Aggregating Automatically Extracted Regulatory Pathway Relations

Byron Marshall, Hua Su, Daniel McDonald, Shauna Eggers, and Hsinchun Chen

***Abstract*—Automatic tools to extract information from biomedical texts are needed to help researchers leverage the vast and increasing body of biomedical literature. While several biomedical relation extraction systems have been created and tested, little work has been done to meaningfully organize the extracted relations. Organizational processes should consolidate multiple references to the same objects over various levels of granularity, connect those references to other resources, and capture contextual information. We propose a feature decomposition approach to relation aggregation to support a five-level aggregation framework. Our BioAggregate tagger uses this approach to identify key features in extracted relation name strings. We show encouraging feature assignment accuracy and report substantial consolidation in a network of extracted relations.**

***Index Terms*—Regulatory Pathway Analysis, Relation Parsing, Knowledge Representation**

I. Introduction

THE number of new abstracts appearing each day in the PubMed database rose from an average of 746/day in 1980 to 1,760/day in early 2005. To help researchers leverage this vast and growing collection of documents, several systems have been developed to extract biological relations from free text (see Section II.A). These systems promise decreased costs and increased coverage as compared to the manual curation processes. Considerable attention has been paid to accuracy of these systems considering both the correctness of extracted information (precision) and the coverage of the output (recall). The evaluators ask "is it correct?" and "did we get everything?" As these extraction technologies mature researchers need to go beyond accuracy evaluation and consider system usefulness.

The GeneScene system extracts regulatory pathway triples from MEDLINE abstracts to support search, visualization, knowledge discovery, and automatic analysis algorithms so that researchers can more efficiently leverage available information to gain insight from previous work, generate new hypotheses, and analyze experimental results. To our knowledge, it is the only end-to-end system that automatically extracts pathway relations from abstracts and presents them as a network. We are scaling up our system to handle millions of abstracts. GeneScene users suggested that extracted relations would be more useful in accomplishing these tasks if (1) references to the same substances and functions are indexed appropriately, (2) those references can be directly connected to existing database resources, and (3) important contextual information is included.

In this paper we propose a methodology for meaningfully organizing or "aggregating" the output of biomedical relation extraction systems and present an initial evaluation of the BioAggregate tagger, which identifies a relation's features to support the organization process. Section II describes the output formats of several current systems and relevant lessons learned in biomedical object recognition research. Later sections list research questions, outline the functionality of our aggregation system, describe the testbed used for evaluation, and provide some preliminary evaluation of the effectiveness of our approach.

II. Background

Processing free text into useful molecular pathway networkd is a multifaceted task. In [1] , Rzhetsky et al. summarize many of the related issues in outlining the architecture of the GeneWays system. Two of the key processes are relation extraction and biomedical object recognition.

A. Relation Extraction Output

The systems listed in Table I use various Natural Language Processing (NLP) techniques to extract the relational information from free text [2-9] . The systems at the top of the table extract triples containing two named entities and a labeled connector that describes the relationship. This format is frequently used in the visualization and automatic analysis of biomedical information [10] . Although the systems listed in the bottom half of the table create nested predicates, the GeneWays developers acknowledged that the predicates need to be unwound into binary relations (i.e. relational triples) before they can be organized into a network [5].

Manuscript received December 9, 2004. This work was supported in part by Funded by: NIH/NLM, 1 R33 LM07299-01, 2002-2005, ''GeneScene: a Toolkit for Gene Pathway Analysis''

Byron Marshall is an assistant professor at Oregon State University (e-mail: byron.marshall@bus.oregonstate.edu). The other authors are associated with the University of Arizona's Artificial Intelligence Lab in the Eller College of Business. Hua Su and Shauna Eggers are research scientists. Dan McDonald is an MIS Doctoral Candidate. Hsinchun Chen is McClelland Professor of MIS and the AI Lab Director.

B. Biomedical Object Recognition

Effectively organizing relations depends on correctly matching entities and connectors. Systems that recognize or identify biomedical name strings in text have been the subject of significant research efforts. Entity "recognition" systems find bits of text referring to biomedical objects and "identification" systems associate name strings with known biomedical objects [11] . While recognition tasks can be accomplished with approximately 80% accuracy [12] , [13] reports only 2% - 29% accuracy in matching fly gene and protein name strings to items in a corresponding lexicon of substance names. Using the GENA name dictionary described in [14] Koike and Takagi achieved better than 90% precision in identifying gene/protein/family references in text.

Biomedical named entity recognition systems face three key problems [3] : (1) new or unknown words, (2) compound word recognition, and (3) ambiguous expressions. Biomedical name strings are frequently composed of several terms [15] . To address these problems biomedical information extraction systems employ extensive lexicons and leverage character patterns and frequently occurring words. For example, the PROPER system [16] employs a list of f-terms (e.g. gene and protein) and character patterns (e.g. a numeric digit following 3 alphabetic characters) to identify the word boundaries of phrases that refer proteins.

Nearly all biomedical information extraction systems use lexical resources to identify biological object references. While some available resources (such as the Gene Ontology) implicitly or explicitly identify object classes such genes and gene products, other resources enumerate instances of those classes. For example, LocusLink lists genes and RefSeq lists genes and gene products. These lists are subject to term ambiguity where multiple substances share a common name string. Ambiguity is even more pronounced across several lexicons [17] with reported cross-dataset ambiguity between 4-20% and overlap with common English words from 0% to 2.4% [11]. This kind of ambiguity is more pronounced among those terms that are both included in the lexicons and used in MEDLINE abstracts [18] .

III. Research Questions

Our review suggests that several factors negatively impact the usefulness of the automatically extracted relations: (1) object and connector name strings use synonyms and contain potentially confusing modifiers; (2) some relations involving the same entity pairs seem to conflict with each other, especially when contextual information is ignored; (3) it is difficult to link extracted relations to other data sources (e.g., a genome, publication, or pathway databases).

This study employs a systematic approach to relation aggregation to find out how effectively we can aggregate automatically-extracted biomedical relation triples: indexing multiple references to the same object over expected variations in relational granularity, connecting relations to existing ontological resources, and capturing contextual information. An effective system would reduce the number of items needed to display extracted information, highlight relative importance based on frequency of occurrence in the biomedical texts, and format the relations for use in knowledge discovery algorithms.

IV. System Design

The BioAggregate tagger implements a feature decomposition approach to biomedical concept matching as part of the larger GeneScene system depicted in Figure 1. The Arizona Relation Parser (ARP) [9] extracts relations, the BioAggregate tagger annotates those relations with feature assignments to support aggregation, and the visualizer allows users to view the extracted and organized results.

The BioAggregate tagger decomposes a relation's entity and connector name strings by recognizing words and phrases that signal features. We will refer to such terms as feature-signaling terms. The tagger implements three noval notions: aggregatable substances, pseudo-substances, and residuals. These notions are closely related to the key entity recognition challenges identified in previous research. The key components of the tagger are lexicons and an efficient finite state automata (FSA) algorithm as depicted in Figure 2. Note how both entities and connectors in the relation are decomposed into a set of features.

TABLE I. Relation Extraction Systems

Relation extractions systems generally produce either relational triples or complex predicate relations, complex predicates are "unwound" for aggregation

	System	Method	Output
Relational Triples	Medstract: Pustejovsky et al.	Semantic Automata	• Relational Triples for Inhibition Relations
	Palakal et al.	POS Tags & HMM Co-reference Grouping	• Verb-labeled Relational Triples
	GeneScene: Leroy et al.	Sentence Parsing, FSA Emphasizes closed class words	• Relational Triples With Negation
	Arizona Relation Parser (ARP): McDonald et al.	Hybrid Syntax/Semantic Parsing	• Relational Triples With Negation, name strings frequently include several modifiers
—	GENIES: Friedman et al.	Semantic Extraction Templates	• Complex Predicate Relations are "unwound" by GeneWays into labeled binary statements
Predicates	Park & Kim	Combinatory Categorical Grammar (CCG)	• Predicate Relations e.g., Activates(A,B), or Activates(A,B,C)
	Edgar: Rindflesch et al.	Matching to UMLS	• Appears to be predicate relations
	PASTA: Gaizauauskas et al	Semantic Templates	• Emphasizes "feature" relations e.g., "mutated-p53"

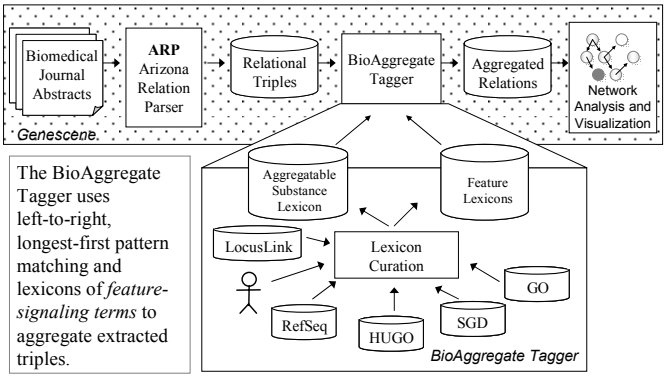

Fig. 1. The GeneScene system supports extraction, organization, and visualization of pathway relations found in the text of MEDLINE abstracts. The BioAggregate tagger organizes the relations to improve research utility.

A. Feature Lexicons

Extensive feature lexicons drive the aggregation tagging process. We manually reviewed a large number of extracted relations while developing a list of desired features. A number of possible features were considered; we preferred features that (1) frequently occur in the words near substance references and (2) correspond to features identified in existing ontologies or databases. The lexicons, which were built from existing biomedical lexicons and human generated lists, include an extensive list of substance names with cross-references to existing lexicons and smaller lists for other features. Implemented features and sources are listed in Table II. For example, the aggregatable substance lexicon was created by merging name string lists from LocusLink, RefSeq, HUGO, and SGD. Since LocusLink was recently superseded by Entrez Gene, future versions of our aggregation system will migrate to Entrez Gene. Our final lexicon includes only substance name strings unambiguously associated with a single human or yeast gene. For more details on the construction and analysis of the feature-signaling term lexicons see [18] .

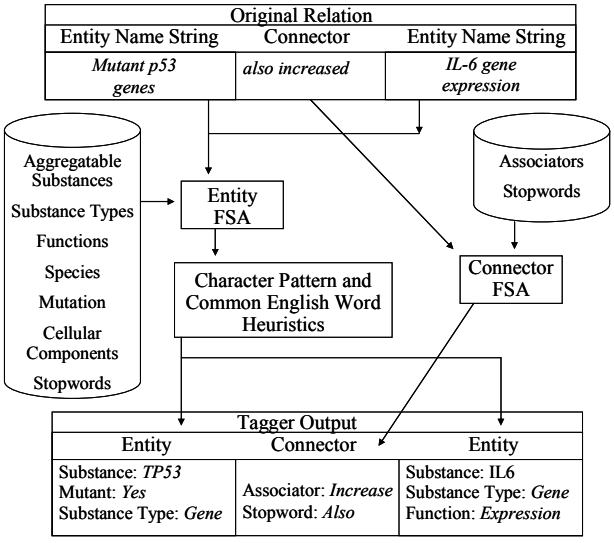

Fig. 2. The BioAggregate decompositional tagger identifies a relation's features by matches its words to terms in extensive feature term lexicons using an efficient Finite State Automata (FSA) algorithm.

Although the tagger's functionality is entity-oriented, connectors are also processed to extract associators which characterize the relations. Using simple verb stems, associators are classified into one of four types: induction, suppression, directional association, and non-directional association. Associators can appear in a relation's connector or entity name strings.

B. Finite State Automata (FSA)

Feature lexicons are loaded into a finite state automata (FSA) to perform left-to-right, longest first pattern matching. This makes the system scalable enough to perform feature identification on more than 180,000 relations in two minutes using a 1.9Ghz processor. When a feature-signaling term is found in an item's name string, it is removed from the name string and the appropriate feature is assigned to the item. Any words remaining after the name string has been processed are saved as a "residual". Because feature synonyms are stored in the lexicons, multiple references to the same substance or process are consolidated. Extracted features also help clarify context and granularity. Figure 2 highlights how decompositional tagging is applied to the entities and connectors in a relation. Longest-first pattern matching is critical, even after appropriate tokenization. For example, an entity containing the term "gene" is likely to be a gene unless the word "gene" appears as part of the phrase "gene product."

C. Aggregatable Substances

Aggregatable substance identifiers play a key role in both our aggregation and lexicon building strategies. We would

TABLE II. FEATURE LEXICONS USED BY BIOAGGREGATE
The feature-signaling term lexicons used by the BioAggregate tagger were largely extracted from existing public sources although manually adjustments were added to correct errors or provide additional needed items [18] .

Feature Lexicon	Generation Methodology
Aggregatable Substance	Compiled from LocusLink, RefSeq, HUGO, and SGD + manual adjustments
Mutation	Manually compiled (e.g., "mutated" = Mutated, "wild-type" = Non-Mutated)
Substance Type	Manually Compiled (e.g., "protein"=Protein, "oncogene"=Gene)
Associator	Stems were identified by inspecting the most common 500 verbs extracted in relations (e.g., stem "induc" identifies induced, induces, and induce) - Verbs beginning with that stem are identified as Associators
Function	*biological_function* list from the GO ontology (e.g., apoptosis) + nominalized forms of associators (e.g., induction)
Species	*common_organism* values from RefSeq (e.g., "human"=human, "baker's yeast"=yeast) + manual adjustments
Cellular Component	*Cellular_components* from the GO ontology (e.g., centriole, extracellular, and membrane) + manual adjustments
Stopword	Manually selected, common words judged to meet this standard: "ignoring this word will not mischaracterize pathway relations".

generally define an *aggregatable substance* as a gene and its related gene products (e.g. the derived RNA transcripts and proteins). The notion implies a group of substances which are similar enough to be grouped together semantically for many analytic purposes. An *aggregatable substance identifier* is a unique gene identifier as recorded in LocusLink.

Previous researchers note that a gene and its related gene products (e.g., proteins) are frequently indicated in text using a shared name string [13] , [1] . Employed lexical resources associate name strings for genes, proteins, mRNA, and other molecules with a corresponding LocusLink identifier. We associate the name strings for a gene and its products with a single aggregatable substance identifier. More than 33,000 of the 100,266 distinct name strings found in the source lists are associated with multiple substances (e.g. a gene and a protein) but are linked to only one aggregatable substance. Omitting these initially ambiguous terms would drastically reduce lexical coverage, but using the aggregateable substance paradigm we are able to retain these entries in our lexicon.

We also use several heuristics in conjunction with our aggregatable substance lexicon. Some aggregatable substance identifiers in the lexicon are also common English words. When such a name string is encountered, a capitalization filter passes only upper or mixed case terms. Thus the terms *cAMP* and *FOR* would be assigned substance identifiers while *for* would not.

The aggregatable substance approach helps address the ambiguous expression problem but also has some important analytic implications. From a user/analysis perspective, a researcher studying pathways for a particular gene might also be interested in information related to the gene's products. Practically, this methodology can assign a substance identifier to a reference even when substance type cues are not available in the text. Of course it is still valuable to distinguish between genes and proteins, so we also systematically extract and record substance types during the aggregation process.

D. Pseudo-Substances

Many useful substance name strings do not appear in our lexicon because they might be newly coined, ambiguous, or left out of the ontological resources we employed. When the tagger has finished extracting all recognized feature-signaling terms, it checks for character patterns (e.g., "starts with a letter, ends with a number".) Some normalization is done (e.g., AP1 is changed to AP-1). We assign the resulting value to the entity's *pseudo-substance* feature. Examples of pseudo-substances found in our relations include stromelysin-3, Gly82, and ESR1. The name string ESR1 maps to a pseudo-substance rather than an aggregatable substance because it is ambiguous in that it refers to both a human and a baker's yeast gene. While we would not un-ambiguously match a pseudo-substance reference to a single item in a related external resource, we can at least associate multiple occurrences of the pseudo-substance with each other. We also would like to suggest that while we do not presently use additional contextual information to resolve potentially ambiguous references, our methodology does not preclude that possibility. We have considered including this functionality either in the ARP or in the tagger. Either way has implications for architecture and performance.

E. Aggregation Levels

Once features have been assigned to the elements in a collection of relations, those relations can be better organized to support analysis. Although users employing the relations in a visual interface will be able to control the details of relation aggregation, a general framework showing increasing levels of abstraction is shown in Table III [18]. These levels of aggregation will be used as defaults in the visualization interface. Table IV shows the original and tagged feature versions of 2 relations extracted from PubMed articles PubMed 8985958, 8436340 and 9707425. The locus link id for the gene p53 is 7157. These relations would match differently at different levels of aggregation.

TABLE IV. AN AGGREGATION EXAMPLE

The relations below (extracted by the Arizona Relation Parser ARP) are equivalent under simple pathway aggregation but aggregation is blocked by the substance type of the p53 entities under typed substance aggregation and by the residual in the connector for aggregatable substance aggregation.

oncosuppressor gene p53 - are known to induce -apoptosis		
Agg. Substance: LL_7157 Subs. Type: Gene Residual : oncosuppressor	Associator: Induc Residual: known, are	Function: Apoptosis
p53 protein induces apoptosis		
Agg. Substance: LL_7157 Subs. Type: Protein	Associator: Induc	Function: Apoptosis

TABLE III. FIVE-LEVEL RELATION AGGREGATION FRAMEWORK

While an effective analysis tool allows the user to control the details of feature-based aggregation, default levels of aggregation provide a starting point for users and for measurement of aggregation effectiveness. This framework was previously proposed in [18] . Relations become more abstract when matching rules nearer the bottom of the table are applied.

Aggregation Level	Matching Rules: Entities	Matching Rules: Connectors	Possible Applications
• Baseline	complete string match		• basis of comparison
• Feature Match	all assigned features and residuals must match		• detailed pathway analysis
• Typed Substance	function or aggregatable substance and substance type must match, and residual must match	morphological forms of the same verb are combined	• pathway analysis – granularity is comparable to some human-curated databases
• Aggregatable Substance	function or aggregatable substance must match and residual must match		• explore the function of a gene and its gene products
• Simple Pathway	function or aggregatable substance must match	connector verbs are placed into one of four categories	• high level overviews and input for automated analysis

F. Comparison with other Biomedical Text Mining Tasks

Our task (decompositional feature assignment) differs from important related work such as the BioCreAtIvE competition: (1) We are focusing on organizing extracted interaction relations not tagging entity mentions. (2) We are interested in helping present pathway network results for analysis not in indexing documents to reflect the gene/protein refereneces. Last year's task 1.A task challenged teams to mark referent phrases in text, task 1.B was to identify a list of entities referenced in a text, and task 2 related functional annotation of proteins and function [12] . Rather than finding substance references in text, our tagger organizes extracted phrases into aggregatable objects. [12] notes that teams who tried to use their systems from 1.A to support other tasks had mixed results (p.2). In addition, several BioCreAtIvE teams noted that matching the exact boundaries in the test set was difficult (p.6). Our system is intended to leverage the output of such a system in a somewhat forgiving manner. This flexibility is important because there are a number of reasonably correct ways to mark up text. In BioCreAtIvE, developers had different ideas about the correct set of markable items (p. 6) and a partial review of one version of the test set resulted in .4% change in the answer key (p. 5). This work may provide some insight into how marked phrases from various sources can be effectively processed for analysis.

G. Multiple Substance Entities

Although simple binary relations among substances are very important, relationships between compound entities are also commonly expressed in the text of PubMed abstracts. Consider, for example, *EGF-mediated activation of Bmk1--requires--MAP-kinase kinase Mek5* from PubMed document 9790194. The first entity could be represented as a complex predicate to show the relationship between *EGF* (LocusLink id 1950), *Bmk1* (LocusLink id 5598), *mediated*, and *activation*. We will not speculate about the details as various representations are possible. In our system we actually capture up to two aggregatable substance identifiers for an entity, recording the last one found as the primary aggregatable substance. Because this kind of entity name is frequently right-headed (the main idea is on the right) we anecdotally observe that this is not a bad heuristic. We indexed this entity as substance LL_5598, associator activate, with secondary substance LL_1950. Depending upon how an application constructed the query, this relation could be found in a search related to either of the substances and matching could be performed on other relations involving the pair of entities. The analysis application would then be able to deal with the results as appropriate.

V. Research Testbed

We used three datasets in evaluating our system: *ARP TP53* Relations, *PROPER* entities, and *AP1* relations. *ARP TP53* is a set of 182,499 automatically-extracted relations generated by the Arizona Relation Parser from a set of 87,903 MEDLINE abstracts related to the gene TP53. The collection was created by selecting all Medline abstracts containing keywords related to TP53. Table V sows some of the relations in the *ARP TP53* dataset. Extracted relations consist of two entities, a connector and a negation indicator. The collection show very little initial overlap. We found 142,974 distinct entity names (case insensitive) and 127,397 distinct related entity pairs (ignoring connector labels and directionality). These relations were used as we designed our system's functions and constructed our lexicons. Later, this large set was used to test the effectiveness of the tagger. Admittedly, this test set is directed at pathways related to a particular gene. However, the large size of the collection should reduce the negative impact of any test set bias.

TABLE V EXAMPLES OF ARP OUTPUT

Original Sentences			
- oncogene mutant p53 suppresses apoptosis			
- mutant p53 blocked E1A-induced apoptosis			
- mutant p53 […] does not induce […] apoptosis			
Resulting Relations			
Entity 1	Negation	Connector	Entity 2
oncogene mutant p53	False	suppresses	Apoptosis
mutant p53	False	blocked	E1A-induced apoptosis
E1A	False	induced	Apoptosis
mutant p53	True	does induce	Apoptosis

The *PROPER* entities set includes 1.6 million entity name strings extracted by the PROPER system [16] from the same 87,903 TP53-related abstracts. This set is used to evaluate our system's usefulness in aggregating entities generated by systems other than ARP. Please note that PROPER does not extract relations so comparison to this data set will evaluate entity matching which is a key component of the larger relation aggregation task.

We tested the coverage of the BioAggregate tagger using 161 "gold standard" pathway relations (*AP1* relations) extracted by a biologist from 50 abstracts randomly drawn from 90,773 PubMed articles related to the AP1 family of transcription factors. These abstracts were not considered during the development of the system's functions and lexicons. We instructed the expert to select interactions between genes, gene products, and biomedical processes. We believe single expert evaluation is adequate because these experiments test feature assignment accuracy rather than relation extraction accuracy.

VI. Experimentation

We tested the BioAggregate tagger using the three previously described datasets to address three questions:

1. How frequently do various features occur in extracted relations and entities?
2. How accurately do we identify those features when they occur in the relations?
3. How much consolidation is accomplished in our network of automatically extracted relations?

A. Feature Occurrence Frequency

We ran the tagger on the *ARP TP53* relations, the entities extracted by PROPER, and on the manually-extracted *AP1* relations. Feature occurrence frequency is tabulated in Table VI. We found that many feature signaling terms are extracted by both PROPER and ARP. Thus, it is not only ARP entities which can be usefully decomposed. Although some features such as mutation occur infrequently (1.3-2.0% of all entities) without decomposition, such entities cannot be appropriately matched. Our approach found a substance, pseudo-substance, or function identifier in more than half of the entities we evaluated. As previously noted, we also extract associators from relation connectors. Most (91.4%) of the connectors in the ARP relations were matched with one of the verbs in our associator lexicon.

TABLE VI. FEATURE OCCURRENCE FREQUENCIES
We found our selecetd features in many of the extracted entities evaluated. The tagging task is relevant to both the ARP and PROPER entities.

Feature	ARP Entities	PROPER Entities
Number of Items Tagged:	364,998	1,600,223
Aggregatable Substance (e.g., *P53*)	30.1%	39.9%
Pseudo-Substance (e.g., *Gly28*)	5.9%	11.9%
Mutation (e.g., *wild-type*)	2.0%	1.3%
Substance Type (e.g., *protein*)	27.9%	17.2%
Function (e.g., *apoptosis*)	19.5%	2.3%
Species (e.g., *human*)	2.8%	1.9%
Cellular Component (e.g., *membrane*)	10.7%	2.6%
Substance, Pseudo-Substance, or Function	51.2%	52.8%

B. Feature Assignment Accuracy

To measure the accuracy of our feature assignments, we randomly selected 100 examples for each feature from the *ARP TP53* relations. Our expert checked the results by reviewing the relations and looking up items in the source documents. In a few cases, the expert was not sure if the item should have been assigned the feature. Excluding these items, feature assignment accuracy was 95% or better for all extracted features. Alternatively, if we consider the ambiguous items to be wrong, function accuracy was still 90% and aggregatable substance accuracy was 87%.

TABLE VII. FEATURE ASSIGNMENT ACCURACY (RECALL AND PRECISION) OF THE BIOAGGREGATE TAGGER ON 161 AP-1 RELATIONS

	Number of Items			Accuracy	
	(A) Gold Standard	(B) Found	(C) Correct	(C) / (B) Precision	(C) / (A) Recall
Entities					
Aggregatable Substance	208	131	107	81.7%	**51.4%**
Substance Type	76	75	73	97.3%	96.1%
Function	37	30	30	100.0%	81.1%
Associator	43	42	41	97.6%	95.3%
Associator Type	34	34	34	100.0%	100.0%
Mutant	6	6	6	100.0%	100.0%
Species	1	1	1	100.0%	100.0%
Cellular Component	13	13	13	100.0%	100.0%
Connectors					
Associator	177	159	150	94.3%	84.7%
Associator Type	123	123	123	100.0%	100.0%

A second consideration in assignment accuracy is coverage. That is, should we have assigned features to more entities? To address this question, our expert reviewed the tagger output for the *AP1* relations. Table VII compares the tagger feature assignments ("Found") to the expert assignments ("Gold Standard"). When the "Found" assignment matches the "Gold Standard" it is considered "Correct". We report 51.4% recall for aggregatable substances. This is an encouraging result given previous results although it should be noted that various methodologies have been used in different studies to evaluate entity recognition so results are not directly comparable. We counted the number of correctly assigned aggregatable substance identifiers and divided by the total number of references to specific genes or gene products. We did not consider pseudo-substance tags to be correct; we only counted correctly associations between aggregatable substance references and LocusLink or RefSeq identifiers.

C. Network Consolidation

While correctly labeling entity features helps capture context, index multiple references to the same substance, and connect extracted relations to external resources, we also hope to see a significant level of network consolidation as a result of relation aggregation. Consolidated networks should result in more focused and concise knowledge representations for visualization and analysis. Because previous studies have not measured biomedical extraction systems from this perspective, we selected a variety of network consolidation measures as a baseline for future evaluation. To get an idea of how much consolidation takes place in a set of aggregated relations we chose a subset for comparison. We chose relations where:

- Every entity has an identified aggregatable substance, a pseudo-substance, or a function.
- Each entity has only one substance or function.
- The connector contains a recognized associator.

These rules control the population for potentially confounding characteristics. The filtered set includes 44,864 of the 182,499 in the *ARP TP53* dataset and might be comparable to those relations that would be relatively important for various analysis tasks.

Figure 3 reports the number of distinct items and relations found at each of the aggregation levels. Please note that we report Typed Substances twice: once as described in the framework and once ignoring residuals (words that were not recognized during the tagging process). We chart these separately to show the strong impact residual words had on entity aggregation. The number of distinct entities decreases somewhat at each successive level of aggregation as does the number of disjoint relations. In this analysis, a disjoint relation is one where neither entity is found in any other relation. An analysis routine would be unable to connect such a relation to other information in the network. Table VIII lists a variety of network consolidation measures that might be of interest to

the reader, including the distinct item measures displayed in Figure 3.

VII. An Example

We queried the aggregated relations for some key substances involved in important pathways related to TP53. Panel A in Figure 4 shows a small part of the unaggregated network (29 out of 277 nodes). Although many relations are displayed, it is difficult to put useful information together because the substances are represented by different strings (p53, wild-type p53, p53 levels, and transcriptional activities of p53, etc.) and redundant relations exist (p73 activated MDM2, p73 transactivate mdm2 promoter, etc.). Panel B depicts the same set of relations after aggregation. The network density is dramatically reduced. Each node in this network represents a unique gene or protein. It is easier to identify all relations between two genes or among multiple genes. For instance, it is straight forward to identify the feedback loop existing between p53 and MDM2 (TP53 Activates Mdm2 and Mdm2 Inhibits TP53). Original relations used to generate the aggregated picture are available by a mouse over, as shown in the box of Panel B for the relation of TP53 Activates p21.

VIII. Discussion and Future Directions

To evaluate extracted knowledge networks algorithms might be applied to rank the credibility of identified relations or detect apparent conflicts to chart changes of understanding over time or generate new hypotheses. Clearly, some relations are stated more than once in the literature, but they are generally stated in different terms. Aggregation allows us to consolidate. We would expect that a relationship found five or more times in the literature is likely to be "true" (that is, not the result of some extraction error and confirmed in more than one study). Table VIII shows that 3% of the "simple pathway" relations were found 5 or more times as compared to virtually no repetition in the "baseline" relations.

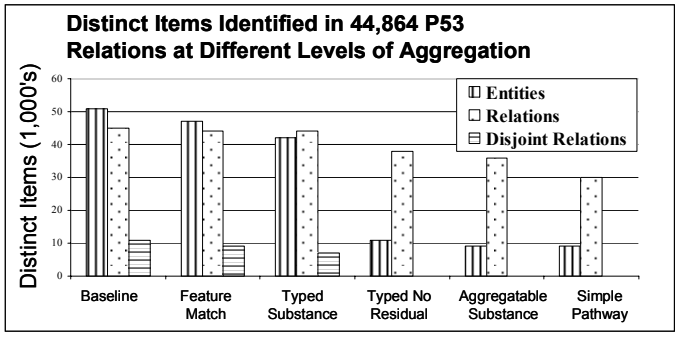

Fig. 3. Distinct items at various levels of aggregation

A substantial number of the relations we evaluated can be correctly mapped to Entrez Gene identifiers using our methodology. This means that a visualization system or analysis application can integrate relations from manually curated databases. Our initial observations suggest that many, or most, of the relations expressed in PubMed abstracts are expressed at higher levels of granularity as compared to those captured in manually created databases. Thus we expect that the two kinds of relations will be complementary.

Aggregation is important if we are to effectively use automatically extracted relations. The experiments and examples we report in this work suggest that decompositional aggregation is a promising methodology:

- Extracted information can be matched to external resources with reasonable accuracy.
- Networks of extracted relations can be significantly consolidated with references to the same biological object indexed at several levels of granularity.
- More concise visualizations of the same information can be created.

Our approach is intended to support flexible query responses. For example, a query related to p53 can be optionally specified to include items where p53 is know to be mutated, know to be normal, or where the mutation feature is unknown. This organizational paradigm adjusts to the ambiguity inherent in free-text sentences and NLP techniques without abandoning the possibility of detailed analysis.

Still, much additional development is needed. Thorough and detailed analysis of matching errors can be used to tune the aggregation process. For example, preliminary evaluation suggests that accounting for homologs in the lexicon construction process would further reduce ambiguity. Because genetic activity is frequently studied in a cross-species environment, this kind of indexing has promising implications. The species issue might also need to be addressed more generally. This important point is highlighted in [14] which uses different resources to identify entities for different species. Cross-species lexical ambiguity is substantial but we use only a single aggregatable substance

TABLE VIII. Network Consolidation Measures

	Baseline	Feature Match	Typed Substance	Typed, no Residual	Aggregatable Substance	Simple Pathway
Distinct Entities	51,033	46,547	41,628	11,362	8,837	8,837
Average occurrences per entity	1.76	1.93	2.16	7.90	10.16	10.16
Distinct entities occurring 5+ times	3.0%	3.5%	4.1%	16.8%	18.9%	21.6%
Distinct relations	44,864	44,494	43,721	38,365	36,051	29,635
Average occurrences per relation	1.00	1.01	1.03	1.17	1.24	1.51
Distinct relations occurring 5+ times	0.00%	0.02%	0.08%	0.86%	1.28%	3.07%
Average number of different name strings per entity	1	1.1	1.2	4.5	5.8	5.8
Network density (linked entity pairs / possible entity pairs * 10,000)	.33	.40	.48	4.4	6.2	6.2
Number of disjoint relations	10,608	8,690	6,817	383	222	222

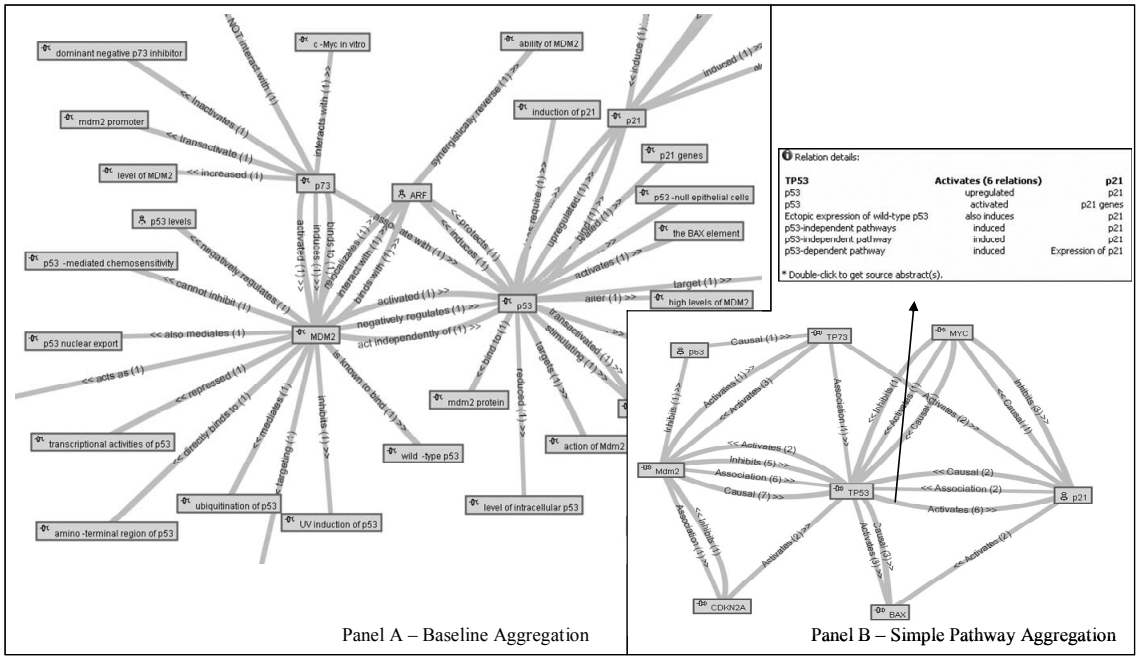

Fig. 4. (A) Visualization of a set of p53 relations before aggregation (baseline); (B) The same set of relations visualized after aggregation (simple pathway).

lexicon focused on human and yeast name strings. Additional contextual information might be effective in identifying the species discussed in an article or referenced in a sentence. This designation could then be used to guide the tagging process to increase the number of correct aggregatable substance matches.

Other extractable and interesting features can be used within the framework. We also plan a more extensive evaluation that addresses how accurately the relations are matched at each level rather than measuring only feature assignment accuracy and the resulting network consolidation.

ACKNOWLEDGMENT

The authors thank Gondy Leroy, Chun-Ju Tseng (Lu), Riyad Kalla and the team who helped establish the early versions of the GeneScene system and interface.

REFERENCES

[1] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. A. Duboue, W. Weng, W. J. WIlbur, V. Hatzivassiloglou, and C. Friedman, "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data," *J. Biomed. Inform.*, vol. 37, pp. 43-53, 2004.

[2] J. C. Park, H. S. Kim, and J. J. Kim, "Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar.," presented at Pac. Symp. Biocomp., Hawaii, USA, pp. 396-407, 2001.

[3] M. Palakal, M. Stephens, S. Mukhopadhyay, R. Raje, and S. Rhodes, "Identification of biological relationships from text documents using efficient computational methods," *J. Bioinform. Comput. Biol.*, vol. 1, pp. 307-342, 2003.

[4] R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett, "Protein structures and information extraction from biological texts: the PASTA system.," *Bioinformatics*, vol. 19, pp. 135-143, 2003.

[5] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles.," *Bioinformatics*, vol. 17, pp. S74-82, 2001.

[6] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, and B. Cochran, "Robust relational parsing over biomedical literature: extracting inhibit relations," presented at Pac. Symp. Biocomput., Lihue, Hawaii, pp. 362-373, 2002.

[7] T. C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter, "EDGAR: extraction of drugs, genes and relations from the biomedical literature," presented at Pac. Symp. Biocomput., Big Island, Hawaii, pp. 517-528, 2000.

[8] G. Leroy, H. Chen, and J. D. Martinez, "A shallow parser based on closed-class words to capture relations in biomedical text," *J. Biomed. Inform.*, vol. 36, pp. 145-158, 2003.

[9] D. M. McDonald, H. Chen, H. Su, and B. B. Marshall, "Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser," Bioinformatics (Forthcoming), 2004.

[10] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," presented at Pac. Symp. Biocomput., Big Island, Hawaii, pp. 300-311, 2004.

[11] O. Tuason, L. Chen, H. Liu, J. A. Blake, and C. Friedman, "Biological nomenclatures: a source of lexical knowledge and ambiguity," presented at Pac. Symp. Biocomp., Big Island, Hawaii, pp. 238-249, 2004.

[12] A. S. Yeh, L. Hirschman, M. Morris, and M. Colosimo, "BioCreAtIve task 1A: gene mention finding evaluation," The MITRE Corporation, Bedford, MA (2004).

[13] L. Hirschman, A. A. Morgan, and A. S. Yeh, "Rutabaga by any other name: extracting biological names," J. Biomed. Inform., vol. 35, pp. 247-259, 2002.

[14] A. Koike and T. Takagi, "Gene/protein/family name recognition in biomedical literature," presented at HLT/NAACL BioLINK Workshop, Boston, pp. 9-16, 2004.

[15] P. V. Ogren, K. B. Cohen, G. K. Acquaah-Mensah, J. Eberlein, and L. Hunter, "The compositional structure of Gene Ontology terms," presented at Pac. Symp. Biocomput., Big Island, Hawaii, USA, pp. 214-225, 2004.

[16] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, "Toward information extraction: identifying protein names from biological papers," presented at Pac. Symp. Biocomput., Big Island, Hawaii, pp. 705-716, 1998.

[17] D. Hanisch, J. Fluck, H. Mevissen, and R. Zimmer, "Playing biology's name game: identifying protein names in scientific text," presented at Pac. Symp. Biocomput., Lihue, Hawaii, USA, pp. 403-414, 2003.

[18] B. B. Marshall, H. Su, D. M. McDonald, and H. Chen, "Linking ontological resources using aggregatable substance identifiers to organize extracted relations," presented at Pac. Symp. Biocomput.. Big Island, Hawaii, 2005.