

Bioinformatics and Grid Computing

Steffen Möller
University of Lübeck
Institute for Neuro- and Bioinformatics



KnowARC



Bioinformatics

- Formal representation of biological knowledge
- Maintenance of biological databases
- Mining of the data and the provisioning of databases of such derived patterns
 - Protein motifs
 - Transcription factor binding sites
- Simulations
 - Molecular dynamics
 - Ligand screening
 - Biochemical pathways



Peculiarities: I/O

- Databases need to be maintained
 - Regular updates (weekly or quarterly)
 - Coherent renewal of indices
- Varying sizes
 - 500k (restriction enzymes)
 - 15 MB (protein interactions)
 - 5GB (sequence families)
 - 10 GB (protein structures)
 - more for complete sequence information



Peculiarities: Variety

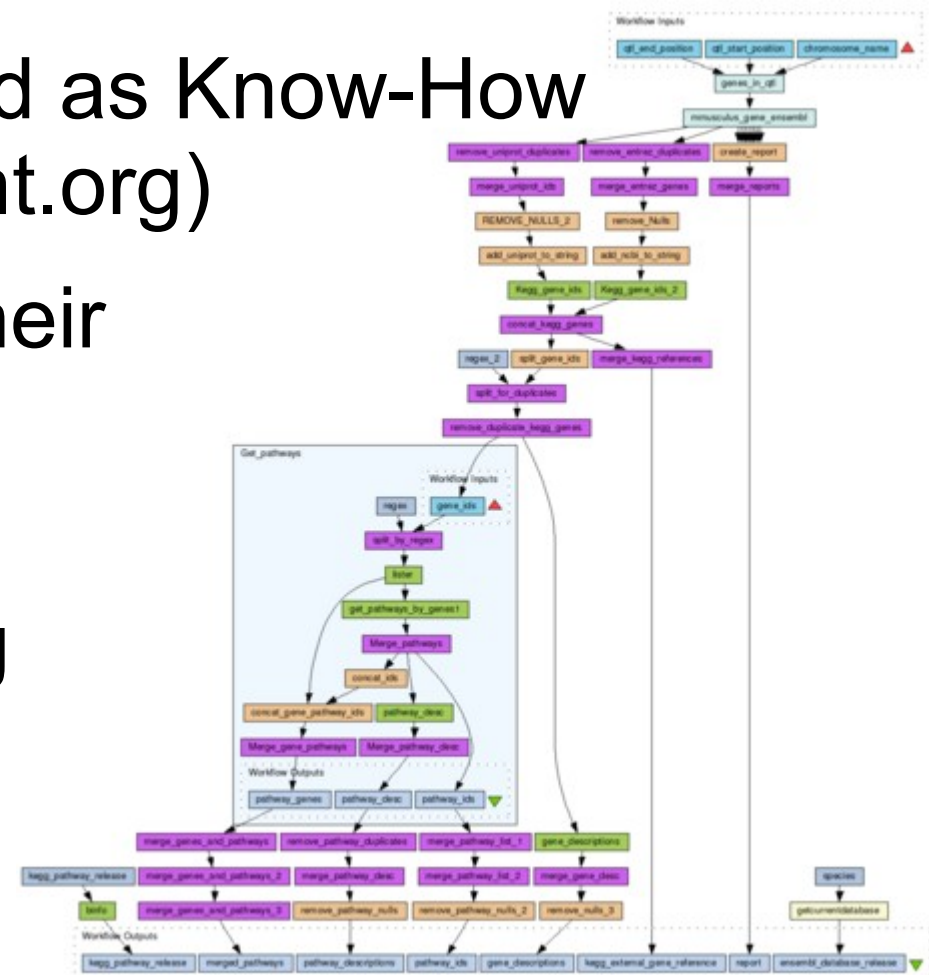
- Huge amount of different computational tools and databases
 - There are databases collecting information on the tools that have been provided
 - Nucleic Acids Research dedicates its annual January issues to describe only the most prominent databases
- Many tools are available only online
 - Often a service to the community to reduce maintenance costs
 - Bioinformatics is forerunner in acceptance of web services

Peculiarities: Complexity

- Higher-level (biological/clinical) bioinformatics
 - Uses *many* tools to address *single* research question
 - Which involves *many* data sources
 - Computational demand is *barely plannable* ahead
- Lower-level (service-oriented)
 - *Very concise* strategy
 - *Regular* schedule
- They all
 - May involve Monte Carlo simulations
 - Are commonly data-parallel

Grids in Bioinformatics

- Historically understood as interaction between web services
- Workflows are exchanged as Know-How (<http://www.myexperiment.org>)
- Research groups have their own medium to high-end research clusters
- Strong in peer computing

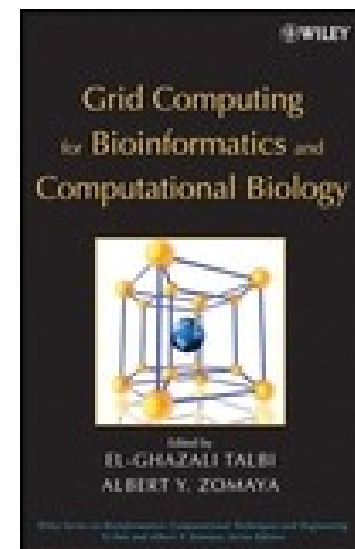


Workflow from myexperiment.org

Computational Grids in Bioinformatics

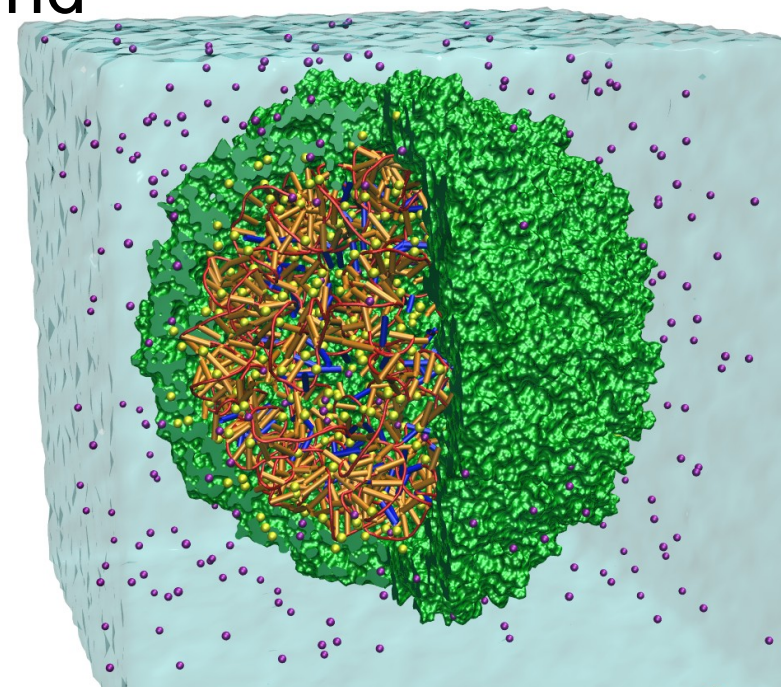
- Steadily gaining grounds
 - Molecular biologists started hiring IT staff
 - Friendly community, Campus Grids
 - Sharing of Know How and maintenance labour
 - Natural extension of (very familiar) local clusters
- But
 - Security sensitive when collaborating with big pharma

Grid Computing for Bioinformatics and Computational Biology
El-Ghazali Talbi (Editor), Albert Y. Zomaya (Editor)
ISBN: 978-0-471-78409-8



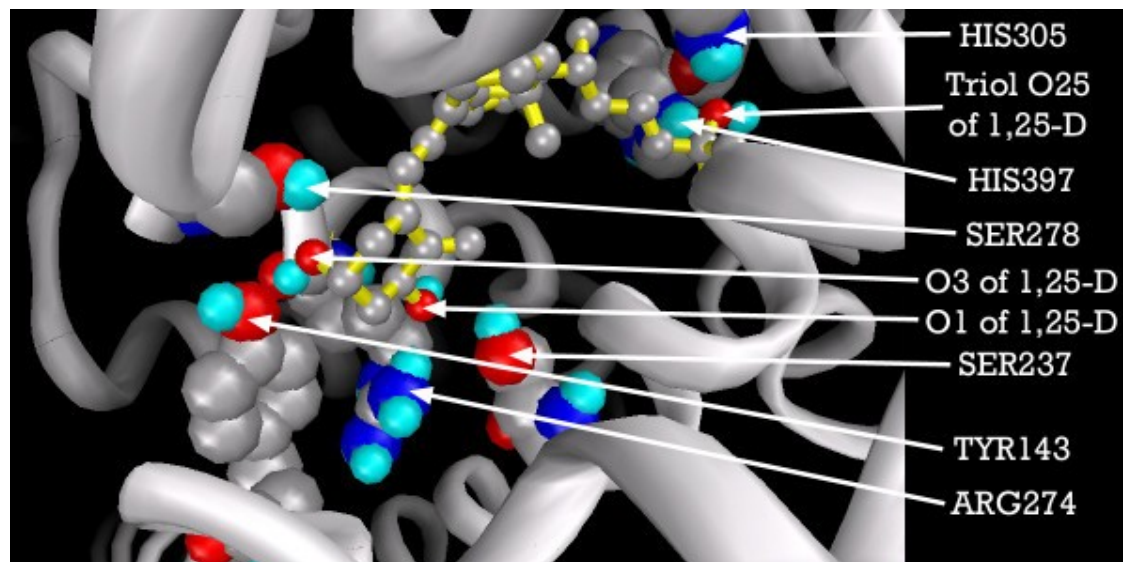
Typical users

- Molecular dynamics calculations
 - Take all CPU time that is available
 - No matter how much you have
 - With job lengths of 3 months being appreciated
 - You do not want them on your grid



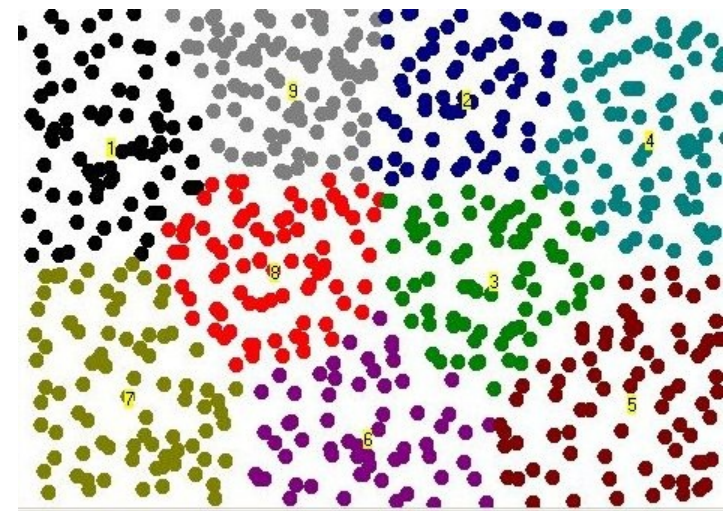
Typical users

- Ligand screening
 - Short job runtime for single compound
 - Millions of compounds to check independently
 - Possibly for many structures
 - Highly compatible with grid computing



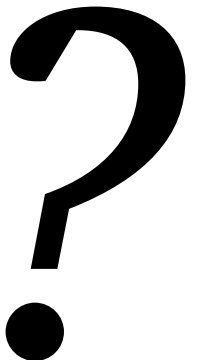
Typical users

- Clustering
 - Gene expression data
 - Gene/RNA/Protein sequence data
- Mining for motifs
 - Genetic backgrounds and disease association
 - Protein sequence and biochemical function
- Combinations of the above



Questions

- How can site maintainers compete with progress in the field
 - Site maintainers should share the burden
 - Grids shall be supported in becoming heterogeneous
- Is there a need for Grid Computing in Bioinformatics or should there just be local clusters
 - Inter-site collaborations – Expertise flocking (horizontally and vertically)
 - Burst-computing



Enhancing ARC grid middlewares

- Specification of Bioinformatics *runtime environments* (ARC provides such)
- Mechanism for their *dynamic installation* (developed 2007 for ARC with Torque)
- Allowing for shared environments
 - Enormous wealth of reliable and trusted applications in Debian Linux or RedHat/Fedora
 - Shared images can be extended dynamically with further runtime environments (prototype)
 - Needs virtualisation of compute clusters (prototype)

Combining Web services with Grid Computing

- Workflow environment Taverna is a prominent tool in bioinformatics
 - Interlinking of
 - WSDL-describe web services
 - Local applications
 - Visualisation
 - Aligned sequences
 - Protein structures
 - Fairly complete access to biological databases
- Taverna recently gained an interface to ARC

Combining Web services with Grid Computing

- Use case database
 - Templates of XRSL job descriptions
 - Name
 - Description
 - Command line
 - Runtime environments
 - Taverna specifies via connections in workflow
 - Input files
 - Output files



KnowARC

File Edit View History Bookmarks Tools Help

http://.inb.uni-luebeck.de:8180/knowarc-webservice-0.0.2/

Getting Started Latest Headlines Lernraum Virtuelle U...

usecaseid	
clustalw protein	Align Protein Secondary with
boxshade	Align protein
t_coffee	T-Coffee Multiple Sequence Align

This XML file does not
shown below.

```

<definitions name="
targetNamespace="
- <message name="
  <part name="cer
  </message>
- <message name="
  <part name="ses
  </message>
- <message name="
  <part name="Mu
  <part name="ses
  </message>
- <message name="
  <part name="gri
  </message>
- <message name="
  <part name="gri
  </message>
- <message name="
  <part name="Pos
  </message>
- <message name="
  <part name="mu

```

Taverna Workbench v1.7.0.0

File Tools Workflows Advanced

Design Results T2 Activity palette preview Taverna 2 preview

Search Watch loads

- Available Processors
 - Local Services
 - Biomart service @ http://www.biomart.org/biomart/martservice
 - Soaplab @ http://www.ebi.ac.uk/soaplab/emboss4/services/
 - Biomoby @ http://moby.ucalgary.ca/moby/MOBY-Central.pl
 - WSDL @ http://soap.genome.jp/KEGG.wsdl
 - WSDL @ http://eutils.ncbi.nlm.nih.gov/entrez/eutils/soap/eutils.wsdl
 - WSDL @ http://www.ebi.ac.uk/ws/services/urn:Dbfetch?wsdl
 - WSDL @ http://www.ebi.ac.uk/xembl/XEMBL.wsdl
 - WSDL @ http://soap.bind.ca/wsdl/bind.wsdl
 - WSDL @ http://pc02.inb.uni-luebeck.de:8180/knowarc-webservice-0.0.2/
 - porttype: JanitorUseCaseWebservicePortType [DOCUMENT]
 - login
 - boxshade
 - boxshadeResult
 - t_coffee
 - t_coffeeResult
 - hmmbuild
 - hmmbuildResult
 - hmmcalibrate
 - hmmcalibrateResult

Advanced model explorer

Workflow Object properties

Add Nested Workflow Offline

Workflow object	Retries	Delay	Backoff	Threads	Critical
Untitled workflow #1					
Workflow inputs					

Summary

- Bioinformatics as an icon for
 - Heterogeneity of research aims in the community
 - Variety in applications used
 - Complexity of workflows
- Ongoing development of and with ARC
 - Dynamic runtime environments
 - Workflow engine Taverna
 - submits jobs to the grid and retrieves results
 - back and forth with regular applications / web services
 - Virtualisation to add computers and be dynamic

Thoughts

- Technology will go hand in hand with the community
 - Official or gentleman's agreements between sites?
 - Accounting and monetary compensation?
- How to accommodate smaller research groups
 - Tit-for-Tat VO is fantastic
 - lower interest because of I/O overhead and the monitor becomes too crowded
 - consequence of specialisation and heterogeneity

Thoughts

- Virtualisation
 - Allows for dynamic deployments
 - Increases reliability, particularly for smaller sites
 - Increases availability, particularly for smaller sites with exceptional applications
 - At the brink to peer computing
 - Windows machines may execute Linux binaries
 - Clients actively seek computational load

Acknowledgements

- Daniel Bayer, Hajo Krabbenhöft in Lübeck
- All the friendly NorduGridders for their collaboration or feedback
 - Copenhagen
 - Lund
 - Helsinki
 - Oslo
 - Kosice

