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*

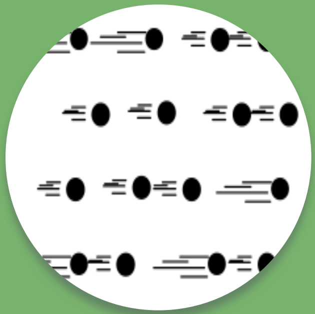
BOSC 2013 – 20 February 2013

BIG DATA CONTEXT

VOLUME

Data at rest

Terabytes to exabytes of existing data to process



VELOCITY

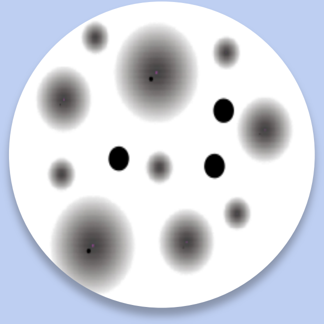
Data in motion

Streaming data, queries take milliseconds to seconds to respond

VARIETY

Data in many forms

Structured, unstructured, text, multimedia



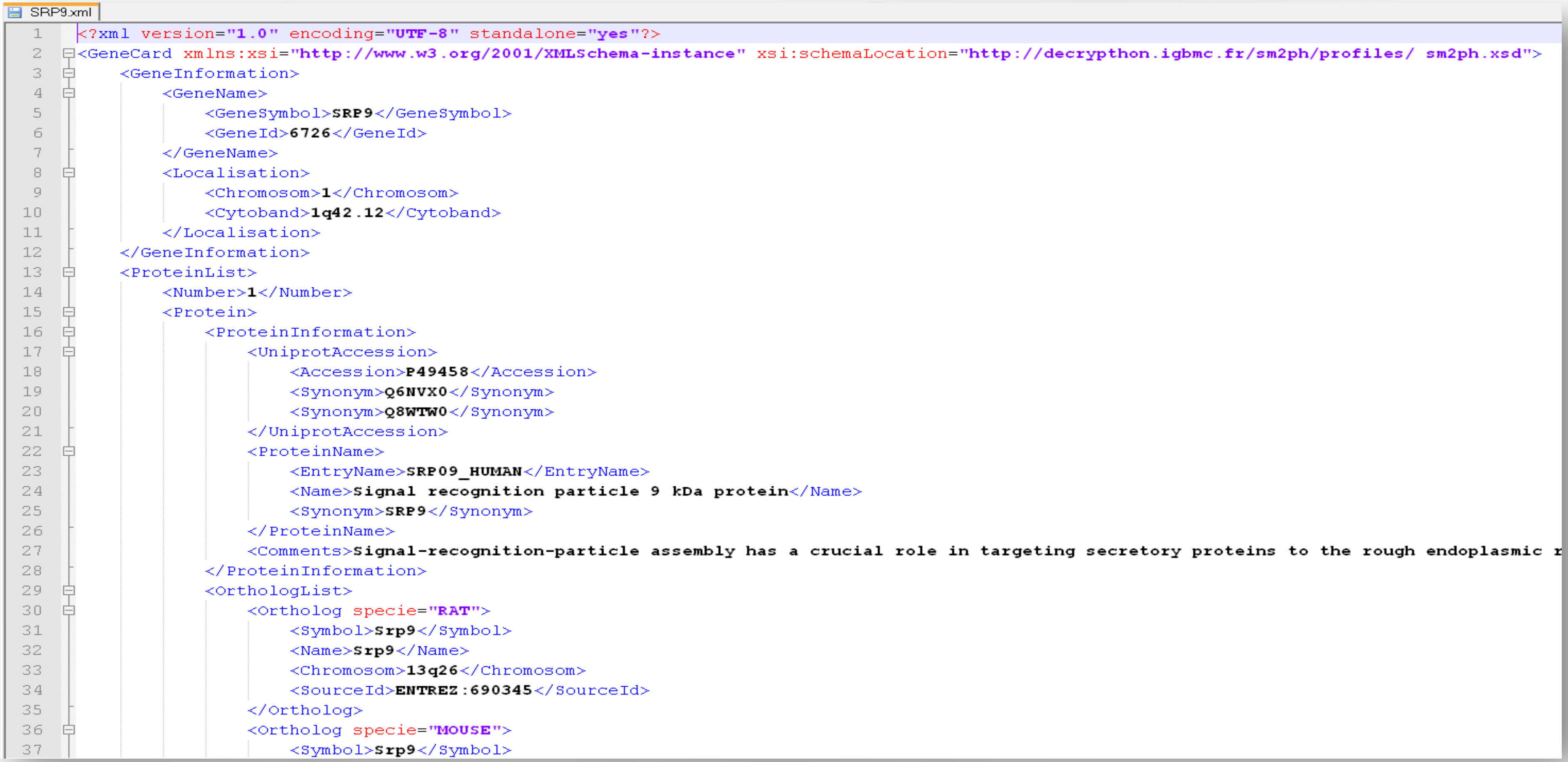
VERACITY

Data in doubt

Uncertainly due to data inconsistency and incompleteness, ambiguities, latency, model approximations

SM2PH – BIG DATA MANAGEMENT

SM2PH (**S**tructural **M**utation to **P**athology **P**henotypes in **H**uman) **http://decrypthon.igbmc.fr/sm2ph/**



```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<GeneCard xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://decrypthon.igbmc.fr/sm2ph/profiles/ sm2ph.xsd">
    <GeneInformation>
        <GeneName>
            <GeneSymbol>SRP9</GeneSymbol>
            <GeneId>6726</GeneId>
        </GeneName>
        <Localisation>
            <Chromosom>1</Chromosom>
            <Cytoband>1q42.12</Cytoband>
        </Localisation>
    </GeneInformation>
    <ProteinList>
        <Number>1</Number>
        <Protein>
            <ProteinInformation>
                <UniprotAccession>
                    <Accession>P49458</Accession>
                    <Synonym>Q6NVX0</Synonym>
                    <Synonym>Q8WTW0</Synonym>
                </UniprotAccession>
                <ProteinName>
                    <EntryName>SRP09_HUMAN</EntryName>
                    <Name>Signal recognition particle 9 kDa protein</Name>
                    <Synonym>SRP9</Synonym>
                </ProteinName>
                <Comments>Signal-recognition-particle assembly has a crucial role in targeting secretory proteins to the rough endoplasmic r
            </ProteinInformation>
            <OrthologList>
                <Ortholog specie="RAT">
                    <Symbol>Srp9</Symbol>
                    <Name>Srp9</Name>
                    <Chromosom>13q26</Chromosom>
                    <SourceId>ENTREZ:690345</SourceId>
                </Ortholog>
                <Ortholog specie="MOUSE">
                    <Symbol>Srp9</Symbol>
```

PRIORITIZATION PROCESS

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

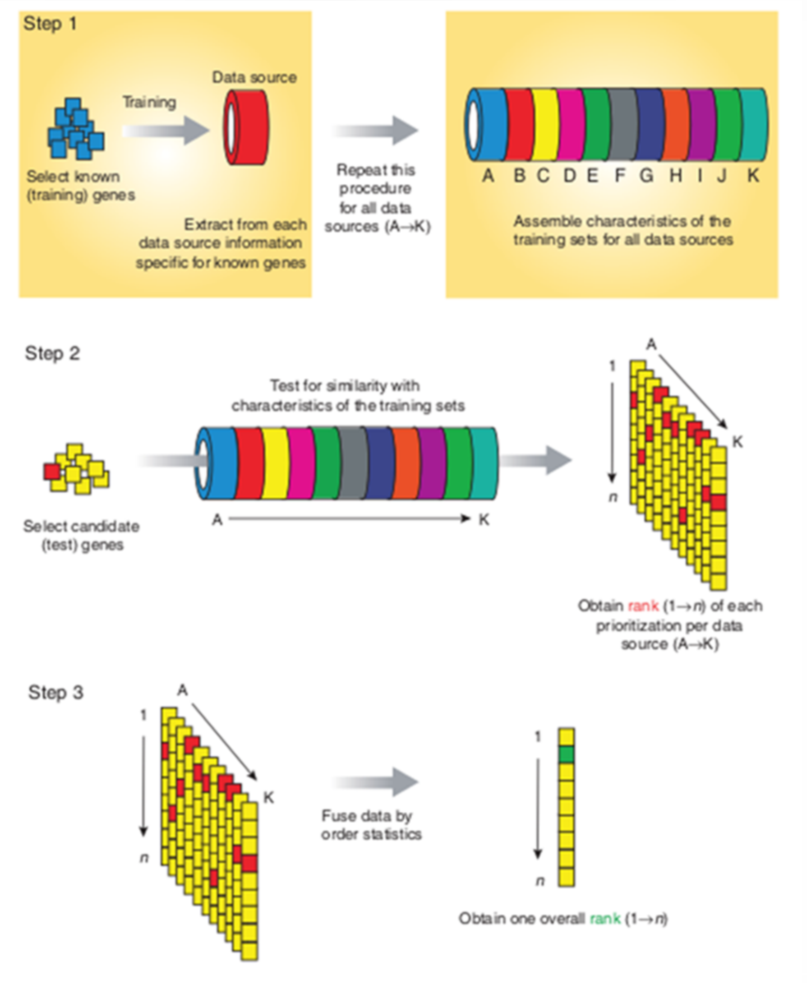
- **3 STEPS PROCESS**

  - Building model for training features

  - **Local** prioritizations
    - Candidate feature profile
    - **Evaluation** according to the similarity with the model (= scoring)
    - **Ranking**

  - **Global** prioritization

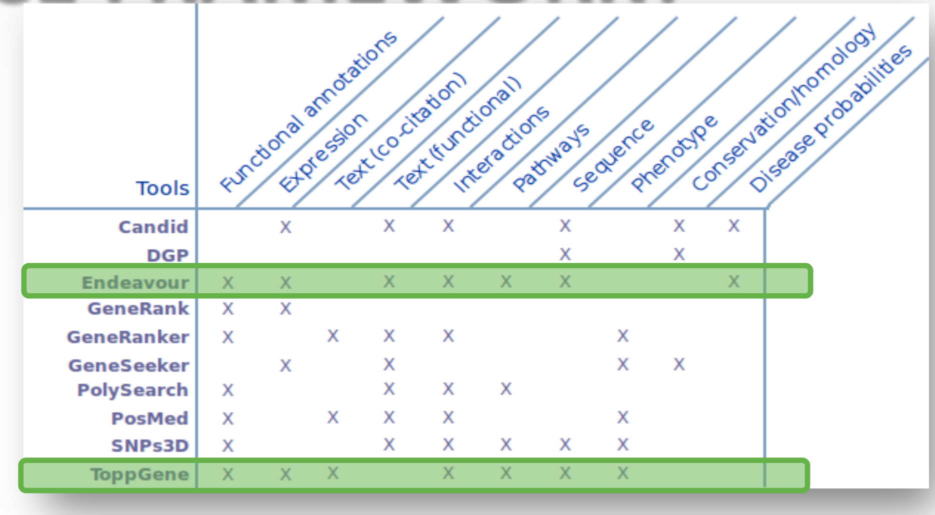
GEPETTO = dedicated to human gene prioritization



Step 1
Training
Data source
Select known (training) genes
Extract from each data source information specific for known genes
Repeat this procedure for all data sources (A→K)
A B C D E F G H I J K
Assemble characteristics of the training sets for all data sources
Step 2
Test for similarity with characteristics of the training sets
Select candidate (test) genes
A
K
1
n
Obtain rank (1→n) of each prioritization per data source (A→K)
Step 3
A
K
1
n
Fuse data by order statistics
1
n
Obtain one overall rank (1→n)

Aerts, S., et al. ***Gene prioritization through genomic data fusion.*** *Nature Biotechnology 24, 537–544.*

WHY AN OPEN-SOURCE FRAMEWORK?



Tools	Functional annotations	Expression	Text (co-citation)	Text (functional)	Interactions	Pathways	Sequence	Phenotype	Conservation/homology	Disease probabilities
Candid		X		X	X		X		X	X
DGP							X		X	
Endeavour	X	X		X	X	X	X			X
GeneRank	X	X								
GeneRanker	X		X	X	X			X		
GeneSeeker		X		X				X	X	
PolySearch	X			X	X	X				
PosMed	X		X	X	X			X		
SNPs3D	X			X	X	X	X	X		
ToppGene	X	X	X		X	X	X	X		

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

- Available on SourceForge platform http://sourceforge.net/projects/gepetto/
- Supported by the community
- Improved by the community
- Completed by the community
  - Adding your own datasources

ACCESSIBILITY

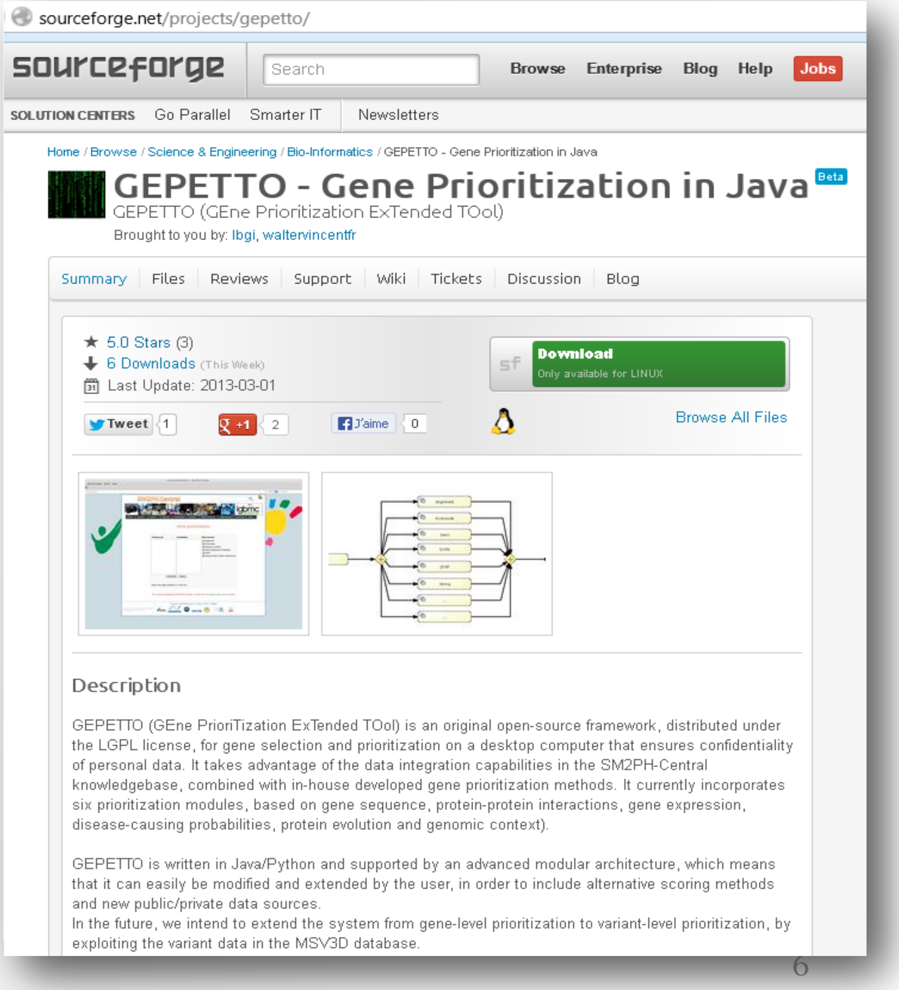
- Biologists and clinicians
  - Web inteface
- Bioinformaticians
  - Standalone application (command line), Java API

STANDARDIZATION / FLEXIBILTY

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

VERSIONING/ CONTRIBUTION

- SVN (via SourceForge)
- GitHub and Mercurial (in progress)



GEPETTO - Gene Prioritization in Java

GEPETTO (GEne Prioritization ExTended TOol)

Brought to you by: lbgi, waltervincentfr

Description

GEPETTO (GEne PrioriTization ExTended TOol) is an original open-source framework, distributed under the LGPL license, for gene selection and prioritization on a desktop computer that ensures confidentiality of personal data. It takes advantage of the data integration capabilities in the SM2PH-Central knowledgebase, combined with in-house developed gene prioritization methods. It currently incorporates six prioritization modules, based on gene sequence, protein-protein interactions, gene expression, disease-causing probabilities, protein evolution and genomic context).

GEPETTO is written in Java/Python and supported by an advanced modular architecture, which means that it can easily be modified and extended by the user, in order to include alternative scoring methods and new public/private data sources.
In the future, we intend to extend the system from gene-level prioritization to variant-level prioritization, by exploiting the variant data in the MSV3D database.

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

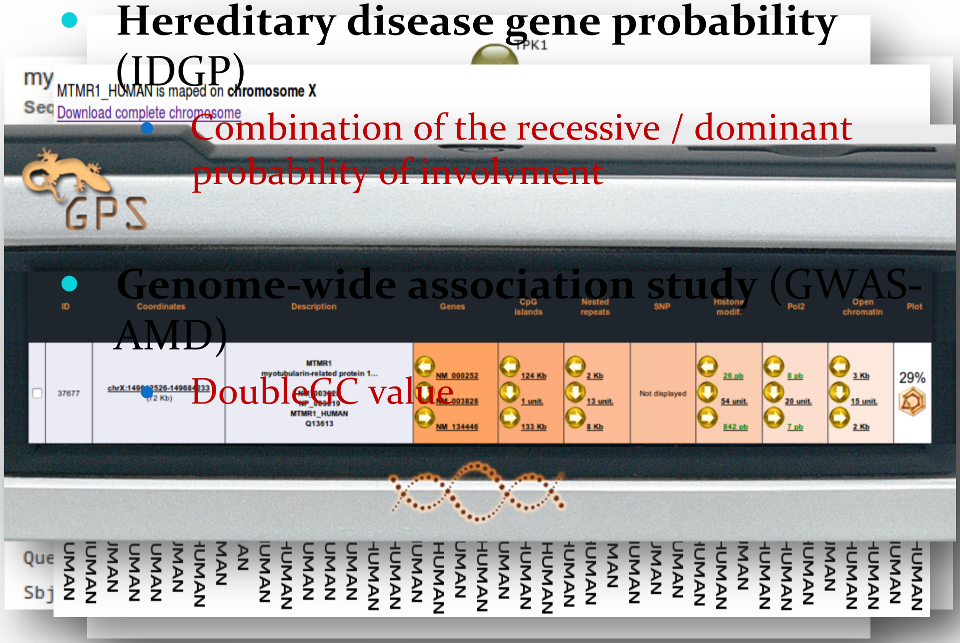
LOCAL PRIORITIZATION MODULES

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



MTMR1_HUMAN is maped on chromosome X

Download complete chromosome

	ID	Coordinates	Description	Genes	CpG islands	Nested repeats	SNP	Histone modif.	Pol2	Open chromatin	Plot
□	37677	chrX:149[illegible]526-149684233 (72 Kb)	MTMR1 myotubularin-related protein 1... NM_00392[illegible] NP_003819 MTMR1_HUMAN Q13613	NM_000252 NM_003828 NM_134446	124 Kb 1 unit. 133 Kb	2 Kb 13 unit. 8 Kb	Not displayed	26 pb 54 unit. 842 pb	8 pb 20 unit. 7 pb	3 Kb 15 unit. 2 Kb	29%

Linard, B., et al. (2011). EvoluCode: **Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data.** Evolutionary Bioinformatics 61.

GLOBAL PRIORITIZATION

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

Aerts S, et al. **Gene prioritization through genomic data fusion. Nat Biotechnol. 2006 May;24(5): 537-44.**

Britto, R., et al. (2012). **GPSy: a cross-species gene prioritization system for conserved biological processes--application in male gamete development.** Nucleic Acids Research.

ORDER STATISTICS

- Able to handle features with missing values.
- Minimizes the bias for known or well-characterized features
- $Q(r_1, r_2, \dots, r_N) = N! \int_0^{r_1} \int_0^{r_2} \dots \int_0^{r_n} ds_N \, ds_{N-1} \dots ds_1$

ROBUST RANK AGGREGATION

- Similar to order statistiss
- More significant result for global prioritization
- $P(r_k^{\prime} \leq r_k) = \sum_{i=k}^{n} \binom{n}{i} \cdot (r_k)^{i} (1 - r_k)^{n-i}$

MALLOWS MODEL

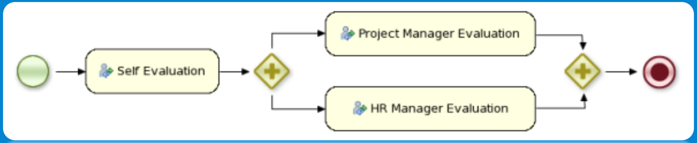
- Looks for a consensus ranking
- $(\pi_1, \dots, \pi_n) = 1/n \cdot \sum_{k=1}^{n} 1_{\{i < \pi_k^j\}}$

GPSy

- Assigns weights to the different criteria to find the optimal weight combination
- $r_i = \sum_{j=1}^{n} w_j \cdot r_{ij} \, / \, \sum_{j=1}^{n} w_j$

GEPETTO WORKFLOW USING JBPM

- A division of an american multinational software company : RedHat
- Specialized in writting 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

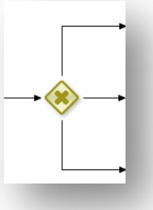
Field usage	• Universal / Not oriented • More flexible
Installation	• Import .JAR libraries • Very easy
Skills required	• No computer programming competences required
Workflow – possibilities	• Complex • Parallel gateway or exclusive gateway (decisionnal workflow)
Workflow – definition	• BPMN2 (Business Process Managment 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

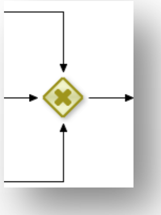
JBPM : EXAMPLE OF SPECIFIC FEATURES

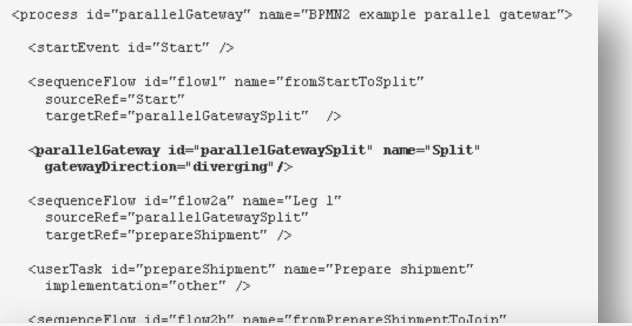
PARALLEL GATEWAY

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

GEPETTO : Used for local prioritization





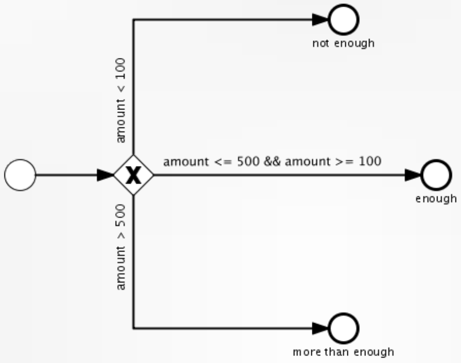


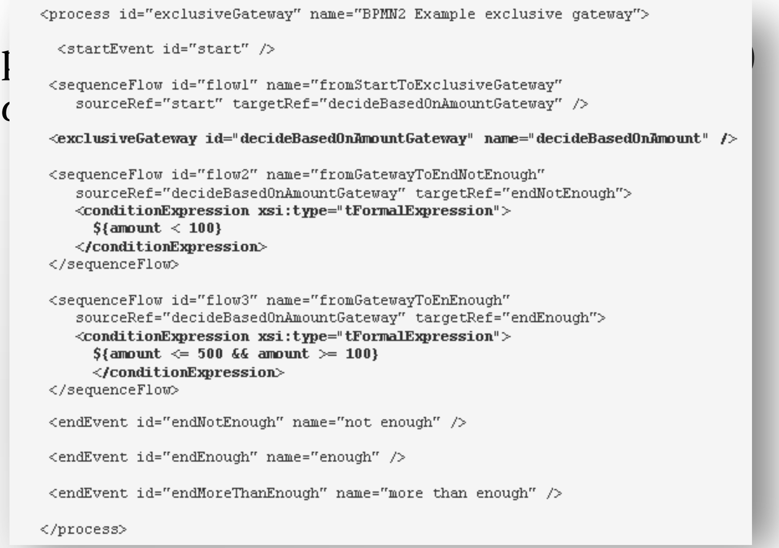
```
<process id="parallelGateway" name="BPMN2 example parallel gatewar">

  <startEvent id="Start" />

  <sequenceFlow id="flow1" name="fromStartToSplit"
    sourceRef="Start"
    targetRef="parallelGatewaySplit"  />

  <parallelGateway id="parallelGatewaySplit" name="Split"
    gatewayDirection="diverging"/>

  <sequenceFlow id="flow2a" name="Leg 1"
    sourceRef="parallelGatewaySplit"
    targetRef="prepareShipment" />

  <userTask id="prepareShipment" name="Prepare shipment"
    implementation="other" />

  <sequenceFlow id="flow2b" name="fromPrepareShipmentToJoin"
```

EXCLUSIVE GATEWAY

- Guides the workflow progression using any p

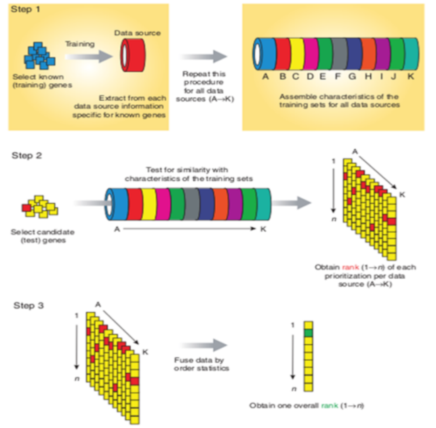
GEPETTO : Used to select the global prioritizatio



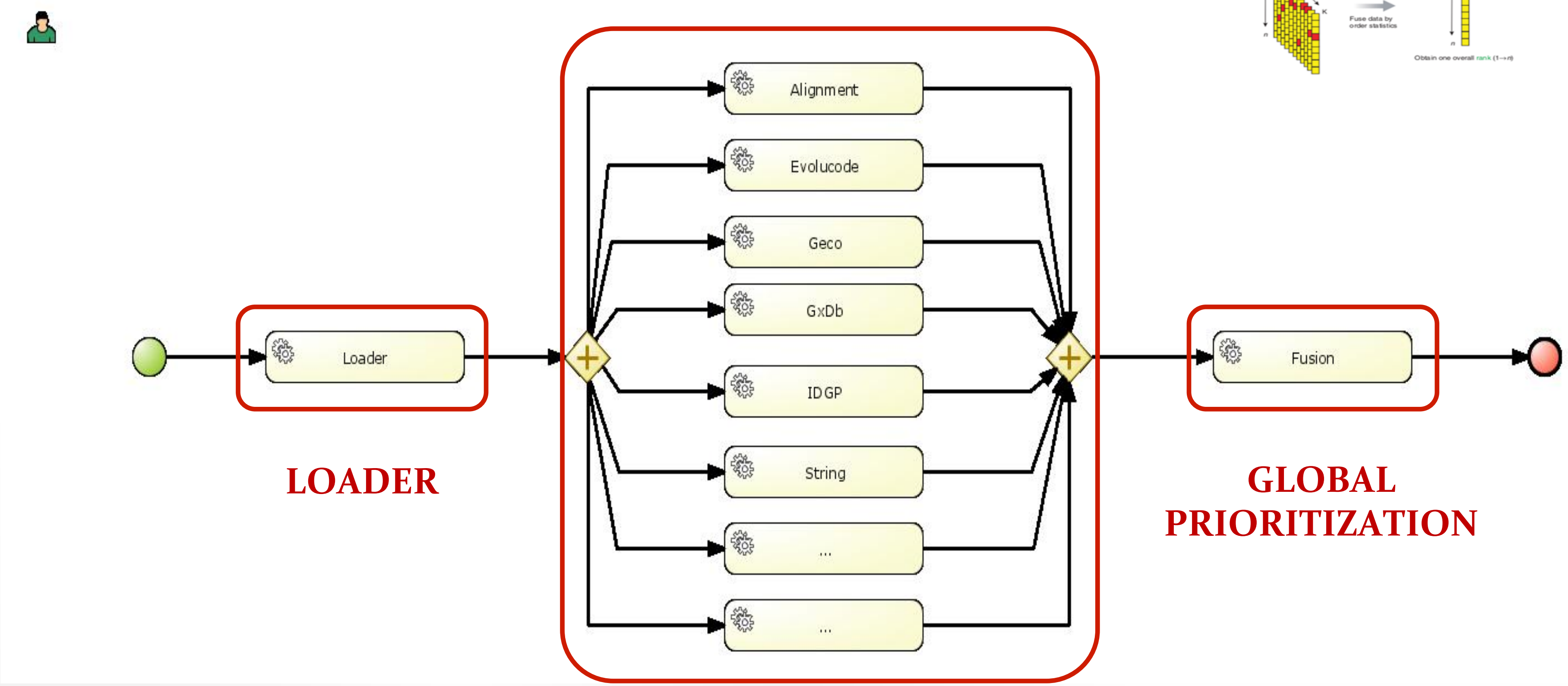


```
<process id="exclusiveGateway" name="BPMN2 Example exclusive gateway">

  <startEvent id="start" />

 <sequenceFlow id="flow1" name="fromStartToExclusiveGateway"
    sourceRef="start" targetRef="decideBasedOnAmountGateway" />

 <exclusiveGateway id="decideBasedOnAmountGateway" name="decideBasedOnAmount" />

 <sequenceFlow id="flow2" name="fromGatewayToEndNotEnough"
    sourceRef="decideBasedOnAmountGateway" targetRef="endNotEnough">
    <conditionExpression xsi:type="tFormalExpression">
      ${amount < 100}
    </conditionExpression>
 </sequenceFlow>

 <sequenceFlow id="flow3" name="fromGatewayToEnEnough"
    sourceRef="decideBasedOnAmountGateway" targetRef="endEnough">
    <conditionExpression xsi:type="tFormalExpression">
      ${amount <= 500 && amount >= 100}
      </conditionExpression>
 </sequenceFlow>

 <endEvent id="endNotEnough" name="not enough" />

 <endEvent id="endEnough" name="enough" />

 <endEvent id="endMoreThanEnough" name="more than enough" />

</process>
```

GEPETTO WORKFLOW USING JBPM


Step 1
Training
Data source
Select known (training) genes
Extract from each data source information specific for known genes
Repeat this procedure for all data sources (A→K)
A B C D E F G H I J K
Assemble characteristics of the training sets for all data sources
Step 2
Test for similarity with characteristics of the training sets
Select candidate (test) genes
A
K
1
n
Obtain rank (1→n) of each prioritization per data source (A→K)
Step 3
Fuse data by order statistics
Obtain one overall rank (1→n)

STEP 2



LOADER

GLOBAL PRIORITIZATION

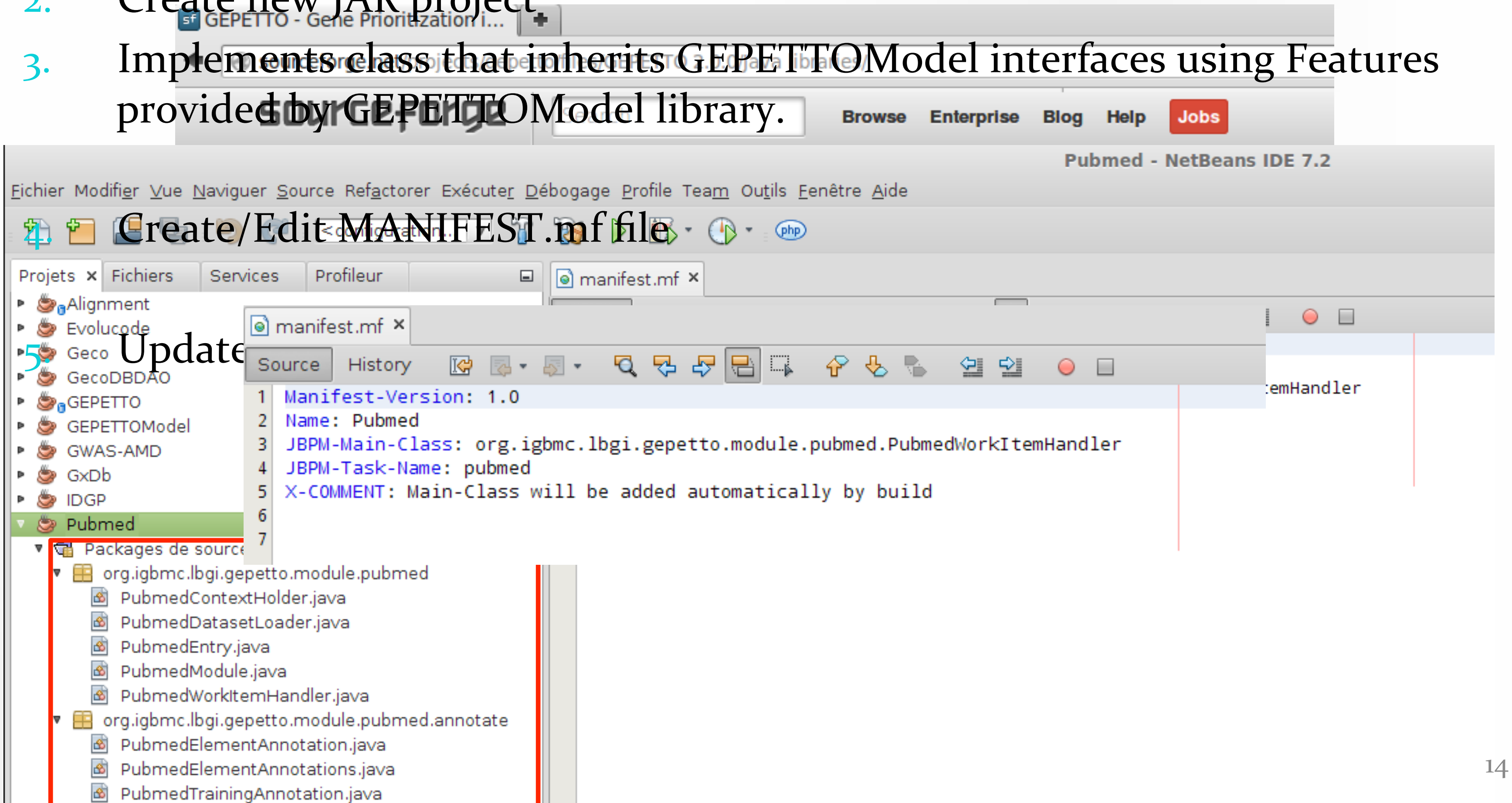
STEP 1

LOCAL PRIORITIZATION

STEP 3

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

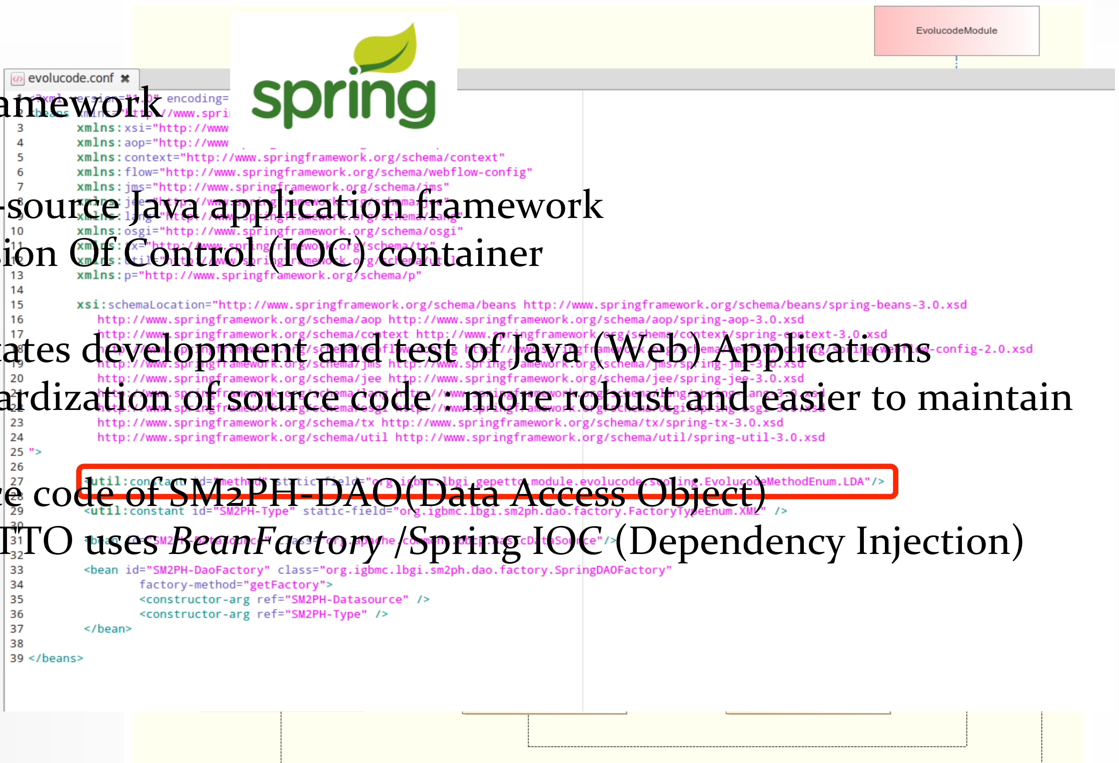

GEPETTO - Gene Prioritization i...
sourceforge
Browse
Enterprise
Blog
Help
Jobs
Pubmed - NetBeans IDE 7.2
Fichier Modifier Vue Naviguer Source Refactorer Exécuter Débogage Profile Team Outils Fenêtre Aide
Projets
Fichiers
Services
Profileur
manifest.mf
Source
History
Alignment
Evolucode
Geco
GecoDBDAO
GEPETTO
GEPETTOModel
GWAS-AMD
GxDb
IDGP
Pubmed
Packages de source
org.igbmc.lbgi.gepetto.module.pubmed
PubmedContextHolder.java
PubmedDatasetLoader.java
PubmedEntry.java
PubmedModule.java
PubmedWorkItemHandler.java
org.igbmc.lbgi.gepetto.module.pubmed.annotate
PubmedElementAnnotation.java
PubmedElementAnnotations.java
PubmedTrainingAnnotation.java

```
Manifest-Version: 1.0
Name: Pubmed
JBPM-Main-Class: org.igbmc.lbgi.gepetto.module.pubmed.PubmedWorkItemHandler
JBPM-Task-Name: pubmed
X-COMMENT: Main-Class will be added automatically by build
```

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)

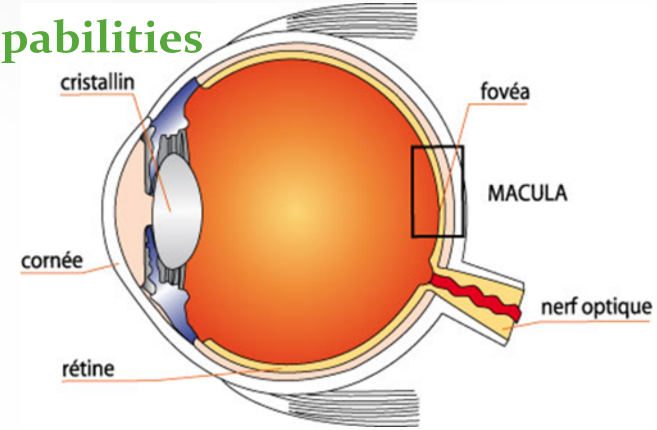


evolucode.conf

```
<?xml version="1.0" encoding=
<beans xmlns="http://www.spri
       xmlns:xsi="http://www
       xmlns:aop="http://www
       xmlns:context="http://www.springframework.org/schema/context"
       xmlns:flow="http://www.springframework.org/schema/webflow-config"
       xmlns:jms="http://www.springframework.org/schema/jms"
       xmlns:jee="http://www.springframework.org/schema/jee"
       xmlns:lang="http://www.springframework.org/schema/lang"
       xmlns:osgi="http://www.springframework.org/schema/osgi"
       xmlns:tx="http://www.springframework.org/schema/tx"
       xmlns:util="http://www.springframework.org/schema/util"
       xmlns:p="http://www.springframework.org/schema/p"

       xsi:schemaLocation="http://www.springframework.org/schema/beans http://www.springframework.org/schema/beans/spring-beans-3.0.xsd
         http://www.springframework.org/schema/aop http://www.springframework.org/schema/aop/spring-aop-3.0.xsd
         http://www.springframework.org/schema/context http://www.springframework.org/schema/context/spring-context-3.0.xsd
         http://www.springframework.org/schema/webflow-config http://www.springframework.org/schema/webflow-config/spring-webflow-config-2.0.xsd
         http://www.springframework.org/schema/jms http://www.springframework.org/schema/jms/spring-jms-3.0.xsd
         http://www.springframework.org/schema/jee http://www.springframework.org/schema/jee/spring-jee-3.0.xsd
         http://www.springframework.org/schema/lang http://www.springframework.org/schema/lang/spring-lang-3.0.xsd
         http://www.springframework.org/schema/osgi http://www.springframework.org/schema/osgi/spring-osgi-3.0.xsd
         http://www.springframework.org/schema/tx http://www.springframework.org/schema/tx/spring-tx-3.0.xsd
         http://www.springframework.org/schema/util http://www.springframework.org/schema/util/spring-util-3.0.xsd
">

  <util:constant id="method" static-field="org.igbmc.lbgi.gepetto.module.evolucode.scoring.EvolucodeMethodEnum.LDA"/>

  <util:constant id="SM2PH-Type" static-field="org.igbmc.lbgi.sm2ph.dao.factory.FactoryTypeEnum.XML" />

  <bean id="SM2PH-Datasource" class="org.apache.commons.dbcp.BasicDataSource"/>

  <bean id="SM2PH-DaoFactory" class="org.igbmc.lbgi.sm2ph.dao.factory.SpringDAOFactory"
          factory-method="getFactory">
          <constructor-arg ref="SM2PH-Datasource" />
          <constructor-arg ref="SM2PH-Type" />
  </bean>

</beans>
```

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)

AMD Gene Consortium, Poch O, Ripp R, Léveillard T among 156 writers, **Seven new loci associated with age-related macular degeneration**. Nat Genet. 2013 Apr;45(4):433-9. doi: 10.1038/ng.2578. Epub 2013 Mar 3.

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)

TRAINING SET	AUC (prioritization of the 23 AMD genes)		
	GEPETTO	ToppGene	Endeavour
Known genes (14)	***0,649***	0,720	0,701
New genes (9)	***0,825***	0,479	0,643
Mixed genes (11)	***0,887***	0,854	0,903

- Conclusion
  - **0,64 < 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

DISCRIMINATION CRITERIA	AUC	
	GEPETTO	Endeavour
TP = AMD known genes ; FP = Other genes (AMD/RP)	**0,967**	1,000
TP = AMD validated genes ; FP = Other genes (AMD/RP)	**0,859**	0,892
TP = AMD candidates genes ; FP = RP genes	**0,752**	0,564

- 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

PARSEC
Poster No44

LBGI
Laboratoire de Bioinformatique et de Génomique Intégratives

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