

U-Compare Species Disambiguation Service

1. BASIC INFORMATION

Service name

U-Compare Species Disambiguation Service

Overview and purpose of the tool

This is a web service that identifies biological named entities and disambiguates them according to species, by assigning a species ID from the NCBI taxonomy. Also identifies sentences and tokens.

A short description of the algorithm

This web service is based on a UIMA-based workflow, created using the U-Compare text mining system¹. The workflow was exported from U-Compare as a web service using the built in functionality (Kontonastios et al., In Press). The workflow was created as part of the work to increase the number of interoperable tools operating on different European languages (Ananiadou et al, 2011).

The workflow consists of the following UIMA-compliant tools

- 1) GENIA Sentence Splitter (University of Manchester)
- 2) GENIA Tagger (with tokenisation) (University of Manchester) (Tsuruoka et al., 2005)
- 3) Species Word Detector (University of Manchester) (Wang et al., 2011)
- 4) ExtractAbbrev (University of California, Berkley)²
- 5) Species Disambiguator (University of Manchester) (Wang et al., 2011)

2. TECHNICAL INFORMATION

Software dependencies and system requirements

This is a web service that can be run from a browser or accessed programmatically. The only basic requirement is an Internet connection.

¹ <http://nactem.ac.uk/ucompare/>

² <http://biotext.berkeley.edu/software.html>

Installation

There is no installation. The web service can be accessed at the following URL:
http://nactem001.mib.man.ac.uk:8080/UCompareWebServices/Species_Disambiguation

The web form available at this URL is shown in Figure 1, with some English text entered into the text box.

Examples

XML document inline XML **stand-off annotation**

Example abstracts

PMC_1804205
PMC_1874608
PMC_2358977
PMC_2651894
PMC_2714965
PMID_1590827
PMID_11393792
PMID_16583246
PMID_17709377
PMID_18264140
PMID_18286479
PMID_18296627
PMID_19609235
PMID_19781662
PMID_20184394

Acid pH activation of the PmrA/PmrB two-component regulatory system of *Salmonella enterica*
Acid pH often triggers changes in gene expression. However, little is known about the identity of the gene products that sense fluctuations in extracytoplasmic pH. The Gram-negative pathogen *Salmonella enterica* serovar Typhimurium experiences a number of acidic environments both inside and outside animal hosts. Growth in mild acid (pH 5.8) promotes transcription of genes activated by the response regulator PmrA, but the signalling pathway(s) that mediates this response has thus far remained unexplored. Here we report that this activation requires both PmrA's cognate sensor kinase PmrB, which had been previously shown to respond to Fe³⁺ and Al³⁺, and PmrA's post-translational activator PmrD. Substitution of a conserved histidine or of either one of four conserved glutamic acid residues in the periplasmic domain of PmrB severely decreased or abolished the mild acid-promoted transcription of PmrA-activated genes. The PmrA/PmrB system controls lipopolysaccharide modifications mediating resistance to the antibiotic polymyxin B. Wild-type *Salmonella* grown at pH 5.8 were > 100 000-fold more resistant to polymyxin B than organisms grown at pH 7.7. Our results suggest that protonation of the PmrB periplasmic histidine and/or of the glutamic acid residues activate the PmrA protein, and that mild acid promotes cellular changes resulting in polymyxin B resistance.

Run

Service Description
Web service created by exporting UIMA-based workflow from the U-Compare text mining system.
Functionality: Identifies biomedical named entities (genes and proteins) in plain text. Also identifies instances.

Usage
POST request should be sent to use the service
1) text -- the value of this parameter is the text to analyze. Expected encoding is UTF-8

References

Application programming interface
String text = "Hello Mr. John Smith I";
String parameters = "text=" + URLEncoder.encode(text, "UTF-8") + "&mode=inline";
URL url = URLUtil.ofTheService;

Figure 1: Web form for the web service

Execution instructions

The web service can be executed by typing or pasting text into the online form and clicking on the “Run” button.

Alternatively, the web service can be executed from within program code, as explained in the “Usage” and “Application programming interface” boxes of the web form.

A POST request should be used to call the service. The following parameters may be used in the request:

- **text** - the value of this parameter is the text to analyze. Expected encoding is UTF-8. This parameter is obligatory.

- **lang** - This parameter sets the language of the text. If this parameter is not provided, then the value "en" will be used
- **mode** - This parameter sets the format of the annotated information returned by the service. If this parameter is not set, XML output will be produced. The two possible types of output are as follows:
 - **inline** – annotations are encoded as inline XML.
 - **xml** – results are output as an XML document containing the annotations added

The following code example shows how the web service can be called from Java code:

```
//Set the input text
String text = "<Text_to_be_analysed>";
//Set the parameter string
String parameters = "text=" + URLEncoder.encode(text,
"UTF-8") + "&mode=inline";
//Create the URL connection
URL url = new
URL("http://nactem001.mib.man.ac.uk:8080/UCompareWebServi
ces/Species_Disambiguation");
URLConnection connection = url.openConnection();
connection.setDoOutput(true);
//Create Output stream
OutputStreamWriter writer = new
OutputStreamWriter(connection.getOutputStream());
//write parameters to output stream
writer.write(parameters);
writer.flush();

//Read the results returned by the service
BufferedReader reader = new BufferedReader(new
InputStreamReader(connection.getInputStream(), "UTF-8"));
String line;
while ((line = reader.readLine()) != null) {
    System.out.println(line);
}
```

}

Input/Output data formats

Input data formats

The input is plain text, UTF-encoded.

Output data format

If the service is run from the web interface, then the output is visualized in the interface using colored highlights in the text to show the individual annotations, and one or more tables of information below, each corresponding to a particular type of annotation.

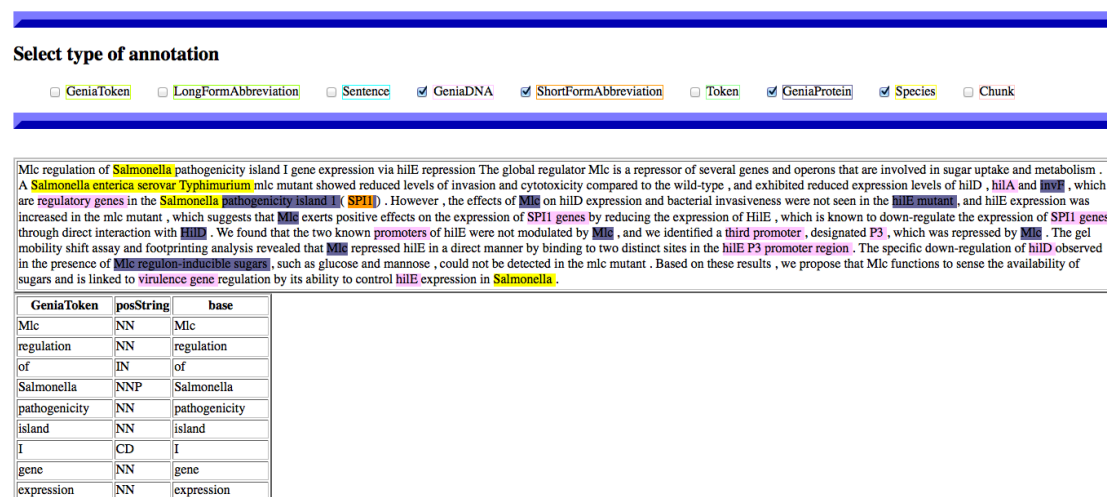
If the service is run programmatically, then the output is provided in XML format. See section 3 for an example.

Integration with external tools

The API allows the functionality of the web service to be embedded in any application.

3. CONTENT INFORMATION

Using the web interface, the output of the service is visualised as shown in Figure 2.



Select type of annotation

GeniaToken LongFormAbbreviation Sentence GeniaDNA ShortFormAbbreviation Token GeniaProtein Species Chunk

Mlc regulation of **Salmonella** pathogenicity island I gene expression via hIIE repression The global regulator Mlc is a repressor of several genes and operons that are involved in sugar uptake and metabolism . A **Salmonella enterica** serovar **Typhimurium** mlc mutant showed reduced levels of invasion and cytotoxicity compared to the wild-type , and exhibited reduced expression levels of hIiD , hIiA and **hIiE** , which are regulatory genes in the **Salmonella pathogenicity island I** (**SPII**). However , the effects of **Mlc** on hIiD expression and bacterial invasiveness were not seen in the **hIiE** mutant , and hIiE expression was increased in the mlc mutant , which suggests that **Mlc** exerts positive effects on the expression of **SP11** genes by reducing the expression of **HII E** , which is known to down-regulate the expression of **SP11** genes through direct interaction with **hIiE** . We found that the two known promoters of hIiE were not modulated by **Mlc** , and we identified a third promoter , designated **P3** , which was repressed by **Mlc** . The gel mobility shift assay and footprinting analysis revealed that **Mlc** repressed hIiE in a direct manner by binding to two distinct sites in the hIiE **P3** promoter region . The specific down-regulation of hIiD observed in the presence of **Mlc** regulation-inducible sugars , such as glucose and mannose , could not be detected in the mlc mutant . Based on these results , we propose that Mlc functions to sense the availability of sugars and is linked to virulence gene regulation by its ability to control hIiE expression in **Salmonella** .

GeniaToken	posString	base
Mlc	NN	Mlc
regulation	NN	regulation
of	IN	of
Salmonella	NNP	Salmonella
pathogenicity	NN	pathogenicity
island	NN	island
I	CD	I
gene	NN	gene
expression	NN	expression

Figure 2: Visualisation of web service output

In Figure 2, the top of the screen has check boxes corresponding to each type of annotation produced by the workflow –Checking one or more of the boxes will cause the annotations to become highlighted in the view of the text below. In Figure 2, annotations corresponding to genes, proteins, species names and abbreviations are highlighted, each using a different colour. Below the text are tables providing further

information about the annotations. The table shown in Figure 2 displays information about tokens, their parts-of-speech and base forms. In Figure 3, the table is shown with species names and their NCBI taxonomy IDs.

Species	speciesId
Salmonella	590
Salmonella enterica serovar Typhimurium	90371
Salmonella	590
Salmonella	590

Figure 3: Species names and IDs

An example of the XML output format, which is more suited to programmatic use, is shown in Figure 4. In the XML, the start and end offsets of each annotation in the text are encoded in the “begin” and “end” attributes. Other features of the annotations are encoded in other attributes of the XML annotations.

```

- <result>
- <Sentence begin="0" end="88">
- <Chunk begin="0" end="14" chunkType="NP">
- <GeniaToken begin="0" end="3" posString="NN" base="Mlc">
  <Token begin="0" end="3">Mlc</Token>
</GeniaToken>
- <GeniaToken begin="4" end="14" posString="NN" base="regulation">
  <Token begin="4" end="14">regulation</Token>
</GeniaToken>
</Chunk>
- <GeniaToken begin="15" end="17" posString="IN" base="of">
- <Token begin="15" end="17">
  <Chunk begin="15" end="17" chunkType="PP">of</Chunk>
</Token>
</GeniaToken>
- <Chunk begin="18" end="67" chunkType="NP">
- <Species begin="18" end="28" speciesId="590">
  - <GeniaToken begin="18" end="28" posString="NNP" base="Salmonella">
    <Token begin="18" end="28">Salmonella</Token>
  </GeniaToken>
</Species>
- <GeniaToken begin="29" end="42" posString="NN" base="pathogenicity">
  <Token begin="29" end="42">pathogenicity</Token>
</GeniaToken>
- <GeniaToken begin="43" end="49" posString="NN" base="island">
  <Token begin="43" end="49">island</Token>
</GeniaToken>
- <GeniaToken begin="50" end="51" posString="CD" base="I">
  <Token begin="50" end="51">I</Token>
</GeniaToken>
- <GeniaToken begin="52" end="56" posString="NN" base="gene">
  <Token begin="52" end="56">gene</Token>
</GeniaToken>
- <GeniaToken begin="57" end="67" posString="NN" base="expression">
  <Token begin="57" end="67">expression</Token>
</GeniaToken>
</Chunk>
- <GeniaToken begin="68" end="71" posString="IN" base="via">
- <Token begin="68" end="71">
  <Chunk begin="68" end="71" chunkType="PP">via</Chunk>
  ...

```

Figure 4: XML output example

3. LICENCE

- a) The web service only is licenced NaCTeM Web Service Licence Agreement (standard non-commercial use) – see “U-Compare-Species-Disambiguation-Service-Licence.pdf” in the “licences” directory. Please contact us using the details below if you require a commercial licence.
- b) The tools used in the workflow on which the web service is based may have their own licences. The NaCTeM Web Service Licence Agreement does NOT apply to these tools.

4. ADMINISTRATIVE INFORMATION

Contact

For further information, please contact Sophia Ananiadou:

sophia.ananiadou@manchester.ac.uk

5. REFERENCES

Ananiadou, S., Thompson, P., Kano, Y., McNaught, J., Attwood, T. K., Day, P. J. R., Keane, J., Jackson, D. and Pettifer, S.. (2011). Towards Interoperability of European Language Resources. *Ariadne*, 67.

Kontonatsios, G., Korkontzelos, I., Kolluru, B., Thompson, P. and Ananiadou, S. (In Press). Deploying and Sharing U-Compare Workflows as Web Services. *Journal of Biomedical Semantics*.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii, (2005) Developing a Robust Part-of-Speech Tagger for Biomedical Text, *Advances in Informatics - 10th Panhellenic Conference on Informatics*, LNCS 3746, pp. 382-392.

Wang, X., Tsujii, J. and Ananiadou, S.. (2010). Disambiguating the Species of Biomedical Named Entities Using Natural Language Parsers. *Bioinformatics*, 26(5), 661--667