

Part II

BACTERIAL DIVERSITY, TAXONOMY, PHYLOGENY & TOOLS

Prokaryotic Nomenclature

- Bacterial and Archaeal Names are covered by an official set of rules:
 - “The International Code of Nomenclature of Prokaryotes”
 - Covers ranks from Class to Subspecies.
 - Ranks of Kingdom and Phylum (Domain) are not covered by the code.

Synonymy

9612 Species Names *But* Only 8062 Species?

New Combination (taxonomic change)

Rhizobium meliloti -> *Ensifer meliloti*

Homotypic Synonym (same type strain)

Aeromonas caviae = *Aeromonas punctata*

Heterotypic Synonym (different type strain)

Wautersia eutropha => *Cupriavidis necator*

LPSN

List of Prokaryotic names with Standing in Nomenclature

Formerly List of Bacterial names with Standing in Nomenclature (LBSN)

J.P. EUZÉBY
SBSV

Author's e-mail

Last full update:
August 06, 2009

Minor changes since
the last full update:
August 08, 2009

URL:
www.bacterio.net

Search

Search LPSN

Google™ Site Search

Search

Taxonomic categories and changes covered by the Rules of the Code

- Genera and suprageneric taxa: [List A-C](#) [List D-L](#) [List M-R](#) [List S-Z](#)
- Suprageneric taxa
- Genera
- Genera of the domain (or empire) of Archaea (or Archaeobacteria)
- Genera of the domain (or empire) of Bacteria (or Eubacteria)
- Approved Lists of Bacterial Names
- Names validly published by announcement in Validation Lists
- Basonyms, new combinations (comb. nov.), nomina nova (nom. nov.)

Taxonomic categories and changes not covered by the Rules of the Code

- Candidatus
- Taxa above the rank of class
- Some prokaryotic names without standing in nomenclature
- Lists of Changes in Taxonomic Opinion

Endorsed prokaryotic names (J.P. Euzéby & D.J. Tiedje)



Identification

Patent and Safe Deposit

► Microorganisms

Plant Viruses

Plant Cell Lines

Human and Animal Cell Lines

Home | DSMZ | Services | Contact | Download | Press

Search

Go

Advanced Search

- >> New Accessions
- >> Head of Department
- >> Staff
- >> Publications
- >> Catalogue
- >> Prices
- >> Ordering Procedure
- >> Conditions of Delivery
- >> Technical Information
- >> Deposit
- >> Special Services
- >> Safety Instructions
- >> Bacterial Nomenclature
- >> Schulgeeignete Kulturen

Microorganisms > Bacterial Nomenclature

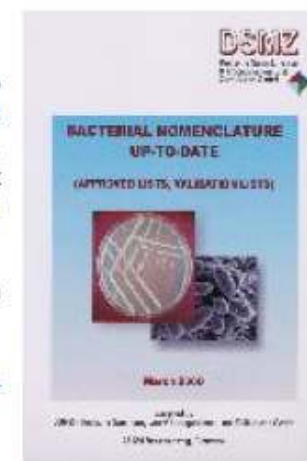
Bacterial Nomenclature Up-to-Date

[Search an alphabetical list of bacterial names A - Z](#)

A complete list of the names can be downloaded from our FTP server: [bactname.pdf](#) (PDF file), [bactname.exe](#) (self-extracting file for MS-DOS systems) or [names.txt](#) (uncompressed ASCII text file). To import the list of names into a database or a spreadsheet programme, you can download [bactname.zip](#) (compressed Excel file).

Other useful information on bacterial or general nomenclature can be found at:

- [J. P. Euzéby, Ecole Nationale Vétérinaire de Toulouse, France](#)
- [TOBA - Taxonomic Outline of the Bacteria and Archaea](#)
- [CCUG, Sweden](#)
- [Species 2000 CheckList](#)



"Bacterial Nomenclature up-to-date" is based on the work of Norbert Weiss, who maintained the database until his retirement in February 2003. The database is based on those names which are validly published according to the Bacteriological Code. The present database is maintained under the supervision of [Dorothea Gleim](#) and [Manfred Kracht](#), who may be consulted on technical aspects (database resp. online access). Queries relating to nomenclature or taxonomic interpretation may be addressed to [Brian J. Tindall](#).

New: For more than 1100 species of the *Actinobacteria* very valuable additional information is available: in the manual [Compendium of Actinomycetales](#) by Joachim Wink most of the known



International Code of Nomenclature of Bacteria

(1990 Revision)

[Short Contents](#) | [Full Contents](#)[Other books @ NCBI](#)

Search

☒ This book ☐ All books
☐ PubMed

Navigation

About this book

1. General Considerations

[General Consideration 1](#)[General Consideration 2](#)[General Consideration 3](#)[General Consideration 4](#)[General Consideration 5](#)[→ General Consideration 6](#)[General Consideration 7](#)

International Code of Nomenclature of Bacteria (1990 Revision) → 1. General Considerations

General Consideration 6


This Code is divided into Principles, Rules, and Recommendations.

1. The *Principles* ([Chapter 2](#)) form the basis of the Code, and the Rules and Recommendations are derived from them.
2. The *Rules* ([Chapter 3](#)) are designed to make effective the Principles, to put the nomenclature of the past in order, and to provide for the nomenclature of the future.
3. The *Recommendations* ([Chapter 3](#)) deal with subsidiary points and are appended to the Rules which they supplement. Recommendations do not have the force of Rules; they are intended to be guides to desirable practice in the future. Names contrary to a Recommendation cannot be rejected for this reason.
4. Provisions for emendations of Rules, for special exceptions to Rules, and for interpretation of the Rules in doubtful cases have been made by the establishment of the International Committee on Systematic Bacteriology (ICSB) and its Judicial Commission, which acts on behalf of the ICSB (see [Rule 1b](#) and Statutes of the International Committee on Systematic Bacteriology, pp. 137–158 of this volume). Opinions issued by the Judicial Commission become effective after receipt of ten or more favorable votes from Commissioners, but may be rescinded by the ICSB as provided in ICSB Statute 8c (2). The official journal of the ICSB is the *International Journal of Systematic Bacteriology* (IJSB), formerly the *International Bulletin of Bacteriological Nomenclature and Taxonomy* (IBNT). (Some other journal could be specified by the ICSB if required. Such possible future specification is implicit in the use of

Prokaryotic Taxonomy

- There is no formal Prokaryotic taxonomy
Taxonomy is a matter of opinion!
- Changes in opinion may require changes in nomenclature
Rhizobium meliloti -> *Ensifer meliloti*
- Every taxon must be circumscribed
-

NCBI

Taxonomy
Browser

PubMedEntrezBLASTOMIMTaxonomyStructure

Search forAScomplete name☐lockGoClear

Taxonomy
browser

Archaea

Bacteria

Eukaryota

Viroids

Viruses

Taxonomy
common tree

Taxonomy
information

Taxonomy
resources

Taxonomic
advisors

Genetic codes

Taxonomy
Statistics

Taxonomy
Name/Id Status
Report

The NCBI Taxonomy Homepage

These are direct links to some of the organisms commonly used in molecular research projects:

Arabidopsis thaliana	Escherichia coli	Pneumocystis carinii
Bos taurus	Hepatitis C virus	Rattus norvegicus
Caenorhabditis elegans	Homo sapiens	Saccharomyces cerevisiae
Chlamydomonas reinhardtii	Mus musculus	Schizosaccharomyces pombe
Danio rerio (zebrafish)	Mycoplasma pneumoniae	Takifugu rubripes
Dictyostelium discoideum	Oryza sativa	Xenopus laevis
Drosophila melanogaster	Plasmodium falciparum	Zea mays

Comments and questions to info@ncbi.nlm.nih.gov

The Taxonomic Outline of Bacteria and Archaea

[HOME](#) [ABOUT](#) [LOG IN](#) [REGISTER](#) [SEARCH](#) [CURRENT](#) [ARCHIVES](#)
[ANNOUNCEMENTS](#) [SITE MAP](#)

[OPEN JOURNAL
SYSTEMS](#)

[Journal Help](#)

Home > **TOBA Release 7.7**

The Taxonomic Outline of Bacteria and Archaea

TOBA Release 7.7

This release of the outline was prepared as part of a discussion on sequencing the genomes of the type strains of Bacteria and Archaea.

Table of Contents

The Outline

Introduction to the Taxonomic Outline of Bacteria and Archaea (TOBA) Release 7.7

George M Garrity, Timothy G. Lilburn, James R. Cole, Scott H. Harrison, Jean Euzéby, Brian J Tindall

[PDF](#)

1-5

Part 1 – The Archaea, Phyla Crenarchaeota and Euryarchaeota

George M. Garrity, Timothy G. Lilburn, James R. Cole, Scott H. Harrison, Jean Euzéby, Brian J Tindall

[PDF](#)

6-31

Part 2 – The Bacteria: Phyla Aquificae, Thermotogae, Thermodesulfobacteria, Deinococcus-Thermus, Chrysiogenetes, Chloroflexi, Thermomicrobia, Nitrospira, Deferribacteres, Cyanobacteria, and Chlorobi

[PDF](#)

George M. Garrity, Timothy G. Lilburn, James R. Cole,

32-51

USER

Username

Password

☐ Remember me

JOURNAL CONTENT

Search

All

Browse

- [By Issue](#)
- [By Author](#)
- [By Title](#)

FONT SIZE

INFORMATION

- [For Readers](#)
- [For Librarians](#)

Bergey's Manual Trust

[Home](#)

[About the Trust](#)

[Editorial Offices](#)

[Publications](#)

[Taxonomic outlines](#)

[Instructions for
Authors](#)

[Trust newsletter](#)

[Resources](#)

Bergey's Taxonomic Outlines

The taxonomic outlines for Volumes 3 and 4 of *Bergey's Manual* are available here. Please download the PDF files below.

- [Volume 3](#)
- [Volume 4](#)
- [Volume 5 - The Actinobacteria](#)

Earlier releases of the complete taxonomic outline are available here:

- [Release 1.0 \(April 2001\)](#)
- [Release 2.0 \(January 2002\)](#)
- [Release 3.0 \(July 2002\)](#)
- [Release 4.0 \(October 2003\)](#)
- [Release 5.0 \(May 2004\)](#)

Bergey's Manual is a registered trademark of Bergey's Manual Trust.

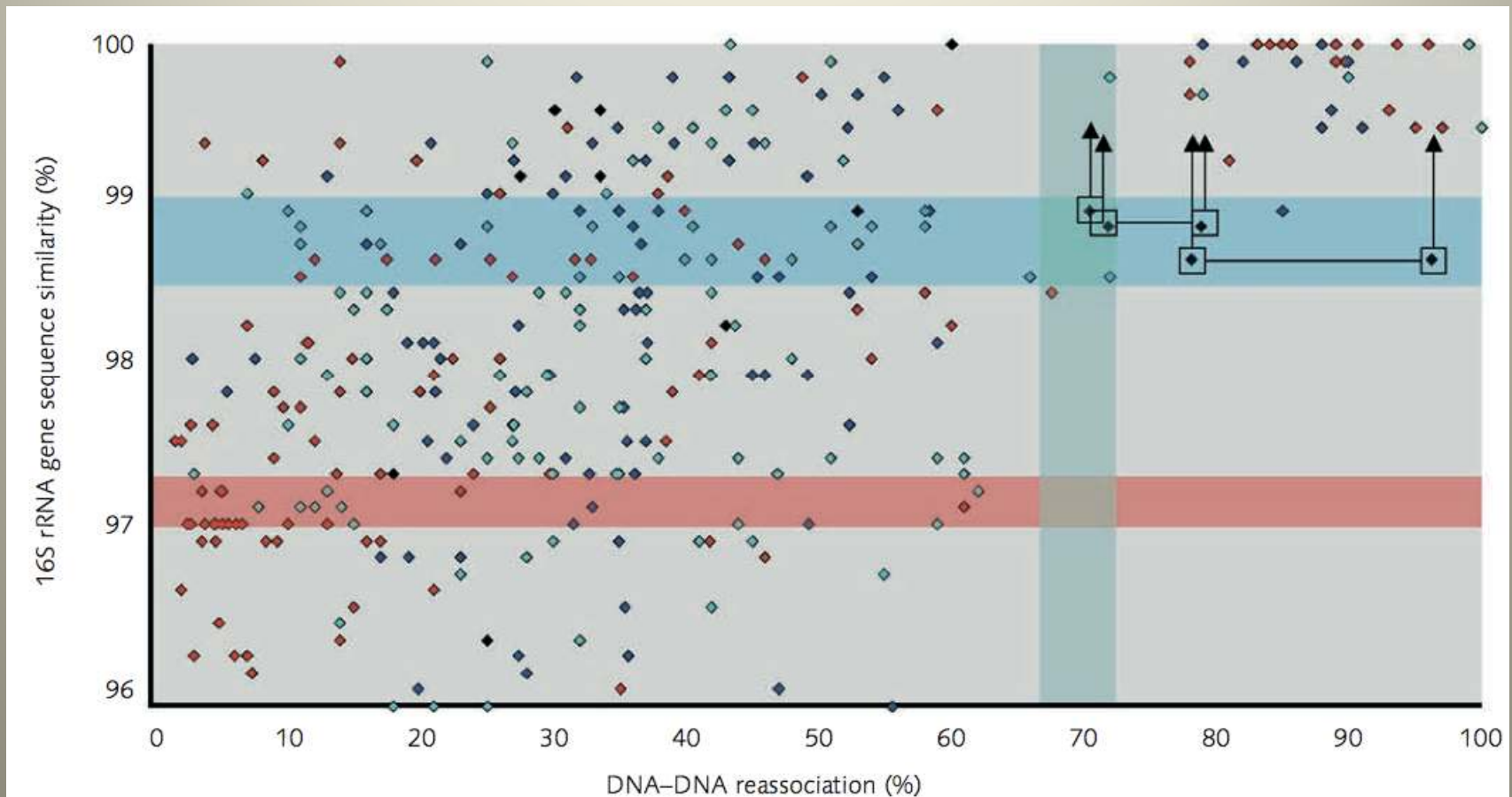
© 2008 Bergey's Manual Trust

This site last updated 11 August 2009

Bacterial Species

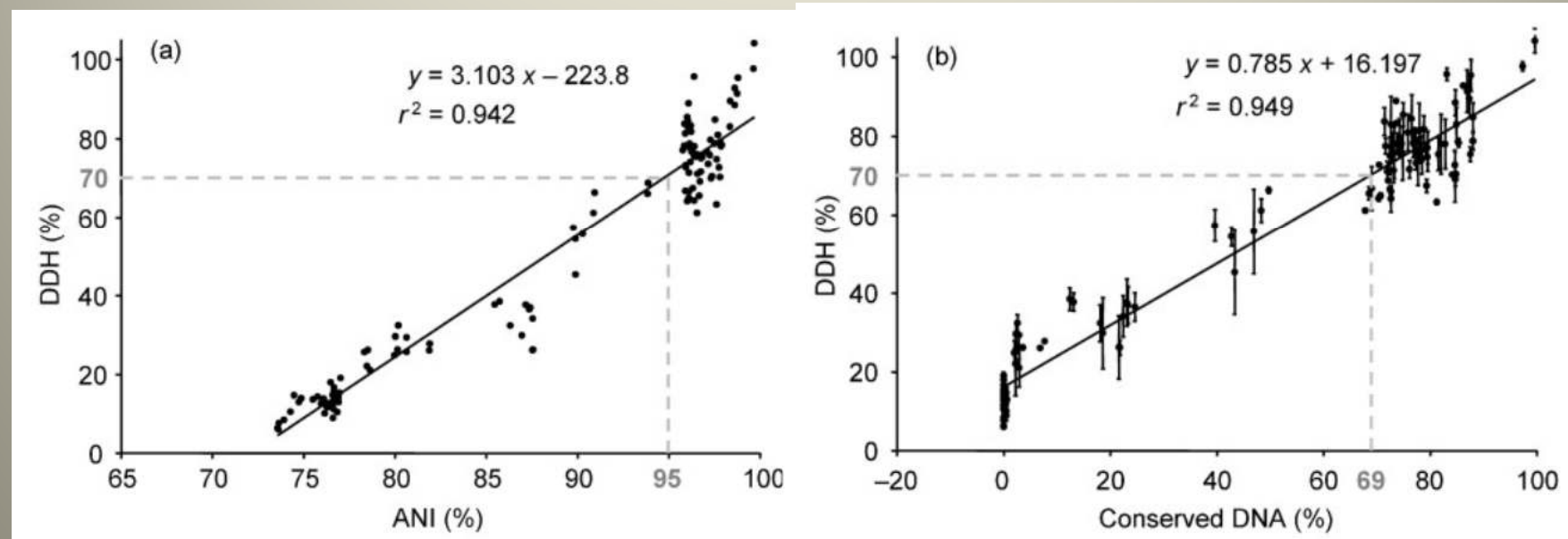
- One strain of a species is designated as the “type strain”. Other closely related strains are of the same species.
- There is no completely objective method for species delimitation.
- The current accepted standard is 70% DNA-DNA hybridization (DDH)

DDH vs 16S Similarity



Stackebrandt, E. and J. Ebers. (2006) Taxonomic parameters revisited: tarnished gold standards. *Microbiology Today* :152-155.

DDH vs Genome Similarity



Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M. (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. **IJSEM** 57:81–91. doi:10.1099/ij.s.0.64483-0

Bacterial Species Definition

- The species concept in bacteria is not well defined.
- As an ad-hoc measure, 70% DHH has several limitations.
- Less than 98.5% 16S similarity indicates different species, but greater than 98.5% does not indicate the same species.
- ANI may be a good candidate to replace DHH as an ad-hoc species definition.

Nomenclatural References

- The International Code of Nomenclature of Prokaryotes:
<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=icnb>
- List of Prokaryotic Names with Standing in Nomenclature
<http://www.bacterio.cict.fr/index.html>
- Bacterial Nomenclature Up-to-Date
http://www.dsmz.de/microorganisms/bacterial_nomenclature.php

Taxonomy References

- NCBI Taxonomy
<http://www.ncbi.nlm.nih.gov/Taxonomy/>
- TOBA
<http://www.taxonomicoutline.org/>
- Bergey's Taxonomy
<http://www.bergeys.org/outlines.html>

Phylogenetic Inference

Use simplest method that works.

Why rRNA?

- Universal
- No ortholog - paralog problem
- No horizontal gene transfer
- Easy to amplify and sequence
- Large database

Jukes & Cantor

To:	A	G	C	T	

	A	1-3a	a	a	a
From:	G	a	1-3a	a	a
	C	a	a	1-3a	a
	T	a	a	a	1-3a

$$a = u \, dt$$

$$p = \frac{3}{4} (1 - e^{-\frac{4}{3} u t})$$

$$ut = -\frac{3}{4} \log_e (1 - \frac{4}{3} p)$$

(ut = evolutionary distance)

Substitution Models

- Jukes & Cantor (0 parameters)
 - All rates equal
- Kimura (2 parameters)
 - Transition & transversion different
- HKY (6 parameters)
 - Plus different frequency of four nucleotides
- Generalized time reversible
 - 8 parameters

Alignment Mask

- Use to remove unstable regions from phylogenetic analysis
 - Inserts and deletes make homolog assignment difficult
- Rapidly changing regions add noise
- Should be reported in publication (but often are not)
- Alignment masks should be provided with alignment
- Can be created by using ad-hoc rules
 - Eg 50% residues in column

Phylogenetic Methods

- UPGMA
 - Assumes clock
- Neighbor Joining
 - No clock, distance based
- Parsimony
 - Slow, used in other fields
- Maximum Likelihood
 - Slow, but fast converging, allows different rates
- Bayesian
 - Can be faster than maximum likelihood

Confidence Estimation

- Bootstrapping alignment columns
 - Estimates confidence for individual branch order
 - Sensitive to “unstable” sequences
- Kishino-Hasegawa-Templeton test
 - Compares whole trees
 - Asks one is significantly better
 - Uses Maximum Likelihood

Methods

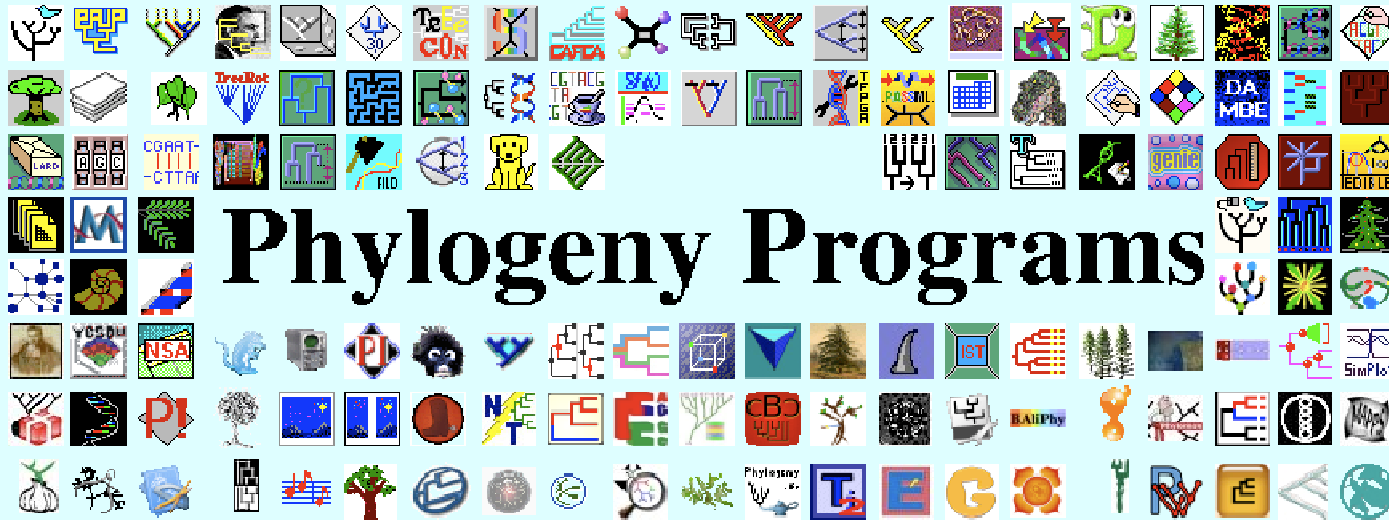
By computer

Cross-referenced

Data types

New programs

Submitting



Changes

Waiting list

Other lists

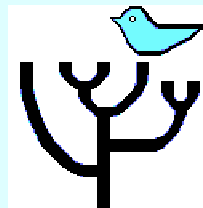
Old programs

Not listed

???

Here are 385 phylogeny packages and 52 [free servers](#), all that I know about. It is an attempt to be completely comprehensive. I have not made any attempt to exclude programs that do not meet some standard of quality or importance. Updates to these pages are made roughly weekly. [Here](#) is a "waiting list" of new programs waiting to have their full entries constructed. Many of the programs in these pages are available on the web, and some of the older ones are also available from [ftp server machines](#).

The programs listed below include both free and non-free ones; in some cases I do not know whether a program is free. I have listed as free those that I knew were free; for the others you have to ask their distributor. Usually when I say that a program is downloadable from a web site, this means that it is available free.



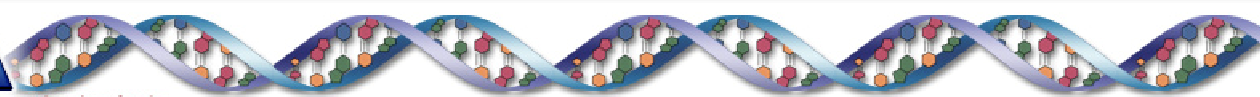

PHYLIP

PHYLIP version 3.68 was released on 13 August 2008. This is a "stable" bug-fix release. An "alpha" pre-release of PHYLIP 3.7, called 3.7a, will be released in a few months.

PHYLIP is a *free* package of programs for inferring phylogenies. It is distributed as source code, documentation files, and a number of different types of executables. These Web pages, by [Joe Felsenstein](#) of the [Department of Genome Sciences](#) and the [Department of Biology](#) at the [University of Washington](#), contain information on **PHYLIP** and ways to transfer the executables, source code and documentation to your computer.

- A [general description](#) of **PHYLIP**.
- [Programs](#) in the **PHYLIP** package
- About the [Executables](#)
- About the [Source code](#) ... compiling it yourself
- The documentation web pages for **PHYLIP** can be read [here](#)
- [Get me PHYLIP](#)
- [How to install PHYLIP](#)
- [Frequently asked questions](#)
- An excellent guide to using PHYLIP with molecular data is available [here](#).
- [PHYLIP on the web](#) (HTML documentation, server services)
- [Current and future versions of PHYLIP \(including new features\)](#)
- [Older versions of PHYLIP, including version 2.5](#)

MEGA – Molecular Evolutionary Genetics Analysis


Molecular Evolutionary Genetics Analysis

Download MEGA ➡ Windows DOS/Win Mac Linux PDF Manual

Home

Overview

Features

Update History

About the Authors

Example Data

Online Manual

PDF Manual

📦 A Walk Through MEGA

Links

FAQ

Fixed Bugs


Report Bugs

User Discussion Forum

Suggestions Box


Acknowledgements

Contact Us



MEGA 4
© 1993-2008

KOICHIRO TAMURA
JOEL DUDLEY
MASATOSHI NEI
SUDHIR KUMAR



MEGA 4.1
BETA
Click to Download

New Features

- Excel and CSV Output
- Update Notification
- Interface Improvements

MEGA 4: Molecular Evolutionary Genetics Analysis

MEGA is an integrated tool for conducting automatic and manual sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, and testing evolutionary hypotheses.

MEGA 4 has been tested on the following Microsoft Windows® operating systems:

Windows 95/98, NT, 2000, XP, and Vista.

New Features:

- **Real-Time Caption Expert Engine**
A unique facility to generate detailed captions for different types of analyses and results. These captions are intended to provide detailed, natural language descriptions of the methods and models used in

Kumar S, Dudley J, Nei M & Tamura K (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics* 9: 299-306.

📄 [Download PDF](#)

Clearcut

The reference implementation for Relaxed Neighbor Joining (RNJ)

Evans, J., Sheneman, L., Foster, J.A., (2006) Relaxed Neighbor-Joining: A Fast Distance-Based Phylogenetic Tree Construction Method, *Journal of Molecular Evolution*, **62**:785-792.

Extremely efficient phylogenetic tree reconstruction



Neighbor joining (NJ) is a popular distance-based phylogenetic tree reconstruction method. It has nice theoretical properties, but suffers from an $O(N^3)$ time complexity. Popular implementations of traditional NJ cannot process datasets with more than a few thousand taxa.

The WEIGHBOR Homepage

Weighbor: Weighted Neighbor Joining

Created by William J. Bruno, Nicholas D. Socci, and Aaron L. Halpern.

New:

Weighbor 1.2 is here! Better and faster than weighbor 1.0. Upgrade now! [Read more.](#)

Please cite: William J. Bruno, Nicholas D. Socci, and Aaron L. Halpern *Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction*, [Mol. Biol. Evol. 17 \(1\): 189-197 \(2000\).](#)

Weighbor is a weighted version of Neighbor Joining that gives significantly less weight to the longer distances in the distance matrix. The weights are based on variances and covariances expected in a simple Jukes-Cantor model. The criterion for which pair is joined is based on a likelihood function on the distances. The resulting trees are less perturbed by adding distant taxa compared to Neighbor Joining, and negative branch lengths are avoided. The method does not suffer from long branch attraction as maximum parsimony and other methods do. The method is much faster than maximum likelihood, usually faster than maximum parsimony, and a lot slower than Neighbor Joining.

HOW TO USE WEIGHBOR

Weighbor (or weighted neighbor joining) is PHYLIP compatible. You must create a distance matrix, such as by using the [PHYLIP](#) program DNADIST, or, the [least-squares method of Goldstein and Pollock](#).

Distances should always be given in units of substitutions per site; scaling distances by a constant can radically change the tree weighbor makes! If your distance matrix represents percent change, the values must be multiplied by .01 before passing them to weighbor. If your data includes a distance of infinity, it should be



MrBayes: Bayesian Inference of Phylogeny

[Home](#)

[Download](#)

[Manual](#)

[Online](#)

[Help](#)

[Bug](#)

[Report](#)

[Authors](#)

[Wiki](#)

[Links](#)

MrBayes is a program for the Bayesian estimation of phylogeny. Bayesian inference of phylogeny is based upon a quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observations. The conditioning is accomplished using Bayes's theorem. The posterior probability distribution of trees is impossible to calculate analytically; instead, MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees.

The program takes as input a character matrix in a NEXUS file format. The output is several files with the parameters that were sampled by the MCMC algorithm. MrBayes can summarize the information in

Phylogeny Programs

- Phylogeny Program List:
<http://evolution.genetics.washington.edu/phylip/software.html>
- Phylip
<http://evolution.genetics.washington.edu/phylip>
- MEGA
<http://www.megasoftware.net/>
- Clearcut
<http://bioinformatics.hungry.com/clearcut/>
- Weighbor
<http://www.t6.lanl.gov/billb/weighbor/index.html>
- Mr. Bayes
<http://mrbayes.csit.fsu.edu/>

The Ribosomal Database Project

Sequences and tools

Aligned and annotated sequences

Ribosomal Database Project II

ABOUT | ANNOUNCEMENTS | CITATION | CONTACTS | RESOURCES | RELATED SITES

Release 9.57 :: Jan 7, 2008 :: 471,792 16S rRNAs
(More on Release 9 and monthly is available in the [release notes](#).

tutorials and help

RDP video tutorials

myRDP login

upload and align your own 16S sequences in your private myRDP space and use the new analysis Pipeline

News

01/07/2008
RDP 9, Update 57 Release includes a new Genome viewing sequences from genome project. [\[more\]](#)

08/08/2007
RDP 9, Update 53 Released using TOBA Release 7.7 (The Taxonomic Outline of Bacteria and Archaea).

interactive online tools

RDP Analysis Tools

- myRDP** - Align and Classify your 16S rRNA sequences. Use the RDP Pipeline to process sequence libraries from raw sequencer output to analysis.
- Tree Builder** Create a phylogenetic tree.
- Hierarchy Browser** - Browse a phylogenetic hierarchy of sequences for download or use.
- Classifier** - Assign 16S rRNA sequences to our taxonomical hierarchy.
- Library Compare** - Compare two sequence libraries.
- Sequence Match** - Upload your sequence and search.
- Probe Match** - See what your probe targets in the database.
- Other Resources** - Alignment files, ASM posters, etc.
- Release 8.1** - The oldest release in the database.

powerful search and selection features

The Ribosomal Database Project (RDP) provides ribosomal sequence data for the scientific community, including online data analysis tools for small-subunit 16S rRNA sequences.

Sponsors:

National Science Foundation
Office of Biological and Environmental Research
National Institutes of Health

RDP HOME | BROWSER | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE | myRDP | seqCART

Hierarchy Browser

7 sequences selected; 7 match your data set

[start over | help | publication view | genomes | download]

Display depth: Auto
e coli
146 sequences

view by publication or genome

click node to return it to hierarchy view):

Bacteria (7/164872/146) ; Proteobacteria (7/56215/122) ; Gammaproteobacteria (7/26787/91)

Hierarchy View:

+	order	Enterobacteriales (7/5454/73)	(selected/total/search matches)	[options]
+	family	Enterobacteriaceae (7/5454/73)		
+	genus	Buchnera (0/56/1)		
+	genus	Citrobacter (0/428/3)		
+	genus	Photothabdus (7/93/7)		

<input checked="" type="checkbox"/>	S000528568	Photothabdus luminescens subsp. laumondii TTO1; BX571859
<input checked="" type="checkbox"/>	S000528570	Photothabdus luminescens subsp. laumondii TTO1; BX571860
<input checked="" type="checkbox"/>	S000528572	Photothabdus luminescens subsp. laumondii TTO1; BX571861

Probe Match : CCTTCGCCACCGGCCTTCC

Display depth: 10

Errors Allowed: 0

Probe: 5'CCTTCGCCACCGGCCTT
Target: 5'GGAAGGCCGGTGGCGAA

Lineage (click node to return it to hierarchy view):

Bacteria (352/273300)

Hierarchy View:

phylum Nitrospira (338/937) (hits/total searched) [list results for this phylum]
class Nitrospira (338/937)
order Nitrospirales (338/937)
family Nitrospiraceae (338/937)
▶ genus Nitrospira (175/489)
▼ genus Leptospirillum (96/186)

S000005187 Leptospirillum ferrooxidans; X72852
S000007286 unidentified bacterium; OS7; X86773
S000012329 unidentified bacterium; OS4; X86770
S000014647 Leptospirillum ferrooxidans (T); L15; DSM 2705; X86776
S000016917 Leptospirillum sp. LA; AJ237902
S000022717 Leptospirillum sp. DSM 2391; AJ237903
S000136476 uncultured Leptospirillum sp.; DGGE B42; A1517443

fast search algorithm,
limit searches to sequences spanning specific regions,
change depth and edit distance

Classifier - Start

Introduction

Use our classifier to assign 16s rRNA sequences to the taxonomical hierarchy proposed in release 6.0 of the nomenclature. The classifier uses a Bayesian rRNA classifier to assign sequences to the taxonomical hierarchy proposed in release 6.0 of the nomenclature. The classifier uses a Bayesian rRNA classifier to assign sequences to the taxonomical hierarchy proposed in release 6.0 of the nomenclature.

Help topics: Taxonomy

place sequences into bacterial taxonomy,
works well with partial or full-length sequences,
bootstrap confidence estimate,
prior alignment not required

Please enter your sequence

Did you know you can select sequences from myRDP and Hierarchy Browser to do seqmatch. Percent identity scores will be reported for aligned sequences (limited to 2000).

Choose a file to upload: no file selected

Cut and paste sequence(s) (in Fasta, GenBank, or EMBL format):

>ndw1_A04.folder=treeHole length=780
CAGGCCTAACACATGCAAGTCGAGGGGTATAGTTCTTCGGAAGTACA

Seqmatch :: Selectable Matches for Query Seq

Query Sequence: S000432340|AF544016.1:<1..>1440, 1

Match hit format:

short ID, orientation, similarity score, S_ab score, unique com

Lineage:

domain Bacteria (9/20/100467) (selected/match/total RDP sequences)
phylum Proteobacteria (9/20/37273)
class Alphaproteobacteria (9/20/9032)
order Rhizobiales (0/11/4012)
family Rhizobiaceae (0/2/1007)
genus Rhizobium (0/1/732)
S000016490 0.901536 0.628 1398 Zoogloea ramigera; ATCC 19
unclassified_Rhizobiaceae (0/1/66)
S000661373 0.896648 0.635 1405 uncultured organism; ctg_NI
family Phyllobacteriaceae (0/7/385)
genus Mesorhizobium (0/7/249)
S000427998 0.893007 0.630 1393 Mesorhizobium amorphae (T); ACCC 19665; AF041442

finds nearest neighbor,
more accurate than BLAST,
uses "q-gram" matching method

Hierarchy Browser

0 sequences selected; 0 match your data set

[[start over](#) | [help](#) | [publication view](#) | [genomes](#) | [download](#)]

Display depth:

[More search tips](#)

25 sequence(s) matched your query.

Lineage (click node to return it to hierarchy view):

Bacteria (0/169768/25) ; Firmicutes (0/53000/25) ; Clostridia (0/33723/25)
(0/32454/25) ; Clostridiaceae (0/11326/25)

Browse and select
from taxonomic hierarchy

Hierarchy View:

± ▼ genus Clostridium (0/219)

- ☐ S000000299 Clostridium a
- ☐ S000002330 Clostridium a
- ☐ S000128471 Clostridium a
- ☐ S000129304 Clostridium a
- ☐ S000129666 Clostridium a
- ☐ S000129667 Clostridium s
- ☐ S000380962 Clostridium a
- ☐ S000390414 Clostridium fe
- ☐ S000390415 Clostridium fe
- ☐ S000437208 Clostridium a
- ☐ S000437209 Clostridium a

examine sequences
by publication

Hierarchy Browser - View by Publication

[[start over](#) | [browse](#) | [help](#)]

Filter by: Sequence Count >=

This table lists the publications for the sequences. The first column count indicates the number of sequences published by the reference on the same row.

	Count	Citation	PMID
View	18339	Ley R.E., Turnbaugh P.J., Klein S., Gordon J.I.; Nature 444(7122):1022-1023(2006).	17183309
View	15111	Frank D.N., St Amand A.L., Feldman R.A., Boedeker E.C., Harpaz N., Pace N.R.; Proc. Natl. Acad. Sci. U.S.A. 104(34):13780-13785(2007).	17699621
View	15111	Frank D.N., StAmand A.L., Feldman R.A., Boedeker E.C., Harpaz N., Pace N.R.; Submitted (21-JUN-2007) to the EMBL/GenBank/DDBJ databases. Molecular, Cellular, and Developmental Biology, University of Colorado, Porter Biosciences, Boulder, CO 80309-0347, USA	
	1831	Eckburg P.B., Bik E.M., Bernstein C.N., Purdom E., Dethlefsen L., Sargent M., Gill S.R., Nelson K.E., Relman D.A.; Science 308(5728):1635-1638(2005).	15831718
View	11575	Eckburg P.B., Bik E.M., Bernstein C.N., Purdom E., Dethlefsen L., Sargent M., Gill S.R., Nelson K.E., Relman D.A.; Submitted (28-MAR-2005) to the EMBL/GenBank/DDBJ databases. Division of Infectious Diseases & Geographic Medicine, Stanford	

Genome: *Clostridium acetobutylicum* ATCC 824

[[help](#)] [[back to genome browser](#)]

This organism is a known **Type Strain**.

RDP Taxonomy Lineage: *domain* Bacteria ; *phylum* Firmicutes ; *class* Clostridia ; *order* Clostridiales ; *family* Clostridiaceae ; *genus* Clostridium

NCBI Taxonomy Lineage: *no rank* root ; *no rank* cellular organisms ; *superkingdom* Bacteria ; *phylum* Firmicutes ; *class* Clostridia ; *order* Clostridiales ; *family* Clostridiaceae

Sequencing Status from NCBI: COMPLETE

Estimated Genome size: 4.13 Mbp

rRNA sequence from
genome projects

[References \(1\)](#) [Outside Links \(5\)](#) [16S rRNA Sequences \(11\)](#)



Links provided by the Genomic Rosetta Stone

Genomes OnLine Database: Gc00060	GOLD: Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding complete and ongoing genome projects around the world.
Genome Catalogue: 000550_GCAT	Database holding genome reports that conform to the MIGS/MIMS specification, as defined by the Genomic Standards Consortium
NCBI Taxonomic Database: 272562	The NCBI taxonomy database contains the names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence.
NCBI Genome Project : 77	The NCBI Entrez Genome Project collection of complete and draft genome assemblies, annotation, and analysis.
Integrated Microbial Genomes (IMG): 637000076	The Integrated Microbial Genomes (IMG) system (Nucleic Acids Research, 2006, Vol. 34, Database issue D344-D348) provides a framework for comparative analysis of the genomes sequenced by the Joint Genome Institute. Its goal is to facilitate the visualization and exploration of genomes from a functional and evolutionary perspective.

Provides links to other
genome-related information

Ribosomal Database Project-II

myRDP Personal/Collaborative Workspace



RDP HOME | BROWSER | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE | myRDP | seqCART

Welcome, RDP Demo! [\[account info\]](#) [\[logout\]](#)

[overview](#) | [upload](#) | [download](#) | [pipeline](#) | [help](#)

Overview

Clicking or **Selects** (Adds) - Clicking **Deselects** (Removes) for download and analysis

Aligned - Failed - Unaligned

group name (selected)	submitter id	date	project	total	pending	A	F	U
aphanizomenon (3)	rdpdemo@demo.edu	07 Nov, 06	cyli	3	0	2	1	0
T00955_1 (0)	rdpdemo@demo.edu	31 Oct, 06						
Mikro (0)	rdpdemo@demo.edu	26 Oct, 06	macro					
Test (0)	rdpdemo@demo.edu	26 Oct, 06	Test					
Tim (0)	rdpdemo@demo.edu	26 Oct, 06	Tim					
CO1a (0)	rdpdemo@demo.edu	25 Oct, 06						
CO2 (0)	rdpdemo@demo.edu	25 Oct, 06						
CO1 (0)	rdpdemo@demo.edu	25 Oct, 06						
Aeromonas hydrophila (0)	rdpdemo@demo.edu	25 Oct, 06						
Tim (0)	rdpdemo@demo.edu	25 Oct, 06	test					
S. citri (0)	rdpdemo@demo.edu	19 Oct, 06						

View Group List

gene: Bacteria 16S rRNA

group name: aphanizomenon

submitter id: rdpdemo@demo.edu

submit date: 07 Nov, 06

project: cyli

note:

total sequences: 3
2 seqs aligned successfully
1 seqs failed alignment

alignment status:
A aligned
F failed
U unaligned
P pending

upload and align your own
16S sequences in your
myRDP space

your gateway to the
high-throughput pipeline

share specific data
with research buddies

List of Sequences:

	status	seqname	description
<input checked="" type="checkbox"/>	F	AY335547	[org=AY335547 [org=AY335547]]
<input checked="" type="checkbox"/>	A	KA11	[org=KA11 [org=KA11]]
<input checked="" type="checkbox"/>	A	AF044269	[org=Aphanizomenon ovalisporum] Aphanizomenon ovalisporum Bacteria; Cyanobacteria; Nostocales; Nostocaceae; Aphanizomenon.

Ribosomal Database Project-II *myRDP* High-Throughput Pipeline

Library Run Stats

Logge [rdpdemo@dem

summary dotur/estimateS help Return to n

Review or Request Analysis of Your Run Plates

Library Run ID: LR00000096
Submitter Name: Demo
Lab Name: testlab1
Library Name Abbrev: demo
Library Name: demo
Vector: Invitrogen_pCR4-TOPO
Account Number:
Project: demo only
Folder:
Primer Name:
Gene: Bacteria 16S rRNA
chromatograms map to wells: Yes
Note(s): this lib run is for demo only

Download Options:

☒ Trimmed
☐ Full

Internal Masking
(Base Score < | error probability >)
20 | 0.01

Download Raw Sequence

Cutoffs:

0.8 Fraction > Q20
600 Length

Update

alignment status counts

ALIGNED	132
LOW-QUALITY SEQ	12

PLATE ID (aligned) # samples # passed avg. Qscore avg. length

T100057 (94)	96	92	1.0	756.8
T100058 (38)	48	35	0.8	606.7

[Load More Data](#)

Library-level information

Return to

Plate T100057 Detail

summary seq cart (2) help

[Return to this Library Run](#)

	1	2	3	4	5	6	7	8	9	10	11	12
A	A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12
B	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11	B12
C	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12
D	D01	D02	D03	D04	D05	D06	D07	D08	D09	D10	D11	D12
E	E01	E02	E03	E04	E05	E06	E07	E08	E09	E10	E11	E12
F	F01	F02	F03	F04	F05	F06	F07	F08	F09	F10	F11	F12
G	G01	G02	G03	G04	G05	G06	G07	G08	G09	G10	G11	G12
H	H01	H02	H03	H04	H05	H06	H07	H08	H09	H10	H11	H12

green=passed, red=failed, black=

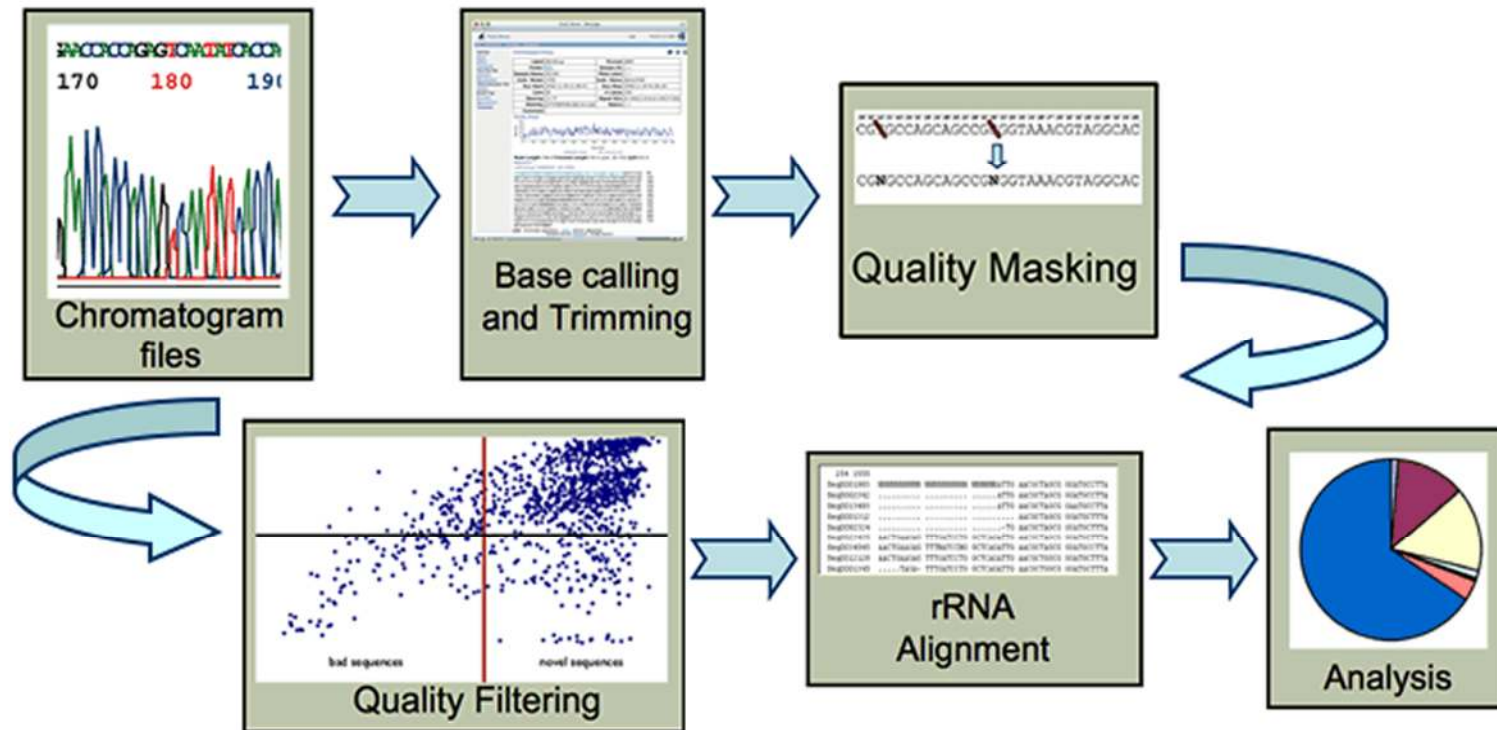
Alignment Status Legend

A ALIGNED
F FAILED ALIGNMENT
L LOW-QUALITY SEQ
N NEW SEQ
R READY FOR ALIGNMENT
S SUBMITTED FOR ALIGNMENT

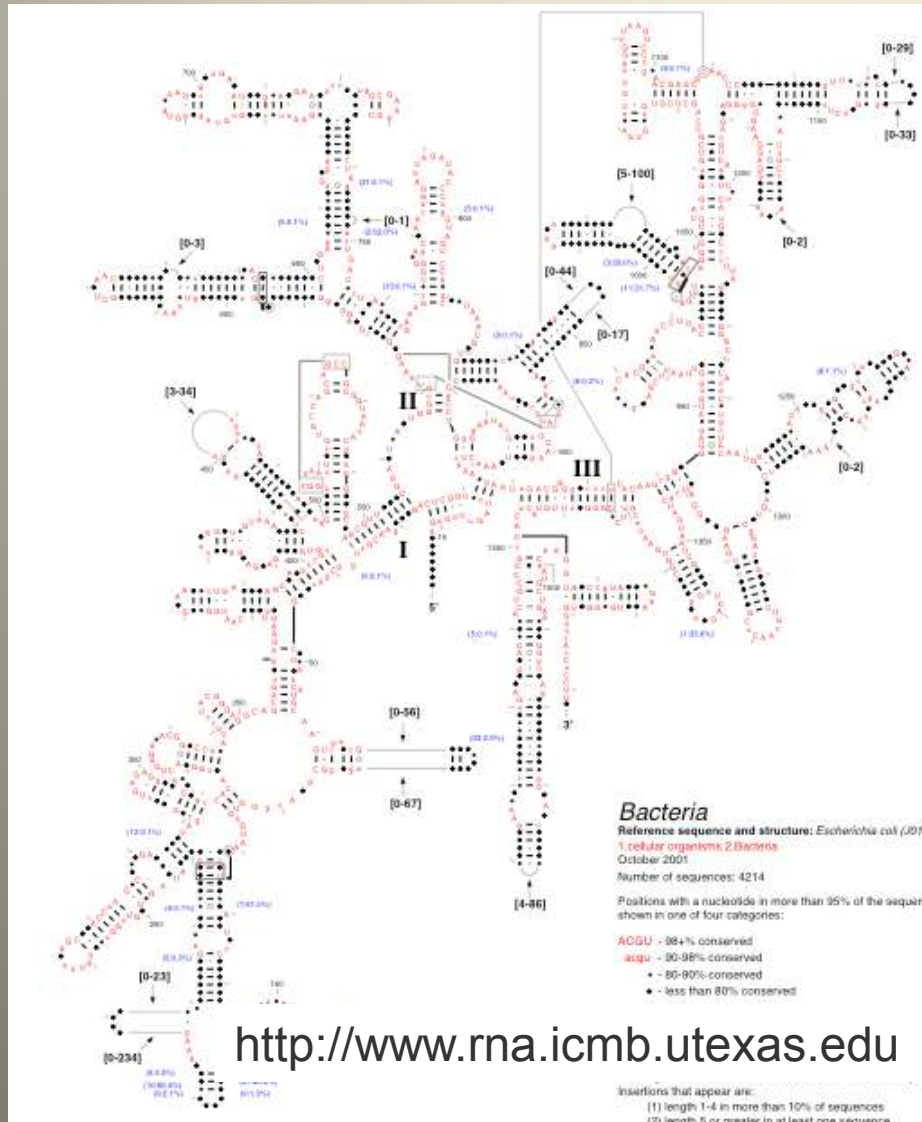
[Update Cart](#)

	Alignment Status	Well	QScore	Length	Passed	Vector
<input type="checkbox"/>	A	A01	0.9911	790	Y	
<input type="checkbox"/>	A	B01	0.9913	802	Y	
<input checked="" type="checkbox"/>	A	C01	0.9935	619	Y	
<input checked="" type="checkbox"/>	A	D01	0.9899	794	Y	top
<input type="checkbox"/>	A	E01	0.9912	798	Y	
<input type="checkbox"/>	A	F01	0.9875	797	Y	
<input type="checkbox"/>	A	G01	0.9898	787	Y	
<input type="checkbox"/>	A	H01	0.9948	766	Y	
<input type="checkbox"/>	A	A02	0.9813	642	Y	
<input type="checkbox"/>	A	B02	0.9912	797	Y	

Ribosomal Database Project-II *myRDP* Single-Read Pipeline

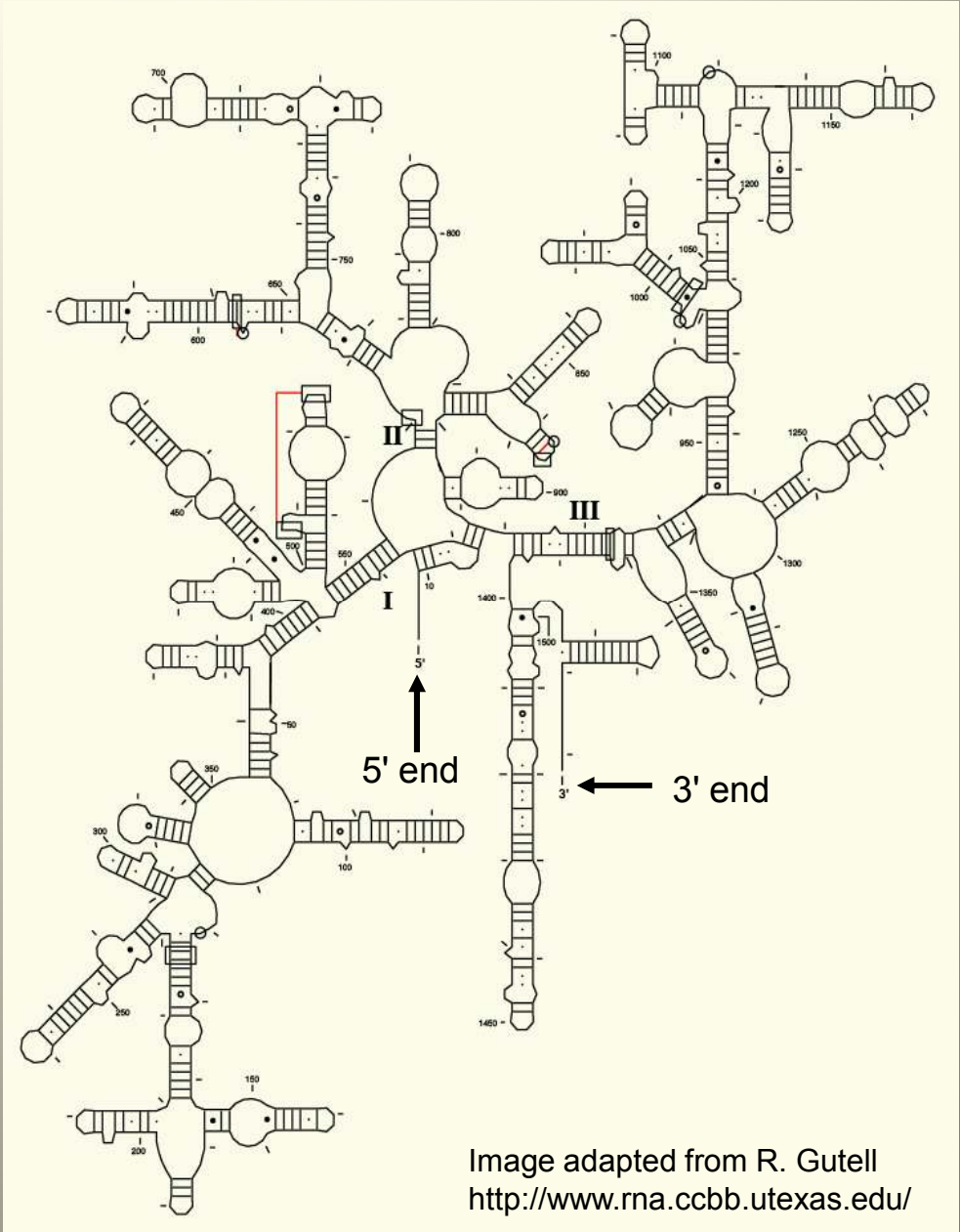


RDP 10 Alignment



- INFERNAL Stochastic Context-Free Grammar Aligner (Eddy)
- INFERNAL (Eddy)
 - Fast
 - Replaces RNAcad (Brown)
- Incorporates secondary structure information
- Probabilistic model
- New training set

Secondary structure
of small
subunit ribosomal
RNA



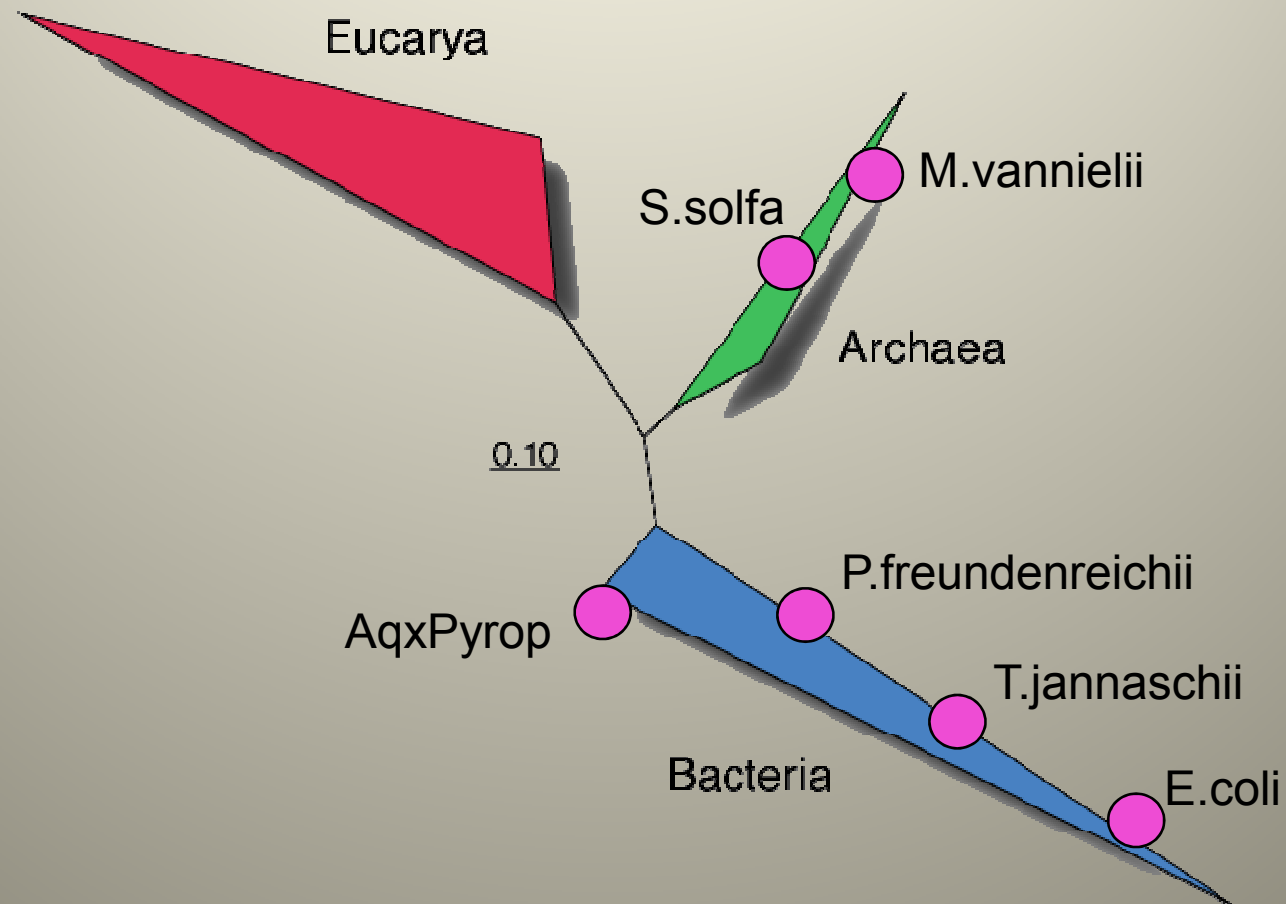
Unaligned rRNA sequences in a multiple alignment editor

[illegible]

Aligned rRNA sequences in editor

[illegible]

The Biosphere



References

Acknowledgement of rRNA secondary structure image:

- Cannone J.J., Subramanian S., Schnare M.N., Collett J.R., D'Souza L.M., Du Y., Feng B., Lin N., Madabusi L.V., Müller K.M., Pande N., Shang Z., Yu N., and Gutell R.R. (2002). The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and Other RNAs. *BioMed Central Bioinformatics*, 3:2. [Correction: *BioMed Central Bioinformatics*. 3:15.]
- Smith T.F., Gutell R., Lee J., and Hartman H. 2008. The origin and evolution of the ribosome. *Biology Direct*, 3:16.
- Woese CR. 1987. Bacterial evolution. *Microbiol Rev.* 1987 51(2):221-71.
- Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *J Theor Biol.* 8(2):357-66.
- Cole, J., Wang, Q., Cardenas, E., Fish, J., Chai, B ., Farris, R., Kulam-Syed-Mohideen, A., McGarrell, D., Marsh, T., Garrity, G. and Tiedje, J. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acid Research*. 2009. In press.

Sequence Alignment

Accuracy, Time, Memory

Multiple Sequence Alignment

- Pairwise dynamic programming
 - Smith-Waserman, Needleman Wunsch
 - Can be transformed into probabilistic framework
- Multidimensional dynamic programming
 - Not practical
- Progressive alignment
 - Muscle, ClustalW
 - Both are progressive iterative

BLAST

- Heuristic search strategy
- Locate high-scoring short matches
 - 3aa or 5 to 11 bases
- Extend short matches
- Determine significance using extreme value distribution statistics

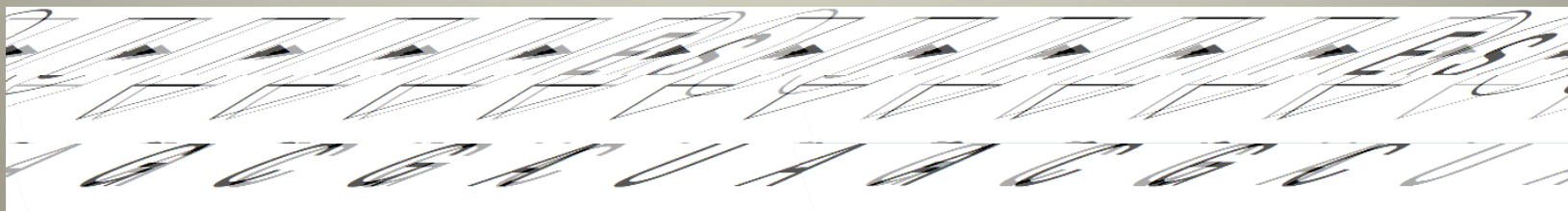
BLAST (cont.)

- E value
 - Database dependent
- Bits
 - Database independent
- % Similarity (identity)
 - For aligned segment s
 - NOT overall % identity

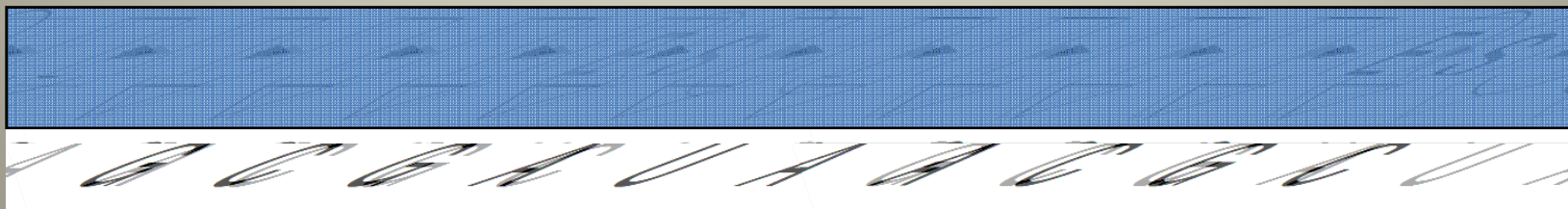
Model Based Alignment

- Profile Hidden Markov Models
 - Protein and nucleic acid
 - Models primary sequence
- Stochastic Context-Free Grammars
 - Incorporates RNA secondary structure

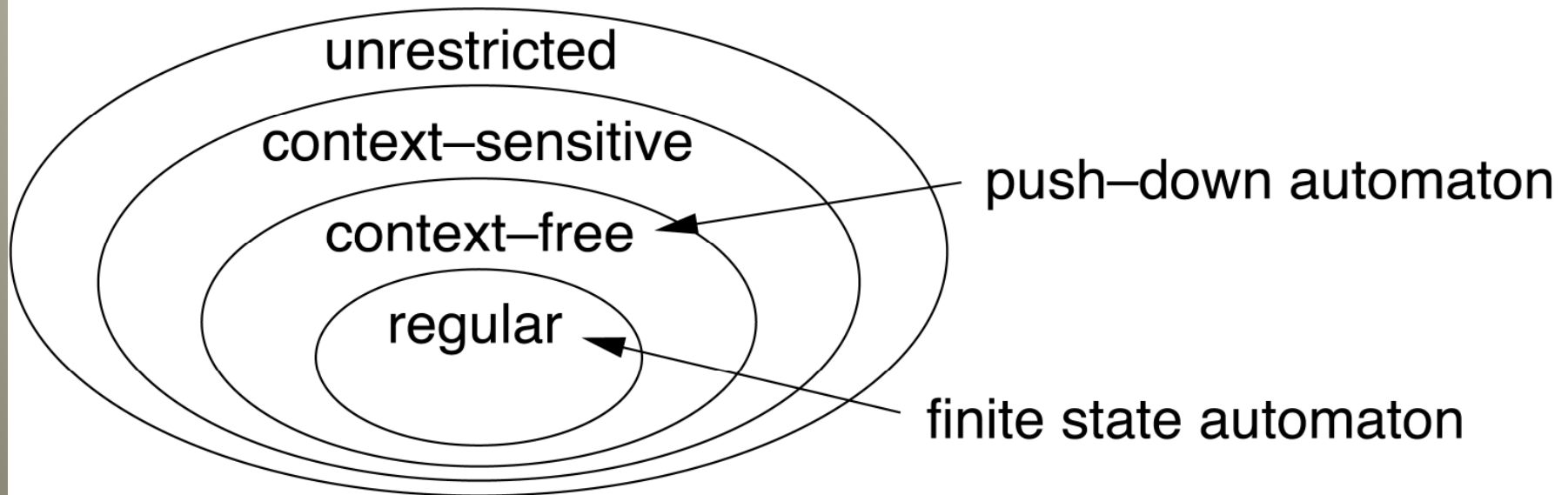
Hidden Markov Model



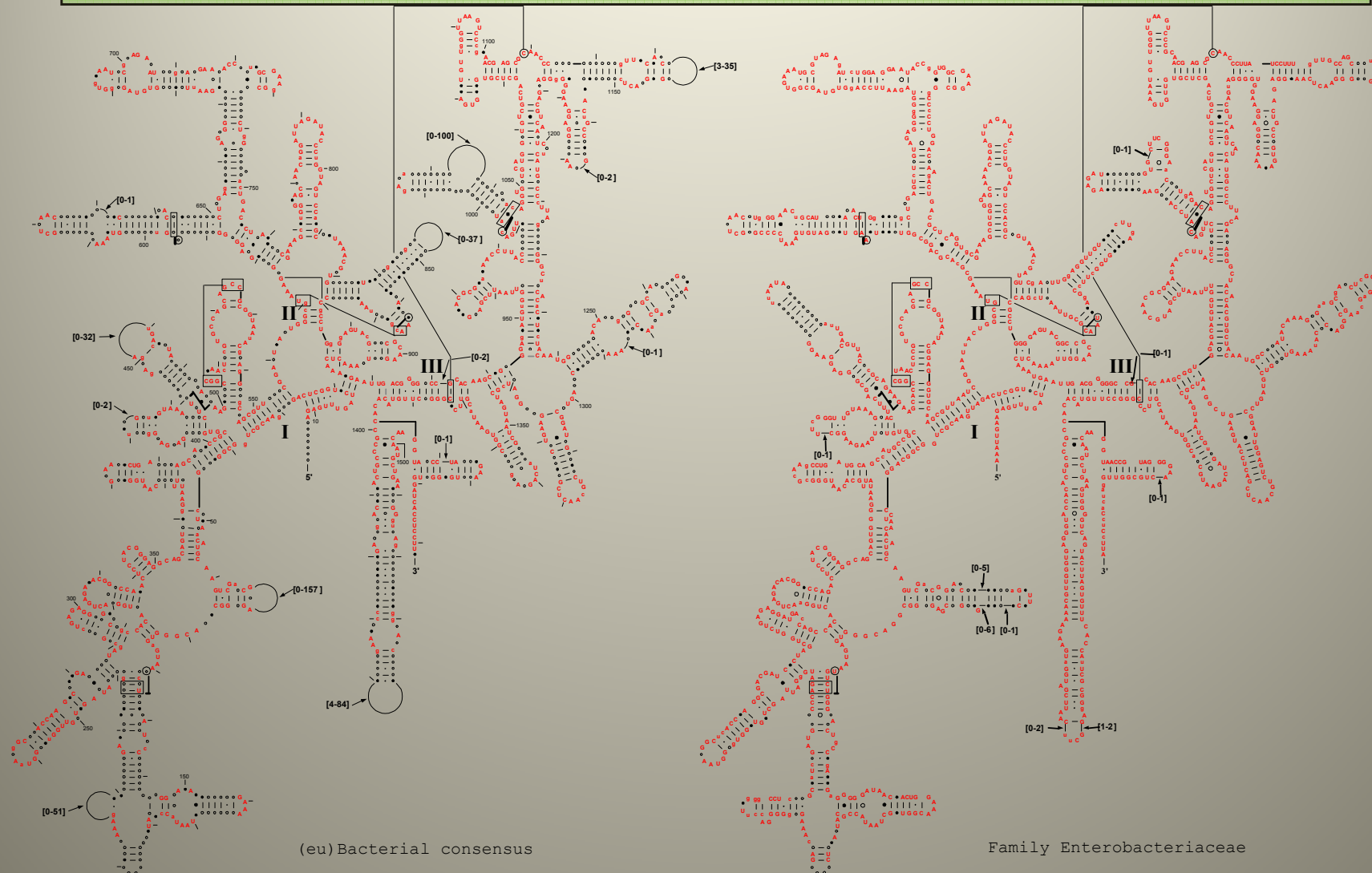
Hidden Markov Model



Chomsky Transformational Grammars



2D Structure Conserved from Domain to Family



Diagrams from the Gutell Lab Comparative RNA Web Site (<http://www.rna.icmb.utexas.edu>)

Aligner References

- MUSCLE
<http://www.drive5.com/muscle/>
- BLAST
<http://blast.ncbi.nlm.nih.gov/>
- HMMER
<http://hmmer.janelia.org/>
- INFERNAL
<http://infernal.janelia.org/>

Distance Calculation

- Phylogenetic methods only score base substitution, not insertion or deletion.
- Score comparable positions
 - Mask out unaligned regions, insertions
 - Ignore positions with deletion

Other Common Distances

- Hamming distance
 - No gap - insert
 - Original Blast
- Edit distance
 - Penalize for gaps
 - RDP Probe Match
- Matching word percentage (q-gram)
 - Does not require alignment
 - RDP Sequence Match

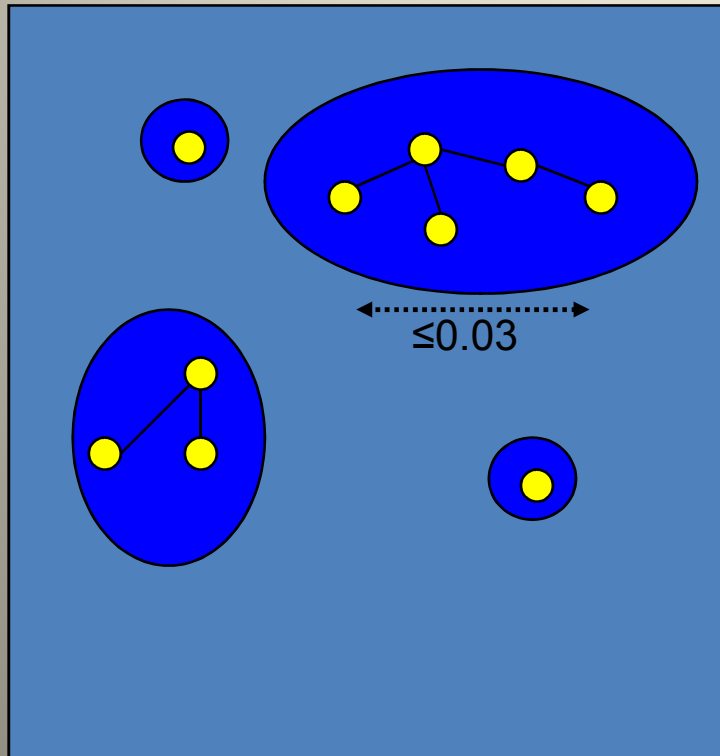
Clustering

Accuracy, Time, Memory

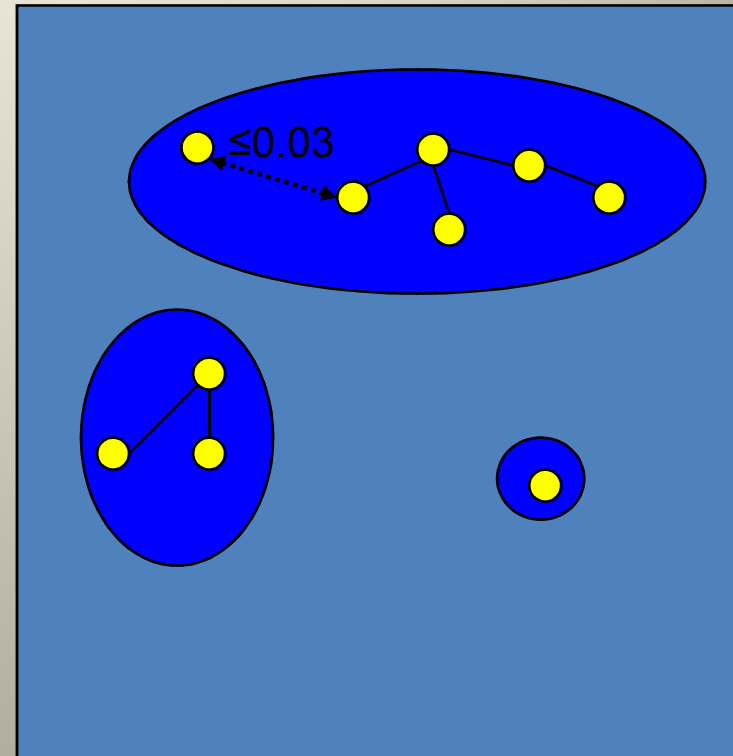
Unsupervised Classification (Clustering)

- Hierarchical Agglomerative
 - Single Linkage (Nearest neighbor)
 - Average Linkage (UPGMA)
 - Complete Linkage (Furthest Neighbor)
- Partitional Clustering
 - K-Means
 - Not often used in this field
- Self Organizing Maps
 - Using word frequency

Hierarchical Clustering



Complete Linkage



Single Linkage

CD-HIT



**Representative
Sequences.**

Contents

- [Download!](#)
- [News & Update](#)

- [CD-HIT Home](#)
- [Author's Home](#)
- [Manual](#)

- [Project Page](#)
- [Project CVS](#)
- [Project FTP](#)
- [Project BUGS](#)

- [References](#)
- [Related Resources](#)

- [Support](#)
- [Thanks!](#)



Bioinformatics.Org



Welcome to the CD-HIT Project Main Page

CD-HIT stands for Cluster Database at High Identity with Tolerance. The program (cd-hit) takes a [fasta format](#) sequence database as input and produces a set of 'non-redundant' (nr) [representative sequences](#) as output. In addition cd-hit outputs a [cluster](#) file, documenting the sequence 'groupies' for each nr sequence representative. The idea is to reduce the overall size of the database without removing any sequence information by only removing 'redundant' (or highly similar) sequences. This is why the resulting database is called [non-redundant](#) (nr). Essentially, cd-hit produces a set of closely related [protein families](#) from a given fasta sequence database.

CD-HIT uses a 'longest sequence first' [list removal algorithm](#) to remove sequences above a certain [identity](#) threshold. Additionally the algorithm implements a very fast heuristic to find high identity segments between sequences, and so can avoid many costly full alignments.

With recent developments, cd-hit package offers new programs for DNA sequence clustering and comparing two databases. It also has lots of new options for clustering control.

CD-HIT was originally written by [Weizhong Li](#) and is now an [open source](#) project!

Bugs

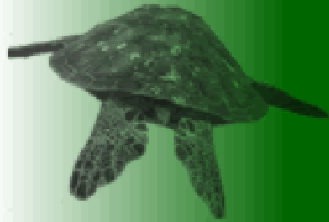
There are a number of [outstanding bugs](#) in the current implementation. We are always looking for hard working and enthusiastic volunteers (people like [Luc Ducazu](#)) to shoot these problems down.

Sub Projects

The CD-HIT project provides a number of opportunities for interesting research activities. If one of these sub-projects takes your interest why not join up and take part? We are especially keen to work closely with bioinformatics MSc students working on their MSc projects.

- [MyCD-HIT](#). A CD-HIT implementation embedded in a [MySQL UDF](#)!
- [Clustering Benchmarks](#). Develop and implement benchmarks to test clustering behavior.
- CD-HIT CGI. On-line access to the algorithm.

Related Resources



FastGroupII

[Tools Home](#)[FastGroupII](#)[Database](#)[Calculation](#)[User's Guide](#)

FastGroupII

[FastGroupII analysis](#)[ClustalW analysis](#)

Download

[FastGroup1.0](#)[Converter](#)

Welcome to FastGroupII

Very little is known about prokaryotic diversity on coral reefs, the most biodiverse of all marine ecosystems. To address this issue we have been sequencing 16S rDNAs from *Archaea* and *Bacteria* associated with corals ([1](#), [2](#)). To help with the analyses, the FastGroup project was started in 2001. FastGroup will take large 16S rDNA datasets, trim them, and determine community structure (e.g., dereplication).

The first version of FastGroup ([3](#)) was a Java program and can be downloaded [here](#). FastGroupII is a web-based tool. It dereplicates large 16S rDNA libraries using several different algorithms, including the original FastGroup Percentage Sequence Identity (PSI), PSI with Gaps, Tree-parsing (ClustalW global alignment) ([4](#)), and Seq-Match (Sequence Match in [Ribosomal Database Project](#)) ([5](#)). FastGroupII also automatically calculates standard diversity and richness indexes, including the Shannon-Wiener index ([6](#)), Chao1 ([7](#)), and rarefaction ([8](#)).

Full online article: *FastGroupII: A web-based bioinformatics platform for analyses of large 16S rDNA libraries*, Yanan Yu, Mya Breitbart, Pat McNairnie and Forest Rohwer, *BMC Bioinformatics* 2006,7:57. Click [here](#).

To use FastGroupII, click [here](#).

Supervised Classification

- K-Nearest Neighbors
 - SeqMatch, Megan, easyTaxon
 - Last Common Ancestor
- Bayesian
 - RDP Classifier
- Kernel methods
 - Support Vector Machines

Thirty-One Years of rRNA Sequencing

Proc. Natl. Acad. Sci. USA
Vol. 75, No. 10, pp 4801-4805, October 1978
Biochemistry

Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*

(recombinant plasmids/DNA sequence analysis/*rrnB* cistron)

JÜRGEN BROSIUS, MARGARET L. PALMER, POINDEXTER J. KENNEDY, AND HARRY F. NOLLER

Thimann Laboratories, University of California, Santa Cruz, California 95064

Communicated by Robert L. Sinsheimer, July 26, 1978

ABSTRACT The complete nucleotide sequence of the 16S RNA gene from the *rrnB* cistron of *Escherichia coli* has been determined by using three rapid DNA sequencing methods. Nearly all of the structure has been confirmed by two to six independent sequence determinations on both DNA strands. The length of the 16S rRNA chain inferred from the DNA sequence is 1541 nucleotides, in close agreement with previous estimates. We note discrepancies between this sequence and the most recent version of it reported from direct RNA sequencing [Ehresmann, C., Stiegler, P., Carbon, P. & Ebel, J. P. (1977) *FEBS Lett.* 84, 337-341]. A few of these may be explained by heterogeneity among 16S rRNA sequences from different cistrons. No nucleotide sequences were found in the 16S rRNA gene that cannot be reconciled with RNase digestion products of mature 16S rRNA.

sequence have been confirmed, and additional errors have been found involving oligonucleotide sequences, ordering of oligonucleotides, and, in one instance, the location of a larger section of the primary structure. No nucleotide sequences were found that cannot be accounted for from the RNase digestion products of mature 16S rRNA.

METHODS

Cloning and Mapping of DNA. The 16S rRNA gene from the *rrnB* cistron of *E. coli* was cloned from two *EcoRI* restriction fragments of λ rif^d18 (17, 18) in the ColE1 plasmid vector. Determination of the location of the 16S rRNA sequences and restriction enzyme cleavage sites will be described elsewhere.

rRNA is becoming increasingly important in our current per-

Twenty-Eight Years Later

Proc. Natl. Acad. Sci., USA
Vol. 103, No. 32, pp 12115–12120, August 2006

www.pnas.org/cgi/doi/10.1073/pnas.0605127103

Microbial diversity in the deep sea and the underexplored “rare biosphere”

Mitchell L. Sogin^{*†}, Hilary G. Morrison^{*}, Julie A. Huber^{*}, David Mark Welch^{*}, Susan M. Huse^{*}, Phillip R. Neal^{*}, Jesus M. Arrieta^{*§}, and Gerhard J. Herndl[‡]

^{*}Josephine Bay Paul Center, Marine Biological Laboratory at Woods Hole, 7 MBL Street, Woods Hole, MA 02543; and [†]Royal Netherlands Institute for Sea Research, P.O. Box 59, 1790 AB, Den Burg, Texel, The Netherlands

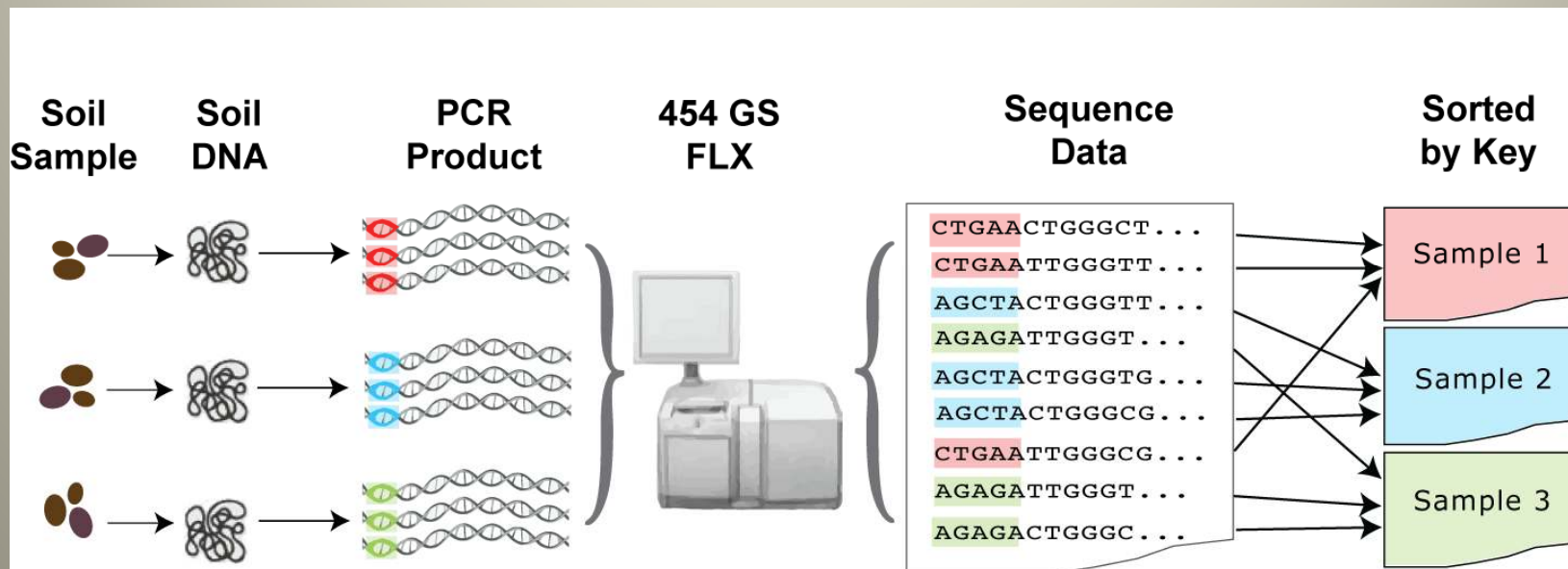
Communicated by M. S. Meselson, Harvard University, Cambridge, MA, June 20, 2006 (received for review May 5, 2006)

The evolution of marine microbes over billions of years predicts that the composition of microbial communities should be much greater than the published estimates of a few thousand distinct kinds of microbes per liter of seawater. By adopting a massively parallel tag sequencing strategy, we show that bacterial communities of deep water masses of the North Atlantic and diffuse flow hydrothermal vents are one to two orders of magnitude more complex than previously reported for any microbial environment. A relatively small number of different populations dominate all samples, but thousands of low-abundance populations account for most of the observed phylogenetic diversity. This “rare biosphere” is very ancient and may represent a nearly inexhaustible source of genomic innovation. Members of the rare biosphere are highly divergent from each other and, at different times in earth’s history, may have had a profound impact on shaping planetary processes.

biodiversity | low abundance | marine | microbes | rarefaction

Gene sequences, most commonly those encoding rRNAs, provide a basis for estimating microbial phylogenetic diversity (5, 7, 14–18) and generating taxonomic inventories of marine microbial populations (5, 7, 14–18). Evolutionary distances between orthologous sequences (19) or similarities to database entries identified through BLAST (20), FASTA (21), or Bayesian classifiers (22) identify operational taxonomic units (OTUs) that correspond to species or kinds of organisms. A variety of parametric and nonparametric methods extrapolate information from observed frequencies of OTUs or species abundance curves to predict the number of different microbial taxa in a local sample (23–26). Richness estimates of marine microbial communities through comparisons of rRNAs range from a few hundred phylotypes per ml in the water column (19) to as many as 3,000 from marine sediments (27, 28). One of the largest water column surveys (1,000 PCR amplicons) described the presence of only 516 unique sequences and estimated occurrence of

Multiplexed Amplicon Pyrosequencing





RDP'S PYROSEQUENCING PIPELINE

About the RDP's Pyrosequencing Pipeline

The Ribosomal Database Project's Pyrosequencing Pipeline aims to simplify the processing of large 16S rRNA sequence libraries obtained through pyrosequencing. This site processes and converts the data to formats suitable for common ecological and statistical packages such as SPADE, EstimateS, and R.

NCBI/EBI Submission Tools:

- **myRDP SRA Prekit Beta** - helps prepare and manage the complicated xml documents required for submitting your Pyrosequencing data to NCBI and EBI's Short Read Archive.
- **Fastq** - A java web start program that creates a Fastq file from a fasta and quality file. Requires Java 5 or above.

Data Processing Steps:

- **Pipeline Initial Process** - sort and trim the raw reads, filter low quality sequences.
- **Aligner** - align sequences using the fast, secondary-structure aware Infernal aligner.
- **Complete Linkage Clustering** - cluster sequences by the complete-linkage clustering method.

Formats for Common Programs:

- **SPADE Formatter** - make a SPADE compatible input format.
- **R Formatter** - make a R compatible input format.
- **EstimateS Formatter** - make an EstimateS compatible input format. Can also be used with PAST.
- **Mothur: Column Distance Matrix** - create a column distance matrix compatible with Mothur.
- **Mothur: Phylip Distance Matrix** - create a matrix and sample group file compatible with Mothur's LIBSHUFF function.

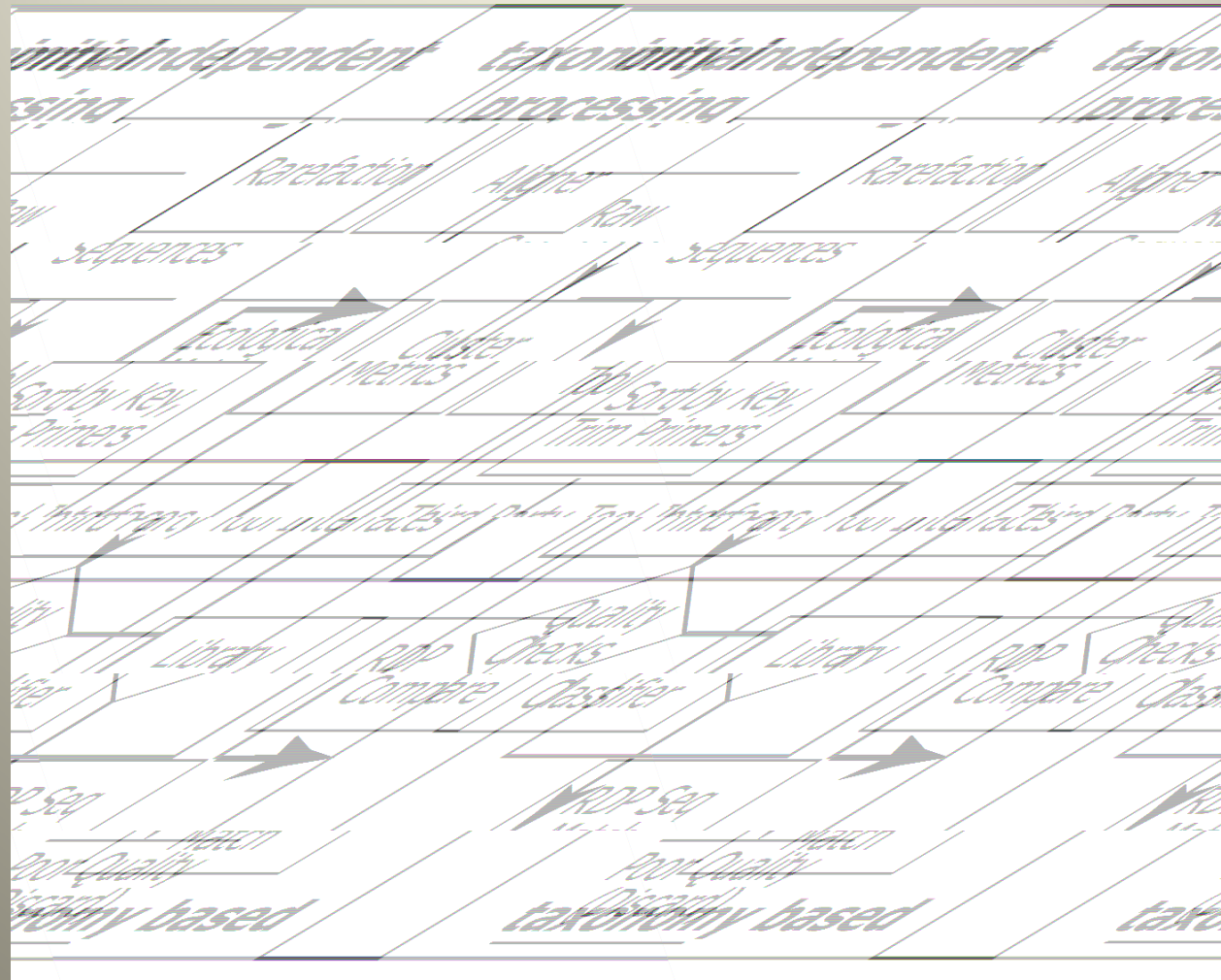
Analysis Tools:

- **Shannon & Chao1 Index** - calculate Shannon Index & Chao1 estimator from a single sample file.
- **Rarefaction** - calculate Rarefaction from a single sample file.
- **RDP Classifier** - assign 16S rRNA sequences to our taxonomical hierarchy.
- **RDP LibCompare** - compare two sequence libraries using the RDP Classifier.

Miscellaneous Utilities:

- **Alignment Merger** - merge multiple alignment files into one alignment.
- **Dereplicate** - use this tool to make representative sequences.
- **FASTA Sequence Selection** - make a sub-selection of FASTA sequences from the original sequence file.

RDP Pyrosequencing Pipeline



Initial Processing Steps

- Sort by barcode (key)
- Quality filter
 - Forward & (optional) reverse primers
 - Ambiguities
 - Length
- Trim key & primer sequences

Two Analysis Tracks

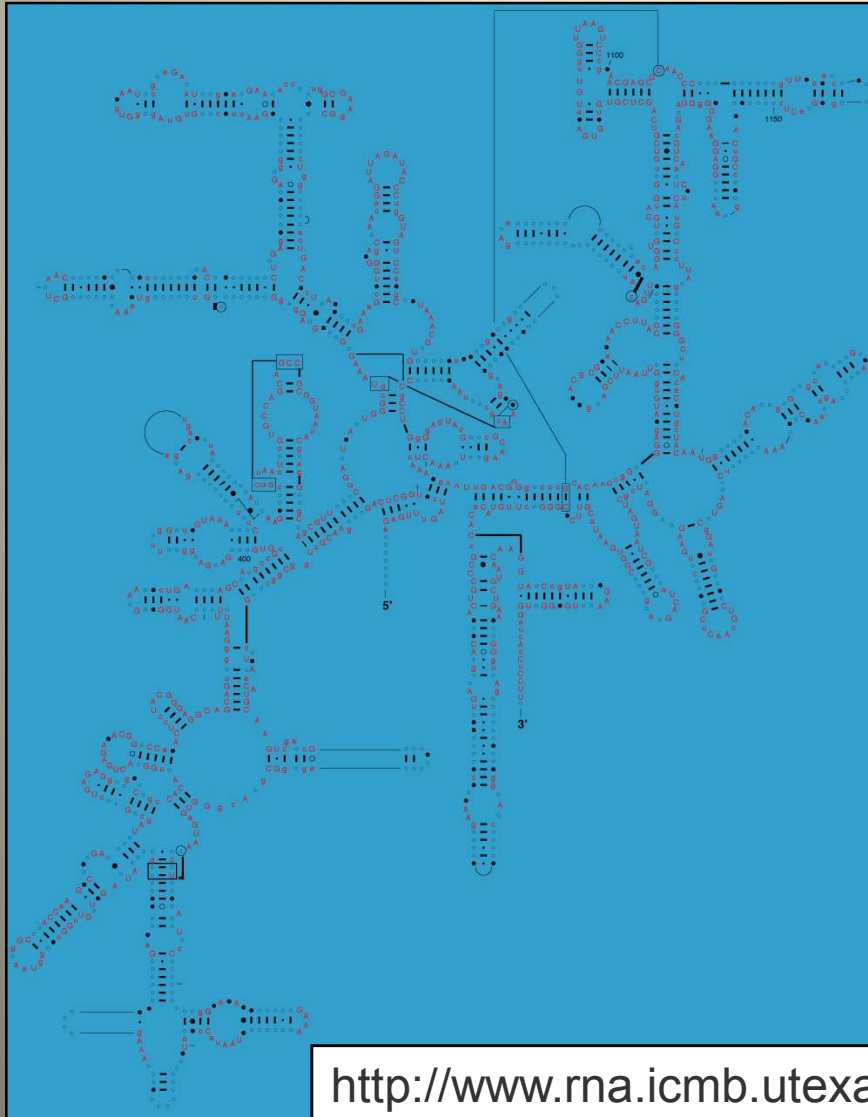
Taxonomy Independent

- Global Alignment
- Cluster Based OTU Assignment
- Standard Ecological Metrics
- Many 3rd Party Data Formats

Taxonomy Dependent

- RDP Classifier
- Sequence Match
- Many 3rd Party Data Formats

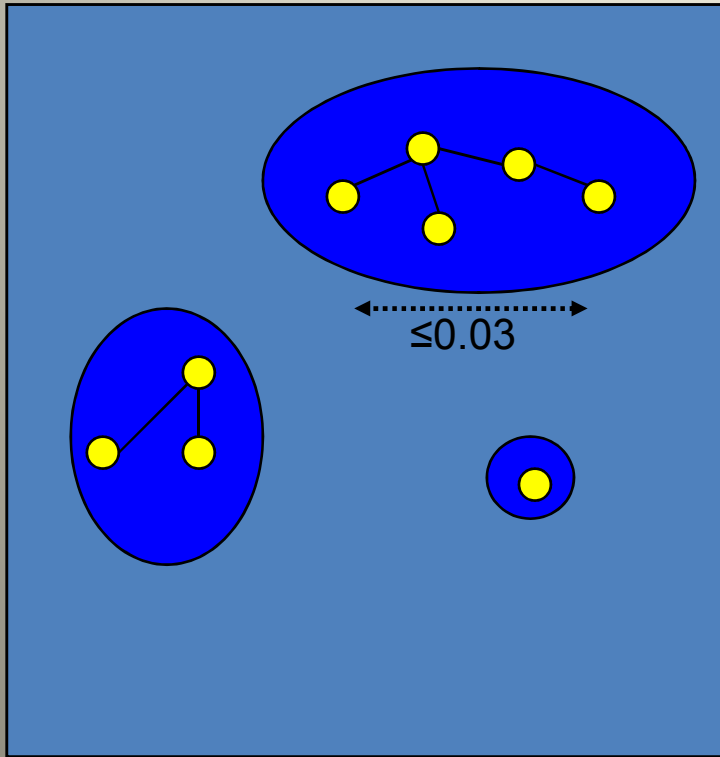
Model Based Alignment



<http://www.rna.icmb.utexas.edu>

- Infernal Aligner
 - (Nawrocki and Eddy. 2007, PLoS Comput Biol)
- Fast - 500/min
- Probabilistic Model
 - Model describes shared features
- Incorporates 2d Structure
 - Cannone et al. 2002, BioMed Central Bioinformatics

Complete Linkage Clustering (Operational Taxonomic Units)

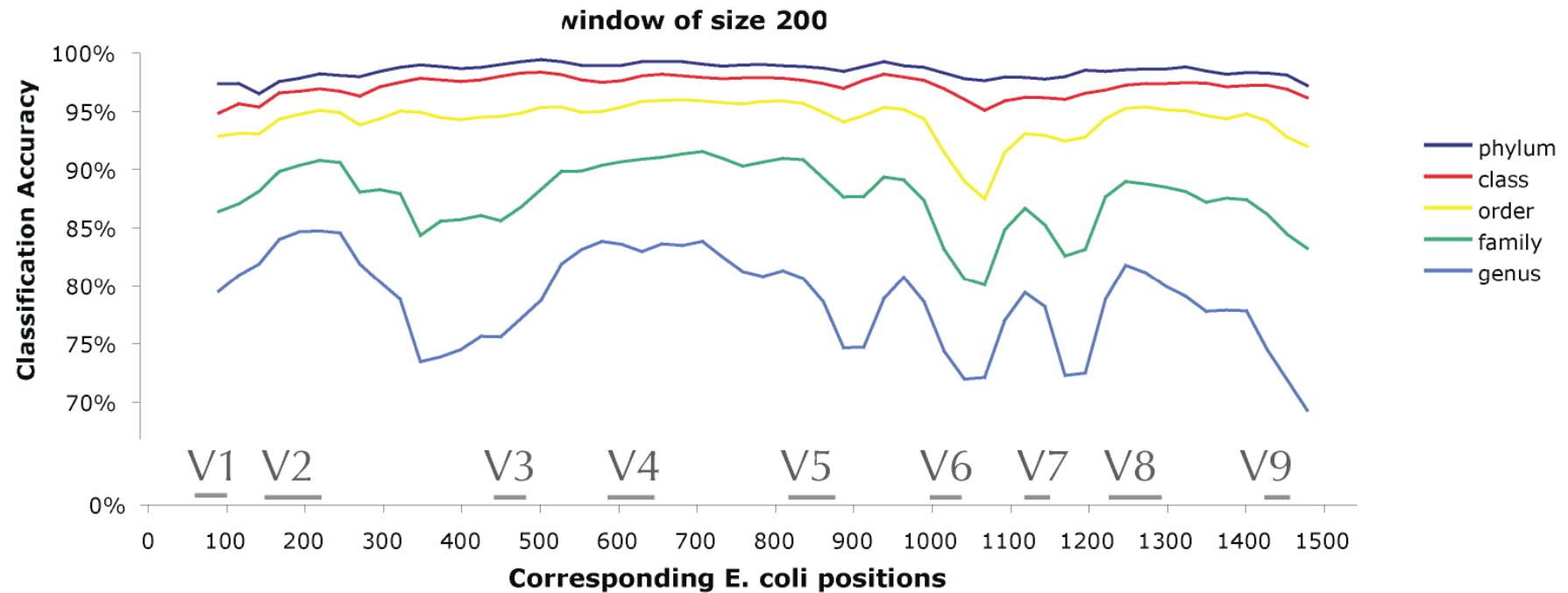


- Distance based method
- Guaranteed intra-cluster distance
- N^2 algorithm
- Current online limit 150,000 unique reads
- Memory-efficient version in testing

RDP Naive Bayesian Classifier

- Fast - 3000/min
- Places sequences into bacterial taxonomy
- Works well on partial or full-length sequences
- Does not require alignment
- Easily re-trained to match new taxonomies
- Bootstrap confidence estimates
- Online GUI - Soap service - Open source

Classifier Accuracy on 200 bp Regions



From Wang et. al., AEM, 2007

RDP Classifier Bootstrap Performance (Genus Level - Short Reads)

		V3				V6				V4	
Bootstrap cutoff	0%	50%	80%		0%	50%	80%		0%	50%	80%
Human Gut											
% classified	100	92.4	82.3		100	73.5	40.4		100	97.0	87.9
% matching	92.0	95.0	98.1		79.0	96.5	98.7		92.8	94.5	95.7
Soil											
% classified	100	71.3	48.3		100	32.7	16.7		100	74.4	56.3
% matching	70.0	85.5	94.6		48.0	80.0	84.3		84.1	93.3	96.8