27622: Introduction to Bioinformatics, Turbo version

Henrik Nielsen, Associate professor
Center for Biological Sequence Analysis
What is bioinformatics?
What are bioinformaticians up to, actually?

- **Manage** molecular biological data
  - Store in *databases*, organise, formalise, describe...
- **Compare** molecular biological data
- **Find** *patterns* in molecular biological data
  - *phylogenies*
  - *correlations* (sequence / structure / expression / function / disease)

**Goals:**
- *characterise* biological patterns & processes
- *predict* biological properties
  - low level data ⇒ high level properties
    (eg., sequence ⇒ function)
Bioinformatics: neighbour disciplines

• Computational biology
  – Broader concept: includes computational ecology, physiology, neurology etc...

• -omics:
  – Genomics
  – Transcriptomics
  – Proteomics

• Systems biology
  – Putting it all together...
  – Building models, identify control & regulation
A view of Systems Biology

- Cell
- Tissue
- Macro-molecule
- Biomedicine
- Bioinformatics
- Metabolite
- Health
- Biotechnology
Bioinformatics: prerequisites

• **Bio-side:**
  – Molecular biology
  – Cell biology
  – Genetics
  – Evolutionary theory

• **informatics side:**
  – Computer science
  – Statistics
  – Theoretical physics
• DNA sequences

>alpha-D
ATGCTGACCCGACCTCTGAGAAGCTGGTCCTGCAGGTGTGGGAGAAGGTGATCCGCCACCCAGACTGTGGAGCCGAGGCCCTGGAGAGGTGCGGGCTGAGCTTGGGGAAACCATGGGCAAGGGGGCGACTGGGTGGGAGCCCTACAGGGCTGCTGGGGGTTGTTCGGCTGGGGGTCAGCACTGACCATCCCGCTCCCGCAGCTGTTCACCACCTACCCCCAGACCAAGACCTACTTCCCACTTCGACTTGCACCATGGCTCCGACCAGGTCCGCAACCACGGCAAGAAGGTGTTGGCCGCCTTGGGCAACGCTGTCAAGAGCCTGGGCAACCTCAGCCAAGCCCTGTCTGACCTCA

>alpha-A
ATGTCGCTGTCTGCCAACGACAAGAGCAACGTGAAGGCCGTCTTCGGCAAAATCGGCGGCAGGCCGGTGACTTGGGTGGTGAAGCCCTGGAGAGGTATGTGGTCATCCGTCATTACCCCATCTCTTGTCTGTCTGTGACTCCATCCCATCTGCCCCCATACTCTCCCCCATCCATATCTGCCTGGCTCTGTCTCATCTGTCCCCAACTGTCCCTGATTGCCTCTGTCCCCCAGGTTGTTCATCACCTACCCCCAGACCAAGACCTACTTCCCCCACTTCGACCTGTCACATGGCTCCGCTCAGATCAAGGGGCACGGCAAGAAGGTGGCGGAGGCACTGGTTGGAGGCTGCCAACACCACATCGATGACATCGCTGGTGCCCTCTCCAAGCTGAGCGACCTCCACGGCCAACAAAGCTCCGTGTGGACCCCGTCAACTTCAAAGTGAGCATCTGGGAAGGGGTGACCATGCTCTGGCTCCCCTCCTGCACACACCTCTGGCTACCCCCTCACCTCACCCCCTTGCTCACCATCTCCTTTTGCCTTTCAGCTGCTGGGTCACTGCTTCCTGGTGGTCGTGGCCGTCCACTTCCCCTCTCTCCTGACCCCGGAGGTCCATGCTTCCCTGGACAAGTTCGTGTGTGCCGTGGGACACCGTCCTTACTGCCAAGTACCGTTAAA
Molecular biology data...

- Amino acid sequences
- Protein structure:
  - X-ray crystallography
  - NMR
• Subcellular localization
Cell biology & proteomics data...

protein-protein interactions
Transcriptomics: DNA microarray technology
Phenotype data: human diseases

ACHOO SYNDROME

Alternative titles: symbols

AUTOSOMAL DOMINANT COMPPELLING HELIOOPHTHALMIC OUTBURST SYNDROME
PHOTIC SNEEZE REFLEX
SNEEZING FROM LIGHT EXPOSURE
PEROUTKA SNEEZE

TEXT

Collie et al. (1978) described a 'disorder' characterized by nearly uncontrollable paroxysms of sneezing provoked in a reflex fashion by the sudden exposure of a dark-adapted subject to intensely bright light, usually sunlight. The number of successive sneezes was usually 2 or 3, but could be as many as 43. The 4 authors were the probands of the 4 families they reported. Several instances of male-to-male transmission were noted. Sneezing in response to bright light was said by Peroutka and Peroutka (1984) to be a common yet poorly understood phenomenon. Photic sneeze reflex was suggested as the appropriate designation by Everett (1954), who found it in 23% of Johns Hopkins medical students. In a poll of 25 neurologists at Johns Hopkins, Peroutka and Peroutka (1984) found the phenomenon in 9, but only 2 of the respondents knew that such a specific reflex exists. The Peroutkas (father and daughter) reported the reflex in 3 generations of their family: grandfather, the father (the proband), his brother and his daughter. The index subject (S.J.P.) invariably sneezes twice when he moves from indoors into bright sunlight. Lewkowia (1969) described sneezing as a complication of slit lamp examination.
Prediction methods

• Homology / Alignment
• Simple pattern (“word”) recognition
• Statistical methods
  – Weight matrices: calculate amino acid probabilities
  – Other examples: Regression, variance analysis, clustering
• Machine learning
  – Like statistical methods, but parameters are estimated by iterative training rather than direct calculation
  – Examples: Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machines (SVM)
• Combinations
The computer

- *Everything* can be reduced to bits (0 or 1)
• A byte = 8 bits
0 1 0 0 0 0 0 1

Can be interpreted as
• The number 65
• The letter ”A”
• Part of a machine code instruction
• Part of a colour specification
• Part of a sound encoding
• …
A text file is a file where every byte is interpreted as a character.

**Examples**

- Plain text .txt
- Program settings .ini
- C source code .c
- Python script .py
- \( \text{\LaTeX} \) source .tex
- Web page source .html
- Sequences .fasta

### The ASCII Table

<table>
<thead>
<tr>
<th>Dec</th>
<th>Hex</th>
<th>Char</th>
<th>Dec</th>
<th>Hex</th>
<th>Char</th>
<th>Dec</th>
<th>Hex</th>
<th>Char</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>00</td>
<td>Null</td>
<td>32</td>
<td>20</td>
<td>Space</td>
<td>64</td>
<td>40</td>
<td>Space</td>
</tr>
<tr>
<td>1</td>
<td>01</td>
<td>Start of heading</td>
<td>33</td>
<td>21</td>
<td>!</td>
<td>65</td>
<td>41</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>02</td>
<td>Start of text</td>
<td>34</td>
<td>22</td>
<td>”</td>
<td>66</td>
<td>42</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>03</td>
<td>End of text</td>
<td>35</td>
<td>23</td>
<td>#</td>
<td>67</td>
<td>43</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>04</td>
<td>End of transmit</td>
<td>36</td>
<td>24</td>
<td>$</td>
<td>68</td>
<td>44</td>
<td>D</td>
</tr>
<tr>
<td>5</td>
<td>05</td>
<td>Enquiry</td>
<td>37</td>
<td>25</td>
<td>%</td>
<td>69</td>
<td>45</td>
<td>E</td>
</tr>
<tr>
<td>6</td>
<td>06</td>
<td>Acknowledge</td>
<td>38</td>
<td>26</td>
<td>&amp;</td>
<td>70</td>
<td>46</td>
<td>F</td>
</tr>
<tr>
<td>7</td>
<td>07</td>
<td>Bell</td>
<td>39</td>
<td>27</td>
<td>^</td>
<td>71</td>
<td>47</td>
<td>G</td>
</tr>
<tr>
<td>8</td>
<td>08</td>
<td>Backspace</td>
<td>40</td>
<td>28</td>
<td>(</td>
<td>72</td>
<td>48</td>
<td>H</td>
</tr>
<tr>
<td>9</td>
<td>09</td>
<td>Vertical tab</td>
<td>41</td>
<td>29</td>
<td>)</td>
<td>73</td>
<td>49</td>
<td>I</td>
</tr>
<tr>
<td>10</td>
<td>0A</td>
<td>Line feed</td>
<td>42</td>
<td>2A</td>
<td>*</td>
<td>74</td>
<td>4A</td>
<td>J</td>
</tr>
<tr>
<td>11</td>
<td>0B</td>
<td>Form feed</td>
<td>43</td>
<td>2B</td>
<td>+</td>
<td>75</td>
<td>4B</td>
<td>K</td>
</tr>
<tr>
<td>12</td>
<td>0C</td>
<td>Form space</td>
<td>44</td>
<td>2C</td>
<td>,</td>
<td>76</td>
<td>4C</td>
<td>L</td>
</tr>
<tr>
<td>13</td>
<td>0D</td>
<td>Carriage return</td>
<td>45</td>
<td>2D</td>
<td>-</td>
<td>77</td>
<td>4D</td>
<td>M</td>
</tr>
<tr>
<td>14</td>
<td>0E</td>
<td>Shift out</td>
<td>46</td>
<td>2E</td>
<td>.</td>
<td>78</td>
<td>4E</td>
<td>N</td>
</tr>
<tr>
<td>15</td>
<td>0F</td>
<td>Shift in</td>
<td>47</td>
<td>2F</td>
<td>/</td>
<td>79</td>
<td>4F</td>
<td>O</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>Digits (0-9)</td>
<td>48</td>
<td>30</td>
<td>0</td>
<td>80</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>Device control</td>
<td>49</td>
<td>31</td>
<td>1</td>
<td>81</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>Device control</td>
<td>50</td>
<td>32</td>
<td>2</td>
<td>82</td>
<td>52</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>13</td>
<td>Device control</td>
<td>51</td>
<td>33</td>
<td>3</td>
<td>83</td>
<td>53</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>14</td>
<td>Device control</td>
<td>52</td>
<td>34</td>
<td>4</td>
<td>84</td>
<td>54</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>15</td>
<td>Negative acknowledge</td>
<td>53</td>
<td>35</td>
<td>5</td>
<td>85</td>
<td>55</td>
<td>5</td>
</tr>
<tr>
<td>22</td>
<td>16</td>
<td>Synchronous idle</td>
<td>54</td>
<td>36</td>
<td>6</td>
<td>86</td>
<td>56</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
<td>End of medium</td>
<td>55</td>
<td>37</td>
<td>7</td>
<td>87</td>
<td>57</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>18</td>
<td>Cancel</td>
<td>56</td>
<td>38</td>
<td>8</td>
<td>88</td>
<td>58</td>
<td>8</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
<td>End of line</td>
<td>57</td>
<td>39</td>
<td>9</td>
<td>89</td>
<td>59</td>
<td>9</td>
</tr>
<tr>
<td>26</td>
<td>1A</td>
<td>Substitution</td>
<td>58</td>
<td>3A</td>
<td>:</td>
<td>90</td>
<td>5A</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>1B</td>
<td>Escape</td>
<td>59</td>
<td>3B</td>
<td>\</td>
<td>91</td>
<td>5B</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>1C</td>
<td>File separator</td>
<td>60</td>
<td>3C</td>
<td>&lt;</td>
<td>92</td>
<td>5C</td>
<td>\</td>
</tr>
<tr>
<td>29</td>
<td>1D</td>
<td>Group separator</td>
<td>61</td>
<td>3D</td>
<td>=</td>
<td>93</td>
<td>5D</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1E</td>
<td>Record separator</td>
<td>62</td>
<td>3E</td>
<td>&gt;</td>
<td>94</td>
<td>5E</td>
<td>^</td>
</tr>
<tr>
<td>31</td>
<td>1F</td>
<td>Unit separator</td>
<td>63</td>
<td>3F</td>
<td>?</td>
<td>95</td>
<td>5F</td>
<td></td>
</tr>
</tbody>
</table>
The are *many* ways to interpret characters with values above 127. Here, you see two of them.

**Mac OS Roman** Encoding:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>032</td>
<td>!</td>
<td>&quot;</td>
<td>#</td>
<td>$</td>
<td>%</td>
<td>&amp;</td>
<td>'</td>
<td>(</td>
<td>)</td>
<td>*</td>
<td>+</td>
<td>,</td>
<td>-</td>
<td>.</td>
<td>/</td>
<td></td>
</tr>
<tr>
<td>048</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>:</td>
<td>&lt;</td>
<td>=</td>
<td>&gt;</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>064</td>
<td>@</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
<td>O</td>
</tr>
<tr>
<td>080</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>[</td>
<td>\</td>
<td>]</td>
<td>^</td>
<td>_</td>
</tr>
<tr>
<td>096</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
<td>e</td>
<td>f</td>
<td>g</td>
<td>h</td>
<td>i</td>
<td>j</td>
<td>k</td>
<td>l</td>
<td>m</td>
<td>n</td>
<td>o</td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>p</td>
<td>q</td>
<td>r</td>
<td>s</td>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td>{</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>A</td>
<td>A</td>
<td>Ç</td>
<td>E</td>
<td>Ê</td>
<td>Ë</td>
<td>Ì</td>
<td>Ï</td>
<td>Ï</td>
<td>Ï</td>
<td>Ï</td>
<td>Ï</td>
<td>Ï</td>
<td>Ï</td>
<td>Ï</td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>é</td>
<td>é</td>
<td>í</td>
<td>í</td>
<td>ì</td>
<td>ó</td>
<td>ö</td>
<td>ö</td>
<td>õ</td>
<td>õ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>ñ</td>
<td>ñ</td>
<td>è</td>
<td>è</td>
<td>é</td>
<td>â</td>
<td>â</td>
<td>â</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
<td>±</td>
</tr>
<tr>
<td>192</td>
<td>ñ</td>
<td>ñ</td>
<td>è</td>
<td>è</td>
<td>é</td>
<td>â</td>
<td>â</td>
<td>â</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
<td>ã</td>
</tr>
<tr>
<td>208</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>224</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
<td>ñ</td>
</tr>
</tbody>
</table>

Windows-1252, sometimes called incorrectly "ANSI". Blue dots indicate unused or control characters.
Text files—line endings

- UNIX standard (including Mac OS X):
  - 10 — LF ("Line feed" char).
- Old Mac (System 9 and before):
  - 13 — CR ("Carriage Return" char).
- DOS/Windows:
  - 13, 10 — both CR and LF.

A good text editor can handle all three systems. Notepad for Windows cannot!
jEdit is a mature programmer's text editor with hundreds of years of development behind it. To download, install, and set up jEdit as quickly and painlessly as possible, go to the Quick Start page.

While jEdit boasts many expensive development tools for features and ease of use, it is released as free software with full source code, provided under the terms of the GPL 2.0.

The jEdit core, together with a large collection of plugins is maintained by a world-wide developer team.

Some of jEdit's features include:

- Written in Java, so it runs on Mac OS X, OS/2, Unix, VMS and Windows.
- Built-in macro language, extensible plugin architecture. Hundreds of macros and plugins available.
- Plugins can be downloaded and installed from within jEdit using the "plugin manager" feature.
- Auto indent, and syntax highlighting for more than 200 languages.
- Supports a large number of character encodings including UTF8 and Unicode.
- Folding for selectively hiding regions of text.
- Word wrap.
- Highly configurable and customizable.
- Every other feature, both basic and advanced, you would expect to find in a text editor. See the Features page for a full list.