Genome Annotation Pipeline in PATRIC

Marcus Nguyen

What is happening during your annotation job?

RAST tool kit customized for PATRIC

OPEN

SUBJECT AREAS:
COMPARATIVE GENOMICS
BIOINFORMATICS

Received
12 November 2014

Accepted
2 January 2015

Published
10 February 2015

Correspondence and requests for materials should be addressed to J.J.D. (jimdavis@uchicago.edu)

RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes

Thomas Brettin[1,2], James J. Davis[1,2], Terry Disz[3], Robert A. Edwards[4,5], Svetlana Gerdes[1,3], Gary J. Olsen[6], Robert Olson[2,4], Ross Overbeek[1,3], Bruce Parrello[1,3], Gordon D. Pusch[1,3], Maulik Shukla[7], James A. Thomason III[8], Rick Stevens[1,2,9], Veronika Vonstein[1,3], Alice R. Wattam[7] & Fangfang Xia[2,4]

[1]Computing, Environment and Life Sciences, Argonne National Laboratory, Argonne IL, 60439, USA, [2]Computation Institute, University of Chicago, Chicago, Illinois, 60637, USA, [3]Fellowship for Interpretation of Genomes, Burr Ridge, IL, 60527, USA, [4]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, 60439, USA, [5]Department of Computer Science, San Diego State University, San Diego, California, 92182, USA, [6]Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA, [7]Virginia Bioinformatics Institute, Virginia Tech University, Blacksburg, VA, 24060, USA, [8]USDA-ARS Laboratory at Cold Spring Harbor Laboratory, Cold Spring Harbor NY, 11724, USA, [9]Department of Computer Science, University of Chicago, Chicago, Illinois, 60637, USA.

The RAST (Rapid Annotation using Subsystem Technology) annotation engine was built in 2008 to annotate bacterial and archaeal genomes. It works by offering a standard software pipeline for identifying genomic features (i.e., protein-encoding genes and RNA) and annotating their functions. Recently, in order

What is happening during your annotation job?

- Calling rRNAs (16S, 23S, 5S)
- Calling tRNAs with tRNAscanSE
  - (Lowe & Eddy 1997)
- Searching for repeat regions
- Finding special proteins
  - Selenoproteins
  - Pyrrolysylproteins
- Calling CRISPRs
  - clustered regularly interspaced short palindromic repeats

What is happening during your annotation job?

- Calling protein-encoding genes
  - Prodigal (Hyatt et al. 2010)
  - Glimmer3 (Delcher et al. 2007)
- Assigning function
  - First attempt: annotates against CoreSEED
  - Second attempt: annotates against FIGFams
  - Third attempt: BLAST against close relatives
- Overlapping genes are resolved

What is happening during your annotation job?

- Annotates matches to:
  - ARDB (Liu & Pop 2009)
  - CARD (McArthur et al. 2013)
  - VFDB (Chen et al. 2012)
  - Victors (Xiang et al. 2007)
  - PATRIC virulence factors (Mao et al. 2014)
  - DrugBank (Law et al. 2014)
  - TTD (Qin et al. 2014)
  - Human homologs
- Assigns proteins to families
- Finds closest neighbors

AMR Predictions

- SIR prediction based on AdaBoost models
- Only models > 70% accuracy run
- Limits genera that can be predicted
  - Based on available SIR data
  - Lots of resistant genomes
  - Few suseptible

What Genomes Will Have AMR Annotations?

- Acinetobacter baumanni
- Klebsiella pneumoniae
- Mycobacterium tuberculosis
- Peptoclostridium difficile
- Pseudomonas aeruginosa
- Staphylococcus aureus
- Streptococcus pneumoniae

Questions Comments?

- If not, let's look at some annotations

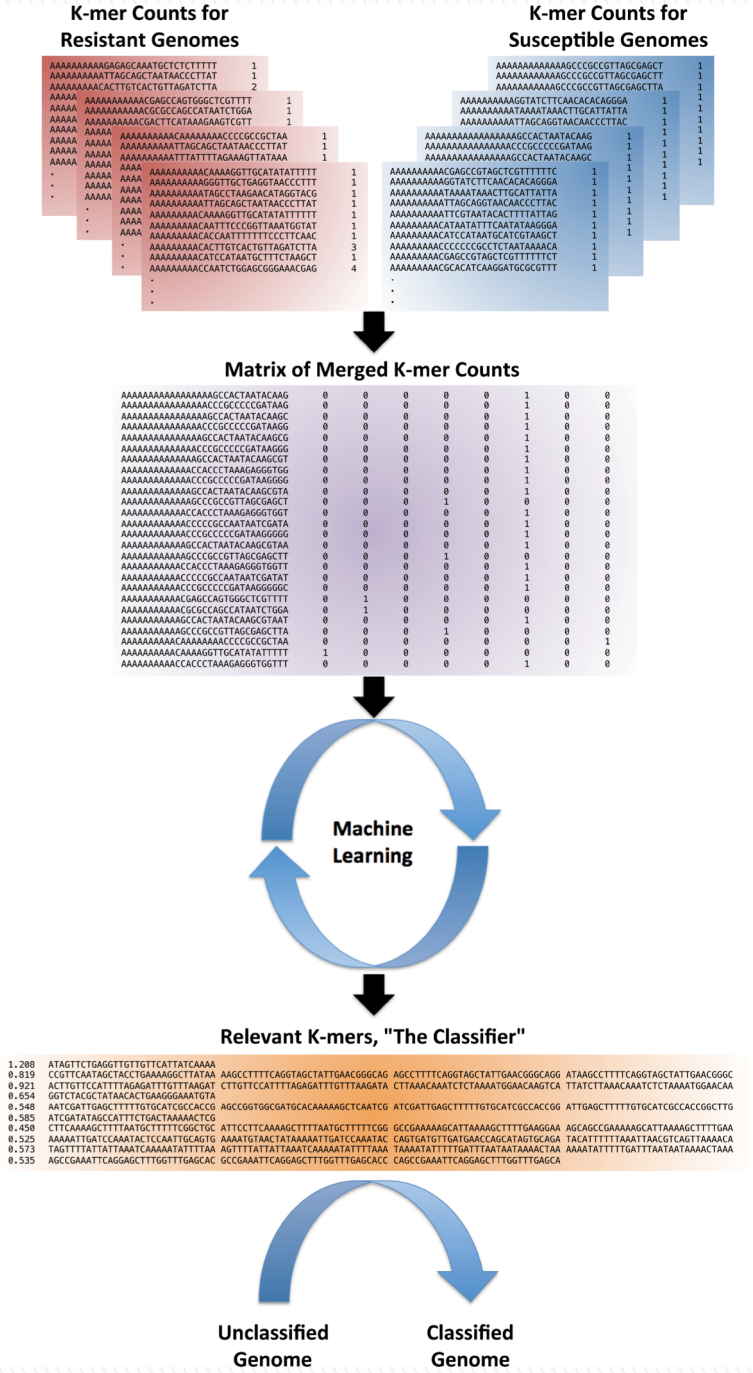
Extra Slides

How Do the Models Work?

- Machine learning
- Give computer
  - Lots of data about genomes
  - And label for genome (S or R)
- Computer finds correlations
  - Between data and label
  - Predict label in unseen genomes

Our Approach

- Give computer
  - Contig 15-mers
  - S or R label
- Computer finds
  - 15-mers related to S or R
  - Uses machine learning technique Adaboost
- Take top 10 15-mers
  - Make S or R prediction



Adaboost

- Stands for *adaptive boosting*
- For each k-mer
  - Sees which k-mer accurately predicts S or R
- Selects best k-mer
- Loop
  - Select best k-mer
    - Predicts well what previous could not

Adaboost Example

15-mer	% S	% R
AATCGACTAA…	0.75	0.25
AATCGCCGTT…	0.05	0.95
ATATGGCATA…	0.45	0.55
ATATATTACG…	0.76	0.24
TTGACAGATA…	0.33	0.67
CGTAGACTAG…	0.11	0.89
TGACATACCA…	0.72	0.28
GTACTACCCA…	0.50	0.50
CGTACCGACT…	0.62	0.38
GATAGATCCG…	0.77	0.23
GATTAAGGCC…	0.20	0.80

15-mer list

Selected 15-mers

Adaboost Example

15-mer	% S	% R
AATCGACTAA…	0.75	0.25
AATCGCCGTT…	0.05	0.95
ATATGGCATA…	0.45	0.55
ATATATTACG…	0.76	0.24
TTGACAGATA…	0.33	0.67
CGTAGACTAG…	0.11	0.89
TGACATACCA…	0.72	0.28
GTACTACCCA…	0.50	0.50
CGTACCGACT…	0.62	0.38
GATAGATCCG…	0.77	0.23
GATTAAGGCC…	0.20	0.80

15-mer list

- AATCGCCGTT…

Selected 15-mers

Adaboost Example

15-mer	% S	% R
AATCGACTAA…	0.75	0.25
AATCGCCGTT…	0.05	0.95
ATATGGCATA…	0.45	0.55
ATATATTACG…	0.76	0.24
TTGACAGATA…	0.33	0.67
CGTAGACTAG…	0.11	0.89
TGACATACCA…	0.72	0.28
GTACTACCCA…	0.50	0.50
CGTACCGACT…	0.62	0.38
GATAGATCCG…	0.77	0.23
GATTAAGGCC…	0.20	0.80

- AATCGCCGTT…
- GATAGATCCG…

15-mer list

Selected 15-mers

Adaboost Example

15-mer	% S	% R
AATCGACTAA…	0.75	0.25
AATCGCCGTT…	0.05	0.95
ATATGGCATA…	0.45	0.55
ATATATTACG…	0.76	0.24
TTGACAGATA…	0.33	0.67
CGTAGACTAG…	0.11	0.89
TGACATACCA…	0.72	0.28
GTACTACCCA…	0.50	0.50
CGTACCGACT…	0.62	0.38
GATAGATCCG…	0.77	0.23
GATTAAGGCC…	0.20	0.80

15-mer list

- AATCGCCGTT…
- GATAGATCCG…
- ATATATTACG…

Selected 15-mers

Adaboost Example

15-mer	% S	% R
AATCGACTAA…	0.75	0.25
AATCGCCGTT…	0.05	0.95
ATATGGCATA…	0.45	0.55
ATATATTACG…	0.76	0.24
TTGACAGATA…	0.33	0.67
CGTAGACTAG…	0.11	0.89
TGACATACCA…	0.72	0.28
GTACTACCCA…	0.50	0.50
CGTACCGACT…	0.62	0.38
GATAGATCCG…	0.77	0.23
GATTAAGGCC…	0.20	0.80

15-mer list

- AATCGCCGTT…
- GATAGATCCG…
- ATATATTACG…
- CGTAGACTAG…
- …

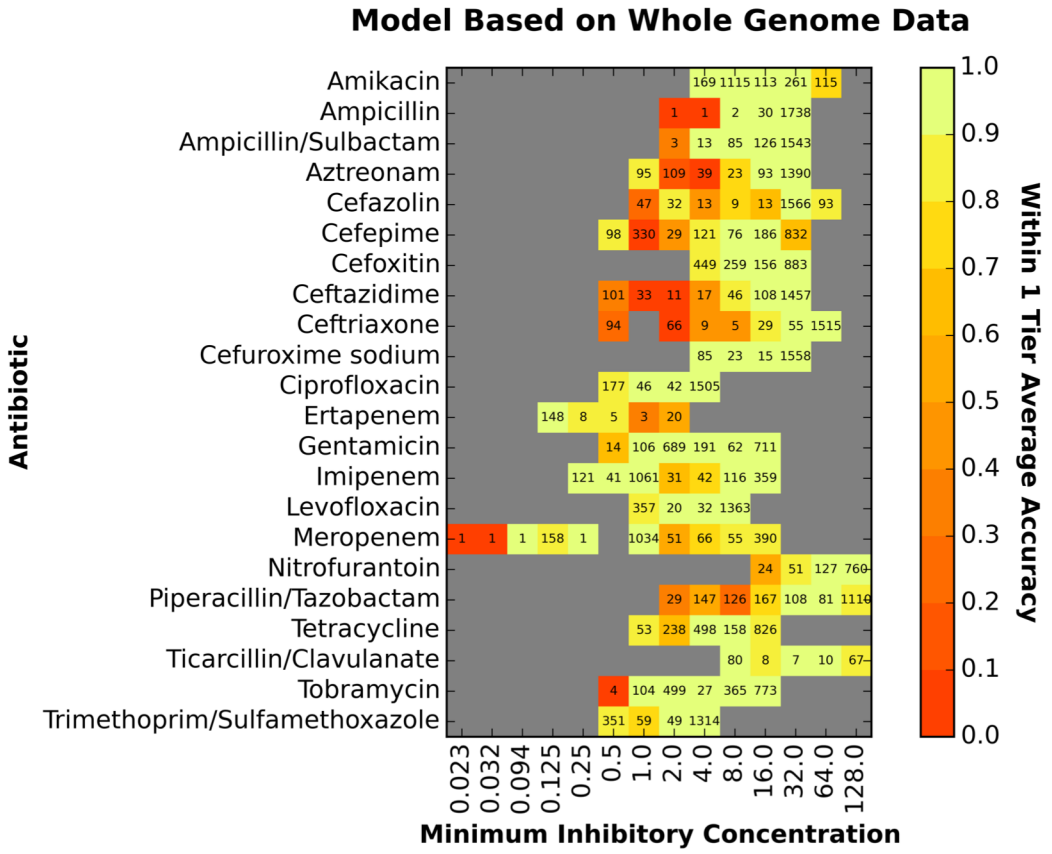
Selected 15-mers

Once 15-mers selected

- Each 15-mer "votes"
  - Susceptible
  - Resistant
- Most "votes" = predicted label
  - If genome has more top-10 resistant k-mers
    - Labeled resistant
  - If genome has more top-10 susceptible k-mers
    - Labeled susceptible

Future Work (predicting MIC)

- Given genome, antibiotic, MIC
  - Train model Using 10-mers
  - Predict MIC
- Building model for *Klebsiella Pneumoniae*
  - Uses gradient boosted trees
  - Overall accuracy (93%)
  - Varies across MIC values and antibiotics



Future Work (predictions with reads)

- Predict AMR using raw reads
  - Susceptibility vs Resistance
  - MIC?
- Clinical setting idea
  - Use MinION
  - Feed reads to model
  - Predict AMR (S, I, R, MIC, etc.)