



生物信息学

第七章 多序列比对

多序列比对



- 序列保守 -> 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
- 可构建HMM模型，搜索更多的同源序列
- 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）

DmPka-C2	100	ARFPFLIYLVDSTKCFYLYLILPLVNGGELFSYHRRVRKFNKHFARFYAAQVALALEYMHKMHLMYRD	168
DmCG12069	102	MTFPNTVNLIASYKDFDSLVLPLIGGGELFTYHRKVRKFTEKQARFYAAQVF LALDY LHHCSLLYRD	170
ScTPK2	125	VHHPFLIRMWGTFQDARNIFMVM DYIEGGELFSLLRKSQRFPNPVAKFYAAEVI LALDY LHSHTLIYRD	193
ScTPK1	142	VTHPFLIRMWGTFQDAQQIFMIMDYIEGGELFSLLRKSQRFPNPVAKFYAAEVC LALDY LHSKDLIYRD	210
ScTPK3	143	VSHPFLIRMWGTFQDSQQVFMIMDYIEGGELFSLLRKSQRFPNPVAKFYAAEVC LALDY LHSKDLIYRD	211
Cekin-1	136	LDFFFLVMMTFSFKDMSNLYMVLEFISGGEMFSLHRRVGRFSEPHSRFYAAQIVLAFDY LHS LGLIYRD	204
DmPka-C1	101	LDFFFLVSLRYHFVKDMSNLYMVLEYPGGEMFSLHRRVGRFSEPHSRFYAAQIVLAFDY LHY LGLIYRD	169
HsPKACg	99	LDFFFLVKLDFSEFKDMSYLYVMMEYPGGEMFSRLQVGRFSEPHACFYAAQVVLAVQY LHS LGLIHRD	167
HsPKACa	99	VNFFFLVKLDFSEFKDMSNLYMVMEYPGGEMFSLHRRVGRFSEPHARFYAAQIVLTFDY LHS LGLIYRD	167
HsPKACb	99	VNFFFLVRLDEYAFKDMSNLYMVMEYPGGEMFSLHRRVGRFSEPHARFYAAQIVLTFDY LHS LGLIYRD	167
DmPKA-C3	329	LRHHPFVISLEWSTKDESNLYMIFDYVCGGELFTYLRNAGKFTSQTSNFYAAEIVSALDY LHS LQIVYRD	397
HsPRKX	104	VSHPFLIRLFWTWHERFLYMLMEYPGGELFSYLRNRGRFSSTTGLFYSAEIIICADY LHSKEIVYRD	172
HsPRKY	104	VSHPFLIRLFWTWHERFLYMLMEYPGGELFSYLRNRGHFSSTTGLFYSAEIIICADY LHSKEIVYRD	172

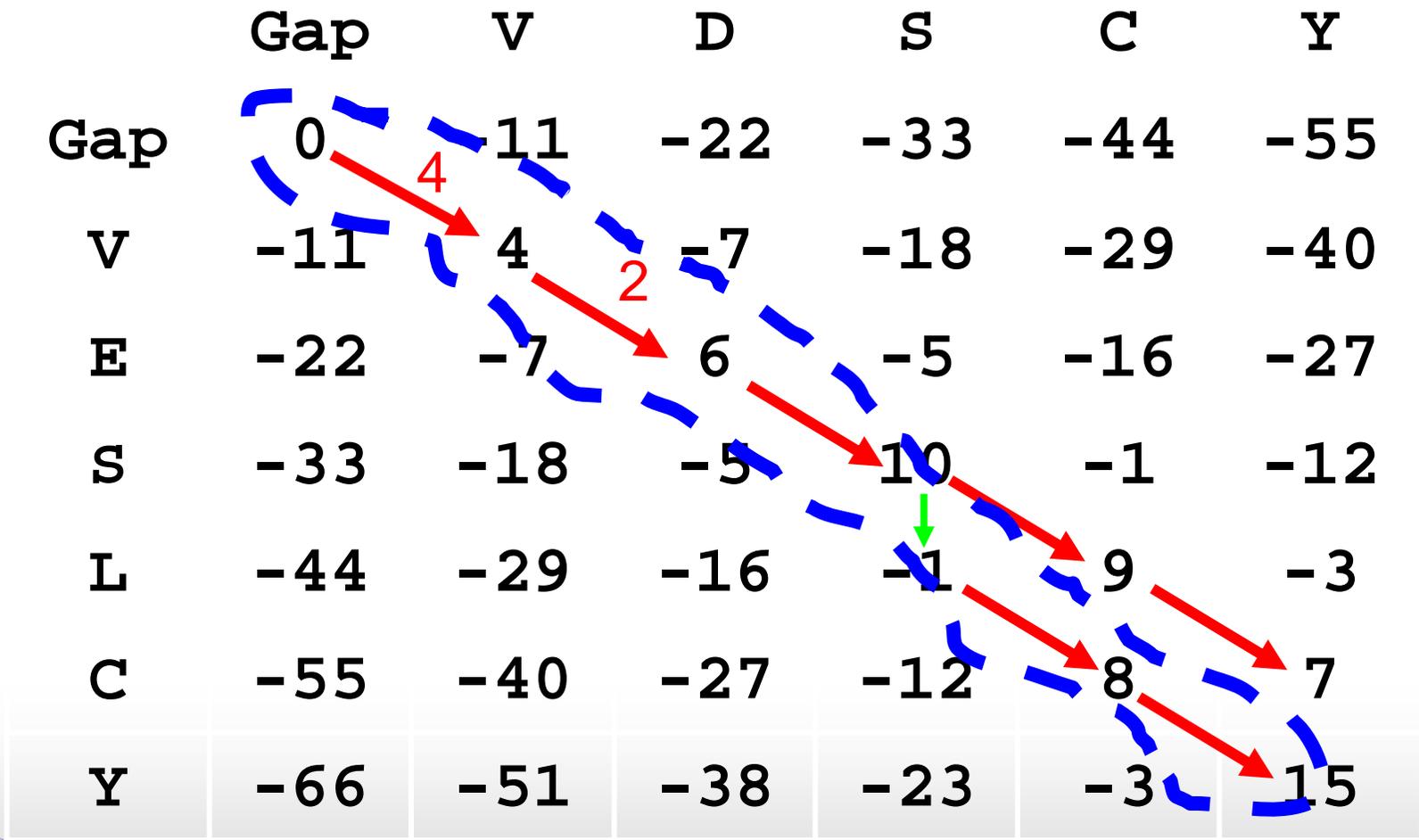
Made by GENEDOC

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



双序列比对的时间复杂度

时间复杂度: $O(n^2)$



多序列比对：最优算法



ARDFSHGLENKLLGCD SMRWE
GRDYKMALLEQWILGCD-MRWD
SRDW--ALIEDCMV-CNEFRWD

多项式时间复杂度： $\leq O(n^3)$

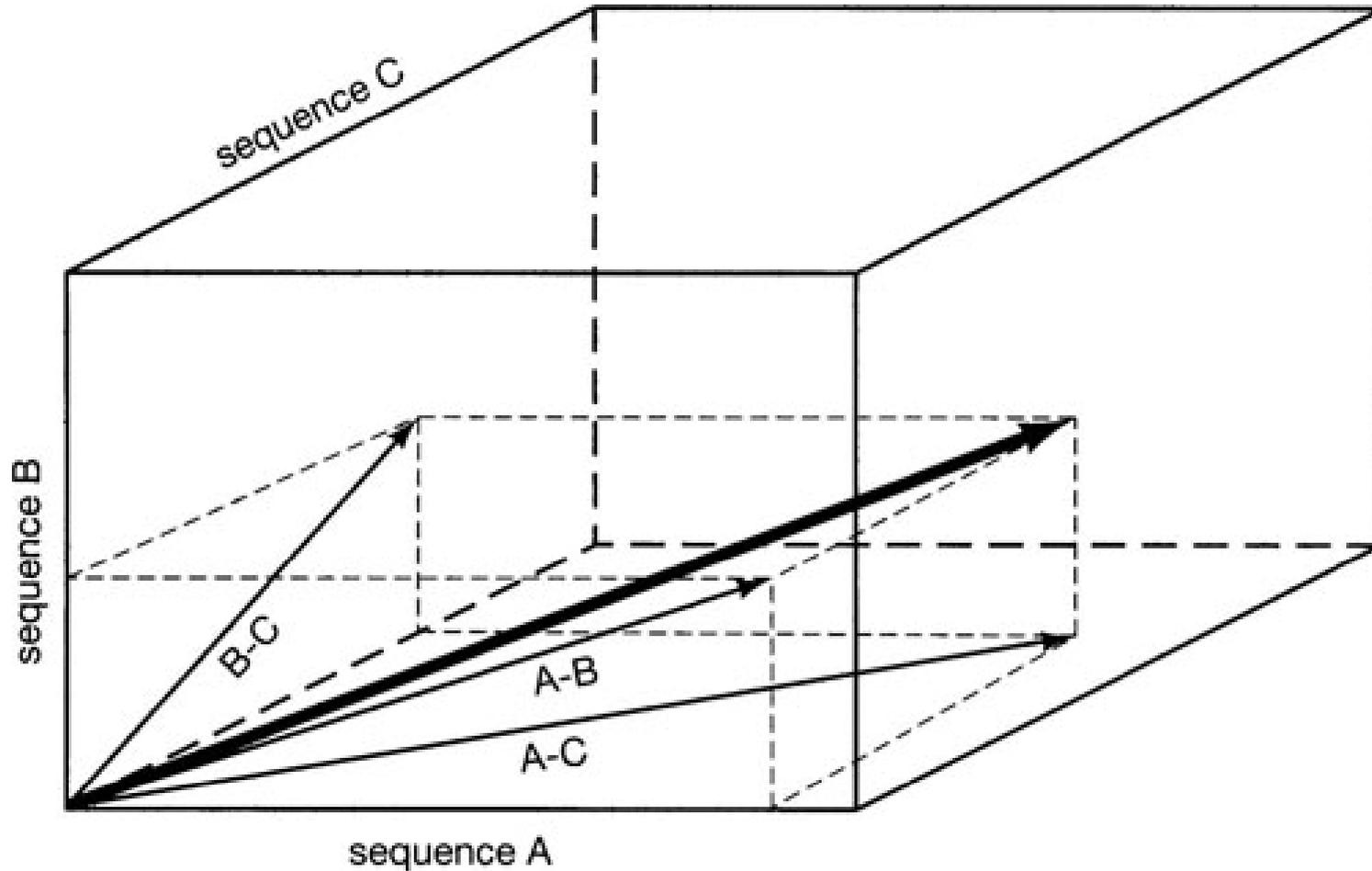
三条序列：时间复杂度： $O(lmn) = O(n^3)$

四条序列：时间复杂度： $O(n^4)$ ，非多项式时间！

...

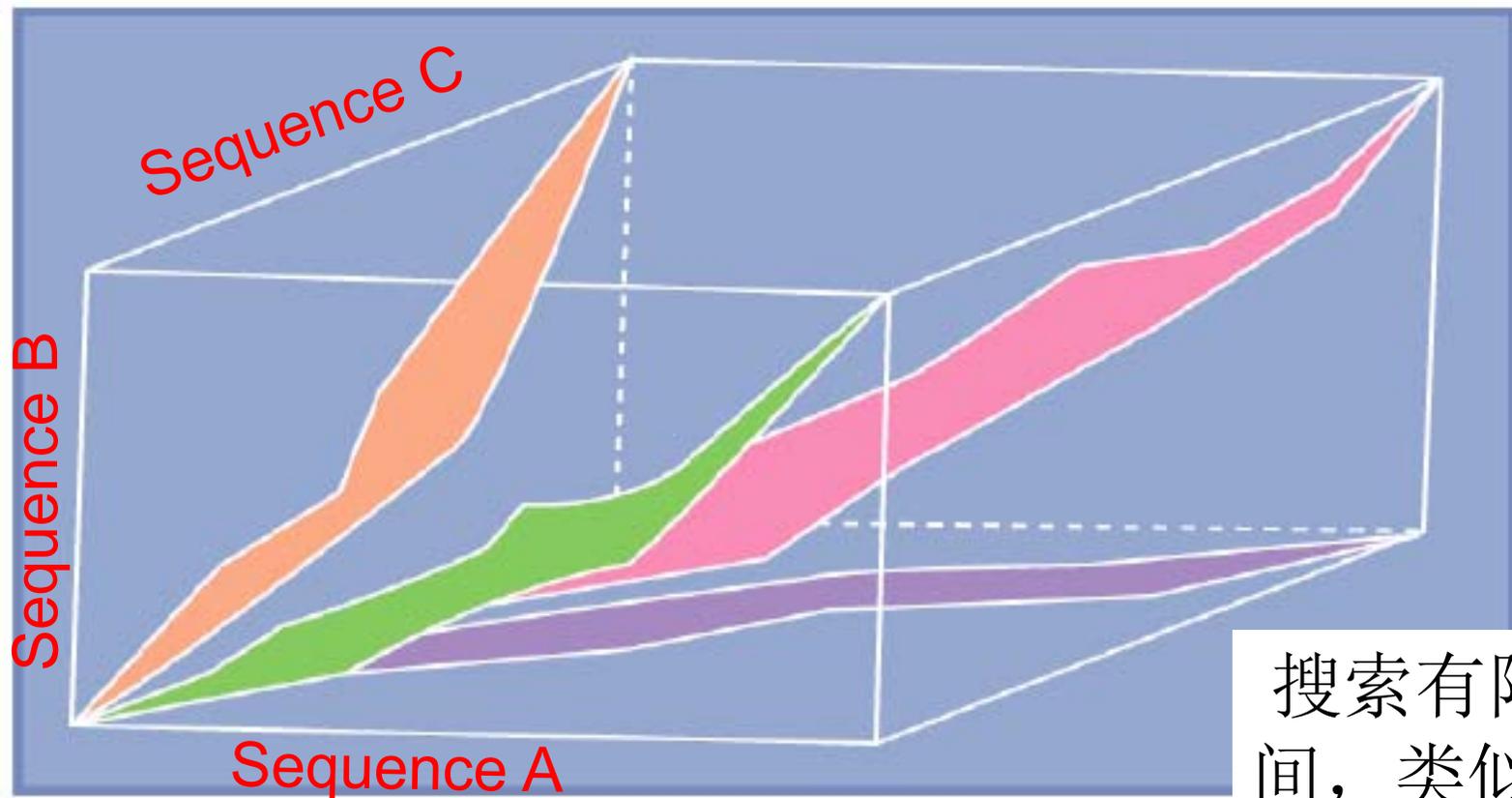
m条序列：时间复杂度： $O(n^m)$ ，指数时间！

动态规划算法：全空间



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

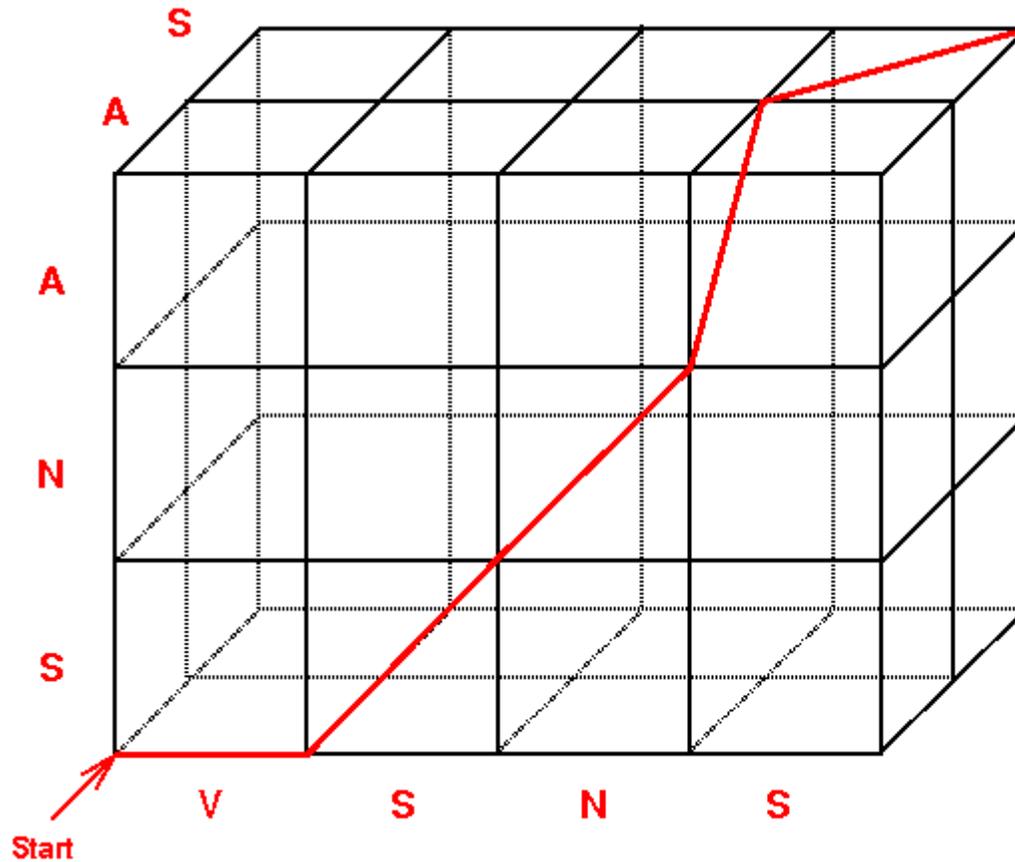
动态规划算法：优化算法



搜索有限空间，类似于
BLAST算法

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

动态规划算法：hyperlattice



V S N - S
- S N A -
- - - A S

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

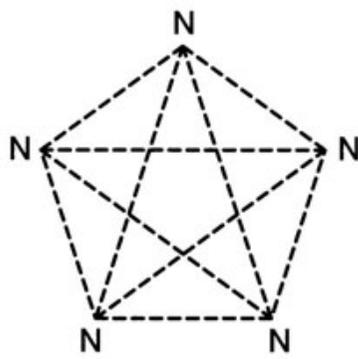
```
_ S N A _
_ _ _ A S
```

非最优的双序列比对

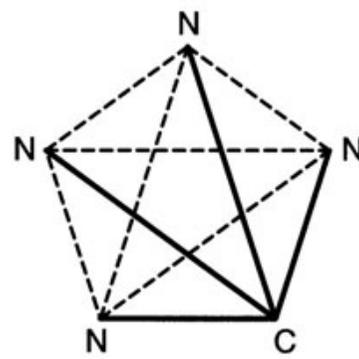


MSA: 多序列比对的打分策略

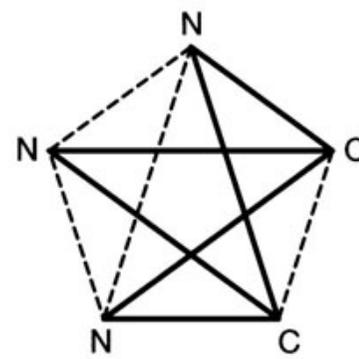
Sequence	Column A	Column B	Column C
1N.....N.....N.....
2N.....N.....N.....
3N.....N.....N.....
4N.....N.....C.....
5N.....C.....C.....



Column A



Column B



Column C

No. of N-N matched pairs (each scores 6):

10	6	4
----	---	---

No. of N-C matched pairs (each scores -3):

0	4	6
---	---	---

BLOSUM62 score :

60	24	6
----	----	---

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

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 1994

Method (Developer)	Algorithm	Matrix*	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. Higgins)	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

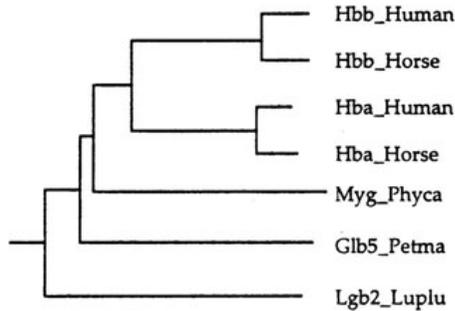


- 将所有序列两两比对，计算进化距离（差异）矩阵
- 使用邻接法（neighbor-joining）构建指导树（guide tree）
- 将进化距离最近的两条序列用全局动态规划算法进行比对
- “渐进”地加上其他序列

Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

Pairwise alignment:
Calculate distance matrix

两两比对，构建距离矩阵



Rooted neighbor-joining tree (guide tree)

指导树的构建

```

-----VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYFWTQRFESFGDLST
-----VQLSGEKAAVLALWDKVN--EEEVGGGEALGRLLVVYFWTQRFESFGDLSN
-----VLSPADKTNVKAAWGKVGAGHAGEYGAEALERMFLSFHTTKTYFPHFDLS--
-----VLSAADKTNVKAWSKVGGHAGEYGAEALERMFSGFHTTKTYFPHFDLS--
-----VLSSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHHETLEKFDKFKHLKT
PIVDVTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTFAAQEFFPKFKGLTT
-----GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAFAAKDLFSFLKGTSE
  *      *      *      *

```

Progressive alignment:
Align following the guide tree

渐进比对

```

PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGTFAATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVND-----DMPNALSALSDLHAHKLRLVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLD-----DLPGALSNSLDLHAHKLRLVDPVNFKL
EAEMKASEDLKKHGVTVLTAALGAILKKGK-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKPADVRRWHAERIINAVNDAVASMDDT--EKMSMKLRDLGSKHAKSFQVDPQYFKV
VP--QNNPELOAHAGKVFELVYEAATLQVTVGVVVTATLKNLGSVHVSKG-VADAHFPV
  *      *      *      *

```

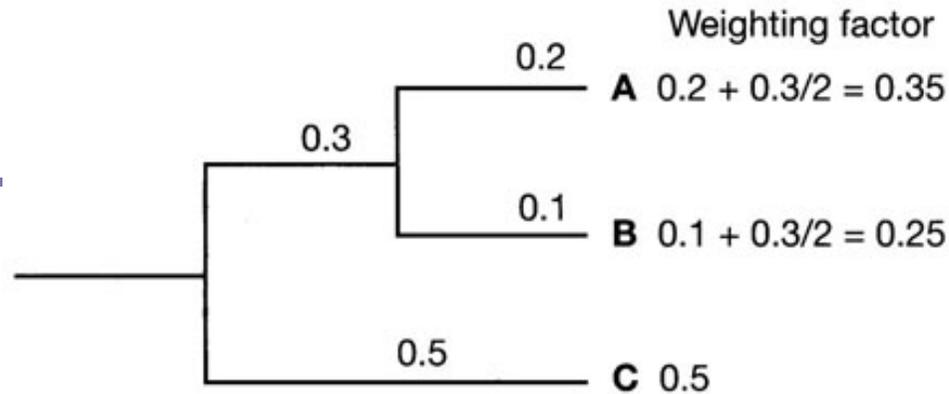
```

LGNVLVCVLAHHEFGKEFTPPVQAAYQKVVAGVANALAHKYH-----
LGNVLVVVLAHHEFGKDFTPPELQASVYQKVVAGVANALAHKYH-----
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR-----
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTISKYR-----
ISEAIIHVLHSRHPGDFGADAQCAMNKALELFRKDIAAKYKELGYQG
LAAVIADTVAAG-----DAGFEKLSMICILLRSAY-----
VKEAAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKEMNDAA---

```



A. Calculation of sequence weights



← 每条序列的权值

B. Use of sequence weights

Column in alignment 1

Sequence A (weight a)K.....

Sequence B (weight b)I.....

Column in alignment 2

Sequence C (weight c)L.....

Sequence D (weight d)V.....

ClustalW的打分原则

Score for matching these two columns in an msa =

$$[a \times c \times \text{score}(K,L) + a \times d \times \text{score}(K,V) + b \times c \times \text{score}(I,L) + b \times d \times \text{score}(I,V)] / 4$$

Score: BLOSUM62的分数

HUST

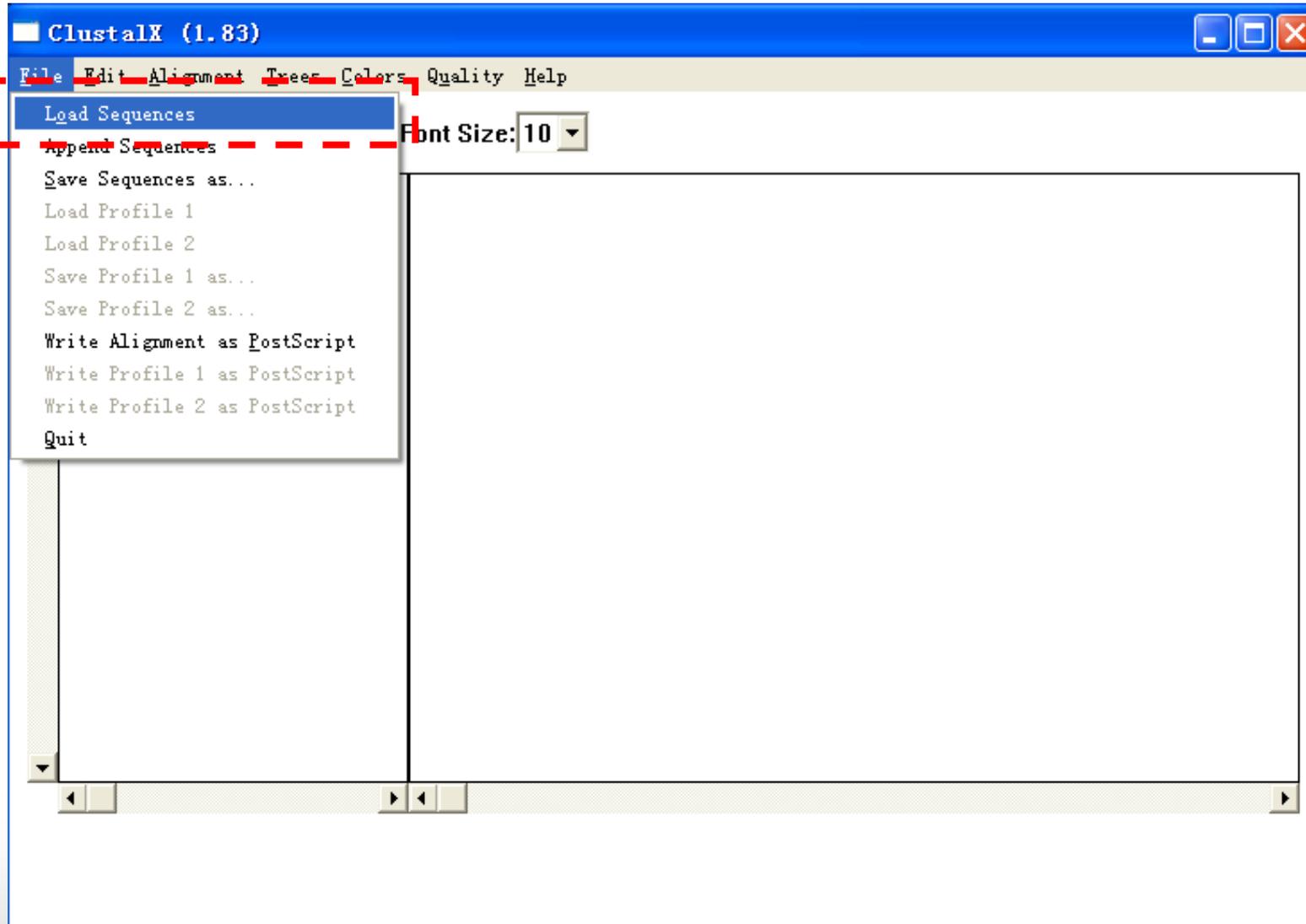
ClustalX: 使用指南



□ FASTA序列格式，多序列

```
>ScTPK1
MSTEEQNGGGQKSLDDRQGEESQKGETSERETTATESGNEKSKSVEKEGGETQEKPKQPHV
TYYNEEQYKQFIAQARVTSGKYSLQDFQILRTLGTGSFGRVHLIRSRHNGRYYAMKVLKK
EIVVRLKQVEHTNDERLMLSIVTHPFIIRMWGTQDAQQIFMIMDYIEGGELFSLLRKSQ
RFPNPVAKFYAAEVC LALEYLHSDKIIYRDLKPENILLDKNGHIKITDFGFAKYVPDVTY
TLCGTPDYIAPEVVSTKPYNKSIDWWSFGILIEMLAGYTPFYDSNTMKTYEKILNAELR
FPPFFNEDVKDLLSRLITRDLSQLGNLQNGTEDVKNHPWFKEVVWEKLLSRNIETPYEP
PIQQGQGDTSQFDKYPEEDINYGVGEDPYADLFRDF
>ScTPK2
MEFVAERAQPVGQTIQQQNVNTYGGVLPQPHDLQQRQQQQQQRQHQQLLTSQLPQKSLV
SKGKYTLHDFQIMRTLGTGSFGRVHLVRSVHNGRYAIKVLKKQQVVKMKQVEHTNDERR
MLKLVEHPFLIRMWGTQDARNIFMVMDYIEGGELFSLLRKSQRFPNPVAKFYAAEVI LA
LEYLHAHNIIYRDLKPENILLDRNGHIKITDFGFAKEVQTVTWTLCGTPDYIAPEVITTK
PYNKSVDWWSLGVLIYEMLAGYTPFYDTPMKTYEKILQGKVYPPYFHPDVVDLLSKLI
TADLTRRIGNLQSGSRDIKAHPWFSEVVWERLLAKDIETPYEPPITSGIGDTSLFDQYPE
EQLDYGIQGDDPYAEYFQDF
>ScTPK3
MYVDPMNNNEIRKLSITAKTETTPDNVGDIPVNAHSVHEECSNTPVEINGRNSGKLKE
EASAGICLVKKPMLQYRDTSGKYSLSDFQILRTLGTGSFGRVHLIRSNHNGRFYALKTLK
KHTIVKLKQVEHTNDERRMLSIVSHPFIIRMWGTQDSQQVFMVMDYIEGGELFSLLRKS
QRFPNPVAKFYAAEVC LALEYLHSDKIIYRDLKPENILLDKNGHIKITDFGFAKYVPDVT
YTLGTPDYIAPEVVSTKPYNKSVDWWSFGVLIYEMLAGYTPFYNSNTMKTYENILNAEL
KFPFFHPPDAQDLLKKLITRDLSERLGNLQNGSEDVKNHPWFNEVIWEKLLARYIETPYE
PPIQQGQGDTSQFDRYPEEEFNIGIYGEDPYMDLMKEF
>Cekin-1
MPTRLDIVGNLQFSSSTDNGDEDQEADVTAFCVLPSPSSFSKLSILDDPVEDFKFLDKA
REDFKQRWENPAQNTACLDDFDRIKTLGTGSFGRVMLVKHKQSGNYYAMKILDKQKVVKL
KQVEHTLNEKRI LQAIDFPFLVNMTFSFKDNSNLYMVLEFISGGEMFSHLRRIGRFSEPH
SRFYAAQIVLAFEYLHSLDLIYRDLKPENLLIDSTGYLKITDFGFAKRVKGRWTWLCGTP
EYLAPEIILSKGYNKAVDWWALGVLIYEEMAAGYPPFFADQPIQIYEKIVSGKVKFPPSHFS
NELKDLLKNLLQVDLTKRYGNLKNGVADIKNHKWFSTDWIAIYQKKITPPSFSKGESNG
RLF EALYPRVDGPADTRHFVEEVQEPTEFVIAATPQLEELFVEF
```

导入序列文件





执行比对

The screenshot shows the ClustalX (1.83) software interface. The menu bar includes File, Edit, Alignment, Trees, Colors, Quality, and Help. The 'Alignment' menu is open, showing options such as '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', and 'Output Format Options'. The main window displays a multiple sequence alignment of 14 sequences, with a ruler at the bottom indicating positions from 1 to 60. The sequences are color-coded by amino acid type.

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:
I:\tools\多序列比对\clustalx1.83\PKA_all.dnd

Output Alignment Files:
Clustal: I:\tools\多序列比对\clustalx1.83\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

SGNESKSWEKEGGETQEKPKQPHV
RQQQQQQRHQQLITSQLPKSLV
SVHEECSSNTFVEINGRNSGKLKE
PSSFSKLSILDDFVEDFKEFLDKA
PTNTAALDFERIKTLGTGSFGRV
SPYTNLENYITRAVLGNGSFGTVM
CGDHDSASGLRAGLATPTQRGKAT
TSPSTGLDYEIKATLGSFSFGK
NTAHLQDFERIKTLGTGSFGRVMI
NNAGLEDFERKKTGTGSFGRVMI
NTASSDQFERLRTLGMGSFGRVMI
LSPEPPVYSLQDFDTLAVGTGTE
RSPEAPAYRLQDCDALVTMGTGTE
LDLITEQYDNLSRALGRPLNVLDI

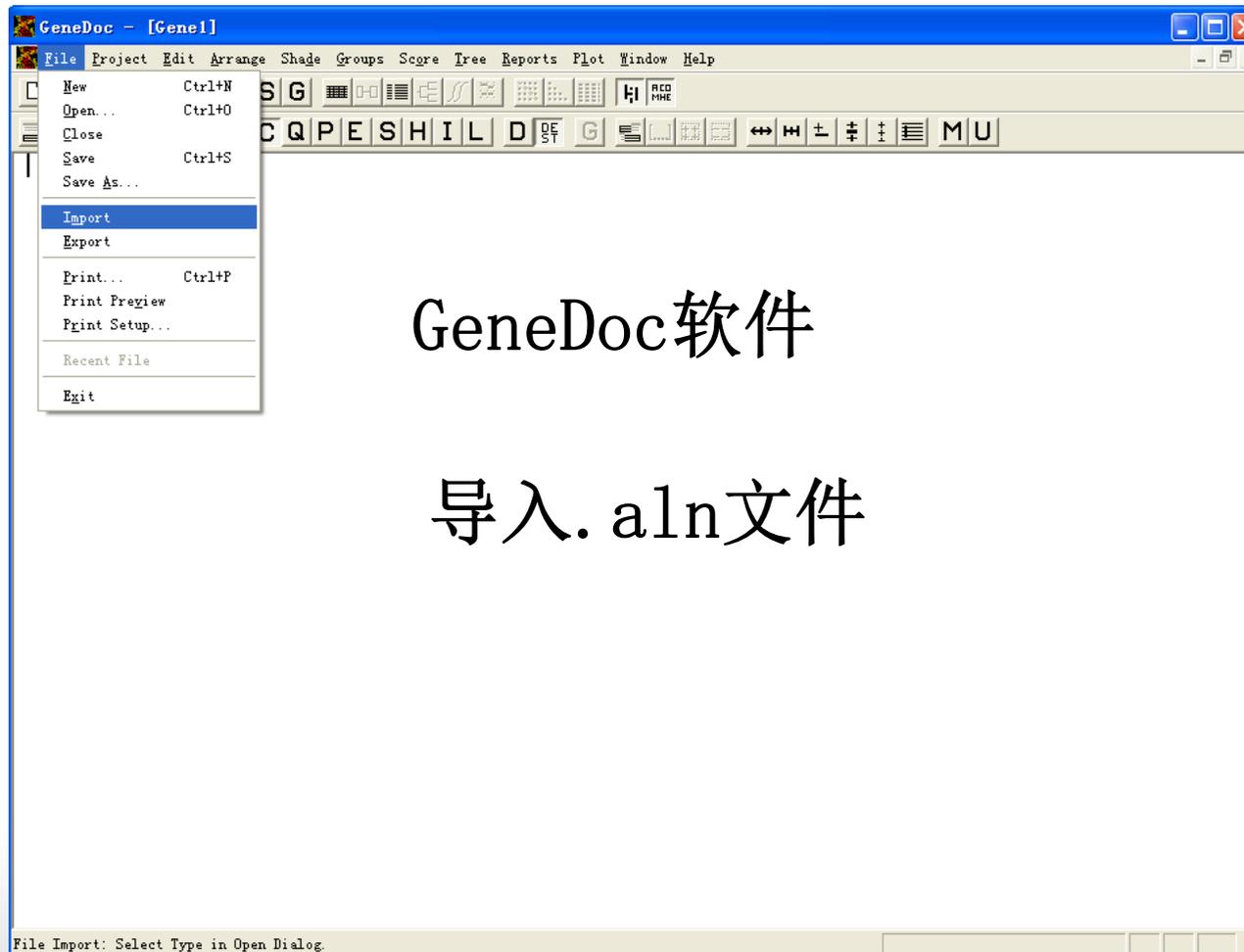
ruler 1 10 20 30 40 50 60

File F:\useful tools\多序列比对\clustalx1.83\PKA_all.seq loaded.

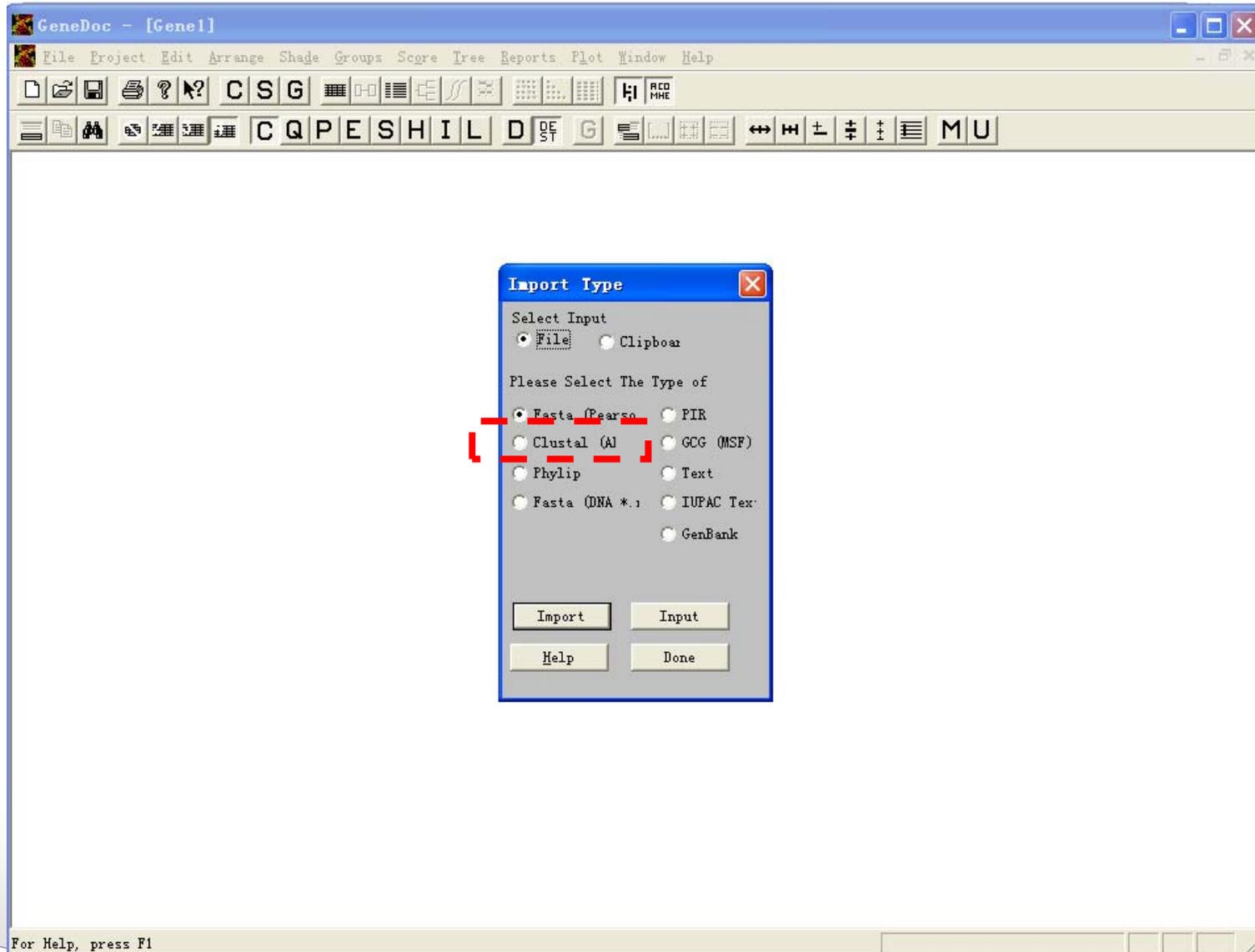
多序列比对：结果处理



□ GeneDoc, BioEdit等软件



选择文件格式



成功导入文件



GeneDoc - [Gene1]

File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help

C S G C Q P E S H I L D G M U

```
DmPKA-C3 : ADDATHDSSEIEEDDGNEDDDEDDDESESSSVQAKGVRKYHLDYQI-IKTVGIGTEGRVCLCRDRISEKY : 299
EcORF708 : --VSQPDDPRELIEQCAFYRLIGFDTHLSPVPRPMYLVSNHRVLIINDENQPFQHWQNPYAGAGLAHKRSRRYE : 244
t g g 5g v 6

DmPka-C2 : YAAKMMSKEDLVRL-----KQVAHVHNEKHVLSNAARFPFLIYLV DSTKC-FDYLYLILPLVNGGELFSY : 133
DmCG12069 : YASKQLSKDQIVKT-----KQVSHVMSEKNVLRSMTPFNNTVNLIASYKD-FDSLYLVPLIGGGELFTY : 135
HsPKACa : YAMKILDQKQVVKL-----KQIEHTLNEKRILQAVNFPFLVKLEFSFKD-NSNLYMVMYVPGGEMESH : 132
HsPKACb : YAMKILDQKQVVKL-----KQIEHTLNEKRILQAVNFPFLVRLLEYAFKD-NSNLYMVMYVPGGEMESH : 132
HsPKACg : YAMKILNKQKQVVKM-----KQVEHILNEKRILQATDFPFLVKLQFSFKD-NSYLYLVMEYVPGGEMESR : 132
DmPka-C1 : YAMKILDQKQVVKL-----KQVEHTLNEKRILQATQFPFLVSLRYHF KD-NSNLYMVLE YVPGGEMESH : 134
Cekin-1 : YAMKILDQKQVVKL-----KQVEHTLNEKRILQATDFPFLVNMTF SFKD-NSNLYMVLEBETISGGEMESH : 169
ScTPK1 : YAMKVLKKEIVVRL-----KQVEHTNDERLMLSIIVTHPFIIRMWGTFFQD-AQQIFMIMDYIEGGELFSL : 175
ScTPK3 : YALKTLKRHTIVKL-----KQVEHTNDERRMLSIIVSHPFIIRMWGTFFQD-SQQVFMVMDYIEGGELFSL : 176
ScTPK2 : YAIKVLKQKQVVKM-----KQVEHTNDERRMLKLVHHPFLIRMWGTFFQD-ARNIFMVMDYIEGGELFSL : 158
HsPRKX : FALKVMSIPDVIRL-----KQEQHVHNEKSVLKEVSHPFLIRLFWTWHD-ERFLYMLMEYVPGGELFSY : 137
HsPRKY : FALKVMSIPDVIRL-----KQEQHVHNEKSVLKEVSHPFLIRLFWTWHE-ERFLYMLMEYVPGGELFSY : 137
DmPKA-C3 : CAMKILAMTEVIRL-----KQIEHVKNERNILREIRHPFVISLEWSTKD-DSNLYMIFDYVCGGELFTY : 362
EcORF708 : EGEDYVCRFFYYDMPHGILTAESQRNKHELHNEIKFLTQPPAGE DAPAVLAHGENAQSGLVMEKLEP-RLLSD : 318
a k 6 k 4q h E L pf d 566 6 Gge6f3

DmPka-C2 : HRRVRKFNKHFYAAQVLALEYLHCSLLYRDLKPENILLDQRGYIKITDFG-FTKRVDGRSTLTCGTF--- : 204
DmCG12069 : HRKVRKFEKQARFYAAQVELALEYLHCSLLYRDLKPENIMMDKNGYLKVTDFG-FAKRVETRTMTLCGTF--- : 206
HsPKACa : LRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQGGYIQVTDG-FAKRVKGRWTWTLGTF--- : 203
HsPKACb : LRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQGGYIQVTDG-FAKRVKGRWTWTLGTF--- : 203
HsPKACg : LQRVGRFSEPHACFYAAQVLAQVYLSLDLIRDLKPENLLIDQGGYLVQVTDG-FAKRVKGRWTWTLGTF--- : 203
DmPka-C1 : LRKVGSRFSEPHSRFYAAQIVLAFEYLHSLDLIYRDLKPENLLIDSTGYLKITDFG-FAKRVKGRWTWTLGTF--- : 205
Cekin-1 : LRRIGRFSEPHSRFYAAQIVLAFEYLHSLDLIYRDLKPENLLIDSTGYLKITDFG-FAKRVKGRWTWTLGTF--- : 240
ScTPK1 : LRKSQRF PNPVAKFYAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFG-FAKYVPDVTYTLGTF--- : 246
ScTPK3 : LRKSQRF PNPVAKFYAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFG-FAKYVPDVTYTLGTF--- : 247
ScTPK2 : LRKSQRF PNPVAKFYAAEVLLALEYLHAHNLIYRDLKPENILLDRNGHIKITDFG-FAKEVQTVTWTLCGTF--- : 229
```

For Help, press F1



选择需要拷贝的行

The screenshot shows the GeneDoc software interface with a sequence alignment. A context menu is open over the row for 'DmPKA-C3', with the option 'Select Blocks for Copy' highlighted. The alignment shows several sequences with gaps and asterisks indicating conserved regions. The menu options include: Find (F9), Find Next (F10), Replace (Ctrl+R), Select Blocks for Copy (Ctrl+E), Copy Selected Blocks to, Copy Alignment 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, and Clear Man Comments.

```
GeneDoc - [Gene1]
File Project Edit Arrange Shade Groups Score Tree Reports Plot Window Help
Fairwise Alignment
Find F9
Find Next F10
Replace Ctrl+R
Select Blocks for Copy Ctrl+E
Copy Selected Blocks to
Copy Alignment 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

HsPRKX : ---KOEQHVHNEKSVLKEVSHPPFLIRLFWTWHD-ERFLYMLMEYVPGGELEFSY : 137
HsPRKY : ---KOEQHVHNEKSVLKEVSHPPFLIRLFWTWHE-ERFLYMLMEYVPGGELEFSY : 137
DmPKA-C3 : ---KQIEHVKNERNILREIRHPFVLSLEWSTKD-DSNLYMIFDYVCGGELEFTY : 362
EcORF708 : ESQRNKHELHNEIKFLTQPPAGE-DAPAVLAHGENAQSGWLVMEKLPG-RLLSD : 318
4q h E L pf d 566 6 Gge6f3

DmPka-C2 : EYMHKMHLMYRDLKPENILLDQRGYIKITDFG-FTKRV DGRSTLTCGTP--- : 204
DmCG12069 : EYLLHCSLLYRDLKPENIMMDKNGYLVKVTDFG-FAKKVETRMTLTCGTP--- : 206
HsPKACa : EYLSHSLDLIYRDLKPENLLIDQQGYIQVTDG-FAKRVKGRWTWTLTCGTP--- : 203
HsPKACb : EYLSHSLDLIYRDLKPENLLIDHQGYIQVTDG-FAKRVKGRWTWTLTCGTP--- : 203
HsPKACg : VQYLSHSLDLIYRDLKPENLLIDQQGYLVKVTDFG-FAKRVKGRWTWTLTCGTP--- : 203
DmPka-C1 : EYLYHSLDLIYRDLKPENLLIDSQGYLVKVTDFG-FAKRVKGRWTWTLTCGTP--- : 205
Cekin-1 : EYLSHSLDLIYRDLKPENLLIDSTGYLKITDFG-FAKRVKGRWTWTLTCGTP--- : 240
ScTPK1 : LRKSQRF PNPVAKFYAAEVCLALEYLHSKDIYRDLKPENILLDKNGHIKITDFG-FAKYVPDVTYTLTCGTP--- : 246
ScTPK3 : LRKSQRF PNPVAKFYAAEVCLALEYLHSKDIYRDLKPENILLDKNGHIKITDFG-FAKYVPDVTYTLTCGTP--- : 247
ScTPK2 : LRKSQRF PNPVAKFYAAEVILALEYLHANNIYRDLKPENILLDRNGHIKITDFG-FAKEVQTVTWTLTCGTP--- : 229
HsPRKX : LRNRGRFSSTTGLFYSAEIIICALEYLHSEIYVYRDLKPENILLDRDGHIKLTDG-FAKKLVDRTWTLTCGTP--- : 208
HsPRKY : LRNRGHFSSTTGLFYSAEIIICALEYLHSEIYVYRDLKPENILLDRDGHIKLTDG-FAKKLVDRTWTLTCGTP--- : 208
DmPKA-C3 : LRNAGKFTSQTSNFYAAEIVSALAYLHSLQIVYRDLKPENLLINRDGHLKITDFG-FAKKLRDRTWTLTCGTP--- : 433
EcORF708 : MLAAG--EEIDREKILGSLRLSLAALEKQGFWHDDVREPMVVMVDARQHARLLIDFGSIVTTPQDCS WPTNLVQSFF : 391
r f fyaa 6 a y6h yrD64PeN6661 g 6tDFG fak 3 tlcgtp

DmPka-C2 : EYLAPEIVQLRPNY-----KSV DWWAFGILVYEFVAGRS PFAIHN R----- : 245
DmCG12069 : EYLPPEIIQSKPYG-----TSVDWWAFGVLVVEFVAGHSPFSAHN R----- : 247
HsPKACa : EYLAPEIILSKGYN-----KAVDWWALGVLIYEMAAGYPPFFADQ----- : 243
HsPKACb : EYLAPEIILSKGYN-----KAVDWWALGVLIYEMAAGYPPFFADQ----- : 243
HsPKACg : EYLAPEIILSKGYN-----KAVDWWALGVLIYEMAVGFPFFYADQ----- : 243
DmPka-C1 : EYLAPEIILSKGYN-----KAVDWWALGVLIYEMAAGYPPFFADQ----- : 245
Cekin-1 : EYLAPEIILSKGYN-----KAVDWWALGVLIYEMAAGYPPFFADQ----- : 280
ScTPK1 : DYI APEVVS TKPN-----KSIDWWSEGILLIYEMLAGYT PFFYDSN----- : 286

Select an area to copy. ScTPK2:
```



比对结果的美化和后处理

Cekin-2	SGGRRTGISAE	89
ScBCY1	NAQRRTISVSGE	149
DmPka-R2	ASSRRKSVFAE	88
HsPRKAR2A	N--RRVSVCAE	103

Cekin-2	DYFGEIALLDRPRAATVVAKTH	329
ScBCY1	DYFGEVALLNDLPRQATVTATKR	386
DmPka-R2	QYFGELALVTHRPRASVYATGG	329
HsPRKAR2A	QYFGELALVTNKPRASAYAVGD	356

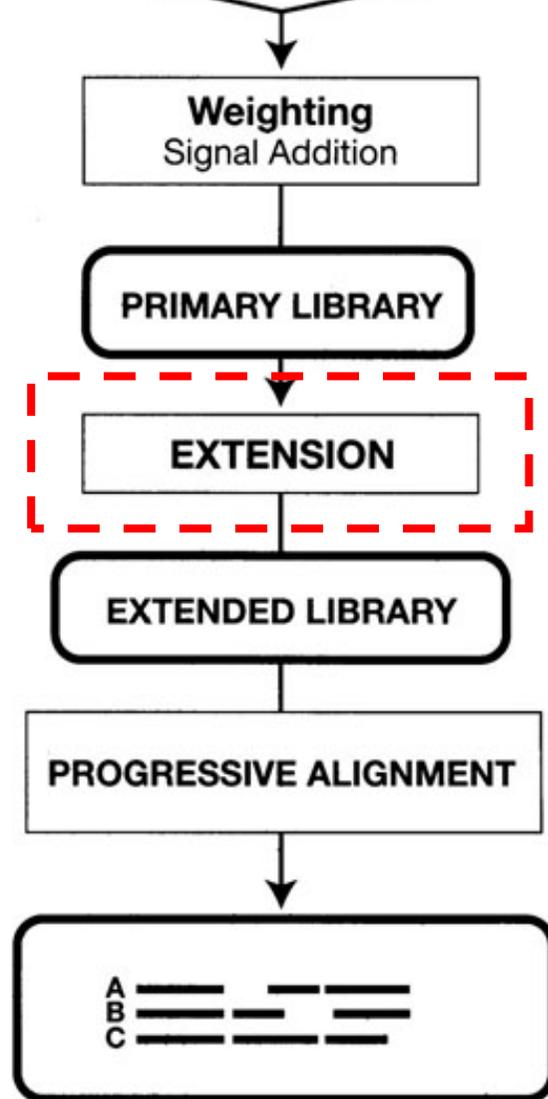
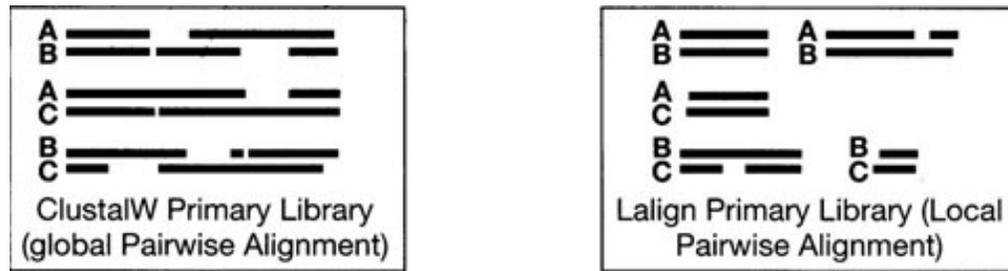
T-Coffee



- 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

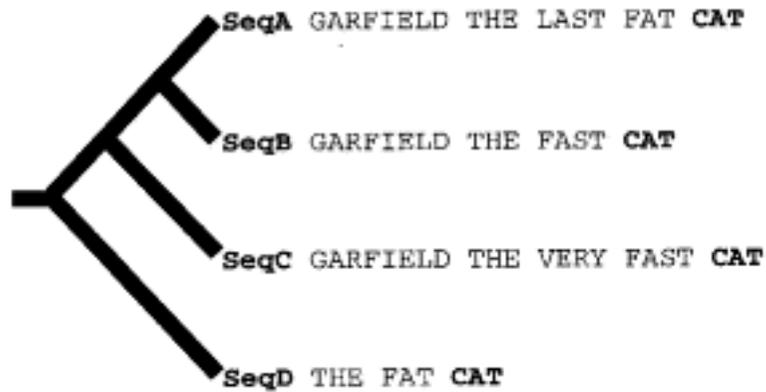


对以上打分的结果设计权重系统，找到序列中最保守的部分

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



a) Regular Progressive Alignment Strategy



```

SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD ----- THE ---- FA-T CAT
  
```

通常的“渐进”算法

b) Primary Library

```

SeqA GARFIELD THE LAST FAT CAT Prim. Weight = 88
SeqB GARFIELD THE FAST CAT ---

SeqA GARFIELD THE LAST FA-T CAT Prim. Weight = 77
SeqC GARFIELD THE VERY FAST CAT

SeqA GARFIELD THE LAST FAT CAT Prim. Weight = 100
SeqD ----- THE ---- FAT CAT
  
```

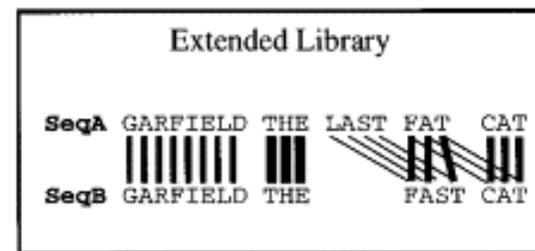
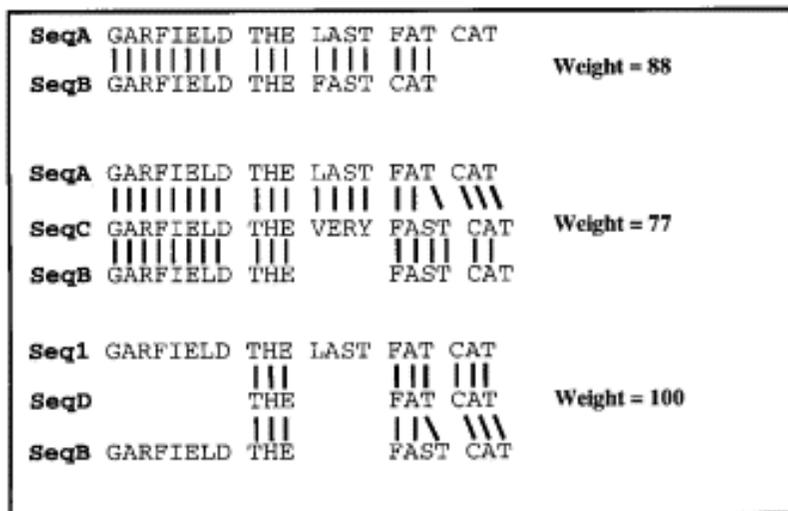
```

SeqB GARFIELD THE ---- FAST CAT Prim Weight = 100
SeqC GARFIELD THE VERY FAST CAT

SeqB GARFIELD THE FAST CAT Prim. Weight = 100
SeqD ----- THE FA-T CAT

SeqC GARFIELD THE VERY FAST CAT Prim. Weight = 100
SeqD ----- THE ---- FA-T CAT
  
```

c) Extended Library for seq1 and seq2



Dynamic Programming

```

SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE ---- FAST CAT
  
```

基于指导库的修正

渐进方法存在的问题



- 启发式算法（Heuristic algorithm）
 - ✿ 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

例如



□ 三条序列:

Seq1: ARKCV

Seq2: ARCV

Seq3: AKCV

□ 若Seq1, 2先比对,
再加入Seq3:

ARKCV

AR-CV

A-KCV

□ Seq1, 3先比对,
再加入Seq2:

ARKCV

A-RCV

A-KCV

□ Seq2, 3先比对,
再加入Seq1:

ARKCV

AR-CV

AK-CV

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

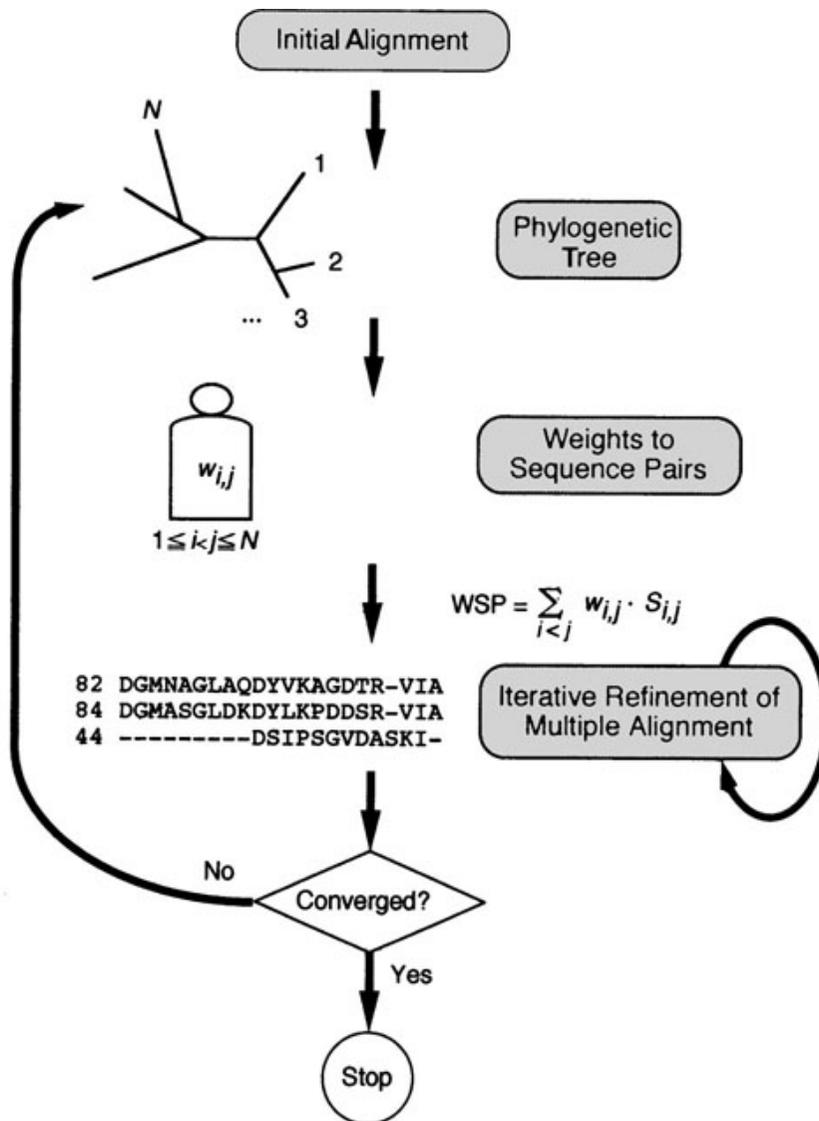
❑ PRRP/PRRN

🌸 <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

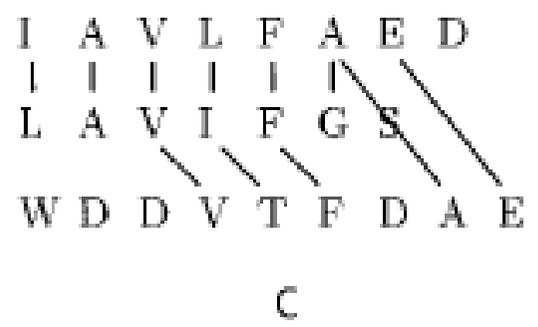
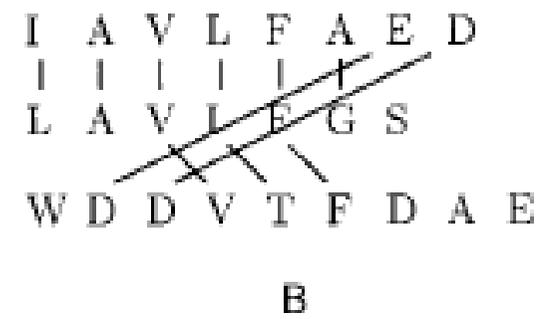
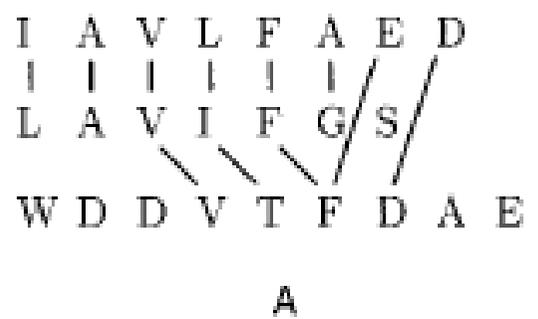


- 对所有序列进行两两之间的局部最优比对
- 找到所有能够匹配的部分M1；将重叠的、前后一致的（consistency）匹配部分连接起来为M2
- 将剩下的未比对的序列重新比对，再发现能够匹配的部分，构成新M1，将一致的部分构成M2
- 重复上述步骤，直到结果收敛



一致的 vs. 不一致的

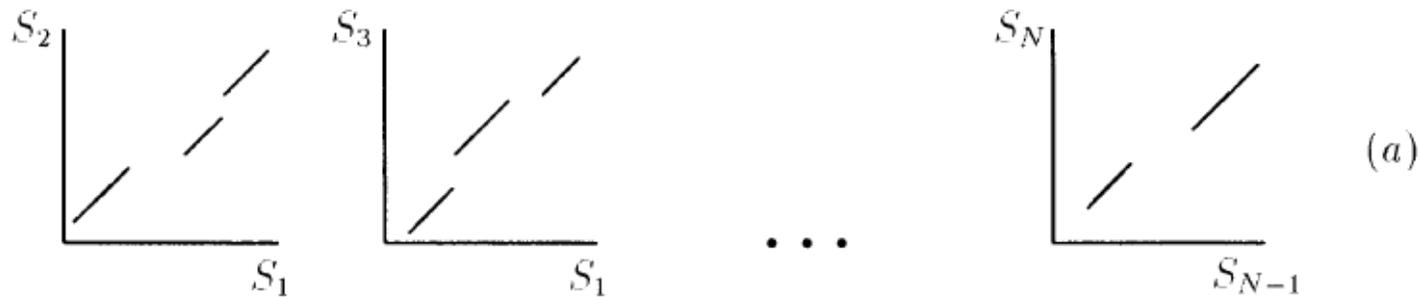
不一致的 (Non-consistent)



一致的 (Consistent)

最终的比对结果

DIALIGN: 算法流程



\mathcal{M}_1

Overlap weights (b)



Sort diagonals (c)

\mathcal{M}_1



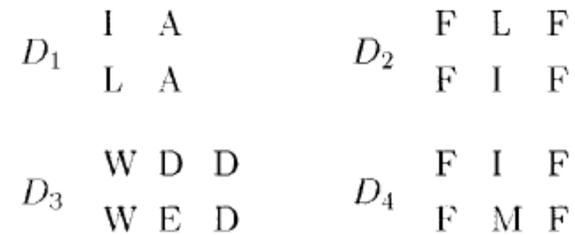
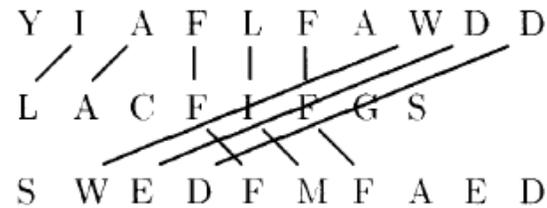
Consistency! (d)

\mathcal{M}_2



迭代过程

1. iteration step



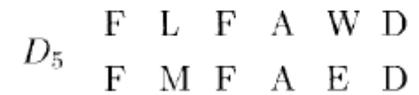
weight scores:

	D_1	D_2	D_3	D_4
weight	0.2	2.6	4.7	2.2
overlap weight	0.2	5.3	4.7	4.9

$M1=\{D1, D2, D3, D4\}$

$M2=\{D1, D2, D4\}$

2. iteration step



$M1=\{D1, D2, D4, D5\}$

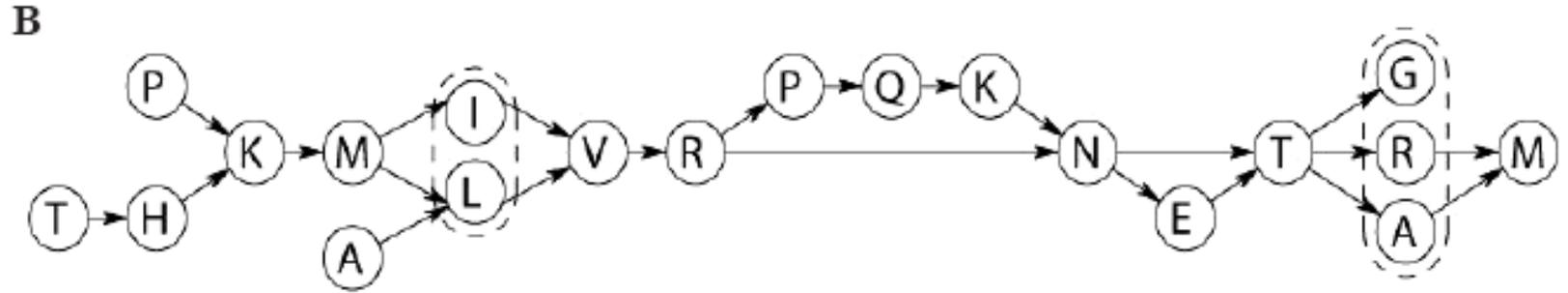


部分有向图算法：POA

- ❑ <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



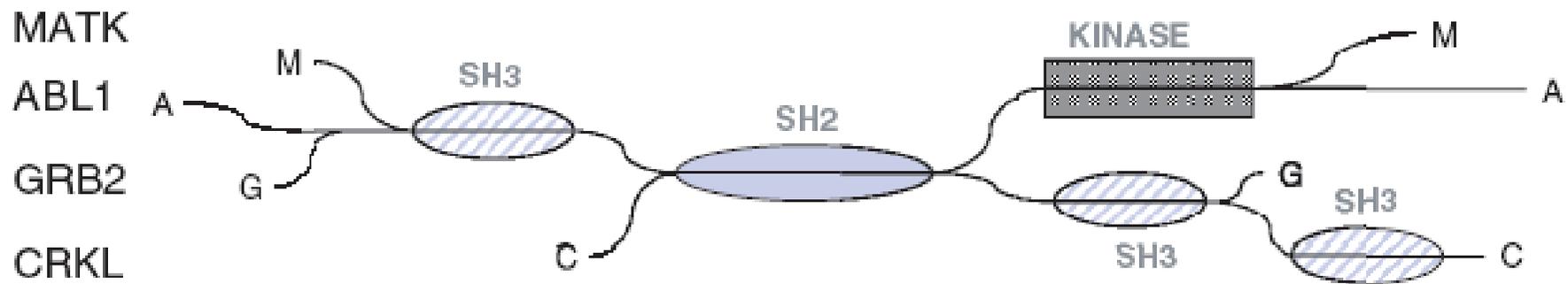
(a)



(b)

CONSENS1TGTAONT.GTTTGTGAGG.CTA
CONSENS0	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S663801	A.GTTCCTGC.TGCGTTTGCTGGACTTATGACTT.GTTTGTGAGG.CAA
Hs#S337687	AA G TTTCCTGC.TGCGTTTGCTGGACTGATGACTT CG TTTGT GNAG CAA
Hs#S629177	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGT N AGG.CAA
Hs#S672957	A.GTTCCTGC.TGCGTTTGCT.....
Hs#S672182	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTT.....
Hs#S674099	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S196113	A.GTT N CT GN T GN GTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S994400GTACONT.GTTTGTGAGG.CTA
Hs#S550772	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S80460	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S39701	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S1988018	A.GTTCCTGC.TG C TTTGCTGGACTGATGACTT. CA TTGTGAGG.CAA
Hs#S341915	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S1794113	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S4698	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGT C GG.CAA
Hs#S813765	A.GT C CTGC. G CGTTTGC G GA C GATGACTT.GTT G GTGAGG.CAA
Hs#S1184845G.CAA
Hs#S1577463GG.CAA
Hs#S914987CTGATGACTT.GTT G GTGAG G CAA
Hs#S1985364	A.GTTCCTGC.TGCGTTTGCTGGACTGATGACTT.GTTTGTGAGG.CAA
Hs#S1465644	.GTT C .TG C CTGC G CGTTTGCT G AACTGATGACTT.GTTAGT. AG .CAA
Hs#S1850471	C .GTTACTGC. C CGTTTGCTGGACT CA T G .ACTT G T NG T.AGG.CAA

激酶的多序列比对



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

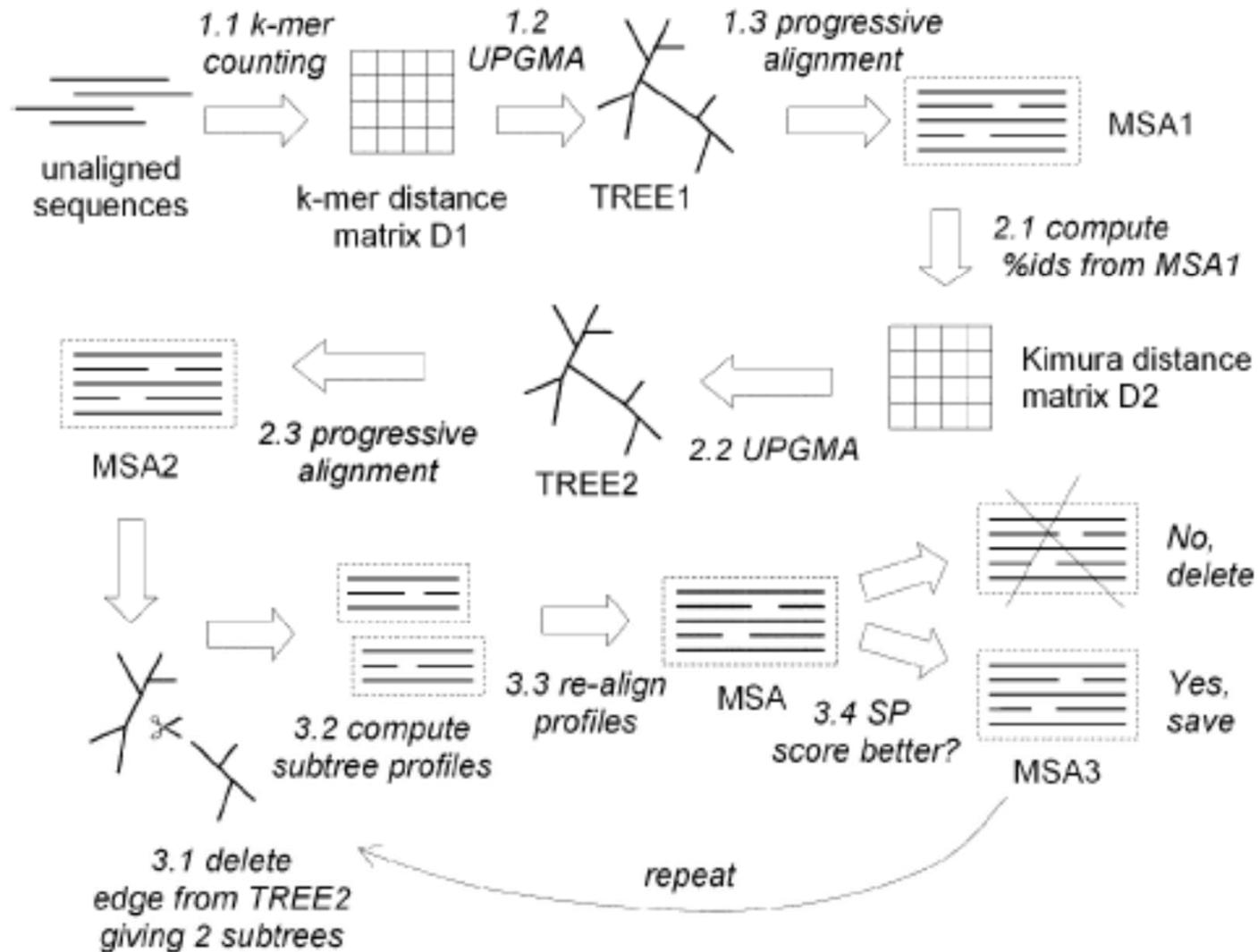
MUSCLE (3)



□ 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



[Home](#) [Software](#) [Services](#) [About](#) [Contact](#)

Muscle5

MUSCLE has been cited by
49,171 papers
[Google scholar](#)
Last updated 26 Jan 2023

[Download](#)

[Documentation](#)

[Support and feedback](#)

[MUSCLE v3](#)

Next-generation MUSCLE

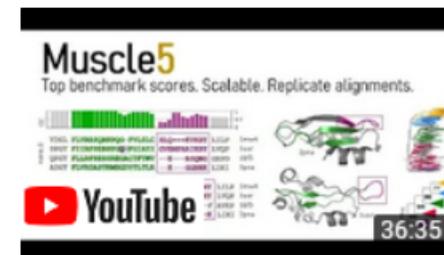
Muscle v5 is a major re-write of MUSCLE based on new algorithms.

Highest accuracy, scalable to thousands of sequences

Compared to previous versions, Muscle v5 is much more accurate, is often faster, and scales to much larger datasets. At the time of writing (late 2021), Muscle v5 has the highest scores on multiple alignment benchmarks including Balibase, Bralibase, Prefab and Balifam. It can align tens of thousands of sequences with high accuracy on a low-cost commodity computer (say, an 8-core Intel CPU with 32 Gb RAM). On large datasets, Muscle v5 is 20-30% more accurate than MAFFT and Clustal-Omega.

Alignment ensembles

Muscle v5 can generate ensembles of high-accuracy alternative alignments. All replicates have equal average accuracy on benchmark test, including the MSA made with default parameters. By comparing results of downstream analysis (trees, structure prediction...) on different replicates, you can assess the effects of alignment errors on your study.



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

MUSCLE: 使用说明



```
C:\ 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/>



Clustal Omega

"The last alignment program you'll ever need"



[Home](#)

[Webservers](#)

[Download](#)

[Documentation](#)

[Contact](#)

[News](#)

Introduction

Clustal Omega is the latest addition to the Clustal family. It offers a significant increase in scalability over previous versions, allowing hundreds of thousands of sequences to be aligned in only a few hours. It will also make use of multiple processors, where present. In addition, the quality of alignments is superior to previous versions, as measured by a range of popular benchmarks.

Please note that Clustal Omega is currently a command line-only tool.

A full description of the algorithms used by Clustal Omega is available in the Molecular Systems Biology paper [Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega](#). Latest additions to Clustal Omega are described in [Clustal Omega for making accurate alignments of many protein sciences](#)

[Webservers](#)

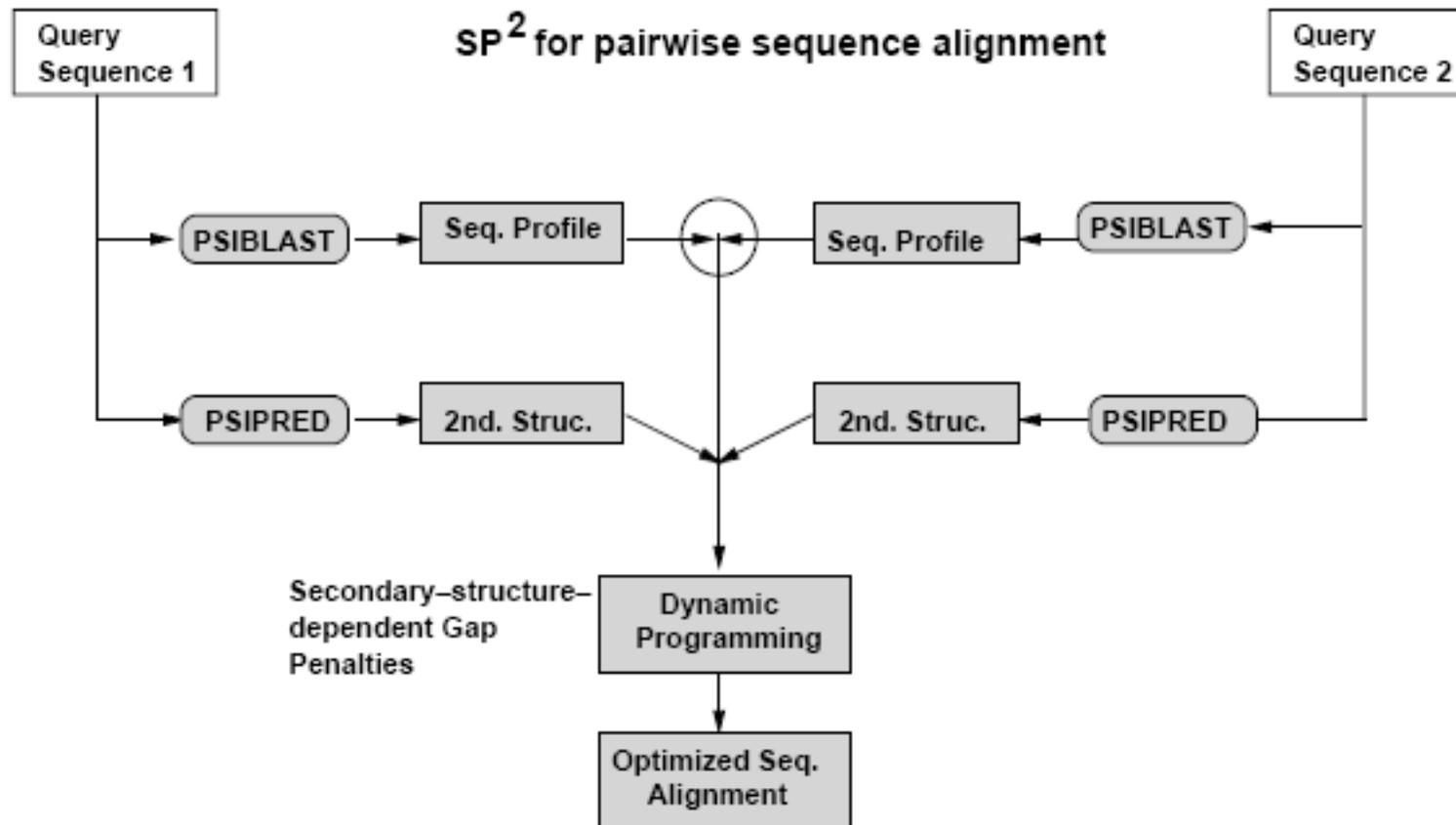
[Download Clustal Omega](#)

结构特征



□ 2005, SPEM

□ 空位罚分：结合蛋白质的二级结构信息

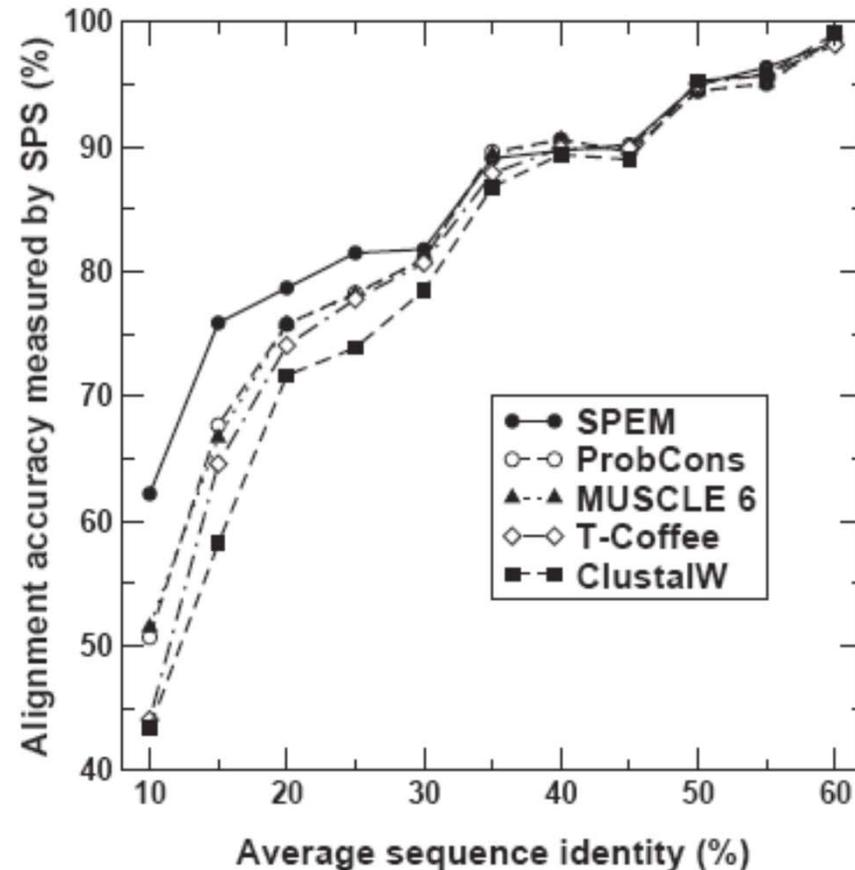
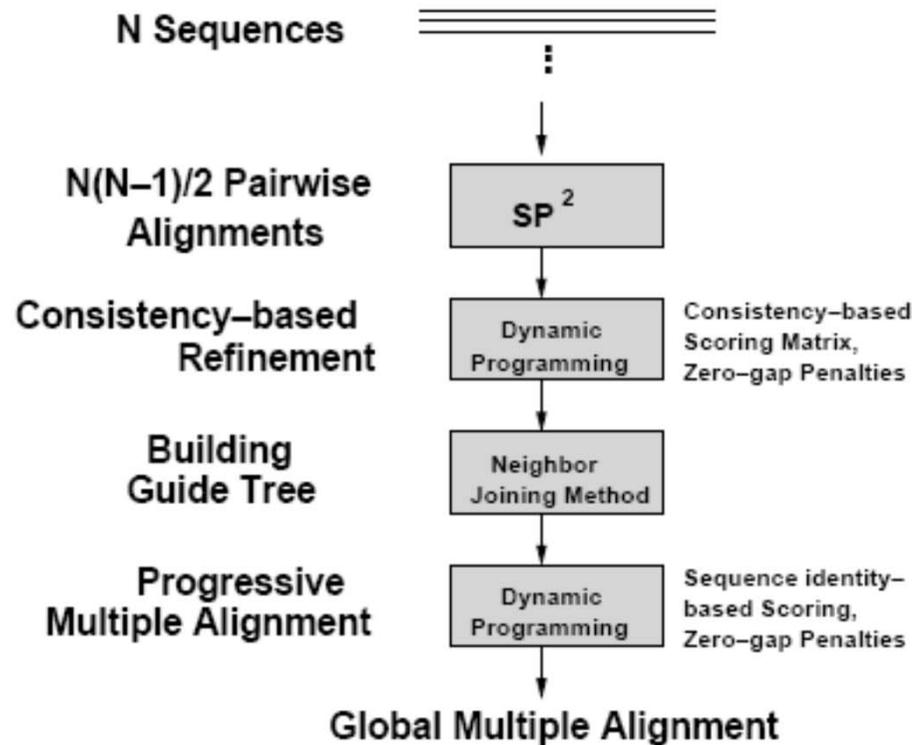


SPEM



- 序列相似度高：准确性大致相当
- 序列相似度低：比其他工具高7~15%

SPEM for multiple sequence alignment

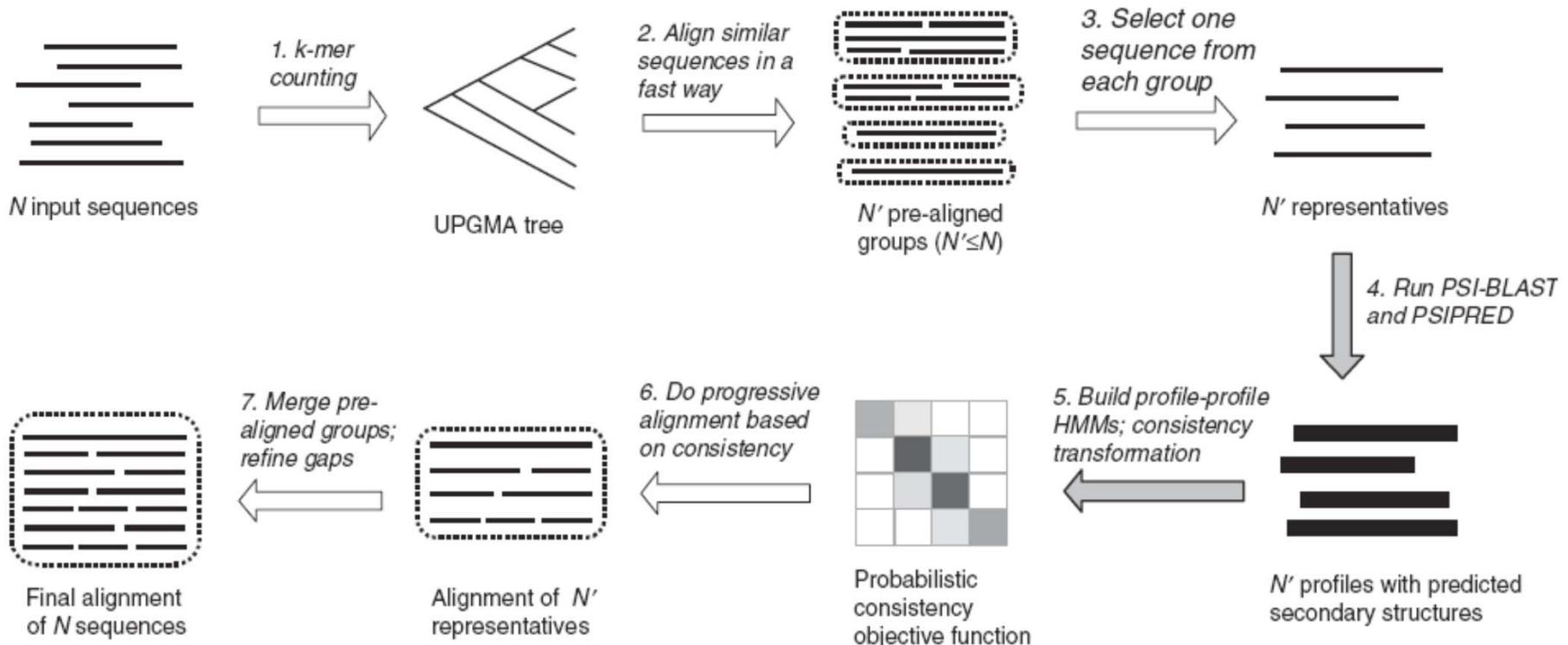


PROMALS



□ 2007年，序列identity<10%的多序列比对

- ❁ 数据库搜索更多同源序列
- ❁ 预测二级结构
- ❁ 隐马尔科夫模型：考虑氨基酸的打分和二级结构
- ❁ 渐进算法的概率打分



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

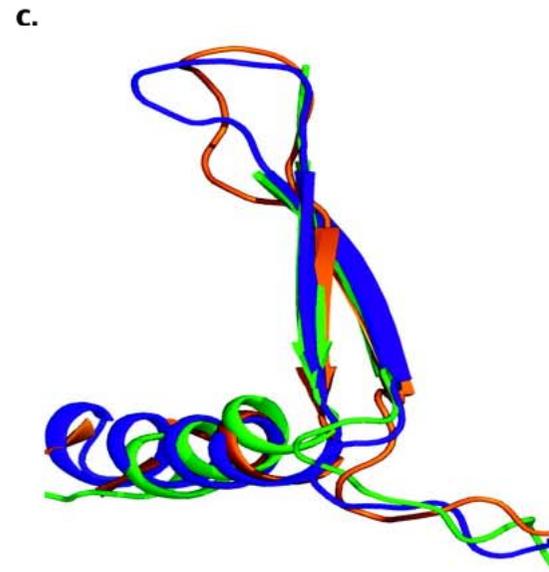
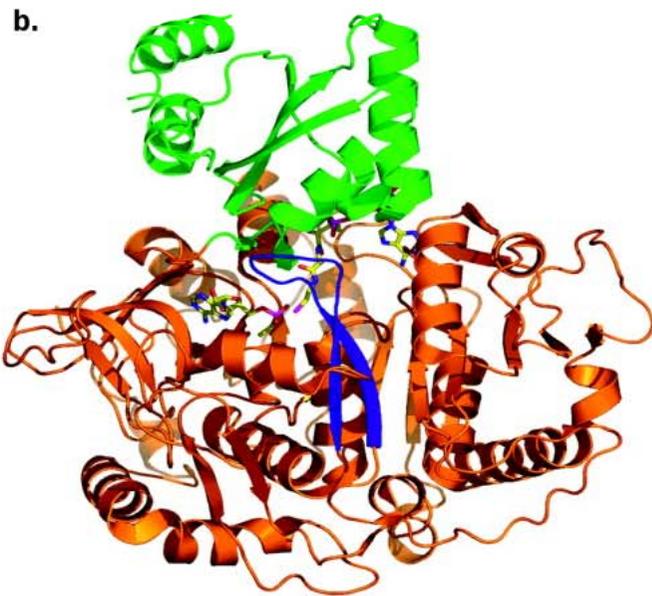
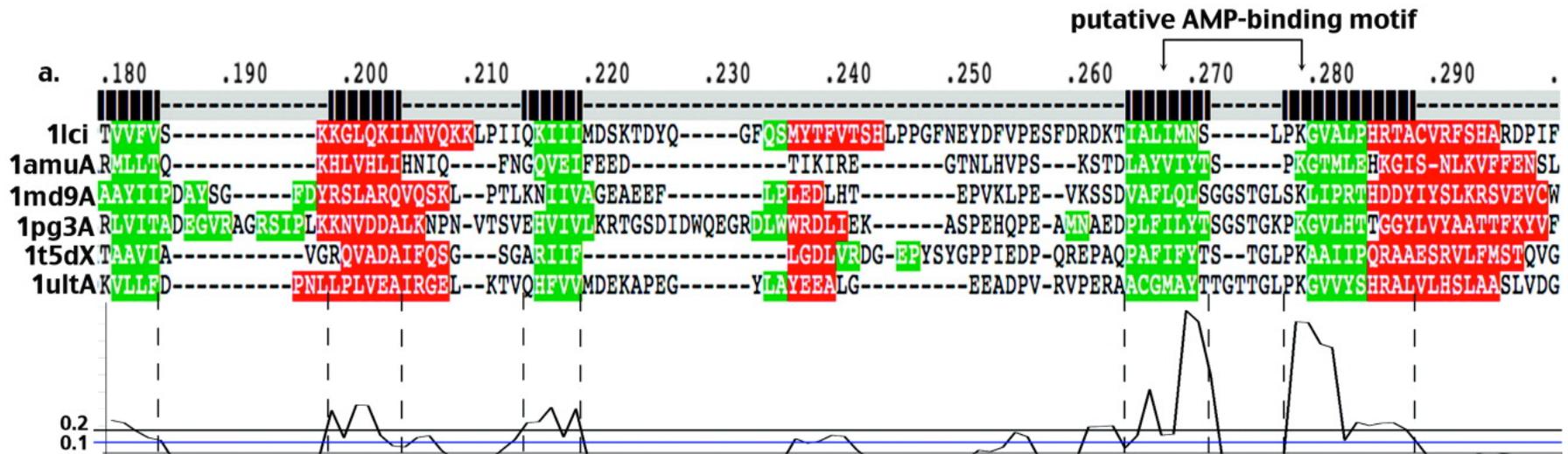
- ❑ **BAliBASE**: 基于蛋白质三级结构，将同一家族的蛋白质序列进行多序列比较
- ❑ 多序列比对工具的性能检验：能否与BAliBASE中的比对结果相吻合

🌸 <http://www.lbggi.fr/balibase/>

Welcome to BAliBASE 4
download the whole benchmark by html

Reference 1: variability, length
Reference 1: variability, length
Reference 2: orphans
Reference 3: sub-families
Reference 4: extensions
Reference 5: insertions
References 6,7,8: Repeat, Transmembrane, Circ. permutation
Reference 9: linear motifs
Reference 10: mixed

AMP结合酶的结构/序列比较



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

