Research Projects in Biomedicine

AUGUSTUS – a MediGRID pilot application

Berlin, 2007/04/18

Thomas Lingner

Department of Bioinformatics

University of Göttingen

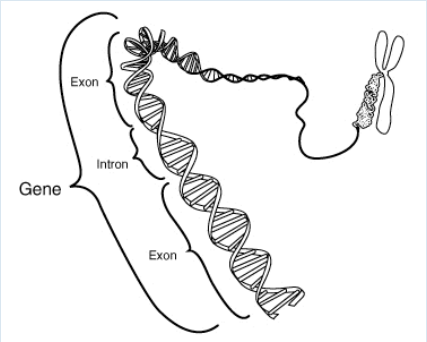
MediGRID

Bundesministerium
für Bildung
und Forschung

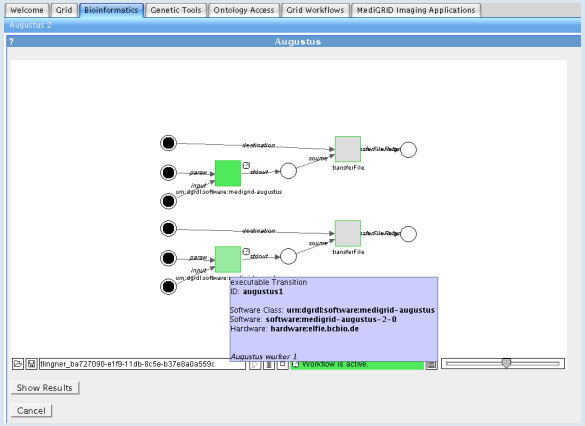
Outline

Research Projects in Biomedicine

using the example of gene prediction



- gene prediction with AUGUSTUS
- AUGUSTUS in biomedical projects
- AUGUSTUS@MediGRID
- related projects
- summary



Gene prediction on DNA sequences

„in silico“ gene finding

- find important functional regions in the genome

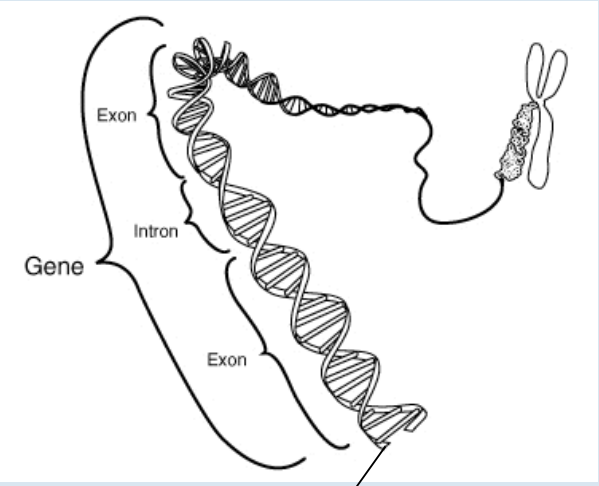
  ➔ genes, promoters ...

- determine coding sequences of gene products

  ➔ proteins, regulatory elements ...

- “wet lab” gene finding is tedious and expensive

- computational methods today commonly used

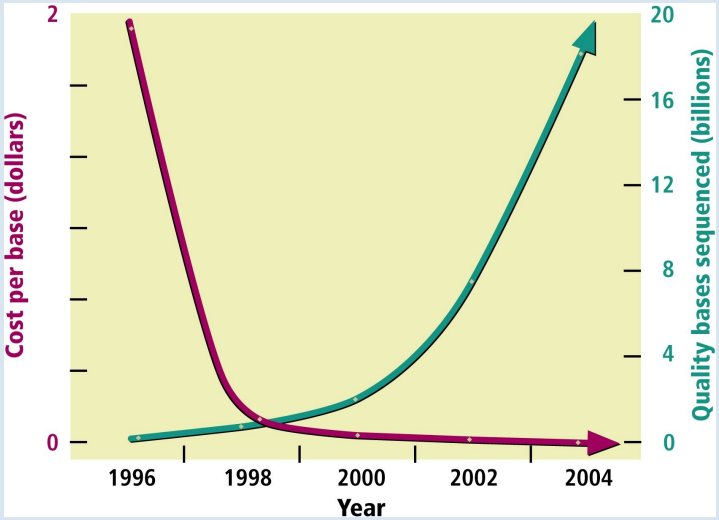


...ATAGCTAGCTAGCTGATCG...

Genome sequencing

"information overload"

- growing number of sequenced genomes
- wide-spread high-throughput techniques



high-throughput shotgun sequencing
(picture by Steve Jurvetson)

development of sequencing quality and cost
(graph from Human Genome Project)

Gene prediction in eukaryotes

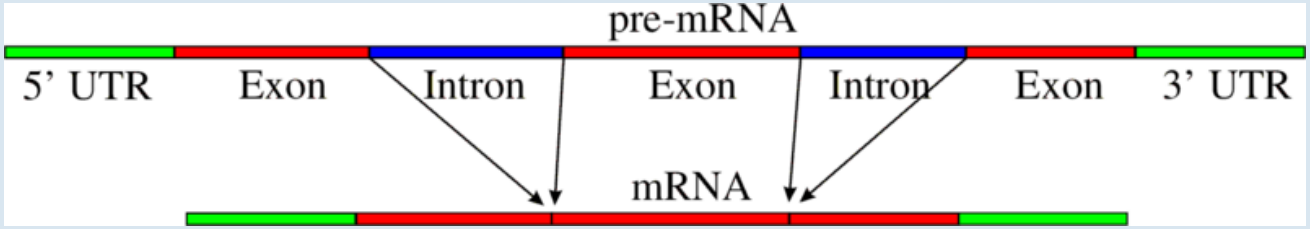
(eukaryotes: organisms with cell nucleus)

- gene structure differs between organisms
- eukaryotic gene structure is complex

  ➔ precise statistical models of gene structure required
- AUGUSTUS uses sophisticated statistical models
- AUGUSTUS developed by Mario Stanke (in C++)

  ✓ constant program improvement and extension

AUG(USTUS) =>
ATG =>
gene start



Community and results

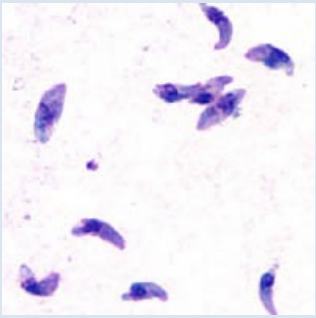
- wide-spread community without computer science background and low computing power
- excellent performance of AUGUSTUS in several comparative studies
  - ✓ AUGUSTUS is used in several international genome projects in collaboration with M. Stanke

AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome
Mario Stanke , Ana Tzvetkova and Burkhard Morgenstern
Genome Biology 2006, **7**(Suppl 1):S11

AUGUSTUS: ab initio prediction of alternative transcripts
Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern
Nucleic Acids Res. 2006 July 1; **34**(Web Server issue): W435–W439.

Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources
Mario Stanke, Oliver Schöffmann, Burkhard Morgenstern, and Stephan Waack
BMC Bioinformatics 2006; **7**: 62.

AUGUSTUS in biomedical projects



Toxoplasma gondii
(Toxoplasmosis)

can cause in enfants:

- *central nervous system disorders*
- *enlargement of the liver and spleen*
- *blindness*
- *mental retardation*

Schistosoma mansoni
(Schistosomiasis)

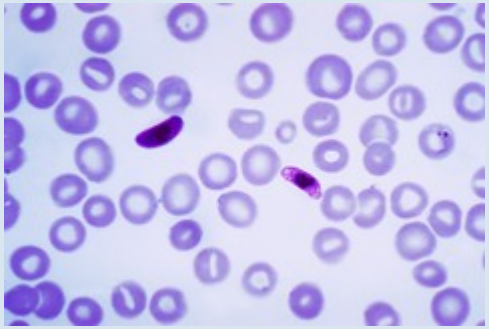
Brugia malayi
(Filariasis)

Elephantiasis of the legs due to filariasis (CDC).

can cause Elephantiasis

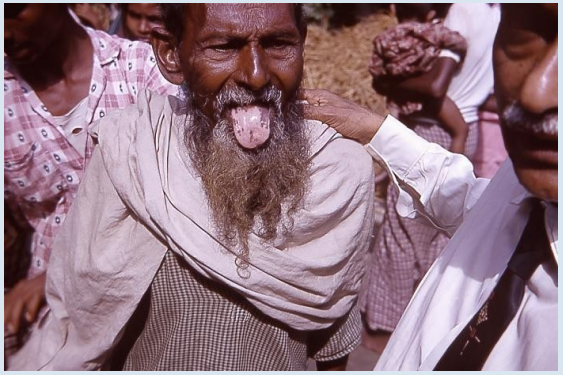
high mortality rate in developing countries

AUGUSTUS in biomedical projects



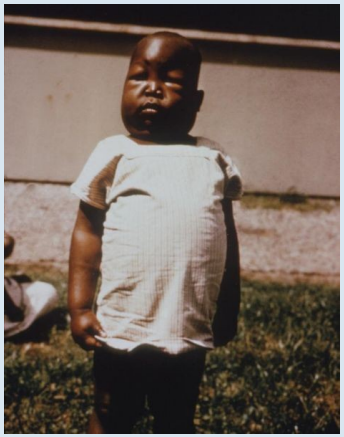
Plasmodium falciparum (Malaria)

accounts for 80% of all human malarial infections and 90% of the deaths



Anaemia caused by malaria

Culex pipiens (transfers Dengue fever, Ross River Fever, Malaria)

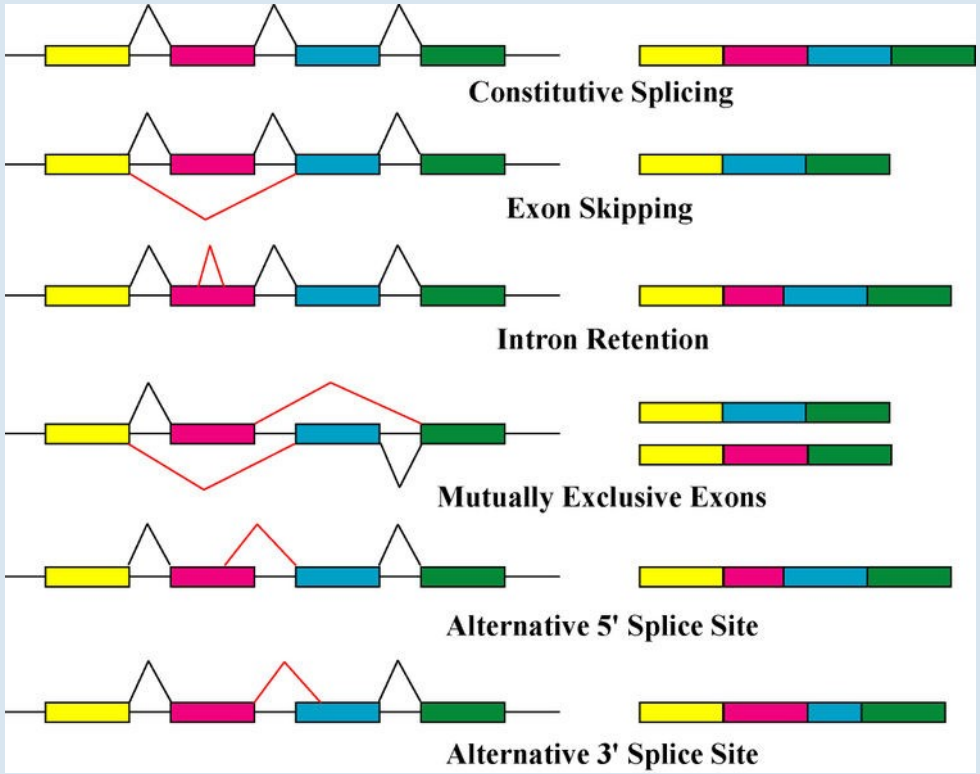


edemata due to chronic renal failure caused by malaria

Aedes aegypti (transfers Yellow fever, Dengue fever)

future projects with medical applications

- effect of single nucleotide mutation on formation of carcinogenic genes in human DNA
  - ➔ mutations can cause modified gene products by alternative splicing
  - ✓ AUGUSTUS can predict alternative splicing



alternative splicing can cause different gene products from a single gene

Genomes are big!

- large amount of input and output data
- centralized results for different job configurations
- large sequence databases can improve prediction

→ ***grid provides massive storage capacity***

Species	***Genome size (Mb)***	***Number of genes***
Saccharomyces cerevisiae	12	5800
Caenorhabditis elegans	97	19000
Arabidopsis thaliana	125	25500
Drosophila melanogaster	180	13700
Oryza sativa	466	45-55000
Mus musculus	2500	29000
Homo sapiens	2900	27000

Why grid? II

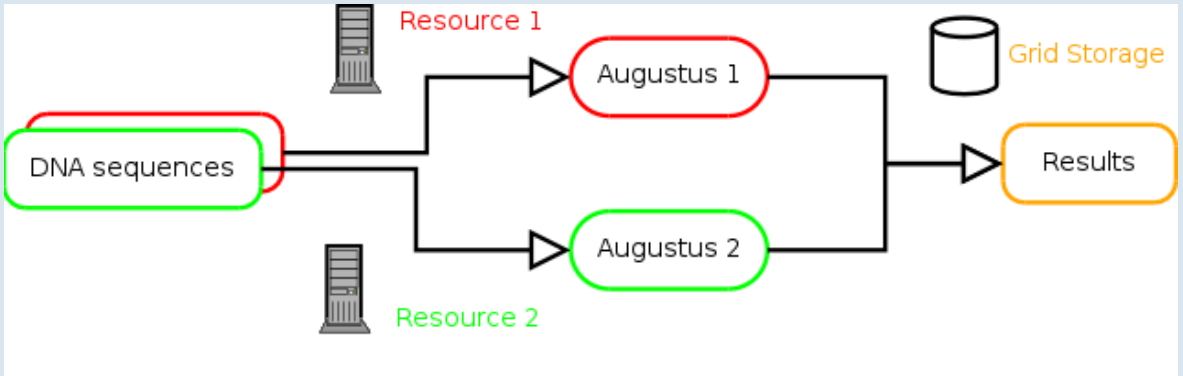
Precise prediction is computationally demanding!
AUGUSTUS uses sophisticated statistical methods

- long running time (up to weeks for genomes)

but:

- easy to split into parallel processes
- discontinuous load

→ ***grid provides distributed resources and balancing***



Why grid? III

Grid portals provide:

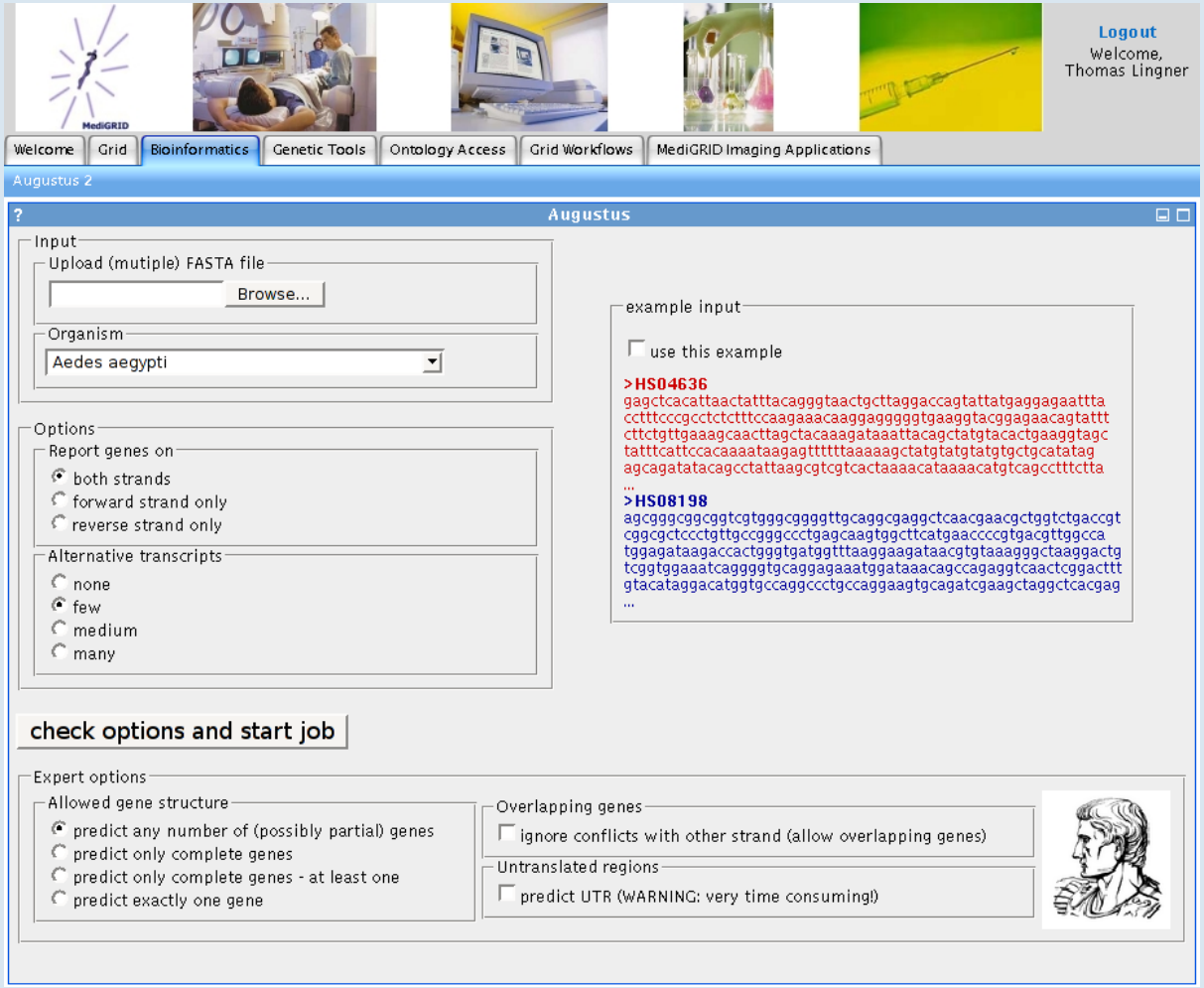
- usability: *complex software - simple interface*
- security: *flexible security infrastructure*

  ➔ *gene prediction requires very few security*
- accessibility: *collaborative environment*
- flexibility: *interchangeable components*
- diversity: *analysis pipelines*
- personal configuration: *“my MediGRID”*
- synergy: *links to related applications etc.*
- portability*: reuse know-how for faster integration*
- *...*

AUGUSTUS@MediGRID

Job Configuration

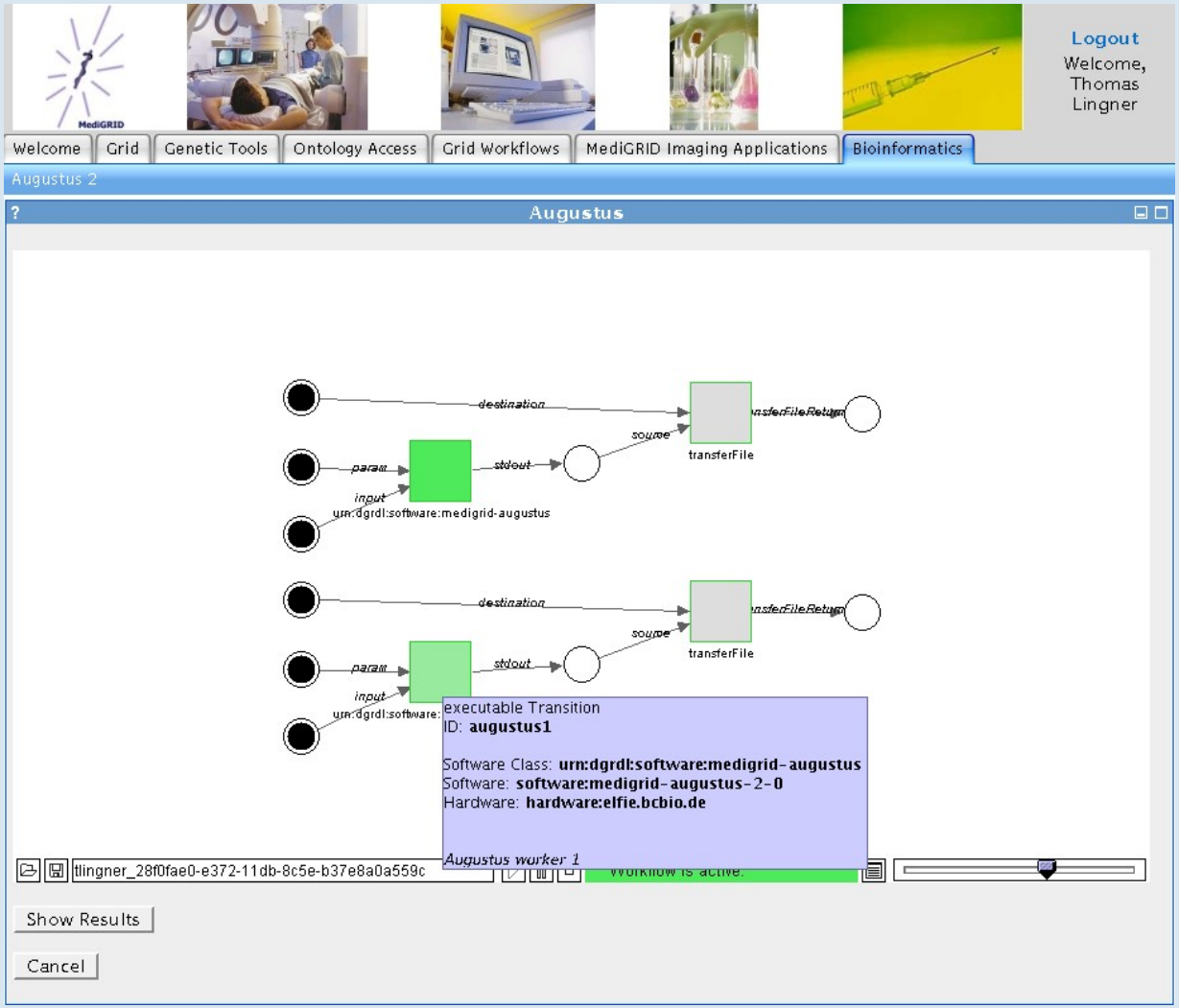
- make usage as easy as possible
- hide grid complexity from user
- check for appropriate options
- support different languages with Gridsphere



AUGUSTUS@MediGRID

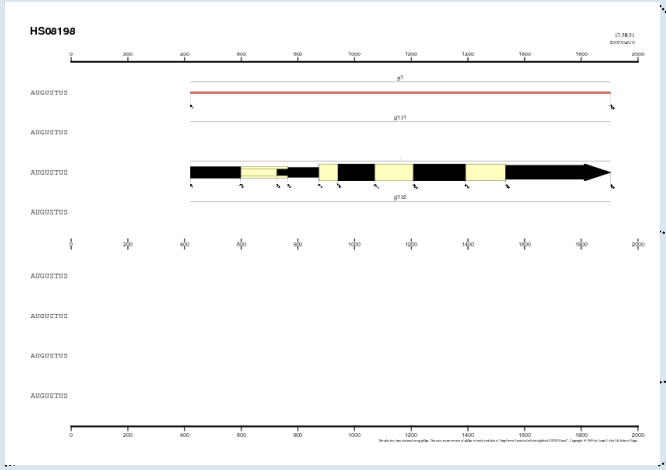
Workflow monitoring and control

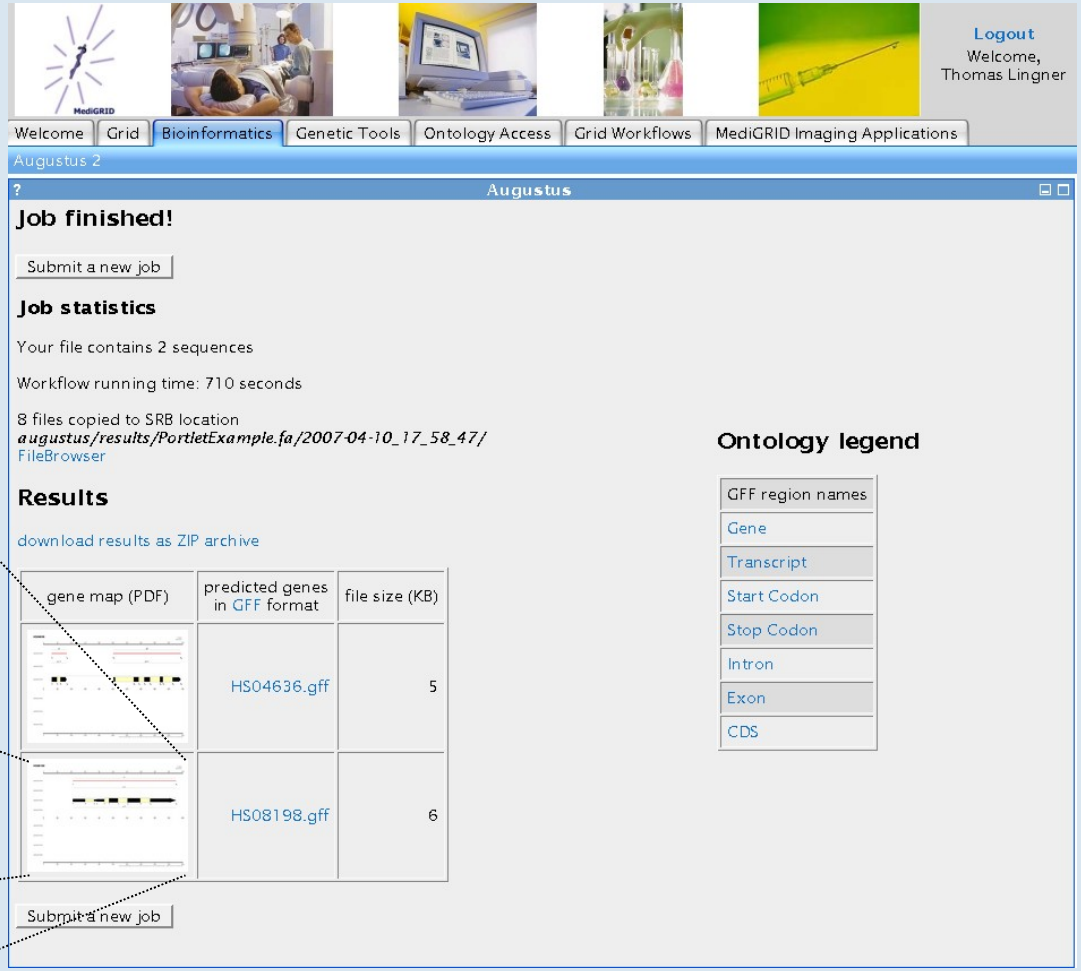
- select resources automatically with Grid Workflow Execution Service and Globus Toolkit 4
- show progress and used resources
- allow modification?



Job results

- provide intuitively interpretable results
- long term file storage with SRB (Storage Resource Broker)



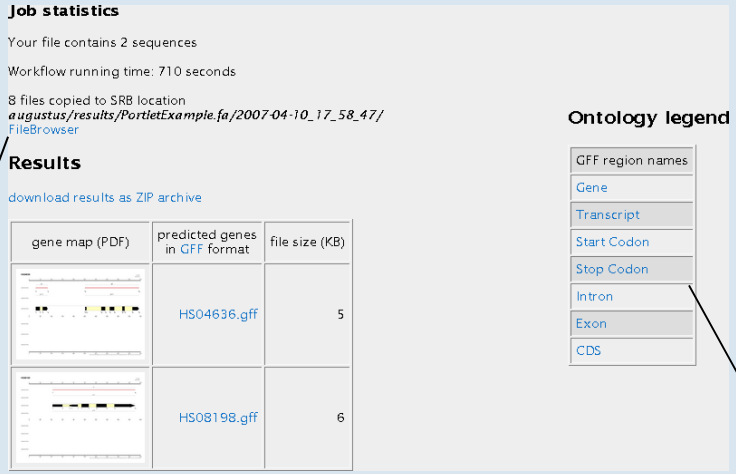


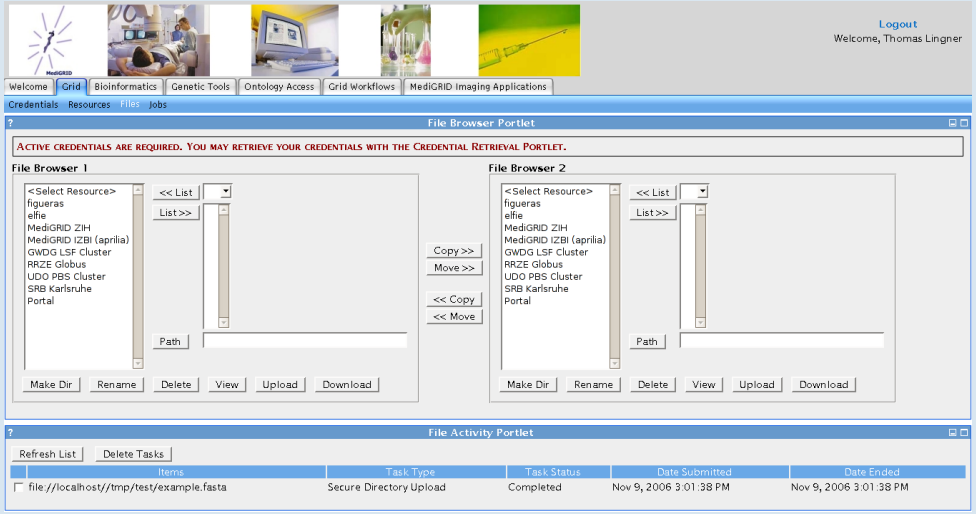
Portal synergy

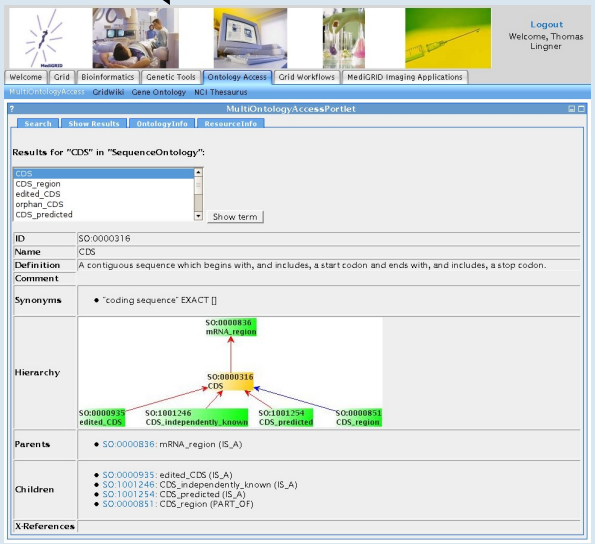
- links to other MediGRID applications

link to grid file browser portlet

link to ontology portlet





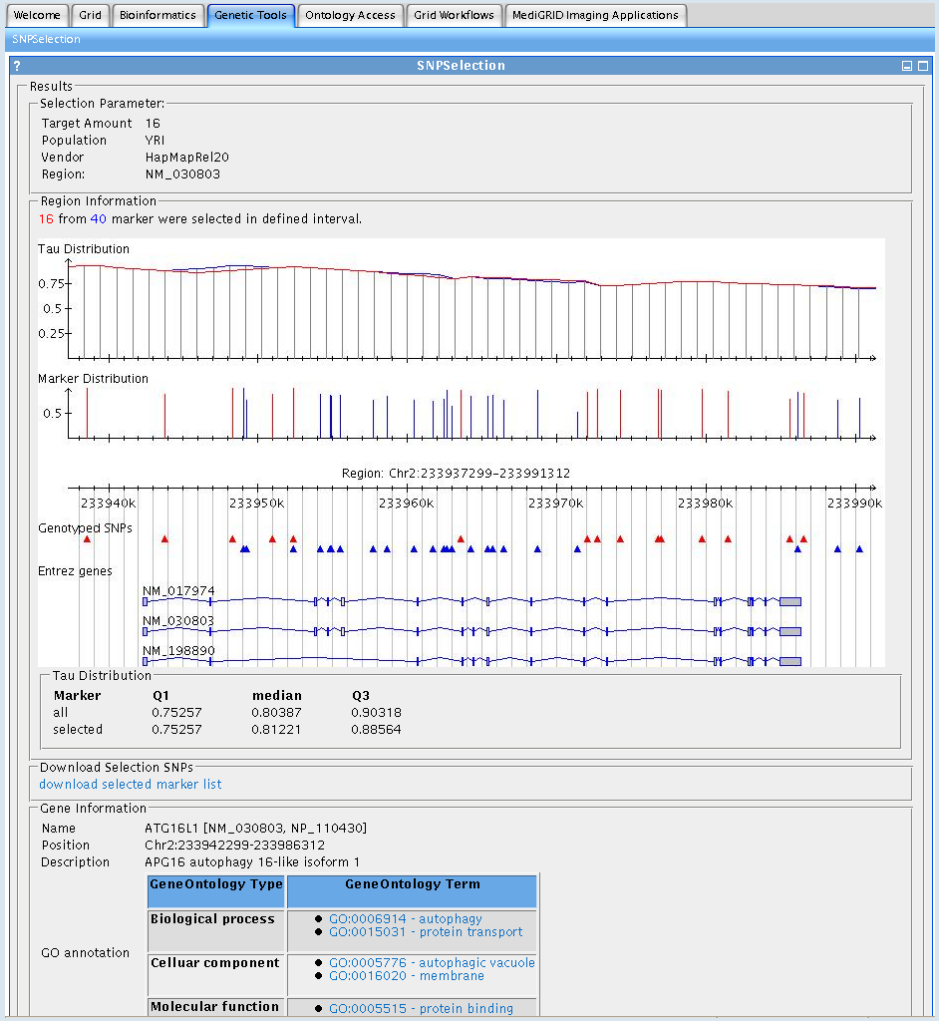


Related projects in MediGRID

SNPSelection

SNP: single nucleotide polymorphism

- entropy-based selection of genetic markers for genome-wide disease association studies
  - ✓ *detected Sarcoidosis gene*



Sarcoidosis is associated with a truncating splice site mutation in BTNL2

Ruta Valentonyte, Jochen Hampe, Klaus Huse, Philip Rosenstiel, Mario Albrecht, Annette Stenzel, Marion Nagy, Karoline I Gaede, Andre Franke, Robert Haesler, Andreas Koch, Thomas Lengauer, Dirk Seegert, Norbert Reiling, Stefan Ehlers, Eberhard Schwinger, Matthias Platzer, Michael Krawczak, Joachim Müller-Quernheim, Manfred Schürmann & Stefan Schreiber

Nature Genetics **37**, 357 - 364 (2005)

Related projects in MediGRID

Coming soon:

- ***DiAlign***: high quality DNA/protein sequence alignments
  - → *Genome comparison for analysis of molecular evolution*
- ***Genomizer****:* resolve individual genetic risk profiles
  - → *analysis of genome-wide disease association studies*
- ***SequCorr****:* platform for correlation analysis of biological data
  - → *find correlation between sequence and image data*
- ***RNAi annotation pipeline***: analysis of molecular pathways
  - → *predict protein structure properties and function*

applications profit from grid differently

Summary

Biomedical projects in grid environments

- allow use of distributed computing power and storage
- make tools and results world-wide accessible
- are easily extendable

→ ***can significantly speed up medical research***

Thanks for your attention!