MCB 5472

Phylogenetic reconstruction
PHYLIP, phyml

Peter Gogarten
Office: *BSP 404*
phone: *860 486-4061,*
Email: *gogarten@uconn.edu*

Terminology - reminder

Related terms:
autapomorphy = a derived character that is only present in one group; an autapomorphic character does not tell us anything about the relationship of the group that has this character ot other groups.

homoplasy = a derived character that was derived twice independently (convergent evolution). ***Note that the characters in question might still be homologous (e.g. a position in a sequence alignment, frontlimbs turned into wings in birds and bats).***

paraphyletic = a taxonomic group that is defined by a common ancestor, however, the common ancestor of this group also has decendants that do not belong to this taxonomic group. Many systematists despise paraphyletic groups (and consider them to be polyphyletic). Examples for paraphyletic groups are reptiles and protists. Many consider the archaea to be paraphyletic as well.

holophyletic = same as above, but the common ancestor gave rise only to members of the group.

Terminology- reminder

- Branches, splits, bipartitions
- In a rooted tree: clades
- Mono-, Para-, polyphyletic groups, cladists and a natural taxonomy

The term cladogram refers to a strictly bifurcating diagram, where each clade is defined by a common ancestor that only gives rise to members of this clade. I.e., a clade is monophyletic (derived from one ancestor) as opposed to polyphyletic (derived from many ancestors). (note you need to know where the root is!)

A clade is recognized and defined by shared derived characters (= synapomorphies). Shared primitive characters (= sympleisiomorphies , aternativie spelling is symplesiomorphies) do not define a clade. (see in class example drawing ala Hennig).

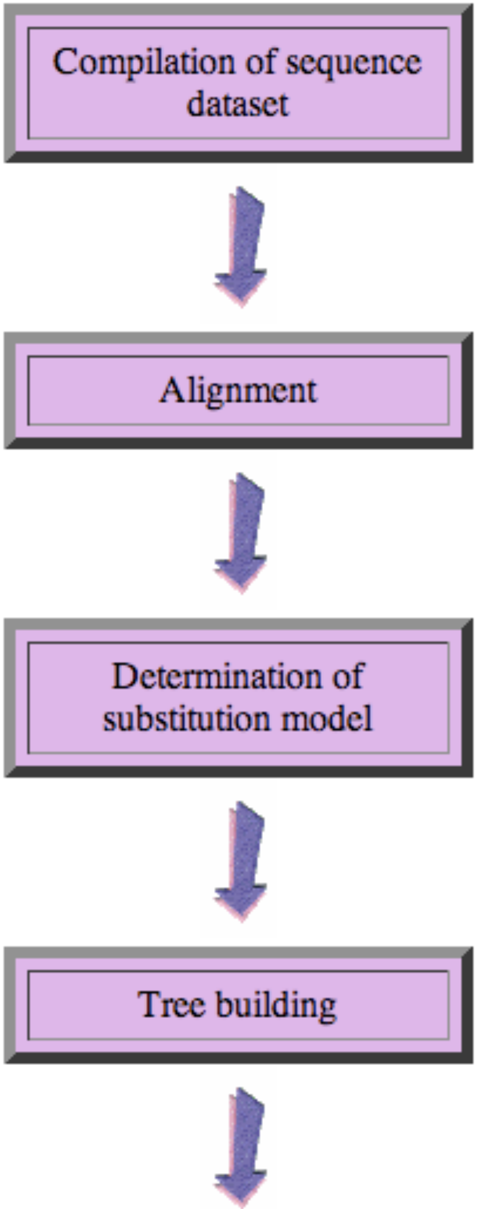
To use these terms you need to have polarized characters; for most molecular characters you don't know which state is primitive and which is derived (exceptions:....).

Steps of the phylogenetic analysis

Phylogenetic analysis is an inference of evolutionary relationships between organisms. Phylogenetics tries to answer the question “How did groups of organisms come into existence?”

Those relationships are usually represented by tree-like diagrams.

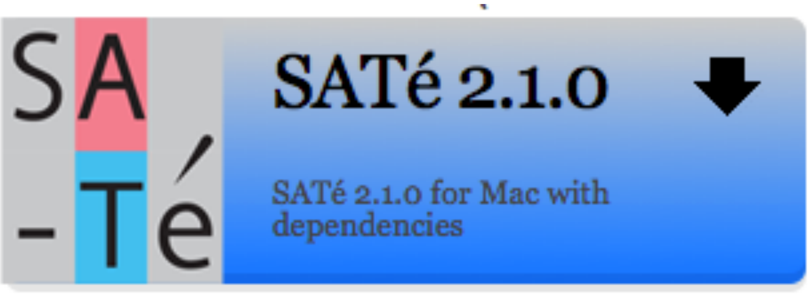
Note: the assumption of a tree-like process of evolution is controversial!


Compilation of sequence dataset
Alignment
Determination of substitution model
Tree building
Tree evaluation

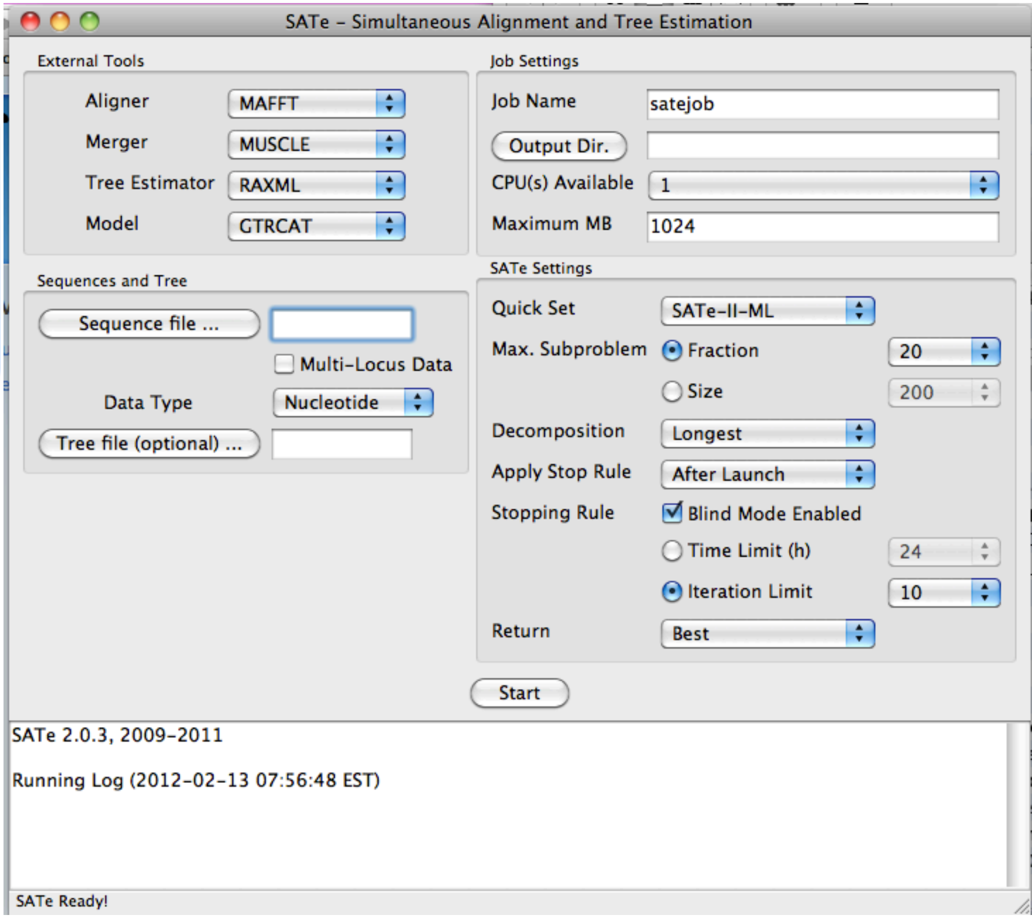
SATé

Simultaneous Alignment and Tree Estimation

http://phylo.bio.ku.edu/software/sate/sate.html



GUI works well on iMacs, but uses only local processors.


SATe – Simultaneous Alignment and Tree Estimation

External Tools

Aligner: MAFFT
Merger: MUSCLE
Tree Estimator: RAXML
Model: GTRCAT

Job Settings

Job Name: satejob
Output Dir.:
CPU(s) Available: 1
Maximum MB: 1024

Sequences and Tree

Sequence file ...:
□ Multi-Locus Data
Data Type: Nucleotide
Tree file (optional) ...:

SATe Settings

Quick Set: SATe-II-ML
Max. Subproblem: ● Fraction 20
○ Size 200
Decomposition: Longest
Apply Stop Rule: After Launch
Stopping Rule: ☑ Blind Mode Enabled
○ Time Limit (h) 24
● Iteration Limit 10
Return: Best

```
SATe 2.0.3, 2009-2011

Running Log (2012-02-13 07:56:48 EST)
```

SATe Ready!

Phylogenetic reconstruction - **How**

Distance analyses

calculate pairwise distances
(different distance measures, correction for multiple hits, correction for codon bias)

make distance matrix (table of pairwise corrected distances)

calculate tree from distance matrix

i) using optimality criterion
(e.g.: smallest error between distance matrix
and distances in tree, or use
ii) algorithmic approaches (UPGMA or neighbor joining) B)

Phylogenetic reconstruction - **How**

Parsimony analyses

find that tree that explains sequence data with minimum number of substitutions

(tree includes hypothesis of sequence at each of the nodes)

Maximum Likelihood analyses

given a model for sequence evolution, find the tree that has the highest probability under this model.

This approach can also be used to successively refine the model.

Bayesian statistics use ML analyses to calculate posterior probabilities for trees, clades and evolutionary parameters. Especially MCMC approaches have become very popular in the last year, because they allow to estimate evolutionary parameters (e.g., which site in a virus protein is under positive selection), without assuming that one actually knows the "true" phylogeny.

Else:
spectral analyses, like evolutionary parsimony, look only at patterns of substitutions,

Another way to categorize methods of phylogenetic reconstruction is to ask if they are using

an optimality criterion (e.g.: smallest error between distance matrix and distances in tree, least number of steps, highest probability), or

algorithmic approaches (UPGMA or neighbor joining)

Packages and programs available: PHYLIP, phyml, MrBayes, Tree-Puzzle, PAUP*, clustalw, raxml, PhyloGenie, PyPhy

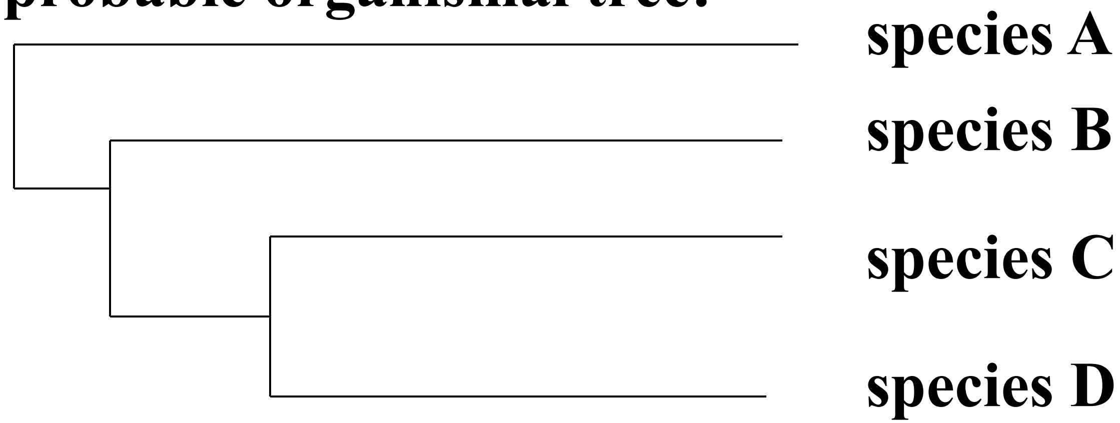
Assessing reliability

- For a discussion of Bootstrapping go here http://web.uconn.edu/gogarten/mcb221_2003/class28.html
- Alternatives: posterior probabilities, aLRT, quartet puzzling support values
- Note: for the trained eye, branch lengths are an important criterion to assess reliability

Trees – what might they mean?

Calculating a tree is comparatively easy, figuring out what it might mean is much more difficult.

If this is the probable organismal tree:



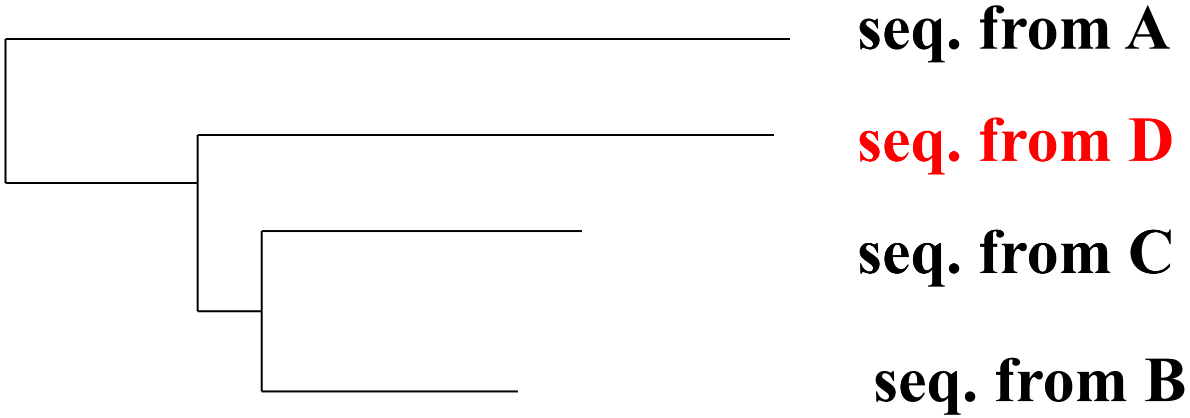
species A

species B

species C

species D

what could be the reason for obtaining this gene tree:



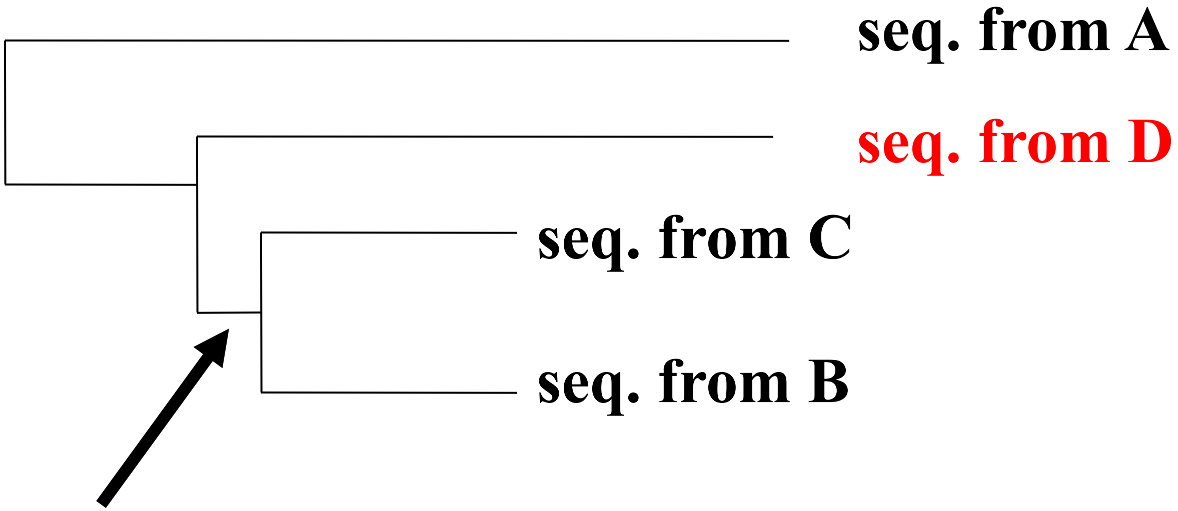
seq. from A

seq. from D

seq. from C

seq. from B

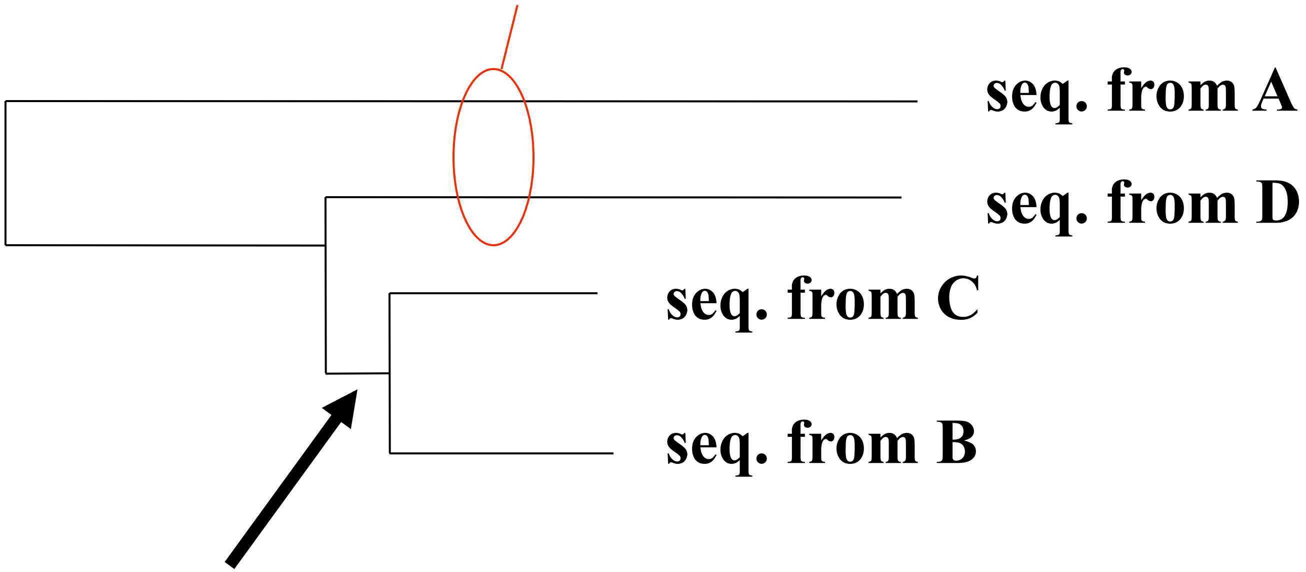
lack of resolution



e.g., 60% bootstrap support for bipartition (AD)(CB)

long branch attraction artifact

the two longest branches join together



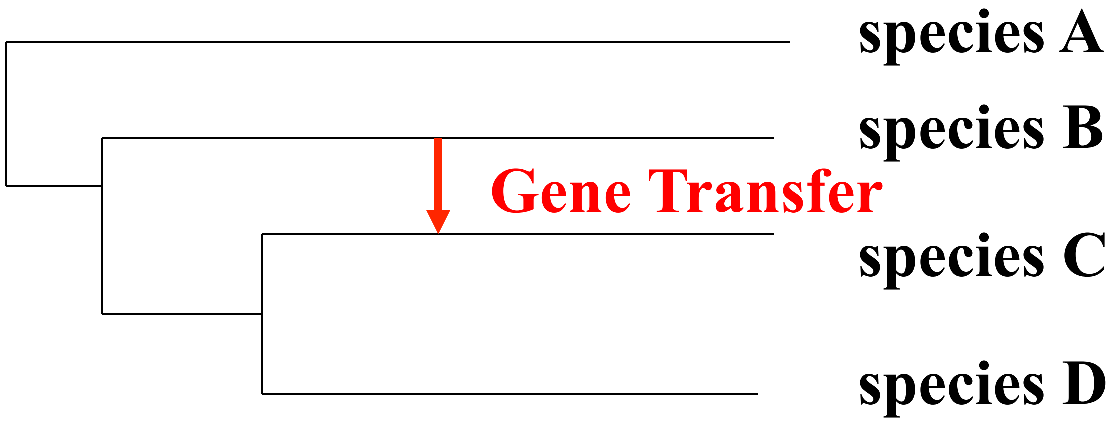
e.g., 100% bootstrap support for bipartition (AD)(CB)

What could you do to investigate if this is a possible explanation?

use only slow positions,
use an algorithm that corrects for ASRV
add sequences that break-up the long branches

Gene transfer

Organismal tree:



species A

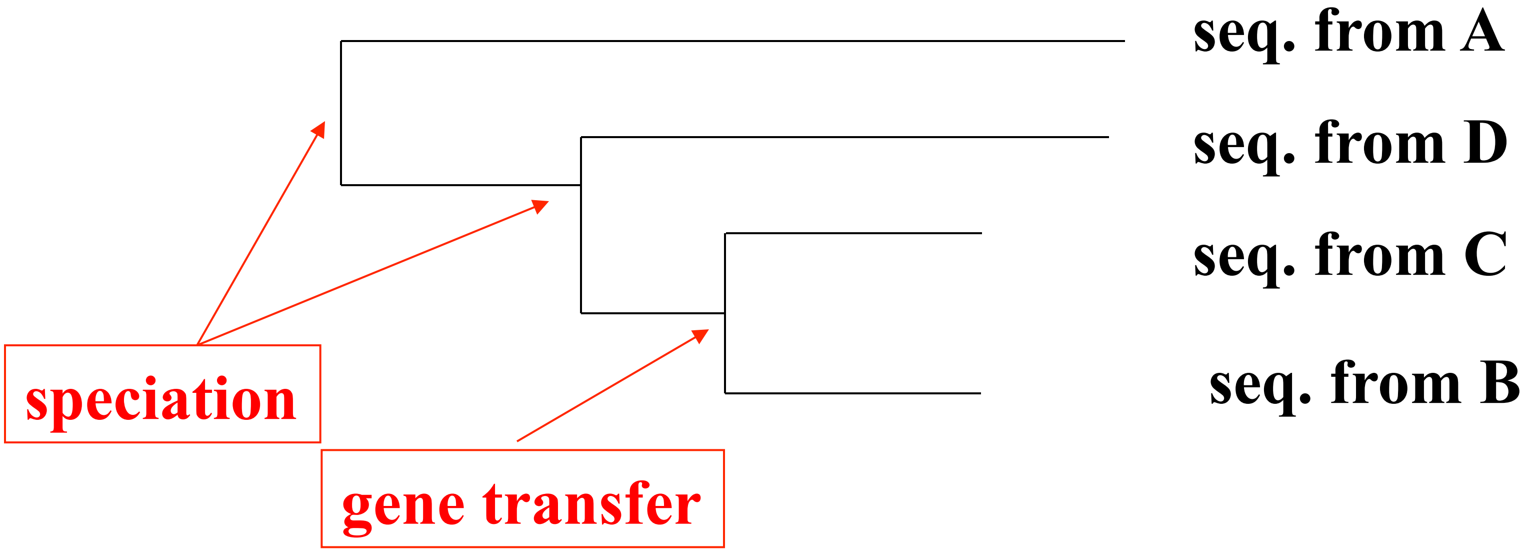
species B

Gene Transfer

species C

species D

molecular tree:



seq. from A

seq. from D

seq. from C

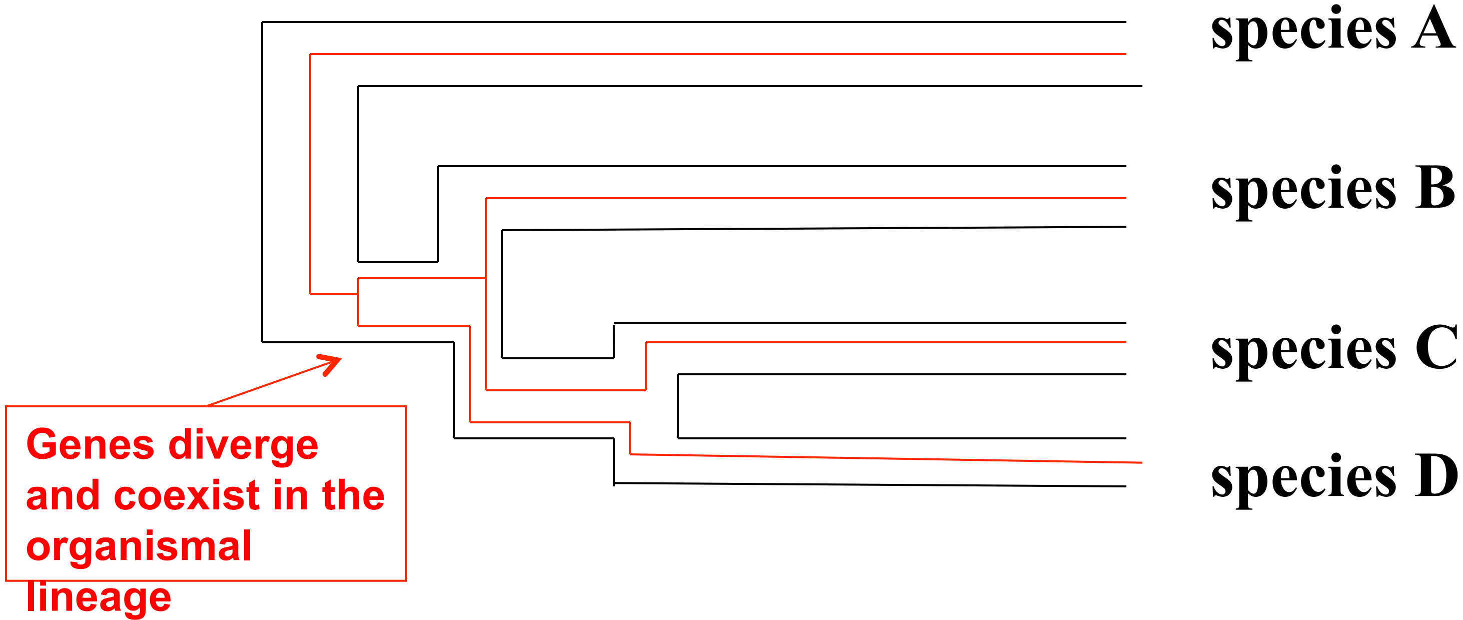
seq. from B

speciation

gene transfer

Lineage Sorting

Organismal tree:



species A

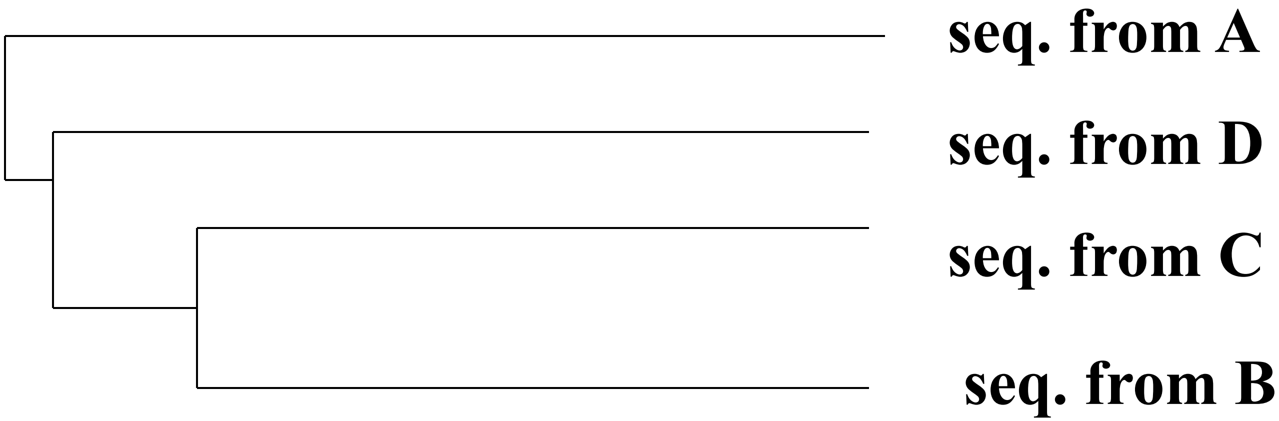
species B

species C

species D

Genes diverge and coexist in the organismal lineage

molecular tree:



seq. from A

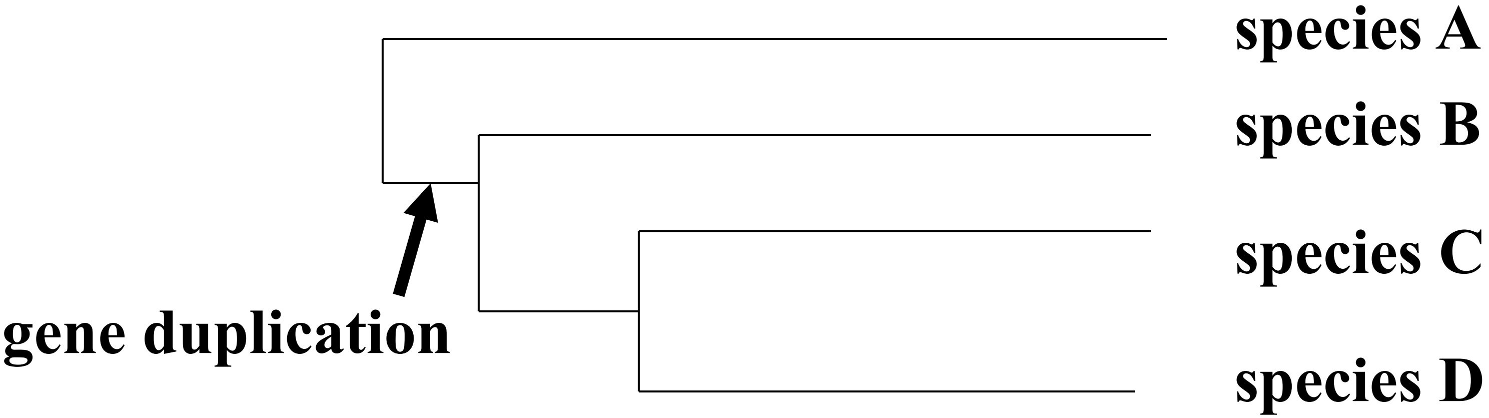
seq. from D

seq. from C

seq. from B

Gene duplication

Organismal tree:



species A

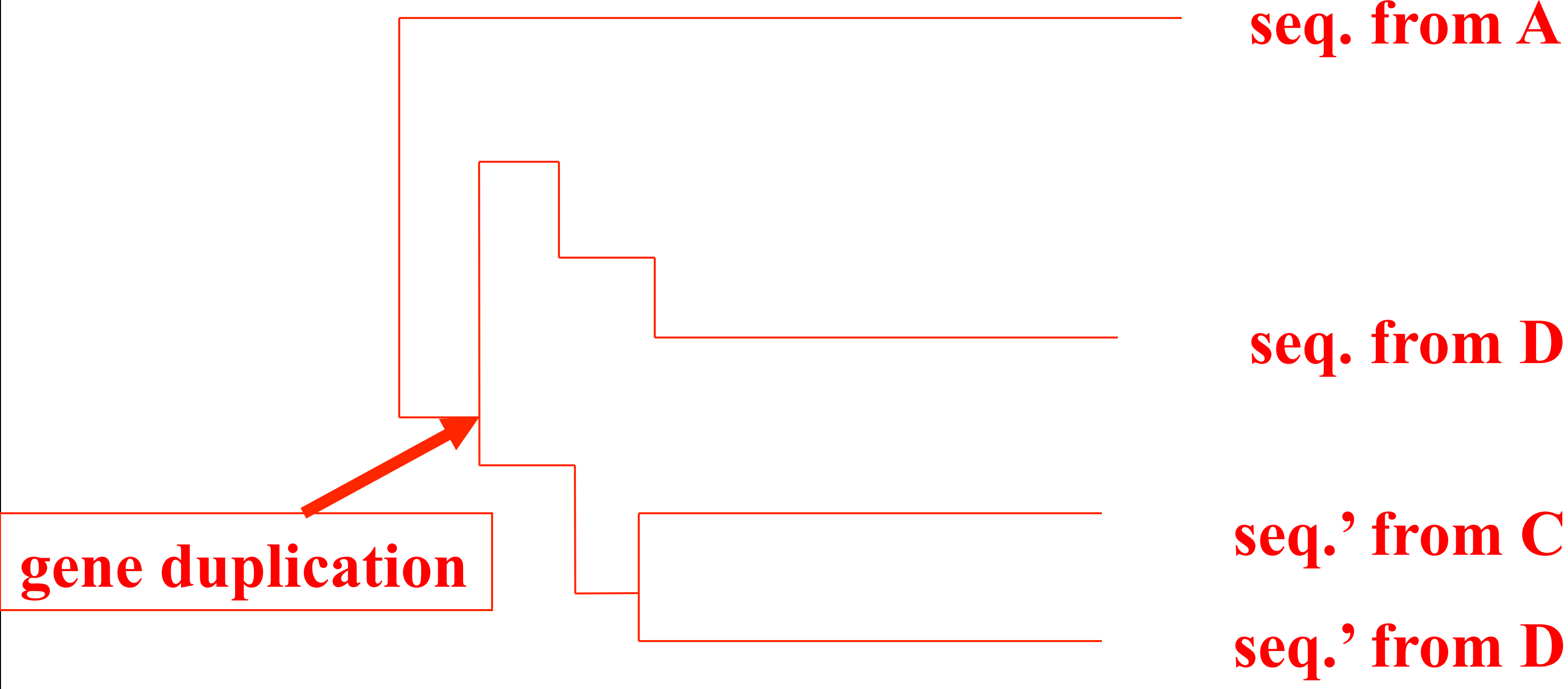
species B

species C

species D

gene duplication

molecular tree:



seq. from A

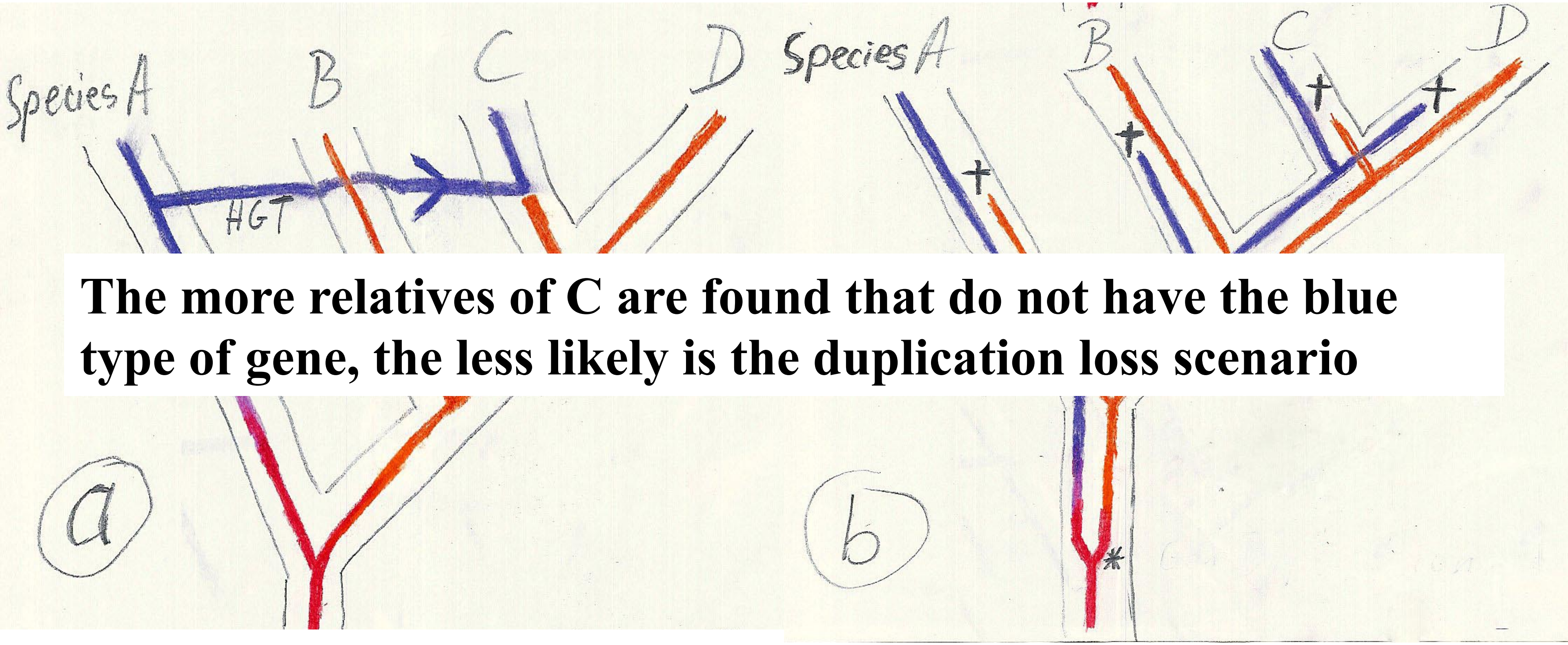
seq. from D

seq.’ from C

seq.’ from D

gene duplication

Gene duplication and gene transfer are equivalent explanations.



The more relatives of C are found that do not have the blue type of gene, the less likely is the duplication loss scenario

Horizontal or lateral Gene

Ancient duplication followed by gene loss

Note that scenario B involves many more individual events than A

1 HGT with orthologous replacement

1 gene duplication followed by 4 independent gene loss events

What is it good for?

Gene duplication events can provide an outgroup that allows rooting a molecular phylogeny.

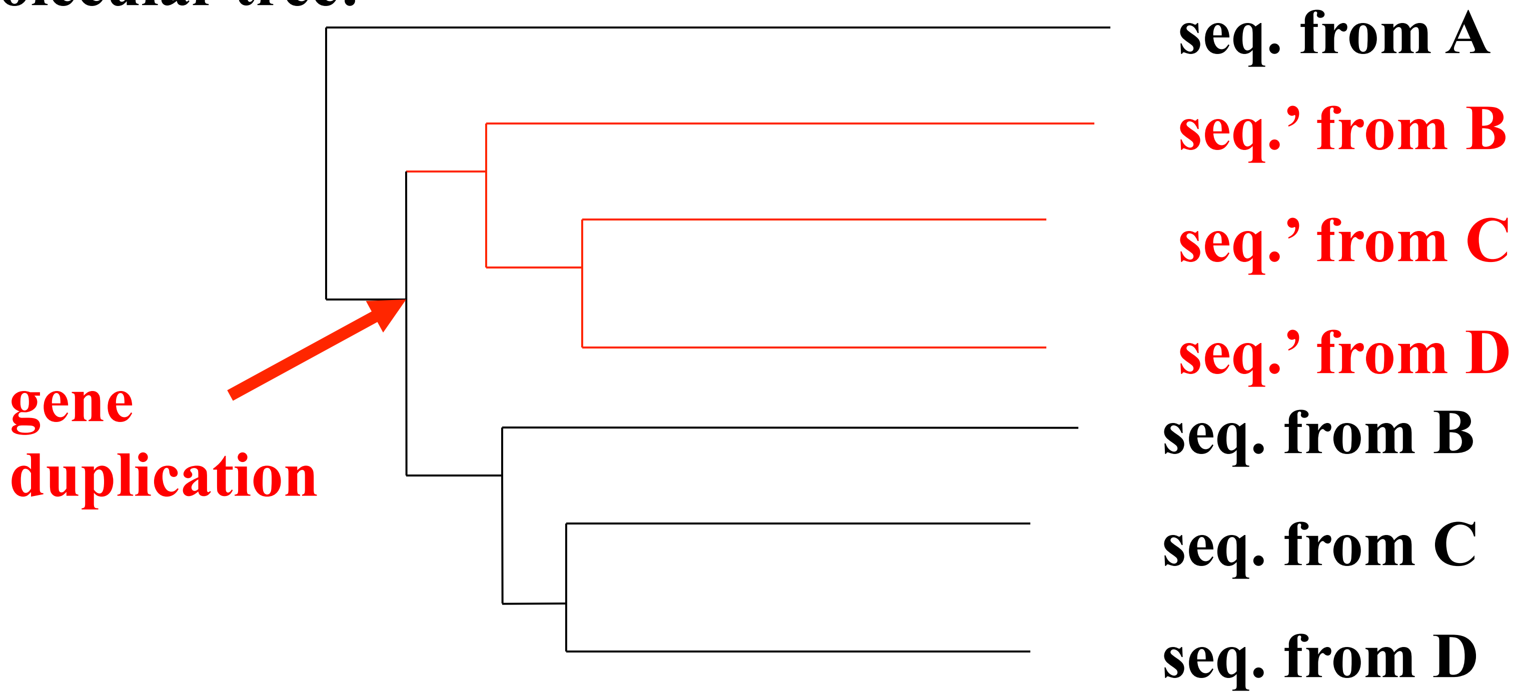
Most famously this principle was applied in case of the tree of life – the only outgroup available in this case are ancient paralogs (see http://gogarten.uconn.edu/cvs/Publ_Pres.htm for more info).

However, the same principle also is applicable to any group of organisms, where a duplication preceded the radiation (example).

Lineage specific duplications also provide insights into which traits were important during evolution of a lineage.

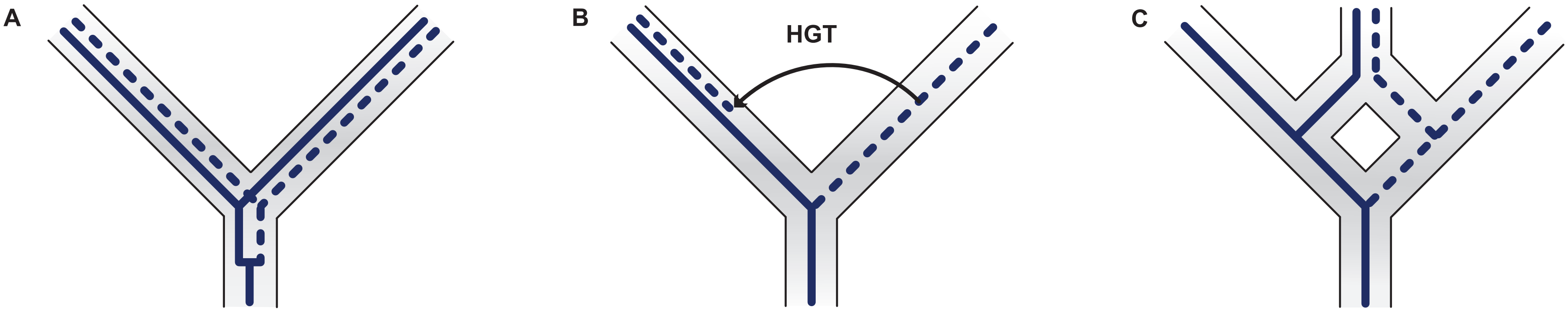
Function, ortho- and paralogy

molecular tree:



The presence of the duplication is a taxonomic character (shared derived character in species B C D).
The phylogeny suggests that seq’ and seq have similar function, and that this function was important in the evolution of the clade BCD.
seq’ in B and seq’in C and D are orthologs and probably have the same function, whereas seq and seq’ in BCD probably have different function (the difference might be in subfunctionalization of functions that seq had in A. – e.g. organ specific expression)

Aside: Gene and genome duplication
versus
Horizontal Gene Transfer



Autochtonous gene/genome duplication

**Gene "duplication" through horizontal gene transfer –
the most common process in prokaryotes**

**Genome "duplication" through allopolypoidization
Another reticulation process that is not often recognized as such.**

Old Assignment

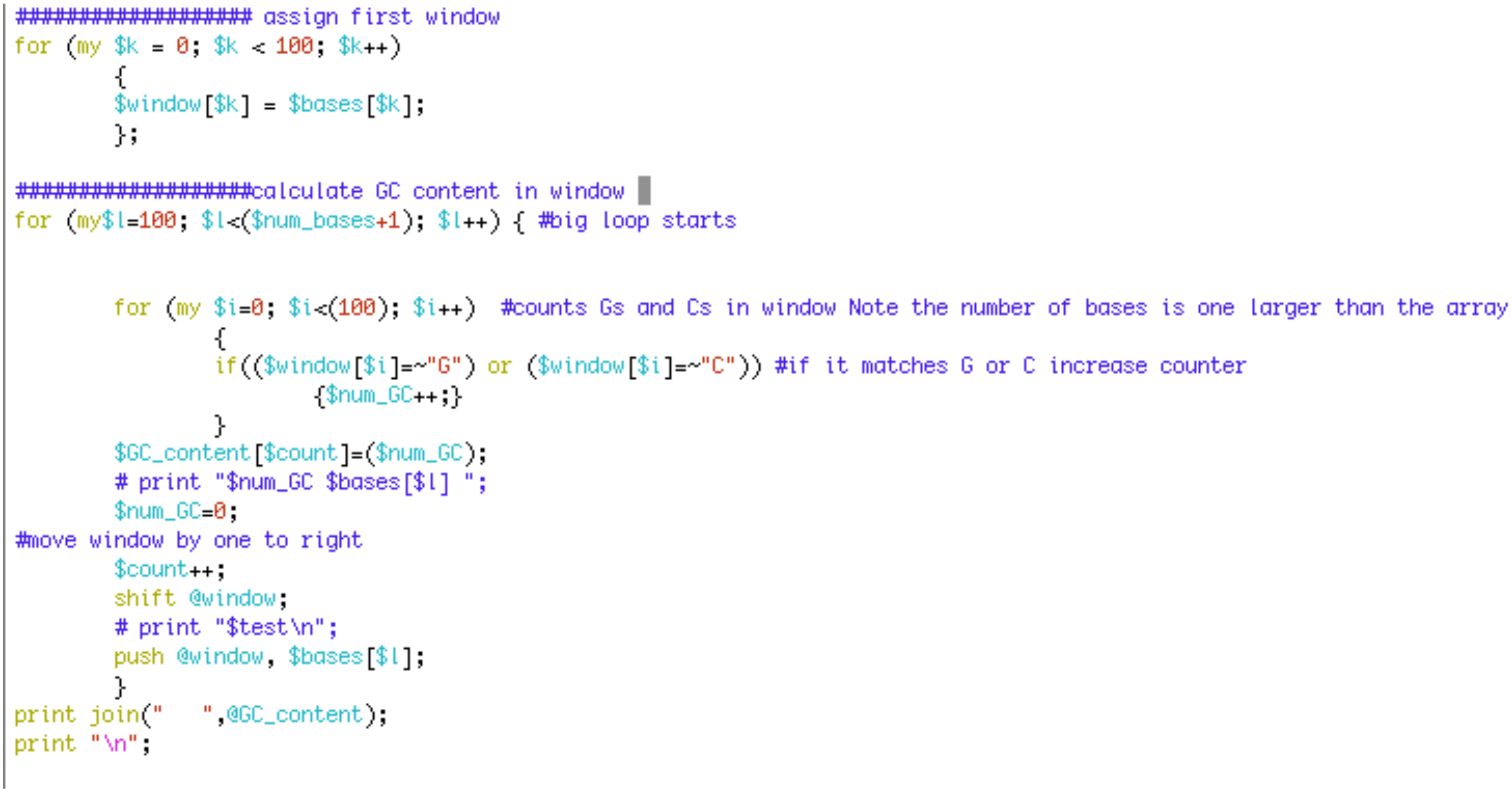
Write a script that takes a genome (fna file) and calculates the GC content.

Modify the script to count tetra- and penta-nucleotide frequencies.

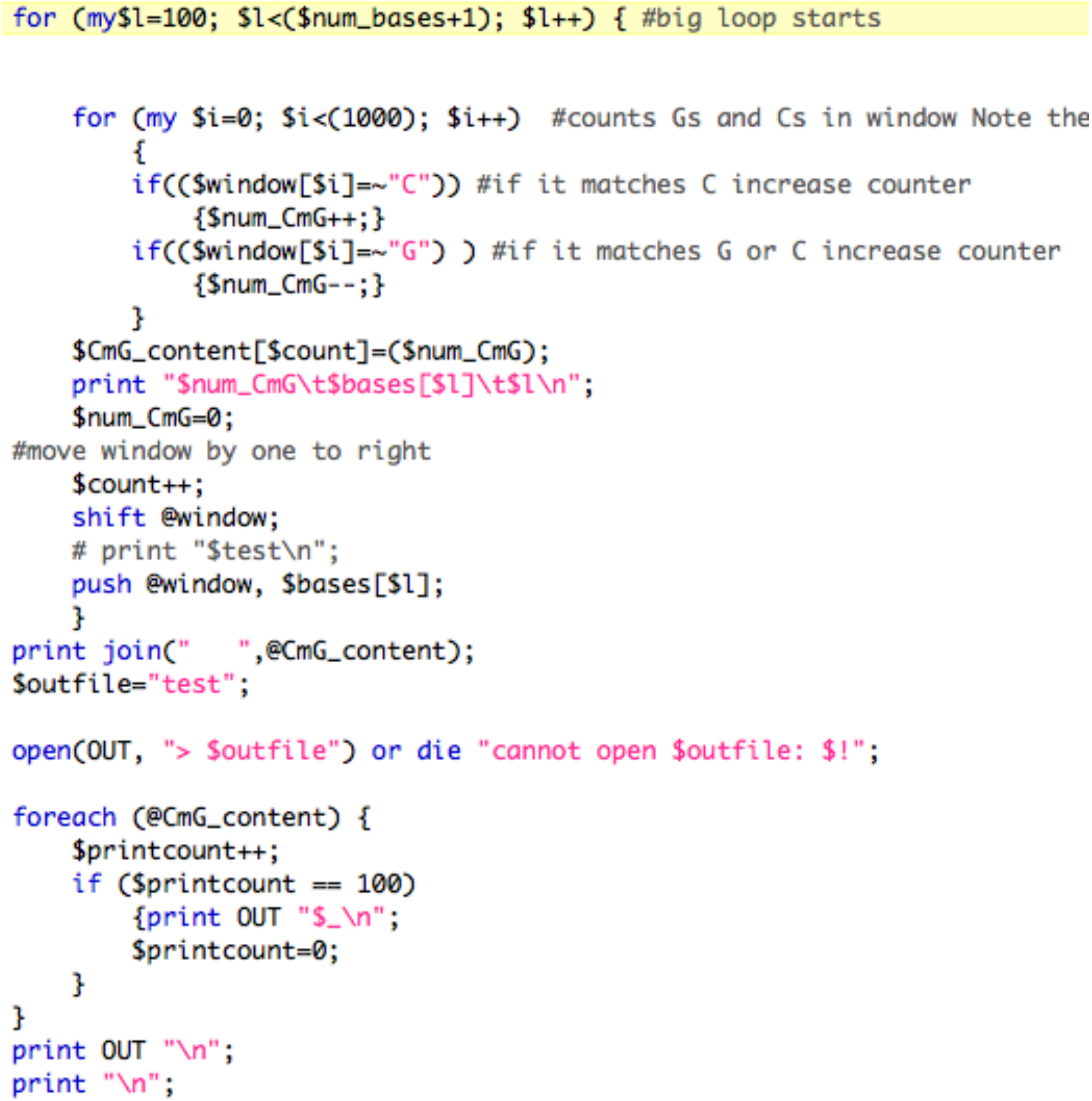
Modify the script so that it creates a table that gives the GC, tetra- and penta-nucleotide content in a sliding window moving through the genome.

(For which of these programs, and for which problems might you want to consider, or correct for strand bias?)

Rolling window example from GC rolling.pl

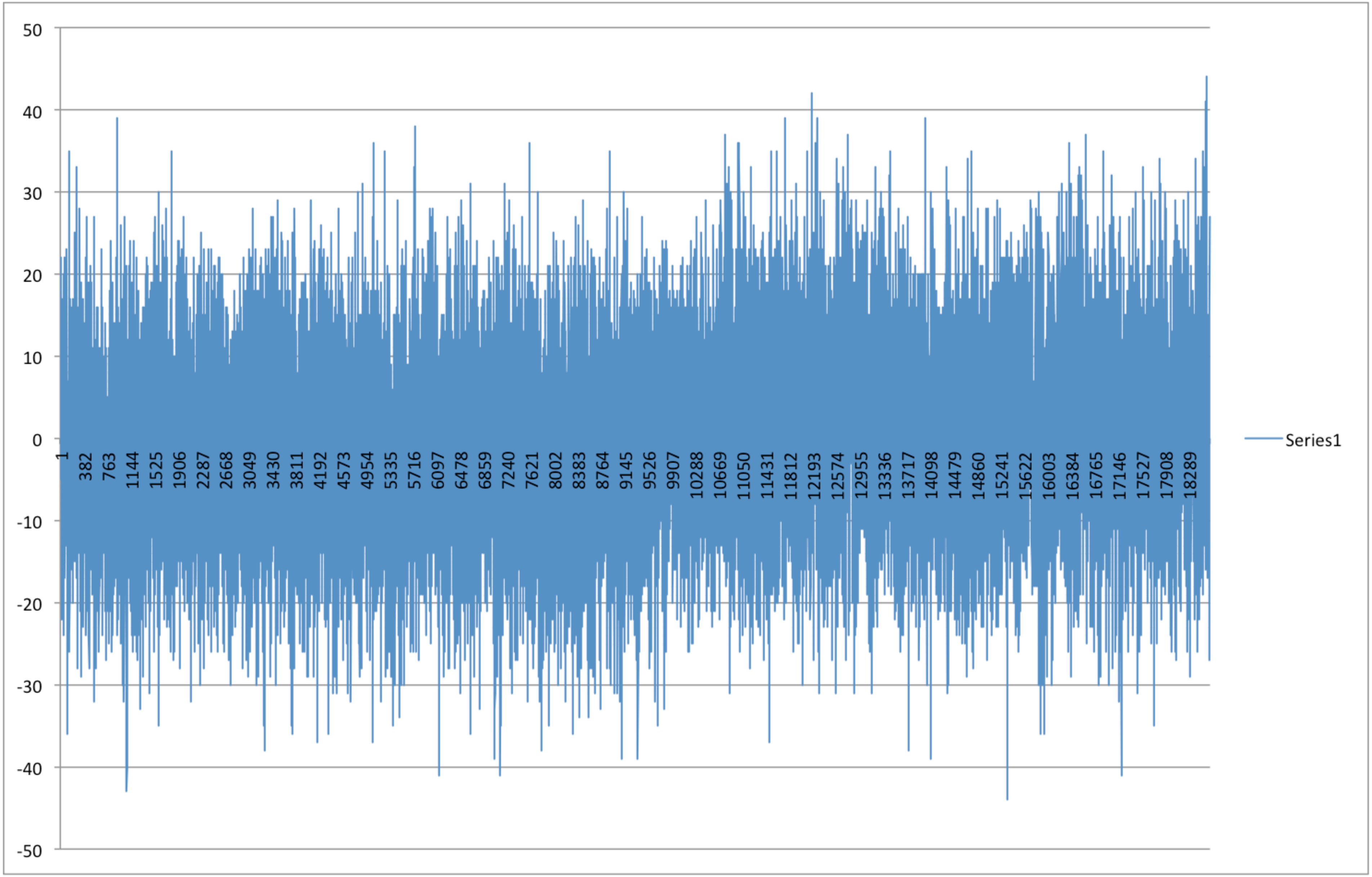


```
################### assign first window
for (my $k = 0; $k < 100; $k++)
        {
        $window[$k] = $bases[$k];
        };

###################calculate GC content in window
for (my$l=100; $l<($num_bases+1); $l++) { #big loop starts


        for (my $i=0; $i<(100); $i++)  #counts Gs and Cs in window Note the number of bases is one larger than the array
                {
                if(($window[$i]=~"G") or ($window[$i]=~"C")) #if it matches G or C increase counter
                        {$num_GC++;}
                }
        $GC_content[$count]=($num_GC);
        # print "$num_GC $bases[$l] ";
        $num_GC=0;
#move window by one to right
        $count++;
        shift @window;
        # print "$test\n";
        push @window, $bases[$l];
        }
print join("   ",@GC_content);
print "\n";
```



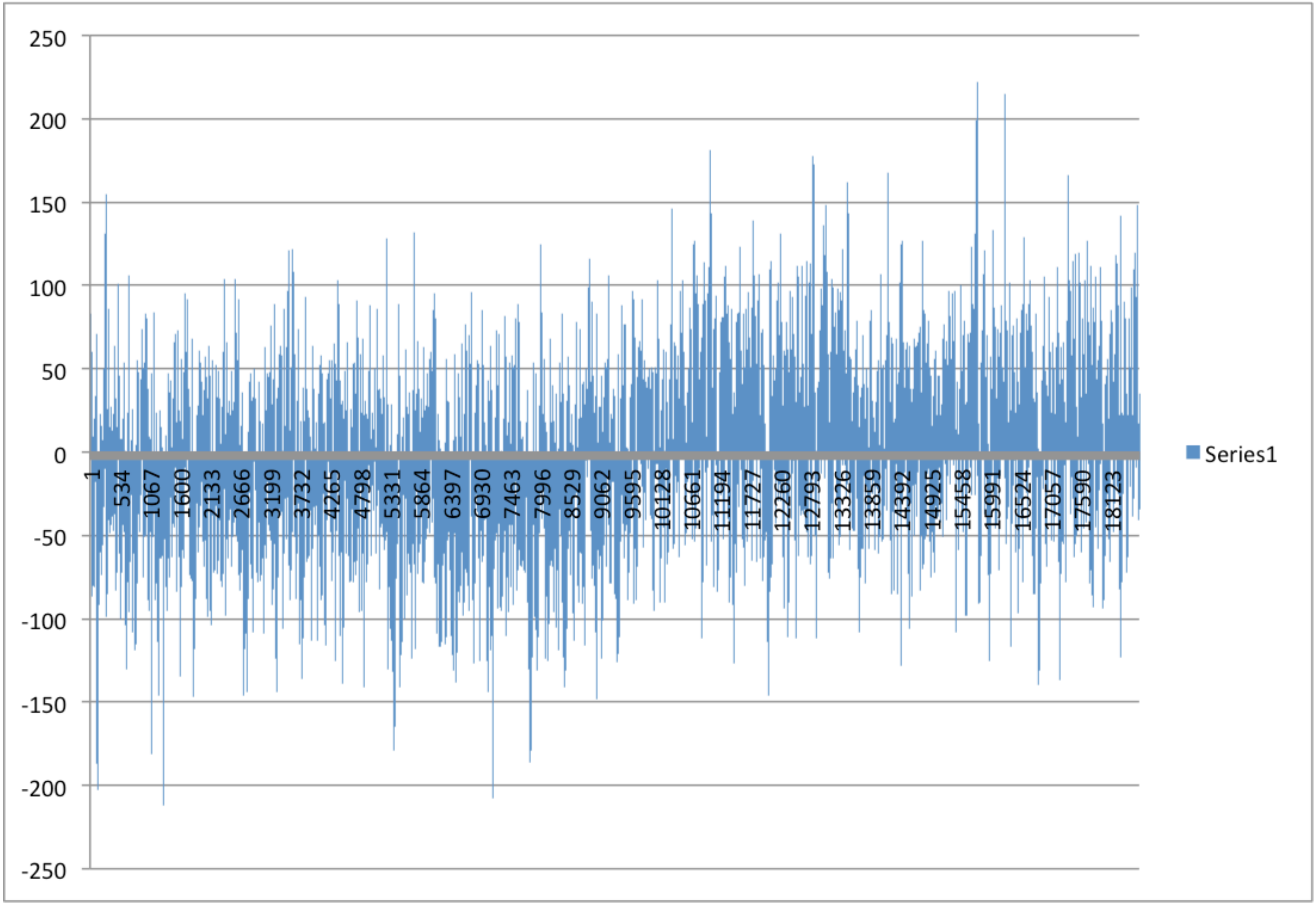
```
for (my$l=100; $l<($num_bases+1); $l++) { #big loop starts


    for (my $i=0; $i<(1000); $i++)  #counts Gs and Cs in window Note the
        {
        if(($window[$i]=~"C")) #if it matches C increase counter
            {$num_CmG++;}
        if(($window[$i]=~"G") ) #if it matches G or C increase counter
            {$num_CmG--;}
        }
    $CmG_content[$count]=($num_CmG);
    print "$num_CmG\t$bases[$l]\t$l\n";
    $num_CmG=0;
#move window by one to right
    $count++;
    shift @window;
    # print "$test\n";
    push @window, $bases[$l];
    }
print join("   ",@CmG_content);
$outfile="test";

open(OUT, "> $outfile") or die "cannot open $outfile: $!";

foreach (@CmG_content) {
    $printcount++;
    if ($printcount == 100)
        {print OUT "$_\n";
        $printcount=0;
    }
}
print OUT "\n";
print "\n";
```

Thermus thermophilus SG0.5JP17-16



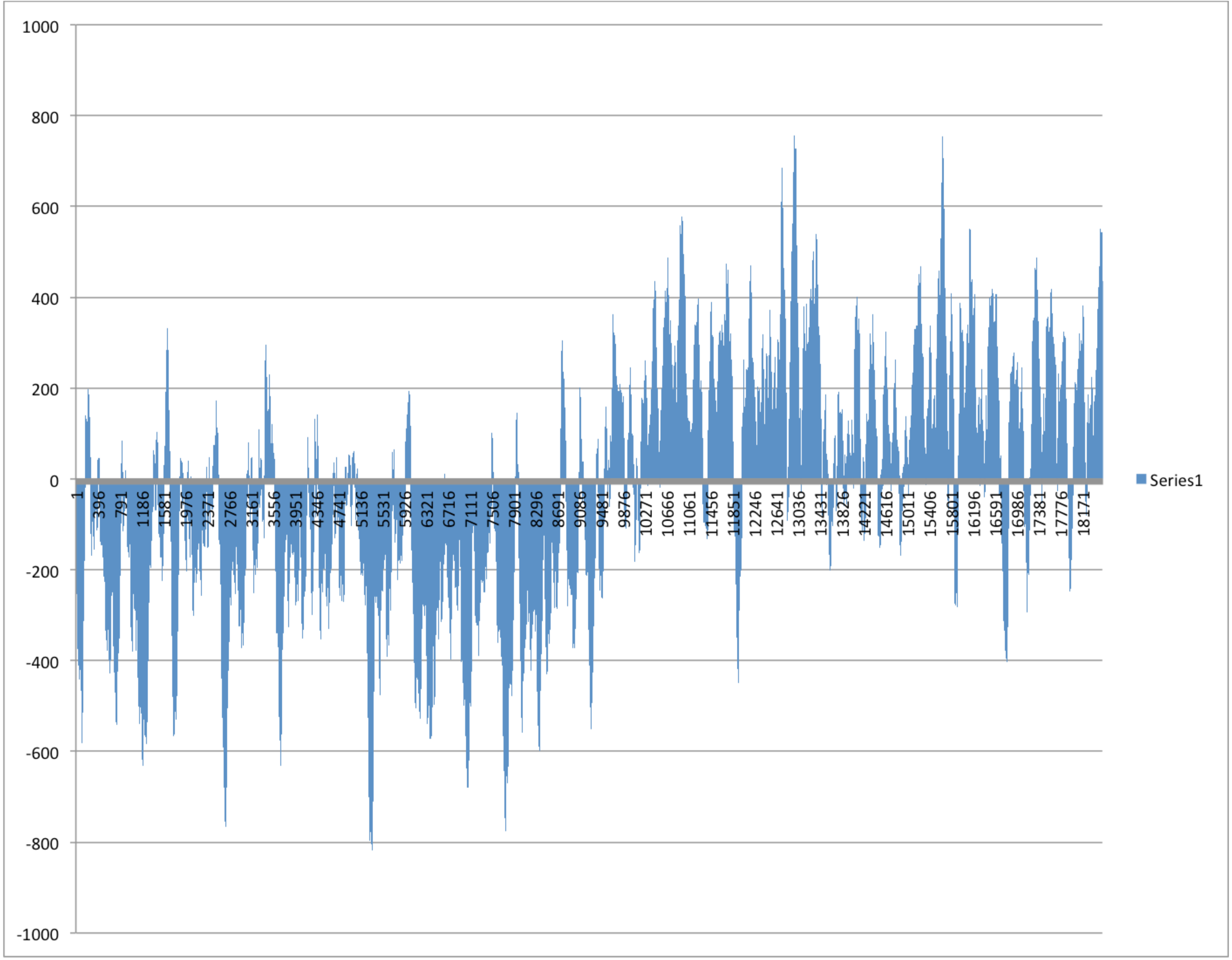
Window=100 , printed every 100

Thermus thermophilus SG0.5JP17-16



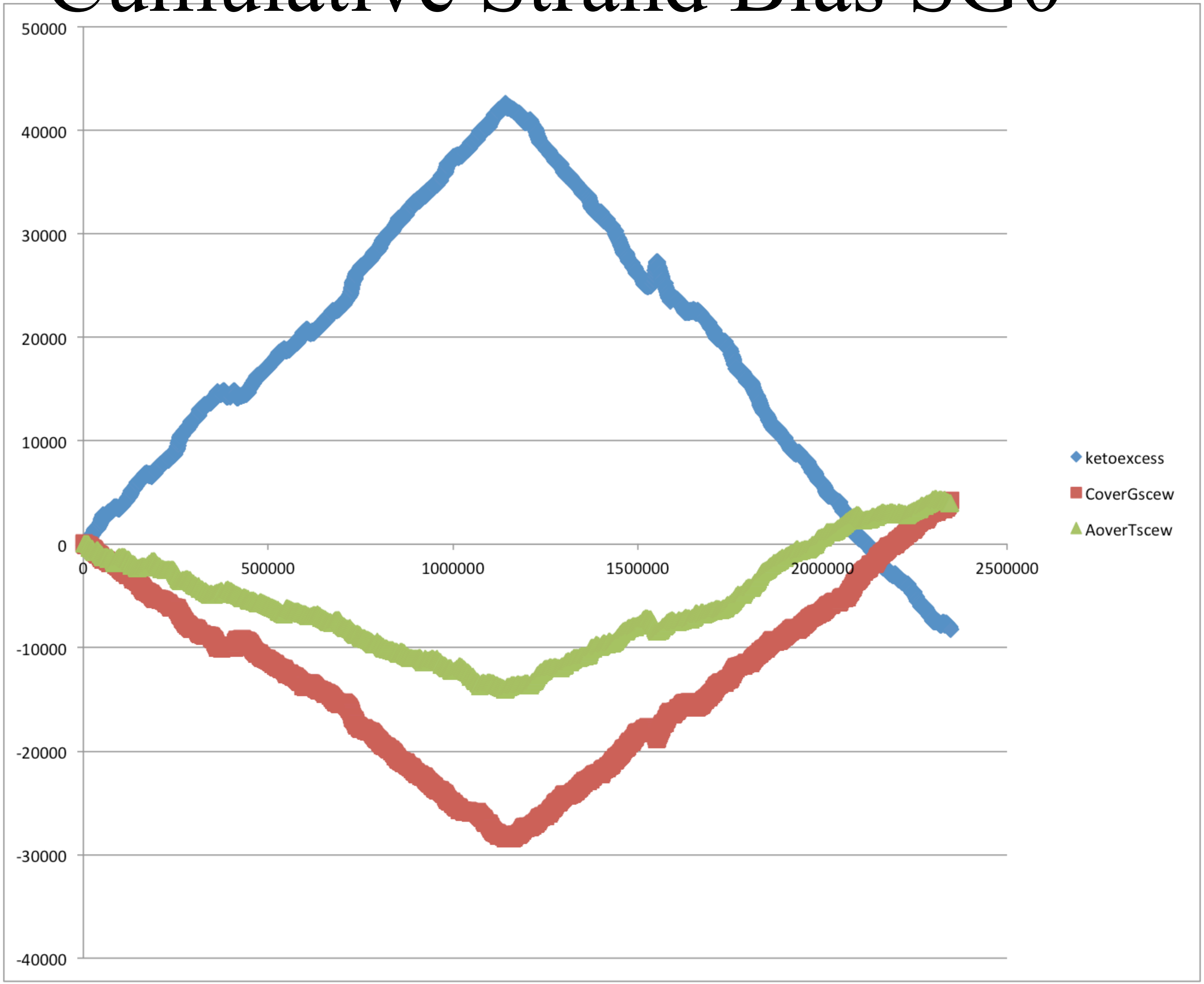
Window=1000 , printed every 100

Thermus thermophilus SG0.5JP17-16



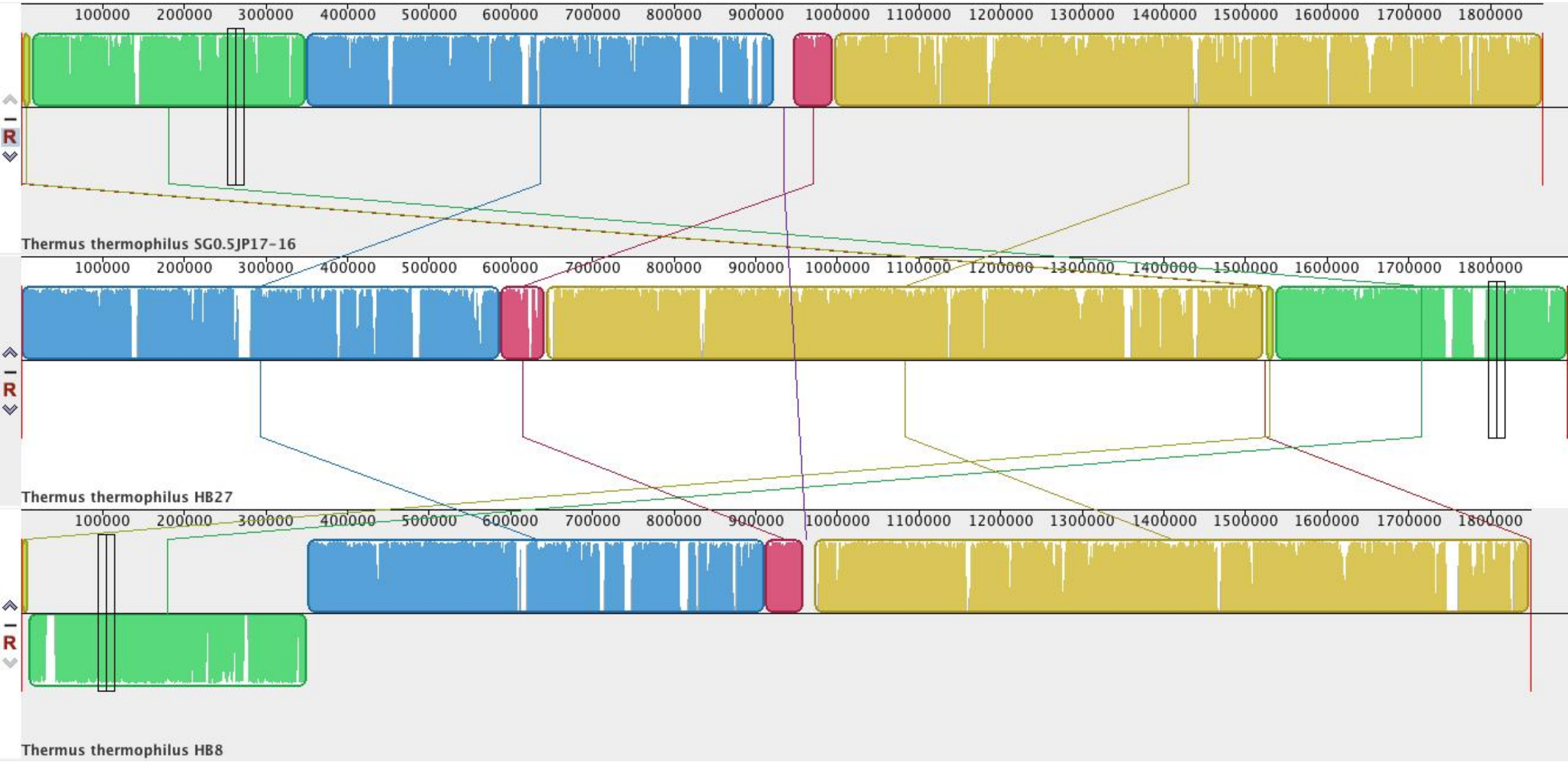
Window=10000 , printed every 100

Cumulative Strand Bias SG0

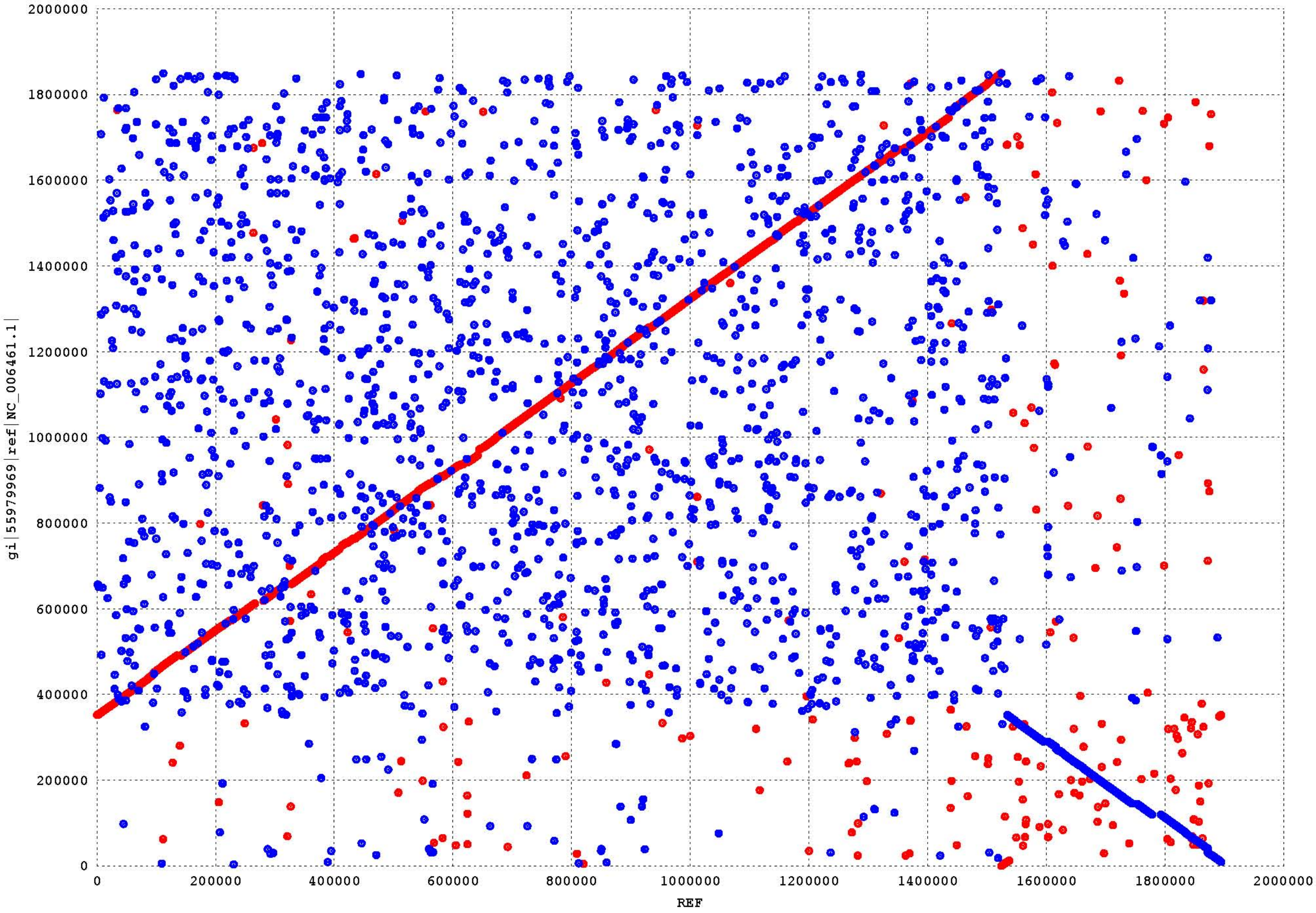


Example on the utility of strand bias

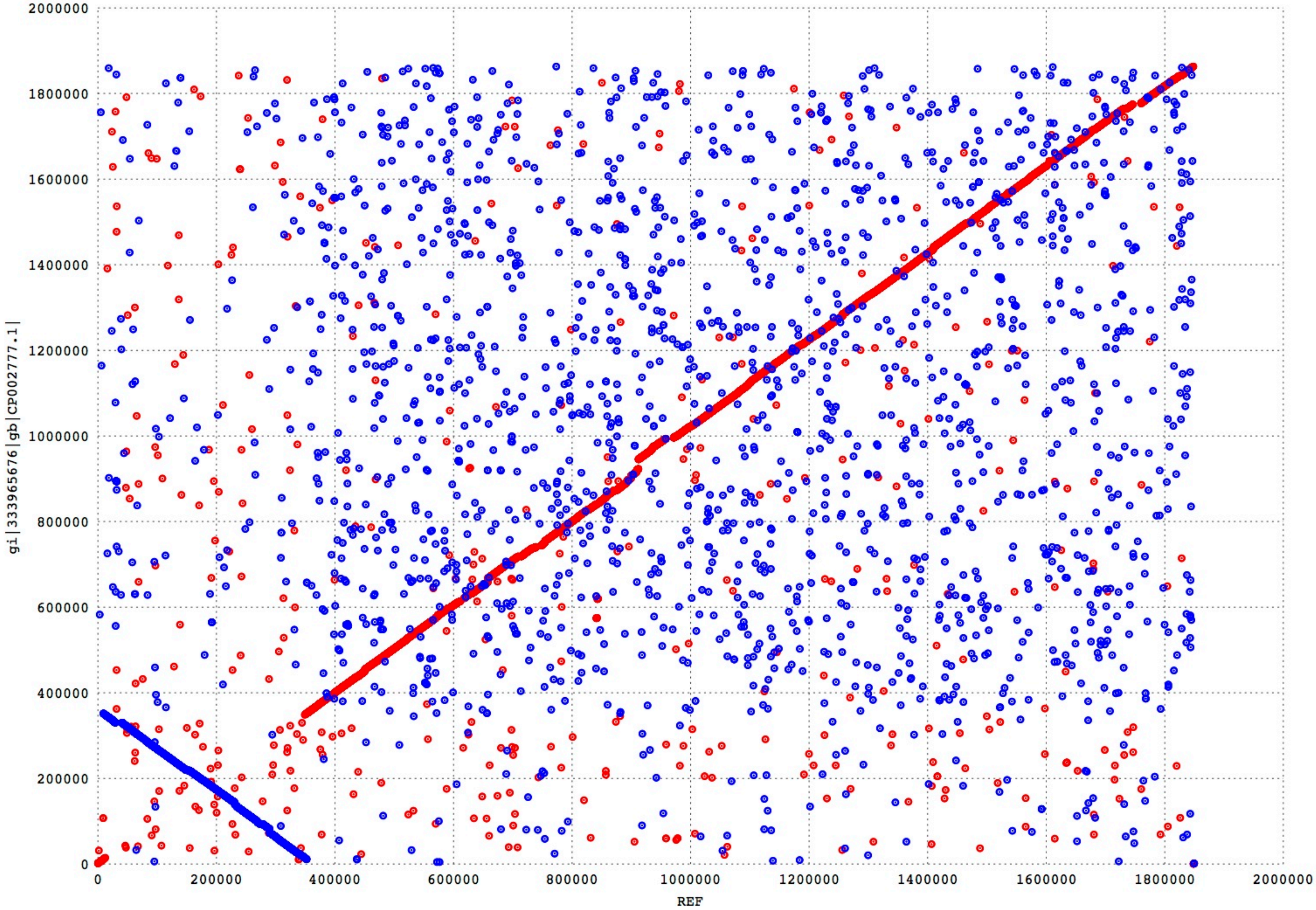
Mauve alignment of three *Thermus thermophilus* genomes:



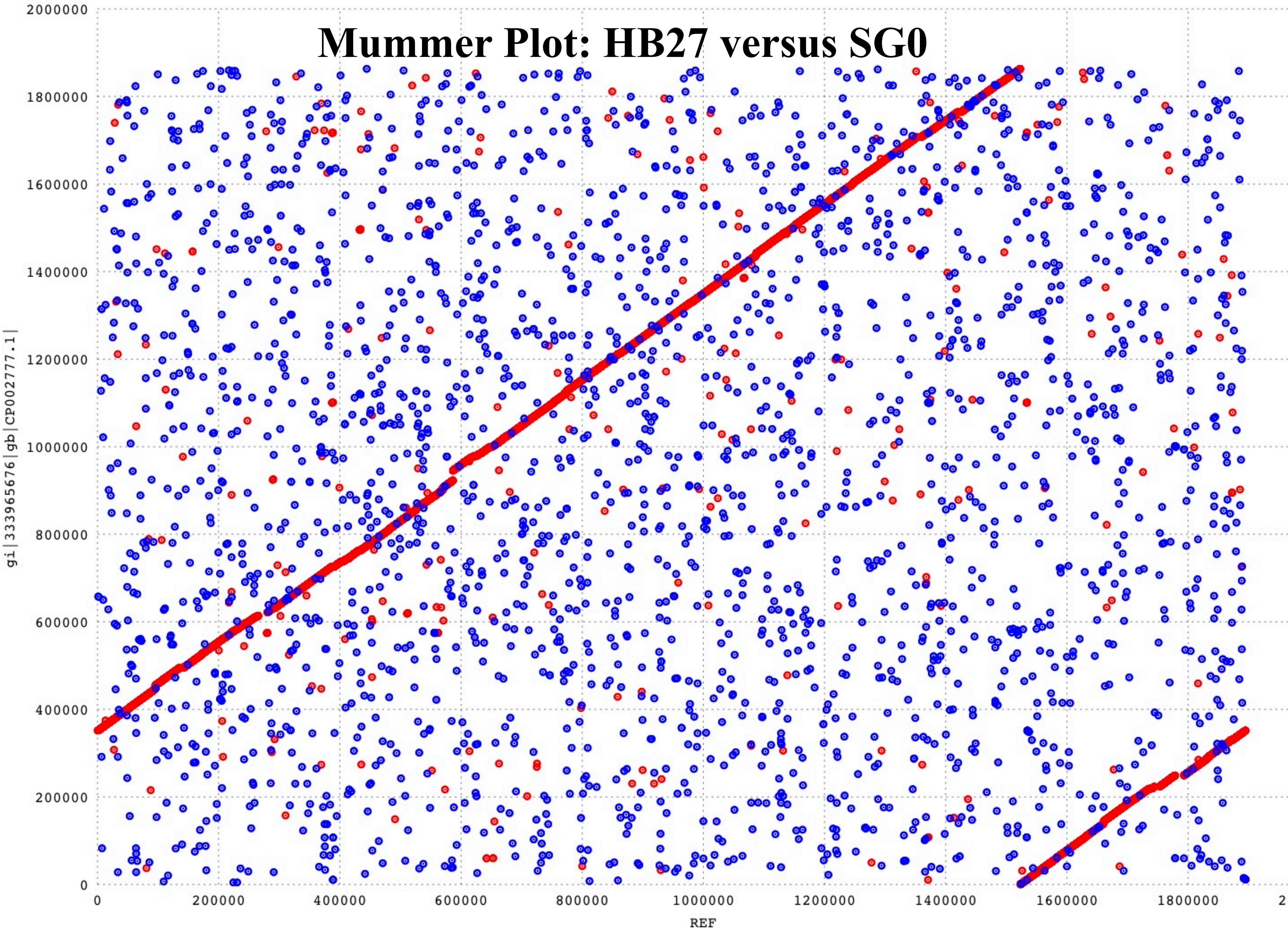
Mummer Plot: HB8 versus HB27



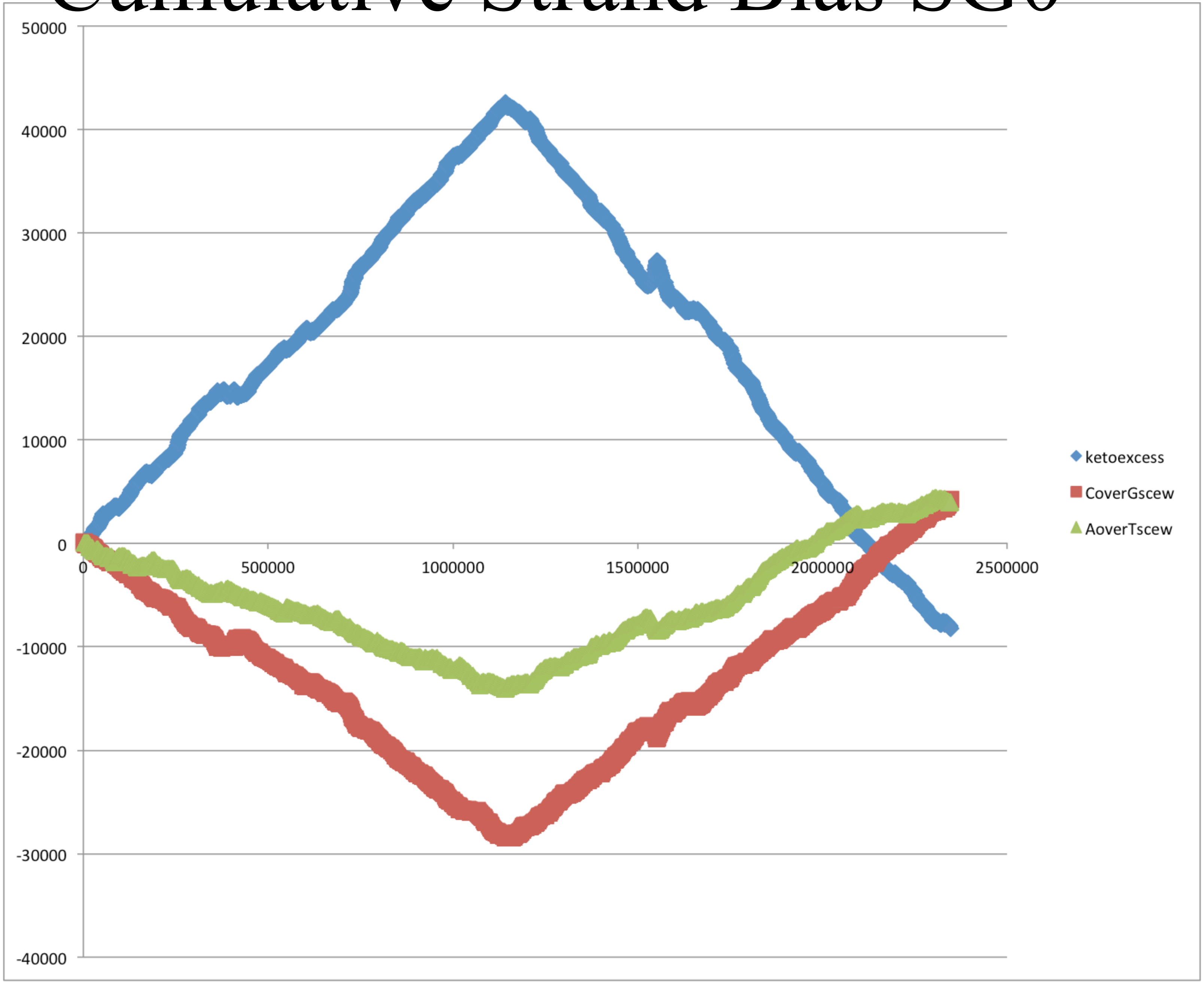
Mummer Plot: HB8 versus SG0



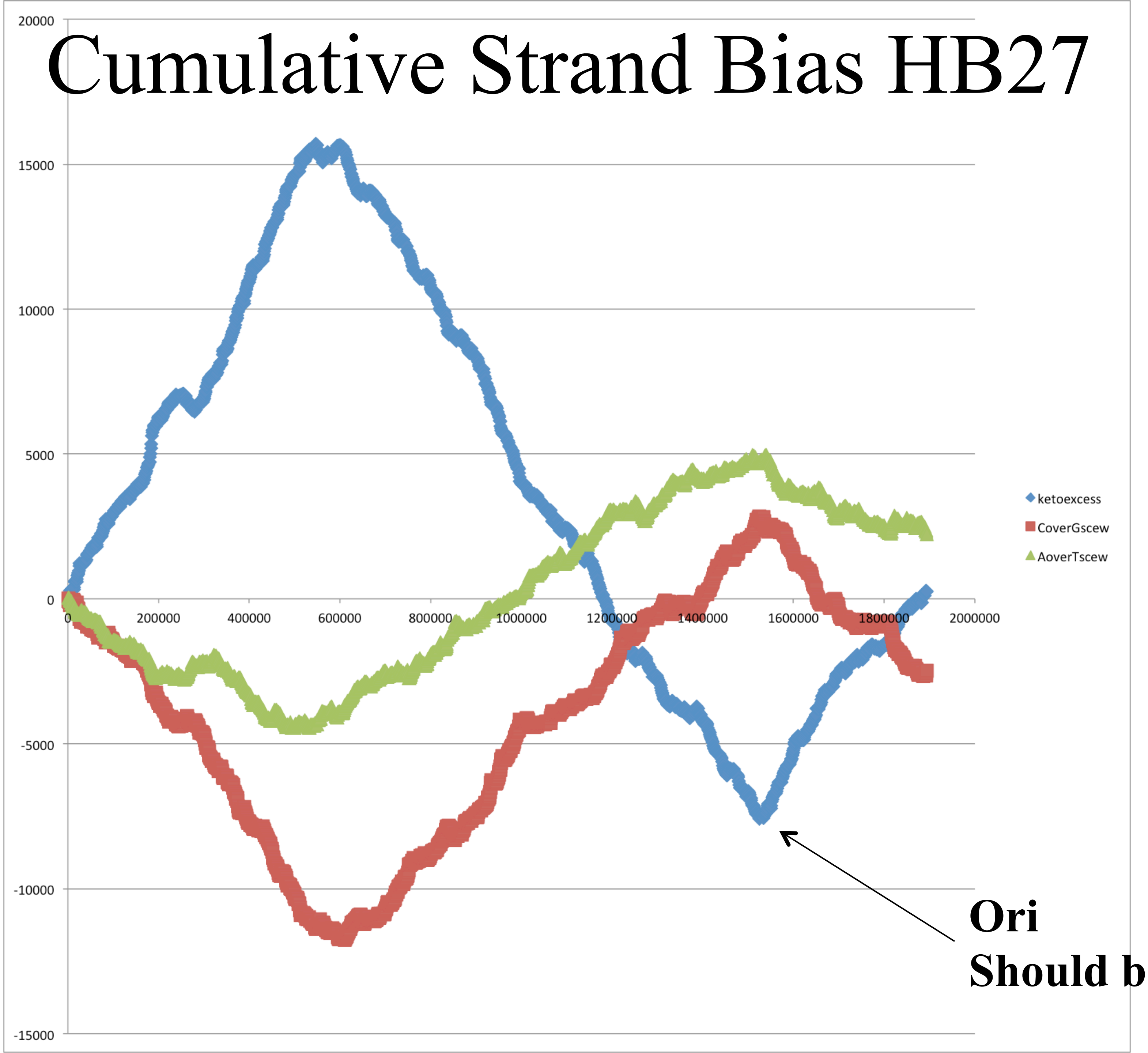
Mummer Plot: HB27 versus SG0



Cumulative Strand Bias SG0

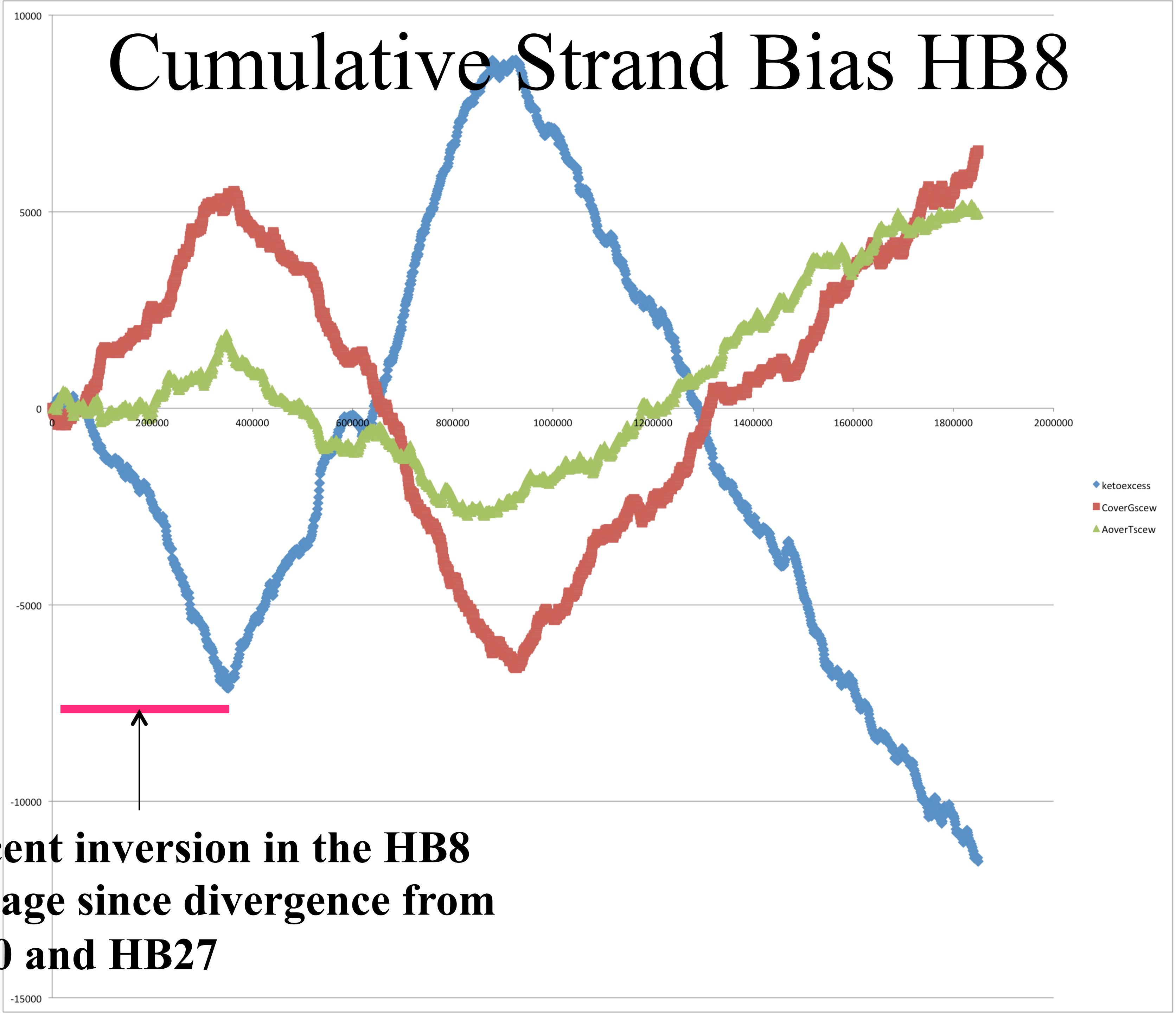


Cumulative Strand Bias HB27



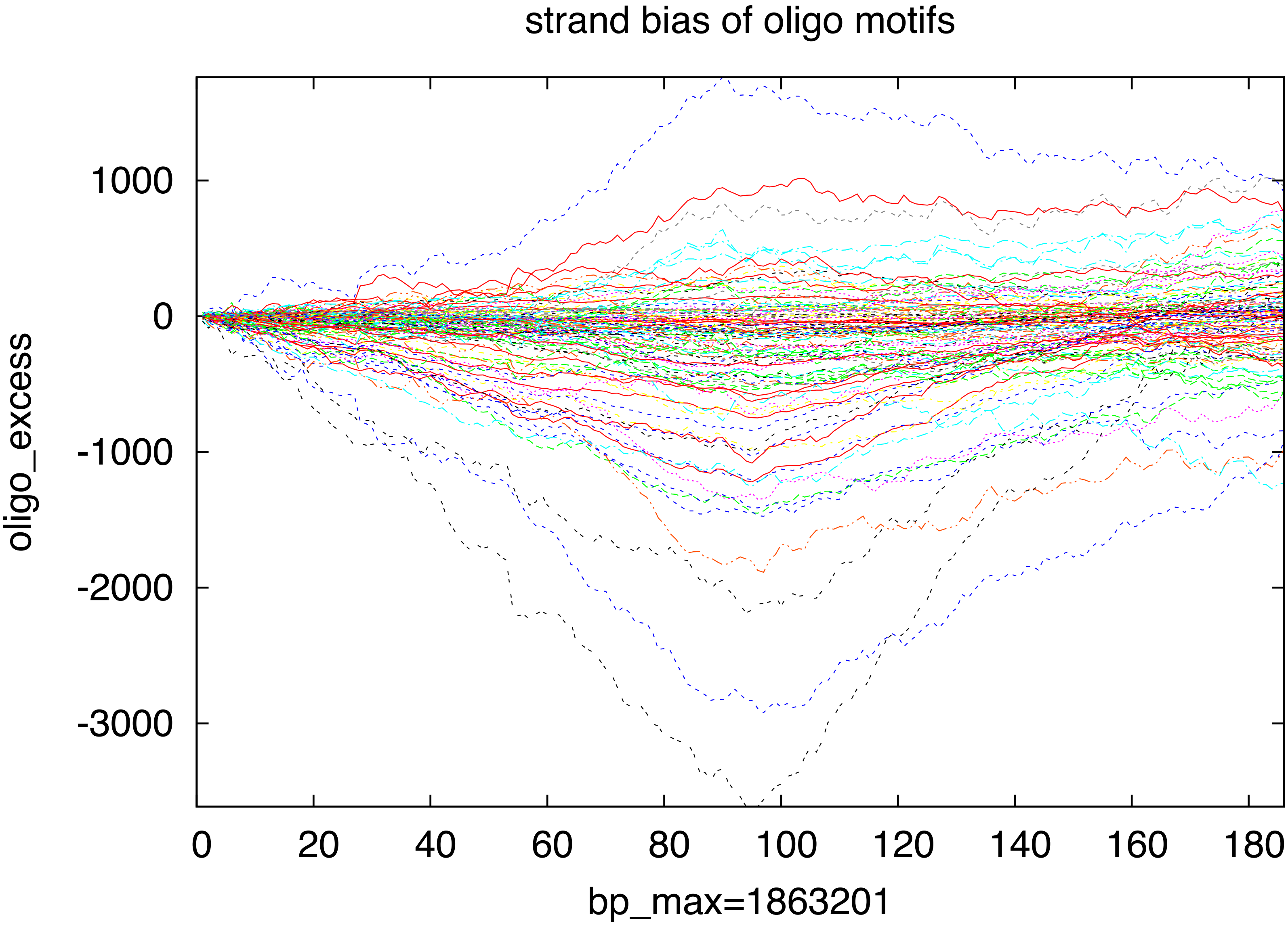
Ori
Should be here

Cumulative Strand Bias HB8

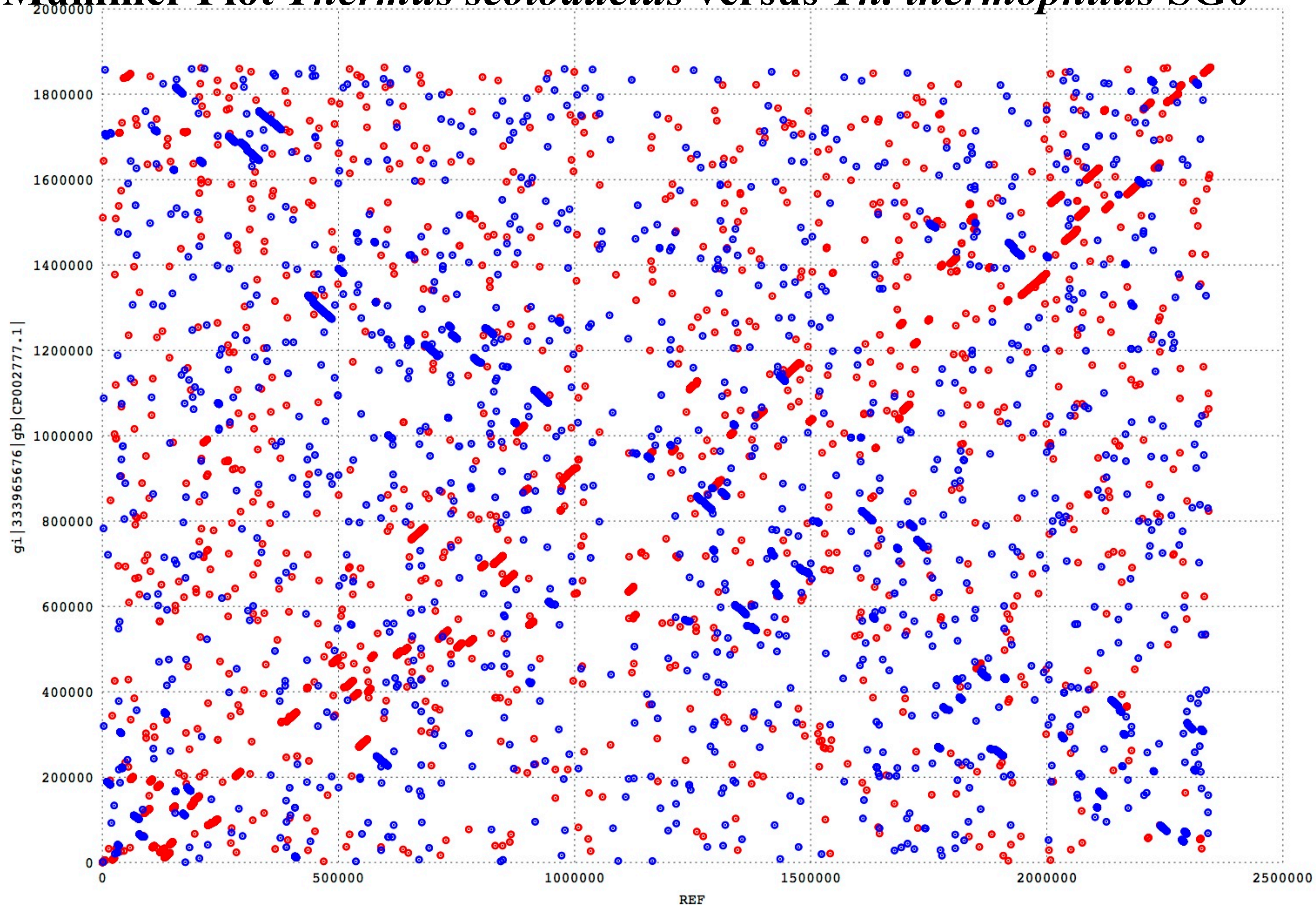


Recent inversion in the HB8 lineage since divergence from SG0 and HB27

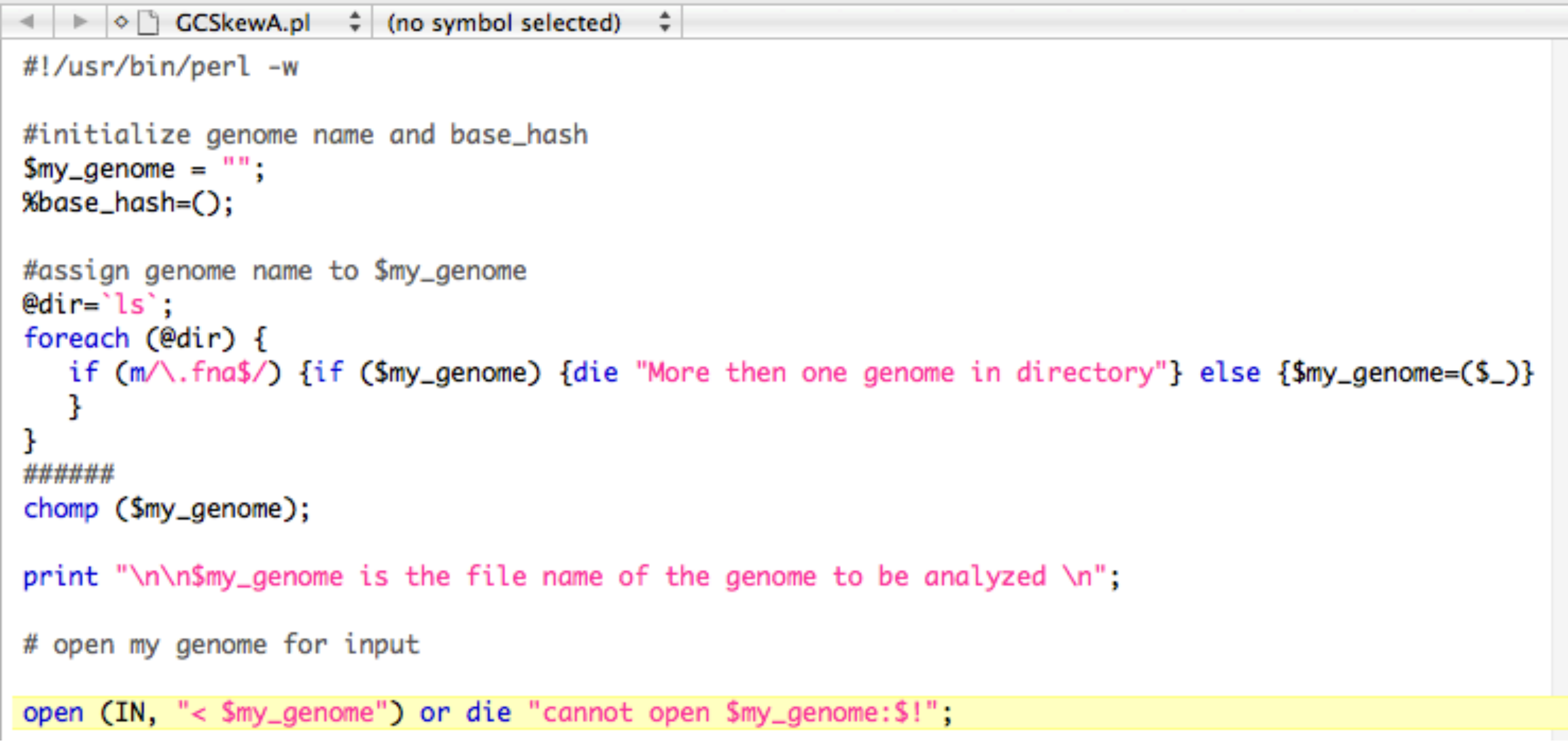
Tetramer bias SG0



Mummer Plot *Thermus scotoductus* versus *Th. thermophilus* SG0



GC Skew

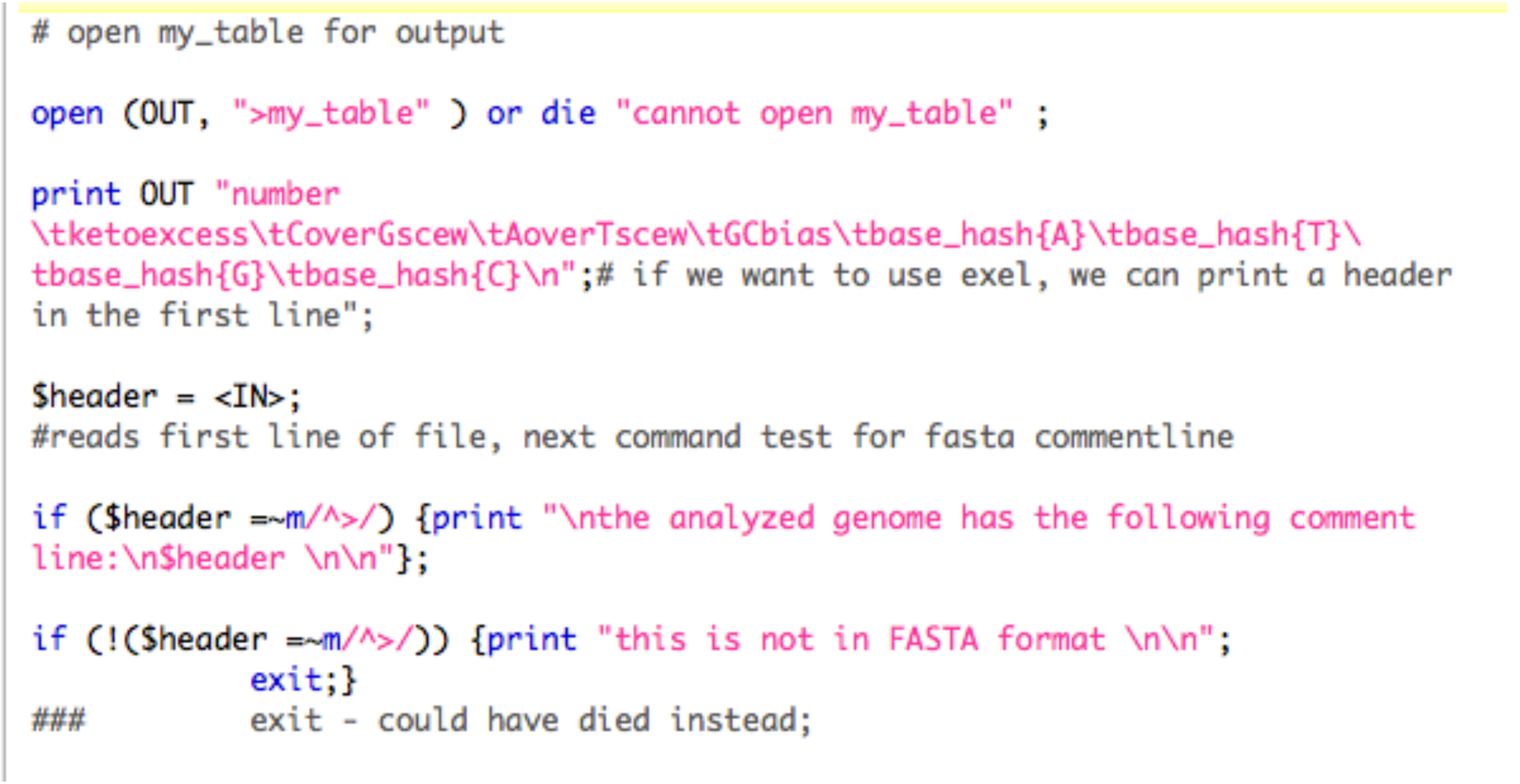


```
#!/usr/bin/perl -w

#initialize genome name and base_hash
$my_genome = "";
%base_hash=();

#assign genome name to $my_genome
@dir=`ls`;
foreach (@dir) {
   if (m/\.fna$/) {if ($my_genome) {die "More then one genome in directory"} else {$my_genome=($_)}
   }
}
######
chomp ($my_genome);

print "\n\n$my_genome is the file name of the genome to be analyzed \n";

# open my genome for input

open (IN, "< $my_genome") or die "cannot open $my_genome:$!";
```

Open input - note, does not use glob

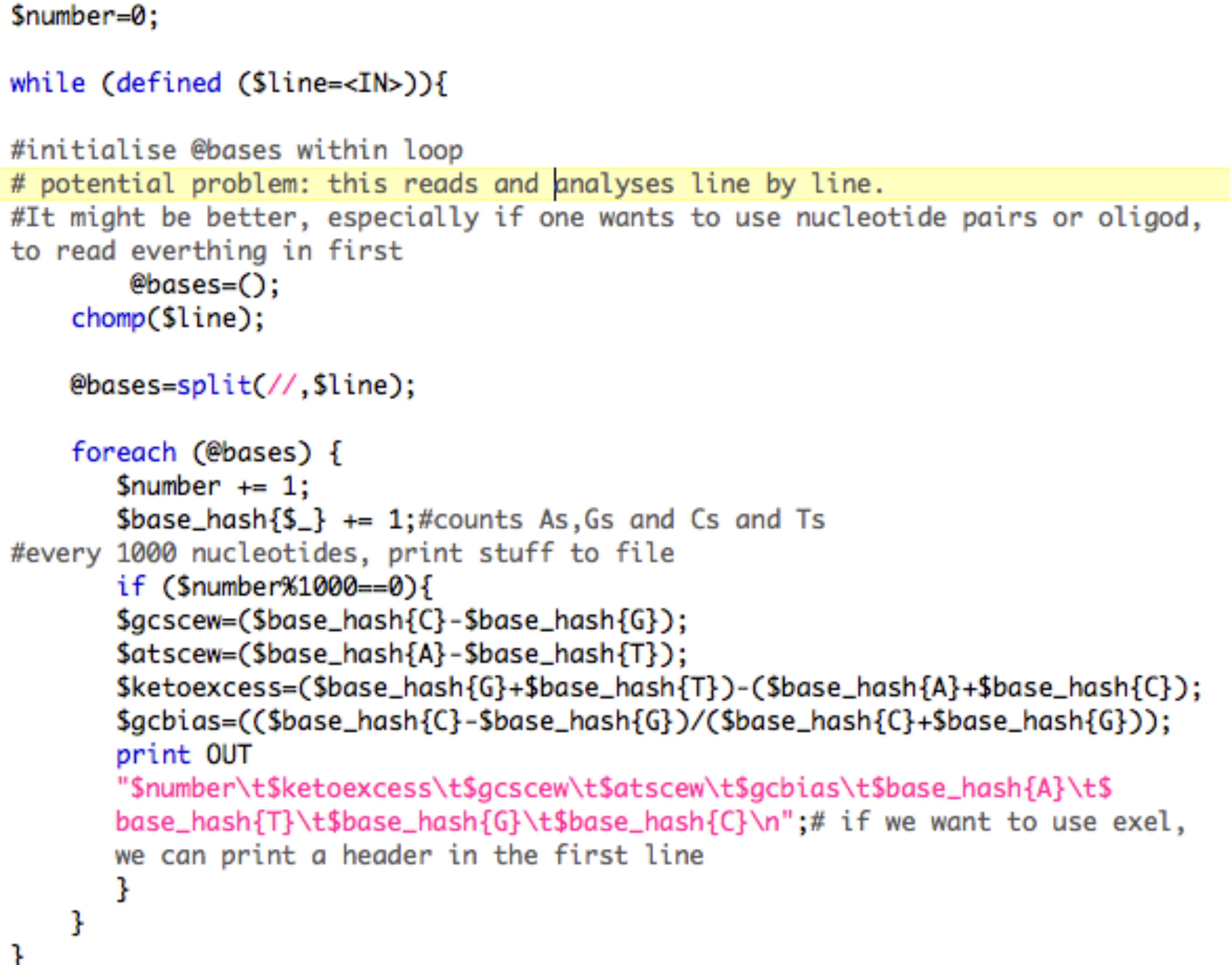
GC Skew - part 2

Start output, read first line of fasta formated genome



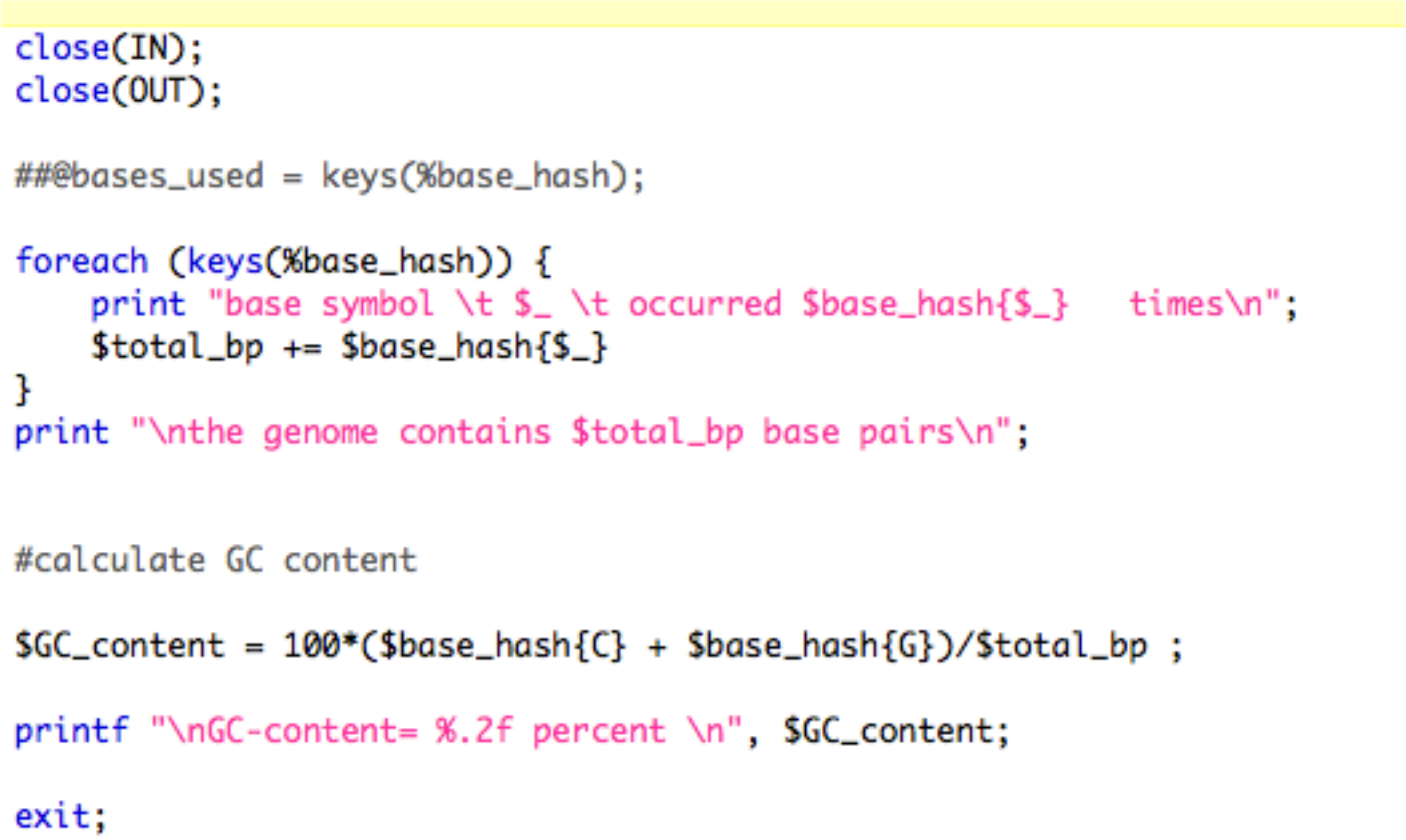
```
# open my_table for output

open (OUT, ">my_table" ) or die "cannot open my_table" ;

print OUT "number
\tketoexcess\tCoverGscew\tAoverTscew\tGCbias\tbase_hash{A}\tbase_hash{T}\
tbase_hash{G}\tbase_hash{C}\n";# if we want to use exel, we can print a header
in the first line";

$header = <IN>;
#reads first line of file, next command test for fasta commentline

if ($header =~m/^>/) {print "\nthe analyzed genome has the following comment
line:\n$header \n\n"};

if (!($header =~m/^>/)) {print "this is not in FASTA format \n\n";
            exit;}
###         exit - could have died instead;
```

GC Skew – part 3



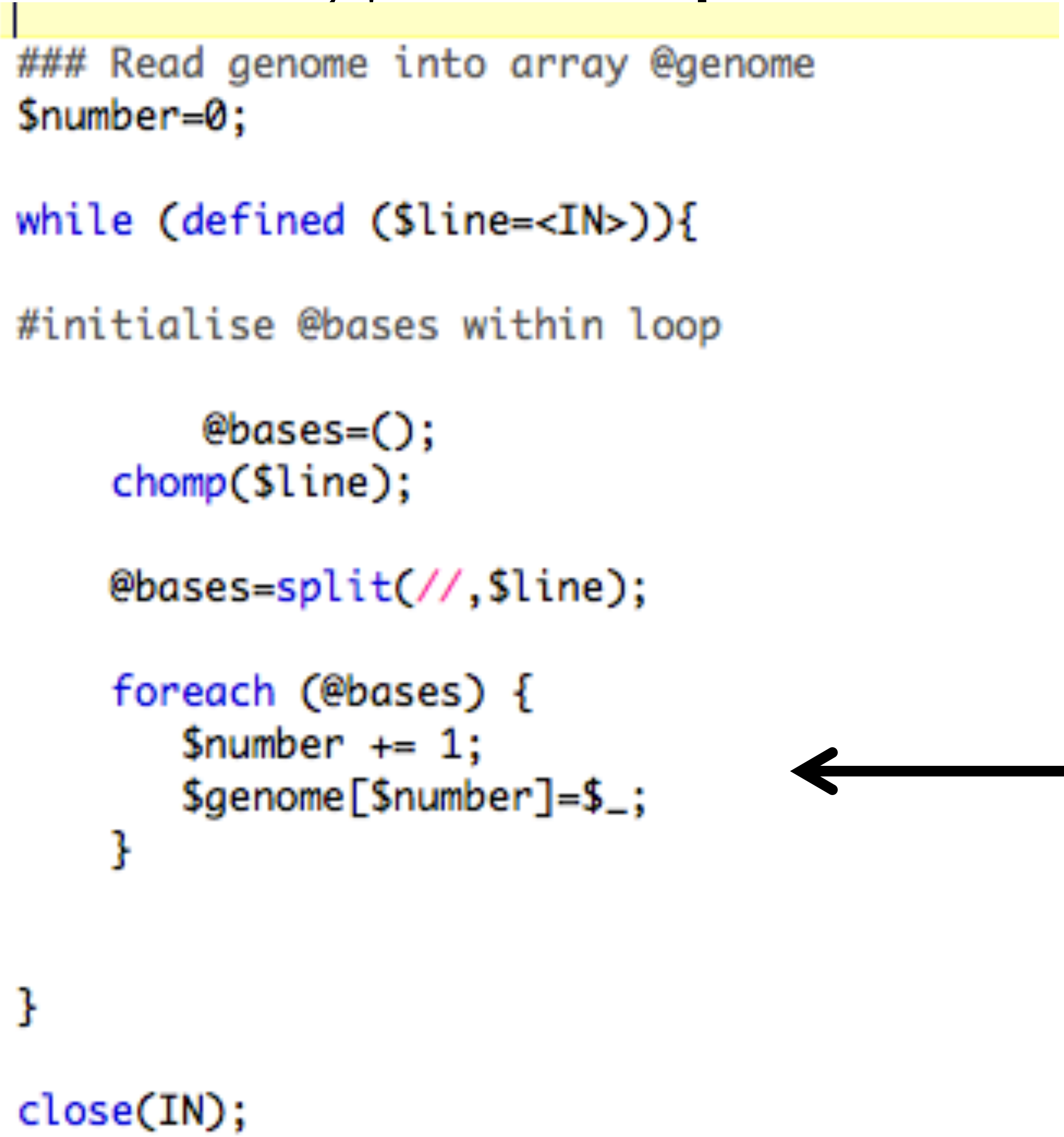
```
$number=0;

while (defined ($line=<IN>)){

#initialise @bases within loop
# potential problem: this reads and analyses line by line.
#It might be better, especially if one wants to use nucleotide pairs or oligod,
to read everthing in first
        @bases=();
    chomp($line);

    @bases=split(//,$line);

    foreach (@bases) {
       $number += 1;
       $base_hash{$_} += 1;#counts As,Gs and Cs and Ts
#every 1000 nucleotides, print stuff to file
       if ($number%1000==0){
       $gcscew=($base_hash{C}-$base_hash{G});
       $atscew=($base_hash{A}-$base_hash{T});
       $ketoexcess=($base_hash{G}+$base_hash{T})-($base_hash{A}+$base_hash{C});
       $gcbias=(($base_hash{C}-$base_hash{G})/($base_hash{C}+$base_hash{G}));
       print OUT
       "$number\t$ketoexcess\t$gcscew\t$atscew\t$gcbias\t$base_hash{A}\t$
       base_hash{T}\t$base_hash{G}\t$base_hash{C}\n";# if we want to use exel,
       we can print a header in the first line
       }
    }
}
```

GC Skew – part 4



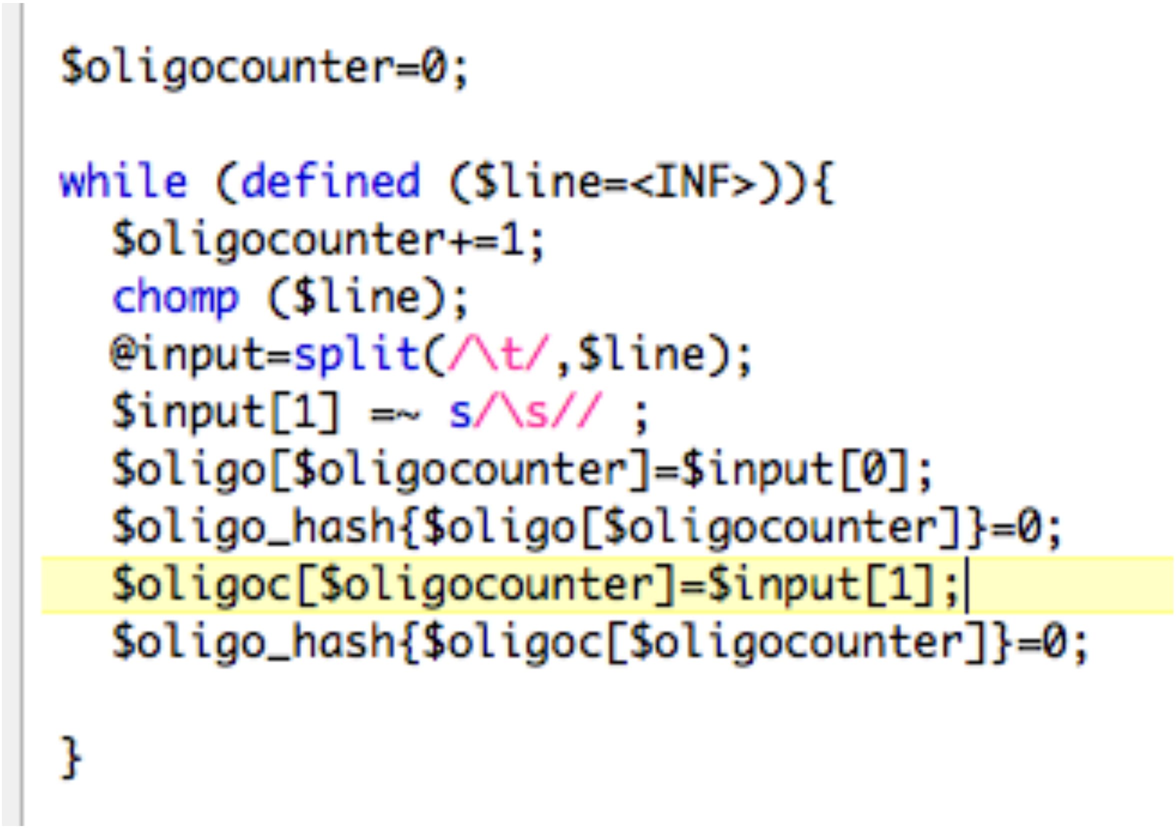
```
close(IN);
close(OUT);

##@bases_used = keys(%base_hash);

foreach (keys(%base_hash)) {
    print "base symbol \t $_ \t occurred $base_hash{$_}   times\n";
    $total_bp += $base_hash{$_}
}
print "\nthe genome contains $total_bp base pairs\n";


#calculate GC content

$GC_content = 100*($base_hash{C} + $base_hash{G})/$total_bp ;

printf "\nGC-content= %.2f percent \n", $GC_content;

exit;
```

Oligo Skew /oligo_scan.pl



```
### Read genome into array @genome
$number=0;

while (defined ($line=<IN>)){

#initialise @bases within loop

        @bases=();
    chomp($line);

    @bases=split(//,$line);

    foreach (@bases) {
       $number += 1;
       $genome[$number]=$_;
    }


}

close(IN);
```

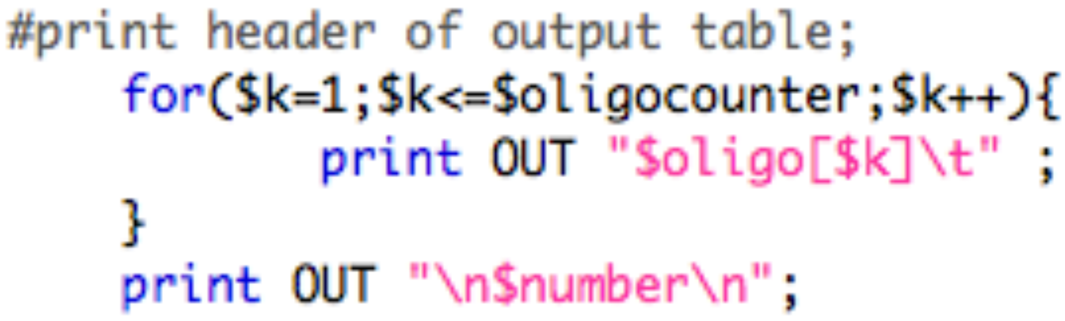
Oligo Skew /oligo_scan.pl – part 2



```
$oligocounter=0;

while (defined ($line=<INF>)){
  $oligocounter+=1;
  chomp ($line);
  @input=split(/\t/,$line);
  $input[1] =~ s/\s// ;
  $oligo[$oligocounter]=$input[0];
  $oligo_hash{$oligo[$oligocounter]}=0;
  $oligoc[$oligocounter]=$input[1];
  $oligo_hash{$oligoc[$oligocounter]}=0;

}
```

Reads in table of oligonucleotides – the table should list oligo and its complement per line seperated by a \t .

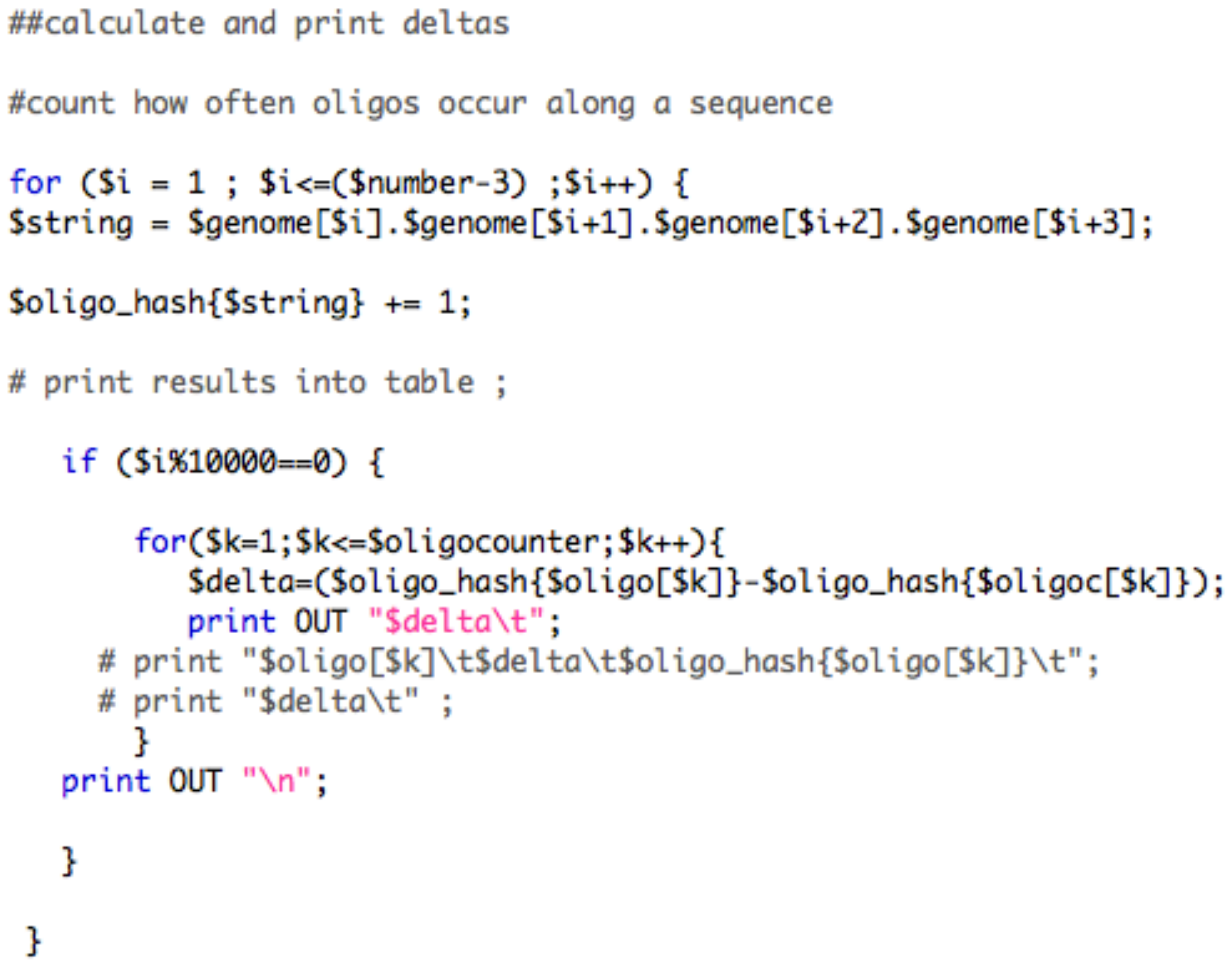
Oligo Skew /oligo_scan.pl – part 3



```
#print header of output table;
    for($k=1;$k<=$oligocounter;$k++){
            print OUT "$oligo[$k]\t" ;
    }
    print OUT "\n$number\n";
```

Writes header for output file

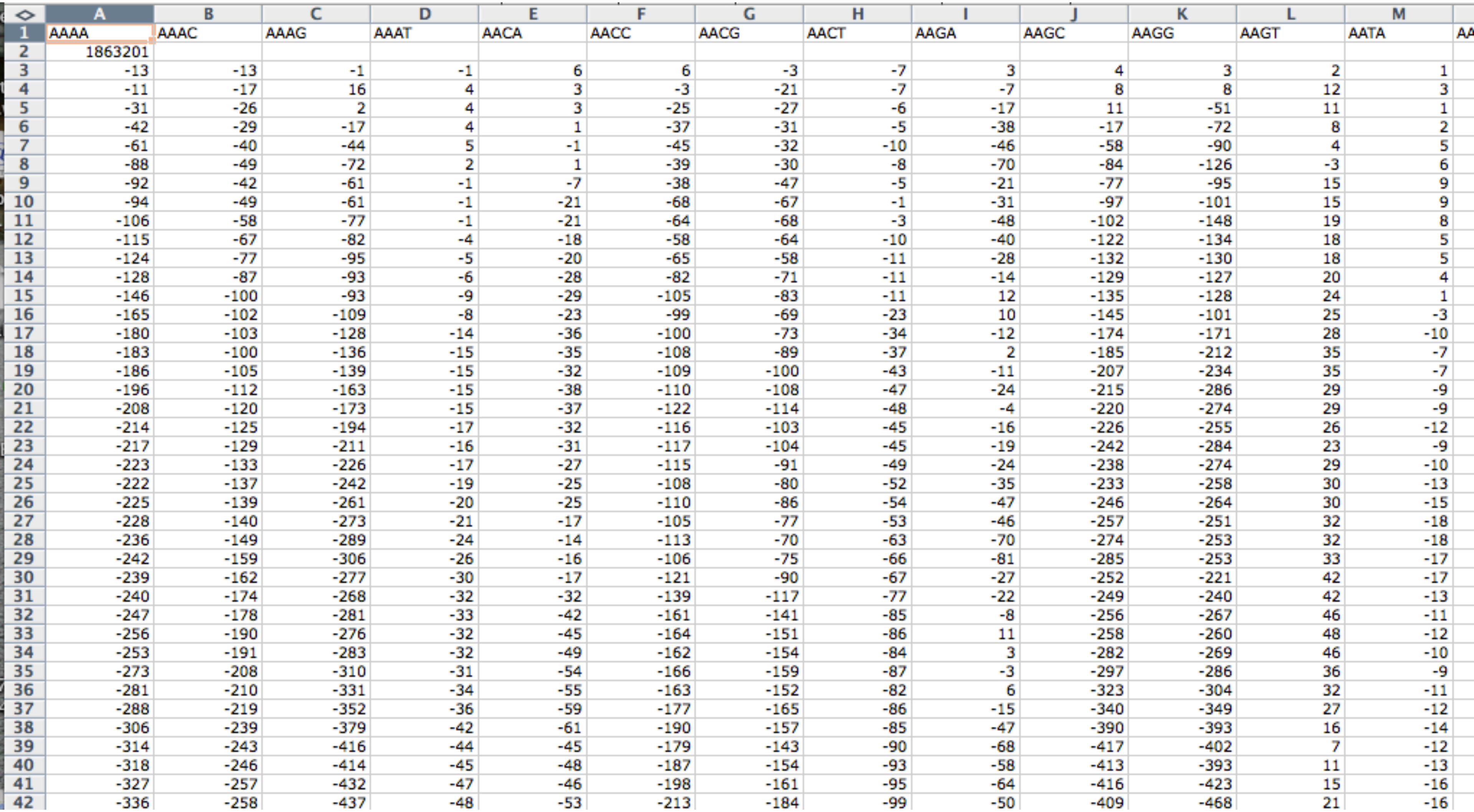
Oligo Skew /oligo_scan.pl – part 4



```
##calculate and print deltas

#count how often oligos occur along a sequence

for ($i = 1 ; $i<=($number-3) ;$i++) {
$string = $genome[$i].$genome[$i+1].$genome[$i+2].$genome[$i+3];

$oligo_hash{$string} += 1;

# print results into table ;

   if ($i%10000==0) {

       for($k=1;$k<=$oligocounter;$k++){
          $delta=($oligo_hash{$oligo[$k]}-$oligo_hash{$oligoc[$k]});
          print OUT "$delta\t";
     # print "$oligo[$k]\t$delta\t$oligo_hash{$oligo[$k]}\t";
     # print "$delta\t" ;
       }
   print OUT "\n";

   }

 }
```

Calculates oligo bias and writes table

Oligo Skew /oligo_scan.pl – part 5

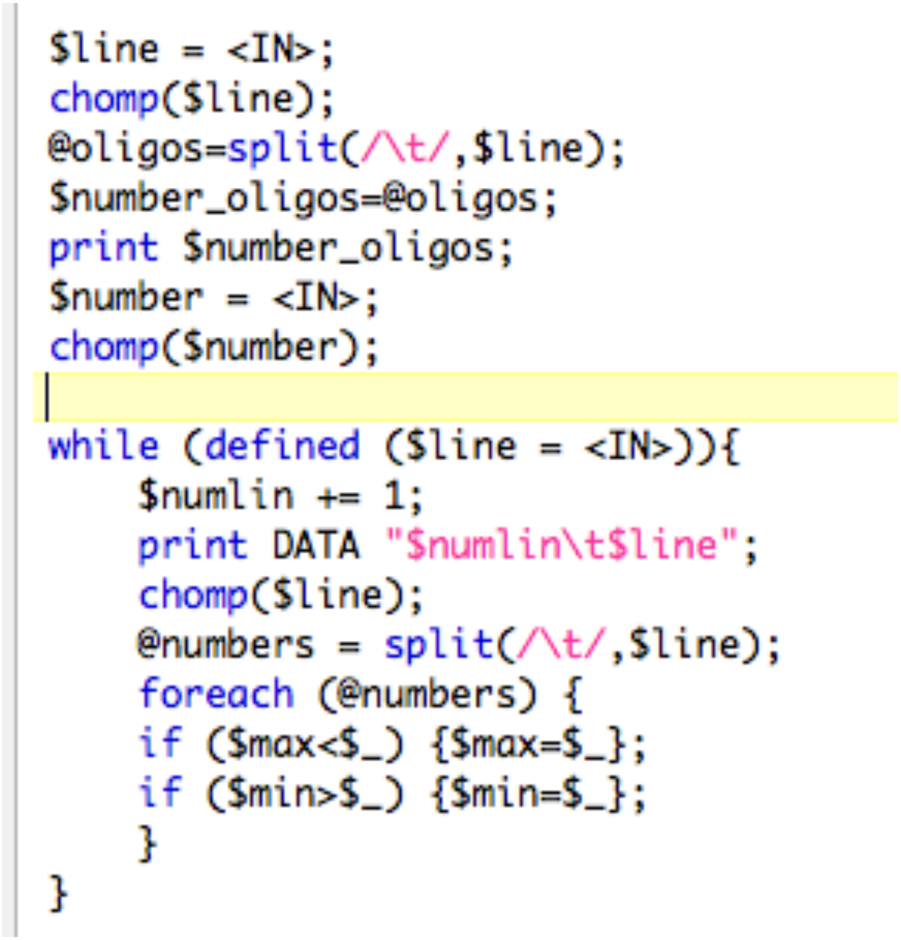


◇	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	AAAA	AAAC	AAAG	AAAT	AACA	AACC	AACG	AACT	AAGA	AAGC	AAGG	AAGT	AATA	AA
2	1863201													
3	-13	-13	-1	-1	6	6	-3	-7	3	4	3	2	1	
4	-11	-17	16	4	3	-3	-21	-7	-7	8	8	12	3	
5	-31	-26	2	4	3	-25	-27	-6	-17	11	-51	11	1	
6	-42	-29	-17	4	1	-37	-31	-5	-38	-17	-72	8	2	
7	-61	-40	-44	5	-1	-45	-32	-10	-46	-58	-90	4	5	
8	-88	-49	-72	2	1	-39	-30	-8	-70	-84	-126	-3	6	
9	-92	-42	-61	-1	-7	-38	-47	-5	-21	-77	-95	15	9	
10	-94	-49	-61	-1	-21	-68	-67	-1	-31	-97	-101	15	9	
11	-106	-58	-77	-1	-21	-64	-68	-3	-48	-102	-148	19	8	
12	-115	-67	-82	-4	-18	-58	-64	-10	-40	-122	-134	18	5	
13	-124	-77	-95	-5	-20	-65	-58	-11	-28	-132	-130	18	5	
14	-128	-87	-93	-6	-28	-82	-71	-11	-14	-129	-127	20	4	
15	-146	-100	-93	-9	-29	-105	-83	-11	12	-135	-128	24	1	
16	-165	-102	-109	-8	-23	-99	-69	-23	10	-145	-101	25	-3	
17	-180	-103	-128	-14	-36	-100	-73	-34	-12	-174	-171	28	-10	
18	-183	-100	-136	-15	-35	-108	-89	-37	2	-185	-212	35	-7	
19	-186	-105	-139	-15	-32	-109	-100	-43	-11	-207	-234	35	-7	
20	-196	-112	-163	-15	-38	-110	-108	-47	-24	-215	-286	29	-9	
21	-208	-120	-173	-15	-37	-122	-114	-48	-4	-220	-274	29	-9	
22	-214	-125	-194	-17	-32	-116	-103	-45	-16	-226	-255	26	-12	
23	-217	-129	-211	-16	-31	-117	-104	-45	-19	-242	-284	23	-9	
24	-223	-133	-226	-17	-27	-115	-91	-49	-24	-238	-274	29	-10	
25	-222	-137	-242	-19	-25	-108	-80	-52	-35	-233	-258	30	-13	
26	-225	-139	-261	-20	-25	-110	-86	-54	-47	-246	-264	30	-15	
27	-228	-140	-273	-21	-17	-105	-77	-53	-46	-257	-251	32	-18	
28	-236	-149	-289	-24	-14	-113	-70	-63	-70	-274	-253	32	-18	
29	-242	-159	-306	-26	-16	-106	-75	-66	-81	-285	-253	33	-17	
30	-239	-162	-277	-30	-17	-121	-90	-67	-27	-252	-221	42	-17	
31	-240	-174	-268	-32	-32	-139	-117	-77	-22	-249	-240	42	-13	
32	-247	-178	-281	-33	-42	-161	-141	-85	-8	-256	-267	46	-11	
33	-256	-190	-276	-32	-45	-164	-151	-86	11	-258	-260	48	-12	
34	-253	-191	-283	-32	-49	-162	-154	-84	3	-282	-269	46	-10	
35	-273	-208	-310	-31	-54	-166	-159	-87	-3	-297	-286	36	-9	
36	-281	-210	-331	-34	-55	-163	-152	-82	6	-323	-304	32	-11	
37	-288	-219	-352	-36	-59	-177	-165	-86	-15	-340	-349	27	-12	
38	-306	-239	-379	-42	-61	-190	-157	-85	-47	-390	-393	16	-14	
39	-314	-243	-416	-44	-45	-179	-143	-90	-68	-417	-402	7	-12	
40	-318	-246	-414	-45	-48	-187	-154	-93	-58	-413	-393	11	-13	
41	-327	-257	-432	-47	-46	-198	-161	-95	-64	-416	-423	15	-16	
42	-336	-258	-437	-48	-53	-213	-184	-99	-50	-409	-468	21	-16	

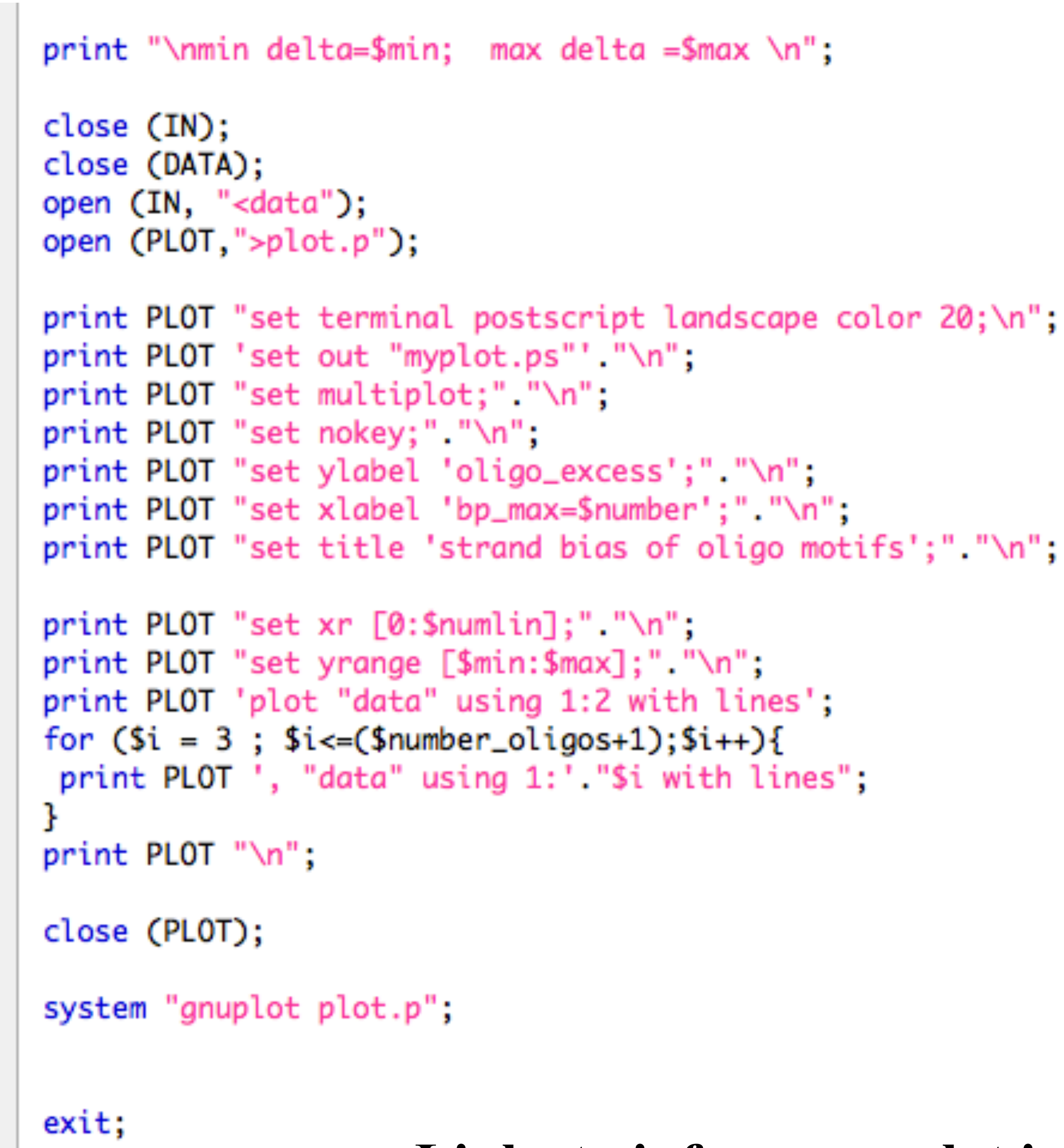
Output table

Oligo Skew /oligo_scan.pl – part 6

Plotscan.pl takes the table, creates a file called data (no header), and creates a file that can be used by gnuplot to create an image.



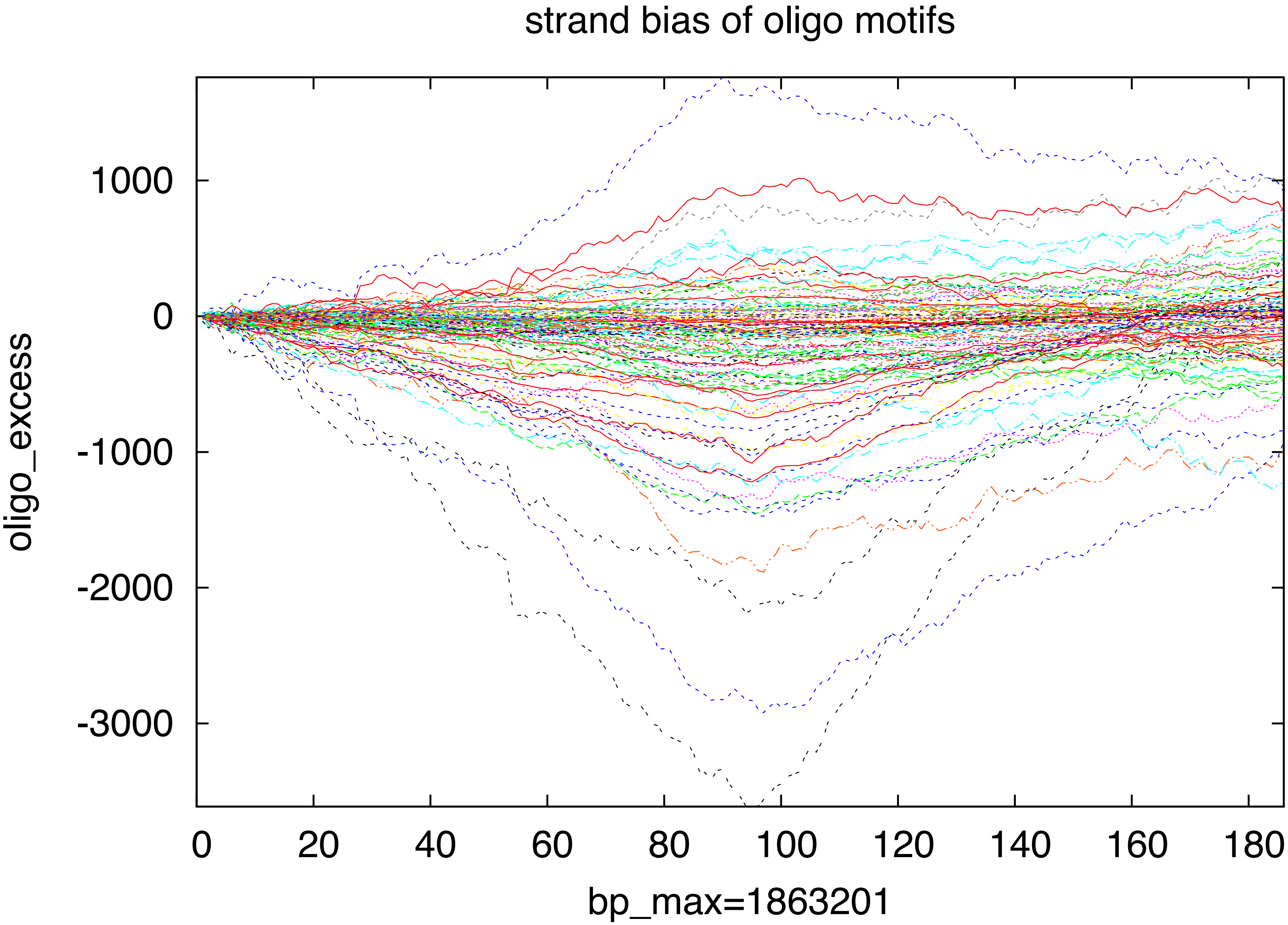
```
$line = <IN>;
chomp($line);
@oligos=split(/\t/,$line);
$number_oligos=@oligos;
print $number_oligos;
$number = <IN>;
chomp($number);

while (defined ($line = <IN>)){
     $numlin += 1;
     print DATA "$numlin\t$line";
     chomp($line);
     @numbers = split(/\t/,$line);
     foreach (@numbers) {
     if ($max<$_) {$max=$_};
     if ($min>$_) {$min=$_};
     }
}
```



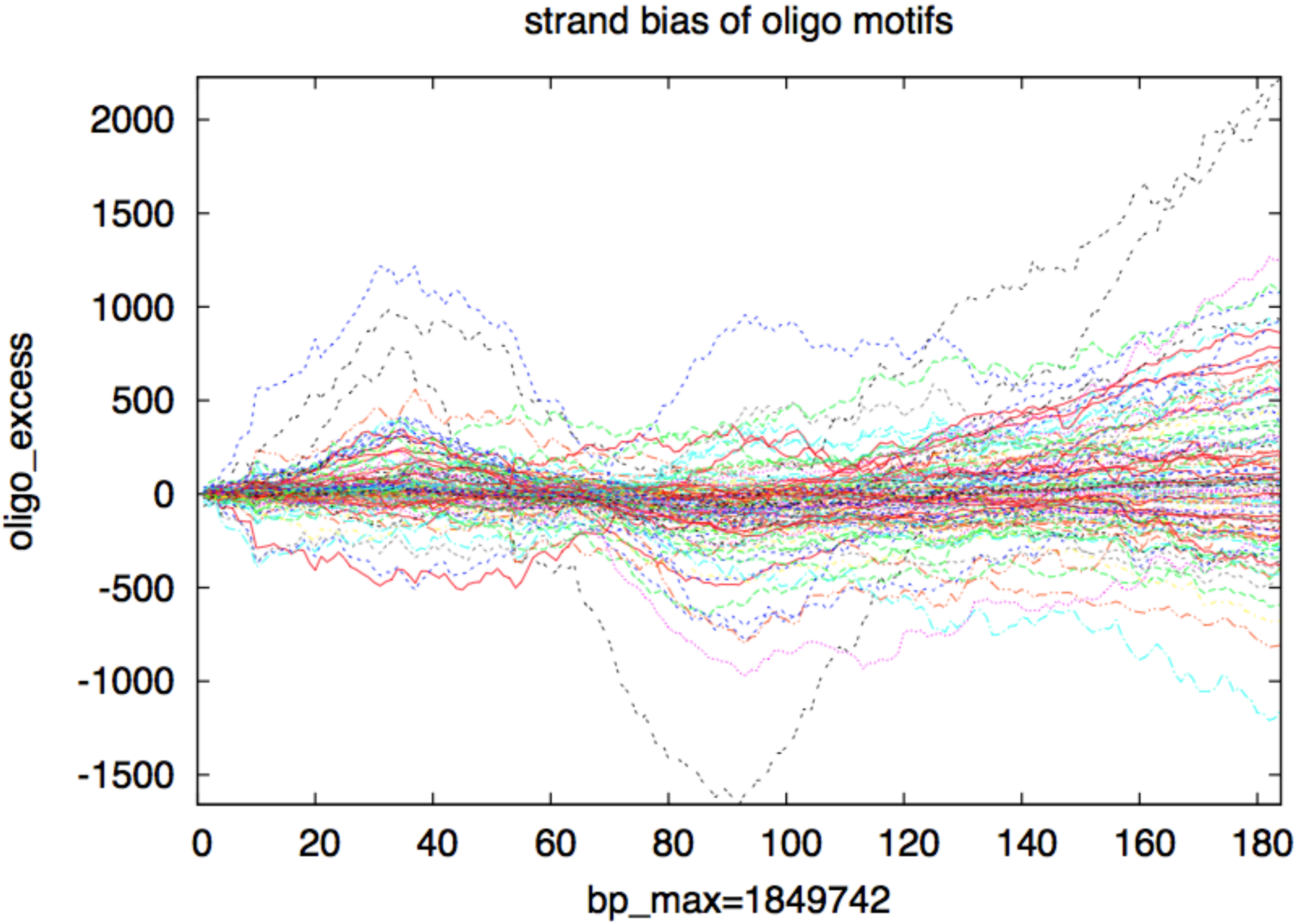
```
print "\nmin delta=$min;  max delta =$max \n";

close (IN);
close (DATA);
open (IN, "<data");
open (PLOT,">plot.p");

print PLOT "set terminal postscript landscape color 20;\n";
print PLOT 'set out "myplot.ps"'."\n";
print PLOT "set multiplot;"."\n";
print PLOT "set nokey;"."\n";
print PLOT "set ylabel 'oligo_excess';"."\n";
print PLOT "set xlabel 'bp_max=$number';"."\n";
print PLOT "set title 'strand bias of oligo motifs';"."\n";

print PLOT "set xr [0:$numlin];"."\n";
print PLOT "set yrange [$min:$max];"."\n";
print PLOT 'plot "data" using 1:2 with lines';
for ($i = 3 ; $i<=($number_oligos+1);$i++){
 print PLOT ', "data" using 1:'."$i with lines";
}
print PLOT "\n";

close (PLOT);

system "gnuplot plot.p";


exit;
```

Links to info on gnuplot is here

Tetramer bias SG0



Tetramer bias HB8



Phylip **written and distributed by Joe Felsenstein and collaborators (some of the following is copied from the PHYLIP homepage)**

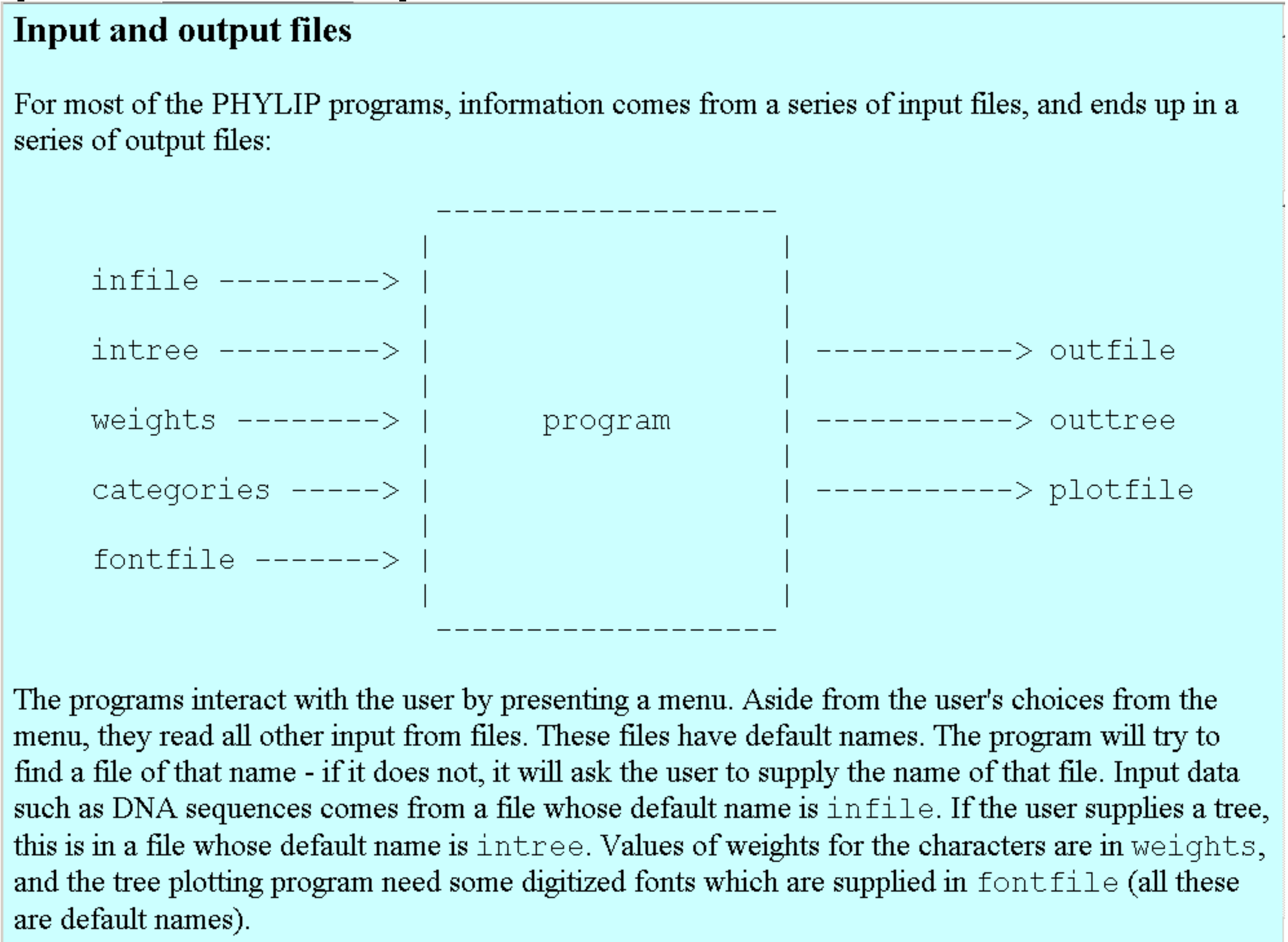
PHYLIP (the *PHYL*ogeny *I*nference *P*ackage) is a package of programs for inferring phylogenies (evolutionary trees).

PHYLIP is the most widely-distributed phylogeny package, and competes with PAUP* to be the one responsible for the largest number of published trees. PHYLIP has been in distribution since 1980, and has over 15,000 registered users.

Output is written onto special files with names like "outfile" and "outtree". Trees written onto "outtree" are in the Newick format, an informal standard agreed to in 1986 by authors of a number of major phylogeny packages.

Input is either provided via a file called “infile” or in response to a prompt.

input and output



Input and output files

For most of the PHYLIP programs, information comes from a series of input files, and ends up in a series of output files:

```
                       -------------------
                      |                   |
    infile ---------> |                   |
                      |                   |
    intree ---------> |                   | -----------> outfile
                      |                   |
    weights --------> |      program      | -----------> outtree
                      |                   |
    categories -----> |                   | -----------> plotfile
                      |                   |
    fontfile -------> |                   |
                      |                   |
                       -------------------
```

The programs interact with the user by presenting a menu. Aside from the user's choices from the menu, they read all other input from files. These files have default names. The program will try to find a file of that name - if it does not, it will ask the user to supply the name of that file. Input data such as DNA sequences comes from a file whose default name is `infile`. If the user supplies a tree, this is in a file whose default name is `intree`. Values of weights for the characters are in `weights`, and the tree plotting program need some digitized fonts which are supplied in `fontfile` (all these are default names).

What’s in PHYLIP

Programs in PHYLIP allow to do parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

Phylip works well with protein and nucleotide sequences
Many other programs mimic the style of PHYLIP programs.
(e.g. TREEPUZZLE, phyml, protml)

Many other packages use PHYIP programs in their inner workings (e.g., PHYLO_WIN)

PHYLIP runs under all operating systems

Web interfaces are available

Programs in PHYLIP are Modular

For example:

SEQBOOT take one set of aligned sequences and writes out a file containing bootstrap samples.

PROTDIST takes a aligned sequences (one or many sets) and calculates distance matices (one or many)

FITCH (or NEIGHBOR) calculate best fitting or neighbor joining trees from one or many distance matrices

CONSENSE takes many trees and returns a consensus tree

…. modules are available to draw trees as well, but often people use treeview or njplot

The Phylip Manual **is an excellent source of information.**

Brief one line descriptions of the programs are here

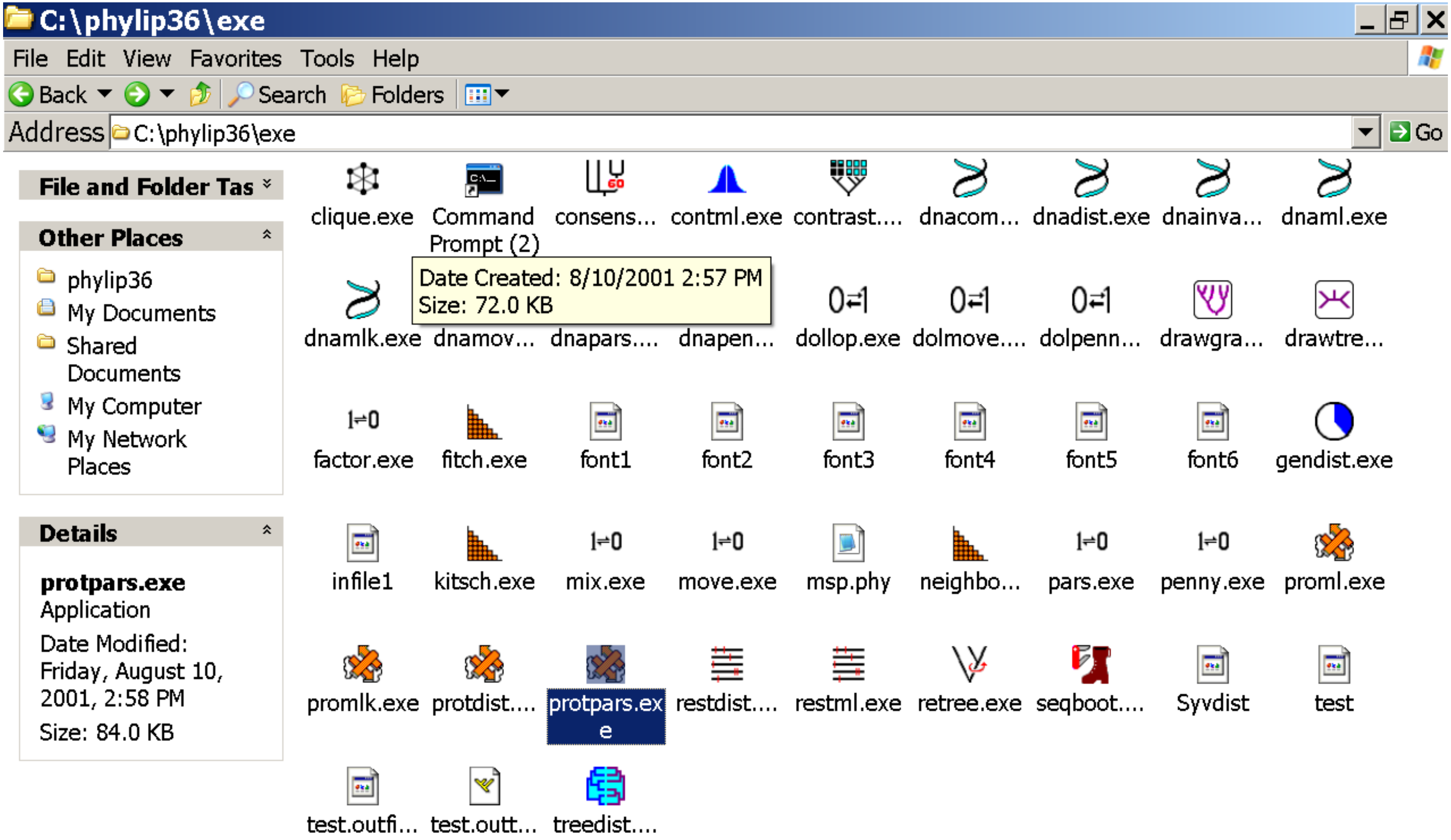
The easiest way to run PHYLIP programs is via a command line menu (similar to clustalw). The program is invoked through clicking on an icon, or by typing the program name at the command line.

```
> seqboot
> protpars
> fitch
```

If there is no file called infile the program responds with:

```
[gogarten@carrot gogarten]$ seqboot
seqboot: can't find input file "infile"
Please enter a new file name>
```

program folder



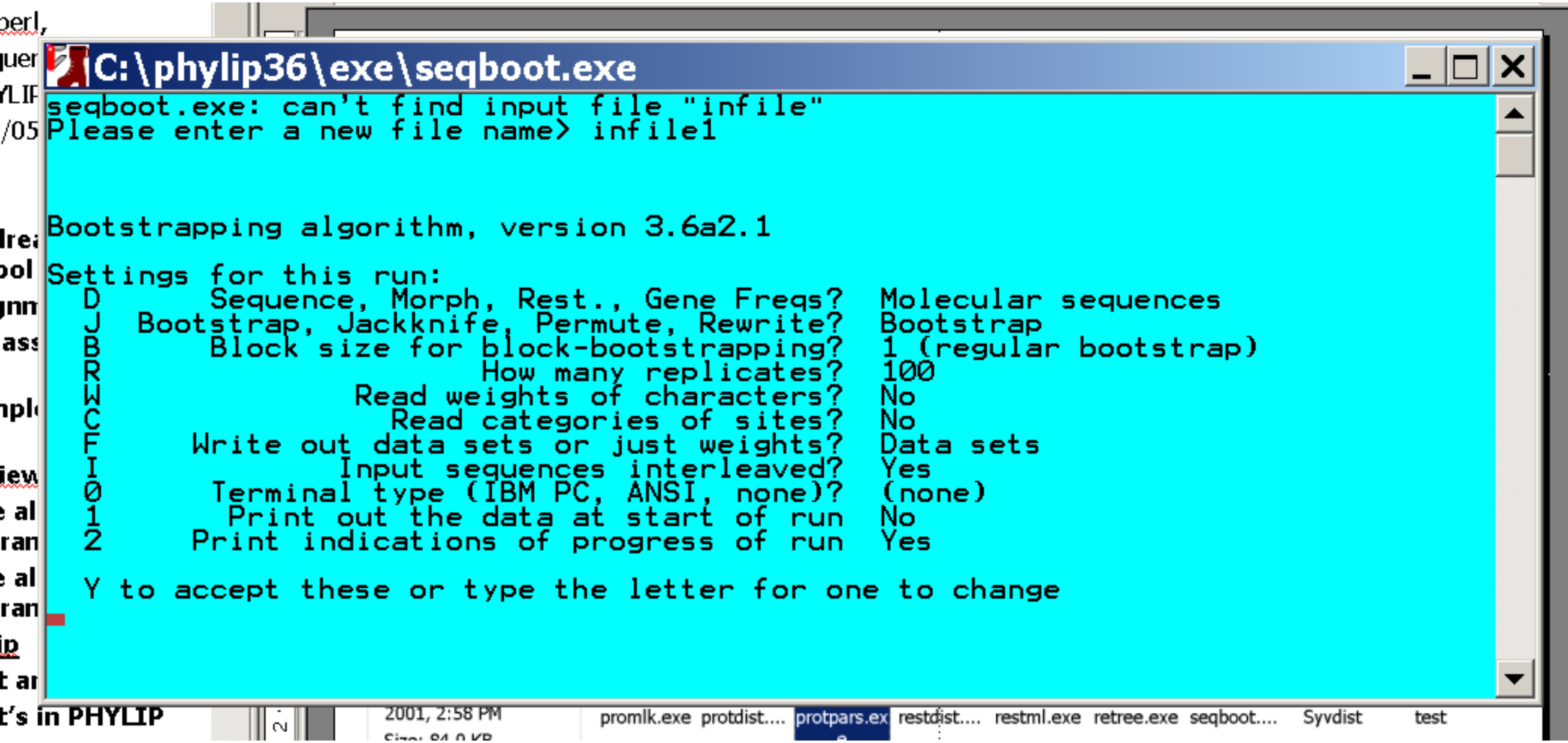
Date Created: 8/10/2001 2:57 PM
Size: 72.0 KB

protpars.exe
Application

Date Modified: Friday, August 10, 2001, 2:58 PM

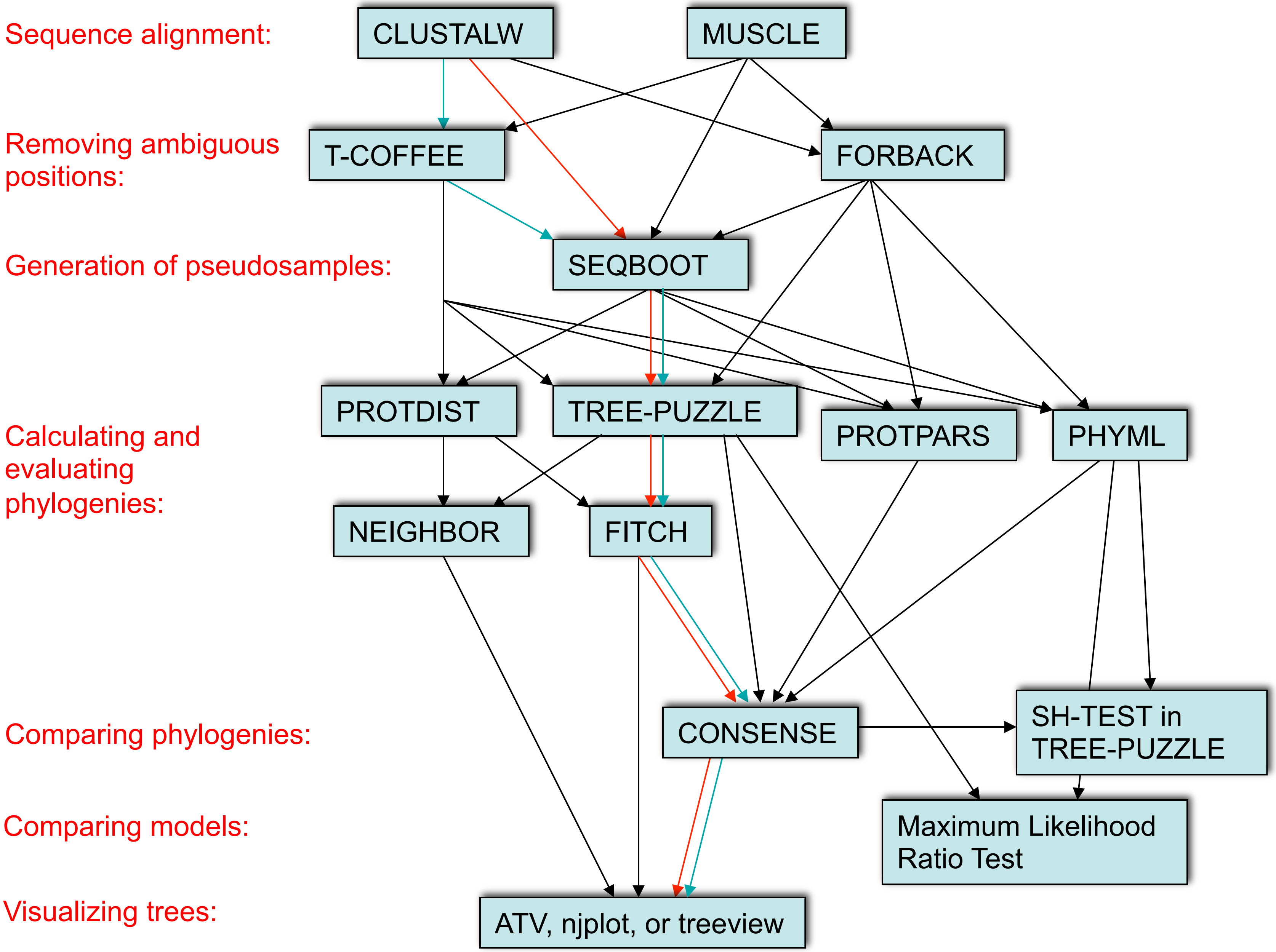
Size: 84.0 KB

menu interface



```
seqboot.exe: can't find input file "infile"
Please enter a new file name> infile1


Bootstrapping algorithm, version 3.6a2.1

Settings for this run:
  D      Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
  J  Bootstrap, Jackknife, Permute, Rewrite?  Bootstrap
  B      Block size for block-bootstrapping?  1 (regular bootstrap)
  R                     How many replicates?  100
  W              Read weights of characters?  No
  C                Read categories of sites?  No
  F     Write out data sets or just weights?  Data sets
  I             Input sequences interleaved?  Yes
  0      Terminal type (IBM PC, ANSI, none)?  (none)
  1       Print out the data at start of run  No
  2     Print indications of progress of run  Yes

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

example: seqboot and protpars on infile1



Sequence alignment: CLUSTALW MUSCLE

Removing ambiguous positions: T-COFFEE FORBACK

Generation of pseudosamples: SEQBOOT

Calculating and evaluating phylogenies: PROTDIST TREE-PUZZLE PROTPARS PHYML NEIGHBOR FITCH

Comparing phylogenies: CONSENSE SH-TEST in TREE-PUZZLE

Comparing models: Maximum Likelihood Ratio Test

Visualizing trees: ATV, njplot, or treeview

Phylip programs can be combined in many different ways with one another and with programs that use the same file formats.

Example 1 Protpars

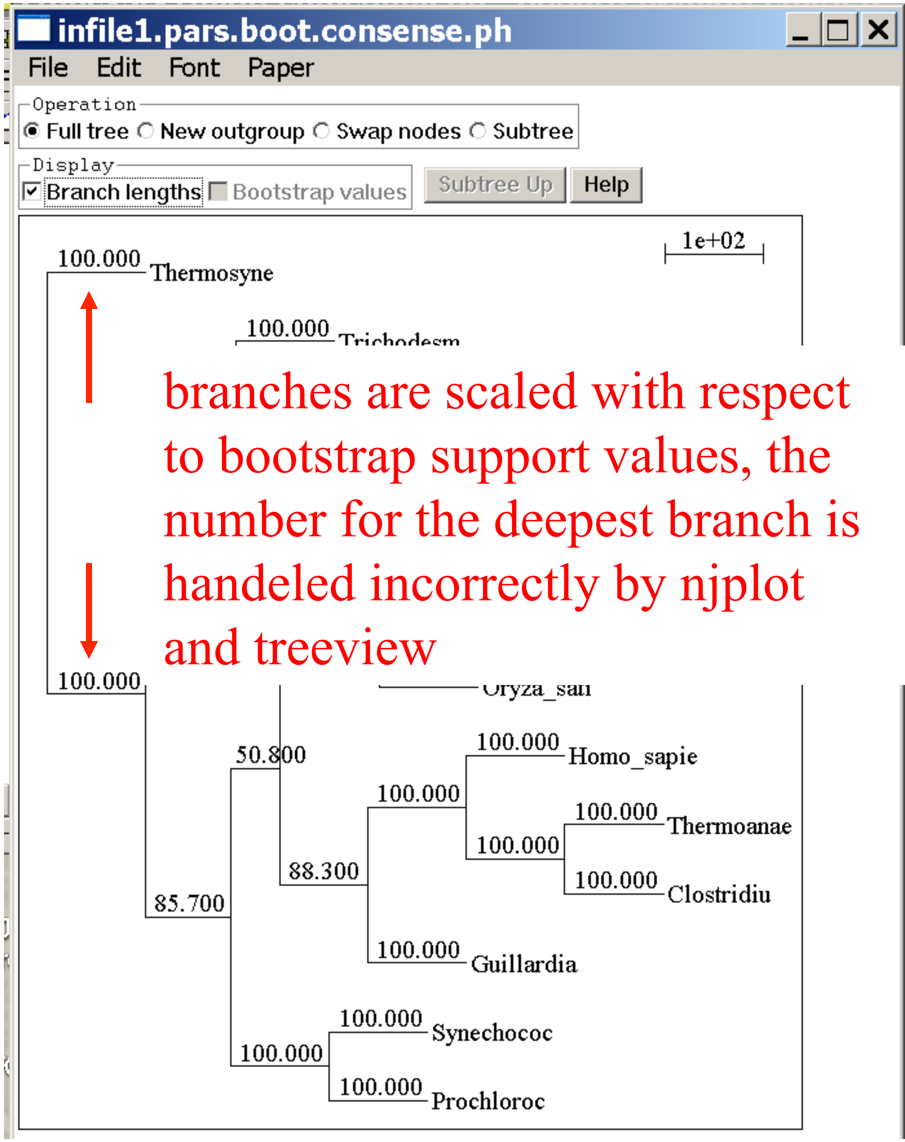
example: seqboot, protpars, consense on infile1

NOTE the bootstrap majority consensus tree does not necessarily have the same topology as the “best tree” from the original data!

threshold parsimony,
gap symbols - versus ?
(in vi you could use `:%s/-/?/g` to replace all – ?)
outfile
outtree compare to distance matrix analysis

protpars (versus distance/FM)

```
Extended majority rule consensus tree

CONSENSUS TREE:
the numbers on the branches indicate the number
of times the partition of the species into the two sets
which are separated by that branch occurred
among the trees, out of 100.00 trees

                                            +------Prochloroc
                +----------------------100.-|
                |                           +------Synechococ
                |
                |             +--------------------Guillardia
         +-85.7-|             |
         |      |      +-88.3-|             +------Clostridiu
         |      |      |      |      +-100.-|
         |      |      |      +-100.-|      +------Thermoanae
         |      +-50.8-|             |
         |             |             +-------------Homo sapie
  +------|             |
  |      |             |                    +------Oryza sati
  |      |             +---------------100.0-|
  |      |                                  +------Arabidopsi
  |      |
  |      |                    +--------------------Synechocys
  |      |                    |
  |      +---------------53.0-|             +------Nostoc pun
  |                           |      +-99.5-|
  |                           +-38.5-|      +------Nostoc sp
  |                                  |
  |                                  +-------------Trichodesm
  |
  +------------------------------------------------Thermosyne


  remember: this is an unrooted tree!
```

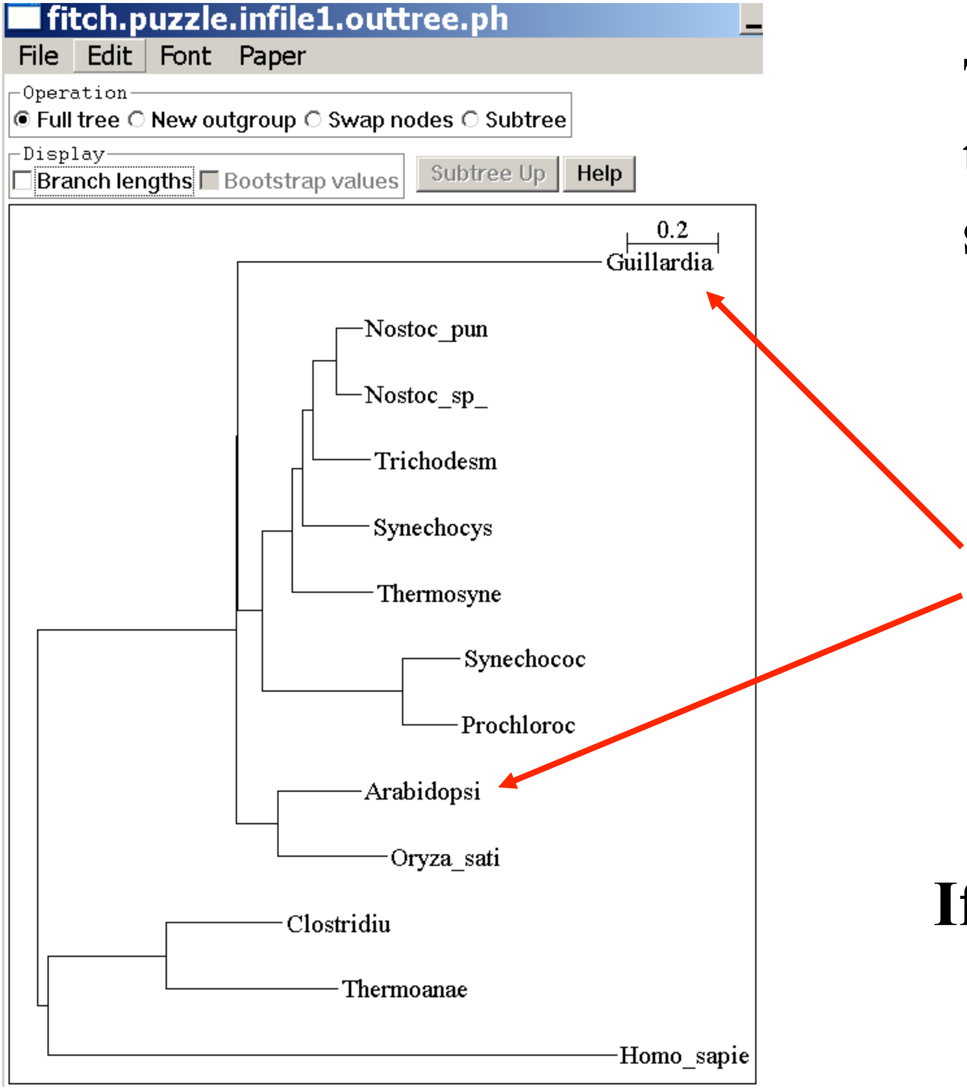


Operation
● Full tree ○ New outgroup ○ Swap nodes ○ Subtree

Display
☑ Branch lengths □ Bootstrap values

branches are scaled with respect to bootstrap support values, the number for the deepest branch is handeled incorrectly by njplot and treeview

(protpars versus) distance/FM



Tree is scaled with respect to the estimated number of substitutions.

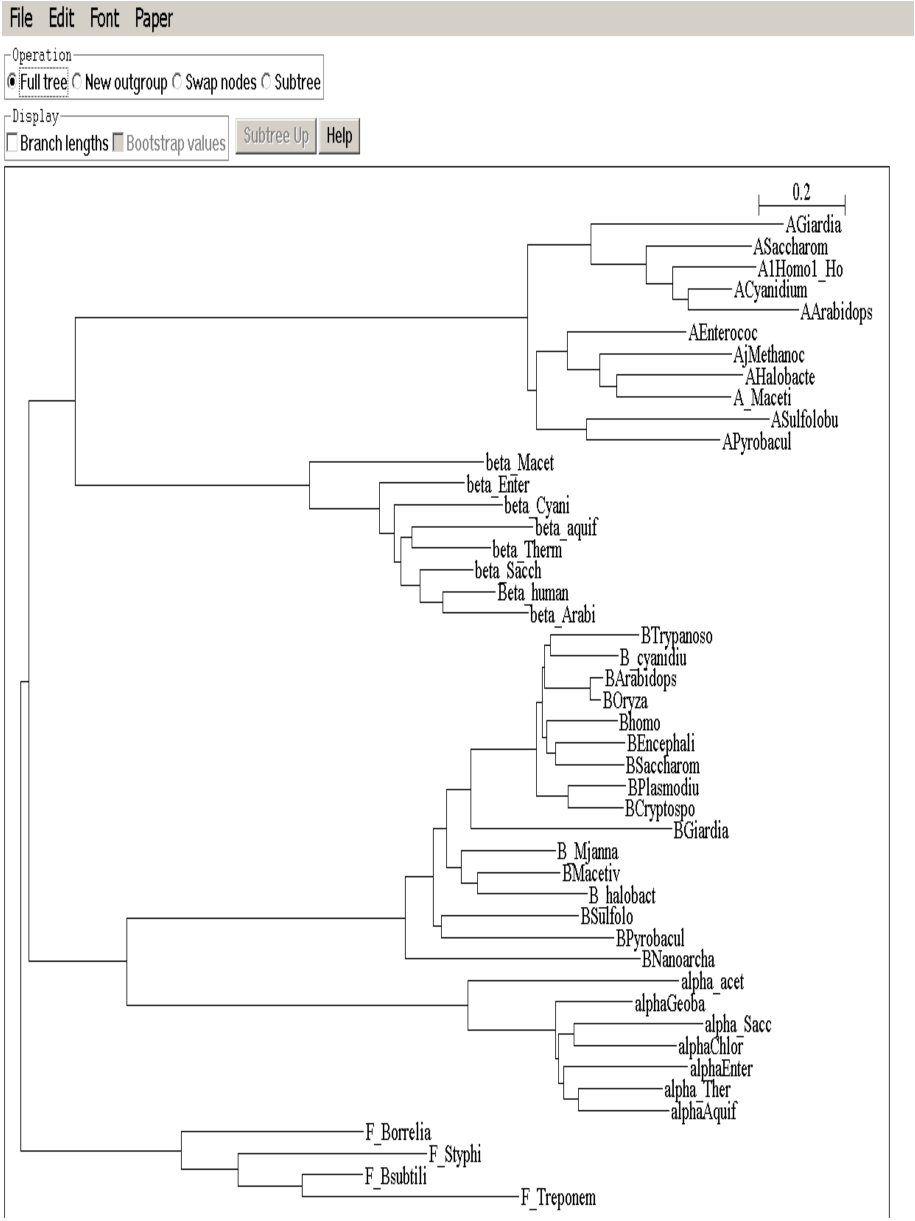
what might be the explanation for the red algae not grouping with the plants?

If time: demo of njplot

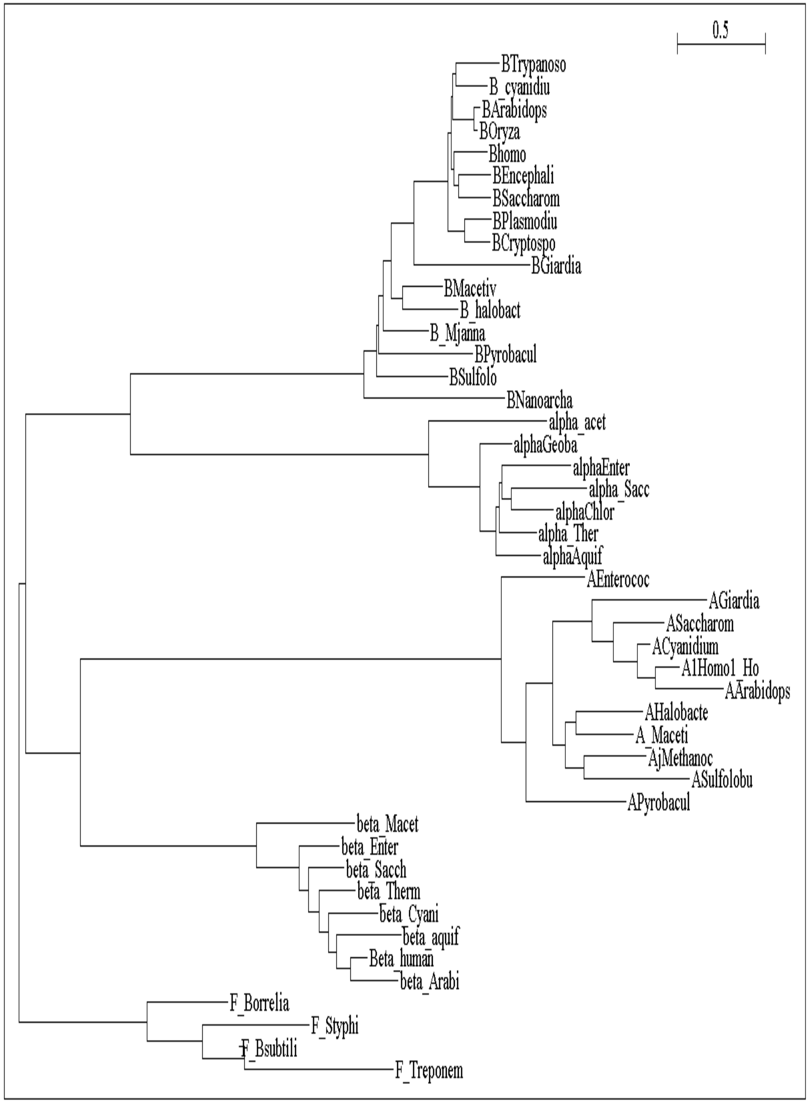
protdist

```
PROTdist
Settings for this run:
  P  Use JTT, PMB, PAM, Kimura, categories model?  Jones-Taylor-Thornton matrix
  G  Gamma distribution of rates among positions?  No
  C           One category of substitution rates?  Yes
  W                    Use weights for positions?  No
  M                   Analyze multiple data sets?  No
  I                  Input sequences interleaved?  Yes
  0                 Terminal type (IBM PC, ANSI)?  ANSI
  1            Print out the data at start of run  No
  2          Print indications of progress of run  Yes
```

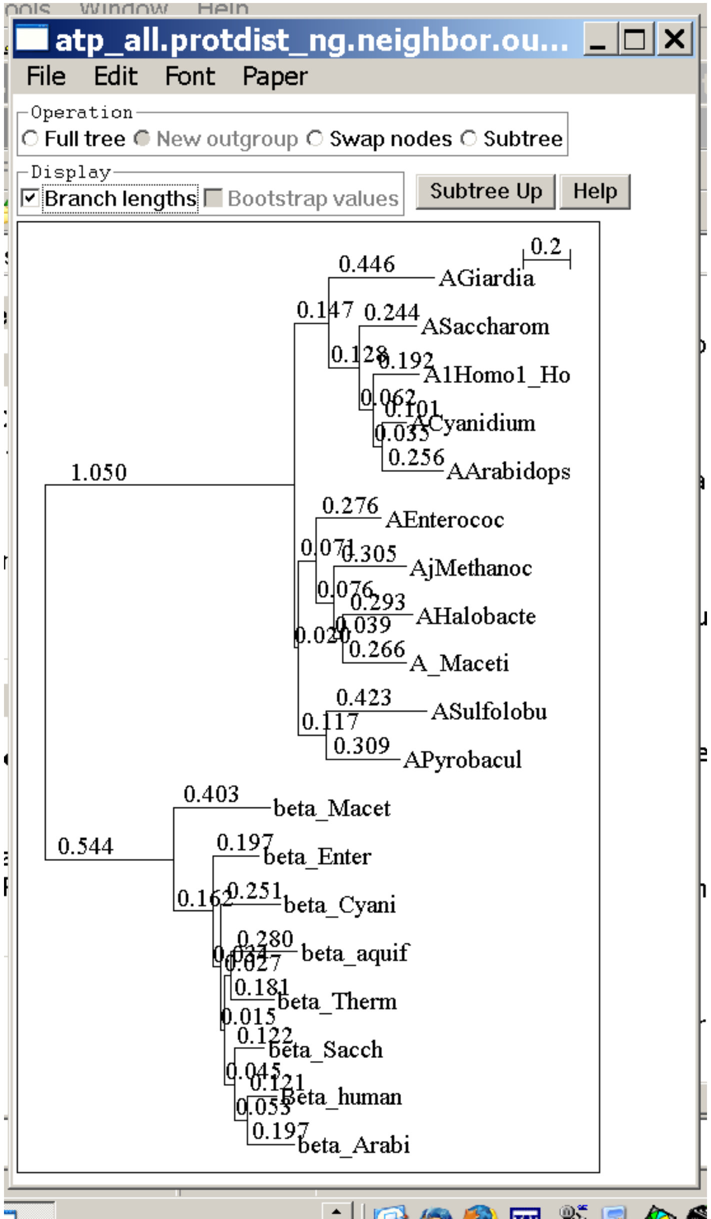
without and with correction for ASRV

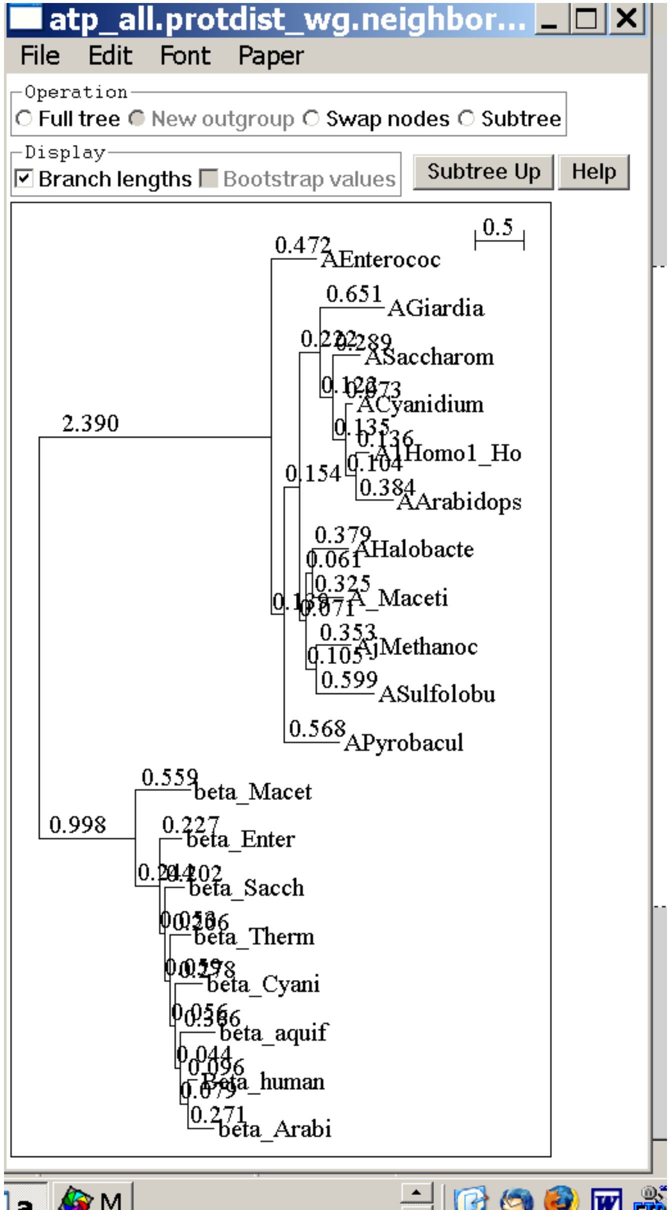

Operation
● Full tree ○ New outgroup ○ Swap nodes ○ Subtree

Display
□ Branch lengths □ Bootstrap values

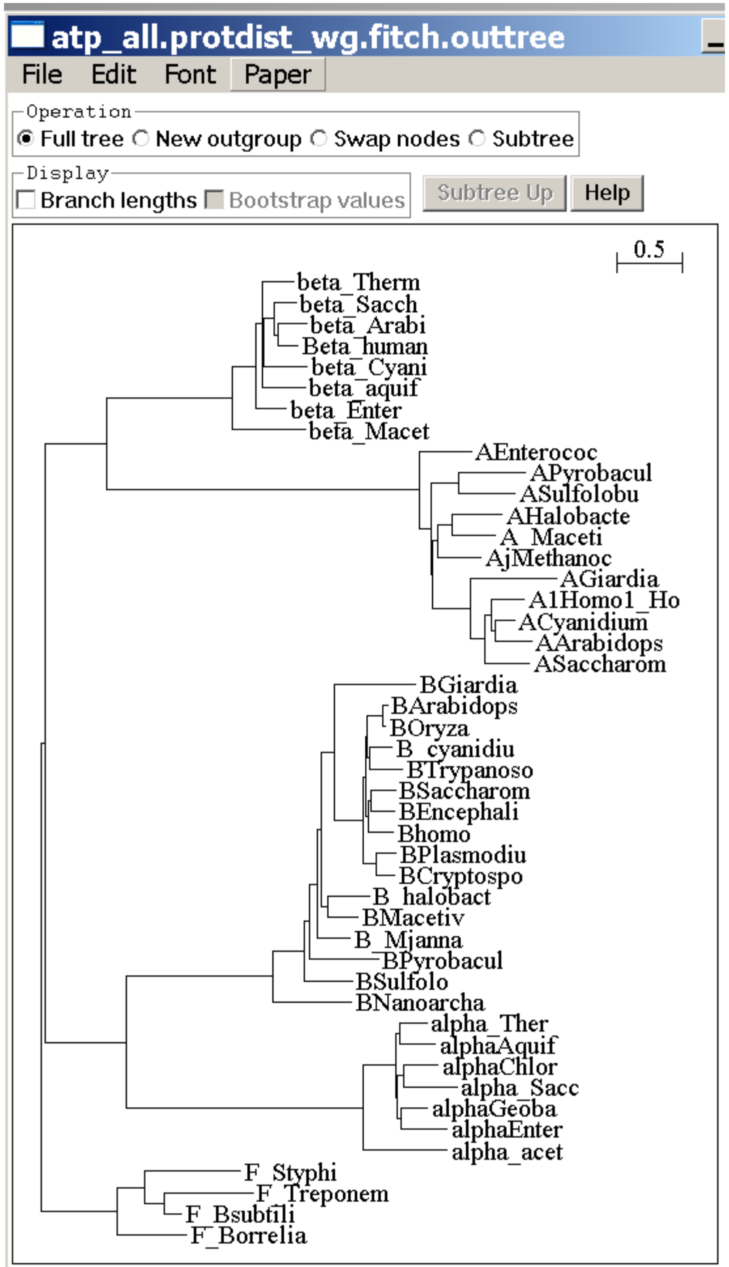


subtree with branch lengths
without and with correction for ASRV





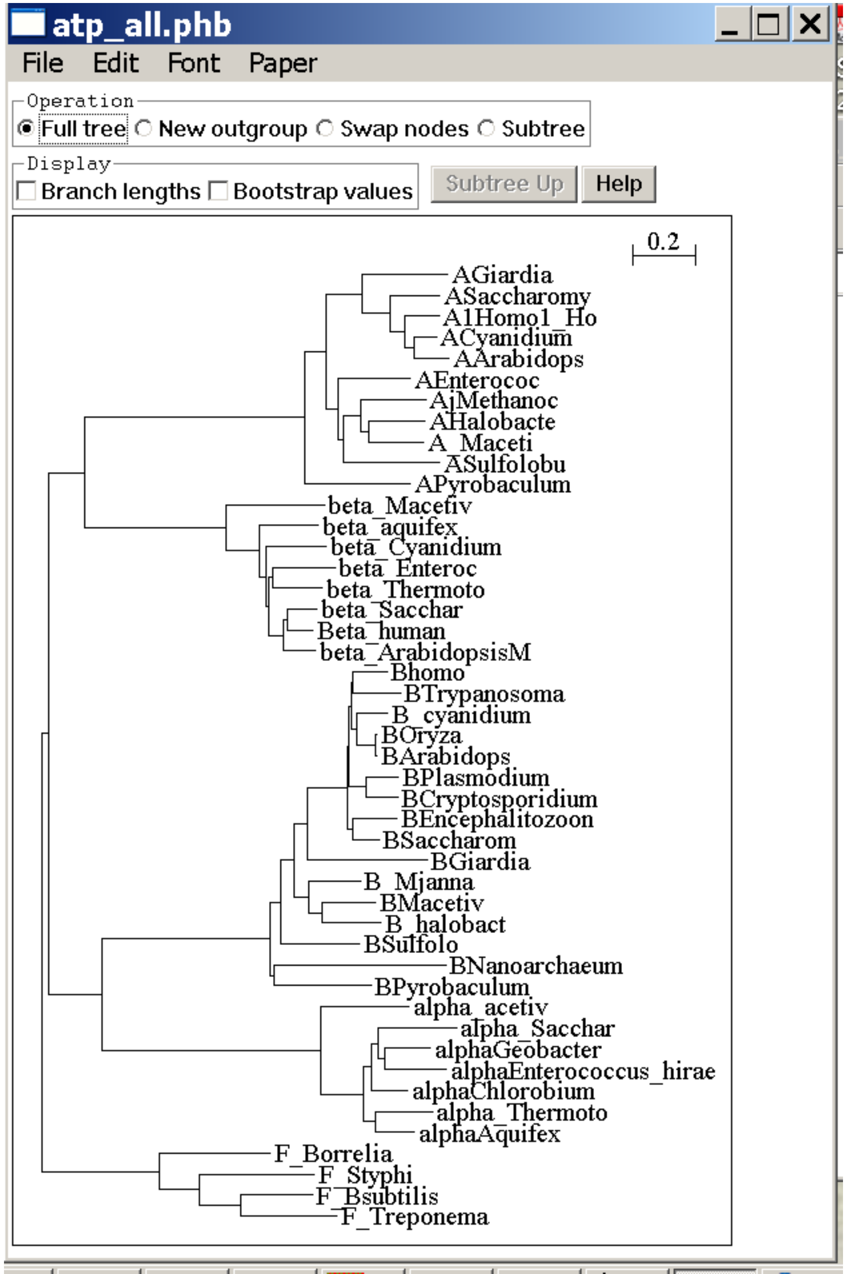
compare to trees with FITCH and clustalw – same dataset



atp_all.protdist_wg.fitch.outtree

Operation: ● Full tree ○ New outgroup ○ Swap nodes ○ Subtree

Display: □ Branch lengths □ Bootstrap values

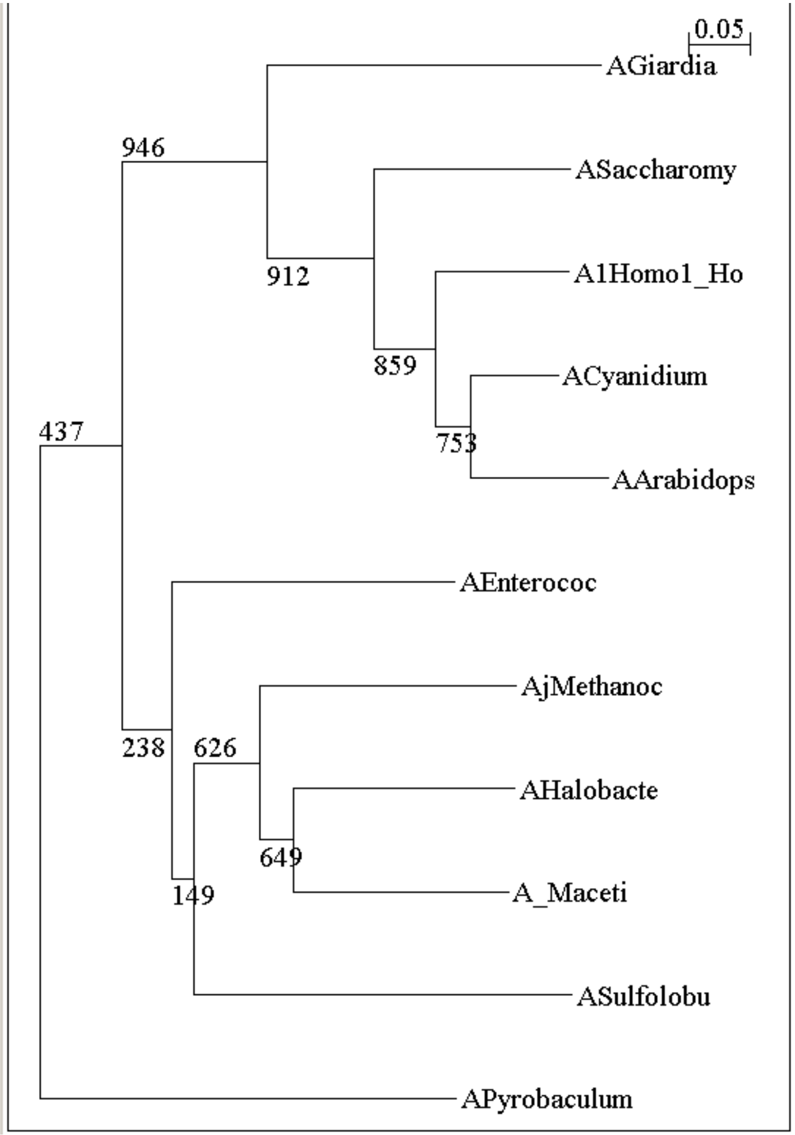


atp_all.phb

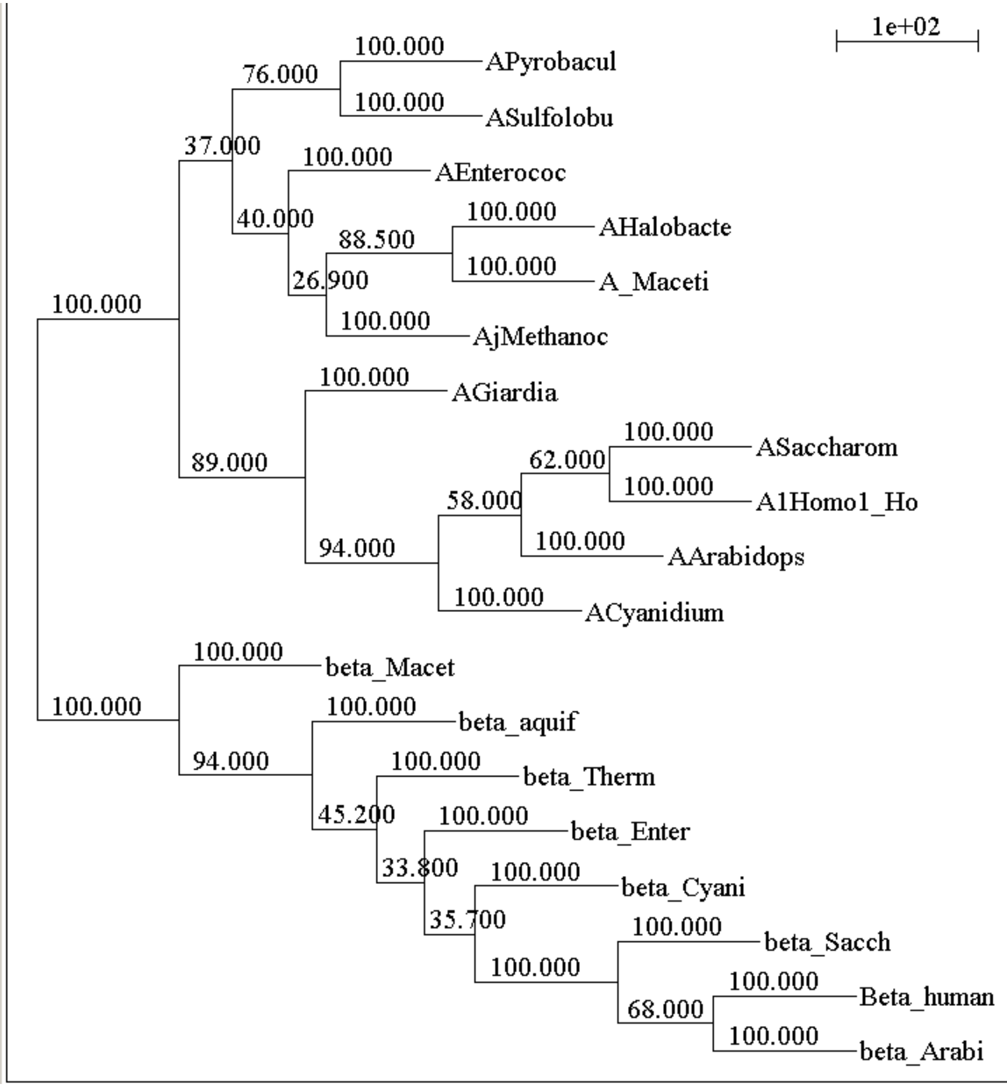
Operation: ● Full tree ○ New outgroup ○ Swap nodes ○ Subtree

Display: □ Branch lengths □ Bootstrap values

bootstrap support ala clustal



protpars (gaps as ?)



phyml

PHYML - A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood

An online interface is here ;
there is a command line version that is described here (not as straight forward as in clustalw);
a phylip like interface is automatically invoked, if you type "phyml" – the manual is here.

Phyml is installed on bbcxsrv1.

Do example on atp_all.phy
Note data type, bootstrap option within program, models for ASRV (pinvar and gamma), by default the starting tree is calculated via neighbor joining.

TreePuzzle ne PUZZLE

TREE-PUZZLE is a very versatile maximum likelihood program that is particularly useful to analyze protein sequences. The program was developed by Korbian Strimmer and Arnd von Haseler (then at the Univ. of Munich) and is maintained by von Haseler, Heiko A. Schmidt, and Martin Vingron

(contacts see http://www.tree-puzzle.de/).

TREE-PUZZLE

- allows fast and accurate estimation of ASRV (through estimating the shape parameter alpha) for both nucleotide and amino acid sequences,
- It has a “fast” algorithm to calculate trees through quartet puzzling (calculating ml trees for quartets of species and building the multispecies tree from the quartets).
- The program provides confidence numbers (puzzle support values), which tend to be smaller than bootstrap values (i.e. provide a more conservative estimate),
- the program calculates branch lengths and likelihood for user defined trees, which is great if you want to compare different tree topologies, or different models using the **maximum likelihood ratio test**.
- Branches which are not significantly supported are collapsed.
- TREE-PUZZLE runs on "all" platforms
- TREE-PUZZLE reads PHYLIP format, and communicates with the user in a way similar to the PHYLIP programs.

Maximum likelihood ratio test

If you want to compare two models of evolution (this includes the tree) given a data set, you can utilize the so-called maximum likelihood ratio test.
If L_1 and L_2 are the likelihoods of the two models, $d = 2(\log L_1 - \log L_2)$ approximately follows a Chi square distribution with n degrees of freedom. Usually n is the difference in model parameters. I.e., how many parameters are used to describe the substitution process and the tree. In particular n can be the difference in branches between two trees (one tree is more resolved than the other).
In principle, this test can only be applied if on model is a more refined version of the other. In the particular case, when you compare two trees, one calculated without assuming a clock, the other assuming a clock, the degrees of freedom are the number of OTUs – 2 (as all sequences end up in the present at the same level, their branches cannot be freely chosen) .

To calculate the probability you can use the CHISQUARE calculator for windows available from Paul Lewis.

TREE-PUZZLE allows (cont)

- TREEPUZZLE calculates distance matrices using the ml specified model. These can be used in FITCH or Neighbor.
  PUZZLEBOOT automates this approach to do bootstrap analyses – WARNING: this is a distance matrix analyses!
  The official script for PUZZLEBOOT is here – you need to create a command file (puzzle.cmds), and puzzle needs to be envocable through the command puzzle.
  Your input file needs to be the renamed outfile from **seqboot**
  A slightly modified working version of puzzleboot_mod.sh is here, and here is an example for puzzle.cmds . Read the instructions before you run this!
- Maximum likelihood mapping is an excellent way to assess the phylogenetic information contained in a dataset.
- ML mapping can be used to calculate the support around one branch.

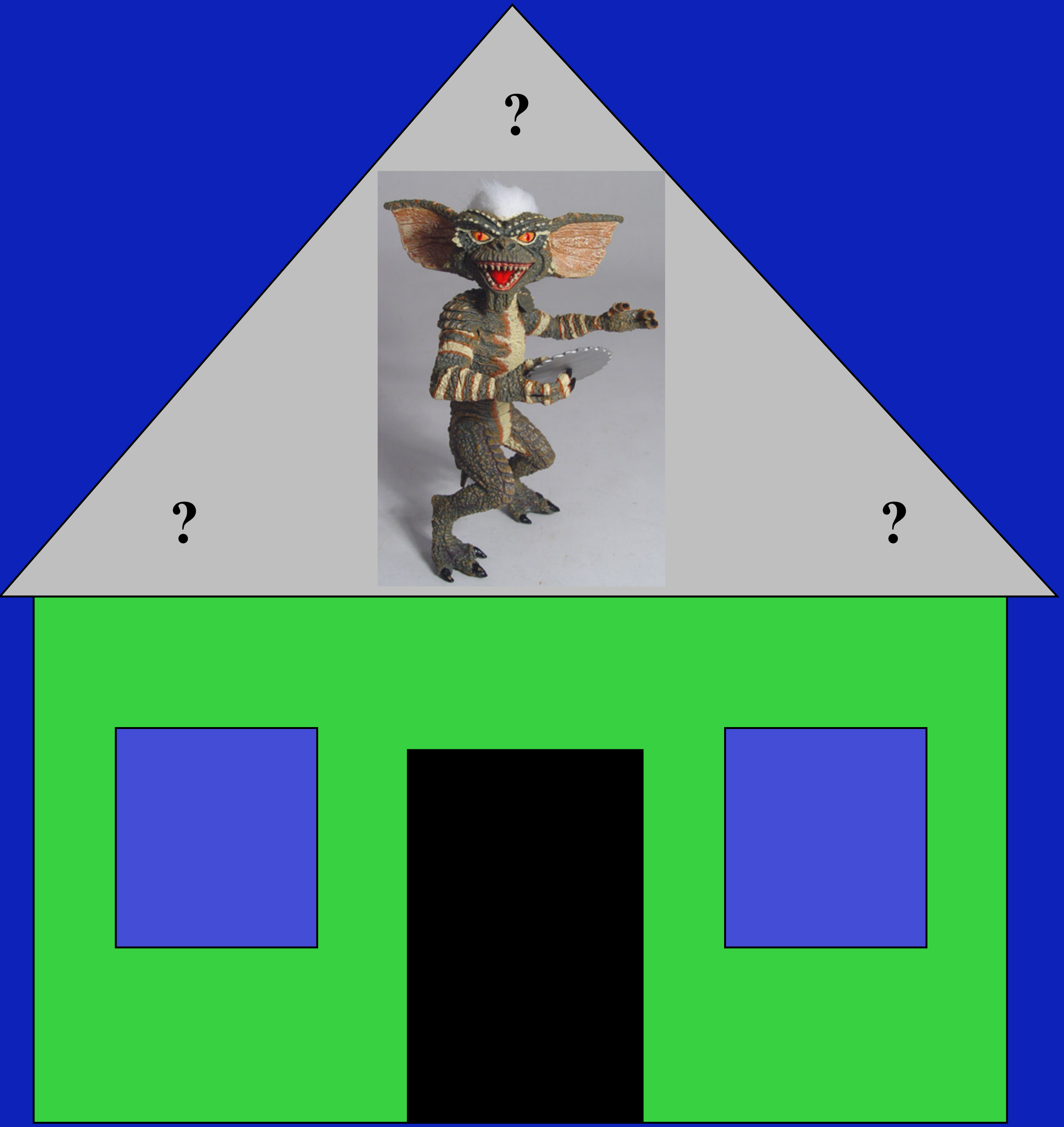
  @@@ Puzzle is cool, don't leave home without it! **@@@**

TREE-PUZZLE – PROBLEMS/DRAWBACKS

- The more species you add the lower the support for individual branches. While this is true for all algorithms, in TREE-PUZZLE this can lead to completely unresolved trees with only a few handful of sequences.

- Trees calculated via quartet puzzling are usually not completely resolved, and they do not correspond to the ML-tree: The determined multi-species tree is not the tree with the highest likelihood, rather it is the tree whose topology is supported through ml-quartets, and the lengths of the resolved branches is determined through maximum likelihood.

Elliot Sober’s Gremlins



Observation: Loud noise in the attic

Hypothesis: *gremlins in the attic playing bowling*

Likelihood =
P(noise|gremlins in the attic)

P(gremlins in the attic|noise)

Bayes' Theorem

Reverend Thomas Bayes
(1702-1761)

$$\underbrace{P(model|data, I)}_{\text{Posterior Probability}} = \underbrace{P(model, I)}_{\text{Prior Probability}} \frac{\overbrace{P(data|model, I)}^{\text{Likelihood}}}{\underbrace{P(data,I)}_{\text{Normalizing constant}}}$$

Likelihood

describes how well the model predicts the data

Posterior Probability

represents the degree to which we believe a given **model** accurately describes the situation given the available **data** and all of our prior information **I**

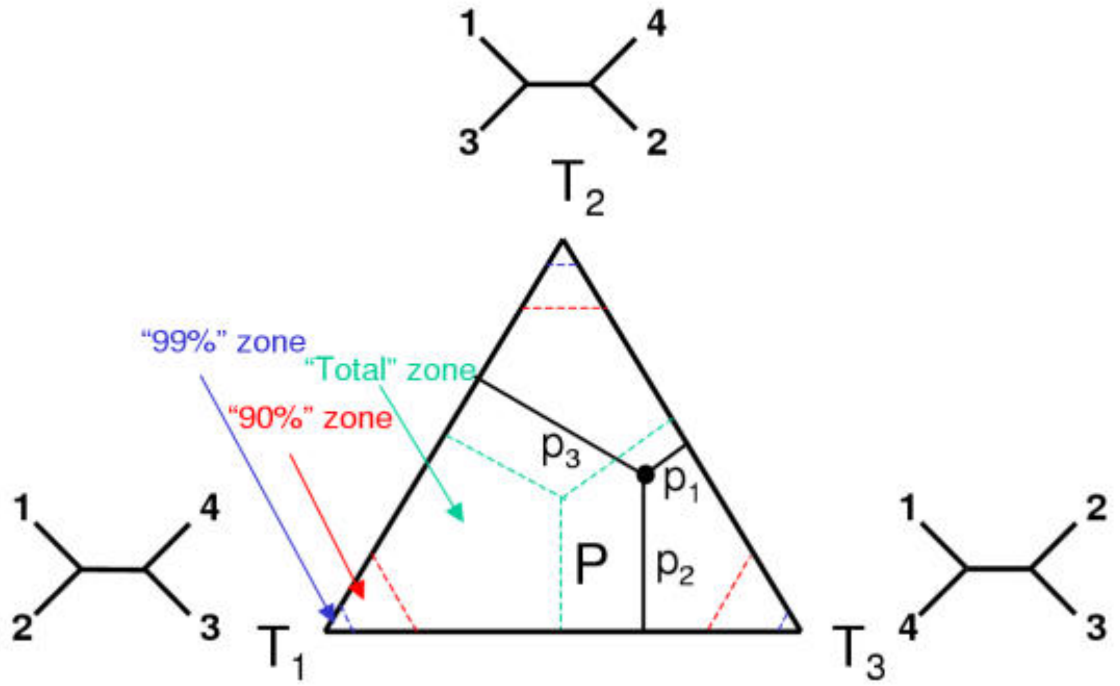
Prior Probability

describes the degree to which we believe the model accurately describes reality based on all of our prior information.

Normalizing constant

ml mapping

-



From: Olga Zhaxybayeva and J Peter Gogarten *BMC Genomics* 2002, 3:4

ml mapping

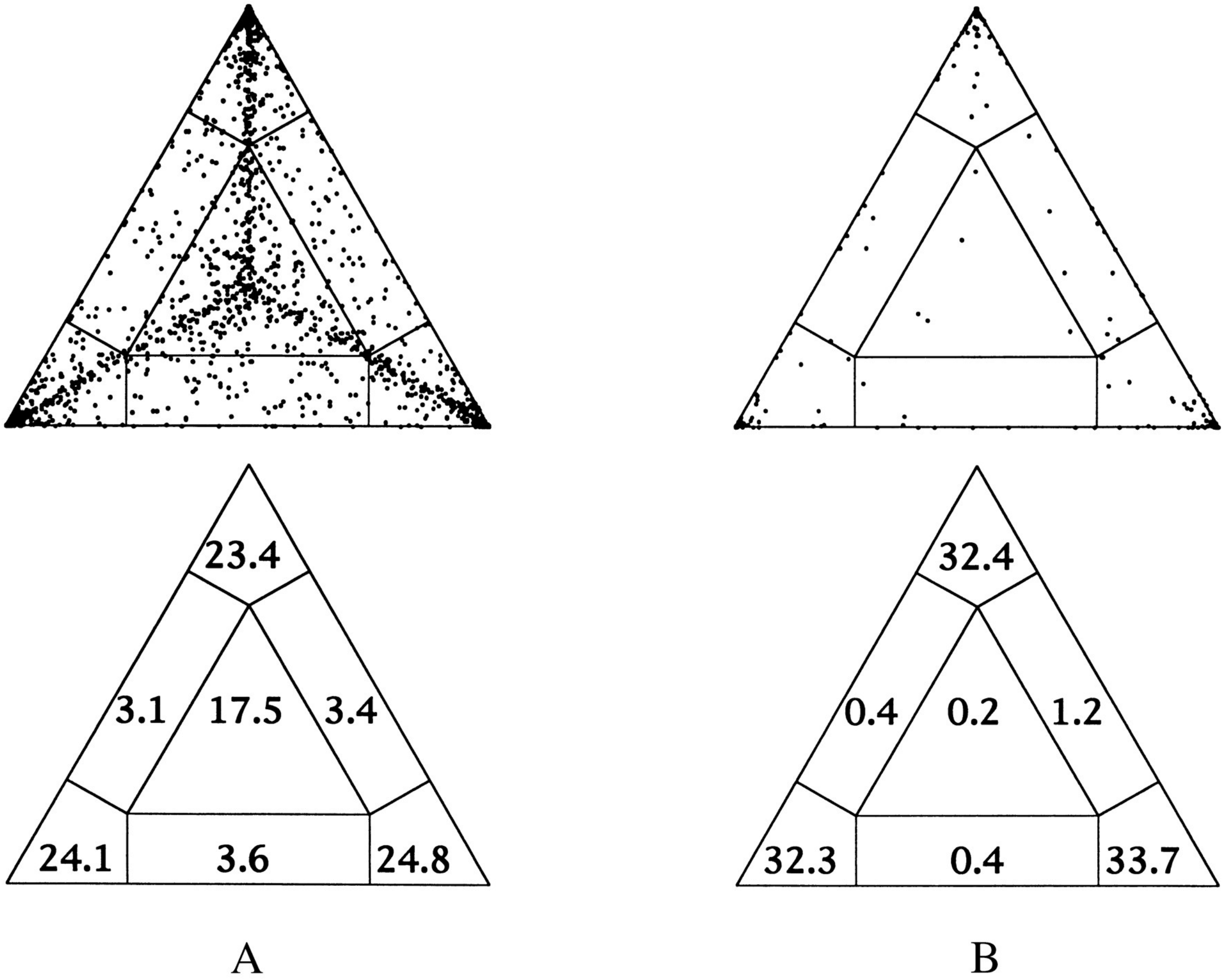
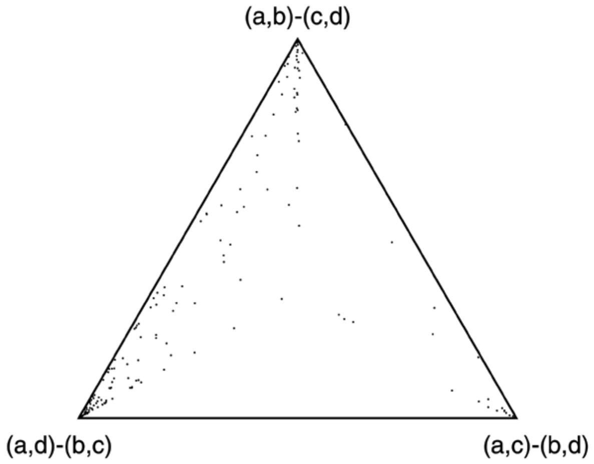
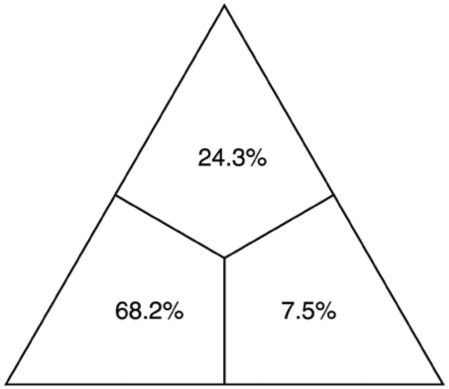


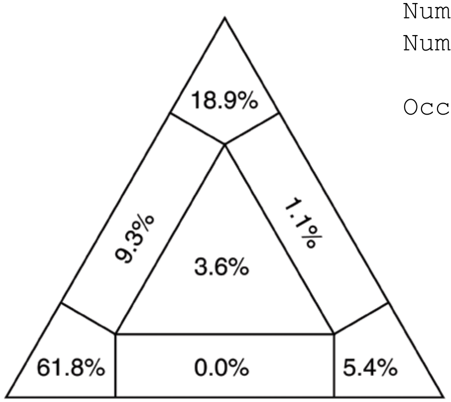
Figure 5. Likelihood-mapping analysis for two biological data sets. (*Upper*) The distribution patterns. (*Lower*) The occupancies (in percent) for the seven areas of attraction.

(*A*) Cytochrome-*b* data from ref. 14. (*B*) Ribosomal DNA of major arthropod groups (15).

From: Korbinian Strimmer and Arndt von Haeseler Proc. Natl. Acad. Sci. USA Vol. 94, pp. 6815-6819, June 1997







Cluster a: 14 sequences
outgroup (prokaryotes)

Cluster b: 20 sequences
other Eukaryotes

Cluster c: 1 sequences
Plasmodium

Cluster d: 1 sequences
Giardia

```
                 (a,b)-(c,d)
                      /\
                     /  \
                    /    \
                   /   1  \
                  / \    / \
                 /   \  /   \
                /     \/     \
               /  3    :   2  \
              /        :       \
             /__________________\
   (a,d)-(b,c)                  (a,c)-(b,d)

Number of quartets in region 1: 68 (= 24.3%)
Number of quartets in region 2: 21 (= 7.5%)
Number of quartets in region 3: 191 (= 68.2%)

Occupancies of the seven areas 1, 2, 3, 4, 5, 6, 7:

                 (a,b)-(c,d)
                      /\
                     /  \
                    /  1 \
                   / \  / \
                  /   /\   \
                 / 6 /  \ 4 \
                /   / 7  \   \
               / \ /______\ / \
              / 3  :   5  :  2 \
             /__________________\
   (a,d)-(b,c)                  (a,c)-(b,d)

Number of quartets in region 1: 53 (= 18.9%)
Number of quartets in region 2: 15 (= 5.4%)
Number of quartets in region 3: 173 (= 61.8%)
Number of quartets in region 4: 3 (= 1.1%)
Number of quartets in region 5: 0 (= 0.0%)
Number of quartets in region 6: 26 (= 9.3%)
Number of quartets in region 7: 10 (= 3.6%)
```

Alternative Approaches to Estimate Posterior Probabilities

Bayesian Posterior Probability Mapping with MrBayes

(Huelsenbeck and Ronquist, 2001)

Problem:

Strimmer's formula $$p_i = \frac{L_i}{L_1+L_2+L_3}$$

only considers 3 trees (those that maximize the likelihood for the three topologies)

Solution:

Exploration of the tree space by sampling trees using a biased random walk
(Implemented in MrBayes program)

Trees with higher likelihoods will be sampled more often

$$p_i \approx \frac{N_i}{N_{total}}$$

,where N_i - number of sampled trees of topology *i, i*=1,2,3

N_{total} – total number of sampled trees (has to be large)

Illustration of a biased random walk

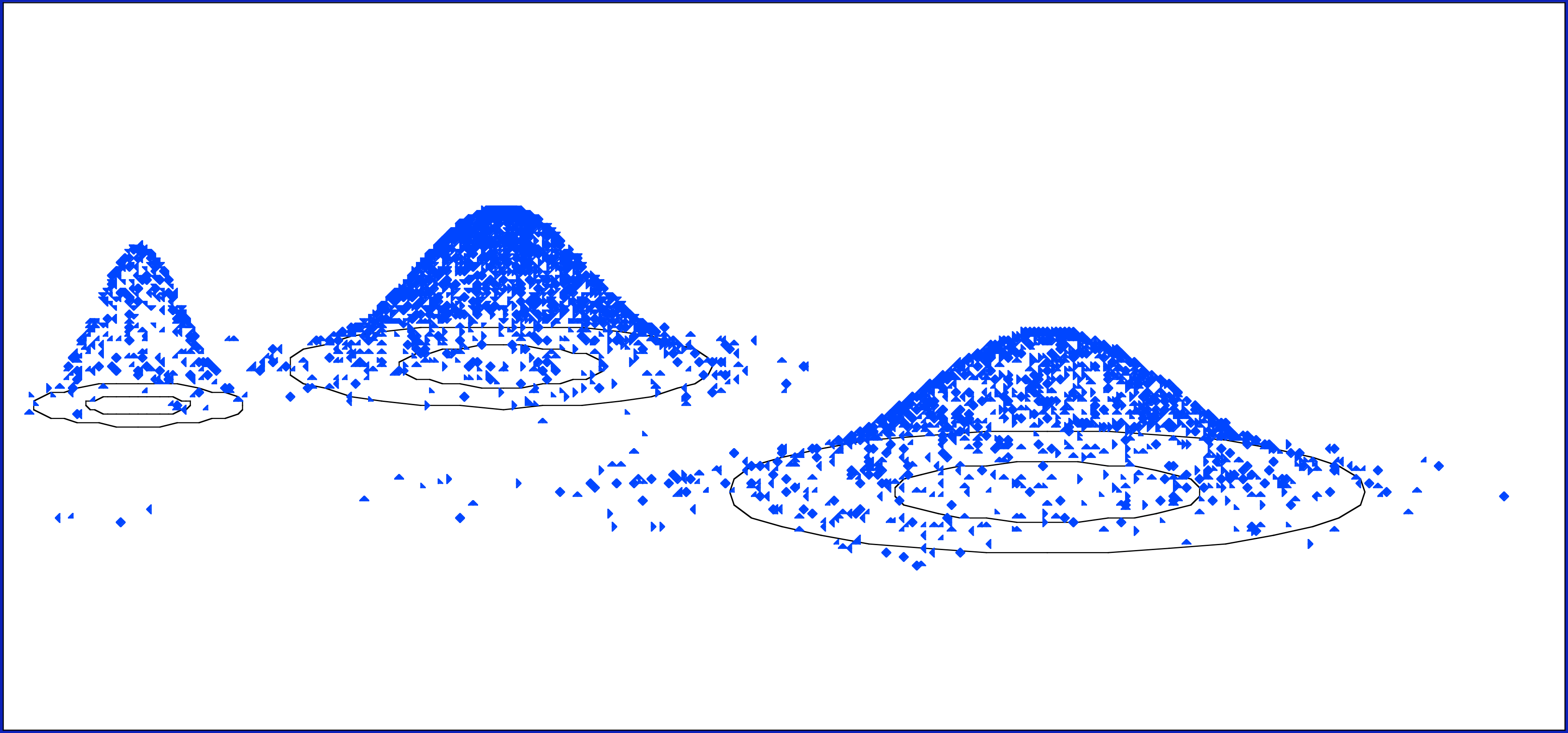
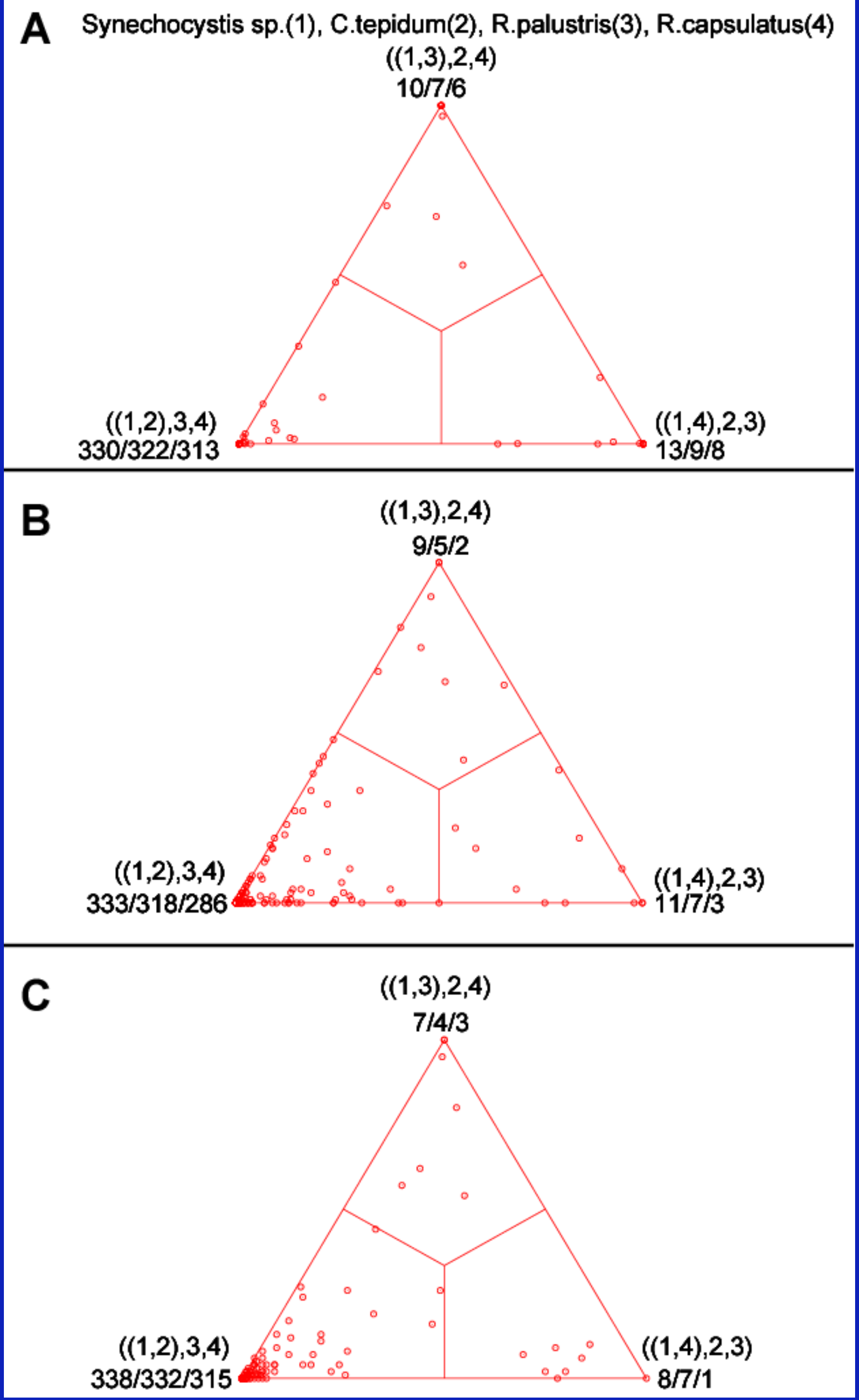


Figure generated using MCRobot program (Paul Lewis, 2001)



Zhaxybayeva and Gogarten, *BMC Genomics* 2003 4: 37

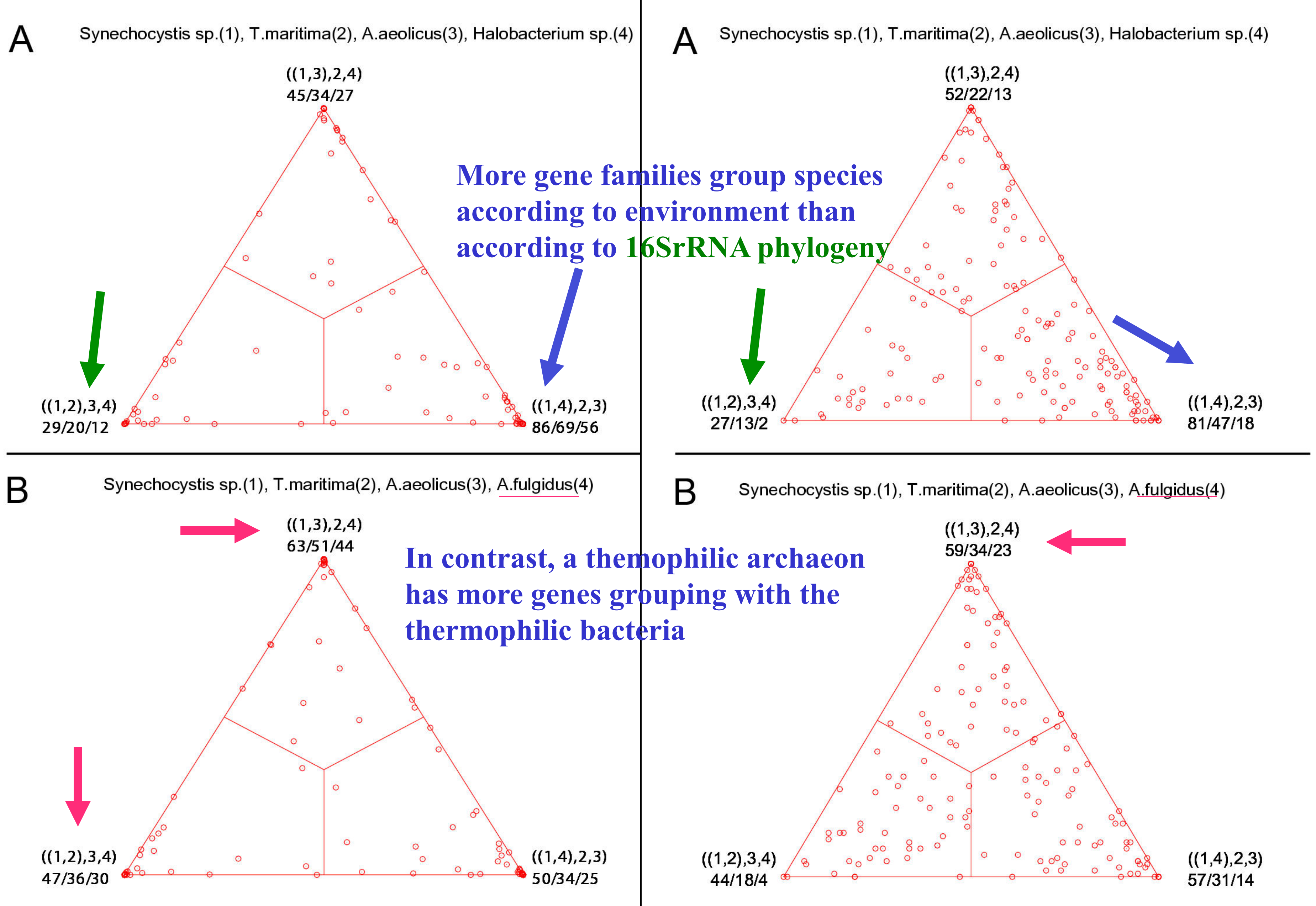
COMPARISON OF DIFFERENT SUPPORT MEASURES

A: mapping of posterior probabilities according to Strimmer and von Haeseler

B: mapping of bootstrap support values

C: mapping of bootstrap support values from extended datasets

ml-mapping *versus* bootstrap values from extended datasets



More gene families group species according to environment than according to 16SrRNA phylogeny

In contrast, a themophilic archaeon has more genes grouping with the thermophilic bacteria

New Assignments:

- **Write a script that determines the number of elements in a %ash.**

- **Write a script (or subroutine) that prints out a hash sorted on the keys in alphabetical order.**

- **How can you remove an entry in a hash (key and value)?**

- **Write a program that it uses hashes to calculates mono-, di-, tri-, and quartet-nucleotide frequencies in a genome.**

Perl assignments

Write a script that takes all phylip formated aligned multiple sequence files present in a directory, and performes a bootstrap analyses using maximum parsimony.

Files you might want to use are A.fa, B.fa, alpha.fa, beta.fa, and atp_all.phy. BUT you first have to convert them to phylip format AND you should replace some or all gaps with ?
(In the end you would be able to answer the question “does the resolution increase if a more related subgroup is analyzed independent from an outgroup?)

hints

Rather than typing commands at the menu, you can write the responses that you would need to give via the keyboard into a file (e.g. your_input.txt)

You could start and execute the program protpars by typing

protpars < your_input.txt

your input.txt might contain the following lines:

```
infile1.txt
r
t
10
y
r
r
```

in the script you could use the line

```
system (“protpars < your_input.txt”);
```

The main problem are the owerwrite commands if the oufile and outtree files are already existing. You can either create these beforehand, or erase them by moving (mv) their contents somewhere else.

create *.phy files

the easiest (probably) is to run clustalw with the phylip option:
For example (here):

```
#!/usr/bin/perl -w
print "# This program aligns all multiple sequence files with names *.fa \n
# found in its directory using clustalw,  and saves them in phyip format.\n“;
while(defined($file=glob("*.fa"))){
            @parts=split(/\./,$file);
            $file=$parts[0];
            system("clustalw -infile=$file.fa -align -output=PHYLIP");
            };
# cleanup:
system ("rm *.dnd");
exit;
```

Alternatively, you could use a web version of readseq – this one worked great for me ☺

Alternative for entering the commands for the menu:

```
#!/usr/bin/perl -w
        system ("cp A.phy infile");
        system ("echo -e 'y\n9\n'|seqboot");
exit;
```

echo returns the string in \` \`, i.e., y\n9\n.
The -e options allows the use of \n
The | symbol pipes the output from echo to seqboot