

#3

Multiple Alignment

•Topics:

- Sequence Alignment for Proteins

Similarity Matrix

- Multiple sequence alignment

Sum of Pairs (SP) score

Multiple alignment by DP

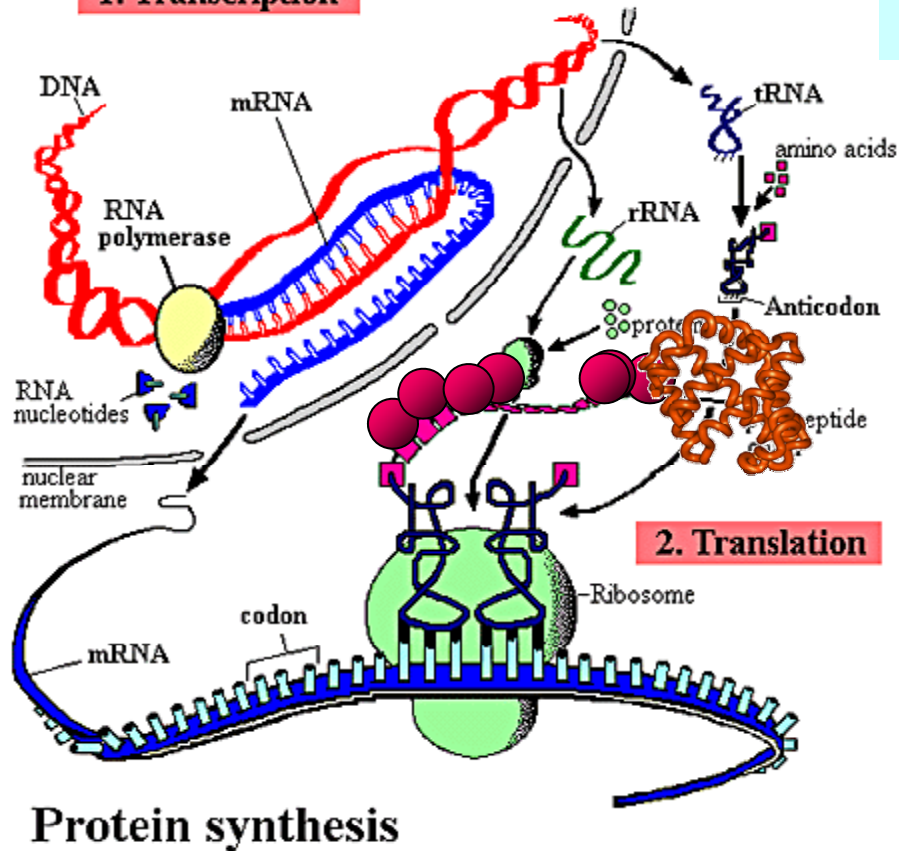
- Heuristic approaches for multiple alignment

Center star method, Progressive (Tree-based) method,

Iterative improvement method

DNA → RNA → Proteins

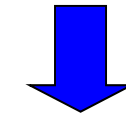
1. Transcription



2. Translation

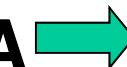
sequence of 4 nucleobases

DNA

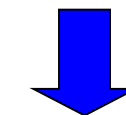


1. Transcription

RNA



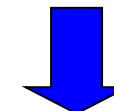
functional RNA



2. Translation

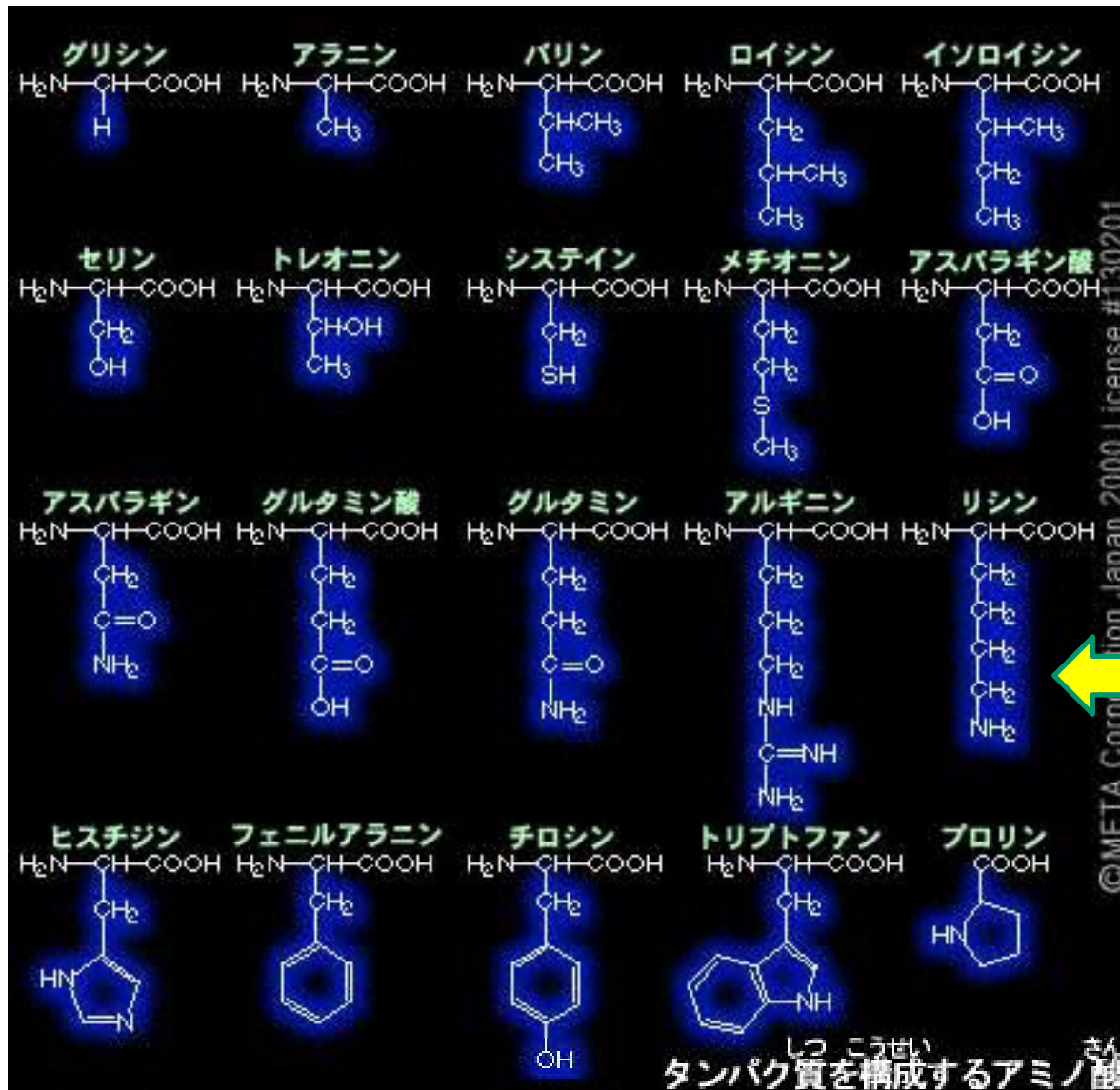
		2nd base in codon					
		U	C	A	G		
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G	3rd base in codon
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G	
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G	
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G	

“Codon table”



sequence of 20 amino acids

20 residues as elements of protein



Protein molecule is a linear chain composed of amino acid residues.
(details shown in Lesson 11)

Amino acid residues
(1 character code)

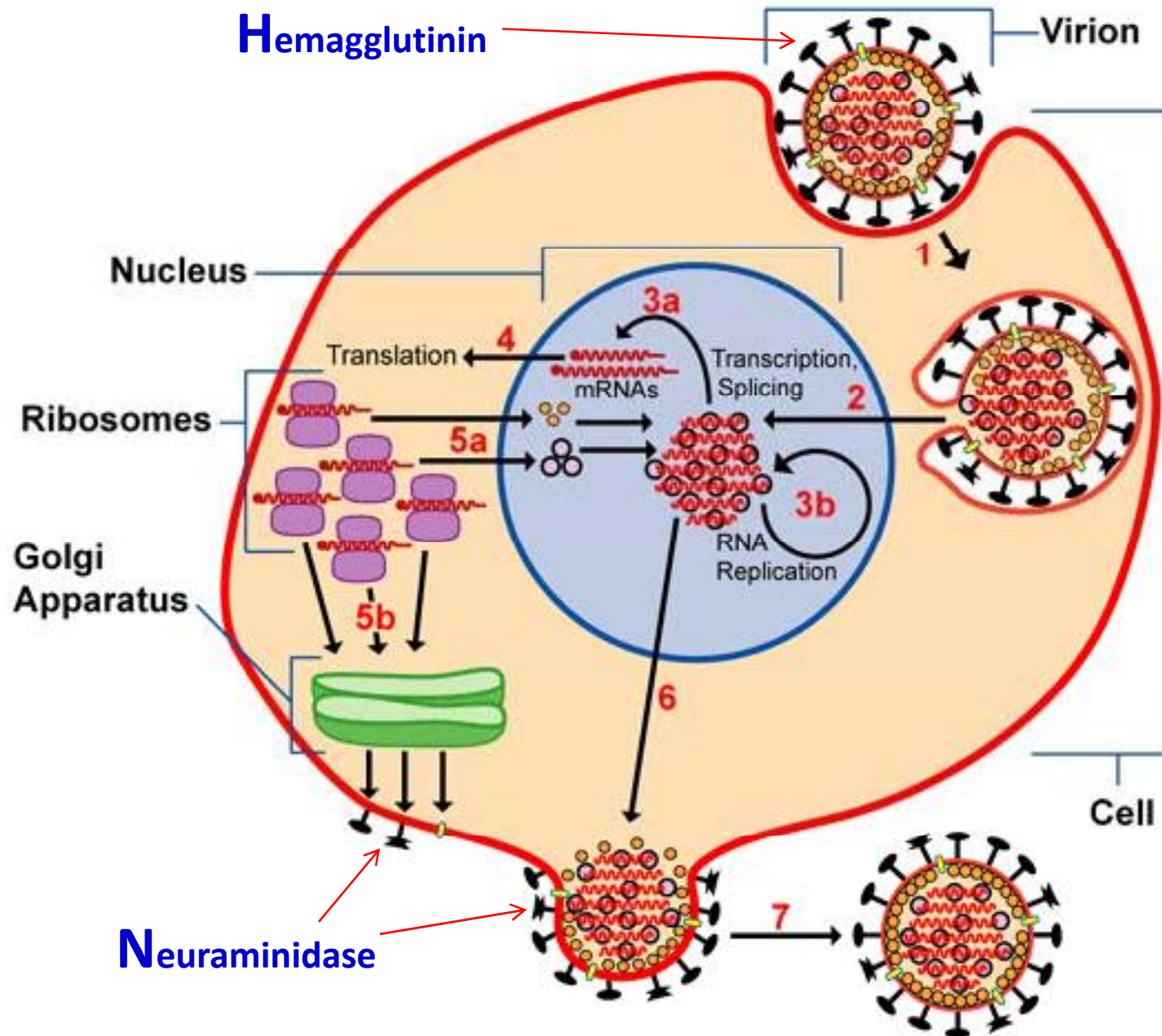
G, A, V, L, I,
S, T, C, M, D,
N, E, Q, R, K,
H, F, Y, W, P

Standard 20 residues.
(Exceptionally also
21st, 22nd residues)

Codon Table

		Second base				Third base
First base		U	C	A	G	
U	UUU Phenylalanine (Phe)	UCU Serine (Ser)	UAU Tyrosine (Tyr)	UGU Cysteine (Cys)	U	
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C	
	UUA Leucine (Leu)	UCA Ser	UAA STOP	UGA STOP	A	
	UUG Leu	UCG Ser	UAG STOP	UGG Tryptophan (Trp)	G	
C	CUU Leucine (Leu)	CCU Proline (Pro)	CAU Histidine (His)	CGU Arginine (Arg)	U	
	CUC Leu	CCC Pro	CAC His	CGC Arg	C	
	CUA Leu	CCA Pro	CAA Glutamine (Gln)	CGA Arg	A	
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G	
A	AUU Isoleucine (Ile)	ACU Threonine (Thr)	AAU Asparagine (Asn)	AGU Serine (Ser)	U	
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C	
	AUA Ile	ACA Thr	AAA Lysine (Lys)	AGA Arginine (Arg)	A	
	AUG Methionine (Met) or START	ACG Thr	AAG Lys	AGG Arg	G	
G	GUU Valine Val	GCU Alanine (Ala)	GAU Aspartic acid (Asp)	GGU Glycine (Gly)	U	
	GUC (Val)	GCC Ala	GAC Asp	GGC Gly	C	
	GUA Val	GCA Ala	GAA Glutamic acid (Glu)	GGA Gly	A	
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G	

Three consecutive bases (=codon) on RNA is translated to one amino acid residue.
Third base is relatively tolerant for mutation (synonymous substitution = 同義置換).



From NCBI "Entrez Genome" WWW page (Viral Genomes)

Swine = hog, pig

Avian = bird's

Flu = influenza

図1 高病原性鳥インフルエンザウイルスに感染したニワトリ

(a) 眼周囲に腫脹が認められる。

(b) 右は健康なニワトリ、左は感染したニワトリ。肉冠に壊死が認められる。

(c) 右は健康なニワトリ、左は感染したニワトリ。脚に皮下出血が認められる。



high-pathogenic avian influenza

low-pathogenic virus

high-pathogenic virus

trypsin-like protease



FURIN enzyme

From:
“Medical Bio”,
Ohmsha,
Jan., 2009,
“メディカル バイオ”
オーム社 7

1. Visit NCBI's **Influenza Virus Resource** site at <http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi>
2. Set **Species=Influenza virus A**, **Host=any**, **Country=any**, **Segment/Protein=NA (neuraminidase)**, Date Range year=2005-2010 for example.
3. Check Full-length .. only box, and Remove identical ... box.

NCBI **Influenza Virus Resource**
Information, Search and Analysis

HOME SEARCH SITE MAP Flu home Database Genome Set Alignment Tree BLAST Annotation FTP Help Contact us

Main Page>>Database

What are you looking for? Select one name each from the lists provided, and/or fill in the boxes. Multiple queries can be built by clicking the "Add to Query Builder" button every time a new query is made, and queries in any combination from the Query Builder can be selected to get sequences in the database. An advanced search tool is available [here](#).

show: ☒ Protein sequence ☐ Coding region ☐ Nucleotide sequence

Virus Species	Host	Country/Region	Segment/Protein	Date Range
any Influenzavirus A Influenzavirus B Influenzavirus C	any Avian Blow fly	any Africa Asia Europe	HA NP NA M1	year month day From: 2005 To: 2010

Subtype Min. length Max. length
H1N1 Search by a string [Help](#)

☒ Full-length sequences only [Help](#) ☒ Remove identical sequences [Help](#) ☐ Sequences from the FLU project only [Help](#) ☐ Include Lab strains [Help](#)

Add to Query Builder Clear selection Find sequence by Accession

Query Builder

Get sequences

4. Then press "Get sequences" button

1. Check how many sequences are matched in total. (353 in this example)
2. Set order as “host > year > country”, and then push “**Reorder sequences**” button.
3. Then click the first checkbox (shown below), in order to cancel default checks below.
4. Choose some sequences from the list, by clicking check box on each line.
Check 4 from Avian, 10 from Human, and 4 from Swine (in total 18 sequences)

NCBI **Influenza Virus Resource**
Information, Search and Analysis

HOME SEARCH SITE MAP Flu home Database Genome Set Alignment Tree BLAST Annotation FTP Help Contact us

Main Page>>Database
Select/de-select sequences from the list below. Click on an action button to proceed.

Show Query Builder

Ordered by the following fields

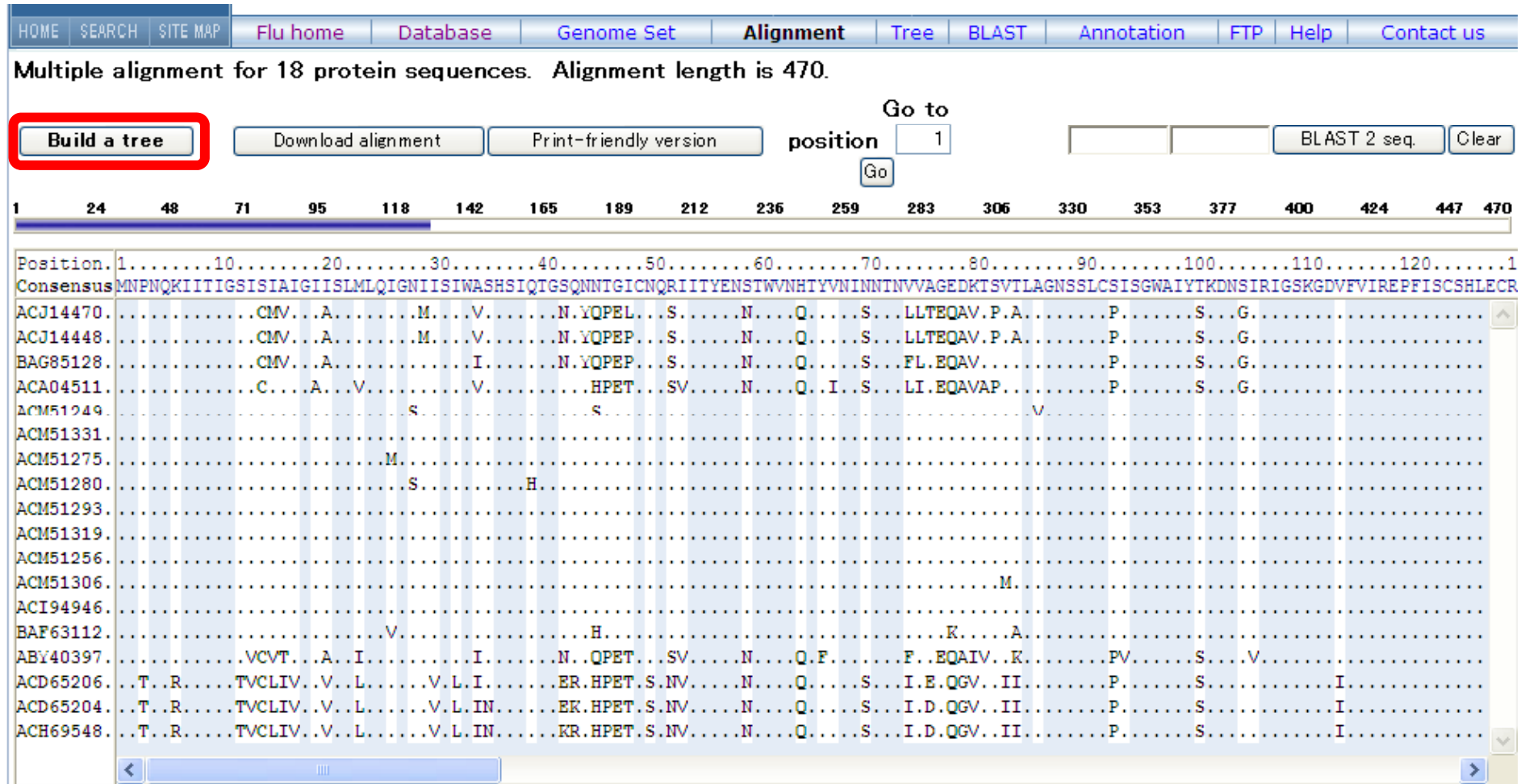
host country year Reorder sequences Add your own sequences

Do multiple alignment Build a tree - Select FASTA sequences to download - Select accession list to download -

<input checked="" type="checkbox"/>	accession	length	host	protein	subtype	country	year	Virus name	353 protein sequences	Age	Gender
<input checked="" type="checkbox"/>	ACJ14470	469	Avian	NA	H1N1	Italy	2006	Influenza A virus (A/duck/Italy/281904/2006(H1N1))			
<input checked="" type="checkbox"/>	ACJ14448	469	Avian	NA	H1N1	Italy	2007	Influenza A virus (A/duck/Italy/69238/2007(H1N1))			
<input checked="" type="checkbox"/>	BAG85128	469	Avian	NA	H1N1	Japan	2007	Influenza A virus (A/duck/Hokkaido/w73/2007(H1N1))			
<input checked="" type="checkbox"/>	ACA04511	469	Avian	NA	H1N1	USA	2005/03/01	Influenza A virus (A/muscovy duck/New York/21211-5/2005(H1N1))		Adult	
<input checked="" type="checkbox"/>	ABO52107	469	Environment	NA	H1N1	USA	2005/08/10	Influenza A virus (A/environment/Ohio/1007/2005(H1N1))			
<input checked="" type="checkbox"/>	ACB36679	470	Giant anteater	NA	H1N1	USA	2007	Influenza A virus (A/giant anteater/Tennessee/UTCVM07-733/2007(H1N1))			
<input checked="" type="checkbox"/>	ABJ16678	470	Human								

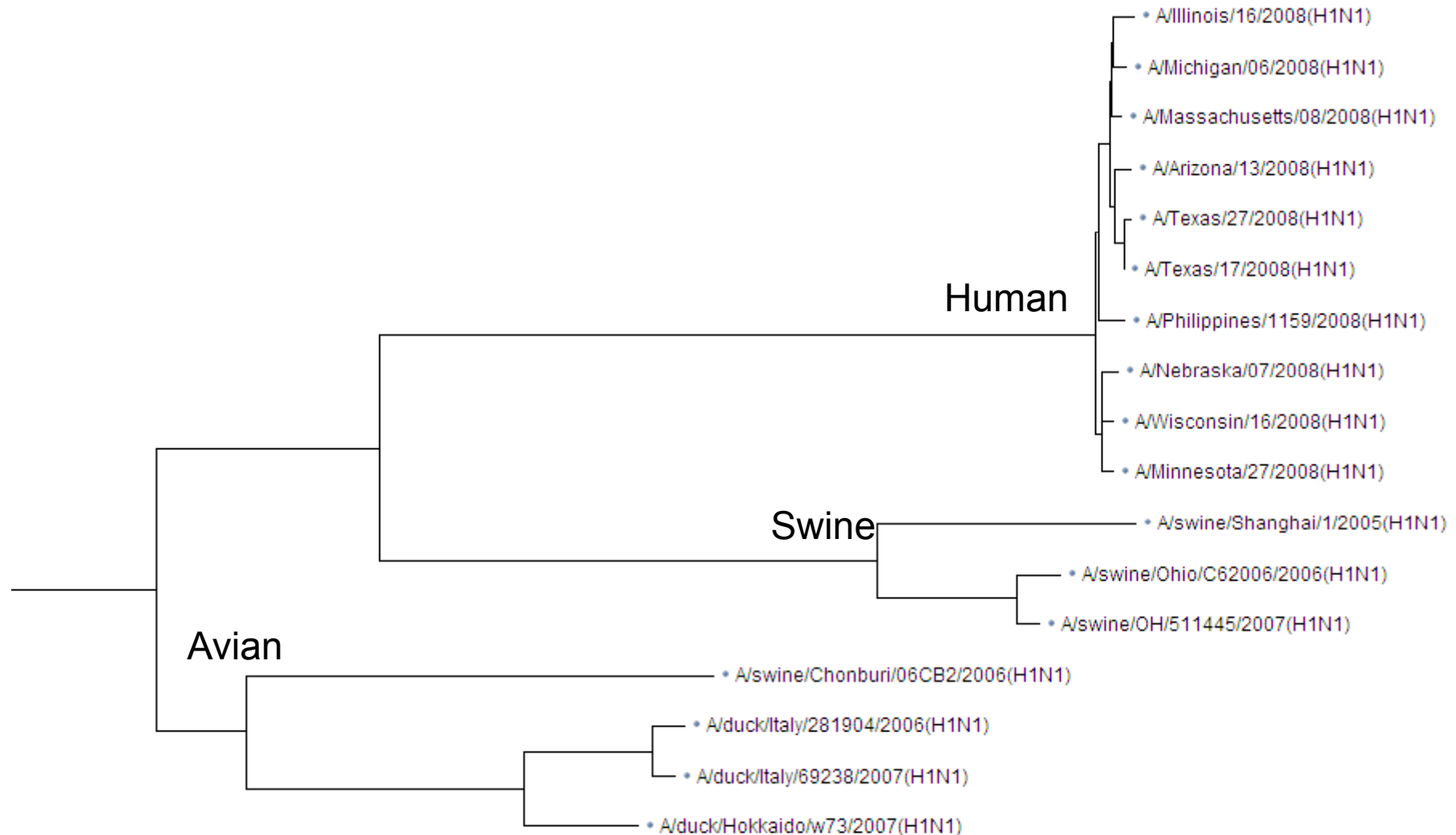
5. Then press “**Do multiple alignment**” button

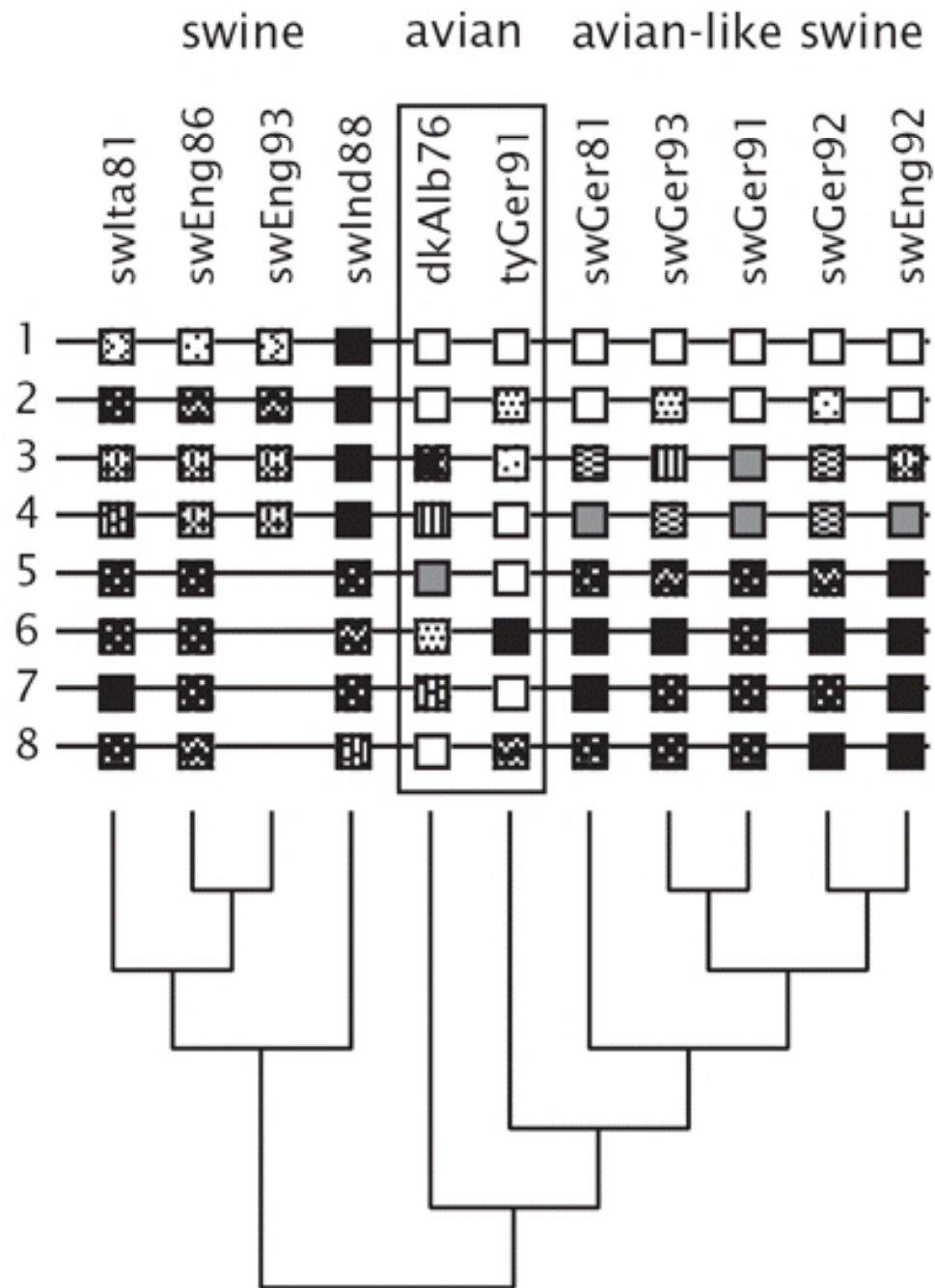
1. **Multiple alignment** for selected sequences are obtained in a few minutes.
2. Browse the whole sequences by using horizontal scroll bar.



First 4 sequences are from Avian, middle 10 are from Human, and last 4 from Swine. Avian neuraminidase and Swine neuraminidase are much different from Human's. Important catalytic site residues are still conserved (116,149,276,292,374,409,428).

1. Press “Build a tree” button, and then click “Next step >>”.
2. Select Clustering Algorithm=Neighbor Joint, and Distance=F84 matrix.
3. Click “Next step >>” and then you will get a calculated **Phylogenetic Tree**.

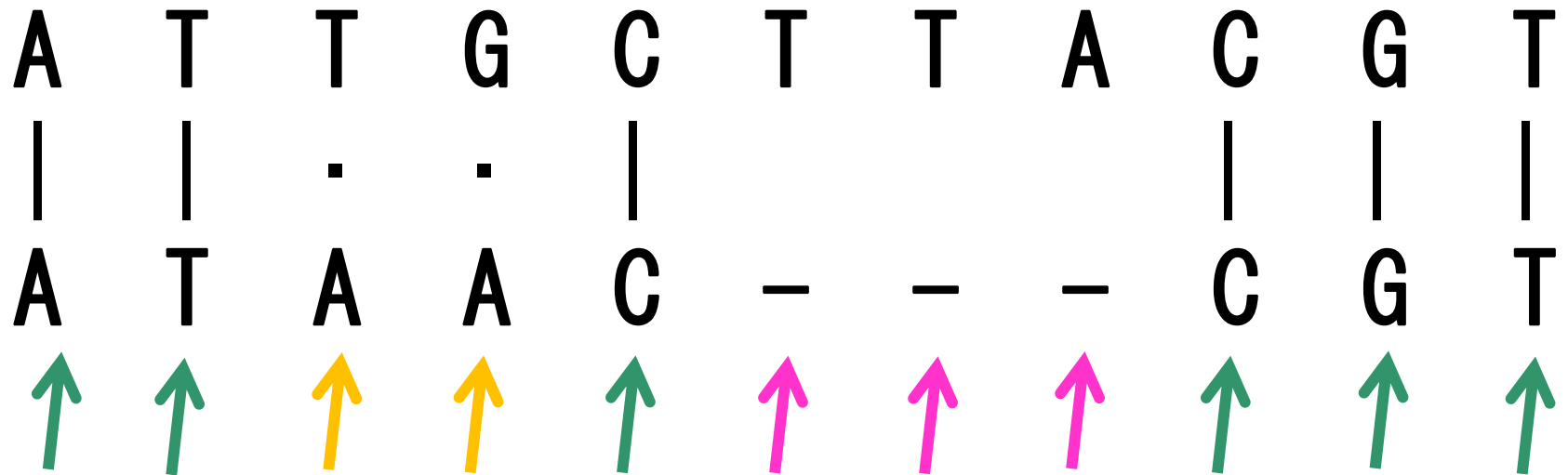




“avian-like swine”
influenza viruses
have been reported.

Alignment score

alignment



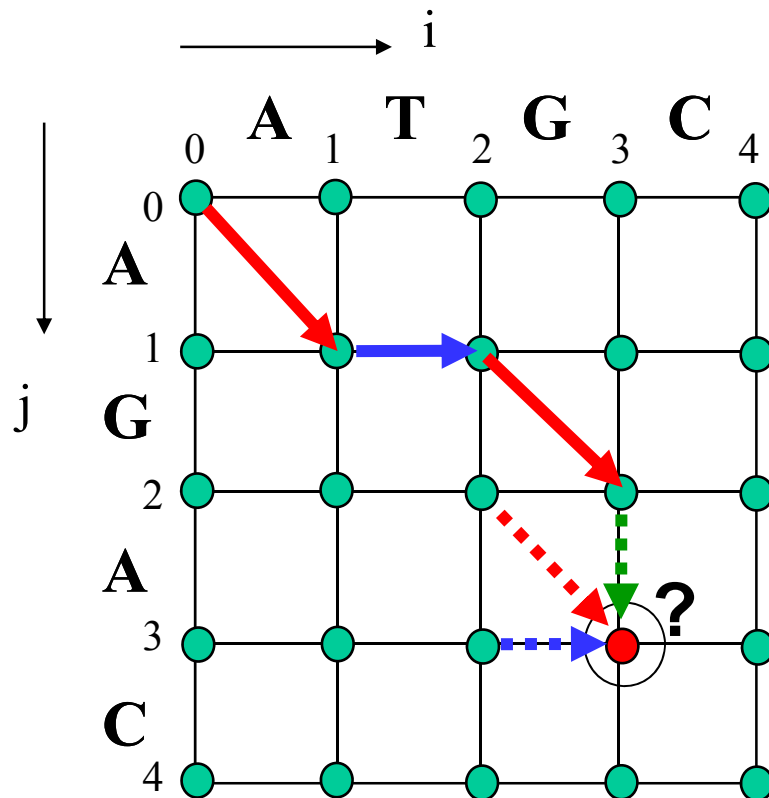
$$2 + 2 + (-1) + (-1) + 2 + (-1) + (-1) + (-1) + 2 + 2 + 2 = +7$$

**Example
score
for DNA**

match: +2
unmatch: -1
gap: -1

Scoring values are subject to change,
depending to the purpose of study,
and/or nature of subjects.

Sequence alignment by DP



for $i > 0, j > 0$,

$$M[i, j] \leftarrow \max \begin{cases} M[i-1, j] + w & \text{blue arrow} \\ M[i, j-1] + w & \text{green arrow} \\ M[i-1, j-1] + S[i, j] & \text{red arrow} \end{cases}$$

where $S[i, j]$ is the match/unmatch score,
 w is gap penalty constant (≤ 0), and
 $M[i, j]$ is **accumulated score** until $[i, j]$.

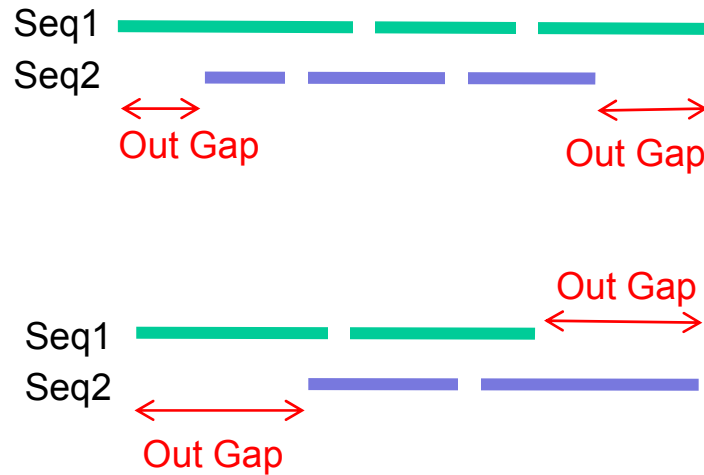
Boundary setting.

$$M[i, 0] \leftarrow 0, \quad M[0, j] \leftarrow 0$$

(note: here “outgap” has no penalty)

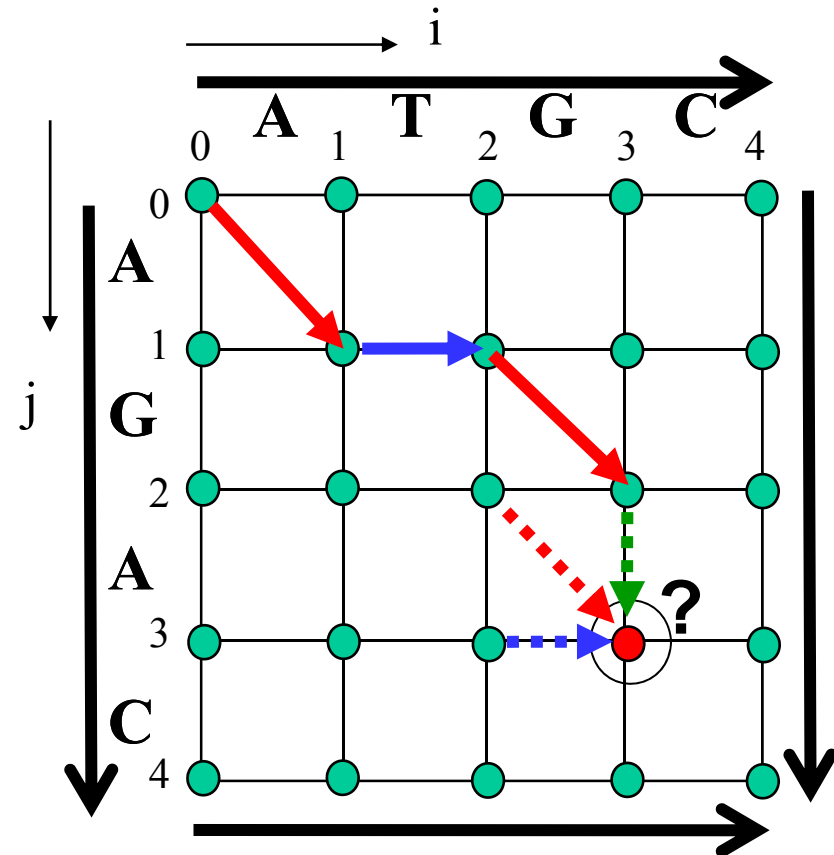
$w = 0$ for last column and last row. 14

Out Gap



Outer trim area is called as an “Out Gap”.

Out Gap penalty is sometimes set to zero.
It means that any movement on four edges (black arrow in the fig.) can be done without penalty.



Similarity Score Matrix for Protein

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM62

default matrix
for protein
sequence
comparison

Multiple Sequence Alignment

エントリ名	位置	1	10	20	30	40	50	60
tyrocidine-I_[ty	1	47	LCIGGVGLARGYWNRPDLTAEKF	VDNPFVPGE	KMYRTGDLAKWLT	DG		
gramicidin-S-I_	1	47	LCIGGEGLARGYWKRPELTSQKF	VDNPFVPGE	KLYKTGDQARWLP	DG		
gramicidin-S-II_	1	47	LYISGANVGRGYLNNQELTAEKF	FADPFRPNE	RMYRTGDLARWLP	DG		
ACV-domain-1_Pen	1	55	LHIGGLGISKGYLNRPELTPHRF	IPNPFQTDCEKQLGINS	LMYKTGDLARWLP	NG		
ACV-domain-1_Asp	1	55	LHIGGLGISKGYLNRPDLTQRF	IPNPFQTDHEKELGLNQL	MYKTGDLARWLP	NG		
ACV-domain-2_Pen	1	57	LYLGGEVVRGYHNRADVTAERF	IPNPFQSEEDKREGRNSRL	YKTGDLVRWIPGSSG			
ACV-domain-2_Asp	1	57	LYLGGEVARGYHNRPEVTAERF	LRNPFQTDSEGRNGRNSRL	YRTGDLVRWIPGSNG			
ACV-domain-3_Pen	1	55	LYLAGDSVTRGYLNQPLLTDQRF	IPNPFCKEEDIAMGRFARLY	KTGDLVRSRF	NR		
ACV-domain-3_Asp	1	55	LYLAGDCVARGYLNQPVLTGDRF	IQNPFQTEQDIACGSYPRL	YRTGDLFRCL	DR		
enterobactin_[en	1	41	LMTRGPYTFRGYKSPQHNASAF	DANGF	YCSGDLISIDP	EG		
angR-protein_[an	1	44	LWIGGDGIALGYFDDDELKTQAQFLHIDGHAW		YRTGDMGCYWP	DG		
luciferase_Photi	1	41	LCVRGPMIMSGYVNNPEATNALI	DKDGW	LHSGDIAYWDE	DE		
luciferase_Lucio	1	41	VCVKGPMMLKGYVNNPEATKELI	DEEGW	LHTGDIGYDE	EK		
luciferase-green	1	41	LCIKGPMVSKGYVNNVEATKEAI	DDDGW	LHSGDFGYDE	DE		
luciferase-y-gre	1	41	LCVKGPMVSKGYVNNVEATKEAI	DDDGW	LHSGDFGYDE	DE		
luciferase-yello	1	41	LCIKGPMVSKGYVNNVEATKEAI	DDDGW	LHSGDFGYDE	DE		
luciferase-orang	1	41	LCIKGPMVSKGYVNNVATKEAI	DDDGW	LHSGDFGYDE	DE		
antigen_[octapep	1	41	LLIKSDSMFSGYFLEKESTEHAFF	TNDGY	FKTGDIVQIND	NG		
acyl-CoA_rat_[lo	1	41	VCVKGANVFKGYLKDPARTAEAL	DKDGW	LHTGDIGKWLP	NG		
acyl-CoA_human_	1	41	VCVKGPNVFQGYLKDPAKTAEL	DKDGW	LHTGDIGKWLP	NG		
CoA-ligase_rice_	1	41	ICIRGQQIMKGYLNNPEATKNTI	DAEGW	LHTGDIGYVDD	DD		
4CL1-CoA-ligase_	1	41	ICIRGDQIMKGYLNDPESTRTTI	DEEGW	LHTGDIGFIDD	DD		
4CL2-CoA-ligase_	1	41	ICIRGDQIMKGYLNDPESTRTTI	DEEGW	LHTGDIGFIDD	DD		
St4CL-2a-CoA-lig	1	41	ICIRGDQIMKGYLNDPEATARTI	EKEGW	LHTGDIGFIDD	DD		
St4CL-1-CoA-liga	1	41	ICIRGDQIMKGYLNDPEATARTI	EKEGW	LHTGDIGFIDD	DD		
St4CL-2b-CoA-lig	1	41	ICIRGDQIMKGYLNDPEATARTI	EKEGW	LHTGDIGFIDD	DD		
acetyl-CoA_Neuro	1	26	RYMETYLH	VYKGY	YFTGDGAARDH	EG		
acetyl-CoA_Asp	1	26	RYMDTYLQ	VYKGY	YFTGDGAGRHD	EG		

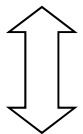
Multiple alignment example: 28 Luciferase proteins from fireflies (part)

Multiple Alignment Score

Sum of Pairs (SP) score

				position i			
Seq ₁	...	L	H	S	G	D	...
Seq ₂	...	L	H	T	G	D	...
Seq ₃	...	Y	K	T	G	D	...

$$\begin{aligned}
 \text{Score}(\text{Seq}_1(i), \text{Seq}_2(i), \dots \text{Seq}_N(i)) &= \sum_{k=1}^N \sum_{l=1}^{k-1} S(\text{Seq}_k(i), \text{Seq}_l(i)) \\
 &= S('T', 'S') + S('T', 'S') + S('T', 'T')
 \end{aligned}$$



cf. **Minimum entropy score**
(Theoretically more favorable)

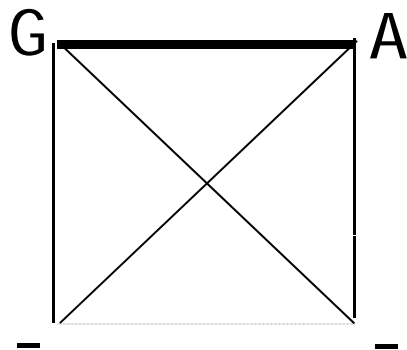
$S(X, Y)$ is similarity score,
 $S(X, '-') = w$ ($w \leq 0$, gap penalty)
 and $S('-', '-') = 0$.

Multiple Alignment Score

for DNA sequences

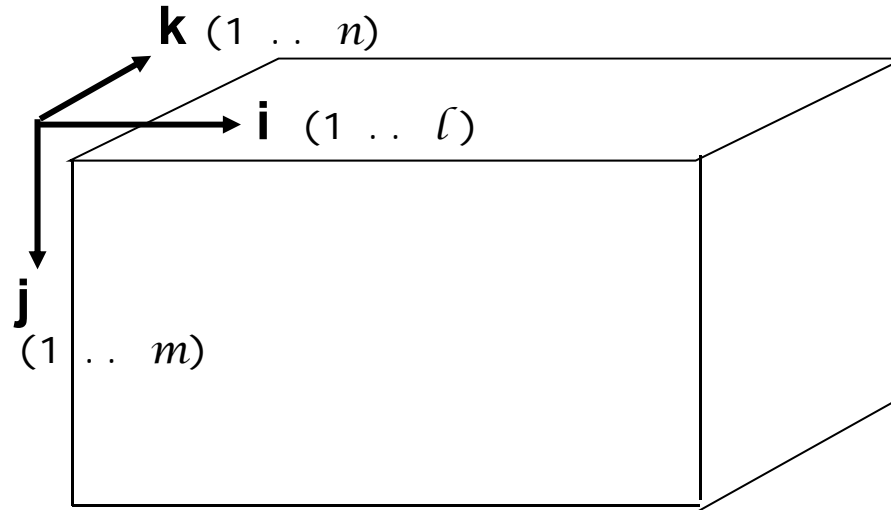
				position i			
Seq ₁	...	A	T	G	C	C	...
Seq ₂	...	A	T	A	C	C	...
Seq ₃	...	A	T	-	C	C	...
Seq ₄	...	A	T	-	C	C	...

$$\text{Score}(\text{Seq}_1(i), \text{Seq}_2(i), \dots, \text{Seq}_N(i)) = \sum_{k=1}^N \sum_{l=1}^{k-1} S(\text{Seq}_k(i), \text{Seq}_l(i))$$



$$\begin{aligned}
 &= S('A', 'G') + \\
 &\quad S('-', 'G') + S('-', 'A') + \\
 &\quad S('-', 'G') + S('-', 'A') + S('-', '-') \\
 &= (-1 \text{ unmatch}) + 4 \times (-1 \text{ gap}) + 0 \\
 &= -5
 \end{aligned}$$

Multiple Alignment by DP



3-dimensional Multiple Alignment

Time complexity:

$$O(l \ m \ n)$$

Space complexity:

$$O(l \ m \ n)$$

$$M[i, j, k] \leftarrow \max \begin{cases} M[i-1, j-1, k-1] + \text{Score}(\text{Seq1}(i), \text{Seq2}(j), \text{Seq3}(k)) \\ M[i, j-1, k-1] + \text{Score}('-', \text{Seq2}(j), \text{Seq3}(k)) \\ M[i-1, j, k-1] + \text{Score}(\text{Seq1}(i), '-', \text{Seq3}(k)) \\ M[i-1, j-1, k] + \text{Score}(\text{Seq1}(i), \text{Seq2}(j), '-') \\ M[i, j, k-1] + \text{Score}('-', '-', \text{Seq3}(k)) \\ M[i-1, j, k] + \text{Score}(\text{Seq1}(i), '-', '-') \\ M[i, j-1, k] + \text{Score}('-', \text{Seq2}(j), '-') \end{cases}$$

Note: **N- dimentional** direct DP with length L will consume $O(L^N)$ time and space.

Heuristic approaches for multiple alignment

(1) Star method

Seq₁ A G G A
Seq₂ A T G C G T
Seq₃ A T G C G A

similarity scores

	Seq 1	Seq 2	Seq 3	Sum
Seq 1		3	6	9
Seq 2	3		9	12
Seq 3	6	9		15 ★

Seq₁ A - G - G A
Seq₂ A T G C G T

$$\text{Score} = (+2) \times 3 + (-1) \times 3 = 3$$

Seq₁ A - G - G A
Seq₃ A T G C G A

$$\text{Score} = (+2) \times 4 + (-1) \times 2 = 6$$

Seq₂ A T G C G T
Seq₃ A T G C G A

$$\text{Score} = (+2) \times 5 + (-1) \times 1 = 9$$

✗
not used

(1) Choose one sequence Seq_C that maximizes
 $\sum_{i \neq c} \text{sim}(\text{Seq}_i, \text{Seq}_C)$

(2) Prepare pairwise alignments (Seq_i, Seq_C) for $i \neq c$.

(3) Combine those alignments based on Seq_C.

A - G - G A
A T G C G T
A T G C G A

Heuristic approaches for multiple alignment

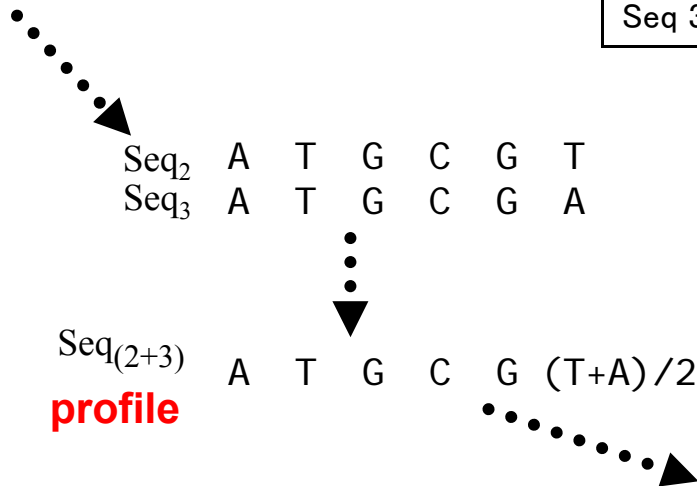
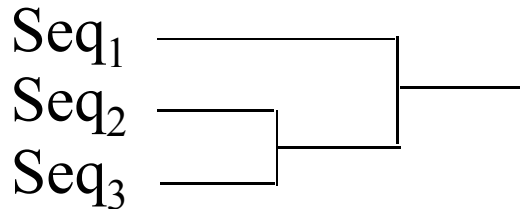
(2) Progressive (Tree-based) method

Seq₁ A G G A
Seq₂ A T G C G T
Seq₃ A T G C G A

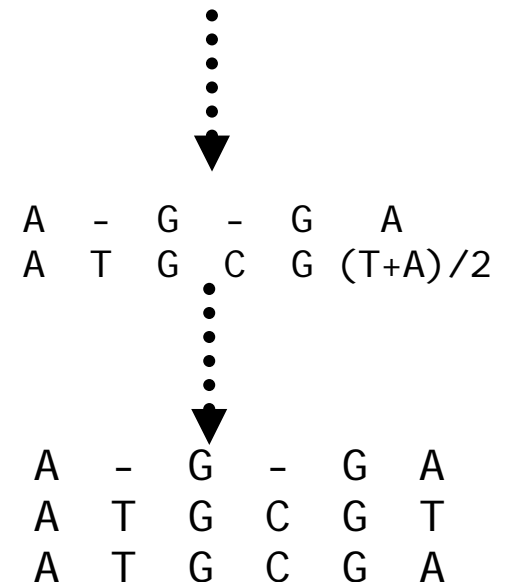
similarity scores

	Seq 1	Seq 2	Seq 3
Seq 1		3	6
Seq 2	3		9
Seq 3	6	9	

“Guide Tree”



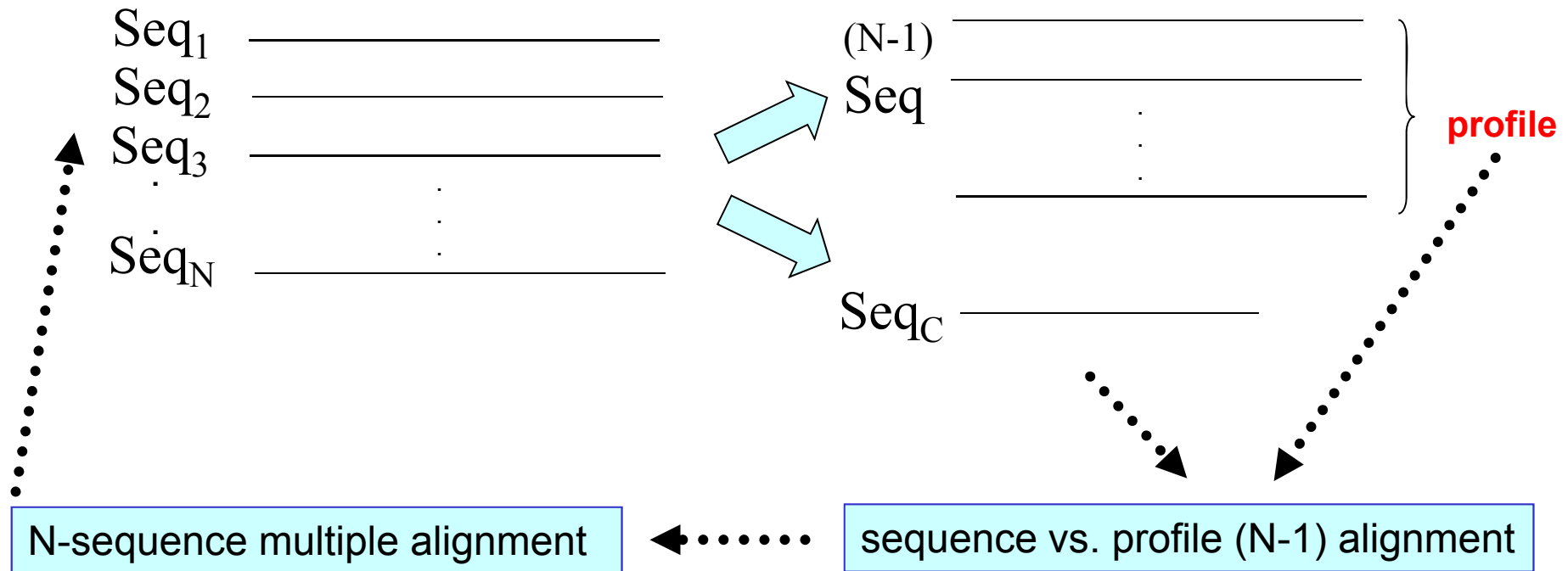
Seq₁ A G G A



- (1) Make a “Guide Tree” based on mutual similarity scores.
- (2) Start from pairwise alignment with most inner pair.
- (3) Following the “Guide Tree”, add sequences step by step.
Perform **sequence-profile (or profile-profile) alignment**.
- (4) go to (3) unless all sequence has been aligned.

Heuristic approaches for multiple alignment

(3) Iterative improving method



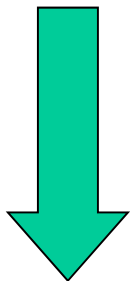
- (1) Make an initial multiple alignment.
- (2) Choose one sequence Seq_C (randomly)
- (3) Perform **sequence-profile alignment** between Seq_C and the profile made from rest of sequences.
- (4) go to (2) unless no progress obtained or iteration count reached to the limitation.

By random selection, it is expected to escape from local minimum.

finding conserved motifs by multiple sequence alignment

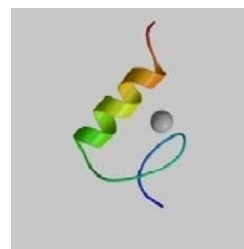
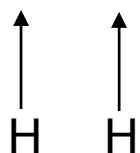
```
>copia    ILDFHEKLLHPGIQKTTKLFGETYYFPNSQLLIQNIINECSICNLAK
>MMULV    LLDFLLHQLTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVN
>HTLV     LQLSPAELHSFTHCGQTALTQGATTTEASNILRSCHACRGGN
>RSV      YPLREAKDLHTALHIGPRALSKACNISMQQAREVVQTCPHCNSA
>MMTV     IHEATQAHTLHHLNAHTLRLLYKITREQARDIVKACKQCVVAT
>SMRV     LESAQESHALHHQNAAALRFQFHITREQAREIVKLCPCPCPDWGS
```

http://www.icot.or.jp/ARCHIVE/Museum/SOFTWARE/GIP/gene_alignment.html



Multiple Alignment (Iterative Improvement method)
on PAPIA service (<http://mbs.cbrc.jp/papia/>)

エントリ名	位置	1	10	20	30	40	50	60
copia	1 47	----	ILDFHEKLLHPGIQKTTKLFGET	---	YYFPNSQLLIQNIINECSICNLAK	---		
MMULV	1 51	---	LLDFLLHQLTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVN	---				
HTLV	1 43	-	LQLSPAELHSFTHCGQTALTQGATT	-----	TEASNILRSCHACRGGN	---		
RSV	1 44	Y	PLREAKDLHTALHIGPRALSKACNIS	-----	MQQAREVVQTCPHCNSA	---		
MMTV	1 43	--	IHEATQAHTLHHLNAHTLRLLYKIT	-----	REQARDIVKACKQCVVAT	---		
SMRV	1 44	--	LESAQESHALHHQNAAALRFQFHIT	-----	REQAREIVKLCPCPCPDWGS	---		



zinc finger motif