```
(* © Orly Alter 2017 *)
```

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

```
(* Lab 3: 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 $\binom{n}{m}$. ≫

```
? Sum
```

Sum[*f*, {*i*, i_{max}}] evaluates the sum $\sum_{i=1}^{i_{max}} f$.

Sum[*f*, {*i*, i_{min}, i_{max}}] starts with $i = i_{min}$.
Sum[*f*, {*i*, i_{min}, i_{max}, *di*}] uses steps *di*.
Sum[*f*, {*i*, {i_1, i_2, ...}}] uses successive values i_1, i_2, ...

Sum[*f*, {*i*, i_{min}, i_{max}}, {*j*, j_{min}, j_{max}}, ...] evaluates the multiple sum $\sum_{i=i_{min}}^{i_{max}} \sum_{j=j_{min}}^{j_{max}} \ldots f$.

Sum[*f*, *i*] gives the indefinite sum $\sum_i f$. ≫

```
(* 2. Define a 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. ≫

```
? 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. ≫

```
(* 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. *)
```

```
(* 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. *)
```

```
(* 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. *)
```

```
(* 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*$_i$. ≫