BLAST 算法

降帅

北京大学生命科学学院 生物信息中心

jiangs@mail.cbi.pku.edu.cn

Outline

- 全局比对中的动态规划
- 局部比对中的动态规划
- Blast 算法
- PSI-BLAST

Outline

- 全局比对中的动态规划
- 局部比对中的动态规划
- Blast 算法
- PSI-BLAST

序列比对

- 相似的序列 → 相似的结构 → 相似的功能

  推断这个未知新序列的可能的功能

- 相似的序列 → 同源

  在演化分析中用来构建演化树的重要依据

双序列比对

Broad bean leghemoglobin I	QLRATGEVVLDGK
Horse beta globin	LHSFGEGVHHLDN

```
QLRATGEVV LDGK
| | | | | | | | | | | | | |
 LHSFGEGVHHLDN
```

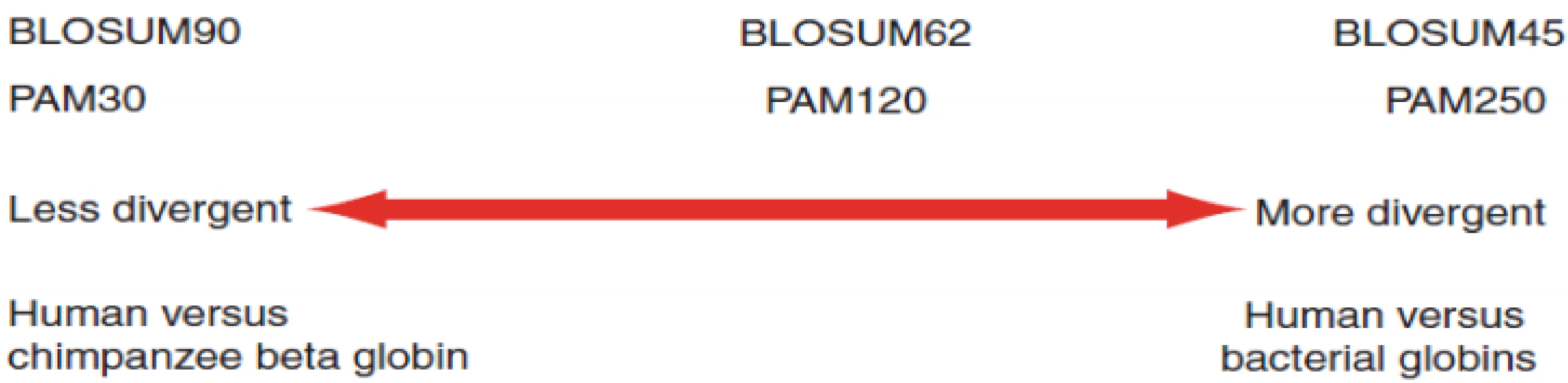
```
QLRATG EVV LDGK -
| | | | | | | | | | | | | | |
 - LHSFGEGV HHLDN
```

```
QLRATG EVV - - LDGK
| | | | | | | | | | | | | | | |
 - LHSFGEGV HHLDN -
```

打分矩阵

- PAM
  - 基于近相关蛋白数据构建的
- BLOSUM
  - 基于实际观测到的远相关蛋白构建的

因此在比对较远蛋白时，应选BLOSUM矩阵



(Jonathan Pevsner, *Bioinformatics and Functional Genomics*)

From 亢雨笺

BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

双序列比对

Broad bean leghemoglobin I	QLRATGEVVLDGK
Horse beta globin	LHSFGEGVHHLDN

```
QLRATGEVV LDGK
| | | | | | | | | | | | |
LHSFGEGVHHLDN
```

```
QLRATG EVV LDGK -
| | | | | | | | | | | | | | |
- LHSFGEGV HHLDN
```

```
QLRATG EVV - - LDGK
| | | | | | | | | | | | | | | |
- LHSFGEGV HHLDN -
```

双序列比对

Gap existence: -11

QLRATGEVV LDGK

| | | | | | | | | | | | | |

LHSFGEGVHHLDN

-2
-3
-1
-2
-2
-2
-2
4
-3
-3
-4
-1
0

-21

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

双序列比对

QLRATG EVV LDGK -

| | | | | | | | | | | | | | |

- LHSFGEGV HHLDN

-11
4
0
1
-2
6
5
-3
4
-3
-1
-4
-1
-11

-16

Gap existence: -11

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

双序列比对

QLRATG EVV - - LDGK

- LHSFGEGV HHLDN -

-11
4
0
1
-2
6
5
-3
4
-11
-1
4
6
0
-11

-9

Gap existence: -11

Gap extension: -1

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

双序列比对

Gap existence: -11　　Gap extension: -1

```
QLRATGEVV LDGK
| | | | | | | | | | | | | |
LHSFGEGVHHLDN
```

-21

```
QLRATG EVV LDGK -
| | | | | | | | | | | | | | |
- LHSFGEGV HHLDN
```

-16

```
QLRATG EVV - - LDGK
| | | | | | | | | | | | | | | | |
- LHSFGEGV HHLDN -
```

-9

双序列比对

QLRATGEVV LDGK

LHSFGEGVHHLDN

-2
-3
-1
-2
-2
-2
-2
4
-3
-3
-4
-1
0

-21

Gap existence: -15

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

双序列比对

Gap existence: -15

QLRATG EVV LDGK -

| | | | | | | | | | | | | |

- LHSFGEGV HHLDN

-15
4
0
1
-2
6
5
-3
4
-3
-1
-4
-1
-15

-24

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

双序列比对

QLRATG EVV - - LDGK

| | | | | | | | | | | | | | |

- LHSFGEGV HHLDN -

-15
4
0
1
-2
6
5
-3
4
-15
-1
4
6
0
-15

-21

Gap existence: -15

Gap extension: -1

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

双序列比对

Gap existence: -15 Gap extension: -1

QLRATGEVV LDGK

LHSFGEGVHHLDN

-21

QLRATG EVV LDGK -

- LHSFGEGV HHLDN

-24

QLRATG EVV - - LDGK

- LHSFGEGV HHLDN -

-21

最优双序列比对

S1和S2的最优比对

序列 S1
序列 S2

最大分数

枚举法

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

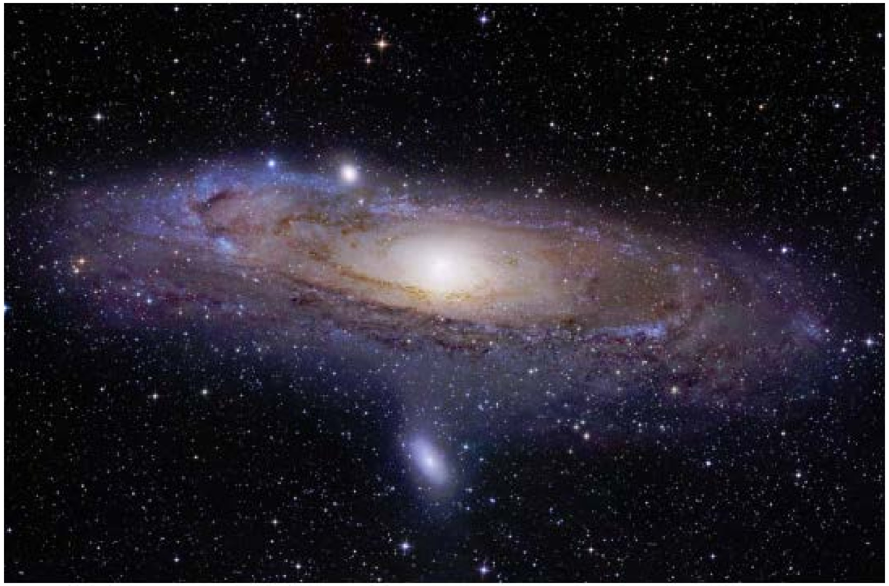
枚举法

LSPADK	L-SPADK	L-SPADK
LTPEEK	LTPEEK-	LT-PEEK

(From 高歌)

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} = \frac{(2*300)!}{(300!)^2} \approx 7\times 10^{88}$$

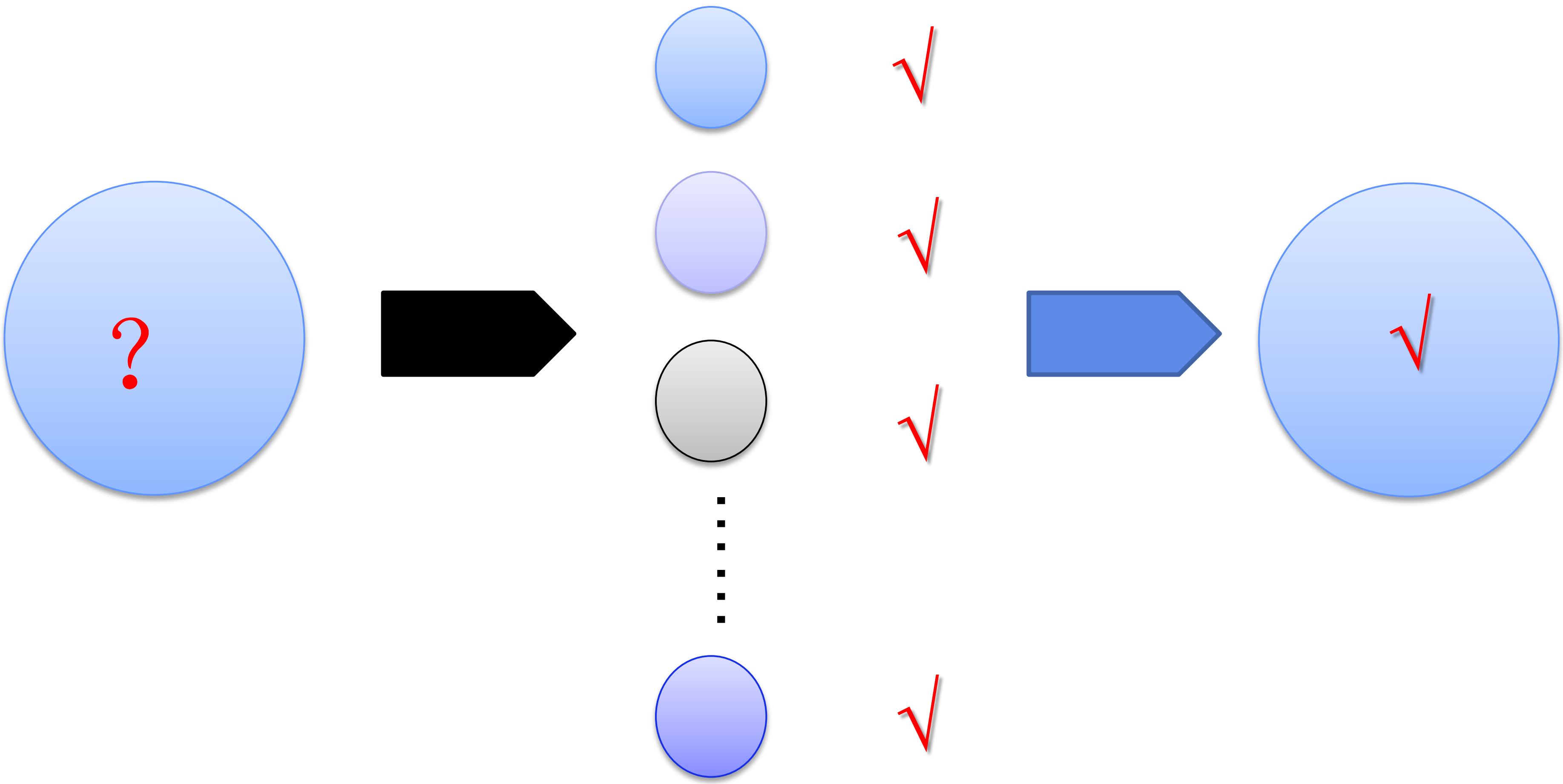
目前可见宇宙中所有原子的一亿倍！



$10^{78} \sim 10^{80}$ atoms

(Source: http://discoverystudentadventures.blogspot.com/2011/11/10-mind-bending-facts-about-universe.html)

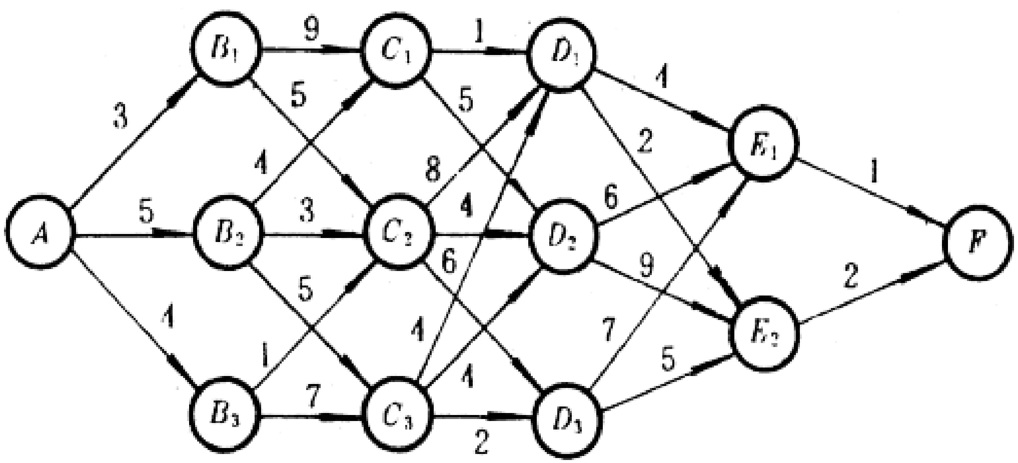
动态规划



动态规划

一个大问题可以分成若干个子问题

寻找每个子问题的最优解，就是最终的最优解



From 亢雨笺

全局比对中的动态规划

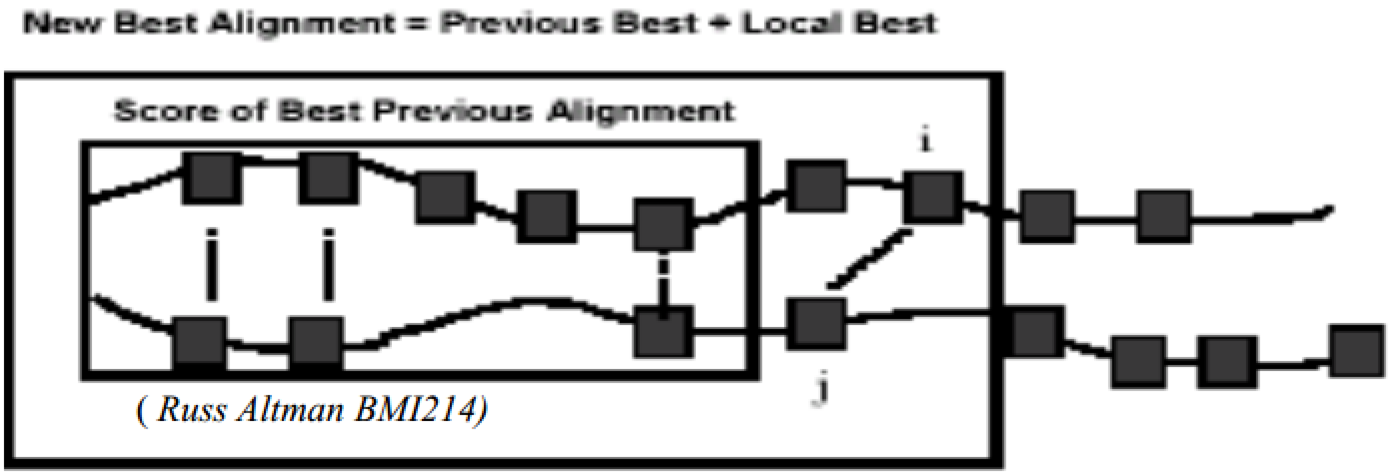
```
MV-LSP

MVHLTP
```

```
HBA_HUMAN          1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D   48
                     || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
HBB_HUMAN          1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD   48

HBA_HUMAN         49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR   93
                     ||     .|:.:||.|||||..|.::.:||:|::.....:.||:||..||.
HBB_HUMAN         49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH   98

HBA_HUMAN         94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR  142
                     |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
HBB_HUMAN         99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH  147
```

(From 高歌)

全局比对中的动态规划

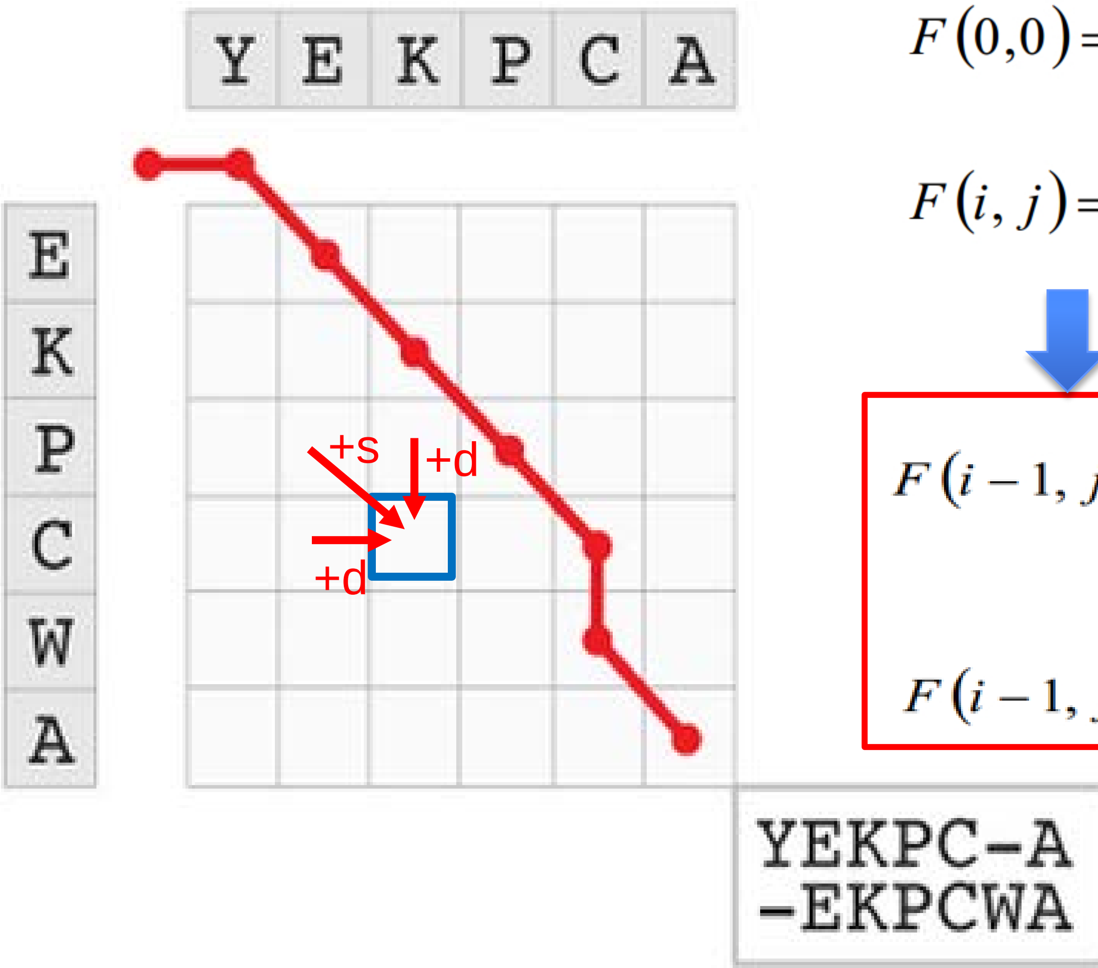
最好的比对 = 之前最好的比对+当前最好的比对



全局比对中的动态规划

http://blog.thegrandlocus.com/2014/06/once-upon-a-blast

全局比对中的动态规划



$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) + d \\ F(i,j-1) + d \end{cases}$$

$$\begin{tikzcd}
F(i-1,j-1) \arrow[dr] & & F(i,j-1) \arrow[d] \\
& s(x_i, y_j) \arrow[dr] & d \arrow[d] \\
F(i-1,j) \arrow[r] & d \arrow[r] & F(i,j)
\end{tikzcd}$$

全局比对中的动态规划

A A G –
– A G C

A A G –
A – G C

		A	A	G
	0	-5		
A		2	-3	
G				-1
C				-6

0+2=2   -5+(-5)=(-10)   -5+(-5)=(-10)

全局比对中的动态规划

- Needleman-Wunsch

Outline

- 全局比对中的动态规划
- 局部比对中的动态规划
- Blast 算法
- PSI-BLAST

局部比对

全局比对

$$
\begin{tikzcd}
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] \\
& s(x_i, y_j) \arrow[dr] & d \arrow[d] \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j)
\end{tikzcd}
$$

局部比对

$$
\begin{tikzcd}
& 0 \arrow[dddr] & \\
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] \\
& s(x_i, y_j) \arrow[dr] & d \arrow[d] \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j)
\end{tikzcd}
$$

局部比对

A G

A G

A

A

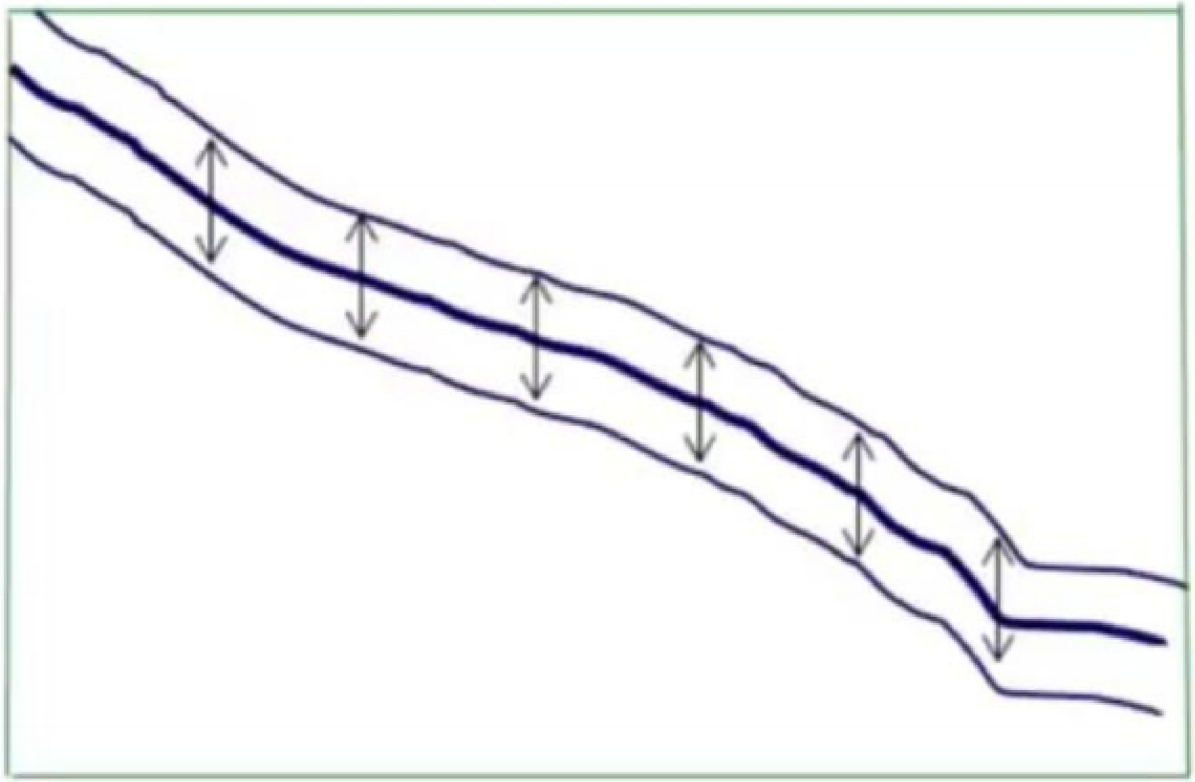
		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0

局部比对

- Smith-Waterman

局部比对

		A	A	G
	0	0	0	0
A	0	2	2	0
G	0	0	0	4
C	0	0	0	0



Outline

- 全局比对中的动态规划
- 局部比对中的动态规划
- Blast 算法
- PSI-BLAST

Basic Local Alignment Search Tool

- Smith-Waterman

  Best alignment

  Slow

- Blast

  Not the best, but good enough

  Fast

Seeding-and-extending

- 寻找 Seed序列
- 在数据库中定位seed序列
- 延伸匹配 → HSP

  只在特定的区域进行局部比对
- 评估HSP的统计显著性

Seeding

Query Sequence

Seed words of length k

可根据用户的
需要自行调整

Protein : k = 3
DNA : k = 11

种子越短： 灵敏度越高 计算速度越慢

Find “neighborhood words”

Query Sequence M V L S P A D K T N V K A A W

Total words : $20^3 = 8000$

DKT
| | |
8000 XXX

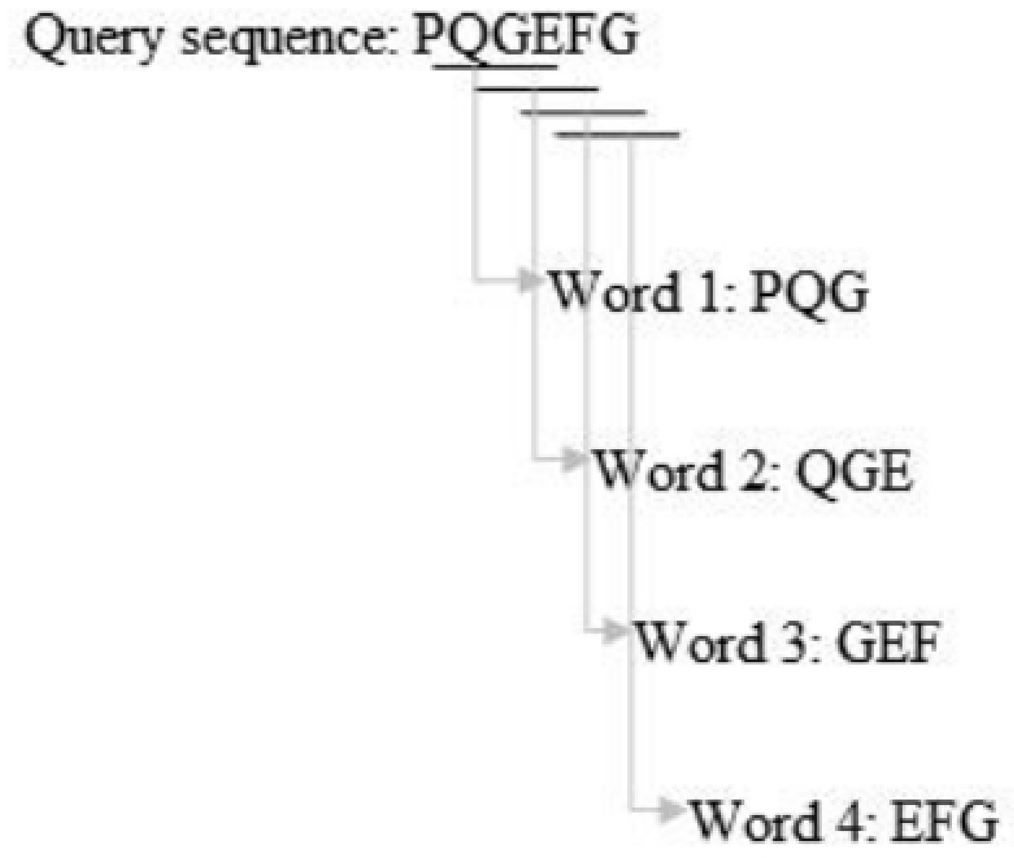
只有在源程序中才可以调

可根据用户的需要自行调整

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

DKT 16
DRT 13
DET 12
DKS 12
DQT 12
EKT 12
DKA 11
DKN 11
DKV 11
DNT 11
DST 11
NKT 11
DAT 10
DDT 10
DHT 10
DKC 10
DKD 10
DKE 10
DKI 10
DKK 10
DKL 10
DKM 10
DKP 10
DKQ 10
DKR 10
DMT 10
DPT 10
DTT 10
QKT 10
SKT 10

Seeding


Query sequence: PQGEFG
Word 1: PQG
Word 2: QGE
Word 3: GEF
Word 4: EFG

Find “neighborhood words”

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	**9**																				C
S	-1	**4**																			S
T	-1	1	**5**																		T
P	-3	-1	-1	**7**																	P
A	0	1	0	-1	**4**																A
G	-3	0	-2	-2	0	**6**															G
N	-3	1	0	-2	-2	0	**6**														N
D	-3	0	-1	-1	-2	-1	1	**6**													D
E	-4	0	-1	-1	-1	-2	0	2	**5**												E
Q	-3	0	-1	-1	-1	-2	0	0	2	**5**											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	**8**										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	**5**								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

查询序列LNKCKTPQGQR

PQG 7+5+6=18 字串

PEG 7+2+6=15

PRG 7+1+6=14

PKG 7+1+6=14 邻居字串

PNG 7+0+6=13

临界值T=13 PMG 7+0+6=13

PQA 7+5+0=12

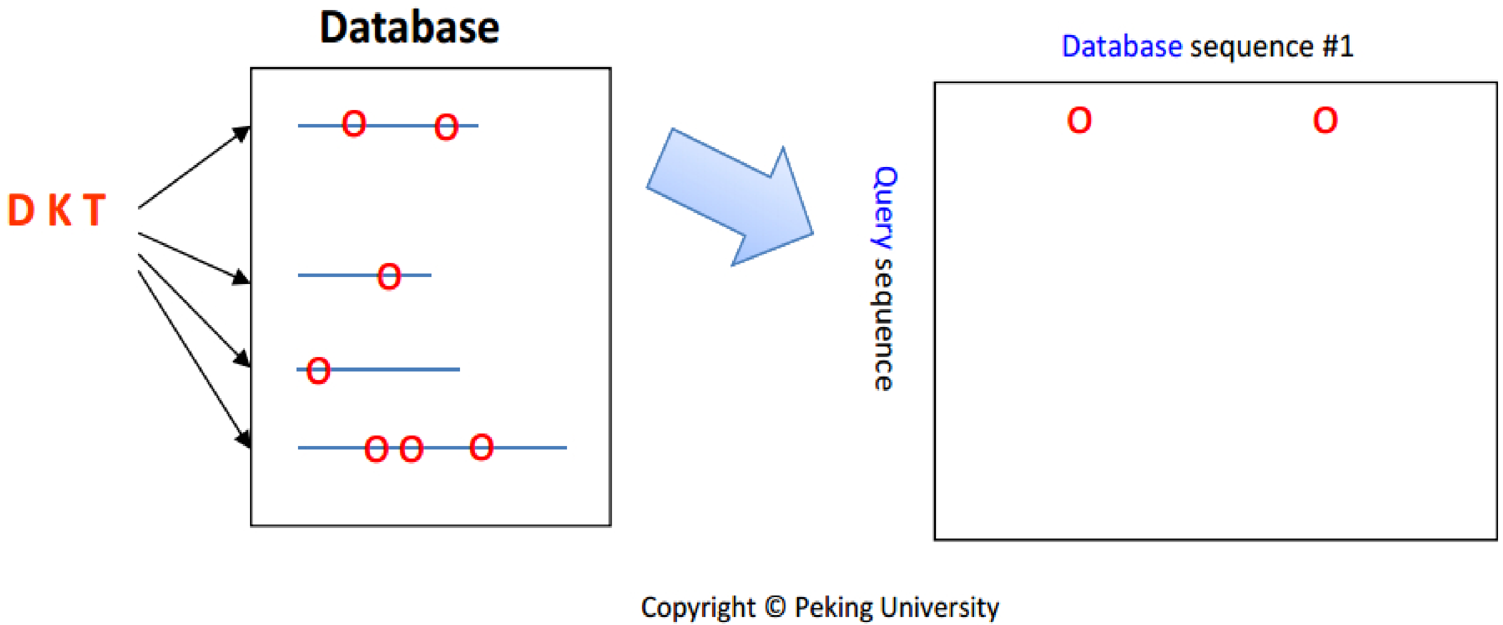
PQN 7+5+0=12

etc.

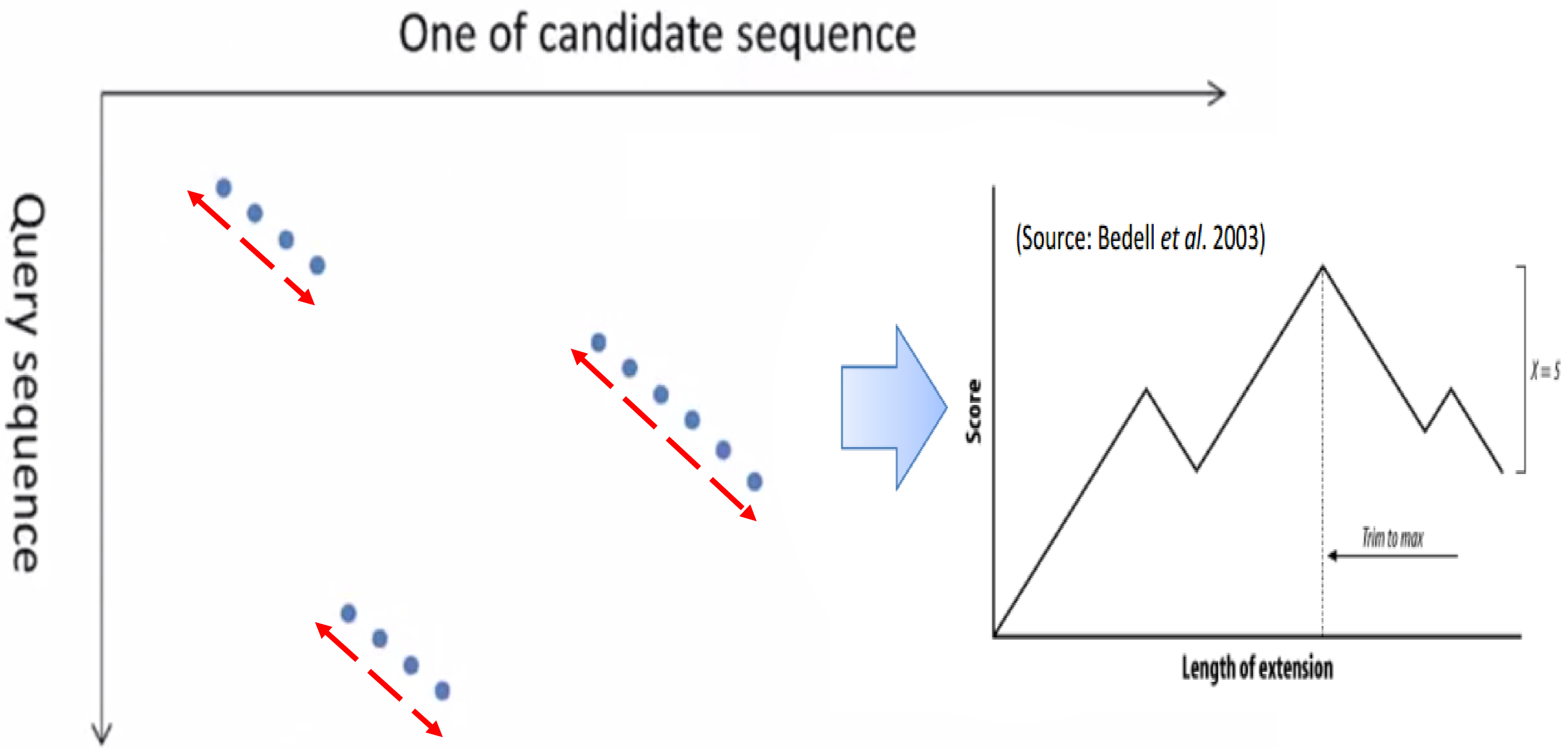
BLAST2：T=11

Index database

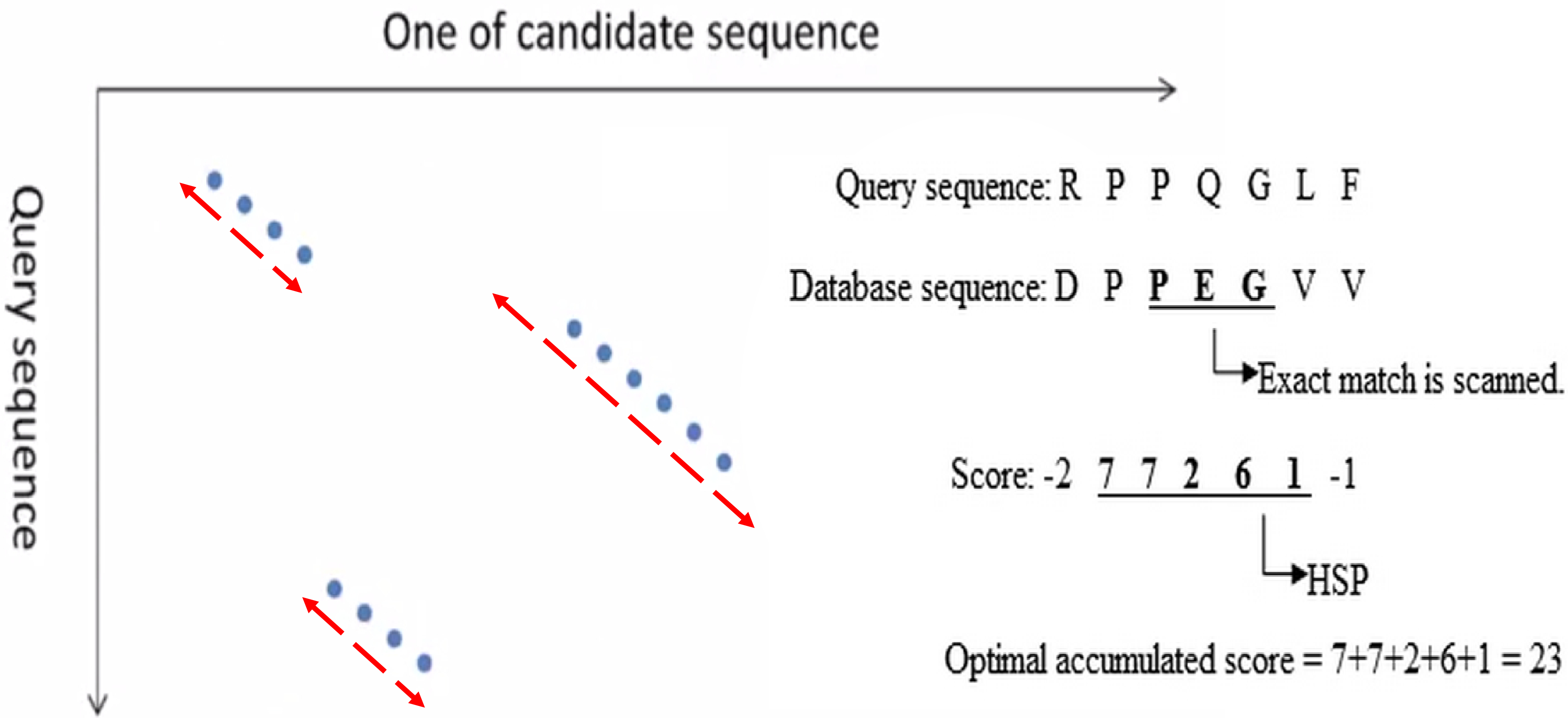
The database was pre-indexed to quickly locate all positions in the database for a given seed.



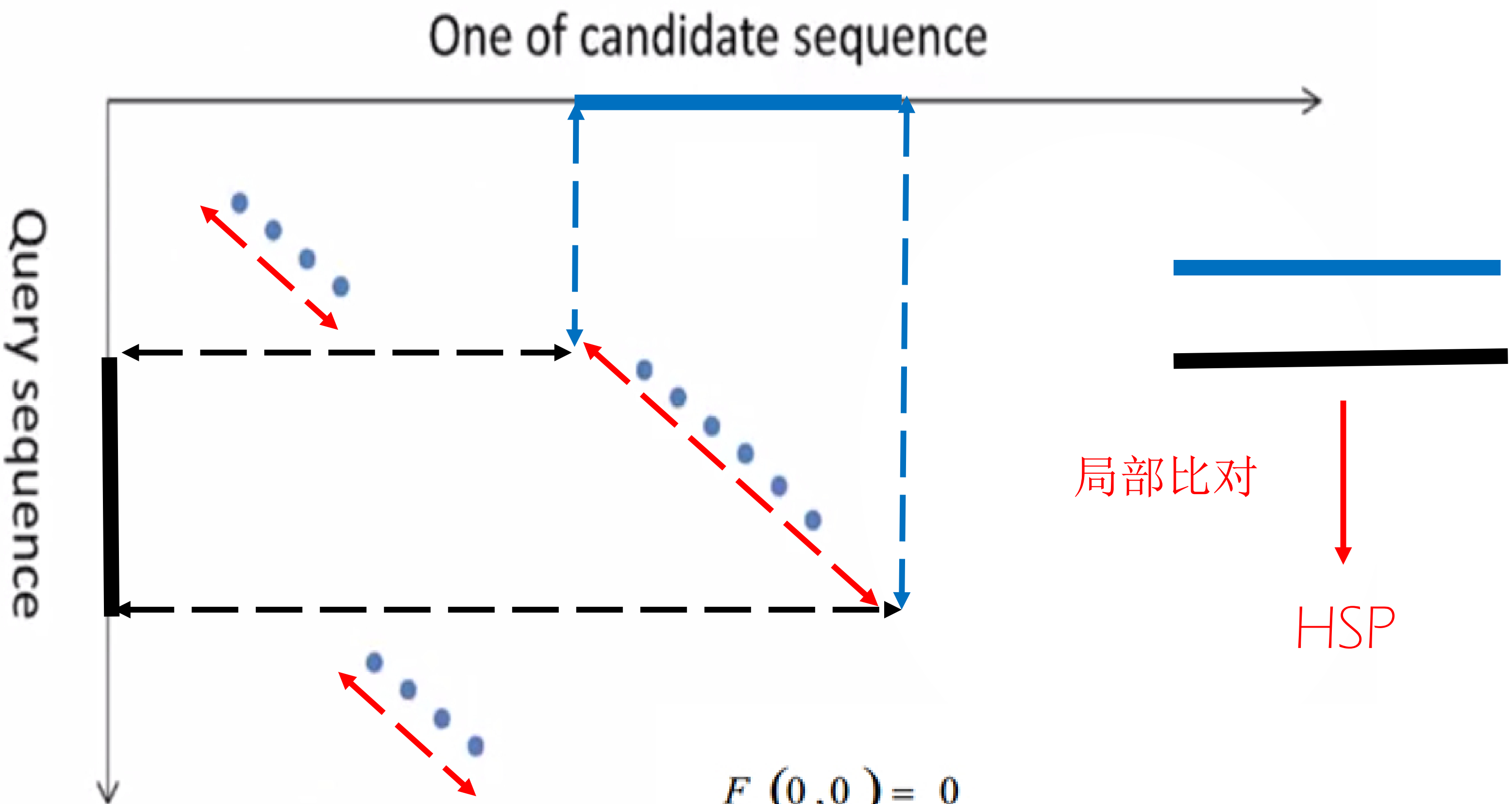
(From Gaog)


One of candidate sequence
Query sequence
(Source: Bedell et al. 2003)
Score
X = 5
Trim to max
Length of extension

(From 高歌)

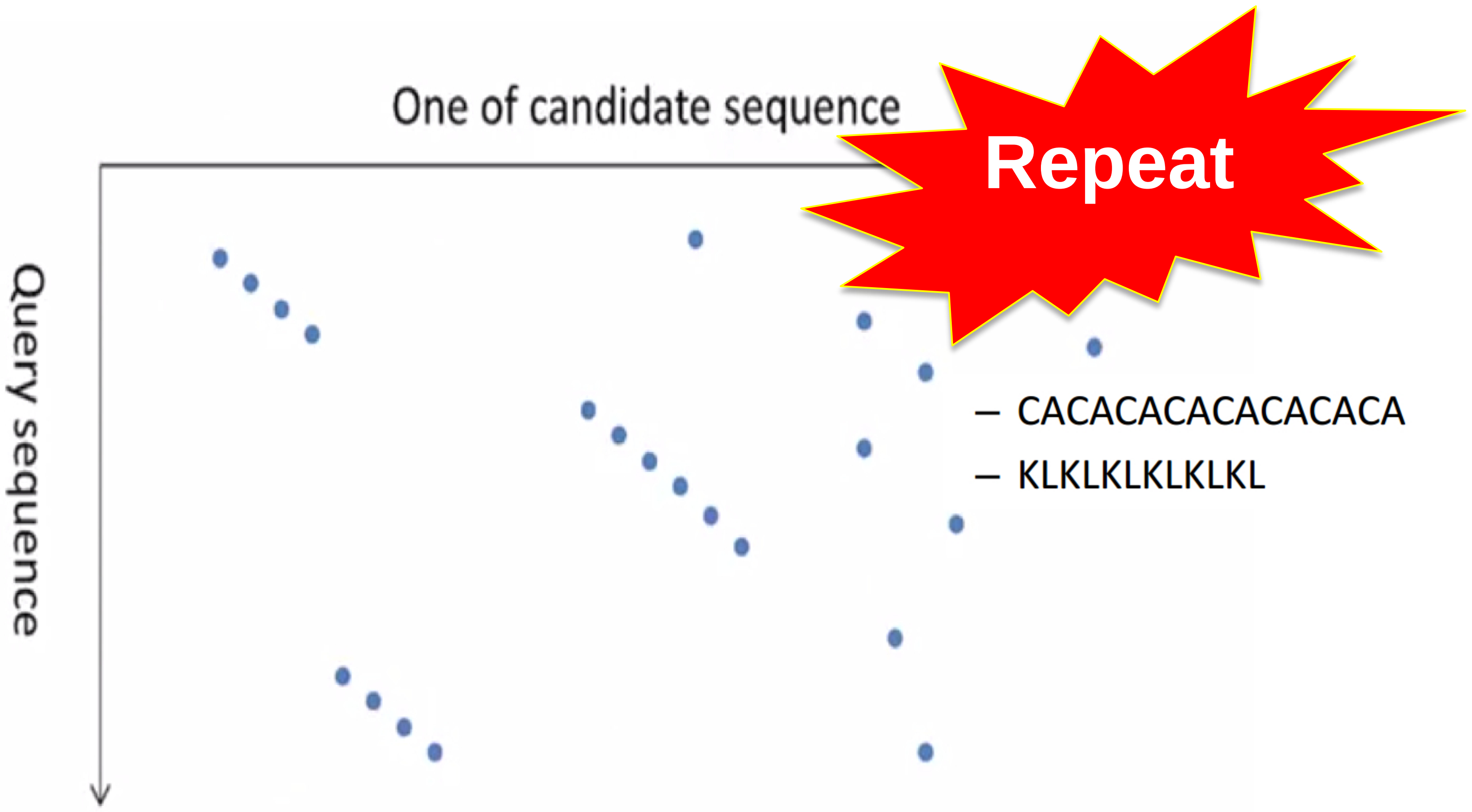

One of candidate sequence
Query sequence
Query sequence: R P P Q G L F
Database sequence: D P P E G V V
Exact match is scanned.
Score: -2 7 7 2 6 1 -1
HSP
Optimal accumulated score = 7+7+2+6+1 = 23

(From 高歌)



$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$



One of candidate sequence
Query sequence
Repeat
– CACACACACACACACA
– KLKLKLKLKLKL

Seeding

Mask low-complexity region

Query Sequence

Seed words of length k

Protein : k = 3
DNA : k = 11

Blast output

Query = N1
>subject S1 Hit
Sequence HSP1
Sequence HSP3
Sequence HSP2
>subject ...

HSP的统计显著性检验

- An amino acid sequence : length L

  Random match: $(1/20)^L$

- Search in Swiss-prot

  $(1/20)^6 * 192,206,270 = 3.00$

E-value

- Evalue
  - 随机情况下，获得当前或者比当前更高比对分数的可能比对条数
- E=10
  - 就意味着会有10个随机的匹配获得与当前比对相等或者更高的分数。

E-value

打分矩阵的标准化系数

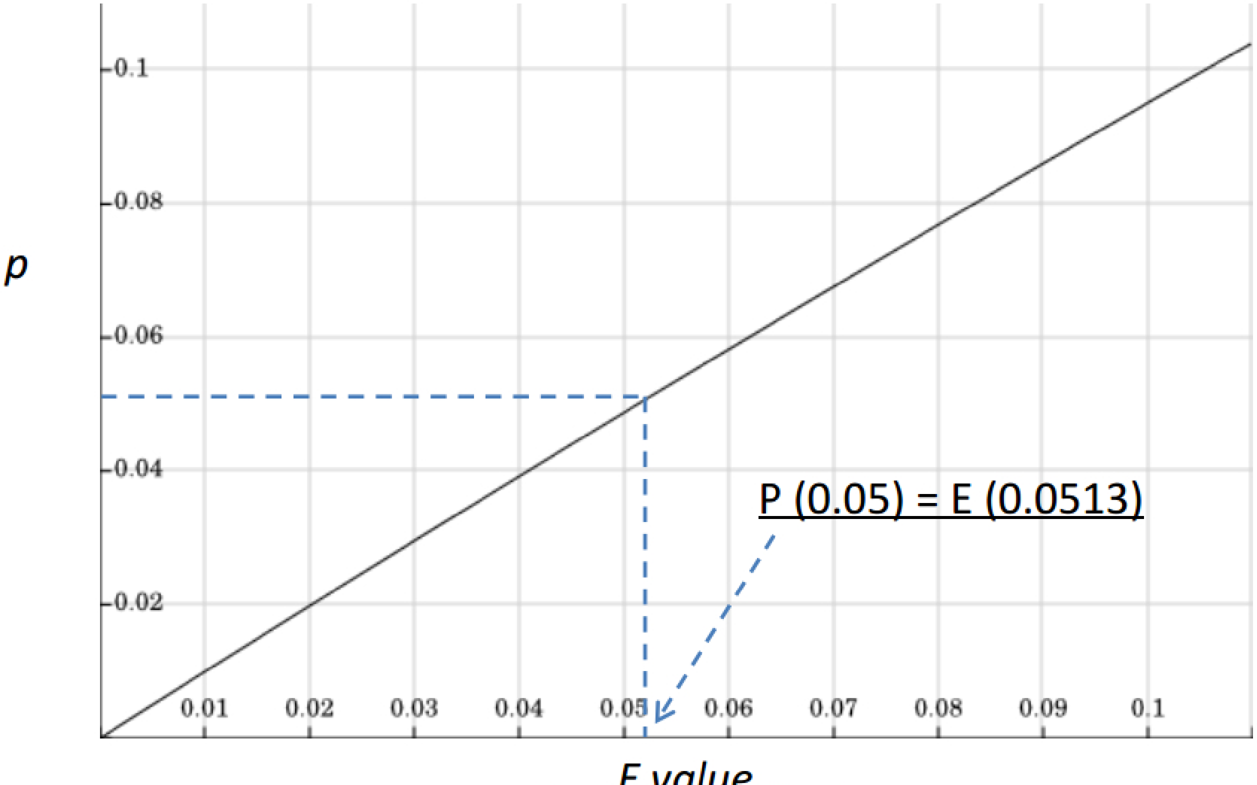
$$E = kmne^{-\lambda S}$$

HSP的打分值

查询序列的长度　　数据库的大小

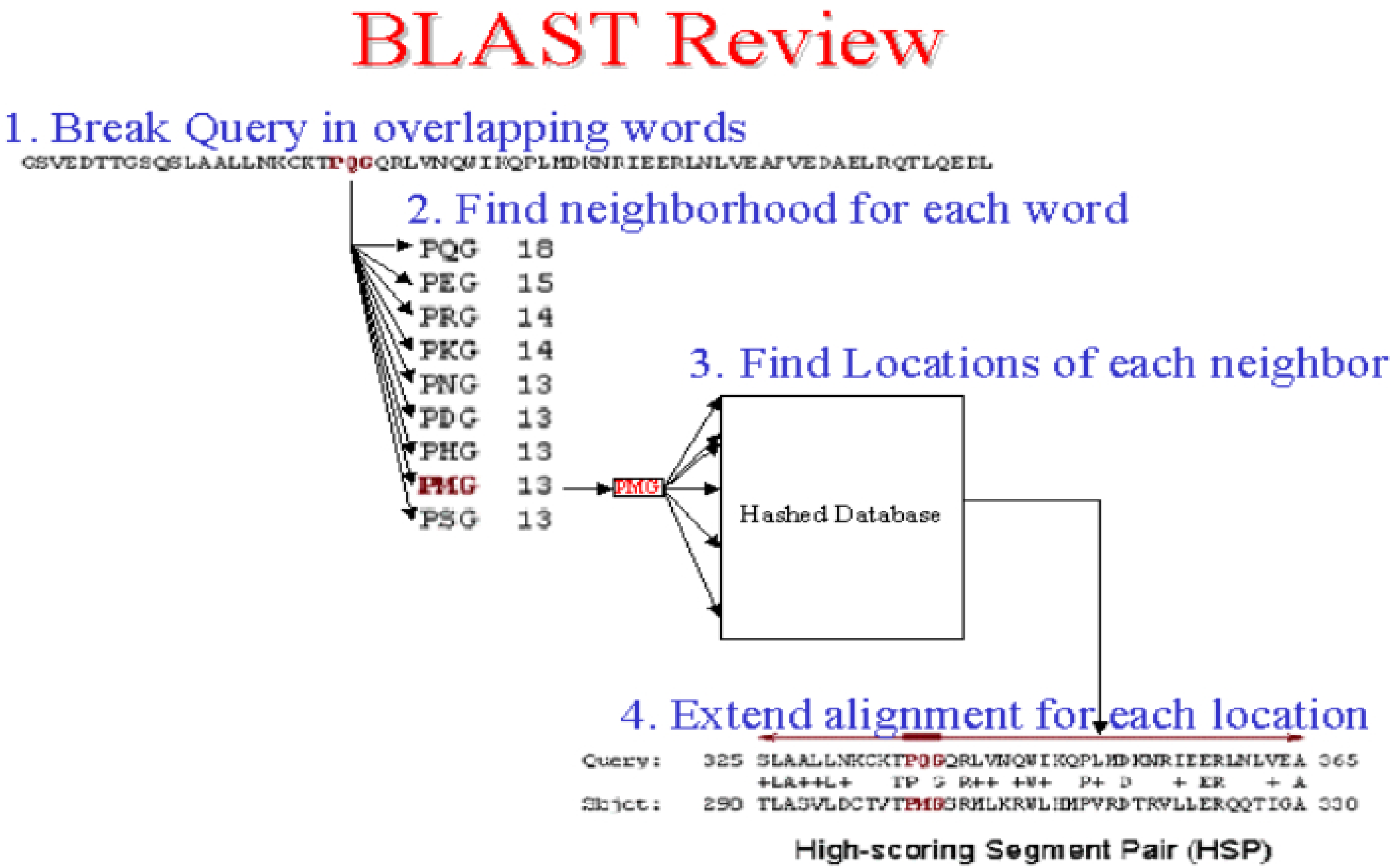
$$E = kmne^{-\lambda S}$$

$$p = 1 - e^{-E}$$



(From Gaog)

Summary



From 亢雨笺

Summary

- Seeding and extending

  选择性的进行局部比对

- Speed vs. sensitivity

Outline

- 全局比对中的动态规划
- 局部比对中的动态规划
- Blast 算法
- PSI-BLAST

PSI-BLAST

- **Position-Specific Iterated BLAST**

- **PSSM (**Position-Specific Scoring Matrix),

  Amino acid substitution scores are given separately for each position in a protein multiple sequence alignment.

-

  http://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi

Sequences:

MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFPHFDL
DKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKT
AVTALWGKVNVDEVGGEALGRLLVVYPWTQ
EKTAVNALWGKVNVDAVGGEALGRLLVVYP
EDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRF
TSLWGKVNVEDAGGETLGRLLVVYPWTQR
ADEVIESGLVQDFDASLSGIGQELGAGAYSMSDVLALP
LDFSEHFNQLELLETHGHLIPTGTQSLWVGNSD
HEAQFGAELLLRLFTVYPSTKVYFPHLSACQDATQLLSH
LALLQVFQLAAHLVYWGKAIIIYPLCENNVYMLSPNASVCLYSPLAEQFSHQFPSHDLPSV
EEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSS
LWKKLGSNVGVYTTEALERTFLAFPATKTYFSHLDLS

Alignment:

PVNFKLLSHCLLVTLAAHLPAEFTP
PASFQLLGHCLLVTLARHYPGDFSP
PVNFKLLSHCLLVTLAARFPADFTA
PANFPLLIQCFHVVLASHLQDEFTV
PENFRLLGNVLVCVLAHHFGKEFTP
PENFRLLGNVLVCVLARNFGKEFTP
PENFKLLGNVLVTVLAIHFGKEFTP
PENFKLLGNVLVTVLAIHFGKEFTP
PENFKLLGNVMVIILATHFGKEFTP
PEDLRMFARLLHYFRGRHHLEEIMY
KDTLELLLMNRYVKPGLKNNLEETA
GTFFVYHAIYLEELTAVELTEKIAQ

Frequency matrix:

Amino Acid	1	2	3	4	5	6	7
N	0.17	0.04	0.04	0.04	0.04	0.04	0.04
T	0.04	0.09	0.09	0.04	0.04	0.04	0.04
E	0.04	0.04	0.09	0.04	0.13	0.09	0.04
G	0.04	0.04	0.04	0.13	0.09	0.04	0.09
W	0.04	0.04	0.04	0.04	0.04	0.09	0.09
I	0.04	0.13	0.04	0.04	0.04	0.04	0.04
H	0.04	0.04	0.04	0.04	0.04	0.04	0.04
R	0.04	0.04	0.04	0.09	0.04	0.04	0.04
A	0.04	0.04	0.09	0.04	0.04	0.04	0.04
C	0.04	0.04	0.04	0.04	0.04	0.09	0.09
...	...	...	...	...	...	...	...

Score matrix:

Amino Acid	1	2	3	4	5	6	7
N	0.54	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
T	-0.06	0.24	0.24	-0.06	-0.06	-0.06	-0.06
E	-0.06	-0.06	0.24	-0.06	0.42	0.24	-0.06
G	-0.06	-0.06	-0.06	0.42	0.24	-0.06	0.24
W	-0.06	-0.06	-0.06	-0.06	-0.06	0.24	0.24
I	-0.06	0.42	-0.06	-0.06	-0.06	-0.06	-0.06
H	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
R	-0.06	-0.06	-0.06	0.24	-0.06	-0.06	-0.06
A	-0.06	-0.06	0.24	-0.06	-0.06	-0.06	-0.06
C	-0.06	-0.06	-0.06	-0.06	-0.06	0.24	0.24
...	...	...	...	...	...	...	...

Frequency Matrix

```
NTEGEWI
NITRGEW
NIAGECC
```

The observed frequency of 'I' occurring at position 2 is 2/3, or .67

$$Freq = \frac{N_{obs}}{\sum N_{obs}}$$

Amino Acid	1	2	3	4	5	6	7
N	1.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.00	0.33	0.33	0.00	0.00	0.00	0.00
E	0.00	0.00	0.33	0.00	0.67	0.33	0.00
G	0.00	0.00	0.00	0.67	0.33	0.00	0.33
W	0.00	0.00	0.00	0.00	0.00	0.33	0.33
I	0.00	0.67	0.00	0.00	0.00	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.00	0.00	0.00	0.33	0.00	0.00	0.00
A	0.00	0.00	0.33	0.00	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.33	0.33
…	…	…	…	…	…	…	…

(From 高歌)

- Limit of sequence number → 0 appearance in specific site

```
PVNFKLLSHCLLVTLAAHLPAEFTP
PASFQLLGHCLLVTLARHYPGDFSP
PVNFKLLSHCLLVTLAARFPADFTA
PANFPLLIQCFHVVLASHLQDEFTV
PENFRLLGNVLVCVLAHHFGKEFTP
PENFRLLGNVLVCVLARNFGKEFTP
PENFKLLGNVLVTVLAIHFGKEFTP
PENFKLLGNVLVTVLAIHFGKEFTP
PENFKLLGNVMVIILATHFGKEFTP
PEDLRMFARLLHYFRGRHHLEEIMY
KDTLELLLMNRYVKPGLKNNLEETA
GTFFVYHAIYLEELTAVELTEKIAQ
```

$$\hat{Freq} = \frac{(N_{obs} + \boxed{1})}{\sum N_{obs} + \boxed{20}} = \frac{Freq \times \sum N_{obs} + 1}{\sum N_{obs} + 20}$$

Amino Acid	1	2	3	4	5	6	7
N	1.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.00	0.33	0.33	0.00	0.00	0.00	0.00
E	0.00	0.00	0.33	0.00	0.67	0.33	0.00
G	0.00	0.00	0.00	0.67	0.33	0.00	0.33
W	0.00	0.00	0.00	0.00	0.00	0.33	0.33
I	0.00	0.67	0.00	0.00	0.00	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.00	0.00	0.00	0.33	0.00	0.00	0.00
A	0.00	0.00	0.33	0.00	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.33	0.33
...	...	...	...	...	...	...	...

NTEGEWI

NITRGEW

NIAGECC

$$\frac{3\times.067+1}{3+20}\approx 0.13$$

Amino Acid	1	2	3	4	5	6	7
N	0.17	0.04	0.04	0.04	0.04	0.04	0.04
T	0.04	0.09	0.09	0.04	0.04	0.04	0.04
E	0.04	0.04	0.09	0.04	0.13	0.09	0.04
G	0.04	0.04	0.04	0.13	0.09	0.04	0.09
W	0.04	0.04	0.04	0.04	0.04	0.09	0.09
I	0.04	0.13	0.04	0.04	0.04	0.04	0.04
H	0.04	0.04	0.04	0.04	0.04	0.04	0.04
R	0.04	0.04	0.04	0.09	0.04	0.04	0.04
A	0.04	0.04	0.09	0.04	0.04	0.04	0.04
C	0.04	0.04	0.04	0.04	0.04	0.09	0.09
...	...	...	...	...	...	...	...

(From 高歌)

Scoring Matrix: PSSM

$$odds\ ratio = \frac{P(x \mid M)}{P(x \mid R)}$$

$$Score = \log odds\ ratio = \log\left(\frac{P(x \mid M)}{P(x \mid R)}\right) = \log(P(x \mid M) - \log(P(x \mid R))$$

PSSM

Amino Acid	1	2	3	4	5	6	7
N	0.17	0.04	0.04	0.04	0.04	0.04	0.04
T	0.04	0.09	0.09	0.04	0.04	0.04	0.04
E	0.04	0.04	0.09	0.04	0.13	0.09	0.04
G	0.04	0.04	0.04	0.13	0.09	0.04	0.09
W	0.04	0.04	0.04	0.04	0.04	0.09	0.09
I	0.04	0.13	0.04	0.04	0.04	0.04	0.04
H	0.04	0.04	0.04	0.04	0.04	0.04	0.04
R	0.04	0.04	0.04	0.09	0.04	0.04	0.04
A	0.04	0.04	0.09	0.04	0.04	0.04	0.04
C	0.04	0.04	0.04	0.04	0.04	0.09	0.09
...	...	...	...	...	...	...	...

NTEGEWI

NITRGEW

NIAGECC

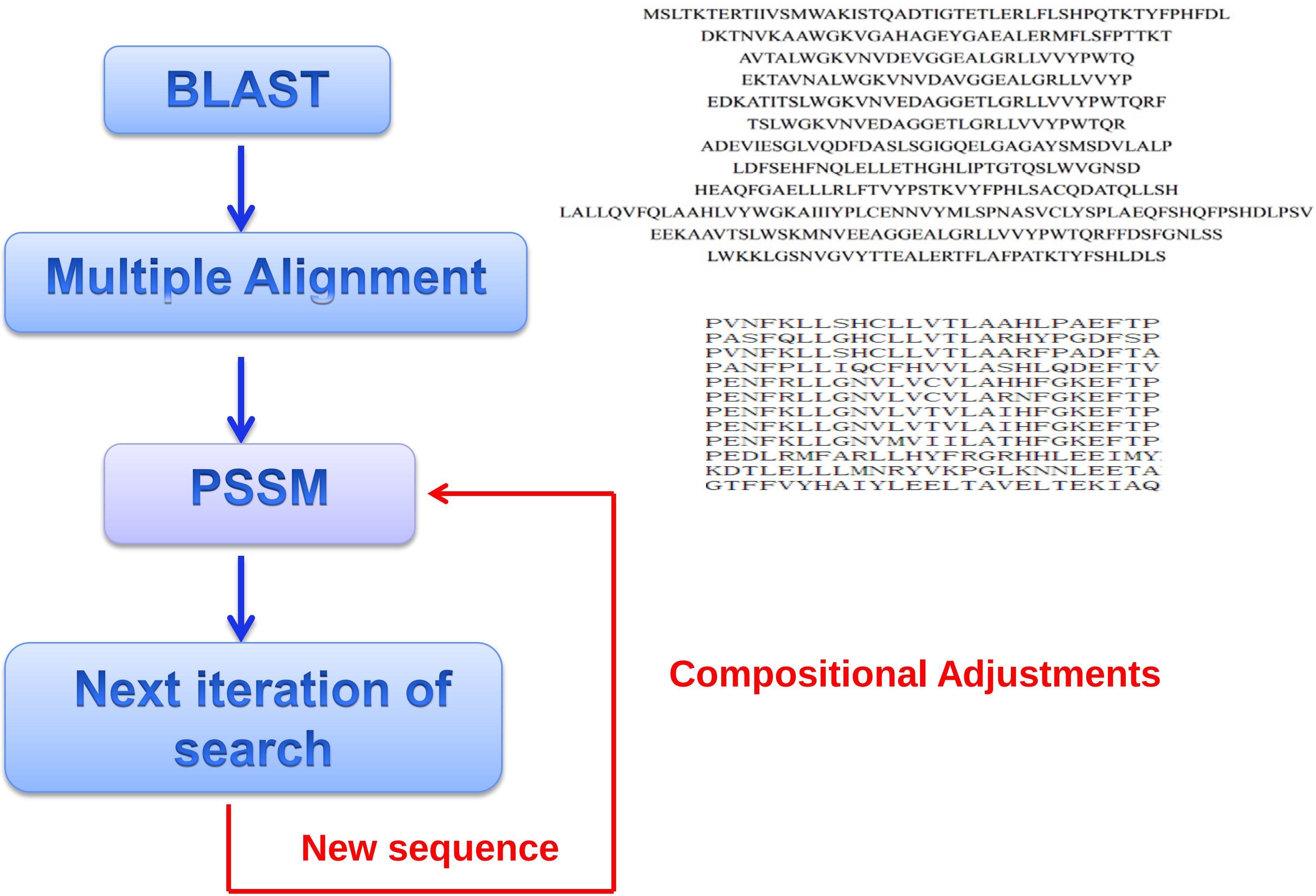
$$\log\left(\frac{0.13}{0.05}\right) \approx 0.42$$

$$Score = \log odds\ ratio = \log\left(\frac{P(x \mid M)}{P(x \mid R)}\right)$$

PSSM

Amino Acid	1	2	3	4	5	6	7
N	0.54	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
T	-0.06	0.24	0.24	-0.06	-0.06	-0.06	-0.06
E	-0.06	-0.06	0.24	-0.06	0.42	0.24	-0.06
G	-0.06	-0.06	-0.06	0.42	0.24	-0.06	0.24
W	-0.06	-0.06	-0.06	-0.06	-0.06	0.24	0.24
I	-0.06	0.42	-0.06	-0.06	-0.06	-0.06	-0.06
H	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06	-0.06
R	-0.06	-0.06	-0.06	0.24	-0.06	-0.06	-0.06
A	-0.06	-0.06	0.24	-0.06	-0.06	-0.06	-0.06
C	-0.06	-0.06	-0.06	-0.06	-0.06	0.24	0.24
...	...	...	...	...	...	...	...

PSI-BLAST


BLAST
Multiple Alignment
PSSM
Next iteration of search
New sequence
MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFPHFDL
DKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKT
AVTALWGKVNVDEVGGEALGRLLVVYPWTQ
EKTAVNALWGKVNVDAVGGEALGRLLVVYP
EDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRF
TSLWGKVNVEDAGGETLGRLLVVYPWTQR
ADEVIESGLVQDFDASLSGIGQELGAGAYSMSDVLALP
LDFSEHFNQLELLETHGHLIPTGTQSLWVGNSD
HEAQFGAELLLRLFTVYPSTKVYFPHLSACQDATQLLSH
LALLQVFQLAAHLVYWGKAIIIYPLCENNVYMLSPNASVCLYSPLAEQFSHQFPSHDLPSV
EEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSS
LWKKLGSNVGVYTTEALERTFLAFPATKTYFSHLDLS
PVNFKLLSHCLLVTLAAHLPAEFTP
PASFQLLGHCLLVTLARHYPGDFSP
PVNFKLLSHCLLVTLAARFPADFTA
PANFPLLIQCFHVVLASHLQDEFTV
PENFRLLGNVLVCVLAHHFGKEFTP
PENFRLLGNVLVCVLARNFGKEFTP
PENFKLLGNVLVTVLAIHFGKEFTP
PENFKLLGNVLVTVLAIHFGKEFTP
PENFKLLGNVMVIILATHFGKEFTP
PEDLRMFARLLHYFRGRHHLEEIMY
KDTLELLLMNRYVKPGLKNNLEETA
GTFFVYHAIYLEELTAVELTEKIAQ
Compositional Adjustments

Reference

- 高歌老师MIB课的课件

  (http://mib.cbi.pku.edu.cn/)

- ABC网站上(http://abc.cbi.pku.edu.cn/)刘欢,谢忱,亢雨笺关于BLAST的介绍。

- 侯玫关于BLAST的介绍

Thanks !