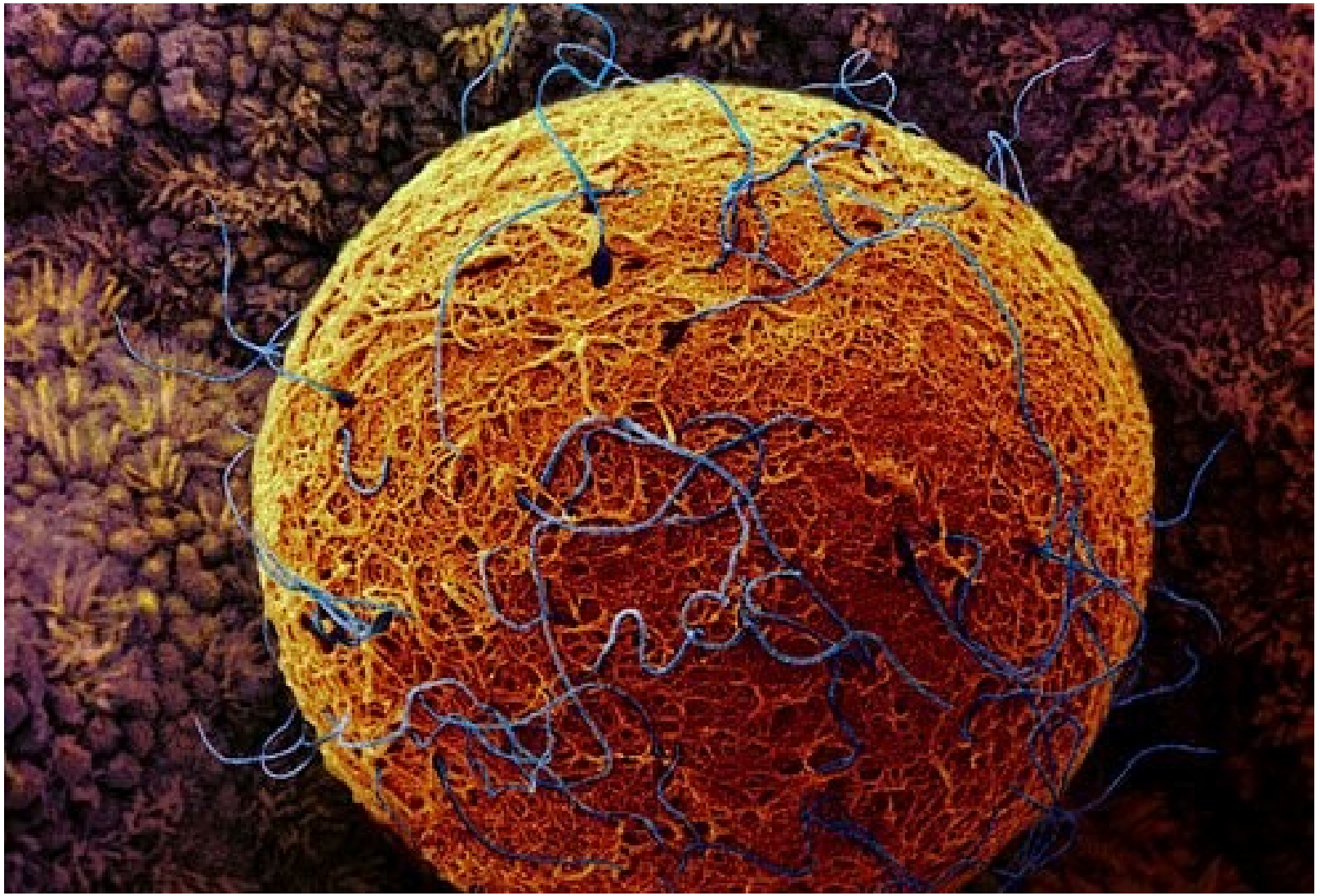
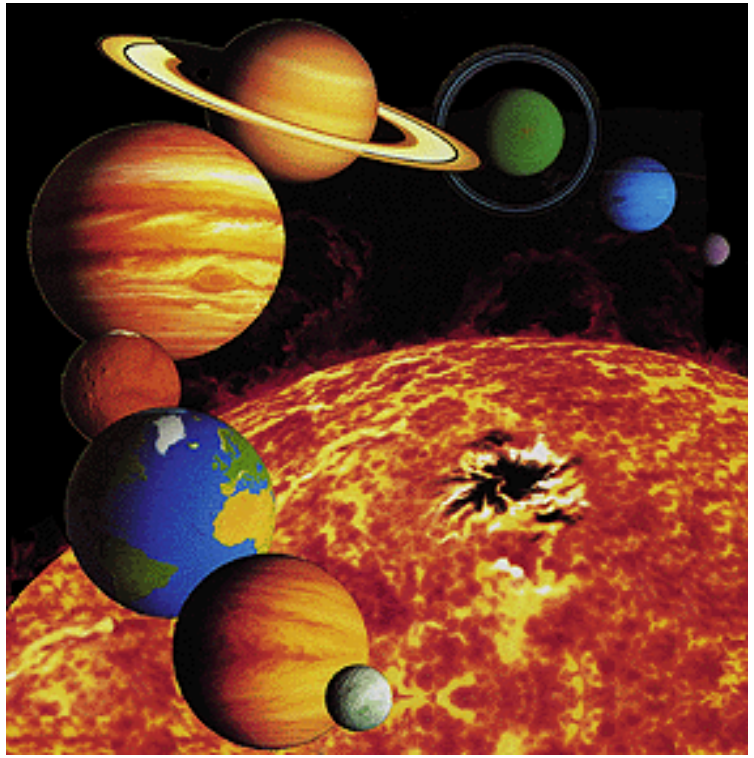
Computational biology

Il materiale di seguito riportato è destinato solo ed unicamente all'attività didattica nell'ambito del CdLM in Biotecnologie della Riproduzione, Università di Teramo

fertility

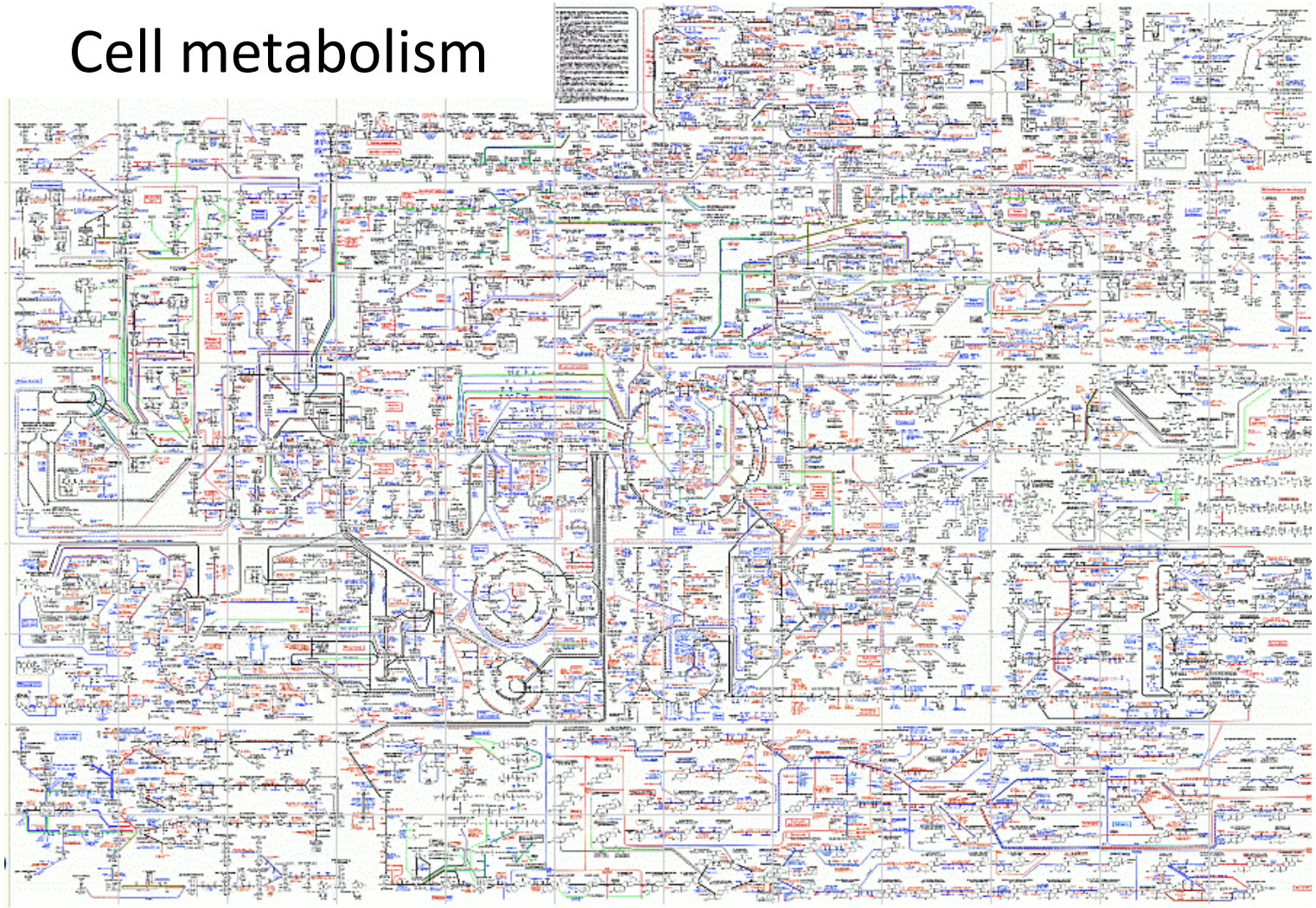


Complicated vs. complex



Biological complexity

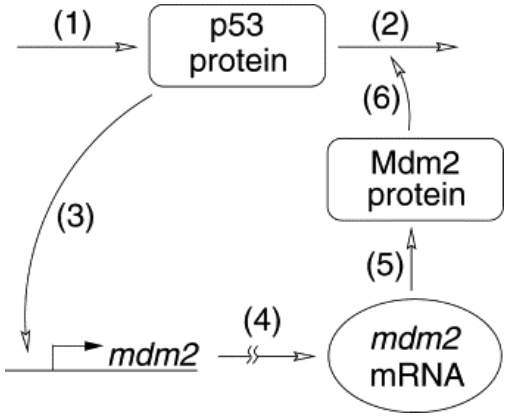
Cell metabolism

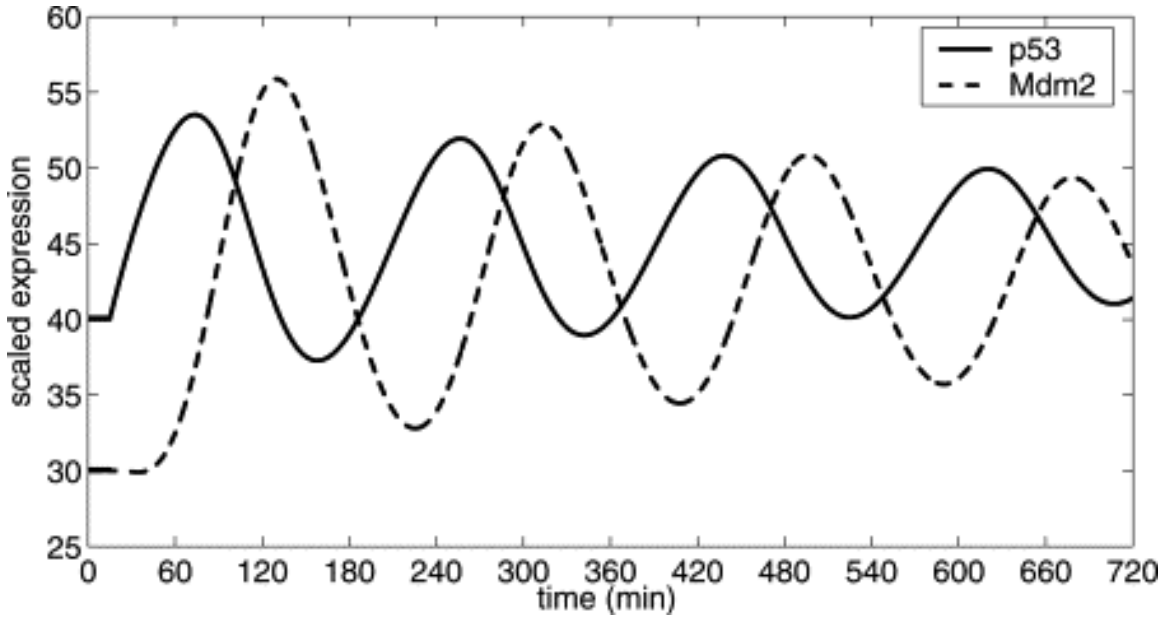


Complexity:

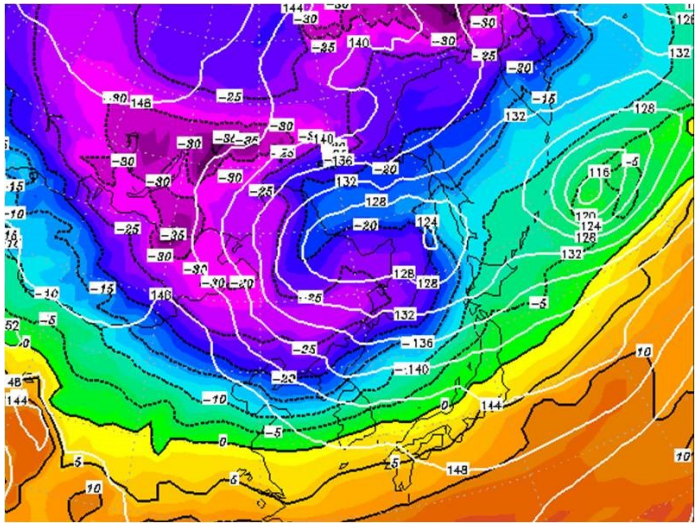
non-linearity of interactions:

An example…

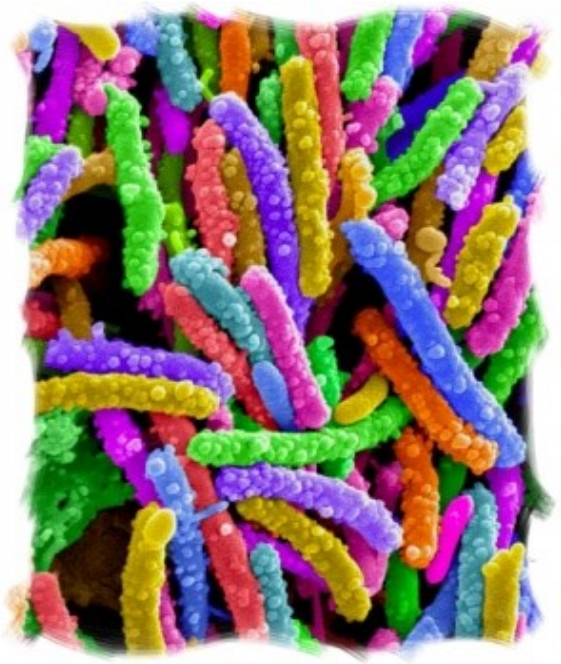

(1)
p53 protein
(2)
(6)
Mdm2 protein
(3)
(5)
(4)
mdm2
mdm2 mRNA

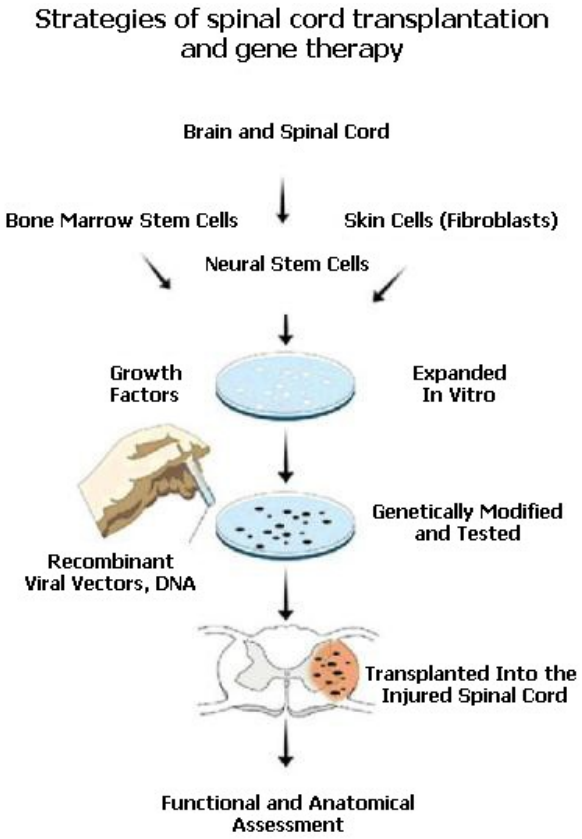

p53
Mdm2
scaled expression
time (min)
60
55
50
45
40
35
30
25
0
60
120
180
240
300
360
420
480
540
600
660
720

unpredictability



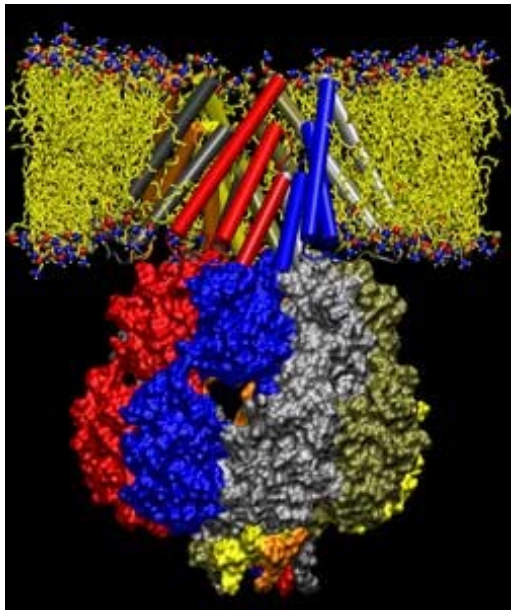
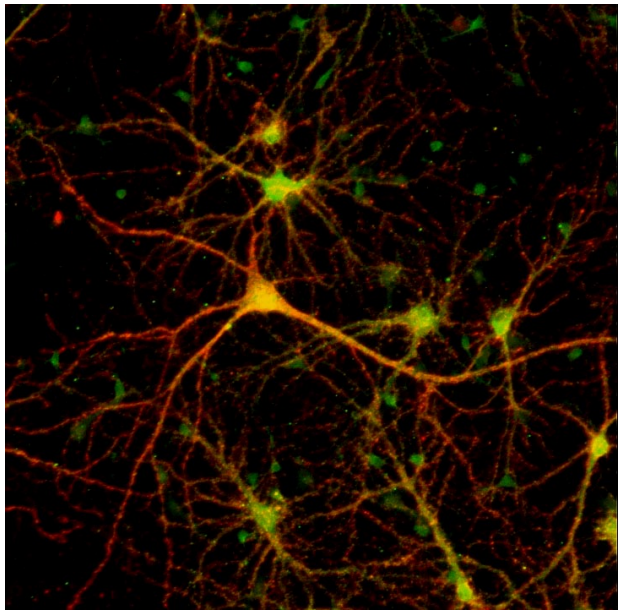
unpredictability





Butterfly effect

Emergence of proprieties

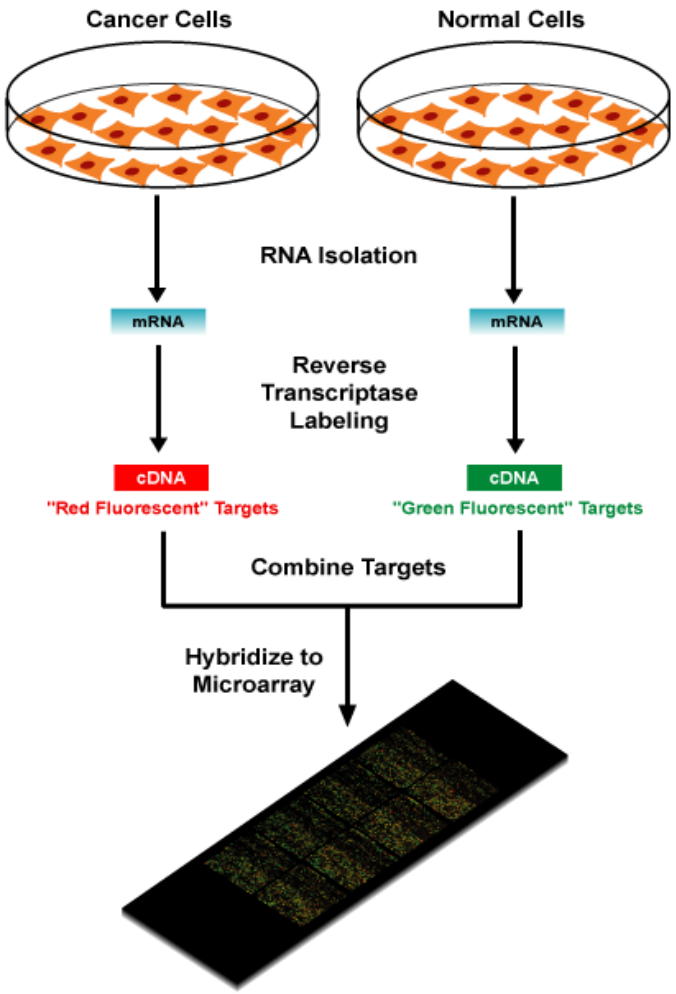

The whole is more (different) than the sum of the individual components

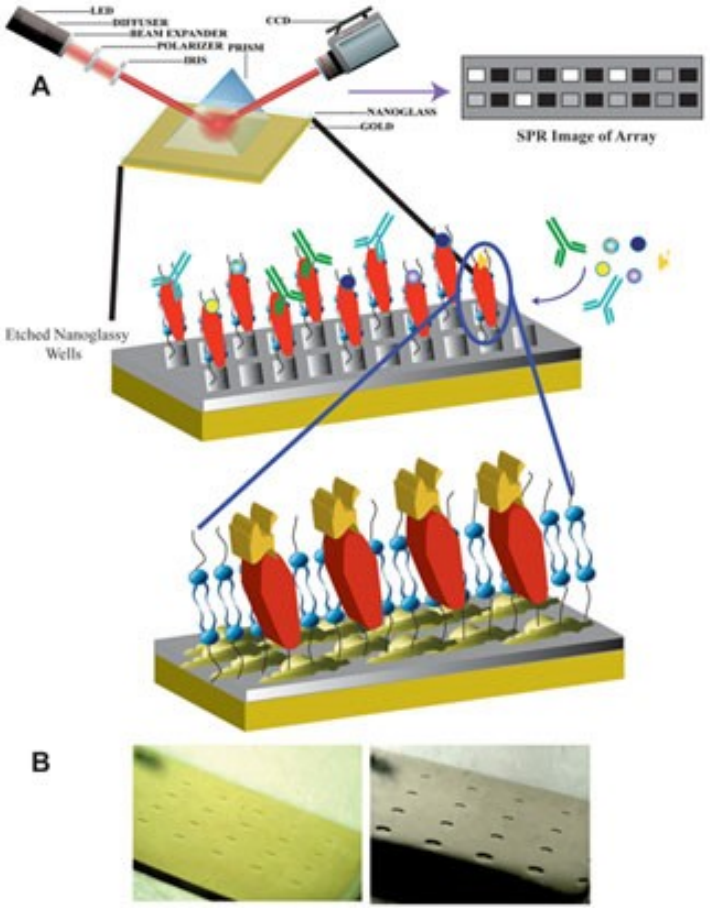
Human Genome Project

Begun formally in 1990, the U.S. Human Genome Project was a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the completion date to 2003. Project goals

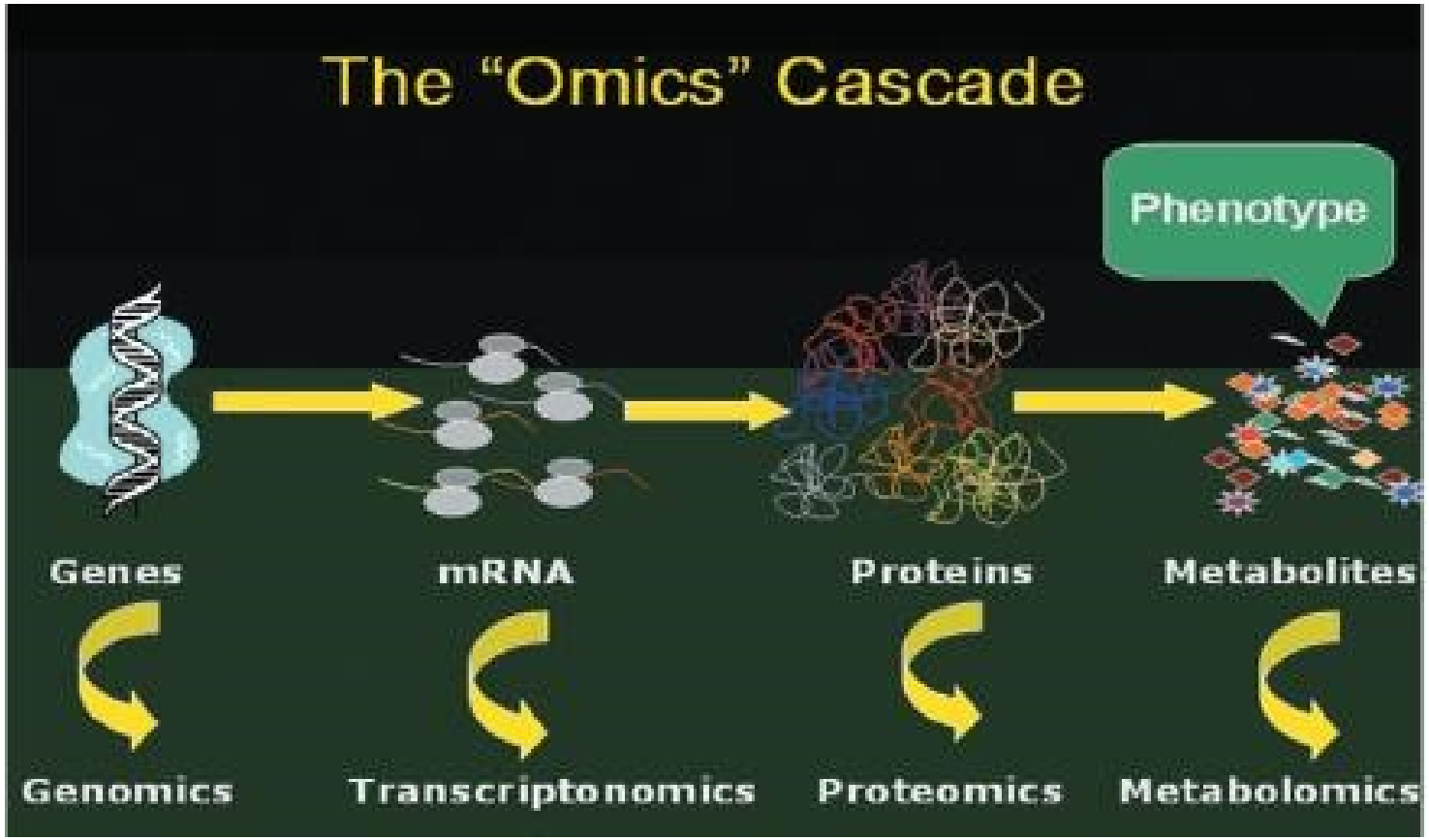
- *identify* all the approximately 20,000-25,000 genes in human DNA,
- *determine* the sequences of the 3 billion chemical base pairs that make up human DNA,
- *store* this information in databases,
- *improve* tools for data analysis,
- *transfer* related technologies to the private sector, and
- *address* the ethical, legal, and social issues (ELSI) that may arise from the project.

- omics

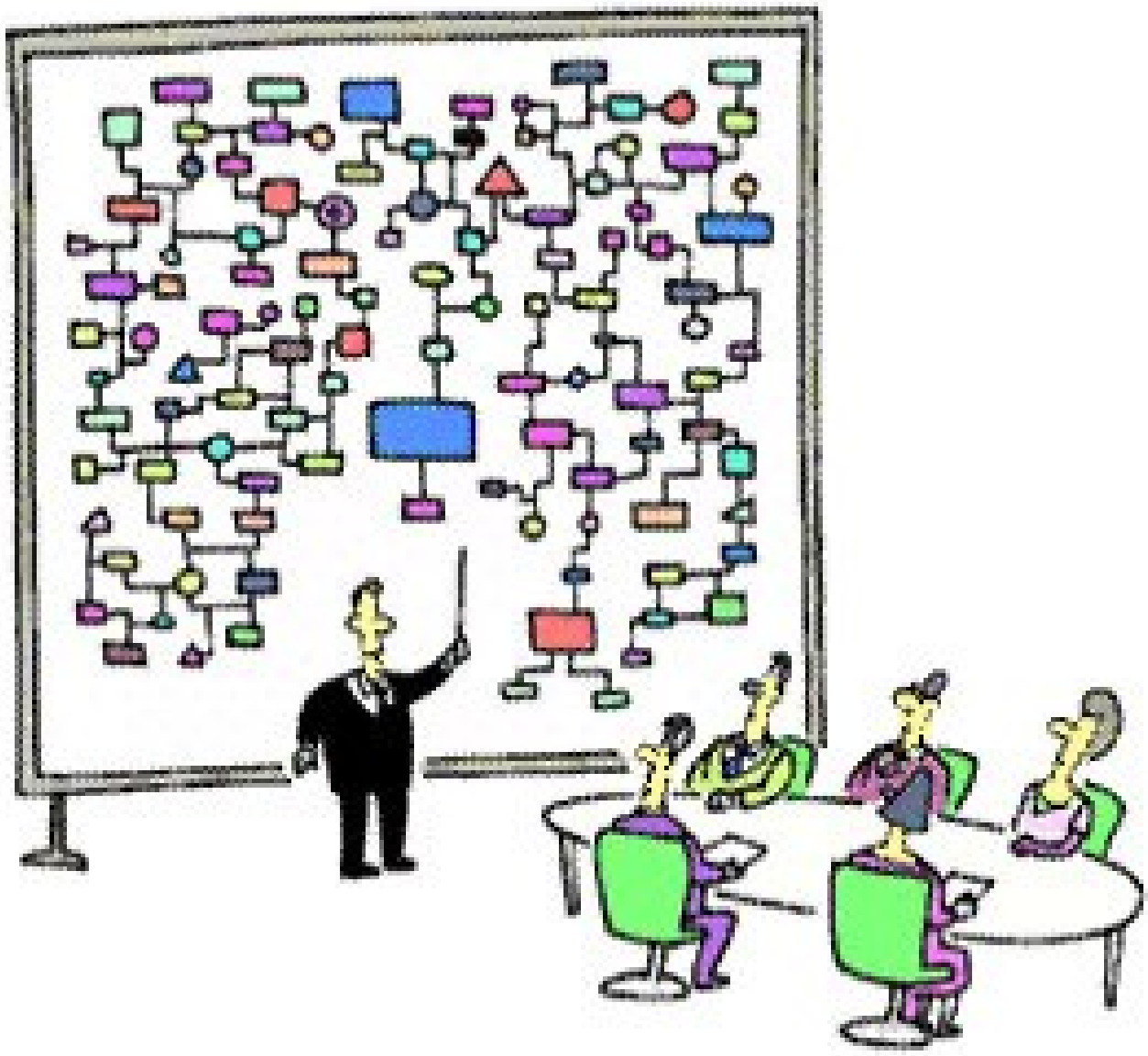

Cancer Cells
Normal Cells
RNA Isolation
mRNA
mRNA
Reverse
Transcriptase
Labeling
cDNA
"Red Fluorescent" Targets
cDNA
"Green Fluorescent" Targets
Combine Targets
Hybridize to
Microarray


A
LED
DIFFUSER
BEAM EXPANDER
POLARIZER
IRIS
PRISM
CCD
NANOGLASS
GOLD
SPR Image of Array
Etched Nanoglassy
Wells
B

- OMICS

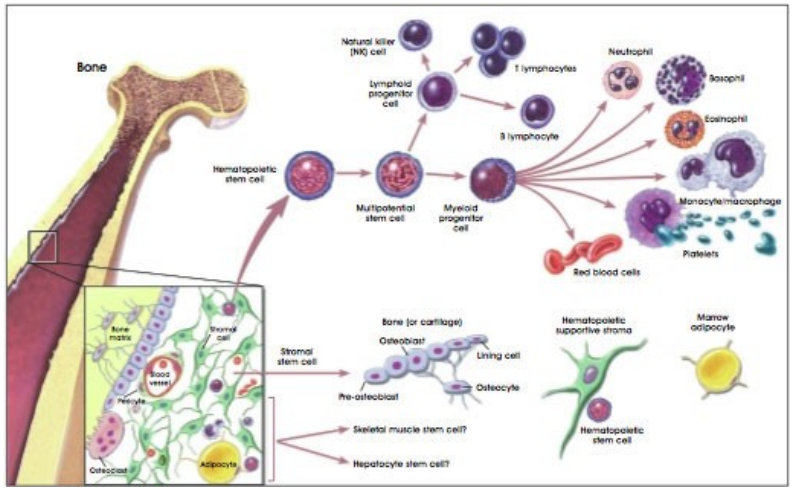

The "Omics" Cascade
Phenotype
Genes
mRNA
Proteins
Metabolites
Genomics
Transcriptonomics
Proteomics
Metabolomics

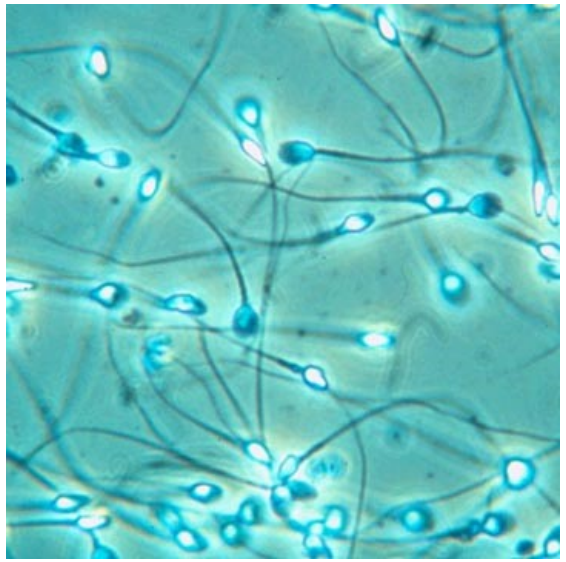

cluster 1 (404)
cluster 2 (260)
cluster 3 (219)
cluster 4 (164)
cluster 5 (144)
cluster 6 (131)
cluster 7 (124)
cluster 8 (113)
cluster 9 (111)
cluster 10 (105)
cluster 11 (99)
cluster 12 (94)
cluster 13 (86)
cluster 14 (78)
cluster 15 (74)
cluster 16 (71)
cluster 17 (70)
cluster 18 (67)
cluster 19 (65)
cluster 20 (57)
cluster 21 (56)
cluster 22 (56)
cluster 23 (55)
cluster 24 (55)
cluster 25 (55)
cluster 26 (55)
cluster 27 (52)
cluster 28 (50)
cluster 29 (48)
cluster 30 (46)
cluster 31 (44)
cluster 32 (42)
cluster 33 (41)
cluster 34 (40)
cluster 35 (39)
cluster 36 (39)
cluster 37 (38)
cluster 38 (35)
cluster 39 (35)
cluster 40 (34)
cluster 41 (32)
cluster 42 (32)
cluster 43 (31)
cluster 44 (30)
cluster 45 (29)
cluster 46 (29)
cluster 47 (29)
cluster 48 (29)
cluster 49 (28)
cluster 50 (28)
cluster 51 (28)
cluster 52 (28)
cluster 53 (27)
cluster 54 (27)
cluster 55 (26)
cluster 56 (25)
cluster 57 (25)
cluster 58 (25)
cluster 59 (25)
cluster 60 (23)
cluster 61 (23)
cluster 62 (23)
cluster 63 (23)
cluster 64 (23)
cluster 65 (22)
cluster 66 (22)
cluster 67 (22)
cluster 68 (21)
cluster 69 (21)
cluster 70 (21)
cluster 71 (21)
cluster 72 (21)
cluster 73 (21)
cluster 74 (21)
cluster 75 (20)
cluster 76 (20)
cluster 77 (20)
cluster 78 (20)
cluster 79 (20)
cluster 80 (20)
cluster 81 (20)
cluster 82 (19)
cluster 83 (19)
cluster 84 (19)
cluster 85 (19)
cluster 86 (19)
cluster 87 (18)
cluster 88 (18)
cluster 89 (18)
cluster 90 (18)
cluster 91 (18)
cluster 92 (17)
cluster 93 (17)
cluster 94 (17)
cluster 95 (16)
cluster 96 (16)
cluster 97 (16)
cluster 98 (16)
cluster 99 (16)
cluster 100 (17)
cluster 101 (16)
cluster 102 (15)
cluster 103 (15)
cluster 104 (15)
cluster 105 (15)
cluster 106 (15)
cluster 107 (14)
cluster 108 (14)
cluster 109 (14)
cluster 110 (14)
cluster 111 (14)
cluster 112 (14)
cluster 113 (14)
cluster 114 (14)
cluster 115 (14)
cluster 116 (14)
cluster 117 (14)
cluster 118 (14)
cluster 119 (14)
cluster 120 (13)
cluster 121 (13)
cluster 122 (13)
cluster 123 (13)
cluster 124 (13)
cluster 125 (13)
cluster 126 (13)
cluster 127 (13)
cluster 128 (13)
cluster 129 (13)
cluster 130 (12)
cluster 131 (12)
cluster 132 (12)
cluster 133 (12)
cluster 134 (12)
cluster 135 (12)
cluster 136 (12)
cluster 137 (12)
cluster 138 (12)
cluster 139 (12)
cluster 140 (12)
cluster 141 (14)
cluster 142 (11)
cluster 143 (11)
cluster 144 (11)
cluster 145 (11)
cluster 146 (11)
cluster 147 (11)
cluster 148 (11)
cluster 149 (11)
cluster 150 (11)
cluster 151 (11)
cluster 152 (11)
cluster 153 (11)
cluster 154 (11)
cluster 155 (11)
cluster 156 (11)
cluster 157 (11)
cluster 158 (11)
cluster 159 (11)
cluster 160 (10)
cluster 161 (11)
cluster 162 (11)
cluster 163 (10)
cluster 164 (10)
cluster 165 (10)
cluster 166 (10)
cluster 167 (10)
cluster 168 (10)
cluster 169 (10)
cluster 170 (10)
cluster 171 (10)
cluster 172 (10)
cluster 173 (10)
cluster 174 (10)
cluster 175 (10)
cluster 176 (10)
cluster 177 (10)
cluster 178 (9)
cluster 179 (9)
cluster 180 (10)
cluster 181 (9)
cluster 182 (9)
cluster 183 (9)
cluster 184 (9)
cluster 185 (9)
cluster 186 (9)
cluster 187 (9)
cluster 188 (9)
cluster 189 (9)
cluster 190 (9)
cluster 191 (9)
cluster 192 (9)
cluster 193 (9)
cluster 194 (9)
cluster 195 (9)
cluster 196 (9)
cluster 197 (9)
cluster 198 (9)
cluster 199 (9)
cluster 200 (9)
N
ms
pm

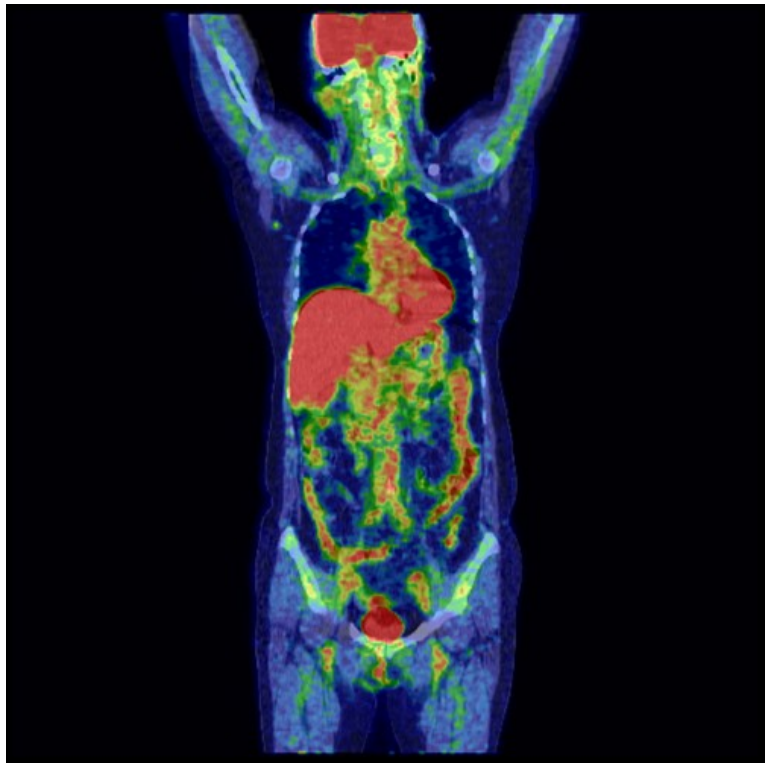


"And that's why we need a computer."

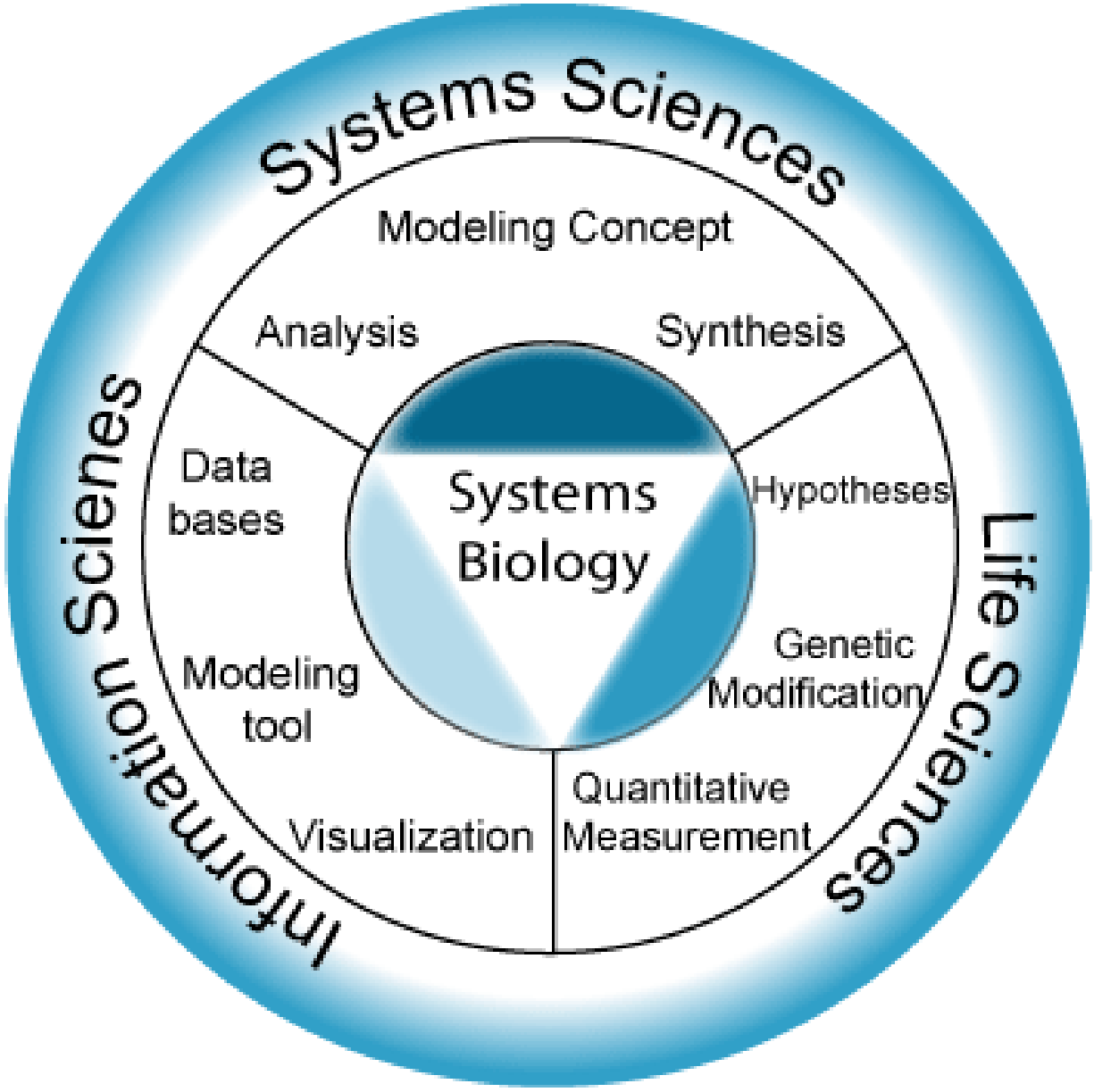
Computational models in biology and medicine







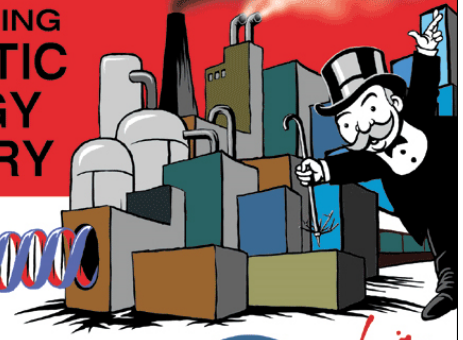
Systems biology


Systems Sciences
Modeling Concept
Analysis
Synthesis
Data bases
Systems Biology
Hypotheses
Modeling tool
Genetic Modification
Visualization
Quantitative Measurement
Information Sciences
Life Sciences

SYNDUSTRY

The news of "Synthia," the world's first human-made species, is just the latest from a rapidly growing artificial life industry. Synthetic biology (or "Syn Bio") aims to profit from the design and construction of industrially useful life-forms.

THE EMERGING SYNTHETIC BIOLOGY INDUSTRY



Syn Bio's Big Shots

Global corporations are investing in synthetic biology labs and partnering with start-up companies.

"Over the next 20 years synthetic genomics is going to become the standard for making anything." ~ Craig Venter

Cargill
Agribusiness giant. Supports synthetic biology R&D.

BP Energy giant. $500 million partnership on synthetic biology with University of California Berkeley; holds equity stake in Craig Venter's Synthetic Genomics, Inc.

Du Pont
Chemical giant. Developed first commercial syn bio product with Genencor and sugar giant Tate & Lyle – a fibre called Sorona.

Pfizer
Pharma giant. Conducts in-house syn bio research for drug development.

Virgin Group
Includes Virgin Fuels, investor in synthetic biology. Controlled by celebrity billionaire Richard Branson.

Synthetic Startups

A bevy of 'pure play' syn bio companies is attempting to design synthetic microbes for fuel, chemicals and drugs. Many are university spin-offs.

Gevo
(USA) Developing synthetic biofuels with support from Virgin.

Mascoma
(USA) Developing synthetic biofuels.

Synthetic Genomics
(USA) Constructing synthetic life forms for biofuels and carbon sequestration.

LS9
(USA) Developing synthetic biofuels and industrial chemicals.

Amyris Biotech
(USA) Developing cellular factories to produce drugs, fuels and industrial chemicals.

ProtoLife
(Italy) Developing synthetic living systems.

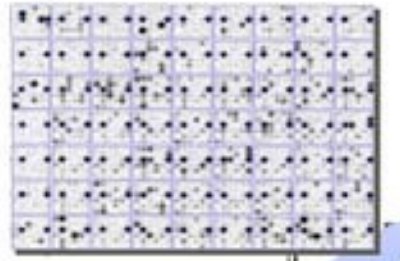
DNA Synthesis Foundries

DNA foundries produce the raw material for creating artificial life: synthetic DNA (sDNA).

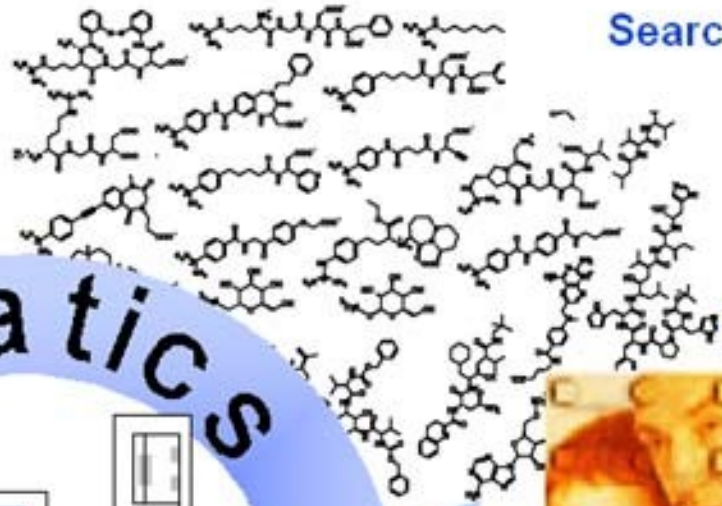
Over 70 DNA foundries worldwide manufacture sDNA for genetic engineers and synthetic biologists. The market for sDNA already exceeds a billion dollars annually. Even long DNA sequences – entire genes, for example – can be ordered over the Internet and delivered within two weeks. The speed of producing accurate DNA sequences is doubling every two years and costs are halving even faster.

Published by ETC Group Dec 2007 Artwork by Shtig
www.etcgroup.org

DNA chips: comparison of cell states

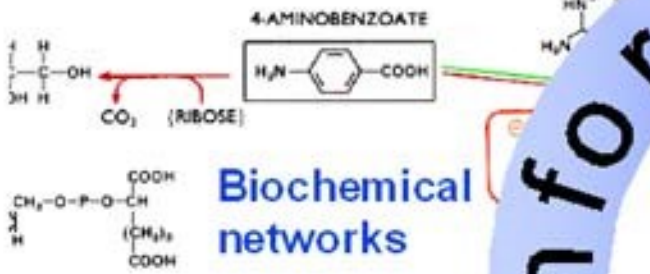


Search for new drugs



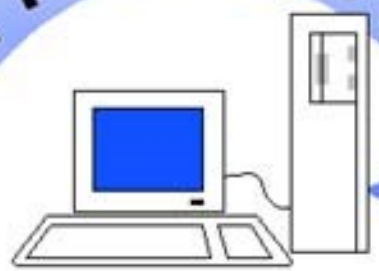
Genetic variations

Biochemical networks



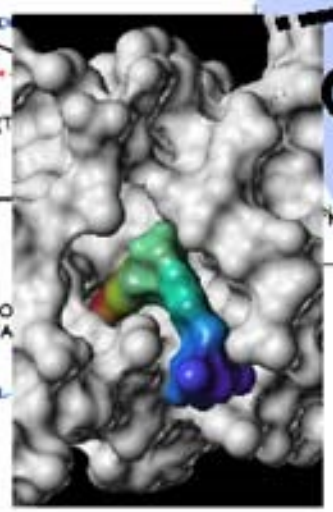
bioinformatics

Data handling, Algorithms
Statistics, Visualisation



Optimizing therapies

Molecular Interactions



Structure prediction

Genomes

ccctgtggagccacaccctagggttggcca
atctactcccaggagcaggagggcaggag...

Proteins

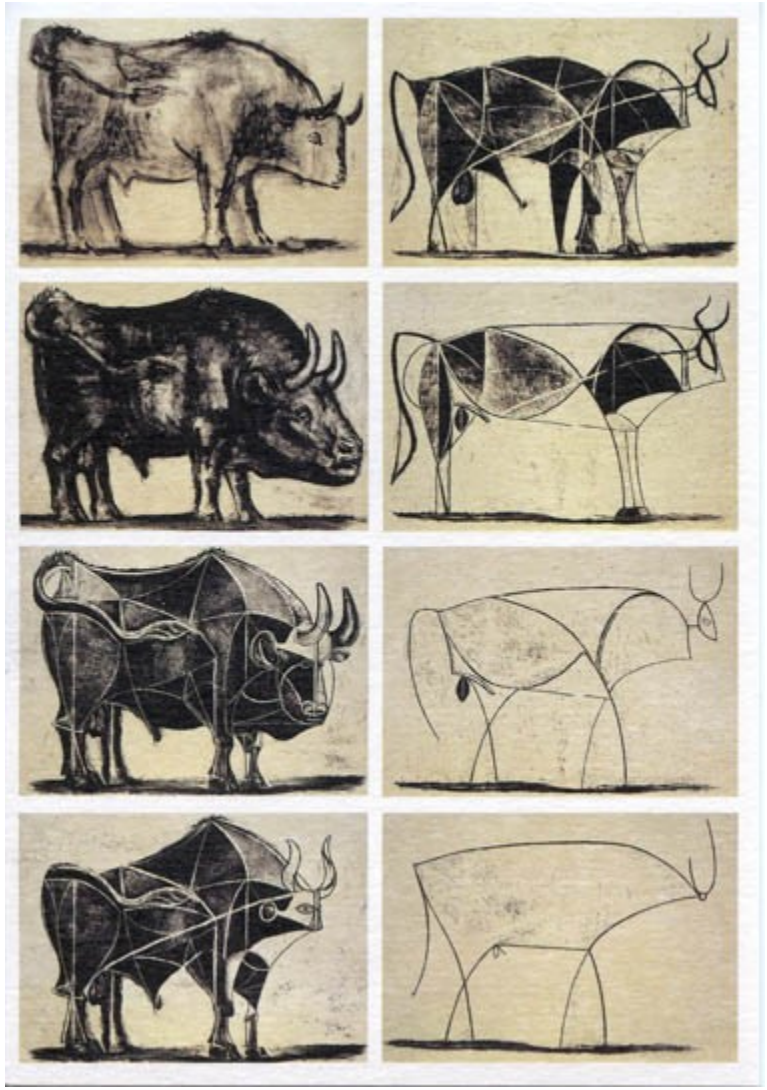
MTNRNFRQIINLLDLRWQRRVPVIHQTETA
ECGLACLAMICGHFGKNIDIIYLRRKFNLS...

Sequence analysis

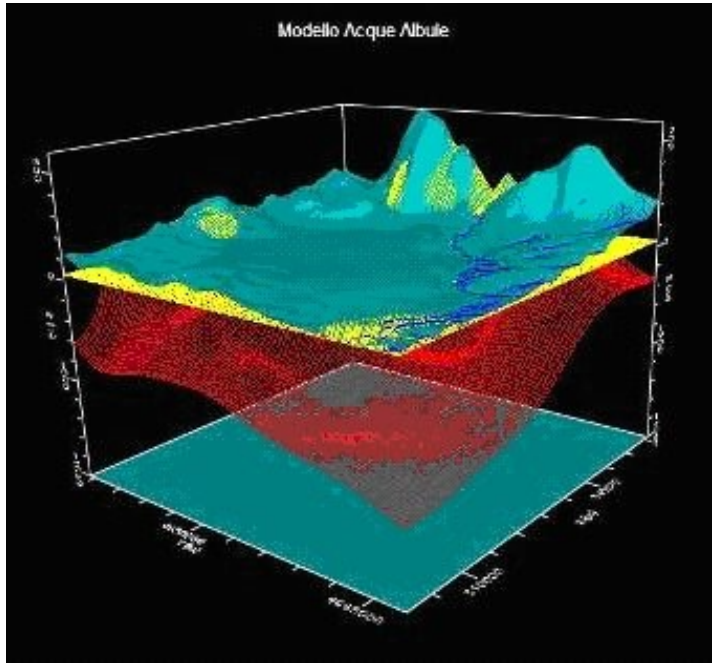
Computational biology

What is a model?

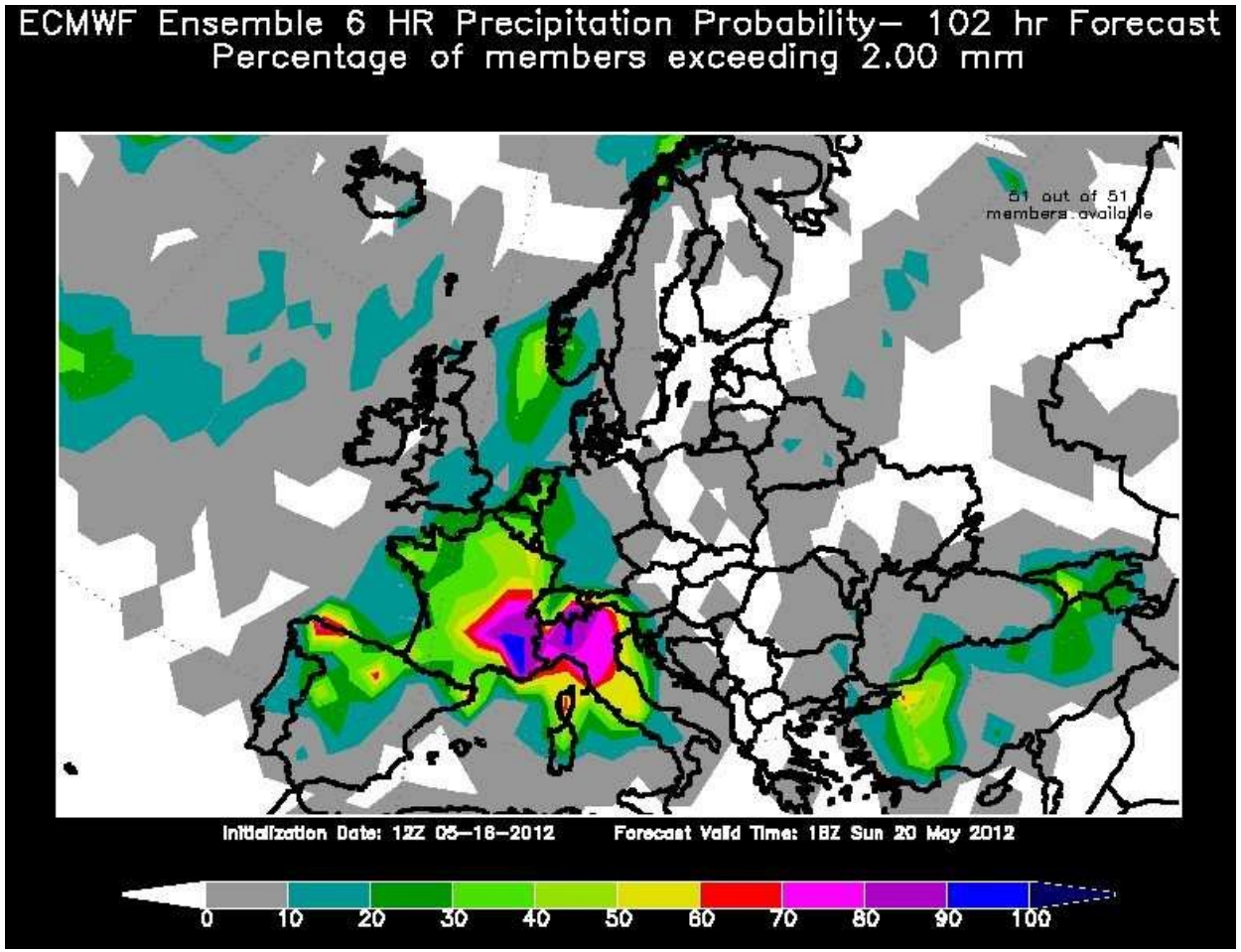
What kind of model we need?



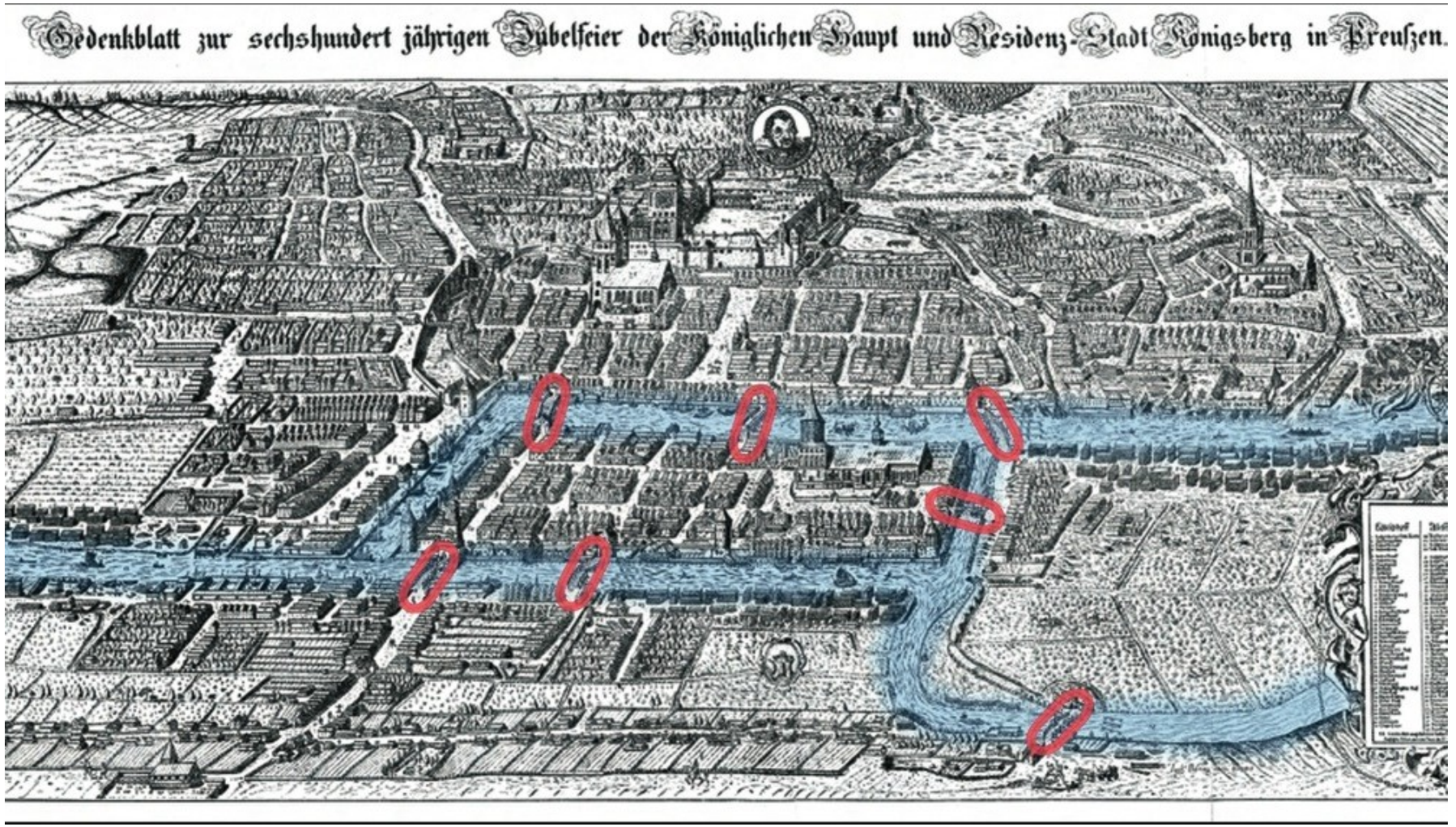
Numerical models



forecasting model



Networks as model



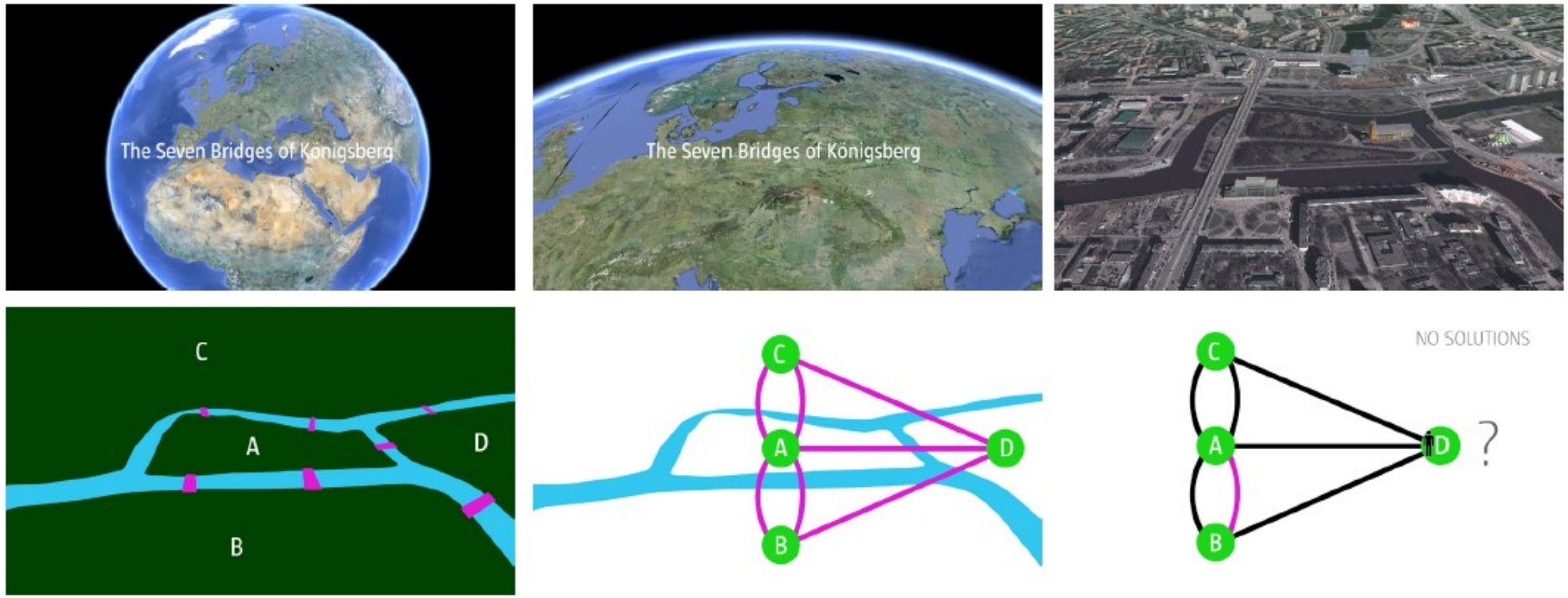


Image 2.1

The bridges of Königsberg.

From the contemporary map of Königsberg (now Kaliningrad, Russia) to Euler's graph. The graph constructed by Euler consists of four nodes (A, B, C, D), each corresponding to a patch of land, and seven links, each corresponding to a bridge. Euler showed in 1736 that there is no continuous path that would cross seven the bridges while never crossing the same bridge twice. The people of Königsberg agreed with him, gave up their fruitless search and in 1875 they built a new bridge between B and C, increasing the number of links of these two nodes to four. Now only one node was left with an odd number of links and it became rather straightforward to find the desired path.

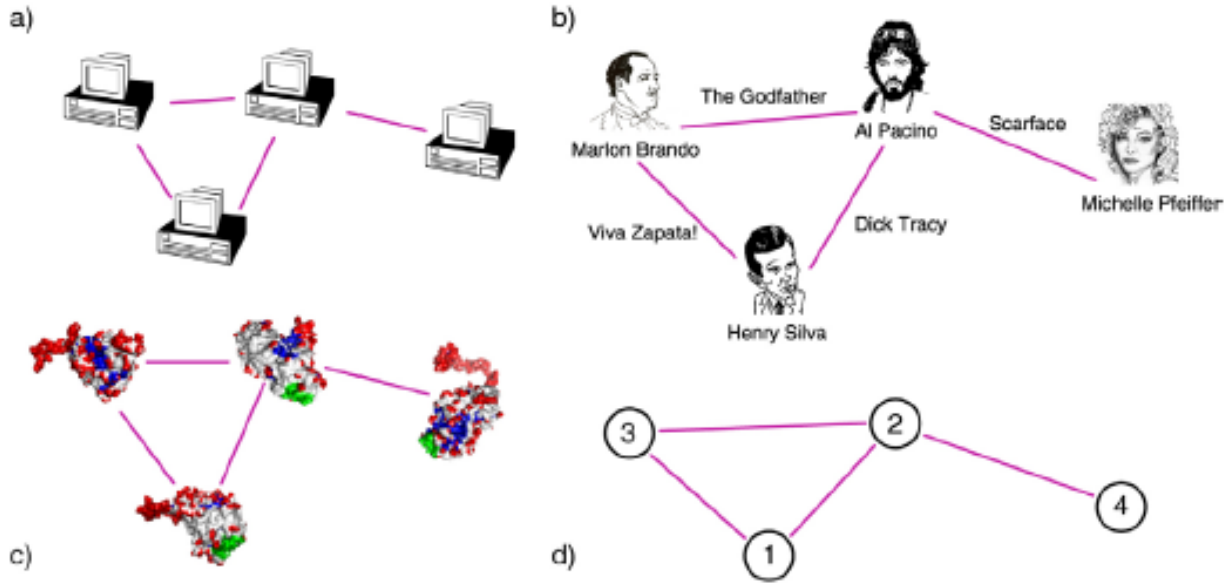
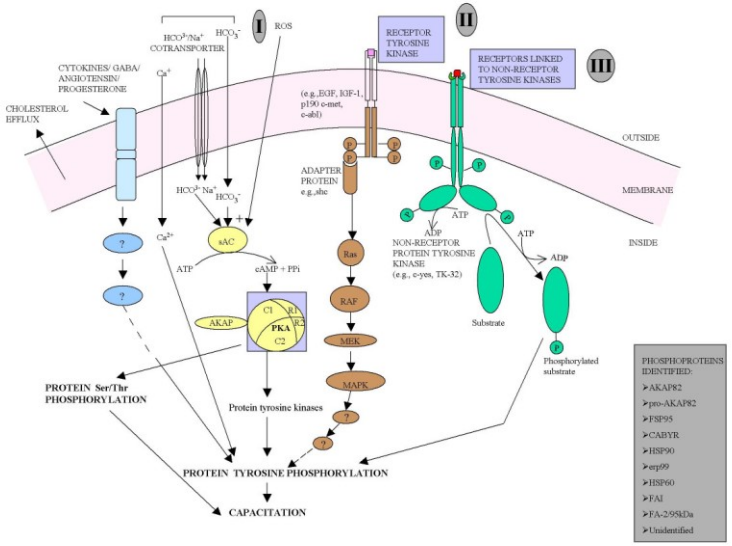


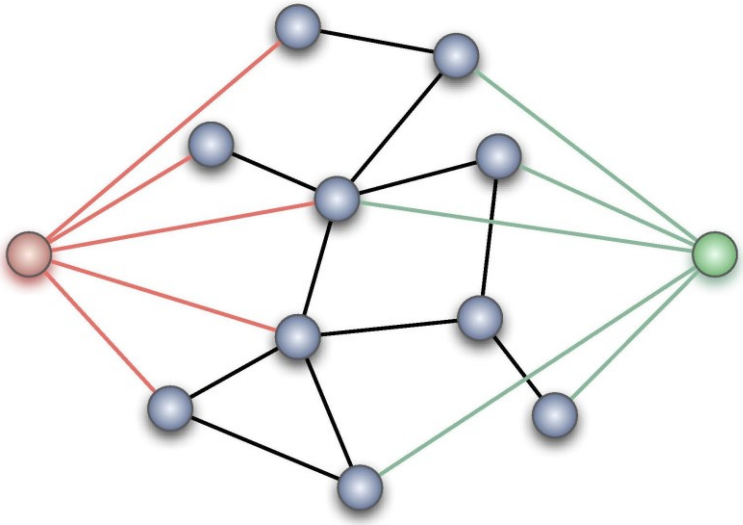
Image 2.3

Real systems of quite different nature can have the same network representation.

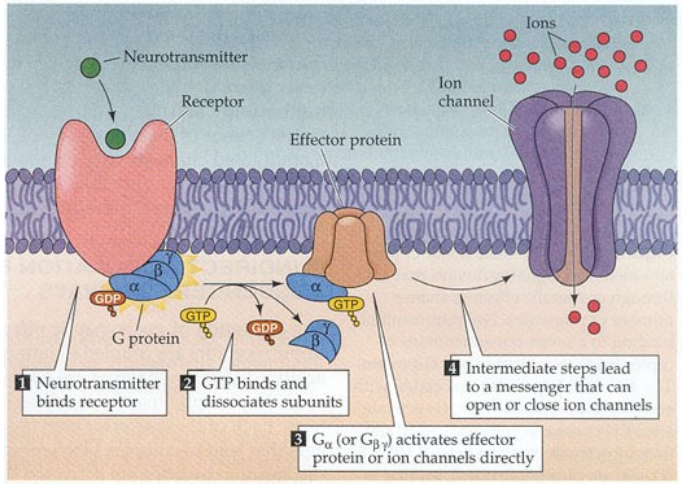
In the figure we show a small subset of (a) the *Internet*, where routers (specialized computers) are connected to each other; (b) the *Hollywood actor network*, where two actors are connected if they played in the same movie; (c) a *protein-protein interaction network*, where two proteins are connected if there is experimental evidence that they can bind to each other in the cell. While the nature of the nodes and the links differs widely, each network has the same graph representation, consisting of $N = 4$ nodes and $L = 4$ links, shown in (d).

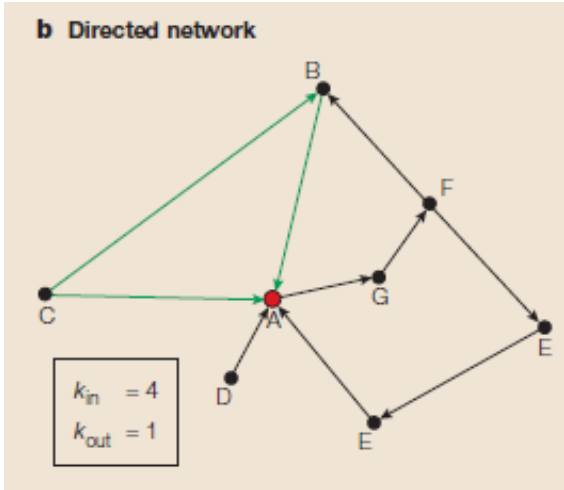
The node

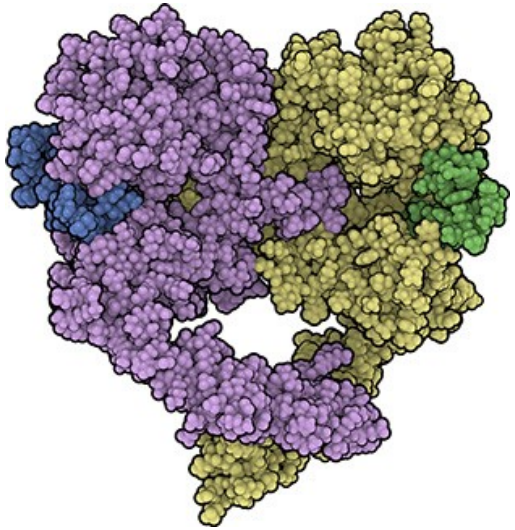

I
ROS
HCO3-/Na+
COTRANSPORTER
HCO3-
CYTOKINES/ GABA/
ANGIOTENSIN/
PROGESTERONE
Ca2+
CHOLESTEROL
EFFLUX
II
RECEPTOR
TYROSINE
KINASE
III
RECEPTORS LINKED
TO NON-RECEPTOR
TYROSINE KINASES
(e.g.,EGF, IGF-1,
p190 c-met,
c-abl)
OUTSIDE
MEMBRANE
INSIDE
ADAPTER
PROTEIN
e.g.,shc
HCO3- Na+ HCO3-
Ca2+
sAC
ATP
cAMP + PPi
Ras
RAF
MEK
MAPK
AKAP
C1
R1
R2
PKA
C2
ATP
ADP
NON-RECEPTOR
PROTEIN TYROSINE
KINASE
(e.g., c-yes, TK-32)
ATP
ADP
Substrate
Phosphorylated
substrate
PROTEIN Ser/Thr
PHOSPHORYLATION
Protein tyrosine kinases
PROTEIN TYROSINE PHOSPHORYLATION
CAPACITATION
PHOSPHOPROTEINS
IDENTIFIED:
➢AKAP82
➢pro-AKAP82
➢PSP95
➢CABYR
➢HSP90
➢ezr99
➢HSP60
➢FAI
➢FA-2.9SkDa
➢Unidentified

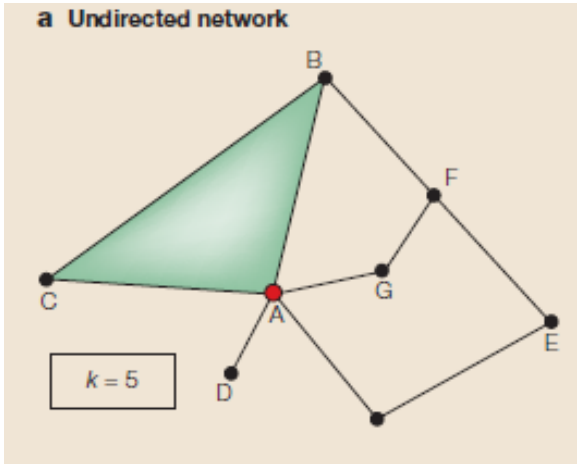


The link


Neurotransmitter
Receptor
Ions
Ion channel
Effector protein
G protein
GDP
GTP
1 Neurotransmitter binds receptor
2 GTP binds and dissociates subunits
3 G_α (or $G_{\beta\gamma}$) activates effector protein or ion channels directly
4 Intermediate steps lead to a messenger that can open or close ion channels


b Directed network
B
F
C
A
G
E
D
E
k_{in} = 4
k_{out} = 1



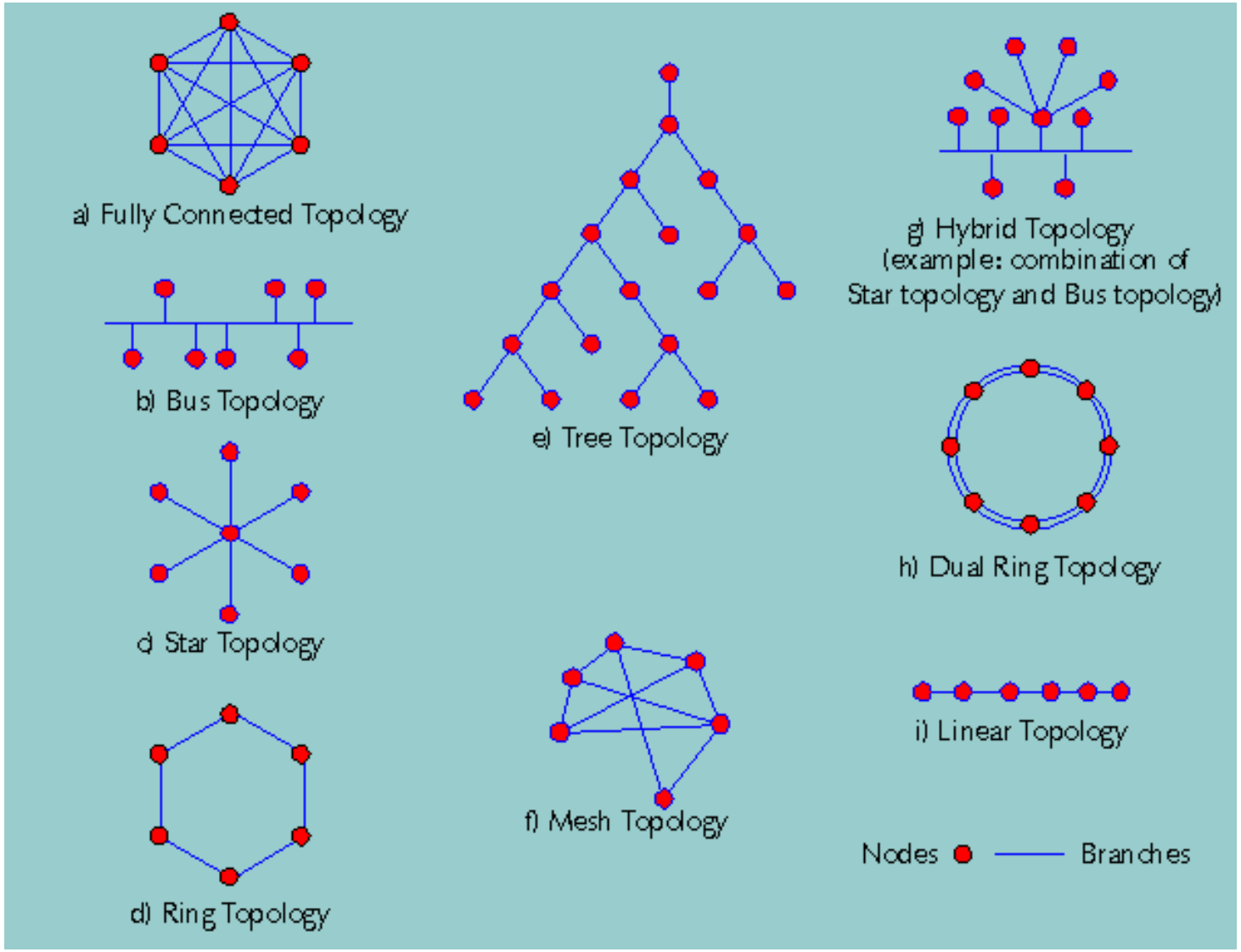

a Undirected network
B
F
C
A
G
E
D
k = 5

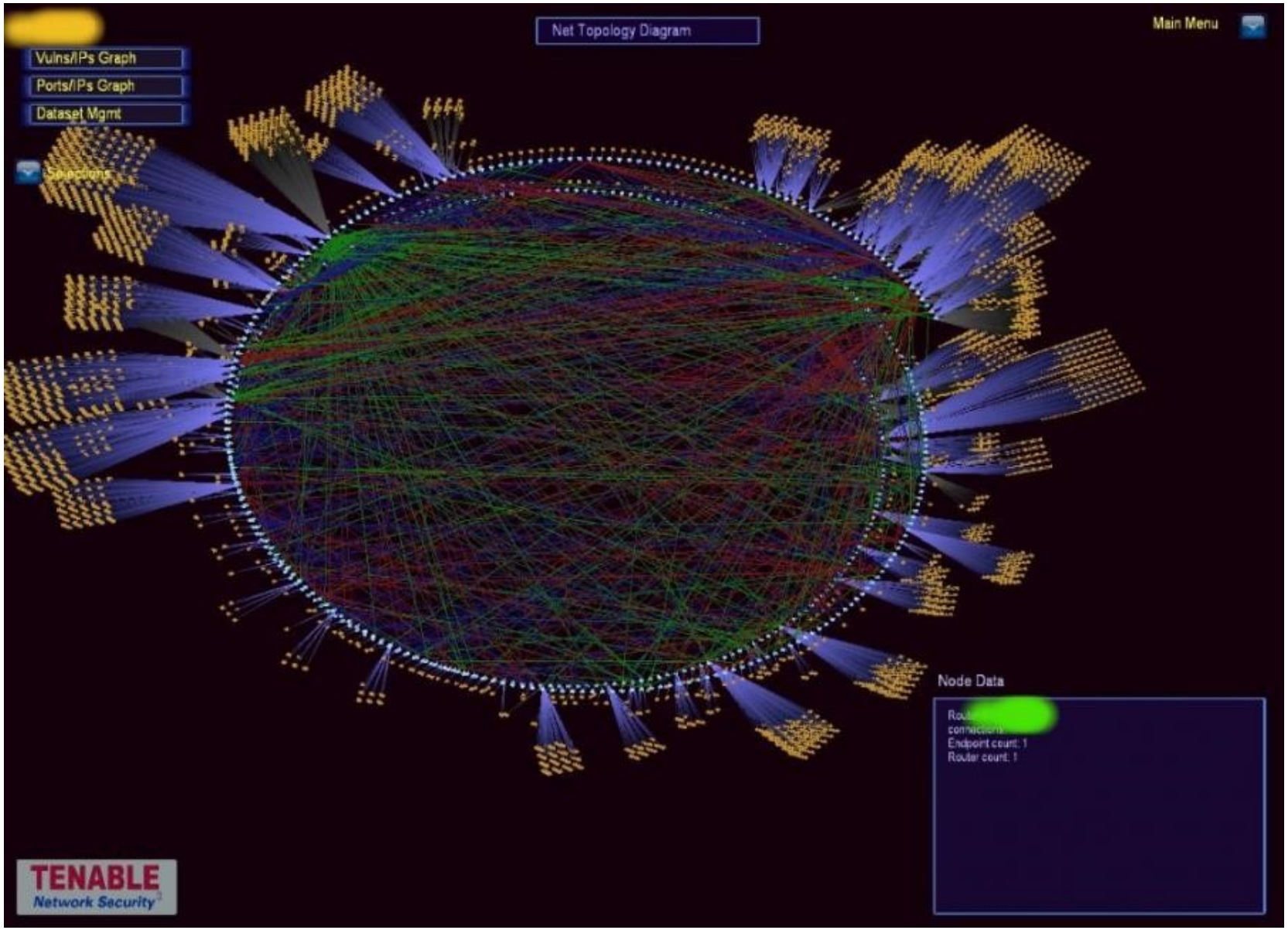
NETWORK NAME	NODES	LINKS	DIRECTED/ UNDIRECTED	N	L	‹K›
Internet	routers	Internet Connections	Undirected	192,244	609,066	2.67
WWW	webpages	links	Directed	325,729	1,497,134	4.60
Power Grid	power plants, transformers	cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	subscribers	calls	Directed	36,595	91,826	2.51
Email	email addresses	emails	Directed	57,194	103,731	1.81
Science Collaboration	scientists	co-authorships	Undirected	23,133	186,936	16.16
Actor Network	actors	co-acting	Undirected	212,250	3,054,278	28.78
Citation Network	papers	citations	Directed	449,673	4,707,958	10.47
E. coli Metabolism	metabolites	chemical reactions	Directed	1,039	5,802	5.84
Yeast Protein Interactions	proteins	binding interactions	Undirected	2,018	2,930	2.90

Table 2.1

Network maps and their basic properties.

Network topology




Net Topology Diagram
Main Menu
Vulns/IPs Graph
Ports/IPs Graph
Dataset Mgmt
Selections
Node Data
Endpoint count: 1
Router count: 1
TENABLE
Network Security

Topological parameters

the number of nodes: which represents the total number of molecules involved;

In an undirected network total number of links, L, can be expressed as the sum of the node degrees:

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i \qquad (1)$$

Here the 1/2 factor corrects for the fact that in the sum (1) each link is counted twice.

the number of edges: which represents the total number of interaction among nods within the network;

the node degree (or connectivity): which indicates how many links each node has to other nodes;

$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N} \qquad (7)$$

In directed networks we distinguish between *incoming degree*, k_i^{in}, representing the number of links that point node i, and *outgoing degree*, k_i^{out}, representing the number of links that point from the node i to other nodes and the *total degree*, k_i, given by

$$k_i = k_i^{in} + k_i^{out} \qquad (8)$$

the node degree distribution P(k): which represents the probability that a selected node has exactly *k* links;

$$p_k = \frac{N_k}{N}$$

the clustering coefficient: it is a measure of how the nodes tend to form clusters: the more the clustering coefficient is higher, the more the presence of clusters will increase;

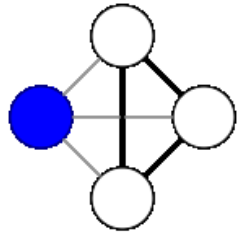
$$C_I = 2n_I/k(k-1),$$

where n_I is the number of links connecting the k_I neighbours of node I to each other

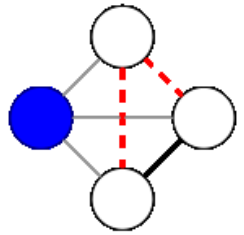
clustering coefficient

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triples of vertices}}$$

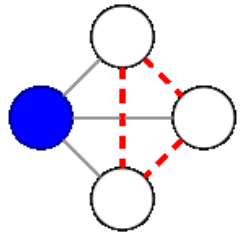
$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i.$$



c = 1



c = 1/3

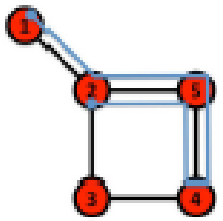


c = 0

the network diameter: which is the largest distance between two nodes;

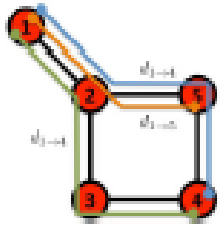
the averaged number of neighbours: which is the mean number of connection of nodes;

the characteristic path length: which is the expected distance between two random individuated connected nodes.



Path

PATH: A sequence of nodes such that each node is connected to the next node along the path by a link. A path always consists of *n* nodes and *n* - *1* links. The length of a path is defined as the number of its links, counting multiple edges multiple times.

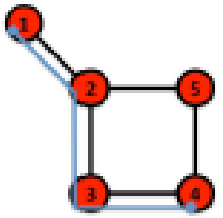


Shortest Path

$d_{1\to4} = 3$

$d_{1\to5} = 2$

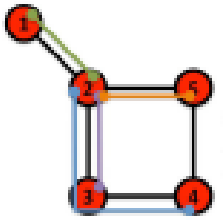
SHORTEST PATH (geodesic path, *d*): the path with the shortest distance *d* between two nodes. We will call it the distance between two nodes.



Diameter

$d_{1\to4} = 3$

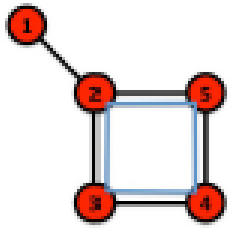
DIAMETER (d_{max}): the longest shortest path in a graph, or the distance between the two furthest away nodes.



Average Path Length

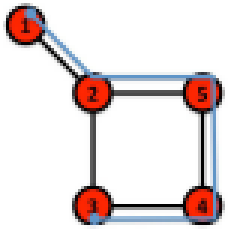
$(d_{1\to2} + d_{1\to3} + d_{1\to4} + d_{1\to5} + d_{2\to1} + d_{2\to3} + d_{2\to4} + d_{2\to5} + d_{3\to1} + d_{3\to2} + d_{3\to4} + d_{3\to5} + d_{4\to1} + d_{4\to2} + d_{4\to3} + d_{4\to5} + d_{5\to1} + d_{5\to2} + d_{5\to3} + d_{5\to4})/10 = 1.6$

AVERAGE PATH LENGTH ($\langle d \rangle$): the average of the shortest paths between all pairs of nodes.



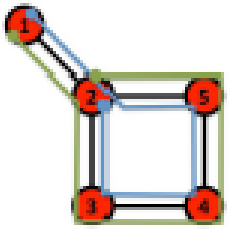
Cycle

CYCLE: a path with the same start and end node.



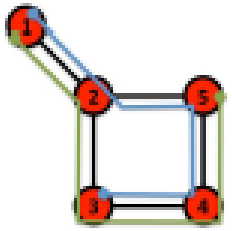
Self-avoiding Path

SELF-AVOIDING PATH: a path that does not intersect itself, i.e. the same node or link does not occur twice along the path.



Eulerian Path

EULERIAN PATH: a path that traverses each link exactly once.



Hamiltonian Path

HAMILTONIAN PATH: a path that visits each node exactly once.

multiple networks

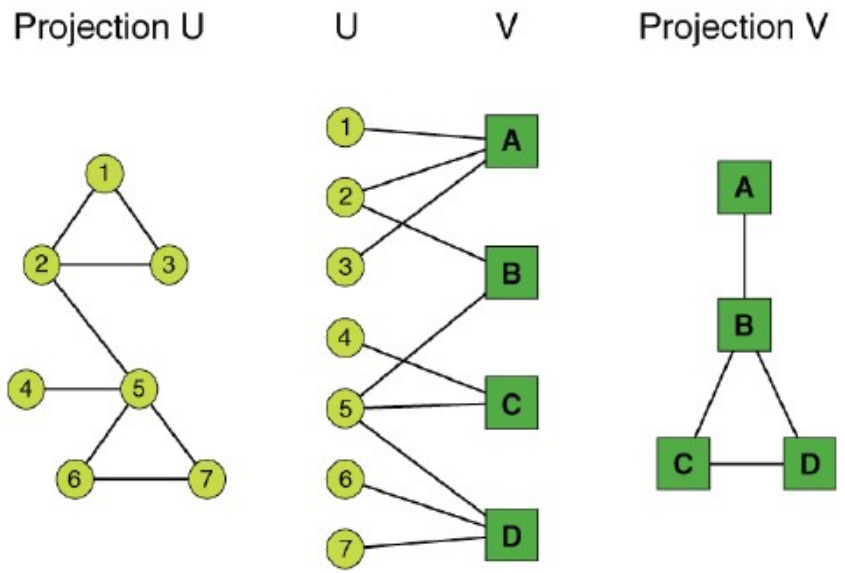


Image 2.9a
Bipartite network.

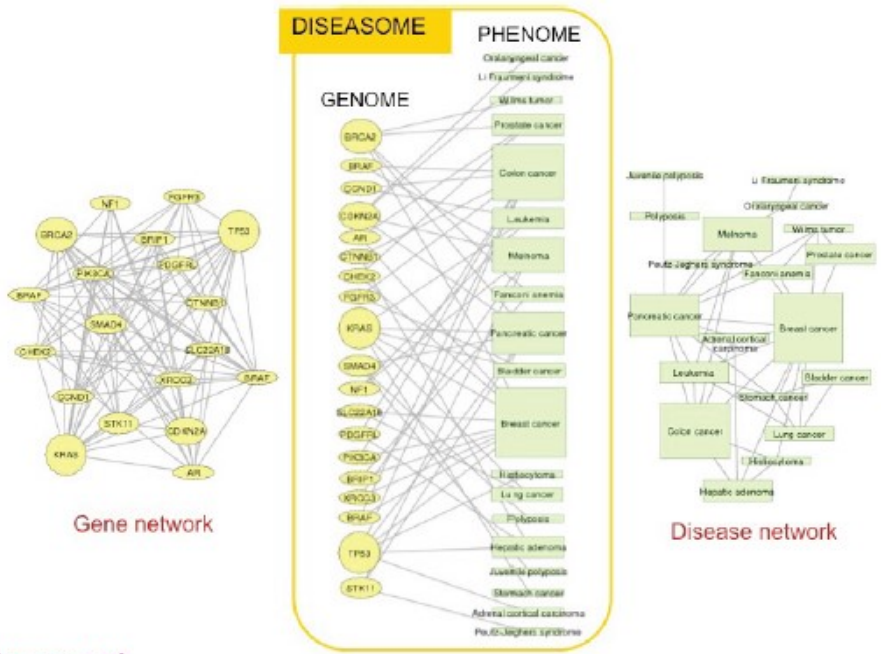


Image 2.9b
Bipartite network.

The tripartite recipe–ingredient–compound network, in which one set of nodes are recipes, like Chicken Marsala, the second set corresponds to the ingredients each recipe has (like flour, sage, chicken, wine, and butter for Chicken Marsala), and the third set captures the flavor compounds, or chemicals that contribute to the taste of a particular ingredient.

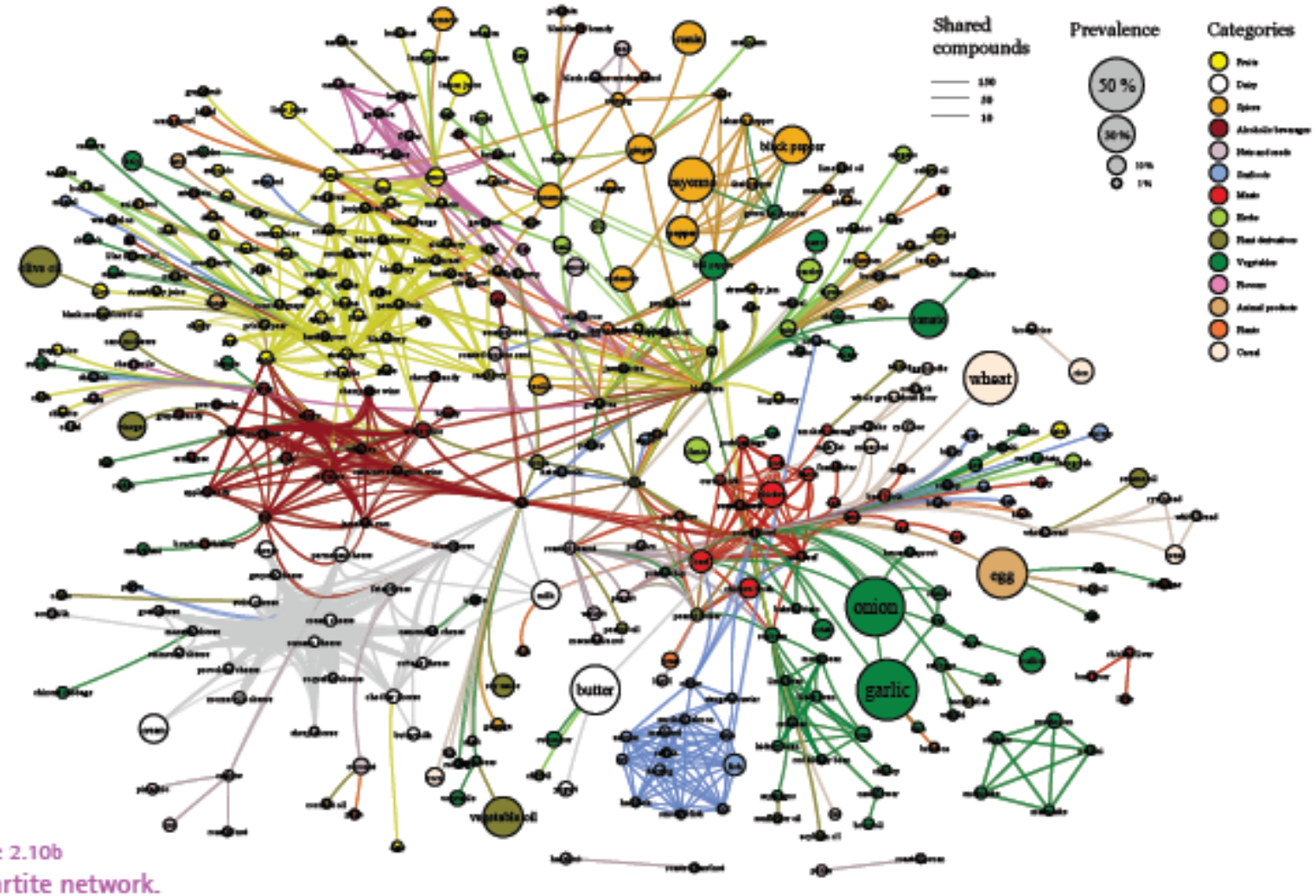


Image 2.10b

Tripartite network.

A projection of the tripartite network, resulting in the ingredient network, often called the flavor network. Each node denotes an ingredient; the node color indicating the food category and node size reflects the ingredient prevalence in recipes. Two ingredients are connected if they share a significant number of flavor compounds, link thickness representing the number of shared compounds between the two ingredients (After [12]).