Кафедра математических и информационных технологий, СПбАУ

Язык запросов для геномного браузера GemlBee

Выполнил: Егор Горбунов
Руководитель: Олег Шпынов

20 мая 2016 г.

Геномный браузер

- Геномный браузер – приложение, отображающее различную информацию о геноме, привязанную к конкретным локациям этого генома
- USCS, Ensembl, BioViz, . . ., GemlBee, . . .

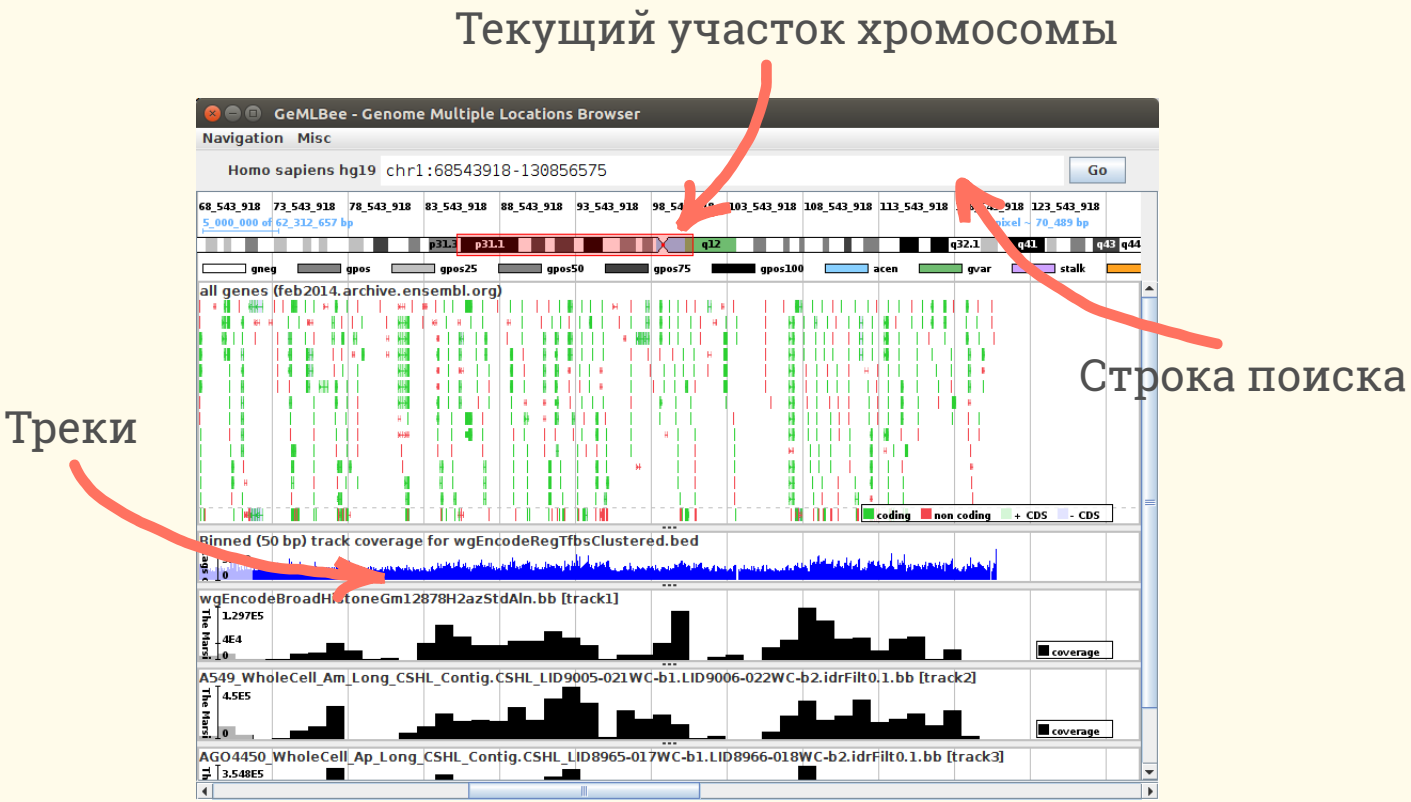

Рис. 1: GemlBee

Треки

- Трек — отдельная дорожка, отображаемая геномным браузером, заключающая в себе какую-то информацию про геном
- Бывают разные, в зависимости от этой информации
  - Какая-либо статистическая информация о нуклеотидах в тех или иных частях днк
  - Карты с конкретными генами с привязкой к участкам днк
  - . . .
- Нам интересны треки, которые полезны при исследовании эпигенетических изменений (экспрессии генов)

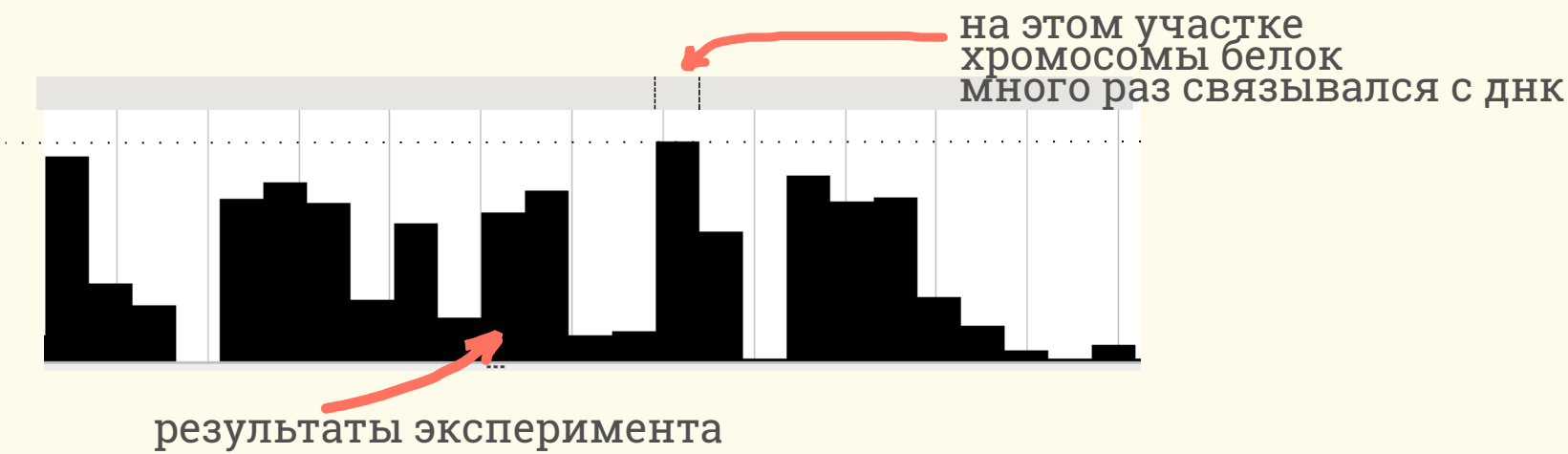

Рис. 2: Пример трека, который нас интересует

Динамическая генерация треков

Частая последовательность действий:

Знаем из статей, что событе X (например, сигнал $>$ t) вероятно связано с событием Y. Хотим посмотреть на хромосоме места, где верно X и визуально оценить Y.

Идея:

Имея результаты нескольких экспериментов, представленных в виде треков, было бы удобно уметь как-то комбинировать эти результаты в новые треки и совершать над ними операции с целью упрощения визуального анализа результатов треков целом.

Нужно ли что-то делать?

List of genome browsers [edit]

- Alamut A gene browser that handles HGVS nomenclature[1] and integrates missense and splicing[2] prediction tools for mutation interpretation.
- Annmap A genome browser that shows Affymetrix Exon Microarray hit locations alongside the gene, transcript and exon data using the Leafletjs Maps API
- Apollo Genome Annotation Curation Tool A cross-platform, Java-based standalone genome viewer with enterprise-level functionality and customizations. The standard for many model organism databases.[3]
- Argo Genome Browser A free and open-source standalone Java-based genome browser for visualizing and manually annotating whole genomes.[4]
- Artemis Genome Browser A free and open-source standalone genome browser (Wellcome Trust Sanger Institute) for visualizing and manually annotating whole genomes.[5] It can also be used to visualize next generation sequencing data.[6]
- Avadis NGS combines a genome browser and set of data analysis tools for ChIP-Seq, RNA-Seq, and genomic variation experiments, developed by Strand Life Sciences
- BugView Free cross-platform desktop browser for visualizing genomes, especially suited for comparing prokaryotic genomes.
- Celera Genome Browser, developed at Celera Genomics as part of Celera's sequencing and annotation of the human genome, and released as open source in 2006.
- Chipmonk A Java-based tool to visualise and analyse ChIP-on-chip array data, developed at the Babraham Institute in Cambridge.
- Dalliance Javascript-based genome browser
- DiProGB: The Dinucleotide Properties Genome Browser
- DNAnexus Flash-based interactive genome browser, as well as next-gen sequence analysis and visualization.
- Ensembl The Ensembl Genome Browser (Wellcome Trust Sanger Institute and EBI)[7][8]
- ERGO The ERGO Bioinformatics Suite developed by Igenbio, Inc
- Gaggle Genome Browser A java-based genome browser developed at Institute for Systems Biology (ISB) for high-throughput data integration.
- GeneWall - Mobile genome browser
- GBrowse The GMOD GBrowse Project[9][10]
- Genestack web-based genomics operating system
- GenomeView is a next-generation stand-alone genome browser and editor specifically designed to visualize and manipulate a multitude of genomics data.[11]
- Genome Maps implements HTML5, scalable vector graphics, displaying genes, transcripts, exons, regulatory features, SNPs etc. Allows the local upload of large genomic data files.[12]
- Genome Wowser An iPad-enabled view of the human genome. The app, developed by the Center for Biomedical Informatics (CBMi) at The Children's Hospital of Philadelphia, provides a functional presentation of the popular UCSC Genome Browser.[13]
- HuRef - stand-alone browser for navigating individual human genome
- The Genomic HyperBrowser focuses on statistical analysis of elements along the genome,[14] built on the Galaxy platform.
- Genostar GenoBrowser is a standalone application to display and explore genomic data from any kind of file (BAM, GenBank, Fasta, GFF, ...)
- Genoverse interactive genome browser, built using JavaScript and HTML5 canvas element
- GenPlay A genome viewer and analyzer developed in Java at Albert Einstein College of Medicine.[15]
- Golden Helix GenomeBrowse A free genome browser for exploring sequencing pile-up and coverage data with numerous annotation tracks hosted on the cloud.
- Integrated Genome Browser (IGB) Open-source and free Java-based desktop genome viewer for visualizing next-gen sequence and microarray data.
- Integrative Genomics Viewer IGV A high-performance visualization tool for interactive exploration of large, integrated genomic datasets.[16] A lite version for the iPad is available in the Apple App Store.
- Integrated Microbial Genomes (IMG) system by the DOE-Joint Genome Institute
- JBrowse a JavaScript genome browser by the open-source Generic Model Organism Database project.[17]
- MGV - Microbial Genome Viewer
- myKaryoView - A Direct-to-consumer oriented genomic browser [18]
- MochiView Genome Browser
- NextBio Genome Browser - an interactive application that lets visualization of physical relationship between private or public biosets and different types of genomic elements, including genes, miRNA targets, CNVs, CpG islands, SNPs, GWAS associations, and LD blocks]
- Pathway Tools Genome Browser[19]
- Persephone Next-generation genome visualization and exploration software.[20]
- Plant GDB - Plant genome browser
- Savant Genome Browser for visual analytics of high-throughput sequencing data
- SEED viewer for visualizing and interrogating the SEED database of complete microbial genomes
- STAR: An Integrated Solution to Management and Visualization of Sequencing Data
- Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.[21]
- TGAC Browser visualisation solutions for big data in the genomic era. An open-source Genome Browser developed at The Genome Analysis Centre, UK works with Ensembl Data set and many more.
- Trackster Galaxy's visualization and visual analysis environment [22][23]
- UCSC Genome Browser and Tools (UCSC Genome Bioinformatics) at UC Santa Cruz).[24] Browser for more than 240 genomes: vertebrates and model invertebrates.
- UGENE visualizes sequences and annotations on a local computer
- Viral Genome Organizer (VGO) A genome browser providing visualization and analysis tools for annotated whole genomes from the eleven virus families in the VBRC (Viral Bioinformatics Resource Center) databases
- VISTA genome browser a comprehensive suite of programs and databases for comparative analysis of genomic sequences. There are two ways of using VISTA - you can submit your own sequences and alignments for analysis (VISTA servers) or examine pre-computed whole-genome alignments of different species.
- WashU EpiGenome Browser web-based visual exploration of genomics and epigenomics data sets[25][26]
- CGView

Не поддерживают требуемую функциональность

Постановка задачи

- Изучить предметную область
- Предложить операции, которые можно совершать над треками
- Сделать язык, позволяющих описывать эти операции
- Встроить этот язык в геномный браузер GemlBee

Операции над треками

- Каждый трек — это вектор из N чисел
- Арифметические опреации ($\circ \in \{+, -, \cdot, /\}$):
  $(a_1, a_2, \ldots, a_N) \circ (b_1, b_2, \ldots, b_N) = (a_1 \circ b_1, a_2 \circ b_2, \ldots, a_N \circ b_N)$
- Операции сравнения треков: $<$, $>$, ... Результат — **предикат** — множество отрезков генома, на котором результат сравнения истинный
- Логические связки $\vee, \wedge, \neg$
- Условная генерация трека **предикату**

Язык

```
a := 2 + 2

b := if (a < 1) then 1 else 3

pred := (a < b) AND (b == c)

c := (b + (if (pred) then b else a) * b) / a

show pred

show c
```

Пример

- 2 разных эксперимента по нахождению регуляторных участков днк медуллобластомы
- генерация треков помогает смотреть на различия результатов и пр.

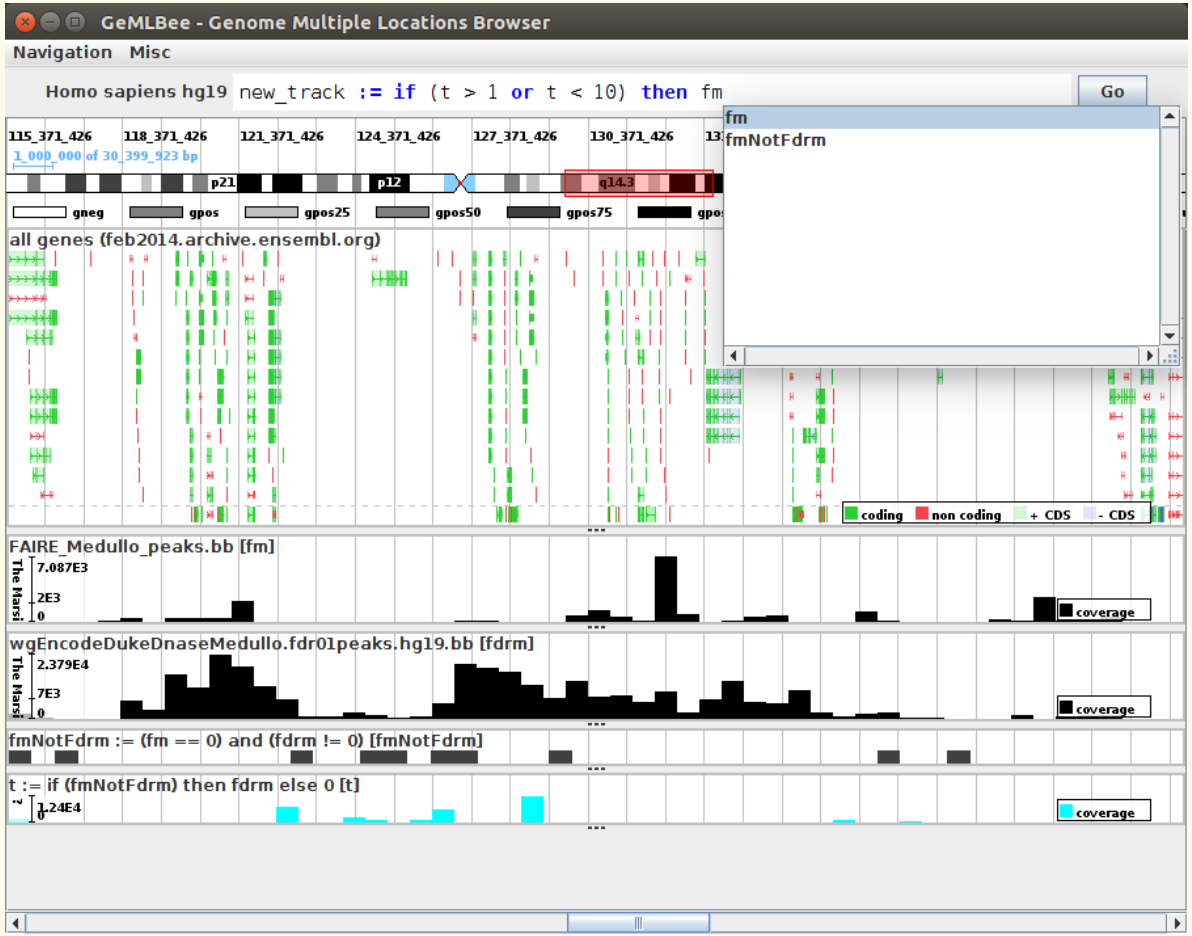

Рис. 4: Скриншот программы

Результаты

Что сделано:

- Реализован интерпретатор языка
- Язык встроен в десктопную версию GemlBee
- Добавлена подсветка синтаксиса и автодополнения
- https://github.com/egorbunov/gemlbee

Что дальше:

- Добавить поддержку в web версии GemlBee

Что узнал и что использовал:

- Kotlin
- Swing
- Parsing expression grammars