Journal of Biomedical Informatics 42 (2009) 967–977

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Collaborative text-annotation resource for disease-centered relation extraction from biomedical text ☆

C. Cano [a], T. Monaghan [b], A. Blanco [a], D.P. Wall [b], L. Peshkin [b,*]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, 18071 Granada, Spain
[b] Center for Biomedical Informatics, Harvard Medical School, 200 Longwood Ave., Boston, MA 02115, USA

ARTICLE INFO

Article history:
Received 28 May 2008
Available online 14 February 2009

Keywords:
Information extraction
Information retrieval
Collaborative annotation
Corpus annotation
Text mining
Relation extraction
Protein–protein interaction
Gene–disease association
Autism
Disease evidence network
Clinical informatics

ABSTRACT

Agglomerating results from studies of individual biological components has shown the potential to produce biomedical discovery and the promise of therapeutic development. Such knowledge integration could be tremendously facilitated by automated text mining for relation extraction in the biomedical literature. Relation extraction systems cannot be developed without substantial datasets annotated with ground truth for benchmarking and training. The creation of such datasets is hampered by the absence of a resource for launching a distributed annotation effort, as well as by the lack of a standardized annotation schema. We have developed an annotation schema and an annotation tool which can be widely adopted so that the resulting annotated corpora from a multitude of disease studies could be assembled into a unified benchmark dataset. The contribution of this paper is threefold. First, we provide an overview of available benchmark corpora and derive a simple annotation schema for specific binary relation extraction problems such as protein–protein and gene–disease relation extraction. Second, we present BioNotate: an open source annotation resource for the distributed creation of a large corpus. Third, we present and make available the results of a pilot annotation effort of the autism disease network.

1. Introduction

Recent disease networks studies on ataxia [1,2] and Huntington's disease [3] demonstrated that the integration of biological knowledge from several sources could lead to biomedical discovery. It is particularly attractive to connect the dots using the wealth of knowledge in the published literature. The entire body of biomedical literature, dubbed the "bibliome", represents a significant resource for understanding the genetic basis of disease. It contains high quality and high-confidence information on genes that have been studied for decades, including the gene's relevance to a disease, its reaction mechanisms, structural information and well-characterized interactions.

Once a clinical study of a disease produces a list of genes, it is necessary to place these in a wider context of cellular mechanisms and molecular interactions. There are numerous academic and commercial projects striving to correctly represent the wealth of entities and relations extracted from the bibliome. However, these resources are bound to be either reliable but scarce and outdated—in the case of manual curation, or extensive but unacceptably inaccurate—in the case of automated extraction. Often this situation leaves clinical disease researchers sifting through the literature with the naked eye and using ad hoc methods to represent relevant knowledge in a disease evidence network. In this paper we describe an open-access tool that facilitates this process, while enabling the gradual development of advanced text mining algorithms.

Recently high-throughput methods in biology have produced a rising volume of scientific publications which analyze these new data and extract biological knowledge from them [4,5]. The scientific community now faces the problem of scaling methods for representation of and search over such a volume of information. For example, a direct search in PubMed [6] for the term *autism* returns over 11,000 related papers. A search for the gene *p53* returns almost 45,000 articles. The biomedical literature grows at an exponential rate [7] and currently MEDLINE contains more than 16 million publications. Naturally, there has been a growing interest in text-mining techniques to automatically extract expert knowledge from the literature.

One common idea to the multitude of advanced computational linguistics approaches is that the semantics of a relation being expressed in a piece of text has to fall out from the syntactic analysis of that text. The syntactic structure of a sentence is often represented as a parse tree capturing the interaction of phrasal constituents within the formalism of Dependency Grammar.

☆ Availability: Source code, documentation and pilot corpus results are available at: http://sourceforge.net/projects/bionotate/. BioNotate is running at http://bionotate.sourceforge.net.

* Corresponding author.
E-mail addresses: ccano@decsai.ugr.es (C. Cano), peshkin@gmail.com (L. Peshkin).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.02.001

In this paper we focus on extracting relationships between biomedical entities, such as interactions between genes and proteins or associations between genes and diseases. We propose a cross-fertilization of two efforts: (1) human-curated compilation of literature into disease evidence networks; and (2) automated information extraction. This work constitutes the first step towards a system where the creation of curated disease networks (see Fig. 1) is seamlessly augmented to collect the curator's judgments about syntactic and semantic features of the texts supporting the relationships represented in the network. Such annotations would be juxtaposed to the parsed structures in order to develop and improve the automated relation extraction, which in turn will serve to facilitate the curation process in order to build and maintain current disease networks.

Example associations extracted from biomedical texts are presented in Table 1. The first sentence reports an interaction between two genes: *SCPA* and *C5a*. The second sentence rules out the existence of a relation between a gene (*APOE*) and a disease (*autism*). While many methods have been put forward for automatically extracting these relations between biological entities from the scientific literature (for a review see [4,7–9]), the problem remains unsolved. One reason is that the development and validation of such methods requires a large corpus of correctly annotated text. However, available corpora are still small and poorly annotated. Although recent efforts have made progress [22,27], there is still a great need of large corpora annotated on protein–protein and gene–disease interactions. To analyze the state-of-the-art with respect to this issue, we have compiled the main features of the available corpora which annotate relations between biomedical entities (genes, proteins and diseases) and/or syntactic dependencies in the sentence in Table 2. In addition, Table 3 shows a detailed analysis of the corpora containing protein–protein interactions.

A careful analysis of these tables suggests the following observations:

1. Only BioInfer and LLL05 include annotations of both protein–protein relations and syntactic dependencies of the sentences.
2. There is a high heterogeneity of annotation schemas. Every corpus annotates different information at a different level. For instance, some corpora only provide the name of the interacting proteins, others provide the exact mentions of the proteins involved in the interacting sentence, keywords, the type of the interaction, etc. Furthermore, these annotations are stored in many different formats.
3. Many corpora do not provide negative examples, i.e. sentences in which at least two entity mentions occur but there is no interaction between them. We consider these examples very valuable for the training and testing any system for the automatic identification of relations.
4. Only Wisconsin annotates gene–disease relationships. Their corpus specifies which sentences contain an interaction and the interacting entities, but does not annotate keywords or syntactic dependencies (see Table 3).
5. According to Table 3, many corpora do not provide enough information to retrieve the exact mentions of the interacting proteins and the interaction keywords (i.e. the semantic link connecting the two entities). These two annotations are basic for designing and training a relation extraction tool based on pattern matching or parsing [5]. Although BioCreAtIvE I- PPI, BioIE, BioInfer and GENIA *events* include these features in their annotation schemas, the limited size of these corpora increases the need for new annotation efforts which provide the community with a large set of examples. Also, in an effort to keep the annotation process as simple as possible, the annotation schema we propose does not consider the classification of the interactions into types or the annotation of the role of the arguments.

Several annotation tools have emerged for general-purpose annotation tasks in recent years (Knowtator [29], WordFreak [30], SAFE-GATE [31], iAnnotate [32]). These tools provide the user with flexible mechanisms to define the annotation schema, so they can be customized for the annotation of relationships in biomedical texts. Some BioNLP groups have also created customized annotation tools for their specific annotation tasks such as Xconc Suite's implementation for annotating events in the GENIA corpus [22]. These tools are not intended for distribution or large scale annotation efforts, but for annotation processes carried out by a limited group of trained annotators according to sophisticated annotation schemas.

In the biomedical field, tools for collaborative annotation have been developed, such as WikiGene [41], CBioC [42] and WikiProteins [43]. WikiGene and WikiProteins are two collaborative frameworks built on wiki-based systems. Users can edit pages associated to the entities of interest and share their knowledge with the community. WikiGene is focused on genes and gene regulatory events. WikiProteins is a more ambitious effort that allows the annotation of many concepts from the biomedical literature (i.e. genes, proteins, drugs, tissues, diseases, etc.) and their relationships with other concepts (creating what the authors called *Knowlets*). While these efforts provide the community with means to access and modify a large amount of information indexed by biological entities of interest, they are not intended for the creation of a corpus explicitly stating the text that supports the relationships between the entities.

Our work is largely inspired by the recent distributed and collaborative annotation efforts that have emerged in the image analysis domain. These efforts have shown a great potential since they

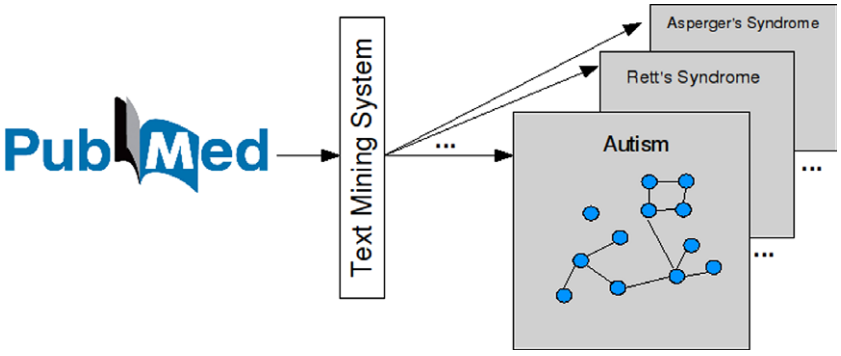

Fig. 1. The final goal is to build a text mining system to extract relationships from the literature. Extracted knowledge may be represented as disease networks, capturing the genes/proteins involved in a disease and the interactions between them.

Table 1
Snippets used as running examples of biomedical text.

1	The action of SCPA enzymatically inhibits the chemotactic activity of C5a by cleaving its neutrophil binding site. [PMID: 12964111]
2	Three promoter, one intronic, and one 3′ UTR single nucleotide polymorphisms (SNPs) in the APOE gene [...] as well as the APOE functional polymorphism (E2, E3, E4) were examined and failed to reveal significant evidence that autism is associated with APOE. [PMID: 14755445]
3	While stimulation of the D2 receptor increased branching and extension of neurites, stimulation of the D1 receptor reduced neurite outgrowth, suggesting that hormones and neurotransmitters may be capable of controlling the development of specific types of neurones.

allow any interested user to contribute in the annotation task. In particular, Label-Me [33] is a tool for tracing and labeling boundaries of objects in images, and Google™ Image Labeler [34] for labeling images to improve search results.

Our approach is similar to that implemented by CBioC. This tool allows the user to annotate relationships between biomedical concepts while browsing PubMed records. The user is presented with potential relationships from the current record extracted by automated tools or suggested by other users. Registered users can add new relationships and vote for suggested relationships. For a given PubMed record, a relationship is defined by providing the literals of the two interacting entities and the keywords of the interaction. However, CBioC does not allow to highlight the exact mentions of these words in the text. Furthermore, the users cannot access the whole corpus of annotations until it is made publicly available by the CBioC team.

We offer an intuitive framework for distributed annotation of extracts of biomedical text with interactions between biological entities of interest. Our system, which we call BioNotate, allows disparate research groups to perform literature annotation to suit their individual research needs, while at the same time contributing to the large-scale effort. There are multiple levels of integration built into the system. At one level, several annotators could collaborate on processing statements from a single corpus on their own server. At another level, multiple corpora could be created on different servers, and the resulting corpora could be integrated into a single overarching resource.

BioNotate provides the community with an annotation tool to harness the great collaborative power of biomedical community over the internet and create a substantially sized corpus as a baseline for research on biomedical relation extraction. Specifically, we focus on the annotation of protein-protein and gene–disease relationships. However, the proposed tool and annotation schema promise to be reusable for a variety of relation types.

2. Approach

We tackle the creation of this collaborative annotation resource in two phases: Phase One involves the design of the annotation schema and protocol; Phase Two involves the design and implementation of the annotation tool.

2.1. Annotation schema and protocol

According to the recently developed convention in the field, throughout this paper we use the term *snippet* to mean a small chunk of text anywhere from a few to a few dozens words, which does not have to fall on sentence boundaries. The annotation process we propose is based on focusing the attention of the annotator on two particular biomedical named entities which appear in a snippet. In this context, we ask the annotator to answer the following question:"*does this snippet imply a direct interaction between the provided entities?*".The Yes/No answer to this question allows the snippets to be classified into positive (existence of an interaction) and negative (absence of an interaction). To further help future research and to enrich the annotation of the corpus, the annotator is also asked to provide the minimal phrase in the snippet that supports his answer. For example, consider again the snippets from Table 1. The first snippet reports an interaction between the genes *SCPA* and *C5a*. Therefore, the annotator would answer *Yes* to the proposed question and highlight *inhibits* as the minimal phrase that supports this answer. The second snippet reports a negative evidence that gene *APOE* is associated with *autism*, but while the two words of interest (*APOE* and *autism*) are syntactically connected by the phrase *is associated with*, the phrase that provides the main message of the discourse is *failed to reveal*. Therefore, the annotator would mark-up *failed to reveal* as the minimal phrase that supports the *No* answer. Note that our focus is on the contextual *semantics* of the message.

If the entities only co-occur in the snippet without any explicit relation being reported between them, the answer would be *No* (the two entities are not related). Since there is no explicit support in the text for either positive or negative evidence of interaction, nothing needs to be highlighted to justify the answer in this case. An example of this can be found in the third snippet in Table 1. Complete annotations of these snippets are provided below.

We consider these two simple annotations: the Yes/No answer to the previous question and the highlighted text supporting the answer the most valuable knowledge the annotator can transfer to the corpus for identifying relations between the provided entities. In addition, this protocol is simple and intuitive enough to be embedded in an annotation tool opened to voluntary collaboration.

2.1.1. Definition of snippet

For our purposes, a snippet is a small chunk of text that may confirm or rule-out a relationship between two known entities (genes, proteins or diseases). We are particularly interested in two types of snippets:

(A) those reporting a direct interaction between a gene and a disease (gene–disease interaction);
(B) those reporting a direct interaction between two genes/proteins (gene–gene/protein–protein interaction).

By our definition, a "relationship" or "interaction" (either positive or negative) between two entities that co-occur in a snippet exists only if there is text in the snippet that explicitly supports that relationship.

Also, we are only interested in *direct* interactions between pairs of entities. For example, these sentences:

- *Gene X regulates both A and B*
- *A and B play a role in autism*
- *A regulates the expression of X. X is associated to B*

do not imply a direct interaction between *A* and *B*.

2.1.2. Annotation Process

The annotator will be shown a snippet and a pair of entities of interest: gene–gene or gene–disease. One mention of each of the two entities of interest is highlighted in the text of the snippet in advance.

Table 2
Summary of the features of different corpora with annotations on protein–protein/gene–disease relationships and syntax. *Type*: type of annotation (*Interactions*: protein–protein or gene–gene interactions, *syntax*: syntactic dependencies). *Object of the annotation*: *PPI* (protein–protein interactions), *NE* (*named entities*: proteins, genes or diseases). *Length*: size of the corpus (in number of abstracts, sentences, interactions or full papers). +/−: the corpus includes positive examples (examples with relation or +), negative examples (without relation or −) or both (+/−).

Type	Corpus name	Object of the annotation	Length	+/−	Format
Interactions	BioText [10]	PPI/disease–treatment	2143 interactions	+	Own
	Wisconsin [11]	PPI/prot-cell loc./Gene–disease	52,000/7900/13412sent	+/−	Own stand-off
	PICorpus [12]	PPI	10271 sent	+	XML/WordFreak
	Fetch prot corpus [13]	PPI	190 full texts	+/−	Stand-off XML
	HIV-1 human PI [14]	PPI	2224 interacting proteins	+	Own
	BioCreAtIvE I - PPI [15,16]	PPI/ NE	255 int/1000 sent	+/−	Stand-off XML
	SPIES corpus [17]	PPI/NE	963 sent	+	Own
	BioIE [18]	PPI/NE	250 sent	+	HTML
	Yapex [19]	PPI/NE	200 abstracts	+/−	XML
	BioContrasts [20]	Prot-Prot contrasts	100 abstracts	+	HTML
	AIMED [21]	PPI/NE	225 abstracts	+/−	XML
	GENIA events [22]	PPI/NE	1000 abstracts	+/−	XML
	BioNotate	PPI/gene–disease	–	+/−	Stand-off XML
Syntax	PennBioIE [23]	NE/syntax (constituents)	642 abstracts		XML/WordFreak
	GENIA treebank [24]	NE/syntax (constituents)	500 abstracts		XML/PTB
	Brown GENIA [25]	Syntax (constituents)	21 abstracts/ 215 sent		PTB
	DepGENIA [26]	Syntax (dependencies)	All GENIA corpus		XML
Interactions & syntax	BioInfer [27]	NE/PPI/syntax (dependencies)	1100 sent/2662 rel	+/−	XML
	LLL 05 [28]	NE/PPI/syntax (dependencies)	80 sent	+	Own stand-off

Table 3
Features of corpora with Protein-Protein interaction annotations. 1.—What proteins are marked-up? (*all* or *interacting proteins* only). 2.—For a particular interaction, are the related proteins provided? 3.—In case several mentions of a protein appear in the text, is the exact mention involved in the interaction provided? 4.—Are the keywords of the interaction provided? 5.—Are the interactions classified into different types or groups (e.g. 'inhibition', 'activation', etc.)? 6.—Are the roles of the two interacting proteins provided?

Corpora	1. annotated proteins	2. interacting proteins	3. exact mention	4. keywords	5. interaction type	6. role arguments
BioText	*interacting proteins*	✓	×	×	✓	×
Wisconsin	*all*	✓	✓	×	×	×
PICorpus	*interacting proteins*	×	×	✓	×	×
Fetch Prot Corpus	*interacting proteins*	✓	×	×	×	×
HIV-1 HUMAN PI	*interacting proteins*	✓	×	×	✓	×
BioCreAtIvE I- PPI	*all*	✓	✓	✓	✓	✓
SPIES Corpus	*all*	×	×	×	×	×
BioIE	*interacting proteins*	✓	✓	✓	×	×
Yapex	*all*	×	×	×	×	×
BioContrasts	*interacting proteins*	×	×	×	×	×
AIMED	*all*	✓	✓	×	×	×
BioInfer	*all*	✓	✓	✓	✓	✓
LLL05	*interacting proteins*	✓	✓	×	✓	✓
GENIA events	*all*	✓	✓	✓	✓	✓
BioNotate	*interacting proteins*	✓	✓	✓	×	×

For a given snippet, the annotator is asked to:

1. Indicate Yes/No whether the text implies that there is a direct interaction between the provided genes/diseases.
2. Highlight the minimal and most important phrase in the text (if any) that supports this Yes/No decision. This text should be labeled as INTERACTION.
3. Locate and highlight the one mention of each of the entities of interest which is essential to the relation of interest. These are the mentions which, if altered, would result in a phrase which no longer conveyed the same relation. For example, consider the following snippet:

```
Gene: Protein A
Gene: Protein B
Snippet: Protein A is found in tissue T. Protein A
interacts with protein B in the presence of catalyst
C to produce D.
```

changing the first mention of *Protein A* to *protein E* would not alter the relation being expressed, while changing the second mention to *protein E* would. Therefore, the second occurrence of *Protein A* should be highlighted, together with the mention of *protein B*.

Also, in the case where a pronoun refers to the entity of interest and links the entity to the INTERACTION phrase, the pronoun and not the entity mention should be marked up with the corresponding label (GENE or DISEASE). For example consider the following snippet:

```
Gene: RELN
Disease: Autism
Snippet: Gene RELN was studied in various disorders.
It turned out to be causing autism.
```

It should be marked-up as a GENE since it refers to the gene of interest *RELN* and changing this pronoun to a mention of another entity would alter the relation being expressed between this gene and the other entity of interest: *autism*.

This also applies to noun phrases that refer to one of the entities of interest. For example, in the following snippet:

```
Gene: FXR1
Gene: FMRP
Snippet: Recently, two proteins homologous to FMRP
were discovered: FXR1 and FXR2. These novel proteins
interact with FMRP and with each other. (PubMed
009259278)
```

"These novel proteins" should be marked up as a gene since it refers to one of the genes of interest ("FXR1") and changing this noun phrase to a mention of another entity would not convey the same relationship.

Only one mention of every gene/disease of interest should be highlighted in each snippet. The annotator should check whether the highlighted regions comply with these guidelines, and correct annotations that do not.

The resulting set of available tags for the annotation is the following:

- GENE: for gene and protein mentions (e.g. RELN, GRM8, WNT2);
- DISEASE: for disease mentions (e.g. autistic disorder, AutD, ASD);
- INTERACTION: minimal, most relevant phrase that supports the Yes/No decision (e.g. "binds to", "phosphorylates").

The complete annotation of the snippets from Table 1 is shown in Table 4.

Detailed annotation guidelines and more annotation examples are provided at the Sourceforge.net project site.

2.2. Annotation tool

Since the task of annotating the snippets will be carried out simultaneously by many annotators, we have implemented an annotation tool with the following features (see Section 2.2.2 for more details):

- Support for parallel and simultaneous annotations by different users.
- Annotator management. The system tracks all the annotations being made by every user. Also, the system allows anonymous annotations.
- Distribution of annotating tasks among annotators. When an annotator logs in the system and requests a snippet to annotate, he is presented with a snippet he has not previously seen from the pool of documents pending annotation. Once the user annotates one snippet, this snippet will never be served again to the same user. To insure the quality of the resultant corpus, we require for every snippet that at least k annotations (performed by different users) significantly agree. More details are provided in Section 2.2.1.
- Access to the annotation system is available from any computer with internet access and a modern web browser. Annotators do not need to install any extra software. Our browser-based system allows the annotators to log in the system from any machine and add new annotations at any time.
- Freely available software. Users can contribute to the annotation of the current corpus hosted on our servers or download the software to annotate their own corpus.

2.2.1. Distribution of snippets among annotators

When an annotator logs in the system and requests a snippet to annotate, he is assigned a new one from the pool of documents pending annotation. The assigned document is picked at random from the documents not previously annotated by this user. Each snippet is annotated by at least k different annotators. If the k annotations of a snippet do not meet a minimum degree of agreement, the snippet is presented to another annotator at random. The process continues until at least k annotations performed on the snippet meet a minimum degree of agreement (see Fig. 2). We have initially established $k = 2$ for the current annotation effort. However, this value can be increased when more annotators join the effort.

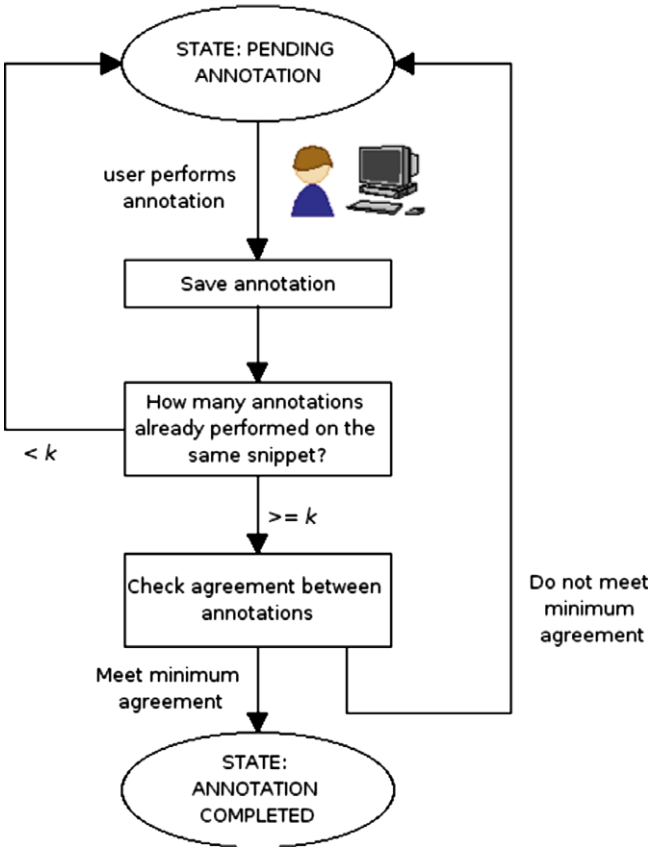

Fig. 2. State diagram representing the annotation of a snippet: from state pending annotation to state annotation completed. The annotation of one snippet is completed when it has been annotated by at least k different users and at least k of these annotations meet a minimum agreement. When the snippet reaches the state annotation completed it will not be served again for annotation.

Table 4
Snippets from Table 1 with annotations. Green indicates tokens labeled as GENE or DISEASE and yellow indicates INTERACTION. Note that in snippet (3), the two entities of interest (*D1* and *D2*) co-occur with no explicit interaction (either positive or negative) being reported in the text. Therefore, the answer to the proposed question is'No' and **nothing** needs to be highlighted as INTERACTION in this case.

1	Does this snippet imply a direct interaction between the provided entities?: Yes. The action of SCPA enzymatically inhibits the chemotactic activity of C5a by cleaving its neutrophil binding site. [PMID: 12964111]
2	Does this snippet imply a direct interaction between the provided entities?: No Three promoter, one intronic, and one 3′ UTR single nucleotide polymorphisms (SNPs) in the APOE gene [...] as well as the APOE functional polymorphism (E2, E3, E4) were examined and failed to reveal significant evidence that autism is associated with APOE. [PMID: 14755445]
3	Does this snippet imply a direct interaction between the provided entities?: No While stimulation of the D2 receptor increased branching and extension of neurites, stimulation of the D1 receptor reduced neurite outgrowth, suggesting that hormones and neurotransmitters may be capable of controlling the development of specific types of neurones.

k given annotations are said to meet the minimum agreement if they satisfy the following three conditions:

1. The Yes/No answer is the same.
2. The token sequences highlighted with labels GENE and/or DISEASE completely overlap.
3. The token sequences highlighted with label INTERACTION overlap (up to 1 different token with respect to the shortest highlighting is allowed between every pair of the *k* annotations).

For example, consider again the snippet (1) from Table 1. If Annotator1 highlights "inhibits" as INTERACTION and Annotator2 highlights "inhibits the activity of" with the same label, the two annotations would agree in terms of the interaction phrase since none of the tokens from the shortest interaction phrase ("inhibits") are different from those in the largest interaction phrase ("inhibits the activity of"). If a new annotator, Annotator3, highlights "enzymatically inhibits", this would also agree with both Annotator1 (same reason above) and with Annotator2: there is only one token ("enzymatically") from the shortest interaction phrase (by Annotator3) which is not included in the largest interaction phrase (by Annotator2). If a new annotator, Annotator4, highlights "action of SCPA enzymatically inhibits" as interaction, this annotation would not agree with that of Annotator2, but would agree with Annotator1 and Annotator3 .

2.2.2. Technical features of the annotation tool

Our annotation tool is a web-based client/server platform implemented in Javascript. On the client side, the application consists of an intuitive user interface where snippets are displayed and the user can perform annotations on the snippets by highlighting arbitrary chunks of text and assigning any of the available labels: GENE, DISEASE or INTERACTION to them. A snapshot of the application is provided in Fig. 4. Some technical details about the user interface are:

- The user can mark-up any arbitrary chunk of text from the snippet. This highlighting can extend across any HTML tags, for example, it can start in a paragraph and end in another one.
- There are two different colors for the highlighting: gene/disease (green) and interaction (yellow). Two or more highlightings can overlap. In this case, the color of the common area is a combination of the colors of the overlapping selections.
- The panel in the right margin of the snippet allows the annotator to keep track of his highlighted selections in the current snippet (see Fig. 4). Each selection has an associated entry in this panel. This panel also allows the user to delete any annotation.
- The small panel in the top allows the user to log in at any time. The user can also perform the annotations anonymously. The system records which user has performed every annotation.

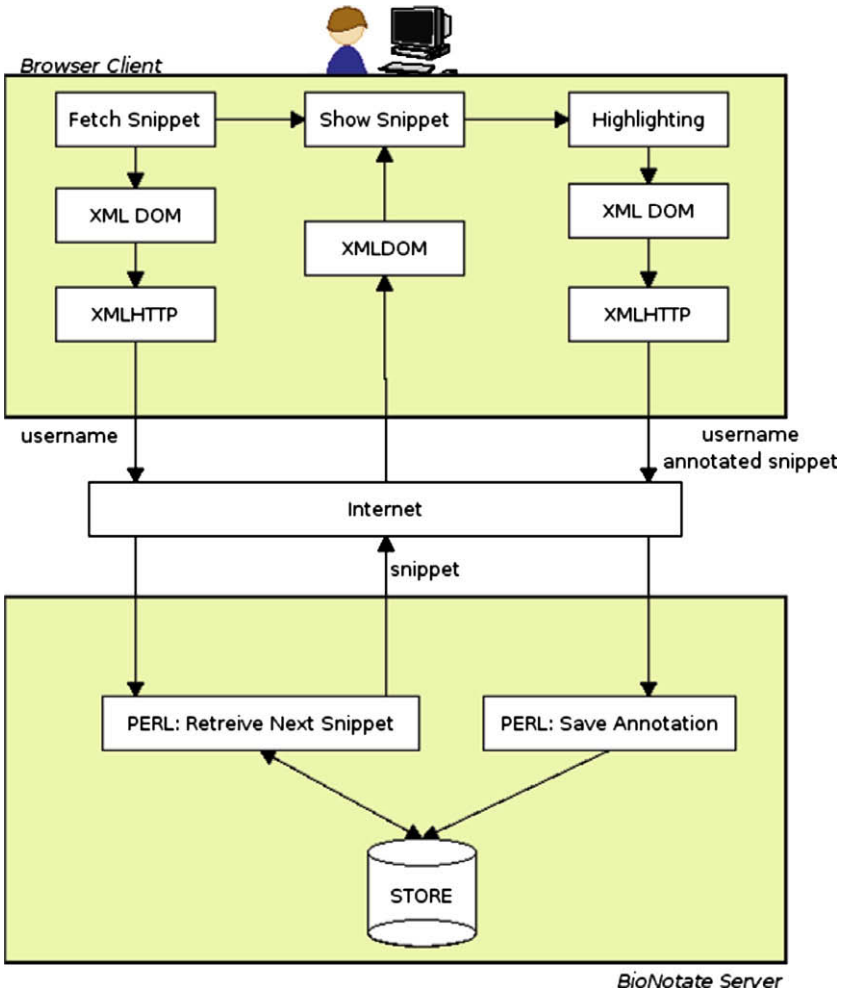

Fig. 3. Components and communications in the annotation tool. The client side iterates in the loop: Fetch the next snippet for the current annotator—present it to the user—allow him to add the annotations and save them. The process continues until the annotator closes the browser window. Each of these modules in the client side communicates with the server side. On the server side, one CGI Perl script serves snippets to the client browsers and another script attends requests for saving annotations.

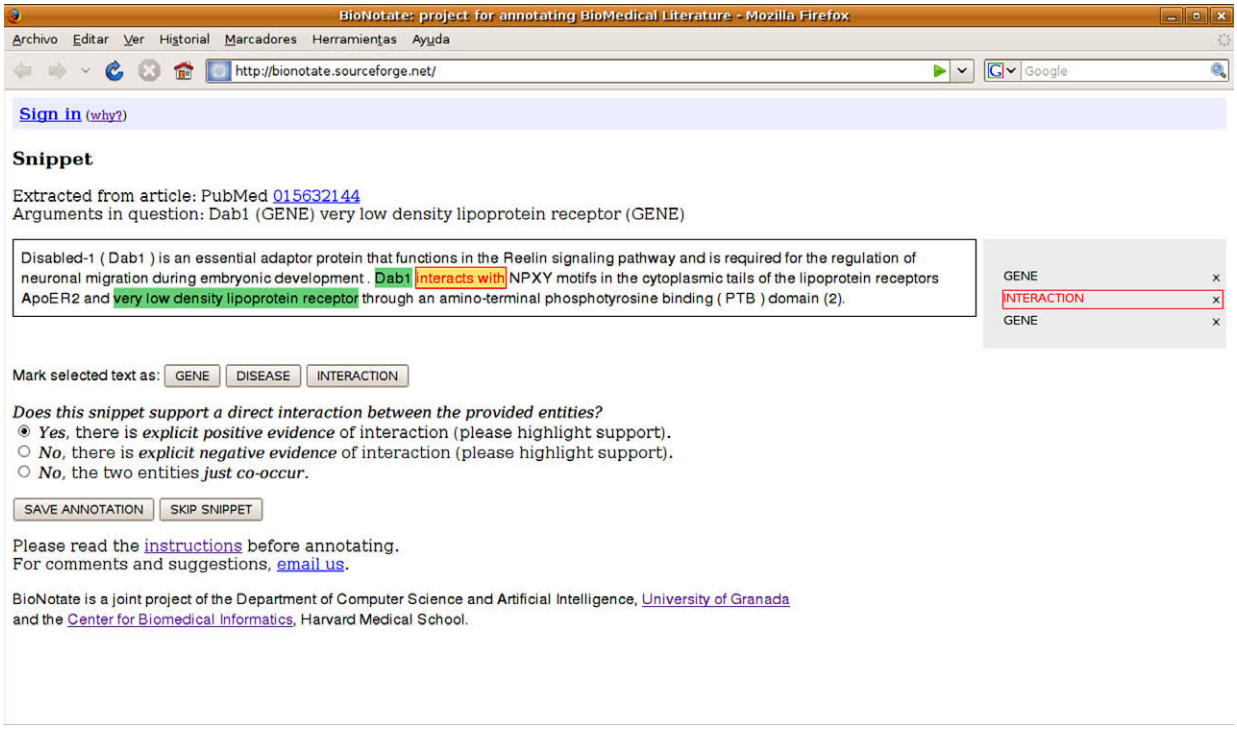

Fig. 4. Snapshot of the annotation tool shows two entities and a relation.

- The user can skip the annotation of a snippet if the text is confusing or he is not confident about his answer.
- The system is fully functional in variety of different browsers and has been tested in *Firefox*, *Internet Explorer*, *Opera* and *Safari*.

The text highlighting and the management of the annotations in the user interface requires intensive use of Javascript and the *Data Object Model* (DOM). The user interface was built using code from the open source project *Marginalia* [38].

The server side is implemented in Perl and consists of two scripts. The first script serves snippets not yet annotated by the user. The second script saves user's annotations and checks whether annotations by different users meet the minimum level of agreement. Both the original (unannotated) snippets and the stand-off annotations are stored as XML files, so no database support is required. A list of snippets pending annotation with the information of which users annotated them is stored in a plain text file that the Perl scripts read and modify as the annotations are completed. The communication between the client and the server is implemented using asynchronous JavaScript and XML (AJAX) requests.

The system works as follows. When an annotator visits a BioNotate application, the javascript client code requests a new snippet to annotate from the server. In that request, the username of the annotator (or 'anonymous' if the user chose not to log in) is also provided. The first Perl script running on the server receives this request and returns a snippet not yet annotated by that user, if one is available. When the user's browser receives the snippet, it is rendered in the annotation user interface so the annotation can be performed. Once the user annotates the snippet and confirms that he wants to save the annotation, the client sends the annotated snippet to the server. The second Perl script manages this incoming annotated snippet. The script stores the snippet as XML and confirms agreement or disagreement with prior annotations of the same snippet, as described in Section 2.2.1. A schematic diagram of this process is shown in Fig. 3. Since the annotated snippet is labeled with the annotator identifier, the snippets annotated by a particular user can be retrieved easily when the corpus is provided to the community.

Fig. 5 shows the information flow in and out of BioNotate. The system must be fed XML-formatted snippets in which two entities of interest have been identified and marked. The resultant annotations performed by the users are also available in the form of XML files. The XML formats used for the original snippets and the annotations can be seen in the figure. The system also generates a plain text file with a list of references to the annotations that agree for every snippet. A full description of the XML formats used by BioNotate and step-by-step configuration instructions can be found in the project webpage.

All the components of BioNotate are fully available in the project webpage, including the javascript software that performs the annotations in the clientside and the Perl scripts managing the annotations and user's request on the server. The scripts for formatting the snippets to be loaded into BioNotate are also available.

3. Case study: pilot corpus on autism

As an example of the use of the BioNotate system, this section presents a pilot effort to annotate a corpus of interactions between genes related to autism. It provides a description of the methods for creating the corpus and the first results of the annotation effort.

3.1. Sources and methods for the creation of a pilot corpus

Our main source of data is PubMed, a widely used biomedical information search tool which includes over 16 million citations and abstracts in its database: MEDLINE. As we previously mentioned, a search of PubMed for extended genes, proteins or diseases returns a huge number of results. To narrow our search for papers reporting protein–protein or protein–disease interactions, we have used publicly available tools and databases such as STRING [35].

STRING is a database of known and predicted protein–protein interactions, which are mainly derived from PubMed. The input is one or more protein names, for which STRING returns a graph where the nodes are proteins and the edges are the relationships between pairs found in the literature. For every edge in this graph, the list of publications which support this relation is provided.

Our goal with this case of study is to build a disease evidence network for autism. Therefore, for the creation of a pilot corpus

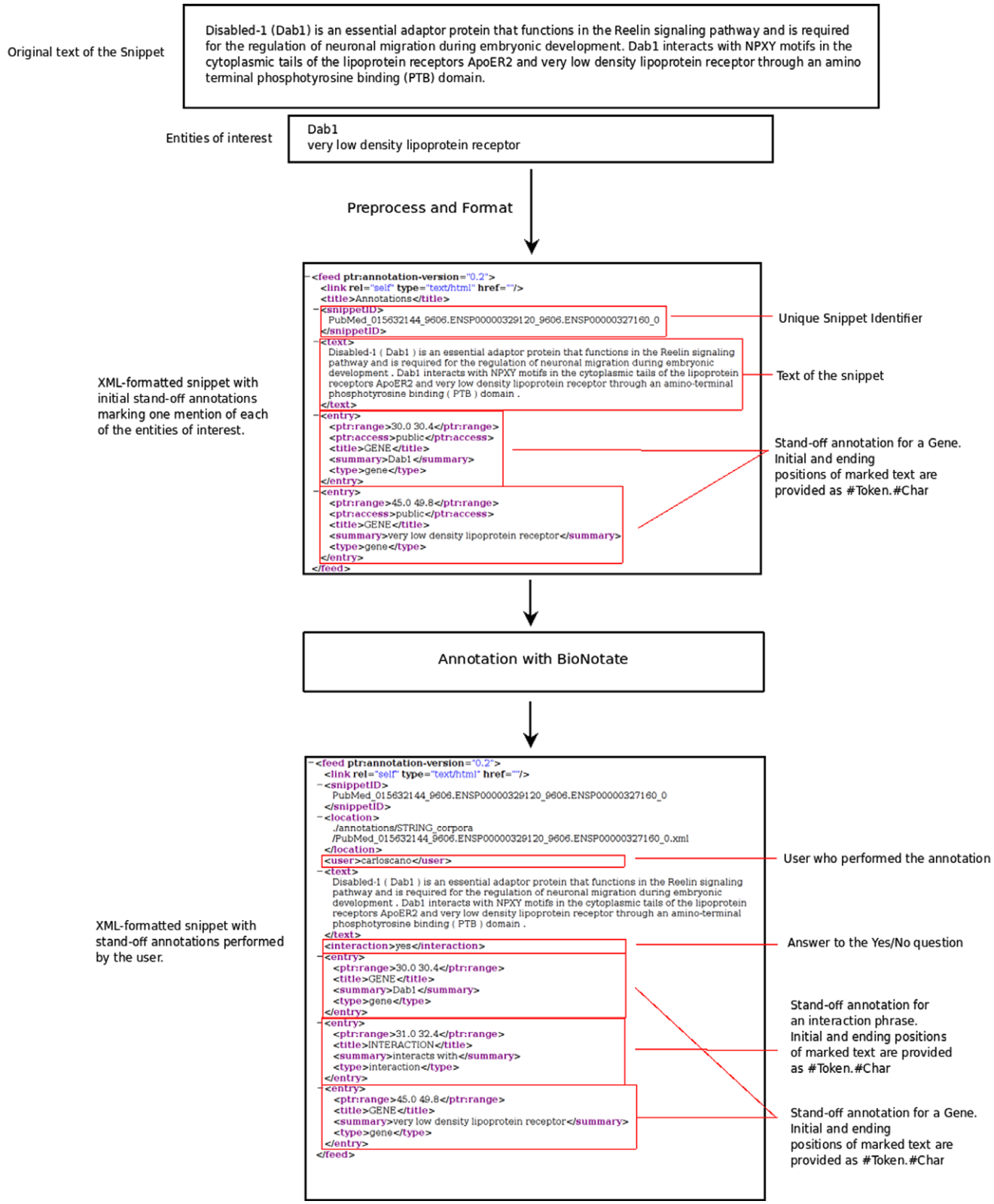

Fig. 5. Information flow in and out of BioNotate. XML-formatted snippets with two marked entities of interest are loaded to BioNotate. Resultant stand-off annotations are also provided in XML format. A full description of the XML formats can be found in the project site.

to start our annotation effort we are focusing on genes and proteins involved in autism and the relationships between them.

To obtain the pilot corpus, we proceeded as follows:

1. We queried OMIM [36] and GeneCards [37] for genes and proteins associated with the keyword "autism". We then generated a combined list by taking the union of the two result sets.
2. We then queried STRING[1] to retrieve all the publications supporting a relationship between every pair of protein names from the list obtained in the previous step. Specifically, the information obtained from STRING for each one of these publications is the PubMedID, the Ensembl Peptide IDs of the two proteins of interest, the text of the abstract and all the mentions of the two proteins in this text. Since STRING already performs a search and identification of the entities of interest in the text, we did not perform any additional Named Entity Recognition (NER).
3. From every abstract supporting a relationship between every pair of proteins, extract all the interesting snippets (chunks of text) which can directly support the interaction. The process of extracting snippets from text is detailed in Section 3.2. The extracted snippets constitute the pilot corpus.

The resultant corpus contains snippets supporting 168 relationships between 127 proteins. A total of 2053 abstracts was processed, yielding 1819 snippets.

3.2. Algorithm for extracting snippets from text

Once all the mentions of the genes/diseases of interest in a text have been identified (by STRING in our case), we create one snippet for every pair of mentions of different genes/diseases which are close enough to each other in the text. Each snippet will contain the text in between the two mentions, and a small amount of text before and after to provide the annotator with some context around the segment of interest. The following schema describes the steps of the algorithm:

1. Retrieve two entities of interest: X,Y within MAX_LENGTH_CORE tokens.
2. All the text in between X and Y is included in the SNIPPET.
3. Extend SNIPPET from X back to the start of the sentence containing X, and from Y forward to the end of the sentence containing Y, up to MAX_LENGTH_TOTAL tokens.
4. If the length of the SNIPPET is under MIN_LENGTH_TOTAL, extend SNIPPET back to the previous sentence and forward to the following sentence, without exceeding MAX_LENGTH_TOTAL tokens.
5. return SNIPPET.

Therefore, the maximum length for a snippet will be given by MAX_LENGTH_TOTAL, and MAX_LENGTH_CORE will provide the maximum numbers of tokens allowed between the two entities of interest in the snippet. We also use the constant MIN_LENGTH_TOTAL to guarantee that all the snippets have enough context for a better comprehension in the annotation process, and thus achieve a higher annotation accuracy. The establishment of values for these constants depends on the type of text being analyzed. We have established experimentally that a MAX_LENGTH_TOTAL of 300 tokens is appropriate for snippets extracted from biomedical texts, with MAX_LENGTH_CORE of 240 tokens and MIN_LENGTH_TOTAL of 40.

In case there is more than one mention of the entities of interest in a snippet we marked-up the two mentions that occur nearest each other to create the snippets that were loaded into BioNotate.

[1] STRING version 6.3

3.3. Results of the pilot corpus

As part of this publication we provide the resulting corpus of our pilot annotation effort on literature related to autism. To date, it consists of one thousand snippets annotated by one of the authors, though we expect it to grow rapidly as we involve multiple annotators. The resulting corpus consists of archived original snippets as well as marked-up snippets in XML format and some post-processing, namely compiled lexicons of all 200 entities encountered in the corpus along with all quotes and supporting relations.

Our annotation reveals that only 116 of the original thousand snippets contain positively identified relationships, i.e. there is roughly 89% error rate on the text support, assuming that all abstracts were meant to identify a positive relation in the STRING database. Not all 200 entities are in fact distinct, e.g. synonymous ways to address the same gene (VLDL-R, VLDLR, VLDLr) and (5-HT-2A, 5-HT2A, 5HT2A) were not merged. As for the phrases supporting the relationships, we encountered many action verbs e.g. "associated with", "docks to", "binds", "phosphorylated by". Naturally, there were many inconclusive cases like "probably unrelated genes", "may interact with" and "little is known about". Phrases supporting an interaction span anywhere from 1 to 28 words, with an average of about 4 words.

In order to evaluate the agreement between different annotators we performed a test on a reduced corpus. For this test, we selected the snippets for which the previous annotator made "interaction" highlightings, i.e. we selected the snippets which explicitly supported either a positive or a negative relationship between the entities of interest (according to that single annotator). We focused on this collection of snippets because they potentially supported interactions, since one annotator already reported so, and therefore they were an interesting test of the effectiveness of our approach. The resultant corpus contains 139 snippets. We involved three more annotators to complete the annotation of this corpus, with the goal of finding agreement among two annotators for every snippet according to the criteria presented in Section 2.2.1.

The results are shown in Table 5. Previous annotation efforts on gene identification and normalization reported agreement rates ranging from 91% to as low as 69% for certain contexts [44]. In our case, the averaged percent of agreement per annotation is over 75% and the task involves annotating interacting entities and interaction keywords in the snippets. This agreement rates are thus similar to other annotation tasks and show that the approach we propose is effective for collaborative annotation.

Disagreement analysis revealed some errors inadvertently introduced by the annotators, such as a negative answer accompanied by a highlighted interaction clearly implying a positive relationship. Another frequent reason for disagreement was the presence of several distinct interaction phrases in the same snippet. For example:

"The KH domains of FXR1 and FMR1 are almost identical, and the two proteins have similar RNA binding properties in vitro.

Table 5
Number of annotated snippets (N Annotations), number of annotated snippets with agreement and % of agreement per annotator according to the agreement criteria specified in Section 2.2.1. The size of the corpus is 139 snippets.

Annotator	N. Annotations	N. Annotations with agreement	% agreement
1	139	94	0.676
2	138	111	0.804
3	48	38	0.792
4	44	35	0.795
Total	369	278	0.753

However, FXR1 and FMR1 have very different carboxy-termini. [. . .] These findings demonstrate that FMR1 and FXR1 are members of a gene family and suggest a biological role for FXR1 that is related to that of FMR1".

Sometimes long interaction phrases can also cause disagreement among annotators, e.g.:

"By immunoblotting, we found that a marked reduction in FMRP levels is associated with a modest increase in FXR1P" (PubMed 012112448).

"No association between the very low density lipoprotein receptor gene and late-onset Alzheimer's disease nor interaction with the apolipoprotein E gene in population-based and clinic samples" (PubMed 009181358).

Another source of disagreement is the highlighting of pronouns and noun phrases that refer to one of the entities of interest according to the guidelines provided in Section 2.1.2. For example, consider the following sentence:

"The biological role of the very low density lipoprotein receptor (VLDL-R) in humans is not yet elucidated. This cellular receptor binds apolipoprotein E (apoE)-containing lipoparticles and is mainly expressed in peripheral tissues" (PubMed 009409253).

In this case one annotator highlighted the mention of the gene "very low density lipoprotein receptor" while another two annotators highlighted the noun phrase "This cellular receptor" which refers to this gene and whose replacement with another entity would alter the relationship being expressed.

Since we require at least two annotators to substantially agree on their annotations, inadvertent errors and disagreements are discarded and the resultant corpus with agreement annotations reflects good quality relationships.

This analysis allowed us to additionally improve the annotations guidelines as well as to enrich the documentation available on the project site with illustrative examples.

According to the gold standard established by the annotators, 110 out of the 139 snippets contain a positive relationship, another 8 contain explicit negative interactions and the remaining 19 do not contain any relationship. In the 76% of the snippets with interaction, the interacting entities were those highlighted in advance (the two mentions that occur nearest each other were marked in advance as described in Section 3.2). In another 18% the interacting mentions were not those highlighted in advance, but the literals were the same as those provided. In the remaining 6% the interacting entities were synonyms of those provided (like'methyl CpG binding protein 2', 'MECP2'), pronouns or noun phrases which referred to the entities (like "This cellular receptor").

4. Discussion and conclusions

We have presented BioNotate, an open source resource for supporting distributed collaborative annotation efforts via a simple interface over a standard internet browser-based client/server system. This resource provides a way to create a substantial benchmark corpus for the evaluation and development of automated relation extraction methods for biomedical literature mining. Additionally, we presented a study of existing definitions of relations between genes and suggested a method to merge existing definitions. It is our hope that the resource presented in this paper will enable rapid progress at the intersection of biology, computational linguistics and knowledge representation.

We also described a method for the creation of a pilot corpus focused on the relations between proteins associated with autism. The annotation tool we built can easily consume new sets of snippets, facilitating similar efforts for the creation of new corpora on this or other diseases. The resultant snippets can be contributed at any time to the current annotation effort. Furthermore, since the tool is freely available, it can be downloaded and deployed anywhere, allowing parallel annotation efforts to be undertaken. Our aim is to facilitate many small, distributed annotation efforts whose output could be integrated into a single and uniform resource.

The consistency of the resulting integrated corpus is guaranteed since every annotated interaction must have explicit verbal support in the text of the snippet. Note that our system offers a process to assemble disjoint annotation efforts into a single corpus consistent from the point of view of computational linguistics, but not necessarily from the biological point of view. That is, decisions about the meaning of text should be consistently supported by the annotation, but generally speaking various annotators do not have to be in agreement as to what constitutes an interaction. More stringent consistency requirements however could be observed by re-constituting corpora from disjoint sets coming from groups once such agreement within a group is reached.

There are several ways for our annotation system to evolve as a result of a tight feedback loop between manual annotation, training of automated text mining tools on a larger corpus, and the use of automated tools to facilitate manual annotation. One issue we encountered was the unbalanced constitution of our corpus, since most of the snippets did not support a relation. Currently, we are using more sophisticated tools for the automatic selection of snippets and the identification of named entities to create better-balanced corpora. We are also developing methods to automatically classify snippets using unsophisticated heuristics in order to enrich the annotated material for positive snippets. This is just one form of active learning [39]. Another related improvement is the use of shallow parsers and noun phrase chunkers to learn to hypothesize the phrase supporting the relationship, relying on the user for corrections, rather than selecting it from scratch.

Once a substantial corpus have been assembled, annotated and analyzed, additional annotations will benefit from the existing one. As we learn about typical disagreements between annotators on the same snippet, and ambiguous phrases and acronyms, we will provide helpful on-the-fly tips and shortcuts to the annotator. Future directions include the integration of knowledge from other biological sources into the disease evidence networks, such as gene expression data and sequence-derived knowledge. An additional improvement aimed at extending the use of resultant annotated corpora can be made by adopting the Distributed Annotation System (DAS) protocol [40] to provide the annotated relationships and their text support to other tools and servers.

We also plan to improve the available software and extend it by providing tools for further post-processing the annotated snippets, such as retrieving all the annotations above varying levels of agreement.

In conclusion, while this resource was developed with binary gene-gene and gene-disease relations in mind, we would welcome applications outside of the original domain. In principle, the tool may be used to create annotated benchmark corpora for arbitrary domains, as long as named entities can be reliably identified.

Acknowledgements

C.C. and A.B. are supported by the projects P08-TIC-4299 of J. A., Sevilla and TIN2006-13177 of DGICT, Madrid. L.P. is supported by the Milton foundation. D.P.W. was supported in part by the National Science Foundation under Grant No. 0543480.

References

[1] Lim J, Hao T, Shaw C, Patel AJ, Szabo G, Rual JF, et al. A protein–protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. Cell 2006;125(4):801–14.
[2] Humbert S, Saudou F. The ataxia-ome: connecting disease proteins of the cerebellum. Cell 2006;125(4):645–7.
[3] Giorgini F, Muchowski PJ. Connecting the dots in Huntington's disease with protein interaction networks. Genome Biol 2005;6(3):210.

[4] Shatkay H. Hairpins in bookstacks: information retrieval from biomedical text. Briefing Bioinfo 2005;6(3):222–38.
[5] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 2006;7:119–29.
[6] PubMed. Available from: http://www.ncbi.nlm.nih.gov/pubmed.
[7] Hunter L, Bretonnel Cohen K. Biomedical language processing: What's beyond PubMed? Mol Cell 2006;21:589–94.
[8] Ananiadou A, Mc Naught J, editors. Text mining for biology and biomedicine. Boston and London: Artech House; 2006.
[9] Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Briefings Bioinfo 2007;8(5):358–75.
[10] Rosario B, Hearst M. Multi-way relation classification: application to protein–protein interaction, In: Proceedings of the HLT-NAACL 2005, Vancouver. http://biotext.berkeley.edu/data.html.
[11] Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. In: Proceedings of the 7th international conference on intelligent systems for molecular biology, 1999. p. 77–86. http://www.biostat.wisc.edu/~craven/ie/.
[12] Johnson HL, Baumgartner WA, Krallinger M, Cohen KB, Hunter L. Refactoring corpora, In: Proceedings of the HLT-NAACL BioNLP workshop on linking natural language and biology, 2006. p. 116–117. http://bionlp-corpora.sourceforge.net/picorpus/index.shtml.
[13] Fetch Prot Corpus. Available from: http://www.sics.se/humle/projects/fetchprot/Corpus/Release20051011/.
[14] HIV-1 Human Protein Interaction Database. Available from: http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html.
[15] Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinfo 2005;6(Suppl. 1):S1.
[16] BioCreAtIvE-I task 1A corpus enriched with annotations for interactions between genes/proteins. Corpus available at: http://www2.informatik.hu-berlin.de/~hakenber/corpora/.
[17] Huang M, Zhu X, Hao Y, Payan DG, Qu K, Li M. Discovering patterns to extract protein–protein interactions from full texts. Bioinformatics 2004;20(18):3604–12.
[18] Kim J, Park J. BIOIE: retargetable information extraction and ontological annotation of biological interactions from the literature. J Bioinfo Comput Biol 2004;2(3):551–68. http://bioie.biopathway.org/.
[19] Yapex Corpus. Available from: http://www.sics.se/humle/projects/prothalt/.
[20] Kim J, Zhang Z, Park J, Ng S. BioContrasts: extracting and exploiting protein–protein contrastive relations from biomedical literature. Bioinformatics 2005;22(5):597–605. http://biocontrasts.i2r.a-star.edu.sg/BioContrasts-testcorpus.html.
[21] AIMed corpus. Available from: ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/.
[22] Kim J, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinfo 2008;9:10. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi.
[23] PennBioIE Corpus. Available from: http://bioie.ldc.upenn.edu/.
[24] Yuka T, Yakushiji A, Ohta T, Tsujii J. Syntax Annotation for the GENIA corpus, In: Proceedings of the IJCNLP 2005, Jeju Island, Korea, 2005. p. 222–7. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi.
[25] Lease M, Charniak E. Parsing Biomedical Literature, Proceedings of the IJCNLP 2005, Jeju Island, Korea. http://bllip.cs.brown.edu/resources.shtml.
[26] DepGENIA Corpus. http://www.ifi.unizh.ch/cl/kalju/download/depgenia/.
[27] Pyysalo S, Ginter F, Heimonen J, Bjorne J, Boberg J, Jarvinen J, et al. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinfo 2007;8:50. Available from: http://www.it.utu.fi/BioInfer .
[28] Genic Interaction Extraction Challenge LLL Workshop 05. http://genome.jouy.inra.fr/texte/LLLchallenge/.
[29] Knowtator http://knowtator.sourceforge.net.
[30] WordFreak http://wordfreak.sourceforge.net.
[31] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th anniversary meeting of the association for computational linguistics 2002. Available from: http://www.gate.ac.uk.
[32] iAnnotate. Available from: http://www.dbmi.columbia.edu/~cop7001/iAnnotateTab/iannotate.htm.
[33] Russell B, Torralba A, Murphy K, Freeman W. LabelMe: a database and web-based tool for image annotation. Intl J Comput Vis 2008;77:157–73.
[34] Google™ Image Labeler. Available from: http://images.google.com/imagelabeler/.
[35] Von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, et al. STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acid Res 2007;35.
[36] OMIM. Available from: http://www.ncbi.nlm.nih.gov/omim/.
[37] GeneCards. Available from: http://genecards.org/.
[38] Marginalia Web Annotation Project. Available from: http://www.geof.net/code/annotation.
[39] Thompson CA, Califf MA, Mooney RJ. Active learning for natural language parsing and information extraction. In: Proc 16th Int Conf on Machine Learning, 1999.
[40] Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L. The distributed annotation system. BMC Bioinfo 2001;2:7.
[41] Maier H, Dohr S, Grote K, O'Keeffe S, Werner T, Hrabe M, et al. WikiGene: identifying problem-related key players of gene regulation using publication abstracts. Nucleic Acid Res 2005;33:W779–82.
[42] Baral C, Gonzalez G, Gitter A, Teegarden C, Zeigler A. CBioC: beyond a prototype for collaborative annotation of molecular interactions from the literature. Comput Syst Bioinfo Conf 2007;6:381–4.
[43] Mons B et al. Calling on a million minds for community annotation in WikiProteins. Genome Biol 2008;9:R89.
[44] Colosimo ME, Morgan AA, Yeh AS, Colombe JB, Hirschman L. Data preparation and interannotator agreement: BioCreAtIvE Task 1B. BMC Bioinfo 2005;6:S12.