BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 5 2006, pages 597–605
doi:10.1093/bioinformatics/btk016

Data and text mining

BioContrasts: extracting and exploiting protein–protein contrastive relations from biomedical literature

Jung-jae Kim[1], Zhuo Zhang[2], Jong C. Park[1] and See-Kiong Ng[2,*]

[1]Computer Science Division & AITrc, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701 South Korea and [2]Knowledge Discovery Department, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore

Received on September 5, 2005; revised on December 15, 2005; accepted on December 16, 2005
Advance Access publication December 20, 2005
Associate Editor: Alex Bateman

ABSTRACT

Motivation: Contrasts are useful conceptual vehicles for learning processes and exploratory research of the unknown. For example, contrastive information between proteins can reveal what similarities, divergences and relations there are of the two proteins, leading to invaluable insights for better understanding about the proteins. Such contrastive information are found to be reported in the biomedical literature. However, there have been no reported attempts in current biomedical text mining work that systematically extract and present such useful contrastive information from the literature for exploitation.

Results: Our BioContrasts system extracts protein–protein contrastive information from MEDLINE abstracts and presents the information to biologists in a web-application for exploitation. Contrastive information are identified in the text abstracts with contrastive negation patterns such as 'A but not B'. A total of 799 169 pairs of contrastive expressions were successfully extracted from 2.5 million MEDLINE abstracts. Using grounding of contrastive protein names to Swiss-Prot entries, we were able to produce 41 471 pieces of contrasts between Swiss-Prot protein entries. These contrastive pieces of information are then presented via a user-friendly interactive web portal that can be exploited for applications such as the refinement of biological pathways.

Availability: BioContrasts can be accessed at http://biocontrasts.i2r.a-star.edu.sg. It is also mirrored at http://biocontrasts.biopathway.org

Supplementary information: Supplementary materials are available at Bioinformatics online.

Contact: skng@i2r.a-star.edu.sg; park@cs.kaist.ac.kr

1 INTRODUCTION

Contrasts are effective conceptual vehicles for learning processes. For example, in pedagogical learning, contrasts are often used for teaching complex concepts to students. In expository writing, contrastive techniques are also frequently used to highlight, compare and group central concepts. Even in machine learning, the typical goal of many machine learning methods (e.g. support vector machines) is to find the distinguishing contrasts between classes of objects.

Contrastive relations are useful for exploring the unknown as they can provide much invaluable insights into the observed phenomena for guiding further knowledge discovery. For example, contrastive information between proteins in terms of their biological interactions can reveal what similarities, divergences and relations there are of the two contrasted proteins, supplying revealing insights into the underlying functional nature of the proteins that had brought about the differences in their observed biological phenomena. Oftentimes, contrastive information also contain negative information (e.g. in the typical contrastive relations expressed as 'A but not B') that are also found to be highly useful in guiding scientific research (Knight, 2003) and machine learning (Li *et al*., 2005).

In order to dissect the vast and mostly unknown interactomes, biologists have expended much efforts to compile various protein and interaction databases [e.g. Swiss-Prot (Bairoch *et al*., 2005), BIND (Alfarano *et al*., 2005), DIP (Salwinski *et al*., 2004) and KEGG (Kanehisa *et al*., 2004)] from experimental data and the published literature. However, the information captured in these resources are typically individual positive facts of the kind such as 'protein A binds to protein B'. As far as we know, none of the existing protein and interaction databases systematically capture the useful contrastive information mentioned above. In this work, we seek to address this void by extracting contrastive information between proteins from the biomedical literature to augment the information in current protein databases.

Given the inherent complexities in molecular biology and the common usage of contrastive techniques to explain complex concepts in expository writing, it is not surprising that the biomedical literature can be a rich resource for contrastive information. However, current biomedical text mining work (Hirschman *et al*., 2002; Shatkey and Feldman, 2003; Cohen and Hunter, 2004) has not focused much (if at all) on the extraction of contrastive information from the literature. In this paper, we present our BioContrasts system that extracts protein–protein contrastive relations from MEDLINE abstracts and presents the information to biologists in a user-friendly web-application for exploitation of such knowledge.

*To whom correspondence should be addressed.

Currently, we focus on identifying contrastive expressions in the MEDLINE abstracts that contain typical contrastive negation patterns such as 'A but not B'. In order to create a useful database of contrastive relations for proteins, we also link the protein names in the extracted contrasts to Swiss-Prot entries with a name grounding procedure.

The rest of this paper is structured as follows: Section 2 provides further background information about protein–protein contrasts, while Section 3 describes the methodologies employed in our current BioContrasts system. Section 4 describes the implementation of the BioContrasts system and the web portal, and Section 5 shows evaluation results on the system performance. Finally, we explore some uses of the extracted contrasts for refining protein pathway databases in Section 6.

2 BACKGROUND

In our work (Kim and Park, 2005) on mining for negative information in MEDLINE abstracts, we have observed an interesting phenomenon that the negative 'not' often encodes contrasts between biomedical objects in forms of negation patterns in the biomedical literature. Take, for instance, the following sentences (1) and (2) from two MEDLINE abstracts:

(1) NAT1 binds *eIF4A* but not *eIF4E* and inhibits both cap-dependent and cap-independent translation (PMID: 9030685[1]).

(2) *Truncated N-terminal mutant huntingtin* repressed transcription, whereas *the corresponding wild-type fragment* did not repress transcription (PMID:11739372).

In addition to expressing the primary negative information about protein eIF4E, the negative 'not' in the sentence (1) also encodes a contrast between the proteins eIF4A and eIF4E in terms of their abilities to bind to NAT1 through the coordinating conjunction 'but'. Similarly, with a combination of 'not' and the subordinate conjunction 'whereas', the sentence (2) also expresses a contrast in addition to the corresponding negative information—in this case, the contrast is between wild-type huntingtin and its mutant with respect to the biological activity of transcription repression.

Each piece of such contrastive information is made up of two parts: (i) a contrastive pair of two or more objects that are so contrasted (e.g. {eIF4A, eIF4E}, {wild-type huntingtin, mutant huntingtin}), called focused objects and (ii) a biological property or process that the contrast is based on (e.g. binding to NAT1, transcription repression), called presupposed property. We call the contrast a protein-protein contrast and the proteins focused proteins if the two focused objects of a contrast are proteins, as in (1) and (2). We also call the linguistic expressions for focused objects contrastive expressions and those for focused proteins contrastive protein names.

In terms of knowledge representation, the contrasts can be computationally represented as follows: the presupposed property as a variable-containing (or 'open') proposition, as schematized in (3a), and the focused objects as instantiations of the variable, as illustrated in (3b) (Prince, 1992).

(3) a. NAT1 binds to X
b. X = eIF4A, X ≠ eIF4E

The focused objects can be further classified into two groups: (i) the object or objects that are positively involved in the biological event of the presupposed property, called positive objects (e.g. eIF4A, mutant huntingtin) and (ii) those that are not involved in the event, called negative objects (e.g. eIF4E, wild-type huntingtin). Note that for contrasts to be meaningful, there should also exist a certain level of implicit similarity between the contrasted objects in addition to the explicit differences being communicated (Umbach, 2004). For example, the focused objects of the contrast encoded in (1) are all eukaryotic translation initiation factors, while in (2) one of the focused proteins is actually a mutant of the other protein. A contrast is therefore a highly informative unit of knowledge that not only explicitly represents a difference between its focused objects but also implicitly indicates that the focused objects are semantically similar.

3 METHODOLOGY

BioContrasts is an information extraction system that extracts protein–protein contrasts from the literature. Currently, we focus on extracting contrastive information that are encoded by the negative 'not' in the literature. Contrastive expressions are identified from the text based on a two-step approach proposed in our earlier system (Kim and Park, 2005): (i) Matching a subclausal coordination pattern (e.g. 'A but not B') to an input sentence and then (ii) identifying parallelism between a negated clause and an affirmative clause in the sentence or adjacent sentences, as exemplified in the sentence (2).

The current system focuses on identifying protein–protein contrasts from the contrastive information extracted by the earlier system to create a useful database of contrastive relationships between Swiss-Prot protein entries. The extracted information are then presented via a user-friendly interactive web-portal for the exploitation of such informative knowledge. In this paper, we provide detailed descriptions (including pseudo-codes in 'Supplementary Materials') for the information extraction procedures in the context of protein–protein contrast extraction, and we show how such knowledge can be exploited for exploratory research applications to help dissect the vastly unknown interactomes.

Figure 1 shows a sample procedure of the BioContrasts system. Given a MEDLINE abstract, the system first locates sentences that contain the negative 'not'. It then identifies contrastive expressions from these sentences using either subclausal coordination or clause-level parallelism. If the contrastive expressions are, or can be reduced to, protein names, the system produces a contrast between the two proteins. It then cross-links (i.e. grounds) the contrastive protein names with entries of a standard protein database (namely, Swiss-Prot). The net result is a database of useful biological contrastive relations between actual Swiss-Prot entries.

We will describe each step of the procedure further in the following subsections. In addition, Appendix A of 'Supplementary Materials' shows the pseudo-codes of important functions of the BioContrasts system, while Appendix B shows short descriptions of the various subfunctions involved. The pseudo-code of the main algorithm of the system can be found in A.1. The reader is encouraged to refer to the appendices for more rigorous descriptions of the procedures in BioContrasts.

[1]The number following the string 'PMID:' is the identifier of a MEDLINE abstract in PubMed (http://www.ncbi.nlm.nih.gov/PubMed/).

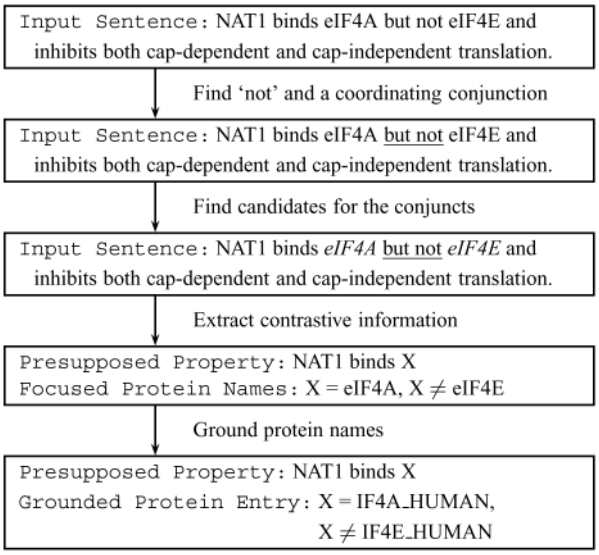

Fig. 1. An example procedure for contrast extraction.

Table 1. Subclausal coordination pattern examples

General patterns	Specific patterns	
A but not B	NP but not NP	PREP NP but not PREP NP
	V NP but not V NP	V PREP NP but not V PREP NP
not A but B	not NP but NP	not PREP NP but PREP NP
	not V NP but V NP	not V PREP NP but V PREP NP
A, not B	NP, not NP	PREP NP, not PREP NP

A and B denote the pair of focused objects in a general subclausal coordination pattern. NP indicates a noun phrase, PREP a preposition, V a verb, and ADJ an adjective.

Table 2. Similarity and hypernymy among verbs in biomedical texts

(a) ActivateVerb: activate, enhance, induce, stimulate, upregulate
(b) InhibitVerb: abolish, downregulate, inhibit, prevent, reduce, suppress
(c) CauseVerb: affect, cause
(d) BindVerb: associate, bind, complex, interact
(e) Hypernymy: ActivateVerb ⇒ CauseVerb, InhibitVerb ⇒ CauseVerb

'Class A ⇒ Class B' indicates that the words in Class A also belong to Class B.

3.1 Identifying subclausal coordinations

Given a sentence that contains a 'not', the system first tries to identify contrastive expressions using subclausal coordination patterns as defined in Table 1[2]. For example, the system matches the sentence (1) to the general pattern 'A but not B' with the variables for focused objects, A and B, matched to 'eIF4A' and 'eIF4E', respectively.

The use of the coordinating conjunction 'but' implies that the coordinated expressions corresponding to the variables A and B should be both syntactically and semantically similar. As such, we formulated more specific patterns as shown in the column on the right of Table 1. Note that the two corresponding syntactic elements in both sides of the specific patterns are always identical (e.g. 'PREP NP' of the pattern 'PREP NP but not PREP NP')—this establishes the syntactic similarity between the coordinated expressions. The syntactic patterns can then be straightforwardly matched to the input sentences with the help of a part-of-speech (POS) tagger and a noun phrase recognizer. For example, the system matches the sentence (1) to the syntactic pattern 'NP but not NP' with 'eIF4A' and 'eIF4E' matching the two NP variables syntactically.

In this work, we have developed (a) a novel POS tagger that assigns to each word its most frequent POS tag by looking up a manually curated POS dictionary[3] together with some domain-specific correction rules and (b) a novel noun phrase recognizer that looks for noun phrases that begin or end at predetermined words or POS, since the NP variables in the patterns are adjacent to a predetermined word (e.g. 'but', 'not') or a POS (e.g. PREP, V). For recognition of such NPs, we first locate a base noun phrase that is adjacent to the predetermined word or POS by utilizing a regular expression, namely '[(DT+|PR)?CD*(RB*ADJ+)?NN+]|PR'.[4] Our method then merges the base noun phrase to its adjacent prepositional phrases and coordinated noun phrases either on its left or on its right. In this way, the system can recognize complex noun phrases with recursive structures[5].

The system then analyzes semantic similarity between the coordinated expressions by analyzing word-level similarity between the expressions. Consider the following example that matches the pattern 'not V NP but V NP':

(4) In contrast, IFN-gamma priming did not affect the expression of *p105 transcripts* but enhanced the expression of *p65 mRNA* (2-fold) (PMID:8641346).

The system first matches the V variables to the verbs 'affect' and 'enhanced', and the NP variables to the noun phrases 'the expression of p105 transcripts' and 'the expression of p65 mRNA'. It then analyzes the similarity between the verbs and the similarity between the noun phrases.

Our system analyzes the word-level similarity by checking whether the variable-matching phrases are semantically identical or at least in a subsumption relation. To check for the similarity between verbs and between adjectives, it utilizes the synonymy and hypernymy relations in WordNet (Fellbaum, 1998) and also in a hand-made resource for semantic similarity between biomedical terms. The latter was constructed by examining 166 'not'-containing abstracts; we show some example verbs in the resource in Table 2.

To check for the similarity between noun phrases, the system analyzes the similarity between their headwords. For simplicity, we

[2]See Appendix A.2 for the pseudo-code for coordination identification.
[3]The POS dictionary contains 71 206 POS tags for 57 702 common English words by compiling the POS assignments in Penn TreeBank (http://www.cis.upenn.edu/ treebank/) with additional manual curation for the purpose of biomedical text mining. However, it does not include biomedical named entities such as gene and protein names.
[4]A base noun phrase is a noun phrase that has no recursive structure. DT indicates a determiner, PR a pronoun, CD a digit, RB an adverbial and NN a noun. TERM? indicates the optionality of the term, TERM* the fact that the term may occur zero or more times and TERM+ the fact that the term should occur at least once. A|B indicates a disjunctive choice between A and B.
[5]See the subfunctions FindStartofNounPhrase and FindEndofNounPhrase in Appendix B for further details.

assume that the headword of a noun phrase is the last noun (or nouns) of the first base noun phrase in the noun phrase. The system then analyzes the similarity between nouns by checking (a) whether the two nouns denote the same kind of biological objects (e.g. 'eIF4A' and 'eIF4E' in example (1) are protein names) by looking up well-known biomedical databases such as Swiss-Prot and (b) whether they are semantically identical [e.g. 'transcription' in example (2)] by utilizing WordNet and our resource for semantic similarity. For example (4), the system extracts a contrastive relation between the two noun phrases that match the NP variables (i.e. 'the expression of p105 transcripts' and 'the expression of p65 mRNA'), as the verb 'affect' is a hypernym of the verb 'enhance' and the two noun phrases have the same headword 'expression'.

Next, the system determines the presupposed property for the focused proteins by extracting the subject phrase and the verb whose object phrases correspond to the focused proteins (i.e. 'NAT1 binds CONTRAST_OBJ', where CONTRAST_OBJ indicates the variable for focused objects.). In this way, a protein–protein contrast between the proteins eIF4A and eIF4E in terms of their respective (or rather, contrastive) binding capabilities with NAT1 is thus extracted by BioContrasts from the example sentence (1).

3.2 Identifying clause-level parallelisms

If none of the subclausal coordination patterns can be matched to the input sentences in the previous step, the system then attempts to identify any parallelism expressed with the sentences[6]. Parallelism for a contrast refers to a pair of a negated clause and an affirmative clause that are either semantically identical or at least in a subsumption relation (ignoring the negative 'not' in the negated clause and the actual expressions for the focused objects). For example, the main clause and the subordinate clause in sentence (2) together encode parallelism for a contrast, because the two verbs are identical ('repress') except for negation, the two object phrases are also identical ('transcription') and the subject phrases which correspond to focused objects denote proteins.

Table 3 shows a compilation of various parallelism patterns used in BioContrasts for identifying clause-level parallelisms. A parallelism pattern is formed using any one from the positive patterns together with any one from the negative patterns in the table. To match an input sentence to a parallelism pattern, the system checks whether (a) the linguistic expressions that match the variables with the same subscript (e.g. $\{V_1, V_1'\}$) are either semantically identical (e.g. {'repress', 'repressed'}) or are in a subsumption relation (e.g. {'affect', 'activate'}) and (b) the variables with the subscript '$_C$' (e.g. $\{Subj_C, Subj_C'\}$), which indicate focused objects of the pattern, are matched to semantically similar expressions (e.g. {'eIF4A', 'eIF4E'}). Again, to deal with these word-level similarities, the system utilizes WordNet and our resource for semantic similarity between biomedical terms as described in Section 3.1.

Let us go through an example of how BioContrasts identifies a clause-level parallelism using the pattern <$Subj_C$ not V_1 Obj_2, $Subj_C'$ V_1' Obj_2'> (i.e. the pattern formed by combining the first positive pattern and the first negative pattern in Table 3).

[6]See A.3 and A.4 for the pseudo-codes for parallelism identification from one sentence and adjacent sentences respectively, and A.5 for the core function shared by these two kinds of parallelism identification.

Table 3. Parallelism pattern examples

Negative patterns	Positive patterns
$Subj_C$ not V_1 Obj_2	$Subj_C'$ V_1' Obj_2'
$Subj_2$ not V_1 Obj_C	$Subj_2'$ V_1' Obj_C'
$Subj_C$ not V_1 $PREP_3$ Obj_2	$Subj_C'$ V_1' $PREP_3'$ Obj_2'
$Subj_2$ not V_1 $PREP_3$ Obj_C	$Subj_2'$ V_1' $PREP_3'$ Obj_C'
$Subj_C$ BeV not ADJ_1 $PREP_3$ NP_2	$Subj_C'$ BeV ADJ_1' $PREP_3'$ NP_2'
$Subj_2$ BeV not ADJ_1 $PREP_3$ NP_C	$Subj_2'$ BeV ADJ_1' $PREP_3'$ NP_C'
NN_1 of NP_C with NP_2 not V	N_1' of NP_C' with NP_2' V
NN_1 between NP_C and NP_2 not V	NN_1' between NP_C' and NP_2' V

Subj and Obj denote the subject phrase and the object phrase in a clause, respectively. BeV indicates a be-verb. The variables with the same subscript, e.g., V_1 and V_1', match linguistic expressions that are semantically identical or one of which entails the other. The variables with the subscript '$_C$', which indicate focused objects of the pattern, match semantically similar expressions.

(i) First, the system locates a verb in a clause that is negated by the negative 'not' to match V_1 of the negative pattern. The system also tries to look for a verb for V_1' of the positive pattern in a neighboring clause. Note that the system chooses only main verbs but not particles in this process.

(ii) Next, the system identifies the corresponding subject phrases and the object phrases of the verbs, using the noun phrase recognizer previously employed for coordination identification together with several heuristics that account for complementizers, clausal conjunctions and sentence-modifying adverbials.

(iii) After that, the system identifies if there is indeed a contrast between the subject phrases by checking whether the headwords of the two object phrases are identical or are in a subsumption relation, and whether those of the two subject phrases are semantically similar. Finally, to determine the associated presupposed property, the system extracts the verb and the object phrase from the affirmative clause and adds the tag for focused objects ('CONTRAST_OBJ') at the subject position of the verb, thus extracting 'CONTRAST_OBJ V_1' Obj_2''.

For example, for sentence (2), the system first locates the verb 'repress' in the subordinate clause which is negated by 'not'. It also locates the positive verb 'repressed' of the main clause. Next, it identifies the corresponding subject phrases and the object phrases in the two clauses. Here, the subject phrase in the main clause (i.e. $Subj_C'$) is 'Truncated N-terminal mutant huntingtin', the object phrase (i.e. Obj_2') 'transcription', the subject phrase of the subordinate clause (i.e. $Subj_C$) 'the corresponding wild-type fragment', and the object phrase (i.e. Obj_2) 'transcription'. The system also checks that the two verb phrases and the two object phrases are all semantically identical. By parallelism identification, the contrastive relation extracted here is one between the two protein names at the corresponding subject positions (thus semantically similar) with respect to the presupposed biological property of 'CONTRAST_OBJ repressed transcription'.

3.3 Extracting contrastive protein names

The two preceding steps produce contrastive expressions that match the variables for focused objects (e.g. NP variables in

Table 1, $Subj_C$ and $Subj'_C$ in Table 3). Our earlier system (Kim and Park, 2005) terminates its extraction process at this point. Here, our BioContrasts system goes further into checking whether the contrastive expressions can be resolved into Swiss-Prot protein entries. To do so, the system must first handle individual protein names as well as coordinations between protein names. Consider the following example:

(5) Sepharose-6B gel filtration experiments with combinations of nonlabeled and 14C reductively methylated initiation factors (eIF-4A, eIF-4B, eIF-4F, and eIF-3) provide evidence that both *eIF-4B and eIF-4F*, but not *eIF-4A*, interact with ribosomes in the presence of specific factors and ATP (PMID:8460942).

If a contrastive expression includes a coordinating conjunction [e.g. 'both eIF-4B and eIF-4F' in (5)], the system decomposes the expression into a list of coordinated phrases (e.g. 'eIF-4B' and 'eIF-4F') and checks that all of them denote proteins before generating the corresponding protein–protein contrast.

Next, the system looks for repeated phrases in the extracted pair of contrastive expressions. For example, the coordination identification step extracted contrastive information between 'the expression of p105 transcripts' and 'the expression of p65 mRNA' from (4). By noting that the phrase 'the expression of' is repeated in the two contrastive expressions and that the pattern '<gene name> transcripts' is semantically identical to '<gene name> mRNA', the system extracts contrastive information between two protein names 'p105' and 'p65'. To address the word-level semantic similarity as required for this step, we utilize WordNet and our semantic resource as explained in Section 3.1.

In BioContrasts, we recognize protein names only after matching the contrastive patterns to sentences. If the system were to recognize protein names before the pattern matching step, it would not be able to effectively deal with complex structures such as cascaded noun phrases [e.g. (4)] and coordination [e.g. (5)]. Thus, our system first extracts contrastive expressions and then checks whether they are protein names or not.

3.4 Grounding protein names

As it is our objective to construct a useful knowledge base of protein–protein contrasts, we feel that it is important to cross-link the contrastive protein names from the literature with entries in standard protein databases such as Swiss-Prot. This cross-linking process is also known as protein name grounding or term identification (Kim and Park, 2004; Krauthammer and Nenadic, 2004). Table 4 shows some examples of protein name grounding. Note that a single protein name may be grounded with multiple Swiss-Prot entries—whenever this is the case, the current system treats contrasts between non-homologous protein entries from different species as unmeaningful and keeps only contrasts between protein entries from the same species or those between homologous protein entries (determined by BLASTP with a threshold of *E*-value e−10) from different species.

We have previously developed a protein name grounding module (Kim and Park, 2004) that matches protein names to Swiss-Prot entries with manually constructed patterns that deal with variations of protein names concerning special characters, indices, plurals, discardable words and acronyms, as exemplified in Table 4. It then outputs only those Swiss-Prot entries whose species information is identical to those found in the same MEDLINE abstracts. For example, given two Swiss-Prot entries, IF3A_SCHPO and IF3A_YEAST, which are both matched to the protein name *Rpg1p/Tif32p*, the module grounds the noun phrase '*Saccharomyces cerevisiae Rpg1p/Tif32p*' only with IF3A_YEAST.

In BioContrasts, we have further enhanced our name grounding module for better performance. Since the ambiguity in protein names had particularly affected the performance of our earlier module, BioContrasts now treats ambiguous protein names in a stricter fashion:

(i) The module first attempts to identify the full names of any abbreviated protein names by analyzing acronyms and appositive structures. Among the Swiss-Prot entries whose entry names are identical to the abbreviated protein names, the grounding module only outputs the entries that also have the exact full names of the abbreviations as their entry names. For example, while the abbreviation 'GR' can be matched to both GCR_HUMAN (Glucocorticoid receptor) and GSHR_HUMAN (Glutathione reductase) on its own, the module will ground the abbreviation only with GCR_HUMAN when there is a full name 'the glucocorticoid receptor (GR)' in the same abstract.

(ii) For better accuracy (currently, accuracy is our priority over coverage in BioContrasts), the module also discards protein names that are easily confused with other terms such as normal English words of lower case (e.g. mass, stellate) and two-character words with one digit (e.g. S1, T1).

The process of matching full names of abbreviations to Swiss-Prot entry names is often more complicated than simple

Table 4. Protein name grounding examples

Protein names from texts	Swiss-Prot entry names	Swiss-Prot IDs
D(2)	D2	D2_DICDI, D2_ONCVO
S receptor Kinase	S-receptor kinase	SRK6_BRAOE
RNase P	RNase P protein	RNPA_BACSU, RNPA_ECOLI, ...
Thioredoxin h (THL1)	Thioredoxin h-type 1	TRXH1_ARATH, TRXH1_BRANA, ...
Fibroblast Growth Factor-2	Heparin-binding growth factor 2 precursor; Basic fibroblast growth factor	FGF2_HUMAN, FGF2_MOUSE, FGF2_XENLA, ...
BFA-inhibited GEFs	Brefeldin A-inhibited guanine nucleotide-exchange protein 1	BIG1_BOVIN, BIG1_HUMAN

The strings of the form 'A; B' among Swiss-Prot entry names indicate two entry names of the same Swiss-Prot entry, i.e. A and B.

Table 5. Comparison of tagger performance

	Brill's tagger	MedPost	Our tagger
Precision (%)	92.8	96.9	92.1
Execution time (s)	24	165	13

pattern matching. First, the process for resolution of abbreviations can be a recursive one. For example, the abbreviated protein name 'TNF R2' in a MEDLINE abstract (PMID:12799919) is identified as a name of the Swiss-Prot entry TNR1B_MOUSE. However, the full name of the abbreviation, 'TNF receptor 2', found in the abstract is not identical to full name of the Swiss-Prot entry, namely, 'Tumor necrosis factor receptor 2'. Our grounding process must recursively match the full name found in the text to the Swiss-Prot entry by resolving the abbreviation 'TNF' with its full name 'Tumor necrosis factor' in turn. Second, the full name of an abbreviation may sometimes be identified only with multiple names in a Swiss-Prot entry. For example, the abbreviation 'FGF-2' and its full name 'fibroblast growth factor-2' found in the abstract (PMID:14598292) can be matched to the Swiss-Prot entry FGF2_XENLA, which has multiple entry names {'FGF-2', 'Basic fibroblast growth factor', 'Heparin-binding growth factor 2 precursor'}. Notice that the substring 'fibroblast growth factor' of the full name is matched to the second entry name 'Basic fibroblast growth factor', but the index '2' is found in the third entry name 'Heparin-binding growth factor 2 precursor' for the SWISS-Prot entry. Our BioContrasts system deals with these issues by employing recursive abbreviation resolution and by matching a protein name to multiple Swiss-Prot entry names, as discussed earlier.

4 IMPLEMENTATION

We have implemented our current BioContrasts' extraction module in the Python language. For natural language processing, instead of adopting existing generic tools into our BioContrasts system, we have developed a novel POS tagger and a novel noun phrase recognizer as described in Section 3.1. We also manually constructed the contrastive negation patterns shown in Tables 1 and 3, as well as the resource for semantic similarity in Table 2 using the 166-abstract training corpus described previously in Section 3.1.

We have chosen to develop such simple modules of natural language processing in order to enhance the efficiency of our system. For example, our POS tagger is designed to perform with reasonable precision rapidly. Table 5 shows the comparative result of our tagger with other taggers [namely, Brill's tagger (Brill, 1995) and MedPost (Smith *et al*., 2004)] on the GENIA corpus 3.02p8 (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA) that consists of 2000 annotated abstracts. We measured the precision of the taggers by comparing the output POS tags with the curated POS tags of the GENIA corpus up to the second level (e.g. 'VB' of 'VBZ'). We also compared the execution time by running the taggers on the same linux machine with two 2.8 GHz CPUs. As shown in Table 5, the precision of our tagger is comparable with Brill's tagger, and the speed of our tagger is the fastest (in fact, much faster than MedPost). With these modules, the BioContrasts system currently performs at ~0.038 s per abstract on average on Sun Fire V440.

Table 6. Example output of the BioContrasts system

PMID	9030685
Sentence	'NAT1 binds eIF4A but not eIF4E and inhibits both cap-dependent and cap-independent translation.'
Positive protein name	eIF4A
Positive S–P entries	IF4A_MOUSE IF4A_RABIT
Negative protein name	eIF4E
Negative S–P entries	IF4E_MOUSE IF4E_RABIT
Presupposed property	NAT1 binds CONTRAST_OBJ
Protein-protein contrasts	<IF4A_MOUSE, IF4E_MOUSE>, <IF4A_RABIT, IF4E_RABIT>

The BioContrasts database is publicly accessible via a user-friendly interactive web portal at http://biocontrasts.i2r.a-star.edu.sg/. The BioContrasts web portal is designed for easy exploitation of the contrastive relations between proteins extracted from the literature for biological knowledge discovery. Some of its key functionalities include:

(1) Users can search for contrasts of proteins of interest with their Swiss-Prot IDs or names.

(2) Users can browse and navigate networks of protein–protein contrasts graphically.

(3) Users can search for contrasts that are associated with KEGG pathways, InterPro domain entries, and Gene Ontology concepts, which may be useful for enhancement of KEGG pathway, inference over contrasts between protein domains, and subcategorization of Gene Ontology concepts.

5 EVALUATION

We compiled a large corpus of 2.5 million 'not'-containing MEDLINE abstracts to extract protein–protein contrasts reported in the literature. Our BioContrasts system extracted a total of 799 169 pairs of contrastive expressions from this corpus. Among them, the system identified 11 284 pairs of contrastive protein names and had the contrastive protein names successfully cross-linked to protein entries in the Swiss-Prot database, resulting in 41 471 contrasts between Swiss-Prot (S–P) entries—the larger number in the resulting protein–protein contrasts owes to the fact that a protein name may be grounded with multiple Swiss-Prot entries. Table 6 shows an example output of literature-reported protein–protein contrasts by the system for the sentence (1) from which two contrasts between Swiss-Prot entries were generated from a single contrastive relationship reported in the text.

To gain an indicative precision of the current BioContrasts system, we examined a set of 100 pairs of contrastive proteins randomly selected from the BioContrasts database[7]. We found that the system correctly extracted 97 pairs (97.0% precision) [cf. the performance of the earlier system: 85.7% precision and 61.5% recall (Kim and Park, 2005)]. We expect that the recall for the current system to be lower than that of our earlier system, as the system currently discards contrasts whose protein names are not grounded

[7]The test corpus of the 100 protein–protein contrasts is available at http://biocontrasts.i2r.a-star.edu.sg/BioContrasts-testcorpus.html

with Swiss-Prot entries. However, this is fine as our current priority is precision over recall for the practical usefulness of our database. In terms of the system's pattern usage, the main contrastive pattern 'A but not B' was used to extract 91 pairs (all correct), while the parallelism patterns were used 5 times (2/5 correct or 40% precision) in this evaluation.

In addition to evaluating the BioContrasts system in terms of its performance in extracting pairs of contrastive protein names from the texts, we were also interested in its performance in grounding the contrastive protein names with Swiss-Prot entries. Out of the 200 focused protein names in the test protein contrasts, 182 names were expressed with adequate clarifying descriptions in their respective abstracts for grounding. Among these 182 protein names, the system correctly grounded 164 (90.1% precision) and linked 6 additional protein names with partially correct entries (3.3% partial precision). Note that the performance of the grounding module is much enhanced from the earlier one (59.5% precision and 40.7% recall) (Kim and Park, 2004). The partially correct entries refer to those grounded to multiple Swiss-Prot entries where some (but not all) are correct cross-links. Note that the 18 names (all abbreviations, incidentally) excluded in our evaluation were also grounded by the system but their correctness could not be determined for reasons explained earlier.

In our error analysis of the system, we have found that all the errors (3 pairs) owed to inadequacy in the linguistic analysis for parallelism identification. Take the sentence below as an example.

(6) Interestingly, IL-15 stimulated relatively low level of expression of CD18, a beta2 integrin molecule related to lymphocyte apoptosis in A-NK cells (11.45%), whereas *IL-2* exerted a strong effect on CD18 expression (87.54%). IL-11b was only expressed at A-NK cell induced by IL-2 (49.56%), *IL-15* did not exert any stimulating effect on CD11b expression (PMID:12969549).

The system at present incorrectly extracted a contrast between 'IL-2' and 'IL-15'. This error has resulted from an incorrect semantic analysis that the two noun phrases, 'a strong effect on CD18 expression' and 'any stimulating effect on CD11b expression', were considered semantically identical because of the common headword 'effect'. Instead, the noun phrases should be considered neither semantically identical nor in a subsumption relation because of the different protein names, 'CD18' and 'CD11b' within the prepositional phrases. However, because our current parallelism identification process only considers the headwords, it was not able to detect the semantic inconsistency here. In order to deal with this case, we need to consider not only headwords but also the syntactic structures so as to determine the semantic similarity between such noun phrases. As noun phrases may show an equally complex syntactic structure as full sentences, we leave this problem for future work.

6 APPLICATIONS

Expressions of contrasts encode highly compact units of insightful information that not only explicitly describe the observed differences between the focused objects (in our case, proteins) but also implicitly indicate certain underlying semantic similarity or relationships between the focused objects being contrasted. In this section, we describe how the extracted protein contrasts can be exploited to construct more complete protein pathways.

6.1 Refining pathway roles of similar proteins

The primary information explicitly communicated by protein–protein contrasts are the observed functional differences of the focused proteins. These information can be used for distinguishing between the finer biological roles of seemingly similar proteins. Such knowledge is particularly useful for today's explorative research, given that many pathways of biology are still not yet at the desired granularity of specific molecules—often, the functional roles of functionally similar proteins are not adequately differentiated in existing pathways. Take, for an example, the pathway for well-studied Huntington's disease (HD) in the standard pathway database KEGG (Kanehisa *et al*., 2004). A key node in the pathway was labeled generically as 'caspase'—resolving this pathway node into more specific caspase molecules is important for understanding how the disease pathway operates, as it is involved in the critical functional process of cleaving of the key Huntingtin (Htt) protein. A further query of the KEGG database revealed that this node could possibly be resolved as caspase-3 and/or caspase-6. However, it is not clear whether both caspases are expected to behave in the same way in the HD pathway or not.

The contrastive information in the BioContrasts database can be used to help refine the functional roles of these two caspases. Indeed, there is a protein–protein contrast between caspase-3 and caspase-6 extracted by BioContrasts from the source sentence (7) shown below (Note that Yama is an alias for caspase-3 while Mch2 is an alias for caspase-6.):

(7) Importantly, *Mch2*, but not *Yama* or LAP3, is capable of cleaving lamin A to its signature apoptotic fragment, indicating that Mch2 is an apoptotic laminase (PMID:8663580).

The existence of such a contrast between the two proteins suggests the possibility that the two proteins may not function identically in the biological operations of the disease. Indeed, by searching for MEDLINE with three keywords, {caspase-3, caspase-6, Huntington's disease}, we found an article which specifically explains the difference between the two proteins in terms of the cleavage sites at Htt, as indicated in sentence (8):

(8) We have previously shown that Htt is cleaved *in vitro* by *caspase-3* at amino acids 513 and 552, and by *caspase-6* at amino-acid position 586 (PMID:10770929).

In this way, the contrastive information in BioContrasts can help guide a biologist in understanding and refining biological pathways of interest. Let us take another major disease pathway for an example. In the KEGG pathway for Alzheimer's disease (AD), it was presented that the two homologous enzymes BACE1 and BACE2 were involved in the derivation of Amyloid beta protein (Abeta) from amyloid beta precursor protein (APP). On querying about these two enzymes for contrastive information in our BioContrasts database, we found a reported contrast between the two enzymes extracted from the sentence (9), which in turn leads to (on further relevant literature search in MEDLINE as provided on the BioContrasts web portal) a telling sentence (10) which presents evidence that BACE2 may actually not have the function of Abeta production (Note that the name BACE was grounded to BACE1 by BioContrasts here.):

(9) The Flemish missense mutation of APP, implicated in a form of familial Alzheimer's disease, is adjacent to this latter site and

markedly increases Abeta production by *BACE2* but not by *BACE* (PMID:10931940).

(10) Our data argue against *BACE2* being involved in the formation of neuritic plaques in AD (PMID:15857888).

In this case, a contrast between homologous proteins that belong to a KEGG pathway led to an important revelation that one of the proteins may actually not have the reported function in the pathway.

6.2 Identifying implicitly similar proteins

Being in a contrastive relationship can implicate some level of implicit similarity or functional relationships between the focused objects being contrasted. Such implicit similarity information can be useful in understanding the interactome (e.g. functional annotation of proteins), especially if the similarity between these proteins are non-obvious (i.e. not readily detected by conventional means).

To verify that there were indeed functional similarities between the focused proteins in the contrasts extracted by BioContrasts, we performed a functional analysis on the contrasted proteins in the BioContrasts database. The analysis was conducted using the functional annotation of proteins based on their Gene Ontology or GO (Harris *et al*., 2004) assignments. Out of the 5407 protein contrasts that involved functionally annotated proteins[8], we have found 4948 contrasts (91.7%) between proteins that shared at least one GO code at level 2. This suggested that there are implicit functional similarities between the focused proteins being contrasted. In addition, we also found that only a low 10% (1720 contrasts) of the protein–protein contrasts in BioContrasts are between homologous protein entries. This means that the implicit similarities encoded by the protein contrasts cannot be easily detected by such conventional methods as sequence homology.

The obvious exploitation of such implicit similarity information is in the functional annotation of proteins. In addition, while the explicit differences expressed in the protein–protein contrasts were shown to be useful for differentiating the finer roles of homologous proteins in a pathway, the implicit similarities encoded in the protein–protein contrasts may also be used to propose candidate non-homologous proteins in incomplete biological pathways. Take the KEGG pathway for Huntington's disease for an example again, and let us focus on one of the pathway member proteins, the neurotronic factor BDNF. We found that there was a protein-protein contrast between BDNF and another neurotronic factor CNTF in our BioContrasts database. The implicit functional similarity between BDNF and CNTF suggests a possible involvement of CNTF in HD (probably in a slightly different way). On further relevant literature search in MEDLINE, this hypothesis was confirmed by another work as follows:

(11) For example, we have shown using an *in vitro* neuronal model of **HD** that *CNTF* and *BDNF* block polyQ-huntingtin-induced cell death (Saudou et al., 1998). *In vivo*, *CNTF* has also been shown to be neuro-protective in rats and monkeys following excitotoxic lesions that reproduce **HD** (for a review, see Brouillet *et al*., 1999) (PMID:12062094).

[8]Note the low percentage (12.8%) of proteins that have GO assignments, reflecting the current lack of (and need for) functional annotational information for the proteins. The implicit similarity information in BioContrasts can help in the functional annotation of proteins.

7 CONCLUSIONS

We have shown that contrasts are richly informative units of information that can be exploited for guiding knowledge discovery in explorative research. In particular, contrastive negation patterns such as 'A but not B' not only explicitly represent differences between the focused objects but also implicitly indicate that the focused objects are semantically similar.

Current biological databases have thus far collected mostly individual declarative positive facts and interactions. Our BioContrasts system captures, for the first time, useful contrastive and non-positive biological relationships between proteins by extracting contrastive expressions in the MEDLINE abstracts containing the contrastive negation patterns. By using pattern-driven text mining approaches, we were able to extract much contrastive information from a large corpus of MEDLINE abstracts, resulting in a useful resource for exploitation—the similarities, divergences, and relations between the contrasted proteins can provide much invaluable insights for the biologists to build more complete pathways for various biological phenomena.

In this work, we have focused on extracting contrasts that were expressed with linguistic structures containing the negative 'not'. However, contrasts can also be represented without the negative 'not'; for example, using contrasting word pairs such as {suppress, support}, as in the sentence:

(12) Transient transfection experiments show that *p97* suppresses both cap-dependent and independent translation, while *eIF4G* supports both translation pathways (PMID:9049310).

In our future work, we will be looking into the various challenges of extracting contrastive information expressed in linguistic structures other than 'A but not B', such as the use of more efficient mechanisms (e.g. a Finite State Automaton) for pattern matching and introducing new patterns to the extraction system. At the same time, we would also like to explore other ways for exploiting the extracted contrasts for biological knowledge discovery. For example, it may be possible that the presupposed properties (biological activities) of the same pairs or groups of focused proteins are also related to each other. It would be interesting to explore how such information may lead to the discovery of new knowledge.

ACKNOWLEDGEMENTS

This work was partially supported by MOST/KOSEF through AITrc and Grants for Interdisciplinary Research (R01-2005-000-10824-0).

Conflict of Interest: none declared.

REFERENCES

Alfarano,C. *et al*. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*., **33**, D418–D424.

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.

Brill,E. (1995) Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Comput. Linguistics*, **21**, 543–565.

Cohen,K.B. and Hunter,L. (2004) Natural language processing and systems biology. In Dubitzky and Pereira (eds), *Artificial Intelligence Methods and Tools for Systems Biology*. Springer Verlag, Dordrecht, Netherlands.

Fellbaum,C. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Harris,M.A. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.

Hirschman,L. *et al*. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553–1561.

Kanehisa,M. *et al*. (2004) The KEGG resources for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.

Kim,J.J. and Park,J.C. (2004) BioAR: Anaphora Resolution for Relating Protein Names with Proteome Database Entries. In *Proceedings of the Reference Resolution and its Application Workshop in Conjunction with ACL 2004,* Barcelona, Spain, pp. 79–86.

Kim,J.J. and Park,J.C. (2006) Extracting contrastive information from negation patterns in biomedical literature. *ACM Transactions on Asian Language Information Processing, Special Issue on Text Mining and Management in Biomedicine*, (in press).

Knight,J. (2003) Negative results: Null and void. *Nature*, **422**, 554–555.

Krauthammer,M. and Nenadic,G. (2004) Term identification in the biomedical literature. *J. Biomed. Inform*., **37**, 512–526.

Li,X.-L., Tan,S.-H. and Ng,S.-K. (2005) Protein interaction prediction using inferred domain interactions and biologically-significant negative dataset. In *Proceedings of the First International Workshop on Data Mining and Bioinformatics,* Singapore.

Prince,E. (1992) The ZPG letter: Subjects, definiteness and information-status. In Mann,W. and Thompson,S. (eds), *Discourse Description: Diverse Analyses of a Fund-Raising Text*. John Benjamins, Amsterdam, pp. 295–325.

Salwinski,L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451.

Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol*., **10**, 821–855.

Smith,L. *et al.* (2004) MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics*, **20,** 2320–2321.

Umbach,C. (2004) On the notion of contrast in information structure and discourse structure. *J. Semantics*, **21**, 155–175.