PIST

Program for the Informative-Sites Test

Version 1.0

(c) Copyright, 2001 Andrew Rambaut and Michael Worobey

Department of Zoology, University of Oxford

South Parks Road, Oxford OX1 3PS, U.K.

e-mail: michael.worobey@zoo.ox.ac.uk

Introduction

PIST is an application for running the informative-sites test, a method for detecting recombination from gene sequence data sets. What follows here is a brief account of how to use the software provided. The full details of the procedure can be found in the following paper:

Worobey, M. (2001) A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. *Mol. Biol. Evol.* **18,** 1425-1434.

PIST is a Power Macintosh executable that takes as input a sequence alignment and estimated maximum likelihood phylogenetic tree, as well as nucleotide substitution model parameters specified in a command line. It simulates replicate data sets along the specified tree, under the specified model of sequence evolution. These clonally evolved data sets can then be used to test whether the real data set, used to estimate the tree and model parameters, shows evidence of non-clonal evolution. Sequence alignments that carry a history of recombination will tend to be rich in conflicting phylogenetic information compared with clonally generated data sets supporting the same tree shape and model parameters, and will tend to have a larger value of *q*—defined as the proportion of two-state parsimony informative sites to all polymorphic sites. PIST outputs *q* values (as well as the tree-length, and a breakdown of various sorts of polymorphic sites) for each replicate simulated data set.

The method is intended for use with unambiguously-aligned, gap-stripped, third codon position sites from non-overlapping reading frames of protein coding nucleotide sequences.

Usage

To run PIST, first you need an alignment, as well as a phylogenetic tree and parameter estimates derived from that alignment. I will outline the overall procedure using as an example the dengue virus data set that is reported in the paper.

1. Get the alignment. Here, I've shown the NEXUS format alignment of 774 third positions from the 7 dengue virus sequences:

d1cme.3rd.nex

```
#NEXUS

[Created by Se-Al]
BEGIN DATA;
    DIMENSIONS NTAX=7 NCHAR=774;
    FORMAT MISSING=? GAP=- DATATYPE=DNA;
    MATRIX
D1Singa90       GCCAAAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTCAACAATACACAAAATGTAGCACGGT
D1Thai80        GCCAGAGGTCGTCTGGACGACCGATTAGGGAACAAAGTACAACGAGGGTCAATAATACACAAAATGAAGCACGGT
D1Phili84       GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGAGTTAACAATACATAAAATGTAGCACGGT
D1Jamai77       GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTCAATAATACACAAAATGTAGCACGGT
D1Brazi90       GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTCAATAATACACAAAATGTAGCACGGT
D1Nauru74       GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTTAACAATACATAAAATGTAGCACGGT
D1FGuiana89     GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTTAATAATACACAAAATGTAGCACGGT
    ;
END;
```

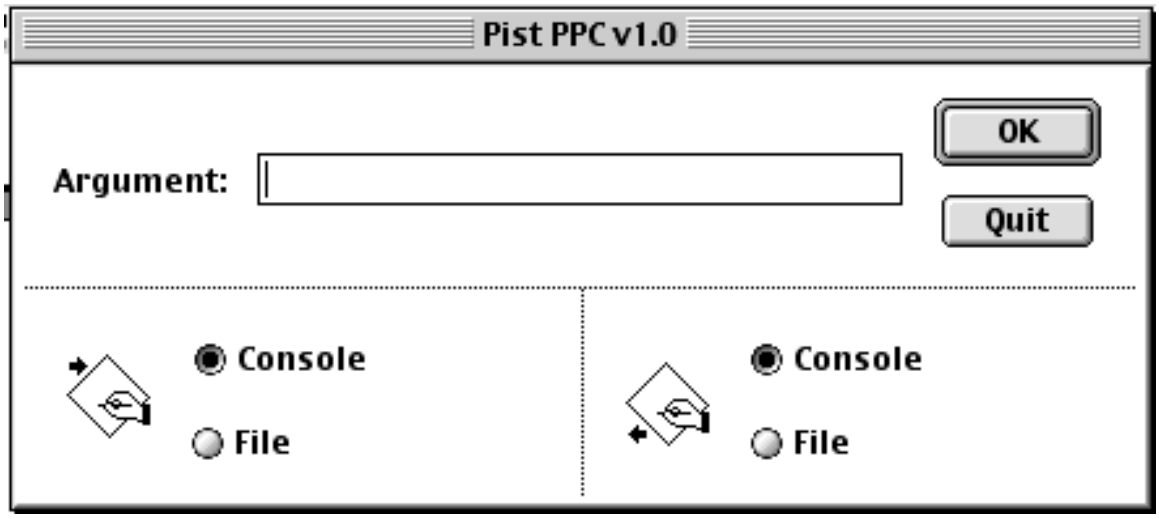
2. Estimate the maximum likelihood tree and substitution model parameters using a program such as PAUP*, to produce an input file for PIST:

d1cme.3rd.phy

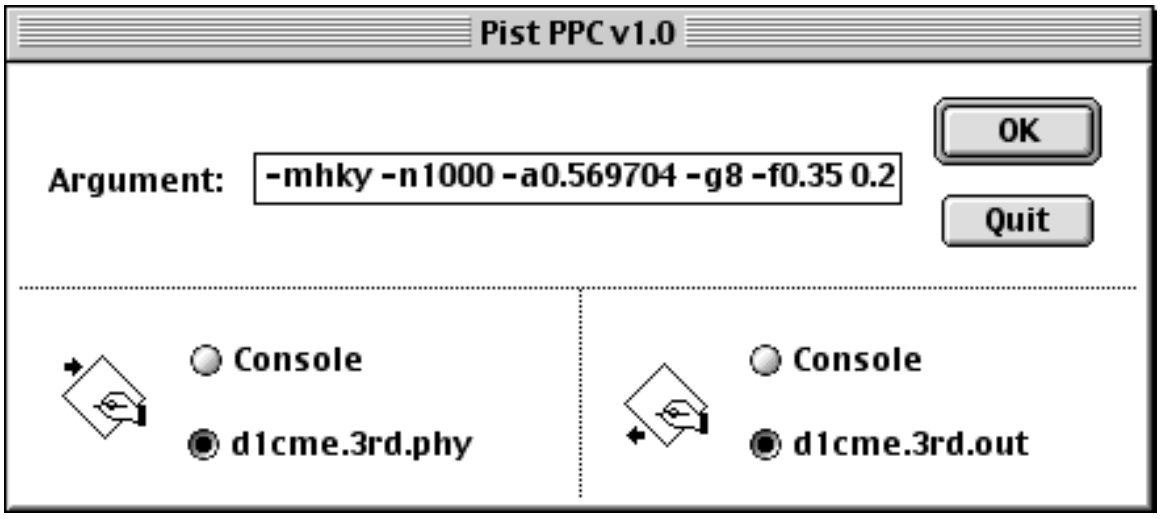
```
 7 774
D1Singa90       GCCAAAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTCAACAATACACAAAATGTAGCACGGTAGCAGAGTCGAA
D1Thai80        GCCAGAGGTCGTCTGGACGACCGATTAGGGAACAAAGTACAACGAGGGTCAATAATACACAAAATGAAGCACGGTAGCCAAGTCAAG
D1Phili84       GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGAGTTAACAATACATAAAATGTAGCACGGTAGCAGAGTCAAA
D1Jamai77       GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTCAATAATACACAAAATGTAGCACGGTAGCAGAGTCAAG
D1Brazi90       GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTCAATAATACACAAAATGTAGCACGGTAACAGAGTCAAG
D1Nauru74       GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTTAACAATACATAAAATGTAGCACGGTAGCAGAGTCGAA
D1FGuiana89     GCCAGAGGTAGTCTGGACGACCGATTAGGGGACAAAGTACAACGAGGGTTAATAATACACAAAATGTAGCACGGTAACAGAGTCAAG
1
(D1Singa90:0.099053,((D1Thai80:0.014686,(D1Phili84:0.006234,D1Nauru74:0.030720):0.054964):0.017499,D1
```

This input file should be in PHYLIP format, and it should include the estimated ML tree with branch lengths. I often use an HKY + gamma substitution model, but more general models (such as REV + gamma + invariant sites) can also be used, as they are supported by the PIST software.

3. Run PIST. A dialogue box will appear that looks like this:



Click on the "File" radio button (bottom left) and choose the PHYLIP format input file you have prepared. Then click on the other "File" button (bottom right) to give a name to your output file. Finally, specify the parameters to control PIST in the command line:

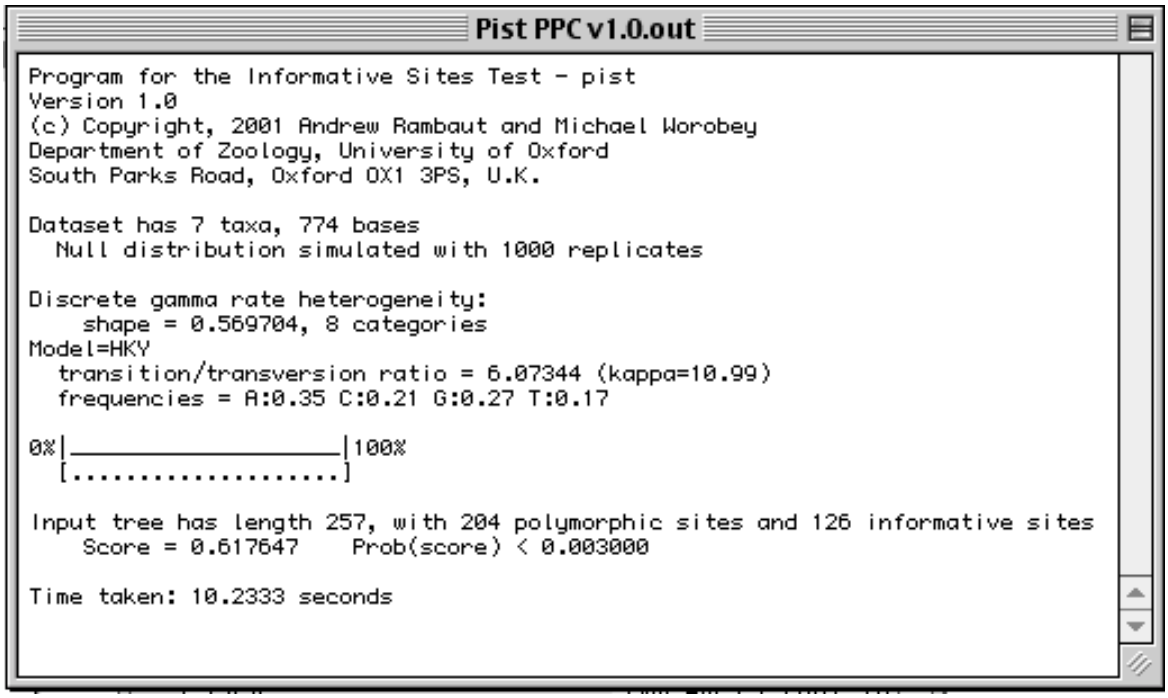


The full command line is not visible, but it looks like this:

-mhky -n1000 -a0.569704 -g8 -f0.35 0.21 0.27 0.17 -t6.07344 -p100

See **Appendix 1 – Parameters to control PIST,** below, for a description of the parameters, or type "-h" in the command line to get a list of the parameters.

After pressing “OK”, and allowing the program to run, a box will appear:



This shows that the observed value of *q* (0.618) was greater than all but two of the 1000 clonal replicates, indicating a history of recombination.

Finally, to view the results for each replicate data set, open the output file (d1cme.3rd.out) in, say, Excel:

	A	B	C	D	E
1	SimNo	TreeLength	PolymorphicS	InformSites	PropInfSites
2	1	238	192	97	0.505208
3	2	220	180	107	0.594444
4	3	254	206	112	0.543689
5	4	259	211	104	0.492891
6	5	270	214	111	0.518692
7	6	270	215	119	0.553488
8	7	271	209	96	0.45933
9	8	260	208	120	0.576923
10	9	245	200	97	0.485
11	10	251	191	95	0.497382
12	11	255	209	91	0.435407
13	12	269	211	107	0.507109
14	13	266	207	114	0.550725

(Just the first 14 are shown here.)

Appendix 1 - Parameters to control PIST

Options and parameters to PIST are supplied on the command line. The general format is a minus sign followed by a code letter. If required, the values of any parameters come after the code, separated from both code and each other with either a comma or a space. Some options act like switches and require no parameters. The case of the options is ignored. Most of the code for PIST is borrowed from Seq-Gen (Andrew Rambaut, http://evolve.zoo.ox.ac.uk/software/) and the reader is referred to the Seq-Gen manual for further details of the Monte Carlo simulation of nucleotide sequences.

Model [default = HKY]
This option sets the model of nucleotide substitution with a choice of either *F84*, *HKY* (also known as *HKY85*) or *REV* (markov general reversable model). The first two models are quite similar but not identical. They both require a transition transversion ratio and relative base frequencies as parameters. Other models such as *K2P*, *F81* and *JC69* are special cases of *HKY* and can be obtained by setting the nucleotide frequencies equal (for *K2P*) or the transition transversion ratio to 1.0 (for *F81*) or both (for *JC69*). The usage is:

-m <MODEL>

Where <MODEL> is a three letter code: HKY ,F84 or REV.

Number of Acceptable Simulated Replicate Datasets [default = 1000]
This option specifies how many simulated acceptable datasets should be simulated.

-n <NUMBER_OF_DATASETS>

Where <NUMBER_OF_DATASETS> is an integer number that corresponds to the number of datasets desired.

Tree Length Acceptance Interval (%) [default = 0.5]

-p <ACCEPTANCE_INTERVAL>

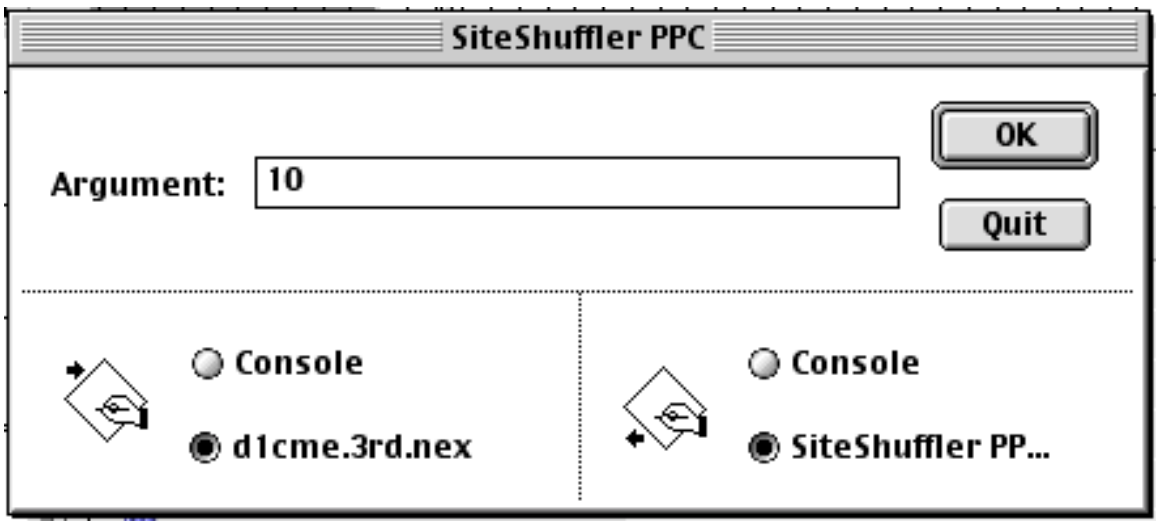
Where <ACCEPTANCE_INTERVAL> is an integer number from 0 to 100 that corresponds to % deviation in the tree length of each simulated data set set compared with the input data set. The default setting is 0.5, so only data sets that have a tree length very similar to the input tree are kept. If <ACCEPTANCE_INTERVAL> is set to 100, then all data sets are kept.

Average Tree Length for Randomised Data [default = none]
-l <RANDOMIZED_TREE_LENGTH>

This option allows the user to input the randomized tree length value ($\hat{t}_r$) used to calculate the ISI (informative-sites index; refer to the paper for more information) which is expected to be about 0 for clonal data and 1 for data at linkage equilibrium. I use the program SITESHUFFLER to get this value. It takes a nexus format alignment as input, then randomly permutes the characters at each position to remove linkage between sites. I produce ten such data sets, then perform a heuristic parsimony search on each using the default settings in PAUP*, and take the average to get $\hat{t}_r$. The ISI is then calculated as follows (see paper):

$$\text{ISI} = \frac{(q - \hat{q}_c)/t}{(1 - \hat{q}_c)/t \quad - \quad (1 - q)/\hat{t}_r}$$

(SITESHUFFLER is included as a PowerMac executable with PIST:



Here, I have specified 10 randomized replicates, starting with the original alignment, d1cme.3rd.nex. The output can then be analyzed in a program such as PAUP*. Typically, the average tree length after randomization is somewhat greater than before.)

Codon-Specific Rate Heterogeneity [default = none]
Using this option the user may specify the relative rates for each codon position. This allows codon-specific rate heterogeneity to be simulated. The default is no site-specific rate heterogeneity.

-c <CODON_POSITION_RATES>

Where <CODON_POSITION_RATES> is three decimal numbers that specify the relative rates of substitution at each codon position, separated by commas or spaces. This option is not used for the informative-sites test.

Gamma Rate Heterogeneity [default = none]
Using this option the user may specify a shape for the gamma rate heterogeneity called alpha.

-a <ALPHA>

Where <ALPHA> is a real number >0 that specifies the shape of the gamma distribution to use with gamma rate heterogeneity. If this is used with the **-g** option, below, then a discrete model is used, otherwise a continuous one. This is a crucial parameter since, in effect, the informative-sites test attempts to distinguish between "real" rate heterogeneity, and artifactual rate heterogeneity.

Discrete Gamma Rate Heterogeneity [default = continuous]
Using this option the user may specify the number of categories for the discrete gamma rate heterogeneity model. The default is no site-specific rate heterogeneity (or the continuous model if only the **-a** option is specified).

-g <NUM_CATEGORIES>

Where <NUM_CATEGORIES> is an integer number between 2 and 32 that specifies the number of categories to use with the discrete gamma rate heterogeneity model.

Proportion of Invariable Sites [default = 0.0]
Using this option the user may specify the proportion of sites that should be invariable. These sites will be chosen randomly with this expected frequency. The default is no invariable sites. Invariable sites are sites that cannot change as opposed to sites which don't exhibit any changes due to chance (and perhaps a low rate).

-i <PROPORTION_INVARIABLE>

Where <PROPORTION_INVARIABLE> is an real number >= 0.0 and <1.0 that specifies the proportion of invariable sites.

Transition Transversion Ratio default = 2.0
This option allows the user to set a value for the transition transversion ratio (TS/TV). This is only valid when either the HKY or F84 model has been selected.

-t <TRANSITION_TRANSVERSION_RATIO>

Where <TRANSITION_TRANSVERSION_RATIO> is a decimal number greater than zero.

General Reversible Rate Matrix [default = all 1.0]
This option allows the user to set 6 values for the general reversible model's rate matrix. This is only valid when either the REV model has been selected.

-r <RATE_MATRIX_VALUES>

Where <RATE_MATRIX_VALUES> are size decimal numbers for the rates of transition from A to C, A to G, A to T, C to G, C to T and G to T respectively, separated by spaces or commas. The matrix is symmetrical so the reverse transitions equal the ones set (e.g. C to A equals A to C) and therefore only six values need be set. These values will be scaled such that the last value (G to T) is 1.0 and the others are set relative to this.

Relative Nucleotide Frequencies [default = all equal]
This option is used to specify the relative frequencies of the four nucleotides. By default, PIST will assume these to be equal. If the given values don't sum to 1.0 then they will be scaled so that they do.

-f <NUCLEOTIDE_FREQUENCIES>

Where <NUCLEOTIDE_FREQUENCIES> are four decimal numbers for the frequencies of A, C, G and T respectively, separated by spaces or commas.

Write Sequences to File **'pist.seq'**
This option allows the user to obtain the simulated sequences for each replicate.

-w

Output File Format [default = PHYLIP]
This option selects the format of the output file. The default is PHYLIP format.

-op

PHYLIP format.

-or

Relaxed PHYLIP format: PHYLIP format expects exactly 10 characters for the name (padded with spaces if the name is actually less than 10). With this option the output file can have up to 64 characters in the name, followed by a single space before the sequence. The longer taxon names are read from the tree. Some programs can read this and it keeps long taxon names.

-on

NEXUS format: This creates a NEXUS file which will load into PAUP. It generates one DATA block per dataset. It also includes the simulation settings as comments which will be ignored by PAUP.

Minimum Information
This option prevents any output except the final trees and any error messages.

-q

Help
This option prints a help message describing the options and then quits.

-h