FUZZY CLUSTERING IMPROVES PHYLOGENETIC RELATIONSHIPS RECONSTRUCTION FROM METABOLIC PATHWAYS

J. Casasnovas (UIB) J. C. Clemente (JAIST)
J. Miró (UIB) F. Rosselló (UIB)
K. Satou (JAIST) G. Valiente (UPC)

IPMU 2006, Paris, July 2–7, 2006

Plan of the talk

- Motivation

Plan of the talk

- Motivation
- Background

Plan of the talk

- Motivation
- Background
- Similarity of metabolic pathways

Plan of the talk

- Motivation
- Background
- Similarity of metabolic pathways
- Fuzzy clustering

Plan of the talk

- Motivation
- Background
- Similarity of metabolic pathways
- Fuzzy clustering
- Results

Plan of the talk

- Motivation
- Background
- Similarity of metabolic pathways
- Fuzzy clustering
- Results
- Conclusion and future work

Motivation

Evolutionary relationships among species have been mainly understood through the "molecular approach."
But...

Motivation

Evolutionary relationships among species have been mainly understood through the "molecular approach." But...

Recent results related to horizontal gene transfer suggest that phylogenetic reconstruction cannot be determined conclusively from sequence data.

Motivation

Evolutionary relationships among species have been mainly understood through the "molecular approach." But...

Recent results related to horizontal gene transfer suggest that phylogenetic reconstruction cannot be determined conclusively from sequence data.

The increasing amount of available information on metabolic pathways for several species motivates the use of similarities among such pathways to infer phylogenetic trees not based exclusively in sequence data.

Background

Metabolic pathways

- Series of chemical reactions occurring within a cell, catalyzed by enzymes, to achieve formation of specific metabolic products from set of substrates

Background

Metabolic pathways

- Series of chemical reactions occurring within a cell, catalyzed by enzymes, to achieve formation of specific metabolic products from set of substrates
- Usually represented as hypergraphs, with hyperedges being reactions that connect set of substrates, products and enzymes

Background

Metabolic pathways

- Series of chemical reactions occurring within a cell, catalyzed by enzymes, to achieve formation of specific metabolic products from set of substrates
- Usually represented as hypergraphs, with hyperedges being reactions that connect set of substrates, products and enzymes
- KEGG (Kyoto Encyclopedia of Genes and Genomes): over 32 000 pathways, 382 species

Background

Phylogenetic Trees

- Evolutionary relationships among several species believed to have a common ancestor

Background

Phylogenetic Trees

- Evolutionary relationships among several species believed to have a common ancestor
- Assumption: similar metabolic pathways can reflect similar evolutionary history

Background

Phylogenetic Trees

- Evolutionary relationships among several species believed to have a common ancestor
- Assumption: similar metabolic pathways can reflect similar evolutionary history
- Develop a measure of pathway similarity based on component chemical substances and enzymes

Background

Phylogenetic Trees

- Evolutionary relationships among several species believed to have a common ancestor
- Assumption: similar metabolic pathways can reflect similar evolutionary history
- Develop a measure of pathway similarity based on component chemical substances and enzymes
- Improve phylogenetic tree reconstruction methods using these similarities

Similarity of metabolic pathways

Structural similarity of metabolic pathways entails:

- a hypergraph representation of a metabolic pathway

Similarity of metabolic pathways

Structural similarity of metabolic pathways entails:

- a hypergraph representation of a metabolic pathway
- a similarity measure between individual reactions, enzymes, and compounds

Enzyme Similarity Measures

- Enzyme Commission (EC) numbers

Enzyme Similarity Measures

- Enzyme Commission (EC) numbers
  - Hierarchical classification of enzymes based on functional categories

Enzyme Similarity Measures

- Enzyme Commission (EC) numbers
  - Hierarchical classification of enzymes based on functional categories
  - Each enzyme is described by a four digit identifier

Enzyme Similarity Measures

- Enzyme Commission (EC) numbers
  - Hierarchical classification of enzymes based on functional categories
  - Each enzyme is described by a four digit identifier
- Hierarchical similarity, HIER

Enzyme Similarity Measures

- Enzyme Commission (EC) numbers
  - Hierarchical classification of enzymes based on functional categories
  - Each enzyme is described by a four digit identifier
- Hierarchical similarity, HIER
  - Number of common most significant EC digits over 4

Enzyme Similarity Measures

- Enzyme Commission (EC) numbers
  - Hierarchical classification of enzymes based on functional categories
  - Each enzyme is described by a four digit identifier
- Hierarchical similarity, HIER
  - Number of common most significant EC digits over 4
  - Shortest path in the hierarchy between enzymes

Enzyme Similarity Measures

- Enzyme Commission (EC) numbers
  - Hierarchical classification of enzymes based on functional categories
  - Each enzyme is described by a four digit identifier
- Hierarchical similarity, HIER
  - Number of common most significant EC digits over 4
  - Shortest path in the hierarchy between enzymes
- Information content similarity, INFO

Enzyme Similarity Measures

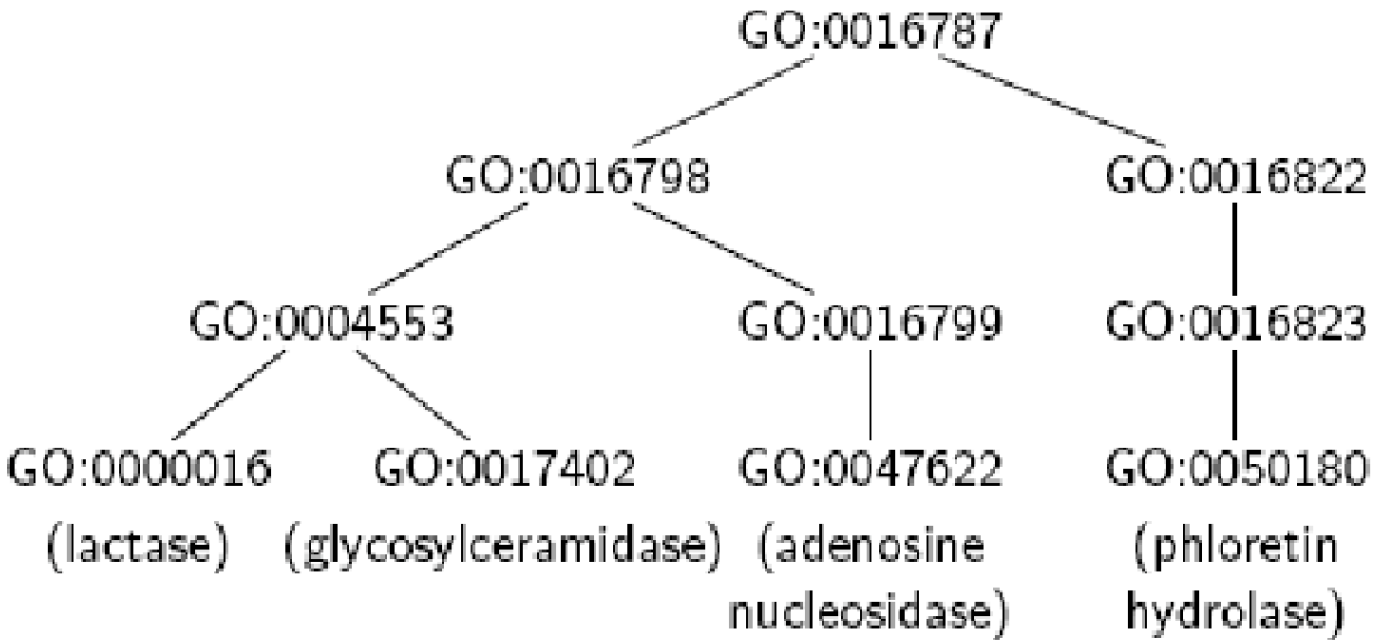
- Enzyme Commission (EC) numbers
  - Hierarchical classification of enzymes based on functional categories
  - Each enzyme is described by a four digit identifier
- Hierarchical similarity, HIER
  - Number of common most significant EC digits over 4
  - Shortest path in the hierarchy between enzymes
- Information content similarity, INFO
  - $sim(e_i, e_j) = 1 - log_2 E(lca(e_i, e_j))/k$

Enzyme Similarity Measures

- Enzyme Commission (EC) numbers
  - Hierarchical classification of enzymes based on functional categories
  - Each enzyme is described by a four digit identifier
- Hierarchical similarity, HIER
  - Number of common most significant EC digits over 4
  - Shortest path in the hierarchy between enzymes
- Information content similarity, INFO
  - $sim(e_i, e_j) = 1 - log_2 E(lca(e_i, e_j))/k$
  - Size of subtree rooted at least common ancestor of the enzymes in the hierarchy

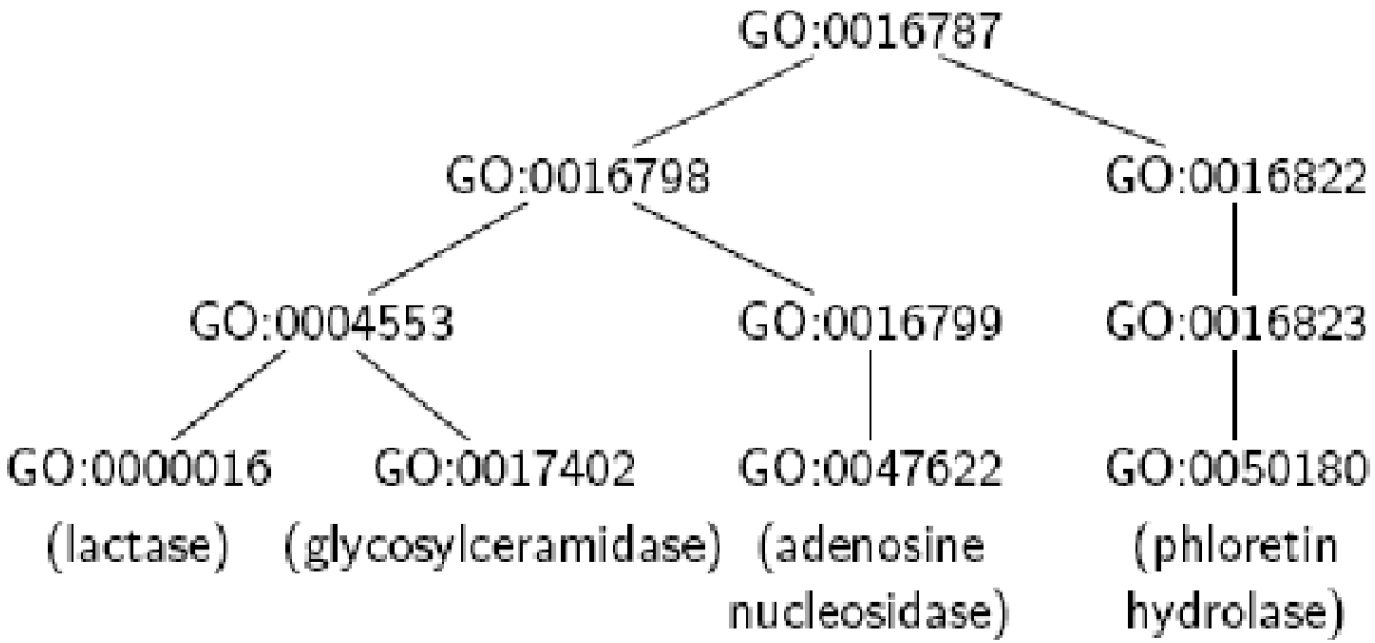
Enzyme Similarity Measures

- Gene Ontology similarity, GO



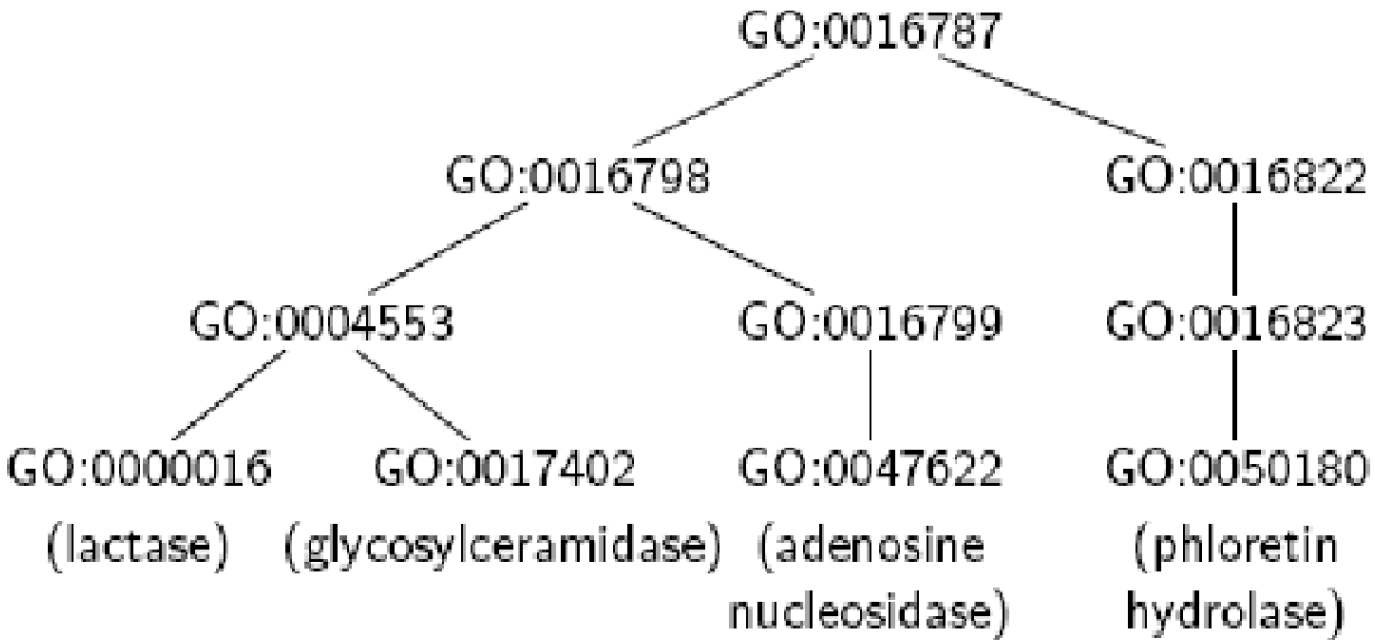
Enzyme Similarity Measures

- Gene Ontology similarity, GO
  - Mapping EC-GO molecular function



Enzyme Similarity Measures

- Gene Ontology similarity, GO
  - Mapping EC-GO molecular function
  - Shortest path distance in the GO molecular function ontology



Pathway Similarity

Similarity of pathways

- Let $P = (R)$, $Q = (S)$ be two pathways, with R, S the respective set of reactions:

$$sim(P,Q) = \frac{1}{|R \cup S|}\Big(\sum_{R \in \mathbf{R} \cap \mathbf{S}} \max_{S \in \mathbf{R} \cap \mathbf{S}} sim(R,S) + \sum_{R \in \mathbf{R} \setminus \mathbf{S}} \max_{S \in \mathbf{S}} sim(R,S) + \sum_{S \in \mathbf{S} \setminus \mathbf{R}} \max_{R \in \mathbf{R}} sim(R,S) \Big)$$

Pathway Similarity

Similarity of reactions

- Let $R = (C, E)$, $S = (D, F)$ be two reactions, with C, D the set of compounds and E, F the sets of enzymes involved

$$sim(R, S) = \frac{1 - \alpha}{|C \cup D|} cpdsim(R, S) + \frac{\alpha}{|E \cup F|} enzsim(R, S)$$

Pathway Similarity

Similarity of reactions

- Let $R = (C, E)$, $S = (D, F)$ be two reactions, with C, D the set of compounds and E, F the sets of enzymes involved

$$sim(R, S) = \frac{1 - \alpha}{|C \cup D|} cpdsim(R, S) + \frac{\alpha}{|E \cup F|} enzsim(R, S)$$

- Reaction similarity is determined by calculating how similar the sets of compounds (*cpdsim*) and enzymes (*enzsim*) are

Pathway Similarity

Similarity of reactions

- Let $R = (C, E)$, $S = (D, F)$ be two reactions, with C, D the set of compounds and E, F the sets of enzymes involved

$$sim(R, S) = \frac{1 - \alpha}{|C \cup D|} cpdsim(R, S) + \frac{\alpha}{|E \cup F|} enzsim(R, S)$$

- Reaction similarity is determined by calculating how similar the sets of compounds (*cpdsim*) and enzymes (*enzsim*) are
- The α parameter establishes relative weight of compounds and enzymes in the similarity of reactions

Pathway Similarity

Similarity of set of enzymes

- Similarity of pair of enzymes calculated using HIER, INFO or GO similarity measure

Pathway Similarity

Similarity of set of enzymes

- Similarity of pair of enzymes calculated using HIER, INFO or GO similarity measure
- Similarity of set of enzymes involved in reactions R, S:

$$enzsim(R,S) = \frac{1}{|E \cup F|}\Big(\sum_{E \in \mathbf{E} \cap \mathbf{F}} \max_{F \in \mathbf{E} \cap \mathbf{F}} sim(E,F) + \sum_{E \in \mathbf{E} \setminus \mathbf{F}} \max_{F \in \mathbf{F}} sim(E,F) + \sum_{F \in \mathbf{F} \setminus \mathbf{E}} \max_{E \in \mathbf{E}} sim(E,F) \Big)$$

Pathway Similarity

Similarity of set of compounds

- Two compounds are either similar ($sim(C, D) = 1$) or dissimilar ($sim(C, D) = 0$)

Pathway Similarity

Similarity of set of compounds

- Two compounds are either similar ($sim(C, D) = 1$) or dissimilar ($sim(C, D) = 0$)
- Similarity of set of compounds involved in reactions R, S:

$$cpdsim(R, S) = \frac{1}{|C \cup D|}\Big(\sum_{C \in \mathbf{C} \cap \mathbf{D}} \max_{D \in \mathbf{C} \cap \mathbf{D}} sim(C, D) + \sum_{C \in \mathbf{C} \setminus \mathbf{D}} \max_{D \in \mathbf{D}} sim(C, D) + \sum_{D \in \mathbf{D} \setminus \mathbf{C}} \max_{C \in \mathbf{C}} sim(C, D) \Big)$$

Reconstruction and validation of phylogenetic relationships

- Create similarity matrix for n species from pathway similarity measure

Reconstruction and validation of phylogenetic relationships

- Create similarity matrix for n species from pathway similarity measure
- Apply clustering algorithm to similarity matrix to obtain phylogenetic tree

Reconstruction and validation of phylogenetic relationships

- Create similarity matrix for n species from pathway similarity measure
- Apply clustering algorithm to similarity matrix to obtain phylogenetic tree
- Compare obtained phylogenetic trees versus NCBI taxonomy using different tools

Reconstruction and validation of phylogenetic relationships

- Create similarity matrix for n species from pathway similarity measure
- Apply clustering algorithm to similarity matrix to obtain phylogenetic tree
- Compare obtained phylogenetic trees versus NCBI taxonomy using different tools

Reconstruction and validation of phylogenetic relationships

- Create similarity matrix for n species from pathway similarity measure
- Apply clustering algorithm to similarity matrix to obtain phylogenetic tree
- Compare obtained phylogenetic trees versus NCBI taxonomy using different tools

Clemente et al (2005) used standard Average Link Hierarchical (ALH) clustering (specifically, Perl `Bio::Tree::DistanceFactory` implementation of UPGMA)

FER clustering

Fuzzy Equivalence Relations (FER) clustering was introduced by Zadeh (1971), and applied in phylogenetic resconstruction by Luo et al (1995).

FER clustering

Fuzzy Equivalence Relations (FER) clustering was introduced by Zadeh (1971), and applied in phylogenetic resconstruction by Luo et al (1995).

- Compute the fuzzy equivalence relation E generated by the similarity matrix

FER clustering

Fuzzy Equivalence Relations (FER) clustering was introduced by Zadeh (1971), and applied in phylogenetic resconstruction by Luo et al (1995).

- Compute the fuzzy equivalence relation E generated by the similarity matrix
- For each t in E, take the partition induced by the t-cut equivalence relation, obtained by replacing in E every entry $< t$ by 0 and every entry $\geqslant t$ by 1

FER clustering

Fuzzy Equivalence Relations (FER) clustering was introduced by Zadeh (1971), and applied in phylogenetic resconstruction by Luo et al (1995).

- Compute the fuzzy equivalence relation E generated by the similarity matrix
- For each t in E, take the partition induced by the t-cut equivalence relation, obtained by replacing in E every entry $< t$ by 0 and every entry $\geqslant t$ by 1
- These partitions, together with the hierarchy induced by the increasing order of t, yield a classification tree.

Phylogenetic Relationships

FER clustering: Example

Compute the similarity of the Glycolysis pathways of the organisms

AFU	*A.fulgidus*
MJA	*M.jannaschii*
MGE	*M.genitalum*
HIN	*H.influenzae*
MTU	*M.tuberculosis*
ECO	*E.coli*

using GO similarity and $\alpha = 0.8$.

The similarity matrix:

	MGE	HIN	MTU	MJA	ECO	AFU
MGE	1.00	0.33	0.07	0.02	0.17	0.22
HIN	0.33	1.00	0.33	0.32	0.34	0.27
MTU	0.07	0.33	1.00	0.09	0.20	0.20
MJA	0.02	0.32	0.09	1.00	0.18	0.24
ECO	0.17	0.34	0.20	0.18	1.00	0.32
AFU	0.22	0.27	0.20	0.24	0.32	1.00

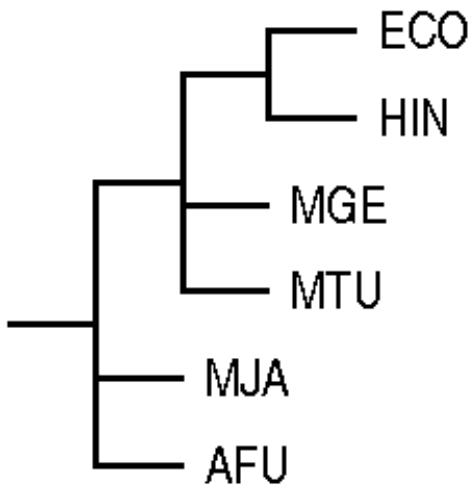
The fuzzy equivalence relation it generates (through max-min transitive closure) is:

	MGE	HIN	MTU	MJA	ECO	AFU
MGE	1.00	0.33	0.33	0.32	0.33	0.32
HIN	0.33	1.00	0.33	0.32	0.34	0.32
MTU	0.33	0.33	1.00	0.32	0.33	0.32
MJA	0.32	0.32	0.32	1.00	0.32	0.32
ECO	0.33	0.34	0.33	0.32	1.00	0.32
AFU	0.32	0.32	0.32	0.32	0.32	1.00

The hierarchy of partitions defined by the t-cuts is:

t	Partition corresponding to the t-cut
1.00	$\{MGE\}\{HIN\}\{MTU\}\{MJA\}\{ECO\}\{AFU\}$
0.34	$\{HIN, ECO\}\{MGE\}\{MTU\}\{MJA\}\{AFU\}$
0.33	$\{MGE, HIN, MTU, ECO\}\{MJA\}\{AFU\}$
0.32	$\{MGE, HIN, MTU, MJA, ECO, AFU\}$

This hierarchical clustering yields the classification tree



very close to the NCBI taxonomy.

Results

To compare the performance of FER clustering and ALH clustering:

- We have computed the similarities of the Glycolysis pathways of a model set of 16 organisms, downloaded from the KEGG server

Results

To compare the performance of FER clustering and ALH clustering:

- We have computed the similarities of the Glycolysis pathways of a model set of 16 organisms, downloaded from the KEGG server
- We have computed the phylogenetic trees generated by these similarities by using both the ALH and the FER clustering

Results

To compare the performance of FER clustering and ALH clustering:

- We have computed the similarities of the Glycolysis pathways of a model set of 16 organisms, downloaded from the KEGG server
- We have computed the phylogenetic trees generated by these similarities by using both the ALH and the FER clustering
- We have compared the resulting trees to the NCBI taxonomy of the 16 organisms.

Results: cousins

The cousins tool measures the similarity of phylogenetic trees at the ground level (roughly, it compares sets of triples consisting of pairs of leaves and their distance, up to a certain distance:

Results: cousins

The cousins tool measures the similarity of phylogenetic trees at the ground level (roughly, it compares sets of triples consisting of pairs of leaves and their distance, up to a certain distance: here, up to second cousins).

Results: cousins

The cousins tool measures the similarity of phylogenetic trees at the ground level (roughly, it compares sets of triples consisting of pairs of leaves and their distance, up to a certain distance: here, up to second cousins).

- GO similarity: FER outperforms ALH always except for $\alpha = 0.7$ and $\alpha = 1$.

Results: cousins

The cousins tool measures the similarity of phylogenetic trees at the ground level (roughly, it compares sets of triples consisting of pairs of leaves and their distance, up to a certain distance: here, up to second cousins).

- GO similarity: FER outperforms ALH always except for $\alpha = 0.7$ and $\alpha = 1$.
- HIER similarity: FER outperforms ALH for $\alpha \leqslant 0.6$.

Results: cousins

The cousins tool measures the similarity of phylogenetic trees at the ground level (roughly, it compares sets of triples consisting of pairs of leaves and their distance, up to a certain distance: here, up to second cousins).

- GO similarity: FER outperforms ALH always except for $\alpha = 0.7$ and $\alpha = 1$.
- HIER similarity: FER outperforms ALH for $\alpha \leqslant 0.6$.
- INFO similarity: FER outperforms ALH for $\alpha \leqslant 0.4$.

Results: cousins

The cousins tool measures the similarity of phylogenetic trees at the ground level (roughly, it compares sets of triples consisting of pairs of leaves and their distance, up to a certain distance: here, up to second cousins).

- GO similarity: FER outperforms ALH always except for $\alpha = 0.7$ and $\alpha = 1$.
- HIER similarity: FER outperforms ALH for $\alpha \leqslant 0.6$.
- INFO similarity: FER outperforms ALH for $\alpha \leqslant 0.4$.

Probably because INFO is more "fine grained"

Results: F-measure

The F-measure measures the similarity of phylogenetic trees at first clustering level.

Results: F-measure

The F-measure measures the similarity of phylogenetic trees at first clustering level.

FER always outperforms ALH:

F-measure for all ALH-trees is 0.88

F-measure for all FER-trees is 0.92

Conclusions

- We have recalled a new measure for pathway similarity

Conclusions

- We have recalled a new measure for pathway similarity
  - Based on enzyme and compound similarity

Conclusions

- We have recalled a new measure for pathway similarity
  - Based on enzyme and compound similarity
  - Time quadratic in number of compounds, enzymes and reactions

Conclusions

- We have recalled a new measure for pathway similarity
  - Based on enzyme and compound similarity
  - Time quadratic in number of compounds, enzymes and reactions
  - Outperform previous best measure

Conclusions

- We have recalled a new measure for pathway similarity
  - Based on enzyme and compound similarity
  - Time quadratic in number of compounds, enzymes and reactions
  - Outperform previous best measure
- Fuzzy clustering

Conclusions

- We have recalled a new measure for pathway similarity
  - Based on enzyme and compound similarity
  - Time quadratic in number of compounds, enzymes and reactions
  - Outperform previous best measure
- Fuzzy clustering
  - A test on 16 organisms shows FER outperforms ALH

Conclusions

- We have recalled a new measure for pathway similarity
  - Based on enzyme and compound similarity
  - Time quadratic in number of compounds, enzymes and reactions
  - Outperform previous best measure
- Fuzzy clustering
  - A test on 16 organisms shows FER outperforms ALH
  - Similar results for a larger (73) set of organisms (unpublished)

Conclusions

- We have recalled a new measure for pathway similarity
  - Based on enzyme and compound similarity
  - Time quadratic in number of compounds, enzymes and reactions
  - Outperform previous best measure
- Fuzzy clustering
  - A test on 16 organisms shows FER outperforms ALH
  - Similar results for a larger (73) set of organisms (unpublished)
  - But we are still far from obtaining a fully correct taxonomy

Future Work

- Pathway similarity

Future Work

- Pathway similarity
  - Compound similarity measure

Future Work

- Pathway similarity
  - Compound similarity measure
  - Further study influence of α parameter

Future Work

- Pathway similarity
  - Compound similarity measure
  - Further study influence of α parameter
- Fuzzy clustering

Future Work

- Pathway similarity
  - Compound similarity measure
  - Further study influence of α parameter
- Fuzzy clustering
  - Fuzzy c-means hierarchical clustering

Future Work

- Pathway similarity
  - Compound similarity measure
  - Further study influence of α parameter
- Fuzzy clustering
  - Fuzzy c-means hierarchical clustering
- Further experiments

Future Work

- Pathway similarity
  - Compound similarity measure
  - Further study influence of α parameter
- Fuzzy clustering
  - Fuzzy c-means hierarchical clustering
- Further experiments
  - Extend results to other pathways

Future Work

- Pathway similarity
  - Compound similarity measure
  - Further study influence of α parameter
- Fuzzy clustering
  - Fuzzy c-means hierarchical clustering
- Further experiments
  - Extend results to other pathways
  - Larger sets of organisms

The organisms

Organisms studied, classified by domain (A: Archaea, B: Bacteria, E: Eukaryota), together with their identifier in the NCBI taxonomy

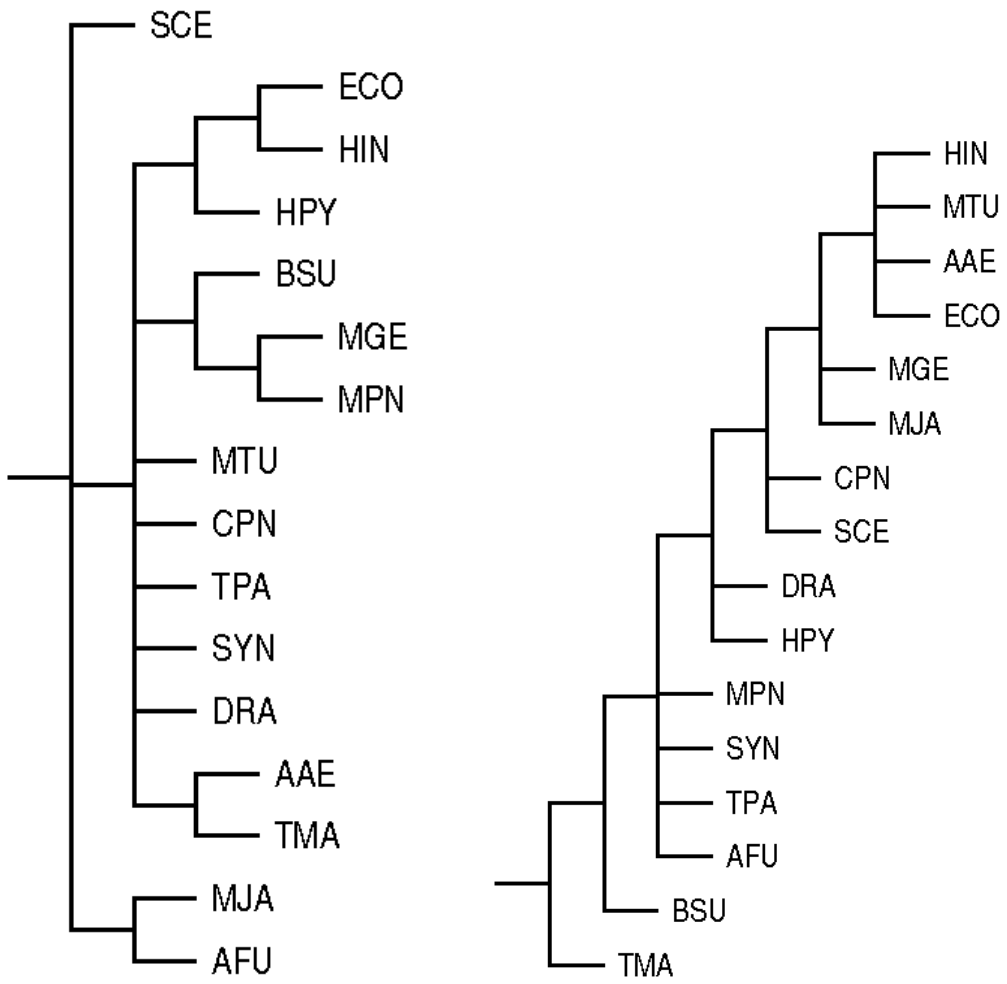
AFU	*A.fulgidus*	A	224325
MJA	*M.jannaschii*	A	243232
CPN	*C.pneumoniae*	B	115713
MGE	*M.genitalum*	B	243273
MPN	*M.pneumoniae*	B	272634
HIN	*H.influenzae*	B	71421
SYN	*Synechocystis*	B	1148
DRA	*D.radiodurans*	B	243230
MTU	*M.tuberculosis*	B	83332
TPA	*T.pallidum*	B	243276
BSU	*B.subtilis*	B	224308
AAE	*A.aeolicus*	B	224324
TMA	*T.maritima*	B	243274
ECO	*E.coli*	B	83333
HPY	*H.pylori*	B	85962
SCE	*S.cerevisiae*	E	4932

F-measure

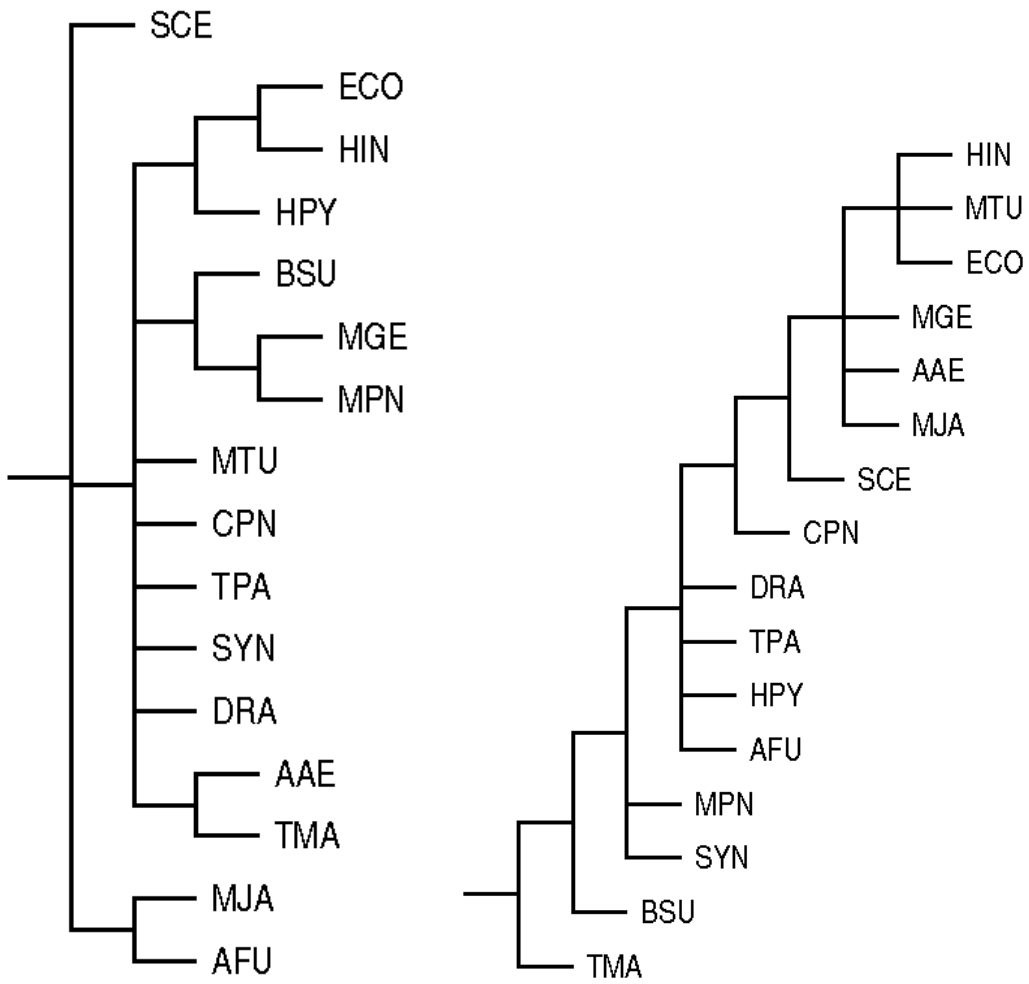
F-measure combines precision and recall
It is defined as

$$F(C) = \sum_{t \in T} \frac{N_t}{N} \max_{C_k \in C} \frac{2P_{tk}R_{tk}}{(P_{tk} + R_{tk})}$$

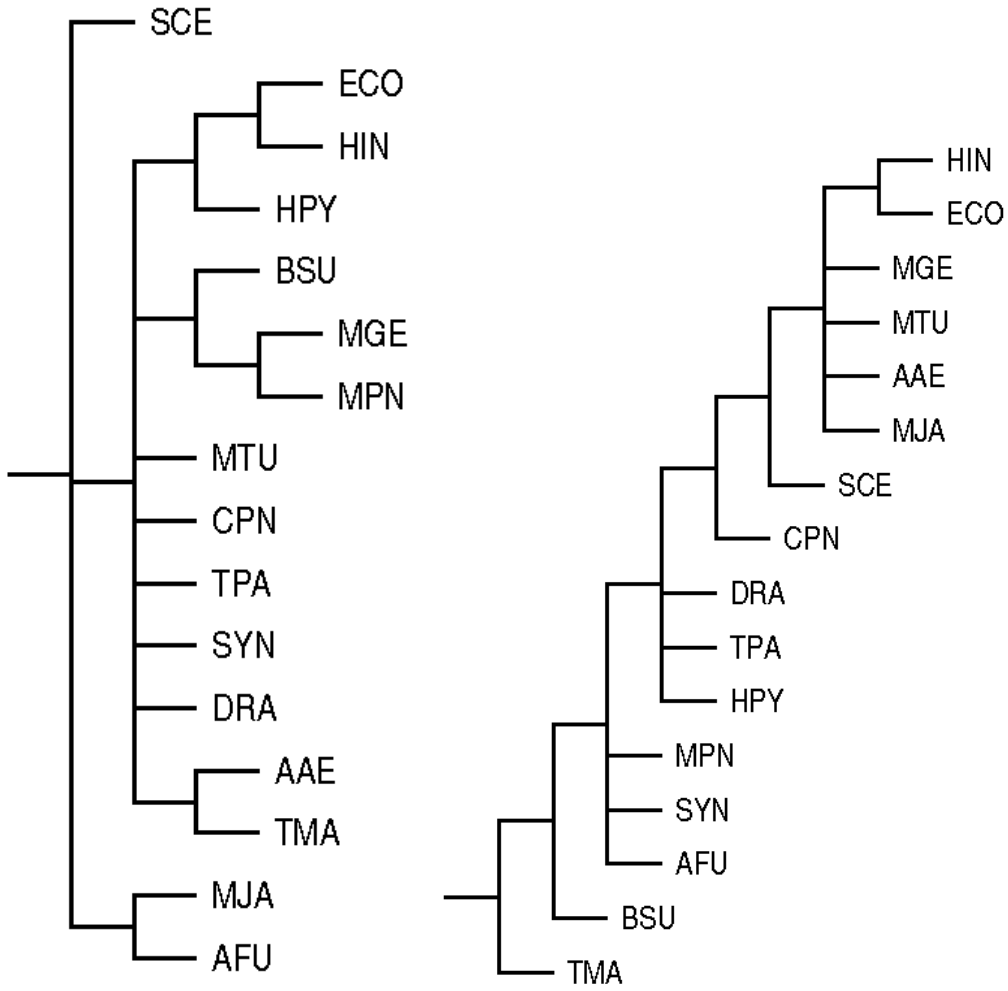
where N_{tk} is the number of elements of class t within cluster C_k, precision is defined as $P_{tk} = N_{tk}/N_k$, and recall is defined as $R_{tk} = N_{tk}/N_t$.



NCBI taxonomy (left) and best tree obtained with FER clustering and GO similarity ($\alpha = 0.2$) (right)



NCBI taxonomy (left) and best tree obtained with FER clustering and HIER similarity ($\alpha = 0.2$) (right)



NCBI taxonomy (left) and best tree obtained with FER clustering and INFO similarity ($\alpha = 0.1$) (right)