Lecture 2: Biopython Tutorial

Case Studies in Biological Sequence Analysis

Yen-Yi Ho

Department of Statistics

Biopython: Python for Genomic Data Scinece

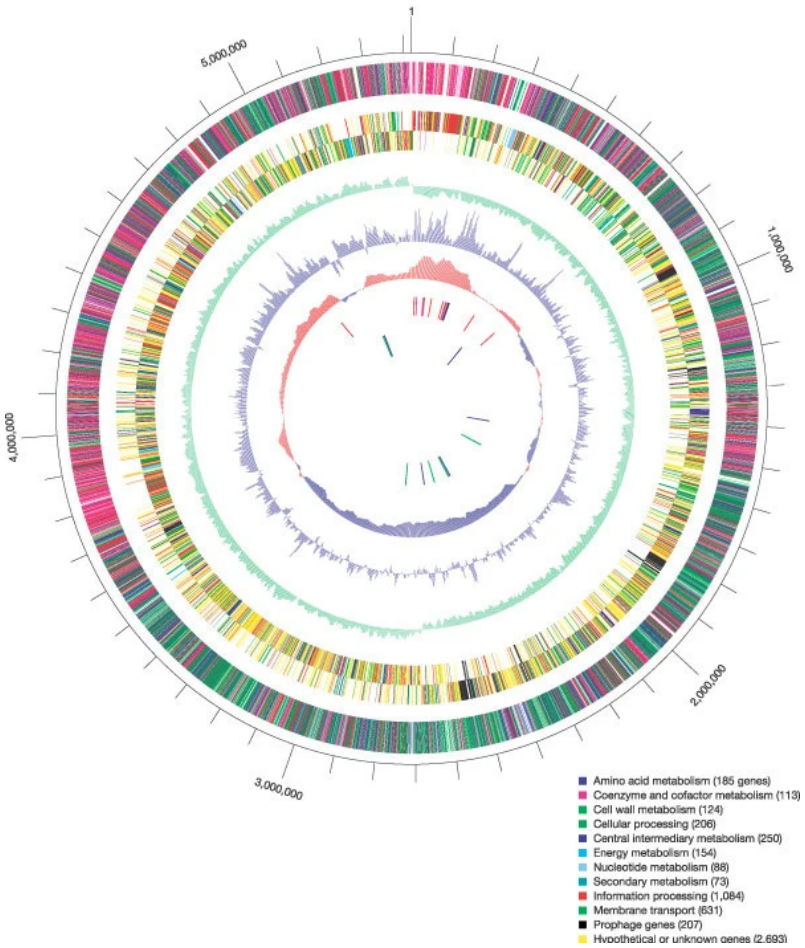

Figure adapted from: Ivanova, N., Sorokin, A., Anderson, I. et al. Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature 423, 87–91 (2003). https://doi.org/10.1038/nature01582

The Biopython Project

The Biopython Project founded in 1999 is an international, open-source collaboration dedicated to developing Python-based tools for computational biology and bioinformatics.
The goal of Biopython is to make it as easy as possible to use Python for bioinformatics.

- https://biopython.org/: an online resource for modules and scripts for bioinformatics use and research.
- Biopython includes parsers for various bioinformatics file formats (such as FASTA, GenBank), access to online services like NCBI Entrez or PubMed databases, interfaces to common bioinformatics programs such as BLAST, ClustalW, and others.

Installing Biopython

- Runs on many platforms: Windows, Mac, and on the various flavors of Linux and Unix
- supports both Python 2 and Python 3
- In a terminal

```
pip install biopython
```

Checking if Biopython Is Installed

```
import Bio
print(Bio.__version__)
```

A Biopython Usage Simple Example

Problem: Find out from what species an unknown DNA sequence came from.

myseq.fa

```
>sequence_unknown
CATGCTACGGTGCTAAAAGCATTACGCCCTATAGTGATTTTCGAGACATACTGTGTTT
TTAAATATAGTATTGCC
```

Motif Finding with Biopython

Outline

- DNA sequence motif
- Understand what a PSSM is
- Construct a PSSM step by step
- Understand the role of pseudocounts
- Learn how sequences are scored using a PSSM

Input Sequences

A DNA sequence motif is a short, recurring pattern of nucleotides (typically 6–12 base pairs) found within DNA sequences that possesses a specific biological function.
Aligned examples of the same biological motif:

```
ATGCA
ATGGA
ATGTA
ATGCA
ATGCC
```

- Number of sequences: 5
- Length of motif: 5

Step 1: Position Frequency Matrix (PFM)

Count how many times each base appears at each position.

Position	A	C	G	T
1	5	0	0	0
2	0	0	0	5
3	0	0	5	0
4	0	3	1	1
5	4	1	0	0

PFM contains **raw counts**.

Why Pseudocounts Are Needed

Problem:

- Zero counts lead to zero probabilities
- Log-odds score becomes $-\infty$ (ex: $\log_2 \frac{0}{5}$)

Solution:

- Add a small value (pseudocount) to each cell
- Prevents overconfidence from small samples

Step 2: Add Pseudocounts

Add pseudocount $\alpha = 0.5$ to every entry.

Position	A	C	G	T
1	5.5	0.5	0.5	0.5
2	0.5	0.5	0.5	5.5
3	0.5	0.5	5.5	0.5
4	0.5	3.5	1.5	1.5
5	4.5	1.5	0.5	0.5

Each **row (position)** now sums to 7.0.

Step 3: Position Probability Matrix (PPM)

Convert counts to probabilities:

$$PPM_{i,j} = \frac{count_{i,j}}{\sum_{k \in \{A,C,G,T\}} count_{i,k}}$$

Position	A	C	G	T
1	0.786	0.071	0.071	0.071
2	0.071	0.071	0.071	0.786
3	0.071	0.071	0.786	0.071
4	0.071	0.500	0.214	0.214
5	0.643	0.214	0.071	0.071

Each **row (position)** sums to 1.

Step 4: Background Frequencies

Assume uniform background distribution:

$$b(A) = b(C) = b(G) = b(T) = 0.25$$

Background frequencies represent random expectation.

PSSM Scoring Function

Position-Specific Scoring Matrix (PSSM) or
Position Weight Matrix (PWM)

$$\text{Score} = \sum_{i=1}^{L} \log_2 \left(\frac{P(\text{base}_i \mid \text{motif})}{P(\text{base}_i \mid \text{background})} \right)$$

"How much more likely is this sequence under the motif model than under a random background model?"

Step 5: Compute the PSSM

PSSM is a log-odds matrix:

$$PSSM_{i,j} = \log_2 \left(\frac{PPM_{i,j}}{b_j} \right)$$

- Positive score: favored over background
- Negative score: disfavored

Final PSSM (Log-Odds Scores)

Position	A	C	G	T
1	1.65	-1.81	-1.81	-1.81
2	-1.81	-1.81	-1.81	1.65
3	-1.81	-1.81	1.65	-1.81
4	-1.81	1.00	-0.22	-0.22
5	1.36	-0.22	-1.81	-1.81

Step 6: Scoring a Sequence

Score sequence `ATGCA` by summing PSSM values:

- Position 1 (A): $+1.65$
- Position 2 (T): $+1.65$
- Position 3 (G): $+1.65$
- Position 4 (C): $+1.00$
- Position 5 (A): $+1.36$

$$\text{Total score} = 7.32$$

How to Interpret PSSM Scores

- Higher total score = better motif match
- Threshold determines sensitivity vs specificity
- Used for genome scanning and motif discovery

Key Takeaways

- PSSM is a position-specific log-odds scoring model
- Built from aligned sequences
- Pseudocounts prevent zero probabilities
- Widely used in motif scanning and alignment

Drawbacks of Biopython PSSM Method

- Assumes **independence between positions**
  - Cannot model interactions between bases
- Represents a **single motif** only
  - No modeling of motif combinations or spacing constraints
- Requires **pre-aligned motif instances**
  - Poor performance on unaligned or noisy data
- Limited predictive power
  - Descriptive rather than quantitative binding prediction
- Uses **simple background models**
  - Often assumes uniform base frequencies
- Cannot capture **long-range dependencies**
  - Motif length typically short (< 20 bp)

More Help with Biopython

- Biopython Tutorial and Cookbook: `https://biopython.org/docs/latest/Tutorial/index.html`
- Biopython FAQ: `https://biopython.org/wiki/The_Biopython_Structural_Bioinformatics_FAQ`