Architektura dużych projektów bioinformatycznych

Bartek Wilczyński

bartek@mimuw.edu.pl
http://www.mimuw.edu.pl/~bartek

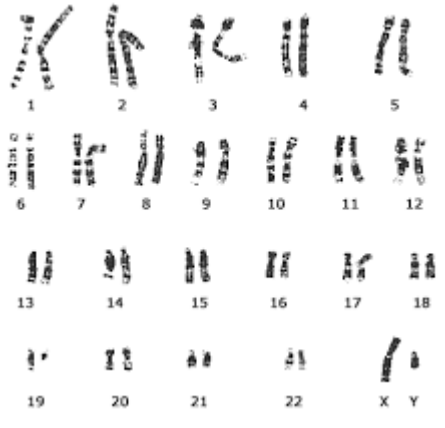
Wykład 4. - Przeglądarki genomów
11.IV. 2018

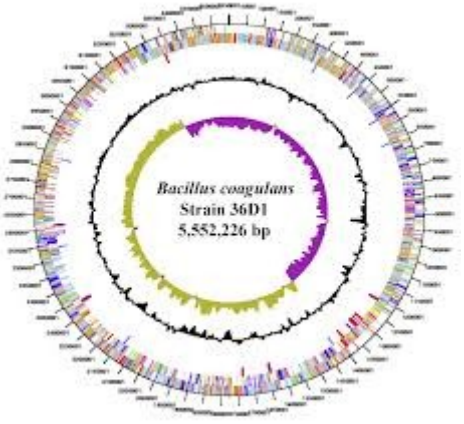
Tematy na dziś

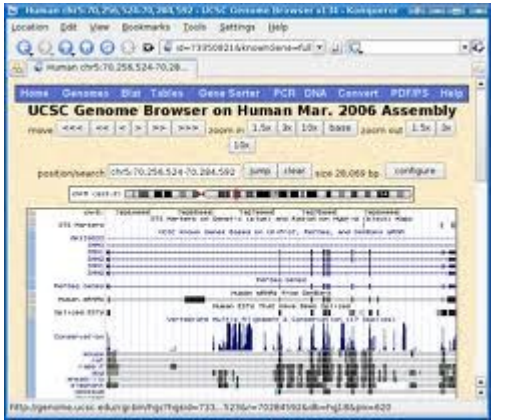
- Po co są przeglądarki genomowe
- Co zawierają opisy genomów (annotation)
- UCSC genome browser
- Gbrowse
- ENSEMBL
- IGB, IGV i podobne
- GenomeDiagram i podobne

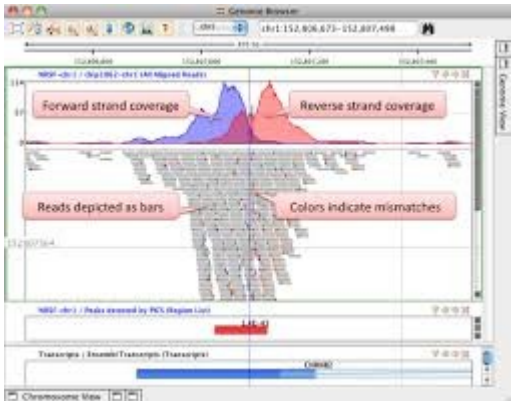
Przeglądarki genomowe

- Od czasu, kiedy dostępne są całe genomy organizmów zachodzi potrzeba wizualizacji
- Genomy bakteryjne często były wizualizowane w całości
- Wraz z pojawieniem się dużych genomów zaszła potrzeba innej wizualizacji
- Teraz także duże zbiory danych stawiają nowe wymagania







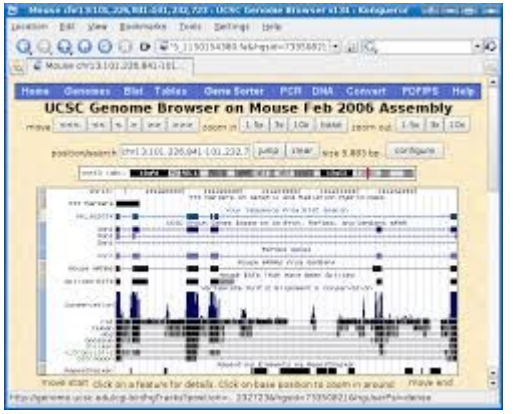


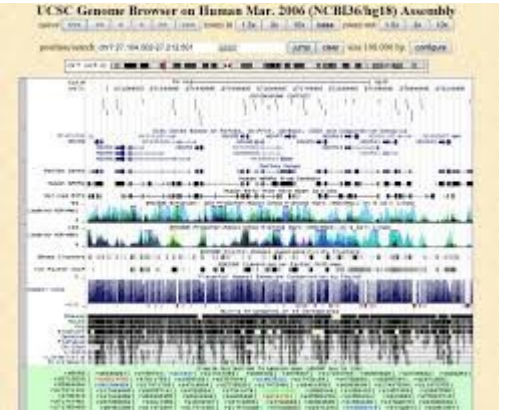
Historycznie

- 1976 - Pierwszy genom RNA bakteriofag MS2
- 1977 – Pierwszy genom DNA (5386 bp)
- 1995
  - *Heamophilus Influenzae* – bakteria 1.8M bp
  - *Saccharomyces Cerevisiae* – eukariont – 12.1 M bp
- 1996 – archea *Methanocaldococcus jannaschii*
- 1997 – *E. coli*
- 2000 – *H. sapiens*

UCSC browser

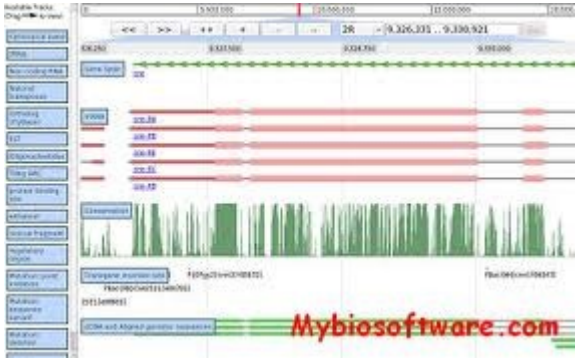
- Jim Kent, 2000
- Napisany w C
- Udostępniany darmowo dla akademickich zastosowań
- Komercyjna licencja Kent Informatics
- Ciągle jedna z bardziej użytecznych przeglądarek
- W zasadzie brak innych developerów

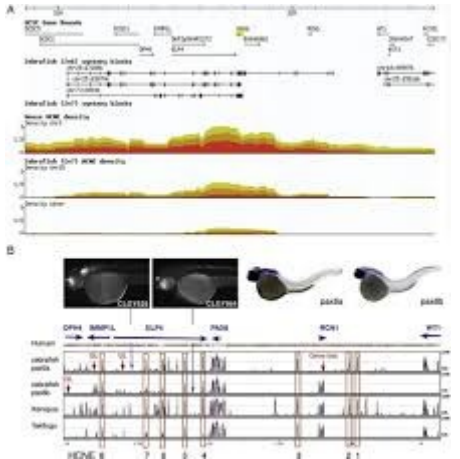




Gbrowse

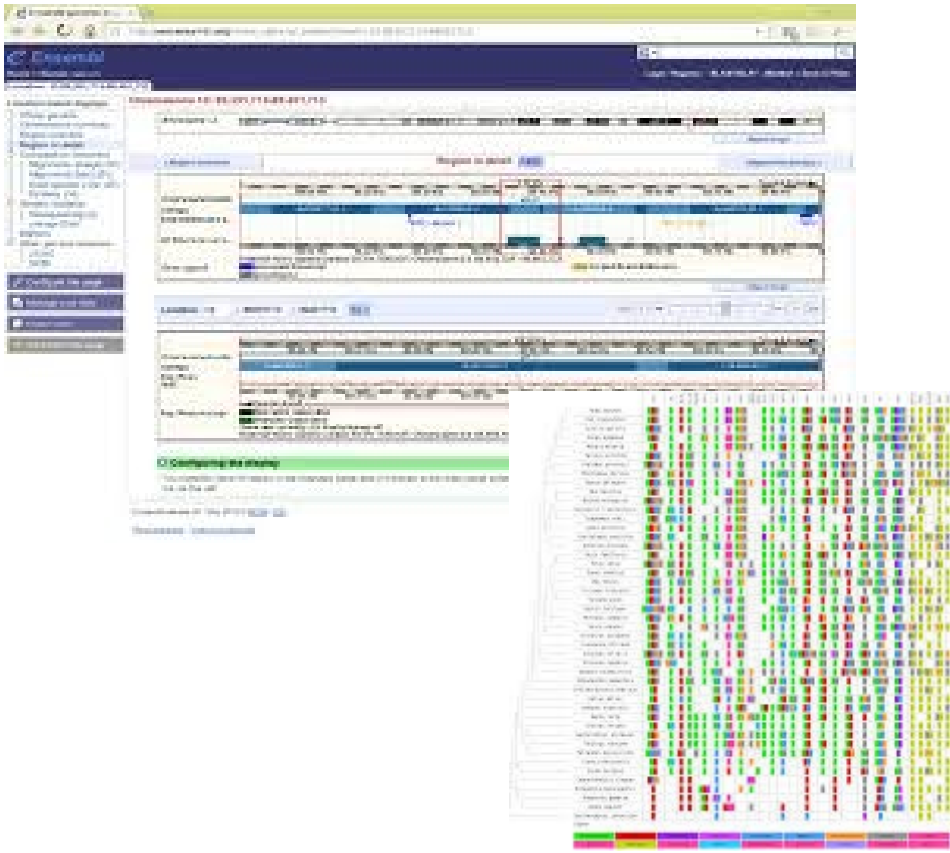
- Część projektu GMOD
- Rozpoczęty w 2002
- PERL artistic license
- Rozwijany głównie w PERLu
- Coraz więcej java-scriptu, począwszy od 2007r.
- Ogromna liczba instalacji
- Jbrowse, WebGbrowse





ENSEMBL

- European Nucleotide Service from EMBL
- Bazuje w EBI niedaleko Cambridge
- Duża baza kodu w Perlu, schemat bazy danych w MySQL
- Licencja Apache
- Do niedawna niewiele genomów, obecnie ogromna liczba genomów I dużo informacji porównawczej

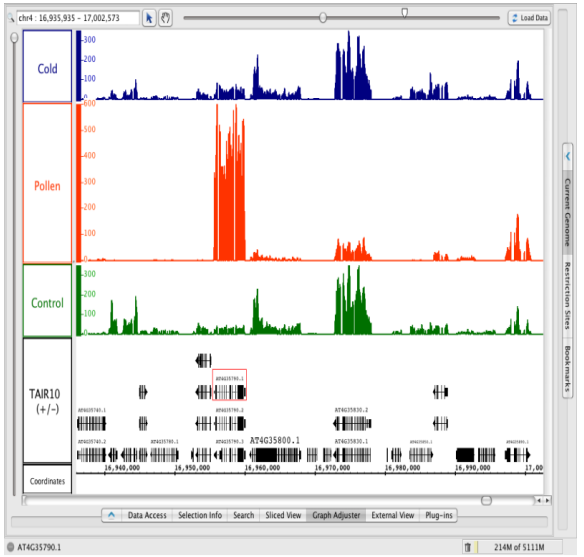


Przeglądarki offline

- Nie zawsze możemy używać przeglądarek online
- Np. możemy nie chcieć udostępniać danych na zewnątrz, albo możemy mieć za dużo danych, aby wysłać je na zewnętrzny serwer
- Rozwiązanie – przeglądarka korzystająca z danych lokalnych, wyświetlająca dane z dysku, ale pobierająca kontekst opisu genomu (geny, transkrypty, itp) ze zdalnego serwera (DAS)
- Typowo napisane w Javie – dostęp do grafiki I przenośność między Mac/Windows/Linux

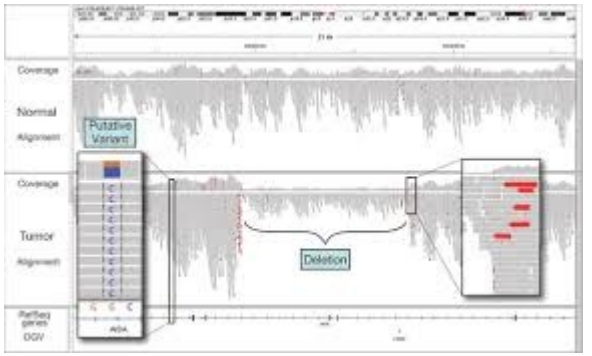
Integrated Genome Browser

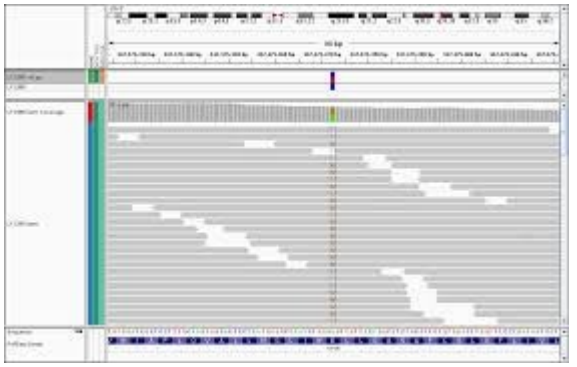
- Kod stworzony początkowo przez firmę Affymetrix do wizualizacji danych z macierzy “kafelkowych” (tiling arrays)
- “Porzucony” przez firmę I obecnie rozwijany w środowisku akademickim (UNC Charlotte)
- Academic free license
- Napisany w Javie, potem usunięto część kodu Affymetrixu

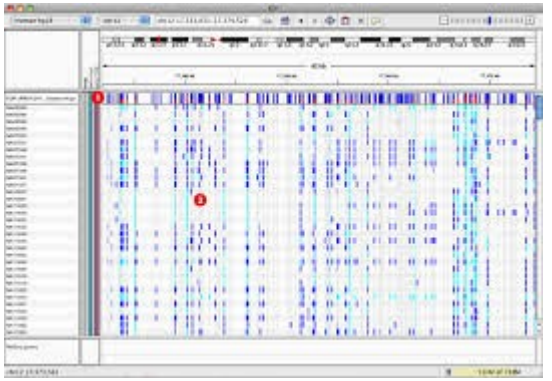


Broad Integrative Genome Viewer

- W Zasadzie klon IGB, choć baza kodu zupełnie nowa
- Rozwój spowodowany brakiem możliwości komercjalizacji przez Broad i problemami technicznymi
- Używany w częściowo komercyjnym genome space
- LGPL licencja







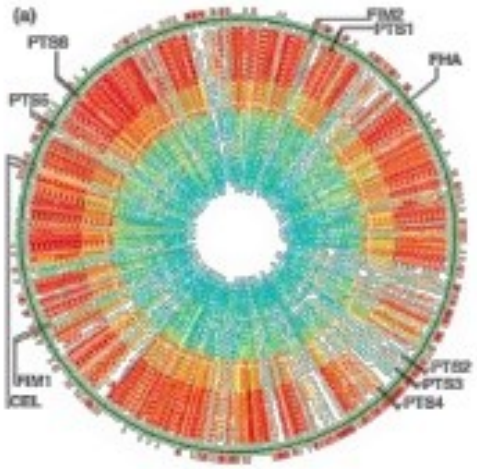
Genome Diagram

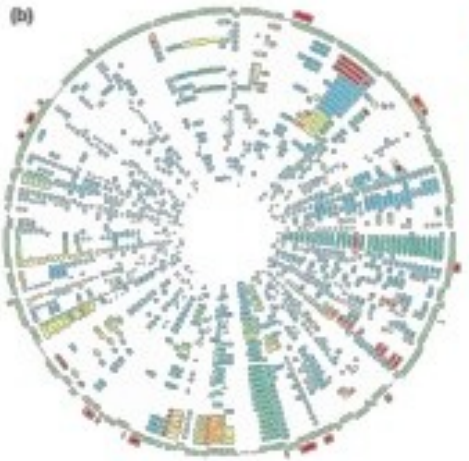
GenomeDiagram

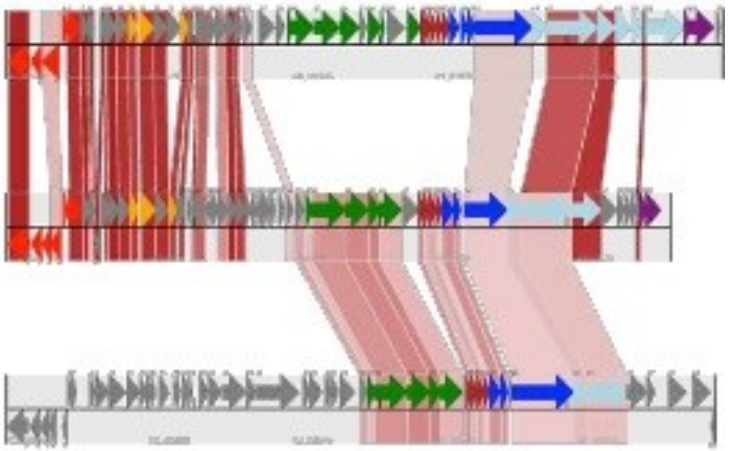
BIOPYTHON

The James Hutton Institute

- Comparative genomics visualisation package
- Developed in 2003 for *Pba* sequencing, later incorporated into Biopython







http://www.biopython.org

Pritchard *et al.* (2006) *Bioinformatics* doi:10.1093/bioinformatics/btk021.

Potencjalne tematy projektów

- http://obf.github.io/GSoC/ideas/
- https://github.com/scikit-learn/scikit-learn/wiki/Google-summer-of-code-(GSOC)-2017
- http://python-gsoc.org/
- http://dreamchallenges.org/
- https://www.kaggle.com/