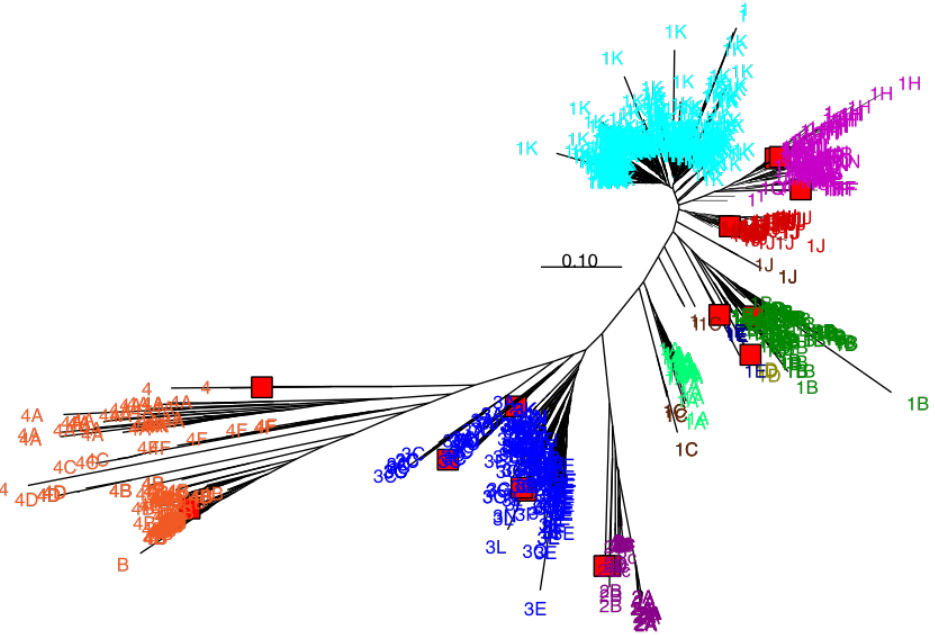
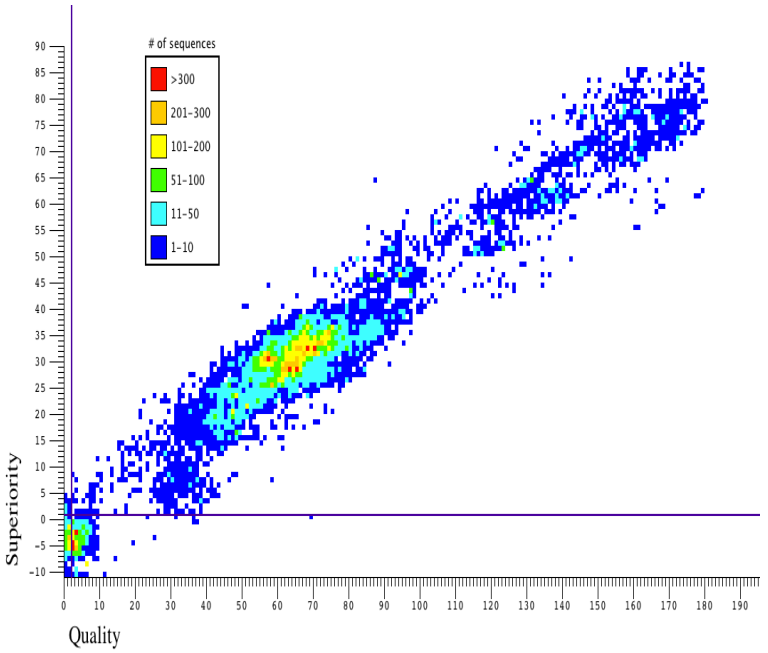
Bio/CS 123B Spring 2020
Module 8: Data Mining GenBank Conserved Domains and ARBitrator



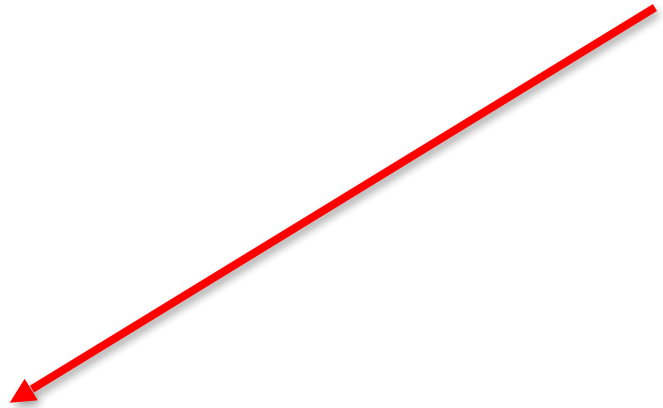


NASA GeneLab Boot Camp and Internships

- 6 internships (lightly paid) … must attend Boot Camp
- 24 extra spaces in Boot Camp, no internship
- Boot Camp: June 7 - June 11
- Internship: June 14 - Aug 18
- Apply at

  https://nams.usra.edu/programs/education/student-r-d-opportunities/

Student R&D Opportunities



Apply Now

Biosciences

Evaluation and Development of NASA GeneLab Data Processing Pipelines

Overview

Up to 6 interns will be selected for this internship and will be split into 3 groups. Each group will work on one of the following 3 projects. In your cover letter, please rank the 3 projects listed below from most to least interesting to you, why you are interested in this internship, what you hope to gain from the internship, and what makes you qualified for this internship. Please provide links to a cover letter, your CV and an unofficial transcript. Provide a link to your unofficial transcript in the last question (i.e. additional materials).

Projects

After you click “Apply Now”…

- Q3 (Nationality) – US citizenship is not required
- Q4 (Internship period) - June 14 - Aug 18
- Q5 (Technical area) - Biosciences
- Q6 (Capstone) - No
- Q7 (How did you hear about) - “SJSU Bioinformatics program”
- Q8 (Cover letter)
  - 100% written by you
  - Include Bioinformatics Minor status (declared, plan to declare, won’t declare)
  - Include SJSU i.d.
  - State if you are applying for boot camp + internship, or just boot camp
- Q9 (C.V.) - CV = “Curriculum Vitae” = resume
- Q10 (Additional materials) – Unofficial transcript from MySJSU

Recommendation email

- From a faculty member, but not your 123A or 123B instructor

Instructions to your recommender (applying for boot camp only)

Please write a short paragraph (50-100 words) explaining why you recommend this student for a 1-week boot camp to learn about NASA's GeneLab RNA-seq analysis pipeline. State the strength of your recommendation on a scale of 1 (weakest) to 10 (strongest). Email your recommendation to philip.heller@sjsu.edu by the end of day on Friday April 16.

Instructions to your recommender (applying for boot camp and internship)

Please write a short paragraph (100-200 words) explaining why you recommend this student for a summer internship at NASA's GeneLab, analyzing gene expression data from Space Shuttle missions and the International Space Station. State the strength of your recommendation on a scale of 1 (weakest) to 10 (strongest). Email your recommendation to philip.heller@sjsu.edu by the end of day on Friday April 16.

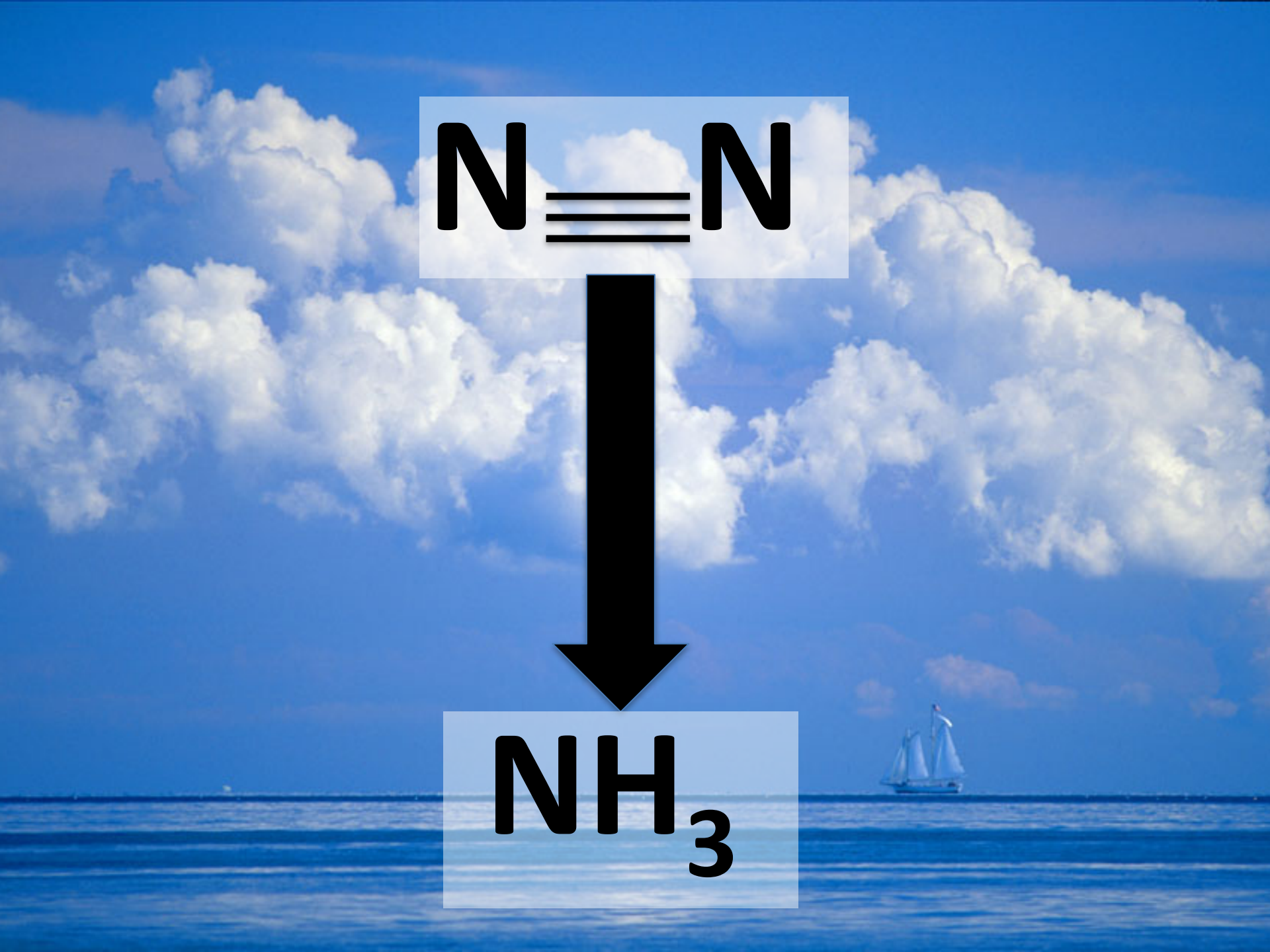
Application Deadline

- End of day Friday April 16
- Be time zone aware
- Late or incomplete applications will not be accepted for any reason

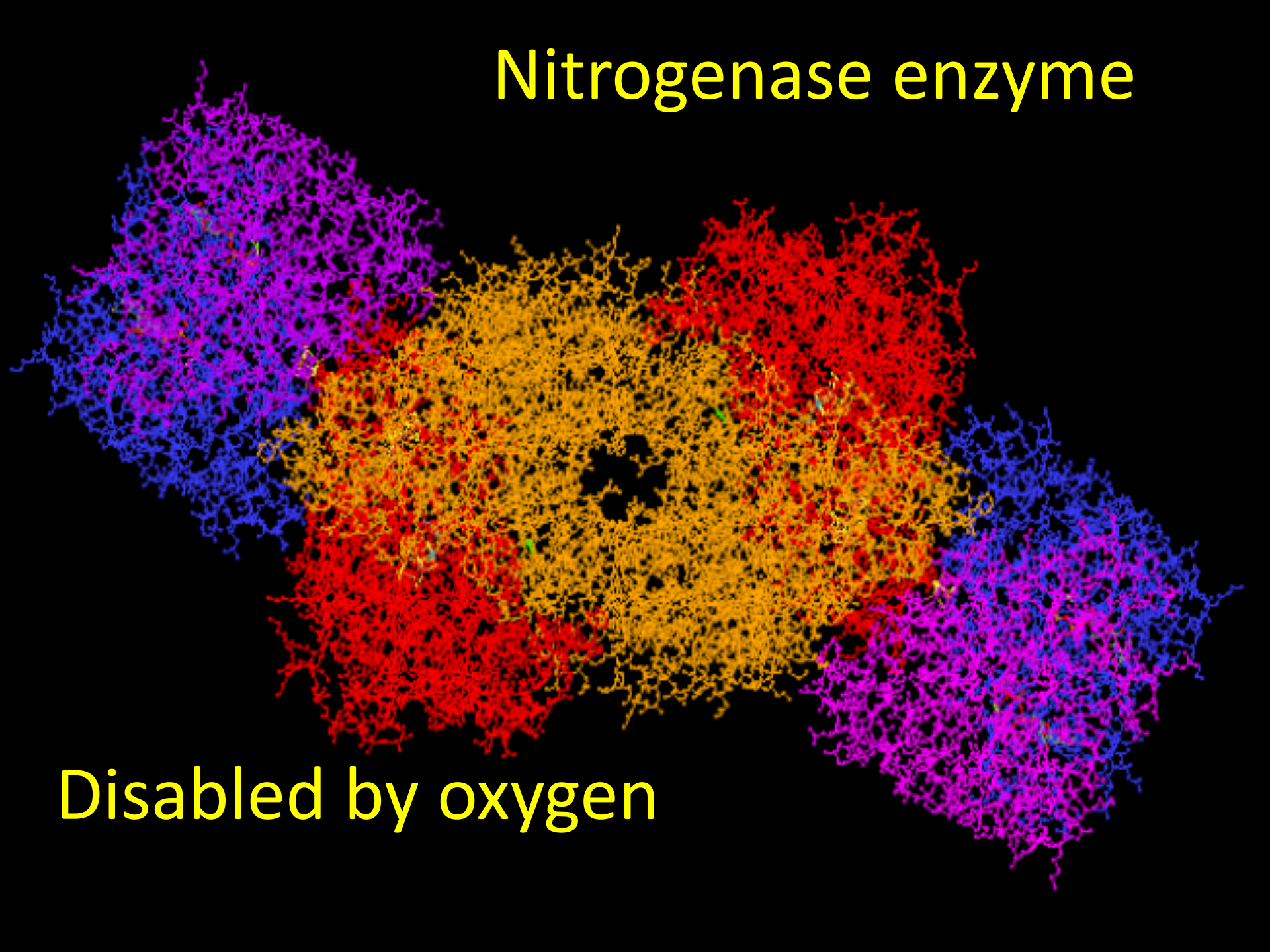
Why we care about nitrogen fixation

- Nitrogen is necessary for making DNA, RNA, and nucleic acids.
- Atmospheric nitrogen isn't useful to life unless it's reduced to a bioavailable form like ammonium

- Life in the ocean annually transforms (“fixes”, “reduces”) 4×10^{21} g of atmospheric nitrogen


N≡N
NH3

Nitrogenase enzyme

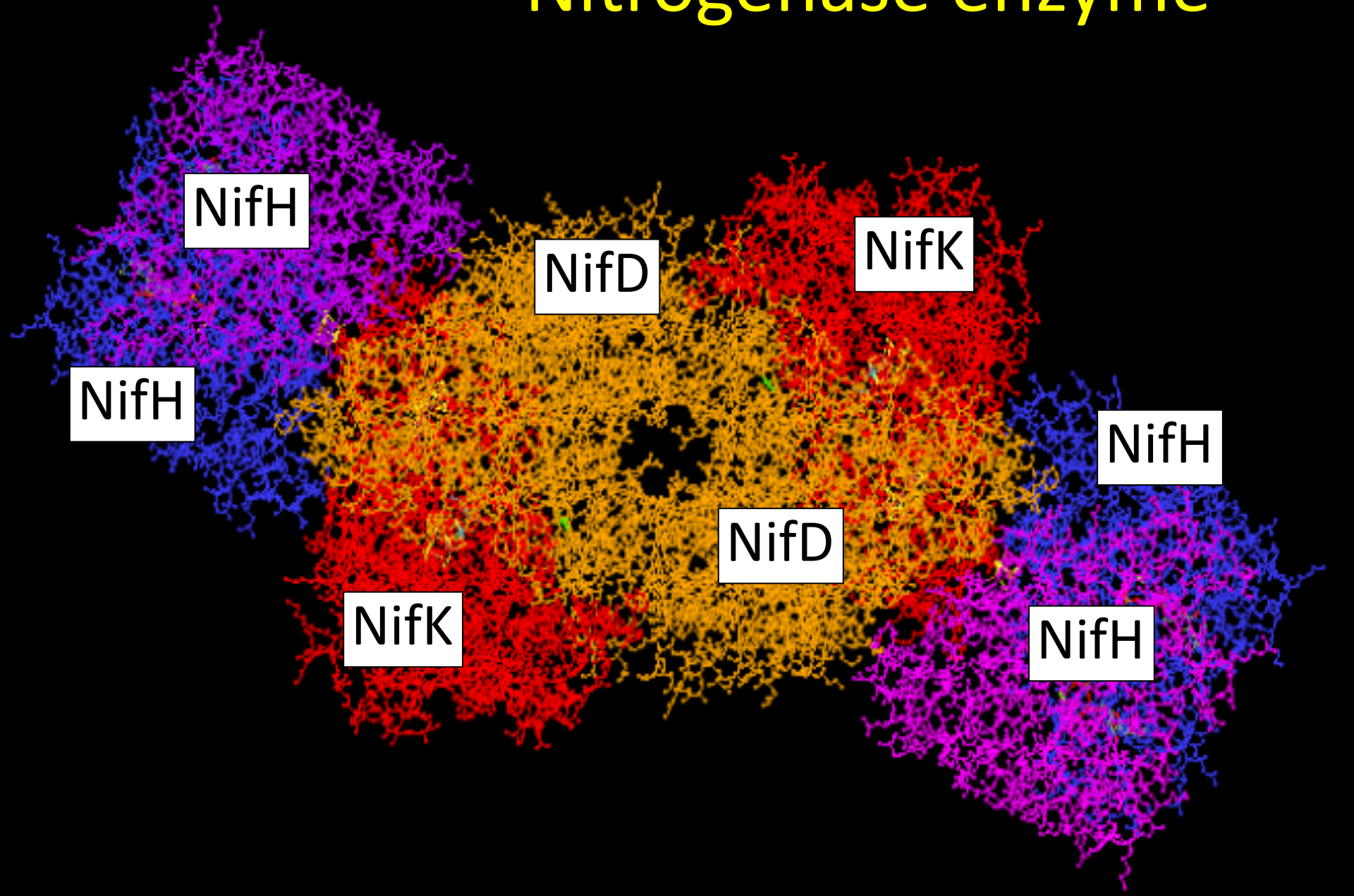


Disabled by oxygen

Ecology 101

- When ecologists say “nutrients”, we usually mean fixed nitrogen. E.g. California coastal nutrient upwelling
- In the open ocean, fixed nitrogen is usually the “limiting nutrient”:
  - Fixed nitrogen in the water enables abundant life
  - Including photosynthesizing plankton and bacteria that remove CO2 from the atmosphere and replace it with O2
- So marine ecologists care about nitrogen fixation

Nitrogenase enzyme

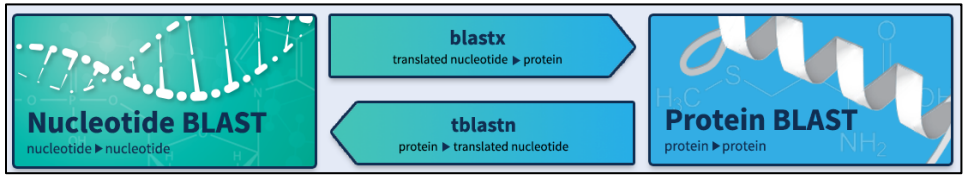


NifH is a “molecular proxy” for nitrogen fixation

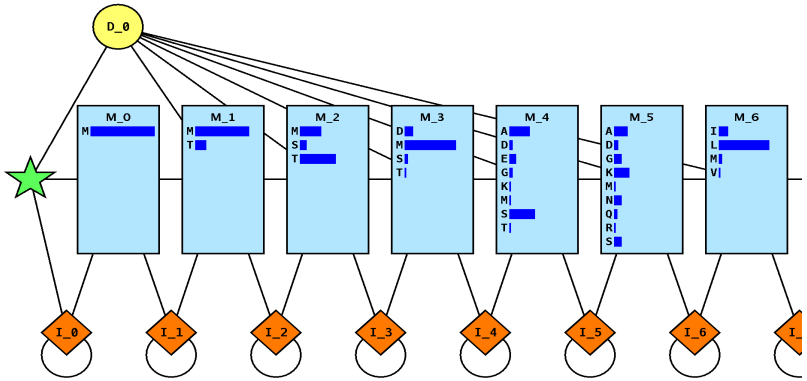
- The most conserved of the 3 nitrogenase proteins
- “Proxy”: when you want to study nitrogen fixation in a habitat, look for NifH
- Can’t directly study which nitrogen fixing species are present
  - 1 liter of seawater can contain 1 billion bacteria across 30,000 species, most of which have never been classified
  - Bacterial morphology (visible features) is too vague to be a reliable basis of identification
- Presence of the gene usually indicates nitrogen fixation
- Presence of the transcript *always* indicates nitrogen fixation
- Filter seawater, retain the bacteria, extract and sequence the DNA, hire a bioinformatician

How to identify a sequence

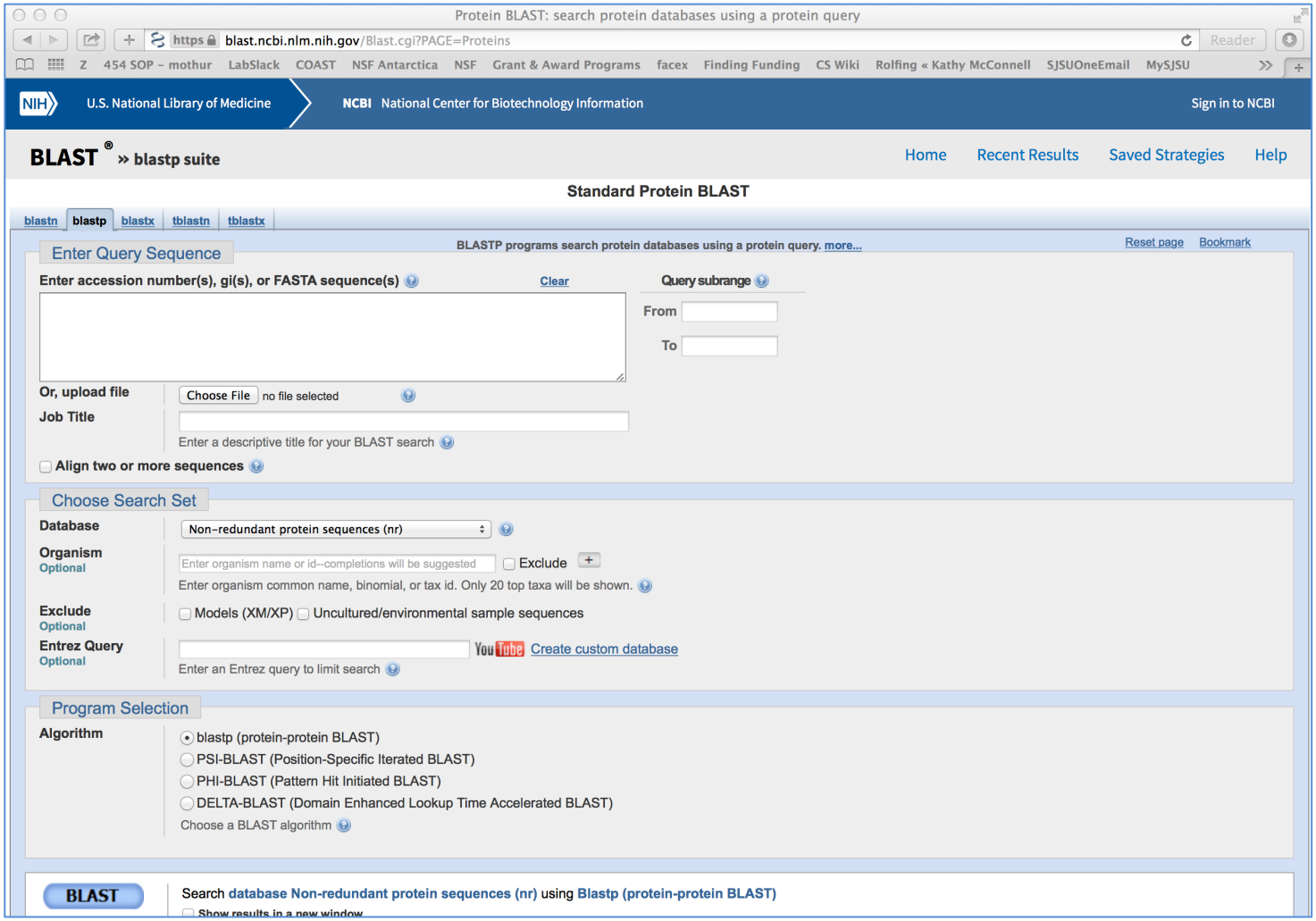
- BLAST (123A)



- HMM (123B)

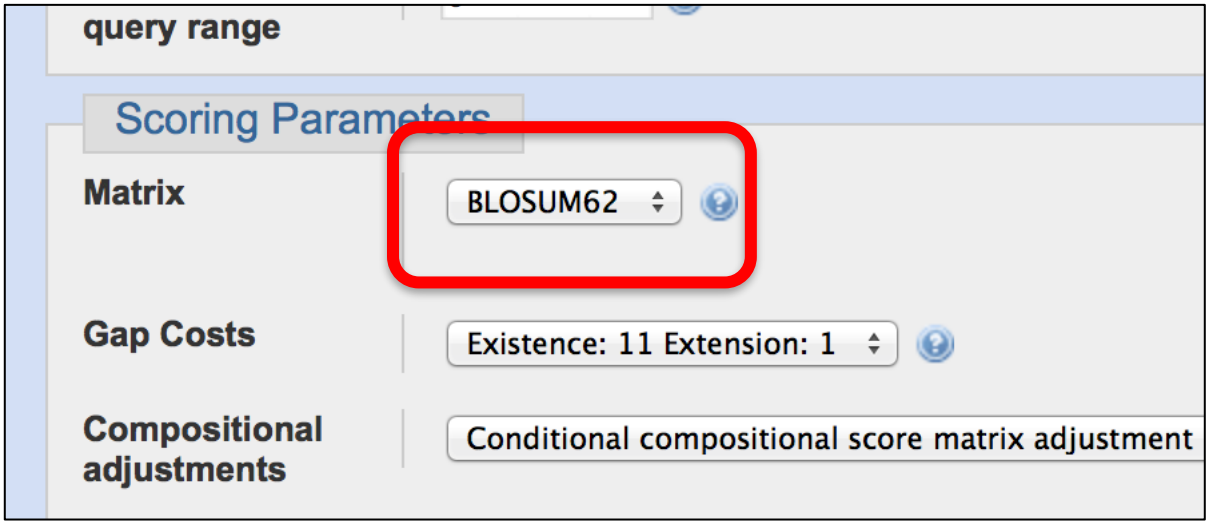


The trouble with BLAST …



BLASTp: Choose a matrix

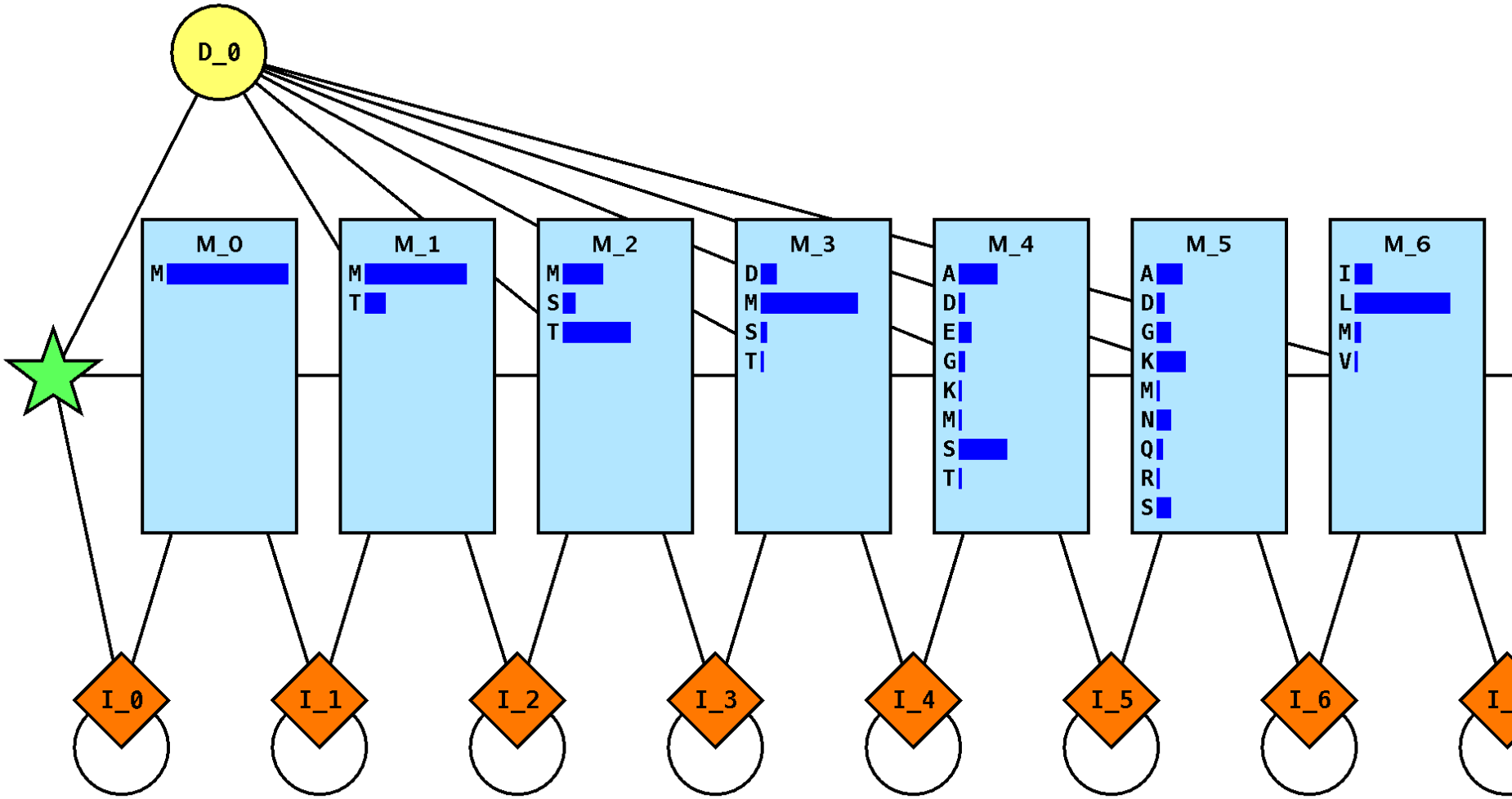
- PAM
  - 30, 70, 250
- BLOSUM
  - 45, 50, 62, 80, 90
- Matrix determines score of all alignment columns



But with protein identification, some columns are more diagnostic than others

- Diagnostic column: related sequences usually have the same amino acid in this column
- Not diagnostic column: related sequences might not have the same amino acid in this column
- BLAST treats all alignment columns alike
- So we use pHMMs
- Conserved regions: most states have only a few aas with significant emission probability
  - those aas are often chemically similar
- Variable regions: most states have many aas with significant emission probability

And that often works really well



But not always!

Step 1: Collect a training set

You will need some sequences that are known to be NifH. The FunGene data base provides list of various microbial gene sequences. For each gene, some of the sequences (the "seeds") have been submitted by experts and are believed with very high confidence to be what FunGene says they are.

1) Go to http://fungene.cme.msu.edu.
2) Under "Biogeochemical cycles", click "Nifh".
3) Near the top of the page, click on a small blue "download seeds" link. A file called nifH.seeds.txt will be downloaded to your computer. These trusted sequences will be the training set for your pHMM.

From your NifH HMM lab

The Fungene Method

- Subject matter experts submit lists of “seed” sequences for genes they know a lot about.
  - These are what the experts say they are, with very high probability
  - That’s why you used the nifH seeds in your last lab
- For each gene, build a pHMM.
- Evaluate members of GenBank’s nr protein database, report sequences with high scores.
- Analysis casts doubt on ~6% of Fungene’s NifH sequences … too many for comfort.
- That’s why you used the trusted seed sequences in lab, and not sequences retrieved by the Fungene HMM.

Why don’t HMMs work well in identifying *NifH*?

- Unknown
- I don’t think anyone has looked into it
- I have my suspicions … paralogs?

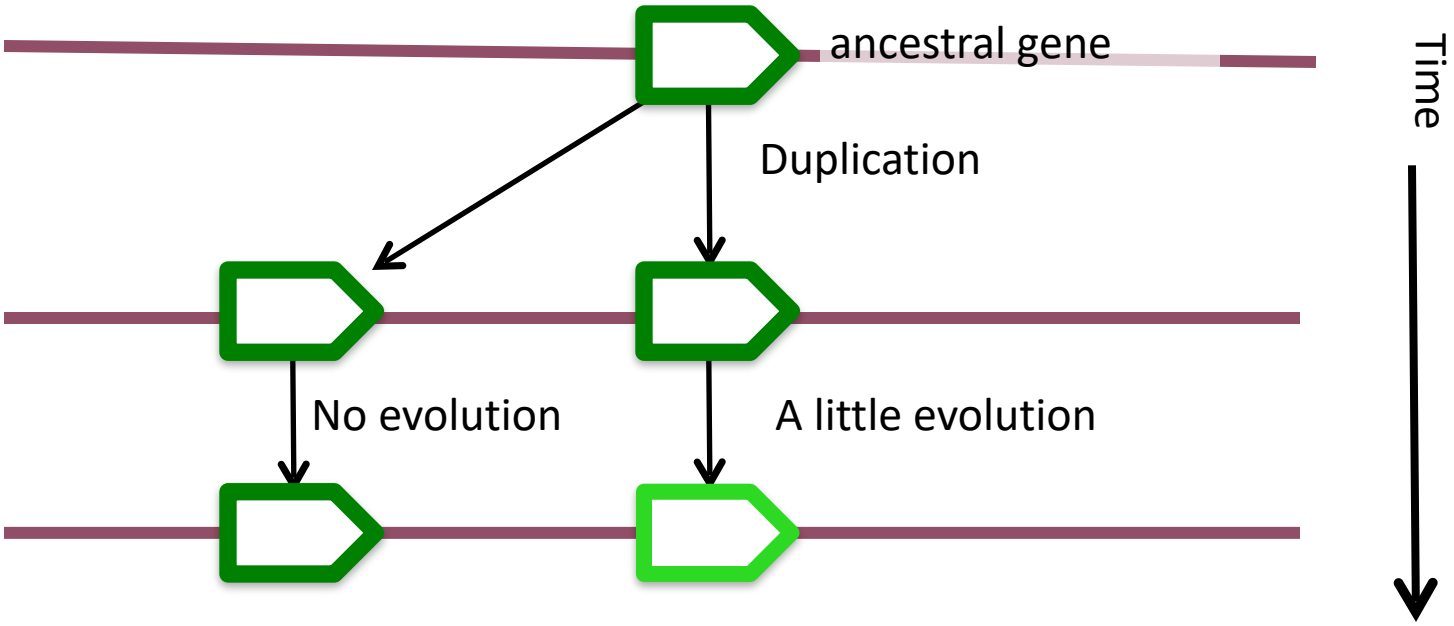
Kinds of Sequence Similarity

- Similarity
  - Coincidence
  - Homology
    - Orthology (orthologs): 2 different species
    - Paralogy (paralogs): 2 different genes in same genome

Origins of similarity

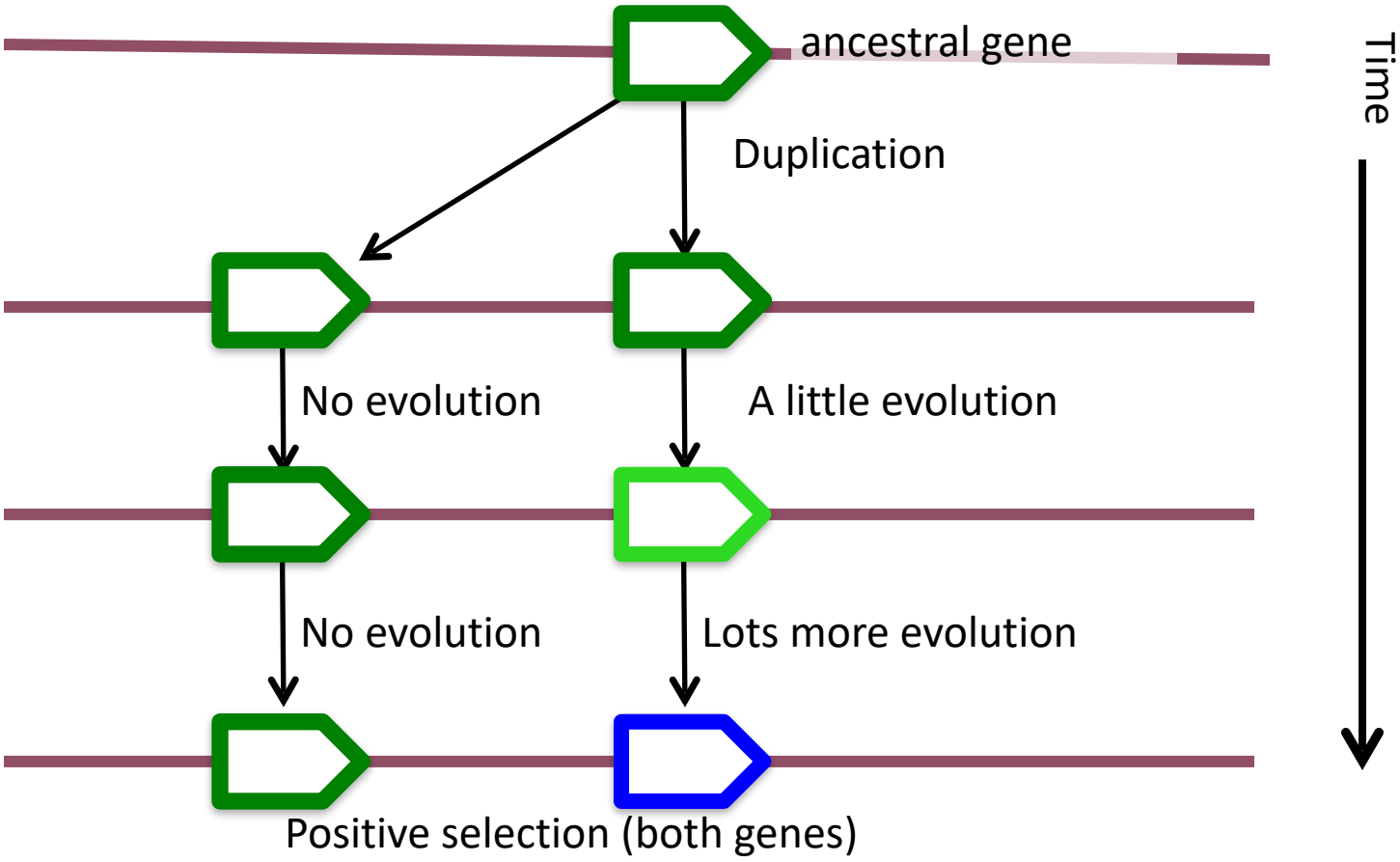
	Orthologs	Paralogs
Event type	Speciation	Gene duplication
Location of similar genes	2 different species	2 genes in same organism

Duplication

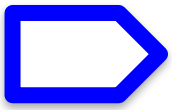


is stabilized by selection

Duplication: 1 possible destiny

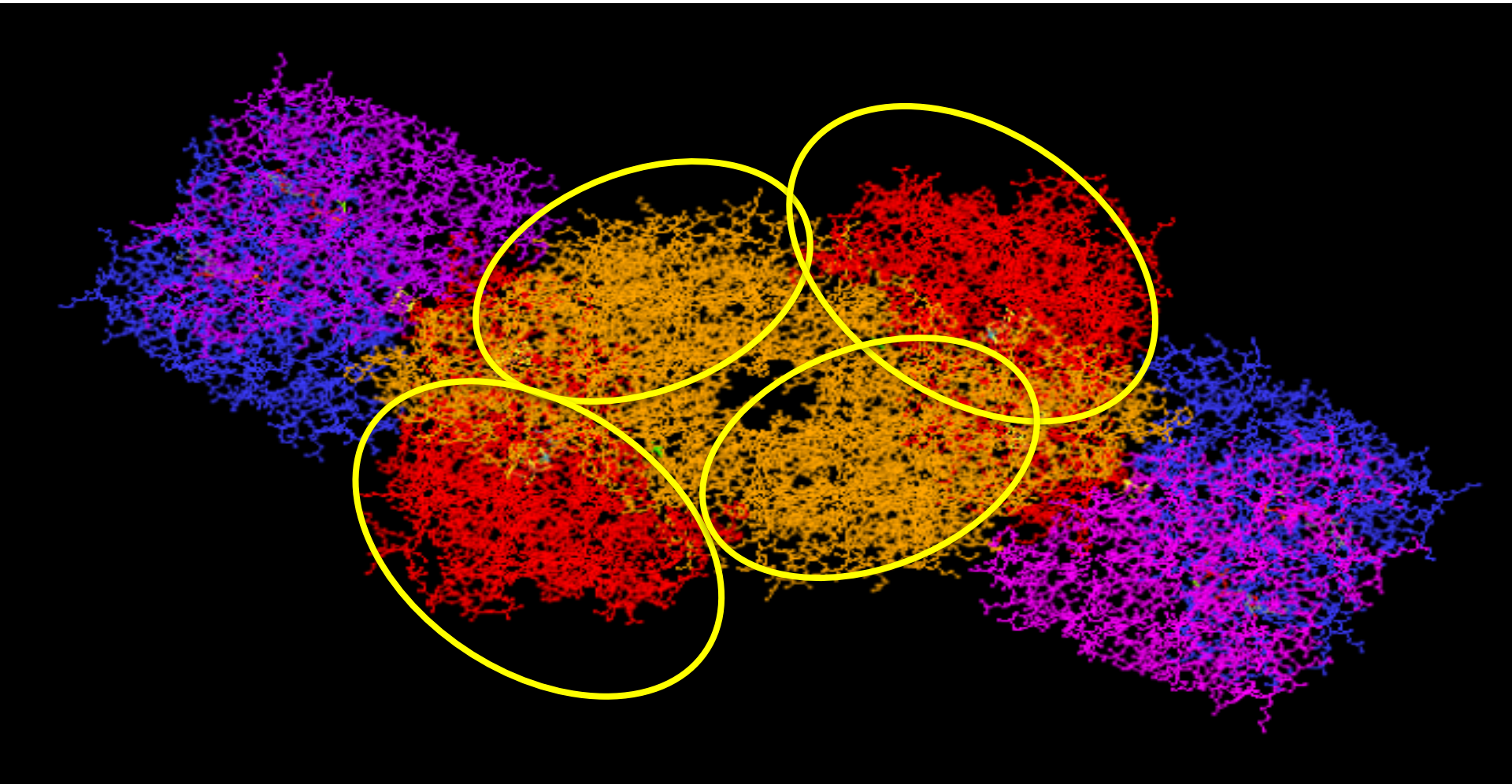


has evolved to provide a new & beneficial function

 and  are *paralogs*:

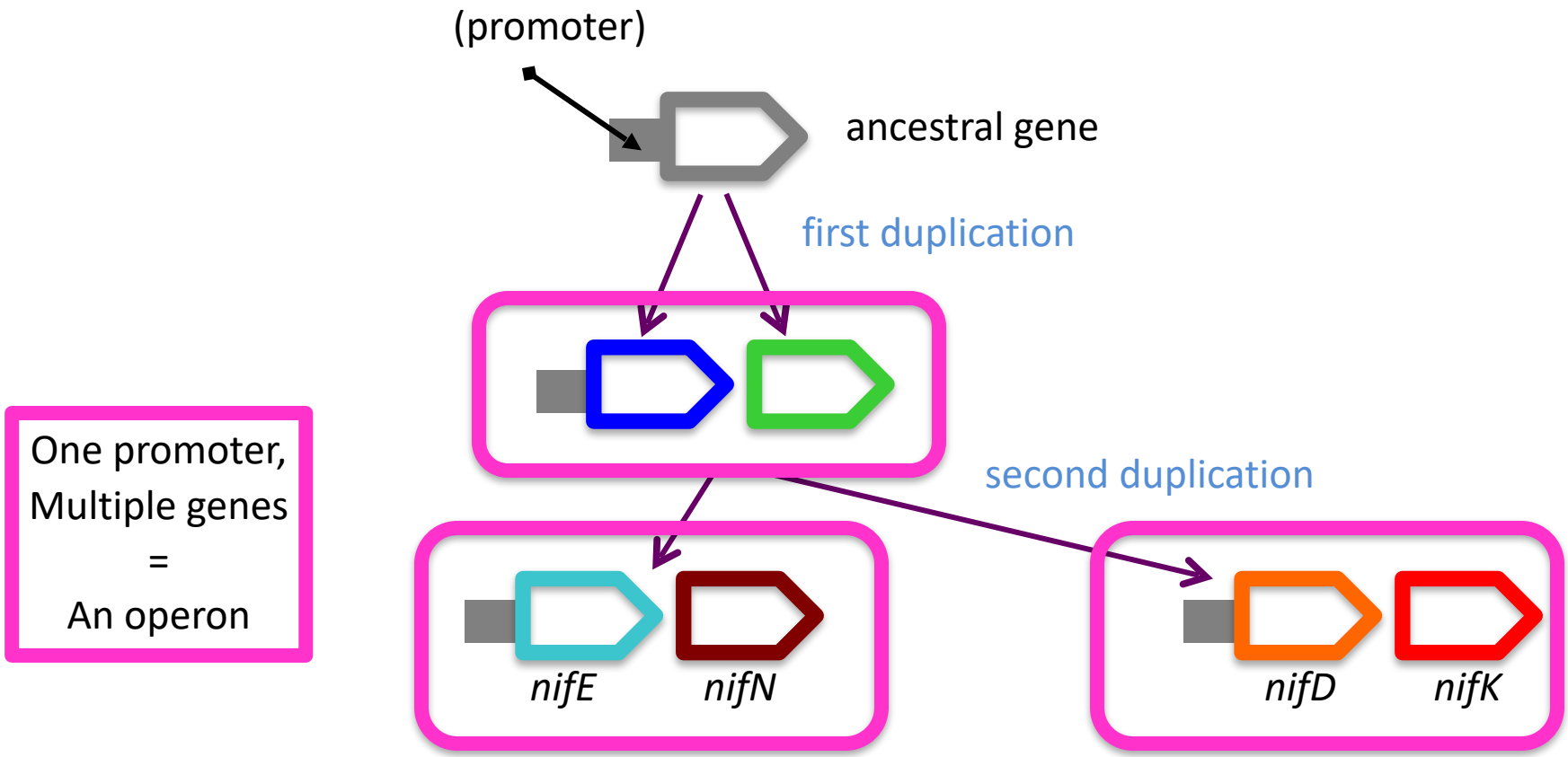
- Sequences are similar but different
- Different function, but maybe related
- Common genome
- Related by descent from a common ancestor, via a duplication event

Paralog Example: *nifE/N/D/K*



Paralog Example: *nifE/N/D/K*

- *nifD* and *nifK* are components of nitrogenase



Trouble at FunGene

- 2010: Fungene NifH database wasn't very good
  - Not sensitive (didn't report known NifH sequences in GenBank)
  - Not specific (reported known not-NifH sequences in GenBank)

Some Classifier Terminology

- Sensitive = able to positively identify what you're looking for
  - Whether or not you mistakenly positively identify other stuff also
- Specificity = ability to avoid incorrect identifications
  - Whether or not you mistakenly overlook correct identifications

Example: panning for gold

Panning for gold

- Panning = searching “database” of river sediment
- Finding gold = a positive identification
- False positive error: you think it’s gold, but it’s not
- False negative error: you throw gold back in the river
- Sensitive = find all the gold in the river
  - Low false negative rate
- Specific = reject dirt, pyrite, rocks, etc.
  - Low false positive rate

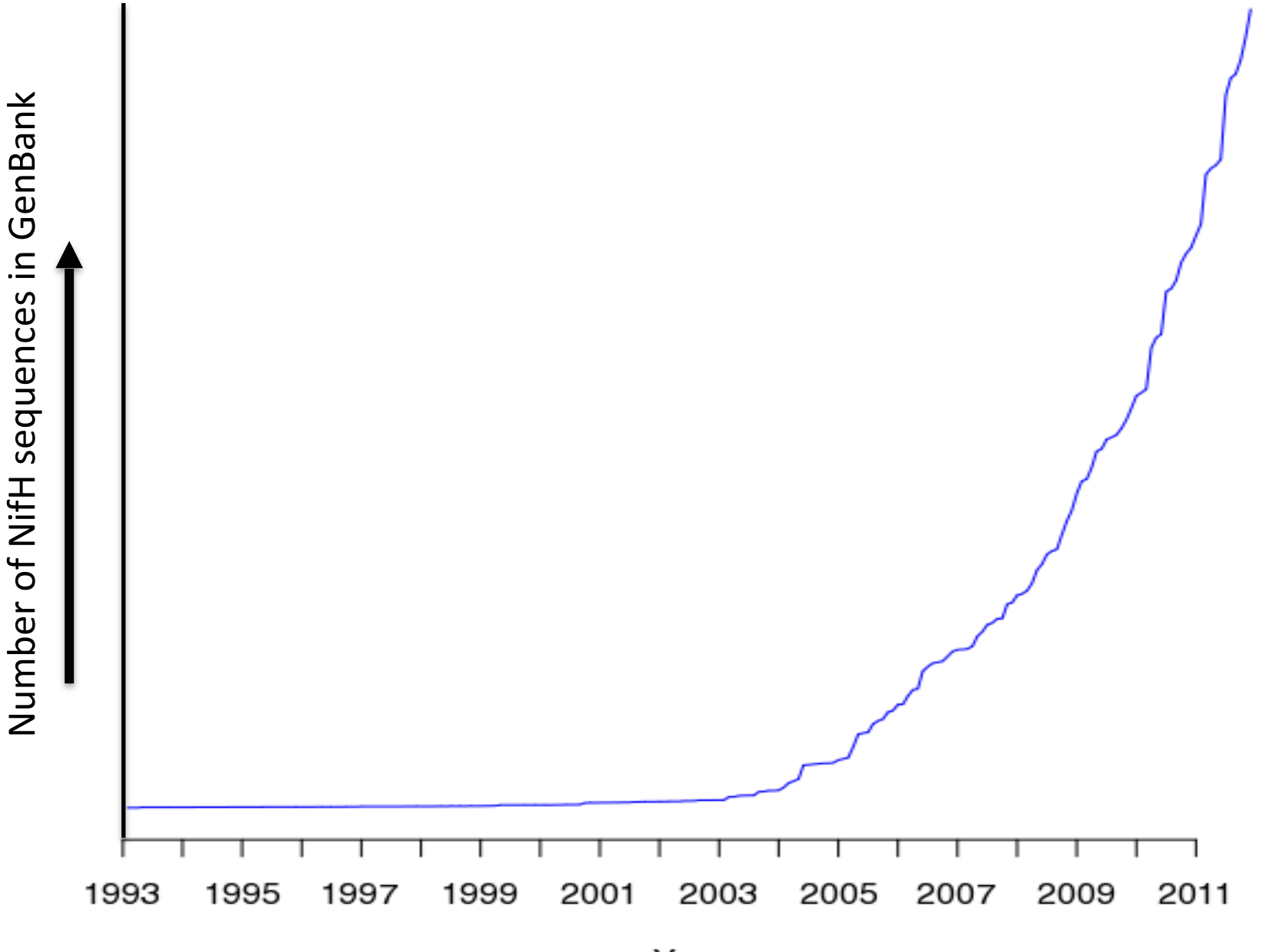
Data mining GenBank for NifH

- True positive: a sequence that you correctly accept as NifH
- True negative: a sequence that you correctly reject as not NifH
- False positive error: you think it's NifH, but it's not
- False negative error: you reject NifH
- Sensitive = find all the NifH in GenBank
  - Low false negative rate
- Specific = reject not-NifD, not-NifK, etc.
  - Low false positive rate

The nifH Catastrophe

- 2010 Fungene nifH database wasn't very good
  - Not sensitive (too many false negatives)
  - Not specific (too many false positives) ***??? Paralogs ???***
- Interest in NifH was increasing worldwide
  - Ecological importance
    - Nitrogen fixation provides most common limiting nutrient in many environments
    - The marine carbon pump
  - Economic importance
    - Agriculture, including soybean root nodules and rice roots
- The world needed a high quality database of NifH seqs

nifH Science: Victim of its own success



The NifH Catastrophe

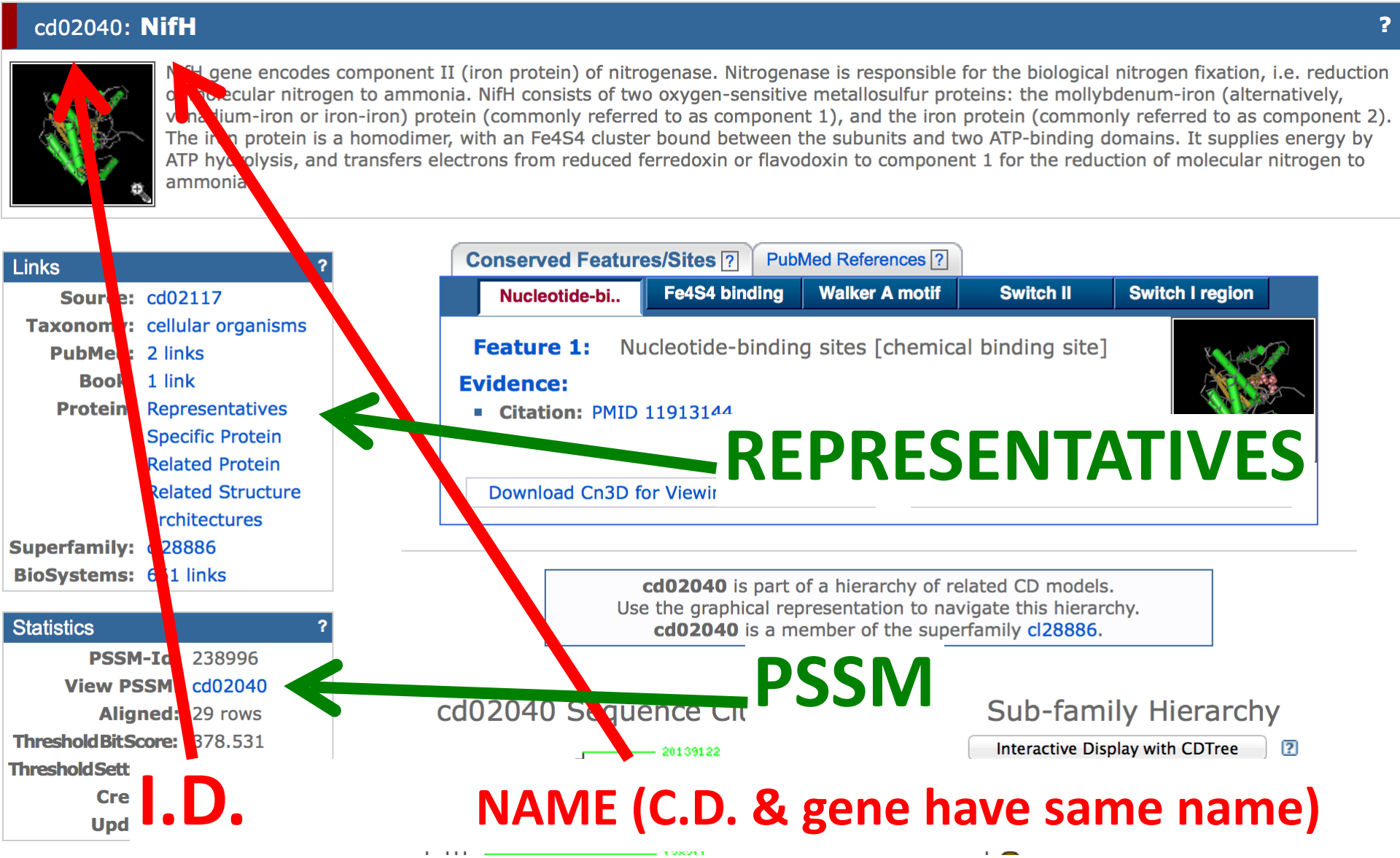
- Manual curation no longer tractable
  - Can't keep up with growth rate of nifH sequences
  - Only as good as the annotations
- Automated curation (e.g. Fungene HMMs) wasn't sensitive or specific enough
- We needed a better algorithm to drive automated curation
  - ARBitrator
  - Uses GenBank's conserved domain database, so let's cover that next

BLASTp uses the same scoring formula on all alignment columns

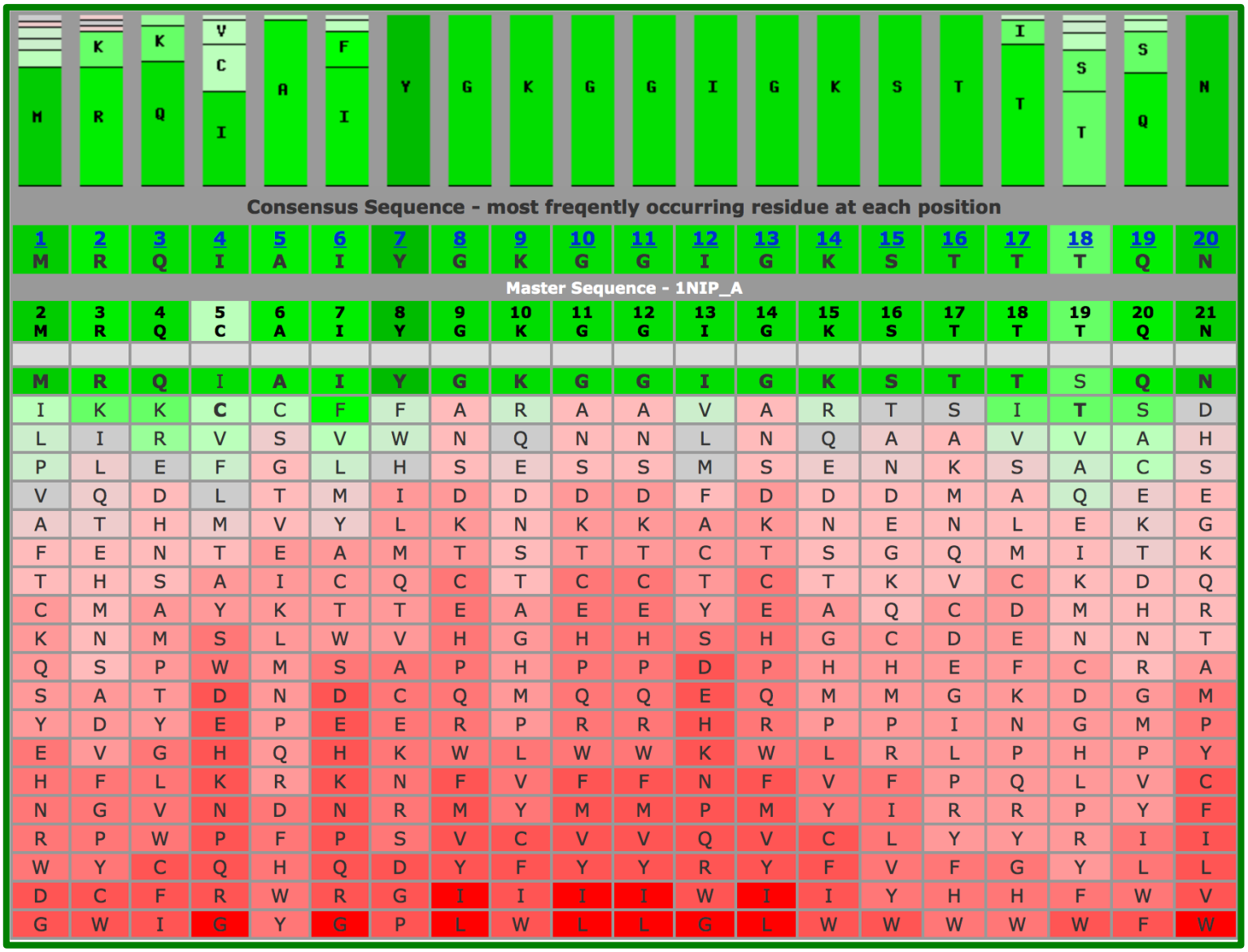
- PAM or BLOSUM
- What if it didn't?
- What if mutations in conserved columns were penalized more heavily than mutations in variable columns?
  - Least survivable mutations are penalized the worst

PSI-BLAST and RPSBLAST

- PSI-BLAST= Position-Specific Iterated BLAST
- RPSBLAST = Reverse PSI-BLAST
  - At NCBI: https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
  - Query is a protein sequence
  - Database is a collection of conserved domains, built from alignments of expert-curated representative sequences
  - Results look like BLASTn or BLASTp results
    - Scores, which we don't about
    - E-values, which we do care about
  - But subjects are conserved domains (CDs) rather than sequences

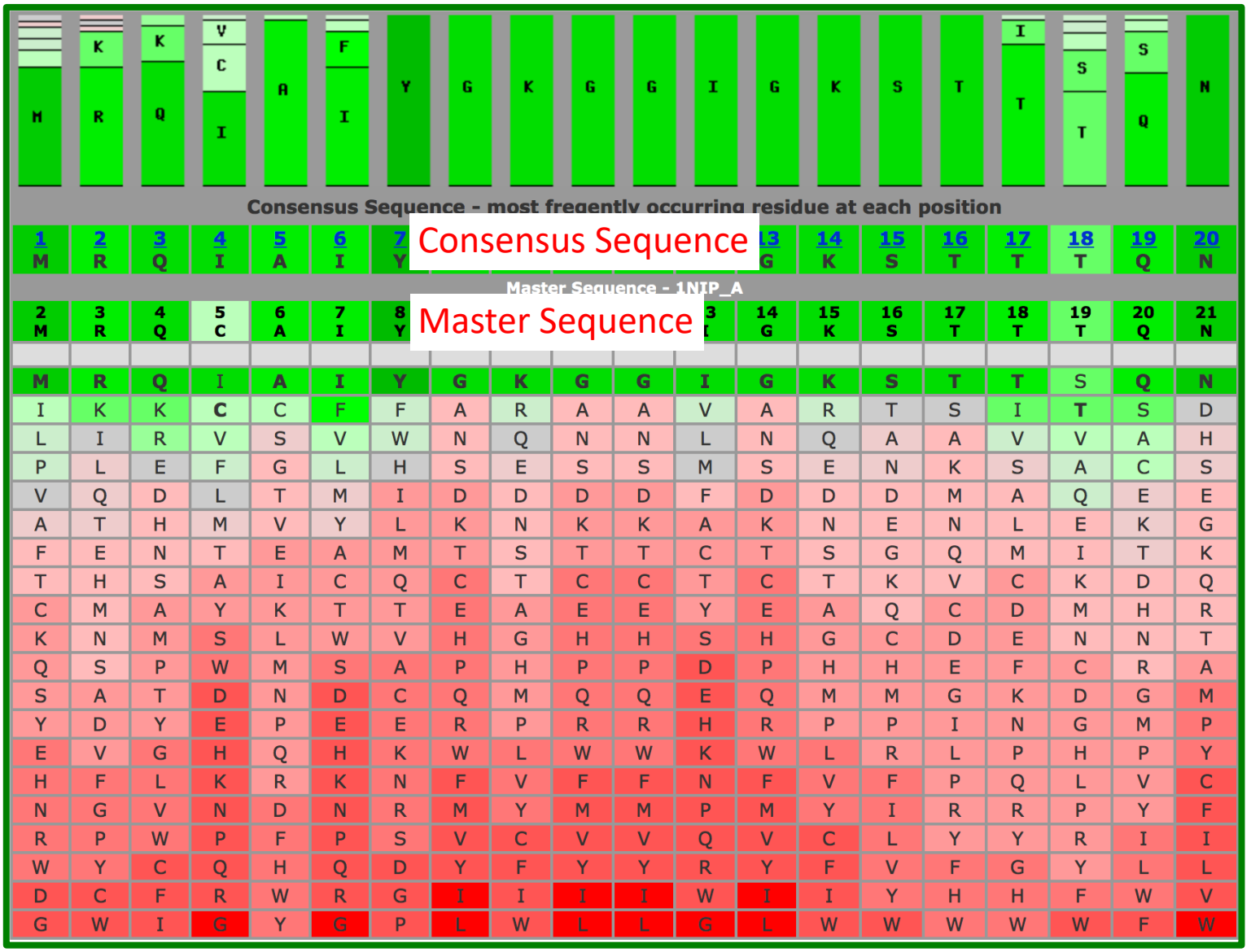

cd02040: NifH
?
NifH gene encodes component II (iron protein) of nitrogenase. Nitrogenase is responsible for the biological nitrogen fixation, i.e. reduction
ecular nitrogen to ammonia. NifH consists of two oxygen-sensitive metallosulfur proteins: the mollybdenum-iron (alternatively,
dium-iron or iron-iron) protein (commonly referred to as component 1), and the iron protein (commonly referred to as component 2).
The i n protein is a homodimer, with an Fe4S4 cluster bound between the subunits and two ATP-binding domains. It supplies energy by
ATP hy olysis, and transfers electrons from reduced ferredoxin or flavodoxin to component 1 for the reduction of molecular nitrogen to
ammonia
Links
?
Sour e: cd02117
Taxonom : cellular organisms
PubMe : 2 links
Book 1 link
Protein Representatives
Specific Protein
Related Protein
Related Structure
rchitectures
Superfamily: 28886
BioSystems: 1 links
Statistics
?
PSSM-Id 238996
View PSSM cd02040
Aligned: 29 rows
ThresholdBitScore: 378.531
ThresholdSett
Cre
Upd
Conserved Features/Sites
?
PubMed References
?
Nucleotide-bi..
Fe4S4 binding
Walker A motif
Switch II
Switch I region
Feature 1: Nucleotide-binding sites [chemical binding site]
Evidence:
Citation: PMID 11913144
Download Cn3D for Viewi
REPRESENTATIVES
cd02040 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.
cd02040 is a member of the superfamily cl28886.
PSSM
cd02040 Sequence Cl
20139122
Sub-family Hierarchy
Interactive Display with CDTree
?
I.D.
NAME (C.D. & gene have same name)

PSSM – Position-Specific Scoring Matrix



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

PSSM – Position-Specific Scoring Matrix

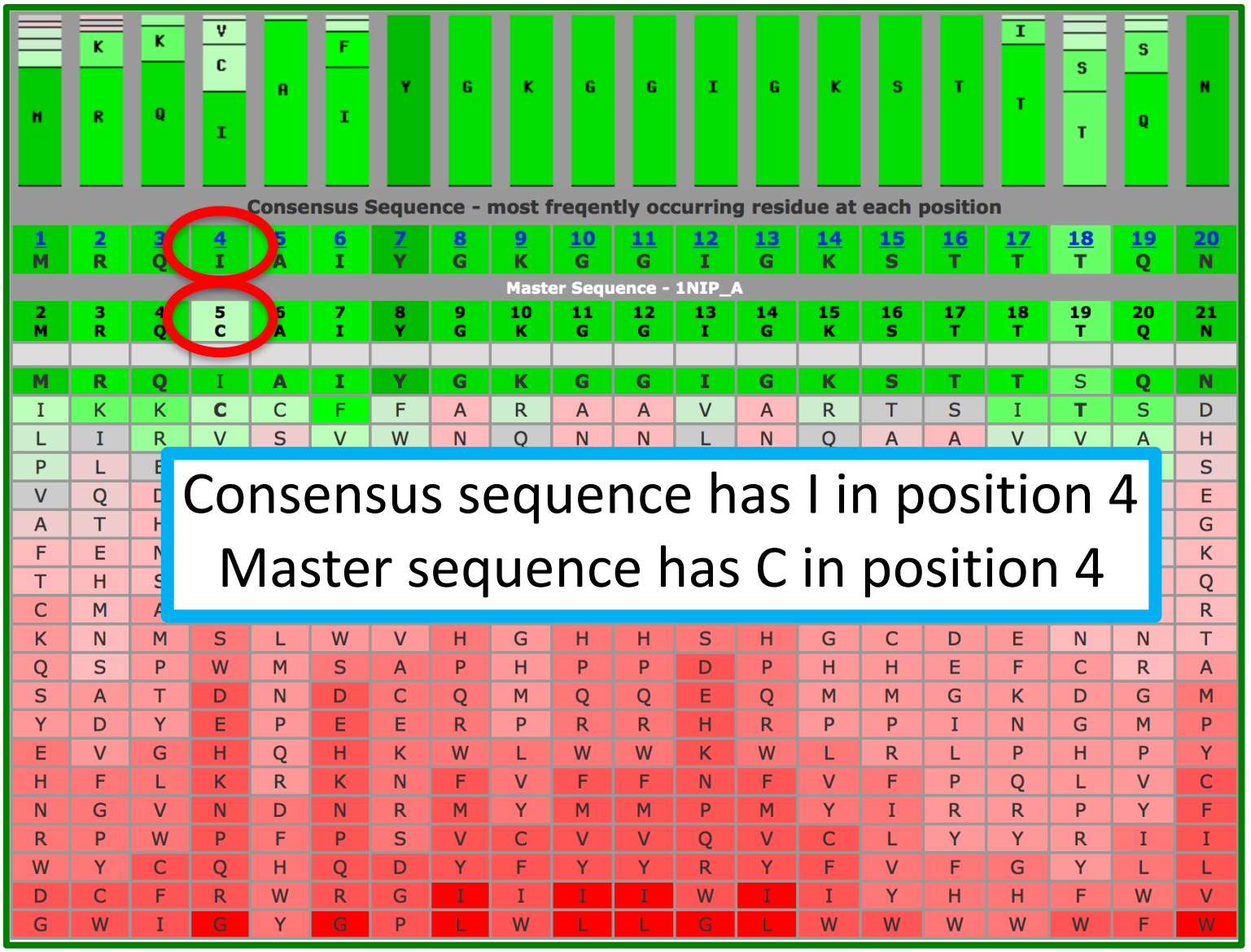


1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
M	R	Q	I	A	I	Y	G	K	G	G	I	G	K	S	T	T	T	Q	N
2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
M	R	Q	C	A	I	Y	G	K	G	G	I	G	K	S	T	T	T	Q	N
M	R	Q	I	A	I	Y	G	K	G	G	I	G	K	S	T	T	S	Q	N
I	K	K	C	C	F	F	A	R	A	A	V	A	R	T	S	I	T	S	D
L	I	R	V	S	V	W	N	Q	N	N	L	N	Q	A	A	V	V	A	H
P	L	E	F	G	L	H	S	E	S	S	M	S	E	N	K	S	A	C	S
V	Q	D	L	T	M	I	D	D	D	D	F	D	D	D	M	A	Q	E	E
A	T	H	M	V	Y	L	K	N	K	K	A	K	N	E	N	L	E	K	G
F	E	N	T	E	A	M	T	S	T	T	C	T	S	G	Q	M	I	T	K
T	H	S	A	I	C	Q	C	T	C	C	T	C	T	K	V	C	K	D	Q
C	M	A	Y	K	T	T	E	A	E	E	Y	E	A	Q	C	D	M	H	R
K	N	M	S	L	W	V	H	G	H	H	S	H	G	C	D	E	N	N	T
Q	S	P	W	M	S	A	P	H	P	P	D	P	H	H	E	F	C	R	A
S	A	T	D	N	D	C	Q	M	Q	Q	E	Q	M	M	G	K	D	G	M
Y	D	Y	E	P	E	E	R	P	R	R	H	R	P	P	I	N	G	M	P
E	V	G	H	Q	H	K	W	L	W	W	K	W	L	R	L	P	H	P	Y
H	F	L	K	R	K	N	F	V	F	F	N	F	V	F	P	Q	L	V	C
N	G	V	N	D	N	R	M	Y	M	M	P	M	Y	I	R	R	P	Y	F
R	P	W	P	F	P	S	V	C	V	V	Q	V	C	L	Y	Y	R	I	I
W	Y	C	Q	H	Q	D	Y	F	Y	Y	R	Y	F	V	F	G	Y	L	L
D	C	F	R	W	R	G	I	I	I	I	W	I	I	Y	H	H	F	W	V
G	W	I	G	Y	G	P	L	W	L	L	G	L	W	W	W	W	W	F	W

Consensus and Master Sequences

- Consensus sequence: position x = most frequent amino acid at position x.
- Master sequence: the naturally occurring sequence most similar to the consensus sequence.

PSSM – Position-Specific Scoring Matrix



Conserved domain families

- Combinations of multiple related conserved domains
- Ex: “Fer4_NifH” family:
  - Both genes involve binding to iron
  - The domains are similar
- Superfamilies
  - Combinations of multiple related families

Gradebook – SP18: CS-46B Se | 195B grades – philip.heller@s | NCBI CDD Conserved Protein | Philip

Secure https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?hslf=1&uid=238941&#seqhrch

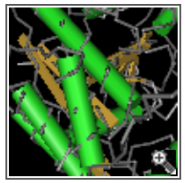
Apps | R Histograms | wiki | MySJSU | IMG Submission | Sample download... | Shoreline Wind | Online Survey | Bu... | III

NCBI

Conserved Protein Domain Family
Fer4_NifH

HOME | SEARCH | SITE MAP | Entrez | CDD | Structure | Protein | Help

cd01983: Fer4_NifH



The Fer4_NifH superfamily contains a variety of proteins which share a common ATP-binding domain. Functionally, proteins in this superfamily use the energy from hydrolysis of NTP to transfer electron or ion.

Links

Source: Pfam
Taxonomy: root
PubMed: 10 links
Protein: Representatives
Specific Protein
Related Protein
Related Structure
Architectures
Superfamily: cl28886

Statistics

PSSM-Id: 238941
View PSSM: cd01983
Aligned: 942 rows
Threshold Bit Score: 27.7526
Threshold Setting Gi: 10175642
Created: 12-Dec-2003
Updated: 17-Jan-2013

PubMed References

- Molybdenum nitrogenases: a crystallographic and mechanistic view. *Met Ions Biol Syst 2002; 39:75-119*
- Purification and characterization of membrane-associated CooC protein and its functional role in the insertion of nickel into carbon monoxide dehydrogenase from Rhodospirillum rubrum. *J. Biol. Chem. 2001 Oct 19; 276(42):38602-38609*
- fleN, a gene that regulates flagellar number in Pseudomonas aeruginosa. *J. Bacteriol. 2000 Jan; 182(2):357-364*
- The Rhodobacter capsulatus chlorin reductase-encoding locus, bchA, consists of three genes, bchX, bchY, and bchZ. *J. Bacteriol. 1993 Apr; 175(8):2407-2413*
- Reconstitution of light-independent protochlorophyllide reductase from purified bchl and BchN-BchB subunits. In vitro confirmation of nitrogenase-like features of a bacteriochlorophyll biosynthesis enzyme. *J. Biol. Chem. 2000 Aug 4; 275(31):23583-23588*
- The three-dimensional structure of septum site-determining protein MinD from Pyrococcus horikoshii OT3 in complex with Mg-ADP. *Structure 2001 Sep; 9(9):817-826*
- Crystal structure of the bacterial cell division regulator MinD. *FEBS Lett. 2001 Mar 9; 492(1-2):160-165*
- Structure-function relationships in an anion-translocating ATPase. *Biochem. Soc. Trans. 2000; 28(4):520-526*
- Conformational changes in four regions of the Escherichia coli ArsA ATPase link ATP hydrolysis to ion translocation. *J. Biol. Chem. 2001 Aug 10; 276(32):30414-30422*
- Structure of the ArsA ATPase: the catalytic subunit of a heavy metal resistance pump. *EMBO J. 2000 Sep 1;*

Structure View

Program: Cn3D

Drawing: All Atoms

Aligned Rows: up to 10

Download Cn3D

Hierarchy ?

Interactive Display

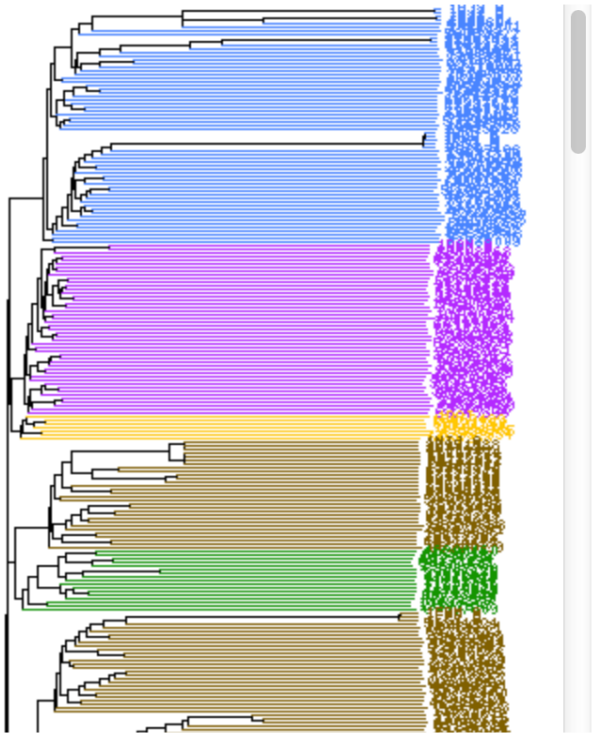
Download CDTree

LinkOut - more resources

cd01983 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.
cd01983 is a member of the superfamily cl28886.

cd01983 Sequence Cluster

Zoom In



Sub-family Hierarchy

Interactive Display with CDTree

- cd01983 Fer4_NifH
  - cd00477 FTHFS
  - cd02034 CooC
  - cd02035 ArsA
  - cd02036 MinD
  - cd02037 MRP-like
  - cd02038 FleN-like
  - cd02042 ParA
  - cd02117 NifH_like
    - cd02032 Bchl_like
    - cd02033 BchX
    - cd02040 NifH
  - cd03108 AdSS
  - cd03109 DTBS
  - cd03110 Fer4_NifH_child
  - cd03111 CpaE_like
  - cd03112 CobW_like
  - cd03113 CTGs
  - cd03114 ArgK-like

Sequence Alignment include consensus sequence ?

Reformat | Format: Hypertext | Row Display: up to 10 | Color Bits: 2.0 bit | Type Selection: the most diverse members

10 20 30 40 50 60 70 80

Back to the problem …
Mining GenBank for some gene of interest: What didn't work

- Pick some representative sequences
- Use them as BLASTp queries
- Report hits with E-value <= some threshold
- Often fooled by paralogs
  - Sensitive but not specific
- True in general, especially true for NifH

Blasting representatives → Sensitive, not specific

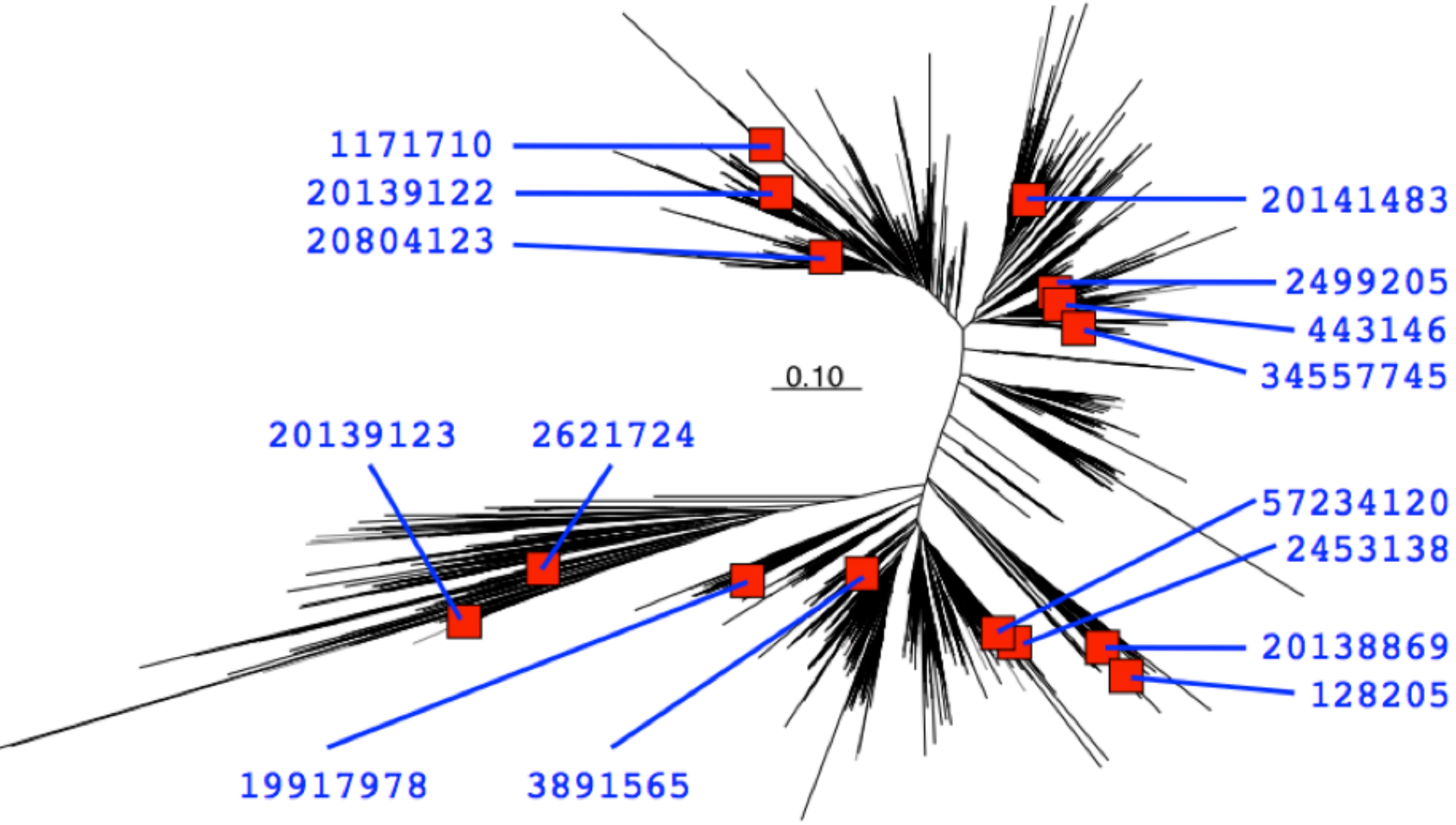
- Lots of false positives, few false negatives.
- Results contain all the gold you want, mixed up with a lot of dirt.
- Call this Phase 1.
- Invent a Phase 2 that filters out the dirt.
  - 2009: Most promising approach involves RPSBLAST and conserved domains.
  - Annotations are not reliable, classifier should only compute about sequences.
- And that's when I knocked on the door …

Training Sets and Representatives

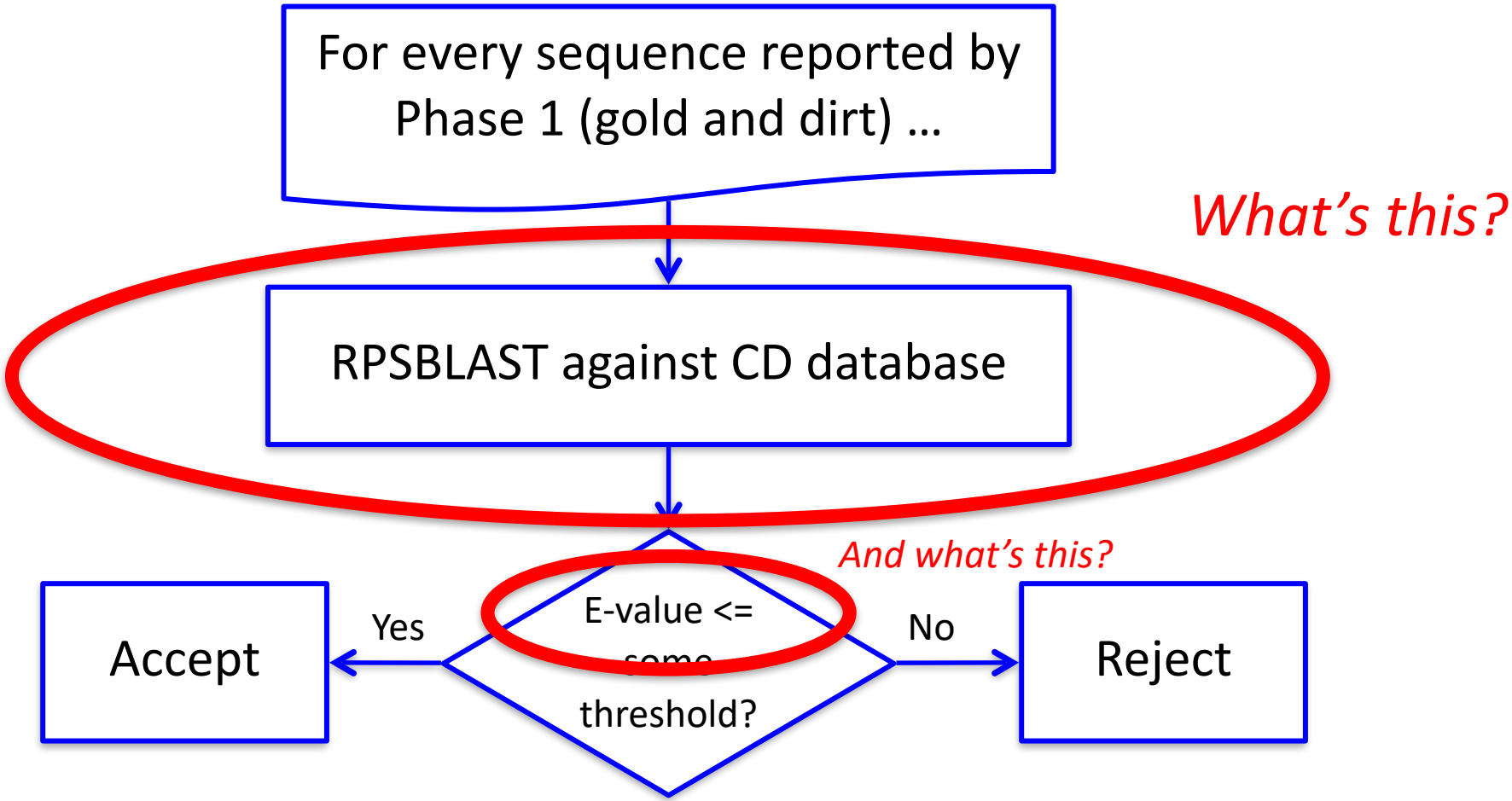
- Positive training set: ~15K sequences, believed with high confidence to be NifH.
- Negative training set: ~750 sequences, with some similarity to NifH, believed with high confidence *not* to be NifH.
- Representatives: 15 members of the positive training set, chosen to represent the diversity of NifH

Representatives

- BLAST queries for phase 1



Phase 2: Trial and error with the training sets (and error and more error)



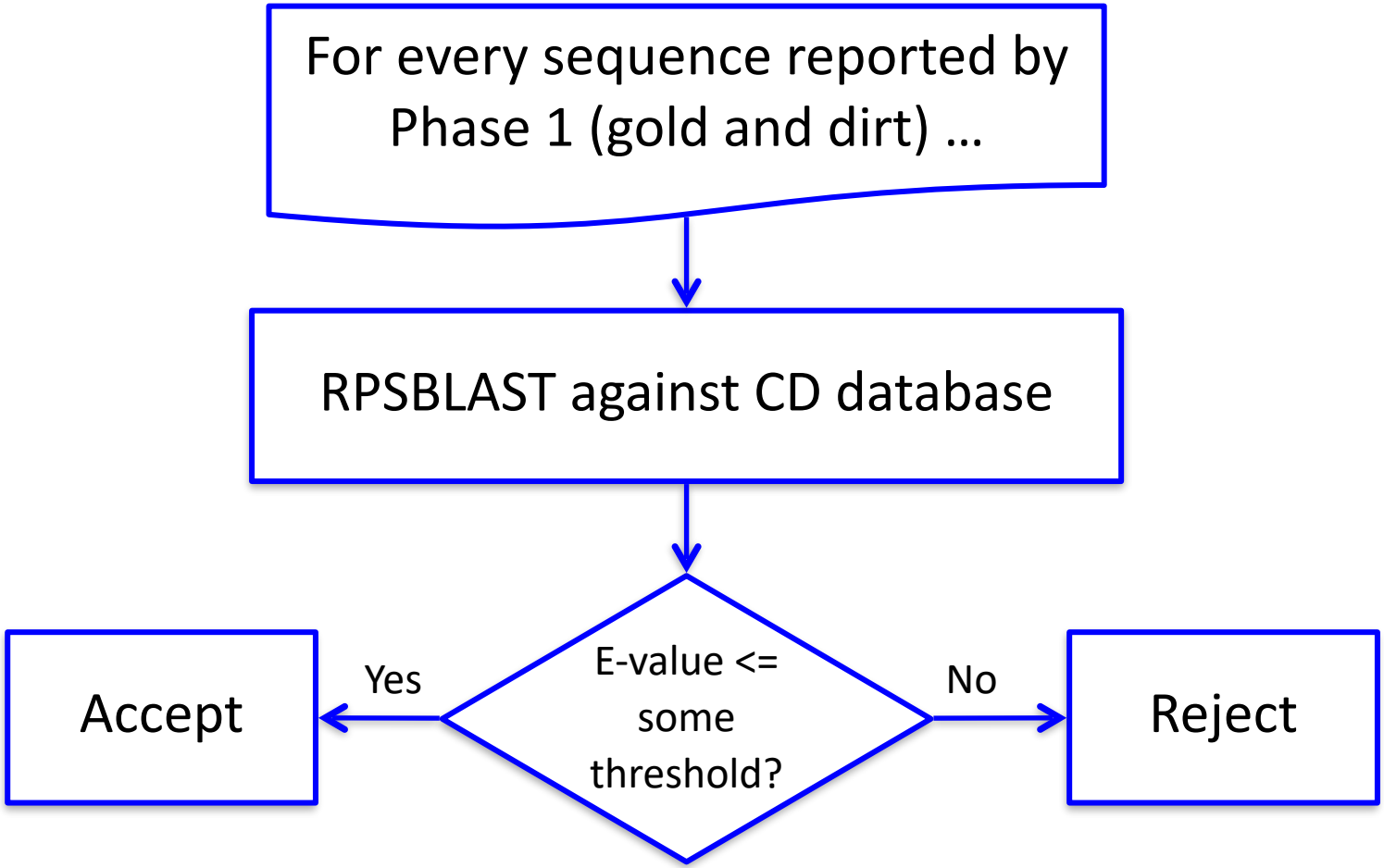
What's RPSBlast?

- Another form of blast, like blastn, blastp, blastx, etc.
- Aligns query sequence (almost always aas) against every member of a database of Conserved Domains.
- Different scoring for each column.
- Reports hits with score, E-value, more
  - E-value is most useful
- Hits can be to Conserved Domains, families, superfamilies

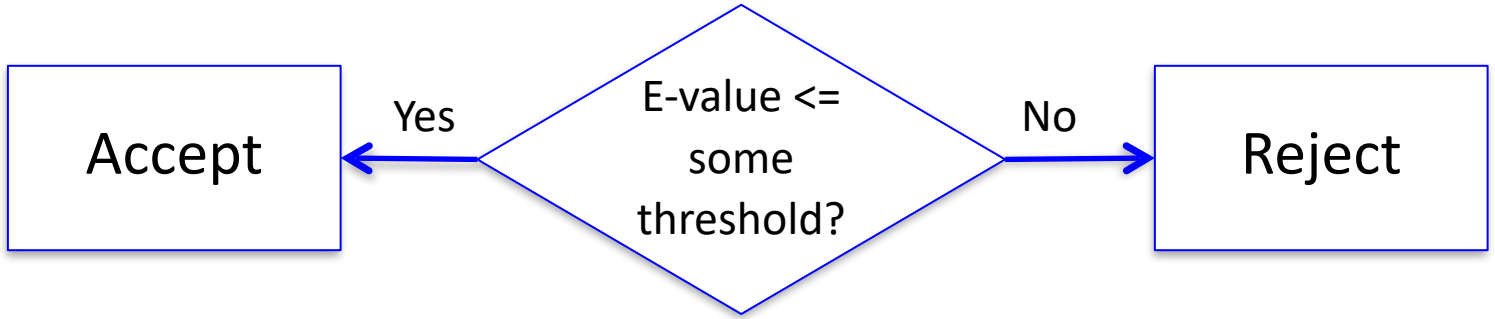
E-value definition

- Given:
  - A query sequence of length L
  - Which you blast against some database DB
  - And you get a hit with score S …
- The E-value of your hit is the probability of:
  - Blasting a random query of length L
  - Against database DB
  - With score >= S

Phase 2: Trial and error with the training sets (and error and more error)



Phase 2: The catastrophe continues



- There is no good threshold
- Too low → algorithm isn't sensitive enough
  - False negatives = erroneous rejections of *NifH*
- Too high → algorithm isn't specific enough
  - False positives = erroneous acceptance of not *NifH*

Now what?

When you can't see your way forward, get more information.

- CONSERVED DOMAIN similarity (e-value) doesn't classify well enough …
- Is there some other trait, besides rpsBLAST score or E-value, that differentiates *NifH* from not-*NifH*?

What the true positives and false negatives had in common

- Best conserved domain hit was the nifH conserved domain (as expected).
- E-value of best hit to some conserved domain *other than nifH* was at least 10x worse than e-value of best hit.
- Example:
  - For a certain *NifH* sequence …
  - Best hit is NifH CD, e-value = 2.0e-48
  - 2nd-best hit is iron-sulfur cluster binding protein conserved domain, e-value = 2.0e-47 (10x worse)

Simplify the numbers: Superiority

- In phase 2, reject any sequence whose best rpsBLAST subject isn't the NifH CD
- Superiority only refers to sequences whose best rpsBLAST subject *is* the NifH CD
- Define Superiority =

log10(E-value of 2nd-best hit)

Minus log10(E-value of best hit)

= by how many orders of magnitude (OOMs) is the NifH hit better than the next best hit?

Superiority Example 1: In phase 2, rpsBLAST a sequence and get the following hits

Subject conserved domain	E-value of hit Subject C.D.
NifH conserved domain	1.0E-80
[illegible] domain	[illegible]
[illegible] domain	[illegible]

Superiority = log10(1.0E-60) – log10(1.0E-80)

= -60 – -80 = 20

Hit to NifH conserved domain is 20 O-O-Ms better than 2nd-best hit

Superiority Example 2: In phase 2, rpsBLAST a sequence and get the following hits

Subject conserved domain	E-value of hit Subject C.D.
NifH conserved domain	2.7E-35
NifH conserved domain	1.0E-32
Cellulose biosynthesis domain	2.7E-20

Superiority = log10(2.7E-20) – log10(2.7E-35)
= -20 – -35 = 15

Ignore the worse NifH conserved domain hit

Superiority Example 3: In phase 2, rpsBLAST a sequence and get the following hits

Subject conserved domain	E-value of hit Subject C.D.
NifH conserved domain	8.9E-61
Fer4_NifH family	1.0E-59
Cellulose biosynthesis domain	8.9E-55

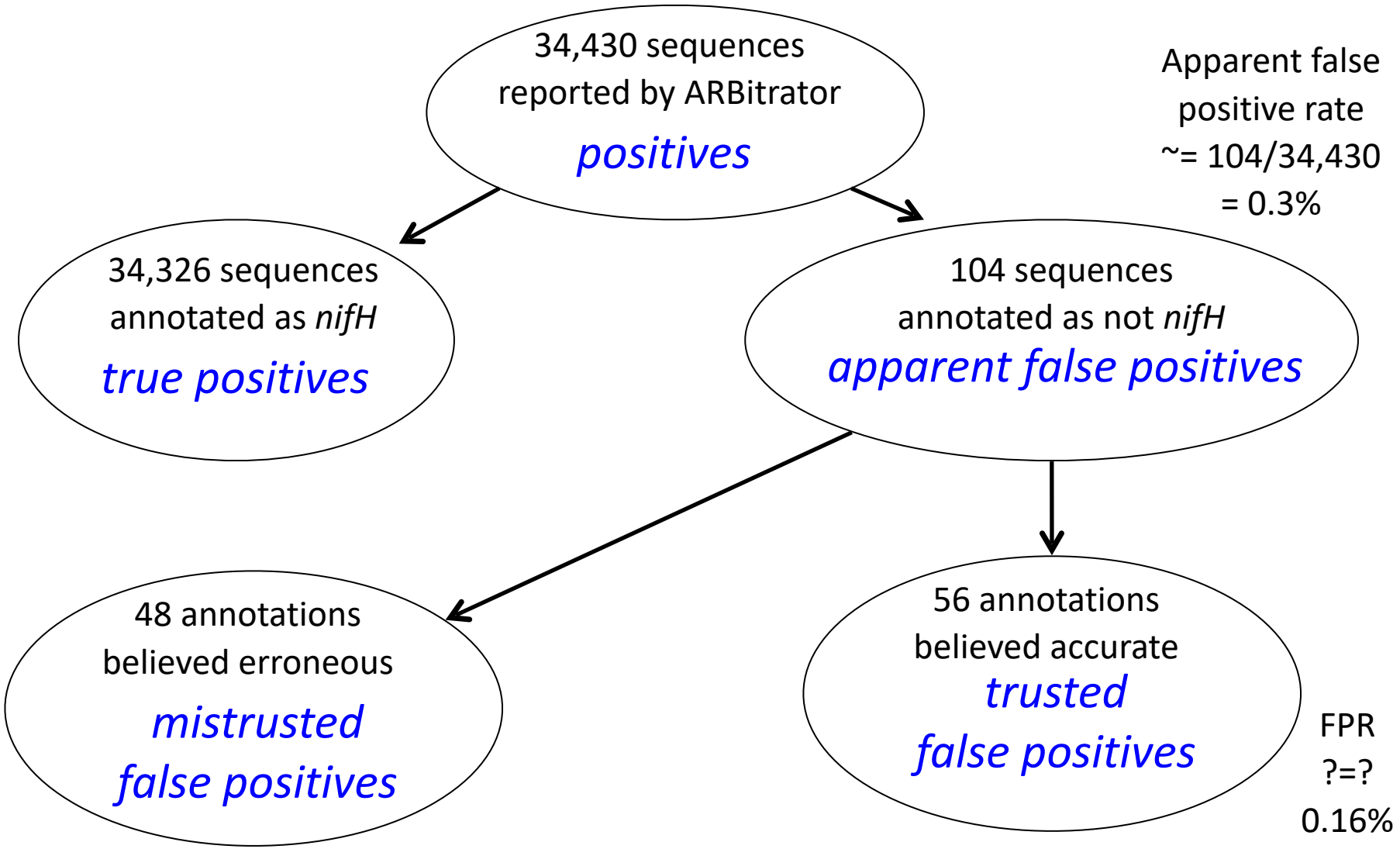
Superiority = log10(8.9E-55) – log10(8.9E-61)
= -55 – -61 = 6

Ignore the family hit

Is there a good Superiority threshold that correctly classifies the positive and negative training sets?

- Yes!
- Threshold = 1
- Tiny error rates

False-Positive Rate



Example of a mistrusted false positive

- GI = CAK43123.1
- Quality is >> threshold
- Annotation: light-independent protochlorophyllide reductase iron-sulfur ATP-binding protein, partial
- AUTHORS: Villadas, P.J. & al.
- TITLE: The rhizosphere bacterial community of the leguminous trees Eperua falcata and Dicorynia guianensis from tropical rainforest in French Guiana

Blast this seq against nr, look at top 20 hits

The sequence hitting itself

Description	Max score	Total score	Query cover	E value	Max ident
light-independent protochlorophyllide reductase iron-sulfur ATP-binding protein [uncultured proteobacterium]	207	207	100%	5e-67	100%
nitrogenase reductase [uncultured bacterium]	204	204	100%	1e-65	98%
light-independent protochlorophyllide reductase iron-sulfur ATP-binding protein [uncultured proteobacterium] >emb|CA...|43	203	203	99%	2e-65	99%
nitrogenase reductase [uncultured bacterium] >gb|ADI6...589.1| nitrogenase reductase [uncultured bacterium] >gb|ADI6165	204	204	100%	2e-65	98%
nitrogenase reductase [uncultured bacterium]	204	204	100%	2e-65	98%
nitrogenase reductase [uncultured bacterium] >gb|AD...61625.1| nitrogenase reductase [uncultured bacterium] >gb|ADI6164	204	204	100%	2e-65	98%
nitrogenase reductase [uncultured bacterium] >gb|A...I61578.1| nitrogenase reductase [uncultured bacterium] >gb|ADI6160	204	204	100%	2e-65	98%
nitrogenase reductase [uncultured bacterium] >gb|A...I61712.1| nitrogenase reductase [uncultured bacterium]	204	204	100%	2e-65	98%
nitrogenase reductase [uncultured bacterium] >gb|ADI61714.1| nitrogenase reductase [uncultured bacterium]	204	204	100%	2e-65	98%

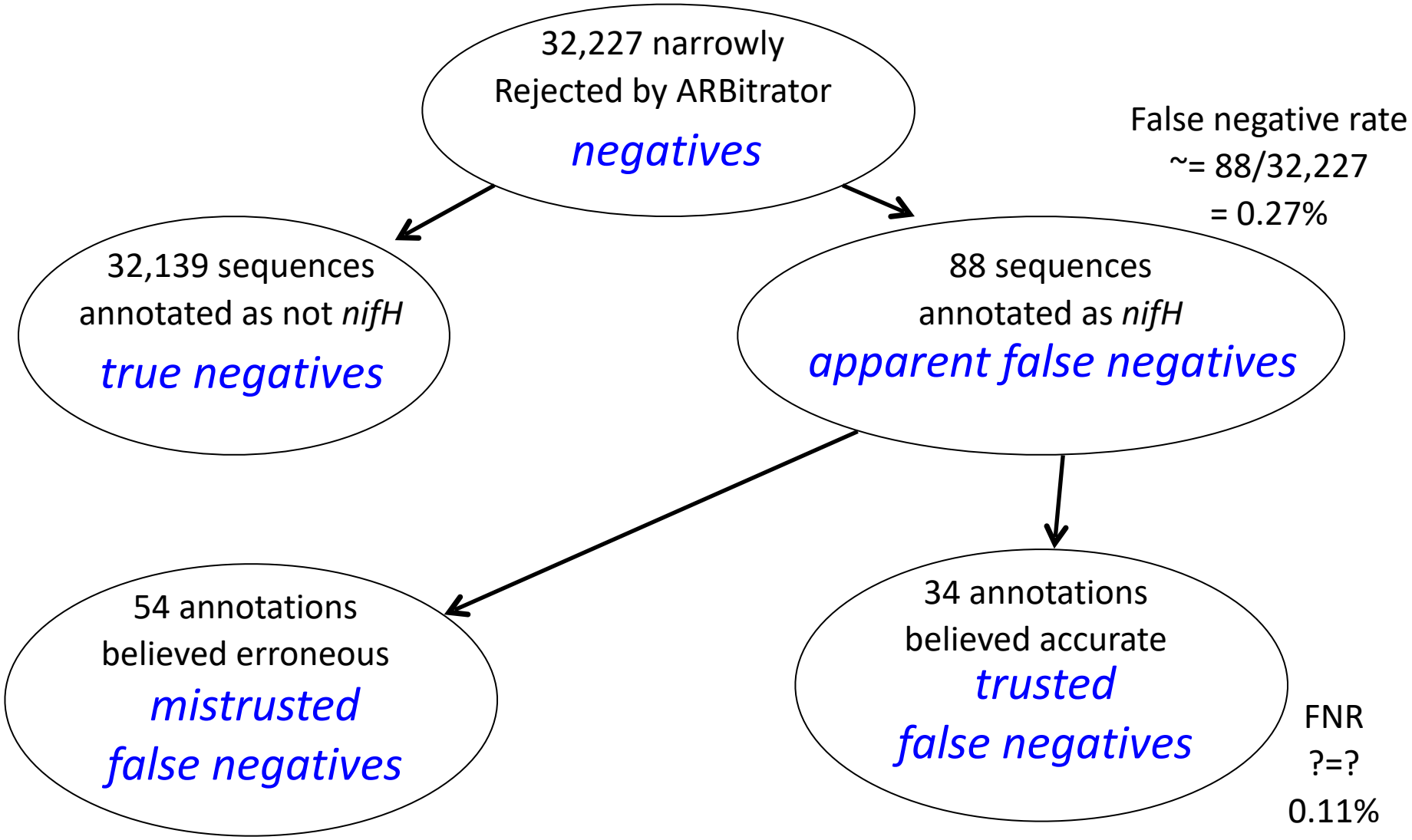
• • •

- The only non-self hit among the top 20 that isn't annotated as *nifH*
- Author: Villadas P.J. & al
- Title: The rhizosphere bacterial community of the leguminous trees Eperua falcata and Dicorynia guianensis from tropical rainforest in French Guiana
- Either ARBitrator is right and these 2 annotations from same study are wrong - OR - ARBitrator and 18 of 20 other annotations from various studies are wrong

False-Negative Rate

- 32,227 sequences were rejected by a small margin
- All annotations of those sequences were checked computationally for “NifH” synonyms

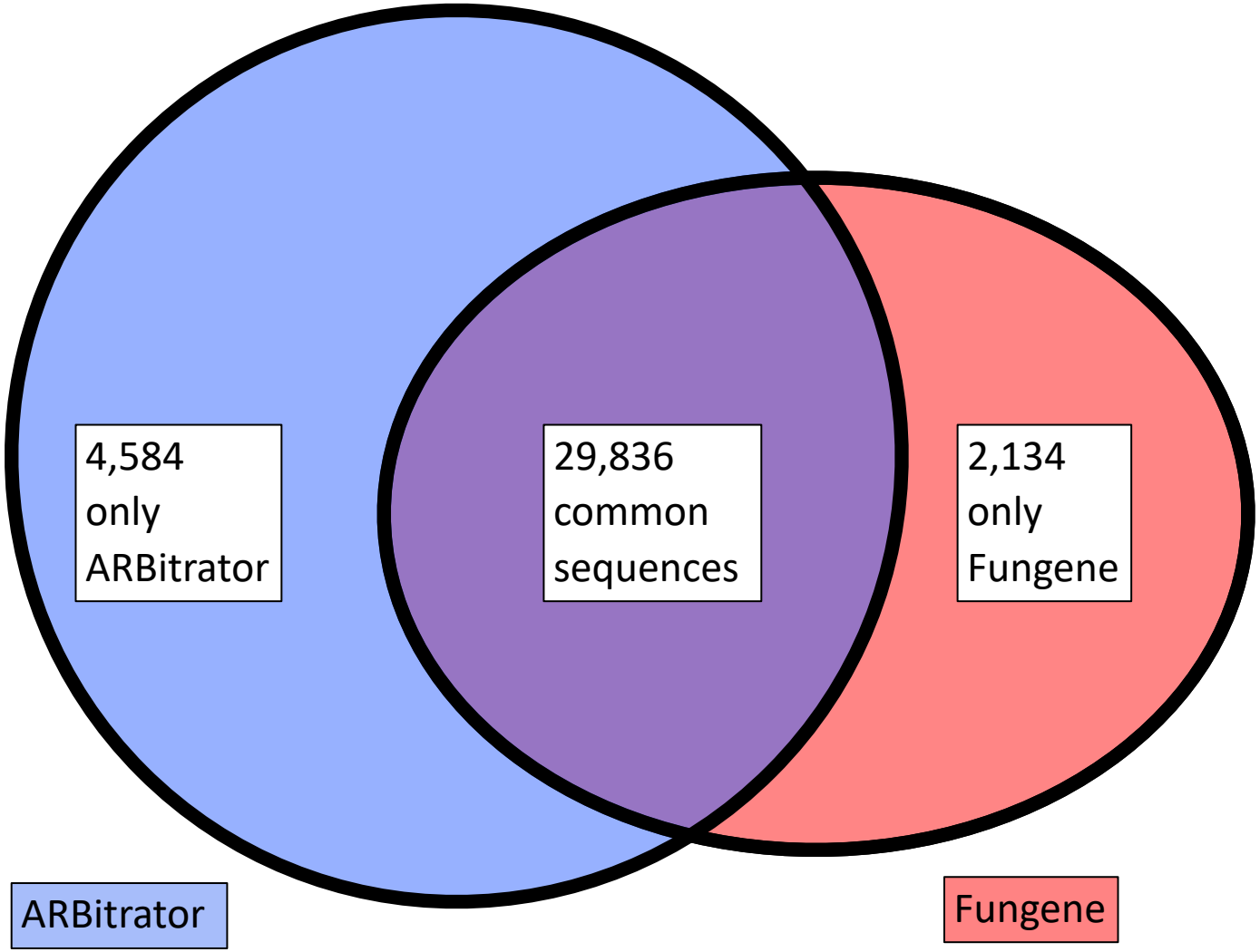
False-Negative Rate (assuming correct annotations)



Error rates

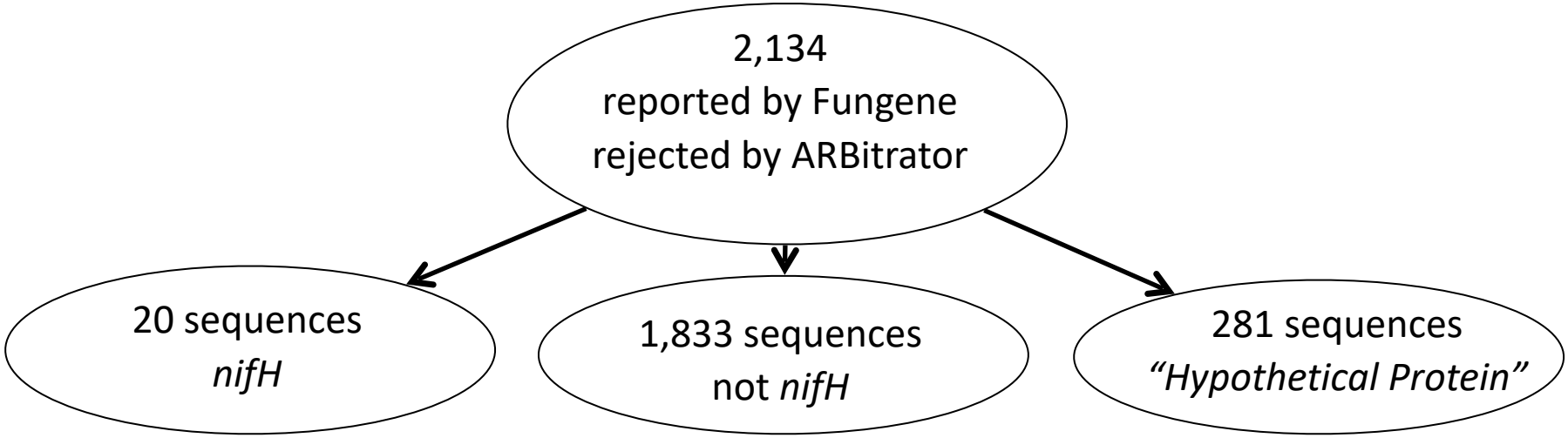
- False positive rate, assuming correct annotations, = .16%
  - P (an ARBitrator sequence isn't NifH | it is correctly annotated) = 0.0016
- False negative rate, assuming correct annotations, = .11%
  - P (a sequence that ARBitrator doesn't report is actually NifH | it is correctly annotated) = 0.0011
- These look like very low rates, but how do they compare to Fungene?

ARBitrator vs. Fungene



How did ARBitrator miss 2,134 sequences?

Annotations of Fungene-only sequences



Among these genes, annotation-based false positive rate = 92%

CO-ARBitrator

- Re-purposing ARBitrator to retrieve COI sequences
- Challenges
  1. Over 1M COI sequences in GenBank, vs 34K NifH at time of 1st ARBitrator run
  2. COI diversity
     - Supposedly different for *every* animal species
     - Hard to find representatives for phase-1 blast
  3. No COI paralogs
     - Hard to find negative training set

CO-ARBitrator results

- 1,054,973 sequences reported in January 2018
  - Article published in Nature Scientific Data.
- False-positive rate ~= 0.0034%.
- False negative rate ~= 0.0018%.