PPI: Protein Interaction Sentences Sub-task 3 (ISS)

1 Premise

In practice, protein-protein interaction information for a given pair of proteins might be mentioned several times throughout a full text article. To produce a protein interaction summary, for instance, it is useful to select the most relevant sentence expressing interaction information for a given protein pair. Also for human interpretation, natural language text passages describing a given interaction are useful. Therefore in this sub-task, the interaction sentence sub-task (ISS), we ask participants to provide, for each protein interaction pair, a ranked list of (maximum 5) evidence passages describing their interaction.

2 System Input

Collection of full text articles which contain protein interaction information curated by IntAct and MINT.

3 System output

For each protein interaction pair, a ranked list of maximum 5 text passages (containing at most 3 sentences per passage) describing their interactions has to be returned.

4 Evaluation

For the evaluation, pooling methods will be used, as follows: all the interaction evidence passages (sentences) from all the systems for each document are collected. Then a curator will categorize these into relevant and irrelevant sentences. This eliminates duplicates. This way the "best sentences" don't have to be exhaustively pre-selected; it also means that there will be limited training data available. The predictions will be evaluated in terms of:
a) Percentage of interaction relevant sentences with respect to the total number of predicted (submitted) sentences.
b) Mean reciprocal rank (MRR) of the ranked list of interaction evidence passages with respect to the manually chosen best interaction sentence.

5 Tentative release dates

The test set of this subtask will be released after the due date of the result submission of PPI subtask 1 (detection of protein interaction curation relevant articles).

Training set PPI subtasks 1-4:	July and September 2006
Test set PPI subtask ISS:	October 15, 2006
Test set prediction due for ISS:	October 22, 2006

6 Training data

Both MINT and IntAct are producing a collection of manually extracted sentences derived from

full text articles which describe their interactions (contain evidence of the interaction). These sentences were mainly extracted using cut and paste of the phrase in the text which indicates that this interaction occurs. In principle the text should not be altered in any way and only obvious phrases should be provided, meaning that if the interaction was only apparent from a table of figure this topic should not be entered, but figure and table legends may be used. Both phrases as well as full sentences and sentence passages can be included. There are cases were several alternative evidences sentences for a given interaction are provided.

Examples of extracted interaction evidence phrases:
(1) Human Mitochondrial DNA Polymerase {gamma} Forms a Heterotrimer
(2) Using a biochemical approach to search for such co-regulatory factors, we identified hGCN5, TRRAP, and hMSH2/6 as BRCA1-interacting proteins.
(3) SRp30c specifically binds to both ESE3 and ESE4

Also a collection of additional resources will be provided consisting in interaction related sentences from: (1) the Anne Lise Veuthey corpus, (2) the Christine Brun corpus, (3) the Prodisen interaction corpus and the (4) GeneRif interaction sentences. Note that these interaction evidence sentences do not necessarily follow the annotation criteria of MINT and IntAct, so they can contain also genetic interactions which are not exhaustively annotated by these two databases.

7 Test data

The interaction databases MINT and IntAct are holding back a set of curated records to produce the test set for the BioCreAtIvE contest. The test data set will consist of articles belonging to this collection of previously manually curated records. For those articles the database curators have extracted manually the evidence passages. These passages consist of phrases, sentences and sentence passages which indicate the actual protein interaction. In principle, this evidence passage should be the most obvious text segment indicating the interaction according to the database curator.
The ranked list of predicted evidence text passages returned by the participants are also compared to the manually curated ones by the expert annotators.
A total of 300 publications are expected to be part of this test set collection. Note that the test set of this task will be released after the due date for subtask 1 (detection of protein interaction curation relevant articles).

8 Data Selection

Note that the text passages referring to the protein interactions are not restricted to abstracts but in practice can be derived from any part of the full text article, including figure and table legends.

9 Submission format

Each run of predictions has to be provided as a single file with xml-like format, containing all the submitted interaction evidence sentences for the interaction pairs extracted from an article. A sample prediction entry for the correct submission format is shown below:

<ENTRY>
<PPI_SUB_TASK_ID> BC2_PPI_ISS **</PPI_SUB_TASK_ID>**
<TEAM_ID> T1_BC2_PPI **</TEAM_ID>**
<RUN_NR> 1 **</RUN_NR>**
<PMID> 10924507 **</PMID>**
<INTERACTION_PAIR>
<INTERACTOR_1> Q08211 **</INTERACTOR_1>**
<INTERACTOR_2> Q9UBU9 **</INTERACTOR_2>**
</INTERACTION_PAIR>
<SENTENCE_RANK> 1 **</SENTENCE_RANK>**
<SENTENCE_PASSAGE>
Specific interaction between RNA helicase A and Tap, two cellular proteins that bind to the constitutive transport element of type D retrovirus.
</SENTENCE_PASSAGE>
</ENTRY>

Where:
1) ENTRY: corresponds to a single evidence passage prediction
2) PPI_SUB_TASK_ID: the identifier of the interaction sentence sub-task, i.e. BC2_PPI_ISS
3) TEAM_ID: the identifier of the team (as provided to each participating team)
4) RUN_NR: the number of the submission run (maximum of three runs)
5) PMID: corresponds to the PubMed identifier of the article
6) INTERACTOR_1 : corresponds to the UniProt accession number of the interactor protein 1
7) INTERACTOR_2: corresponds to the UniProt accession number of the interactor protein 2
8) SENTENCE_RANK: corresponds to the sentence rank (1 to 5)
9) SENTENCE_PASSAGE: corresponds to the actual interaction evidence text passage (maximum 3 sentences).

NOTE: the predicted interaction sentences must come from the test set HTML full text articles, meaning that you should assure that these text segments can be directly matched to the HTML documents! We will not assure evaluation of predictions which can not be matched directly to the full text HTML articles.

10 Training data release

People who intend to participate in the protein-protein interaction (PPI) task of the second BioCreAtIvE challenge should send the following information:
1) Team contact e-mail (one per team).
2) Tentative list of participant team members (name and e-mail).
3) Institutions.

to: mkrallinger@cnio.es