Modeling and Simulation 3 and Finalize Lectures

BCB 5250 Introduction to Bioinformatics II

Spring 2020

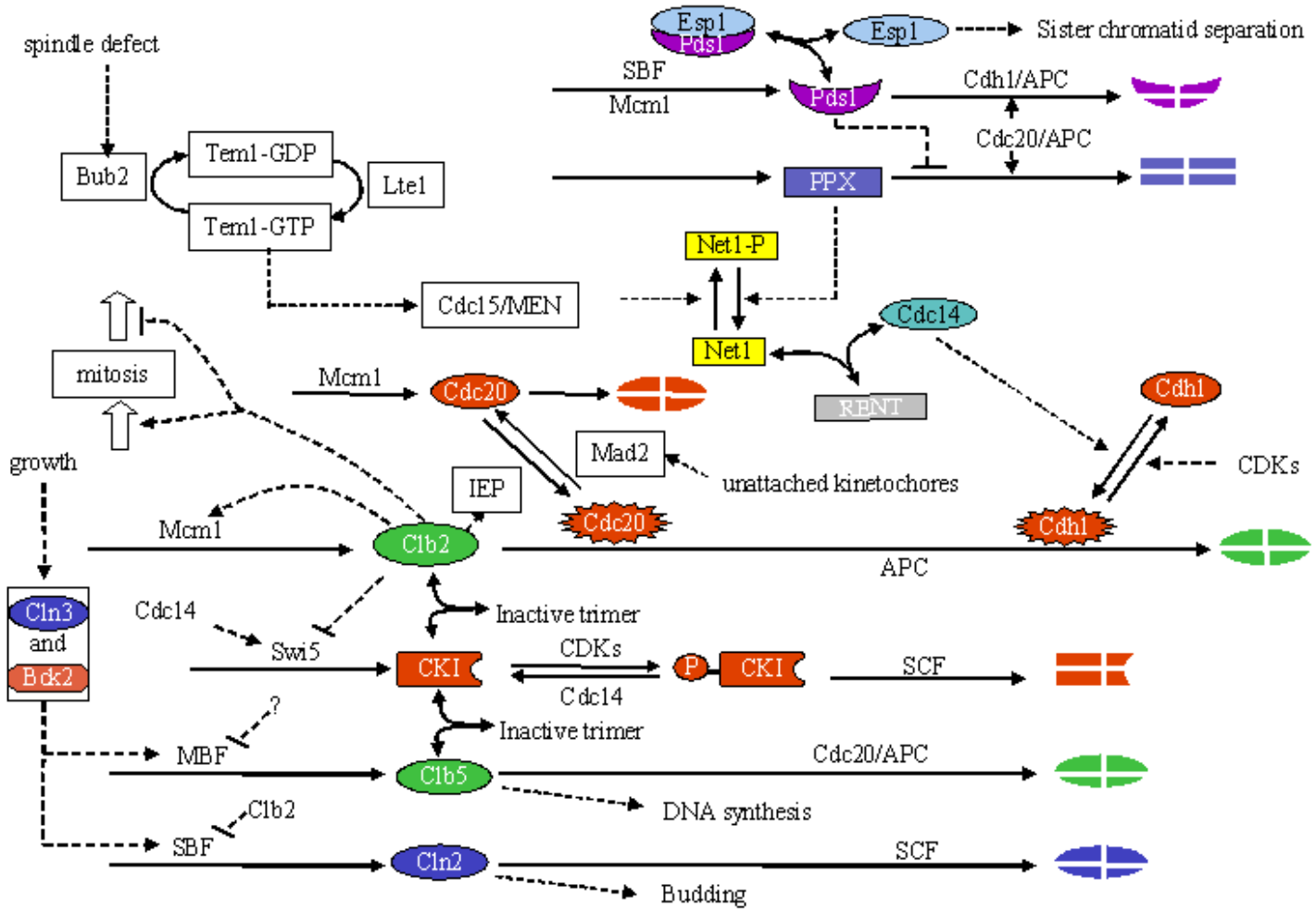
Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

Stochastic Simulation

- Stochastic Simulation

Budding Yeast Cell Cycle Model



The three major cell cycle events - DNA replication, mitosis and cell division, are controlled by **cyclin-dependent kinases (Cdks)**, which are active only when bound to a cyclin partner.

SBML

```
<?xml version="1.0" encoding="UTF-8"?>
<sbml xmlns="http://www.sbml.org/sbml/level2" lev
<model id="Ec_iAF1260" name="Ec_iAF1260">
<listOfCompartments>
  <compartment id="in" outside="out"/>
  <compartment id="out" />
</listOfCompartments>
<listOfSpecies>
  <species id="A_out" name="A_out" boundaryCondit
  <species id="A_in" name="A_in" boundaryConditio
  <species id="B_in" name="B_in" boundaryConditio
  <species id="B_out" name="B_out" boundaryCondit
</listOfSpecies>
<listOfReactions>
  <reaction id="R1" name="R1" reversible="false":
    <listOfReactants>
      <speciesReference species="A_out" stoichion
    </listOfReactants>
    <listOfProducts>
      <speciesReference species="A_in" stoichiome
    </listOfProducts>
  </reaction>
  <reaction id="R2" name="R2" reversible="false":
    <listOfReactants>
      <speciesReference species="A_in" stoichiome
    </listOfReactants>
    <listOfProducts>
      <speciesReference species="B_in" stoichiome
    </listOfProducts>
  </reaction>
  <reaction id="R3" name="R3" reversible="false":
    <listOfReactants>
      <speciesReference species="B_in" stoichiome
    </listOfReactants>
    <listOfProducts>
      <speciesReference species="B_out" stoichion
    </listOfProducts>
  </reaction>
</listOfReactions>
</model>
</sbml>
```

5 Units Definitions
A named definition of a new unit of measure, or a redefinition of an existing SBML default unit.

1 Compartments
A well-stirred container of a particular type and finite size where species may be located.

163 Parameters
A quantity with a symbolic name. In SBML, the term parameter is used in a generic sense to refer to named quantities regardless of whether they are constants or variables in a model.

54 Species
A pool of entities of the same species type located in a specific compartment.

94 Reactions
A statement describing some transformation, transport or binding process that can change the amount of one or more species.

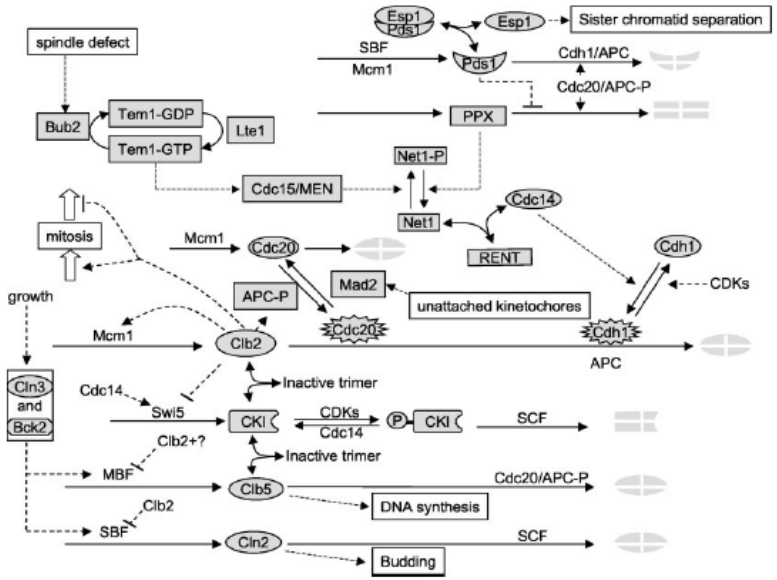
35 Rules
A mathematical expression added to the set of equations constructed based on the reactions defined in a model.

6 Functions
A named mathematical function that may be used throughout the rest of a model.

4 Events
A statement describing an instantaneous, discontinuous change in a set of variables of any type when a triggering condition is satisfied.

DE System
For each specie, sum of rections kinetic laws, taken with "+" if the specie is produced or with "-" if the specie is consumed.

View All



Why Stochastic?

- The continuous deterministic approach: adequate for understanding the average behavior of cells.
- At the microscopic, or molecular level, events are discrete – variables are not continuous.
- In the small volume of a bacterial cell (0.5 μm^3), no more than 10's to perhaps 100's of control proteins - random fluctuations can be very important to outcomes.
- A successful mathematical model based on computational cell biology has to explain not only a single wild-type strain but also various mutants that cannot be explained well by the deterministic method.
- Thus, the discrete stochastic approach can provide more accurate results than the continuous deterministic one for many biological systems.

A discrete modeling framework

- In developing a stochastic modeling framework for chemical reactions networks, we will continue to assume spatial homogeneity, and a fixed volume.
- The abundance of each chemical species will be described by the number of molecules in the reaction volume.
- The state of the system is then the vector *N* of molecule counts. (Recall, in a differential equation-based model, the state is the vector s of species concentrations.)
- As the stochastic dynamics proceed, the molecule counts will change their values in discrete jumps (in contrast to the smooth changes in concentration values that occur in differential equation models).

A discrete modeling framework

- We will characterize each reaction in the network by a stoichiometry vector *s*, and a propensity function *a*.
- For each reaction, the stoichiometry vector indicates the identity and number of reactants and products in a reaction: the *j*-th component of this vector is the net number of molecules of species *j* produced or consumed in the reaction. The propensity is a description of reaction rate.

A discrete modeling framework

- To illustrate these ideas, consider the network composed of the two reactions

$$R_1 = A + B \xrightarrow{k_1} C \qquad R_2 = C \xrightarrow{k_2}$$

- The state of this system will be described by the numbers of molecules of species *A*, *B*, and *C* present at any given time. The stoichiometry vectors are

$$s_1 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \begin{matrix} \leftarrow A \\ \leftarrow B \\ \leftarrow C \end{matrix} \qquad s_2 = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \begin{matrix} \leftarrow A \\ \leftarrow B \\ \leftarrow C \end{matrix}$$

A discrete modeling framework

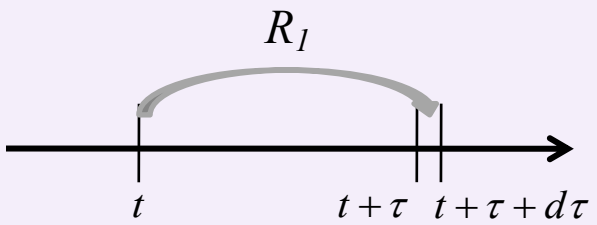
- The reaction propensities are functions of reactant abundance.
- We will assume that the probability of a reaction event is proportional to the product of the abundance of each reactant species (as in mass action).
- The propensities for this example are then

$$a_1(N) = k_1 N_A N_B \qquad a_2(N) = k_2 N_C$$

Gillespie Algorithm

- Stochastic Simulation Algorithm (SSA), which is also called Gillespie Algorithm, generates a statistically correct trajectory of a stochastic equation using Monte Carlo Methods.

- When will the next reaction occur?
- What kind of reaction it will be?



- N species $\{S_1,...,S_N\}$ M reactions $\{R_1,...,R_M\}$.
- $X(t)=(X_1(t),...,X_N(t))$ number of molecules.
- $a_j(x)\,dt$ is the probability, given $X(t)=x$
- R_j will be fired in next dt.
- The time for the next reaction to occur is
- $$\tau=\frac{1}{a_0(x)}\ln\left(\frac{1}{r_1}\right).$$
- The next reaction index j is given by
- $$\sum_{j'=1}^{j}a_{j'}(x)>r_2a_0(x).$$
- Update system $X(t+\tau):=x+\nu_j$,and iterate.

Determining the next reaction

- The probability that a particular reaction will occur is proportional to the propensity of the reaction.
- Consider a network that involves three reactions, R_1, R_2, and R_3, with propensities a_1, a_2, and a_3.
- Let $P(R = R_i)$ denote the probability that R_i will be the next reaction to occur. Probability $P(R = R_i)$ is proportional to the propensity a_i of reaction R_i.
- Together, these probabilities sum to one. The probability distribution is:

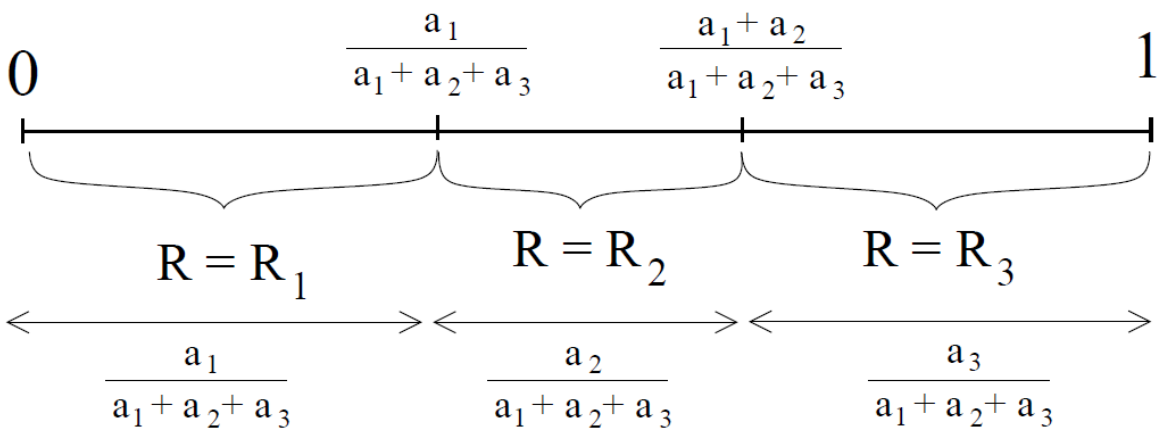
$$P(R = R_1) = \frac{a_1}{a_1 + a_2 + a_3}$$

$$P(R = R_2) = \frac{a_2}{a_1 + a_2 + a_3}$$

$$P(R = R_3) = \frac{a_3}{a_1 + a_2 + a_3}$$

Determining the next reaction

- In order to implement a simulation of this network's behavior, we need to sample from this probability distribution.
- Most numerical software packages have built-in functions that generate random numbers drawn uniformly between zero and one.
- We divide the zero-to-one interval into three subintervals—one for each reaction as



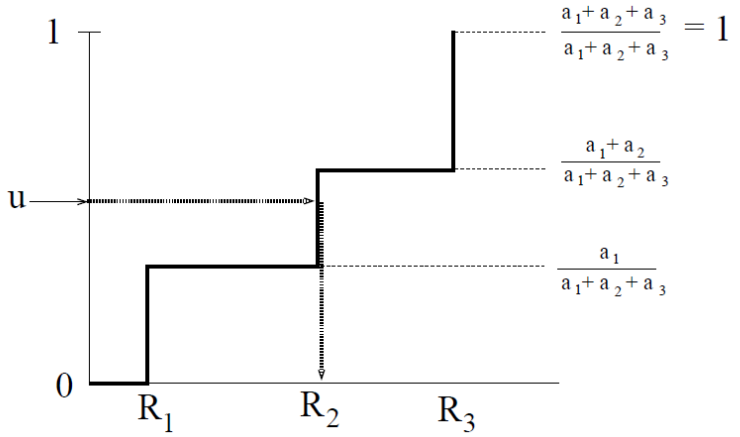
- The length of each subinterval is equal to the probability of the corresponding reaction.

Determining the next reaction

- Cumulative distribution function for the random variable *R*.
- The height of the staircase graph corresponds to the cumulative probability as in

$$\begin{aligned}
&\text{if} \quad 0 \leq u \leq \frac{a_1}{a_1+a_2+a_3}, \quad \text{then we set } R = R_1 \\
&\text{if} \quad \frac{a_1}{a_1+a_2+a_3} < u \leq \frac{a_1+a_2}{a_1+a_2+a_3}, \quad \text{then we set } R = R_2 \\
&\text{if} \quad \frac{a_1+a_2}{a_1+a_2+a_3} < u \leq \frac{a_1+a_2+a_3}{a_1+a_2+a_3} = 1, \quad \text{then we set } R = R_3.
\end{aligned}$$

A reaction is chosen by selecting a number u from the uniform distribution on the horizontal axis and then extending a horizontal line to the staircase graph.

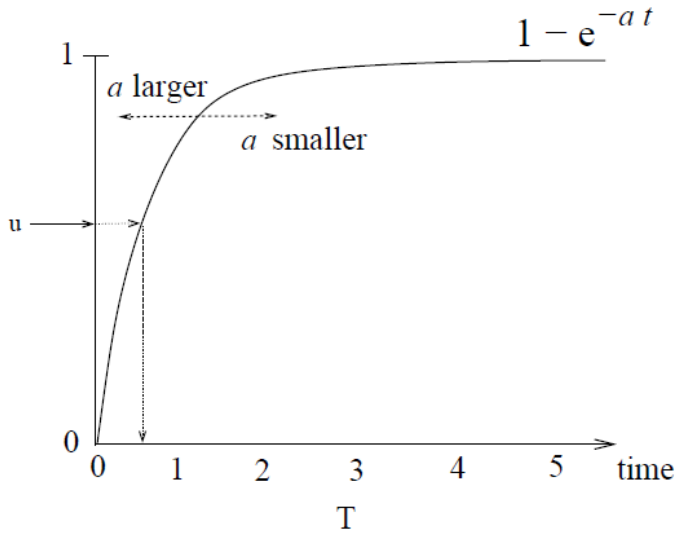


Determining the time to the next reaction

- The time *T* that elapses between reactions is also a random variable.
- The cumulative distribution function for *T* is given by:

$$P(0 \leq T \leq t) = 1 - e^{-at}$$

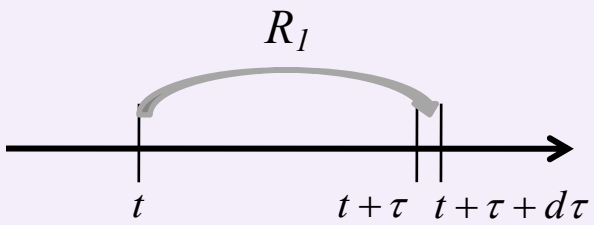
where a is the sum of the reaction propensities.



Gillespie Algorithm

- Stochastic Simulation Algorithm (SSA), which is also called Gillespie Algorithm, generates a statistically correct trajectory of a stochastic equation using Monte Carlo Methods.

- When will the next reaction occur?
- What kind of reaction it will be?



- N species $\{S_1,...,S_N\}$ M reactions $\{R_1,...,R_M\}$.
- $X(t)=(X_1(t),...,X_N(t))$ number of molecules.
- $a_j(x)\,dt$ is the probability, given $X(t)=x$
- R_j will be fired in next dt.
- The time for the next reaction to occur is
- $$\tau=\frac{1}{a_0(x)}\ln\left(\frac{1}{r_1}\right).$$
- The next reaction index j is given by
- $$\sum_{j'=1}^{j}a_{j'}(x)>r_2a_0(x).$$
- Update system $X(t+\tau):=x+\nu_j$,and iterate.

Model Conversion

- Deterministic model is written in normalized concentration form, because normalized concentrations are what experimenters measure in the lab.
- Stochastic simulation requires the model to be in terms of populations because they account for individual molecules.
- Requires "Scaling Factor" for species and volume information

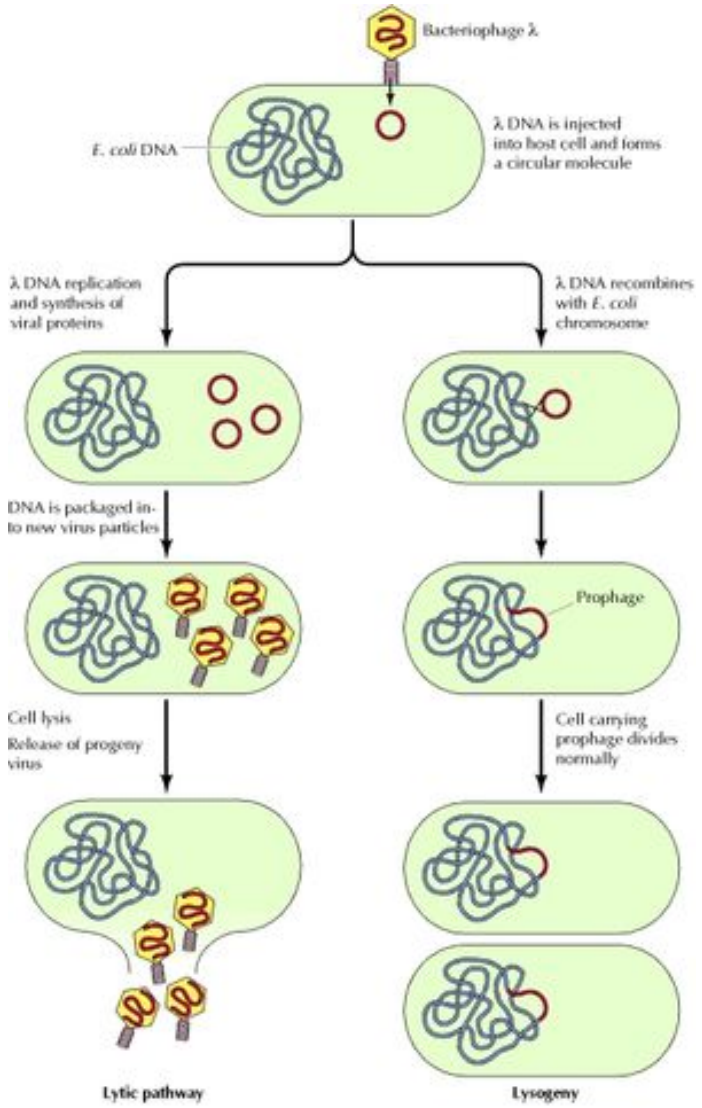
Notation of Species S

Notation	Description
$[S]_n$	Normalized concentration
$[S]$	The real concentration
N_S	The number of molecules (population) of species S

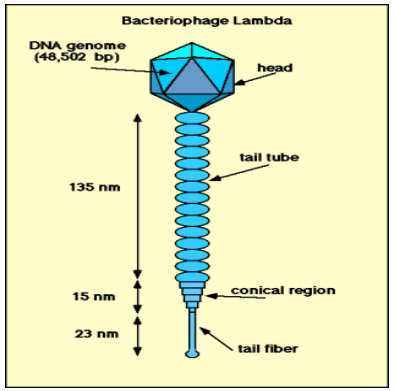
$$[S] = c_S[S]_n$$

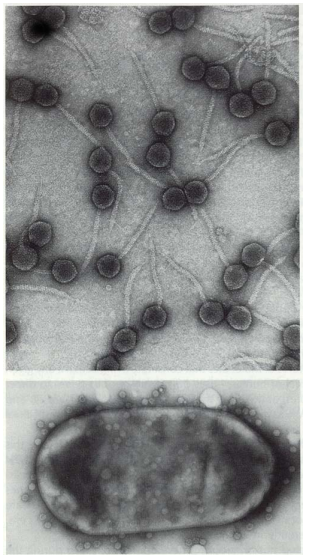
$$N_S = V[S]$$

Lambda-phage affected *E.coli*



- Build a comprehensive and efficient model that replicates the bacteriophage lambda genetic switch between lysis and lysogeny.





E. coli

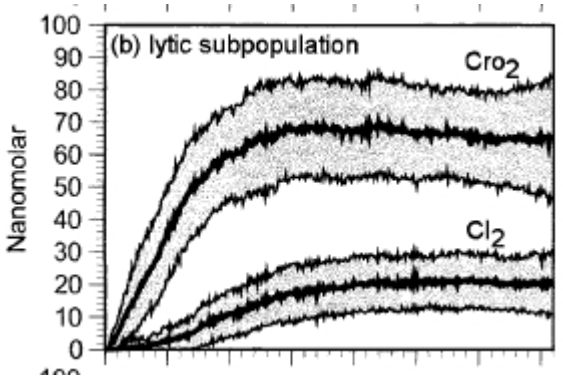
From: Ptashne 1995

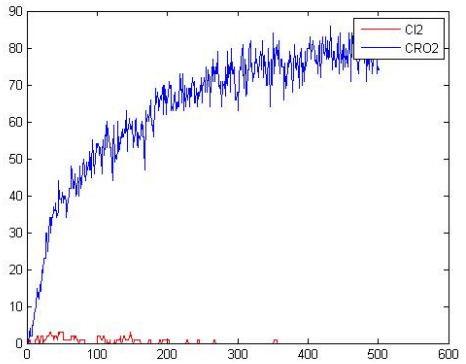
Reaction and Parameters for Lambda Switch

12 Species	21 Reactions		C_i	C_{-i}
Cl	$\rightarrow$		0.015	
Cl + Cl	$\leftrightarrow$	Cl_2	0.01	0.25
D + Cl_2	$\leftrightarrow$	D_1	9.768	15.0
D_1 + Cl_2	$\leftrightarrow$	D_2	9.768	15.0
D_2 + Cl_2	$\leftrightarrow$	D_3	9.768	222.38
D_2 + P	$\rightarrow$	D_2 + P + Cl	0.5	
CRO	$\rightarrow$		0.005	
CRO + CRO	$\leftrightarrow$	CRO_2	0.01	0.25
D + CRO_2	$\leftrightarrow$	D'_3	9.768	29.77
D'_3 + CRO_2	$\leftrightarrow$	D'_2	9.768	245.371
D'_2 + CRO_2	$\leftrightarrow$	D'_1	9.768	245.371
D + P	$\rightarrow$	D + P + CRO	0.5	
D'_3 + P	$\rightarrow$	D'_3 + P + CRO	0.5	

Lambda Switch Results

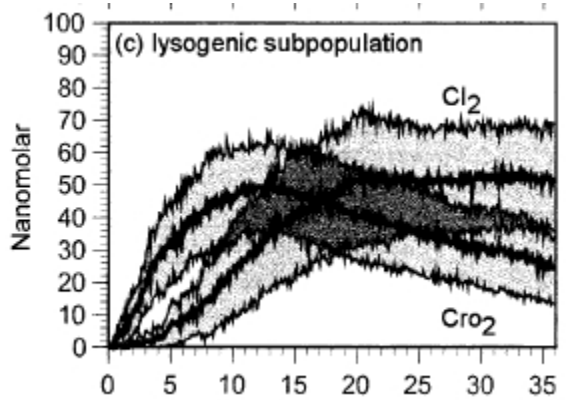
Lysis

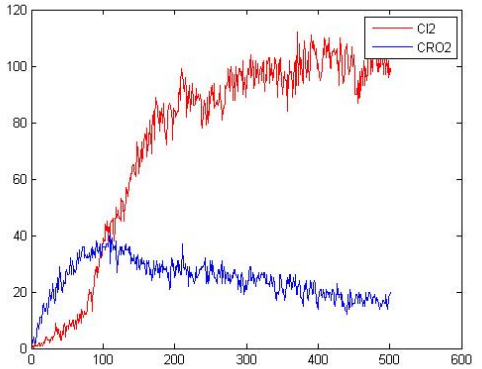




9times go to Lysis in 20 simulations

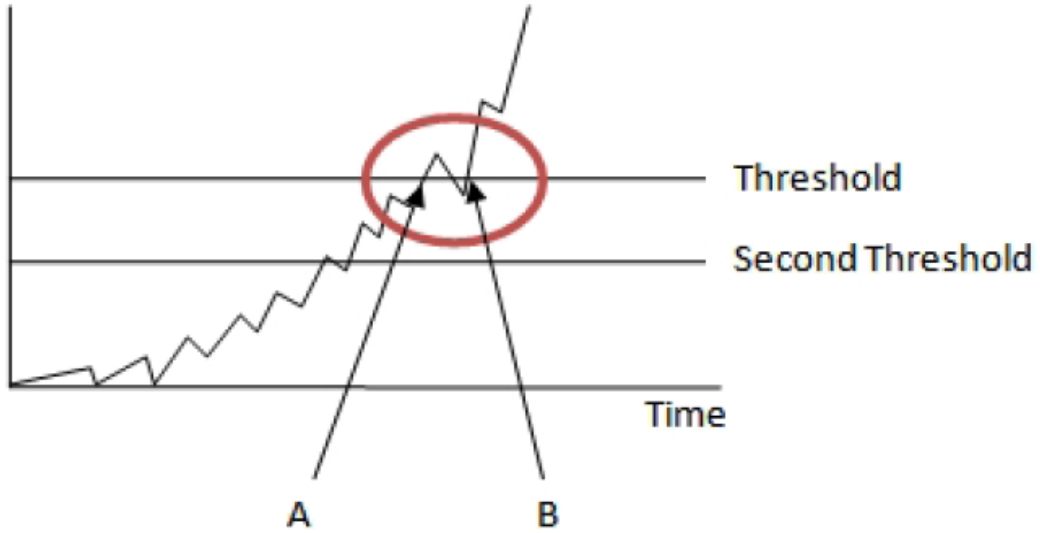
Lysogeny



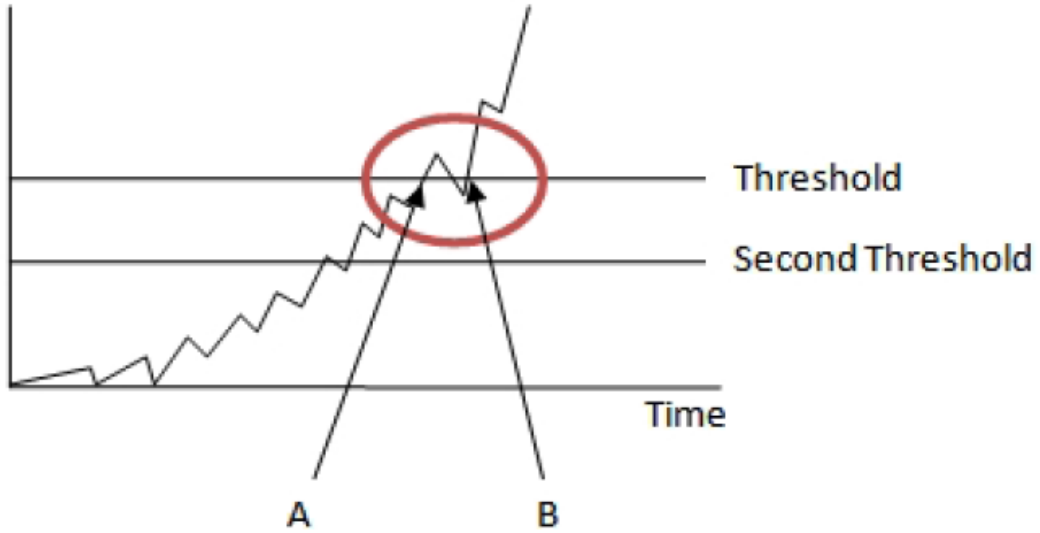


11times go to Lysogeny in 20 simulations

Difficulty of Stochastic Simulation



Difficulty of Stochastic Simulation

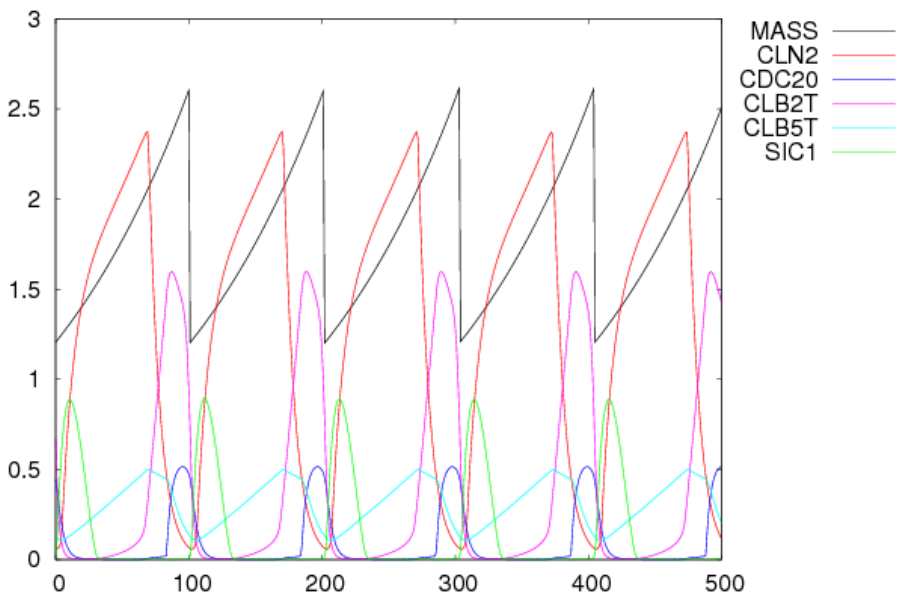


Event handling algorithm

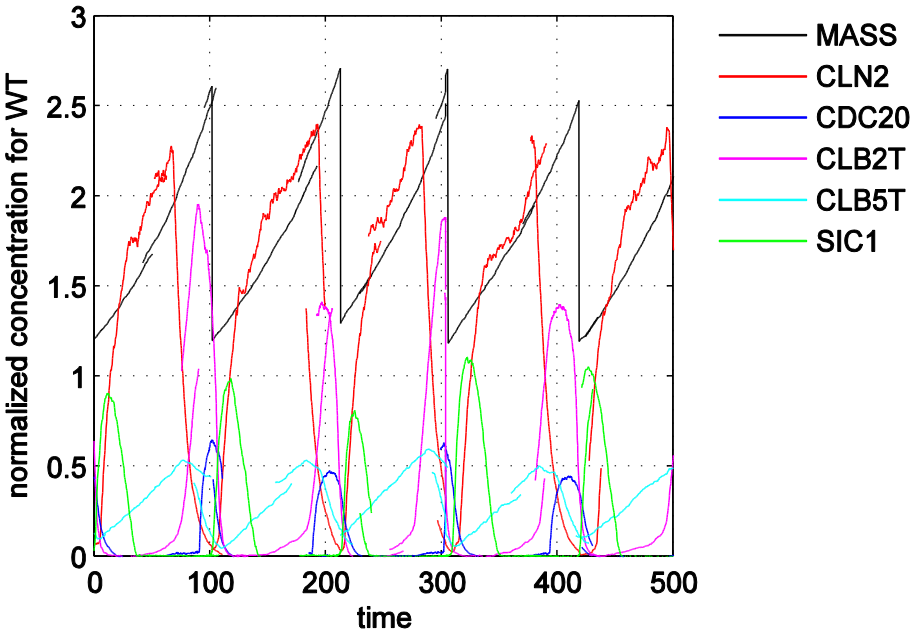
```
if (X < second threshold)
   then (EventFlag ← TRUE)
if (X > threshold AND EventFlag = TRUE)
   then (event is triggered;
         EventFlag ← FALSE)
```

SSA with Wild-Type Budding Yeast Model

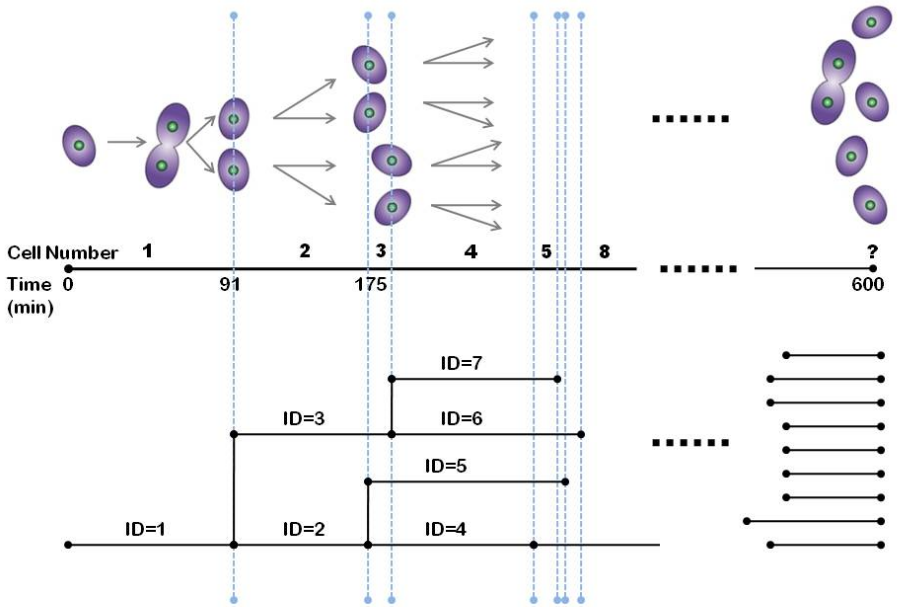
Deterministic simulation trajectories

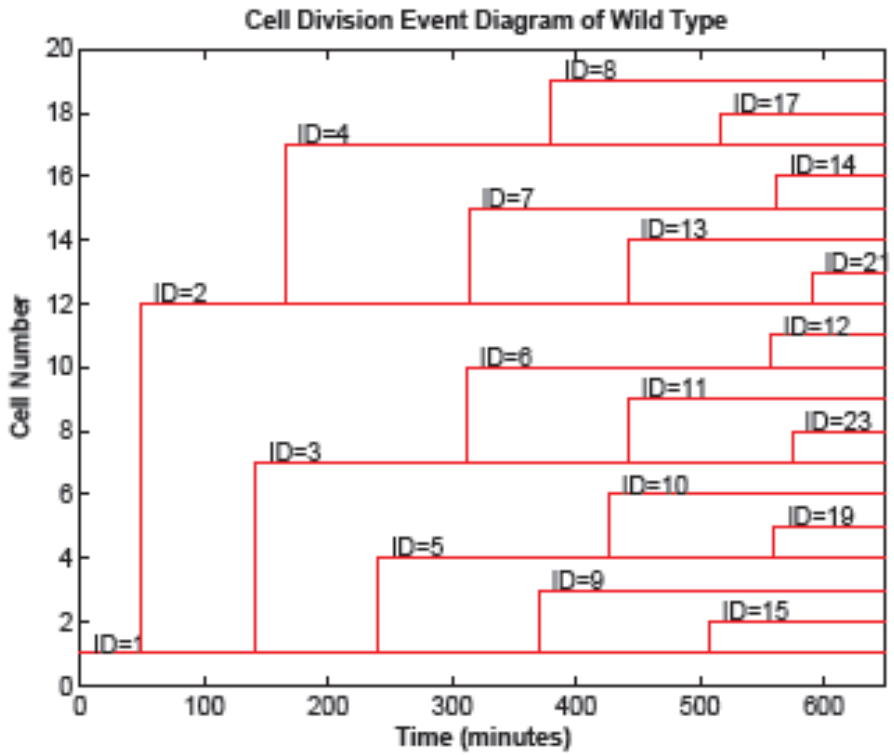


Stochastic simulation trajectories

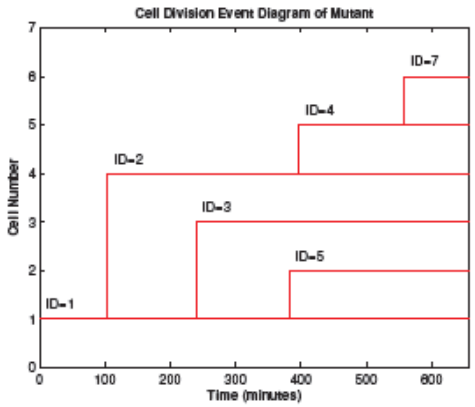


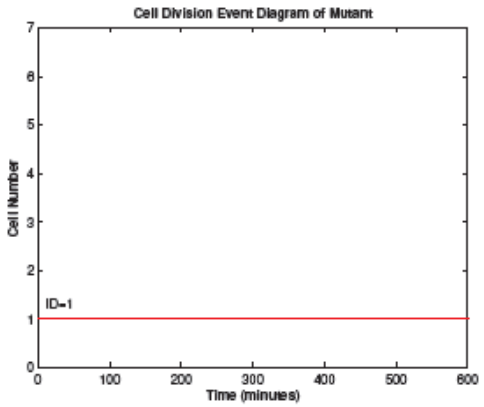
Multistage cell cycle tracking



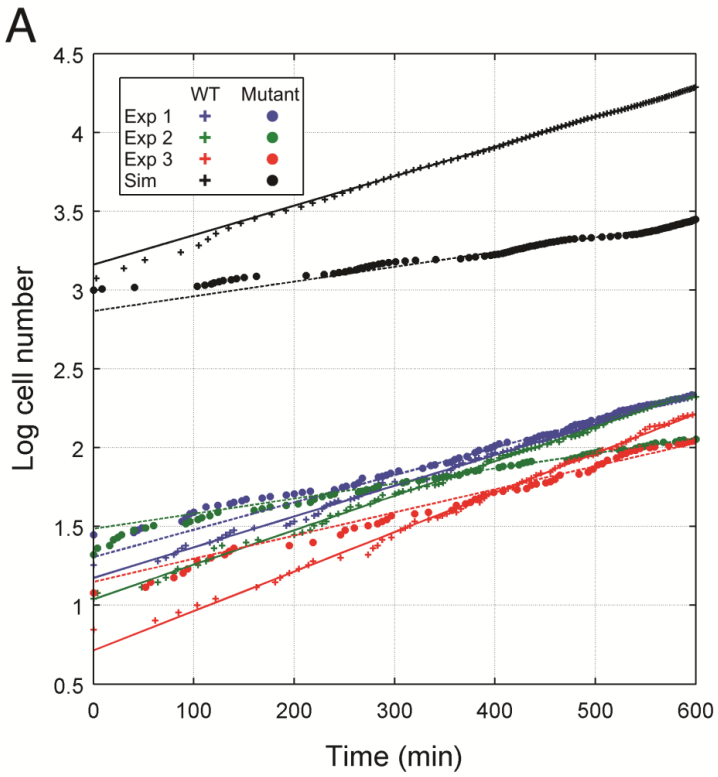


Extraordinary Mutant Simulation





- The *Clb2-db clb5* mutant is inviable in glucose but partially viable on slower growth media such as raffinose.
- This effect can be explained well by stochastic simulation.



A. Comparison of wild-type and mutant cell growth rate in raffinose by wet-lab experiments and stochastic simulations

REPORT

Cell Cycle 10:6, 1-11; March 15, 2011; © 2011 Landes Bioscience

Stochastic exit from mitosis in budding yeast

Model predictions and experimental observations

David A. Ball,[1] Tae-Hyuk Ahn,[2] Pengyuan Wang,[2] Katherine C. Chen,[3] Yang Cao,[2] John J. Tyson,[3] Jean Peccoud[1] and William T. Baumann[4,*]

Stochastic Simulation Service

http://stochss.org/

StochSS
Stochastic Simulation Service

Download About Documentat

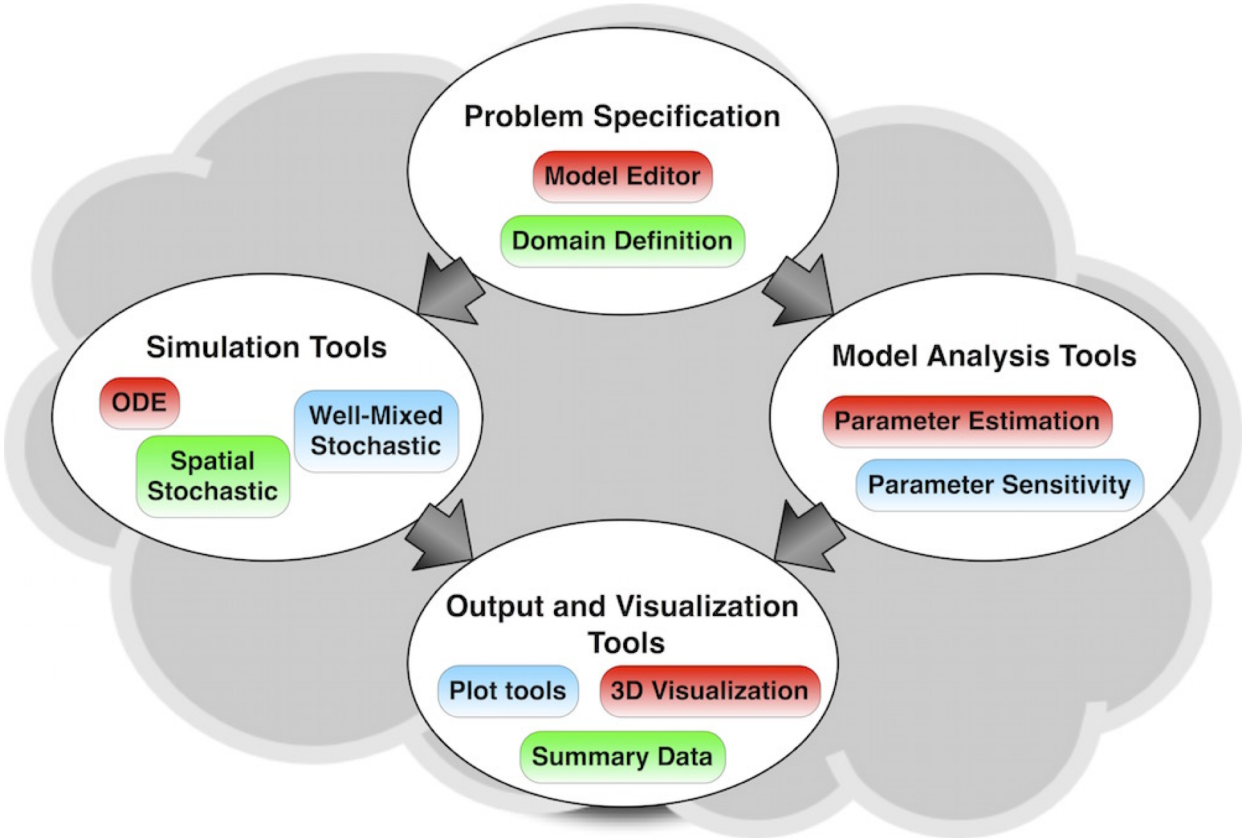
Click Here to Try StochSS Online

We have recently published a paper about StochSS in PLOS Computational Biology. It should be of interest to anyone who wants to know more about how StochSS is designed or how it can be used for modeling and simulation.

Stochastic Simulation Service: Bridging the Gap between the Computational Expert and the Biologist

StochSS is an integrated development environment (IDE) for simulation of biochemical networks:

- Define models in an easy-to-use model editor.
- Interactive 3D visualization of models and results.
- Scale your simulations to the cloud with a click-of-a-button.
- StochSS automatically converts your models from ODE, to well-mixed stochastic, to spatial stochastic.
- Supports parameter estimation and parameter sensitivity.
- Supports parameter sweeps.
- Multi-user environment makes collaborations easier.



COPASI Stochastic Simulation

http://copasi.org/Support/User_Manual/Methods/Time_Course_Calculation/Stochastic_Simulation/

The Direct-Method

This stochastic simulation method implements the Gillespie's direct method as described in [Gillespie76].

Options for the Direct-Method

Max Internal Steps
This parameter is a positive integer value specifying the maximal number of internal steps the integrator is allowed to take before the next desired reporting time. The default value is '1000000'.

Use Random Seed
This flag can be '0' or '1' and determines if the user-defined random seed should be used for the calculation. The default is '0' meaning that the random seed is set to a random value before each run and consecutively calculated trajectories will be different. If the value of this flag is set to '1', the user-defined random seed will be used and each calculated trajectory will be the same for the same value of the given random seed.

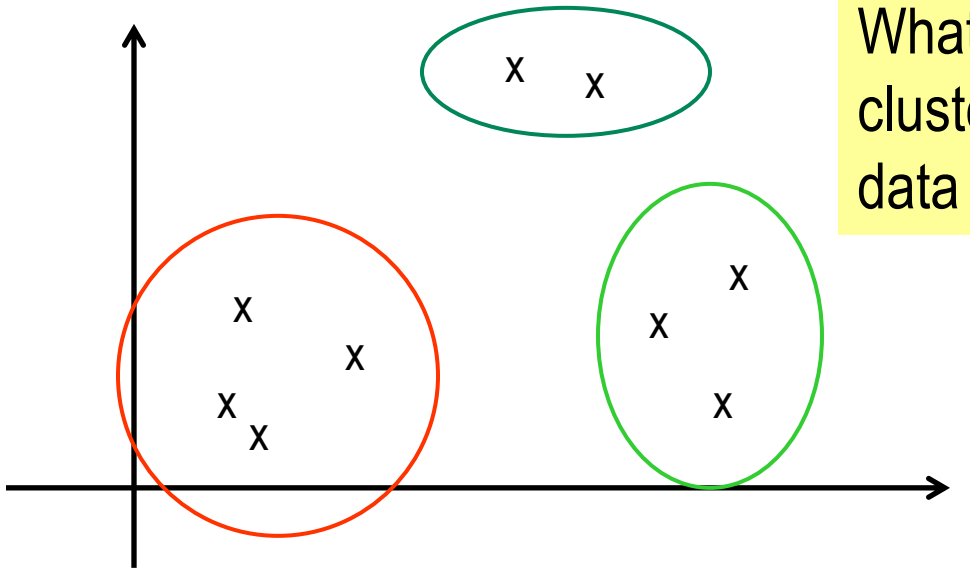
Random Seed
This unsigned integer is used as random seed in the calculations, if the flag Use Random Seed is set to '1'. The default value is '1'.

Important Topics in Modeling and Simulations

Clustering

Clustering - Intro

- The goal of clustering is to
  - group data points that are close (or **similar**) to each other
  - identify such groupings (or clusters) in an **unsupervised** manner
    - Unsupervised: no information is provided to the algorithm on which data points belong to which clusters
- Example



What should the clusters be for these data points?

Clustering Example in Bioinformatics

- Suppose genes A and B are grouped in the same cluster, then we hypothesis that genes A and B are involved in similar function.
  - If we know the role of gene A is apoptosis
  - but we do not know if gene B is involved in apoptosis
  - we can do experiments to confirm if gene B indeed is involved in apoptosis.
- Suppose genes A and B are grouped in the same cluster, then we hypothesize that proteins A and B might interact with each other.
  - So we can do experiments to confirm if such interaction exists.
- So clustering RNA-Seq data in a way helps us make hypotheses about:
  - potential functions of genes
  - potential protein-protein interactions

Clustering in Single-cell RNA-Seq Data

- In recent years, the advances in single-cell RNA-seq techniques have enabled us to perform large-scale transcriptomic profiling at single-cell resolution in a high-throughput manner.
- Unsupervised learning such as data clustering has become the central component to identify and characterize novel cell types and gene expression patterns

https://arxiv.org/pdf/2001.01006.pdf

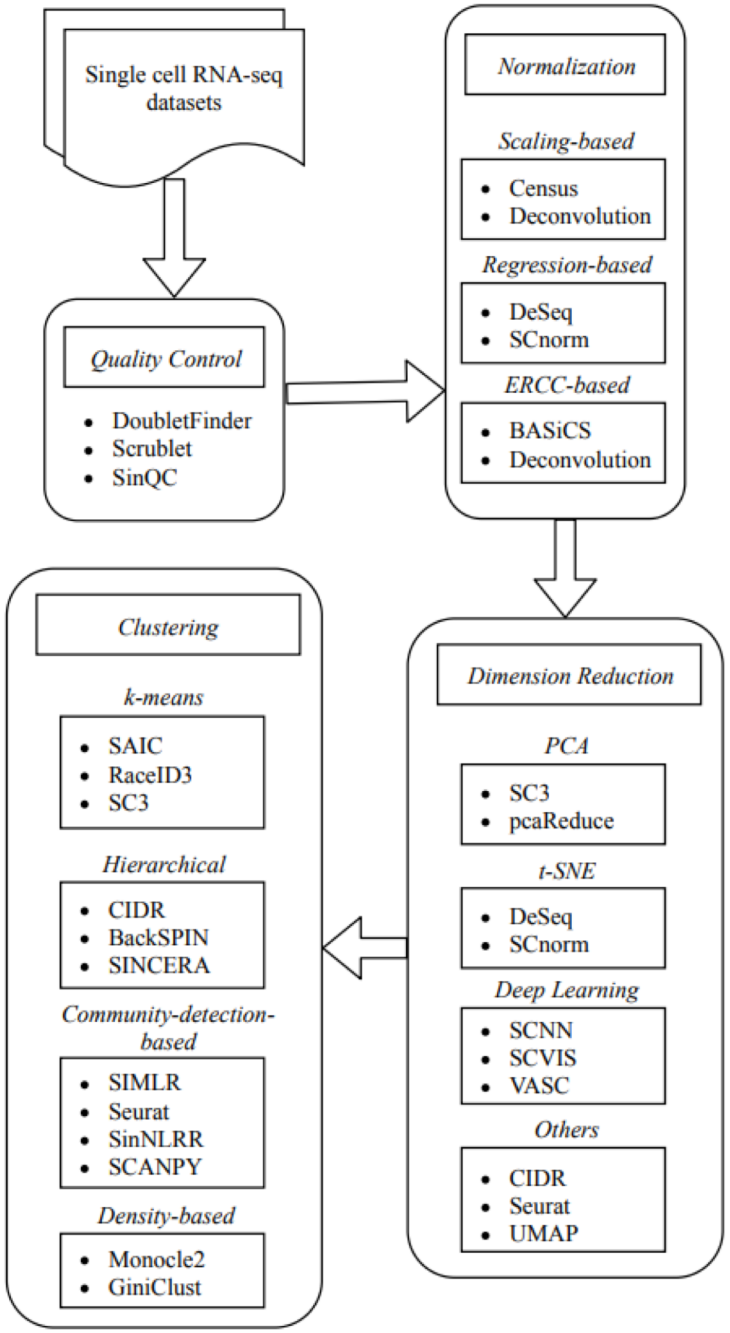


Figure 1: Workflow of single cell RNA-seq data clustering.

Key Terms in Cluster Analysis

- Distance & Similarity measures
- Hierarchical & non-hierarchical
- Single/complete/average(centroid) linkage
- Dendrograms & ordering

Euclidean Distance

- Distance between gene x and y, given n samples (or distance between samples x and y, given n genes)
- Euclidean distance metrics detect similar vectors by identifying those that are closest in space.

$$d(x,y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2}$$

Euclidean Distance via R

- Distance between gene x and y, given n samples (or distance between samples x and y, given n genes)
- Euclidean distance metrics detect similar vectors by identifying those that are closest in space.

$$d(x,y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2}$$

- Using R

```
> x <- rnorm(30)
> y <- rnorm(30)
> dist(rbind(x,y))
```

Pearson's correlation coefficient

- The Pearson's correlation coefficient is a widely used statistical score to measure the co-variance of two variables.
- It is particularly used in the case of hierarchical clustering applied to microarray or RNA-Seq data.
- The Pearson's correlation coefficient can be computed by dividing the covariance of the two variables by the product of their standard deviations.

$$r_{x,y} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{cov(x,y)}{s_x s_y}$$

Lab: Correlation Test using R

- http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

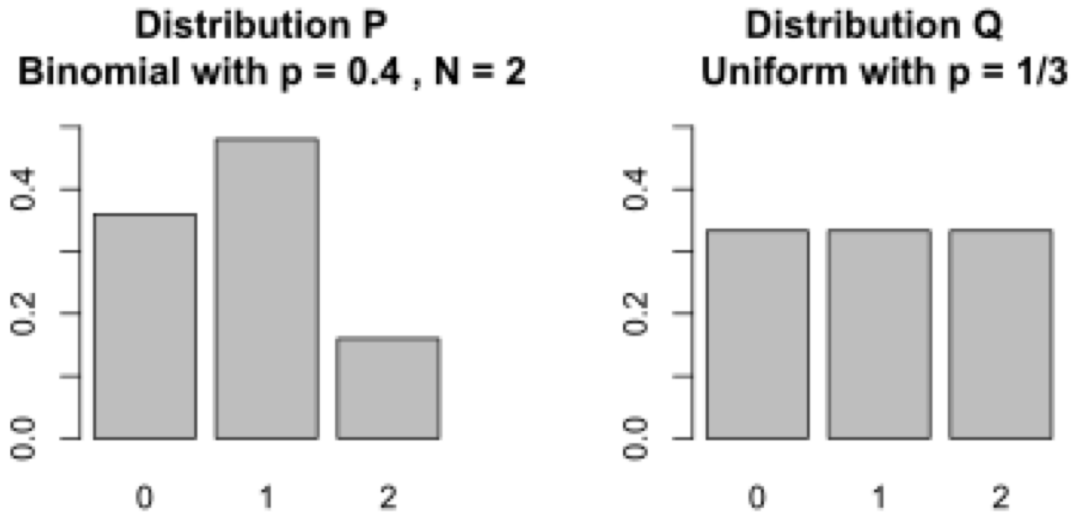
- What is correlation test?
- Install and load required R packages
- Methods for correlation analyses
- Correlation formula
  - Pearson correlation formula
  - Spearman correlation formula
  - Kendall correlation formula
- Compute correlation in R
  - R functions
  - Import your data into R
  - Visualize your data using scatter plots
  - Preleminary test to check the test assumptions
  - Pearson correlation test
    - Interpretation of the result
    - Access to the values returned by cor.test() function
  - Kendall rank correlation test
  - Spearman rank correlation coefficient
- Interpret correlation coefficient
- Online correlation coefficient calculator
- Summary
- Infos

Kullback–Leibler divergence (KL Divergence)

- It is a measure of how one probability distribution is different from the second.
- It is also called relative entropy.
- It is not symmetric in P and Q. In applications, P typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while Q typically represents a theory, model, description, or approximation of P.

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}.$$

KL Divergence Example



x	0	1	2
Distribution *P*(x)	0.36	0.48	0.16
Distribution *Q*(x)	0.333	0.333	0.333

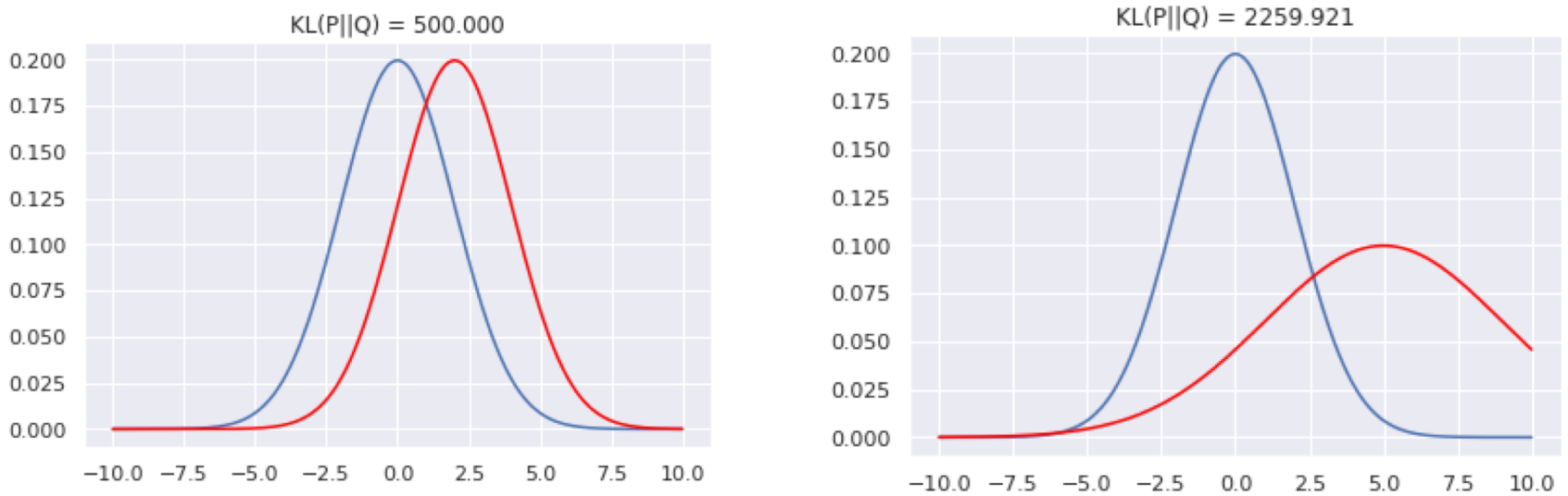
The KL divergences $D_{\mathrm{KL}}(P \parallel Q)$ and $D_{\mathrm{KL}}(Q \parallel P)$ are calculated as follows. This example uses the natural log with base *e*, designated $\ln$ to get results in nats (see units of information).

$$
\begin{aligned}
D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \ln\left(\frac{P(x)}{Q(x)}\right) \\
&= 0.36 \ln\left(\frac{0.36}{0.333}\right) + 0.48 \ln\left(\frac{0.48}{0.333}\right) + 0.16 \ln\left(\frac{0.16}{0.333}\right) \\
&= 0.0852996
\end{aligned}
$$

$$
\begin{aligned}
D_{\mathrm{KL}}(Q \parallel P) &= \sum_{x \in \mathcal{X}} Q(x) \ln\left(\frac{Q(x)}{P(x)}\right) \\
&= 0.333 \ln\left(\frac{0.333}{0.36}\right) + 0.333 \ln\left(\frac{0.333}{0.48}\right) + 0.333 \ln\left(\frac{0.333}{0.16}\right) \\
&= 0.097455
\end{aligned}
$$

KL Divergence Interpretation

- In the context of machine learning, the D(P||Q) is often called the information gain achieved if Q is used instead of P.
- Please read a few blogs or examples for this
  - https://towardsdatascience.com/kl-divergence-python-example-b87069e4b810



The lower the KL divergence, the closer the two distributions are to one another. Therefore, as in the case of t-SNE and Gaussian Mixture Models, we can estimate the Gaussian parameters of one distribution by minimizing its KL divergence with respect to another.

Jensen–Shannon divergence

- The Jensen–Shannon divergence (JSD) is a symmetrized and smoothed version of the Kullback–Leibler divergence

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}.$$

$$D_{\mathrm{JS}} = \tfrac{1}{2}D_{\mathrm{KL}}\left(P\|M\right) + \tfrac{1}{2}D_{\mathrm{KL}}\left(Q\|M\right)$$

where M is the average of the two distributions,

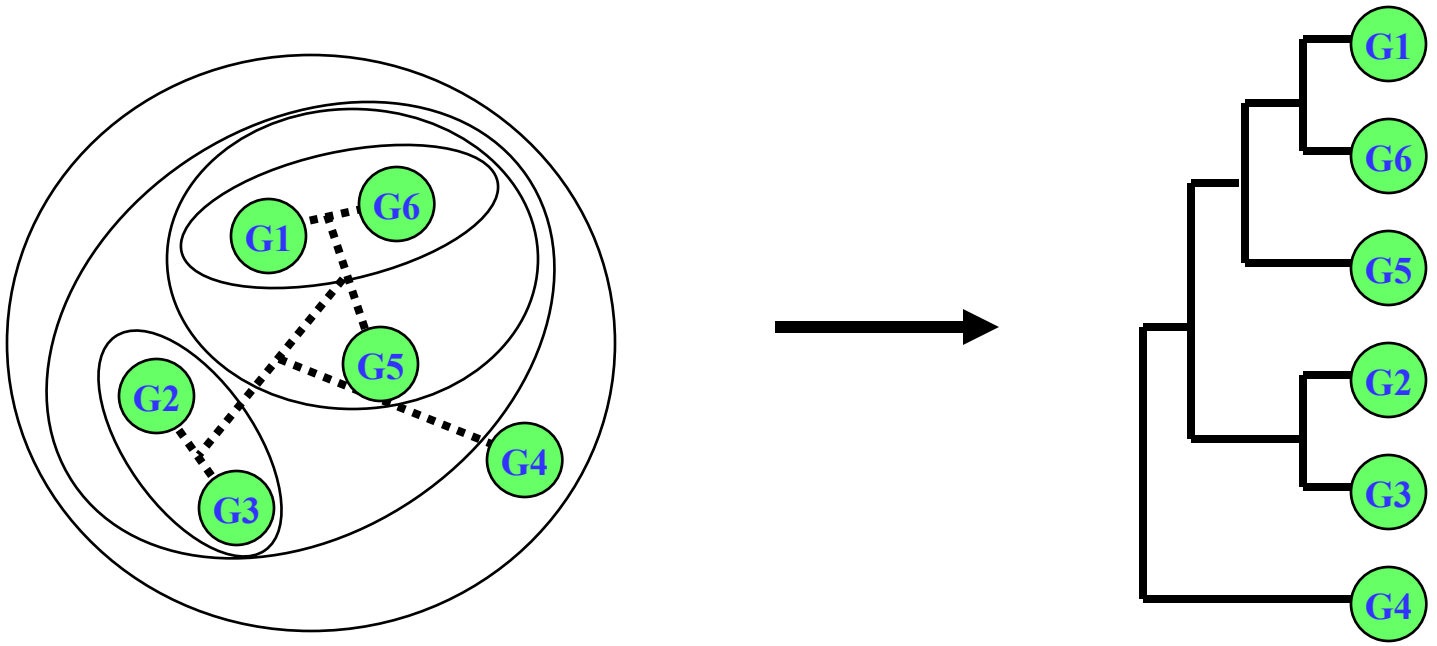
$$M = \tfrac{1}{2}(P+Q).$$

Agglomerative Hierarchical Clustering

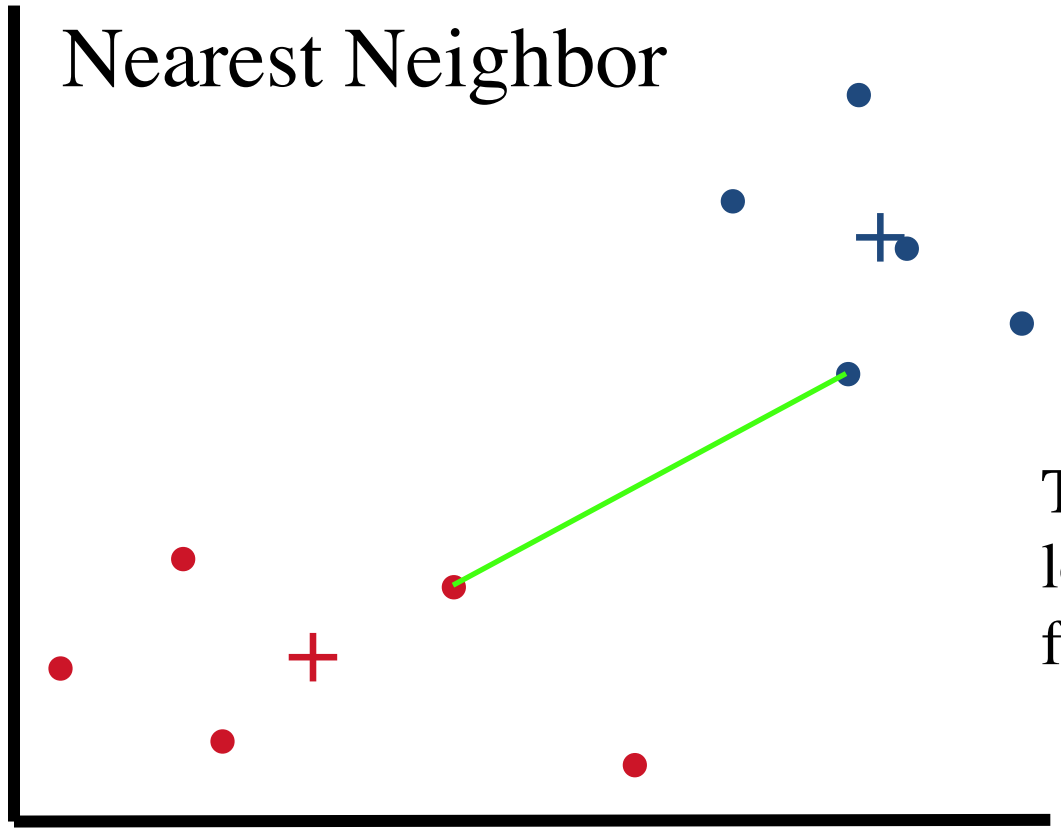
1. Compare all expression patterns to each other.
2. Join patterns that are the most similar out of all patterns.
3. Compare all joined and unjoined patterns.
4. Go to step 2, and repeat until all patterns are joined.

Need a rule to decide how to compare clusters to each other

Visualization of Hierarchical Clustering

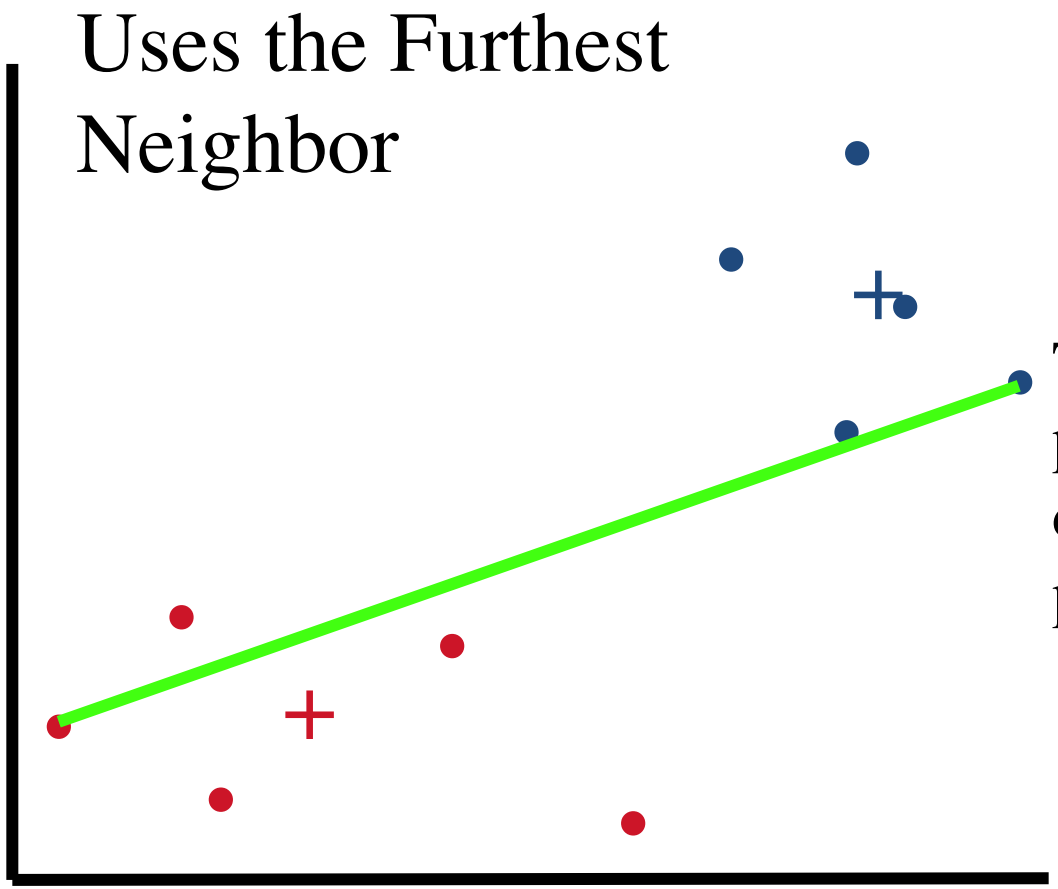


Single linkage Clustering



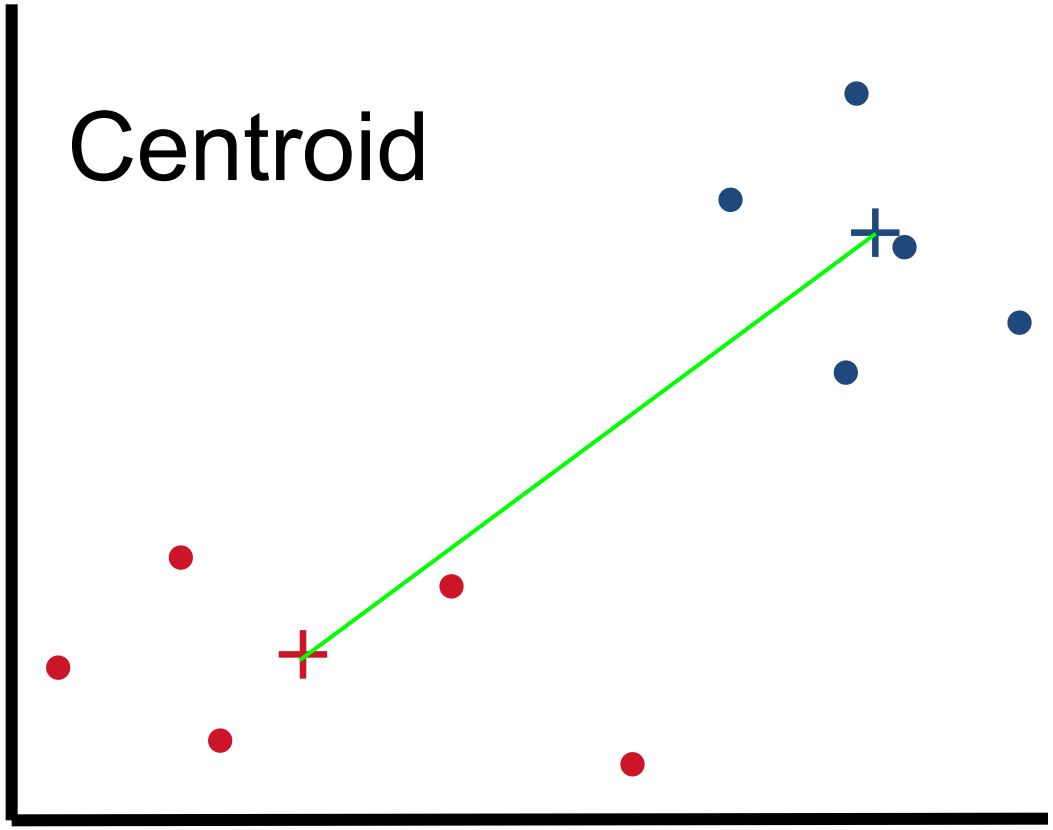
This method produces long chains which form straggly clusters.

Complete Linkage Clustering



This method tends to produce very tight clusters of similar patterns

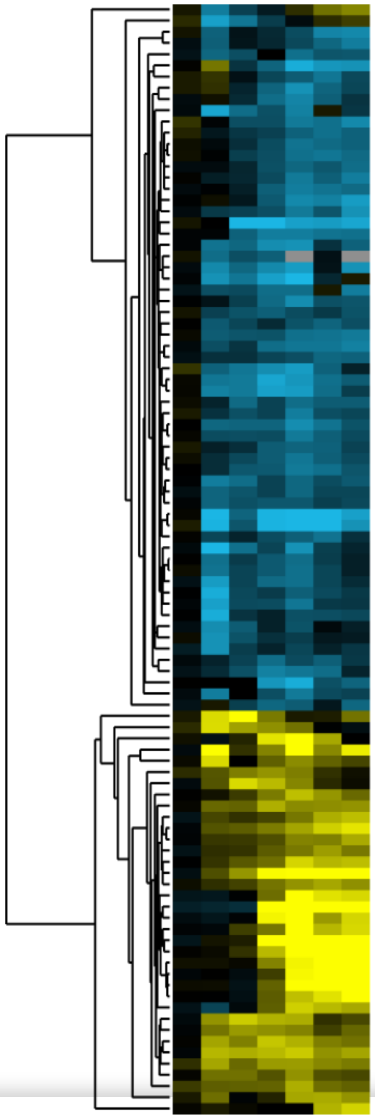
Average Centroid Linkage Clustering



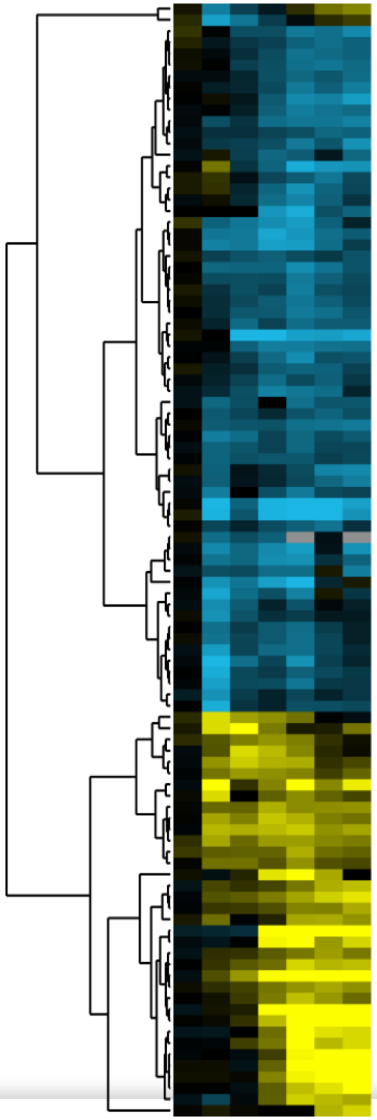
The red and blue '+' signs mark the centroids of the two clusters.

And we get a cluster

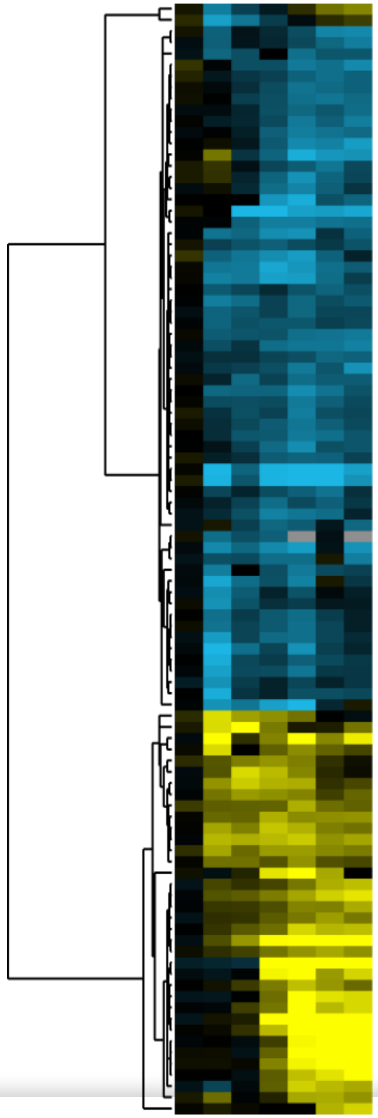
Single



Complete



Average(Centroid)



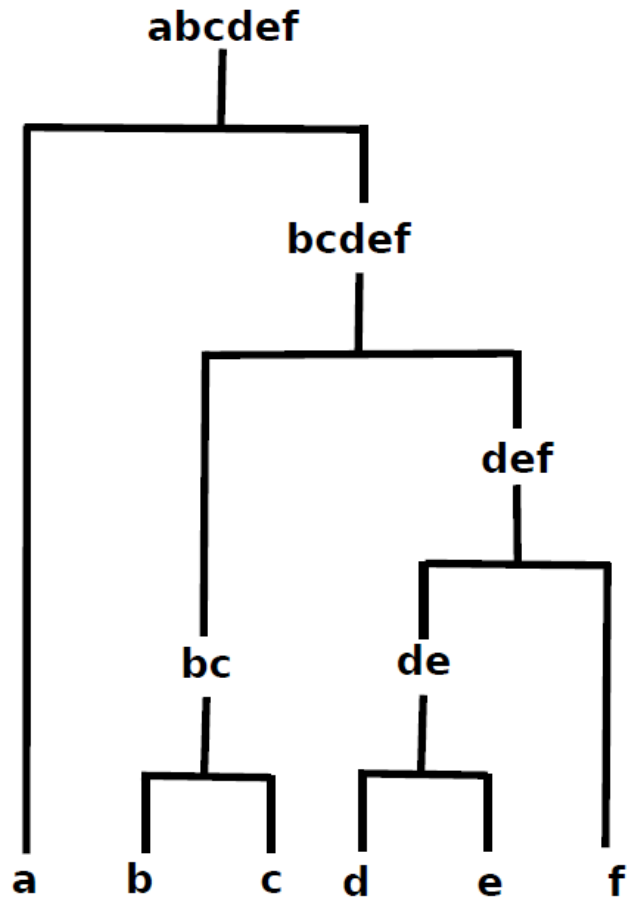
Hierarchical clustering

Construction

- **Agglomerative** approach (bottom-up)
  - Start with every element in its own cluster, then iteratively join nearby clusters
- **Divisive** approach (top-down)
  - Start with a single cluster containing all elements, then recursively divide it into smaller clusters

Hierarchical clustering

Dendograms



In biology

- Phylogenic trees
- Sequences analysis
  infer the evolutionary history of sequences being compared

Hierarchical clustering

Advantages

- Simple
- Easy to implement
- Easy to visualize (Good interpretability)
- Does not require to set the number of clusters

Disadvantages

- Can lead to artifacts
- Computationally intensive $O(n^2 \log n^2)$
- Hard to decide at which level of the hierarchy to stop
- Lack of robustness

Hierarchical Clustering in R

```
> x1 <- c(rnorm(20,sd=.05),
          rnorm(20,mean=1,sd=.05),
          rnorm(20,mean=1.5,sd=.05))

> x2 <- 4+c(rnorm(20,sd=2),
            rnorm(20,mean=8,sd=1.5),
            rnorm(20,mean=8,sd=1))

> mydata.df <- as.data.frame(cbind(x1,x2))

> clusters <- hclust(dist(mydata.df))

> plot(clusters)
```

Partitioning Methods

- Split data up into smaller, more homogenous sets
- Should avoid artifacts associated with incorrectly joining dissimilar vectors
- Can cluster each partition independently of others, by genes and arrays
- Self-Organizing Maps and K-means clustering are two possible partitioning methods

K-means Clustering

- Input: n objects (or points) and a number k
- Algorithm
  - Randomly place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
  - Assign each object to the group that has the closest centroid.
  - When all objects have been assigned, recalculate the positions of the K centroids.
  - Repeat Steps 2 and 3 until the stopping criteria is met.

K-means Clustering

- Stopping criteria:
  - No change in the members of all clusters
  - when the squared error is less than some small threshold value α
    - Squared error (*se*)

$$se = \sum_{i=1}^{k} \sum_{p \in cluster} \|p - m_i\|^2$$

      - where m_i is the mean of all instances in cluster c_i
    - $se^{(j)} < \alpha$
- Properties of k-means
  - Guaranteed to converge
  - Guaranteed to achieve local optimal, not necessarily global optimal.

K-means Clustering

Advantages

- Low complexity
  - complexity is O(nkt), where t = #iterations

Disadvantages

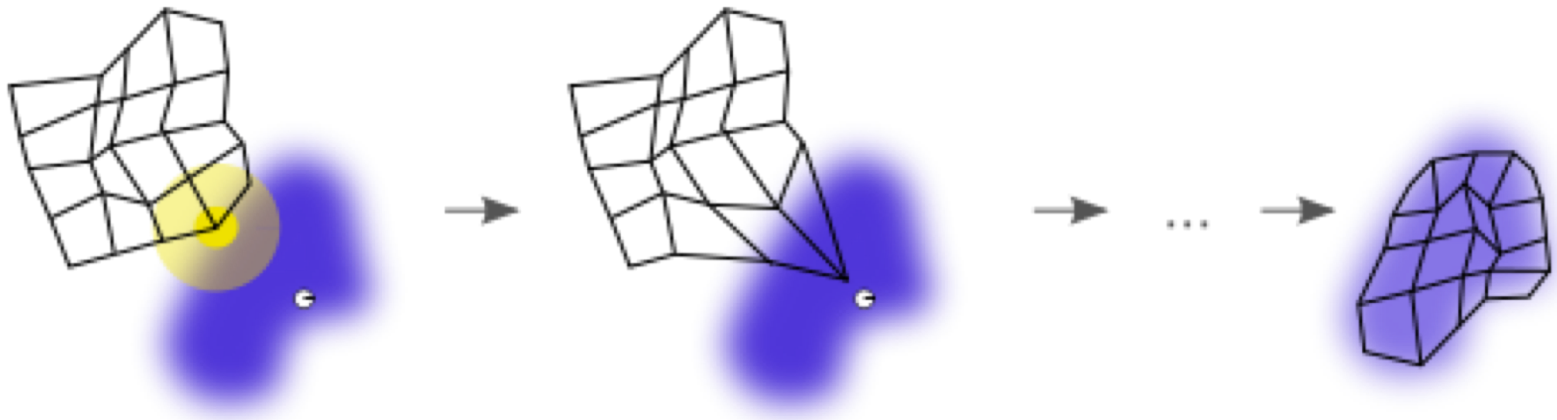
- Necessity of specifying k
- Sensitive to noise and outlier data points
  - Outliers: a small number of such data can substantially influence the mean value)
- Clusters are sensitive to initial assignment of centroids
  - K-means is not a deterministic algorithm
  - Results depend on the initial, random partition

K-means Clustering in R

```
> x1 <- c(rnorm(20,sd=.05),
          rnorm(20,mean=1,sd=.05),
          rnorm(20,mean=1.5,sd=.05))

> x2 <- 4+c(rnorm(20,sd=2),
            rnorm(20,mean=8,sd=1.5),
            rnorm(20,mean=8,sd=1))

> mydata.df <- as.data.frame(cbind(x1,x2))

> clusters <- kmeans(mydata.df, 3, nstart = 20)

> plot(mydata.df, col = clusters$cluster)
```

Self-Organizing Maps (SOM)

- Create a ‘Map’ of ‘n’ partitions, that is modeled on the expression data, where each partition in the map has an associated vector.
- Genes’ expression vectors are assigned to the partition with the most similar associated vector.
- Neighboring partitions are more similar to each other than they are to distant partitions.

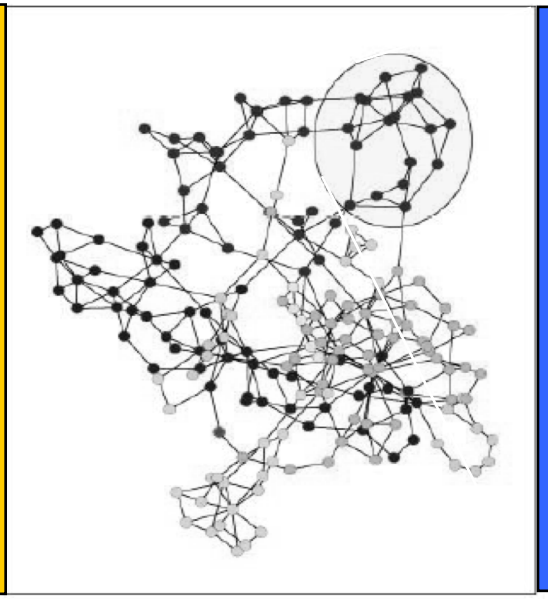


By Clustering

- Identify elements (genes, molecules, cells, …)
- Ascertain their relationships (co-expressed, interacting, …)
- Integrate information to obtain view of system as a whole

Large (genomic) systems

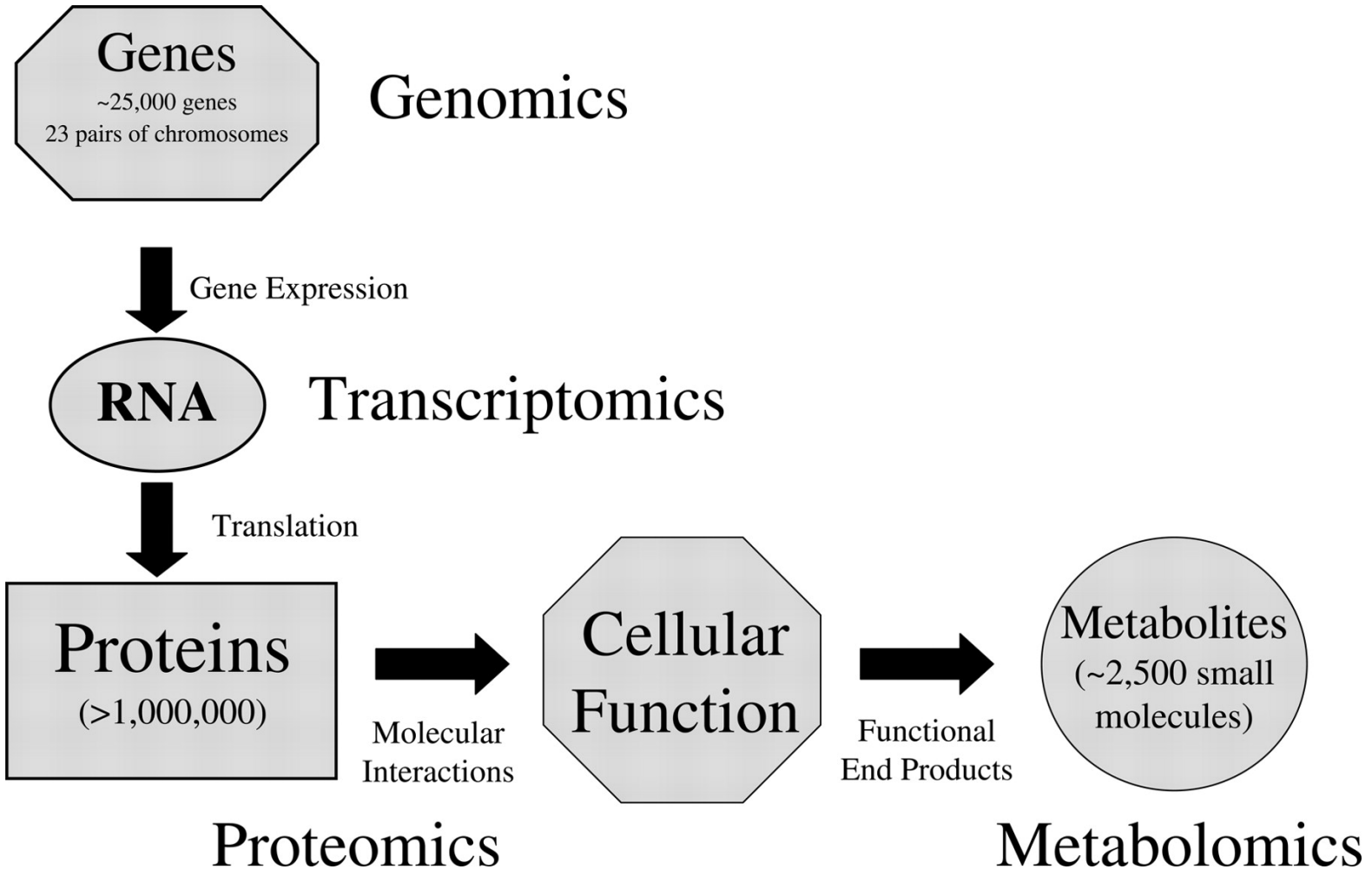
- many uncharacterized elements
- relationships unknown
- *computational analysis* should:
  - improve annotation
  - reveal relations
  - reduce complexity



Small systems

- elements well-known
- many relationships established
- *quantitative modeling* of systems properties like:
  - Dynamics
  - Robustness
  - Logics

Other Important Topics: Proteomics, Metabolomics, ….

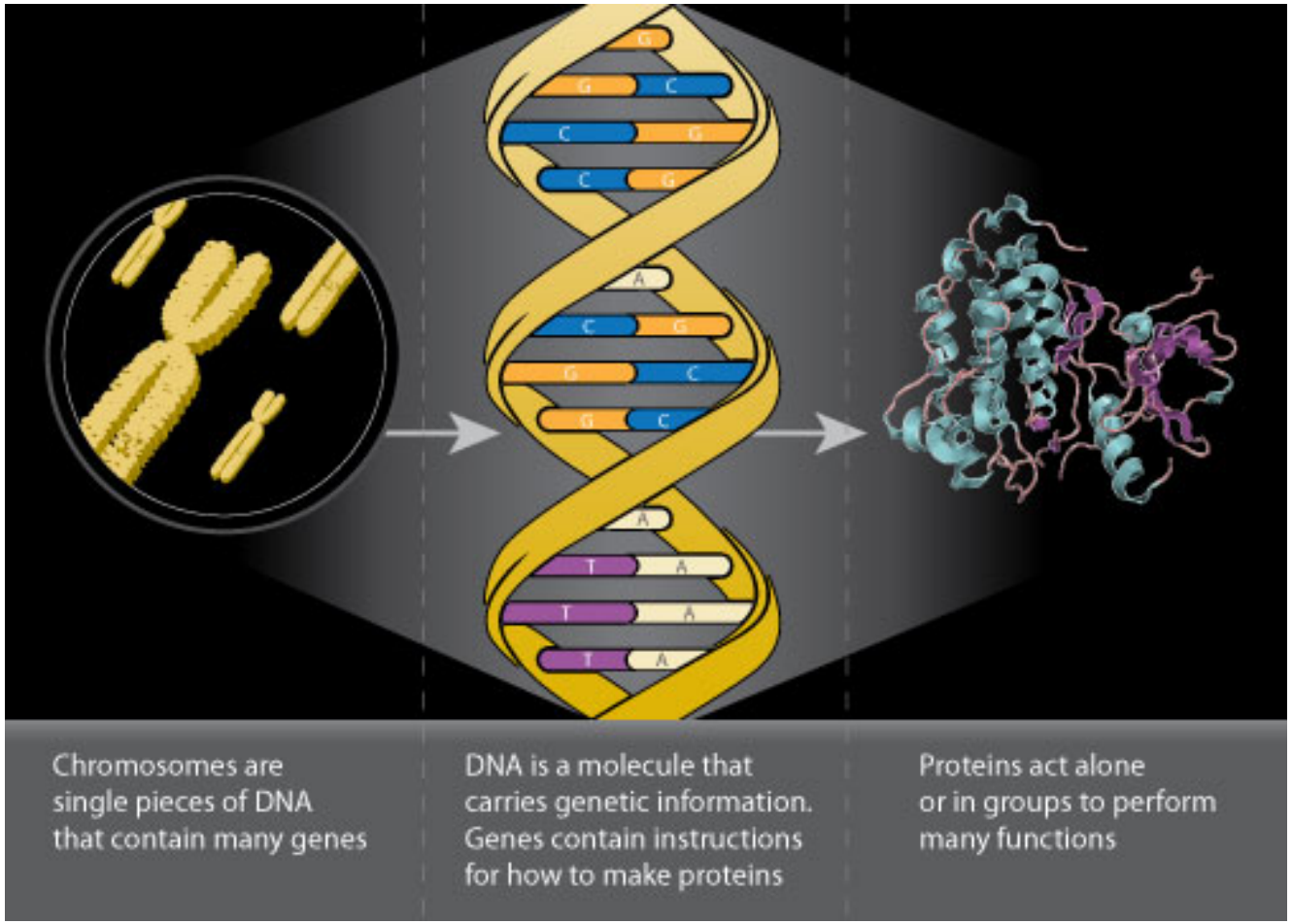


http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2666223/

Proteomics: A Definition

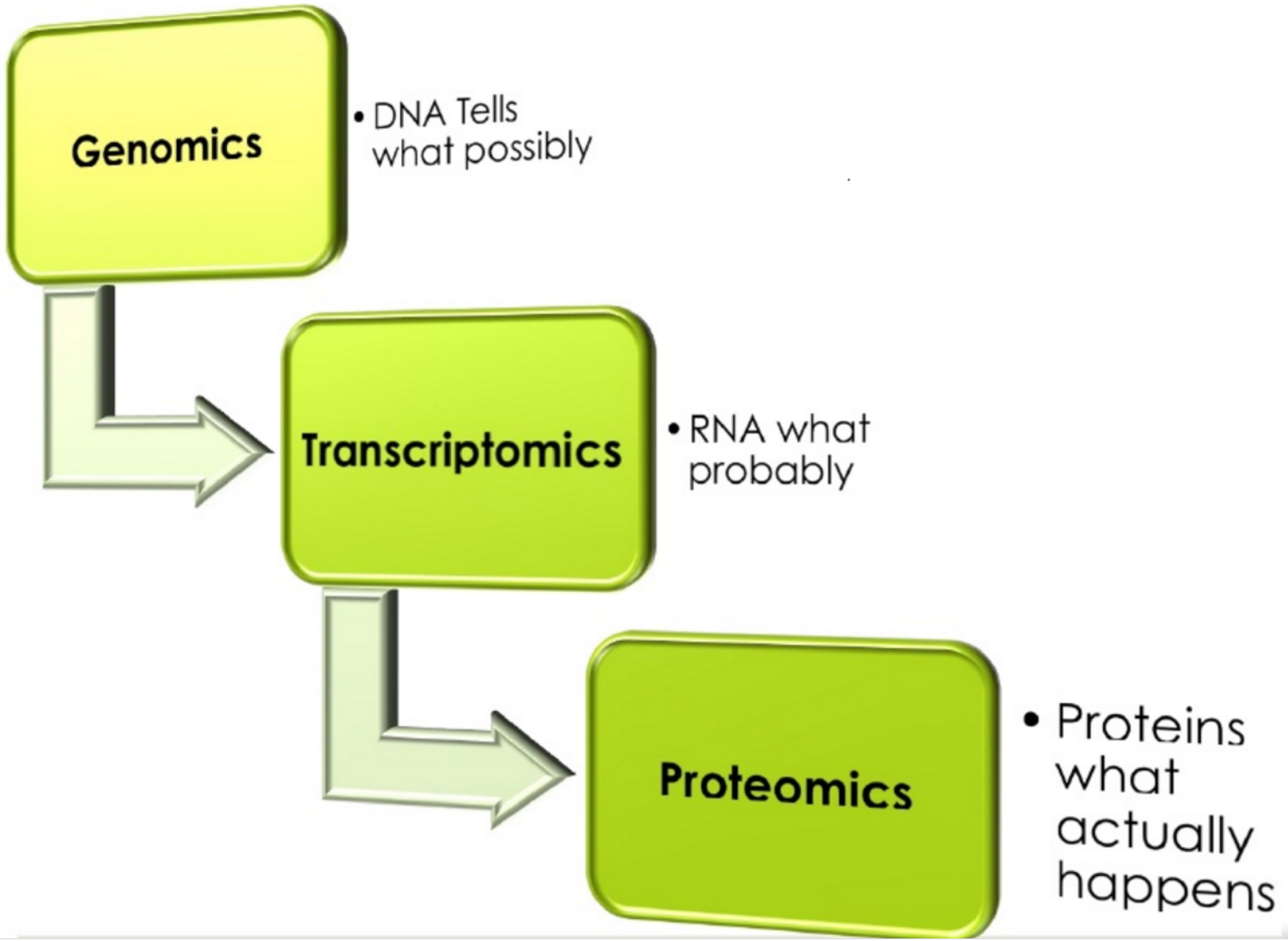
- The term "**proteome**" and "**proteomics**" were coined by Mark Wilkins (Prof. UNSW Australia) in 1994.
- **Proteome**: The entire complement of proteins, including the modifications made to a particular set of proteins, produced by an organism or a cellular system.
- **Proteomics**: A study of entire protein systems (proteomes): what are the component proteins, how they interact with each other, what kinds of metabolic networks or signaling networks they form

Why Proteomics?

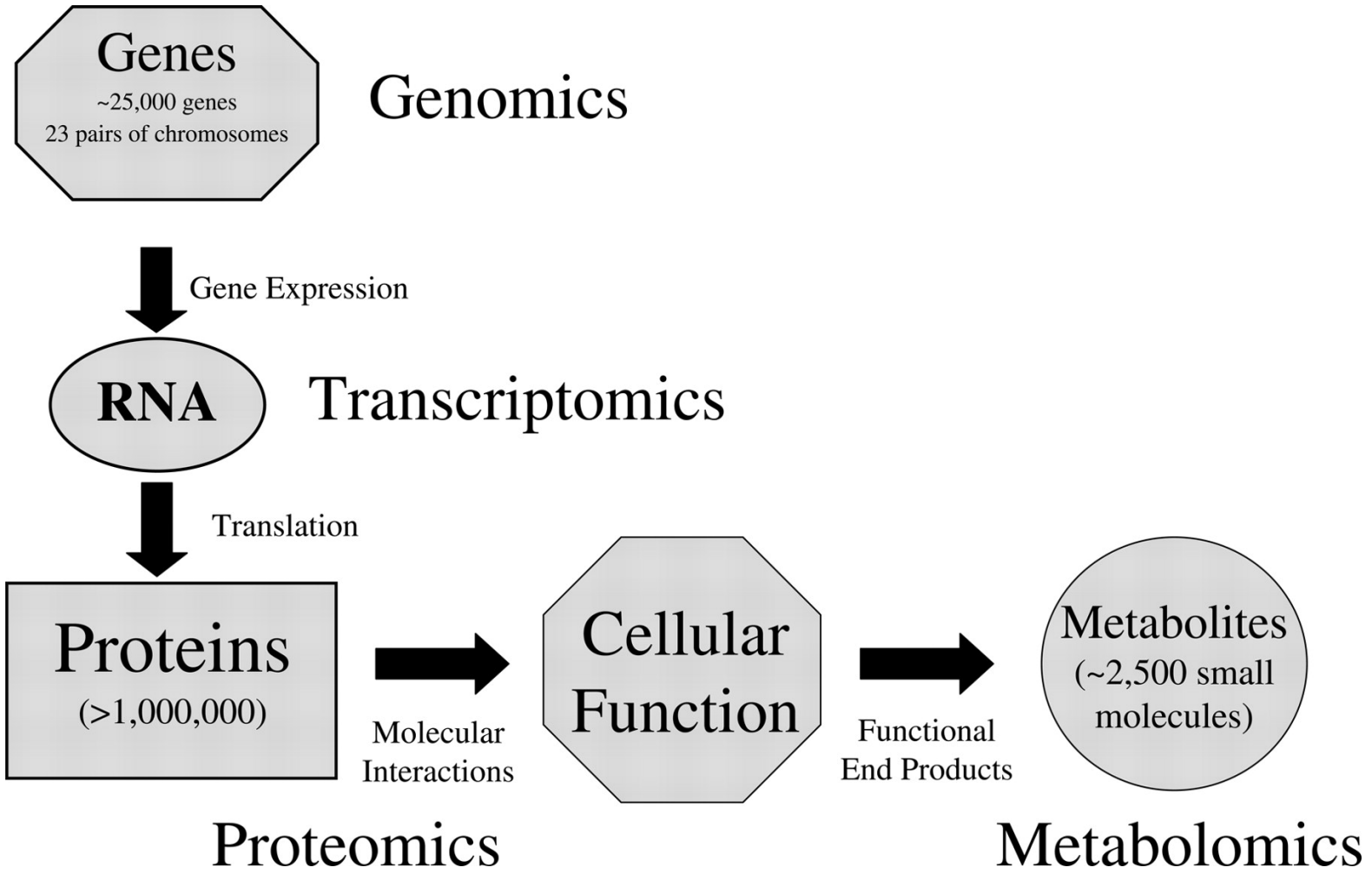


http://proteomics.cancer.gov/whatisproteomics

Why Proteomics?



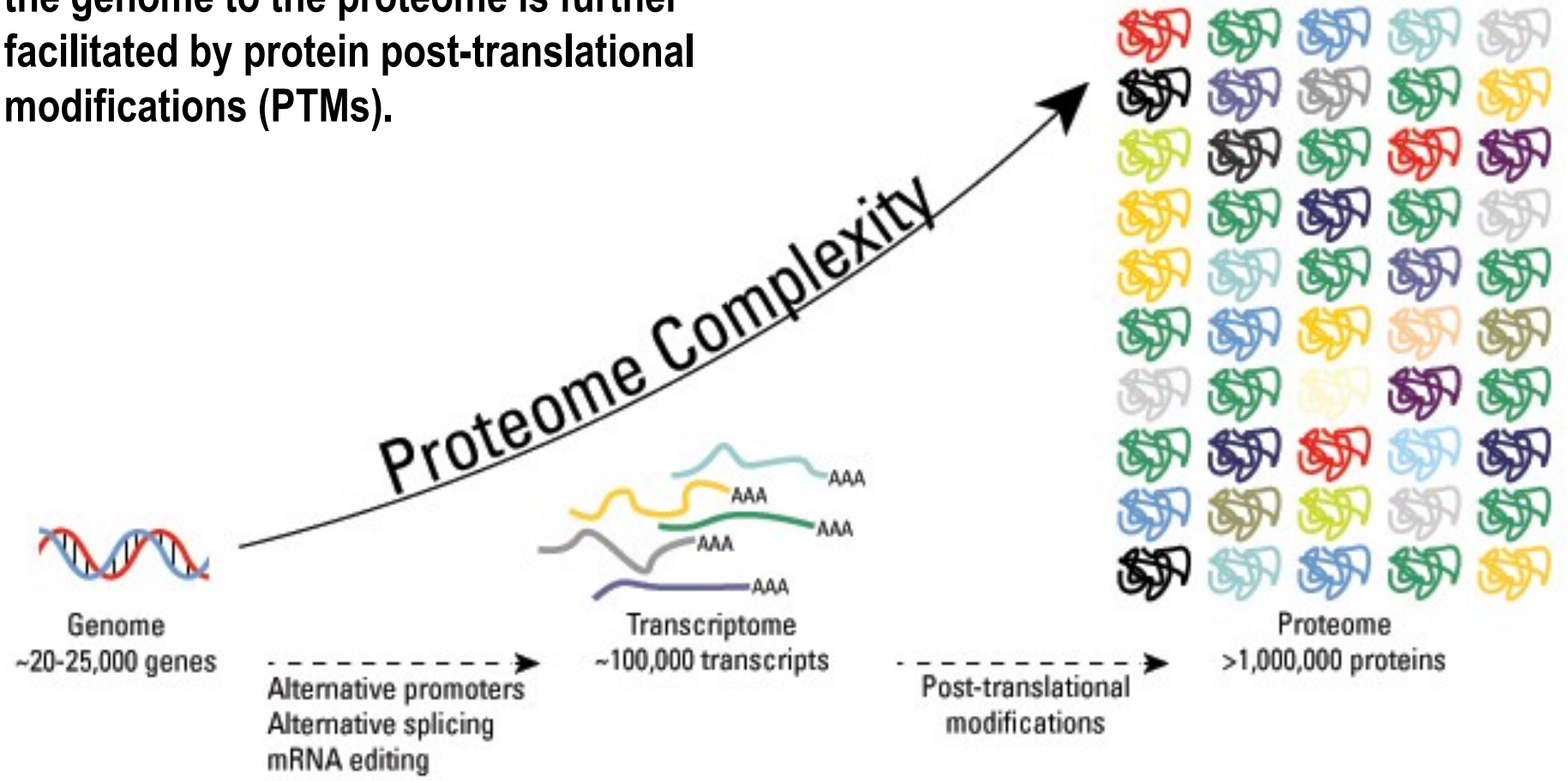
The Omics Cascade



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2666223/

Proteome Complexity

The increase in complexity from the level of the genome to the proteome is further facilitated by protein post-translational modifications (PTMs).



https://www.thermofisher.com/

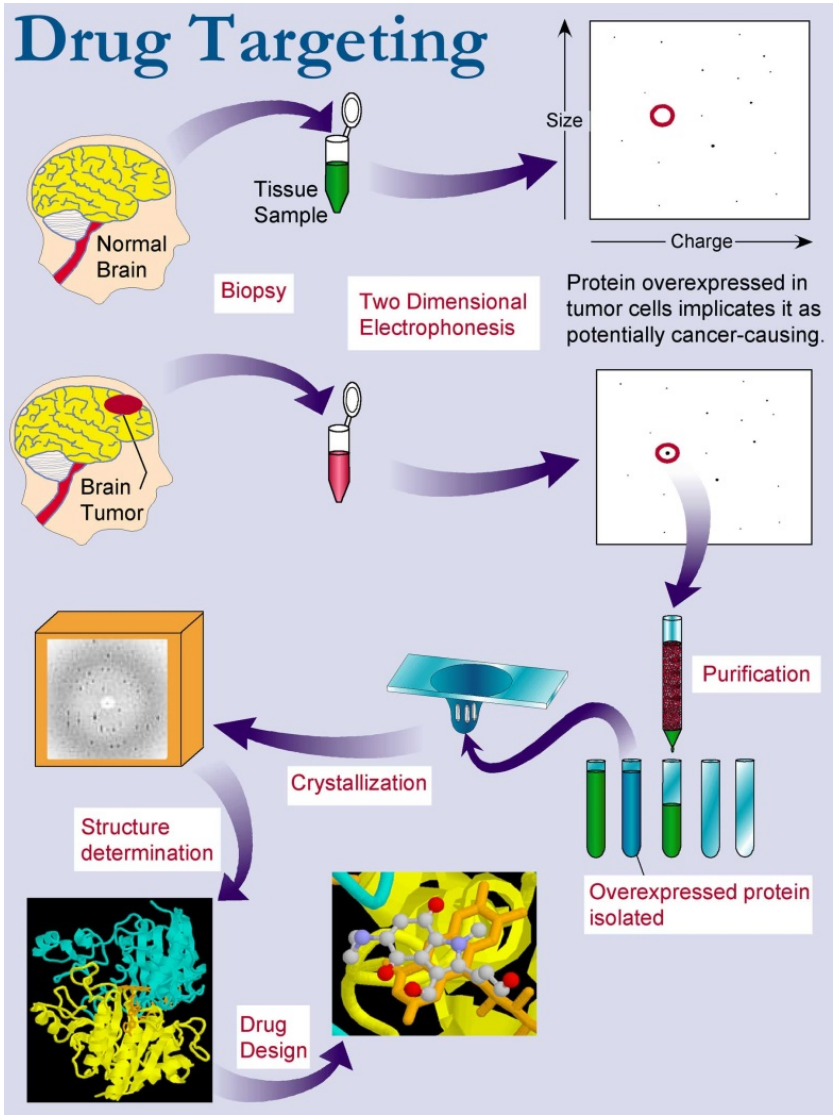
Proteome Complexity

- One gene can encode more than one protein (even up to 1,000).
- Proteins are dynamic.
  - Proteins are continually undergoing changes, e.g., binding to the cell membrane, partnering with other proteins to form complexes, or undergoing synthesis and degradation. The genome, on the other hand, is relatively static.
- Proteins are co- and post-translationally modified.
  - As a result, the types of proteins measured can vary considerably from one person to another under different environmental conditions, or even within the same person at different ages or states of health.
- Proteins exist in a wide range of concentrations in the body.

https://www.thermofisher.com/

Rational Drug Design

- Compare proteome of healthy and cancerous tissue.
- Identify protein linked to onset of cancer.
- Determine 3-D shape.
- Design drug to alter protein function.
- Specifically targeted molecules may have fewer side effects.



Types of Proteomics Research

- Structural proteomics
  - 3D configuration of the protein. Understanding the protein's structure aids in identification of the protein's interactions and function.
  - X-ray crystallography and NMR spectroscopy.
- Expression proteomics
  - Identify proteins in a particular sample, and those proteins differentially expressed in related samples (e.g., diseased vs. healthy tissue)
  - Mass spectrometry and protein microarray
- Interaction proteomics
  - Determine the protein functions and it also explains the way proteins assemble in bigger complexes.
  - Mass spectrometry and yeast two-hybrid system.

Protein Bioinformatics: Protein sequence analysis

- Helps characterize protein sequences *in silico* and allows prediction of protein structure and function
- Statistically significant BLAST hits usually signifies sequence homology
- Homologous sequences may or may not have the same function but would always (very few exceptions) have the same structural fold
- Protein sequence analysis allows protein classification

Bioinformatics for proteomics

- Protein identification
- Protein structure
- Post-translational modifications
- Protein biomarkers

Check available database and tools for proteomics at http://www.expasy.org/proteomics

Protein Research Paradigm Shifts in Bioinformatics

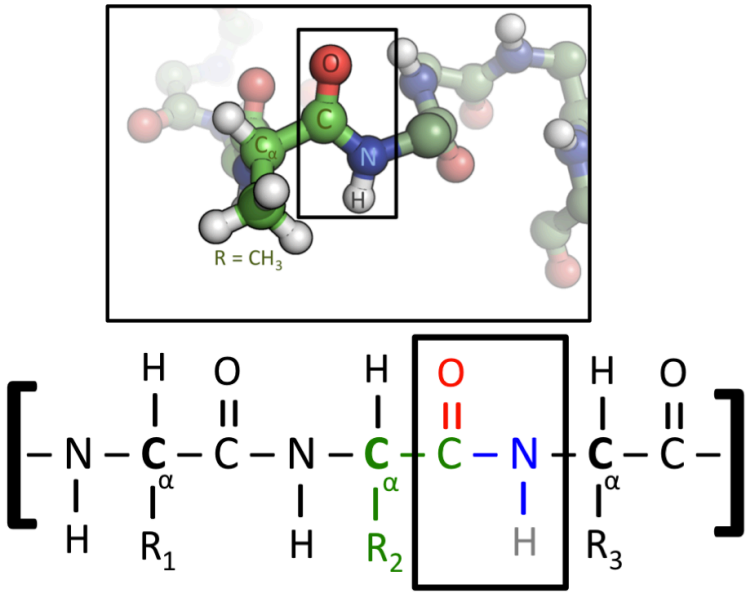
- Sequencing (1980's to early 1990's)
  - DNA/RNA/Protein sequence analysis and storage
- 3-D Protein Structure Prediction (Mid 1980's to late 1990's)
  - Databases of protein structures
- DNA/RNA Microarray Expression (Mid 1990's to 2000's)
  - Databases of expression data
- Protein Interaction Experiments (Early 2000's to Present)
  - Databases with pairwise interactions
- Mass Spec Proteomics Profiling (Early 2000's to Present)
  - Databases with mass spec, protein identifications, proteomic patterns
- Integration of Multiple Omics Data (ongoing and future)

Development of Protein Databases

- Atlas of protein sequence and structure – Dayhoff (1966) first sequence database (pre-bioinformatics). Currently known as Protein Information Resource (PIR)
- Protein data bank (PDB) (http://www.rcsb.org/pdb/home/home.do) – structural database (1972) remains most widely used database of structures
- UniProt (http://www.uniprot.org/) – The United Protein Databases (2003) is a central database of protein sequence and function created by joining the forces of the SWISS-PROT, TrEMBL and PIR protein database activities
  - SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.
  - TrEMBL is a computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.
- Pfam (http://pfam.xfam.org/) - The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).

Proteins

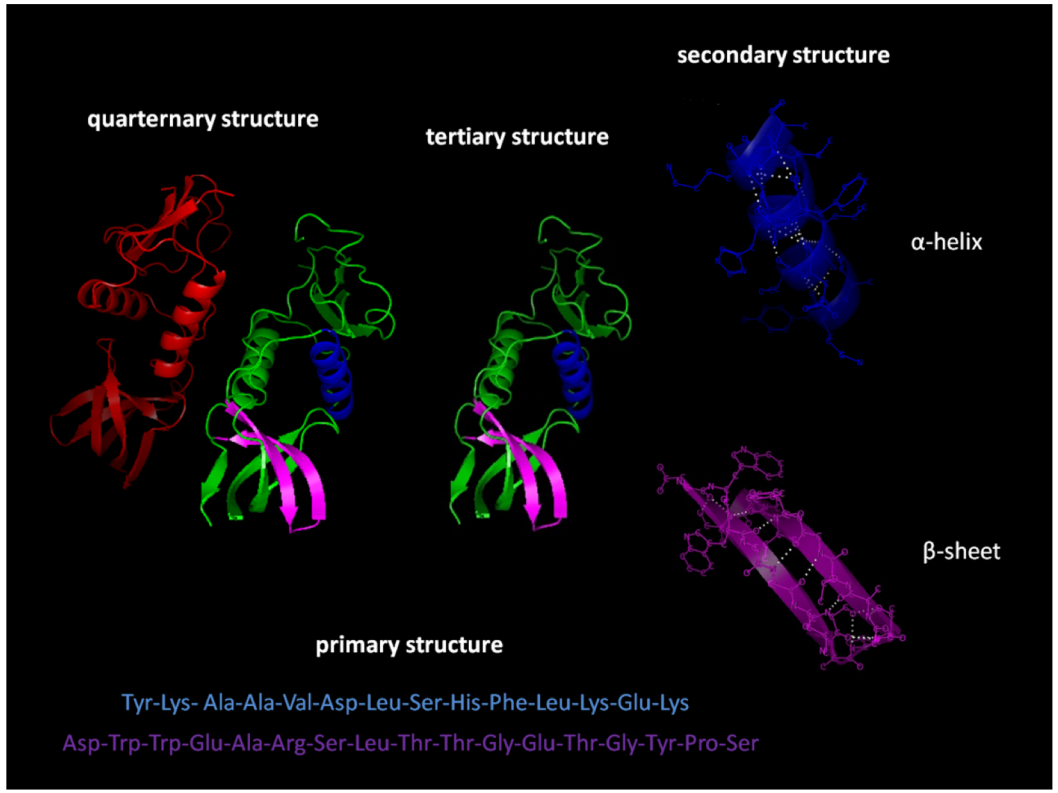
- Biochemical compounds
- A linear chain of amino acid residues is called a polypeptide
- Bonded together by peptide bonds

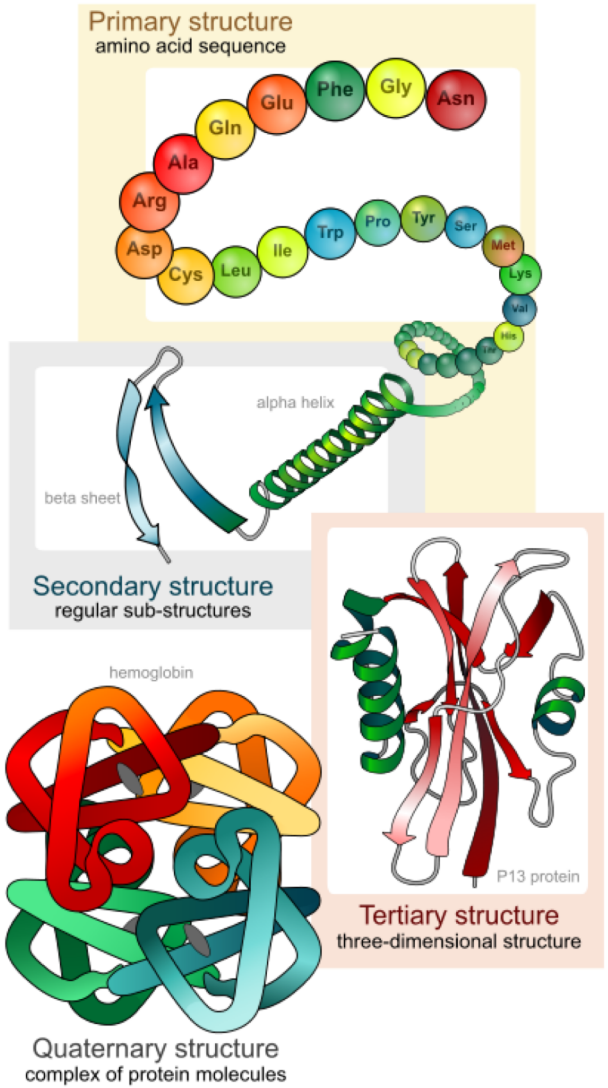


Chemical structure of the peptide bond (bottom) and the three-dimensional structure of a peptide bond between an alanine and an adjacent amino acid (top/inset)

https://en.wikipedia.org/wiki/Protein

Protein Structure





https://en.wikipedia.org/wiki/Protein

Protein Structure Prediction

- Accurate determination of the three-dimensional shape of a protein from its amino acid sequence.
- Understand (structure dependent) function of proteins.
- Protein structure needed for drug design.

I-TASSER

- Widely used protein structure & function predictions web server
- http://zhanglab.ccmb.med.umich.edu/I-TASSER/
- Test it with a sample input provided in the webpage.

Why do we analyze proteins?

- Proteins play crucial roles in nearly all biological processes. These many functions of proteins are a result of the folding of proteins into many distinct 3D structures.
- Protein analysis tries to explore how amino acid sequences specify the structure of proteins and how these proteins bind to substrates and other molecules to perform their functions.
- Protein analysis allows us to understand the function of the protein based on its structure.

Protein Profiling

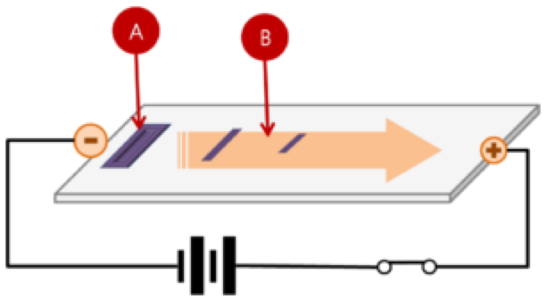
- Determination of the proteins that make up a given proteome.
- Challenges of Protein Profiling
  - Proteomes vary by cell type.
  - Proteomes vary by stage of cell development.
  - Some proteins abundant, others very rare.
  - Some biologically important proteins are tiny and difficult to detect.

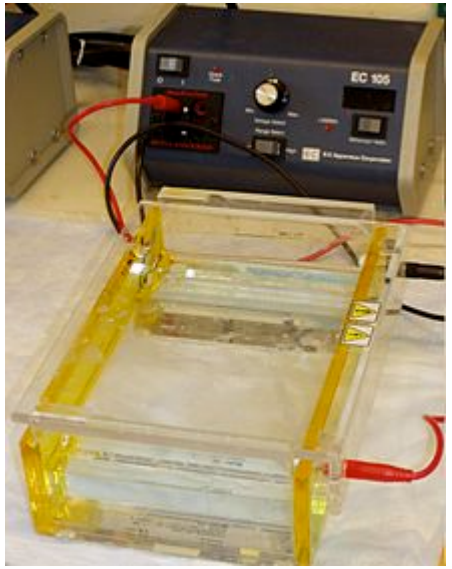
Protein Profiling Techniques

- Two-dimensional gel electrophoresis.
- Chemical protein sequencing.
- Protein sequencing by mass spectrometry.

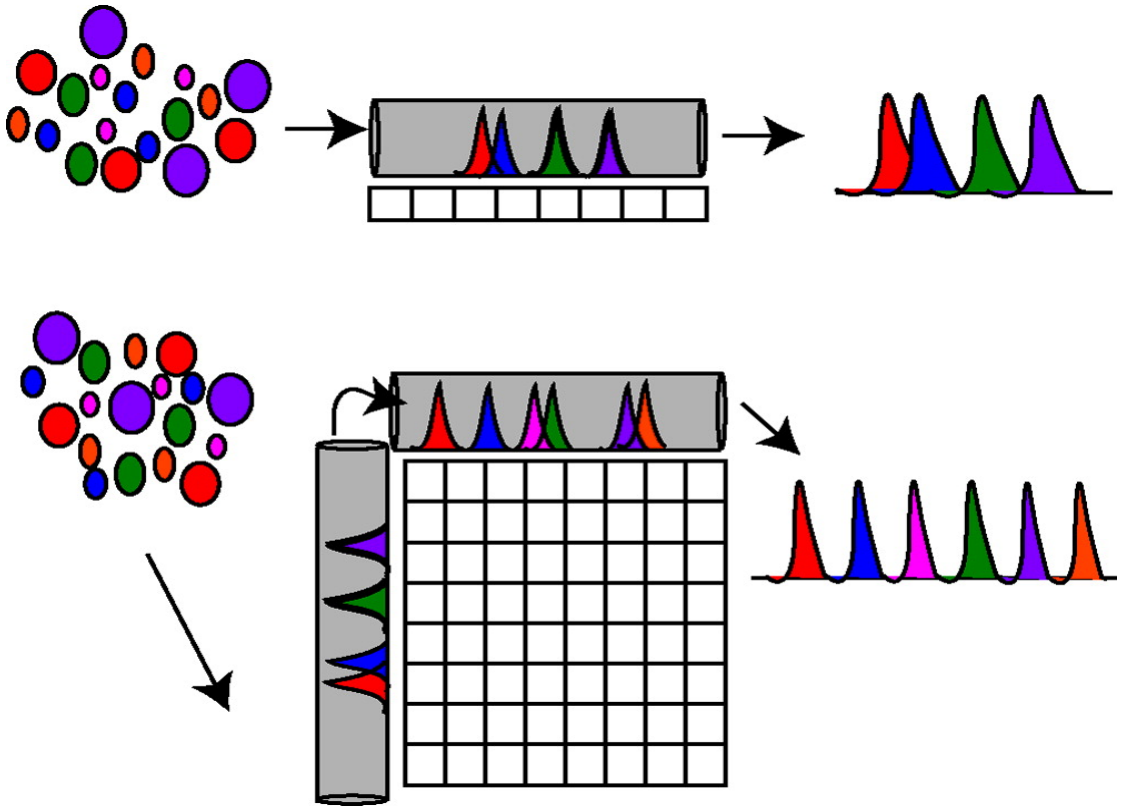
Principle of Gel Electrophoresis

- Gel electrophoresis is a method for separation and analysis of macromolecules (DNA, RNA and proteins) and their fragments, based on their size and charge.
- In simple terms, proteins move in electric field. Their relative speed depends on the charge, size, and shape of the protein.





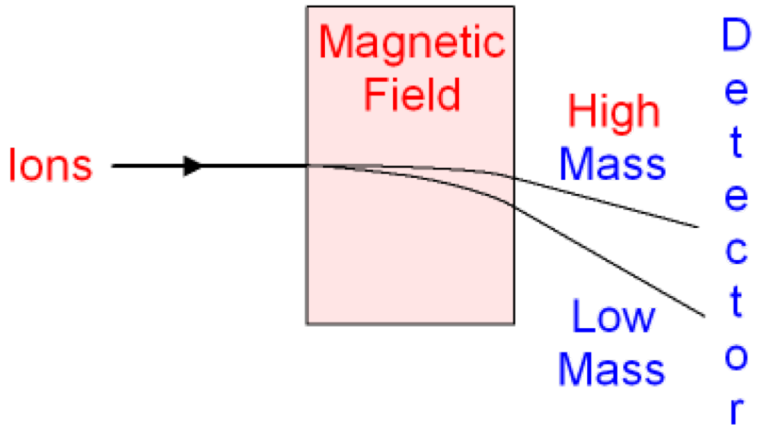
Visualization of 1D and 2D separations



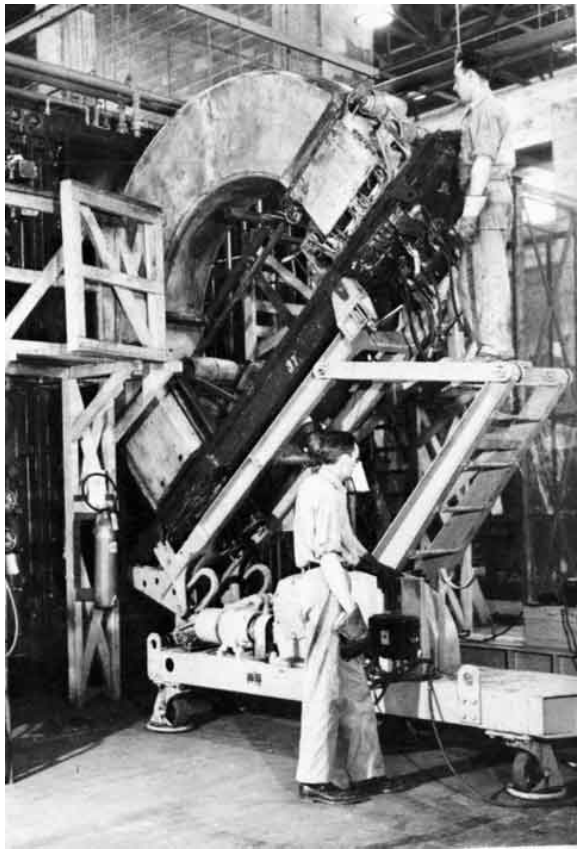
Identical samples of six species are separated by 1D and 2D techniques. Although the column shown for 1D separation has a theoretical peak capacity of eight (indicated by the boxes below the column), the 1D technique is able to clearly resolve only four distinct peaks. The addition of a second chromatographic dimension greatly improves the theoretical peak capacity (8 8 = 64) as shown in the boxes below the columns. The second column is able to improve the separation of overlapped peaks so that clearly resolved peaks from all six species can be clearly identified.

What is Mass Spectrometry (MS)?

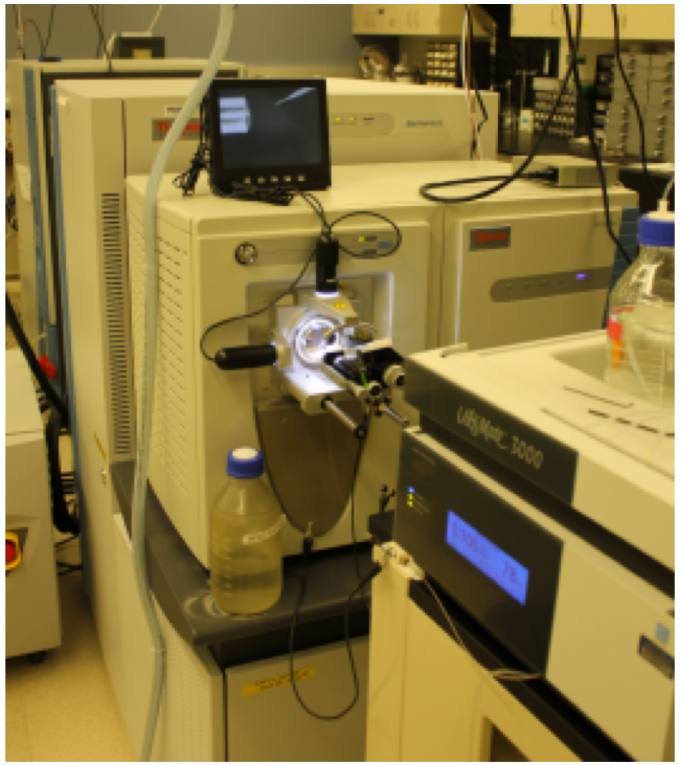
- An analytical technique that measures the mass-to-charge ratio (m/z) of a molecule to determine its identity and quantify its abundance.
  - The m refers to the molecular or atomic mass number and z to the charge number of the ion; however, the quantity of m/z is dimensionless by definition. An ion of 100 atomic mass units (m = 100) carrying two charges (z = 2) will be observed at m/z = 50.



Mass Spectrometer

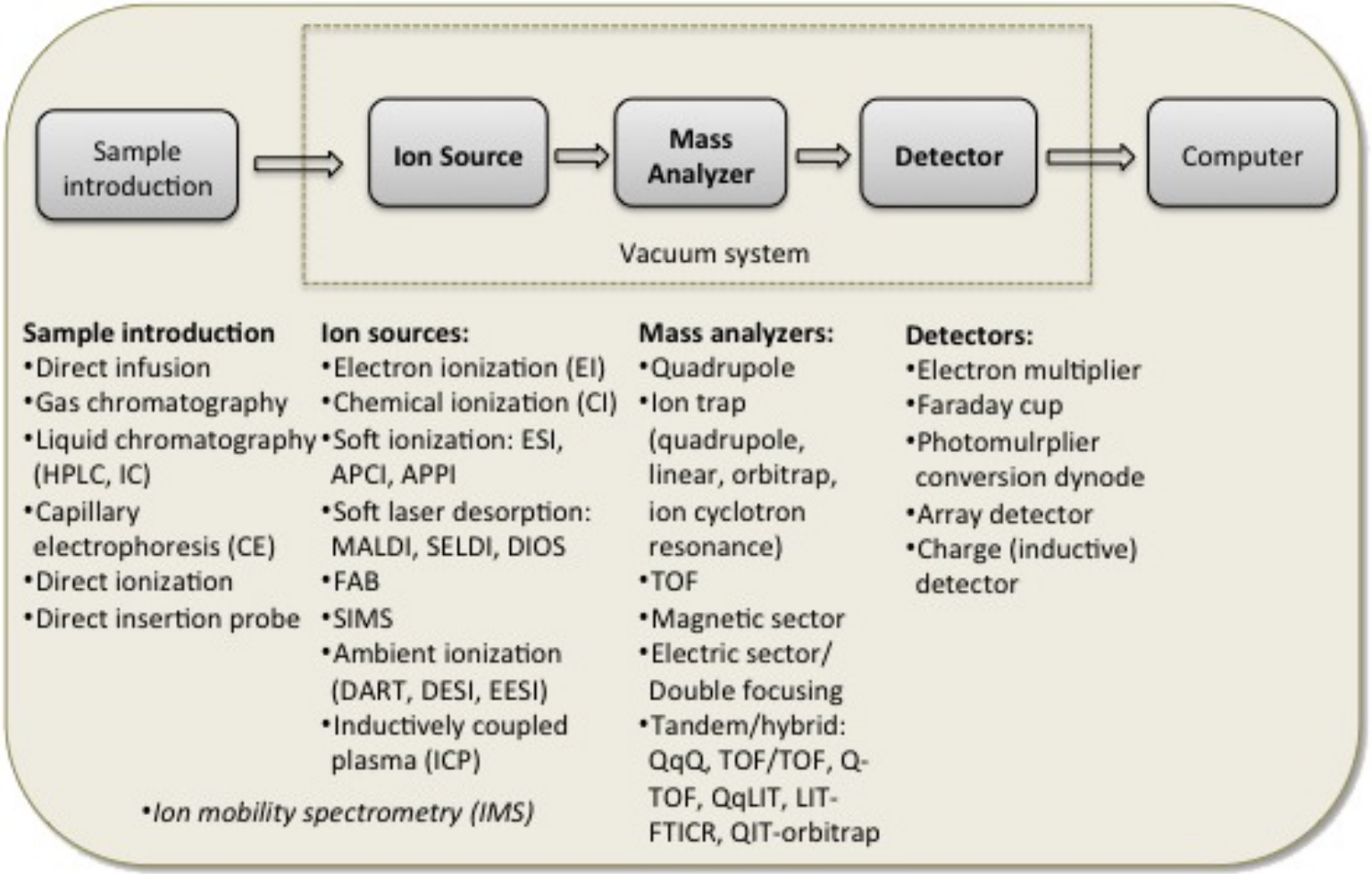


Calutron mass spectrometers were used in the Manhattan Project for uranium enrichment at Oak Ridge National Lab during World War II.



Mass spectrometer at Oak Ridge in current (2015).

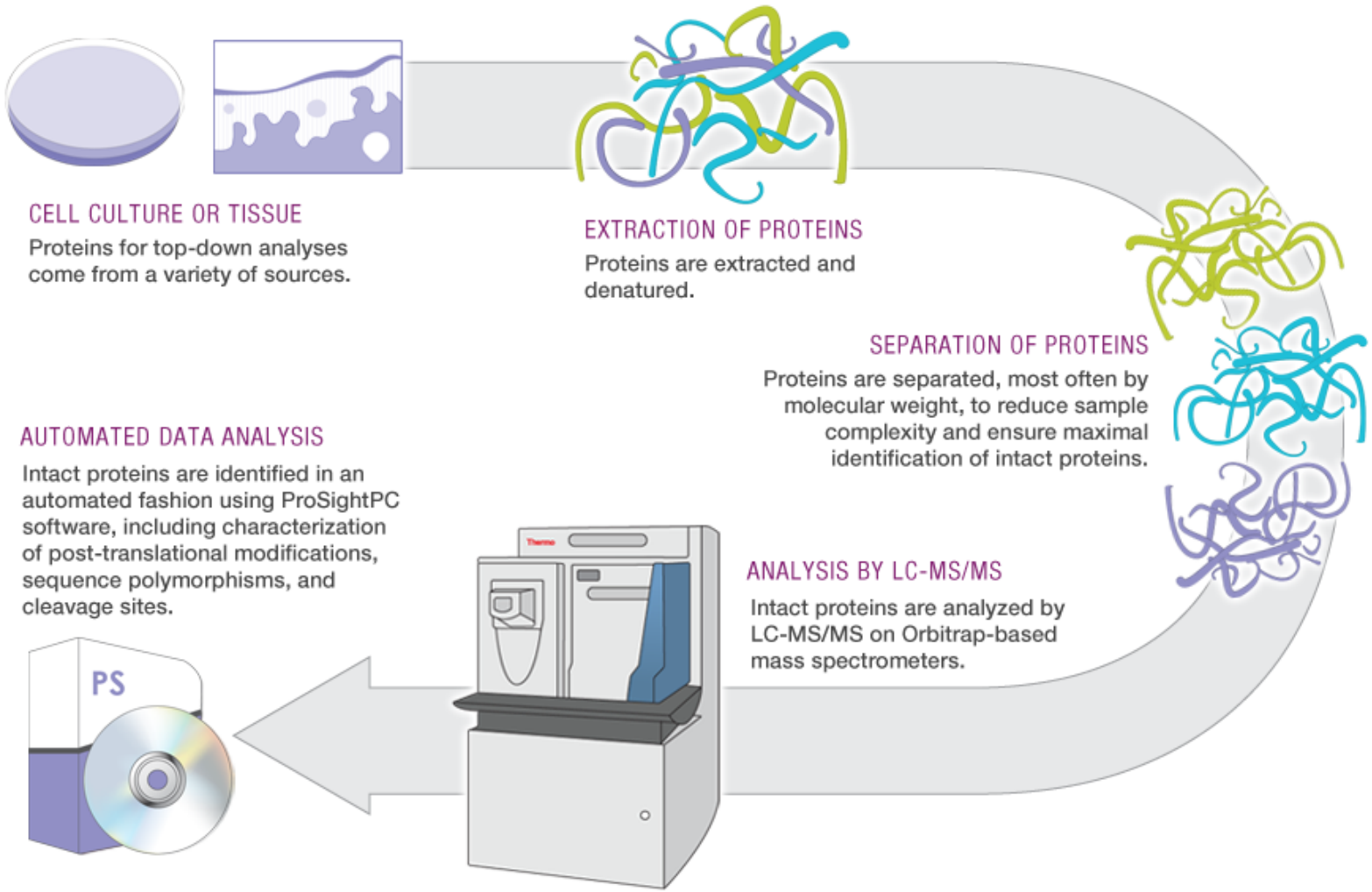
Components of a Mass Spectrometer



Mass Spectrometry

- Key Points:
  - Mass spectrometers work on samples in a gaseous state.
  - The gaseous samples are ionized by an ion source (Ionized).
  - Mass analyzers separate ionized samples according to their mass-to-charge ratio (Selected).
  - A particle's mass can be calculated very accurately based on parameters such as long it takes to travel a certain distance or its angle of travel (Detected).

Top-down analysis



http://planetorbitrap.com/top-down-proteomics#tab:overview

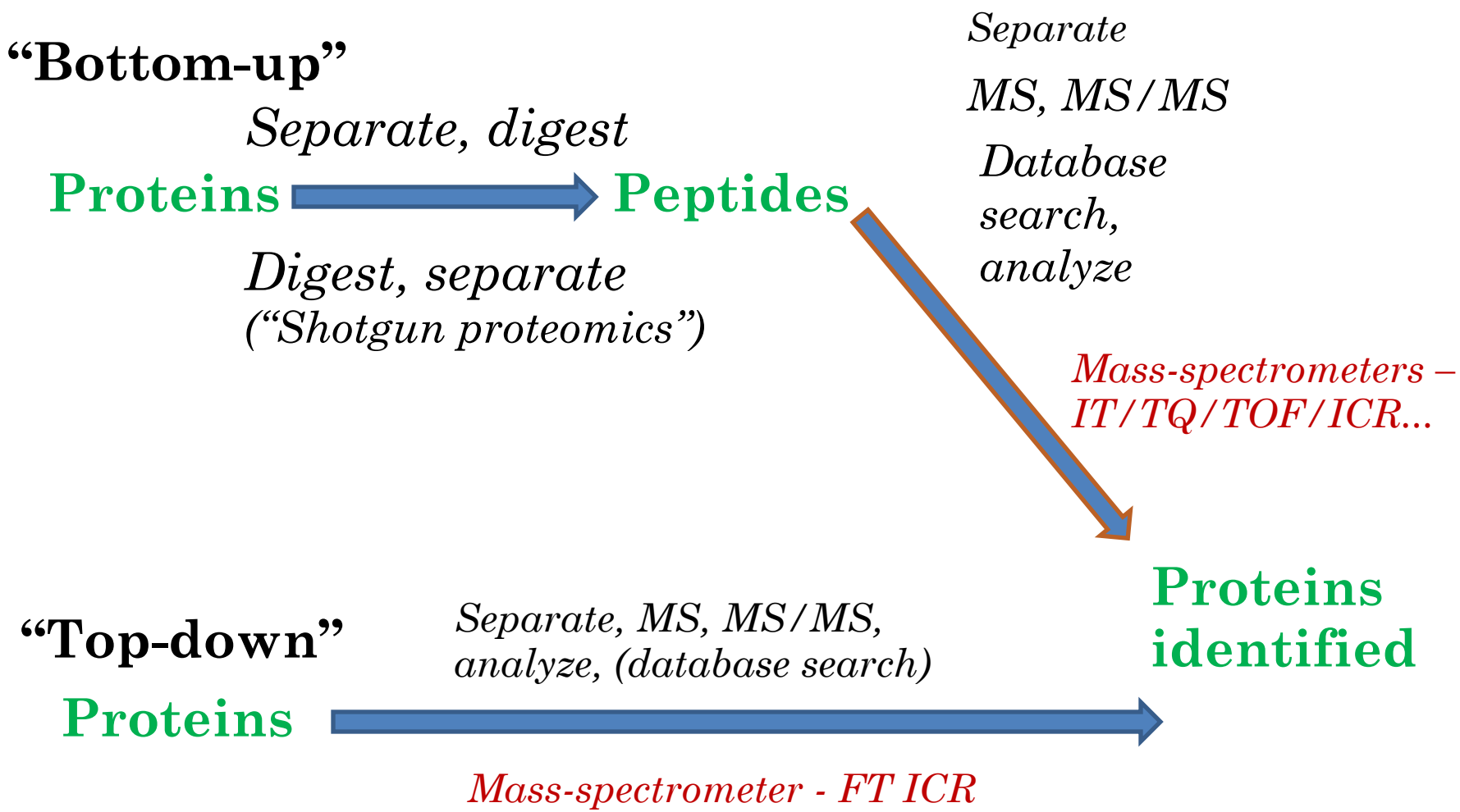
Top-down Advantages

- Top-down proteomics allows for analysis of an entire protein molecule without digestion and also allows for low mass protein detection. 100% protein sequence coverage is possible - potential access to the complete protein sequence and the ability to locate and characterize PTMs.
- The main advantages of the top-down approach include the ability to detect degradation products, sequence variants, and combinations of post-translational modifications.
- Characterization of small proteins represents a significant challenge for bottom up proteomics due to the inability to generate sufficient tryptic peptides for analysis. Top-down proteomics allows for low mass protein detection, thus increasing the repertoire of proteins known

Top-down Disadvantages

- The top down approach was relegated to analysis of individual proteins or simple mixtures, while complex mixtures and proteins were analyzed by more established methods such as Bottom-up proteomics.
- Top-down proteomics interrogation can overcome problems for identifying individual proteins, but has not been achieved on a large scale due to a lack of intact protein fractionation methods that are integrated with tandem mass spectrometry.
- The favored instrumentation (FT-ICR, hybrid ion trap FT-ICR or hybrid ion trap–orbitrap) are expensive to purchase and operate.
- bioinformatics tools for top-down proteomics are primitive compared to those for bottom-up

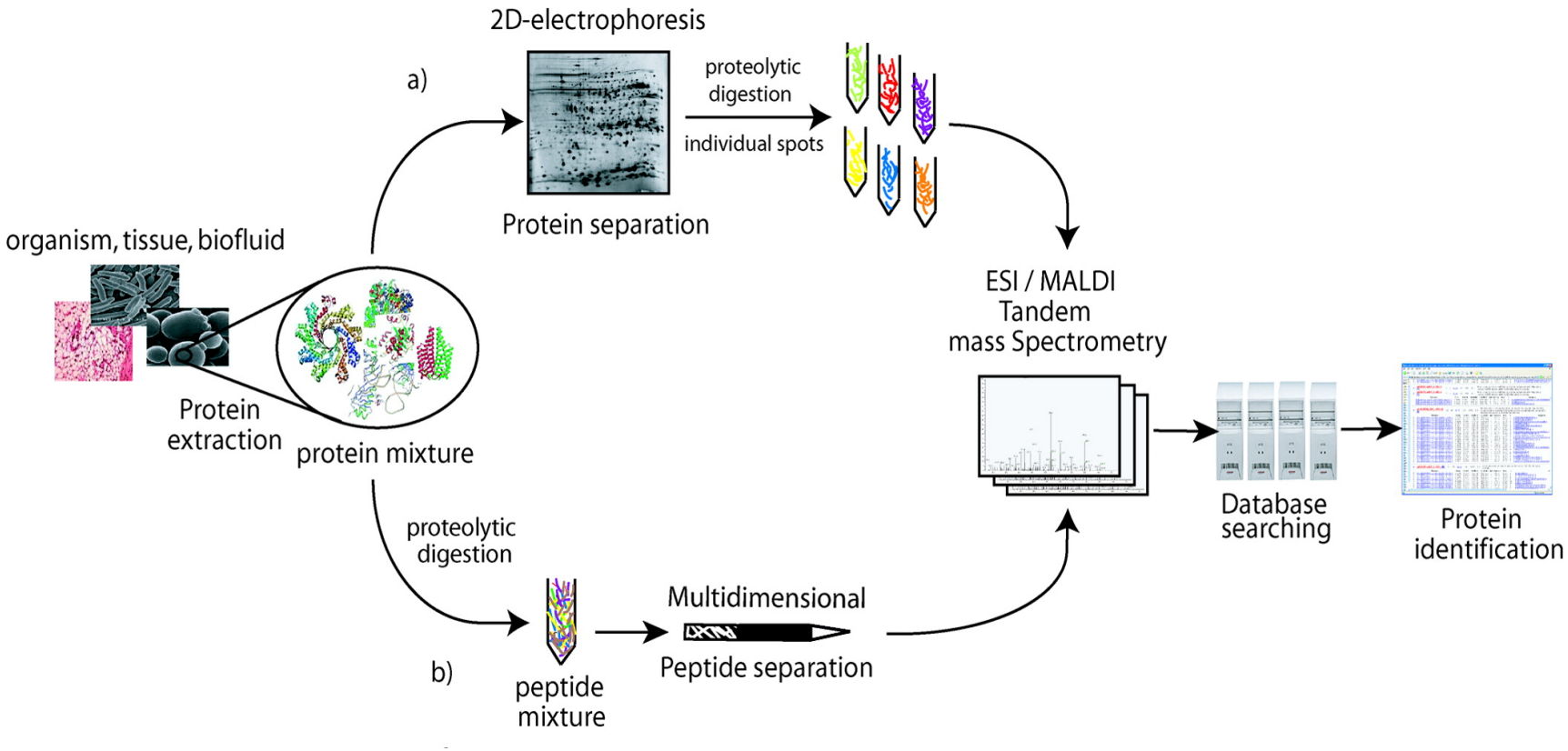
Proteomics: Various Approaches



Bottom-up Proteomics

- Bottom-up proteomics is a common method to identify proteins and characterize their amino acid sequences and PTMs by enzymatic digestion of proteins prior to analysis by mass spectrometry
- The proteins may first be purified (e.g., Gel electrophoresis) resulting in one or a few proteins in each enzymatic digest.
- Alternatively, the crude protein extract is digested directly, followed by one or more dimensions of separation of the peptides by liquid chromatography coupled to mass spectrometry ("shotgun proteomics")

Bottom-up Proteomics: Two Approaches



M. L. Fournier, J. M. Gilmore, S. A. Martin-Brown, and M.P. Washburn, "Multidimensional Separations-Based Shotgun Proteomics", Chem. Rev. 2007, 107, 3654-3686

Bottom-up Advantages

- Bottom-up proteomics is the most mature and most widely used approach for protein identification and characterization.
- Less sophisticated instrumentation and expertise.
- High throughput.
- Commercial instruments with control software and open-source bioinformatics tools optimized for bottom-up applications are available from several vendors.

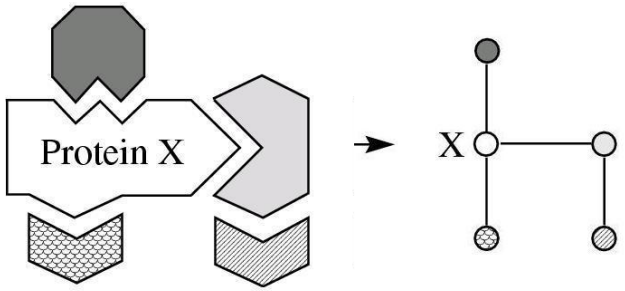
Bottom-up Disadvantages

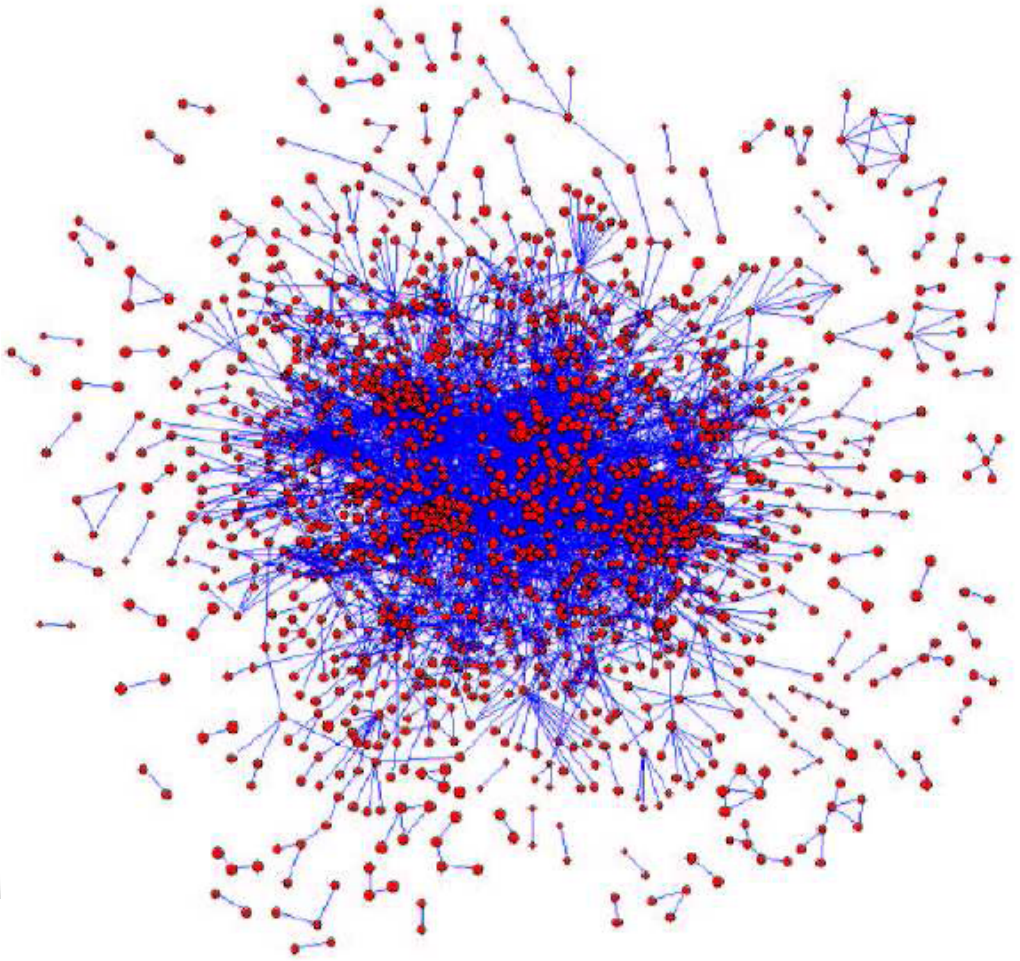
- Most importantly, only a fraction of the total peptide population of a given protein is identified. Therefore, information on only a portion of the protein sequence is obtained - PTM and isoform information is often lost.
- Confidence in protein ID strongly depends on restriction criteria and reference database.
- Other problems include the loss of information about low-abundance peptides in mass spectra dominated by high-abundance species.

Biological networks

- **Biological nets**

  E.g., Protein-protein interaction (PPI) networks





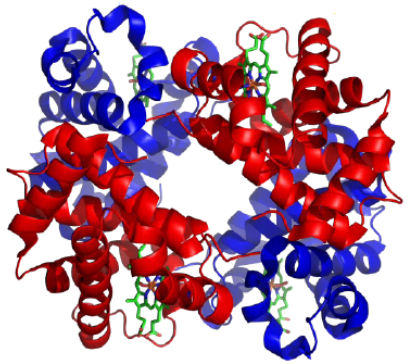
Protein-protein interaction (PPI) networks

- A *protein-protein interaction (PPI)* usually refers to a physical interaction, i.e., binding between proteins
- Can be other associations of proteins such as functional interactions – e.g., synthetic lethality: type of a "genetic interaction".

Protein-protein interaction (PPI) networks

- PPIs are very important for structure and function of a cell:
  - Participate in signal transduction (*transient interactions*)
    - Play a role in many diseases (e.g., cancer)
  - Can be *stable interactions* forming a *protein complex*

    (a form of a quaternary protein structure, set of proteins which bind to do a particular function, e.g., ribosome, hemoglobin)



- PPIs are essential to almost every process in a cell
- Thus, understanding PPIs is crucial for understanding life, disease, development of new drugs (most drugs affect PPIs)

Protein-protein interaction (PPI) networks

Methods to detect PPIs

- Biological and computational approaches
- None are perfect
  - High rates of *false positives*
    - Interactions present in the data sets that are not present in reality
  - High rates of *false negatives*
    - Missing true interactions

Protein-protein interaction (PPI) networks

Methods to detect PPIs

- PPIs initially studied individually by small-scale biochemical techniques (SS)
- However, large-scale (high-throughput) interaction detection methods (HT) are needed for high discovery rates of new protein interactions
- SS of better "quality," i.e., less noisy than HT
- However, HT are more standardized, while SS are performed differently each time
- SS are biased – the focus is on the subsets of proteins interesting to particular researchers
- HT – view of the entire proteome

Protein-protein interaction (PPI) networks

Methods to detect PPIs

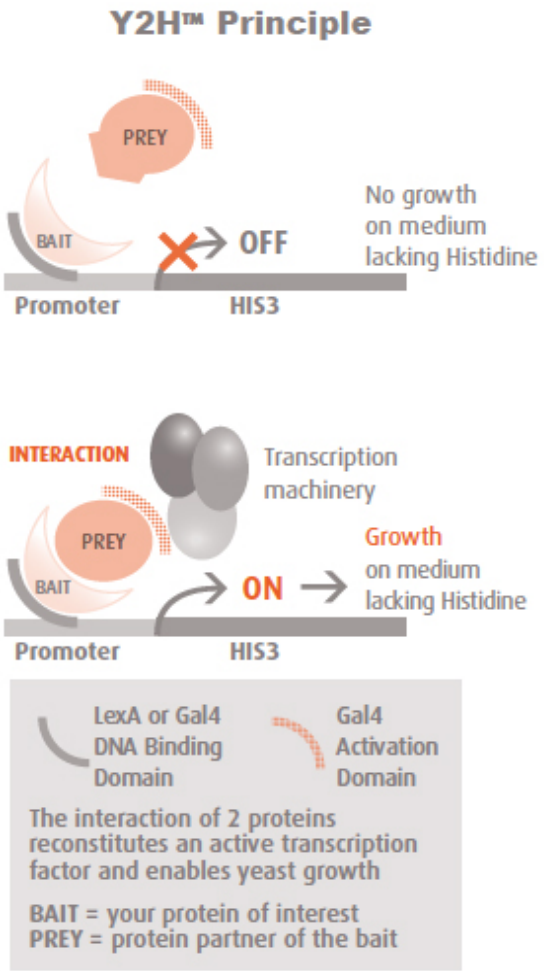
- Physical binding
  - Yeast 2-hybrid (Y2H) screening
  - Mass spectrometry of purified complexes
- Functional associations
  - Correlated mRNA expression profiles
  - Genetic interactions
  - In silico (computational) methods
- In many cases, functional associations do take the form of physical binding

Protein-protein interaction (PPI) networks

Yeast two-hybrid assay

- *Binary* PPIs
- Pairs of proteins to be tested for interaction are expressed as *artificial* (genetically engineered) *fusion proteins* in yeast:
  - One protein is fused to a *reporter gene* (a gene attached to another gene of interest)
  - The other is fused to a *transcription factor*
  - Any interaction between them is detected by the transcriptional activation of the reporter gene

http://www.sumanasinc.com/webcontent/animations/content/yeasttwohybrid.html



Protein-protein interaction (PPI) networks

Yeast two-hybrid assay

- This method is scalable to the entire proteome
- Directly tests a protein pair for an interaction
- But high noise rate (50%, even up to 70%)
  - Because Y2H investigates interactions between:
    - artificial, fusion proteins
    - in the yeast
    - in the yeast's *nucleus*
  - **Each of these steps is noisy**
  - Proteins need to be in their native environment, not in nucleus
    - E.g., although proteins can physically bind, they never do so inside a cell, because of different localization, or because they are never simultaneously expressed

Protein-protein interaction (PPI) networks

Mass spectrometry of purified complexes

- Individual proteins are tagged and used as hooks to biochemically purify whole protein complexes
- Complexes separated and components identified by mass spectrometry (MS)
  - MS measures mass-to-charge ratio of ions
- TAP (Tandem Affinity Purification)
- HMS-PCI (High-Throughput MS Protein Complex Identification)
- Not binary but co-complex data

Protein-protein interaction (PPI) networks

Mass spectrometry of purified complexes

- Pros:
  - Detects real complexes in their physiological settings
  - Consistency check is possible by tagging several members of a complex
  - Good for screening permanent/stable interactions
- Cons:
  - Might miss some complexes that are not present under given cellular conditions
  - Tagging may disturb complex formation
  - Loosely associated components can be washed off during purification

Protein-protein interaction (PPI) networks

Functional associations

- Correlated mRNA expression profiles
  - Results in a gene expression correlation network
- Co-expression means that resulting proteins *could* interact
- Co-expression overlaid over PPI data, e.g. tool KeyPathwayMiner (http://tomcat.compbio.sdu.dk/keypathwayminer/)

Protein-protein interaction (PPI) networks

Functional associations

- Genetic interactions
  - Two non-essential genes that cause lethality when mutated at the same time form a *synthetic lethal* interaction
  - Such genes are often functionally associated and their encoded proteins may also interact physically

Protein-protein interaction (PPI) networks

Functional associations

- Genetic interactions

Cell

Leading Edge

Perspective

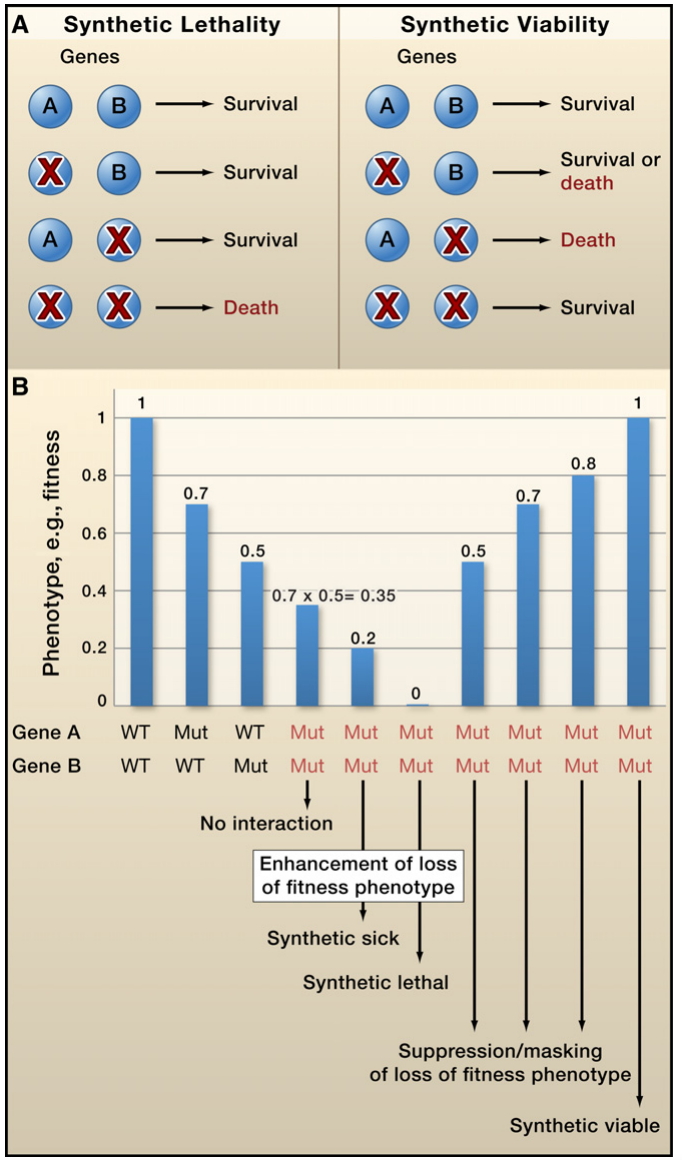
Genetic Interactions in Cancer Progression and Treatment

Alan Ashworth,[1,2,*] Christopher J. Lord,[1,2,*] and Jorge S. Reis-Filho[1,2,*]
[1]The Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, Fulham Road, London SW3 6JB, UK
[2]All authors contributed equally to this work
*Correspondence: alan.ashworth@icr.ac.uk (A.A.), chris.lord@icr.ac.uk (C.J.L.), jorge.reis-filho@icr.ac.uk (J.S.R.-F.)
DOI 10.1016/j.cell.2011.03.020

Figure 1. Gene Interactions in Cancer

(A) Extreme forms of genetic interaction are defined by synthetic lethality (in which a combination or synthesis of gene mutations causes cell death) and the reverse scenario, synthetic viability (in which a combination of gene effects rescues the lethal effects of a single gene change).

(B) Different modes of genetic interaction defined by quantitative effects on a phenotype, such as cell fitness. Here the value 1 represents the maximal fitness of cells, and the individual effects of changes in genes A (0.7) or B (0.5) are shown. When no interaction between genes A and B exists, the simple combination of effects (shown here as 0.7 × 0.5 = 0.35) is expected; any deviation from this value suggests an interaction between genes A and B.



Protein-protein interaction (PPI) networks

Quality and completeness of PPI data

- Data sets produced by different methods are often complementary
- Even data sets obtained by the same technique complement each other to some (large) extent
- Completeness of data sets:
  - Yeast: ~50% (~6K proteins, ~30K-60K interactions)
  - Human: ~10% (~25K proteins, ~260K interactions; ~300 million pairs to test)
  - Fly
  - Worm
  - Recently, herpes viruses (genome-wide coverage)

Protein-protein interaction (PPI) networks

PPI databases

- Biological General Repository for Interaction Datasets (BioGRID)
- Human Protein Reference Database (HPRD)
- *Saccharomyces* Genome Database (SGD)
- Munich Information Center for Protein Sequences (MIPS)
- Database of Interacting Proteins (DIP)
- Molecular Interactions Database (MINT)
- Online Predicted Human Interaction Database (OPHID) → **I2D**
- VirusMINT

- The lack of standardization
  - Different databases use different naming conventions
  - Inconsistencies in mapping between them
  - This can seriously jeopardize network topological analyses

*Distinguish between binary and co-complex data.