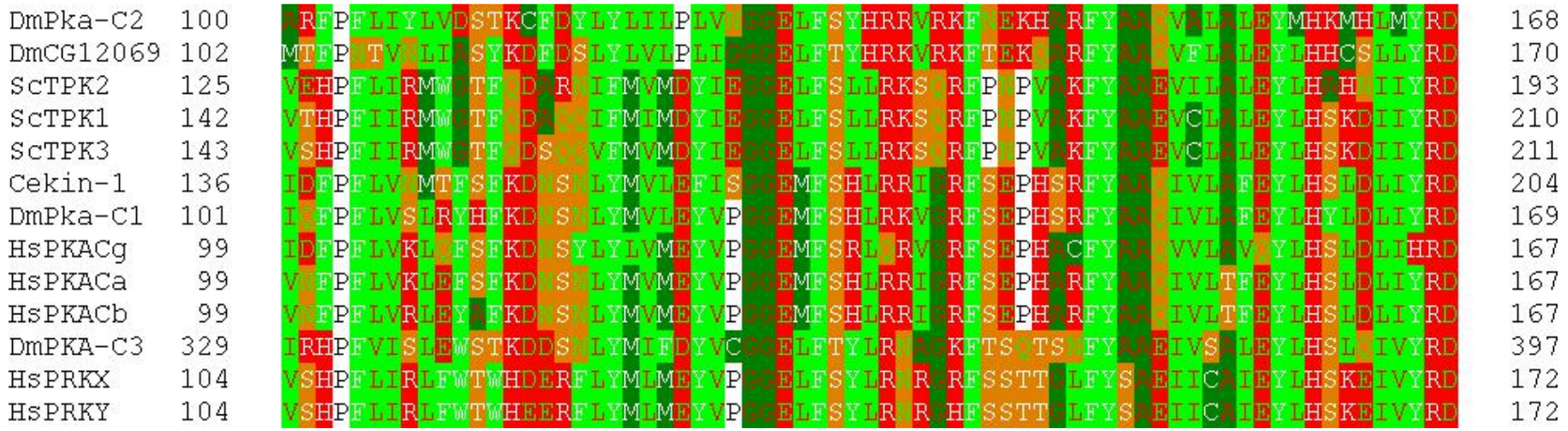
多序列比对-工具篇

为什么进行多序列比对

- 序列保守 意味着 潜在的功能保守
  - 不同物种中的同源基因，功能保守，序列相似性较高
  - 通过多条序列的比较，发现保守与变异的部分
- 构建分子进化树的必须步骤
- 比较基因组学研究的基础
- 两类：全局或局部的多序列比对
- 本章：全局多序列比对

全局多序列比对

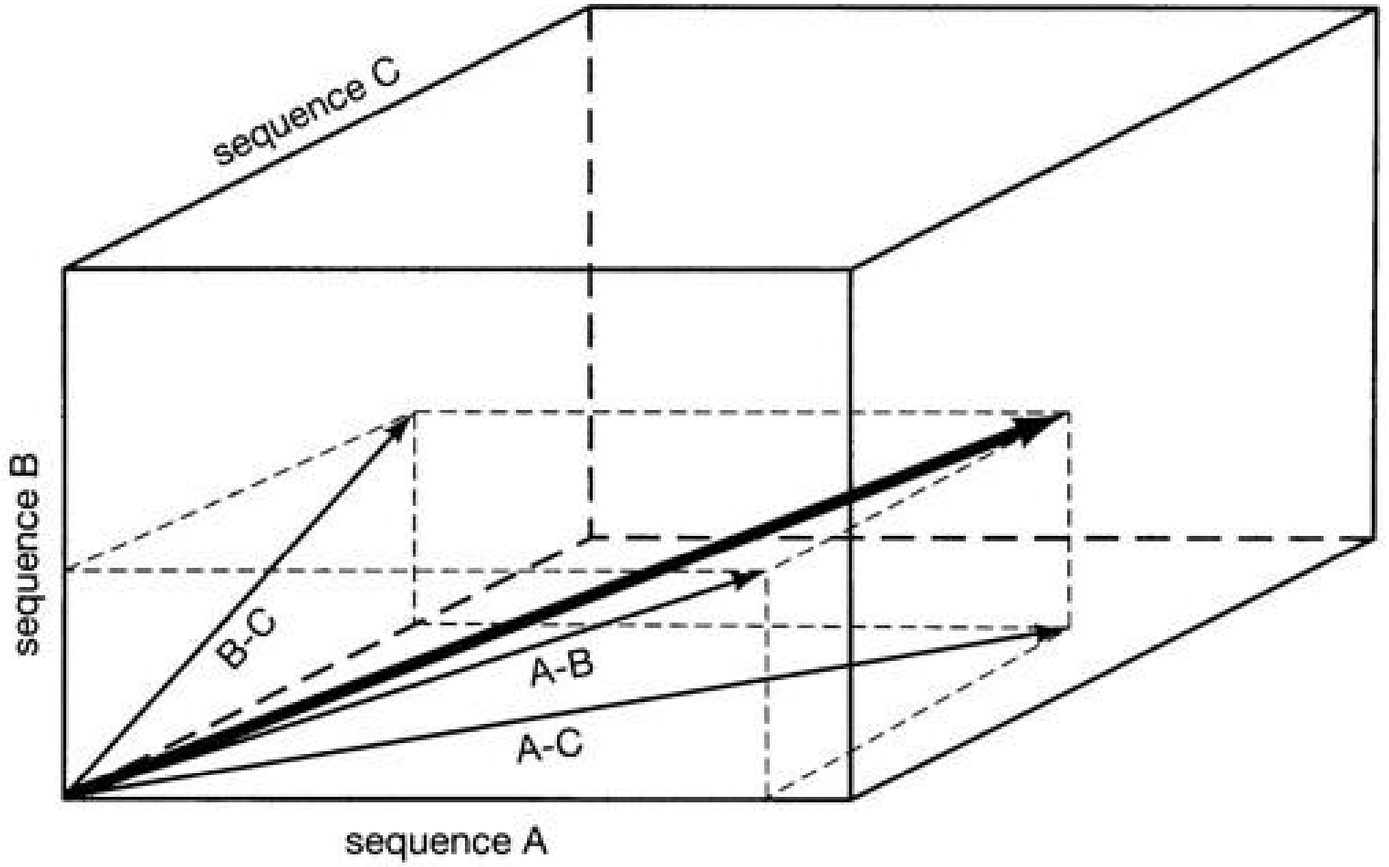
- 蛋白激酶**PKA**家族的多序列比对结果（部分）



Made by GENEDOC

http://genedoc.software.informer.com/

动态规划算法：全空间



http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html

注意

- 最优的多序列比对，其两两序列之间的比对 不一定最优

```
V S N _ S
_ S N A _
_ _ _ A S
```

最优的多序列比对

```
_ S N A _
_ _ _ A S
```

非最优的双序列比对

多序列比对的计算方法

- 渐进方法：**Progressive methods**
- 迭代方法：**Iterative refinement**
- 部分有向图算法
- 隐马尔科夫模型：**HMM profile-profile**
- 整合算法：**Meta-methods**
- 结构特征

渐进方法

Progressive methods

- 渐进方法：Pairwise alignment
- ClustalW/X: "Classic Clustal"
  - http://www.clustal.org/
  - http://www.clustal.org/clustal2/
- T-Coffee
  - http://tcoffee.org/
  - http://tcoffee.crg.cat/apps/tcoffee/all.html

ClustalW/X

- Clustal: 1988年开发
- ClustalW: 1994年，Julie D. Thompson 等人改进、开发
- ClustalX: 1997年，图形化软件

Table 1
Multiple Alignment Methods

Method (Developer)	Algorithm	Matrix[a]	Indels	Limits[b]	Assumptions[c]	Features[d]	Data Type[e]
Global:							
AMULT (G. Barton)	NW	Any	C		Y, S	R, SE	P
ASSEMBLE (M. Vingron)	Dot matrix NW	Log odds	I+E		Y, S		P
CLUSTAL V (D. Higgens)	WL	Any	I+E			I	P, N
DFALIGN (D.-F. Feng)	NW	Log odds	C	UP	Y, E, O		P
GENALIGN[f] (H. Martinez)	CW, NW	UM	I+E			SE	P, N
MSA (S. Altschul)	CL	PAM250	I+E	ROS	N	B, FA	P
MULTAL (W. Taylor)	NW	UM, PAM250	C		S	AP, FA	P
MWT (J. Kececioglu)	maximum weight trace	Any	C	ROS	N		P
TULLA (S. Subbiah)	NW	Any	RGW	10 sequences	S	R, SE	P
Local:							
MACAW (G. Schuler)	SW	PAM250		DOS	Y	SE, FA, MD	P
PIMA (P. Smith)	SW	AACH	I+E		Y	MD	P
PRALIGN (M. Waterman)	CW	PAM250	I+E[g]		Y	MD, MC	P, N[h]

ClustalW/X：计算过程

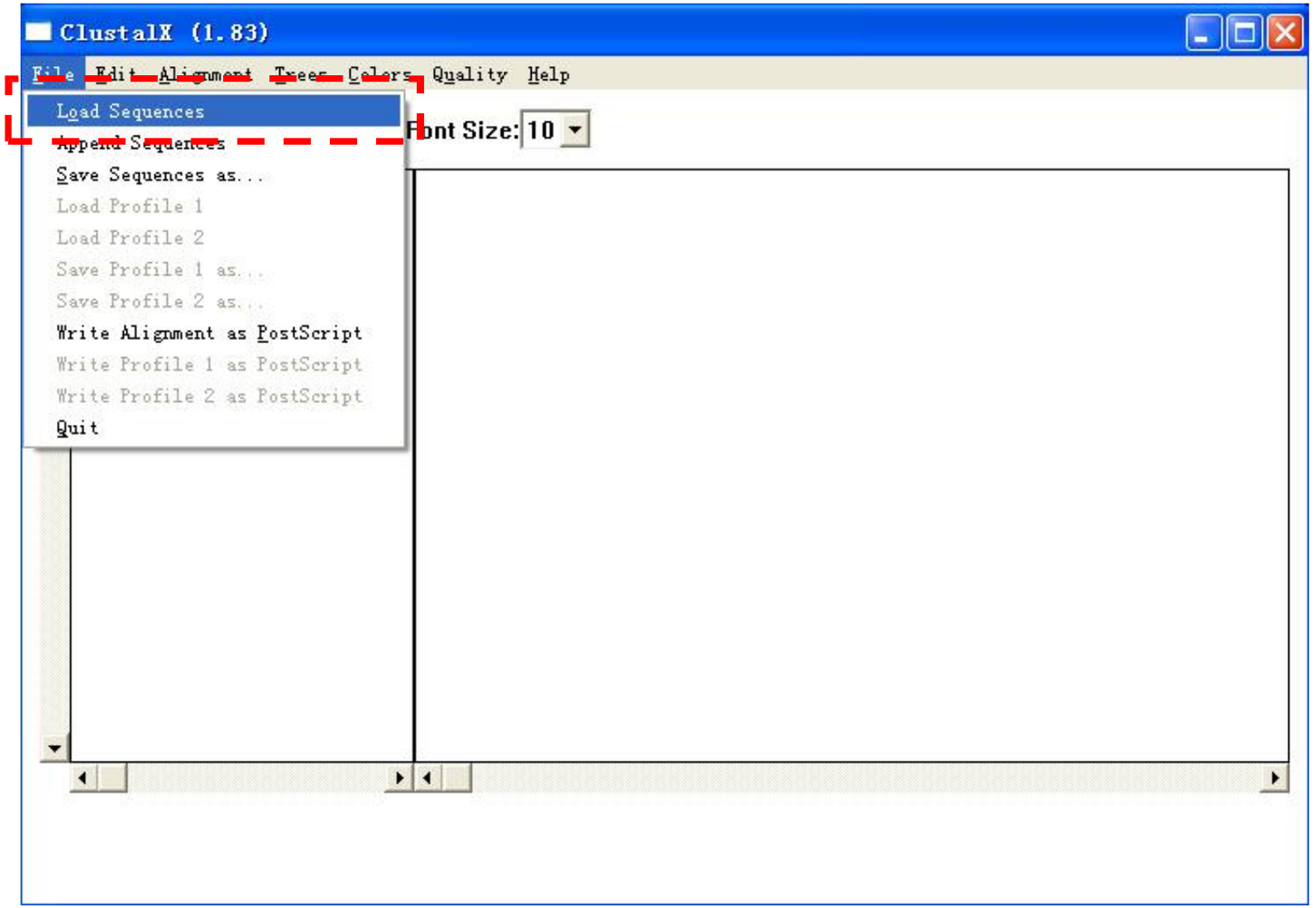
- 将所有序列两两比对，计算进化距离（差 异）矩阵
- 使用邻接法（**neighbor-joining**）构建指导树（**guide tree**）
- 将进化距离最近的两条序列用全局动态规 划算法进行比对
- “渐进”地加上其他序列

ClustalX: 使用指南

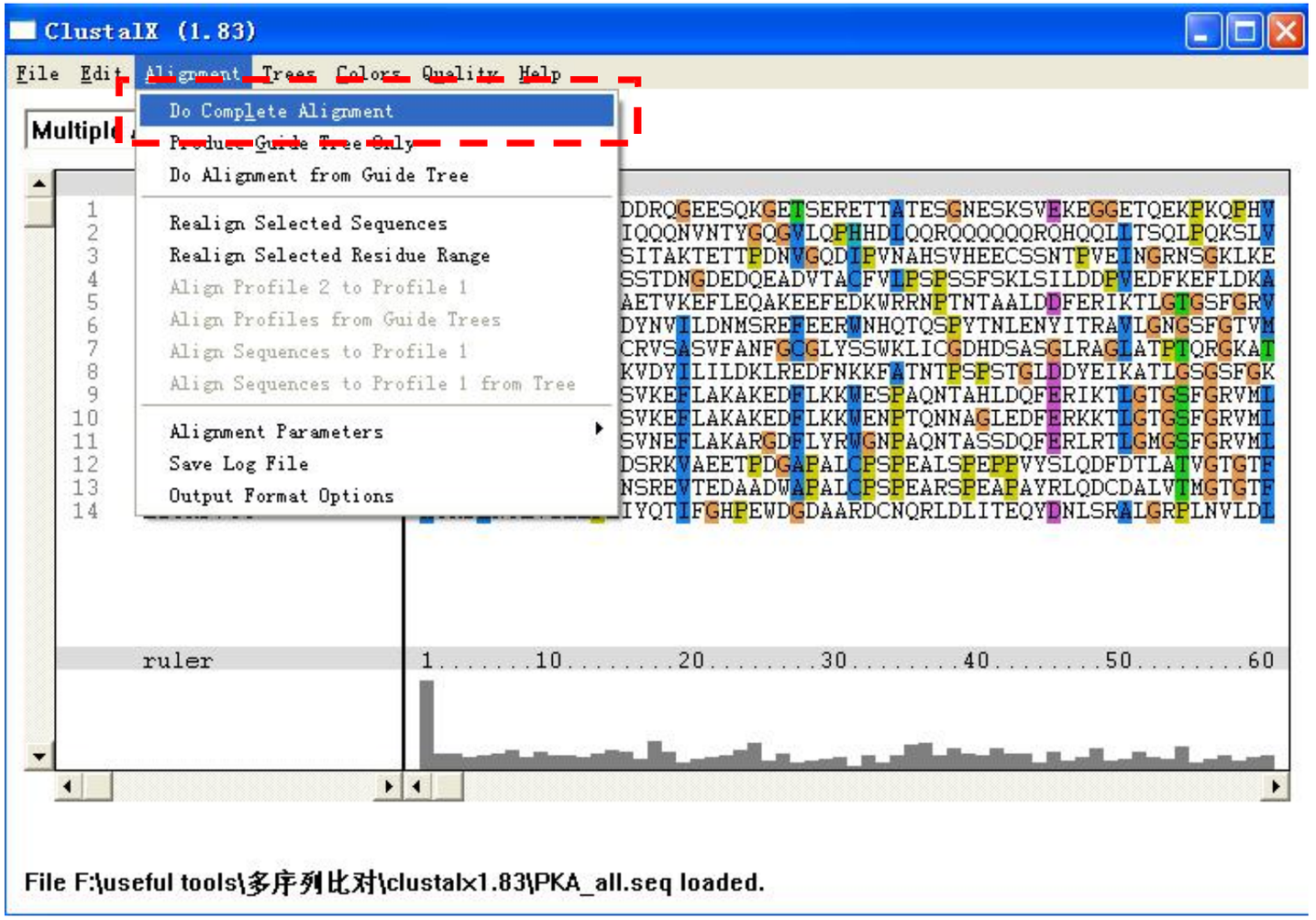
- **FASTA序列格式，多序列**

```
>ScTPK1
MSTEEQNGGGQKSLDDRQGEESQKGETSERETTATESGNESKSVEKEGGETQEKPKQPHV
TYYNEEQYKQFIAQARVTSGKYSLQDFQILRTLGTGSFGRVHLIRSRHNGRYYAMKVLKK
EIVVRLKQVEHTNDERLMLSIVTHPFIIRMWGTFQDAQQIFMIMDYIEGGELFSLLRKSQ
RFPNPVAKFYAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFGFAKYVPDVTY
TLCGTPDYIAPEVVSTKPYNKSIDWWSFGILIYEMLAGYTPFYDSNTMKTYEKILNAELR
FPPFFNEDVKDLLSRLITRDLSQRLGNLQNGTEDVKNHPWFKEVVWEKLLSRNIETPYEP
PIQQGQGDTSQFDKYPEEDINYGVQGEDPYADLFRDF
>ScTPK2
MEFVAERAQPVGQTIQQQNVNTYGQGVLQPHHDLQQRQQQQQQRQHQQLLTSQLPQKSLV
SKGKYTLHDFQIMRTLGTGSFGRVHLVRSVHNGRYYAIKVLKKQQVVKMKQVEHTNDERR
MLKLVEHPFLIRMWGTFQDARNIFMVMDYIEGGELFSLLRKSQRFPNPVAKFYAAEVILA
LEYLHAHNIIYRDLKPENILLDRNGHIKITDFGFAKEVQTVTWTLCGTPDYIAPEVITTK
PYNKSVDWWSLGVLIYEMLAGYTPFYDTTPMKTYEKILQGKVVYPPYFHPDVVDLLSKLI
TADLTRRIGNLQSGSRDIKAHPWFSEVVWERLLAKDIETPYEPPITSGIGDTSLFDQYPE
EQLDYGIQGDDPYAEYFQDF
>ScTPK3
MYVDPMNNNEIRKLSITAKTETTPDNVGQDIPVNAHSVHEECSSNTPVEINGRNSGKLKE
EASAGICLVKKPMLQYRDTSGKYSLSDFQILRTLGTGSFGRVHLIRSNHNGRFYALKTLK
KHTIVKLKQVEHTNDERRMLSIVSHPFIIRMWGTFQDSQQVFMVMDYIEGGELFSLLRKS
QRFPNPVAKFYAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFGFAKYVPDVT
YTLCGTPDYIAPEVVSTKPYNKSVDWWSFGVLIYEMLAGYTPFYNSNTMKTYENILNAEL
KFPPFFHPDAQDLLKKLITRDLSERLGNLQNGSEDVKNHPWFNEVIWEKLLARYIETPYE
PPIQQGQGDTSQFDRYPEEEFNYGIQGEDPYMDLMKEF
>Cekin-1
MPTRLDIVGNLQFSSSTDNGDEDQEADVTACFVLPSPSSFSKLSILDDPVEDFKEFLDKA
REDFKQRWENPAQNTACLDDFDRIKTLGTGSFGRVMLVKHKQSGNYYAMKILDKQKVVKL
KQVEHTLNEKRILQAIDFPFLVNMTFSFKDNSNLYMVLEFISGGEMFSHLRRIGRFSEPH
SRFYAAQIVLAFEYLHSLDLIYRDLKPENLLIDSTGYLKITDFGFAKRVKGRTWTLCGTP
EYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVKFPSHFS
NELKDLLKNLLQVDLTKRYGNLKNGVADIKNHKWFGSTDWIAIYQKKITPPSFSKGESNG
RLFEALYPRVDGPADTRHFVEEVQEPTEFVIAATPQLEELFVEF
```

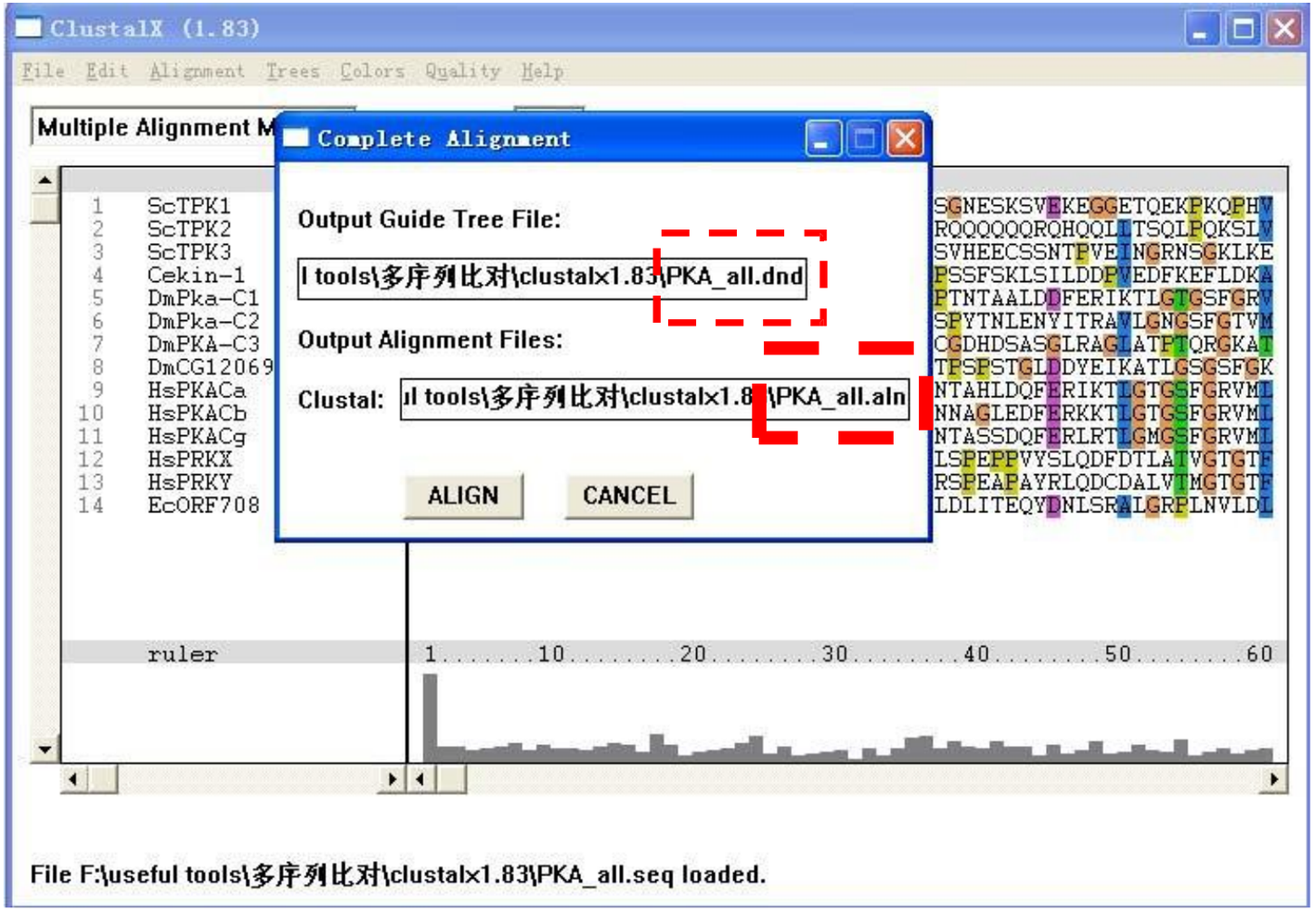
导入序列文件


ClustalX (1.83)
File Edit Alignment Trees Colors Quality Help
Load Sequences
Append Sequences
Save Sequences as...
Load Profile 1
Load Profile 2
Save Profile 1 as...
Save Profile 2 as...
Write Alignment as PostScript
Write Profile 1 as PostScript
Write Profile 2 as PostScript
Quit
Font Size: 10

执行比对

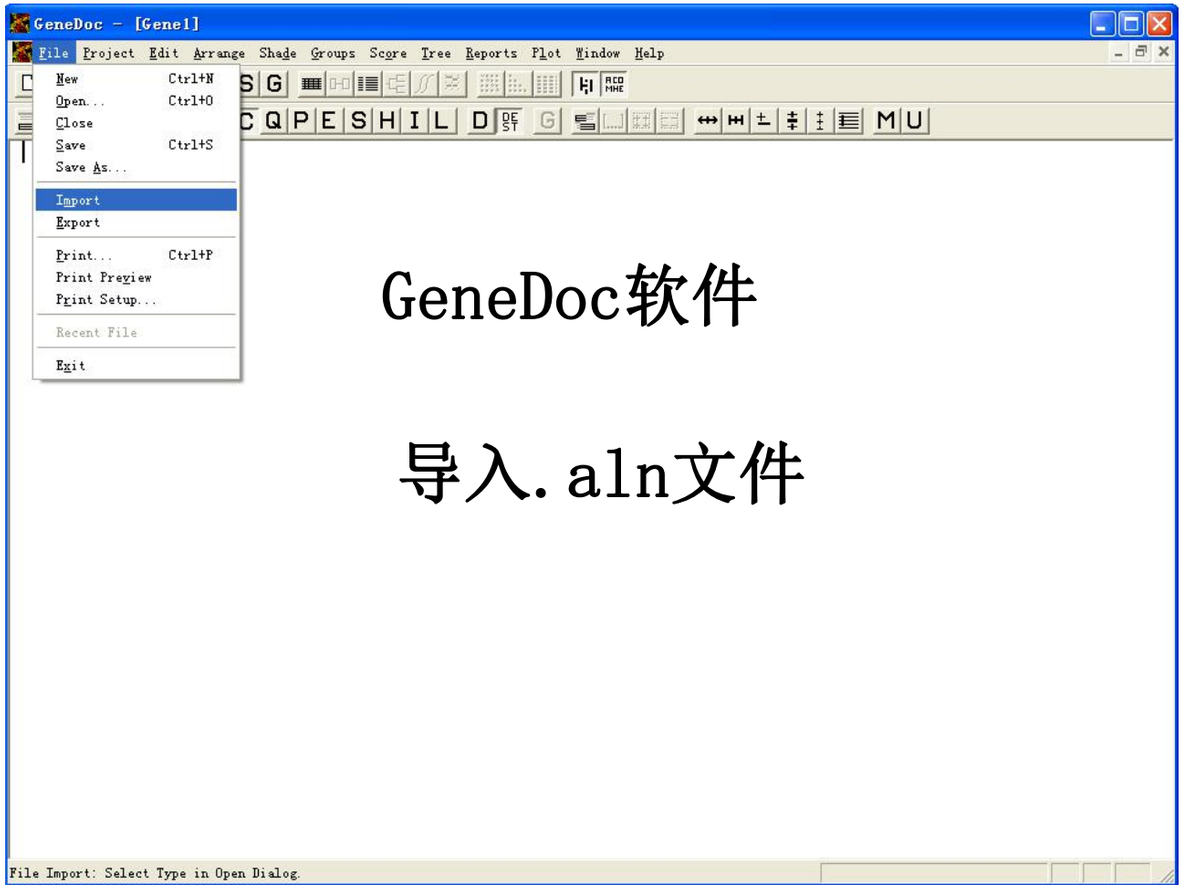

ClustalX (1.83)
File Edit Alignment Trees Colors Quality Help
Do Complete Alignment
Produce Guide Tree Only
Do Alignment from Guide Tree
Realign Selected Sequences
Realign Selected Residue Range
Align Profile 2 to Profile 1
Align Profiles from Guide Trees
Align Sequences to Profile 1
Align Sequences to Profile 1 from Tree
Alignment Parameters
Save Log File
Output Format Options
ruler
1.....10.....20.....30.....40.....50.....60
File F:\useful tools\多序列比对\clustalx1.83\PKA_all.seq loaded.

文件导出


ClustalX (1.83)
File Edit Alignment Trees Colors Quality Help
Multiple Alignment M
Complete Alignment
Output Guide Tree File:
l tools\多序列比对\clustalx1.83\PKA_all.dnd
Output Alignment Files:
Clustal:
ıl tools\多序列比对\clustalx1.8 \PKA_all.aln
ALIGN
CANCEL
1 ScTPK1
2 ScTPK2
3 ScTPK3
4 Cekin-1
5 DmPka-C1
6 DmPka-C2
7 DmPKA-C3
8 DmCG12069
9 HsPKACa
10 HsPKACb
11 HsPKACg
12 HsPRKX
13 HsPRKY
14 EcORF708
ruler
1.....10.....20.....30.....40.....50.....60
File F:\useful tools\多序列比对\clustalx1.83\PKA_all.seq loaded.

多序列比对：结果处理

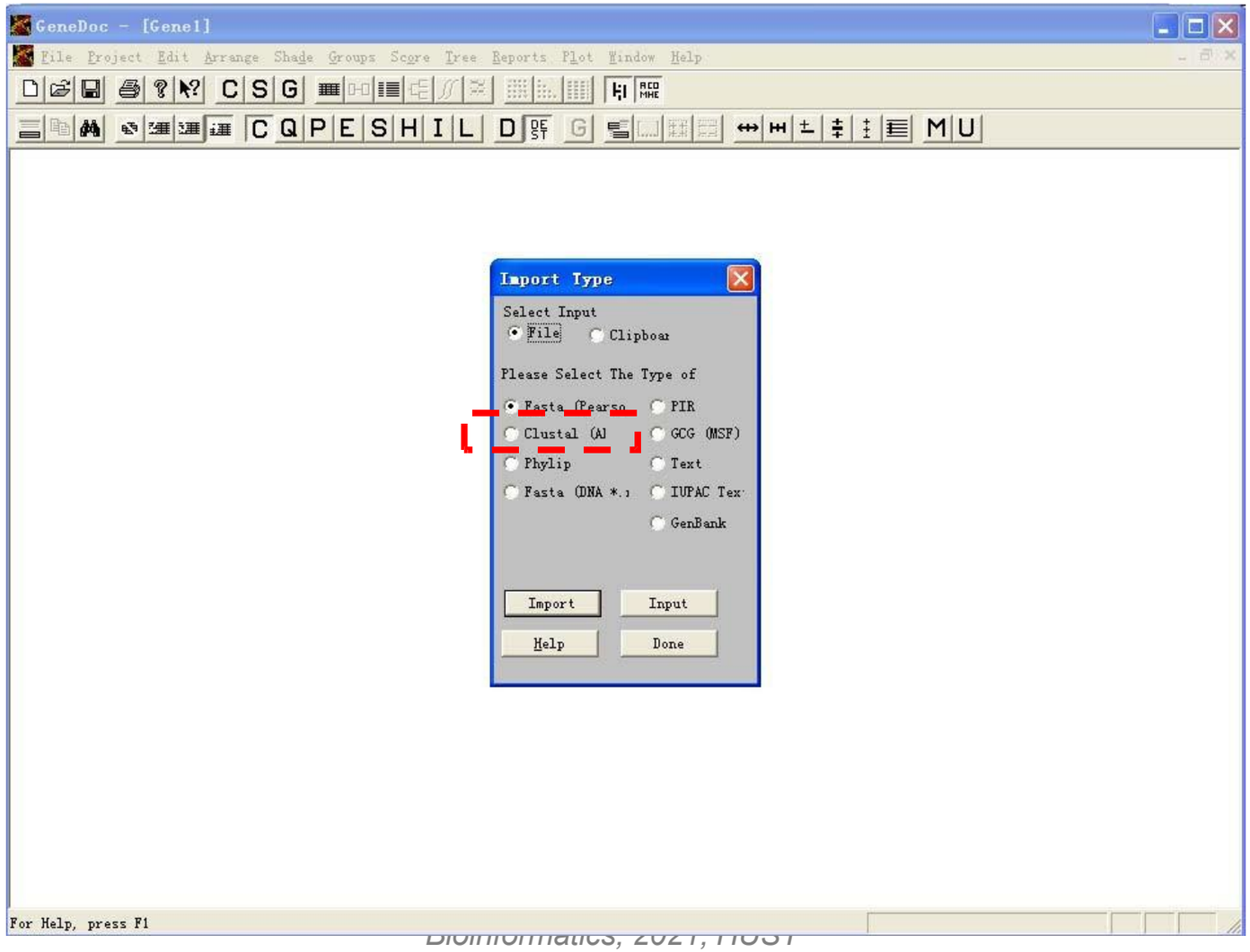
- GeneDoc, BioEdit等软件



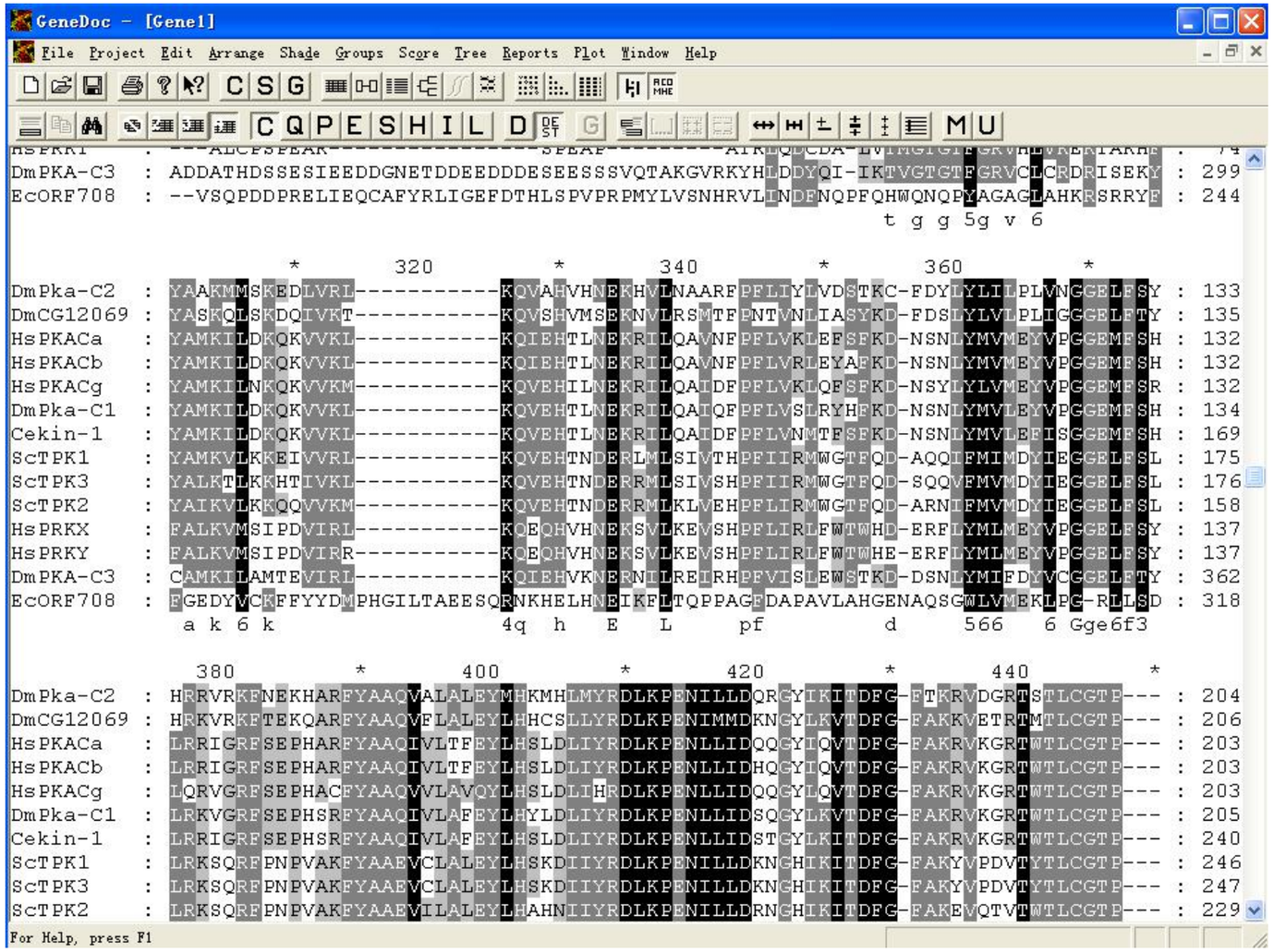
GeneDoc软件

导入.aln文件

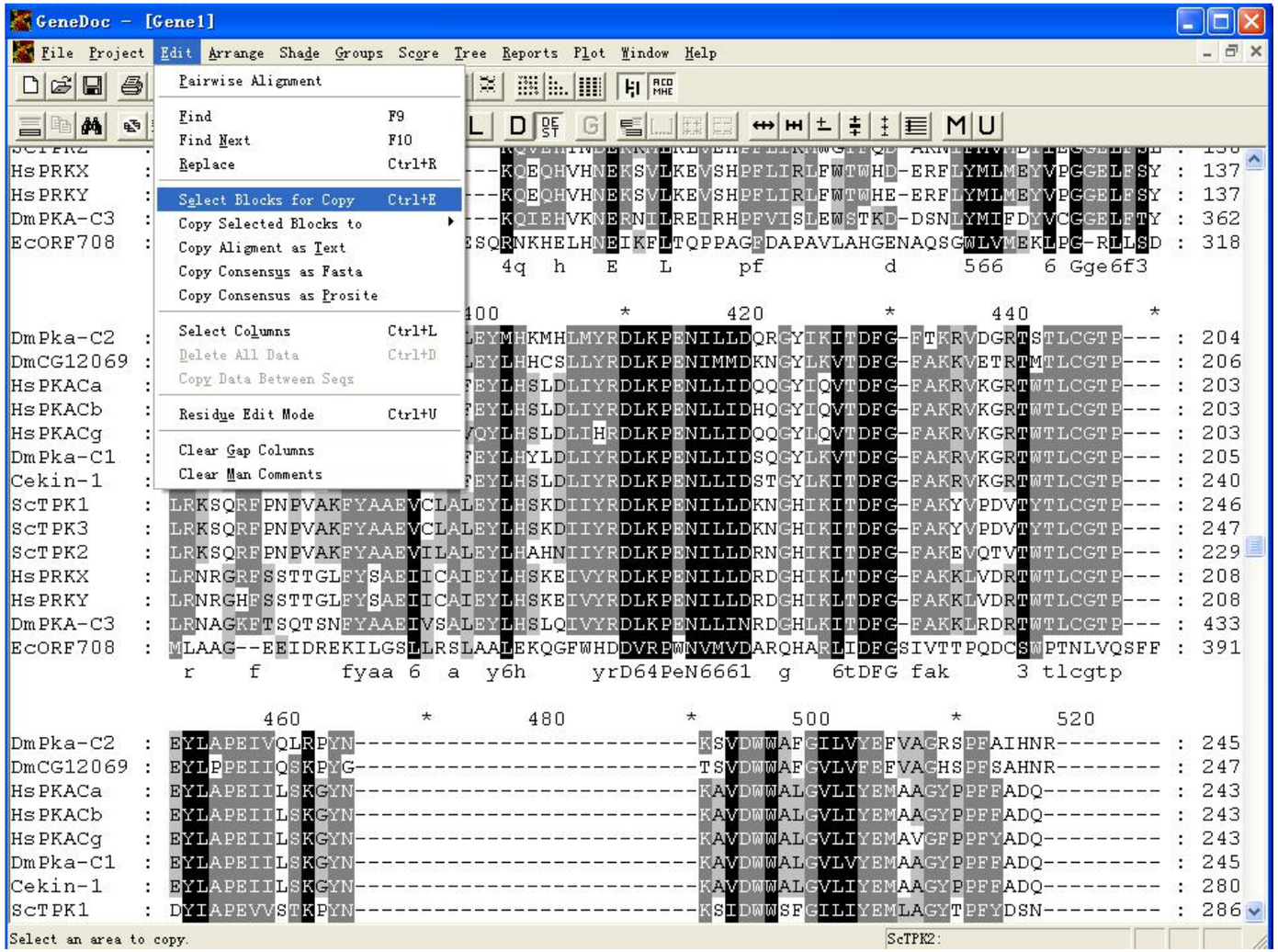
选择文件格式


GeneDoc - [Gene1]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
Import Type
Select Input
File
Clipboar
Please Select The Type of
Fasta (Pearso
Clustal (Al
Phylip
Fasta (DNA *.)
PIR
GCG (MSF)
Text
IUPAC Tex
GenBank
Import
Input
Help
Done
For Help, press F1

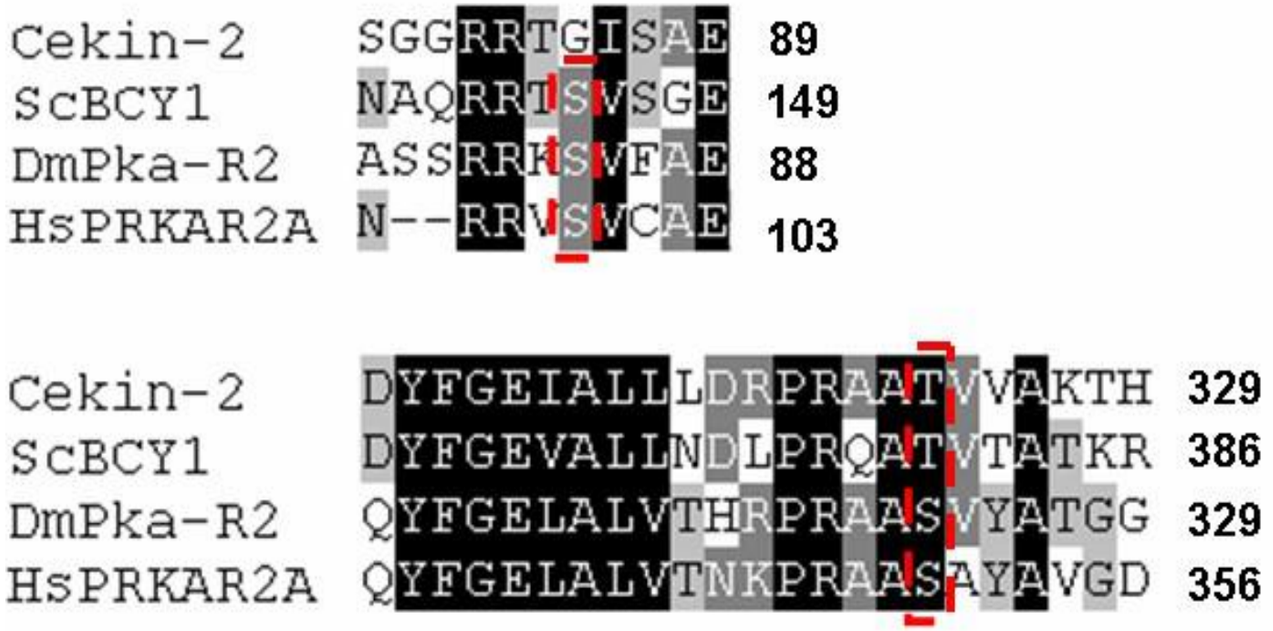
成功导入文件


GeneDoc - [Gene1]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
DmPKA-C3 : ADDATHDSSESIEEDDGNETDDEEDDDESEESSSVQTAKGVRKYHLDDYQI-IKTVGTGTFGRVCLCRDRISEKY : 299
EcORF708 : --VSQPDDPRELIEQCAFYRLIGEFDTHLSPVPRPMYLVSNHRVLLNDENQPFQHWQNQPYAGAGLAHKRSRRYF : 244
DmPka-C2 : YAAKMMSKEDLVRL----------KQVAHVHNEKHVLNAARFPFLIYLVDSTKC-FDYLYLILPLVNGGELFSY : 133
DmCG12069 : YASKQLSKDQIVKT----------KQVSHVMSEKNVLRSMTFPNTVNLIASYKD-FDSLYLVLPLIGGGELFTY : 135
HsPKACa : YAMKILDKQKVVKL----------KQIEHTLNEKRILQAVNFPFLVKLEFSFKD-NSNLYMVMEYVPGGEMFSH : 132
HsPKACb : YAMKILDKQKVVKL----------KQIEHTLNEKRILQAVNFPFLVRLEYAFKD-NSNLYMVMEYVPGGEMFSH : 132
HsPKACg : YAMKILNKQKVVKM----------KQVEHILNEKRILQAIDFPFLVKLQFSFKD-NSYLYLVMEYVPGGEMFSR : 132
DmPka-C1 : YAMKILDKQKVVKL----------KQVEHTLNEKRILQAIQFPFLVSLRYHFKD-NSNLYMVLEYVPGGEMFSH : 134
Cekin-1 : YAMKILDKQKVVKL----------KQVEHTLNEKRILQAIDFPFLVNMTFSFKD-NSNLYMVLEFISGGEMFSH : 169
ScTPK1 : YAMKVLKKEIVVRL----------KQVEHTNDERLMLSIVTHPFIIRMWGTFQD-AQQIFMIMDYIEGGELFSL : 175
ScTPK3 : YALKTLKKHTIVKL----------KQVEHTNDERRMLSIVSHPFIIRMWGTFQD-SQQVFMVMDYIEGGELFSL : 176
ScTPK2 : YAIKVLKKQQVVKM----------KQVEHTNDERRMLKLVEHPFLIRMWGTFQD-ARNIFMVMDYIEGGELFSL : 158
HsPRKX : FALKVMSIPDVIRL----------KQEQHVHNEKSVLKEVSHPFLIRLFWTWHD-ERFLYMLMEYVPGGELFSY : 137
HsPRKY : FALKVMSIPDVIRR----------KQEQHVHNEKSVLKEVSHPFLIRLFWTWHE-ERFLYMLMEYVPGGELFSY : 137
DmPKA-C3 : CAMKILAMTEVIRL----------KQIEHVKNERNILREIRHPFVISLEWSTKD-DSNLYMIFDYVCGGELFTY : 362
EcORF708 : EGEDYVCKFFYYDMPHGILTAEESQRNKHELHNEIKFLTQPPAGFDAPAVLAHGENAQSGWLVMEKLPG-RLLSD : 318
DmPka-C2 : HRRVRKFNEKHARFYAAQVALALEYMHKMHLMYRDLKPENILLDQRGYIKITDFG-FTKRVDGRTSTLCGTP--- : 204
DmCG12069 : HRKVRKFTEKQARFYAAQVFLALEYLHHCSLLYRDLKPENIMMDKNGYLKVTDFG-FAKKVETRTMTLCGTP--- : 206
HsPKACa : LRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFG-FAKRVKGRTWTLCGTP--- : 203
HsPKACb : LRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDHQGYIQVTDFG-FAKRVKGRTWTLCGTP--- : 203
HsPKACg : LQRVGRFSEPHACFYAAQVVLAVQYLHSLDLIHRDLKPENLLIDQQGYLQVTDFG-FAKRVKGRTWTLCGTP--- : 203
DmPka-C1 : LRKVGRFSEPHSRFYAAQIVLAFEYLHYLDLIYRDLKPENLLIDSQGYLKVTDFG-FAKRVKGRTWTLCGTP--- : 205
Cekin-1 : LRRIGRFSEPHSRFYAAQIVLAFEYLHSLDLIYRDLKPENLLIDSTGYLKITDFG-FAKRVKGRTWTLCGTP--- : 240
ScTPK1 : LRKSQRFPNPVAKFYAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFG-FAKYVPDVTYTLCGTP--- : 246
ScTPK3 : LRKSQRFPNPVAKFYAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFG-FAKYVPDVTYTLCGTP--- : 247
ScTPK2 : LRKSQRFPNPVAKFYAAEVILALEYLHAHNIIYRDLKPENILLDRNGHIKITDFG-FAKEVQTVTWTLCGTP--- : 229
For Help, press F1

选择需要拷贝的行

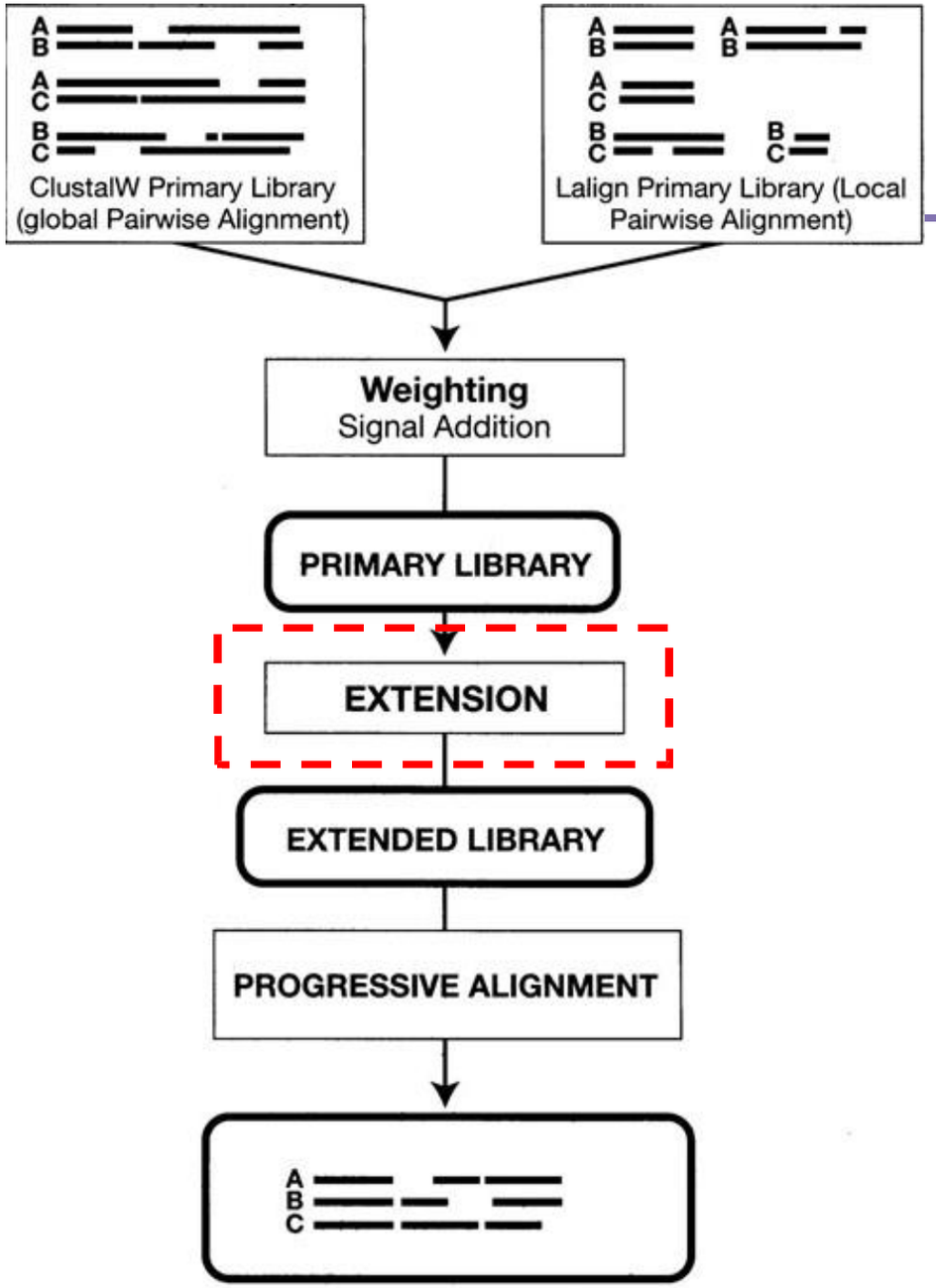

GeneDoc - [Gene1]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
Pairwise Alignment
Find F9
Find Next F10
Replace Ctrl+R
Select Blocks for Copy Ctrl+E
Copy Selected Blocks to
Copy Aligment as Text
Copy Consensus as Fasta
Copy Consensus as Prosite
Select Columns Ctrl+L
Delete All Data Ctrl+D
Copy Data Between Seqs
Residue Edit Mode Ctrl+U
Clear Gap Columns
Clear Man Comments
Select an area to copy.
ScTPK2:

比对结果的美化和后处理



T-Coffee

- 采用Clustal程序计算两两序列之间的全局 最优比对结果
- 采用LALIGN程序计算两两序列之间的局 部最优比对的结果
  - https://www.ebi.ac.uk/Tools/psa/lalign/
- 设计加权系统，综合考虑上述两两部分结 果，构建指导库
- 采用渐进算法，得到最终的结果



同时进行全局和局部的双序列比对

对以上打分的结果设计权重系统，找到序列中最保守的部分

渐进比对，基于上述计算得到的指导库（**primary library**）

渐进方法存在的问题

- 通常使用启发式策略选择初始序列
  - 最终结果可能受初始选定的序列的影响
- 距离最近的，有两组序列**AB**和**CD**，哪组最 先比对？两种方案：
  - **A.** 分别、同时比对。究竟应以**AB**为准，加入 **CD**，然后再加上其他序列，还是以**CD**为准？ 结果可能出入很大
  - **B.** 随机挑选一组作为基准
- 当序列之间差异较大时，上述问题更加明显

- 三条序列：

```
Seq1: ARKCV
Seq2: ARCV
Seq3: AKCV
```

- 若Seq1, 2先比对，再加入Seq3：

```
A R K C V
A R - C V
A-KCV
```

- Seq1, 3先比对，再加入Seq2:

```
A R K C V
A - R C V
A-KCV
```

- Seq2, 3先比对，再加入Seq1:

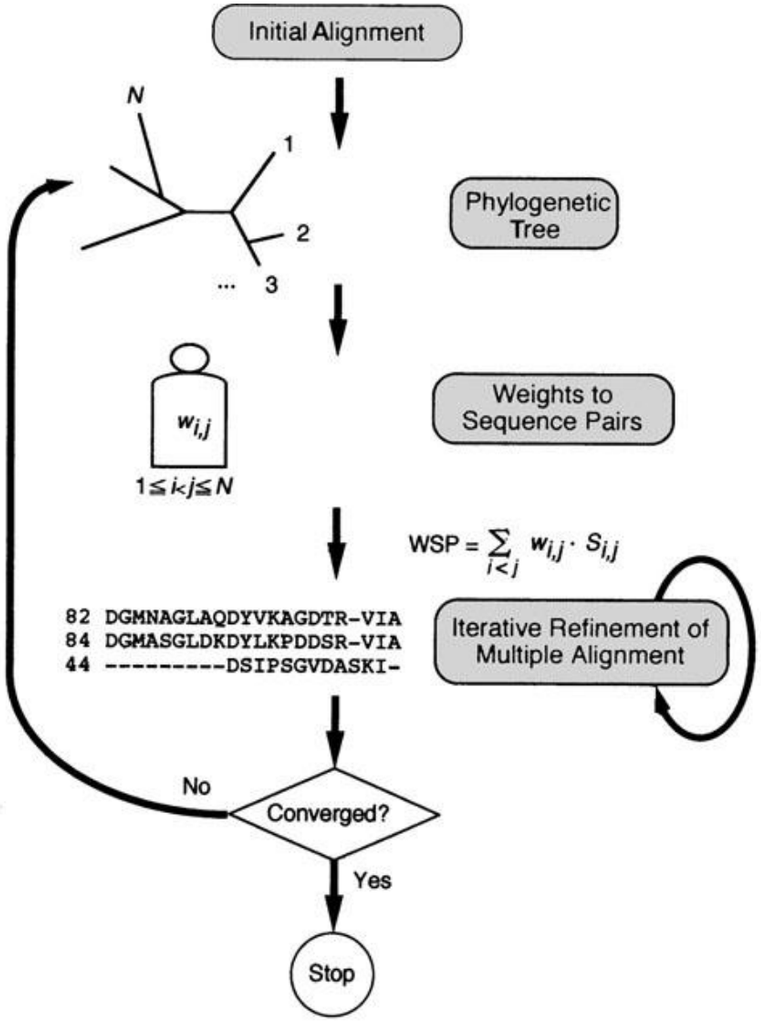
```
A R K C V
A R - C V
AK-CV
```

迭代算法

迭代算法

- 部分解决渐进算法存在的问题，主要是ClustalW/X存在的问题
- PRRP/PRRN
  - https://www.genome.jp/tools-bin/prrn
- DIALIGN
  - http://dialign.gobics.de/

PRRP/PRRN



1. 先用“渐进”算法进行多序列比对

2. 基于多序列比对的结果构建进化树

3. 重新计算序列之间的进化距离，再用“渐进”算法进行多序列比对

4. 重复上述步骤，直到结果不再发生改变为止

DIALIGN

- 对所有序列进行两两之间的局部最优比对
- 找到所有能够匹配的部分**M1**；将重叠的、前后一致的（**consistency**）匹配部分连接起来为**M2**
- 将剩下的未比对的序列重新比对，再发现能够匹配的部分，构成新**M1**，将一致的部分构成**M2**
- 重复上述步骤，直到结果收敛

有向图算法

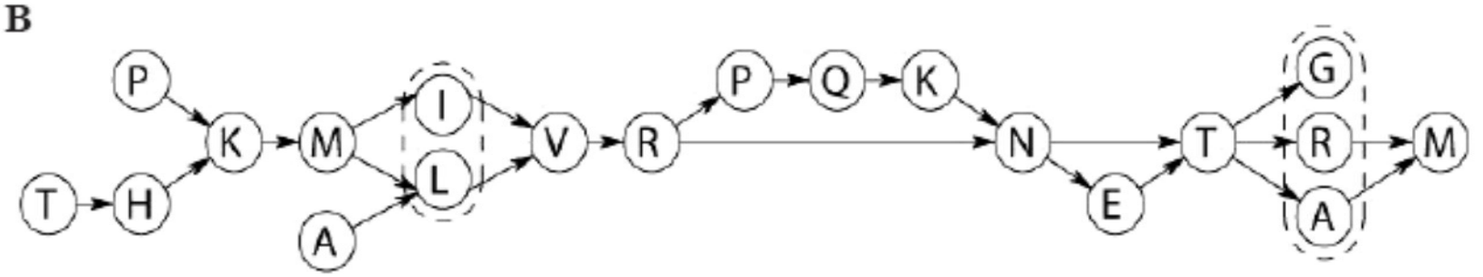
部分有向图算法：POA

- https://simpsonlab.github.io/2015/05/01/understanding-poa/
- https://sourceforge.net/projects/poamsa/

A

```
.  .  P  K  M  .  I  V  R  P  Q  K  N  E  T  G  .
.  .  .  .  .  A  L  V  R  P  Q  K  N  .  T  R  M
T  H  .  K  M  .  L  V  R  .  .  .  N  E  T  A  M
```

B



隐马尔可夫以及整合算法

隐马尔科夫模型: ProbCons

- http://probcons.stanford.edu/
- 主要改进：
- 所有序列的两两比对，通过profile HMM的方法进行双序列比对
- 将渐进算法与迭代算法整合

整合算法MUSCLE

- 算法分为三个部分，每个部分相对独立
- **Draft progressive:**
  - **(1)** 对两条序列，计算距离采用***k*-mer**的思想；
  - **(2)** 用**UPGMA**算法构建引导树
  - **(3)** 使用渐进算法进行多序列比对
- 优点：两条序列之间的距离不采用动态规划算法进行比对，节省时间

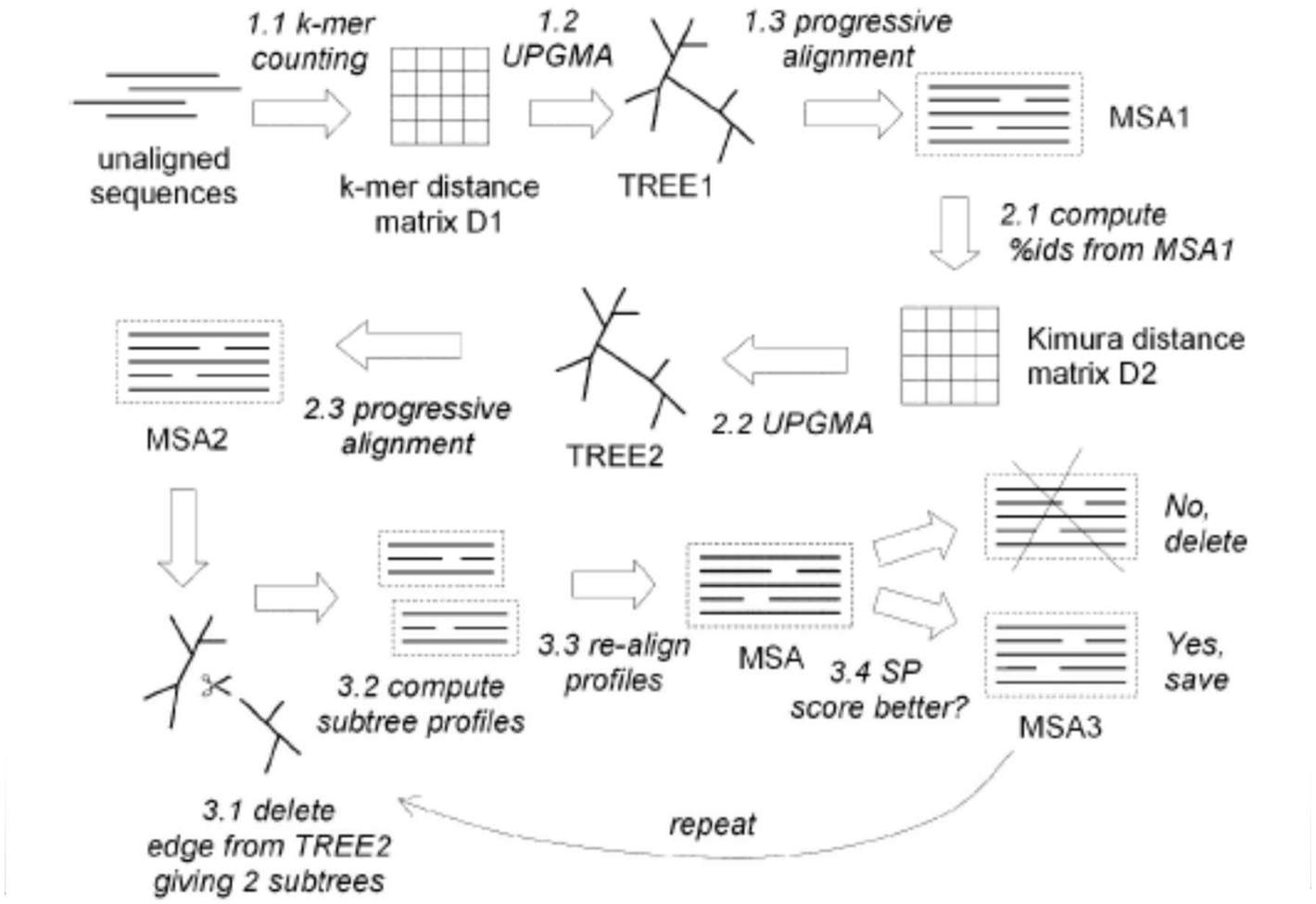
MUSCLE (2)

- **Improved progressive:**
  - **(1) 基于*k*-mer得到的树可能会产生次优结果， 因此，采用Kimura距离的方法对*k*-mer产生 的树重新计算距离矩阵**
  - **(2) 重新用UPGMA构建进化树**
  - **(3) 使用渐进算法进行多序列比对**

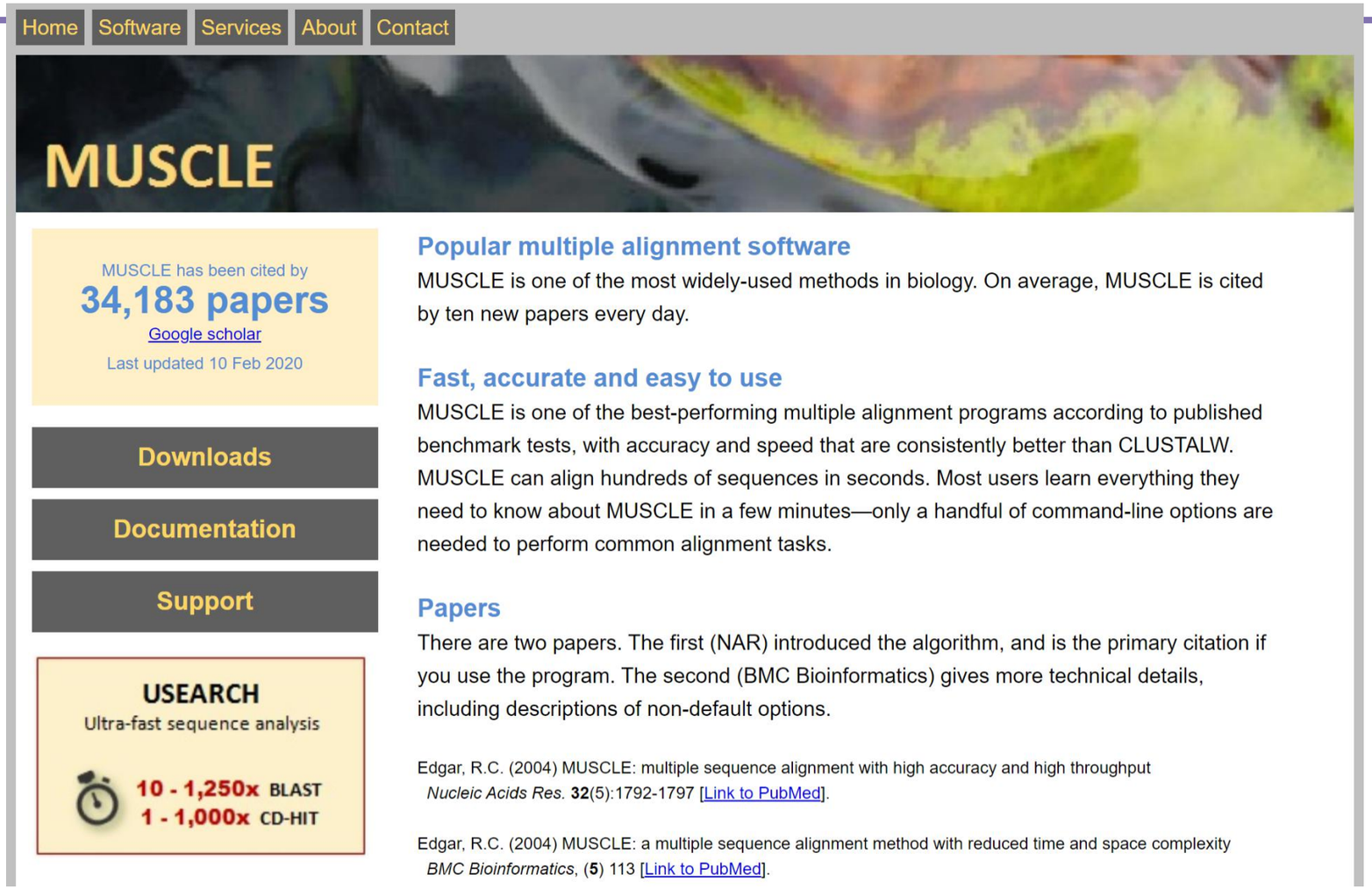
MUSCLE (3)

- **Refinement:**
  - (1) 随机从进化树上挑出一条边，删除
  - (2) 得到两组树，对每组树，计算profile
  - (3) 将两组profile进行比对
  - (4) 如果最终得分提高，保留结果，否则丢弃

MUSCLE的算法流程

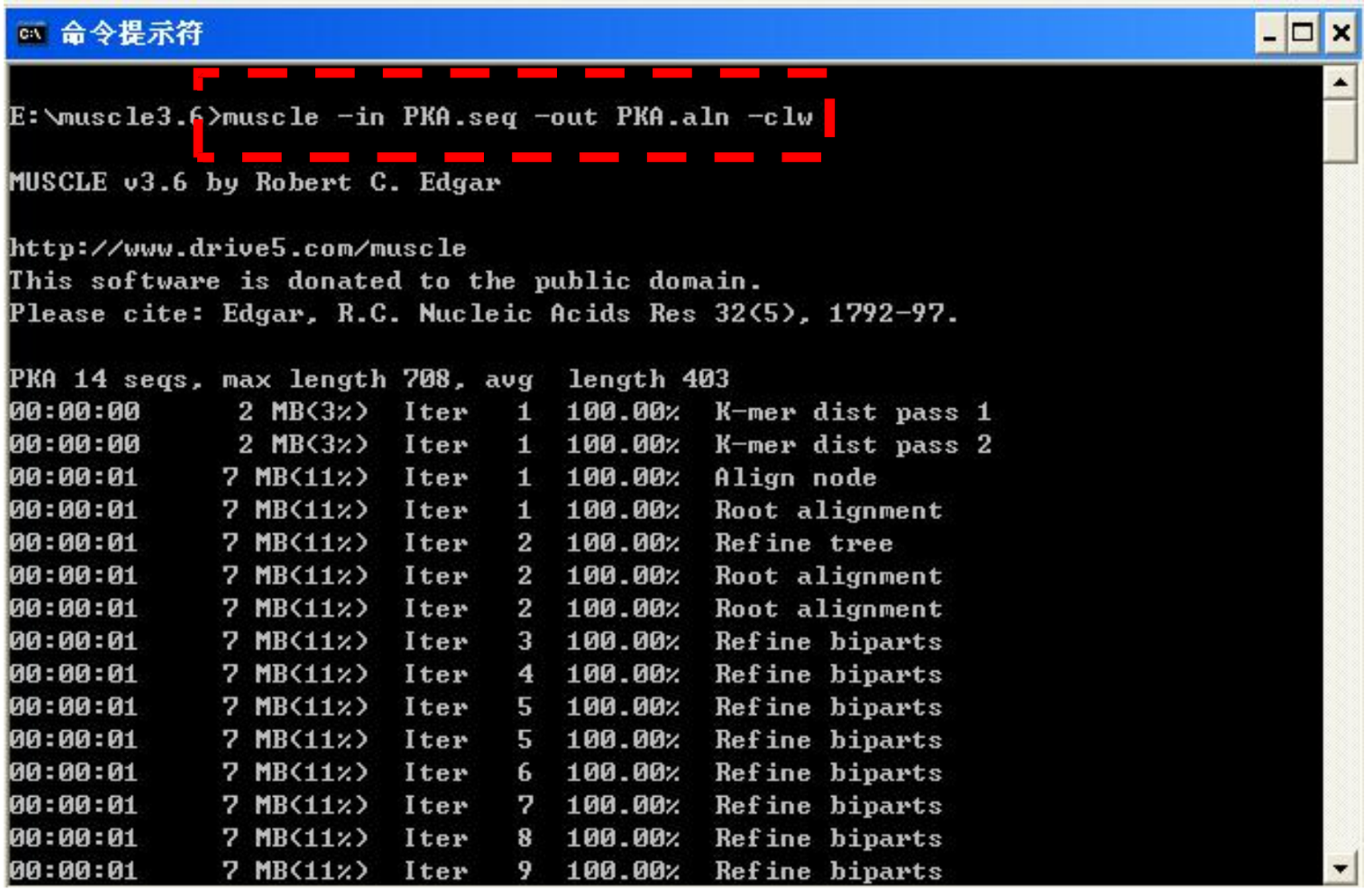


MUSCLE: 使用指南



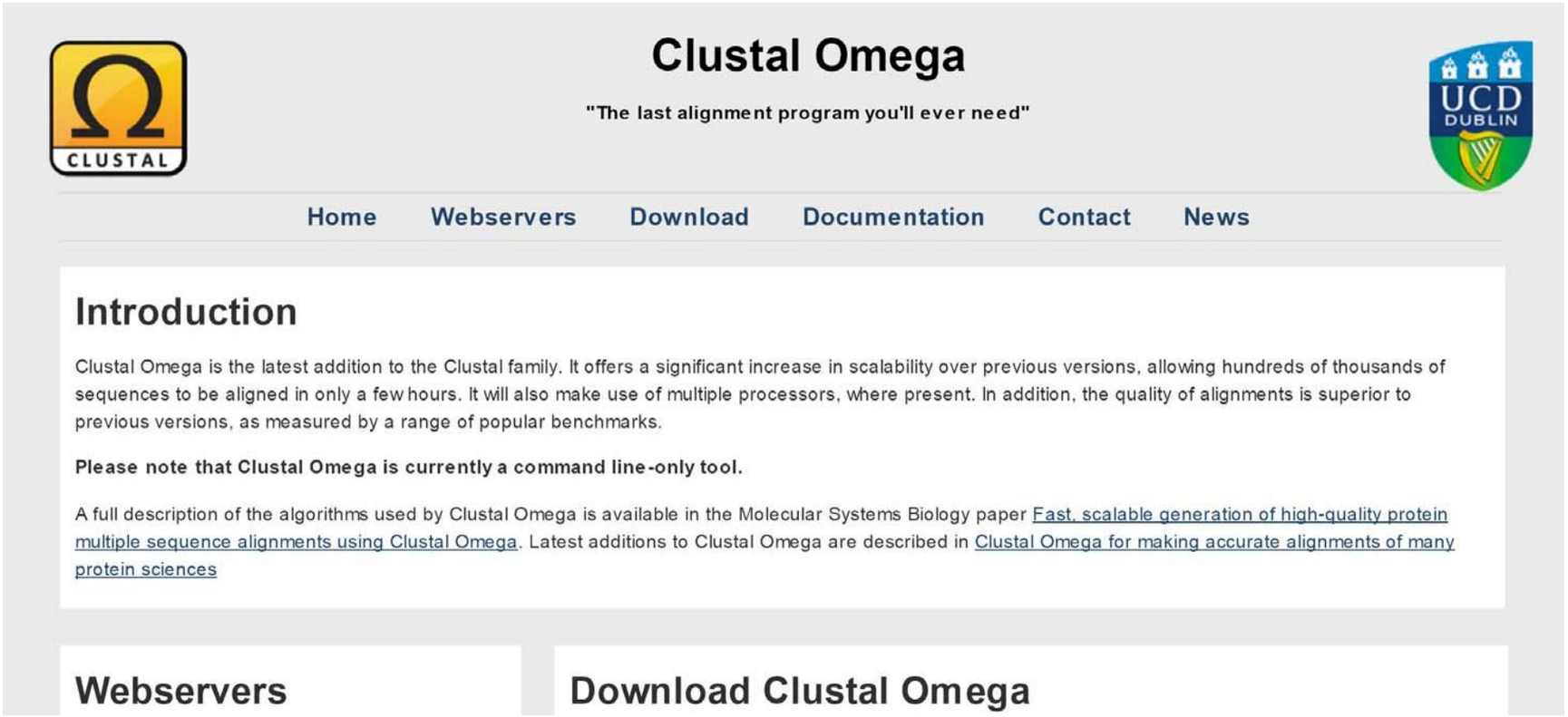
http://www.drive5.com/muscle/

MUSCLE：使用说明


命令提示符
E:\muscle3.6>muscle -in PKA.seq -out PKA.aln -clw
MUSCLE v3.6 by Robert C. Edgar
http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.
PKA 14 seqs, max length 708, avg length 403
00:00:00 2 MB(3%) Iter 1 100.00% K-mer dist pass 1
00:00:00 2 MB(3%) Iter 1 100.00% K-mer dist pass 2
00:00:01 7 MB(11%) Iter 1 100.00% Align node
00:00:01 7 MB(11%) Iter 1 100.00% Root alignment
00:00:01 7 MB(11%) Iter 2 100.00% Refine tree
00:00:01 7 MB(11%) Iter 2 100.00% Root alignment
00:00:01 7 MB(11%) Iter 2 100.00% Root alignment
00:00:01 7 MB(11%) Iter 3 100.00% Refine biparts
00:00:01 7 MB(11%) Iter 4 100.00% Refine biparts
00:00:01 7 MB(11%) Iter 5 100.00% Refine biparts
00:00:01 7 MB(11%) Iter 5 100.00% Refine biparts
00:00:01 7 MB(11%) Iter 6 100.00% Refine biparts
00:00:01 7 MB(11%) Iter 7 100.00% Refine biparts
00:00:01 7 MB(11%) Iter 8 100.00% Refine biparts
00:00:01 7 MB(11%) Iter 9 100.00% Refine biparts

Clustal Omega

- 算法原理类似**MUSCLE**
  - http://www.clustal.org/omega/
  - https://www.ebi.ac.uk/Tools/msa/clustalo/



其他多序列比对工具

- MAFFT：渐进 & 迭代
  - https://www.genome.jp/tools-bin/mafft
- T-Coffee (M-coffee)：整合其他工具的输出结果
  - http://tcoffee.crg.cat/apps/tcoffee/all.html
- Multiple Sequence Alignment
  - https://www.ebi.ac.uk/Tools/msa/

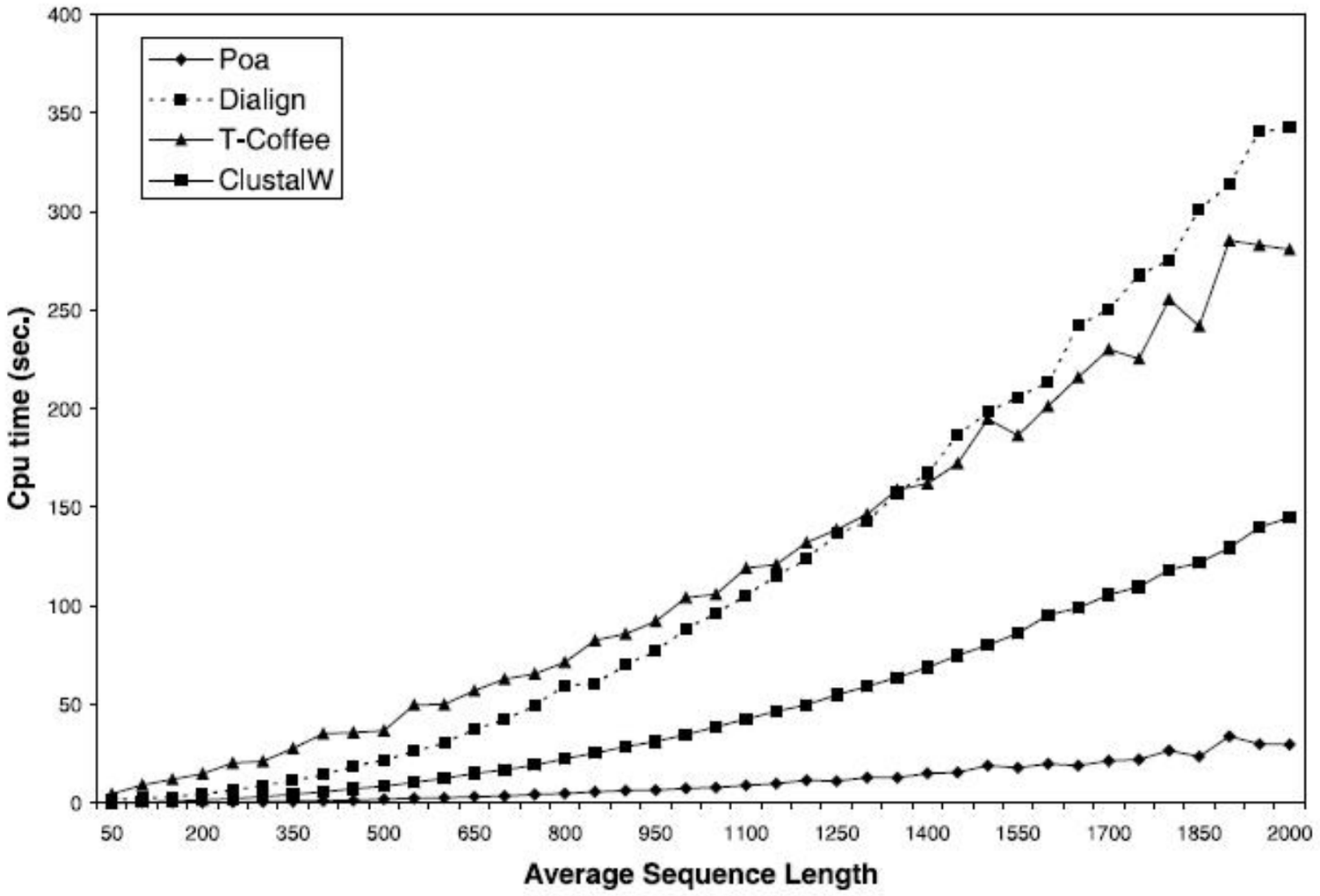
多序列比对：性能检验

- **BAliBASE**：基于蛋白质三级结构，将同一家族的蛋白质序列进行多序列比较
- 多序列比对工具的性能检验：能否与**BAliBASE**中的比对结果相吻合
  - http://www.lbgi.fr/balibase/

性能比较

- ClustalW/X: 最经典、最被广泛接受的工具
- MUSCLE: 最流行的多序列比对工具
- Clustal Omega: 类似MUSCLE
- T-Coffee：序列相似性高时最准确
- DIALIGN: 序列相似性低时较准确
- POA：性能接近T-Coffee和DIALIGN，速度最快（目前主要用于三代测序数据分析）

运算时间比较


Poa
Dialign
T-Coffee
ClustalW
Cpu time (sec.)
Average Sequence Length
0
50
100
150
200
250
300
350
400
50
200
350
500
650
800
950
1100
1250
1400
1550
1700
1850
2000