Documentation for stepwise MaxChi and Phylpro procedures

Jinko Graham, Brad McNeney and Françoise Seillier-Moiseiwitsch

July 26, 2004

Installation

Get the software Download and unpack the tar file stepwise.*version*.tar.gz or the zip file stepwise.*version*.zip, where *version* is the version number (e.g. 0.1). Unix users with access to gnu-tar, may unpack the tar file with

```
tar xzvf stepwise.0.1.tar.gz
```

Windows-users may unpack the tar file or zip file with their favorite archiving software. The source-code and example files will unpack in a directory called `stepwise`.

Compile Change to the directory `stepwise` and compile the source code with the makefile by typing `make`. The makefile has been tested on Linux (RedHat 8.0) and on WindowsXP with MinGW (http://www.mingw.org) version 3.1.0 installed.

Input Input files should be in Phylip input data format. See `http://evolution.genetics.washington.edu/phylip/doc/main.html#inputfiles` for details. Here is a short example, taken from the Phylip website:

```
   6   13
Archaeopt CGATGCTTAC CGC
HesperorniCGTTACTCGT TGT
BaluchitheTAATGTTAAT TGT
B. virginiTAATGTTCGT TGT
BrontosaurCAAAACCCAT CAT
B.subtilisGGCAGCCAAT CAC
```

The first line specifies the number of species and number of sites per species. The lines that follow contain the data, with exactly ten characters for the species name.

Run After executing the makefile, there should be executable files for both MaxChi and Phylpro in the `stepwise` directory (`maxchi` and `phylpro` on Unix systems or `maxchi.exe` and `phylpro.exe` on Windows). The executable files should be run from the command-line. The next sections give examples for how to run the executable files using the example alignment `simulinfile` provided with the source code in the `stepwise` directory.

Example - MaxChi

Suppose the program is installed in a directory called `stepwise`. Let `stepwise>` be the command-line prompt. To start the maxchi program and redirect its output to a file called `step1`, type

```
stepwise> maxchi simulinfile > step1
```

Unix users may have to precede `maxchi` with `./`, as in `./maxchi`, if the current working directory is not in your path. The program will prompt for the number of previously declared breaks (0 to start), the window half-width and the number of Monte Carlo replicates to use for approximating the permutation distribution:

```
Enter number of breaks: 0

Enter window half width to use: 30

Enter number of MC reps to use for the permutation distribution: 1000
```

Larger values of the window half-width lead to more stability in the test statistics under the null hypothesis of no breakpoints (or of no further breakpoints in subsequent steps). However, you will not be able to detect breakpoints within a window half-width of the ends of the alignment (or, in subsequent steps, within a window half-width of the previously declared breakpoints).

Windows users viewing the output file `step1` with a text editor such as Notepad will find that `step1` has Unix-style line breaks, rather than the usual DOS-style line-breaks. Notepad will not view this file properly, but WordPad, Word, or your web browser will. All users should note that their output may not match the output below due to the Monte Carlo error in determining the permutation null distribution. Specifically, critical values may differ slightly from run to run, and so the number of site-specific MaxChi statistics significant at both the 10% and 5% levels may vary as well.

```
-----------------------------------------------
Stepwise MaxChi, version 0.1 output
-----------------------------------------------

Read data from simulinfile: sample size 30 with 1000 bases

0 prior breaks

window half-width=30 sites.

1000 MC reps for the permutation distribution

There are 28 unique sequences in the 30 provided.

There are 239 ungapped polymorphic sites:
3 4 7 10 11 14 15 16 20 23 26 29 43 45 49 55 58 60 62 64
68 71 76 79 85 89 92 93 96 98 100 101 110 116 118 120 126 127 139 144
152 154 155 156 159 166 169 171 181 184 185 199 200 204 206 207 213 216 222 228
230 231 232 237 241 249 250 256 259 260 266 273 274 277 282 283 286 287 288 291
293 300 305 313 314 317 320 321 325 335 342 347 350 366 367 374 375 377 392 401
413 423 425 430 431 432 434 436 437 438 453 463 470 483 486 496 513 526 527 528
531 548 549 556 558 564 568 573 576 579 583 588 604 616 618 620 622 624 633 637
644 649 653 661 667 668 670 674 680 685 694 698 701 706 712 720 729 735 737 741
742 744 746 747 750 752 753 756 760 763 765 768 769 773 776 782 784 790 796 801
802 805 807 810 811 815 818 819 825 826 827 828 832 834 844 851 852 854 857 858
866 869 870 871 883 884 886 887 895 896 908 910 916 921 922 923 927 928 932 934
935 938 941 945 948 949 952 953 954 955 964 971 972 983 984 988 990 997 1000
-----------------------------------------------
There were 17 site-specific MaxChi statistics significant at the
10 percent level (90th percentile = 18.373, 95th percentile = 19.817):
```

```
Number Location  MaxChi   pairs
     1      531  20.000*  (sample17__:sample27__)
     2      548  20.000*  (sample17__:sample27__)
     3      549  20.000*  (sample17__:sample27__)
     4      556  20.000*  (sample17__:sample27__)
     5      616  20.000*  (sample19__:sample30__)
     6      620  18.468   (sample19__:sample30__)
     7      622  18.468   (sample19__:sample30__)
     8      807  19.461   (sample3___:sample8___)
                          (sample3___:sample9___)
     9      810  24.310*  (sample3___:sample8___)
                          (sample3___:sample9___)
    10      811  19.461   (sample3___:sample8___)
                          (sample3___:sample9___)
                          (sample3___:sample11__)
                          (sample3___:sample15__)
                          (sample3___:sample16__)
    11      815  19.461   (sample3___:sample8___)
                          (sample3___:sample9___)
                          (sample3___:sample11__)
                          (sample3___:sample15__)
                          (sample3___:sample16__)
    12      826  19.200   (sample3___:sample14__)
                          (sample3___:sample27__)
    13      827  21.172*  (sample3___:sample14__)
                          (sample3___:sample27__)
    14      828  19.200   (sample4___:sample14__)
                          (sample4___:sample27__)
    15      832  21.172*  (sample4___:sample14__)
                          (sample4___:sample27__)
    16      834  21.172*  (sample4___:sample13__)
                          (sample4___:sample14__)
                          (sample4___:sample27__)
    17      844  21.172*  (sample4___:sample13__)
                          (sample4___:sample14__)
                          (sample4___:sample27__)
-----------------------------------------------
Notes - "Location" is the polymorphic site just before the proposed breakpoint.
      - MaxChi statistics significant at the 5 percent level indicated by a *
```

On the basis of the above output at the first step, suppose we declare a breakpoint between polymorphic sites (531,548), a breakpoint between (616,618) and a breakpoint between (810,811). To run the program for a second step, conditioning on these three declared breakpoints, type

```
stepwise> maxchi simulinfile > step2
```

The program will again prompt for the number of previously declared breakpoints (3 now) along with their locations. Locations are specified by the nearest polymorphic site to the left of the breakpoint. The program will also prompt for the window half-width and the number of Monte Carlo replicates to use for estimating the permutation distribution, as before.

```
Enter number of breaks: 3

Enter the 3 ordered site(s) just before the break(s): 531 616 810
```

```
Enter window half width to use: 30

Enter number of MC reps to use for the permutation distribution: 1000
```

The contents of the file step2 is:

```
-----------------------------------------------
Stepwise MaxChi, version 0.1 output
-----------------------------------------------

Read data from simulinfile: sample size 30 with 1000 bases

3 prior breaks
 Breaks entered: 531 616 810

window half-width=30 sites.

1000 MC reps for the permutation distribution

There are 28 unique sequences in the 30 provided.

There are 239 ungapped polymorphic sites:
3 4 7 10 11 14 15 16 20 23 26 29 43 45 49 55 58 60 62 64
68 71 76 79 85 89 92 93 96 98 100 101 110 116 118 120 126 127 139 144
152 154 155 156 159 166 169 171 181 184 185 199 200 204 206 207 213 216 222 228
230 231 232 237 241 249 250 256 259 260 266 273 274 277 282 283 286 287 288 291
293 300 305 313 314 317 320 321 325 335 342 347 350 366 367 374 375 377 392 401
413 423 425 430 431 432 434 436 437 438 453 463 470 483 486 496 513 526 527 528
531 548 549 556 558 564 568 573 576 579 583 588 604 616 618 620 622 624 633 637
644 649 653 661 667 668 670 674 680 685 694 698 701 706 712 720 729 735 737 741
742 744 746 747 750 752 753 756 760 763 765 768 769 773 776 782 784 790 796 801
802 805 807 810 811 815 818 819 825 826 827 828 832 834 844 851 852 854 857 858
866 869 870 871 883 884 886 887 895 896 908 910 916 921 922 923 927 928 932 934
935 938 941 945 948 949 952 953 954 955 964 971 972 983 984 988 990 997 1000
-----------------------------------------------
There were 1 site-specific MaxChi statistics significant at the
10 percent level (90th percentile = 15.152, 95th percentile = 17.067):

Number Location  MaxChi   pairs
     1      342  17.143*  (sample15__:sample28__)
                          (sample16__:sample28__)
-----------------------------------------------
Notes - "Location" is the polymorphic site just before the proposed breakpoint.
      - MaxChi statistics significant at the 5 percent level indicated by a *
```

On the basis of the above output at the second step, suppose we declare an additional breakpoint between polymorphic sites (342,347). To run the program for a third step, conditioning on this breakpoint and the three others declared in the first step, type

```
stepwise> maxchi simulinfile > step3
```

The program will prompt for the number of previously declared breakpoints (4 now) along with their locations, the window half-width and the number of Monte Carlo replicates.

```
Enter number of breaks: 4

Enter the 4 ordered site(s) just before the break(s): 342 531 616 810

Enter window half width to use: 30

Enter number of MC reps to use for the permutation distribution: 1000
```

The contents of the file step3 is:

```
-----------------------------------------------
Stepwise MaxChi, version 0.1 output
-----------------------------------------------

Read data from simulinfile: sample size 30 with 1000 bases

4 prior breaks
 Breaks entered: 342 531 616 810

window half-width=30 sites.

1000 MC reps for the permutation distribution

There are 28 unique sequences in the 30 provided.

There are 239 ungapped polymorphic sites:
3 4 7 10 11 14 15 16 20 23 26 29 43 45 49 55 58 60 62 64
68 71 76 79 85 89 92 93 96 98 100 101 110 116 118 120 126 127 139 144
152 154 155 156 159 166 169 171 181 184 185 199 200 204 206 207 213 216 222 228
230 231 232 237 241 249 250 256 259 260 266 273 274 277 282 283 286 287 288 291
293 300 305 313 314 317 320 321 325 335 342 347 350 366 367 374 375 377 392 401
413 423 425 430 431 432 434 436 437 438 453 463 470 483 486 496 513 526 527 528
531 548 549 556 558 564 568 573 576 579 583 588 604 616 618 620 622 624 633 637
644 649 653 661 667 668 670 674 680 685 694 698 701 706 712 720 729 735 737 741
742 744 746 747 750 752 753 756 760 763 765 768 769 773 776 782 784 790 796 801
802 805 807 810 811 815 818 819 825 826 827 828 832 834 844 851 852 854 857 858
866 869 870 871 883 884 886 887 895 896 908 910 916 921 922 923 927 928 932 934
935 938 941 945 948 949 952 953 954 955 964 971 972 983 984 988 990 997 1000
-----------------------------------------------
There were 2 site-specific MaxChi statistics significant at the
10 percent level (90th percentile = 13.303, 95th percentile = 15.152):

Number Location  MaxChi   pairs
     1      155  14.067   (sample3___:sample12__)
                          (sample4___:sample12__)
     2      156  14.067   (sample3___:sample12__)
                          (sample4___:sample12__)
-----------------------------------------------
Notes - "Location" is the polymorphic site just before the proposed breakpoint.
      - MaxChi statistics significant at the 5 percent level indicated by a *
```

Since there are no further declared breakpoints significant at the 5% level, we end the stepwise search.

Example - Phylpro

To start the phylpro program and redirect its output to a file called `step1`, type

```
stepwise> phylpro simulinfile > step1
```

Unix users may have to precede `phylpro` with `./`, as in `./phylpro`, if the current working directory is not in your path. As was the case for Maxchi, the program will prompt for the number of previously declared breaks (0 to start), the window half-width and the number of Monte Carlo replicates for approximating the permutation distribution:

```
Enter number of breaks: 0

Enter window half width to use: 30

Enter number of MC reps to use for the permutation distribution: 1000
```

Again, all users should note that their output may not match the output below due to the Monte Carlo error in determining the permutation null distribution. Specifically, critical values may differ slightly from run to run, and so the number of site-specific MinCor statistics significant at both the 10% and 5% levels may vary as well.

```
-----------------------------------------------
Stepwise Phylpro, version 0.1 output
-----------------------------------------------

Read data from simulinfile: sample size 30 with 1000 bases

0 prior breaks

window half-width=30 sites.

1000 MC reps for the permutation distribution

There are 28 unique sequences in the 30 provided.

There are 239 ungapped polymorphic sites:
3 4 7 10 11 14 15 16 20 23 26 29 43 45 49 55 58 60 62 64
68 71 76 79 85 89 92 93 96 98 100 101 110 116 118 120 126 127 139 144
152 154 155 156 159 166 169 171 181 184 185 199 200 204 206 207 213 216 222 228
230 231 232 237 241 249 250 256 259 260 266 273 274 277 282 283 286 287 288 291
293 300 305 313 314 317 320 321 325 335 342 347 350 366 367 374 375 377 392 401
413 423 425 430 431 432 434 436 437 438 453 463 470 483 486 496 513 526 527 528
531 548 549 556 558 564 568 573 576 579 583 588 604 616 618 620 622 624 633 637
644 649 653 661 667 668 670 674 680 685 694 698 701 706 712 720 729 735 737 741
742 744 746 747 750 752 753 756 760 763 765 768 769 773 776 782 784 790 796 801
802 805 807 810 811 815 818 819 825 826 827 828 832 834 844 851 852 854 857 858
866 869 870 871 883 884 886 887 895 896 908 910 916 921 922 923 927 928 932 934
935 938 941 945 948 949 952 953 954 955 964 971 972 983 984 988 990 997 1000
-----------------------------------------------
There were 36 site-specific minimum correlation statistics significant at the
10 percent level (10th percentile = -0.300, 5th percentile = -0.374):

Number Location  MinCor   targets
     1      437  -0.330   sample16__
     2      463  -0.391*  sample28__
```

```
     3      470  -0.384*  sample28__
     4      483  -0.314   sample28__
     5      496  -0.442*  sample28__
     6      513  -0.549*  sample28__
     7      526  -0.656*  sample28__
     8      527  -0.627*  sample28__
     9      528  -0.583*  sample28__
    10      531  -0.648*  sample28__
    11      548  -0.696*  sample28__
    12      549  -0.691*  sample28__
    13      556  -0.684*  sample28__
    14      558  -0.635*  sample28__
    15      564  -0.631*  sample28__
    16      568  -0.625*  sample28__
    17      573  -0.641*  sample28__
    18      576  -0.528*  sample28__
    19      579  -0.539*  sample28__
    20      583  -0.536*  sample28__
    21      588  -0.616*  sample28__
    22      604  -0.509*  sample28__
    23      616  -0.475*  sample28__
    24      618  -0.415*  sample28__
    25      620  -0.342   sample28__
    26      622  -0.356   sample28__
    27      735  -0.492*  sample29__
    28      737  -0.406*  sample29__
    29      741  -0.449*  sample29__
    30      742  -0.381*  sample29__
    31      744  -0.460*  sample29__
    32      746  -0.345   sample29__
    33      747  -0.336   sample29__
    34      750  -0.323   sample29__
    35      752  -0.329   sample29__
    36      756  -0.342   sample29__
-----------------------------------------------
Notes - "Location" is the polymorphic site just before the proposed breakpoint.
      - MinCor statistics significant at the 5 percent level indicated by a *
```

On the basis of the above output at the first step, suppose we declare breakpoints between polymorphic sites (548,549) and (735,737). To run the program for the second step, conditioning on these two declared breakpoints, type

```
stepwise> phylpro simulinfile > step2
```

The program will again prompt for the number of previously declared breakpoints (2 now) along with their locations, the window half-width and the number of Monte Carlo replicates:

```
Enter number of breaks: 2

Enter the 2 ordered site(s) just before the break(s): 548 735

Enter window half width to use: 30

Enter number of MC reps to use for the permutation distribution: 1000
```

The contents of the file step2 is:

```
-----------------------------------------------
Stepwise Phylpro, version 0.1 output
-----------------------------------------------

Read data from simulinfile: sample size 30 with 1000 bases

2 prior breaks
 Breaks entered: 548 735

window half-width=30 sites.

1000 MC reps for the permutation distribution

There are 28 unique sequences in the 30 provided.

There are 239 ungapped polymorphic sites:
3 4 7 10 11 14 15 16 20 23 26 29 43 45 49 55 58 60 62 64
68 71 76 79 85 89 92 93 96 98 100 101 110 116 118 120 126 127 139 144
152 154 155 156 159 166 169 171 181 184 185 199 200 204 206 207 213 216 222 228
230 231 232 237 241 249 250 256 259 260 266 273 274 277 282 283 286 287 288 291
293 300 305 313 314 317 320 321 325 335 342 347 350 366 367 374 375 377 392 401
413 423 425 430 431 432 434 436 437 438 453 463 470 483 486 496 513 526 527 528
531 548 549 556 558 564 568 573 576 579 583 588 604 616 618 620 622 624 633 637
644 649 653 661 667 668 670 674 680 685 694 698 701 706 712 720 729 735 737 741
742 744 746 747 750 752 753 756 760 763 765 768 769 773 776 782 784 790 796 801
802 805 807 810 811 815 818 819 825 826 827 828 832 834 844 851 852 854 857 858
866 869 870 871 883 884 886 887 895 896 908 910 916 921 922 923 927 928 932 934
935 938 941 945 948 949 952 953 954 955 964 971 972 983 984 988 990 997 1000
-----------------------------------------------
There were 11 site-specific minimum correlation statistics significant at the
10 percent level (10th percentile = 0.080, 5th percentile = 0.007):

Number Location  MinCor   targets
     1      154  0.027   sample13__
     2      199  0.079   sample11__,sample12__,sample13__,sample30__
     3      826  0.016   sample18__
     4      827  -0.041*  sample18__
     5      828  -0.047*  sample18__
     6      832  -0.097*  sample18__
     7      834  -0.087*  sample18__
     8      844  -0.093*  sample18__
     9      851  0.038   sample18__
    10      854  0.058   sample18__
    11      858  0.072   sample18__
-----------------------------------------------
Notes - "Location" is the polymorphic site just before the proposed breakpoint.
      - MinCor statistics significant at the 5 percent level indicated by a *
```

On the basis of the above output at the second step, suppose we declare an additional breakpoint between polymorphic sites (832,834). To run the program for a third step, conditioning on this breakpoint and the two others declared in the first step, type

```
stepwise> phylpro simulinfile > step3
```

The program will prompt for the number of previously declared breakpoints (3 now) along with their locations, the window half-width and the number of Monte Carlo replicates.

```
Enter number of breaks: 3

Enter the 3 ordered site(s) just before the break(s): 548 735 832

Enter window half width to use: 30

Enter number of MC reps to use for the permutation distribution: 1000
```

The contents of the file step3 is:

```
-----------------------------------------------
Stepwise Phylpro, version 0.1 output
-----------------------------------------------

Read data from simulinfile: sample size 30 with 1000 bases

3 prior breaks
 Breaks entered: 548 735 832

window half-width=30 sites.

1000 MC reps for the permutation distribution

There are 28 unique sequences in the 30 provided.

There are 239 ungapped polymorphic sites:
3 4 7 10 11 14 15 16 20 23 26 29 43 45 49 55 58 60 62 64
68 71 76 79 85 89 92 93 96 98 100 101 110 116 118 120 126 127 139 144
152 154 155 156 159 166 169 171 181 184 185 199 200 204 206 207 213 216 222 228
230 231 232 237 241 249 250 256 259 260 266 273 274 277 282 283 286 287 288 291
293 300 305 313 314 317 320 321 325 335 342 347 350 366 367 374 375 377 392 401
413 423 425 430 431 432 434 436 437 438 453 463 470 483 486 496 513 526 527 528
531 548 549 556 558 564 568 573 576 579 583 588 604 616 618 620 622 624 633 637
644 649 653 661 667 668 670 674 680 685 694 698 701 706 712 720 729 735 737 741
742 744 746 747 750 752 753 756 760 763 765 768 769 773 776 782 784 790 796 801
802 805 807 810 811 815 818 819 825 826 827 828 832 834 844 851 852 854 857 858
866 869 870 871 883 884 886 887 895 896 908 910 916 921 922 923 927 928 932 934
935 938 941 945 948 949 952 953 954 955 964 971 972 983 984 988 990 997 1000
-----------------------------------------------
There were 6 site-specific minimum correlation statistics significant at the
10 percent level (10th percentile = 0.112, 5th percentile = 0.027):

Number Location  MinCor   targets
     1      126  0.104   sample13__
     2      144  0.094   sample13__
     3      154  0.027*  sample13__
     4      155  0.110   sample13__
     5      185  0.083   sample11__,sample12__,sample13__,sample30__
     6      199  0.079   sample11__,sample12__,sample13__,sample30__
-----------------------------------------------
Notes - "Location" is the polymorphic site just before the proposed breakpoint.
      - MinCor statistics significant at the 5 percent level indicated by a *
```

On the basis of this output at the third step, suppose we declare an additional breakpoint between polymorphic sites (154,155). To run the program for a fourth step conditioning on this breakpoint and the three others declared in the first and second steps, type

```
stepwise> phylpro simulinfile > step4
```

The program will prompt for the number of previously declared breakpoints (4 now) along with their locations, the window half-width and the number of Monte Carlo replicates.

```
Enter number of breaks: 4

Enter the 4 ordered site(s) just before the break(s): 154 548 735 832

Enter window half width to use: 30

Enter number of MC reps to use for the permutation distribution: 1000
```

The contents of the file step4 is:

```
-----------------------------------------------
Stepwise Phylpro, version 0.1 output
-----------------------------------------------

Read data from simulinfile: sample size 30 with 1000 bases

4 prior breaks
 Breaks entered: 154 548 735 832

window half-width=30 sites.

1000 MC reps for the permutation distribution

There are 28 unique sequences in the 30 provided.

There are 239 ungapped polymorphic sites:
3 4 7 10 11 14 15 16 20 23 26 29 43 45 49 55 58 60 62 64
68 71 76 79 85 89 92 93 96 98 100 101 110 116 118 120 126 127 139 144
152 154 155 156 159 166 169 171 181 184 185 199 200 204 206 207 213 216 222 228
230 231 232 237 241 249 250 256 259 260 266 273 274 277 282 283 286 287 288 291
293 300 305 313 314 317 320 321 325 335 342 347 350 366 367 374 375 377 392 401
413 423 425 430 431 432 434 436 437 438 453 463 470 483 486 496 513 526 527 528
531 548 549 556 558 564 568 573 576 579 583 588 604 616 618 620 622 624 633 637
644 649 653 661 667 668 670 674 680 685 694 698 701 706 712 720 729 735 737 741
742 744 746 747 750 752 753 756 760 763 765 768 769 773 776 782 784 790 796 801
802 805 807 810 811 815 818 819 825 826 827 828 832 834 844 851 852 854 857 858
866 869 870 871 883 884 886 887 895 896 908 910 916 921 922 923 927 928 932 934
935 938 941 945 948 949 952 953 954 955 964 971 972 983 984 988 990 997 1000
-----------------------------------------------
There were 1 site-specific minimum correlation statistics significant at the
10 percent level (10th percentile = 0.194, 5th percentile = 0.066):

Number Location  MinCor   targets
     1      347  0.169   sample29__
-----------------------------------------------
Notes - "Location" is the polymorphic site just before the proposed breakpoint.
      - MinCor statistics significant at the 5 percent level indicated by a *
```

Since there are no further declared breakpoints significant at the 5% level, we end the stepwise search.