**Project C1**
**Feature selection in high dimensional data for risk prognosis in oncology**

Prof. Dr. Alexander Schramm, Prof. Dr. Sven Rahmann

---

### New Dimension of Data Volume: Whole genome & nanopore sequencing

Features are derived from molecular probes or sequences (reads).

**Goals:**
- Identifying molecular biomarkers for risk prognosis
- Modeling prediction functions

**Challenges:**
- data volume (100s of GBs sequence or ion current data)
- limited number of samples vs. an extremely high number of features
  ($n < p$ problem)
- resource-efficient feature generation from raw data

---

### Efficient Whole Genome Analysis with DNA k-mers

Use genomewide-unique $k$-mers ($k \in \{21, 23, 25, \ldots \}$) for:
- single nucleotide variant (SNV) discovery
- copy number variants (CNVs)
- structural variants (translocations, fusions)
- methylation analysis from WGS data
- gene expression analysis from RNA-seq

**Feature generation from whole genomes on a standard laptop:**
- Output only unique k-mers that deviate from expected count: new k-mers, lost k-mers, surprising copy number, ...
- Project deviant k-mers to biological entities (regions, genes, transcripts, pathways)
- Detect enrichment of deviant k-mers and deviant biological entities in tumour samples

**Feature reduction:**
- Aggregation: Variant $\rightarrow$ Gene $\rightarrow$ Pathway
- Clustering of similar features with graph-based methods

---

### Key Data Structure: Efficient DNA k-mer Key-Value Store

**Challenge:** small hash table and fast look-ups.

**Speed bottleneck:** cache misses during memory look-ups

**New proposal:** 3-way bucketed Cuckoo hashing with quotienting

Maximal fill rates of hash table for different numbers of hash functions ($H: 2$ or $3$), bucket sizes ($x$-axis, $1$–$15$) and bounds on random walk length during insertion ($W: 100, 500, 1000, 5000, 10000$).

Look-up needs $H$ cache misses in the worst case.

---

### Analysis of Ion Current Data

**Establishment of the technology and preliminary experiments on microbiomes:**

- Validation of CRISPR / Cas9 based knock-out and overexpression of PRKCI.
- Biological Target Validation

**Computational challenge:** Lightweight conversion of ion current signal to DNA sequence
- Signal segmentation: Fused LASSO; given signal $y = (y_j)$,

$$\min \{ f(x) \} = \frac{1}{2} \sum_{j=1}^{n} (y_j - x_j)^2 + \lambda \sum_{i=1}^{m} |x_i|$$

- Discretisation of signal levels; new efficient algorithms for discretised fused LASSO, where $x_i$ must be from a finite known level set $\mathcal{L}$.
- Learn mapping between k-mers of level set $\mathcal{L}$ to (modified) DNA sequence

**Alternative approach:**
- Work with $k$-mers of discretised signal space $\mathcal{L}$ directly (richer representation)
- Discover variants as for WGS analysis in $\mathcal{L}$-space

---

### Biological Target Validation

**Validation of CRISPR / Cas9 based knock-out and overexpression by Western Blot analysis**

3D culture reveals decreased spheroid formation ability and invasiveness upon PRKCI knock-out, while over-expression of PRKCI increases the invasiveness of SH-EP cells (neuroblastoma cell line).