

Research Projects in Biomedicine

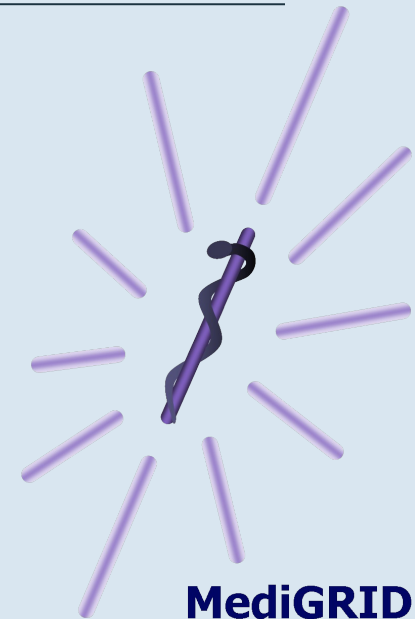
AUGUSTUS – a MediGRID pilot application

Berlin, 2007/04/18

Thomas Lingner

Department of Bioinformatics

University of Göttingen

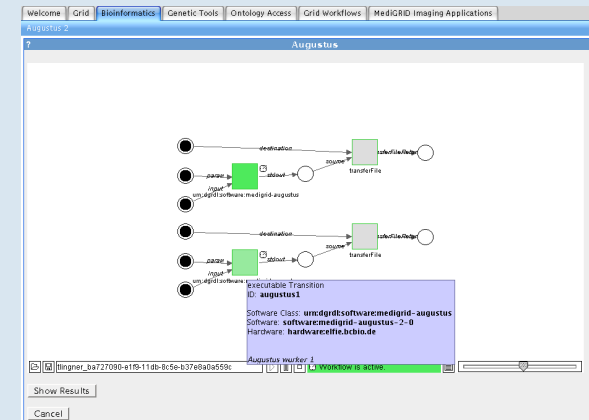
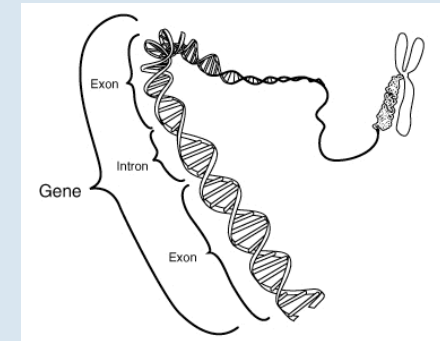


MediGRID

Research Projects in Biomedicine

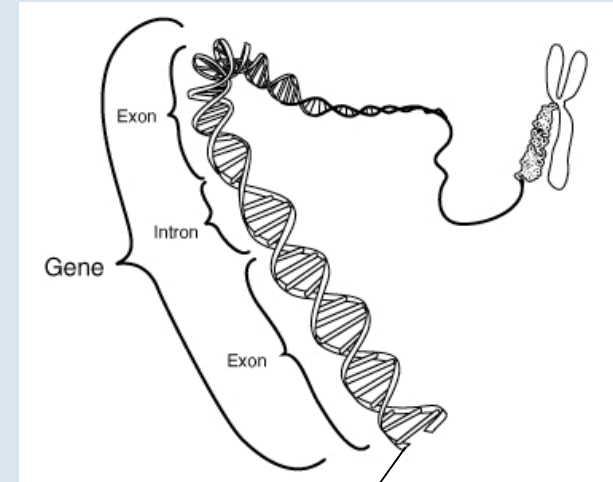
using the example of gene prediction

- 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

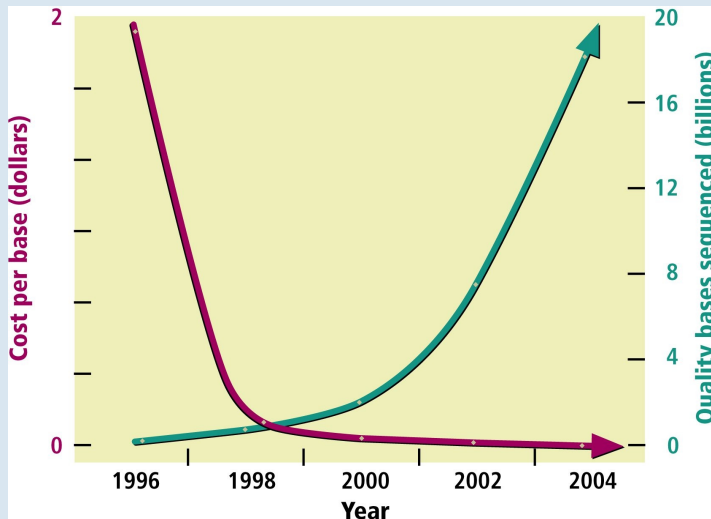
“information overload”

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



*high-throughput shotgun sequencing
(picture by Steve Jurvetson)*

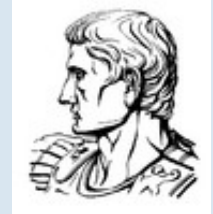
*development of sequencing quality and cost
(graph from Human Genome Project)*



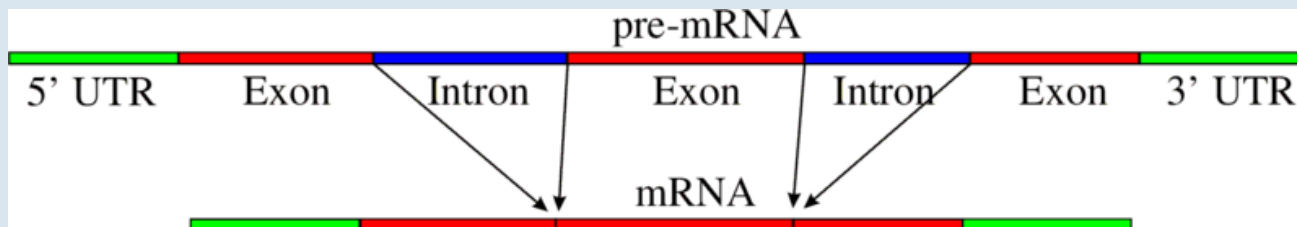
Gene prediction in eukaryotes

(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

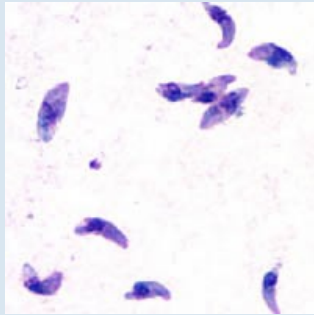
Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern

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.



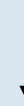
Toxoplasma gondii
(Toxoplasmosis)

can cause in infants:

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



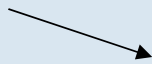
Schistosoma mansoni
(Schistosomiasis)



high mortality rate in developing countries

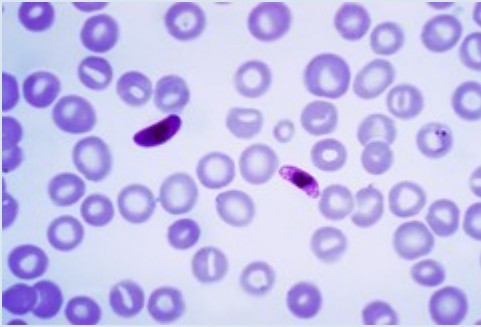


Brugia malayi
(Filariasis)



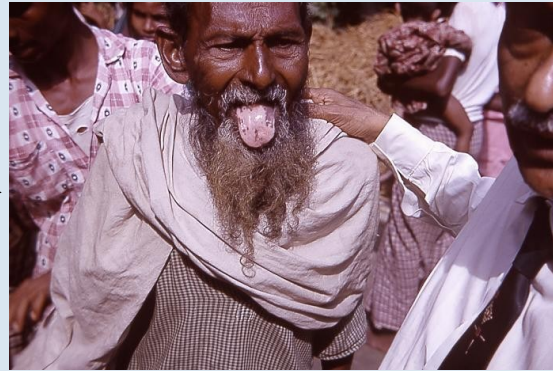
Elephantiasis of the legs due to filariasis (CDC).

can cause Elephantiasis



Plasmodium falciparum (Malaria)

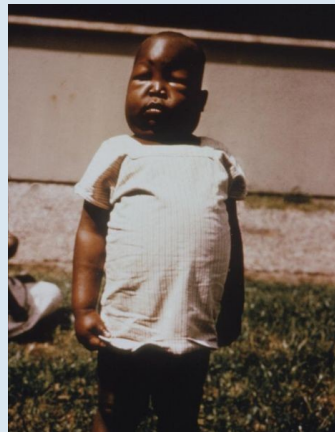
accounts for 80% of all human malarial infections and 90% of the deaths



Anaemia caused by malaria



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



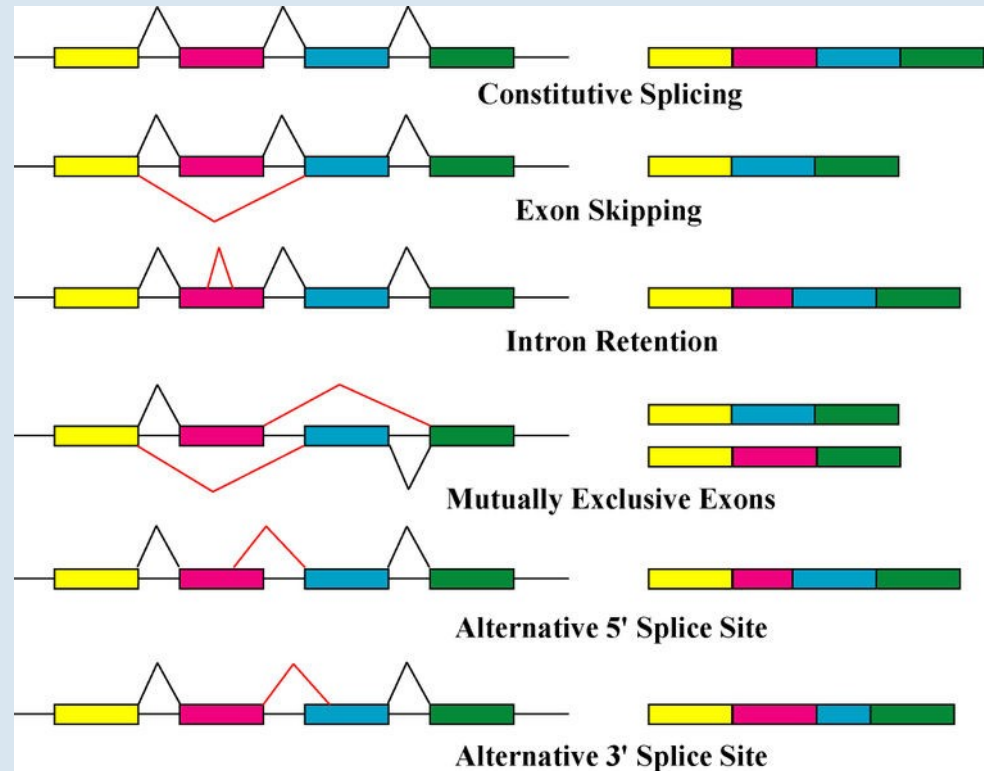
edemata due to chronic renal failure caused by malaria



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

Precise prediction is computationally demanding!

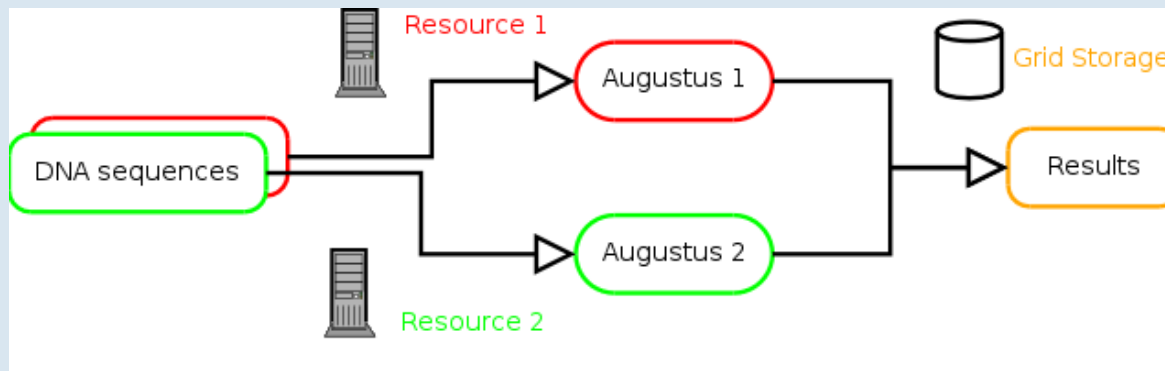
AUGUSTUS uses sophisticated statistical methods

- long running time (up to weeks for genomes)

but:

- easy to split into parallel processes
- discontinuous load

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

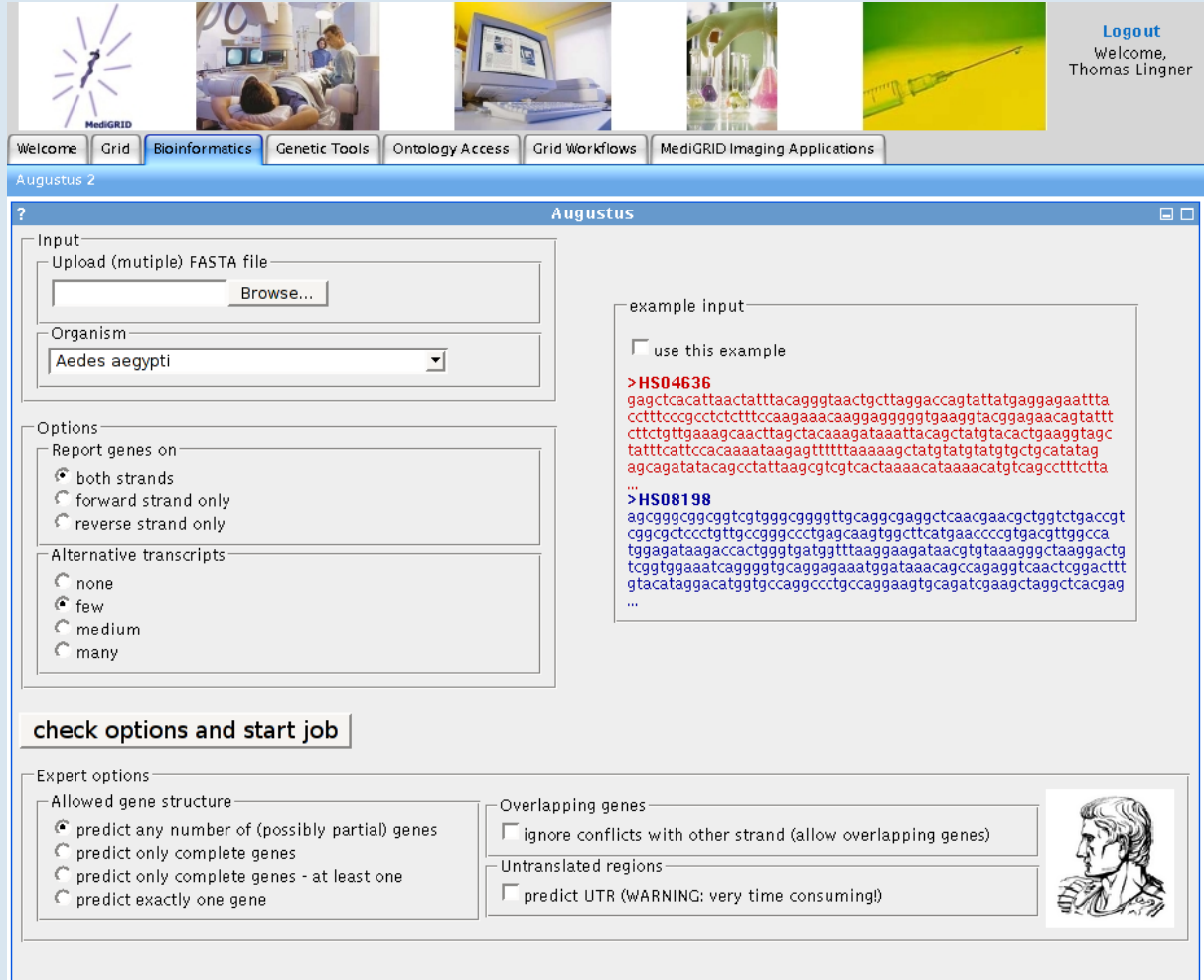


Grid portals provide:

- 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*
- ...

Job Configuration

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



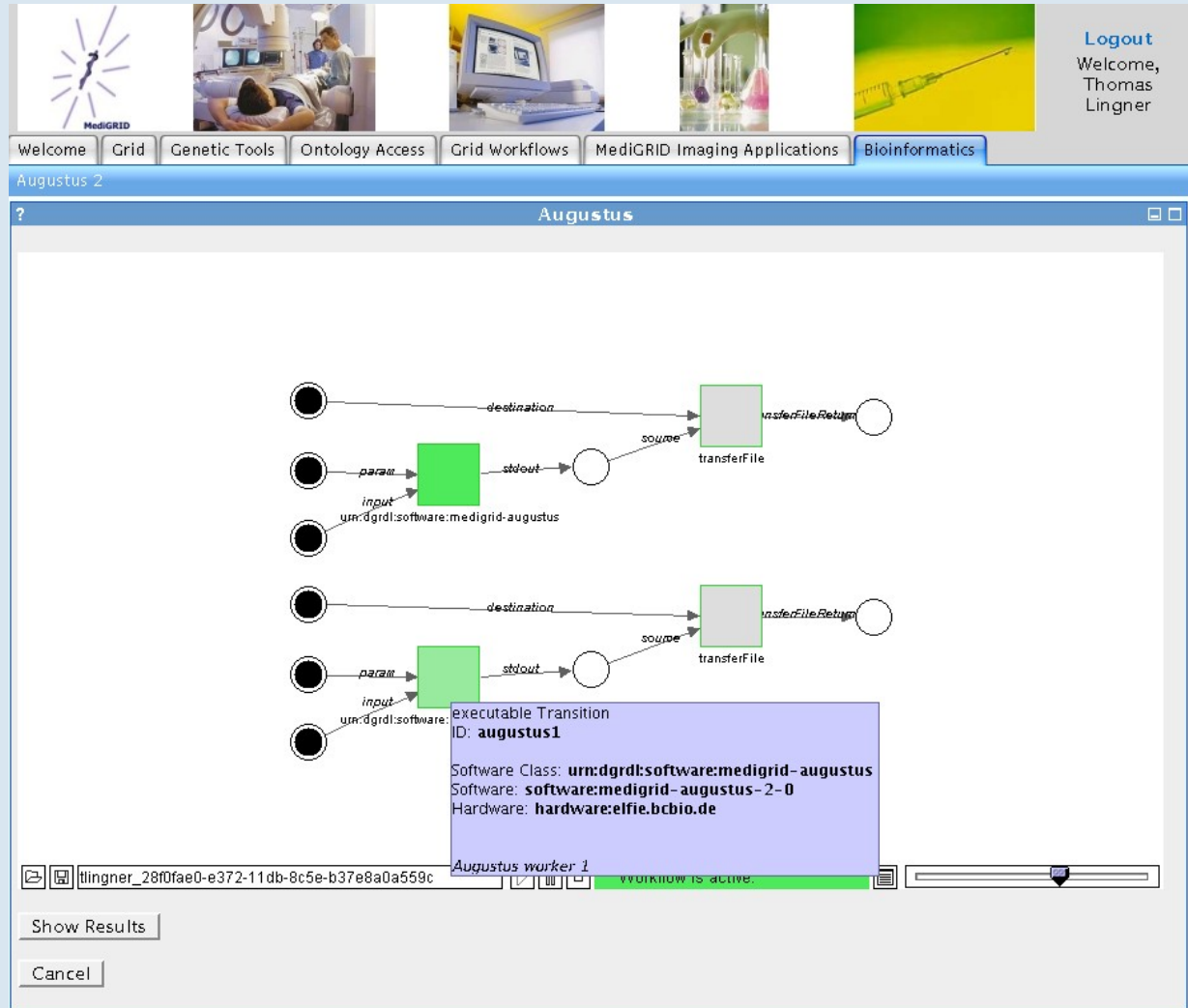
The screenshot shows the Augustus web interface. At the top, there are navigation tabs: Welcome, Grid, Bioinformatics (selected), Genetic Tools, Ontology Access, Grid Workflows, and MediGRID Imaging Applications. The main content area is titled "Augustus" and contains the following sections:

- Input:**
 - Upload (multiple) FASTA file: [text input] Browse...
 - Organism: Aedes aegypti (dropdown menu)
- Options:**
 - Report genes on:
 - both strands
 - forward strand only
 - reverse strand only
 - Alternative transcripts:
 - none
 - few
 - medium
 - many
- check options and start job** (button)
- Expert options:**
 - Allowed gene structure:
 - predict any number of (possibly partial) genes
 - predict only complete genes
 - predict only complete genes - at least one
 - predict exactly one gene
 - Overlapping genes:
 - ignore conflicts with other strand (allow overlapping genes)
 - Untranslated regions:
 - predict UTR (WARNING: very time consuming!)

On the right side, there is a "Logout" button and a welcome message: "Welcome, Thomas Lingner". Below the expert options, there is a small portrait of a man.

Workflow monitoring and control

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



Welcome | Grid | Genetic Tools | Ontology Access | Grid Workflows | MediGRID Imaging Applications | Bioinformatics

Augustus 2

Augustus

executable Transition
 ID: **augustus1**
 Software Class: **urn:dgrid:software:medigrid-augustus**
 Software: **software:medigrid-augustus-2-0**
 Hardware: **hardware:elfie.bcbio.de**

Augustus worker 1

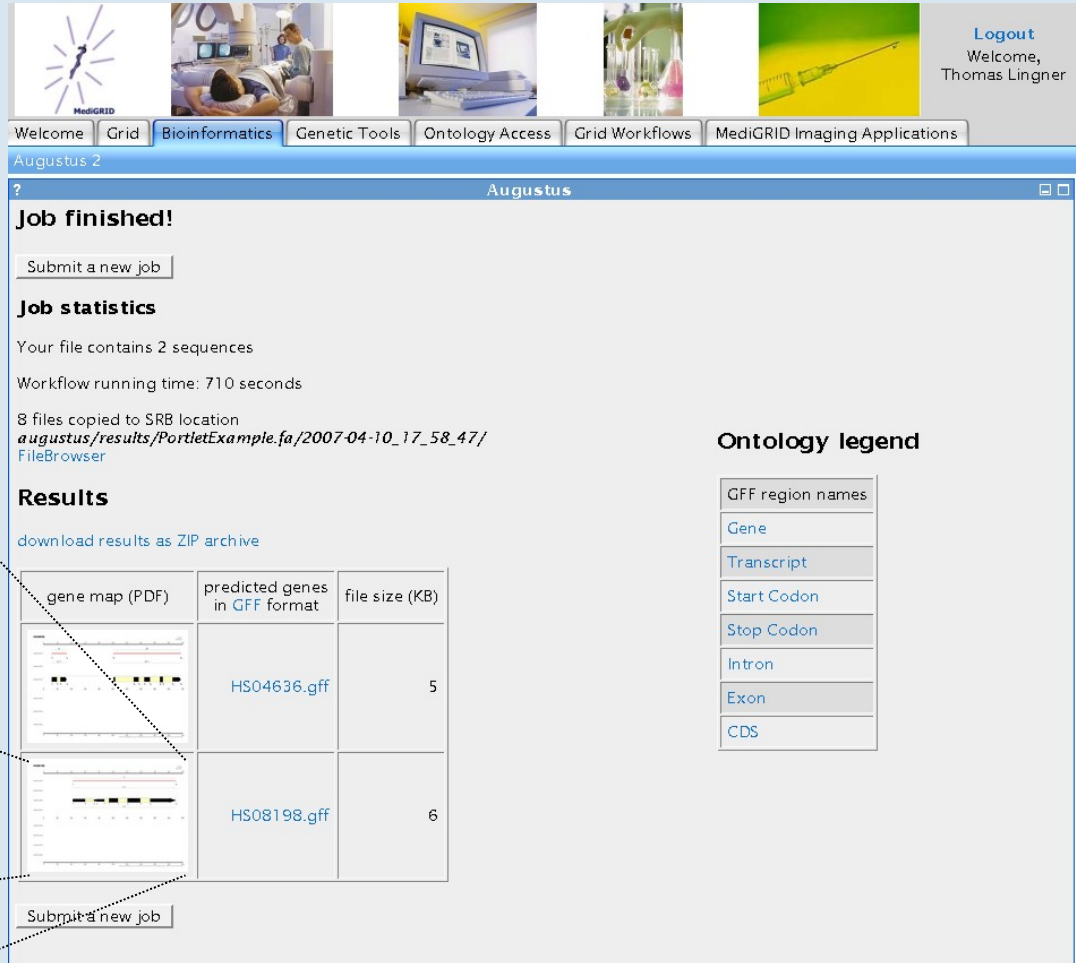
workflow is active.

Show Results

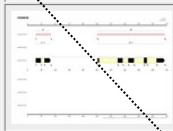
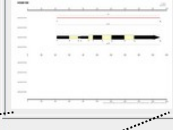
Cancel

Job results

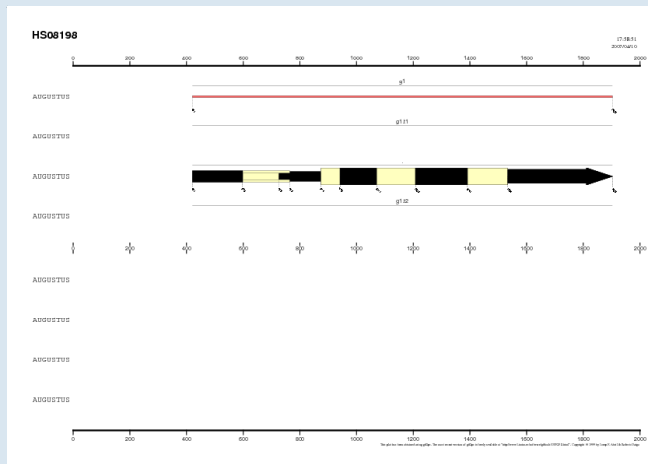
- provide intuitively interpretable results
- long term file storage with SRB (Storage Resource Broker)



The screenshot shows the Augustus web interface. At the top, there are navigation tabs: Welcome, Grid, Bioinformatics, Genetic Tools, Ontology Access, Grid Workflows, and MediGRID Imaging Applications. The main content area displays a "Job finished!" message with a "Submit a new job" button. Below this, "Job statistics" indicate that the file contains 2 sequences and the workflow running time was 710 seconds. A link to "FileBrowser" is provided for the results. The "Results" section includes a "download results as ZIP archive" link and a table of predicted genes.

gene map (PDF)	predicted genes in GFF format	file size (KB)
	HS04636.gff	5
	HS08198.gff	6

On the right side, an "Ontology legend" lists GFF region names: Gene, Transcript, Start Codon, Stop Codon, Intron, Exon, and CDS.



- links to other MediGRID applications

link to grid file browser portlet

Job statistics

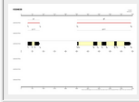
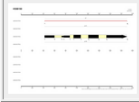
Your file contains 2 sequences

Workflow running time: 710 seconds

8 files copied to SRB location
[augustus/results/PortletExample_fa/2007-04-10_17_58_47/](#)
[FileBrowser](#)

Results

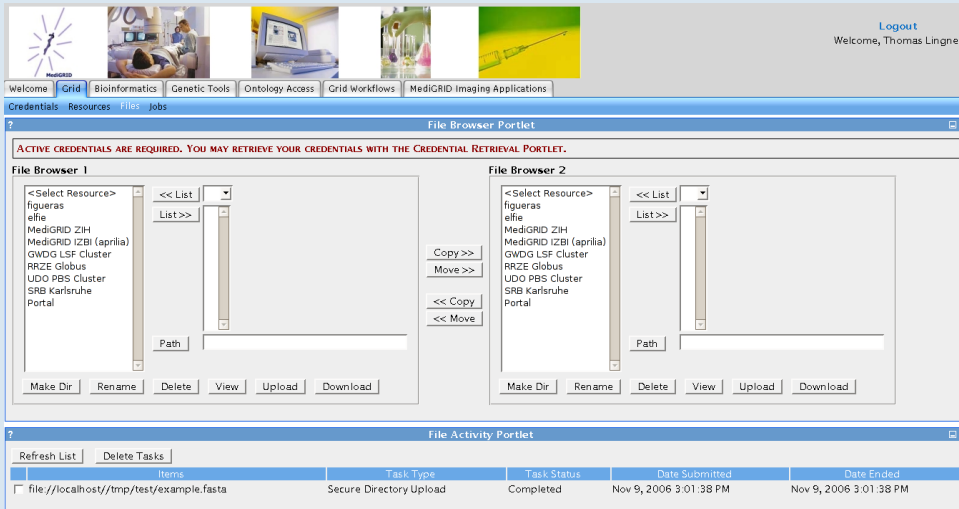
[download results as ZIP archive](#)

gene map (PDF)	predicted genes in GFF format	file size (KB)
	HS04636.gff	5
	HS08198.gff	6

Ontology legend

GFF region names
Gene
Transcript
Start Codon
Stop Codon
Intron
Exon
CDS

link to ontology portlet



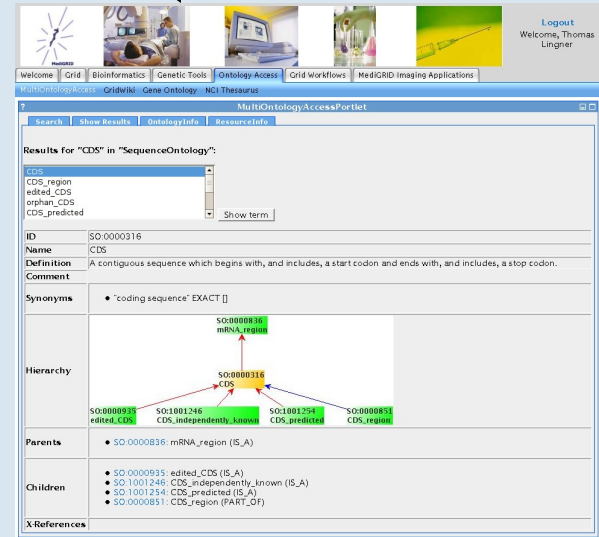
Welcome | Grid | Bioinformatics | Genetic Tools | Ontology Access | Grid Workflows | MediGRID Imaging Applications | Logout
Welcome, Thomas Lingner

File Browser Portlet

ACTIVE CREDENTIALS ARE REQUIRED. YOU MAY RETRIEVE YOUR CREDENTIALS WITH THE CREDENTIAL RETRIEVAL PORTLET.

File Activity Portlet

Items	Task Type	Task Status	Date Submitted	Date Ended
file //localhost/tmp/test/example.fasta	Secure Directory Upload	Completed	Nov 9, 2006 3:01:38 PM	Nov 9, 2006 3:01:38 PM



Welcome | Grid | Bioinformatics | Genetic Tools | Ontology Access | Grid Workflows | MediGRID Imaging Applications | Logout
Welcome, Thomas Lingner

MultiOntologyAccessPortlet

Results for "CDS" in "SequenceOntology":

- CDS
- CDS_region
- edited_CDS
- orphan_CDS
- CDS_predicted

ID: SO:0000316
Name: CDS
Definition: A contiguous sequence which begins with, and includes, a start codon and ends with, and includes, a stop codon.
Comment:

Synonyms:

- "coding sequence" EXACT []

Hierarchy:

```

graph TD
    SO0000316[CDS] --> SO0000316[mRNA_region]
    SO0000316 --> SO0000316[edited_CDS]
    SO0000316 --> SO0000316[CDS_independently_known]
    SO0000316 --> SO0000316[CDS_predicted]
    SO0000316 --> SO0000316[CDS_region]

```

Parents:

- SO:0000316: mRNA_region (IS_A)

Children:

- SO:0000316: edited_CDS (IS_A)
- SO:1001246: CDS_independently_known (IS_A)
- SO:1001254: CDS_predicted (IS_A)
- SO:0000316: CDS_region (PART_OF)

X-References:

SNPSelection

*SNP: single nucleotide
polymorphism*

- entropy-based selection of genetic markers for genome-wide disease association studies

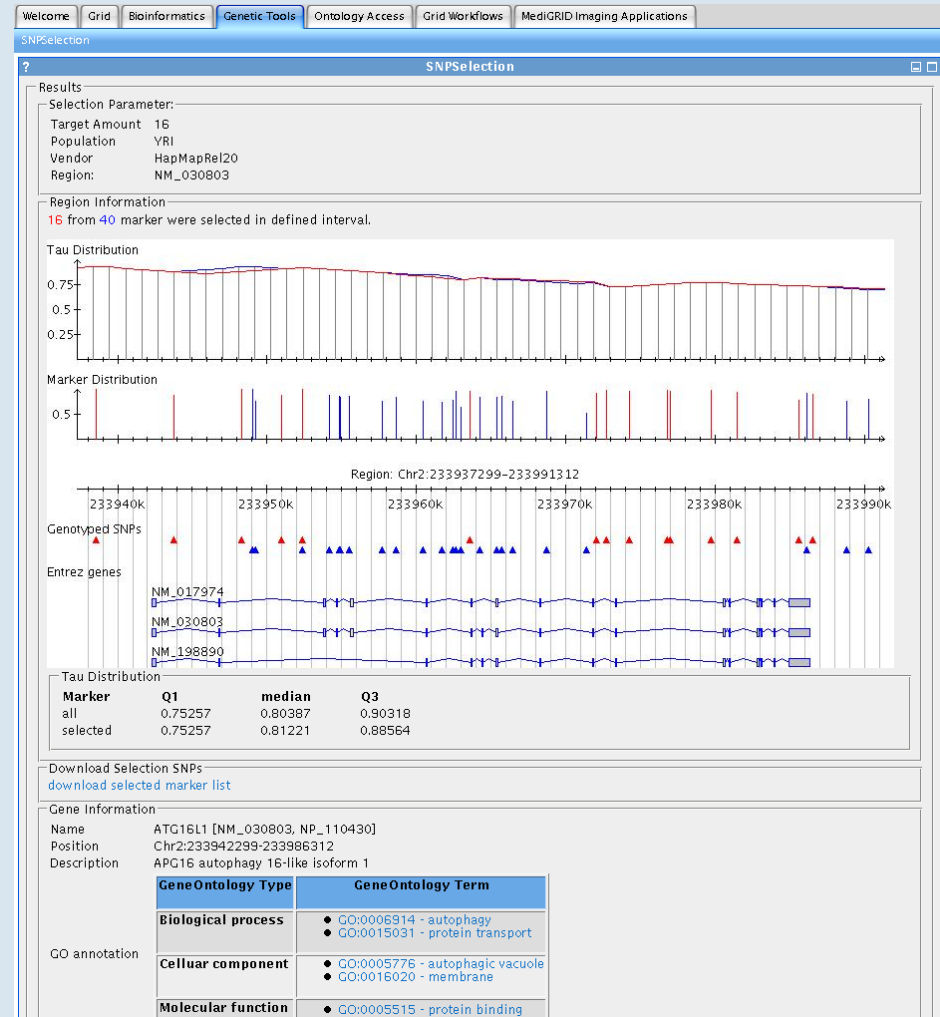
✓ *detected*

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 Seeger, Norbert Reiling, Stefan Ehlers, Eberhard Schwinger, Matthias Platzer, Michael Krawczak, Joachim Müller-Quernheim, Manfred Schürmann & Stefan Schreiber

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



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



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!