

# Bioinformatics

## Introduction

Sami Khuri  
 Department of Computer Science  
 San José State University  
 San José, California, USA  
 Sami.Khuri@sjsu.edu  
 www.cs.sjsu.edu/faculty/khuri

©2013 Sami Khuri

## Outline

- Ten Questions
- Top 25 Questions for the next 25 years [Science]
- Genetics in Medicine – Sixth and Seventh Editions
- **Workshop: Introduction to Bioinformatics**
- Biology Review & Introduction to Bioinformatics
- Pairwise and Multiple Sequence Alignments
- Phylogenetic Tree Construction
- RNA Secondary Structure
- Gene Prediction
- Human Genome Variation

©2013 Sami Khuri

Sequence alignment of HBA\_HUMAN, HBA\_HORSE, HBA\_CHICK, HBB\_HUMAN, HBB\_BOSMU, HBB\_HORSE, HBB\_MACGI, MYG\_PHYCA, GLB5\_PETMA, and LGB2\_LUPLU. Includes a **Question One** box.

©2013 Sami Khuri

## Quagga: Zebra or Horse?



Died in Amsterdam zoo in 1883.

**Question Two**

©2013 Sami Khuri

Sequence alignment of MYH16 (Myosin Heavy Gene) across various species including Non-human, Woreley monkey, Ptilin macaque, Rhesus, Orang-utan, Gorilla, Bonobo, Chimpanzee, Human, Africa (pygmy), Spain (Berque), Iceland, Japan, Russia, and South America. Includes a **Question Three** box.

©2013 Sami Khuri

MYH16: Myosin Heavy Gene  
 Accession Number: BK001410 at NCBI (exon 19)

**Question Three**

Table 1 Types and Frequencies of  $\beta$ -Thalassemia Mutations and  $\beta$ -Globin variants in Lebanon

Mutation	Phenotype	Number of Chromosomes	Number of homozygotes/heterozygotes	Frequency (%)
IVS-I-110 (G>A)	$\beta^+$	178	65/48	34.2
IVS-I-1 (G>A)	$\beta^0$	78	29/20	15.0
IVS-I-6 (T>C)	$\beta^+$	75	28/19	14.4
cd 29 (C>T)	$\beta^+$	50	22/6	9.6
IVS-II-1 (G>A)	$\beta^0$	45	17/11	8.6
cd 5 (-C>T)	$\beta^0$	26	9/8	5.0
cd 30 (G>C)	$\beta^0$	14	6/2	2.7
cd 8 (-AA)	$\beta^0$	13	5/3	2.5
cd 44 (-C)	$\beta^0$	8	3/2	1.5
IVS-II-745 (C>G)	$\beta^+$	6	3/0	1.1
$\beta^0$	$\beta^+$	5	0/5	1.0
-87 (C>G)	$\beta^+$	4	1/2	0.8
IVS-I-5 (G>C)	$\beta^+$	4	2/0	0.8
-88 (C>T)	$\beta^+$	3	1/1	0.6
290 bp deletion	$\beta^0$	3	1/1	0.6
25 bp deletion	$\beta^0$	2	1/0	0.4
$\beta^0$ -thalassemia (Sicilian type)	$\beta^0$	2	1/0	0.4
cd 8/9 (+G)	$\beta^0$	1	0/1	0.2
cd 36/37 (-T)	$\beta^0$	1	0/1	0.2
cd 39 (C>T)	$\beta^0$	1	0/1	0.2
Unknown		1	0/1	0.2
Total		520	194/132	100

"Genetic heterogeneity of beta thalassemia in Lebanon reflects historic and recent population migration" by N. J. Makhoul, et al.  
 "Annals of Human Genetics" in 2005 (issue 69, pages 55 to 66).

**Question Four**

©2013 Sami Khuri

TABLE 2. FREQUENCY DISTRIBUTION OF THE MOST COMMON  $\beta$ -THALASSEMIA MUTATIONS IN MOROCCO AND IN ARAB AND MEDITERRANEAN COUNTRIES

Mutations	Morocco (160) <sup>a</sup>	Algeria (239)	Tunisia (233)	Portugal (561)	Spain (324)	Italy (325)	Egypt (337)	Lebanon (520)	Turkey (795)
Codon 39 (C → T)	26.58	27.6	40	36.8	36	40	1.5	0.5	3.8
F5C-8 (-AA)	13.91	-	0.9	-	0.4	0.1	1.8	2.5	5.4
IVS-II-745 (C → G)	7.6	0.9	2.5	-	-	5	5.6	1.2	5
-29 (A → G)	6.33	3.8	-	-	-	-	-	-	-
F5C-6 (-A)	5.7	17	6.65	1	1	1.9	0.9	-	0.4
IVS-I-110 (G → A)	5.7	24.7	20.5	10	13	19.9	32.9	34.2	39.2
IVS-I-2 (T → C)	5.06	3.3	0.76	-	-	-	-	-	-
IVS-I-1 (G → A)	5.06	11.7	1	28	35	10.2	11.3	15	5
Total	76	89	72.3	75.8	85.4	77	54	53.4	58.8
References <sup>b</sup>	1	2,3	4	5,6	7	8	9	10	

<sup>a</sup>Values in parenthesis indicate the total number of chromosomes studied.  
<sup>b</sup>1, Bennani *et al.* (1994); 2, Fattoum *et al.* (1991); 3, Haj Khellil *et al.* (2004); 4, Faustino *et al.* (1999); 5, Anselem *et al.* (1988); 6, Ribeiro *et al.* (1997); 7, Rostelli *et al.* (1992b); 8, Waye *et al.* (1999); 9, Makhoul *et al.* (2005); 10, Tadmouri *et al.* (1998).

"Molecular basis of beta thalassemia in Morocco: possible origins of the molecular heterogeneity" by I. Agouti *et al.* In "Genetic Testing" Volume 12, Number 4, 2008.

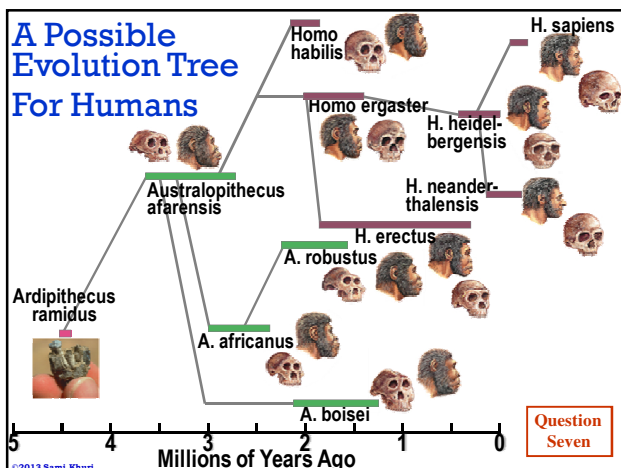
**Question Five**

### Aligning BRCA1 Sequences

```

Wombat : KVNELTRSDHLASNSNGRSHQSALEPSALEDGHDTAEGNSSVSSEKTD : 52
Opossum : KVNELTRSDHLAPDTSVRSHQAEATNALEVGHYET-DGNSSISEKTD : 51
Armadillo : KVNELTRSDHLITSDSHRGSEDLAEVAGALKV---SRVDEYSSFSEKTD : 50
Sloth : KVNELTRSDHLITSDSHRGSEDLAEVAGALKV---SRVDEYSSFSEKTD : 50
Dugong : KVNELTRSDGL---DDLHQRGSEDLAEVAGALE---SRVDEYSSFSEKTD : 47
Hyrax : KVNELTRSDDL---SDSPEGSEDLNGKQAGPVK---SRVDEYSSFSEKTD : 47
Aardvark : KVNELTRSDGL---DGSHQRGSEDLAEVAGALE---SRVDEYSSFSEKTD : 47
Tenrec : KVNELTRSDHGL---GDSRQGRPSGADLAVAFEV---SRVDEYSSFSEKTD : 47
Rhinoceros : KVNELTRSDDLITSDSHRGSEDLAEVAGALE---SRVDEYSSFSEKTD : 50
Pig : KVNELTRSDDLITSDSHRGSEDLAEVAGALE---SRVDEYSSFSEKTD : 50
Hedgehog : KVNELTRSDDLITSDSHRGSEDLAEVAGALE---SRVDEYSSFSEKTD : 50
Human : KVNELTRSDDLITSDSHRGSEDLAEVAGALE---SRVDEYSSFSEKTD : 50
Rat : KVNELTRSDDLITSDSHRGSEDLAEVAGALE---SRVDEYSSFSEKTD : 50
Hare : KVNELTRSDDLITSDSHRGSEDLAEVAGALE---SRVDEYSSFSEKTD : 50
    
```

**Question Six**



### The Mighty Mouse

**Mouse versus Human**

Approximately the same number of chromosomes and local gene order in mammals.

Insights into mouse genetics are likely to illuminate human genetics as well.

**Question Eight**

### Pseudogenes

"The Real Life of Pseudogenes" by Mark Gerstein and Deyou Zheng Scientific American, August 2006.

CHROMOSOMES of humans and mice carry a very similar array of functional genes (orange) but reveal distinct differences in their pseudogenes (blue), which can highlight important turning points in an organism's evolutionary history. For example, the counterpart of a mouse gene called *Gulo* has become a pseudogene (*Vgulo*) in humans and other primates. *Gulo* makes an enzyme that is the last element in a biochemical pathway for synthesizing vitamin C. Most mammals possess the active gene, but the primate lineage seems to have lost it more than 40 million years ago. When the *Gulo* gene became a pseudogene, primates became dependent on food sources of vitamin C to avoid **scurvy**.

**Question Nine**

### Azidothymidine (AZT) Blocks Reverse Transcriptase in HIV

Potential Drugs

**Question Ten**

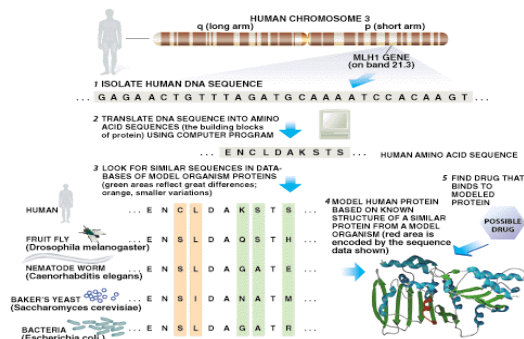
## Why Study Bioinformatics (I)

- Bioinformatics is intrinsically interesting.
- Bioinformatics offers the prospect of finding better drug targets earlier in the drug development process.
  - By looking for genes in model organisms that are similar to a given human gene, researchers can learn about protein the human gene encodes and search for drugs to block it.

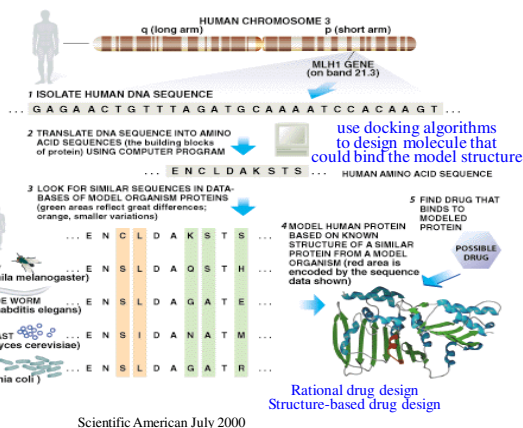


©2013 Sami Khuri

## How can Bioinformatics Help?



©2013 Sami Khuri



Scientific American July 2000

## Why Study Bioinformatics (II)

- Molecular biology is the new frontier of 21<sup>st</sup> century science.
  - DNA, RNA, genes, stem cells, etc. are everywhere in the news.
- Science Magazine celebrated its 125<sup>th</sup> anniversary by issuing twenty five big questions facing science over the next quarter-century.



[www.sciencemag.org/sciext/125th](http://www.sciencemag.org/sciext/125th)

©2013 Sami Khuri

## Science: Top 25 Questions (I)

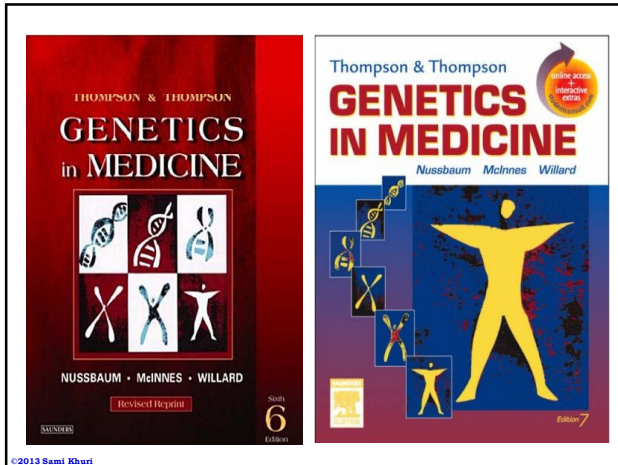
- \* What Is the Universe Made Of?
- \* What is the Biological Basis of Consciousness?
- Why Do Humans Have So Few Genes?**
- To What Extent Are Genetic Variation and Personal Health Linked?**
- \* Can the Laws of Physics Be Unified?
- \* How Much Can Human Life Span Be Extended?
- What Controls Organ Regeneration?**
- How Can a Skin Cell Become a Nerve Cell?**
- How Does a Single Somatic Cell Become a Whole Plant?**
- \* How Does Earth's Interior Work?
- \* Are We Alone in the Universe?
- \* How and Where Did Life on Earth Arise?

©2013 Sami Khuri

## Science: Top 25 Questions (II)

- What Determines Species Diversity?**
- What Genetic Changes Made Us Uniquely Human?**
- \* How Are Memories Stored and Retrieved?
- How Did Cooperative Behavior Evolve?**
- How Will Big Pictures Emerge from a Sea of Biological Data?**
- \* How Far Can We Push Chemical Self-Assembly?
- \* What Are the Limits of Conventional Computing?
- Can We Selectively Shut Off Immune Responses?**
- \* Do Deeper Principles Underlie Quantum Uncertainty and Nonlocality?
- Is an Effective HIV Vaccine Feasible?**
- \* How Hot Will the Greenhouse World Be?
- \* What Can Replace Cheap Oil -- and When?

©2013 Sami Khuri



## Preface of the Seventh Edition

Much has changed, however, since the last edition of this book. Completion of the HGP provides us with a catalogue of all human genes, their sequence, and an extensive, and still growing, database of human variation. Genomic information has stimulated the creation of powerful new tools that are changing human genetics research and medical genetics practice. We therefore have expanded the scope of the book to incorporate the concepts of “**Personalized Medicine**” into *Genetics in Medicine* by providing more examples of how genomics is being used to identify the contributions made by genetic variation to disease susceptibility and treatment outcomes.