Exploring the diversity of the nitrogen fixation gene *nifH*

Isabella Asselstine, Xinran Li, Aakanx Panchal, Dan Quesnelle

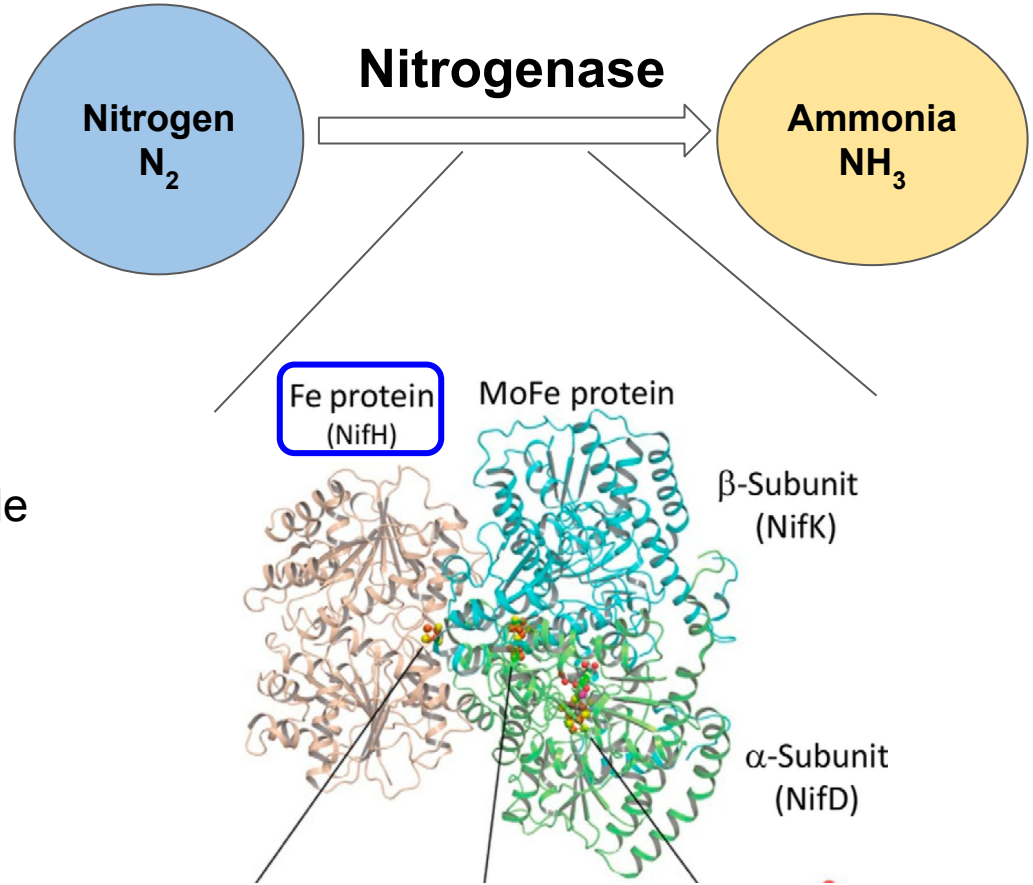
Group 1

Queen's UNIVERSITY

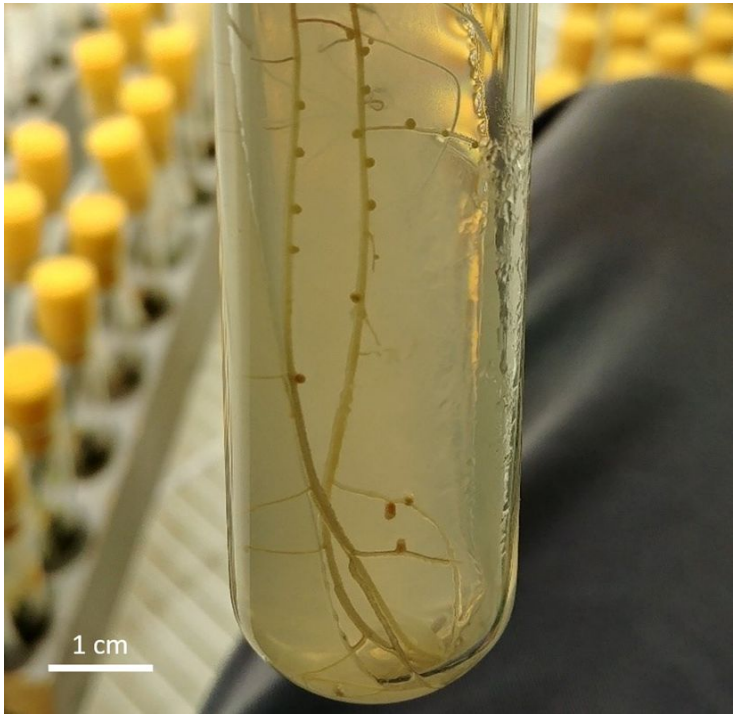
Background

Nitrogen fixation

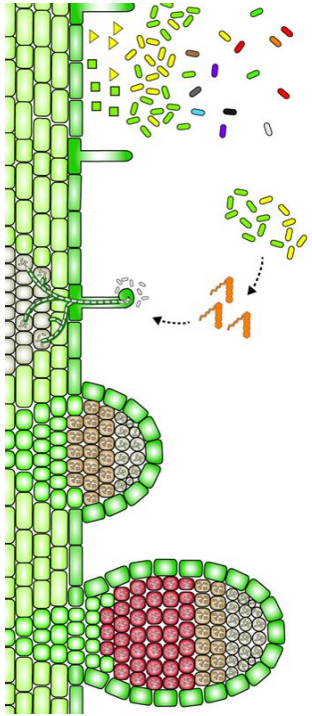
- Nitrogen fixation is the conversion of stable nitrogen into a more reactive form such as ammonia
- Bacterial nitrogenases are responsible for over 90% of nitrogen fixation on earth
- Nitrogenases are coded by *nif* genes
- *nifH* encodes a nitrogenase subunit



Sinorhizobium meliloti 2011



Sinorhizobium is housed in legume nodules



rhizobia produce Nodulation Factors that bind legume Nodulation Factor Receptors

nitrogen fixation:

$N_2 + 8H^+ + 8e^- + 16ATP \rightarrow 2NH_3 + H_2 + 16ADP + 16P_i$

Questions to be addressed

1. What interesting insights can we glean from an alignment, distance matrix, and phylogeny of 100 *nifH* gene sequences?

2. Are there non-rhizobium *nifH* sequences that are closer to the *nifH* sequence of the model rhizobium *Sinorhizobium meliloti* 2011 than the *nifH* sequences of other rhizobia?

Methods

nifH gene sequences were retrieved using NCBI Gene IDs

Search results

Items: 1 to 20 of 453

<< First < Prev Page 1 of 23 Next > Last >>

See also 690 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases
nifH ID: 11970903	nitrogenase iron protein [*Methanocella conradii HZ254*]	NC_017034.1 (137701..138519)	MTC_RS00730, Mtc_0142
nifH ID: 9742706	nitrogenase iron protein [*Methanolacinia petrolearia DSM 11571*]	NC_014507.1 (276717..277538, complement)	MPET_RS01330, Mpet_0263
nifH ID: 5144036	nitrogenase iron protein [*Methanocella arvoryzae MRE50*]	NC_009464.1 (1392058..1392882, complement)	RCI_RS07010, RCIX1606

nifH gene sequences were retrieved using NCBI Gene IDs

Number	Gene_ID	Name
1	11970903	nitrogenase iron protein [Methanocella conradii HZ254]
2	9742706	nitrogenase iron protein [Methanolacinia petrolearia DSM 11571]
3	5144036	nitrogenase iron protein [Methanocella arvoryzae MRE50]
4	5411589	nitrogenase iron protein [Methanoregula boonei 6A8]
5	31573395	nitrogenase iron protein [Paenibacillus odorifer]
6	25012506	nitrogenase iron protein [Sinorhizobium meliloti GR4]
7	66135341	nitrogenase iron protein [Methanosarcina mazei]
8	66132306	nitrogenase iron protein [Methanobrevibacter arboriphilus]
9	65565472	nitrogenase iron protein [Methanospirillum hungatei]
10	65097490	nitrogenase iron protein [Methanospirillum sp. J.3.6.1-F.2.7.3]

gene_IDs.R

Output list of gene IDs

gene_id_parser.sh

Retrieve gene sequences

nifh_sequences1.fasta

Isolation of *nifH* NCBI Gene IDs in R

```
get_gene_IDs = function(InputName) {

  gene_IDs = read.csv(file =InputName)

  ids = as.numeric(paste(gene_IDs$Gene_ID, sep = " "))

  return(ids)
}
```

Script (gene_IDs.R)

```
11970903  9742706  5144036  5411589 31573395 25012506 66135341 66132306
65565472 65097490 58027523 57209642 66686341 64294505 66558925 60795032
66379201 64336658 44996749 31489522 29419855 29418613 24886632  7271977
 1479061 61599259 61514986 29763669  9735509  1475788 56451760 70914963
69754478 69732137 69525103 66895772 66892402 66478756 66432510 66344908
66343528 66149407 64067074 64021609 61614856 61428909 61055496 58409220
48977766 48977636 45960649 45960611 24876744 24863695 24846099 24822259
24801823  3624474 61408688 65304567 61408688 70906580 70583870 68876202
68377734 67488582 66643971 66566559 66429654 66392686 66392181 66144139
62372264 61929281 61929156 61283970 58726859 57504914 55820408 49323801
45960570 41607343 24873336 24862159 24829851 24826380 24809358 24790299
 1451768 70363091 69043340 67488558 66686368 60431491 58788210 24824788
 3625965 10981576 44086063 14654102
```

Output (Gene IDs)

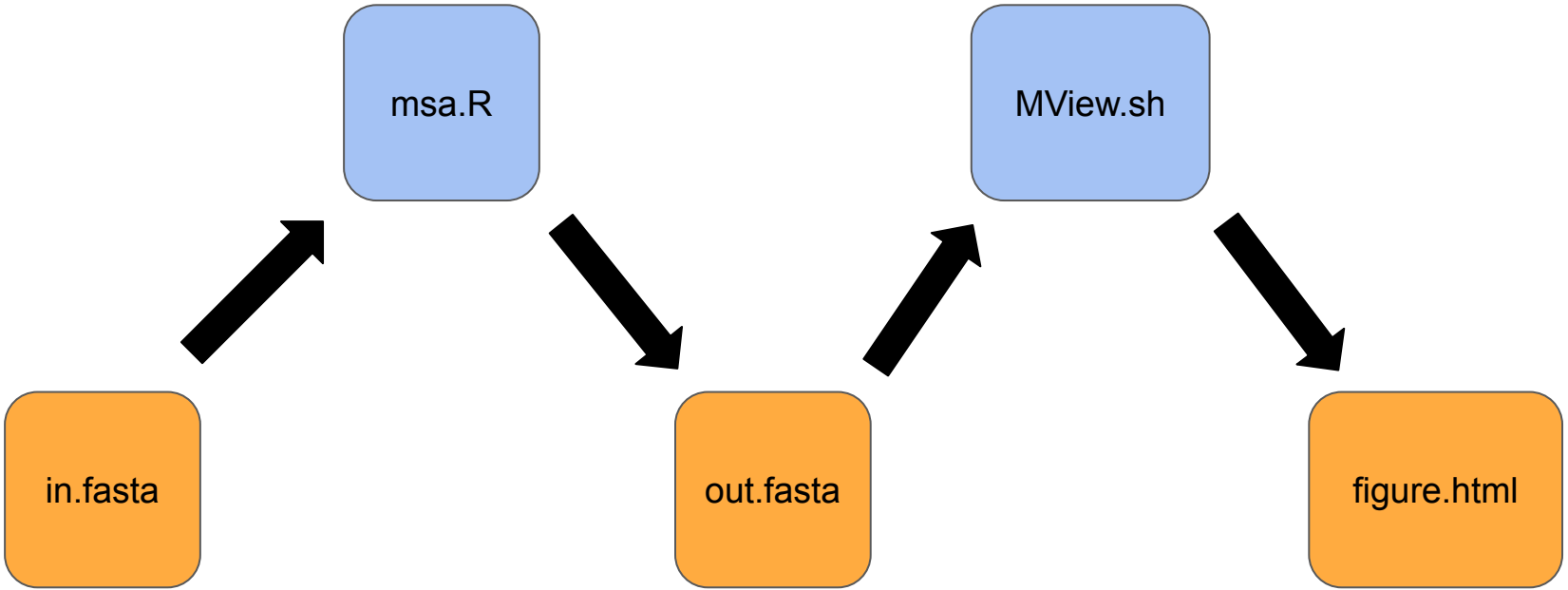
Retrieval of *nifH* gene sequences using Entrez Direct

Script: gene_id_parser.sh

```
# This script comes from the EDirect cookbook repository on Githib
# Link: https://github.com/NCBI-Hackathons/EDirectCookbook/blob/master/EDirect_Cookbook.txt

# The NCBI gene IDs from gene_IDs.R are copied and pasted into the for loop.
# For each gene ID, this script will find and retrieve the gene sequence.
# It will then paste this sequence into the file nifh_sequences1.fasta for use in the alignment.

for value in { 11970903  9742706  5144036  5411589 31573395 25012506 66135341 66132306 65565472 65097490 58027523 57209642 66686341
do

efetch -db gene -id $value -format native -mode xml \
  | xtract -pattern Entrezgene-Set \
    -group Gene-commentary \
    -if Gene-commentary_type@value -equals "genomic" \
    -element Gene-commentary_accession, Gene-commentary_version \
    -block Gene-commentary_seqs \
    -element Seq-interval_from,Seq-interval_to,Na-strand@value \
  | awk 'BEGIN{FS="\t";OFS="\t"}{print $1"."$2,$3,$4,$5}' \
  | while read -r chrom start stop strand ; do
    efetch -db nuccore -id $chrom -chr_start $start -chr_stop $stop -strand $strand -format fasta >> nifh_sequences1.fasta
    done
done
```

Output: FASTA file containing all 100 *nifH* sequences

multiple sequence alignment



R package msa for producing an alignment

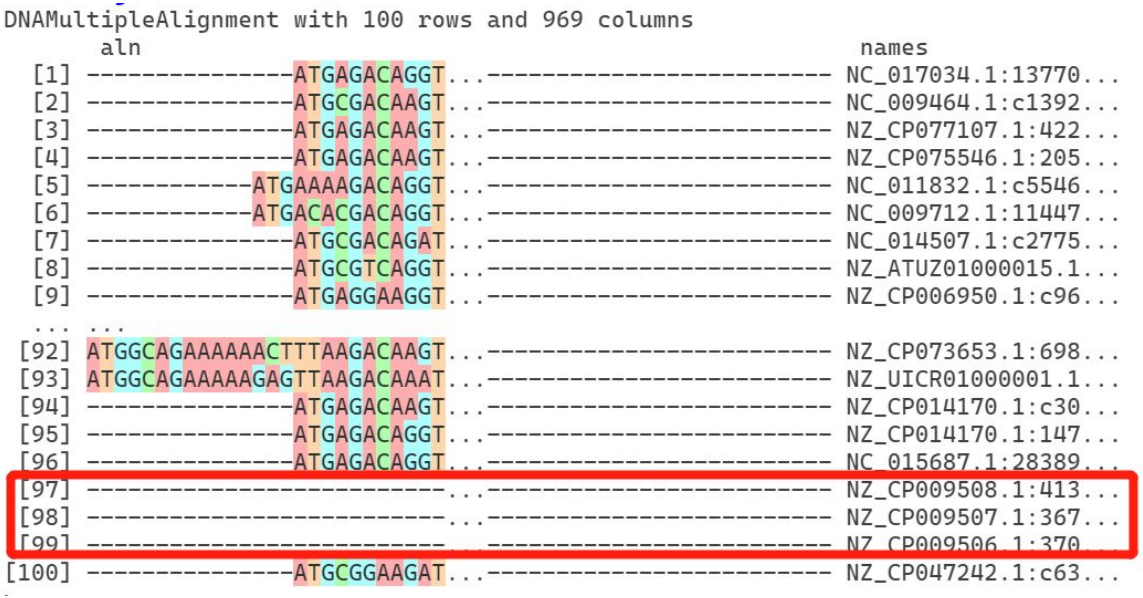
```
library(msa)

alignment <- msa(readDNAStringSet(sequences), method = "ClustalW")
# fasta to DNAStringSet for msa()
# can also select ClustalOmega, Muscle as alignment tools

msa2fasta <- function(alignment, filename) {
  sink(filename) # divert output to filename
  for(i in 1:length(rownames(alignment))) {
    cat(paste0('>', rownames(alignment)[i]), "\n",
        toString(unmasked(alignment)[[i]]), "\n")
        # write description line, sep = "" in paste0() by default and sep = " " in paste()
        # convert sequence (DNAString object) in alignment to string object
  }
  sink(NULL) # end diversion to filename
}

msa2fasta(alignment, filename)
```

MView for visualizing our alignment

```
mview.bat -in fasta -width 100 -html head -bold -css on -coloring mismatch -colormap red -ref 32 -sort pid $outfile > ./2-MSA/figure2.html
```

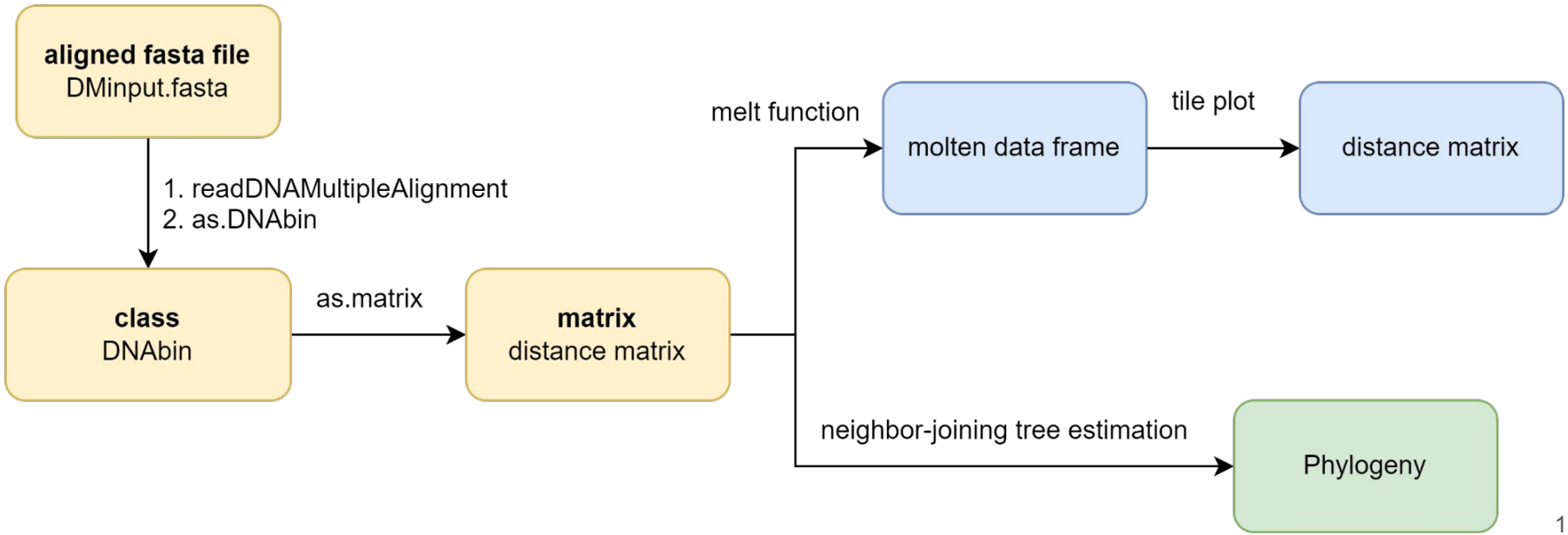
Alignment fasta preprocessing

- Remove sequences with large gaps (>60%)
- Prevent numerical errors
- Script: ratio_test.py
- Input -> output: out.fasta (100) -> DMinput.fasta(92)



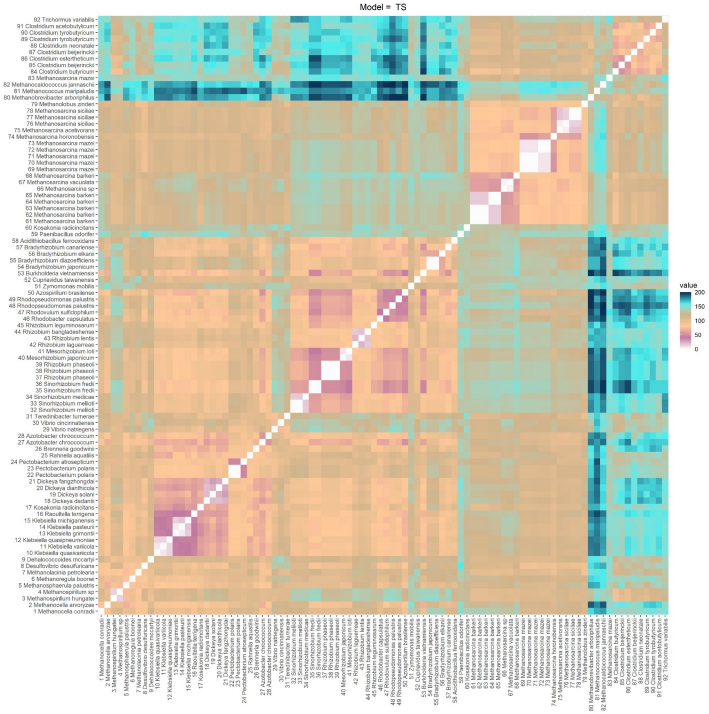
Visualization

- Plot distance matrix and phylogeny
- Script: distance_matrix_and_tree.R



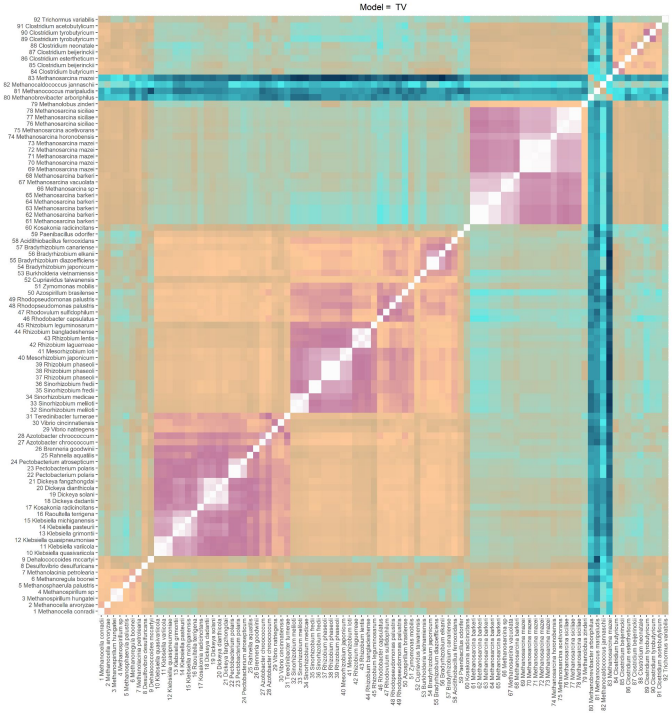
Results

Numbers of transitions & transversions



Transition

- Substitution of a purine by another: C<->T
- Substitution of a pyrimidine by another: A<->G



Transversion

- Substitution of a purine by a pyrimidine, or vice-versa:
  A<->C, A<->T, C<->G, G<->T

Number of indels

- Counts the number of sites where there is an insertion/deletion gap in one sequence and not in the other

Distance matrix: JC69

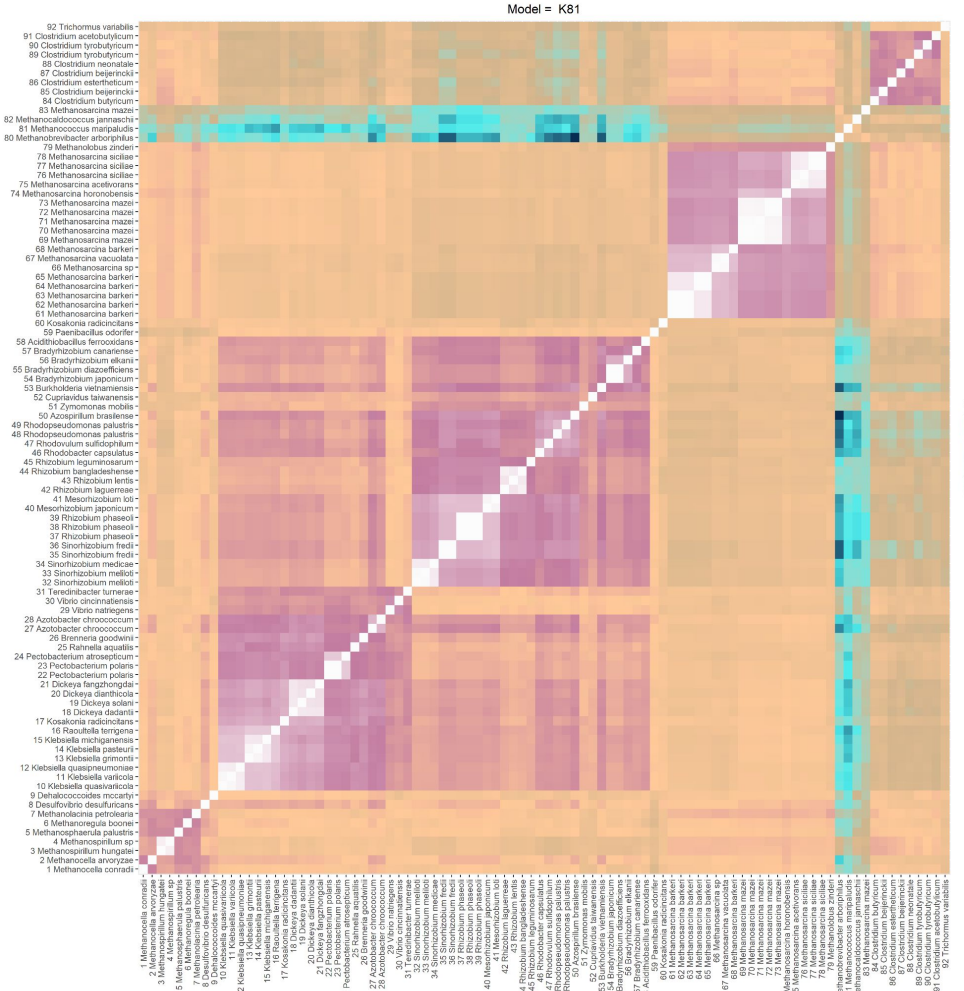
- Developed by Jukes and Cantor (1969)
- Assumptions:
  - Equal probabilities for all substitutions
  - Equal probabilities for all sites along the DNA sequence
  - Balanced base frequencies (0.25)

Distance matrix: K80

- Developed by Kimura (1980)
- Kimura's 2-parameters distance
- Assumptions:
  - Same underlying assumptions as JC69
  - Except two kinds of substitution are considered
    - Transitions
    - Transversions

Distance matrix: K81

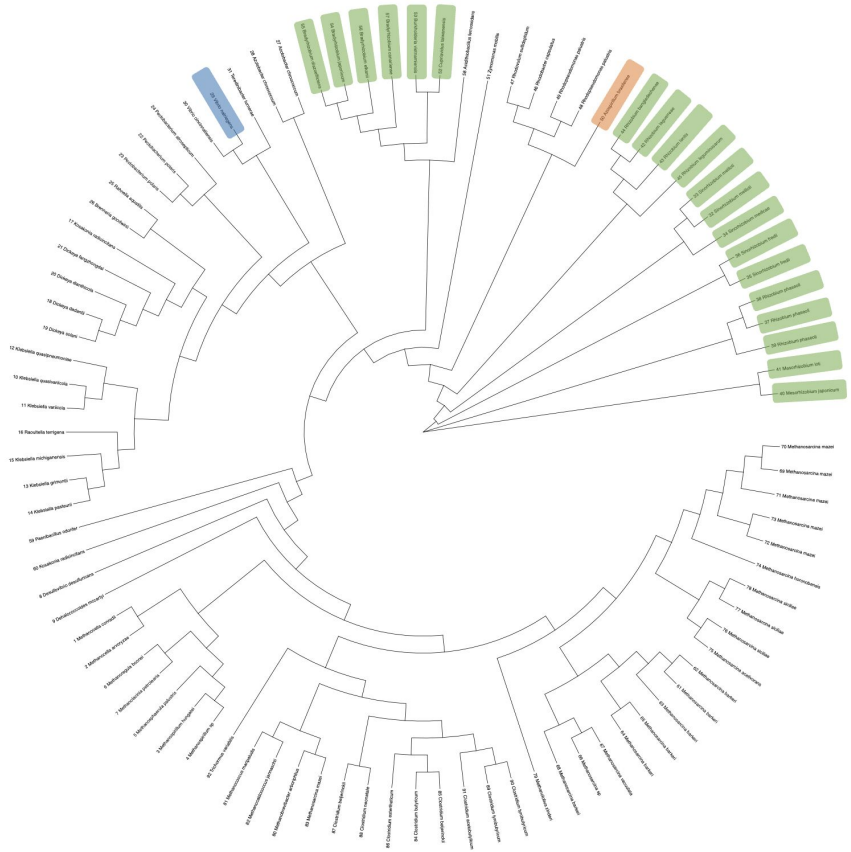
- Generalization of K80 by Kimura (1981)
- Kimura's 3-parameters distance
- Three substitution types model
- Assumptions:
  - Different rates for two kinds of transversions
    - A<->C, G<->T
    - A<->T, C<->G



Distance matrix: TN93

- Developed by Tamura and Nei (1993)
- Assumptions:
  - Distinct rates for both kinds of transitions and transversions
  - Base frequencies are estimated from data

Phylogenetic Tree

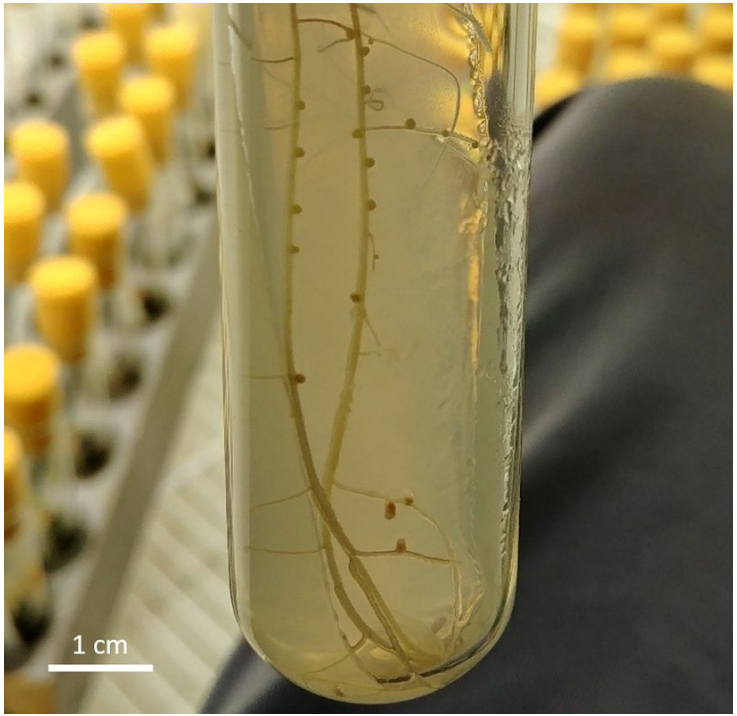


Rhizobia

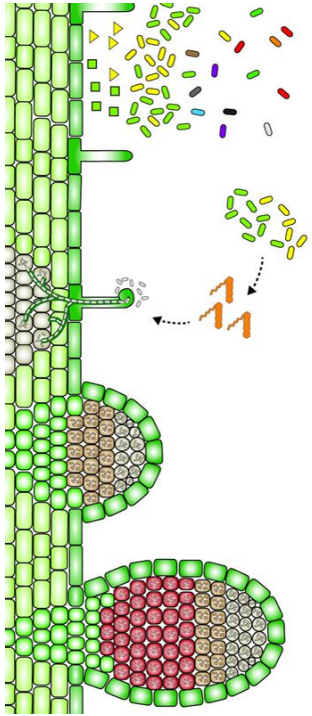
Azospirillum brasilense

Vibrio natriegens

Sinorhizobium meliloti 2011



Sinorhizobium is housed in legume nodules



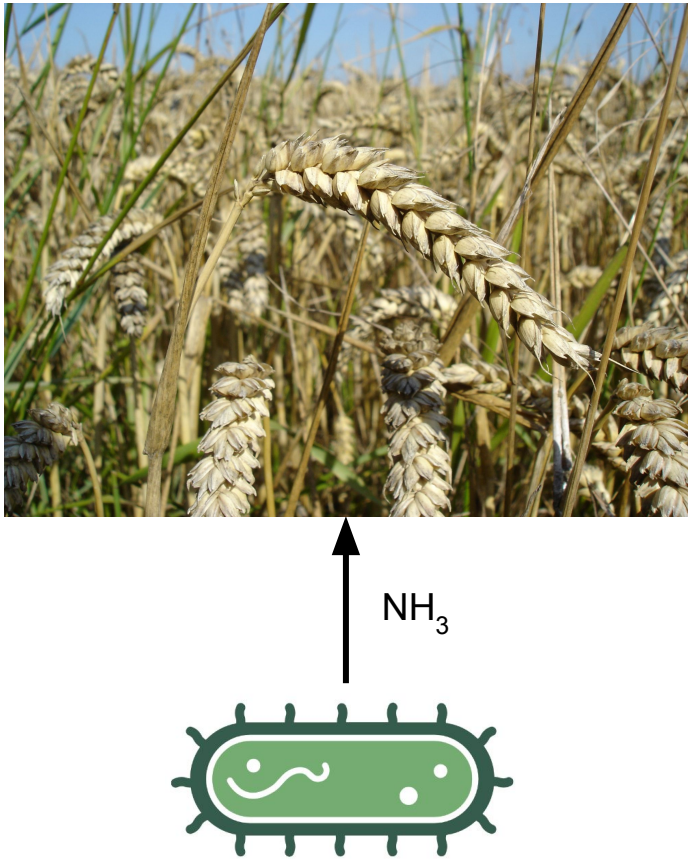
rhizobia produce Nodulation Factors that bind legume Nodulation Factor Receptors

nitrogen fixation:

$N_2 + 8H^+ + 8e^- + 16ATP \rightarrow 2NH_3 + H_2 + 16ADP + 16P_i$

Azospirillum brasilense

- Closest *nifH* sequence to that of the model rhizobium *S. meliloti* 2011 that belongs to a non-rhizobium, in terms of percent identity (80.3%)
- *A. brasilense* is a plant growth promoting rhizobacterium (PGPR) that associates closely with grasses
- Although free-living, proximity to plants leads to transfer of fixed nitrogen, unlike the intracellular transfer of fixed nitrogen by rhizobia with legumes



Rhizobia sharing less percent identity than *A. brasilense* with *S. meliloti* 2011:

(coverage for all sequences >98.7%, so only looking at percent identity)

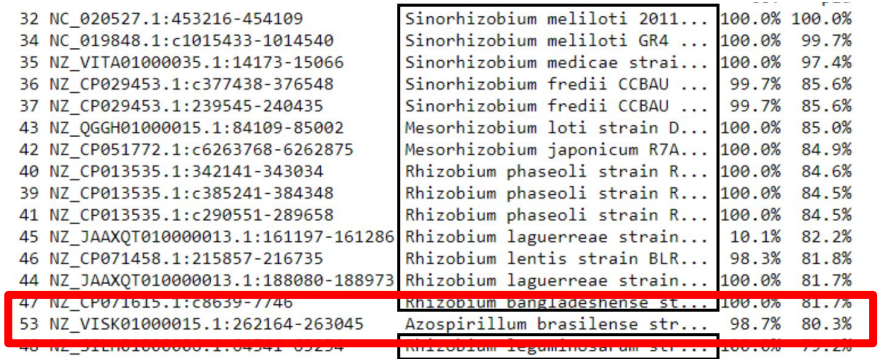
non-rhizobium:

- *Azospirillum brasilense*: 80.3%

rhizobia:

- *Rhizobium leguminosarum*: 79.2%
- *Burkholderia vietnamiensis*: 73.9%
- *Bradyrhizobium canariense*: 73.7%
- *Bradyrhizobium elkanii*: 72.2%
- *Bradyrhizobium japonicum*: 71.1%
- *Bradyrhizobium diazoefficiens*: 71.1%
- *Cupriavidus taiwanensis*: 67.8%

suggesting a recent Horizontal Gene Transfer event?

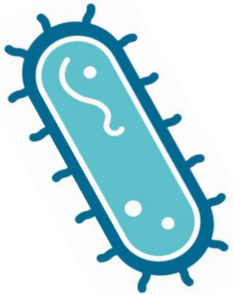


32	NC_020527.1:453216-454109	Sinorhizobium meliloti 2011...	100.0%	100.0%
34	NC_019848.1:c1015433-1014540	Sinorhizobium meliloti GR4 ...	100.0%	99.7%
35	NZ_VITA01000035.1:14173-15066	Sinorhizobium medicae strai...	100.0%	97.4%
36	NZ_CP029453.1:c377438-376548	Sinorhizobium fredii CCBAU ...	99.7%	85.6%
37	NZ_CP029453.1:239545-240435	Sinorhizobium fredii CCBAU ...	99.7%	85.6%
43	NZ_QGGH01000015.1:84109-85002	Mesorhizobium loti strain D...	100.0%	85.0%
42	NZ_CP051772.1:c6263768-6262875	Mesorhizobium japonicum R7A...	100.0%	84.9%
40	NZ_CP013535.1:342141-343034	Rhizobium phaseoli strain R...	100.0%	84.6%
39	NZ_CP013535.1:c385241-384348	Rhizobium phaseoli strain R...	100.0%	84.5%
41	NZ_CP013535.1:c290551-289658	Rhizobium phaseoli strain R...	100.0%	84.5%
45	NZ_JAAXQT010000013.1:161197-161286	Rhizobium laguerreae strain...	10.1%	82.2%
46	NZ_CP071458.1:215857-216735	Rhizobium lentis strain BLR...	98.3%	81.8%
44	NZ_JAAXQT010000013.1:188080-188973	Rhizobium laguerreae strain...	100.0%	81.7%
47	NZ_CP071615.1:c8639-7746	Rhizobium bangladeshense st...	100.0%	81.7%
53	NZ_VISK01000015.1:262164-263045	Azospirillum brasilense str...	98.7%	80.3%
48	NZ_SJEH01000000.1:84341-85234	Rhizobium leguminosarum str...	100.0%	79.2%

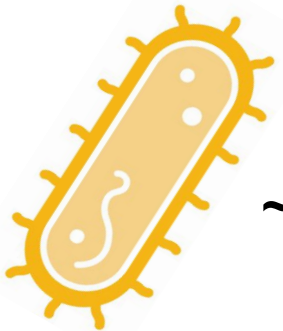
Vibrio natriegens

- *nifH* sequence with percent identity of 62.7% with that of *S. meliloti* 2011
- *V. nat.* is a marine, Gram-negative bacterium that was first isolated in 1958 in salt marsh mud
- It has been reported to have an exceptionally fast growth rate, with a doubling time of <10 min

V. nat.
<10 min



vs



E. coli
~20 min

Future Steps

- *nifHDK* concatenated alignment, accounting for all three nitrogenase subunits
- NifHDK protein alignment and tree
  - Avoid wobble bases changing relatively easily (3rd positions in codons)
    - Could be encoding same amino acid still
  - Because convergence of amino acid sequences (20 unique aa's) is expected to be more rare than convergence of DNA sequences (4 unique nt's)
    - Greater correlation between sequence similarity and homology
    - Greater signal:noise ratio
- NifHDK protein structure alignment

References

Britannica, The Editors of Encyclopaedia. "nitrogen fixation". *Encyclopedia Britannica*, 9 Jul. 2021, https://www.britannica.com/science/nitrogen-fixation. Accessed 17 April 2022.

Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. in *Mammalian Protein Metabolism*, ed. Munro, H. N., pp. 21--132, New York: Academic Press.

Kans J. (2013) Entrez Direct: E-utilities on the Unix Command Line. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16, 111--120.

Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences USA*, 78, 454--458.

Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10, 512--526.