Multiple Sequence Alignment

Andrew Michael

Gene I chose to study

envZ gene from *Escherichia coli*

A histidine kinase that responds to osmolarity changes in the medium of the cell by regulating phosphorylation.

Gene sequence:

```
ATGAGGCGATTGCGCTTCTCGCCACGAAGTTCATTTGCCCGTACGTTATTGCTCATCGTCACCTTGCTGT
TCGCCAGCCTGGTGACGACTTATCTGGTGGTGCTGAACTTCGCGATTTTGCCGAGCCTCCAGCAGTTTAA
TAAAGTCCTCGCGTACGAAGTGCGTATGTTGATGACCGACAAACTGCAACTGGAGGACGGCACGCAGTTG
GTTGTGCCTCCCGCTTTCCGTCGGGAGATCTACCGTGAGCTGGGGATCTCTCTCTACTCCAACGAGGCTG
CCGAAGAGGCAGGTCTGCGTTGGGCGCAACACTATGAATTCTTAAGCCATCAGATGGCGCAGCAACTGGG
CGGCCCGACGGAAGTGCGCGTTGAGGTCAACAAAAGTTCGCCTGTCGTCTGGCTGAAAACCTGGCTGTCG
CCCAATATCTGGGTACGCGTGCCGCTGACCGAAATTCATCAGGGCGATTTCTCTCCGCTGTTCCGCTATA
CGCTGGCGATTATGCTATTGGCGATAGGCGGGGCGTGGCTGTTTATTCGTATCCAGAACCGACCGTTGGT
CGATCTCGAACACGCAGCCTTGCAGGTTGGTAAAGGGATTATTCCGCCGCCGCTGCGTGAGTATGGCGCT
TCGGAGGTGCGTTCCGTTACCCGTGCCTTTAACCATATGGCGGCTGGTGTTAAGCAACTGGCGGATGACC
GCACGCTGCTGATGGCGGGGGTAAGTCACGACTTGCGCACGCCGCTGACGCGTATTCGCCTGGCGACTGA
GATGATGAGCGAGCAGGATGGCTATCTGGCAGAATCGATCAATAAAGATATCGAAGAGTGCAACGCCATC
ATTGAGCAGTTTATCGACTACCTGCGCACCGGGCAGGAGATGCCGATGGAAATGGCGGATCTTAATGCAG
TACTCGGTGAGGTGATTGCTGCCGAAAGTGGCTATGAGCGGGAAATTGAAACCGCGCTTTACCCCGGCAG
CATTGAAGTGAAAATGCACCCGCTGTCGATCAAACGCGCGGTGGCGAATATGGTGGTCAACGCCGCCCGT
TATGGCAATGGCTGGATCAAAGTCAGCAGCGGAACGGAGCCGAATCGCGCCTGGTTCCAGGTGGAAGATG
ACGGTCCGGGAATTGCGCCGGAACAACGTAAGCACCTGTTCCAGCCGTTTGTCCGCGGCGACAGTGCGCG
CACCATTAGCGGCACGGGATTAGGGCTGGCAATTGTGCAGCGTATCGTGGATAACCATAACGGGATGCTG
GAGCTTGGCACCAGCGAGCGGGGCGGGCTTTCCATTCGCGCCTGGCTGCCAGTGCCGGTAACGCGGGCGC
AGGGCACGACAAAAGAAGGGTAA
```

Using BLAST

BLASTx search translated nucleotide -> protein

Citrobacter farmeri (96.22% identity)

Salmonella bongori (95.56% identity)

Yokenella regensburgei (93.71% identity)

Using BLAST

tBLASTn search translated protein -> translated nucleotide

Citrobacter farmeri

Salmonella bongori

Yokenella regensburgei

T-Coffee

Genes:

MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric_Notredame
SCORE=733
*
 BAD AVG GOOD
*
Escherichia    :  85
Citrobacter    :  67
Salmonella     :  85
Yokenella      :  59
cons           :  73

Escherichia    ----------------ATGAGGC---------GATTGCGCTTCTCGCCACGAAG---------------
Citrobacter    ----------------ATGAGGC---------GAATGCGCTTCTCGCCACGAAG---------------
Salmonella     ----------------ATGAGGC---------GAATGCGCTTCTCGCCGCGAAG---------------
Yokenella      ACTCTCTTTCTGTCCCGCGATGCGCGATACCGGAATCGGTAGCCAGGCGCGAATCAACAAGCCACCCCG

cons                          ** **          ** *  *   *  * * ****

Escherichia    TT---------CATTTGCCCGTACGTTAT--TGCTCATCGTCACCTTGCTGTTCGCCAGCCTGGTGACG
Citrobacter    TT---------CATTTGCCCGCACGCTGC--TGCTCATCGTCACTCTGCTGTTTGTCAGCCTGGTGACG
Salmonella     TT---------CATTTGCTCGCACGCTGT--TGCTCATCGTCACCTTGCTGTTCGTTAGCCTGGTGACG
Yokenella      TTCGCTGGTGCCAATTTCCAGCAGACCGTTATGGTTATCGATAATACGCTGCACAATCGCCAGACCGAG

cons           **          ** ** *  * *       ** * ****  *    ****     *** *     *

Escherichia    ACTTAT-CTGGTGGTGCTGAACTTCGCGATTTTGCCGAGCCTCCAGCAGTTTAATAAAGTCCTCGCGTA
Citrobacter    ACTTAT-CTGGTGGTGCTGAACTTTGCGATTCTGCCGAGTCTCCAGCAGTTTAATAAGGTCCTGGCCTA
Salmonella     ACCTAC-CTGGTGGTGCTGAACTTCGCGATCTTACCGAGCCTCCAGCAGTTTAATAAGGTTCTGGCTTA
Yokenella      CCCCGTACCGCTGGTGCTGCGC---GCGCTGTCGCCGCGAA-CAAACGGCTGGAACAGGT----GCTTA

cons            *   * * ******** *   *** *    *** *   * * * * *  * * **     ** **

Escherichia    CGAAGTGCGTATGTTGAT-----GACCGACAAACTGCAACTGGAGGACGGCACGCAGTTGGTTGTGCCT
Citrobacter    CGAAGTGCGTATGCTGAT-----GACCGATAAACTGCAACTGGAGGACGGCACGCAACTGGTGGTGCCT
Salmonella     TGAAGTCCGTATGCTGAT-----GACCGATAAGCTGCAACTGGAGGACGGTACGCAATTAGTTGTGCCT
Yokenella      CGCTGCTCGGGT-TTGATCCCCGGGCCGTCATCTTCCACCTGGAA--CCAGGCGCGATTCGGTTCACTC
```

T-Coffee

Proteins:

E. coli

C. farmeri

S. bongori

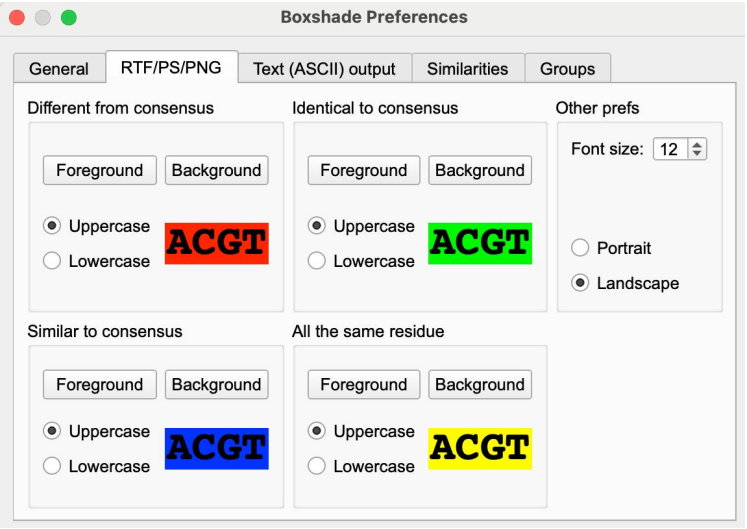
Y. regensburgei

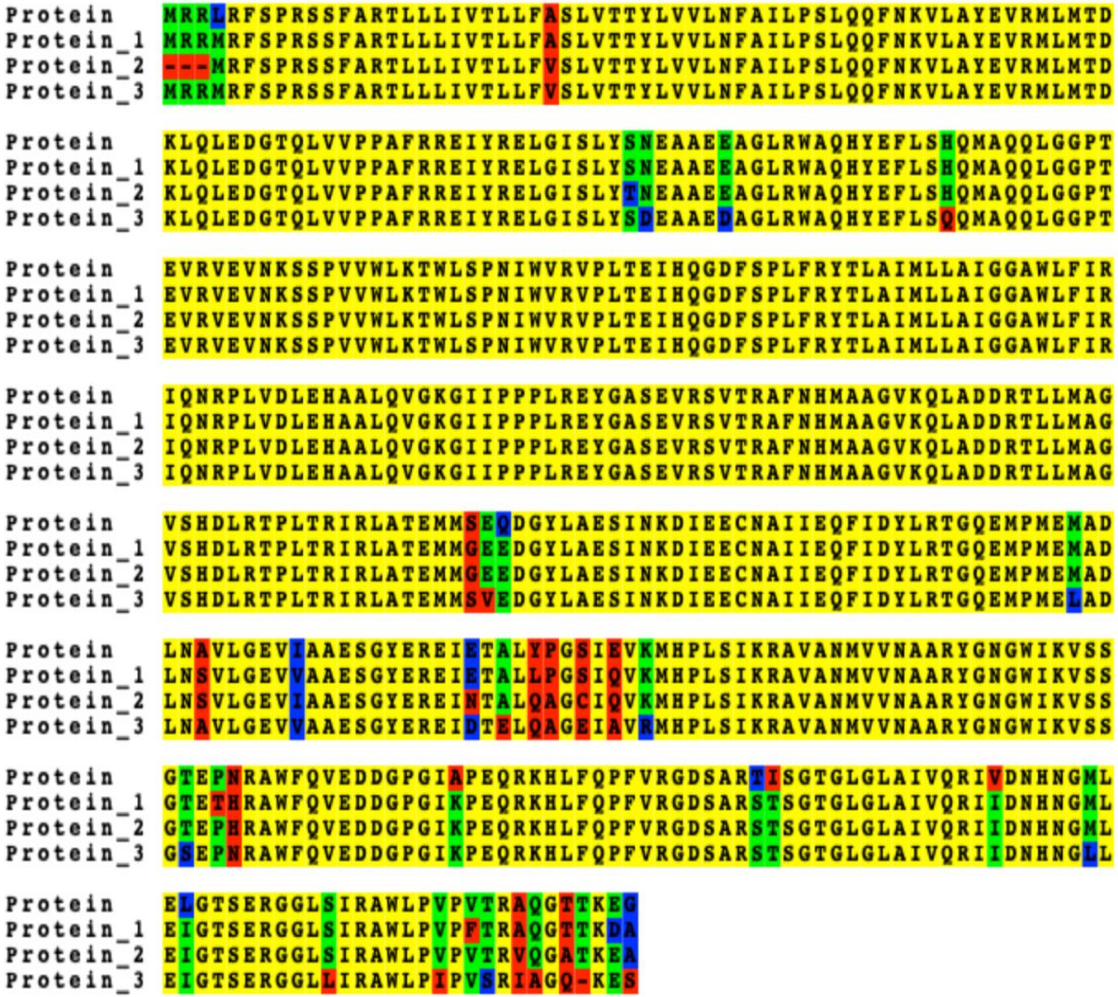
MSA

The multiple sequence alignment result as produced by T-coffee.

```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
SCORE=996
*
 BAD AVG GOOD
*
Protein      :  99
Protein_1    :  99
Protein_2    :  99
Protein_3    :  99
cons         :  99

Protein      MRRLRFSPRSSFARTLLLIVTLLFASLVTTYLVVLNFAILPSLQQFNKVLAYEVRMLMTDKLQLEDGTQ
Protein_1    MRRMRFSPRSSFARTLLLIVTLLFASLVTTYLVVLNFAILPSLQQFNKVLAYEVRMLMTDKLQLEDGTQ
Protein_2    ---MRFSPRSSFARTLLLIVTLLFVSLVTTYLVVLNFAILPSLQQFNKVLAYEVRMLMTDKLQLEDGTQ
Protein_3    MRRMRFSPRSSFARTLLLIVTLLFVSLVTTYLVVLNFAILPSLQQFNKVLAYEVRMLMTDKLQLEDGTQ

cons            :******************.****************************************

Protein      LVVPPAFRREIYRELGISLYSNEAAEEAGLRWAQHYEFLSHQMAQQLGGPTEVRVEVNKSSPVVWLKTW
Protein_1    LVVPPAFRREIYRELGISLYSNEAAEEAGLRWAQHYEFLSHQMAQQLGGPTEVRVEVNKSSPVVWLKTW
Protein_2    LVVPPAFRREIYRELGISLYTNEAAEEAGLRWAQHYEFLSHQMAQQLGGPTEVRVEVNKSSPVVWLKTW
Protein_3    LVVPPAFRREIYRELGISLYSDEAAEDAGLRWAQHYEFLSQQMAQQLGGPTEVRVEVNKSSPVVWLKTW

cons         ******************:::****:************:**************************

Protein      LSPNIWVRVPLTEIHQGDFSPLFRYTLAIMLLAIGGAWLFIRIQNRPLVDLEHAALQVGKGIIPPPLRE
Protein_1    LSPNIWVRVPLTEIHQGDFSPLFRYTLAIMLLAIGGAWLFIRIQNRPLVDLEHAALQVGKGIIPPPLRE
Protein_2    LSPNIWVRVPLTEIHQGDFSPLFRYTLAIMLLAIGGAWLFIRIQNRPLVDLEHAALQVGKGIIPPPLRE
Protein_3    LSPNIWVRVPLTEIHQGDFSPLFRYTLAIMLLAIGGAWLFIRIQNRPLVDLEHAALQVGKGIIPPPLRE

cons         ************************************************************************

Protein      YGASEVRSVTRAFNHMAAGVKQLADDRTLLMAGVSHDLRTPLTRIRLATEMMSEQDGYLAESINKDIEE
Protein_1    YGASEVRSVTRAFNHMAAGVKQLADDRTLLMAGVSHDLRTPLTRIRLATEMMGEEDGYLAESINKDIEE
Protein_2    YGASEVRSVTRAFNHMAAGVKQLADDRTLLMAGVSHDLRTPLTRIRLATEMMGEEDGYLAESINKDIEE
Protein_3    YGASEVRSVTRAFNHMAAGVKQLADDRTLLMAGVSHDLRTPLTRIRLATEMMSVEDGYLAESINKDIEE
```

Boxshade


Boxshade Preferences
General
RTF/PS/PNG
Text (ASCII) output
Similarities
Groups
Different from consensus
Foreground
Background
Uppercase
Lowercase
ACGT
Identical to consensus
Foreground
Background
Uppercase
Lowercase
ACGT
Other prefs
Font size: 12
Portrait
Landscape
Similar to consensus
Foreground
Background
Uppercase
Lowercase
ACGT
All the same residue
Foreground
Background
Uppercase
Lowercase
ACGT


Protein MRRLRFSPRSSFARTLLLIVTLLFASLVTTYLVVLNFAILPSLQQFNKVLAYEVRMLMTD
Protein_1 MRRMRFSPRSSFARTLLLIVTLLFASLVTTYLVVLNFAILPSLQQFNKVLAYEVRMLMTD
Protein_2 ---MRFSPRSSFARTLLLIVTLLFVSLVTTYLVVLNFAILPSLQQFNKVLAYEVRMLMTD
Protein_3 MRRMRFSPRSSFARTLLLIVTLLFVSLVTTYLVVLNFAILPSLQQFNKVLAYEVRMLMTD
Protein KLQLEDGTQLVVPPAFRREIYRELGISLYSNEAAEEAGLRWAQHYEFLSHQMAQQLGGPT
Protein_1 KLQLEDGTQLVVPPAFRREIYRELGISLYSNEAAEEAGLRWAQHYEFLSHQMAQQLGGPT
Protein_2 KLQLEDGTQLVVPPAFRREIYRELGISLYTNEAAEEAGLRWAQHYEFLSHQMAQQLGGPT
Protein_3 KLQLEDGTQLVVPPAFRREIYRELGISLYSDEAAEIAGLRWAQHYEFLSQQMAQQLGGPT
Protein EVRVEVNKSSPVVWLKTWLSPNIWVRVPLTEIHQGDFSPLFRYTLAIMLLAIGGAWLFIR
Protein_1 EVRVEVNKSSPVVWLKTWLSPNIWVRVPLTEIHQGDFSPLFRYTLAIMLLAIGGAWLFIR
Protein_2 EVRVEVNKSSPVVWLKTWLSPNIWVRVPLTEIHQGDFSPLFRYTLAIMLLAIGGAWLFIR
Protein_3 EVRVEVNKSSPVVWLKTWLSPNIWVRVPLTEIHQGDFSPLFRYTLAIMLLAIGGAWLFIR
Protein IQNRPLVDLEHAALQVGKGIIPPPLREYGASEVRSVTRAFNHMAAGVKQLADDRTLLMAG
Protein_1 IQNRPLVDLEHAALQVGKGIIPPPLREYGASEVRSVTRAFNHMAAGVKQLADDRTLLMAG
Protein_2 IQNRPLVDLEHAALQVGKGIIPPPLREYGASEVRSVTRAFNHMAAGVKQLADDRTLLMAG
Protein_3 IQNRPLVDLEHAALQVGKGIIPPPLREYGASEVRSVTRAFNHMAAGVKQLADDRTLLMAG
Protein VSHDLRTPLTRIRLATEMMSEQDGYLAESINKDIEECNAIIEQFIDYLRTGQEMPMEMAD
Protein_1 VSHDLRTPLTRIRLATEMMGEEDGYLAESINKDIEECNAIIEQFIDYLRTGQEMPMEMAD
Protein_2 VSHDLRTPLTRIRLATEMMGEEDGYLAESINKDIEECNAIIEQFIDYLRTGQEMPMEMAD
Protein_3 VSHDLRTPLTRIRLATEMMSVEDGYLAESINKDIEECNAIIEQFIDYLRTGQEMPMELAD
Protein LNAVLGEVIAAESGYEREIETALYPGSIEVKMHPLSIKRAVANMVVNAARYGNGWIKVSS
Protein_1 LNSVLGEVVAAESGYEREIETALLPGSIQVKMHPLSIKRAVANMVVNAARYGNGWIKVSS
Protein_2 LNSVLGEVIAAESGYEREINTALQAGCIQVKMHPLSIKRAVANMVVNAARYGNGWIKVSS
Protein_3 LNAVLGEVVAAESGYEREIDTELQAGEIAVRMHPLSIKRAVANMVVNAARYGNGWIKVSS
Protein GTEPNRAWFQVEDDGPGIAPEQRKHLFQPFVRGDSARTISGTGLGLAIVQRIVDNHNGML
Protein_1 GTETHRAWFQVEDDGPGIKPEQRKHLFQPFVRGDSARSTSGTGLGLAIVQRIIDNHNGML
Protein_2 GTEPHRAWFQVEDDGPGIKPEQRKHLFQPFVRGDSARSTSGTGLGLAIVQRIIDNHNGML
Protein_3 GSEPNRAWFQVEDDGPGIKPEQRKHLFQPFVRGDSARSTSGTGLGLAIVQRIIDNHNGLL
Protein ELGTSERGGLSIRAWLPVPVTRAQGTTKEG
Protein_1 EIGTSERGGLSIRAWLPVPFTRAQGTTKDA
Protein_2 EIGTSERGGLSIRAWLPVPVTRVQGATKEA
Protein_3 EIGTSERGGLLIRAWLPIPVSRIAGQ-KES

Pairwise Distance

	1	2	3	4
1. Protein sequence envZ from E. coli				
2. Protein sequence envZ from C. farmeri	0.0341			
3. Protein sequence envZ from S. bongori	36.2194	35.6468		
4. Protein sequence envZ from Y. regensburgei	0.0701	0.0628	36.2741	

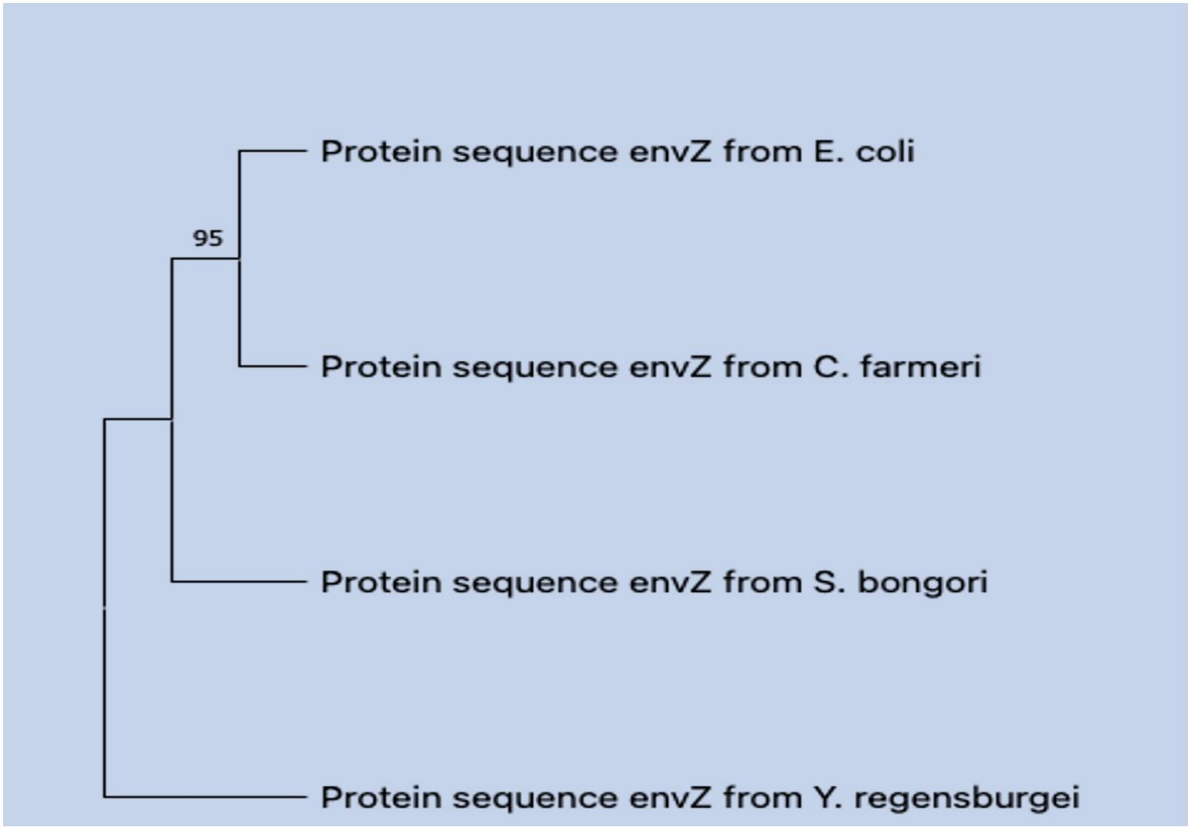
Phylogenetic Trees

M11: Analysis Preferences

Phylogeny Reconstruction

Option	Setting
Statistical Method	*Maximum Likelihood*
Test of Phylogeny	*Bootstrap method*
No. of Bootstrap Replications	*1000*
Substitutions Type	*Amino acid*
Model/Method	*Jones-Taylor-Thornton (JTT) model*
Rates among Sites	*Uniform Rates*
No of Discrete Gamma Categories	Not Applicable
Gaps/Missing Data Treatment	*Use all sites*
Site Coverage Cutoff (%)	Not Applicable
ML Heuristic Method	*Nearest-Neighbor-Interchange (NNI)*
Initial Tree for ML	*Make initial tree automatically (Default - NJ/Bio*
Initial Tree File	Not Applicable
Branch Swap Filter	*None*
Number of Threads	*3*

Help Cancel OK



Phylogenetic Trees

M11: Analysis Preferences

Phylogeny Reconstruction

Option	Setting
Scope	*All Selected Taxa*
Statistical Method	*Neighbor-joining*
Test of Phylogeny	*Bootstrap method*
No. of Bootstrap Replications	*10000*
Substitutions Type	*Amino acid*
Model/Method	*Jones-Taylor-Thornton (JTT) model*
Rates among Sites	*Uniform Rates*
Gamma Parameter	Not Applicable
Pattern among Lineages	*Same (Homogeneous)*
Gaps/Missing Data Treatment	*Pairwise deletion*
Site Coverage Cutoff (%)	Not Applicable
Number of Threads	*3*

Help | Cancel | OK