

User manual of MK

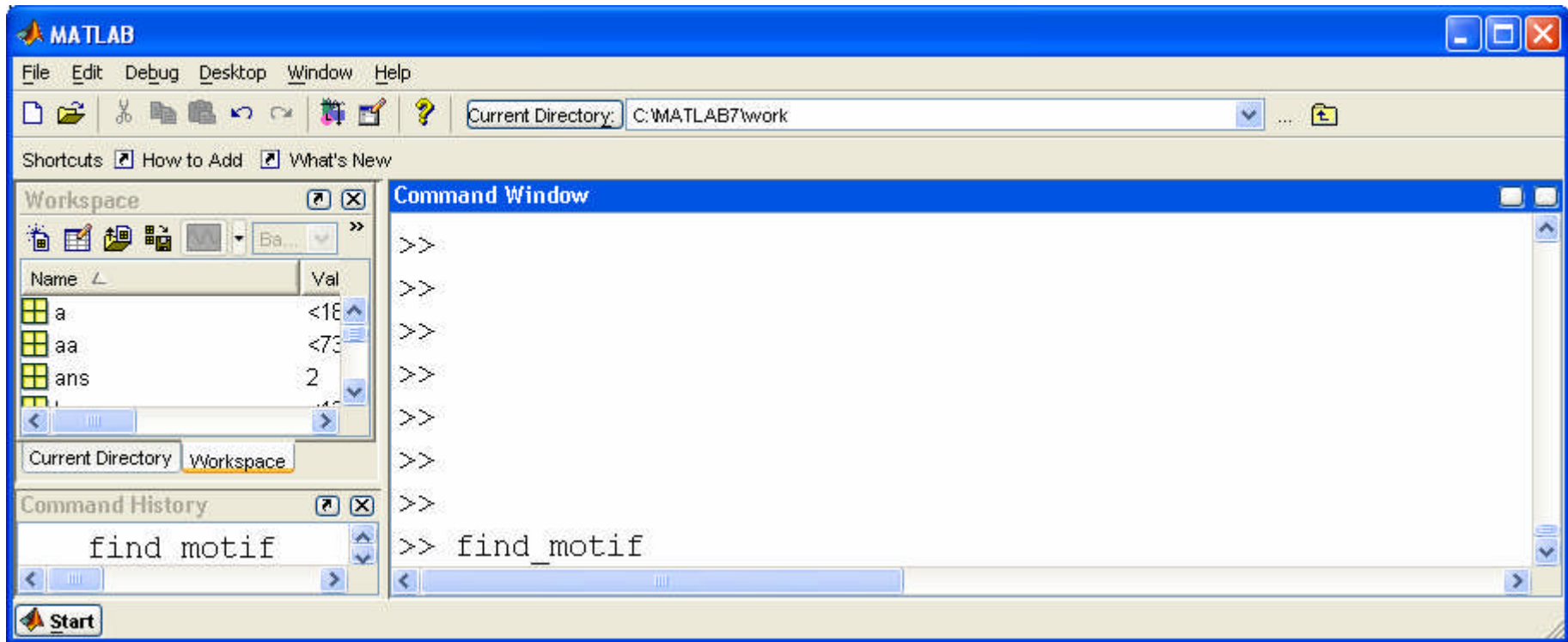
Prepared by
Abdullah Mueen and Eamonn
Keogh

The *MK* Code

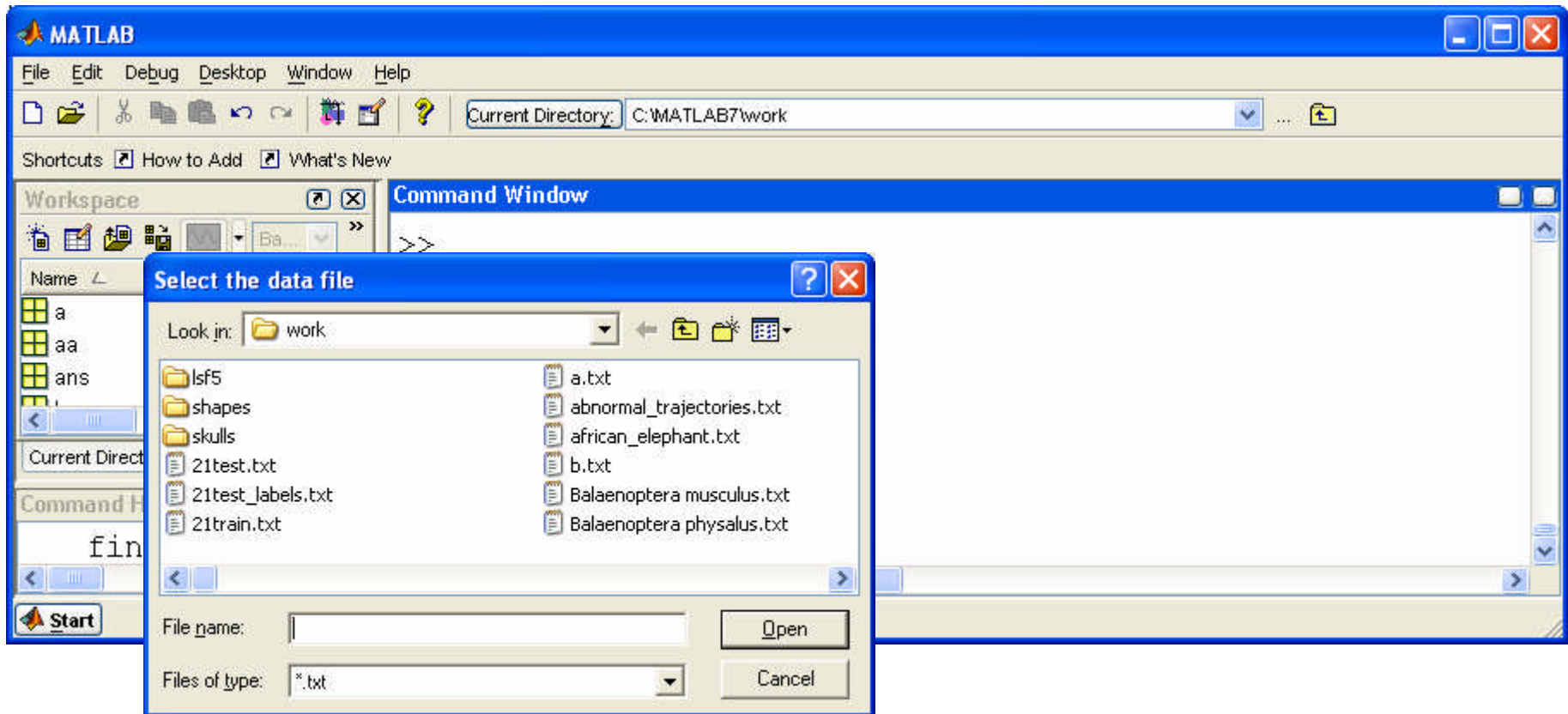
- *MK* is coded in C. It is compiled by gcc and executable in any platform. The executable file is also included with this package.
- The code dynamically allocates memory whenever necessary. If it fails to allocate memory it prints an error message in the stdout and stops execution.
- To the best of our knowledge it is bug free as long as the input is in the correct format. Please report bugs if you encounter any.

A Sample Run

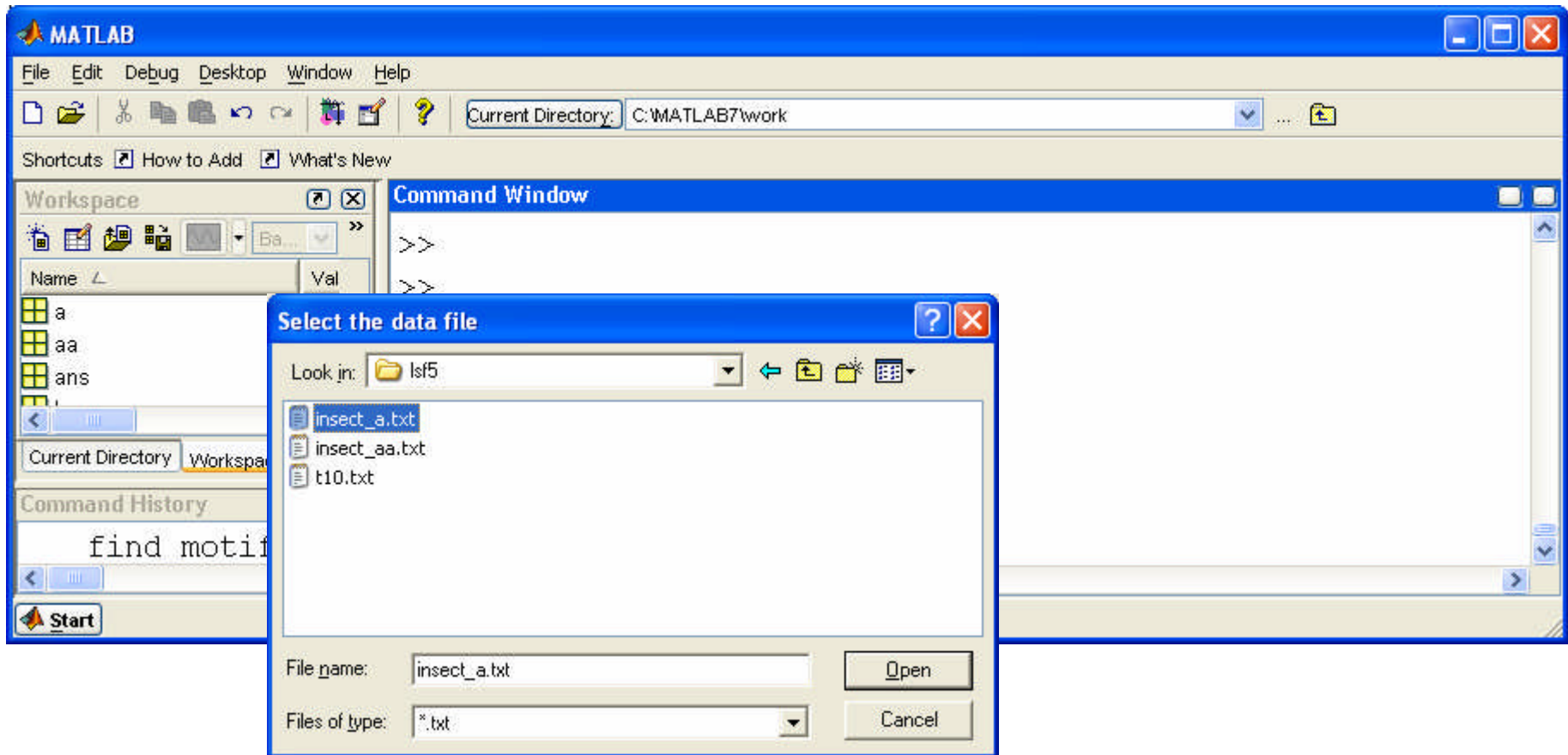
- Please read this sample run, then redo it on your own machine (the sample data is provided) before doing anything else.
- Put all the files in the matlab work directory, then boot up matlab..



From the matlab window, type find_motif



Point to the time series you want to explore



In this case, insect_a.txt

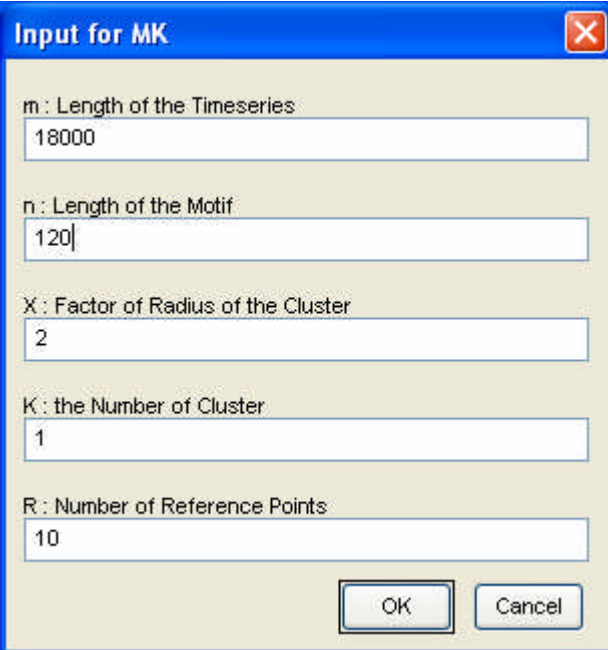
This number should be less than or equal to the full length of the time series. If it is less than the full length, the code truncates off the remainder

The length of the motifs you wish to find.

X must be at least 1. As it gets larger, many more time series tend to be in the motif cluster. We suggest you start small (say 2) and increase it a little (say to 3) in the next run.

The number of distinct motifs to find. Use 1 for the first few times.

Number of reference points. We strongly suggest you use the default value of 10.



The screenshot shows a dialog box titled "Input for MK" with a close button (X) in the top right corner. It contains five input fields, each with a label and a value:

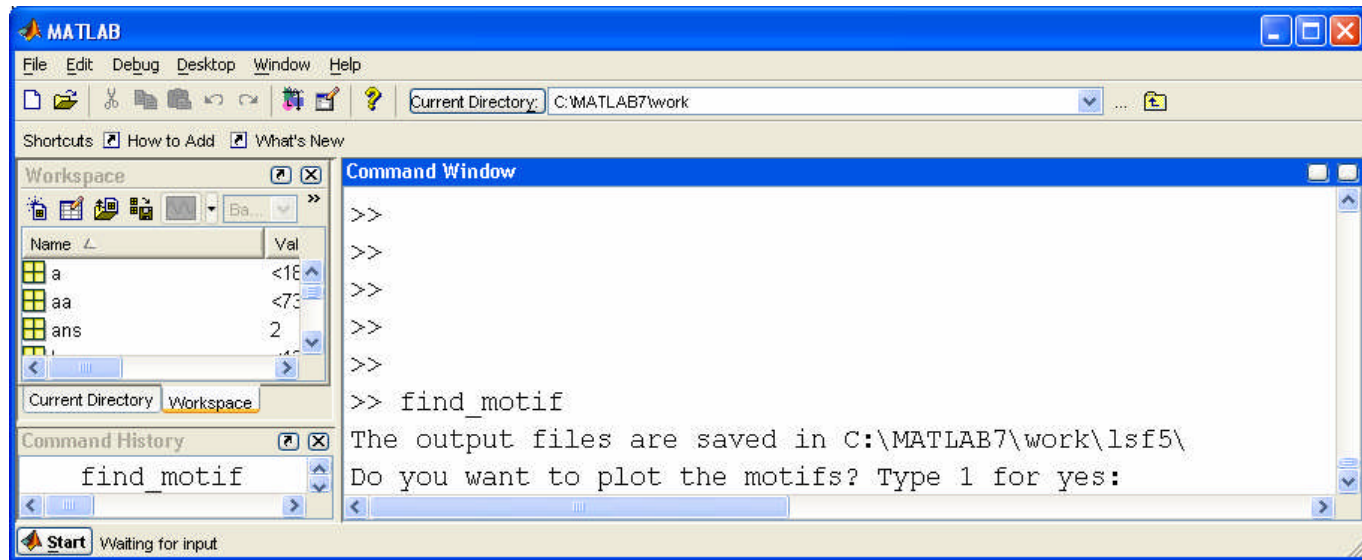
- m : Length of the Timeseries: 18000
- n : Length of the Motif: 120
- X : Factor of Radius of the Cluster: 2
- K : the Number of Cluster: 1
- R : Number of Reference Points: 10

At the bottom right of the dialog box are two buttons: "OK" and "Cancel".

When you are ready, click OK

What's happening?

- The code is running. If you have less than a 20,000 length time series, and a motif length less than 500, this should be a few seconds.
- If you have very long time series 50,000+ or very long motifs 1000+, or very noisy data, this could take minutes.
- When the code is done, you will see...



The code is inviting you to plot the output

The output file is named such that you can tell which experiment it came from: In this case it is..

insect_a_txt_18000_120_2.0_1.txt

The source time series

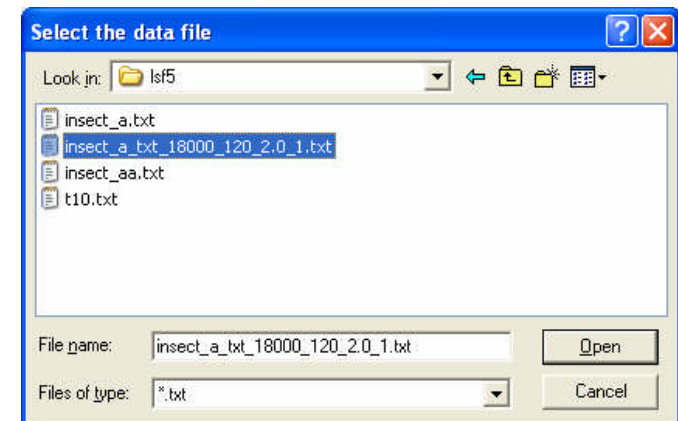
How long of section you looked at

The motif length

The radius

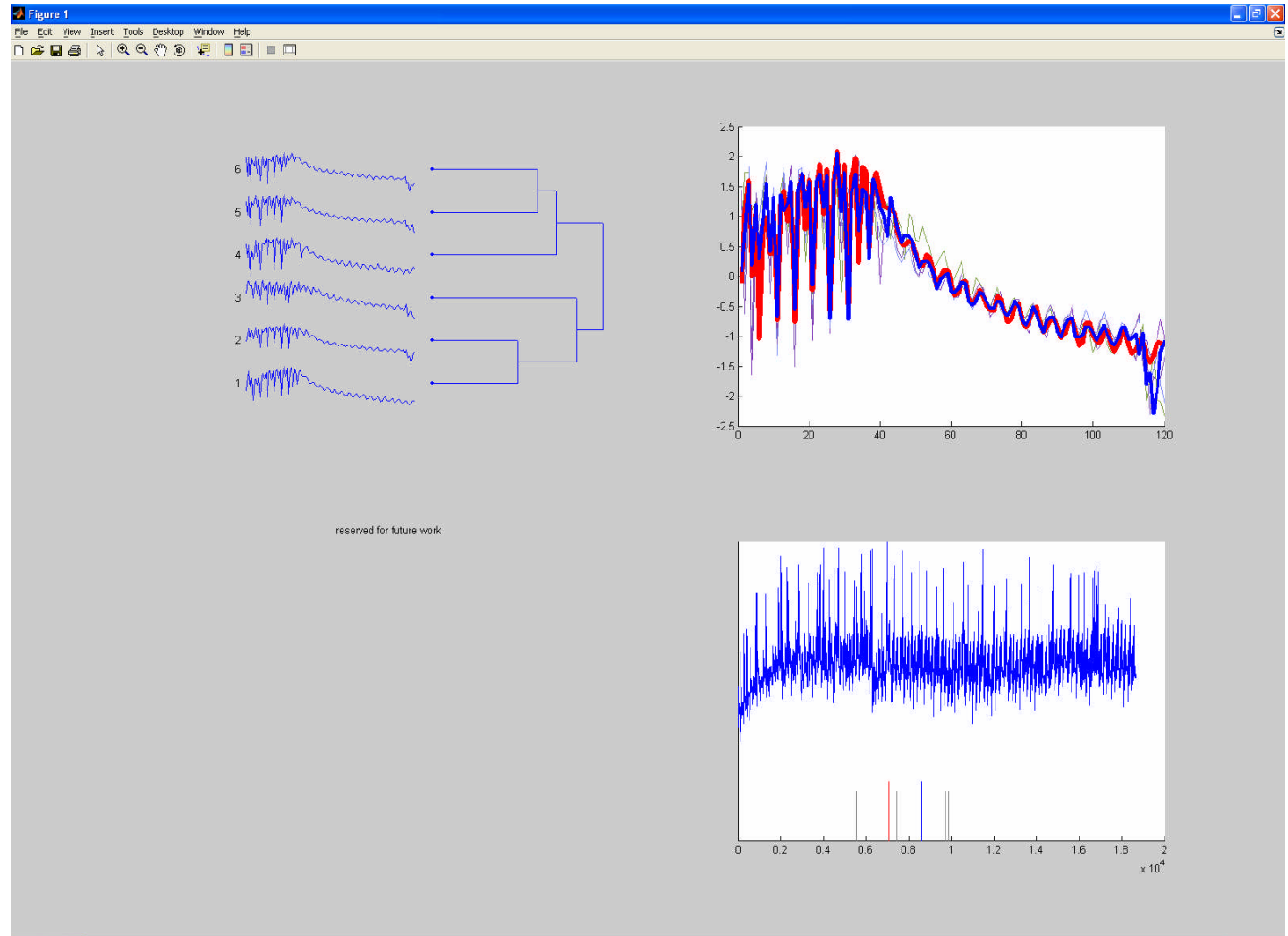
This is the kth motif

Lets say “yes”, and plot the output, a dialogue box appears, and we find the file...



Here is a dendrogram (single linkage) of the motifs discovered. Time series “1” and “2” are the “red” and “blue” time series. If too many motifs are returned, a message “dendrogram suppressed due to size” will appear.

Here are the motifs plotted on top of each other. The two “seed” motifs are in red and blue



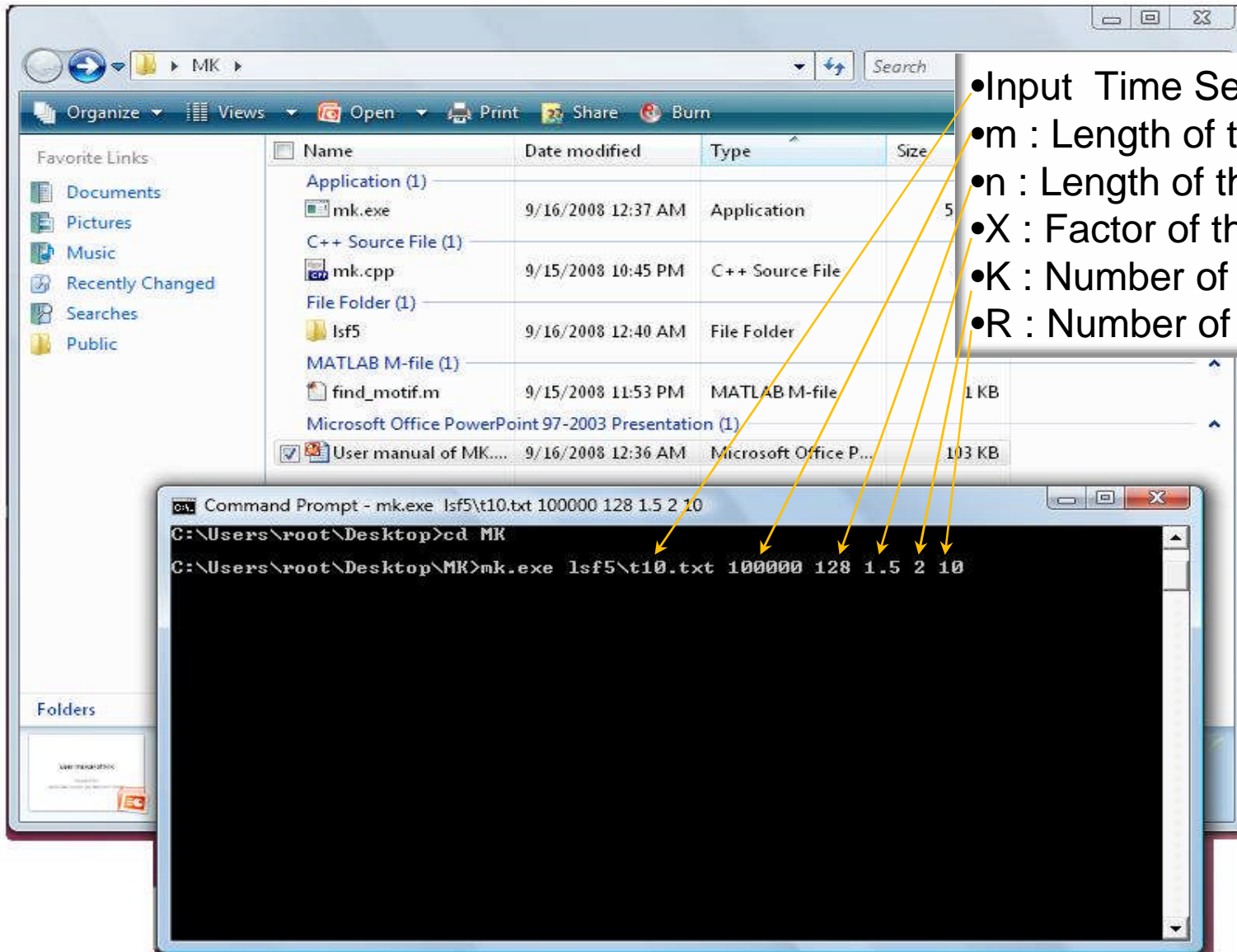
Here are the locations of the motifs for context



Stand Alone Code

- The previous slides show how to use the matlab “wrapper” we wrote for the main motif finding code.
- The main code is in C.
- If you want to, you can call this code directly, the next few slides tell you how.

Input



- Input Time Series File
- m : Length of the Time series
- n : Length of the Motif
- X : Factor of the Cluster Radius
- K : Number of Cluster
- R : Number of Reference Points.

Input (Contd.)

```
t10.txt - WordPad
File Edit View Insert Format Help
Courier New 10 Western
-7.6150635e+003
-7.6918154e+003
-7.6645298e+003
-7.6574150e+003
-7.6471948e+003
-7.6373833e+003
-7.6288066e+003
-7.6239624e+003
-7.6171494e+003
-7.6070859e+003
-7.5931694e+003
-7.5852114e+003
-7.5845767e+003
-7.5812876e+003
-7.5664038e+003
-7.5501704e+003
-7.5411353e+003
-7.5366299e+003
-7.5251660e+003
-7.5063730e+003
-7.4897842e+003
-7.4878247e+003
-7.4906245e+003
-7.4896299e+003
-7.4863696e+003
-7.4901074e+003
-7.4996890e+003
-7.5051108e+003
-7.5040664e+003
-7.5037085e+003
-7.4986958e+003
-7.4922554e+003
-7.4825459e+003
For Help, press F1
```

The input file contains m real numbers representing the Time Series. Numbers Can be separated by space or lines. They can also be in any real number format.

- m and n must be positive integer and $4 < n \ll m$.
- X can be any real number. Default value is 2 and can be omitted.
- The parameters K and R are integers. Default values are 1 and 10 respectively. Can be omitted also. $R \ll m-n$

```
Command Prompt - mk.exe lsf5\t10.txt 100000 128 1.5 2 10
C:\Users\root\Desktop>cd MK
C:\Users\root\Desktop\MK>mk.exe lsf5\t10.txt 100000 128 1.5 2 10
```

Output

- Output will be K files.
- The output files are named by concatenating all the input parameters separated by '_'.
- The last number denotes the rank of the motif cluster.

- Each of them has a set of subsequence time series printed in lines.
- The first number is the location of the subsequence in the original time series.
- The subsequence time series are **z-normalized**.

The screenshot shows a Windows file explorer window displaying a folder named 'MK' containing three text files: 't10.txt', 't10_txt_100000_128_1.5_1.txt', and 't10_txt_100000_128_1.5_2.txt'. The file 't10_txt_100000_128_1.5_1.txt' is selected. Below the file explorer, a WordPad window is open, displaying the content of the selected file. The content consists of two lines of z-normalized time series data, each starting with a location number followed by a series of values.

Name	Date modified	Type	Size
t10.txt	8/27/2008 10:22 PM	Text Document	1,758 KB
t10_txt_100000_128_1.5_1.txt	9/16/2008 12:43 AM	Text Document	3 KB
t10_txt_100000_128_1.5_2.txt	9/16/2008 12:46 AM	Text Document	4 KB

```
78865 -1.814236 -1.956496 -2.039228 -2.080241 -2.088035 -2.071412 -2.048211 -2.013429 -1.959688 -1.906668 -1.867136 -:  
83860 -1.814236 -1.956496 -2.039228 -2.080241 -2.088035 -2.071412 -2.048211 -2.013429 -1.959688 -1.906668 -1.867136 -:
```

Ready to Find Motifs?

- All you need to do is read next page and email Eamonn Keogh requesting the password
- Why do we make you request the password?

We want to track how many people are using our code.

We want to encourage others to share their datasets (as we have)

We want to encourage others to share their code (as we have)

Note: the current code is main memory only, sometime in 2009 we plan to release a disk aware version that can handle 50,000,000+ time series. If you have a pressing need for such scalability now, let us know.

Email to eamonn@cs.ucr.edu

1. I am requesting the password for the MK code
2. I promise that if I publish a paper that uses the MK code, I will make every effort to make the data I test on publicly available.
3. I promise that if I publish a paper that uses the MK code, I will make every effort to make the code I use publicly available.

If you disagree with the above, I will still give you the code, but you need to explain why in detail.