Genes, Proteins and Chemistry

Protein Prediction
Part 1: Structure
TOC today

- Biology introduction
  - Organisms
  - Genes
  - Central dogma

- Protein introduction
  - Amino acids
  - Protein structure
  - Bonds & energies

- NEXT lecture (Thursday after easter)
  - 3D comparisons
Life is diverse

Trilobite Bergeroniellus spinosus
Lena River Gorge, Siberia
http://www.emory.edu/COLLEGE/PHYS/research/ichnology
Descriptive definition of life

- Organization - unit: cells
- Growth
- Adaptation
- Response to stimuli
- Reproduction
- Metabolism (transfer of energy)
- Homeostasis (regulation of internal environment to maintain constant state)
Descriptive definition of life

- Organization - unit: cells
- Growth
- Adaptation
- Response to stimuli
- Reproduction
- Metabolism (transfer of energy)
- Homeostasis (regulation of internal environment to maintain constant state)

Viruses: life?
Mimivirus

- dsDNA virus
- 90% coding capacity
- 10% Junk DNA
- 1.2 million base pairs
- ~911 protein coding genes
- additional genes (inc. aminoacyl tRNA synthetases: sugar, lipid, and amino acid metabolism)

© Wikipedia

Central dogma of molecular biology

- DNA Polymerase (DNA → DNA)
- RNA Polymerase (DNA → RNA)
- Ribosome (RNA → Protein)

© Burkhard Rost (TU Munich)
DNA sequencing

© yourgenome.org
Translation in action

www.dnalc.org
## Genome sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>470</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>1,740</td>
</tr>
<tr>
<td>Methanococcus jannaschi</td>
<td>1,738</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4,288</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae - yeast</td>
<td>6,600</td>
</tr>
<tr>
<td>Drosophila melanogaster - fruit fly</td>
<td>13,600</td>
</tr>
<tr>
<td>Caenorhabditis elegans - worm</td>
<td>19,000</td>
</tr>
<tr>
<td>Arabidopsis thaliana - mustard</td>
<td>26,735</td>
</tr>
<tr>
<td>Oryza sativa - rice</td>
<td>50,000</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>* Estimate from 1999</td>
</tr>
</tbody>
</table>

* Estimate from 1999
## Genome sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>470</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>1,740</td>
</tr>
<tr>
<td>Methanococcus jannaschi</td>
<td>1,738</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4,288</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae - yeast</td>
<td>6,600</td>
</tr>
<tr>
<td>Drosophila melanogaster - fruit fly</td>
<td>13,600</td>
</tr>
<tr>
<td>Caenorhabditis elegans - worm</td>
<td>*21,000</td>
</tr>
<tr>
<td>Arabidopsis thaliana - mustard</td>
<td>*26,735</td>
</tr>
<tr>
<td>Oryza sativa - rice</td>
<td>*50,000</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>*25,000</td>
</tr>
</tbody>
</table>

* Estimates from 2010
Now we know it all?

Like for every good manual:
You hardly ever find what you look for;
when you find it, it is difficult to understand

http://static.open.salon.com
Central dogma of molecular biology

DNA Polymerase

DNA

replication
(DNA -> DNA)

RNA Polymerase

RNA

transcription
(DNA -> RNA)

Ribosome

translation
(RNA -> Protein)

Protein

Function

Structure

dhorspool@en.wikipedia
Central dogma of molecular biology

DNA → RNA → Protein

information, code, library, manual → intermediate step → machinery of life
FANTOM 3: Functional annotation of mouse

Science 2005 309:1559-63
The Transcriptional Landscape of the Mammalian Genome

Numbers:
- Total transcripts: 181,047
- New protein-coding transcripts: 16,247
- New proteins: 5,154
- Multiple splice variants: 65%
- 1.35 5’ start sites for each 3’ end
- 1.83 3’ ends for each 5’ end
Central dogma of molecular biology

DNA  RNA  Protein
information, code, library, manual
intermediate step
machinery of life

Still the central dogma, but we know that reality is more complicated
Data deluge: what do we want?

- DNA
- ORF
- Protein
- Active protein
- Domains = smallest functional / structural subunits
- 3D structure
- Function
  - cellular function
  - physiological function
  - substrate binding sites
  - protein-protein interfaces
  - activity
  - specificity
  - docking
  - localisation

© Burkhard Rost (TU Munich)
## Data deluge: what do we have?

<table>
<thead>
<tr>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;180</td>
<td>entire organisms</td>
</tr>
<tr>
<td>&gt;14,000,000</td>
<td>genes (GenBank)</td>
</tr>
<tr>
<td>&gt;50,000,000</td>
<td>proteins (TrEMBL)</td>
</tr>
<tr>
<td>&gt;70,000</td>
<td>domains (DALI)</td>
</tr>
<tr>
<td>&gt;90,000</td>
<td>structures (PDB)</td>
</tr>
<tr>
<td>100,000</td>
<td>annotations (SWISS-PROT)</td>
</tr>
</tbody>
</table>

- cellular function
- physiological function
- substrate binding sites
- protein-protein interfaces
- activity
- specificity
- docking
- localisation

Exon | Intron
---|---

Expressed?
Growth of protein sequence data

- # entries in TREMBL (total sequence number)
- 1E6
- 2E6
- 3E6
- 4E6
- 5E6
- 1996
- 2000
- 2004
- 2008
- 2012

- # entries in Swissprot (curated sequences)
- 5E5
- 1E5
- 6E4 known structures (ca 10 %)
What are proteins?
A gallery of proteins (2)

Proteins are extremely diverse

David Goodsell
Protein functions

- **Defense** (e.g. antibodies)
- **Structure** (e.g. collagen)
- **Enzymes – metabolism, catabolism**
- **Communication / Signaling** (e.g. insulin)
- **Ligand binding / Transport** (e.g. hemoglobin)
- **Storage** (e.g. ferritin)
Proteins = gene products = machinery of life

From the book: “DNA: The Secret of Life” by James Watson and Andrew Berry
Amino acids

Twenty-One Amino Acids

A. Amino Acids with Electrically Charged Side Chains
   Positive
   - Arginine (Arg)
   - Histidine (His)
   - Lysine (Lys)

   Negative
   - Aspartic Acid (Asp)
   - Glutamic Acid (Glu)

B. Amino Acids with Polar Uncharged Side Chains
   - Serine (Ser)
   - Threonine (Thr)
   - Asparagine (Asn)
   - Glutamine (Gln)

C. Special Cases
   - Cysteine (Cys)
   - Selenocysteine (Sec)
   - Glycine (Gly)
   - Proline (Pro)

D. Amino Acids with Hydrophobic Side Chain
   - Alanine (Ala)
   - Isoleucine (Ile)
   - Leucine (Leu)
   - Methionine (Met)
   - Phenylalanine (Phe)
   - Tryptophan (Trp)
   - Tyrosine (Tyr)
   - Valine (Val)

pKa Data: CRC Handbook of Chemistry, v. 2010

Dan Copson, Department of Medical Biophysics, University of Toronto 2009
© Burkhard Rost (TU Munich)
Proteins - genetic code

From the book: “DNA: The Secret of Life” by James Watson and Andrew Berry © Burkhard Rost (TU Munich)
Kingdoms similar in amino acids usage

| Kingdom          | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
| Archae           |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 15%| 10%| 5% |   |
| Prokaryotes      |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 10%| 5% | 5% |   |
| Eukaryotes       |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | 10%| 5% | 5% |   |

| Number of codons | 4 | 2 | 2 | 2 | 2 | 4 | 2 | 3 | 2 | 6 | 1 | 2 | 4 | 2 | 6 | 6 | 4 | 4 | 1 | 2 |
| Hypothetical age  | 1 | 15| 4 | 6 | 16| 1 | 17| 12| 13| 8 | 18| 10| 7 | 14| 11| 5 | 9 | 3 | 20| 19|   |

Organisms *dissimilar* in codon usage

Codontable 1 (red):
    Homo_sapiens

Codontable 2 (black):
    Escherichia coli K12

Mean difference: 12.02 %
# Amino acid structure

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>H₂C—CO₂H</td>
<td>Gly, G</td>
</tr>
<tr>
<td>Alanine</td>
<td>H₃C—CO₂H</td>
<td>Ala, A</td>
</tr>
<tr>
<td>Valine</td>
<td>H₃C—CO₂H</td>
<td>Val, V</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cysteine</td>
<td>H₅S—CO₂Hₙ</td>
<td>Cys, C</td>
</tr>
<tr>
<td>Methionine</td>
<td>H₃C—S—NH₂</td>
<td>Met, M</td>
</tr>
<tr>
<td>Lysine</td>
<td>H₂N—CH₃—CO₂Hₙ</td>
<td>Lys, K</td>
</tr>
</tbody>
</table>

**Predominant Species at pH=6.0**

- Alanine: pI=6.01
- Isoleucine: pI=6.02

Images: [https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/proteins.htm](https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/proteins.htm)
Excursion: Physical Chemistry
Covalent bond

H·H

H: H

H - H

© Burkhard Rost (TU Munich)
Covalent bond – energy

© http://www.webchem.net/notes/chemical_bonding/covalent_bonding.htm
Double bond

Bond energy:
C-C 347 kJ/mol
C=C 611 kJ/mol

Energy barrier to rotation:
C-C 13 - 26 kJ/mol
C=C 264 kJ/mol

© Burkhard Rost (TU Munich)
Delocalisation

$\text{H} - \text{C} - \text{O}^{\ominus}$

$\text{H} - \text{C} - \text{O}^{\ominus}$

A

$\text{H} - \text{C} - \text{O}^{\ominus}$

$\text{H} - \text{C} - \text{O}^{\ominus}$

B

$\text{H} - \text{C} - \text{O}^{\ominus}$

$\text{H} - \text{C} - \text{O}^{\ominus}$

$\text{H} - \text{C} - \text{O}^{\ominus}$

$\text{H} - \text{C} - \text{O}^{\ominus}$

© Burkhard Rost (TU Munich)
Polar bonds

Difference in electronegativity
(= potential to attract electrons)

⇒ Polar bond, dipole
Ionic interactions

\[ V_{21} = \frac{1}{4\pi \varepsilon_0} q_2 \frac{1}{r} \]

Coulomb potential

Lenard Jones Potential

\[ V(r) = 4\epsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right] \]

Attractive \(-B/r^6\)

Repulsive \(+A/r^{12}\)

## Strength of non-bonded interactions

### Interactions in gas phase

<table>
<thead>
<tr>
<th>Molecules</th>
<th>ΔG [kJ/mol]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH₄ ... CH₄</td>
<td>-2</td>
</tr>
<tr>
<td>C₂H₆ ... C₂H₆</td>
<td>-10</td>
</tr>
<tr>
<td>H₂O ... H₂O</td>
<td>-22</td>
</tr>
<tr>
<td>NH₃ ... NH₃</td>
<td>-18</td>
</tr>
<tr>
<td>Na⁺ ... H₂O</td>
<td>-90</td>
</tr>
<tr>
<td>NH₄⁺ ... CH₃COO⁻</td>
<td>&lt;-400</td>
</tr>
</tbody>
</table>

Source: *Wirkstoffdesign: Entwurf und Wirkung von Arzneistoffen* by Gerhard Klebe
Non-covalent interactions in proteins

Covalent (electron sharing) < Hydrogen bond < Van der Waals

1 \quad 1/10 \quad 1/100-1/1000

From Wikipedia
Salt bridges

Arg, Lys, (His) -> Glu, Asp

From Wikipedia, Author: Chem540f09grp6
Salt bridges

Arg, Lys, (His) -> Glu, Asp

π-π interactions


Trp, Tyr, Phe -> Trp, Tyr, Phe
π-π interactions

PDBid: 2eve

Trp, Tyr, Phe -→ Trp, Tyr, Phe
Summary of typical interactions

• Hydrogen bonds

• Ionic interactions / salt bridges

• Metal complexes

• Cation - π
Back to proteins
## Amino acid structure

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Abbreviations</th>
<th>Name</th>
<th>Formula</th>
<th>Abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>H₂C(_\text{O})OH</td>
<td>Gly G</td>
<td>Cysteine</td>
<td>HS(<em>\text{C}</em>\text{O})OH</td>
<td>Cys C</td>
</tr>
<tr>
<td>Alanine</td>
<td>H₃C(_\text{O})OH</td>
<td>Ala A</td>
<td>Methionine</td>
<td>H₃C(<em>\text{S}</em>\text{O})OH</td>
<td>Met M</td>
</tr>
<tr>
<td>Valine</td>
<td>H₃C(_\text{O})OH</td>
<td>Val V</td>
<td>Lysine</td>
<td>H₂N(<em>\text{C}</em>\text{O})OH</td>
<td>Lys K</td>
</tr>
</tbody>
</table>

Predominant Species at pH=6.0

- Alanine pI=6.01
- Isoleucine pI=6.02

Images: https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/proteins.htm

© Burkhard Rost (TU Munich)
Side chain properties
Side chain properties

- Alanine
- Arginine
- Asparagine
- Aspartic Acid
- Cysteine
- Glycine
- Glutamine
- Glutamic Acid
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine
Negatively charged amino acids

- Aspartic Acid
- Glutamic Acid

- Alanine
- Arginine
- Cysteine
- Glycine
- Glutamine
- Histidine
- Isoleucine
- Leucine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine
amino acids

- Alanine
- Arginine
- Asparagine
- Aspartic Acid
- Cysteine
- Glycine
- Glutamine
- Glutamic Acid
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine
Polar amino acids

Ala-NH₂\(^2\)-CH₃
Arg-H-NH₂\(^2\)-CH\(_2\)CH\(_2\)CH\(_2\)NHCH\(_2\)H
Cys-H-NH₂\(^2\)-CH\(_2\)SH
Glu-H-NH₂\(^2\)-CH\(_2\)CH\(_2\)CH\(_2\)NHCH\(_2\)H
Gln-H-NH₂\(^2\)-CH\(_2\)CH\(_2\)CH\(_2\)NHCH\(_2\)H
His-H-NH₂\(^2\)-CH\(_2\)\[N\]
Ile-H-NH₂\(^2\)-CH\(_2\)CH\(_2\)CH\(_3\)
Leu
Met-H-NH₂\(^2\)-CH\(_2\)CH\(_2\)SCH\(_3\)
Phe-H-NH₂\(^2\)-CH\(_2\)\[C₆H₄\]
Prl-H-NH₂\(^2\)-CH\(_2\)CH\(_2\)CH\(_2\)NHCH\(_2\)H
Ser-H-NH₂\(^2\)-CH\(_2\)OH
Thr-H-NH₂\(^2\)-CH\(_2\)\[C₅H₅N\]
Tyr-H-NH₂\(^2\)-CH\(_2\)\[C₆H₄\]
Val-H-NH₂\(^2\)-CH\(_3\)CH\(_3\)
amino acids
Peptides

\[
\text{R-OH} + \text{H}_2\text{N-}R' \rightarrow \text{R-NH-}R' + \text{H}_2\text{O}
\]

N-terminus

C-terminus

C–N double bond character in amide (peptide) bonds

planar peptide bond segments

Images: https://www2.chemistry.msu.edu/faculty/reusch/virttxtjml/protein2.htm
Protein sequence

>gi|16128674|ref|NP_415226.1| potassium translocating ATPase, subunit A [Escherichia coli K12]
MAAQGFLLLATFLLVLMVLARPLGSGLARLINDIPLPGTTGVERVLFRALGVSR
EMNWKQYLCAILGLNMLGLAVLFFMLLGQHYLPLNPQQLPGLSWDLALNTAVSF
VTNTNWQSYSGETTLSYFSQMAGLTQNVNFLSAAASGIAVIFALIRAFTRQSMSTL
GNAWVDLLRITLWLVPVALLIALFIQQQGALQNFYPYQAVNTVEGAQOQLPMG
PVASQEAIKMLGTNGGGFFANSSHPFENPTALTNTFQMLAIFLIPTALCFAFG
EVMGDRRQGRMLLMWAMSVIFVICVGVMWAVEVQGNPHLLALGTDSSINMEGKES
RFGVLYSSLFAVVTAAASCIAVIAMHDSFTALGGMVPMWLMQIGEVVFVGGVGSG
LYGMMLFLVLLAVFIAGLMIGRTPEYLKIDVREMKLTLAILVTPTLVLMGAA
LAMMTDAGRSAMLNPGPHGFSEVLYAVSSAANNNSAFAGLSANSPFWNCLLAFC
CMFVGRFGVIIIPVMAIAAGSVKLSSQASSGTLPHTGPLLFFVGLLIGTVLVLGAL
TFIPALALGPVAEYLS
Protein structure

Branden & Tooze, Introduction to Protein Structure
Secondary structure: $\alpha$-Helix

Dihedral Angles

\[
\begin{align*}
\Phi &= C-N-C^\alpha C \\
\Psi &= N-C-C^\alpha N
\end{align*}
\]

one turn of Ala-Thr-Gly-Ala-Phe-Leu-Ala-Phe-Ser-Ile-Gly

Images: https://www2.chemistry.msu.edu/faculty/reusch/virtxtjml/protein2.htm
Secondary structure: $\beta$-Sheet

Images: https://www2.chemistry.msu.edu/faculty/reusch/virtxtjml/protein2.htm
Spot the secondary structure
Spot the secondary structure
Spot the secondary structure

from 3goe, structure at 0.97 Å resolution
GN Ramachandran  (Gopalasamudram Narayana Iyer Ramachandran)

- 8 October 1922 - 7 April 2001
- 1942: Electrical engineering -> physics (Master with Raman) / optics
- 1949: PhD: Cavendish Lab in Cambridge: crystallography
- 1952: Madras: shift to biological molecules: triple helical structure of collagen
- 1963 (JMB): analysis of backbone confirmations: Ramachandran plots
Secondary structure: Ramachandran

from: http://boscoh.com/protein/the-mysterious-regions-of-the-ramachandran-plot

Ramachandran plot (Ramachandran, 1963, JMB 7:99)

© Burkhard Rost (TU Munich)
Protein structure cartoons
Protein structure

Primary

Secondary

Tertiary

Quaternary

From secondary to tertiary structure

Branden & Tooze, Introduction to Protein Structure
Disulfide bridges

Cys -> Cys

From Wikipedia, author: Benjah-bmm27

From Wikipedia, author: Burkhard Rost (TU Munich)
Reality and images
Marco Punta contributed the slides

- PhD in Trieste (MD for membrane proteins)
- Postdoc @ Columbia Univ in the City of New York (contact predictions)
- Senior scientist in NYCOMPS (Target selection for membrane proteins)
- IAS Fellow @ TUM
- Project manager @ Pfam @ Sanger Inst. Hinxton (Cambridgeshire)
Georges Braque - Houses at L’Estaque
Where is that?

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
Mycoplasma genitalium

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
**Mycoplasma genitalium**

**Protein synthesis (labels in black)**
1. DNA
2. DNA polymerase
3. single-stranded-DNA binding protein (protects single-stranded portions during replication)
4. RNA polymerase
5. messenger RNA
6. ribosome
7. transfer RNA (in pink) and elongation factor Tu (in blue)
8. elongation factor Tu and Ts
9. elongation factor G
10. aminoacyl-tRNA synthetases
11. topoisomerases
12. Rec system for DNA repair: a) RecA, b) RecBC
13. chaperonin GroEL (helps folding of new proteins)
14. proteasome ClpA (destroys old proteins)

**Enzymes for energy production (labels in red)**
15. glycolytic enzymes
16. pyruvate dehydrogenase complex

**Membrane proteins (labels in blue)**
17. ATP synthase
18. secretory proteins
19. sodium pump
20. zinc transporter
21. magnesium transporter
22. ABC transporter
   (different ABC transporters transport different types of molecules-ABC is short for "ATP-binding cassette")
23. magnesium transporter
24. lypoglycan
   (long carbohydrate chains connected to lipid in the membrane)

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
Doyle et al. (1998) Science 280:69-77 - The structure of the potassium channel: molecular basis of K+ conduction and selectivity
Alcohol dehydrogenase (ADH)

http://www.proteopedia.org/wiki/images/7/7b/1htb2.png

homodimer ADH5


http://upload.wikimedia.org/wikipedia/commons/thumb/a/a5/Protein_ADH5_PDB_1m6h.png/800px-Protein_ADH5_PDB_1m6h.png
Umberto Boccioni - Dynamism of a soccer player
Different levels of abstraction

(a) Wu et al. unpublished

(b) Photograph: Filippo Monteforte/AFP/Getty Images
“components” of protein structure: domains
Domain from introns?

RNA splicing

Gene product = protein

domais = introns?
Domain merger

prokaryote P, protein A

prokaryote P, protein B

prokaryote P2, protein C
Domains
Domains
Domains
3D modules

Multiple 3D alignment identifies consensus secondary structure
Guessing domains from sequence

protein A
protein B
protein C
protein D
protein E
protein F
Guessing domains from sequence

protein A
protein B
protein C
protein D
protein E
protein F
Guessing domains from sequence

- protein A
- protein B
- protein C
- protein D
- protein E
- protein F

domain 1  domain 2
Problem: biology is not that simple
Domain length distribution

some common assumptions and facts about proteins
Some facts about proteins

- How many in human?
  - 20-25K in human

- How long are they?
  - ~35-30k, median around 400

- Do they consist of units?
  - most proteins have more than 2 domains

- Longer proteins are rarer than shorter ones?
  - Yes

- Prokaryotic proteins shorter than eukaryotic?
  - Yes
Kingdoms similar in length

TOC today

- Biology introduction
  - Organisms
  - Genes
  - Central dogma
- Protein introduction
  - Amino acids
  - Protein structure
  - Bonds & energies
- NEXT lecture (Thursday after easter)
  - 3D comparisons