Curso fundamenteal de Inferencia Filogenética Molecular

Pablo Vinuesa (vinuesa@ccg.unam.mx)

Progama de Ingeniería Genómica, CCG, UNAM

http://www.ccg.unam.mx/~vinuesa/

Tutor: PDCBM, Ciencias Biológicas, PDCBioq. y
Profesor de la Lic. Ciencias Genómicas y posgrado

- **Tutorial de uso del paquete PHYLIP**

1. Uso de PHYLIP de modo interactivo
2. Uso de PHYLIP desde scripts de Perl
3. Estima de máxima verosimilitud del parámetro alpha de la distribución gamma
4. Selección de modelos de sustitución de proteína bajo máxima verosimilitud

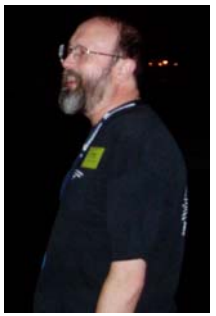
Inferencia filogenética usando el paquete

PHYLIP

(phylogeny inference package)

por

Joseph Felsenstein



Department of Genome Sciences,
University of Washington
Seattle, WA, USA

Inferencia filogenética usando el paquete

PHYLIP

- Distribuído como código fuente en C, desde 1980

http://evolution.gs.washington.edu/phylip.html

- también existen ejecutables para PCs y Macs (multiplataforma)
- Infiere filogenias por MP, compatibilidad, métodos de matrices de distancias, y ML
- También calcula árboles consenso, distancias entre árboles, hace remuestreo de datos (bootstrap), imprime y edita árboles, calcula matrices de distancias
- Maneja como datos alineamientos de nt y aa, matrices de frecuencias génicas, sitios de restricción, fragmentos de restricción, caracteres discretos y contínuos
- Es de los paquetes más utilizados para inferir filogenias y ES GRATIS, con excelente documentación

PHYLIP
– conversión de formatos y técnicas de remuestreo

SEQBOOT

Reads in a data set, and produces multiple data sets from it by **bootstrap resampling**. Since most programs in the current version of the package allow processing of multiple data sets, this can be used together with the consensus tree program CONSENSE to do bootstrap (or delete-half-jackknife) analyses with most of the methods in this package. This program also allows the Archie/Faith technique of permutation of species within characters. It can also **rewrite a data set to convert it from between the PHYLIP Interleaved and Sequential forms**, and into a preliminary version of a new XML sequence alignment format which is under development and which is described in the RETREE documentation web page.

Inferencia filogenética usando el paquete

PHYLIP - distancias

DNADIST

Computes four different **distances between species from nucleic acid sequences**. The distances can then be used in the distance matrix programs. The distances are the Jukes-Cantor formula, one based on Kimura's 2-parameter method, Jin and Nei's distance which allows for rate variation from site to site, and a maximum likelihood method using the model employed in DNAML (**F84**). The latter method of computing distances can be very slow.

PROTDIST

Computes a **distance measure for protein sequences**, using maximum likelihood estimates based on the **JTT**, Dayhoff PAM matrix, Kimura's 1983 approximation to it, or a model based on the genetic code plus a constraint on changing to a different category of amino acid. Rate variation from site to site is also allowed. The distances can be used in the distance matrix programs.

Inferencia filogenética usando el paquete

PHYLIP - distancias

FITCH

Estimates **phylogenies from distance matrix** data under the "**additive tree model**" according to which the distances are expected to equal the sums of branch lengths between the species. Uses the **Fitch-Margoliash criterion and some related least squares criteria**. Does **not** assume an evolutionary **clock.** This program will be useful with distances computed from molecular sequences, restriction sites or fragments distances, with DNA hybridization measurements, and with genetic distances computed from gene frequencies.

KITSCH

Estimates **phylogenies from distance matrix data** under the "**ultrametric**" model which is the same as the additive tree model except that an **evolutionary clock is assumed**. The **Fitch-Margoliash criterion and other least squares criteria are assumed**. This program will be useful with distances computed from molecular sequences, restriction sites or fragments distances, with distances from DNA hybridization measurements, and with genetic distances computed from gene frequencies.

NEIGHBOR

An implementation by Mary Kuhner and John Yamato of Saitou and Nei's "**NJ Method**," and of the **UPGMA** (Average Linkage clustering) method. **Neighbor Joining** is a distance matrix method producing an unrooted tree **without** the assumption of a **clock**. **UPGMA does assume a clock**. The branch lengths are not optimized by the least squares criterion but the methods are very fast and thus can handle much larger data sets.

PHYLIP - árboles

DRAWGRAM

Plots **rooted phylogenies, cladograms, and phenograms** in a wide variety of user-controllable formats. The program is interactive and allows previewing of the tree on PC or Macintosh graphics screens, and Tektronix or Digital graphics terminals. Final output can be to a file formatted for one of the drawing programs, on a laser printer (such as Postscript or PCL-compatible printers), on graphics screens or terminals, on pen plotters (Hewlett-Packard or Houston Instruments) or on dot matrix printers capable of graphics (Epson, Okidata, Imagewriter, or Toshiba).

DRAWTREE

Similar to DRAWGRAM but plots **unrooted phylogenies**.

CONSENSE

Computes **consensus trees** by the **majority-rule consensus tree** method, which also allows one to easily find the strict consensus tree. Is not able to compute the Adams consensus tree. Trees are input in a tree file in standard nested-parenthesis notation, which is produced by many of the tree estimation programs in the package. This program can be used as the final step in doing bootstrap analyses for many of the methods in the package.

RETREE

Reads in a tree (with branch lengths if necessary) and allows you **to reroot the tree**, to flip branches, to change species names and branch lengths, and then write the result out.
Can be used to convert between rooted and unrooted trees, and to write the tree into a preliminary version of a new XML tree file format which is under development and which is described in the RETREE documentation web page.

Inferencia filogenética usando el paquete

PHYLIP - MP

PROTPARS

Estimates phylogenies from **protein** sequences (input using the standard one-letter code for amino acids) using the **parsimony** method, in a variant which counts only those nucleotide changes that change the amino acid, on the assumption that silent changes are more easily accomplished.

DNAPARS

Estimates phylogenies by the **parsimony** method using **nucleic acid sequences**. Allows use the full IUB ambiguity codes, and estimates ancestral nucleotide states. Gaps treated as a fifth nucleotide state. Can use 0/1 weights, reconstruct ancestral states, and infer branch lengths.

DNAPENNY

Finds all **most parsimonious phylogenies for nucleic acid sequences by branch-and-bound search**. This may not be practical (depending on the data) for more than 15 species or so.

PHYLIP - ML

DNAML
Estimates phylogenies from nucleotide sequences by maximum likelihood. The model employed allows for unequal expected frequencies of the four nucleotides, for unequal rates of transitions and transversions, and for different (prespecified) rates of change in different categories of sites, with the program inferring which sites have which rates. It also allows different rates of change at known sites.

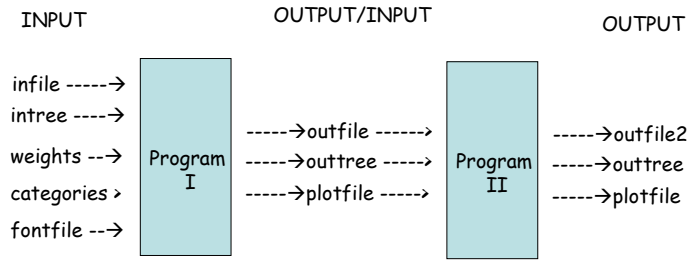
DNAMLK
Same as DNAML but assumes a molecular clock. The use of the two programs together permits a likelihood ratio test of the molecular clock hypothesis to be made.

PROML
Estimates phylogenies from protein amino acid sequences by maximum likelihood. The PAM or JTTF models can be employed. The program can allow for different (prespecified) rates of change in different categories of amino acid positions, with the program inferring which posiitons have which rates. It also allows different rates of change at known sites.

PROMLK
Same as PROML but assumes a molecular clock. The use of the two programs together permits a likelihood ratio test of the molecular clock hypothesis to be made.

Inferencia filogenética usando el paquete PHYLIP – manejo de archivos y programas

- Phylip tiene una estructura modular: consta de muchos programas, cada cual hace un(os) análisis particular(es) (similar a los programas de UNIX/LINUX)
- Para hacer unos análisis complejos hemos de construir una "pipeline" en la que la salida de un programa se usa como entrada del siguiente



Esquema de una "pipiline" de programas del paquete PHYLIP

Inferencia filogenética usando el paquete PHYLIP – manejo de archivos y programas

PHYLIP programs and documentation

PHYLIP, the PHYLogeny Inference Package, consists of 35 programs.
There are documentation files for each program, in the form of web pages in HTML 3.2. There are also documentation web pages for each group of programs, and a main documentation file that is the basic introduction to the package.
Before running any of the programs you should read it.
Below you will find a list of the programs and the documentation files.
The names of the documentation files are highlighted as links that will take you to those documentation files.

http://evolution.genetics.washington.edu/phylip/phylip.html

Esta es la liga a la documentación en formato HTML 3.2. Es una documetnación muy buena. Este URL lo encuentras también en las ligas de nuestra página del curso

Ejercicios – métodos de distancia

- Usando el archivo 5_atpD+recA_phy.phy (ya alineado y en formato phylip)

1. Vamos a calcular un árbol NJ con 1000 pseudoréplicas de bootstrap

Seqboot  seqboot.exe

DNAdist (F84 ó ML)  dnadist.exe

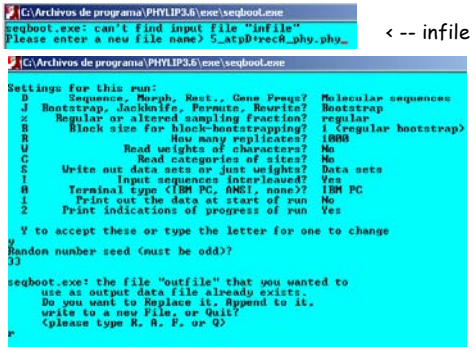
Neighbor  neighbor.exe

Consense  consense.exe

Drawgram  drawgram.exe

Los ejecutables de PHILIP: vista en formato windows

PHYLIP – Seqboot



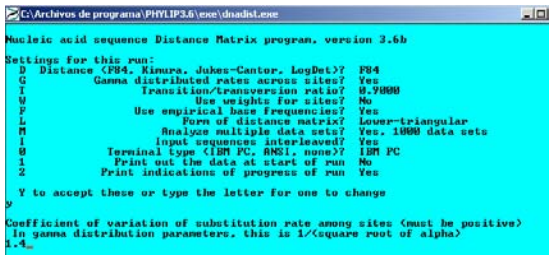
OPCIONES

R – 1000 (Con la opción R (réplicas) podemos modificar el valor a 1000)
y (con Y aceptamos las condiciones para el análisis y lo iniciamos)

outfile -- > contiene 1000 pseudoréplicas de bootstrap de nuestro alineamiento original; renombramos outfile como **1000boot**

PHYLIP – dnadist

dnadist.exe

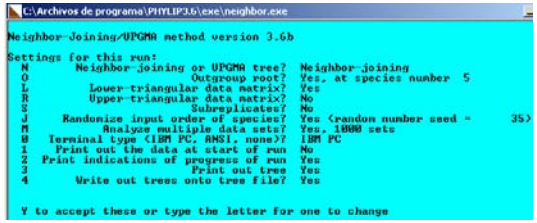


Opciones:
T – 0.9 (estimado por ML)
G – yes, CV = 1.4 (α = 0.39, estimado por ML)
L – lower-triangular matrix
M – mutlitple data sets (=1000)

renombramos outfile como 1000F84+G

PHYLIP – neighbor

neighbor.exe



Opciones:
O – redefinimos el taxon 5 como outgruop
L – lower-triangular matrix
J – randomizamos la entrada de taxa en el algoritmo
M – mutlitple data sets (=1000)
Y

renombramos outfile como 1000NJtrees

PHYLIP – consense

consense.exe

```
C:\Archivos de programa\PHYLIP3.6\exe\consense.exe
consense.exe: can't find input tree file "intree"
Please enter a new file name> 1000NJtrees
```

```
C:\Archivos de programa\PHYLIP3.6\exe\consense.exe

Consensus tree program, version 3.6b

Settings for this run:
 C         Consensus type (MRe, strict, MR, Ml):  Majority rule (extended)
 O                    Outgroup root:  Yes, at species number  5
 R               Trees to be treated as Rooted:  No
 I            Terminal type (IBM PC, ANSI, none):  IBM PC
 1                  Print out the sets of species:  Yes
 2          Print indications of progress of run:  Yes
 3                        Print out tree:  Yes
 4              Write out trees onto tree file:  Yes

Are these settings correct? (type Y or the letter for one to change)
y
```

Opciones:
O – redefinimos el taxon 5 como outgruop
y

renombramos outfile como 1000NJcns

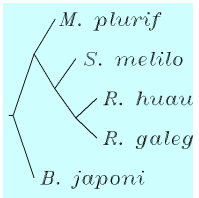
PHYLIP – drawgram

drawgram.exe

```
C:\Archivos de programa\PHYLIP3.6\exe\drawgram.exe
drawgram.exe: can't find input tree file "intree"
Please enter a new file name> 1000NJcns
DRAWGRAM from PHYLIP version 3.6b
Reading tree ...
Tree has been read.
Loading the font ...
drawgram.exe: can't find font file "fontfile"
Please enter a new file name> font1
```

```
C:\Archivos de programa\PHYLIP3.6\exe\drawgram.exe

Rooted tree plotting program version 3.6b

Here are the settings:
 0  Screen type (IBM PC, ANSI):  IBM PC
 P       Final plotting device:  Postscript printer
 U          Previewing device:  MS Windows display
 H                  Tree grows:  Horizontally
 S                  Tree style:  Curvogram
 B          Use branch lengths:  No
 L              Angle of labels:  90.0
 R      Scale of branch length:  Automatically rescaled
 D        Depth/Breadth of tree:  0.53
 T     Stem-length/tree-depth:  0.05
 C    Character ht / tip space:  0.3333
 A             Ancestral nodes:  Weighted
 F                        Font:  Times-Roman
 M          Horizontal margins:  1.65 cm
 M            Vertical margins:  2.16 cm
 #              Pages per tree:  one page per tree

Y to accept these or type the letter for one to change
y
```

Opciones:
O – taxon 5 como outgruop
S – curvogram
B – no
y



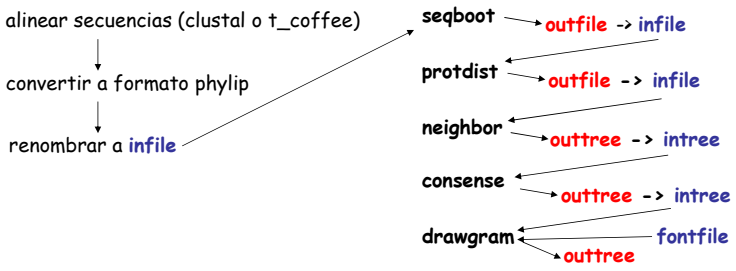
No hay outfile, sólo el display --- >

Desde luego que hay mejores editores de árboles, en particular **TreeView**

PHYLIP – una tubería de análisis desde la línea de comandos (UNIX)

- **Problema a resolver**:
  Vamos a hacer un análisis de NJ con 100 pseudoréplicas de bootstrap bajo modelo JTT de cada uno de los archivos de ejemplo de GDPs de eucariontes y procariones, así como del alineamiento de perfiles de ambos sets.

1. Debemos de pensar en la secuencia de análsis a realizar. Veamos el pseudocódigo:



PHYLIP – una tubería de análisis desde la línea de comandos (Perl)

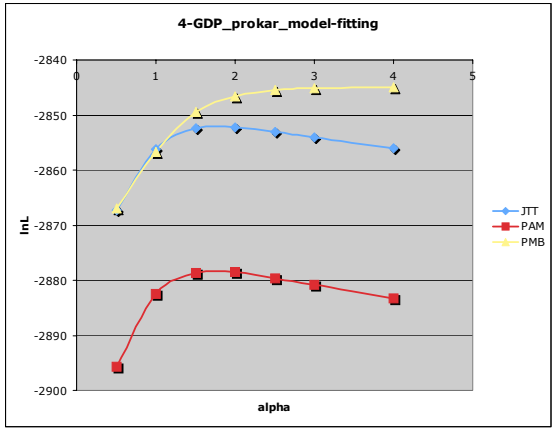
- Este tipo de análsis repetitivo es **fácilmente automatizable empleando** lenguajes de scripting como los que ofrecen los shell (bash, tcsh ...) u otros lenguajes más poderosos como son **Perl** o Python.

```
#!/usr/bin/perl -w

##################################################################
#  phylip_seqbootNeighbor_V01.pl written by P. Vinuesa 26-04-2006
#  vinuesa@ccg.unam.mx
#  Centro de Ciencias Genomicas-UNAM, Mexico
##################################################################
# This script will run seqboot neighbor and consense on all *phy placed in the current directory.
# The user will be asked if she/he wants to perform a bootstrap analysis with 100 pseudoreplicates.
#
# usage: perl phylip_seqbootNeighbor_V01.pl
##################################################################


print "# $0 runs seqboot, protdist, neighbor and consense
# on all multiple sequence files in phylip format (*.phy)
# found within the current working directory.\n\n";

use strict;

my($file, @parts, $answer);
```

CONTINÚA EN LA SIGUIENTE PÁGINA

```
while(defined($file=glob("*.phy"))){

        @parts=split(/\./,$file);

#       the last line splits the file name at the period and assigns the parts
#       to an array @parts ($parts[0], $parts[1] ...). The next line assigns only
#       the first part to $file and overwrites the earlier assignment.

        $file=$parts[0];
        print "\tDo you want to perform a bootstrap analysis with 100 pseudoreplicates for file: $file?
        Type y or n\n";
        chomp($answer=<STDIN>);
        if($answer =~ /y/i){
           system ("cp $file.phy infile");
           system ("seqboot < seqboot.cmd");
           system ("mv outfile infile");
           system ("protdist < protdist_plus100boot.cmd");
           system ("mv outfile infile");
           system ("neighbor < neighbor_100datasets.cmd");
           system ("rm outfile");
           system ("mv outtree intree");
           system ("consense < consense.cmd");
           system ("mv outtree $file.outtree");
           system ("mv outfile $file.outfile");
           system ("rm intree");
    }else{
           system ("cp $file.phy infile");
           system ("protdist < protdist_0boot.cmd");
           system ("mv outfile infile");
           system ("neighbor < neighbor_1datasets.cmd");
           system ("mv outtree $file.outtree");
           system ("mv outfile $file.outfile");
        }

};
# cleanup:
system ("rm infile");
exit;
```

PHYLIP – selección del modelo de sustitución y estima del parámetro alpha para secs. de prot. bajo ML



PHYLIP – una tubería de análisis desde la línea de comandos (UNIX)

- Para poder correr el script necesitamos escribir además los archivos de comandos (*.cmd) requeridos por cada programa phylip
- estos archivos son muy sencillos: contienen en cada renglón el comando que usarías al manejar phylip desde el menú gráfico. Sólo anota el orden exacto de la secuencia de comandos y sus valores en un archivo de texto plano, y ponle por ej. la extensión cmd para identificarlos.

seqboot.cmd

```
r
100
y
99
```

protdist_plus100boot.cmd

```
m
d
100
y
```

neighbor_100datasets.cmd

```
j
77
m
100
99
y
```

consense.cmd

```
y
```

PHYLIP – selección del modelo de sustitución y estima del parámetro alpha para secs. de prot. bajo ML

- Existen dos programas en el paquete PHYLIP para inferir filogenias de ML a partir de secuencias de proteína: **proml** y **promlk**
- Ambos programas implementan tres matrices empíricas de sustitución:

	JTT	PMB (Blossum-like)	PAM
comando:	defecto	P	P
			P

- Se pueden construír modelos +G, +G+I (por defecto no asume HTSES)

	+G	+G+I
comando:	R	R
		R

PHYLIP – selección del modelo de sustitución y estima del parámetro alpha para secs. de prot. bajo ML

- Si quieres obtener una estima de ML del valor del parámetro alpha (α) de la distribución gamma (Γ) bajo un modelo particular (p. ej. JTT), debes de probar con una serie de valores de CV, donde **CV = 1/ (α)$^{\frac{1}{4}}$**

- Por ejemplo, podemos evaluar los siguentes valores de CV (y α correspondientes):

CV	**1.41**	**1.00**	**0.82**	**0.71**	**0.63**	**0.58**	**0.50**
α	**0.50**	**1.00**	**1.50**	**2.00**	**2.50**	**3.00**	**4.00**

- Finalmente PROML (o PROMLK) les van a pedir el número de categorías con el que quieran aproximar (discretamente) la distribución gamma. Un valor de **4 categorías** es generalmente suficiente.

PHYLIP – selección del modelo de sustitución y estima del parámetro alpha para secs. de prot. bajo ML

- Este problema es nuevamente muy tedioso y tardado de hacer manualmente. Por ello vamos a usar el script **proml_modelfit_V01.pl** para ejecutarlo. El script toma todos los alineamientos *.phy de un directorio y calcluar los valores de verosimilitud global de las filogenias resultantes bajo cada uno de los tres modelos empíricos de sustitución que implementados en PROML/PROMLK bajo los siete valores de CV mostrados en la página anterior. Evaluamos **3 modelos X 7 valores de CV = 21 filogenias de ML** por alineam.

- El script **proml_modelfit_V01.pl** abre cada archivo de salida (outfile) de PROML para **parsearlo**. Es decir, capturamos de cada outfile los datos que nos interesan: (CV, alpha y -lnL). El programa **imprime en pantalla los resultados del parseo y además los escribe en archivos**. Ello va a facilitar poder hacer un **análisis gráfico de la función de verosimilitud dados alpha y -lnL**

PHYLIP – selección del modelo de sustitución y estima del parámetro alpha para secs. de prot. bajo ML

#1CV	alpha	-lnL	for JTT
1.414000	0.500151	-2867.31600	
1.000000	1.000000	-2856.13225	
0.816000	1.501826	-2852.46224	
0.707000	2.000604	-2852.25738	
0.632000	2.503605	-2853.03578	
0.577000	3.003643	-2854.04273	
0.500000	4.000000	-2855.96898	
#2CV	alpha	-lnL	for PMB
1.414000	0.500151	-2867.00406	
1.000000	1.000000	-2856.74978	
0.816000	1.501826	-2849.46126	
0.707000	2.000604	-2846.65591	
0.632000	2.503605	-2845.50671	
0.577000	3.003643	-2845.05301	
0.500000	4.000000	-2844.91106	
#3CV	alpha	-lnL	for PAM
1.414000	0.500151	-2895.76260	
1.000000	1.000000	-2882.49693	
0.816000	1.501826	-2878.60925	
0.707000	2.000604	-2878.58367	
0.632000	2.503605	-2879.62446	
0.577000	3.003643	-2880.89181	
0.500000	4.000000	-2883.26476	