

# Evaluation of Signaling Cascades Based on the Weights from Microarray and ChIP-seq Data

by

Zerrin Işık

Volkan Atalay

Rengül Çetin-Atalay



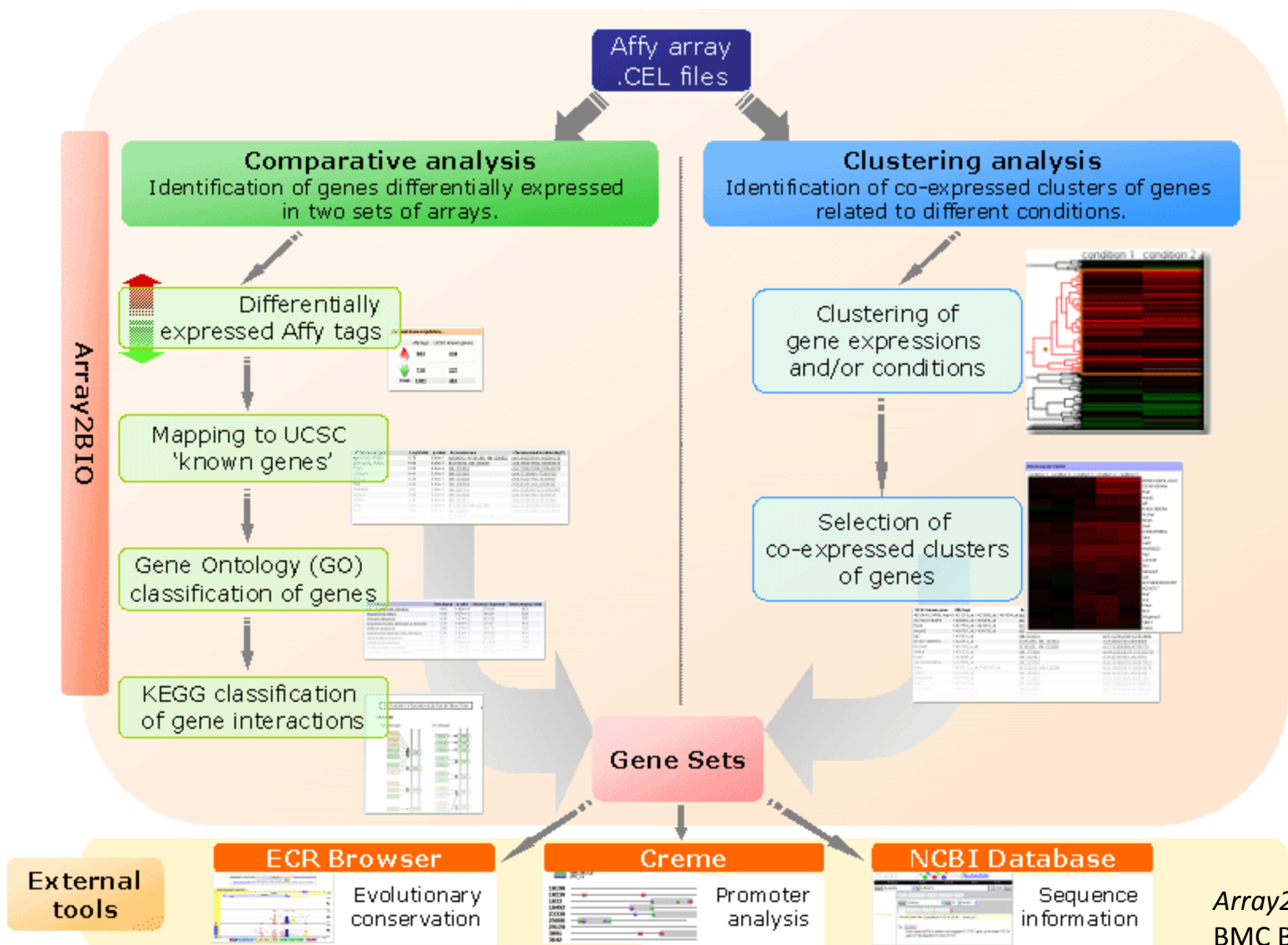
*Middle East Technical University and Bilkent University*  
*Ankara - TURKEY*



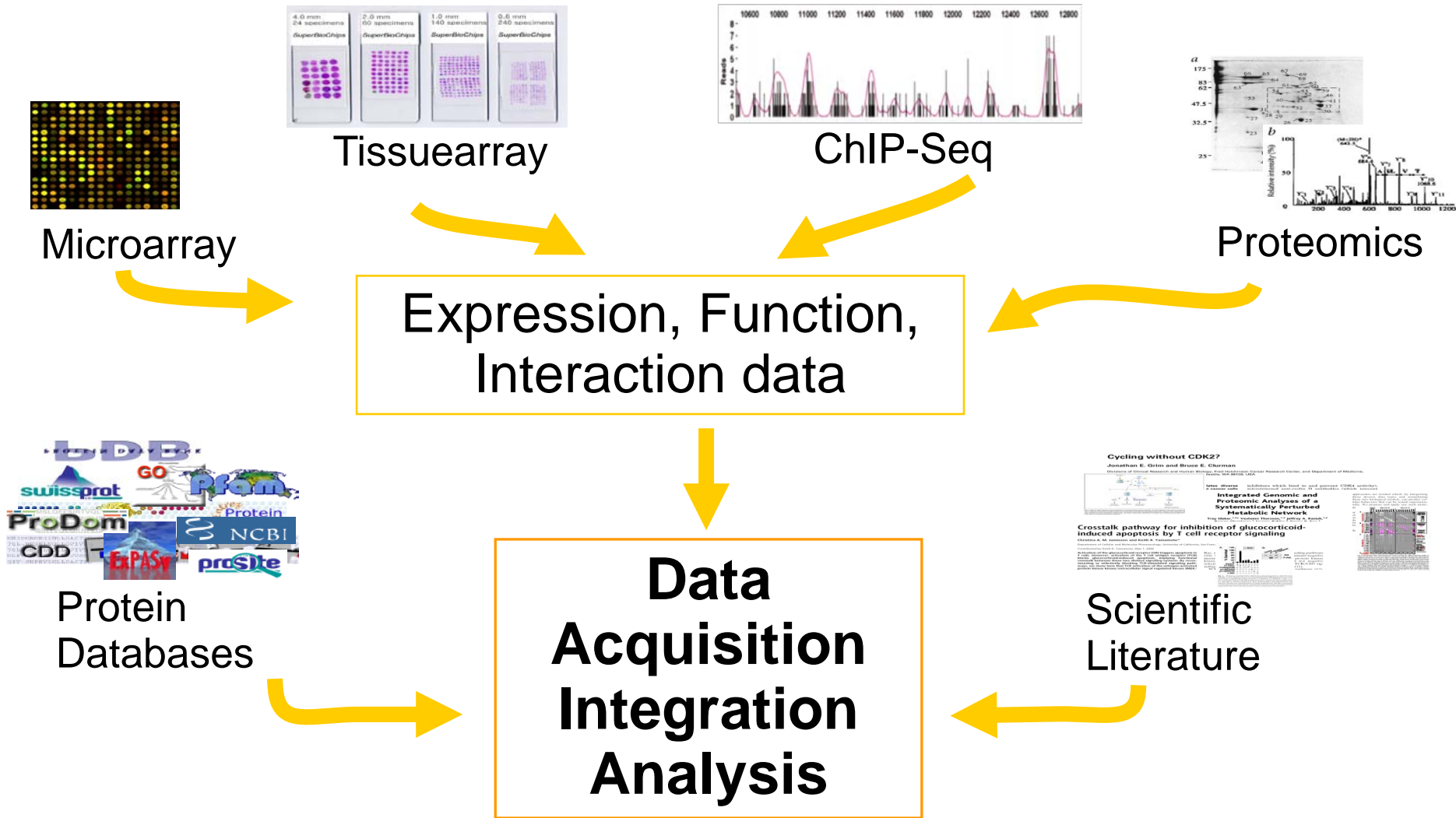
# *Content*

- ✓ Analysis of Microarray Data
- ✓ ChIP-Seq Data
- ✓ Data Processing & Integration
- ✓ Scoring of Signaling Cascades
- ✓ Results

# Traditional Analysis of Microarray Data

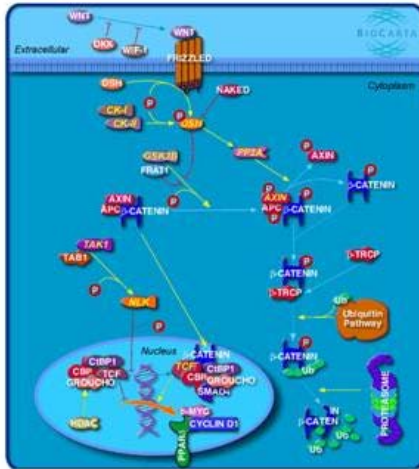


# Traditional Analysis of Microarray Data

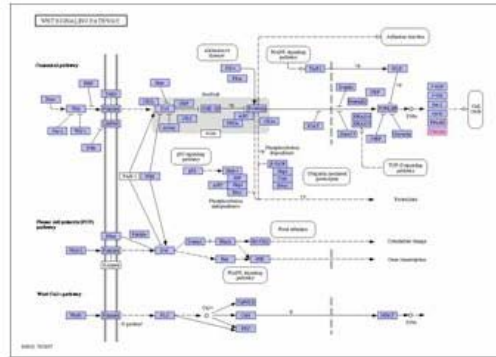


# Traditional Analysis of Microarray Data

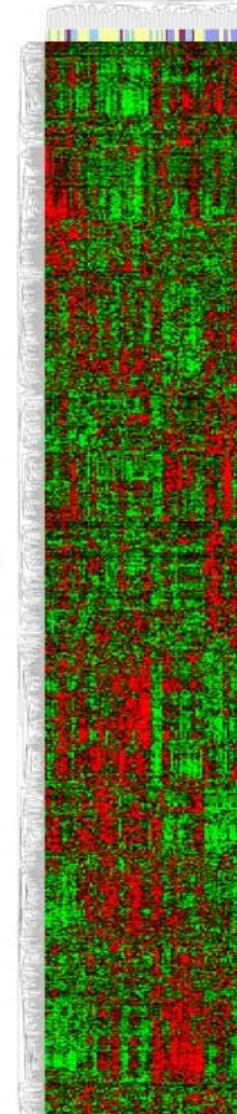
**BioCarta**



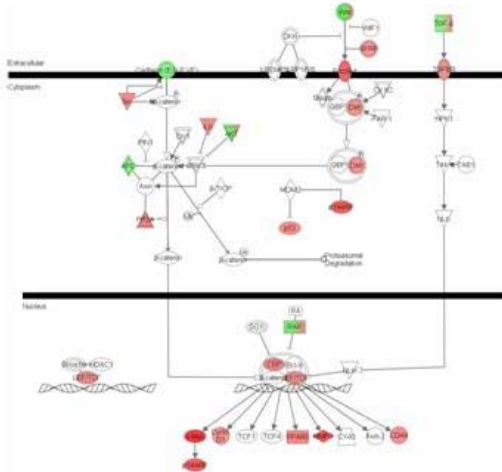
**KEGG**



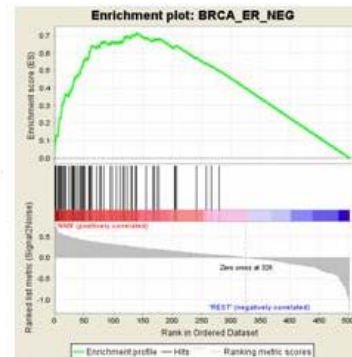
**Microarray Data**



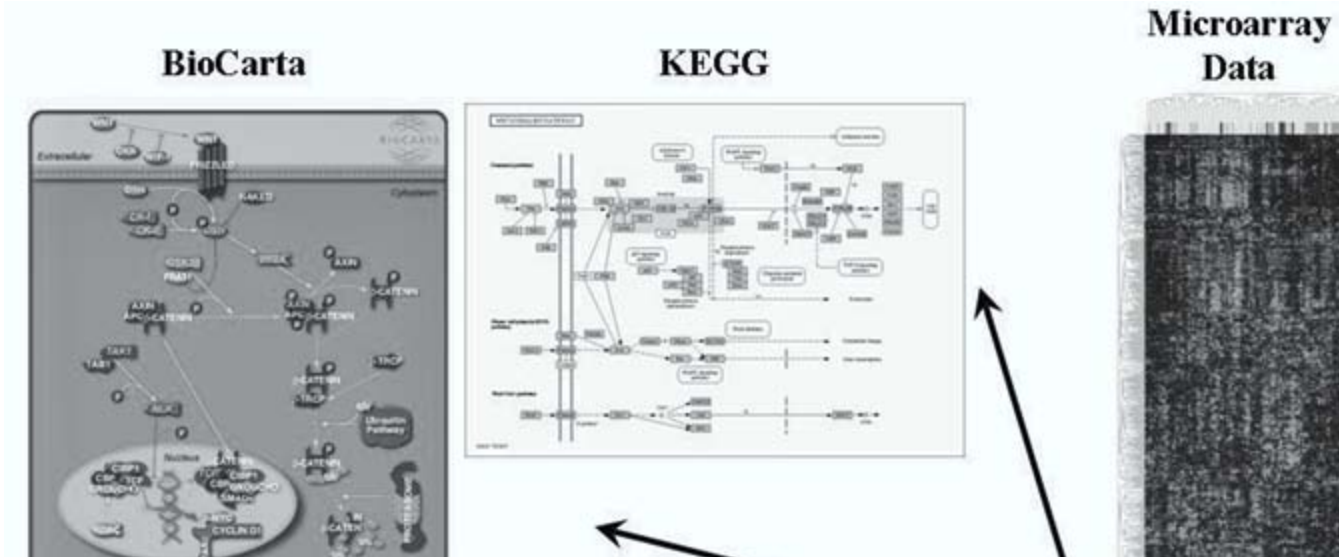
**Ingenuity Pathway Analysis (IPA)**



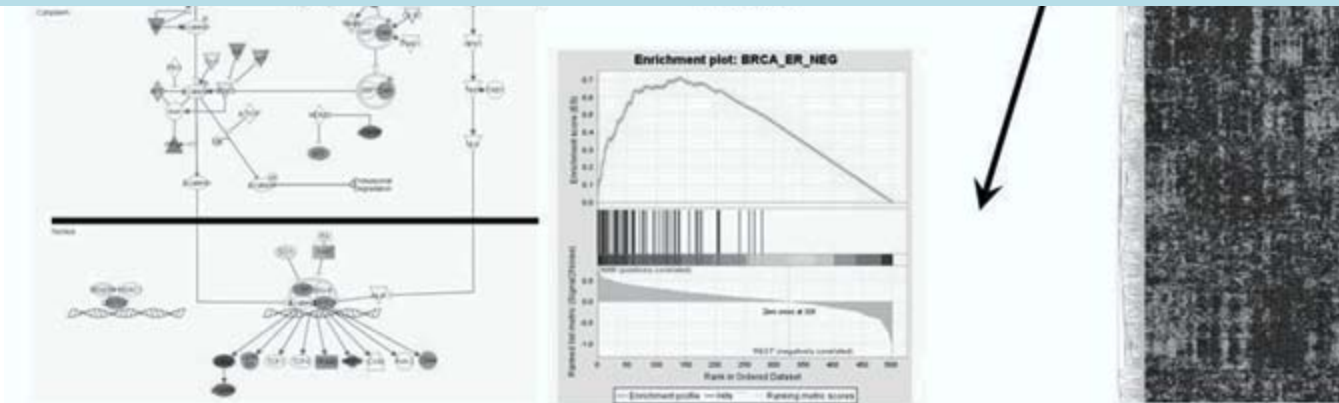
**GSEA**



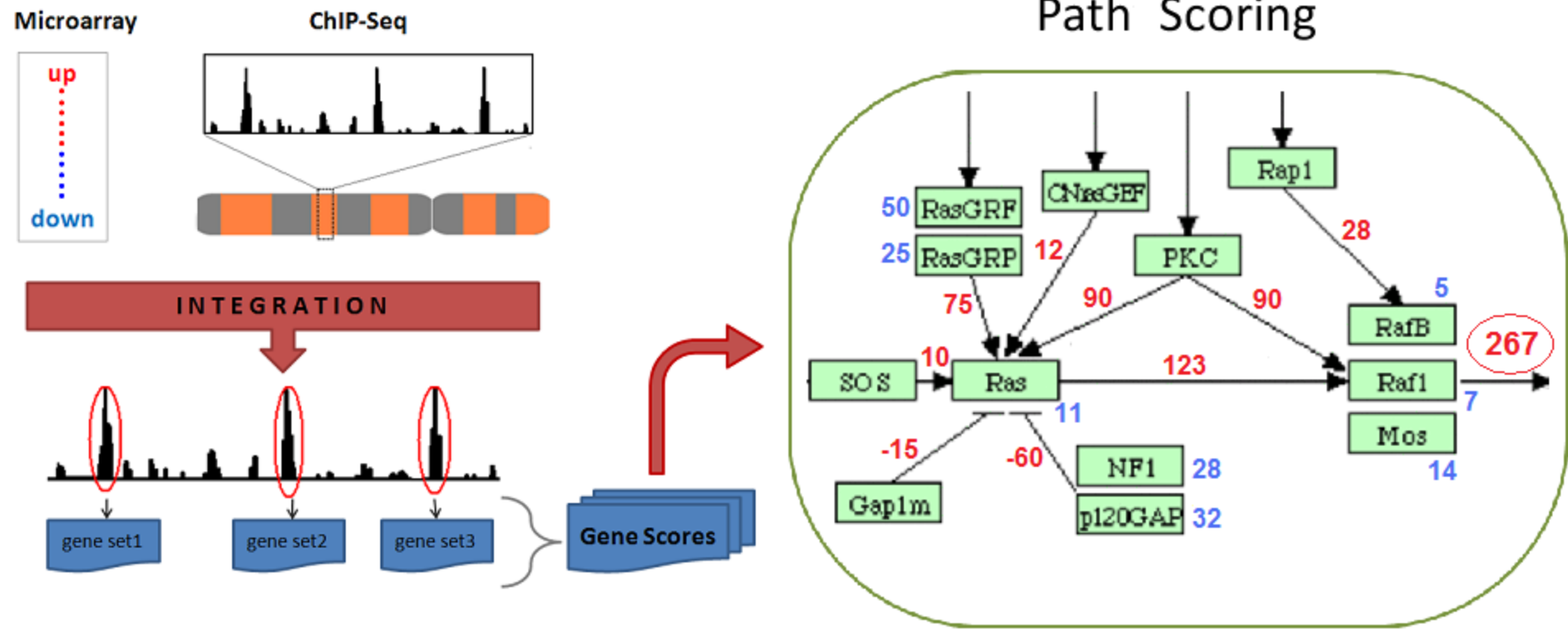
# Traditional Analysis of Microarray Data



**These tools depend on the primary significant gene lists!**



# Our Framework

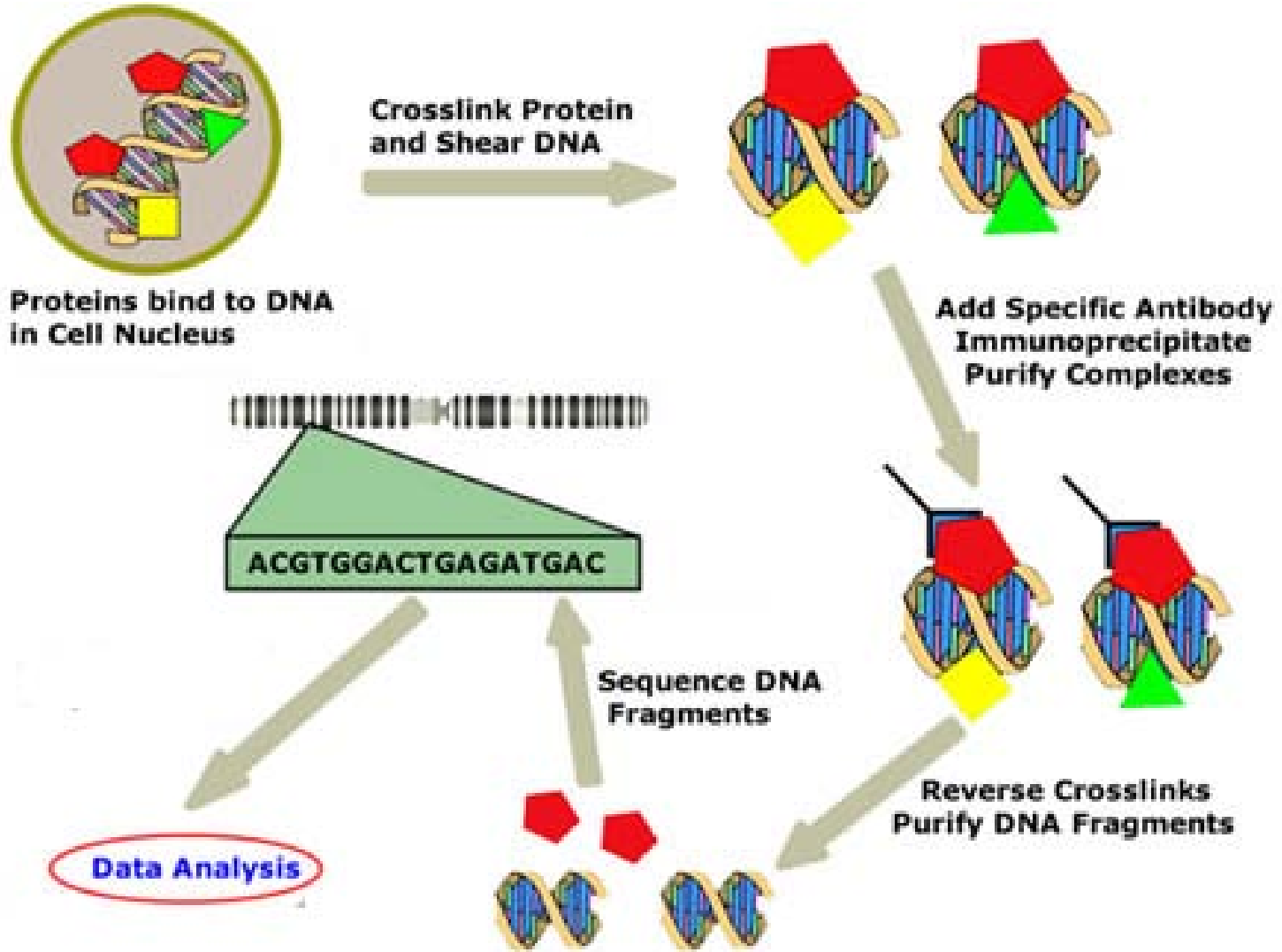


# *Content*

- ✓ Analysis of Microarray Data
- ✓ **ChIP-Seq Data**
- ✓ Data Processing & Integration
- ✓ Scoring of Signaling Cascades
- ✓ Results



# Chromatin Immunoprecipitation



# ChIP-Sequencing

- Chromatin Immunoprecipitation (ChIP) combined with genome re-sequencing (ChIP-seq) technology provides protein DNA interactome data.
- Generally, ChIP-seq experiments are designed for target transcription factors to provide their genome-wide binding information.

# Analysis of ChIP-seq Data

- Several analysis tools available:
  - QuEST: peak region detection
  - SISSRs : peak region detection
  - CisGenome: system to analyse ChIP data
    - visualization
    - data normalization
    - peak detection
    - FDR computation
    - gene-peak association
    - sequence and motif analysis

# Analysis Steps of ChIP-seq Data

- Align reads to the reference genome.

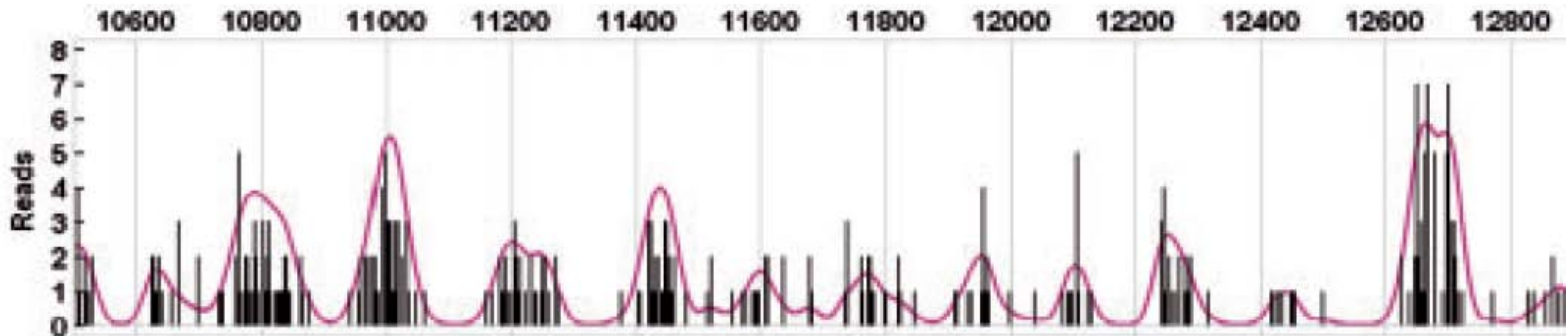
```
<< < > >>
31060 31070 31080 31090 3110
+84904 s_7_0066_282 CAGAGGTGGTGC AATTCCTTCTTGCCCAT
-84905 s_7_0007_217 AGAGGTGGTGC AATTCCTTCTTGCCAATG
-84906 s_7_0019_2486 AGAGGTGGTGC AATTCCTTCTTGCCAATG
+84907 s_7_0056_533 AGAGGTGGTGC AATTCCTTCTTGCCATG
+84908 s_7_0056_2056 AGAGGTGGTGC AATTCCTTCTTGCCATG
-84909 s_7_0017_4043 GAGGGGGTGC AATTCCTTCTTGCCAATGA
-84910 s_7_0023_1554 GCGGCGGTGC AACCTTCTTGCCAATGA
+84911 s_7_0023_3423 GAGGTGGTGC AATTCCTTCTTGCCAATGA
-84912 s_7_0030_776 GAGGTGGTGC AATTCCTTCTTGCCAATGA
-84913 s_7_0046_4984 GAGGAGGTGC ATTTCCTTCTTGCCAATGA
-84914 s_7_0026_4287 CGGCGGTGC AATTCCTTCTTGCCAATGAA
-84915 s_7_0057_3331 AGGTGGTGC AATTCCTTCTTGCCAATGAA
+84916 s_7_0007_2838 GGTGGTGC AATTCCTTCTTGCCAATGAAA
-84928 s_7_0012_5038 GGTGCAATTC CTTCTTGCCAATGAAATCA
-84929 s_7_0013_1582 GGTGCAATTC CTTCTTGCCAATGAAATCA
+84930 s_7_0018_1291 GGTGCAATTC CTTCTTGCCAATGAAATCA
-84931 s_7_0033_2217 CGTGCAATTC CTTCTTGCCAATGAAATCA
:> CONSENSUS -%%- ACAGAGGTGGTGC AATTCCTTCTTGCCAATGAAATCATGTATGCGTGTG
```

1:17:900:850 AGAACTTGGTGGTCATGGTGGGAAGGGAG

U1 0 1 0 chr2.fa 9391175 F .. 19A

# Analysis Steps of ChIP-seq Data

- Identification of peak (binding) regions.
  - **Peak:** Region has high sequencing read density



- FDR computation of peak regions.
- Sequence and motif analysis.

# Further Analysis of ChIP-Seq Data

- Although there are a few number of early stage analysis tools for ChIP-seq data, gene annotation methods should also be integrated like in the case of microarray data analysis.
- ChIP-seq experiments provide detailed knowledge about target genes to predict pathway activities.

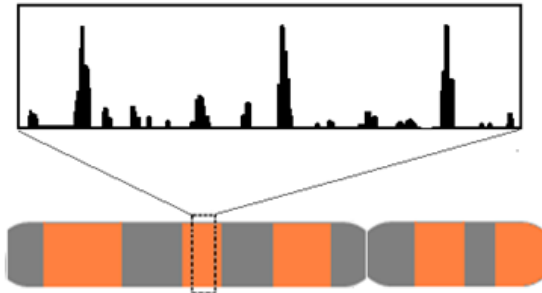
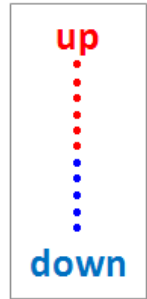
# *Content*

- ✓ Analysis of Microarray Data
- ✓ ChIP-Seq Data
- ✓ **Data Processing & Integration**
- ✓ Scoring of Signaling Cascades
- ✓ Results

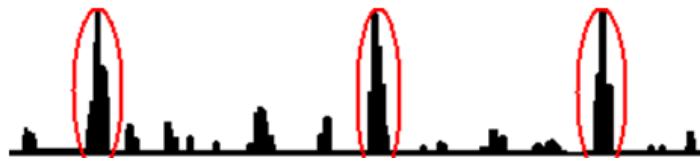
# Our Framework

Microarray

ChIP-Seq



INTEGRATION



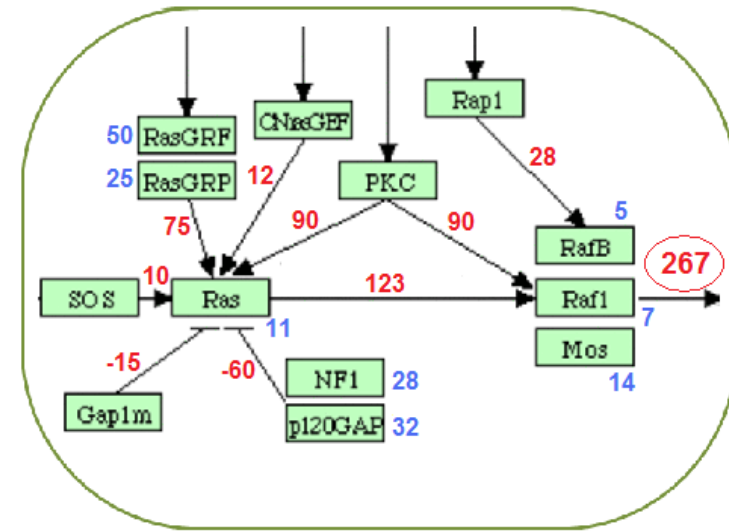
gene set1

gene set2

gene set3

Gene Scores

Pathway Scoring

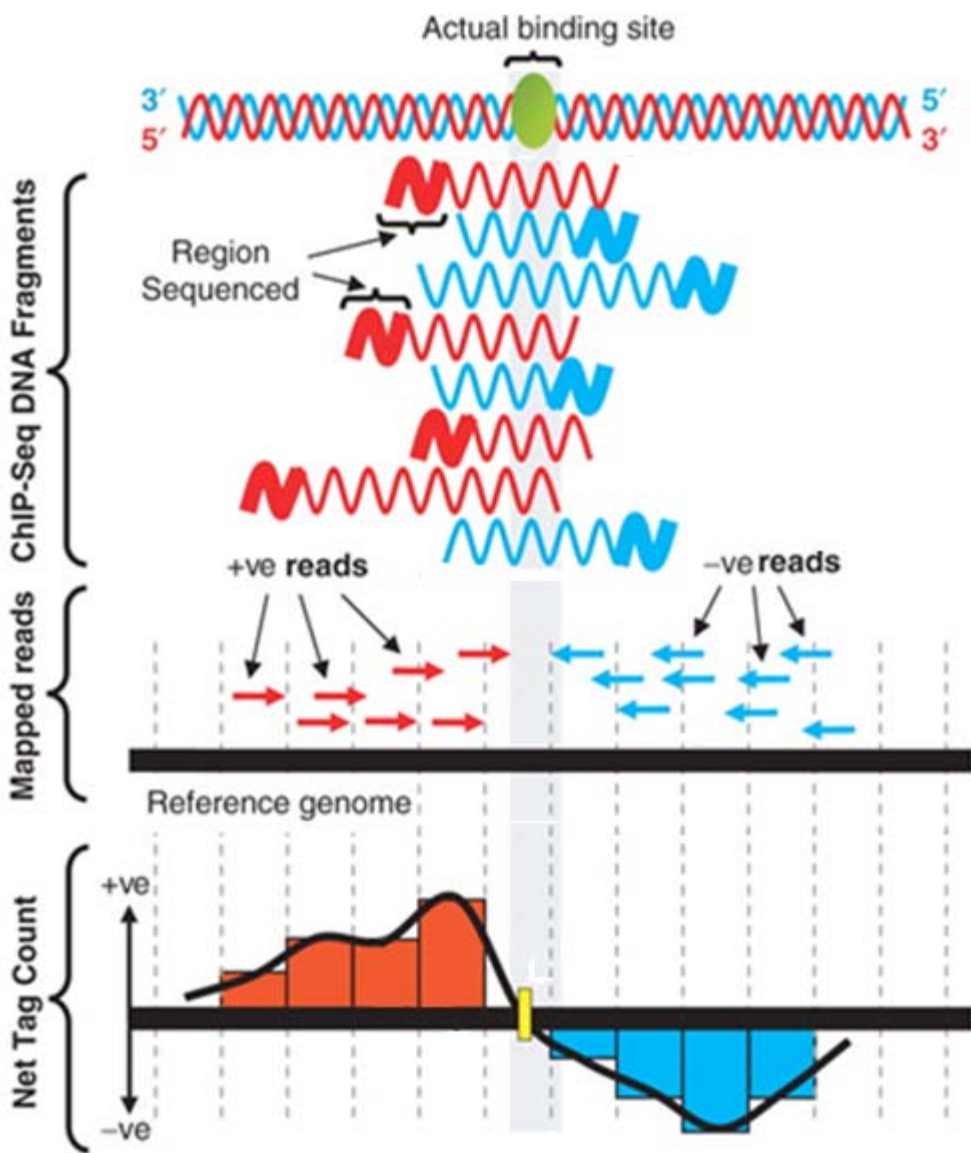




# Data Set

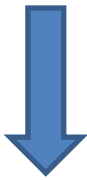
- **ChIP-Seq Data:** *OCT1* (TF)
  - Kang et.al. Genes Dev. 2009 (GSE14283)
  - Performed on human HeLa S3 cells.
  - Identify the genes targeted by OCT1 TF under conditions of oxidative stress.
- **Microarray Data:**
  - Murray et.al. Mol Biol Cel. 2004 (GSE4301)
  - 12800 human genes.
  - oxidative stress applied two channel data.

# Analysis of Raw ChIP-Seq Data



CisGenome software identified peak regions of OCT1 data.

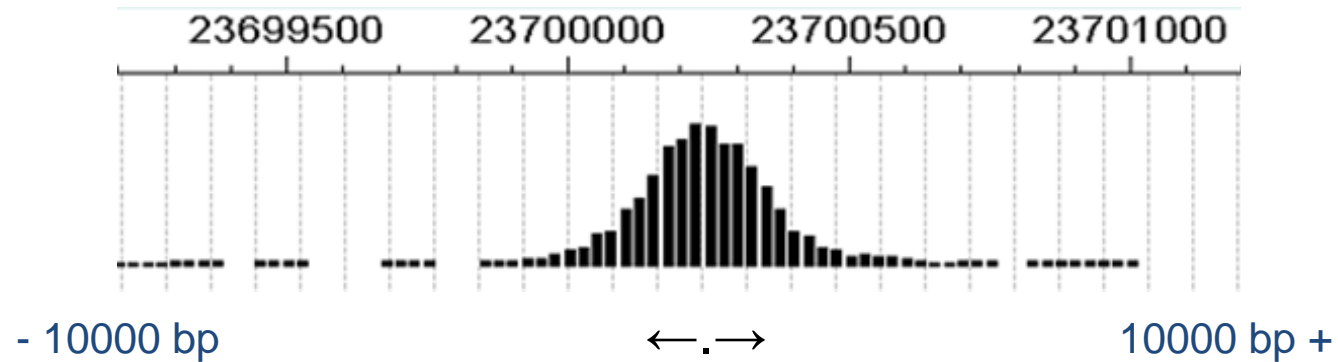
3.8 million reads



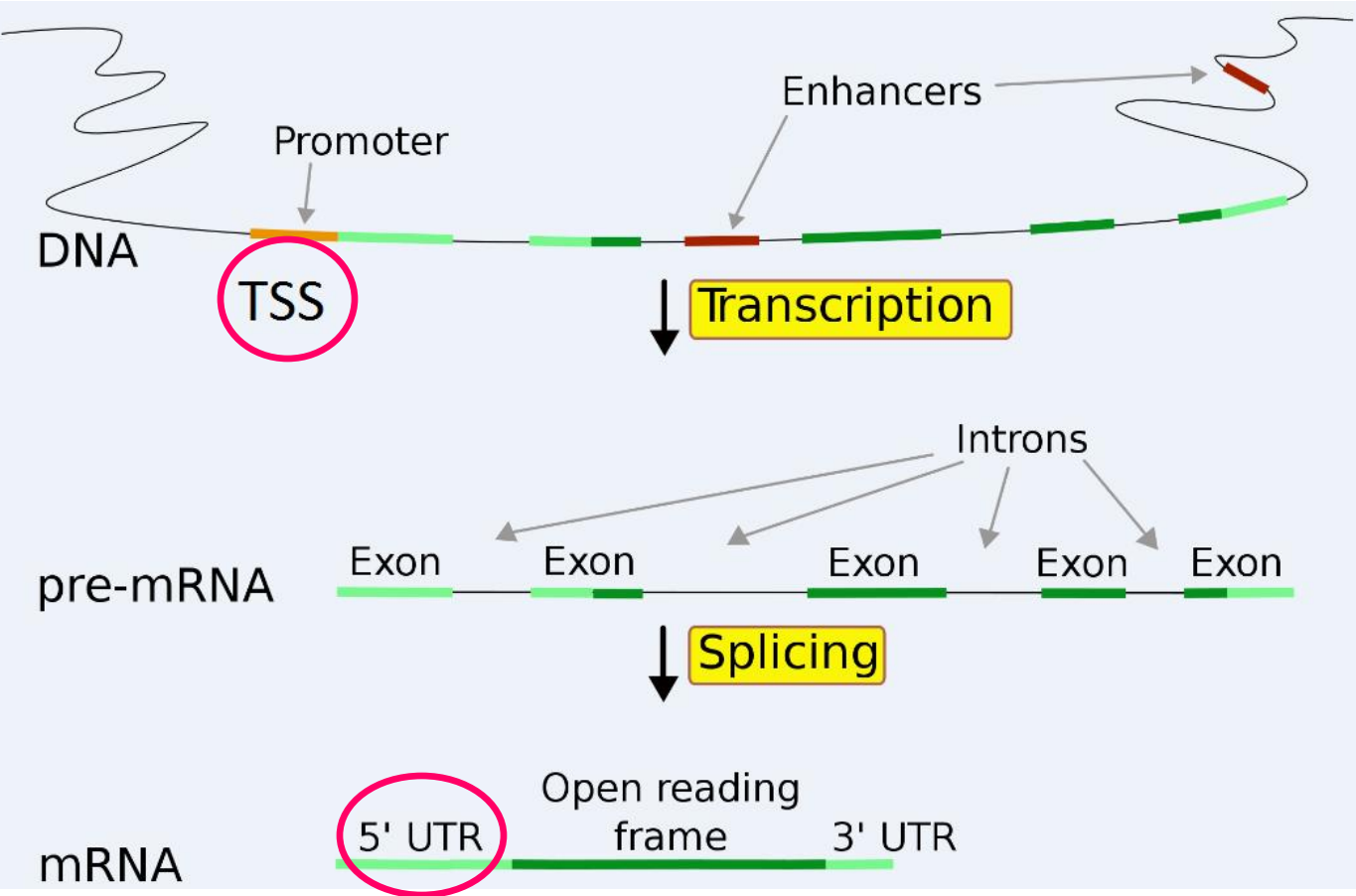
5080 peak regions

# Analysis of Raw ChIP-Seq Data

Identify neighboring genes of peak regions.



# Analysis of Raw ChIP-Seq Data



Total # of genes  
2843



# selected genes  
260



# ChIP-Seq Data Ranking

Percentile rank of each peak region is computed:

$$\textit{ReadRank}(r) = \frac{cf_l + 0.5(f_r)}{T}$$

*cf<sub>l</sub>* : cumulative frequency for all scores lower than score of the peak region *r*

*f<sub>r</sub>* : frequency of score of peak region *r*

*T* : the total number of peak regions

# Microarray Data Analysis

- Two channel data
- Use limma package of R-Bioconductor
  - Apply background correction
  - Normalize data between arrays
  - Compute fold-change of gene  $x$  :

$$FoldChange(x) = \log_2\left(\frac{\overline{ch2_x}}{\overline{ch1_x}}\right)$$

# Microarray Data Ranking

Set a percentile rank value for each gene :

$$\textit{ExpRank}(x) = \frac{cf_l + 0.5(f_x)}{T}$$

$cf_l$  : cumulative frequency for all fold-change values lower than the fold - change of the gene  $x$

$f_x$  : frequency of the fold-change of the gene  $x$

$T$  : the total number of genes in chip

# Integration of ChIP-Seq and Microarray Data

Scores were associated by taking their weighted linear combinations.

$$Score(x) = c_{chip}ReadRank(x) + c_{exp}ExpRank(x)$$



# Integration of ChIP-Seq and Microarray Data

Scores were associated by taking their weighted linear combinations.

$$Score(x) = c_{chip}ReadRank(x) + c_{exp}ExpRank(x)$$

| Gene name | Score(x) | ReadRank | ExpRank |
|-----------|----------|----------|---------|
| SPRY3     | 0.2565   | 0.000    | 0.513   |
| CNTFR     | 0.2215   | 0.233    | 0.210   |
| OSMR      | 0.5100   | 0.802    | 0.218   |
| PRLR      | 0.8460   | 0.712    | 0.980   |
| PIK3CA    | 0.3525   | 0.100    | 0.605   |

# *Content*

- ✓ Analysis of Microarray Data
- ✓ ChIP-Seq Data
- ✓ Data Processing & Integration
- ✓ **Scoring of Signaling Cascades**
- ✓ Results

# Scoring of Signaling Cascades

- KEGG pathways were used as the model to identify signaling cascades under the control of specific biological processes.
- Each signaling cascade was converted into a graph structure by extracting KGML files.

# KGML example

```
<entry id="11" name="hsa:1154" type="gene" link=http://www.genome.jp/dbget-bin/www_bget?
  hsa+1154> <graphics name="CISH" fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle"
  x="802" y="283" width="46" height="17"/> </entry>
<entry id="16" name="hsa:6772" type="gene" link=http://www.genome.jp/dbget-bin/www_bget?
  hsa+6772> <graphics name="STAT1..." fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle"
  x="343" y="246" width="46" height="17"/> </entry>
<entry id="21" name="hsa:3716" type="gene" link=http://www.genome.jp/dbget-bin/www_bget?
  hsa+3716> <graphics name="JAK1..." fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle"
  x="208" y="246" width="46" height="17"/> </entry>
<relation entry1="21" entry2="16" type="PPrel"><subtype name="phosphorylation" value="+p"/>
</relation>
<relation entry1="11" entry2="16" type="PPrel"><subtype name="inhibition" value="--|"/>
</relation>
```

# KGML example

```
<entry id="11" name="hsa:1154" type="gene" link=http://www.genome.jp/dbget-bin/www_bget?
  hsa+1154> <graphics name="CISH" fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle"
  x="802" y="283" width="46" height="17"/> </entry>
<entry id="16" name="hsa:6772" type="gene" link=http://www.genome.jp/dbget-bin/www_bget?
  hsa+6772> <graphics name="STAT1..." fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle"
  x="343" y="246" width="46" height="17"/> </entry>
<entry id="21" name="hsa:3716" type="gene" link=http://www.genome.jp/dbget-bin/www_bget?
  hsa+3716> <graphics name="JAK1..." fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle"
  x="208" y="246" width="46" height="17"/> </entry>
<relation entry1="21" entry2="16" type="PPrel"><subtype name="phosphorylation" value="+p"/>
</relation>
<relation entry1="11" entry2="16" type="PPrel"><subtype name="inhibition" value="--|"/>
</relation>
```



---

**Algorithm 1** : Computing Score of Signaling Cascades

---

**Input:** Graph  $\underline{P}$ , has *nodes* and *edges* arrays

*Score*: indicates self score of each node given by our method

*outputScore*: contains output edge score of each node

**Initialization:**

Apply Breadth-First Search algorithm

Extract initialization (ancestor of  $\underline{P}$ ) nodes: *initialNodes* = start node(s) of  $\underline{P}$

*otherNodes* = *nodes* \ *initialNodes*

**Score Computation:**

for  $i = 1$  to *length(initialNodes)* do

*outputScore[initialNodes[i]]* = *Score[initialNodes[i]]*

end for

for  $j = 1$  to *length(otherNodes)* do

*ancestorNodes* = ancestor node(s) of *otherNodes[j]*

*outputScore[j]* = *Score[j]*

    for  $k = 1$  to *length(ancestorNodes)* do

$e = E(k, j)$  {the edge between *ancestorNodes[k]* and *otherNodes[j]*}

        if type of  $e$  is activation then

*sign[k]* = 1 {assign weight of activation edge}

        else

*sign[k]* = -1 {assign weight of inhibition edge}

        end if

*outputScore[j]*+ = *outputScore[k]* \* *sign[k]* {sum up weight of incoming edge}

    end for

    if *outputScore[j]* < 0 then

*outputScore[j]* = 0 {negative score is originated by only inhibition edges}

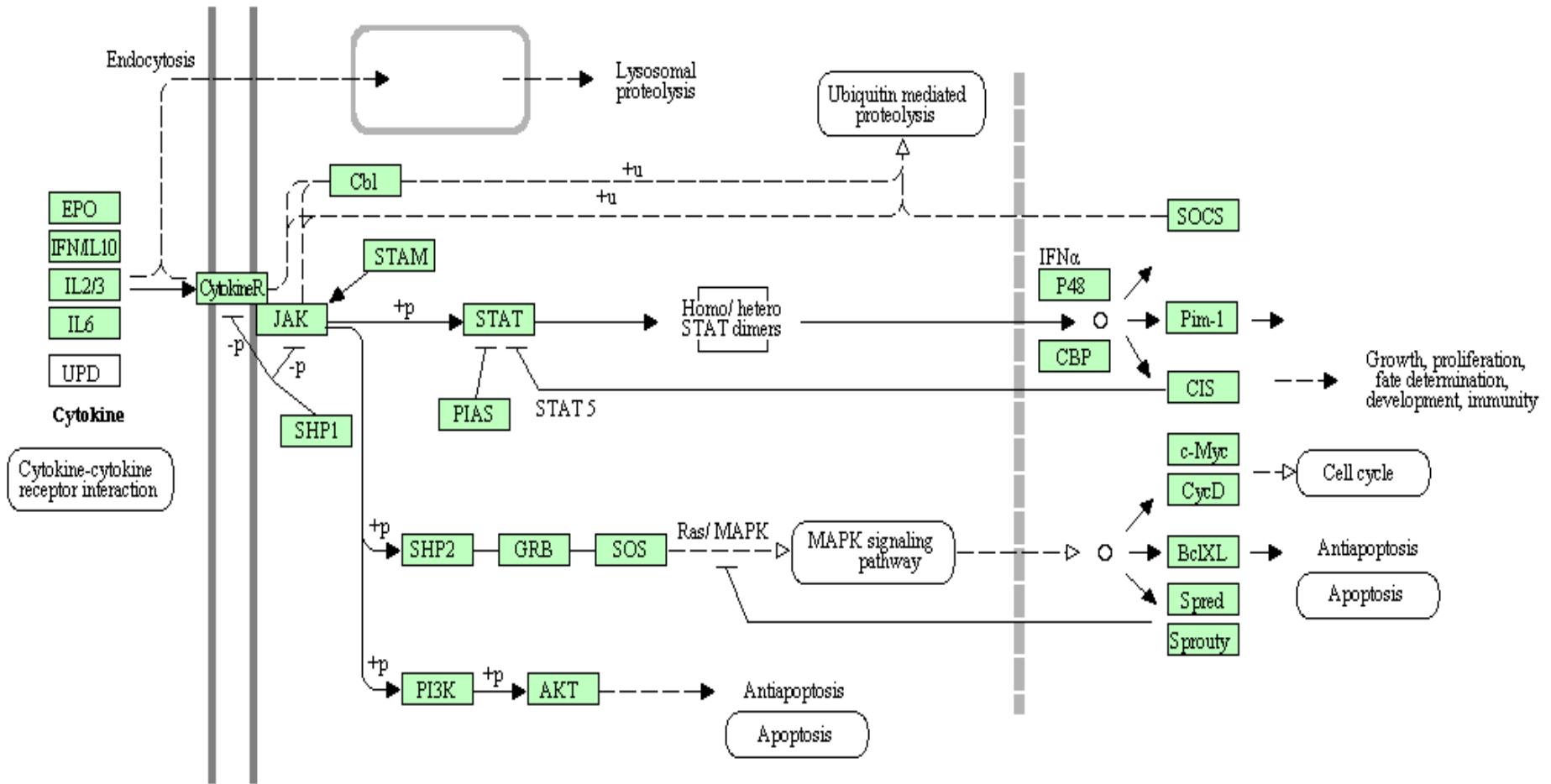
    end if

end for

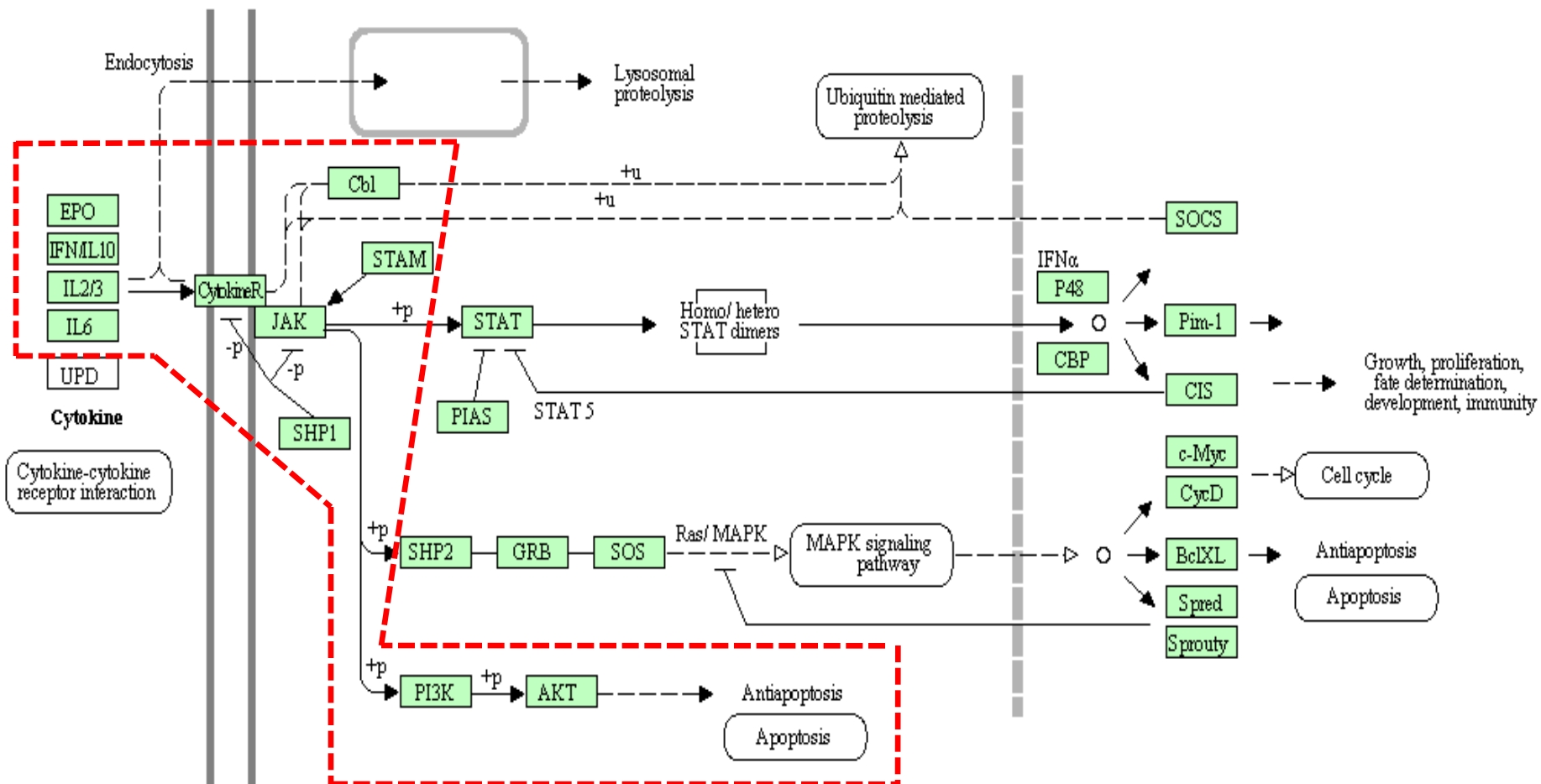
**Output:** *outputScore* of outcome biological processes in graph  $\underline{P}$ .

---

# JAK-STAT SIGNALING PATHWAY

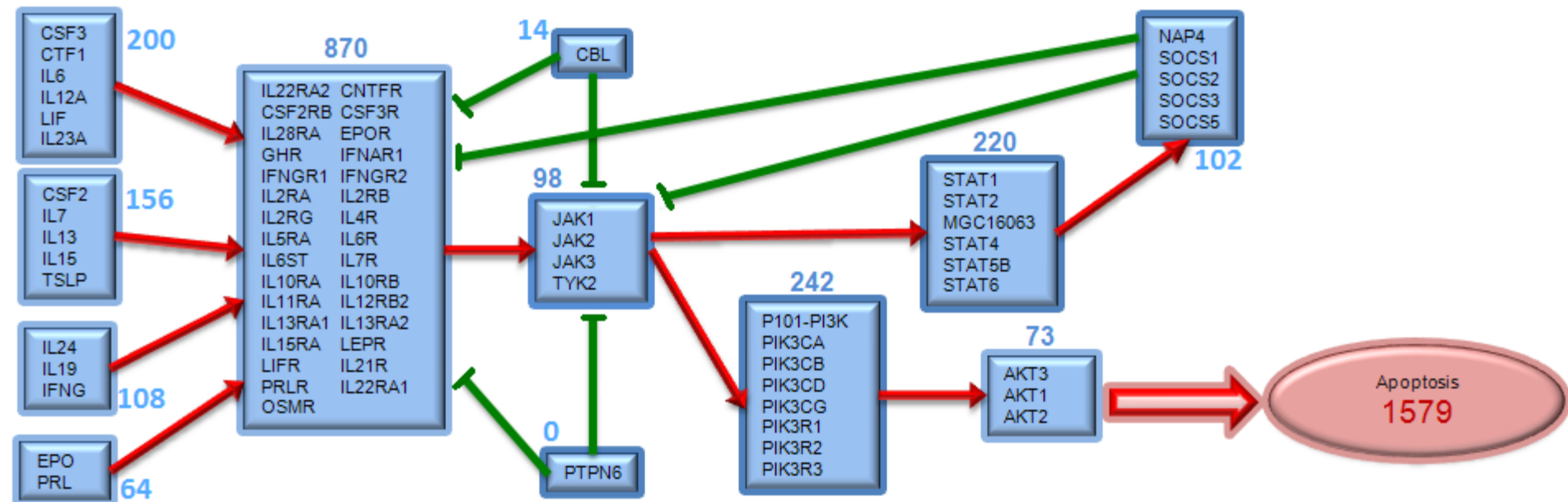


# JAK-STAT SIGNALING PATHWAY

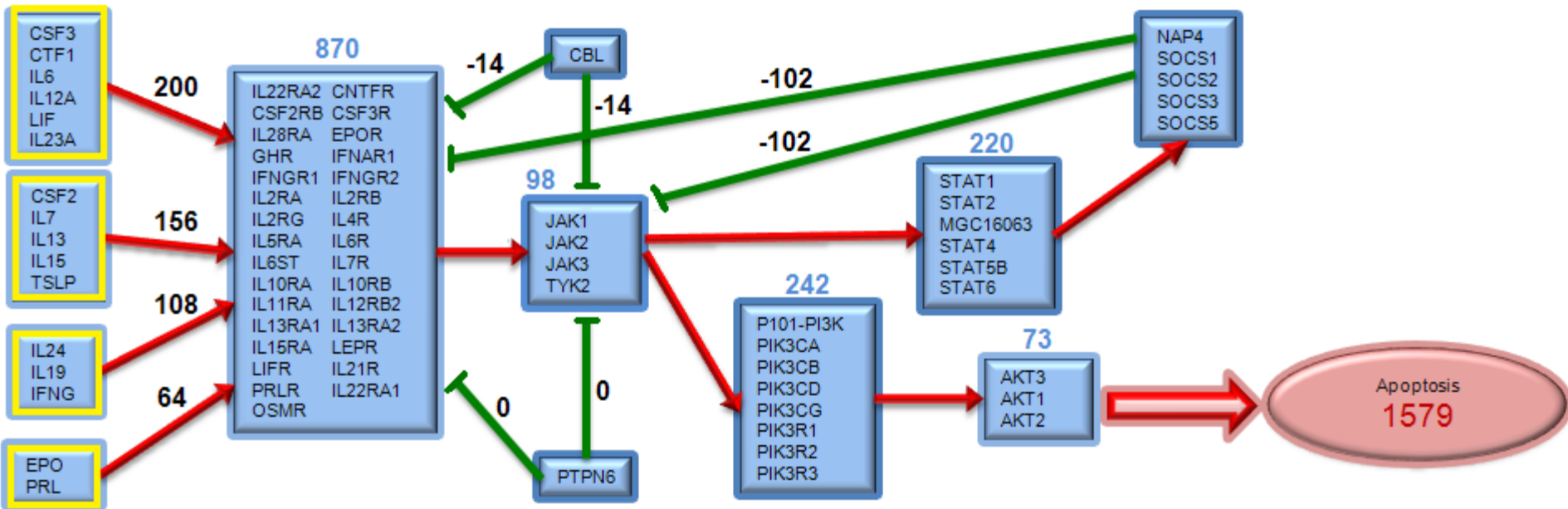




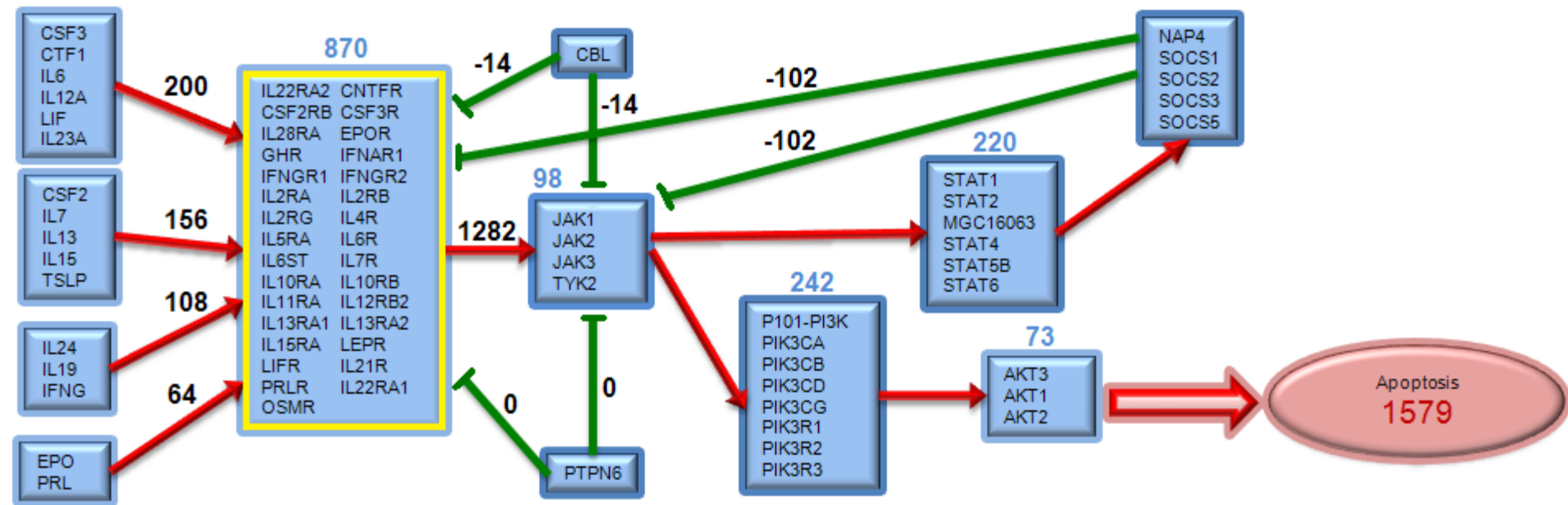
# Score Computation on Graph



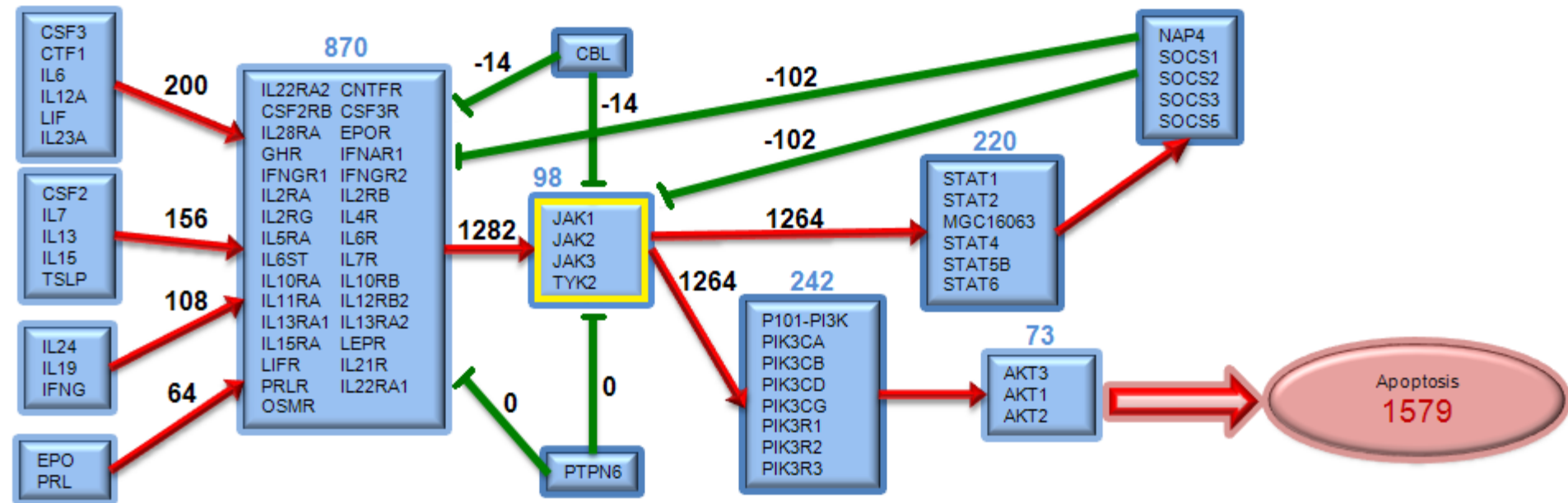
# Score Computation on Graph



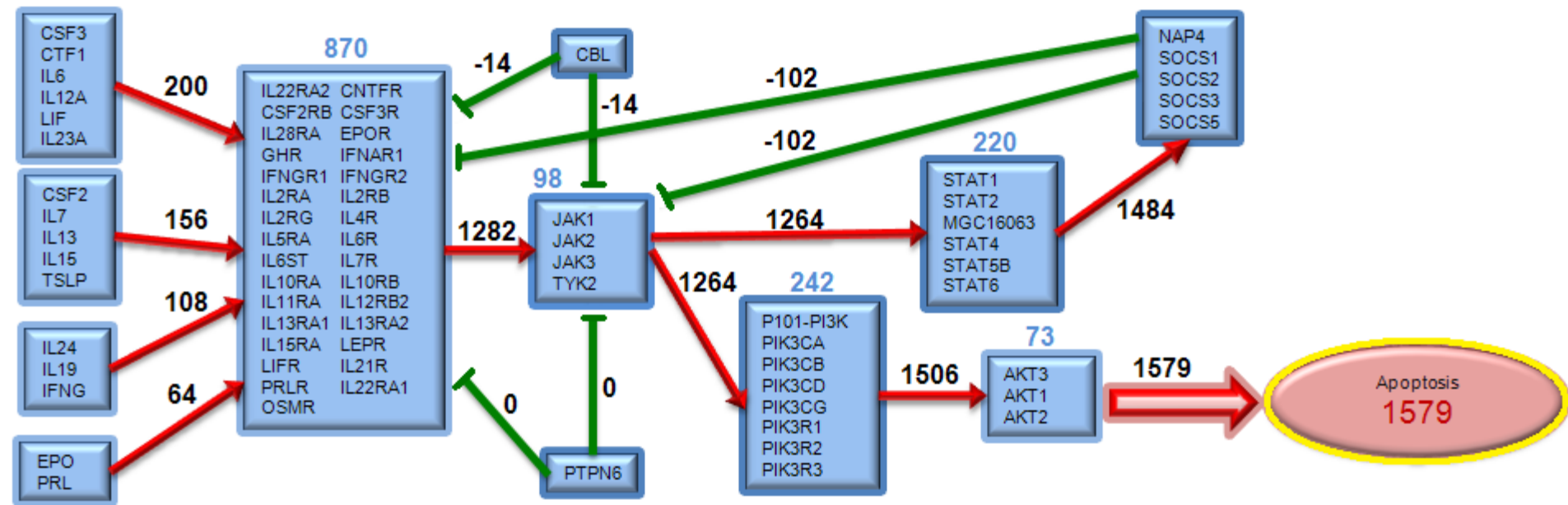
# Score Computation on Graph



# Score Computation on Graph



# Score Computation on Graph



# Scoring Measures of Outcome Process

➤  $TotalScore(P) = \sum_{s=1}^N outputScore(s)$

➤  $AverageScore(P) = \frac{TotalScore(P)}{M}$

# *Content*





- ✓ Analysis of Microarray Data
- ✓ ChIP-Seq Data
- ✓ Data Processing & Integration
- ✓ Scoring of Signaling Cascades
- ✓ **Results**

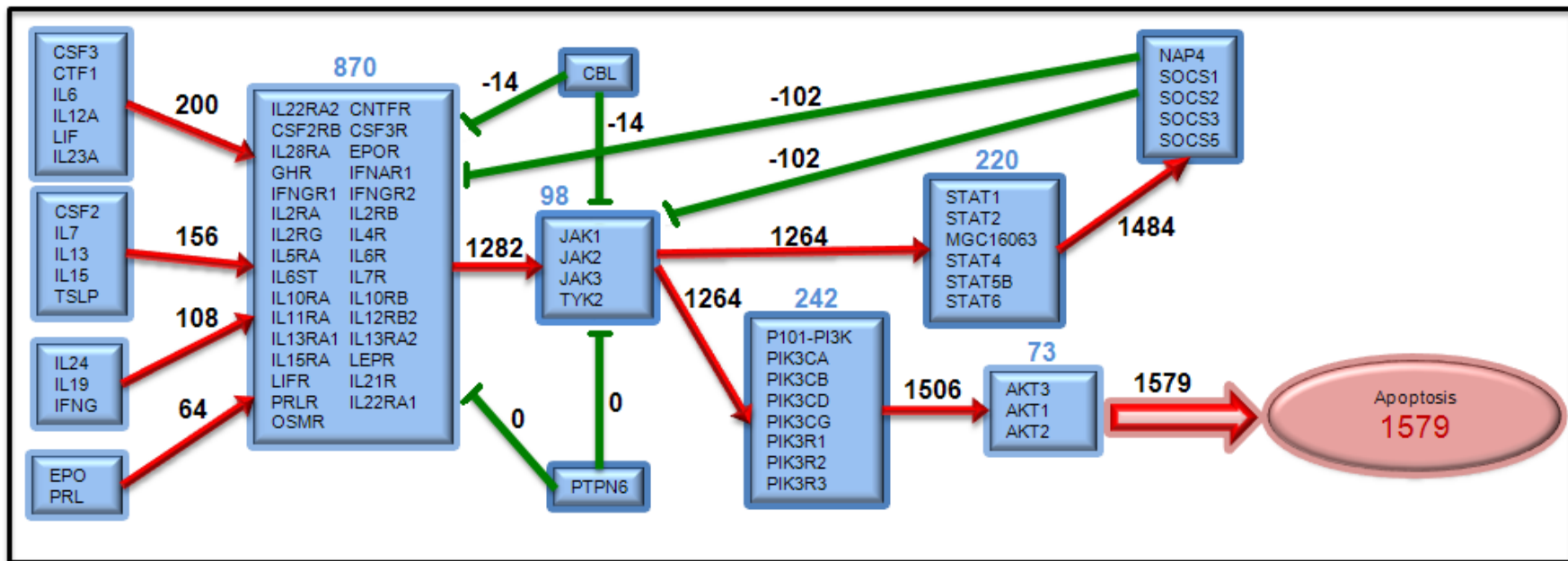
# Evaluated Signaling Cascades

- Jak-STAT
- TGF- $\beta$
- Apoptosis
- MAPK

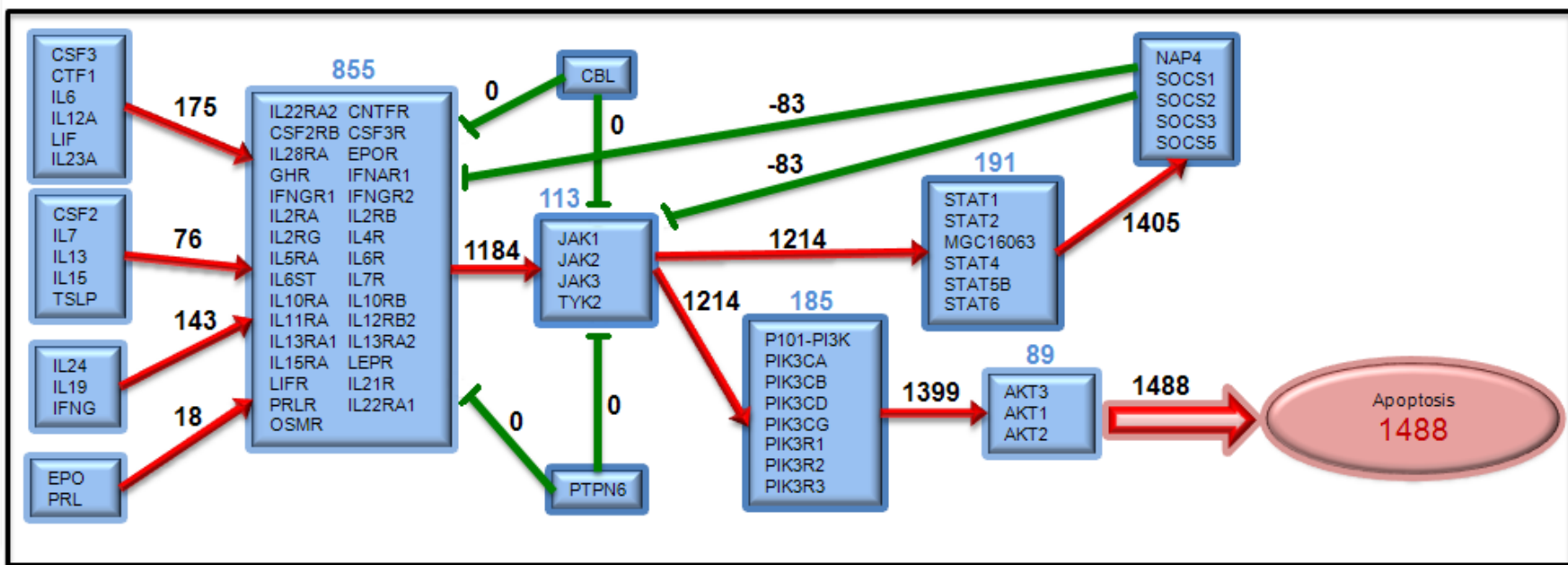


# Evaluated Signaling Cascades

- **Jak-STAT**  Apoptosis  
Cell cycle  
MAPK  
Ubiquitin mediated proteolysis
- **TGF- $\beta$**   Apoptosis  
Cell cycle  
MAPK
- **Apoptosis**  Survival  
Apoptosis  
Degradation
- **MAPK**  Apoptosis  
Cell cycle  
p53 signaling  
Wnt signaling  
Proliferation and differentiation



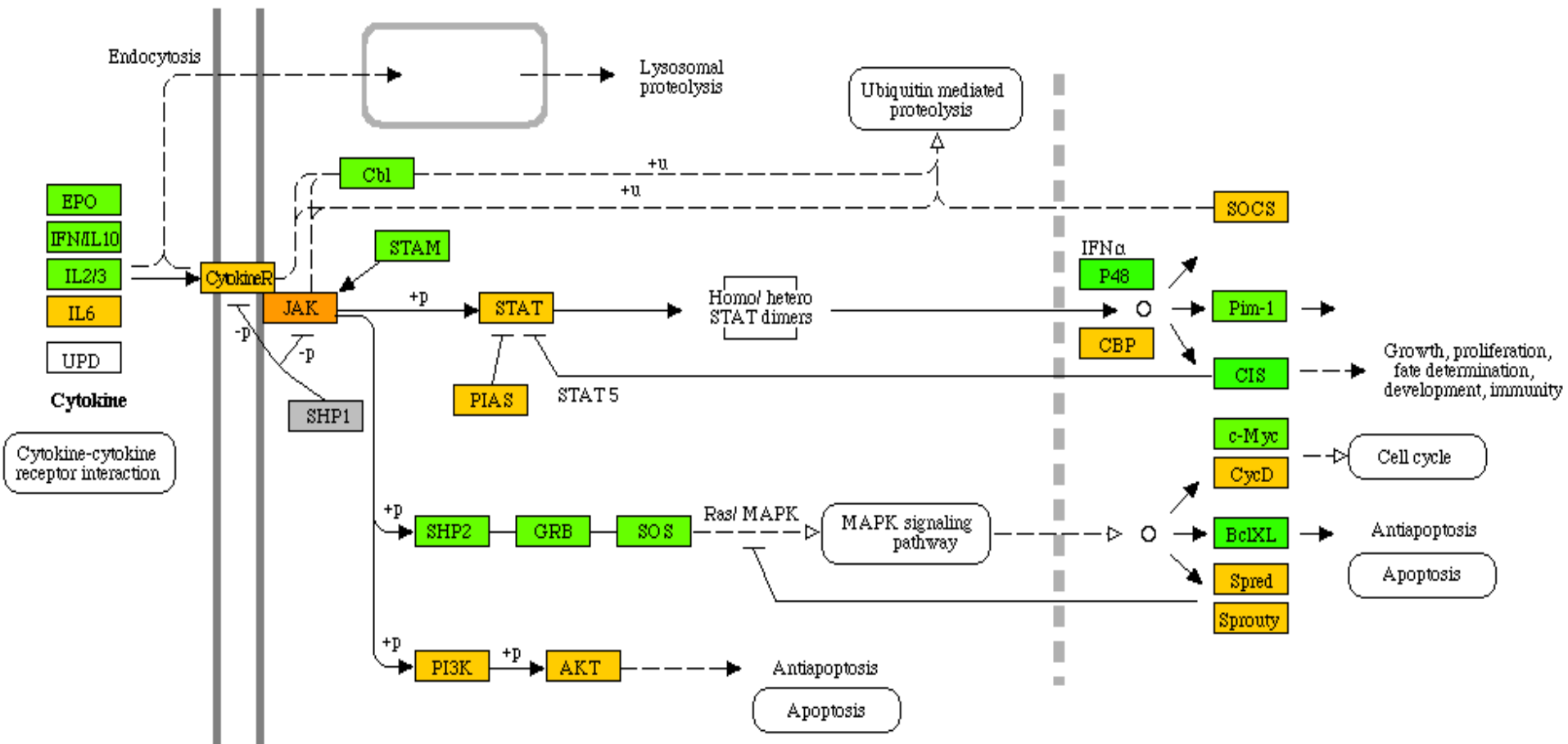
(A) Oxidative stress



(B) Control data

# Result of KegArray Tool

## JAK-STAT SIGNALING PATHWAY



# Enrichment Scores of Outcome Processes

| KEGG ID  | Enriched Pathway                | Control Sample |            | Oxidative Stress |            |
|----------|---------------------------------|----------------|------------|------------------|------------|
|          |                                 | Total Score    | Avg. Score | Total Score      | Avg. Score |
| hsa04630 | Apoptosis                       | 5633           | 217.80     | 5843             | 224.73     |
|          | Cell Cycle                      | 5447           | 209.50     | 5558             | 213.76     |
|          | Ubiquitin mediated proteolysis  | 2587           | 99.5       | 2754             | 105.92     |
|          | MAPK signaling                  | 1336           | 51.38      | 1358             | 52.23      |
| hsa04350 | Cell Cycle                      | 158            | 2.92       | 166              | 3.07       |
|          | MAPK signaling                  | 44             | 0.81       | 76               | 1.40       |
|          | Apoptosis                       | 52             | 0.96       | 66               | 1.22       |
| hsa04210 | Survival                        | 2222           | 37.66      | 2762             | 46.81      |
|          | Apoptosis                       | 2668           | 45.22      | 2709             | 45.91      |
|          | Degradation                     | 1984           | 33.62      | 2188             | 37.08      |
| hsa04010 | Proliferation & differentiation | 19346          | 172.73     | 22315            | 199.24     |
|          | Cell Cycle                      | 2533           | 22.61      | 3771             | 33.66      |
|          | Apoptosis                       | 1652           | 14.75      | 2949             | 26.33      |
|          | p53 signaling                   | 832            | 7.42       | 1135             | 10.13      |
|          | Wnt signaling                   | 185            | 1.65       | 288              | 2.57       |

**Table 1.** Enrichment scores of hsa04630 Jak-STAT signaling, hsa04350 TGF- $\beta$  signaling, hsa04210 Apoptosis, hsa04010 MAPK signaling pathways

# Discussion

- The scores obtained with control experiment are lower compared to oxidative stress scores.
- The most effected biological process under oxidative stress condition and transcription of OCT1 protein was *Apoptosis* process having the highest score between signaling cascades.
- Biologist should perform lab experiment to validate this cause and effect relation.

# Conclusion

- Our hybrid approach integrates large scale transcriptome data to quantitatively assess the weight of a signaling cascade under the control of a biological process.
- Signaling cascades in KEGG database were used as the models of the approach.
- The framework can be applicable to directed acyclic graphs.

# Future Work

- Different ranking methods on the transcriptome data will be analyzed.
- In order to provide comparable scores on signaling cascades, score computation method will be changed.
- Permutation tests will be included to provide significance levels for enrichment scores of signaling cascades.

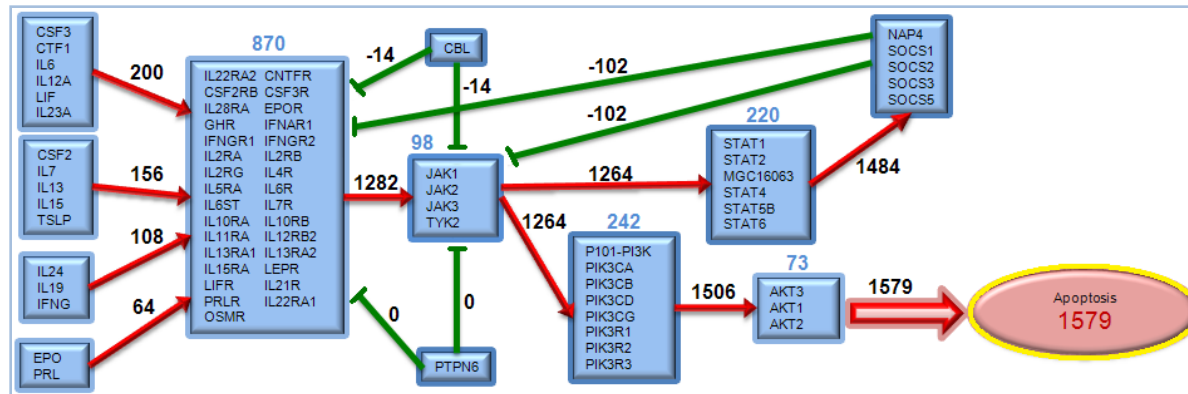
# Acknowledgement

- My colleagues:
  - *Prof.Dr. Volkan Atalay*
  - *Assoc. Prof. MD. Rengül Çetin-Atalay*
- Sharing their raw ChIP-seq data:
  - *Assist. Prof. Dr. Dean Tantin*
- Travel support:
  - *The Scientific and Technological Research Council of Turkey (TÜBİTAK)*



# Evaluation of Signaling Cascades Based on the Weights from Microarray and ChIP-seq Data

*Zerrin Işık, Volkan Atalay, and Rengül Çetin-Atalay*



*Middle East Technical University and Bilkent University  
Ankara - TURKEY*

