Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

1

TODAY'S TOPICS

- Course mechanics
- What is Bioinformatics?

2

Course Web page

http://bioinfolab.unl.edu/emlab/Courses/BIOS877/current/index.html

- Course description
- Course schedule
  - → Check this often! Will be regularly updated.
  - → Readings (chapters from the optional textbook, *etc.*)
  - → Lecture handouts (color PDF version)
  - → Homework assignment due dates
- Useful links (more links will be added)
- For non-biology students
  - → List of molecular biology/genetics keywords (study them!)
  - → Online review materials provided.
  - → No review session for molecular biology during the course

3

Course Home Page on Canvas

https://canvas.unl.edu

- Check Email address for your Canvas account!
- Announcements & Email for communication
- Syllabus
  - → Link to the course Web page
  - → Notes on grading, assignment submission, *etc.*
- Course Content
  - → Lecture slides (full size PDF file)
  - → Related articles (PDF)
  - → "Introduction to Bioinformatics"
    (Chapter 1 of Essential Bioinformatics by J. Xiong) [Lecture 1]
- Grades

4

Course Home Page on Canvas (continued)

https://canvas.unl.edu

- Assignments
  - → Download/submit homework assignments
- Course Survey (non-graded)
  - → Download from Assignments
  - → Submit by Jan 26 (Fri)
  - → This is to ensure that you have no problem for using Assignments page. If you have any problem, we need to fix it ASAP!!

5

Textbook (optional)

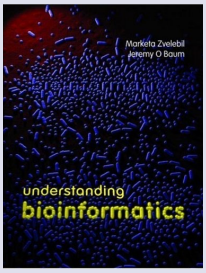
Zvelebil, M. & Baum, J. (2007)

Understanding Bioinformatics

Garland Science
ISBN: 0815340249

Book website:
https://www.routledge.com/Understanding-Bioinformatics/Zvelebil-Baum/p/book/9780815340249

- eBook Rental is also available



6

Required in this course

- You should be familiar with basic molecular biology & genetics terminology
  - → Use the keyword list and online materials to prepare yourself
- You should be able to use/access internet, Web browser, and familiar with general computer tasks
  - → To work on homework assignments
- If you do not have or access to a computer … let me know!

7

Not required in this course

- Programming skill
- Unix/Linux knowledge

But they are very useful if you want to do lots of bioinformatics/genomics analyses in the future!!

Unix/Linux/Programming online tutorials are available on the Course website (see Links page)

8

No computer lab

- It's not necessary to bring your laptop to the class
  - → It will be useful to have it if you want to follow during demonstrations in the class.
- Homework assignments require you to do some web-based bioinformatics analysis

9

Assignments

- No in-class exam 
- Multiple assignments (about every week)
  - manual computation of core methods
  - small sequence analysis
  - article review
    - → for core methods
    - → for bioinformatics research
- You need to commit your time
- Read Notes for assignment submission in the Syllabus page

10

REMEMBER THIS!

- Some assignments take time. Some require manual computation.
- Some assignments require you to use software on the Web.
  - → Web servers could become slow. They may crash.
  - → Canvas does not always cooperate. Submitting assignments may take time, especially just before deadlines!

  We experienced all of these in the past!
- Learning new programs requires time.
  - → Read manuals and online helps.

  Start working on your assignment as soon as possible. DO NOT wait until just before the deadline!

11

Grading

- Read Notes for grading in the Syllabus page
- If you are taking this course as BIOS477, your worst assignment score is excluded from calculating the final score (except for the final assignment)
  - → This means … you can safely skip one assignment 
- Late submission of an assignment will cost 5% (or more) deduction per day (indicated in each assignment)
- Graduate Student Only assignments:
  - → indicated as G1 and G2
  - → undergraduates can use them for bonus points

12

What you will learn from this course

- Core algorithms in bioinformatics
- How they work
  (If you want, you should be able to write your own program!)
- How you can interpret the results for biological significance

GOAL: You should be able to learn new bioinformatics methods by yourself in the future

Bioinformatics fields evolve fast
→ requires continuous self-learning

13

Other Bioinformatics Courses at UNL

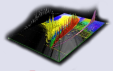
- BIOS 337 Basic Application of Bioinformatics
  → Fall; Basic hands-on bioinformatics course
- BIOS 426/826 Systems Biology
  → Fall, online; Next-gen sequence analysis, protein interaction/regulatory network

- BIOS 967 Introduction to R for Biological Sciences (Fall)
- BIOS 998 Structural Informatics and Biology (Spring)
- AGRO 820 Bioinformatics Applications in Agriculture (Fall)
- BIOC 439/839 Dynamics of Biochemical and Biological Networks (Spring)
- FDST 867: Computational Genomics for Food and Nutritional Sciences (Fall) (including microbiome analysis)
- BSEN 951: Advanced Mathematical Modeling in Biological Engineering (Fall)
- ECEN 450/850 Bioinformatics (Fall)
- ECEN 453/853 Computational and Systems Biology (Spring)
- CSCE 471/871 Computational Methods in Bioinformatics (Spring)
- CSCE 155T/890: Computer Science I: Informatics Focus (Python programming)

A list of more courses is available on the "Links" page (at the bottom)

14

Undergraduate Minor in Computational Biology & Bioinformatics



http://cbb.unl.edu/

[Core courses]

- CSCE 155T Computer Science I: Informatics Focus
- BIOS 337 Basic Application of Bioinformatics
- STAT 218 Introduction to Statistics
- CSCE 311 Data Structures and Algorithms for Informatics

+ 2 Electives
  one from Life Science → BIOS477 (this course) can be used
  one from Computer Science/Mathematics/Statistics/Engineering

15

Useful tricks (more in the Syllabus page)

- Save your sequences in the plain TEXT format
  → The default Word file (.doc or .docx) is not in the text format!
  → Use "Save as" and choose the text format.
  → Use text editors (BBEdit, Notepad++, *etc.*).
- Read the manuals and follow the input format!
- Control your computer. Don't let your computer control you! Don't just double click on files.
- DO NOT blindly believe what you find on internet (*e.g.*, Google, Wikipedia)! Lots of incorrect information. Always double-check with the original information source (journal articles).

16

Any questions?

17

What is Bioinformatics?

- Too many definitions!!
  Everybody defines it differently.
- Luscombe *et al.* (2001) (Xiong, Ch. 1, page 4)

  "An interdisciplinary field involving biology, computer science, mathematics, and statistics to analyze biological sequence data, genome content, and arrangement, and to predict the function and structure of macromolecules"
- National Cancer Institute (NCI's Dictionary of Cancer Terms)

  "A field of science that uses computers, databases, math, and statistics to collect, store, organize, and analyze large amounts of biological, medical, and health information. Information may come from many sources, including genetic and molecular research studies, patient statistics, tissue specimens, clinical trials, and scientific journals. Also called computational biology."

18

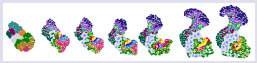
Who coined the term "Bioinformatics"?

- Paulien Hogeweg (in early 1970s)

A Dutch theoretical biologist, who started using the term "Bioinformatics" to refer to "the study of informatic process in biotic systems".

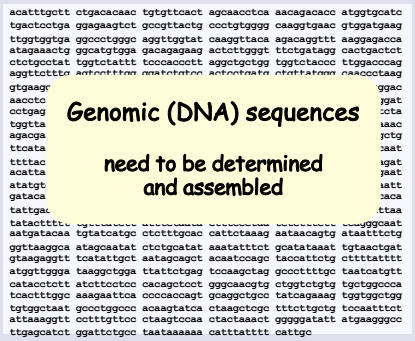
Hogeweg (2011) "The root of bioinformatics in theoretical biology." *PLoS Comp Biol* 7: e1002021. [Available on Canvas/References]

Dr. Hogeweg's lab website: https://www-binf.bio.uu.nl/ph/



19

What is the information in Bioinformatics?



20

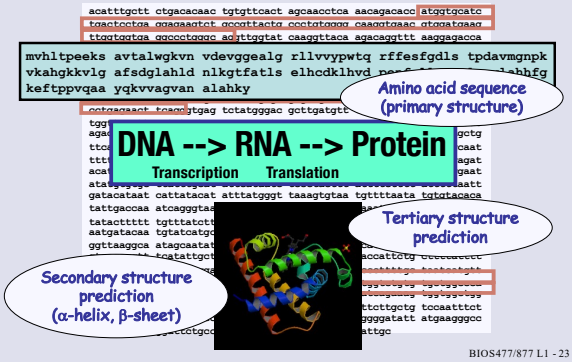
What is the information in Bioinformatics?

21

What is the information in Bioinformatics?

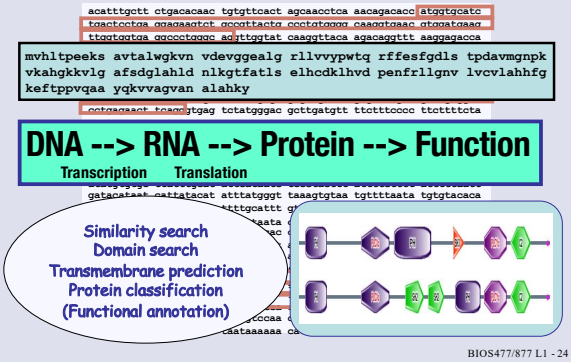
22

What is the information in Bioinformatics?



23

What is the information in Bioinformatics?



24

How do we analyze biological sequences?

Single sequence analysis

```
mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg
keftppvqaa yqkvvagvan alahky
```

Gene prediction (DNA)
Gene expression (DNA)
Primary structure prediction (from DNA to protein)
Transmembrane region prediction (protein)
Secondary structure prediction (protein)

25

How do we analyze biological sequences?

Multiple sequence analysis

```
mvhltpeeks avtalwgkvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg
keftppvqaa yqkvvagvan alahky
```

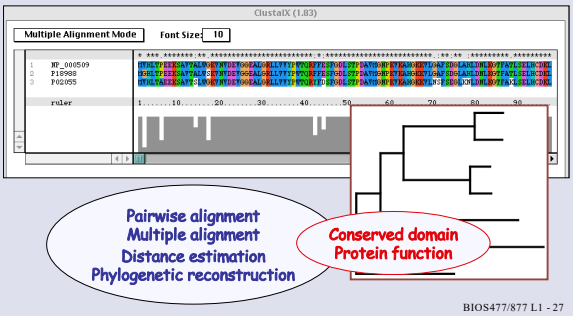
```
mghltpeeks avtalwskvn vdevggealg rllvvypwtq rffesfgdls tpdavmgnpk
vkahgkkvlg afsdglahld nlkgtfatls elhcdklhvd penfrllgnv lvcvlahhfg
keftppvqaa yqkvvagvan alahkyh
```

```
mvhltaeeks avtslwgkvn vdevggealg rllvvypwtq ryfdsfgdls tpdavmgnpk
vkahgkkvln sfseglknld nlkgtfakls elhcdklhvd penfkllgnv lvcvlahhfg
keftpqvqaa yqkvvagvan alahkyh
```


26

How do we analyze biological sequences?

Multiple sequence analysis



Pairwise alignment
Multiple alignment
Distance estimation
Phylogenetic reconstruction

Conserved domain
Protein function

27

More Topics in Bioinformatics

- Large-scale fragment/contig assembly
  - ➜ whole genome shotgun sequencing
- Next-gen sequencing data analysis
- RNA-seq (gene expression) analysis
- Metagenomic/transcriptomic analysis
- RNA secondary structure prediction
  - ➜ rRNA, tRNA, small RNA, miRNA, RNA viruses, *etc.*
- Tertiary structure prediction (*ab initio*)

These topics are not covered in this course ...

28

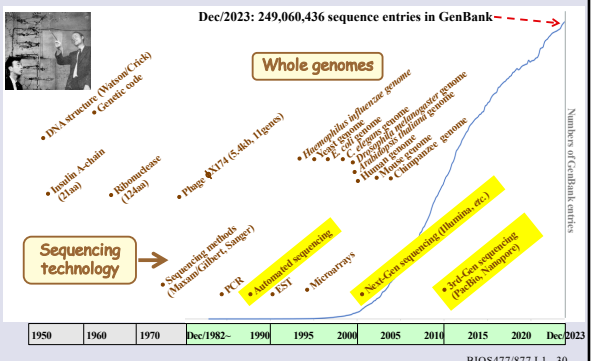
Limitations of Bioinformatics

(Read Xiong Ch1 page 7)

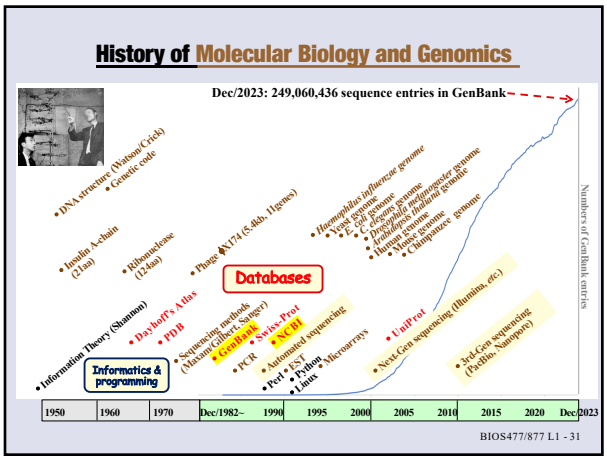
- Bioinformatics and experimental biology are complementary.
- Raw data are produced by experimental science.
  - ➜ The quality of bioinformatics predictions depends on the quality of data.
- Bioinformatics predictions are not formal proofs.
  - ➜ They need to be tested experimentally.
- Bioinformatics provides interpretation of experimental data.
  - ➜ Facilitates generating a hypothesis for further experimental research.

29

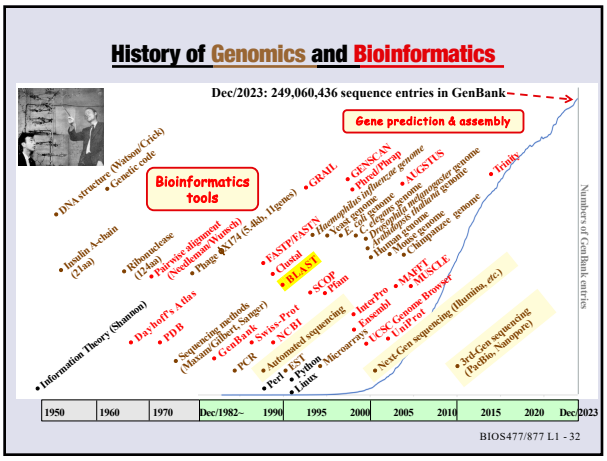
History of Molecular Biology and Genomics



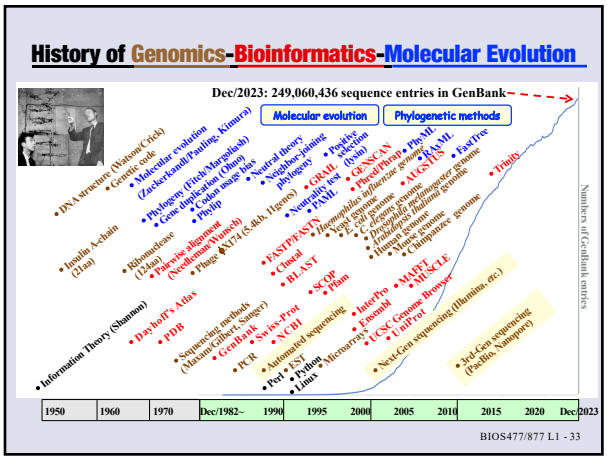
30



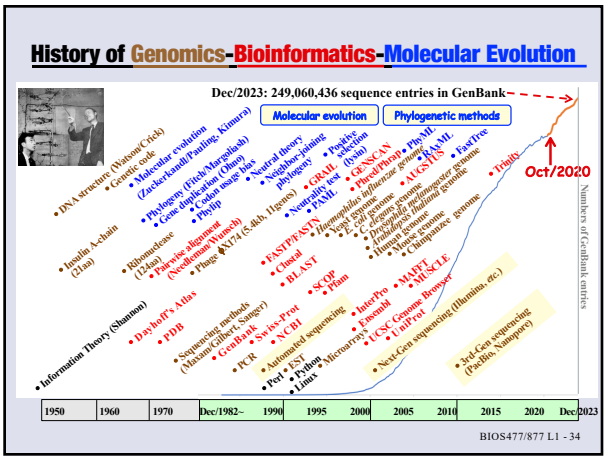
31



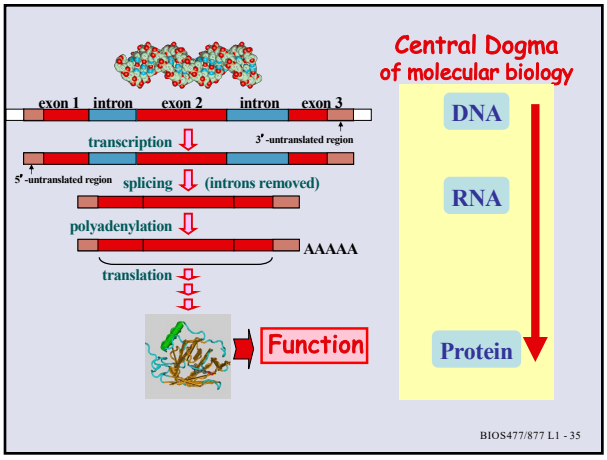
32



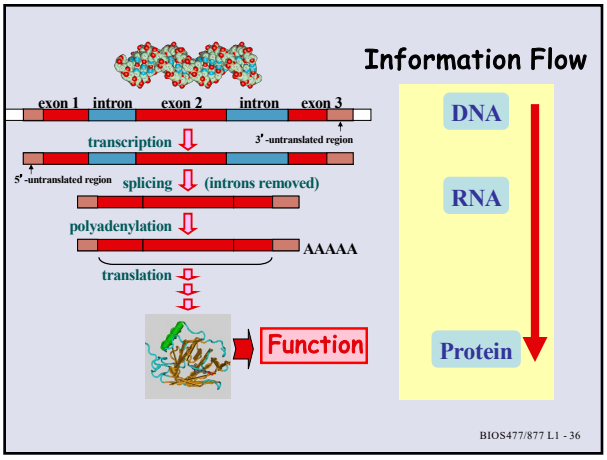
33



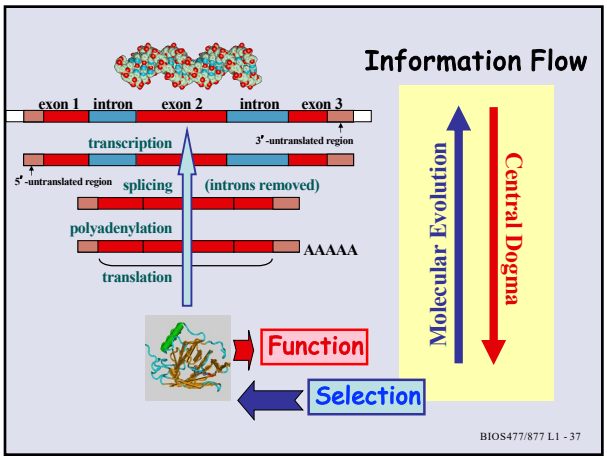
34



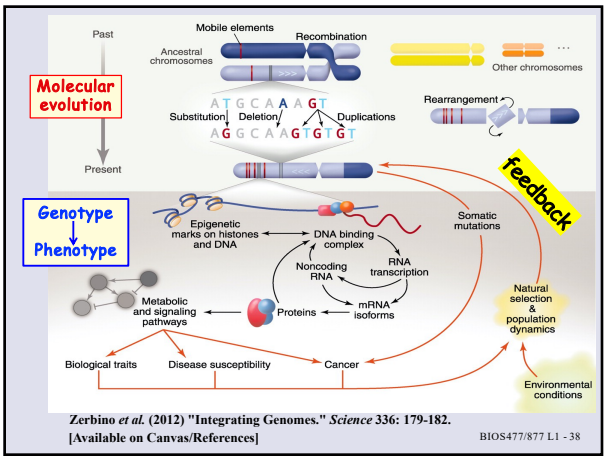
35



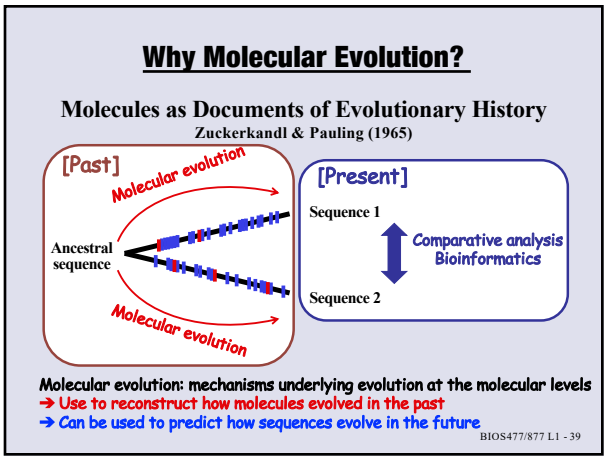
36


Information Flow
exon 1
intron
exon 2
intron
exon 3
transcription
3' -untranslated region
5' -untranslated region
splicing
(introns removed)
polyadenylation
AAAAA
translation
Function
Selection
Molecular Evolution
Central Dogma
BIOS477/877 L1 - 37

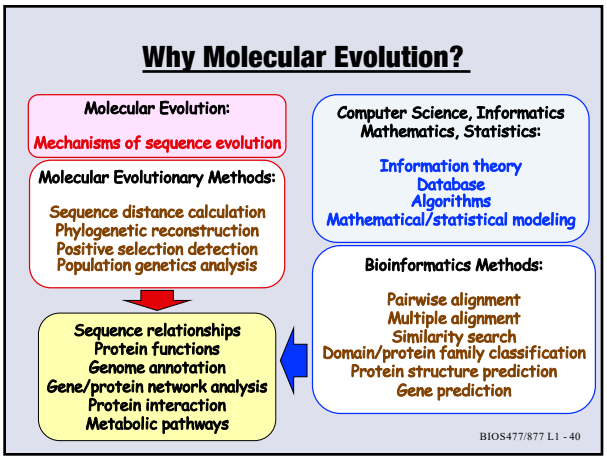
37


Past
Present
Molecular evolution
Mobile elements
Recombination
Ancestral chromosomes
Other chromosomes
Substitution
Deletion
Duplications
Rearrangement
Genotype
Phenotype
Feedback
Epigenetic marks on histones and DNA
DNA binding complex
Somatic mutations
Noncoding RNA
RNA transcription
Metabolic and signaling pathways
Proteins
mRNA isoforms
Natural selection & population dynamics
Biological traits
Disease susceptibility
Cancer
Environmental conditions
Zerbino et al. (2012) "Integrating Genomes." Science 336: 179-182.
[Available on Canvas/References]
BIOS477/877 L1 - 38

38


Why Molecular Evolution?
Molecules as Documents of Evolutionary History
Zuckerkandl & Pauling (1965)
[Past]
Molecular evolution
Ancestral sequence
Molecular evolution
[Present]
Sequence 1
Comparative analysis
Bioinformatics
Sequence 2
Molecular evolution: mechanisms underlying evolution at the molecular levels
➔ Use to reconstruct how molecules evolved in the past
➔ Can be used to predict how sequences evolve in the future
BIOS477/877 L1 - 39

39


Why Molecular Evolution?
Molecular Evolution:
Mechanisms of sequence evolution
Molecular Evolutionary Methods:
Sequence distance calculation
Phylogenetic reconstruction
Positive selection detection
Population genetics analysis
Computer Science, Informatics Mathematics, Statistics:
Information theory
Database
Algorithms
Mathematical/statistical modeling
Bioinformatics Methods:
Pairwise alignment
Multiple alignment
Similarity search
Domain/protein family classification
Protein structure prediction
Gene prediction
Sequence relationships
Protein functions
Genome annotation
Gene/protein network analysis
Protein interaction
Metabolic pathways
BIOS477/877 L1 - 40

40