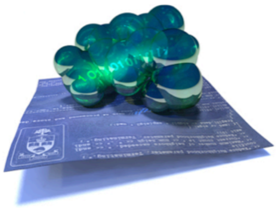
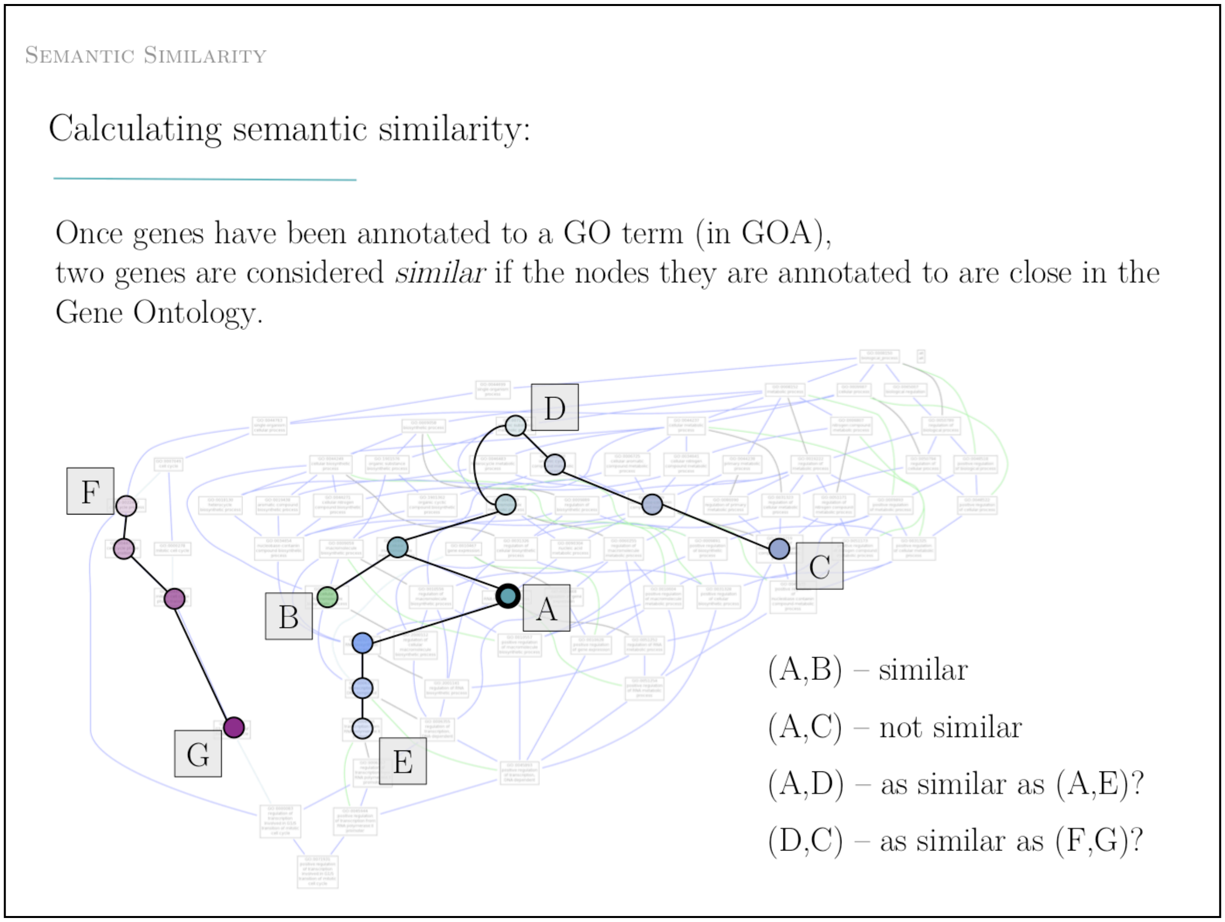
A
Bioinformatics
Course

Semantic Similarity



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto



Semantic Similarity

Calculating semantic similarity:

Once genes have been annotated to a GO term (in GOA),
two genes are considered *similar* if the nodes they are annotated to are close in the Gene Ontology.

(A,B) – similar

(A,C) – not similar

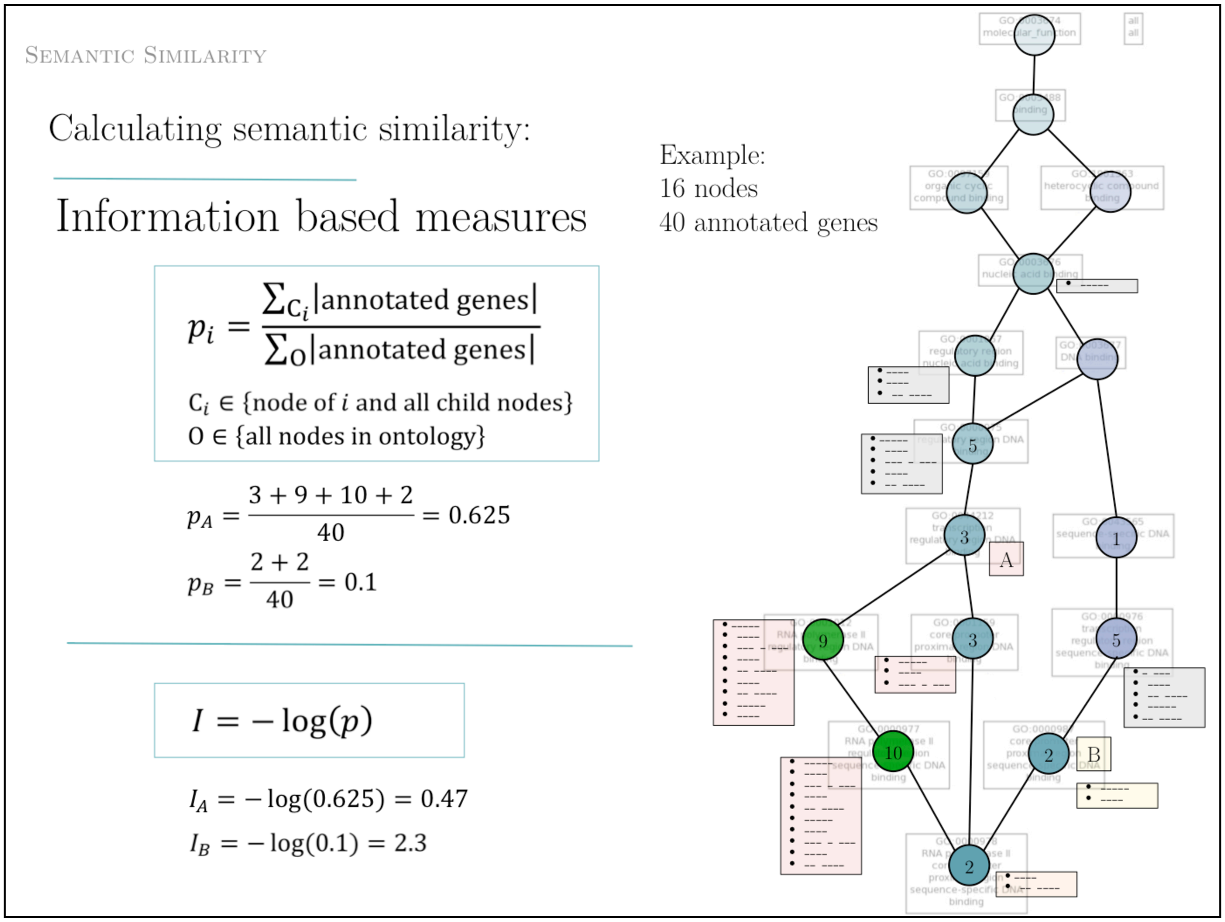
(A,D) – as similar as (A,E)?

(D,C) – as similar as (F,G)?

It is easy to see that nodes A and B are similar since they share the same parent, whereas A and C are not – their lowest common ancestor is three steps away. However matters are not so clear–cut if we are asking about relationships at different levels of the tree (as between (A,D) and (A,E)), or in parts of the tree that are broad (as with (D,C)) as compared to parts that are narrow as with (F,G).

A unifying concept is to ask how *specific* each parent is, i.e. which portion of the tree it has beneath it. This can be expressed as *information* (in the information-theoretical sense).

Nb.: The GO is not actually a tree but a DAG (Directed Acyclic Graph). This means that nodes can have more than one parent but the connections are directed such that there are no cycles in the graph. A tree *is* a DAG, but not all DAGs are trees. This is necessary, because many concepts can be valid children of more than one parent concept, but this arrangement also creates difficulties since there may be multiple pathwaysto higher nodes and to the root.



Semantic Similarity

Calculating semantic similarity:

Information based measures

$$p_i = \frac{\sum_{C_i} |\text{annotated genes}|}{\sum_{O} |\text{annotated genes}|}$$

$C_i \in$ {node of i and all child nodes}

$O \in$ {all nodes in ontology}

$$p_A = \frac{3+9+10+2}{40} = 0.625$$

$$p_B = \frac{2+2}{40} = 0.1$$

$$I = -\log(p)$$

$$I_A = -\log(0.625) = 0.47$$

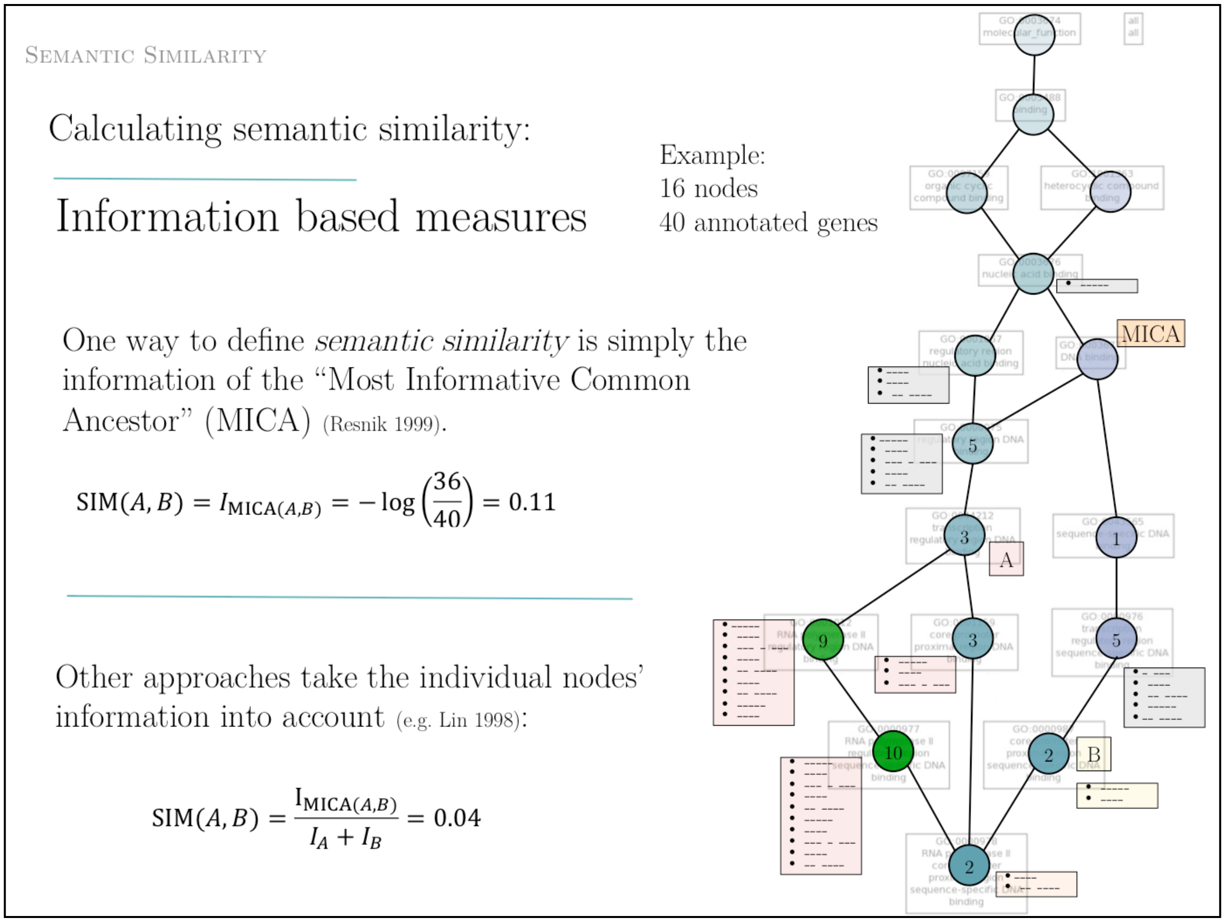
$$I_B = -\log(0.1) = 2.3$$

Example:
16 nodes
40 annotated genes

GO:0003674 molecular_function
all all
binding
organic cyclic compound binding
heterocyclic compound binding
nucleic acid binding
regulatory region nucleic acid binding
DNA binding
regulatory region DNA binding
5
transcription regulatory region DNA binding
3
A
sequence-specific DNA binding
1
9
3
5
10
2
B
2

The "probability" p_{A} of a gene A is the number of genes annotated to it and its children, divided by the number of genes in the tree.

"Information" is the negative log of the probability.



Calculating semantic similarity:

Information based measures

Example:
16 nodes
40 annotated genes

One way to define *semantic similarity* is simply the information of the "Most Informative Common Ancestor" (MICA) (Resnik 1999).

$$\mathrm{SIM}(A,B) = I_{\mathrm{MICA}(A,B)} = -\log\left(\frac{36}{40}\right) = 0.11$$

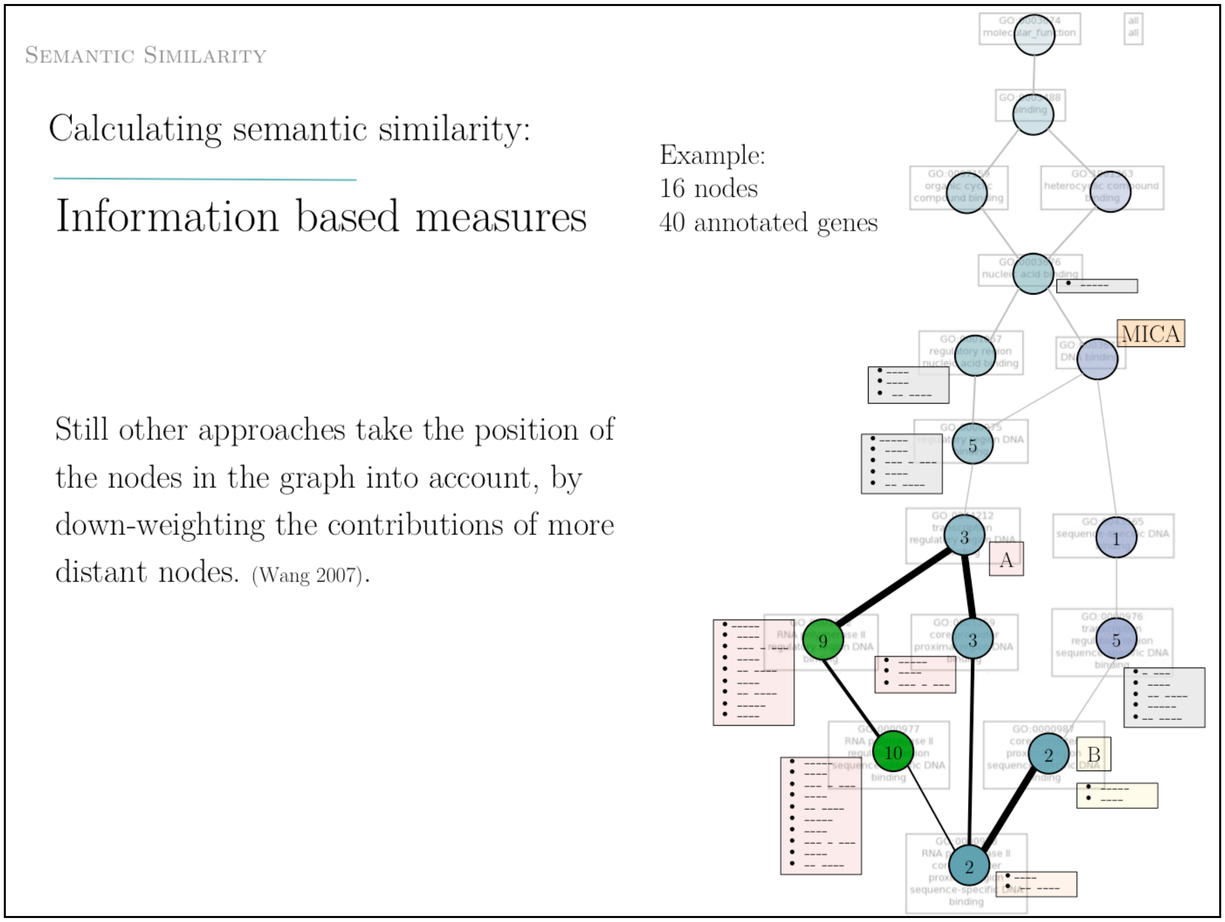
Other approaches take the individual nodes' information into account (e.g. Lin 1998):

$$\mathrm{SIM}(A,B) = \frac{\mathrm{I}_{\mathrm{MICA}(A,B)}}{I_A + I_B} = 0.04$$

GO:0003674 molecular_function
all all
GO:0005488 binding
GO:0097159 organic cyclic compound binding
GO:1901363 heterocyclic compound binding
GO:0003676 nucleic acid binding
GO:0001067 regulatory region nucleic acid binding
GO:0003677 DNA binding
MICA
GO:0000975 regulatory region DNA binding
5
GO:0044212 transcription regulatory region DNA binding
3
A
GO:0043565 sequence-specific DNA binding
1
GO:0001012 RNA polymerase II regulatory region DNA binding
9
GO:0001159 core promoter proximal region DNA binding
3
GO:0000976 transcription regulatory region sequence-specific DNA binding
5
GO:0000977 RNA polymerase II regulatory region sequence-specific DNA binding
10
GO:0000987 core promoter proximal region sequence-specific DNA binding
2
B
GO:0000978 RNA polymerase II core promoter proximal region sequence-specific DNA binding
2

Resnik, P (1995). Semantic similarity in a taxonomy: An Information-Based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98), Jude W. Shavlik (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 296-304



Semantic Similarity

Calculating semantic similarity:

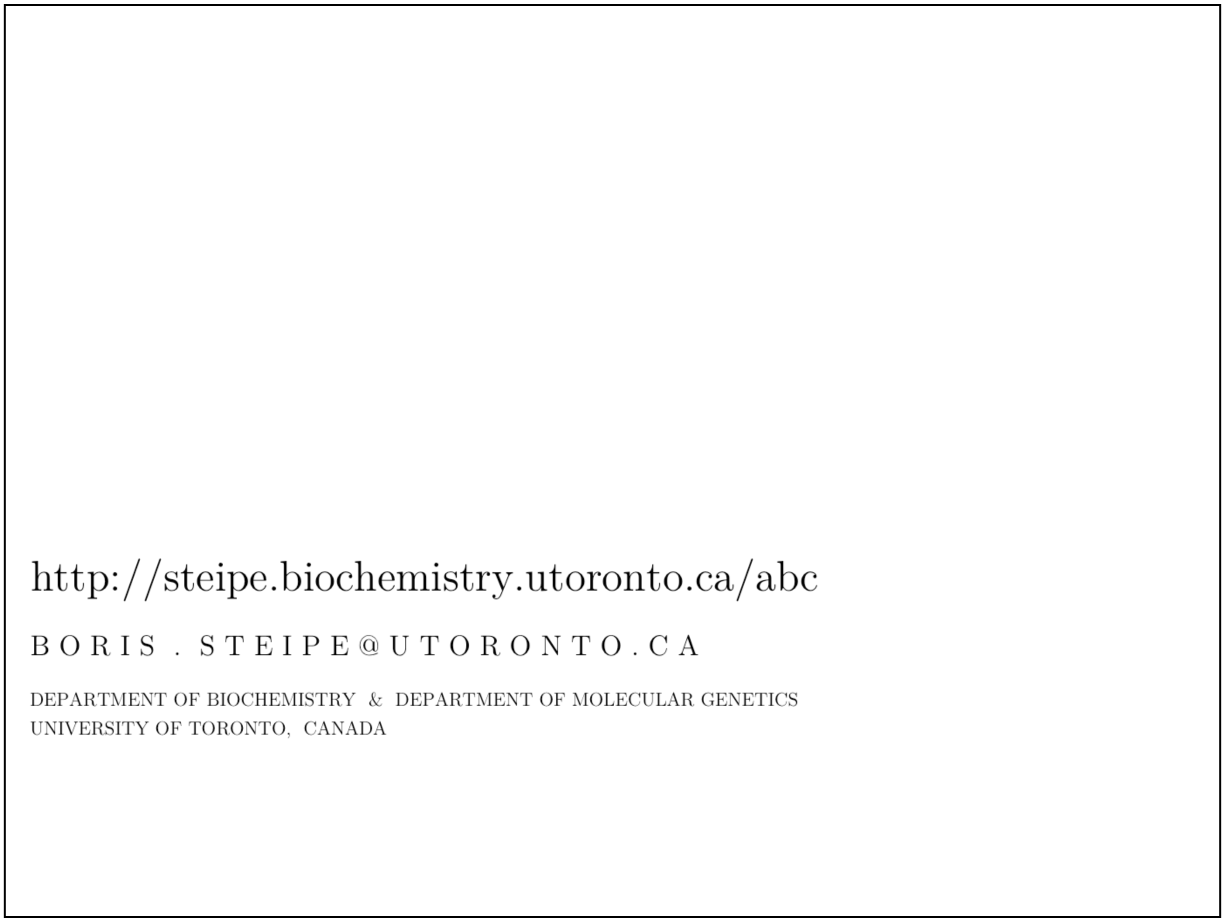
Information based measures

Example:
16 nodes
40 annotated genes

Still other approaches take the position of the nodes in the graph into account, by down-weighting the contributions of more distant nodes. (Wang 2007).

molecular_function
all
binding
organic cyclic compound binding
heterocyclic compound binding
nucleic acid binding
MICA
regulatory region nucleic acid binding
DNA binding
5
3
1
A
9
3
5
10
2
B
2
GO:0000977
RNA polymerase II
GO:000098
core promoter proximal region sequence-specific DNA binding

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. (2007). A new method to measure the semantic similarity of gO terms. Bioinformatics (Oxford, England) 23, 1274–1281.



http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA