Welcome to Practical Bioinformatics !!!

2020/09/16

Li Wang (wangli03@caas.cn)

个人简介

- **2002 – 2006** 武汉大学　　生物技术基地班　　学士
- **2006 – 2011** 中国科学院植物研究所　　植物学　　博士

2009.2 - 2010.11 德国哥廷根大学　　(联合培养博士生)

2009.12 ; 2010.4 英国自然历史博物馆 (短期访问学者)

- **2012.3 – 2014.3**　美国德州理工大学　　博士后　　Dr. Matt Olson
- **2014.4 – 2019.6**　美国爱荷华州立大学　　博士后　　Dr. Matt Hufford
- **2018.7 – 2019.6**　美国加州大学戴维斯分校　　博士后　　Dr. Jeffrey Ross-Ibarra
- **2019.7 –** 现在　　中国农科院深圳农业基因组研究所

团队成员

博士后人员：

李诚	博士	浙江大学
孙士超	博士	石河子大学
陈雪青	博士	上海师范大学
柳小莉	博士	成都中医药大学

科研助理： 周微

博士研究生：

韩笑雨 陈新连（中山大学）

硕士研究生：

姬姣姣 臧兰兰 宋依婷

website: wanglilab.github.io

Who are we?

Dr. Yuwen Liu	liuyuwen@caas.cn
Dr. Li Wang	wangli03@caas.cn
Dr. Wenlong Ma	mawenlong_nwsuaf@163.com
Dr. Cheng Li	licheng@caas.cn
Dr. Yi Zou	zouyi@caas.cn
Weigang Zheng	zhengweigang@caas.cn
Chao Wang	netwc@qq.com
Yang Fu	yfu1116@163.com
Xuezhu Liao	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/WangliLab/CAAS_PracticalBioinformatics_2020Aut

please join our wechat group

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

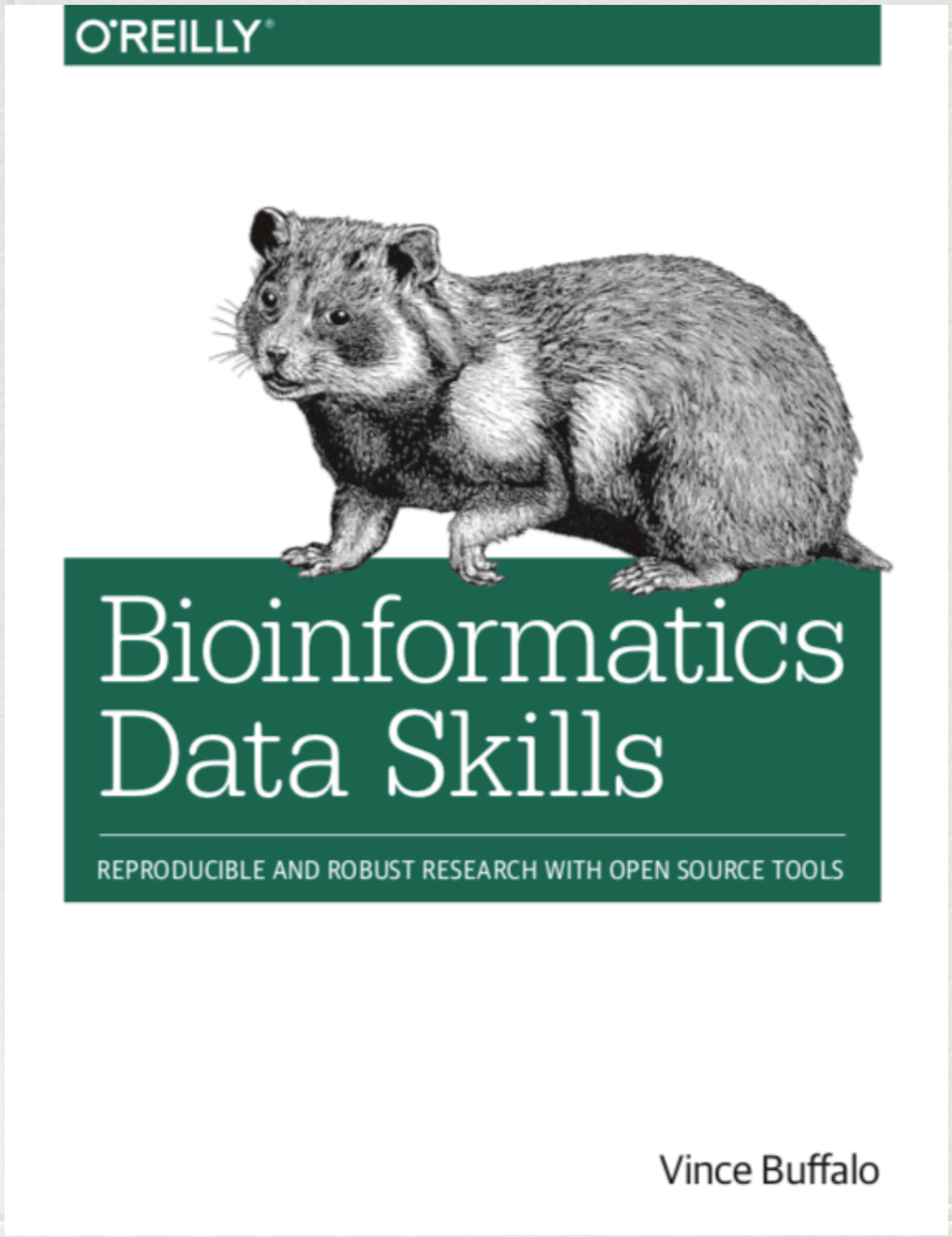
Assignment 1: Unix 10%

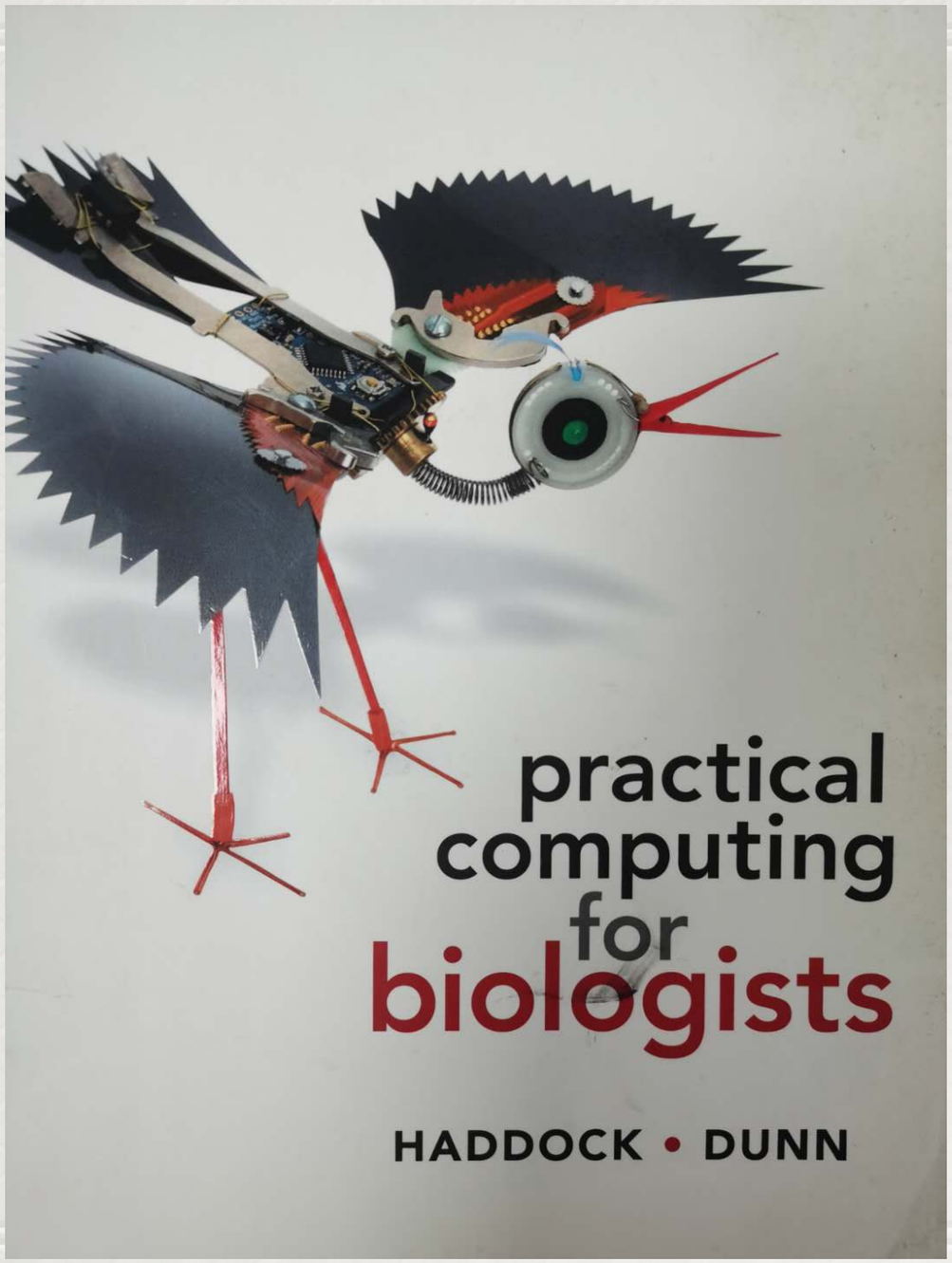
Assignment 2: R 10%

Assignment 3: Python 10%

Group Project and Presentation 70%

Additional materials to read





Assignment

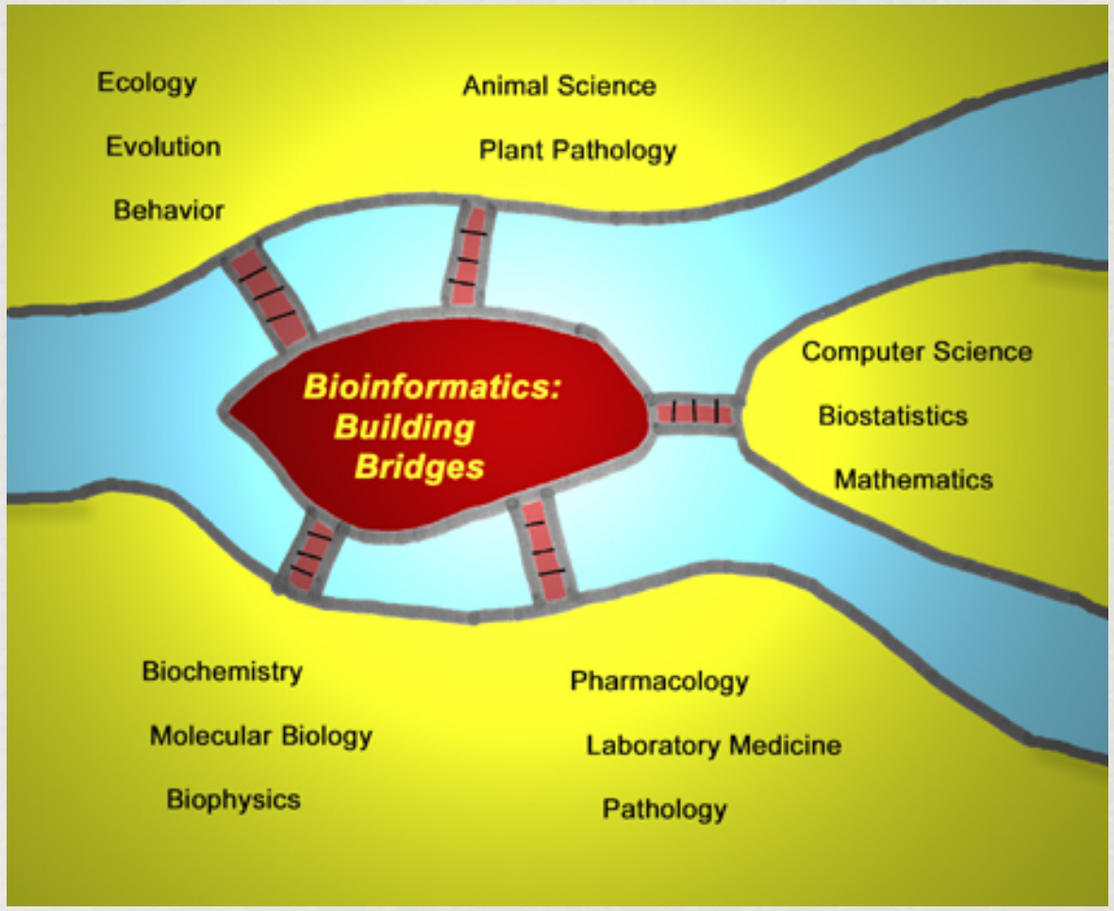
Read Chapters 1,2,4,5 of Buffalo

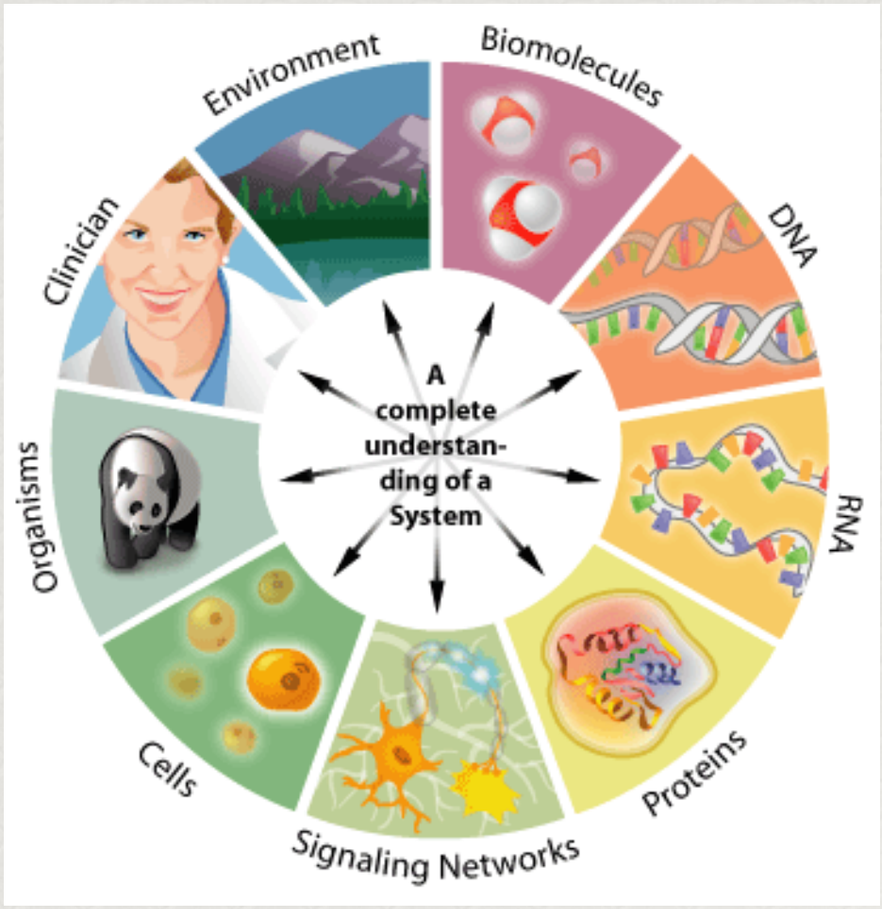
Introduction to Bioinformatics

2020/09/16

Li Wang (wangli03@caas.cn)

What is bioinformatics?





What is bioinformatics?

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

Computer scientists
(+Mathematicians,
Statisticians, etc.)
Develop tools, softwares,

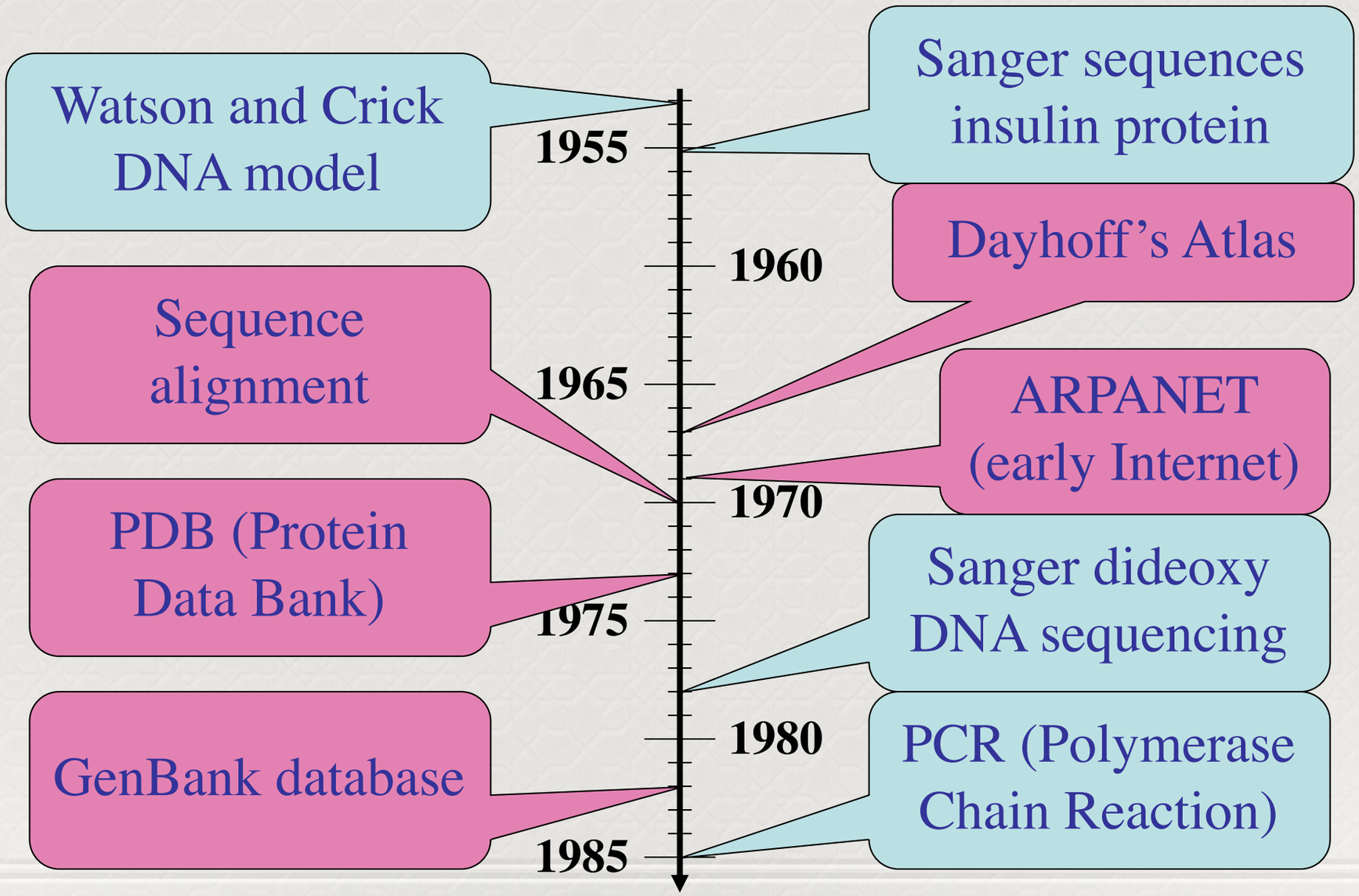
Bioinformaticians
Study biological questions by
analyzing molecular data

What is bioinformatics?

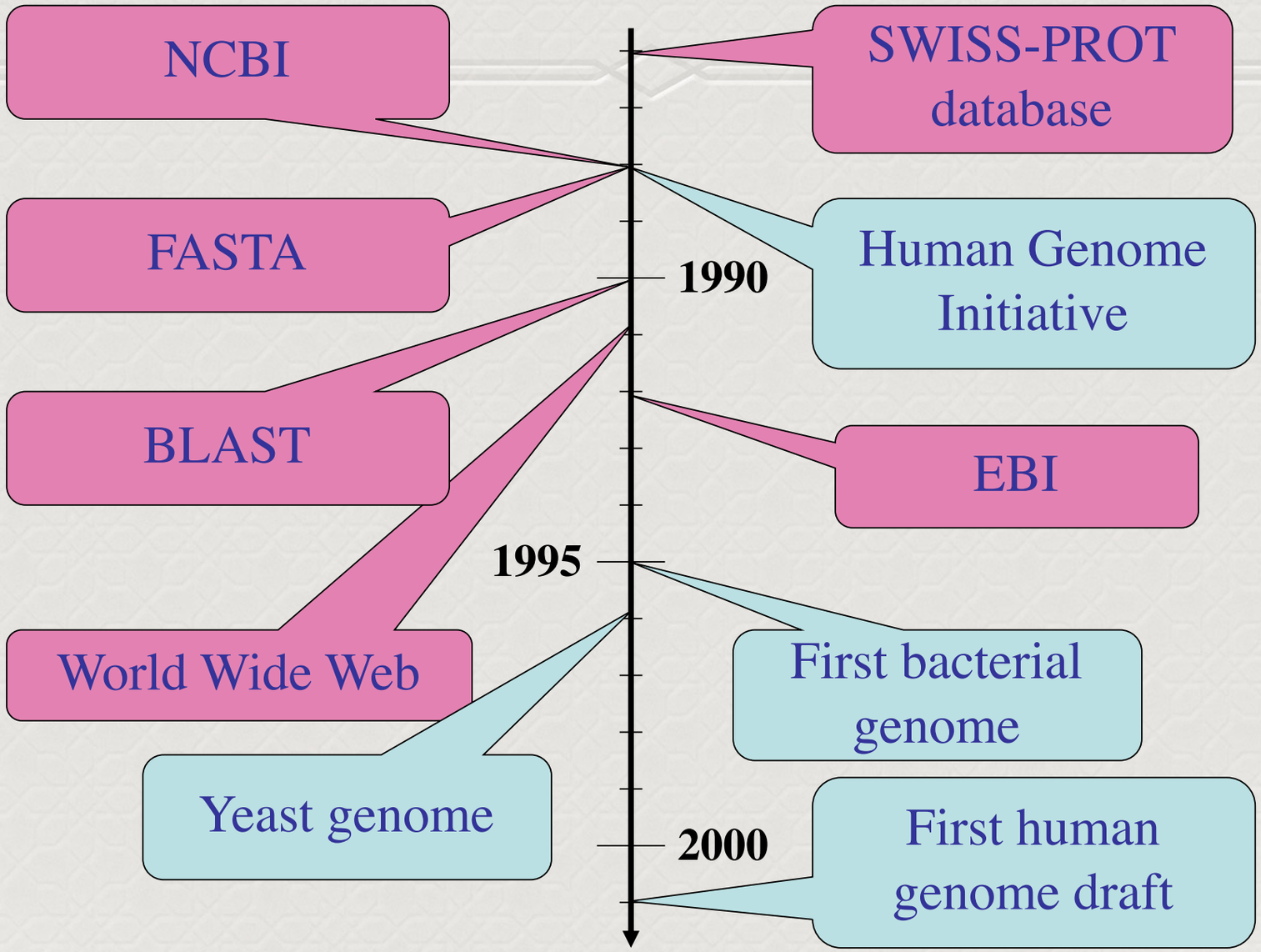
- First introduced in 1990s.
- It involves the computational tools and methods used to manage, analyze and manipulate volumes and volumes of biological data.
- It is an interdisciplinary approach requiring advanced knowledge of computer mathematics and statistical methods for understanding of biological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

From DNA to Genome

How many years has it taken from the discovery of DNA structure to the first published genome?



From DNA to Genome



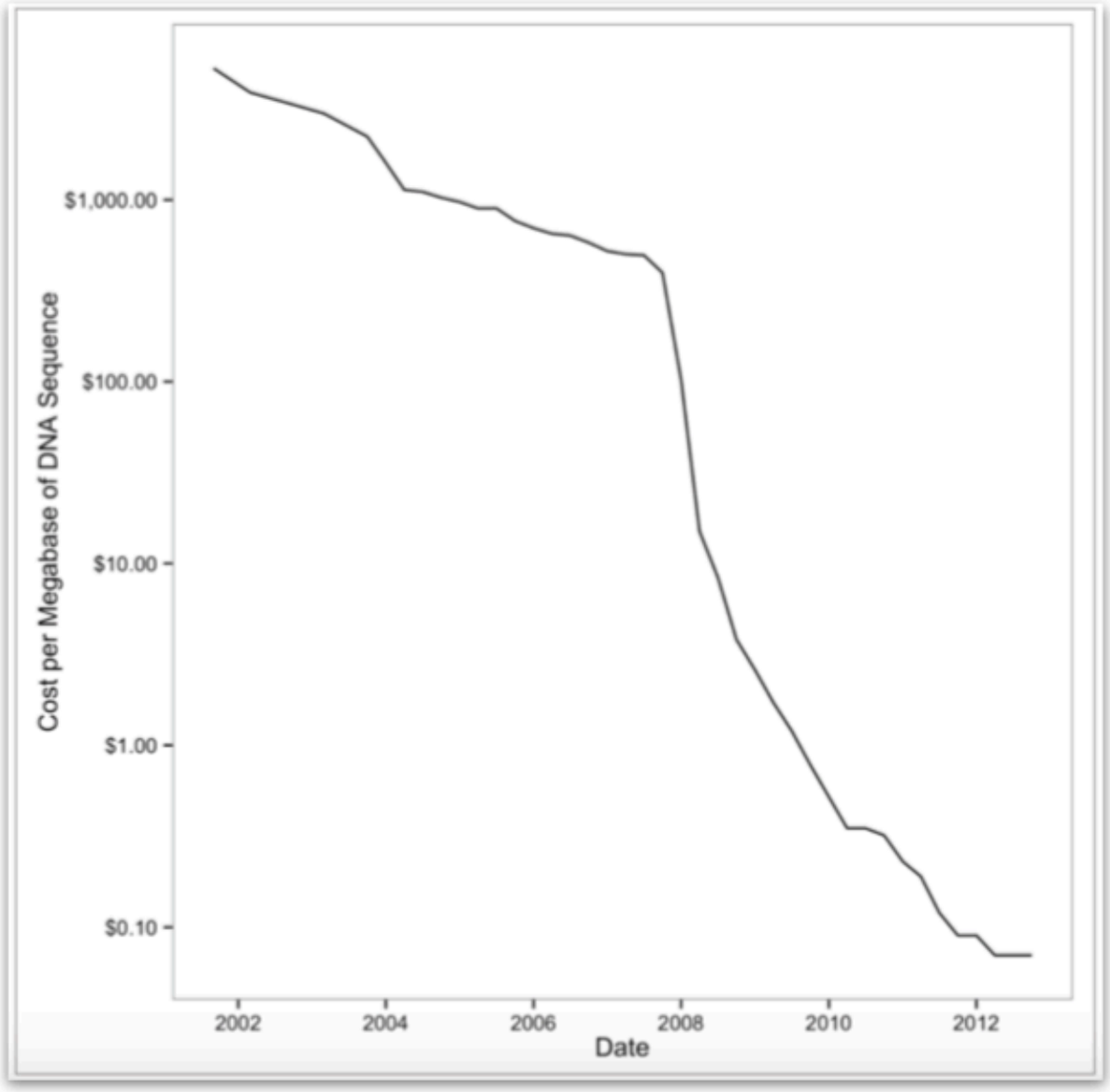
https://www.nature.com/articles/s41588-019-0570-0

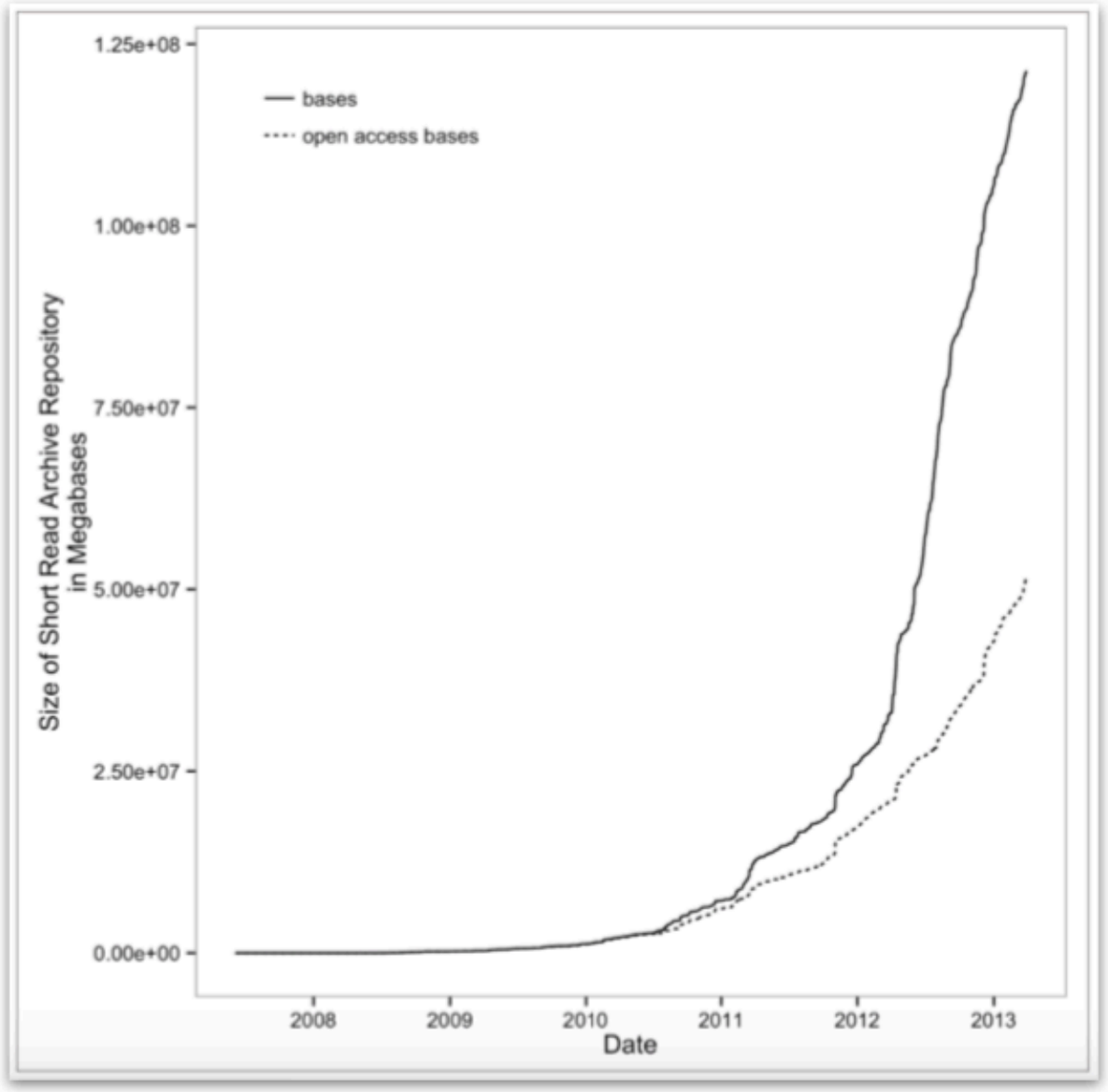
Why is bioinformatics necessary?

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

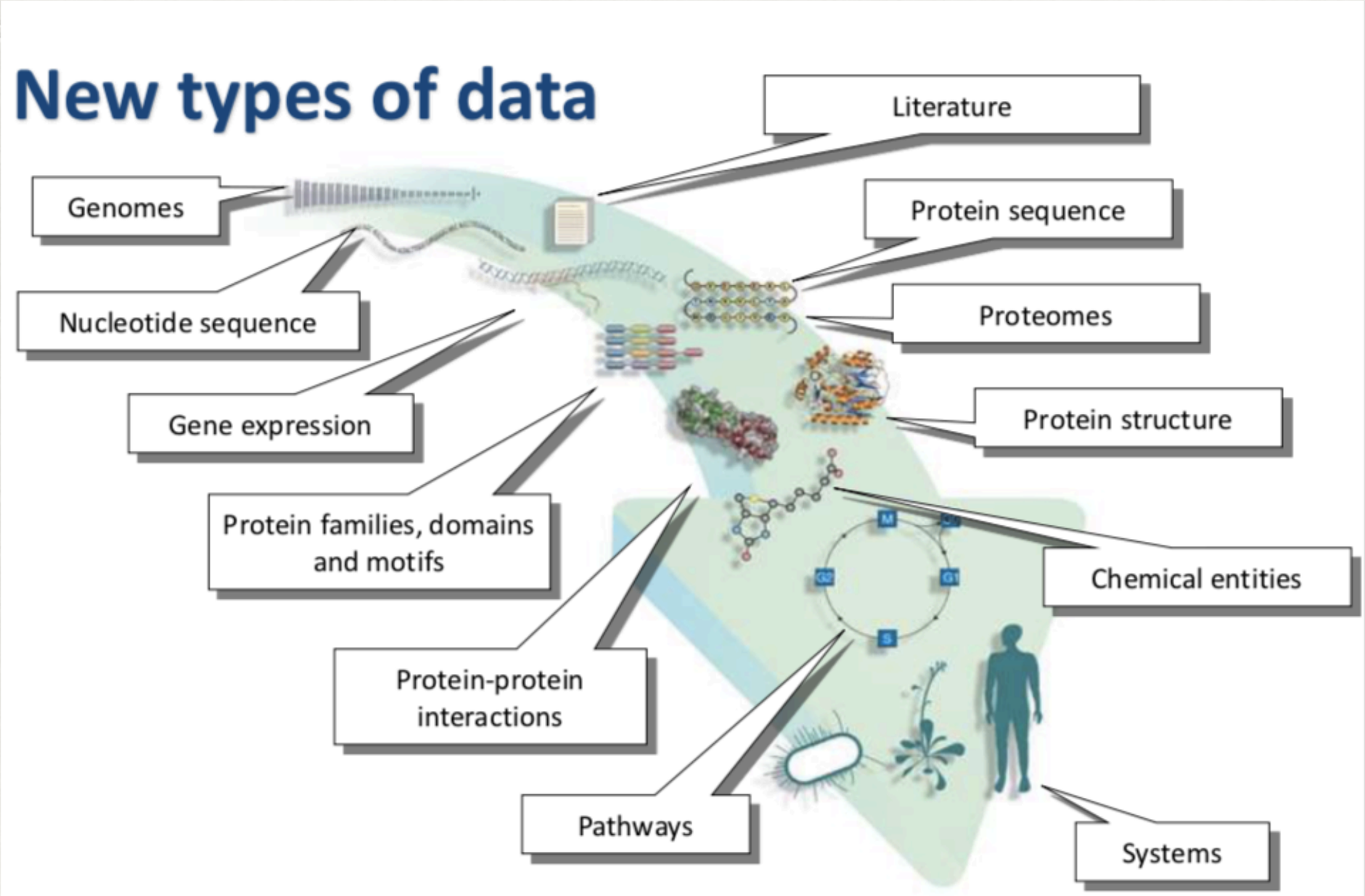
- Data explosion
- New types of data
- High-throughput biology
- Emphasis on systems, not reductionism

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.
- ∘ ∘ ∘ ∘ ∘ ∘

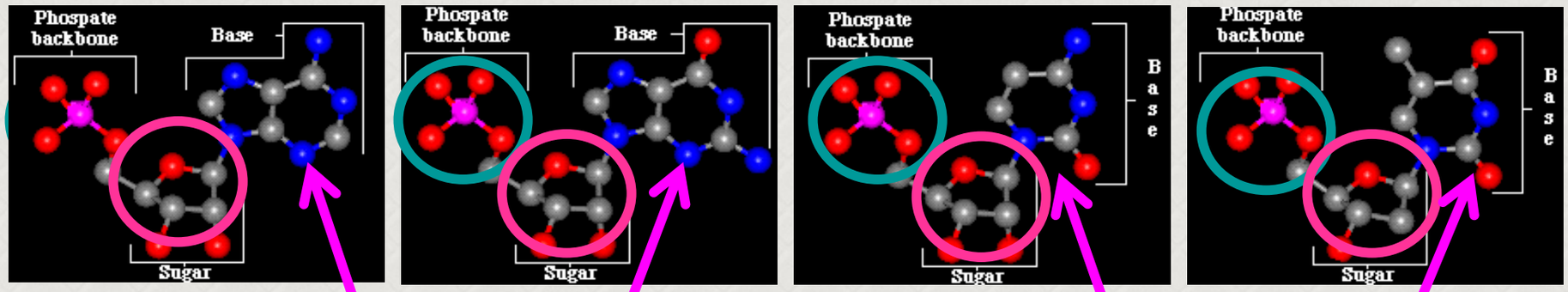
Molecular bioinformatics

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.

2 purines:
嘌呤

2 pyrimidines:
嘧啶



adenine (A) guanine (G) cytosine (C) thymine (T)

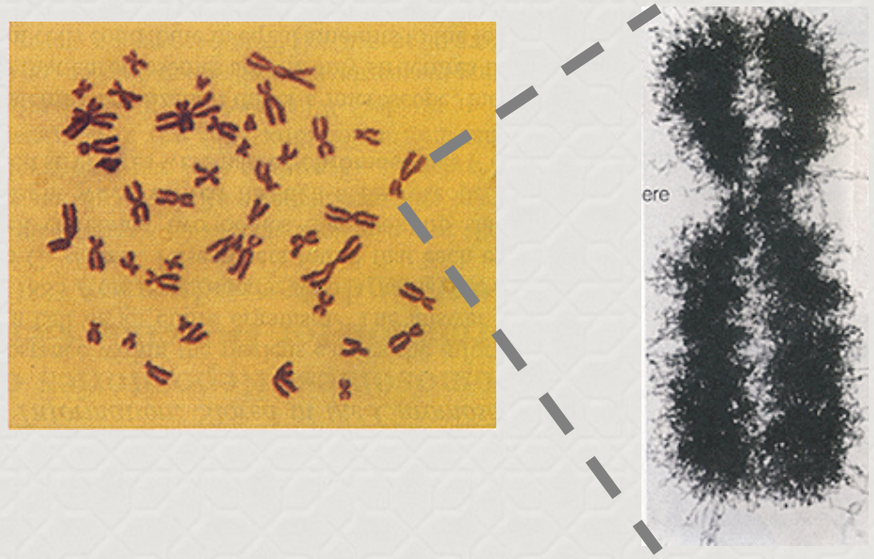
two rings

one ring

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

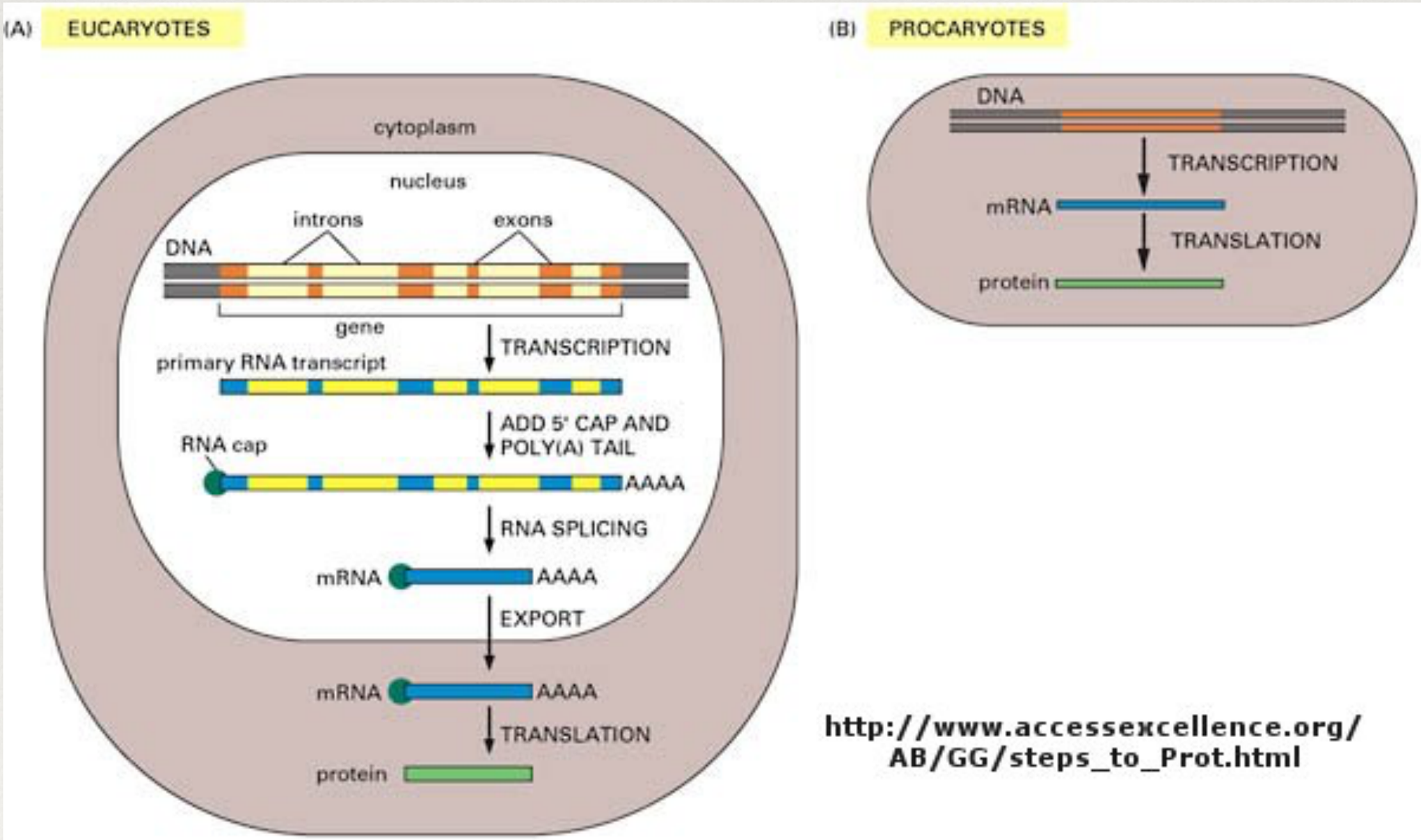
Eukaryotes may have up to 3 subcellular genomes:

1. Nuclear
2. Mitochondrial
3. Plastid



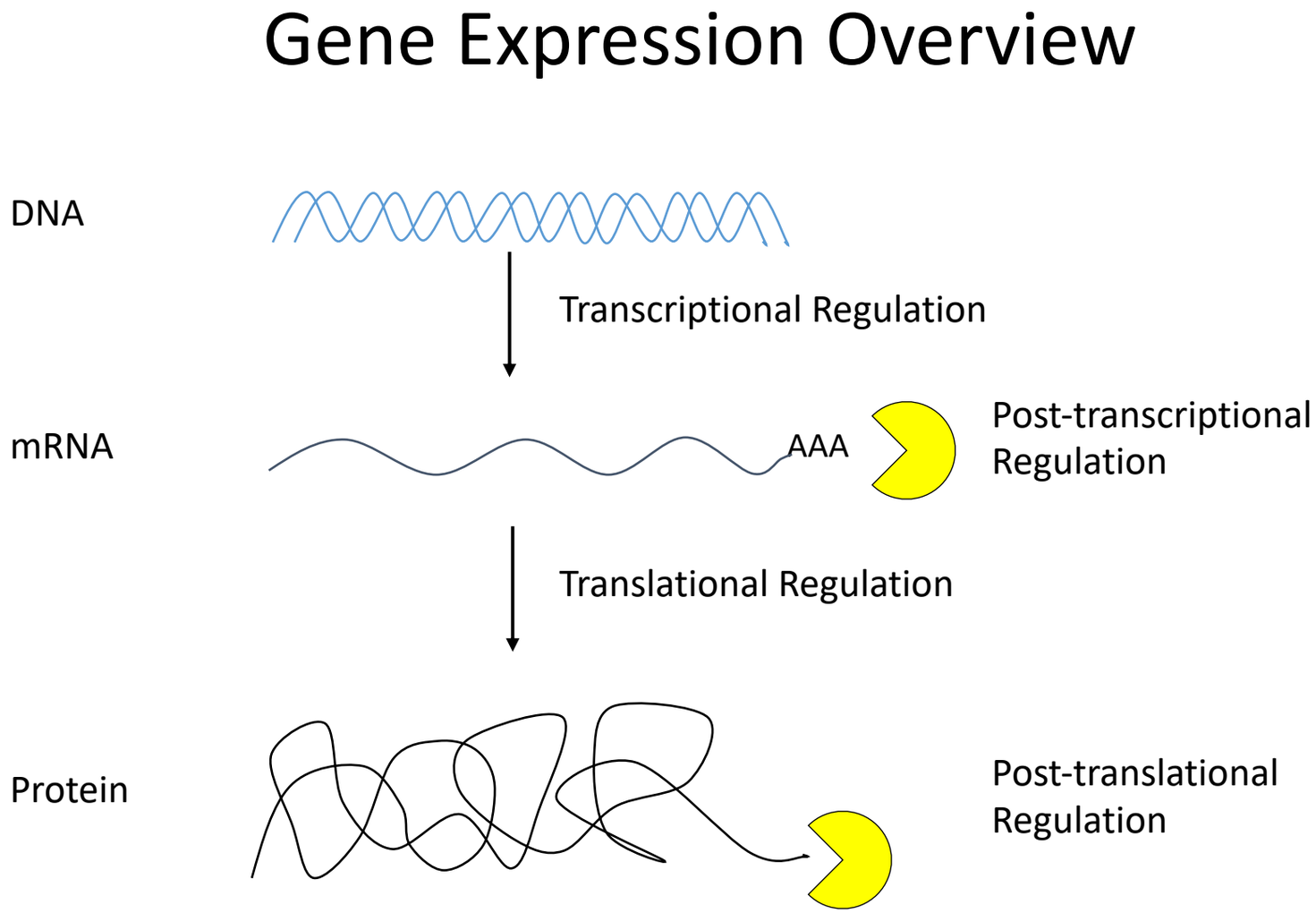
Bacteria have either circular or linear genomes and may also carry plasmids

Central dogma: DNA makes RNA makes Protein



http://www.accessexcellence.org/AB/GG/steps_to_Prot.html

Central dogma: DNA makes RNA makes Protein



Amino acids - The protein building blocks

Second letter

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

Protein folding

a human hemoglobin

How does it all looks like on a computer monitor?

a DNA sequence in fasta format

>gi|14456711|ref|NM_000558.3| **Homo sapiens hemoglobin, alpha** 1 (HBA1), mRNA
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCG
CCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACC
ACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGA
CGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACA
AGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCC
GAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCG
TTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGT
ACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC

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

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACC**ATG**GTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCC
GCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCAC
CACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCG
ACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCAC
AAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGC
CGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACC
GT**TAA**GCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCC
GTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCGGC

a protein sequence

>gi|4504347|ref|NP_000549.1| **alpha 1 globin [Homo sapiens]**

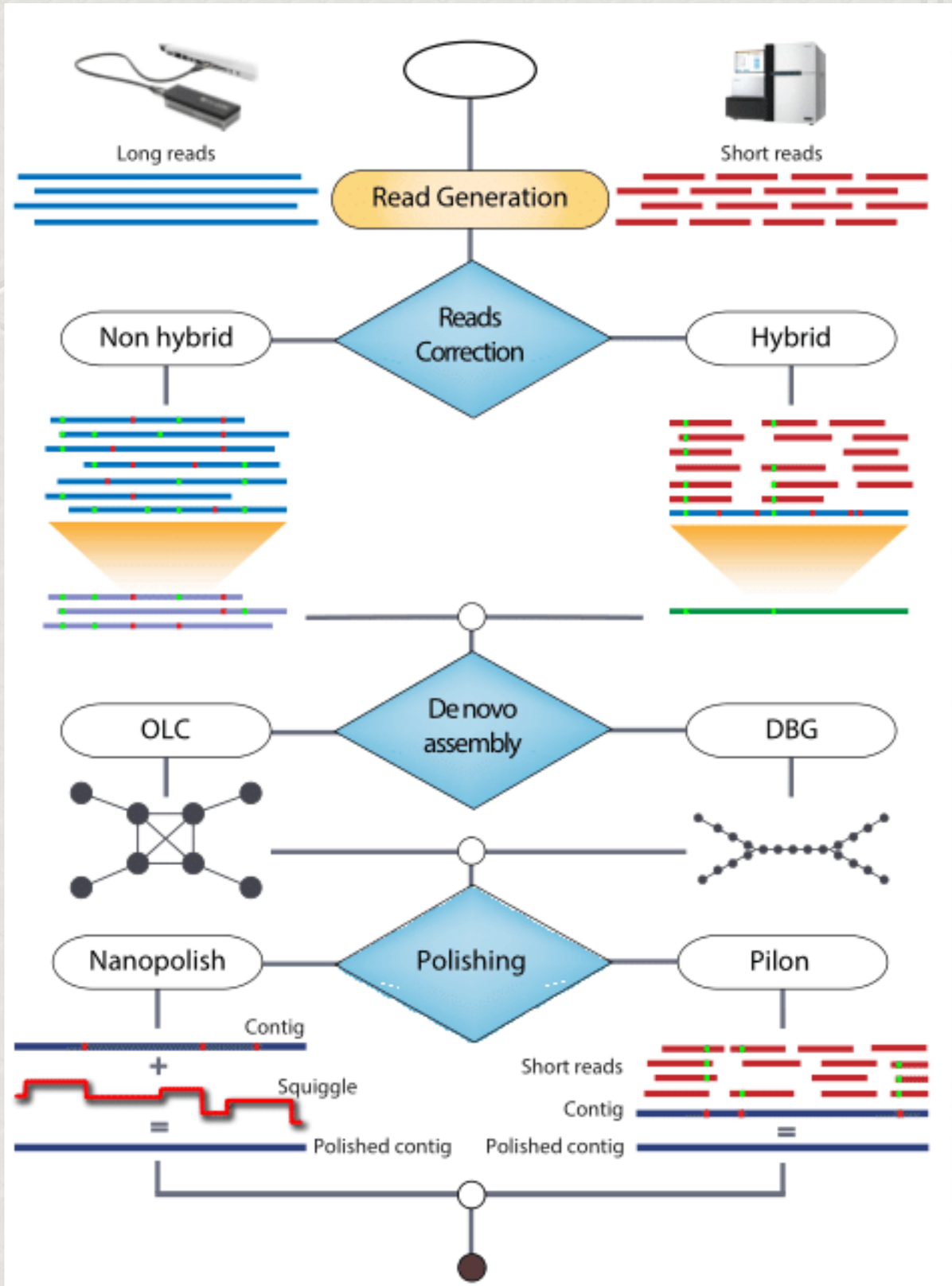
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAH
VDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

How big is a human genome?

ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGG
GGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGACCT
ACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGC
CGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTC
AACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCT
CCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTTCT
TGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGGCG
GCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCT
GGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAGAC
CTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAAC
GCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGG
TCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGC
CTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGCTT
CTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTGGG
CGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGC
CTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACCACCAAG
ACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCA
ACGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCC
GGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCAC
GCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCCATGC
TTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTGAGTG
GGCGGCGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGG
ACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGT
GCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTGGCC
ATGCTTCTTGCCCCTTGGGCCTCCCCCCAGCCCCTCCTCCCCTTCCTGCACCCGTACCCCCGTGGTCTTTGAATAAAGTCTG
AGTGGGCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAA
GGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCCACC
ACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCG...

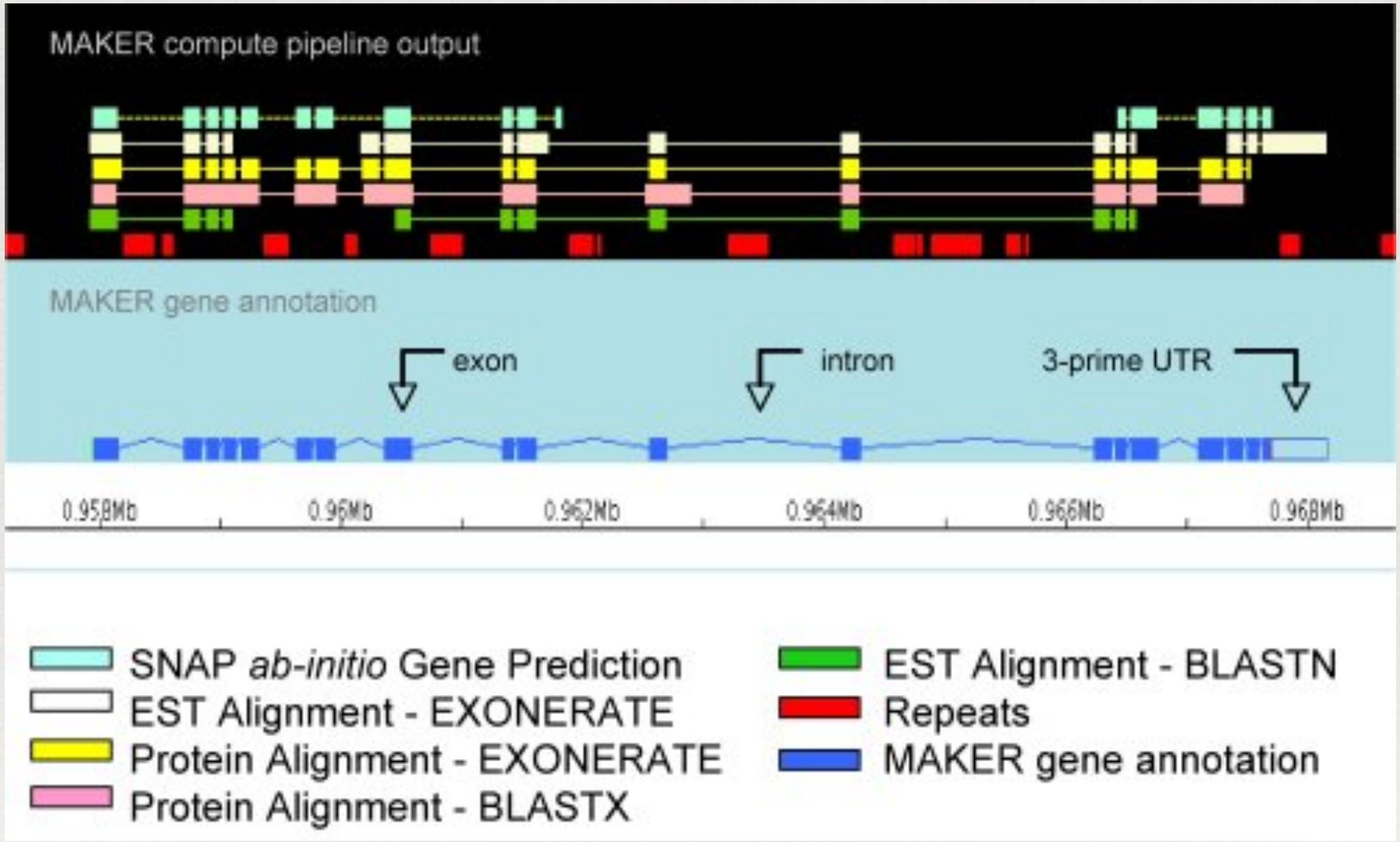
What do we actually do with bioinformatics?

Genome assembly



What do we actually do with bioinformatics?

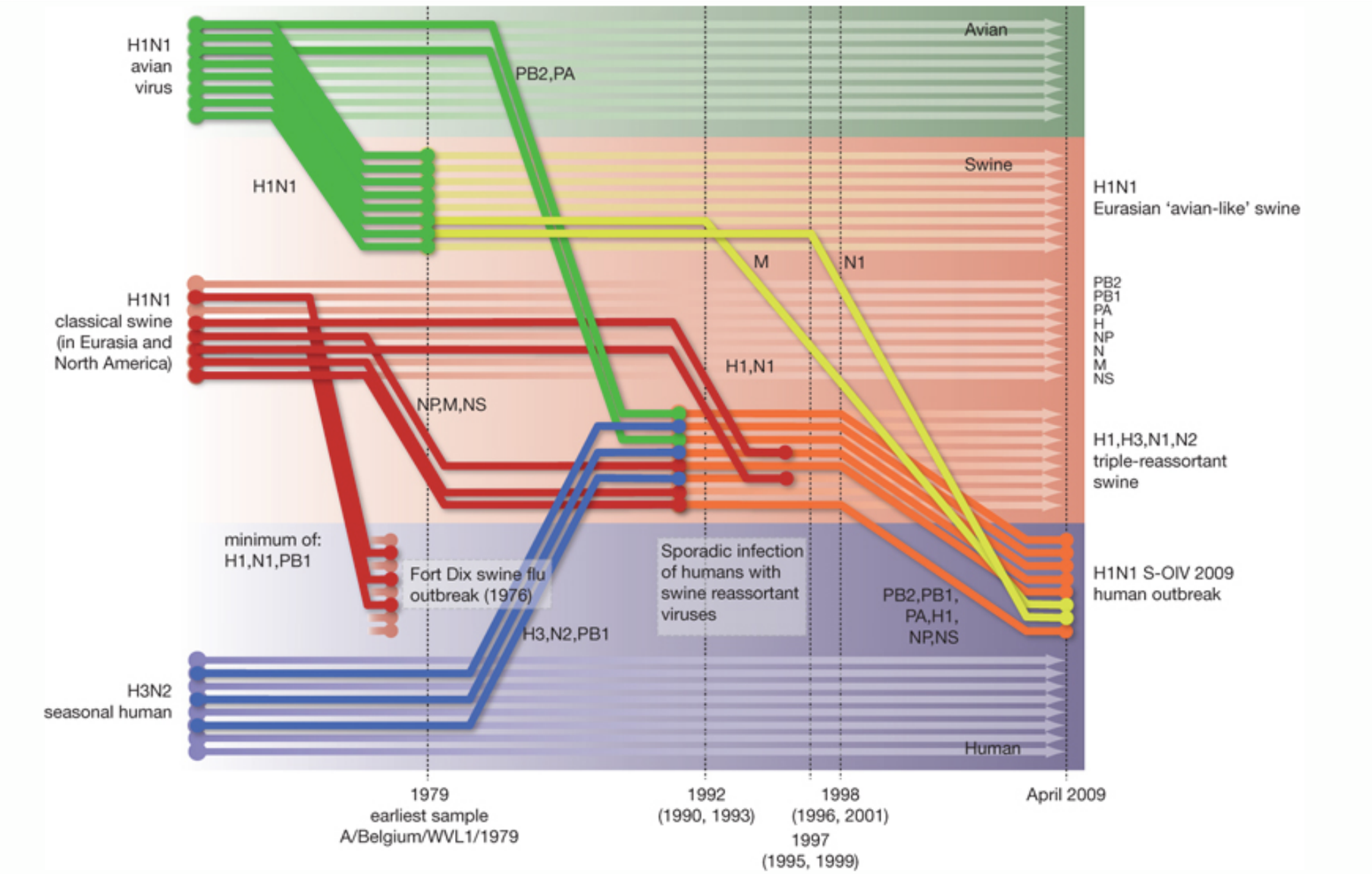
Genome annotation



What do we actually do with bioinformatics?

Molecular evolution

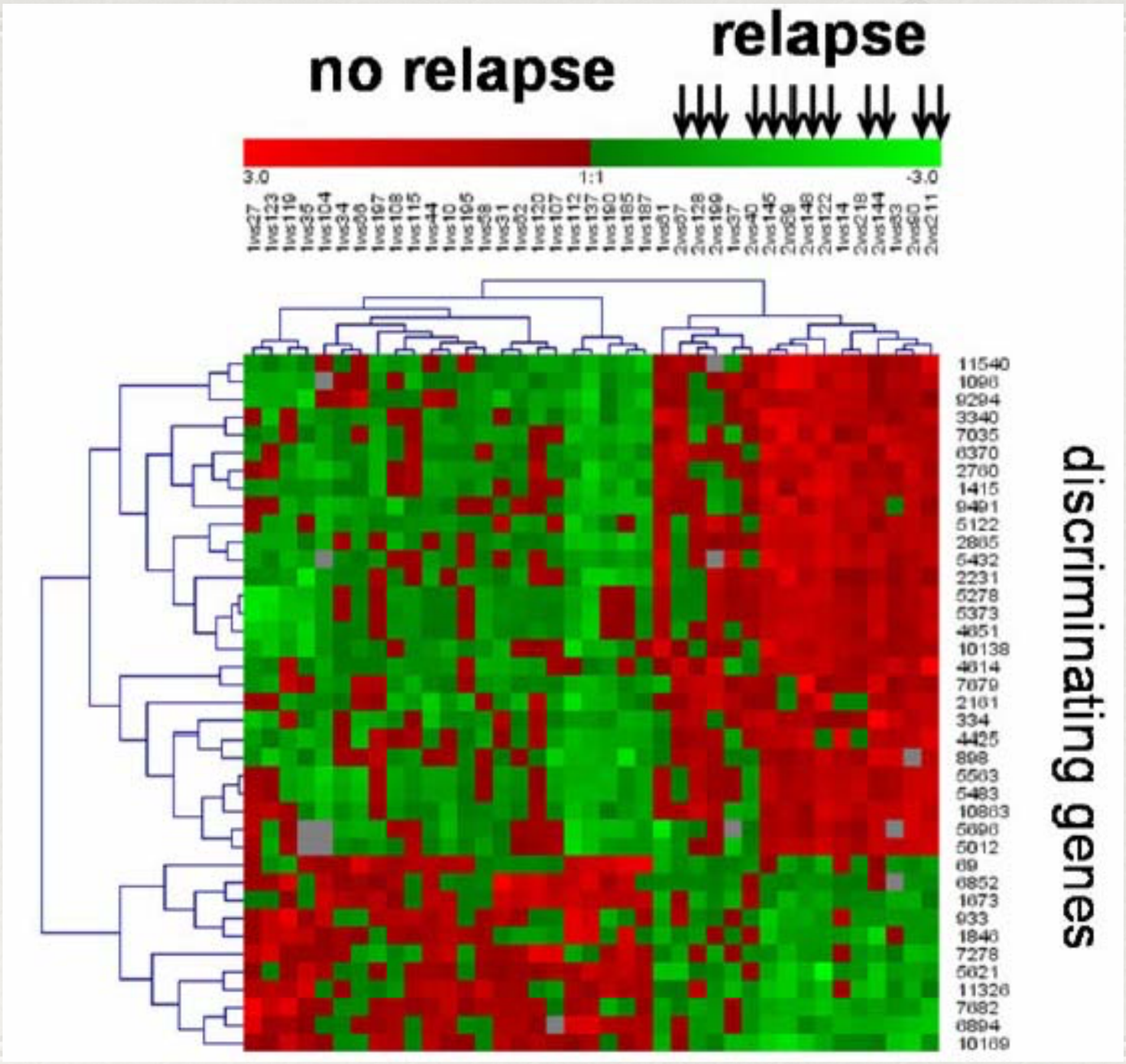
Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic



Smith et al. (2009) *Nature* ***459, 1122-1125***

What do we actually do with bioinformatics?

Gene expression analysis

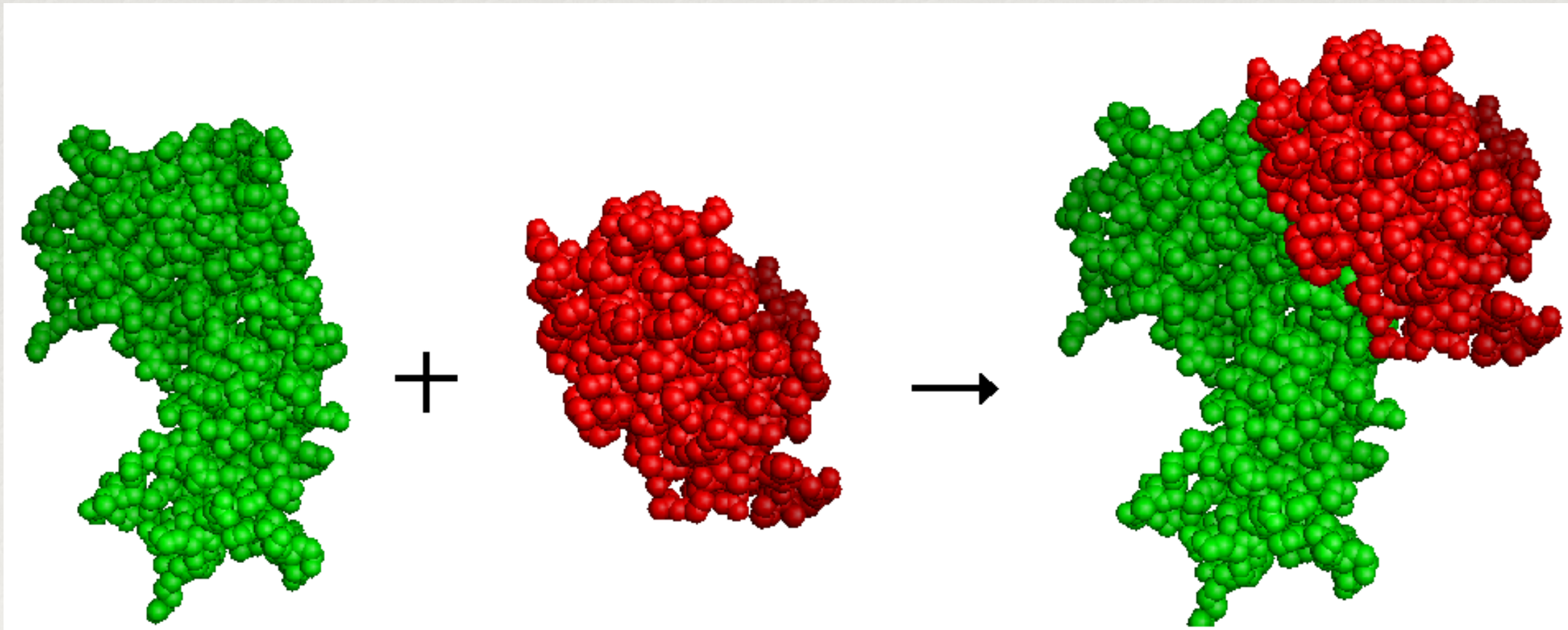


A set of 39 genes discriminates between the two classes of tumors.

What do we actually do with bioinformatics?

Protein structure prediction

Protein docking



Online Learning Resources

生信技能树：
http://www.biotrainee.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/UMTIwNTc0MjU2OA==

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