

Bridging Physics and Biology

let's get a flavour for why today's biology can appeal to physicists

(a) It's a physics "playground":

Soft Matter – it's the stuff of life

Polymers, amphiphiles, diffusion, interactions in water, fluctuations....

For example, active research on: “phase transitions / critical behaviour” and “multivalent membrane interactions”

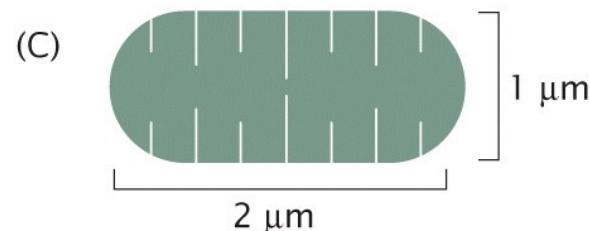
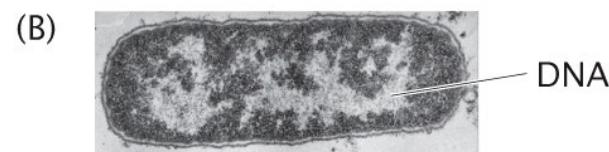
(b) Biology is ripe as a “hard/quantitative science”.

A lightning tour of:

- “Model organisms”
- Organisms & Questions
- Size, Space, Dilutions

E.coli bacterium is one of the best understood “model organisms”.

As well as
importance
of “models”,
many reasons
to study
bacteria and
E.coli in
particular.



What, and how much, stuff is in there ?

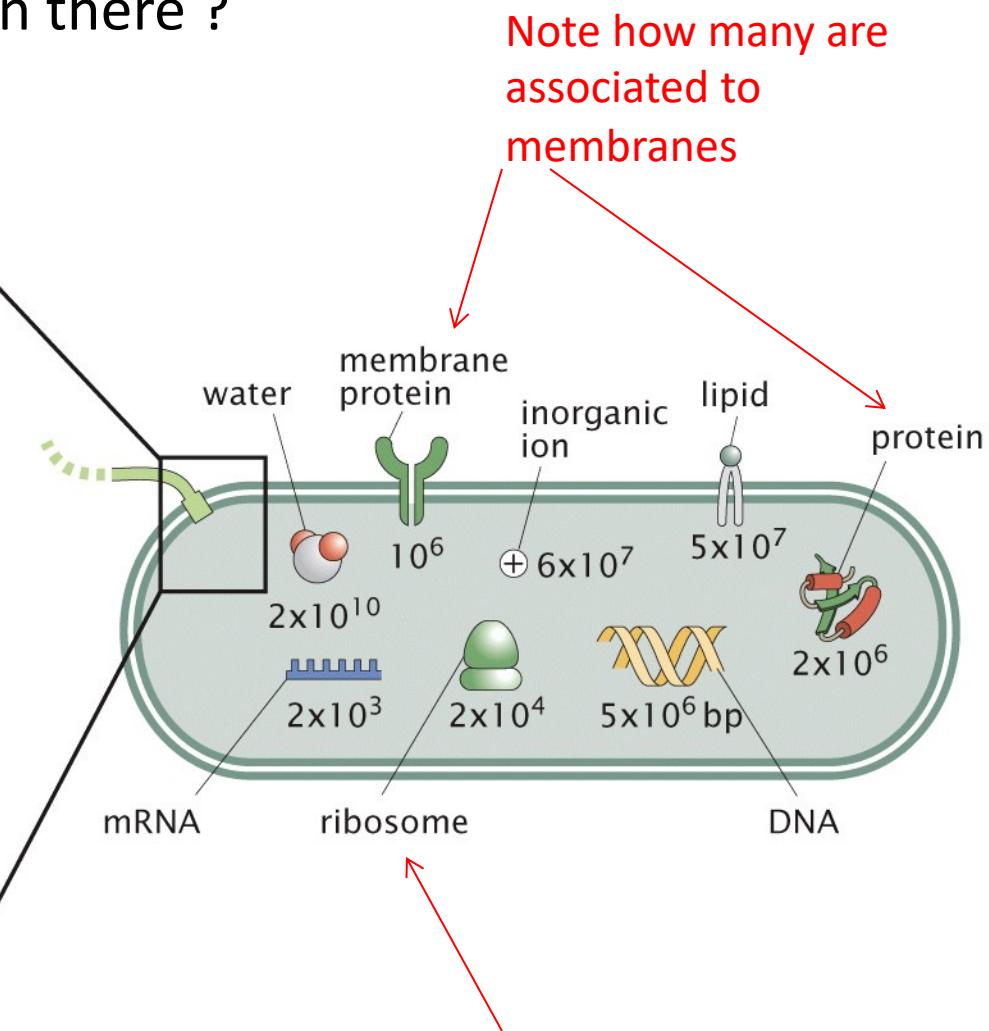
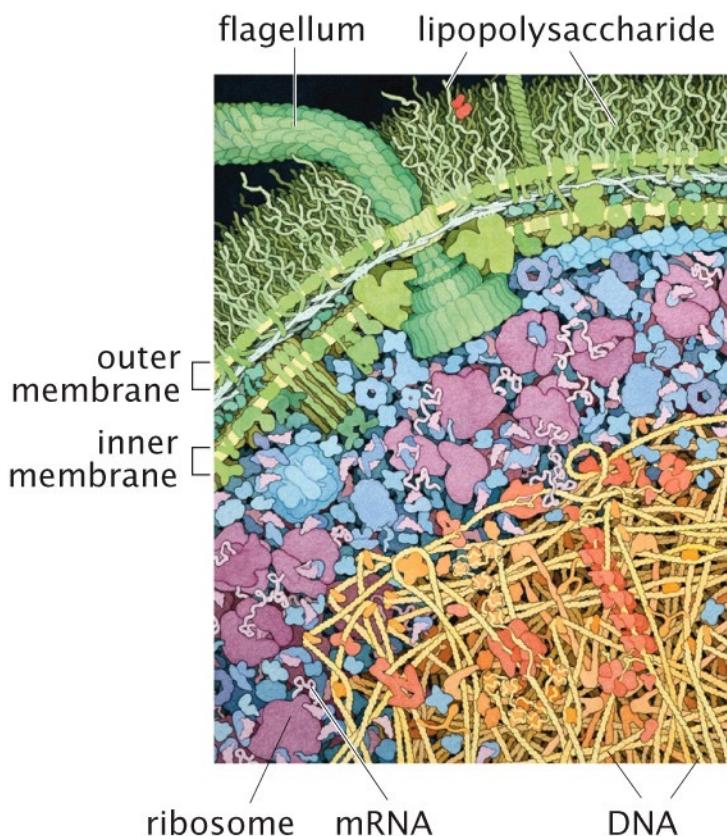


Figure 2.4 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

In *E.coli*, 1 molecule per cell corresponds to a concentration of 2nM

Note how many are associated to membranes

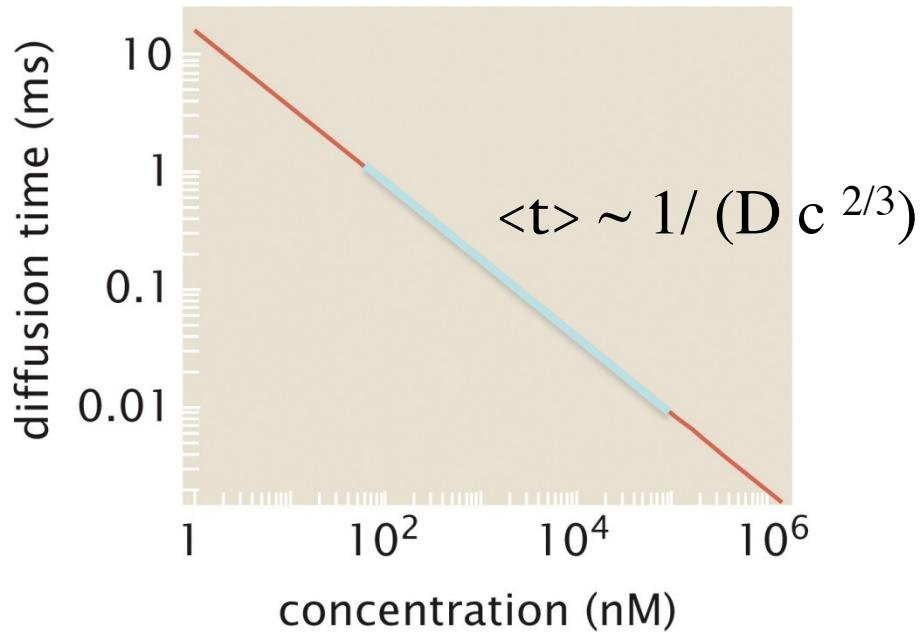
20% of proteins in a cell are making up ribosomes

Table 2.1: Observed macromolecular census of an *E. coli* cell. (Data from F. C. Neidhardt et al., Physiology of the Bacterial Cell, Sinauer Associates, 1990 and M. Schaechter et al., Microbe, ASM Press, 2006.)

Substance	% of total dry weight	Number of molecules
Macromolecules		
Protein	55.0	2.4×10^6
RNA	20.4	
23S RNA	10.6	19,000
16S RNA	5.5	19,000
5S RNA	0.4	19,000
Transfer RNA (4S)	2.9	200,000
Messenger RNA	0.8	1,400
Phospholipid	9.1	22×10^6
Lipopolysaccharide (outer membrane)	3.4	1.2×10^6
DNA	3.1	2
Murein (cell wall)	2.5	1
Glycogen (sugar storage)	2.5	4,360
Total macromolecules	96.1	
Small molecules		
Metabolites, building blocks, etc.	2.9	
Inorganic ions	1.0	
Total small molecules	3.9	

Diffusion times (we come back to this later)

+ crowding



+ Effect of size (of cell, and of object diffusing)

Protein diffusion times across cell:

0.1s bacteria 10s yeast 100s mammalian

Small molecule diffusion times across cell:

0.001s bacteria 0.01s yeast 0.1s mammalian

Spacing between molecules

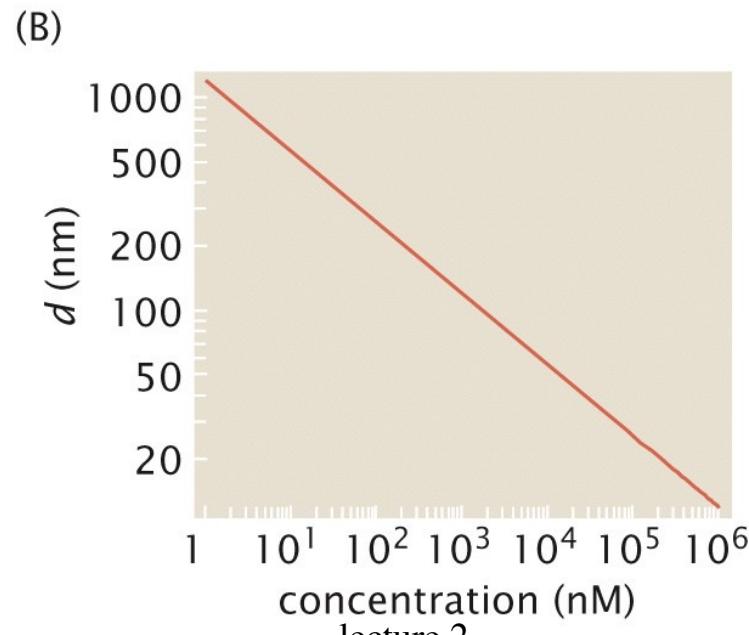
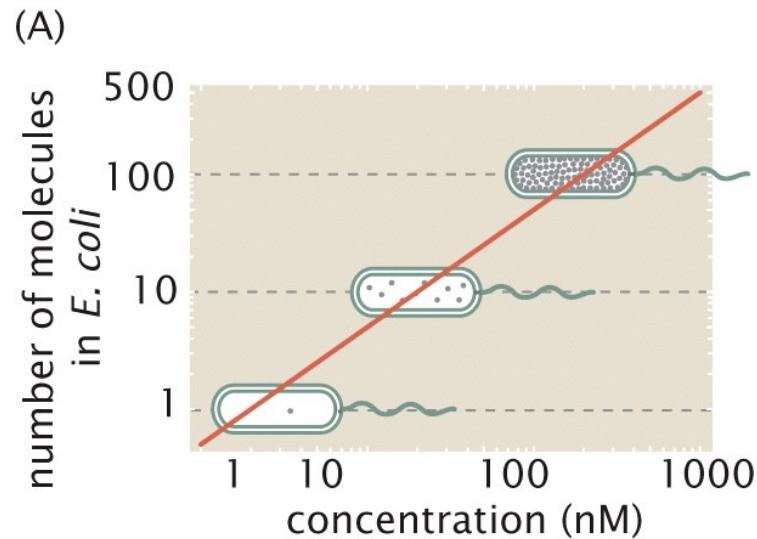
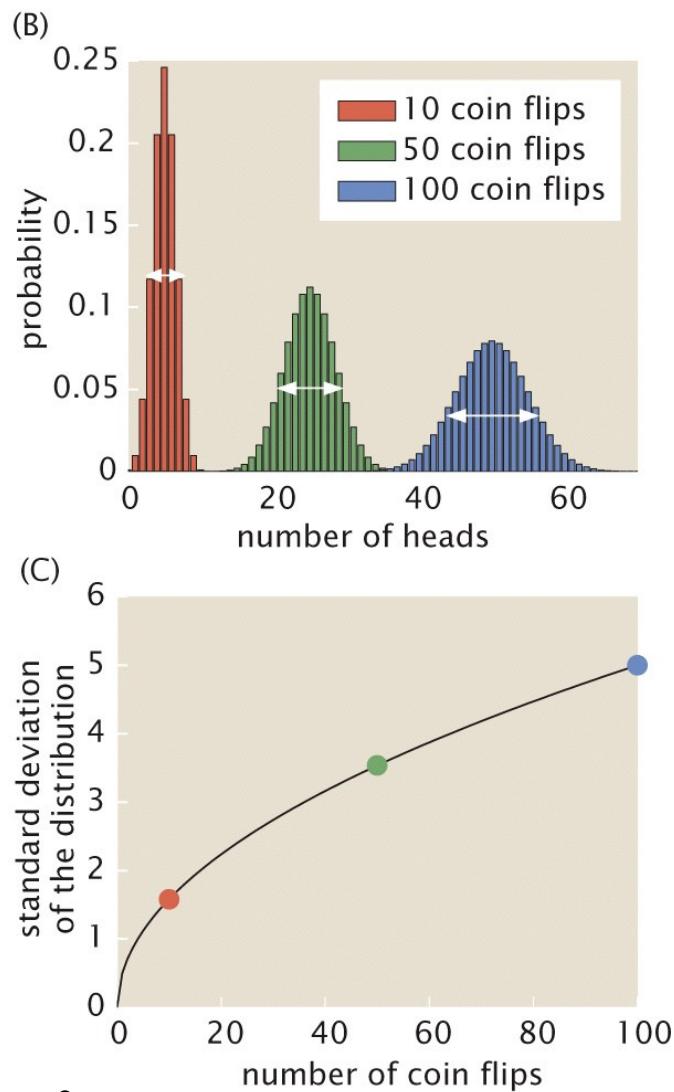
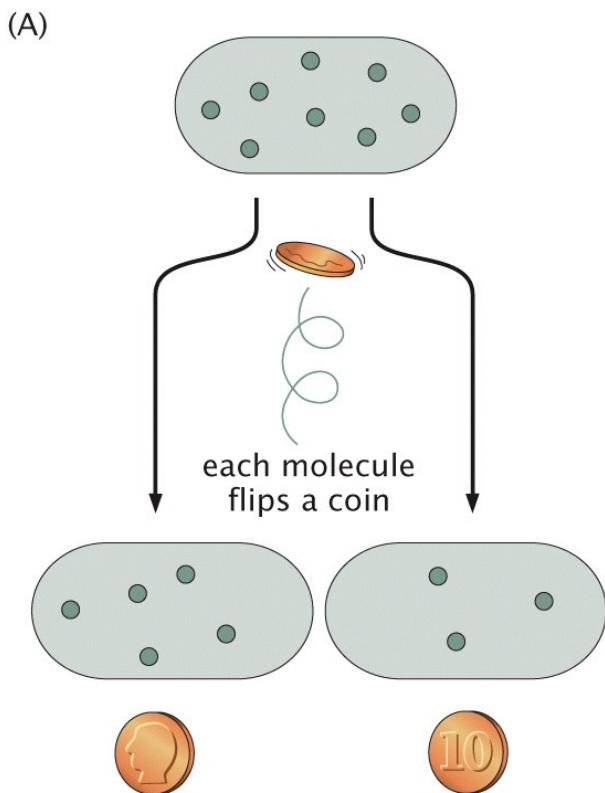


Figure 2.12 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Keep fresh in mind these “trivial” scalings.
(e.g. different for things bound and diffusing along membranes or along filaments).

Cell is crowded – but of a particular protein there might be few.



$$\text{If } p=q, \langle n \rangle = Np, \langle n^2 \rangle - \langle n \rangle^2 = Npq$$

$$\text{Sqrt}(\langle n^2 \rangle - \langle n \rangle^2) / \langle n \rangle = 1/\sqrt{N}$$

So rule of thumb stochasticity

Important if $N < 100$

This is one source of stochasticity called “biochemical noise” and gives binomial distributions

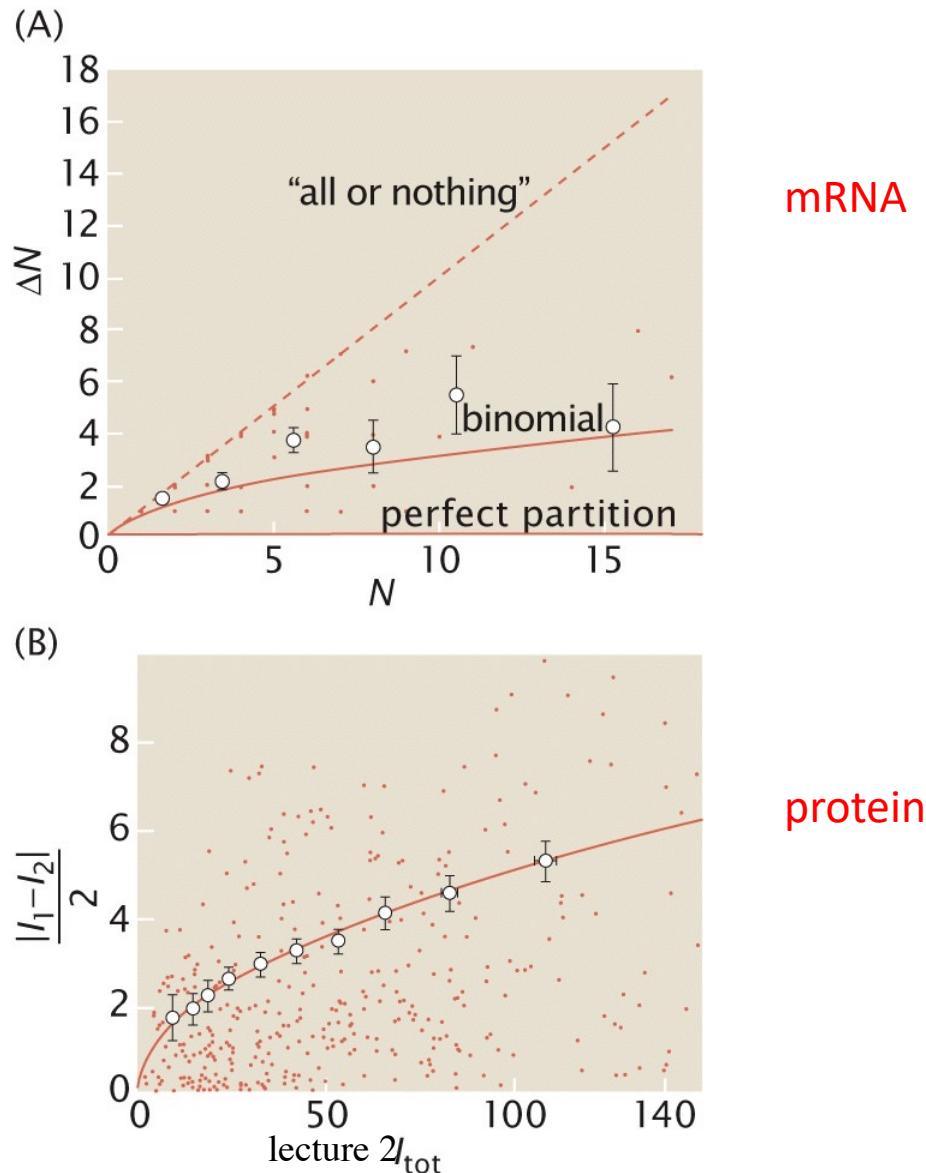


Figure 2.9 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Nice quantitative experiments on this

– but how to calibrate α in $I_{\text{tot}} = \alpha N_{\text{tot}}$?

For random partitioning, $\langle (I_1 - I_2)^2 \rangle = \alpha I_{\text{tot}}$ (*proof QS exercise*)

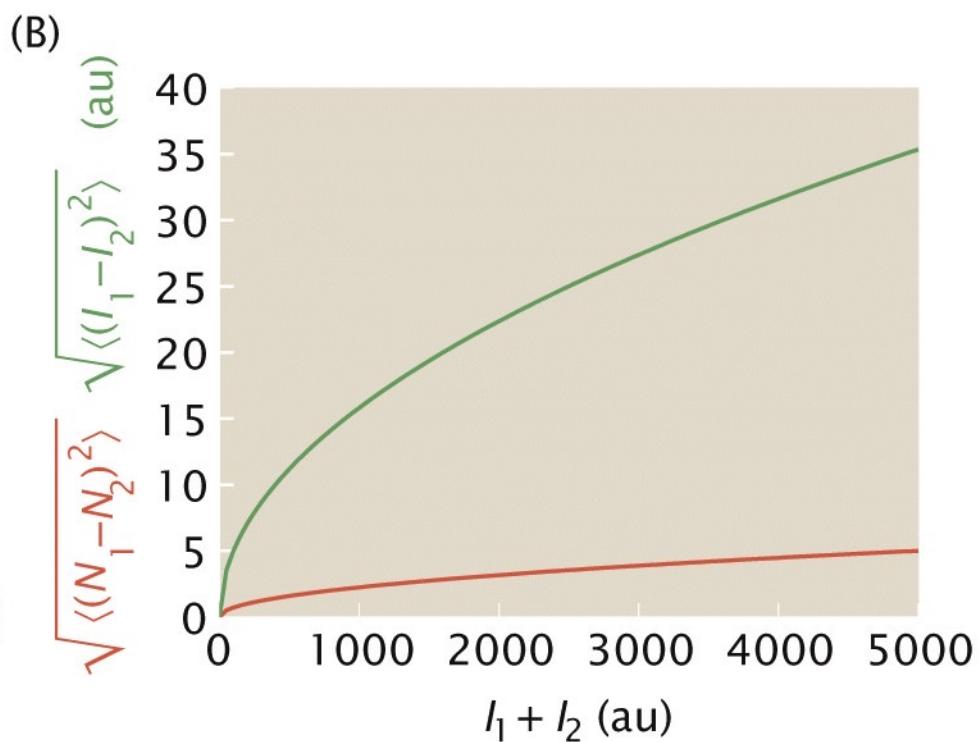
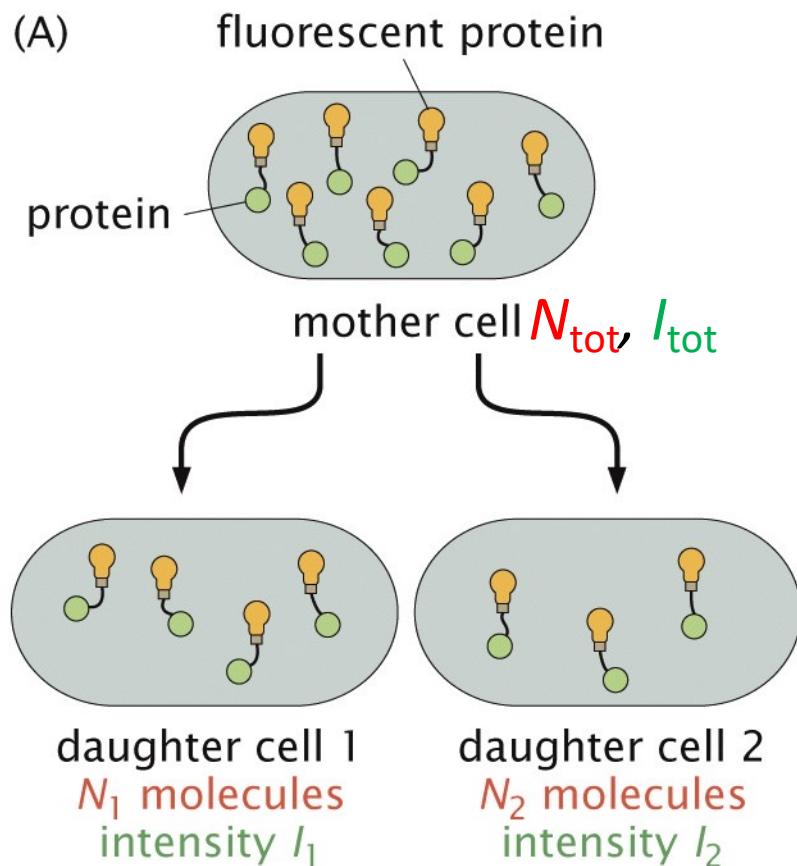
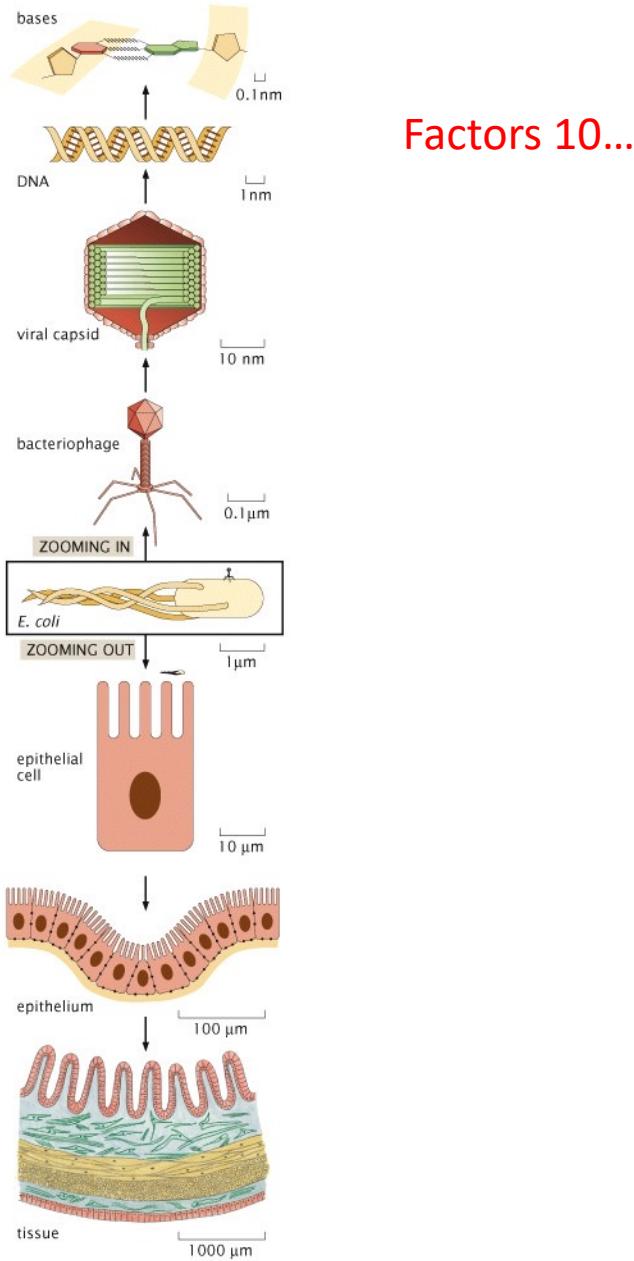


Figure 2.10 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

MatLab simulation here

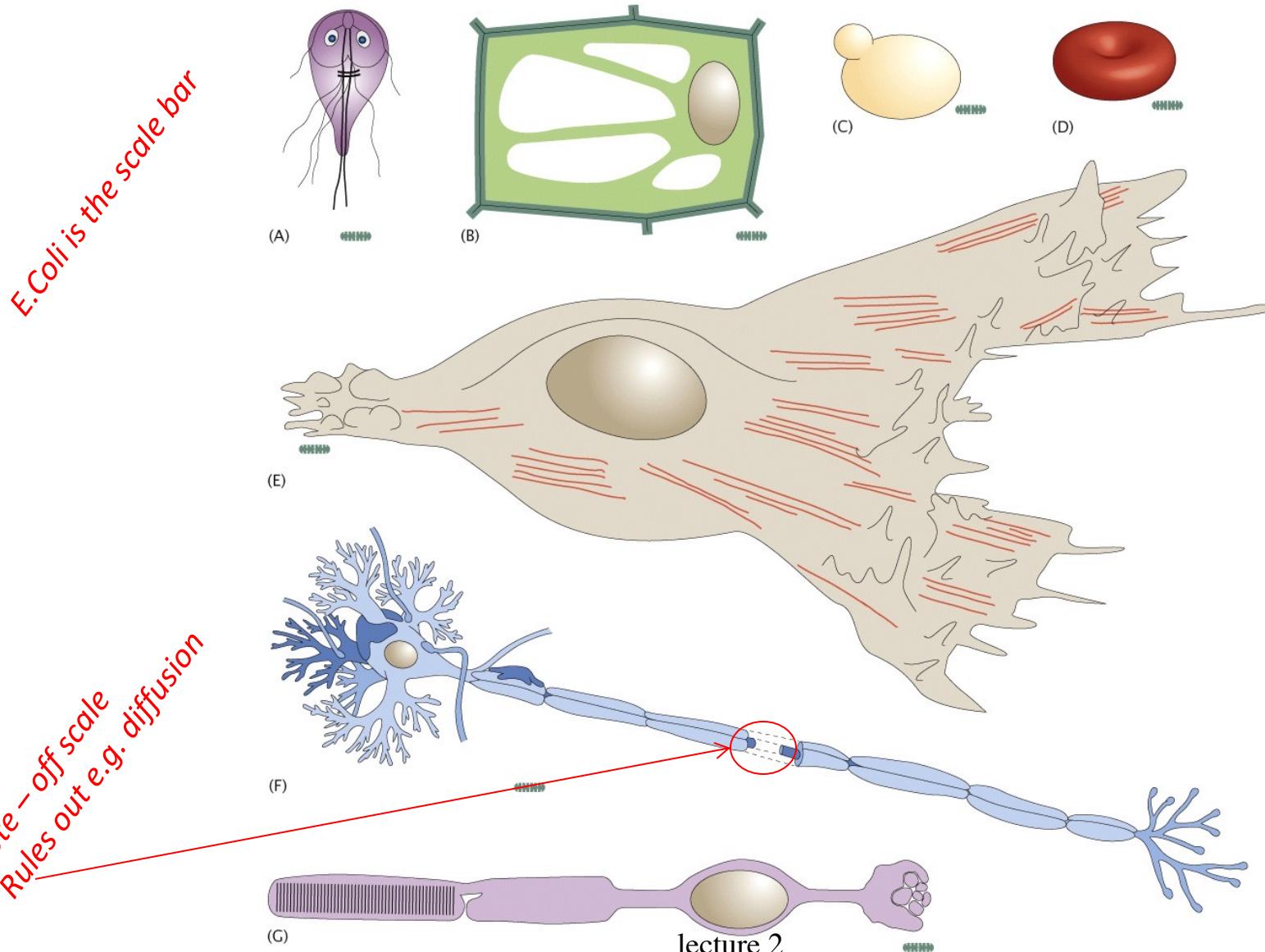
Idea of lengthscales - 1



Factors 10...

lecture 2
Figure 2.15 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Idea of lengthscales - 2



lecture 2

Even a snapshot in time shows considerable complexity in a cell

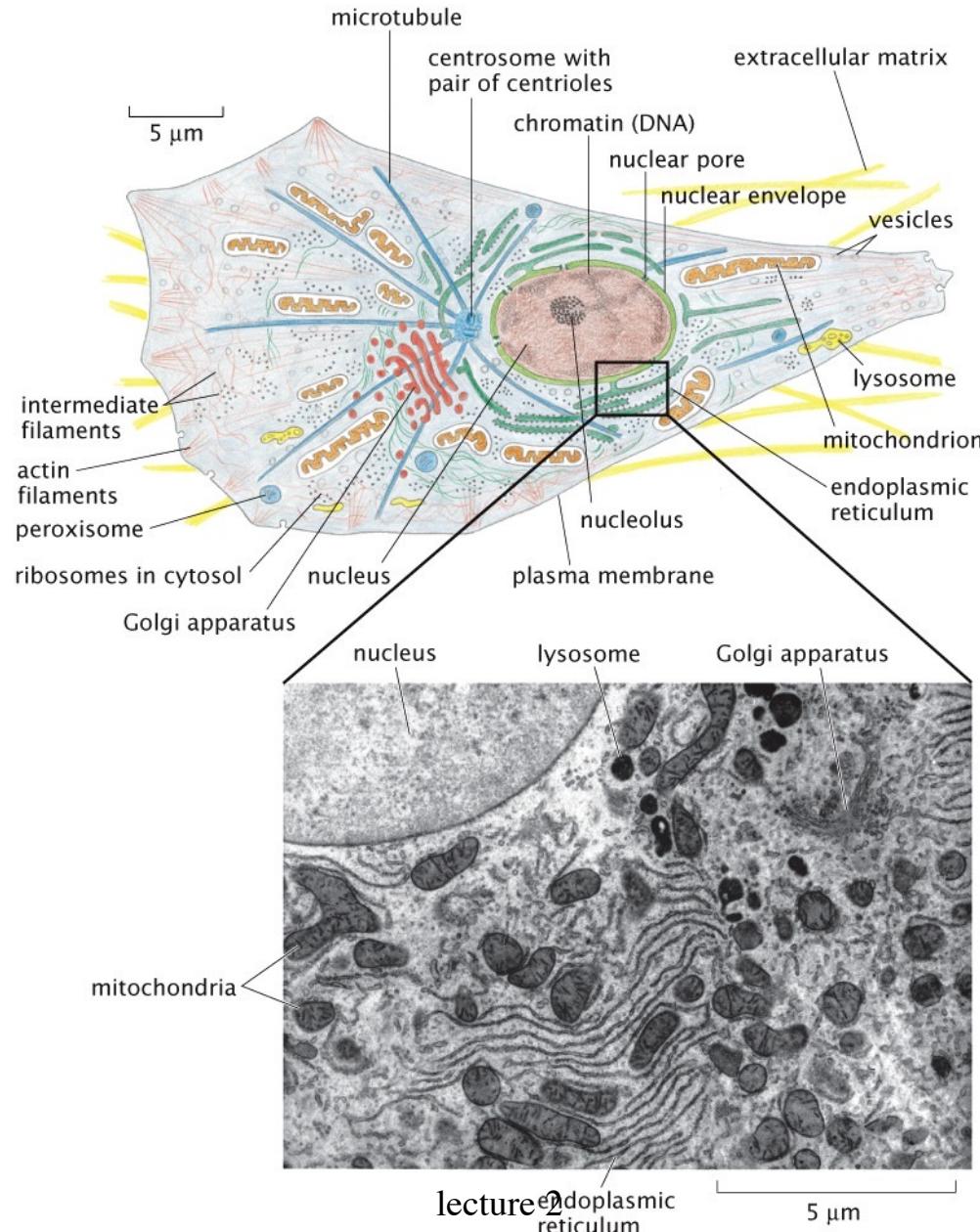


Figure 2.23 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Lipid/protein structures (organelles) Dynamic, deformable, and self-assembled at least at each cell division

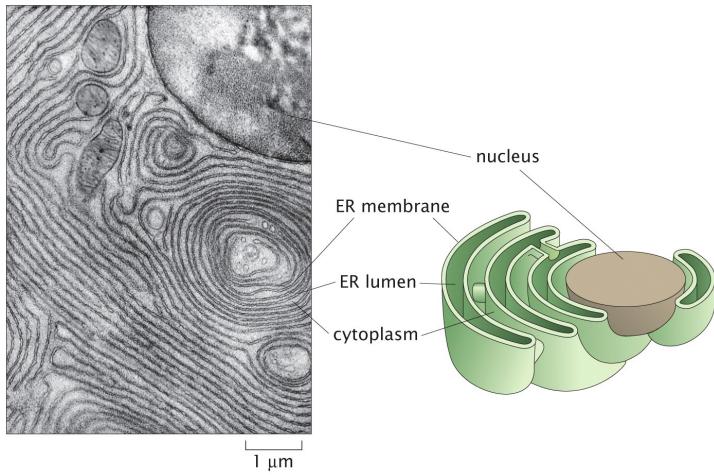


Figure 2.24 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

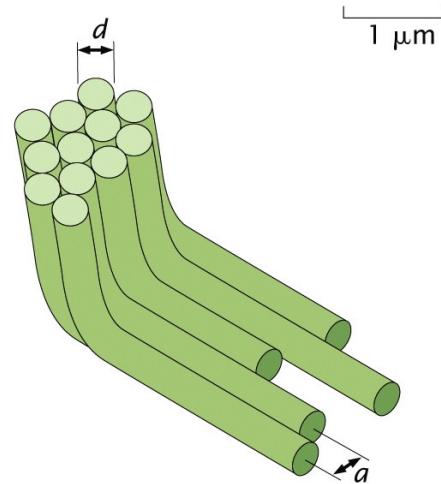
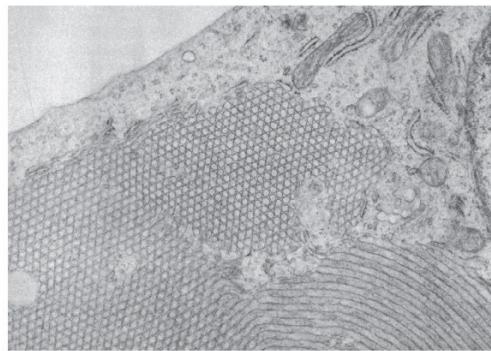


Figure 2.25b Physical Biology of the Cell, 2ed. (© Garland Science 2013)

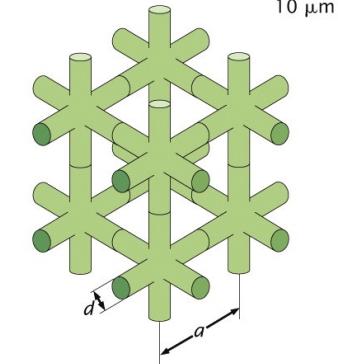
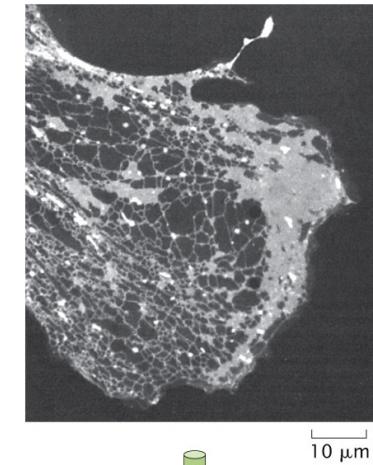


Figure 2.25a Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Key functional units are protein (or protein RNA) assemblies

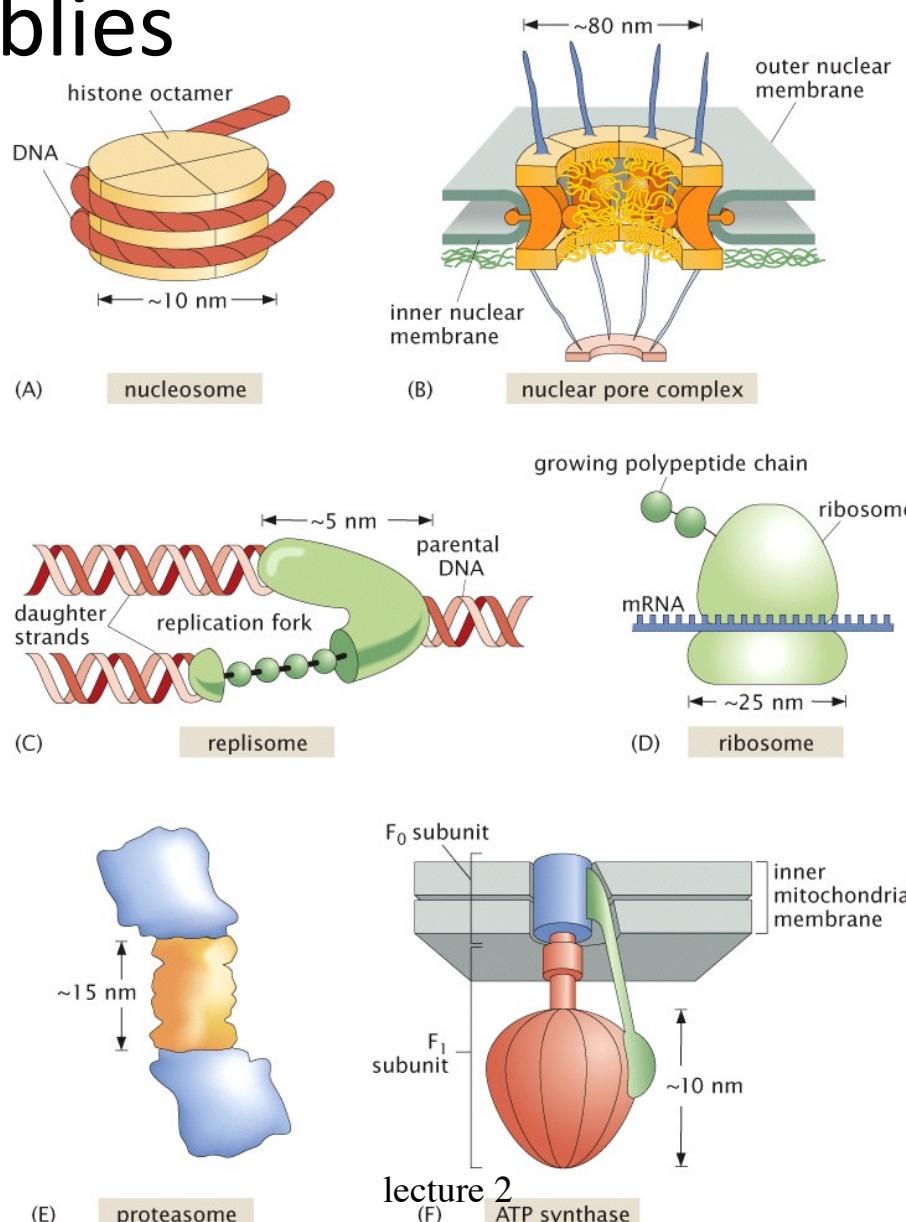


Figure 2.26 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Self assembly and hierarchy of scales is clearly visible in many structural and functional units.

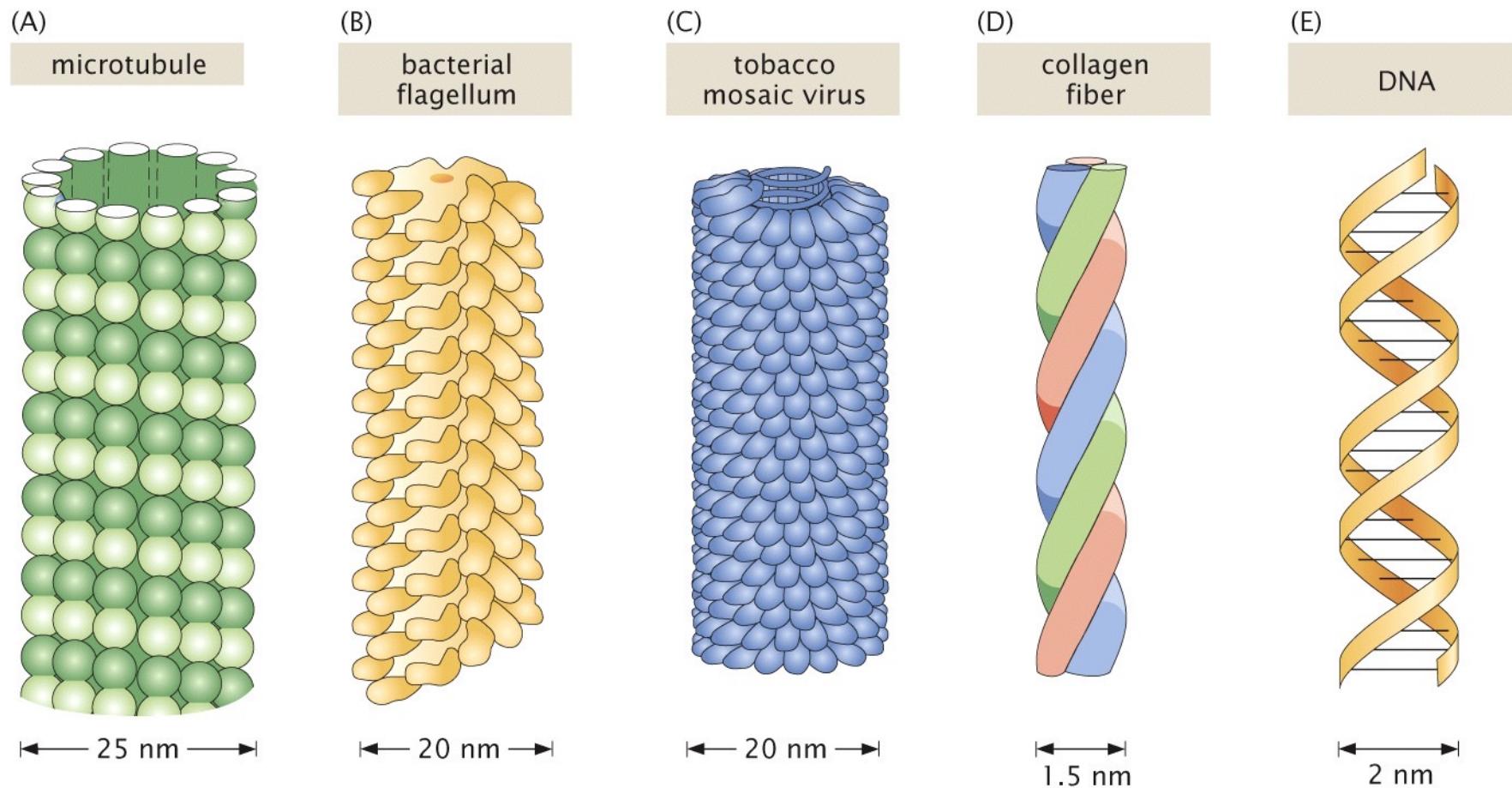
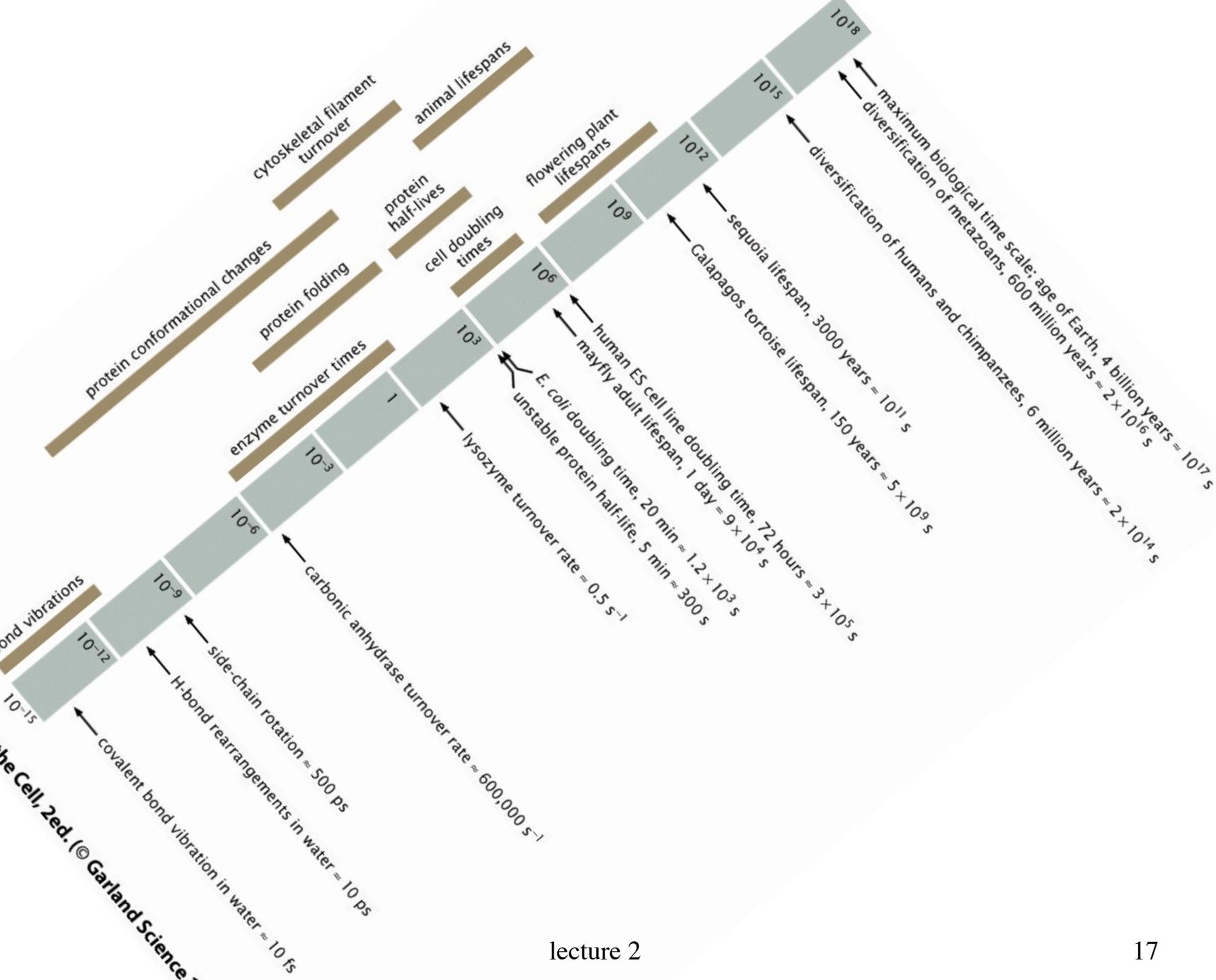


Figure 2.27 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

“Model organisms”. Organisms & Questions

Time



Development of *Drosophila*

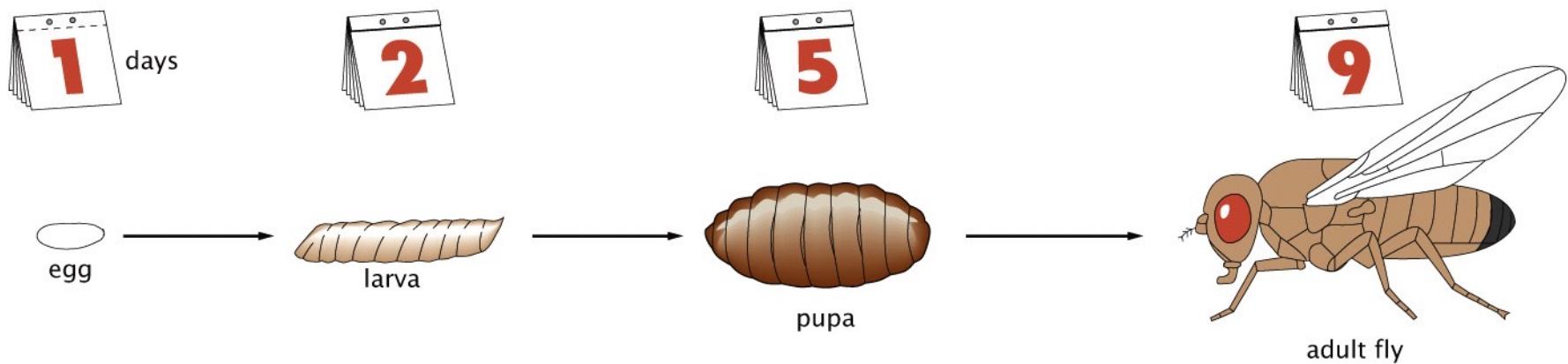


Figure 3.2a Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Early development of *Drosophila*

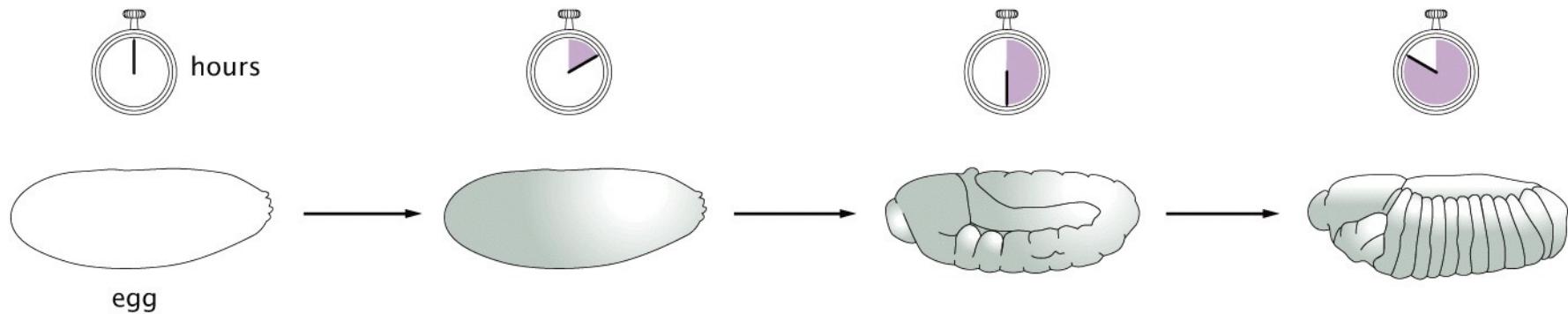


Figure 3.2b Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Bacterial cell division

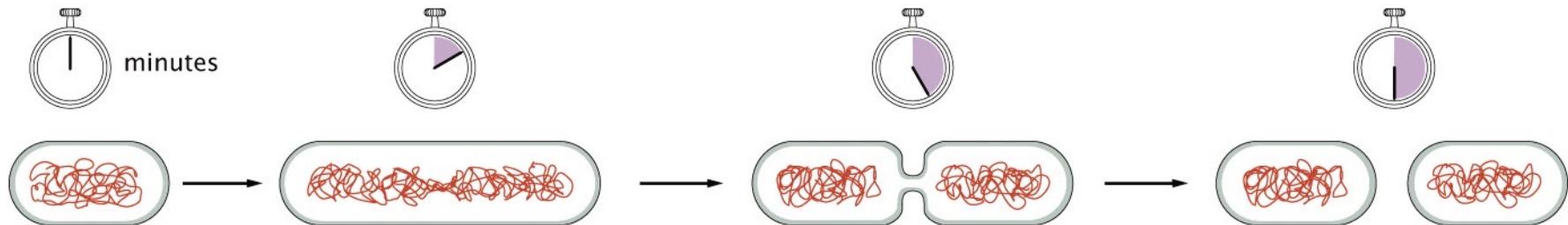


Figure 3.2c Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Cell generation times:

30min-few hrs bacteria

20hrs to non-dividing mammalian

2hr- many hrs yeast

Cell movements

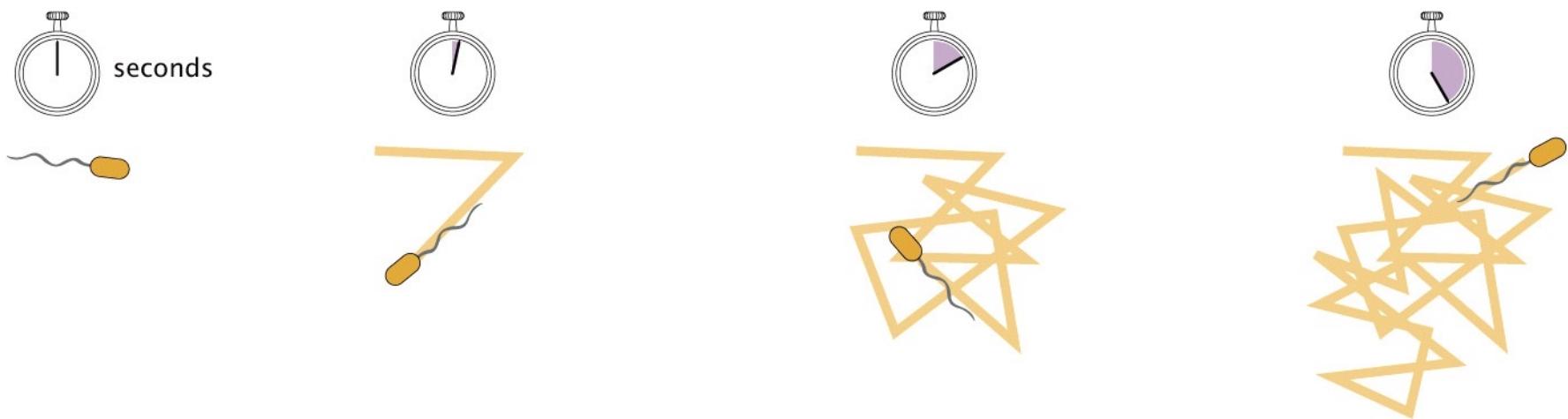


Figure 3.2d Physical Biology of the Cell, 2ed. (© Garland Science 2013)

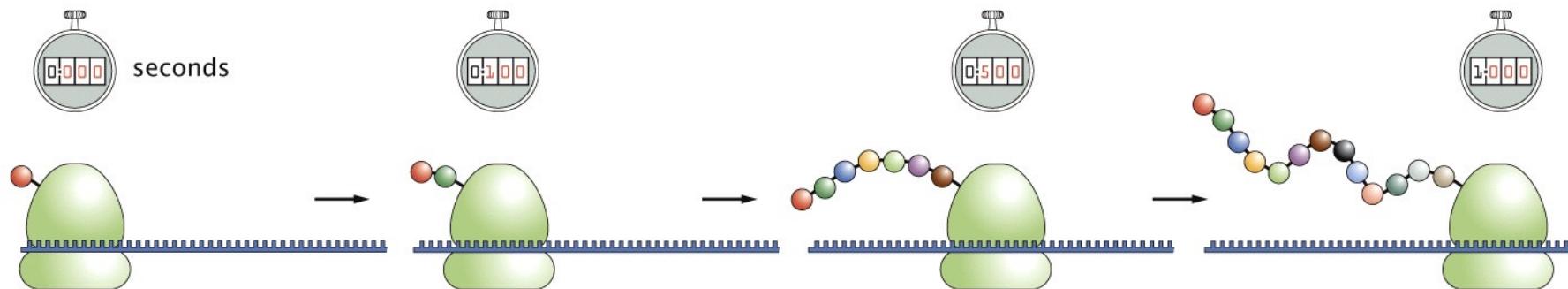


Figure 3.2e Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Protein translation time:

~2min in bacteria or yeast (40aa/sec),
and ~ 30 mins in mammalian cells, including mRNA nuclear export

Typical mRNA lifetime:

2-5mins bacteria

10mins-1hr yeast

10min-10hrs mammalian

Cell volumes: $1\mu\text{m}^3$ bacteria $1000\mu\text{m}^3$ yeast $10000\mu\text{m}^3$ mammalian**Considering the cell volume, concentrations of 1 protein per cell are:**

1nM bacteria

1pM yeast

0.1pM mammalian

Protein per cell are: $4 \cdot 10^6$ bacteria $4 \cdot 10^9$ yeast $4 \cdot 10^{10}$ mammalian**Ribosomes per cell are:** 10^4 bacteria 10^7 yeast 10^8 mammalian

Transcription

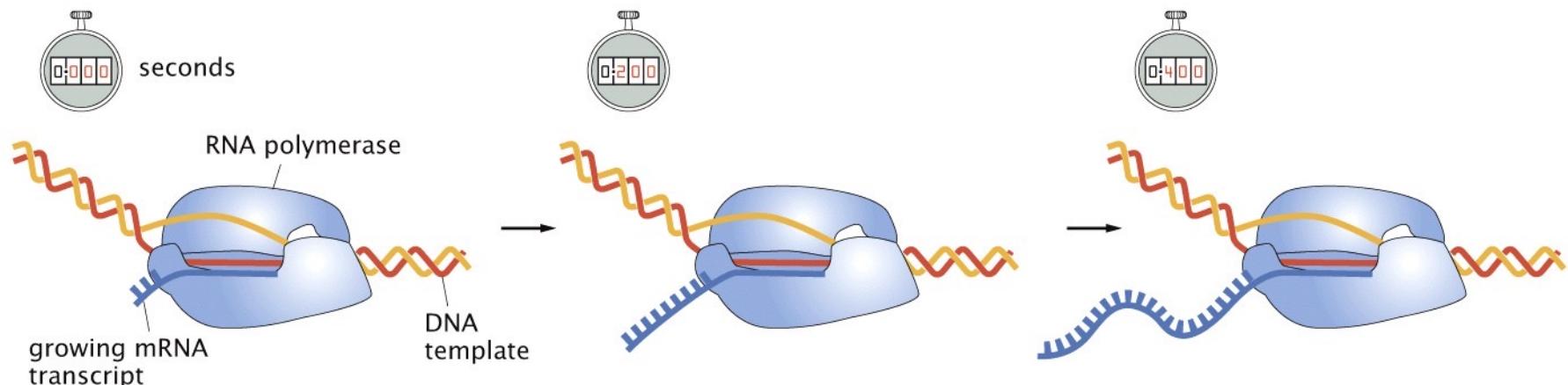


Figure 3.2f Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Gene transcription times:

1min bacteria (80bp/sec)

1min yeast

30min mammalian (including mRNA processing)

Genome sizes:

4.6 10^6 bp bacteria

4500 genes

1.3 10^7 bp yeast

6600 genes

3 10^9 bp mammalian

30000 genes

Gating of ion channels

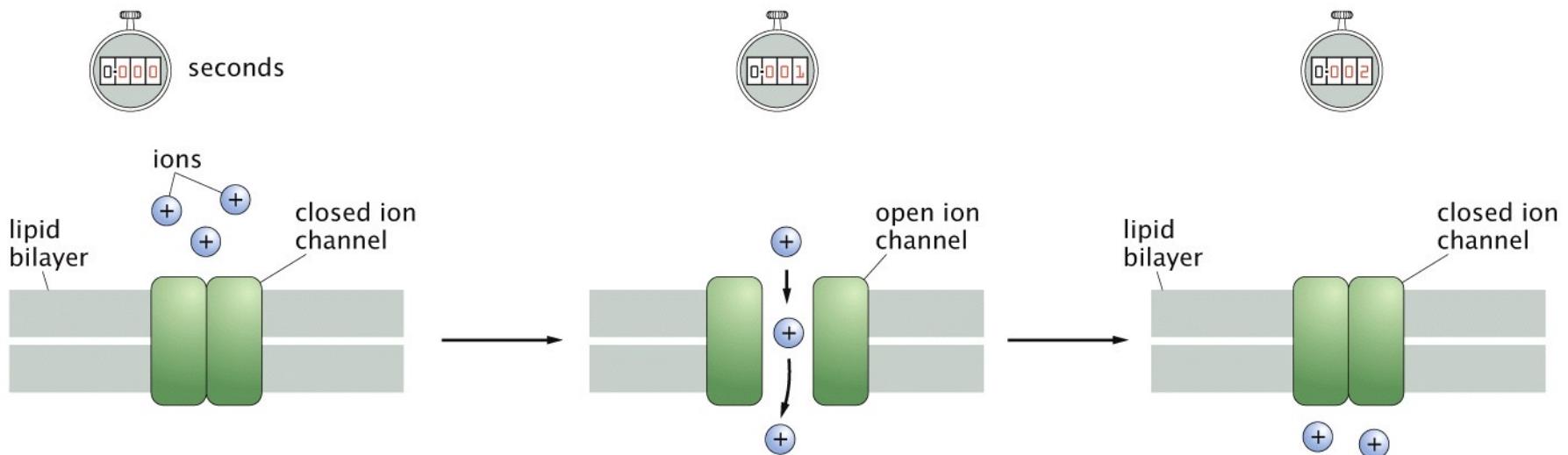


Figure 3.2g Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Enzyme catalysis

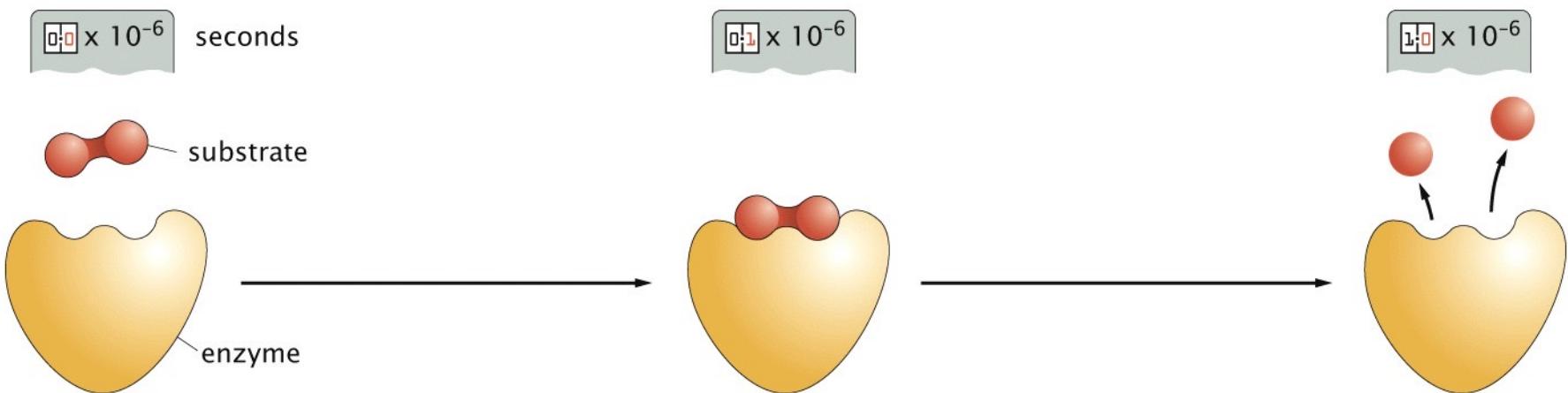


Figure 3.2h Physical Biology of the Cell, 2ed. (© Garland Science 2013)

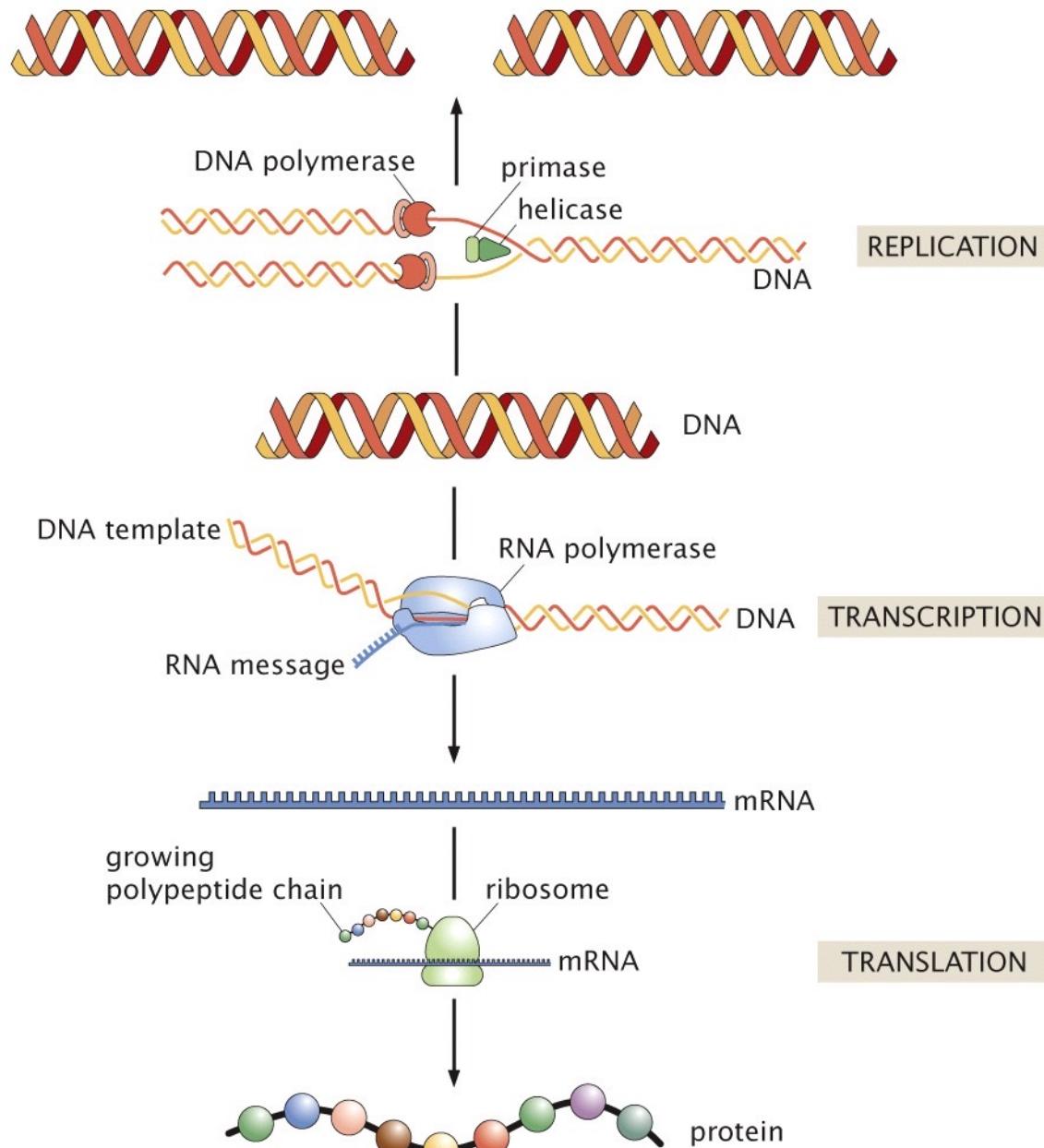


Figure 3.12 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

mRNA lifetimes

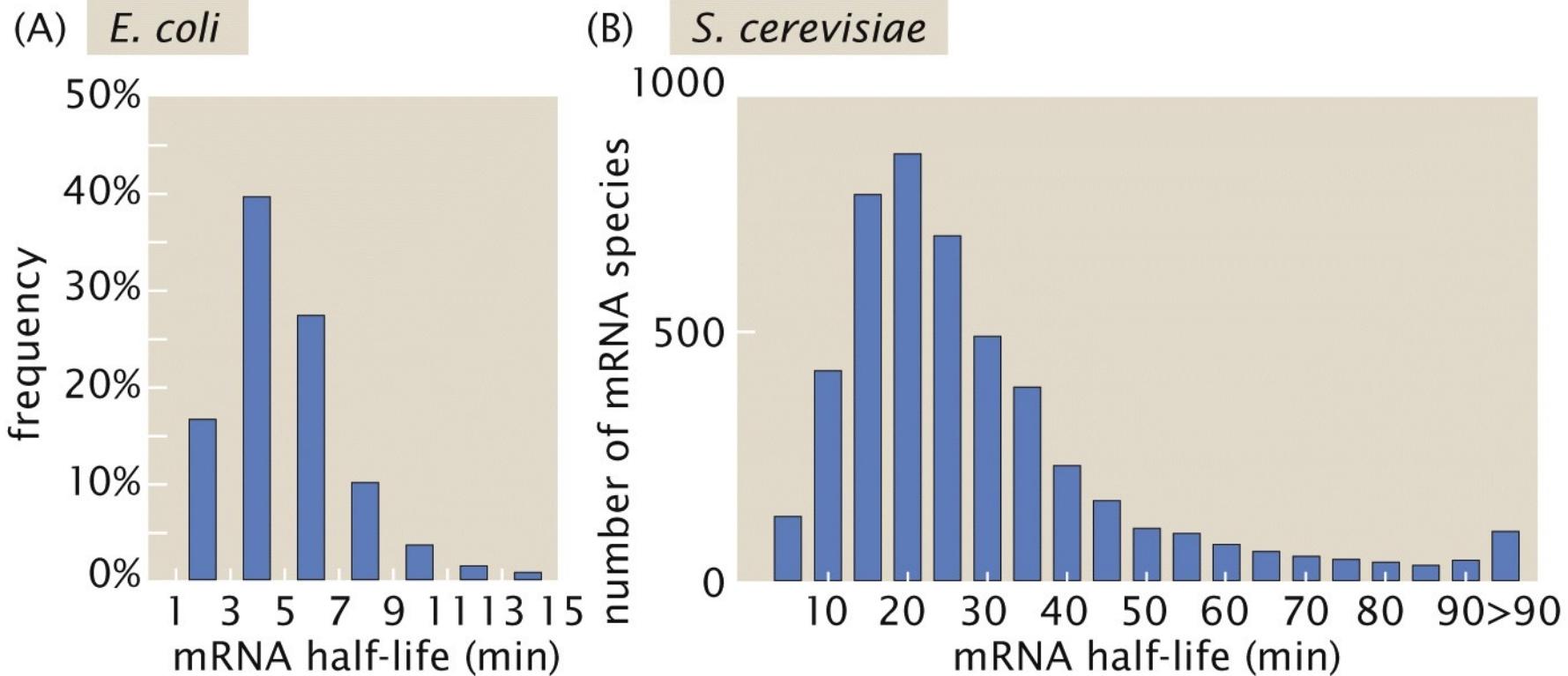
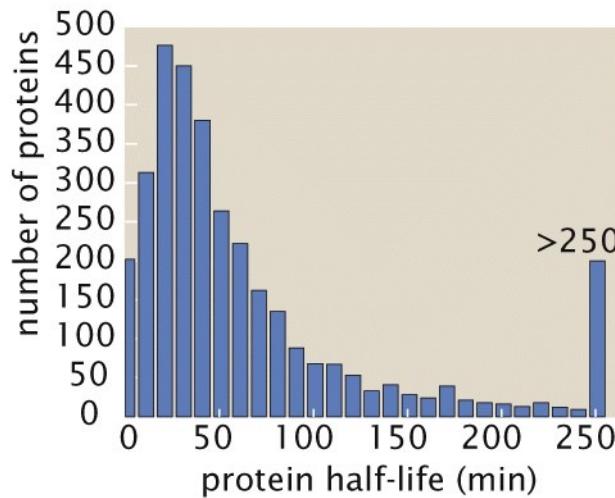


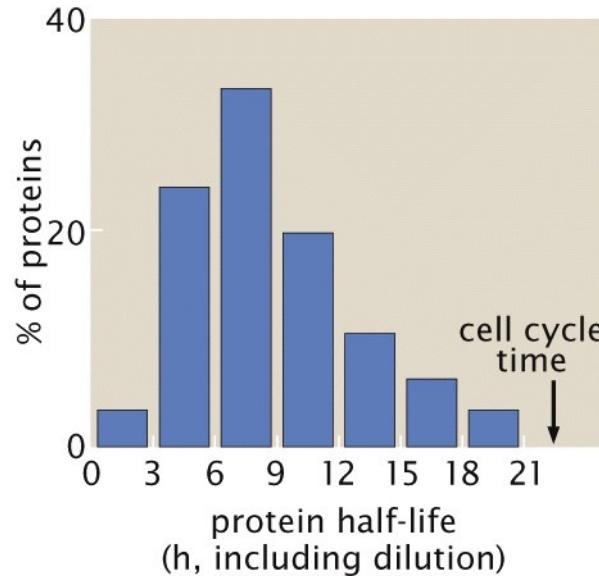
Figure 3.14 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Protein lifetimes

(A) *S. cerevisiae*



(B) human cancer cell line



STRUCTURE OF NETWORKS

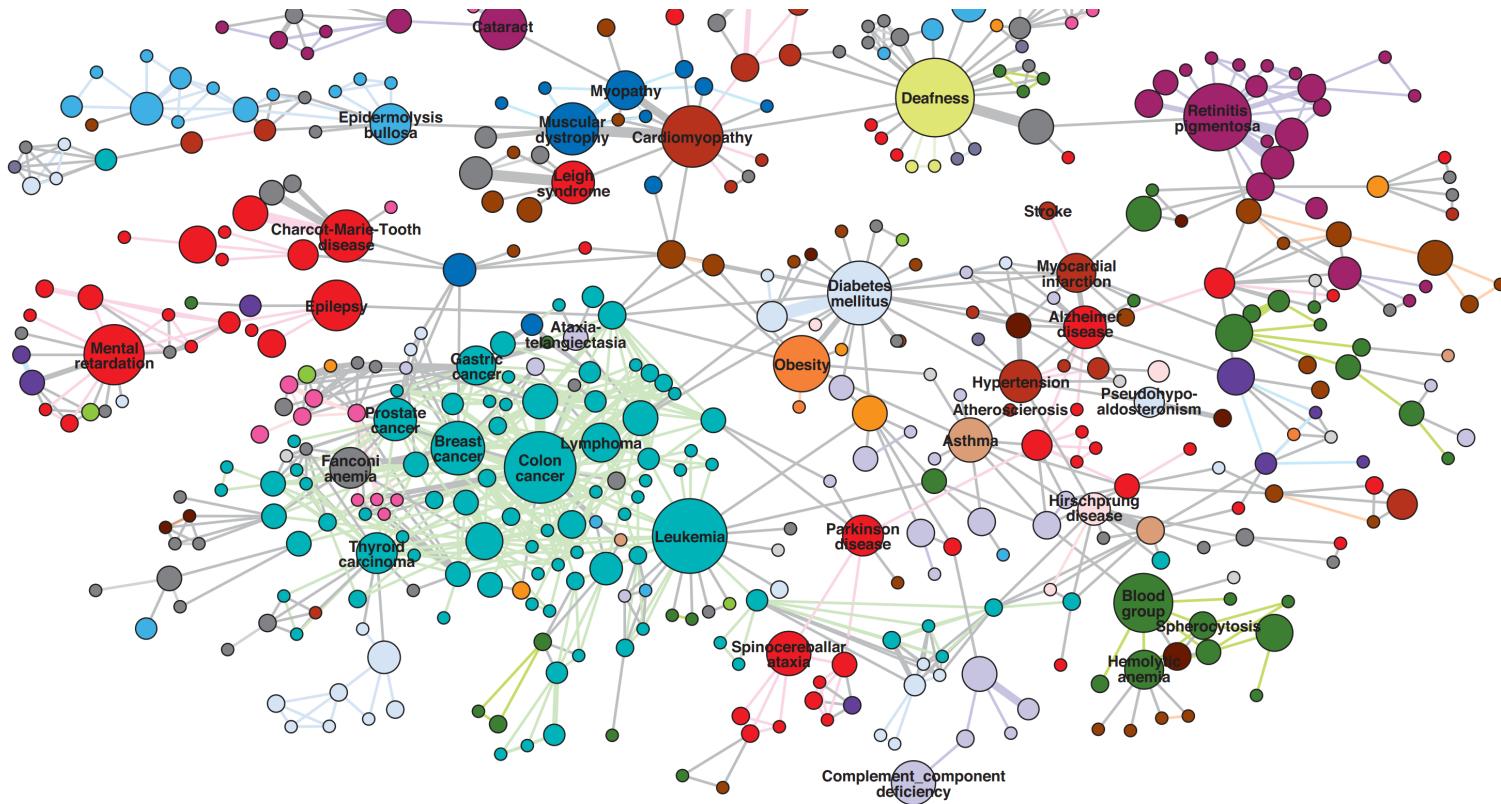
credit – Dr Sebastian Ahnert

Networks and network data surround us:

- Mobile phone networks
- Transport networks
- Power grids
- Online social networks (e.g. Facebook, Twitter)
- Gene regulatory networks
- Protein interaction networks
- Neural networks

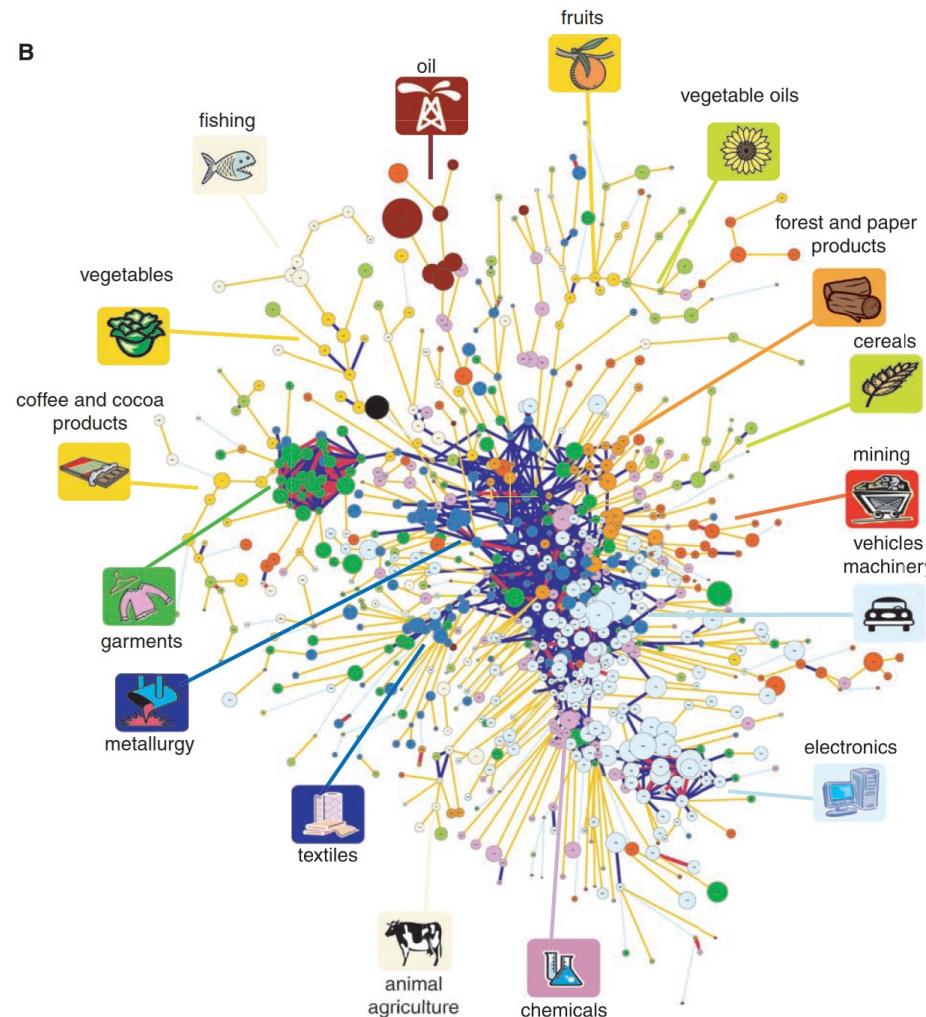
Networks – disease network

Goh, K.-I. et al. The human disease network. PNAS 104, 8685–8690 (2007).



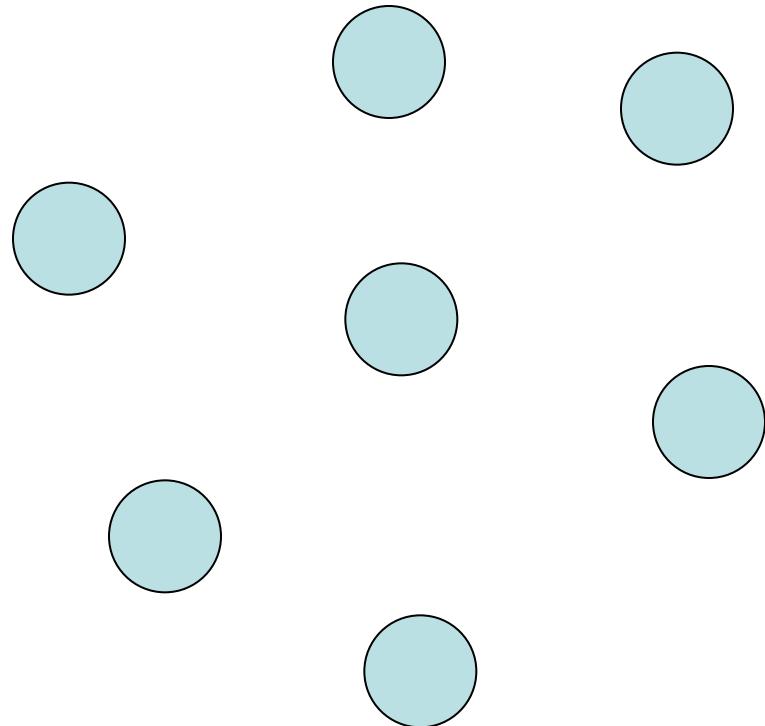
Networks – products

Hidalgo, C. A., Klinger, B., Barabasi, A.-L. & Hausmann, R. The product space conditions the development of nations. Science 317, 482–487 (2007).



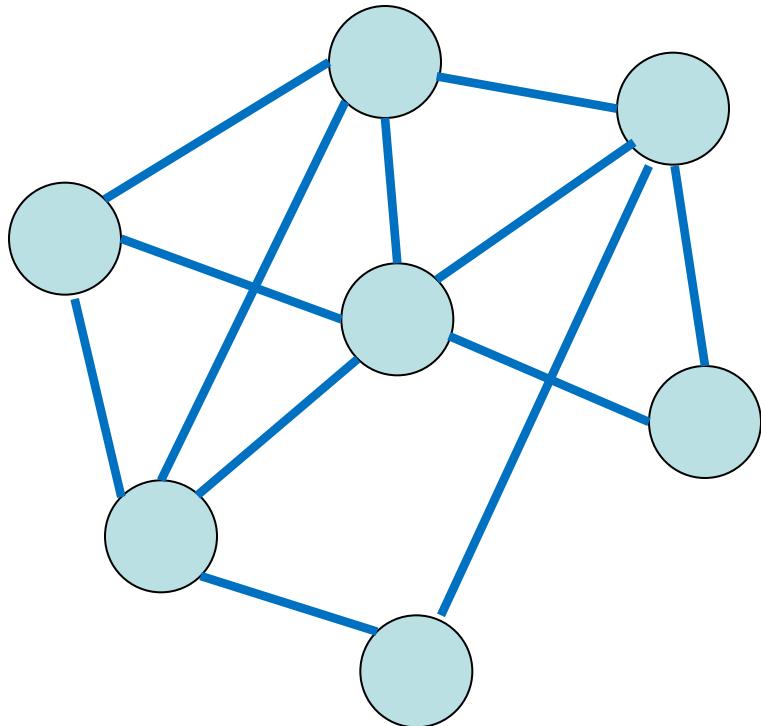
Networks – basics

A network consists of **nodes**



Networks – basics

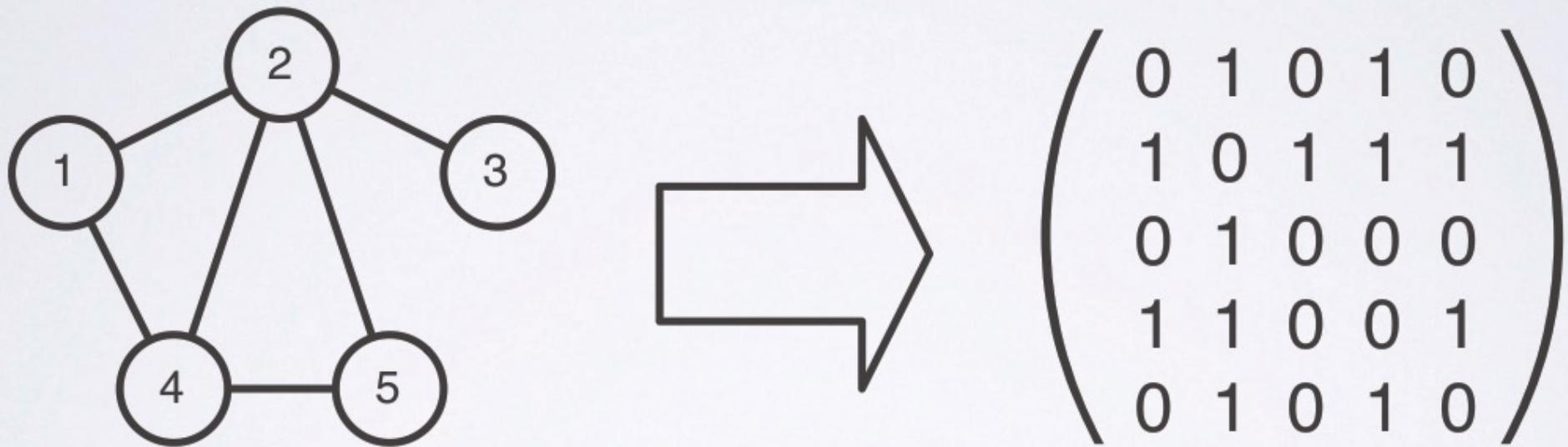
A network consists of **nodes** and **edges**



This abstract framework allows us to examine a wide range of networks with the same tools.

Networks – basics

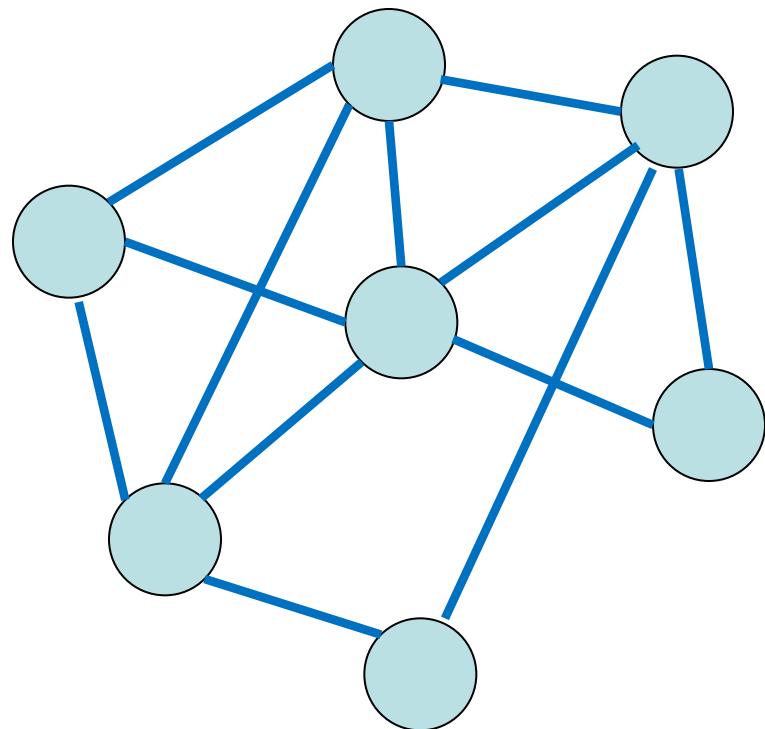
The most convenient way of describing the **structure** of a network is the adjacency matrix a_{ij} .



A link between node i to node j is recorded by a '1' in the i th row and the j th column.

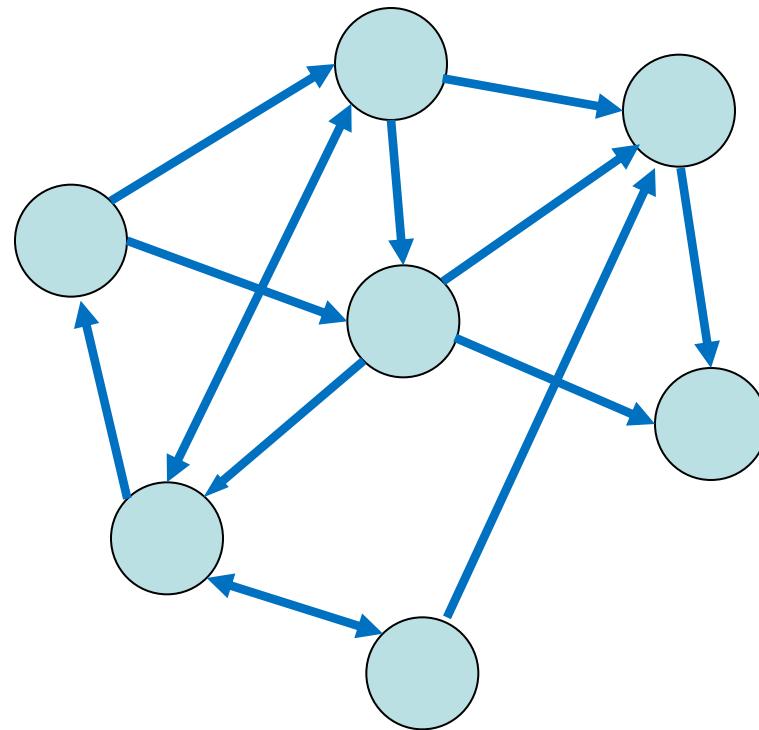
Networks – basics

Networks can be **undirected**



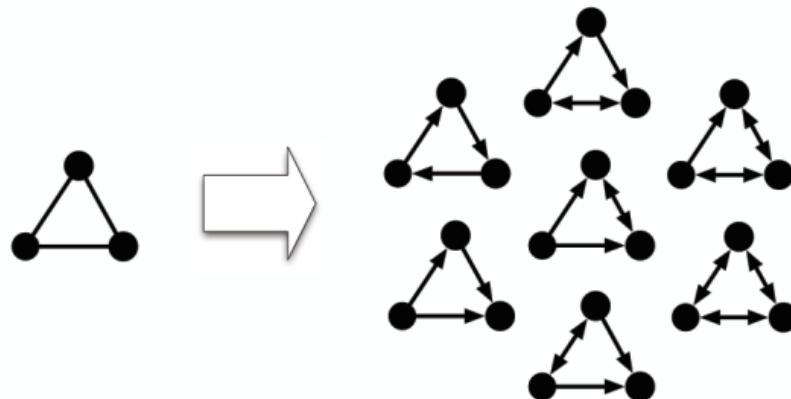
Networks – basics

Networks can be undirected **or** directed



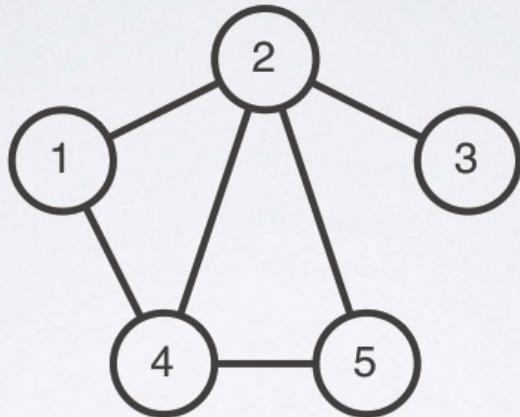
Directed networks have twice as many potential connections and have a much higher complexity.

Consider this:

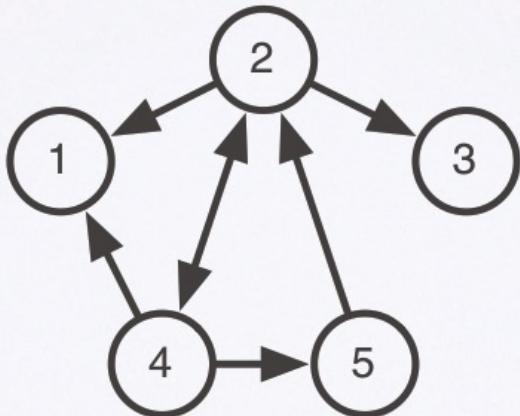


Networks – basics

Undirected networks have a *symmetric* adjacency matrix $a_{ij} = a_{ji}$



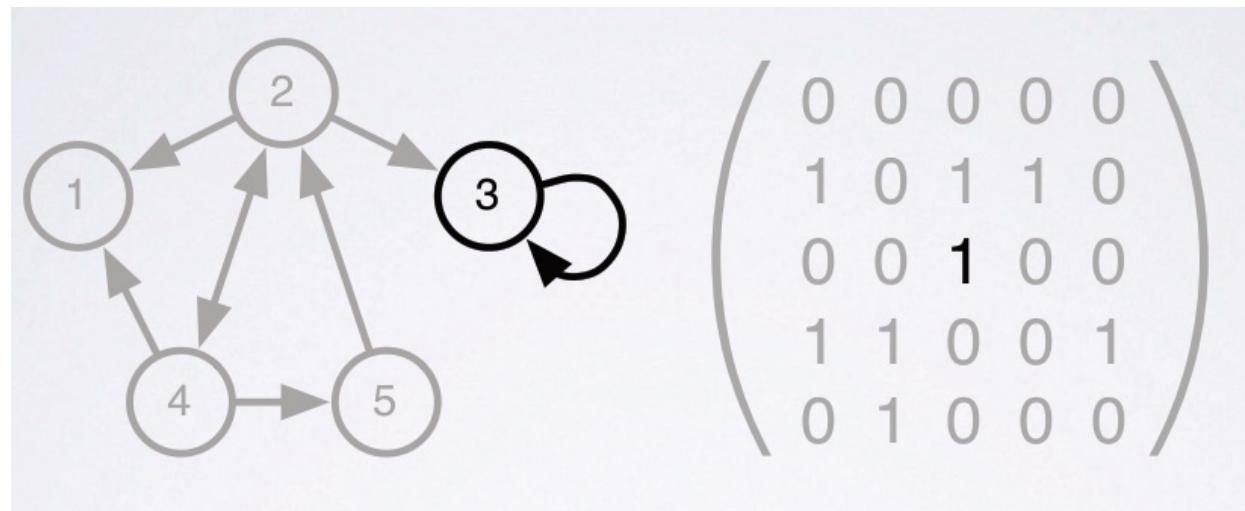
$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$



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

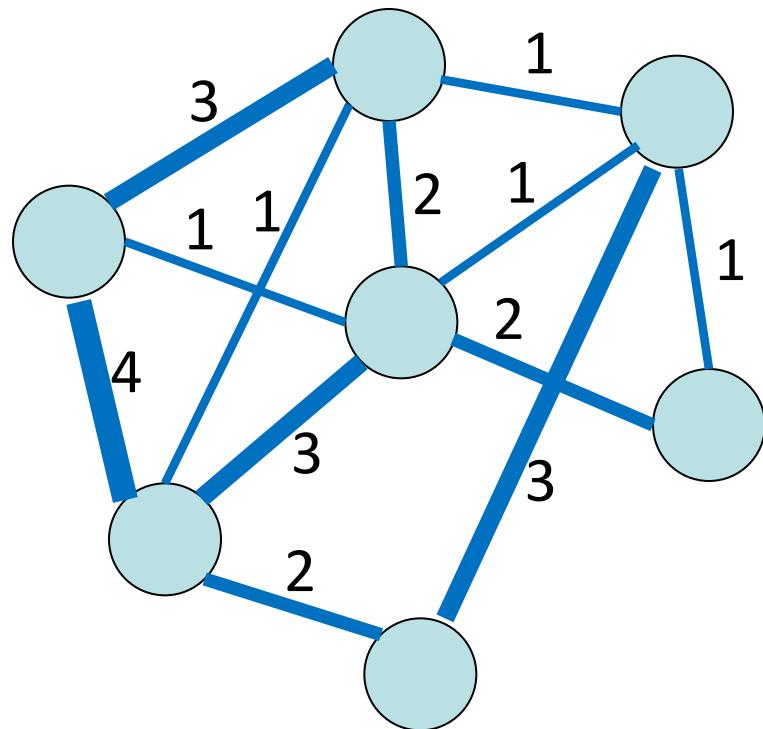
Networks – basics

Networks also can have **self-interactions**, which correspond to the diagonal entries a_{ii}



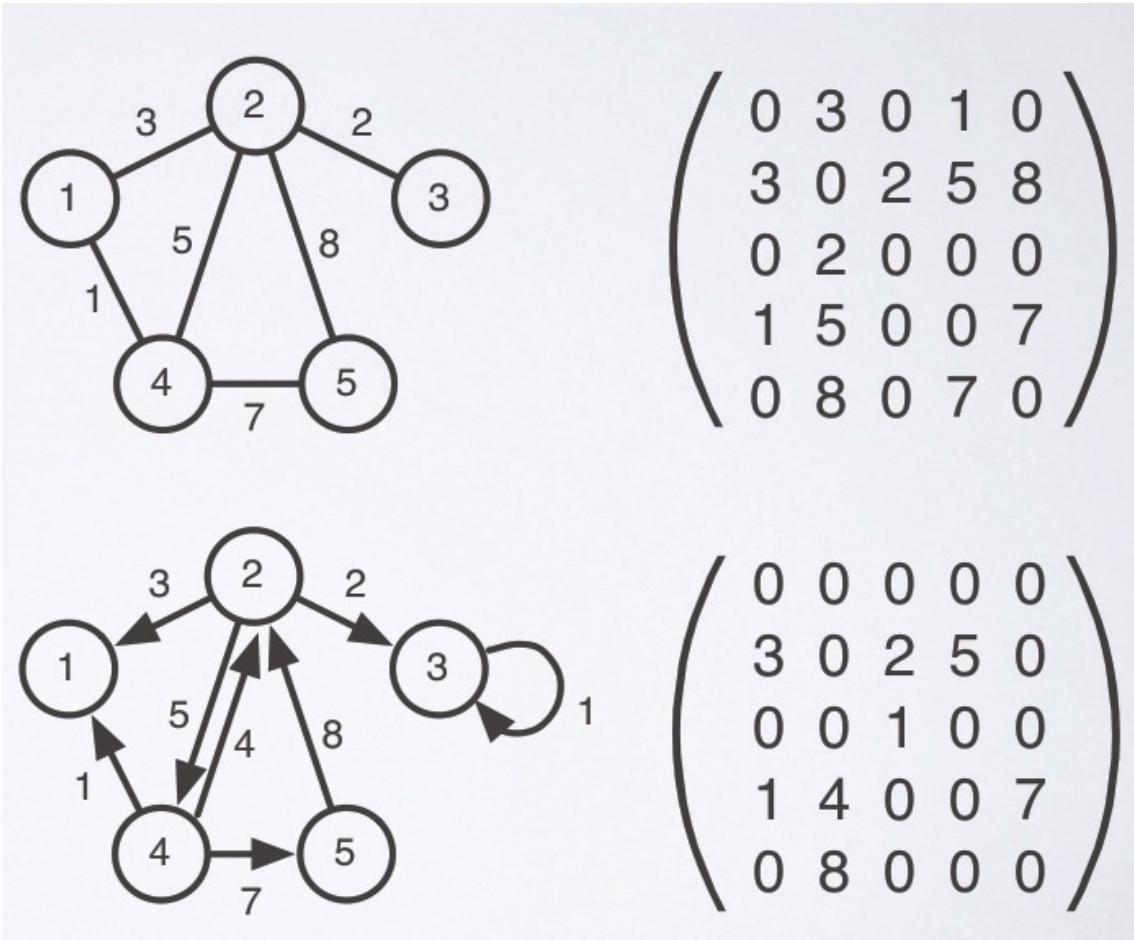
Networks – basics

Networks also can also be **weighted**, this means that some connections are stronger than others. Weighted network measures are harder to define.



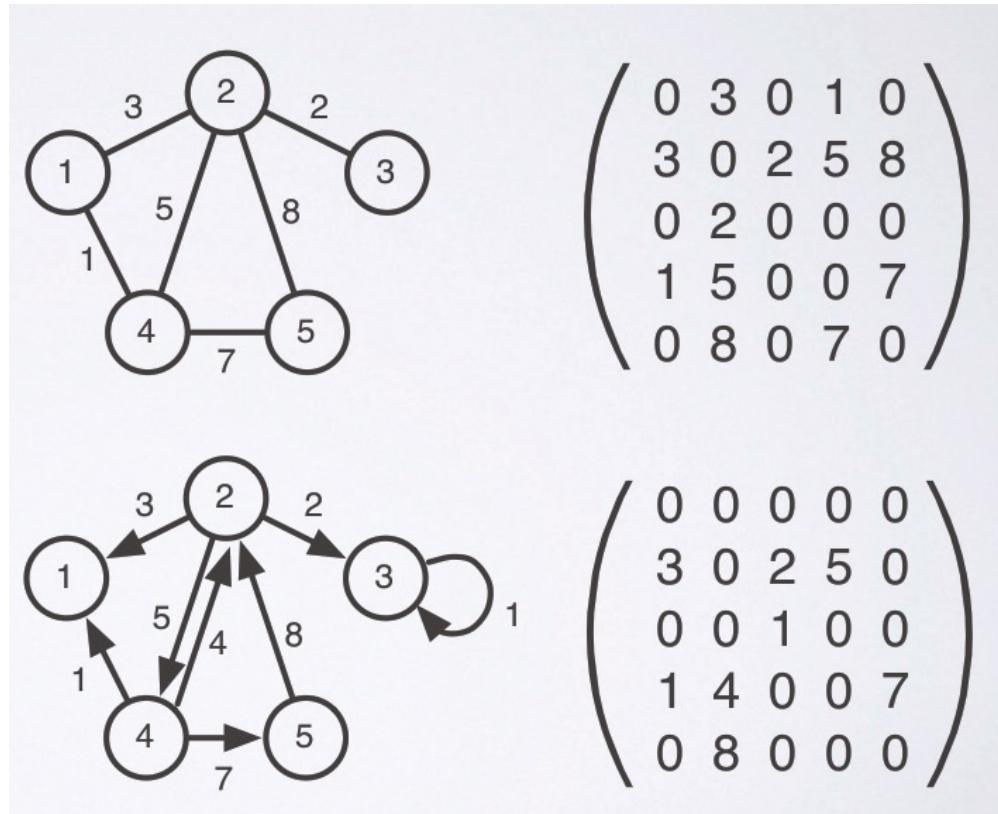
Networks – basics

Can also be weighted and directed.



Networks – basics

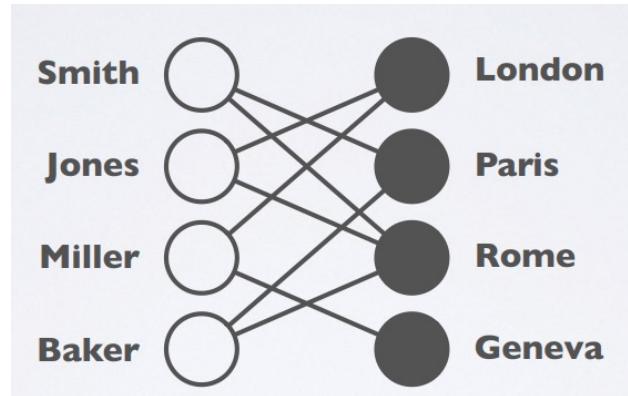
Can also be weighted and directed.



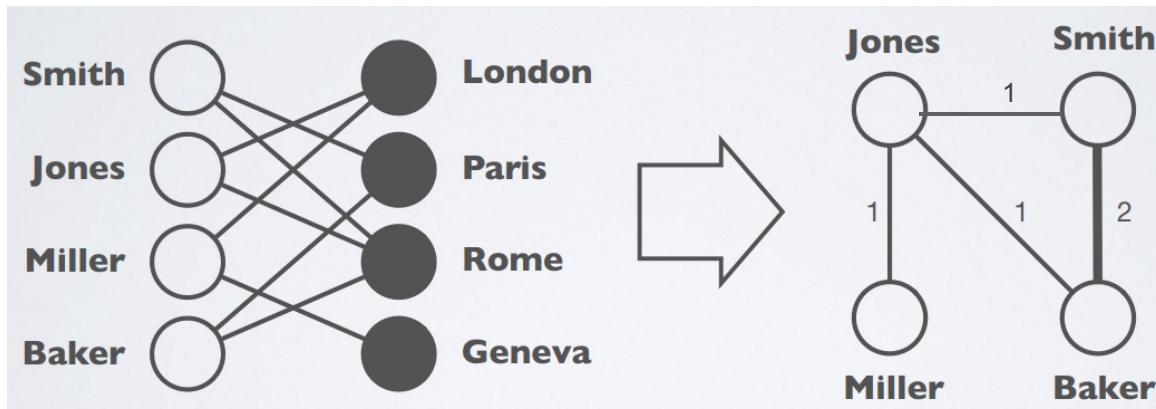
For a weighted network
we get a real adjacency
matrix, often denoted
as w_{ij} .

Networks – basics

Bipartite networks: we have two types of nodes, and edges can only connect nodes of different types.



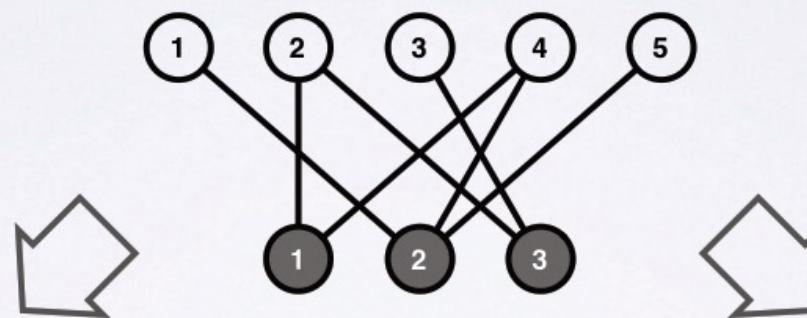
We can perform a ‘one-mode projection’ onto either node type. What this means is that we create a weighted network of one of the node types.



An edge in this new network signifies how many nodes of the other type are shared.

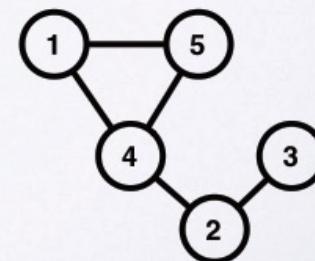
Networks – basics

Bipartite networks: In terms of the bipartite adjacency matrix the two projections are simply $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$:



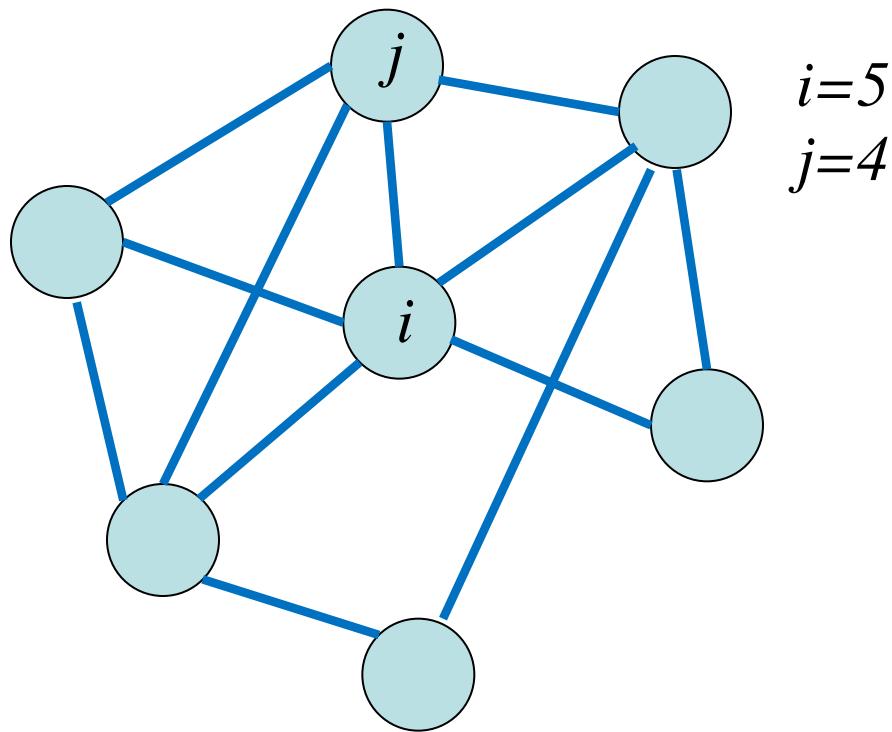
$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 2 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix}$$



Networks – basics

The number of connections a node has is its **degree** k .



In an undirected network we can write the degree in terms of the adjacency matrix as:

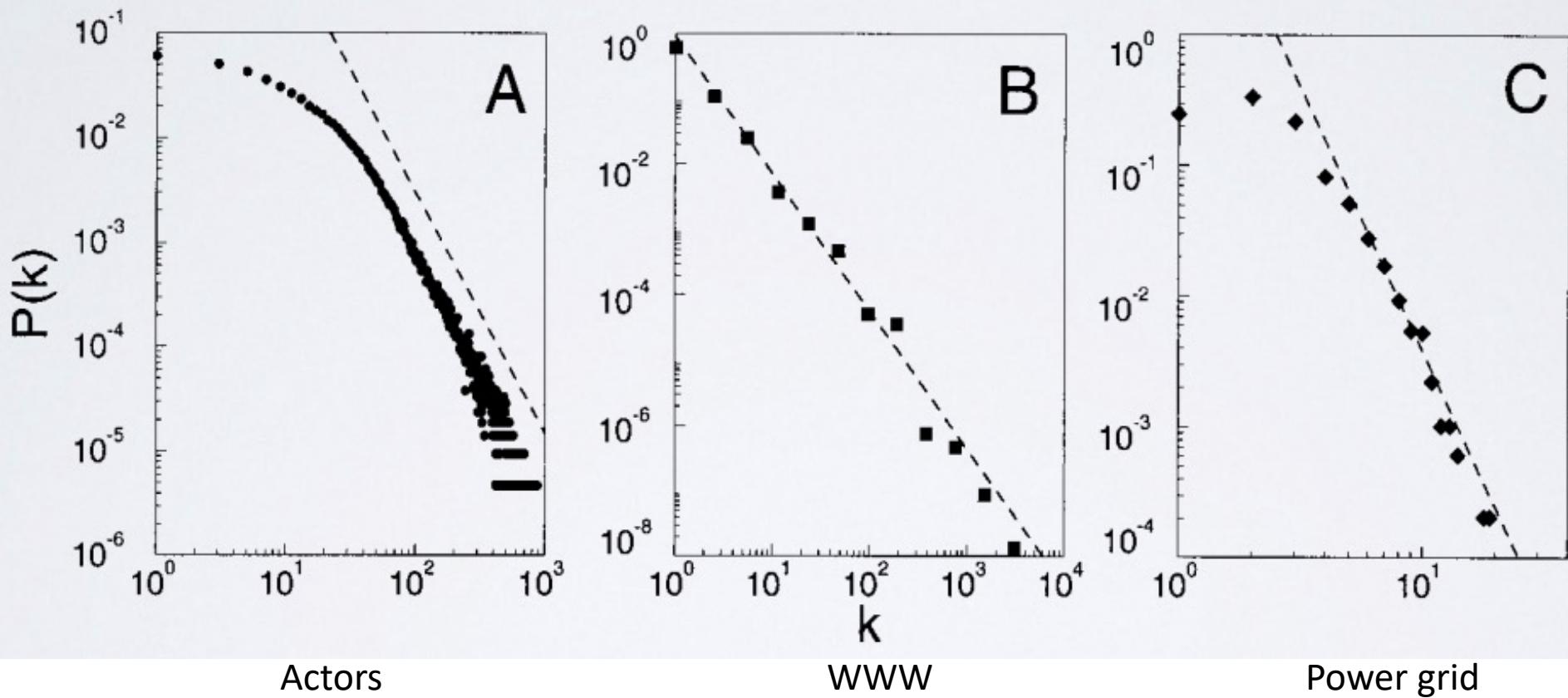
$$k_i = \sum_j a_{ij}$$

In a directed network we have to distinguish: the in-degree (out-degree) of a node i are the number of directed edges pointing to (from) node i :

$$k_i^{(in)} = \sum_j a_{ji} \quad k_i^{(out)} = \sum_j a_{ij}$$

Networks – basics

Many real-world networks have similar properties, such as a scale-free degree distribution.
Barabasi, A. & Albert, R. Emergence of scaling in random networks. Science 286, 509–512 (1999)



$$P(k) = k^{-\gamma}$$

Networks – basics

From Barabasi, A. & Albert, R.

Probabilistic network growth model produces scale-free networks.

Question sheet exercise – make a script to add new node and attach it to m existing nodes, where the probability of attaching it to a particular node i is:

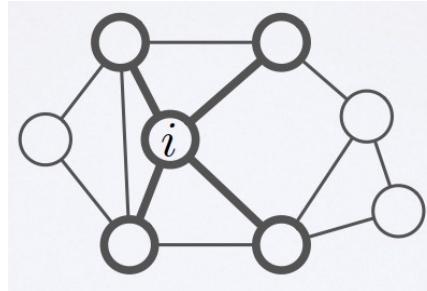
$$p_i = \frac{k_i}{\sum_j k_j}$$

What is the degree distribution for the Barabási-Albert model? You can work this out analytically by considering k as a continuous variable.

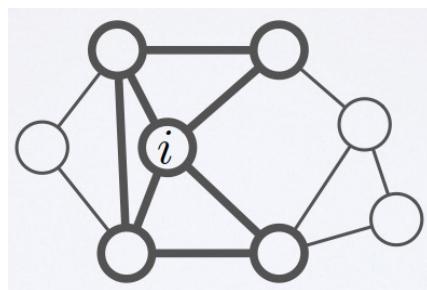
Networks – basics

Clustering coefficients: c_i measures how densely the local neighbourhood of node i is connected.

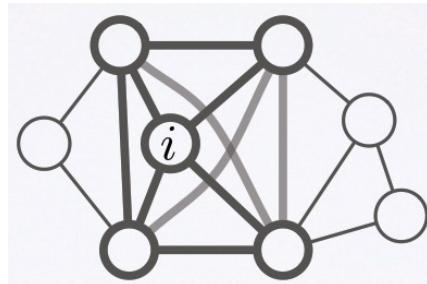
Consider the neighbourhood of a node i with four neighbours.



We then consider how many pairs of these neighbours are connected - in other words we count triangles that contain node i .



Finally we consider how many triangles there could have possibly been. With four neighbours there are six possible triangles.



The clustering coefficient is the fraction of possible triangles that is actually realised. In this case it would be $c_i = 3/6 = 1/2$

Networks – basics

Recap.

We have seen the basic measures for describing local **structural** characteristics of individual nodes and edges:

The *degree* is a local measurement. It is one way of defining the importance of a node.

The *clustering coefficient* is another local measurement, and measures how densely the neighbourhood of a node is connected.

There are also global characteristics for the whole network:

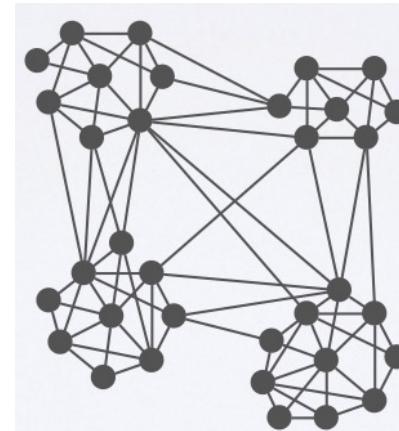
The *average shortest path length* is an example of a global measurement.

There are measures that connect local to global:

The *betweenness* is a measure of the importance of an individual node or edge, but it depends on the global properties of the network.

Communities: more densely connected parts of a network.

And then there is **dynamics on networks** – each node has a state, and “rules” to update based on the states of its connected neighbours.



Networks in biology

If limited to one “role” per protein, the roughly 30,000 Human genes would have limited utility. The key to diversity of behaviour is:

- (i) the combinatorial power from many genes acting in concert;
- (ii) the time profile of expressing and suppressing genes;
- (iii) localization/compartmentalization of proteins in different locations;
- (iv) interactions with the resources and stimuli from the environment.

Various forms of behaviour can then emerge from a palette of few elements.

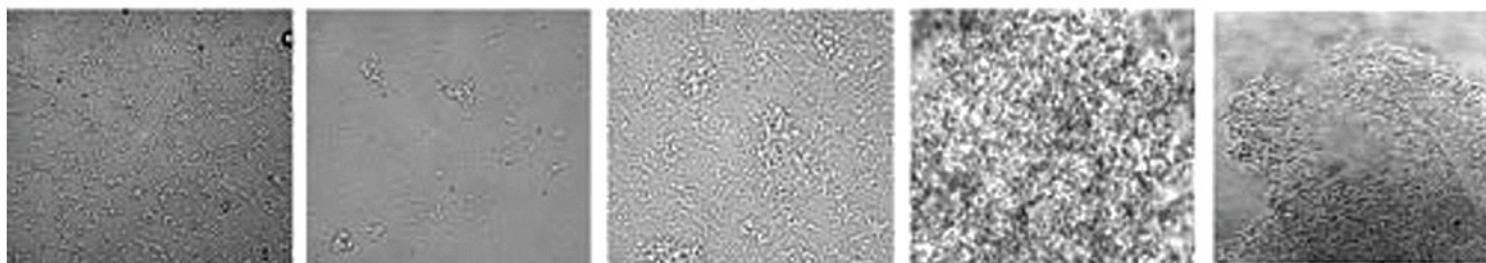
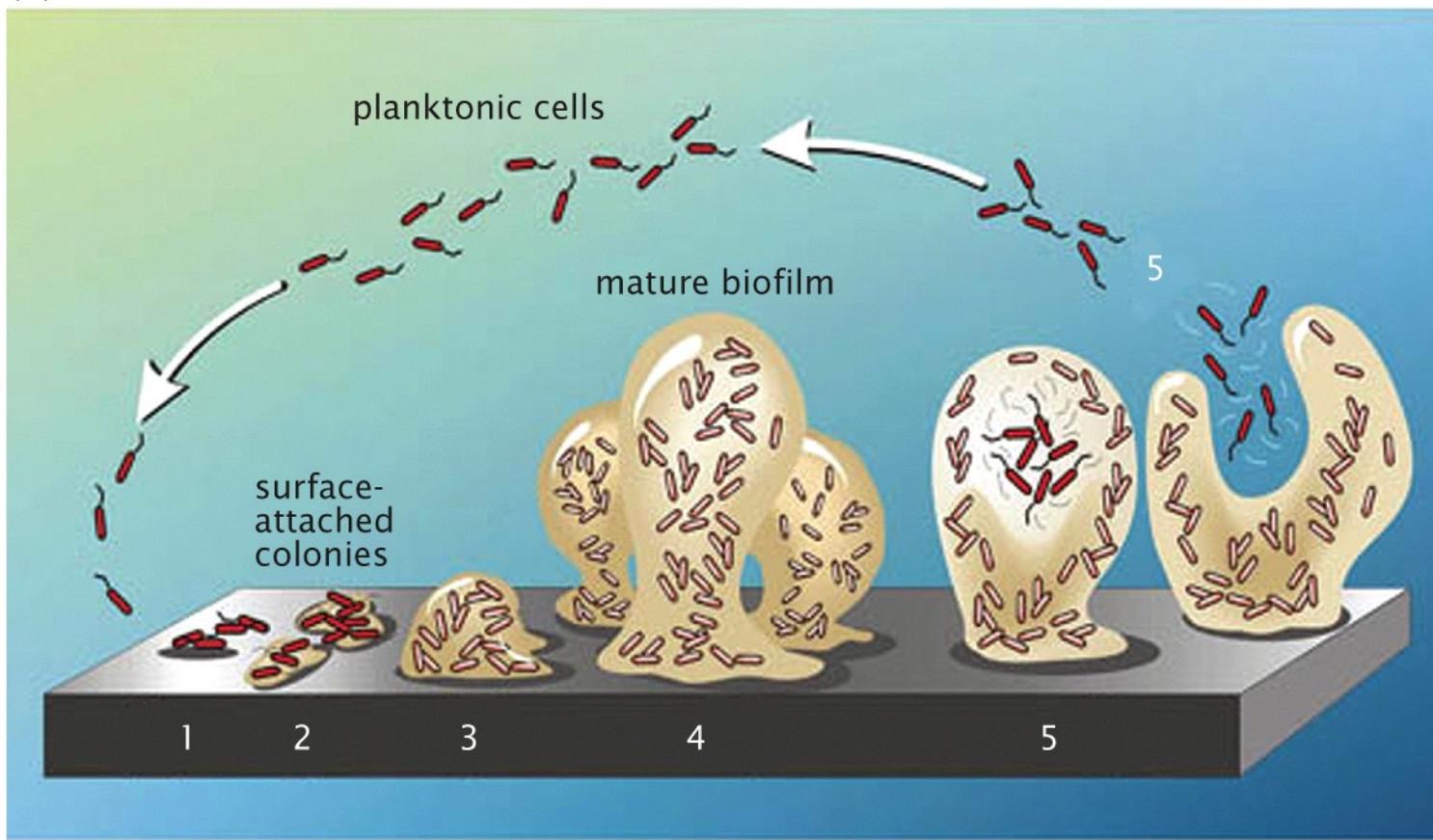
These associations, and interdependencies (connections across scales), are at the heart of a quantitative and fundamental understanding of biological processes.

There is a lot of “physics” in there, both “old” and “new”, and challenges of both theoretical and experimental nature.

A network representation can be a useful tool to study complex systems. The nodes of the network can be a set of genes, or proteins, or metabolic products (sugars, lipids) in the cell. Links identify dependencies.

Examples of processes that go beyond “one gene = one effect” while still appearing tractable to us...

(A) Bacteria – many can take “decision” to make a biofilm



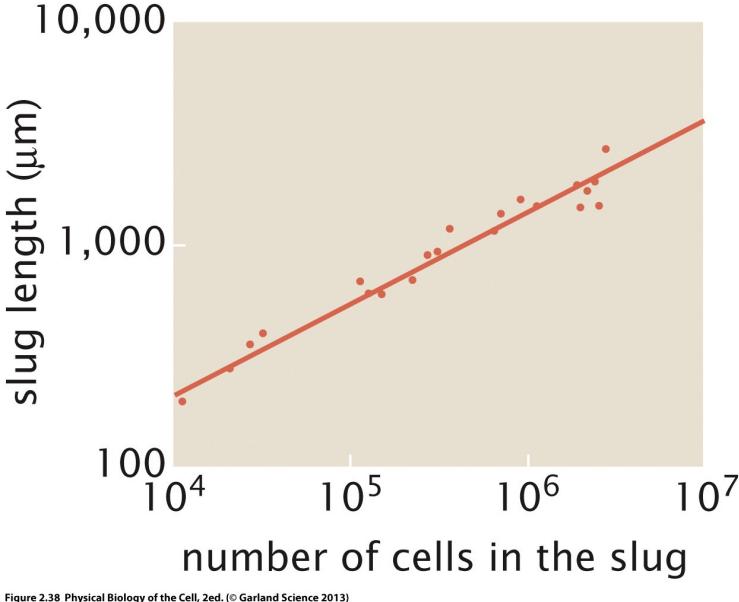
(B)

20 μm

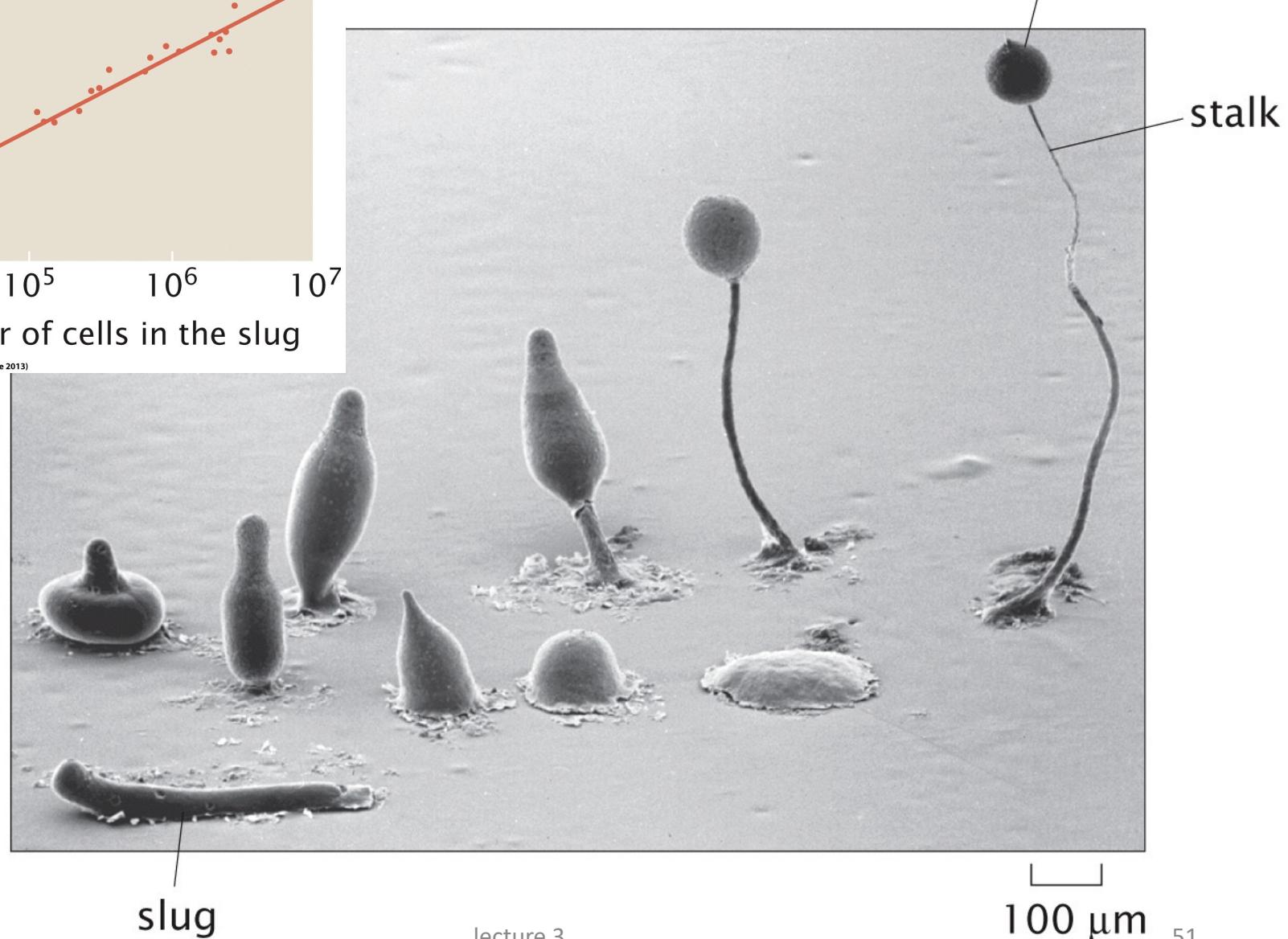
Figure 2.36 Physical Biology of the Cell, 2ed. (© Garland Science 2013) lecture 3

Examples cont....

Figure 2.38 Physical Biology of the Cell, 2ed. (© Garland Science 2013)



Dictyostelium – can make “decision” to differentiate cells

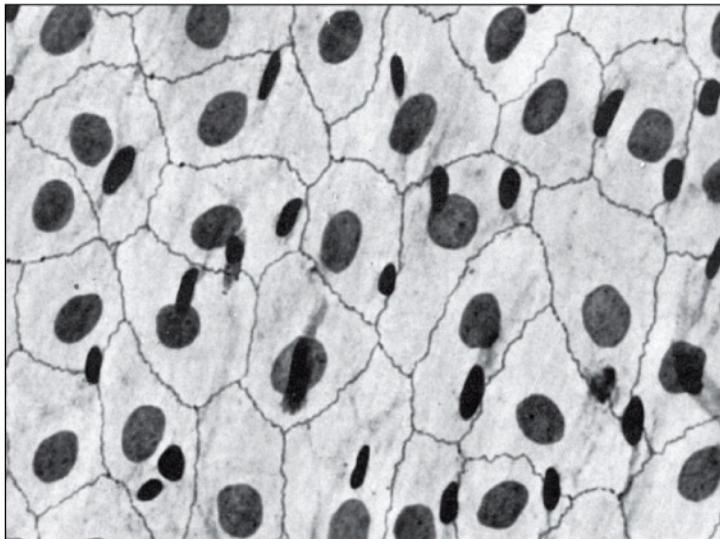


lecture 3

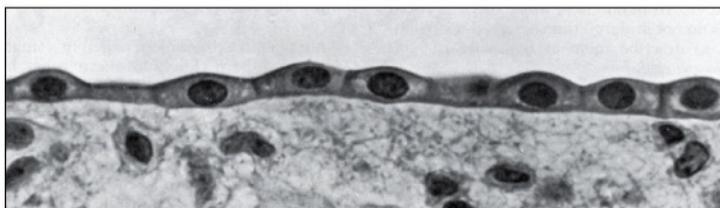
Figure 2.37 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Epithelium – a system that is under constant renewal.

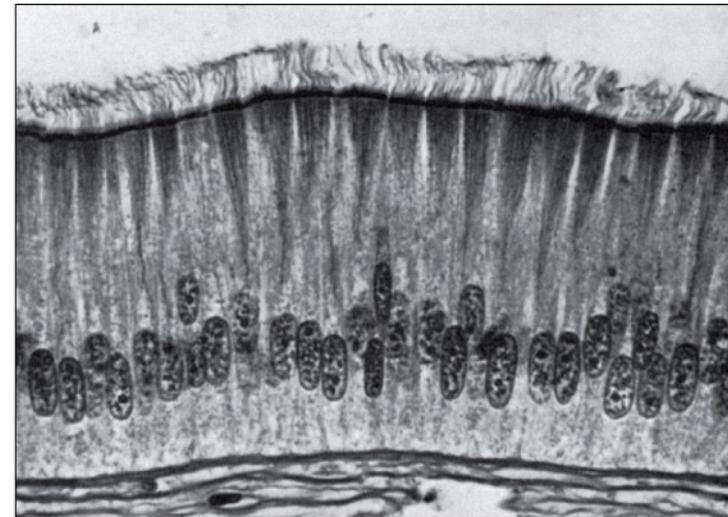
Examples cont.....



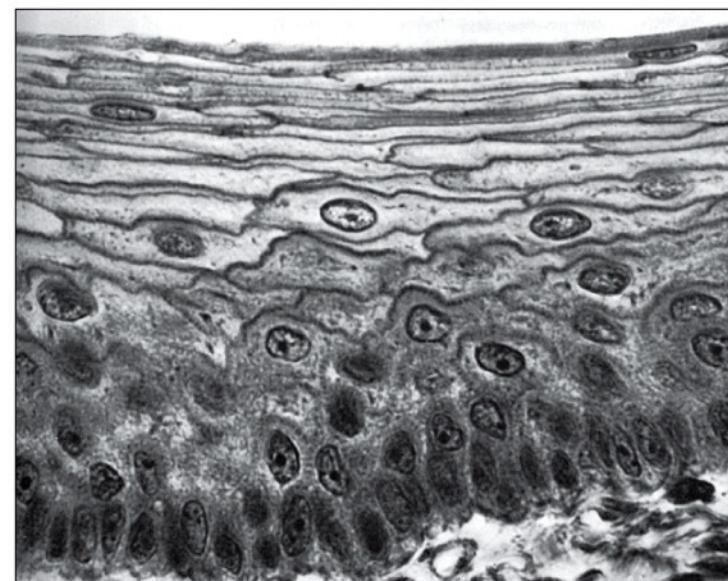
(A)



(B)



(C)



(D)

~10 μm

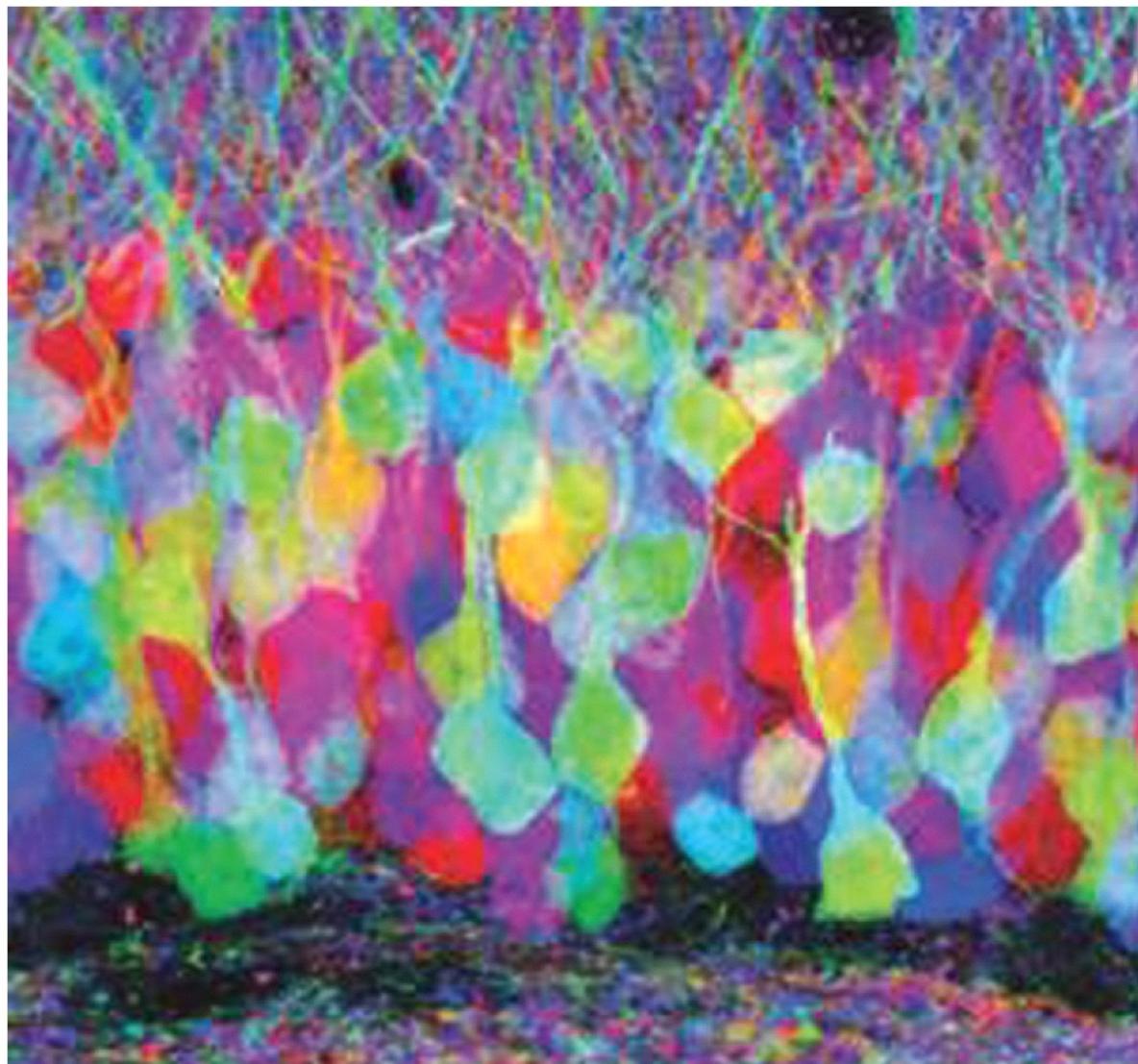
Examples cont.....

Intestinal villi – epithelium meets mechanical instability



Figure 2.41 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Architecture of neurons



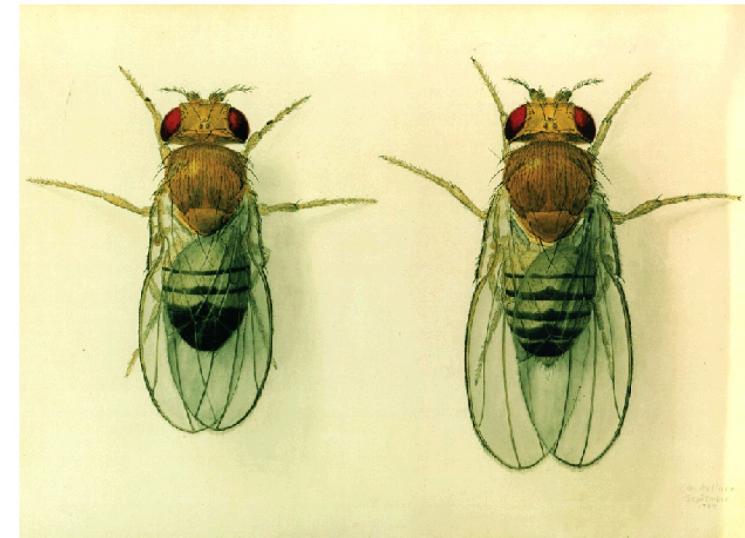
10 μm

lecture 3

Figure 2.42 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

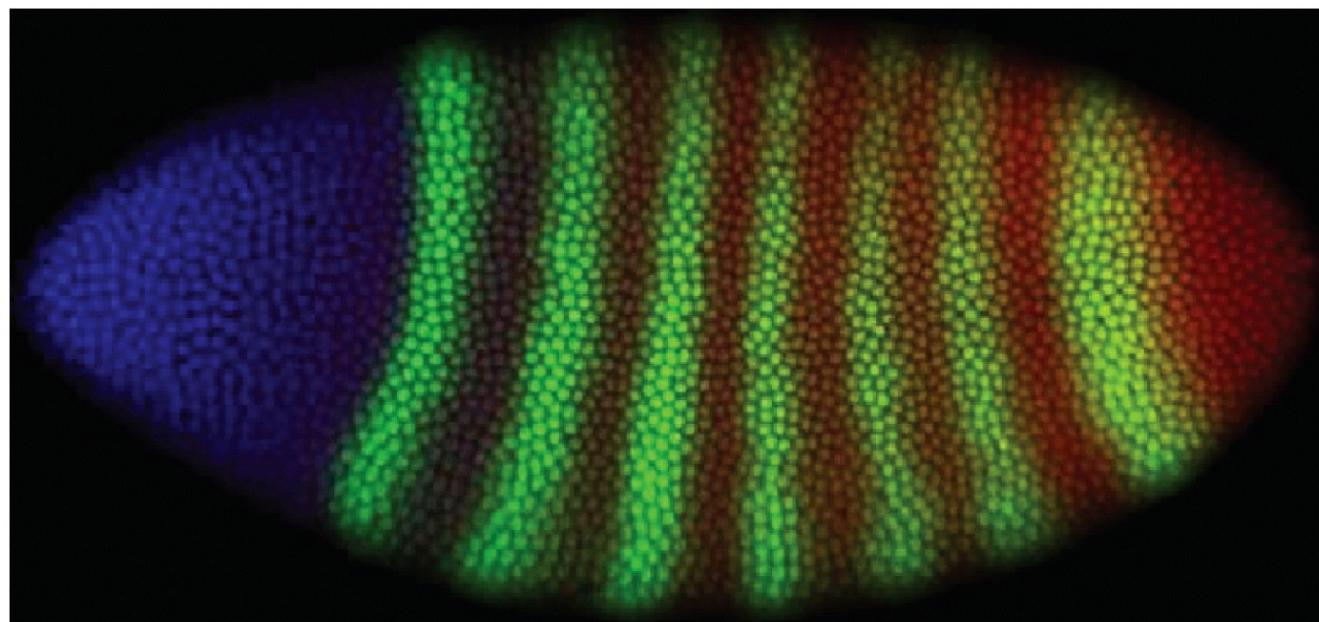
Drosophila

– one of the model systems to study development



© Garland Science
2013

“Turing patterns”



100 μm

lecture 3

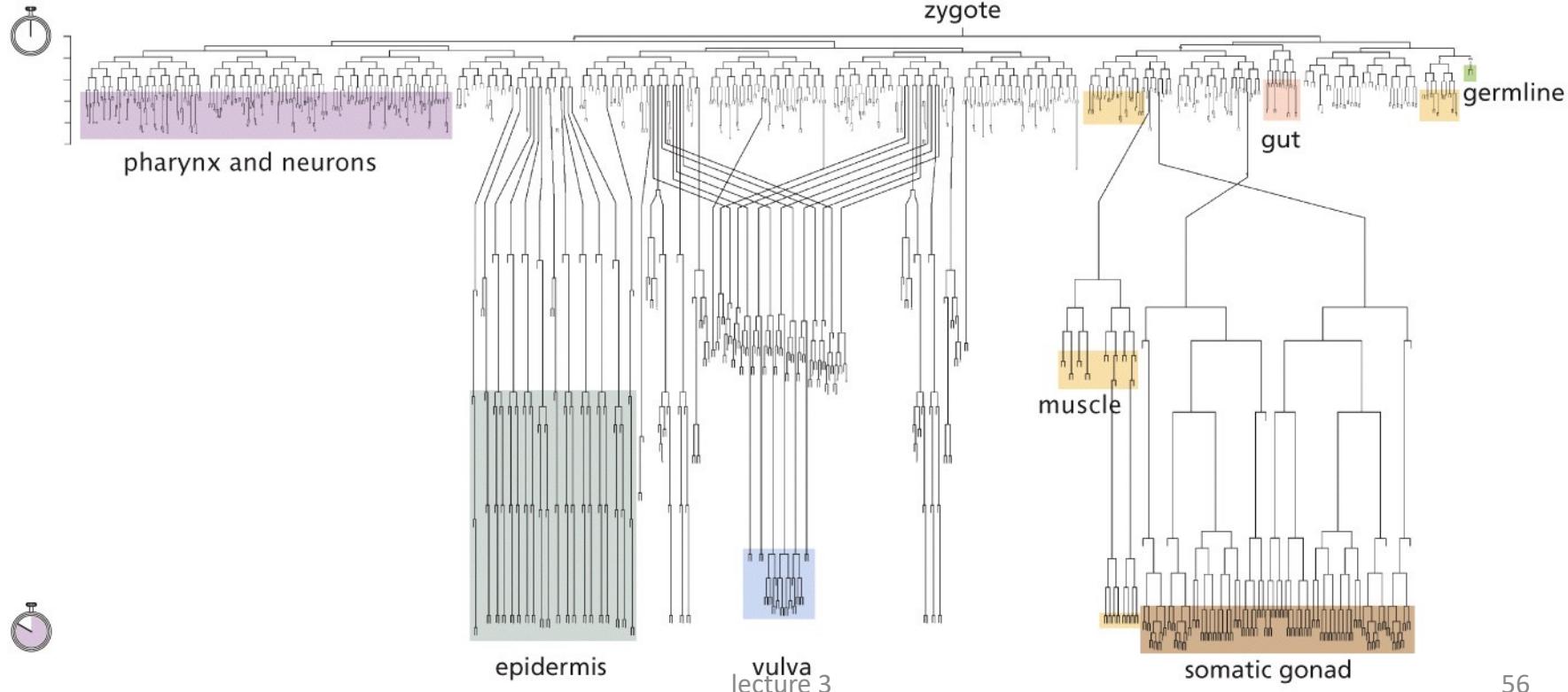
55

C.Elegans – development and
“connections” are all “deterministic”:
959 cells in adult hermaphrodite
1031 in adult male



Figure 2.46 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

hours



Examples cont.....

Figure 2.47 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Examples cont.....

The eukaryotic cell cycle

(bacteria are much simpler)

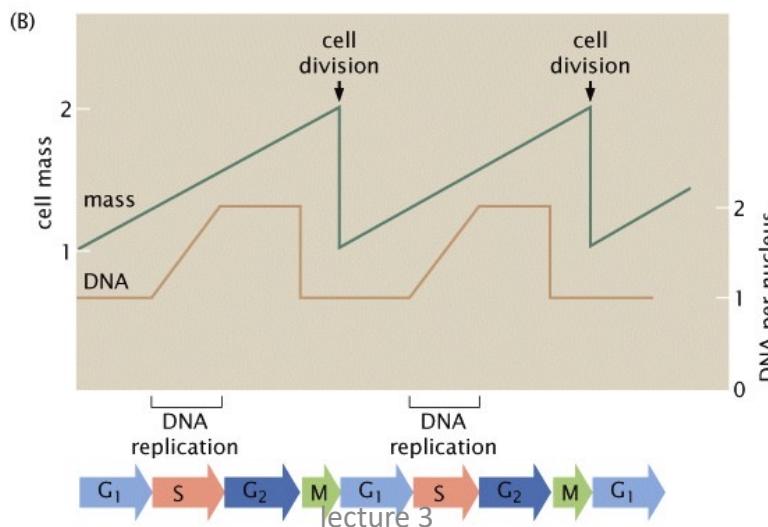
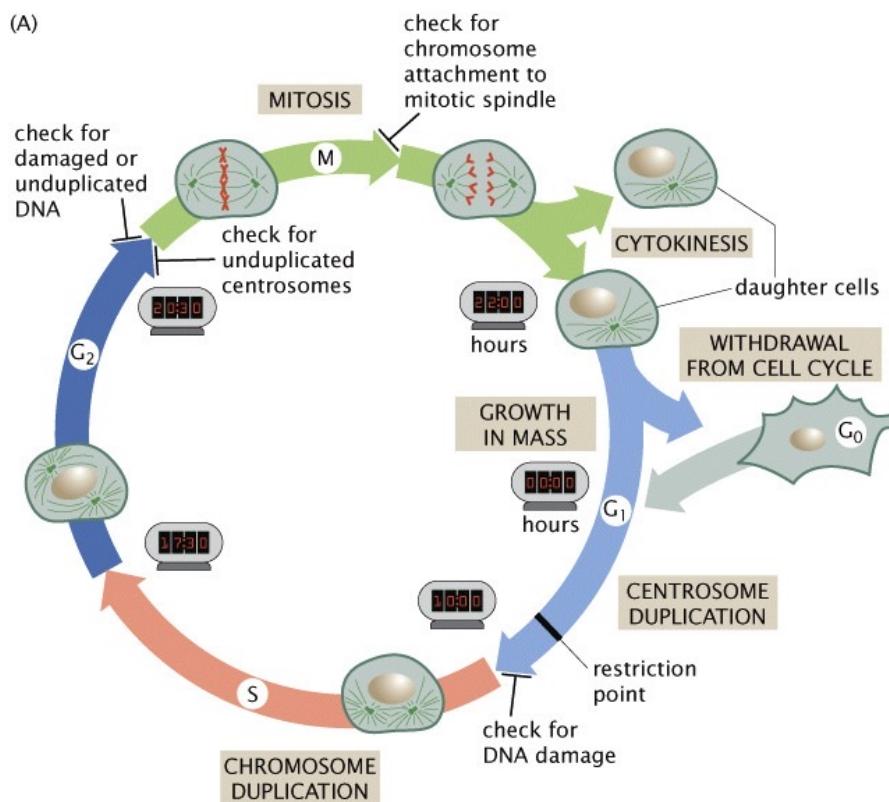
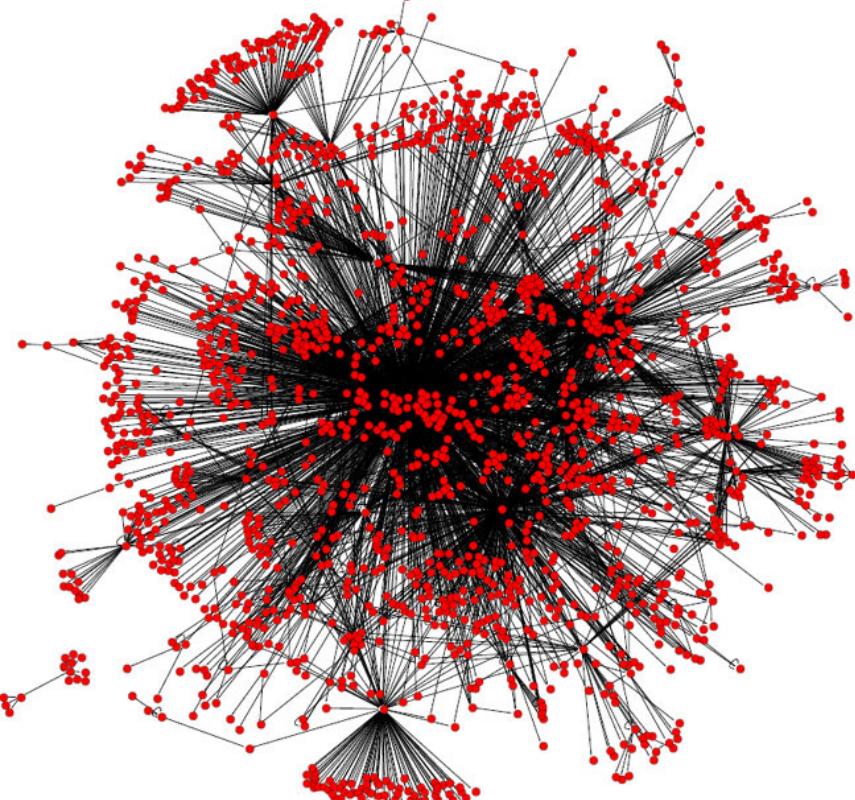


Figure 3.21 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Representation of *E.coli* transcriptional regulatory **network**, from the public database regulonDB



Genetic **network** controlling early development in sea urchin embryo

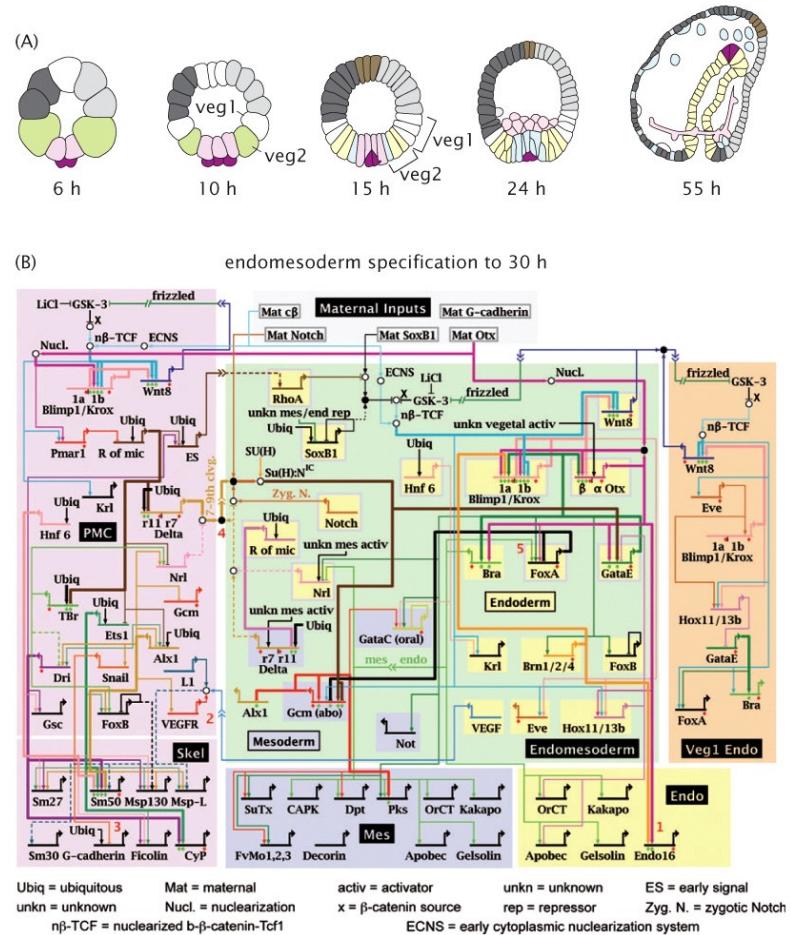


Figure 19.1 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

The transcription networks of *E. coli*, or yeast (yeast network in notes handout) are very complex.

How can we tell if buried in this information there are interesting statistical properties?

One possible route to extracting information from such data is to look for specific motifs, subgroups of several nodes, that can cooperate in simple functions (e.g. a feedforward loop). A particular motif can be significant if it appears more (or less) frequently than expected.

In these complex networks, physicists have indeed discovered an excess of certain “network motifs”: functional *modular* units. Modularity probably make sense considering this functionality has arisen from evolution.

Especially in fast growing organisms, like bacteria, the regulation networks are constantly evolving, and selected under constant evolutionary pressure – this kind of logic tends to suggest that there have to be good reasons for any statistical property that can be found.

Need a simple model whose expectations can be compared with biological data.

Random graphs, introduced by Erdos and Renyi, serve this purpose:

The ER model consists of N nodes,

with any pair connected at random and independently, with probability p .

“a mathematician is a device to turn coffee into theorems”

Note that you can obtain the expected number of subgraphs of n nodes and l links as a product of the number of ways of picking n points and connecting them with l links, and a factor that accounts for the number of ways of connecting the points into the desired graph:

$$\mathfrak{N}(n, l) = \binom{N}{n} p^l \times \frac{n!}{(\text{symmetry factors})}$$

For example, there are $n!/2$ ways to string n points along a straight line with $l = (n - 1)$, and the expected number of such linear pathways is:

$$\mathfrak{N}(n \text{ in a line}) = \frac{N!}{(N - n)!} \frac{p^{n-1}}{2}$$

[the numbers of cyclical graphs, and complete graphs, are worked out in notes]

Average degree of a node is $\langle k \rangle = 2l / N = p(N-1) \sim pN$ and the degree distribution is:

$$P_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$$

The autoregulation network motif

In *E. coli* transcription network there is an excess of self-edges, the vast majority of which are repressors that implement negative autoregulation.

How can this conclusion be reached?

Need to compare with the expected number of self-edges in a random network.

With N nodes, there are $N(N - 1)/2$ possible pairs of nodes that can be connected by an edge. Each edge can point in one of two directions, for a total of $N(N - 1)$ possible places to put a directed edge. An edge can also begin and end at the same node, so there are a total of N possible self-edges. Total number of edges is thus: $N(N - 1) + N = N^2$.

In the ER model, the E edges are placed at random in the N^2 possible positions, so each possible edge position is occupied with probability $p = E/N^2$.

Probability of having k self edges in an ER network:

- a self edge needs to choose its node of origin as a destination, out of the possible N , so $p_{self} = 1/N$
- Since the E edges are placed at random, the probability of having k self edges is approx binomial:

$$P(k) = \binom{E}{k} p_{self}^k (1 - p_{self})^{E-k}$$

- average number of self-edges is E times the probability of being a self edge, i.e.

$$\langle N_{self} \rangle_{rand} = E p_{self} = E/N$$

- with a standard deviation that is approximately (because binomial approx Poisson) the square root of the mean, so

$$\sigma_{rand} \simeq \sqrt{E/N}$$

- Alon considers data where $N=424$, and $E=519$, and in which there are 40 self edges (34 are repressors). The random graph expectation is

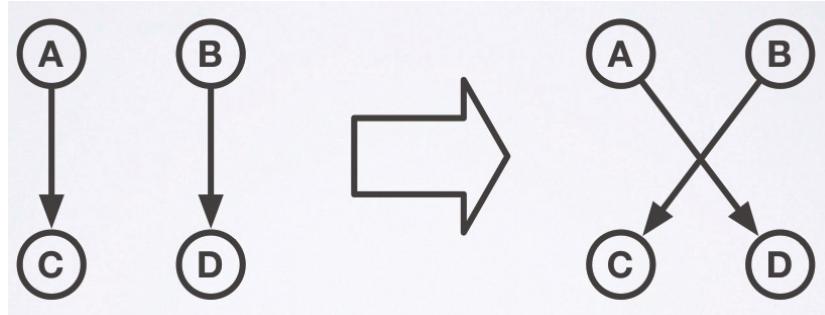
$$\langle N_{self} \rangle_{rand} = E/N = 1.2 \text{ with } \sigma_{rand} \simeq \sqrt{1.2} = 1.1$$

- => Negative autocorrelation is indeed way more common than “random”
- We shall see later why it is a useful feature in regulation of gene expression.

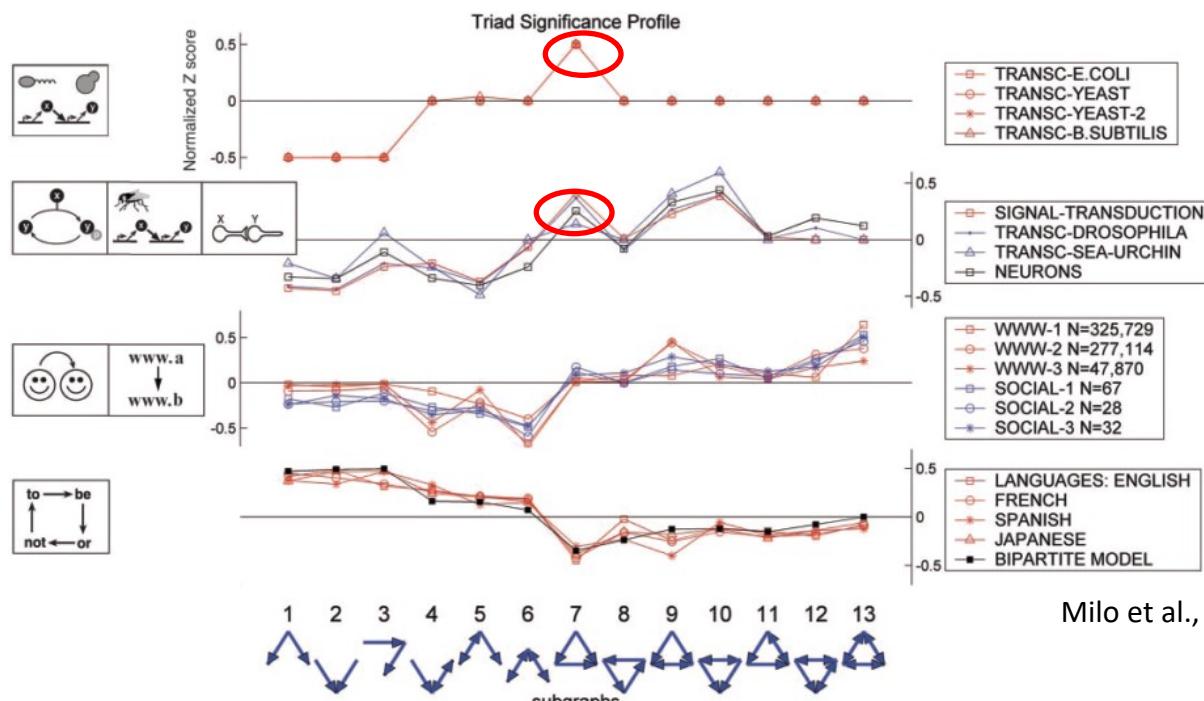
Very significant!

Identifying motifs in network data

To evaluate whether their number is higher than would be expected by chance, the networks are randomized by swapping two inputs or two outputs.



This gives rise to a network with the same in- and out-degrees as the original network.



Frequency signatures of network motifs classify networks into superfamilies.

The percolation transition in random networks

As one increases the average connectivity, there is a percolation transition. This is a “geometric” phase transition, with the formation of a macroscopic cluster that takes up a finite (not small) fraction of nodes.

For the random graph this transition can be treated approximately, for large N :

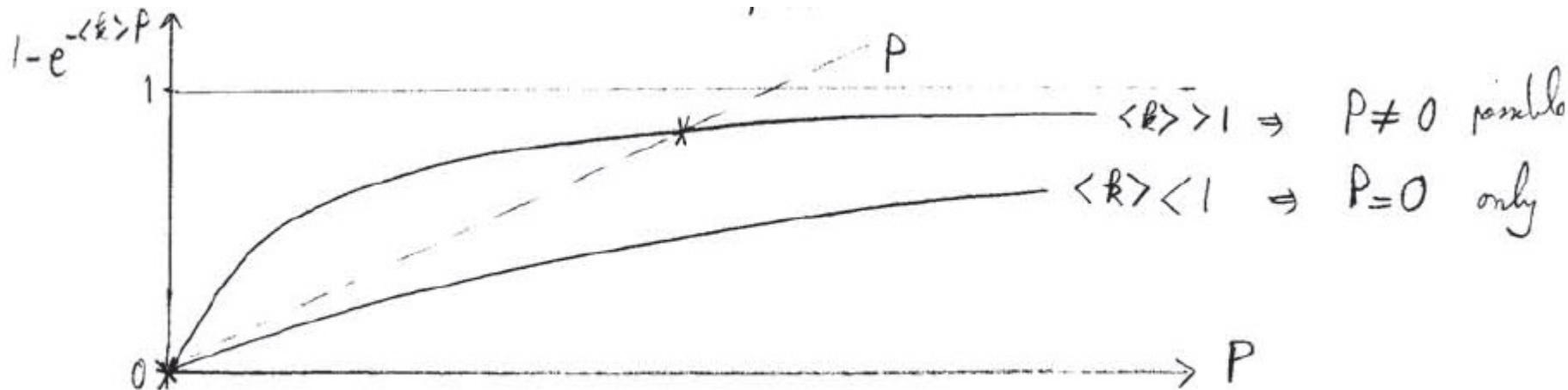
- p is the probability that a given pair of nodes is linked.
- Above the percolation transition, the number of nodes M in the largest cluster also goes to infinity, proportionately to the number of nodes, such that there is a finite percolation probability $P(p) = \lim_{N \rightarrow \infty} \frac{M}{N}$ (the probability for a node to belong to the infinite cluster)
- $P(p)$ can be calculated from a self-consistency argument.
- Take a particular site and consider the probability that it is not connected to the infinite cluster. This is the case if none exist of the $(N - 1)$ edges emanating from this site potentially connecting it to the large cluster.
- A particular edge connects to the infinite cluster with probability $pP(p)$ (that the edge exists, and that the adjoining site is on the large cluster), and hence

$$\begin{aligned} 1 - P(p) &= (\text{prob of no connections to any edge})^{N-1} \\ &= (1 - pP)^{N-1}. \end{aligned}$$

- We take the limit $N \rightarrow \infty$, but also at the same time $p \rightarrow 0$ such that we keep $p(N - 1) = \langle k \rangle$, where $\langle k \rangle$ is the (finite) average number of edges per node.
- then the equation above can be expressed as

cont.... The percolation transition in random networks

- Solve graphically $1 - P(p) = e^{-\langle k \rangle p}$



- Two regimes! If $\langle k \rangle < 1$, then the only solution is $P=0$.
- If $\langle k \rangle > 1$, then there is a non-zero P solution, i.e. an infinite cluster.
- Close to the percolation transition, at $\langle k \rangle_c$, P is small and we can expand the last expression, to get
$$P \approx \frac{2(\langle k \rangle - 1)}{\langle k \rangle^2} \approx 2(\langle k \rangle - 1)$$
- Percolation on random graphs (& lattices) has all kinds of applications: sol-gel transition in material networks, epidemics, social sciences, behaviour of the WWW.