Complex-to-Pairwise Mapping of Biological Relationships using a Semantic Network Representation

Juho Heimonen,[1] **Sampo Pyysalo,**[2] **Filip Ginter**[1] and **Tapio Salakoski**[1,2]
[1]Department of Information Technology, University of Turku
[2]Turku Centre for Computer Science (TUCS)
Joukahaisenkatu 3–5
20520 Turku, Finland
first.last@utu.fi

Abstract

This study examines representations of protein–protein interactions focusing on the mapping between simple, pairwise annotation and complex, structured annotation. A simple semantic network representation equivalent to the BioInfer predicate formalism is introduced and used to transform the complex annotation of BioInfer into pairwise annotation through hand-written rules. Evaluation shows that this binarisation can be largely validly performed with limited loss of information, but also reveals specific challenges. The binarised BioInfer is the first corpus of this type where the inclusion rules are formalised to the level of a computational implementation and is freely available at http://www.it.utu.fi/BioInfer.

1 Introduction

The identification of protein-protein interactions (PPI) from free text is one of the most important and widely studied information extraction tasks in biomedical natural language processing. Automatic PPI extraction would benefit a wide range of applications, from advanced search engines to automated pathway database construction.

The great majority of PPI extraction methods and annotated corpora have cast the task as one of identifying pairs of protein names for which some relationship is stated. While the simplest case of extracting unordered pairs is the most widely studied, approaches targeting e.g. ordered pairs or pairs with a connecting relationship type (e.g. Ding et al. (2002), Nédellec (2005)) have also been published, as have some methods for extracting n-ary (for $n > 2$) relations (McDonald et al., 2005). However, pairwise approaches remain the norm and the information extracted by these constitutes only a small part of the knowledge in biomedical literature.

Recently two corpora that contain PPI annotation considerably more detailed than pairwise relations have been introduced. These resources, the BioInfer (Pyysalo et al., 2007) and GENIA Event (Kim et al., 2008) corpora, aid the development of extraction systems that capture complex PPI—here, understood to refer to n-ary interactions of proteins and to include also structured (nested) relations where, for example, a protein affects the interaction of other proteins. This paper explores the relationship between this type of complex annotation and the prevailing pairwise annotation.

First, it is argued that a representation capable of capturing the core of information in complex relationships while remaining practical to extract is needed in complex PPI extraction. In this paper, protein relationships are represented as semantic networks. Since they are based on the BioInfer annotation, these networks follow the textual expressions of the statements of those relationships and are capable of expressing complex PPI. While this is not a fully formal knowledge representation, it aims to support automatic, consistent derivation of simpler, more easily extracted targets and serve as a practical intermediate between textual expressions and formal biological knowledge.

Second, the representation is applied together with a transformation ruleset tailored for the task of transforming the complex relationships in the BioInfer corpus into typed binary (i.e. pairwise) relationships where the types preserve considerably more information regarding the nature of

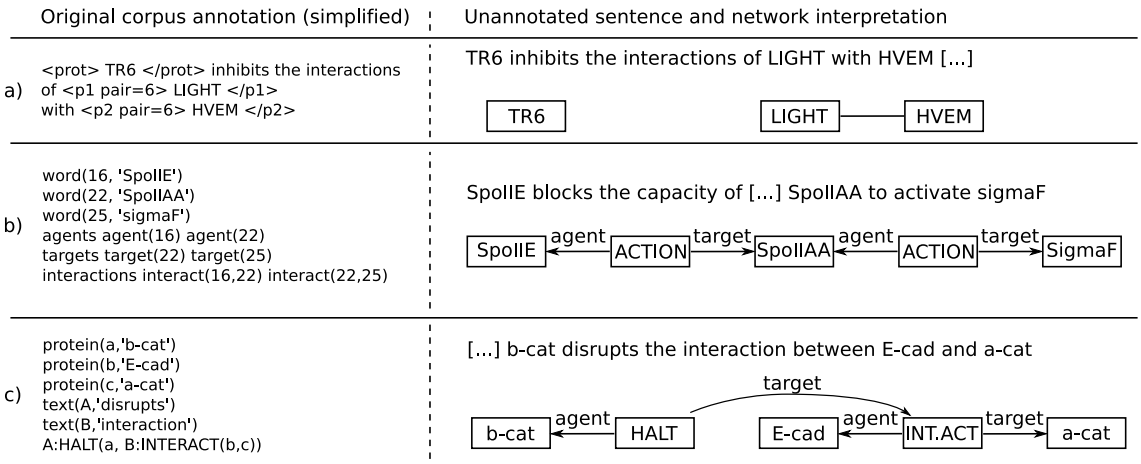

Figure 1: Examples of annotation in a) AIMed, b) LLL and c) BioInfer as semantic networks. Note that the original annotation does not include this representation.

PPI than simple protein pairs. This transformation aims to capture all (and only) biologically meaningful relationships in the original annotation. The transformation is evaluated in a detailed analysis where the magnitude and properties of the information loss necessarily entailed by such a simplification is further discussed along with its significance to the PPI extraction task.

2 Representing biomedical knowledge

The PPI annotation schemes in most domain corpora aim at capturing simple facts about proteins rather than serving as a knowledge representation in the sense of a computable model that supports deductive inference.

Figure 1 illustrates the information contents of annotations in the AIMed (Bunescu et al., 2005), LLL (Nédellec, 2005) and BioInfer corpora as informal semantic networks[1]. AIMed and LLL model interactions as pairwise relationships while BioInfer allows complex relationships. Furthermore, AIMed is not annotated for direction or type while LLL and BioInfer are. The key limitation of pairwise relationship annotation is its incapability to express complex structured relationships. Thus, the annotation involves decomposition that leads to approximations and loss of information. For example, in the LLL annotation in Figure 1b, the effect of SpoIIE on sigmaF is not explicitly annotated and cannot be inferred from the annotation shown in the figure, which is indistinguishable from the annotation that would be given, for example, to *SpoIIE activates SpoIIAA which binds SigmaF*.

In addition to loss of information, the decomposition can lead to inconsistencies. There is large variation in annotation principles (see e.g. Pyysalo et al. (2008)) which evidently leads to annotation of a variety of interaction types across domain resources. For individual corpus annotation efforts, inconsistencies in decomposition principles may contribute to low inter-annotator agreement (see e.g. Alex et al. (2008)).

Despite the limitations of pairwise annotation, pairwise relationships may be necessary in applications such as querying for interactions between two proteins. Assuming that complex relationships are a useful target for information extraction efforts and that simple relationships have benefits in post-extraction applications, a mapping from complex to simple relationships is needed. Further, significant challenges still remain even in pairwise PPI extraction (Krallinger et al., 2007), and while carefully hand-crafted systems extracting complex PPI have been introduced (Friedman et al., 2001), reliable machine-learning approaches to complex PPI extraction may not emerge in the near future. A reliable mapping of the BioInfer and GENIA annotations to pairwise annotations would thus serve to increase the applicability of these resources to presently available extraction methods.

[1]Note that not all the information in Figure 1 is explicitly represented in the corpora: for example, interaction types in LLL are found as comments in the corpus file.

3 Methods and resources

3.1 Corpora

BioInfer was the first domain corpus to introduce the annotation of complex protein relationships. It consists of 1100 sentences annotated for protein names, their relationships, and dependency syntax and uses a predicate formalism in its PPI annotation (see Figure 1c). The GENIA event corpus contains similar annotation, but its relationship annotation of 1000 PubMed abstracts was published late during the present study, which thus focuses on the BioInfer corpus. The essential features of the PPI annotation of the BioInfer and GENIA corpora are largely identical: complex relationships are annotated, participants in relationships are not restricted to protein names but refer to the actual participants even when these are e.g. abstract entities such as *gene expression*, and the annotation is fully bound to the text. Therefore, the methods described in this paper could well be applied to GENIA in a future study.

3.2 Semantic network representation

The term *semantic network* can refer to a variety of graphical representations of knowledge which differ in expressive strength and complexity. A graph representation is a natural choice for semantics, and several well-developed and powerful formalisms have been introduced (Sowa, 1976; Mel'čuk, 1988). However, their complexity makes them difficult targets for automated extraction. An ideal representation for PPI extraction would be as simple as possible, yet capable of capturing all PPI statement types in natural language, and formally well-founded.

In the context of this paper, a semantic network is understood to refer to a directed graph in which the nodes represent biological concepts and the edges represent the stated roles of these concepts. As the applied networks derive from the BioInfer predicate annotation, the graphs are further acyclic, that is, DAGs. The nodes are bound to their corresponding textual expressions through *text bindings* following the original BioInfer annotation. A relationship is defined as a directed subtree with at least two leaves, and a relationship composed of an entire subtree rooted at a source (DAG "root") is termed a complete relationship. In this model a binary relationship is defined as a relationship containing exactly three nodes, two of which are leaves, and a complex relationship is one that is not binary.

type	meaning
agent	agent in an asymmetric process
patient	patient in an asymmetric process
participant	participant in a symmetric process
sub	substructure or member
super	superstructure, family or group
identity	identical entities
possessor	possessor of a property

Table 1: Edge types used in the semantic network.

The nodes and the edges in the network can represent any concept of interest and any semantically sound role, respectively. However, the set of valid edge types is restricted by the type of the predecessor. For example, *actin* (a physical entity) can have an agent or patient role in *depolymerisation* (a process) but not in *filaments* (another physical entity). A controlled vocabulary or, ideally, an ontology must be employed to accurately and formally express the knowledge.

A predicate representation such as that of BioInfer can be directly mapped into an equivalent semantic network where the node types correspond to predicates and their arguments and the edge types only distinguish between the argument positions (1^{st}, 2^{nd} etc.). In case of BioInfer, the node types thus correspond to types in the BioInfer ontologies. Further, edge types (shown in Table 1) are indirectly obtained from the description of the nesting and the predicates (see Section 3.3.1). Thus, the network representation can capture the same general set of biomedical relationships as the original BioInfer annotation. However, the network representation has several practical advantages over the predicate representation of BioInfer. Biological concepts, which can be either physical, such as molecules or cell components, or abstract, such as processes, properties or relationships, are represented in a unified manner, unlike in the predicate representation that differentiates between predicates (relationships) and entities. Further, the participant roles are explicitly represented, facilitating processing of relationships. Finally, the network representation is naturally extensible: for example, information regarding cell type could be added simply by attaching additional edges to the network.

Figure 2 provides an example of a semantic network that uses the BioInfer ontologies.

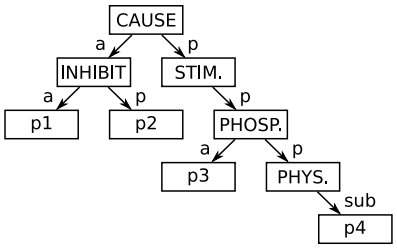

Figure 2: Example of a semantic network representing the sentence *Inhibition of B by A causes stimulation of phosphorylation of D filaments by C. Agent* is abbreviated as *a* and *patient* as *p*.

The fact that no agent is stated for the node *STIM.* (STIMULATE) renders this particular relationship unexpressable in the BioInfer formalism without adding an anonymous entity.

3.3 Binarisation process

Here, binarisation is defined as a process of mapping a complex relationship into a set of (typed) binary relationships, aiming at sound (valid, truth-preserving) inference as well as to preserve the key biological information of the original relationship. This is achieved through a corpus-specific set of hand-written inference rules. Instead of formal inference (as understood in logic) aiming at finding new (unstated) knowledge, the purpose of the inference rules is to reduce original annotation into binary annotation by applying transformations that generate the most accurate approximation of the original information content.

The validity of inference is evaluated with respect to biologists' understanding of whether the generated binary relationships describe relations stated in the text. Ideally, the binary annotation includes all (and only) pairwise PPI that are biologically relevant, along with appropriate types. Note that not all protein pairs forming a relationship generate biologically relevant binary relationships: for example, no such relationship can be validly inferred between p_1 and p_3 from the statement p_1 *prevents the phosphorylation of* p_2 *by* p_3. By contrast, for p_1 *prevents the binding of* p_2 *to* p_3, a p_1–p_3 relationship could be inferred because *bind* is a symmetric relationship.

Before binarisation, the semantic network is preprocessed to simplify the binarisation process and to separate the binarisation from refinement of relationships.

3.3.1 Preprocessing of the network

The BioInfer corpus contains annotation for a number of non-biological relationship types, such as equality and coreference, which are used to detail the expression of other, biological, relationships. Non-biological relationships are excluded from the binarised corpus. However, to preserve as much biological information as possible, these relationships are resolved by graph transformations following their interpretations, as given in (Pyysalo et al., 2007).

For example, in BioInfer the EQUAL predicate is used to express identity relationships, mostly in abbreviations and synonym definitions, and the COREFER predicate is used to express coreference. Only the first argument of these predicates is then used in other relationships, and thus in the network these relationships are introduced for the second argument by copying edges and nodes, as illustrated in Figure 3.

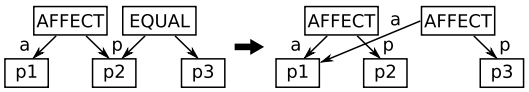

Figure 3: Preprocessing EQUAL predicates. The annotation AFFECT($p1$,$p2$) EQUAL($p2$,$p3$) for the expression $p1$ *affects* $p2$ *(also called* $p3$*)* is preprocessed into AFFECT($p1$,$p2$) AFFECT($p1$,$p3$).

In the BioInfer entity annotation, entities can be nested, i.e. contain other entities: for example, *p1 subunit* is annotated as two entities, *p1 subunit* and the nested *p1*. However, the annotation does not specify the type of the relations implied by nesting. These relations are represented as edges in the network and their types can be resolved reliably by heuristics based on the types and text bindings of the end nodes of the edges. For example, in *[depolymerisation of [[actin] filaments]]* the edge from *depolymerisation of* to *filaments* is resolved into *patient* (rule: physical entity nested in a process with *of* in its text binding) and the edge from *filaments* to *actin* is resolved into *sub* (rule: physical entity nested in larger physical entity). The special predicate REL-ENT, implying indirect nesting, is resolved similarly.

3.3.2 Extraction of binary relationships

Binary relationships are extracted in a two-step process. First, candidate relationships are generated from the original graph by forming all possible relationships with exactly two proteins

as leaves. In order to determine the polarity of the resulting binary relationship, all adjacent nodes of type NOT are included into the relationship. Since the edges are explicitly labeled with roles whose interpretation is independent of other edges, such a subgraph is sufficient to preserve all the details of the relationship between the two selected proteins while being easier to process than the entire graph.

Second, the relationships are transformed with a set of rules that reduce them into binary relationships. Each rule defines a transformation that aims to preserve the information content while simplifying the relationship by removing nodes and/or altering the types of the nodes and edges. Unlike in formal inference, each transformation produces an approximated relationship, and the validity of the inference is not guaranteed. To minimise the overall extent of approximations and to avoid invalid inference, the rules are manually ordered so that more reliable and less approximative rules have priority.

Rules including the root determine the final relationship type and are applied first. Essentially these rules process nodes representing verbs with little semantic content as well as determine the overall regulatory effect. Rules applying to leaves remove nodes whose information content cannot be included in the final relationship, and are applied only if other rules do not match. In most cases, the removed information concerns the details of the exact types of the physical entities. By iteratively applying the first matching rule, each relationship is transformed until a binary relationship is obtained or none of the rules match. The semantic network representing all valid binary relationships is simply the union of the binary relationships obtained in this step.

Figure 4 illustrates the transformation process. In step a), a node representing the verb *cause* is removed. This is a minor approximation since the node (*CAUSE*) indicates that *p1* is (indirectly) an agent in the stimulation process. Similarly, an agent of a regulatory process (*INHIBIT*) causing another process (*STIM.*) is indirectly the agent of that other process. Hence, *INHIBIT* is removed in step b). Step c) is a rearrangement of nodes: a regulatory process (*STIM.*) is processed into the effect attribute (see Section 4.1) of the affected physical process (*PHOS.*). In step d), it is approximated that anything that is stated for a physical entity (*PHYS.*) is also valid for its component (*p4*). In this example the resulting relationship is REG(+)_PHOSPHORYLATE(p1,p4).

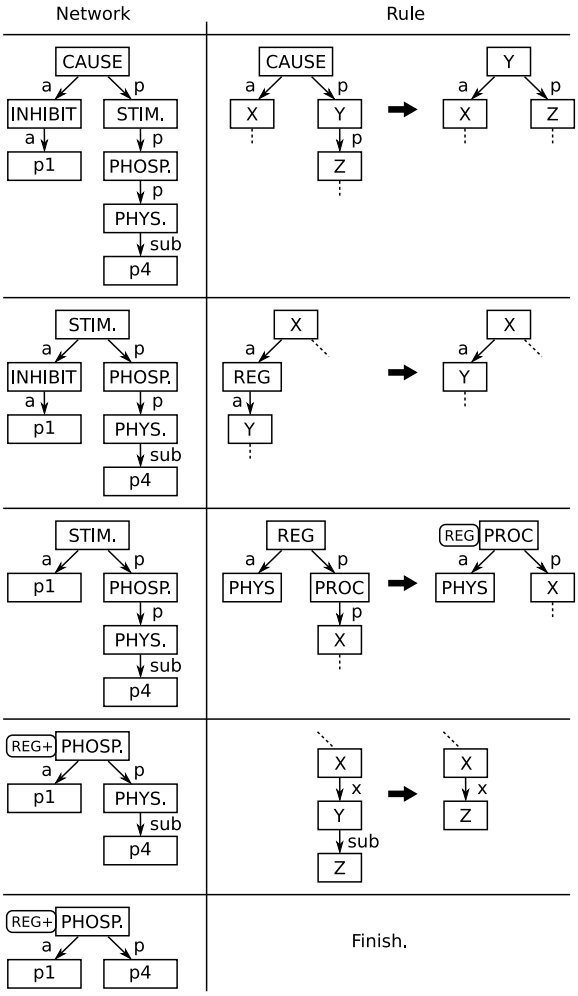

Figure 4: An example of candidate relationship processing. See Figure 2 for description of notation and Section 4.1 for REG(+) attribute description.

3.4 Development and evaluation protocol

In order to be able to fairly evaluate the effect of the binarisation process on previously-unseen data, the software and rules were developed on a random sample of 437 sentences. The process was then applied to the complete BioInfer corpus and all relations in a random sample of 50 previously-unseen sentences of the binarised BioInfer were analysed by a biologist to determine the quality of the binarisation.

In the error analysis, instances of information loss were counted and their causes examined. The losses were categorised as follows, in decreasing order of severity: missing interaction, invalid inference, invalid interaction text binding,

approximated interaction type, and lost interaction detail. The latter two were considered as approximations while the other as errors. The lost interaction details were divided into three categories (process/property, structure/membership, identity) and evaluated by counting the entities that did not contribute to the corresponding binarised relationship.

The applied proof-of-concept software is implemented in Python and Prolog. Any similar programming language or inference tool would be equally good provided that it supports the ordering of the rules and the search for the first sequence (based on the rule order) of transformations leading to a binary relationship.

4 Results and discussion

4.1 Binarisation details

Single BioInfer predicate types are not alone sufficient to summarise complex relationships. In particular, polarity needs to be preserved to separate explicit negative statements, originally annotated with the NOT predicate, from unannotated (i.e. non-existing) statements (see Pyysalo et al. (2007)). In addition, complex relationships can combine aspects of regulation to the primary effect: for example, the annotation for p_1 *suppresses the polymerisation of* p_2 includes both the SUPPRESS and POLYMERIZE types but neither alone is sufficient to express the whole relationship. To make it possible to preserve negation and regulatory aspects, the predicates are augmented with *polarity* and *effect* attributes.

The base predicate specifies the relevant biological process while the effect attribute describes how this process is affected by the agent. The effect can be positive, negative, or unspecified regulation or a direct action. For simplicity, when polarity or effect have their "default" values (positive and direct action, respectively) these are omitted from the augmented predicate: thus, instead of POS_DIRECT_INHIBIT simply INHIBIT is used as the name. Hence, for example, NEG_POLYMERIZE indicates the agent does not polymerise the patient, REG(-)_BIND indicates that the agent negatively regulates the binding of the patient (to an unspecified entity).

The BioInfer ontologies are modified to better support the binarisation as follows. The Process_entity subtree in the entity ontology is mapped to the relationship ontology: for example, the process entity DEPOLYMERIZATION is mapped to the predicate DEPOLYMERIZE. In addition, to be able to determine the effect attribute in the binarisation, relationship types considered regulatory (Dynamics and Amount subtrees and the PREVENT type) were flagged.

4.2 Statistics

This section briefly summarises the key statistics relating to the binarisation. The original BioInfer corpus in the graph representation contains 2662 complete relationships, 942 of which are binary. Note that some of these binary relationships (such as EQUAL) are preprocessed into other relationships. The binarised BioInfer contains 2762 relationships of which 94.4% (vs. 93.9% in the original) have positive polarity and 89.7% direct action effect.

During the binarisation process, the rules matched 4794 times in total: the fraction of rules involving the root is 39.7% and those involving leaves 51.6%. The most applied root-matching rules were those processing CAUSE, regulatory relationship types, and CONTAIN (10.3%, 9.8%, 8.4% resp.) while leaf-matching rules were applied mostly to remove edges of identity (21.3%) or structure/membership (17.3%) types.

The distributions of predicates in the original and binarised BioInfer are clearly different. In the binarised corpus, general predicates (for example PARTICIPATE, AFFECT, and CONDITION) have nearly all been removed while the number of predicates in the Change-subtree has increased 63% even though the number of predicates in its Dynamics-subtree have decreased 25%. The former two observations confirm that the general predicates have been transformed to biologically relevant ones, as intended. The last observation corresponds to the regulatory predicates being reinterpreted as effect attributes.

4.3 Error analysis

Table 2 shows the observed errors and approximations in the sample. For those types that can occur only once per relationship, the expected number per relationship in the binarised BioInfer is shown. For the lost interaction details, the expected number per non-leaf entity in the original BioInfer is shown.

Three of the observed missing interactions are

error type	count	E
missing interaction	7	0.07
invalid inference	13	0.12
invalid interaction text binding	0	0.00
total	20	0.19

approximation type	count	E
approximated interaction type	8	0.08
lost entity (process/property)	9	0.06
lost entity (structure/membership)	15	0.09
lost entity (identity)	7	0.04
total	31	0.19

Table 2: The errors and approximations observed in the analysed sample of the binarised BioInfer. Expectation E for errors and approximated interaction types given per-relationship, other approximations per-relation, where per-relationship expectations refer to the binarised corpus and per-entity expectations to the original corpus.

duplicates of existing interactions. For example, two regulatory relationships would be annotated in the sentence *Actin regulates cofilin phosphorylation and dephoshorylation*, but the binary annotation cannot express the difference and hence produces only one relationship. Another three missing interactions are deliberately removed as self-interactions (which are not relationships in the applied semantic network model). The last missing interaction is due to the failure in nesting role resolution, caused by an invalid nesting in a phrase *actin-bound nucleotide exchange*. The nesting is technically allowed by the BioInfer annotation but the role of *actin* in *exchange* cannot be expressed with a single edge.

For the majority of the observed invalid inferences the cause is an incorrectly identified effect attribute. In six cases, the regulatory effect of a node is missed or falsely assumed. For example, in the sentence *Addition of profilin caused actin depolymerisation*, the process *addition* (annotated as INCREASE) does not refer to positive regulation but rather to an experimental setup. The two other effects are misidentified due to a similar case of nesting as described in the previous paragraph (consider the phrase *concentration required for polymerisation*). In the remaining five cases, the true agent is an unexpressed process while the claimed agent (protein) has an unstated relationship with the patient. This renders the binarised relationship invalid. Consider the sentence *Dephosphorylation of cofilin leads to actin depolymerisation* as an example in which *dephosphorylation* causes *depolymerisation* while the effect of *cofilin* as such on *actin* is unstated.

The expectations for losing information in entities is surprisingly low given that leaf-targeting rules were the most applied. Moreover, since words carrying little biologically relevant information, such as "protein" and "function", are included in these numbers, the biological information loss is even less. The observed approximations in the interaction types are minor, such as the type INITIATE being generalised to positive regulation in mapping to an attribute.

In short, the error analysis reveals some weaknesses of the original BioInfer annotation scheme, especially nesting, while the binarisation fails mostly on identifying a regulatory effect. Given that regulatory relationships are a small minority, the effect attribute could be completely dismissed.

5 Conclusions

This paper has provided the first study of the relationship between the pairwise annotations commonly used to annotate PPI and the complex annotations in recent corpora such as BioInfer and the GENIA Event corpus. A simple semantic network representation was presented, and the BioInfer predicate annotation was mapped into this representation. This mapping unifies some arguably unnecessary distinctions in the original annotation, such as the mirroring of some relationship types with entity types (e.g. PHOSPRORYLATE vs. PHOSPHORYLATION), and explicitly represents all relationships between entities, including relationships whose type is unspecified in the original annotation (e.g. sub/superstructure). The semantic network thus provides a more consistent representation of the relevant information, facilitating rule-based inference.

The binarisation of the BioInfer relationship annotation was implemented as a set of graph transformation rules. This transformation aimed to determine which biologically relevant relationships between two proteins can be inferred from the full semantic network and how much of the original information content can be preserved with BioInfer relationship types augmented with polarity and effect (direct/regulatory) attributes. A study of the resulting binary PPI indicated that

while the original annotation and the chosen representation are, in general, capable of supporting this form of inference, a number of errors were produced in the process. The study of these errors suggested some weaknesses in the original annotation and further indicated that while the existence of relationships was inferred correctly, the effect attribute could not always be reliably determined. The evaluation further provided an estimate of the approximations inherent to binary annotation even when regulatory effects are separately captured.

The results suggest that it is sufficient to summarise the relationships between proteins with a pairwise annotation for use in various applications. However, information extraction could benefit from the details available in complex relationships. Thus, together with the possibility to transform complex relationship into binary ones, the extraction of semantic networks could prove to be a feasible approach to PPI information extraction.

The similarities between the network representation considered here and the conceptual graph (CG) model of Sowa (1976) suggest that the CG model could be adopted as a knowledge representation for PPI extraction. As a well-founded formalism, the CG model would provide a means to robustly express extracted relationships. However, the CG model may need to be adjusted to address the linguistic aspects of information extraction in the biomedical domain.

The created binary BioInfer is the first corpus with pairwise PPI annotation where the rationale for including or excluding a particular pair is formalised to the level of computationally implemented rules. As binary PPI annotation is still dominant in particular in machine-learning-based PPI extraction, this resource can provide valuable data to a field where annotation consistency has been a challenge. Similarly, the semantic network form of the corpus can provide a more approachable target for automatic PPI extraction than the original predicate form. The software tools and the data (in the original BioInfer format) produced in this study are freely available from `http://www.it.utu.fi/BioInfer`.

Acknowledgements

This work has been supported by the Academy of Finland.

References

Bea Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin, and Xinglong Wang. 2008. The ITI TXM corpora: Tissue expressions and protein-protein interactions. In *Proceedings of LREC'08*.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun Kumar Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. 2002. Mining MEDLINE: abstracts, sentences, or phrases? In *Proceedings of PSB'02*, pages 326–337.

Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. 2001. GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl. 1):S74–S82.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

Martin Krallinger, Florian Leitner, and Alfonso Valencia. 2007. Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proceedings of BioCreative II*, pages 41–54.

Ryan McDonald, Fernando Pereira, Seth Kulick, Scott Winters, Yang Jin, and Pete White. 2005. Simple algorithms for complex relation extraction with applications to biomedical IE. In *Proceedings of ACL'05*, pages 491–498.

Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.

Claire Nédellec. 2005. Learning language in logic - genic interaction extraction challenge. In *Proceedings of LLL'05*.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl. 3):S6.

John F. Sowa. 1976. Conceptual graphs for a database interface. *IBM Journal of Research and Development*, 20(4):336–357.