max planck institut informatik

Computational methods for analyzing metagenomics data

2012 Workshop on Genomics

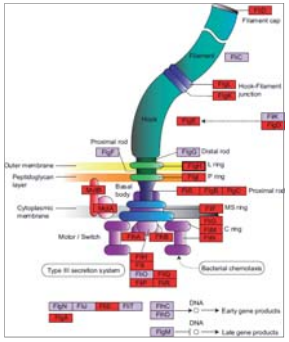
Alice C. McHardy

Heinrich Heine University Düsseldorf

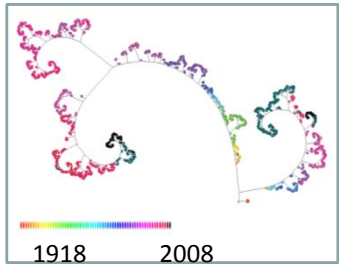
Max Planck Institute Informatics

Methods Overview

- Taxonomic assignment of metagenome sequences



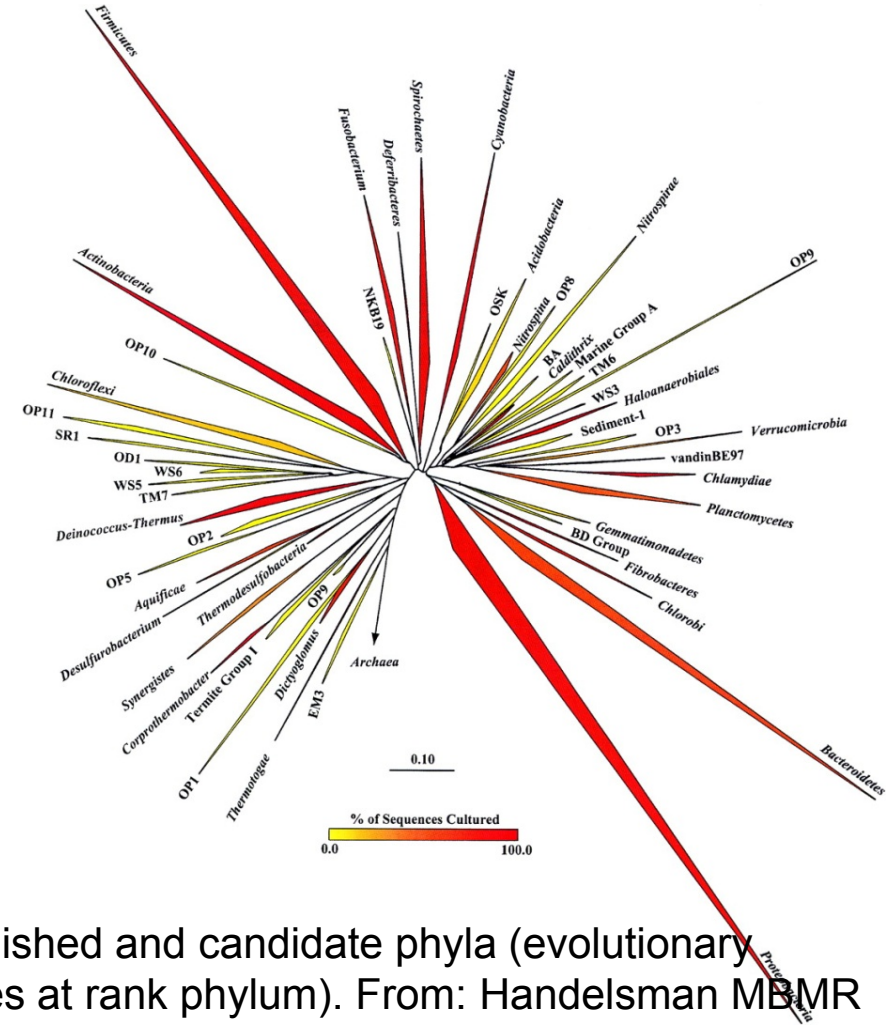
- Inference of functional and phenotypic relationships for gene families



- Searching for the imprint of selection

Metagenomics: Insight into the uncultured world

- Insight into little studied phyla
- Novel genome reconstruction
- Study of communities as a whole
- Detection of antibiotic resistance genes, degradative enzymes...



Established and candidate phyla (evolutionary classes at rank phylum). From: Handelsman MBMR 2004

Genome Analysis

Sequencing

GAGAGATATTTTAGACCCACGATGATCC
AGCGATTTGACACGATGATAG
AGCGATTTGACACGGAGAGTATAGAA
AGCGATTTGTAGTATATCGAGGGG

Assembly

Finishing

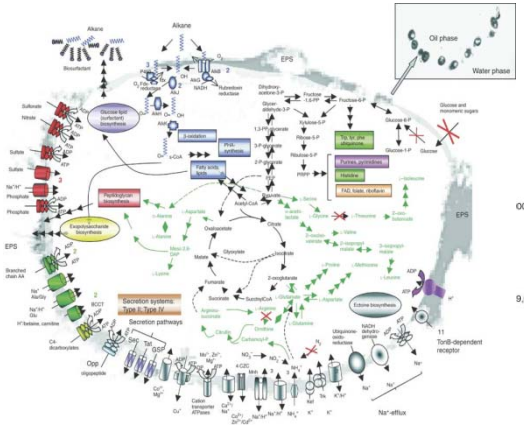
[...]sdATGACGATTCCGAAACACGAGCCGCGCGAGGTCTTCGATCGGGCGATCGAGCATACGCGGGCGCTTTCCC
CGGCCACTTTCGATCAGTGGTTCGGGGGAGTTCAGTTCGATGATCTGACCGACGGCGTCCTGACGCTGCG
CGAGCAGCAGCGCGCCATAGCGCGCCAGCACGTCGCGGGCCCGCGCCATGCGCCTCGCCGCGTCGTCCTC
GGATCCCGTGAAGATGAGCACGAGCGCCGTGACCTCGCGCCCGCTCGCCGAGCCCGCGCGCGGCCGGGCCG
CTGTCGGCCTTCTCGAGCAGCTGGAGGAGCCCCGACGCCGAGCGGCTGCGCGTGCTCGTCGTGTGCGGCA
CCTCGACGGGGGTGCGCTCGTTGGCCCCGTTCGGCTCCTCTTGGAGCACGGCGCTCGCCTGGATCTCGGC
GCTCTTCTCCTCGTGGAACGGCTCCAGGAAATCGGCGAGGTCGTTCGCCCCGAACGGCTCGGCCGTCGCG
TAGAAGAGCCCAGGAGCGCCTCGTGCAGCCGCCCGGCGTCGGGGAACCGATCCTCCGGGTTCTTCGCGA
GCATCGTGCTCACGACGTCGGCGAGCGACCGGGGCACGTCGGGGCGCACGAGCGCGAGCGGCGGATACTC
GCTCGCCTGCAGCCGCCGCATGATCTCGAACGCCGTCGGCGCCGCGAACGGGTTCGACCCGGCGAGCATC
GATGGTCGTGTTCCGCGGGATGAGCTTGAAGCTCACGCCGCCGCCGCGTCTCGACGCCGAGCGACAGCGGC
GTCACGTCGAGGAGCAGCACCTCGTCCACCTGCCCGGAGAGCGCGGCGCCCTGCAGCGCAGCGCCGACCG
CGACGACCTCGTCGGGGTTGACGCCCTTGTTCGGCTCCCGGCCGAAGAACTCGCGCACCGCCGCCTGCAC
CGCGGGCATCCGGGTCATGCCGCCGACGAGGACCACCGTGTTGACGGCGGAGACGGGGAGCTTGGCGTCC
CCGAGCGTCGCGCGACAGACGTCGATGGTCCGCCGGATGAGGCCCTCGCAGAGCATCTCGAGCTCGTTGC
GCCGCATCGTCCGCTCGAGGTGGAGCGGCCCTCCGCCCGGGCCGACGGCGATGAAGGGGATGTTGATCTC
GGTCTCGAGCGACGACGAGAGCTCGTGCTTCGCCTTCTCGGCCGCCTCCTTGAGCCGCTGCAGCGCCATG
CGATCCCGGCGCAGATCGATGCGGCTCTTCCCCTCGAACTCGTCGGCGAGCAGGTCGATGATCCGCTGGT
CGAAGTCCTCGCCGCCGAGGTGCGTGTCGCCGCCGGTCGCCTTGACGCTGAAGACCCCGCTCGCGATCTC
GAGGATCGAGATATCGAACGTGCCTCCGCCGAGATCGTAGACCGCGATCGTCTCGGCCTTCACCTTGTCG [...]

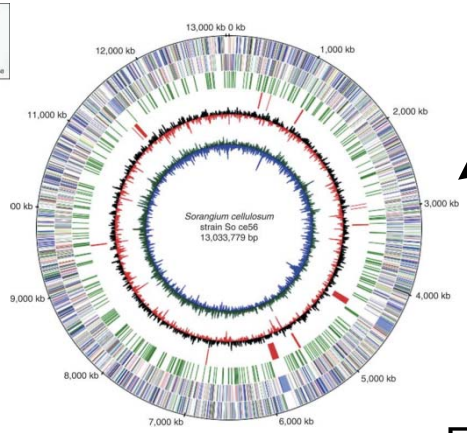
Annotation

Gene prediction

Functional annotation

Process level annotation





From: Schneiker *et al.* Nature Biotech. 2007 (left), 2006 (right)

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Metagenome Analysis

Sequencing

GAGAGATATTTTAGACCCACGATGATCC

AGCGATTTGACACGATGATAG

AGCGATTTGACACGGAGAGTATAGAA

AGCGATTTGTAGTATATCGAGGGG

Assembly

[...]sdATGACGATTCCGAAACACGAGCCGCGCGAGGTCTTCGATCGGGCGATCGAGCATACGCGGGCGCTTTCCC
CGGCCACTTTCGATCAGTGGTTCGGGGGAGTTCAGTTCGATGATCTGACCGACGGCGTCCTGACGCTGCG
CGAGCAGCAGCGCGCCATAGCGCGCCAGCACGTCGCGGGCCCGCGCCATGCGCCTCGCCGCGTCGTCCTC
GGATCCCGTGAAGATGAGCACGAGCGCCGTGACCTCGCGCCGCTCGC

CCTCGACGGGGGTGCGCTCGTTGGCCCCGTTCGGCTCCTCTTGGAGCACGGCGCTCGCCTGGATCTCGGC

GCTCTTCTCCTCGTGGAACGGCTCCAGGAAATCGGCGAGGTCGTTCGCCCCGAACCGCTCGGCCGTCGCG
TAGAAGAAGCCGAGGAGCGCCTCGTGCAGCCGCCCGGCGTCGGGGAACCGATCCTCCGGGTTCTTCGCGA
GCATCGTGCTCACGACGTCGGCGAGCGACCGGGGCACGTCGGGGCGCACGAGCGCGAGCGGCGGATACTC

GCTCGCCTGCAGCCGCCGCATGATCTCGAACGCCGTCGGCGCCGCGAACGGGTTCGACCCGGCGAGCATC
GATGGTCGTGTTCCGCGGGATGAGCTTGAAGCTCACGCCGCCGCCGGTCTCGACGCCGAGCGACAGCGGC

GTCACGTCGAGGAGCAGCACCTCGTCCACCTGCCCGGAGAGCGCGGCGCCCTGCAGCGCAGCGCCGACCG
CGACGACCTCGTCGGGGTTGACGCCCTTGTTCGGCTCCCGGCCGAAGAACTCGCGCACCGCCGCCTGCAC
CGCGGGCATCCGGGTCATGCCGCCGACGAGGACCACCGTGTTGACGGCGGAGACGGGGAGCTTGGCGTCC
CCGAGCGTCGCCGCGACAGACGTCGATGGTCCGCCGGATGAGGCCCTCGCAGAGCATCTCGAGCTCGTTGC
GCCGCATCGTCCGCTCGAGGTGGAGCGGCCCTCCGCCCGGGCCGACGGCGATGAAGGGGATGTTGATCTC

GGTCTCGAGCGACGACGAGAGCTCGTGCTTCGCCTTCTCGGCCGCCTCCTTGAGCCGCTGCAGCGCCATG
CGATCCCGGCGCAGATCGATGCGGCTCTTCCCCTCGAACTCGTCGGCGAGCAGGTCGATGATCCGCTGGT
CGAAGTCCTCGCCGCCGAGGTGCGTGTCGCCGCCGGTCGCCTTGACGCTGAAGACCCCGCTCGCGATCTC
GAGGATCGAGATATCGAACGTGCCTCCGCCGAGATCGTAGACCGCGATCGTCTCGGCCTTCACCTTGTCG [...]

Taxonomic characterization

[...]sdATGACGATTCCGAAACACGAGCCGCGCGAGGTCTTCGATCGGGCGATCGAGCATACGCGGGCGCTTTC
CC CGGCCACTTTCGATCAGTGGTTCGGGGGAGTTCAGTTCGATGATCTGACCGACGGCGTCCTGACGCTGCG
CGAGCAGCAGCGCGCCATAGCGCGCCAGCACGTCGCGGGCCCGCGCCATGCGCCTCGCCGCGTCGTCCTC
GGATCCCGTGAAGATGAGCACGAGCGCCGTGACCTCGCGCCGCTCGC

CCTCGACGGGGGTGCGCTCGTTGGCCCCGTTCGGCTCCTCTTGGAGCACGGCGCTCGCCTGGATCTCGGC

GCTCTTCTCCTCGTGGAACGGCTCCAGGAAATCGGCGAGGTCGTTCGCCCCGAACCGCTCGGCCGTCGCG
TAGAAGAAGCCGAGGAGCGCCTCGTGCAGCCGCCCGGCGTCGGGGAACCGATCCTCCGGGTTCTTCGCGA
GCATCGTGCTCACGACGTCGGCGAGCGACCGGGGCACGTCGGGGCGCACGAGCGCGAGCGGCGGATACTC

GCTCGCCTGCAGCCGCCGCATGATCTCGAACGCCGTCGGCGCCGCGAACGGGTTCGACCCGGCGAGCATC
GATGGTCGTGTTCCGCGGGATGAGCTTGAAGCTCACGCCGCCGCCGGTCTCGACGCCGAGCGACAGCGGC

GTCACGTCGAGGAGCAGCACCTCGTCCACCTGCCCGGAGAGCGCGGCGCCCTGCAGCGCAGCGCCGACCG
CGACGACCTCGTCGGGGTTGACGCCCTTGTTCGGCTCCCGGCCGAAGAACTCGCGCACCGCCGCCTGCAC
CGCGGGCATCCGGGTCATGCCGCCGACGAGGACCACCGTGTTGACGGCGGAGACGGGGAGCTTGGCGTCC
CCGAGCGTCGCCGCGACAGACGTCGATGGTCCGCCGGATGAGGCCCTCGCAGAGCATCTCGAGCTCGTTGC
GCCGCATCGTCCGCTCGAGGTGGAGCGGCCCTCCGCCCGGGCCGACGGCGATGAAGGGGATGTTGATCTC

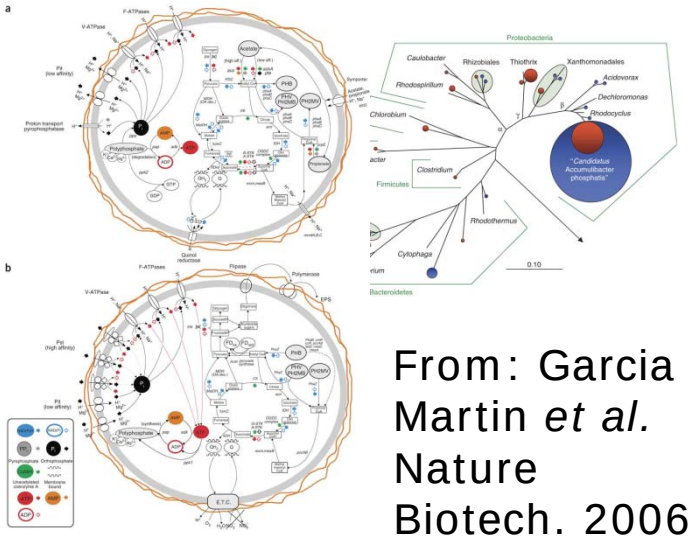
GGTCTCGAGCGACGACGAGAGCTCGTGCTTCGCCTTCTCGGCCGCCTCCTTGAGCCGCTGCAGCGCCATG
CGATCCCGGCGCAGATCGATGCGGCTCTTCCCCTCGAACTCGTCGGCGAGCAGGTCGATGATCCGCTGGT
CGAAGTCCTCGCCGCCGAGGTGCGTGTCGCCGCCGGTCGCCTTGACGCTGAAGACCCCGCTCGCGATCTC
GAGGATCGAGATATCGAACGTGCCTCCGCCGAGATCGTAGACCGCGATCGTCTCGGCCTTCACCTTGTCG [...]

Annotation

Gene prediction

Functional Annotation

Process-level Annotation

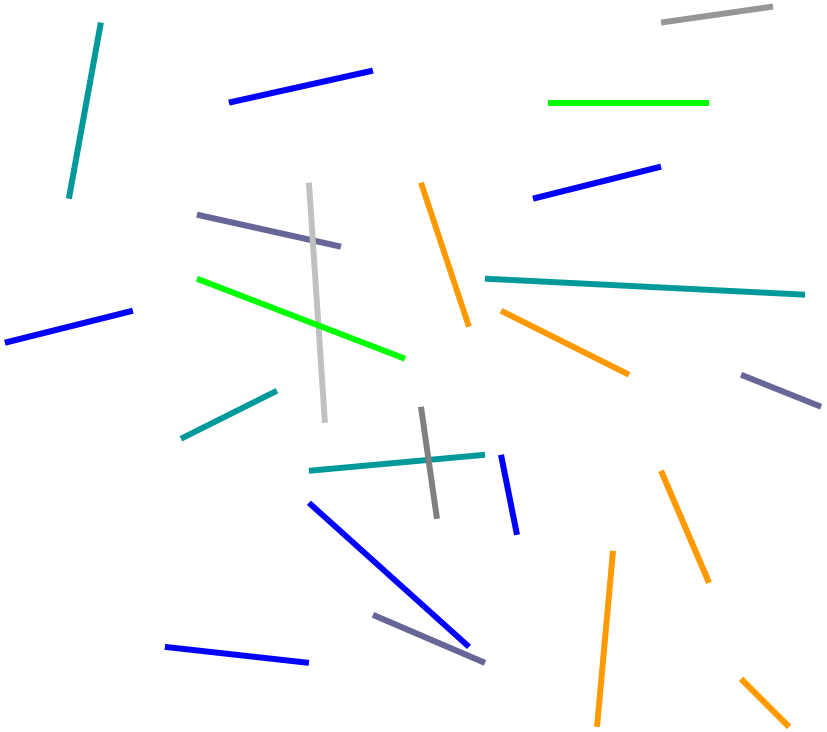


From: Garcia Martin *et al.* Nature Biotech. 2006

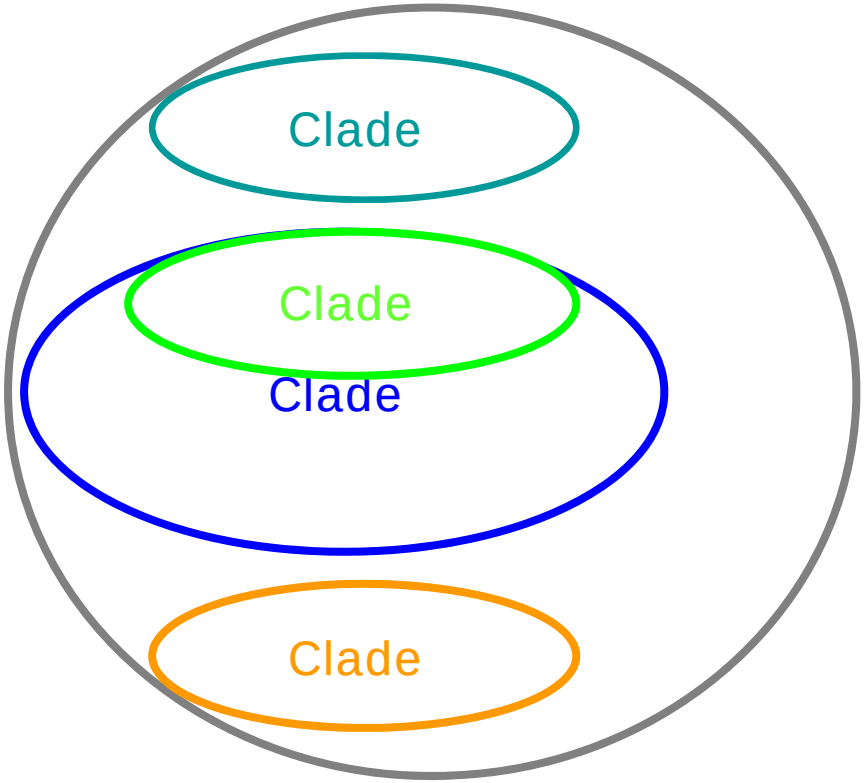
HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Problem: Taxonomic classification of metagenome sequences



Metagenome sequence fragments (contigs) of unknown origin.

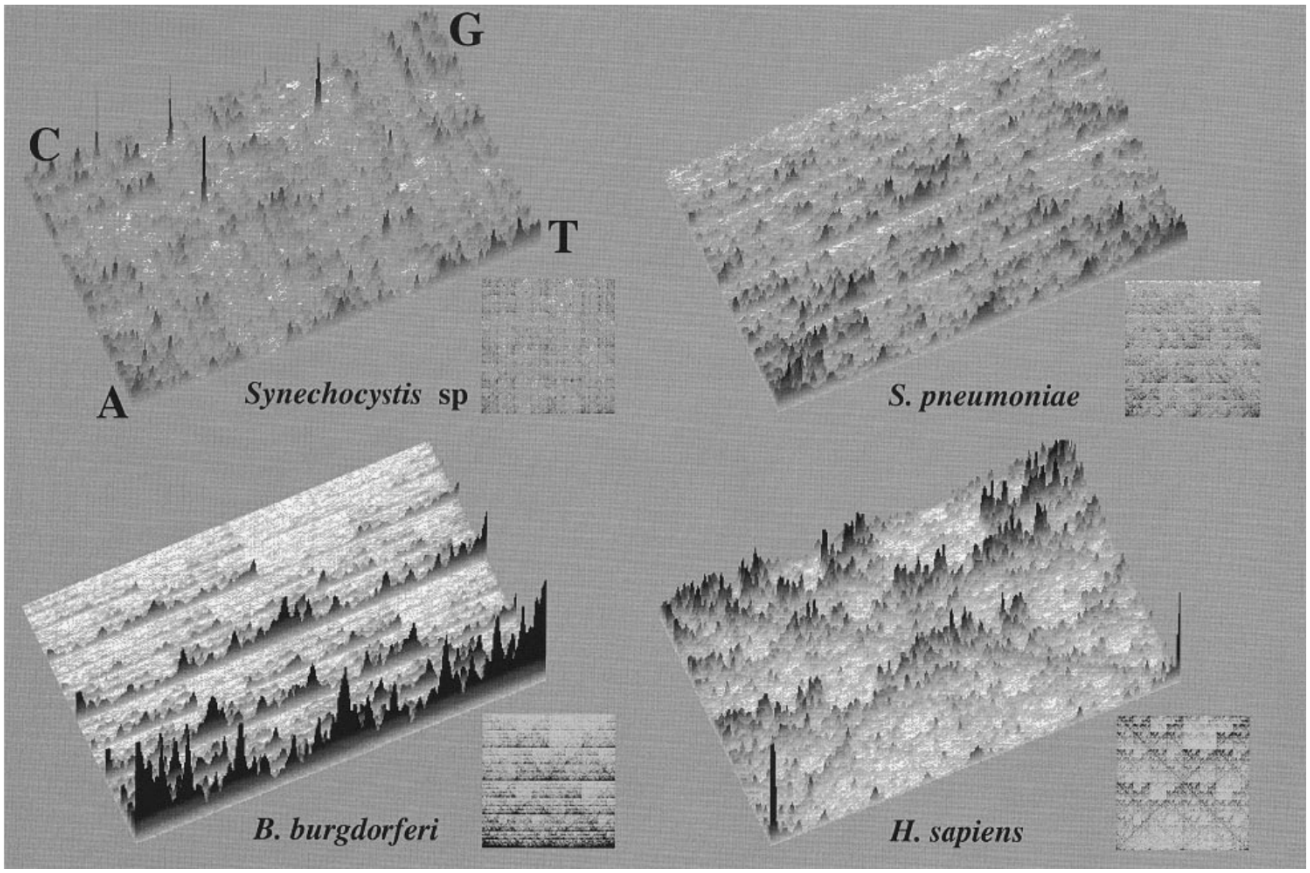


Taxonomic classes (clades) representing sample populations

Information used for taxonomic classification

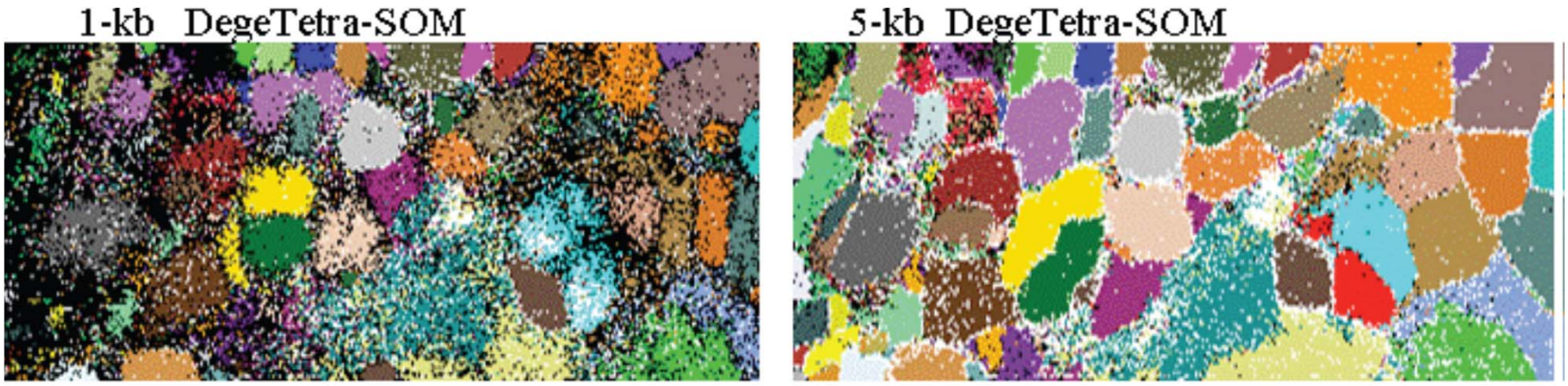
Technique	Information	Applicability
Conserved genes (Woese and Fox PNAS 1977)	Sequence conservation	< 1% of fragments
Clade-specific genes (Graham *et al.* PNAS 2000)	" "	< 1% of fragments
Sequence similarity (Huson et al. Genome Research 2011)	Sequence conservation	Dependent on available reference genomes
Genome signatures (Karlin & Burge, Trends Genet. 1995, Sandberg *et al.* Genome Research 2001)	Sequence composition	Fragments above length cut-off (method dependent)

The ‘genome signature’
(Karlin & Burge, Trends Genet., 1995)



3-D display of 7-letter frequencies for 100 kb segments of 4 organisms. *From*: Deschavanne *et al.*, *Mol. Biol. Evol.* 1999

Genome Signatures Visualized



Self-Organizing Maps of tetranucleotide usage for fragments of 81 prokaryotic genomes (coloring by species). *From*: Abe *et al. DNA Research* 2005

Molecular basis of genome signatures?

- DNA replication and repair mechanisms
- Context-dependent mutation biases
- Dinucleotide stacking stability
- Methylation modifications
- Avoidance of restriction enzyme sites (palindromic sequences)

Karlin *et al.* J. Bac 1997

Abundant Populations and observed Phylotypes in Different Metagenome Studies

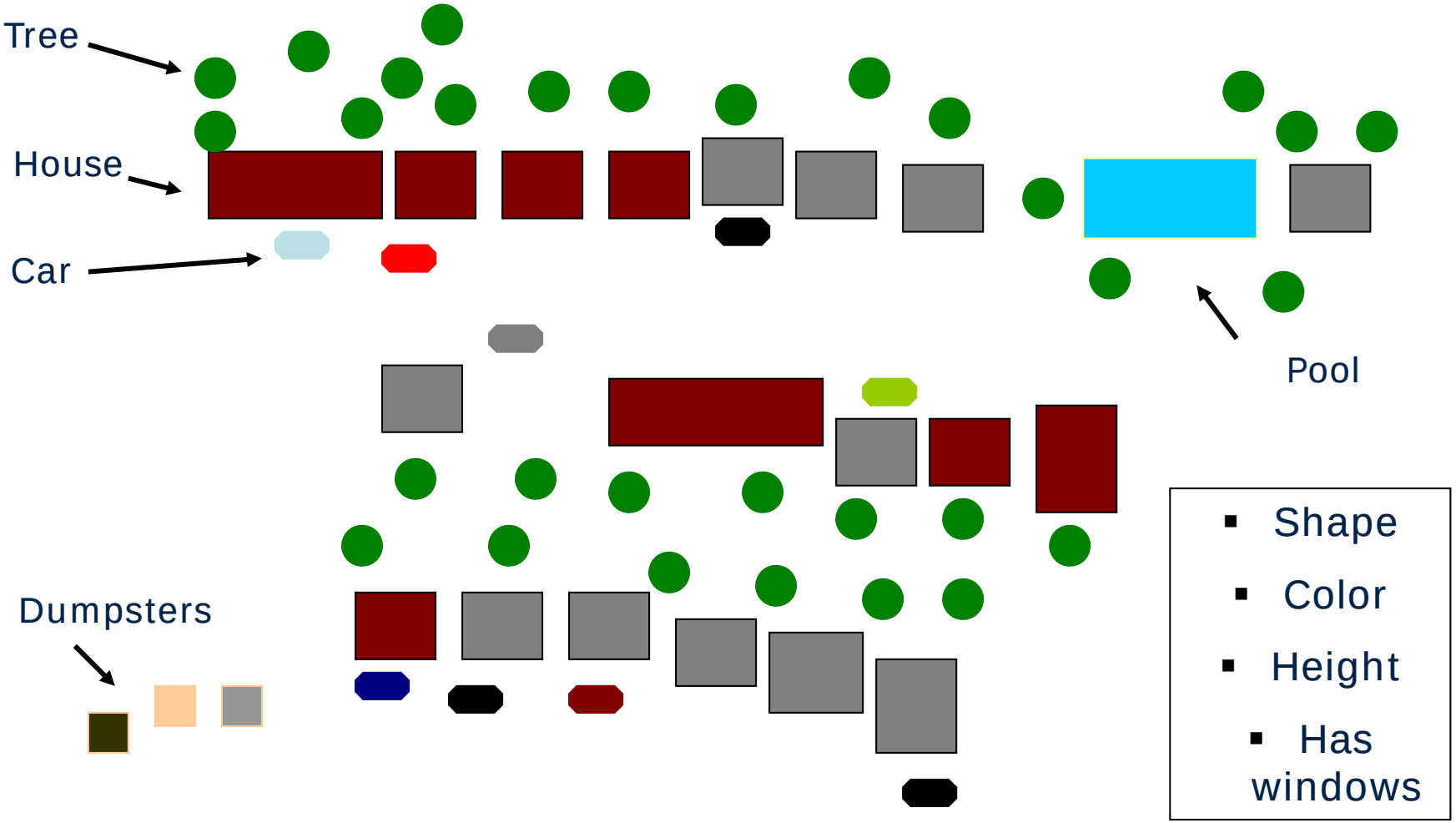
Sample	Dominant populations	Total
Acid mine drainage biofilm (Tyson *et al.* Nature 2004)	2	6
EBPR sludge (Garcia Martin *et al.* Nature Biotech. 2006)	1	>13
Wood-feeding higher termite Hindgut (Warnecke *et al* Nature 2007)	(24 phylotypes of 2 phyla)	~270
Minnesota soil (Tringe *et al.* Science 2005)	0	847

Clustering versus Classification

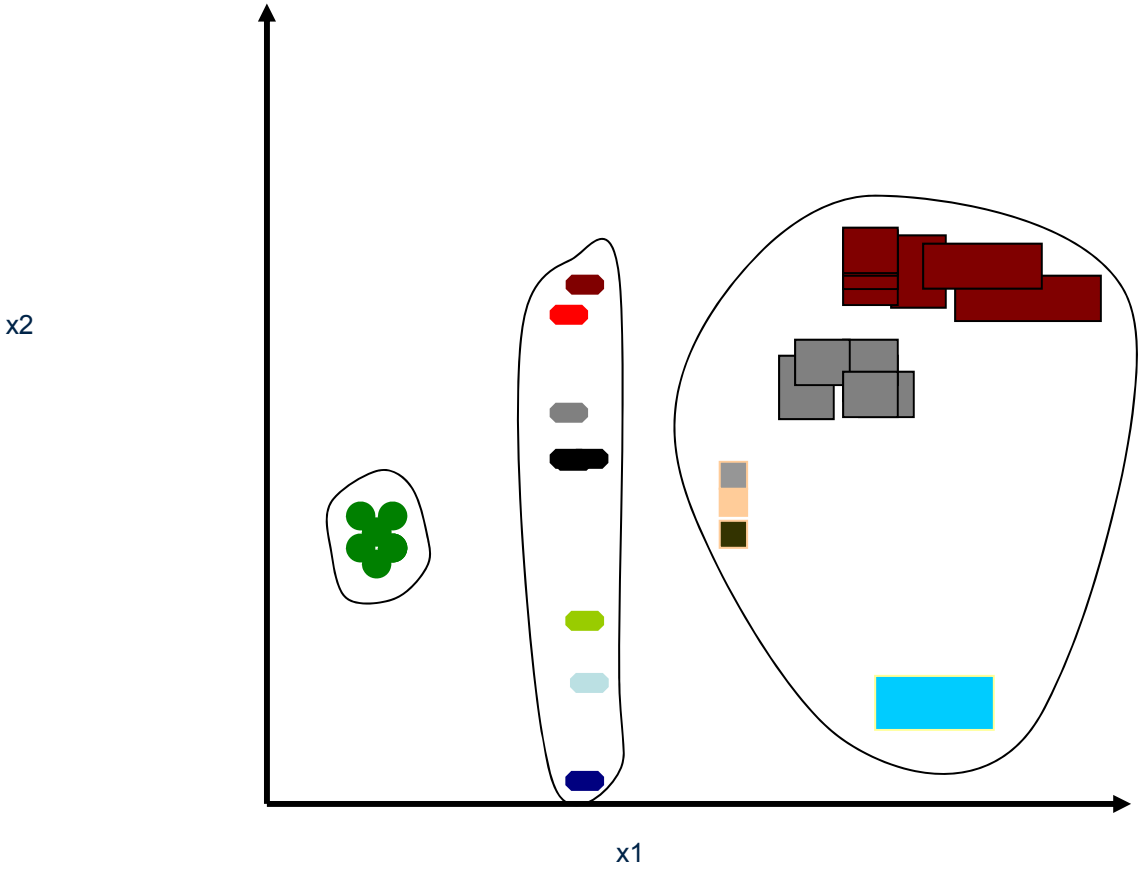
- Clustering
  - Exploratory data analysis
  - No knowledge of labels
  - Identify distinct structures in input space to partition data
  - Requires specification of number of clusters

- Classification
  - If training data is available (items of known origin: label)
  - Use labeled items to learn a model for classes
  - Classify data set with the model (assign items to classes)

Task: Classify items on a city map



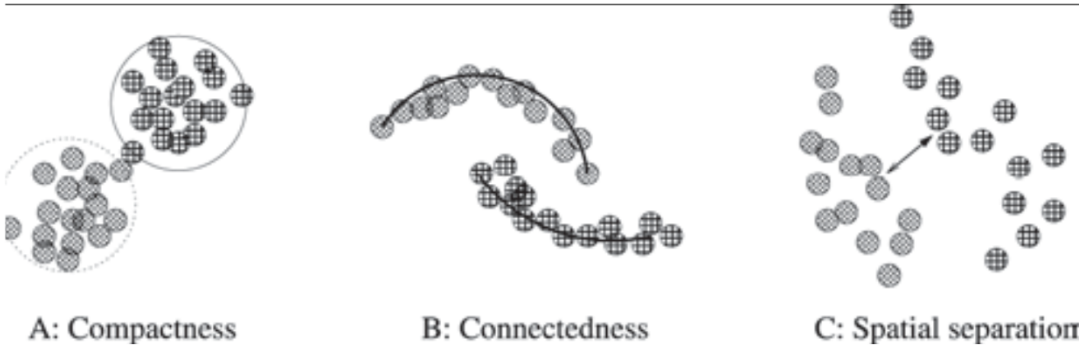
Clustering



Choice of k?

Clustering Criteria and Algorithms

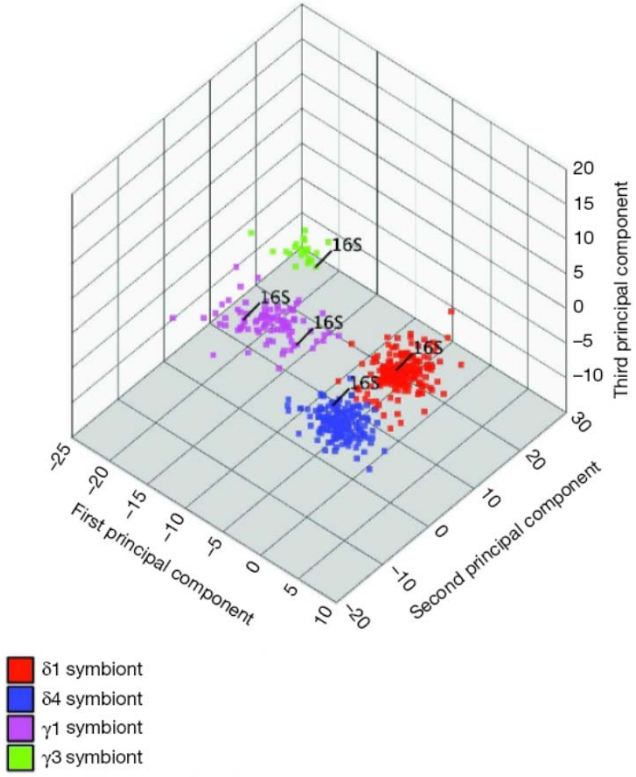
- Compactness (k-means, average link agglomerative clustering, Self-Organizing Map)
- Connectedness (density-based methods, single-link agglomerative clustering)
- Spatial separation (with compactness/balance of cluster sizes: optimization with simulated annealing, evolutionary algorithms)



From: Handl *et al.* Bioinformatics 2005

Clustering Methods for Metagenome Taxonomic Binning

- TETRA (Teeling et al. Env. Microbiol 200
- MetaClust (Woyke et al. Nature 2006)
- SOM (Abe et al. DNA Research 2005)
- CompostBin (Chatterji et al. RECOMB 20
- GSOM (Chan et al. BMC Bioinformatics 2



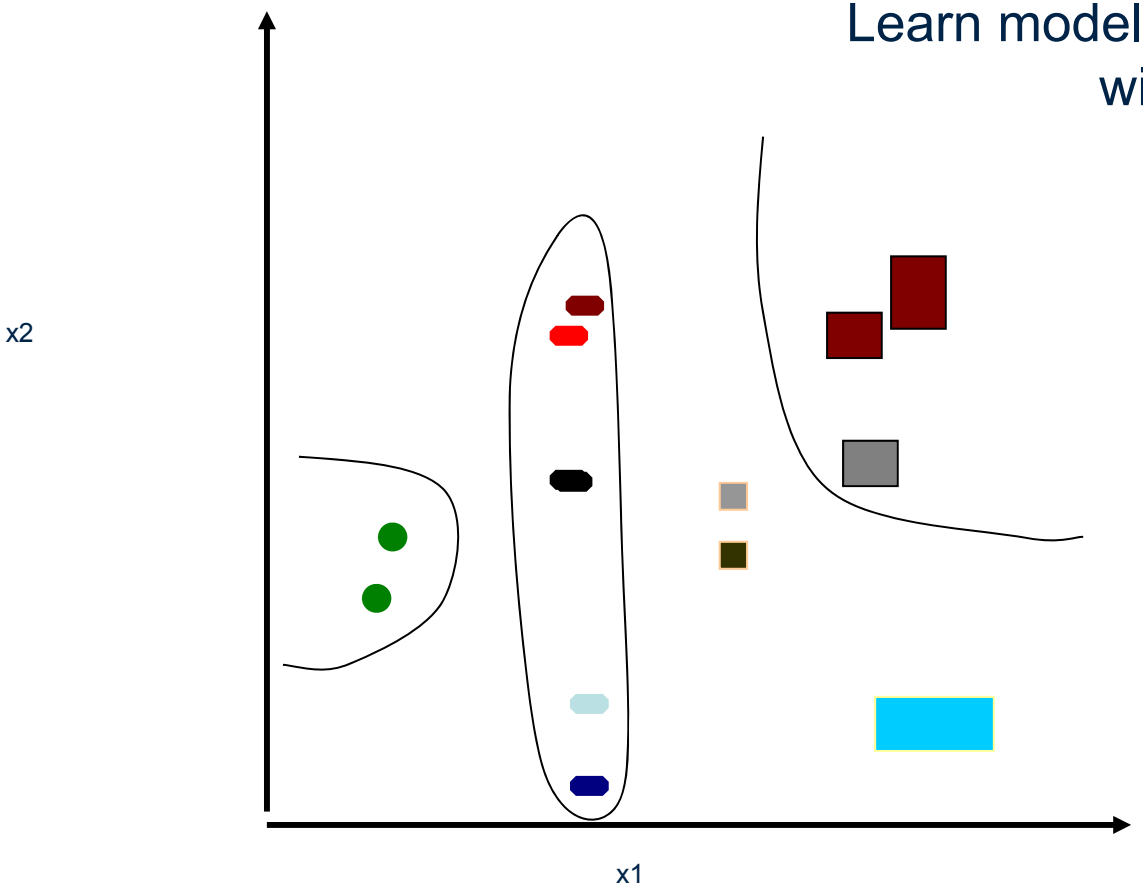
Clustering of the *O. algarvensis* symbiont scaffolds. *From*: Woyke *et al.* Nature 2006

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

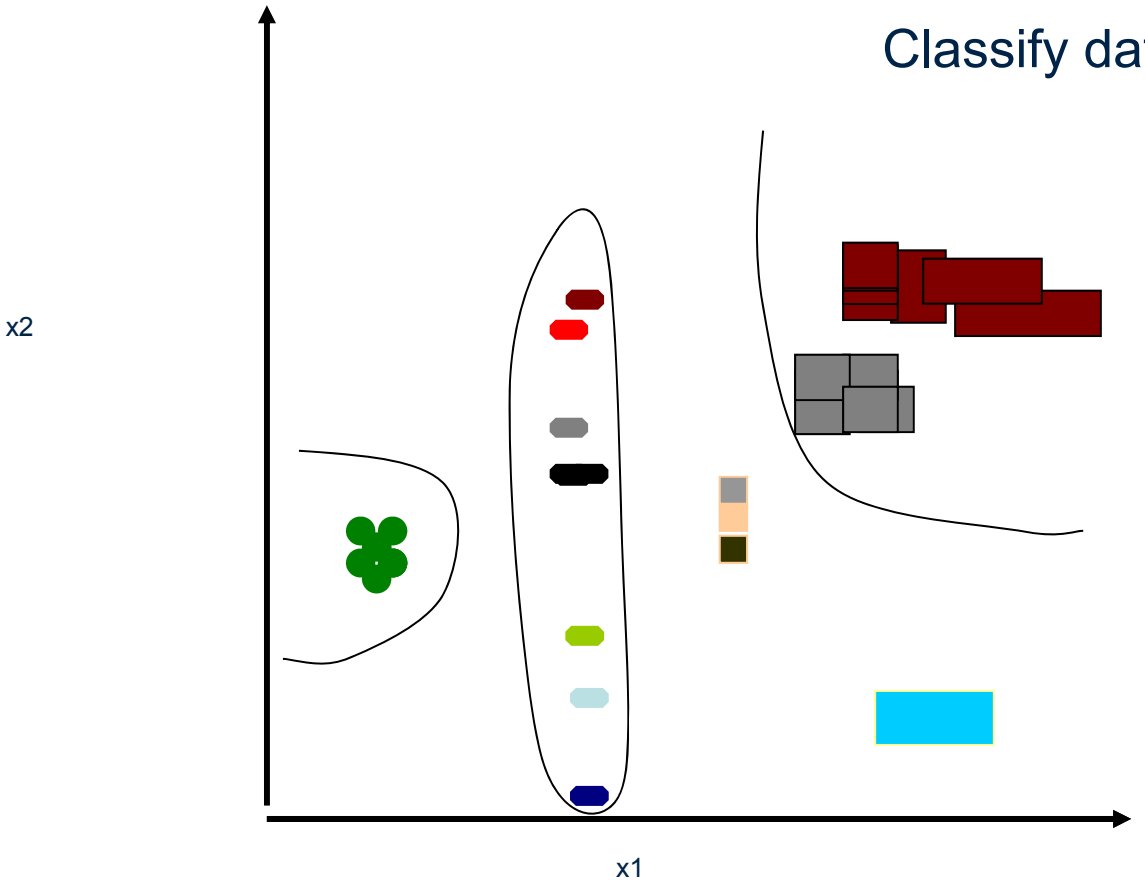
Classification

Learn model (houses, cars, trees)
with labeled data



Classification

Classify data with model



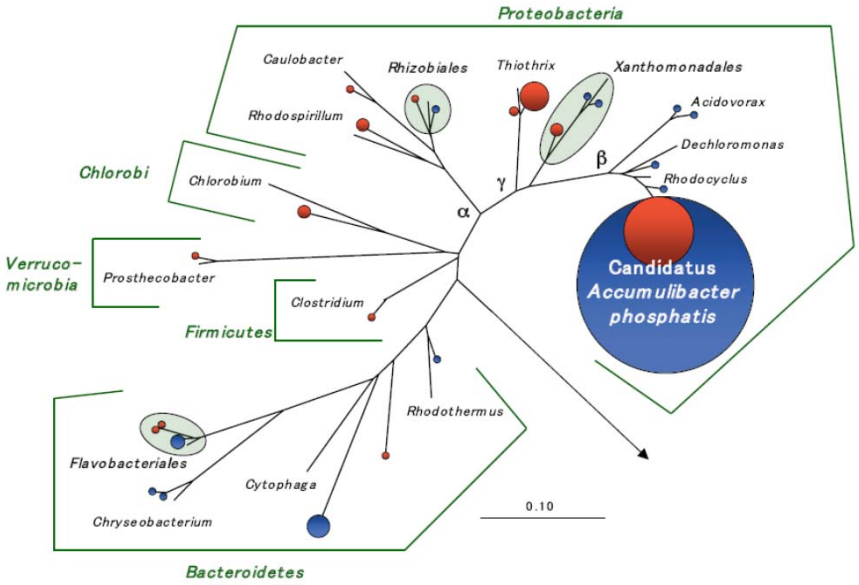
Modeling taxonomic fragment assignment as a classification problem

- **Why**? If good training data is available, classification is more accurate than clustering
- Accuracy is a function of sample complexity (# of taxa), fragment length, and availability of good reference data
- **Classes**: Phylogenetic clades (species for abundant populations and higher-level clades)
- **Training data**: Sample-derived (contigs with marker genes) and publicly available sequences

PhyloPythia (McHardy et al. Nature Methods 2007)

Tools for Taxonomic Metagenome Sequence Assignment (J. Droege, I. Gregor, K. Patil)

- *PhyloPythia*
  - Sequence composition-based Support Vector Machine model of evolutionary classes and populations in a microbial community
- *PhyloPythiaS*
  - Structural SVM
  - Faster
  - Web server for creating sample-specific models and assignment: *http://binning.bioinf.mpi-inf.mpg.de/*



From: Garcia Martin *et al.* Nature Biotech. (2006)

References:

McHardy et al. Nature Methods (2007)

Mavrommatis et al. Nature Methods (2008)

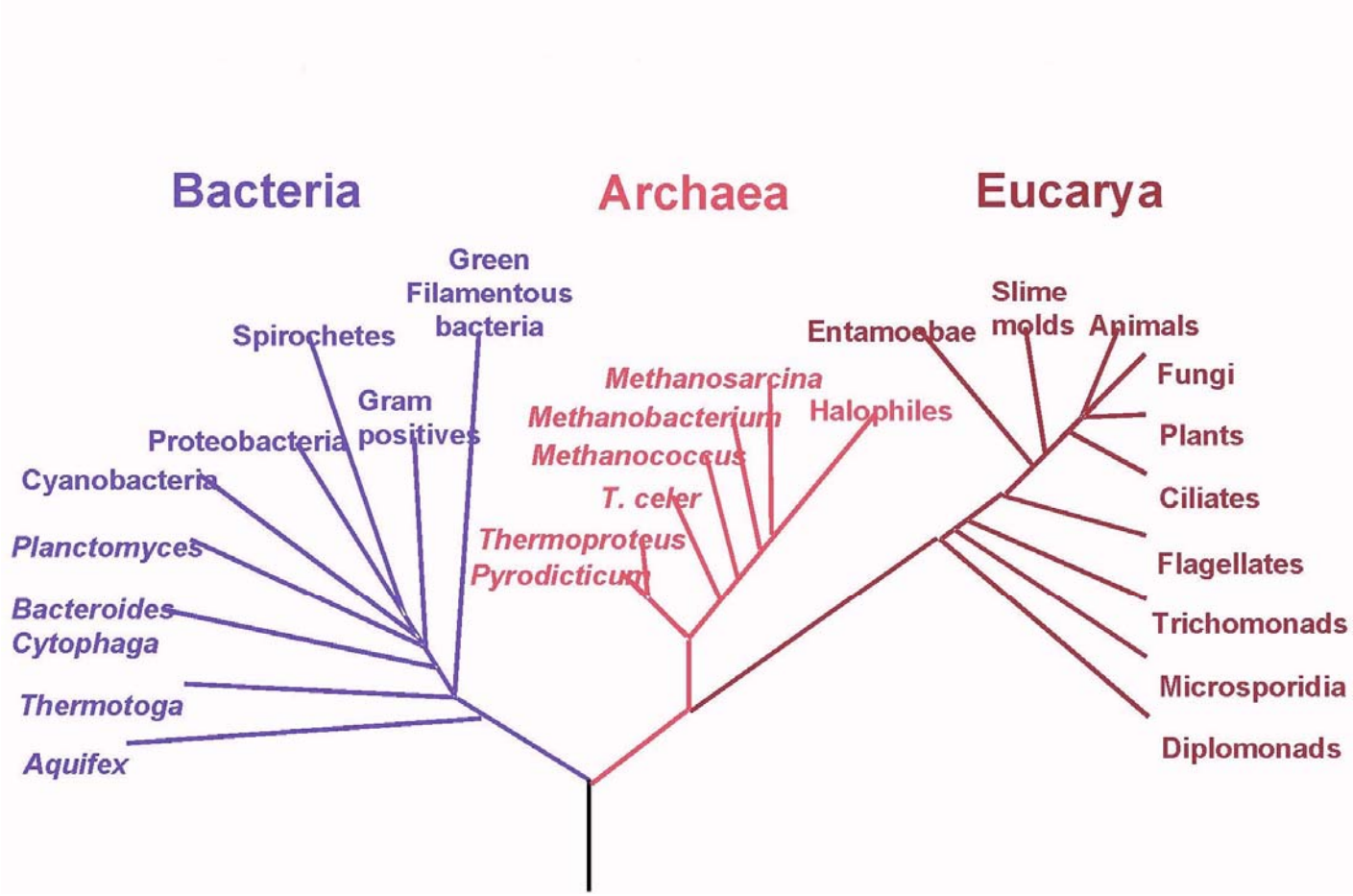
McHardy, Rigoutsos Curr. Opin. Microbiol. (2008)

Patil et al. Nature Methods (2011)

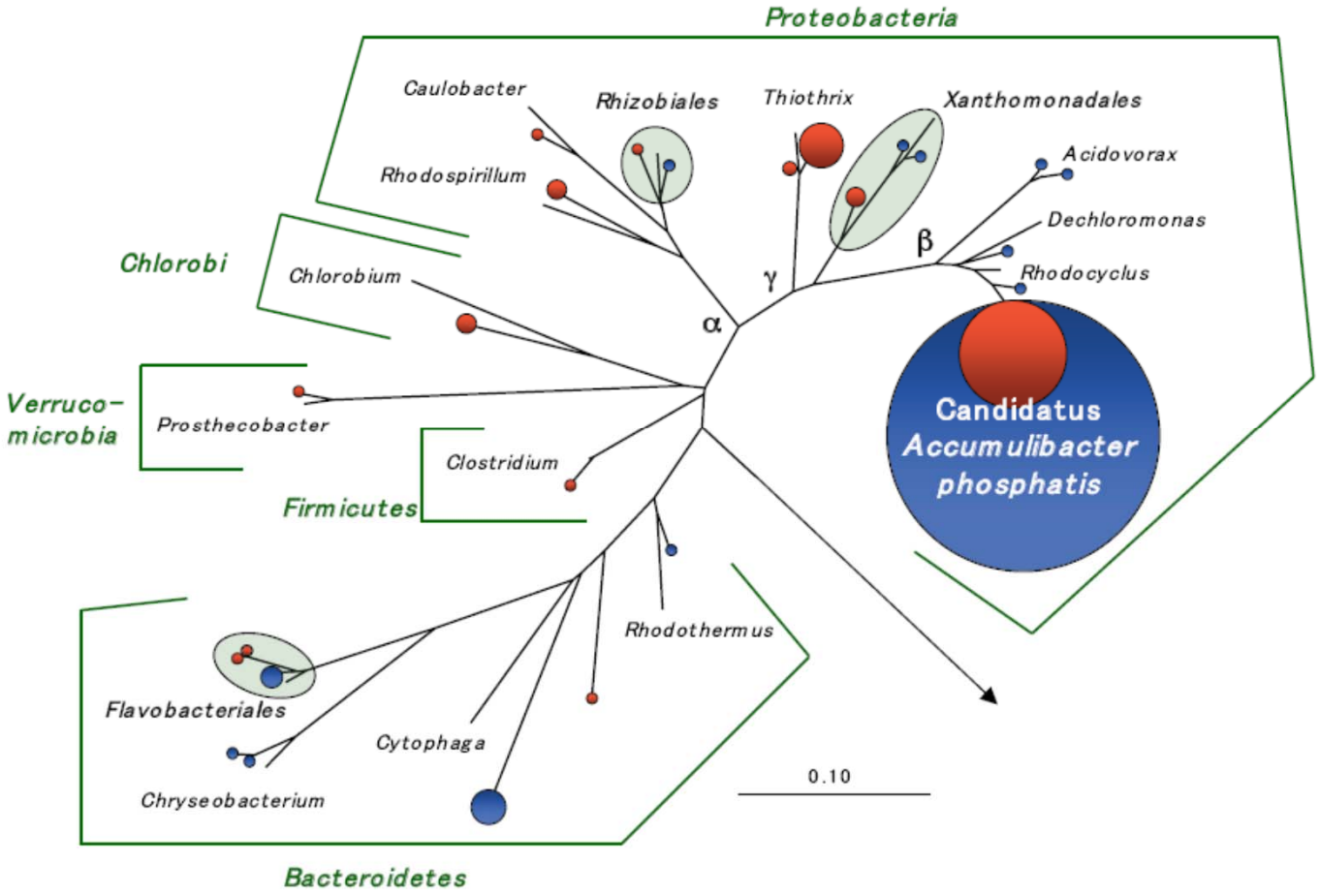
HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Reference System: Phylogenetic Taxonomy



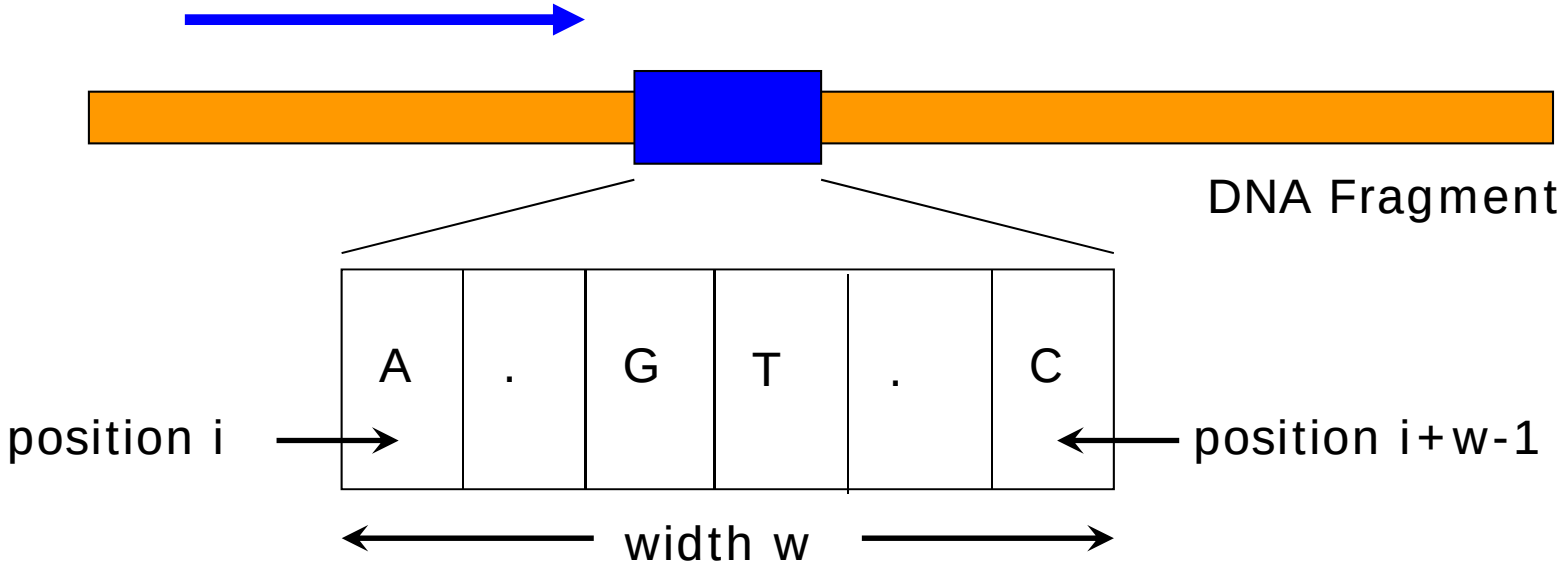
Clades to model: EBPR Sludge Metagenome



From: Garcia Martin *et al.* Nature Biotech. 2006

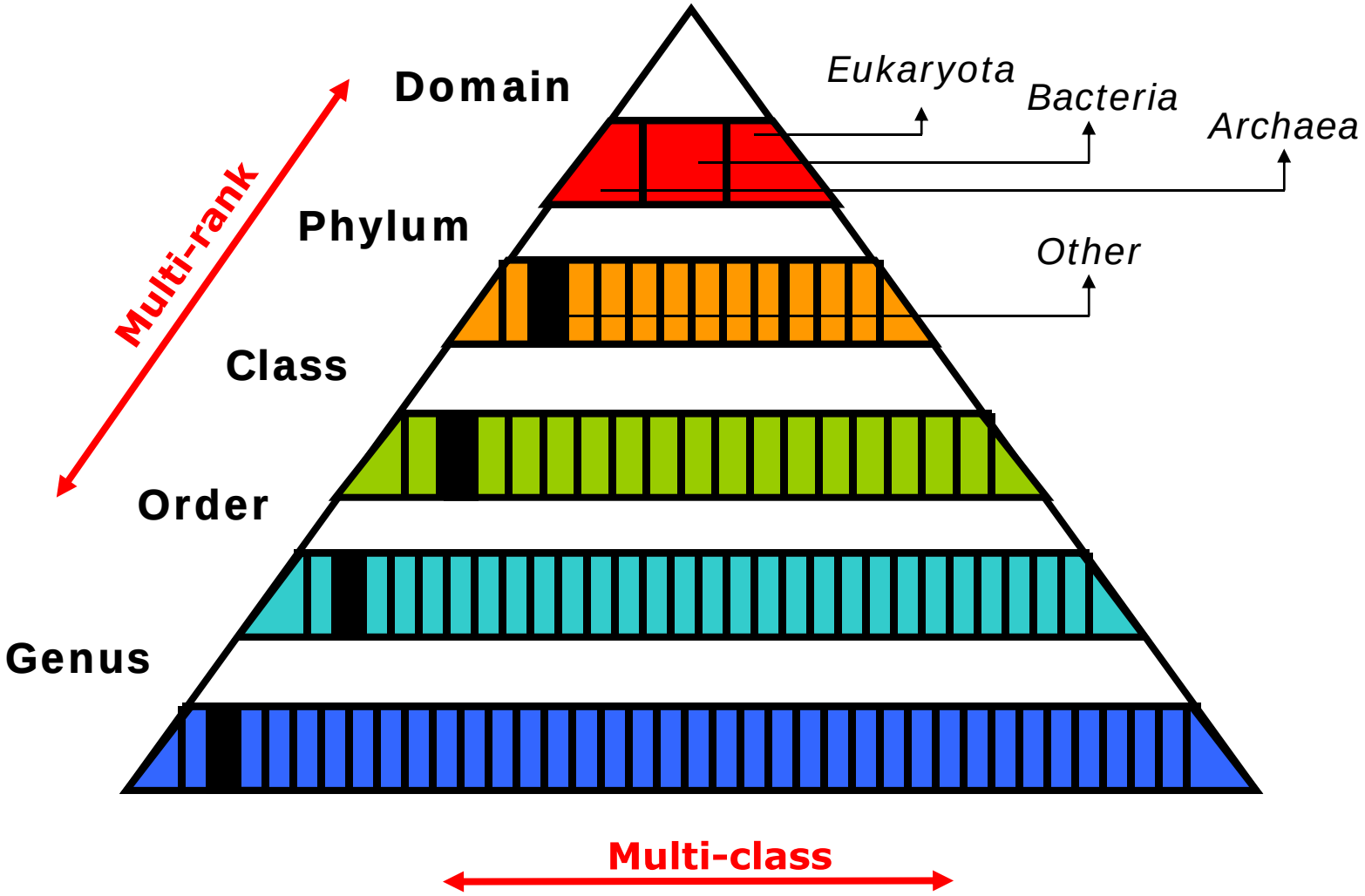
Input: DNA Fragments

- Map sequence to vector of sequence pattern frequencies
- Compositional signature of fragment



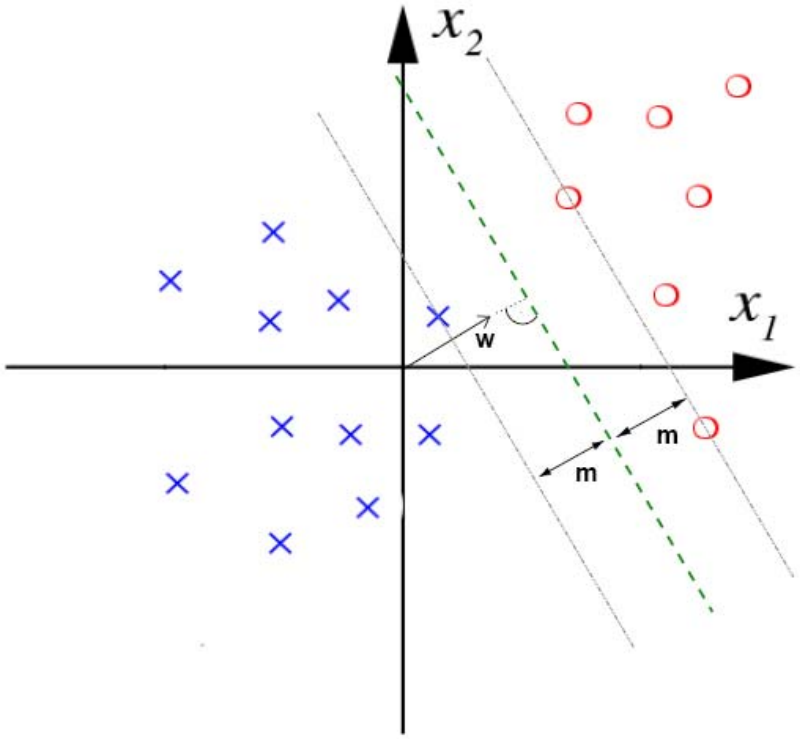
Exemplary <*w*, *l*>-pattern, where *w* is 6 and the number of literals *l* is 4.

Output: Taxonomic Assignments

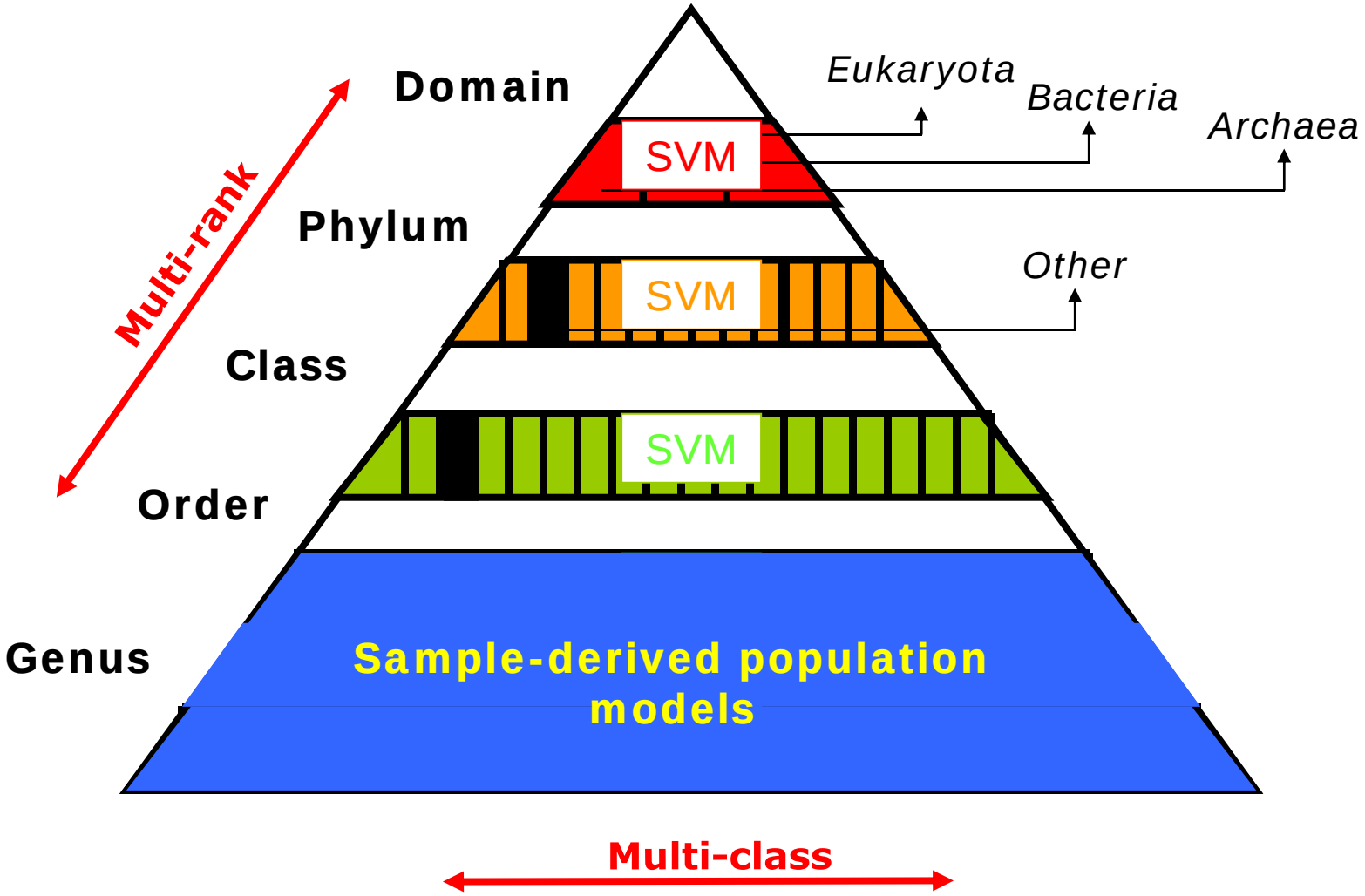


Classifier: Support Vector Machine

- Maximum margin classifier, soft margin: Good generalization ability. Sparse feature spaces, small training sets... (Boser et al. 1992, Cortes and Vapnik 1995)
- Kernel trick: Non-linear classifier
- Multi-class capable by extension to One-vs.-All (OVA), All-vs.-All (AVA), directed acyclic graph (DAG) architecture...

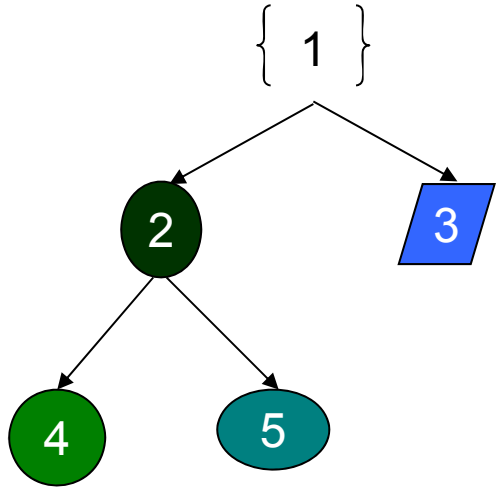


PhyloPythia Model Architecture



PhyloPythiaS Model Architecture

- Phylogenetic classification using structured output SVMs (Altun et al. 2003).
- Taxonomic hierarchy specifies relationships between clades (structure in the output space)
- Use knowledge of taxonomic relationships between clades
- Advantages over multi-class SVM
  - Training and testing times
  - Sparse models



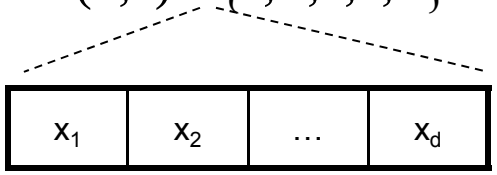
Output structure of taxonomic assignments

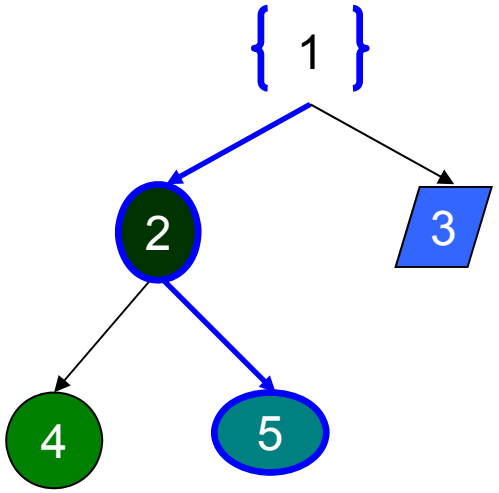
PhyloPythiaS: Joint feature space Ψ

- Map input-output pairs to a joint feature space Ψ
  - Generalization across output

$$\Lambda^T(5) = \{1,1,0,0,1\}$$

$$\Psi^T(x,5) = \{x, x, 0, 0, x\}$$





Output path in the tree (blue bordered)

$$F(x,5) : \langle w, \Psi(x,5) \rangle = \langle w_1, x \rangle + \langle w_2, x \rangle + \langle w_5, x \rangle$$

Compatibility score

Other Classifiers (list incomplete)

Composition-based and hybrids

- TACOA (Nearest Neighbor Classifier)
- PhymmBL (Interpolated Markov models + BLAST)
- NBC: Naïve Bayes Classifier
- Megan4 (Lowest common ancestor on blastx versus nr results, combined with composition-based NBC)

Similarity/domain searches

- CARMA (taxonomic analysis of Pfam domains)

EVALUATION

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

max planck institut
informatik

How to evaluate your results

- What do results in a method paper mean?
  - Many measures are used: Accuracy, Precision (Specificity), Recall (Sensitivity), ROC Curves

- How is a taxonomic assignment method performing on your data set?

  -Create your own reference: Identify and classify contigs in your data set with marker genes (16S, 23S rRNA, other conserved markers): Compare with predicted taxonomic assignments

  - Look at the consistency of taxonomic assignments for contigs belonging to the same scaffold

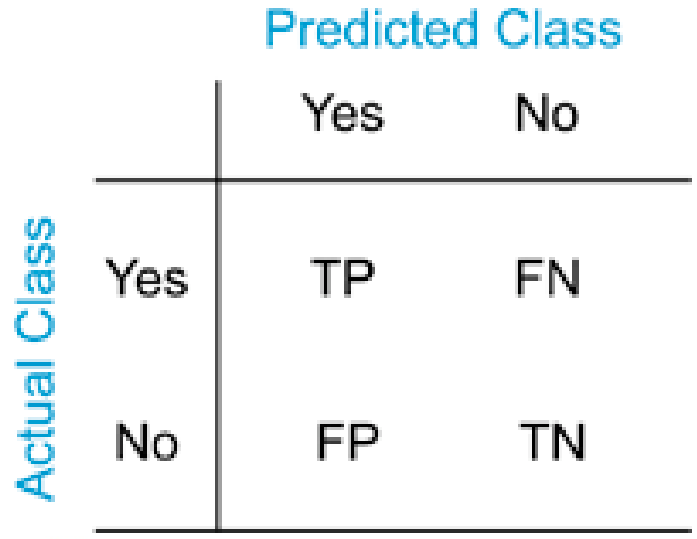
Evaluation with ‘standard of truth’

- Data set of N fragments of **known taxonomic origin** for evaluation:
  - Simulated data (create mixture of sequence fragments from multiple sequenced genomes)
  - Contigs that carry marker genes which you can reliably assign *based on other information than what the classifier uses*
- Assign with method

- Compare predicted taxonomic IDs with correct taxonomic IDs
  - True positive (**TP**) :Contig is of positive class, assigned positive class
  - False Negative (**FN**): Contig is of positive class, assigned to other class
  - True Negative (**TN**):Contig is of negative class, assigned to negative class:
  - False positive (**FP**): Contig is of negative class, assigned to positive class

Evaluation Measures

- Example: Community with only two species in it: Species 1 = Positive class; Species 2 = Negative class
- Evaluation works similarly for more than two species



- **Accuracy**: Fraction of overall data set classified correctly

  Acc = (TP + TN)/N

- **Sensitivity** (Recall): Fraction of a class that has been classified correctly:

  Sn = TP/(TP+FN)

- **Specificity** (Precision): Fraction of class assignments that are correct:

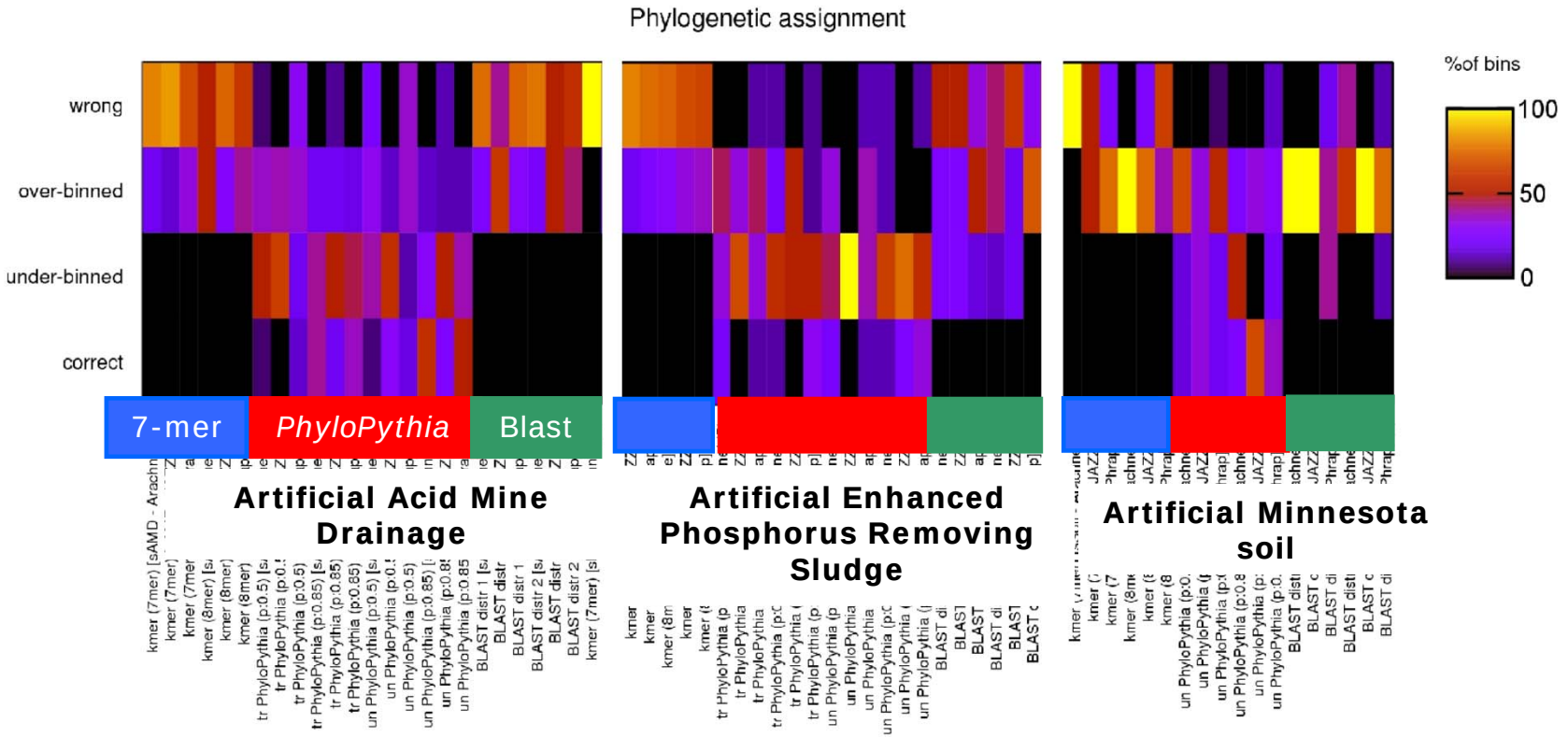
  Sp = TP/(TP+FP)

Example: Assessing assignment performance at species level

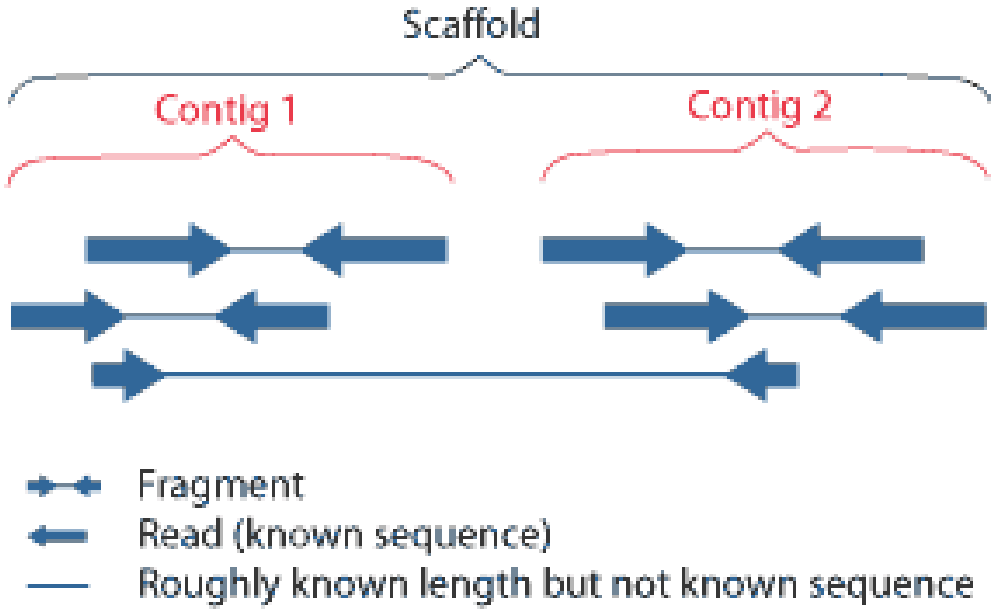
		Predicted Class: WG-1	WG-2	N.A.	Recall / Sensitivity
Actual Class	**WG-1**	10	0	2	10/12
	WG-2	8	28	0	28/36
	N.A.	3	2	89	89/94
	Precision / Specificity	10/21	28/30	89/91	119/142 (Accuracy)

PhyloPythia performed best in simulation study

(Mavromatis *et al.* Nature Methods 2007)



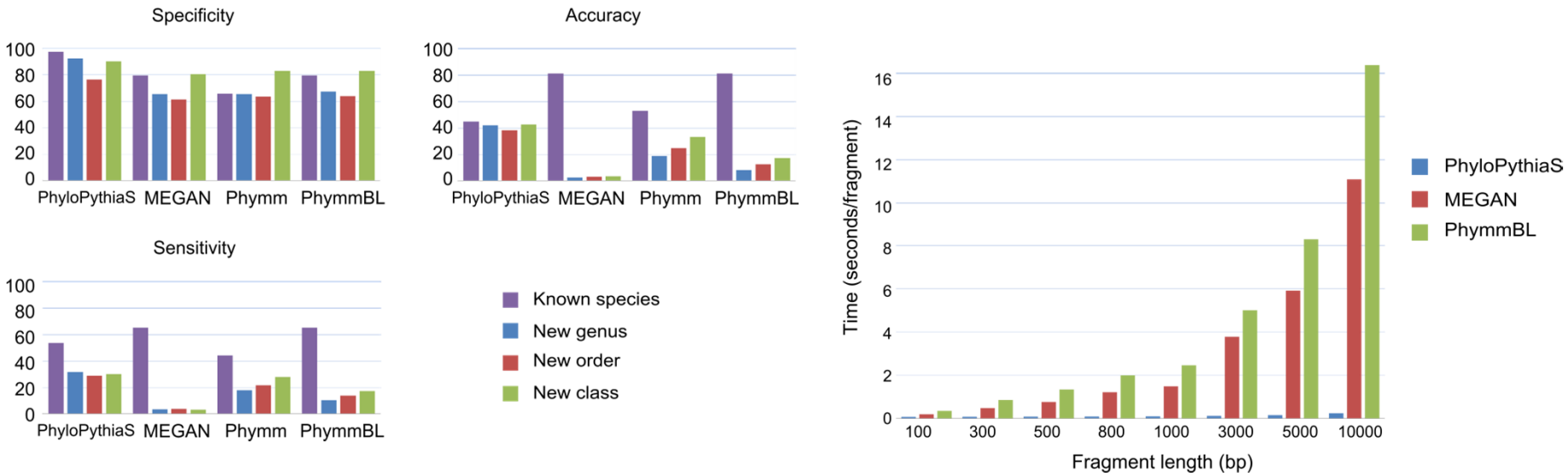
If you don’t have reference data: Look at contig scaffold consistency for contig assignments



From: JGI Genome Portal

PhyloPythiaS: Assignment Accuracy and Execution Time

- 0.08 – 0.1 sec for 0.1 to 10kb fragments (evaluated on a Linux machine with 3GHz processor and 4 GB main memory)
- 85-fold (Megan), 106-fold (PhymmBL) faster (evaluated on 13 Mb sample)
- Higher assignment accuracy for populations of novel clades when little reference data is available (>= 100 kb)



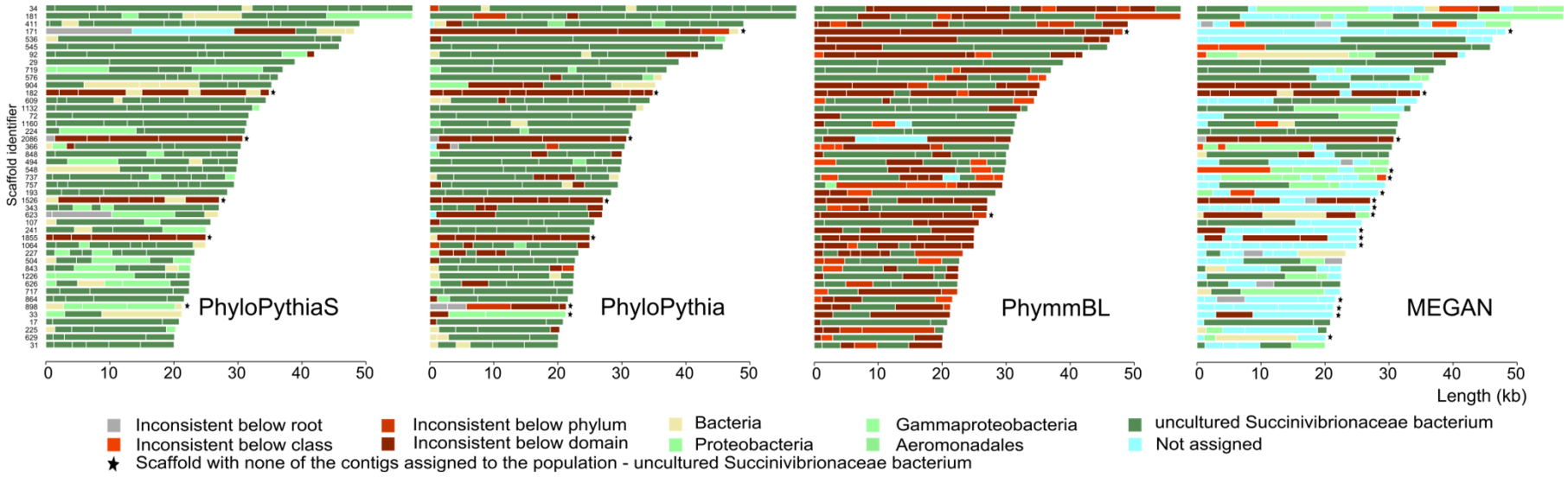
Genus-level assignment accuracy of different taxonomic assignment methods for ‘simMC’ data set (Mavromatis et al. 2007),comprising dominant populations of three different genera and execution time comparison.

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Assignment Consistency for Tammar Wallaby Gut Metagenome

Scaffold assignment consistency for a dominant population of a novel family (Succinivibrionaceae)



From: Patil *et al.* Nature Methods (2011)

5995 contigs (in scaffolds), ~13.57 Mb sequence

WG-1: Scaffold-contig consistency

PhyloPythia: 2.37 Mb (89.75%)

PhyloPythiaS: 2.51 Mb (97.20%)

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

CASE STUDIES

Metagenome Case Studies

Microbial communities from hot spring thermal gradient in Yellowstone Park *(Hugenholtz, Tringe, JGI)*

Tammar Wallaby foregut *(Pope et al., PNAS 2010, Science 2011)*

Human gut *(Turnbaugh et al. PNAS 2010)*

Microbial communities from the termite hindgut *(Warnecke et al. Nature 2007;Brune, Hugenholtz)*

Norwegian reindeer gut *(Pope et al. under review; Morrison; Vaaje-Kolstad)*

EBPR wastewater treatment sludge *(Garcia Martin et al. Nature Biotechnol. 2006)*

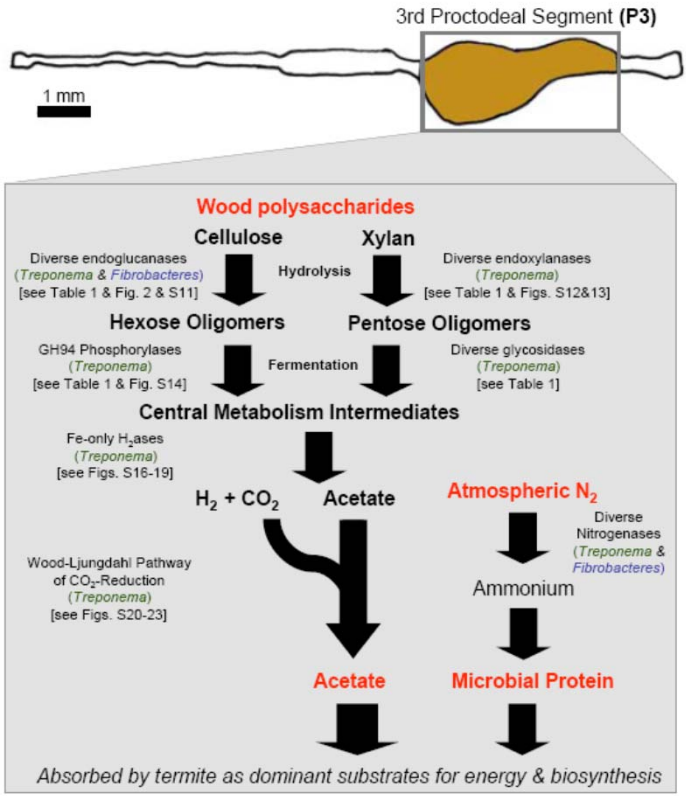
Terephthalate-degrading community *(Liu, Singapore U; Hugenholtz, JGI)*

Lake Washington Methylotrophs *(Kalyuzhnaya et al. Nature Biotechnol. 2008)*

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Taxonomic assignment of the metagenome from the hindgut of a wood-degrading higher termite (with P. Hugenholtz, J. Leadbetter)



Model of nutritional symbiosis-relevant mechanism by *Nasutitermes* paunch bacteria.

Warnecke *et al.* Nature (2007)


NATUR & WISSENSCHAFT

FRANKFURTER ALLGEMEINE ZEITUNG — MITTWOCH, 28. NOVEMBER 2007 · NR. 277 · SEITE N 1

Euphorie war gestern

Die Nanotechnik hat noch einen langen Weg vor sich

Kirschblüten

Holz als nahrhaftes Hauptgericht

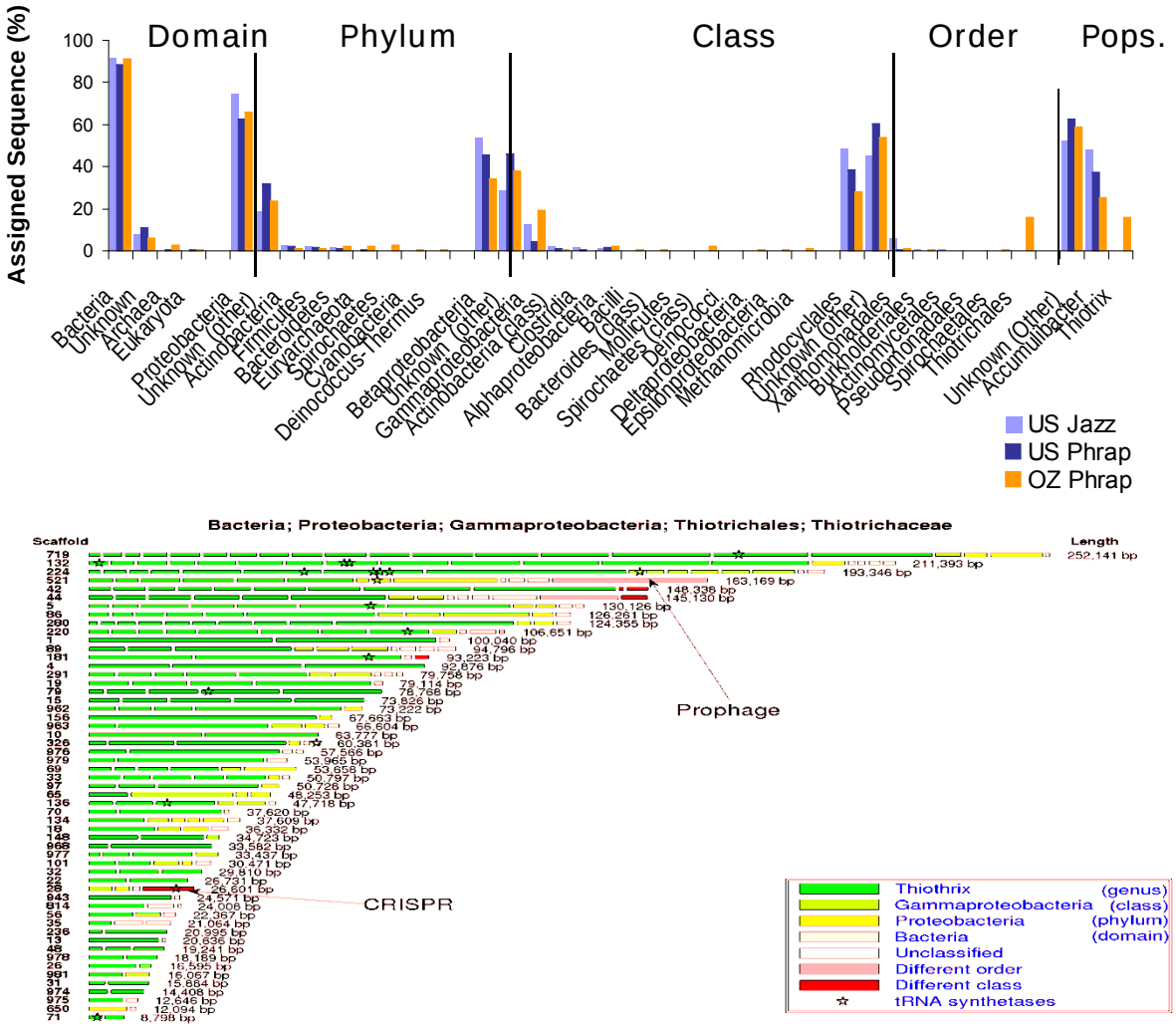
FAZ, 28.11.2007

Taxonomic group	# fragments	%[a]	# bp	%[a]
Bacteria	18,869	96	28,048,863	96
Bacteroidetes	189	1	287,830	1
Fibrobacteres	434	2	828,179	3
Firmicutes	640	3	963,754	3
Clostridia	39	<1	98,404	<1
Proteobacteria	1,614	8	2,579,469	9
Betaproteobacteria	197	1	287,933	1
Gammaproteobacteria	47	<1	92,769	<1
Deltaproteobacteria	43	<1	85,112	<1
Epsilonproteobacteria	111	1	125,930	<1
Spirochaetes	4,648	24	9,487,929	33
Spirochaetales				
Treponema	4,500	23	9,288,738	32
Archaea	584	3	762,065	3
Euryarchaeota	196	1	269,315	1
Other	251	1	311,181	1
Total	**19,720**		**29,164,892**	

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Taxonomic assignments of two EBPR sludge metagenomes (Garcia Martin *et al*. Nature Biotech. 2006)



- All assignments for rRNA marker carrying contigs > 2kb correct (with diff. degree of spec.)
- Retrieved 75% (additional 3.7 Mb sequence) of the genome of a novel organism (*Thiothrix*-like species)
- 97% consistency in scaffold assignments

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

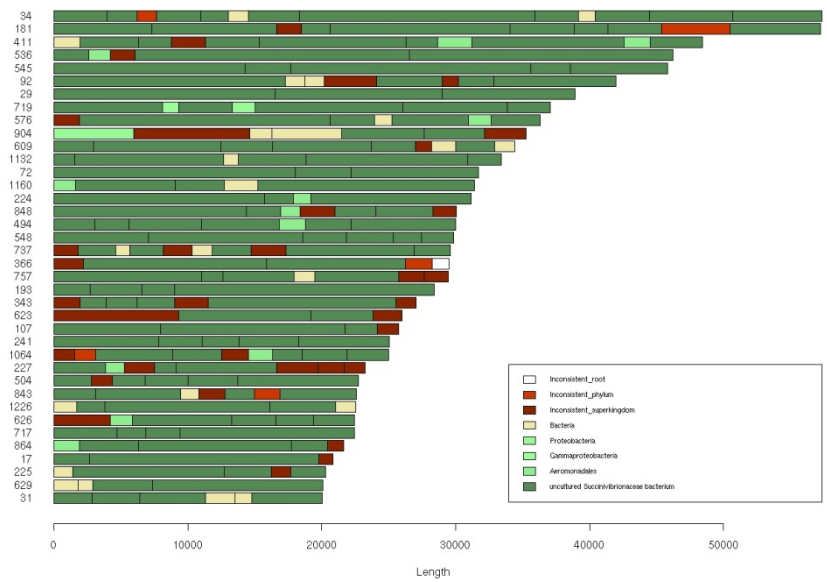
Taxonomic assignment of two deeply sequenced gut microbiomes of human twins (with P. Turnbaugh, J. Gordon)

- Pair of obese human twins
- Pyrosequencing of two gut microbiomes (TS28/TS29)
  - 6.3 / 3.8 Gbp DNA shotgun reads
- Microbial diversity: 878 / 768 OTUs
- Can such a complex community be accurately binned?
- Models of 15 genus-level and 14 family-level clades with sample-specific data, combination with higher level taxonomic models
- 89% / 94% of scaffolds assigned to genus- and family-level bins (25/24 bins, respectively)

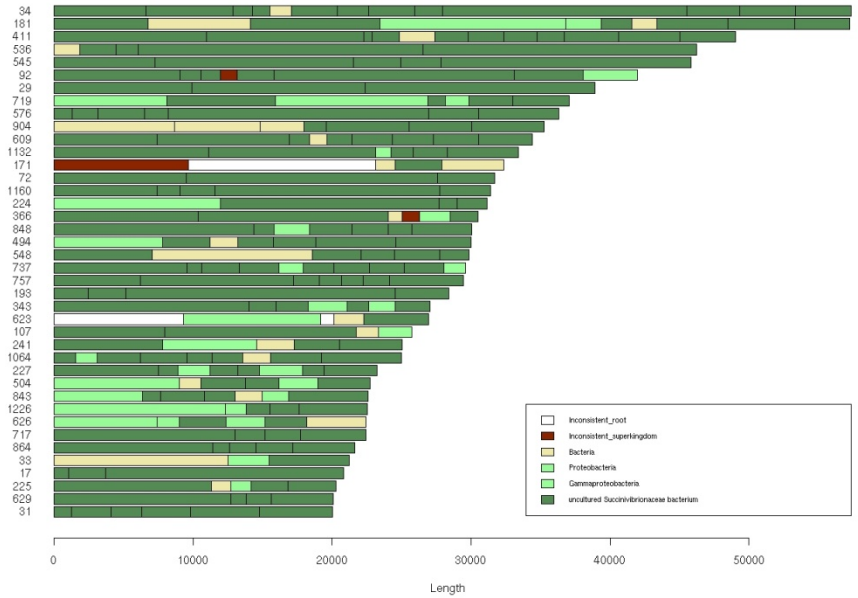
From: Bajzer et al. Nature (2006)

Turnbaugh *et al.* PNAS (2010)

Taxonomic Assignment of Tammar Wallaby Gut Metagenome (with P. Pope, M. Morrison)



PhyloPythia



PhyloPythiaS

5995 contigs (in scaffolds), ~13.57 Mb sequence

Scaffold-contig consistency for WG-1

PhyloPythia: 2.37 Mb (89.75%)
PhyloPythiaS: 2.51 Mb (97.20%)

Consistency with WG-1 genome sequence Bin contig to genome alignment with NUCmer

PhyloPythia: 1.79 Mb (90.09% sequence)
PhyloPythiaS: 1.80 Mb (85.77% sequence)

Pope *et al.* PNAS (2010)

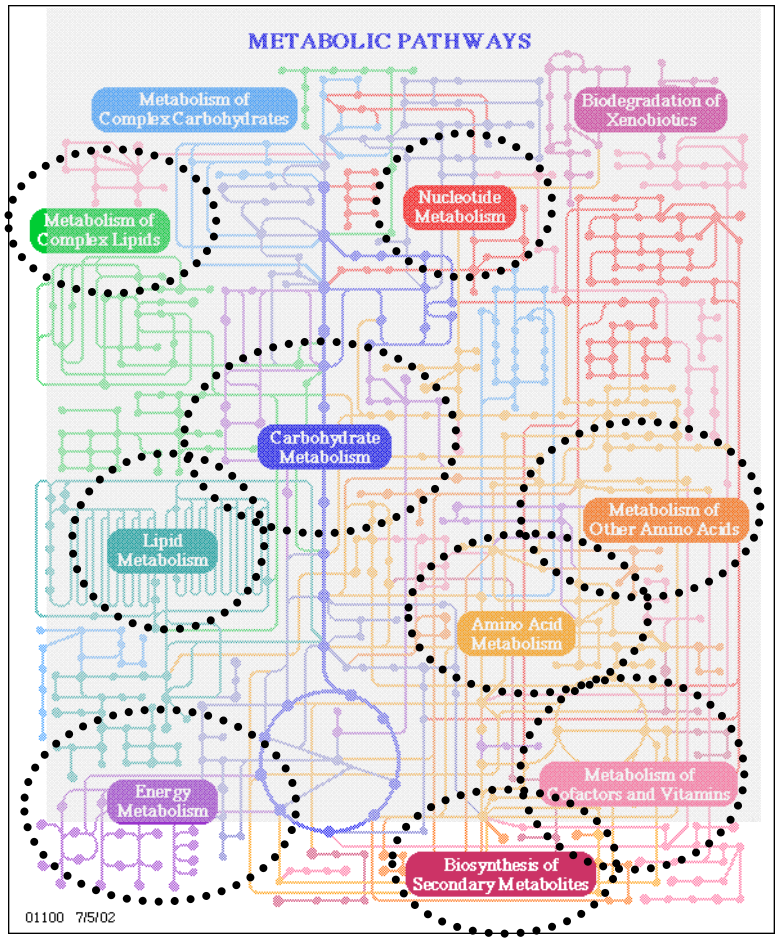
Pope *et al.* Science (2011)

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Problem: Functional Module Inference and Protein Family Annotation

?



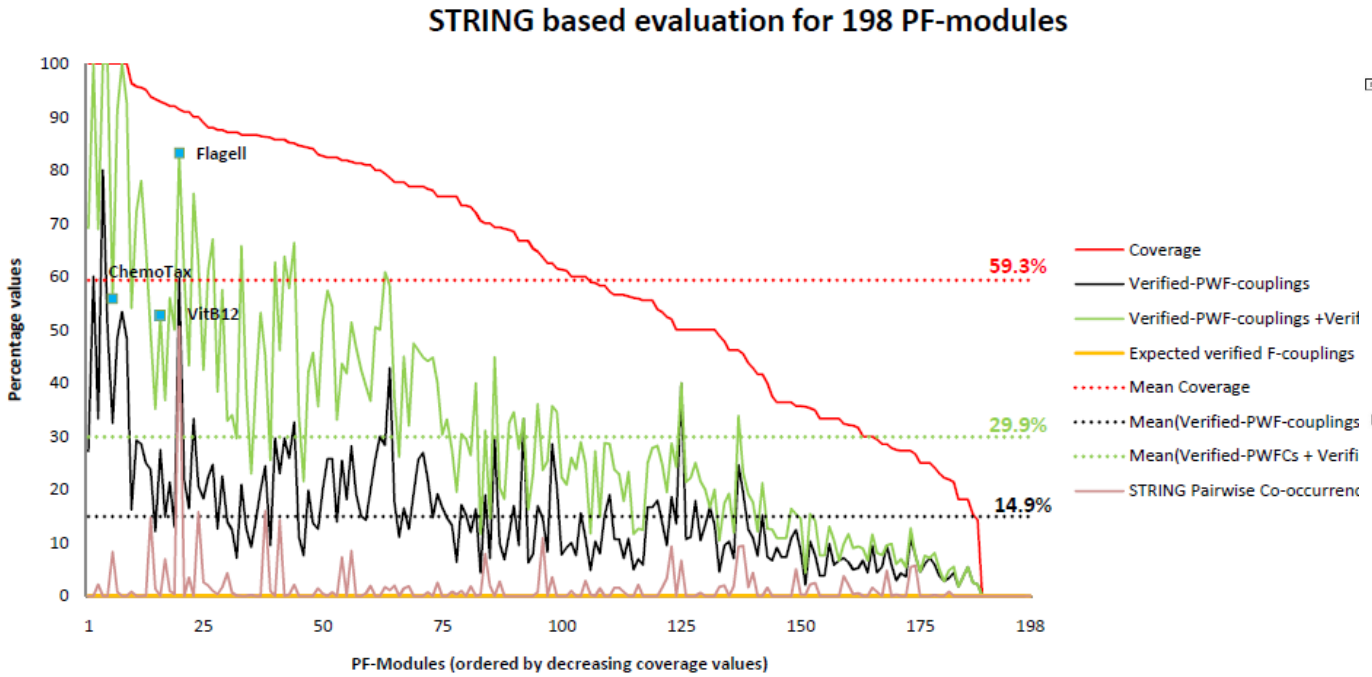
Inference of functional modules from protein family co-occurrence patterns across sequence samples with Bayesian graphical models

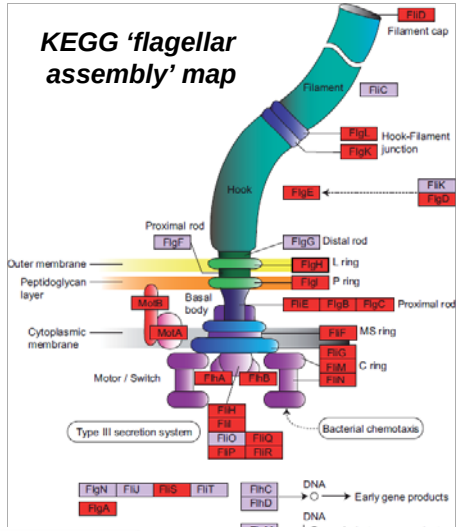
HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

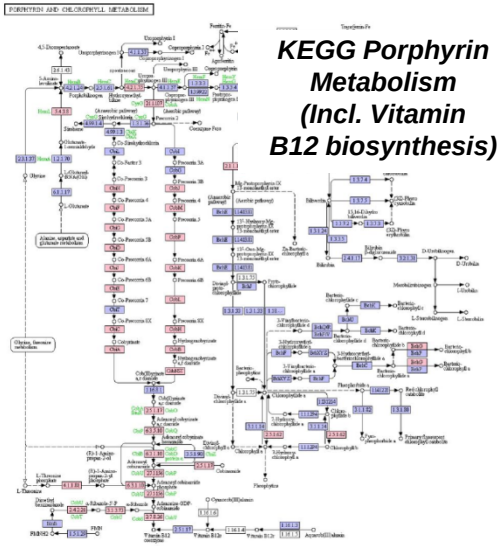
max planck institut informatik

Functional Modules identified from Microbial Genomes (S. Konietzny)

Coverage of a module with respect to the STRING functional network and KEGG pathways





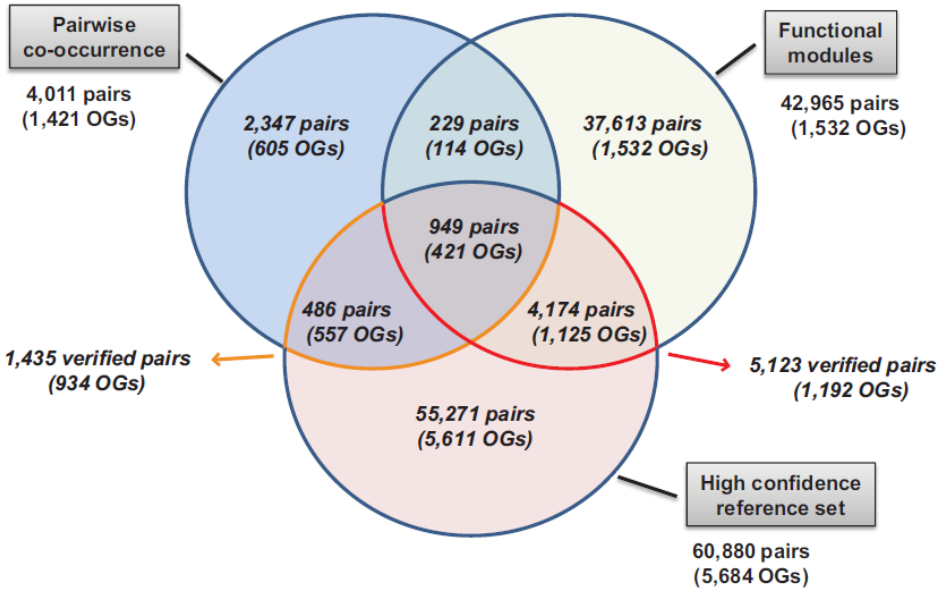


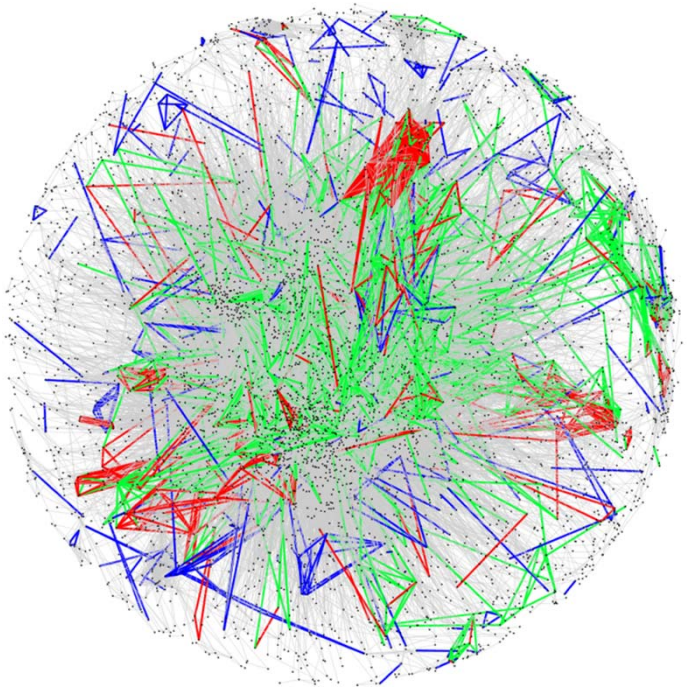
Konietzny et al. BMC Bioinformatics (2011)

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Support by known Pairwise Interactions





Left: Known functional interactions from STRING database identified by functional module inference and pairwise co-occurrence.
Right: Network representation of identified interactions: functional modules (green), pairwise co-occurrence (blue), both (red).

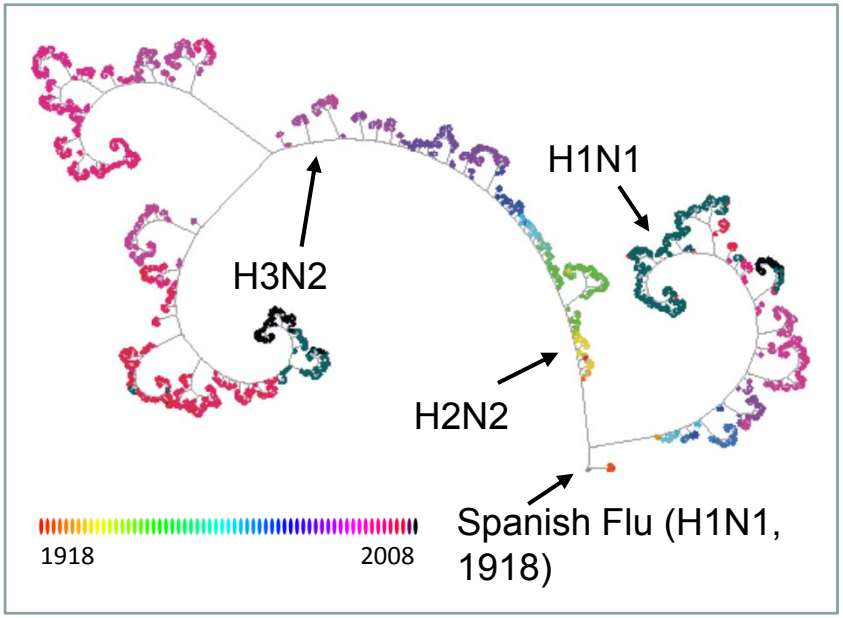
HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Problem: Searching for the Imprint of Selection

- Identify adaptive selection (genotype-phenotype associations) based on evolutionary patterns
- Measures to detect and quantify selection
- Viruses as a model system

- **Problems**
  - Antigenic drift
  - Capability of human-to-human transmission of influenza viruses
  - Microbial habitat adaptation
  - …
  - ..

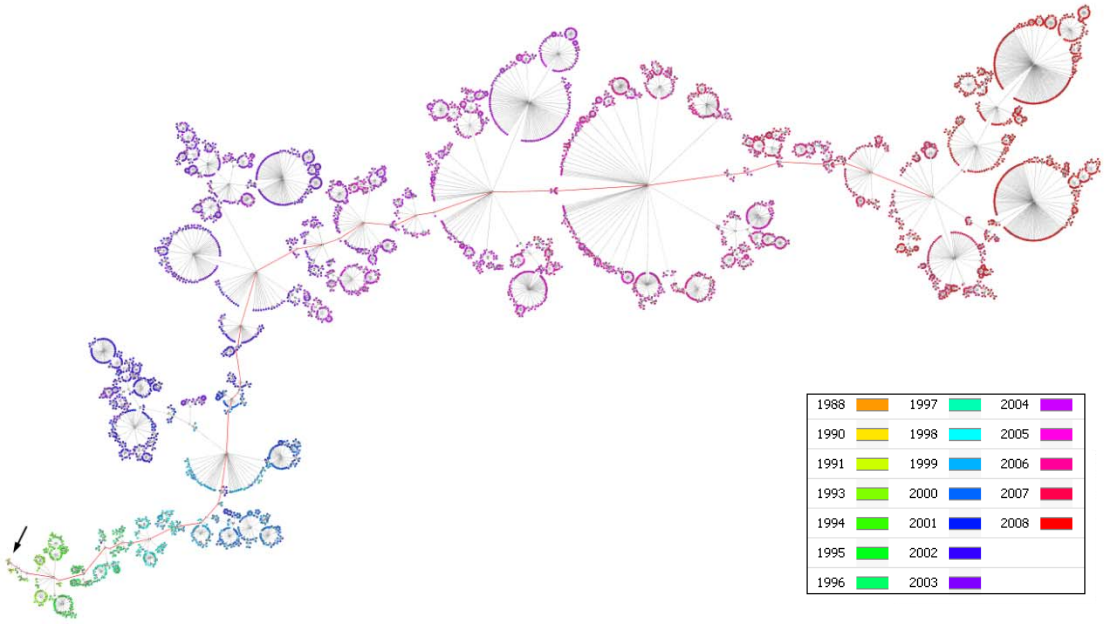


Maximum likelihood tree for PB2 sequences of human influenza A viruses of the subtypes H1N1, H2N2 und H3N2.

Adams, McHardy PLoS Pathogens (2009)
Steinbrück, McHardy NAR (2010)
Adams, McHardy Proc Biol Sci. (2010)

Allele Dynamics Plots
(L. Steinbrück)

- Visualization of evolutionary dynamics for time-series of population-level sequence samples



To what branch in the phylogeny do mutations map?

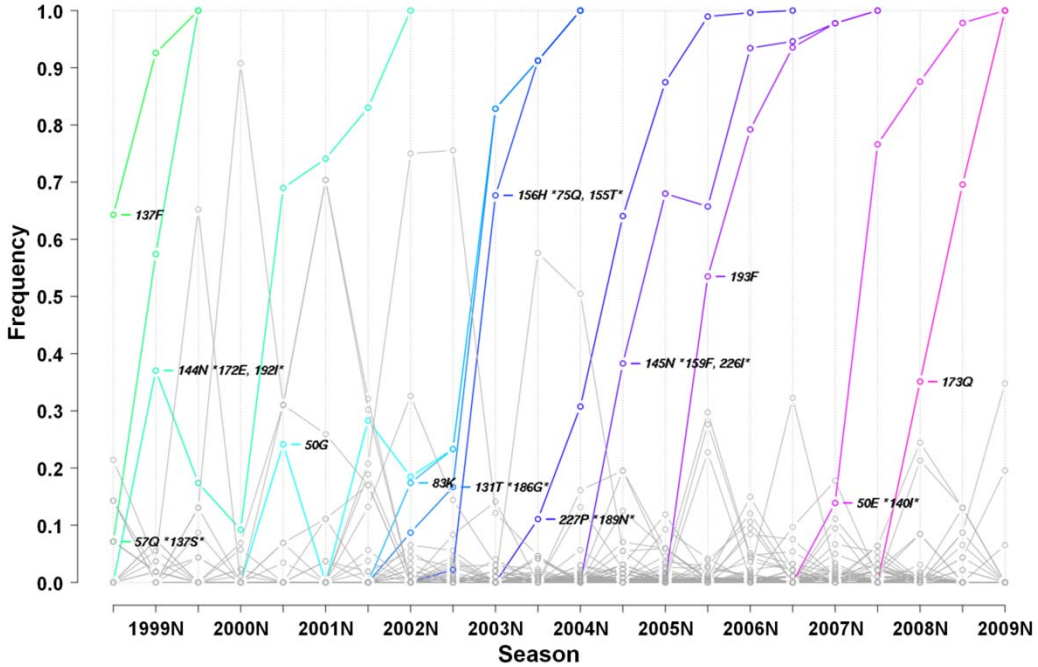
Which mutations influence viral evolution?

Phylogenetic tree for the HA segment of ~4000 influenza A (H3N2) viruses

Steinbrück, McHardy NAR (2010)

Allele Dynamics Plots
(L. Steinbrück)

- Visualization of evolutionary dynamics for time-series of population-level sequence samples



To what branch in the phylogeny do mutations map?

Which mutations influence viral evolution?

Gene allele plot for the HA segment of ~4000 influenza A (H3N2) viruses

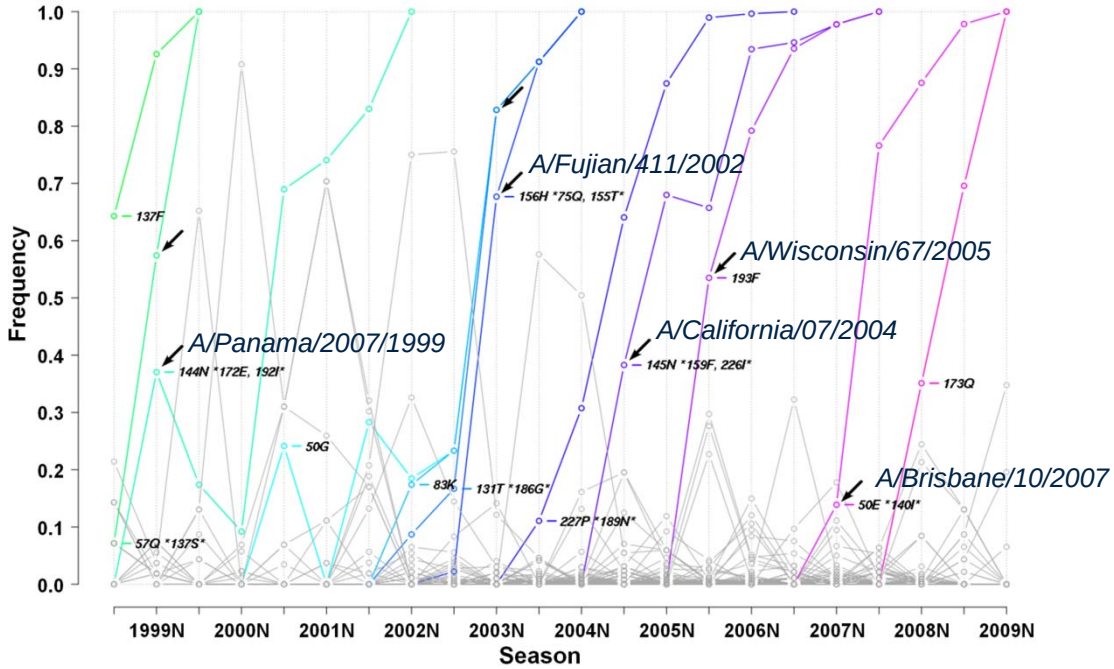
Steinbrück, McHardy NAR (2010)

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Allele Dynamics Plots
(L. Steinbrück)

- Visualization of evolutionary dynamics for time-series of population-level sequence samples
- Identification of alleles that are on the rise to predominance



Gene allele plot for the HA segment of ~4000 influenza A (H3N2) viruses

To what branch in the phylogeny do mutations map?

Which mutations influence viral evolution?

Steinbrück, McHardy NAR (2010)

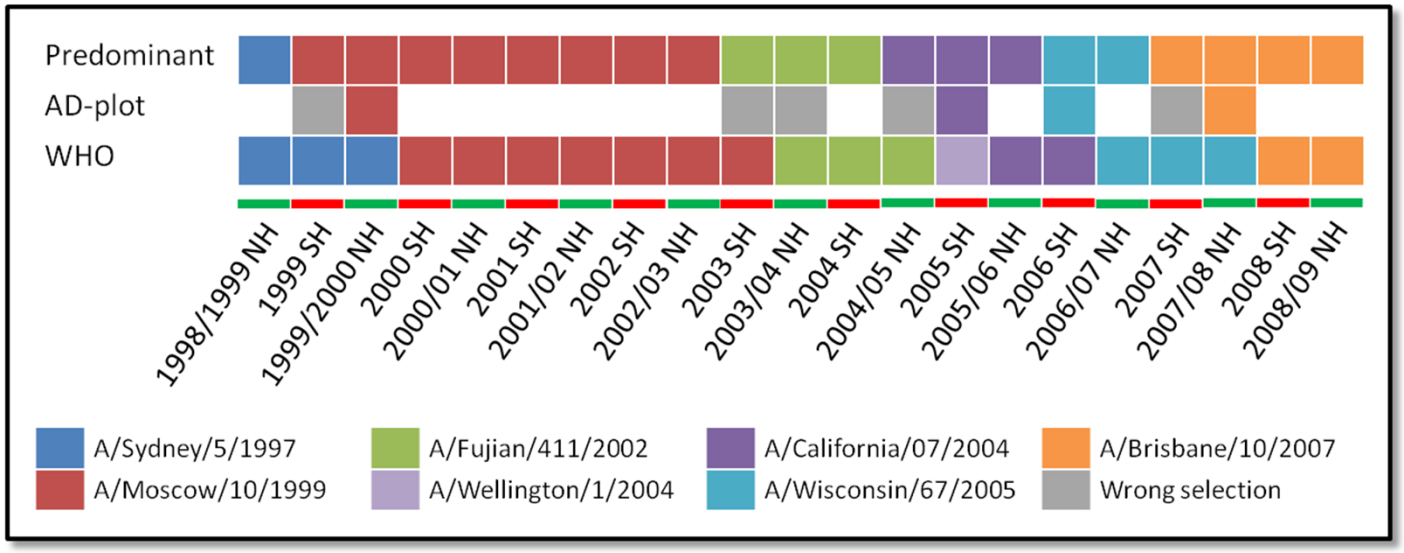
HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Allele Dynamics Plots

(L. Steinbrück)

- Visualization of evolutionary dynamics for time-series of population-level sequence samples
- Identification of alleles that are on the rise to predominance
- Applicable to identify future predominant and novel influenza A (H3N2) virus strains

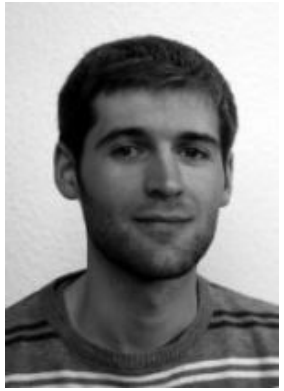


Steinbrück, McHardy NAR (2010)

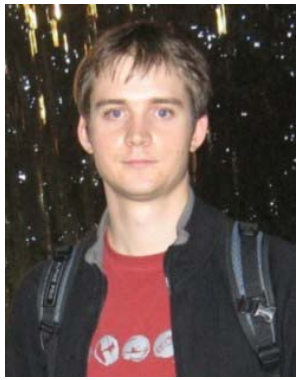
HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

McHardy Lab



Johannes Droege



Ivan Gregor

Sebastian Konietzny

Kaustubh Patil

Lars Steinbrueck

Christina Tusche

Aaron Weimann

Former lab members: Ben Adams, Yulia Trukhina, Phillip Muench

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Acknowledgements

Joint Genome Institute: **P. Hugenholtz** (now Australian Center for Ecogenomics), F. Warnecke, S. Tringe, H. Garcia-Martin, K. Mavrommatis, N. Kyrpides
MPI Informatics: T. Lengauer, F. Domingues, L. Dietz
MPI Molecular Genetics: M. Schweiger, C. Röhr, M. Kerrick
MPI Terrestrial Microbiology: A. Brune,W. Ohtsubo
MPI Plant Research: P. Schulze-Lefert
HHU: A. Borkhardt

IBM: **I. Rigoutsos** (now Jefferson U.), A.Tsirigos, T. Huynh
Washington University / Harvard : M.Chistoserdova, J. Gordon, **P. Turnbaugh**
CSIRO / UMB: **M. Morrison**, **P. Pope**
U. Potsdam : T. Scheffer, P. Haider
U. Singapore: W. Liu
CalTech: J. Leadbetter
ANL/U. Chicago: J. Gilbert

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

max planck institut informatik

Open Positions

- http://www.cs.uni-duesseldorf.de/AG/AlgBio/Jobs

- Looking for a postdoctoral research with interest in computational metagenomics (taxonomic assignment, plant biomass degradation, genotype-phenotype relationships)

- Looking for a PhD student with interest in reassortment events in influenza A virus evolution

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

max planck institut
informatik

LAB SESSION


HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF


max planck institut
informatik

Taxonomic assignment of a metagenome from the Tammar Wallaby gut

- Our data set: Assembled sequence data
- 5995 contigs (in scaffolds)
- ~13.57 Mb sequence

1) *Identify marker genes and taxonomic placement*
2) *Decide on what clades to model with PhyloPythiaS*
3) *Create tree file (which clades to model), file with sample-specific training data*
4) *Train PhyloPythiaS classifier (create your own model)*
5) Classify your data with PhyloPythiaS using a model
6) Visualize results

Optional (recommended for best performance)

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF
max planck institut informatik

Expert-identified sample-derived training sequences

- **538960**: Bacteria; Proteobacteria; Gammaproteobacteria; Aeromonadales; Succinivibrionaceae; environmental samples; **uncultured Succinivibrionaceae bacterium** (WG-1)
- **297314**: Bacteria; Firmicutes; Clostridia; Clostridiales; Lachnospiraceae; environmental samples; **uncultured Lachnospiraceae bacterium** (WG-2)
- **336130**: Bacteria; Firmicutes; Erysipelotrichi; Erysipelotrichales; Erysipelotrichaceae; environmental samples; **uncultured Erysipelotrichaceae bacterium** (WG-3)
- **171549**: Bacteria; Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; **Bacteroidales**

PhyloPythiaS webserver

- http://binning.bioinf.mpi-inf.mpg.de/
  - Build your own or use the generic model to assign your data
  - Create your own model with clades you are interested in by uploading sample-specific training data and specifying clades

You need:

- Your data set as a multi-FASTA file: Prediction file
- Sample-specific mode, in addition to prediction file:
  - 'Tree file/Clade list': Specify species and higher-level clades for model
  - 'Sample specific data': Additional sequences from clades as multi-FASTA file
  - Recommended: >100 kb of sequence per clade (or represented by >2 species in public databases)

Webserver Training data File Formats

- 'Sample specific data': Training sequences as multi-Fasta file with the following header:

  >ID label:TaxID

  acgttttacccccccttcgatagcgata…

- 'Tree file/Clade list': Clades to model (species level or above)
  - Either: Plain text file with list of taxonomic IDs (see below)
  - Needs only leaf node taxonomic IDs according to NCBI Taxonomy
  - Or: Full tree structure in Newick format (if you want to include nodes not represented in NCBI Taxonomy)

Example Tree files

List of tax IDs

538960
106588
297314
186806
541000
331630
31977
909930
203492
171549
2159

Newick file:

((((((538960)83763)135624)1236)
1224,((((297314)186803,186806,
31977,541000)186802)186801,(((
331630)128827)526525)526524)
1239,(((203492)203491)203490)3
2066,(((((106588)816)815)171549
)200643)976)2,((((2159)2158)183
925)28890)2157)1;

Practical Part

- Web server session:
  http://tiny.cc/0kaow

- Commandline tutorial for PhyloPythiaM: See distributed pdf.

PhyloPythiaM command line options

- usage: run.py [-h] -c config.cfg [-n] [-g] [-j J [J ...]] [-o O [O ...]] [-t] [-p P [P ...]] [-r] [-s]
- optional arguments:

-h, --help #show this help message and exit

-c config.cfg, --config config.cfg #configuration file of the pipeline

-n, --run-rrna16S #run hidden markov model and classify according to the 16S, 23S, and 5S g

-s, --summary #Summary

-g, --run-marker-gene-analysis #run hidden markov model and classify based on marker genes (from Amphora)

-t, --pps-train #run PhyloPythiaS "train script„

-p P [P ...], --pps-predict P [P ...] # run PhyloPythiaS "predict script" (c) for contigs, (s) for scaffolds, (v) run cross validation

-r, --read-pps-out #Reads the output/placements of PPM

PhyloPythiaS Configuration File Options

#directory where processed NCBI data is stored, provide empty directory to create new

#REUSABLE

NCBI_PROCESSED_DIR:

#Directory containing NCBI taxonomy in SQlite3 format with file name "ncbitax_sqlite.db"

#provide empty directory to create new database

#REUSABLE

NCBI_TAX_DIR:

#project directory, must be empty

PROJECT_DIR:

#a file containing a tree in newick format (see restrictions in INSTALL.txt)

#OR a file with ncbi taxon ids (one id per line) to create a tree from

TREE_FILE:

Configuration File Options (optional)

#a directory with sample specific fasta files (file names must start with appropriate organism/species ncbi taxonomic id). leave empty if you don't have any

SAMPLE_SPECIFIC_DIR:

#kmer feature space for multiple kmers use kmer_min-kmer_max

KMER:6

#C values for SVM, if single value is given then models will be build with that value.

#If comma separated (no space) values are given then cross-validation will be performed.

#If a single value is provided, all models will be built with it. Our experience shows that in general. values less than 1 (e.g. 0.01 and 0.1) do not provide good models.

C_GRID:1000

#Fragment lengths for different models (comma separated, no space)

FRAGMENT_LEN:1000,3000,5000,10000,15000,50000

#kmer feature. use reverse complement for computing kmer features?

REV_COMPLEMENT:1

Configuration File Options (optional)

#0:disabled, 1:sequence length, 2:sequence_length-k+1, 3:embedded monomer frequency

KMER_NORMALIZATION:1

#Taxonomic ranks (comma separated, no space) statring at the lowest rank. Please make sure that "root" is there at the end.

TAXONOMY_RANKS:genus,family,order,class,phylum,superkingdom,root

#Number of examples per training file

NUMBER_EXAMPLES:10000

#step size for sample specific data; either a single number (for all fragment lengths) or an array separated with ","

SAMPLE_SPECIFIC_STEP:100,300,500,1000,1500,5000

#should the models be built in parallel (please make sure that you have enough number of processors and main memory)

PARALLEL_MODELS:FALSE

Configuration File Options (optional)

#allowed file extrensions

EXTENSIONS:

#genomes to exclude: file containing one ncbi tax_id per line

GENOMES_EXCLUDE:

#if the training data is already there then just build models (TRUE/FALSE)

ONLY_MODELS:FALSE

#clean-up the data (sampled_fasta and train_data direcories) created after training? TRUE/FALSE

CLEAN_UP_TRAIN:FALSE

##prediction configuraion

#number of classifiers to use, keep this odd to avoid ties

N_CLASSIFIERS:3

#Create Pie charts for every taxonomic rank TRUE/FALSE (in prediction)

#slice colors are determined automatically so no color consistency is guaranteed

PIE_CHARTS:TRUE