File Formats for Microarray Data
Version 2
Charlie Kim


## Introduction

Several file formats are in common use for microarray data. Development of the formats has been driven by software requirements. The formats described below are in common use in our lab, and are required for programs such as Cluster and Treeview.

## .txt (text)

This is a standard text file, usually tab-delimited. This means that there is a tab character in between each column. This makes for easy import into spreadsheet programs such as Excel. The columns are generally a unique ID (UNIQID), an annotation column (NAME), and data columns (EXPT1, EXPT2, …). The row format includes a header line (UNIQID, NAME, EXPTs) immediately followed by data lines.

| UNIQID | NAME | EXPT1 | EXPT2 | EXPT3 |
|---|---|---|---|---|
| UNIQID1 | Gene 1 | -0.039 | 0.558 | -0.752 |
| UNIQID2 | Gene 2 | -0.045 | 0.214 | -1.133 |
| UNIQID3 | Gene 3 | 0.053 | -0.151 | -1.621 |
| UNIQID4 | Gene 4 | 0.011 | 0.174 | -1.389 |
| UNIQID5 | Gene 5 | -0.06 | 0.498 | -2.569 |
| UNIQID6 | Gene 6 | -0.147 | 0.595 | -3.456 |
| UNIQID7 | Gene 7 | -0.115 | 0.117 | -1.228 |
| UNIQID8 | Gene 8 | -0.003 | -0.274 | -1.267 |

## .pcl (pre-clustering)

This format was developed for loading into Eisen's Cluster program. It is simply a tab-delimited text file with a set of specific fields. It incorporates the GWEIGHT column and EWEIGHT row for flexibility in pre-specifying which experiments or spots should have more weight (i.e. have more influence in driving the clustering). The format is the same as a .txt file, except that it includes a GWEIGHT column after the NAME column, and the EWEIGHT row after the header row. This format can be prepared and/or manipulated in Excel. See Eisen's Cluster software for additional information.

| UNIQID | NAME | GWEIGHT | EXPT1 | EXPT2 | EXPT3 |
|---|---|---|---|---|---|
| EWEIGHT | | | 1 | 1 | 1 |
| UNIQID1 | Gene 1 | 1 | -0.039 | 0.558 | -0.752 |
| UNIQID2 | Gene 2 | 1 | -0.045 | 0.214 | -1.133 |
| UNIQID3 | Gene 3 | 1 | 0.053 | -0.151 | -1.621 |
| UNIQID4 | Gene 4 | 1 | 0.011 | 0.174 | -1.389 |
| UNIQID5 | Gene 5 | 1 | -0.06 | 0.498 | -2.569 |
| UNIQID6 | Gene 6 | 1 | -0.147 | 0.595 | -3.456 |
| UNIQID7 | Gene 7 | 1 | -0.115 | 0.117 | -1.228 |
| UNIQID8 | Gene 8 | 1 | -0.003 | -0.274 | -1.267 |

**.cdt (cluster data tree – I'm not sure if this is correct, but that's how I remember what it is)**

This format is the output generated by Eisen's Cluster. The format is identical to the .pcl format (tab-delimited text), except that it includes an additional column (GID) and row (AID). The GID and AID fields are required by Treeview for drawing the dendrograms representing the relationships between arrays and/or genes. This format can be prepared and/or manipulated in Excel.

| GID | UNIQID | NAME | GWEIGHT | EXPT1 | EXPT2 | EXPT3 |
|---|---|---|---|---|---|---|
| AID | | | | ARRY0X | ARRY14X | ARRY13X |
| EWEIGHT | | | | 1 | 1 | 1 |
| GENE1682X | UNIQID1 | Gene 1 | 1 | -0.039 | 0.558 | -0.752 |
| GENE478X | UNIQID2 | Gene 2 | 1 | -0.045 | 0.214 | -1.133 |
| GENE1244X | UNIQID3 | Gene 3 | 1 | 0.053 | -0.151 | -1.621 |
| GENE3199X | UNIQID4 | Gene 4 | 1 | 0.011 | 0.174 | -1.389 |
| GENE4488X | UNIQID5 | Gene 5 | 1 | -0.06 | 0.498 | -2.569 |
| GENE1628X | UNIQID6 | Gene 6 | 1 | -0.147 | 0.595 | -3.456 |
| GENE2226X | UNIQID7 | Gene 7 | 1 | -0.115 | 0.117 | -1.228 |
| GENE163X | UNIQID8 | Gene 8 | 1 | -0.003 | -0.274 | -1.267 |