BIOE 505: Computational Bioengineering

What this class is all about?

Instructor

Name: Sergei Maslov

 Professor of Bioengineering, Physics, Carl R. Woese
 Institute for Genomic Biology, and
 National Center for Supercomputing
 Applications

 Office: 3103 Carl Woese Institute for Genomic Biology and sometimes 3146C Everitt Laboratory (both by appointment)

- E-mail: maslov@illinois.edu
- Phone: 217-265-5705



Questions and Suggestions:

maslov@Illinois.edu

Start subject with [BIOE505]

Grading

- Midterm exam 40%
- Final exam 60%
- Homework (ungraded) will be posted online.
 Solutions will be posted in a week.
- Homework will build on topics covered in lectures and will consist of problem sets related to topics covered in lectures
- Useful to prepare for exams

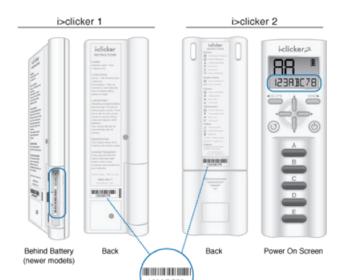
Course Website

https://courses.engr.illinois.edu/bioe505 Grades will be on https://my.bioen.illinois.edu/gradebook

\leftarrow \rightarrow C \triangle \triangleq courses.grainge	er.illinois	.edu/bioe5	05/fa2019/index.html					
🏥 Apps ★ 🕙 🔶 🔥 Recent 💰	Group	🛟 Librai	ry@RC 🧭 🚟 Ludwig • Fi	nd your 🧵 WebEdit 🤅	BIOE			
	$\leftarrow \rightarrow$	C 🗅	Courses.grainger.illinois	.edu/bioe505/fa2019/schedu	le.html		🖈 🔶 I 📷	5 🕨 🖣
BIOE 505 - Comp	Apps	5 ★ 🕙	🔶 🝐 Recent 💧 Group	🐉 Library@RC 🧭 🚟 Lu	udwig • Find your 👖 🤉	WebEdit 🚯 BIOE 505 Prof.	Ser I BIOE 505 Cour	rse E
<u>Schedule</u>	BIO)E 50	5 - Computat	ional Bioengi	neering			
Instructor								
Sergei Maslov: <u>maslov@illinois.edu</u> Office: IGB 3406 Office hours: by appointment								
Logistics	#	Date	Topics	Slides	Matlab	Homework	Exams	
Tuesdays: 12:00AM - 1:50AM Thursdays: 12:00AM - 1:50PM								
106B8 Engineering Hall	1	Aug 27						
I WANT YOU TO BRING YOUR i-CL]							
I WANT YOU TO BRING YOUR OWN LAPTOPS SHOULD HAVE MATLAB	-							
STATISTICS AND MACHINE LEARN	2	Aug 29						
Description								

Bring your iClickers to my lectures

- Who knows what is an iClicker?
- Show of hands: who has an iClicker?
- I would like you all to have an iClicker and bring it to every class.
 On amazon.com a new iClicker (1st generation is OK) costs around \$40. It is also sold at UIUC Bookstore. The used ones are cheaper.
- An alternative solution is using a mobile app: <u>https://www.iclicker.com/students/apps-and-remotes/apps</u>
- Your answers <u>WILL NOT</u> be used for grading.
 I need them to see if I lost some of you and what could
 I rephrase to better explain the material



Who has Matlab?

- A. Already have it installed on my laptop
- B. Will install it (starting this year it is free!)
- C. Plan to access it on EWS via CITRIX
- D. I don't know yet
- E. I will never use Matlab! Why don't we use Python?

Get your i-clickers

We will use Matlab in class

- Bring your laptops to class
- Need to have Matlab installed and know the basic user interface (inline commands, plotting)
- We will use Statistics and Machine Learning Toolbox and Bioinformatics Toolboxes
- Good news! Now all faculty and graduate students get Matlab for free. See <u>offering on the WebStore</u> site and follow the <u>detailed instructions</u>.
- .m files and .mat with Matlab commands and data will be on the website after the lecture

Possible alternative to purchasing Matlab and toolboxes is to use campus resources.

Both Engineering Workstations (EWS) and ACES computers have Matlab. I don't think all of them offer the statistics and bioinformatics toolboxes (EWS should, ACES computers may not..).

See the following to access:

Citrix for EWS, Matlab, and ACES computers -- links for all

https://it.engineering.illinois.edu/ews/lab-information/remote-connections/connecting-citrix https://it.engineering.illinois.edu/services/instructional-services/remote-connections-citrix

Accessing Engineering Workstations (EWS)

https://it.engineering.illinois.edu/ews

Accessing ACES Academic Computing Workstations

http://acf.aces.illinois.edu/remote/ http://acf.aces.illinois.edu/remote/pc.html

To access off campus use:

CISCO Virtual Private Network -- <u>For off-campus access to campus computer and network resources</u> (software programs, files saved on the network, etc.)

https://techservices.illinois.edu/services/virtual-private-networking-vpn/download-and-set-up-the-vpnclient

CISCO VPN CLIENT

https://webstore.illinois.edu/shop/product.aspx?zpid=2600

CISCO AnyConnect VPN https://webstore.illinois.edu/shop/product.aspx?zpid=1222

What will you learn in this course?

- Basics of probability and statistics
 - Basic concepts of probability, Bayes theorem
 - Discrete and continuous probability distributions
 - Multivariate statistics
 - Sampling distributions
 - Parameter estimation
 - Hypothesis testing
 - Regression
- How it is applied to biological data
 - Basics of genomics
 - Systems biology (gene expression, networks)

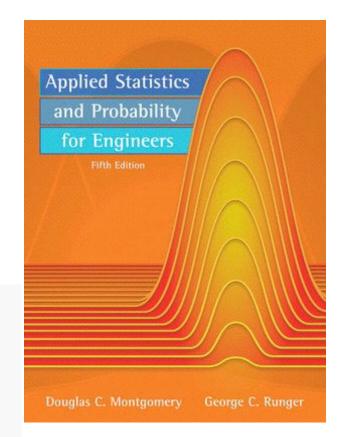
The main Probability/Statistics Textbook

Applied Statistics and Probability for Engineers, 5th Edition D. C. Montgomery and G. C. Runger John Wiley & Sons, Inc. (2011)

You can also use other editions from 4^{th} (2007) to 6^{th} (2014)

5th edition is available for free at our library





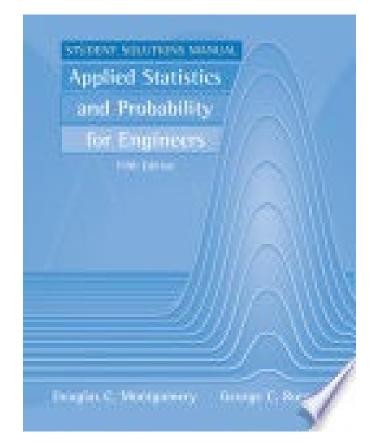
Problems for our main Probability/Statistics Textbook

Student Solutions Manual Applied Statistics and Probability for Engineers, 5th Edition D. C. Montgomery and G. C. Runger John Wiley & Sons, Inc. (2010)

You can also use other editions from 4th (2007) to 6th (2014)

5th edition is available for free at our library





Probability/Statistics for Bioengineering with Matlab exercises

Statistics for Bioengineering Sciences with MATLAB and WinBUGS Support

Brani Vidakovic

Department of Biomedical Engineering, Georgia Tech

(2011) Springer, New York

It is constantly updated with the newest version at the link below.

Springer Texts in Statistics		
Brani Vidakovic		
Statistics		

for Bioengineering Sciences

With MATLAB and WinBUGS Support

Deringer

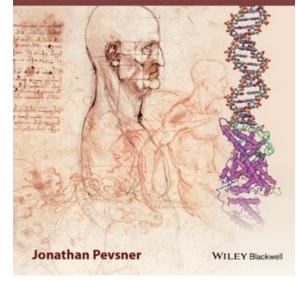
Free as a PDF eBook at <u>http://statbook.gatech.edu/statb4.pdf</u> Matlab exercises and datasets are at <u>http://springer.bme.gatech.edu</u>

Genomics/Systems Biology Textbook

J Pevsner
 Bioinformatics and functional genomics
 Wiley-Blackwell,
 2nd edition [2009] exists in electronic form
 3rd edition [2015] has up-to-date
 information on NGS: RECOMMENDED
 (about \$60 on amazon)

 2nd edition is available for free in electronic form in our library





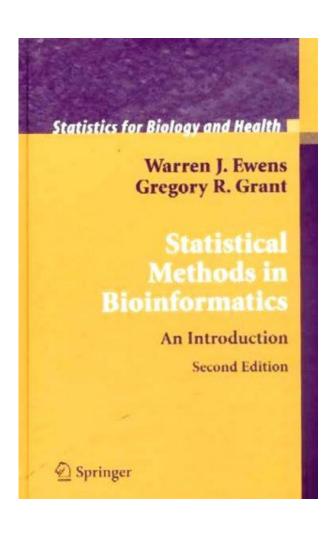


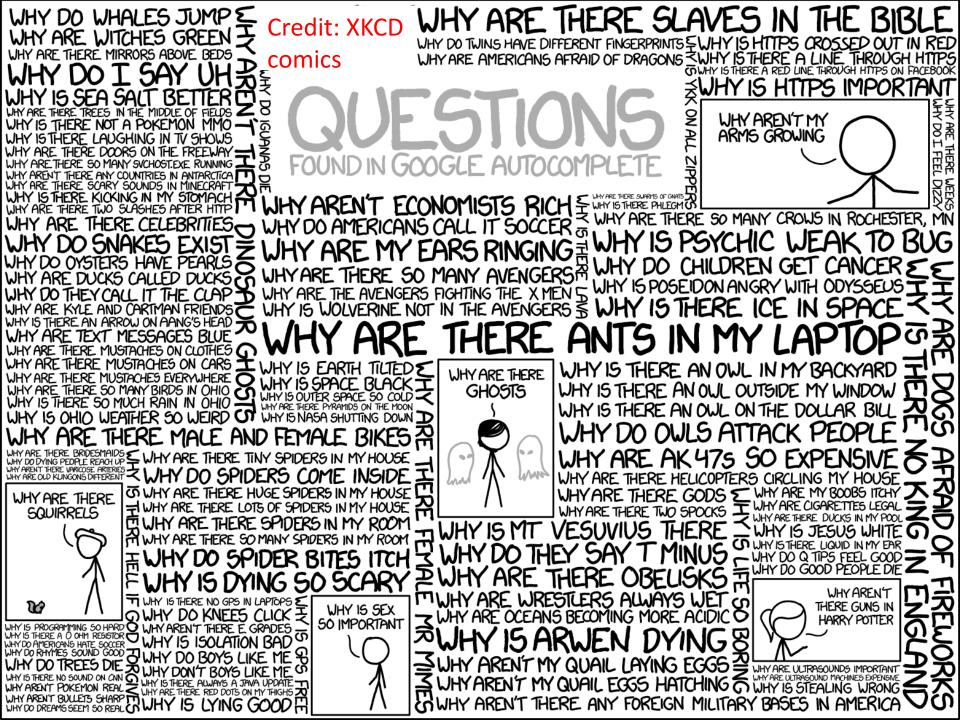
Another Bioinformatics/Statistics Textbook

 Ewens, WJ and Grant, GR Statistical Methods in Bioinformatics: An Introduction, 2nd ed, Springer, 2005.

2nd edition as PDF eBook







This course is about biological data and probability theory, and statistics concepts needed for its analysis

What biological data will be discussed?

<u>Will be covered in lectures or Matlab exercises:</u>

- <u>Genomic data</u>: strings of letters ACGT
- <u>Gene Expression data</u>: messenger RNA copy numbers transcribed from genes
- <u>Proteomic data</u>: protein abundances
- <u>Network data</u>: pairs of interacting genes or proteins and protein-protein interaction strengths

Will not be covered:

• Imaging data such as e.g. fMRI brain scans, Brain connectome data, Ecosystem dynamics data Why do you need probability and statistics to analyze modern biological data?

Definition of probability theory by Encyclopedia Britannica

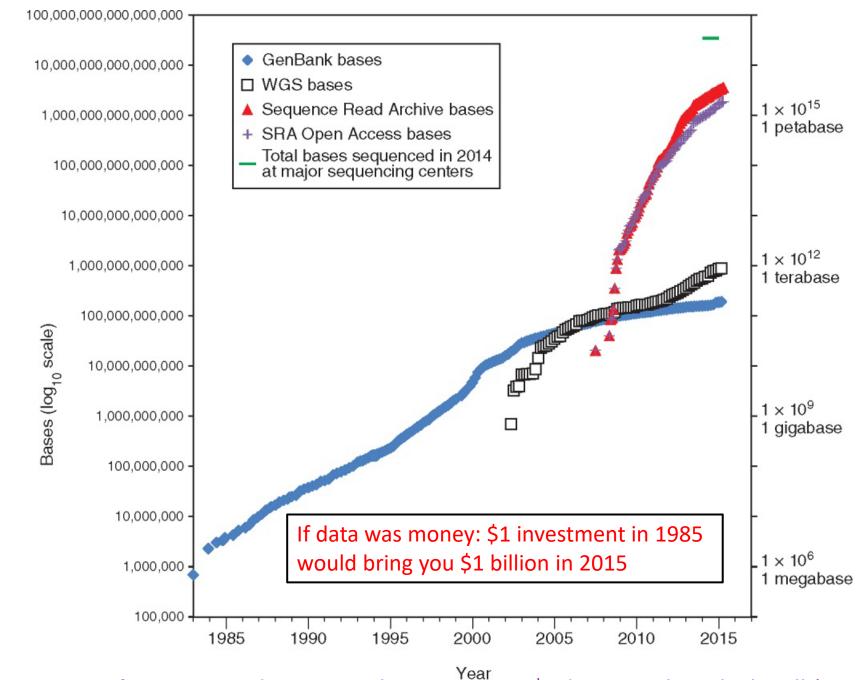
a branch of mathematics concerned with the analysis of random phenomena

Definition of *statistics* by Merriam-Webster

1: a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data

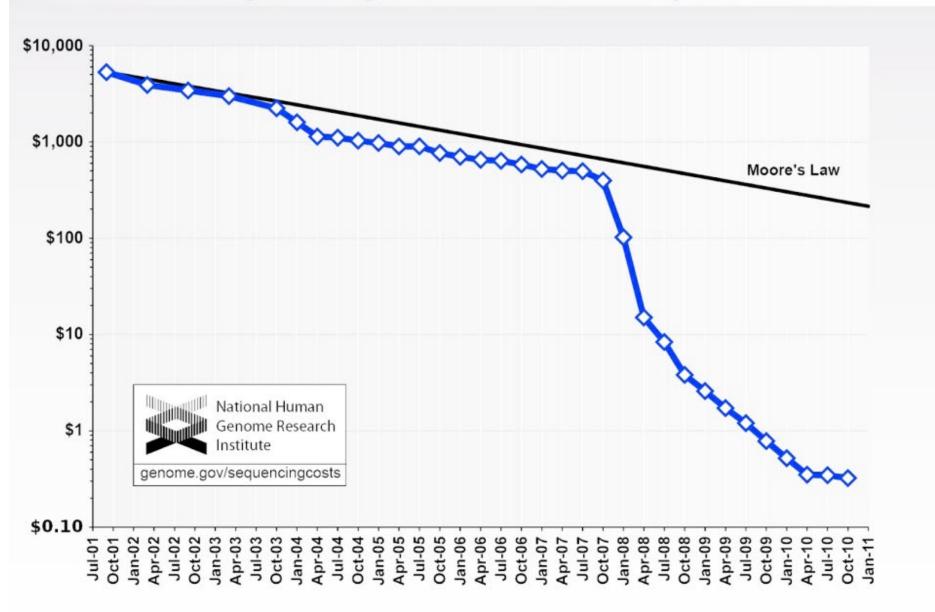
Why do you need probability and statistics to analyze modern biological data?

Reason 1: Biology now has Lots of Data



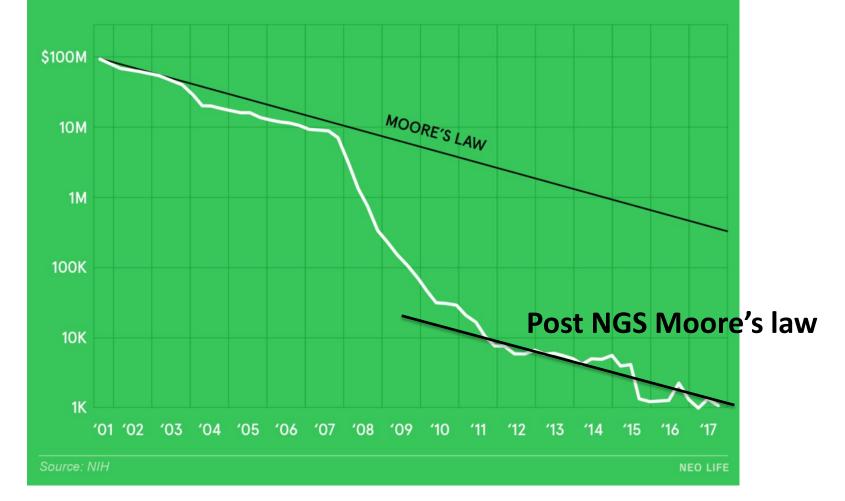
J. Pevsner, Bioinformatics and Functional Genomics, 3rd edition, Wiley-Blackwell (2015)

Cost per Megabase of DNA Sequence



Cost per Genome Sequenced

The cost of sequencing a human genome compared with the reductions that would be expected at the rate Moore's law predicts for computer chips. Over the past decade, next-generation sequencing and cloud computing drove the figure down. The average bumped higher in recent years because of brief slowdowns in production.



Who will have bigger data by 2025?

Data Phase	Astronomy	Twitter			
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year			
Storage	1 EB/year	1–17 PB/year			
Peta=1	.0 ¹⁵ Exa=10 ¹⁸	Zetta=10 ²¹			
YouTube	<u>)</u>	Genomics			
500-900	million hours/year	1 zetta-bases/year			
1–2 EB/y	ear	2–40 EB/year			

Z. Stephens, S. Lee, F. Faghri, R. Campbell, C. Zhai, M. Efron, R. Iyer, M. Schatz, S. Sinha, and G. Robinson (2015) PLoS Biol 13: e1002195.

Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	A, C, G, T = 2 bits = 0.25 bytes
1000	1 kilobase pair	1 kb	
1,000,000	1 megabase pair	1 Mb	
10°	1 gigabase pair	1 Gb	3
1012	1 terabase pair	1 Tb	
1015	1 petabase pair	1 Pb	

Size	Abbreviation	No. bytes	Examples
Bytes	_	1	1 byte is typically 8 bits, used to encode a single character of text
Kilobytes	1 kb	10 ³	Size of a text file with up to 1000 characters
Megabytes	1 MB	10 ⁶	Size of a text file with 1 million characters
Gigabytes	1 GB	10 ⁹	600 GB: size of GenBank (uncompressed flat files) ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt (WebLink 2.84)
Terabytes	1 TB	1012	385 TB: United States Library of Congress web archive (http://www.loc.gov/ webarchiving/faq.html) (WebLink 2.85)
			464 TB: Data generated by the 1000 Genomes Project (http://www.1000genomes.org/ faq/how-much-disk-space-used-1000-genomes-project) (WebLink 2.86)
Petabytes	1 PB	1015	1 PB: size of dataset available from The Cancer Genome Atlas (TCGA)
			5 PB: size of SRA data available for download from NCBI
			15 PB: amount of data produced each year at the physics facility CERN (near Geneva) (http://home.web.cern.ch/about/computing) (WebLink 2.87)
Exabytes	1 EB	1018	2.5 exabytes of data are produced worldwide (Lampitt, 2014)

J. Pevsner, Bioinformatics and Functional Genomics, 3rd edition, Wiley-Blackwell (2015)

What makes genomic data so big?

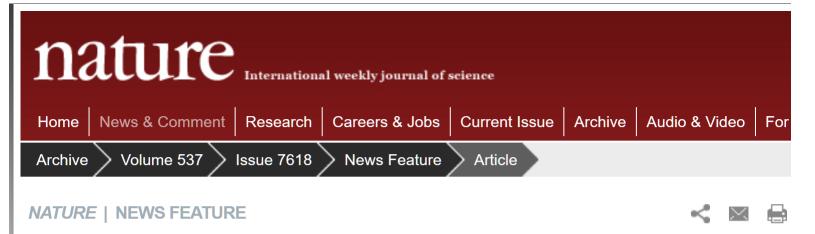
- There are ~9 millions species each with its own genome
- Each of us humans (7.5 billions and counting) has unique DNA: we want to compare them all to each other
- Each cell has just 1 genome (DNA) but multitude of transcriptomes (RNA levels) and proteomes (protein levels)
- Cancer cells acquire mutations in their genomes: need to track multiple lineages in a tumor vs time to understand cancer
- DNA was proposed as a long-term storage medium of information

Farfetched? Storage standards evolve fast but DNA standard remained unchanged for 4 billion years

Note: Nature article started the comparison with a hard drive and flash memory skipping the floppy disk







How DNA could store all the world's data

Modern archiving technology may hold an answer to that pro

STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

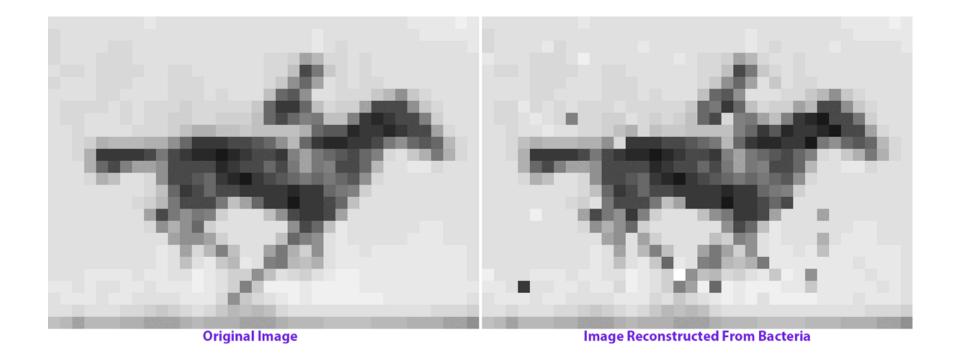
Andy Extance

31 August 2016

	Hard	Flash	Bacterial DNA	WEIGHT OF DNA NEEDED TO STORE WORLD'S
Read-write speed (µs per bit) >	~3,000– 5,000	~100	<100	DATA
Data retention (years) >	>10	>10	>100	Ă
Power usage (watts per gigabyte) >	~0.04	~0.01–0.04	<10-10	~1 kg
Data density (bits per cm ³) >	~1013	~1016	~1019	©nature ture.com/2bRwUIL

- Prof Olgica Milenkovic from Electrical and Computer Engineering UIUC is a local expert on this topic
- Profs. George Church and Sri Kosuri (Harvard Medical School) explains a potential use of DNA as storage medium in 2012
- <u>https://www.youtube.com/watch?v=IJAdqAVjQqY</u>

Fast-forward from 2012 to 2017



Shipman SL, Nivala J, Macklis JD, Church GM. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature. 2017;547: 345–349. doi:10.1038/nature23017 Why do you need probability and statistics to analyze modern biological data?

Reason 2: Life is random and messy

Show video "Cell organelles"

- Made at the Walter and Eliza Hall Institute of Medical Research at Victoria, Australia
- Animated by award-winning artist Dr. Drew Berry
- Go to <u>https://www.wehi.edu.au/wehi-tv</u> for other videos

Life is messy, random, and noisy

Yet it is beautifully complex and has many parts (see statistics)

Why life is so random?

- Biomolecules are very small (nano- to micro-meters) → Brownian noise
- # molecules/cell is often small → Large cell-to-cell variations
- Genomic data comes from biological evolution

 the Mother of all random processes
- Genomic data involves (random) samples
 - We have genomes of some (not all) organisms
 - We have tissue samples of some (not all) cancer patients

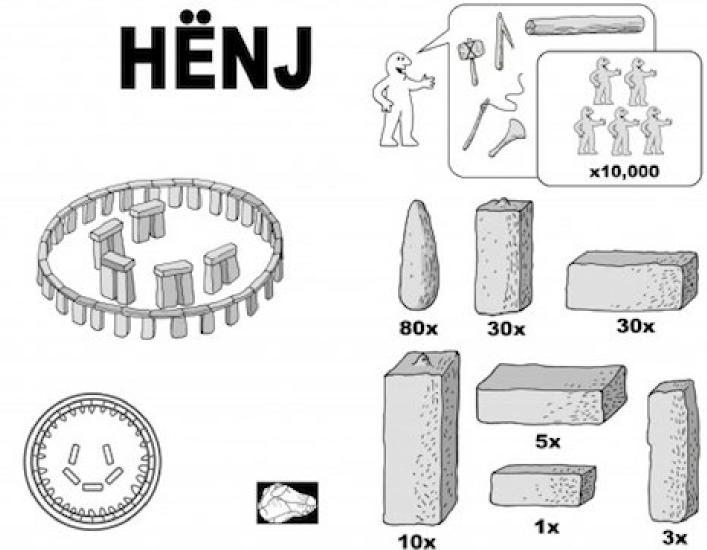
Why life is so complex?

Primer on complex system

Complex systems have many interacting parts

- All parts are different from each other
 - 10s thousands (10⁴) types of proteins in an organism
 - 100 thousands (10⁵) organizations (AS) in the Internet
 - 1 billion (10⁹) people on Facebook
 - -10 billion (10¹⁰) web pages in the WWW
 - 100 billion (10¹¹) neurons in a human brain
 - NOT 10²³ electrons or quarks studied by physics: they are all the same and boring!
- Yet they share the same basic design
 - All proteins are strings of the same 20 amino acids
 - All WWW pages use HTML, JavaScript, etc.
 - All neurons generate and receive electric spikes

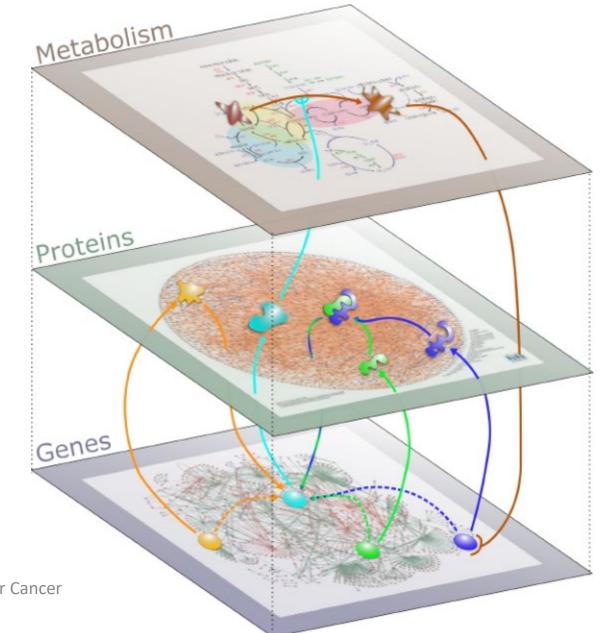
Example: a complex system with many parts



Justin Pollard, http://www.designboom.com

Parts interact \rightarrow they need to be assembled to work - Aller 100m 2 Miles **x30** 30x

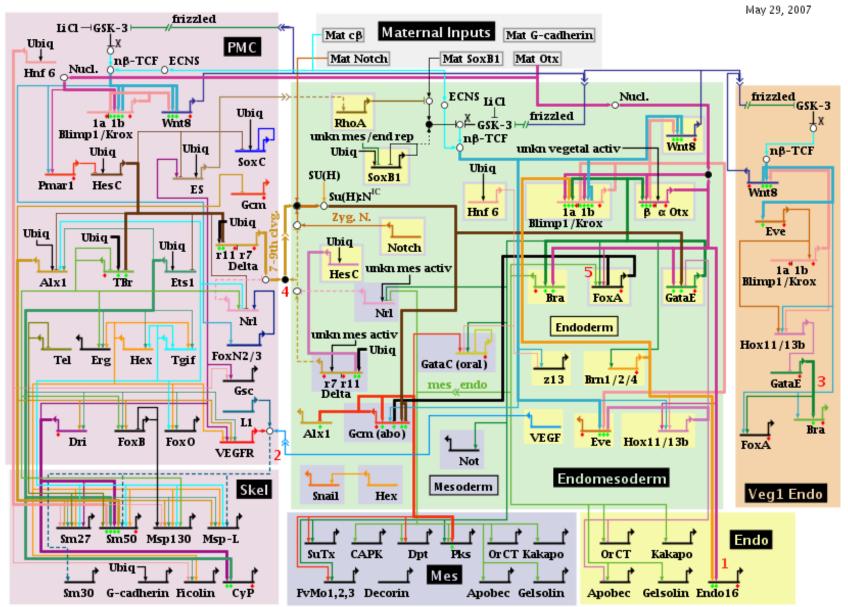
Justin Pollard, http://www.designboom.com Intra-cellular Networks operate on multiple levels



Slides by Amitabh Sharma, PhD

Northeastern University & Dana Farber Cancer Institute

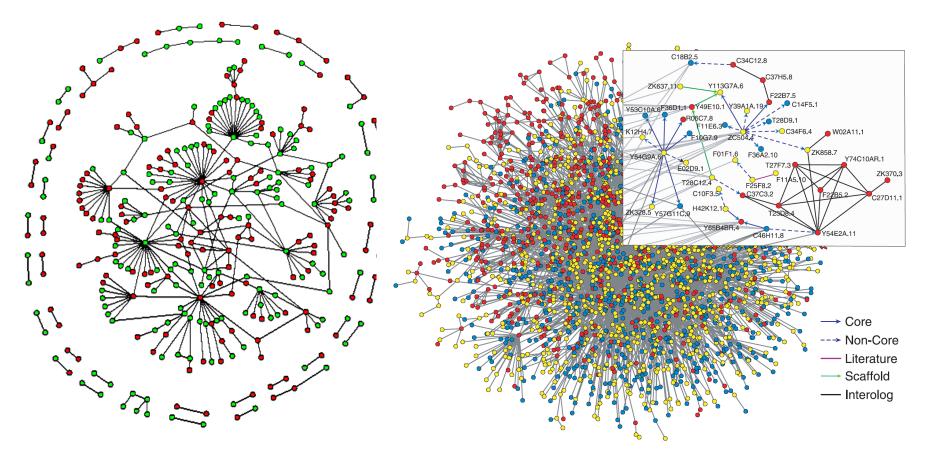
Sea urchin embryonic development (from endomesoderm up to 30 hours) by Davidson's lab



Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor; unkn = unknown; Nucl. = nuclearization; $\chi = \beta$ -catenin source; n β -TCF = nuclearized b- β -catenin-Tcf1; ES = early signal; ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Copyright © 2001–2007 Hamid Bolouri and Eric Davidson

Protein-Protein binding IntAct Database (Dec 2015) Interactions: 577,297 Proteins: 89,716



Baker's yeast *S. cerevisiae* (only nuclear proteins shown) From S. Maslov, K. Sneppen, Science 2002 Worm *C. elegans* From S. Lee et al , Science 2004

Metabolic pathway chart by ExPASy: **5702 reactions as of December 2015**

	Α	В	С	D	E	F	G	Н	I	J	К	L
												10
1						Salaras						
•	THI	WILE E	TYEME A		an an an an							
2				No worker	THE TENE							En in
-									The Carlos		Fraction and	
3			YP		182 1					TRANGE		
	Can P	TT- OB									are and a	
4	Sta Stand									N KOZ		
5							A BE	The sur the				and a standard stand Standard standard stan
-	天公										and and a state	
6												
Ŭ		1 Arger		Received and the second				and damage				ndicide s p lactic
7												
										的复		
8	WARN C.								a the factor and			
				an gran a gan						-32-6-10-		
9			A Inter		Contraction of the second		Part and					
					Harter Alternet							
10						E E	and the second				- 38 - 18 - 19 - 19 - 19 - 19 - 19 - 19 - 1	
	T.					C-2-5-						

Brain and nerves of a worm



- Worm (C. elegans) has 302 neurons
- Our brain has 100 billion (10¹¹) neurons

