

# Algorithms in Bioinformatics

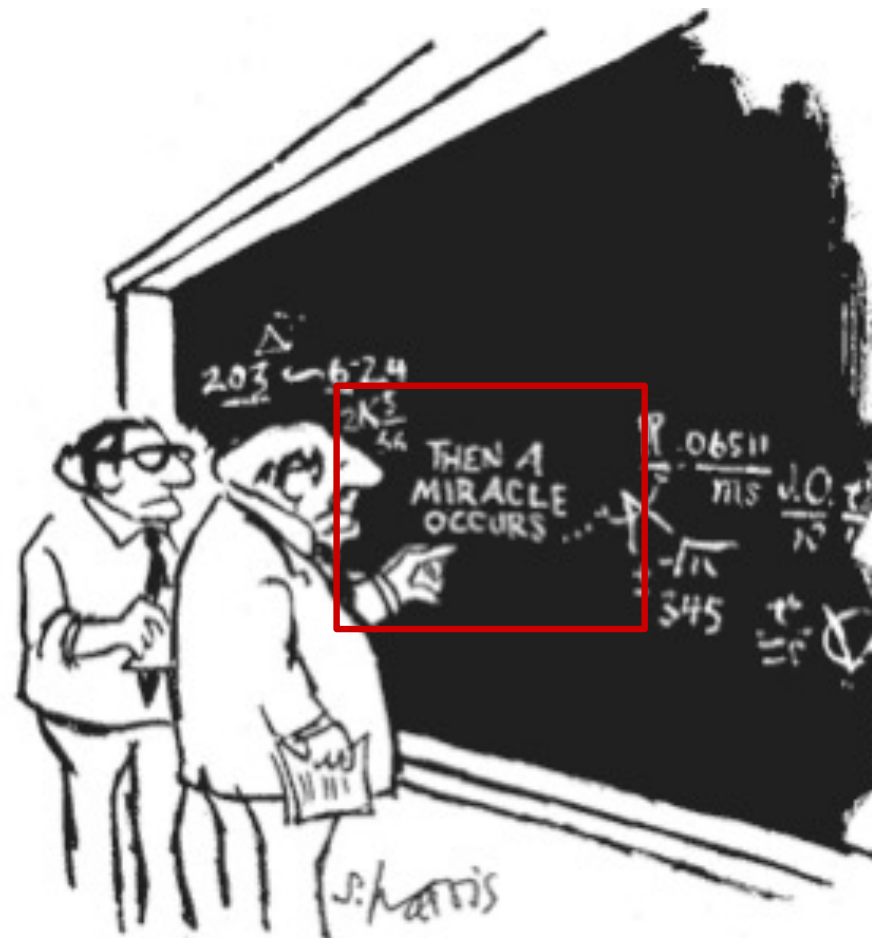
## #22125/#22175



Morten Nielsen  
Department of Health Technology  
DTU

# Course objective

---

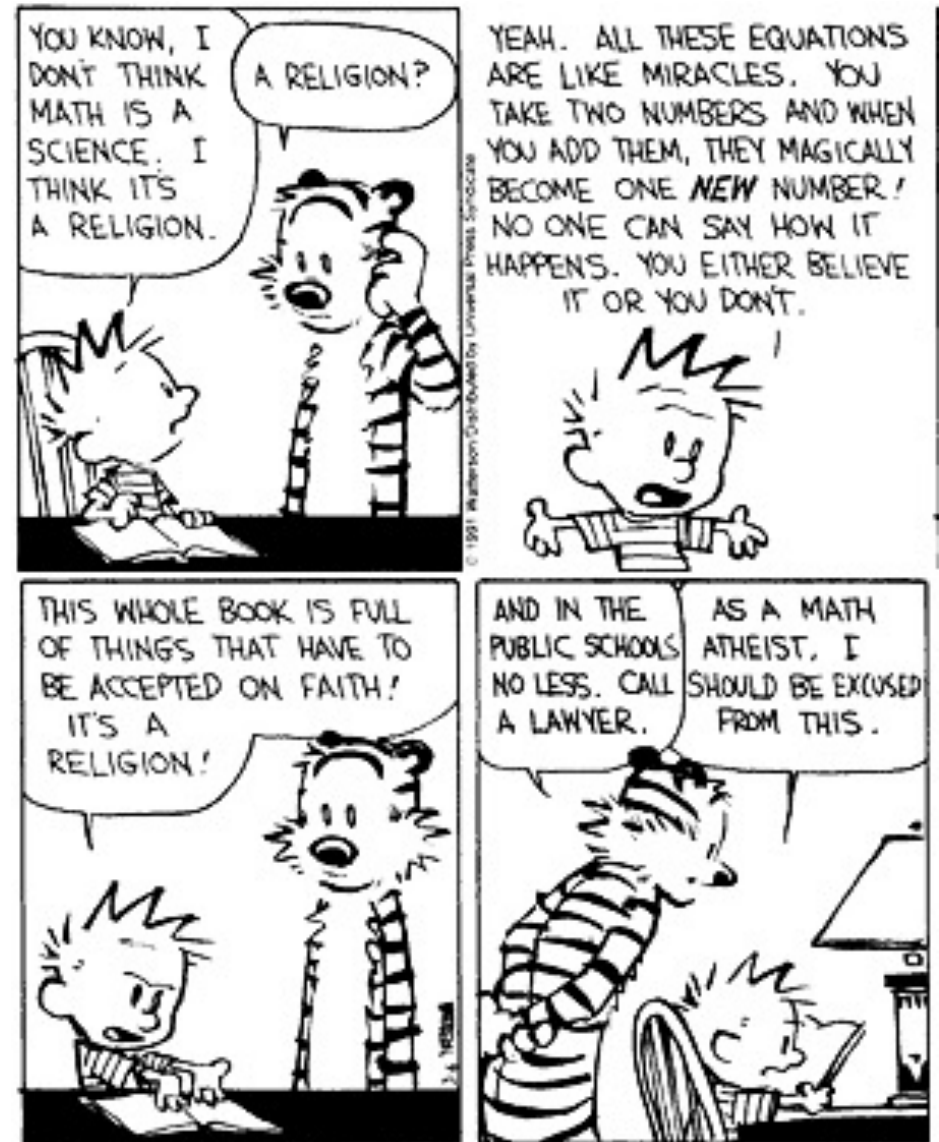


"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

---

# Algorithms are black-boxes

- No one knows how a neural network is trained
- No one knows how a PSSM is constructed
- Often no software exists that does exactly what you need



# Where conventional algorithms fail ..

---

- Sequence alignment

```

1PMY._ 4 VKMLNSGPGGMMVFDPALVRLKPGDSIKFLPTDKG--HNVETIKGMAPDG
      :   :   :   :   :   :   :   :   :   :   :   :
1PLC._ 0 IDVLLGADDGSLAFVPSEFSISPGEKIVF-KNNAGFPHNIVFDEDSIPSG

1PMY._ 54 ADYVKTTVGQEA-----VKFDKEGVYGFKCAPHYMMGMVALVVV
      :   :   :   :   :   :   :   :   :   :   :   :
1PLC._ 50 VDASKISMSEEDLLNAKGETFEVALSNKGEYSFYCSPHQGAGMVGKVTV
  
```

- Gaps should more likely be placed in loops and not in secondary structure elements
    - No conventional alignment algorithm can do this
-

# Sequence motif identification

---

- Say you have 10 ligands known to bind a given receptor. Can you accurately characterize the binding motif from such few data?
- HMM and Gibbs samplers might do this, but what if you know a priori that some positions are more important than others for the binding?
  - Then no conventional method will work

```

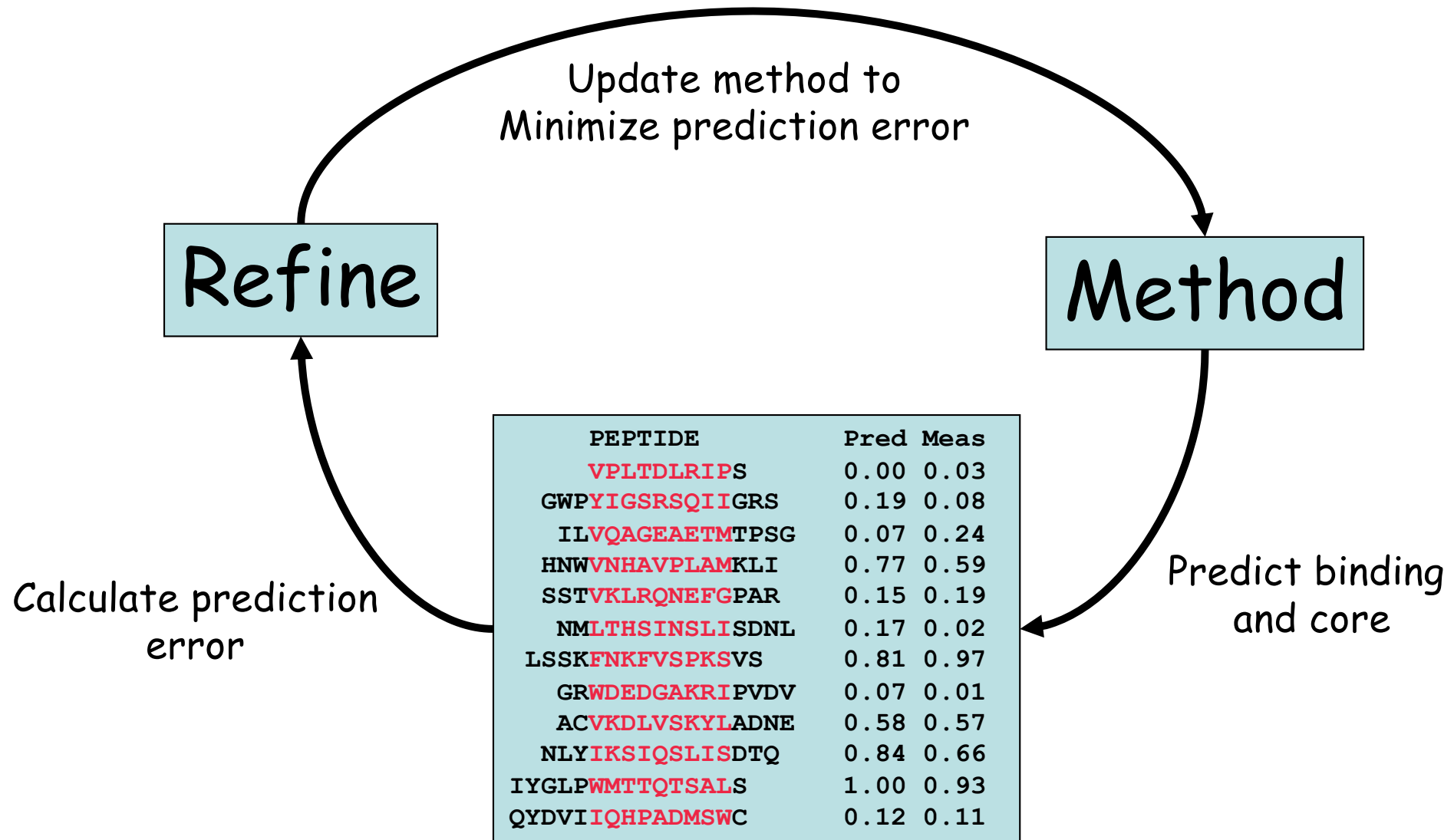
RFFGGDRGAPKRG
YLDPLIRGLLARPAKLQV
KPGQPPRLLIYDASNRATGIPA
GSLFVYNITTNKYKAFLDKQ
SALLSSDITASVNCAK
PKYVHQNTLKLAT
GFKGEQGPKGEP
DVFKELKVHHANENI
SRYWAIRTRSGGI
TYSTNEIDLQLSQEDGQTIE
  
```

# Artificial neural networks

Could an ANN be trained to simultaneously identify the binding motif and binding strength of a given peptide?

PEPTIDE	IC50 (nM)
VPLTDLRIPS	48000
GWPYIGSRSQIIGRS	45000
ILVQAGEAETMTPSG	34000
HNW <b>VNHAVPLAM</b> KLI	120
SSTVKLRQNEFGPAR	8045
NMLTHSINSLISDNL	47560
LSSK <b>FNK FVSPKS</b> VS	4
GRWDEDGAKRIPVDV	49350
AC <b>VKDLVSKYL</b> ADNE	86
NLY <b>IKSIQSLIS</b> DTQ	67
IYGLP <b>WMTTQTSALS</b>	11
QYDVIIQHPADMSWC	15245

# The Bioinformatical approach. NN-align



# Course objective

---

- To provide the student with an overview and in-depth understanding of bioinformatics machine-learning algorithms.
  - Enable the student to first evaluate which algorithm(s) are best suited for answering a given biological question and next
  - Implement and develop prediction tools based on such algorithms to describe complex biological problems such as immune system reactions, vaccine discovery, disease gene finding, protein structure and function, post-translational modifications etc.
-



# Course program

---

- Weight matrices
  - Sequence alignment
  - Hidden Markov Models
  - Sequence redundancy
  - Gibbs sampling
  - Stabilization matrix method
  - Artificial neural networks
  - Project
-

# The Mission

---

- When you have completed the course, you will have
    - Worked in great detail on all the most essential algorithms used in bioinformatics
    - Have a folder with program templates implementing these algorithms
    - When you in your future scientific carrier need to implement modifications to conventional algorithms, this should give you a solid starting point
-

# Course structure

---

- Mornings
    - Lectures and small exercises introducing the algorithms
  - Afternoons
    - Exercise where the algorithms are implemented
  - Project work in groups of 2-3 persons
    - The 1 week project work where a biological problem is analyzed using one or more of the algorithms introduced in the course
-

# Course structure

---

9.00 - 9.15

Introduction to course

[Introduction to course \[PDF\]](#)

9.15 - 9.30

[Introduction to the immune system \[PDF\]](#)

9.30 - 11.20 (coffee break included)

[Weight matrix construction \[PDF\]](#) . [\[PPTX\]](#)

[Logo Handout](#)

[Handout. Estimation of pseudo counts](#)

11.20 - 12.00

[Some notes on sequence alignment \[PDF\]](#)

12.00 - 13.00

Lunch

13.00 - 13.30

Questions to the mornings lectures and other general issues

Checking that we all have python and jupyter-notebook installed and running

13.30 - 17.00

A brief introduction to Python programming and Jupyter-notebooks

[Python intro](#)

Implementation of PSSM construction from pre-aligned sequences including pseudo count correction for low counts and sequence cl

[PSSM construction and evaluation](#)

---

# Programming language

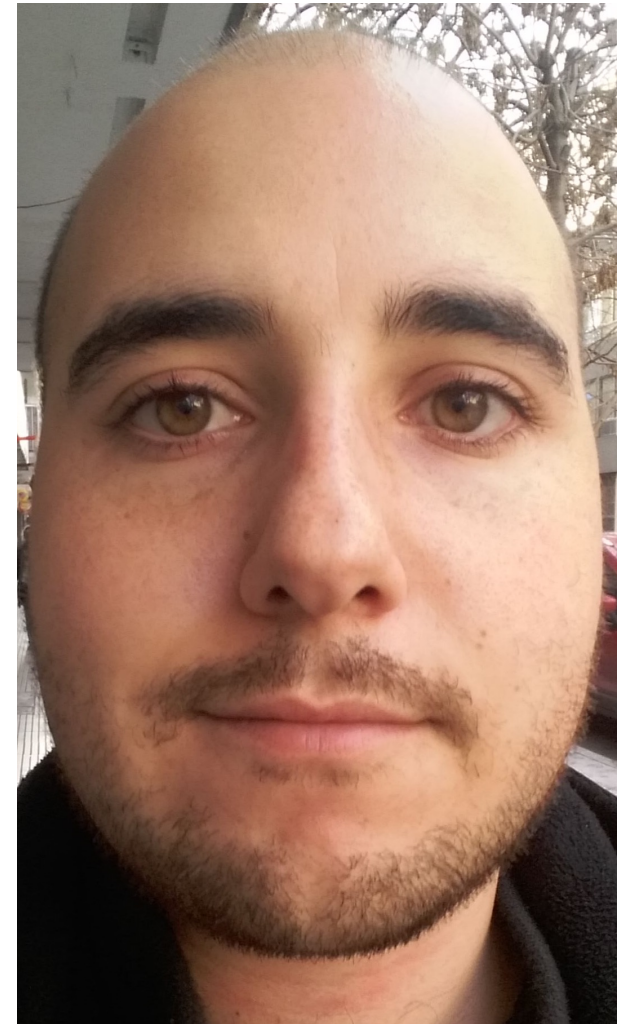
---

- I would have loved to do this in C
    - C is 10-50 times faster than Python
  - But, C is hard both to read and write
  - So for pedagogical reasons, I have decided to use Python, and jupyter-notebooks
-

# Programming language

---

- C code translated to Python by PhD student Brno Alvarez ([brunoalvarez89@gmail.com](mailto:brunoalvarez89@gmail.com))



# Course material

---

- Lund et al, MIT, chapter 3 and 4.
- Research papers
  - Check course program website for updates to course material

# Course Evaluation

---

- Written examination, and poster
  - Evaluation of poster and oral examination
  - Exam form
    - Group presentation of project
    - Written exam based on weekly exercises and the material of the course lectures
-