

多序列比对与进化树

Content



1. How to construct a MSA?

- a. ClusterW
- b. PSI-BLAST

2. Sequence profile & profile alignments

- a. What is a sequence profile?
- b. Profile-sequence alignment
- c. Profile-profile alignment

多序列比对

Multiple sequence alignment (MSA)

Content

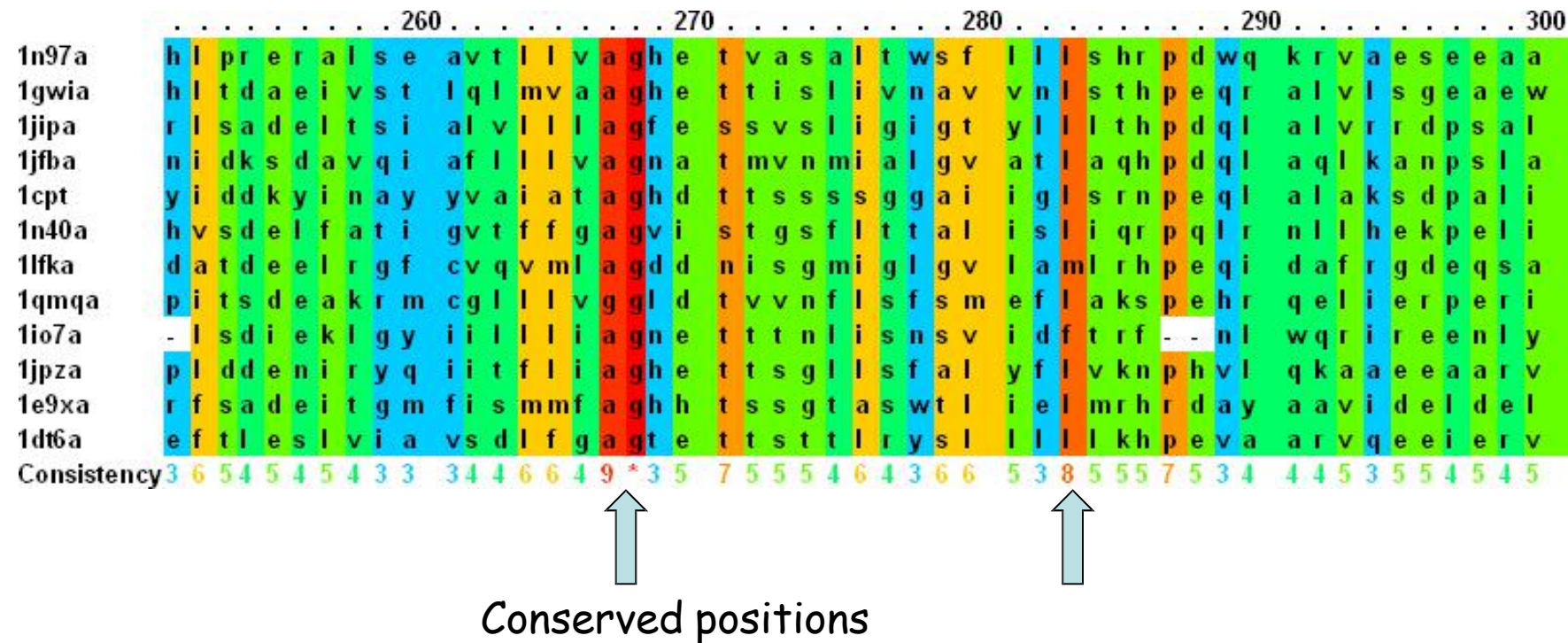
1. Phylogenetic tree
2. UPGMA & neighboring-joining methods
3. How to construct a MSA?



- a. ClusterW
- b. PSI-BLAST

4. Sequence profile & profile alignments
 - a. What is a sequence profile?
 - b. Profile-sequence alignment
 - c. Profile-profile alignment

Multiple sequence alignment

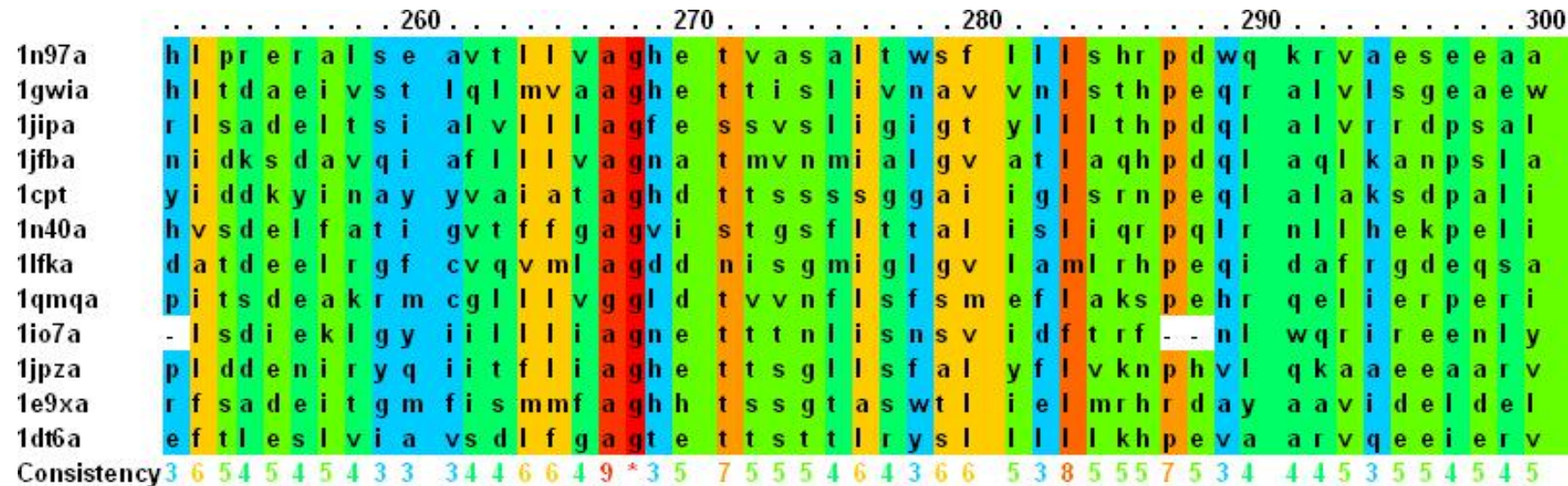


Multiple sequence alignment: Take three or more sequences and align them so that the greatest number of similar characters are aligned in the same column of the alignment.

Multiple sequence alignment

Protein 1 MSAADLLRLVGPRWVRPRRLGRIPDQPIVHAVRETAPGMLADQLSDHLATIVPHAELHVGDAARGTERERSVQVRTLLDTAVL
Protein 2 GLREHDSWPRIGRLQFPRYALTSWLLKQNLPAELNHAPHSNIRDLLHDFLNSRRRPGRGK
Protein 3 QNAREAAAWTSMTEQLPWYLFLLSLVAFPFYYALWVRGKVPRWFLRQQYLAPR
Protein 4 ESADFPSFVRRLLITTPSERESAEQVRRLLVHAFSLSDSHSRRLLWRWRWPKDCYPVLLLKDRLPGTIGETLVRLVNNVRNETGARDPLLVVATGEQPLEDGE
Protein 5 TPRAPVTLEQWERDLQAARRKRSPTAWYVPLRIADEPADALDYDRFGALGRAHLPLKRSKLVRRTPLLLVLVLLVGSTAGYAGYLRTHCGQWWPYQNSDIGEVDGECIGV
Protein 6 SDDTTSRFFSAHDARMVAAQEKIAEQNEEAERRWEDQPNLPHPTVVYFSTFPSSDDDPPTLAGIADEL DGVAVMQRESLGRNVLMKVVLAN
Protein 7 GGLRMKHGPRVAADVAELVGRDDSVVAVAGLGGSWQATVDTIEALEAGVPMVGTTISADLLESSPLFYQVAPSNABEA
Protein 8 KVVANYIAAGPVDPRGTGAPRRPDNVL IYSNPRDLYSHDLAQLTAGELRARGIEPMPDSDRIPCGKQNLVFFAGR
Protein 9 ANDLATFLTKMPPECGKPENYPQLLAGDDTSKLVDDAMDDHEGVVLDHVSFTGRSAWDPQSQQGTPLRGRGLLARDALEVIALAVQ
...

How?



多序列比对的计算方法



- ❑ 渐进方法: **Progressive methods**
 - ❑ **ClustalW**
- ❑ 迭代方法: **Iterative refinement**
 - ❑ **PSI-BLAST**, **DIALIGN**, **PRRP**
- ❑ 部分有向图算法:
 - ❑ **POA**
- ❑ 隐马尔科夫模型: **HMM profile-profile**
 - ❑ **ProbCons**
- ❑ 整合算法: **Meta-methods**
 - ❑ **MUSCLE**

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 - academic.oup.com

Abstract The sensitivity of the commonly used progressive multiple sequence alignment method has been greatly improved for the alignment of divergent protein sequences. Individual weights are assigned to each sequence in a partial alignment in order to

☆ 被引用次数 : 57276 相关文章 所有 58 个版本



The screenshot shows the top navigation bar of the Nature journal website. The main header reads 'nature International weekly journal of science'. Below this is a navigation menu with links: Home, News & Comment, Research, Careers & Jobs, Current Issue, and Archive. A secondary navigation bar includes links for Archive, Volume 514, Issue 7524, News Feature, and Article. The main content area is titled 'NATURE | NEWS FEATURE' and 'عربي'. It features a section titled 'The top 100 papers' with the subtitle 'Nature explores the most-cited research of all time.' and lists the authors 'Richard Van Noorden, Brendan Maher & Regina Nuzzo'. The date '29 October 2014' is displayed at the bottom.

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

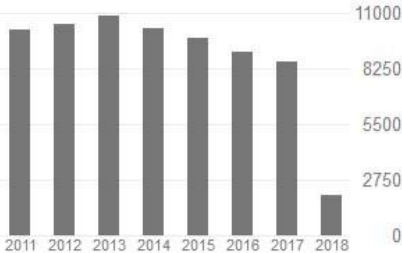


标题	引用次数	年份
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix ch... JD Thompson, DG Higgins, TJ Gibson Nucleic acids research 22 (22), 4673	57668	1994
The CLUSTAL_X windows interface: flexible strategies 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

创建我的个人资料

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

- ❑ 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

Progressive method: ClustalW

Step 1:

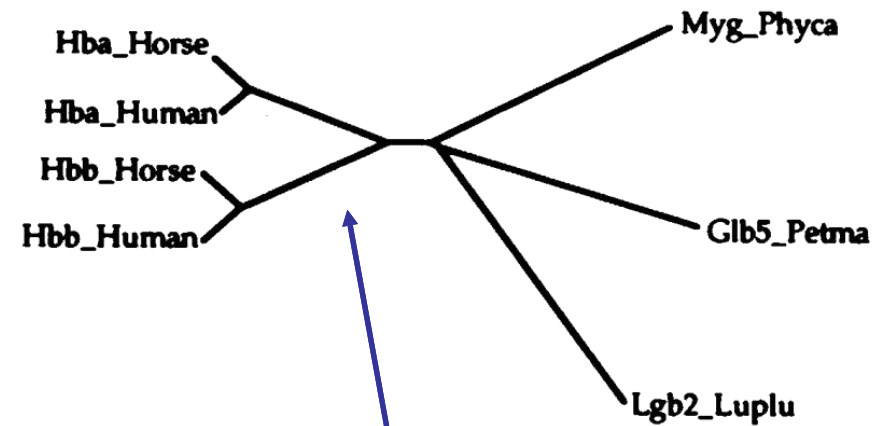
Pairwise alignment:
Calculate distance matrix

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
		1	2	3	4	5

1-SID

Step 2:

Unrooted Neighbor-Joining tree



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

Progressive method: ClustalW



$$0.221=0.081+0.226/2+0.061/4+0.015/5+0.062/6$$
$$0.221=0.081+0.226/2+0.061/4+0.015/5+0.062/6$$

Progressive method: ClustalW

Dynamic programming scoring function

1 peeksavtal
2 geekaavlaal
3 padktnvkaa
4 aadktnvkaa

5 egewqlvlhv
6 aaektkirsa



Without sequence Weights:


$$\begin{aligned} \text{Score} = & M(t, v) \\ & + M(t, i) \\ & + M(1, v) \\ & + M(1, i) \\ & + M(k, v) \\ & + M(k, i) \\ & + M(k, v) \\ & + M(k, i) / 8 \end{aligned}$$

With sequence Weights W_i :

$$\begin{aligned} \text{Score} = & M(t, v) * W_1 * W_5 \\ & + M(t, i) * W_1 * W_6 \\ & + M(1, v) * W_2 * W_5 \\ & + M(1, i) * W_2 * W_6 \\ & + M(k, v) * W_3 * W_5 \\ & + M(k, i) * W_3 * W_6 \\ & + M(k, v) * W_4 * W_5 \\ & + M(k, i) * W_4 * W_6 / 8 \end{aligned}$$

$M(i, j)$: 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. Phylogenetic 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

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



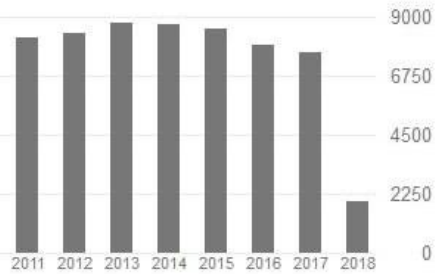
Stephen Frank Altschul (born February 28, 1957) is an American [mathematician](#) who has designed [algorithms](#) that are used in the field of [bioinformatics](#) (the Karlin-Altschul algorithm^[2] and its successors^[3]). Altschul is the co-author of the [BLAST](#) algorithm used for [sequence analysis](#) 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	66139	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



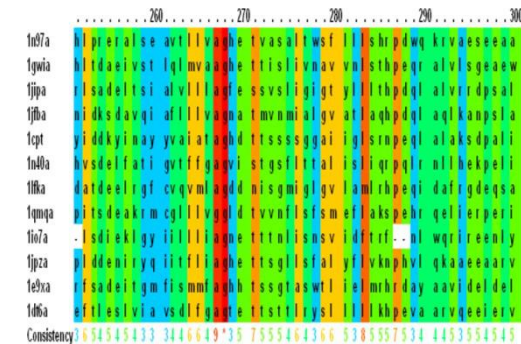
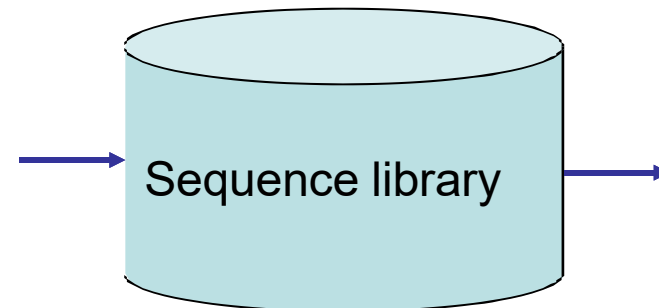
[合著作者](#) [查看全部](#)

- 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

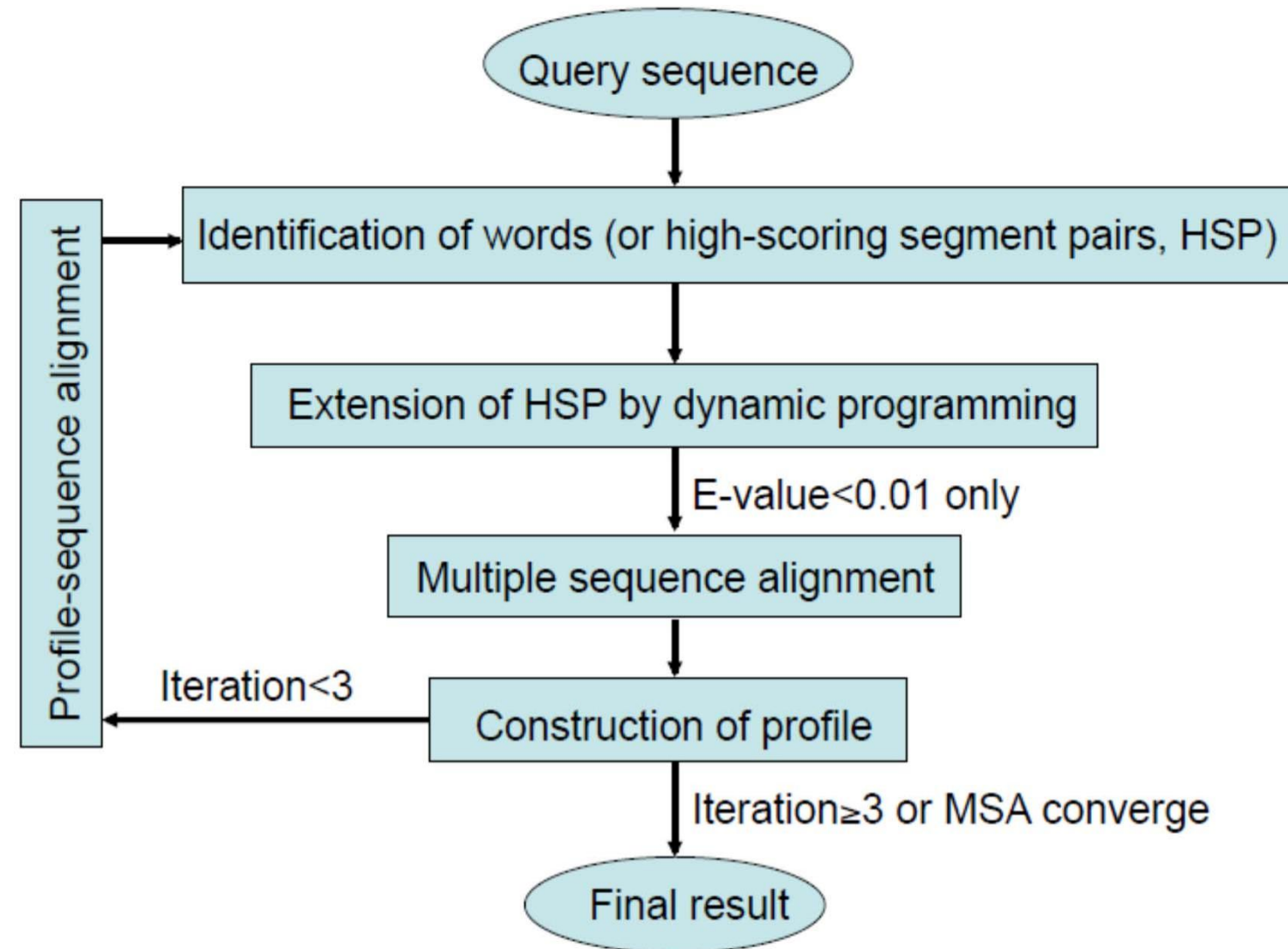
Problem to solve: how to identify a set of sequences from a library, which are all homologous to the query sequence of interest?

MVLSEGEWQLVLHVWAKVEADVA
GHGQDILIRLFKSHPETLEKFDRVK
HLKTEAEM



MSA of related sequences

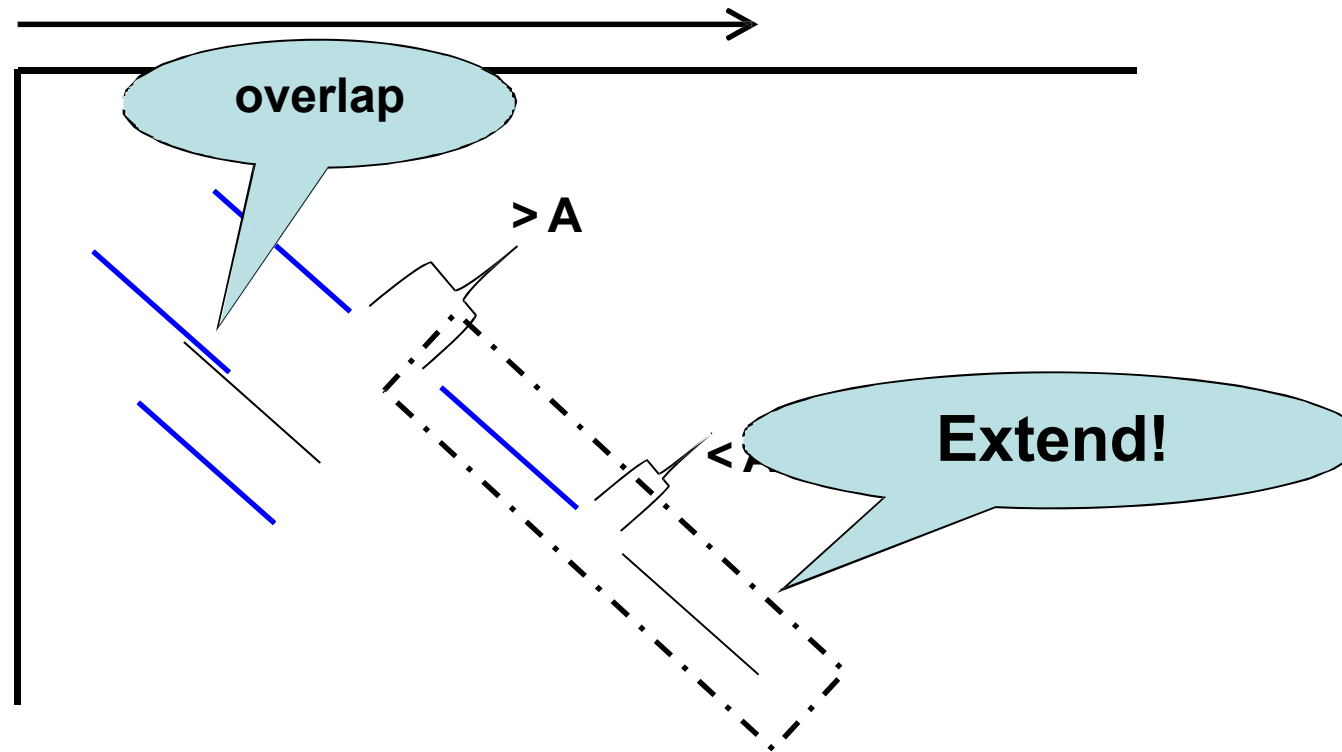
Flowchart of PSI-BLAST



Flowchart of PSI-BLAST

Difference between BLAST and PSI-BLAST:

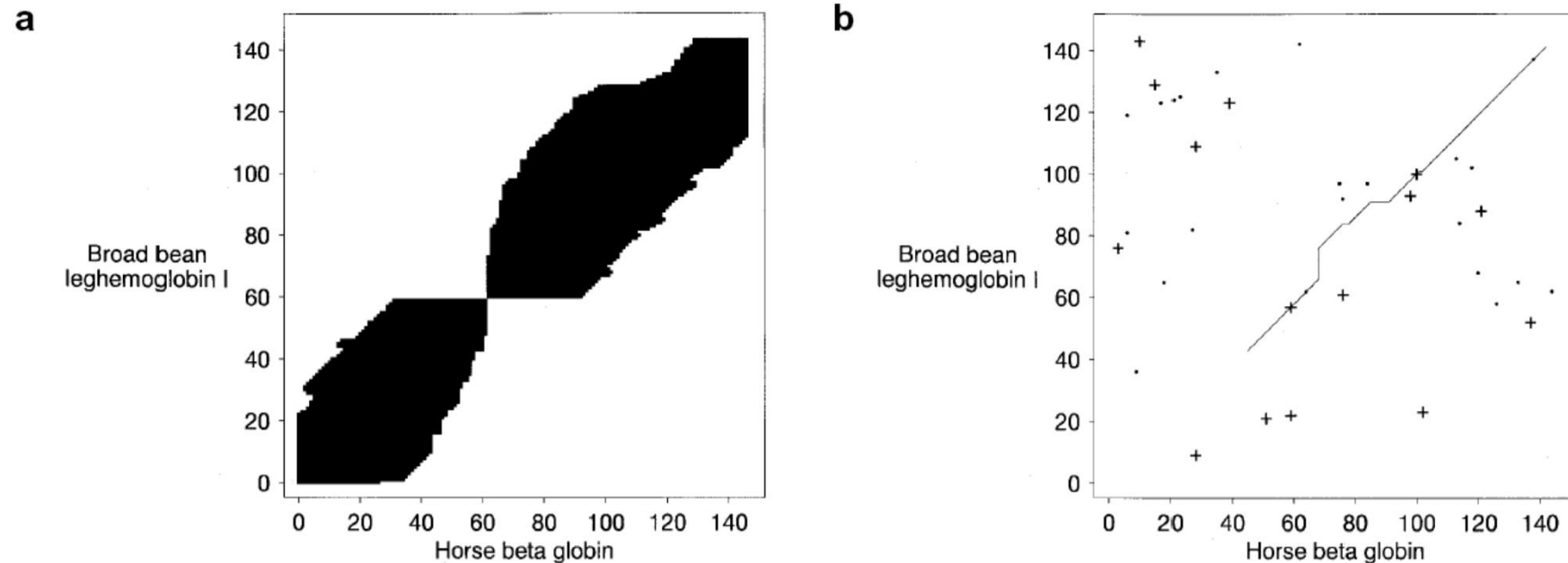
1. Starting with **two** diagonal HSPs in PSI-BLAST rather than **one** HSP in BLAST



Flowchart of PSI-BLAST

Difference between BLAST and PSI-BLAST:

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



蚕豆的血红蛋白

马beta
球蛋白

C

Leghemoglobin	43	FSFLKDSAGVVDSPKLGHAHEKVFGMVRDSAVQLRATGEVV--LDGKDGS-----	90
		F L + V+ +PK+ AH +KV L + GE V LD G+	
Beta globin	45	FGDLSNPGAVMGNPVKVKAHGKKV-----LHSFGEVGHLDNLKGTFAALSE	90

Leghemoglobin	91	IHIQKGVLDP-HFVVVKEALLKTIKEASGDKWSEELSAWEVAYDGLATAI	140
		+H K +DP +F ++ L+ + G ++ EL A+++ G+A A+	
Beta globin	91	LHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVAVAGVANAL	141

Flowchart of 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

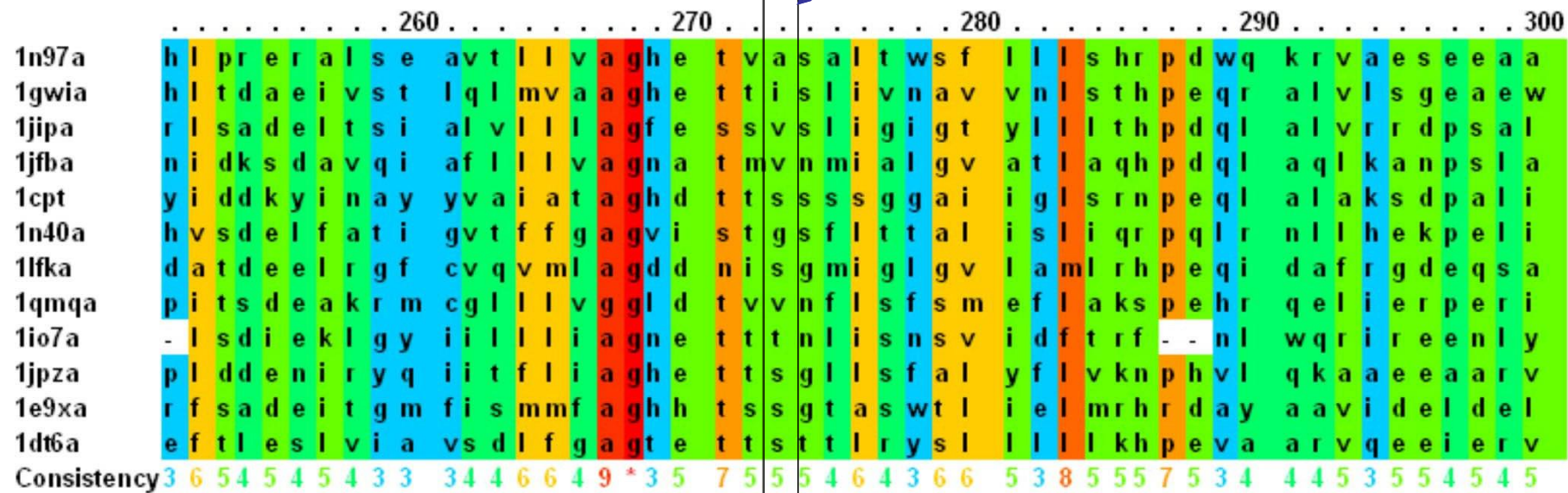
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-2	-2	-3	-4	-2	-1	-3	-3	-2	1	2	-2	8	0	-3	-2	-1	-2	-2	0
2	K	-2	4	-1	-2	-4	1	0	-2	-1	-3	-3	5	-2	-4	-2	-1	-1	-4	-2	-3
3	I	-2	-4	-4	-4	-2	-3	-4	-5	-4	6	1	-3	1	-1	-3	-3	-1	-3	-2	2
4	P	-1	-3	-3	-2	-4	-2	-2	-3	-3	-3	-4	-2	-3	-4	8	-1	-2	-4	-4	-3
5	K	-1	4	-1	-1	-4	1	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	3	-4	-2	-2	2	8	-2	-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 ($i=1,2,\dots,20$)

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVL

Position j ($j=1,2,\dots,L$)



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

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

P_i : background probability

如何估计 Q_{ij} ，一种思路直接从MSA中计算。但这样忽略了样本量小，和残基之间的先验知识最好的估计方法是Dirichlet mixtures 方法。但在实际使用中我们使用以下伪计数方法。

PSSM derivation

Position j ($j=1,2,\dots,L$) Residue A_i ($i=1,2,\dots,20$)

pseudocount

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

$\alpha = N_c - 1$ is the number of different residues
 $\beta = 10$

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

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

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

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

target frequency $q_{ij} = P_i P_j e^{\lambda S(i,j)}$,
其中s (i,j) 来自Log-odds of PAM250

$\alpha = N_c - 1$ is the he number of different residues
 $\beta = 10$

pseudocount frequency

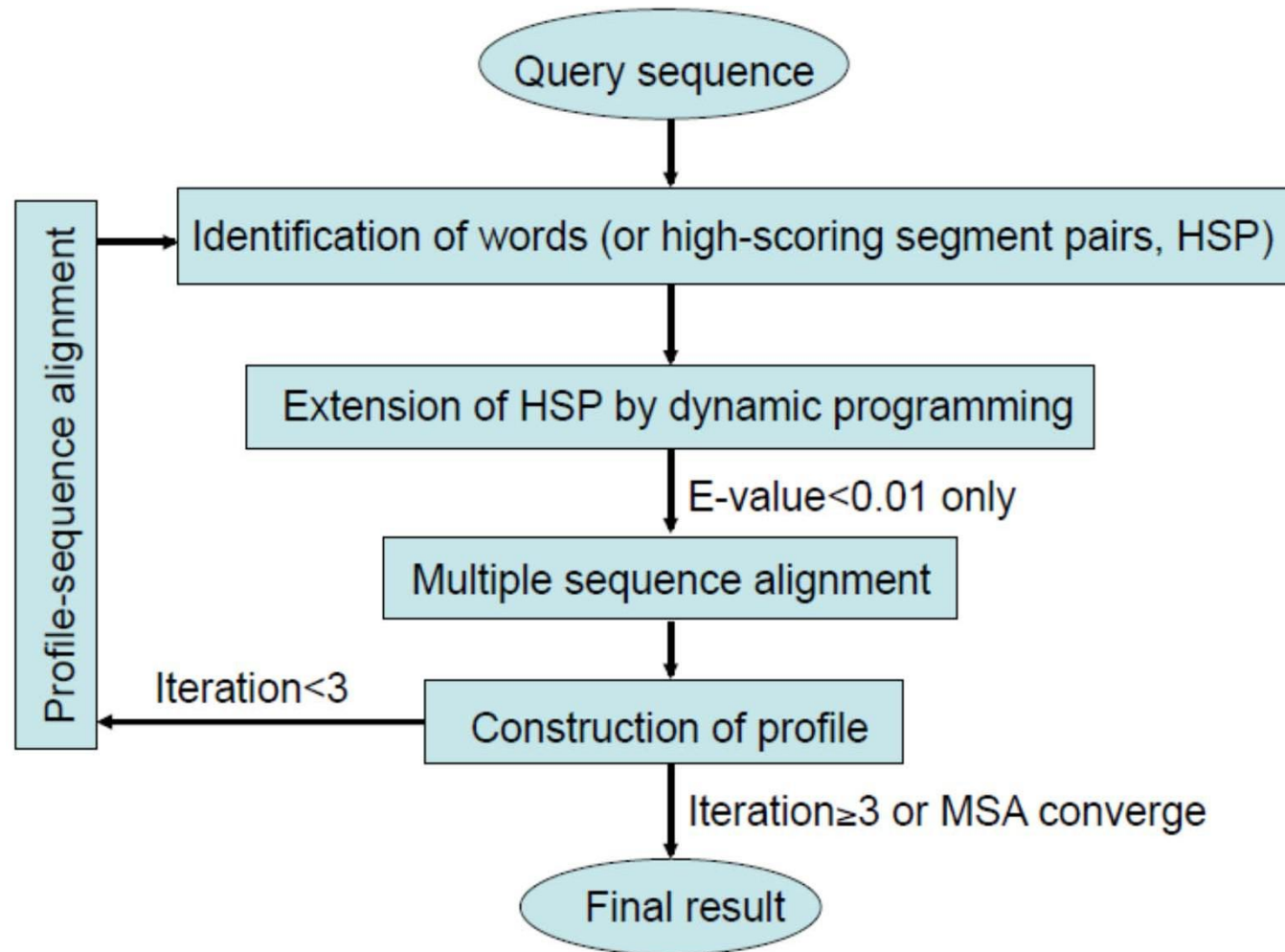
$$g_i = \sum_j \frac{f_j}{P_j} q_{ij}$$

其中 f_j 是观察到的频率

$$Q_i = \frac{\alpha f_j + \beta j}{\alpha + \beta}$$

PSI-BLAST iteration

5. Perform sequence-profile alignment (it's not described how this was done in the paper)



Content

1. Phylogenetic 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

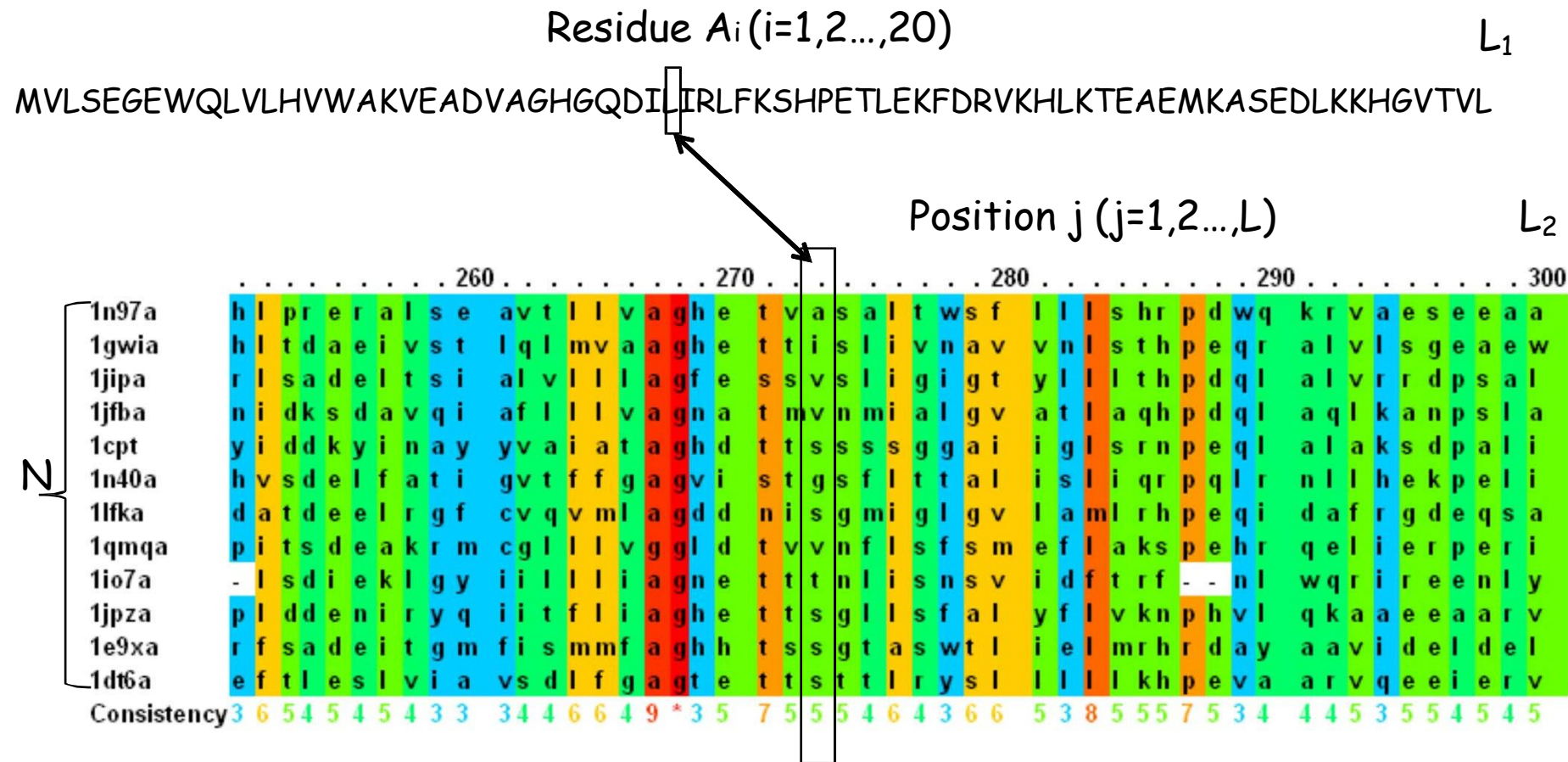


What is a sequence profile?

Sequence Profile:

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

What is a sequence profile?



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

What is a sequence profile?

20种氨基酸 A, R, N, D, C, ..., V

The alignment score for residue A_i and position j

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

Amino acid	3-letter abbreviation	1-letter symbol	Chemical characteristics
Alanine	ALA	A	Nonpolar, hydrophobic
Arginine	ARG	R	Polar, hydrophilic
Asparagine	ASN	N	Polar, hydrophilic
Aspartic acid	ASP	D	Polar, hydrophilic
Cysteine	CYS	C	Polar, hydrophilic
Glutamine	GLN	Q	Polar, hydrophilic
Glutamic acid	GLU	E	Polar, hydrophilic
Glycine	GLY	G	Polar, hydrophilic
Histidine	HIS	H	Polar, hydrophilic
Isoleucine	ILE	I	Nonpolar, hydrophobic
Leucine	LEU	L	Nonpolar, hydrophobic
Lysine	LYS	K	Polar, hydrophilic
Methionine	MET	M	Nonpolar, hydrophobic
Phenylalanine	PHE	F	Nonpolar, hydrophobic
Proline	PRO	P	Nonpolar, hydrophobic
Serine	SER	S	Polar, hydrophilic
Threonine	THR	T	Polar, hydrophilic
Tryptophan	TRP	W	Nonpolar, hydrophobic
Tyrosine	TYR	Y	Polar, hydrophilic
Valine	VAL	V	Nonpolar, hydrophobic

If we list $p(j, a)$ for all 20 possible amino acids at position j , we will get a $L_2 \times 20$ matrix. This matrix is called sequence profile of the N sequences

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	T	-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	F	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	0	3	0	3	6	4	9
28	E N 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

Sequence	Position					Weight	
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
GRDGF	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

每列之和为1

i=1...4 四条序列
j=1...5 五列

$$w_i = \sum_{j=1}^L w_{ij} = \sum_{j=1}^L \frac{1}{n_i * f_{ij}}$$

Number of amino acid
types at jth position

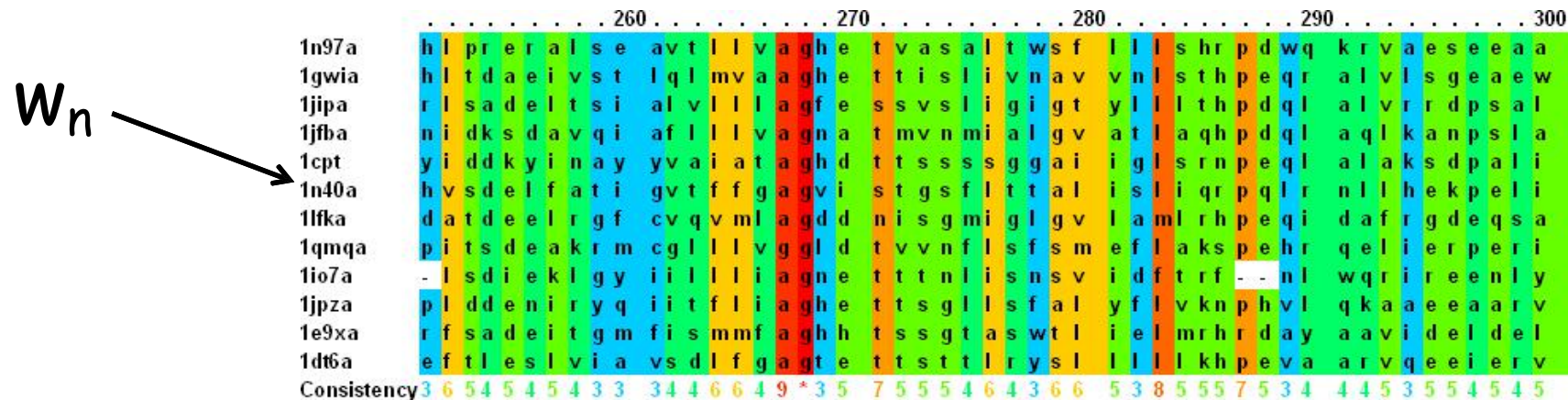
Number of occurrence for
A_i at jth position

在第j位置上,氨
基酸A_i出现的
次数

在第j位置上氨基酸类型的个数

Profile with sequence weight


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



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. Phylogenetic 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

Query sequence

Template profile

				A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-	
M I L G	A R R V G	H I G F	P F G L	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	-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	0	-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	

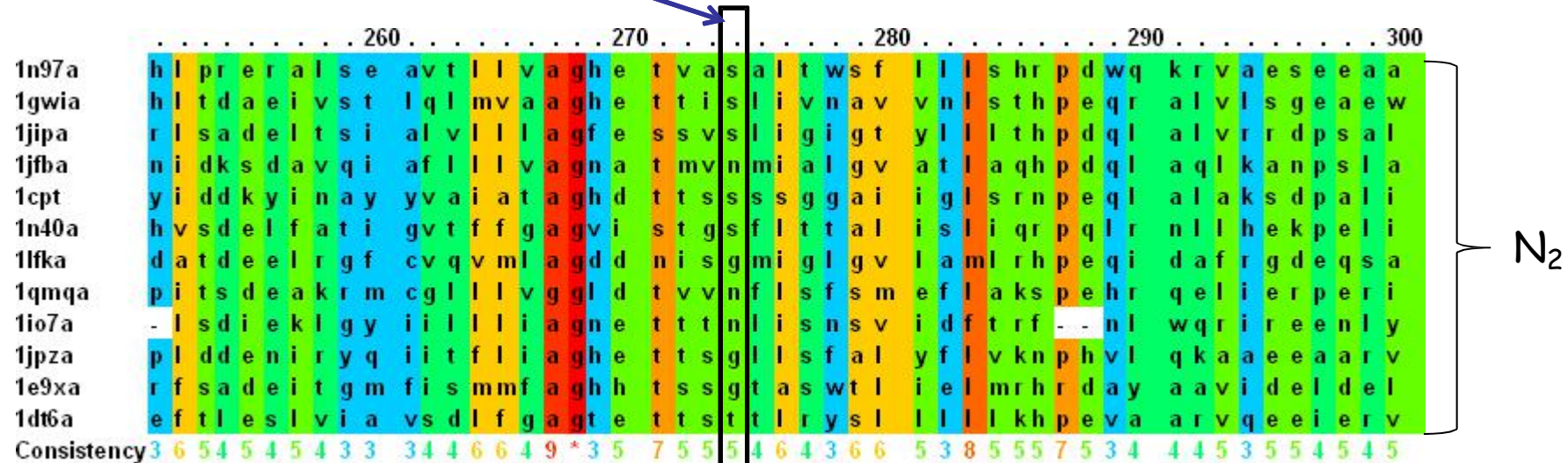
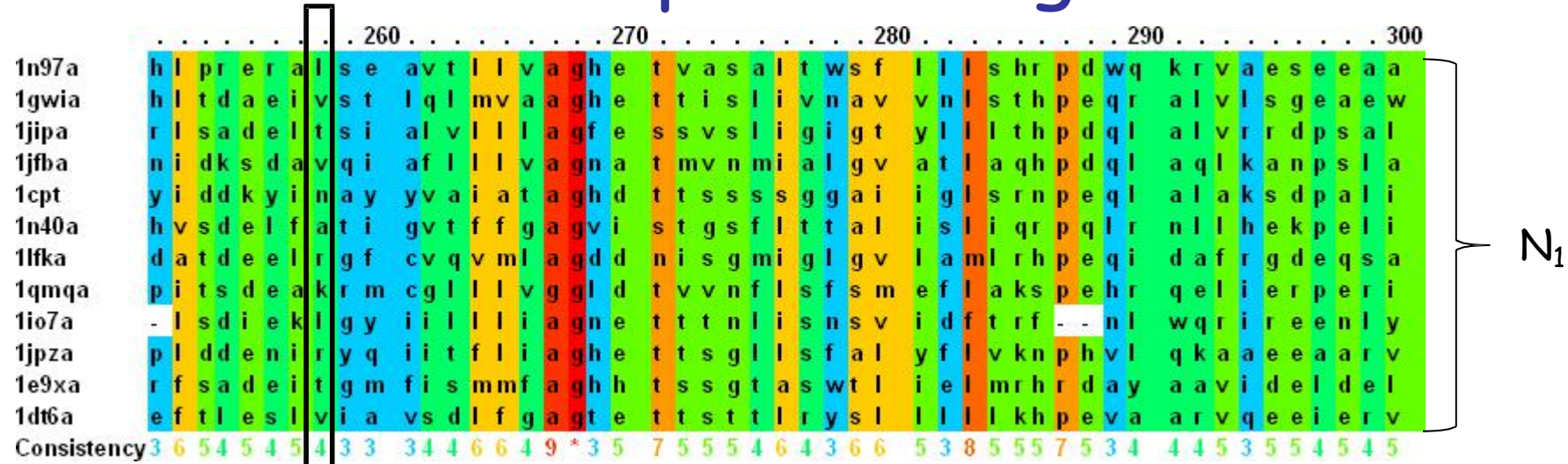
Algorithm: Dynamic programming

Content

1. Phylogenetic 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



Profile-profile alignment



Algorithm: Dynamic programming

Profile-profile alignment

References

•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

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

Frequency vector \longrightarrow $\overrightarrow{f_i} \cdot \overrightarrow{p_j}$ \longleftarrow Log-odds vector

Profile-profile alignment

Query sequence (MSA)

Template profile

10th →

MAHPPQIRIPATYLRGTSKGVFRLEDLPEDRLFMRVIGSPD	1	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
MVLSGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETL--	2	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
EMKASEDLKKHGVTVLT--ALGA-LKKKGHEAEKKGHEAEEL	3	2	2	-2	-2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9	
SRWWCN-DGRTPGSRNLCNIPCSALLSEAEKGEFELKG----	4	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	1	3	6	0	-6	-4	9	
TASVNCACKIVSDGNGMNAWAWNRNCKGTDVQAFIR--GCRL	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	-1	-1	-1	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

$$i = 9, j = 5; S(i, j) = S(9, 5) = 3 * (-2) + 1 * 0 + 1 * 2 = -4$$

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]

作业：

Project

1. Install PSI-BLAST and nr database in your PC and run PSI-BLAST for at least one protein sequence randomly selected from PDB (e.g., from the first project).

PSI-BLAST is available at:

<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/> nr database (~2G)

is available at the course website:

<http://yanglab.nankai.edu.cn/teaching/bioinformatics/nr.tar.bz2>

You should report the MSA and the PSSM matrix of your query sequence.

2. Deduce the formula of weighted profile-profile alignment, i.e.,

$$S(i, j) = \vec{f_i} \cdot \vec{p_j}$$

Project

下面是选做题，仅对编程感兴趣的同学

3. Write a program for profile-profile alignment (the inputs to your program are two MSAs)

Content

1. Bioinformatics databases
2. Sequence alignment and database searching
3. Phylogenetic tree and multiple sequence alignment
- ➡ 4. Protein structure alignment
5. Protein secondary structure prediction
6. Protein tertiary structure prediction
7. Protein function 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 (Wang Yixian)

- **TM-score**

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

Papers to read

DALI:

Holm and Sander. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993, 233: 123-28 (Yuan Han, Zhao Yixin)

CE:

Shindyalov and Bourne, Protein structure alignment by incremental combinatorial extension (CE) of optimal path. Prot Eng, 1998, 11 739-747 (Chen Yu, Guo Yunhang)

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). (Miao Zhen)