Protein Prediction I for Computer Scientists

Biological Background
May 11/16th, Summer Term 2017
Burkhard Rost & Lothar Richter

PP1CS SoSe 17
Lecture and exercise

- [link](https://www.rostlab.org/teaching/ss17/pp1cs)
- Announcements, slides and videos
- **Lecture** Tuesdays (10:00-11:30 am) and Thursdays (10:00 – 11:30 am)
- Room MW 1801 (Mechanical Engineering)
- **Exercise** Thursdays 12:30 – 14.00 pm
  Room Hörsaal 3 (MI 00.06.011, Lecture hall 3) and mostly MW2250 on Tuesday 13-15
- **Register** for the lecture and exam in [TUM online](#)
Exercise

- Exercise wiki
  https://i12r-studfilesrv.informatik.tu-muenchen.de/sose17/pp4cs1/index.php/Main_Page
Exercise – Topics and Schedule

<table>
<thead>
<tr>
<th>Slot</th>
<th>Thursday</th>
<th>Tuesday</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>May 4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>May 9&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Structure of the Exercise / Biological Background</td>
</tr>
<tr>
<td>2</td>
<td>May 11&lt;sup&gt;th&lt;/sup&gt;</td>
<td>May 16&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Biological background</td>
</tr>
<tr>
<td>3</td>
<td>May 18&lt;sup&gt;th&lt;/sup&gt;</td>
<td>May 23&lt;sup&gt;rd&lt;/sup&gt;</td>
<td>Protein structures</td>
</tr>
<tr>
<td>4</td>
<td>Jun 1&lt;sup&gt;st&lt;/sup&gt;</td>
<td>Jun 6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Resources for Biological Information / Formats</td>
</tr>
<tr>
<td>5</td>
<td>Jun 8&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Jun 13&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Alignments</td>
</tr>
<tr>
<td>6</td>
<td>Jun 22&lt;sup&gt;nd&lt;/sup&gt;</td>
<td>Jun 27&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Machine Learning incl. Tricks / Secondary Structure Prediction</td>
</tr>
<tr>
<td>7</td>
<td>Jun 29&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Jul 4&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Homology Modeling / Prediction of Other Protein Features</td>
</tr>
<tr>
<td>8</td>
<td>Jul 6&lt;sup&gt;th&lt;/sup&gt;</td>
<td></td>
<td>Wrap Up – Questions</td>
</tr>
<tr>
<td></td>
<td>WED</td>
<td>Jul 12&lt;sup&gt;th&lt;/sup&gt;</td>
<td>EXAM</td>
</tr>
</tbody>
</table>
Molecular Mass

- derived from the atomic nuclear composition
- 1 u resp. 1 Da corresponds to the mass of a single nucleon (either proton or neutron)
- Definition: 1 u is one twelfth of the mass of an unbound neutral atom of carbon-12 (¹²C) in its nuclear and electronic ground state and at rest.
- $1.66\ldots \times 10^{-27}$ kg $\approx$ 1 yoctogram

Amount of Substance, the mole [mol]

- SI base unit
- The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon-12; its symbol is "mol".
- When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, other particles, or specified groups of such particles.
Amount of Substance, the mole [mol]

The definition of the mole also determines the value of the universal constant that relates the number of entities to amount of substance for any sample. This constant is called the Avogadro constant, symbol $N_A$ or $L$, and has the value $6.022... \times 10^{23}$ mol$^{-1}$ (entities per mole).
### Chemical Elements

(Mendeleev's) Periodic Table of Chemical Elements via TikZ

<table>
<thead>
<tr>
<th>Element</th>
<th>Atomic Number</th>
<th>Mass Number</th>
<th>Mass</th>
<th>Period</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrogen</td>
<td>1</td>
<td>1</td>
<td>0.0079</td>
<td>1</td>
<td>1 IA</td>
</tr>
<tr>
<td>Lithium</td>
<td>3</td>
<td>6</td>
<td>6.941</td>
<td>2</td>
<td>2IA</td>
</tr>
<tr>
<td>Sodium</td>
<td>11</td>
<td>23</td>
<td>22.99</td>
<td>3</td>
<td>3IA</td>
</tr>
<tr>
<td>Magnesium</td>
<td>12</td>
<td>24</td>
<td>24.305</td>
<td>4</td>
<td>4VA</td>
</tr>
<tr>
<td>Silicon</td>
<td>14</td>
<td>28</td>
<td>28.086</td>
<td>4</td>
<td>13IA</td>
</tr>
<tr>
<td>Phosphorus</td>
<td>15</td>
<td>31</td>
<td>30.974</td>
<td>5</td>
<td>15VA</td>
</tr>
<tr>
<td>Sulfur</td>
<td>16</td>
<td>32</td>
<td>32.06</td>
<td>6</td>
<td>16VA</td>
</tr>
<tr>
<td>Chlorine</td>
<td>17</td>
<td>35</td>
<td>35.45</td>
<td>7</td>
<td>17VIIA</td>
</tr>
<tr>
<td>Arsenic</td>
<td>33</td>
<td>75</td>
<td>74.92</td>
<td>4</td>
<td>18VIIA</td>
</tr>
<tr>
<td>Bromine</td>
<td>35</td>
<td>79</td>
<td>79.904</td>
<td>5</td>
<td>18VIIA</td>
</tr>
<tr>
<td>Iodine</td>
<td>53</td>
<td>127</td>
<td>126.9</td>
<td>6</td>
<td>18VIIA</td>
</tr>
<tr>
<td>Xe</td>
<td>54</td>
<td>131</td>
<td>131.3</td>
<td>6</td>
<td>18VIIA</td>
</tr>
</tbody>
</table>

**Non-metal**

- Caesium
- Sodium
- Lanthanide/Actinide
- Non-metal
- Metalloid
- Alkaline Earth Metal

**Metal**

- Potassium
- Rubidium
- Cs
- Barium
- La-Lu
- Scandium
- Actinium
- Ac
- Ti
- Ta
- W
- Mo
- Re
- Os
- Ir
- Pt
- Au
- Hg
- Tl
- Pb
- Bi
- Po
- At
- Rn

**Metalloid**

- Germanium
- Antimony
- Tellurium
- Lutetium

**Alkaline Earth Metal**

- Magnesium
- Barium

**Alkaline Earth Metal**

- Calcium
- Strontium
- Barium

**Transition Metal**

- Iron
- Nickel
- Cobalt
- Tungsten
- Molybdenum
- Tc
- Re
- Nb
- Pd
- Ru
- Rh
- Cd
- In
- Sn
- Sb
- Te
- I
- Xe

**Actinide**

- Uranium
- Neptunium
- Protactinium
- Neptunium
- Plutonium
- Americium
- Curium
- Americium
- Bk
- Cf
- Es
- Fm
- Md
- No

**Post-transition Metal**

- Gallium
- Indium
- Thallium

**Noble Gas**

- Neon
- Argon
- Krypton
- Xenon

**Uninherited Elements**

- Uninherited Elements

**Uninherited Elements**

- Uninherited Elements

**Uninherited Elements**

- Uninherited Elements

### Notes

- Uninherited Elements
- Uninherited Elements

### Links

Spread of Space and Time

- $1 \, \text{Å} = 100.000 \, \text{fm} = 100 \, \text{pm} = 0,1 \, \text{nm} = 10^{-4} \, \mu\text{m} = 10^{-7} \, \text{mm} = 10^{-8} \, \text{cm} = 10^{-10} \, \text{m}$
- the radius of isolated atoms is in the range of $0.3$-$3$ Å
- C-C: $154$ pm, C=C: $134$ pm, C-H: $109$ pm, C-O: $143$ pm, C=O: $120$ pm, C-N: $147$ pm, N-H: $101$ pm
Spread of Space and Time

- average molecular weight of an amino acid is (100)-110 Da
- typical soluble proteins are in the range of 3-6 nm (10^{-9}m)
- ~300 aa for prokaryotic proteins
- ~400 aa for eukaryotic protein
- visible light: 380-750 nm
Spread of Space and Time

- E. coli cell: 1x2 µm => ≈1 µm³
- S. cerevisiae: φ 4-6 µm
- H. sapiens: φ10-100 µm but:
  - 200 different cell types with very different shape
  - volume range of 30-4,000,000µm³
Some Giants

- Blue Whale: up to 190 tons, 24 m
- Lion's mane jellyfish: bell diameter of 2.5 m, 150 kg, the tentacles as long as 37 m
- Honey fungus of the species Armillaria ostoyae: spanning 8.9 km² of area, est. 605 tons, age 2400 years
- Phellinus ellipsoideus, a fungus with the largest fruiting body masses up to 500 kg
"Relative scale" by TimVickers - Wikimedia

PP1CS SoSe 17
Tree of Life

Laura A. Hug1, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas and Jillian F. Banfield

A highly resolved Tree of Life, based on completely sequenced genomes [1]. The image was generated using iTOL: Interactive Tree Of Life [2], an online phylogenetic tree viewer and Tree Of Life resource. Eukaryotes are colored red, archaea green and bacteria blue.


Cell Morphology - Terms

- Membrane: A membrane is a selective barrier. Biological membranes include cell membranes (outer coverings of cells or organelles that allow passage of certain constituents) and consist of a phospholipid bilayer.

- cytoplasm: semi-liquid matrix/medium inside the cytoplasm membrane

- cytoplasm(atic) membrane: the membrane defining the cell and separate inside from outside
Cell Morphology - Terms

- **cell compartment**: a region inside the cell mostly enclosed by a membrane, structural unit

- **cell organelle**: a special type of compartment most likely evolved from bacterial endosymbionts

- **nucleus**: if present, a membrane enclose volume of the cell which contains (most of) the genomic material

- **cell wall**: some cells may have an extra-cellular, rigid layer composed out of various polymers
Us (humans), plants and animals

Cells

Bacteria and Archaea

"Celltypes" by Science Primer - Wikimedia

PP1CS SoSe 17
Prokaryotic Cell

- no membrane enclosed compartments
- genomic material located in nucleoid region
- no organelles
Eukaryotic Cell

- membrane enclosed compartments
- genomic material is encapsulated in the nucleus
- organelles: mitochondria, chloroplasts
- further compartments: endoplasmatic reticulum, Golgi-apparatus
- ...somes: Peroxisome, Lysosome, ... . Special purpose membrane vesicles.
Hydro...

- hydrophilic: likes to interact with water molecules because of electrical (partial) charges
- hydrophobic: tries to avoid interaction with water because of no distinct charged points
Lipo...

- lipophilic: likes to interact with fatty/unpolar molecules because of no distinct charged points
- (lipophobic: tries to avoid interaction with fatty/unpolar molecules because of electrical (partial) charges)
Biomolecules et al.

- typically polymers, i.e. constructed from many identical or similar residues
- occur also in hybrid forms
- mass range of kDa to MDa
- can form complexes
- can contain hydrophilic as well as lipophilic parts (amphiphilic)
Carbohydrates

- biological molecule consisting of carbon (C), hydrogen (H) and oxygen (O) atoms, usually with a hydrogen–oxygen atom ratio of 2:1 (as in water): $C_m(H_2O)_n$ where $m$ could be different from $n$

- hydrophilic

- mono-, di-, oligo-, polysaccharide(s)

- mono- and disaccharides: typically called sugars

- important polysaccharides: starch, glycogen, cellolose, chitin
Carbohydrates

- biologically important: $C_5$ (pentoses) and $C_6$ (hexoses) sugars

- purposes:
  - fuel (e.g. glucose)
  - energy storage (starch, glycogen)
  - metabolites serves as substrates for other molecules
  - structural components (e.g. in nucleic acids)
  - modifications for proteins and lipid (e.g. glycoproteins, e.g. Hemagglutinin, the protein to which the flu virus docks)
Lipids

- In biology, lipid is a loosely defined term for substances of biological origin that are soluble in nonpolar solvents.
- Include fats, waxes, sterols, fat-soluble vitamins (such as vitamins A, D, E, and K), various glycerolipids.
- Functions of lipids include storing energy, signaling and acting as structural components of cell membranes.
Glycerolipids

- contains a glycerol as “bridge” with three docking slots and (hydrophobic) fatty acids (14-24 C)
- three fatty acids: triglycerides, storing energy
- two fatty acids, one (hydrophilic) phosphate group: phospholipid, major building block of membranes
Structure of the Cell Membrane

Outside of cell

Lipid Bilayer

Proteins

Carbohydrate chains

Transport Protein

Phospholipids

Inside of cell (cytoplasm)

4nm
Nucleic Acids

- directed (5’->3’), filamentous molecules
- polymers of nucleotides (base+sugar+phosphate)
- stores and transmits genetic information
- can adopt well defined geometry in double helix (B-DNA: 2.0 nm)
- hydrophobic nucleobases on the inside
- hydrophilic backbone: phosphate and sugar
Nucleic Acids

- RNA: typical single stranded but can adopt complex secondary structure with itself or other RNA molecules
- DNA: typical double stranded (double helix) formed by a strand with its reverse complement
- Bonds: non-covalent Hydrogen bonds (weak)

taken from https://en.wikipedia.org/wiki/Nucleic_acid_structure
Nucleic Acids

- the two strands are anti-parallel
- each “Watson-Crick” canonical base pair has diameter of 2nm
- one turn takes 10 nucleotides and 3.4 nm
- major and minor groove

taken from https://en.wikipedia.org/wiki/Nucleic_acid_structure
Proteins

- Polymerized amino acids
- Linear filaments of amino acids
- Only L-amino acids used in biological proteins
- "Work horse" or "machinery of life"
- Many functions: catalysis of chemical reactions (enzymes), structural elements (e.g. collagen fibers)
- Sensing (light, chemical signals)
- Immune system and more
Genetics

- Originally developed from observation of different phenotypes by Gregor Mendel (19th Century) -> statistics

- Today: based on the access to underlying molecular mechanisms, like sequencing and \textit{in-vitro} recombination
Central Dogma of Molecular Biology

- Describes the flow of information from DNA ➔ RNA ➔ Protein

1. DNA Replication
2. Transcription
3. Translation
4. Reverse Transcription (Viruses)
5. RNA Replication (Viruses)


PP1CS SoSe 17
Replication

- Duplication of DNA molecules
- Replication is highly controlled and starts at defined points (called origin or ori)
- for each strand in the double helix there is a complementary strand synthesized
- the existing single strand serve as template
- synthesis takes place in 5’->3’ direction
- carried out by multi-protein complex
Transcription

- “Creation of working copy of genes”
- Synthesis of single stranded RNA with DNA as template starting after a special region (Promotor)
- in Eukaryotes this takes place in the Nucleus
- before Translation (m)RNA undergoes several maturation steps
- carried out by a multi-protein complex
- (x)RNA: x denotes the function, m: translated into protein, r: component of the ribosome, s(n)...: regulatory functions, t: amion acid carrier for translation

PP1CS SoSe 17
Translation

- Conversion of the information coded in a mRNA into a protein (amino acid sequence), i.e. actual protein synthesis (polymerization of amino acids)
- by default: the sequences of a mRNA and a protein are noted in the same direction
- carried out in the cytosol or at the endoplasmatic reticulum (ER)
- carried by ribosomes (multi protein/RNA complex)
- requires a ribosome bindings site
- starts with firsth AUG codon
Gene – Definition

- dictionary.com: The basic physical unit of heredity; a linear sequence of nucleotides along a segment of DNA that provides the coded instructions for synthesis of RNA, which, when translated into protein, leads to the expression of hereditary character.

- wikipedia: a broad, modern working definition of a gene is any discrete locus of heritable, genomic sequence which affect an organism's traits by being expressed as a functional product or by regulation of gene expression

  -> this comprises the protein coding regions as well as non-coding and regulatory regions
Gene – Structure

# Genetic Code

- Translation table to translate a nucleic acid sequence into an amino acid sequence

<table>
<thead>
<tr>
<th>1st base</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU</td>
<td>(Phe/F) Phenylalanine</td>
<td>UAU</td>
<td>Tyr(Y) Tyrosine</td>
</tr>
<tr>
<td></td>
<td>UUC</td>
<td></td>
<td>UAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUA</td>
<td></td>
<td>UAA</td>
<td>Stop (Ochre)</td>
</tr>
<tr>
<td></td>
<td>UUG</td>
<td>(Leu/L) Leucine</td>
<td>UAG</td>
<td>Stop (Amber)</td>
</tr>
<tr>
<td>C</td>
<td>CUU</td>
<td></td>
<td>CAU</td>
<td>His(H) Histidine</td>
</tr>
<tr>
<td></td>
<td>CUC</td>
<td></td>
<td>CAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUA</td>
<td></td>
<td>CAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUG</td>
<td></td>
<td>CAG</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>AUU</td>
<td>(Ile/I) Isoleucine</td>
<td>AUC</td>
<td>Asn(N) Asparagine</td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td></td>
<td>AAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUA</td>
<td></td>
<td>AAG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUA</td>
<td>Met(M) Methionine</td>
<td>ACA</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GUU</td>
<td>(Val/V) Valine</td>
<td>GUC</td>
<td>Thr(T) Threonine</td>
</tr>
<tr>
<td></td>
<td>GUC</td>
<td></td>
<td>GCC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUA</td>
<td></td>
<td>GCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUG</td>
<td>Ala(A) Alanine</td>
<td>GAG</td>
<td>Asp(D) Aspartic acid</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Glu/E) Glutamic acid</td>
</tr>
</tbody>
</table>

### Amino acids
- Biochemical properties: nonpolar, polar, basic, acidic
- Taken from [https://en.wikipedia.org/wiki/Genetic_code](https://en.wikipedia.org/wiki/Genetic_code)
Inheritance / Mutations

- Inheritance happens when a parent transfers genetic material to an offspring individual
- Vegetative reproduction: the genetic material of parent and child is assumed to be identical (clones)
- Sexual reproduction: the genetic material of the child is a combination of parental genetic material
Inheritance / Mutations

- copy errors lead to the rise of evolution (via the constant iteration of mutation and selection)

- mutation types:
  - (long range) rearrangements
  - point mutations
Rearrangements

- rearrangement of DNA segments (real molecule fragments) due to a recombination process in an undesired fashion (wrong order, loss of fragments, ...)
- induced by: meiosis (reduction fission during germ cell genesis), DNA damage (chemical, radiation), virus infections
- either loss of complete gene (functions) or loss of controls (\(\rightarrow\) uncontrolled proliferation,...)
Point Mutations

- change of a single DNA residue
- Frame-shift: gain or loss of a residue (single insertion or deletion)
- Substitution:
  - transition: A<->G, C <-> T
  - transversion: A,G <-> C,T
- mostly due to replication errors, facilitated by chemicals and radiation
Effect of Point Mutations

- **Frame Shifts:** Depending on the position this might result in a complete loss of the gene function, rather unpredictable

- **Substitutions:**
  - silent: residue is functional or yields a redundant codon
  - missense: leads to the change of an amino acid
  - nonsense: introduced a stop codon, further effect depends on the specific position
  - affects a splicing site
Gene Regulation

- Regulations takes place on several levels:
  - transcription rate
  - stability of mRNA
  - translation
Metabolism / Physiology

- Anabolism: Reactions aim to the synthesis of new substance
- Catabolism: Reactions aim to the degradation of substance typically for energy generation or removal of damaged / aged structures
- steady state / equilibrium: degradation and synthesis rates are balanced, i.e. the system is dynamically stable (not statically!)
Metabolism / Physiology

- Housekeeping or constitutive genes/ expression: all (mandatory) efforts to guarantee vital functions
- Facultative: optional activation/function as the specific conditions allow or recommend it
- Enzymes: Proteins which catalyse a chemical reaction
- Prosthetic group: typically covalent bound non-protein compound of an enzyme needed for function
- Co-enzyme/-factor: as above, not covalently bound
Proteins

- more about proteins and amino acids in the dedicated exercise