```
(* © Orly Alter 2016 *)
```

```
(* BIOEN 6770: Genomic Signal Processing *)
```

```
(* Lab 3 Solutions: Probability *)
```

```
(* General Commands *)
```

```
Clear["Global`*"]
```

```
(* Hypergeometric Distribution *)
```

```
(* 1. Define the P-Value Function by Using the Built-In Binomial Coefficient *)
```

```
? :=
```

lhs := *rhs* assigns *rhs* to be the delayed value of *lhs*. *rhs* is maintained in an unevaluated form. When *lhs* appears, it is replaced by *rhs*, evaluated afresh each time. ≫

```
? Binomial
```

Binomial[*n*, *m*] gives the binomial coefficient $\binom{n}{m}$. ≫

```
pValue1[k_, K_, m_, M_] :=
  N[Sum[Binomial[K, i] * Binomial[M - K, m - i], {i, k, m}] / Binomial[M, m]];
```

```
pValue1[84, 86, 86, 251]
pValue1[(251 - 86) - (86 - 84), 251 - 86, 251 - 86, 251]
```

8.01057×10^{-62}

8.01057×10^{-62}

```
pValue2[k_, K_, m_, M_] :=
  N[Sum[Binomial[K, i] * Binomial[M - K, m - i], {i, k, K}] / Binomial[M, m]];
```

```
pValue2[84, 86, 86, 251]
```

8.01057×10^{-62}

```
(* 2. Define the P-Value Function by Using the Built-In Hypergeometric Distribution *)
```

```
? HypergeometricDistribution
```

HypergeometricDistribution[n, n_{succ}, n_{tot}] represents a hypergeometric distribution. ≫

```
? PDF
```

PDF[*dist*, x] gives the probability density function for the symbolic distribution *dist* evaluated at x.
PDF[*dist*, {x_1, x_2, ...}] gives the multivariate
probability density function for a symbolic distribution *dist* evaluated at {x_1, x_2, ...}.
PDF[*dist*] gives the PDF as a pure function. ≫

```
pValue3[k_, K_, m_, M_] := N[PDF[HypergeometricDistribution[m, K, M], k]];

pValue3[84, 86, 86, 251]
```

8.00827×10^{-62}

```
? CDF
```

CDF[*dist*, x] gives the cumulative distribution function for the symbolic distribution *dist* evaluated at x.
CDF[*dist*, {x_1, x_2, ...}] gives the multivariate cumulative
distribution function for the symbolic distribution *dist* evaluated at {x_1, x_2, ...}.
CDF[*dist*] gives the CDF as a pure function. ≫

```
pValue4[k_, K_, m_, M_] := N[1 - CDF[HypergeometricDistribution[m, K, M], k - 1]];

pValue4[84, 86, 86, 251]
```

8.01057×10^{-62}

```
(* 3. Read the data by Spellman et al. –
 http://www.alterlab.org/teaching/BIOEN6770/labs/Spellman_Cell_Cycle.txt –
 and interpret each time point by calculating the P-value of the enrichment of the time
   point in overexpressed and underexpressed M/G1, G1, S, S/G2 and G2/M genes.
   Explain your steps and comment on your results. *)

(* Read Spellman_Cell_Cycle.txt *)

a = 1;
b = 8;

stream = "http://www.alterlab.org/teaching/BIOEN6770/labs/Spellman_Cell_Cycle.txt";
matrix = Import[stream, "Table"];
{genes, arrays} = Dimensions[matrix] - {a, b}
Clear[stream];
```

```
{565, 18}
```

```
arraynames = Transpose[Drop[Transpose[Drop[matrix, {a + 1, a + genes}]], {1, b}]];
matrix = Drop[matrix, 1];

matrix = Transpose[matrix];
genenames = Drop[matrix, {b + 1, b + arrays}];
matrix = Drop[matrix, {1, b}];
matrix = Transpose[matrix];
Clear[a, b];
```

```
(* Compute P-Value of Enrichment of Annotations Assuming Hypergeometric Distribution *)

(* Cell Cycle Annotations *)

annotations = genenames[[7]];
stages = {"M/G1", "G1", "S", "S/G2", "G2/M"};

most = 56;
numbers = Flatten[
  Table[{Count[Flatten[annotations], stages[[a]]]}, {a, 1, Dimensions[stages][[1]]}]]
{77, 216, 51, 87, 134}

counter = Table[{a}, {a, 1, arrays}];
parallelprobability = Table[{0}, {a, 1, Dimensions[counter][[1]]}];
antiprobability = Table[{0}, {a, 1, Dimensions[counter][[1]]}];

Do[{
  pattern = Transpose[Sort[Transpose[Join[{Transpose[matrix][[c]]}, {annotations}]],
      OrderedQ[{{#2}, {#1}}] &]][[2]],
  table = Table[{
     stages[[a]],
     numbers[[a]],
     Count[Flatten[Drop[pattern, {most + 1, genes}]], stages[[a]]]},
    {a, 1, Dimensions[stages][[1]]}],
  parallelprobability[[c]] = Table[
    ScientificForm[
     Sum[N[Binomial[table[[a, 2]], b] * Binomial[genes - table[[a, 2]], most - b] /
          Binomial[genes, most]], {b, table[[a, 3]], most}], 2],
    {a, 1, Dimensions[stages][[1]]}],
  table = Table[{
     stages[[a]],
     numbers[[a]],
     Count[Flatten[Drop[pattern, {1, genes - most}]], stages[[a]]]},
    {a, 1, Dimensions[stages][[1]]}],
  antiprobability[[c]] = Table[
    ScientificForm[
     Sum[N[Binomial[table[[a, 2]], b] * Binomial[genes - table[[a, 2]], most - b] /
          Binomial[genes, most]], {b, table[[a, 3]], most}], 2],
    {a, 1, Dimensions[stages][[1]]}]},
 {c, 1, Dimensions[counter][[1]]}]
```

```
table = Transpose[Join[
    Transpose[Join[{{"Arrays"}}, Transpose[arraynames]]],
    Transpose[Join[{stages}, parallelprobability]]]];
TableForm[table]
```

Arrays	M/G1	G1	S	S/G2	G2/M
0_min	1.3×10^{-4}	9.6×10^{-1}	$9.\times10^{-1}$	9.8×10^{-1}	3.4×10^{-1}
7_min	3.2×10^{-5}	1.	5.9×10^{-1}	$5.\times10^{-1}$	8.2×10^{-1}
14_min	4.4×10^{-3}	7.7×10^{-4}	1.	1.	1.
21_min	3.5×10^{-1}	1.2×10^{-7}	1.	1.	1.
28_min	9.9×10^{-1}	$2.\times10^{-3}$	2.1×10^{-2}	1.	9.5×10^{-1}
35_min	1.	1.8×10^{-1}	1.4×10^{-4}	9.5×10^{-1}	8.2×10^{-1}
42_min	1.	1.	2.1×10^{-2}	2.1×10^{-3}	8.4×10^{-2}
49_min	1.	1.	5.9×10^{-1}	$7.\times10^{-4}$	1.3×10^{-6}
56_min	9.6×10^{-1}	1.	1.	$5.\times10^{-1}$	1.2×10^{-15}
63_min	1.2×10^{-2}	1.	1.	9.8×10^{-1}	1.2×10^{-8}
70_min	3.2×10^{-5}	$6.\times10^{-1}$	1.	1.	4.6×10^{-1}
77_min	2.8×10^{-2}	$2.\times10^{-3}$	1.	1.	8.2×10^{-1}
84_min	8.1×10^{-1}	7.7×10^{-4}	7.7×10^{-1}	1.	7.2×10^{-1}
91_min	9.6×10^{-1}	1.2×10^{-1}	5.3×10^{-2}	1.	4.6×10^{-1}
98_min	1.	7.1×10^{-1}	2.1×10^{-2}	1.3×10^{-1}	7.2×10^{-1}
105_min	9.9×10^{-1}	9.8×10^{-1}	2.1×10^{-2}	3.5×10^{-1}	8.4×10^{-2}
112_min	9.9×10^{-1}	1.	7.7×10^{-1}	1.5×10^{-5}	$2.\times10^{-5}$
119_min	8.1×10^{-1}	1.	$9.\times10^{-1}$	6.6×10^{-1}	5.5×10^{-11}

```
table = Transpose[Join[
    Transpose[Join[{{"Arrays"}}, Transpose[arraynames]]],
    Transpose[Join[{stages}, antiprobability]]]];
TableForm[table]
```

Arrays	M/G1	G1	S	S/G2	G2/M
0_min	5.1×10^{-1}	9.2×10^{-1}	2.2×10^{-3}	1.	2.3×10^{-1}
7_min	9.6×10^{-1}	9.9×10^{-1}	7.1×10^{-3}	6.6×10^{-1}	4.5×10^{-2}
14_min	9.1×10^{-1}	1.	5.3×10^{-2}	2.3×10^{-1}	2.3×10^{-2}
21_min	2.8×10^{-2}	1.	$9. \times 10^{-1}$	9.8×10^{-1}	$2. \times 10^{-5}$
28_min	5.3×10^{-8}	1.	9.7×10^{-1}	1.	2.2×10^{-4}
35_min	1.2×10^{-9}	1.	1.	1.	1.1×10^{-2}
42_min	2.2×10^{-12}	$6. \times 10^{-1}$	1.	1.	9.9×10^{-1}
49_min	2.3×10^{-13}	4.9×10^{-1}	1.	1.	1.
56_min	1.2×10^{-2}	8.3×10^{-6}	$9. \times 10^{-1}$	1.	1.
63_min	6.7×10^{-1}	$1. \times 10^{-2}$	$3. \times 10^{-5}$	1.	1.
70_min	1.	9.9×10^{-1}	1.4×10^{-7}	5.8×10^{-3}	1.
77_min	1.	1.	5.8×10^{-4}	2.1×10^{-4}	2.3×10^{-2}
84_min	9.6×10^{-1}	1.	1.	$7. \times 10^{-2}$	9.7×10^{-13}
91_min	4.4×10^{-3}	1.	1.	1.	5.4×10^{-6}
98_min	$3. \times 10^{-7}$	9.6×10^{-1}	1.	1.	8.4×10^{-2}
105_min	5.3×10^{-8}	$8. \times 10^{-1}$	9.7×10^{-1}	1.	7.2×10^{-1}
112_min	1.6×10^{-6}	2.1×10^{-2}	1.	1.	1.
119_min	2.2×10^{-1}	2.9×10^{-5}	3.9×10^{-1}	1.	1.

```
(* 4. Compute the SVD of the Spellman data,
and interpret each of the five most significant "eigenarrays" by calculating the P-
 value of the enrichment of the eigenarray in overexpressed and underexpressed M/G1,
G1, S, S/G2 and G2/M genes.
  Explain your steps and comment on your results. *)

(* General Commands *)

Clear["Global`*"]

(* Read Spellman_Cell_Cycle.txt *)

a = 1;
b = 8;

stream = "http://www.alterlab.org/teaching/BIOEN6770/labs/Spellman_Cell_Cycle.txt";
matrix = Import[stream, "Table"];
{genes, arrays} = Dimensions[matrix] - {a, b}
Clear[stream];
```

```
{565, 18}
```

```
arraynames = Transpose[Drop[Transpose[Drop[matrix, {a + 1, a + genes}]], {1, b}]];
matrix = Drop[matrix, 1];

matrix = Transpose[matrix];
genenames = Drop[matrix, {b + 1, b + arrays}];
matrix = Drop[matrix, {1, b}];
matrix = Transpose[matrix];
Clear[a, b];

(* SVD of the Measured Cell Cycle Data *)

{rows, columns} = Dimensions[matrix]
{u, sigma, v} = SingularValueDecomposition[matrix];
rank = columns - Count[Diagonal[sigma], 0.]
{u, sigma, v} = SingularValueDecomposition[matrix, rank];
Dimensions[u]
Dimensions[sigma]
Dimensions[v]
```

```
{565, 18}
```

```
18
```

```
{565, 18}
```

```
{18, 18}
```

```
{18, 18}
```

```
(* Compute P-Value of Enrichment of Annotations Assuming Hypergeometric Distribution *)

(* Cell Cycle Annotations *)

annotations = genenames[[7]];
stages = {"M/G1", "G1", "S", "S/G2", "G2/M"};

most = 56;
numbers = Flatten[
  Table[{Count[Flatten[annotations], stages[[a]]]}, {a, 1, Dimensions[stages][[1]]}]]
{77, 216, 51, 87, 134}

counter = Table[{a}, {a, 1, 5}];
parallelprobability = Table[{0}, {a, 1, Dimensions[counter][[1]]}];
antiprobability = Table[{0}, {a, 1, Dimensions[counter][[1]]}];

Do[{
  pattern = Transpose[Sort[Transpose[Join[{Transpose[u][[c]]}, {annotations}]],
      OrderedQ[{{#2}, {#1}}] &]][[2]],
  table = Table[{
     stages[[a]],
     numbers[[a]],
     Count[Flatten[Drop[pattern, {most + 1, genes}]], stages[[a]]]},
    {a, 1, Dimensions[stages][[1]]}],
  parallelprobability[[c]] = Table[
    ScientificForm[
     Sum[N[Binomial[table[[a, 2]], b] * Binomial[genes - table[[a, 2]], most - b] /
          Binomial[genes, most]], {b, table[[a, 3]], most}], 2],
    {a, 1, Dimensions[stages][[1]]}],
  table = Table[{
     stages[[a]],
     numbers[[a]],
     Count[Flatten[Drop[pattern, {1, genes - most}]], stages[[a]]]},
    {a, 1, Dimensions[stages][[1]]}],
  antiprobability[[c]] = Table[
    ScientificForm[
     Sum[N[Binomial[table[[a, 2]], b] * Binomial[genes - table[[a, 2]], most - b] /
          Binomial[genes, most]], {b, table[[a, 3]], most}], 2],
    {a, 1, Dimensions[stages][[1]]}]},
 {c, 1, Dimensions[counter][[1]]}]
```

```
table = Transpose[Join[
    Transpose[Join[{{"Eigenarrays"}}, counter]],
    Transpose[Join[{stages}, parallelprobability]]]];
TableForm[table]
```

Eigenarrays	M/G1	G1	S	S/G2	G2/M
1	1.5×10^{-3}	9.6×10^{-1}	9.7×10^{-1}	1.	2.3×10^{-2}
2	1.	$8.\times10^{-1}$	2.1×10^{-2}	8.9×10^{-1}	4.5×10^{-2}
3	1.	1.	9.7×10^{-1}	3.4×10^{-2}	9.7×10^{-13}
4	1.	$6.\times10^{-1}$	1.4×10^{-7}	3.4×10^{-2}	1.
5	8.1×10^{-1}	1.	2.3×10^{-1}	$5.\times10^{-1}$	1.3×10^{-6}

```
table = Transpose[Join[
    Transpose[Join[{{"Eigenarrays"}}, counter]],
    Transpose[Join[{stages}, antiprobability]]]];
TableForm[table]
```

Eigenarrays	M/G1	G1	S	S/G2	G2/M
1	3.5×10^{-1}	4.9×10^{-1}	$9.\times10^{-1}$	9.8×10^{-1}	8.4×10^{-2}
2	7.4×10^{-6}	9.8×10^{-1}	7.7×10^{-1}	9.8×10^{-1}	5.9×10^{-1}
3	3.5×10^{-1}	1.4×10^{-11}	1.	1.	1.
4	1.7×10^{-10}	9.6×10^{-1}	1.	1.	4.6×10^{-1}
5	4.4×10^{-3}	$4.\times10^{-2}$	5.3×10^{-2}	1.	1.

```
(* 5. Compute the SVD of the data by Nielsen et al. –
 http://www.alterlab.org/teaching/BIOEN6770/labs/Nielsen_Sarcoma.txt –
 and interpret each of the five most significant "eigenarrays" by using GOrilla at –
 http://cbl-gorilla.cs.technion.ac.il/ .
      Explain your steps and comment on your results. *)

(* Read Nielsen_Sarcoma.txt *)

stream = "http://www.alterlab.org/teaching/BIOEN6770/labs/Nielsen_Sarcoma.txt";
matrix = Import[stream, "Table"];

(* Drop Header Rows *)

matrix = Drop[matrix, 1];

(* Save Gene Symbols *)

geneSymbol = Transpose[matrix][[2]];
Dimensions[geneSymbol]
```

```
{5520}
```

```
(* Drop Header Columns *)

matrix = Transpose[matrix];
matrix = Drop[matrix, {1, 4}];
matrix = Transpose[matrix];

{rows, columns} = Dimensions[matrix]
Clear[stream];
```

```
{5520, 46}
```

```
(* SVD of the Measured Sarcoma Data *)

{rows, columns} = Dimensions[matrix]
{u, sigma, v} = SingularValueDecomposition[matrix];
rank = columns - Count[Diagonal[sigma], 0.]
{u, sigma, v} = SingularValueDecomposition[matrix, rank];
Dimensions[u]
Dimensions[sigma]
Dimensions[v]
```

```
{5520, 46}
```

```
46
```

```
{5520, 46}
```

```
{46, 46}
```

```
{46, 46}
```

```
(* Write Eigenarrays *)

Dimensions[u]
Dimensions[geneSymbol]
```

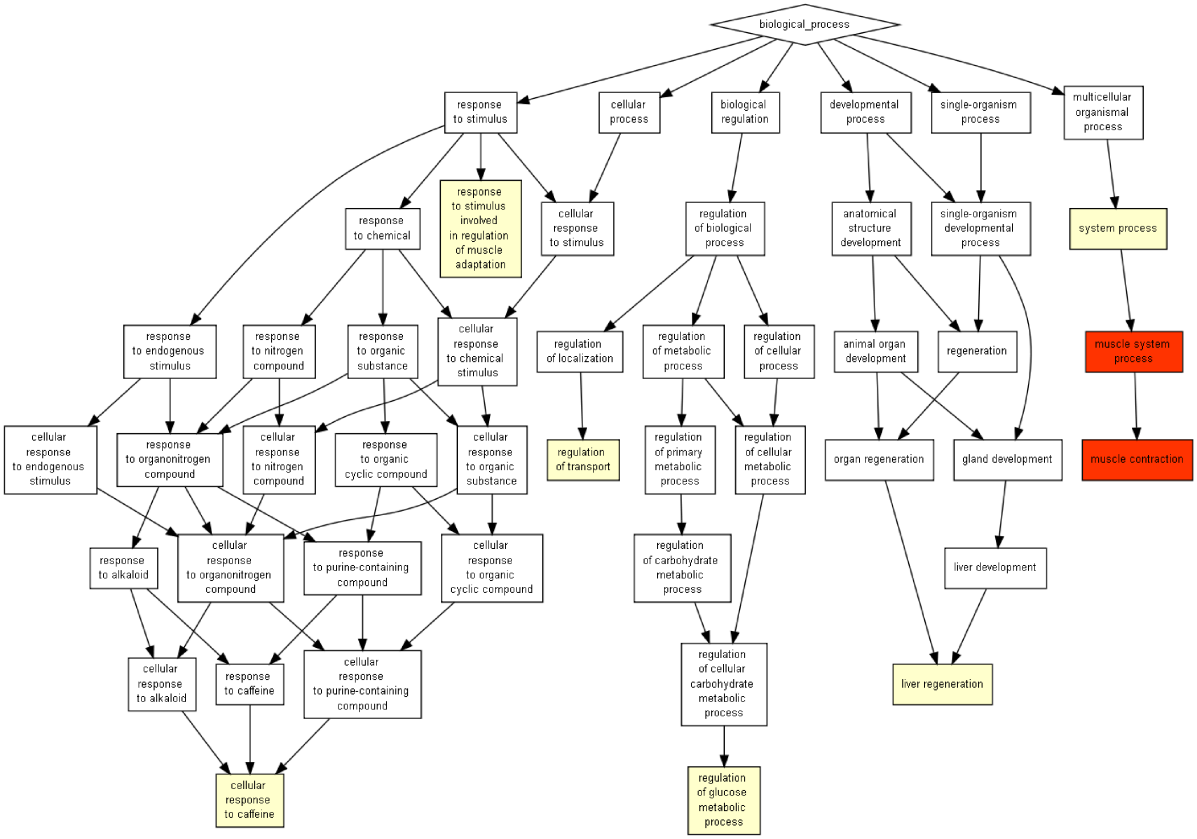
```
{5520, 46}
```

```
{5520}
```

```
u = Transpose[Join[{geneSymbol}, Transpose[u]]];
u = Join[{Table[a, {a, 0, 46}]}, u];
u[[1, 1]] = "Eigenarray"; Export["Desktop/Eigenarrays_Nielsen_Sarcoma.txt", u, "Table"]
```

```
Desktop/Eigenarrays_Nielsen_Sarcoma.txt
```

```
(* At a cutoff of n=200, the most significant GO process
 enrichment among the overexpressed genes in Eigenarray 3 is GO:0006936,
muscle contraction, with a hypergeometric P-value < 10^(-9). *)
```



```
(* 6. Extra Credit: Visualize the numerical results from Problem 3. *)
```

```
? BarChart
```

BarChart[{y_1, y_2, ...}] makes a bar chart with bar lengths y_1, y_2, ...
BarChart[{..., w_i[y_i, ...], ..., w_j[y_j, ...], ...}] makes a bar chart with bar features defined by the symbolic wrappers w_k.
BarChart[{$data_1$, $data_2$, ...}] makes a bar chart from multiple datasets $data_i$. ≫