Multiple protein sequence alignment with ClustalW

Multiple sites: pir.georgetown.edu
www.genome.ad.jp
Personal computer: download ClustalW

BLOSUM 62 Scoring Matrix

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

Example of how to score an alignment: Write down two sequences:

sequence#1	V	D	S	–	C	Y
sequence#2	V	E	S	L	C	Y
Score from sub. Matrix	4	2	4	-11	9	7

Score = Σ(AA pair scores) – gap penalty = 15

Selecting the right protein-scoring matrix

http://www.ncbi.nlm.nih.gov/pubmed/18792939

- Different scoring matrices are only optimal for detecting different classes of alignments.
- One needs to consider the similarity and amino acid composition of compared sequences.

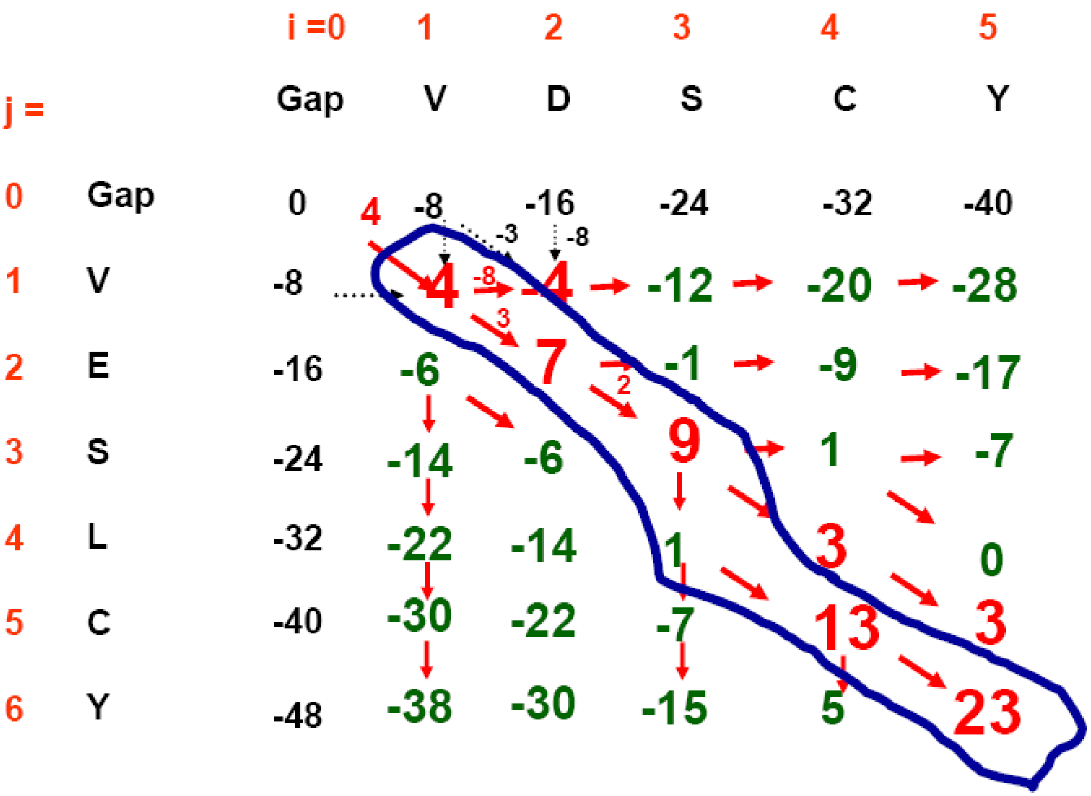
For general alignment purpose, BLOSUM62 or PAM120 are recommended.

For PAMx matrices, higher x detects more divergent sequences;
For BLOSUMn matrices, lower n detects more divergent sequences.

BLOSUM 80 | BLOSUM62 | BLOSUM45
PAM30 | PAM120 | PAM250

Less divergent ↔ more divergent

Gap penalty



j =		i =0	1	2	3	4	5
		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4	-4	-12	-20	-28
2	E	-16	-6	7	-1	-9	-17
3	S	-24	-14	-6	9	1	-7
4	L	-32	-22	-14	1	3	0
5	C	-40	-30	-22	-7	13	3
6	Y	-48	-38	-30	-15	5	23

http://en.wikipedia.org/wiki/Gap_penalty

Using nucleotide sequence databases

Gene-centric databases

Genome-centric resources
- Prokaryotic
- Eukaryotic

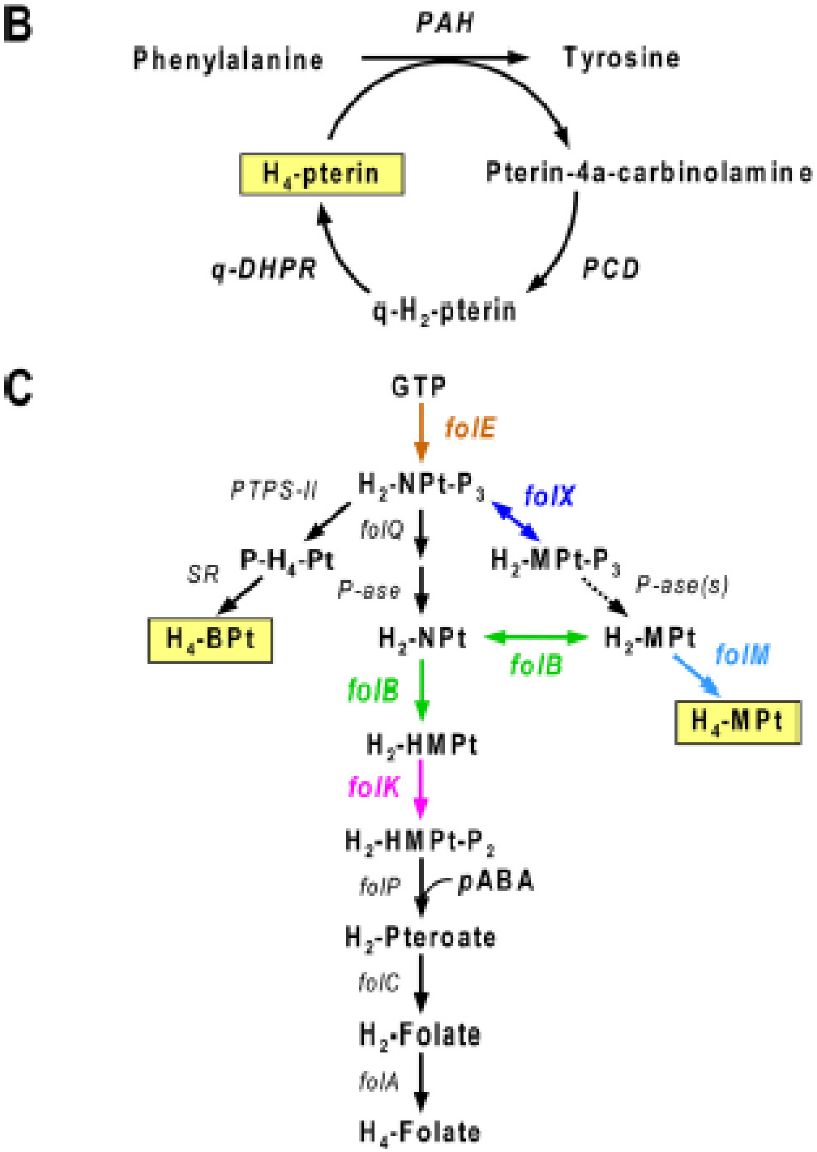
NCBI: www.ncbi.nlm.nih.gov

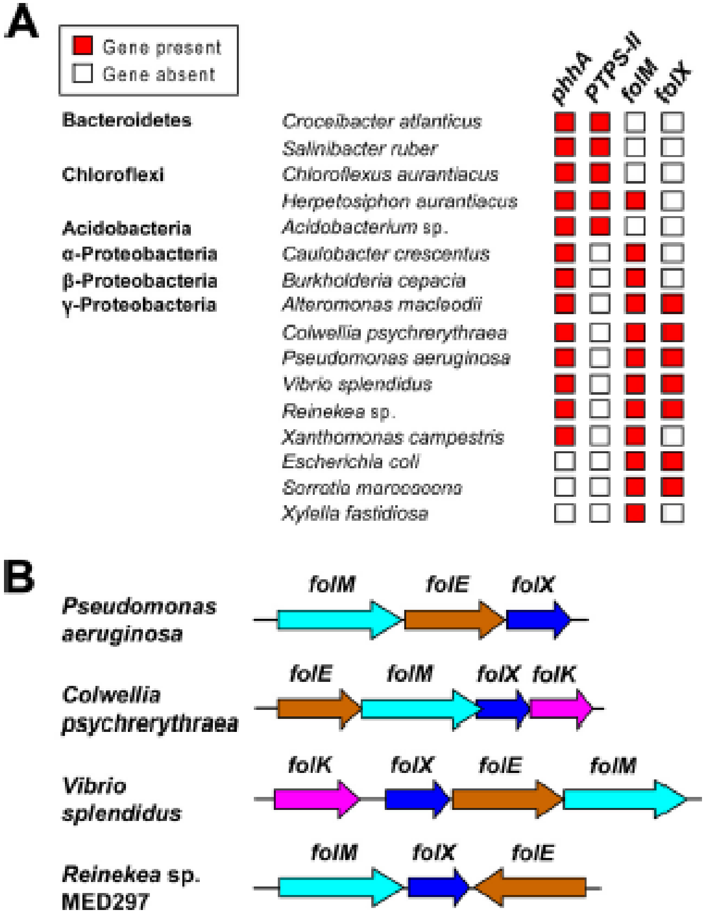
Ensembl: www.ensembl.org

TIGR: www.tigr.org
    http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi

Microbial genome: http://mbgd.genome.ad.jp/

FolX and FolM Are Essential for Tetrahydromonapterin Synthesis in *Escherichia coli* and *Pseudomonas aeruginosa*[▽][†]

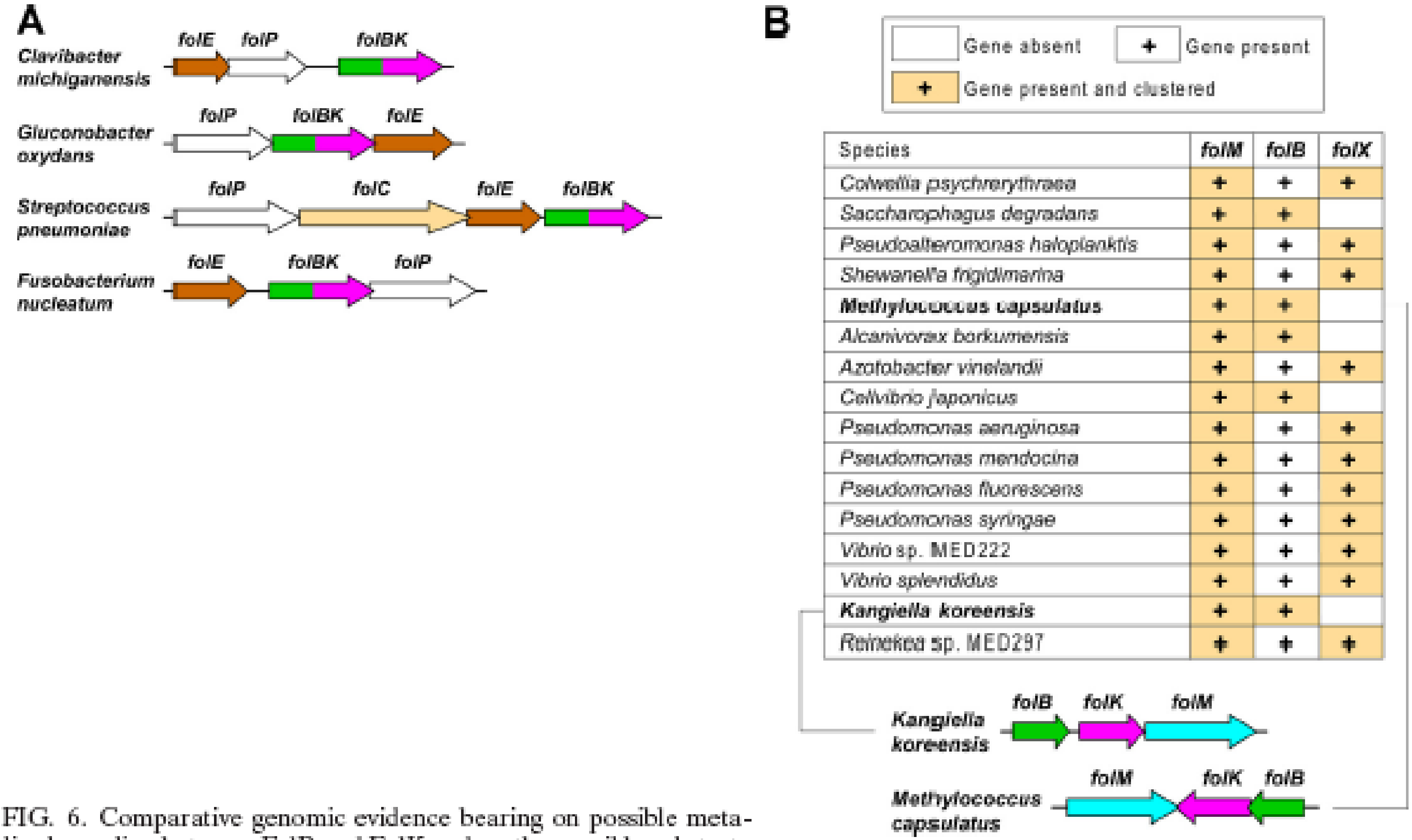




🟥 Gene present
☐ Gene absent

		phhA	PTPS-II	folM	folX
Bacteroidetes	*Croceibacter atlanticus*	🟥	🟥	☐	☐
	Salinibacter ruber	🟥	🟥	☐	☐
Chloroflexi	*Chloroflexus aurantiacus*	🟥	🟥	☐	☐
	Herpetosiphon aurantiacus	🟥	🟥	🟥	☐
Acidobacteria	*Acidobacterium* sp.	🟥	🟥	☐	☐
α-Proteobacteria	*Caulobacter crescentus*	🟥	☐	🟥	☐
β-Proteobacteria	*Burkholderia cepacia*	🟥	☐	🟥	☐
γ-Proteobacteria	*Alteromonas macleodii*	🟥	☐	🟥	🟥
	Colwellia psychrerythraea	🟥	☐	🟥	🟥
	Pseudomonas aeruginosa	🟥	☐	🟥	🟥
	Vibrio splendidus	🟥	☐	🟥	🟥
	Reinekea sp.	🟥	☐	🟥	🟥
	Xanthomonas campestris	🟥	☐	🟥	☐
	Escherichia coli	☐	☐	🟥	🟥
	Serratia marcescens	☐	☐	🟥	🟥
	Xylella fastidiosa	☐	☐	🟥	☐

FIG. 2. Comparative genomic evidence implicating *folX* and *folM* in tetrahydromonapterin synthesis. (A) Phyletic distribution of phenylalanine hydroxylase (*phhA*), PTPS-II, *folM*, and *folX* genes among representative genomes. Note the anticorrelation between PTPS-II and *folM* or *folX*. Note also that *folM* sometimes is unaccompanied by *folX*, and that *folM* and *folX* sometimes occur without *phhA*. (B) Clustering of *folX* and *folM* with each other and with *folE* in diverse *Gammaproteobacteria*. In two genomes these genes also cluster with *folK*. Genes are colored to match Fig. 1C. Arrows indicate the transcriptional direction; overlaps between arrows indicate translational coupling.



☐ Gene absent
\+ Gene present
🟧 + Gene present and clustered

Species	*folM*	*folB*	*folX*
Colwellia psychrerythraea	🟧 +	+	🟧 +
Saccharophagus degradans	🟧 +	🟧 +	
Pseudoalteromonas haloplanktis	🟧 +	+	🟧 +
Shewanella frigidimarina	🟧 +	+	🟧 +
Methylococcus capsulatus	🟧 +	🟧 +	
Alcanivorax borkumensis	🟧 +	🟧 +	
Azotobacter vinelandii	🟧 +	+	🟧 +
Cellvibrio japonicus	🟧 +	🟧 +	
Pseudomonas aeruginosa	🟧 +	+	🟧 +
Pseudomonas mendocina	🟧 +	+	🟧 +
Pseudomonas fluorescens	🟧 +	+	🟧 +
Pseudomonas syringae	🟧 +	+	🟧 +
Vibrio sp. MED222	🟧 +	+	🟧 +
Vibrio splendidus	🟧 +	+	🟧 +
Kangiella koreensis	🟧 +	🟧 +	
Reinekea sp. MED297	🟧 +	+	🟧 +

FIG. 6. Comparative genomic evidence bearing on possible metabolic channeling between FolB and FolK and on the possible substrate for FolM in bacteria without FolX. (A) The widespread occurrence in folate synthesis gene clusters of *folB-folK* fusion genes. The species shown represent four diverse phyla (*Actinobacteria*, *Alphaproteobacteria*, *Firmicutes*, and *Fusobacteria*). (B) Clustering patterns of *folM* genes in *Gammaproteobacteria* with or without *folX* genes. Note that *folM-folB* clusters occur only if *folX* is absent. The *folM-folB* clusters also include *folK*, as illustrated by two representative examples (*Kangiella koreensis* and *Methylococcus capsulatus*).

Assignment

Confirm your sequences through multiple alignment

Human sequence
- chromosome
- gene structure

Microbial sequence
- neighboring sequences