Network inference for biological systems

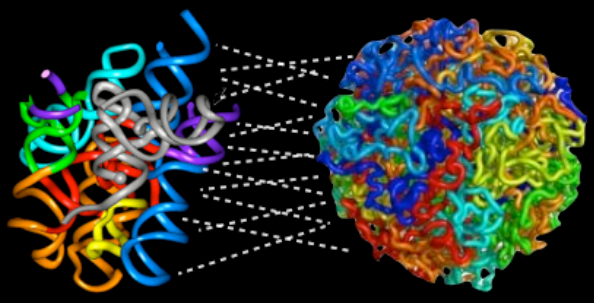
To reveal molecular mechanisms

By Joe Song
Visiting Professor, Department of Computer Science
Czech Technical University
Professor, Department of Computer Science
New Mexico State University

Updated: May 9, 19, 2019

The Interactome

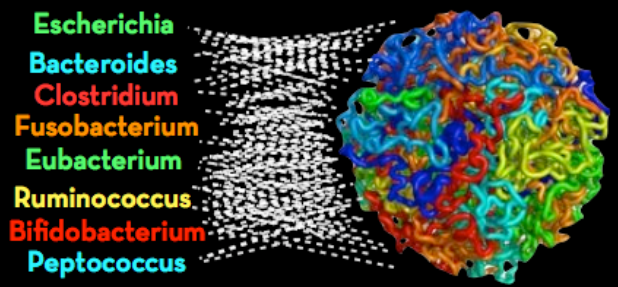
1,443 direct interactions



HIV
19 proteins

Homo sapiens
100,000 proteins

$10^{??}$ interactions



Gut microbiome
9 million unique genes

Homo sapiens
100,000 proteins

PMID: 19025396

• Interactome

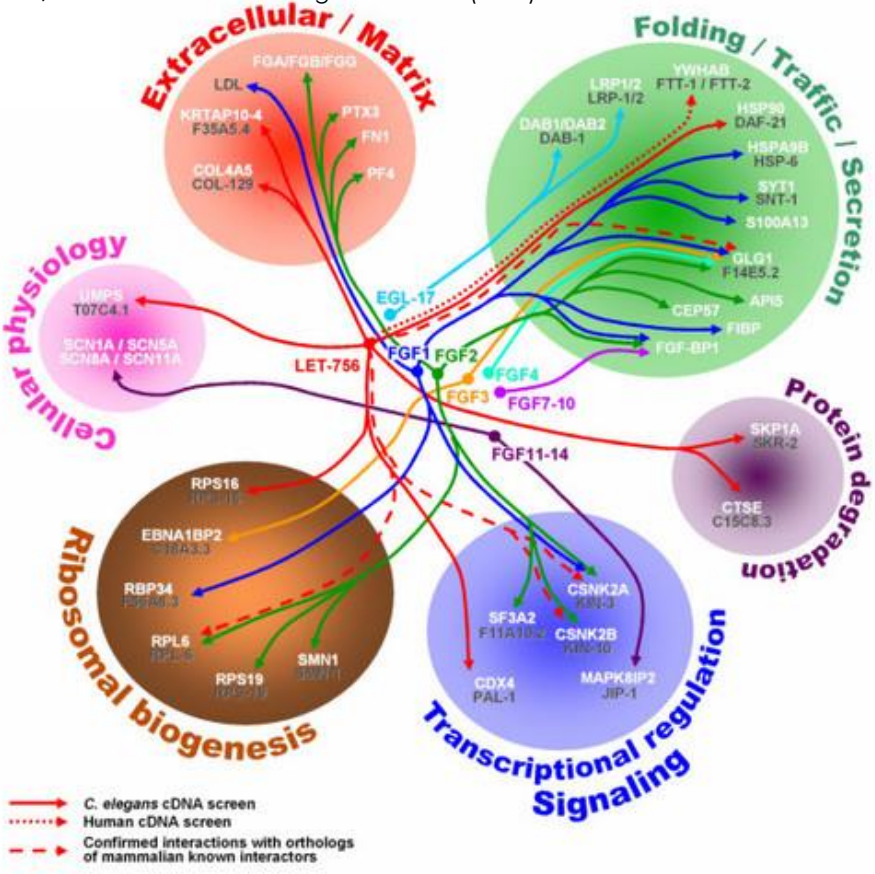
- Definition:
  - Interactome is a network of all molecular interactions in a cell
- From static:
  - What types of molecule are in a cell?
  - How many of each type are there?
- To dynamic:
  - How do molecules inside a cell work together via physical, genetic, and functional interactions to give rise to specific phenotypes?

• Experimental approach

One protein at a time–slow.

What other proteins do **fibroblast growth factor (FGF)** LET-765 interact with?

Popovici, Cornel, et al. "Direct and heterologous approaches to identify the LET-756/FGF interactome." *BMC genomics* 7.1 (2006): 105.



Interactome of FGF family:

- Use LET-765 as bait in two-hybrid cDNA library from worm.
- Interactors of various FGFs were grouped in six functional categories.
- Human (white) and worm *C. elegans* (grey) orthologs are indicated.

• Physical interactions in a cell

Directional	Non-directional
Protein → DNA: • Transcription regulation • Chromatin modeling Protein → RNA: • Post-transcription regulation via RNA binding proteins RNA → DNA • Epigenetic regulation RNA → RNA • lncRNA targeting at mRNA • microRNA suppressing mRNA RNA → Protein • lncRNA binding proteins Enzyme → metabolite: • Enzymatic reactions	Protein—protein: • Protein complex via chemical reactions • Phosphorylation / Dephosphorylation

• Physical interactions involving ncRNAs

- Recently understood

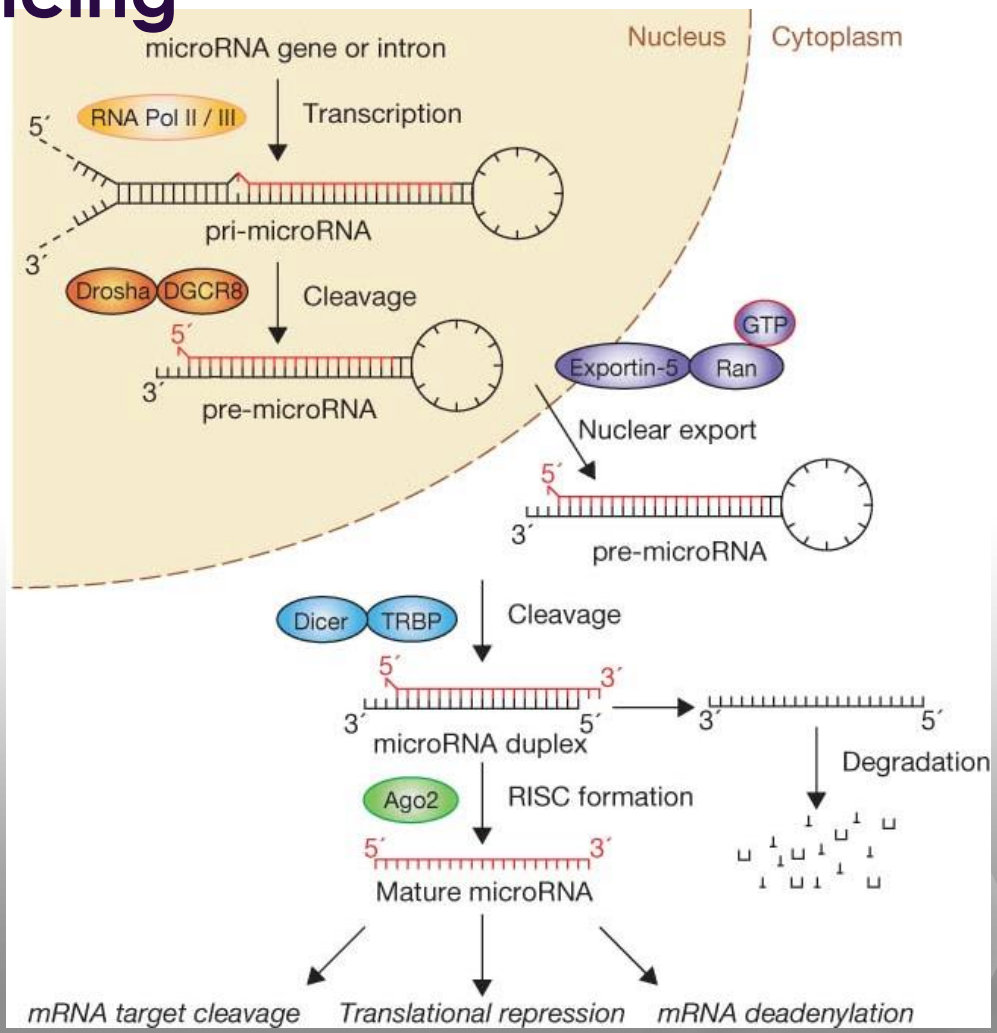
Directional
ncRNA → RNA: • RNA silencing: ncRNA suppresses expression of mRNA • RNA editing: insertion, deletion, and base substitution of nucleotides within the edited RNA molecule **ncRNA→ Protein (Chromatin modifying complexes) → DNA:** • DNA methylation: reduces accessibility to genomic DNA • Chromatin remodeling: changes access to condensed genomic DNA region by transcription factors

Non-coding RNA (ncRNA) is believed to be at the center of all this

• Functional interactions

- Definition: The causal effect of a molecule on phenotypes
- Examples of genetic interactions:
  - Does mutation in a gene cause cancer?
  - Is there a gene responsible for higher level of intelligence?
- Examples of epigenetic interactions:
  - Can methylation of a gene lead to developmental disorder?

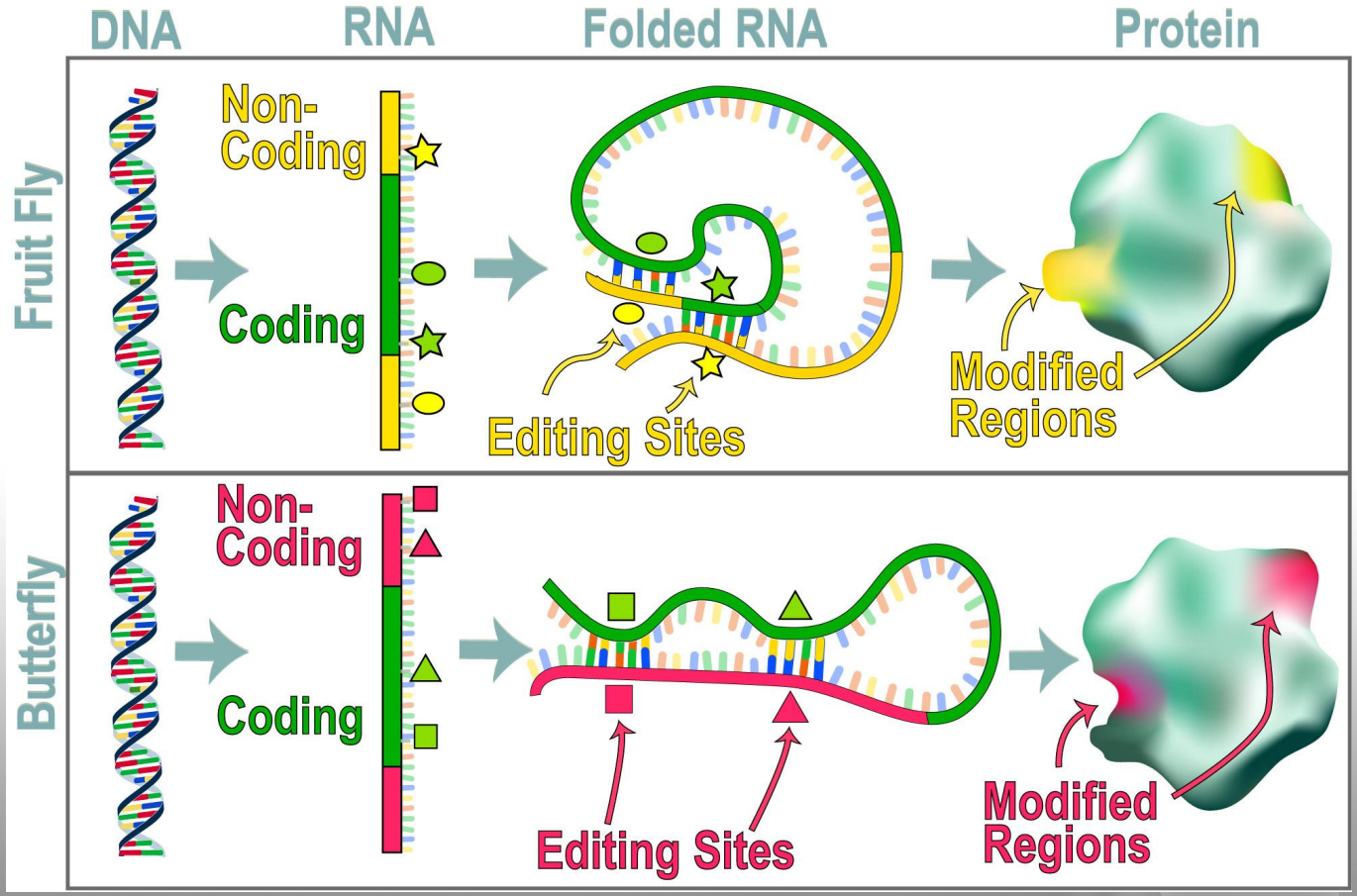
• RNA silencing



• RNA silencing

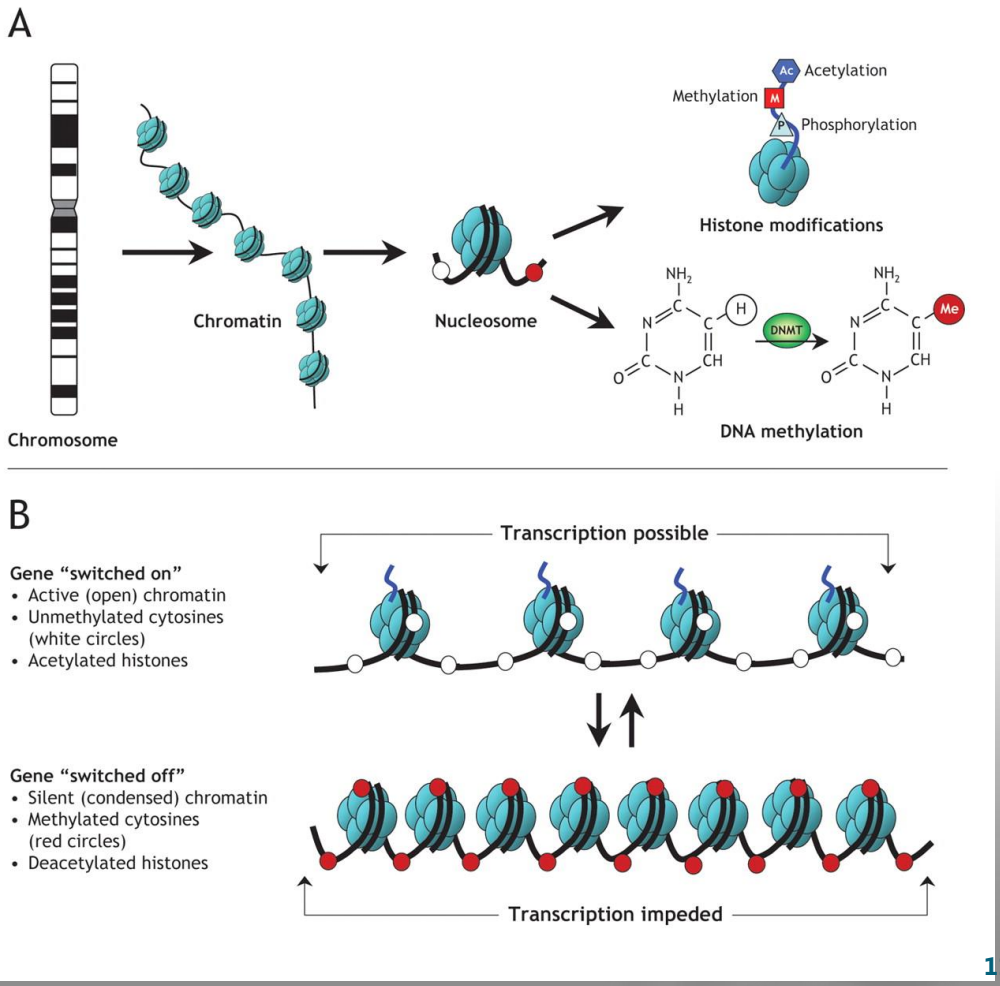
- mRNA target cleavage
  - mRNA was completed cut
- Translational repression
  - Imperfect match between microRNA and target mRNA.
  - mRNA is not cleaved
  - mRNA is bounded by Ago2
  - Free mRNA is reduced for translation
- mRNA deadenylation
  - polyA tails of mRNA is shortened
  - The life span of an mRNA positively depends on the length of polyA tail
  - The availability of mRNA is thus reduced in time

• RNA editing



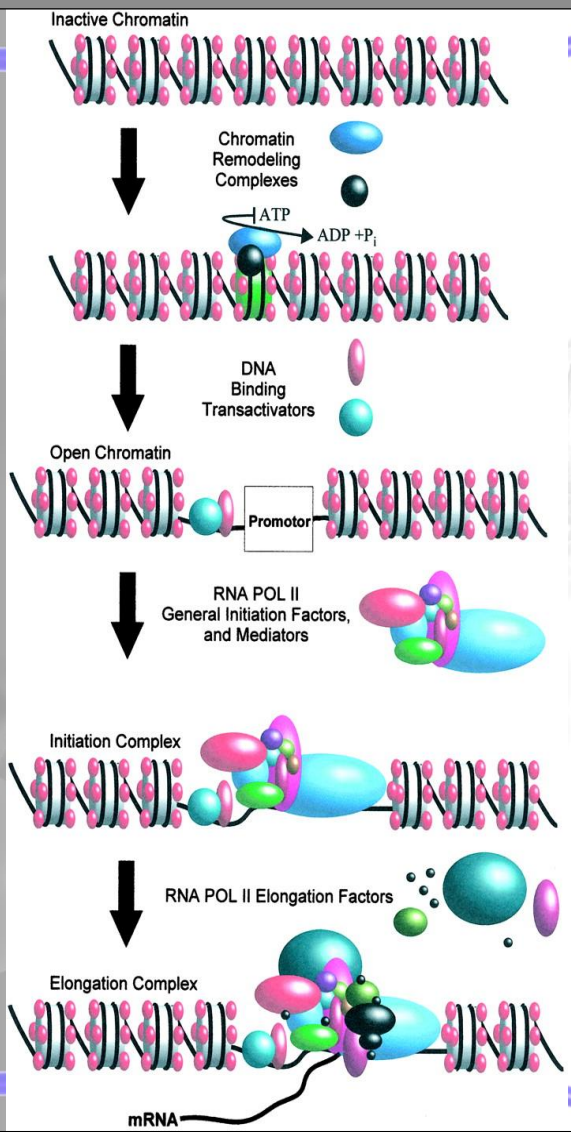
http://www.nsf.gov/news/mmg/media/images/1_rna_editing_f1.jpg

• DNA methylation and histone modification



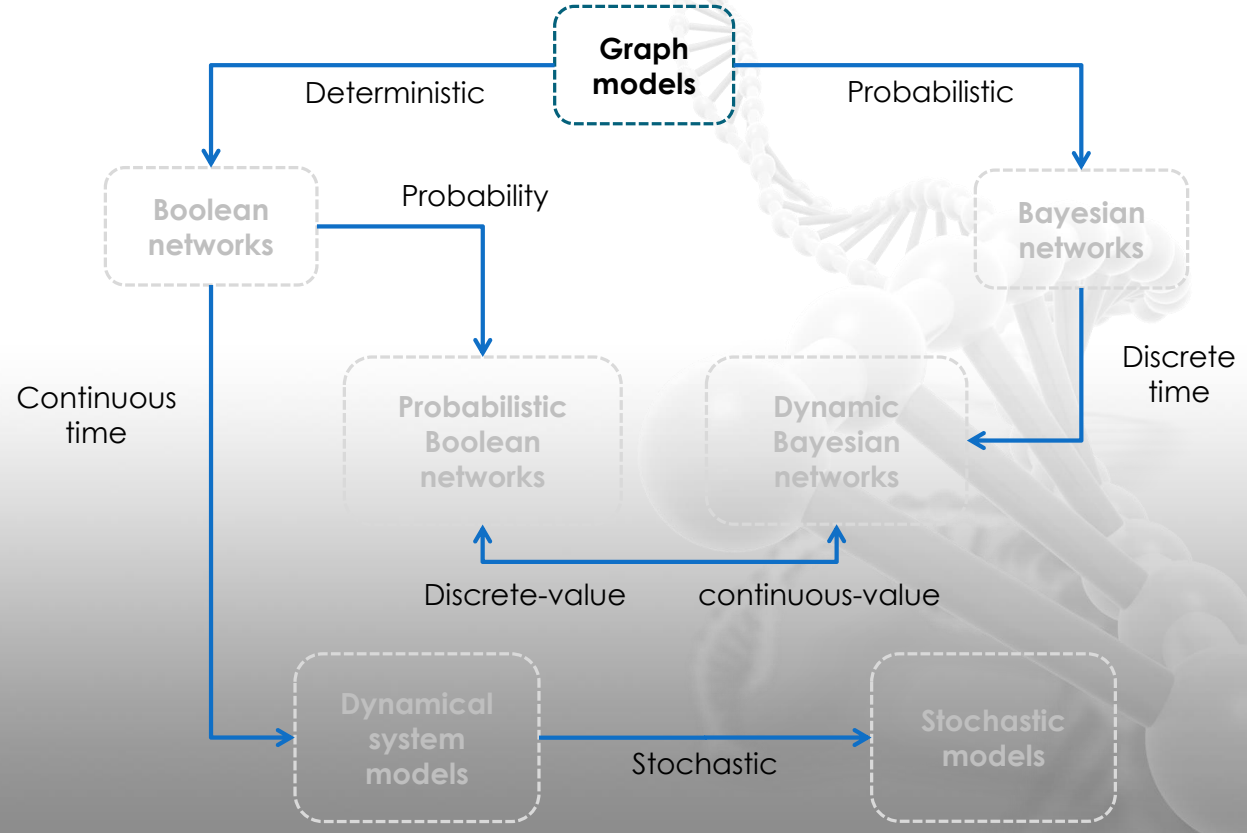
http://cnx.org/content/m26565/latest/graphics35.jpg

• Chromatin remodeling



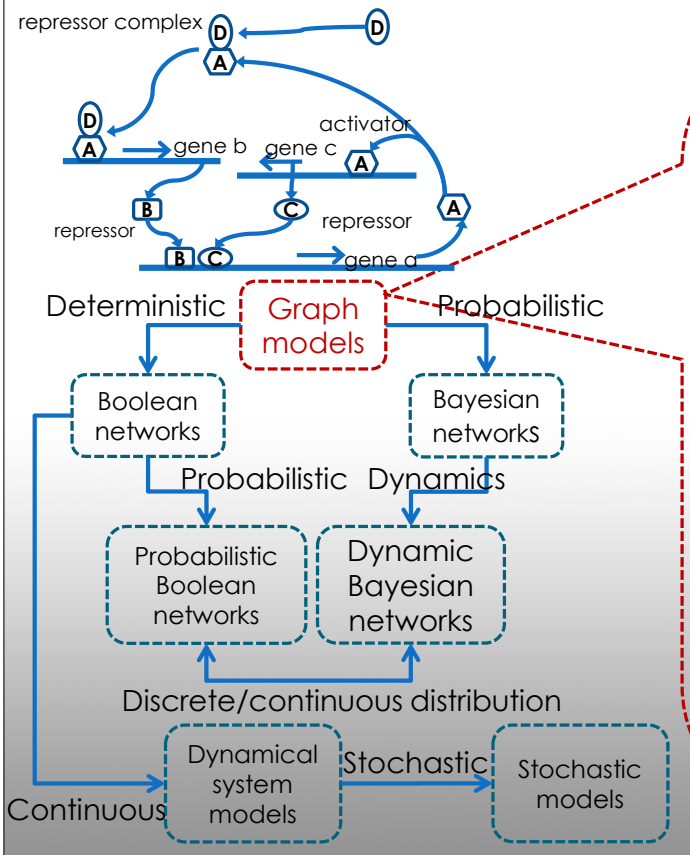
http://pharmrev.aspetjournals.org/content/54/1/129/F5.large.jpg

• Mathematical Models of Biological Systems



• Mathematical Models of Biological Systems

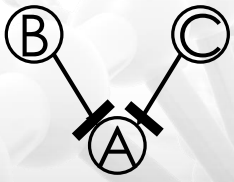
• Graph models



Positive (+) or negative (-) relationships as interaction:

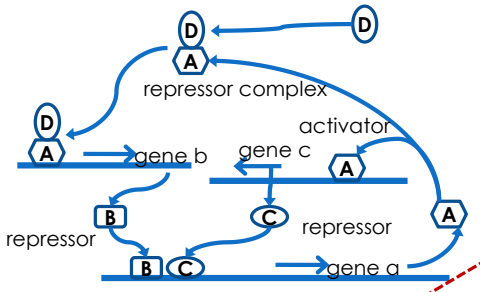
$$(a,\{b,c\},-)$$

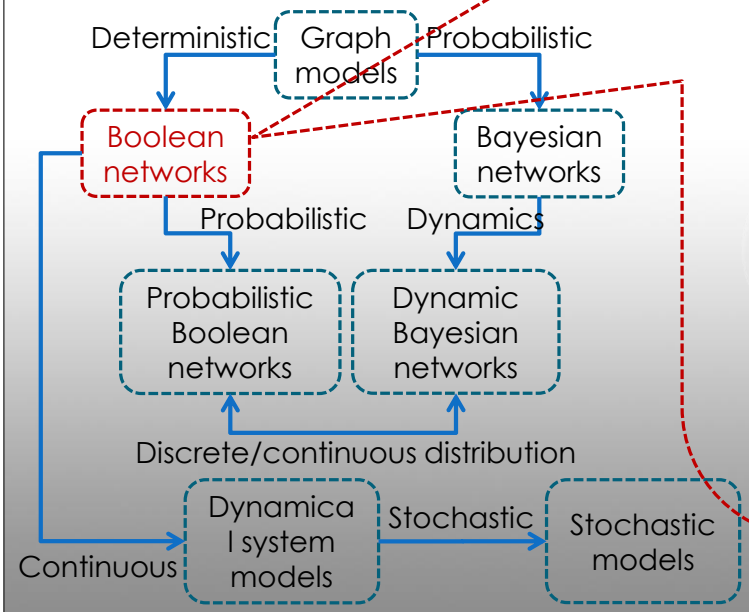
An example:



• Mathematical Models of Biological Systems

• Boolean networks





Boolean equations as interaction:

$$A[t] = \overline{B[t-1]C[t-1]}$$

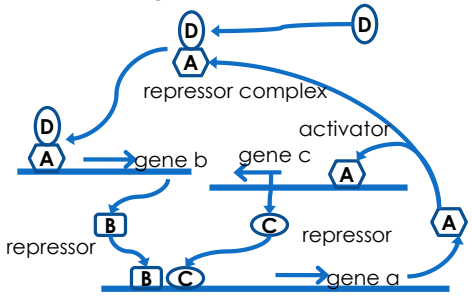
Example:

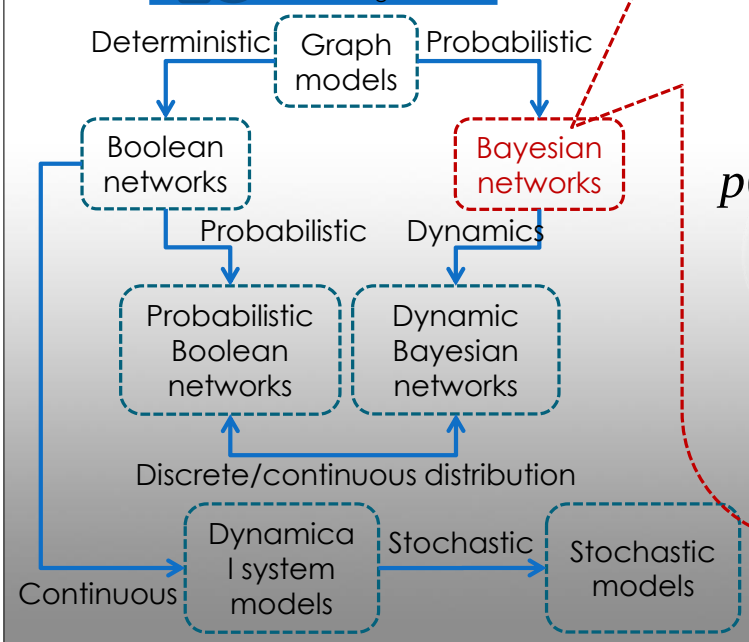
B[t-1]	C[t-1]	A[t]
0	0	1
0	1	1
1	0	1
1	1	0

(Kauffman, Journal of Theoretical Biology, 1969)

• Mathematical Models of Biological Systems

• Bayesian networks





- Conditional Boolean probability as interaction:

$$p(A \mid B, C)$$

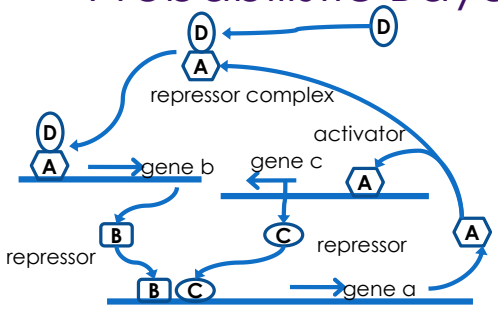
- No cycle is allowed
- Example:

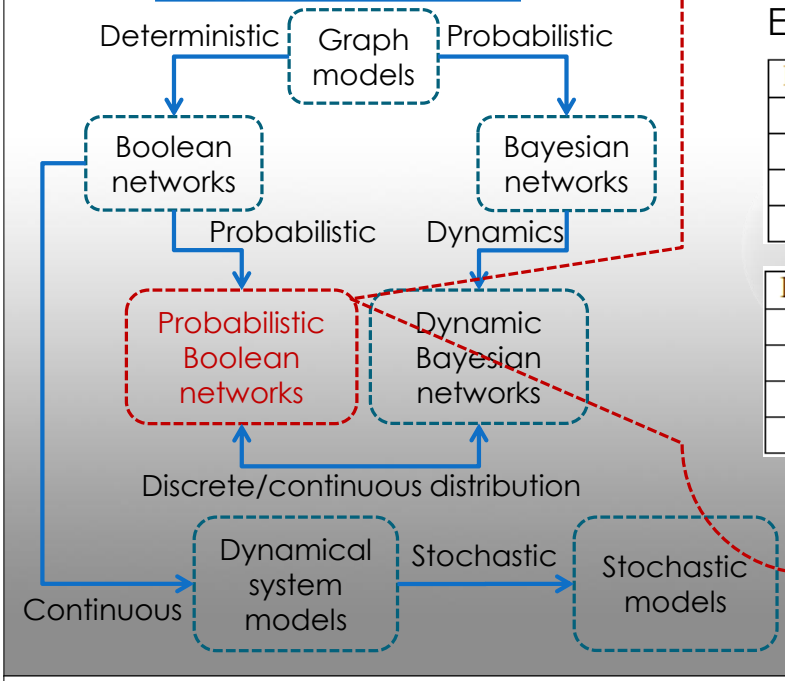
$$p(A \mid B = b, C = c) = N\left(\frac{1}{b \cdot c + 1}, \sigma^2\right)$$

(Friedman, Science, 2004)

• Mathematical Models of Biological Systems

• Probabilistic Bayesian Networks





Conditional Boolean probability with time as interaction:

$$p(A[t] \mid \overline{B[t-1]C[t-1]})$$

Example:

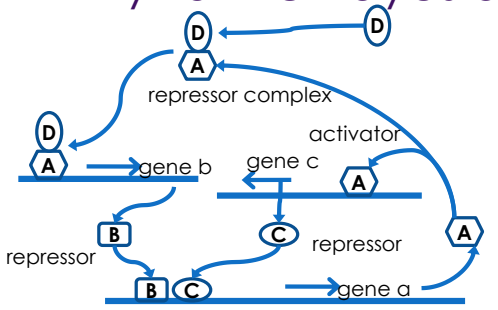
B[t-1]	C[t-1]	A[t] with probability 0.9
0	0	1
0	1	1
1	0	1
1	1	0

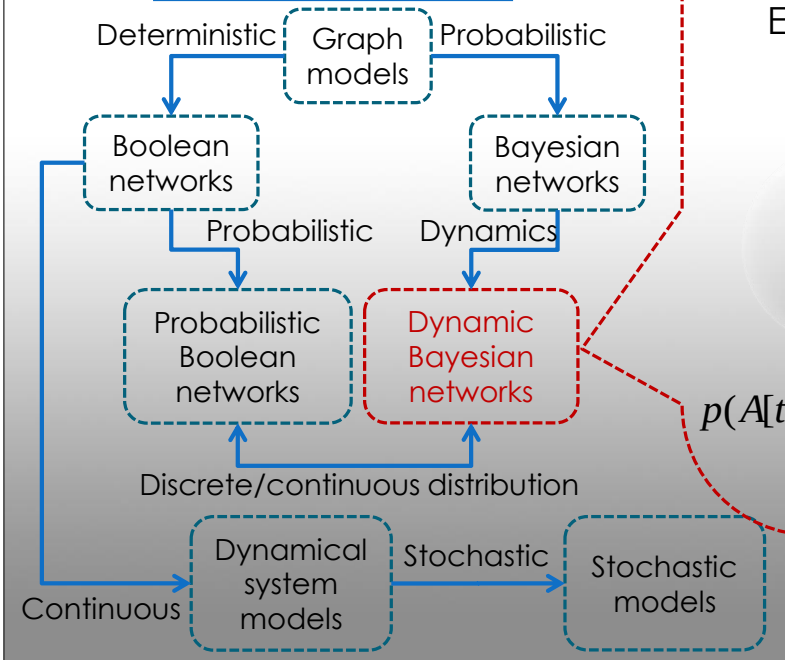
B[t-1]	C[t-1]	A[t] with probability 0.1
0	0	0
0	1	0
1	0	0
1	1	1

(Shmulevich ,*et. al.*, Bioinformatics, 2002)

• Mathematical Models of Biological Systems

• Dynamic Bayesian Networks

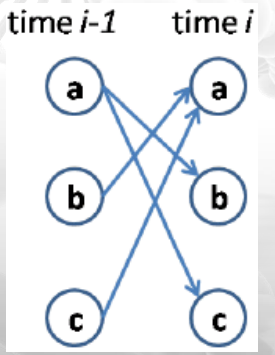




Conditional distributions with time as Interaction:

$$p(A[t] \mid B[t-1], C[t-1])$$

Example:

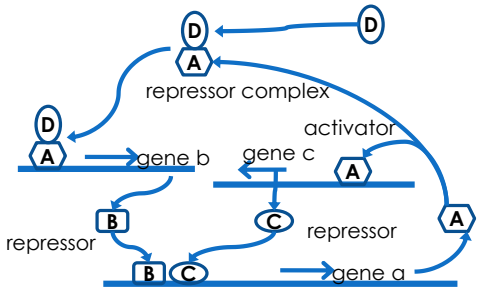


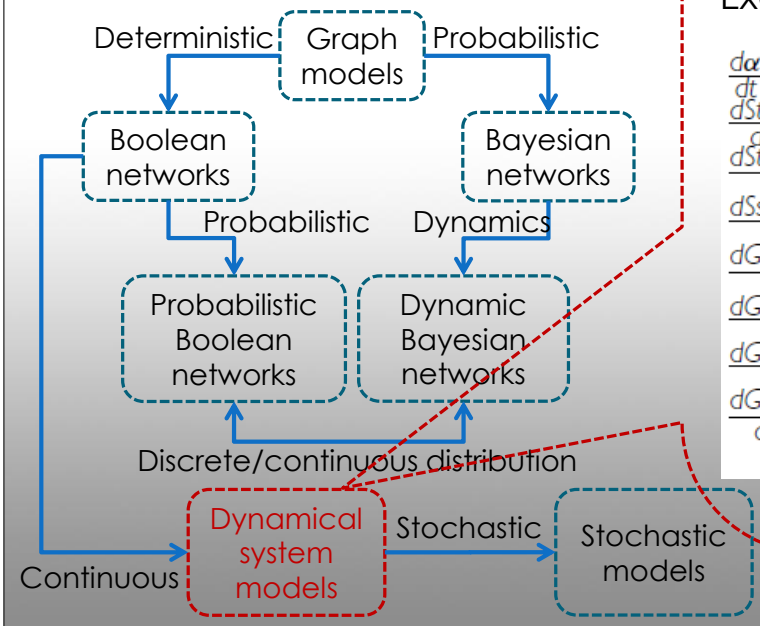
$$p(A[t] \mid B[t-1]=b, C[t-1]=c) = N(\frac{1}{b \cdot c+1}, \sigma^2)$$

(Murphy and Mian, 1999)

• Mathematical Models of Biological Systems

• Dynamical System Models





(De Jong, Journal of computational biology, 2002)

Differential equations as Interaction:

$$\frac{dA[t]}{dt} = f(B[t], C[t])$$

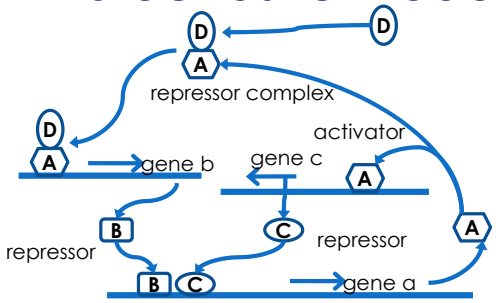
Example:

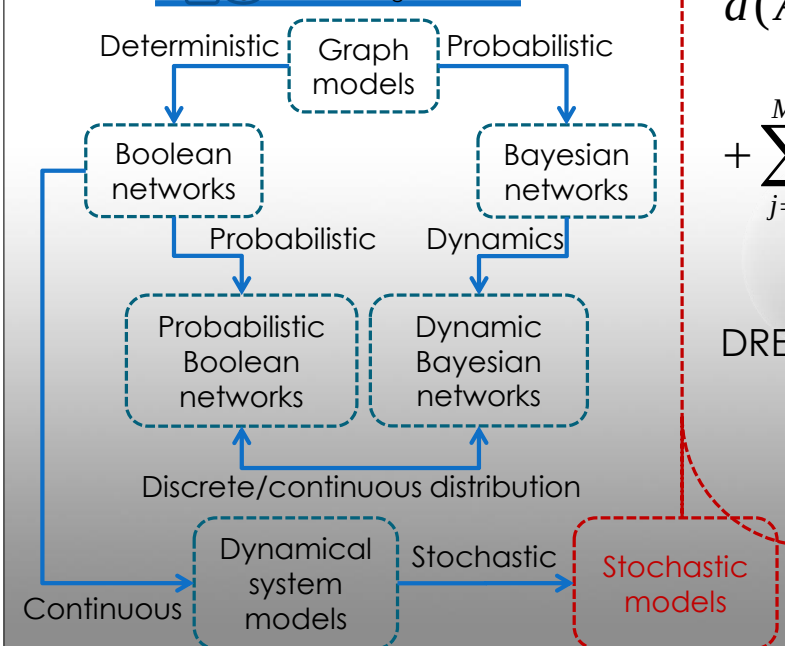
$$\frac{d\alpha}{dt} = -\nu_1$$
$$\frac{dSte2}{dt} = -\nu_2 + \nu_3 - \nu_5$$
$$\frac{dSte2_{active}}{dt} = \nu_2 - \nu_3 - \nu_4$$
$$\frac{dSst2_{active}}{dt} = \nu_{46} - \nu_{47}$$
$$\frac{dG\alpha\beta\gamma}{dt} = -\nu_6 + \nu_9$$
$$\frac{dG\alpha GTP}{dt} = \nu_6 - \nu_7 - \nu_8$$
$$\frac{dG\alpha GDP}{dt} = \nu_7 + \nu_8 - \nu_9$$
$$\frac{dG\beta\gamma}{dt} = \nu_6 - \nu_9 - \nu_{10} + \nu_{11} + \nu_{21} + \nu_{23} + \nu_{25} + \nu_{27} + \nu_{32} - \nu_{42} + \nu_{43}$$

Kofahl and Klipp, *Yeast*, 2004)

• Mathematical Models of Biological Systems

• Stochastic models





Stochastic differential equations as interaction:

$$d(A[t]) = \sum_{j=1}^{M} f_j(A[t], B[t], C[t], v_j)dt$$

$$+ \sum_{j=1}^{M} f_j'(A[t], B[t], C[t], v_j)dW_j[t]$$

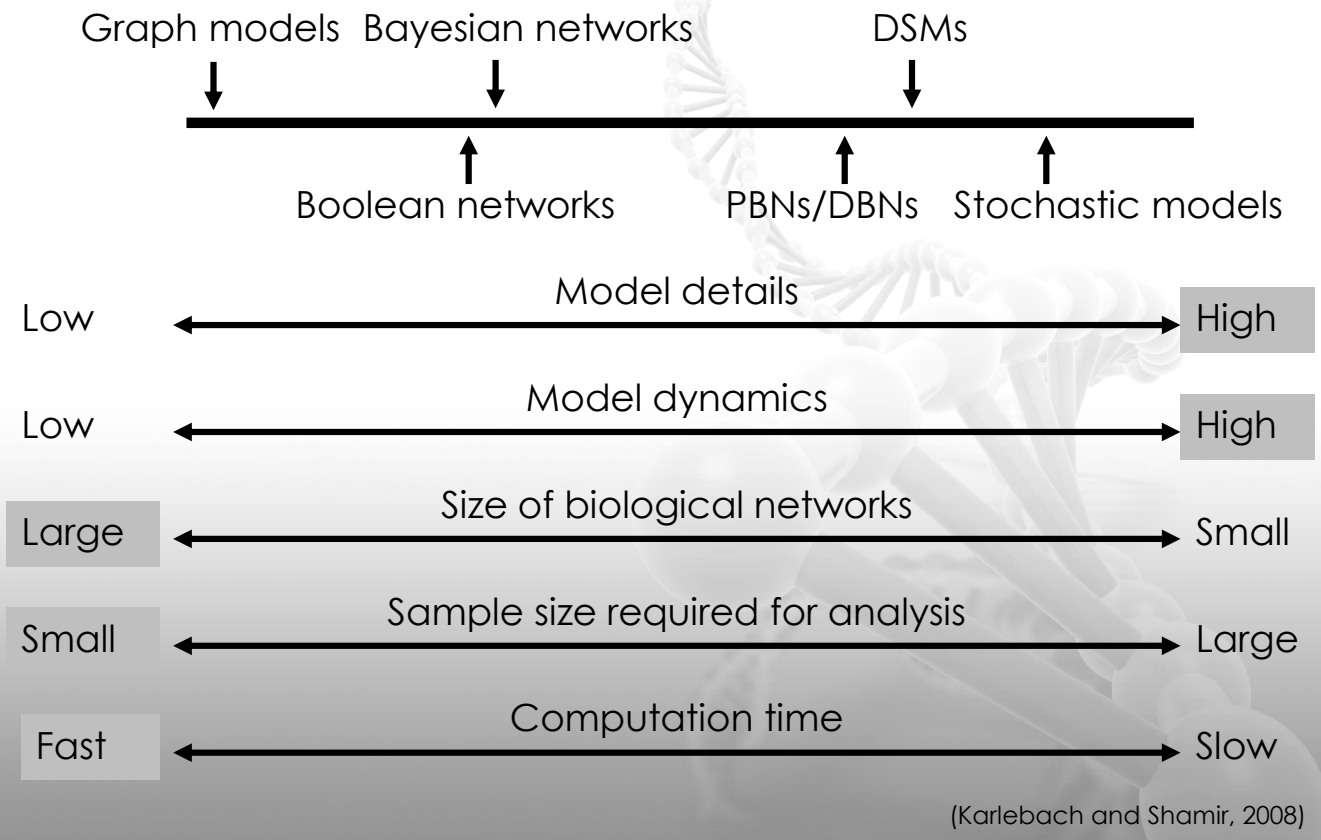
(Higham, SIAM REVIEW,2008)

DREAM 4 (2009)

$$\frac{dA[t]}{dt} = f(B[t], C[t]) + N(\mu, \sigma^2)$$

Brownian motion: seemingly random movement of particles suspended in a fluid.

• Comparison of network models



(Karlebach and Shamir, 2008)

• Network inference

- **Qualitative**: Determine edges in a network from observed snapshots of a system
  - Directional / undirected interaction
  - Direct / indirect interaction
  - Parents of a node
  - Often mathematically ill-conditioned
- Quantitative relationships among nodes are not required.

• Network Inference / Model Reconstruction

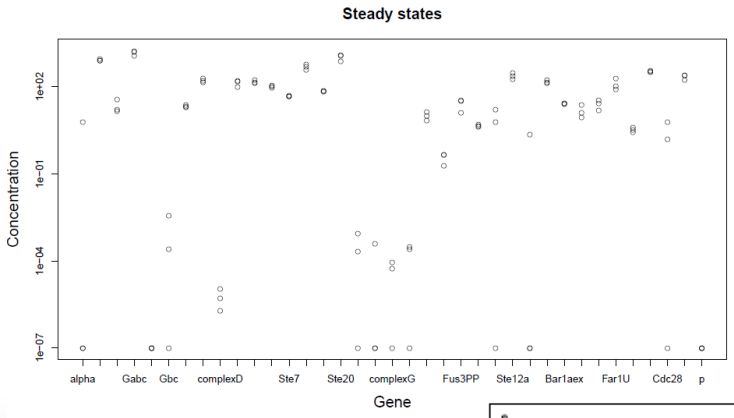
• Input

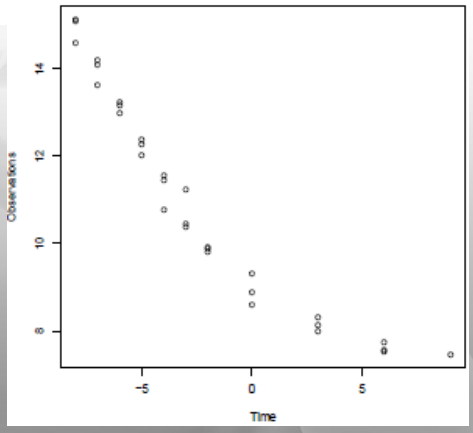
- Proteins
- mRNAs
- Metabolites

- Steady states
- Time courses

- Microarray
- Molecular imaging
- RNA-Seq (bulk or single-cell)

- Yeast Two Hybridization
- Chip-Seq

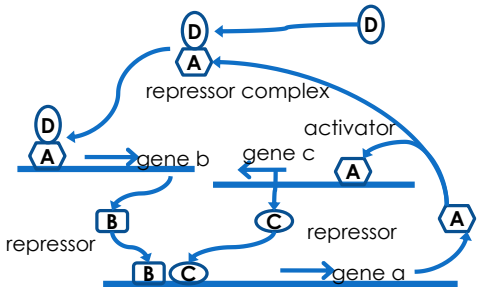


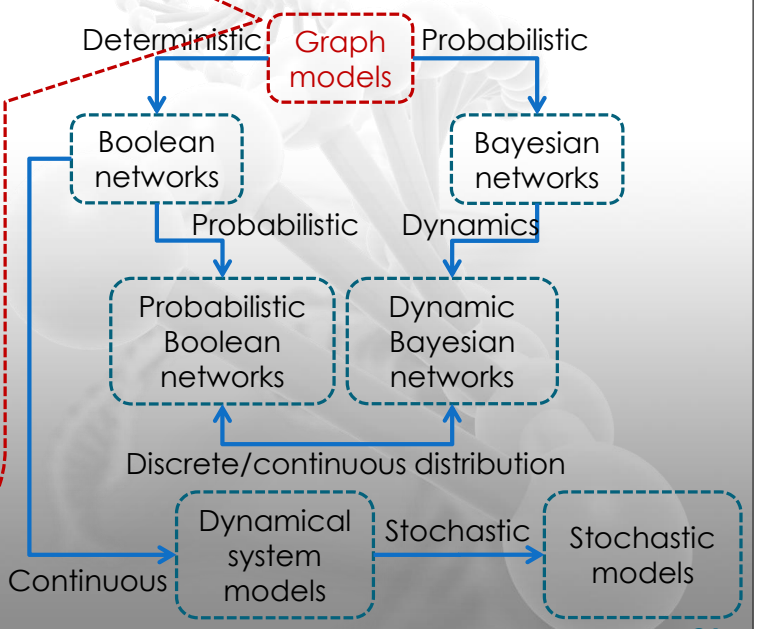


• Model Reconstruction

• Graph models

1. Correlation analysis
2. Mutual information (Basso et al., 2005)
3. Partial correlation (Castelo and Roverato, 2006)
4. Chi-square tests





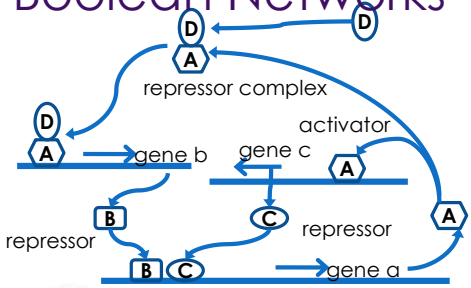
• Model Reconstruction

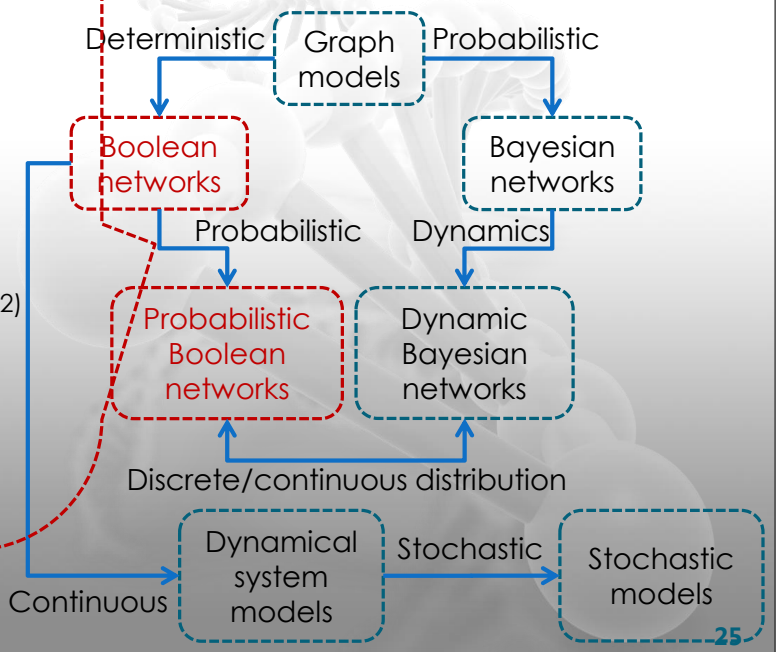
• Boolean Networks and Probability Boolean Networks

1. Quantization
2. Truth table
3. List all states which will result in specific node active at next time; {000,001,010,100,101,110}
4. Develop Boolean function

$$\bar{A}\bar{B}C + \bar{A}B\bar{C} + A\bar{B}\bar{C} + A\bar{B}C + AB\bar{C} = \bar{B}C$$

5. Probabilistic Boolean networks (Shmulevich et al., 2002)





• Model Reconstruction

• Bayesian networks and dynamic Bayesian networks

1. Define candidate graphs
2. Find the best conditional probabilities

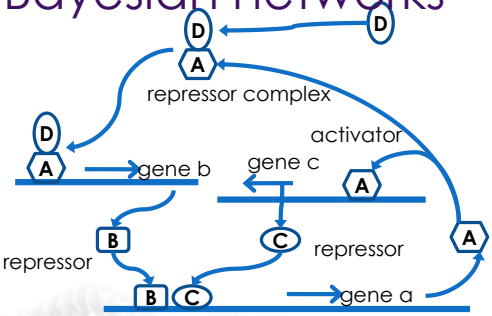
$$\max_{p(X_A|Pa(X_A)),p(X_B|Pa(X_B)),p(X_C|Pa(X_C))} \prod_i \prod_j p(X_i = x_{ij} \mid Pa(X_i) = \alpha_{ij})$$

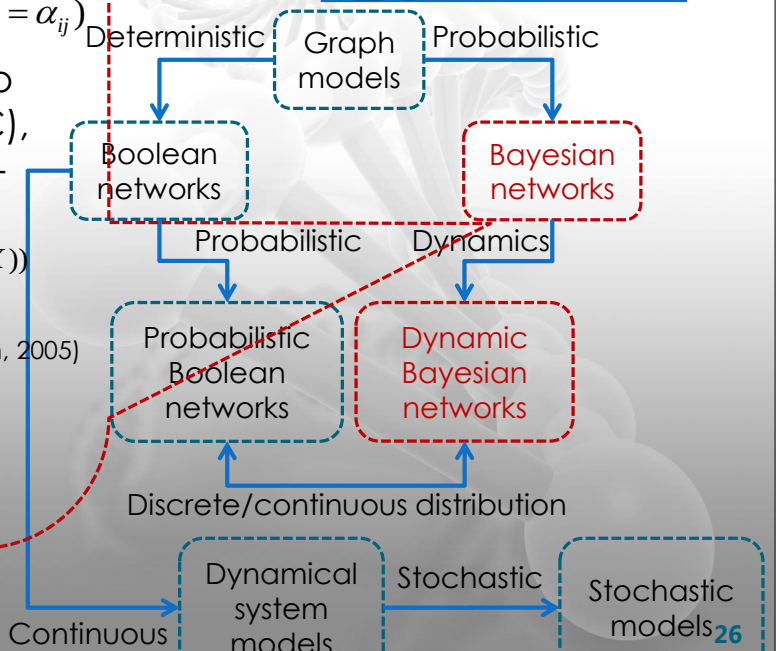
3. Choose the network according to Bayesian information criterion(BIC), which measures the goodness-of-fitting

$$\ln p(Pa(X) \mid X) \approx BIC = \ln p(X \mid Pa(X), \hat{\Theta}) - \frac{\ln N}{2}(\#Pa(X))$$

(Murphy and Mian, 1999)

4. DBN (more nodes) (Zou and Conzen, 2005)
   - Idea (Xing and Wu, 2006)
   - Accuracy
   - Higher order





• Model Reconstruction

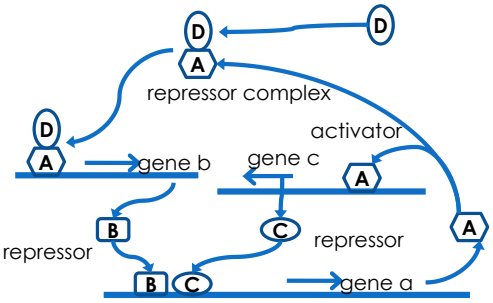
• Dynamical system models

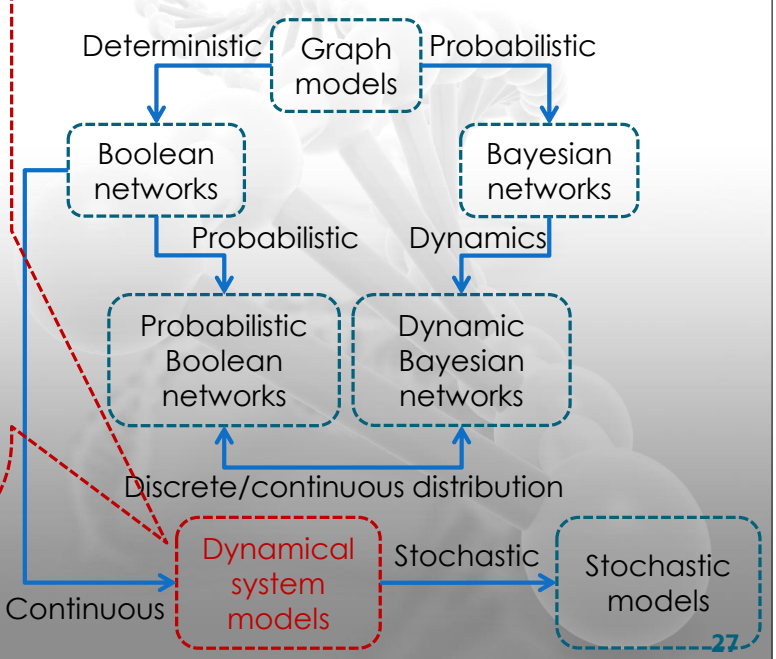
1. Estimate derivatives (Bansal et al., 2006)
2. Define regulation terms (Cooperative binding) (Yagil, 1971)

$$h(x,\theta,m) = \frac{x^m}{x^m + \theta^m}$$

m - cooperative TF binding
Θ – dissociation constant (half occupation)

3. Identify regulators (Song and Liu, 2007)
4. Optimize model parameters

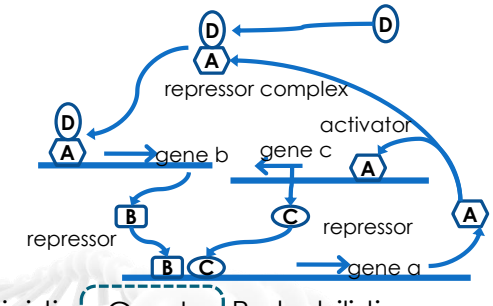


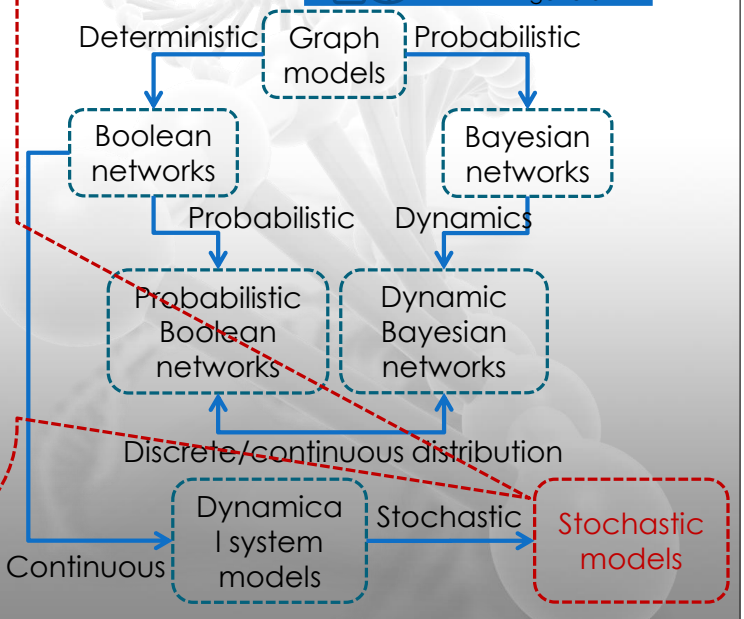


• Model Reconstruction

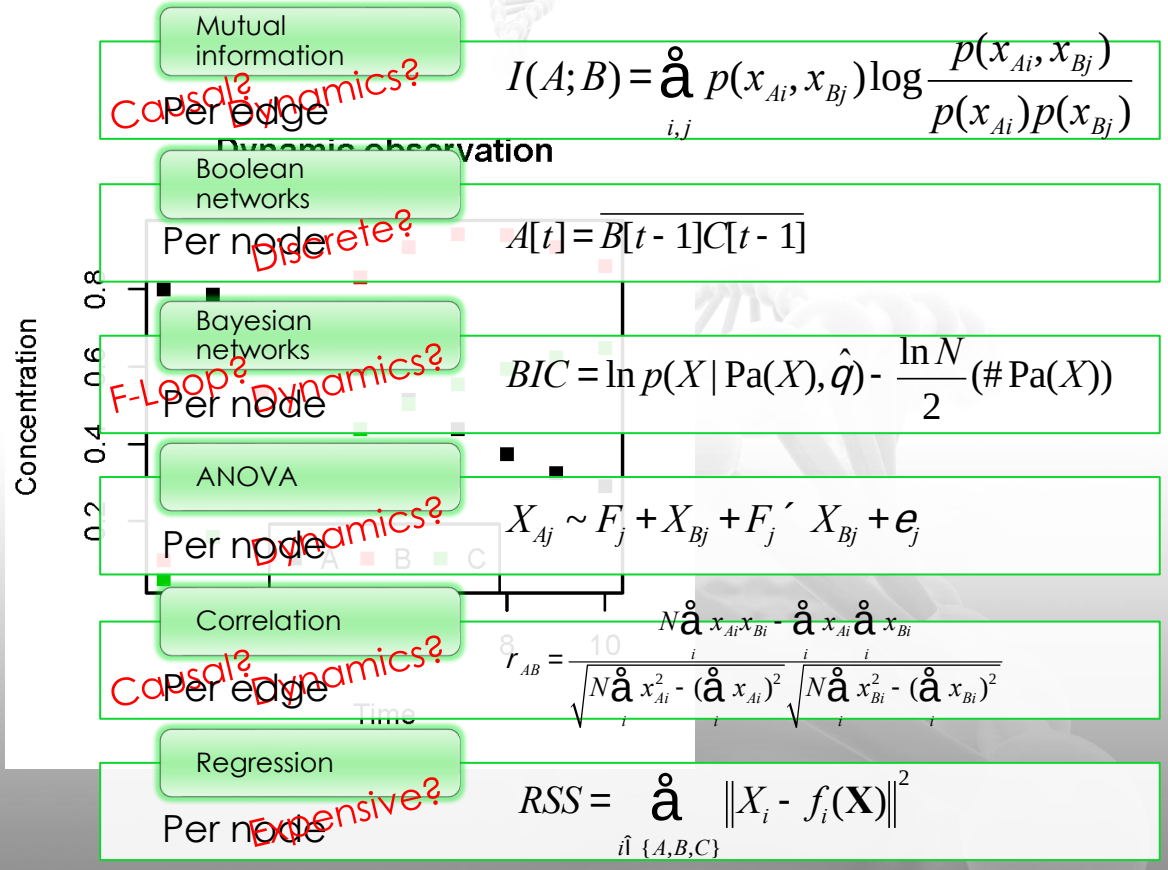
• Stochastic models

1. Simulation study (Higham, 2008)
2. Small discrete stochastic model reconstruction fill into probabilistic model reconstruction (Yu et al., 2004)
3. Small continuous stochastic model reconstruction by using sparse vector autoregressive model (Opgen-Rhein and Strimmer, 2007)





• Association and Causality



• Mutual information

$$I(X;Y) = \sum_x \sum_y p(X=x, Y=y) \log \frac{p(X=x, Y=y)}{p(X=x)p(Y=y)}$$

I(X;Y): mutual information between X and Y
X, Y: two random variables
p(X,Y): joint probability function between X and Y
p(X), p(Y): marginal probability function of X and Y

Can be used on both continuous and discrete data

Assumptions and limitations:

- Joint and marginal probability distribution functions have to be estimated—difficult on data of small sample sizes
- A statistic insensitive to sample size
- Variant: G-Test which derives a chi-square distribution for discrete data

• Entropy and conditional entropy

- Entropy—Level of uncertainty in X:

$$H(X) = -\sum_{x} p(x) \log p(x)$$

When the log is 2-based, H(X) has a unit of number of bits.

- Joint entropy—Level of joint uncertainty of X and Y:

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y)$$

- Conditional Entropy—Level of uncertainty of Y given X:

$H(Y|X) = H(X,Y) - H(X)$ [Chain rule; think conditional probability]

H(Y | X) is minimized to zero if and only if Y is a function of X.

• Properties of mutual information

- I(X; Y)
- Nonnegativity: I(X;Y)≥0
- Symmetry: I(X;Y)=I(Y;X)
- I(X;Y)=H(X)+H(Y)-H(X,Y)
- I(X;Y)= H(Y) - H(Y | X) = H(X) - H(X | Y)

• Coefficient of determination and correlation

Pearson correlation coefficient

x_i, y_i: The i-th observation of random variable X and Y

$$\rho_{XY} = \frac{N\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{N\sum_i x_i^2 - \left(\sum_i x_i\right)^2} \cdot \sqrt{N\sum_i y_i^2 - \left(\sum_i y_i\right)^2}}$$

N: Sample size

Coefficient of determination R^2

$$R^2 = \frac{Explained\ variance}{Total\ variance} = 1 - \frac{model\ residue}{total\ variance}$$

- For linear models with intercept, $R^2 = \rho_{XY}^2 \in [0,1]$
- For other models, $R^2 \in (-\infty, 1]$ (yes it can be negative)

• Linear models

- Correlation is based on linear models
- Linear models are the only choice at small sample sizes:
  - $Y = aX + b$
- In modern applications of large sample sizes, linear models carry high biases to miss patterns.

• Bayesian information criterion

$$BIC = \log L - \frac{\log N}{2} d$$

Likelihood: $L = \tilde{O}_i \, p(X = x_i \mid \mathrm{Pa}(X = x_i), \hat{q})$

- X – child
- Pa(X) – parents of X
- N – sample size
- d – number of parameters used in p(X | Pa(X))
- $\hat{\theta}$ – the estimated parameter of the model

• Bayesian information criterion

- Assumptions and limitations

- Likelihood based approaches may not be sensitive to interactions
- No asymptotic null distribution is available

• Analysis of variance (ANOVA)

$$Y \sim X_1 + X_2 + X_1 \times X_2$$
$$Y = \mu_i + \mu_j + \mu_{ij} + \epsilon$$

X_1, X_2: discrete parent variables
Y: continuous child variable
ϵ : noise term

μ_i: contribution to Y when $X_1 = i$
μ_j: contribution to Y when $X_2 = j$
μ_{ij}: contribution to Y when $X_1 = i, X_2 = j$

- Good for detecting potential causal relationships
- The most widely used statistical method
- The interaction term allows for complex nonlinear interactions from X_1, X_2 to Y

Assumptions and limitations:

- Linear additive assumption if no interaction terms $X_1 \times X_2$ are used
- Noise is assumed to be normally distributed
- Piece-wise constant model with Gaussian noise
- Always optimized if group means are all equal despite different distributions

• Conditional independence

- Definition: X and Y are called statistically independent conditioned on Z when their probability functions satisfy any of the equations:
  - p(X, Y | Z) = p(X | Z) p(Y | Z)
  - p(X | Y, Z) = p(X | Z)
  - p(Y | X, Z) = p(Y | Z)
- Examples
  - X → Z → Y
  - From the graph: P(X,Y,Z)=P(X)P(Z | X)P(Y | Z)=P(X,Z)P(Y | Z)=P(X | Z)P(Z)P(Y | Z)

    By definition: P(X,Y,Z)=P(X,Y | Z)P(Z)

    Then we have P(X | Z)P(Y | Z)=P(X,Y | Z).

    Therefore X and Y are independent given Z
  - Y → Z → X

•Example—Boys Scouts and Juvenile Delinquency

- Ref: https://onlinecourses.science.psu.edu/stat504/node/112/

This is a 3 × 2 × 2 table. It classifies *n* = 800 boys according to
S: socioeconomic status
B: whether they are a boy scout
D: whether they have been labeled as a juvenile delinquent

Socioeconomic status	Boy scout	Delinquent	
		Yes	No
Low	Yes	11	43
	No	42	169
Medium	Yes	14	104
	No	20	132
High	Yes	8	196
	No	2	59

Is delinquency D independent of boy scout status B given socioeconomic status S?
Equivalently, is p(D, B | S) = p(D | S) p(B | S)? B→S→D

• Are delinquency and boy scout status (unconditionally) dependent?

- Pearson chi-square test of independence

Test statistic: $\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$

Expected count in cell (i,j):

$$e_{ij} = \frac{(sum\ of\ row\ i)(sum\ of\ column\ j)}{sum\ of\ counts\ in\ all\ cells}$$

Observed count in cell (i,j): n_{ij}

Degrees of freedom: (#rows – 1)(#columns – 1)

Boy scout	Delinquent	
	Yes	No
Yes	33	343
No	64	360

We use the **Pearson's chi-square test** for this independence between *B* and *D*.

Expected counts are printed below the observed counts:
χ^2 = *3.477 + 0.480 + 3.083 + 0.425 = 7.465*, where each value in the sum is a contribution (squared Pearson residual) of each cell to the overall Pearson χ^2 statistic.

With degrees of freedom *df* = (2-1)*(2-1)=1,
p-value=1-pchisq(7.465,1)=0.006,
rejecting the marginal independence of *B and D*.

In R, use function chisq.test() to perform the test.

	Delinquent=Yes	Delinquent=No	Total
Boy Scout=Yes	33 45.59	343 330.41	376
Boy Scout=No	64 51.41	360 372.59	424
Total	97	703	800

• Test of conditional independence

Socioeconomic status	Boy scout	Delinquent	
		Yes	No
Low	Yes	11	43
	No	42	169
Medium	Yes	14	104
	No	20	132
High	Yes	8	196
	No	2	59

$(\chi^2 \mid \text{S=Low}) = 0.0058$ df = 1

$(\chi^2 \mid \text{S=Medium}) = 0.101$ df = 1

$(\chi^2 \mid \text{S=High}) = 0.0534$ df = 1

To test the conditional independence of (BS, DS) we can add these up to get the overall chi-squared statistic:

$$0.0058 + 0.101 + 0.0534 = 0.160.$$

Each of the individual tests has 1 degree of freedom, so the total number of degrees of freedom is 3. The *p*-value is [illegible], indicating that the conditional independence model fits extremely well. As a result, we will not reject this model here.

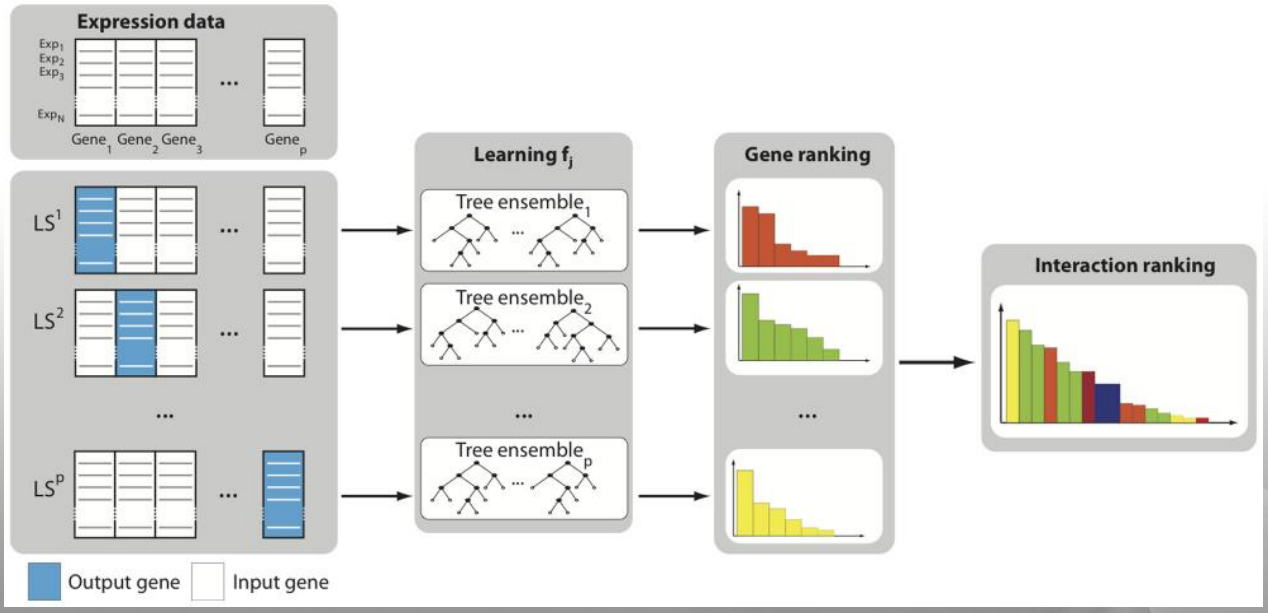
• Handling high dimensionality

- Penalized linear regression
  - Ridge regression, penalizing on the sum of coefficient squares
  - Lasso, penalizing on the sum of absolute values of coefficient
  - Elastic-net, penalizing on both sums
- Tree-based nonlinear regression
  - Decision trees
  - Random forests. E.g. GENIE3
  - Gradient boosting machines. E.g. GRNBoost

• GENIE3

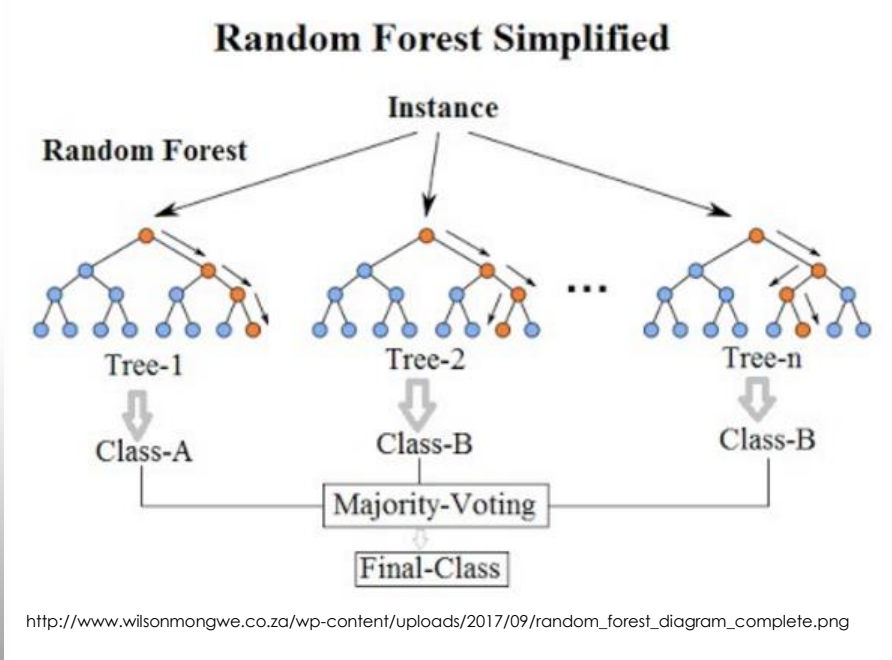
• GEne Network Inference with Ensemble of trees

- GENIE3 employs tree-based ensemble (Random Forest) nonlinear regressions
- Each gene has a predictive model based on parent (regulatory) genes (transcription factors).
- The regression models act as a feature selection mechanism.



References: Huynh-Thu V, Irrthum A, Wehenkel L, Geurts P (2010). "Inferring regulatory networks from expression data using tree-based methods." *PLoS ONE*, **5**(9), e12776. doi: 10.1371/journal.pone.0012776.

• Random forests



http://www.wilsonmongwe.co.za/wp-content/uploads/2017/09/random_forest_diagram_complete.png

• GRNBoost

- Motivation: to reduce runtime and enhance data size capability
- Using gradient-boosting machines (GBM) xgboost.
  - GBM: an ensemble learning algorithm that uses boosting to combine multiple weak learners, like shallow trees, into a strong one.
  - Random forests (used by GENIE3) use bagging (bootstrap aggregation) for model averaging to improve regression accuracy.
- It uses gradient-boosted stumps (regression trees of depth 1) as the base learner.
- It casts multiple regression into a Map/Reduce framework based on Apache Spark.
- Available on Github: https://github.com/aertslab/GRNBoost
- Published as part of SCENIC

• Gradient boosting

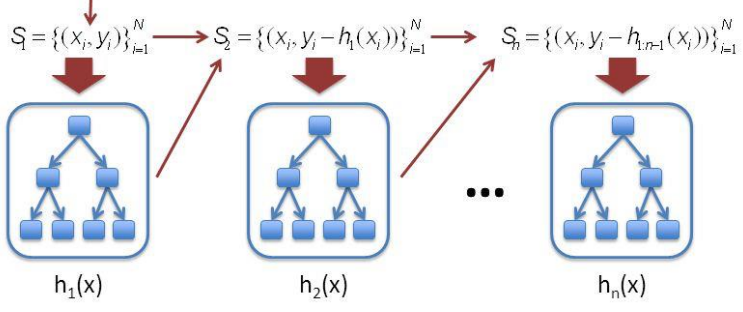
Gradient Boosting (Simple Version)

(Why is it called "gradient"?)
(Answer next slides.)

(For Regression Only)

$S = \{(x_i, y_i)\}_{i=1}^{N}$

$h(x) = h_1(x) + h_2(x) + \ldots + h_n(x)$



http://statweb.stanford.edu/~jhf/ftp/trebst.pdf

https://dimensionless.in/wp-content/uploads/2016/12/slide_24.jpg

24

• The GRNBoost algorithm

Input: Gene expression matrix
Output: Regulators for each target gene

1. Partition gene expression vectors over nodes in the compute cluster, using a Spark RDD.
2. Construct a predictor matrix of expression values for all **candidate** regulator genes.
3. Broadcast the predictor matrix to different compute partitions, using a Spark broadcast variable.
4. Map phase:
   a) Iterate over gene expression vectors
   b) Use the predictor matrix to train the XGBoost regression models with vectors as training labels.
   c) Extract, from the trained models, strengths of regulator–target relationships and emit as a set of gene network edges.
5. Reduce phase: combine all sets of edges into a final gene network.

• Summary

- Reaching a plateau in applications
  - Insufficiently perturbed data
  - High dimension but small sample sizes (10—1,000)
- A sudden demand from single-cell experiments
  - Dynamic can be rich
  - Sample size can be large (100,000-1,000,000)
- Opportunities
  - Fundamentally new statistics, such as FunChisq
  - Fast methods for finding parents from high dimension and large sample size data
  - New applications (various single-cell data: scRNA-seq, sc cytomerty, etc)