

Integrating Sun Grid Engine into the Genome Annotation System GenDB

Alexander Goesmann

Bioinformatics Resource Facility (BRF)

Center for BioTechnology (CeBiTec)

Bielefeld University

Germany

June 22nd 2004

Heterogeneous Data from Genomic Explorations

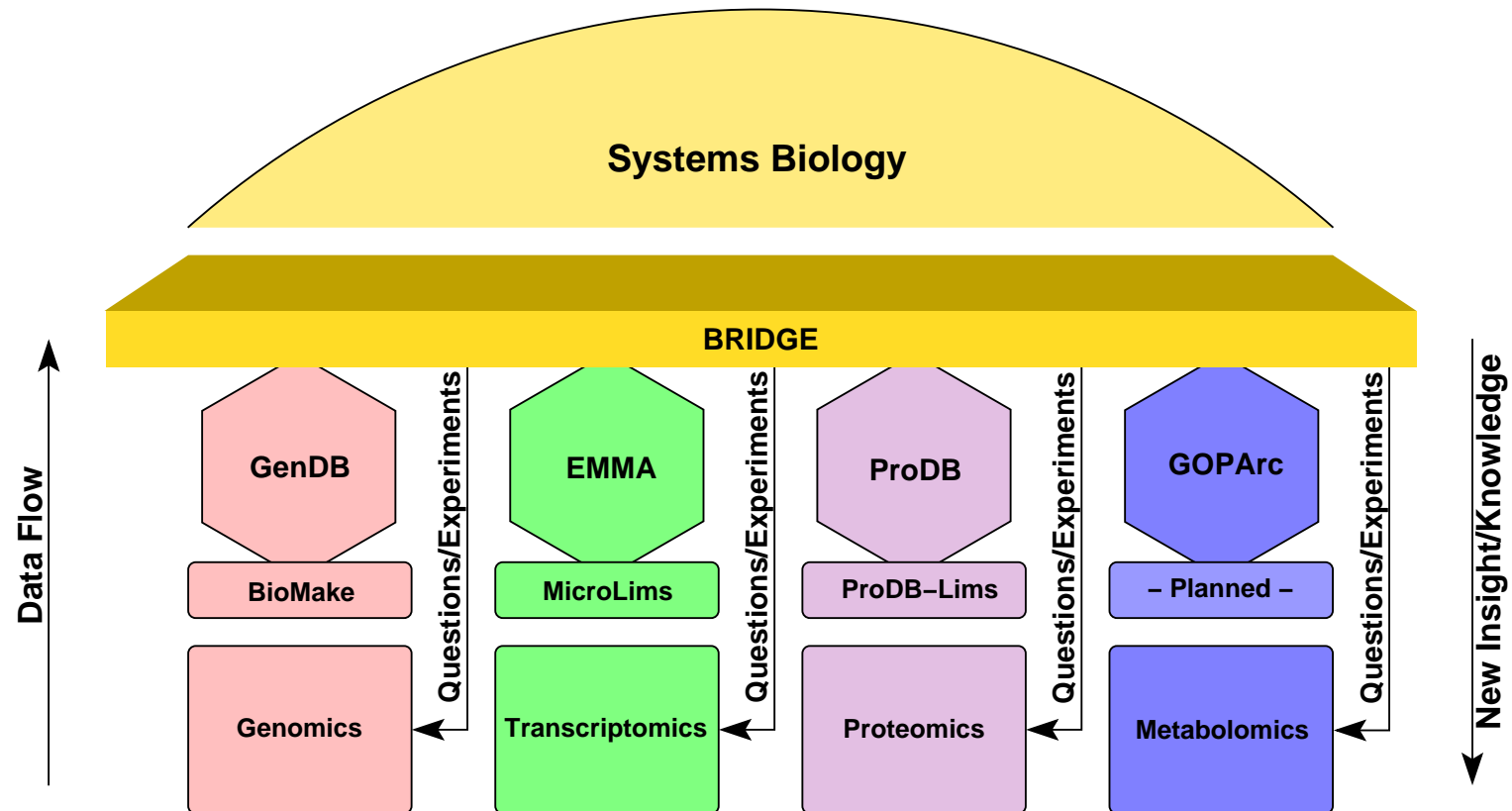
- more than 1000 genomes finished / in progress
- genome annotation has become a standard task
- transcriptome & proteome data is available
- metabolome data is coming up

⇒ data integration is indispensable

BRIDGE – A Platform for Systems Biology

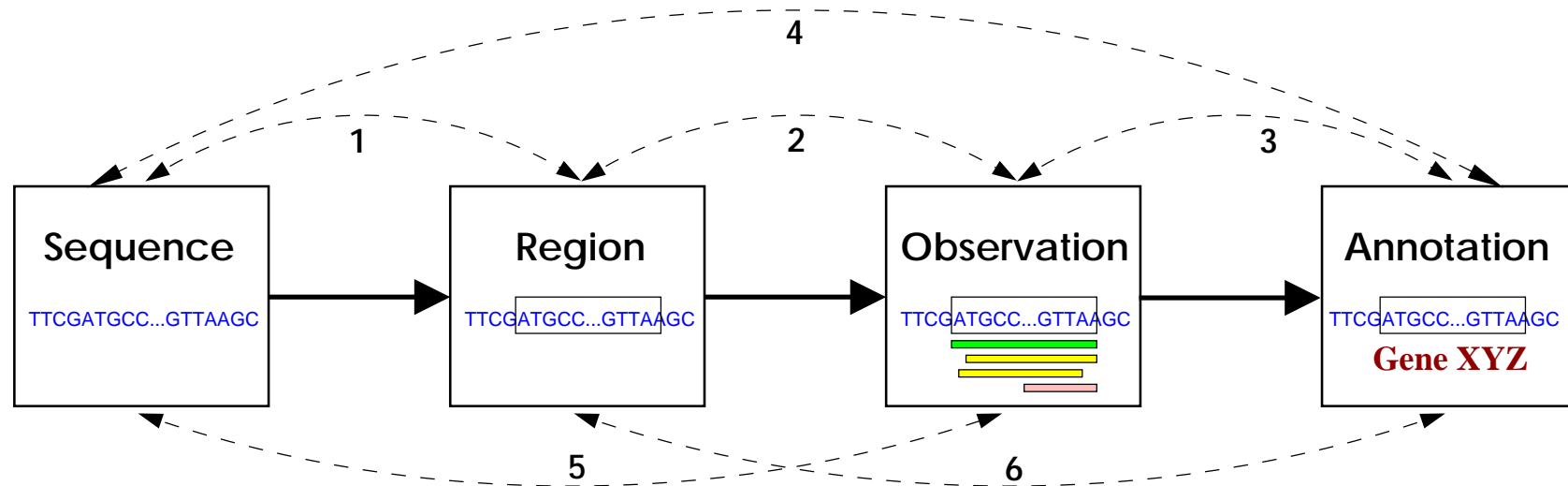
- applications developed @CeBiTec:
GenDB, EMMA, ProDB, GOPArc
 - specialized components for separate scopes
- ⇒ idea: build common interface for programmers & users
- top level layer for data integration: BRIDGE

BRIDGE – A Platform for Systems Biology



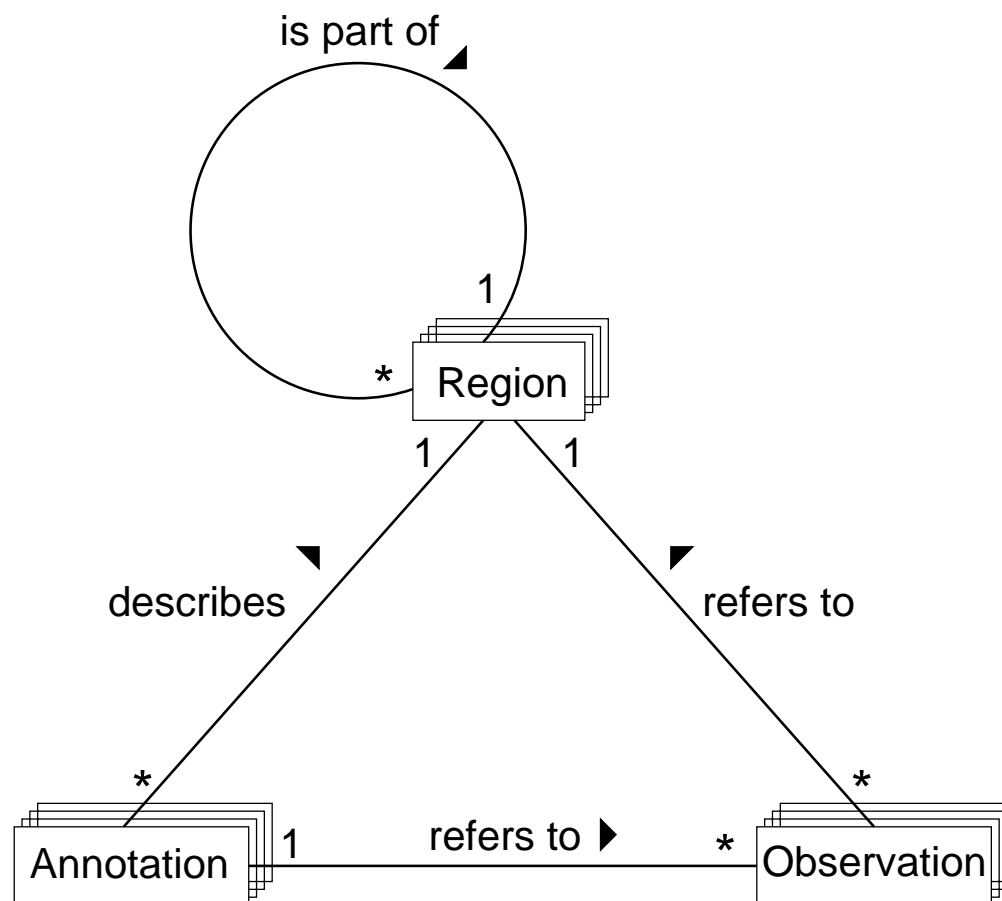
Goesmann *et al.*, *J. Biotechnol.* 2003

A Genome Annotation Pipeline



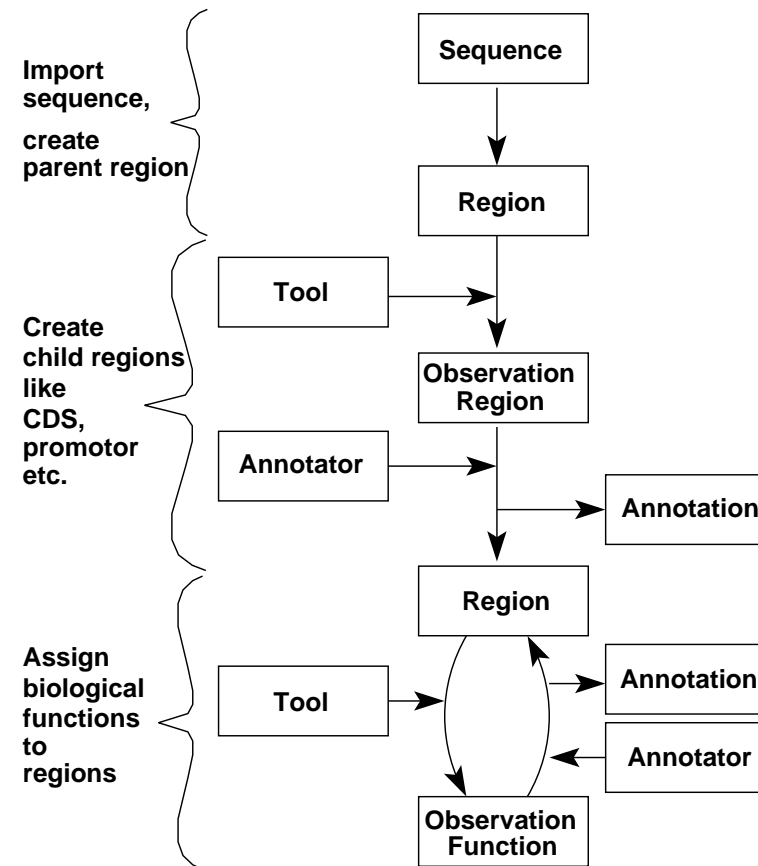
⇒ this process is automated by a genome annotation system

The GenDB Data Model



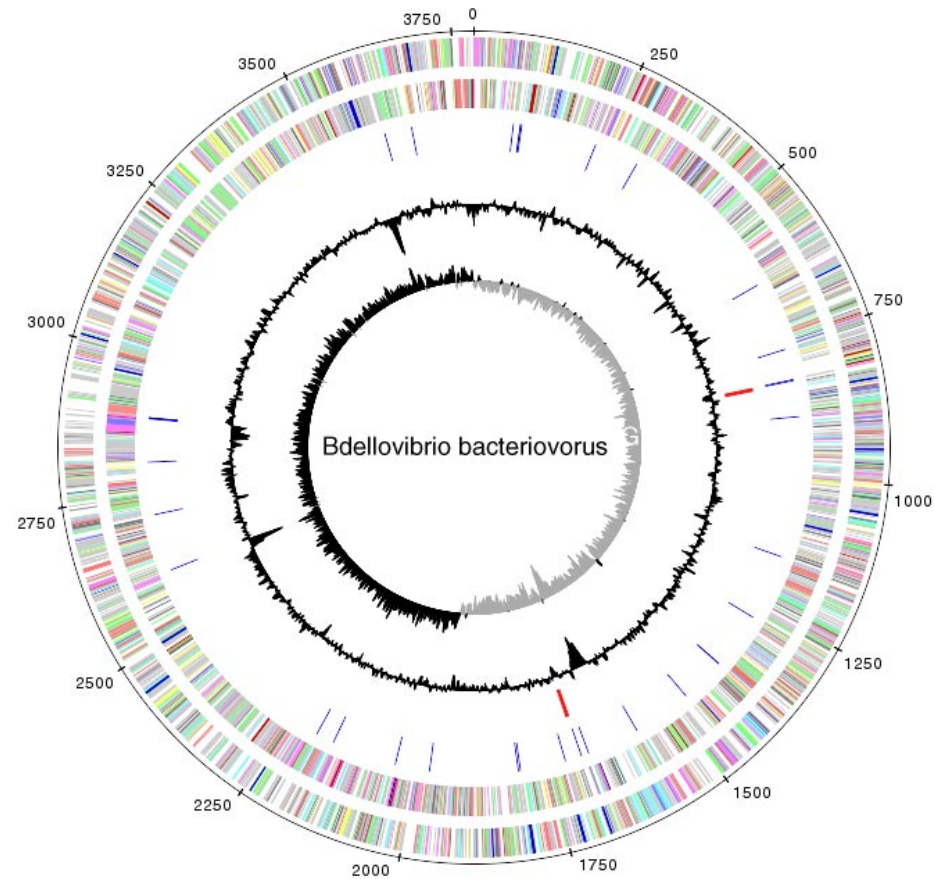
⇒ a hierarchy of more than 170 classes in GenDB-2.1

A typical GenDB Workflow



⇒ region creation & function assignment are modeled as annotations

An Example for an Annotated Genome

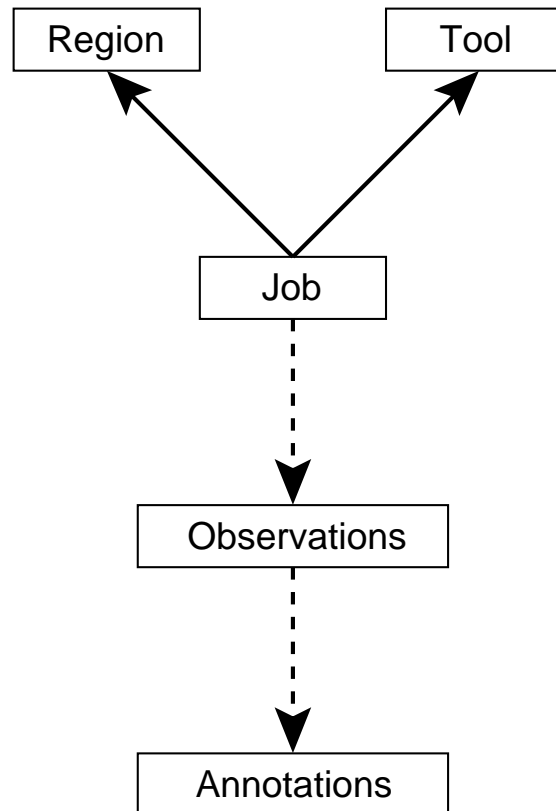


Rendulic *et al.*, *Science* 2004

Computational Requirements

- bacterial genomes range from 1 - 12 million basepairs
 - approx. 1000 genes per 1 million basepairs
 - 1-10 Tools used for gene prediction (per genome)
 - 10-100 Tools used for function prediction (per gene)
 - ~ 50 Observations / tool
 - 20 Tools for approx. 12.000 genes in *S. cellulosum* annotation
 - 240.000 Jobs are run → 12 million Observations
 - **not a one time effort**
- ⇒ extensive need for automation

The GenDB Tool and Job Concept



⇒ each Tool implements the same interface:
`run()` & `auto_annotate()`

A sample Tool: Blast (Altschul et al., NAR 1997)

- task: find similar sequences in large sequence databases
- result: alignments of query and hit sequences
- main work is implemented in `run()` method:
 - create FASTA file with sequence of given `Region`
 - run Blast and parse output using BioPerl
 - create an `Observation` for each hit result
- optional: automatic function assignment

N.B.: GenDB stores only minimal subset & recomputes the full alignment on demand

The GenDB Tool Configuration Wizard

Tool Configuration for GenDB_Azoarcus_sp

Blast configuration:

(Short) Tool name: Blast

Tool description: PSI-Blast vs. the SwissProt database

Select the blast application: psiblast

Blast database: /vol/biodb/fasta/sprot.fas

Index file: /vol/biodb/bioperl_index/sprot.fas

SRS databases: SWALL

File with pattern for PHI-Blast initialization:

Number of PSI-Blast iterations: 3

Command-line options:

Configure auto annotator:

Run auto annotator

Select annotation mode: Update latest annotation

Set threshold: 1e-50

Set regular expression:

Set region state: automatically annotated

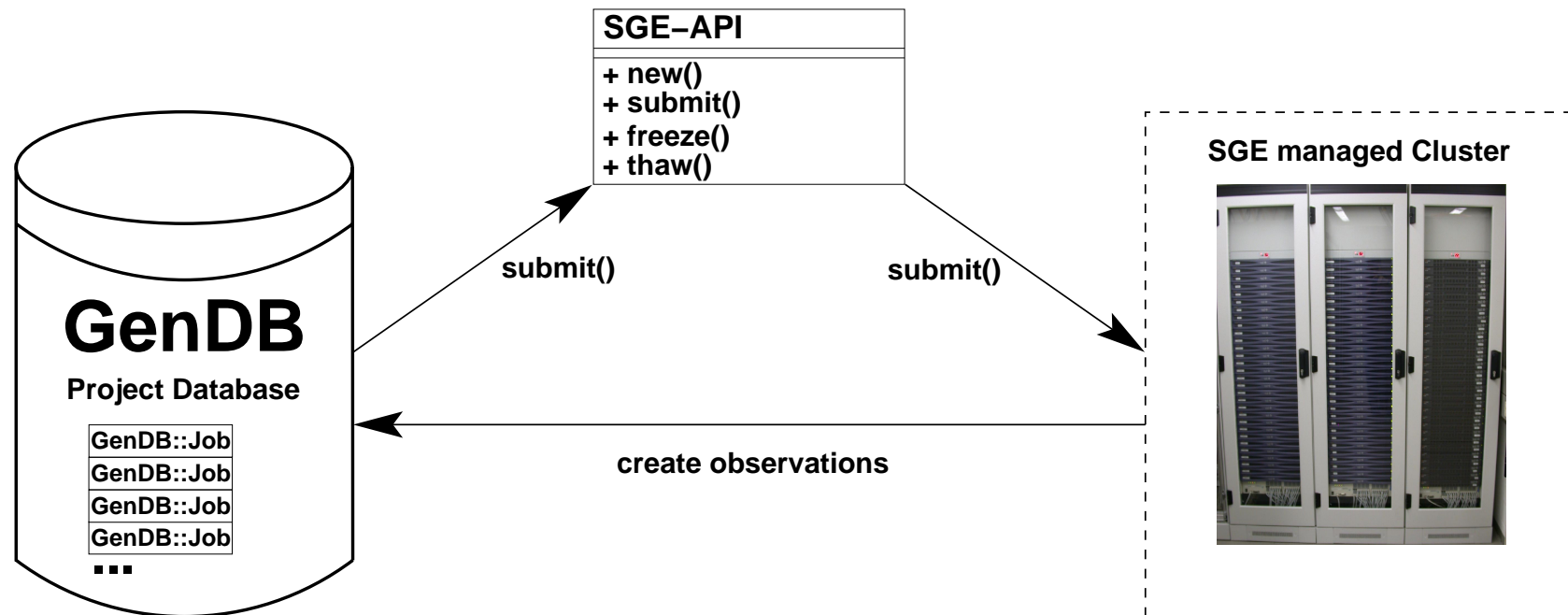
Back Next Cancel

⇒ instances of a Tool are configured and stored in a project

Definition of the GenDB Class "Job"

- Job = unique combination of a Tool and Region
- used to manage all Jobs within a GenDB project
- provides status information:
pending, submitted, running, finished, failed, cancelled
- stores date requested, submitted, finished
- can provide individual priorities
- single Jobs are submitted to SGE scheduler in `submit()` method via Perl SGE API

Integrating SGE into GenDB



The SGE API

- provides an easy to use, object-oriented interface to SGE
- each `Job` is computed via the `runtool.pl` Perl script
- tasks are scheduled by adding their commandline:

```
runtool.pl -p GenDB_Demo -j 12345 [-a]
```

- call method `submit()` to send tasks as jobs to SGE
- array jobs can be submitted by `freeze()` and `thaw()`
- supports email notification, syslog, arbitrary SGE options

Batch Submission of Jobs

- via Job-Submitter Wizard:
 - create Jobs for valid combinations of all Regions & Tools
 - submits Jobs to SGE
 - allows to restart submitted, failed, finished Jobs
- submission via Perl script or from GUI

Job Status Information

The left screenshot shows a summary of job counts for 'GenDB-B.-petrii Jobs':

pending:	0
submitted:	25026
running:	61
cancelled:	0
finished:	33740
failed:	0

The right screenshot shows a detailed table of job status information:

Job	Tool	Region	Date	Notify
Jobs				
pending				
submitted				
running				
running	Blast2n vs nt	Bor_new.chd_2671	Wed Jun 16 19:16:28 2004	
running	Pfam	Bor_new.chd_2671	Wed Jun 16 19:16:28 2004	
running	InterPro	Bor_new.chd_2671	Wed Jun 16 19:16:28 2004	
running	Blast2n vs nt	Bor_new.chd_2672	Wed Jun 16 19:16:28 2004	
running	Pfam	Bor_new.chd_2672	Wed Jun 16 19:16:28 2004	
running	InterPro	Bor_new.chd_2672	Wed Jun 16 19:16:28 2004	
running	Blast2n vs nt	Bor_new.chd_2673	Wed Jun 16 19:16:28 2004	
running	Pfam	Bor_new.chd_2673	Wed Jun 16 19:16:28 2004	
running	InterPro	Bor_new.chd_2673	Wed Jun 16 19:16:28 2004	
running	Blast2n vs nt	Bor_new.chd_2674	Wed Jun 16 19:16:28 2004	
running	Pfam	Bor_new.chd_2674	Wed Jun 16 19:16:28 2004	
running	Blast2n vs nt	Bor_new.chd_2675	Wed Jun 16 19:16:28 2004	
running	Blast2p vs nr	Bor_new.chd_2675	Wed Jun 16 19:16:28 2004	
running	PSI-Blast SP	Bor_new.chd_2675	Wed Jun 16 19:16:28 2004	
running	PSI-Blast COG	Bor_new.chd_2675	Wed Jun 16 19:16:28 2004	
running	Pfam	Bor_new.chd_2675	Wed Jun 16 19:16:28 2004	
running	TIGRFAM	Bor_new.chd_2675	Wed Jun 16 19:16:28 2004	
running	InterPro	Bor_new.chd_2675	Wed Jun 16 19:16:28 2004	
running	Blast2n vs nt	Bor_new.chd_2676	Wed Jun 16 19:16:28 2004	

⇒ failed Jobs can be resubmitted

GenDB Observations

Observations

Current Selection: CDS: Sce_765, 1516 Observations (14 visible)

Observation	Score	E-Value	Tool	DB	Start-Stop	Go	Description
942	0.0		Blast2Self	sce_new_4402	1 - 1413	-	dnaA chromosomal replication initiator protein (GenDB-ID=27618)
542.5	3.8e-160		Pfam	Bac_DnaA	402 - 1341	-	Bacterial dnaA protein
468.7	6.5e-139		TIGRFAM	DnaA	15 - 1404	-	DnaA: chromosomal replication initiat
368	1e-100		Blast2p vs nr	sprot DNAA_BACSU	58 - 1404	-	(P05648) Chromosomal replication initiator protein dnaA
368	1e-101		Blast2p vs KEGG	bsu:BG10065	58 - 1404	-	dnaA, dnaH, dnaJ, dnaK; chromosomal replication initiator protein
365	1e-100		PSI-Blast SPROT	sprot DNAA_BACSU	58 - 1404	-	(P05648) Chromosomal replication initiator protein dnaA
365	1e-100		PSI-Blast COG	BS_dnaA	58 - 1404	-	
332	1e-91		tBlastn Myxo	GMX_asmb_lid_580	85 - 1410	-	(closed molecule)
299	4e-82		Blast2p P.aer	AAG03391.1	400 - 1404	-	(dnaA) (chromosomal replication initiator protein DnaA)
264	7e-72		Blast2p Bdello	CAE77685.1	82 - 1404	-	(dnaA) (chromosomal replication initiator protein dnaA) hypothetical protein
			TMHMM				
	0.006088	-	InterPro	PR00051	199 - 213	3677, 3688, 5524, 6270, 6275	Bacterial chromosomal replication initiator protein, DnaA
	4.7e-137	-	InterPro	TIGR00362	5 - 468	3677, 3688, 5524, 6270, 6275	Bacterial chromosomal replication initiator protein, DnaA
		2e-13	Blast2n vs nt	embl AF071023	703 - 784	-	() Streptomyces reticuli DNA replication initiator protein (dnaA) gene,complete cds

Observation History Level_1 Level_2 Level_3 Level_4 Level_5 To Window Next Previous

⇒ Observations can be sorted by level, score, start, stop, ...

GenDB Tool Pipeline

1. sequence import
2. Region prediction: Glimmer, Critica, tRNA-scan, ...
3. automatic Region creation (Region Annotation)
4. function prediction: Blast, Pfam, InterPro, SignalP, ...
5. automatic function assignment (Function Annotation)
6. export, e.g. into EMBL or GFF file

Summary

- after sequence import the GenDB genome annotation pipeline can be run without further user interaction
- complete standard annotation of 3 Mb genome on:
 - 96 CPU cluster Sun Netra t1 & Netra X1
 - 8 CPU database server (SunFire 880)**IS possible within 24 hours**

Outlook

- GenDB-2.2 will feature incremental function annotation pipeline
- modified Tool concept for EMMA & ProDB
- update to Sun Grid Engine 6
- extension of cluster: 120 Dual Opteron
(later + 360 Dual Opteron)

Acknowledgements

Bioinformatics Resource Facility (Center for Biotechnology):

- Thomas Bekel
- Torsten Kasch
- Burkhard Linke
- Oliver Rupp
- Dr. Folker Meyer

Thank you for your attention!