Метод максимального правдоподобия

Прикладная генетика для зоологов, лекция 7

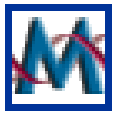
Мюге Н.С.

Принципы кладистики: Хенниг (1966) «Филогенетическая систематика»

- 1. Кладограммы (филогенетические схемы) строятся по дихотомическому принципу.
- 2. Таксоны выделяются только по вертикальному принципу.
- 3. Ранг таксонов определяется последовательностью их ответвления на кладограмме, понижаясь от основания кладограммы к вершине; таким образом, степень родства таксонов соответствует времени их разделения.
- 4. Все признаки, характеризующие таксон, подразделяются на плезиоморфные (унаследованные, примитивные) и апоморфные (производные, прогрессивные).
- 5. Таксоны выделяются только по апоморфным признакам.
- 6. Критерием родства является синапоморфия; соответственно последовательность обособления различных таксонов на кладограмме определяется путем сопоставления их апоморфных признаков.
- 7. Пары таксонов, исходящие на кладограмме из одной точки, образуют «сестринские группы», связанные друг с другом максимальным родством и характеризующиеся наиболее полной синапоморфией.
- 8. Из пары сестринских групп одна обычно сохраняет значительно большее сходство с предковым таксоном, чем другая (правило девиации); обоим сестринским таксонам придается тем не менее одинаковый ранг.
- 9. Предковый таксон, давая начало двум сестринским, исчезает, что определяется требованиями дихотомического принципа построения кладограмм.

Три основных метода реконструкции филогении:

- Парсимония (Parsimony) (PAUP, MEGA, Phylip)
- Максимального правдоподобия (maximum likelihood) - (PAUP, Phylip)
- Обратной вероятности, байезиан (bayesian) – (MrBayes)



Правдоподобие (Likelihood)

- В модели, использующейся для анализа нуклеотидных последовательностей методом правдоподобия, определяется вероятность перехода за определенное время от одной последовательности к другой в результате мутаций.

Программы для анализа парсимонии:

Phylip	http://evolution.genetics.washington.edu/phylip.html
PAUP	http://paup.csit.fsu.edu/
PhyML	http://atgc.lirmm.fr/phyml/

Работа с программой PAUP (Phylogenetic Analysis Using Parsimony)

- Создать .nex файл (save as.. In MegAlign)
- Аутгруппа должна быть первой (или задать как outgroup=7,8
- Убрать все тире в названиях
- Выбрать критерий анализа (set criterion = likelihood, set criterion = parsimony)
- Bootstrap nreps=1000
- Savetree from=1 to=1 treefile=NNNN.tre

- **JC69 model (Jukes and Cantor, 1969)**
- JC69 is the simplest substitution model. There are several assumptions. It assumes equal base frequencies (

$$\pi_T = \pi_C = \pi_A = \pi_G = \frac{1}{4}$$

-) and equal mutation rates. The only parameter of this model is therefore μ, the overall substitution rate.

- **K80 model (Kimura, 1980)**
- The K80 model distinguishes between transitions (A <-> G, i.e. from purine to purine, or C <-> T, i.e. from pyrimidine to pyrimidine) and transversions (from purine to pyrimidine or vice versa) (α/β).
- It also assumes equal base frequencies

$$\pi_T = \pi_C = \pi_A = \pi_G = \frac{1}{4}$$

F81 model (Felsenstein 1981)

Unequal base frequencies ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$)

Rate matrix $Q = \begin{pmatrix} * & \pi_T & \pi_T & \pi_T \\ \pi_C & * & \pi_C & \pi_C \\ \pi_A & \pi_A & * & \pi_A \\ \pi_G & \pi_G & \pi_G & * \end{pmatrix}$

HKY85 model (Hasegawa, Kishino and Yano 1985)

The HKY85 model distinguishes between transitions and transversions (α/β).

It allows unequal base frequencies ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$).

Rate matrix $Q = \begin{pmatrix} * & \kappa\pi_T & \pi_T & \pi_T \\ \kappa\pi_C & * & \pi_C & \pi_C \\ \pi_A & \pi_A & * & \kappa\pi_A \\ \pi_G & \pi_G & \kappa\pi_G & * \end{pmatrix}$

T92 model (Tamura 1992)

One frequency only π_{GC}

$$\pi_G = \pi_C = \frac{\pi_{GC}}{2}$$

$$\pi_A = \pi_T = \frac{(1-\pi_{GC})}{2}$$

Rate matrix $Q = \begin{pmatrix} * & \kappa(1-\pi_{GC})/2 & (1-\pi_{GC})/2 & (1-\pi_{GC})/2 \\ \kappa\pi_{GC}/2 & * & \pi_{GC}/2 & \pi_{GC}/2 \\ (1-\pi_{GC})/2 & (1-\pi_{GC})/2 & * & \kappa(1-\pi_{GC})/2 \\ \pi_{GC}/2 & \pi_{GC}/2 & \kappa\pi_{GC}/2 & * \end{pmatrix}$

The evolutionary distance between two noncoding sequences according to this model is given by

$$d = -h\ln(1-\frac{p}{h}-Q) - \frac{1}{2}(1-h)\ln(1-2Q)$$

where $h = 2\theta(1-\theta)$ where $\theta \in (0,1)$ is the GC content.

TN93 model (Tamura and Nei 1993)

The TN93 model distinguishes between the two different types of transition - i.e. (A <-> G) is allowed to have a different rate to (C<->T). Transversions are all assumed to occur at the same rate, but that rate is allowed to be different from both of the rates for transitions

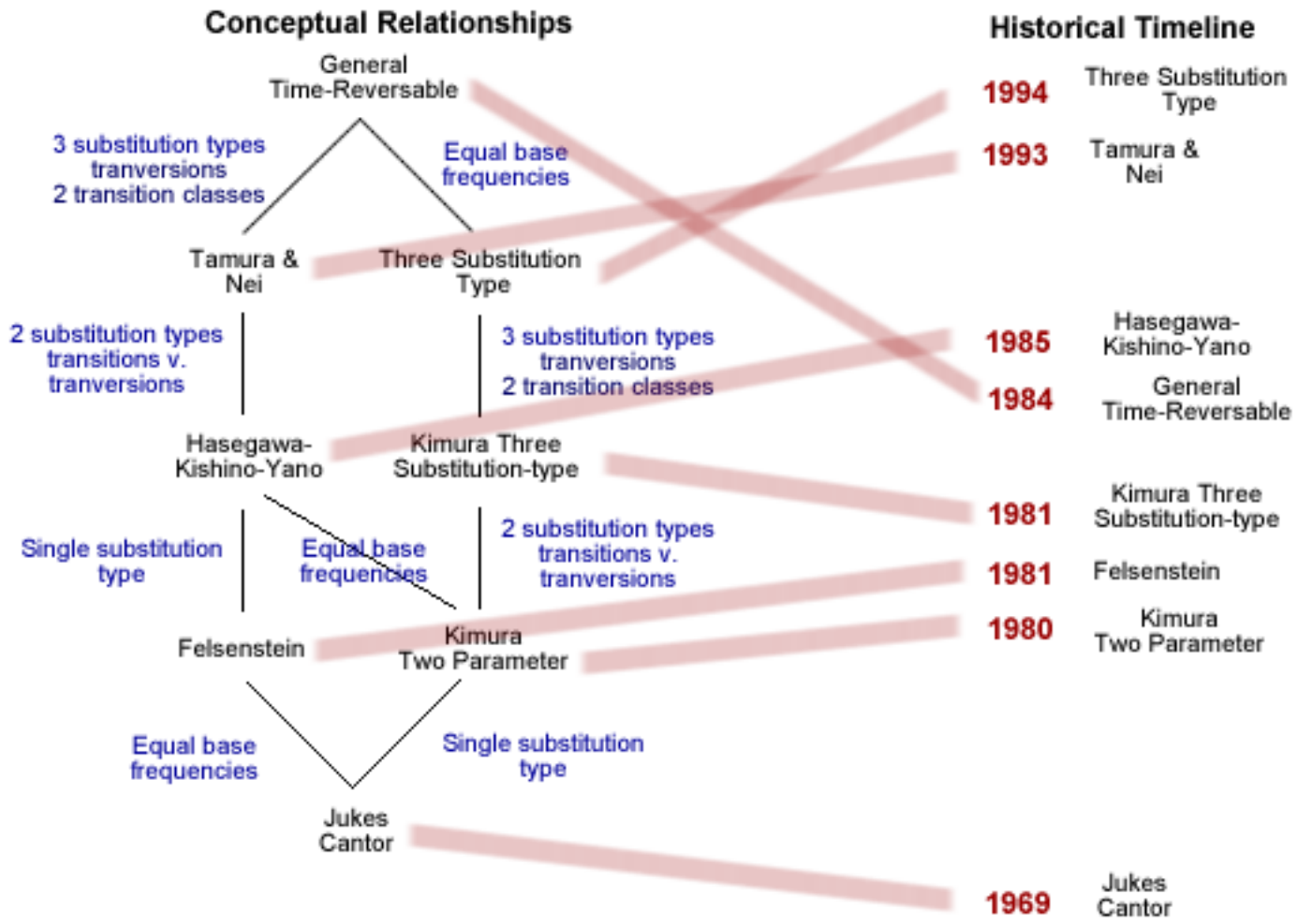
TN93 also allows unequal base frequencies ($\pi_T \neq \pi_C \neq \pi_A \neq \pi_G$).

Rate matrix $Q = \begin{pmatrix} * & \kappa_1\pi_T & \pi_T & \pi_T \\ \kappa_1\pi_C & * & \pi_C & \pi_C \\ \pi_A & \pi_A & * & \kappa_2\pi_A \\ \pi_G & \pi_G & \kappa_2\pi_G & * \end{pmatrix}$

GTR: Generalised time reversible

$$Q = \begin{pmatrix} -\left(\frac{\pi_1 x_1}{\pi_2} + \frac{\pi_1 x_2}{\pi_3} + \frac{\pi_1 x_3}{\pi_4}\right) & \frac{\pi_1 x_1}{\pi_2} & \frac{\pi_1 x_2}{\pi_3} & \frac{\pi_1 x_3}{\pi_4} \\ x_1 & -\left(x_1 + \frac{\pi_2 x_4}{\pi_3} + \frac{\pi_2 x_5}{\pi_4}\right) & \frac{\pi_2 x_4}{\pi_3} & \frac{\pi_2 x_5}{\pi_4} \\ x_2 & x_4 & -\left(x_2 + x_4 + \frac{\pi_3 x_6}{\pi_4}\right) & \frac{\pi_3 x_6}{\pi_4} \\ x_3 & x_5 & x_6 & -\left(x_3 + x_5 + x_6\right) \end{pmatrix}$$

Эволюция моделей эволюции ДНК



Для анализа ML – выбрать и задать модель

MODELTEST: A tool to select the best-fit model of nucleotide substitution

© 1998-2006 David Posada

Current version is 3.7.

MODELTEST is program for the selection the model of nucleotide substitution that best fits the data. The program chooses among 56 models, and implements three different model selection frameworks: hierarchical likelihood ratio tests (hLRTs), Akaike information criterion (AIC), and Bayesian information criterion (BIC). The program also implements the assesment of model uncertainty and tools for model averaging and calculation of parameter importance, using the AIC or the BIC.

Operative systems
Executables are provided for macintosh and windows. Source code and a makefile are provided for compilation in any OS with a C compiler.

Links
These are some useful links related to Modeltest:

- MTgui: a windows and linux interface for modeltest. By Paulo Nuin.
- MrModeltest : a version of Modeltest modified for its use with MrBayes. By Johan Nylander.
- Instructions for running Modeltest on Windows. By Bevan Weir
- FindModel : web server to choose among 28 nucleotide models with the AIC at Los Alamos National Laboratory

программа

Пошаговые инструкции

Modeltest online

Citation
Posada D and Crandall KA 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14 (9): 817-818.

See also: Posada D and Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. Systematic Biology 53: 793-80

Modeltest registration form

http://darwin.uvigo.es/software/modeltest.html

Modeltest tutorial

Modeltest guide
MrBayes guide
Using DOS
Downloads

Related Links

Modeltest website
Maximum likelihood
PAUP*
ClustalX

The step by step guide to Modeltest

This guide focuses on using Modeltest to find the best parameters to construct a maximum likelihood phylogenetic tree. This is not a substitute for a bioinformatics education, make sure you know what you are doing. [Updated 23rd July 2006]

MODELTEST is a computer program by David Posada and Keith A Crondall. One use of this program is to calculate the best model for DNA evolution using Maximum Likelihood (ML). For more information read their publication (pdf) or visit the Modeltest website. When I started to use this program I found there was little detailed information about to help newbies use it, so I wrote this tutorial to remind me of the steps, and to help some graduate students in my lab.

The latest version of Modeltest is 3.7 released on July 29, 2005

To use the Modeltest program you need the application itself (available from the authors website), and the phlyogenetics package **PAUP*** (this is not free unfortunately, hopefully your university or company has a license). I will be referring to the windows command line version in this document, but these files will work on all versions. Finally you will need a fast computer, I use a 2.4 GHz Pentium 4 laptop, on this it takes about an hour to work through 50 taxa of 1300 bp. You also need two files: modelblockPAUPb10.txt and ML-search.txt which you can download from my website. modelblockPAUPb10.txt was written by David Posada and is a copy of the file found in the paupblock folder that is included in the download. ML-search was written by me, modify this for you own use.

I also assume that at this point you have a sequence alignment in the NEXUS format (this is critical). The easiest way to do this is use the "save as" command in ClustalX Open up this file in a text editor and replace the "-" (gap) symbol with the missing data symbol "?" where appropriate. It is getting out of the scope of this document to explain this, but typically this is at the ends of the alignment when you don't have full sequence for all taxa. One handy hint don't use the "-" symbol in taxa names, use "_" instead.

I use my own file-naming scheme to keep track of things. I suggest replacing the word "test" below with the name of your gene, i.e.: "p53.ML.search.nex"

Finally I highly recommend you read the documentation file: modeltest3.6.pdf found in the doc folder of the download

Step by step MODELTEST for the windows version

1. Start with a nexus file of your alignment in this document it will be called "test.aln.nex"
2. Add the PAUP*-block to the end of this file. The sample block is called "modelblockPAUPb10.txt" This has changed from previous versions, by adding the command "default lscores longfmt=yes;" this is due to a bug in PAUP* 4.10b. If you use the PAUP-block that comes with earlier versions of Modeltest it won't work!
3. Save this file as "test.model.nex" (keeping the original). Drag this new file on to the paupstar.exe executable. Or if you have a version with a graphical user interface (GUI), execute the file using the menu system.
4. All going well a DOS-like window will open up and PAUP* will begin to test your data against 56 different models of DNA evolution.

http://www.rhizobia.co.nz/phylogenetics/modeltest.html

- Testing models of evolution - Modeltest 3.7
- Confidence level = 0.01

- Equal base frequencies
- Null model = JC -lnL0 = 4917.6050
- Alternative model = F81 -lnL1 = 4859.0439
- 2(lnL1-lnL0) = 117.1221 df = 3
- P-value = <0.000001
- Ti=Tv
- Null model = F81 -lnL0 = 4859.0439
- Alternative model = HKY -lnL1 = 4747.9785
- 2(lnL1-lnL0) = 222.1309 df = 1
- P-value = <0.000001
- Equal Ti rates
-

- Null model = HKY -lnL0 = 4747.9785
- Alternative model = TrN -lnL1 = 4741.0918
- 2(lnL1-lnL0) = 13.7734 df = 1
- P-value = 0.000206
- Equal Tv rates
- Null model = TrN -lnL0 = 4741.0918
- Alternative model = TIM -lnL1 = 4738.0659
- 2(lnL1-lnL0) = 6.0518 df = 1
- P-value = 0.013892
- Equal rates among sites
- Null model = TrN -lnL0 = 4741.0918
- Alternative model = TrN+G -lnL1 = 4363.6865
- 2(lnL1-lnL0) = 754.8105 df = 1
- Using mixed chi-square distribution
- P-value = <0.000001
- No Invariable sites

- Null model = TrN+G -lnL0 = 4363.6865
- Alternative model = TrN+I+G -lnL1 = 4359.4595
- 2(lnL1-lnL0) = 8.4541 df = 1
- Using mixed chi-square distribution
- P-value = 0.001821

- Model selected: TrN+I+G
- -lnL = 4359.4595
- K = 7
- Base frequencies:
- freqA = 0.2899
- freqC = 0.2059
- freqG = 0.1375
- freqT = 0.3667
- Substitution model:
- Rate matrix
- R(a) [A-C] = 1.0000
- R(b) [A-G] = 3.7327
- R(c) [A-T] = 1.0000
- R(d) [C-G] = 1.0000
- R(e) [C-T] = 4.2004
- R(f) [G-T] = 1.0000
- Among-site rate variation
- Proportion of invariable sites (I) = 0.4317
- Variable sites (G)
- Gamma distribution shape parameter = 1.2196

- PAUP* Commands Block: If you want to implement the previous estimates as likelihod settings in PAUP*, attach the next block of commands after the data in your PAUP file:

- [!
- Likelihood settings from best-fit model (TrN+I+G) selected by hLRT in Modeltest 3.7 on Tue Sep 18 02:41:29 2007
-]

- BEGIN PAUP;
- Lset Base=(0.2899 0.2059 0.1375) Nst=6 Rmat=(1.0000 3.7327 1.0000 1.0000 4.2004) Rates=gamma Shape=1.2196 Pinvar=0.4317;
- END;

- --

HCV sequence database

Housekeeping | Retrieving data | Analysis | Tools | Links | Search

Please read an important announcement about the future of the HCV database here.

FindModel

Purpose: Findmodel analyzes your alignment to see which phylogenetic model best describes your data; this model can then be used to generate a better tree.

Explanation: Findmodel was developed from a web implementation of the Modeltest script written by David Posada and Keith

NOTE: There is a downloadable interface for the original Modeltest code. It is available from Genedrift.org. Thanks to Stuart Ray for this information.

Input

Paste your input here [SAMPLE INPUT]

or upload your file — Обзор...

Options

use all the 28 models ☐

construct the initial tree using:
- (•) Weighbor
- ( ) PAUP*
- ( ) MrBayes

Submit | Reset

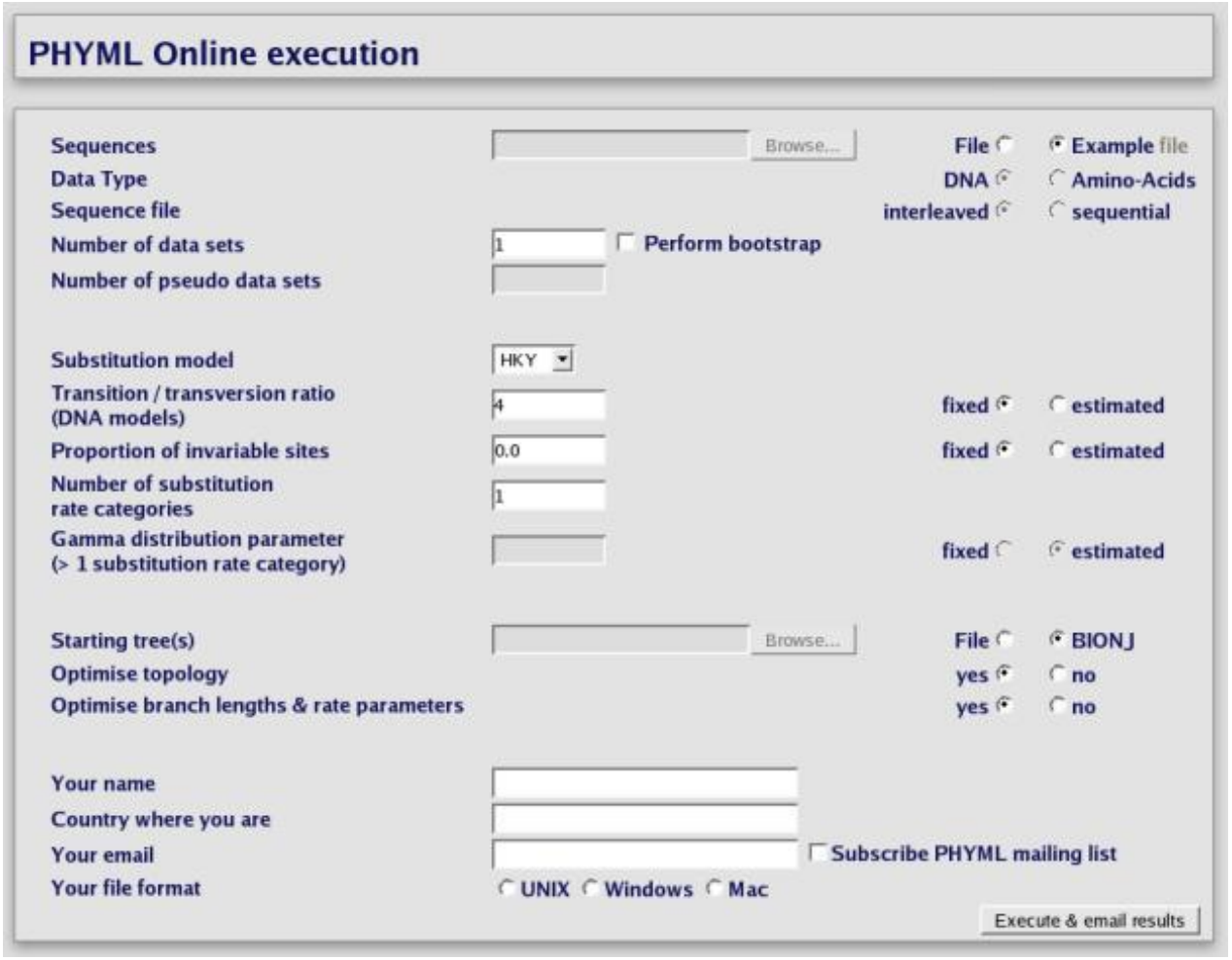
This program is computationally intensive and may take a while to run; please don't resubmit your request!

http://hcv.lanl.gov/content/hcv-db/findmodel/findmodel.html

Работа с программой PAUP (Phylogenetic Analysis Using Parsimony)

- Создать .nex файл (save as.. In MegAlign)
- Аутгруппа должна быть первой (или задать как outgroup=7,8
- Убрать все тире в названиях
- Выбрать критерий анализа (set criterion = likelihood, set criterion = parsimony)
- Bootstrap nreps=1000
- Savetree from=1 to=1 treefile=NNNN.tre

PhyML в интернете



Руководство пользователя : http://atgc.lirmm.fr/phyml/usersguide.html

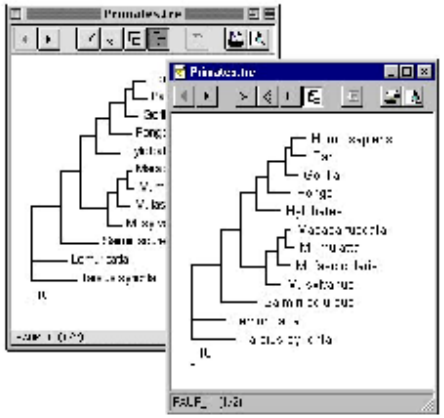
Программа для работы с деревьями- TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html)

Taxonomy and Systematics at Glasgow

TreeView

Tree drawing software for Apple Macintosh and Windows

(and now Linux and Unix)



TreeView is a simple program for displaying phylogenies on Apple Macintosh and Windows PCs. It has the following features:

Phylogeny Programs

Changes | Waiting list | Other lists | Old programs | Not listed | ???

Here are 362 phylogeny packages and 50 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.

Email addresses in these pages have had the @ symbol replaced by (at) and also surrounded by invisible confusing tags and blank characters in hopes of foiling spambots that harvest email addresses.

If you discover any inaccuracies, or feel that I have left any important programs or facts out, or if links do not work properly, please e-mail me at: (joe (at) gs.washington.edu). You can also use the submission form here to submit new entries.

Owing to past NSF support of these pages, I am required to note that any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation (NSF supported these pages from 1995-2003).

List of packages arranged ...

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