Relation Characterization using Ontological Concepts

Muhammad Abulaish[#]
Center of Excellence in Information Assurance
King Saud University
Riyadh, Kingdom of Saudi Arabia
e-mail: mAbulaish@ksu.edu.sa

Abstract – **This paper presents a method using the concept of AND-OR tree to characterize relations, mined from MEDLINE abstracts, using biological ontology concepts. A biological relation is expressed as a binary relation associated to two molecular biology concepts as defined in the GENIA ontology. Since a biological relation may relate different pairs of biological concepts and vice-versa, the strength of a relation which reflects the relative frequency of occurrence of the specific association within the corpus, is calculated and stored in the underlying ontological structure. A biological relation along with the degree of association is termed as fuzzy biological relation.**

Keywords – web intelligence; biological relation mining; biological relation characterization; fuzzy ontology structure.

I. INTRODUCTION

The sheer enormity of the collection of text documents in the Molecular Biology domain necessitates design of automated content analysis systems, without which the assimilation of knowledge from this vast repository is becoming practically impossible. *Ontology-guided information extraction and query processing mechanisms* have been successfully applied for extracting information from biomedical documents. The general practice is to annotate text elements with ontology concept or biological process tags. These tags can be exploited for information retrieval. Reasoning about contents of a text document however needs more than identification of the ontological concepts present in it. Identification of the interactions among various core molecular biology concepts is essential for exploiting knowledge contained within the vast repository of research articles effectively. Biological relations defined between biological entities can establish the context of the entities in a document. Hence, it is important that the biological relations among the ontological concepts present in a text are also extracted and interpreted correctly. Sekimizu and Tsujii [7] stated how some of the most commonly occurring relations in the domain of Molecular Biology like *activate, bind, interact, regulate, encode, signal* and *function* can be located within text. However, it is expensive and labour-intensive to pre-define all such relations exhaustively. Literature Based Discovery (LBD) aims at extracting these relations automatically. Rinaldi *et al.* [1] proposed a LBD mechanism to further characterize the seven relations of [7] in terms of the participating entities.

In this paper, we present a mechanism to characterize biological relations extracted from text documents. Our work is significantly different from those reported in [1] and [7] since we do not pre-suppose any set of relations; rather discover these relations from a corpus. The relations can be effectively used for indexing the corpus to aid in efficient information retrieval. Also, we attempt to characterize biological relations as interactions between molecular biology ontology concepts rather than between entities. Concept level characterization is a more generic characterization than entity-level characterization. The set of relations are mined from tagged MEDLINE abstracts which are a part of the GENIA corpus, using a text mining approach which we had proposed in [3]. Since the abstracts are tagged using leaf-level concepts from GENIA ontology, all mined relations are initially defined between leaf-level concepts only. We describe in this paper how each mined relation is subjected to feasibility analysis to determine its most generic representation. Thus a biological concept pair associated to a relation need not be only leaf-level concepts, rather could be concepts defined at any level of specificity in GENIA ontology. The extension of the underlying ontology, termed as *fuzzy biological relation ontology*, is also proposed to store these generic relations along with their degree of associations as interactions between biological concepts. The fuzzy biological relation ontology can be viewed as a supplement to the GENIA concept ontology, which helps in analyzing the biological relations prevalent in the domain. An ontology is not a database, and hence should not be a store-house for relation instances. The proposed fuzzy biological relation ontology adheres to this principle and stores knowledge about the various categories of relations occurring in the corpus at appropriate levels of conceptualization rather than every instance of relations mined. The relation-concept pair association is many-many, and it is observed that all possible combinations do not occur equally frequently within the text corpus. Hence a fuzzy ontological structure is the most appropriate

[#] *On leave from Jamia Millia Islamia (A Central University), Delhi, India* representation to store the relations, where each relation R in conjunction with a concept-pair C_1 and C_2, is associated with a strength μ that reflects the frequency of association $R(C_1, C_2)$ within the corpus. In some sense, this ontology represents the importance of the relations within a corpus and hence reflects the focus of research at a given point in time.

The remaining paper is organized as follows. Section II presents a brief review of the related works in the area of biological relation extraction and characterization. Section III introduces the relations mined from GENIA corpus using text mining. In section IV, we present the biological relation characterization process. In section V, we present the creation of the fuzzy biological relation ontology to accommodate biological relations and their fuzzy strengths representing the degree of associations. Finally, section VI concludes the paper with future directions.

II. Related Work

In this section, we present an overview of some of the earlier works reported in the area of biological relation extraction and characterization. The general approach has been to focus on certain verbs that represent biological relations. Thomas *et al.* [2] modified a pre-existing parser based on cascaded finite state machines to fill templates with information on protein interactions for three verbs – *interact with, associate with, bind to*. Sekimizu *et al.* [7] designed mechanisms for locating a pre-defined collection of seven verbs – *activate*, *bind*, *interact*, *regulate*, *encode*, *signal,* and *function* in a collection of abstracts and developed partial and shallow parsing techniques to find the verbs' subject and objects. Rinaldi *et al.* [1] have proposed an approach towards automatic extraction of subject and object for this set of seven relations in the domain of Molecular Biology, based on a complete syntactic analysis of an existing corpus. The PASTA system is a more comprehensive system that extracts relations between proteins, species and residues [5]. Ono *et al.* [6] reports a method for extraction of protein-protein interactions based on a combination of syntactic patterns. They employ a dictionary look-up approach to identify proteins in the document to analyze, and then select sentences that contain at least two proteins, which are then parsed with POS matching rules. The rules are triggered by a set of keywords, which are frequently used to name protein interactions like *associate, bind* etc.

It can be observed that most of the earlier works are designed to extract a pre-defined set of relations or relations that occur between a pre-defined set of elements like proteins or genes. The proposed work is more generic in nature since there is no underlying assumption about the biological relations or ontology concepts that are taking part in a relation.

III. Feasible Biological Relations Mined From GENIA Corpus

Discovering the interactions between genes and proteins is a core task in Molecular Biology. A biological relation is expressed as a binary relation between two biological concepts [6]. Gene Ontology (GO) defines an exhaustive set of biological processes, functions and relations. Since manual definition is laborious and also may cause problems of integrity and consistency whenever there is an update, a better approach is to mine for such relations from literature using automated means. We propose to characterize biological relations mined from the GENIA corpus which contains 2000 tagged MEDLINE abstracts. Tags are leaf concepts in GENIA ontology. Tags may be nested whereby a tagged biological entity in conjunction with other entities or processes may be tagged as a different leaf concept. A relation is identified as a biological activity co-occurring with a pair of tags.

Using a text mining framework [3], we have extracted the frequently occurring biological relational verbs co-occurring with a pair of ontology tags from the GENIA corpus. After a feasibility analysis 24 basic biological relational verbs were identified. Considering these 24 seed relations, their morphological variants and also preposition associations, we have mined 4162 unique biological relation triplets from the GENIA corpus [3]. A feasible biological relation is represented as $R(C_i,C_j)$, where R denotes a frequently occurring biological relation located within leaf-level tags C_i and C_j in the corpus. The relation $R(C_i, C_j)$ is not commutative.

The relation to concept-pair association is many-many. For example, the relation "*activated with*" occurs between concept-pairs <*multi_cell, protein_molecule*>, <*protein_molecule, protein_molecule*>, <*cell_type, protein_molecule*>, <*cell_line, protein_molecule*> and <*cell_type, other_organic_compound*>. The concept-pair <*multi_cell, protein_molecule*> is also associated with other biological relations like "*stimulated with*", "*binding*" etc. Hence, each feasible biological relation R is associated with a strength $\mu_{(Ci,\ Cj)}(R)$ which is a function of co-occurrence of the relation R in conjunction with the pair of leaf concepts C_i and C_j and is computed using the formula shown in equation 1.

$$\mu_{(Ci,Cj)}(R) = \frac{1}{2}\left\{ \frac{|R(C_i, C_j)|}{\sum_{a,b} |R(C_a, C_b)|} + \frac{|R(C_i, C_j)|}{\sum_{r} |R_r(C_i, C_j)|} \right\} \qquad (1)$$

IV. Biological Relation Characterization

In this section, we present how we generalize the mined relations, introduced in the previous section, and calculate their degree of associations. The aim is to generalize a relation at appropriate level of specificity and not store every instance of relations in the fuzzy biological relation ontology, which may include noise.

According to the GENIA ontology all biological concepts can be broadly categorized into two categories –

source and *substance*. Hence we categorize the mined relations into the following all possible four major categories depending on the nature of the participating concepts.

- *source-source*
- *source-substance*
- *substance-source*, and
- *substance-substance*

For example, a relation which is obtained as an association between the leaf pair <*cell_type, protein_molecule*> is categorized as a relation belonging to the <*source, substance*> category. Further, this relation can be also looked upon as a relation between concept pairs <*natural source, amino acid*> or <*source, amino acid*> or <*source, organic*> etc., which are all defined as generalizations of the respective leaf concepts along the GENIA ontology.

We establish the need for generalization through an example. On consolidating the information about the relation "*expressed in"* in the corpus, it is observed that out of total 170 occurrences of the relation, all of them occur in association to the concept pair *(compound, source)*. Of these, 127 occurrences are in conjunction with *(compound, natural source),* 91 of which can be traced to occur with the pair *(amino acid, natural source), 35* with the pair *(nucleic acid, natural source)* and only 1 occurs with the pair *(lipid, natural source)*. Similarly, the 43 occurrences of the relation that are with *(compound, artificial source)* can be further specialized to *(amino acid, artificial source)* for 34 of them and the remaining 9 are associated with the pair *(nucleic acid, artificial source)*. On the basis of these figures, we conclude that the relation ontology can store the relation "*expressed in"* in association to *(amino acid, natural source)* as a *strong* association, and that between the pairs *(amino acid, artificial source)* and *(nucleic acid, natural source)* as weak associations respectively. All other associations for this relation may be ignored. Our aim is to generate such appropriate concept pair associations for all feasible biological relations mined.

A. AND-OR Tree Creation

To generate the complete list of arguments for any relation *R* at the optimal level, we first create an AND-OR tree of concept-pairs, termed as AND-OR concept-pair tree, that stores all possible concept-pairs generated from the underlying concept ontology. For an ontology with *m* concepts *m(m-1)* concept-pairs can be generated. The number of times a concept-pair <C_i, C_j> occurs in the concept-pair tree is obtained using the recursive equation 2 where i and j represents the level of concepts C_i and C_j respectively in the concept ontology tree.

$$N(i,j)=\begin{cases}N(i-1,j)+N(i,j-1), & \forall\ i,j\geq 2\\ 1, & otherwise\end{cases} \tag{2}$$

The total number of nodes in a concept-pair tree, N, is given by the formula in equation 3. In this equation, l_1 and l_2 are depths of the left and right concept ontology trees respectively, n_i is the number of nodes at level i in the concept ontology tree.

$$N=\sum_{i=1}^{l_1}[n_i\times\sum_{j=1}^{l_2}n_j\times N(i,j)] \tag{3}$$

There are four instances of AND-OR tree, created with the two *source* and *substance* sub-trees from the GENIA ontology. Every node in the concept-pair tree has two constituent concepts denoted as <LEFT, RIGHT>. For every node, two sets of child nodes are created as follows. The first set of child nodes is created by expanding the LEFT concept to consider all its child nodes in the GENIA ontology, while keeping the RIGHT concept unchanged. The second set of child nodes is created by keeping the LEFT concept unchanged while expanding the RIGHT concept in the GENIA ontology. The complete tree is generated recursively. Fig. 1 shows the concept-pair tree resulting from the merging of concepts from two hypothetical sub-trees. The concept-pair tree is analyzed as an AND-OR tree for obtaining the most specific level at which a relation is definable with sufficient strength using an information-loss based approach. Since the mined relation instances are defined for leaf concept pairs, hence for any given relation, the frequency of the relation for a leaf node in the concept-pair tree can be computed from these. The instance frequencies are used in a bottom-up approach to compute the frequencies at the inner nodes up to the root node. However, as is obvious from Fig. 1, there are alternative paths to reach the root from the leaf nodes, and these have to be used in disjunction. In order to derive the most appropriate levels of specificity for describing any relation, the concept-pair tree is traversed treating it as an AND-OR tree. The total number of relations definable for root concept pair <*a*, *d*> can be obtained either as the summation of relations defined for pairs <*b*, *d*> and <*c*, *d*>

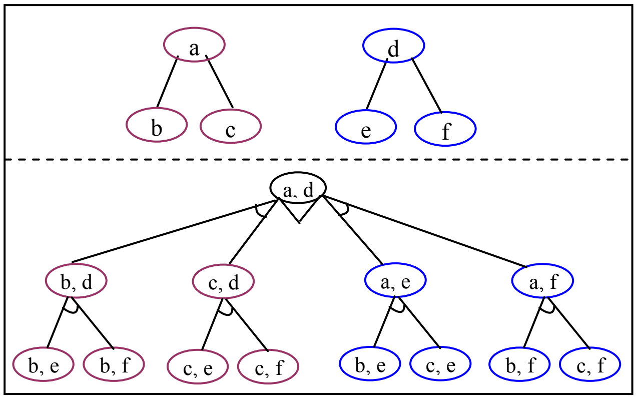

Figure 1. Sample AND-OR concept-pair tree

or for the pairs <*a*, *e*> and <*a*, *f*>, which in turn can be computed as summations of the instances defined at the leaf nodes below them. It is obvious that the summation of <*b*, *d*> and <*c*, *d*> has to be equal to that of <*a*, *e*> and <*a*, *f*> since each collection is ultimately composed of the same leaf concept pair nodes. Hence for all internal nodes, the total number of relations at a node (LEFT, RIGHT) is denoted by |R(LEFT, RIGHT)|, and is obtained by summing up the counts at its child nodes generated either by recursively expanding the LEFT concept or the RIGHT concept. A function to create AND-OR tree is shown in Fig. 2.

```
struct ontologyConcept {
    string conceptName;
    int numberOfChildConcept;
    struct ontologyConcept *child[MAXCHILDS];
}
struct andOrTreeNode {
    struct ontologyConcept *concept1;
    struct ontologyConcept *concept2;
    string conceptPair;
    int L1Length;
    int L2Length;
    struct andOrTreeNode *L1[];
    struct andOrTreeNode *L2[];
}
function andOrTreeCreation (struct ontologyConcept *root1, struct
ontologyConcept *root2){
struct andOrTreeNode *T;
T->concept1 := root1;
T->concept2 := root2;
T->conceptPair := merge(root1->conceptName, root2->conceptName);
int leftIndex := 1;
While (leftIndex ≤ root1->numberOfChildConcept)
    T->L1[leftIndex] := andOrTreeCreation (root1->child[leftIndex], root2);
    leftIndex++;
T->L1Length = leftIndex;
int rightIndex := 1;
While (rightIndex ≤ root2->numberOfChildConcept)
    T->L2[rightIndex] = andOrTreeCreation (root1, root2->child[rightIndex]);
    rightIndex++;
T->L2Length = rightIndex;
Return T;
}
```

Figure 2. AND-OR concept-pair tree creation function

B. *Relation Characterization*

After the frequency of a relation is determined for each node in the concept-pair tree, an information loss function based on set-theoretic approach is applied at each node to determine the appropriateness of defining the relation at that level. This process follows a top-down scanning of the AND-OR tree. Starting from the root node, the aim is to determine those branches and thereby those nodes which can account for sufficiently large number of relation instances. When the frequency of a relation drops beyond a threshold at a node, its descendents are not considered for the relation conceptualization. The information loss at each node *N* other than the root node is computed using equation 4 in which η represents the number of instances of relation *R* at the node *N* and θ represents the number of instances of *R* at parent node of the node *N*.

$$InfoLoss(N) = \frac{|\theta - \eta|}{|\theta + \eta|} \tag{4}$$

For any given relation, starting from the root node of a concept pair tree, the information loss incurred at each internal node is considered in a top-down fashion to decide the most appropriate level for generalizing the relation. If the information loss at a node *N* is above a threshold, it is implied that the node itself accounts for a very small percentage of the relation instances that are defined for its parent. Hence any subtree rooted at this node may be pruned off from further consideration while deciding the appropriate level of concept pair association for a relation. We have used a threshold of 10%, i.e., any node which accounts for less than 10% of its parent's relations is pruned off. Since a node may have at most two alternative paths denoted by the expansion of LEFT and RIGHT respectively, along which a relation may be further specialized, the choice of appropriate path to follow is done as follows. For each subtree at a node, total error for the subtree is computed as the average over information loss for each retained child. If the error values of the two subtrees are close to each other then both the subtrees are pruned off, and the node serves as the appropriate level of specification. Otherwise, the tree with less total error is retained for further specialization, while the one with the higher error is pruned off. The set of concept-pairs retained are used for conceptualizing the relations.

The strength of each generalized relation is computed using equation 5, where *G* denotes one of the four concept pair trees. T^G denotes the total count of all relations that are defined for the tree under consideration, and N_R^G denotes the total number of relation instances of type *R* for the given tree.

$$\mu^{G}_{(C_i,C_j)}(R) = \frac{1}{2}\left\{\frac{|R(C_i,C_j)|}{N^{G}_{R}} + \frac{|R(C_i,C_j)|}{T^{G}}\right\} \qquad (5)$$

V. Fuzzy Biological Relation Ontology Creation

It has been established earlier that generic relations are fuzzy in the sense that a relation can be defined between different concept-pairs with varying degrees of strength and vice-versa. This is best done through the use of linguistic qualifiers that express the strength of a relation to a varying degree. Thus, a fuzzy biological relation *R* is defined as a collection of triplet <C_i, C_j, μ>, where C_i and C_j are either source or substance concepts defined in the GENIA ontology and μ is the fuzzy strength in terms of linguistic qualifiers *weak*, *moderate* and *strong* that are obtained after fuzzyfication of the numeric values of strength obtained from equation 5. Since there are four different concept-pair categories, hence each relation is defined with multiple inheritances.

Table 1 shows some of the candidate biological relations along with their associated concept pairs and strengths identified for creating fuzzy relational ontology structure. Since GENIA ontology stores information about biological concepts only, it cannot be exploited for representing biological interactions. Hence, we consider extending this ontology by adding the generic relations to this.

Table I. Instances Stored in Fuzzy Biological Relation Ontology

Relation Name	**Concept-pair Category**	**Concept-pairs**	**Strength**
Expressed in	Source-Source	<natural,organism> <natural,tissue> <natural,cell_type>	Weak Weak Moderate
	Substance-Substance	<DNA,organic> <protein,amino_acid> <protein,nucleic_acid> <RNA,other_organic_compound>	Weak Moderate Weak Weak
	Substance-Source	<organic,source>	Strong
Activates	Source-Substance	<natural,amino_acid> <cell_line,amino_acid> <cell_line,nucleic_acid>	Moderate Weak Weak
	Source-Source	< source,source>	Strong
	Substance-Substance	< protein,amino_acid> < protein,nucleic_acid>	Moderate Weak
	Substance-Source	< organic,source>	Strong
Inhibits	Source-Substance	<natural,amino_acid> <natural,nucleic_acid>	Strong Moderate
	Source-Source	< cell_type, artificial> <cell_type, natural>	Strong Strong
	Substance-Substance	<substance, compound>	Strong
	Substance-Source	<lipid, cell_type> <protein_family_or_group, cell_type> <protein_molecule, cell_type> <DNA_domain_or_region, cell_type>	Weak Weak Moderate Weak

In order to accommodate generic biological relations and their strengths, three generic classes - *concetPair*, *genericRelation,* and *fuzzyStrength*, in addition to existing GENIA ontology classes, are added to the GENIA relational ontology structure. The *conceptPair* class has two properties, *hasLeftConcept* and *hasRightConcept*, whose values are the instances of the GENIA concept classes. The *fuzzyStrength* class stores the fuzzy quantifiers to be associated with the generic relations to represent their strength. The *fuzzyStrength* class consists of a single property *termSet* which is defined as a symbol and contains the fuzzy quantifiers *weak*, *moderate* or *strong*. The *genericRelation* class has two properties – *leftRightActors* and *strength*. The *leftRightActors* property is a kind of OWL object property which range is bound to the *conceptPair* class. This is also restricted to store exactly one value, an instance of the *conceptPair* class, for every instance of a generic relation. The *strength* property is also a kind of OWL object property for which the range is bound to the *fuzzyStrength* class. This property is also restricted to store exactly one value for every instance of the generic relations. All mined generic relations are defined as instances of the class *genericRelation*. In order to incorporate the mined relations, their strength and domain and range sets we have used Protégé[1].

VI. Conclusion and Future Work

In this paper, we have present a method using the concept of AND-OR tree to characterize biological relations mined from MEDLINE abstracts. The mined relations which are always defined between a pair of leaf-level concepts in the GENIA ontology are generalized using a novel technique. The generalization task is framed as an optimization problem over an AND-OR concept-pair tree. Since the relations occur with varying strengths, creation of fuzzy biological relation ontology is also presented to store the biological relations and their fuzzy strengths. The fuzzy biological relational ontology can supplement the existing GENIA ontology of molecular biology concepts. Presently, relation-based indexing of text corpora is being explored to answer biomedical queries over text documents in an efficient way.

[1] http://protégé.stanford.edu

References

[1] F. Rinaldi, G. Scheider, C. Andronis, A. Persidis, O. Konstani, Mining Relations in the GENIA Corpus, in: Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy, 24 September 2004.

[2] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, M. Carroll, Automatic Extraction of Protein Interactions from Scientific Abstracts, in: Pacific Symposium on Biocomputing, 2000, pp. 538-549.

[3] M. Abulaish, and L. Dey, Biological Ontology Enhancement with Fuzzy Relations: A Text-Mining Framework, in: *Proceedings of the 2005 IEEE/WIC/ACM Int'l Conference on Web Intelligence*, France, 2005, pp. 379-385.

[4] M. Wallace, and Y. Avrithis, "Fuzzy Relational Knowledge Representation and Context in the Service of Semantic Information Retrieval", *In Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Budapest, Hungary, July 2004.

[5] R. Gaizauskas, G. Demetriou, P. J. Artymiuk, P. Willett, Protein Structures and Information Extraction from Biological Texts: the PASTA System, Bioinformatics 19(1), 2003, pp. 135-143.

[6] T. Ono, H. Hishigaki, A. Tanigami, T. Takagi, Automated Extraction of Information on Protein-Protein interactions from the Biological Literature, Bioinformatics 17(2), 2001, pp. 155-161.

[7] T. Sekimizu, H.S. Park, and J. Tsujii, "Identifying the interaction between genes and genes products based on frequently seen verbs in Medline abstract", *Genome Informatics 9*, 1998, pp. 62–71.