Review

A Few General Skills

- Design a good scoring function
  - Log likelihood ratio
  - Evaluate the significance (Bayesian or p-value)
- Dealing with noisy data
  - FDR
- Dynamic programming
  - Sequence, tree, set, mass …
- Trade off speed and accuracy/sensitivity
  - Filtration (Spaced seed)
- Useful models/data structure
  - HMM.
  - Suffix tree & array.
  - Information distance.

Scoring Function

- Design a good scoring function
  - Log likelihood ratio
  - Estimate foreground and background probability (BLOSUM)
  - Relative entropy
  - Evaluate the significance (Bayesian, p-value, E-value)
  - Machine learning

Dealing with Noisy Data

- Result validation
  - Optimality doesn’t mean reality.
  - Should throw away garbage results.
  - FDR

Dynamic Programming

- Build optimal solution from the optimal solution of a smaller sub-problem.
- A partial order is needed on the concerned sub-problems.
  - Sequence (alignment, HMM)
  - Tree (ancestor reconstruction)
  - Set (spaced seed sensitivity)
  - Mass (de novo sequencing)

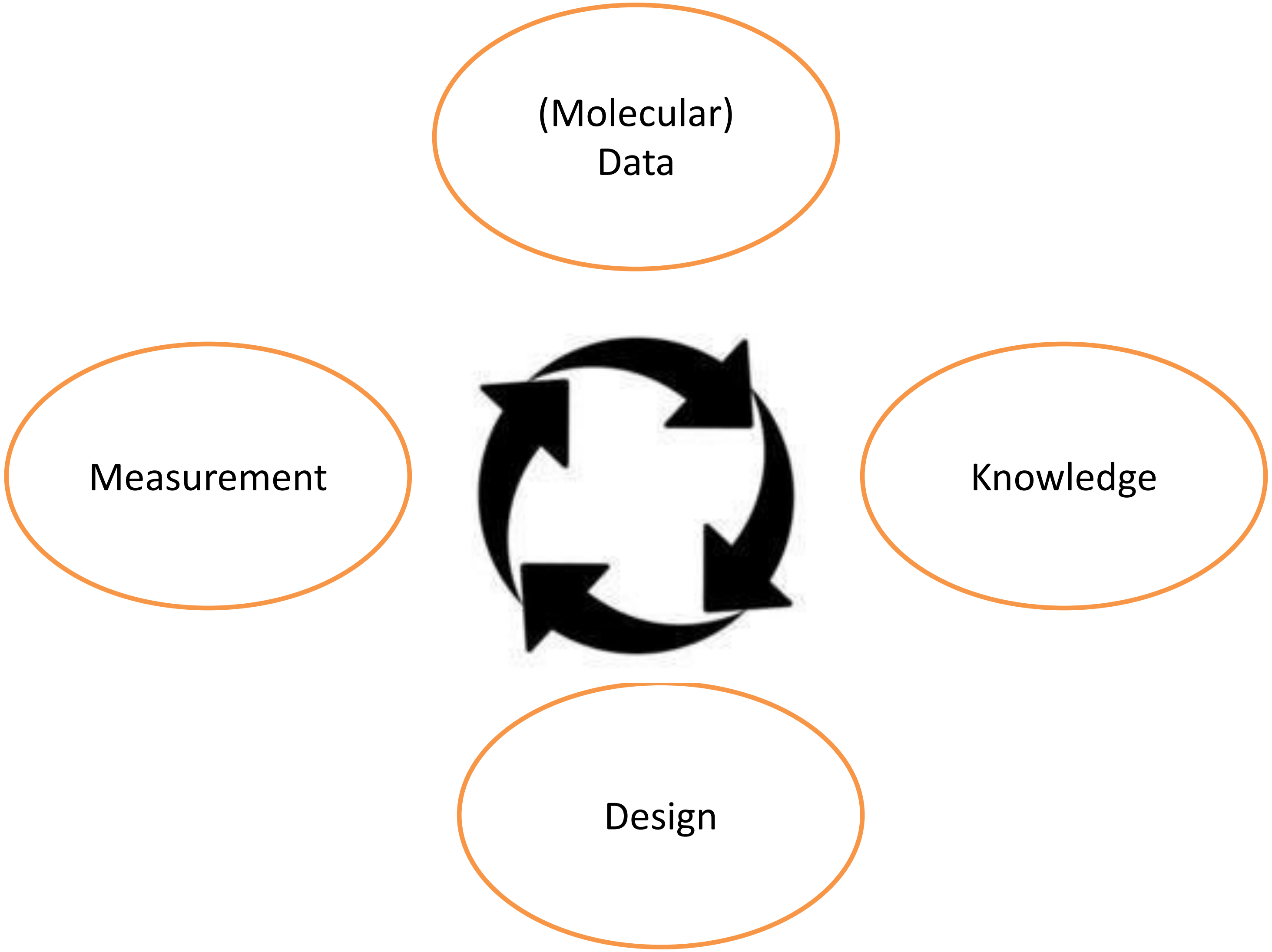
Speed/Sensitivity Tradeoff

- Efficient algorithm is ideal (no need to trade off).
- Filtration and better filtration.
  - Seed
  - 2-hit
  - HSP extension
  - Spaced seed
  - Multiple spaced seed

Useful Models/Data Structures

- Hiddel Markov Model
  - Find the path of hidden states to best explain the emitted symbols.
- Suffix Tree and Array
  - Linear space, linear time construction.
  - Support many efficient string operations
    - Substring query
    - Longest common substring
    - Maximal repeats
    - Maximal unique match
- Information Distance
  - Universal distance but noncomputable

Bioinformatics



(Molecular) Data

Measurement

Knowledge

Design

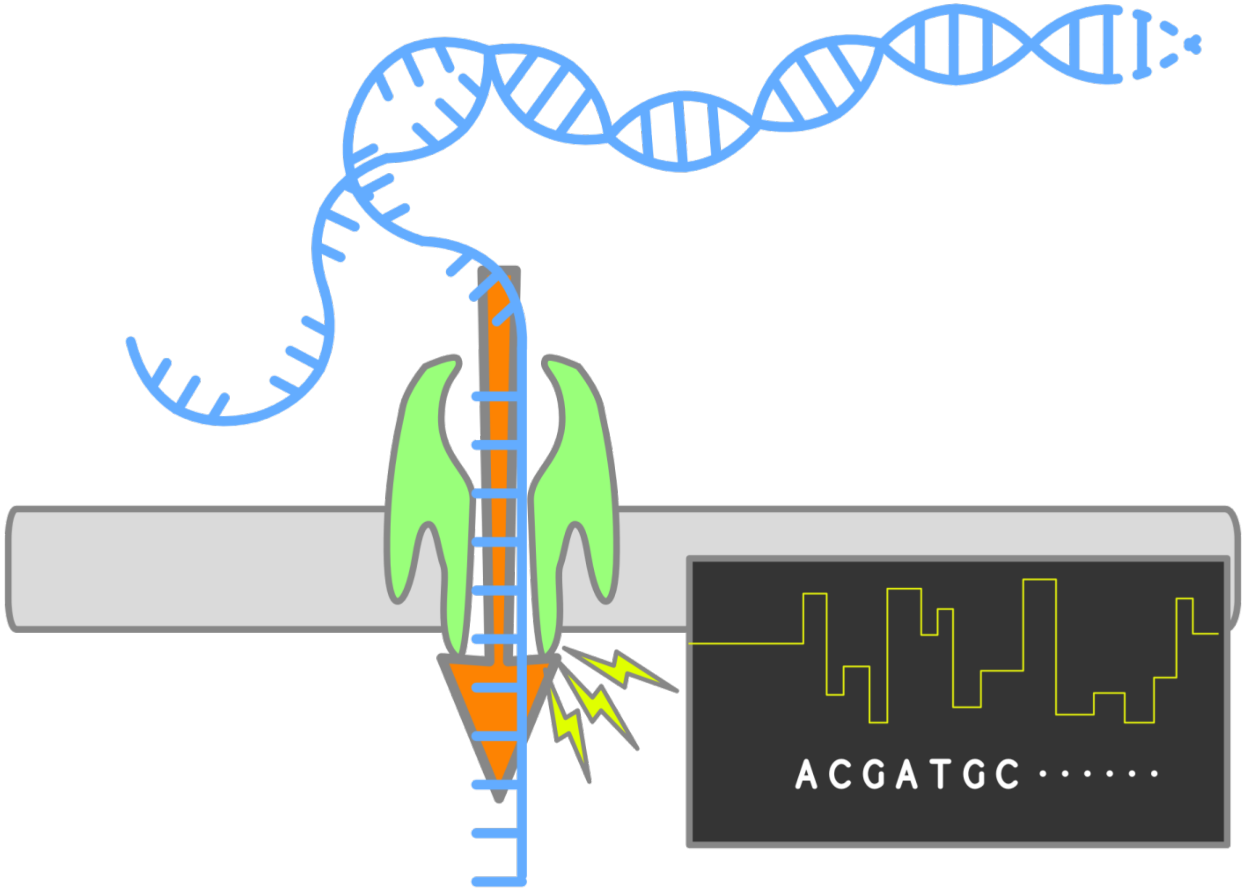
Bioinformatics: Data Analysis

- Sequence alignment
- Multiple sequence alignment
- Homology search
- Gene prediction
- Phylogeny
- Protein structure prediction

Bioinformatics: Data Generation

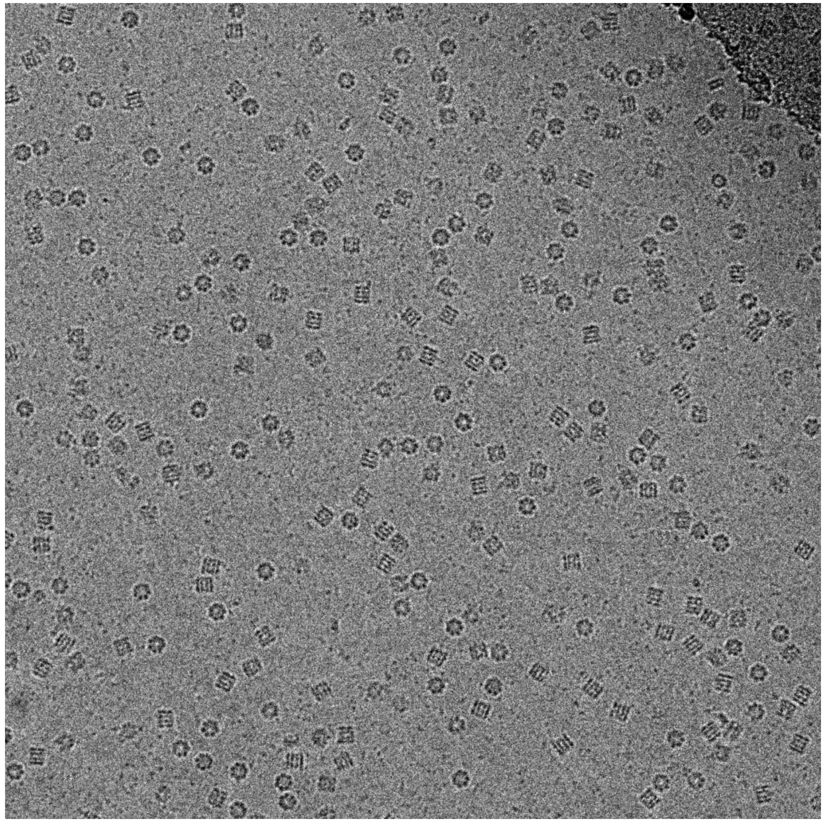
- E.g. Mass spectrometry based proteomics
  - Peptide identification via database search
  - De novo peptide sequencing
  - De novo protein sequencing
  - Quantification
  - Multiple myeloma
- Other examples not studied
  - DNA sequencing
  - Cryogenic electron microscopy (cryo-EM)

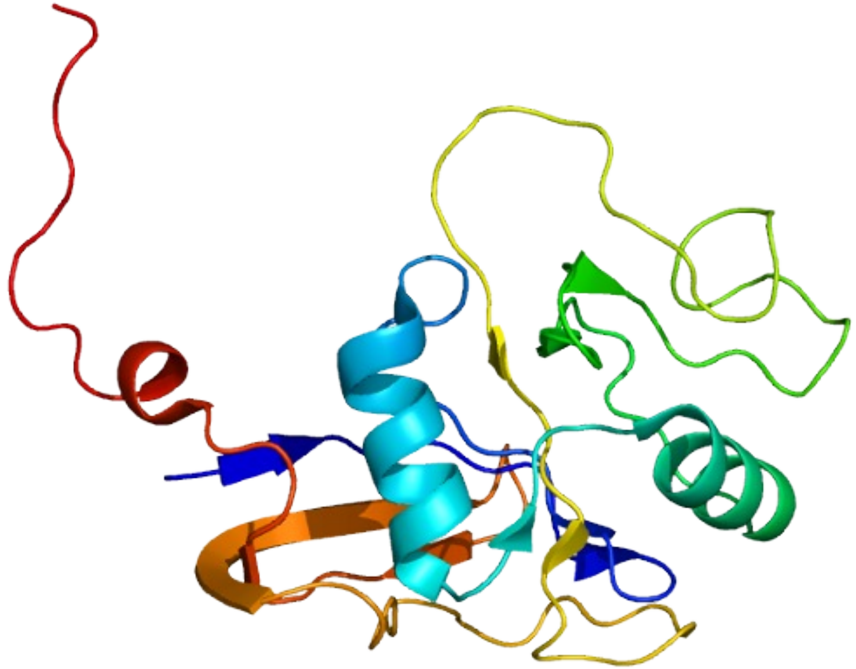
Nanopore sequencing



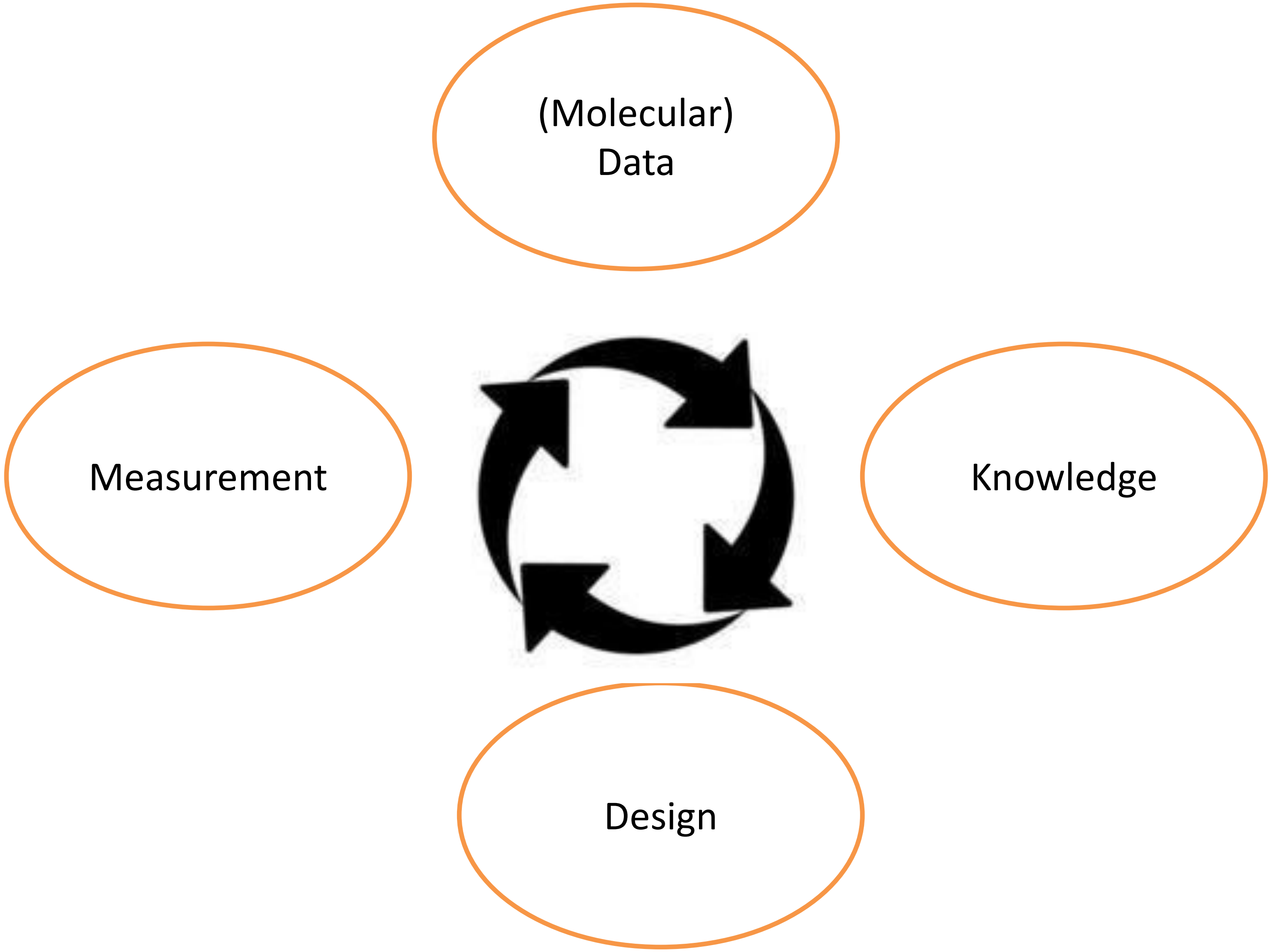
- Signal processing
- Error correction

Cryo EM





Bioinformatics



(Molecular) Data

Measurement

Knowledge

Design