Comparison of Metabollic Pathways by Considering Potential Fluxes

Marta Simeoni

Università Ca' Foscari Venezia

Workshop BioPPN 2012 - Amburg, June 25, 2012

Joint work with
Paolo Baldan (Univ. Padova) and Nicoletta Cocco (Univ. Venezia)

Motivation

Comparison of metabolic pathways of different species may be useful for

- understanding metabolic functions
- giving interesting information on their evolution

Motivation

Comparison of metabolic pathways of different species may be useful for

- understanding metabolic functions
- giving interesting information on their evolution

Many techniques have been proposed in the recent literature. Each approach

- chooses a metabolic pathway representation
- proposes a distance measure
- possibly supplies a tool to perform the comparison

Motivation

The distances in the literature generally focus on static, topological information of the pathways...

Motivation

The distances in the literature generally focus on static, topological information of the pathways...
disregarding the fact that they represent dynamic processes

Motivation

The distances in the literature generally focus on static, topological information of the pathways...
disregarding the fact that they represent dynamic processes

We propose to consider both

- structural and
- behavioural

aspects in the definition of a distance between pathways.

Motivation

The distances in the literature generally focus on static, topological information of the pathways...
disregarding the fact that they represent dynamic processes

We propose to consider both

- structural and
- behavioural

aspects in the definition of a distance between pathways.

To this aim, we take advantage of a Petri net representation of metabolic pathways

Metabolic pathways (MPs)

Metabolism: the chemical system which generates the essential components for life

Metabolic pathways:

- subsystems dealing with some specific function
- represented as a network of chemical *reactions* catalised by one or more *enzymes* where some molecules (*reactants or substrates*) are transformed into others (*products*)
- the *stoichiometric matrix* identifies the pathways components and their relations
- kinetics represented by the rate equation associated with each reaction

KEGG pathways

Metabolic pathways information are collected in many different databases (e.g. KEGG, Biomodels, Metacyc)

We consider the KEGG pathway database.

- At present it contains around 93000 pathways
- Pathways are represented by maps with additional information
- Models are coded in KGML (KEGG Markup Language)
- A web service for querying the KEGG system from users programs is available

Petri net (PN) representation of metabolic pathways

Metabolic pathways can be naturally modelled with PNs:

- Places are associated to molecular species (metabolites, enzymes)
- Transitions correspond to chemical reactions
  - Input places are substrates
  - Output places are products
- The incidence matrix of the PN is identical to the stoichiometric matrix of the system of chemical reactions
- The number of tokens in each place of the PN indicates the amount of substance associated with that place

Petri net (PN) representation of metabolic pathways

Metabolic pathways can be naturally modelled with PNs:

- Places are associated to molecular species (metabolites, enzymes)
- Transitions correspond to chemical reactions
  - Input places are substrates
  - Output places are products
- The incidence matrix of the PN is identical to the stoichiometric matrix of the system of chemical reactions
- The number of tokens in each place of the PN indicates the amount of substance associated with that place

We use an extension of our tool MPath2PN to automatically translate metabolic pathways into Petri nets

Our approach

We propose a comparison technique for MPs which considers

- structural aspects: by considering homology of enzymes/reactions and
- behavioural aspects: by considering a measure of the similarity of flows in the pathways

Our approach

We propose a comparison technique for MPs which considers

- structural aspects: by considering homology of enzymes/reactions and
- behavioural aspects: by considering a measure of the similarity of flows in the pathways $\Rightarrow$ T-invariants

Our approach

We propose a comparison technique for MPs which considers

- structural aspects: by considering homology of enzymes/reactions and
- behavioural aspects: by considering a measure of the similarity of flows in the pathways $\Rightarrow$ T-invariants

Minimal (semi-positive) T-invariants correspond to elementary flux modes of a metabolic pathway, i.e. minimal sets of reactions that can operate at a steady state

The set of semi-positive T-invariants has a unique basis, the Hilbert basis, consisting of the minimal T-invariants

Our approach

We propose a comparison technique for MPs which considers

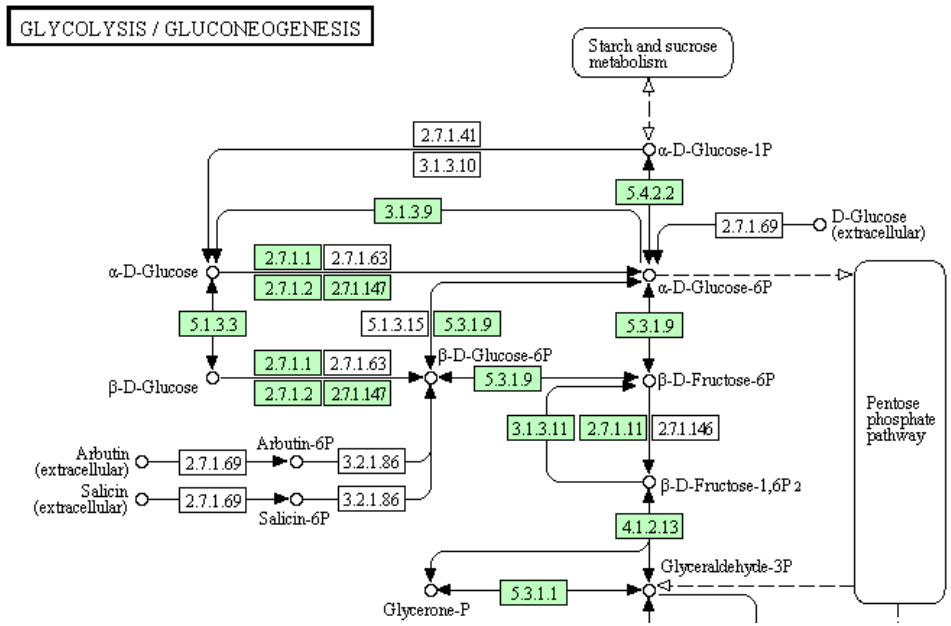
- structural aspects: by considering homology of enzymes/reactions and
- behavioural aspects: by considering a measure of the similarity of flows in the pathways $\Rightarrow$ T-invariants

Minimal (semi-positive) T-invariants correspond to elementary flux modes of a metabolic pathway, i.e. minimal sets of reactions that can operate at a steady state

The set of semi-positive T-invariants has a unique basis, the Hilbert basis, consisting of the minimal T-invariants
$\Rightarrow$ characteristic of the net

Structural aspects

We consider a distance between pathways based on their representation as multisets of reactions/enzymes



Structural aspects

Let X_1 and X_2 be the multisets of reactions of two pathways P_1 and P_2

- Reactions are represented by their EC numbers and the similarity considered between them is the identity
- As a similarity index we choose the Sørensen index on multisets

$$S\_index(X_1, X_2) = \frac{2|X_1 \cap X_2|}{|X_1| + |X_2|}$$

- The distance based on reactions is:

$$d_R(P_1, P_2) = 1 - S\_index(X_1, X_2).$$

Behavioural aspects

We focus on the Hilbert bases $\mathcal{B}(P_1)$ and $\mathcal{B}(P_2)$:

- we see each T-invariant as a multiset of reactions
- the similarity score between two invariants is given by the Sørensen index:

Then $\text{I\_Score}(P_1, P_2)$ is computed by performing an heuristic for the best match between the two bases $\mathcal{B}(P_1)$ and $\mathcal{B}(P_2)$
The induced distance is then:

$$d_I(P_1, P_2) = 1 - \text{I\_Score}(P_1, P_2)$$

Behavioural aspects

function I_Score(P_1, P_2);
  input: two metabolic pathways P_1 and P_2;
  output: the similarity measure between $\mathcal{B}(P_1)$ and $\mathcal{B}(P_2)$;
begin
  $I_1 = \mathcal{B}(P_1)$; $I_2 = \mathcal{B}(P_2)$;
  $score = 0$;
  $card = \max\{|I_1|, |I_2|\}$;
  while $(I_1 \neq \emptyset \ \wedge \ I_2 \neq \emptyset)$ **do**
  begin
    $(X_1, X_2) =$ Find_max_Sim(I_1, I_2); {Returns a pair of T-invariants, (X_1, X_2),
      in $I_1 \times I_2$ such that $S\_index(X_1, X_2)$
      is maximum}
    $score = score + S\_index(X_1, X_2)$;
    $I_1 = I_1 - \{X_1\}$;
    $I_2 = I_2 - \{X_2\}$;
  end;
  $score = score/card$;
  return *score*
end

A family of distances

The two distances d_R and d_I are combined into a weighted sum:

$$d_D(P_1, P_2) = \alpha\ d_R(P_1, P_2) + (1 - \alpha)\ d_I(P_1, P_2)$$

The weight $\alpha \in [0, 1]$ allow the analyst to move the focus between static ($\alpha = 1$) and behavioural ($\alpha = 0$) aspects.

A family of distances

- Two organisms O_1 and O_2 can be compared by considering their similarity on n chosen MPs: $P_1, \ldots, P_n$
- In this case the distances between the two organisms with respect to the various MPs need to be combined
- We adopt a simple solution, which consists in taking the average distance

$$d_D(O_1, O_2) = \frac{\sum_{j=1}^{n} d_D(P_j^1, P_j^2)}{n}$$

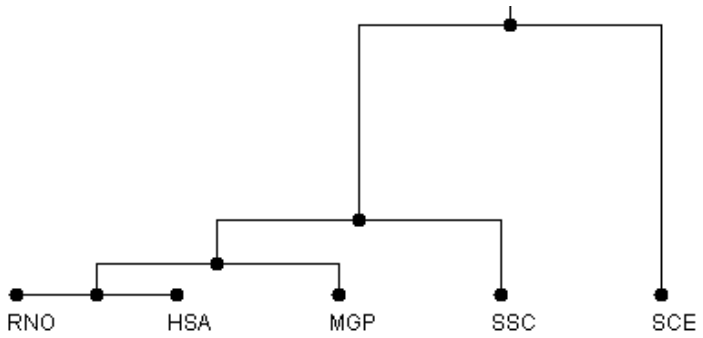
The prototype tool CoMeta

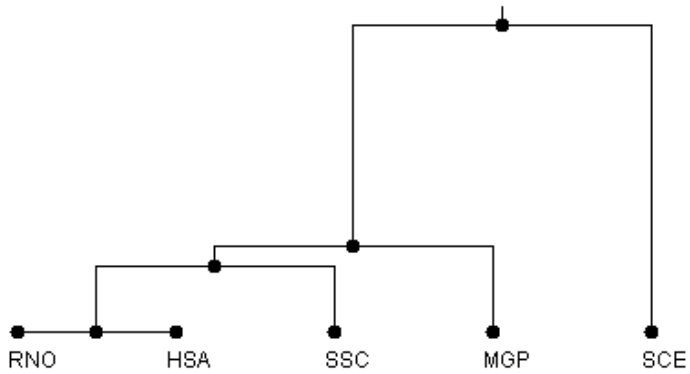
CoMeta (COmparing METAbolic pathways) has been developed to validate our proposal
Its main features are:

- download of the information on the specified organisms and pathways from KEGG
- translate the MPs into corresponding PNs (MPath2PN)
- compute the combined distance for each pair of organisms and build the corresponding distance matrix (INA)
- build and display a phylogenetic tree (UPGMA or Neighbour Joining methods)

Experimenting with CoMeta

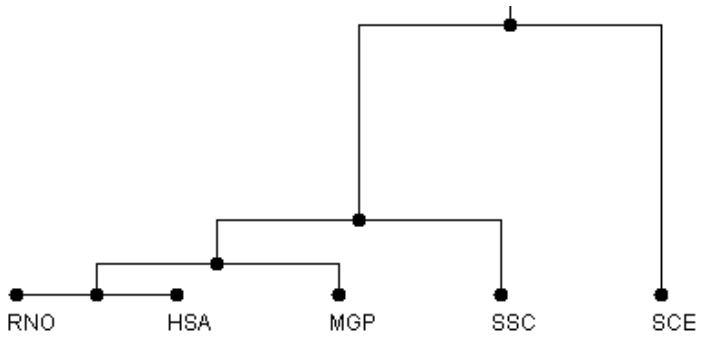
- Consider the glycolysis pathway for *Homo sapiens* (HSA), *Rattus norvegicus* (RNO), *Meleagris gallopavo* (MGP), *Sus scrofa* (SSC), *Saccharomyces cerevisiae* (SCE)
- Compute the distance for $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$
- The resulting UPGMA trees for $\alpha > 0.5$ (left) and $\alpha \leq 0.5$ (right) are:

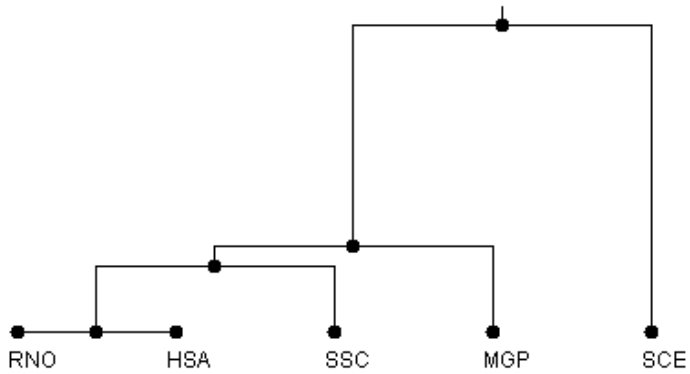




Experimenting with CoMeta

- Consider the glycolysis pathway for *Homo sapiens* (HSA), *Rattus norvegicus* (RNO), *Meleagris gallopavo* (MGP), *Sus scrofa* (SSC), *Saccharomyces cerevisiae* (SCE)
- Compute the distance for $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$
- The resulting UPGMA trees for $\alpha > 0.5$ (left) and $\alpha \leq 0.5$ (right) are:

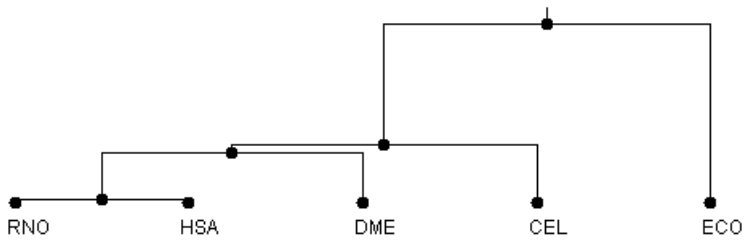


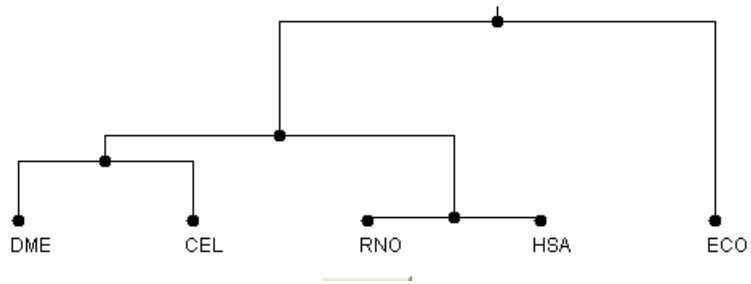


The distance based on invariants produce a better classification

Experimenting with CoMeta

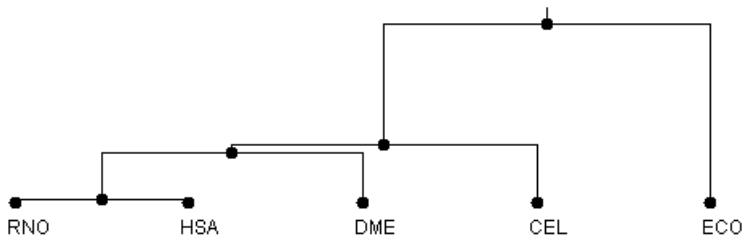
- Consider the glycolysis, pyruvate metabolism and purine metabolism pathways for *Homo sapiens* (HSA), *Rattus norvegicus* (RNO), *C. elegans* (CEL), *Drosophila melanogaster* (DME) and *E. coli* (ECO)
- Compute the distance for $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$
- The resulting UPGMA trees for $\alpha = 1$ (left) and $\alpha \leq 0.75$ (right) are:

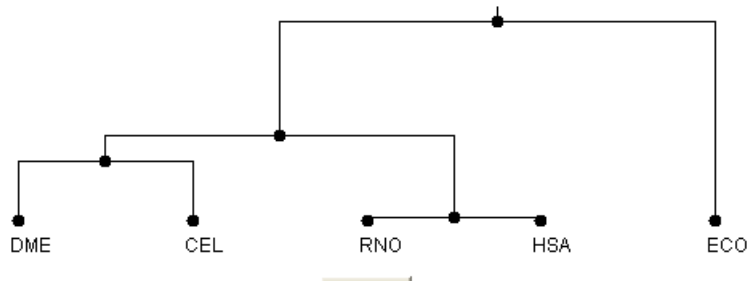




Experimenting with CoMeta

- Consider the glycolysis, pyruvate metabolism and purine metabolism pathways for *Homo sapiens* (HSA), *Rattus norvegicus* (RNO), *C. elegans* (CEL), *Drosophila melanogaster* (DME) and *E. coli* (ECO)
- Compute the distance for $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$
- The resulting UPGMA trees for $\alpha = 1$ (left) and $\alpha \leq 0.75$ (right) are:



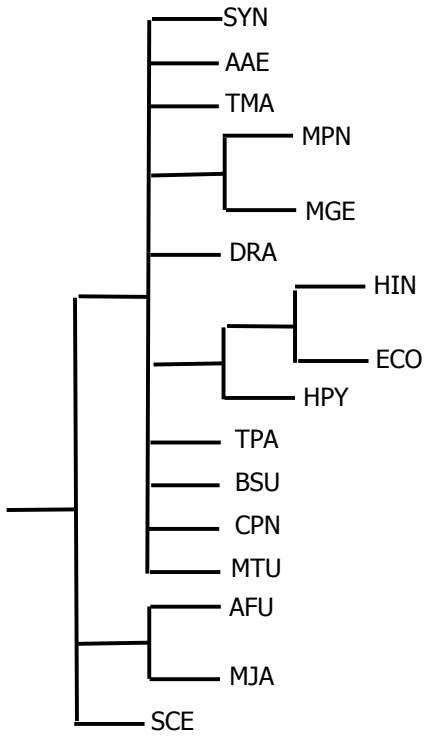


The distance based on invariants produce a worst classification (according to the NCBI taxonomy)

Experimenting with CoMeta

Consider the glycolysis pathway for the following organisms and their NCBI taxonomy:

Cod.	Organism	Reign
afu	A. fulgidus	Archea
mja	M. jannaschii	Archea
cpn	C. pneumoniae	Bacteria
mge	M. genitalum	Bacteria
mpn	M. pneumoniae	Bacteria
hin	H. influenzae	Bacteria
syn	Synechocystis	Bacteria
dra	D. radiodurans	Bacteria
mtu	M. tuberculosis	Bacteria
tpa	T. pallidum	Bacteria
bsu	B. subtilis	Bacteria
aae	A. aeolicus	Bacteria
tma	T. maritima	Bacteria
eco	E. coli	Bacteria
hpy	H. pylori	Bacteria
sce	Saccharomyces cerevisiae	Eucaryotes

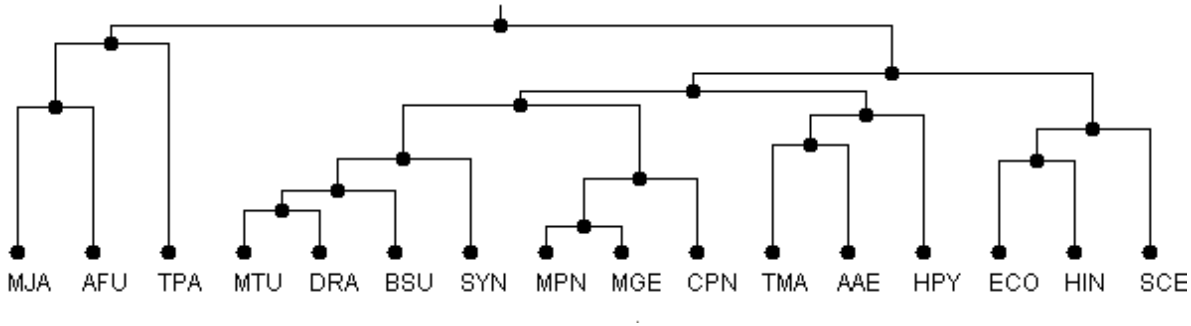


Compute the distance for $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$

Experimenting with CoMeta

Similarity values computed with cousins and best UPGMA tree:

α	Similarity value
0.00	0.25
0.25	0.2673267
0.50	0.3163265
0.75	0.3163265
1.00	0.2621359



- Our results cannot be immediately compared with those in the literature since
  - NCBI classification and
  - KEGG data

  have been changing in the meantime
- Neverthenless, our technique produces results which are
  - consistent with the reference classification
  - comparable with those in the literature

Concluding remarks

A framework for MP comparison:

- MPs are represented as PNs
- static and behavioural aspects are combined into a family of distance measures
- the prototype tool CoMeta is available

Experiments made with CoMeta shows that:

- Our combined measure produces valid phylogenetic classifications
- Measures based on more sophisticated representations of a pathway not necessarily give better results than our combined measure
- However, we need to perform more experiments to determine which combination of the two proposed distances gives the best results

Future work

- We are extending of CoMeta by
  - Adding a more refined similarity measure on EC numbers
  - Adding the Tanimoto index, beside the Sørensen one
- We are performing extensive studies on the distributions of the two proposed distances
- It would be very interesting to compare different organisms by considering their whole metabolic networks

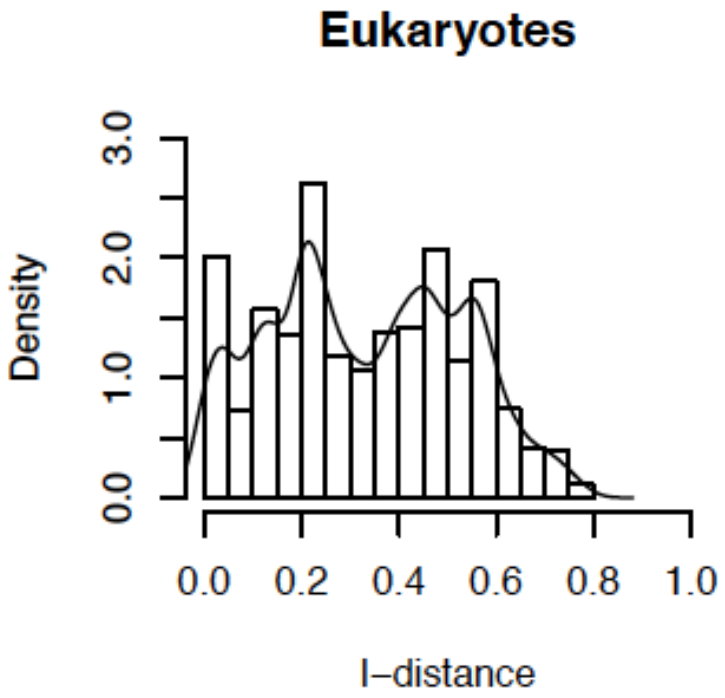
Future work

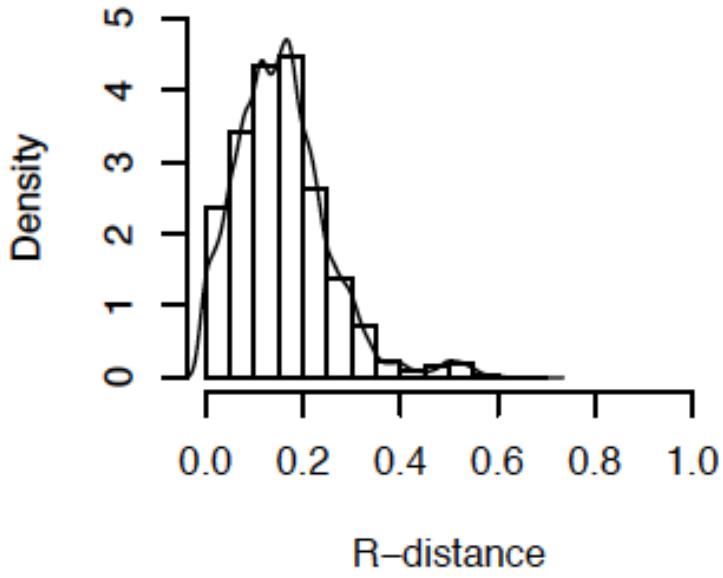
- We are extending of CoMeta by
  - Adding a more refined similarity measure on EC numbers
  - Adding the Tanimoto index, beside the Sørensen one
- We are performing extensive studies on the distributions of the two proposed distances
- It would be very interesting to compare different organisms by considering their whole metabolic networks
  ...but the Hilbert bases can be exponential in the size of the original net

Studying the distribution of the proposed distances

- We explored the metabolic pathways in KEGG with CoMeta, in order to validate the tool and analyse the significance of our proposed distances d_R and d_I.
- We considered different pathways and different classes of organisms.
- For each class we studied the distribution of the values of our two distances for all the pairs of organisms in it.
- We report here some preliminary results regarding the Glycolysis pathway and the Sørensen index only
- We consider the class of Eukaryotes, and its subclasses Animals, Vertebrates and Mammals (each subclass is included into the previous one).

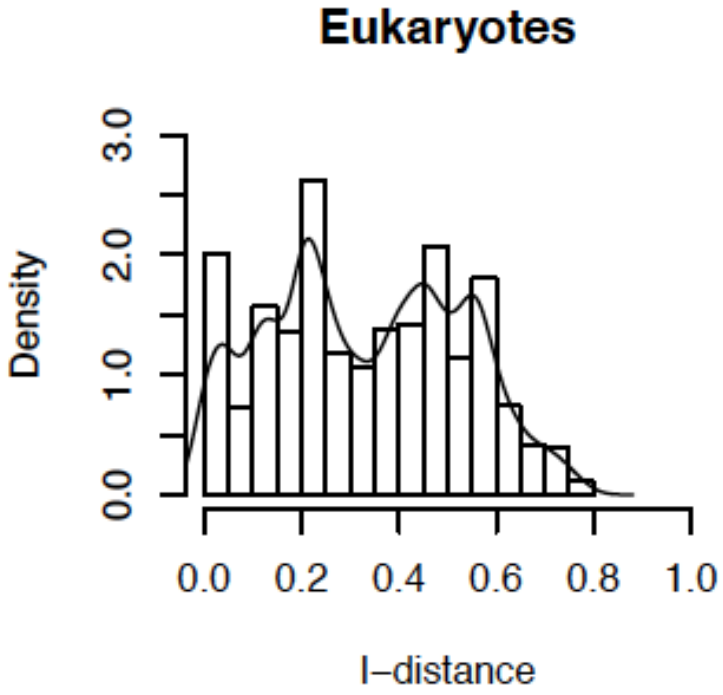
Eukaryotes

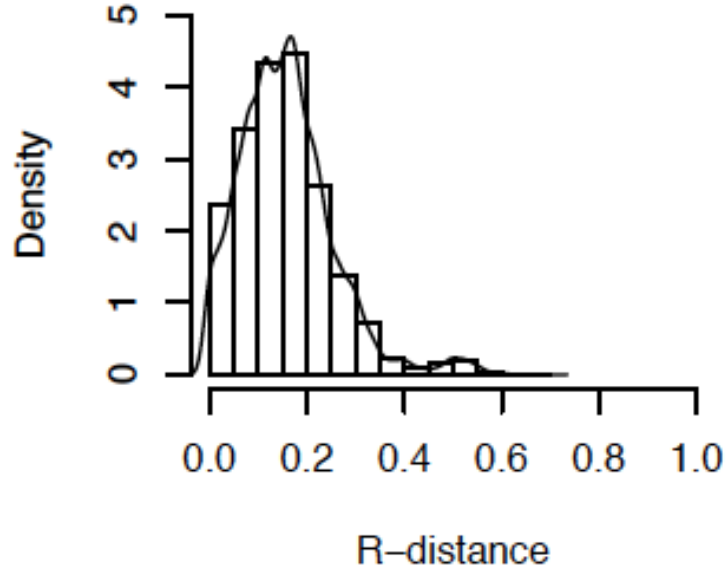




- I-distance shows a rather flat distribution ranging from 0 to 0.8
- R-distance takes values in the interval $[0, 0.55]$, mostly concentrated in $[0.05, 0.25]$

Eukaryotes

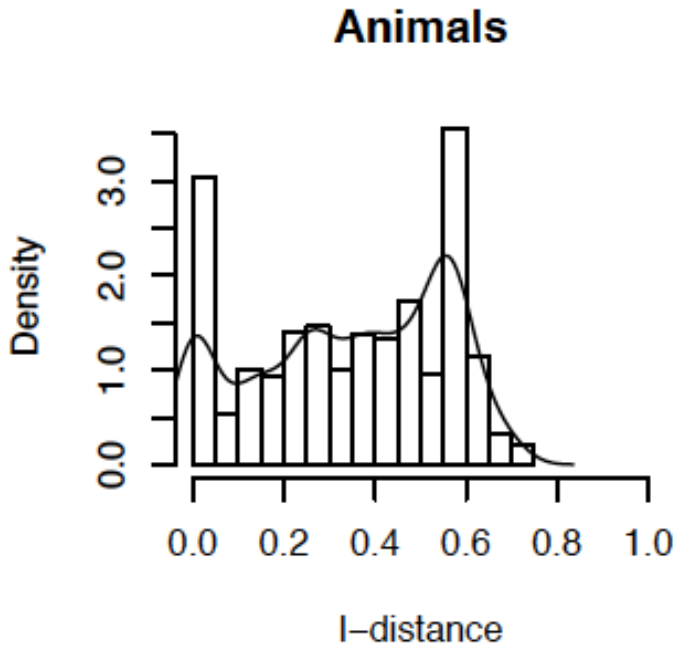


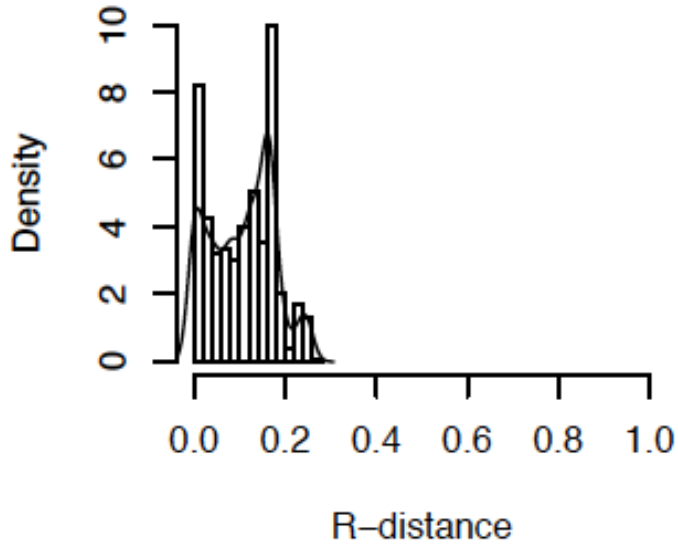


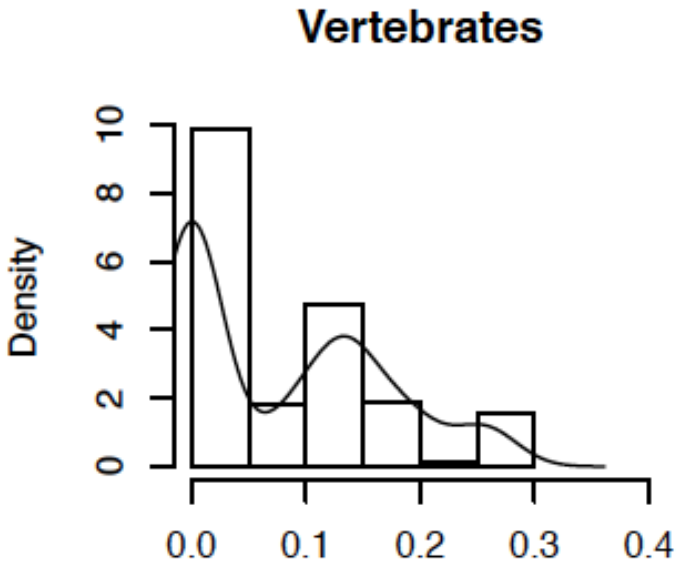
- I-distance shows a rather flat distribution ranging from 0 to 0.8
- R-distance takes values in the interval $[0, 0.55]$, mostly concentrated in $[0.05, 0.25]$

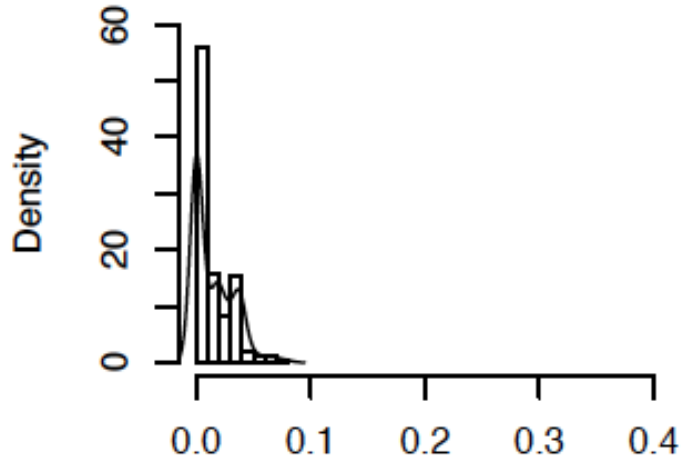
This suggests that, in this case, the I-distance discriminates more than the R-distance

Animals, Vertebrates...

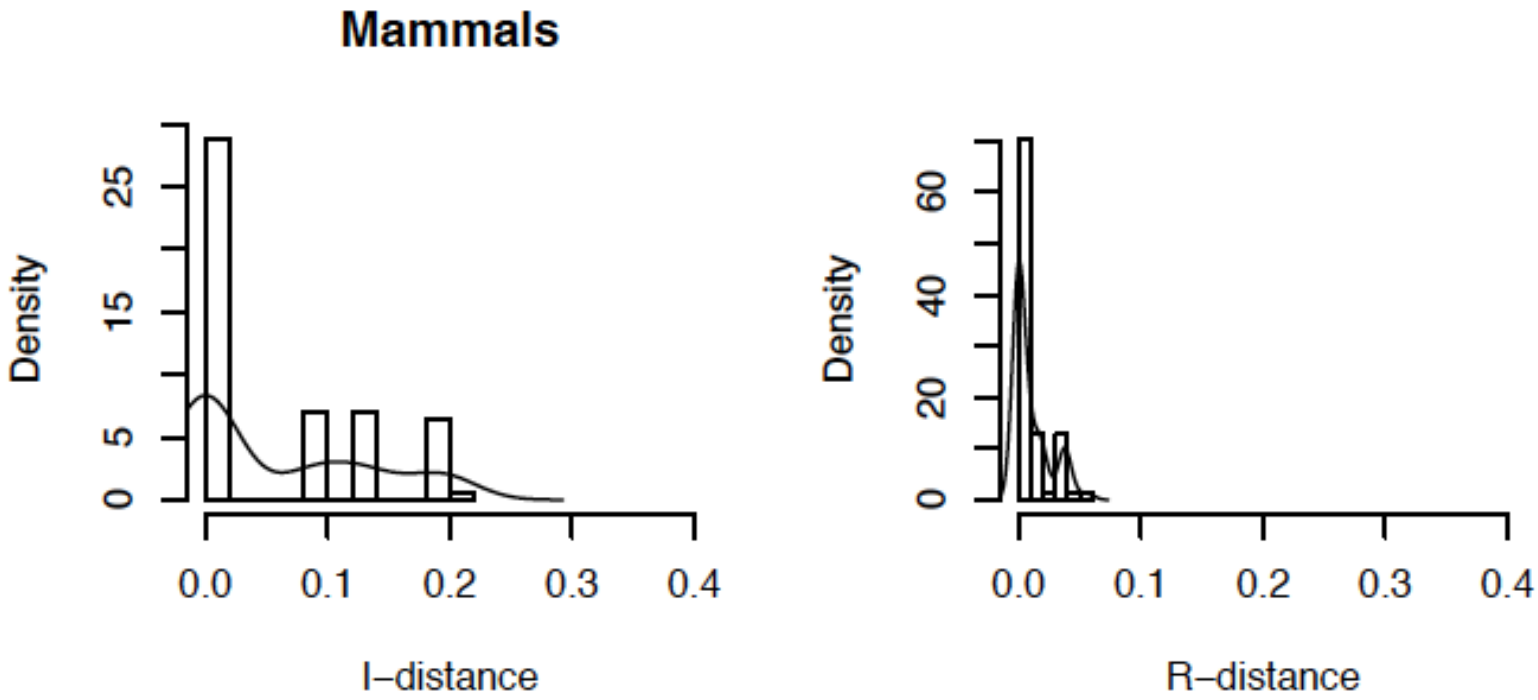








... and Mammals



- The classes become more and more homogeneous
- This is correctly represented by our two distances which become narrower in range and more and more similar between themselves