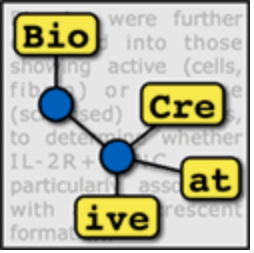
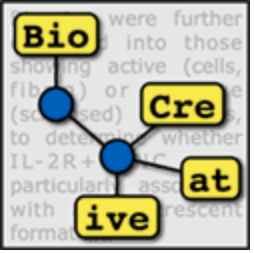


BioCreative III IAT Task Overview

Organizers: Arighi/Hirschman/Wu



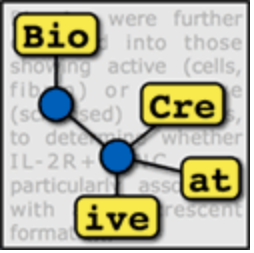
What is IAT?

- It is a demonstration interactive task.

  By demonstration we mean no challenge

  By interactive we mean teams should provide some interface for user to accomplish task

  Can we come up with ideas and metrics for next BioCreative?



Motivation: Stop hearing comments like these....

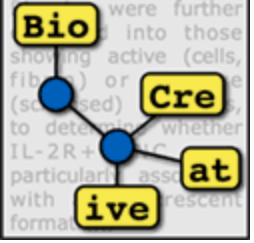
- “Express the problem that it makes sense to each other” Florance
- “Talk to your users” Spengler
- “What problem you want to solve and whether this is the best technology” Sever
- “Develop tools that people may want to use” Chanda

IAT is one step towards this goal

The subliminal task:
Find PPIs “Patient developers-Patient curators Interactions”

Patient developers set:
BioCreative teams

Patient curators set:
User Advisory Group

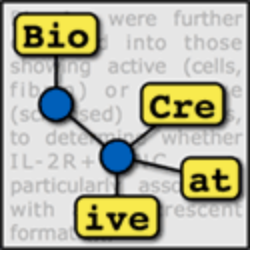


1-Establish UAG

- What is UAG? User advisory group

  Co-Chaired by Cecilia Arighi and Zhiyong Lu

  - Gianni Cesarini, MINT, University of Rome Tor Vergata, Italy
  - **Andrew Chatr-aryamontri, BioGrid, University of Edinburgh, UK**
  - **Pascale Gaudet , dictyBase, Northwestern University, USA**
  - **Michele Gwinn Giglio, University of Maryland, USA**
  - Ian Harrow, Pfizer, UK
  - Eva Huala,TAIR, Stanford University, USA
  - Pankaj Jaiswal, Gramene Database and Plant Ontology, Oregon State University, USA
  - **Donghui Li, TAIR, Stanford University, USA**
  - **Lois Maltais, MGI, The Jackson Laboratory, USA**
  - **Phoebe Roberts, Pfizer, USA**
  - **Livia Perfetto, MINT, University of Rome Tor Vergata, Italy**
  - Paul Sternberg, Wormbase, Caltech, USA
  - Luca Toldo, Merck KGaA, Germany
  - Jean-Francois Tomb, Dupont, USA

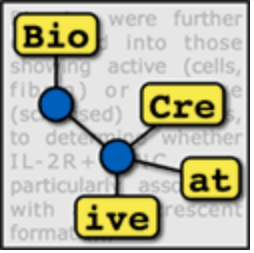


What for?

- **Develop end user requirements for interactive text mining tools:** the UAG provided the guidelines for the requirements delivered to the participants in the BioCreative III interactive task

- **Serve as users for the interactive task:** UAG members tested the systems, and provided feedback

- **Provide gene normalization annotation of a corpus of full text articles** for use in developing baseline statistics (inter-annotator agreement, and time for task completion) as well as a gold standard of articles correctly annotated for gene/protein normalization



2-What task?

Diverse Community

- Different interests:

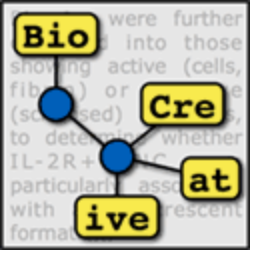
  -Curation of genes for a given organism (MOD)

  -Curation of protein-protein interactions

  -Curation of phenotypes

  -GO Annotations

  -Curation of pathways



Keep the task aligned with other BioCreative III tasks

So based on this, a common theme that fits:

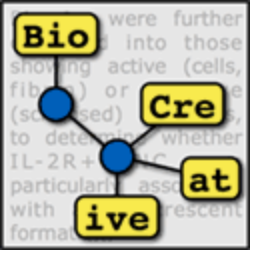
-Identify genes that are “primary/central” in the context of the article

-Link to Database ID

-Retrieve articles for which a given gene is “primary/central”

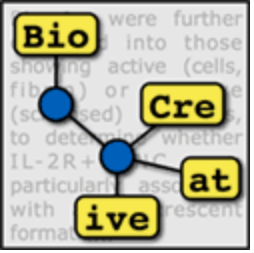
The UAG defined the concept of primary/central gene

Considered primary genes those that had experimental support but they also had biological significance in the context of the article



Setting system specifications

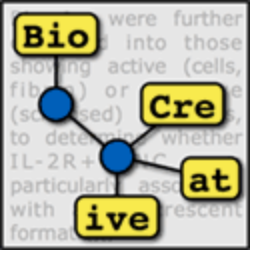
- Data set: full text articles in XML from the PubMed Central Open Access collection
- User-friendly web-based interface
- Indexing Task: Given a PMCID, provide Gene list (normalized) and ranking (based on centrality)
- Retrieval Task: Retrieve documents for which a given gene is central.



Other specifications

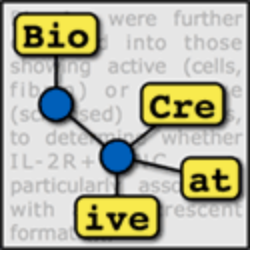
Display will include:

- editable list of gene/protein identifiers, including names, identifier linked to appropriate standard database (EntrezGene, UniProt), species, links to one or more mentions in the abstract or text.
- window showing full text, including annotations of genes



More wishes...

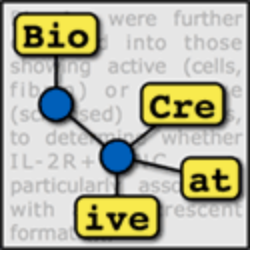
- Ability to sort gene list based on frequency (how many times it is mentioned), location (in what sections it is mentioned), experimental evidence (whether it is studied in an experiment) or their combinations
- Support to identify gene/protein mentions in text and link a mention to a unique identifier
- Support for interactive disambiguation of gene/protein mentions based on context (e.g., other genes, species, chromosomal location) to enable the user to manually select the correct unique identifier from a set of possibilities (or to enter in the identifier if it is not present in the list)
- Ability to select a gene from the list and retrieve full text articles from PubMed Central that provide further information on the selected gene (for the retrieval subtask)
- Ability to collect event and timing information at the session level (and ideally at a finer granularity of user action)
- Newly added: ability to export results (e.g. tab-delimited file) containing the following information: PMCID|Gene|DB ID|ranking is encouraged.



Timeline

Late February/Early March	Announcement of interactive task, including detailed task description
July 26	Pre-test validation
August 27	Systems ready for test by User Advisory Group
September 8	UAG return results

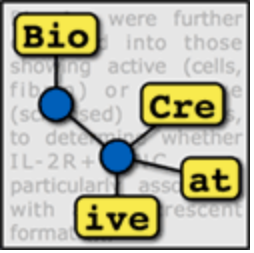
- Pre-test validation to ensure that the system will be ready by the time of testing
- Users have 1 week to get familiar with system and work with two articles
- Timeline too tight!!!!



Setting up protocol for system checking

Needed to address all sort of questions:

- Should we test system previous to the workshop? Should we do curation on site?
- How many systems?
- How many curators?
- How many articles?
- What articles?
- How to report results?
  - Report on system usability
  - Report on system performance based on the task
  - Do the systems assist in the task?

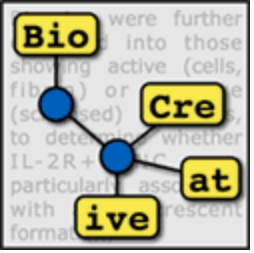


What articles to pick?

- Curators often struggle with following cases:
  1. many species (PMCID: 2680910)
  2. new genes described (PMCID:2764847)
  3. Articles with ambiguities:

     multiple genes with common gene name
     PMCID:2275796-> GLUT9 (SLC2A9 and SLC2A6)

     names shared with non-gene terminology



What systems dared to the challenge?

Team 61	MyMiner
Team 65	ODIN
Team 68	Gene View
Team 78	System University of Iowa
Team 89	System University of Wisconsin
Team 93	GNSuite

Some exchange with systems at pre-validation stage included:

-Begging to have the system ready by the time of testing

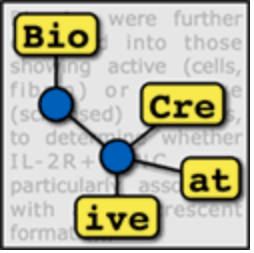
-Reporting bugs

-Suggesting some feature

Some exchange with systems after release included:

-Reporting bugs, missing data

-Requesting documentation



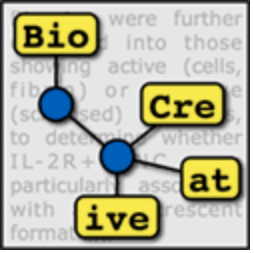
This is just a modest attempt to link the two communities

- We are aware that we do not have numbers to get to conclusions with statistically support

- Curators have little time to get familiar about the systems so training is needed. This may influence time to finish task

- Curators have different levels of expertise

- SO....This is just a collection of views from various curators to provide feedback to the systems and evaluate what can be done reasonably next time



Challenges from Organization

-Will anyone show up?

Reach a reasonable number of systems

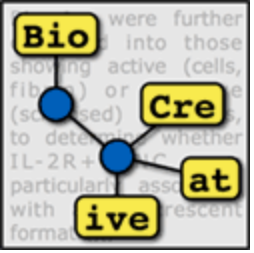
Have enough curators to test

-Timeline:

Too tight to allow system to comfortably deliver

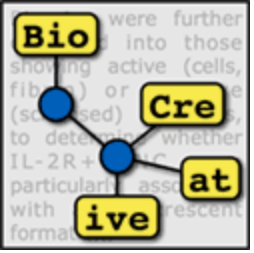
Too tight for the curators to inspect the systems

-Compiling data and check some of the curation to find what are false positives, what are false negatives. Example: case in which the manual curation have not found an entity but a curator after using the system had. Is that assertion correct?



Some of the Observations about the systems: (to discuss during UAG session):

- User found the interfaces easy to use
- They appreciate the highlighting of full text of both genes and organism
- They like the easy link to the Databases
- They like flexibility to select-unselect organism and/or genes
- They like that you can export results

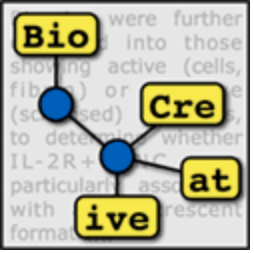


Some of the Observations about the systems: (Continuation)

- In many cases systems identify correct gene mentions as primary.
- Species assignment is not as good. In addition, the user would like to know source for species assignment
- In other cases the primary genes were not found, or found with low frequency (allowing user input would really help in improving this, some systems do have partial implementation of this feature)
- In some of the systems we see false positives which are not genes, some of them which are highly ranked (the context information should be considered)

  PMCID:2275796 ->CAD (coronary artery disease)

- In some specific cases we see some ambiguity in Retrieval task, or retrieval of irrelevant articles



Looking forward

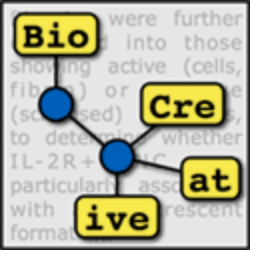
-The systems should find user at the developing phase
UAG/BioCreative could help in finding partner?

-Design timeline properly to also include user training session

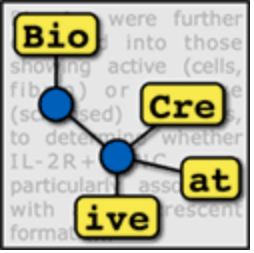
-To make it more significant for the user, they should curate papers of their own domain

-Brainstorm to improve task description

-Next workshop: Biocuration and Text Mining: 2011 to prepare for the BioCreative IV (2012)



At least I think the Developer-Curator Interaction task has been accomplished!!!
(Not sure they are patient anymore)



Unlimited Thanks to....

UAG members

Participating teams

Zhiyong Lu, NCBI, NIH

Qinghua Wang, CBCB, University of Delaware

BioCreative Organizers

Ben Carterette, CIS, University of Delaware

Oana Tudor, CIS, University of Delaware

Funded by NSF

And ALL of YOU!!!