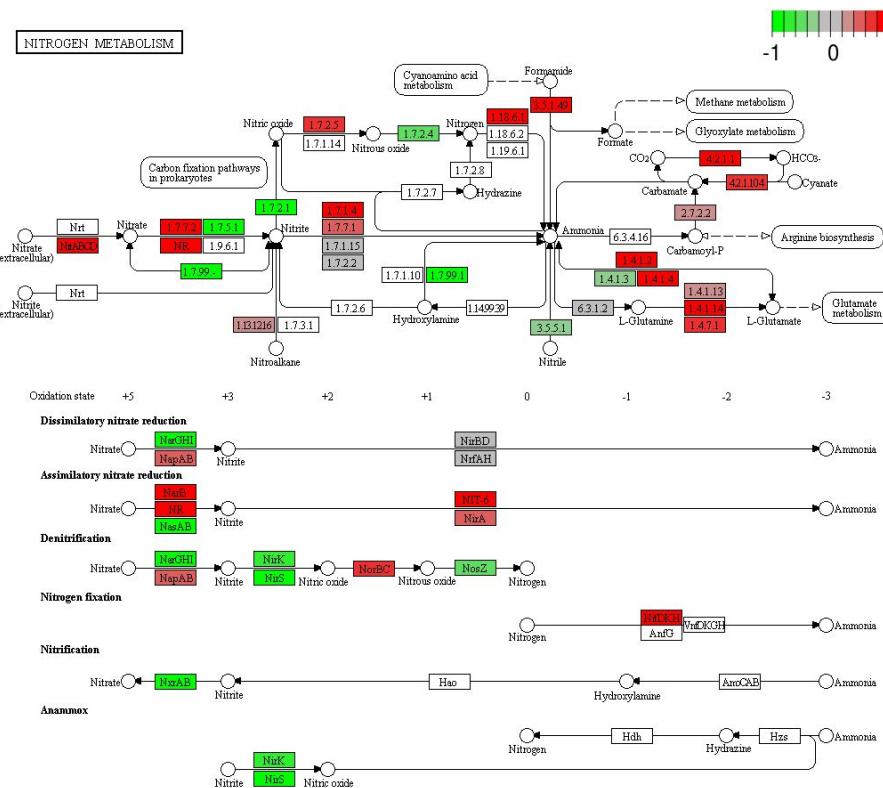


Annotation of genomes



By Dr. Richard Allen White III

Lecture 5 - Sep 24th, 2019

Zoom! 404-899-586

Annotation

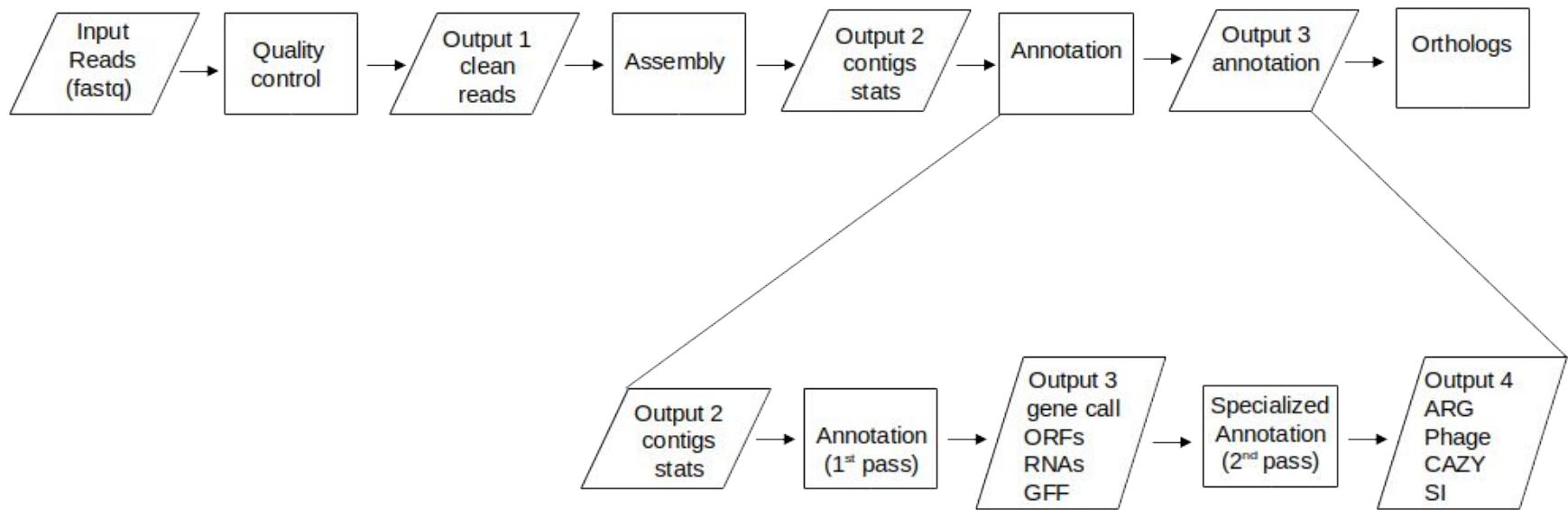
Concepts:

- Open reading frame (ORF) calling
- Reference databases
- Automatic annotation

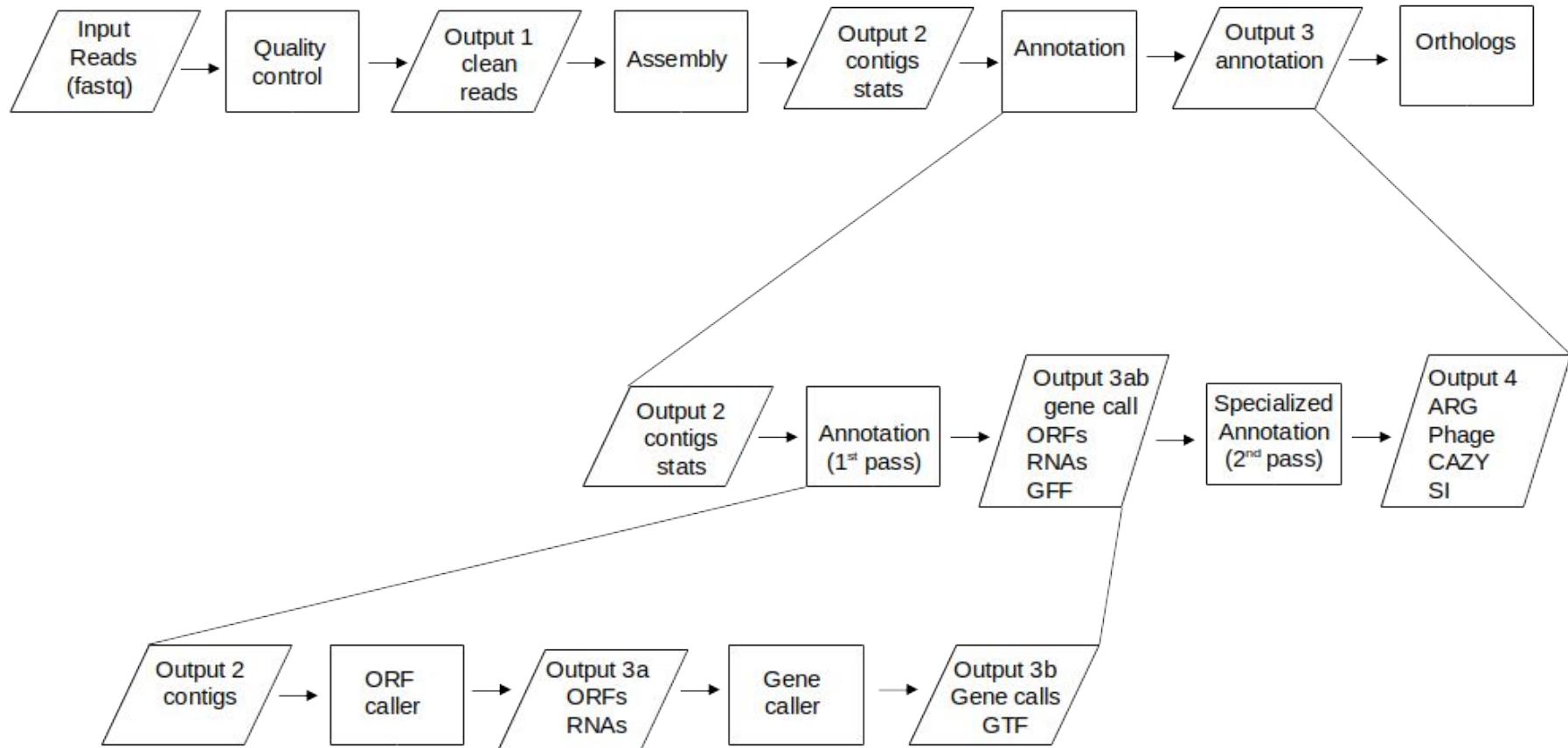
Learning Objectives:

- Ways and programs to call ORFs
- Your only as good as your database
- Finding the best annotation

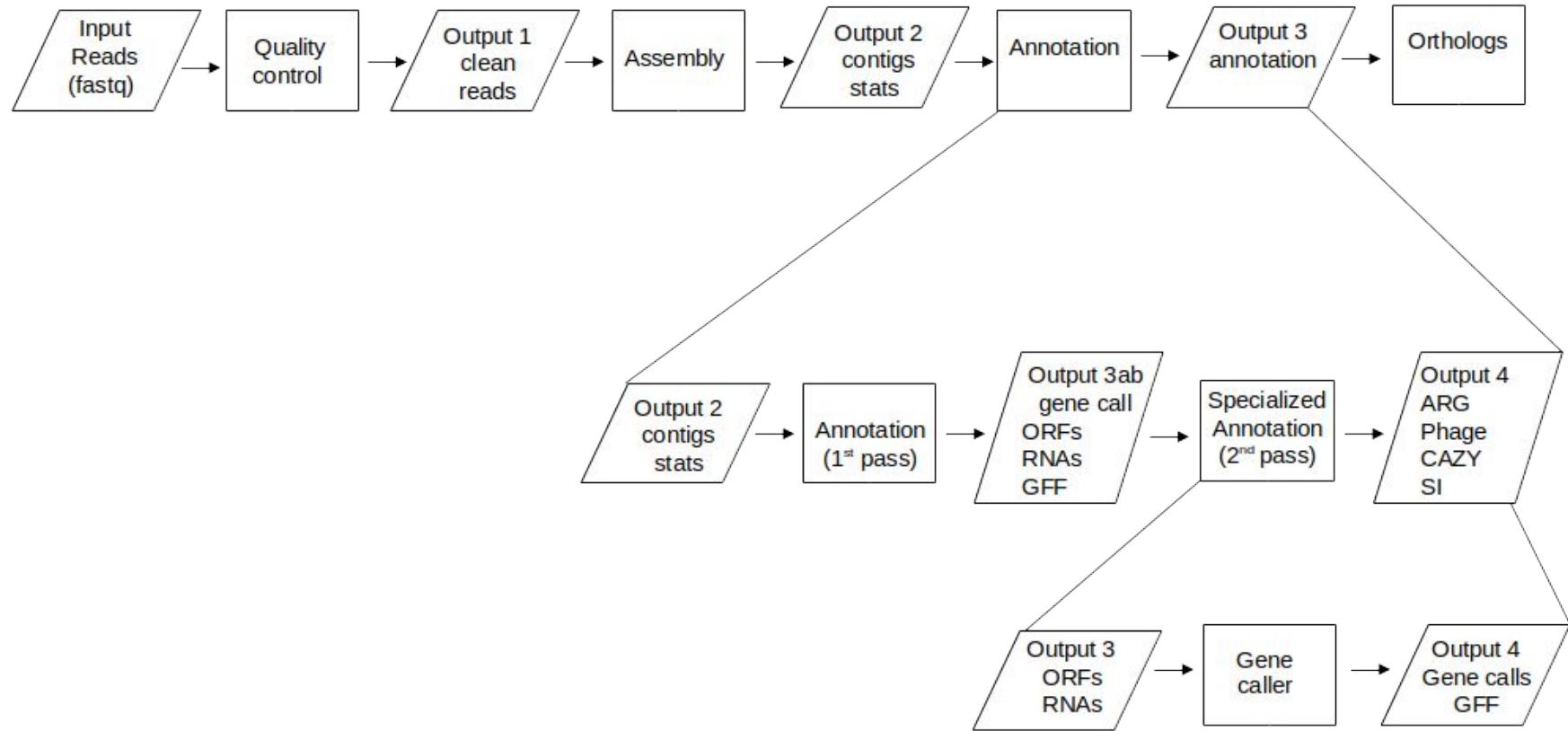
Annotation - flowgraph



Annotation - flowgraph



Annotation - flowgraph

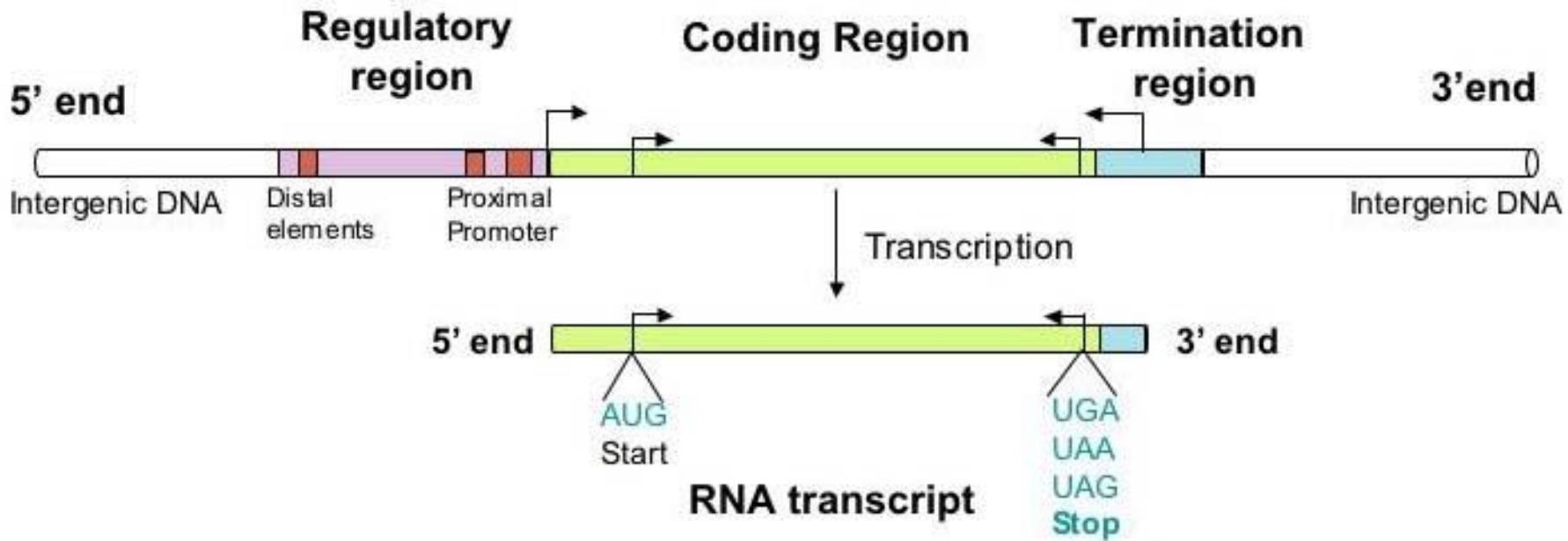


Annotation step 1 -ORF calling

		Second letter					
		U	C	A	G		
First letter	U	UUU } Phe UUC UUA } Leu UUG }	UCU } Ser UCC UCA UCG }	UAU } Tyr UAC UAA Stop UAG Stop }	UGU } Cys UGC UGA Stop UGG Trp }	U C A G	Third letter
	C	CUU } Leu CUC CUA CUG }	CCU } Pro CCC CCA CCG }	CAU } His CAC CAA } Gln CAG }	CGU } Arg CGC CGA CGG }	U C A G	
	A	AUU } Ile AUC AUA AUG Met	ACU } Thr ACC ACA ACG }	AAU } Asn AAC AAA } Lys AAG }	AGU } Ser AGC AGA } Arg AGG }	U C A G	
	G	GUU } Val GUC GUA GUG }	GCU } Ala GCC GCA GCG }	GAU } Asp GAC GAA } Glu GAG }	GGU } Gly GGC GGA GGG }	U C A G	

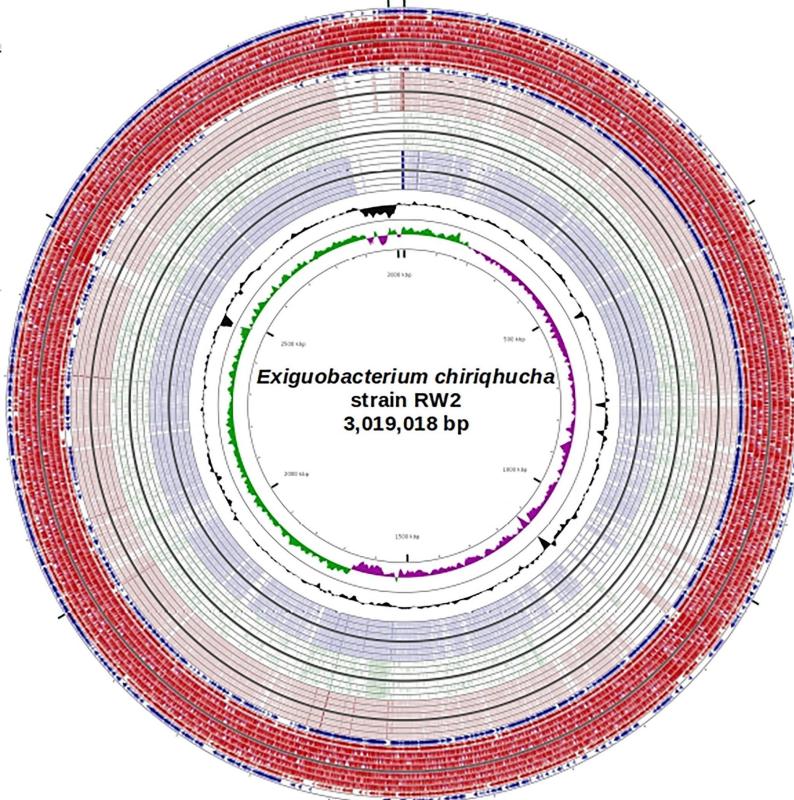
- Conversion of nucleotide contigs to gene calls
 - Find genes (start and stop codons), 1 start with 3 stops
 - Convert into amino acid sequences
 - Make map of location and size of genes (GFF or GTF)

Annotation step 1 -ORF calling



- Prokaryotic gene structure
 - High gene density (85% coding in *E. coli* K12)
 - Six-base consensus sequence is AGGAGG for translation is?
 - Is ORF a gene?
 - 29.5% have an imputed function, 2.1% have a phenotype and 19.5% have no function assignment (*E. coli* K12)
 - How much is coding in the human genome? (10%, 90%, 1%)

Annotation step 1 -ORF calling



- Finding ORFs
 - More ORFs than genes
 - RW2
 - >4,500 ORFs
 - 3,075 genes/3,022 proteins
 - K12
 - >6000 ORFs
 - 4,566 genes/4,242 proteins
 - If random, one stop per every 64 divided by 3 = 21 codons on average
 - Average protein = ~300 codons
 - Problems
 - Hard to search and find long ORFs
 - Short genes
 - Overlapping on opposite strands

White 3rd et al., 2019.

<https://www.frontiersin.org/articles/10.3389/fmicb.2018.03189/full>

Annotation step 1 -ORF calling

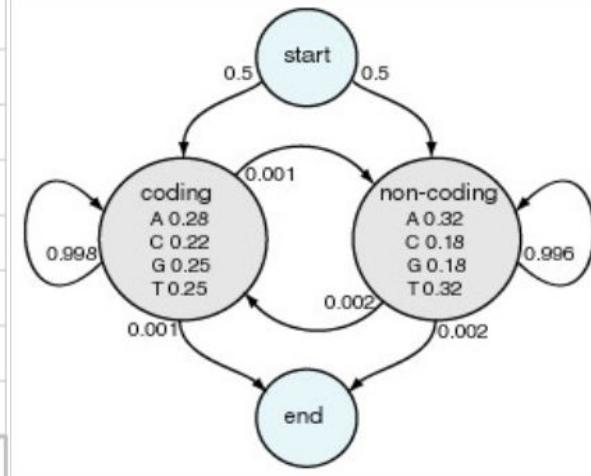
- Methods for ORF calling
 - 6-frame translation
 - Sixpack
(<http://emboss.sourceforge.net/apps/cvs/emboss/apps/sixpack.html>)
 - Interpolated Markov models (IMMs)
 - Glimmer (<http://ccb.jhu.edu/software/glimmer/index.shtml>)
 - GC bias information + Shine-Dalgarno RBS motif
 - Prodigal (<https://github.com/hyattpd/Prodigal>)
 - Heuristic in homogeneous Markov models with hidden Markov model (HMM)
 - Genemark (<http://opal.biology.gatech.edu/GeneMark/genemarks.cgi>)
 - Sequencing error models and codon usages in a hidden Markov model (HMM)
 - Fraggenescan (<https://sourceforge.net/projects/fraggenescan/>)
 - Aho–Corasick algorithm to find regions uninterrupted by stop codons
 - orfM (<https://github.com/wwood/OrfM>)

Annotation step 2 - Gene calling

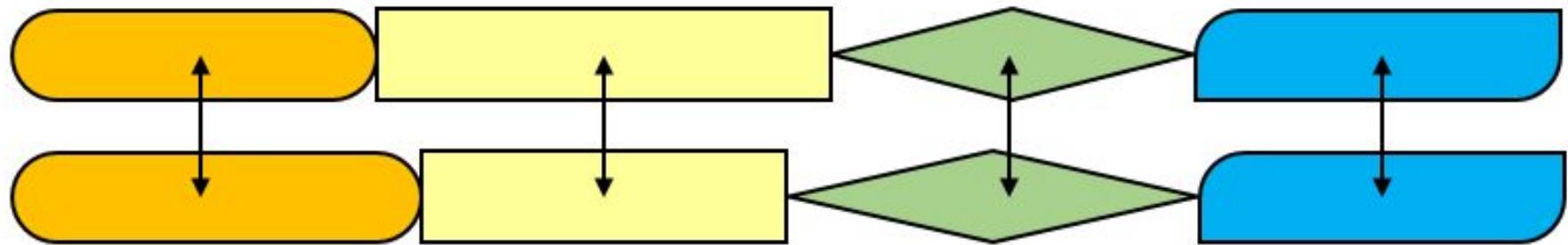
- Gene calling or Gene prediction
 - Figure out what the proteins are coding for based on reference
 - Alignment based (BLAST - local alignments)
 - Deep searching (hmmer - HMMs)

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

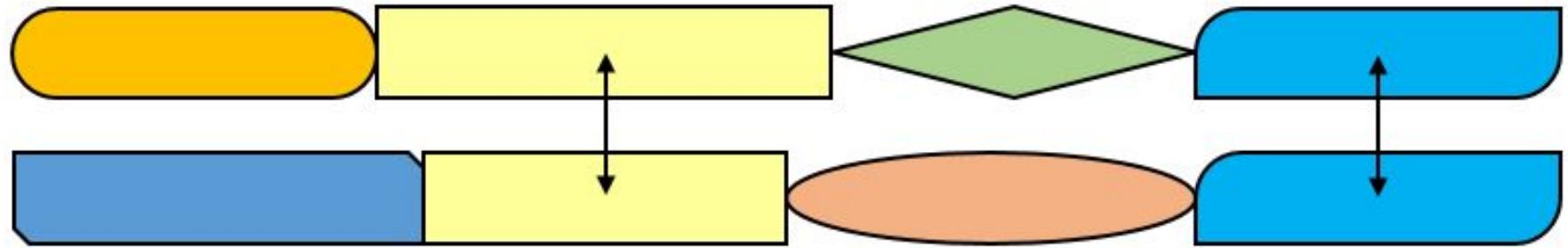
	match = 1		mismatch = -1		gap = -1			
	G	C	A	T	G	C	U	
G	0	-1	-2	-3	-4	-5	-6	-7
A	-1	1	0	-1	-2	-3	-4	-5
C	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0



Annotation step 2 -Gene calling

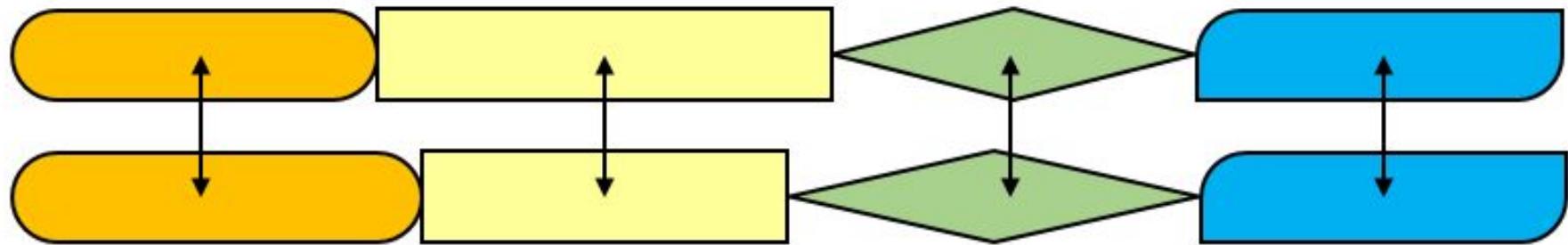


Type? Algorithm?

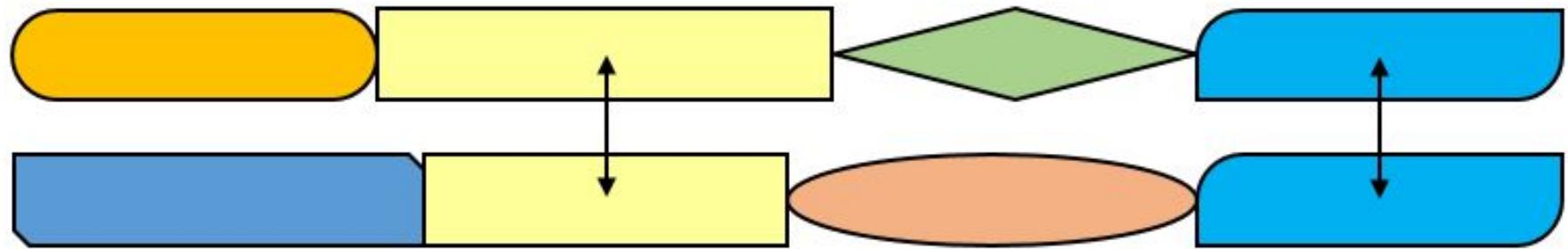


Type? Algorithm?

Annotation step 2 -Gene calling

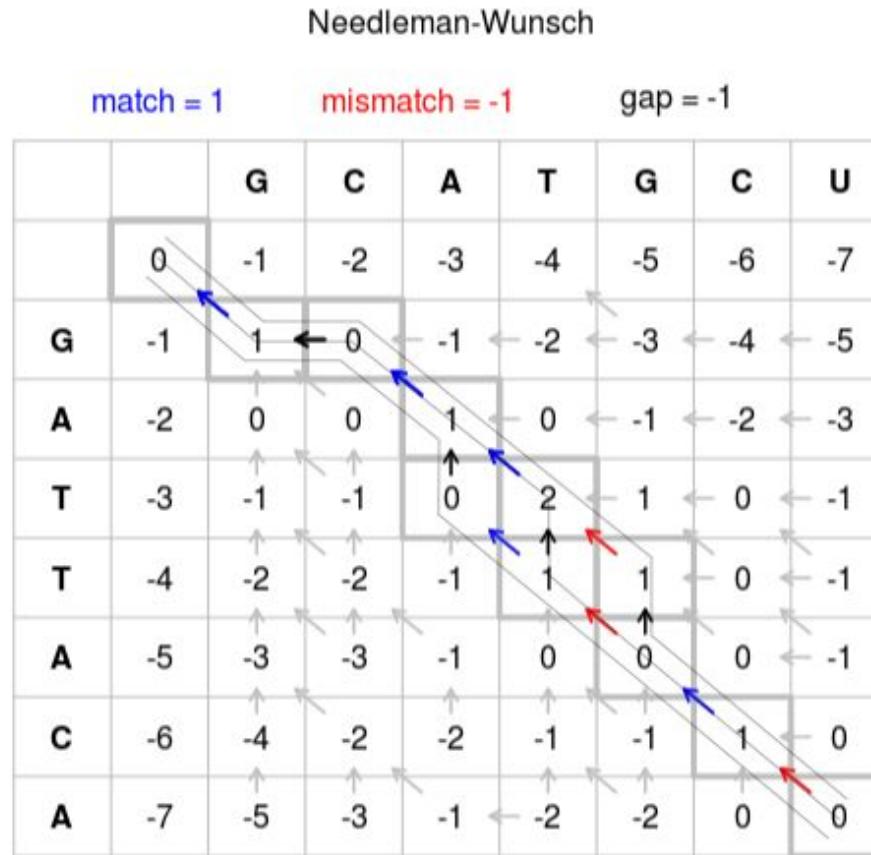


Global, Algorithm: Needleman-Wunsch



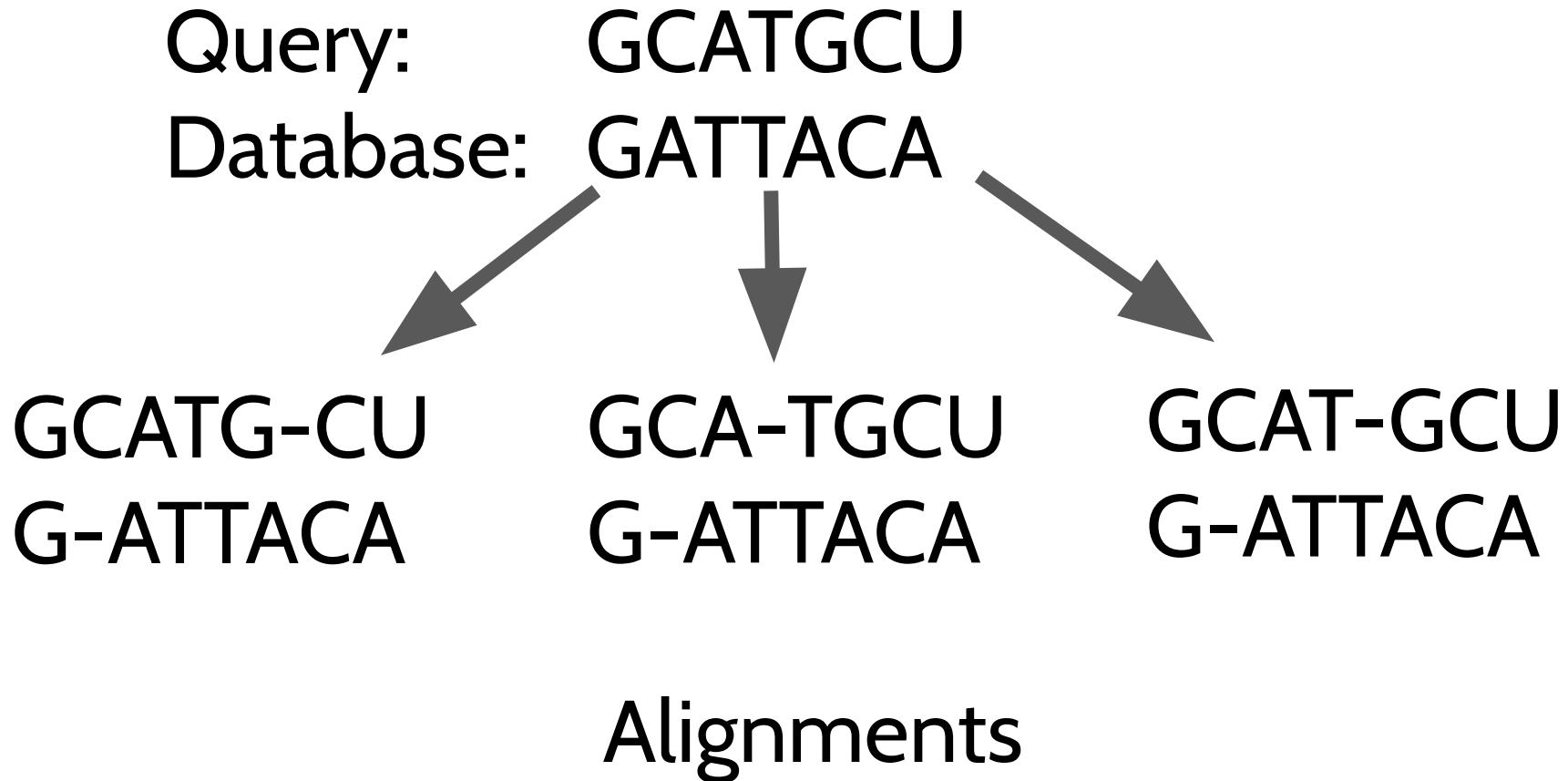
Local, Algorithm: Smith-waterman

Annotation step 2 -Gene calling



Global Algorithm: Needleman-Wunsch

Annotation step 2 -Gene calling



Global, Algorithm: Needleman-Wunsch

Annotation step 2 -Gene calling

Alignment: +GCATGCU
+GATTACA

Score: 0 // Handle matches handle, doesn't win any score

Alignment: +GCATGCU
+GATTACA

Score: 0 // 1 gap, score -1

Alignment: +GCATGCU
+GATTACA

Score: 0 // 2 gaps, score -2

Alignment: +GCATGCU
+GATTACA

Score: 0 // 3 gaps, score -3

Alignment: +GCATGCU
+GATTACA

Score: 0 // 4 gaps, score -4

*Global,
Algorithm: Needleman-Wunsch scoring*

Annotation step 2 -Gene calling

Initialize the scoring matrix

	T	G	T	T	A	C	G	G
T	0	0	0	0	0	0	0	0
G	0							
G	0							
T	0							
T	0							
G	0							
A	0							
C	0							
T	0							
A	0							

Substitution matrix:

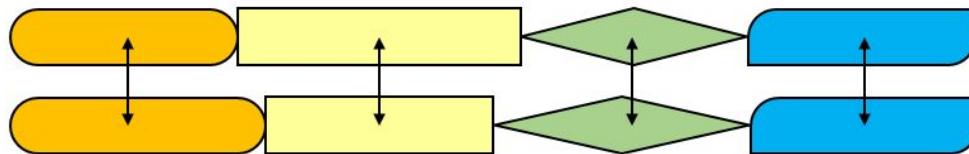
$$S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$$

Gap penalty:

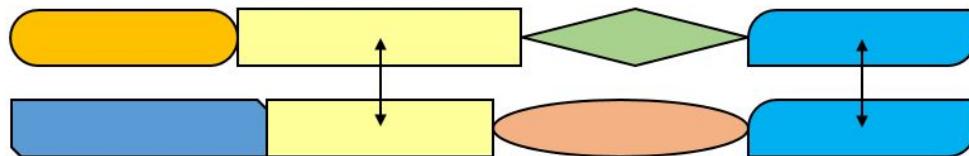
$$W_k = kW_1$$
$$W_1 = 2$$

Local Algorithm: Smith-Waterman

Annotation step 2 -Gene calling



Global Alignment

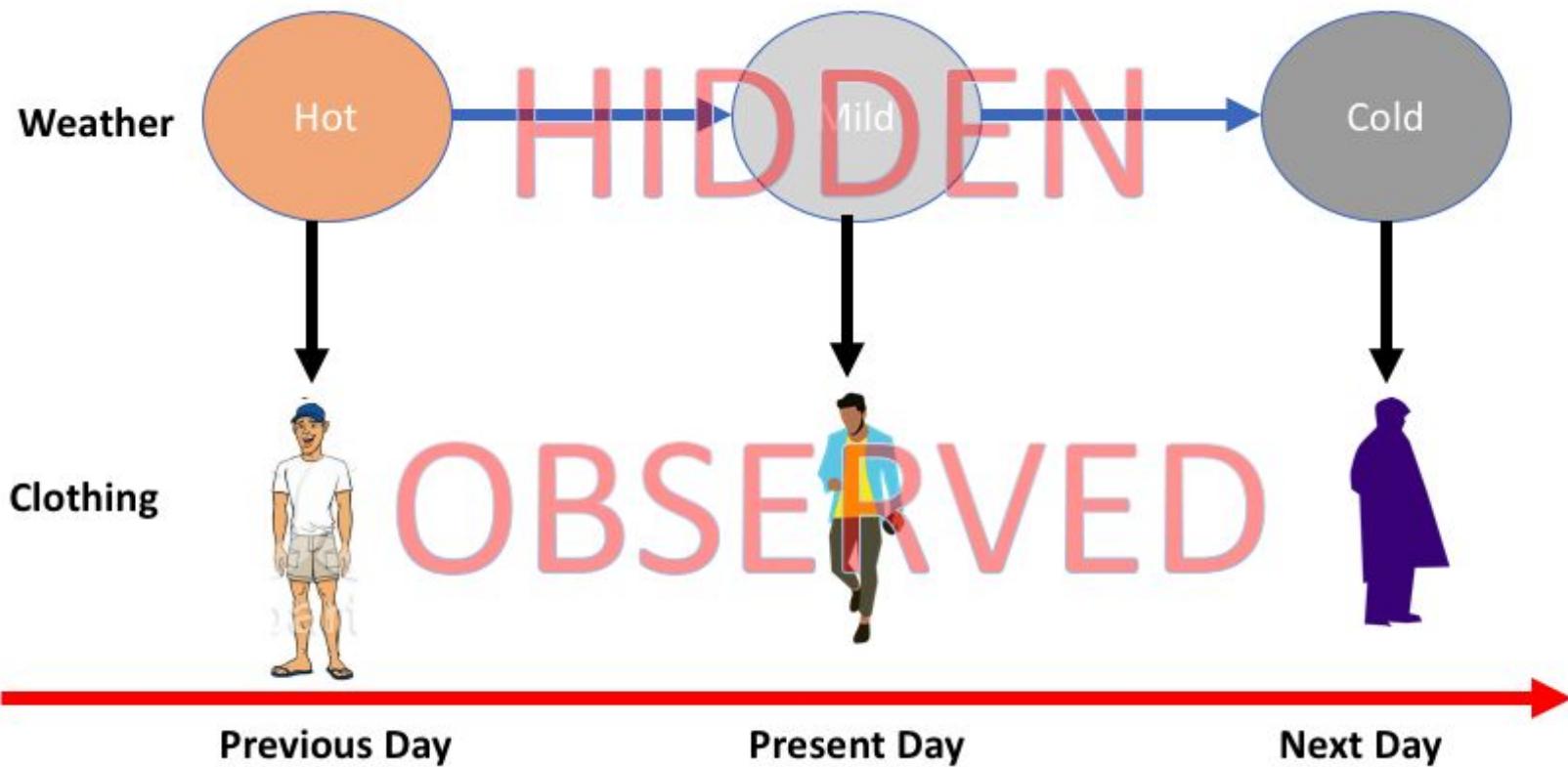


Local Alignment

	Smith-Waterman algorithm	Needleman-Wunsch algorithm
Initialization	First row and first column are set to 0	First row and first column are subject to gap penalty
Scoring	Negative score is set to 0	Score can be negative
Traceback	Begin with the highest score, end when 0 is encountered	Begin with the cell at the lower right of the matrix, end at top left cell

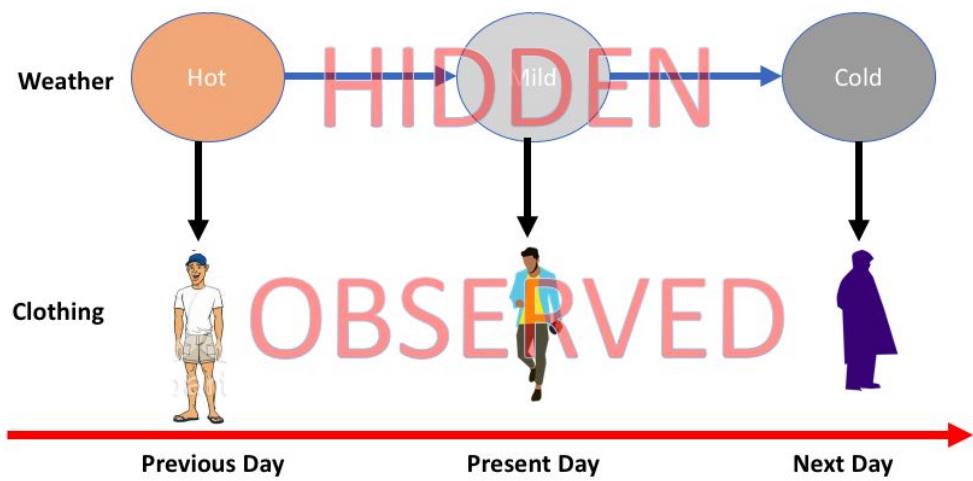
Local vs. Global

Annotation step 2 -Gene calling (HMM)



Annotation step 2 -Gene calling (HMM)

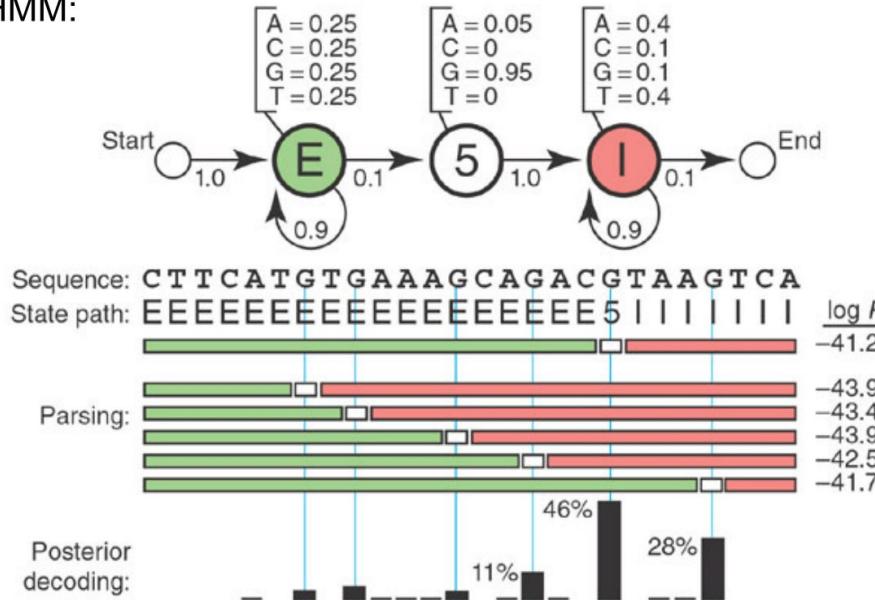
- 1) Transition data — the probability of transitioning to a new state conditioned on a present state.
- 2) Emission data — the probability of transitioning to an observed state conditioned on a hidden state.
- 3) Initial state information — the initial probability of transitioning to a hidden state. This can also be looked at as the prior probability.



Annotation step 2 -Gene calling (HMM)

§ A **Markov model** (or **chain**) is a probability model comprised of one or more **states** that generate a sequence of symbols (e.g., DNA: A, C, G, T)

Example HMM:



§ In a first order or simple Markov chain, the next state chosen depends only on the current state (not on any past states)

§ In a **Hidden Markov Model (HMM)**, we cannot directly determine the current state from the generated (or emitted) symbol

Figure from SR Eddy *Nature Biotechnology* 22, 1315 - 1316 (2004)

Annotation step 2 -Gene calling (HMM)

§ A protein sequence profile HMM, the states correspond to **matches**, **insertions**, or **deletions** at different positions in the profile

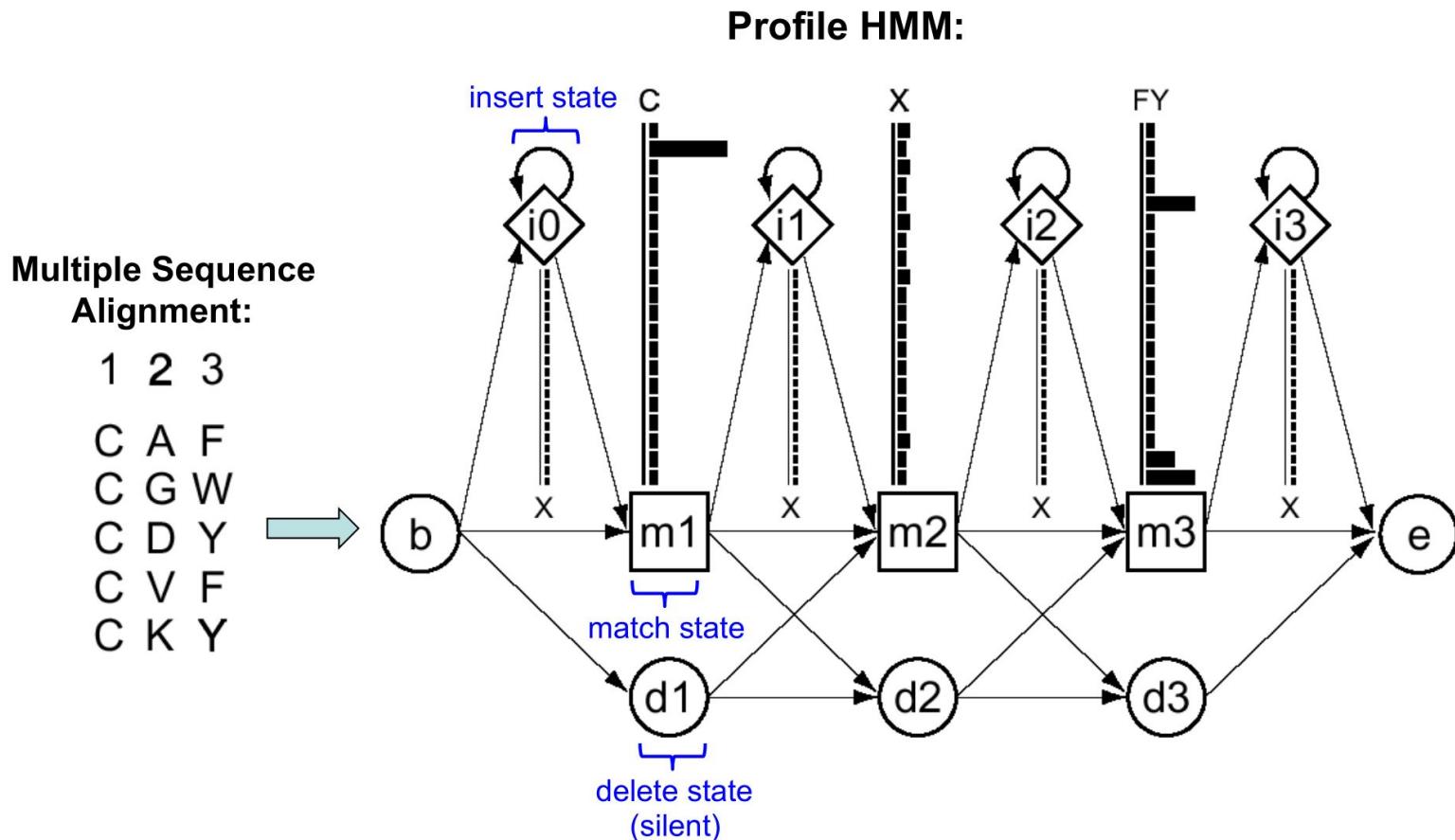
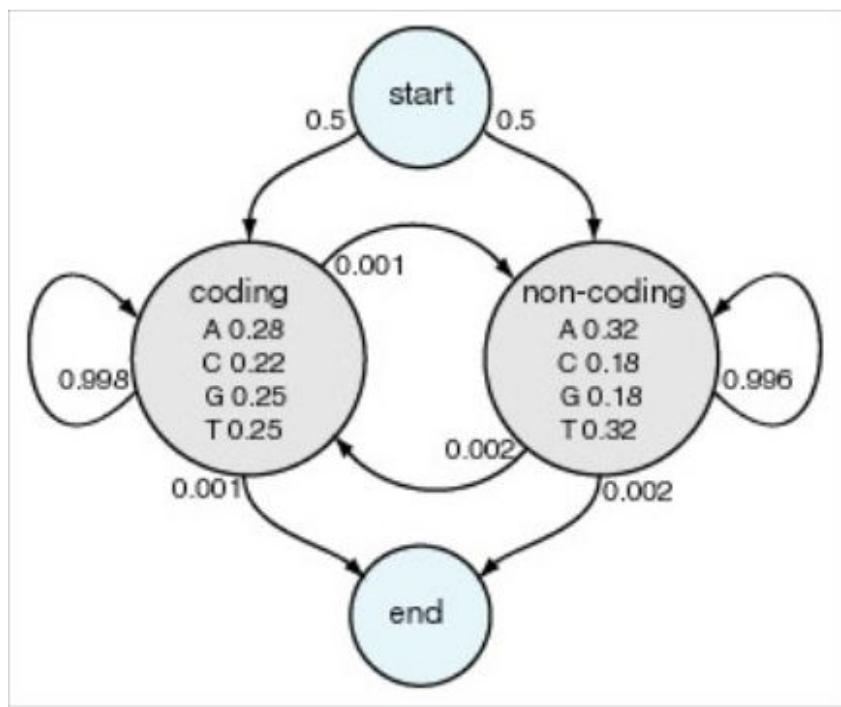


Figure from SR Eddy (1998) *Bioinformatics* 14:755-763

downloaded from: http://www.quretec.com/u/vilo/edu/2005-06/Text_Algorithms/L10_Probabilistic/HMM_intro_2.png

Annotation step 2 -Gene calling (HMM)



$$\Phi = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0.998 & 0.002 & 0 \\ 0.5 & 0.001 & 0.996 & 0 \\ 0 & 0.001 & 0.002 & 0 \end{bmatrix}$$

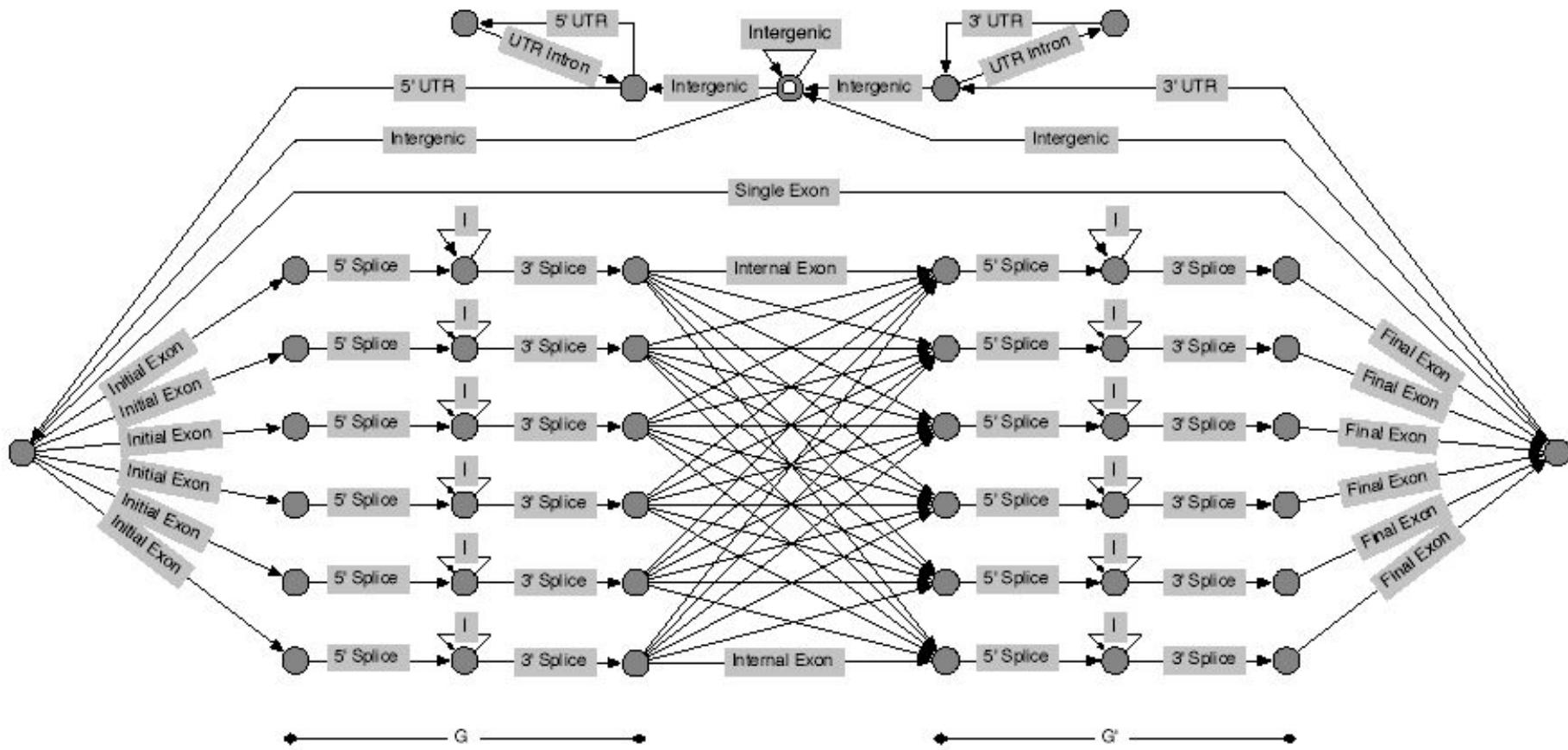
$$H = \begin{bmatrix} 0.28 & 0.32 \\ 0.22 & 0.18 \\ 0.25 & 0.18 \\ 0.25 & 0.32 \end{bmatrix}$$

$x_m(i)$ = probability of being in state m at position i;

$H(m, y_i)$ = probability of emitting character y_i in state m;

Φ_{mk} = probability of transition from state k to m.

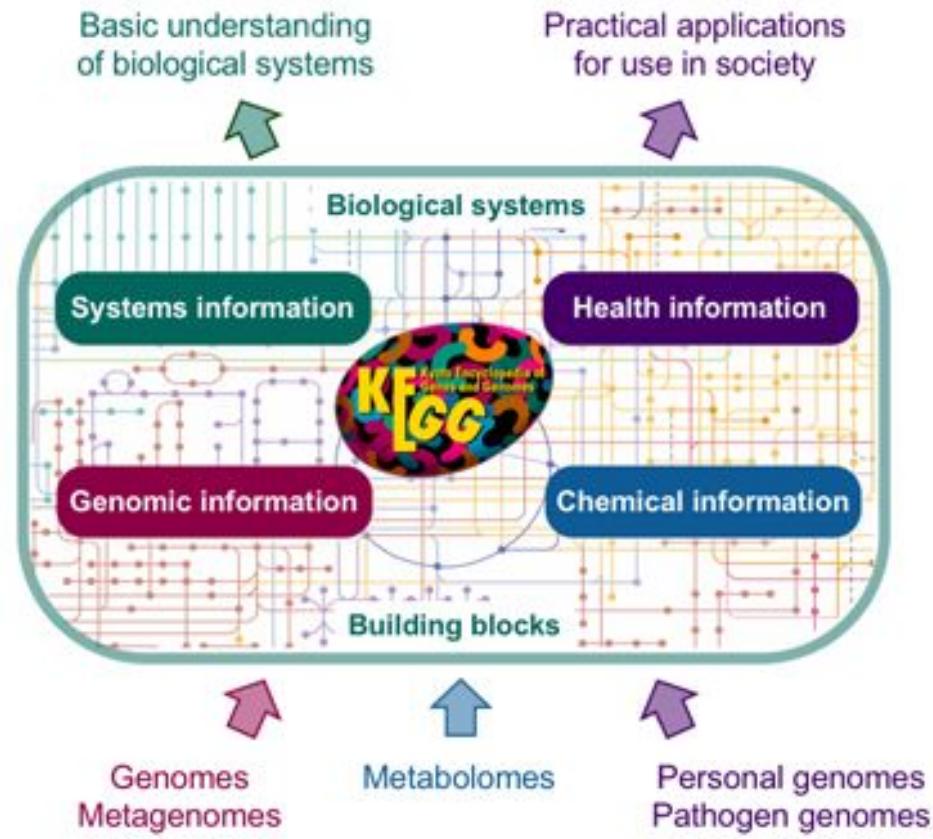
Annotation step 2 -Gene calling (HMM)



Only as good as your database!!

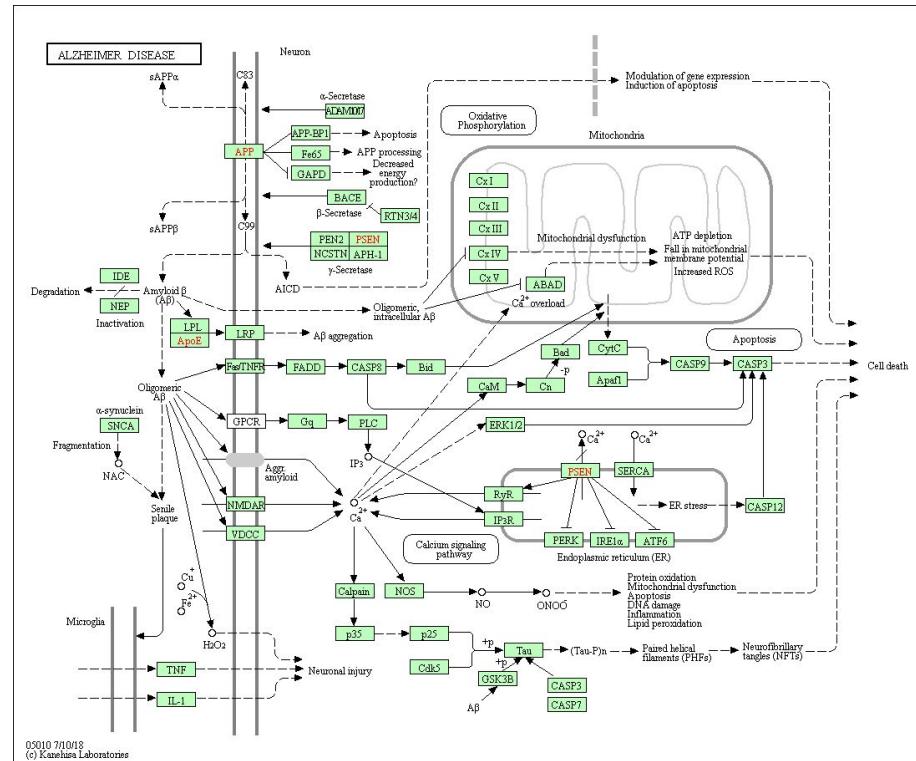
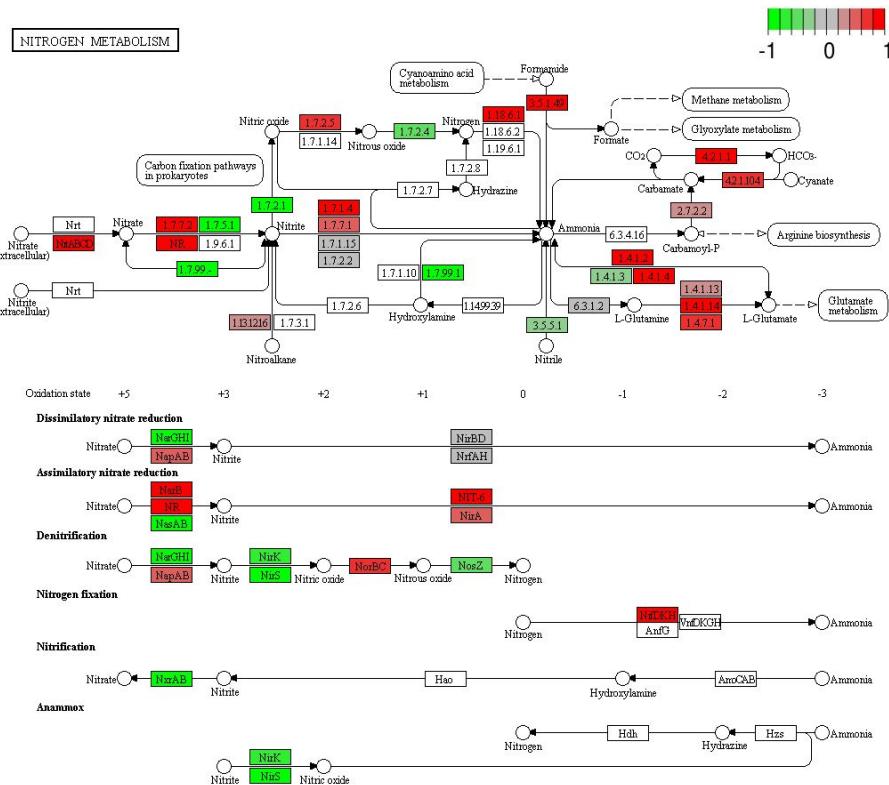
- Only as good as your database reference
 - Functional gene database
 - SEED
 - KEGG (KO)
 - COG
 - NOG, now eggNOG
 - Uniprot
 - CaZy
 - MetaCyc
 - Taxonomic identification database
 - GenBank
 - RefSeq
 - Uniprot
 - GTDB

Only as good as your database (KEGG)



<https://www.genome.jp/kegg/kegg1a.html>

Only as good as your database (KEGG)



<https://www.genome.jp/kegg/kegg1a.html>

\$5k a year!

Only as good as your database (Uniprot)

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (560,823)
Manually annotated and reviewed.
Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (171,501,488)
Automatically annotated and not reviewed.
Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations, Taxonomy, Subcellular locations, Cross-ref. databases, Diseases, Keywords.

UniProt data

Download latest release, Get the UniProt data, Statistics, View Swiss-Prot and TrEMBL statistics, How to cite us, The UniProt Consortium, Submit your data, Submit your sequences and annotation updates.

News

Forthcoming changes, Planned changes for UniProt, UniProt release 2019_08, Magnetic personalities | Cross-references to DrugCentral, Pharos and MassIVE | Change of UniRef clustering method from CD-HIT to MMseqs2 ..., UniProt release 2019_07, The enemy of my enemy is my friend | Cross-references to ChEBI in the pmlist.txt document file | Retirement of UniProt decoy databases [...], News archive.

Protein spotlight

A Sense Of Direction, September 2019

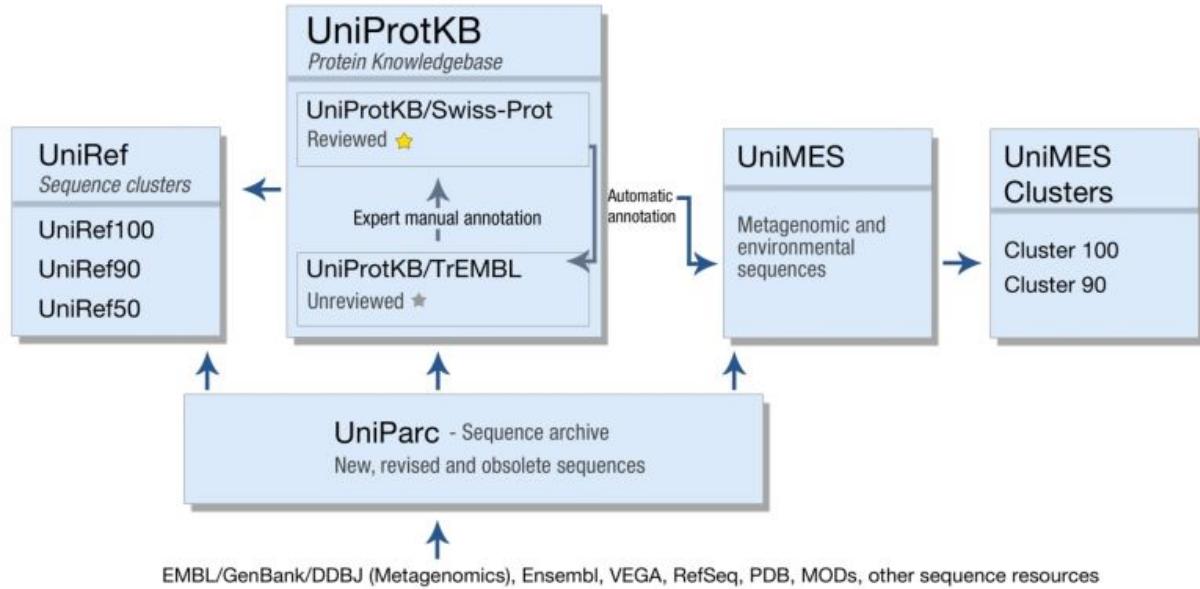
Survival depends on cues, mobility and a medium to evolve in. Cues - such as scents, sounds or colours for example - will attract organisms towards food, mating grounds and an environment in which they feel protected and are happy to stay. Thanks to them, organisms usually head off in a direction they expect will be to their advantage, using the means of locomotion they have, to cross all sorts of media...

We'd like to inform you that we have updated our Privacy Notice to comply with Europe's new General Data Protection Regulation (GDPR) that applies since 25 May 2018.

Do not show this banner again

<https://www.uniprot.org/>

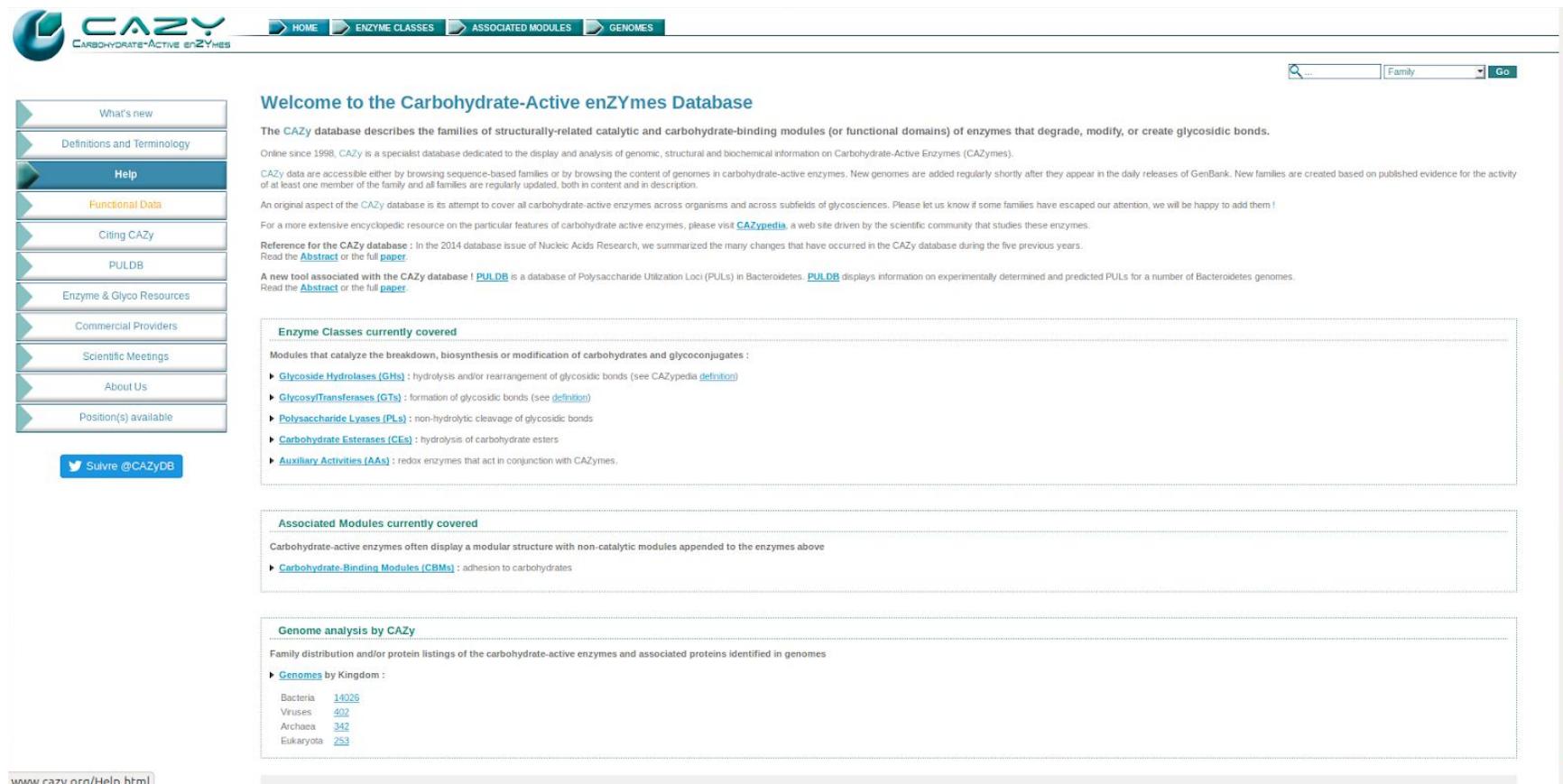
Only as good as your database (Uniprot)



<https://www.uniprot.org/>

FREE!!!

Only as good as your database (CAZY)



The screenshot shows the CAZY database homepage. At the top, there's a navigation bar with links to HOME, ENZYME CLASSES, ASSOCIATED MODULES, and GENOMES. A search bar at the top right contains the text "Family" with a "Go" button. On the left, a sidebar lists various links: What's new, Definitions and Terminology, Help (selected), Functional Data, Citing CAZY, PULDB, Enzyme & Glyco Resources, Commercial Providers, Scientific Meetings, About Us, and Position(s) available. Below the sidebar is a Twitter link: "Suivre @CAZyDB". The main content area has three sections: "Enzyme Classes currently covered" (listing GHs, GTs, PLs, CEts, and AAAs), "Associated Modules currently covered" (listing CBMs), and "Genome analysis by CAZy" (showing a table of genome counts by kingdom). The footer includes the URL "www.cazy.org/Help.html" and a copyright notice: "Last update: 2019-06-21 © Copyright 1998-2019".

<http://www.cazy.org/>

FREE!!! BUT?

Only as good as your database (CAZy)

 dbCAN meta server: automated CAZyme annotation

Home | Annotate | Download | Help | About us

What is dbCAN2 meta server?

Cite us: NAR/gky418 and gks479

dbCAN2 meta server is a web server for automated [Carbohydrate-active enzyme ANnotation](#), funded by the National Science Foundation (DBI-1652164). Similar resources on the web include CAZy, CAT (obsolete), and Hotpep. dbCAN2 meta server is an updated version of the original dbCAN web server, and has the following [new features](#) (thanks to dbCAN users all over the world for suggestions):

- dbCAN2 meta server allows submission of nucleotide sequences: prokaryotic genomic sequences (fna file) of draft genomes and metagenomes; for eukaryotic genomes, please [still submit protein seqs \(faa file\)](#)
- dbCAN2 meta server integrates three state-of-the-art tools/databases for automated CAZyme annotation:
 1. HMMER for annotated CAZyme domain boundaries according to the [dbCAN CAZyme domain HMM database](#)
 2. DIAMOND for fast blast hits in the [CAZy database](#)
 3. Hotpep for short conserved motifs in the [PPR library](#)
- dbCAN2 meta server can identify transcription factors (TFs), transporters (TCs), and further CAZyme gene clusters (CGCs) using [CGC-Finder](#) if users submit faa+gff files or fna file
- dbCAN2 meta server combines the results from the three tools and allows visualization as venn diagram and detailed results as graphs

dbCAN2 meta server will be updated once a year to use the most updated CAZy database, dbCAN HMM database and Hotpep peptide database

News

- 8/08/2019: dbCAN HMMdb v8 is released (based on CAZyDB 7/26/2019). Now the HMMdb contains 641 CAZyme HMMs (421 family HMMs + 3 cellulose HMMs + 217 subfamily HMMs). The CAZyDB for Diamond search is also updated, containing in total 1,386,849 fasta sequences. See [readme](#) for details.
- 4/01/2019: dbCAN2 has a [docker](#) version written by Haidong Yi.
- 3/19/2019: dbCAN2 web server has moved to UNL and has a new [URL](#).
- 1/20/2019: dbCAN2 standalone package is available on [github](#); if you prefer to still use the old hmmscan way, the data are available in the [download page](#)
- 8/25/2018: dbCAN HMMdb v7 is released (based on CAZyDB 7/31/2018): HMMs of 15 new families were added (AA14, AA15, CBM82, CBM83, GH146, GH147, GH148, GH149, GH150, GH151, GH152, GH153, GT105, GT106, PL28). GT2 family HMM now is replaced with 8 Pfam HMMs (GT2_Chitin_synth_1, GT2_Chitin_synth_2, GT2_Glycos_transf_2, GT2_Glyco_transf_2_2, GT2_Glyco_transf_2_3, GT2_Glyco_transf_2_4, GT2_Glyco_transf_2_5, GT2_Glyco_trans_2_3)
- 5/2/2018: dbCAN2 meta server paper is accepted to publish at [Nucleic Acids Research](#)
- 8/15/2017: Tanner and Le Huang begin to work on dbCAN2 meta server
- 7/1/2017: Yanbin is awarded the NSF CAREER grant for CAZyme bioinformatics research

1,723 Pageviews
Aug. 24th - Sep. 24th



Copyright 2017 © YIN LAB, UNL. All rights reserved. Designed by Tanner Yohe and Le Huang. Maintained by Yanbin Yin.

<http://bcb.unl.edu/dbCAN2/>

Only as good as your database (MetaCyc)

My Most Visited  Getting Started

 A member of the KEGG database collection

Sites ▾ | Search ▾ | Genome ▾ | Metabolism ▾ | Analysis ▾ | SmartTables ▾ | Help ▾ |

MetaCyc Metabolic Pathway Database

MetaCyc is a curated database of experimentally elucidated metabolic pathways from all domains of life. MetaCyc contains 2722 pathways from 3009 different organisms.

MetaCyc contains pathways involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes, and genes. The goal of MetaCyc is to catalog the universe of metabolism by storing a representative sample of each experimentally elucidated pathway.

MetaCyc applications include:

- Online encyclopedia of metabolism
- Predict metabolic pathways in sequenced genomes
- Support metabolic engineering via enzyme database
- Metabolite database aids metabolomics research

[Guide To MetaCyc](#)

Accessing MetaCyc Data

MetaCyc data can be accessed in several ways:

- [Search](#) for pathways, enzymes, reactions, and metabolites through MetaCyc.org
- [Download](#) MetaCyc data files
- [Browse](#) a list of all MetaCyc metabolic pathways
- Add MetaCyc to your browser search bar for [instant MetaCyc searching](#)


Empty? by Arkadiusz R. Used under creative common license

Chitin is the second most common polymer in the world, after cellulose. Our understanding of its degradation was significantly boosted with the discovery of LPMOs. And what are LPMOs?

[Learn More](#)

1 2 3 4 5 6 7 8 9 10

How to Cite MetaCyc

Please cite MetaCyc as Caspi et al 2018, "The MetaCyc database of metabolic pathways and enzymes", *Nucleic Acids Research* 46(D1):D633-D639

Funding Sources

The development of MetaCyc is funded by grant GM080746 from the NIH National Institute of General Medical Sciences.

<https://metacyc.org/>

Only as good as your database (MetaCyc)

Sites ▾ | Search ▾ | Genome ▾ | Metabolism ▾ | Analysis ▾ | SmartTables ▾ | Help ▾ |

Provide Feedback

Add to SmartTable

MetaCyc Pathway: phenylmercury acetate degradation

Enzyme View: All Organisms ▾ More Detail Less Detail

This view shows enzymes only for those organisms listed below, in the list of taxa known to possess the pathway. If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.

Some taxa known to possess this pathway include : Bacillus sp. RC607, Escherichia coli J53-1, Pseudomonas sp. K-62, Staphylococcus aureus

Expected Taxonomic Range: Bacteria

Superclasses: Detoxification → Mercury Detoxification

Summary:
Mercury exists naturally in small amounts in the environment, though its levels have been increasing due to anthropogenic activities such as coal burning, the use of mercurial fungicides and the use of mercury as catalyst in industry. Mercury is extremely toxic, mostly due to its affinity to thiol groups, with which it forms tight coordinate bonds.

Mercury resistance is the most widespread antimicrobial-resistance system, and is found in a wide variety of Gram-negative and Gram-positive bacteria, as well as archaea [Summers72, Clark77a, Weiss77, Schottel78, Mirgain89, Kiyono97]. Mercury resistance is often divided into two categories: broad-spectrum (resistance to inorganic mercury as well as a wide range of organomercurials) and narrow spectrum (resistance to inorganic mercury and a very limited number of organomercurials). The key enzyme in both categories is **mercury(II) reductase** (MerA), which reduces the more toxic mercuric ion (Hg^{2+}) to the relatively non-toxic and volatile elemental mercury (Hg^0), a reaction that occurs in the cytoplasm [Schiering91].

Organisms that possess broad-spectrum resistance also utilize a second enzyme, **organomercury lyase** (MerB) that breaks the mercury-carbon bond in the organomercurials and releases the mercury in the form of Hg^{2+} , which is then reduced by **mercury(II) reductase**.

Unification Links: Eawag-BBD-Pathways:ogm

Credits:
Created 23-Apr-2001 by Pellegrini-Toole A, Marine Biological Laboratory
Revised 12-Jan-2007 by Caspi R, SRI International

References

Clark77a: Clark DL, Weiss AA, Silver S (1977). "Mercury and organomercurial resistances determined by plasmids in *Pseudomonas*." J Bacteriol 132(1):186-96. PMID: 410779

<https://metacyc.org/>

BioCyc - Pay model like KEGG

Only as good as your database (GenBank)

The screenshot shows the NCBI GenBank homepage. At the top, there's a navigation bar with links for NCBI, Resources, How To, raw937, My NCBI, and Sign Out. Below the navigation is a search bar with dropdown menus for Nucleotide, GenBank, Submit, Genomes, WGS, Metagenomes, TPA, TSA, INSDC, and Other. The main content area has several sections:

- GenBank Overview**: Describes GenBank as the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. It mentions the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. It notes that these organizations exchange data on a daily basis.
- Access to GenBank**: Provides information on how to search and retrieve data from GenBank, including links to Entrez Nucleotide, BLAST, and NCBI e-utilities.
- GenBank Data Usage**: States that the GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. It notes that NCBI places no restrictions on the use or distribution of the GenBank data, but some submitters may claim patent, copyright, or other intellectual property rights.
- Confidentiality**: Discusses the policy regarding the appearance of data in GenBank prior to publication. It states that upon request, NCBI will withhold release of new submissions for a specified period of time. However, if the accession number or sequence data appears in print or online prior to the specified date, your sequence will be released. It encourages authors to inform NCBI of the appearance of their published data and to send full publication data to the following address: update@ncbi.nlm.nih.gov.
- Privacy**: Notes that if you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source. It states that GenBank assumes that the submitter has received any necessary informed consent authorizations required prior to submitting sequences.

<https://www.ncbi.nlm.nih.gov/genbank/>

FREE!!! BUT!!

Only as good as your database (GenBank)



<https://www.ncbi.nlm.nih.gov/genbank/>

Landfill of data - can have errors (phiX)

Only as good as your database (GTDB)

The screenshot shows the GTDB homepage with a background image of a tree. At the top, there are two informational banners: one about preprints and another about the update to GTDB R04-RS89. Below this, a bar chart displays the taxonomic distribution of 143,512 bacteria, with the total count of 23,458 shown above the chart. A pie chart below the bacteria chart shows the distribution of named clusters (35.4%), unnamed clusters (64.4%), isolates (59.2%), MAGs (38.9%), and SAGs (1.8%). The main content area features a welcome message, the title "GENOME TAXONOMY DATABASE", the number of genomes (145,904), and the release date (Release 04-RS89 (19th June 2019)). To the right, there is a sidebar titled "Tweets by @ace_gtdb" with two tweets from the official account. The first tweet links to a bioRxiv preprint about species clusters. The second tweet announces the release of GTDB-Tk v0.3.2, highlighting improved output tables and reduced classification time. At the bottom right, there is a logo for "Ecogenomics/GTDBTk" and links for "Embed" and "View on Twitter".

*** Preprint describing GTDB species clusters is out in [bioRxiv](#) ***
*** GTDB-Tk has been updated to use GTDB R04-RS89: More info [here](#) ***

BACTERIA (143,512)

SPECIES 23,458
GENUS 7,372
FAMILY 1,969
ORDER 816
CLASS 296
PHYLUM 112

ARCHAEA (2,392)

16 PHYLUM
36 CLASS
96 ORDER
238 FAMILY
534 GENUS
1,248 SPECIES

Welcome to GTDB

GENOME TAXONOMY DATABASE

145,904 genomes

Release 04-RS89 (19th June 2019)

Tweets by @ace_gtdb

GTDB @ace_gtdb Preprint describing the methodology used to establish species clusters in the GTDB is out in @biorkxivpreprint: biorxiv.org/content/10.110...

GTDB @ace_gtdb GTDB-Tk v0.3.2 has been released. It features improved output tables and a ~40 min reduction in classification time. This version is the basis of the upcoming GTDB-Tk manuscript and was validated on >10,000 MAGs. [github.com/Ecogenomics/GT...](#)

GTDB TK Ecogenomics/GTDBTk

Embed View on Twitter

<https://gtdb.ecogenomic.org/>

Use this taxonomy!!

Only as good as your database (GTDB)

The screenshot shows the GitHub repository page for **Ecogenomics / GTDBTk**. The page includes navigation links for Code, Issues (13), Pull requests (0), Wiki, Security, and Insights. A header bar shows Watch (20), Star (99), Fork (13), and a search bar. Below the header, a summary box displays 393 commits, 5 branches, 18 releases, 6 contributors, and the license GPL-3.0. A dropdown menu for the branch is set to "stable". Buttons for "New pull request", "Create new file", "Upload files", "Find File", and "Clone or download" are visible. The main content area lists recent commits by user **dspark1134**, with details like commit message, file changes, and timestamp.

GTDB-Tk: a toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes.

taxonomy species-assignments phylogenetics archaea bacteria nomenclature

393 commits 5 branches 18 releases 6 contributors GPL-3.0

Branch: stable ▾ New pull request Create new file Upload files Find File Clone or download ▾

dspark1134 Optimized gtdb_to_ncbi_majority_vote.py script. Latest commit ef6dae6 on Aug 5

File	Commit Message	Time Ago
bin	gtdbtk test writes log file to out_dir and exits cleaner.	2 months ago
docs	add notes for release 0.3.0	3 months ago
gtdbtk	Symlink the unrooted tree when running infer.	2 months ago
scripts	Optimized gtdb_to_ncbi_majority_vote.py script.	2 months ago

<https://github.com/Ecogenomics/GtbdTk>

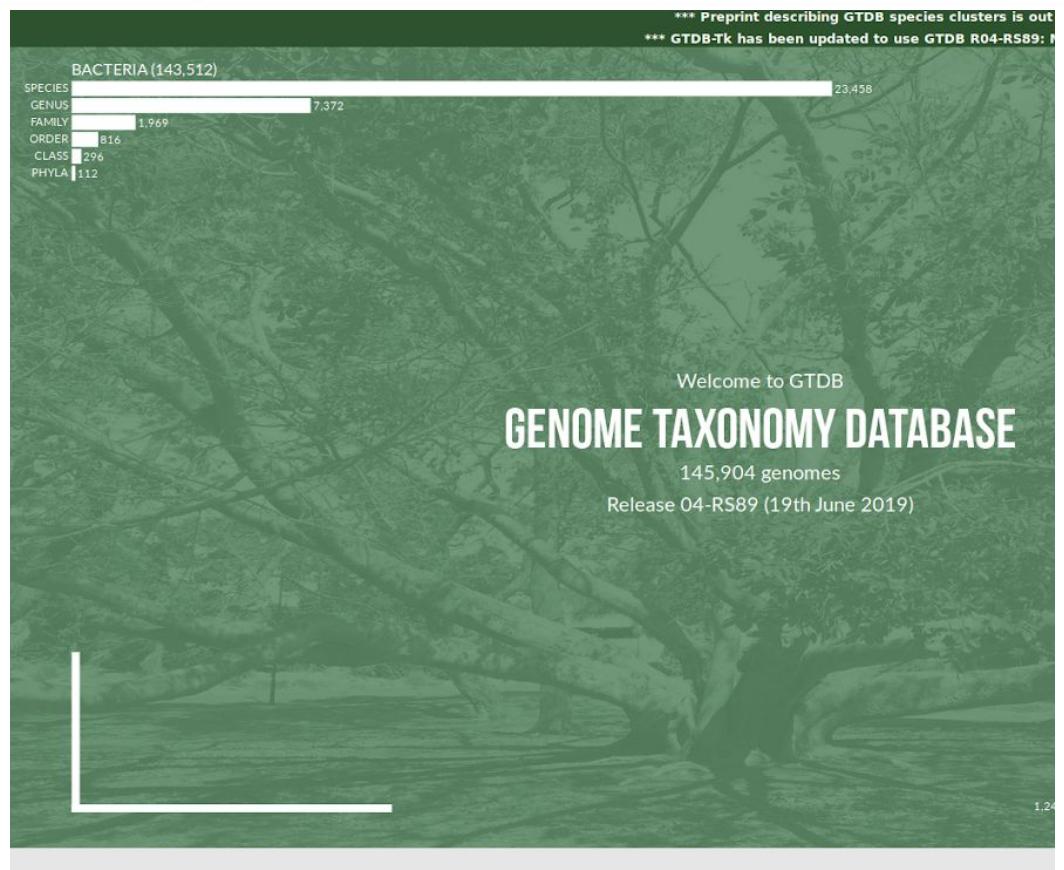
Use this tool!

Only as good as your database (GTDB)

Name	Last modified	Size	Description
Parent Directory			
FILE_DESCRIPTIONS	11-Sep-2019 01:12	3.0K	
METHODS	17-Sep-2019 04:16	1.8K	
RELEASE_NOTES	09-Aug-2019 22:55	1.3K	
VERSION	20-Jun-2019 12:12	30	
ar122_metadata_r89.tsv	20-Jun-2019 12:12	2.9M	
ar122_msa_individual_genes_r89.tar.gz	19-Jun-2019 14:45	13M	
ar122_msa_marker_info_r89.tsv	20-Jun-2019 12:12	8.6K	
ar122_msa_mask_r89.txt	19-Jun-2019 14:45	32K	
ar122_msa_r89.faa	19-Jun-2019 14:45	6.1M	
ar122_r89.sp_labels.tree	17-Sep-2019 04:03	68K	
ar122_r89.tree	09-Aug-2019 22:48	56K	
ar122_ssu_r89.fna	20-Jun-2019 12:12	1.2M	
ar122_taxonomy_r89.tsv	20-Jun-2019 12:12	334K	
bac120_metadata_r89.tsv	20-Jun-2019 12:12	197M	
bac120_msa_individual_genes_r89.tar.gz	19-Jun-2019 14:45	343M	
bac120_msa_marker_info_r89.tsv	20-Jun-2019 12:12	8.2K	
bac120_msa_mask_r89.txt	19-Jun-2019 14:45	40K	
bac120_msa_r89.faa	19-Jun-2019 14:45	113M	
bac120_r89.sp_labels.tree	17-Sep-2019 04:03	1.2M	
bac120_r89.tree	19-Jun-2019 14:45	1.0M	
bac120_ssu_r89.fna	20-Jun-2019 12:12	26M	
bac120_taxonomy_r89.tsv	20-Jun-2019 12:12	20M	
gtdb_uba_mags.tar.gz	19-Jun-2019 14:46	2.1G	
gtdb_uba_mags_arc.tar.gz	19-Jun-2019 14:46	93M	
gtdbtk_r89_data.tar.gz	19-Jun-2019 17:53	26G	
metadata_field_desc.tsv	13-Sep-2019 00:18	6.6K	
ncbi_vs_gtdb_r89_archaea.xls	30-Aug-2019 07:45	31K	
ncbi_vs_gtdb_r89_bacteria.xls	30-Aug-2019 07:45	281K	
sp_clusters_r89.tsv	20-Jun-2019 12:12	6.7M	
ssu_r89.fna	20-Jun-2019 12:12	401M	
synonyms_r89.tsv	30-Aug-2019 11:42	74K	

Apache/2.2.15 (CentOS) Server at data.ace.uq.edu.au Port 443

<https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0/>



Automatic annotator - RAST

The RAST system is back online (sort of). Several of the nodes that ran services that RAST depends on failed to reboot after the scheduled downtime, and had to be repaired or replaced. In particular, the BLAST compute-cluster will be down for some time, causing all RASTik jobs to stall at the BLAST step, and crippling RAST's throughput for "Classic RAST". . .

After some deliberation, we have reluctantly elected to disable the BLAST stage of RAST rather than allow jobs to accumulate, so that at least jobs will be able to run to completion minus BLAST similarities.

The main impacts that disabling similarities will have are that:

- 1.) Genes in "Classic RAST" jobs with no Kmer hits and whose translations are not already in our nonredundant database will be annotated "Hypothetical protein".
- 2.) Genes whose translations are not already in our nonredundant database will not be able to retrieve homologous regions in other genomes within the "Compare Regions" display of the SEED-Viewer.

If a gene *did* have Kmer hits and could be placed into a PATtyFam, we can partially compensate for (2.) above by constructing "Compare Regions" based on PATtyFam membership rather than BLAST similarity, so we will switch "Compare Regions" to default to aligning regions based on PATtyFam membership.

If you feel you **do** need BLAST sims against the NR and/or BLAST-based SEED-viewer companions, you should select "Classic RAST" rather than default to RASTik — but be advised that, because BLAST throughput is crippled, "Classic RAST" jobs are going to run slower than usual. :-(

We are continuing to work on the system, and we hope to bring RAST back to full capacity as soon as the compute cluster can be restored. (We are also working on a stopgap measure that will compute BLAST similarities against just the nearest-neighbor genomes, but it will probably be at least a week before that feature can go "live".)

We apologize that the scheduled maintenance outage took longer than originally anticipated and that the system did not come back up cleanly afterwards. We thank you for your patience during RAST's recovery period.

Welcome to RAST

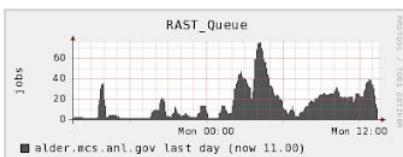
[» Register for a new account, service, or user-group](#)

[» Forgot your password?](#)

Login

Password

RAST Job Load, last 24 hours



What is RAST?

RAST (Rapid Annotation using Subsystem Technology) is a fully-automated service for annotating complete or nearly complete bacterial and archaeal genomes. It provides high quality genome annotations for these genomes across the whole phylogenetic tree.

We have a number of presentations and tutorials available:

- [Registering for RAST](#)
- [The IRIS/Automated Assembly/RASTik Workshop Presentations and Tutorials](#)
- [The SEED/Classic-RAST Workshop presentations and Tutorials](#)
- [Downloading and installing the RASTik Toolkit](#)
- [Downloading and installing the myRAST Toolkit](#)

<http://rast.nmpdr.org/>

Automatic annotator - RAST

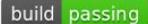
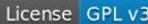
The screenshot shows the RAST homepage. At the top, there's a green banner with links to "Tutorials" and "Help". Below this, a blue header bar contains the RAST logo and the text "Rapid Annotation using Subsystem Technology version 2.0". A green sidebar on the left contains a "Info: NOTICE: RAST is Back Online (sort of)" message. This message explains that the RAST system was back online after a scheduled downtime, but due to issues with the BLAST stage, it was disabled. It also mentions that some genes without Kmer hits or translations were annotated as "Hypothetical protein". The main content area has a white background with a blue header "Welcome to RAST". It features a login form with fields for "Login" and "Password" and a "Login" button. To the right of the login form is a chart titled "RAST Job Load, last 24 hours" showing the number of jobs over time.

Methods for RAST

- Glimmer
 - ORFs
- BLAST
 - Gene calling
- Database
 - FigFams
- Framework
 - SEED subsystems

Automatic annotator - PROKKA

 README.md

 build passing  License GPL v3  DOI 10.1093/bioinformatics/btu153  Language Perl 5

Prokka: rapid prokaryotic genome annotation

Introduction

Whole genome annotation is the process of identifying features of interest in a set of genomic DNA sequences, and labelling them with useful information. Prokka is a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files.

Installation

Brew

If you are using the [MacOS Brew](#) or [LinuxBrew](#) packaging system:

```
brew install brewsci/bio/prokka
```

<https://github.com/tseemann/prokka>

Automatic annotator - PROKKA

build passing License GPL v3 DOI 10.1093/bioinformatics/btu153 Language Perl 5

Prokka: rapid prokaryotic genome annotation

Introduction

Whole genome annotation is the process of identifying features of interest in a set of genomic DNA sequences, and labelling them with useful information. Prokka is a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files.

Installation

Brew

If you are using the [MacOS Brew](#) or [LinuxBrew](#) packaging system:

```
brew install brewsci/bio/prokka
```

Methods for PROKKA

- Prodigal
 - ORFs
- BLAST+ and hmmer
 - BLAST+
 - 70%
 - Hmmer
 - 30%
- Database
 - UniProtKB/COG
- Framework
 - COG/UniProt

Automatic annotator - PROKKA

build passing License GPL v3 DOI 10.1093/bioinformatics/btu153 Language Perl 5

Prokka: rapid prokaryotic genome annotation

Introduction

Whole genome annotation is the process of identifying features of interest in a set of genomic DNA sequences, and labelling them with useful information. Prokka is a software tool to annotate bacterial, archaeal and viral genomes quickly and produce standards-compliant output files.

Installation

Install:

```
conda install -c conda-forge -c bioconda -c defaults prokka
```

Command:

```
prokka Bradyrhizobium_USDA3458.fna --cpu N --outdir USDA3458 --prefix  
USDA3458 --rfam
```

predicted 56 tRNAs, 1 transfer-messenger RNA (tmRNA), 37 noncoding RNAs (misc_RNA), 1 copy of a 5S-16S-23S operon, 0 CRISPRs, and 8,018 coding genes

White III et al., 2019. <https://mra.asm.org/content/8/38/e00813-19.full>

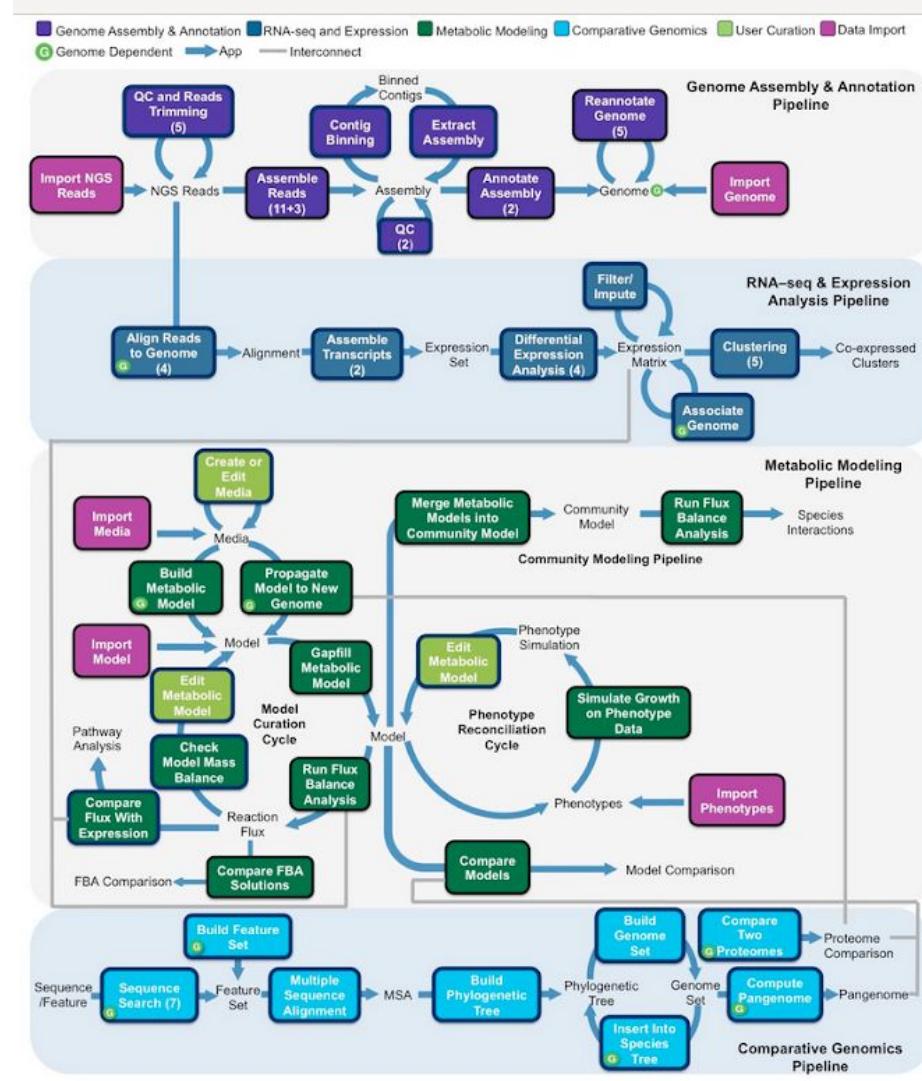
Automatic annotator - kBASE

The screenshot shows the KBase website homepage. At the top, there is a navigation bar with links for About, Data & Tools, Research, Documentation, and Help. On the right side of the navigation bar are buttons for Sign Up and Sign In. A red banner at the top indicates "System maintenance in 2 days" and specifies the time as "Wed Sep 25 from 2:00pm to 4:00pm". The main content area features a large image of a grassy field under a blue sky. Overlaid on this image is the KBase logo, which consists of three overlapping circles in yellow, green, and blue, with the word "KBase" written across them. Below the logo, the text reads "A collaborative, open environment for systems biology of plants, microbes and their communities". A blue button labeled "Use KBase" is centered in the image. At the bottom of the main content area, the text "KBase: The U.S. Department of Energy Systems Biology Knowledgebase" is visible. Below this, a section titled "In KBase you can..." lists six functions, each accompanied by an icon:

- Search, organize and perform large-scale analysis of biological data
- Assemble and annotate microbial genomes
- Build and validate metabolic models
- Analyze RNA-seq and expression data
- Carry out comparative and phylogenetic analysis
- Make your research reproducible, accessible and

<https://kbase.us/>

Automatic annotator - kBASE



<https://kbase.us/>