Introduction to biopython

Mathilde Bonnemaison
June 2nd
Boston Python Meetup

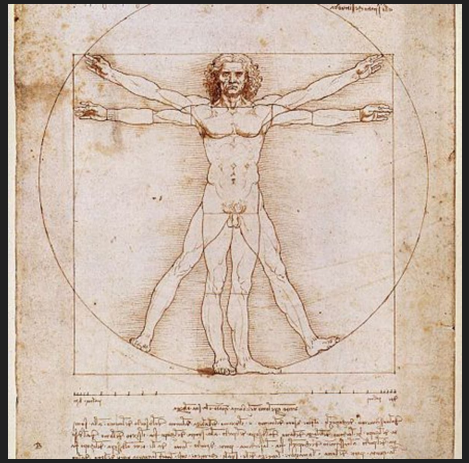
What is Biopython?

- Biopython is a set of libraries providing bioinformatic tools for biologists.
- Biopython website: https://biopython.org/
  - Instructions to download, install & get started
  - Open-source project
- Examples of Biopython functionalities:
  - Sequence handling
  - 3D structure
  - Population Genetics

What is insulin?

- Insulin instructs cells throughout the body to take up sugar in the blood
- No insulin = diabetes
- 1920’s - 2000’s: Diabetic patients used porcine insulin

Human insulin vs. Porcine insulin

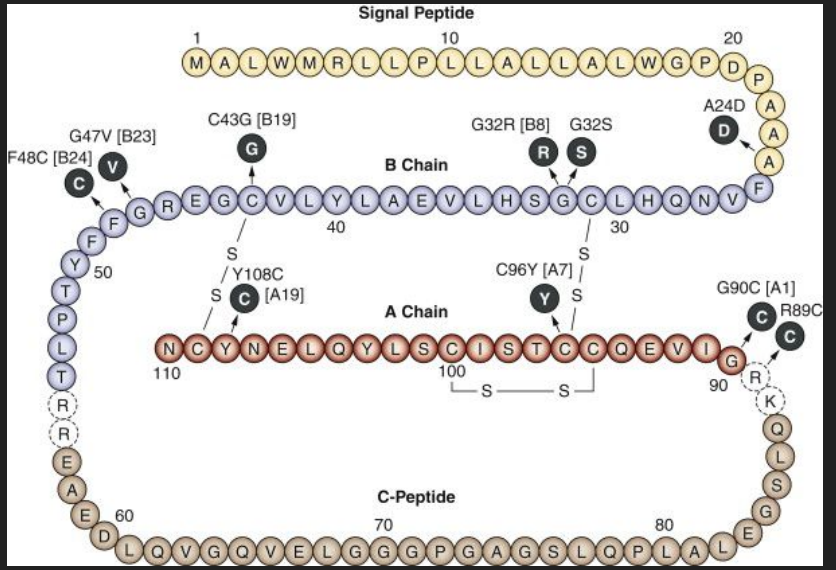


How come this treatment worked?

We’re going to see an explanation using biopython

How are proteins represented?

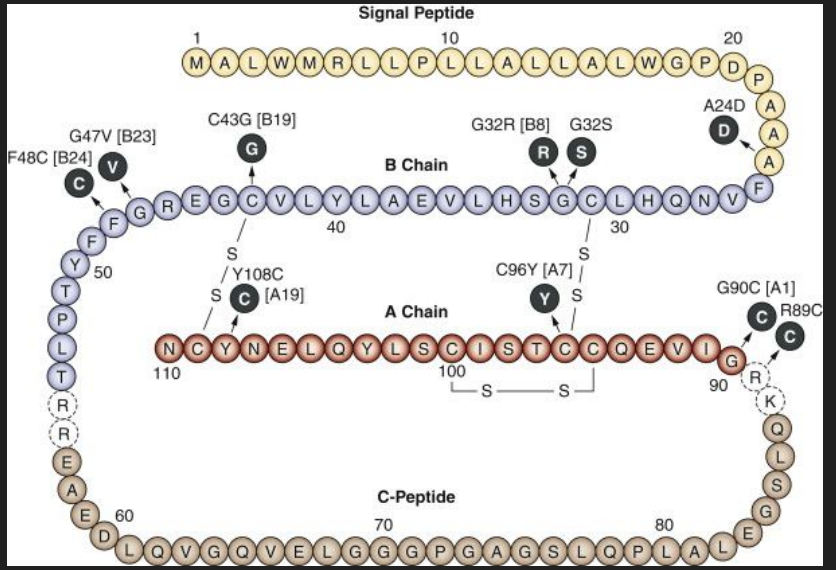
- A protein is like a pearl necklace where each pearl is an amino acid.
- Each amino acid can be represented with a letter.



How are proteins represented?

- A protein is like a pearl necklace where each pearl is an amino acid.
- Each amino acid can be represented with a letter.

MALWMRLLPLLALLALWGPDPAAAFVNQHLC
GSHLVEALYLVCGERGFFYTPKTRREAEDLQ
VGQVELGGGPGAGSLQPLALEGSLQKRGIVE
QCCTSICSLYQLENYCN

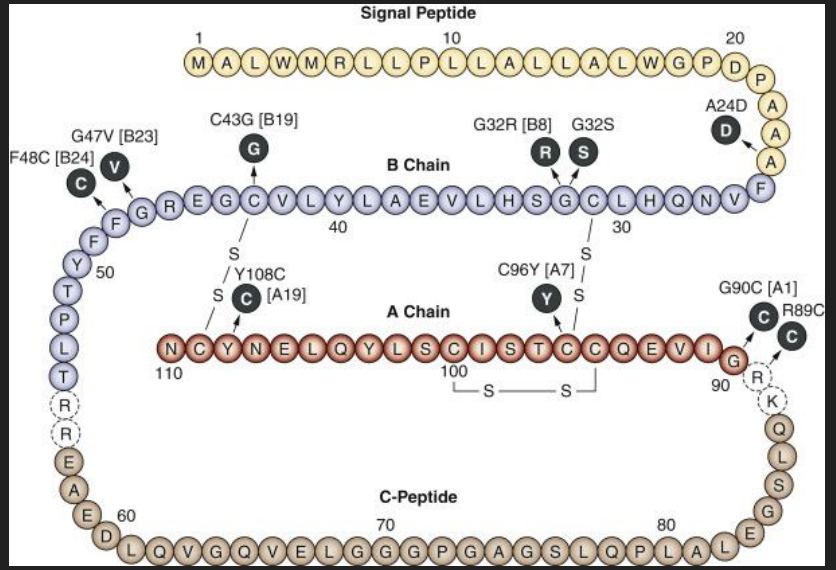


How are proteins represented?

- A protein is like a pearl necklace where each pearl is an amino acid.
- Each amino acid can be represented with a letter.

MALWMRLLPLLALLALWGPDPAAAFVNQHLC
GSHLVEALYLVCGERGFFYTPKTRREAEDLQ
VGQVELGGGPGAGSLQPLALEGSLQKRGIVE
QCCTSICSLYQLENYCN

- Biopython represents sequences of DNA, RNA and proteins using the Seq class.




```
[1]: from Bio.Seq import Seq
     protein = Seq('MALWMRLLPLLALLALWGPDPAAAFVNQHLC')

[2]: print(protein)

[2]: Seq('MALWMRLLPLLALLALWGPDPAAAFVNQHLC')

[3]: type(protein)

[3]: Bio.Seq.Seq
```

FASTA files in biopython

Protein sequences are stored in FASTA files on the NCBI website.

FASTA files contain more information than just the protein sequence.

Example: human insulin fasta file

```
>AAA59172.1 insulin [Homo sapiens]
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

FASTA files in biopython

The `SeqIO.read()` function allows you to read the fasta file:

```
[11]: from Bio import SeqIO
      human = SeqIO.read("insulin_human.fa", "fasta")

[12]: print(human)

      ID: AAA59172.1
      Name: AAA59172.1
      Description: AAA59172.1 insulin [Homo sapiens]
      Number of features: 0
      Seq('MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKT...YCN')

[15]: type(human)

[15]: Bio.SeqRecord.SeqRecord

[13]: print(human.seq)

      MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN

[14]: print(human.id)

      AAA59172.1
```

Compare human and porcine insulin sequences

Compare human and porcine insulin sequences

```
from Bio import pairwise2
human = SeqIO.read('insulin_human.fa', 'fasta')
pig = SeqIO.read('insulin_pig.fa', 'fasta')
alignments = pairwise2.align.globalxx(human.seq, pig.seq)
```

Compare human and porcine insulin sequences

```
from Bio import pairwise2
human = SeqIO.read('insulin_human.fa', 'fasta')
pig = SeqIO.read('insulin_pig.fa', 'fasta')
alignments = pairwise2.align.globalxx(human.seq, pig.seq)
```

For a nice printout, use the `format_alignment` method of the pairwise2 module:

```
from Bio.pairwise2 import format_alignment
print(format_alignment(*alignments[0]))
```

```
MALWM-RLLPLLALLALWGPD-PA-A-AFVNQHLCGSHLVEALYLVCGERGFFYTPKT-RREAEDL--QV-GQ-VELGGGP-GAGSLQP-LALEGSL--QKRGIVEQCCTSICSLYQLENYCN
||||  ||||||||||||    || | ||||||||||||||||||||||||||||||  |||||    |  |  ||||||  | | ||  |||||    |||||||||||||||||||||||
MALW-TRLLPLLALLALW---APAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPK-ARREAE--NPQ-AG-AVELGGG-LG-G-LQ-ALALEG--PPQKRGIVEQCCTSICSLYQLENYCN
  Score=95
```

Compare human and porcine insulin sequences

```
from Bio import pairwise2
human = SeqIO.read('insulin_human.fa', 'fasta')
pig = SeqIO.read('insulin_pig.fa', 'fasta')
alignments = pairwise2.align.globalxx(human.seq, pig.seq)
```

For a nice printout, use the `format_alignment` method of the pairwise2 module:

```
from Bio.pairwise2 import format_alignment
print(format_alignment(*alignments[0]))
```

```
MALWM-RLLPLLALLALWGPD-PA-A-AFVNQHLCGSHLVEALYLVCGERGFFYTPKT-RREAEDL--QV-GQ-VELGGGP-GAGSLQP-LALEGSL--QKRGIVEQCCTSICSLYQLENYCN
||||  ||||||||||||    || | ||||||||||||||||||||||||||||||  |||||    |  |  ||||||  | | ||  |||||    ||||||||||||||||||||||||
MALW-TRLLPLLALLALW---APAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPK-ARREAE--NPQ-AG-AVELGGG-LG-G-LQ-ALALEG--PPQKRGIVEQCCTSICSLYQLENYCN
  Score=95
```

Human and porcine insulin sequences are very similar explaining why treating diabetic patients with porcine insulin worked.

Biopython can do a lot more...

- Convert a DNA sequence into an RNA or protein sequence
- BLAST: compare a sequence to a database
- 3D representation of proteins, nucleic acids
- Phylogenetic trees