Flexible Enrichment with Cortana – Software Demo

Keywords: subgroup discovery, enrichment, software

Marvin Meeng
LIACS, Leiden University, Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands

Arno Knobbe
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, e.g., genes in the case of gene set enrichment [6], or they are concerned with only a single source of background knowledge, e.g., 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.

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
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 (e.g., 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.

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

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

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.
Enrichment with Cortana

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


