Introduction to Network Modelling in Genomics

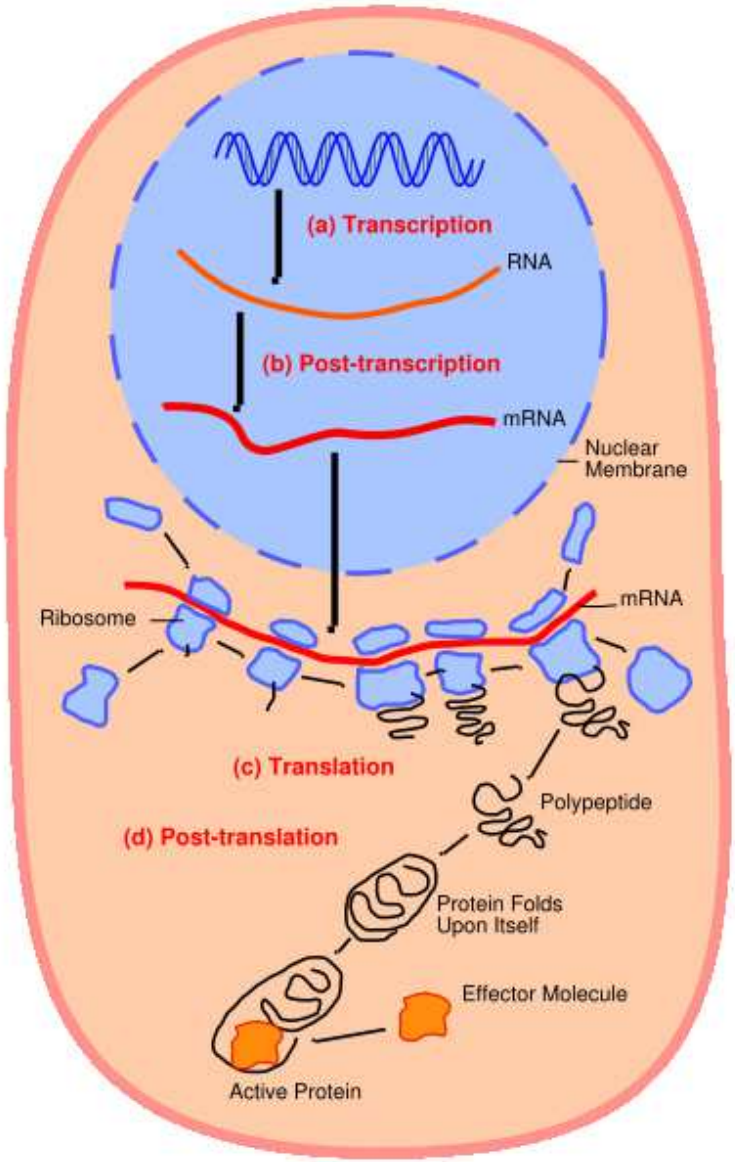
Ernst Wit
University of Groningen

e.c.wit@rug.nl
http://www.math.rug.nl/∼ernst

21 January 2014

rijksuniversiteit groningen

How does the genome operate?



Every cell within every organism contains its full genetic programme:

its genome or DNA

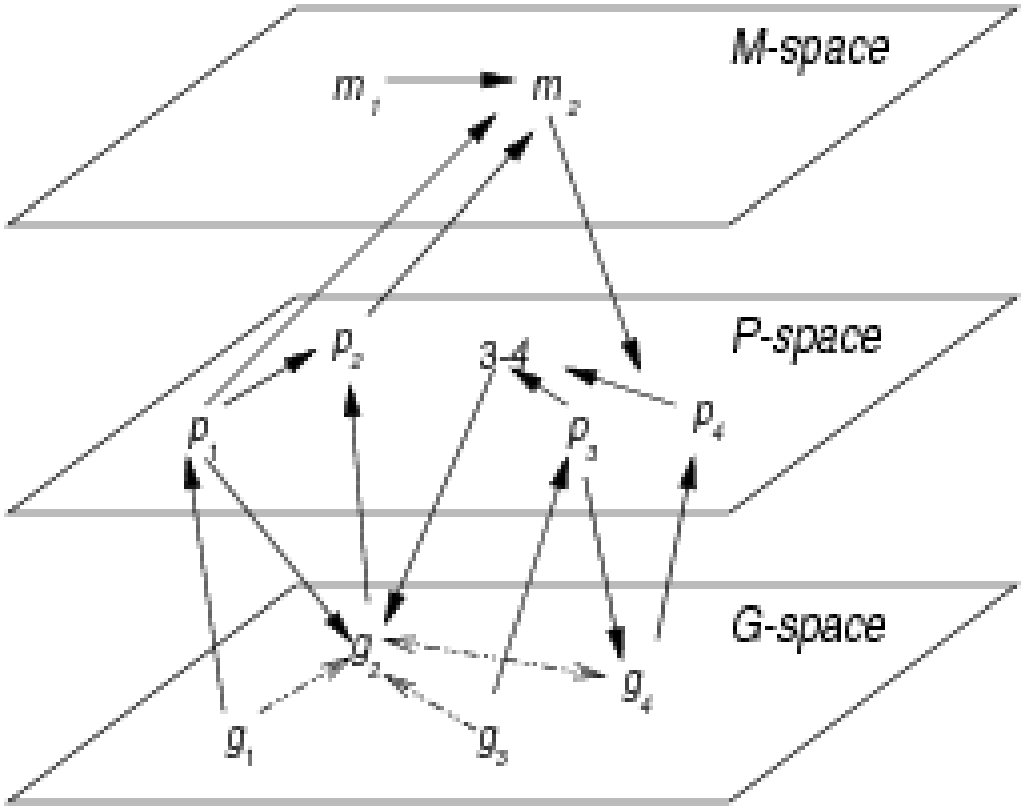
Depending on the internal and external conditions of that cell, it will activate certain genes more or less.

If a particular gene is needed, its DNA will copied/transcribed into RNA.

The amount of RNA of a gene is called **gene expression**

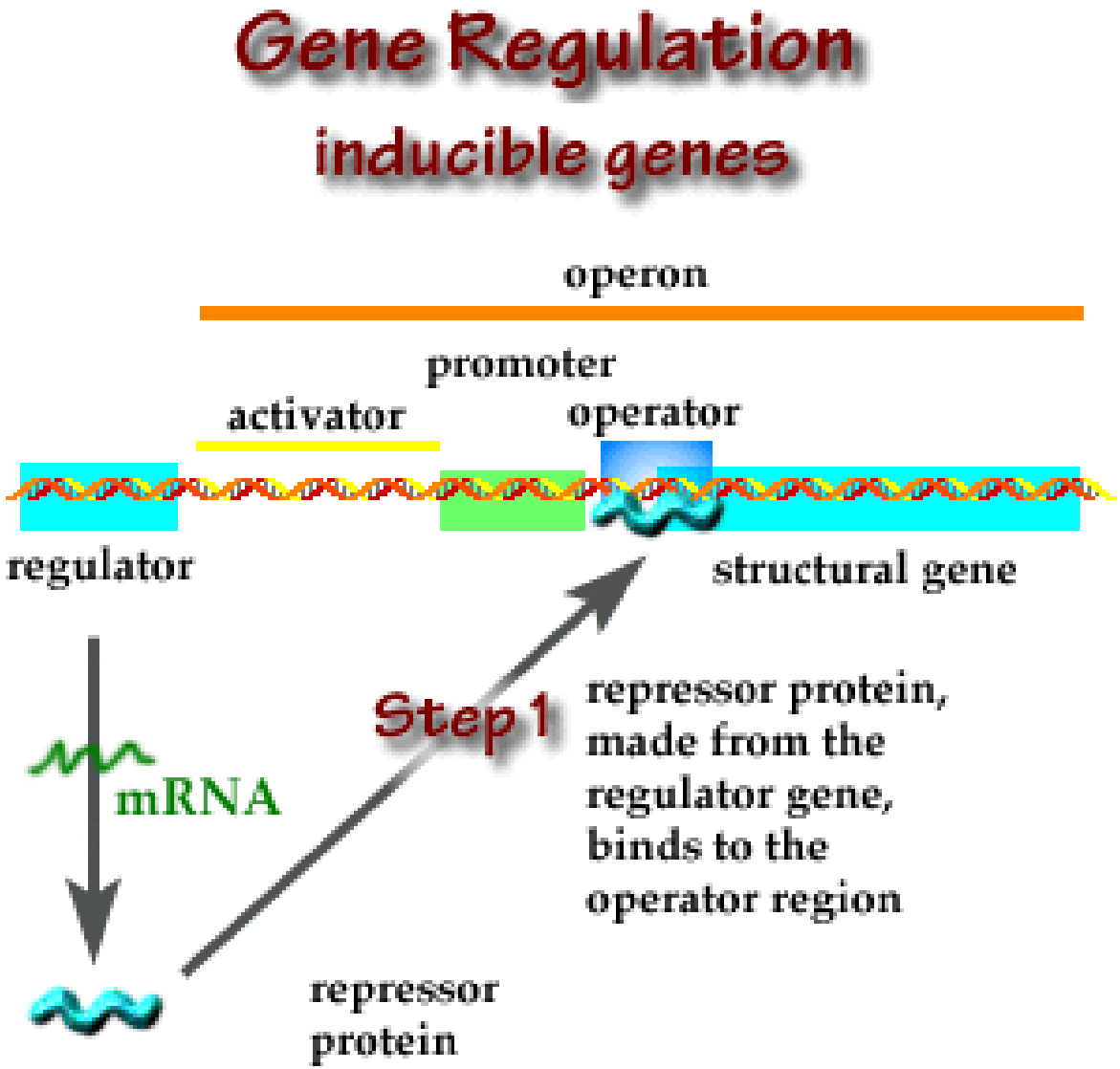
rijksuniversiteit groningen

What is a gene regulatory network?

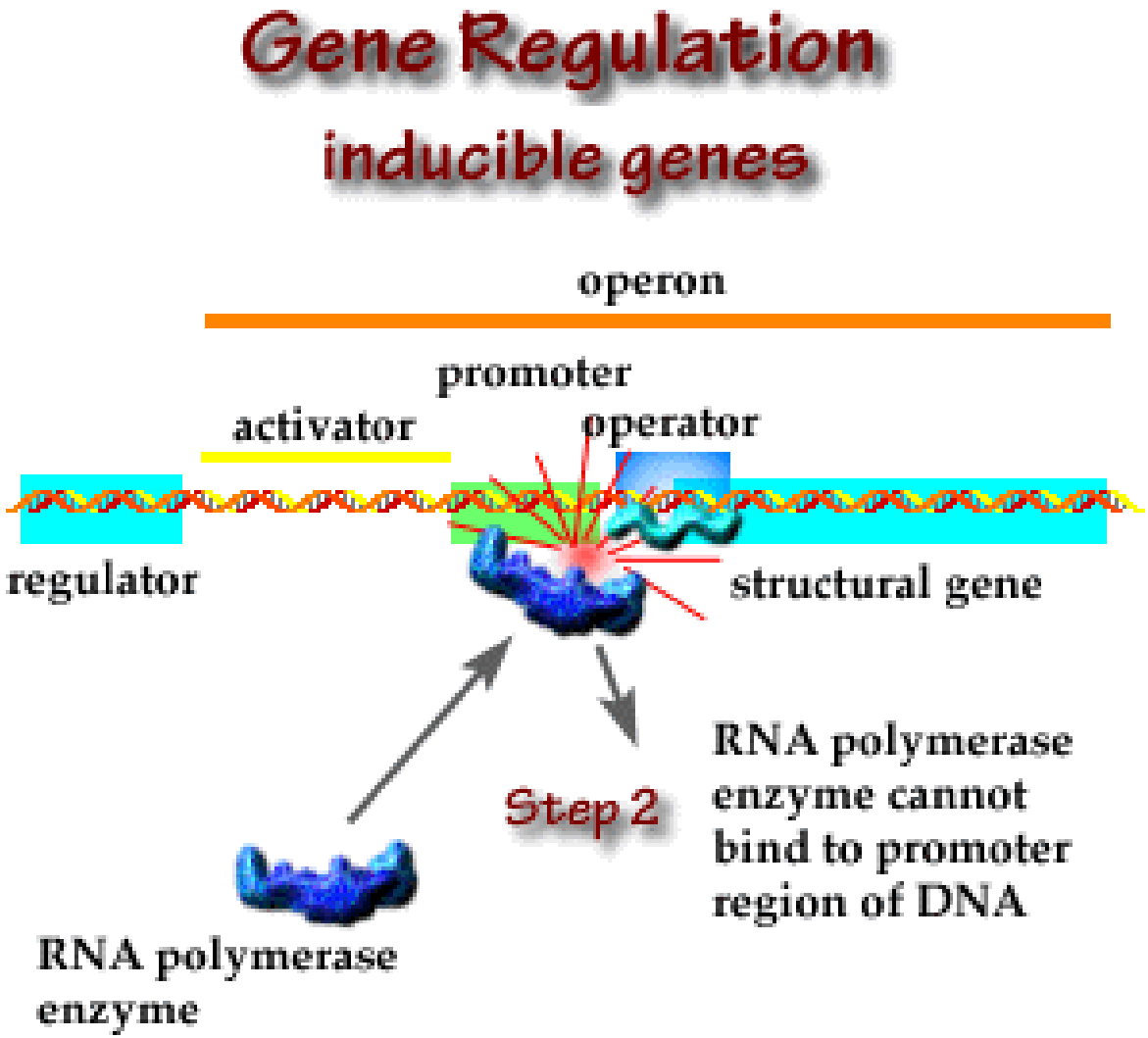


"A collection of DNA segments in a cell which "interact" with each other via their RNA or proteins and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed."

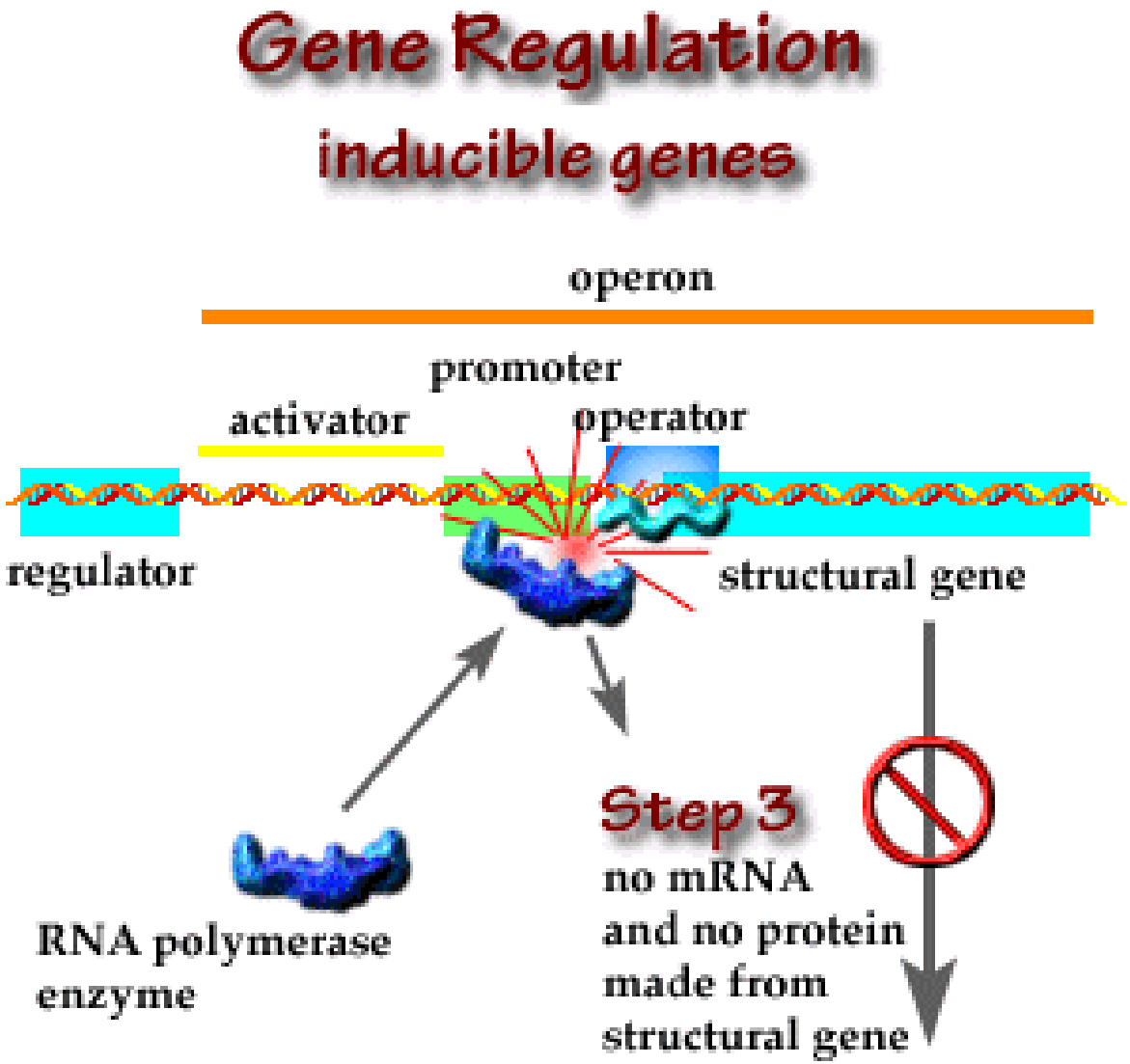
Gene regulation



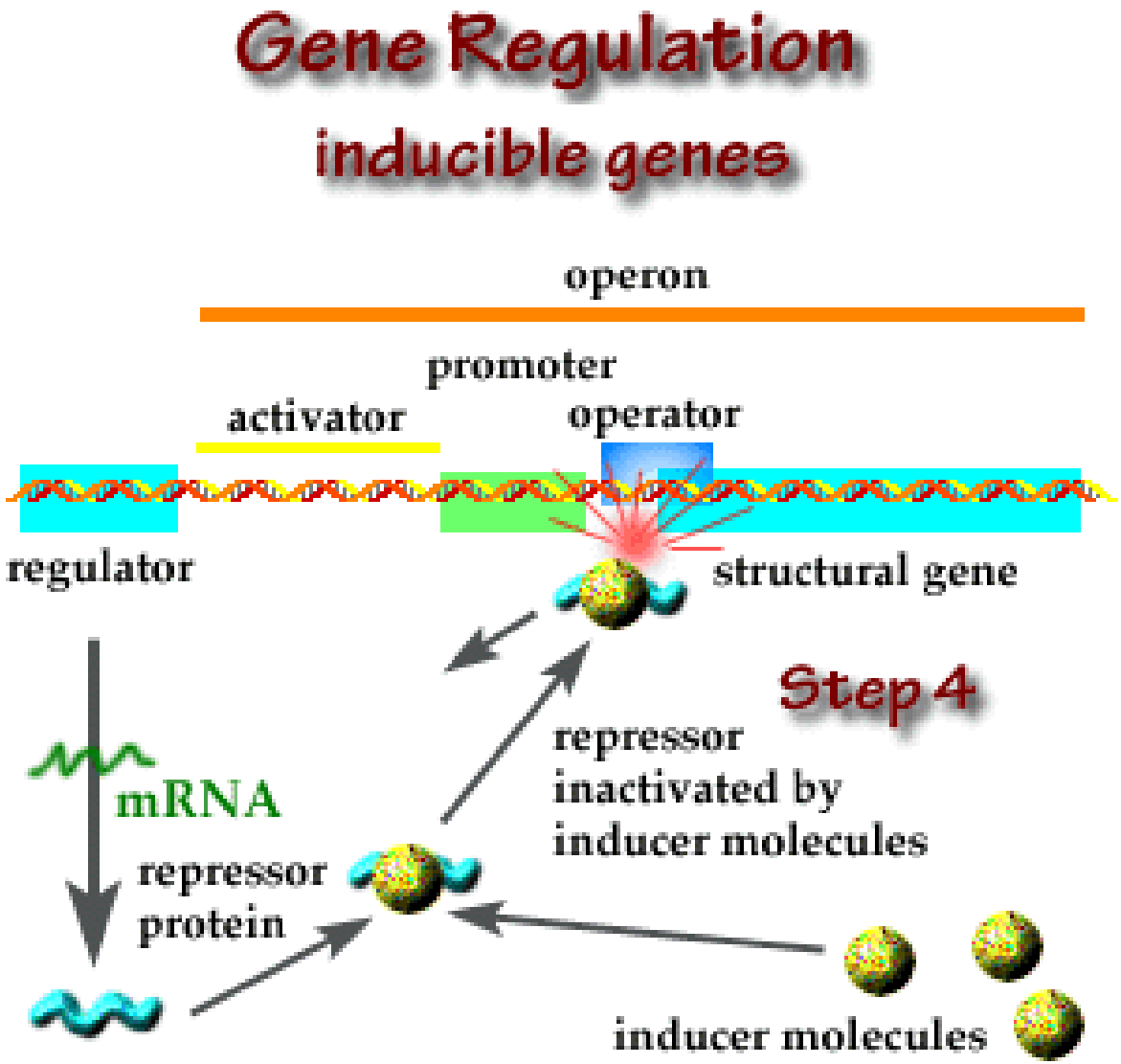
Gene regulation



Gene regulation

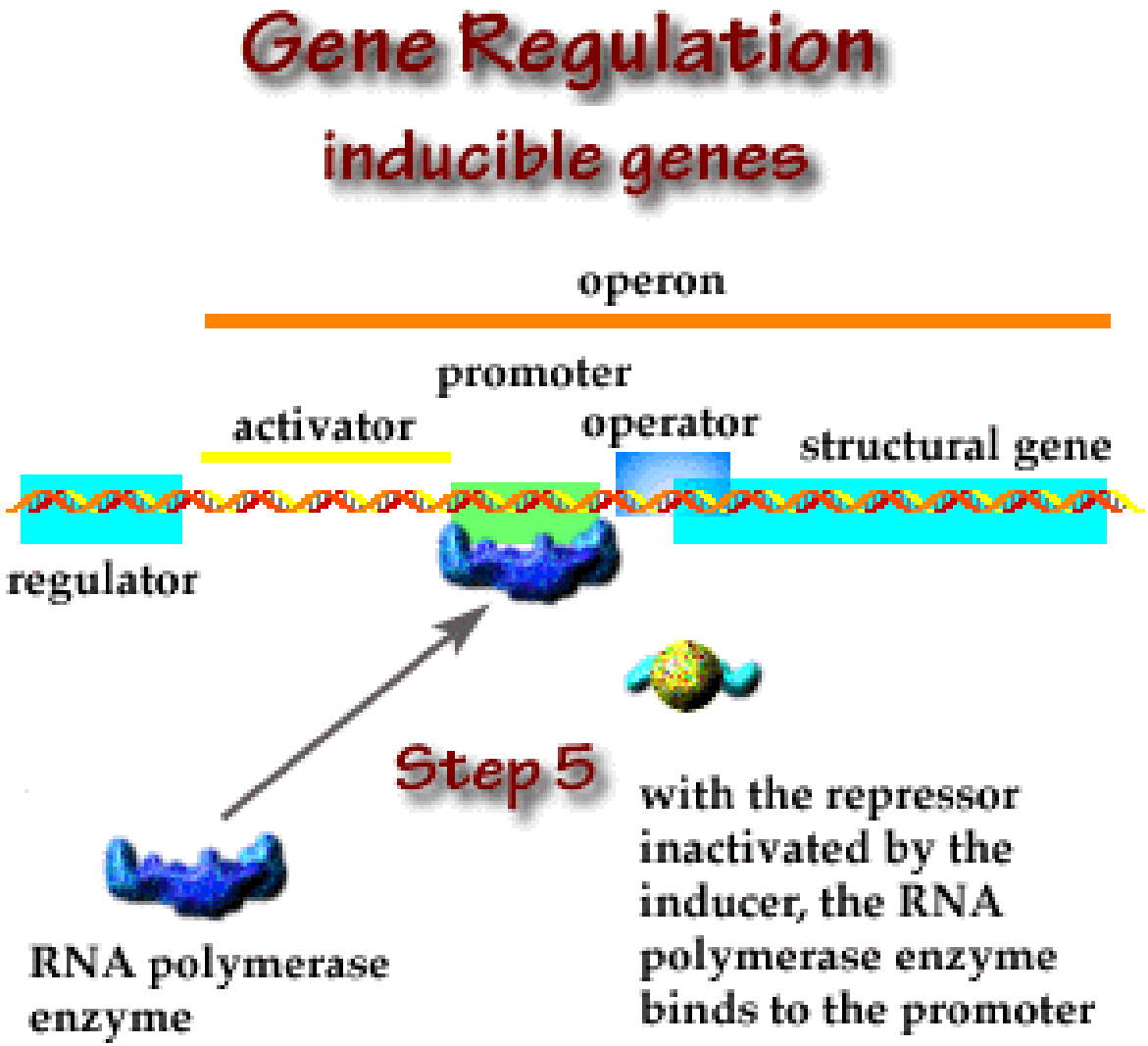


Gene regulation



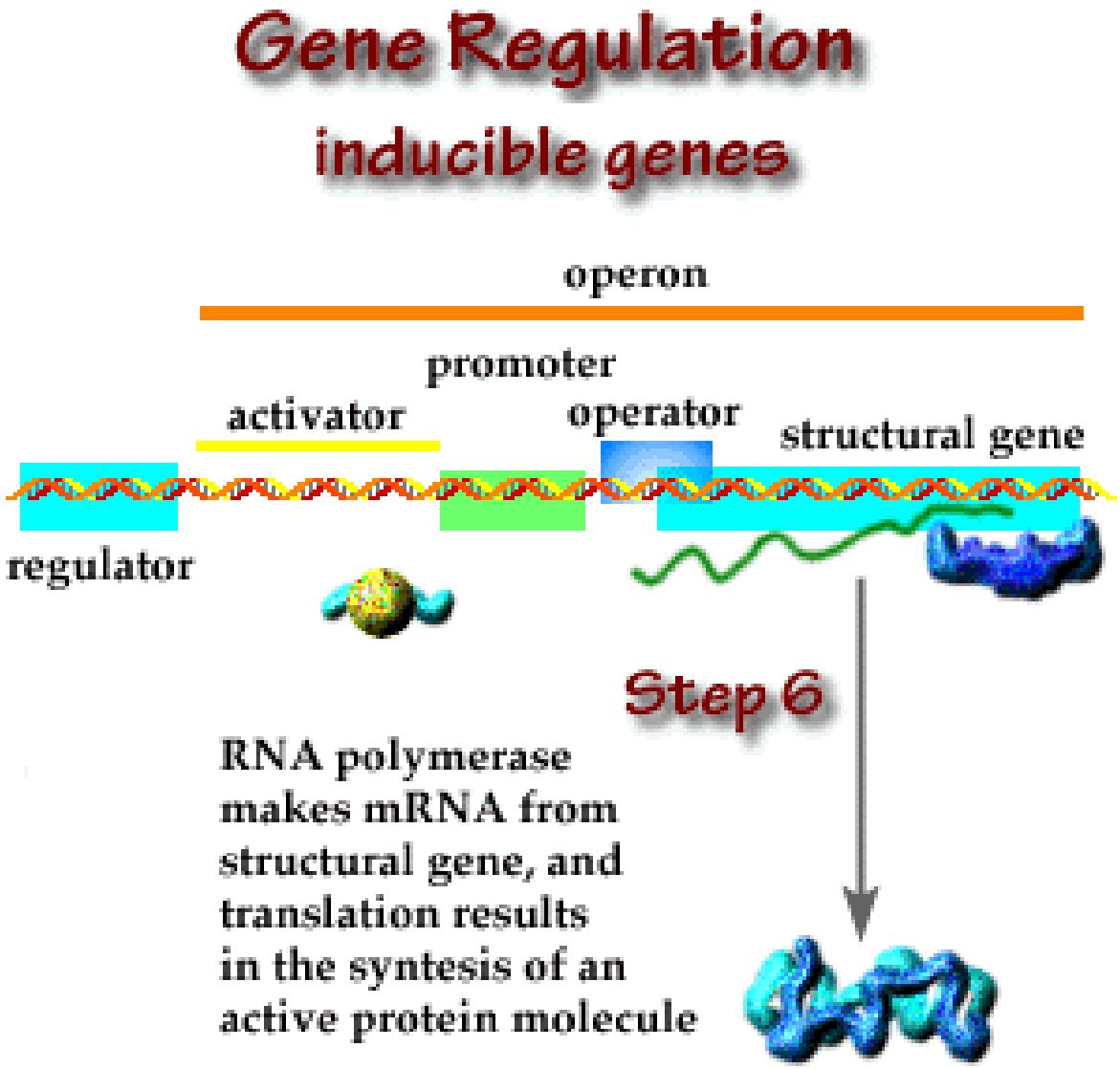
rijksuniversiteit groningen

Gene regulation



rijksuniversiteit groningen

Gene regulation

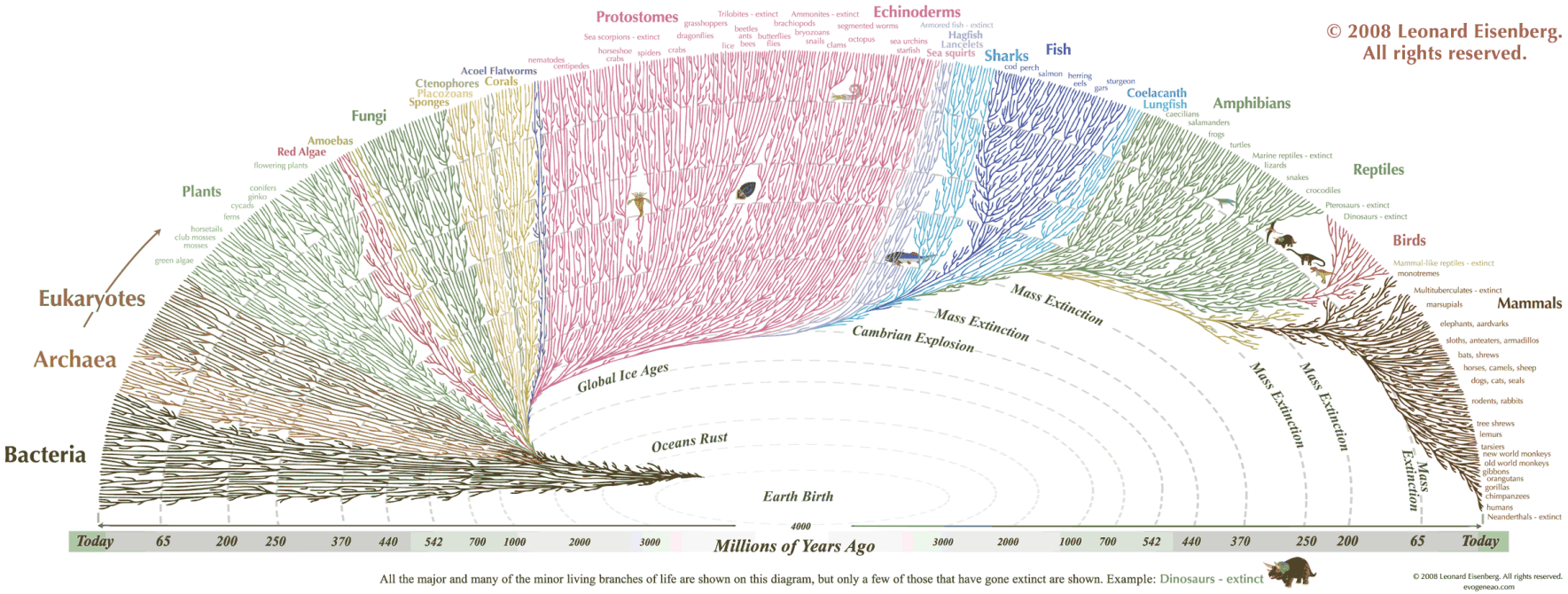


What do we want to know about the genetic network?

1. Evolution and history
2. Overall topology
3. Links and link strengths
4. Dynamics (next lecture)

1. Evolution and history

- Within-species, between-species
- Branching process: tree structure
- Question: Most recent common ancestor?
- Methods: clustering, hierarchical trees, coalescent trees.



rijksuniversiteit groningen

2. “Topology”

- Clustering coefficient
- Characteristic path length, Diameter
- Degree distribution
- Hubs, small world, scale-free
- Robustness

a. Degree distribution

Degree/connectivity: number of interactions of node in network.

Two kinds of connectivity:

- k_{in}: *incoming/arriving connectivity* or *in-degree*
- k_{out}: *outgoing/departing connectivity* or *out-degree*

Degree distribution:
(empirical) distribution $p(k)$ of degree size k within a single network, e.g.,

- random,
- scale-free
- hierarchical.

Two types of random networks

Erdos-Renyi network

- taking $\binom{N}{2}$ draws from a Bernoulli(p) distribution.
- draw a link between jth pair if j draw was a success.

Both in- and out-degree distribution of each node modelled as

$$\text{Binomial}(N-1, p).$$

Scale-free network

- At iteration K, sample among the $\binom{N}{2} - K$ remaining links.
- Sampling probability proportional to number of links at receiving link.

Resulting in-degree has power-law distribution with exponent γ:

$$p_k \propto k^{-\gamma}$$

b. Clustering Coefficient: inter-connectivity of node

Fraction of existing links between node's neighbours:

$$C_i = \frac{2e_i}{k_i(k_i - 1)}, \quad i = 1, \ldots, N$$

- e_i: total number of links between neighbours of node i.
- k_i: the degree, i.e. number of links, of ith node

Average Clustering Coefficient (C):

$$C = \frac{\sum_{i=1}^{N} 2e_i/(k_i(k_i - 1))}{N}.$$

rijksuniversiteit groningen

c. Characteristic path length and diameter

Let d_{ij}: the shortest path length between the ith and the jth node

Characteristic path length:

$$L = \frac{2\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}}{N(N-1)}$$

Diameter:

longest distance among all path lengths in a system,

$$D = \max\{d_{ij}\}(i, j \in N).$$

Characteristic path length and diameter

- **Scale-free network**: small L
- **Random network**: no highly connected nodes, resulting in $L \approx C$.

Information of L, C and γ can be used to understand whether the network has **small-world** or **ultra-small world** property:

- *Small-world property*: small L, large C and power exponent term $\gamma > 3$.
- *Ultra small-world property*: even smaller L and $2 < \gamma < 3$ (also indicative of modular structure in system).

d. Existence of hubs and robustness

In many (biological) networks:
many genes have few connections, few genes have many connections.

Hubs: highly connected nodes in system, often *global regulators*.

- Hubs are typical in scale-free networks.
- Hubs are atypical in random networks.

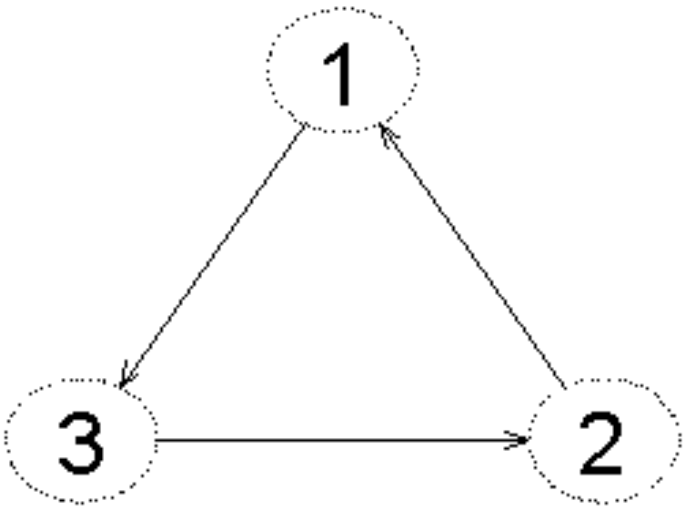
Network Robustness:
invariance of network from random removal of nodes/links.

- scale-free networks are robust, except from removal of hubs.

These few connections among hubs is known as **centrality principle**.
Their ability to control the whole system is called **lethality principle**.

rijksuniversiteit groningen

3. Links



Genetic network

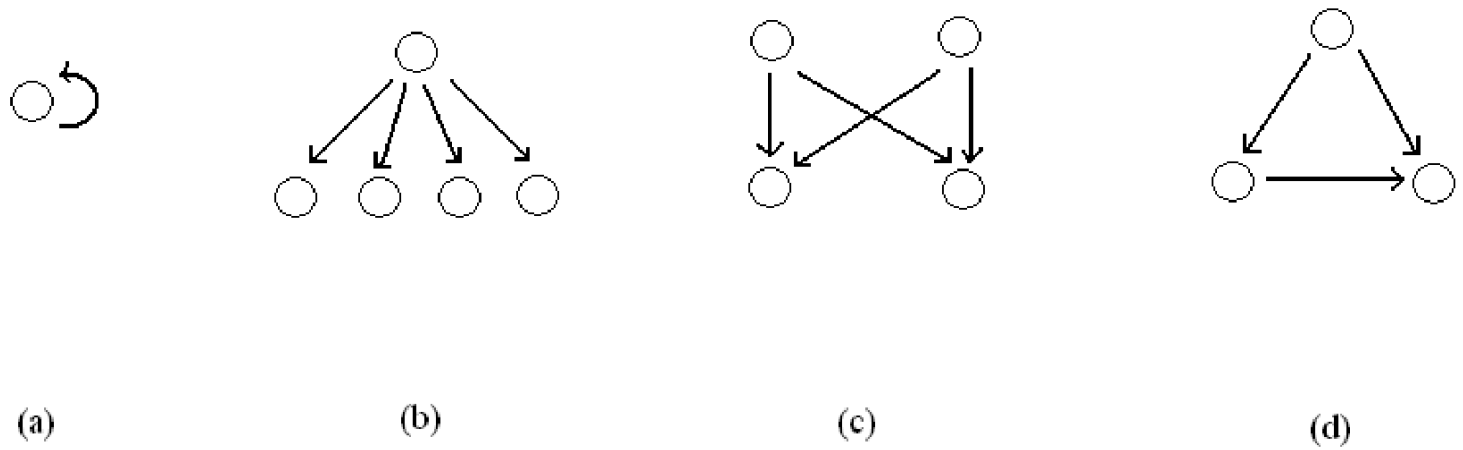
$$X = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Adjacency matrix

The aim of this network analysis is to infer the **adjacency matrix**.

Motif

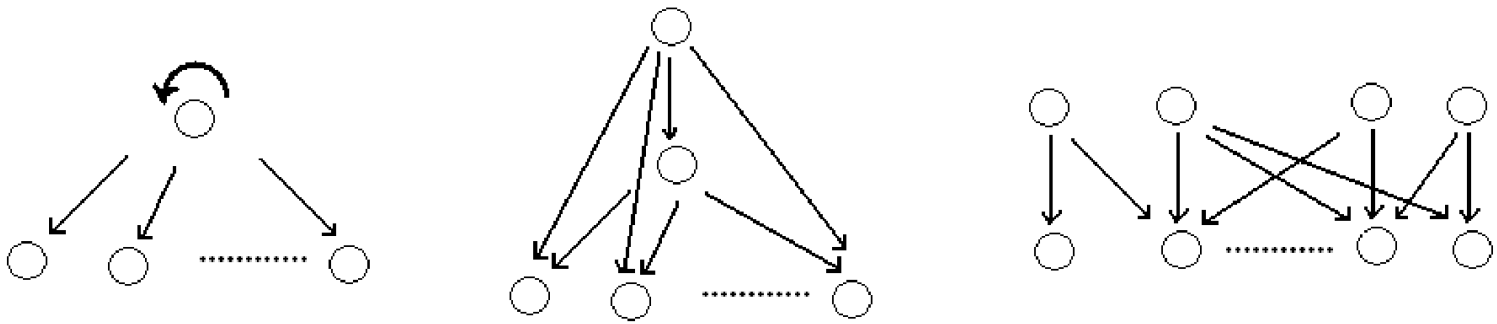
A local sub-structure between TFs and target genes that is observed commonly in biological systems.



(a) auto-regulation (b) single input (c) multiple input (d) feed-forward loop.

Module

Modules are composed of interconnecting motifs, that work together to perform a specific action in the system.



(a) (b) (c)

(a) single input module (b) multi-output FFL (c) dense overlapping regulons.

Modularity coefficient

Idea:
Observed fraction of links is compared to mean fraction of the connections under "independence".

Evolutionary argument:

Evolution has worked towards modularity by maintaining dense connections within modules and sparse external connections between modules.

Objective function:
plausible modularity of system is structure that *minimizes MC*.

Definition of Modularity coefficient

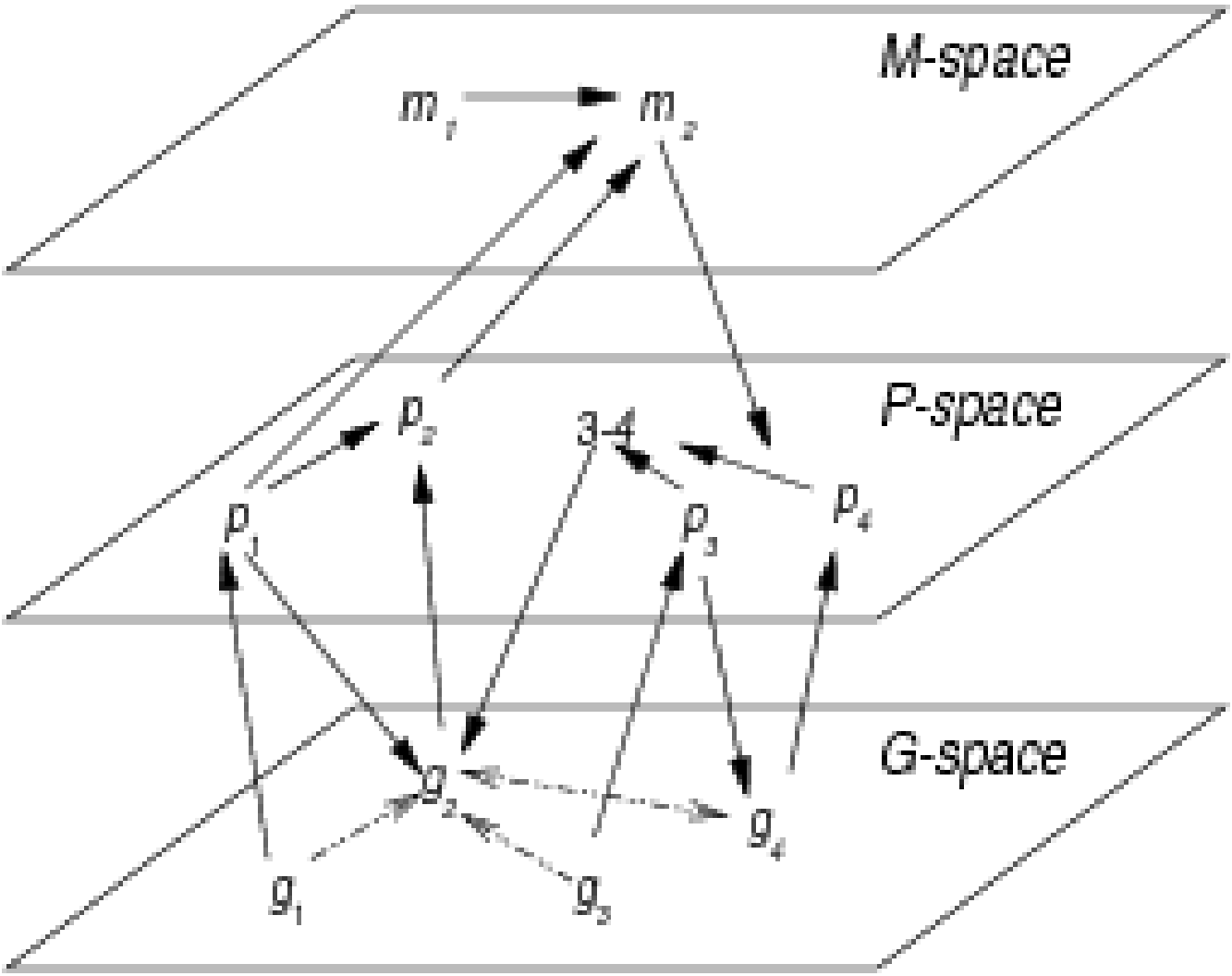
MC = difference between observed and "expected" links

Modularity coefficient

$$MC = \sum_{j=1}^{N_m} \sum_{i} \left(\frac{k_{ij}}{K} - \left(\frac{d_j}{2K} \right) \right)^2$$

- k_{ij}: number of links of ith node in jth module
- d_j: sum of degrees within jth module, i.e. $d_j = \sum_i k_{ij}$.
- K: total number of links in system, $K = \sum_j d_j/2$.
- N_m: the number of modules in the system

Interpret Network as Separation Statements



Graphoids: $A \perp_\sigma B | C$ = "C separates A and B"

An *independence model* $\perp_\sigma$ is a ternary relation over subsets of a finite set V. It is *semi-graphoid* if for all subsets A, B, C, D:

(S1) if $A \perp_\sigma B \,|\, C$ then $B \perp_\sigma A \,|\, C$ (symmetry);

(S2) if $A \perp_\sigma (B \cup D) \,|\, C$ then $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, C$ (decomposition);

(S3) if $A \perp_\sigma (B \cup D) \,|\, C$ then $A \perp_\sigma B \,|\, (C \cup D)$ (weak union);

(S4) if $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, (B \cup C)$, then $A \perp_\sigma (B \cup D) \,|\, C$ (contraction).

It is a *graphoid* if (S1)–(S4) holds and

(S5) if $A \perp_\sigma B \,|\, (C \cup D)$ and $A \perp_\sigma C \,|\, (B \cup D)$ then $A \perp_\sigma (B \cup C) \,|\, D$ (intersection).

It is *compositional* if also

(S6) if $A \perp_\sigma B \,|\, C$ and $A \perp_\sigma D \,|\, C$ then $A \perp_\sigma (B \cup D) \,|\, C$ (composition).

niversiteit
igen

Admissions to Berkeley

Here are three variables A: Admitted?, S: Sex, and D: Department.

Department	Sex	Whether admitted	
		Yes	No
I	Male	512	313
	Female	89	19
II	Male	353	207
	Female	17	8
III	Male	120	205
	Female	202	391
IV	Male	138	279
	Female	131	244
V	Male	53	138
	Female	94	299
VI	Male	22	351
	Female	24	317

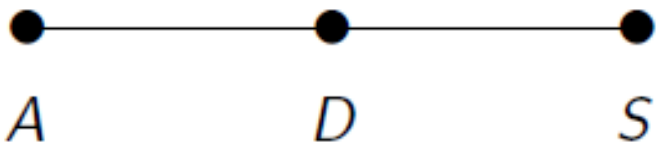
When dealing with complex systems of many random variables, we must have a concept which is more sophisticated, but equally fundamental: that of *conditional independence*.

rijksuniversiteit groningen

Conditional independence

For three variables it is of interest to see whether independence holds for fixed value of one of them, e.g. *is the admission independent of sex for every department separately?*

We denote this as $A \perp\!\!\!\perp S \mid D$ and display it graphically as



Algebraically, this corresponds to the relations

$$p_{ijk} = p_{i+|k}\, p_{+j|k}\, p_{++k} = \frac{p_{i+k}\, p_{+jk}}{p_{++k}}.$$

Florida murders

Sentences in 4863 murder cases in Florida over the six years 1973-78

	Sentence	
Murderer	Death	Other
Black	59	2547
White	72	2185

The table shows a greater proportion of white murderers receiving death sentence than black (3.2% vs. 2.3%), although the difference is not big, the picture seems clear.

Controlling for colour of victim

		Sentence	
Victim	Murderer	Death	Other
Black	Black	11	2309
	White	0	111
White	Black	48	238
	White	72	2074

Now the table for given colour of victim shows a very different picture. In particular, note that 111 white murderers killed black victims and none were sentenced to death.

Fundamental properties of conditional independence

For random variables X, Y, Z, and W it holds

(C1) If $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$;

(C2) If $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \mid Z$;

(C3) If $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;

(C4) If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \mid Z$;

If density w.r.t. product measure $f(x, y, z, w) > 0$ also

(C5) If $X \perp\!\!\!\perp Y \mid (Z, W)$ and $X \perp\!\!\!\perp Z \mid (Y, W)$ then $X \perp\!\!\!\perp (Y, Z) \mid W$.

rijksuniversiteit groningen

Definition of Graphical Model through Markov properties

Let $\mathcal{G} = (V, E)$ simple undirected graph and let $\perp_\sigma$ be an independence model. We say $\perp_\sigma$ satisfies

(P) *the pairwise Markov property* w.r.t. $\mathcal{G}$ if

$$\alpha \not\sim \beta \Rightarrow \alpha \perp_\sigma \beta \,|\, V \setminus \{\alpha, \beta\};$$

(L) *the local Markov property* w.r.t. $\mathcal{G}$ if

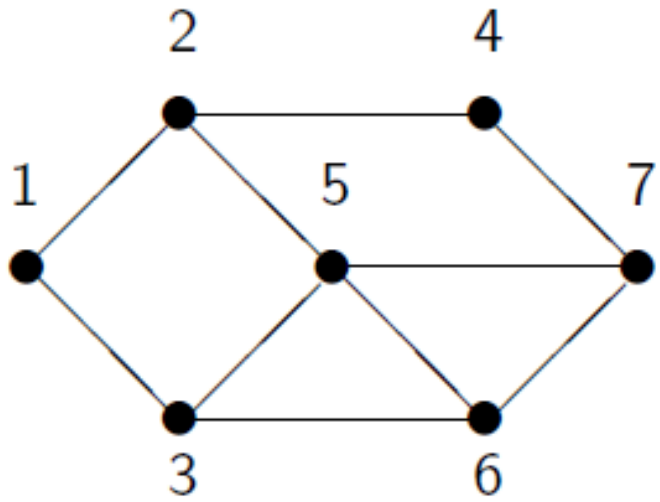
$$\forall \alpha \in V : \alpha \perp_\sigma V \setminus \mathrm{cl}(\alpha) \,|\, \mathrm{bd}(\alpha);$$

(G) *the global Markov property* w.r.t. $\mathcal{G}$ if

$$A \perp_{\mathcal{G}} B \,|\, S \Rightarrow A \perp_\sigma B \,|\, S.$$

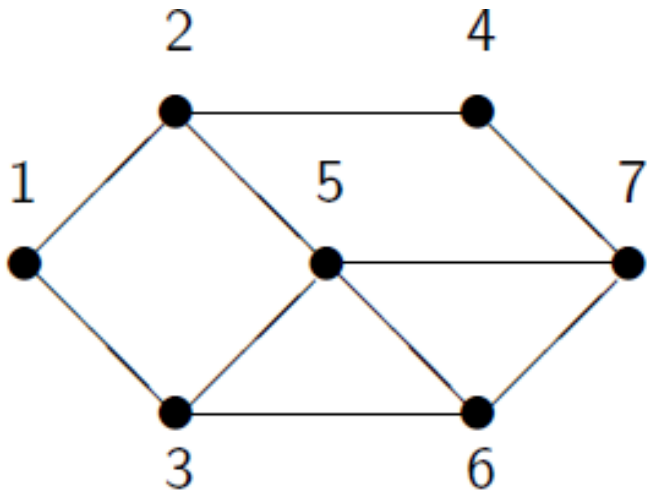
rijksuniversiteit groningen

Pairwise Markov property



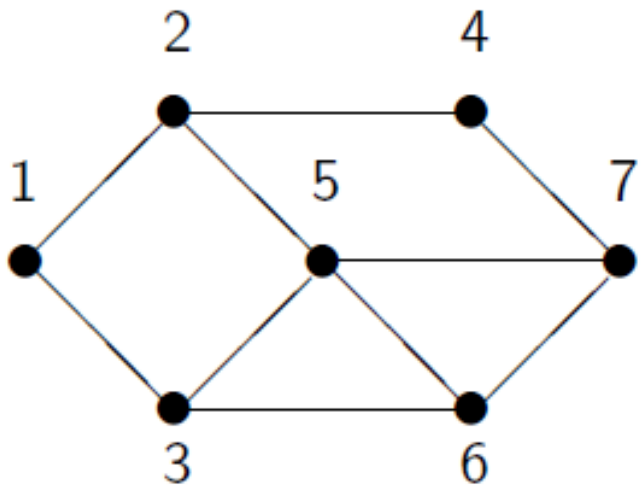
Any non-adjacent pair of random variables are conditionally independent given the remaning.
For example, $1 \perp_\sigma 5 \,|\, \{2,3,4,6,7\}$ and $4 \perp_\sigma 6 \,|\, \{1,2,3,5,7\}$.

Local Markov property



Every variable is conditionally independent of the remaining, given its neighbours.
For example, $5 \perp_\sigma \{1,4\} \,|\, \{2,3,6,7\}$ and $7 \perp_\sigma \{1,2,3\} \,|\, \{4,5,6\}$.

Global Markov property



To find conditional independence relations, one should look for separating sets, such as $\{2,3\}$, $\{4,5,6\}$, or $\{2,5,6\}$
For example, it follows that $1 \perp_\sigma 7 \,|\, \{2,5,6\}$ and $2 \perp_\sigma 6 \,|\, \{3,4,5\}$.

Relations among Markov Properties

For any semigraphoid it holds that

$$(\mathsf{G}) \Rightarrow (\mathsf{L}) \Rightarrow (\mathsf{P})$$

If $\perp_\sigma$ *satisfies graphoid axioms* it further holds that

$$(\mathsf{P}) \Rightarrow (\mathsf{G})$$

so that *in the graphoid case*

$$(\mathsf{G}) \iff (\mathsf{L}) \iff (\mathsf{P}).$$

The latter holds in particular for $\perp\!\!\!\perp$ *, when* $f(x) > 0$.

rijksuniversiteit groningen

Factorization definition

Assume density f w.r.t. product measure on $\mathcal{X}$.
For $a \subseteq V$, $\psi_a(x)$ denotes a function which depends on x_a only, i.e.

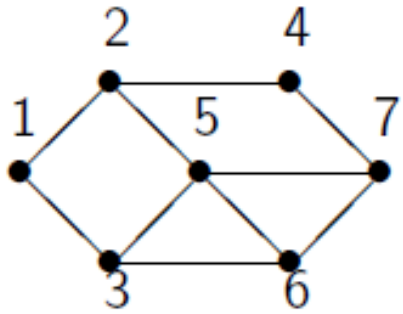
$$x_a = y_a \Rightarrow \psi_a(x) = \psi_a(y).$$

We can then write $\psi_a(x) = \psi_a(x_a)$ without ambiguity.
The distribution of X *factorizes w.r.t.* $\mathcal{G}$ or satisfies (F) if

$$f(x) = \prod_{a \in \mathcal{A}} \psi_a(x)$$

where $\mathcal{A}$ are *complete* subsets of $\mathcal{G}$.
Complete subsets of a graph are sets with all elements pairwise neighbours.

Cliques: maximal complete subgraphs



The *cliques* of this graph are the maximal complete subsets $\{1,2\}$, $\{1,3\}$, $\{2,4\}$, $\{2,5\}$, $\{3,5,6\}$, $\{4,7\}$, and $\{5,6,7\}$. A complete set is any subset of these sets.

The graph above corresponds to a factorization as

$$\begin{aligned} f(x) &= \psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{25}(x_2,x_5) \\ &\times \psi_{356}(x_3,x_5,x_6)\psi_{47}(x_4,x_7)\psi_{567}(x_5,x_6,x_7). \end{aligned}$$

Hammersley-Clifford Theorem

Let (F) denote the property that f factorizes w.r.t. $\mathcal{G}$ and let (G), (L) and (P) denote Markov properties w.r.t. $\perp\!\!\!\perp$. *It then holds that*

$$(\mathrm{F}) \Rightarrow (\mathrm{G})$$

and further: *If* $f(x) > 0$ *for all* x, $(\mathrm{P}) \Rightarrow (\mathrm{F})$.

The former of these is a simple direct consequence of the factorization whereas the second implication is more subtle and known as the *Hammersley–Clifford Theorem*.

Thus in the case of positive density (but typically only then), *all the properties coincide:*

$$(\mathrm{F}) \iff (\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P}).$$

rijksuniversiteit groningen

Conclusion

1. A genomic network is defined by:
   - Evolution
   - "Topology"
   - Links and link strengths
   - Dynamics (later)
2. Graphical models are useful ways to describe genomic networks,
   - in terms of conditional independence relationships.
   - Hammersley-Clifford is key to statistical modelling of GMs.