Morten Nielsen
Department of Health Technology
DTU
Course objective

"I think you should be more explicit here in step two."

"Then a miracle occurs..."
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  VKMLNSGPGGMMVFDPALVRLKPDBSIKFLPTDKG--HNVETIKGMAPDG
  1PLIC._ 0  IDVLLGADDGSALFVPSEFSISPGEKIVF-KNNAGFPHNIVFDEDSIPSG
  1PMY._ 54  ADYVKTTVQEQEAV---------VKFDKEGVYGFCKAPHYMMGVMVALVVV
  1PLIC._ 50  VDASKISMSSEDLNLNAPTFEVALSNKGEYSFHYCSPHQAGMGVGVTV

- **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

```
RF FGGDRGAPKR G
YLDPLIRGLLARPAKLQV
KPGQPPRLLIYDASNRATGIPA
GSLFVYNITTNKYKAFLDKQ
SALLSSDITASVNCAK
PKYVHQNTLKLAT
GFKGEQGPKGEP
DVFKELKVHHAENI
SRYWAIRTRSGGI
TYSTNEIDLOLSQEDGQTIE
```
Artificial neural networks

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

<table>
<thead>
<tr>
<th>PEPTIDE</th>
<th>IC50 (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPLTDLRIPS</td>
<td>48000</td>
</tr>
<tr>
<td>GWPYIGSRSQIIGRS</td>
<td>45000</td>
</tr>
<tr>
<td>ILVQAGEAETMTPSG</td>
<td>34000</td>
</tr>
<tr>
<td>HNWVNHAVPLAMKLI</td>
<td>120</td>
</tr>
<tr>
<td>SSTVKLRQNEFGPAR</td>
<td>8045</td>
</tr>
<tr>
<td>NMLTHSINSLISDNL</td>
<td>47560</td>
</tr>
<tr>
<td>LSSKYFNFVSPKSVS</td>
<td>4</td>
</tr>
<tr>
<td>GRWDEDGAKRIPVDV</td>
<td>49350</td>
</tr>
<tr>
<td>ACVKDLVSKYLADNE</td>
<td>86</td>
</tr>
<tr>
<td>NLYIKSIQSLISDTQ</td>
<td>67</td>
</tr>
<tr>
<td>IYGLPWMTTQTSALS</td>
<td>11</td>
</tr>
<tr>
<td>QYDVIIQHPADMSWC</td>
<td>15245</td>
</tr>
</tbody>
</table>
The Bioinformatical approach. NN-align

Refine

Calculate prediction error

Update method to
Minimize prediction error

Method

Predict binding and core

<table>
<thead>
<tr>
<th>PEPTIDE</th>
<th>Pred</th>
<th>Meas</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPLTDLRIPS</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td>GWPYIGSRSQIIGRS</td>
<td>0.19</td>
<td>0.08</td>
</tr>
<tr>
<td>ILVQAGEAETTPSG</td>
<td>0.07</td>
<td>0.24</td>
</tr>
<tr>
<td>HNWVNHAVPLAMKL</td>
<td>0.77</td>
<td>0.59</td>
</tr>
<tr>
<td>SSTVKLQNEFGPAR</td>
<td>0.15</td>
<td>0.19</td>
</tr>
<tr>
<td>NMLTHSINSLISDNL</td>
<td>0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>LSSKFNKFVSPKSVS</td>
<td>0.81</td>
<td>0.97</td>
</tr>
<tr>
<td>GRWDEDGAKRIPVDV</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>ACVKDLVSKYLADNE</td>
<td>0.58</td>
<td>0.57</td>
</tr>
<tr>
<td>NLYIKSIQSLISDTQ</td>
<td>0.84</td>
<td>0.66</td>
</tr>
<tr>
<td>IYGLPWMTTQTSAI</td>
<td>1.00</td>
<td>0.93</td>
</tr>
<tr>
<td>QYDVIIQHPADMSC</td>
<td>0.12</td>
<td>0.11</td>
</tr>
</tbody>
</table>
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

**Thursday, 4. June**

*Introduction to course, UNIX and Python programming crash course 101*

*Morten Nielsen*

## BACKGROUND TEXTS

- Python for Non-Programmers
- Python Numpy introduction

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00</td>
<td>9.15 &quot;Recorded&quot; Introduction to course</td>
</tr>
<tr>
<td>9.15</td>
<td>9.35 &quot;Recorded&quot; Introduction to the immune system</td>
</tr>
<tr>
<td>9.35</td>
<td>9.45 &quot;Recorded&quot; Performance measures</td>
</tr>
<tr>
<td>9.45</td>
<td>10.00 Coffe break</td>
</tr>
<tr>
<td>10.00</td>
<td>10.15 &quot;Recorded&quot; Some notes on sequence alignment</td>
</tr>
<tr>
<td>10.15</td>
<td>12.00 &quot;Online&quot; Checking that we all have python and jupyter-notebook installed and running</td>
</tr>
<tr>
<td>12.00</td>
<td>13.00 Lunch</td>
</tr>
<tr>
<td>13.00</td>
<td>17.00 &quot;Online&quot; Weight matrix construction</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*A UNIX/Linux crash course*

*Handout. Estimation of pseudo counts*
Programming language

• C
  - C
    • C
      - C
• Why C?
  - Nothing compares to C in speed
    • Least of all Perl and Python
  - Nothing comes for free in C
    • You in most cases cannot simply use “library” sub routines
Programming language

• C
  - C
    • C
    - C
    » C
• Why C?
  - Nothing compares to C in speed
    • Least of all Perl and Python
  - Ex. A Gibbs sampler coded in Perl runs 50 times slower than the same method coded in C
But, C is hard both to read and write
So for pedagogical reasons, I have decided to switch to python, and jupyter-notebooks
Programming language

- C code translated to Python by PhD student Brno Alvarez (brunoalvarez89@gmail.com)
• Python is NOT my preferred programming language
  - You are my test rabbits
  - Please give feedback
  - Python is slow :(
Course material

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

- Oral examination, and report
- Evaluation of report and oral examination
- Exam form
  - Group presentation of project
  - Individual “Portfolio exam” based on weekly exercises and the material of the course lectures