Towards dynamic database infrastructures for mouse genetics

software methods, models and tools

Morris Swertz (m.a.swertz@rug.nl)
Athens, October 10, 2008
BIBE 7.2.4

university medical center groningen
department of medical genetics

university of groningen
groningen bioinformatics center

...and partners
Outline

› Pilot 1: extensible genotype and phenotype db (xGaP)
  • Systematic extensible data model and
  • MOLGENIS model-to-software generator
  • To produce family of de novo databases

› Pilot 2: integrative workflow
  • Wrap existing dbs with MOLGENIS and BioMart
  • So we can use web services as lingua franca
  • To automate integration of heterogeneous data

› Conclusions
Challenge

Complex and large scale phenotype/genotype experiments

Main work flow
Data dependency
Biomaterial/result
Lab/analysis process
Scale of information
Associated data files

strains

inbreed

markers

individuals

genotype

hybridize

expressions

microarrays

correlate

network

preprocess

norm exprs.

map

QTL profiles

genotypes

genotype

100

10,000

100,000

1,000,000

10.000

10.000
Need ‘cyber’ infrastructure

“Towards a cyber infrastructure for the biological sciences”
Two CASIMIR pilots

“Towards a cyber infrastructure for the biological sciences”
Pilot 1: Results

- Extensible database for genotype and phenotype experiments; http://www.xgap.org

Swertz et al, CASIMIR consortium, GEN2PHEN consortium (submitted)
Pilot 1: Results

- Extensible database for genotype and phenotype experiments;
  
  [http://www.xgap.org](http://www.xgap.org)

Swertz et al, CASIMIR consortium, GEN2PHEN consortium *(submitted)*
Pilot 1: Results

- Common core model based on FuGE,
- Systematic extension points

Swertz et al, CASIMIR consortium, GEN2PHEN consortium (submitted)
Pilot 1: Results

- R services
- Very useful for plugins...

```r
source("http://localhost:8080/xgap/api/R")
MOLGENIS is connected

#download data
traits <- get.metabolitedata(name="mytraits")
25 metabolite downloaded in 30ms

genotypes <- get.markerdata(name="mygenotypes")
744 marker downloaded in 30ms

#calculate ...

#upload results for others to use
add.data(qtls, name="myqtls")
18,600 data items added in 2sec
```

Swertz et al, CASIMIR consortium, GEN2PHEN consortium \textbf{(submitted)}
Pilot 1: Results

- Web services
- See also pilot 2.

Swertz et al, CASIMIR consortium, GEN2PHEN consortium *(submitted)*
Issues with traditional development

1. Time intensive
2. Duplicated efforts, reinventing wheels
3. Incompatible syntax, structures and semantics

http://www.molgenis.org
Alternative strategy

Beyond standardization:

Dynamic software infrastructures for systems biology

Morris A. Swertz and Ritsert C. Jansen

Abstract | Progress in systems biology is stalled by a lack of suitable software infrastructures. Biological information connects to work that is done in other labs, and this knowledge is helpful. However, the infrastructure needed to make this work is currently lacking. We argue that a software tool called a generator, can be used to automate the creation of software infrastructures that systems biologists can easily extend.

http://www.molgenis.org

Standard software models

Software factory

Blueprint model

bioinformatician

softw engineer

biologist

http://www.molgenis.org

Standard software models

New Biology

Software factory

http://www.molgenis.org
1 Standard software models

New Informatics

Software factory

http://www.molgenis.org
Biological extensions to those models

http://www.molgenis.org
2 Biological extensions to those models

MOLGENIS configuration file:

```
<entity name="Data">
  <field name="Assay" type="xref" xref_field="Assay.Id" />
  <field name="Row" type="xref" xref_field="Row.Id" />
  <field name="Col" type="xref" xref_field="Col.Id" />
  <field name="Value" type="decimal" />
  <unique fields="Assay,Col,Row" />
</entity>
```

http://www.molgenis.org

2 Biological extensions

MOLGENIS configuration file:

```xml
<molgenis name="gg">
    <entity name="Identifiable" abstract="true">
        <field name="Id" type="autoid" />
        <field name="Name" type="varchar" />
    </entity>
    <entity name="Experiment" implements="Identifiable">
        <unique fields="Name" />
    </entity>
    <entity name="Trait" implements="Identifiable">
        <field name="Experiment" type="xref" xref_field="Experiment.Id" />
        <unique fields="Name,Experiment" />
    </entity>
    <entity name="Subject" implements="Identifiable">
        <field name="Species" type="varchar"/> 
        <field name="Experiment" type="xref" xref_field="Experiment.Id" />
        <unique fields="Name,Experiment" />
    </entity>
    <entity name="Assay" implements="Identifiable">
        <field name="Experiment" type="xref" xref_field="Experiment.Id" />
        <unique fields="Name,Experiment" />
    </entity>
    <entity name="Data">
        <field name="Assay" type="xref" xref_field="Assay.Id" />
        <field name="Col" type="xref" xref_field="Trait.Id" />
        <field name="Row" type="xref" xref_field="Subject.Id" />
        <field name="Value" type="decimal" />
        <unique fields="Assay,Col,Row" />
    </entity>
</molgenis>
```
Dynamic software applications
Works in Eclipse or via batch program

Download and customize...

Model file

Generate

FormGen
MenuGen
MatrixGen
JTypeGen
JListGen
JDatabaseGen
MySQLGen
TreeGen
PluginGen
JDBCMapGen
JReadCsvGen
JReadCsvGen
JListGen
RLjstGen
JDBogen
RMMatrixGen
HSQLGen
WSGen

MyScript

Plugins

APIs in Java, R, Web services and HTTP

user interaction infrastructure

Communication infrastructure

Data infrastructure

http://www.molgenis.org
Dynamic software applications
Works in Eclipse or via batch program

Each Generator is a Template:

```java
public class ${entity.Name}Mapper
    extends JDBC Mapper<${entity.Name}>
{
    public String addSql(${entity.Name} e)
    {
        return String.format(
            "insert into ${entity.Name} ( ",
            +""#csv($entity.Fields, "$i.Name")"
            +") values (",
            +""#csv($entity.Fields, "%s")"
            + "",
            +""#csv($entity.Fields, e.get${i.Name}())"
        );
    }
} ...
```

http://www.molgenis.org

Federation of dynamic software applications
Reuse and integration is eased by the standard generated interfaces

http://www.molgenis.org
Towards a catalogue of models

Genotype-to-Phenotype (http://www.xgap.org)

Proteomics (mass spec)

Animal observatory

HT sequencing

Biobanking and LSDBs

Metabolomics

MOLGENIS biological software family

Plus international standardization initiatives

Plus complementary model-based biosoftware

FuGE, ISA-TAB, MAGE-TAB, OBO, caBIO?

http://www.molgenis.org


Taverna, BioMOBY? Galayx? for processing
“Towards a cyber infrastructure for the biological sciences”

Two CASIMIR pilots

user interaction infrastructure

data infrastructure

processing infrastructure

researchers

Pilot 2

Communication + integration infrastructure
Pilot 2: Integration of existing data

› Doesn’t scale to genome-wide by hand (word, excel)

Pilot 2: Automation is hard

› Non-existent or incompatible interfaces....

Pilot 2: Strategy and tools

› “wrap” sources to have web services (ws)

MOLGENIS wrapping an existing data

- MPD in original structure (based on download files)

BioMART wrapping of existing data

- Ensembl Mouse transformed to warehouse structure
- [http://www.biomart.org](http://www.biomart.org)

Pilot 2: Integrative workflow

Taverna workflow

- uses
- select data (once per query)

MOLGENIS

- uses
- copy all data (once per update)

Web Services

- KEGG
- Kyoto University, Japan

MPD

- MPD mirror
- transform all data (once per update)

BioMart

- Ensembl
- Jackson Laboratories, USA

- Sanger Institute, UK

- University of Groningen, Netherlands

[!] Clearly, common data elements and vocabularies would greatly improve the situation

Summary

› Pilot 1: extensible genotype and phenotype db (xGaP)
  • Systematic extensible data model and
  • MOLGENIS model-to-software generator
  • To produce family of *de novo* databases

› Pilot 2: integrative workflow using Taverna
  • Wrap existing dbs with MOLGENIS and BioMart
  • So we can use web services as lingua franca
  • To automate integration of heterogeneous data

› Towards a catalogue of software models
› Towards common data elements/ontologies
Thank you

m.a.swertz@rug.nl
Morris A. Swertz
Rudi Alberts
Damian Smedley
Katy Wolstencroft
Andrew R. Jones
Bruno M. Tesson
Richard A. Scheltema
Gonzalo Vera

Paul Schofield
Klaus Schughart
John M. Hancock
Helen E. Parkinson
Carole Goble
Engbert O. de Brock
Cisca Wijmenga
Ritsert C. Jansen
the GEN2PHEN consortium
the CASIMIR consortium