Bioinformatics & Machine Learning

Daniel Glez-Peña

IP Leiria, June 3 2009

Agenda

1. Bioinformatics

   Definition, major research areas, databases

2. Machine Learning for bioinformatics

   Algorithm types, examples in bioinformatics

3. DNA Microarrays

   Technological overview

4. Applications

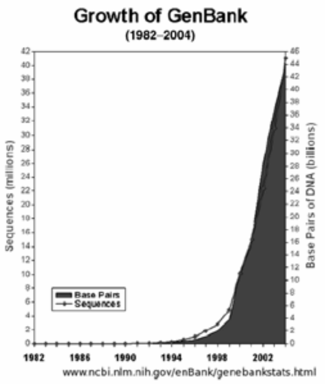
   GeneCBR and WhichGenes?

Bioinformatics

Bioinformatics

- “Application of the Information Technologies to the field of molecular biology”
- Creation and enhancement of:
  - Databases with biological information
  - Algorithms
  - Statistical techniques

…to solve formal and practical problems arising from the management and analysis of biological data


Growth of GenBank

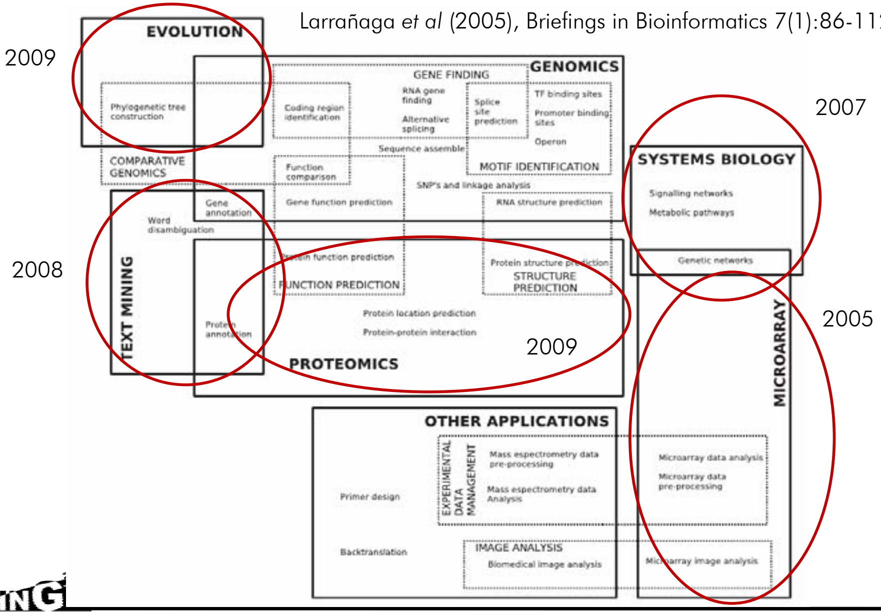
(1982–2004)

www.ncbi.nlm.nih.gov/enBank/genebankstats.html

Major research areas

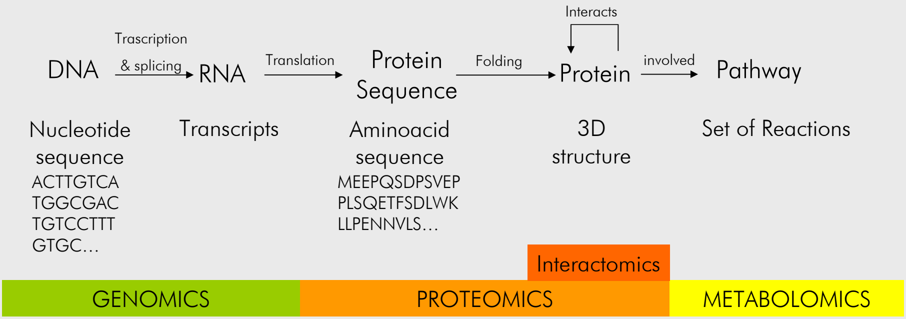
- GENOMICS
  - Sequence analysis
  - Genome annotation
  - Analysis of mutations in cancer
- PROTEOMICS
  - Protein-protein docking
  - Analysis of protein expression
  - Prediction of protein structure
- MICROARRAYS
  - Analysis of gene expression
  - Genetic network induction
- TEXT MINING
  - Gene annotation
  - Protein annotation
  - Relation extraction
- EVOLUTION
  - Phylogenetic reconstruction
  - Comparative genomics
- SYSTEMS BIOLOGY
  - Modelling biological systems
- OTHER
  - Image Analysis

Major research areas



Larrañaga *et al* (2005), Briefings in Bioinformatics 7(1):86-112

Molecular biology dogma



DNA → Trascription & splicing → RNA → Translation → Protein Sequence → Folding → Protein (Interacts) → involved → Pathway

Nucleotide sequence
ACTTGTCA
TGGCGAC
TGTCCTTT
GTGC…

Transcripts

Aminoacid sequence
MEEPQSDPSVEP
PLSQETFSDLWK
LLPENNVLS…

3D structure

Set of Reactions

Interactomics

GENOMICS | PROTEOMICS | METABOLOMICS

Sequence analysis
Genome annotation
Analysis of mutations

Evolution
Phylogenetic reconstruction
Comparative genomics

Gene expression analysis
[DNA microarray]

Protein expression analysis
[mass spectometry]

Protein structure prediction
[folding]

Protein interaction prediction
[3D docking]

Modelling biological systems
Functional analylis

Databases

Genomics

Sequences

Entrez Nucleotide

EMBL Nucleotide Sequence Database

DDBJ DNA Data Bank of Japan

Genomes

Entrez Genome

e! Ensembl

UCSC

TIGR

Cene-centric

Entrez Gene

GeneCards®

Bibliome

PubMed

Proteomics

Proteins

swissprot

TrEMBL Translated EMBL

Entrez Protein

UniProt the universal protein resource

PIR

Structure

PDB Protein Data Bank

SCOP

CATH

Domains

prosite

InterPro

Conserved Domains

Pfam

SMART

Experimental data

GEO Gene Expression Omnibus

AE ArrayExpress

Interactomics & Metabolomics

Prot-Prot interactions

STRING

IntAct

Pathways

KEGG

BIOCARTA

R

PathwayInteractionDatabase

GenMAPP Gene Map Annotator and Pathway Profiler

Ontologies

the Gene Ontology

MeSH

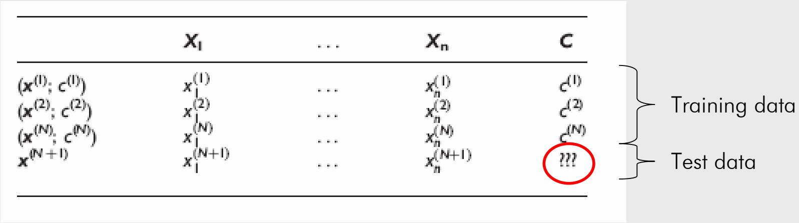
Machine Learning for Bioinformatics

Machine Learning & Bioinformatics

- CLASSIFICATION (SUPERVISED LEARNING)
- CLUSTERING (UNSUPVERVISED LEARNING)
- GRAPHICAL PROBABILISTIC MODELS
- OPTIMIZATION

ML & Bioinformatics: Classification

- Classification (supervised learning)
  - Given a set of “instances”, each one with a set of measured “attributtes” and a “outcome” value we want to train a model that predicts the outcome in further problem instances
    - If the “outcome” is discrete (typical 2 o more different values) we are talking about **classification** (if not: regression)



	X_1	...	X_n	C
$(\mathbf{x}^{(1)}; c^{(1)})$	$x_1^{(1)}$	...	$x_n^{(1)}$	$c^{(1)}$
$(\mathbf{x}^{(2)}; c^{(2)})$	$x_1^{(2)}$	...	$x_n^{(2)}$	$c^{(2)}$
$(\mathbf{x}^{(N)}; c^{(N)})$	$x_1^{(N)}$	...	$x_n^{(N)}$	$c^{(N)}$
$\mathbf{x}^{(N+1)}$	$x_1^{(N+1)}$	...	$x_n^{(N+1)}$	???

Training data

Test data

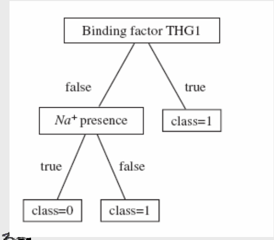
ML & Bioinformatics: Classification

- Classification
  - Feature subset selection.
    - Are all input attributes useful?
    - Advantages: reduced cost in data adquisition, improved uniderstability of the model, faster training, and better accuracy
    - It is a search space problem (2^n-1), in general:
      - 1. Generate a subset
        [brute force, deterministic/not deterministic heuristic search]
      - 2. Evaluate subset
        Statistical estimation: Information Gain, X2, t-test, DFP, CFS
        Wrapper (use classifier accuracy in training set)
      - 3. if (!halt_condition) GOTO 1

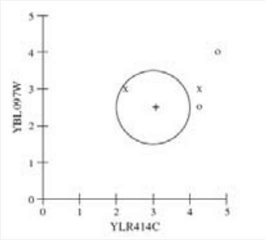
ML & Bioinformatics: Classification

- Classification
  - Popular techniques
    - Logistic regression
    - Linear discriminant analysis (LDA)
    - Bayesian classifiers: Naive Bayes, semi-NB, Tree augmented NB, k dependence Bayesian…
    - Classification trees: CART, C4.5, RandomForest, J48…
    - K-Nearest Neighbours
    - Support Vector Machines
    - Meta: Bagging, Boosting

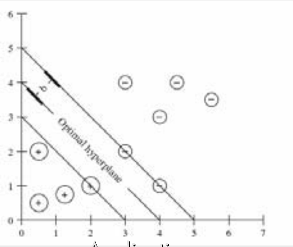
Classification Tree



kNN classifier



Support Vector Machine



ML & Bioinformatics: Classification

- Examples of classification in Bioinformatics (I)
  - Genomics
    - Gene finding (if a sequence is a coding region)
    - Splice site prediction (if a sequence is a splice site)
    - Predict disease genes (from i.e. its sequence length?)
    - Prediction of mutation (SNP) effect
    - Cancer prediction from gene expression (microarrays)
  - Proteomics
    - Prediction of secondary structure (alpha-helix, beta-sheet,etc.)
    - Prediction of sub-cellular location of the protein
    - Cancer prediction from protein expression (mass spectra)

ML & Bioinformatics: Classification

- Examples of classification in Bioinformatics (and II)
  - Systems biology
    - Predict the cell migration speed (high, low) from the phosphorilation levels of signalling proteins
    - Predict a gene regulatory level (up-regulated or down-regulated given the ‘related’ genes expression)
  - Text mining
    - Protein/gene recognition in biomedical literature (is this word a gene/protein given some word features: ortographic, part-of-speech, suffix, trigger words, etc…??)

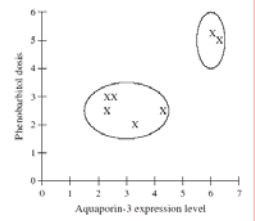
ML & Bioinformatics: Clustering

- Clustering
  - Partition a set of “instances” in several groups (clusters) given the differences between them
    - Their are based on “distances” between instances that is a problem-dependant issue
      - Typical: Euclidean, Pearson, Sperman

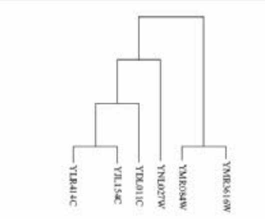
ML & Bioinformatics: Clustering

- Clustering
  - Popular techniques
    - Partition clustering: k-means, SOM, GCS, PAM
    - Hierarchical clustering with single-linkage, complete linkage, centroid linkage and wards-criterion
      - They produce the popular "dendograms"
    - Model-based clustering

Partition clustering



Hierarchical clustering (dendogram)



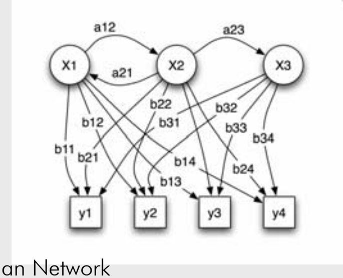
ML & Bioinformatics: Clustering

- Clustering in Bioinformatics
  - Mainly applied to analyze gene expression data
    - Co-Expression detection (group genes with similar expression)
    - Subclass discovery (group samples given the expression of its genes)
    - Expression data visualization/summarization with dendograms

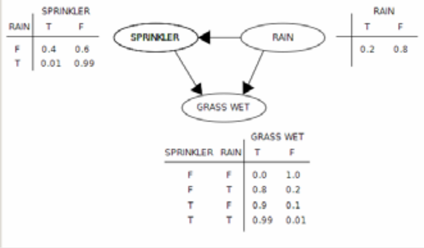
ML & Bioinformatics: Probabilistic graphical models

- DAGs where nodes are random variables and links are probabilities from any kind of conditional dependence
  - Examples
    - Hidden Markov Models
    - Bayesian Networks

Hidden Markov Model



Bayesian Network



RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

ML & Bioinformatics: Probabilistic graphical models

- Probabilistic Graph Models in Bioinformatics
  - Genomics
    - HMM to gene finding (does a gene sequence come from a coding or a non coding DNA region?)
    - Bayesian networks to detect splice sites (does a gene sequence come from a splice-site)
  - Systems Biology
    - Inference of regulatory genetic networks. Bayesian networks to expression pattern recognition (which genes cause other genes to express?)

ML & Bioinformatics: Optimization

- Optimization
  - Search of the best solution in a huge (exponential) space.
  - Popular techniques
    - Exact optimization
      - Brute force
    - Deterministic
      - Hill climbing, local optimization
    - Stochastic
      - Monte Carlo
      - Simulated Annealing
      - Tabu search
      - Evolutionary
        - Genetic algorithms
        - Genetic Programming
        - Estimation of probability

ML & Bioinformatics: Optimization

- Optimization techniques in Bioinformatics
  - Genomics
    - Multiple sequence alignment (used almost all optimization algorithms)
    - Splice site prediction with estimation of distribution algorithms
    - DNA sequencing
    - Cluster microarray data
  - Proteomics
    - Protein folding (predict 3D structure)
    - Protein side-chain prediction (determine the optimal set of ‘angles’ in the 3D structure that minimize the energy)
  - Systems Biology
    - Inference of gene networks and estimate the parameters of bioprocesses
  - Evolution
    - Inference of phylogenetic trees
    - Haplotype reconstruction

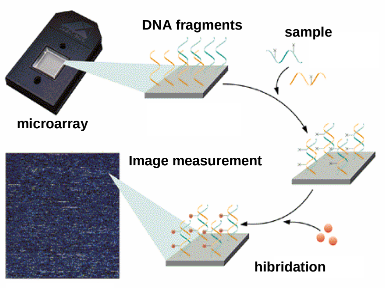
DNA Microarrays

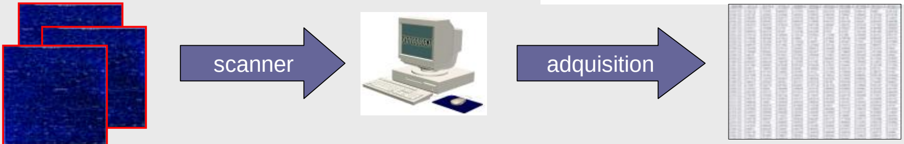
DNA Microarrays

- DNA microarray. Objetive: Measure gene expression
- Description
  - Matrix with measures the expression of thousands of genes simultaneously
  - Gives a “global” vision of gene activity, and allows comparison
    - Between different individuals
    - Same individual at different times
    - Different tissues

DNA Microarrays

- How it works
  - DNA fragments are spotted or printed in probes on the array surface
    - Each probe is a gene
  - Hibridation is performed with a sample putted onto the array
  - A scanner measures the intensity in each probe





Microarray data

Dataset

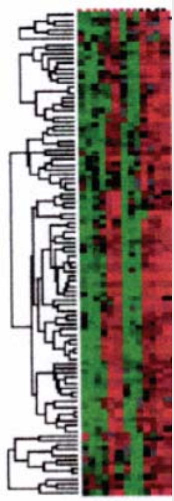
DNA Microarrays

AFFYMETRIX®

- Human Genome U133
  - HG U133A, HG U133B
  - 22.000 probes aprox. (≅1 probe x gen)
- Human Genome U133 plus
  - 44.000 probes (≅2 probes x gen)
- Exon array
  - 1.4 millions of probes (≅16 probes x gen)

DNA microarrays

- Typical analyses & ML Techniques
  - Gene-based analysis
    - Co-expression detection with clustering techniques (unsupervised)
  - Differential gene expression analysis
    - Detect which genes has a significant expression variation among samples of two or more conditions (feature selection)
  - Sample-based analysis
    - Class predicion with classification techniques (supervised)
    - Class discovery with clustering techniques (unsupervised)
  - Problems:
    - Huge number of features (thousands of genes) y low number of samples (dozens) V.S. Machine Learning
    - High false positive rate



DNA microarrays

- Functional interpratation after data analysis
  - Typically we have a list of genes of interest (ie. differentially expressed)
  - Question: who are those genes?
  - Solution: Use the available gene annotations (Gene Ontology, Pathways, etc) and see if there is a correlation with a functional module.
    - They answer to the question: Are my genes significantly chosen from a given gene function? If so, which function?
    - On-line tools
      - List-based: FatiGO, DAVID, Pathjam
      - Gene-set based: GSEA, FatiScan

Sample applications

geneCBR

BMC Bioinformatics

Software

Open Access

DFP: a Bioconductor package for fuzzy profile identification and gene reduction of microarray data

Daniel Glez-Peña[1], Rodrigo Álvarez[2], Fernando Díaz[3] and Florentino Fdez-Riverola*[1]

Address: [1]Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain, [2]Departamento de Informática, University of Vigo, Edificio Fundición, Campus As Lagoas-Marcosende, 36310 Vigo, Pontevedra, Spain and [3]Escuela Universitaria de Informática, University of Valladolid, Plaza Santa Eulalia, 9-11, 40005 Segovia, Spain.

Email: Daniel Glez-Peña - [illegible]; Rodrigo Álvarez - [illegible]; Fernando Díaz - [illegible]; Florentino Fdez-Riverola* - [illegible]

* Corresponding author

Published: 29 January 2009

BMC Bioinformatics 2009, 10:37 doi:10.1186/1471-2105-10-37

Received: 29 September 2008
Accepted: 29 January 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/37

Abstract

Background: Expression profiling assays done by using DNA microarray technology generate enormous data sets that are not amenable to simple analysis. The greatest challenge in maximizing the use of this huge amount of data is to develop algorithms to interpret and interconnect results from different genes under different conditions. In this context, fuzzy logic can provide a systematic and unbiased way to both (i) find biologically significant insights relating to meaningful genes, thereby removing the need for expert knowledge in preliminary steps of microarray data analyses and (ii) reduce the cost and complexity of later applied machine learning techniques being able to achieve interpretable models.

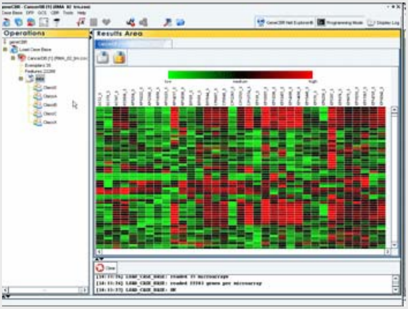
Results: DFP is a new Bioconductor R package that implements a method for discretizing and selecting differentially expressed genes based on the application of fuzzy logic. DFP takes advantage of fuzzy membership functions to assign linguistic labels to gene expression levels. The technique builds a reduced set of relevant genes (FP, Fuzzy Pattern) able to summarize and represent each underlying class (pathology). A last step constructs a biased set of genes (DFP, Discriminant Fuzzy Pattern) by intersecting existing fuzzy patterns in order to detect discriminative elements. In addition, the software provides new functions and visualization tools that summarize achieved results and aid in the interpretation of differentially expressed genes from multiple microarray experiments.

Conclusion: DFP integrates with other packages of the Bioconductor project, uses common data structures and is accompanied by ample documentation. It has the advantage that its parameters are highly configurable, facilitating the discovery of biologically relevant connections between sets of genes belonging to different pathologies. This information makes it possible to automatically filter irrelevant genes thereby reducing the large volume of data supplied by microarray experiments. Based on these contributions geneCBR, a successful tool for cancer diagnosis using microarray datasets, has recently been released.

Page 1 of 8
(page number not for citation purposes)

Translational tool for DNA microarray-based diagnostics

- www.genecbr.org
- Glez-Peña *et al*. BMC Bioinformatics 10:37 2007
- Classification guided by a clustering algorithm GCS



```
[illegible] LOAD_CASE_BASE: readed [illegible] microarrays
[illegible] LOAD_CASE_BASE: readed [illegible] genes per microarray
[illegible] LOAD_CASE_BASE: OK
```

WhichGenes?

Nucleic Acids Research Advance Access published April 30, 2009

Nucleic Acids Research, 2009, 1–6
doi:10.1093/nar/gkp263

WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis

Daniel Glez-Peña[1], Gonzalo Gómez-López[2], David G. Pisano[2] and Florentino Fdez-Riverola[1,3,*]

[1]Higher Technical School of Computer Engineering, University of Vigo, Ourense, [2]Bioinformatics Unit (UBio), Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid and [3]Informatics Department, University of Vigo, Vigo, Pontevedra, Spain

Received January 30, 2009; Revised April 3, 2009; Accepted April 8, 2009

ABSTRACT

WhichGenes is a web-based interactive gene set building tool offering a very simple interface to extract always-updated gene lists from multiple databases and unstructured biological data sources. While the user can specify new gene sets of interest by following a simple four-step wizard, the tool is able to run several queries in parallel. Every time a new set is generated, it is automatically added to the private gene-set cart and the user is notified by an e-mail containing a direct link to the new set stored in the server. WhichGenes provides functionalities to edit, delete and rename existing sets as well as the capability of generating new ones by combining previous existing sets (intersection, union and difference operations). The user can export his sets configuring the output format and selecting among multiple gene identifiers. In addition to the user-friendly environment, WhichGenes allows programmers to access its functionalities in a programmatic way through a Representational State Transfer web service. WhichGenes front-end is freely available at http://www.whichgenes.org/, WhichGenes API is accessible at http://www.whichgenes.org/api/.

INTRODUCTION

During the past several years, bioinformatics enrichment tools have played a very important and successful role contributing to the gene functional analysis of large gene lists (ranging in size from hundreds to thousands of genes) for various high-throughput biological studies (1).

From the large amount of tools that are currently available in the community, two widely used approaches can be identified: (i) individual gene analysis (IGA), which evaluates the significance of individual genes between two groups of samples compared, and (ii) gene set analysis (GSA), free from the problems of the 'cutoff-based' methods (2).

In recent years, GSA approach has received a great deal of attention because, from a biological perspective, functionally related genes often display a coordinated expression to accomplish their roles in the cell. In this direction, GSA methods enable the understanding of cellular processes as an intricate network of functionally related components (3). However, while extensive work has been done during last years in developing new GSA methods, little effort has been put on implementing tools that can help researchers gather, store and manage gene sets containing large 'interesting' gene lists from multiple data sources. To our knowledge, only the Gene Set Builder tool (4) shares the same fundamental concept, giving support to easily handle sets of genes. Compared with Gene Set Builder, which follows a fixed database-driven architecture to give access only to Ensembl and GeneLynx gene catalogs, the tool presented herein follows a more flexible database-free and interactive approach allowing the integration of 14 different data sources for the *Homo sapiens* and *Mus musculus* organisms.

In this article, we present WhichGenes, an online, database-free, web-based tool for easily gathering, building, storing and exporting always updated gene sets coming from multiple data sources. It allows researchers to elaborate custom hypotheses in the form of lists of genes in order to further use them as input in existing GSA tools. WhichGenes currently supports queries about *Homo sapiens* and *Mus musculus* organisms by retrieving up-to-date gene lists directly coming from multiple databases, currently including Ensembl, MSigDB, KEGG/Biocarta/Reactome pathway databases, GeneCards, CancerGenes, Decipher, DiseaseCTD, TargetScan, miRBase, ChemicalCTD, AmiGO and IntAct. Generated gene sets can be

[*]To whom correspondence should be addressed. Tel: +34 988 387015; Fax: +34 988 387001; Email: riverola@uvigo.es

On-line geneset building tool

- Create your own genesets from multiple datasources and use them in your favourite geneset-based analysis tools like GSEA
- www.whichgenes.org
- Glez-Peña *et al*. Nucleic Acids Res (web server issue) 2009

Questions?

SinG33

Upgrading your knowledge TM