PPI: Protein Interaction Article Sub-task 1 (IAS)

1 Premise

In practice, before deciding to curate an article for protein interactions, or detecting protein interaction descriptions in sentences, it is necessary to identify those articles which actually contain relevant information relative to protein interaction annotation. Although this aspect is relevant for subsequent steps, it has often been neglected by previously published protein-interaction extraction systems. Thus this sub-task will be concerned with the classification of whether a given article contains protein interaction information.

2 System Input

The participants will be given a collection of PubMed article abstracts which have been analyzed whether they are useful to derive protein interaction annotations according to the curation standards used by the IntAct and MINT databases. Although the curation was done using the corresponding full text articles, we will only be able to provide PubMed abstracts for this sub-task. This reflects more the actual textual data which is freely available, as in practice there are still serious limitations to obtain large collections of updated full text articles (for many curated biomedical journals). It is also a way to explore the limits of abstract-based detection of annotation relevant articles.

3 System output

For a given collection of articles (abstracts), participants will need to return a ranked list of articles (identifiers) based on their relevance for protein interaction annotation

4 Evaluation

Although in principle the possibility exists to consider additional evaluation metrics (e.g. mean average precision, utility measure, confidence weighted score, etc.), we are thinking of evaluating the participating systems using the AROC (area under the receiver operating characteristic curve) measure based on the ranked predicted collections.

5 Tentative release dates

To avoid the possibility that participants will exploit the overlap between test set articles used in subtask 1 and the subtasks 2-4 for detecting the protein interaction annotation relevant articles, we will have two test set release and result submission dates for the protein interaction task.

First the test set for the interaction article detection subtask will be released. After receiving the results of this sub-task, the test set for the other subtasks will be released.
(Note that the exact dates might be shift slightly).

Tentative time schedule:

Training set PPI subtasks 1-4:	June 2006
Test set PPI subtask IAS:	October 8, 2006
Test set prediction due IAS:	October 13, 2006
Test set PPI subtasks 2-4:	October 15, 2006

6 Training data

The training data 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. (We recommend that you should have a detailed look at this format, especially for the other subtasks).

The training collections for this subtask (referred as PPI-IAS) are of three basic types:

1) True Positives (TP): collection of PubMed article abstracts which are relevant for protein interaction curation in the sense of the annotation process and guidelines used by the MINT and IntAct databases (refer to the annotation manuals of MINT and IntAct). This means that the articles corresponding to these abstracts must contain information which meets the curation standards used to extract protein interaction information. Both databases have basically the same annotation standards and formats. Thus they agree on the curation model, which has been assured using a curator agreement study on 5 full text articles. (Note that interaction types corresponding to genetic interactions are currently not curated by MINT and IntAct!)

2) True Negatives (TN): consists in articles which have been classified by domain expert curators from these two databases as not relevant for protein interaction curation after extensive full text analysis.

3) Likely True Positives (TP*): consists of a collection of PubMed identifiers of articles which have been used for protein interaction annotation by other interaction databases (namely BIND, HPRD, MPACT and GRID). Note that this additional collection is a NOISY data set and thus not part of the ordinary TP collection, as these databases might have different annotation standards compared to MINT and IntAct (e.g. regarding the curation of genetic interactions). Thus be careful when using this additional collection!

Note that in principle there are no restrictions to use additional data collections for purpose of system training. We recommend not to use for training purpose articles with more than 20 annotated interactions, as in the test set no large scale interaction experiment articles will be used.

7 Test data

The test set collection will consist of a collection of PubMed article abstracts in a format compliant with the training collection format. We expect the test collection will have over 600 articles in total (including TP and TN) cases. The participating systems will have to return, given this test collection, two non-overlapping collections of ranked lists, one for the articles predicted as being relevant for curation, and one of the articles predicted to be non-relevant (see system output section). The predictions will then be compared to previously analyzed manual classification (according to whether they are relevant for protein interaction curation), done by MINT and IntAct database curators. The actual classification is hold back for the test set.

8 Data Selection

Given the exhaustive journal curation strategy used by MINT and IntAct, there should be no bias of initial article selection. Note that these databases are not organism specific, so they curate proteins from a number of model organisms.

9 Data set format

The training data for this subtask will be provided in an XML-like format, which is easy to parse.

For a sample entry, please see the example below:

<ENTRY>
<CURATION_RELEVANCE>
1
</CURATION_RELEVANCE>
<PPI_DATABASE>
MINT
</PPI_DATABASE>
<PMID>
10022833
</PMID>
<TITLE>
Socs1 binds to multiple signalling proteins and suppresses steel
factor-dependent proliferation.
</TITLE>
<SOURCE>
EMBO J. 1999 Feb 15;18(4):904-15.
</SOURCE>
<ABSTRACT>
We have identified Socs1 as a downstream component of the Kit receptor tyrosine kinase signalling pathway. We show that the expression of Socs1 mRNA is rapidly increased in primary bone marrow-derived mast cells following exposure to Steel factor, and Socs1 inducibly binds to the Kit receptor tyrosine kinase via its Src

homology 2 (SH2) domain. Previous studies have shown that Socs1 suppresses cytokine-mediated differentiation in M1 cells inhibiting Janus family kinases. In contrast, constitutive expression of Socs1 suppresses the mitogenic potential of Kit while maintaining Steel factor-dependent cell survival signals. Unlike Janus kinases, Socs1 does not inhibit the catalytic activity of the Kit tyrosine kinase. In order to define the mechanism by which Socs1-mediated suppression of Kit-dependent mitogenesis occurs, we demonstrate that Socs1 binds to the signalling proteins Grb-2 and the Rho-family guanine nucleotide exchange factors Vav. We show that Grb2 binds Socs1 via its SH3 domains to putative diproline determinants located in the N-terminus of Socs1, and Socs1 binds to the N-terminal regulatory region of Vav. These data suggest that Socs1 is an inducible switch which modulates proliferative signals in favour of cell survival signals and functions as an adaptor protein in receptor tyrosine kinase signalling pathways.
</ABSTRACT>
</ENTRY>

The curation relevance (CURATION_RELEVANCE tag) corresponds to whether the article is useful for protein interaction curation (1) or is not useful for protein interaction curation (0).

In case of the test set, we will provide the data basically in the same format as the training data, but the following differences in two fields:

<CURATION_RELEVANCE>
NONE
</CURATION_RELEVANCE>
<PPI_DATABASE>
NONE
</PPI_DATABASE>

Where NONE corresponds to the fields which are held back for the test set.

10 Prediction submission format

Regarding the output format, it should consist of tab separated columns with the following information:

<team_id> <run_id> <sub_task_id> <type> <rank> <pmid>

where

team_id: corresponds to the assigned team identifier (provided to each team), e.g. T1_BC2_PPI
run_id: corresponds to the run id (max. of three runs per team), e.g. 1
sub_task_id: the identifier of this subtask, i.e. 'BC2_PPI_IAS'
type: prediction of relevance for protein-protein interaction: 'T' or 'F'
rank: corresponds to the rank of the prediction, must start with 1, up to the total number of articles in the returned subset.

pmid: the PubMed identifier of the prediction.

Be sure that your prediction is compliant with this simple output format.

11 Number of runs

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

12 Training data release

People who intend to participate at 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