



生物信息学

第五章 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	ARFPFLIYLVLDSTKCFDYLYLILPLVNGGELFSYHRRVRKFNEKHARFYAAQVALALEYMHKMHLIMYRD	168
DmCG12069	102	MTFPNTVALIASYKDFDSLYLVLPLIGGGELFTYHRKVRKFTEKQARFYAAQVFLALEYIHHCSLLYRD	170
ScTPK2	125	VEHPFLIRMWGTFQDARNITFMVMDYIEGGELFSLLRKSQRFNPVAKFYAAEVILLAELYIHAHNIIYRD	193
ScTPK1	142	VTHPFIIRMWGTFQDAQQIFMIMDYIEGGELFSLLRKSQRFNPVAKFYAAEVCLALEYIHSKDDIIYRD	210
ScTPK3	143	VSHPFIIRMWGTFQDSQQVFMVMDYIEGGELFSLLRKSQRFNPVAKFYAAEVCLALEYIHSKDDIIYRD	211
Cekin-1	136	IDFPFLVNMTFSEKDNSNLYMVLEFISGGEMFSHLRRIGRFSEPHSRFYAAQIVLAFELYIHSLDLIIYRD	204
DmPka-C1	101	TQFPFLVSLRYHFKDNSNLYMVLEYVPGGEMFSHLRKVGRFSEPHSRFYAAQIVLAFELYIHYLDLIIYRD	169
HsPKACg	99	IDFPFLVKLQFSFKDNSLYLYLVMEMYVPGGEMFSRLQRVGRFSEPHACFYAAQVVLAVCYIHSLDLIIHRD	167
HsPKACa	99	VNFPFLVKLEFSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYIHSLDLIIYRD	167
HsPKACb	99	VNFPFLVRLEYAFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYIHSLDLIIYRD	167
DmPKA-C3	329	TRHPFVISLEWSTKDDSNLYMIEDYVCGGELFTYLRNAGKFTSQTSNFYAAEVSALEYIHSLQIVYRD	397
HsPRKX	104	VSHPFLIRLFWTWHDERFLYMLMEYVPGGELFSYLRNRGRFSSTTGLFYSAEIICAIHEYIHSKEIVYRD	172
HsPRKY	104	VSHPFLIRLFWTWHEERFLYMLMEYVPGGELFSYLRNRGHFSSTTGLFYSAEIICAIHEYIHSKEIVYRD	172

Made by GENEDOC

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

Bioinformatics, 2020, HUST



双序列比对的时间复杂度

时间复杂度: $O(n^2)$

	Gap	V	D	S	C	Y
Gap	0	11	-22	-33	-44	-55
V	-11	4	-7	-18	-29	-40
E	-22	-7	6	-5	-16	-27
S	-33	-18	-5	10	-1	-12
L	-44	-29	-16	-1	9	-3
C	-55	-40	-27	-12	8	7
Y	-66	-51	-38	-23	-3	15



多序列比对：最优算法

ARDFSHGLLENKLLGCD SMRWE
..:.. .:..:.. .:::..:..:..
GRDYKMA LL E QW I LGCD -MRWD
.:.:.:.:.:.:.:.:.:.:.:.:.:
SRDW--ALIEDCMV-CNFFRWD

多项式时间复杂度： $\leq O(n^3)$

三条序列：时间复杂度： $O(lmn) = O(n^3)$

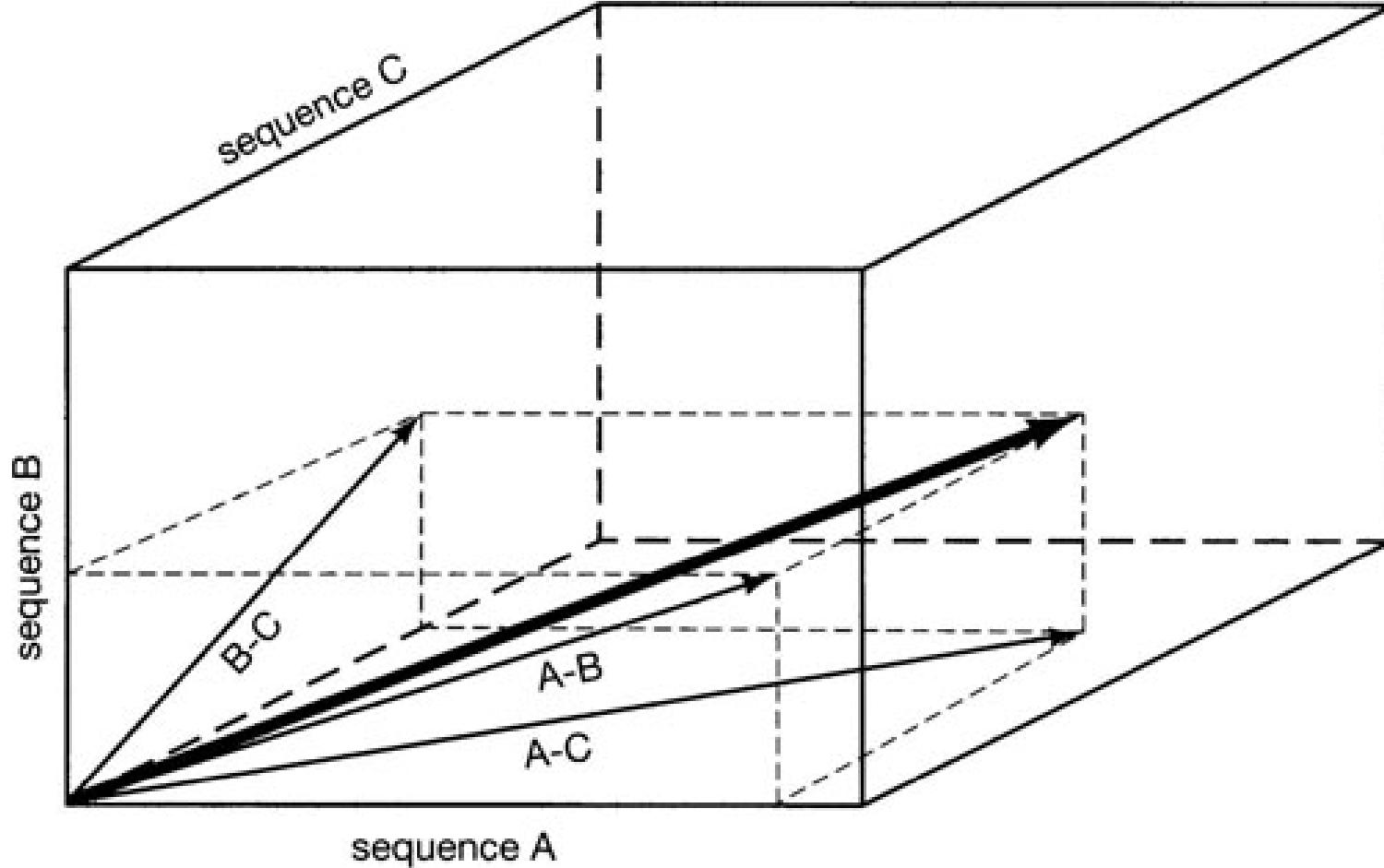
四条序列：时间复杂度： $O(n^4)$ ，非多项式时间！

...

m 条序列：时间复杂度： $O(n^m)$ ，指数时间！



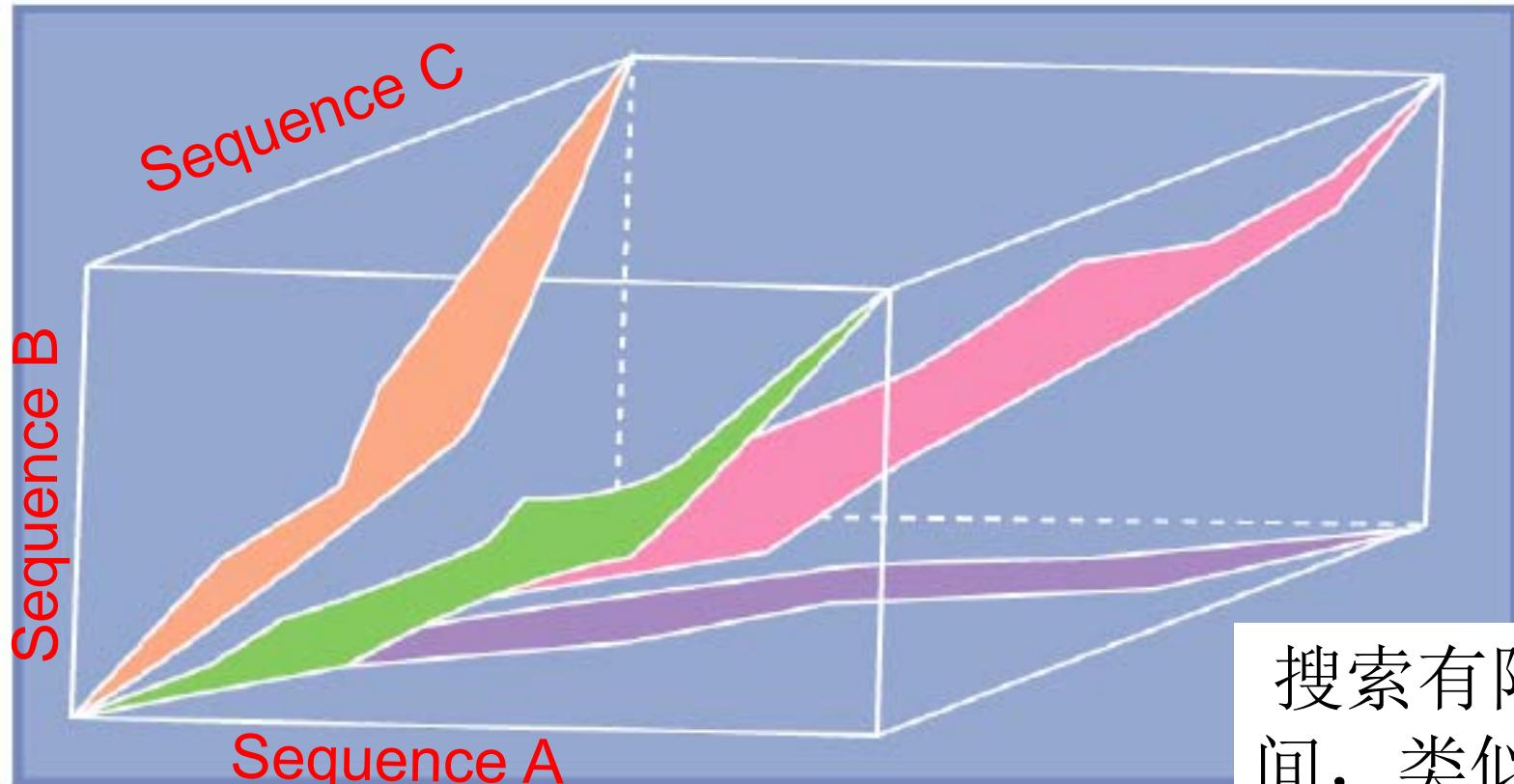
动态规划算法：全空间



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



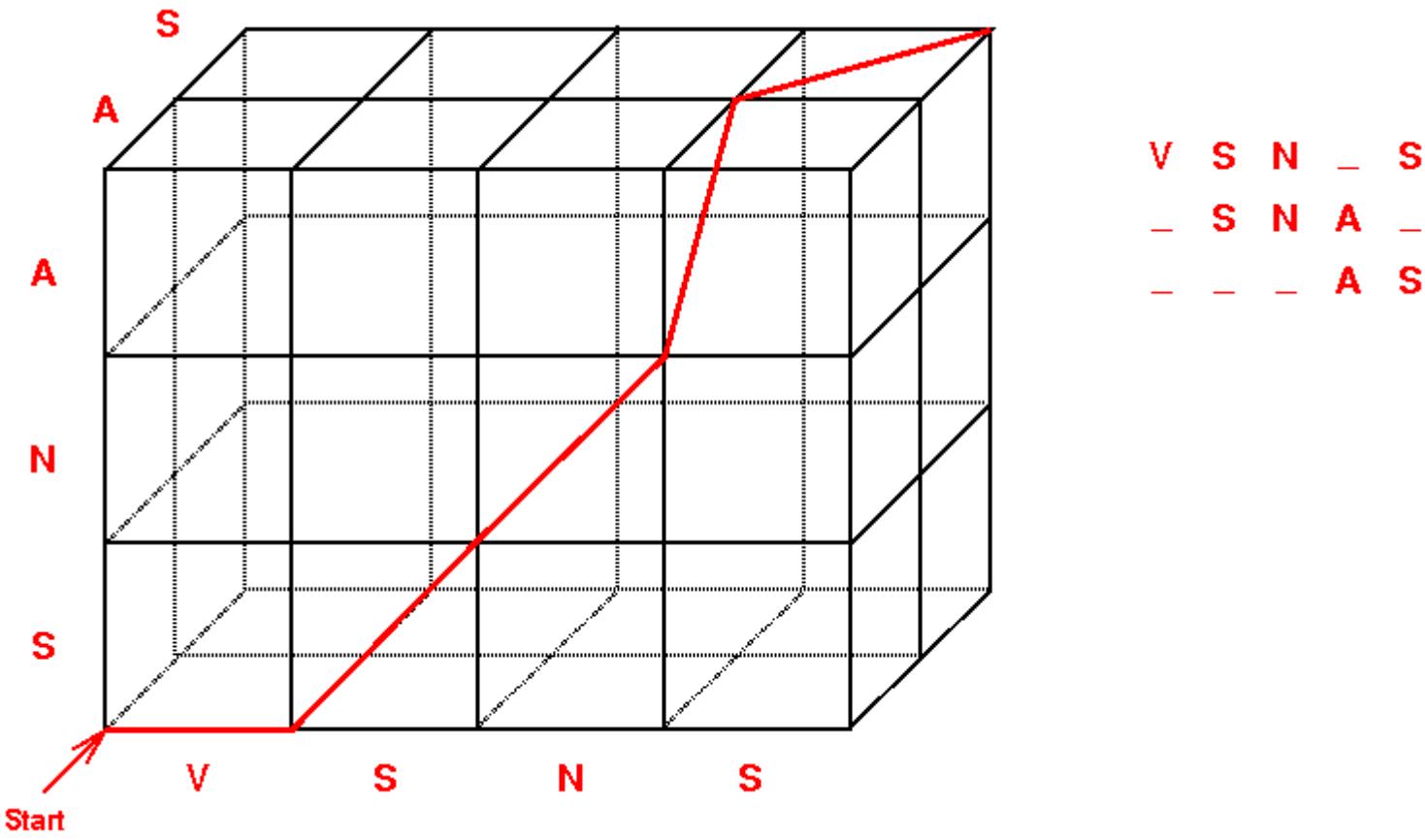
动态规划算法：优化算法



搜索有限空间，类似于
BLAST算法

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

动态规划算法：hyperlattice



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

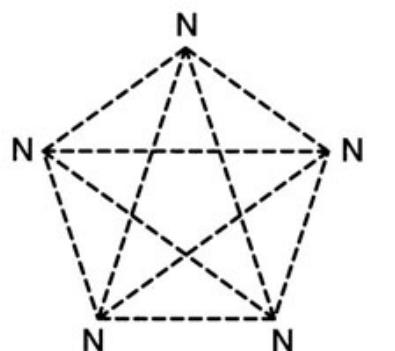
-	S	N	A	-
-	-	-	A	S

非最优的双序列比对

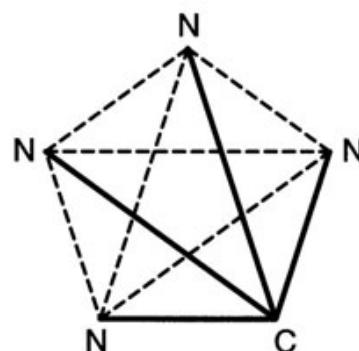


MSA: 多序列比对的打分策略

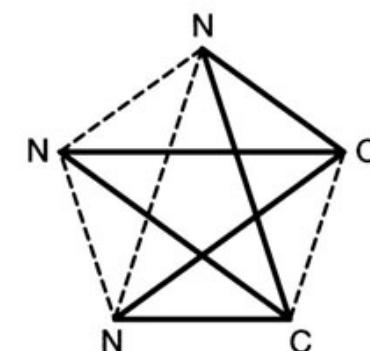
Sequence	Column A	Column B	Column C
1N.....N.....	N
2N.....N.....	N
3N.....	N.....	N
4N.....	N.....	C
5N.....	C.....	C



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

- 漐进方法: **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



- 演进方法：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. Kecelioglu)	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



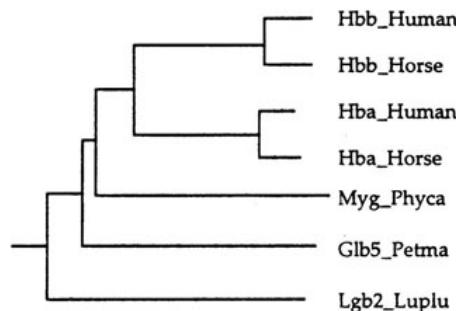
- 将所有序列两两比对，计算进化距离（差异）矩阵
- 使用邻接法（neighbor-joining）构建指导树（guide tree）
- 将进化距离最近的两条序列用全局动态规划算法进行比对
- “渐进” 地加上其他序列



两两比对，构
建距离矩阵

	1	2	3	4	5	6
1	-					
2	.17	-				
3	.59	.60	-			
4	.59	.59	.13	-		
5	.77	.77	.75	.75	-	
6	.81	.82	.73	.74	.80	-
7	.87	.86	.86	.88	.93	.90

Pairwise alignment:
Calculate distance matrix



Rooted neighbor-joining
tree (guide tree)

指导树的构建

```

-----VHLTPEEKSAVTALWGKVN--VDEVGGGEALGRLLVVYWTQRFESFGDLST
-----VQLSGEEKAAVLAWLWDKVN--EEEVGGEALGRLLVVYWTQRFDSFGDLSN
-----VLSPADKTNVKAAWKGKVGAHAGEYGAEEALERMFLSFETTTKTYFPHFDSL-
-----VLSAADKTNVKAAWSKVGHHAGEYGAEEALERMFGLGFETTKTYFPHFDSL-
-----VLSEGEWQLVLHVWAKVEADVAHGQDILIRLFKSHPTELEKFDRFKHLKT
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
-----GALTESQAALVKSSWEENANIPKHTRFFILVLEIAAAKDLFSFLKGTE

```

Progressive
alignment:
Align following
the guide tree

渐进比对

```

PDAVMGNPKVKAHGKKVLGAFSDGIAHLD----NLKGTFAVLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVHLFGEGVHLD----NLKGTFAALSELHCEKLHVDPENFRL
---HGSAQVKGHGKKVADALTNAVAHVD---DMPNALSALSDLHAHKLRLVPVNFKL
---HGSAQVKAHGKKVGDALTLAGVHLD---DLPGALSNLSDLHAHKLRLVPVNFKL
EAEMKASEDLKKHGVTVLTALGAIKKKG---HHEAEELPLAQSHATKKHPIKYLEF
ADQLKKSSADVRWHAERIINAIVNDAVASMDDT---EKMSMKLRLDSGKHAKSFQVDPQYFKV
VP---QNNPELOAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV

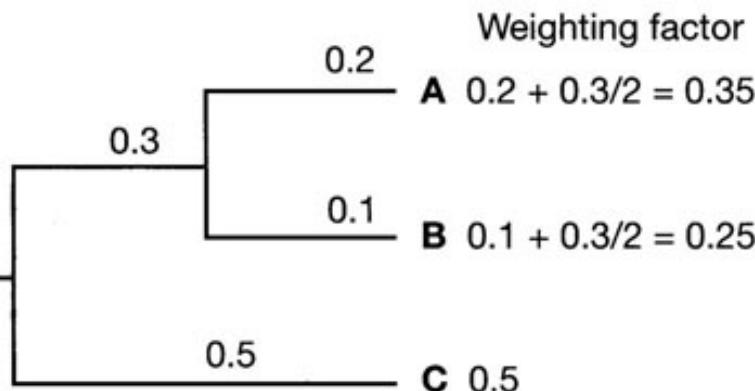
```

```

LGNVLVLCVLAHHFGKEETPPVQARYOKVVAGVANALAHKYH-----
LGNVLVUVLARHFGKDFTPELQASYQKVVAGVANALAHKYH-----
LSHCLLVTLA AHLPAEFTPAVHASLDKF LASVSTVLT SKYR-----
LSHCLLSTLAVHLPNDFTPAVHASLDKF LSSVSTVLT SKYR-----
ISEAI IHVLHSRH PGDFGADAQGAMNKALEFRKDIAAKYKELGYQG
LAAVIADTVAA G-----DAGFEKLMSMICILLRSAY-----
VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---

```

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.....

ClustalW的打分原则

Column in alignment 2

Sequence C (weight c)L.....
Sequence D (weight d)V.....

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



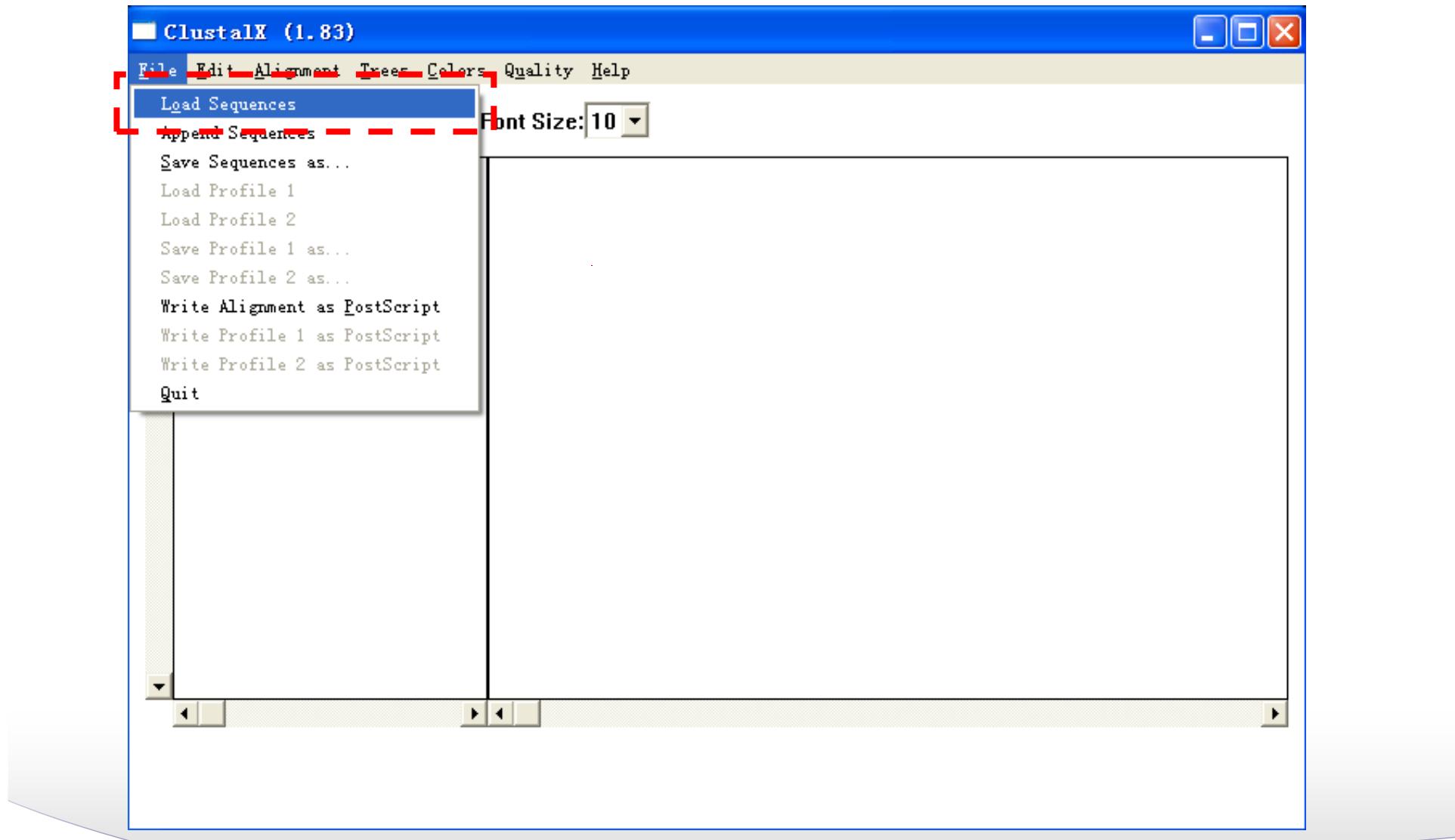
ClustalX: 使用指南

□ FASTA序列格式，多序列

```
>ScTPK1
MSTEEQNGGGQKS LDDRQGEESQKG ETSERETTATESGNESKS VEVKEGGETQEKP KQPHV
TYYNEEQYKQFIAQARVTSGK YSLQDFQ IRLT LGTGSFGRVHLIRSRHNGRYYAMKVLKK
EIVVRLKQVEHTN DERLMLSIVTHPFI IIRMWGTFQD AQQIFMIMDYIEGGELFSLLRKS Q
RFNPVAKFYAAEVCLALEY LHSKDI IYRDLKPE NILLDKNGH IKITDFGFAKYVPDV TY
TLCGTPD YIAPEVVSTK PYNKS IDWW SFGI LIYEMLAGYTPF YDSNTMKT YEKILNAEL R
FPPFFNEDVKD LLSRLITRD LSQR LGNL QNGTEDV KNHPWF KEV VWEK LLSRN IETPYEP
PIQQGQGDT SQFD KYPE EDINYGVQ GEDPYADLF RDF
>ScTPK2
MEFVAERAQPVGQT IQQQN VN TYGQGV LQPHD LQQR QQQQQ QRHQQL LSQL P QKSL V
SKGKYTLHDFQ IMLRT LGTGSFGRVHLVRSV HNGRYYAI KV LKKQ QV VKM KQVE HTNDERR
MLKLVEHPFL I RMWGT FQD ARNIFM VMDYIEGGELFSLLRKS QRF PN PVAKF YAAEVILA
LEYLHAHN IIYRDLKPE NILL DRNGH IKITDFGFAKE VQTV TWTLCGTPD YIAPEV ITTK
PYNKSVDWWSLGV LIYEMLAGYTPF YDTTPMKT YEKILQGKV VVPPYFHPD VVD LLSKLI
TADLTRRIG N LQSGS RDIA KHPWF SEV VWERL LAKD IE TPYE PPITS GIGD TS LFDQ YPE
EQLDYGIQGDDPYAEYFQDF
>ScTPK3
MYVDPMNNNEIRKLSITAKTETTPDNVGQDIPVNAHSVHECSSNTPVEINGRNSGKLKE
EASAGICLVKKPMLQYRDTSGK YLSDFQ IRLT LGTGSFGRVHLIRSNHNGRFYALKTLK
KHTIVKLKQVEHTN DERRML SIVSHPF IIRMWGT FQDSQ QVFM VMDYIEGGELFSLLRKS
QRFPNPVAKFYAAEVCLALEY LHSKDI IYRDLKPE NILLDKNGH IKITDFGFAKYVPDV T
Y TLCGTPD YIAPEVVSTK PYNKSVDWWSFGI LIYEMLAGYTPF YNSNTMKT YENI LNAEL R
KFPPFFHPDAQDLLKKLITRD LSERLG N LQNGSE DVKNHPWF NEVI WEKLLARYIETPYE
PPIQQGQGDT SQFD RYPEEEFNYGIQGEDPYMDLMKEF
>Cekin-1
MPTRL DIVGNLQFSSSTDNGDEDQ EADVTACFVLPS PSSFSKLS I LDDPVEDFKEFLDKA
REDFKQRWENPAQNTAC LDDFDRIK TLGTGSFGRVMLVHKQSGNYYAMK I LDKQKV VKL
KQVEHTLNEKRI LQAIDFPFLVNMTFSFKD NSNLYMVLE FISGGEMF SHLRRIGRFSEPH
SRFYAAQIVLA FEYLHS LD L IYRDLKPE NLLIDSTGYLK I TDFGFAK RVKG RTWTLCGTP
EYLAPEIILSKGYNKA DW W ALGV LIYEMAAGYPPFFADQPIQIYEKIVSGKVKFPSHFS
NELKD LKNL LQV DLT KRYGN LKNGVADIK NHKWFGSTDWIAI YQK KITPPSF SKGESNG
RLFEALYPRVDGPADTRHFVEEVQEPTEFVIAATPQ LEELF VEF
```



导入序列文件





执行比对

ClustalX (1.83)

File Edit Alignment Trees Colors Quality Help

Multiple

- Do Complete Alignment
- Produced 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

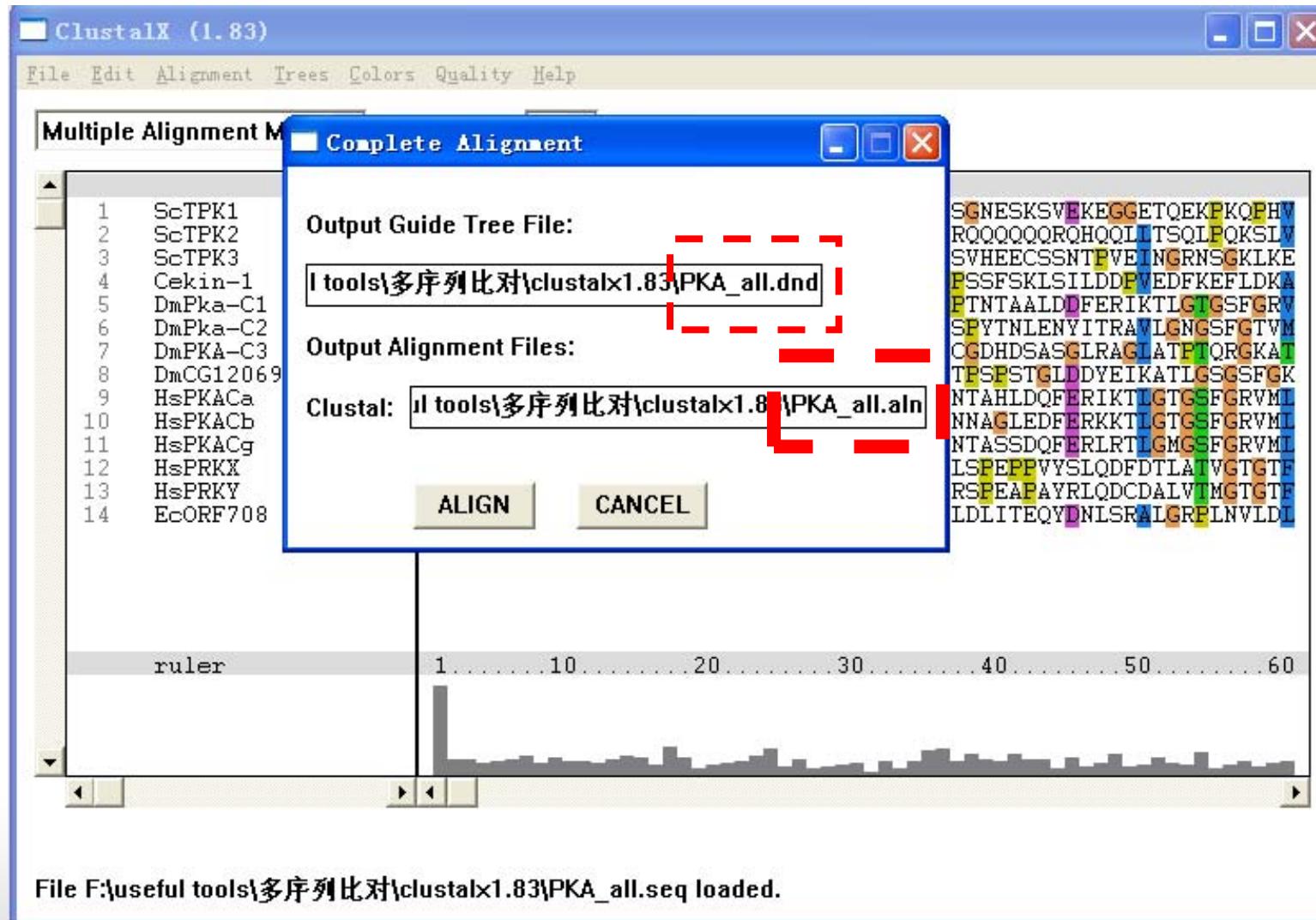
DDROQGEESQK**G**EISERETTATESGNESKS**V**EKEGG**E**TQE**K**PKQ**F**H**V**
IQQQN**V**NNTY**G**Q**G**V**L**Q**F**H**D**I**Q**QRQQQQ**Q**R**H**Q**Q**LI**T**SQL**P**Q**K**SL**V**
S**I**TAK**T**ET**T**PD**N**VG**Q**D**I**PF**V**NA**H**SV**H**EE**C**SS**N**T**P**WE**I**NG**R**NS**G**K**L**KE
S**S**STD**N**G**D**ED**Q**E**A**DT**A**CF**V**IP**S**P**S**FS**K**LS**I**L**D**DP**W**ED**F**KE**F**LD**K**A
A**E**TV**K**EF**L**EQ**A**KE**F**E**D**KW**R**RN**F**T**N**TA**A**LD**D**F**E**RI**K**T**I**GT**G**SF**G**R**V**
D**Y**N**V**I**L**DN**M**S**R**E**E**ER**M**N**H**QT**Q**SP**T**N**L**EN**Y**IT**R**AV**L**GN**G**SF**G**T**V**
CRVSASVF**F**AN**F**CG**G**I**L**YSS**W**K**L**I**C**G**D**H**D**S**A**GL**R**AG**L**AT**P**T**Q**R**G**K**A**
KV**D**Y**I**L**I**L**D**KL**R**ED**F**N**K**K**F**T**N**TP**S**P**S**T**G**LD**D**YE**E**IK**A**T**L**G**S**G**F**K**G**
SV**K**EFLAK**A**K**E**DF**L**KK**M**E**S**P**A**Q**N**T**A**HL**D**Q**F**E**R**I**K**T**I**GT**G**SF**G**R**V**
ML**S**V**K**EFLAK**A**K**E**DF**L**KK**M**E**N**P**T**Q**N**N**A**LED**F**E**R**K**K**T**I**GT**G**SF**G**R**V**
ML**S**V**N**EFLAK**A**K**G**DF**L**Y**R**M**G**N**P**A**Q**N**T**A**S**SD**Q**F**E**R**I**L**R**T**I**GM**G**SF**G**R**V**
ML**S**DS**R**K**V**A**E**E**T**PD**G**A**P**A**C**F**S**P**E**AL**S**P**E**PP**V**Y**S**L**Q**D**F**D**T**LA**T**VG**T**G**T**
ML**S**NS**R**E**V**T**E**DA**A**D**W**A**P**A**C**F**S**P**E**AR**S**P**E**A**P**R**L**Q**D**C**D**AL**V**T**M**G**T**G**T**
ML**S**I**Y**Q**T**I**F**G**H**F**E**WD**G**D**A**ARD**C**N**Q**R**L**DI**T**EQ**Y**D**N**L**S**R**A**LG**R**PL**N**VL**D**I

ruler 1 10 20 30 40 50 60

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



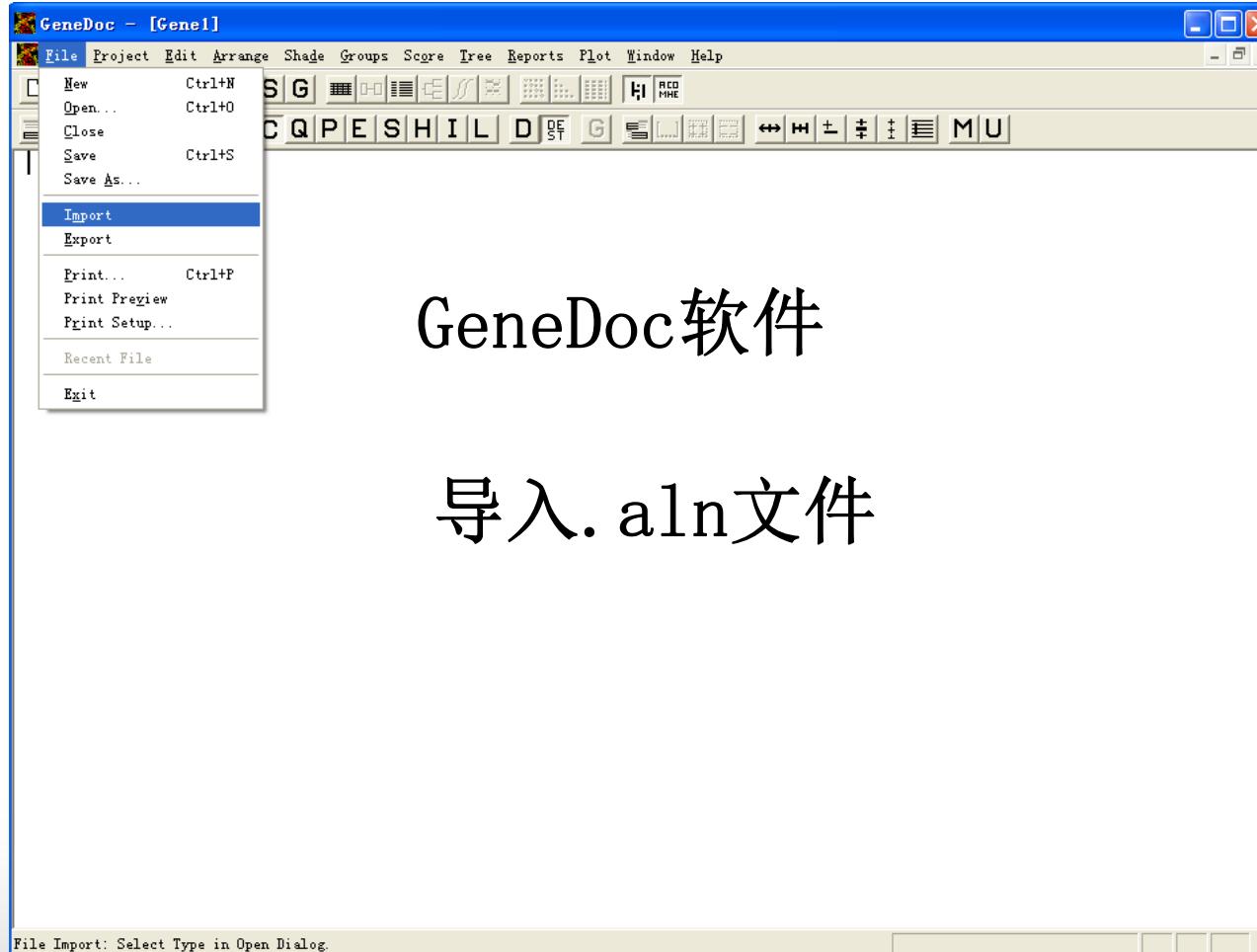
文件导出





多序列比对：结果处理

□ GeneDoc, BioEdit等软件

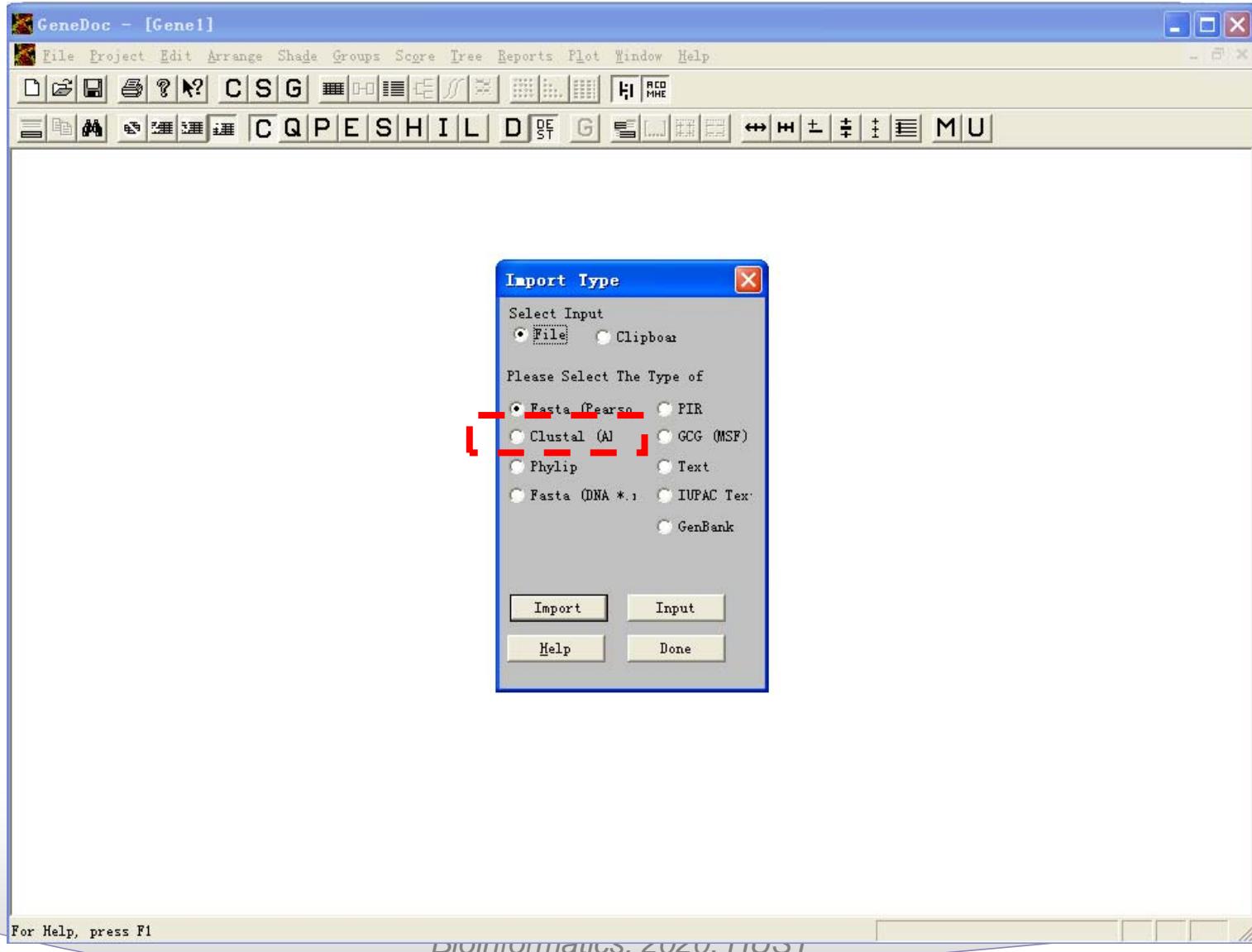


GeneDoc软件

导入. aln文件



选择文件格式



For Help, press F1

BIOMINFORMATICS, 2020, TU DORTMUND



成功导入文件

GeneDoc - [Gene1]

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

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

Hs PRK1 : ADDCFSEPKR-----AIRKQPCDA-DVINGIGIEGRVWVQVNRVIAKRE : 74

Dm PKA-C3 : ADDATHDSSESIEEDDGNETDDEDDDESEEESSVQTAKGVRKYHLDYQI-IKTVGTTFGRVCLCRDRISEKY : 299

EcORF708 : --VSQPDPRELIEQCAFYRLIGEFDTHLSPVPRPMYLVSNHRVLNDENQPFQHWQNQPYAGAGLAHKRSRRYE : 244
t g g 5g v 6

DmPka-C2 : YAAKNMMSKEDLVRL-----KQWAEPVHNEKHVLNAARFPEFLITYLVDSTKC-FDYLYLILPLVNGGEELFSY : 133

DmCG12069 : YASKQLSKDQIVKRT-----KQVSHVMSEKNVLRSMTFPTVNLIASYKD-FDSLYLVPLIGGGELFTY : 135

Hs PKACa : YAMKILDQKVVKL-----KQIEHTLNEKRILQAVNPFPLVKEFSEKD-NSNLYMVMMEYVPGGEMFSH : 132

Hs PKACb : YAMKILDQKVVKL-----KQIEHTLNEKRILQAVNPFPLVREYAFKD-NSNLYMVMMEYVPGGEMFSH : 132

Hs PKACg : YAMKILNKQKVVKM-----KQVEHILNEKRILQAIIDFPFLVKLQFSEKD-NSYLYLVMEYVPGGEMFSR : 132

DmPka-C1 : YAMKILDQKVVKL-----KQVEHTLNEKRILQAIQFPFLVSLRYHFKD-NSNLYMVLLEYVPGGEMFSH : 134

Cekin-1 : YAMKILDQKVVKL-----KQVEHTLNEKRILQAIIDFPFLVNMTEFSKD-NSNLYMVLLEISGGEMFSH : 169

ScTPK1 : YAMKVLKKEIVVRL-----KQVEHTNDERILMSIVTHPFIRMWGTFQD-AQQIFMIMDYIEGGELFSL : 175

ScTPK3 : YALKTLKHTIVK-----KQVEHTNDERRMLSISSHPEIIRMWGTFQD-SQQVFMVMDYIEGGELFSL : 176

ScTPK2 : YAIVKVLKQQVVKM-----KQVEHTNDERRMLKLVEHPFLIRMWGTFQD-ARNIFMVMMDYIEGGELFSL : 158

Hs PRKX : FALKVMSIPDVIRL-----KQEQHVHNEKSVLKEVSHPFIRLFWTWHD-ERFLYMLMEYVPGGELFSY : 137

Hs PRKY : FALKVMSIPDVIRR-----KQEQHVHNEKSVLKEVSHPFIRLFWTWHE-ERFLYMLMEYVPGGELFSY : 137

Dm PKA-C3 : CAMKILAMTEVIRL-----KQIEHVKNERNNLIREIRHPPVISLEWSTKD-DSNLYMIFDIVCGGEELFTY : 362

EcORF708 : FGEDYVCKFFYYDMPHGILTAEESQRNKHHLNEIKFLTQPPPAGEADAPAVLAHGENAQSGRWLVMEKLPG-RLLSD : 318
a k 6 k 4q h E L pf d 566 6 Gge6f3

DmPka-C2 : HRRVRKFNEKHFYAAQVALALEYMHKMLMYRDLKPENILLDQRCGYIKITDFG-FTKRVDGRTSTLCGTB--- : 204

DmCG12069 : HRKVRKFTEKQARFYAAQVFLALEYLHHCSLLYRDLKPENIMMDKNGYLKVTDFG-FAKKVETRTMTLCGTB--- : 206

Hs PKACa : LRRIGRFSSEPHARFYAAQIVLTFEYLSLIDLIVRDLKPENLLIDQQGYIQVTDFG-FAKRVKGRTWTLCGTB--- : 203

Hs PKACb : LRRIGRFSSEPHARFYAAQIVLTFEYLSLIDLIVRDLKPENLLIDHQGYIQVTDFG-FAKRVKGRTWTLCGTB--- : 203

Hs PKACg : LQRVGRFSSEPHACFYAAQVVLAVQYlhsldliHRDLKPENLLIDQQGYLQVTDFG-FAKRVKGRTWTLCGTB--- : 203

DmPka-C1 : LRKVGRFSSEPHSRFYAAQIVLAFELYHYIDLIVRDLKPENLLIDSQGYLKVTDFG-FAKRVKGRTWTLCGTB--- : 205

Cekin-1 : LRRIGRFSSEPHSRFYAAQIVLAFELYHSLDLIVRDLKPENLLIDSTGYLKIKITDFG-FAKRVKGRTWTLCGTB--- : 240

ScTPK1 : LRKSQRFPNPVAKFYAAEVCLALEYLHSKDIYRDLKPENILLDKNCNHIKITDFG-FAKYVPDVYTLCGTB--- : 246

ScTPK3 : LRKSQRFPNPVAKFYAAEVCLALEYLHSKDIYRDLKPENILLDKNCNHIKITDFG-FAKYVPDVYTLCGTB--- : 247

ScTPK2 : LRKSQRFPNPVAKFYAAEVVLALEYLHANNIYRDLKPENILLDRNCHIKITDFG-FAKEVQTVTWTLCGTB--- : 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 Alignment as Text
Copy Consensus as Fasta
Copy Consensus as Prosite
Select Columns Ctrl+L
Delete All Data Ctrl+D
Copy Data Between Segs
Residue Edit Mode Ctrl+U
Clear Gap Columns
Clear Max Comments

L D P F G L E M U

KQEQHVHNEKSVLKEVSHPFLIRLFMTWHD-ERFLYMLMEYVPGGELESY : 137
KQEQHVHNEKSVLKEVSHPFLIRLFMTWHE-ERFLYMLMEYVPGGELESY : 137
KQIEHVKNERNILREIRHPFVISLEWSTKD-DSNLYMIFDYVCGGELETY : 362
ESQRNKHNLHNEIKFLTQPPAGFDAPAVLAHGENAQSGWLVMKLPG-RLLSD : 318

4q h E L pf d 566 6 Gge6f3

400 * 420 * 440 *

LEYMHKKMHLMYRDLKPENILLDQRGYIKITDFG-FTKRVDGRTSTLCGT P--- : 204
LEYLHHCSLLYRDLKPENIMMDKNGYLKVTDFG-FAKKVETRTMTLCGT P--- : 206
FEYLHSLDLIYRDLKPENILLIDQQGYIQVTDFG-FAKRVKGRTWTLCGT P--- : 203
FEYLHSLDLIYRDLKPENILLIDHQGYIQVTDFG-FAKRVKGRTWTLCGT P--- : 203
QYLHSLDLIHRDLKPENILLIDSQGYLQVTDFG-FAKRVKGRTWTLCGT P--- : 203
FEYLHYLDLIYRDLKPENILLIDSQGYLKVTDFG-FAKRVKGRTWTLCGT P--- : 205
FEYLHSLDLIYRDLKPENILLIDSQGYLKVTDFG-FAKRVKGRTWTLCGT P--- : 240
LRKSQRFPNPVAKFYAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFG-FAKYVPDVYTLCGT P--- : 246
LRKSQRFPNPVAKFYAAEVCLALEYLHSKDIIYRDLKPENILLDKNGHIKITDFG-FAKYVPDVYTLCGT P--- : 247
LRKSQRFPNPVAKFYAAEVILALEYLHANIIYRDLKPENILLDRNGHIKITDFG-FAKEVQTVWTLCGT P--- : 229
LRNRGRFSSTTGLFYSAEIIICAIYEYLHSKEIVYRDLKPENILLDRGHIKLTDGF-FAKKLVDRWTLCGT P--- : 208
LRNRGHFSSTTGLFYSAEIIICAIYEYLHSKEIVYRDLKPENILLDRGHIKLTDGF-FAKKLVDRWTLCGT P--- : 208
LRNAGKFTSQTSNFYAAEIVSALEYLHSQIVYRDLKPENILLINRDGHKLKITDFG-FAKKLQRDRWTLCGT P--- : 433
MLAAG--EEIDREKILGSILLRSLAALEKQGFWHDDVRPWNVMVDARQHARLIDFGSIVTTPDQCSMPTNLVQSFF : 391

r f fyaa 6 a y6h yrD64PeN6661 g 6tDFG fak 3 tlcgt p

460 * 480 * 500 * 520

EYLAPEIIVQIRPYN-KSVDWWAFGILVYEFVAGRSPEFAIHNR----- : 245
EYLPEIIIQSKPYG-TSVDWWAFGVLVFEFVAGHSPESAHNR----- : 247
EYLAPEIIILSKGYN-KAVDWWALGVLIYEMAAGYPPFFADQ----- : 243
EYLAPEIIILSKGYN-KAVDWWALGVLIYEMAAGYPPFFADQ----- : 243
EYLAPEIIILSKGYN-KAVDWWALGVLIYEMAVGFPPFYADQ----- : 243
EYLAPEIIILSKGYN-KAVDWWALGVLYEMAAGYPPFFADQ----- : 245
EYLAPEIIILSKGYN-KAVDWWALGVLIYEMAAGYPPFFADQ----- : 280
DYIAPEVVSTKPYN-KSIDWWWSFGILYIYEMLAGYTPEYDSN----- : 286

Select an area to copy.

ScTPK2:



比对结果的美化和后处理

Cekin-2	SGGRRRTG I S S A E	89
ScBCY1	NAQRRT I S V SGE	149
DmPka-R2	ASSRRK I S M FAE	88
HsPRKAR2A	N--RRV I S V CAE	103

Cekin-2	DYFGEIALLL D R P R A A T V V A K TH	329
ScBCY1	DYFGEVALLN D L P R Q A T V T A T KR	386
DmPka-R2	QYFGELALV T HR P R A A S V Y ATGG	329
HsPRKAR2A	QYFGELALV T N K P R A A S A Y Y AVGD	356

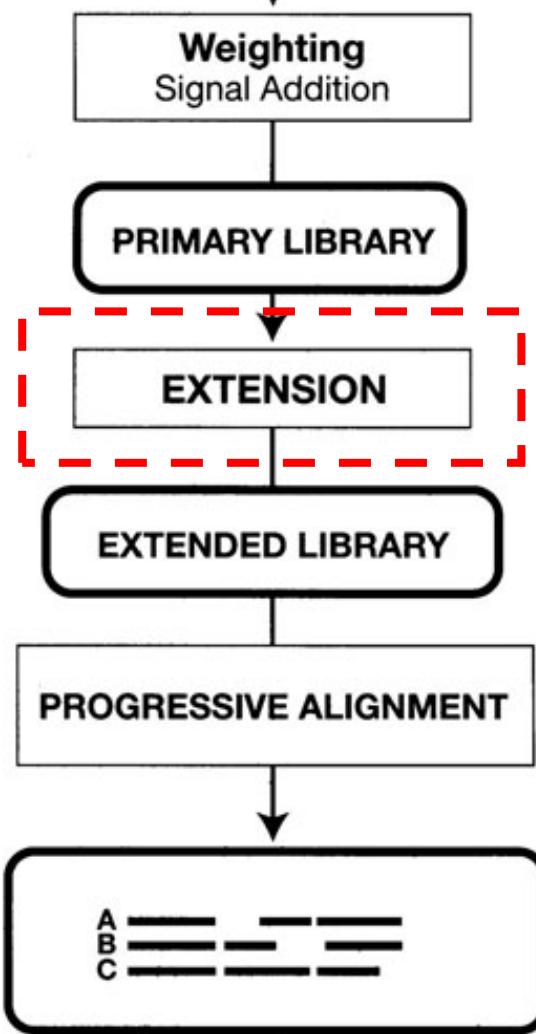
T-Coffee



- 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



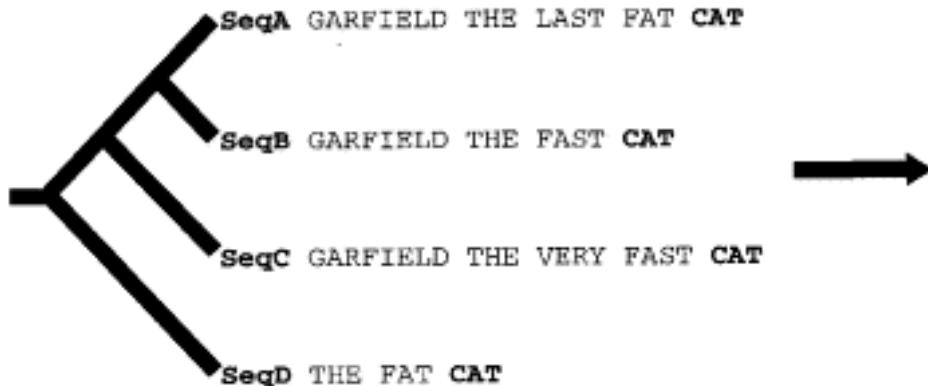
同时进行全局和局部的
双序列比对



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

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

c) Extended Library for seq1 and seq2

```

SeqA GARFIELD THE LAST FAT CAT
      |||||||   |||   ||||   |||           Weight = 88
SeqB GARFIELD THE FAST CAT

```

SeqA GARFIELD THE LAST FAT CAT

```

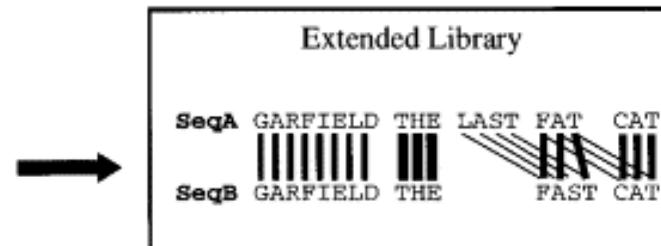
SeqC GARFIELD THE VERY FAST CAT    Weight = 77
      |||||||   |||   |||||   |||    |
SeqB GARFIELD THE           FAST CAT
      |||||||   |||       |||||   |||    |

```

```

Seq1 GARFIELD THE LAST FAT CAT
          |||   |||   |||
SeqD           THE      FAT CAT    Weight = 100
          |||   |||   |||

```



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:
- 若Seq1, 2先比对,
再加入Seq3:
- Seq1, 3先比对,
再加入Seq2:
- Seq2, 3先比对,
再加入Seq1:

Seq1: ARKCV

Seq2: ARCV

Seq3: AKCV

ARKCV

AR-CV

A-KCV

ARKCV

A-RCV

A-KCV

ARKCV

AR-CV

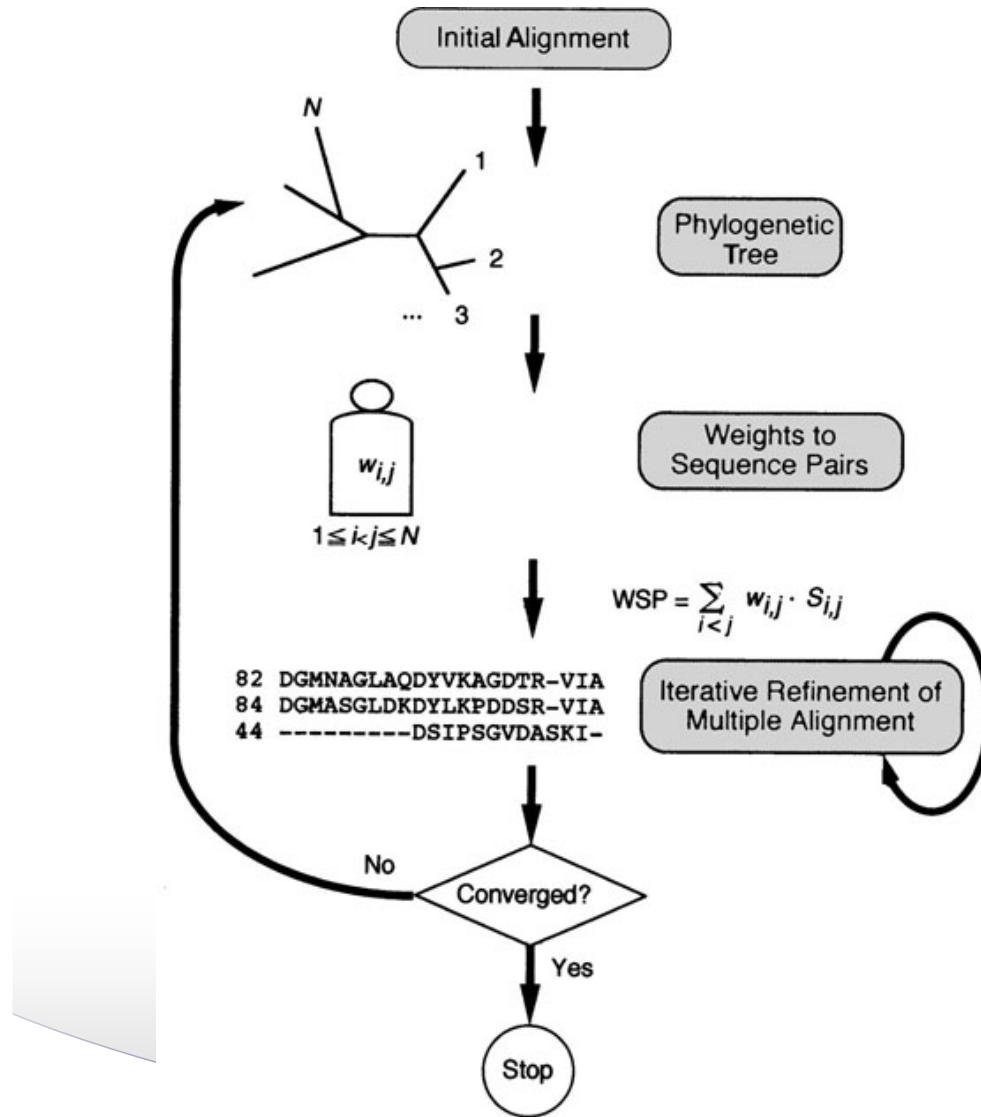
AK-CV

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://dalign.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)

I	A	V	L	F	A	E	D	
						/	/	
L	A	V	I	F	G	S		
W	D	D	V	T	F	D	A	E

A

I	A	V	L	F	A	E	D	
						/	/	
L	A	V	I	E	G	S		
W	D	D	V	T	F	D	A	E

B

I	A	V	L	F	A	E	D	
						/	/	
L	A	V	I	F	G	S		
W	D	D	V	T	F	D	A	E

C

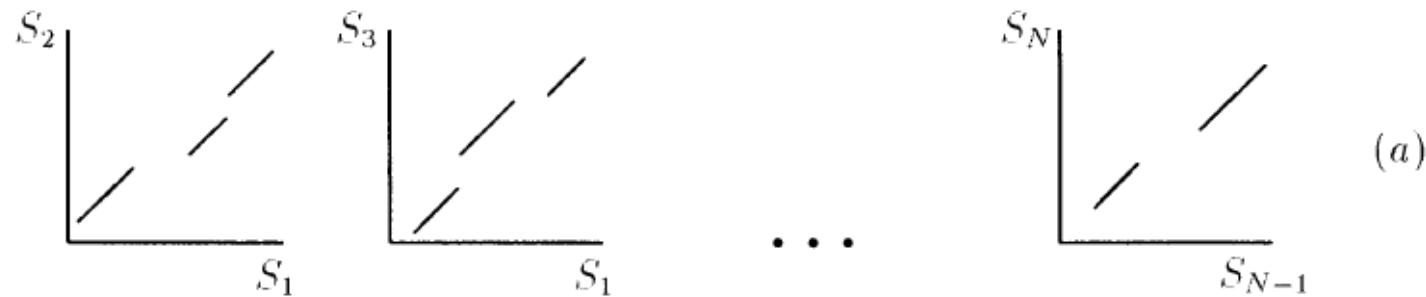
一致的 (Consistent)

I	A	-	V	L	F	-	A	E	d
		-				-			
L	A	-	V	I	F	-	G	S	-
w	d	d	V	T	F	d	A	E	-

D

最终的比对结果

DIALIGN: 算法流程

 \mathcal{M}_1  \mathcal{M}_1  \mathcal{M}_2

Overlap weights

Sort diagonals

(b)

(c)

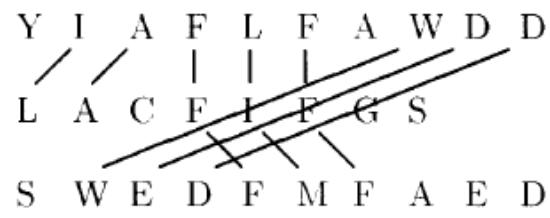
Consistency!

(d)



迭代过程

1. iteration step



$D_1 \quad I \quad A$ $D_2 \quad F \quad L \quad F$
 $L \quad A$ $F \quad I \quad F$

 $D_3 \quad W \quad D \quad D$ $D_4 \quad F \quad I \quad F$
 $W \quad E \quad D$ $F \quad M \quad F$

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



$D_5 \quad F \quad L \quad F \quad A \quad W \quad D$
 $F \quad M \quad F \quad A \quad E \quad D$

M1={D1, D2, D4, D5}



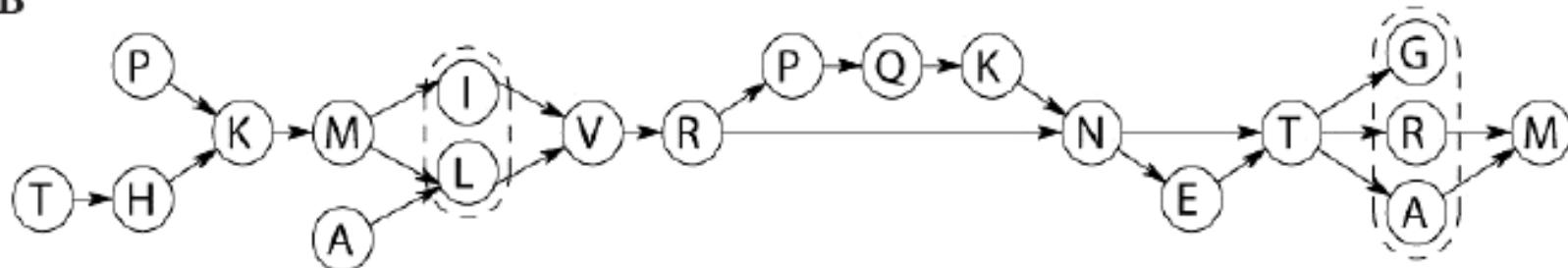
部分有向图算法：POA

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

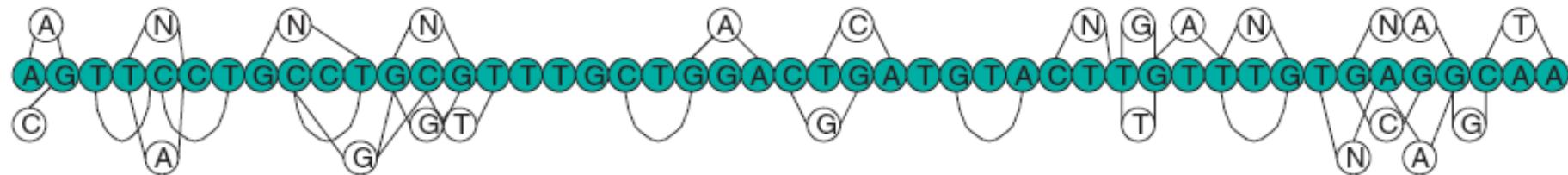
A

.	.	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

B



(a)

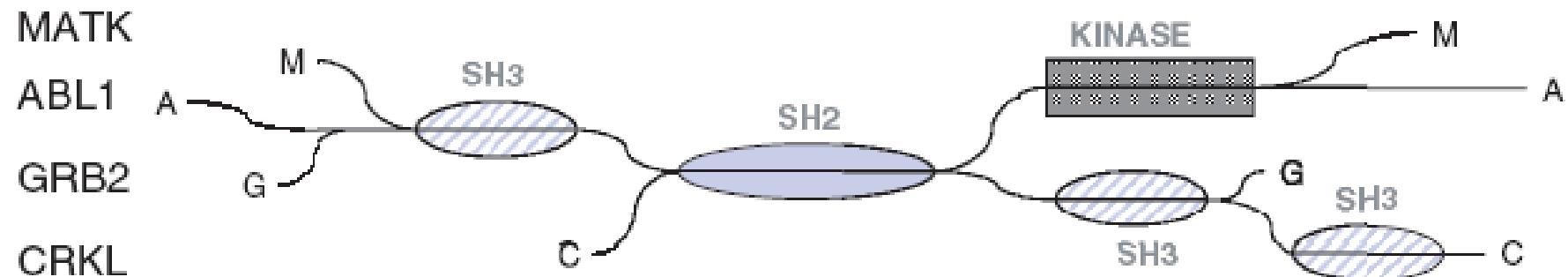


(b)

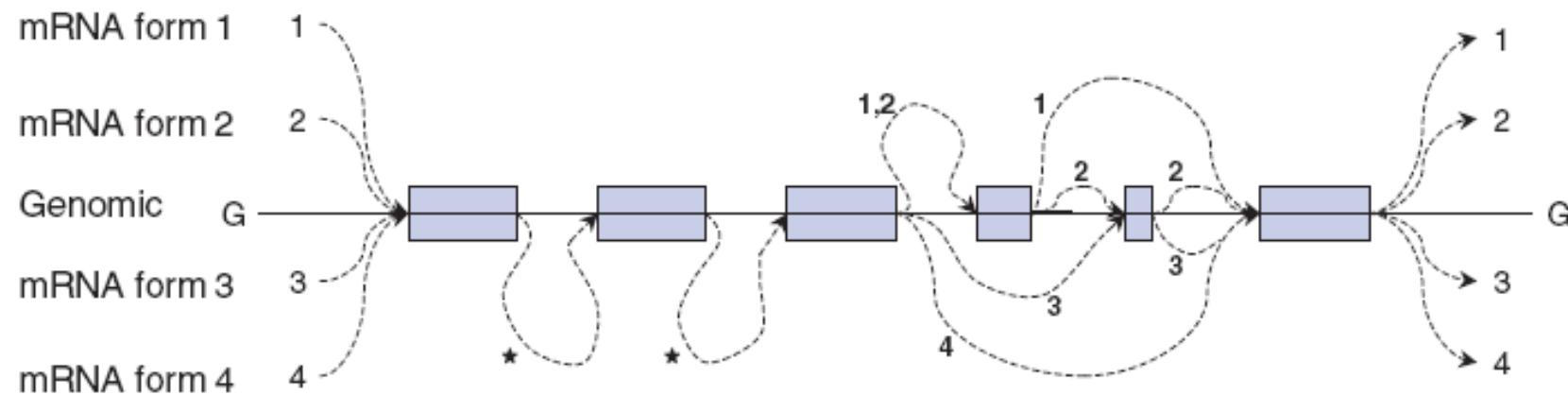
CONSENS1TGTA <u>QNT</u> .GTTTGTGAGG.CTA
CONSENS0	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S663801	A.GTTCCCTGC.TCGGTTGCTGGACTTATGTACTT.GTTTGTGAGG.CAA
Hs#S337687	AAGTTCCCTGC.TCGGTTGCTGGACTGATGTACTT <u>GTTTGTGAGG</u> CNAAGGCAA
Hs#S629177	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S672957	A.GTTCCCTGC.TCGGTTGCT.....
Hs#S672182	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTT.....
Hs#S674099	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S196113	A.GTT <u>N</u> CTG <u>N</u> T <u>G</u> CGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S994400GTA <u>QNT</u> .GTTTGTGAGG.CTA
Hs#S550772	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S80460	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S39701	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1988018	A.GTTCCCTGC.TG <u>G</u> TTT <u>G</u> CTGGACTGATGTACTT. <u>G</u> ATTGTGAGG.CAA
Hs#S341915	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1794113	A.GTTCCCTGC.TCGGCTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S4698	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGT <u>CCGG</u> .CAA
Hs#S813765	A.GT <u>U</u> CC <u>T</u> GC.GCGTTGC.GGACGGATGTACTT.GT <u>U</u> .GTGAGG.CAA
Hs#S1184845G.CAA
Hs#S1577463GG.CAA
Hs#S914987CTGATGTACTT.GT <u>U</u> .GTGAGGCAA
Hs#S1985364	A.GTTCCCTGC.TCGGTTGCTGGACTGATGTACTT.GTTTGTGAGG.CAA
Hs#S1465644	..GTTC.TGC <u>U</u> CTGC <u>U</u> TTGCTGA <u>ACT</u> GATGTACTT.GTTAGT.AAG.CAA
Hs#S1850471	C.GTTACTGC.GCGTTGCTGGACTCATG.AC <u>TT</u> GTT <u>NG</u> T.AGG.CAA



激酶的多序列比对



mRMA的基因组定位





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

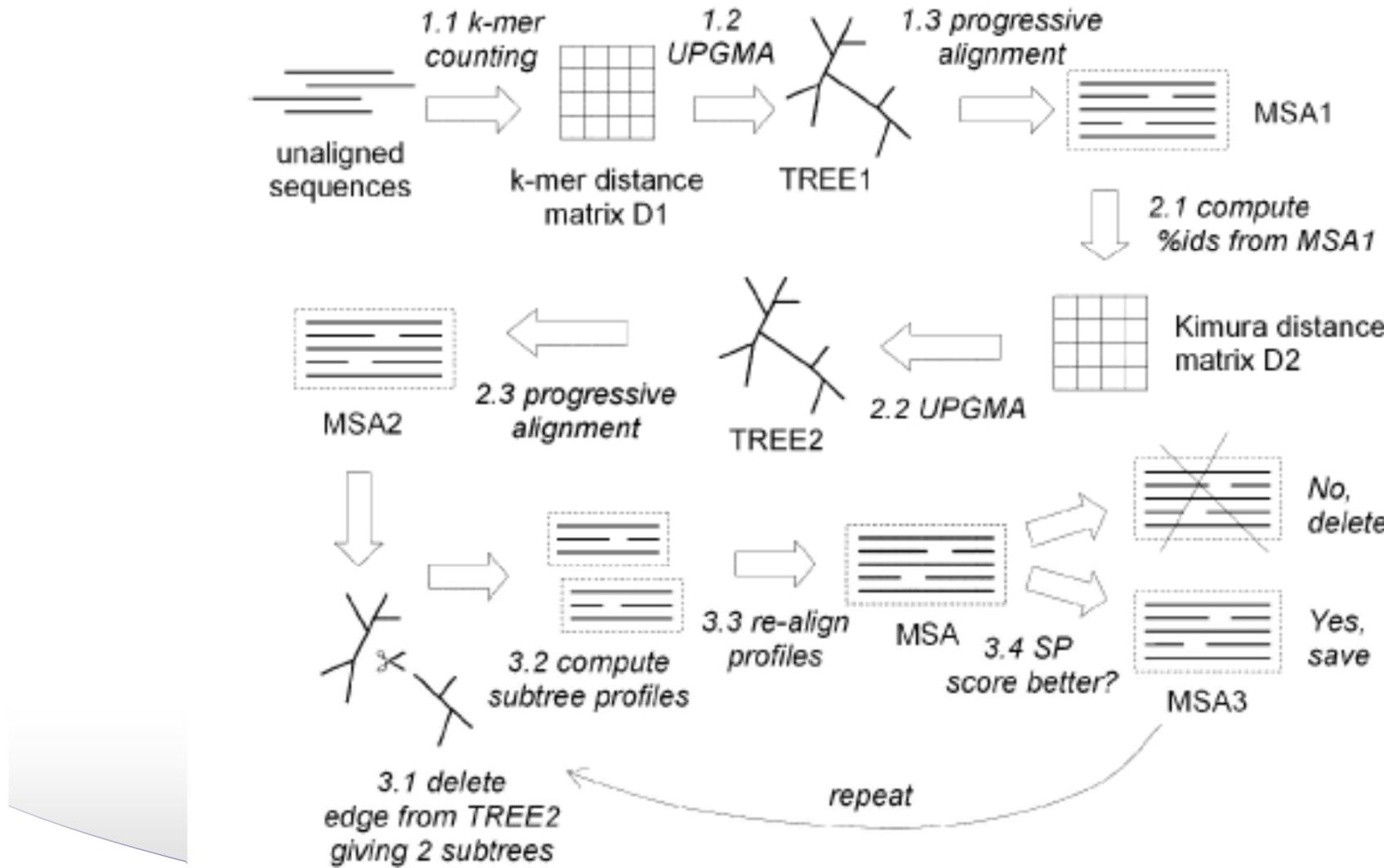


MUSCLE (3)

□ 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

MUSCLE

MUSCLE has been cited by
34,183 papers
[Google scholar](#)

Last updated 10 Feb 2020

Downloads

Documentation

Support

USEARCH

Ultra-fast sequence analysis



10 - 1,250x BLAST
1 - 1,000x CD-HIT

Popular multiple alignment software

MUSCLE is one of the most widely-used methods in biology. On average, MUSCLE is cited by ten new papers every day.

Fast, accurate and easy to use

MUSCLE is one of the best-performing multiple alignment programs according to published benchmark tests, with accuracy and speed that are consistently better than CLUSTALW. MUSCLE can align hundreds of sequences in seconds. Most users learn everything they need to know about MUSCLE in a few minutes—only a handful of command-line options are needed to perform common alignment tasks.

Papers

There are two papers. The first (NAR) introduced the algorithm, and is the primary citation if you use the program. The second (BMC Bioinformatics) gives more technical details, including descriptions of non-default options.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput
Nucleic Acids Res. **32**(5):1792-1797 [[Link to PubMed](#)].

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity
BMC Bioinformatics, **(5)** 113 [[Link to PubMed](#)].



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



MUSCLE: 使用说明

命令提示符

```
E:\>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
```

DIVISION OF COMPUTATIONAL BIOLOGY, ZHEJIANG UNIVERSITY



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.

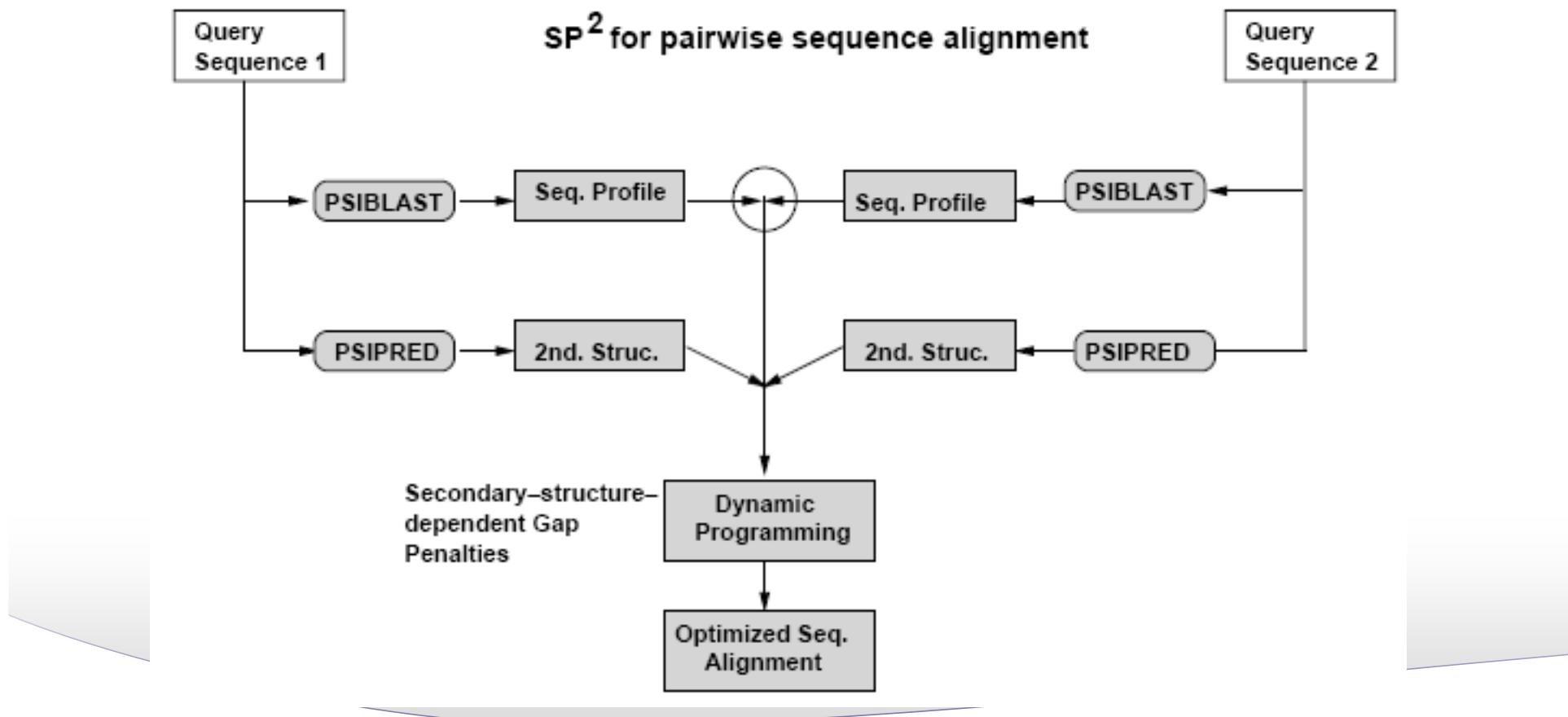
A full description of the algorithms used by Clustal Omega is available in the Molecular Systems Biology paper [Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega](#). Latest additions to Clustal Omega are described in [Clustal Omega for making accurate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

Webservers **Download Clustal Omega**

结构特征



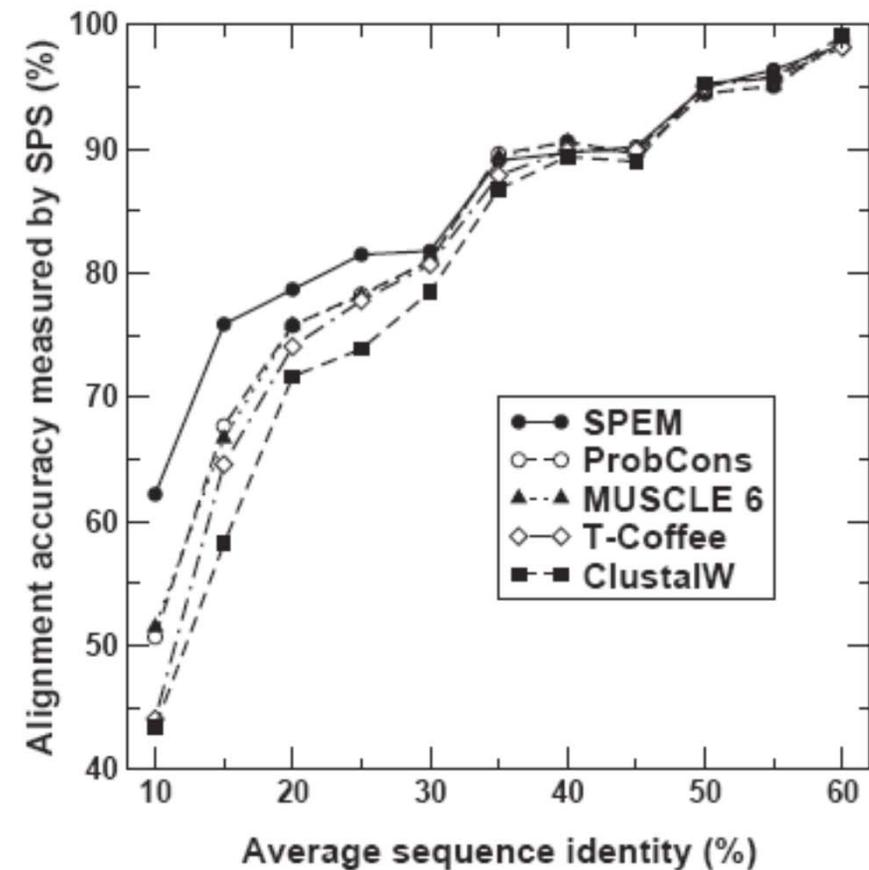
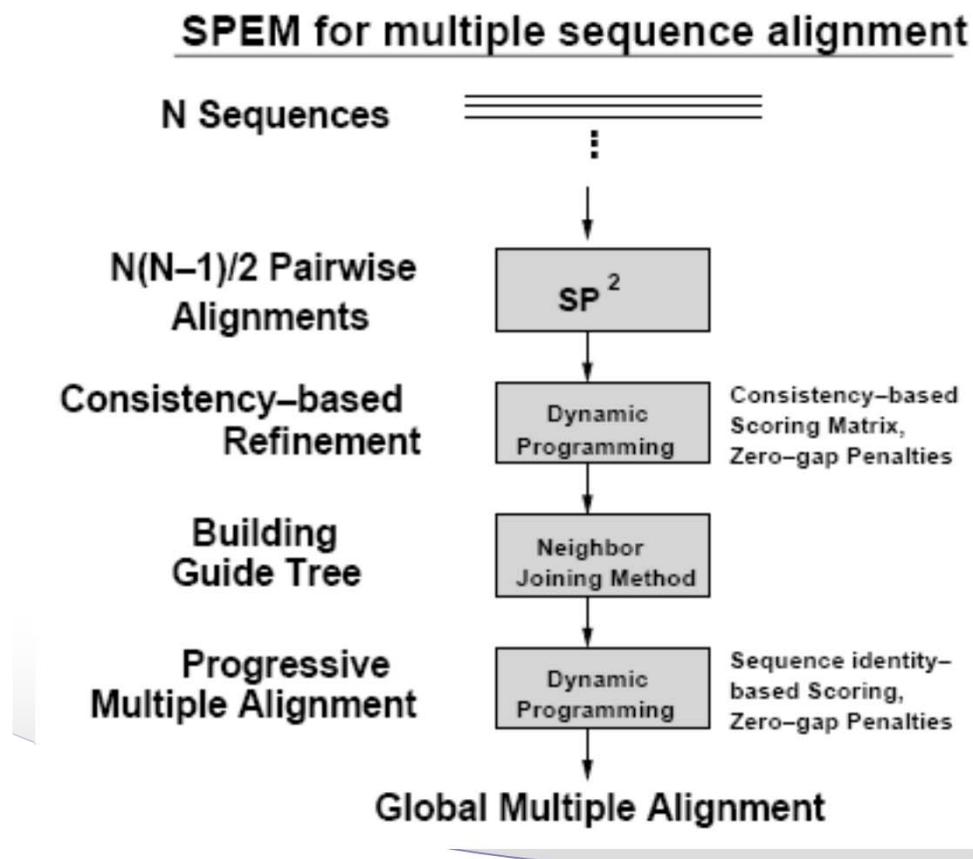
- 2005, SPEM
- 空位罚分：结合蛋白质的二级结构信息



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%

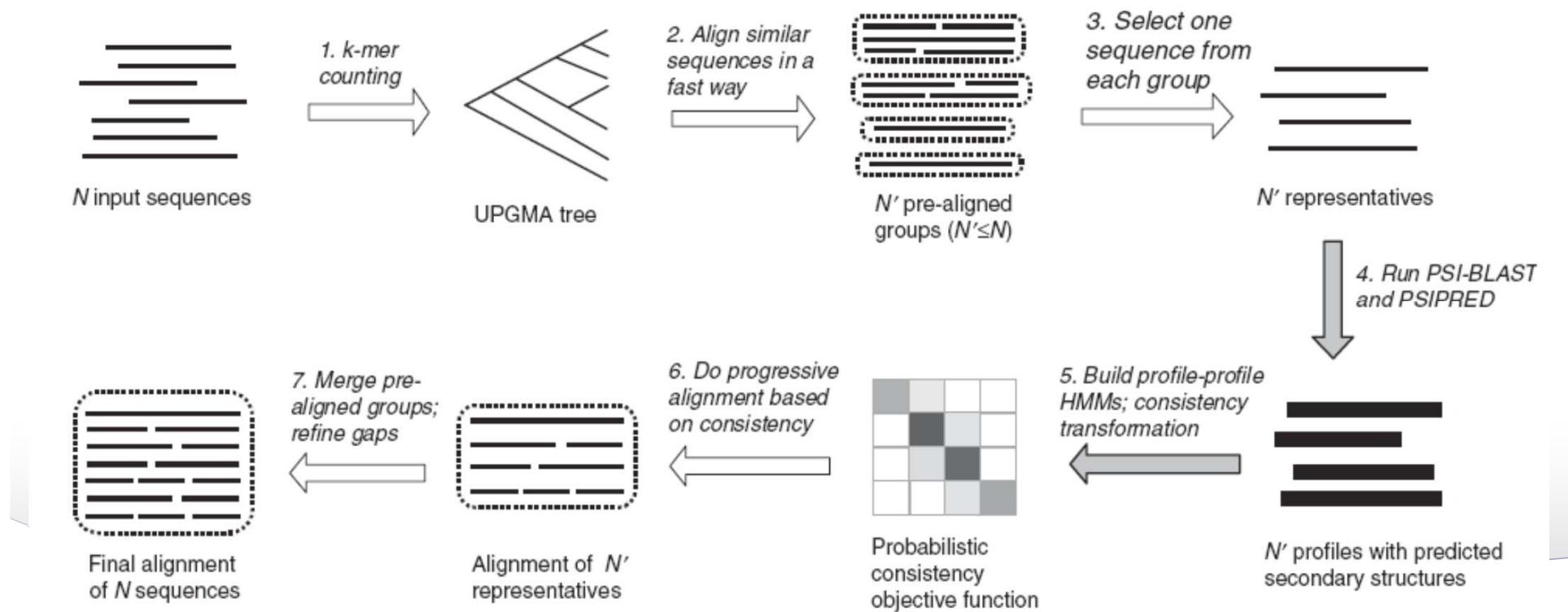


PROMALS



□ 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.lbgi.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

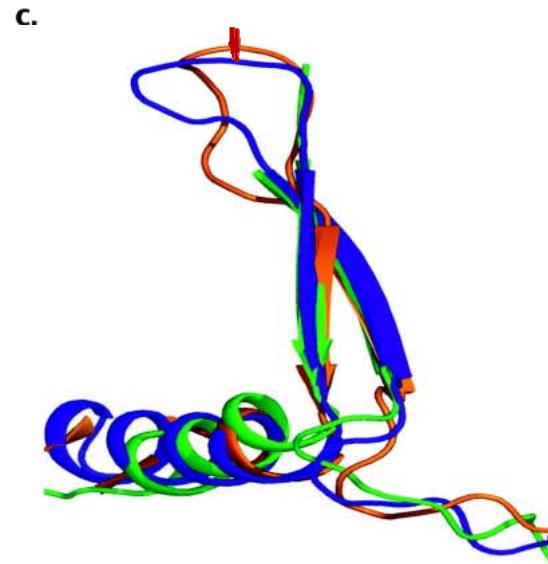
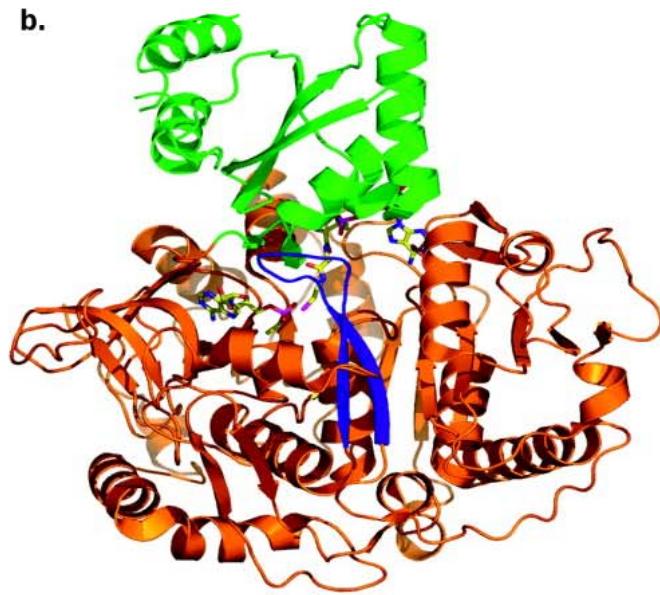
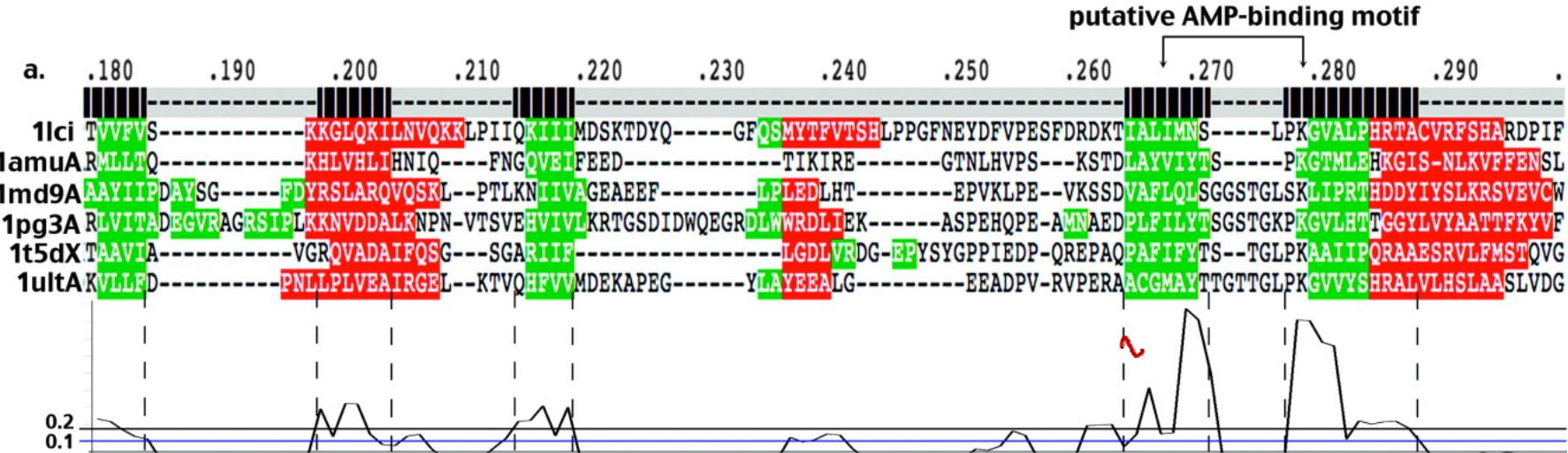
Reference 5: insertions

References 6,7,8
Repeat
Transmembrane
Circ. permutation

Reference 9: linear motifs

The screenshot shows the BALiBASE 4 homepage. At the top, it says "Welcome to BALiBASE 4" and "download the whole benchmark by html". Below this, there is a large area filled with sequence logos representing different reference motifs. At the bottom, there are ten boxes labeled Reference 1 through Reference 10, each with a small diagram and a brief description. The descriptions are: 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.

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

