BLAST算法简介

2008.3.31

谢忱

xiec@mail.cbi.pku.edu.cn

BLAST简介

BLAST算法

BLAST应用

BLAST简介

BLAST算法

BLAST应用

BLAST

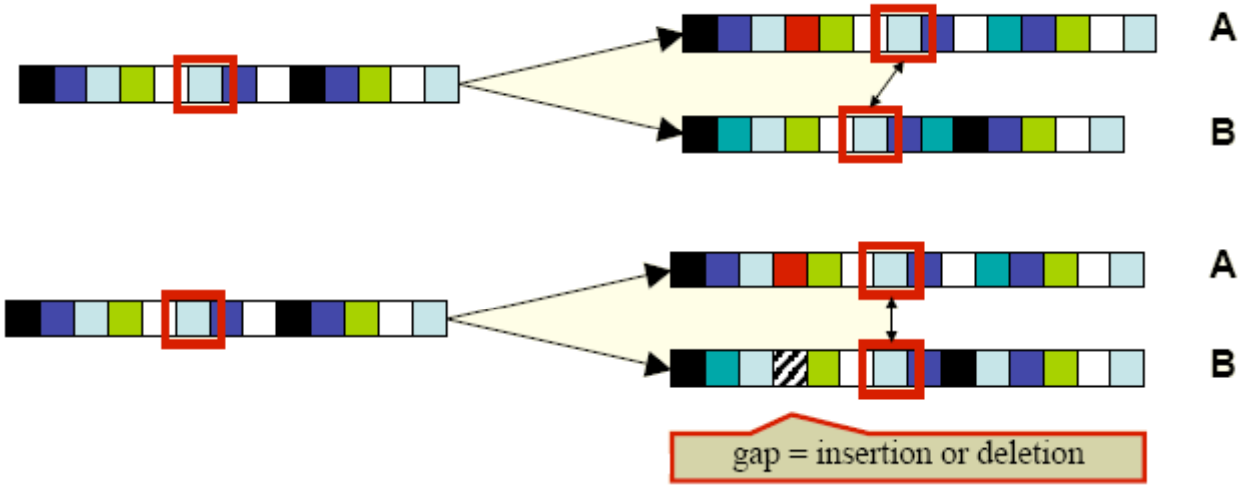
- Basic Local Alignment Search Tool
- 是一种基于成对局部序列比对的数据库相似性搜索工具
- Altschul *et al.* 1990

关于序列比对（Sequence Alignment）

- 序列比对的目的是找出序列间的相近程度。
- 用于推测序列的共同区域；推测该核酸或蛋白功能；以及推测一组序列是否起源于同一祖先。

序列比对的生物学意义

- 序列比对的目的是将一组序列所有位置上的来源于祖先序列上相同位置的碱基或氨基酸残基连配起来。



（From GaoG）

成对序列比对vs多序列比对

- 成对序列比对（Pairwise Sequence Alignment）用来基于序列相似性确定之前未知的生物学关系。
- 多序列比对（Mutiple Sequence Alignment）用于基于已知的一组序列间的生物学关系确定未知的保守区域。

打分函数（Scoring Function）

- 用来测量候选比对的质量
- 包括打分矩阵和空位罚分
  - 打分矩阵（Scoring Matrix）
    - PAM
    - BLOSUM
  - 空位罚分（Gap Penalty)
    - 线性（Linear) & 仿射（Affine）

氨基酸打分矩阵（**Scoring Matrix**）

- PAM：Percent Accepted Mutation

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

Dayhoff 1978

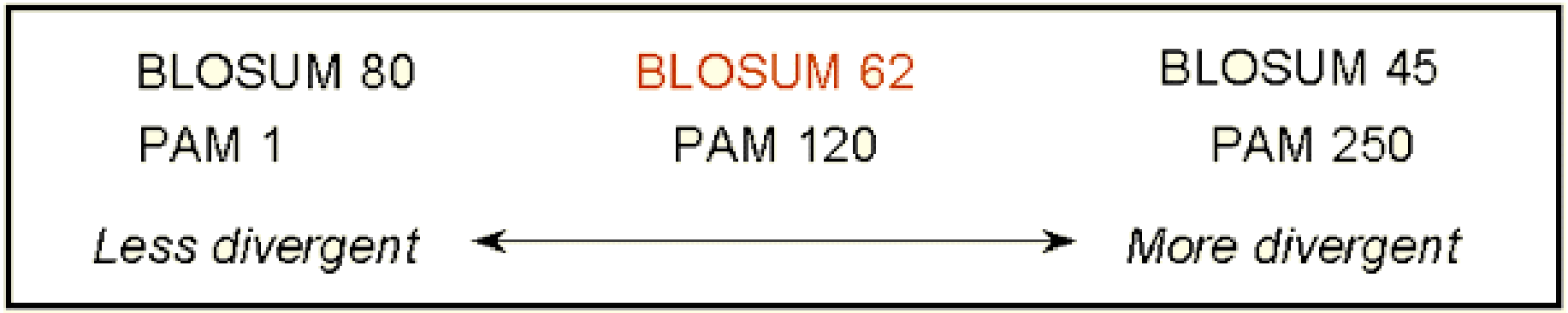
氨基酸打分矩阵（**Scoring Matrix**）

- BLOSUM：BLOcks SUbstitution Matrix

	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	

Henikoff *et al.* 1992

氨基酸打分矩阵（**Scoring Matrix**）



默认矩阵: PAM250, BLOSUM62

- PAM打分矩阵逐渐被BLOSUM取代，其原因包括PAM的数据集较小；PAM是由外推法得到，前提是分子钟恒定。

全局序列比对vs局部序列比对

- 全局比对（Global Alignment）
  - 在整个序列比对，适于长度和相似性较高的序列
  - Needleman-Wunsch 1970
- 局部比对（Local Alignment）
  - 在一段区域比对，适于差异较多的序列
  - Smith-Waterman 1981

序列比对算法

- 列举法（Enumerate）

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

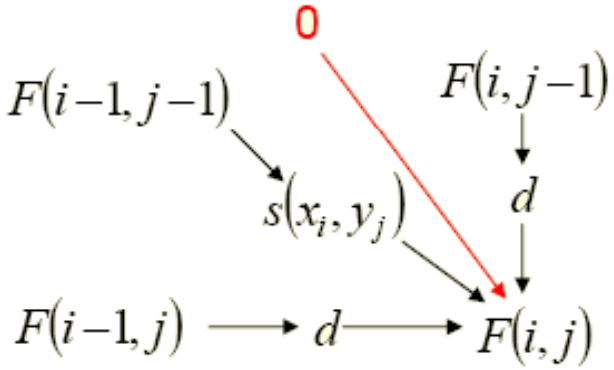
- 动态规划法（Dynamic Programming）
  - 问题 → 子问题 → 子问题的最优解 → 原问题最优解

DP for Local alignment: Example

	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Find the optimal local alignment of AAG and AGC.
Use a linear gap penalty of d = -5.

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



(From GaoG)

数据库相似性搜索

（Database Similarity Searching）

- 是大规模的成对序列比对，查询序列与数据库中的每一条序列进行比对
- 用于确定查询序列与数据库序列之间的相似度
  - 查询序列的结构和功能是未知的
  - 数据库序列的结构和功能是已知的

不同比对算法间出现的矛盾

算法	准确度 （敏感度，特异度）	速度
详尽的（exhaustive）： 动态规划	完美	非常慢
启发式的（heuristic）： FASTA BLAST	稍差	较快

The Great Men Behind BLAST

- 1970 - Needleman &Wunsch, global alignment
- 1978 - Dayhoff et al., PAM scoring matrix
- 1981 - Smith & Waterman, local algorithm
- 1985 - Lipman & Pearson, database search
- 1987 - Lipman & Pearson, FASTA
- 1990 - Altschul & Lipman, BLAST
- 1992 - Steven Henikoff & Jorja G Henikoff, BLOSUM scoring matrix

（From LuoJC）

BLAST简介

BLAST算法

BLAST应用

BLAST算法核心思想

Seeding-Extending

（种子-延伸）

第1步

- 低复杂序列产生假阳性
  - AAAAAAAA
  - KLKLKLKLKLKL

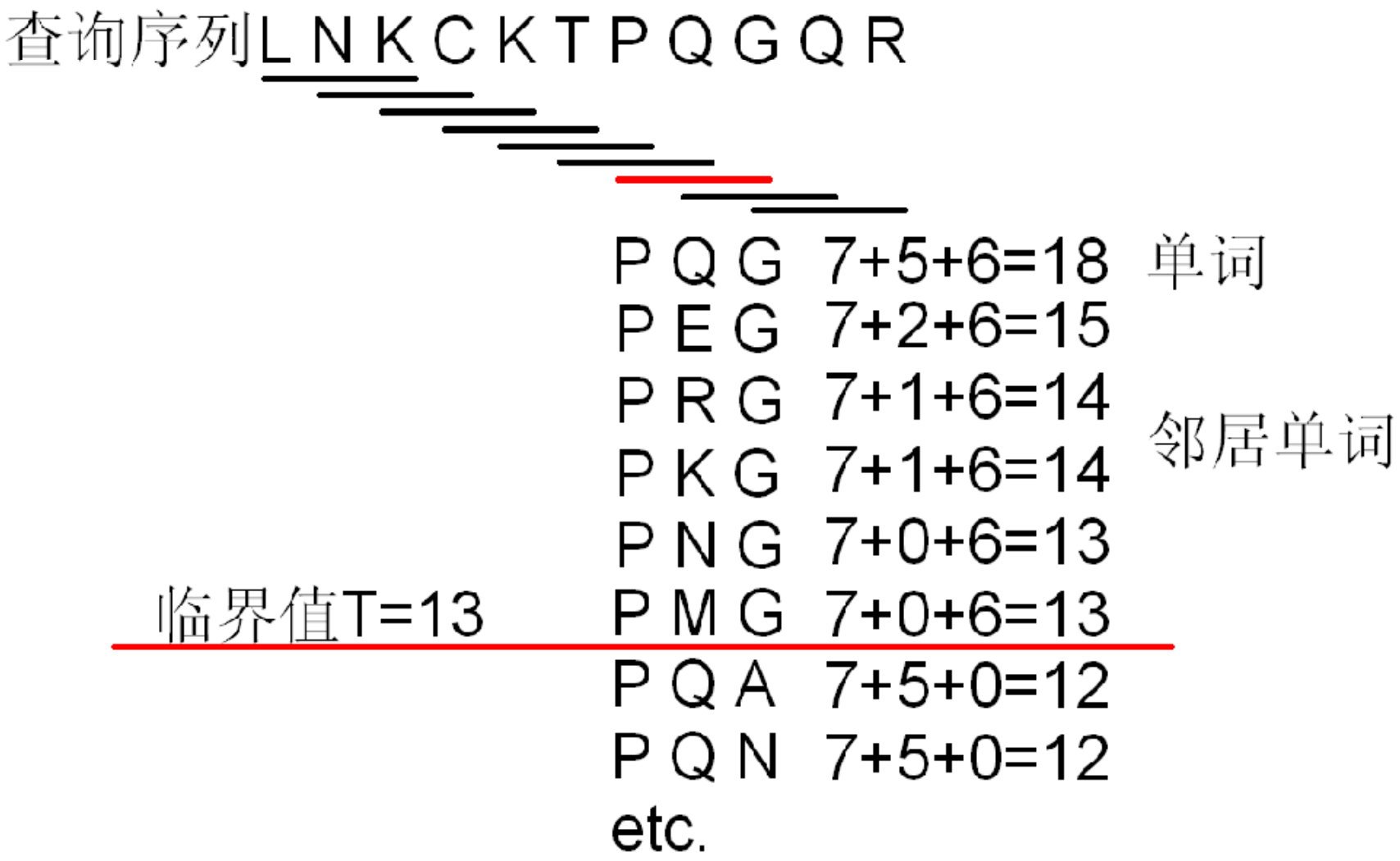
$$K = \frac{1}{L}\log_N\left(\frac{L!}{\prod_i n_i!}\right)$$

- 因此需要用如下字母去掩盖这些区域
  - Ns（for nucleotide residues）
  - Xs（for amino acid residues）

第2步 Seeding

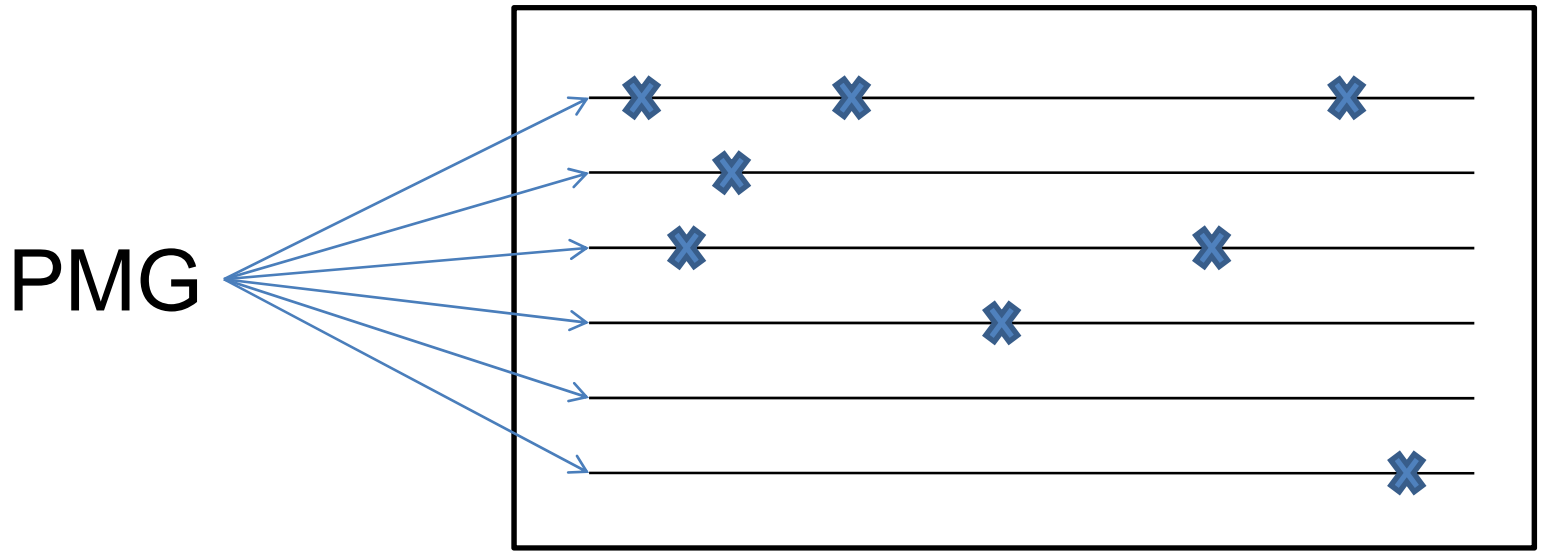
- 对查询序列作如下处理：
- 给定单词（word）长度w（蛋白为3，核酸为11）和打分矩阵，将长度为n的查询序列从第一位到最后一位拿到n-w+1个单词
- 创建一个单词列表，其中的单词为经与上述单词打分后分数高于T的长度为w的单词。

第2步 Seeding


查询序列L N K C K T P Q G Q R
P Q G 7+5+6=18
单词
P E G 7+2+6=15
P R G 7+1+6=14
P K G 7+1+6=14
邻居单词
P N G 7+0+6=13
临界值T=13
P M G 7+0+6=13
P Q A 7+5+0=12
P Q N 7+5+0=12
etc.

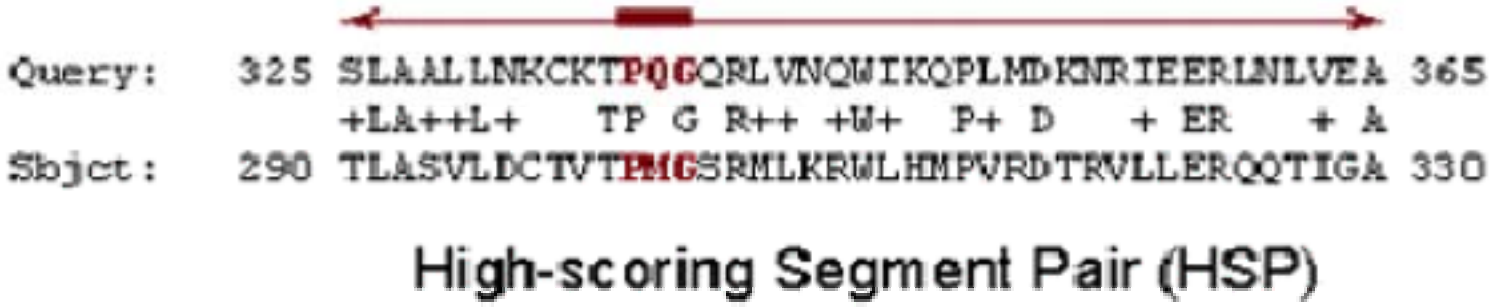
第2步 Seeding

- 对于邻居单词列表中的每一个单词在所有的数据库序列中找到其出现的每一个位置。



第3步 Extending

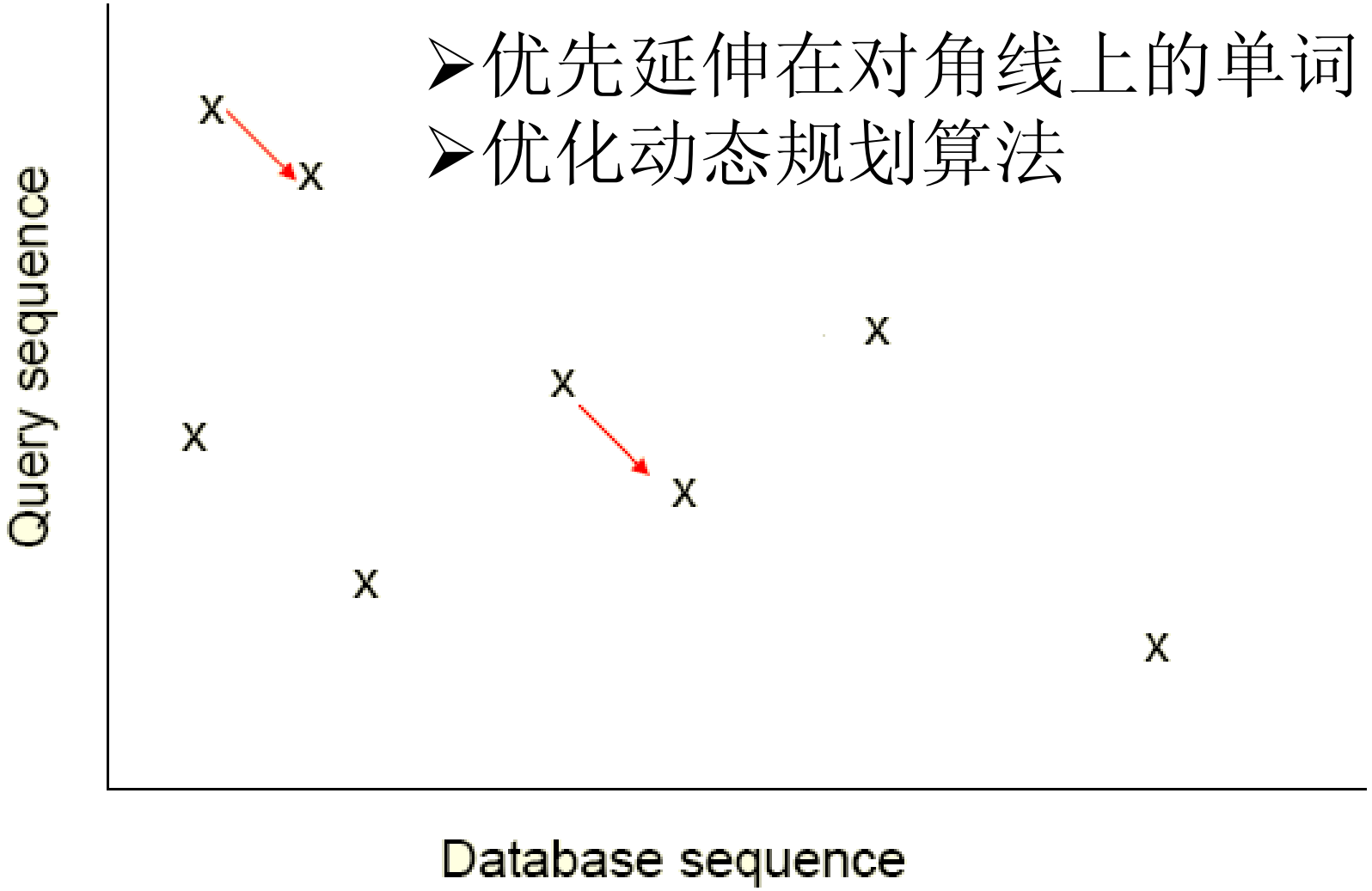
- 用动态规划的方法沿左右两个方向延伸种子直到打分不低于一个临界值（允许暂时低于临界值）。得到的结果称为高分片段对（high-scoring segment pair, HSP）。



High-scoring Segment Pair (HSP)

第3步 Extending

- 优先延伸在对角线上的单词
- 优化动态规划算法



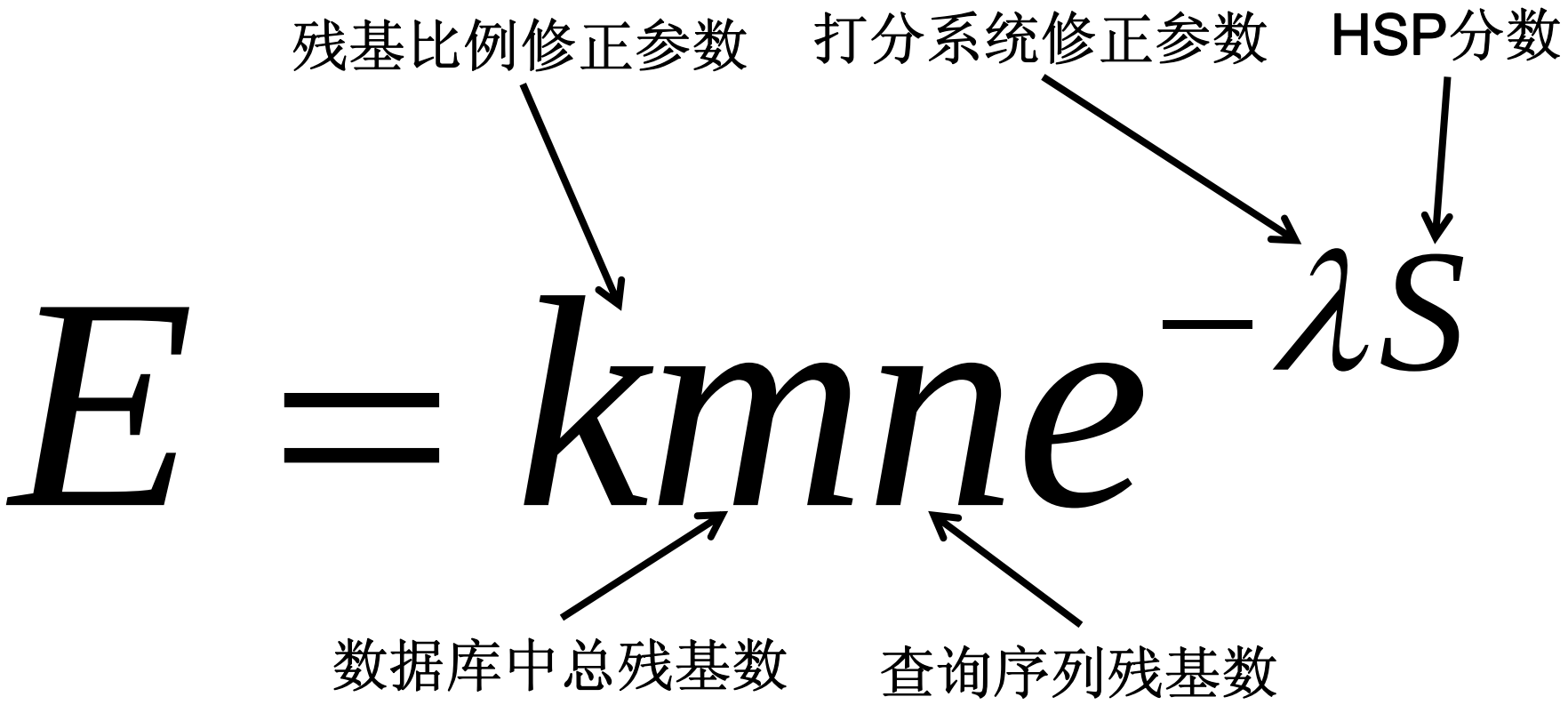
第4步

- BLAST算法得到的结果的统计显著性分析采用E-value。
- E-value：期望值（expect value）。它是指对一个给定的打分值在随机情况下在数据库中搜索比对的结果数目的期望。

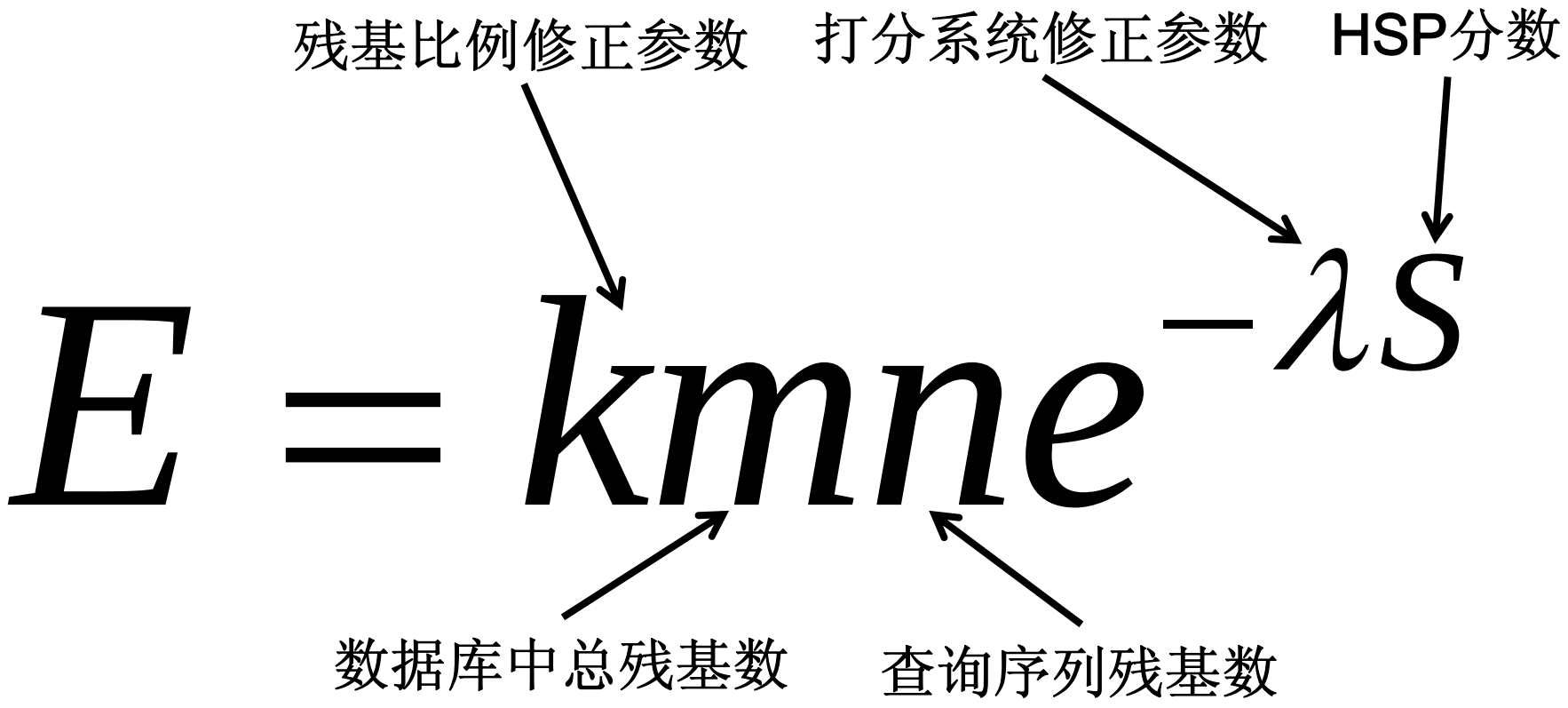
第4步

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

第4步



第4步



（在Xiong的书中 $E = mnP, P = ke^{-\lambda S}$）

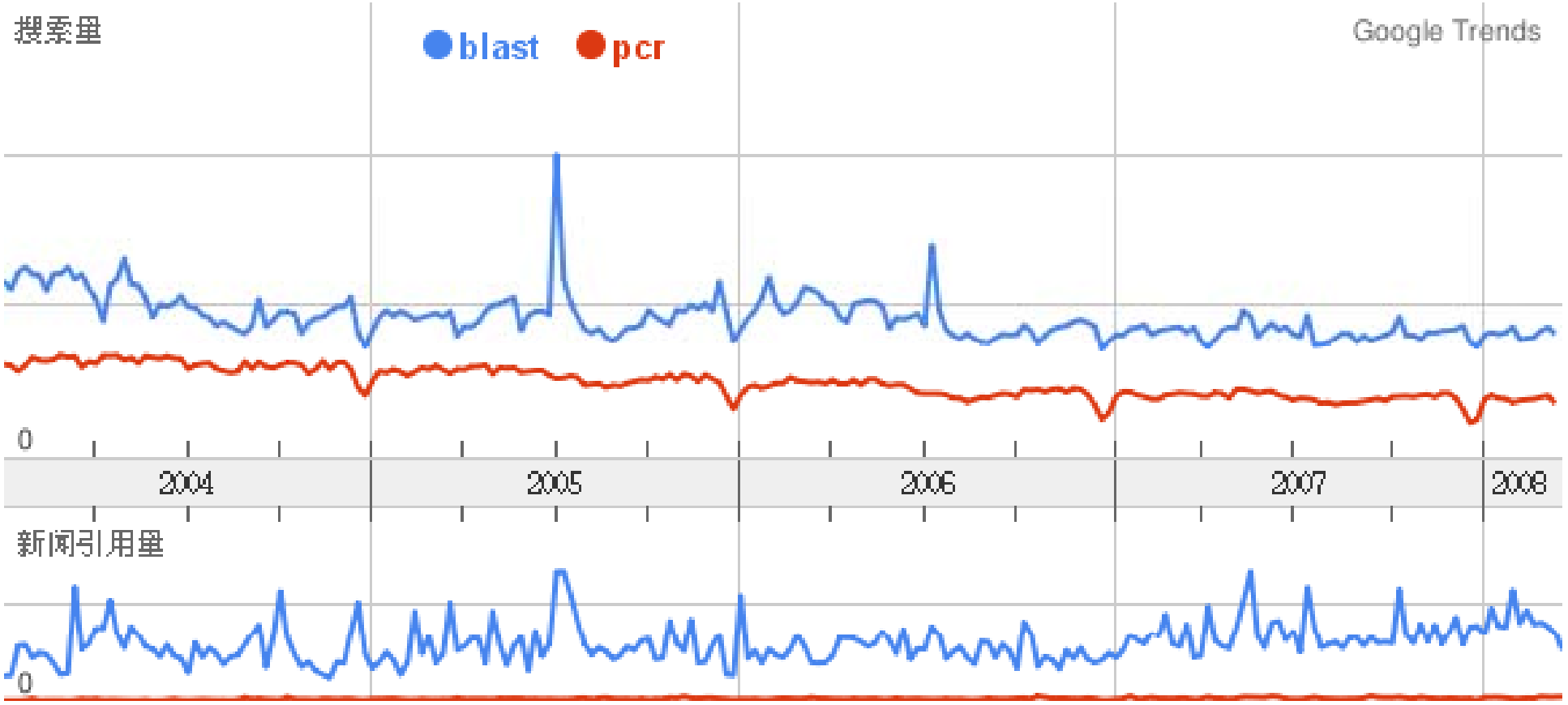
HSP是随机结果的概率

BLAST简介

BLAST算法

BLAST应用

BLAST is popular.



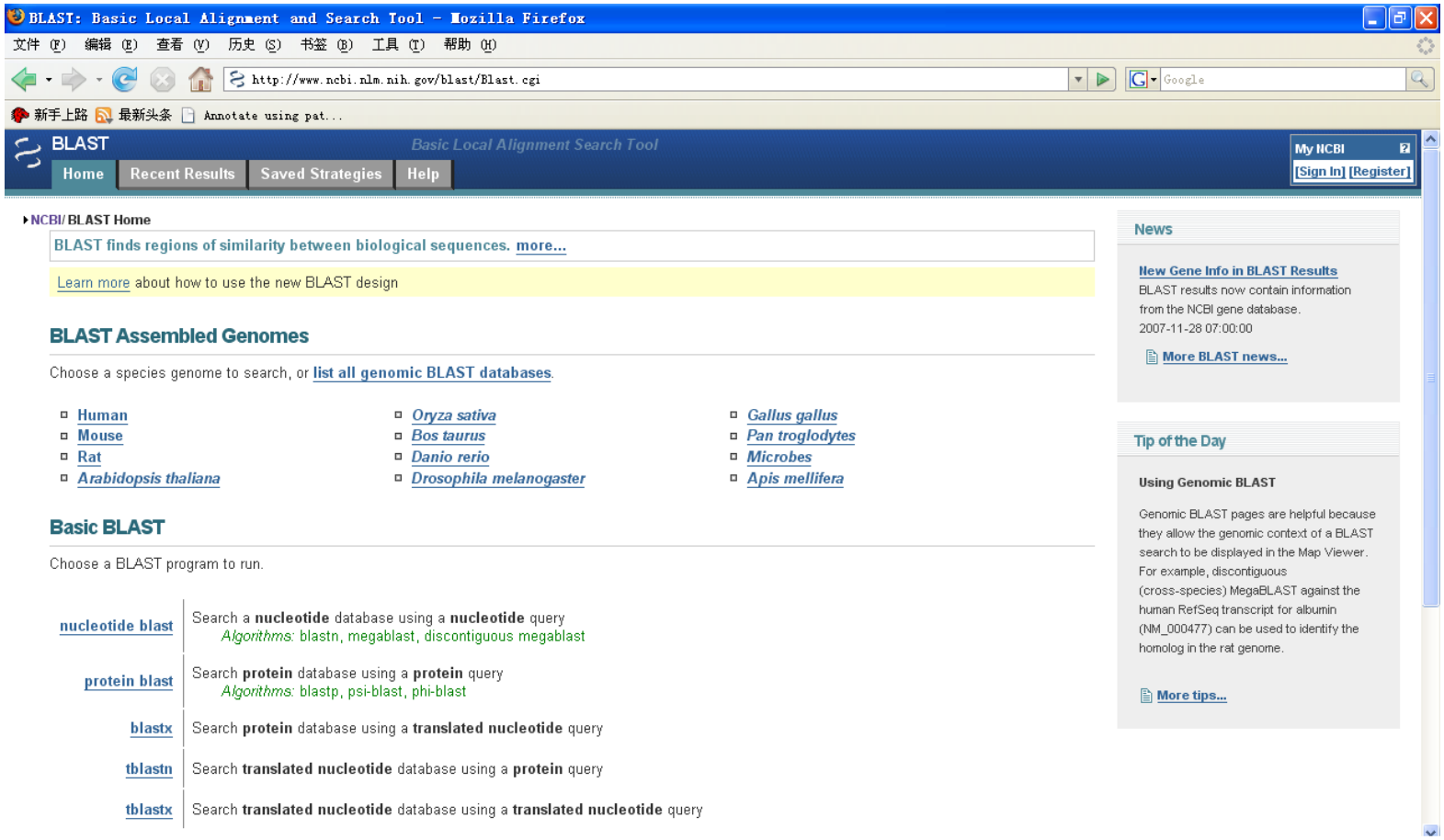
何处**BLAST**

- NCBI - National Center for Biotechnology Information (US)
- EBI - European Bioinfromatics Institute (EU)
- TIGR - The Genome Institute (US)
- Sanger - Sanger Institute (UK)
- UK-CropNet - The UK Crop Plant Bioinformatics Network (UK)
- WU-BLAST - Washington University (US)

（From LuoJC）

NCBI BLAST

- http://www.ncbi.nlm.nih.gov/blast/Blast.cgi



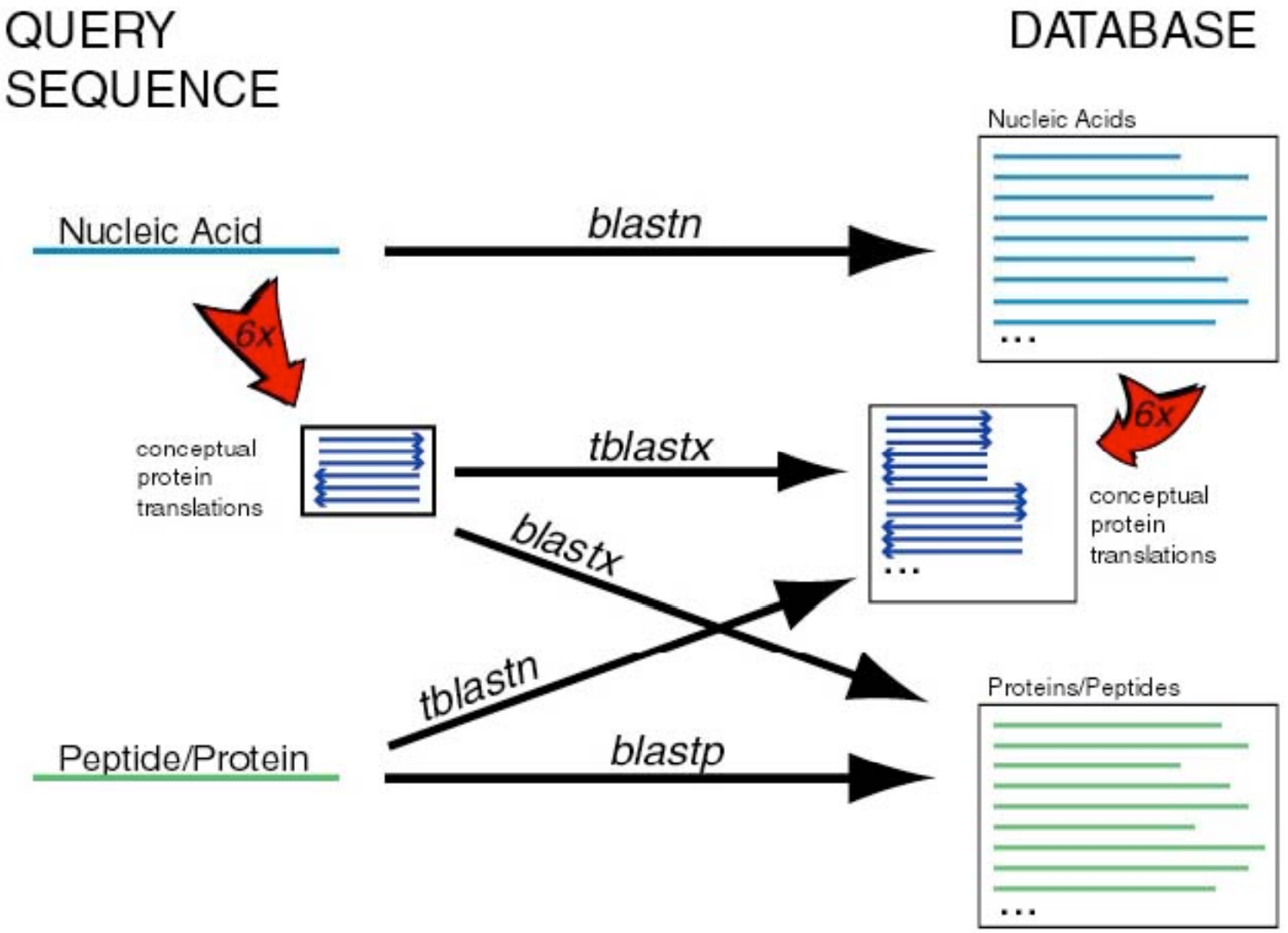
第1步 选择BLAST程序

程序	查询序列	数据库序列
blastp	蛋白	蛋白
blastn	核酸	核酸
blastx	翻译的核酸	蛋白
tblastn	蛋白	翻译的核酸
tblastx	翻译的核酸	翻译的核酸

第1步 选择BLAST程序

程序	典型用途
blastp	确定蛋白共同区域；为系统发育分析收集相关蛋白
blastn	将寡核苷酸、EST、重复序列映射到基因组；确定相关转录本
blastx	在基因组DNA上找编码蛋白基因
tblastn	确定与已知蛋白相似的转录本；将蛋白映射到基因组DNA
tblastx	跨物种基因预测；搜索基因

第1步 选择BLAST程序



(From Joel, H.)

第2步 输入查询序列

- 可以用3种格式输入查询序列：
  - FASTA格式序列
  - NCBI Accession number
  - GI

第3步 选择BLAST数据库（蛋白）

Database	Description
nr	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
month	All new or revised GenBank CDS translation+PDB+SwissProt+PIR released in the last 30 days.
swissprot	The last major release of the SWISS-PROT protein sequence database (no updates). These are uploaded to our system when they are received from EMBL.
patents	Protein sequences derived from the Patent division of GenBank.
yeast	Yeast (Saccharomyces cerevisiae) protein sequences. This database is not to be confused with a listing of all Yeast protein sequences. It is a database of the protein translations of the Yeast complete genome.
E. coli	E. coli (Escherichia coli) genomic CDS translations.
pdb	Sequences derived from the 3-dimensional structure Brookhaven Protein Data Bank.
kabat	Kabat's database of sequences of immunological interest.
alu	Translations of select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences.

第3步 选择BLAST数据库（核酸）

Database	Description
nr	All non-redundant GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or HTGS sequences).
month	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
dbest	Non-redundant database of GenBank+EMBL+DDBJ EST Divisions.
dbsts	Non-redundant database of GenBank+EMBL+DDBJ STS Divisions.
mouse ests	The non-redundant Database of GenBank+EMBL+DDBJ EST Divisions limited to the organism mouse.
human ests	The Non-redundant Database of GenBank+EMBL+DDBJ EST Divisions limited to the organism human.
other ests	The non-redundant database of GenBank+EMBL+DDBJ EST Divisions all organisms except mouse and human.
yeast	Yeast (Saccharomyces cerevisiae) genomic nucleotide sequences. Not a collection of all Yeast nucelotides sequences, but the sequence fragments from the Yeast complete genome.
E. coli	E. coli (Escherichia coli) genomic nucleotide sequences.
pdb	Sequences derived from the 3-dimensional structure of proteins.
kabat	Kabat's database of sequences of immunological interest. For more information http://immuno.bme.nwu.edu/
patents	Nucleotide sequences derived from the Patent division of GenBank.
vector	Vector subset of GenBank(R), NCBI, (ftp://ncbi.nlm.nih.gov/pub/blast/db/ directory).
mito	Database of mitochondrial sequences (Rel. 1.0, July 1995).
alu	Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. It is available at ftp://ncbi.nlm.nih.gov/pub/jmc/alu. See "Alu alert" by epd
gss	Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
htgs	High Throughput Genomic Sequences.

第4步 选择比对参数

- 一般参数
  - 显示序列最大数目，（短查询序列），期望临界值，单词长度
- 打分参数
  - 矩阵，空位罚分，（组成调整）
- 过滤和掩盖
  - 过滤，掩盖

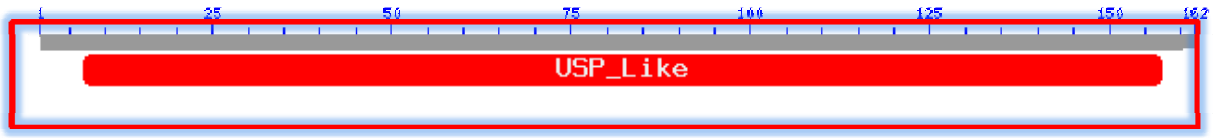
第5步 得到结果

- 主要包含五部分：
  - 保守区域（conserved domains）
  - 图形化的概括框（graphical overview box）
  - 匹配列表（matching list）
  - 比对文本描述（text discription of the alignment）
  - BLAST部分参数

保守区域（**conserved domains**）

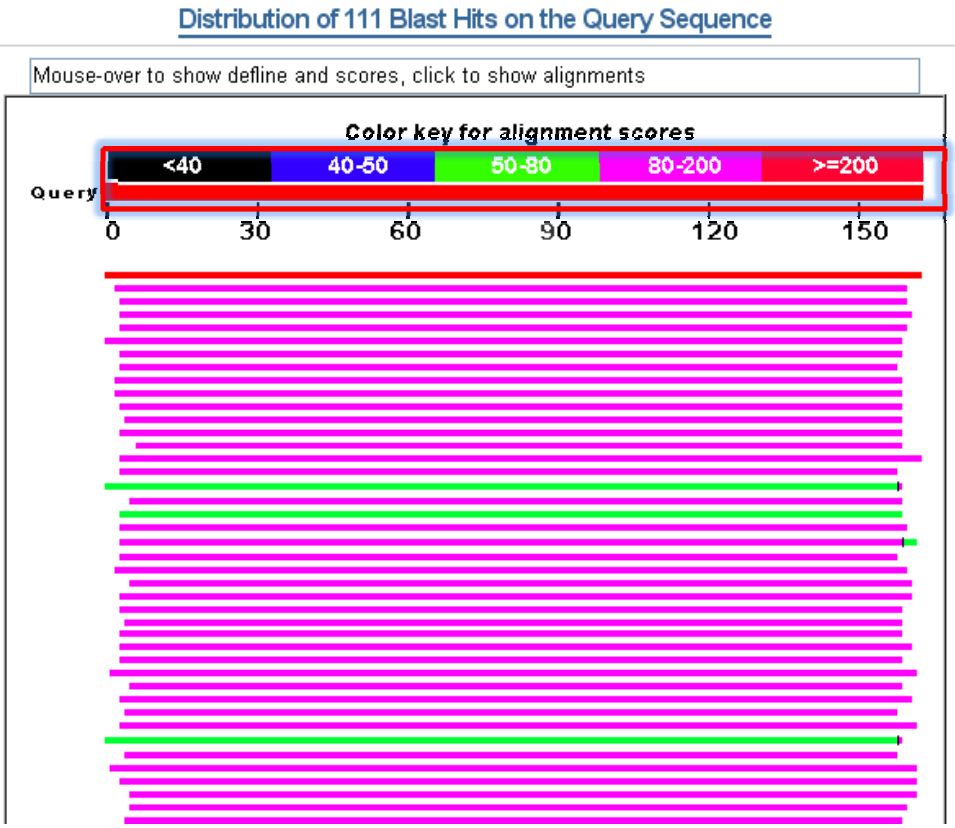
Job Title: gi|2501594|sp|Q57997|Y577_METJA PROTEIN MJ0577

Putative conserved domains have been detected, click on the image below for detailed results.

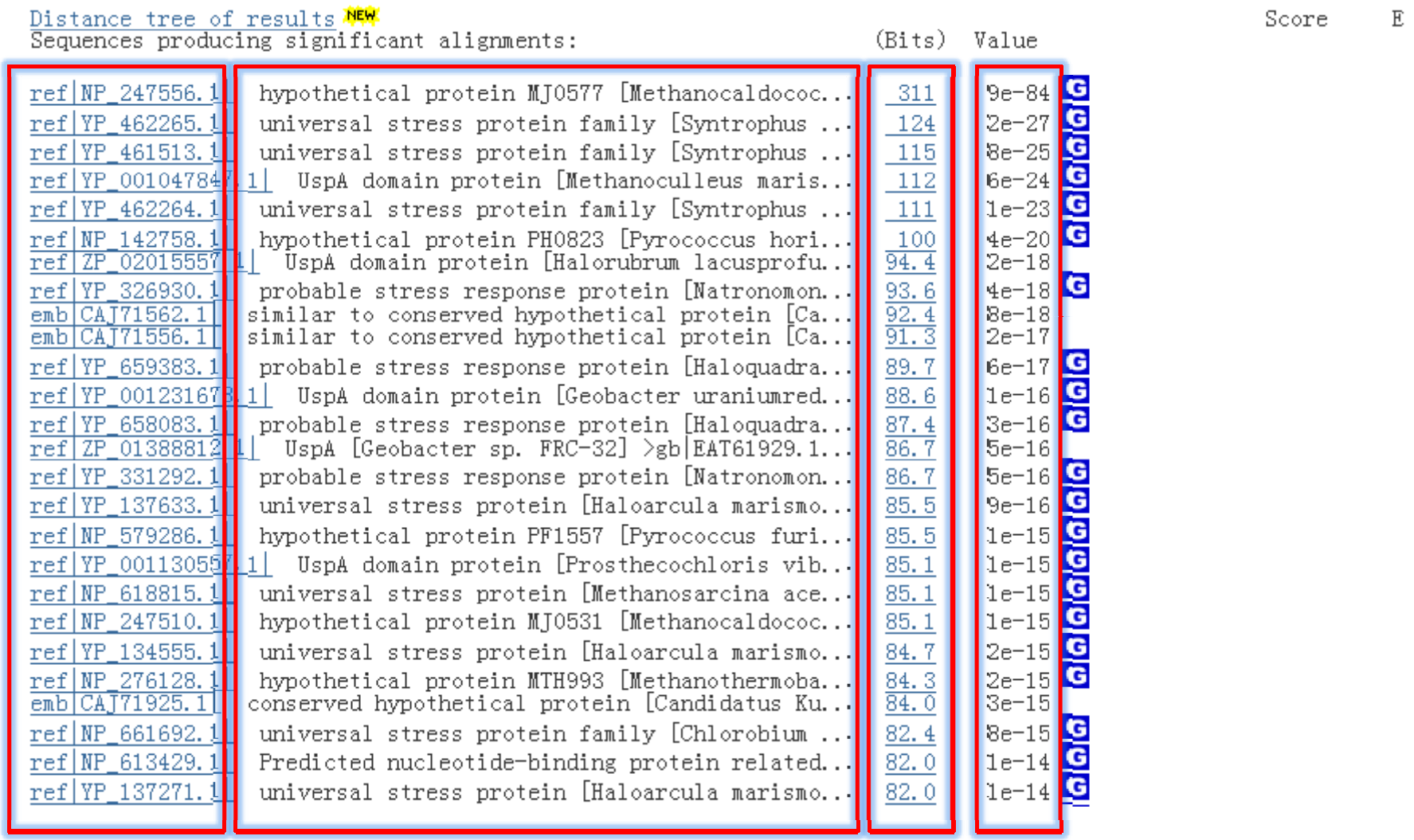


图形化的概括框（graphical overview box）

Query= gi|2501594|sp|Q57997|Y577_METJA PROTEIN MJ0577
Length=162



匹配列表（**matching list**）



Distance tree of results NEW

Sequences producing significant alignments:		(Bits)	Score E Value
ref|NP_247556.1|	hypothetical protein MJ0577 [Methanocaldococ...	311	9e-84
ref|YP_462265.1|	universal stress protein family [Syntrophus ...	124	2e-27
ref|YP_461513.1|	universal stress protein family [Syntrophus ...	115	8e-25
ref|YP_001047847.1|	UspA domain protein [Methanoculleus maris...	112	6e-24
ref|YP_462264.1|	universal stress protein family [Syntrophus ...	111	1e-23
ref|NP_142758.1|	hypothetical protein PH0823 [Pyrococcus hori...	100	4e-20
ref|ZP_02015557.1|	UspA domain protein [Halorubrum lacusprofu...	94.4	2e-18
ref|YP_326930.1|	probable stress response protein [Natronomon...	93.6	4e-18
emb|CAJ71562.1|	similar to conserved hypothetical protein [Ca...	92.4	8e-18
emb|CAJ71556.1|	similar to conserved hypothetical protein [Ca...	91.3	2e-17
ref|YP_659383.1|	probable stress response protein [Haloquadra...	89.7	6e-17
ref|YP_001231673.1|	UspA domain protein [Geobacter uraniumred...	88.6	1e-16
ref|YP_658083.1|	probable stress response protein [Haloquadra...	87.4	3e-16
ref|ZP_01388812.1|	UspA [Geobacter sp. FRC-32] >gb|EAT61929.1...	86.7	5e-16
ref|YP_331292.1|	probable stress response protein [Natronomon...	86.7	5e-16
ref|YP_137633.1|	universal stress protein [Haloarcula marismo...	85.5	9e-16
ref|NP_579286.1|	hypothetical protein PF1557 [Pyrococcus furi...	85.5	1e-15
ref|YP_001130557.1|	UspA domain protein [Prosthecochloris vib...	85.1	1e-15
ref|NP_618815.1|	universal stress protein [Methanosarcina ace...	85.1	1e-15
ref|NP_247510.1|	hypothetical protein MJ0531 [Methanocaldococ...	85.1	1e-15
ref|YP_134555.1|	universal stress protein [Haloarcula marismo...	84.7	2e-15
ref|NP_276128.1|	hypothetical protein MTH993 [Methanothermoba...	84.3	2e-15
emb|CAJ71925.1|	conserved hypothetical protein [Candidatus Ku...	84.0	3e-15
ref|NP_661692.1|	universal stress protein family [Chlorobium ...	82.4	8e-15
ref|NP_613429.1|	Predicted nucleotide-binding protein related...	82.0	1e-14
ref|YP_137271.1|	universal stress protein [Haloarcula marismo...	82.0	1e-14

比对文本描述（text discription of the alignment）

>ref|NP_247556.1| hypothetical protein MJ0577 [Methanocaldococcus jannaschii DSM 2661]
sp|Q57997|Y577_METJA Uncharacterized protein MJ0577
pdb|1MJH|A Chain A, Structure-Based Assignment Of The Biochemical Function Of Hypothetical Protein Mj0577: A Test Case Of Structural Genomics
pdb|1MJH|B Chain B, Structure-Based Assignment Of The Biochemical Function Of Hypothetical Protein Mj0577: A Test Case Of Structural Genomics
gb|AAB98568.1| conserved hypothetical protein [Methanocaldococcus jannaschii DSM 2661]
Length=162

GENE ID: 1451442 MJ0577 | hypothetical protein [Methanocaldococcus jannaschii DSM 2661] **(10 or fewer PubMed links)**

Score = 311 bits (797), Expect = 9e-84, Method: Compositional matrix adjust.
Identities = 162/162 (100%), Positives = 162/162 (100%), Gaps = 0/162 (0%)

```
Query  1    MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA  60
            MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA
Sbjct  1    MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA  60

Query  61   GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG  120
            GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG
Sbjct  61   GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG  120

Query  121  VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS  162
            VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS
Sbjct  121  VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS  162
```

BLAST部分参数

```
  Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding
environmental samples from WGS projects
    Posted date:  Mar 28, 2008  5:59 PM
  Number of letters in database: -2,118,652,067
  Number of sequences in database:  6,373,249

Lambda     K      H
   0.313    0.134    0.349
Gapped
Lambda     K      H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 6373249
Number of Hits to DB: 51131832
Number of extensions: 2065059
Number of successful extensions: 12372
Number of sequences better than 100: 517
Number of HSP's better than 100 without gapping: 0
Number of HSP's gapped: 12283
Number of HSP's successfully gapped: 572
Length of query: 162
Length of database: 2176315225
Length adjustment: 123
Effective length of query: 39
Effective length of database: 1392405598
Effective search space: 54303818322
Effective search space used: 54303818322
T: 11
A: 40
X1: 16 (7.2 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 42 (20.8 bits)
S2: 64 (29.3 bits)
```

Advanced BLAST

- Megablast
  - nucleotide only
  - optimized for large batch searches
- PSI-BLAST
  - constructs PSSMs automatically
  - searches protein database with PSSMs
- RPS BLAST
  - searches a database of PSSMs
  - basis of conserved domain database

(From GaoG)

Some ways to be good

- Check data, whenever possible.
  - Carefulness is good.
- Check the output, carefully
  - Extraordinary claims require extraordinary evidence
- Use your mind.
  - Tool is good, understanding tool is better.

（From GaoG）

知其道 用其妙 THIS IS HOW:

SIEMENS

References:

1. Mount, D., Bioinfomatics Sequence and Genome Analysis, (2002), COLD SPRING HARBAR LABORATORY PRESS.
2. Xiong, J., Essential Bioinformatics, (2006), CAMBRIDGE UNIVERSITY PRESS.
3. http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html
4. ppts from Luo, JC and Gao, G.

Further Reading

- Papers:
  - Needleman, *et al*. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48:** 443-453.
  - Smith, *et al*. (1981). Identification of common molecular subsequences. *J Mol Biol* **147:** 195-197.
  - Dayhoff, *et al*. (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, pp. 345–352.
  - Henikoff, *et al*. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89:** 10915-10919.
  - Altschul, *et al*. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res*. **25**:3389-3402.
  - Pertsemlidis, *et al*. (2001) "Having a BLAST with bioinformatics (and avoiding BLASTphemy)" *Genome Biol* 2, REVIEWS2002
  - McGinnis, *et al*. (2004) "BLAST: at the core of a powerful and diverse set of sequence analysis tools." *Nucleic Acids Res*. **32**:W20-W25.
- Books:
  - Durbin, *et al*. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids,(1998). Cambridge Cambridge University Press.
  - Korf, *et al*. (2003) BLAST (O'Reilly).
- Web:
  - http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

（From GaoG）

Thanks!