UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Kvik: Interactive exploration of genomic data from the NOWAC postgenome biobank

Bjørn Fjukstad

UNIVERSITY OF TROMSØ

Project Advisors and Collaborators

Associate Professor **Lars Ailo Bongo**, Department of Computer Science

Professor **Eiliv Lund**, Department of Community Medicine

Mie Jareid and **Karina Standahl Olsen**, Department of Community Medicine

Overview

Biological Background
Norwegian Women and Cancer (NOWAC)
Challenges
Requirement analysis
Demo
Kvik
Evaluation
Future and Related Work
Conclusion

Biological Background

Cells are the smallest units in our body that still preform a function

All cells store the same genetic information within **DNA**

…gtgcatctgactcctgaggagaag…
…cacgtagactgaggactcctcttc…

g guanine a adenine t thymine c cytosine

Genes are sequences of DNA that code to **proteins**

Biological Background

Cells are the smallest units in our body that still preform a function

All cells store the same genetic information within **DNA**

```
…gtgcatctgactcctgaggagaag…
…cacgtagactgaggactcctcttc…
```

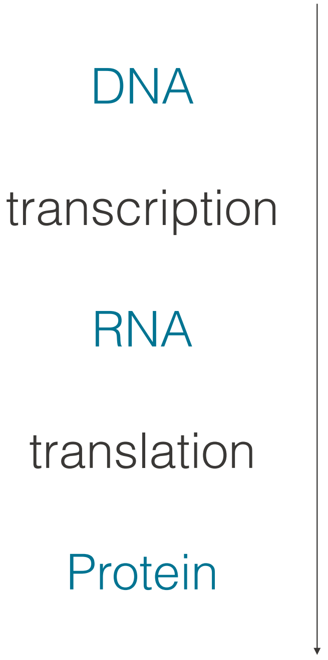
g guanine a adenine t thymine c cytosine

Genes are sequences of DNA that code to **proteins**

Biological Background

Gene expression is the process of transcribing DNA into RNA that translates into proteins

Gene **expression levels** reveal how much RNA is produced



DNA

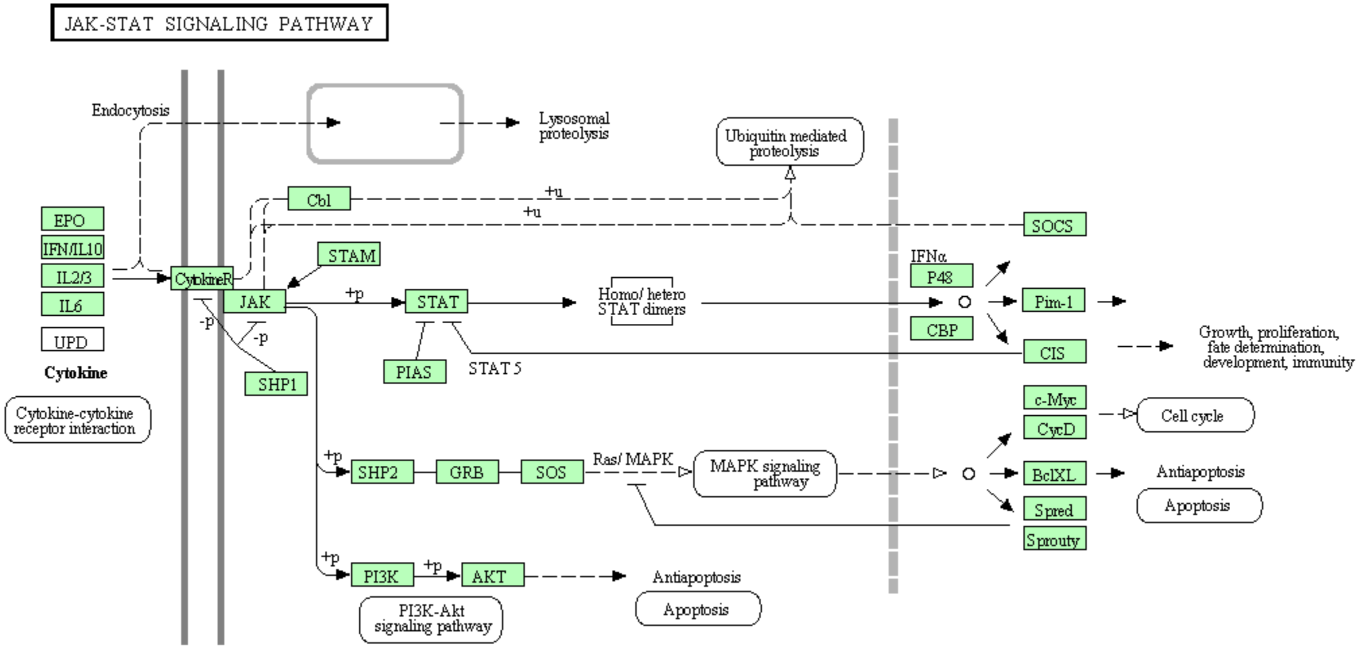
transcription

RNA

translation

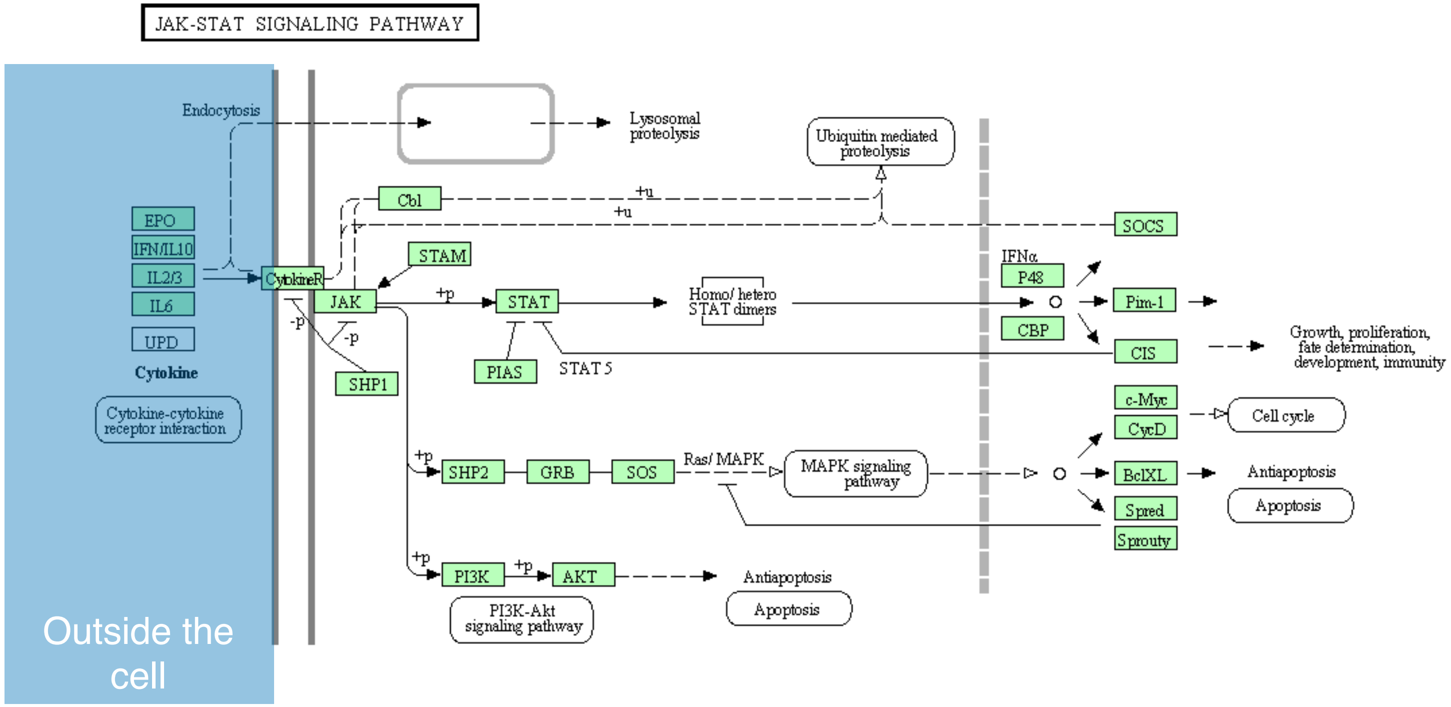
Protein

Pathways



genome.jp/kegg-bin/show_pathway?hsa04630

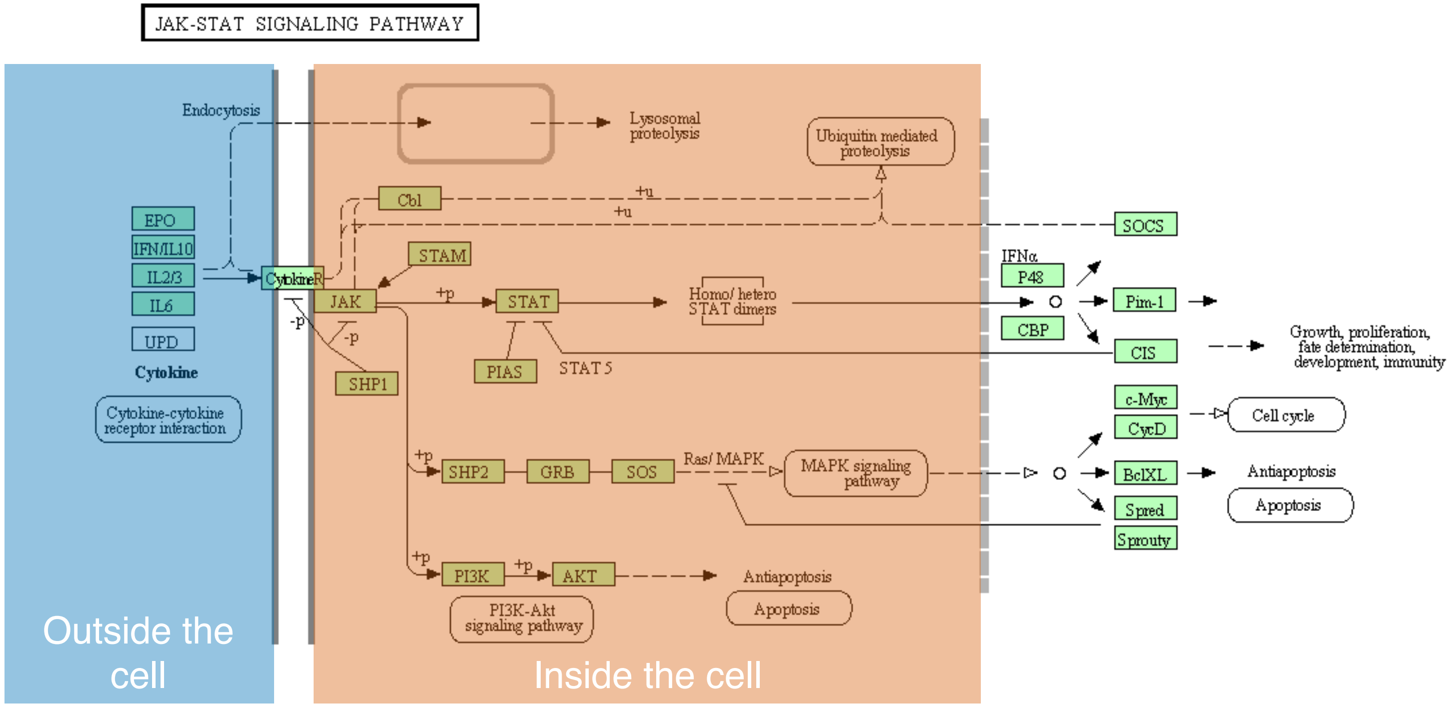
Pathways



Outside the cell

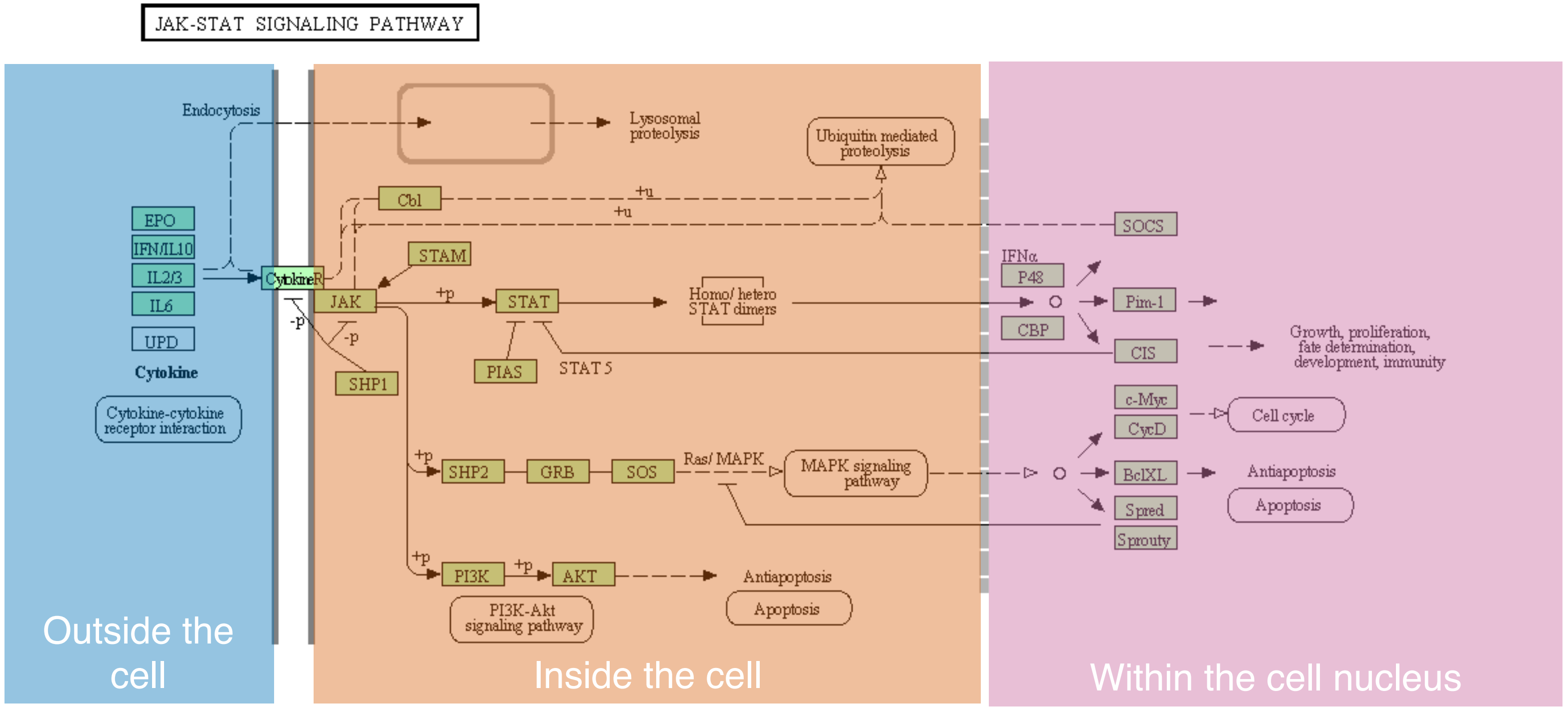
genome.jp/kegg-bin/show_pathway?hsa04630

Pathways



genome.jp/kegg-bin/show_pathway?hsa04630

Pathways



genome.jp/kegg-bin/show_pathway?hsa04630

Traditional Workflow

Complex spreadsheets and **a plethora of databases and applications**

Manual process of looking up and finding relevant pathways.

Comparison and integration of data **between applications**

Kvik

Explore the dynamics of carcinogenesis through biological pathways and gene expression

Integrates state of the art **pathway maps** from the KEGG database and **gene expression data** from the NOWAC postgenome biobank in a single system

Norwegian Women and Cancer

Identify the possible relationships between lifestyle and the risk of cancer.

Started data collection in 1991, now the biobank holds more than 60 000 blood samples and 800 biopsies.

Information about exposure through questionnaires

Large research group with international collaborators

Challenges

Researchers have access to **large quantities of research data** that can enable novel discoveries

Realizing these discoveries require **new systems** for exploratory analyses

To understand complex diseases, systems need to integrate information from **multiple biological levels** and **sources**

Challenges

There are numerous systems that organize and manage research data

Systems that help researchers gain new knowledge are still **largely missing** in the bioinformatics domain

Such systems need to **integrate both advanced statistical models and interactive visualizations** of large-scale datasets

NOWAC Challenges

Researchers explore and analyze the NOWAC biobank **without prior explicit hypothesis**

They **cannot share the NOWAC biobank** outside the research group, making current online tools unsuitable

Overview

Biological Background

Norwegian Women and Cancer (NOWAC)

Challenges

Requirement analysis

Demo

Kvik

Evaluation

Future and Related Work

Conclusion

Requirement Analysis

Identified seven requirements through collaboration with cancer researchers

Interactivity, Scalability, Simplicity, Familiarity,
Heterogeneity, Extendability, and Security

Requirement Analysis

Identified seven requirements through collaboration with cancer researchers

Interactivity, Scalability, Simplicity, Familiarity,
Heterogeneity, Extendability and Security,

Simplicity

Researchers shouldn't bother with installing **additional applications or plug-ins**

Kvik runs in a modern web browser using **HTML5** to provide a platform-independent exploration tool

Familiarity

Researchers work more efficiently using **familiar interfaces and representations**

Kvik follows the **drawing convention in KEGG** to provide familiar pathway maps

Extendability

Researchers require data exploration systems to handle **new analysis methods** and **data processing systems**

Kvik has an extendable analysis backend allowing researchers to **incorporate their own statistical models** into the data exploration

Kvik

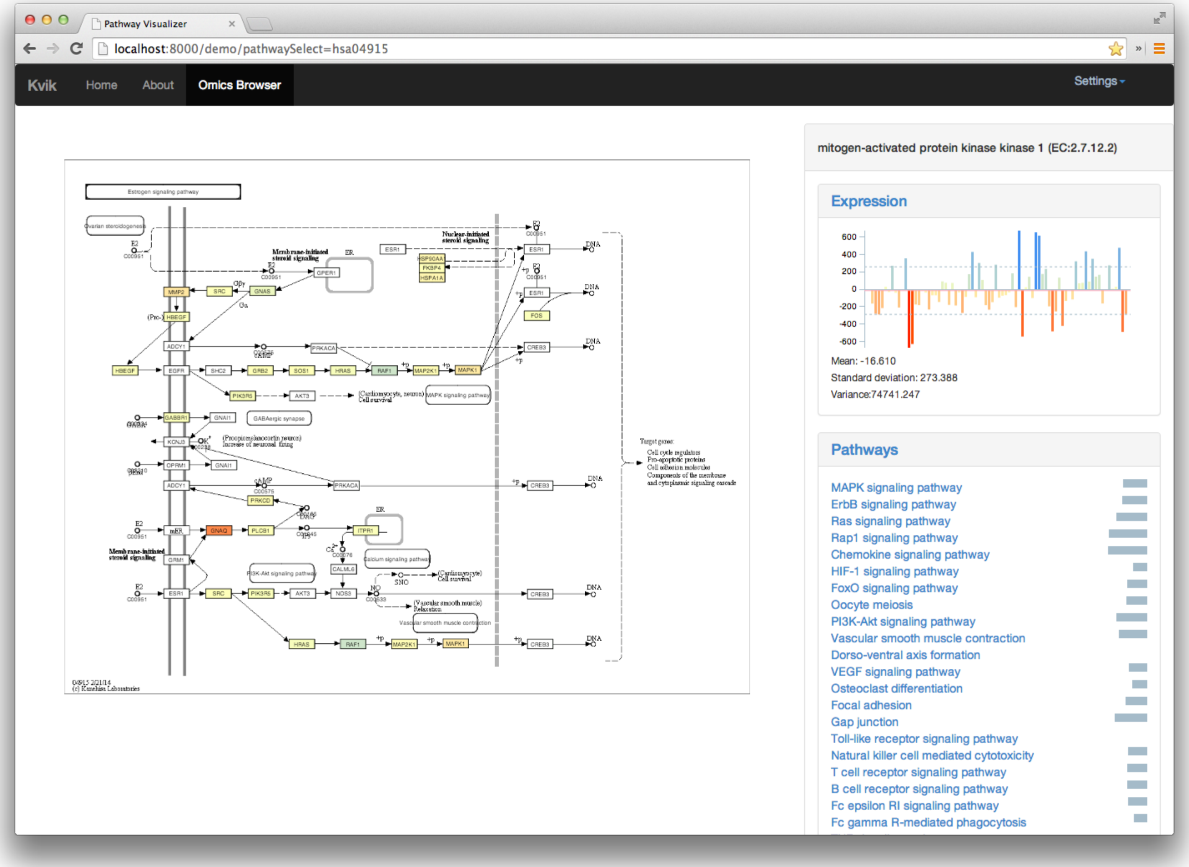
Interactive exploration of the dynamics of carcinogenesis through studies of **biological pathways** and **genomic data**

Allows researchers to navigate and **explore large amounts of research data**

Three-tiered architecture with a lightweight **web application** for data exploration and a powerful backend for statistical analysis

Gene expression from the **NOWAC** biobank and pathways from **KEGG**

Demo



mitogen-activated protein kinase kinase 1 (EC:2.7.12.2)

Mean: -16.610

Standard deviation: 273.388

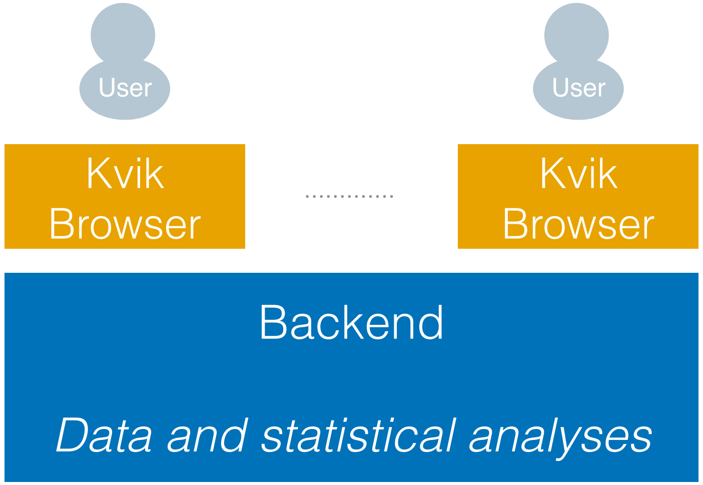
Variance:74741.247

- MAPK signaling pathway
- ErbB signaling pathway
- Ras signaling pathway
- Rap1 signaling pathway
- Chemokine signaling pathway
- HIF-1 signaling pathway
- FoxO signaling pathway
- Oocyte meiosis
- PI3K-Akt signaling pathway
- Vascular smooth muscle contraction
- Dorso-ventral axis formation
- VEGF signaling pathway
- Osteoclast differentiation
- Focal adhesion
- Gap junction
- Toll-like receptor signaling pathway
- Natural killer cell mediated cytotoxicity
- T cell receptor signaling pathway
- B cell receptor signaling pathway
- Fc epsilon RI signaling pathway
- Fc gamma R-mediated phagocytosis

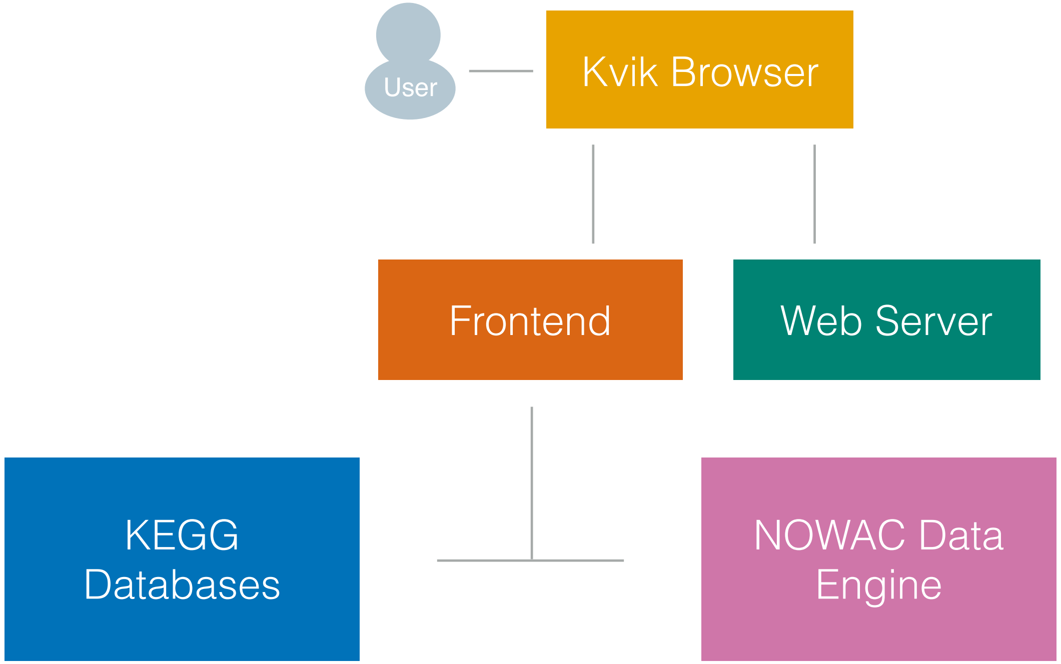
Overview

Biological Background
Norwegian Women and Cancer (NOWAC)
Challenges
Requirement analysis
Demo
Kvik
Evaluation
Future and Related Work
Conclusion

Architecture



Design



User

Kvik Browser

Frontend

Web Server

KEGG Databases

NOWAC Data Engine

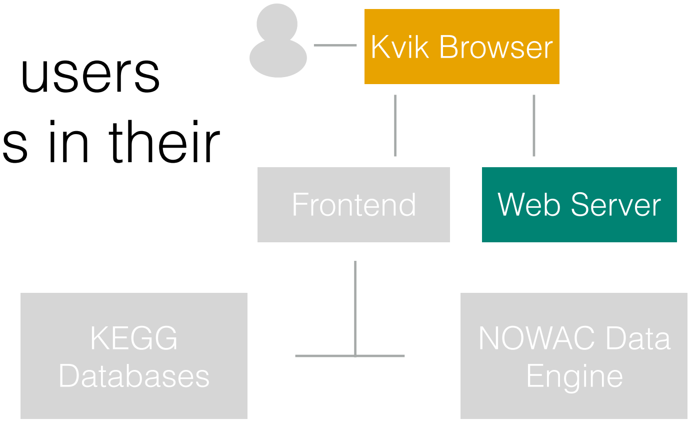
Kvik Browser

Designed and implemented as a web application

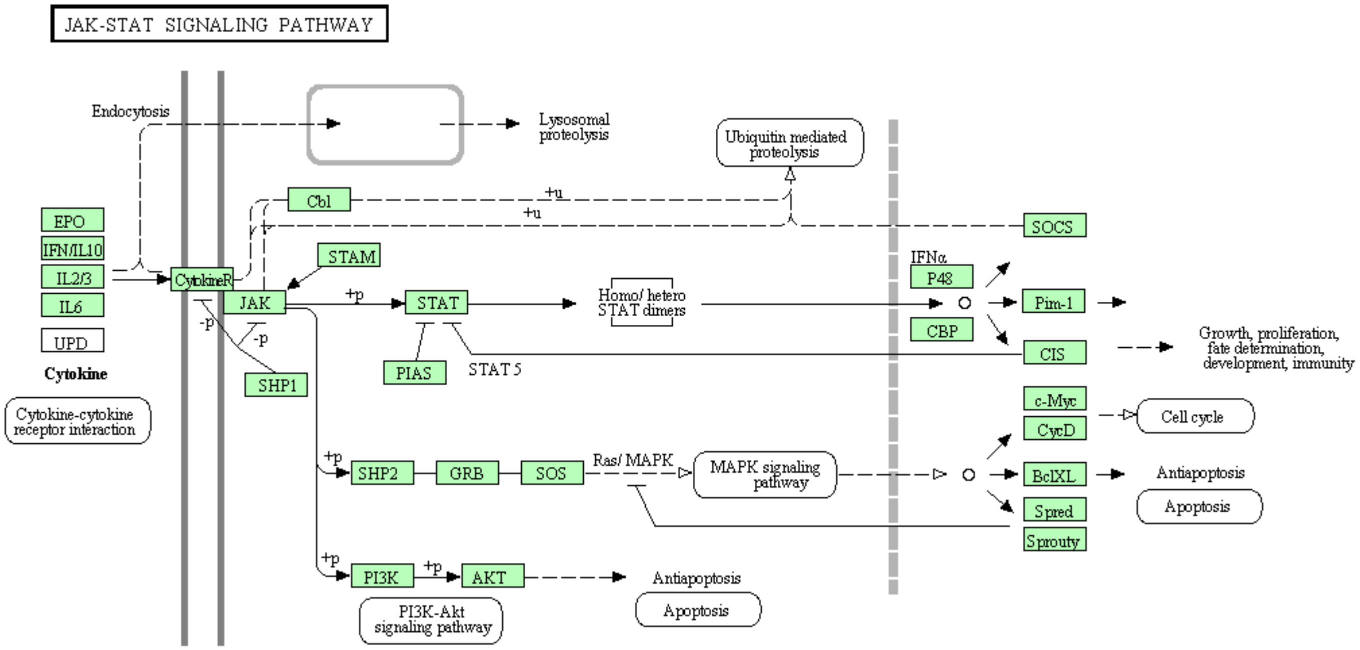
Uses **Cytoscape.js** to visualize biological pathways and **D3** for gene expression data

The web server hosts the application, and multiple users can interact with through the Kvik Browser that runs in their web browser

cytoscape.github.io/cytoscape.js / d3js.org



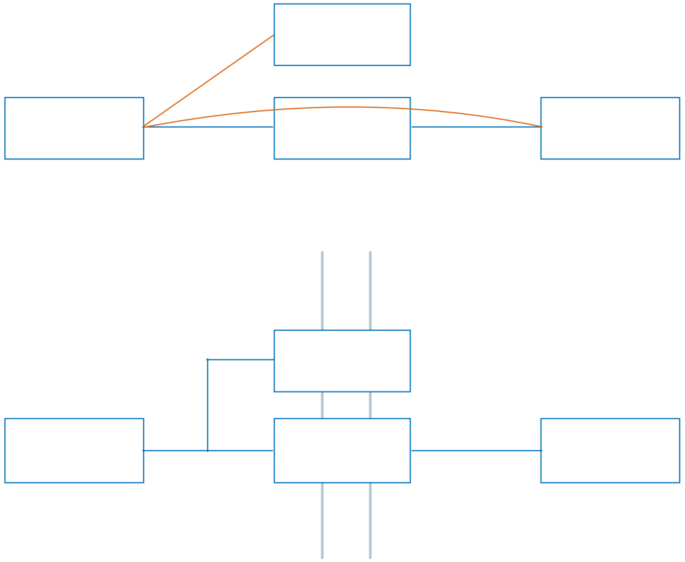
KEGG Pathway Maps



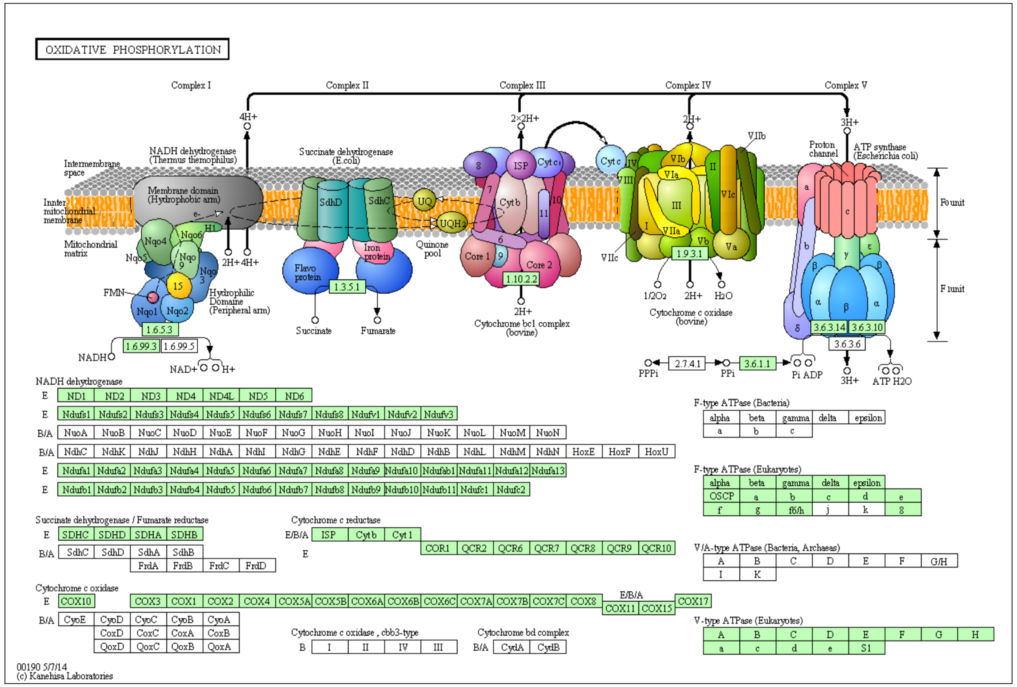
Visualizing Biological Pathways

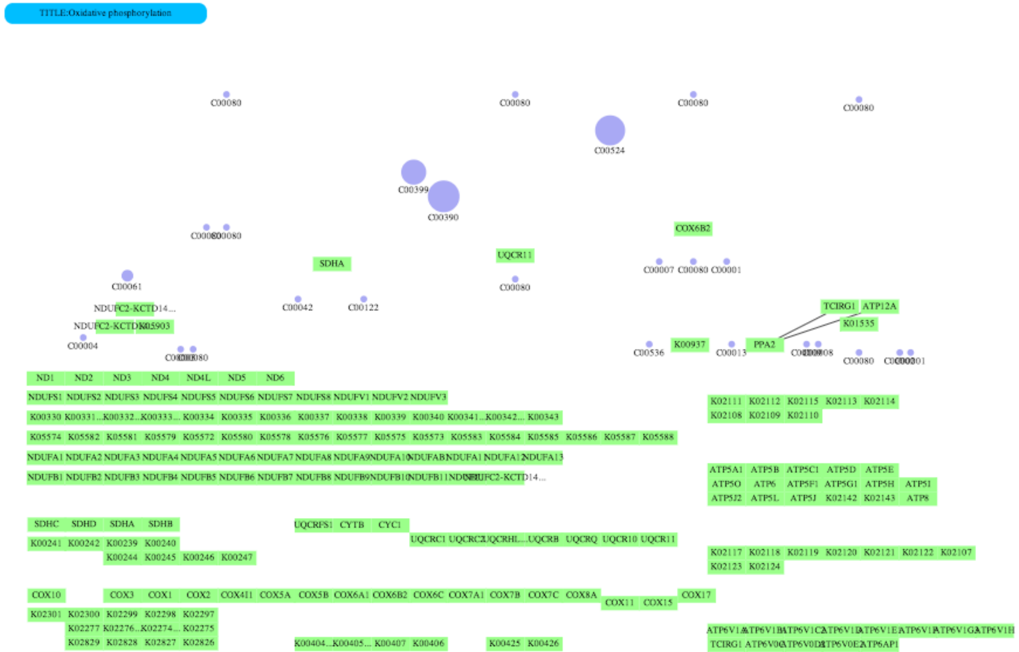
KEGG Markup Language (KGML) representation

```
<pathway name="path:hsa04915" … >
  <entry id=1 name="hsa:2009" x=950  y=12 … >
  .
  .
  .
  <relation entryid1=1 entryid2=3 … >
  .
  .
  .
</pathway>
```

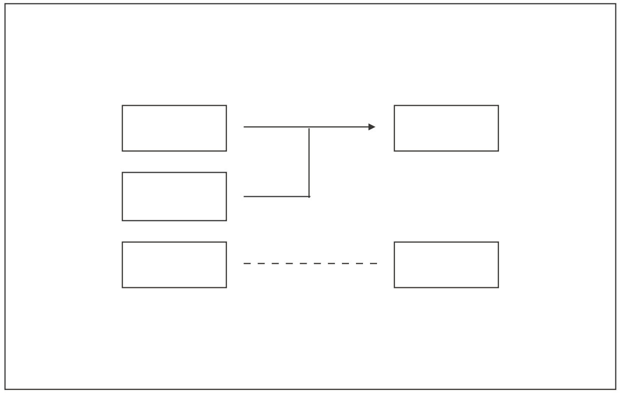


Visualizing Biological Pathways

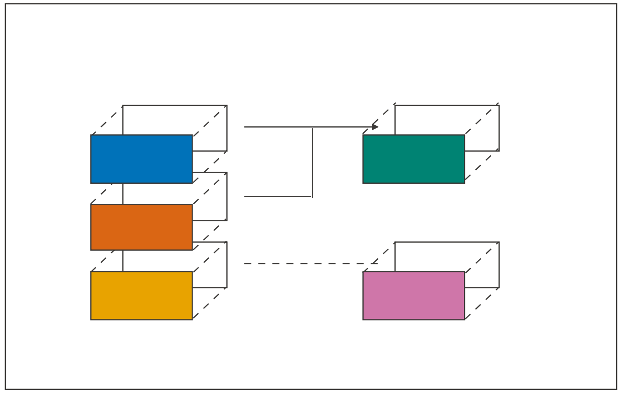

OXIDATIVE PHOSPHORYLATION
Complex I
Complex II
Complex III
Complex IV
Complex V
NADH dehydrogenase (Thermus themophilus)
Succinate dehydrogenase (E.coli)
Intermembrane space
Inner mitochondrial membrane
Mitochondrial matrix
Membrane domain (Hydrophobic arm)
Hydrophilic Domaine (Peripheral arm)
Flavo protein
Iron protein
Quinone pool
Cytochrome bc1 complex (bovine)
Cytochrome c oxidase (bovine)
Proton channel
ATP synthase (Escherichia coli)
F0 unit
F1 unit
NADH
NAD+
Succinate
Fumarate
PPPi
PPi
Pi ADP
ATP H2O
NADH dehydrogenase
Succinate dehydrogenase / Fumarate reductase
Cytochrome c reductase
Cytochrome c oxidase
Cytochrome c oxidase , cbb3-type
Cytochrome bd complex
F-type ATPase (Bacteria)
F-type ATPase (Eukaryotes)
V/A-type ATPase (Bacteria, Archaeas)
V-type ATPase (Eukaryotes)
00190 5/7/14
(c) Kanehisa Laboratories



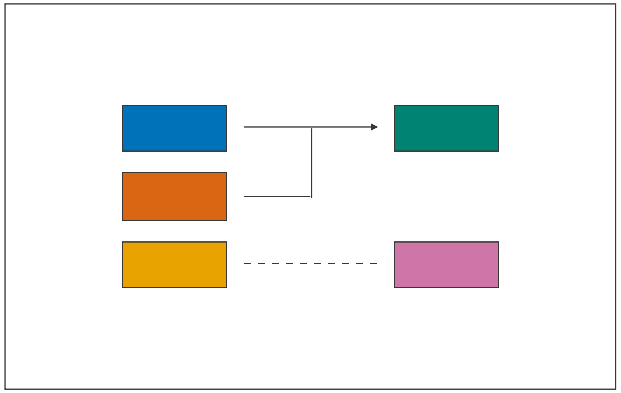
Visualizing gene expression data on biological pathways



Static image from KEGG



Fetch gene expression and draw according to location in KGML representation



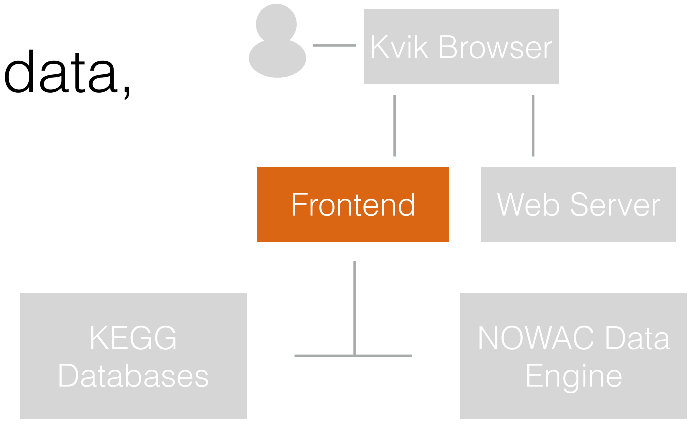
Final visualiation

Frontend

The Frontend **retrieves data from different sources** and provides a simple interface to the Kvik Browser

Implemented as a **HTTP REST** service in Go

It provides the Kvik Browser with gene expression data, pathway maps and information about genes and pathways

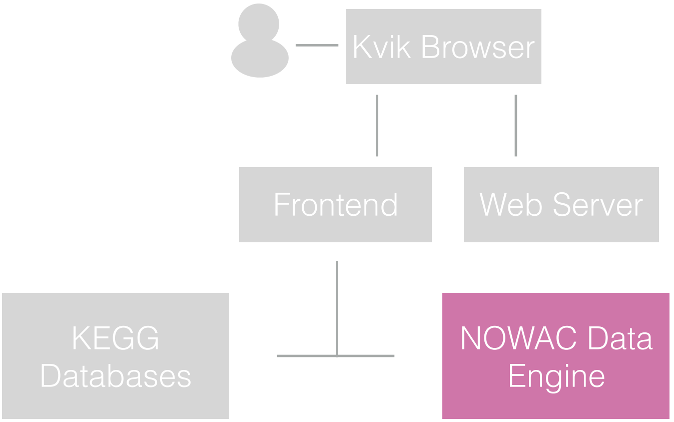


NOWAC Data Engine

A stand-alone service that provides both **storage** and **computational resources**.

Manages a subset of the NOWAC biobank, and performs analysis on it.

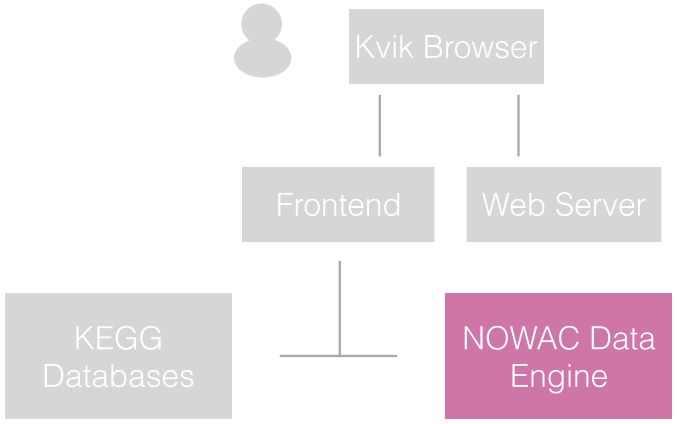
Three parts: **Interface**, **Execution Engine**, and **Data Management system**



Execution Engine

A simple Python service accessible by **RPC**

Our **collaborators can contribute with R-scripts** and new analysis methods, everything accessible from the Kvik Browser



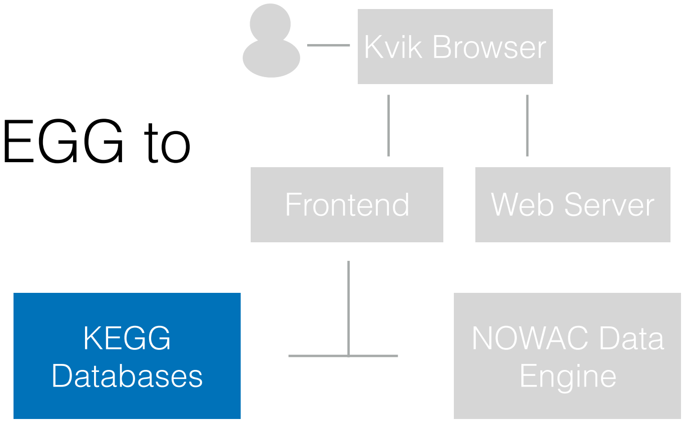
KEGG Databases

A simple library to fetch information about genes and pathways, in addition to pathway maps

Uses the **KEGG REST API** rather than expensive FTP license

Local cache to reduce the number of requests to KEGG to improve latency and respect licensing issues

rest.kegg.jp



Overview

Biological Background

Norwegian Women and Cancer (NOWAC)

Challenges

Requirement analysis

Demo

Kvik

Evaluation

Future and Related Work

Conclusion

Evaluation

1 How long time does it take to load pathways and genes?

2 Can it run on your commodity workstation?

3 How does the backend scale to larger datasets?

4 Can researchers integrate Kvik into their day-to-day workflow?

Experimental Setup

Both **client and backend ran on the same machine**

2013 Mac Mini, 2.66 GHz Intel Core i7 Processor (4 cores), 16 GB RAM, 1 TB Drive

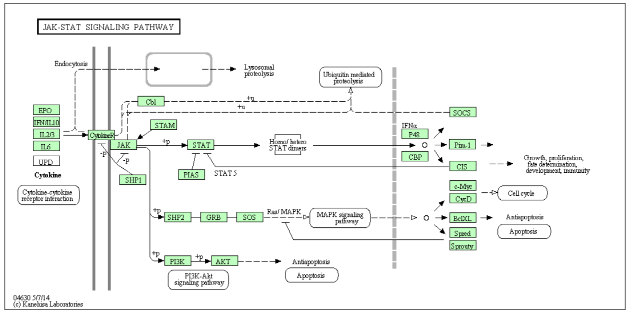
OS X 10.9.2 with Firefox v 29.0

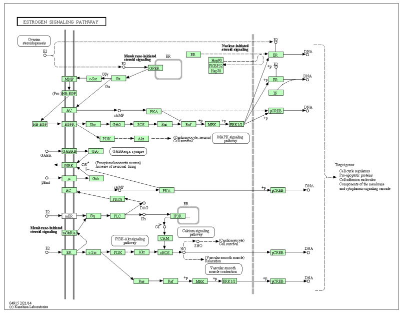
Benchmark.js was used to measure latency in the Kvik Browser, the standard library in Go to measure the backend

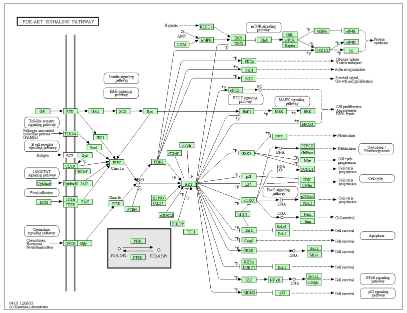
benchmarkjs.com golang.org/pkg/testing#Benchmark

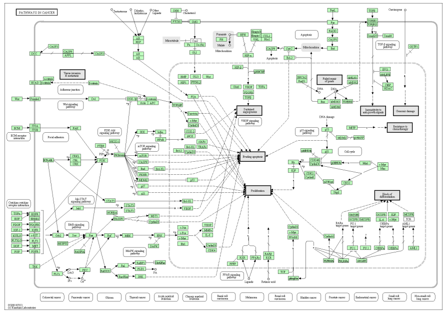
Pathway load time

Id	Name	Number of nodes
hsa04630	Jak-Stat signaling pathway	35
hsa04915	Estrogen signaling pathway	74
hsa4151	PI3K-Akt signaling pathway	120
hsa05200	Pathways in cancer	267

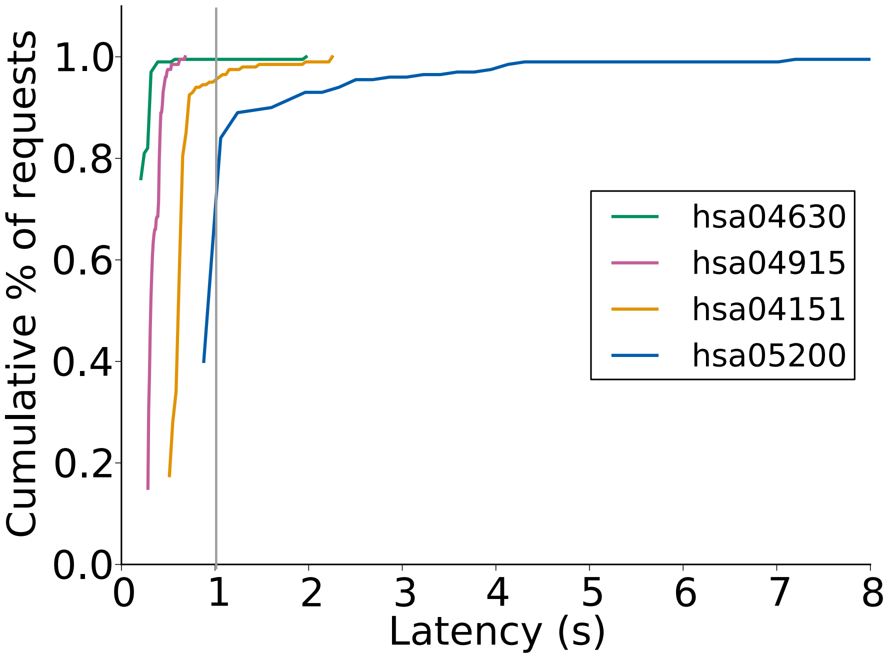








Pathway load time

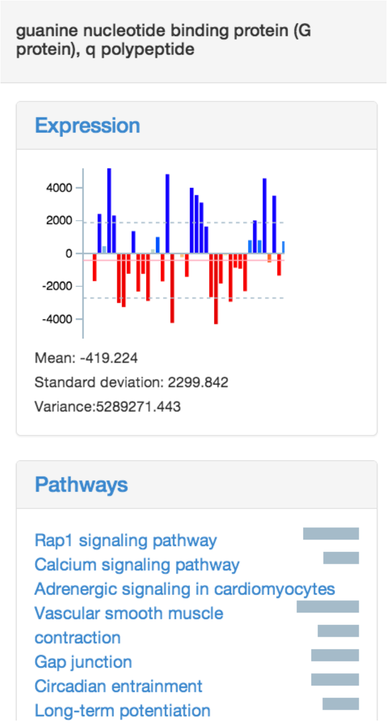


On average Kvik retrieves gene expression data from the NOWAC biobank and visualizes complex pathways within a second

Long tail of requests for large pathways

Gene details load time

Id	Name	Number of pathways
hsa:4313	MMP2	6
hsa:3303	HSPA1A	12
hsa:6654	SOS1	32
hsa:5604	MAP2K1	55



guanine nucleotide binding protein (G protein), q polypeptide

Expression

Mean: -419.224

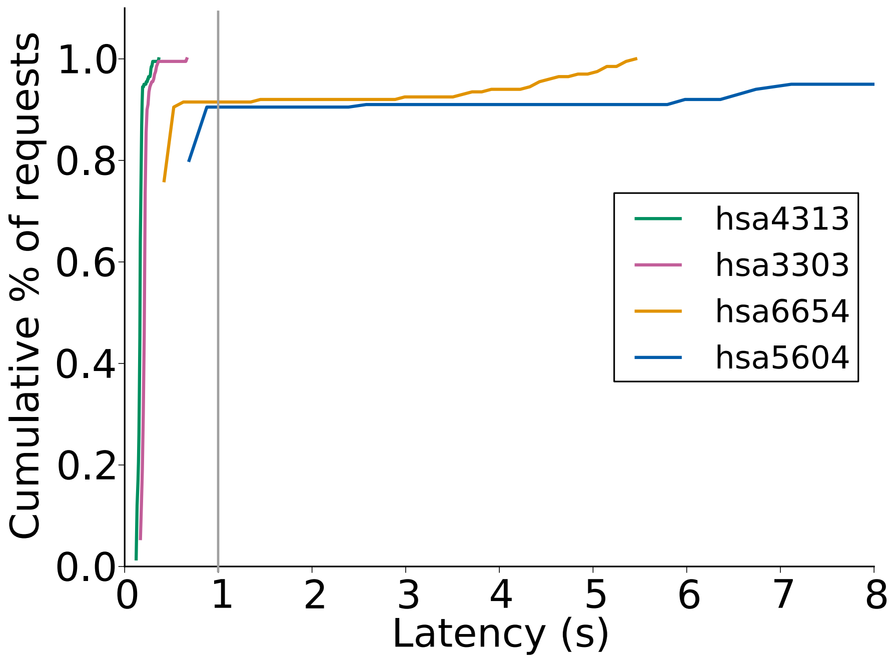
Standard deviation: 2299.842

Variance:5289271.443

Pathways

- Rap1 signaling pathway
- Calcium signaling pathway
- Adrenergic signaling in cardiomyocytes
- Vascular smooth muscle contraction
- Gap junction
- Circadian entrainment
- Long-term potentiation

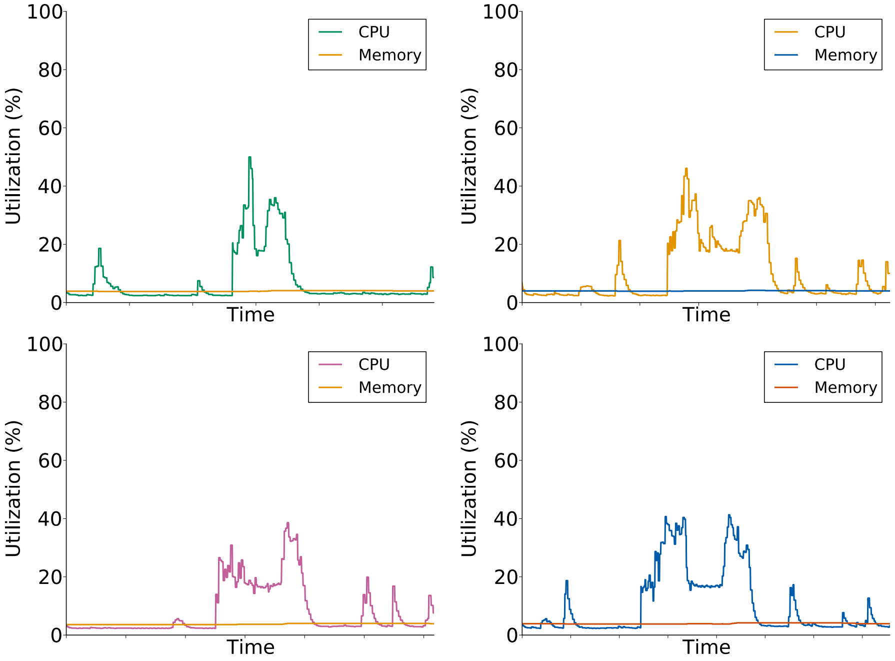
Gene details load time



On average, Kvik generates the gene inspection views within a second

Long tail of requests that requires further inspection

Resource consumption



Kvik Browser uses at peak 50% CPU and less than 4% memory

Early experiments show promising results on mobile devices, such as iPads.

Backend scalability

Evaluated using **real data** from the NOWAC postgenome biobank

77 case control pairs, gene expression values for **9101 genes**

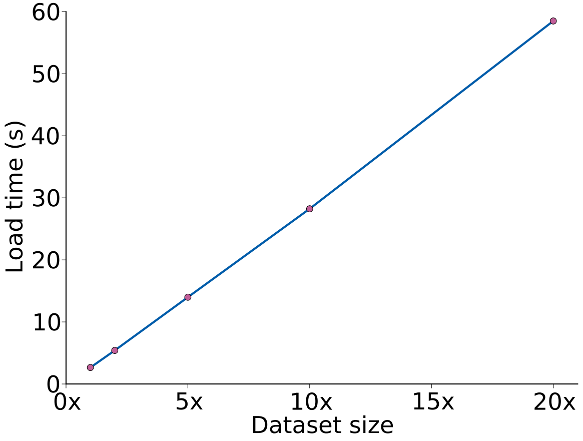
23.8 MB of raw data stored on Stallo compute cluster

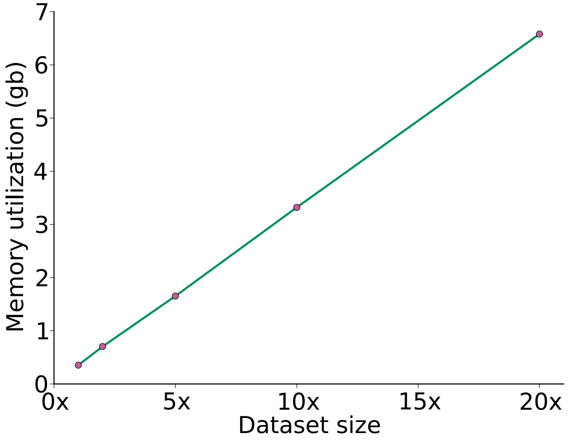
notur.no/hardware/stallo

Backend scalability

The backend scales linearly

Loads the equivalent of half of the NOWAC biobank in less than a minute, using less than 7GB of RAM





Researcher evaluation

Using an iterative development process was a success

Familiar pathway maps from KEGG speed up data exploration

We plan on deploying Kvik in the NOWAC research group based on their positive feedback

Overview

Biological Background

Norwegian Women and Cancer (NOWAC)

Challenges

Requirement analysis

Demo

Kvik

Evaluation

Future and Related Work

Conclusion

Future Work

More advanced statistical analyses through **statistical packages** like Bioconductor

Collaboration and sharing results with other researchers using Kvik

Integrate **new data sources** and **more data** from the NOWAC biobank.

Related Work

There are many pathway databases, for example **KEGG**, **BioCarta** and **WikiPathways**

A wide variety of stand-alone applications such as the **Caleydo Framework**, **Vanted**, and **VisANT**

Some online tools such as **Pathway Projector** and **KEGGViewer**, but hard to integrate with the NOWAC biobank

kegg.jp biocarta.com wikipathways.org
github.com/Caleydo/caleydo visant.bu.edu vanted.ipk-gatersleben.de
github.com/biojs/biojs g-language.org/PathwayProjector

Contributions

A **requirement analysis** for visualization systems for exploring and visualizing data from the NOWAC postgenome biobank

The **design and implementation** of Kvik, a data exploration tool for biological pathways and gene expression data

The **experimental evaluation of Kvik**, demonstrating that researchers can use Kvik for interactive exploration of the full NOWAC biobank and KEGG databases

Concluding Remarks

The first version of Kvik provides **interactive exploration of biological pathways and genomic data**

Kvik is important to enable novel discoveries from the NOWAC postgenome biobank

Open-sourced at github.com/fjukstad/kvik

UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Thank you!

Bjørn Fjukstad
bjorn.fjukstad@uit.no

UNIVERSITY · OF TROMSØ ·