Science of Information Initiative

Wojciech Szpankowski and Ananth Grama

. . . to integrate research and teaching activities aimed at investigating the role of **information** from various viewpoints: from fundamental theoretical underpinnings of information to diverse applications in life sciences, communication networks, economics, and complex systems.

Outline

1. What is Information?
2. Beyond Shannon Information
3. Today's Challenges
4. Some Examples

Interpretation of Information . . .

C. F. Von Weizsäcker:

"Information is only that which produces information" (relativity).
"Information is only that which is understood" (rationality)
"Information has no absolute meaning".

R. Feynman:

". . . Information is as much a property of your own knowledge as anything in the message.
. . . Information is not simply a physical property of a message:
it is a property of the message and your knowledge about it."

Interpretation of Information . . .

J. Wheeler:
"It from Bit" (Information is physical.)

C. Shannon:
"These semantic aspects of communication are irrelevant . . ."

Structural and Biological Information

F. Brooks, jr. (JACM, 50, 2003, "Three Great Challenges for ... CS "):

"Shannon ... gave us a definition of Information and a metric for Information as communicated from place to place. We have **no theory** however that gives us a metric for the Information embodied in structure ... this is the most fundamental gap in the theoretical underpinning of Information and computer science. ... A young information theory scholar willing to spend years on a **deeply fundamental problem** need look no further."

Structural and Biological Information

M. Eigen

'The differentiable characteristic of the living systems is **Information**. Information assures the controlled reproduction of all constituents, thereby ensuring conservation of viability . . . . Information theory, pioneered by Claude Shannon, cannot answer this question . . . in principle, the answer was formulated 130 years ago by Charles Darwin.

Shannon Information . . .

In 1948 C. Shannon created a powerful theory of **information** that served as the backbone to a now classical paradigm of digital communication.

In our setting, Shannon defined:

objective: statistical ignorance of the recipient;
statistical uncertainty of the recipient.

cost: # binary decisions to describe E;
= $-\log P(E)$; $P(E)$ being the probability of E.

Context: the semantics of data is irrelevant . . .

Self-information for E_i: $\text{info}(E_i) = -\log P(E_i)$.
Average information: $H(P) = -\sum_i P(E_i) \log P(E_i)$

Shannon's statistical information tells us how much a recipient of data can reduce their statistical uncertainty by observing data.

Beyond Shannon

Delay: In many communication problems, when information is transmitted over a (biological or computer) network, the amount of delay incurred is important (e.g., complete information arriving late maybe useless).

Space: In networks the spatially distributed components raise fundamental issues of limitations in **information exchange** since the available resources must be shared (e.g., spatial localization is at the heart of cell dynamics).

Information and Control: **Information** is exchanged in space and time for decision making, thus timeliness of information delivery along with reliability and complexity constitute the basic objective (e.g., robust, and self stabilizing system dynamics are key to signaling and regulation in biology).

Beyond Shannon

Dynamic information: In a complex network in a space-time-control environment information is not simply communicated but also processed and generated along the way (e.g., human brain).

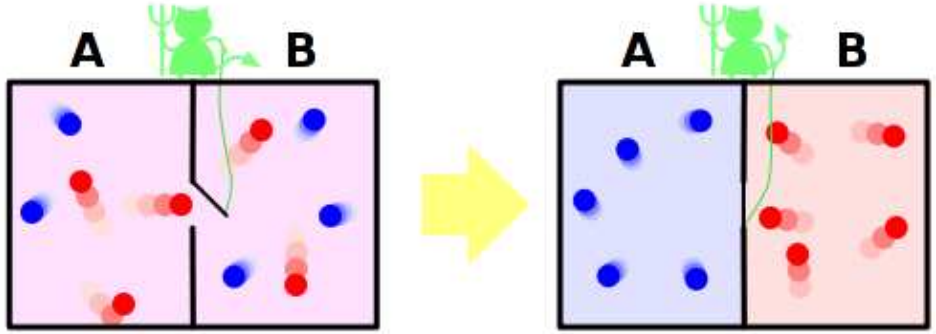
Limited Computational Resources: Often information is limited by available computational resources (e.g., cell phone, living cells, weirdness of quantum world).

Semantics vs Syntax: Semantics of higher order processes are often needed, without precise knowledge of underlying components. How can these semantics be inferred and coded in a context-specific manner?

Information in the Physical Sciences

- Connections between information and the real world have intruigued researchers for centuries.

- Maxwell's thought experiment (1867) in relation to the 2nd Law of Thermodynamics provides an elegant nexus.

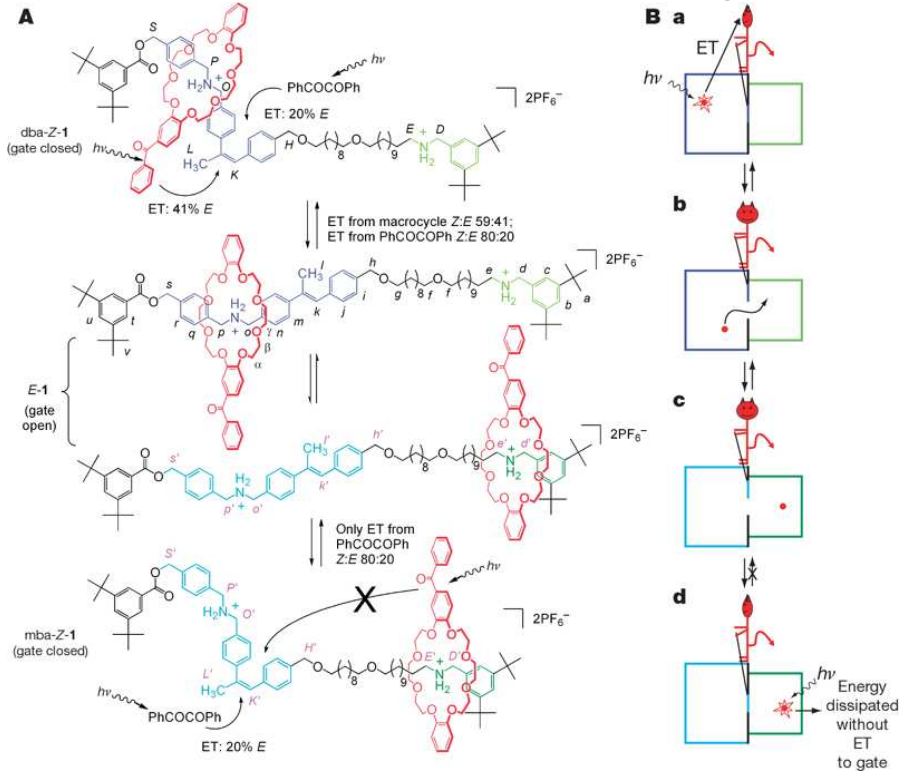


Maxwell's demon was one of the first observations relating energy to information.

- Szilard and Brillouin subsequently used this to equate one bit of information with $K_B T \ln 2$ joules of energy.

Information and the Sciences – a Brief History

Maxwell's Demon Realized (David Leigh, Nature, 2007)



Irradiation of rotaxane 1 at 350 nm in CD_3OD at 298 K interconverts the three diastereomers of 1 and, in the presence of benzil, drives the ring distribution away from the thermodynamic minimum, increasing the free energy of the molecular system without ever changing the binding strengths of the macrocycle or ammonium binding sites.

Information and the Sciences – a Brief History

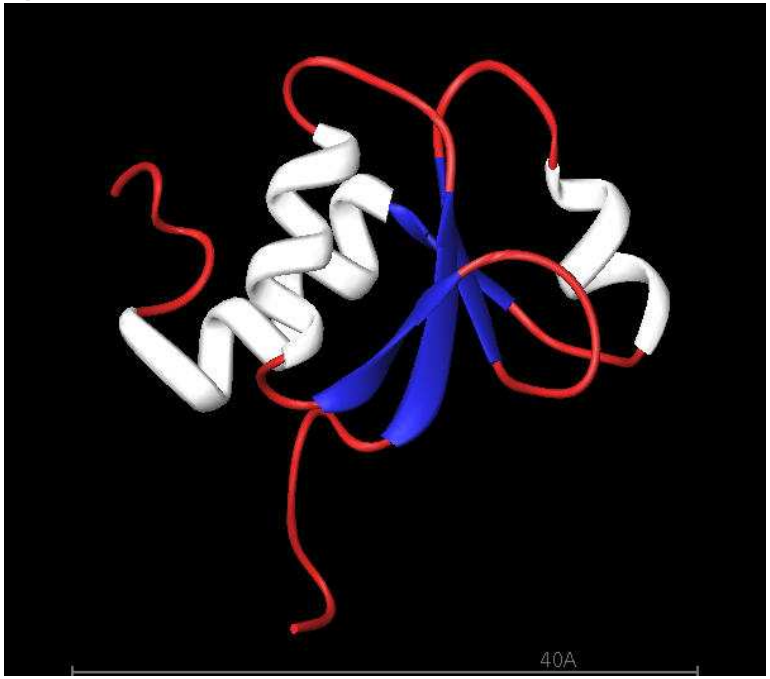
- As the science of information developed, so did interest in correspondence between information theoretic and scientific measures.

- Among the more intuitive and well studied is the correspondence between thermodynamic (Boltzmann-Gibbs) entropy and information theoretic (Shannon-Hartley) entropy.

- Drawing on these, two questions arise:
  - Can we draw on information theoretic formalisms to address foundational questions in scientific disciplines?
  - Can we draw on physical principles to address basic questions in computing?

Role of Information in Scientific Discovery

What is the "informative" component of XRCC1, BRCA-1, and H. sapiens DNA ligase III?



BRCT domain, from T.Thermophilus DNA Ligase.

These are inferred using a variety of information correlation and extraction techniques (sequence analysis, Markov models), and experimentally validated.

Role of Information in Scientific Discovery

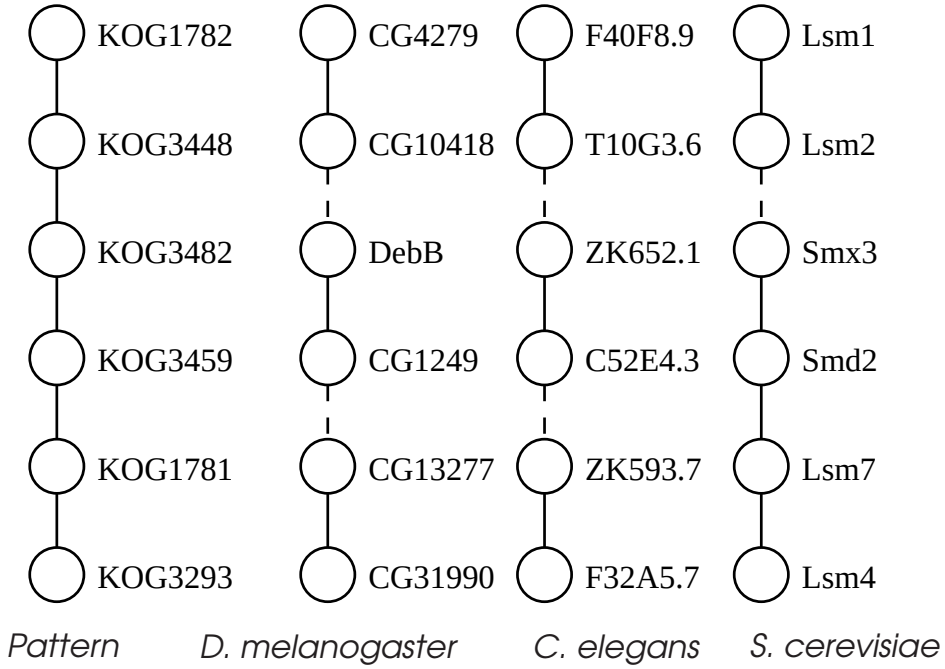
What is the "infomative" component of a set of PPI networks?

- PPI networks for 9 eukaryotic organisms derived from BIND and DIP
  - *A. thaliania, O. sativa, S. cerevisiae, C. elegans, D. melanogaster, H. sapiens, B. taurus, M. musculus, R. norvegicus*
  - # of proteins ranges from 288 (*Arabidopsis*) to 8577 (*fruit fly*)
  - # of interactions ranges from 340 (*rice*) to 28829 (*fruit fly*)

- Ortholog contraction
  - Group proteins according to existing COG ortholog clusters
  - Merge Homologene groups into COG clusters
  - Cluster remaining proteins via BLASTCLUST
  - Ortholog-contracted *fruit fly* network contains 11088 interactions between 2849 ortholog groups

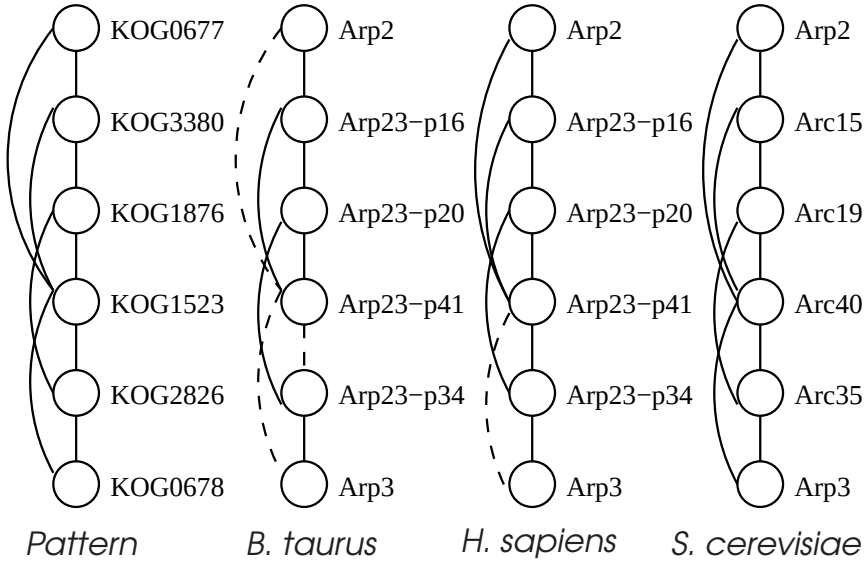
- MULE is available at
  `http://www.cs.purdue.edu/pdsl/`

Conserved Protein Interaction Patterns



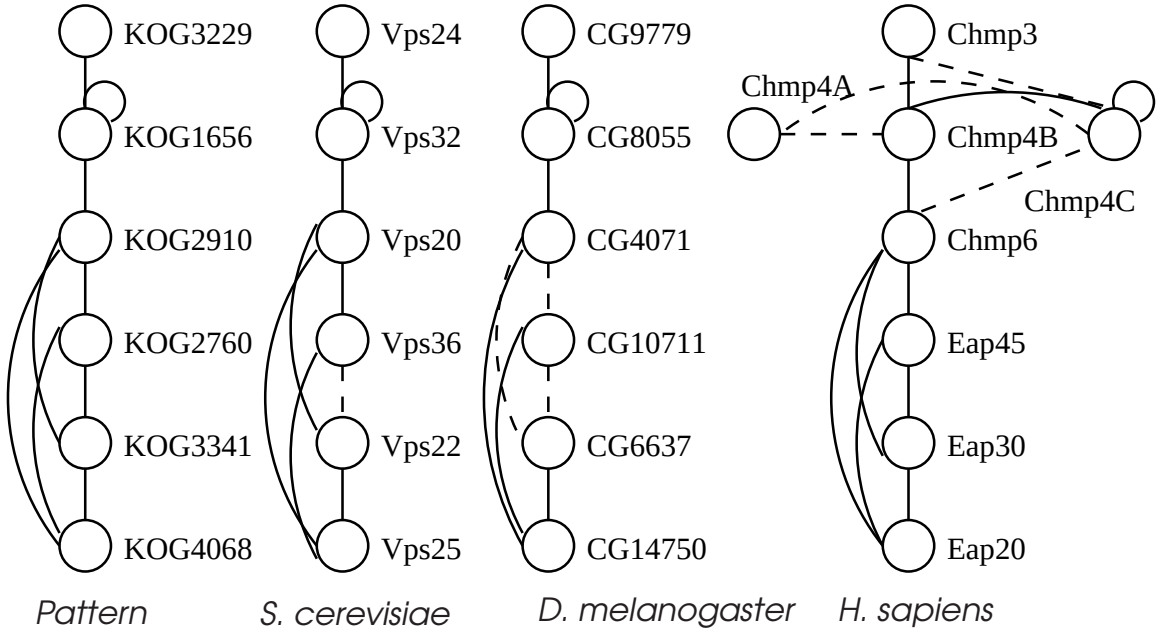
Small nuclear ribonucleoprotein complex ($p < 2e - 43$)

Conserved Protein Interaction Patterns



Actin-related protein Arp2/3 complex ($p < 9e - 11$)

Conserved Protein Interaction Patterns



Endosomal sorting ($p < 1e - 78$)

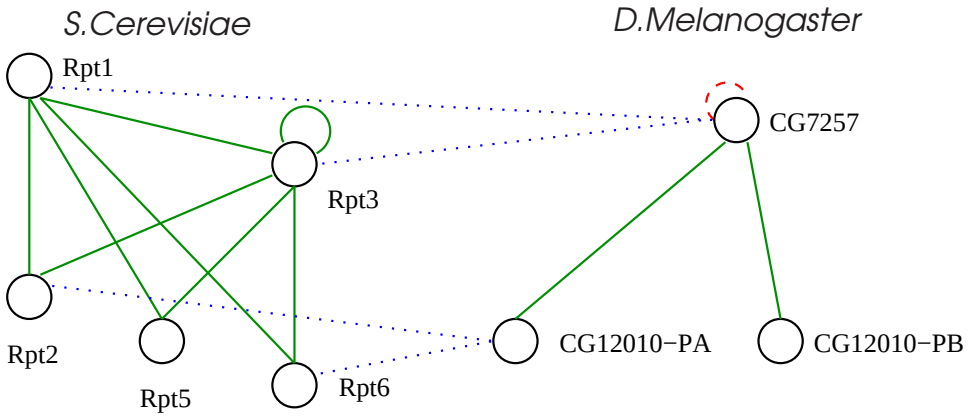
Role of Information in Scientific Discovery

What is the "informative" component shared by two given PPI networks (Yeast and Fruit Fly)?

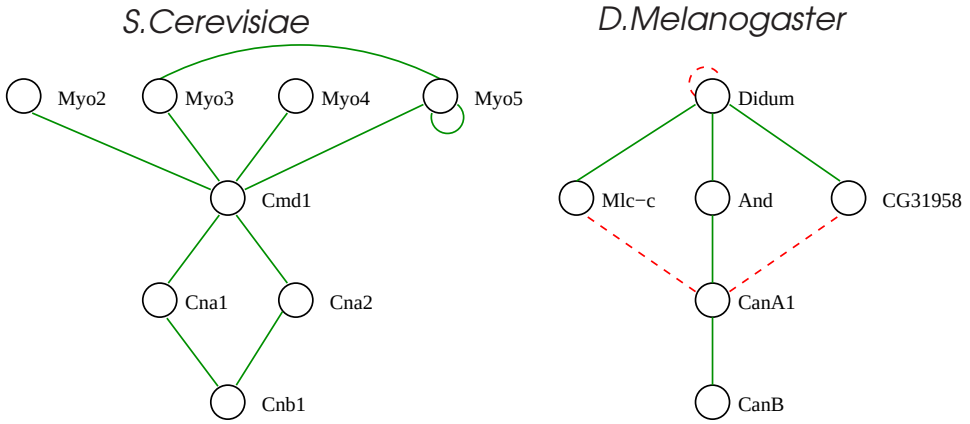
Rank	Score	z-score	# Proteins	# Matches	# Mismatches	# Dups.
1	15.97	6.6	18 (16, 5)	28	6	(4, 0)
	protein amino acid phosphorylation (69%)					
	JAK-STAT cascade (40%)					
2	13.93	3.7	13 (8, 7)	25	7	(3, 1)
	endocytosis (50%) / calcium-mediated signaling (50%)					
5	8.22	13.5	9 (5, 3)	19	11	(1, 0)
	invasive growth (sensu Saccharomyces) (100%)					
	oxygen and reactive oxygen species metabolism (33%)					
6	8.05	7.6	8 (5, 3)	12	2	(0, 1)
	ubiquitin-dependent protein catabolism (100%)					
	mitosis (67%)					
21	4.36	6.2	9 (5, 4)	18	13	(0, 5)
	cytokinesis (100%, 50%)					
30	3.76	39.6	6 (3, 5)	5	1	(0, 6)
	DNA replication initiation (100%, 80%)					

Subnets Conserved in Yeast and Fruit Fly

Proteosome regulatory particle subnet

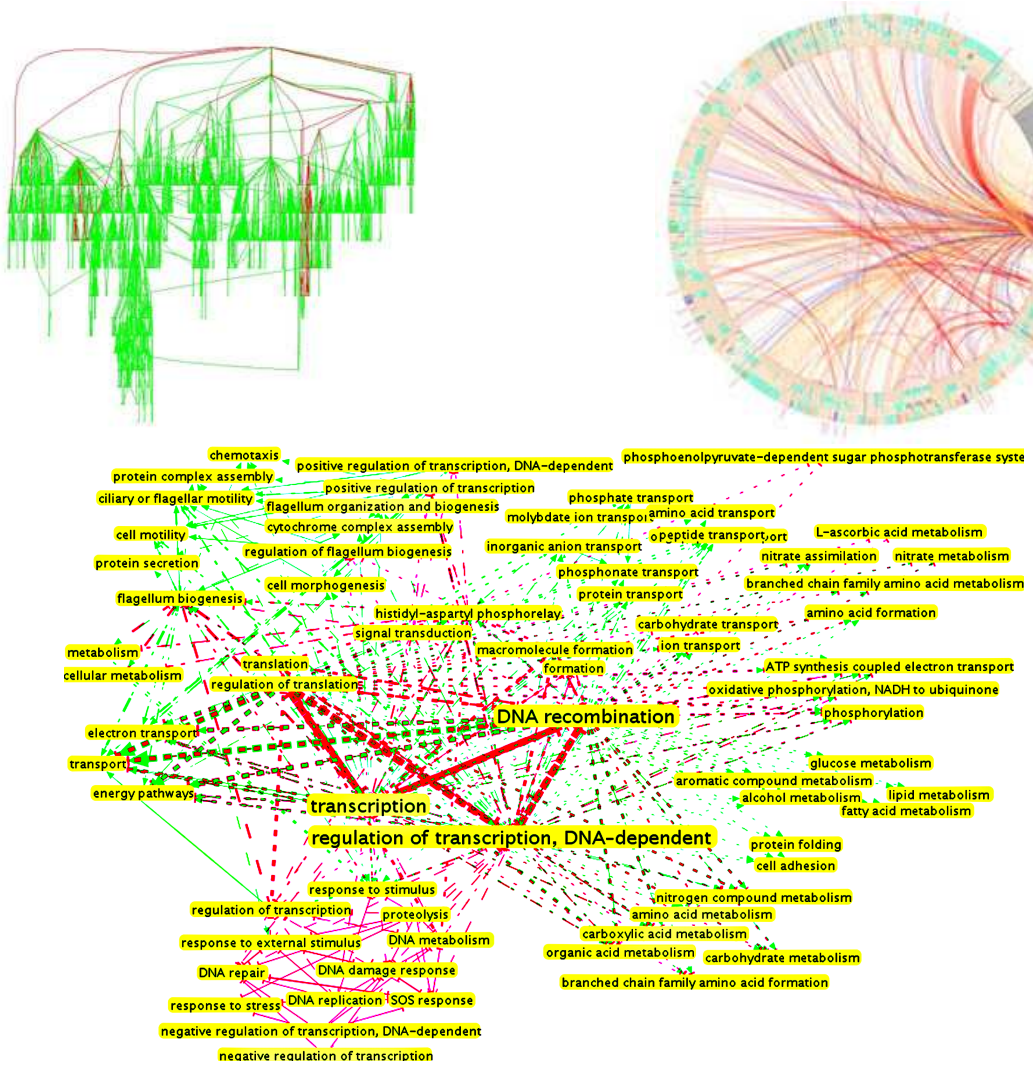


Calcium-dependent stress-activated signaling pathway



Role of Information in Scientific Discovery: Examples

What are statistically significant functional pathways in gene regulatory networks?

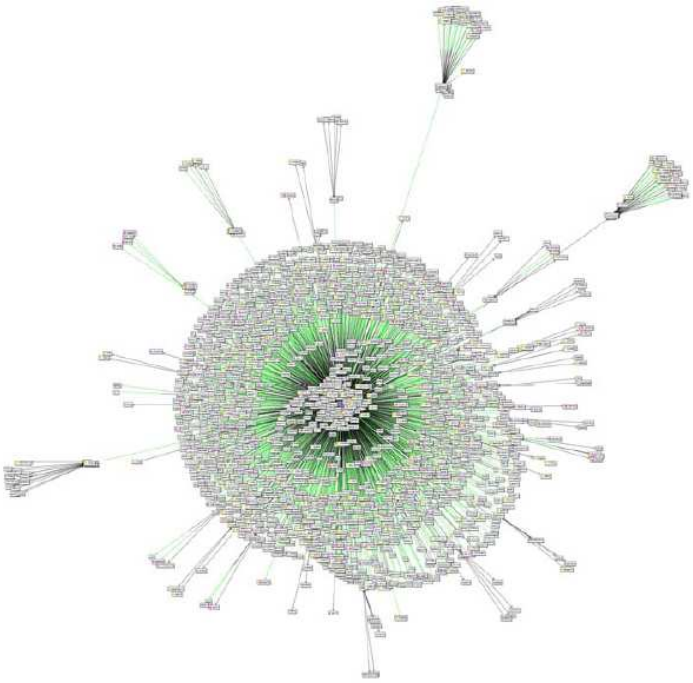


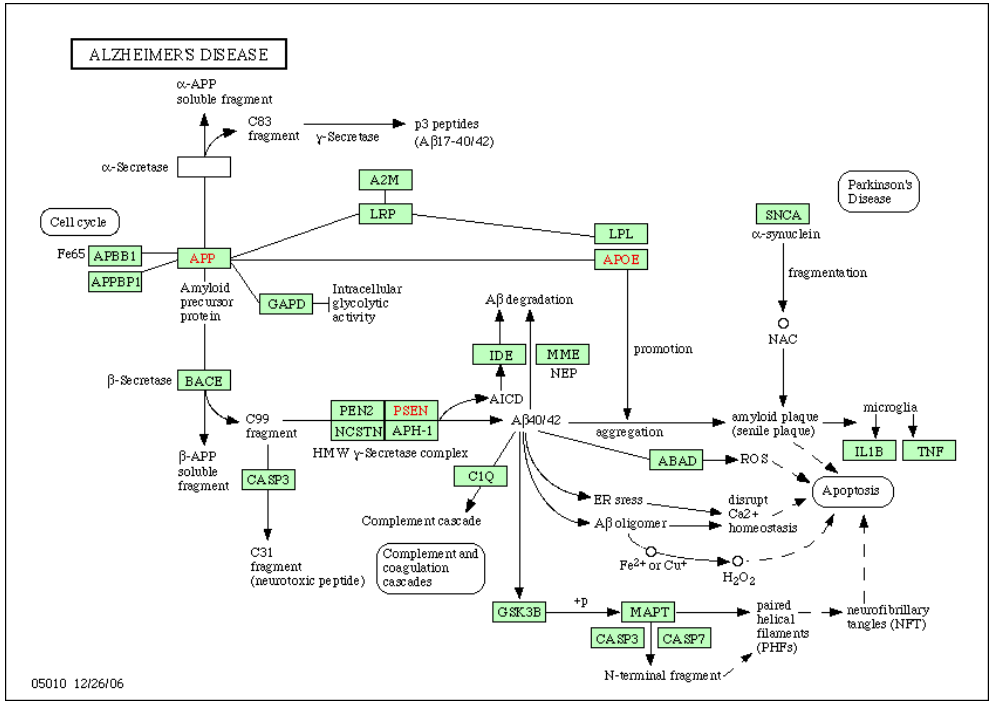
Role of Information in Scientific Discovery: Examples

Frequency	p-value	Pathway
276	5E-94	metabolic process $\dashv$ flagellum biogenesis $\longrightarrow$ transport
136	3.1E-71	regulation of translation $\dashv$ DNA recombination $\longrightarrow$ transport
38	4.9E-47	response to stimulus $\dashv$ transcription $\longrightarrow$ cell motility
36	6.6E-35	flagellum biogenesis $\longrightarrow$ ciliary or flagellar motility
56	1.4E-24	regulation of translation $\dashv$ transcription $\longrightarrow$ carboxylic acid metabolism
178	8.3E-21	signal transduction $\dashv$ transcription $\longrightarrow$ transport
14	8.6E-20	phosphate transport $\longrightarrow$ transcription $\longrightarrow$ phosphonate transport
16	2E-16	SOS response $\dashv$ regulation of transcription $\dashv$ DNA repair
501	1.2E-13	regulation of transcription, DNA-dependent $\longrightarrow$ transport
12	3.6E-10	proteolysis $\dashv$ regulation of transcription $\dashv$ response to external stimulus
15	3.8E-7	nitrate assimilation $\dashv$ cytochrome complex assembly
10	1.4E-6	cell morphogenesis $\dashv$ protein secretion
178	3.8E-4	transcription $\longrightarrow$ carbohydrate metabolic process

Information Sciences in Life Sciences

Characterizing phenotype and disease in networks.





Detecting "informative" parts of networks is an essential aspect of understanding disease and remediation (Alzheimers).

Possibilities Abound