

Technical report (Original)

Connection up- and down-regulation expression analysis of microarrays (CU-DREAM): a physiogenomic discovery tool

Chatchawit Apornthewan^a, Apiwat Mutirangura^b

^aDepartment of Mathematics, Faculty of Science; ^bCenter for Excellence in Molecular Genetics of Cancer and Human Diseases, Department of Anatomy, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand.

Background: Many microarray experiments have been conducted during recent years, and scores of gene expression data have been archived in public databases. The use of data from multiple experiments can provide valuable information. However, there is a lack of convenient tools to compare datasets in this manner.

Objective: Implement software, called CU-DREAM, to compare the datasets of two microarray experiments. CU-DREAM is easy to use and compatible with Gene Expression Omnibus (GEO).

Subjects and methods: Five experiments were used to demonstrate the functionality of CU-DREAM. These are GSE6791, GSE7803, GSE5816, GSE4246, and GSE13638 for studies of cancers and RNA interference.

Results: All six showcases demonstrated the validity of the CU-DREAM approach. One showcase could confirm the regulation of genes identified in two independent experiments on cervical cancer. The statistical significance was lower compared with cervical and lung cancers. In addition, CU-DREAM could identify isoform changes in lung cancer. The last showcase demonstrated that Dicer- and Ago2-depleted cells or Dicer-depleted HeLa and HEK293 cells shared the same gene regulation pathways. CU-DREAM had seven main functions: 1) to identify genes that are up- and down-regulated in an experiment, 2) to validate significantly regulated genes using data from another experiment, 3) to determine if two different diseases have a similar effect on gene regulation, 4) to identify isoform-changed genes, 5) to determine if cells share gene regulation mechanisms, 6) to identify common gene regulation pathways even when comparing two different cell types, and 7) to identify down-stream genes that are regulated by the conditions of the analyzed experiments.

Conclusion: CU-DREAM is an effective tool for the pre-screening of drugs, substances or environmental insults or the identification of the genetic changes that are associated with pathological conditions (CU-DREAM can be downloaded from: <http://pioneer.netserv.chula.ac.th/~achatcha/cu-dream>).

Keywords: CU-DREAM, gene expression, gene expression omnibus, intersection, microarray

The advent of high-throughput microarrays has shifted a paradigm from hypothesis-driven research to data-driven discovery [1]. Instead of setting up hypotheses and testing them one by one, potential hypotheses are mined from genome-wide data collected all at once. Recently, Gene Expression

Omnibus (GEO) maintains data from over 10,000 experiments, and this number is increasing [2-5]. A modern microarray is made up of >50,000 probes. Consequently, a plethora of genome-wide data have been produced and archived for re-analysis. The data collected by an investigator may be re-analyzed by another investigator. However, as it is done from a different point of view, it may yield other important discovery.

Another paradigm shift has taken place with respect to the role of bioinformatics. Bioinformatics

Correspondence to: Chatchawit Apornthewan, Department of Mathematics, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand. Email: Chatchawit.A@chula.ac.th

has long been recognized as part of data analysis. Bioinformaticians are normally not the principal investigators who come up with the questions asked in primary medical discovery. However, Butte [6] has stimulated bioinformaticians to ask questions that no other scientists can ask or answer. He pointed out four reasons to enable bioinformaticians to be questioners. One of his reasons was “the intersection of data across multiple experiments”.

The objective of this study is to explain how to find answers using software, called CU-DREAM. The CU-DREAM stands for connection up- and down-regulation expression analysis of microarrays. It allows researchers to explore diseases, biological processes, and related mechanisms that alter genome-wide gene expression patterns and identify regulators and regulated genes. In the study, we devised CU-DREAM to compare the results of two independent microarray experiments. It was demonstrated that CU-DREAM could perform a comprehensive analysis of expression array databases to determine if there is a connection between the test variables of two expression microarray experiments. Common test variables including diseases, biological processes, environmental factors, and experimental conditions such as siRNA, immunoprecipitation, and genes, were studied.

Methods

CU-DREAM is simple and accessible to the public, including those who are not computer proficient. CU-DREAM is installed and operated as single executable file for use in a command prompt

on Windows operating systems. Users can feed inputs and obtain outputs via Microsoft Excel. Presuming that most users are familiar with Excel, everyone should be able to use CU-DREAM. All programming was performed in Microsoft Visual C# Express. The interprocess communication with Excel was realized through Microsoft Office system Primary Interop Assemblies (PIA). The user manual is available at the project homepage (<http://pioneer.netserv.chula.ac.th/~achatcha/cu-dream>).

Results

Five showcases were selected to demonstrate the functionality of CU-DREAM. All datasets used in the showcases are listed in **Table 1**. These can be downloaded from GEO. The parameter setting for each showcase, including experimental and control groups, is available at project homepage. The first three datasets are from studies of cancer, and the last two are from studies of RNA interference (RNAi).

Showcase 1: CU-DREAM identified genes that were up- and down-regulated.

The experimental group was cervical cancer cells, and the control group was normal cells. CU-DREAM performs a t-test for each probe. A partial result is shown in **Table 2**. A gene is up-regulated if “Mean1” (experimental group) is greater than “Mean2” (control group). Otherwise, the gene is down-regulated. The statistical significance is determined using a p-value threshold. This dataset included 54,675 probes. Note that **Table 2** has been formatted for use in this paper; the actual CU-DREAM outputs are Excel files.

Table 1. All datasets used in the showcases.

| Dataset | Title |
|----------|---|
| GSE6791 | Gene expression profiles of HPV-positive and HPV-negative head/neck and cervical cancers [7] |
| GSE7803 | Preinvasive and invasive cervical squamous cell carcinomas [8] |
| GSE5816 | A genome-wide screen for hypermethylated genes in lung cancer cells [9] |
| GSE4246 | Analysis of transcripts regulated by Dicer and Argonaute proteins in human HEK-293 cells [10] |
| GSE13638 | Ago2 or Dicer knockdown effects on mRNA levels in HeLa cells [11] |

Table 2. A partial result for GSE6791.

| Probe ID | Gene Symbol | Mean1 | Mean2 | Mean1 – Mean2 | P-value |
|-----------|-------------|-------|-------|---------------|----------|
| 1007_s_at | DDR1 | 12.04 | 12.05 | -0.01 | 8.94E-01 |
| 1053_at | RFC2 | 7.47 | 7.10 | 0.37 | 3.94E-02 |
| 117_at | HSPA6 | 7.64 | 7.31 | 0.33 | 2.08E-01 |
| 121_at | PAX8 | 9.89 | 10.57 | -0.68 | 4.98E-05 |

Showcase 2: CU-DREAM validated significantly regulated genes.

We compared two cervical cancer expression libraries, GSE6791 and GSE7803, to identify genes that were up- and down-regulated in both libraries. **Table 3** shows the up-up and down-down intersections between these two studies of cervical cancer. Each entry in the contingency table is the number of genes. The significant p-values and odds ratios >1 indicate the strong association between the two independent studies. More specifically, the genes that were up-regulated in one study were also up-regulated in another study. This result confirmed that the two studies involved a common gene regulation mechanism. Importantly, microarrays include both unique probes (having one corresponding gene) and homology probes (having multiple corresponding genes). In addition, a gene may have multiple corresponding probes. Therefore, counting the number

of genes is not straightforward. CU-DREAM uses a simple count algorithm. This algorithm is documented in the user manual. CU-DREAM also outputs gene symbols in sets **A**, **B**, **C**, and **D**.

Showcase 3: CU-DREAM determined if two different diseases had a similar effect on gene regulation.

We compared lung cancer (GSE5816) and cervical cancer (GSE6791). This showcase deliberately used different tissues, the lungs, and the cervix, as shown in **Tables 4**. The result is similar to that of the previous showcase, but the p-values and odds ratios are less significant for this showcase. This is expected because different types of cancer tissues are less likely to share the same regulated genes and underlying mechanisms than two samples of the same type of cancer.

Table 3. The intersections between two studies of cervical cancer (GSE6791 and GSE7803).

| | | GSE7803 | | | | GSE7803 | |
|--|-----------|-----------|----------|--|-------------|-------------|----------|
| | | Up (0.01) | Not up | | | Down (0.01) | Not down |
| GSE6791 | Up (0.01) | A =1,186 | B =1,967 | GSE6791 | Down (0.01) | 1,024 | 4,029 |
| | Not up | C =800 | D =9,105 | | Not down | 965 | 7,040 |
| P-value: 0.00E+00, Odds ratio: 6.86, Lower 95% CI: 6.20, Upper 95% CI: 7.60. | | | | P-value: 4.77E-37, Odds ratio: 1.85, Lower 95% CI: 1.68, Upper 95% CI: 2.04. | | | |

Partial lists of genes in sets **A**, **B**, **C**, and **D**.

| Set A | Set B | Set C | Set D |
|---------|--------|-------|--------|
| RFC2 | PTPN21 | HSPA6 | DDR1 |
| PTPN11 | MAPK1 | PAX8 | GUCA1A |
| C5orf22 | TMEFF1 | HOXD4 | UBA7 |
| SEC62 | EYA3 | BRF1 | THRA |

Table 4. The intersections between lung cancer (GSE5816) and cervical cancer (GSE6791).

| | | Cervical cancer | | | | Cervical cancer | |
|--|-----------|-----------------|-----------|--|-------------|-----------------|---------|
| | | Up (0.01) | Notup | | | Down (0.01) | Notdown |
| Lung cancer | Up (0.01) | A =95 | B =95 | Lung cancer | Down (0.01) | 451 | 756 |
| | Not up | C =3,882 | D =16,302 | | Not down | 7,760 | 11,407 |
| P-value: 1.75E-26, Odds ratio: 4.20, Lower 95% CI: 3.15, Upper 95% CI: 5.59. | | | | P-value: 3.20E-02, Odds ratio: 0.88, Lower 95% CI: 0.78, Upper 95% CI: 0.99. | | | |

Partial lists of genes in sets *A*, *B*, *C*, and *D*.

| Set A | Set B | Set C | Set D |
|-----------|--------|--------|--------|
| ZC3HAV1L | WDR17 | RFC2 | DDR1 |
| UHRF1BP1L | COBL | PTPN21 | HSPA6 |
| CARD8 | JMJD6 | MAPK1 | PAX8 |
| NEK4 | ZNF114 | VPS18 | GUCA1A |

Showcase 4: CU-DREAM identified isoform-changed genes.

Many genes have more than one isoform as the result of alternative splicing or the use of alternative promoters. CU-DREAM helps identify multi-isoform genes. Expression microarrays detect the majority of genes using multiple corresponding probes. If a gene shows as both up- and down-regulated, then the different isoforms of the gene may be differentially regulated. Here, we determined if there were isoform changes in lung cancer by comparing up- and down-regulated genes in lung cancer cells (GSE5816). We can intersect the up- and down-regulated genes identified in a single study, as shown in **Table 5**. A gene can be counted as both up- and down-regulated because there are multiple corresponding probes. Some probes may show up-regulation and some probes may show down-regulation. The intersection of up and down could reveal “isoforms” that exhibit differential expression between the experimental and control groups. We found that there might be isoform changes for at least 23 genes.

Table 5. The intersection between lung cancer (GSE5816) and lung cancer (GSE5816).

| | | Lung cancer | |
|--------------------|-----------|-------------|---------|
| | | Down (0.05) | Notdown |
| Lung cancer | Up (0.05) | 23 | 756 |
| | Not up | 3,040 | 16,555 |

P-value: 6.56E-22, Odds ratio: 0.17, Lower 95% CI: 0.11, Upper 95% CI: 0.25.

Showcase 5: CU-DREAM determined if cells shared gene regulation mechanisms.

We present the result of two different experiments that experimentally depleted different proteins. However, these proteins are known to form a complex and consequently down-regulate a common set of genes. We compared cells that had lower levels of Dicer and Ago2: HEK293 cells from GSE4246. Dicer and Ago2 form the RISC complex, which generates small RNA to limit mRNA levels at the post-transcriptional level (RNAi) or which activates transcriptional processes (RNAa) [12]. There are four Argonaute proteins, Ago1, Ago2, Ago3 and Ago4. A low Dicer concentration will limit the amount of RISC derived from Dicer and all Argonautes. However, a low Ago2 concentration will limit only the amount of RISC derived from Dicer and Ago2. Therefore, lowering the level of Dicer should reveal up- and down-regulated genes that are also up- and down-regulated by Ago2 depletion. **Table 6** shows the CU-DREAM results for Dicer and Ago2. Both up vs. up and down vs. down intersections had significant odds ratios >1. Obviously, the up-up genes should be regulated by RNAi, but the down-down genes suggest a possibility of another underlying mechanism, is complementary to RNAi. This process is known as RNAa.

Showcase 6: CU-DREAM can identify common gene regulation pathways even when comparing between two different cell types.

Table 6. The intersections between HEK293 Dicer (GSE4246) and HEK293 Ago2 (GSE4246).

| | | HEK293 Ago2 | | | | HEK293 Ago2 | |
|---------------|-----------|-------------|--------|---------------|-------------|-------------|----------|
| | | Up (0.05) | Not up | | | Down (0.05) | Not down |
| HEK293 | Up (0.05) | 177 | 669 | HEK293 | Down (0.05) | 149 | 680 |
| Dicer | Not up | 1,134 | 8,486 | Dicer | Not down | 1,041 | 8,596 |

P-value: 1.42E-14, Odds ratio: 1.98, Upper 95% CI: 1.66, Upper 95% CI: 2.36.

P-value: 4.34E-10, Odds ratio: 1.81, Upper 95% CI: 1.50, Upper 95% CI: 2.18.

We analyzed the mRNA expression of cells that were depleted of Dicer mRNA, but these cells were different cell types, cervical cancer (HeLa), cells and embryonic kidney (HEK293) from GSE13638 and GSE4246, respectively. **Table 7** shows that the intersection of different tissues is plausible, but the p-values are less significant than those for intersections of the same tissue.

Discussion

Expression microarrays yield useful information. The most common information to extract is the identity of the up- and down-regulated genes. We designed CU-DREAM to be a tool not only to identify regulated genes but also to compare different experiments. A comparison within the same experiment between up- and down-regulated genes can be used to identify genes with several isoforms that are differentially regulated under the experimental conditions. For example, at least 23 isoforms were differentially regulated when cells were transformed to become lung cancer.

The main objective of CU-DREAM is to compare data from two different experiments. There are several main benefits of this approach. First, CU-DREAM helps researchers to validate regulated genes under the same conditions, as demonstrated in the showcase comparing two expression arrays for cervical cancer cells. Second, CU-DREAM helps to differentiate biological conditions. For example, there are numerous

cancer cell types, but they may be very similar at the molecular level. If this is the case, the odds ratio should be >1, as demonstrated by the comparison of cervical and lung cancers; a higher odds ratio was obtained when different cervical cancers from different GSEs were compared. This result suggests that the gene regulation is similar among many different types of cancers and cells of the same cancer type share more regulated genes. Therefore, a high CU-DREAM odds ratio between regulated genes in the same direction may group two pathological lesions together. Recently, it has been shown that ovarian clear-cell and ovarian endometrioid carcinomas frequently contain a somatic mutation in the same gene [13]. We analyzed these two carcinomas with CU-DREAM and found odds ratios of 22.7 and 36.8 for the up-up and down-down intersections, respectively (data not shown). Note that the odds ratio has a symmetry property. The group (row) and event (column) can be used interchangeably. Significant odds ratios above one indicate the likelihood that the two experiments share the same gene regulation pathways. On the contrary, odds ratios below one indicate that the event of interest (“up” or “down”) is more likely to happen in the control group for the prior t-test or that the up-stream regulators acted significantly in the opposite direction.

CU-DREAM leads to a better understanding of human biology and of the pathogenesis of diseases by identifying connections between two different

Table 7. The intersections between HeLa Dicer (GSE13638) and HEK293 Dicer (GSE4246).

| | | HEK293 Dicer | | | | HEK293 Dicer | |
|--------------|-----------|--------------|--------|--------------|-------------|--------------|---------|
| | | Up (0.05) | Not up | | | Down (0.05) | Notdown |
| HeLa | Up (0.05) | 52 | 375 | HeLa | Down (0.05) | 58 | 463 |
| Dicer | Not up | 764 | 8,952 | Dicer | Not down | 716 | 8,906 |

P-value: 1.34E-03, Odds ratio: 1.62, Lower 95% CI: 1.20, Upper 95% CI: 2.19.

P-value: 2.00E-03, Odds ratio: 1.56, Lower 95% CI: 1.17, Upper 95% CI: 2.07.

experiments. Here, to simplify the analysis, we presented CU-DREAM results for proteins that are known to be in the same complex. Nevertheless, it is important to emphasize that CU-DREAM can reveal connections between diseases and between in vivo and in vitro experiments. Therefore, CU-DREAM facilitates the identification of drugs, substances or genetic changes that promote, or inhibit pathological conditions.

Acknowledgments

This work was supported by the Center of Excellence in Molecular Genetics of Cancer and Human Diseases, Department of Anatomy, Faculty of Medicine, Chulalongkorn University, TRF-MRG young scientific researcher grant and TRF senior research scholar grant. The authors have no conflict of interest to report.

Availability and requirements of CU-DREAM

- Project name: CU-DREAM
- Project homepage: <http://pioneer.netserv.chula.ac.th/~achatcha/cu-dream>
- Operating system: Microsoft Windows
- Programming Languages: C#
- Other requirements: .NET Framework 3.5 or higher, Microsoft Excel 2007, Microsoft Office system Primary Interop Assemblies (PIA)
- License: GNU General Public License
- Any restriction to use by non-academics: none.

References

1. Smalheiser NR. Informatics and hypothesis-driven research. *EMBO Rep.* 2002; 3:702.
2. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.* 2009; 37(Database issue):D885-90.
3. Edgar R, Barrett T. NCBI GEO standards and services for microarray data. *Nat Biotechnol.* 2006; 24:1471-2.
4. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.* 2007; 35(Database issue):D760-5.
5. Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, et al. NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.* 2005; 33(Database issue):D562-6.
6. Butte AJ. Translational bioinformatics applications in genome medicine. *Genome Med.* 2009; 1:64.
7. Pyeon D, Newton MA, Lambert PF, den Boon JA, Sengupta S, Marsit CJ, et al. Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. *Cancer Res.* 2007; 67:4605-19.
8. Zhai Y, Kuick R, Nan B, Ota I, Weiss SJ, Trimble CL, et al. Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion. *Cancer Res.* 2007; 67:10163-72.
9. Shames DS, Girard L, Gao B, Sato M, Lewis CM, Shivapurkar N, et al. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS Med.* 2006; 3:e486.
10. Schmitter D, Filkowski J, Sewer A, Pillai RS, Oakeley EJ, Zavolan M, et al. Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res.* 2006; 34:4801-15.
11. Han J, Pedersen JS, Kwon SC, Belair CD, Kim YK, Yeom KH, et al. Posttranscriptional crossregulation between Droscha and DGCR8. *Cell.* 2009; 136:75-84.
12. Pushparaj PN, Aarthi JJ, Kumar SD, Manikandan J. RNAi and RNAa—the yin and yang of RNAome. *Bioinformatics.* 2008; 2:235-7.
13. Wiegand KC, Shah SP, Al-Agha OM, Zhao Y, Tse K, Zeng T, et al. ARID1A Mutations in endometriosis-associated ovarian carcinomas. *N Engl J Med.* 10.1056/NEJMoa1008433.