIPF TKIKNQPPV IENPLSNNLRAKFLSEISPP
LFQYNTFYNNQELKMSSLLKKIHRVSRNENKNPIVDFKKTGDLYCIRGELGEGGYATVY
LAESSQGHRLALKVEKPASVWEYIIMSQVEFRLRKSTILKSIINASALHLFLDESYLVN
YASQGTVLDLINLQREKAIDGNGIMDEYLCMFITVELMKVLEKIHEVGIIHGDLKPDNCM
IRLEKPGEP LGAHYMRNGEDGWENKGIYLD FGRSFDMTLLPPGTFKFSNWKADQQDCWE
MRACKPWSYEADYYGLAGVIHSM LFGKF IETIQLQNGRCKLKNPFKRYWKKEIWGVIFDL
LLNSGQASNQALPMTEKIVEIRNLIESHLEQHAENHLRNVL SIEEELSHFQYKGPSSRR
F
```

酵母的同源序列：旁系同源序列



BLAST

Find a protein sequence to run BLAST sequence similarity search by UniProt ID (e.g. P05067 or A4_HUMAN or UPI0000000001).

UniProt IDs



OR

Enter one or more sequences (20 max). You may also [load from a text file](#).

```
>sp|P41695|BUB1_YEAST Checkpoint serine/threonine-protein kinase BUB1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) OX=559292
MNLDLGSTVRGYESDKDTFPQSKGVSSSQKEQHSQLNQTKIAYEQRLNDLEDMDPLDL
FLDYMIWISTSYIEVDSESGQEVLRSTMERCLIIYIQDMETYNDRFLKIIWIYINLFLS
NNFHESENTFKYMFNKGIGTKLSLFYEEFSKLLENAQFFLEAKVLLLEGAENNCRPYNRL
LRSLSNYEDRLREMNIVENQNSVPDSRERLKGRLIYRTAPFFIRKFLTSSLMTDDKENRA
NLNSNVGVGKSAPNVYQDSIVVADFKSETERLNLNSSKQPSNQLKNGNKKTSIYADQKQ
SNMPVYKLINTPGRKPERIVFNFLIYPENDEEFNTEEILAMIKGLYKVQRRGKKHTEDY
TSDKNRKKRKLVDLVERRQDLPSSQPPVVPKSTRIEVFKDDDNPSQSTHHKNTQVQVQTT
```

Your input contains 1 sequence

Target database

UniProtKB Swiss-Prot

Restrict by taxonomy

Enter taxon names or IDs to include

Saccharomyces cerevisiae [4932] x

重置

Run BLAST

Name your BLAST job

Mad3: 旁系同源序列



BLAST Align Peptide search ID mapping SPARQL

Tool results

Advanced | List

Search

Help

Blast parameters

Identity



17

100

Score



60

5323

E-Value



0

9.9

Status

BLAST 193 results found in UniProtKB

Overview Taxonomy Hit Distribution Text Output Input Parameters API Request

BLAST Align Map IDs Download Add Customize columns Resubmit

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	
☐ P41695	BUB1_YEAST	Checkpoint serine/threonine-protein kinase BUB1[...]	BUB1, YGR188C, G7542	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	1,021 AA	100% 5323
☐ P47074	MAD3_YEAST	Spindle assembly checkpoint component MAD3[...]	MAD3, YJL013C, J1341	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	515 AA	35% 495

>SP:P47074 MAD3_YEAST Spindle assembly checkpoint component MAD3 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) OX=559292

GN=MAD3 PE=1 SV=1

Length=515

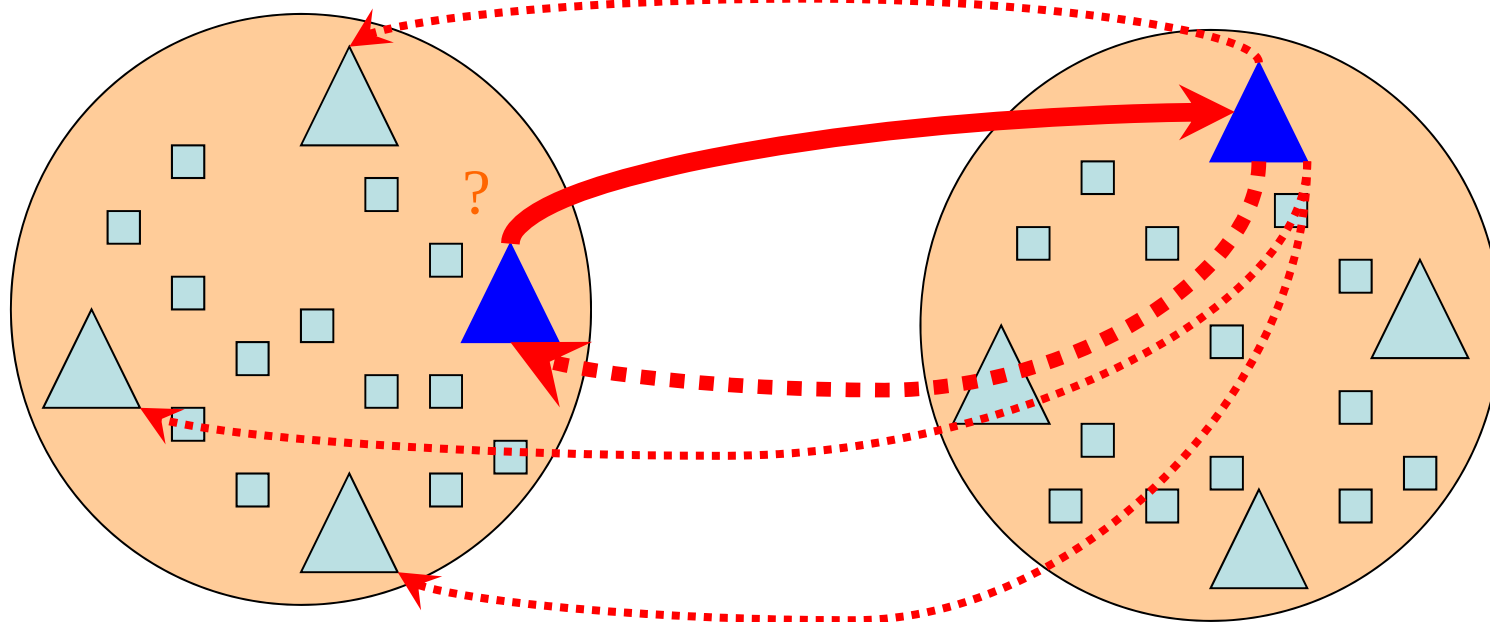
Score = 195 bits (495), Expect = 6e-54

Identities = 124/354 (35%), Positives = 196/354 (55%), Gaps = 33/354 (9%)

```

Query  35  QLNQTKIAYEQRLNDLEDMDPLDLFLDYMIWISTSYIEVDSESGQEVLRSTMERCLIIY  94
      ++NQ K ++EQRL+++L + DP+ L+L+Y+ W++ +Y + S Q + + +ERCL +
Sbjct  55  EINQVKSSFQRLIDELPALSDPITLYLEYIKWLNNAYPQ-GGNSKQSGMLTLLERCLSH  113
    
```

Reciprocal Best Hits



直系同源序列：Reciprocal Best Hits

人类同源序列：直系同源序列



BLAST

Find a protein sequence to run BLAST sequence similarity search by UniProt ID (e.g. P05067 or A4_HUMAN or UPI0000000001).

UniProt IDs



OR

Enter one or more sequences (20 max). You may also [load from a text file](#).

```
>sp|P41695|BUB1_YEAST Checkpoint serine/threonine-protein kinase BUB1 OS=Saccharomyces cerevisiae (strain ATCC 204508 / S288c) OX=559292
MNLDLGSTVRGYESDKDTFPQSKGVSSSQKEQHSQNLNQTAKIAYEQRLNLDLEDMDPLDL
FLDYMIWISTSYIEVDSESGQEVLRSTMERCLIIYIQDMETYNDRFLKIWIWYINLFLS
NNFHESENTFKYMFNKGIGTKLSLFYEEFSKLENAQFFLEAKVLELGAENNCRPYNRL
LRSLSNYEDRLREMNIVENQNSVPDSRERLKGRLIYRTAPFFIRKFLTSSLMTDDKENRA
NLNSNVGVGKSAPNVYQDSIVVADFKESETERLNLNSSKQPSNQRLKNGNKKTSIYADQKQ
SNNPVYKLINTPGRKPERIVFNFLIYPENDEEFNTEELAMIKGLYKVQRRGKKHTEDY
TSDKNRKKRKLVDLVERRQDLPSSQPPVVPKSTRIEVFKDDDNPSQSTHHKNTQVQVQTT
```

Your input contains 1 sequence

Target database

UniProtKB Swiss-Prot

Restrict by taxonomy

Enter taxon names or IDs to include

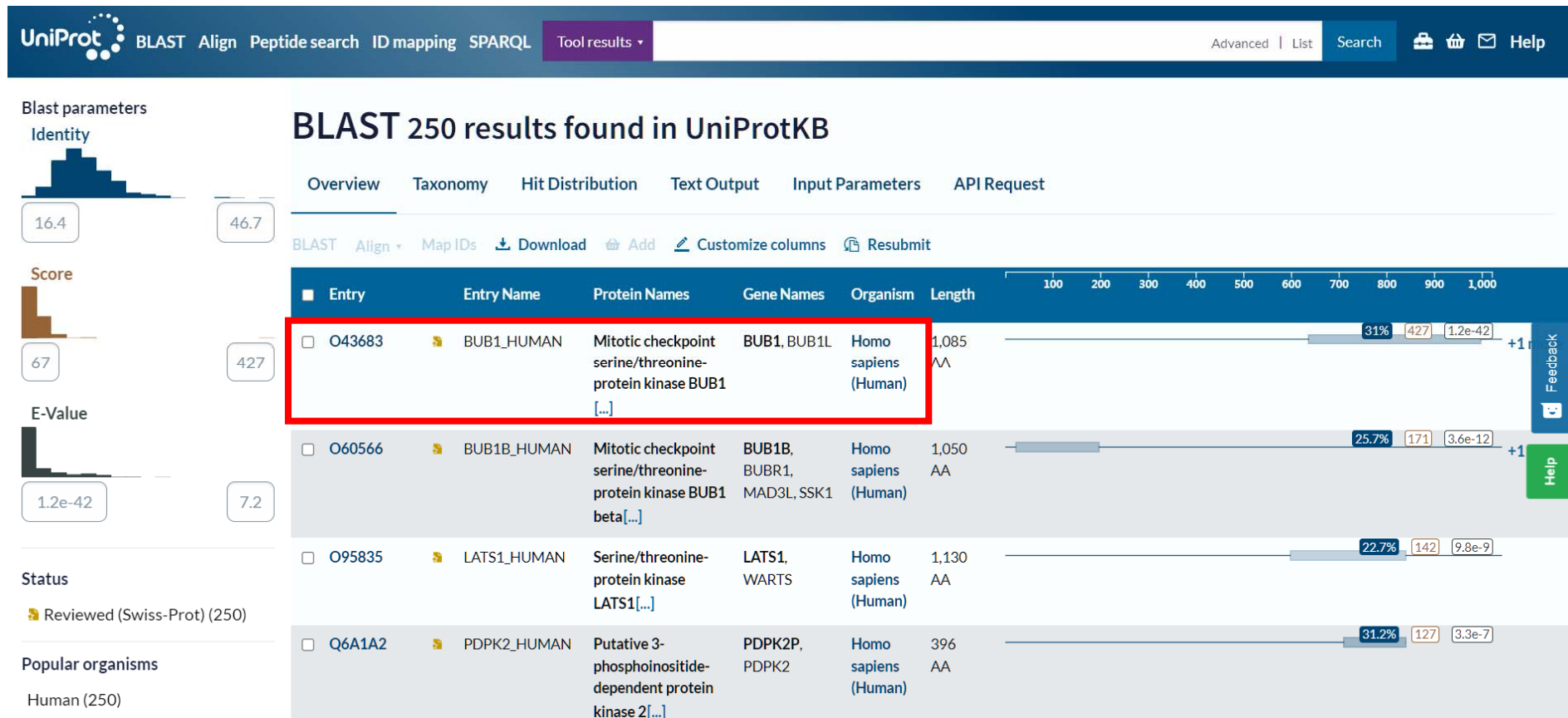
Homo sapiens [9606] x

重置

Run BLAST

Name your BLAST job

人类Bub1?



在酵母中做比对



BLAST

Find a protein sequence to run BLAST sequence similarity search by UniProt ID (e.g. P05067 or A4_HUMAN or UPI0000000001).

UniProt IDs



OR

Enter one or more sequences (20 max). You may also [load from a text file](#).

```
>sp|O43683|BUB1_HUMAN Mitotic checkpoint serine/threonine-protein kinase BUB1 OS=Homo sapiens OX=9606 GN=BUB1 PE=1 SV=1
MDTPENVLQMLEAHMQSYKGN DPLGEWERYIQWVEENFPENKEYLITLLEHLMKEFLDKK
KYHNDPRFISYCLKFAEYNSDLHQFFFLYNHGIGTLSSPLYIAWAGHLEAQGELQHASA
VLQRGIQNQAEPREFLQQQYRLFQTRLTETHLPAQARTSEPLHNVQVLNQMITSKSNPGN
NMACISKNQGSELSGVISSACDKESNMERRVITISKSEYSVHSSLASKVDVEQVVMYCKE
KLIRGESEFSFEELRAQKYNQRRKHEQWVNEDRHYMKRKEANAFEEQLLKQKMDLHKKL
HQVVETSHEDLPASQERSEVNPARGPSVGSQQELRAPCLPVTYQQTPVNMEKNPREAPP
VVPPLANAI SAALVSPATSQSIAPPVPLKAQTVTDSMF AVASKDAGCVNKSTHEFKPQSG
AETKEGCFETHKVA NTSSFHTTPNTSLGMV OATPSKVOPSP TVHTKFAIGFTMMMFQAPTI
```

Your input contains 1 sequence

Target database

UniProtKB Swiss-Prot

Restrict by taxonomy

Enter taxon names or IDs to include



Saccharomyces cerevisiae [4932] x

重置

Run BLAST

Best Hit!



BLAST Align Peptide search ID mapping SPARQL

Tool results ▾

Advanced | List Search

🏠 📁 📧 Help

Blast parameters

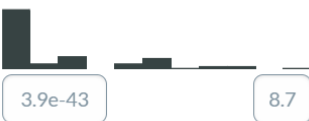
Identity



Score



E-Value



Status

📌 Reviewed (Swiss-Prot) (114)

BLAST 114 results found in UniProtKB

Overview Taxonomy Hit Distribution Text Output Input Parameters API Request

BLAST Align ▾ Map IDs ⬇ Download 📁 Add ✎ Customize columns 🔄 Resubmit

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	
☐ P41695	BUB1_YEAST	Checkpoint serine/threonine-protein kinase BUB1 [...]	BUB1, YGR188C, G7542	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	1,021 AA	31% 427 3.9e-43
☐ P47074	MAD3_YEAST	Spindle assembly checkpoint component MAD3 [...]	MAD3, YJL013C, J1341	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	515 AA	20.5% 151 1.9e-10

```
>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
Length=1021
```

```
Score = 169 bits (427), Expect = 4e-43
Identities = 115/371 (31%), Positives = 190/371 (51%), Gaps = 46/371 (12%)
```

```
Query 734 PNFIVGNPWDDKLIFKLLSGLSKPVSSYPNTF---EWQCKLPAI-----KPK 777
      P I+ NP + L K LS +S P+ Y NTF + K+ ++ P
Sbjct 637 PPVIIENPLSNNLRKFLSEISPLLFQY-NTFYNNQELKMSSLLKKIHRVSRNENKNPI 695
```