GEPETTO : OPEN-SOURCE FRAMEWORK FOR GENE PRIORITIZATION

Laboratoire de Bioinformatique et de Génomique Intégratives (LBGI)
Illkirch, France

Team leader : Olivier POCH
Project Manager : Hoan NGUYEN
Presented by : Vincent WALTER

BOSC 2013 – 20 February 2013
**BIG DATA CONTEXT**

**VOLUME**
Data at rest
Terabytes to exabytes of existing data to process

**VELOCITY**
Data in motion
Streaming data, queries take milliseconds to seconds to respond

**VARIETY**
Data in many forms
Structured, unstructured, text, multimedia

**VERACITY**
Data in doubt
Uncertainly due to data inconsistency and incompleteness, ambiguities, latency, model approximations
SM2PH – BIG DATA MANAGEMENT

SM2PH (Structural Mutation to Pathology Phenotypes in Human) [Link](http://decrypthon.igbmc.fr/sm2ph/)

Data warehouse
- Data driver
  - Gene-centric
  - All human proteins (≈20,000)
  - All human SNPs (≈400,000)

Combines multi-data sources
- Information associated with any protein
- External databases: UniProt, String-DB, etc.
- Lab databases: GxDb, Evolucode, etc.

Open-data
- Downloadable on SM2PH website
- Provides data in XML format (universal)


PRIORITIZATION PROCESS

- Identification of the most promising feature associated to a question (ex. biological process, pathology, network) in dynamic systems

- **3 STEPS PROCESS**
  - Building model for training features
  - **Local** prioritizations
    - Candidate feature profile
    - **Evaluation** according to the similarity with the model (= scoring)
    - Ranking
  - **Global** prioritization

**GEPETTO** = dedicated to human gene prioritization

---

**WHY AN OPEN-SOURCE FRAMEWORK?**

**SEVERAL GENE PRIORITIZATION TOOLS**
- New ones published regularly
- Most of them are web GUI

**BUT MANY DRAWBACKS**
- **Maintenance** – short period
- **Availability** - couple of months

- **Queryability** – not for high-throughput process
- **Integrability**– no libraries for gene prioritization

- **Extensibility** – not open-source
- **Data** – key aspects are not integrated (genomic context, 3D-structure,...)
**GEPETTO FRAMEWORK : SPECIFICATIONS**

**OPEN-SOURCE**
- **Supported** by the community
- **Improved** by the community
- **Completed** by the community
  - Adding your own datasources

**ACCESSIBILITY**
- **Biologists and clinicians**
  - Web interface
- **Bioinformaticians**
  - Standalone application (command line), Java API

**STANDARDIZATION / FLEXIBILITY**
- Facilitated maintenance
- **Modularity** - 1 JAR project by local prioritization
- **Scalability / Customizability**

**VERSIONING / CONTRIBUTION**
- **SVN** (via SourceForge)
- GitHub and Mercurial (in progress)
**GEPETTO FRAMEWORK : TECHNOLOGIES**

**PROGRAMMING LANGUAGES**
- **Java**: modules libraries / core of the application
- **Python**: command line launcher
- **R**: statistical tests for local prioritization

**PLATFORMS**
- Only **UNIX** systems are supported today
  - Debian-based distributions (Ubuntu,...)
    - Débian package to install easily
  - Redhat-based (Fedora,...)
    - Source tarball (tar.gz archive)

**INTEROPERABILITY**
- **Galaxy integration**
  - I/O: Use file as input and can generate file as output
  - Queryable by command line
LOCAL PRIORITIZATION MODULES

GENERIC MODULES

- **Protein sequence** (Alignment)
  - E-value of the best hit with BLASTp

- **Evolutionary barcodes** (Evolucode)
  - Evolutionary histories in vertebrates (16) for all human genes
  - Based on K-means clustering

- **Genomic context** (Geco)
  - Nested repeats, open chromatin, PolII,...

- **Transcriptomic** (GxDB)
  - Tissular expression in 79 human tissues
  - Pearson Correlation and Fisher's omnibus analysis

- **Protein-protein interactions** (String)
  - Physical / functional interactions
  - Size of the overlap with the training model

SPECIFIC MODULES

- **Hereditary disease gene probability** (IDGP)
  - Combination of the recessive / dominant probability of involvement

- **Genome-wide association study** (GWAS-AMD)
  - DoubleGC value

**Global Prioritization**

- Global scoring and ranking
  - Data fusion
  - Based on the different local ranking

**Order Statistics**

- Able to handle features with missing values.
- Minimizes the bias for known or well-characterized features

\[ Q(r_{1}, r_{2}, ..., r_{N}) = \int_{0}^{r_{1}} \cdots \int_{0}^{r_{2}} \cdots \int_{0}^{r_{N}} ds_{1} \cdots ds_{N-1} \cdots ds_{1} \]

**Robust Rank Aggregation**

- Similar to order statistics
- More significant result for global prioritization

\[ P(r_{k} \leq r_{k}') = \sum_{i=k}^{n} \binom{n}{i} \cdot (r_{k}')^{i} \cdot (1-r_{k}')^{n-i} \]

**Mallows Model**

- Looks for a consensus ranking

\[ \pi_{1}, ..., \pi_{n} = \frac{1}{n} \cdot \sum_{k=1}^{n} \sum_{i < \pi_{k} \mid j} 1 \]

**GPSy**

- Assigns weights to the different criteria to find the optimal weight combination

\[ r_{ij} = \frac{\sum_{j=1}^{n} w_{ij} \cdot r_{ij}}{\sum_{j=1}^{n} w_{ij}} \]
**GEPETTO WORKFLOW USING JBPM**

- A division of an American multinational software company: RedHat
- Specialized in writing and supporting open-source middleware software
- Provides tools for Java applications

- Open-source workflow engine
- Since 2003 (Don't reinvent the wheel)
- Complete workflow engine
  - Manages information flows
  - Makes the bridge between biologists (business analysts and end users) and bioinformaticians (developers).
- Provides features not provided by other workflow engines
- Based on BPMN2 standard notation (Business Process Model and Notation)
- Describes the process step by step
## Advantages of JBPM

| Field usage | • Universal / Not oriented  
| • More flexible |
| Installation | • Import .JAR libraries  
| • Very easy |
| Skills required | • No computer programming competences required |
| Workflow – possibilities | • Complex  
| • Parallel gateway or exclusive gateway (decisionnal workflow) |
| Workflow – definition | • BPMN2 (Business Process Management Notation) – standard notation  
| • XML |
| Workflow – graphical editor | • Web Service for JbossAS (Application Server)  
| • Web Service for Apache Tomcat |
| Workflow – passing data mode between nodes | • Java / Object (directly in memory)  
| • Not necessary to handle many files |
| Workflow – languages or programs | • All languages / programs wrapped into WorkItemHandler |
**JBPM : EXAMPLE OF SPECIFIC FEATURES**

**PARALLEL GATEWAY**
- Used to split / synchronize the respectively incoming or outgoing sequence flow
- Executes task in parallel

**GEPETTO : Used for local prioritization**

**EXCLUSIVE GATEWAY**
- Guides the workflow progression using any part of decision (decisional approach)

**GEPETTO : Used to select the global prioritization**
GEPETTO WORKFLOW USING JBPM

STEP 1

LOCAL PRIORITIZATION

STEP 2

STEP 3

GLOBAL PRIORITIZATION

LOADER
ADD NEW LOCAL PRIORITIZATION MODULE

1. Download GEPETTOModel.jar
2. Create new JAR project
3. Implements class that inherits GEPETTOModel interfaces using Features provided by GEPETTOModel library.
4. Create/Edit MANIFEST.mf file
5. Update core to register module
ADD NEW LOCAL SCORING METHOD

- Local scoring method are managed using Polymorphism
- To add a new scoring method
  - Implement a new class which inherit from GenePrioritization or ProteinPrioritization
  - Add this class to the ModulePrioritizationFactory
  - Change the method used in Spring Application Context File (XML file)

- Spring Framework
  - Open-source Java application framework
  - Inversion Of Control (IOC) container
  - Facilitates development and test of Java (Web) Applications
  - Standardization of source code : more robust and easier to maintain
  - Reduce code of SM2PH-DAO(Data Access Object)
  - GEPETTO uses BeanFactory/Spring IOC (Dependency Injection)
APPLICATIOns – AMD USE CASE

- **AMD** (Age-related Macular Degeneration)
  - Macula degeneration
  - Affects *old people* (over 50 years)
  - Causes a significant *weakening* of the *visual capabilities*

- Confidential GWAS-data
  - Provided by the AMD Gene Consortium
    - T. Léveillard, et al.
    - 8,000 patients
    - 8,000 controls

- Data sets
  - Genes involved in AMD (12 SNPs, 14 genes in KEGG, Pubmed)
  - Recently validated new genes (7 SNPs, 9 genes)
  - Candidate genes out of the limit of GWAS detection (21 SNPs, 29 genes)

APPLICATIONS – AMD USE CASE

APPLICATION 1

- Evaluating the ability to successfully detect AMD validated genes

- Comparison against Endeavour / ToppGene
  - Method: ROC-AUC
  - Test set: AMD known and new genes (23) + Genes closed to GWAS limits (29)

<table>
<thead>
<tr>
<th>TRAINING SET</th>
<th>AUC (prioritization of the 23 AMD genes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GEPETTO</td>
</tr>
<tr>
<td>Known genes (14)</td>
<td>0.649</td>
</tr>
<tr>
<td>New genes (9)</td>
<td>0.825</td>
</tr>
<tr>
<td>Mixed genes (11)</td>
<td>0.887</td>
</tr>
</tbody>
</table>

- Conclusion
  - $0.64 < \text{AUC} < 0.89$: GEPETTO uses a good model.
  - Able to detect genes that are experimentally validated (over the limit of GWAS detection)
APPLICATIONS – AMD USE CASE

APPLICATION 2

- Evaluate the ability to **discriminate** AMD target genes and Retinitis Pigmentosa (RP) known genes

- Compared to Endeavour
  - **Method**: ROC-AUC
  - **Training set**: AMD known genes
  - **Test set**: All AMD candidates + RP known genes

<table>
<thead>
<tr>
<th>DISCRIMINATION CRITERIA</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP = AMD known genes ; FP = Other genes (AMD/RP)</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
</tr>
<tr>
<td>TP = AMD validated genes ; FP = Other genes (AMD/RP)</td>
<td>0.859</td>
</tr>
<tr>
<td></td>
<td>0.892</td>
</tr>
<tr>
<td>TP = AMD candidates genes ; FP = RP genes</td>
<td>0.752</td>
</tr>
<tr>
<td></td>
<td>0.564</td>
</tr>
</tbody>
</table>

- **Conclusion**
  - **0.75 < AUC < 0.97**: GEPETTO uses a good model.
  - Very discriminant in some cases
# Perspectives

| Optimisation                                                                 | • Implementation of the 3 new methods of global ranking  
| • Implementation of new methods of local ranking |
| New parameters                                                               | • Using SNP integration from MSV3d *(MisSense Variants mapped to 3D-structures)* / Bayesian networks |
| Extends to other organisms                                                   | • Human-centric gene prioritization  
| • Add the possibility to prioritize genes for other species |
| Pattern extraction                                                           | • Extraction of patterns for hereditary disease causing genes |
| Application                                                                  | • Development of modules devoted to ciliopathies or other rare diseases |
ACKNOWLEDGEMENT

Olivier POCH
Julie THOMPSON
Hoan NGUYEN
Odile LECOMPTE
Luc MOULINIER
Jean MULLER
Laetitia POIDEVIN
Raymond RIPP
Wolfgang RAFFELSBERGER
Alexis ALLOT
Kirsley CHENNEN
Etienne GOFFINET
Alan LAHURE
Can MOREL
Anaïs NICOL
Hélène POLVECHE
Raphaël SCHNEIDER
Carlos BERMJO-DAS-NEVES