Two challenges: Data growth & Data Loss

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady (Moore’s law)
- DNA sequence data is **doubling every 6-8 months** over the last 3 years and looks to continue for this decade

‘Oops, that link was the laptop of my former PhD student’

Nature news, 19 December 2013

Slide courtesy of Jaap Heringa
Bioinformatics Research & Data

- Sequence data & analysis pipelines

- Protein structure, dynamics & interactions
Comparative metagenomics (metatranscriptomics)

Sequencing

Bioinformatics & statistical analysis

9% of sequences found are from pathway \(a\)
5% function \(b\)
etc..

Sequencing

Bioinformatics & statistical analysis

12% of sequences found are from pathway \(p\)
10% pathway \(m\)
etc..

→ Healthy vs. Diseased →

Slide courtesy of Ali May
Orthologous gene group (KO)  
Function e.g. lactate dehydrogenase

Major metabolic pathways

Slide courtesy of Ali May
Result: gum disease subnetwork

7 pathways, 39 KOs (32 up- and 7 downregulated in disease)
Top 10 species that contribute most to the subnetwork

- Well-known periodontal pathogens, as well as some others.
- Species don’t express all KOs identified in a pathway.
- Community interaction through metabolites?

Slide courtesy of Ali May
**Workflow**

- Few platforms that support this kind of workflow
- Total ~1TB intermediate results: cannot be stored
  - Input and output need to be un/zipped on the fly

<table>
<thead>
<tr>
<th>Step</th>
<th>Data Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw sequence data</td>
<td>~350G</td>
</tr>
<tr>
<td>Quality Filtering</td>
<td>~340G</td>
</tr>
<tr>
<td>Contaminant filtering</td>
<td>~300G</td>
</tr>
<tr>
<td>Mapping</td>
<td>~200G</td>
</tr>
</tbody>
</table>

- This case, only ~2k CPUh because we can focus on small number of bacteria
  - Typically 100-1000M CPUh
- Total cost of experiment: ~€7000 (~€200 per sample)

Data & Coffee

15 April 2016

38 samples
Bioinformatics Research & Data

- Sequence data & analysis pipelines

- Protein structure, dynamics & interactions
Two protein molecules binding: $A + B \leftrightarrow AB$
- Calculate from simulation

May et al., Bioinformatics 30:326, 2014
Interaction depends on mutations

Interaction depends on mutations.

![Graph showing PMF vs. center of mass separation](image)

- PMF [kJ mol^{-1}]
- Center of mass separation $r$ [nm]

- WT
- Best
- Worst

PMF values range from -125 to 0 kJ mol^{-1}.

Center of mass separation ranges from 2.5 to 4 nm.
Interaction also depends on the timescale

![Graph showing PMF vs. Center of mass separation (r [nm]) with different timescales (2ns, 2μs).](image)

- **2ns**
  - Coarse-grained (from avg. MF of 20 sims., 2 ns)
  - All-atom (from avg. MF of 10 sims., 2 ns)
  - Experimental ΔG
  - Coarse-grained (from avg. MF of 3 sims., 2 μs)

**Center of mass separation (r [nm])**

**PMF [kJ mol⁻¹]**

**MP1-p14**
Interaction also depends on orientation

- For 15 protein pairs, in total 19,000 orientations
- 1.2M CPUh
- ~100T raw data → not stored, but we now want to analyze it... :-(
Bioinformatics Research & Data

- Sequence data & analysis pipelines
- Protein structure, dynamics & interactions
Two challenges: Data growth & Data Loss

- Computer speed and storage capacity is doubling every 18 months and this rate is steady (Moore’s law)
- DNA sequence data is doubling every 6-8 months over the last 3 years and looks to continue for this decade

‘Oops, that link was the laptop of my former PhD student’
Semantic interoperability through RDF (resource description framework)

Unit of information: the data triple

Linked data cloud

Data interoperability
Bioinformatics

data (& coffee)

Anton Feenstra

Data & Coffee
15 April 2016
Two challenges: Data growth & Data Loss

- Computer speed and storage capacity is doubling every 18 months and this rate is steady (Moore’s law)
- DNA sequence data is doubling every 6-8 months over the last 3 years and looks to continue for this decade

‘Oops, that link was the laptop of my former PhD student’

Slide courtesy of Jaap Heringa
Comparative metagenomics (metatranscriptomics)

Healthy vs. Diseased

Bioinformatics & statistical analysis

9% of sequences found are from pathway a
5% function b etc..

Sequencing

Bioinformatics & statistical analysis

12% of sequences found are from pathway p
10% pathway m etc.

Slide courtesy of Ali May
Orthologous gene group (KO)
Function e.g. lactate dehydrogenase

Major metabolic pathways

mRNA surveillance

RNA degradation

Oxidative phosphorylation

Protein processing in ER

Slide courtesy of Ali May

Orthologous gene group (KO)
Function e.g. lactate dehydrogenase

Major metabolic pathways

mRNA surveillance

RNA degradation

Oxidative phosphorylation

Protein processing in ER

Slide courtesy of Ali May
Result: gum disease subnetwork

7 pathways, 39 KOs (32 up- and 7 downregulated in disease)

Slide courtesy of Ali May

Data & Coffee 15 April 2016
Top 10 species that contribute most to the subnetwork

- Well-known periodontal pathogens, as well as some others.
- Species don’t express all KOs identified in a pathway.
- Community interaction through metabolites?

Slide courtesy of Ali May
Workflow

38 samples

raw sequence data

~350G

Quality Filtering

~340G

Contaminant filtering

~300G

Mapping

~200G

Analysis

Analysis

Analysis

- Few platforms that support this kind of workflow
- Total ~1TB intermediate results: cannot be stored
  - Input and output need to be un/zipped on the fly
- This case, only ~2k CPUh because we can focus on small number of bacteria
  - Typically 100-1000M CPUh
- Total cost of experiment:
  ~€7000 (~€200 per sample)

38 samples

~350G

Quality Filtering

~340G

Contaminant filtering

~300G

Mapping

~200G

Analysis

Analysis

Analysis

- Few platforms that support this kind of workflow
- Total ~1TB intermediate results: cannot be stored
  - Input and output need to be un/zipped on the fly
- This case, only ~2k CPUh because we can focus on small number of bacteria
  - Typically 100-1000M CPUh
- Total cost of experiment:
  ~€7000 (~€200 per sample)
Two protein molecules binding: $A + B \leftrightarrow AB$
- Calculate from simulation

May et al., Bioinformatics 30:326, 2014
Interaction depends on mutations

PMF [kJ mol\(^{-1}\)]

Center of mass separation \(r\) [nm]

WT  
Best  
Worst

PMF range: -125 to 0

WT  
Best  
Worst

WT

Interaction depends on mutations

27 Jan 2014 CG: Realistic binding energies
Interaction also depends on the timescale

Center of mass separation $r$ [nm]

PMF [kJ mol$^{-1}$]

-150 -100 -50 0 50 100 150

2ns 2μs

Coarse-grained (from avg. MF of 20 sims., 2 ns)
All-atom (from avg. MF of 10 sims., 2 ns)
Experimental $\Delta G^\text{eff}$
Coarse-grained (from avg. MF of 3 sims., 2 μs)

Experimental $\Delta G_{\text{off}}$
Interaction also depends on orientation

- For 15 protein pairs, in total 19,000 orientations
- 1.2M CPUh
- ~100T raw data
  → not stored, but we now want to analyze it... :-(
Two challenges: Data growth & Data Loss

- Computer speed and storage capacity is doubling every 18 months and this rate is steady (Moore’s law)
- DNA sequence data is doubling every 6-8 months over the last 3 years and looks to continue for this decade

‘Oops, that link was the laptop of my former PhD student’
Semantic interoperability through RDF (resource description framework)

Unit of information: the data triple

Linked data cloud

Data interoperability