SOFTWARE TOOL ARTICLE

Explain your data by Concept Profile Analysis Web Services
[version 1; referees: 2 approved with reservations]

Kristina Hettne¹, Reinout van Schouwen¹, Eleni Mina¹, Eelke van der Horst¹,
Mark Thompson¹, Rajaram Kaliyaperumal¹, Barend Mons¹, Erik van Mulligen²,
Jan A. Kors², Marco Roos¹

¹Department of Human Genetics, Leiden University Medical Center, Leiden, 2300 RC, The Netherlands
²Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, 3000 CA, The Netherlands

Abstract

The Concept Profile Analysis technology (overlapping co-occurring concept sets based on knowledge contained in biomedical abstracts) has led to new biomedical discoveries, and users have been able to interact with concept profiles through the interactive tool “Anni” (http://biosemantics.org/anni). However, Anni provides no way for users to save their procedures, results, or related provenance. Here we present a new suite of Web Service operations that allows bioinformaticians to design and execute their own Concept Profile Analysis workflow, possibly as part of a larger bioinformatics analysis. The source code can be downloaded from ZENODO at http://www.dx.doi.org/10.5281/zenodo.10963.

This article is included in the International Society for Computational Biology Community Journal gateway.

Corresponding author: Kristina Hettne (K.M.Hettne@lumc.nl)
Competing interests: The authors declare that they have no competing interests.
Grant information: The work in this paper was funded by the Seventh Framework Programme of the European Commission (Digital Libraries and Digital Preservation area ICT-2009.4.1 project reference 270192) (Wi4Ever), and grant agreement No. 305444 (RD-Connect).
Copyright: © 2014 Hettne K et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
How to cite this article: Hettne K, van Schouwen R, Mina E et al. Explain your data by Concept Profile Analysis Web Services [version 1; referees: 2 approved with reservations] F1000Research 2014, 3:173 (https://doi.org/10.12688/f1000research.4830.1)
Introduction
Concept Profile Analysis (CPA) has proven a powerful tool for interpreting and prioritizing results of bioinformatics analysis, and for linking data sets based on the best “educated guess” when precise links are not available. The technology uses the vector space model to relate concepts (such as genes and biological processes) mined from the literature to each other. Vectors can be compared efficiently and transparently\(^1\), and the model yields a measure of the strength of the relationship between concepts. We call these vectors “concept profiles”. The CPA algorithms have for example been successfully applied to compare microarray studies\(^2\), for predicting proteins putatively associated with muscular dystrophy pathways\(^3\), and for associating chemical structures with gene expression data\(^4\).

The standalone application Anni\(^10\) supports a number of standard CPA operations. For example, to perform pathway analysis for a gene expression experiment, a user first provides a list of gene database identifiers for the most significantly expressed genes. Anni uses these identifiers to query the concept profile database for the corresponding concept profiles, and subsequently constructs a “concept set” of these profiles. To match the list of genes with pathways, the user performs the operation “match concept sets” for the gene concept set with a predefined concept set of the category “Gene Ontology (GO) biological process”. Note that we refer here to GO concept profiles. The concept profile matching scores between the two concept sets are calculated by Anni, resulting in a ranked list of GO biological processes for the gene list. Finally, literature evidence in the form of documents containing co-mentions of the gene and biological processes can be retrieved by Anni from a supporting documents database, or from documents providing enough statistical evidence to support the gene-biological process associations without actually mentioning the gene and the biological process together in an abstract.

Here we present a new suite of Web Service (WS) operations that allows bioinformaticians to design and execute their own CPA workflow outside the Anni Web tool, possibly as part of a larger bioinformatics analysis. The WS was designed according to the outcome of an Anni usage analysis, where the common user and machine operations were identified.

Technical specifications
We implemented the CPA WS using Java, Model-View-Controller (MVC) Spring framework, and Apache Tomcat following the Java API for XML WS (JAX-WS) specifications. We compiled the Anni Java code for the different operations into separate libraries, for which wrappers were written in Java. Spring MVC was used as a WS interface to remote applications. The WS was implemented according to the JAX-WS standard, enabling an auto-generated WSDL specification and use of Java Annotations to specify operations. Apache Tomcat was used for deployment. The CPA WS uses a database of indexed PubMed records. The thesaurus behind the Anni Web application was converted to Simple Knowledge Organization System (SKOS), and the SKOS concept IDs were implemented as resolvable Unique Resource Identifiers leading to a Virtuoso Universal Server triple store.

User and machine operations as Taverna workflows
As an example on how to work with the CPA WS we implemented several workflows in the workflow management system Taverna workbench v 2.4\(^3\) following the best practices for workflow design\(^5\). The whole suit of CPA workflows consists of 11 workflows collected in a myExperiment pack [http://www.myexperiment.org/packs/368]. These workflows are of two different types: 1) nine workflows calling one WS operation, and 2) two pipelines of nested workflows calling more than one WS operation. The workflows of type 1 are the building blocks to make pipelines of type 2, and were implemented with re-usability in mind.

Here we describe the workflow “Match concept profiles with predefined set” (Figure 1) in order to illustrate the design and use of the WS and workflows. The workflow invokes the WS operation “getSimilarConceptProfilesPredefined”. The operation takes three input parameters, which can be accessed using the XML splitter function in Taverna. The user specifies the concept(s) to be matched (“Query concept IDs”), the concept set to match against (“Match concept set”), and a cutoff number of matched concepts to return (“Cutoff”).

Opening the “Run workflow” window in Taverna will result in showing the structured annotations for the whole workflow and the input parameters, as well as the example values (Figure 2). WS functional annotations can be accessed via the “Details” tab in Taverna (Figure 3). When the workflow is run, it will produce a ranked list of concepts associated to the query concept(s), and their similarity scores.

The above described workflow executes the core functionality of concept profile matching. The other WS operations implement functionality such as explaining the association found (by listing the common concepts contributing most to the score) and showing the literature evidence (by retrieving the links to the abstracts in PubMed). Workflows implementing these WS can be coupled to the “Match concept profiles with predefined set” workflow to form a pipeline of nested workflows. Examples of such pipelines are the “GWAS to biomedical concept” nested workflow, which performs Single Nucleotide Polymorphism annotation (SNP), and the “Annotate gene list with top ranking concepts” nested workflow for gene annotation (Figure 4).

Discussion
The CPA WS and workflows raise the level of reproducibility of bioinformatics experiments that make use of CPA compared to Anni, and the CPA WS can more easily be used together with other tools. For example, CPA-based SNP annotation can be performed with the CPA WS by coupling an external WS to map the SNP identifiers to Entrez gene identifiers. With Anni, the SNP to Entrez gene identifier analysis would have to be performed separately, decreasing the reproducibility.

Some of the functionalities in Anni have not been migrated to the WS. For example, Anni provides a function for hierarchical clustering of the results. Clustering is not a CPA function by itself, but we are considering to implement workflows that perform this function. We are also working on a workflow implementation of the process.
Figure 1. **Taverna workflow for matching concept(s) with a predefined set of concept profiles.** Blue boxes represent the workflow inputs and outputs, green box the WS invocation, and purple boxes the XML splitters for the inputs and outputs of the WS operation. The workflow is available at [http://www.myexperiment.org/workflows/3396](http://www.myexperiment.org/workflows/3396).

Figure 2. **Taverna run window.** Detailed, structured descriptions for the whole workflow and its input parameters, with example values are shown in the window.
Figure 3. Taverna details window. A detailed description of the function of the WS operation is shown in the window.

Figure 4. Taverna nested workflow for gene annotation. Blue boxes represent input and output parameters, purple boxes the local Taverna worker services, yellow boxes the Xpath services for fast XML parsing, and grey boxes the constant values. The workflow is available at http://www.myexperiment.org/workflows/3921.
that creates the data underlying the Anni WS, possibly using the recently developed text-mining workbench Argo, allowing for more flexibility in performing CPA. Specialization of the underlying resources for services to use in specific research domains, such as plant breeding or metabolomics, is a topic for future work.

Conclusions
By creating a WS building upon the Anni interactive tool, we made available the CPA technology in a way that users can easier integrate the technology with other software and save their procedures, results and related provenance.

Software availability
CPA WS: https://www.biocatalogue.org/services/3559

Latest source code
https://trac.nbic.nl/conceptprofileminingws/
Workflows: http://www.myexperiment.org/packs/368

Archived source code as at the time of publication
http://www.dx.doi.org/10.5281/zenodo.10963

Software license
Apache 2.0

References
Open Peer Review

Naoaki Okazaki
Department of System Information Sciences, Tohoku University, Sendai, Japan

This paper presents a suite of Web services with which users can design and execute their own workflow for Concept Profile Analysis. The paper includes examples of running the Web services on Taverna workbench.

The motivation of promoting the Web services presented in this paper is unclear to me. The abstract says, “However, Anni provides no way for users to save their procedures, results, or related provenance.” I guess that the authors wrote this sentence in order to explain the advantage of using the Web services over Anni. However, I am not sure whether bioinformaticians prefer the Web services to Anni in their analyses. Some may find Anni more useful than the Web services because Anni was designed to perform pre-defined workflows. For this reason, I am also not sure whether the conclusions are justified sufficiently on the basis of the developed Web services. The impact of this paper would be greater if the authors could explain the sales point(s) of this software (especially to bioinformaticians, real users) more concretely.

The explanation of how the Web services work on Taverna workbench is helpful for us to understand what can be achieved by the Web services. However, this paper lacks the comprehensive detail of the Web services, e.g., what kinds of services were implemented, what is the functionality of each service, what is the underlying research technologies for implementing the each service (although software technologies such as Apache Tomcat are explained). These details may be useful for readers to imagine the possible way to integrate the Web services with other tools.

Minor comment:

It would be better if this paper could introduce Concept Profile Analysis (CPA) with references.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 05 August 2014
This article describes a suite of services that bring the previously existing concept profile analysis tool "Anni" to the web. These web services enable a range of bioinformatics analysis tasks, and can be embedded in other workflows. The authors provide examples of implementing workflows within the Taverna workbench which nicely illustrate the applicability of the services.

The authors provide links to both the software and the Taverna workflows used as examples, which is very good.

The article could be improved by providing a more clear distinction between what can be achieved with the web services, and how the operations are implemented. The very first paragraph of the article, for instance, provides a strongly methodological description of the tool (introducing the vector space model, and the notion of comparing vectors). A more application-oriented oriented motivation for the tool would be preferable right at the beginning; the details of e.g. how "pathway analysis" is supported (via the concept profiles) can be provided after the general rationale for the tool. Note that the authors also mention a use case for "linking data sets" but do not appear to provide a concrete example of where this might be relevant. Similarly, the technical specifications mix somewhat user and developer functionality; some reorganisation could improve the manuscript. Where they mention "the user", is this a 'real' user (i.e. someone interested in bioinformatics analysis), or a computational user/developer?

Where the authors state, "The concept profile matching scores between the two concept sets", does this refer to a pairwise comparison between the elements of the two sets, or a more holistic/aggregated comparison across the complete sets? Please clarify.

Figure 2 might be more interesting if it showed an example of a "Match concept set" rather than the "Cutoff" value. How are such concept sets defined? On the other hand, parameters such as "cutoff" should be explained, at least in the figure caption, for completeness.

There are several details provided that are not fully explained. Specifically, there is a mention of "the XML splitter function in Taverna" for input parameters. Again, please consider whether this is important for a user to understand, or rather for a developer to understand the code. Also, what are "WS functional annotations"? (cf. also the reuse of the term "annotation" with possibly a different sense in "Single Nucleotide Polymorphism annotation" and "gene annotation"). What do you refer to with "the data underlying the Anni WS"?

The conclusions are justified, if a bit unspecific. A more targeted summary of the functionality would be preferable, rather than "we made available the CPA technology".

Note a few small English usage issues. (1) "considering to implement" should be "considering implementing" (2) "users can easier integrate" should be "users can more easily integrate".

Competing Interests: No competing interests were disclosed.
I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com