

# **PHYLIP**

Joe Felsenstein

University of Washington, Seattle

# PHYLIP

- Distributed since 1980
- Originally in Pascal, now in C
- Intended to provide “basic transportation”
- Intended to provide a wide variety of methods
- Freely available (unless you try to charge others for it)

# Advantages of PHYLIP

1. Free (in the sense of “free beer”), easily obtainable
2. Runs on all major platforms
3. Very good documentation
4. Lots of people around who know how to use it
5. Runs can be automated by using input redirection and command files
6. Support for PHYLIP-format files by other programs such as ClustalW, MacClade and PAUP\*

Over 20,000 registered users in over 50 countries including: Fiji, Cuba, Papua New Guinea, Iran, Iceland. Large numbers of users in countries such as India, Brazil, Argentina, Russia, and China where even modest cash prices for software would be a major burden.

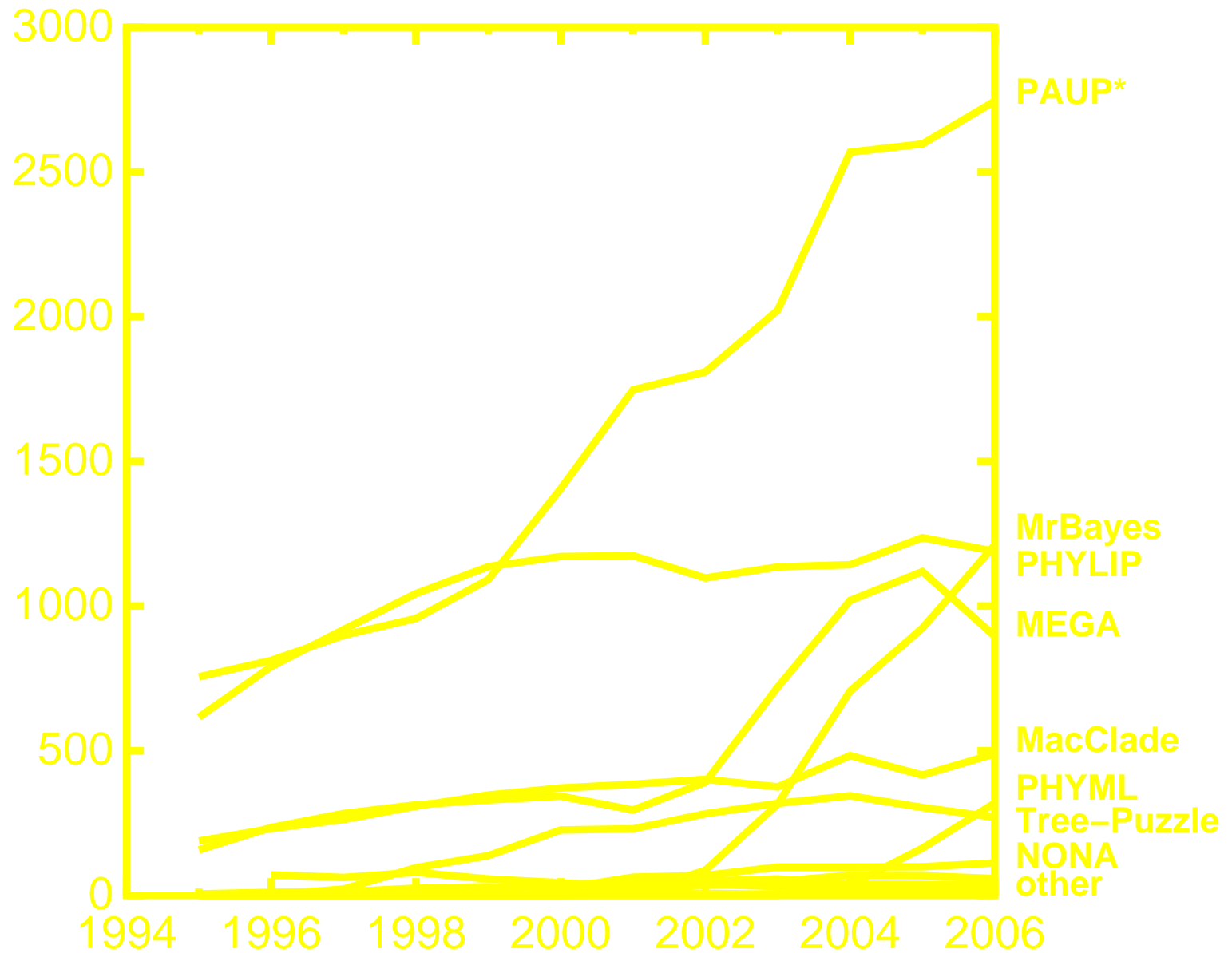
# Disadvantages of PHYLIP

1. Tree search less thorough than some other packages such as PAUP\*.
2. Much, much slower than packages such as PAUP\*
3. Character-mode interface is not mouse/windows GUI
4. Manual steps such as renaming file names can be tedious
5. Not as many options available as in other programs
6. Cannot read NEXUS standard files

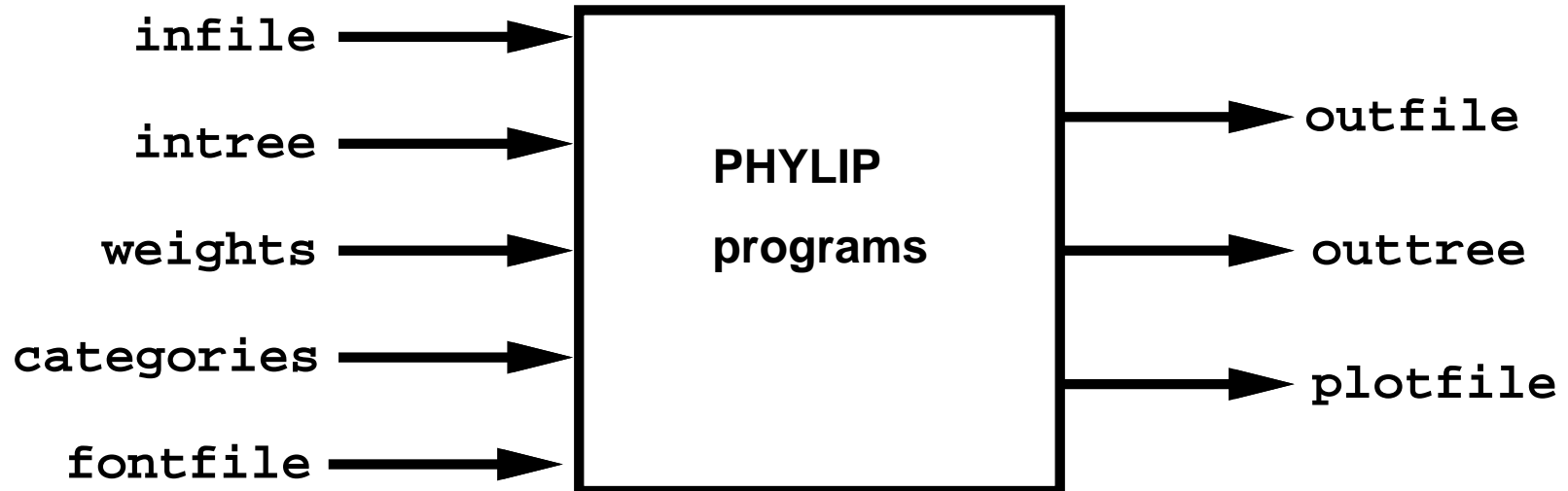
## PHYLIP is

- First in total numbers of copies distributed
- As of late 2005, it was second after PAUP\*, just ahead of Mega, and MrBayes in total numbers of trees published (see next graph).
- By now, PAUP\*, MEGA, and MrBayes are probably battling it out, with PHYLIP in fourth place.

# Numbers (papers) of published phylogenies



# PHYLIP programs



These are the default file names. If the input files do not exist (or if the output files exist and you choose not to overwrite them), you will be asked for the file name. This is not a bug.

# Input format for PHYLIP (DNA, Interleaved)

7 112

Bovine	CCAAACCTGT	CCCCACCATC	TAACACCAAC	CCACATATAC	AAGCTAAACC	AAAAATACCA
Mouse	CCAAAAAAC	ATCCAAACAC	CAACCCCAGC	CCTTACGCAA	TAGCCATACA	AAGAATATTA
Gibbon	CTATACCCAC	CCAACCTCGAC	CTACACCAAT	CCCCACATAG	CACACAGACC	AACAACCTCC
Orang	CCCCACCCGT	CTACACCAGC	CAACACCAAC	CCCCACCTAC	TATACCAACC	AATAACCTCT
Gorilla	CCCCATTTAT	CCATAAAAAC	CAACACCAAC	CCCCATCTAA	CACACAAACT	AATGACCCCC
Chimp	CCCCATCCAC	CCATACAAAC	CAACATTACC	CTCCATCCAA	TATACAAACT	AACAACCTCC
Human	CCCCACTCAC	CCATACAAAC	CAACACCACT	CTCCACCTAA	TATACAAATT	AATAACCTCC
	CCCCAGCCCA	ACACCCTTCC	ACAAATCCTT	AATATACGCA	CCATAAATAA	CA
	TCCCACCAA	TCACCCTCCA	TCAAATCCAC	AAATTACACA	ACCATTAACC	CA
	GCACGCCAAG	CTCTCTACCA	TCAAACGCAC	AACTTACACA	TACAGAACCA	CA
	ACACCCTAAG	CCACCTTCCT	CAAAATCCAA	AACCCACACA	ACCGAAACAA	CA
	ACACCTCAAT	CCACCTCCCC	CAAATACAC	AATTCACACA	AACAATACCA	CA
	ACATCTTGAC	TCGCCTCTCT	CAAACACAC	AATTCACGCA	AACAACGCCA	CA
	ACACCTTAAC	TCACCTTCTC	CAAACGCAC	AATTCGCACA	CACAACGCCA	CA

## Format for trees in tree files (Newick standard)

```
(Mouse:0.87231,Bovine:0.49807,(Gibbon:0.25930,(Orang:0.24166,  
(Gorilla:0.12322,(Chimp:0.13846,  
Human:0.08571):0.06026):0.04405):0.10815):0.39538);  
(Mouse:0.87558,Bovine:0.49718,(Gibbon:0.25698,(Orang:0.24477,  
((Gorilla:0.16328,Chimp:0.13802):0.01842,  
Human:0.08495):0.06610):0.10637):0.39287);  
(Mouse:0.87819,Bovine:0.49461,(Gibbon:0.25837,(Orang:0.24161,  
(Chimp:0.13941,(Gorilla:0.16639,  
Human:0.09533):0.00616):0.06709):10938):0.39630);
```

## Forms of distribution (as of version 3.6)

- Generic C source code and documentation. This can easily be compiled on Linux or Unix systems that have the X Windows windowing system. On Apple's Mac OS X (which is Unix) all but the tree drawing programs can compile without difficulty.
- Windows executables that will run on all Windows systems from Windows95 on, except Windows ce (there are executables of older versions that will run on Windows 3.1 and on 386 DOS systems).
- Native-mode Mac OS X executables that use the Aqua interface. (Note that the GCC compiler in Mac OS X can also be used to compile a version that will run under X Windows).
- Macintosh Mac OS systems running Mac OS 8 or 9. (There are executables of earlier versions that run on earlier "68k" Macs. The Mac OS 8 and 9 executables will probably be discontinued in version 3.7).
- Runs can be automated using "batch" files (command files) on Windows and on Linux/Unix systems (including Mac OS X). One makes a file of keystrokes that you issue to the menu, and gets these to the programs using input redirection.

# PHYLIP guide

A useful guide to using PHYLIP with molecular sequences has been produced by Jarno Tuimala. It can be downloaded as a PDF from

<http://koti.mbnet.fi/tuimala/oppaat/phylip2.pdf>

or using the link to it on the main PHYLIP web page.

## What to do in this exercise

1. Get a DNA or protein sequence data set of aligned sequences. You can use one of the ones provided by the course if you wish.
2. Estimate a tree by parsimony using `Dnapars` or `Protpars`.
3. Look at the tree by looking at the output file `outfile` and also by renaming `outtree` to `intree` and using `Drawgram`.
4. Compute a distance matrix using `Dnadist` or `Protdist`.
5. Rename `outfile` to `infile` and get a distance-based tree using `Fitch` and also using `Neighbor`. Examine each by looking at its `outfile` and also by using `Drawtree`.
6. Returning to the original sequences (you did save them, didn't you?) run `Dnaml` or `Proml`. Use the `R` and `A` options to do a Gamma-distributed rates analysis with a coefficient of variation of rates of 2 and about 6 rate categories. Examine the result the same way as for the distance methods.
7. If there is time, use one of these methods to do a bootstrap analysis (using `Seqboot`, then a tree-making program, then `Consense` and renaming files when needed).

## You can also make nice plots of trees



Virtual Reality Markup Language

## Now for a demo ...

Some of the data sets we have available. Or use your own ...

cytb.prt	Cytochrome B protein sequences
eftualpha.prt	EFTU Alpha protein sequences
fifteen.dna	Large subunit rRNA sequences for 15 eukaryotes
fifteen.dst	A set of distances for the above
hasegawa.dna	Mitochondrial D-loop and adjacent 3rd positions for apes
primates.dna	The same, for a wider range of primates
mammhbb.prt	Mammalian hemoglobin beta sequences
mammhbb.dna	Mammalian hemoglobin beta DNA sequences
mammalspenny.prt	David Penny's concatenation of mammal proteins
sarich.dst	Immunological distances from Sarich 1969 Syst. Zool.
ten.prt	COX II protein sequences of mammals
turbeville.dna	Turbeville et al MBE 1994 large subunit rRNA for chordates
turbeville.wts	0/1 weights for the above identifying well-aligned regions

# How it was done

This presentation was prepared using freeware:

- LaTeX (mathematical typesetting and PDF preparation)
- prosper class for projection slides
- Idraw (drawing program to modify plots and draw figures)
- dvips to prepare Postscript file
- ps2pdf to turn this into a PDF
- Adobe Acrobat Reader (to display the PDF in full-screen mode)
- Linux (operating system)