Users Manual: Horizon Annealing

H.D. Sheets, August 22, 2012 Edition

Introduction:

Horizon Annealing is a method of biostratigraphic correlation, which seeks to use simulated annealing methods to produce an optimal ordination of fossil collection horizons appearing in different sections (localities). This approach is very similar to that used in the CONOP system developed and used by Peter Sadler and colleagues, but rather than ordering FADs and LADs, HA orders collections. The approach of ordering collections is somewhat akin to the methods developed by John Alroy, but HA uses a different optimality criteria and method of searching for solutions.

Before attempting to use this program please read:


And


This is a draft users manual! Please feel free to contact me if you run into difficulty with the program or manual, that will provide incentive for me to put more time into this.

sheets@canisius.edu or hsheets@buffalo.edu
Contents

1.) What does the program do?
2.) Input File Formats
3.) Running the program
3.) Graphic Representation of Results
4.) Output Files
5.) References
1.) What does the program do?

The Horizon Annealing program loads in a set of files that specify the location of local range ends (FADs and LADs) within each of a number of distinct sections (locality). The program then seeks to determine the ordering or sequencing of the horizons in the data set that produces the least amount of unobserved range extensions, summed over all sections. The program searches for the optimal ordering of horizons by proposing a range of different variations on the horizon ordering, and using simulated annealing to search for an optimal solution. Details of this operation are in the Sheets et al. 2012 paper, but largely consist of proposing expansion or contraction of a section relative to the composite solution, insertion of hiatuses (gaps) within sections, and motion of the section upward or downward in time relative to the composite. Whole collections (horizons) are moved as units, and temporal ordering of horizons within a section is always preserved.

At the moment, range extensions are counted in units of horizons, rather than in rock thickness, and the events allowed are limited to species FADs and LADs. The conceptual basis of HA does allow for the use of rock thickness as a measure of range extension, and also allows for the inclusion of radiometric dates or ash beds as data types (with differing impacts or weights on the proposed solutions).

Input data

The program begins by having the user load in a data file, which has numerically coded events in each, with each line specifying a specific event (FAD or LAD), in a specific section, at a specific horizon (identified by a height measurement, and an ordinal position within the section). Two dictionary files specifying the names and id numbers of taxa and of the sections are also loaded. It is also possible to load a previously obtained solution (horizon ordination) to work from, or to specify a random starting configuration. These files are in the same format as the CONOP system uses, the input file formats are discussed below. The user must load each of these three or four files using a menu system prior to running the program.

Annealing Control

At each step of the program, a variant solution is proposed, by randomly selecting a section and then randomly altering its position by expanding or contracting it, sliding it upwards or downward, or inserting a gap. The range extension of the proposed solution (which is the error function for the solution) is then determined, by summing all local range extensions implied by the proposed solution. If the proposed solution is better than the current solution (the working solution), the proposed solution is used as the new working solution. When the proposed solution is worse, then a random value in the range 0-1 is selected and compared with the Boltzmann factor for the solution:

\[ e^{-(\text{increase in range extension})/\text{temperature}} \]

If the random value is less than the Boltzmann factor, then the proposed solution is accepted as the working solution. If the solution has a range extension lower than any seen previously, it is also store as the best solution. This process continues on for many iterations (proposed solutions) at each
temperature value. Large temperatures relative to the range extension will allow the program to accept solutions with increased range extensions relatively often, lower temperatures will result in higher errors being accepted only rarely.

In simulated annealing, the temperature is slowly reduced, so that initially a wide range of solutions is examined, but as the temperature falls the system is no longer allowed to frequently move to higher error solutions.

To run a simulated annealing program one must specify

a.) An initial temperature

b.) A temperature cooling schedule

c.) Some way of specifying the number of trials used

The HA program currently uses a user specified starting temperature, and then uses two loops to control the number of trials. The “inner” loop is the number of trials (proposed solutions) to be generated at each temperature value. There is a user-controlled cooling rate (ranging from 0 to 1), which is used to reduce the temperature at the end of each inner loop, by multiplying the temperature by the cooling rate, resulting in an exponential decay of the temperature. The user also specifies the number of steps of temperature reduction to be made (the “outer” loop). These control options appear on the main screen of the programs, with default values show.

**Graphical Output**

Graphical Information about the annealing process

The program will produce a plot at the end of a trial showing the range extension of the current solution at each temperature reduction step, as a mauve diamond, and the value of the best solution every time the best solution changes as green squares. This is shown in Figure 1, below, which also shows the input and output file controls and the annealing controls on the program.

The plot of range extension versus number of cooling steps allows the user to see how effective the cooling process is. Adjusting the initial temperature and cooling rate will alter the effectiveness of the search.
Figure 1: Image of control panel, showing proposed and best solution as a function of the number of cooling steps. This was a solution trial for the Riley Data set using a random starting configuration, and shows rapid improvement in the best solution found with cooling.

Figure 2: This is the plot of the proposed solution for 200 trials of the Riley solution starting at a low temperature (4), and achieving no solutions better than the initial solution used. The initial solution was the result of roughly 10 trials, always starting from the best known solution and varying the starting temperature, cooling rate and number of steps. This is a “mature” solution at this point, as further trials from the best solution no longer produce improvements in the net range extension.
Graphical Representation of information about the solution

Once a solution is found, the program can produce a variety of graphic displays of the solution.

**Figure 3:** This is a color coded range chart based on the composite solution for the Riley system. Species names are shown in abbreviated form below the ranges. Color codes indicate the number of sections each species was found in, black means 1 section, blue 2 or 3, green 4 or 5 and red 6 or more. The color coding is meant to indicate the amount of evidence the taxa range is based on. The Y-axis is the ordinal position in the composite.

**Figure 4:** Section Range chart. This plot shows the position of individual horizons (colored horizontal lines) within each section (section numbers are across the bottom of the plot). Positions are shown...
relative to the position in the composite. The line colors indicate the number of events appearing on each collection horizon.

**Figure 5**: This is a biplot of the first section (Morgan Creek) in the Riley Formation data set, showing the position of events and horizons in the section (y-axis) against the position in the composite (x-axis). The green circles are the observed FADS in the section, the red “x”s are the observed LADs and the blue squares are the horizon placements, or the line of correlation, of the section with the composite. The HA program is capable of producing these plots for each of the individual sections in the data set.

**Text Output**

The program outputs a series of data files which display the solution in different ways, and have different uses. The user specifies the names for each of these files, as well as the directory they will be stored in. The program has a set of default names for these files, and will use the directory the input data is stored in as a default, there may be no need to ever change these. Please see the output file descriptions that follow for details.
2.) Input File Formats

The program requires three input files

- the event data file
- a section dictionary
- a taxa dictionary

And one optional input, the best solution determined the program to date, which is an output file created by the program starting from a random solution, or a previous best solution. This is called the grand score matrix file.

When you start the program, you will have to specify these files using the buttons in the “Input Files” section of the control panel. The event data file, section dictionary and taxa dictionary all follow the CONOP (version 9) input file format.

I strongly suggest obtaining the Riley data set examples, and examining these files in Notepad or TextEdit. HA does not check file formats, so it will not tolerate input file errors. Running new files first in CONOP may help trap errors.

Event file (Data File)

This file specifies the events (FADs and LADs) that make up the input data for the problem. Each line specifies a specific event, there is no header line in these data files.

The columns are as follows

<table>
<thead>
<tr>
<th>Column</th>
<th>Contains</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Species ID number</td>
</tr>
<tr>
<td>2</td>
<td>Event Code (1=FAD, 2=LAD)</td>
</tr>
<tr>
<td>3</td>
<td>Section ID number</td>
</tr>
<tr>
<td>4</td>
<td>Horizon Height (in cm, m, ft, etc, as per measurements)</td>
</tr>
<tr>
<td>5</td>
<td>Horizon Ordinal position from the lowest position in the section, including only Horizons with events on them in the ordination</td>
</tr>
<tr>
<td>6</td>
<td>Event code again- allows for differential weight</td>
</tr>
<tr>
<td>7</td>
<td>CONOP weight, not used in HA, set this to 1</td>
</tr>
<tr>
<td>8</td>
<td>CONOP weight, not used in HA, set this to 1</td>
</tr>
</tbody>
</table>
Example fragment of the Riley problem input file, RILE7X62.dat

```
1 1 1 1446.00 6 1 1.00 1.00
1 1 5 1085.00 1 1 1.00 1.00
1 2 1 1446.00 6 2 1.00 1.00
1 2 5 1085.00 1 2 1.00 1.00
2 1 2 1247.00 1 1 1.00 1.00
2 2 2 1252.00 2 2 1.00 1.00
3 1 2 1252.00 2 1 1.00 1.00
3 1 6 1279.00 1 1 1.00 1.00
3 2 2 1341.00 3 2 1.00 1.00
3 2 6 1279.00 1 2 1.00 1.00
4 1 1 1222.00 1 1 1.00 1.00
```

The last line of this fragment indicates the FAD of species 4 was seen in section 1 at a horizon height of 1222, which was the 1st horizon in the section. The line above this indicates the LAD of species 3 was seen in section 6 at a height of 1279, which was the first horizon of section 6.

The taxon dictionary list the taxa names, ID numbers and short versions of the taxa names.

Each taxa occupies one line, the taxa ID number appears first, followed by the short form of the name in single quotes, then the full name also in single quotes.
Example file fragment from the Riley Data set, from the file RILEY62.evt

1 'Hol.tene' 'Holcacephalus tenerus'
2 'Mod.owen' 'Modocia cf.oweni'
3 'Bol.well' 'Bolaspidella wellsvillensis'
4 'Bol.burn' 'Bolaspidella burnetensis'
5 'Ced.cord' 'Cedarina cordilleras'
6 'Kor.simp' 'Kormagnostus simplex'
7 'Ced.eury' 'Cedarina eurycheilos'

Section Dictionary

Each section is a single line in this file. The section ID is first, followed by an abbreviated, three-letter form of the section name in single quotes, then the ordinal positions in reverse order (??? not used in HA, see CONOP manual), then the full section name in single quotes and then another CONOP control code, just leave this as 1 when using HA.

Example file from the Riley Data set, this is the file RILEY7.SCT

1 'Mor' 6 'Morgan Creek' 1
2 'Wht' 5 'White Creek' 1
3 'Jms' 4 'James River' 1
4 'Lln' 3 'Little Llano River' 1
5 'Lio' 2 'Lion Mountain' 1
6 'Pon' 1 'Pontotol' 1
7 'Str' 7 'Streeter' 1

The Grand Score Matrix

If you leave the Starting Solution in HA as “Rand”, it will start with a random solution. If you want to start with a previously obtained solution, so you can continue refining that solution, you will need to select a grand capture matrix produced by HA as the starting solution.

The grand capture matrix file has four rows, and a number of columns equal to the total number of all horizons. The entry on the first line (row) is the range extension of the solution. The entries on the next
line indicate the section number, the values on the third line are the horizon number within the section, and the fourth line is the score of that particular horizon in the composite. You probably won’t need to create one of these on your own, HA will generate them. If you want to see the format, run HA using a random starting point on the Riley data set and open the resulting grand capture matrix in excel, using “space” as a delimiter.
3.) Running the Program

Download and install the HA program from the website, following the instructions on the site. Also obtain the Riley example files and try running those first. Select the RILEY7x62.DAT file as the input data file, and the Riley7.SCT and RILEY62.EVT files as the section listing (dictionary) and Taxon Listing (dictionary). Leave the Starting Solution as random.

The program will set the directory you have the input data file in as the default location for the output files. You can choose to alter this if you prefer.

The program should look something like what you see below. Use the default settings to start with.

![Program Interface]

**Figure 6:** HA set up to search for solutions to the Riley data set.

Try running the program using the “Run” button.
Once it has run, use the “Plots” menu button to display the plots shown in Figures 2-5. Plots may be copied to the clipboard or save to a file, although these functions are slightly buggy as of 8-22-2012.

Once you have run the program once, you may want to select the grand capture history created as the starting point, instead of working from random starting points. The program will create time stamped versions of the starting grand capture matrix every time it is run, which will allow you to back-up to a prior version of your best solution if something goes wrong. You may want to periodically delete excess files.

Experiment with the starting temperature and cooling rate to get a feel for how they influence the search for solutions. Likewise alter the number of temperature steps (the outer loop) and the number of trials at each temperature (the inner loop) to see the effects these have.

Also try running the Zhang and Plotnick example set.
4 Graphical Output

-----This section of the manual is currently incomplete- please check for later versions

The graphical outputs are accessed through the plot menu, and should be largely self-explanatory. See Figures 2-6 of this manual to see the figures produced.
5.) Output Files

The HA program produces a variety of output files:

soln.out-

implied_time_matrix.txt

grand_score_matrix.txt

soln_CONOP_format.txt

CONOP_Height.txt

Species_Height.txt

implied_time_annealed.txt

_The following are limited descriptions of the file contents—more to come later_

_soln_CONOP_format.txt_- This is the ordination of FADs and LADs only in the CONOP format. This may be used as input file solution in CONOP allowing for comparisons of HA and CONOP solutions. The first entry on each row is the Taxa number, followed by the event code (1=FAD, 2=LAD) and then the ordinal position in the composite. Ties are not allowed, so each event has a unique position. There is no header line.

_CONOP_Height.txt_- This is a version of the CONOP style output file discussed above, but with the ordinal position of the event in the composite added as a fourth column. This last column does allow for ties in taxa positions. This version will not load into CONOP.

_soln.out_- This file contains a matrix, whose columns are horizons and whose rows are taxa. The first row is the ordinal position of the horizon in the composite, the second row is the section number, the third row is the horizon number, the fourth is an estimated placement of the horizon in the composite, the fifth row is the horizon’s score. Rows after the 5th start with the taxa number, followed by a series of presence/absence codes. 1 indicates presence, 0-absence and -1 means not seen in section. This output is thus a structure “capture history” in CMR style terms.

_grand_score_matrix.txt_- See the discussion of this file and its contents in the earlier discussion of input file formats. This file indicates the position in the composite of all horizons, and so indicates the line of correlation.
6.) References


Sadler, P.M. 1999: Constrained optimisation approaches to the stratigraphic correlation and seriation problems: a user’s guide and reference manual to the CONOP program family: Riverside, California, Peter M. Sadler.


