Gene ontology and functional analysis

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start

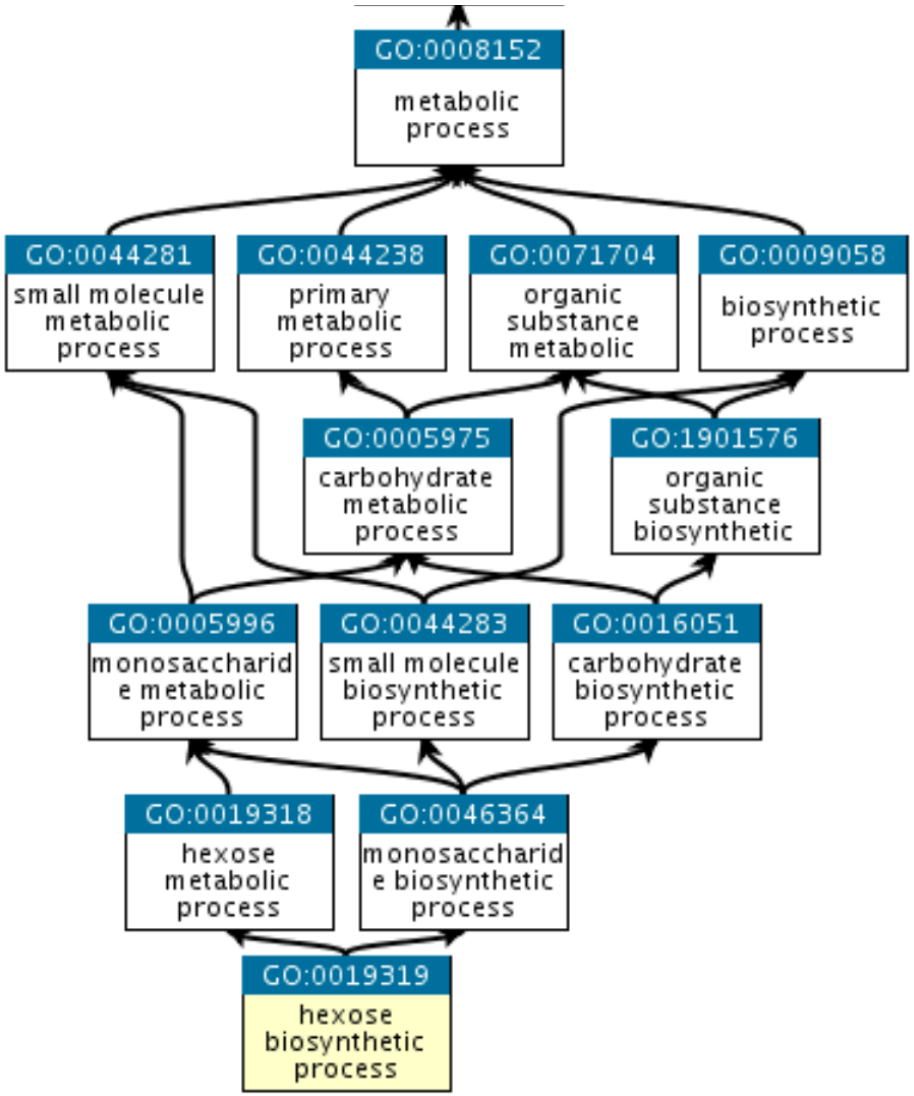
Overview

- Bioinformatics deals with a large amount of measurements
  - these measurements need to be transformed into knowledge,
  - they need to be merged with current knowledge bases,
- gene ontology (GO)
  - describes our knowledge about genes and their products,
  - ontology = a formal specification of concepts and their relationships,
  - other relevant knowledge-bases: BioGrid, KEGG, Disease ontology, ...
- common ways to use GO
  - functional enrichment analysis
    - biological interpretation of gene/protein -omics lists,
    - in here, focus on gene expression data introduced before,
  - automated function prediction (AFP)
    - computationally predict gene/protein function,
    - commonly used as a hypothesis for further biological validation.

Gene ontology (GO)

- Ontology
  - a formal specification of concepts and relationships between them,
  - consists of individuals, classes, attributes, relations, axioms, rules, ...
- gene ontology
  - the world's largest source of information on the functions of genes,
  - over 700,000 experimentally supported annotations,
  - taken from 150,000 published papers,
  - thanks to additional inference over 6 million functional annotations,
  - a diverse set of organisms (animal, plant, microbial genomes).

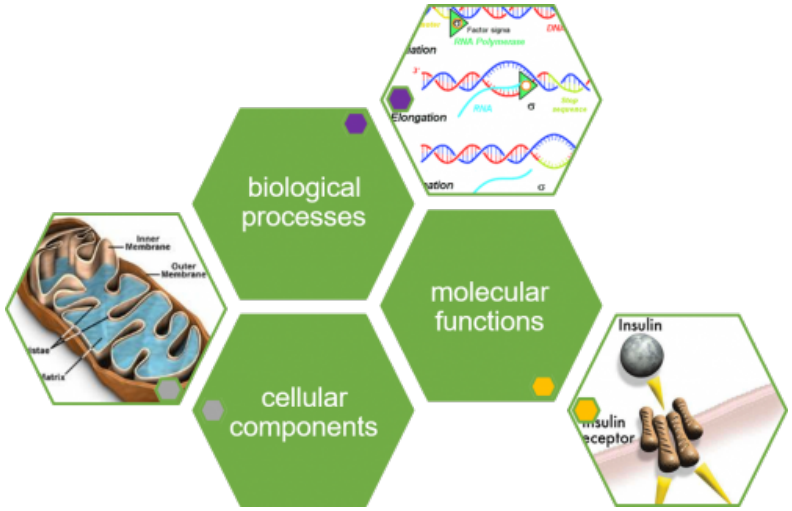
Gene ontology relationships structured as a DAG



http://geneontology.org/docs/ontology-documentation/.

GO distinguishes three aspects (three ontologies)

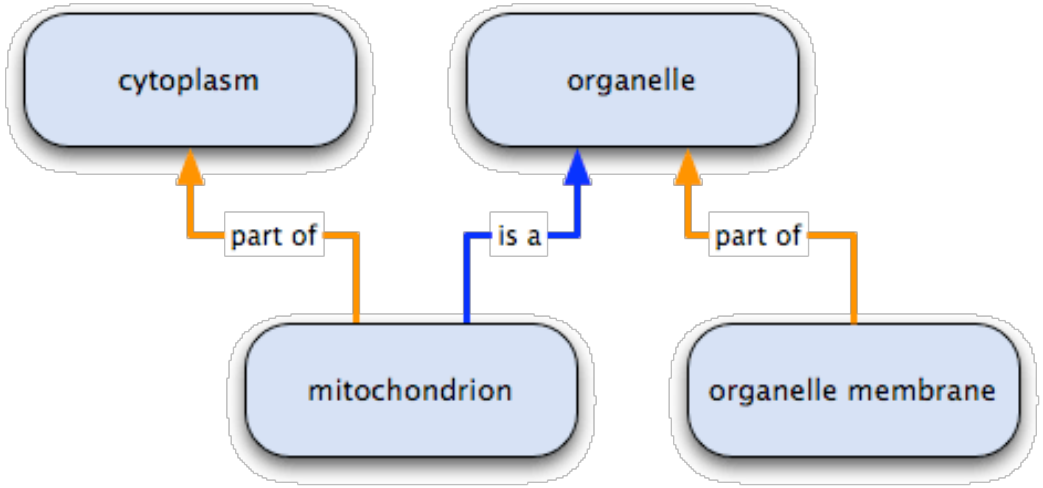
- The ontology covers three domains
  - **molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis,
  - **biological process**, operations or sets of molecular events with a defined beginning and end, such as cell division, metabolic process.
  - **cellular component**, the parts of a cell or its extracellular environment, such as nucleus, ribosome, mitochondrion.



https://www.ebi.ac.uk

GO terms and relationships between them

- The ontologies structured as a directed acyclic graph (DAG)
  - $G = \langle V, E \rangle$, $V = \{t | \text{GO terms}\}$, $E = \{(t, u) | t \in V \text{ and } u \in V\}$,
- types of relationships between GO terms (the graph edges)
  - **is a** subtype relation,
  - **part of** part-whole relation,
  - **regulates** control relation (non-transitive).



http://geneontology.org/docs/ontology-relations/

GO annotations

- In GO we have to distinguish
  - the taxonomy itself, which is a set of terms with their precise definitions and defined relationships between them,
  - the associations between gene products and GO terms (**GO annotations** considered a part of GO too),
  - an example of such a link below (millions of them exist).

Gene	Chr	GO term	Evidence	Inferred from	Reference
Chst15 carbohydrate sulfotransferase 15	7	hexose biosynthetic process	IBA	PTN000404454	J:265628 [PMID:21873635]

Functional enrichment analysis

- Remember gene profiling and differential gene expression (DGE)
  - it reports a list of differentially expressed genes/transcripts,
  - or a ranking of genes with respect to a test statistics,
- potential problems with DGE
  - no genes may be significantly altered $\rightarrow$ no result,
  - many significantly altered genes $\rightarrow$ hard to interpret,
  - multi-functional genes $\rightarrow$ hard to interpret,
  - caused by noise, small samples, small effects (differences between groups),
- **functional enrichment analysis**
  - examines differential expression in terms of well-defined gene sets,
  - hits cumulative effects from many slightly altered biologically related genes,
- the well-defined gene sets
  - in this lecture gene ontology terms/classes,
  - in general, the sets come from any prior biological knowledge.

The Molecular Signatures Database (MSigDB)

- Manually curated database of gene sets
  - in 2021, 32,284 gene sets divided into 9 major collections.

Collections

The MSigDB gene sets are divided into 9 major collections:

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

https://www.gsea-msigdb.org/

Functional enrichment analysis

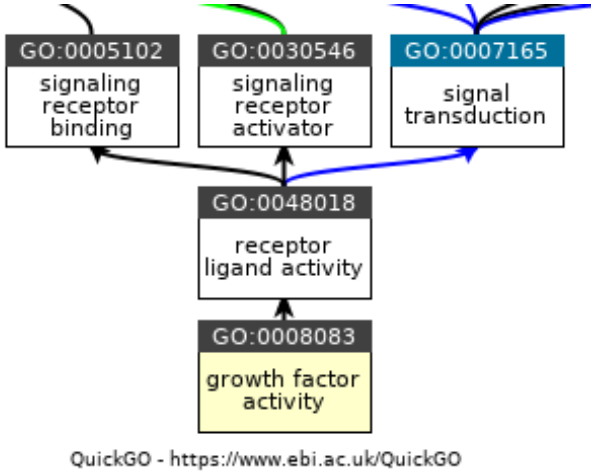
- Over representation analysis
  - identify a set of differentially expressed genes D,
  - take a pre-defined set of genes S,
  - count frequencies in a 2x2 contingency table,
  - do a test of independence
    (chi-squared test, hypergeometric test (Fisher's exact test)).

$$X^2 = \sum_{s \in \{S, S^c\}} \sum_{d \in \{D, D^c\}} \frac{\left(m_{sd} - \frac{m_s m_d}{m}\right)^2}{\frac{m_s m_d}{m}} < \chi^2_{df=1,\alpha}$$

	Differentially expressed gene	Non-differentially expressed gene	Total
In gene set	m_{SD}	m_{SD^c}	m_S
Not in gene set	m_{S^cD}	$m_{S^cD^c}$	m_{S^c}
Total	m_D	m_{D^c}	m

Over representation analysis – example

gname	pvalue	padj	in D
ERRFI1	1.16E-24	2.94E-20	1
LIF	2.43E-15	3.09E-11	1
DUSP1	1.56E-14	1.32E-10	1
...			
NOP56	1.99E-05	0.009009	1
DDX31	2.14E-05	0.009537	1
NUDCD1	2.34E-05	0.01025	0
...			
PSMA6P1	1	1	0
TMSB4Y	1	1	0
BCORP1	1	1	0

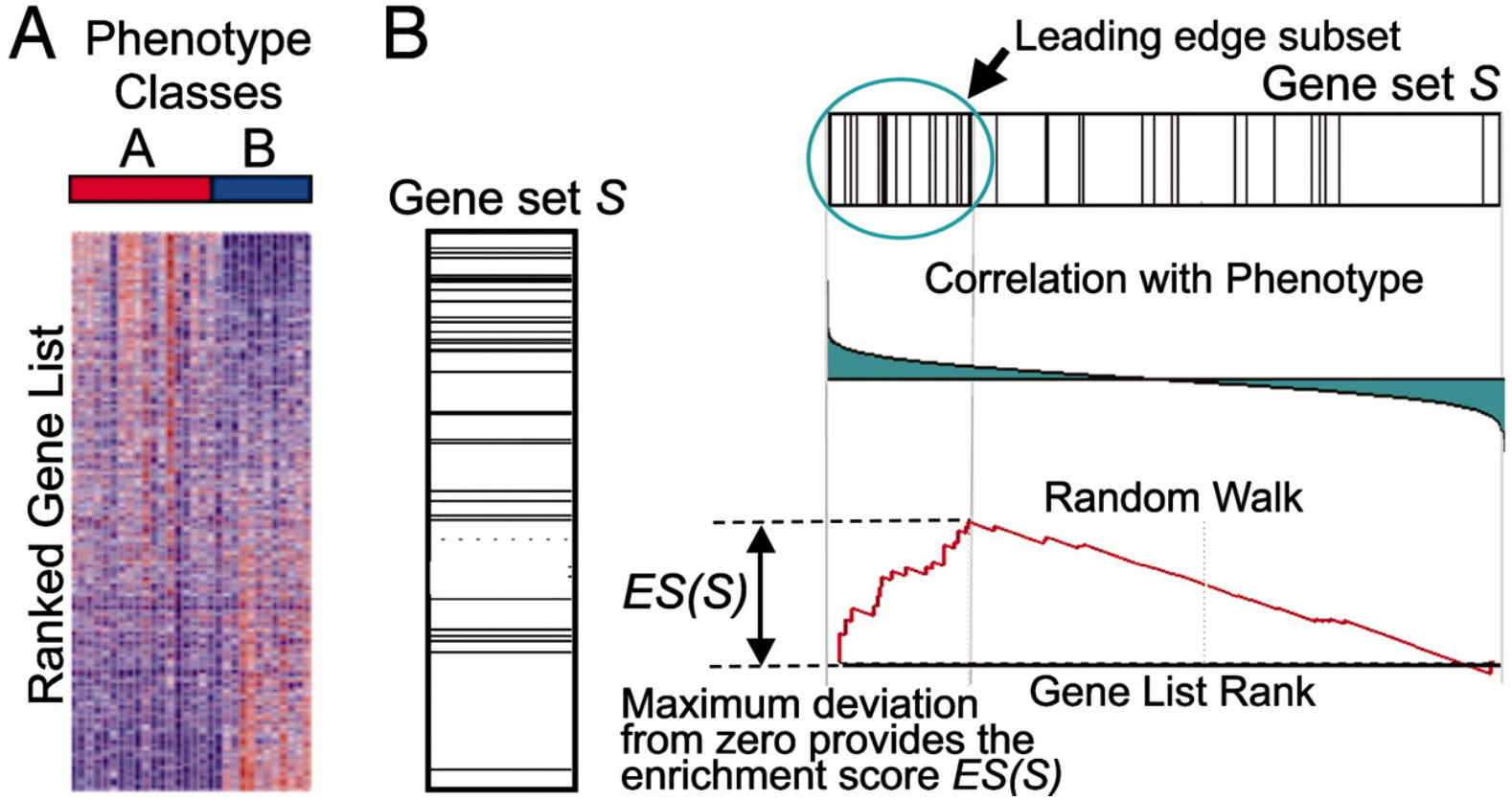


	Differentially expressed gene	Non-differentially expressed gene	Total
Growth factor activity	10	190	200
Without this function	40	9760	9800
Total	50	9950	10000

- Fisher's Exact Test:
  - p-value = 4.19e-08 → growth factor activity enriched in our list.

Functional enrichment analysis

- Gene set enrichment analysis (GSEA)
  - deal with gene scores, calculate an enrichment score for a gene set S.



Subramanian et al.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.

GSEA, details

- The basic idea
  - given a gene set S (e.g., from GO),
  - and a sorted gene list L (e.g., outcome of DGE), r_j is the jth gene score,
  - goal is to find out whether S is randomly distributed in L or stays focused at one of the ends,
- the enrichment score (ES) could be calculated for any position i in L
  - we search for the position in L that maximizes the score,

$$ES(S) = \max_i ES(S,i) = |P_{hit}(S,i) - P_{miss}(S,i)|$$

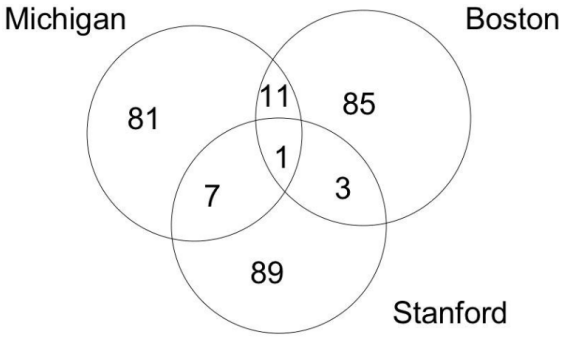
$$P_{hit}(S,i) = \sum_{g_j \in S, j \leq i} \frac{|r_j|^p}{N_R} \qquad P_{miss}(S,i) = \sum_{g_j \notin S, j \leq i} \frac{1}{m - m_S}$$

where p is a parameter, default 1, $N_R = \sum_{g_j \in S} |r_j|^p$

  - significance of $ES(S)$ tested against a large number of random gene sets with size m_S.

GSEA, a case study

- Lung cancer studies in Michigan, Boston and Stanford
  - no genes were significantly associated with cancer outcome,
  - small overlap between top 100 genes found in the studies (S_M, S_B, S_S),
- GSEA outcome for the same data
  - S_B significantly enriched in Michigan data and vice versa,
  - 8 significant gene sets in Boston, 11 in Michigan data, large overlap.



Gene set	FDR
Data set: Lung cancer outcome, Michigan study	
Enriched in poor outcome	
Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

Gene set	FDR
Data set: Lung cancer outcome, Boston study	
Enriched in poor outcome	
Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216

Subramanian et al.: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.

Automated function prediction (AFP)

- Automated function prediction
  - take a protein sequence and predict its function in terms of GO annotations,
  - motivated by a huge gap between the explosive increase of NGS protein sequences and limited number of experimental GO annotations,
  - similar tasks for different input data and annotations exist,
- challenges in AFP
  - many labels per protein $\rightarrow$ multi-label classification problem,
  - structured ontology $\rightarrow$ follow true path rule,
    - annotation at a node must propagate to all ancestor nodes,
  - large variation in the number of GO terms per protein,
- Critical Assessment of Functional Annotation (CAFA) challenge
  - similar structure and goals as CASP mentioned earlier,
  - started in 2010, now CAFA4.

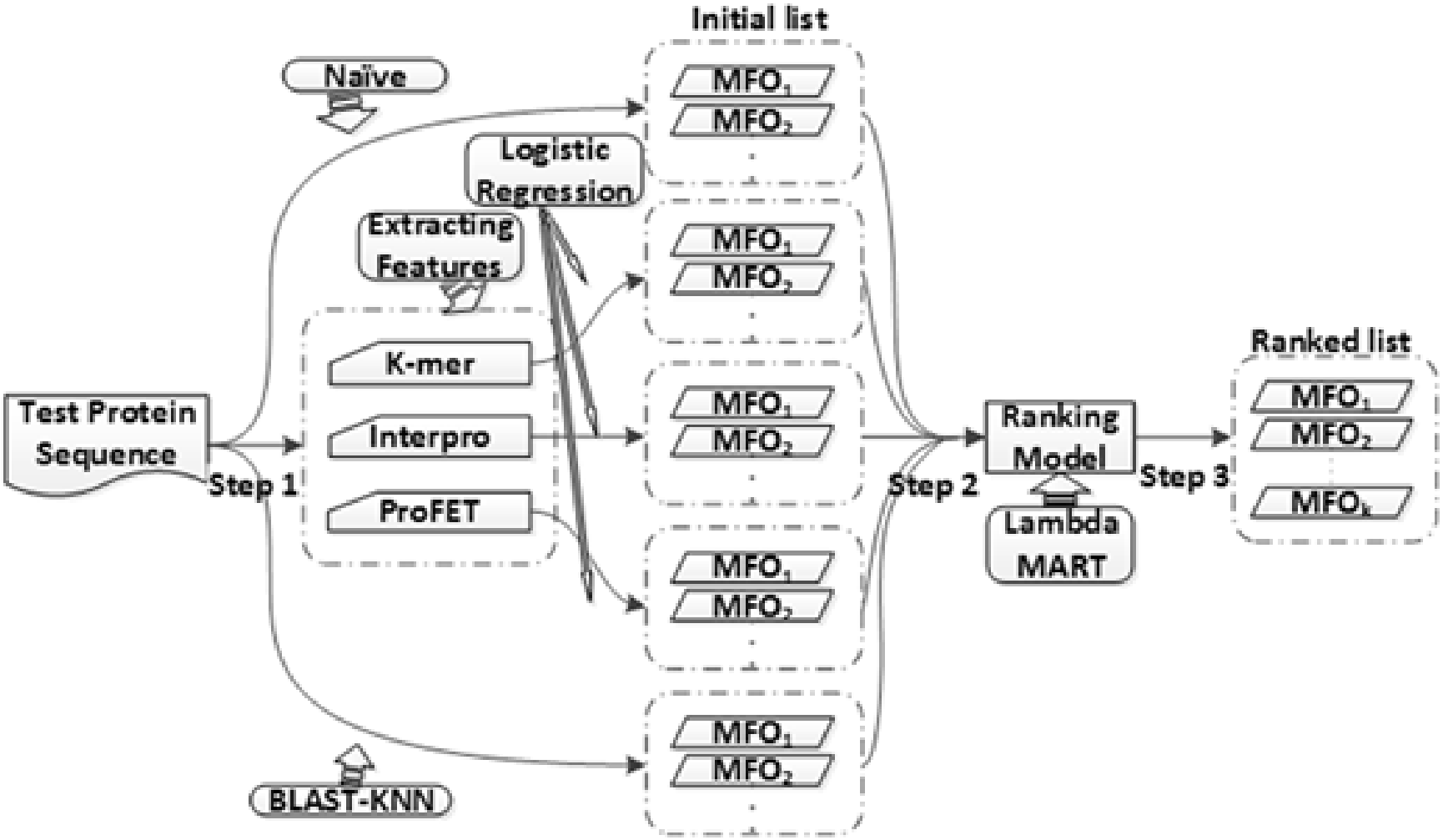
BLAST-KNN – a straightforward AFP solution

- k-nearest neighbor using BLAST results (BLAST-KNN)
  - for given protein P run BLAST to identify a set H of similar proteins,
  - for each GO term G calculate score that P is with G

$$S(G,P) = \frac{\sum_{p \in H} I(G,p) B(P,p)}{\sum_{p \in H} B(P,p)}$$

  - $I(G,p)$ is 0/1 ground-truth indicator whether p is annotated by G,
  - $B(P,p)$ is a similarity score between P and p,
- BLAST-KNN parameters
  - the input protein sequence dataset and annotations,
  - the number of similar proteins (E-value threshold),
  - similarity score between proteins.

GOLabeler – an advanced AFP solution



You et al.: GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank.

GOLabeler – an advanced AFP solution

- The key ideas
  - use more than homology information (sequence alignment)
    - GO term frequency (naïve classification),
    - amino acid trigrams,
    - domains and motifs and biophysical properties,
  - learning to rank (LTR)
    - traditional learning solves 0/1 classification problem for each GO term,
    - solves a ranking problem on a list of items, frequent in search engines,
    - in our case, we rank GO terms wrt their relevance,
    - top-k GO terms considered, the score of parent could be replaced with max score of its children,
    - LambdaMart general LRT algorithm used (Microsoft, good performance in Yahoo challenge).

Summary

- The main topics covered
  - gene ontology – structure, purpose, size,
  - functional enrichment analysis – generalizes differential gene expression,
  - automated function prediction – computationally extend GO annotations,
- other issues
  - GSEA could be applied e.g. in genome-wide associations studies,
    - GSEA-SNP – SNPs contributing to a disease tend to group in genes,
  - methods to remove redundant terms from enriched GO lists
    - utilize hierarchical structure/overlaps between GO terms,
  - advanced ML methods in AFP
    - structured output (kernel) methods, hierarchical ensemble methods, Bayesian corrections,
  - recent progress in protein structure prediction will be reflected in AFP too
    - e.g., TALE: Transformer-based protein function Annotation with joint sequence–Label Embedding, 2020.