Informatics et al.
CONTACT: Lothar Richter richter@rostlab.org

Announcements

Videos: YouTube / www.rostlab.org

THANKS:

EXERCISES:

Special lectures:
• TBD - Thomas Hopf
• TBD - Jonas Reeb

No lecture:
• 05/10 Student assembly (SVV)
• 05/17 Ascension day
• 05/26 Whitsun holiday
• 06/04 Corpus Christi

LAST lecture: June 28 (followed by 2 wrap-up sessions)

Examen: June 30, 2016: lecture time room TBD
• Makeup: TBC: Oct 18 & 20, 2016 - lecture time
Protein Prediction - Part 1: Structure

Computational Biology 1
Definitions

- **Computational Biology**
  - biology replacing experiments by computers (include neurobiology, image processing)

- **Bioinformatics**
  - anything that has to do with storing and using the information about bio-sequences

- [http://www.rostlab.org/talks/](http://www.rostlab.org/talks/)
Who are we?
Burkhard Rost
Some Professional Positions

- **Professor** for Computational Biology, TUM, Dept. Computer Science
- **Fellow** Institute of Advanced Studies, TUM
- **Affiliate professor** @ WZW Freising

- **Affiliate professor** @ Columbia University in the City of New York, College for Physicians and Surgeons, Dept. Biochemistry and Molecular Biophysics
Study complex systems

  Master: Neuronal networks
- 1989-1995: PhD/Postdoc: Application of physics to Computational Biology, EMBL (European Molecular Biology Laboratory, Heidelberg)
- 1995-1996: European Bioinformatics Inst. (EBI), Cambridge, UK
- 1998: LION Biosciences
- 1998-2010: Professor at Columbia University, New York
  (tenure since 2005)
- 2009: Alexander von Humboldt professorship
- 2009-now: Professor at TUM, Chair for Comp. Biology
Little $, but attention, friends, travel, fun
My passion is my job

- Seminars on 5 continents
  Dec 2009 - 2011
  (TV in 4: USA, Mali, Uruguay, Germany)
- 189 invited talks in 27 countries
- 248 publications, 186 peer-reviewed
- >32,000 citations to our work;
  H-index > 65
- google.com “rost”

Nothing makes me happier than to be proven wrong
TOC today

- Biology introduction
  - Organisms
  - Genes
  - Central dogma

- Protein introduction
  - Some numbers/idea

- NEXT lectures
  - Building blocks
  - Domains
  - 3D comparisons
Nobel Prizes in Informatics?
2013 Nobel Prize in Chemistry

“for the development of multiscale models for complex chemical systems” (nobelprize.org)

Martin Karplus
Harvard Univ, USA & Univ Strasbourg, France

Michael Levitt
Stanford Univ, USA

Arieh Warshel
USC, Los Angeles, USA
Protein structure comparisons

**All-alpha**

3sdh

**All-beta**

1bww

**AlphaBeta**

1xne
Predict protein function
Classical physics vs quantum physics

Figure 1
Figure 1 Multi-copper-oxidase embedded in water.¹

Michael Levitt (AP): It’s sort of nice ... to see that computational science, computational biology is being recognized. ... It’s become a very large field and it’s always in some ways been the poor sister, or the ugly sister, to experimental biology.

Arieh Warshel (NobelPrize) [experimental results are] like seeing a watch and wondering how actually it works. ... what we developed is a way which requires a computer to look ... and then to eventually understand how exactly [a protein] does what it does.
Background
Life is diverse

Trilobite Bergeroniellus spinosus
Lena River Gorge, Siberia
http://www.emory.edu/COLLEGE/ENVS/research/ichnology
How does life work?

Leonardo Da Vinci (1452-1519)
Vitruve Luc Viatour (~1492)
How does life work?

Leonardo Da Vinci (1452-1519)
Vitruve Luc Viatour (~1492)
Common 2 Life?

David Pacchioli: Logic of Organisms

Trilobite Bergeroniellus spinosus
Lena River Gorge, Siberia
http://www.emory.edu/COLLEGE/ENVS/research/ichnology
Terms/facts

DNA

protein

RNA
Central dogma of molecular biology

DNA → RNA → Protein

Information, code, library, manual → Intermediate step → Machinery of life

4 nucleotides (GATC) → 4 nucleotides (GAUC) → 20 amino acids (ACDEFGHIKLMNPRSTVWY)

Transcription → Translation

© Wikipedia

© Burkhard Rost ROSTLAB
A gallery of proteins

DNA Polymerase

Myosin

Actin

Microtubule

Collagen

scale reduced
Images
What is similar?
What is similar?
What is similar?

silicone

© Wikipedia

glass

© http://www.lionsons.com
What is this?

Diamond

Graphite/coal

© Wikipedia
What is this?

Diamond

Graphite/coal

carbon

diamond lattice

© Wikipedia  © Burkhard Rost
Which molecule dominates our body weight?
Which molecule makes most of our body?

H2O - water

how much?
What elements make most of our body?

H2O - water

about 50-65% in adults
What elements make most of our body?
Which ones make up life?

<table>
<thead>
<tr>
<th>Element</th>
<th>Atomic Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen</td>
<td>65.00</td>
</tr>
<tr>
<td>Hydrogen</td>
<td>9.70</td>
</tr>
<tr>
<td>Body Weight</td>
<td>96.5</td>
</tr>
</tbody>
</table>

© Wikipedia
Which ones make up life?

<table>
<thead>
<tr>
<th>Element</th>
<th>Atomic Weight</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen (O)</td>
<td>16.00</td>
<td>65.90</td>
</tr>
<tr>
<td>Carbon (C)</td>
<td>12.01</td>
<td>18.60</td>
</tr>
<tr>
<td>Hydrogen (H)</td>
<td>1.01</td>
<td>9.70</td>
</tr>
<tr>
<td>Nitrogen (N)</td>
<td>14.01</td>
<td>3.20</td>
</tr>
<tr>
<td>Calcium (Ca)</td>
<td>40.08</td>
<td>1.80</td>
</tr>
<tr>
<td>Phosphorus (P)</td>
<td>31.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Body weight: 96.5%**

© Wikipedia
Which ones make up life?

- oxygen - O
- carbon - C
- hydrogen - H
- nitrogen - N
- calcium - Ca
- phosphorus - P
- potassium - K
- sodium - Na
- chlorine - Cl
- magnesium - Mg
- sulfur - S
- iron - Fe
- iodine - I

Body weight: 96.5
What is this?
What is this?

Escherichia coli
What is this?

What is this?

Escherichia coli /E coli

What is this?
What is this?

© Wikipedia
What is this?

Salmonella bongori

© Wikipedia
What is this?

What is this?

Avian or bird flu (virus)

What is this?

What is this?

HIV virus

What is this?
Velcro - invented by George de Mestral

George de Mestral (1907-1990)

- Swiss electrical engineer
- Invented velcro 1949

http://www.engr.sjsu.edu/WofMatE/polymers.htm

© Wikipedia
© Burkhard Rost
© ROSTLAB.
Velcro copied from plants

George de Mestral
(1907-1990)

- Swiss electrical engineer
- Invented velcro 1949

© Wikipedia

Burr (seed of fruit)

© Wikipedia
Bacteria differ in shape

© Wikipedia

© Burkhard Rost
ROSTLAB.
TUM
What do they have in common?
Common to bacteria: single cells

- Cocci: coccus, diplococci, encapsulated Pneumococcus, Staphylococci, streptococci, sarcina, tetrad
- Bacilli: coccobacillus, bacillus, diplobacilli, palisades, Streptobacilli
- Budding and appended bacteria: hypha, stalk

![Diagram of bacterial shapes and sizes](image)

- Eukaryotes
- Prokaryotes
- Viruses
- Proteins
- Small molecules
- Atoms
Cells

Eukaryote
- Membrane-enclosed nucleus
- Nucleolus
- Mitochondrion
- Ribosomes
- Cell Membrane

Prokaryote
- Nucleoid
- Capsule (some prokaryotes)
- Flagellum
- Cell Wall (in some eukaryotes)

© Wikipedia
Do we carry bacteria?

Leonardo Da Vinci (1452-1519)
Vitruve Luc Viatour (~1492)
we carry bacteria when we are sick
Our bodies are teeming with bacteria


© NATURE REVIEWS
MICROBIOLOGY 9:244-53,
COPYRIGHT 2011
Common?

Trilobite Bergeroniellus spinosus
Lena River Gorge, Siberia
http://www.emory.edu/COLLEGE/ENVS/research/ichnology
DNA/Gene

DNA
DEOXYRIBONUCLEIC ACID

Cytosine
Guanine
Adenine
Thymine

© Wikipedia

www.chemistryexplained.com
Where do we have genes?
Genes in nucleus and mitochondria

green: cell (actin)  red: mitochondria  blue: nucleus

© Wikipedia
Genes in nucleus and mitochondria

green: cell (actin)
red: mitochondria
blue: nucleus
Central dogma of molecular biology

DNA → RNA → Protein

information, code, library, manual

intermediate step

machinery of life

© Laszlo Kajan, TUM

© Burkhard Rost
1. what is the smallest building block of life that can replicate?
   • protein - quarks - cells - organelles
2. different cells in typical human?
   • 20 - 200 - 400 - 1000
3. what are the parts of cells called?
   • cellular parts - organelles - organs - celluoplasts
4. which part of cells is called the “powerhouse”?
   • nucleus - Golgi apparatus - mitochondria - Endoplasmic reticulum
5. what part of a plant cell is involved with photosynthesis?
   • mitochondria - nucleus - smooth reticulum - chloroplast
6. what is mitosis?
   • cell death - cellular respiration - cell division - cellular communication
7. who first used the term cell?
   • Aristotle - Captain Hooke - Robert Hooke - James Watson
8. how many elements are found in amounts larger than trace amounts (0.01%) in our bodies?
   • 92 - 48 - 13 - 11
9. when communities of living things interact with non-living things they are called ...?
   • population - community - biosphere - ecosystem
10. the most common molecule in the human body is?
1. what is the smallest building block of life that can replicate?
   • protein - quarks - cells - organelles
2. different cells in typical human?
   • 20 - 200 - 400 - 1000
3. what are the parts of cells called?
   • cellular parts - organelles - organs - celluoplasts
4. which part of cells is called the “powerhouse”?
   • nucleus - Golgi apparatus - mitochondria - Endoplasmic reticulum
5. what part of a plant cell is involved with photosynthesis?
   • mitochondria - nucleus - smooth reticulum - chloroplast
6. what is mitosis?
   • cell death - cellular respiration - cell division - cellular communication
7. who first used the term cell?
   • Aristotle - Captain Hooke - Robert Hooke - James Watson
8. how many elements are found in amounts larger than trace amounts (0.01%) in our bodies?
   • 92 - 48 - 13 - 11
9. when communities of living things interact with non-living things they are called ...?
   • population - community - biosphere - ecosystem
10. the most common molecule in the human body is?
    • H2O
Life: The Players
What is life?
Life is diverse

Trilobite Bergeroniellus spinosus
Lena River Gorge, Siberia
http://www.emory.edu/COLLEGE/ENVS/research/ichnology

David Pacchioli: Logic of Organisms
Can you define life?
Descriptive definition of life

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

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

Viruses = life?
## Central dogma as information flow

<table>
<thead>
<tr>
<th>General</th>
<th>Special</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA -&gt; DNA</td>
<td>Replication</td>
<td>RNA -&gt; DNA</td>
</tr>
<tr>
<td>DNA -&gt; RNA</td>
<td>Transcription</td>
<td>RNA -&gt; RNA</td>
</tr>
<tr>
<td>RNA -&gt; protein</td>
<td>Translation</td>
<td>DNA -&gt; protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td>protein -&gt; DNA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>protein -&gt; RNA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>protein -&gt; protein</td>
</tr>
</tbody>
</table>

Narayanese *Centraldogma nodetails* - Wikimedia
Central dogma of molecular biology

DNA Polymerase

DNA -> DNA

RNA Polymerase

DNA -> RNA

Ribosome

RNA -> Protein

Protein

© 2014 David Goodsell & RCSB Protein Data Bank

dhorspool@en.wikipedia
Codon wheel

Amino acid code

- A - Alanine
- C - Cysteine
- D - Aspartic acid
- E - Glutamic acid
- F - Phenylalanine
- G - Glycine
- H - Histidine
- I - Isoleucine
- K - Lysine
- L - Leucine
- M - Methionine
- N - Asparagine
- P - Proline
- Q - Glutamine
- R - Arginine
- S - Serine
- T - Threonine
- V - Valine
- W - Tryptophan
- Y - Tyrosine

Image: C. Brookes, European Bioinformatics Institute
Translation in action

www.dnalc.org
DNA sequencing

Key:
- C - Cytosine
- G - Guanine
- T - Thymine
- A - Adenine

© yourgenome.org

slide: Andrea Schafferhans

© Burkhard Rost

TUM
Next generation sequencing

Figure 1: Conceptual Overview of Whole-Genome Resequeencing

A. Extracted gDNA.
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.
C. Individual sequence reads are reassembled by aligning to a reference genome.
D. The whole-genome sequence is derived from the consensus of aligned reads.

Illumina: MiSeq

- run: 6 hours
- full capacity: raw: \(~5-10\) TB data / day
  reduced: 50-200GB/day

Illumina - Early 2012
Illumina: HiSeq 2000

- run: 1-8 days, 24 GB/day
- 1 human genome @ 30x / day
- full capacity: raw: 25 TB data / day, reduced: 100-500 GB
- 120 x 76 x 94 cm

BGI - Shenzhen
### Illumina HiSeq 2016

#### Table 1: HiSeq X System Sequencing Capacity

<table>
<thead>
<tr>
<th></th>
<th>HiSeq X Ten System</th>
<th>HiSeq X Five System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Number of Instruments</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Annual Genome Capacity</td>
<td>&gt; 18,000</td>
<td>&gt; 9000</td>
</tr>
<tr>
<td>Price per 30x Genome</td>
<td>&lt; $1000</td>
<td>&lt; $1500</td>
</tr>
</tbody>
</table>

~2TB/run (<3 days)
Haemophilus Influenzae (Pfeiffer’s bacillus)

Mycoplasma Genitalium

Saccharomyces cerevisiae

Caenorhabditis elegans, nematode - worm

Drosophila melanogaster fruit fly 195 authors
# 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>1740</td>
</tr>
<tr>
<td>Methanococcus jannaschi</td>
<td>1,738</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4,288</td>
</tr>
<tr>
<td>Sacharomyces 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>*100,000</td>
</tr>
</tbody>
</table>

* Estimate from 1999
Manual for human

☐ not Jan 1, 2000

☐ number of genes/proteins:
  Oct 1999
  (after >5 years):
  100,000
Manual for human

- not Jan 1, 2000

- number of genes/proteins:
  Oct 1999 (after >5 years):
  100,000

- Nov 1999: oops there are only 30,000
# Genome sizes

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

http://i42.tinypic.com
Amino acid sequence determines protein 3D structure

Christian Anfinsen
Nobel Prize in Chemistry 1972
Central dogma of molecular biology

DNA → RNA → Protein

- DNA: information, code, library, manual
- RNA: intermediate step
- Protein: machinery of life
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
What are proteins?
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)
Gallery of proteins 1

Slide: Andrea Schafferhans
Gallery of proteins 2

- Microtubule
- Rhinovirus
- Antibody
- Myosin
- Actin
- Collagen

slide: Andrea Schafferhans
© Burkhard Rost
ROSTLAB.
Protein sequence

>gi|16128674|ref|NP_415226.1| potassium translocating ATPase, subunit A [Escherichia coli K12]
MAAQGFLLIATFLVLMVLARPLGSGLARLINDIPLPGTTGVERVLFRALGVSDREMNWK
QYLCAILGLNMLGLAVLFFMLLGQHYLPLNPQQLPGLSWDALTAVSFVNTNTNWQSYSG
ETTL SYFSQMAGLTVQNFLSAASGIAVIFALIRAFTQRSMSTLGNAWVDLLLRITLWVLVP
VALLIALFFIQQGALQNFLPYQAVNTVEGAQQLLPMGPVASQEAIKMLGTNGGGFFNANS
SHPFNPTALTNFVQLAIFLIPTALCFAFGEVMGDRRQGRMLLWAMSVIFVICVGVVVMW
AEVQGPNPHLLALGTDSSINMEGKESRGFVGLVSSLFAVVTTAASCGAIVAMHDSFTALGGM
VPMWLMQIGEVVFVGGSGLYGMMLFVLLAVFIAGLMIGRTPEYLGKKIDVREMKLTLA
ILVTPTLVLMGAAALAMMTDAGRSMLNPGPHGFSEVLAYVSSAANNGSAFAGLSANSFP
WNCLLAFCMFVGRFGVIIPVMIAIGSLVSKKSQAASSGLPTHGPLFVGLLLGTVVLLVGA
LTFIPALALGPVAEYLS
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

- how many proteins known?
Bio-sequence data explodes

Number of entries in UniProtKB/TrEMBL

© Uniprot EBI Cambridgeshire

© Burkhard Rost

ROSTLAB.
Protein WITH annotations grow less

Number of entries in UniProtKB/Swiss-Prot

© Swiss-Prot Geneva/Lausanne
Raw protein sequence data vs. annotations

Number of entries in UniProtKB/TrEMBL

© Uniprot EBI Cambridgeshire
PDB: Proteins with experimentally known 3D structure

© PDB Rutgers Univ

© Burkhard Rost
ROSTLAB. TUM
PDB: known structure \(\sim 0.2\%\) of sequences

UniProt=all (0.2%) 100,000,000

Swiss-Prot=Annotation (20%) PDB=3D


© PDB Rutgers Univ © Burkhard Rost ROSTLAB. TUM
Proteins = gene products
= machinery of life

From the book: “DNA: The Secret of Life” by James Watson and Andrew Berry
Proteins - genetic code

From the book: “DNA: The Secret of Life” by James Watson and Andrew Berry
Kingdoms similar in amino acids usage

organisms DO differ in the codon usage

nucleotide  amino acid

GCT
GCA
GCC  →  A
GCG
Lecture plan (CB1 structure)

- 01: 04/11 Tue: no lecture
- 02: 04/13 Thu: no lecture
- 03: 04/19 Tue: Organization of lecture: intro into cells & biology
- 04: 04/21 Thu: Intro I - acids/structure - domains
- 05: 04/26 Tue: Intro 2 - domains - 3D comparisons
- 06: 04/28 Thu: Alignment 1
- 07: 05/03 Tue: Alignment 2
- 08: 05/05 Thu: SKIP: Ascension Day
- 09: 05/10 Tue: SKIP: student assembly (SVV)
- 10: 05/12 Thu: Comparative modeling
- 11: 05/17 Tue: SKIP: Whitsun holiday (05/15-17)
- 12: 05/19 Thu: Experimental structure determination / 3D -> 1D: Secondary structure assignment
- 13: 05/24 Tue: 1D: Secondary structure prediction 1
- 14: 05/26 Thu: SKIP: Corpus Christi
- 15: 06/31 Tue: 1D: Secondary structure prediction 2
- 16: 06/02 Thu: 1D: Transmembrane structure prediction 1
- 17: 06/07 Tue: 1D: Transmembrane structure prediction 2 / Solvent accessibility prediction
- 18: 06/09 Thu: 2D prediction 1
- 19: 06/14 Tue: 2D prediction 2 - Thomas Hopf
- 20: 06/16 Thu: 3D prediction / Nobel prize symposium
- 21: 06/21 Tue: 1D: Disorder prediction
- 22: 06/23 Thu: recap 1
- 23: 06/28 Tue: recap 2
- 24: 06/30 Thu: examen, no lecture
- 25: 07/05 Tue: examen alternative, no lecture
- 26: 07/07 Thu: examen, no lecture