

Welcome to Practical Bioinformatics !!!



2020/09/16

Li Wang (wangli03@caas.cn)

Who are we?

Dr. Yuwen Liu

Dr. Li Wang

Dr. Wenlong Ma

Dr. Cheng Li

Dr. Yi Zou

Weigang Zheng

Chao Wang

Yang Fu

Xuezhu Liao

liuyuwen@caas.cn

wangli03@caas.cn

mawenlong_nwsuaf@163.com

licheng@caas.cn

zouyi@caas.cn

zhengweigang@caas.cn

netwc@qq.com

yfu1116@163.com

liaoxuezhu@caas.cn

Who are we?

Name

Hometown

Major

What I am good at?

What I will teach you?

How you could support me?

Share one tip how you learn bioinformatics

Who are you?

Name

Hometown

Where did you graduate?

What is your major ? Who is your supervisor?

What are you good at?

What brings you here?

What do you want to achieve through the course?

**Study group: find your partners
and select your group leader**

How will we play the game?

PBS (Project based study)

Test: Oral presentation
(how you apply the skills to answer
biological questions?)

Presence: >= 13/16

website: <https://github.com/liu-genomics/>

CAAS-Bioinfomatics

please join our wechat group



中国农科院实用生物信息学



该二维码7天内(9月18日前)有效，重新进入将更新

How will we play the game?

Wednesday evenings D104

6:30 — 7:15

7:30 — 8:15

8:25 — 9:10

Server

Account: username@192.168.20.70

Passwords: 123456

Possibility:

Publish a paper

Publish a book

Structure of the course

- ★ Basic bioinformatics
- ★ R
- ★ Python
- ★ Practical bioinformatic tools

Our goals

By the end of this course, you should:

Navigate through your computer, create and modify files and directories, and process data using basic Unix commands

Become familiar with high performance computing resources

Become familiar with basic R syntax and data structures and implement these in data analysis and plotting.

Utilize the Python scripting language for more sophisticated data processing.

Become familiar with several practical bioinformatic tools and learn how to write scripts and analysis pipelines for working with these data.

How will we play the game?

Grading

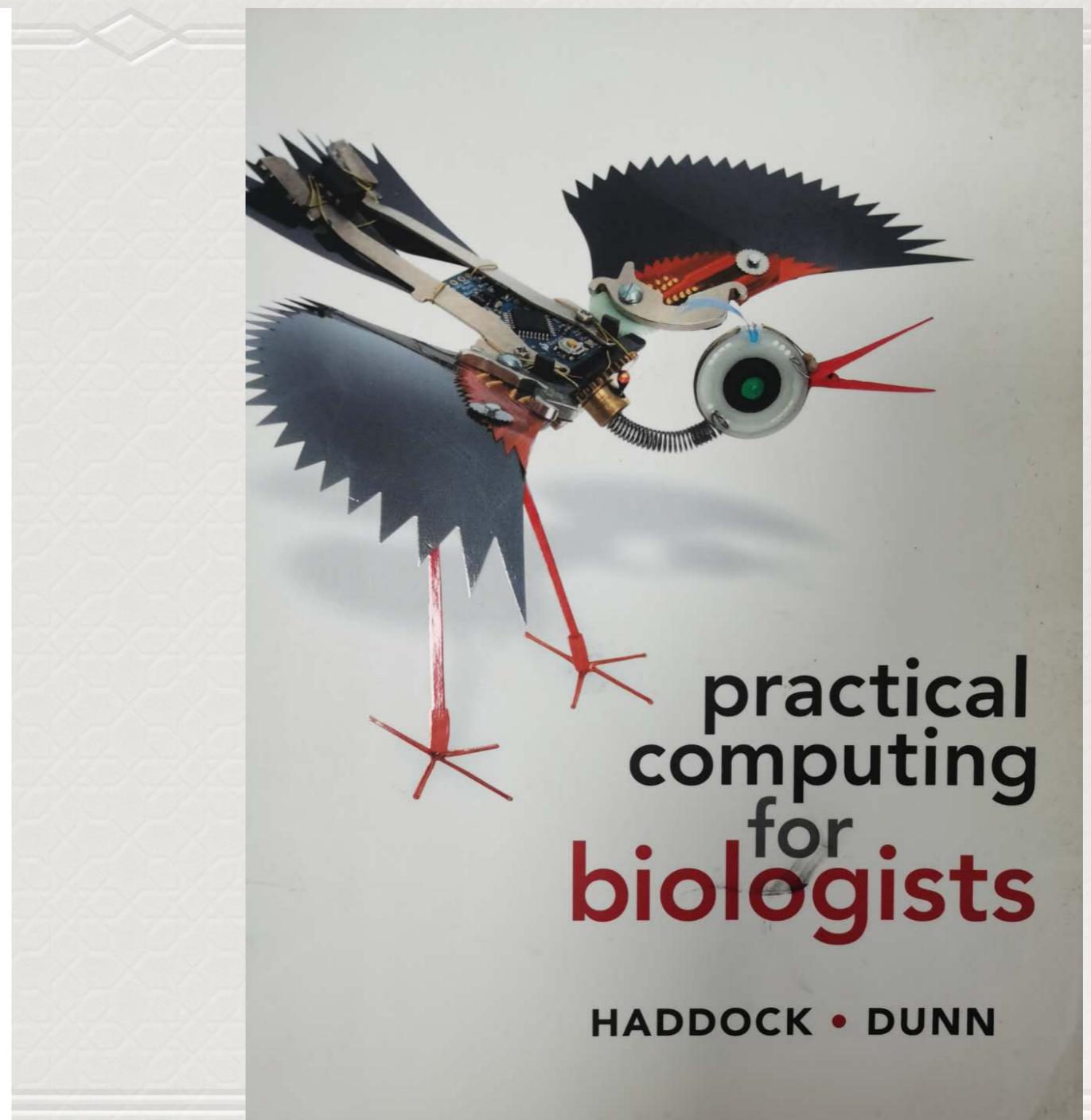
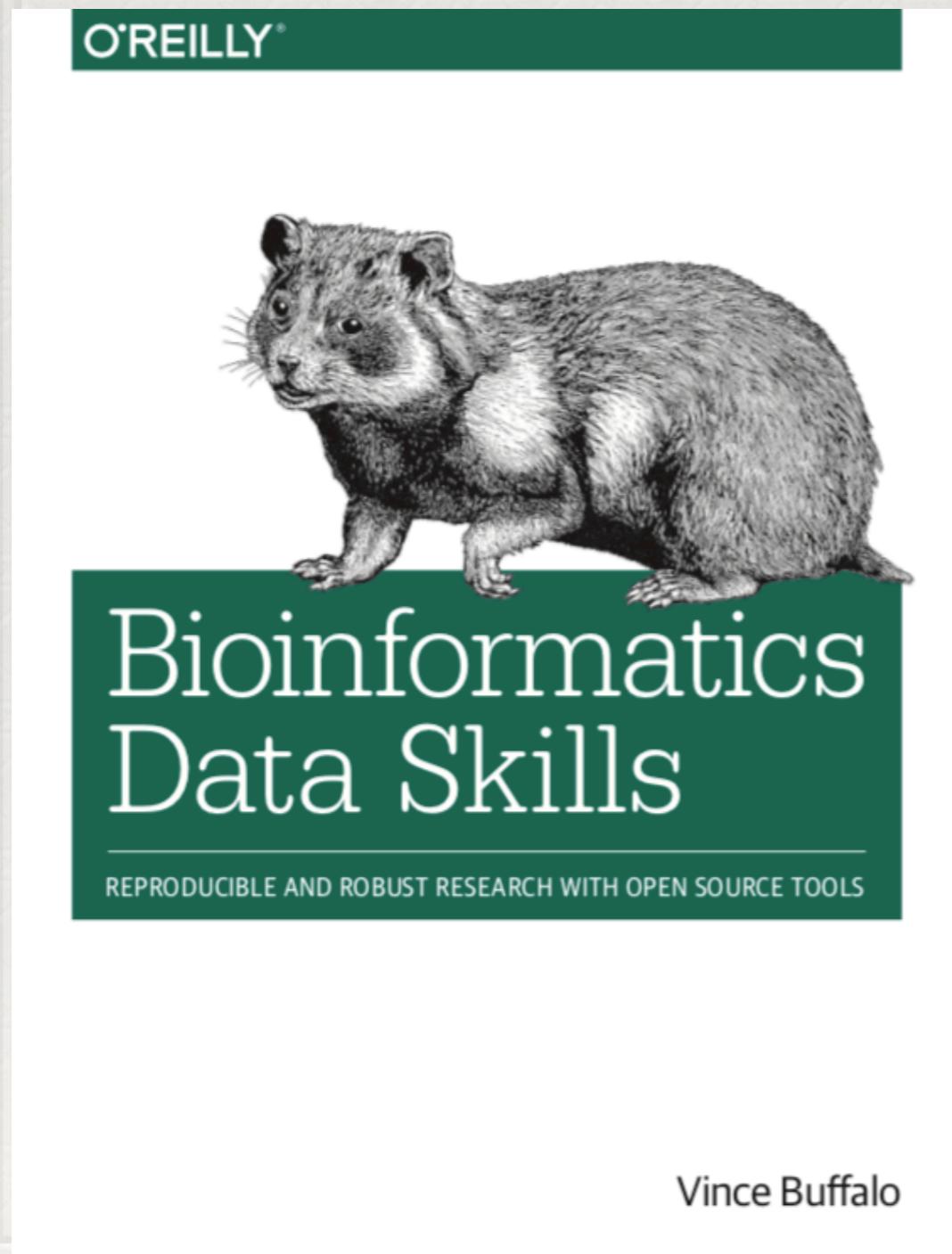
Assignment 1: Unix 10%

Assignment 2: R 10%

Assignment 3: Python 10%

Group Project and Presentation 70%

Additional materials to read



Assignment

Read Chapters 1,2,4,5 of Buffalo

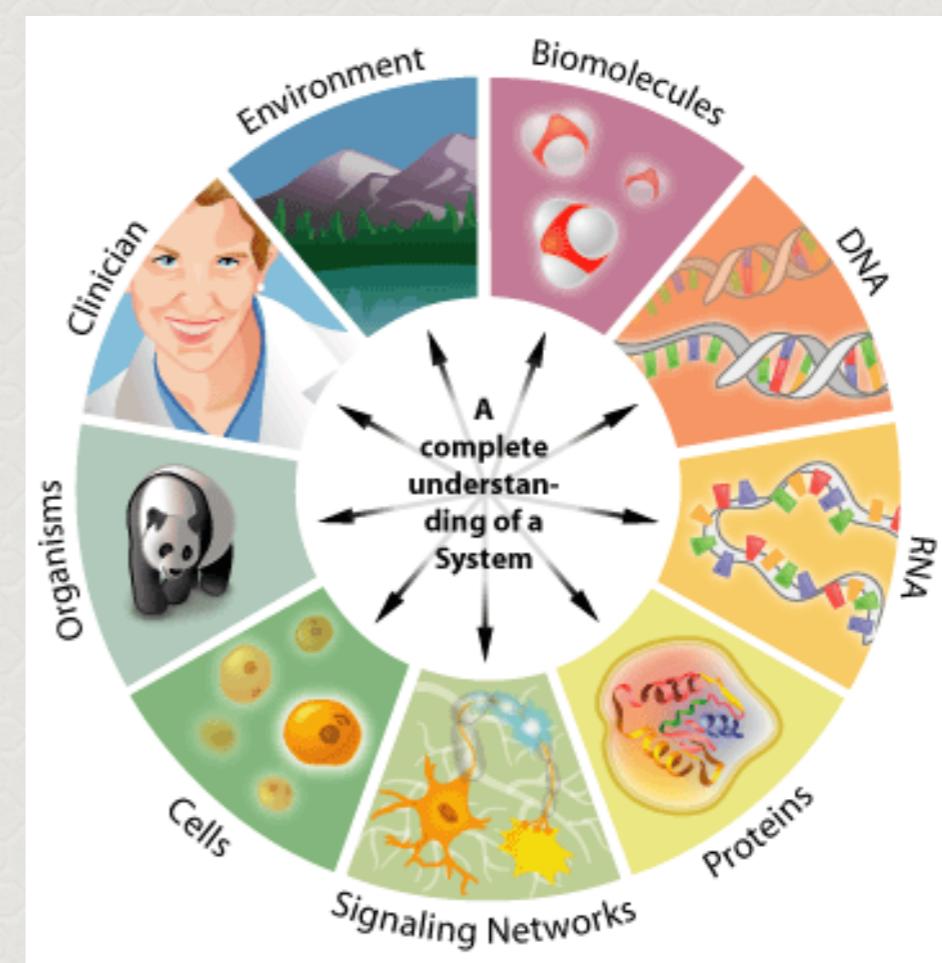
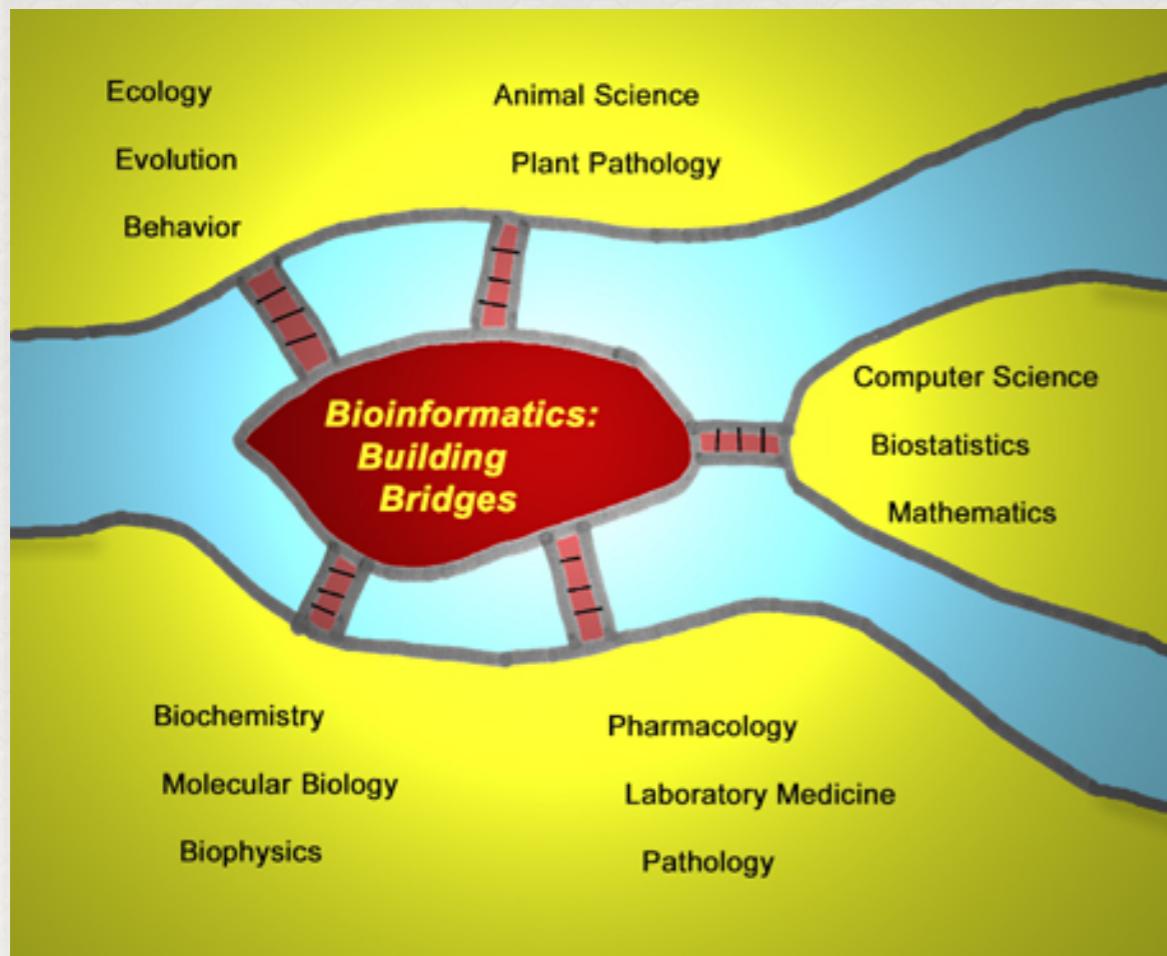
Introduction to Bioinformatics



2020/09/16

Li Wang (wangli03@caas.cn)

What is bioinformatics?



What is bioinformatics?

Biologists

collect molecular data:
DNA & Protein sequences,
gene expression, etc.

Marriage

Bioinformaticians

Study biological questions by
analyzing molecular data

Computer scientists

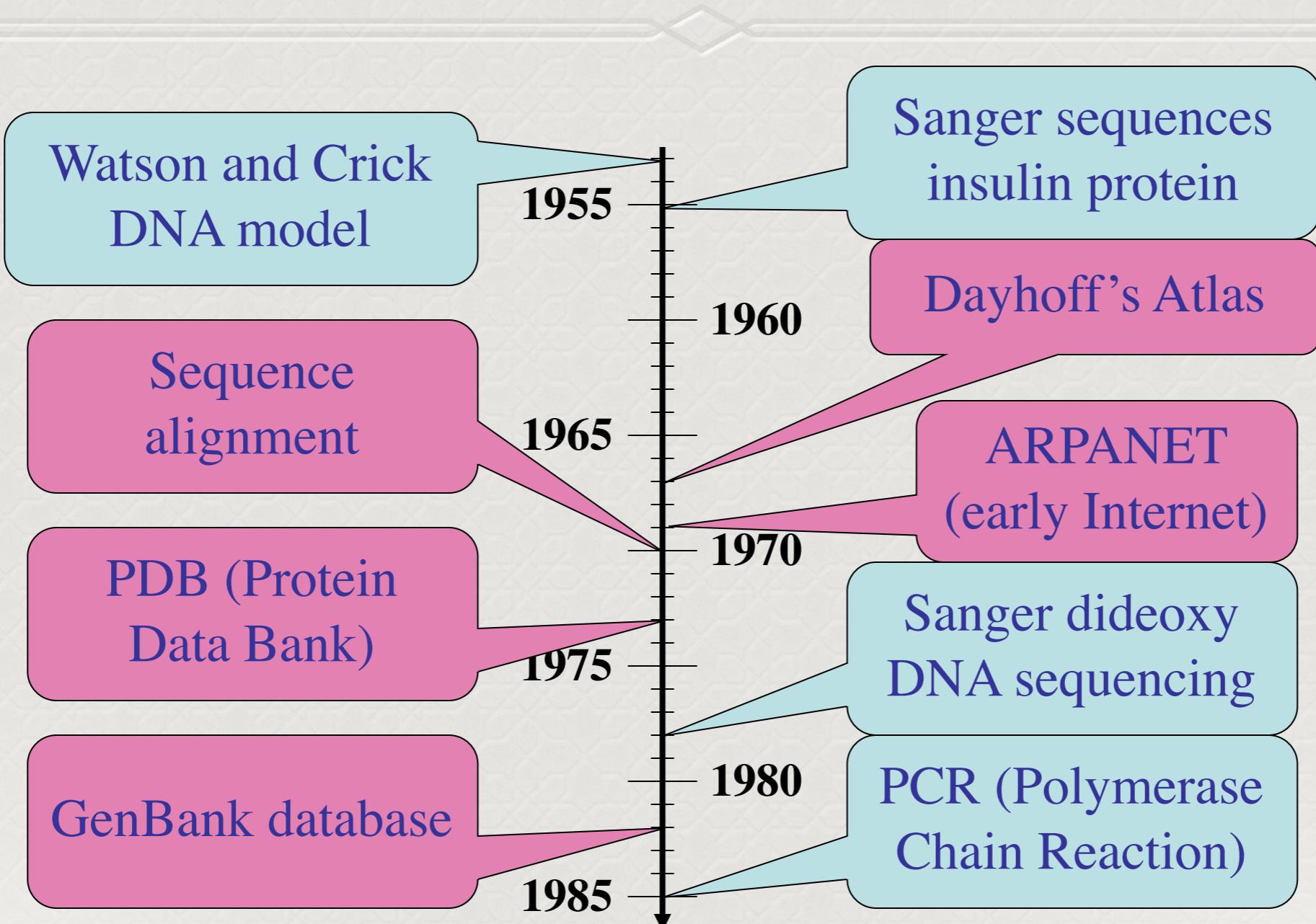
(+Mathematicians,
Statisticians, etc.)

Develop tools, softwares,

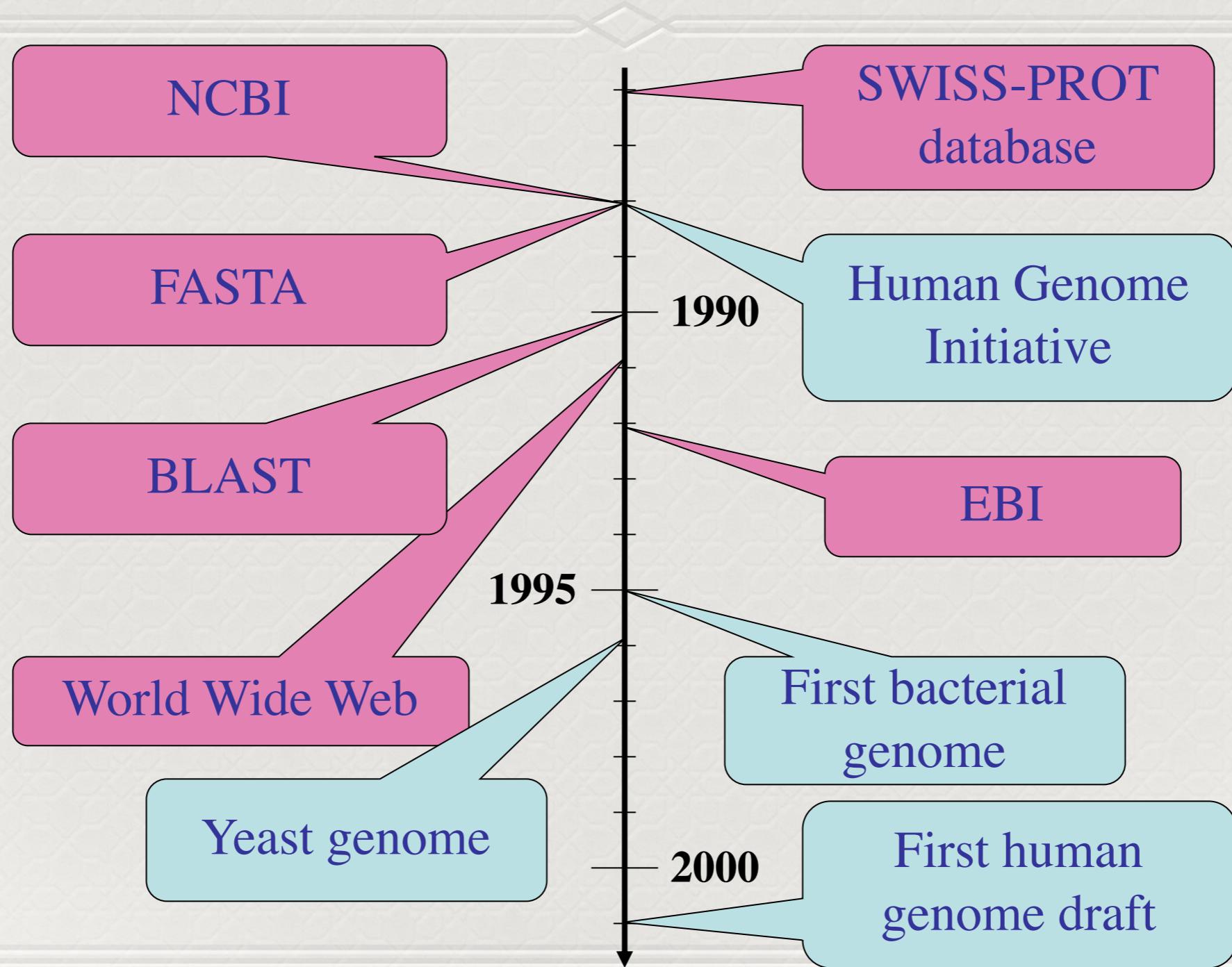
What is bioinformatics?

- It is a recently developed science using information to understand biological phenomenon. First introduced in 1990s.
- It involves the computational tools and methods used to manage, analyze and manipulate volumes and volumes of biological data.
- It is an interdisciplinary approach requiring advanced knowledge of computer mathematics and statistical methods for understanding of biological phenomena at the molecular level.

From DNA to Genome



From DNA to Genome

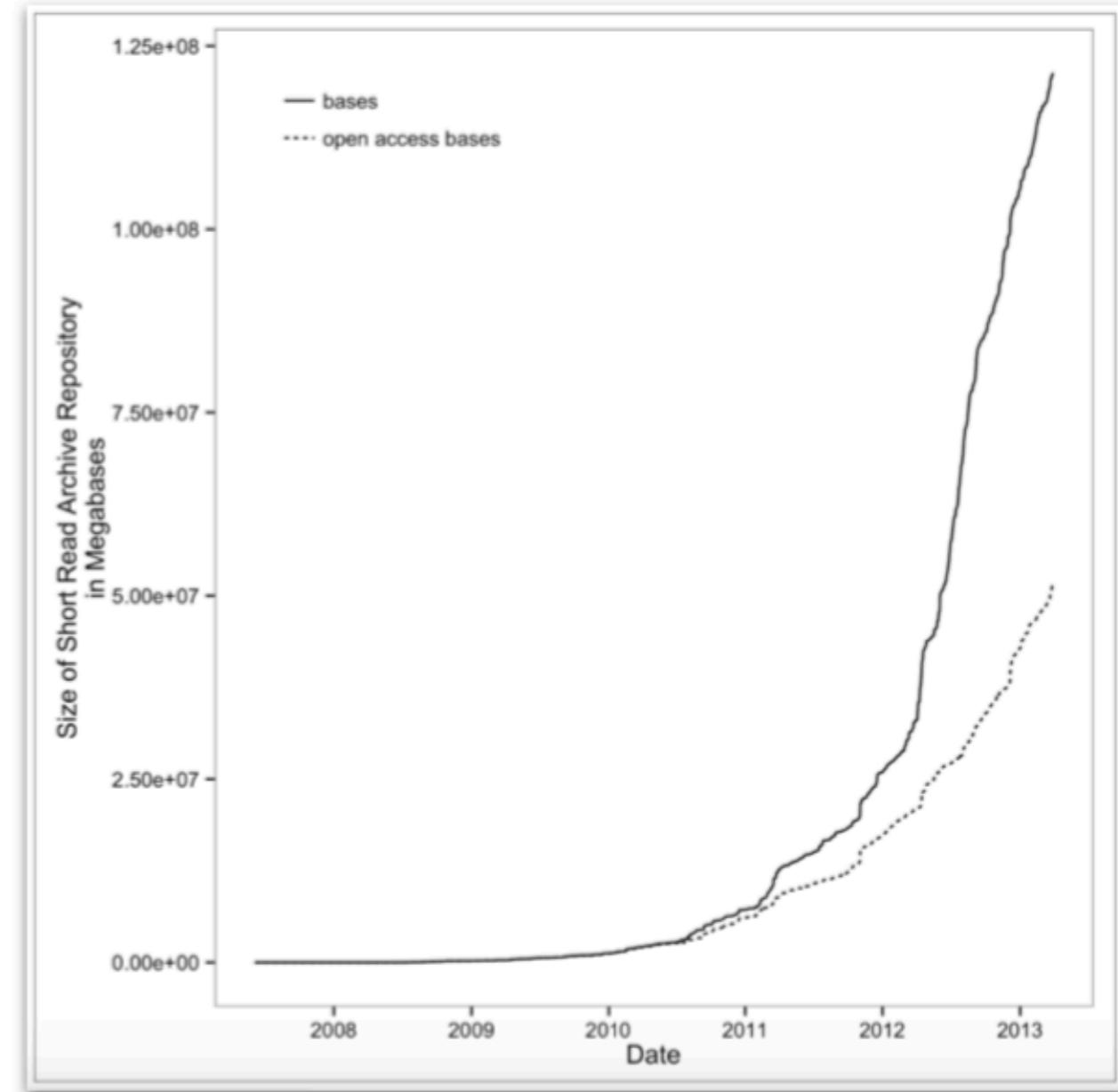
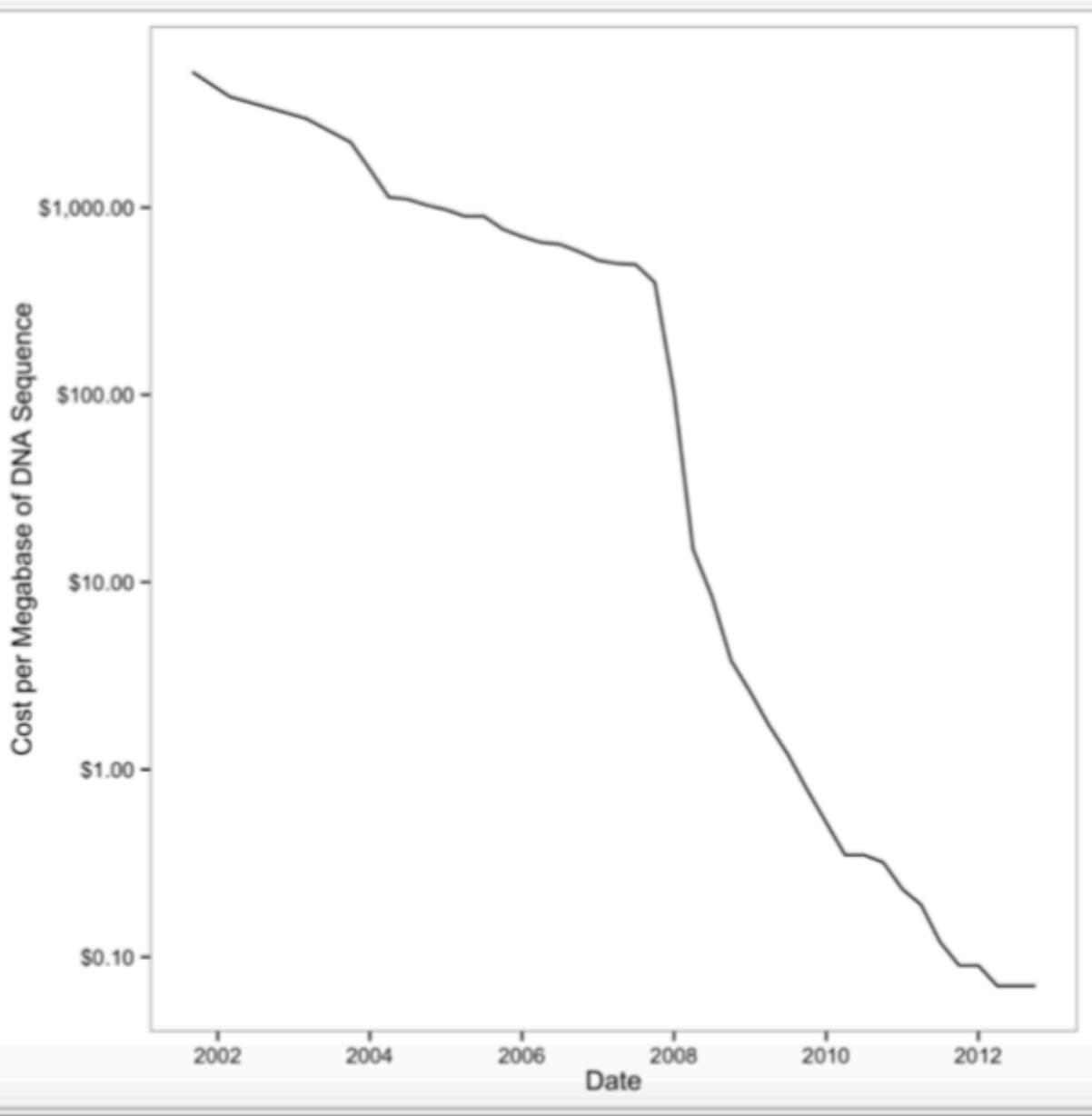


Why is bioinformatics necessary?

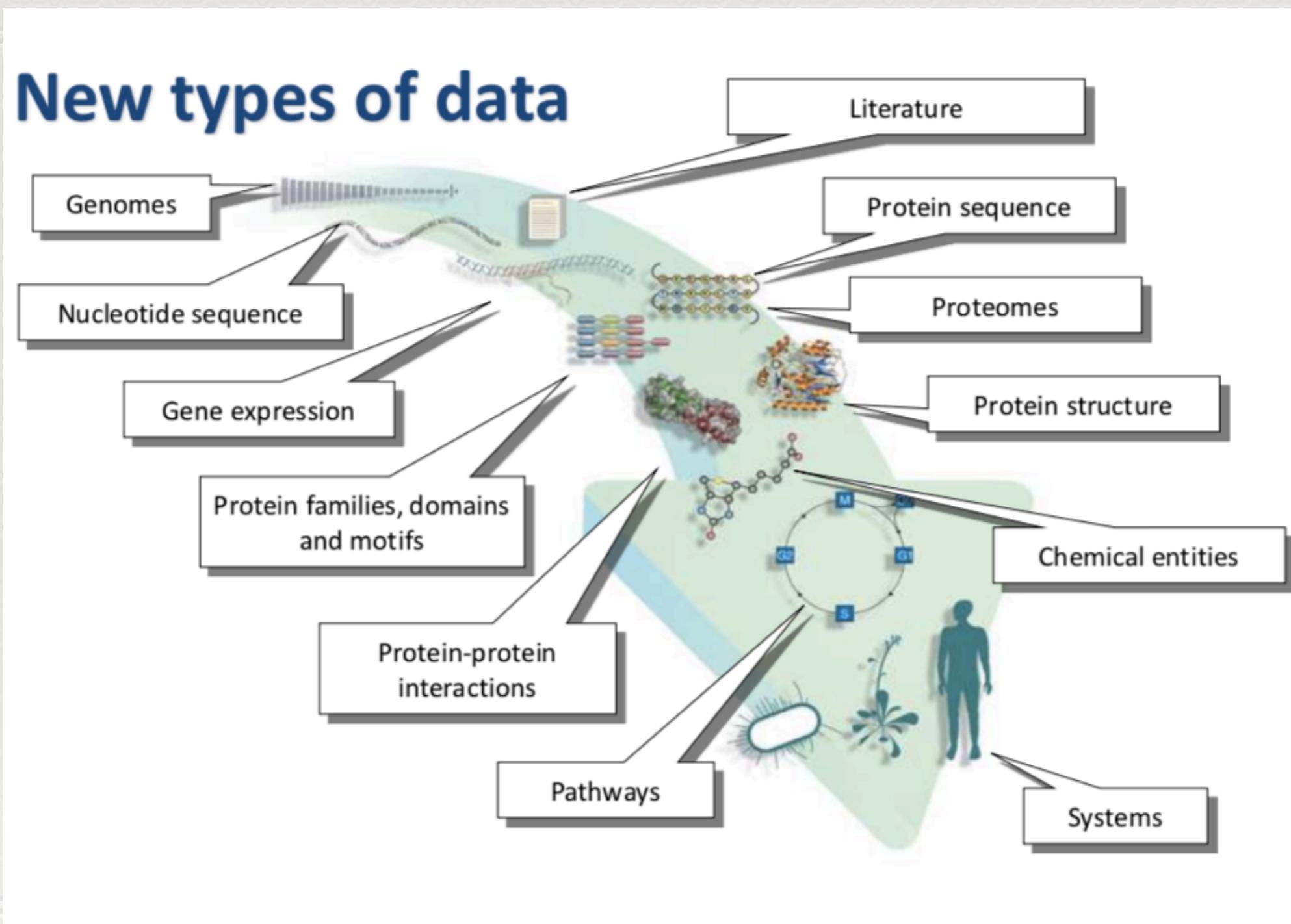
The need for bioinformatics has arisen from the recent explosion of publicly available genomic information, such as resulting from the Human Genome Project.

- Data explosion
- New types of data
- High-throughput biology
- Emphasis on systems, not reductionism

Why is bioinformatics necessary?



Why is bioinformatics necessary?



Why is bioinformatics necessary?

The present **bottlenecks** in bioinformatics include the education of biologists in the use of advanced computing tools, the recruitment of computer scientists into this evolving field, the limited availability of developed databases of biological information, and the need for more efficient and intelligent search engines for complex databases.

Application of bioinformatics

- To uncover the wealth of Biological information hidden in the mass of sequence, structure, literature and biological data.
- In agriculture, it can be used to produce high yield, low maintenance crops.
- It is being used now and in the foreseeable future in the areas of molecular medicine.
- It has environmental benefits in identifying waste and clean up bacteria.
-

Broad coverage of bioinformatics

Comparative genomics:

understanding the genomes of different species or populations

Functional genomics:

identification of genes and their respective functions

gene expression atlas

the regulation of gene expression (methylome, ATAC etc.)

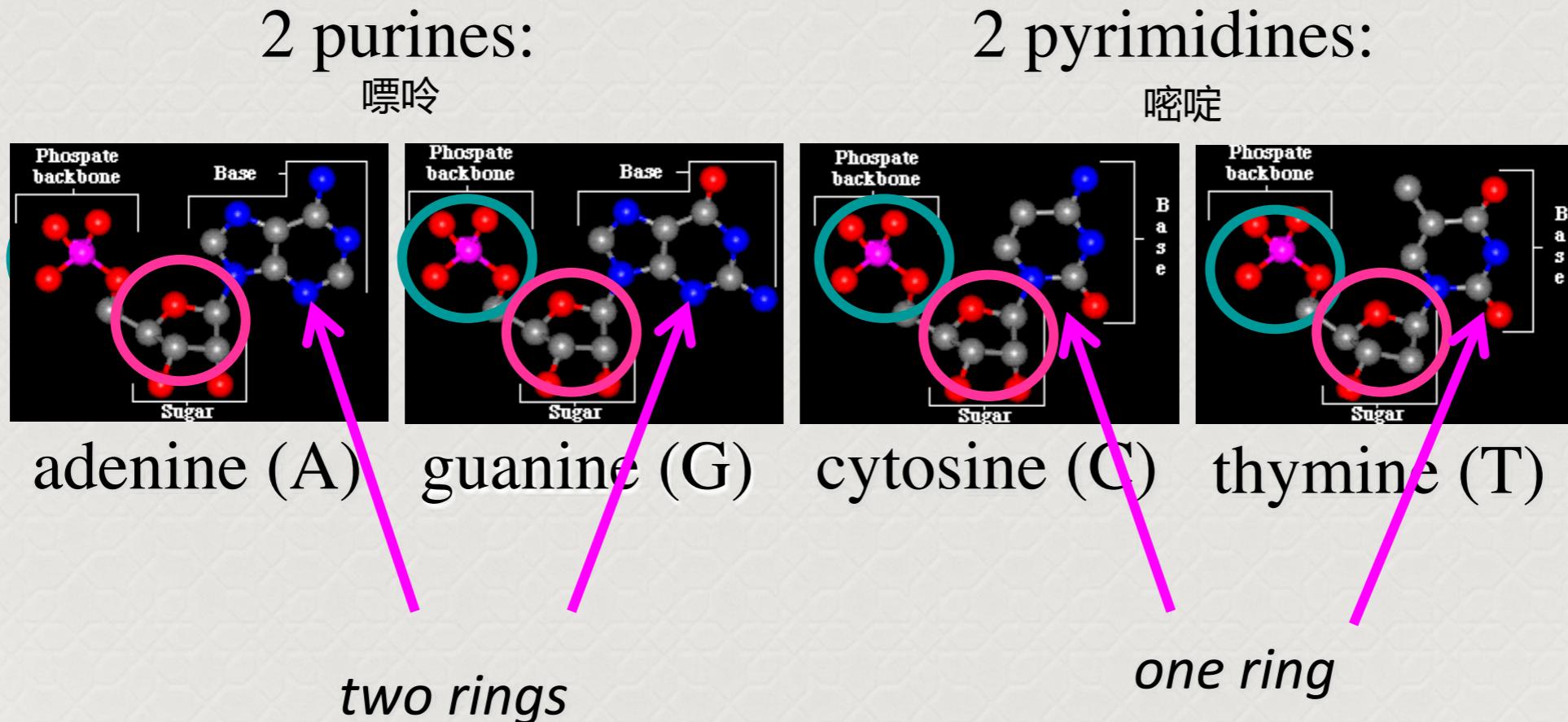
Structural genomics:

predictions related to functions of protein

Molecular bioinformatics

Molecular Bioinformatics involves the use of computational tools to discover new information in complex data sets (from the **one-dimensional information of DNA** through the **two-dimensional information of RNA** and the **three-dimensional information of proteins**, to the **four-dimensional information of evolving living systems**).

The hereditary information of all living organisms, with the exception of some viruses, is carried by deoxyribonucleic acid (**DNA**) molecules.

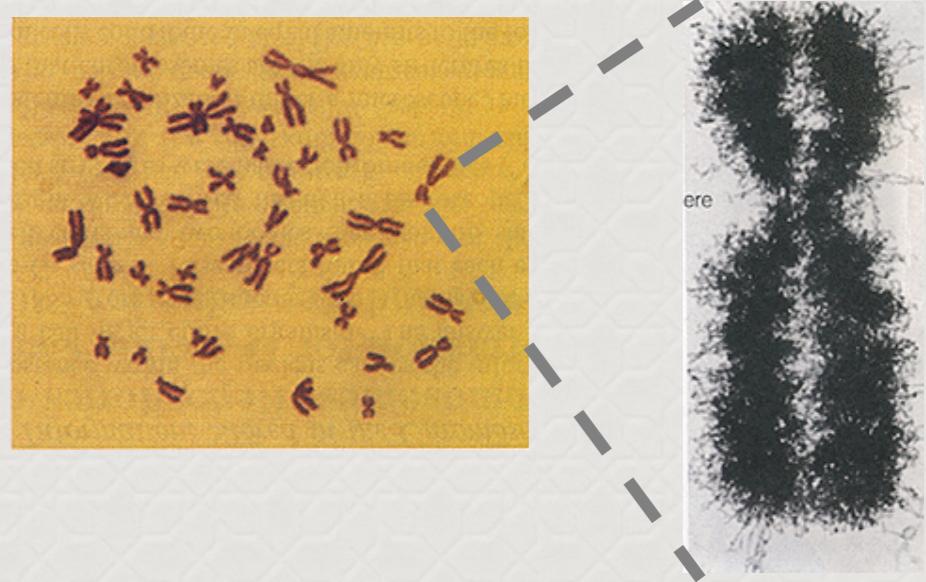


The entire complement of genetic material carried by an individual is called the **genome**.

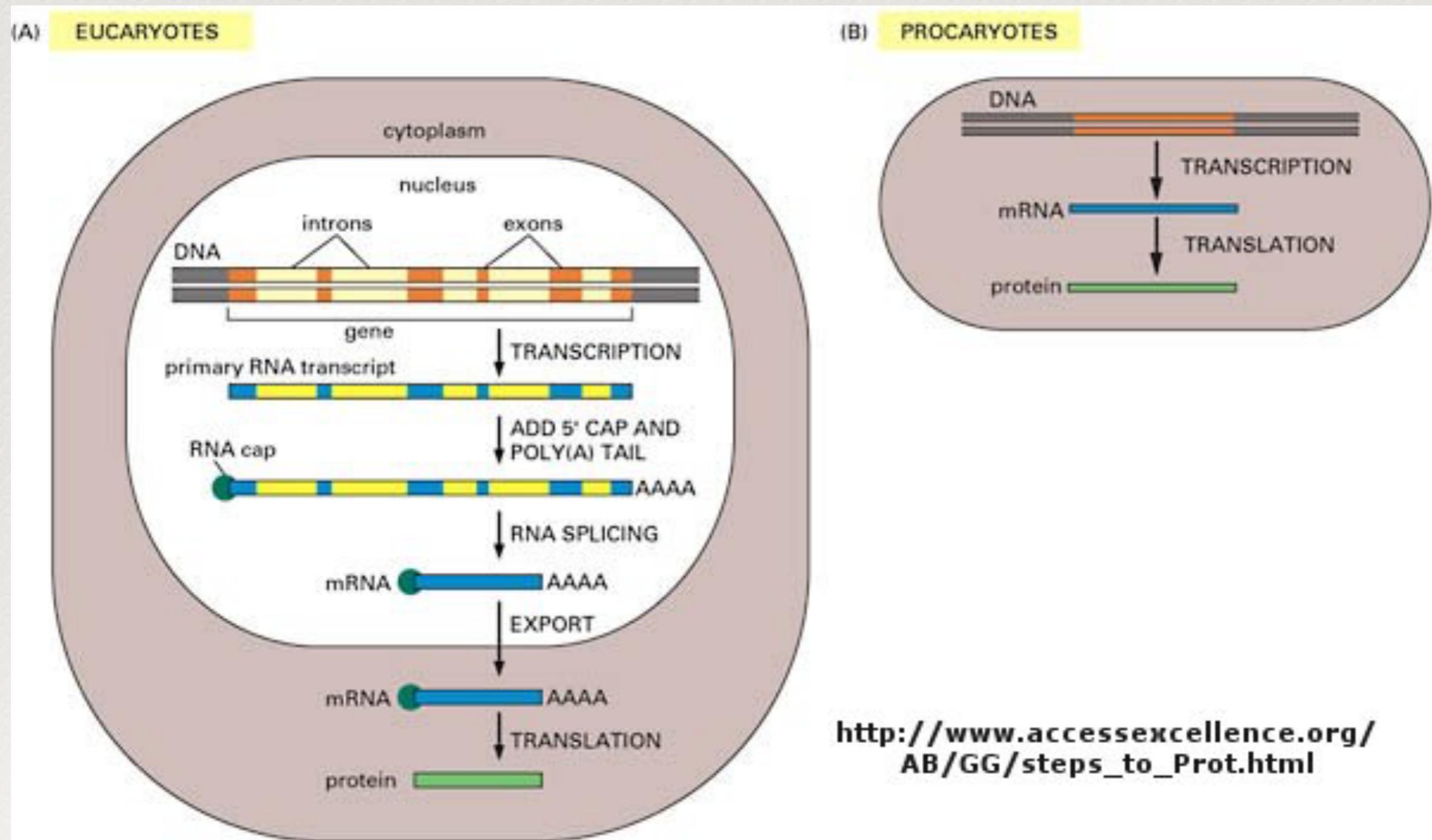
Eukaryotes may have up to 3 subcellular genomes:

1. Nuclear
2. Mitochondrial
3. Plastid

Bacteria have either circular or linear genomes and may also carry plasmids

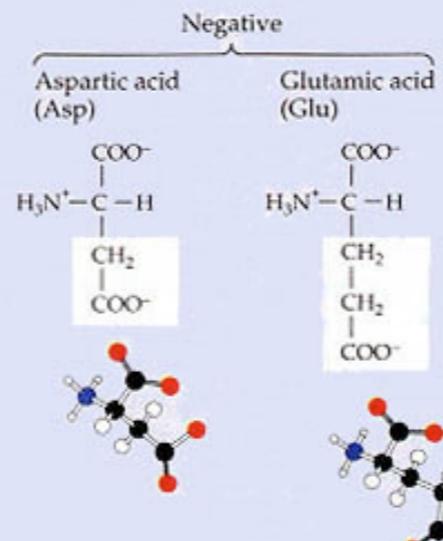
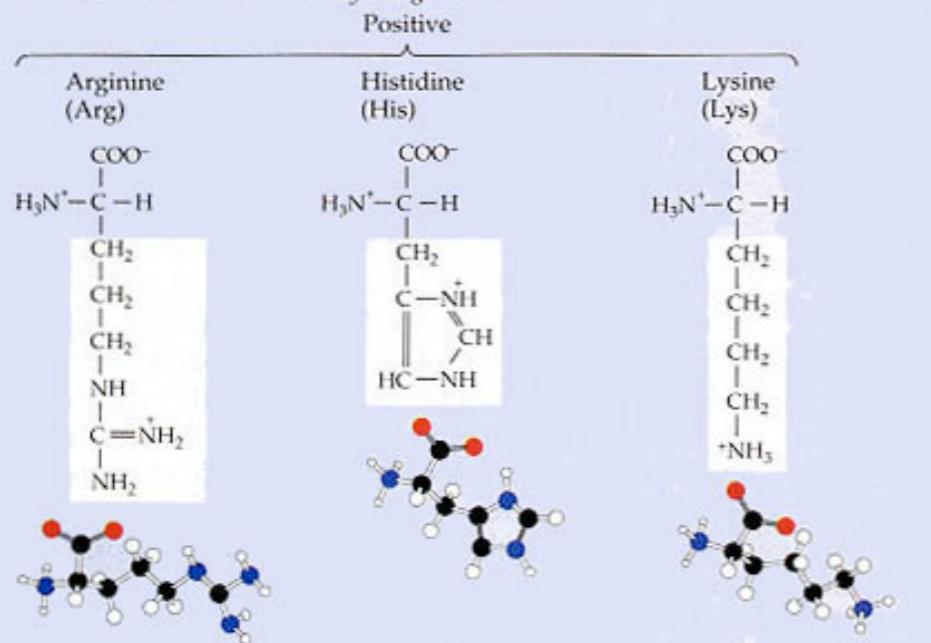


Central dogma: DNA makes RNA makes Protein

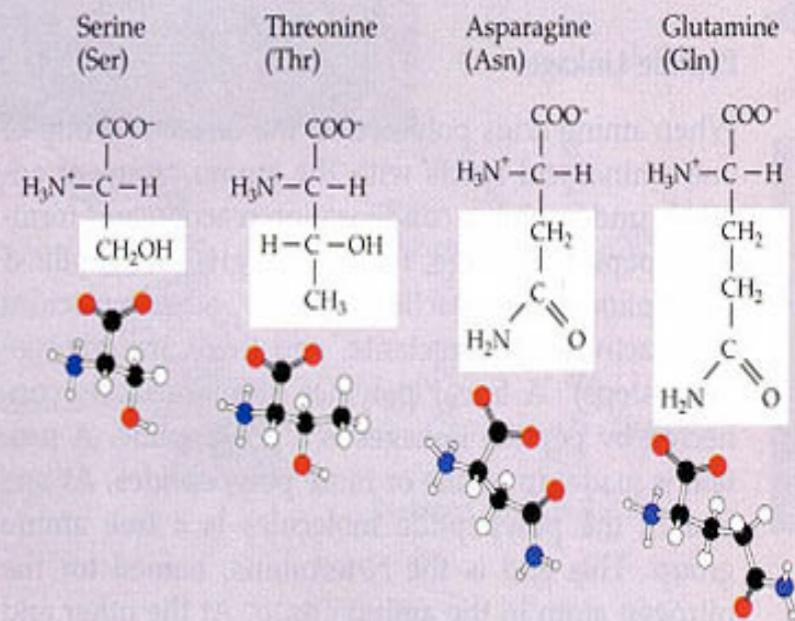


Amino acids - The protein building blocks

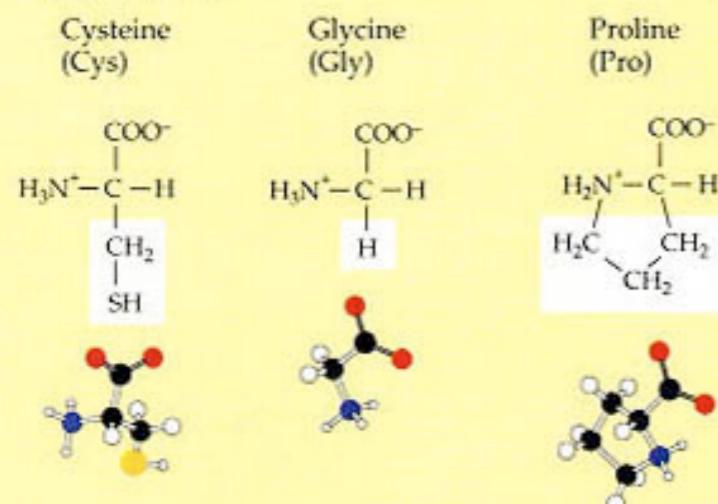
A. Amino acids with electrically charged side chains



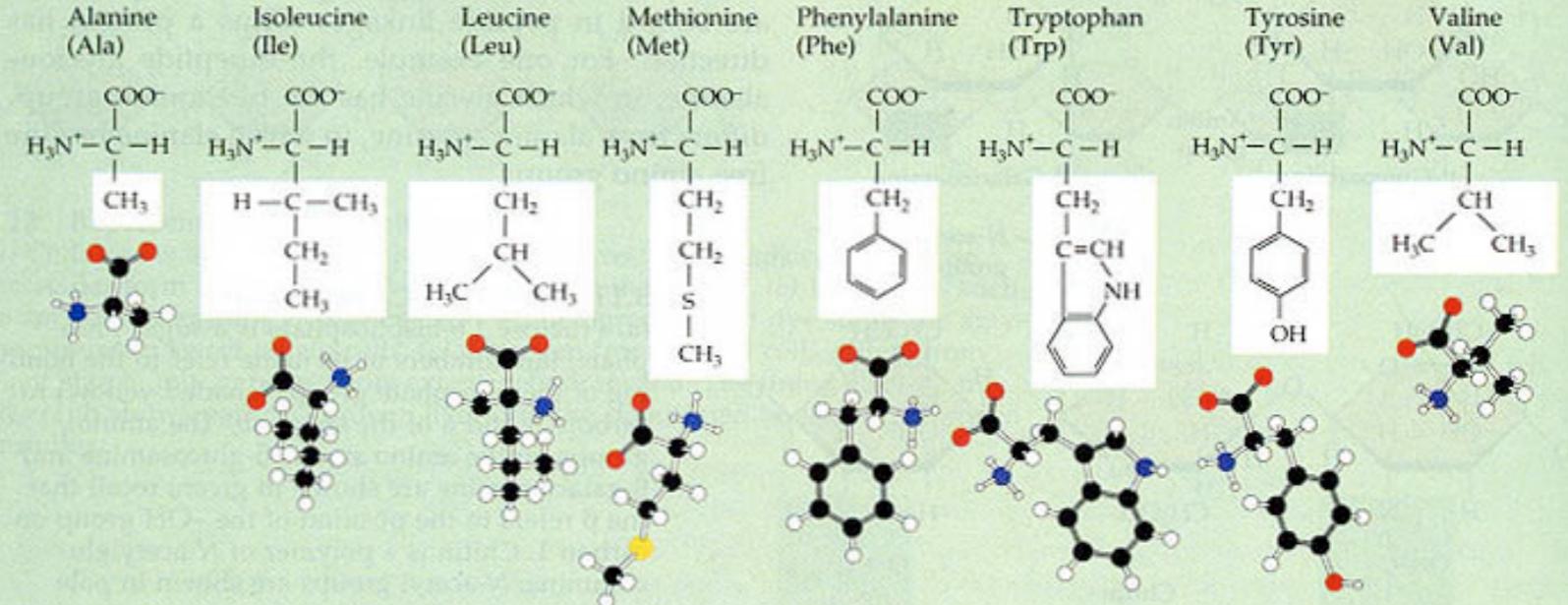
B. Amino acids with polar but uncharged side chains



C. Special cases



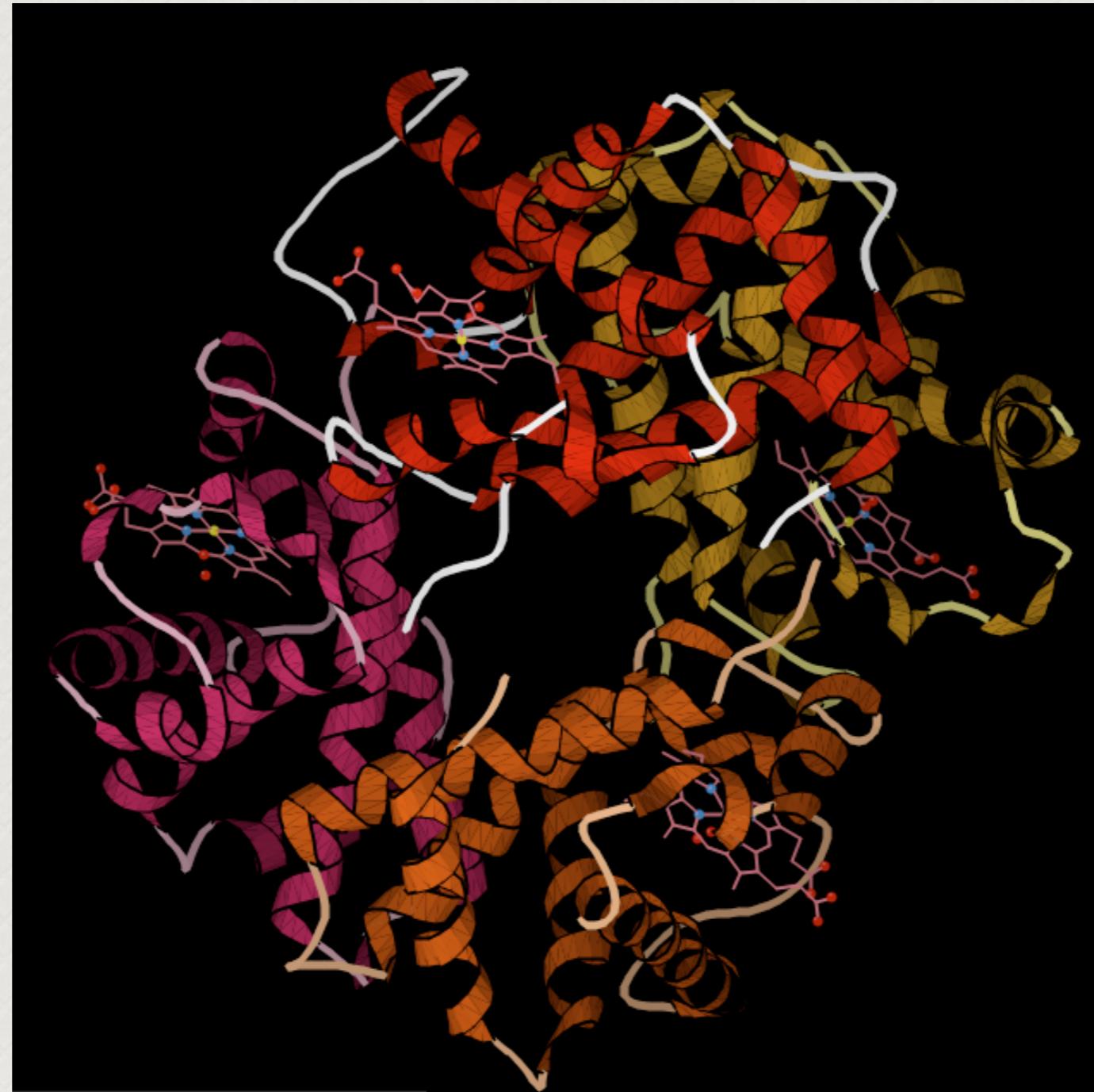
D. Amino acids with hydrophobic side chains



		Second letter					
		U	C	A	G		
First letter	U	UUU } Phe UUC UUU } Leu UUG	UCU } Ser UCC UCA UCG	UAU } Tyr UAC UAA Stop UAG Stop	UGU } Cys UGC UGA Stop UGG Trp	U C A G	Third letter
	C	CUU } CUC CUA } Leu CUG	CCU } CCC CCA CCG	CAU } His CAC CAA } Gln CAG	CGU } CGC CGA CGG	U C A G	
	A	AUU } AUC } Ile AUA AUG Met	ACU } ACC ACA ACG	AAU } Asn AAC AAA } Lys AAG	AGU } Ser AGC AGA AGG	U C A G	
	G	GUU } GUC } Val GUA GUG	GCU } GCC GCA GCG	GAU } Asp GAC GAA } Glu GAG	GGU } GGC GGA GGG	U C A G	

Protein folding

a human hemoglobin



How does it all looks like on a computer monitor?

a DNA sequence

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTGCCGACAAGACCAACGTCAAGGCCG
CCTGGGGTAAGGTGGCGCGACGCTGGCGAGTATGGTGCAGGAGGCCCTGGAGAGGATGTTCTGTCCTCCCCACC
ACCAAGACCTACTTCCCGACTTCGACCTGAGCCACGGCTCTGCCAGGTTAAGGGCACGGCAAGAAGGTGGCCGA
CGCGCTGACCAACGCCGTGGCGACGTGGACGACATGCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGACA
AGCTTCGGGTGGACCCGGTCAACTCAAGCTCTAACGCCACTGCCTGCTGGTACCCCTGGCCGCCACCTCCCCGCC
GAGTTCACCCCTGCGGTGCACGCCCTGGACAAGTTCTGGCTTGTGAGCACCCTGACCTCCAAATACCG
TTAAGCTGGAGCCTCGGTGGCCATGCTTCTGCCCTGGCCCTCCCCCAGCCCTCCCTGCACCCGT
ACCCCGTGGTCTTGAATAAGTCTGAGTGGCGGC
```

How does it all looks like on a computer monitor?

a DNA sequence

```
>gi|14456711|ref|NM_000558.3| Homo sapiens hemoglobin, alpha 1 (HBA1), mRNA
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGtgcgtgtcctggacaaggaccacgtcaaggcc
gcctgggtaaggtcgcgacgcgtggcgagtatggtgcggaggccctggagaggatgttcctgtcctccccac
caccaagacctattcccgcacttcgacctgagccacggctctgcccaggtaagggccacggcaagaagggtggccg
acgcgctgaccaacgcgtggcgacgtggacatgccaacgcgctgtccgcctgagcgacctgcacgcgcac
aagcttcgggtggaccggtaactcaagctctaagccactgcctgctggtgaccctggccgcccacctcccgcc
cgagttcacccctgcggtgacgcctccctggacaagttcctggcttgtgagcaccgtgctgacctccaaatacc
GTAAgctggagcctcggtgcacgcctccctggacaagttcctggcttgtgagcaccgtgctgacctccaaatacc
gtaccccgtaggtttgaataaaagtctgagtggcgcc
```

a protein sequence

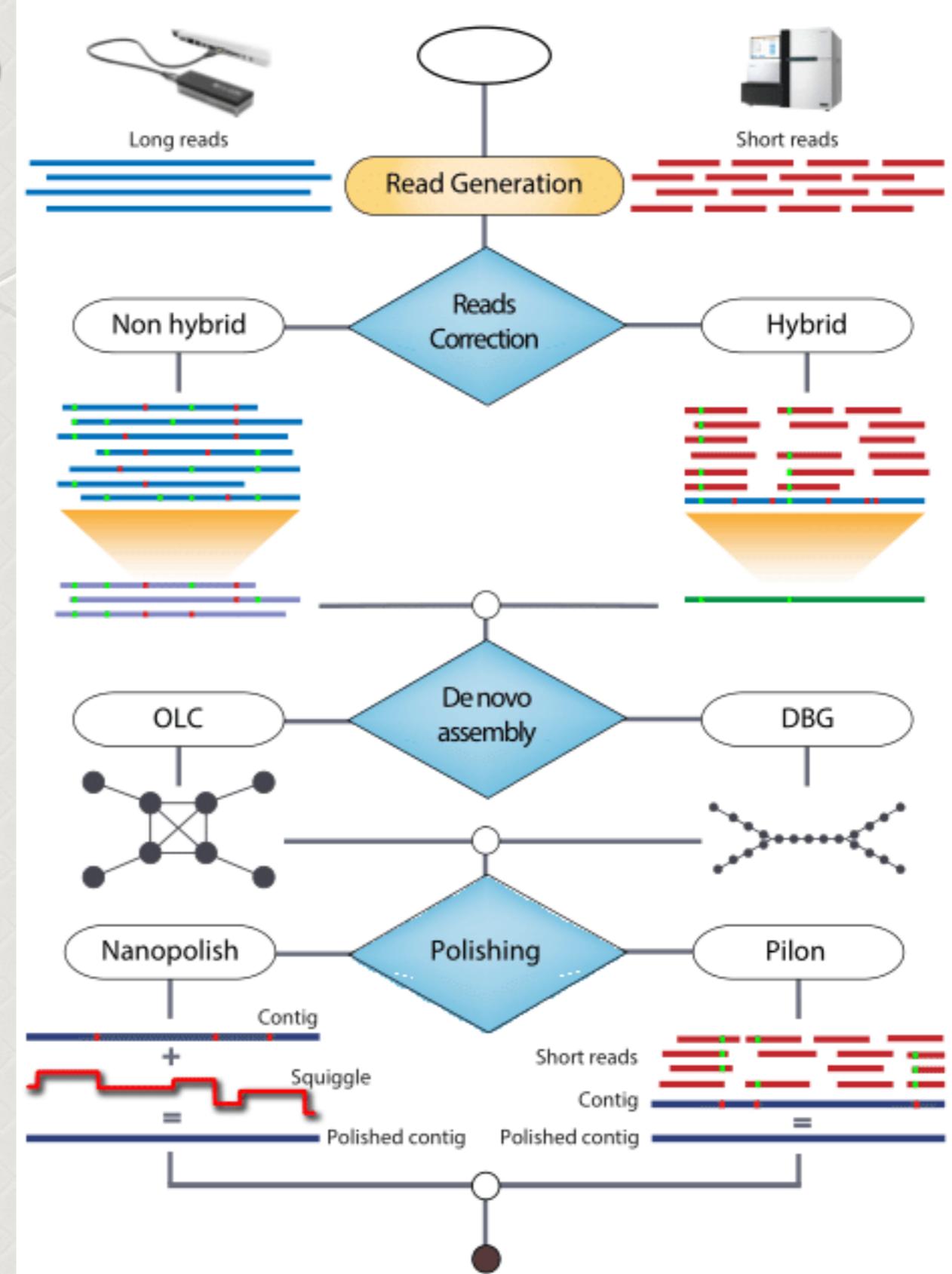
```
>gi|4504347|ref|NP_000549.1| alpha 1 globin [Homo sapiens]
MVLSPADKTNVKAAWGVGAHAGEYGAELERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAH
VDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
```

How big is a human genome?

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGG
GGTAAGGTGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTCCTGTCCTTCCCCACCAAGACCT
ACTTCCCGCACTCGACCTGAGCCACGGCTCTGCCAGGTTAAGGCCACGGCAAGAAGGTGGCCACGCCGTACCCAACGC
CGTGGCGACGTGGACGACATGCCAACGCCGTGTCGCCCTGAGCGACCTGCACGCCACAAGCTCGGGTGGACCCGGTC
AACTCAAGCTCTAAGCCACTGCCTGCTGGTACCCCTGGCGCCACCTCCCCGCCAGTTACCCCTGCCGTGCACGCC
CCCTGGACAAGTCCTGGCTCTGTGAGCACCGTGTGACCTCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCT
TGCCCCCTGGCCTCCCCCAGCCCCCTCCCTGCACCCGTACCCCGTGGTCTTGAATAAAAGTCTGAGTGGCG
GCACTCTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGC
GGGGTAAGGTGGCGCGACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTCCTGTCCTTCCCCACCAAGAC
CTACTTCCGCACCTCGACCTGAGCCACGGCTCTGCCAGGTTAAGGCCACGGCAAGAAGGTGGCCACGCCGTACCAAC
GCCGTGGCGCACGTGGACGACATGCCAACGCCGTGTCGCCCTGAGCGACCTGCACGCCACAAGCTCGGTGGACCCGG
TCAACTTCAAGCTCTAAGCCACTGCCTGCTGGTACCCCTGGCGCCACCTCCCCGCCAGTTACCCCTGCCGTGCACCC
CTCCCTGGACAAGTCCTGGCTTCTGTGAGCACCGTGTGACCTCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTT
CTTGCCCCCTGGCCTCCCCCAGCCCCCTCCCTGCACCCGTACCCCGTGGTCTTGAATAAAAGTCTGAGTGG
CGGCACTCTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCC
CTGGGTAAGGTGGCGCGACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTCCTGTCCTTCCCCACCAAG
ACCTACTTCCGCACCTCGACCTGAGCCACGGCTCTGCCAGGTTAAGGCCACGGCAAGAAGGTGGCCACGCCGTACCA
ACGCCGTGGCGCACGTGGACGACATGCCAACGCCGTGTCGCCCTGAGCGACCTGCACGCCACAAGCTCGGTGGACCC
GGTCAACTTCAAGCTCTAAGCCACTGCCTGCTGGTACCCCTGGCGCCACCTCCCCGCCAGTTACCCCTGCCGTGCAC
GCCTCCCTGGACAAGTCCTGGCTTCTGTGAGCACCGTGTGACCTCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGC
TTCTGCCCCCTGGCCTCCCCCAGCCCCCTCCCTCCCTGCACCCGTACCCCGTGGTCTTGAATAAAAGTCTGAGTGG
GGCGGCCGTGGCGCACGTGGACGACATGCCAACGCCGTGTCGCCCTGAGCGACCTGCACGCCACAAGCTCGGTGG
ACCCGGTCAACTTCAAGCTCTAAGCCACTGCCTGCTGGTACCCCTGGCGCCACCTCCCCGCCAGTTACCCCTGCCGT
GCACGCCCTGGACAAGTCCTGGCTTCTGTGAGCACCGTGTGACCTCAAATACCGTTAAGCTGGAGCCTCGGTGGCC
ATGCTTCTGCCCCCTGGCCTCCCCCAGCCCCCTCCCTCCCTGCACCCGTACCCCGTGGTCTTGAATAAAAGTCTG
AGTGGCGGGCACTCTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAA
GGCCGCCCTGGGTAAGGTGGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTCCTGTCCTTCCCCACC
ACCAAGACCTACTTCCGCACCTCGACCTGAGCCACGGCTCTGCCAGGTTAAGGCCACGGCAAGAAGGTGGCCG...

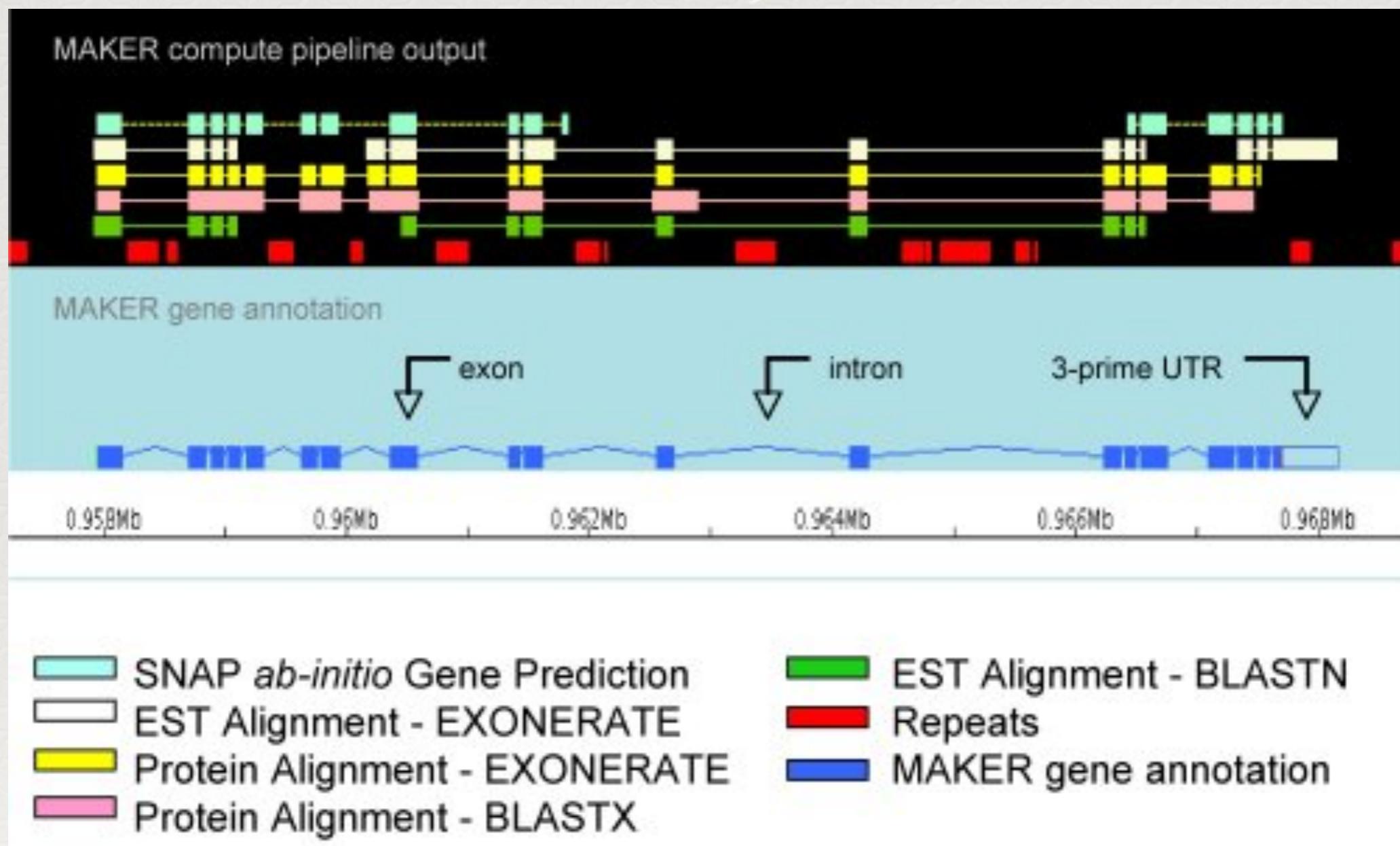
What do we actually do with bioinformatics?

Genome assembly



What do we actually do with bioinformatics?

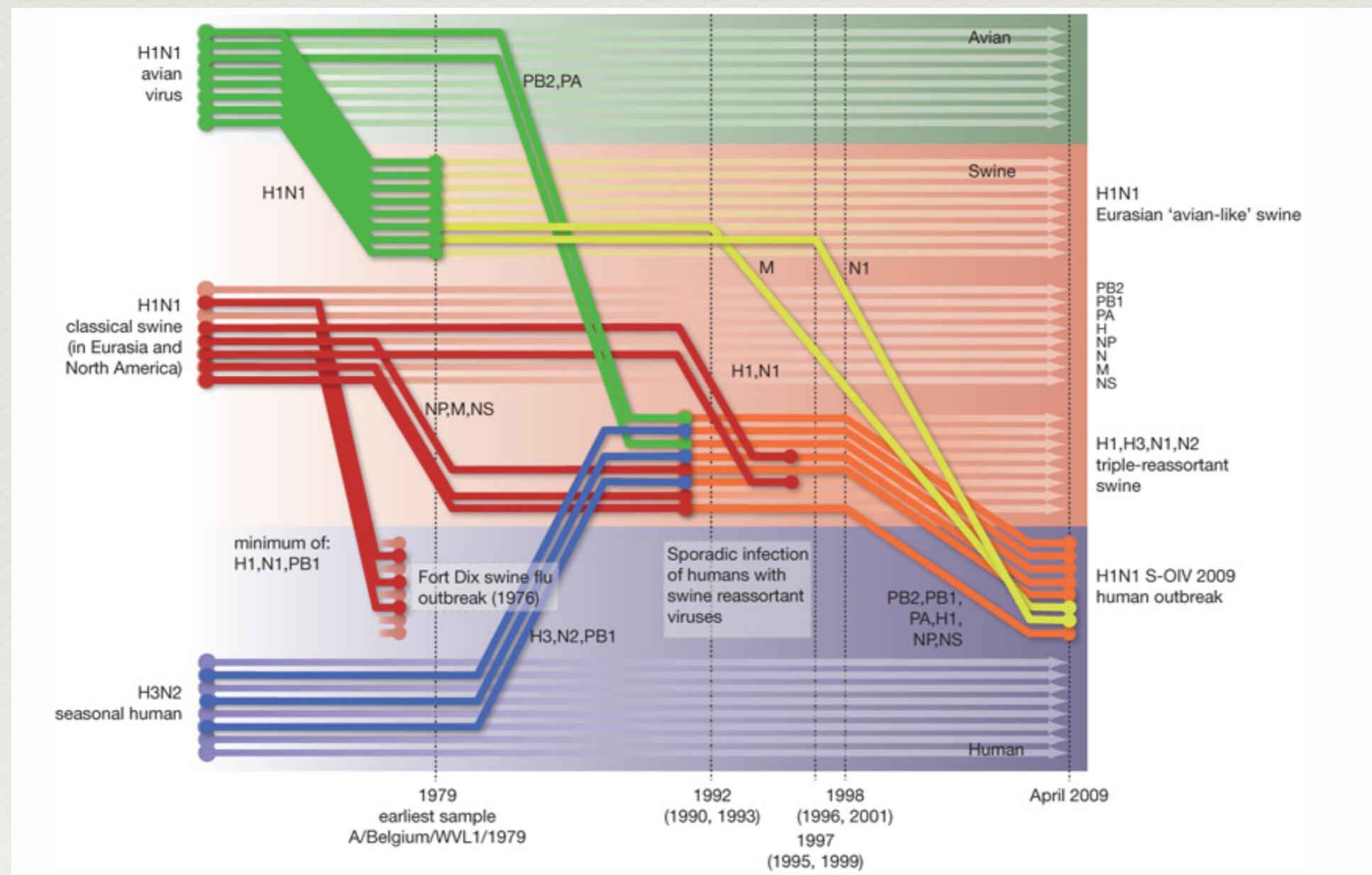
Genome annotation



What do we actually do with bioinformatics?

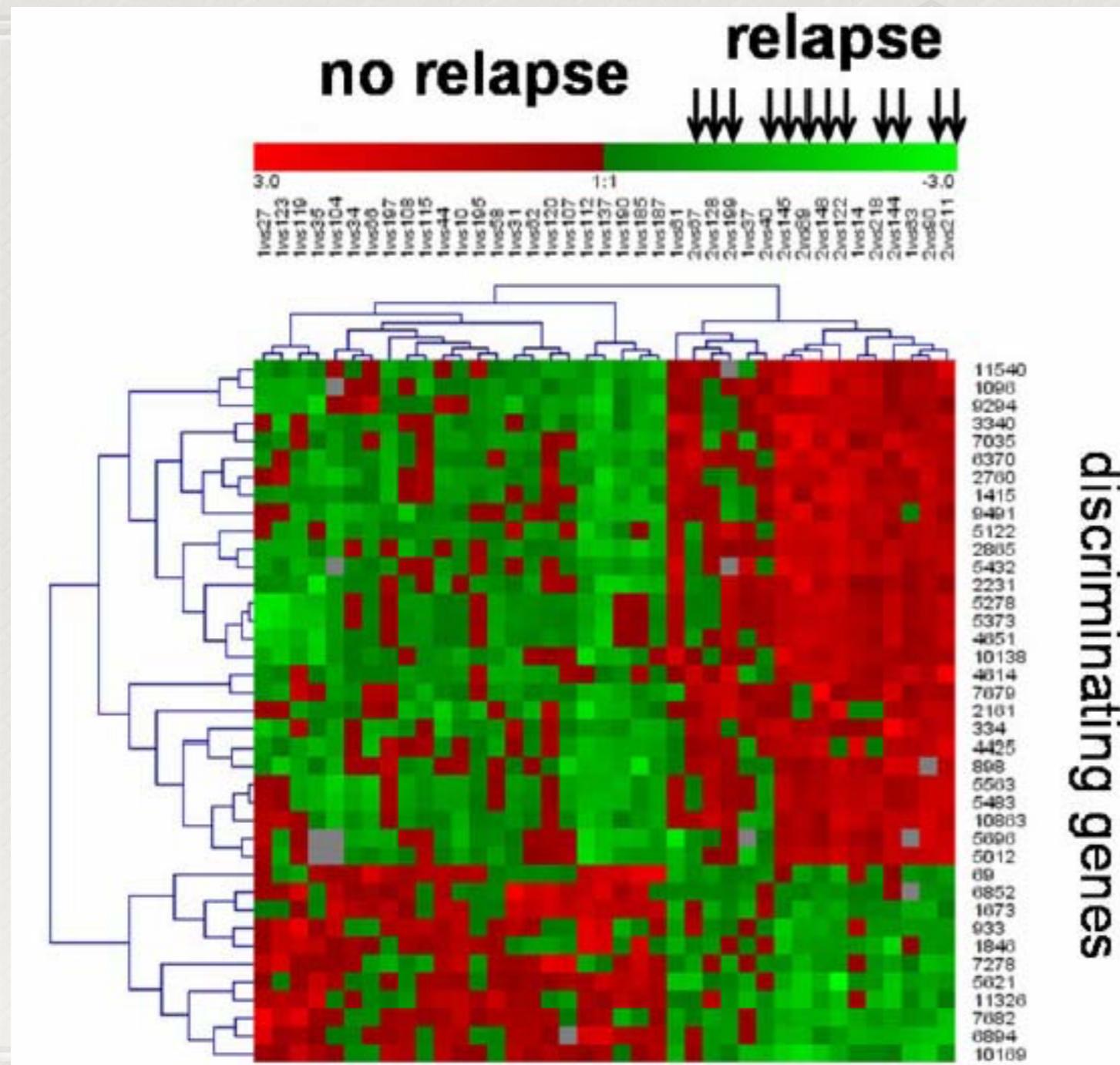
Molecular evolution

Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic



What do we actually do with bioinformatics?

Gene expression analysis

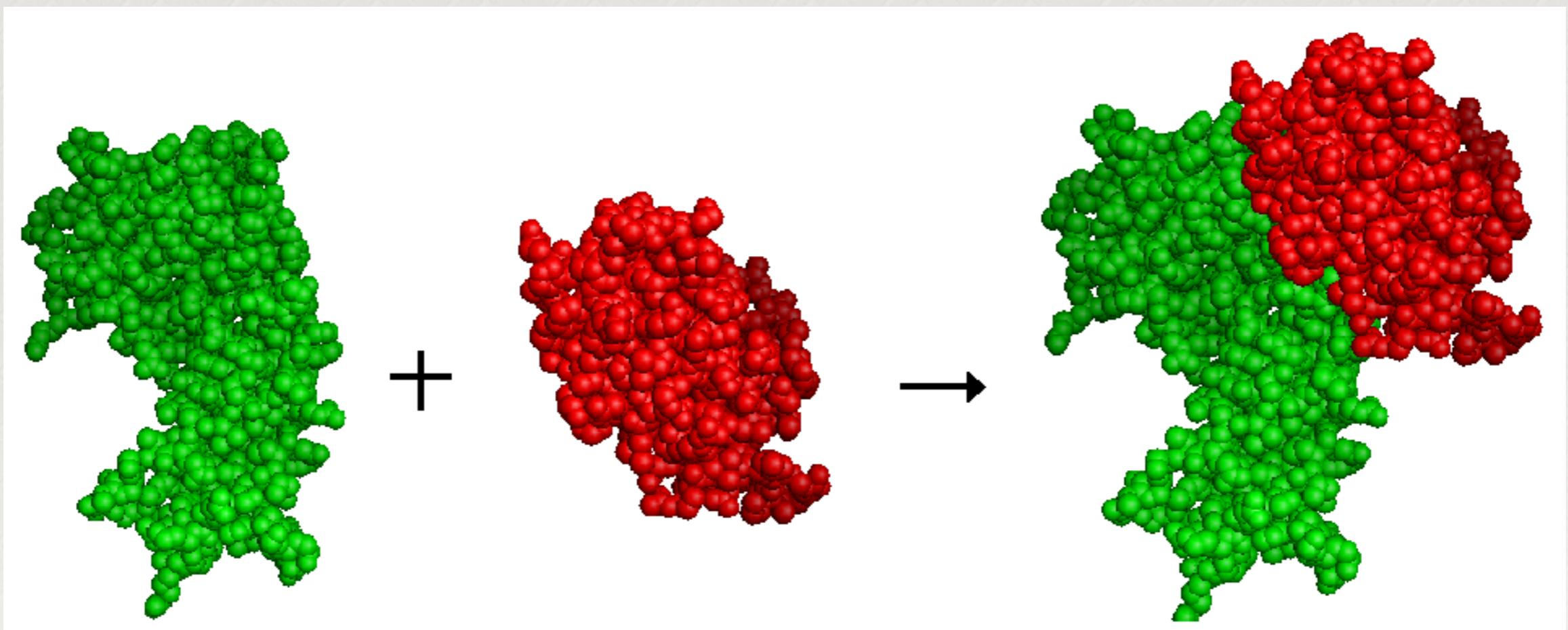


A set of 39 genes discriminates between the two classes of tumors.

What do we actually do with bioinformatics?

Protein structure prediction

Protein docking



Online Learning Resources

生信技能树：

<http://www.bioproject.com/>

<https://www.jianshu.com/u/d645f768d2d5>

基因学苑：

<https://zhuanlan.zhihu.com/p/40497508>

陈巍学基因：

<http://i.youku.com/u/UMTlwNTc0MjU2OA==>

陈连福的生信博客：

<http://www.chenlianfu.com/?cat=3>

Online Glossaries

Bioinformatics :

<http://www.geocities.com/bioinformaticsweb/glossary.html>

<http://big.mcw.edu/>

Genomics:

<http://www.geocities.com/bioinformaticsweb/genomicglossary.html>

Molecular Evolution:

<http://workshop.molecularevolution.org/resources/glossary/>

Biology dictionary:

http://www.biology-online.org/dictionary/satellite_cells

Other glossaries, e.g., the list of phobias:

<http://www.phobialist.com/class.html>