Content

1. Bioinformatics databases
2. Sequence alignment and database searching

➡ 3. Phylogenic tree and multiple sequence alignment

4. Protein structure alignment
5. Protein secondary structure prediction
6. Protein tertiary structure prediction

Phylogenic tree and multiple sequence alignment

杨建益

Email: yangjy@nankai.edu.cn

Webpage: http://yanglab.nankai.edu.cn/

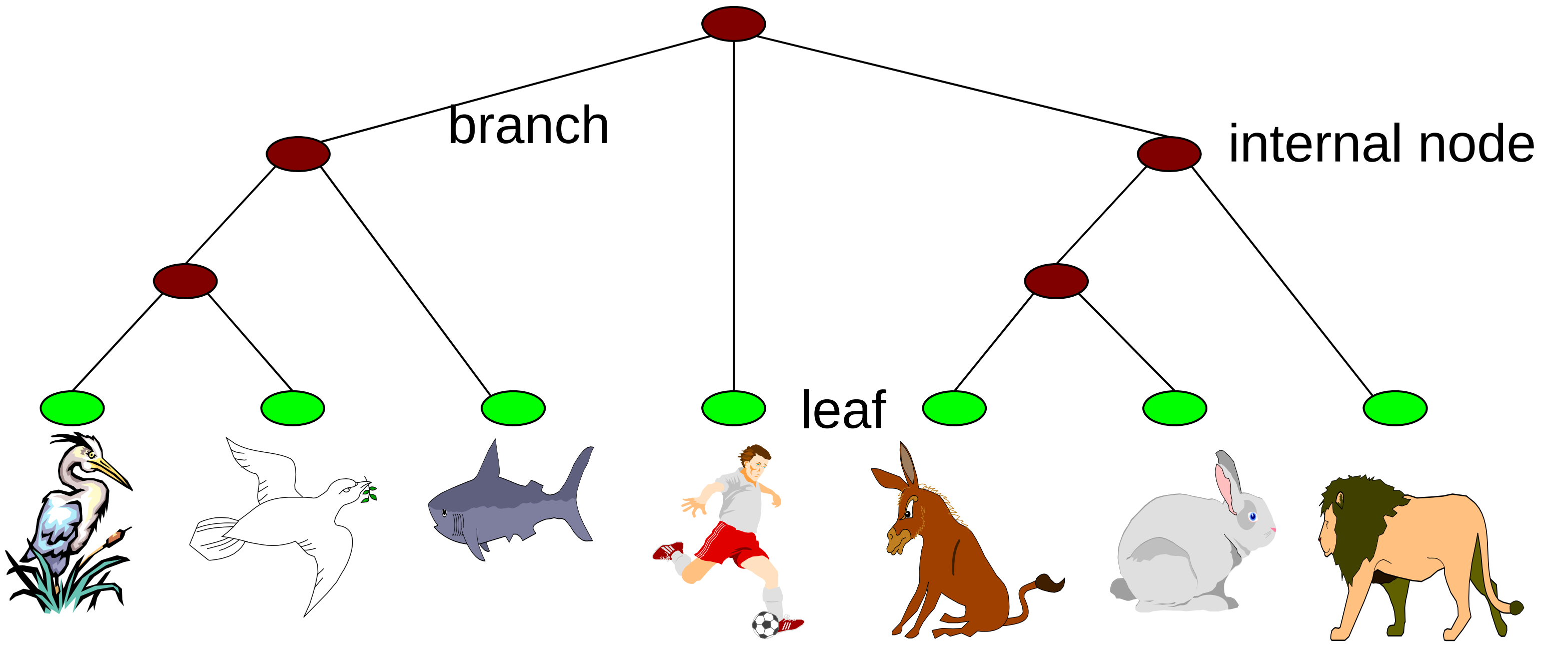
Course: http://yanglab.nankai.edu.cn/teaching/bioinformatics/

Office: 数学科学学院，419室

Content

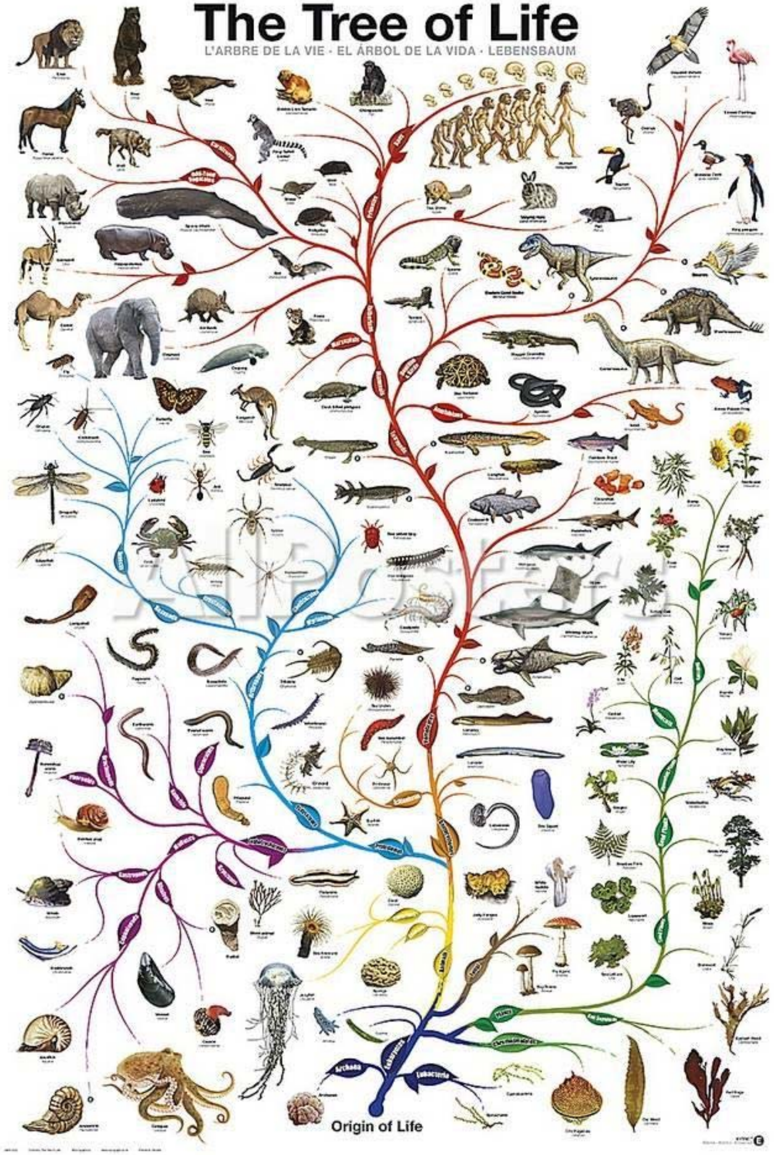
1. Phylogenic tree
2. UPGMA & neighbor-joining methods
3. How to construct a MSA?
   a. ClusterW
   b. PSI-BLAST
4. Sequence profile & profile alignments
   a. What is a sequence profile?
   b. Profile-sequence alignment
   c. Profile-profile alignment

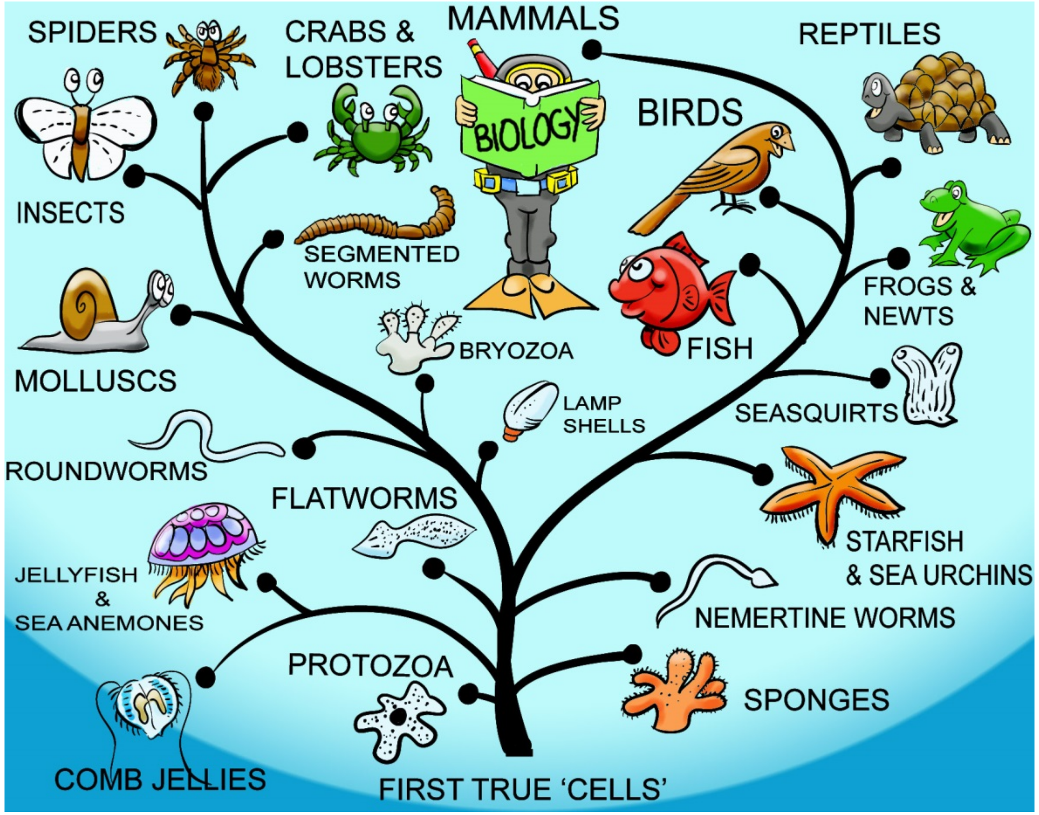
Phylogenetic tree



A phylogenetic tree is a tree showing the evolutionary relationships among various biological species or other entities that are believed to have a common ancestor.

Phylogenetic tree of life





Rooted / Unrooted Tree

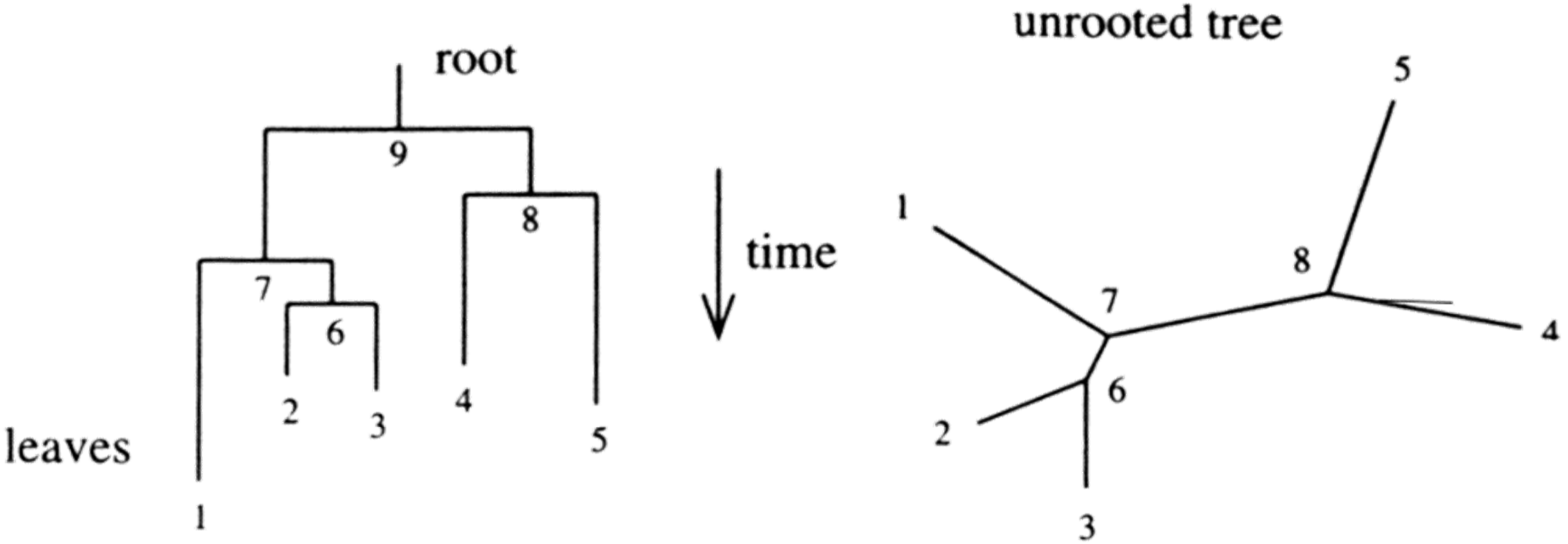
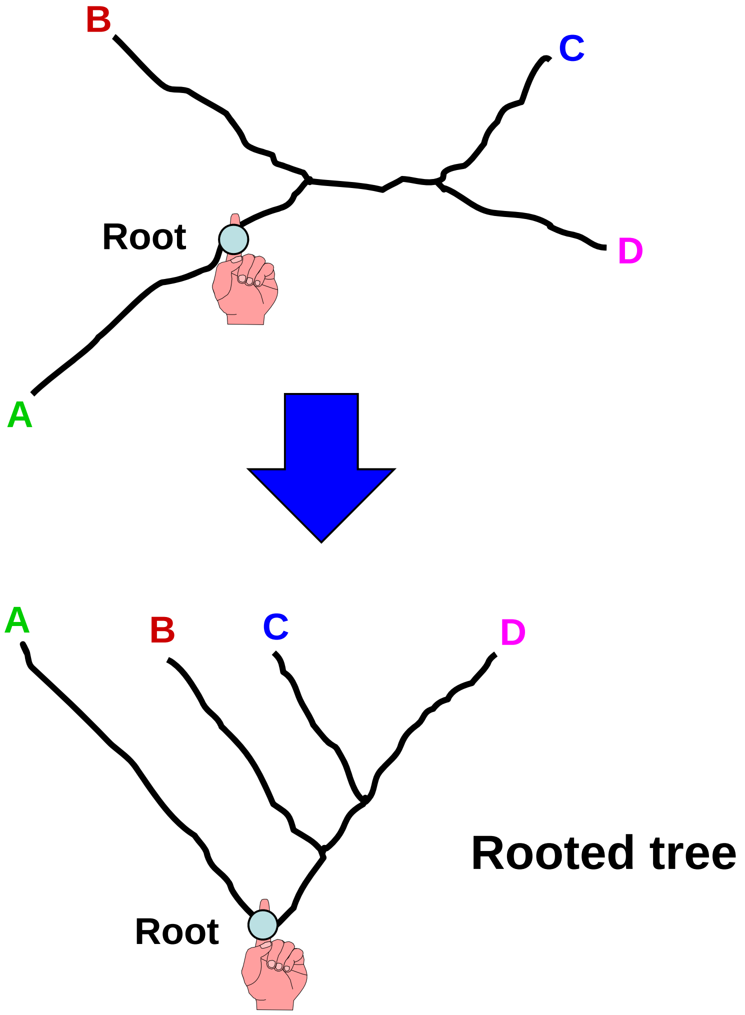


Figure 7.2 *An example of a binary tree, showing the root and leaves, and the direction of evolutionary time (the most recent time being at the bottom of the figure). The corresponding unrooted tree is also shown; the direction of time here is undetermined.*

Rooting the tree

To root a tree mentally, imagine that the tree is made of string. Grab the string at the root and tug on it until the ends of the string (the taxa) fall opposite the root



Note that in this rooted tree, taxon A is no more closely related to taxon B than it is to C or D.

Phylogenetic tree

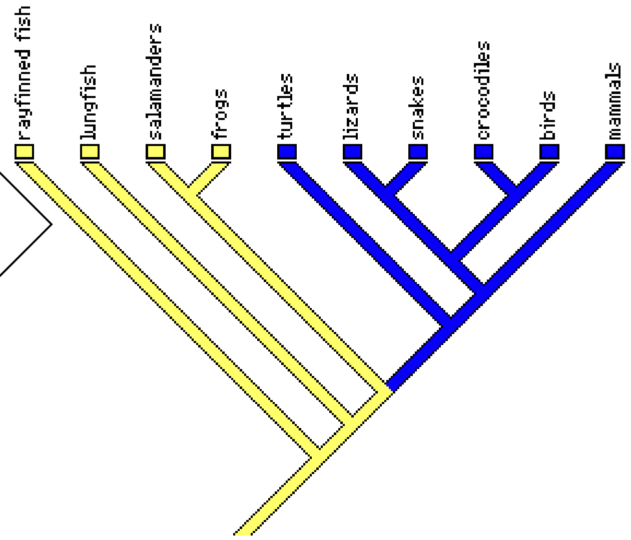
Problem to solve:

Input: pair-wise distances

Output: phylogenetic tree

	RF	LF	SA	FR	TU	LI	SN	CR	BI	MA
RF	0	0.2	0.3	0.1	0.8	0.9	0.3	0.9	0.3	0.2
LF		0	0.6	0.5	0.6	0.6	0.7	0.1	0.6	0.4
SA			0	0.1	0.4	0.9	0.3	0.2	0.4	0.5
FR				0	0.3	0.1	0.2	0.4	0.3	0.2
TU					0	0.2	0.5	0.9	0.5	0.1
LI						0	0.3	0.6	0.3	0.4
SN							0	0.2	0.6	0.8
CR								0	0.1	0.9
BI									0	0.1
MA										0

How?



1. Topology of the phylogenetic tree
2. Revolutionary age (length of each branch)

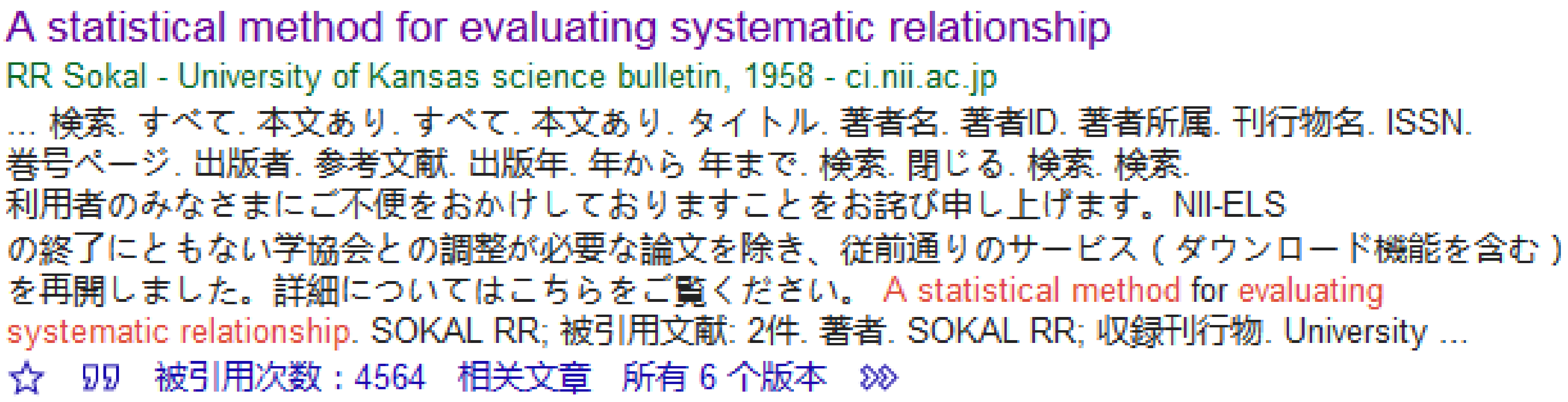
Content

1. Phylogenic tree
2. UPGMA & neighbor-joining methods
3. How to construct a MSA?
    a. ClusterW
    b. PSI-BLAST
4. Sequence profile & profile alignments
    a. What is a sequence profile?
    b. Profile-sequence alignment
    c. Profile-profile alignment

UPGMA (Unweighted Pair Group Method with Arithmetic mean)

Reference:

Sokal R and Michener C (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin. 38: 1409–1438.



A statistical method for evaluating systematic relationship

RR Sokal - University of Kansas science bulletin, 1958 - ci.nii.ac.jp

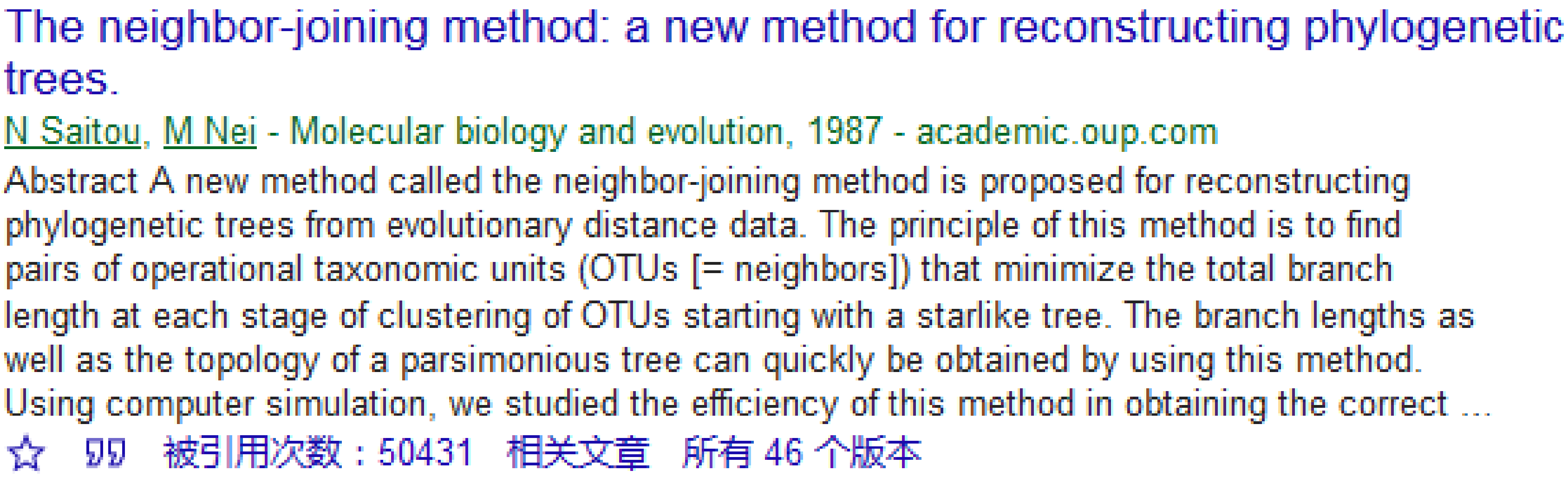
... 検索. すべて. 本文あり. すべて. 本文あり. タイトル. 著者名. 著者ID. 著者所属. 刊行物名. ISSN. 巻号ページ. 出版者. 参考文献. 出版年. 年から 年まで. 検索. 閉じる. 検索. 検索. 利用者のみなさまにご不便をおかけしておりますことをお詫び申し上げます。NII-ELS の終了にともない学協会との調整が必要な論文を除き、従前通りのサービス (ダウンロード機能を含む) を再開しました。詳細についてはこちらをご覧ください。 A statistical method for evaluating systematic relationship. SOKAL RR; 被引用文献: 2件. 著者. SOKAL RR; 収録刊行物. University ...

https://en.wikipedia.org/wiki/UPGMA

Neighbor-joining method

Reference:

N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing hylogenetic tree. Mol Biol Evol. (1987) 4: 406-425.



The neighbor-joining method: a new method for reconstructing phylogenetic trees.

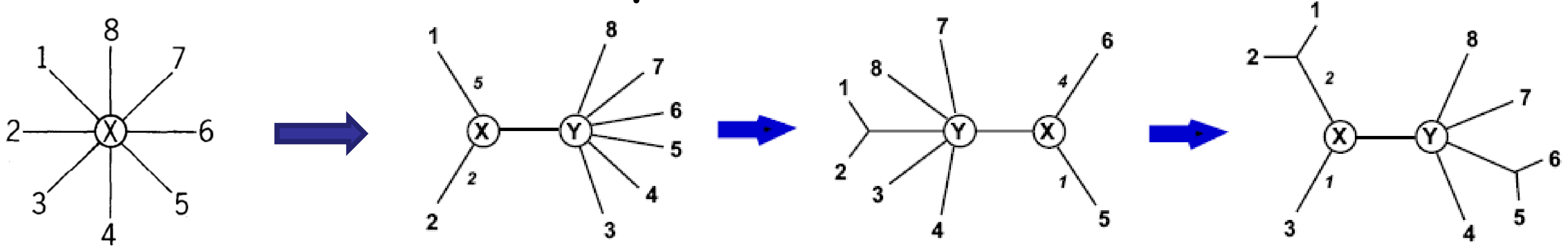
N Saitou, M Nei - Molecular biology and evolution, 1987 - academic.oup.com

Abstract A new method called the neighbor-joining method is proposed for reconstructing phylogenetic trees from evolutionary distance data. The principle of this method is to find pairs of operational taxonomic units (OTUs [= neighbors]) that minimize the total branch length at each stage of clustering of OTUs starting with a starlike tree. The branch lengths as well as the topology of a parsimonious tree can quickly be obtained by using this method. Using computer simulation, we studied the efficiency of this method in obtaining the correct ...

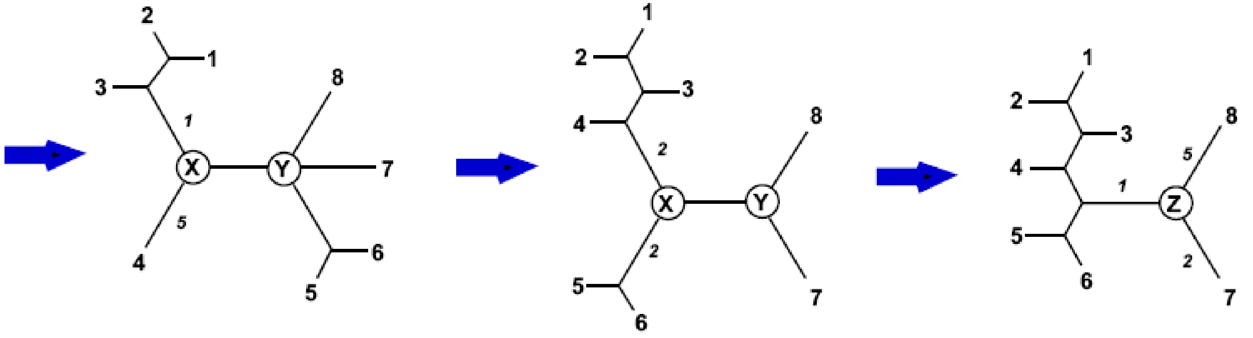
☆ ❞ 被引用次数：50431 相关文章 所有 46 个版本

Neighbor-joining method

Step 1: Join all lives to one point



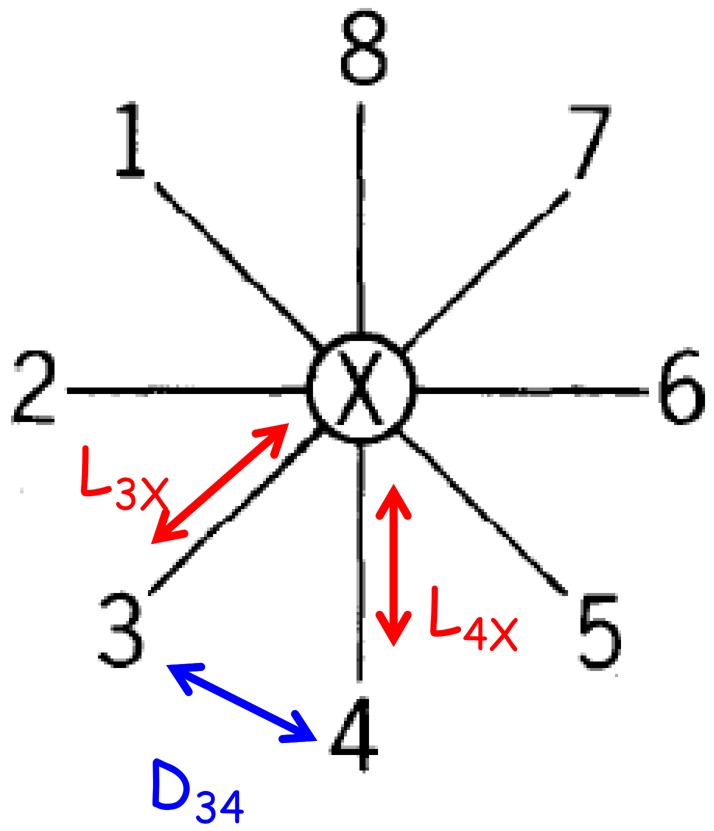
Step 2: Split the points into branches gradually



Strategy: Select two nodes to join at each step

Principle: Keep the total length of all branches minimum

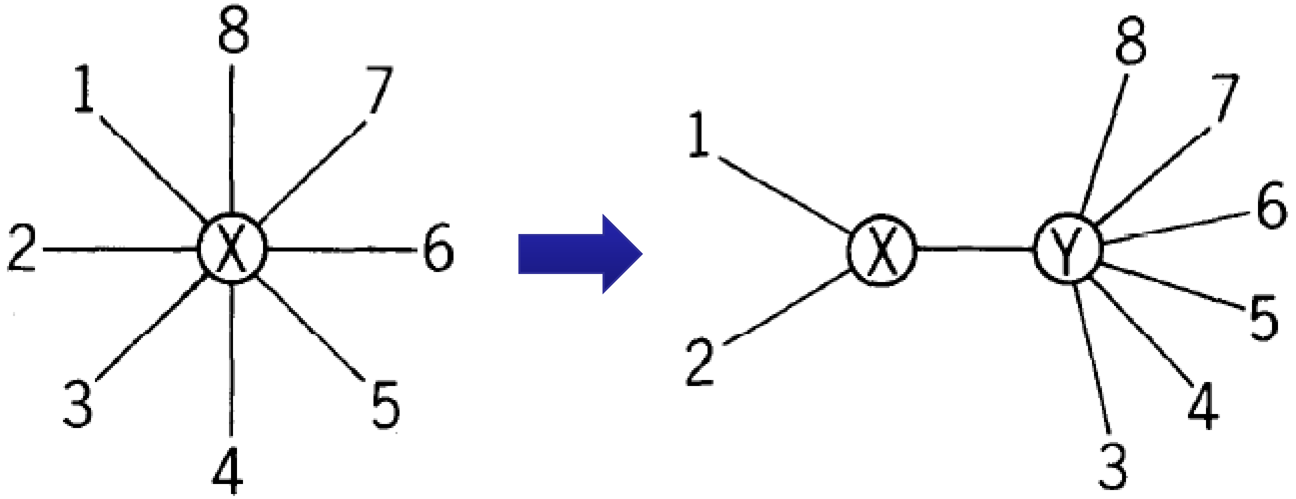
Known and unknown



D_{ij} is known

L_{ix} is unknown

Calculate the total length of all branches

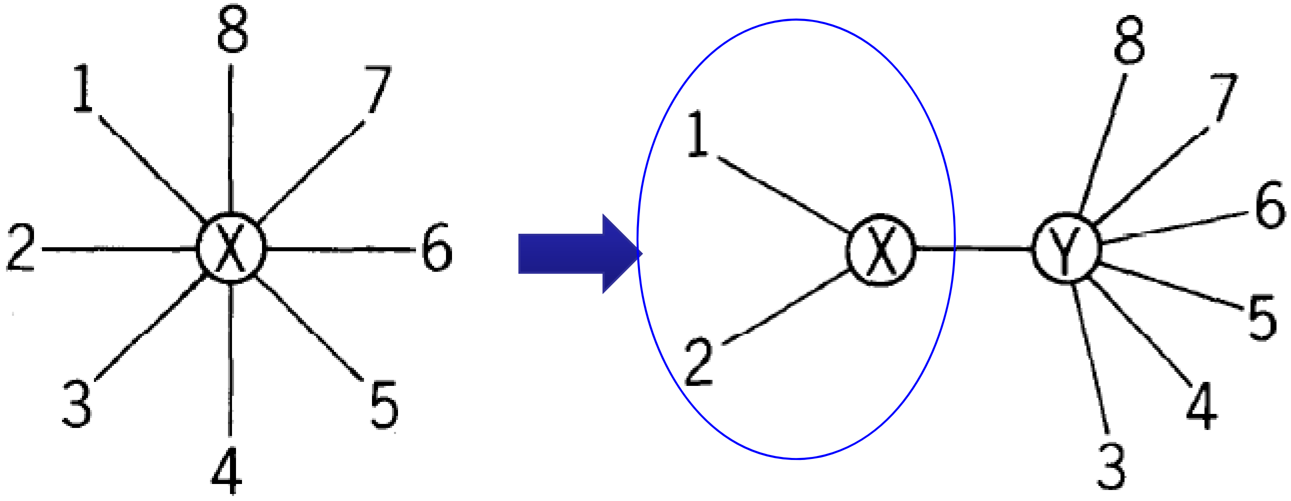


If 1,2 is joined, the total length of all branches is:

$$S_{12} = L_{1X} + L_{2X} + L_{XY} + L_{3Y} + L_{4Y} + L_{5Y} + L_{6Y} + L_{7Y} + L_{8Y}$$

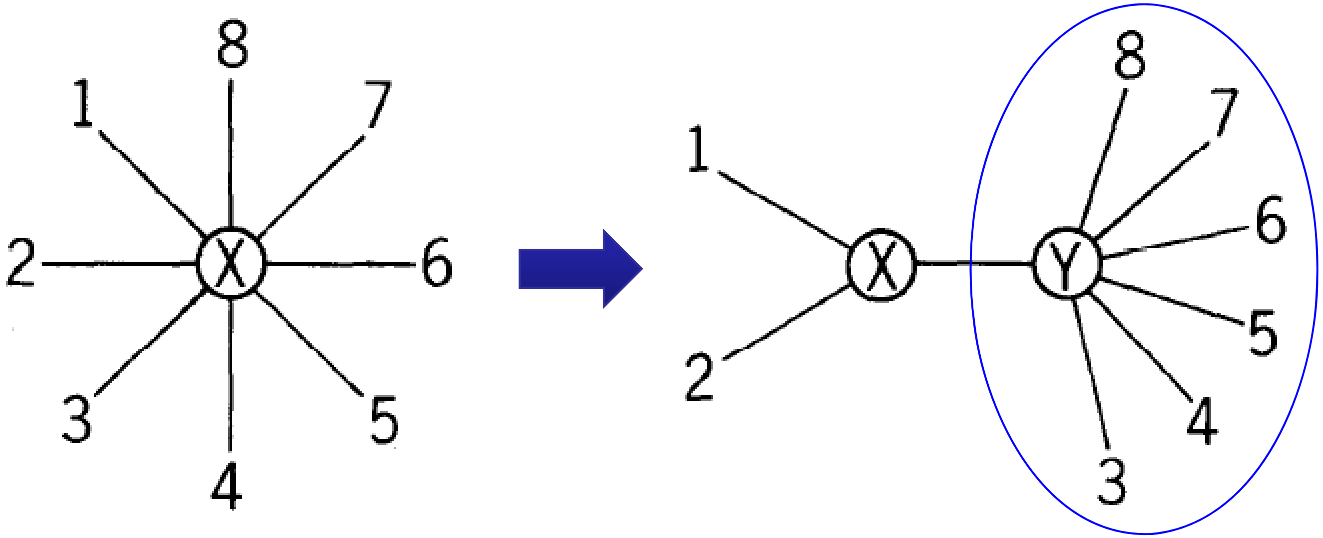
How to calculate S_{12} based on D_{ij}?

Calculate the total length of all branches



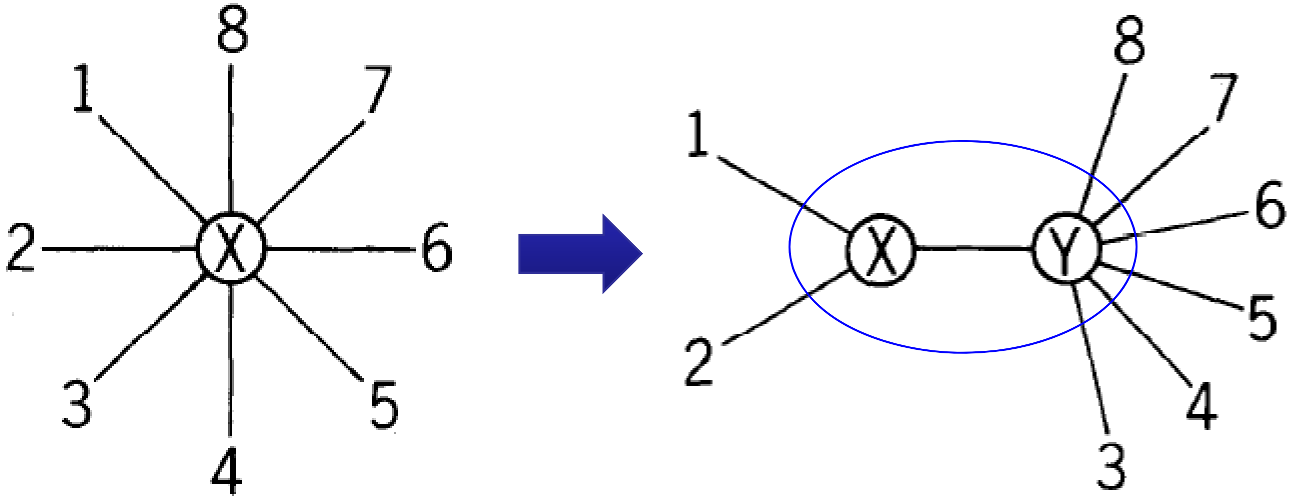
$$L_{1X} + L_{2X} = D_{12}$$

Calculate the total length of all branches



$$\sum_{i=3}^{N} L_{iY} = \frac{1}{N-3} \sum_{2<i<j} D_{ij}$$

Calculate the total length of all branches



$$L_{XY} = \frac{1}{2(N-2)}\left[\sum_{k-3}^{N}(D_{1k}+D_{2k}) - (N-2)(L_{1X}+L_{2X}) - 2\sum_{i=3}^{N} L_{iY}\right]$$

$$= \frac{1}{2(N-2)}\left[\sum_{k-3}^{N}(D_{1k}+D_{2k}) - (N-2)D_{12} - \frac{2}{N-3}\sum_{2<i<j} D_{ij}\right]$$

Calculate the total length of all branches

$$L_{1X} + L_{2X} = D_{12} \qquad \sum_{i=3}^{N} L_{iY} = \frac{1}{N-3} \sum_{2<i<j} D_{ij}$$

$$L_{XY} = \frac{1}{2(N-2)} \left[\sum_{k-3}^{N} (D_{1k} + D_{2k}) - (N-2) D_{12} - \frac{2}{N-3} \sum_{2<i<j} D_{ij} \right]$$

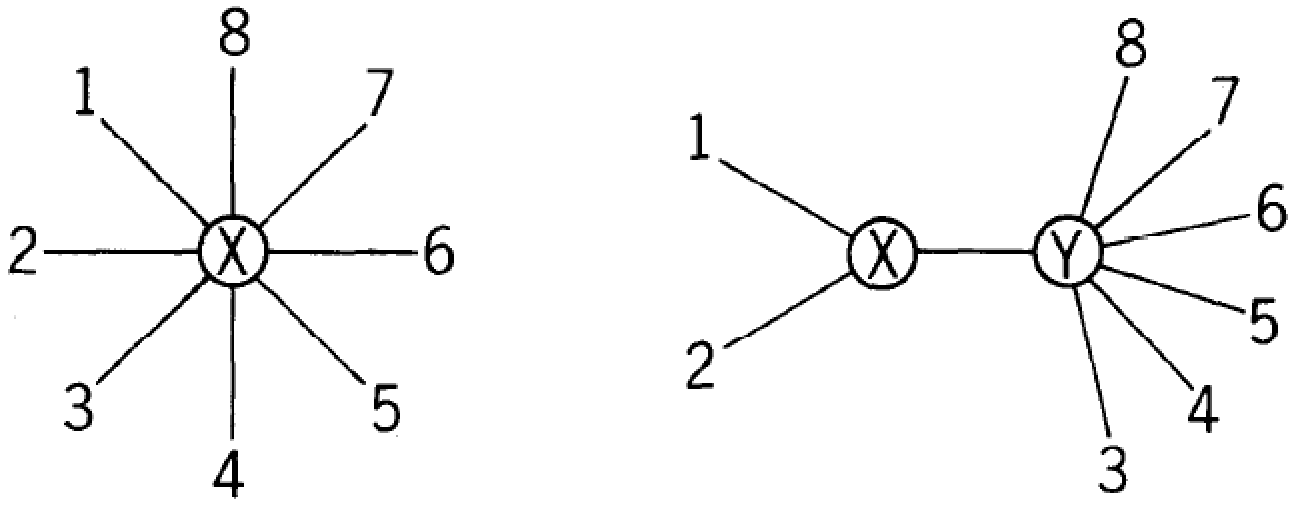
$$S_{12} = L_{1X} + L_{2X} + L_{XY} + L_{3Y} + L_{4Y} + L_{5Y} + L_{6Y} + L_{7Y} + L_{8Y}$$

$$= D_{12} + \frac{1}{2(N-2)} \left[\sum_{k=3}^{N} (D_{1k} + D_{2k}) - (N-2) D_{12} - \frac{2}{N-3} \sum_{2<i<j} D_{ij} \right] + \frac{1}{N-3} \sum_{2<i<j} D_{ij}$$

$$= \frac{1}{2(N-2)} \sum_{k=3}^{N} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{(N-2)} \sum_{2<i<j} D_{ij}$$

$$S_{mn} = \frac{1}{2(N-2)} \sum_{k \neq m,n}^{N} (D_{mk} + D_{nk}) + \frac{1}{2} D_{mn} + \frac{1}{(N-2)} \sum_{i<j, \neq m,n} D_{ij}$$

Calculate the new length

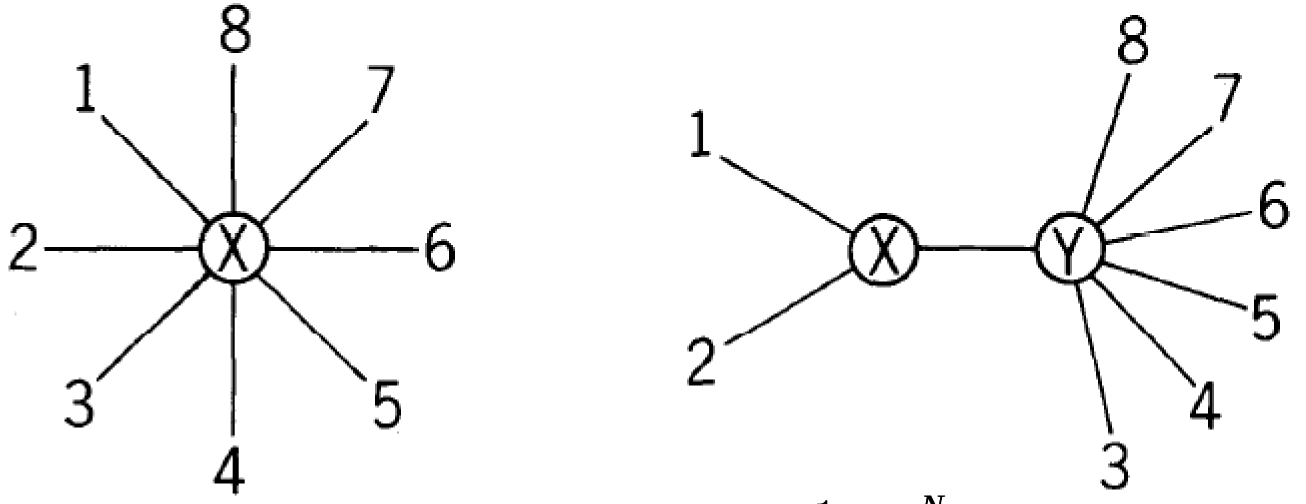


$$\boldsymbol{D_{Xj}} = \frac{\boldsymbol{D_{1j}} + \boldsymbol{D_{2j}} - \boldsymbol{D_{12}}}{2}, \boldsymbol{j} = 3, 4, \cdots, \boldsymbol{N}$$

$$\boldsymbol{L_{1X}} + \boldsymbol{L_{XY}} + \boldsymbol{L_{Yj}} = \boldsymbol{D_{1j}}, \boldsymbol{j} = 3, \cdots, \boldsymbol{N}$$

$$\boldsymbol{L_{2X}} + \boldsymbol{L_{XY}} + \boldsymbol{L_{Yj}} = \boldsymbol{D_{2j}}, \boldsymbol{j} = 3, \cdots, \boldsymbol{N}$$

Calculate the new length



$$L_{1X}=\frac{(L_{1X}+L_{2X})+(L_{1X}-L_{2X})}{2}=\frac{D_{12}+\frac{1}{N-2}\sum_{i=3}^{N}(D_{1i}-D_{2i})}{2}$$

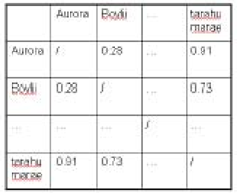
$$L_{1X}+L_{XY}+L_{Yj}=D_{1j},\ j=3,\cdots,N$$
$$L_{2X}+L_{XY}+L_{Yj}=D_{2j},\ j=3,\cdots,N$$

⇨

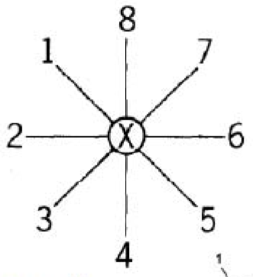
$$(N-2)(L_{1X}-L_{2X})=\sum_{i=3}^{N}(D_{1i}-D_{2i})$$

$$L_{2X}=\frac{(L_{1X}+L_{2X})+(L_{2X}-L_{1X})}{2}=\frac{D_{12}+\frac{1}{N-2}\sum_{i=3}^{N}(D_{2i}-D_{1i})}{2}$$

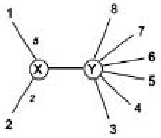
Flowchart for constructing phylogenetic tree



	Aurora	Boyli	...	tarahu marae
Aurora	/	0.28	...	0.91
Boyli	0.28	/	...	0.73
...	...	...	/	...
tarahu marae	0.91	0.73	...	/



Enumerate all possible splits



$$S_{mn} = \frac{1}{2(N-2)} \sum_{K \neq m,n}^{N} (D_{mk} + D_{nk}) + \frac{D_{mn}}{2} + \frac{1}{N-2} \sum_{i<j,\neq m,n} D_{ij}$$

Join two nodes of minimum S_{mn}

$$D_{Xj} = \frac{D_{1j} + D_{2j} - D_{12}}{2}, \quad L_{1X} = \frac{D_{12} + \frac{1}{N-2}\sum_{i=3}^{N}(D_{1i} - D_{2i})}{2}, \quad L_{2X} = \frac{D_{12} + \frac{1}{N-2}\sum_{i=3}^{N}(D_{2i} - D_{1i})}{2}$$

Remaining N>2?

Yes

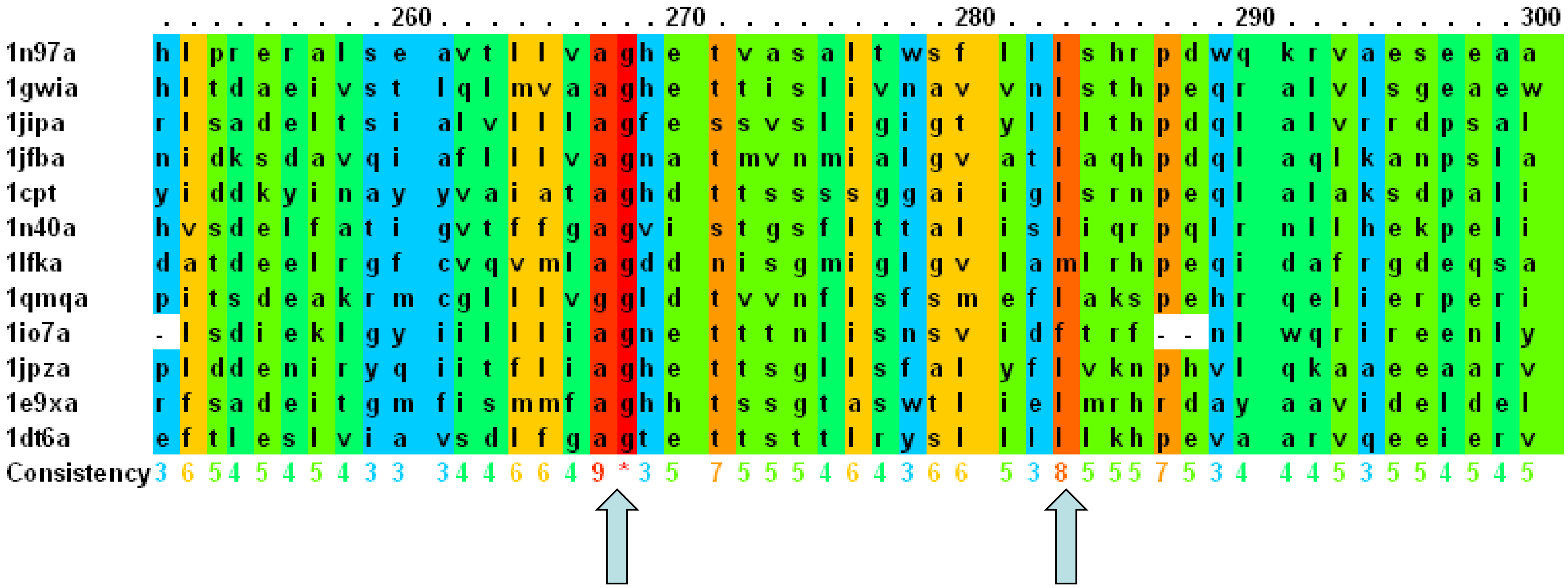
No

Phylogenetic tree with branch lengths

Content

1. Phylogenic tree
2. UPGMA & neighbor-joining methods
3. How to construct a MSA?
   a. ClusterW
   b. PSI-BLAST
4. Sequence profile & profile alignments
   a. What is a sequence profile?
   b. Profile-sequence alignment
   c. Profile-profile alignment

Multiple sequence alignment



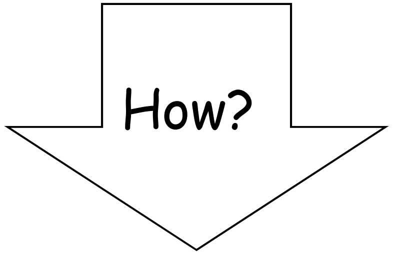
```
             . . . . . . . . . 260 . . . . . . . . . 270 . . . . . . . . . 280 . . . . . . . . . 290 . . . . . . . . . 300
1n97a        h l p r e r a l s e a v t l l v a g h e t v a s a l t w s f l l l s h r p d w q k r v a e s e e a a
1gwia        h l t d a e i v s t l q l m v a a g h e t t i s l i v n a v v n l s t h p e q r a l v l s g e a e w
1jipa        r l s a d e l t s i a l v l l l a g f e s s v s l i g i g t y l l l t h p d q l a l v r r d p s a l
1jfba        n i d k s d a v q i a f l l l v a g n a t m v n m i a l g v a t l a q h p d q l a q l k a n p s l a
1cpt         y i d d k y i n a y y v a i a t a g h d t t s s s s g g a i i g l s r n p e q l a l a k s d p a l i
1n40a        h v s d e l f a t i g v t f f g a g v i s t g s f l t t a l i s l i q r p q l r n l l h e k p e l i
1lfka        d a t d e e l r g f c v q v m l a g d d n i s g m i g l g v l a m l r h p e q i d a f r g d e q s a
1qmqa        p i t s d e a k r m c g l l l v g g l d t v v n f l s f s m e f l a k s p e h r q e l i e r p e r i
1io7a        - l s d i e k l g y i i l l l i a g n e t t t n l i s n s v i d f t r f - - n l w q r i r e e n l y
1jpza        p l d d e n i r y q i i t f l i a g h e t t s g l l s f a l y f l v k n p h v l q k a a e e a a r v
1e9xa        r f s a d e i t g m f i s m m f a g h h t s s g t a s w t l i e l m r h r d a y a a v i d e l d e l
1dt6a        e f t l e s l v i a v s d l f g a g t e t t s t t l r y s l l l l l k h p e v a a r v q e e i e r v
Consistency  3 6 5 4 5 4 5 4 3 3 3 4 4 6 6 4 9 * 3 5 7 5 5 5 4 6 4 3 6 6 5 3 8 5 5 5 7 5 3 4 4 4 5 3 5 5 4 5 4 5
```

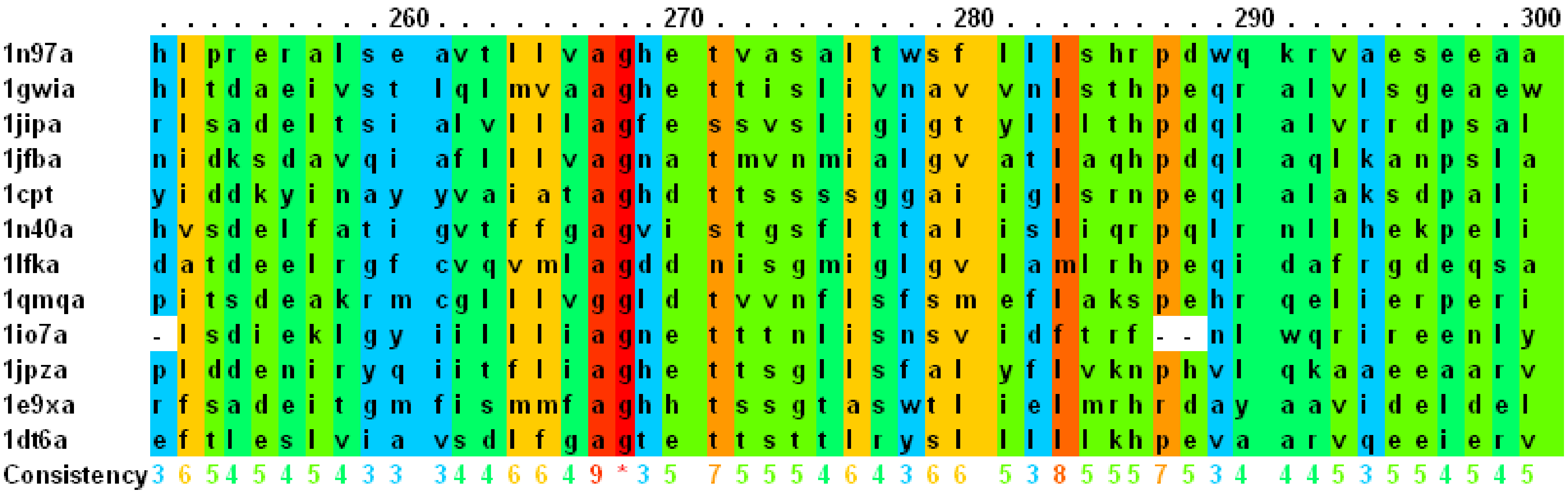
Conserved positions

<u>Multiple sequence alignment</u>: Take three or more sequences and align them so that the greatest number of similar characters are aligned in the same column of the alignment.

Multiple sequence alignment

```
Protein 1 MSAADLLRLVGPRWVRPRRLGRIPDQPIVHAVRETAPGMLADQLSDHLATIVPHAELHVGDAARGTERERSVQVRTLLDTAVL
Protein 2 GLREHDSWPRIGRLQFPRYALTSWLLKQNLRPAELNHAPHSNIRDLLHDFLNSRRRPGRGK
Protein 3 QNAREAAAWTSMTEQLPWYLFLLSLVAFPFYYALWVRRGKVPRWFLRQQYLAPR
Protein 4 ESADFPSFVRRLITTPSERESAEQVRRLLVHAFLSDLSDSHSRRLWRWRWVPKDCYPVLLLKDLRPGTIGETLVRLVNNVRNETGARDPLLVVATGEQPLEDGE
Protein 5 TPRAPVTLEQWERDLQAARRKRSPTAWYVPLRIADEPADALDYDRFGALGRAHLPLKRSKLVRRTPLLLVLLLLVGSTAGYAGYLRTHCGQWWPYQNSDIGEVDGECIGV
Protein 6 SDTTSTSRFFSAHDARMVAAQEKIAEQNEEAERRWEDQPNLPHPTVVYFSTFPSSDDDPPTLAGIADELDGVAVMQRESLGRNVLMKVVLAN
Protein 7 GGLRMKHGPRVAADVAELVGRDDSVVAVAGLGGSWQATVDTIEALEAEGVPMVGTTISADLLSESSPLFYQVAPSNAWEA
Protein 8 KVVANYIAAGPVDPRTGAPRRPDNVLIYSNPRDLYSHDLAQLTAGELRARGIEPMPDSDRIPCGKQNLVFFAGR
Protein 9 ANDLATFLTKMPPECGKPENYPQLLAGDDTSKLVLDDAMDDHEGVVLDHVSFTGRSAWDPQSQQGTPLRGRGLLARDALEVIALAVQ
...
```

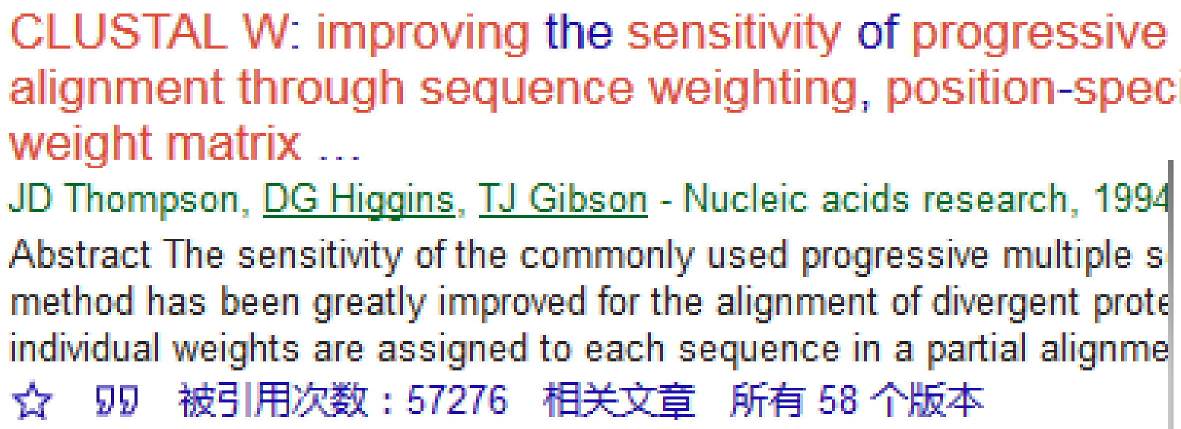




```
            . . . . . . . . . 260 . . . . . . . . . 270 . . . . . . . . . 280 . . . . . . . . . 290 . . . . . . . . . 300
1n97a       hlpreralse avtllvaghe tvasaltwsf lllshrpdwq krvaeseeaa
1gwia       hltdaeivst lqlmvaaghe ttisliynav vnlsthpeqr alvlsgeaew
1jipa       rlsadeltsi alvlllagfe ssvsligigt yllllthpdql alvrrdpsal
1jfba       nidksdavqi afllllvagna tmvnmialgv atlaqhpdql aqlkanpsla
1cpt        yiddkyinay yvaiataghd ttssssggai iglsrnpeql alaksdpali
1n40a       hvsdelfati gvtffgagvi stgsflttal islaqrpqlr nllhekpeli
1lfka       datdeelrgf cvqvmlagdd nisgmiglgv lamlrhpeqi dafrgdeqsa
1qmqa       pitsdeakrm cglllvggld tvvnflsfsm eflakspehr qelierperi
1io7a       -lsdieklgy iillliagne tttnlisnsv idftrf--nl wqrireenly
1jpza       plddeniryq iitfliaghe ttsgllsfal yflvknphvl qkaaeeaarv
1e9xa       rfsadeitgm fismmfaghh tssgtaswtl ielmrhrday aavideldel
1dt6a       eftleslvia vsdlfgagte ttsttlrysl llllkhpeva arvqeeierv
Consistency 3654545433 3446649*35 7555464366 5385557534 4453554545
```

Progressive method: ClustalW

Reference:
Thompson et al. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acid Res. (1994) 22, 4673-4680

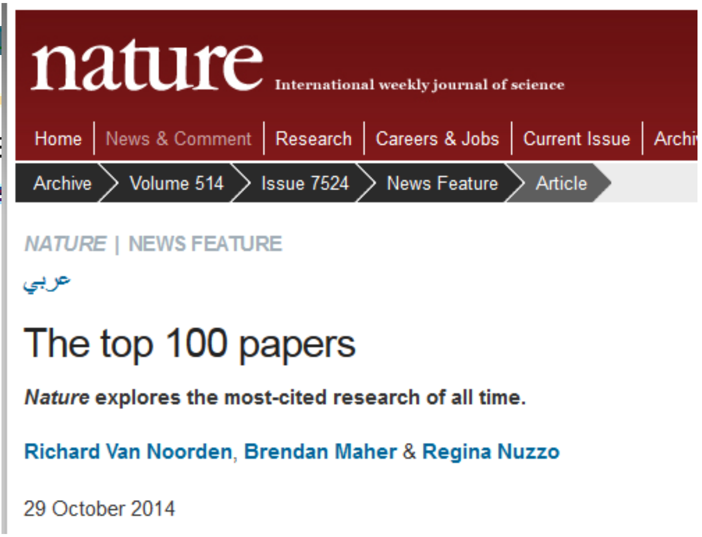


CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix ...

JD Thompson, DG Higgins, TJ Gibson - Nucleic acids research, 1994

Abstract The sensitivity of the commonly used progressive multiple s
method has been greatly improved for the alignment of divergent prote
individual weights are assigned to each sequence in a partial alignme

☆ 被引用次数：57276 相关文章 所有 58 个版本



nature
International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archi
Archive > Volume 514 > Issue 7524 > News Feature > Article

NATURE | NEWS FEATURE

عربي

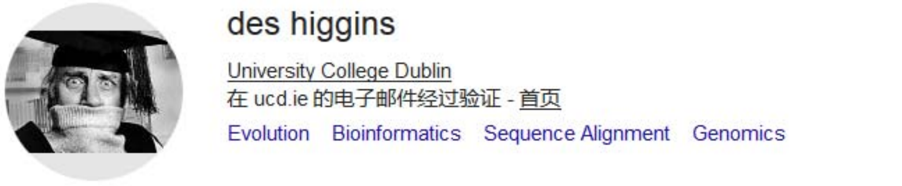
The top 100 papers

Nature **explores the most-cited research of all time.**

Richard Van Noorden, Brendan Maher & Regina Nuzzo

29 October 2014

Des Higgins

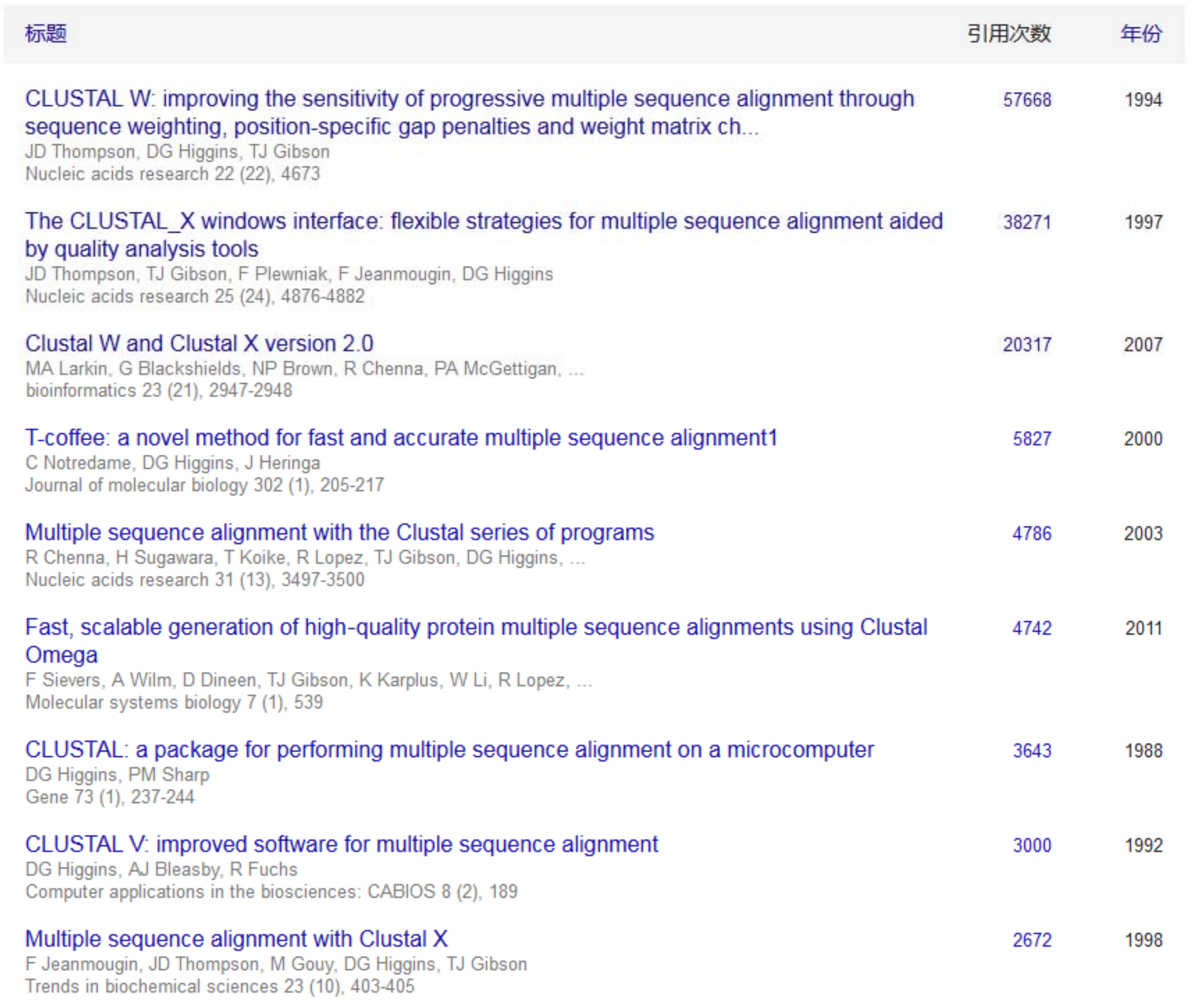


des higgins

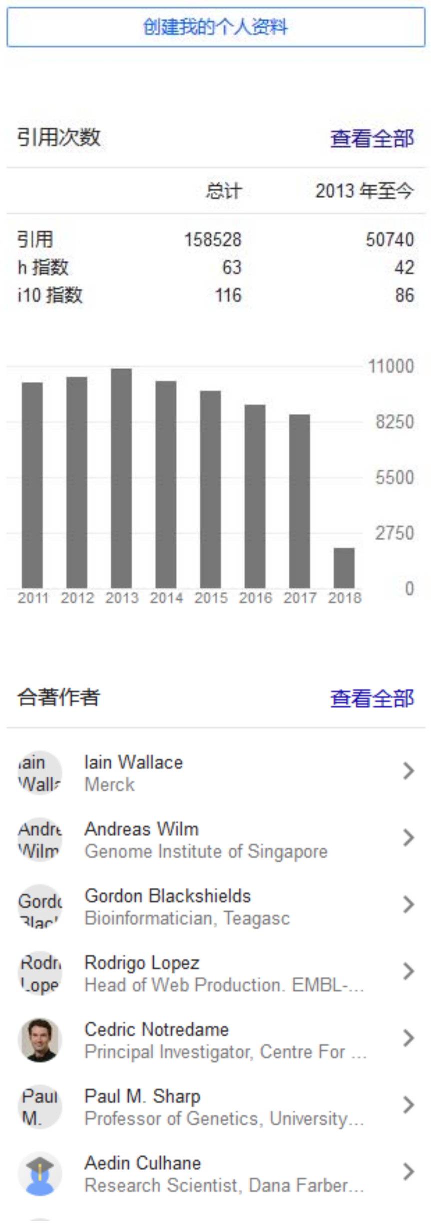
University College Dublin

在 ucd.ie 的电子邮件经过验证 - 首页

Evolution Bioinformatics Sequence Alignment Genomics



标题	引用次数	年份
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix ch... JD Thompson, DG Higgins, TJ Gibson Nucleic acids research 22 (22), 4673	57668	1994
The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools JD Thompson, TJ Gibson, F Plewniak, F Jeanmougin, DG Higgins Nucleic acids research 25 (24), 4876-4882	38271	1997
Clustal W and Clustal X version 2.0 MA Larkin, G Blackshields, NP Brown, R Chenna, PA McGettigan, ... bioinformatics 23 (21), 2947-2948	20317	2007
T-coffee: a novel method for fast and accurate multiple sequence alignment1 C Notredame, DG Higgins, J Heringa Journal of molecular biology 302 (1), 205-217	5827	2000
Multiple sequence alignment with the Clustal series of programs R Chenna, H Sugawara, T Koike, R Lopez, TJ Gibson, DG Higgins, ... Nucleic acids research 31 (13), 3497-3500	4786	2003
Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega F Sievers, A Wilm, D Dineen, TJ Gibson, K Karplus, W Li, R Lopez, ... Molecular systems biology 7 (1), 539	4742	2011
CLUSTAL: a package for performing multiple sequence alignment on a microcomputer DG Higgins, PM Sharp Gene 73 (1), 237-244	3643	1988
CLUSTAL V: improved software for multiple sequence alignment DG Higgins, AJ Bleasby, R Fuchs Computer applications in the biosciences: CABIOS 8 (2), 189	3000	1992
Multiple sequence alignment with Clustal X F Jeanmougin, JD Thompson, M Gouy, DG Higgins, TJ Gibson Trends in biochemical sciences 23 (10), 403-405	2672	1998



引用次数 查看全部

	总计	2013 年至今
引用	158528	50740
h 指数	63	42
i10 指数	116	86

合著作者 查看全部

- Iain Wallace — Merck
- Andreas Wilm — Genome Institute of Singapore
- Gordon Blackshields — Bioinformatician, Teagasc
- Rodrigo Lopez — Head of Web Production. EMBL-...
- Cedric Notredame — Principal Investigator, Centre For ...
- Paul M. Sharp — Professor of Genetics, University...
- Aedin Culhane — Research Scientist, Dana Farber...

Progressive method: ClustalW

Progressive algorithm:

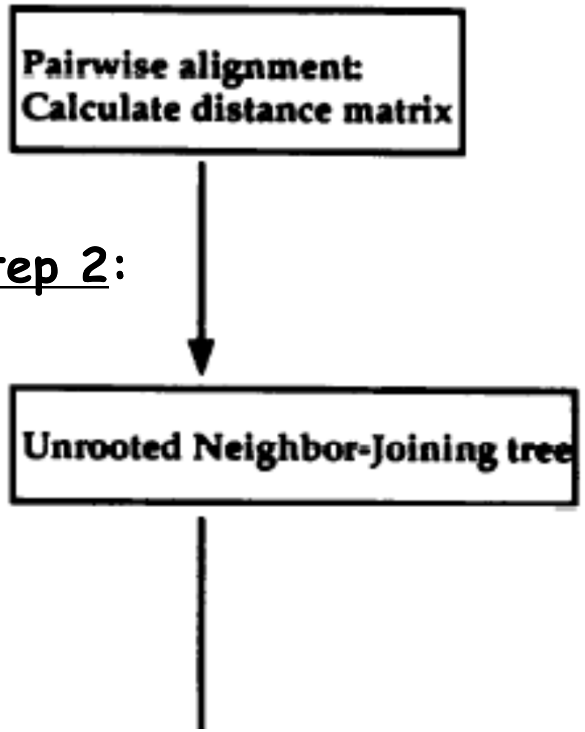
Step 1. All pairs of sequences are aligned separately and a pair-wise distance matrix is obtained

Step 2. Construct a guide tree from the distance matrix

Step 3 .Starting from the closely related sequences, other sequences are progressively aligned by dynamic programming

Progressive method: ClustalW

Step 1:

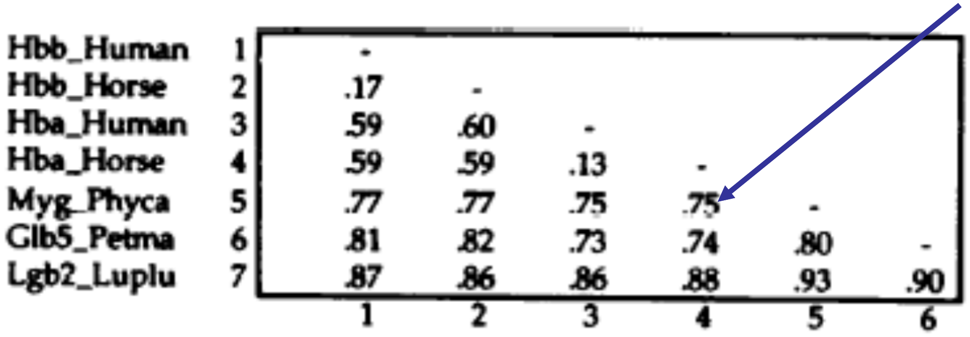


Pairwise alignment:
Calculate distance matrix

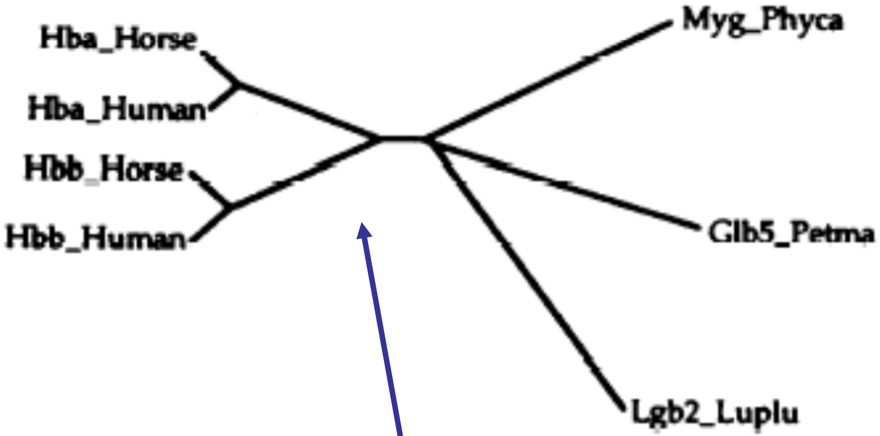
Step 2:

Unrooted Neighbor-Joining tree

1-SID

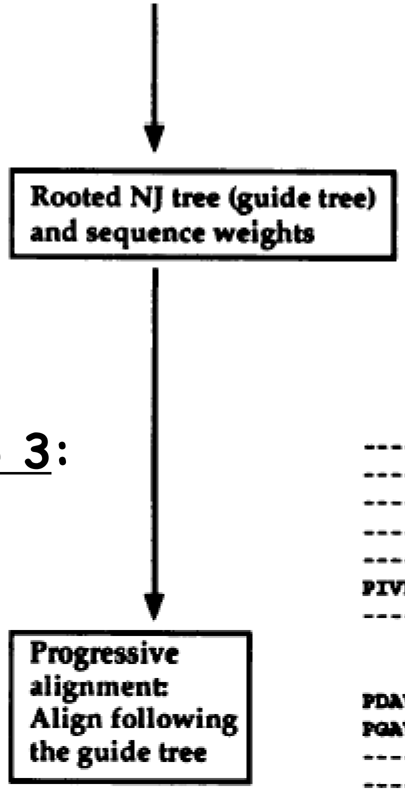


		1	2	3	4	5	6
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90



Neighbor-joining method: Saitou & Nei, Mol Biol Evol (1987) 196, 199-216

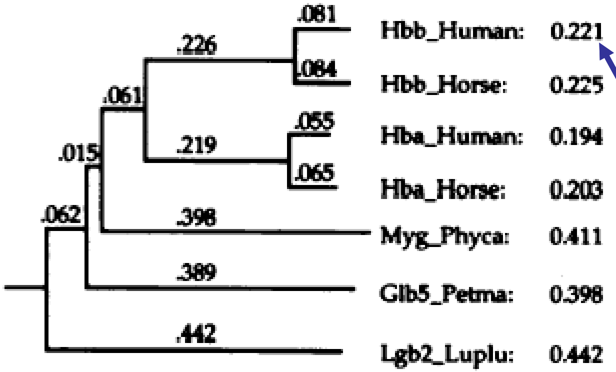
Progressive method: ClustalW



Rooted NJ tree (guide tree) and sequence weights

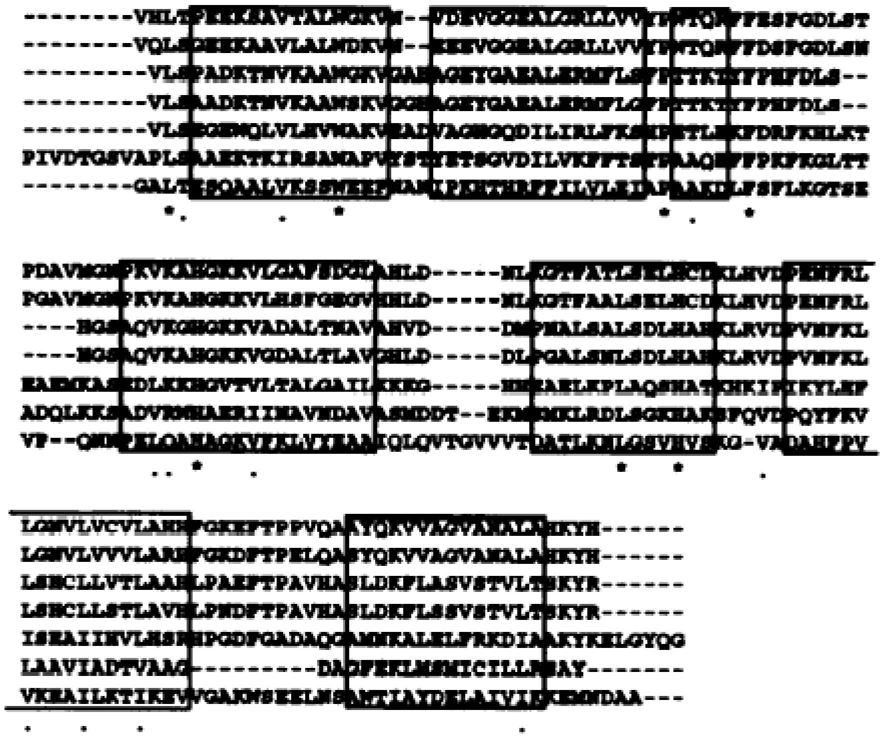
Step 3:

Progressive alignment: Align following the guide tree



Weight= Weight of branch + fraction of parent branch

0. 221=0. 081+0. 226/2 +0. 061/4+0. 015/5+0. 062/6

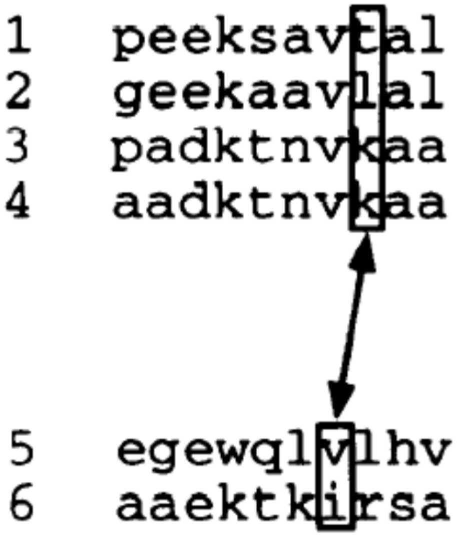


```
--------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
--------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS--
---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS--
---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
--------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
          *.      .   *                      * .   *

PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGTFATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSFGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
EAEMKASEDLKKHGVTVLTALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
         .. *   .                         *   *     .

LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
.   .   .                          .
```

Dynamic programming

Progressive method: ClustalW

Dynamic programming scoring function



```
1   peeksavtal
2   geekaavlal
3   padktnvkaa
4   aadktnvkaa

5   egewqlvlhv
6   aaektkirsa
```

Without sequence Weights:

$$\mathrm{Score} = M(t,v) + M(t,i) + M(l,v) + M(l,i) + M(k,v) + M(k,i) + M(k,v) + M(k,i)/8$$

With sequence Weights W_i:

$$\mathrm{Score} = M(t,v)*W_1*W_5 + M(t,i)*W_1*W_6 + M(l,v)*W_2*W_5 + M(l,i)*W_2*W_6 + M(k,v)*W_3*W_5 + M(k,i)*W_3*W_6 + M(k,v)*W_4*W_5 + M(k,i)*W_4*W_6/8$$

M(i,j): PAM or BLOSUM mutation matrix
W_n: Weight factor of n'th sequence based on guide tree. Groups of closely related sequences receive lower weights because they contain duplicated information

Content

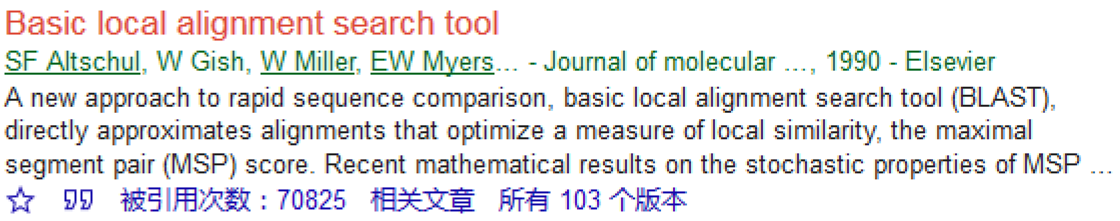
1. Phylogenic tree
2. UPGMA & neighbor-joining methods
3. How to construct a MSA?
   a. ClusterW
   ➡ b. PSI-BLAST
4. Sequence profile & profile alignments
   a. What is a sequence profile?
   b. Profile-sequence alignment
   c. Profile-profile alignment

PSI-BLAST

An iterative sequence-profile alignment algorithm

PSI-BLAST (The most often-used algorithm for sequence-profile alignment tool)

S. F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. (1997) 25, 3389-3402

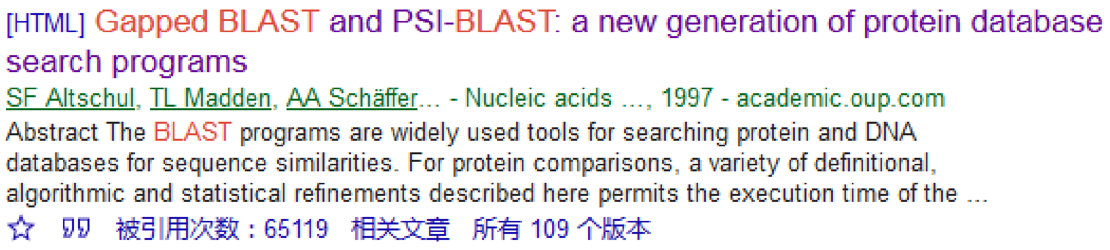


Basic local alignment search tool

SF Altschul, W Gish, W Miller, EW Myers... - Journal of molecular ..., 1990 - Elsevier

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP ...

☆ ” 被引用次数：70825 相关文章 所有 103 个版本



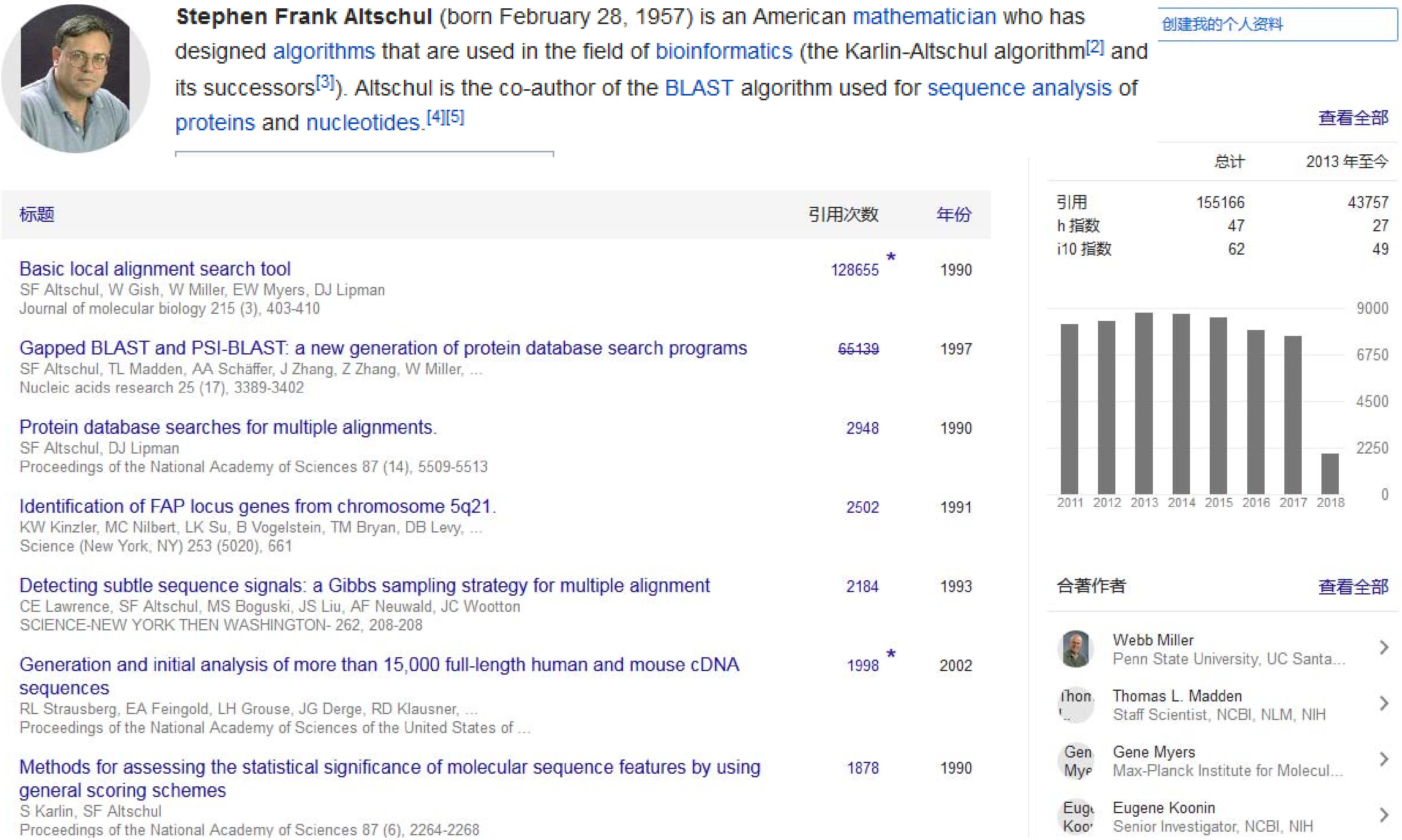
[HTML] Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

SF Altschul, TL Madden, AA Schäffer... - Nucleic acids ..., 1997 - academic.oup.com

Abstract The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarities. For protein comparisons, a variety of definitional, algorithmic and statistical refinements described here permits the execution time of the ...

☆ ” 被引用次数：65119 相关文章 所有 109 个版本

Stephen Altschul



Stephen Frank Altschul (born February 28, 1957) is an American mathematician who has designed algorithms that are used in the field of bioinformatics (the Karlin-Altschul algorithm[2] and its successors[3]). Altschul is the co-author of the BLAST algorithm used for sequence analysis of proteins and nucleotides.[4][5]

创建我的个人资料

标题	引用次数	年份
Basic local alignment search tool SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman Journal of molecular biology 215 (3), 403-410	128655 *	1990
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs SF Altschul, TL Madden, AA Schäffer, J Zhang, Z Zhang, W Miller, ... Nucleic acids research 25 (17), 3389-3402	~~65139~~	1997
Protein database searches for multiple alignments. SF Altschul, DJ Lipman Proceedings of the National Academy of Sciences 87 (14), 5509-5513	2948	1990
Identification of FAP locus genes from chromosome 5q21. KW Kinzler, MC Nilbert, LK Su, B Vogelstein, TM Bryan, DB Levy, ... Science (New York, NY) 253 (5020), 661	2502	1991
Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, JC Wootton SCIENCE-NEW YORK THEN WASHINGTON- 262, 208-208	2184	1993
Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences RL Strausberg, EA Feingold, LH Grouse, JG Derge, RD Klausner, ... Proceedings of the National Academy of Sciences of the United States of ...	1998 *	2002
Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes S Karlin, SF Altschul Proceedings of the National Academy of Sciences 87 (6), 2264-2268	1878	1990

查看全部

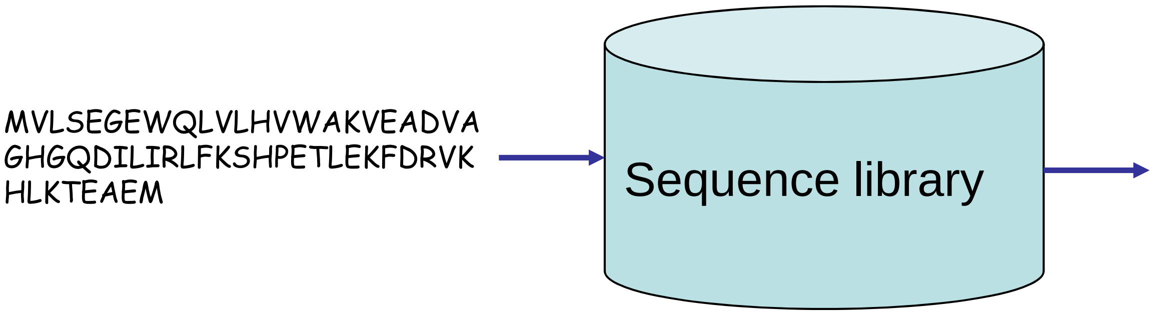
	总计	2013 年至今
引用	155166	43757
h 指数	47	27
i10 指数	62	49

合著作者 查看全部

- Webb Miller — Penn State University, UC Santa...
- Thomas L. Madden — Staff Scientist, NCBI, NLM, NIH
- Gene Myers — Max-Planck Institute for Molecul...
- Eugene Koonin — Senior Investigator, NCBI, NIH

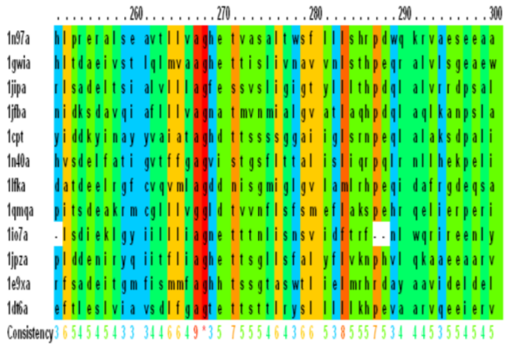
PSI-BLAST

Problem to solve: how to identify a set of sequences from a library, which are all homologous to the query sequence of interest?



MVLSEGEWQLVLHVWAKVEADVA
GHGQDILIRLFKSHPETLEKFDRVK
HLKTEAEM

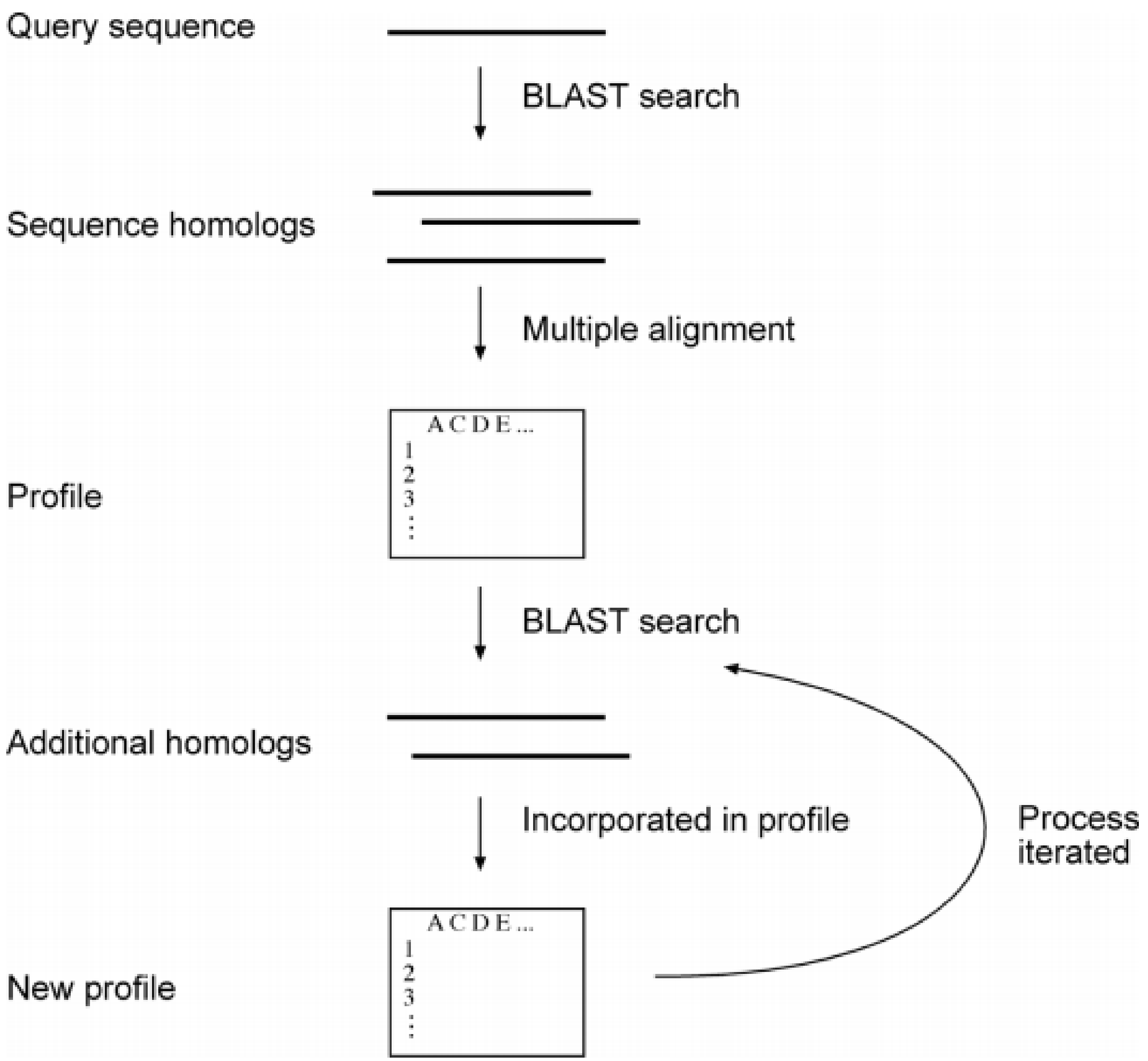
Sequence library



```
           .........260.......270.......280.......290.......300
1n97a       hlpreralse avtllvaghe tvasaltwsf lllshrpdwq krvaeseeaa
1gwia       hltdaeivst lqlmvaaghe ttislivnav vnlsthpeqr alvlsgeaew
1jipa       rlsadeltsi alvlllagfe ssvsligigt ylllthpdql alvrrdpsal
1jfba       nidksdavqi aflllvagna tmvnmialgv atlaqhpdql aqlkanpsla
1cpt        yiddkyinay yvaiataghd ttssssggai iglsrnpeql alaksdpali
1n40a       hvsdelfati gvtffgagvi stgsflttal isliqrpqlr nllhekpeli
1lfka       datdeelrgf cvqvmlagdd nisgmiglgv lamlrhpeqi dafrgdeqsa
1qmqa       pitsdeakrm cglllvggld tvvnflsfsm eflakspehr qelierperi
1io7a       -lsdieklgy iillliagne tttnlisnsv idftrf--nl wqrireenly
1jpza       plddeniryq iitfliaghe ttsgllsfal yflvknphvl qkaaeeaarv
1e9xa       rfsadeitgm fismmfaghh tssgtaswtl ielmrhrday aavideldel
1dt6a       eftleslvia vsdlfgagte ttsttlrysl llllkhpeva arvqeeierv
Consistency 3654545433 3446649*35 7555464366 5385557534 4453554545
```

MSA of related sequences

Flowchart of PSI-BLAST



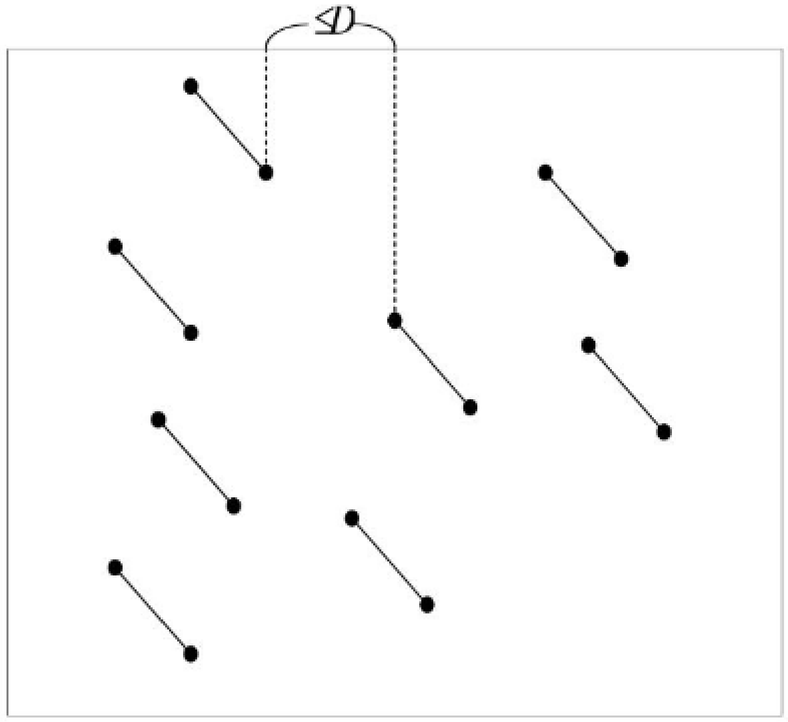
Difference between BLAST and PSI-BLAST

Observation from BLAST1.0:

- Extension step accounts for 90% of the total time in BLAST1.0
- HSP of interest is much longer than a single word pair

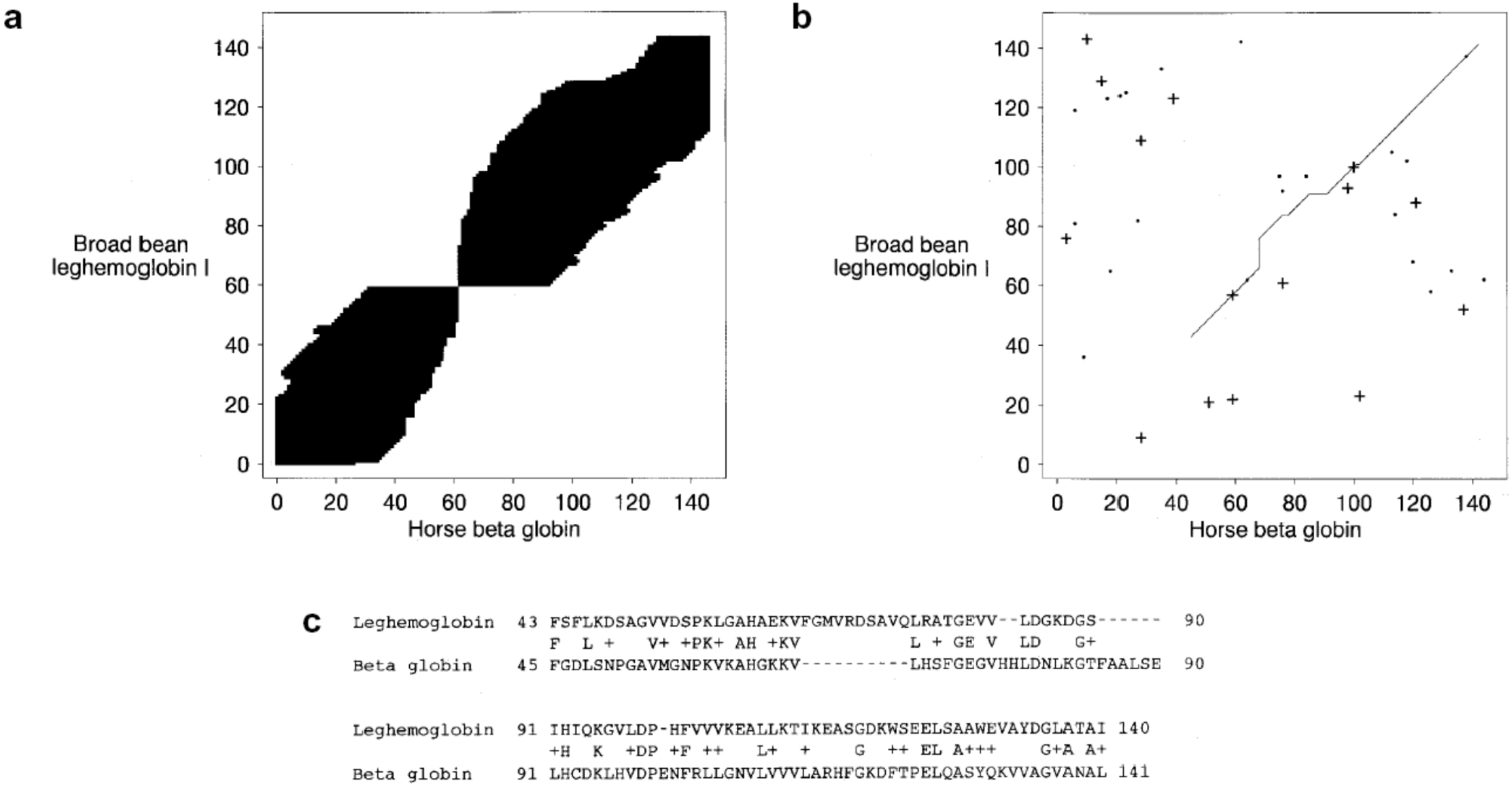
1. two-hit method

Invoke an extension only when two non-overlapping hits are found within distance *D* on the same diagonal



Difference between BLAST and PSI-BLAST

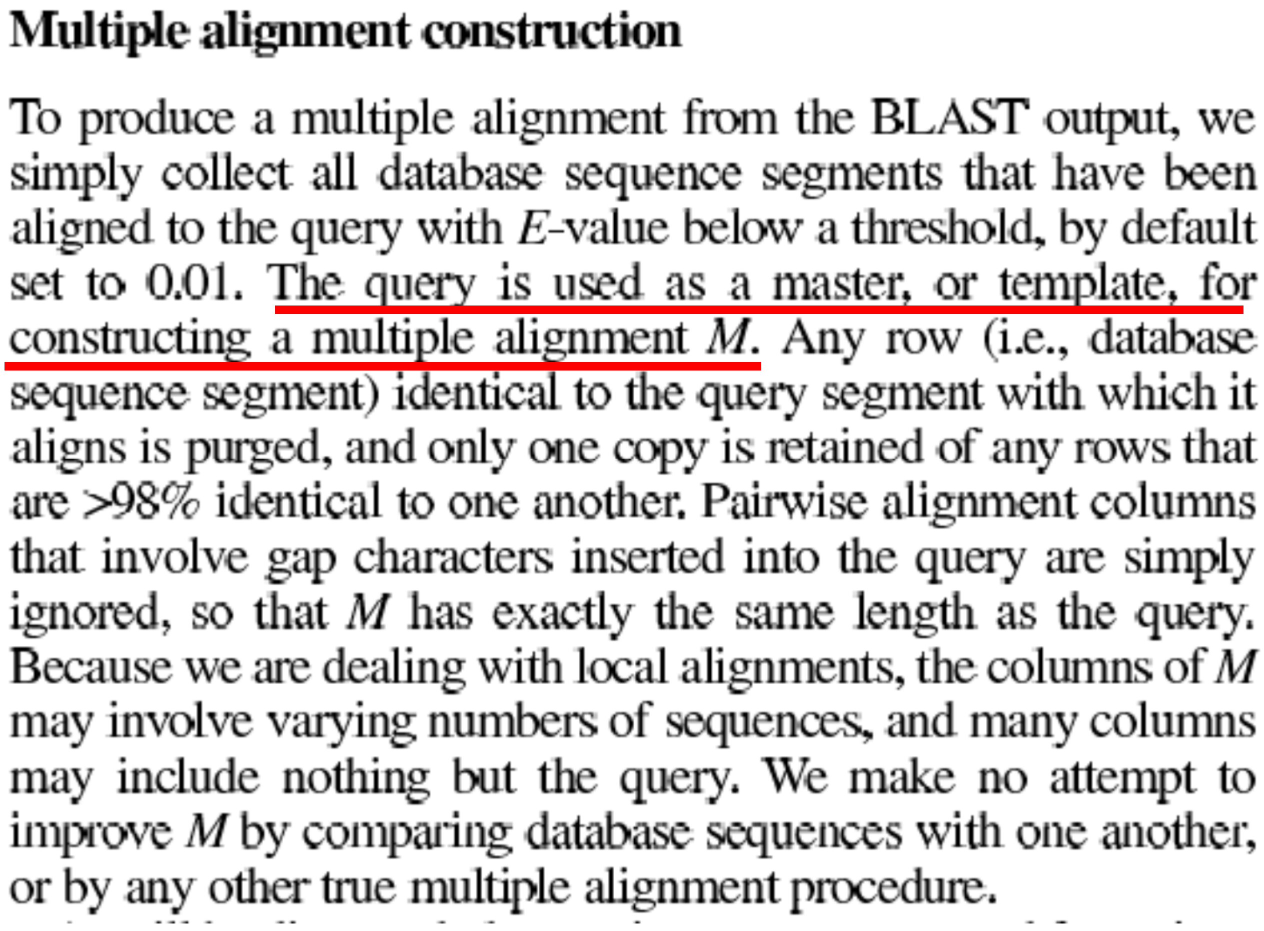
2. Dynamics programming extension of HSP allow gaps (vs. ungapped extension in BLAST1.0).



```
Leghemoglobin  43 FSFLKDSAGVVDSPKLGAHAEKVFGMVRDSAVQLRATGEVV--LDGKDGS------  90
                  F  L +   V+ +PK+ AH +KV          L + GE V  LD   G+
Beta globin    45 FGDLSNPGAVMGNPKVKAHGKKV----------LHSFGEGVHHLDNLKGTFAALSE  90

Leghemoglobin  91 IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLATAI 140
                  +H  K  +DP +F ++   L+  +    G  ++ EL A+++    G+A A+
Beta globin    91 LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANAL 141
```

Difference between BLAST and PSI-BLAST

3. Construct MSA of the homologous sequences based on pairwise alignments



Multiple alignment construction

To produce a multiple alignment from the BLAST output, we simply collect all database sequence segments that have been aligned to the query with *E*-value below a threshold, by default set to 0.01. The query is used as a master, or template, for constructing a multiple alignment *M*. Any row (i.e., database sequence segment) identical to the query segment with which it aligns is purged, and only one copy is retained of any rows that are >98% identical to one another. Pairwise alignment columns that involve gap characters inserted into the query are simply ignored, so that *M* has exactly the same length as the query. Because we are dealing with local alignments, the columns of *M* may involve varying numbers of sequences, and many columns may include nothing but the query. We make no attempt to improve *M* by comparing database sequences with one another, or by any other true multiple alignment procedure.

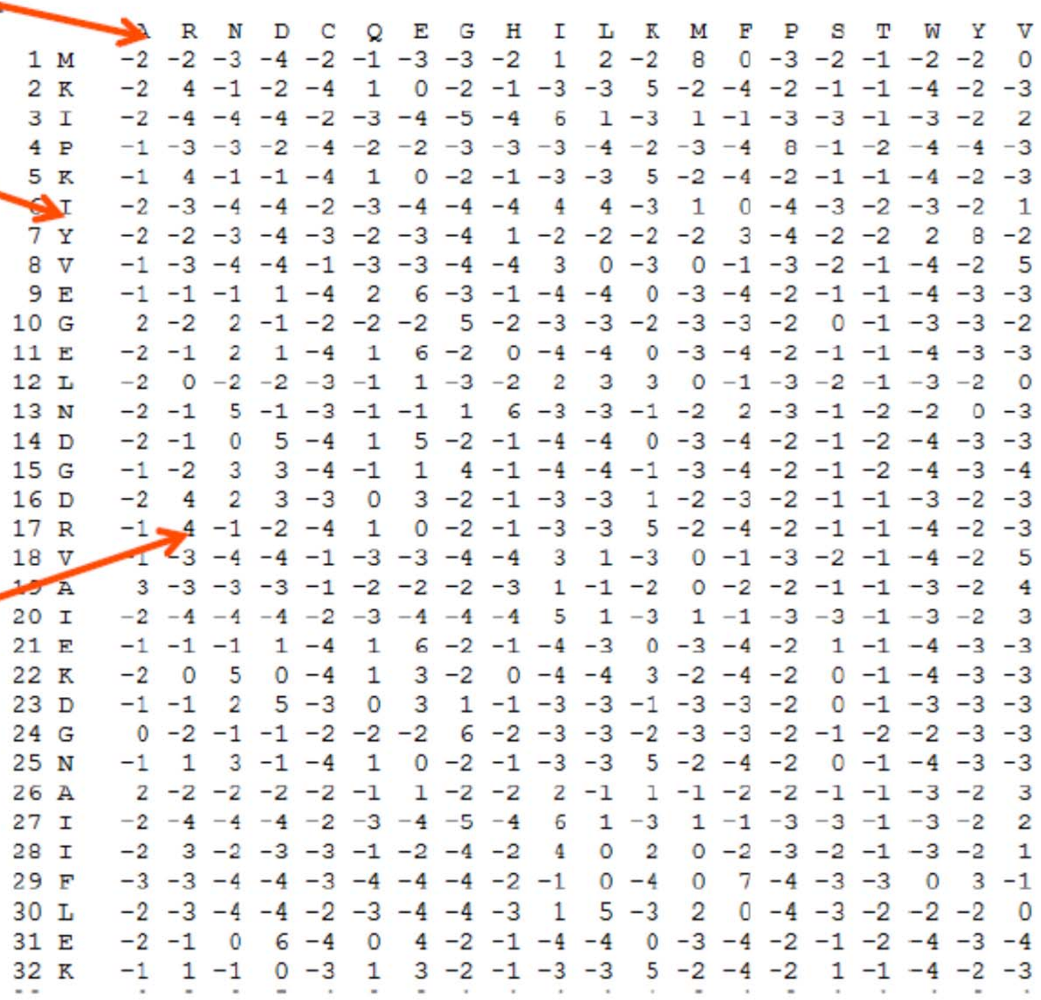
PSI-BLAST Profile

4. How to derive Position-Specific Score Matrix (PSSM)?

20 amino acids

Your query sequence

Log odds of amino acid "R" appears at 18th position of the sequence



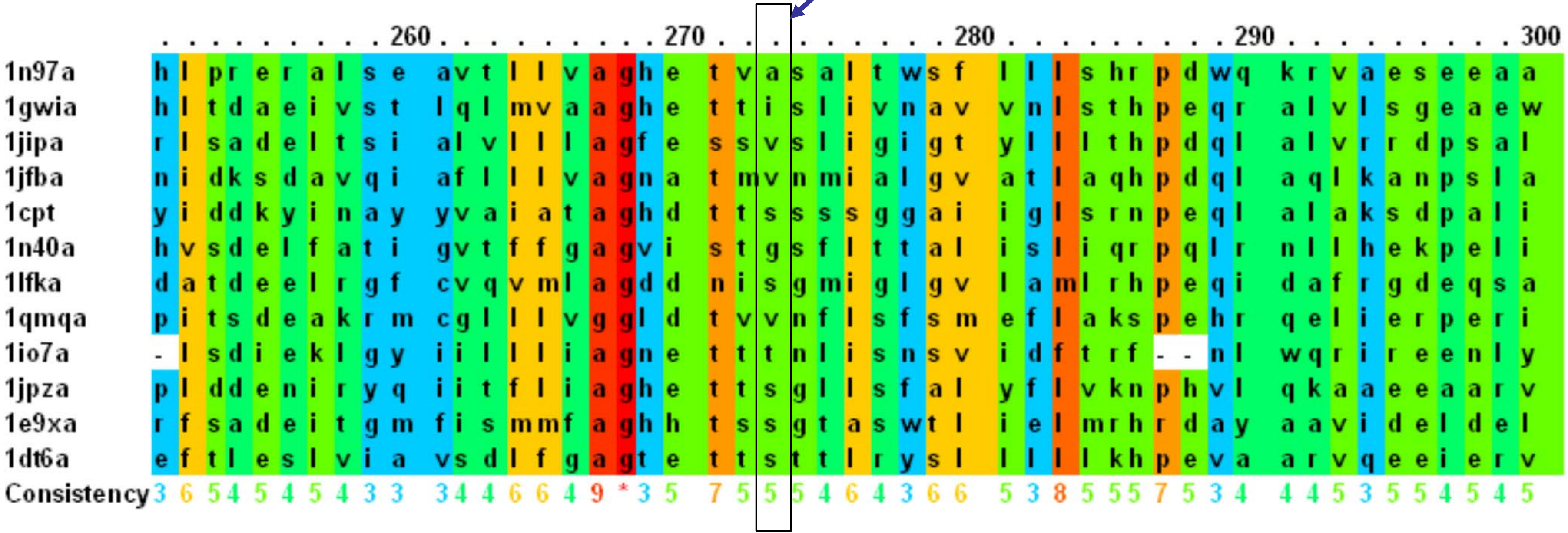
```
          A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
 1 M     -2 -2 -3 -4 -2 -1 -3 -3 -2  1  2 -2  8  0 -3 -2 -1 -2 -2  0
 2 K     -2  4 -1 -2 -4  1  0 -2 -1 -3 -3  5 -2 -4 -2 -1 -1 -4 -2 -3
 3 I     -2 -4 -4 -4 -2 -3 -4 -5 -4  6  1 -3  1 -1 -3 -3 -1 -3 -2  2
 4 P     -1 -3 -3 -2 -4 -2 -2 -3 -3 -3 -4 -2 -3 -4  8 -1 -2 -4 -4 -3
 5 K     -1  4 -1 -1 -4  1  0 -2 -1 -3 -3  5 -2 -4 -2 -1 -1 -4 -2 -3
 6 I     -2 -3 -4 -4 -2 -3 -4 -4 -4  4  4 -3  1  0 -4 -3 -2 -3 -2  1
 7 Y     -2 -2 -3 -4 -3 -2 -3 -4  1 -2 -2 -2 -2  3 -4 -2 -2  2  8 -2
 8 V     -1 -3 -4 -4 -1 -3 -3 -4 -4  3  0 -3  0 -1 -3 -2 -1 -4 -2  5
 9 E     -1 -1 -1  1 -4  2  6 -3 -1 -4 -4  0 -3 -4 -2 -1 -1 -4 -3 -3
10 G      2 -2  2 -1 -2 -2 -2  5 -2 -3 -3 -2 -3 -3 -2  0 -1 -3 -3 -2
11 E     -2 -1  2  1 -4  1  6 -2  0 -4 -4  0 -3 -4 -2 -1 -1 -4 -3 -3
12 L     -2  0 -2 -2 -3 -1  1 -3 -2  2  3  3  0 -1 -3 -2 -1 -3 -2  0
13 N     -2 -1  5 -1 -3 -1 -1  1  6 -3 -3 -1 -2  2 -3 -1 -2 -2  0 -3
14 D     -2 -1  0  5 -4  1  5 -2 -1 -4 -4  0 -3 -4 -2 -1 -2 -4 -3 -3
15 G     -1 -2  3  3 -4 -1  1  4 -1 -4 -4 -1 -3 -4 -2 -1 -2 -4 -3 -4
16 D     -2  4  2  3 -3  0  3 -2 -1 -3 -3  1 -2 -3 -2 -1 -1 -3 -2 -3
17 R     -1  4 -1 -2 -4  1  0 -2 -1 -3 -3  5 -2 -4 -2 -1 -1 -4 -2 -3
18 V     -1 -3 -4 -4 -1 -3 -3 -4 -4  3  1 -3  0 -1 -3 -2 -1 -4 -2  5
19 A      3 -3 -3 -3 -1 -2 -2 -2 -3  1 -1 -2  0 -2 -2 -1 -1 -3 -2  4
20 I     -2 -4 -4 -4 -2 -3 -4 -4 -4  5  1 -3  1 -1 -3 -3 -1 -3 -2  3
21 E     -1 -1 -1  1 -4  1  6 -2 -1 -4 -3  0 -3 -4 -2  1 -1 -4 -3 -3
22 K     -2  0  5  0 -4  1  3 -2  0 -4 -4  3 -2 -4 -2  0 -1 -4 -3 -3
23 D     -1 -1  2  5 -3  0  3  1 -1 -3 -3 -1 -3 -3 -2  0 -1 -3 -3 -3
24 G      0 -2 -1 -1 -2 -2 -2  6 -2 -3 -3 -2 -3 -3 -2 -1 -2 -2 -3 -3
25 N     -1  1  3 -1 -4  1  0 -2 -1 -3 -3  5 -2 -4 -2  0 -1 -4 -3 -3
26 A      2 -2 -2 -2 -2 -1  1 -2 -2  2 -1  1 -1 -2 -2 -1 -1 -3 -2  3
27 I     -2 -4 -4 -4 -2 -3 -4 -5 -4  6  1 -3  1 -1 -3 -3 -1 -3 -2  2
28 I     -2  3 -2 -3 -3 -1 -2 -4 -2  4  0  2  0 -2 -3 -2 -1 -3 -2  1
29 F     -3 -3 -4 -4 -3 -4 -4 -4 -2 -1  0 -4  0  7 -4 -3 -3  0  3 -1
30 L     -2 -3 -4 -4 -2 -3 -4 -4 -3  1  5 -3  2  0 -4 -3 -2 -2 -2  0
31 E     -2 -1  0  6 -4  0  4 -2 -1 -4 -4  0 -3 -4 -2 -1 -2 -4 -3 -4
32 K     -1  1 -1  0 -3  1  3 -2 -1 -3 -3  5 -2 -4 -2  1 -1 -4 -2 -3
..        .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .  .
```

PSSM derivation

Residue A_i (i=1,2…,20)

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVL

Position j (j=1,2…,L)



```
            . . . . . . . . . 260 . . . . . . . . . 270 . . . . . . . . . 280 . . . . . . . . . 290 . . . . . . . . . 300
1n97a       hlpreralse avtllvaghe tvasaltwsf lllshrpdwq krvaeseeaa
1gwia       hltdaeivst lqlmvaaghe ttislivnav vnlsthpeqr alvlsgeaew
1jipa       rlsadeltsi alvlllagfe ssvsligigt yllllthpdql alvrrdpsal
1jfba       nidksdavqi aflllvagna tmvnmialgv atlaqhpdql aqlkanpsla
1cpt        yiddkyinay yvaiataghd ttssssggai iglsrnpeql alaksdpali
1n40a       hvsdelfati gvtffgagvi stgsflttal isliqrpqlr nllhekpeli
1lfka       datdeelrgf cvqvmlagdd nisgmiglgv lamlrhpeqi dafrgdeqsa
1qmqa       pitsdeakrm cglllvggld tvvnflsfsm eflakspehr qelierperi
1io7a       -lsdieklgy iilllliagne tttnlisnsv idftrf--nl wqrireenly
1jpza       plddeniryq iitfliaghe ttsgllsfal yflvknphvl qkaaeeaarv
1e9xa       rfsadeitgm fismmfaghh tssgtaswtl ielmrhrday aavideldel
1dt6a       eftleslvia vsdlfgagte ttsttlrysl llllkhpeva arvqeeierv
Consistency 3654545433 3446649*35 7555464366 5385557534 4453554545
```

Q_{ij}: Estimated probability of A_i to be found at position j.

P_i: background probability

$$S(i,j) = \log\frac{Q_{ij}}{P_i}, \quad 1 \le i \le 20, \;\; 1 \le j \le L$$

PSSM derivation

Position j (j=1,2…,L)

Residue A$_i$ (i=1,2…,20)

pseudocount

$$Q_{ij} = \frac{\alpha f_{ij} + \beta g_{ij}}{\alpha + \beta}$$

$\alpha = N_c - 1$ **is the number of different residues**

$\beta = 10$

$$g_{ij} = \sum_{a=1}^{20} \frac{f_{aj}}{P_a} q_{ia}$$

$$S_{ij} = 2\log_2 \frac{q_{ij}}{e_{ij}}, \quad 1 \le j \le i \le 20$$

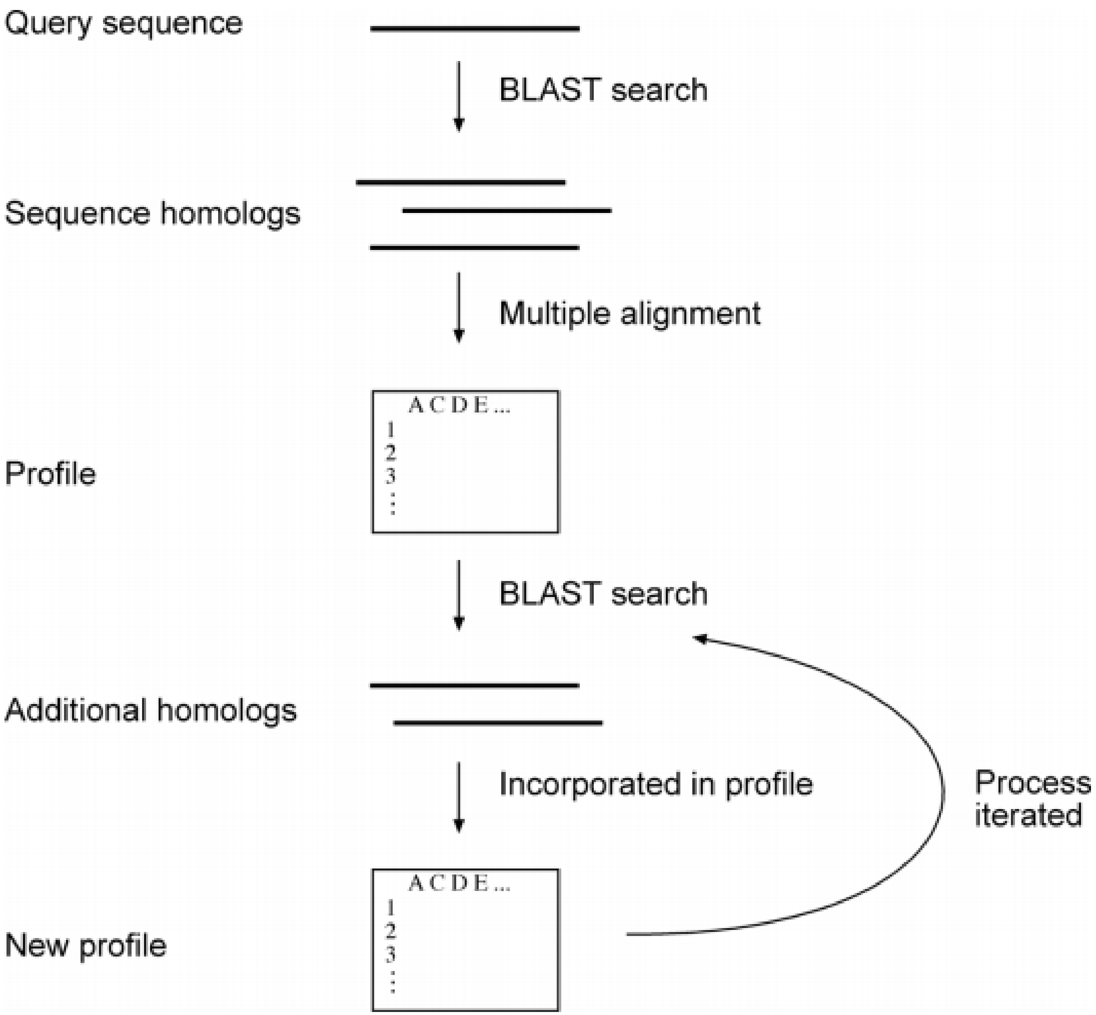
$$q_{ia} = P_i P_a e^{\lambda B(i,a)}$$

B(i, a): BLOSUM

$$\mathrm{S}(i,j) = \log \frac{\mathrm{Q}_{ij}}{P_i} = \log \frac{\alpha f_{ij} + \beta g_{ij}}{P_i(\alpha + \beta)} = \log \frac{\alpha f_{ij} + \beta P_i \sum_{a=1}^{20} f_{aj} e^{\lambda B(i,a)}}{P_i(\alpha + \beta)}$$

PSI-BLAST iteration

5. Perform sequence-profile alignment



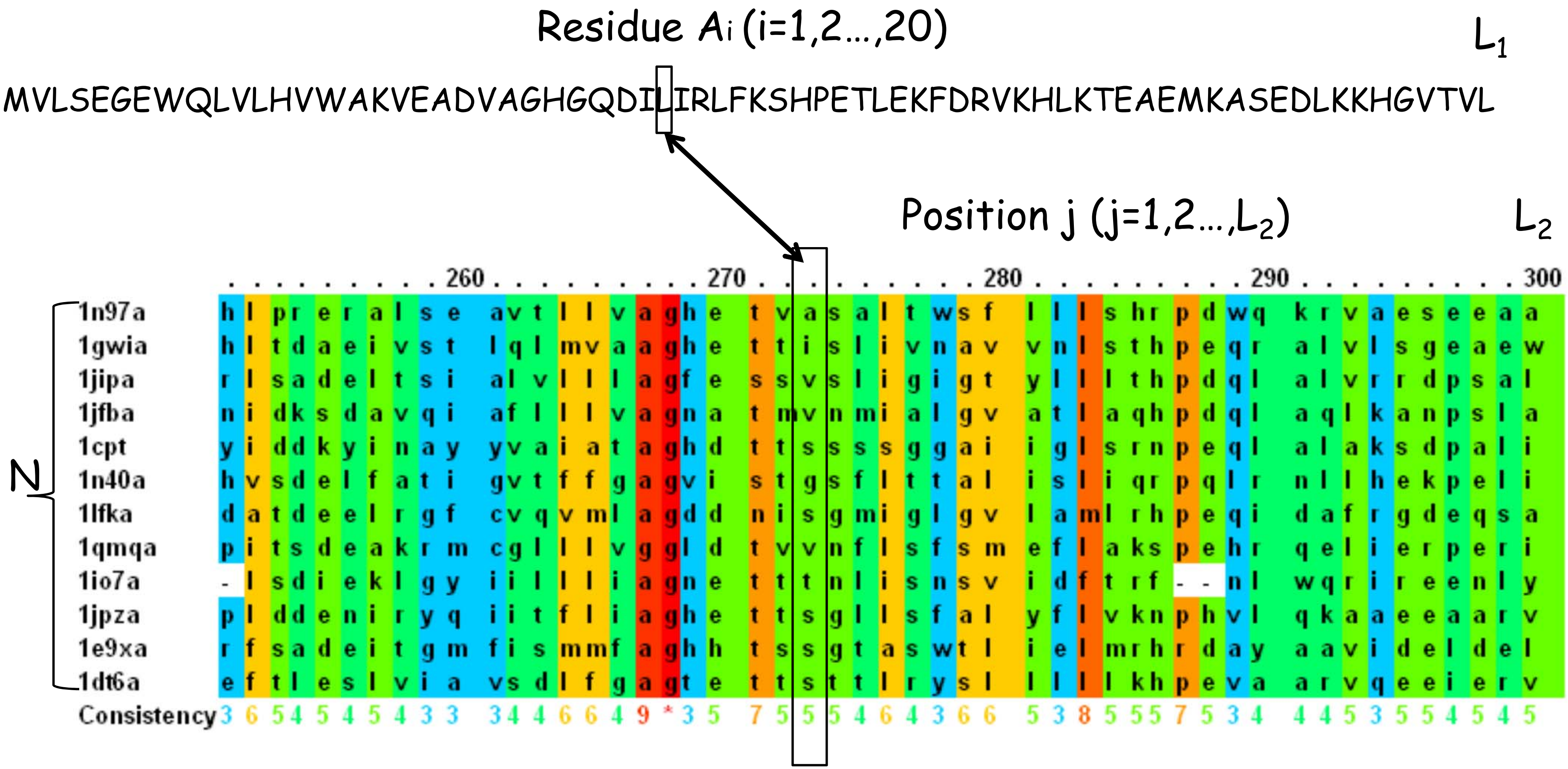
Content

1. Phylogenic tree
2. UPGMA & neighbor-joining methods
3. How to construct a MSA?
    a. ClusterW
    b. PSI-BLAST
4. Sequence profile & profile alignments
    ➡ a. What is a sequence profile?
    b. Profile-sequence alignment
    c. Profile-profile alignment

What is a sequence profile?

Sequence Profile:
Gribskov, Mclanchlan, Eisenberg. Profile analysis: Detection of distantly related proteins. PNAS (1987) 84, 4355-58.

What is a sequence profile?



Residue A_i (i=1,2…,20)

L_1

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVL

Position j (j=1,2…,L_2)

L_2

N

```
             .........260.........270.........280.........290.........300
1n97a        hlprerals e avtllvaghe tvasaltwsf lllshrpdwq krvaeseeaa
1gwia        hltdaeivst lqlmvaaghe ttislivnav vnlsthpeqr alvlsgeaew
1jipa        rlsadeltsi alvlllagfe ssvsligigt yllllthpdql alvrrdpsal
1jfba        nidksdavqi aflllvagna tmvnmialgv atlaqhpdql aqlkanpsla
1cpt         yiddkyinay yvaiataghd ttssssggai iglsrnpeql alaksdpali
1n40a        hvsdelfati gvtffgagvi stgsflttal isliqrpqlr nllhekpeli
1lfka        datdeelrgf cvqvmlagdd nisgmiglgv lamlrhpeqi dafrgdeqsa
1qmqa        pitsdeakrm cglllvggld tvvnflsfsm eflakspehr qelierperi
1io7a        -lsdieklgy iillliagne tttnlisnsv idftrf--nl wqrireenly
1jpza        plddeniryq iitfliaghe ttsgllsfal yflvknphvl qkaaeeaarv
1e9xa        rfsadeitgm fismmfaghh tssgtaswtl ielmrhrday aavidelde l
1dt6a        eftleslvia vsdlfgagte ttsttlrysl llllkhpeva arvqeeierv
Consistency  3654545433 3446649*35 7555464366 5385557534 4453554545
```

Gribskov, Mclanchlan, Eisenberg. Profile analysis: Detection of distantly related proteins. PNAS (1987) 84, 4355-58.

What is a sequence profile?

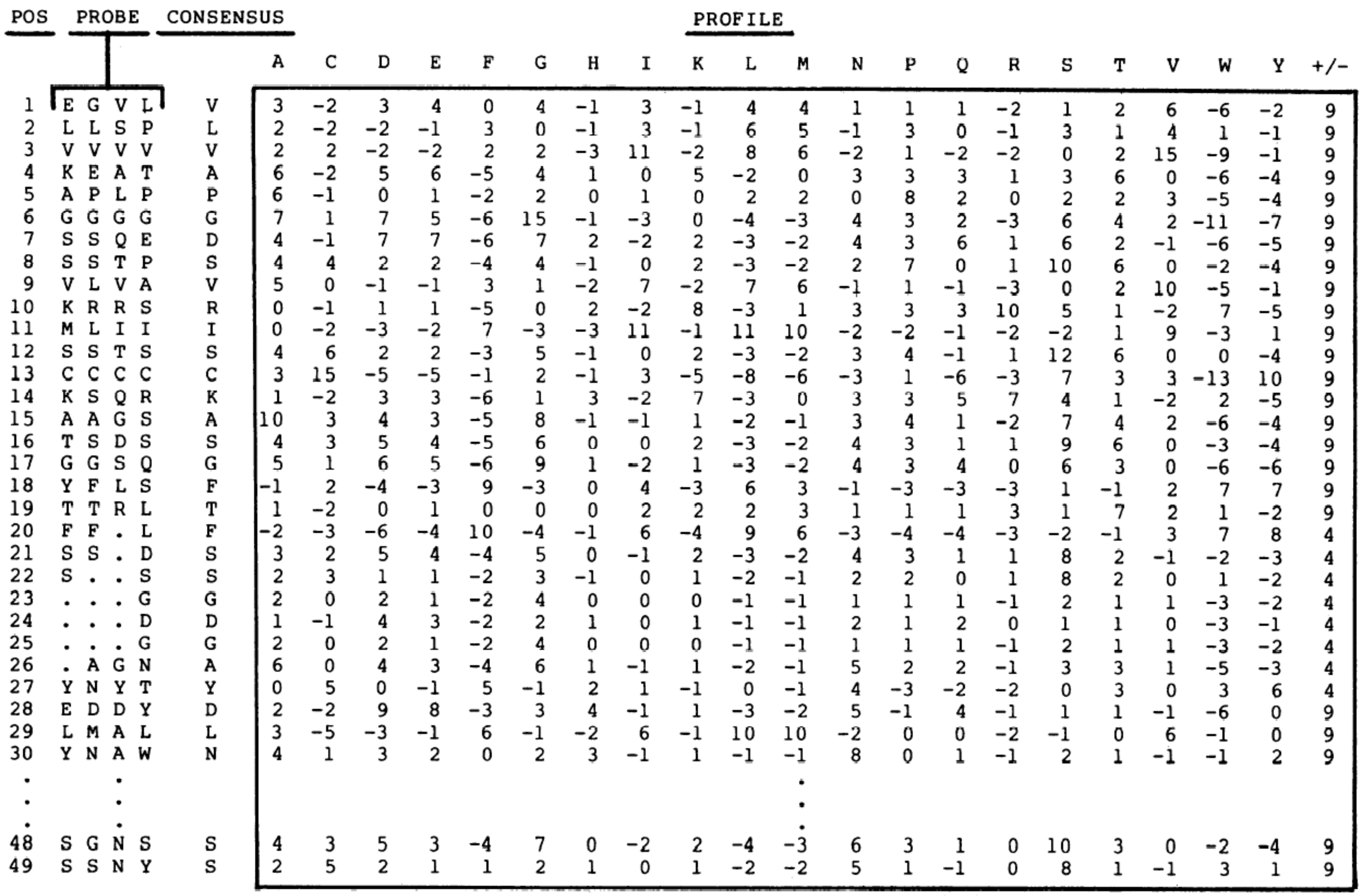
The alignment score for residue A_i and position j

$$
\begin{aligned}
&\mathrm{S}(i, j)\\
&= B(A_i, A_{j1}) + B(A_i, A_{j2}) + \cdots + B(A_i, A_{jN})\\
&= f_{jA} B(A_i, A) + f_{jR} B(A_i, R) + \cdots + f_{jV} B(A_i, V)\\
&= \sum_{a=1}^{20} f_{ja} B(A_i, a)\\
&:= p(j, A_i)
\end{aligned}
$$

If we list p(j,a) for all 20 possible amino acids at position j, we will get a L_2x20 matrix. This matrix is called <u>sequence profile of the N sequences</u>

What is a sequence profile?

An example profile



POS	PROBE	CONSENSUS										PROFILE										
			A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.	.												.										
.	.												.										
.	.												.										
48	S G N S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

How to weight the sequences in MSA?

Example & question:

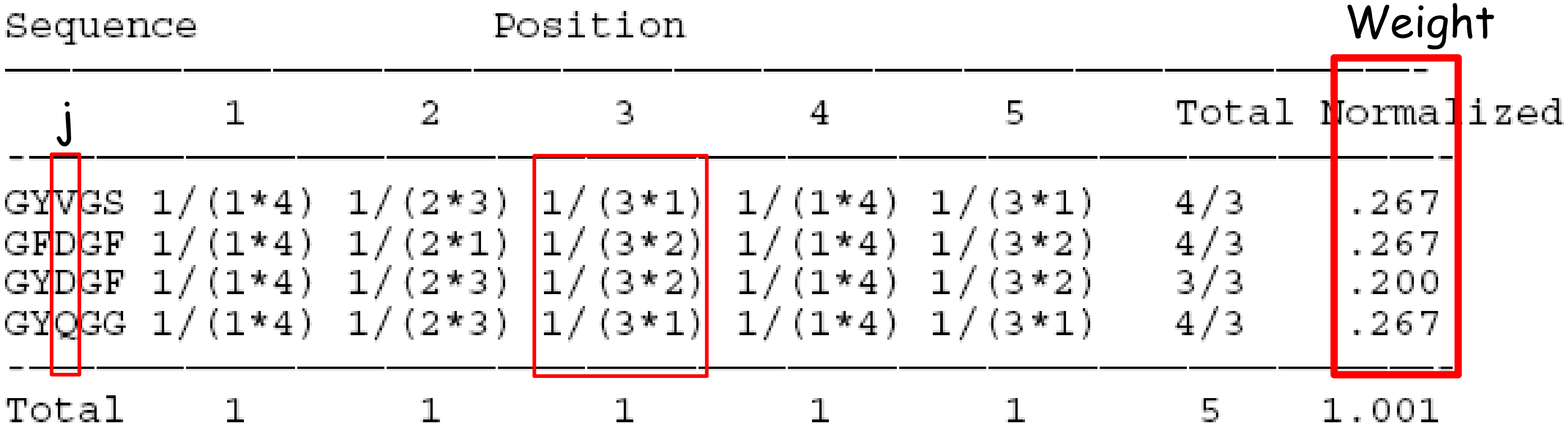
GYVGS
GFDGF
GYDGF
GYQGG

How to assign weight to each of the 4 sequences?

Principle: Give more weight to the non-redundant sequences

Henikoff & Henikoff weight:
Steven Henikoff and Jorja G. Henikoff, Position-based sequence weights, Journal of Molecular Biology. Volume 243, Issue 4, 4 November 1994, Pages 574-578

Henikoff & Henikoff weight



Weight

Sequence	Position						
j	1	2	3	4	5	Total	Normalized
GYVGS	1/(1*4)	1/(2*3)	1/(3*1)	1/(1*4)	1/(3*1)	4/3	.267
GFDGF	1/(1*4)	1/(2*1)	1/(3*2)	1/(1*4)	1/(3*2)	4/3	.267
GYDGF	1/(1*4)	1/(2*3)	1/(3*2)	1/(1*4)	1/(3*2)	3/3	.200
GYQGG	1/(1*4)	1/(2*3)	1/(3*1)	1/(1*4)	1/(3*1)	4/3	.267
Total	1	1	1	1	1	5	1.001

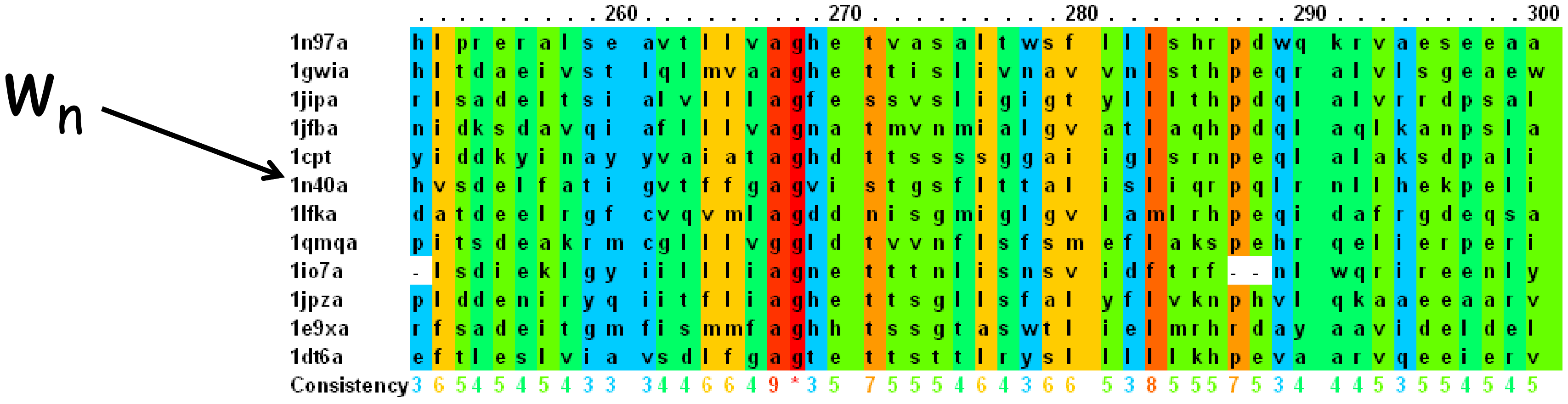
$$W_k = \sum_{j=1}^{L} w_{k_i,j} = \sum_{j=1}^{L} \frac{1}{n_j} \times \frac{1}{f_{k_i,j}}$$

Number of amino acid types at the *j*-th position

Number of occurrence for $A_{k,i}$ at the *j*-th position

Profile with sequence weight

$$
\begin{aligned}
& \mathrm{S}(i, j) \\
& = w_1 B(A_i, A_{j1}) + w_2 B(A_i, A_{j2}) + \cdots + w_N B(A_i, A_{jN}) \\
& = \sum_{n=1}^{f_{jA}} w_n B(A_i, A) + \sum_{n=1}^{f_{jR}} w_n B(A_i, R) + \cdots + \sum_{n=1}^{f_{jV}} w_n B(A_i, V) \\
& = f'_{jA} B(A_i, A) + f'_{jR} B(A_i, R) + \cdots + f'_{jV} B(A_i, V) \\
& = \sum_{a=1}^{20} f'_{ja} B(A_i, a) \\
& := p(j, A_i)
\end{aligned}
$$



```
           . . . . . . . . . 260 . . . . . . . . . 270 . . . . . . . . . 280 . . . . . . . . . 290 . . . . . . . . . 300
1n97a      h l p r e r a l s e a v t l l v a g h e t v a s a l t w s f l l l s h r p d w q k r v a e s e e a a
1gwia      h l t d a e i v s t l q l m v a a g h e t t i s l i v n a v v n l s t h p e q r a l v l s g e a e w
1jipa      r l s a d e l t s i a l v l l l a g f e s s v s l i g i g t y l l l t h p d q l a l v r r d p s a l
1jfba      n i d k s d a v q i a f l l l v a g n a t m v n m i a l g v a t l a q h p d q l a q l k a n p s l a
1cpt       y i d d k y i n a y y v a i a t a g h d t t s s s s g g a i i g l s r n p e q l a l a k s d p a l i
1n40a      h v s d e l f a t i g v t f f g a g v i s t g s f l t t a l i s l i q r p q l r n l l h e k p e l i
1lfka      d a t d e e l r g f c v q v m l a g d d n i s g m i g l g v l a m l r h p e q i d a f r g d e q s a
1qmqa      p i t s d e a k r m c g l l l v g g l d t v v n f l s f s m e f l a k s p e h r q e l i e r p e r i
1io7a      - l s d i e k l g y i i l l l i a g n e t t t n l i s n s v i d f t r f - - n l w q r i r e e n l y
1jpza      p l d d e n i r y q i i t f l i a g h e t t s g l l s f a l y f l v k n p h v l q k a a e e a a r v
1e9xa      r f s a d e i t g m f i s m m f a g h h t s s g t a s w t l i e l m r h r d a y a a v i d e l d e l
1dt6a      e f t l e s l v i a v s d l f g a g t e t t s t t l r y s l l l l l k h p e v a a r v q e e i e r v
Consistency3 6 5 4 5 4 5 4 3 3 3 4 4 6 6 4 9 * 3 5 7 5 5 5 4 6 4 3 6 6 5 3 8 5 5 5 7 5 3 4 4 4 5 3 5 5 4 5 4 5
```

What is a profile - summary

- Profile is a matrix representation of a MSA
- Profile = MSA (+) BLUSOM
- You have to have a MSA before you can construct a profile matrix
- This MSA can be pre-generated by CLUSTALW or PSI-BLAST

Content

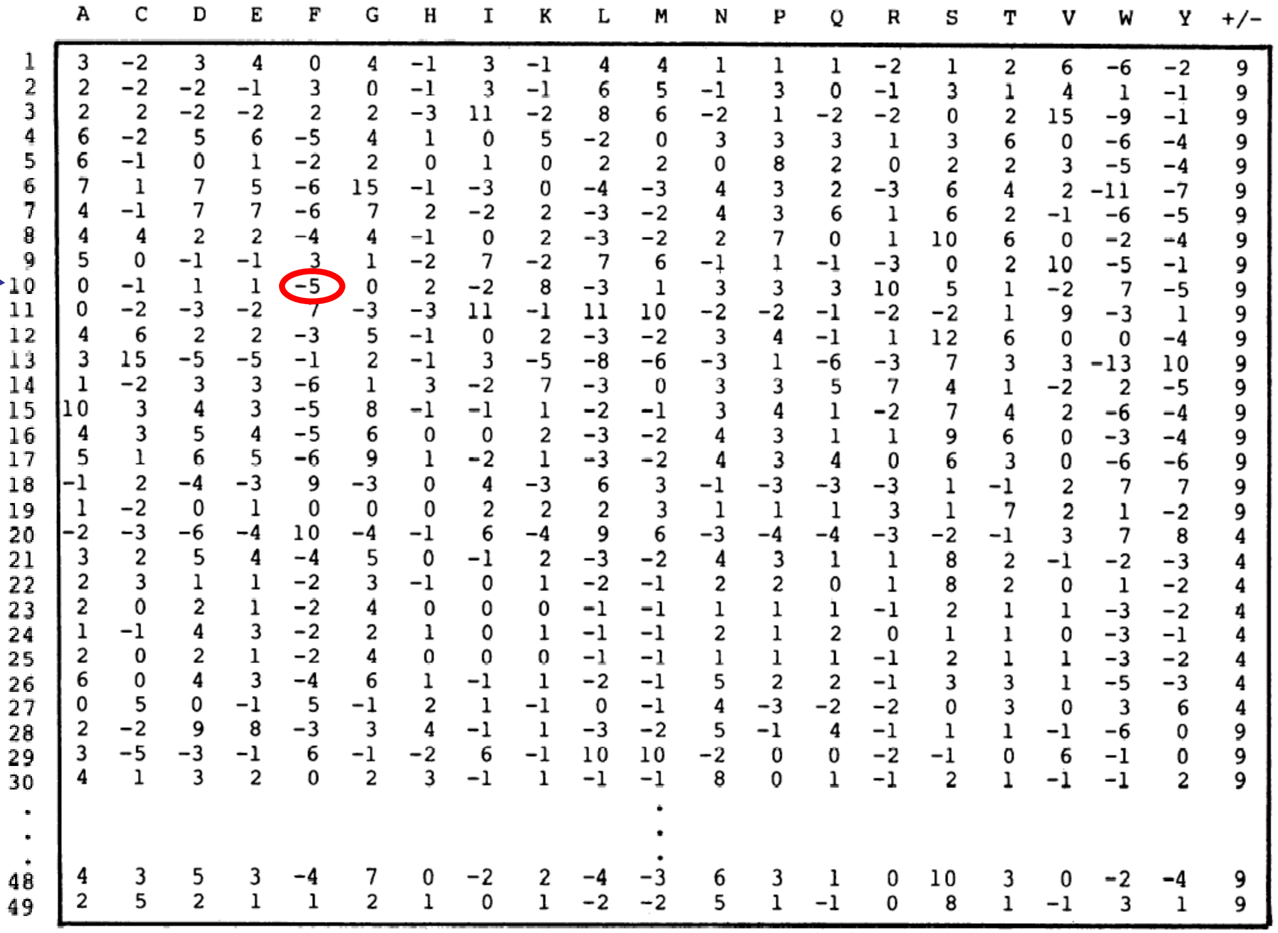
1. Phylogenic tree
2. UPGMA & neighboring-joining methods
3. How to construct a MSA?
   a. ClusterW
   b. PSI-BLAST
4. Sequence profile & profile alignments
   a. What is a sequence profile?
   b. Profile-sequence alignment
   c. Profile-profile alignment

Sequence-profile alignment

Template sequence

M
A
H
P
P
Q
F
I
R
I
P
A
T
Y
L
R
G
G
T
S
K
G
V
F
D

Query profile



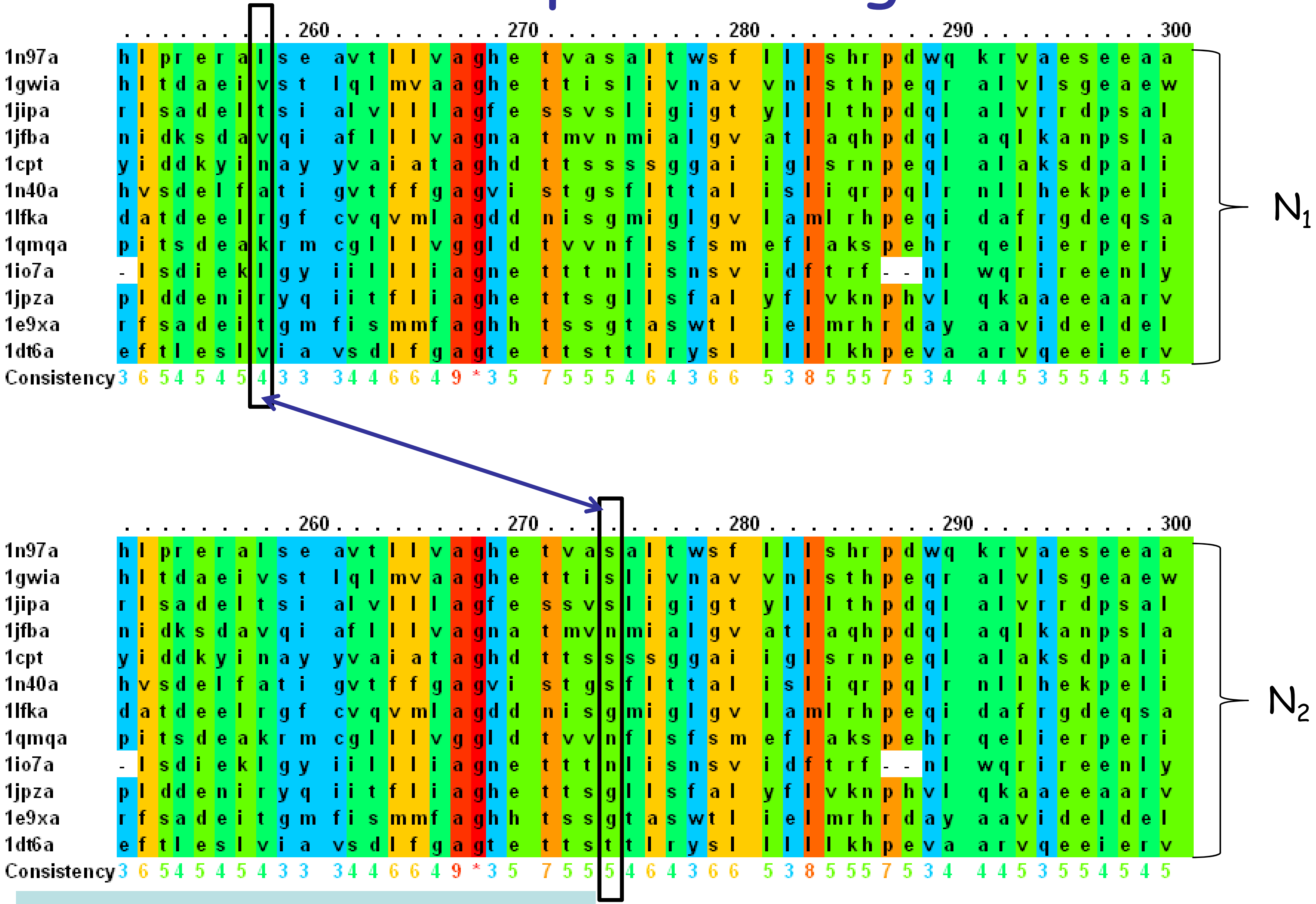
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	3	−2	3	4	0	4	−1	3	−1	4	4	1	1	1	−2	1	2	6	−6	−2	9
2	2	−2	−2	−1	3	0	−1	3	−1	6	5	−1	3	0	−1	3	1	4	1	−1	9
3	2	2	−2	−2	2	2	−3	11	−2	8	6	−2	1	−2	−2	0	2	15	−9	−1	9
4	6	−2	5	6	−5	4	1	0	5	−2	0	3	3	3	1	3	6	0	−6	−4	9
5	6	−1	0	1	−2	2	0	1	0	2	2	0	8	2	0	2	2	3	−5	−4	9
6	7	1	7	5	−6	15	−1	−3	0	−4	−3	4	3	2	−3	6	4	2	−11	−7	9
7	4	−1	7	7	−6	7	2	−2	2	−3	−2	4	3	6	1	6	2	−1	−6	−5	9
8	4	4	2	2	−4	4	−1	0	2	−3	−2	2	7	0	1	10	6	0	−2	−4	9
9	5	0	−1	−1	3	1	−2	7	−2	7	6	−1	1	−1	−3	0	2	10	−5	−1	9
10	0	−1	1	1	−5	0	2	−2	8	−3	1	3	3	3	10	5	1	−2	7	−5	9
11	0	−2	−3	−2	7	−3	−3	11	−1	11	10	−2	−2	−1	−2	−2	1	9	−3	1	9
12	4	6	2	2	−3	5	−1	0	2	−3	−2	3	4	−1	1	12	6	0	0	−4	9
13	3	15	−5	−5	−1	2	−1	3	−5	−8	−6	−3	1	−6	−3	7	3	3	−13	10	9
14	1	−2	3	3	−6	1	3	−2	7	−3	0	3	3	5	7	4	1	−2	2	−5	9
15	10	3	4	3	−5	8	−1	−1	1	−2	−1	3	4	1	−2	7	4	2	−6	−4	9
16	4	3	5	4	−5	6	0	0	2	−3	−2	4	3	1	1	9	6	0	−3	−4	9
17	5	1	6	5	−6	9	1	−2	1	−3	−2	4	3	4	0	6	3	0	−6	−6	9
18	−1	2	−4	−3	9	−3	0	4	−3	6	3	−1	−3	−3	−3	1	−1	2	7	7	9
19	1	−2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	−2	9
20	−2	−3	−6	−4	10	−4	−1	6	−4	9	6	−3	−4	−4	−3	−2	−1	3	7	8	4
21	3	2	5	4	−4	5	0	−1	2	−3	−2	4	3	1	1	8	2	−1	−2	−3	4
22	2	3	1	1	−2	3	−1	0	1	−2	−1	2	2	0	1	8	2	0	1	−2	4
23	2	0	2	1	−2	4	0	0	0	−1	−1	1	1	1	−1	2	1	1	−3	−2	4
24	1	−1	4	3	−2	2	1	0	1	−1	−1	2	1	2	0	1	1	0	−3	−1	4
25	2	0	2	1	−2	4	0	0	0	−1	−1	1	1	1	−1	2	1	1	−3	−2	4
26	6	0	4	3	−4	6	1	−1	1	−2	−1	5	2	2	−1	3	3	1	−5	−3	4
27	0	5	0	−1	5	−1	2	1	−1	0	−1	4	−3	−2	−2	0	3	0	3	6	4
28	2	−2	9	8	−3	3	4	−1	1	−3	−2	5	−1	4	−1	1	1	−1	−6	0	9
29	3	−5	−3	−1	6	−1	−2	6	−1	10	10	−2	0	0	−2	−1	0	6	−1	0	9
30	4	1	3	2	0	2	3	−1	1	−1	−1	8	0	1	−1	2	1	−1	−1	2	9
.											.										
.											.										
.											.										
48	4	3	5	3	−4	7	0	−2	2	−4	−3	6	3	1	0	10	3	0	−2	−4	9
49	2	5	2	1	1	2	1	0	1	−2	−2	5	1	−1	0	8	1	−1	3	1	9

Algorithm: Dynamic programming

Content

1. Phylogenic tree
2. UPGMA & neighbor-joining methods
3. How to construct a MSA?
    a. ClusterW
    b. PSI-BLAST
4. Sequence profile & profile alignments
    a. What is a sequence profile?
    b. Profile-sequence alignment
    ➡ c. Profile-profile alignment

Profile-profile alignment



```
           . . . . . . . . . 260 . . . . . . . . . 270 . . . . . . . . . 280 . . . . . . . . . 290 . . . . . . . . . 300
1n97a       h l p r e r a l s e a v t l l v a g h e t v a s a l t w s f l l l s h r p d w q k r v a e s e e a a
1gwia       h l t d a e i v s t l q l m v a a g h e t t i s l i v n a v v n l s t h p e q r a l v l s g e a e w
1jipa       r l s a d e l t s i a l v l l l a g f e s s v s l i g i g t y l l l t h p d q l a l v r r d p s a l
1jfba       n i d k s d a v q i a f l l l v a g n a t m v n m i a l g v a t l a q h p d q l a q l k a n p s l a
1cpt        y i d d k y i n a y y v a i a t a g h d t t s s s s g g a i i g l s r n p e q l a l a k s d p a l i
1n40a       h v s d e l f a t i g v t f f g a g v i s t g s f l t t a l i s l i q r p q l r n l l h e k p e l i
1lfka       d a t d e e l r g f c v q v m l a g d d n i s g m i g l g v l a m l r h p e q i d a f r g d e q s a
1qmqa       p i t s d e a k r m c g l l l v g g l d t v v n f l s f s m e f l a k s p e h r q e l i e r p e r i
1io7a       - l s d i e k l g y i i l l l i a g n e t t t n l i s n s v i d f t r f - - n l w q r i r e e n l y
1jpza       p l d d e n i r y q i i t f l i a g h e t t s g l l s f a l y f l v k n p h v l q k a a e e a a r v
1e9xa       r f s a d e i t g m f i s m m f a g h h t s s g t a s w t l i e l m r h r d a y a a v i d e l d e l
1dt6a       e f t l e s l v i a v s d l f g a g t e t t s t t l r y s l l l l l k h p e v a a r v q e e i e r v
Consistency 3 6 5 4 5 4 5 4 3 3 3 4 4 6 6 4 9 * 3 5 7 5 5 5 4 6 4 3 6 6 5 3 8 5 5 5 7 5 3 4 4 4 5 3 5 5 4 5 4 5
```

```
           . . . . . . . . . 260 . . . . . . . . . 270 . . . . . . . . . 280 . . . . . . . . . 290 . . . . . . . . . 300
1n97a       h l p r e r a l s e a v t l l v a g h e t v a s a l t w s f l l l s h r p d w q k r v a e s e e a a
1gwia       h l t d a e i v s t l q l m v a a g h e t t i s l i v n a v v n l s t h p e q r a l v l s g e a e w
1jipa       r l s a d e l t s i a l v l l l a g f e s s v s l i g i g t y l l l t h p d q l a l v r r d p s a l
1jfba       n i d k s d a v q i a f l l l v a g n a t m v n m i a l g v a t l a q h p d q l a q l k a n p s l a
1cpt        y i d d k y i n a y y v a i a t a g h d t t s s s s g g a i i g l s r n p e q l a l a k s d p a l i
1n40a       h v s d e l f a t i g v t f f g a g v i s t g s f l t t a l i s l i q r p q l r n l l h e k p e l i
1lfka       d a t d e e l r g f c v q v m l a g d d n i s g m i g l g v l a m l r h p e q i d a f r g d e q s a
1qmqa       p i t s d e a k r m c g l l l v g g l d t v v n f l s f s m e f l a k s p e h r q e l i e r p e r i
1io7a       - l s d i e k l g y i i l l l i a g n e t t t n l i s n s v i d f t r f - - n l w q r i r e e n l y
1jpza       p l d d e n i r y q i i t f l i a g h e t t s g l l s f a l y f l v k n p h v l q k a a e e a a r v
1e9xa       r f s a d e i t g m f i s m m f a g h h t s s g t a s w t l i e l m r h r d a y a a v i d e l d e l
1dt6a       e f t l e s l v i a v s d l f g a g t e t t s t t l r y s l l l l l k h p e v a a r v q e e i e r v
Consistency 3 6 5 4 5 4 5 4 3 3 3 4 4 6 6 4 9 * 3 5 7 5 5 5 4 6 4 3 6 6 5 3 8 5 5 5 7 5 3 4 4 4 5 3 5 5 4 5 4 5
```

Algorithm: Dynamic programming

Profile-profile alignment

<u>**References**</u>

- **Anna R. Panchenko. Finding weak similarities between proteins by sequence profile comparison. Nucleic Acids Research, 2003, Vol. 31, No. 2 683.**
(This paper is to introduce what is the sequence profile-profile alignment. The key is to understand how to derive the alignment scoring function)

- **Edgar & Sjolander, A comparison of scoring functions for protein sequence profile alignment. Bioinformatics (2004) 20, 1301-8**
(This paper is to compare the result of different ways to make the profile-profile alignments. The key is to understand the different formulas for representing profile-profile comparison)

Profile-profile alignment

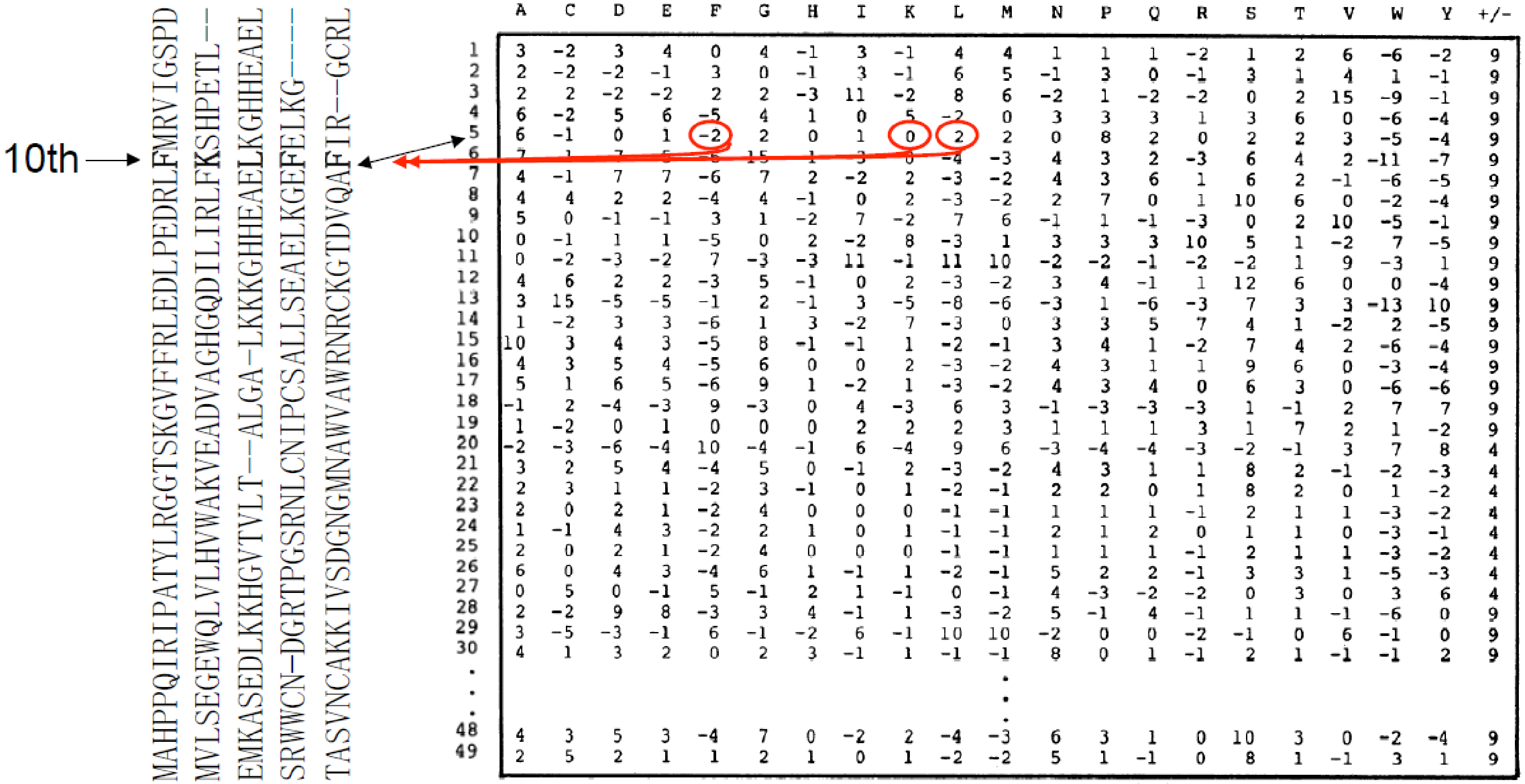
Score of aligning position *i* with position j

$$
\begin{aligned}
\mathbf{S}(i,j) &= \sum_{k=1}^{N_1}\sum_{l=1}^{N_2} B(A_{ik},A_{jl}) \\
&= \sum_{k=1}^{N_1}[B(A_{ik},A_{j1})+B(A_{ik},A_{j2})+\cdots+B(A_{ik},A_{jN_2})] \\
&= \sum_{k=1}^{N_1}[f_{jA}B(A_{ik},A)+f_{jR}B(A_{ik},R)+\cdots+f_{jV}B(A_{ik},V)] \\
&= \sum_{k=1}^{N_1}\sum_{b=1}^{20} f_{jb}B(A_{ik},b) \\
&= \sum_{a=1}^{20}\sum_{b=1}^{20} f_{ia}f_{jb}B(a,b) \\
&= \sum_{a=1}^{20} f_{ia}[\sum_{b=1}^{20} f_{jb}B(a,b)] \\
&= \sum_{a=1}^{20} f_{ia}\,p(j,a) \\
&= \vec{f_i}\bullet\vec{p_j}
\end{aligned}
$$

Frequency vector

Log-odds vector

Profile-profile alignment



Query sequence (MSA)

```
MAHPPQIRIPATYLRGGTSKGVFFRLEDLPEDRLFMRVIGSPD
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETL--
EMKASEDLKKHGVTVLT--ALGA-LKKKGHHEAELKGHHEAEL
SRWWCN-DGRTPGSRNLCNIPCSALLSEAELKGEFELKG----
TASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAFIR--GCRL
```

Template profile

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	7	1	7	5	-6	15	1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
⋮											⋮										
48	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

$S(10,5)=(-2)*3+0*1+2*1=-4$

Profile-profile alignment

Performance:

- **Profile-profile** alignment ~ 3% better than **Profile-sequence** alignment
- **Profile-profile** alignment ~ 40% better than **sequence-sequence** alignment

[Ref: Edgar & Sjolander, Bioinformatics (2004) 20, 1301-8]

Content

1. Bioinformatics databases
2. Sequence alignment and database searching
3. Phylogenic tree and multiple sequence alignment

➡ 4. Protein structure alignment

5. Protein secondary structure prediction
6. Protein tertiary structure prediction

Papers to read

- **RMSD**

W. Kabsch, A solution for the best rotation to relate two sets of vectors Acta Cryst (1976) A32: 922-923

- **TM-score**

Yang Zhang, Jeffrey Skolnick. A scoring function for the automated assessment of protein structure template quality. Proteins, vol 57, 702 (2004).

Papers to read

TM-align:
Yang Zhang, Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, vol 33, 2302 (2005).

mTM-align:
Dong et al. mTM-align: an algorithm for fast and accurate multiple protein structure alignment, Bioinformatics, 2017 .