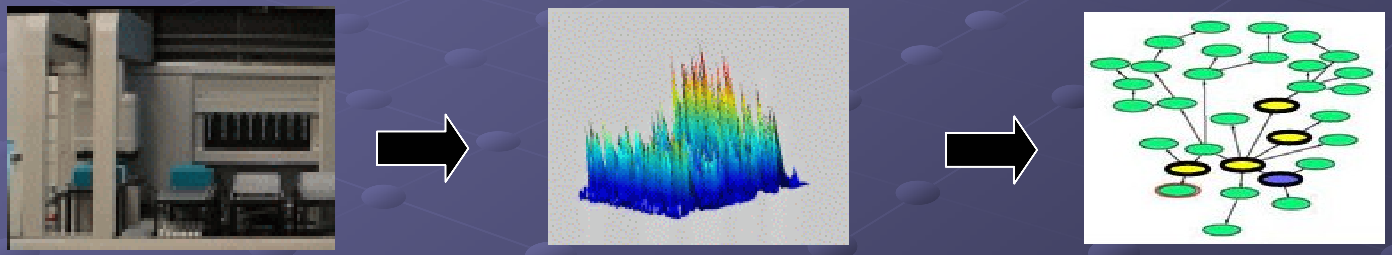
6.092/HST.480
Bioinformatics & Proteomics:
An Engineering-Based Problem Solving Approach



Gil Alterovitz[1], Manolis Kellis[2], Marco Ramoni[1]

[1] Harvard/MIT Division of Health Science and Technology (HST)
[2] Electrical Engineering & Computer Science, MIT

Harvard-MIT
Division of Health
Science & Technology

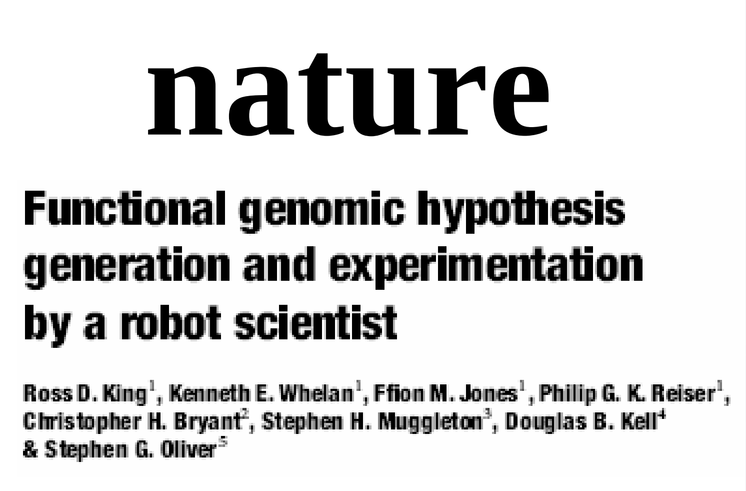
Today

- Introduction- Gil Alterovitz
  - Motivation: Why Bioinformatics?
  - Course Introduction
  - Introduction to Modern Biology: Part I
- Bioinformatics from Industry's Perspective: Mathworks- Rob Henson
  - Bioinformatics in Industry
  - Matlab Bioinformatics Toolbox
  - Clustering and Related Technologies (DeRisi's Microarray Paper)

Motivation: Why Bioinformatics?

Why Engineering and Computer Science?

Robotics/Automation: For lab automation, hypothesis testing and generation

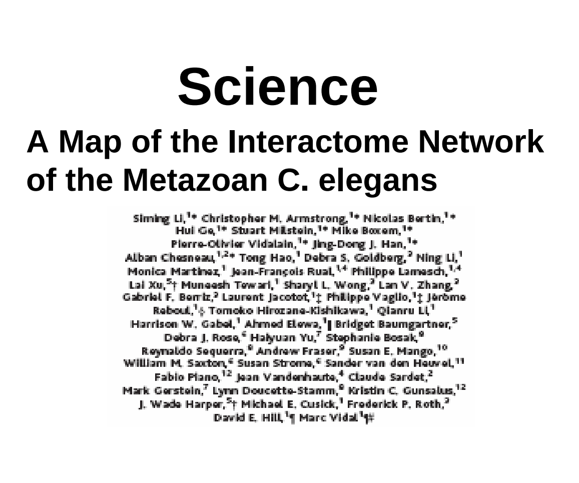


nature

Functional genomic hypothesis generation and experimentation by a robot scientist

Ross D. King[1], Kenneth E. Whelan[1], Ffion M. Jones[1], Philip G. K. Reiser[1],
Christopher H. Bryant[2], Stephen H. Muggleton[3], Douglas B. Kell[4]
& Stephen G. Oliver[5]

Network Theory: Modeling Protein Interaction

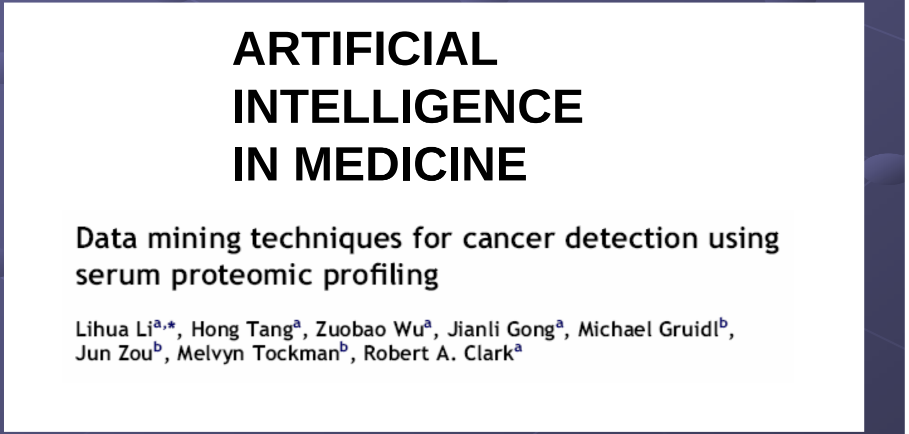


Science

A Map of the Interactome Network of the Metazoan C. elegans

Siming Li,[1*] Christopher M. Armstrong,[1*] Nicolas Bertin,[1*]
Hui Ge,[1*] Stuart Milstein,[1*] Mike Boxem,[1*]
Pierre-Olivier Vidalain,[1*] Jing-Dong J. Han,[1*]
Alban Chesneau,[1,2*] Tong Hao,[1] Debra S. Goldberg,[3] Ning Li,[1]
Monica Martinez,[1] Jean-François Rual,[1,4] Philippe Lamesch,[1,4]
Lai Xu,[5†] Muneesh Tewari,[1] Sharyl L. Wong,[3] Lan V. Zhang,[3]
Gabriel F. Berriz,[3] Laurent Jacotot,[1‡] Philippe Vaglio,[1‡] Jérôme
Reboul,[1§] Tomoko Hirozane-Kishikawa,[1] Qianru Li,[1]
Harrison W. Gabel,[1] Ahmed Elewa,[1‖] Bridget Baumgartner,[5]
Debra J. Rose,[6] Haiyuan Yu,[7] Stephanie Bosak,[8]
Reynaldo Sequerra,[8] Andrew Fraser,[9] Susan E. Mango,[10]
William M. Saxton,[6] Susan Strome,[6] Sander van den Heuvel,[11]
Fabio Piano,[12] Jean Vandenhaute,[4] Claude Sardet,[2]
Mark Gerstein,[7] Lynn Doucette-Stamm,[8] Kristin C. Gunsalus,[12]
J. Wade Harper,[5†] Michael E. Cusick,[1] Frederick P. Roth,[3]
David E. Hill,[1¶] Marc Vidal[1¶#]

Visualization/Image Processing: Protein Expression 3-D Heat Map

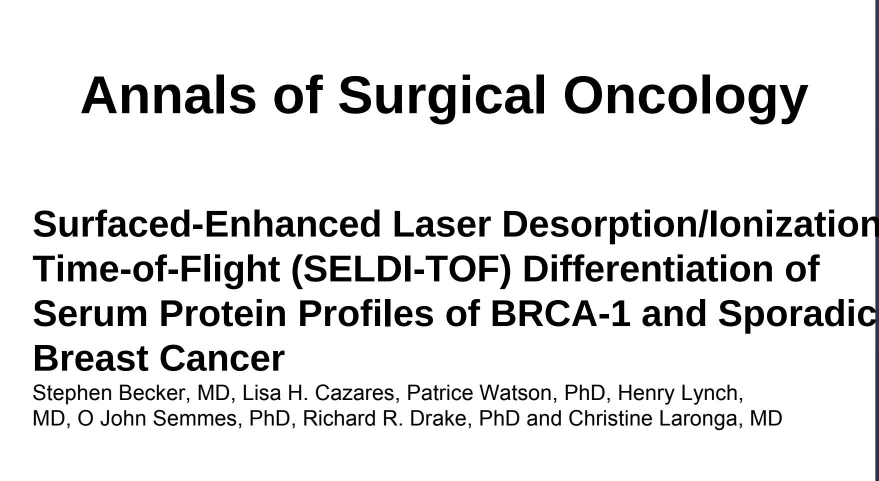


ARTIFICIAL INTELLIGENCE IN MEDICINE

Data mining techniques for cancer detection using serum proteomic profiling

Lihua Li[a,*], Hong Tang[a], Zuobao Wu[a], Jianli Gong[a], Michael Gruidl[b],
Jun Zou[b], Melvyn Tockman[b], Robert A. Clark[a]

New Mass Spectrometry Technologies With Clinical Applications



Annals of Surgical Oncology

Surfaced-Enhanced Laser Desorption/Ionization Time-of-Flight (SELDI-TOF) Differentiation of Serum Protein Profiles of BRCA-1 and Sporadic Breast Cancer

Stephen Becker, MD, Lisa H. Cazares, Patrice Watson, PhD, Henry Lynch,
MD, O John Semmes, PhD, Richard R. Drake, PhD and Christine Laronga, MD

Signal Processing in Biology

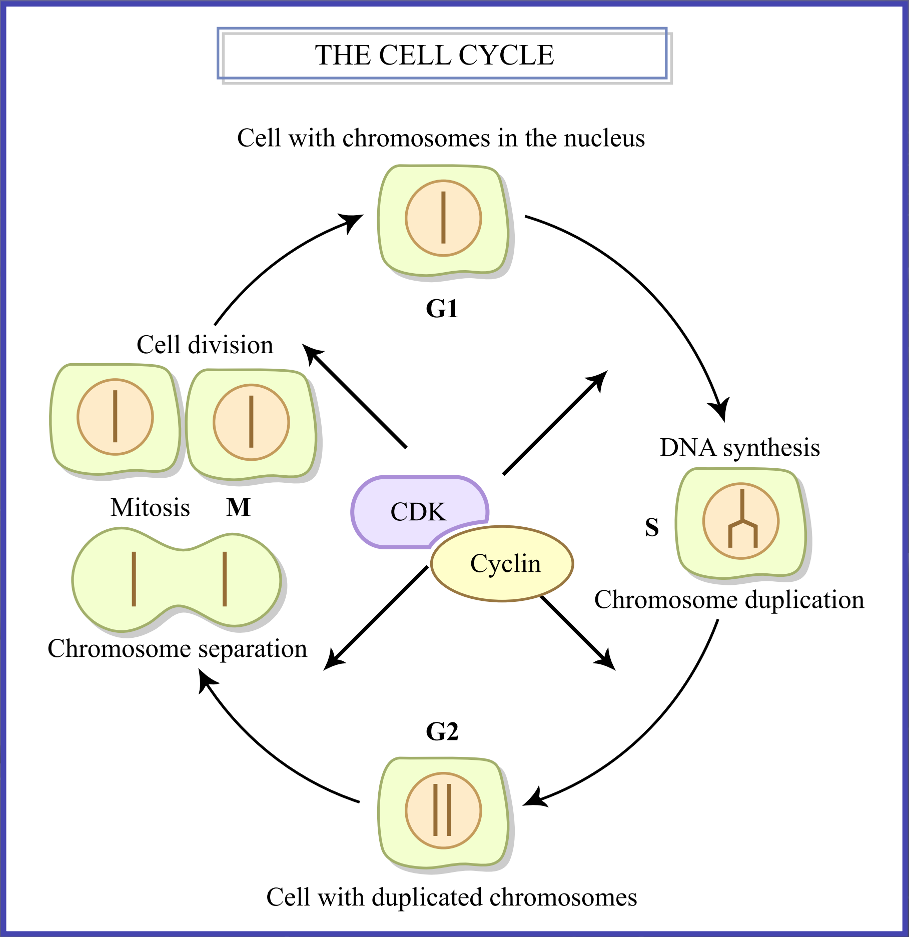


Figure by MIT OCW

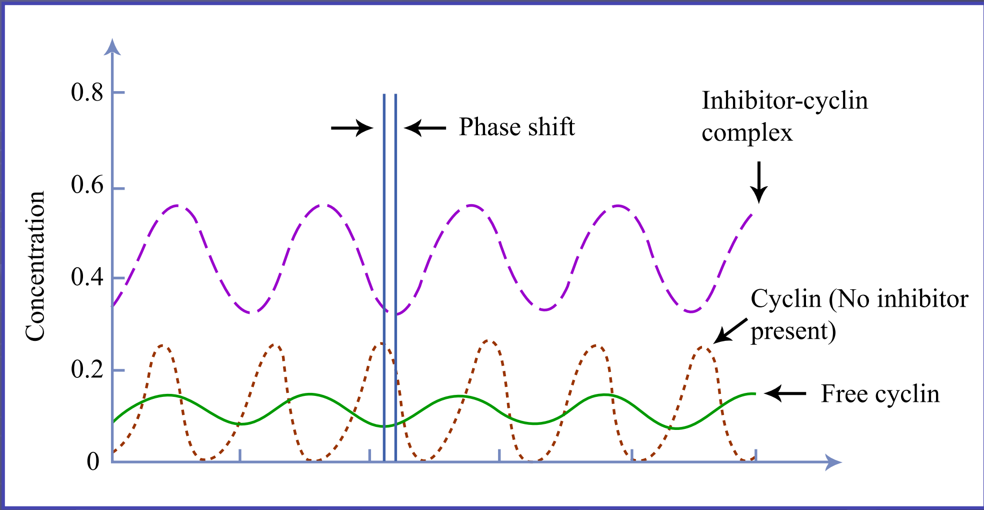


Figure by MIT OCW

Model:

$$\dot{C} = v_i - k_1 \frac{XC}{C + K_5} - k_d C,$$

$$M = \frac{V_1(1 - M)}{(1 - M) + K_1} - \frac{V_2 M}{M + K_2},$$

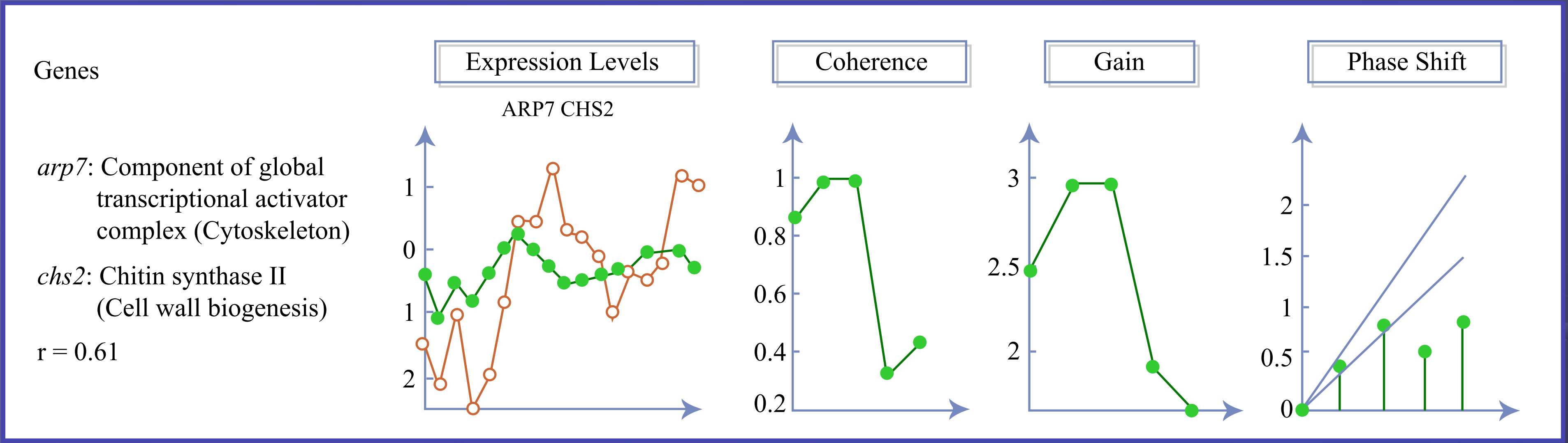
$$\dot{X} = \frac{V_3(1 - X)}{(1 - X) + K_3} - \frac{V_4 X}{X + K_4}$$

$$V_1 = \frac{C}{C + K_6} V_{1'}, V_3 = M V_{3'},$$

Gardner, T. S., Dolnik, M. & Collins, J. J. A theory for controlling cell cycle dynamics using a reversibly binding inhibitor. *Proc Natl Acad Sci* **95**, 14190-5 (1998).

Course Introduction

Signal Processing in Bioinformatics

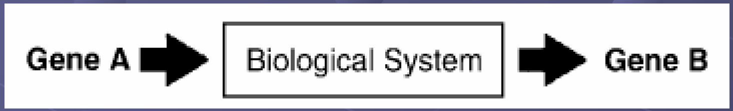


arp7: Component of global transcriptional activator complex (Cytoskeleton)

chs2: Chitin synthase II (Cell wall biogenesis)

r = 0.61

Figure by MIT OCW



$C_{xy}[f] = \frac{|P_{xy}[f]|^2}{P_{xx}[f]P_{yy}[f]}$

$T_{xy}[f] = \frac{P_{xy}[f]}{P_{xx}[f]}$

$\Phi = \frac{2\pi \cdot m}{\lambda},$

$x[t] = A\cos(\Omega_0 t + \varphi),$

power spectral density (P_{xx})

Coherence Function

Transfer Function

Phase

Butte, A. J., Bao, L., Reis, B. Y., Watkins, T. W. & Kohane, I. S. Comparing the similarity of time-series gene expression using signal processing metrics. *J Biomed Inform* **34**, 396-405 (2001).

Harvard-MIT
Division of Health
Science & Technology

Instructors

- **Gil Alterovitz**
  - HST Medical Engineering Medical Physics-Electrical Engineering and Computer Science, Graduate Student/Whitaker Fellow.
  - Proteomics & Computational Biology, Introductory Material
- **Robert Berwick**
  - Professor, Electrical Engineering and Computer Science, MIT
  - Language/Sequence Analysis
- **Rob Henson**
  - Director of Bioinformatics Group, Mathworks (Matlab).
  - Mathematics and Signal Processing, Industrial Experience
- **Manolis Kellis**
  - Assistant Professor, Electrical Engineering and Computer Science, MIT
  - Sequence Analysis
- **Nanguneri Nirmala**
  - Functional Genomics Group, Novartis Institutes for BioMedical Research
  - Expression Analysis, Industrial Experience
- **Marco F. Ramoni**
  - Assistant Professor of Pediatrics and Medicine, Harvard Medical School
  - Expression Analysis, Bayesian Networks
- **Paola Sebastiani**
  - Associate Professor, Department of Biostatistics, Boston University
  - Statistical Methodologies and Bioinformatics

Organization: Levels of Abstraction

- Part I: Sequence
- Part II: Expression
- Part III: Proteomics
- Part IV: Systems/Misc.

Part I / II

- Tue, January 4, 2005, 11:00am-11:45pm
  - Review of Modern Biology- Gil Alterovitz
- Tue, January 4, 2005, 11:45am-12:30pm
  - Introduction to Bioinformatics Laboratory / Bioinformatics in the Computer Industry- Rob Henson / Gil Alterovitz
- Thurs, January 6, 2005, 11:00am-11:45pm
  - Review of Modern Biology II- Gil Alterovitz
- Thurs, January 6, 2005, 11:45am-12:30pm
  - Sequence Analysis: Motif and Regulation- Manolis Kellis
- Tue, January 11, 2005, 11:00am-11:45pm
  - Sequence Analysis: Genes and Genome- Manolis Kellis
- Tue, January 11, 2005, 11:45am-12:30pm
  - Sequence Analysis: Gene Evolution- Manolis Kellis and Robert Berwick
- Thurs, January 13, 2005, 11:00am-11:45pm
  - Microarray Expression Data Analysis- Marco Ramoni
- Thurs, January 13, 2005, 11:45am-12:30pm
  - Machine Learning: Bayesian Methodologies- Marco Ramoni

Part IV / III

- Tue, January 18, 2005, 11:00am-12:00pm
  - Bioinformatics in the Biotech Industry- Nanguneri Nirmala
- Tue, January 18, 2005, 12:00am-12:30pm
  - Control and Feedback in Systems- Gil Alterovitz
- Thurs, January 20, 2005, 11:00am-11:45pm
  - Scale-free Networks I- Paola Sebastiani
- Thurs, January 20, 2005, 11:45am-12:30pm
  - Scale-free Networks II- Paola Sebastiani
- Tue, January 25, 2005, 11:00am-11:45pm
  - Statistical Models and Stochastic Processes in Proteomics- Gil Alterovitz
- Tue, January 25, 2005, 11:45am-12:30pm
  - Signal Processing for Proteomics – Gil Alterovitz
- Thurs, January 27, 2005, 11:00am-12:00pm
  - Biological Methods, Automation, Robotics- Gil Alterovitz
- Thurs, January 27, 2005, 12:00pm-12:30pm
  - Project Discussion and Wrap-up- Gil Alterovitz

Parts III/IV switched to accommodate speakers.

Class Information

- **Date:** Winter 2005
- **Credits/Hours:**
  - *Four weeks:* TR, 11:00 am-12:30 pm
    Total hours: 12 scheduled + estimated 28 outside = 40
    Weekly load: 3 scheduled + estimated 7 outside = 10
  - *Units:* 3 credits (1-0-2), U
  - *Audience:* upper undergraduate/graduate.
- **Student Prerequisites:**
  - 7.012 or equivalent recommended
  - 6.003 or equivalent recommended
  - 6.041 or equivalent recommended
- **Grades:**
  - P/D/F
- **Time/Location:**
  - Lecture TR, 11a-12:30p
  - Lab (optional), 9a-~10:40a (cluster of dual-processor, dual LCD panel Windows-based machines with Matlab pre-installed).

Resources

- Book chapter:
  - G. Alterovitz, E. Afkhami, and M. Ramoni, "Robotics, Automation, and Statistical Learning for Proteomics," in Focus on Robotics and Intelligent Systems Research, vol. 1, F. Columbus, Ed. New York: Nova Science Publishers, Inc., 2005 (In press).
- Reference texts:
  - A. V. Oppenheim, A. S. Willsky, and H. Nawab, *Signals and Systems*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1997.
  - A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*. New York, NY: McGraw-Hill, 2002.
  - I. S. Kohane, A. T. Kho, and A. J. Butte, *Microarrays for an Integrative Genomics*. Cambridge, MA: MIT Press, 2002.
- Other:
  - 17 other papers/resources (*Nature, PNAS, Machine Learning, Bioinformatics, Physical Review E*, etc.)

Academic Information

Labs/Homework

- **3 Labs (homeworks)**
- **Final Project-** Student selected based on one of the four areas.

Grading

- Labs
  - 40%
- Final Project
  - 50%
- Participation
  - 10%

Miscellaneous

- Fill out background sheet and turn them in at the front.

Modern Biology in Two Lectures (Part I Today)

Genes to Proteins

Transcription

Translation

DNA: “Lifetime Plan”

```
5′ATCTACAGATCAGCTACGACGCGACGAT
TTAGCAGCAGCGACGCGACAGCAGCTAGTG
ACGATAGCACATAGTTAGCACAGAGCAGAC
ACAGACAGCACAGCGACAGCGACGACG-3′
```

mRNA: “Task List”

```
5′AUCUACAGAUCAGCUACGACGCGACGAU
UUAGCAGCAGCGACGCGACAGCAGCUAGUG
ACGAUAGCACAUAGUUAGCACAGAGCAGAC
ACAGACAGCACAGCGACAGCGACGACG-3′
```

Protein: Machines

```
MWTRFDSALPRSTPSTAKLVMPOILLLLEE
EDTYESAQYKTWLMVCSDETTTE
```

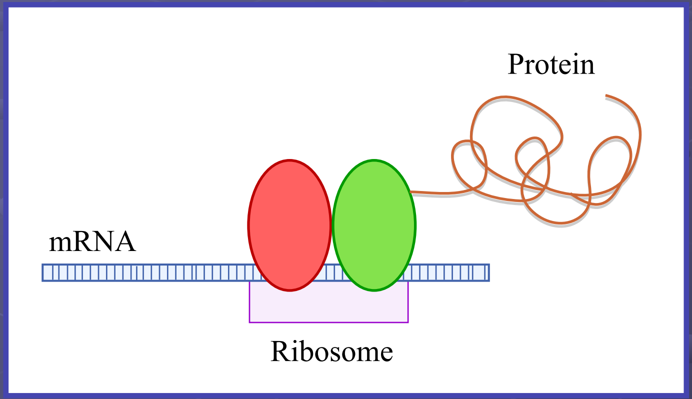


Figure by MIT OCW

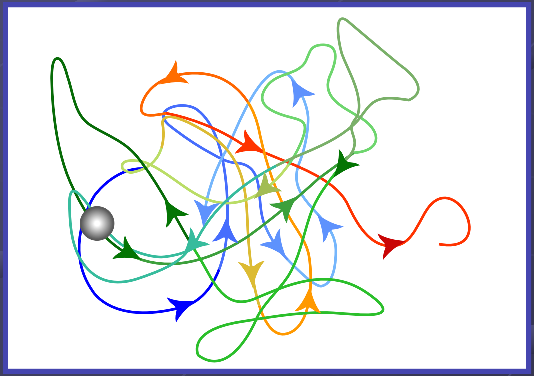


Figure by MIT OCW

DNA Sequencing

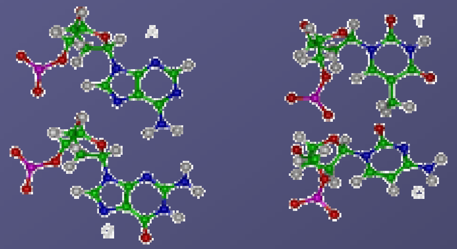
Relative Expression Levels

Identification
Post translation modification
Splicing variants
Relative expression levels

Source: HPCGG

HST Harvard-MIT Division of Health Science & Technology

Transcription

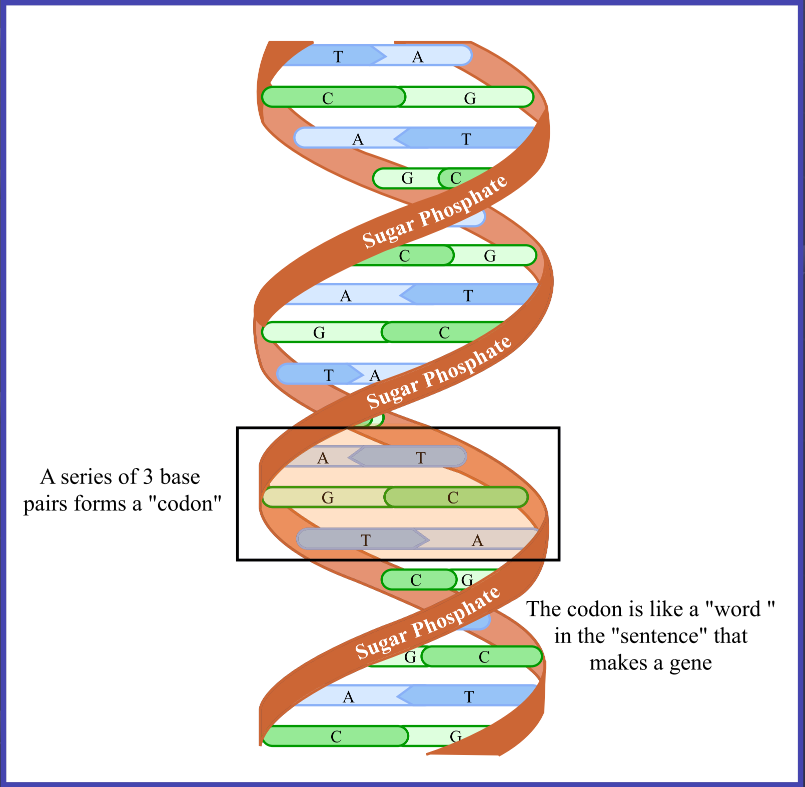


Atom	Color
Oxygen	Red
Nitrogen	Blue
Carbon	Green
Phosphorus	Magenta
Sulfur	Yellow
Hydrogen	Grey

DNA→RNA

- G=Guanine
- C=Cytosine
- A=Adenine,
- T=Thymine (DNA only)
- U=Uracil (RNA only).
- (DNA) T → U (RNA)

- DNA, RNA= sequence of nucleotide bases
- "Parity Bit" Analogy
  - Redundant information in second strand for error correction.



A series of 3 base pairs forms a "codon"

The codon is like a "word " in the "sentence" that makes a gene

Figure by MIT OCW

DNA = deoxyribonucleic acid
RNA = ribonucleic acid

Translation

SECOND BASE

FIRST BASE	U	C	A	G	THIRD BASE
U	UUU UUC Phenylalanine UUA UUG Leucine	UCU UCC UCA UCG Serine ●	UAU UAC Tyrosine UAA Stop codon UAG Stop codon	UGU UGC Cysteine UGA Stop codon UGG Tryptophan	U C A G
C	CUU CUC CUA CUG Leucine ●	CCU CCC CCA CCG Proline ●	CAU CAC Histidine CAA CAG Glutamine	CGU CGC CGA CGG Arginine ●	U C A G
A	AUU AUC AUA Isoleucine AUG Methionine start codon	ACU ACC ACA ACG Threonine ●	AAU AAC Asparagine AAA AAG Lysine	AGU AGC Serine AGA AGG Arginine	U C A G
G	GUU GUC GUA GUG Valine ●	GCU GCC GCA GCG Alanine ●	GAU GAC Aspartic acid GAA GAG Glutamic acid	GGU GGC GGA GGG Glycine ●	U C A G

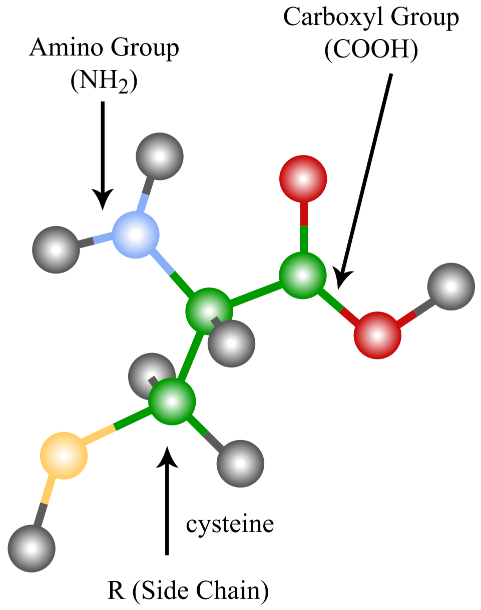


Figure by MIT OCW

RNA→Protein

Protein = Sequence of Amino Acids

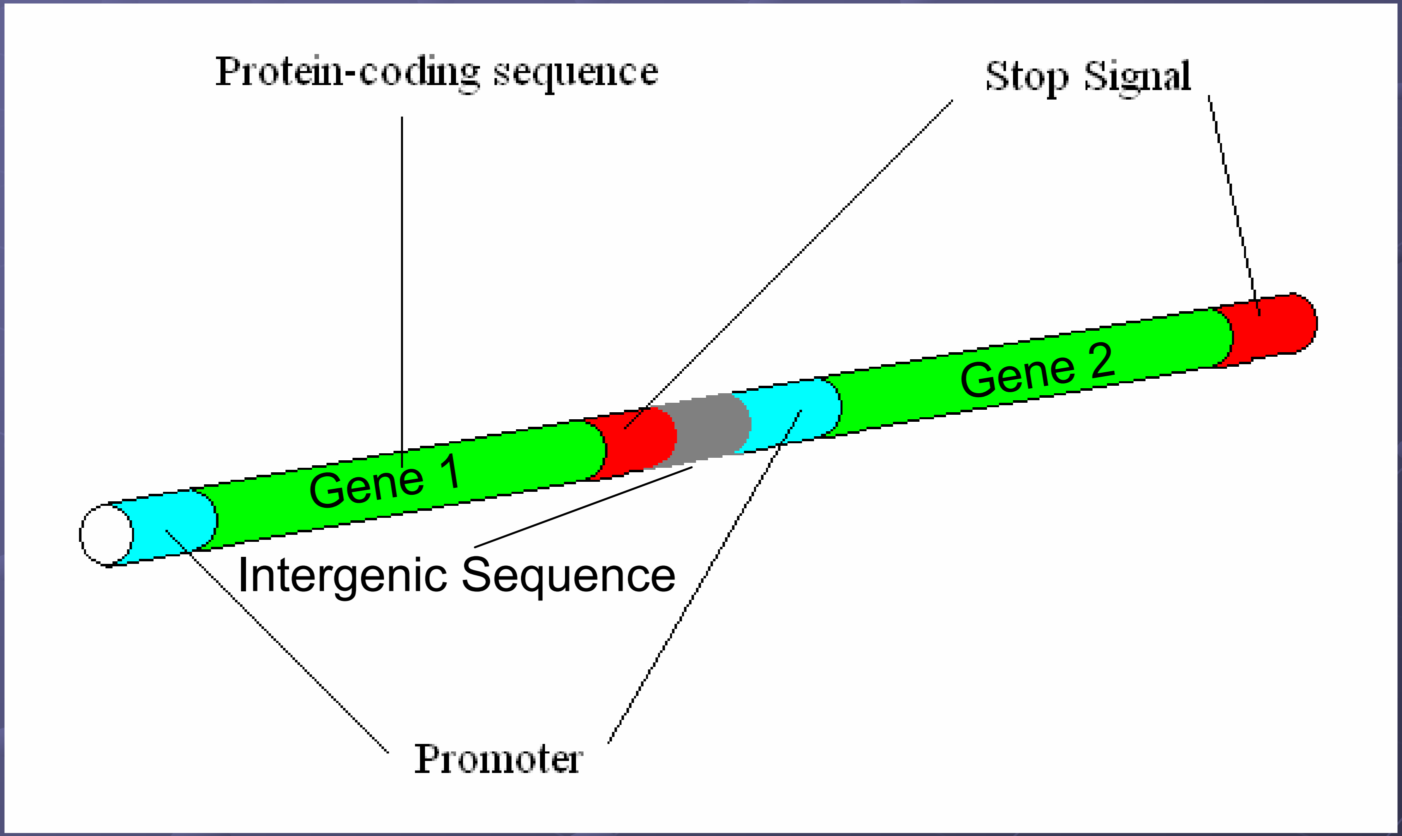
Harvard-MIT
Division of Health
Science & Technology

Name	Symbol	Mass (-H_2O)	Side Chain	Occurrence (%)
Alanine	A, Ala	71.079	CH_3-	7.49
Arginine	R, Arg	156.188	$HN{=}C(NH_2)$-NH-$(CH_2)_3$-	5.22
Asparagine	N, Asn	114.104	H_2N-CO-CH_2-	4.53
Aspartic acid	D, Asp	115.089	HOOC-CH_2-	5.22
Cysteine	C, Cys	103.145	HS-CH_2-	1.82
Glutamine	Q, Gln	128.131	H_2N-CO-$(CH_2)_2$-	4.11
Glutamic acid	E, Glu	129.116	HOOC-$(CH_2)_2$-	6.26
Glycine	G, Gly	57.052	H-	7.10
Histidine	H, His	137.141	N=CH-NH-CH=C-CH_2- |_________|	2.23
Isoleucine	I, Ile	113.160	CH_3-CH_2-$CH(CH_3)$-	5.45
Leucine	L, Leu	113.160	$(CH_3)_2$-CH-CH_2-	9.06
Lysine	K, Lys	128.17	H_2N-$(CH_2)_4$-	5.82
Methionine	M, Met	131.199	CH_3-S-$(CH_2)_2$-	2.27
…				

20 amino acids in total.

Letters- compared to DNA/RNA

Genes



Communication analogy: start, message, stop.

Source: Ehsan Afkhami

Harvard-MIT
Division of Health
Science & Technology

[Slide not shown]
Rob Henson

- Rob Henson comes to us from Mathworks- creators of Matlab software. Rob studied Mathematics at Cambridge University. He spent 7 years in Japan working in the software industry before coming to the US. At Mathworks, he leads the bioinformatics group- which released the newest version of their bioinformatics toolbox a couple months ago. It is my great pleasure to introduce Rob- who will be talking about clustering technologies in bioinformatics and his perspective from industry. Thank you for coming today.

[Rob Henson's lecture (will be posted when available)]

Outline

- Bioinformatics from Industry's Perspective: Mathworks- Rob Henson
    - Bioinformatics in Industry
    - Matlab Bioinformatics Toolbox
    - Clustering and Related Technologies (DeRisi's Microarray Paper)

Modern Biology in Two Lectures (Part II Thurs)

Splicing, Alternative Splicing, Post-Translational Modifications, and Bioinformatics Tools and Databases