PPI: Protein Interaction Pairs Sub-task 2 (IPS)

1 Premise

The characterization of protein interactions has attracted considerable interest within the biology community. A range of experimental techniques are being used to detect such protein interactions. Results of such experimental interaction characterization studies are often described in peer reviewed literature articles. For domain experts to extract manually such protein interactions from the literature is a time consuming process. This sub-task is one of the crucial parts within the protein interaction task. The aim is to identify protein-protein interaction pairs (pairs of interactors) from full text articles. The individual proteins of a given interaction pair should be uniquely identified by their corresponding UniProt accession numbers.
IntAct and MINT curate all interactions which are classified as of interaction types (MI:0190) colocalisations (MI:0403) and Physical interactions (MI:0218), as well as all the corresponding child nodes. This means that e.g. predicted interactions or genetic interactions are not considered in this contest.

2 System Input

As system input (training data), the participants will get a collection of articles with the associated interaction pairs extracted from these articles, as well as the corresponding gene mention symbols and synonyms (see 'alias type' node (MI:0300). For the test set, the system input is a collection of full text articles, for which the participating teams have to predict the interactor protein pairs.

3 System output

The participating teams are requested to provide, for each full text article, a ranked list of protein-protein interaction pairs, namely pairs of UniProt accession numbers of the interactors.

4 Evaluation

The evaluation of the submitted predictions will be in terms of precision and recall of the submitted protein interaction pairs for each article. We will evaluate three aspects:

a) If the interaction pair was correctly extracted in terms of interactor accession numbers compared to the manually assigned ones according to the so-called "spoke" model.
b) If the interaction pair was correctly extracted in terms of interactor accession numbers compared to the manually assigned ones according to the so-called "matrix" model.

The most relevant aspect which will be evaluated is of course a) , the interaction pairs according to the spoke model.

Note on protein complexes: In a TAP experiment, A is bait, and B, C, D are found by mass spectrometry to be associated with A. However, the experiment does not tell us about the detailed topology and hence the pairwise interactions of the complex. It might be for example : BACD, AB , CD , DBCA , etc.

IntAct (and MINT) report one interaction with partners ABCD, and annotate their respective roles as Bait or Prey. Thus, an IntAct interaction may have 20 partners. To expand these into binary interactions, they internally use the so-called "spoke" model, where they assume each Bait forms a pairwise interaction with each of the Preys.

The alternative would be to use the "matrix" model, where each protein in the complex is assumed to interact with each other member of the complex. While the spoke model will not always be right, there are publications arguing it is "less wrong" than the matrix model. Besides, the matrix model produces a combinatorial explosion for larger complexes.

We will evaluate predictions considering both models.

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:	June 2006
Test set PPI IPS:	October 15, 2006
Test set prediction due for IPS:	October 22, 2006

6 Training data

The training data set was derived from the content of the IntAct and MINT databases. The data files of both databases are freely accessible for download and are compliant with the HUPO PSI Molecular Interaction Format.

In principle, any of the data files from the IntAct and the MINT ftp servers would be usable to derive protein interactions pairs (with their corresponding protein accession number of UniProt), but for this sub-task, files released in 2005 and 2006 are used. All the articles contained in these databases were manually reviewed for whether they contain interaction annotation information relevant for this database. They were used to extract manually the protein interactions mentioned, linking each interacting protein to its corresponding unique UniProt accession number. For this sub-task we recommend not using articles on very large scale experiments (i.e. more than 20-30 interactions), as in the test set, no large scale interaction experiments will be encountered. Thus ideally articles with less than 21 interaction pairs should be used for training. We recommend that you check the 'alias type' node (MI:0300) and its child nodes of the MI ontology and also that you read the PPI task relevant questions.

As training data full text articles will be provided in various formats, including pdf, html and plain text (converted from pdf using the pdftotext program). Also the corresponding protein interaction annotations extracted manually from these articles will be provided in standard PSI-MI 2.5 format.

7 Test data

The interaction databases MINT and IntAct are holding back a set of curated records to produce the test set for BioCreAtIvE. Both are doing a considerable annotation effort to produce the test and training data collection. A total of around 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 of subtask 1 (detection of protein interaction curation relevant articles).

8 Data Selection

Note that the interacting proteins are not restricted to a single organism source, so in principle for the linking step to the UniProt database entry, inter-species protein name ambiguity may have to be taken into account.
The IntAct and MINT annotations (training set) are done down to the isoform level. For the test set, participants should not worry about the mapping to the master entries. The percentage of interactions with splice variant annotations is expected to be less than 5 percent. In case remapping to the master entry is an issue for the test set, this will be done by the CNIO evaluators.

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 fro an article.
A sample prediction entry is shown below:

<ENTRY>
<PPI_SUB_TASK_ID> BC2_PPI_IPS **</PPI_SUB_TASK_ID>**
<TEAM_ID> T1_BC2_PPI **</TEAM_ID>**
<RUN_NR> 1 **</RUN_NR>**
<PMID> 10924507 **</PMID>**
<INTERACTION_PAIR>
<RANK> 1 **</RANK>**
<INTERACTOR_1> Q08211 **</INTERACTOR_1>**
<INTERACTOR_2> Q9UBU9 **</INTERACTOR_2>**
</INTERACTION_PAIR>
</ENTRY>

Where:

1) ENTRY:	corresponds to a single evidence passage prediction
2) PPI_SUB_TASK_ID:	The identifier of the interaction pair sub-task, i.e. BC2_PPI_IPS
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) RANK:	the rank of the interaction pair
7) INTERACTOR_1 : protein 1	corresponds to the UniProt accession number of the interactor
8) INTERACTOR_2: protein 2	corresponds to the UniProt accession number of the interactor

10 Number of runs

For this sub-task, each participating team can submit up to three runs .

11 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