Flexible Enrichment with Cortana – Software Demo

Keywords: subgroup discovery, enrichment, software

Marvin Meeng MEENG@LIACS.NL
LIACS, Leiden University, Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands

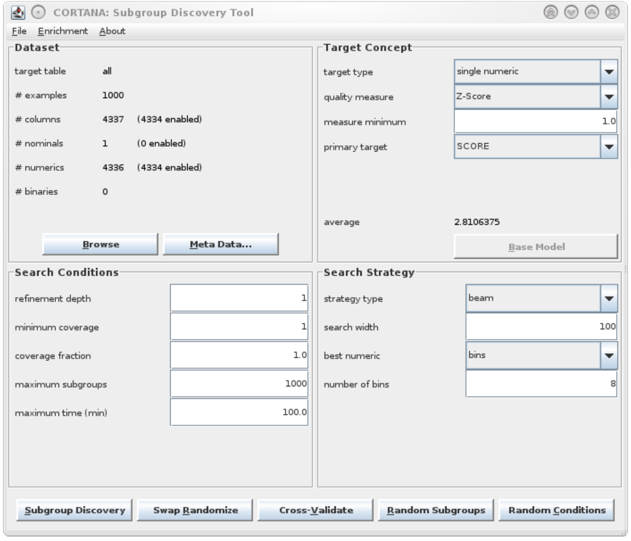
Arno Knobbe KNOBBE@LIACS.NL
LIACS, Leiden University, Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands

Abstract

The software demonstration will introduce an open-source package, called *Cortana*, that simplifies and unifies the procedure of enriching a list, or ranking, of biological entities with background knowledge. A host of software applications already exists that can do this enrichment [2, 5, 9]. However, most are focused on enriching only a single kind of biological entity, eg. genes in the case of gene set enrichment [6], or they are concerned with only a single source of background knowledge, eg. biological processes from GO. As a result, these tools are by design limited in their ability to integrate background knowledge from a variety of domain-crossing sources. The software tool introduced here, *Cortana*, will allow integration of knowledge extracted from both existing sources, like the online knowledge bases [3, 4, 8] that are used by other enrichment tools, as well as custom made ones, created by, or available to, the end user.

1. Cortana - Main

The rationale behind using *Cortana* as the basis for the proposed enrichment procedure is that it is a generic data mining tool. As such, it benefits from recent developments in the data mining field. These include, for example, statistically sound validation methods and a range of well-understood, and well-tested quality measures. Finally, it can deal with a variety of data types, including nominal, numeric, ordinal and binary. Furthermore, as it is not purpose-build, it can be used for a wide variety of data mining tasks, biological enrichment being among them, and this allows for many different primary and background sources to be used.



Dataset
target table all
# examples 1000
# columns 4337 (4334 enabled)
# nominals 1 (0 enabled)
# numerics 4336 (4334 enabled)
# binaries 0

Target Concept
target type single numeric
quality measure Z-Score
measure minimum 1.0
primary target SCORE
average 2.8106375

Search Conditions
refinement depth 1
minimum coverage 1
coverage fraction 1.0
maximum subgroups 1000
maximum time (min) 100.0

Search Strategy
strategy type beam
search width 100
best numeric bins
number of bins 8

Figure 1. Main window with parameters set for mining run.

2. Cortana - Bioinformatics facilities

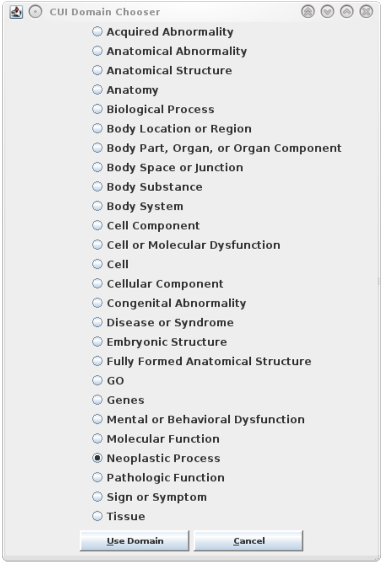
To smoothly integrate the enrichment functionality into *Cortana* a 'bioinformatics module' was added, allowing easy deployment of the tool for enrichment tasks. The typical workflow is described in the next section. The end user can use both existing and self defined, or custom, background sources. The tool is made available with domain files from two knowledge bases. The first consist of the three *GO*-domains: *biological process*, *cellular component* and *molecular function*, the well known binary gene/GO-term associations provided by `http://www.geneontology.org` [3]. The other is a set of so-called *association-matrices* derived from the literature by means of text mining. To

Appearing in *Proceedings of the 20th Machine Learning conference of Belgium and The Netherlands.* Copyright 2011 by the author(s)/owner(s).

create these, a set of common concepts is determined through text mining of a large corpus of PubMed articles [1, 7], and then association scores between these concepts are calculated and stored in a matrix, or actually multiple matrices, each one specific to a certain domain. Alternatively, a user can use custom background sources. However, in this case, the end user will also need to supply a mapping file, allowing the entities in the input ranking to be matched to the entities in a background source.

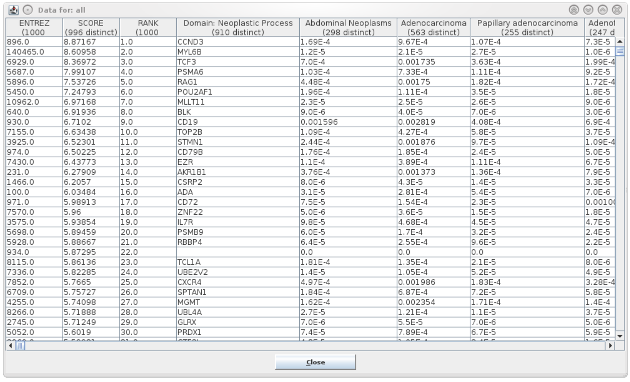
3. Cortana - Workflow

This section describes the typical workflow when enriching a (gene-)ranking with background knowledge. One starts by opening a file containing a ranking of biological entities. Typically this will be the result of an analysis of data obtained in a microarray experiment, yielding a list of *differentially expressed genes*, although other types of biological entities (eg. proteins) can be dealt with analogously. Figure 3 shows part of the ranking used in this example, with details such as the ENTREZ-identifier of the gene in question, its score resulting from the primary data analysis, its resulting rank, and the gene-symbol. Figure 1 shows *Cortana*'s main window, in which some information about the loaded data is displayed, such as the number of genes in the ranking, and the total number of descriptive values per gene that is available for enrichment. After loading the ranking, one or more files containing background knowledge from various domains can be included in the analysis. This is achieved by pressing 'Add CUI Domain', which will show a list of available CUI domains, as seen in Figure 2. Note that adding multiple background sources is possible. If one uses a custom background source, one will be asked to also indicate which file to use, to map the entities in the ranking to the corresponding information in the background source.



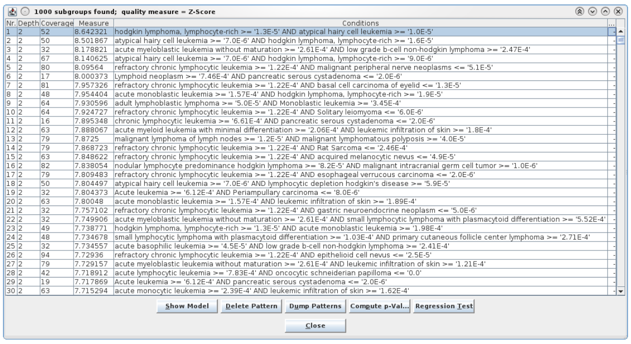
○ Acquired Abnormality
○ Anatomical Abnormality
○ Anatomical Structure
○ Anatomy
○ Biological Process
○ Body Location or Region
○ Body Part, Organ, or Organ Component
○ Body Space or Junction
○ Body Substance
○ Body System
○ Cell Component
○ Cell or Molecular Dysfunction
○ Cell
○ Cellular Component
○ Congenital Abnormality
○ Disease or Syndrome
○ Embryonic Structure
○ Fully Formed Anatomical Structure
○ GO
○ Genes
○ Mental or Behavioral Dysfunction
○ Molecular Function
● Neoplastic Process
○ Pathologic Function
○ Sign or Symptom
○ Tissue

Figure 2. Selecting a domain to add to ALL gene ranking.



ENTREZ (1000	SCORE (996 distinct)	RANK (1000	Domain: Neoplastic Process (910 distinct)	Abdominal Neoplasms (298 distinct)	Adenocarcinoma (563 distinct)	Papillary adenocarcinoma (255 distinct)	Adenoi (247 d
896.0	8.87167	1.0	CCND3	1.69E-4	9.67E-4	1.07E-4	7.3E-5
140465.0	8.60958	2.0	MYL6B	1.2E-5	2.1E-5	2.7E-5	1.0E-6
6929.0	8.36972	3.0	TCF3	7.0E-4	0.001735	3.63E-4	1.99E-4
5687.0	7.99107	4.0	PSMA6	1.03E-4	7.33E-4	1.11E-4	9.2E-5
5896.0	7.53726	5.0	RAG1	4.48E-4	0.00175	1.82E-4	1.72E-4
5450.0	7.24793	6.0	POU2AF1	1.96E-4	1.11E-4	3.5E-5	1.8E-5
10962.0	6.97168	7.0	MLLT11	2.3E-5	2.5E-5	2.6E-5	9.0E-6
640.0	6.91936	8.0	BLK	9.0E-6	4.0E-5	7.0E-6	3.0E-6
930.0	6.7102	9.0	CD19	0.001596	0.002819	4.08E-4	6.9E-4
7155.0	6.63438	10.0	TOP2B	1.09E-4	4.27E-4	5.8E-5	3.7E-5
3925.0	6.52301	11.0	STMN1	2.44E-4	0.001876	9.7E-5	1.09E-4
974.0	6.50225	12.0	CD79B	1.76E-4	1.85E-4	2.4E-5	5.0E-5
7430.0	6.43773	13.0	EZR	1.1E-4	3.89E-4	1.11E-4	6.7E-5
231.0	6.27909	14.0	AKR1B1	3.76E-4	0.001373	1.36E-4	7.9E-5
1466.0	6.2057	15.0	CSRP2	8.0E-6	4.3E-5	1.4E-5	3.3E-5
100.0	6.03484	16.0	ADA	3.1E-5	2.81E-4	5.4E-5	7.0E-6
971.0	5.98913	17.0	CD72	7.5E-5	1.54E-4	2.3E-5	0.00100
7570.0	5.96	18.0	ZNF22	5.0E-6	3.6E-5	1.5E-5	1.8E-5
3575.0	5.93854	19.0	IL7R	9.8E-5	4.68E-4	4.5E-5	4.7E-5
5698.0	5.89459	20.0	PSMB9	6.0E-5	1.7E-4	3.2E-5	2.4E-5
5928.0	5.88667	21.0	RBBP4	6.4E-5	2.55E-4	9.6E-5	2.2E-5
934.0	5.87295	22.0		0.0	0.0	0.0	0.0
8115.0	5.86136	23.0	TCL1A	1.81E-4	1.35E-4	2.1E-5	8.0E-6
7336.0	5.82285	24.0	UBE2V2	1.4E-5	1.05E-4	5.2E-5	4.9E-5
7852.0	5.7665	25.0	CXCR4	4.97E-4	0.001986	1.83E-4	3.28E-4
6709.0	5.75727	26.0	SPTAN1	1.84E-4	6.87E-4	7.2E-5	5.8E-5
4255.0	5.74098	27.0	MGMT	1.62E-4	0.002354	1.71E-4	1.4E-4
8266.0	5.71888	28.0	UBL4A	2.7E-5	1.21E-4	1.1E-5	3.7E-5
2745.0	5.71249	29.0	GLRX	7.0E-6	5.5E-5	7.0E-6	5.0E-6
5052.0	5.6019	30.0	PRDX1	7.4E-5	7.89E-4	6.7E-5	5.9E-5

Figure 3. Browse combined table of ALL gene ranking and domain (*neoplastic process* in this case).



Nr.	Depth	Coverage	Measure	Conditions	...
1	2	52	8.642321	hodgkin lymphoma, lymphocyte-rich >= '1.3E-5' AND atypical hairy cell leukemia >= '1.0E-5'	-
2	2	50	8.501867	atypical hairy cell leukemia >= '7.0E-6' AND hodgkin lymphoma, lymphocyte-rich >= '1.6E-5'	-
3	2	32	8.178821	acute myeloblastic leukemia without maturation >= '2.61E-4' AND low grade b-cell non-hodgkin lymphoma >= '2.47E-4'	-
4	2	67	8.140625	atypical hairy cell leukemia >= '7.0E-6' AND hodgkin lymphoma, lymphocyte-rich >= '9.0E-6'	-
5	2	80	8.09564	refractory chronic lymphocytic leukemia >= '1.22E-4' AND malignant peripheral nerve neoplasms <= '5.1E-5'	-
6	2	17	8.000373	Lymphoid neoplasm >= '7.46E-4' AND pancreatic serous cystadenoma <= '2.0E-6'	-
7	2	81	7.957326	refractory chronic lymphocytic leukemia >= '1.22E-4' AND basal cell carcinoma of eyelid <= '1.3E-5'	-
8	2	48	7.954404	acute monoblastic leukemia >= '1.57E-4' AND hodgkin lymphoma, lymphocyte-rich >= '1.9E-5'	-
9	2	64	7.930596	adult lymphoblastic lymphoma >= '5.0E-5' AND Monoblastic leukemia >= '3.45E-4'	-
10	2	64	7.924727	refractory chronic lymphocytic leukemia >= '1.22E-4' AND Solitary leiomyoma <= '6.0E-6'	-
11	2	16	7.895348	chronic lymphocytic leukemia >= '6.61E-4' AND pancreatic serous cystadenoma <= '2.0E-6'	-
12	2	63	7.888067	acute myeloid leukemia with minimal differentiation >= '2.06E-4' AND leukemic infiltration of skin >= '1.8E-4'	-
13	2	79	7.8725	malignant lymphoma of lymph nodes >= '1.2E-5' AND malignant lymphomatous polyposis >= '4.0E-5'	-
14	2	79	7.868723	refractory chronic lymphocytic leukemia >= '1.22E-4' AND Rat Sarcoma <= '2.46E-4'	-
15	2	63	7.848622	refractory chronic lymphocytic leukemia >= '1.22E-4' AND acquired melanocytic nevus <= '4.9E-5'	-
16	2	82	7.838054	nodular lymphocyte predominance hodgkin lymphoma >= '8.2E-5' AND malignant intracranial germ cell tumor >= '1.0E-6'	-
17	2	79	7.809483	refractory chronic lymphocytic leukemia >= '1.22E-4' AND esophageal verrucous carcinoma <= '2.0E-6'	-
18	2	50	7.804497	atypical hairy cell leukemia >= '7.0E-6' AND lymphocytic depletion hodgkin's disease >= '5.9E-5'	-
19	2	32	7.804373	Acute leukemia >= '6.12E-4' AND Periampullary carcinoma <= '8.0E-6'	-
20	2	63	7.80048	acute monoblastic leukemia >= '1.57E-4' AND leukemic infiltration of skin >= '1.89E-4'	-
21	2	32	7.757102	refractory chronic lymphocytic leukemia >= '1.22E-4' AND gastric neuroendocrine neoplasm <= '5.0E-6'	-
22	2	32	7.749906	acute myeloblastic leukemia without maturation >= '2.61E-4' AND small lymphocytic lymphoma with plasmacytoid differentiation >= '5.52E-4'	-
23	2	49	7.738771	hodgkin lymphoma, lymphocyte-rich >= '1.3E-5' AND acute monoblastic leukemia >= '1.98E-4'	-
24	2	48	7.734678	small lymphocytic lymphoma with plasmacytoid differentiation >= '1.03E-4' AND primary cutaneous follicle center lymphoma >= '2.71E-4'	-
25	2	32	7.734557	acute basophilic leukemia >= '4.5E-5' AND low grade b-cell non-hodgkin lymphoma >= '2.41E-4'	-
26	2	94	7.72936	refractory chronic lymphocytic leukemia >= '1.22E-4' AND epithelioid cell nevus <= '2.5E-5'	-
27	2	79	7.729157	acute myeloblastic leukemia without maturation >= '2.61E-4' AND leukemic infiltration of skin >= '1.21E-4'	-
28	2	42	7.718912	acute lymphocytic leukemia >= '7.83E-4' AND oncocytic schneiderian papilloma <= '0.0'	-
29	2	19	7.717869	Acute leukemia >= '6.12E-4' AND pancreatic serous cystadenoma <= '2.0E-6'	-
30	2	63	7.715294	acute monocytic leukemia >= '2.39E-4' AND leukemic infiltration of skin >= '1.62E-4'	-

Figure 4. Result of enriching ALL gene ranking with Neoplastic Process.

Cortana will now combine the original data with the selected domain(s), the result of which can be checked by selecting 'Browse' (see Figure 3) or 'MetaData'. Then, one sets the parameters to be used for the mining run, and starts the actual enrichment by pressing 'Subgroup Discovery', again see Figure 1. After the mining finishes, a result window like Figure 4 will be shown, listing those subgroups, or concepts, that were found to be interesting, according to the selected parameters. In this case, setting the search parameter *refinement depth* to *2*, see Figure 1, allowed two conditions to be combined to form subgroups.

4. Benefits of the presented tool

Cortana is a flexible data mining tool for exploratory data analysis. With the addition of a bioinformatics module, its range of generic data mining facilities can now be employed for enrichment in the biological and medical domain. The tool comes with a large collection of background knowledge that can be involved in the enrichment process. This collection includes the customary functional annotations from GO (binary concepts), as well as the association matrices describing the level of association between each entity in the ranking and concepts from a number of domains (numeric concepts). In total, the 28 background files comprise 26.8 GB of background information available for enrichment. The bioinformatics module of *Cortana* has been employed in a number of medical and biological applications, including enrichment of neuroblastoma gene rankings, leukaemia (ALL and AML) gene rankings, and metabolic syndrome gene and metabolites rankings.

References

[1] Jelier, Schuemie, Veldhoven, et al., *Anni 2.0: a multipurpose text-mining tool for the life sciences*, Genome Biology 2008, 9(6):R96.

[2] Glynn Dennis Jr. et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery*, Genome Biology, 2003 4(9):R60, `http://david.abcc.ncifcrf.gov/home.jsp`.

[3] The Gene Ontology, 2011, `http://www.geneontology.org`.

[4] Ensembl, 2011, `http://www.ensembl.org`.

[5] Zeeberg B.R. et al., *GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data*, Genome Biology, 2003 4(4):R28, doi:10.1186/gb-2003-4-4-r28, `http://discover.nci.nih.gov/gominer/index.jsp`.

[6] Subramanian, et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, PNAS October 25, 102:43, 2005.

[7] Jelier, Goeman, Hettne, Schuemie, den Dunnen & 't Hoen, *Literature-aided interpretation of gene expression data with the weighted global test*, Briefings in Bioinformatics, doi:10.1093/bib/bbq082, 2010.

[8] KEGG: Kyoto Encyclopedia of Genes and Genomes, 2011, `http://www.genome.jp/kegg/`.

[9] Search for Enriched Gene Sets, 2011, `http://kt.ijs.si/software/SEGS`.

[10] Trajkovski, Lavrač & Tolar, *SEGS: Search for enriched gene sets in microarray data*, Journal of Biomedical Informatics, 41(4), 588–601, 2008, `http://dx.doi.org/10.1016/j.jbi.2007.12.001`.

Cortana

Open Source Subgroup Discovery