9. Cluster Analysis

Background

A new feature in Kintecus V3.9 is the ability to perform complex hierarchical cluster analysis[29] on temporal concentration profiles of the network with/without experimentally obtained temporal concentration profiles. Hierarchical cluster analysis in Kintecus has the ability to group related and unrelated parts of temporal concentration profiles in a meaningful, quantitative way. This grouping allows a user to clearly see patterns that were initially indiscernible or hidden.

"Why should I care about that?"

Cluster Analysis can significantly help in answering questions:

- Analyze ALL species and determine which species or groups of species (or subgroups, etc.) are positively, zero or negatively correlated to each other and with other groups/species in either a pictorially or numeric output or both. This allows one to answer questions such as:
- What concentrations of E and S cause EIS to positively (or negatively or zero) correlate with EI ? or with ES? Or both? Etc. etc. For combustion, you can now finally answer that question that has been bugging you since you were seven years old: If I combinatorially scan H2, O2 and temperature over a wide range, do the O and OH radicals always positively correlate and also do they positively correlate (or negatively or zero) correlate with other species?
- Determine which species in one's experimentally determined concentration profile are positively, zero or negatively correlated with all the modeled species temporal profiles. Again, this can be grouped into a pictorially or numeric output.
- Which species' concentration profiles tend to stay in the same range of concentration values?
- There are other questions one can answer utilizing the myriad of cluster techniques Kintecus provides that the author has not fully examined…

Cluster analysis becomes quite a powerful tool when coupled with combinatorial scanning (see the Scanning Section to see how to set up a Combinatorial Scan).

Implementation

Hierarchical Cluster analysis is accomplished in Kintecus by providing the "-cluster" flag on the Kintecus command line:

>kintecus **–cluster**

The "-cluster" flag has many features that the user may want to change:

-CLUSTER{:A:B:C:D:E:F:G:H:I:J:K}

All options a, b, c, d, e, f, g, h, i, j, and k are all optional, but when setting options b to k all preceding options must be specified. All options have a default setting which can be specified with the letter "D" or "d" (most Kintecus switches have this feature). Here is a breakdown of all the options (defaults are listed in brackets, [] or by entering the letter "D" for an option indicates to use default value.).

a) Option "a" determines the type of cluster analysis to perform. Currently there is only one: Hierarchical Cluster Analysis specified by the number "1" in place of "a". Future versions might incorporate K-means clustering. Adding 100 to this value will state to Kintecus to skip the computation of the concentration/temperature/volume/etc outputs and to immediately start the clustering. The output concentration file, CONC.TXT is assumed to be in the same directory or the file associated with the "-OUT:filename" switch exists and is readable.

b) Option "b" determines the type of computation implemented in the similarity/dissimilarity matrix computation. Option "b" must be between zero to eight (0-8) inclusively. The most common dissimilarity methods you will most likely use is the Euclidean distance (set option "b" to zero, "0") and other the popular method is Correlation (r^2) (six, "6" for option "b"). There are other methods for computing the similarity/dissimilarity matrix:

Option "b" Values	Similarity/Dissimilarity Method	What is it?
[0]	**Sqrt(Sum(distance between $[c_1]_t$ and $[c_2]_t$)^2))**	**Euclidean distance between concentration profiles**
1	Sum(ABS(distance between $[c_1]_t$ and $[c_2]_t$))	Norm of L1 of concentration profiles
2	Max(distance between $[c_1]_t$ and $[c_2]_t$)	Norm of L-at-infinite
3	Mahalanobis distance	Mahalanobis distance
4	Cos(theta between vectors $[c_1]_t$ and $[c_2]_t$)	Dot product between concentration profiles
5	Just theta between vectors $[c_1]_t$ and $[c_2]_t$	Angle
6	**Rho from a plot of $[c_1]_t$ and $[c_2]_t$**	**r, Correlation (-1.0 to +1.0)**
7	ABS of Rho	r, correlation (0.0 to +1.0)
8	Count the number of times values in $[c_1]_t$ match $[c_2]_t$	Matching count

Table 19. Types of ways to compute the similarity matrix for the hierarchical clustering method. Note that the scaling option, "e" has no affect on methods 3-8.

c) Option "c" specifies the type of hierarchical cluster analysis method and can have values from zero to four (0-4). You will most likely set option "c" to zero, "0" in almost all cases. The author hasn't been able to find appropriate uses for the other hierarchical cluster analysis methods.

Option "c" Values	Hierarchical Cluster Analysis Method
[0]	Minimum distance between cluster (by far the most common method)
1	Maximum distance
2	Average distance within clusters
3	Average distance between clusters
4	Ward's method (option "b" should be set to zero=Euclidean distances)

Table 20. Type of hierarchical cluster analysis methods available.

d) Option "d" specifies whether to perform a transformation on the similarity matrix and can have values from 0-2. Option "d" is primarily utilized for Similarity/Dissimilarity Methods 6 and 7 (correlation).

Option "d" Values	Transformation Method
[0]	No transformation
1	Multiply by –1.0 (not really used that often)
2	Convert the values in the similarity/dissimilarity matrix into distances by obtaining the reciprocal of the absolute value: 1/|distance|. This transformation is usually used with Option "b" methods 6 and 7 (correlation matrices).

e) Option "e" specifies whether to scale the data before the calculation of the similarity matrix. Typically it is set at zero, "0", to perform no scaling.

Option "e" values	Type of Scaling
[0]	No scaling, leave concentration data alone.
1	Scale each species concentration temporal profile by the standard deviation of the same concentration temporal profile.
2	Scale each species concentration temporal profile by the range of the concentration temporal profile.

f) Option “f” specifies the number of clusters to form and is NOT USED in the Hierarchical Cluster Analysis Method. It should be left at the numeric value of “2”.

Options g-k are primarily intended in the printing of the cluster tree:

g) Option “g” specifies the page width in characters of the cluster tree output [101].

h) Option “h” specifies the type of cluster printout and ranges from 1 to 3.

i) Option “i” specifies the number of lines printed before each node and can have values from 1 to 10 (one line is the default). You shouldn’t have to change this default value.

j) Option “j” specifies the subtree printing specification and can range from zero [0] up to 100. Zero is the default. You shouldn’t have to change this default value.

k) The final option “k” specifies the number of horizontal slices of tree to print and can range from 1 to 10. The default value is one, “1”. You shouldn’t have to change this default value.

Entering only the “-cluster” switch on the command line is equivalent to the following cluster switch options:

-cluster:1:0:0:0:0:2:178:1:1:0:1

So Now What?

Although the default settings for the “-cluster” switch will always work, it only provides information on which temporal concentration profiles are closest to each other and other groups. You will very likely find the cluster switch with the following settings as the **most useful**:

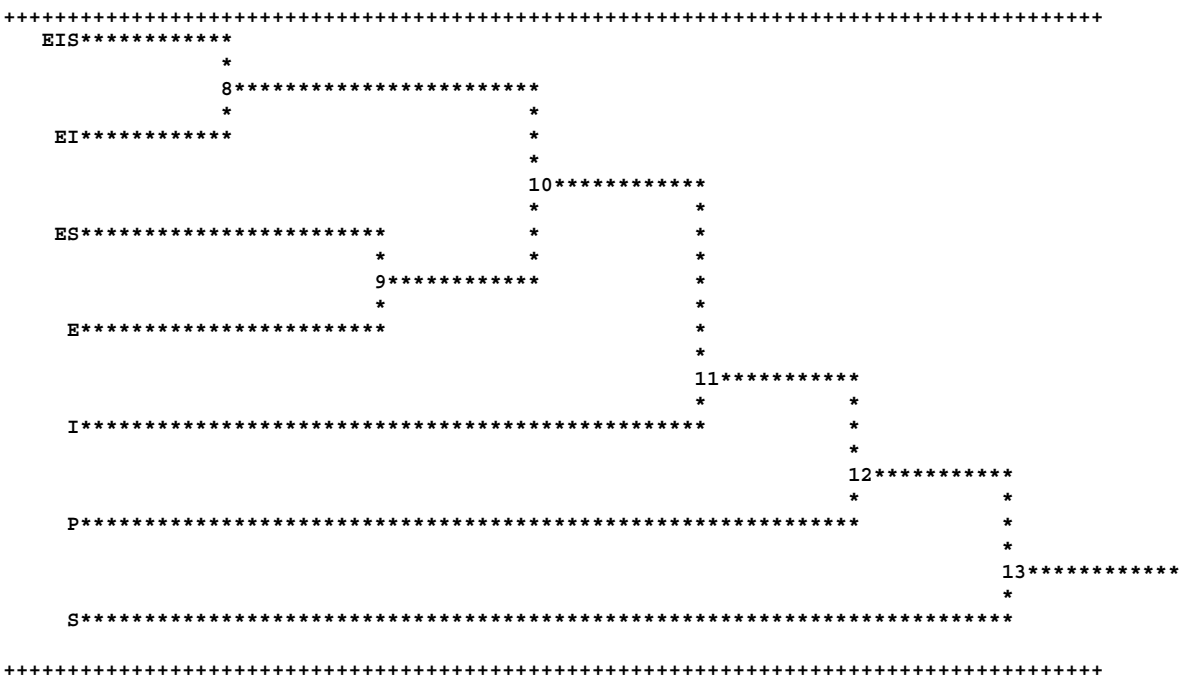
“-cluster:1:6:0:2”

which implements the very useful clustering technique of correlation, r^2, for the similarity matrix. Please see the sample clustering techniques 2 and 3 below on the next page. This very useful correlation technique was not set as the default because occasionally it fails due to some species’ temporal concentration profiles have no change and are near zero. This can sometimes cause this clustering technique based on correlation to fail. A simple correction to this is to NOT DISPLAY the output of the species (or species’) which is not changing and is near zero (go to the Species Description Spreadsheet and set the “Display Species?” from Yes to No).

Sample Plots/Output

Cluster Analysis: Sample 1

The following is a simple sample and is mainly for pedagogical purposes. More "real-world" examples follow this. The first example is utilizes the default setting implied by the "-cluster" switch (Hierarchical Cluster analysis using Euclidean distances for the similarity matrix and a minimum grouping between clusters). Part of the cluster.txt output file contains the cluster graph after clicking RUN on the **Enzyme_Cluster_Analysis.xls** Kintecus-Excel spreadsheet:



We can see that enzymes EIS and EI and ES and E are both in their own clusters, with I, P and S as the "outsiders". Why this grouping? Remembering that cluster analysis builds on the similarity matrix and we are using Euclidean distances for the similarity matrix, the species with the closest concentration temporal profiles will be grouped. A log temporal concentration plot (shown below) of this run shows that indeed EIS, EI are quite close together and all by themselves. Species E and ES show a similar pattern hence their own little cluster. Species I, P and S are in their own branches alone, but species I is closer to the EIS, EI and ES, E clusters than P or S.

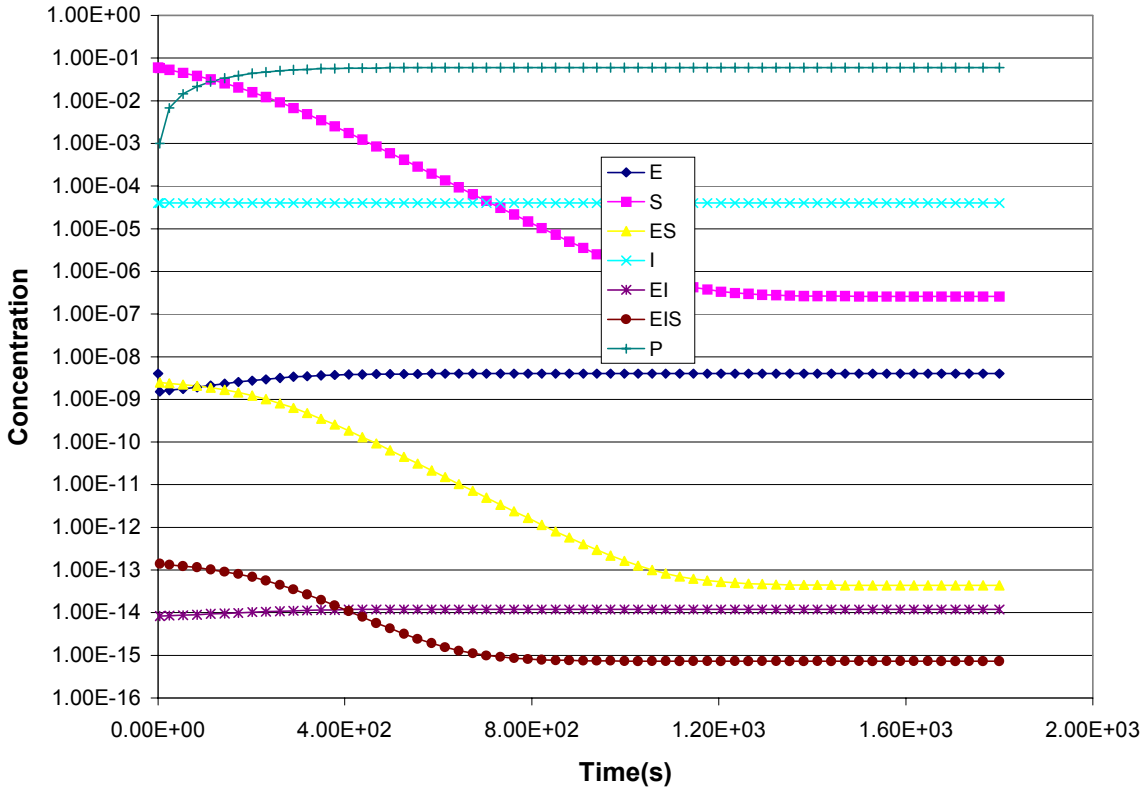
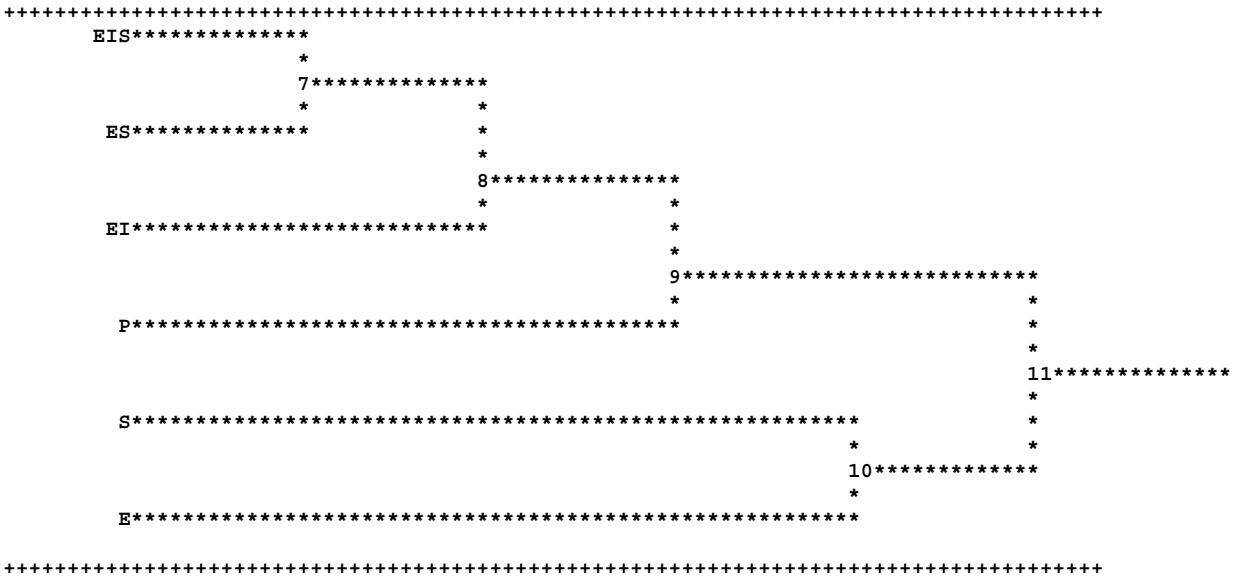


Figure 24. Temporal Concentrations plot of the Sample Enzyme Model. (y-axis is on a logorithmic scale)

Cluster Analysis: Sample 2

Lets see a more useful clustering technique in that we shall use correlation, r^2, for the similarity matrix and we will keep the same minimum clustering technique. Note that we have to convert the r^2's to distances. This is accomplished by transforming the r^2's into distances by taking the reciprocal (1/r^2) which can be automatically done by setting Option "d" (Transformation Method)to the value 2. We arrive at the following cluster switch which will accomplish all this: **"-cluster:1:6:0:2"** . There is one small problem with this distance measure, since the concentration of species "I" never changes, the r^2's with plotting species "I" will always be zero. This will yield infinite distances for the transformation method and the clustering method will fail. To correct this, the output for species "I" has been turned off by selecting "No" for the "Display Species" field in the "Species Description Spreadsheet". Here is the output after clicking RUN on the **Enzyme_Cluster_Analysis_2.xls** Kintecus-Excel spreadsheet:



with the corresponding similarity matrix (located inside the same cluster.txt file with the untransformed r^2's rounded to three decimal places):

```
Transposed Similarity/Dissimilarity Matrix
         2       3       4       5       6
 1   0.296   0.000   0.000   0.000   0.000
 2           0.000   0.000   0.000   0.000
 3                   0.945   1.000   0.296
 4                           0.944   0.591
 5                                   0.293
```

and the numeric id's for the species that appear in the similarity matrix:

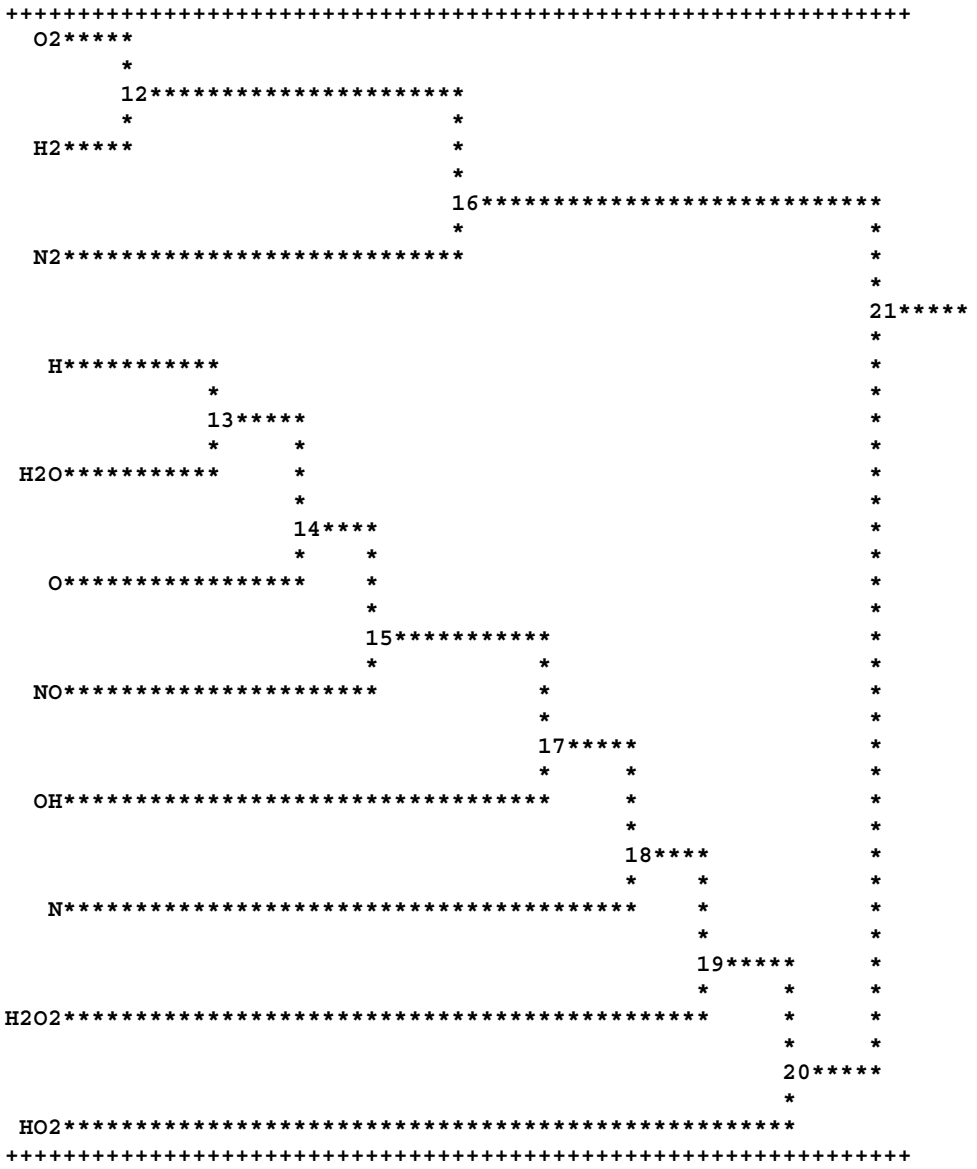
	1	2	3	4	5	6
Time(s)	E	S	ES	EI	EIS	P

Table 21. Numeric id's for the species that appear in the similarity matrix.

We can immediately see why EIS and ES (5 and 3 respectively in the matrix) are clustered together as they have perfect correlation of 1.000 (a plot of EIS versus ES will yield a linear equation y=mx+b with perfect correlation: r^2=1.00). We can also see why EI (4 in the similarity matrix) has been clustered with EIS and ES because EI has an excellent correlation (0.94) with EIS. Species P has some good correlation with Species' ES, EI and EIS (r^2's=0.296, 0.591 and 0.293) so species "P" has been clustered with ES, EI and EIS. The other cluster only has species S and E as they are only correlated with each other (r^2=0.296).

Cluster Analysis: Sample 3

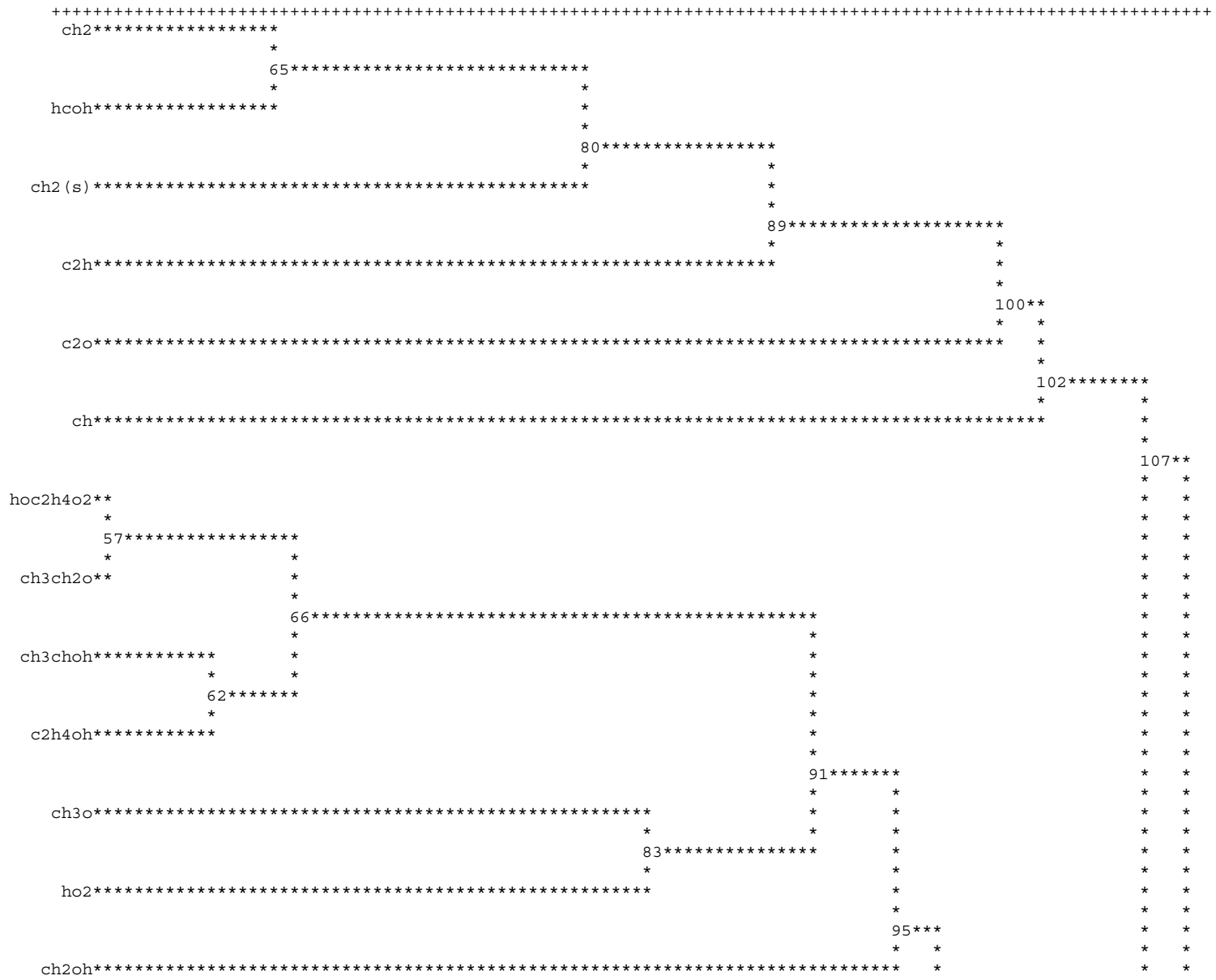
Let's try another Hierarchical Cluster analysis utilizing correlation distances for the similarity matrix and a minimum grouping between clusters on the O_2 and H_2 isobaric (constant pressure) combustion model. Here is the output after clicking RUN on the **Combustion_O2_H2_cluster.xls** Kintecus-Excel spreadsheet:

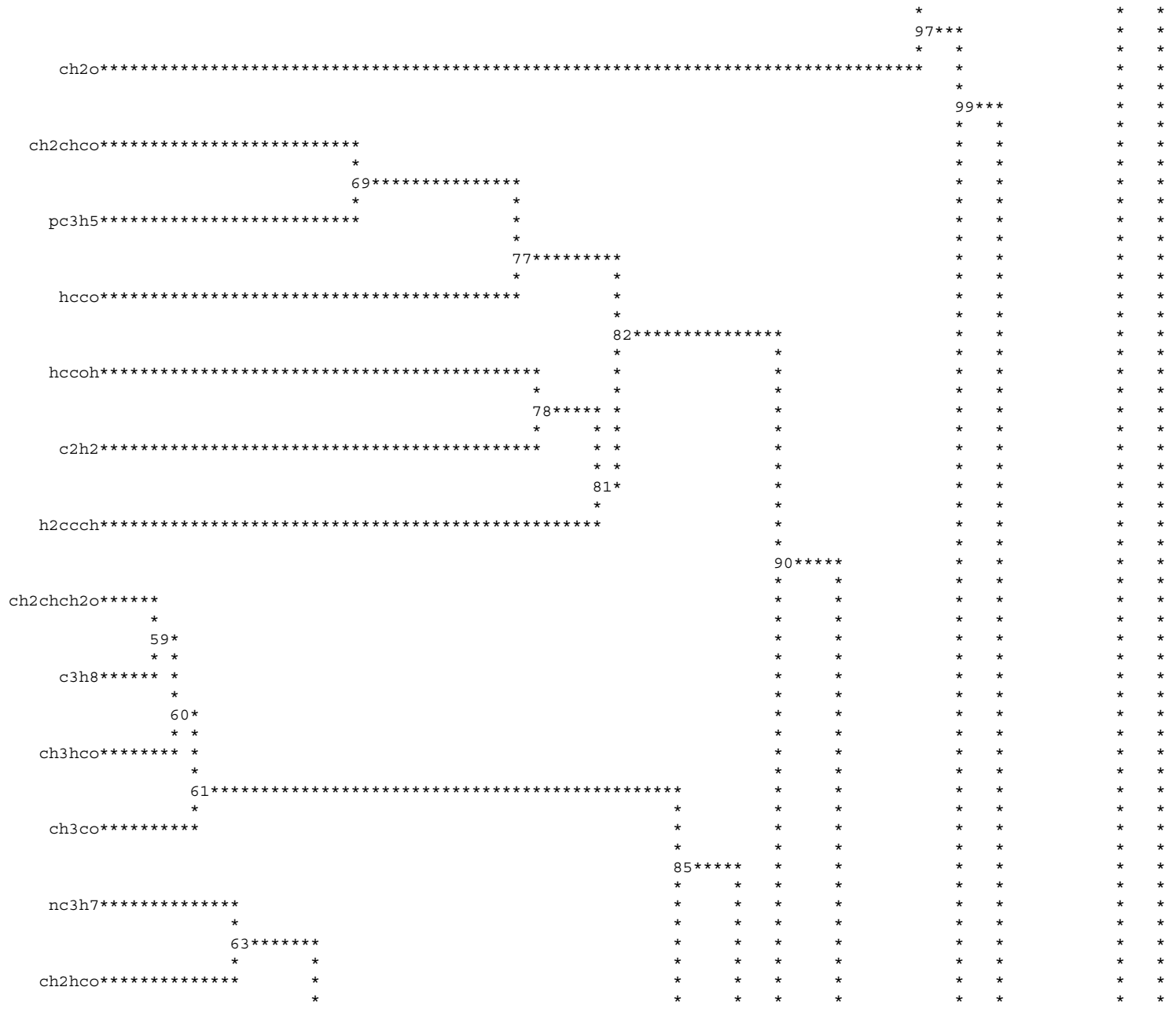


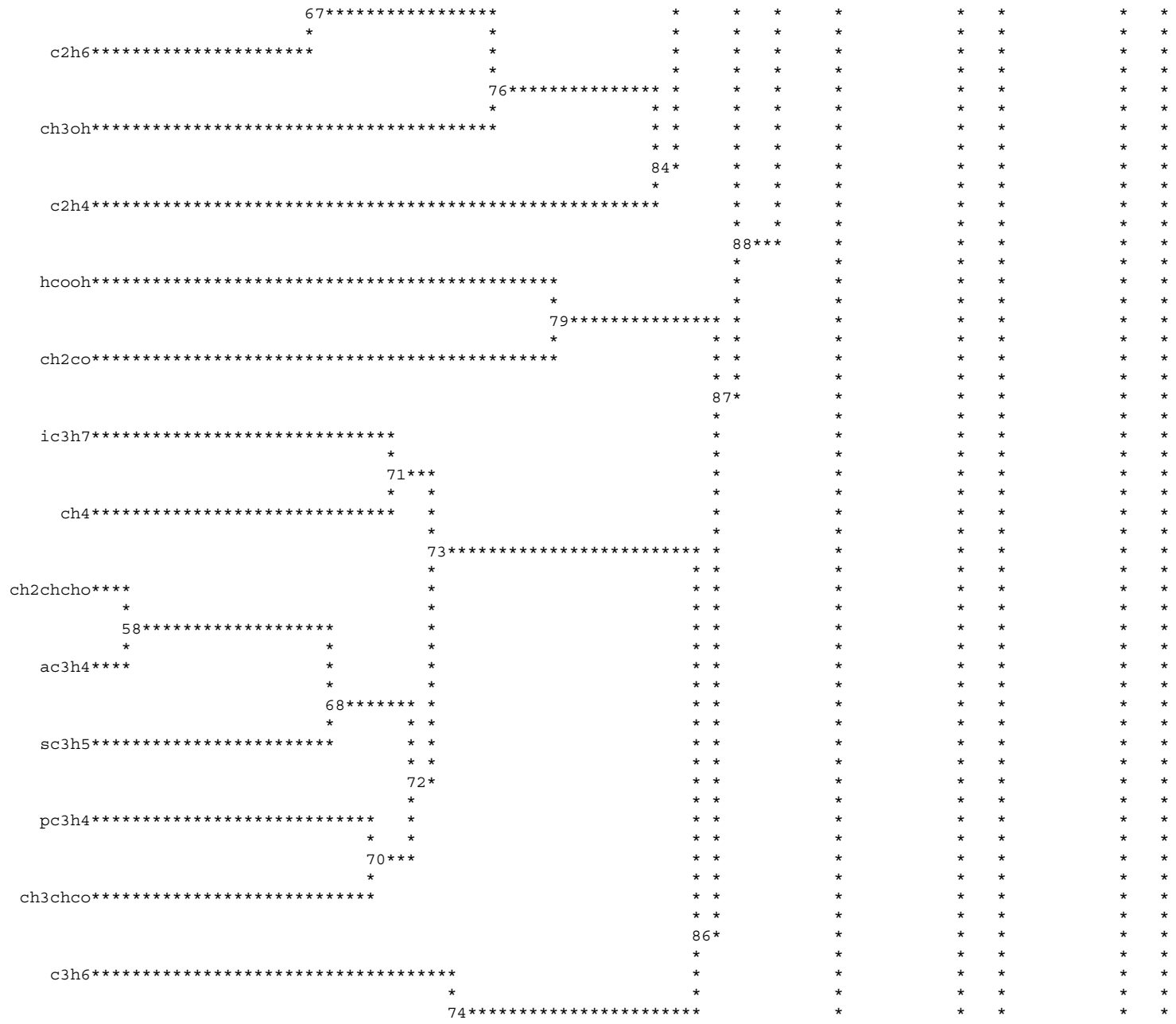
We can see that this cluster makes perfect sense in correlation because all three species O_2, N_2 and H_2 all decrease in concentration and reach equilibrium nearly at the exact time so they are in their "own" cluster, while at the same time all the other species slowly grow in concentration so they are in a separate cluster.

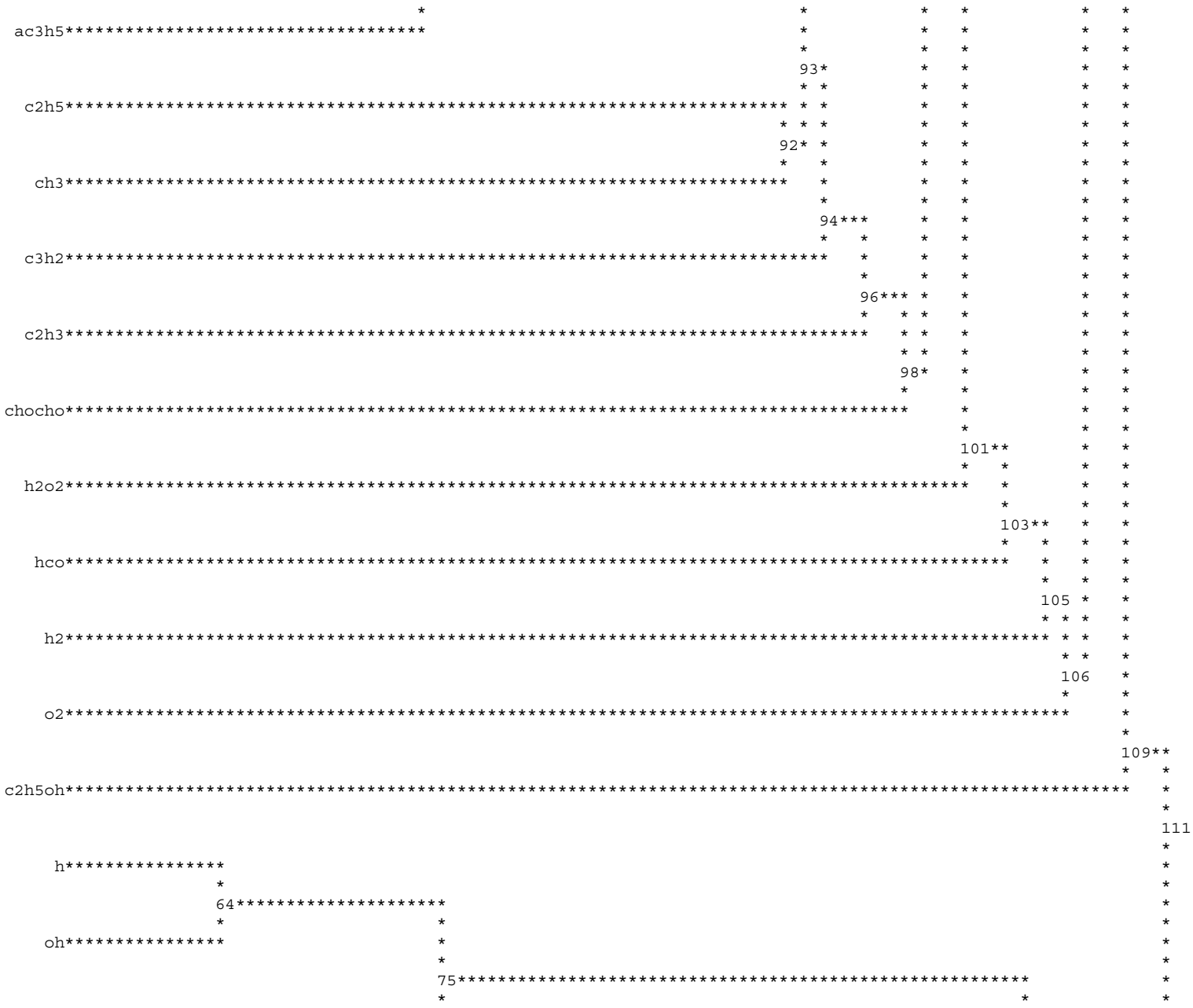
Cluster Analysis: Sample 4

Let's try another Hierarchical Cluster analysis utilizing correlation distances for the similarity matrix and a minimum grouping between clusters on the combustion of ethanol under isochoric (constant volume) conditions. The cluster switch: **"-cluster:1:6:0:2:d:d:127"** was added on the Kintecus command line for the Kintecus-Excel **Ethanol_Combustion.xls** model and executed. The pages that follow are the output from the "cluster.txt" file. It is left as an exercise to the reader to determine why there are three main families of clusters present.


ch2
65
hcoh
80
ch2(s)
89
c2h
100
c2o
102
ch
107
hoc2h4o2
57
ch3ch2o
66
ch3choh
62
c2h4oh
91
ch3o
83
ho2
95
ch2oh


ch2o
ch2chco
pc3h5
hcco
hccoh
c2h2
h2ccch
ch2chch2o
c3h8
ch3hco
ch3co
nc3h7
ch2hco
97
99
69
77
82
78
81
90
59
60
61
85
63


c2h6
ch3oh
c2h4
hcooh
ch2co
ic3h7
ch4
ch2chcho
ac3h4
sc3h5
pc3h4
ch3chco
c3h6
67
76
84
79
71
73
58
68
72
70
88
87
86
74


ac3h5
c2h5
ch3
c3h2
c2h3
chocho
h2o2
hco
h2
o2
c2h5oh
h
oh
92
93
94
96
98
101
103
105
106
109
111
64
75


o
104
co2
108
h2o
110
co

10. Excel Tricks

Kintecus Excel Trick 1:
PUTTING YOUR "STUFF" INTO KINTECUS WORKSHEETS

Although you cannot add extra fields and "stuff" in the MODEL.DAT, SPECIES.DAT and PARM.DAT text files, YOU CAN add extra columns of data (numbers, formulae, entire Works of Shakespeare) to any of those Excel Spreadsheets. For example, in the Excel Species Description Worksheet, numbers and/or data and/or links can be entered in column I (column #9) onto the very ending column IV (column #255). These extra columns will NOT BE WRITTEN into the SPECIES.DAT file. Some Kintecus users have already noticed this and use all those extra columns from converting units to "Goal Seeking" to reading data from a remote Omega TCP/IP device. As a side note, you CAN add drawings, movies, Excel comments, text boxes, graphs, etc. ANYWHERE in any Kintecus Worksheet as floating objects.

Kintecus Excel Trick 2

THE MIGHTY AutoFilter

Since Kintecus has a unique way of defining entire reactions with one row; this allows for some very interesting possibilities with Excel's very powerful AutoFilter. One can perform anything from deleting commented out lines, to removing all bimolecular reactions that referenced Miller 1992. You can view a sample video animation of the procedures below if you go to **www.kintecus.org** . Here are a few demonstrations with the Ethanol Combustion Kintecus Workbook:

<A> Getting Rid of All Those Comments *****

A.1) Click on the MODEL worksheet in the Ethanol Combustion Workbook.
A.2) Turn on the mighty AutoFilter (select Data => Filter => Autofilter). You should see five (5) select boxes (they look like upside triangles) hovering on row 2 over columns A to E. Reselecting