Automated Phenotype-Genotype Table Understanding

Shifta Ansari, Robert E. Mercer, and Peter Rogan

Scholarly writing in the broad area of experimental biomedicine is a genre that has a rhetorical style that exhibits some easily identifiable stylistic features: division of the paper into well-defined sections (Introduction, Methods, Results, Discussion), and the use of tables and figures to organize and express important results. Tables and figures have stylistic features, as well: titles, captions, content.

In addition to these common stylistic features, the community-accepted rhetorical style for authors of scientific papers is to publish their experimental findings in a tabular form, because the quantity of experimental data is large and the tabular arrangement allows for a concise presentation of the relationships among the data and for a rapid understanding of the results. Thus, tables are one of the most important sources of information.

The rapid advancement of knowledge in the biomedical field [1] has led to significant efforts to extract information from papers and interpret it automatically. Our contribution to this effort is a general approach not only to access and extract information from tables, but also *to understand* the information contained in tables by semantically grounding it with the appropriate concepts in an ontology, and to make it available for further use.

We report here our phenotype-genotype table understanding undertaking. To summarize: We have curated papers from the domain of genetics that discuss phenotype, genotype, mutation, and gene, and their relationships, syndrome (constellation of phenotypes), and disease, to design an ontology that captures the concepts needed to understand these tables and to engineer a tool which populates this ontology with data reported in these tables.

Shifta Ansari · Robert E. Mercer
Department of Computer Science, The University of Western Ontario, London, ON, Canada
e-mail: sansar6@uwo.ca, mercer@csd.uwo.ca

Peter Rogan
Department of Biochemistry, The University of Western Ontario, London, ON, Canada
e-mail: progan@uwo.ca

M. Ali et al. (Eds.): *Contemporary Challenges & Solutions in Applied AI*, SCI 489, pp. 47–52.
DOI: 10.1007/978-3-319-00651-2_7

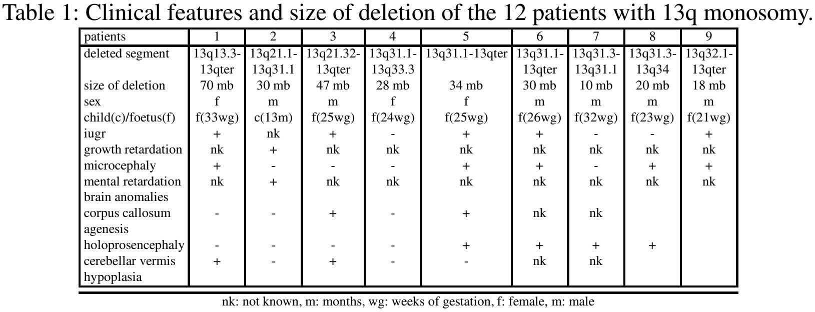

Table 1: Clinical features and size of deletion of the 12 patients with 13q monosomy.

patients	1	2	3	4	5	6	7	8	9
deleted segment	13q13.3-13qter	13q21.1-13q31.1	13q21.32-13qter	13q31.1-13q33.3	13q31.1-13qter	13q31.1-13qter	13q31.3-13q31.1	13q31.3-13q34	13q32.1-13qter
size of deletion	70 mb	30 mb	47 mb	28 mb	34 mb	30 mb	10 mb	20 mb	18 mb
sex	f	m	m	f	f	m	m	m	m
child(c)/foetus(f)	f(33wg)	c(13m)	f(25wg)	f(24wg)	f(25wg)	f(26wg)	f(32wg)	f(23wg)	f(21wg)
iugr	+	nk	+	-	+	+	-	-	+
growth retardation	nk	+	nk	nk	nk	nk	nk	nk	nk
microcephaly	+	-	-	-	+	+	-	+	+
mental retardation	nk	+	nk	nk	nk	nk	nk	nk	nk
brain anomalies									
corpus callosum agenesis	-	-	+	-	+	nk	nk		
holoprosencephaly	-	-	-	-	+	+	+	+	
cerebellar vermis hypoplasia	+	-	+	-	-	nk	nk		

nk: not known, m: months, wg: weeks of gestation, f: female, m: male

Fig. 1 A (horizontal) table from the development set corpus, modified to fit the page

1 Contributions to Table Information Understanding

As an example of what needs to be done to extract information from a table, the table in Fig. 1 shows the three types of information: title, caption (or footnotes), and content. The content is contained in *cells*. The table information understanding problem can be seen as extracting and providing a semantics for this information.

Our contributions are a phenotype-genotype table ontology, a reading of the table cells that maintains relationships among the cells, and a tool that populates the ontology with the information extracted from the phenotype-genotype tables found in scholarly biomedical articles. Details of the first two contributions follow.

The Phenotype-Genotype Table Ontology. The process of building a phenotype-genotype table ontology required the curation of a sufficient number of texts containing a variety of tables in order to provide credence for our ontology. Using 107 tables found in 50 papers curated from our selected domain as our development set, we have engineered a table ontology that extends a subpart of the UMLS ontology with new concepts that are present in these tables (e.g. subjects have an age) as well as other fundamental concepts. This ontology, a portion is shown in Fig. 2, provides a semantics for the table data. In addition to storing the table data, the relationship among the cells (shown by the red lines), which is important information conveyed by the table structure can also be maintained.

Our ontology reproduces appropriate pieces of the UMLS ontology. These parts of our ontology are verified simply by the acceptance of the UMLS ontology. For concepts like genotype and phenotype, which are not in the UMLS ontology we have used an expert's knowledge to assist us. For example, we have added a concept named ORGANISM to accommodate certain important classes, like PATIENT and FAMILY, and certain important attributes for them, like AGE, GENDER, BIRTH WEIGHT, etc. To accommodate cell values that act as an identifier for the data in a column or row (e.g. PATIENTS, as in Fig. 1), we designed a generic concept named IDENTIFICATION ENTITY.

Table Information Understanding. To understand the table information, we begin with Hurst's concept of a reading path [2] to associate the *access cells* and the *data*

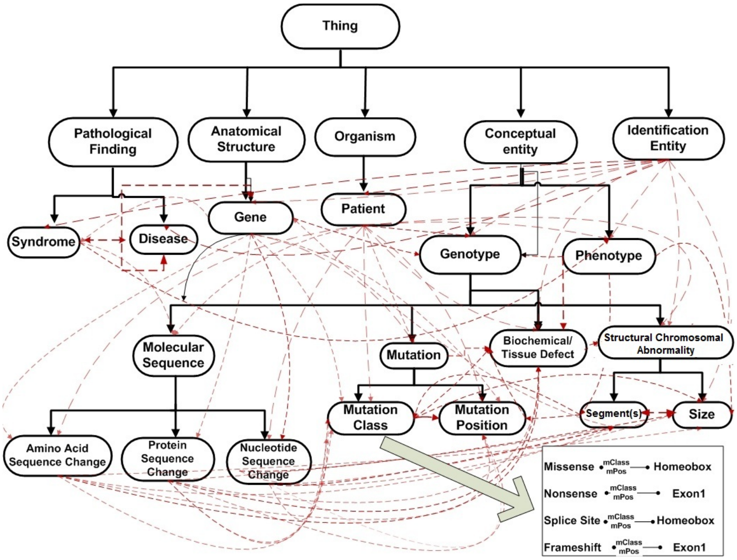

Fig. 2 A portion of the proposed ontology generated from the curated tables

cells. Data cells are characterized as cells where data appears in a terminating role. Access cells, which can be conceptualized as descriptors of the data cells, are the remaining cells. The reading of a table is the reading of all of the data cells.

Hurst views table data cells as independent entities. We realize that for the type of tables that are in our corpus, the data cells are connected (for instance, a column or row represents data about a single patient). These connections need to be maintained in the ontology in order to have a complete interpretation of the table. Hurst's concept of reading path is modified to maintain these relationships. In some cases to achieve the correct interpretation of the data cell values in our corpus, the primary relationship among the access cells may need to be changed from a vertical (column) relationship to a horizontal (row) relationship. To accomplish this, we change the order of the sequence of the reading path. For example, for Fig. 4, the primary headings are in columns and the data are distributed for the subjects down the columns. However, for Fig. 1, the primary headings are in the rows and the data are distributed across the rows. We refer to the previous column-oriented table as a vertical table and the latter row-oriented table as a horizontal table.

Heuristics are required to distinguish horizontal and vertical tables automatically. After observing our collection of tables, we designed the following heuristics: horizontal tables usually have many more columns than vertical tables; the first column heading starts with the Family or Patient ID followed by Age/Gender/Weight/Height attributes of the patient as row headings; and the remaining column headings contain attributes such as patient numbers or Identification numbers rather than alphabetic attributes (see Fig. 1). Very often, the data cell values are expressed as symbols.

Using the reading path concept we can reach a particular data cell. For example, considering the table in Fig. 4 the reading path for the data cell in row 2 column

access cell c_{00} (column/row heading 0)	access cell c_{01} (column heading 1)	access cell c_{02} (column heading 2)	$\cdots$ $\cdots$	access cell c_{0n} (column heading n)
access cell c_{10} (row heading 1)	data cell c_{11}	data cell c_{12}	$\cdots$	data cell c_{1n}
$\cdots$ $\cdots$ $\cdots$ $\cdots$				
access cell c_{m0} (row heading m)	data cell c_{m1}	data cell c_{m2}	$\cdots$	data cell c_{mn}

Fig. 3 Abstract table indicating access cells and data cells (m rows, n columns)

Table 1: HLXB9 Mutations Identified in the Study and Associated Phenotypes

Mutation Class	Mutation Position	Nucleotide Change	Amino Acid Change	Clinical Phenotype	Family or Patient No.
Missense	Homeobox	C→G, nt 4171	R247G	Hemisacrum, ARM, presacral mass, perianal abcess	3
Splice Site	Homeobox	A→G, nt 4889	NA	Hemisacrum, ARM, presacral mass, nonpenetrance	16
Frameshift	Exon 1	Ins C, nt 125-30	NA	Hemisacrum, ARM, presacral mass, neurogenic bladder, nonpenetrance	20

Fig. 4 Example of a vertical table (reduced in size) from the corpus of 115 phenotype-genotype tables

2 (containing value Homeobox) is: Mutation Class → Missense → Mutation Position → Homeobox. Using this simple reading path, we are able to insert the values Missense and Homeobox and the relation between them in the ontology. However, using the reading path concept for each data cell, it would not be possible to retain the relation between Homeobox and the other cells in the row. Instead, we combine all of the reading paths of all cells in a row, taking common terms only once.

According to our revised reading path, the reading path for the second row of the table in Fig. 3 is: cell c_{00} → cell c_{10} → cell c_{01} → cell c_{11} → cell c_{02} → cell c_{12} → ... cell c_{0n} → cell c_{1n}. If we apply this to the table in Fig. 4 we get: Mutation Class → Missense → Mutation Position → Homeobox → Nucleotide Change → C→G, nt 4171 → amino acid change → R247G → Clinical Phenotype → Hemisacrum, ARM, presacral mass, perianal abcess → Family or Patient No. → 3.

Now we search for the first term MUTATION CLASS in the ontology and once we find it we insert the second value "Missense" under the class MUTATION CLASS. Similarly, we enter other concept-value pairs from the reading path. In this way we can populate our ontology appropriately. Moreover, from the reading path we know how data cells are connected with each other. We reflect this connection into our ontology by creating a relationship and connect data cells with it. As an example, to preserve the relationship between data cells under MUTATION CLASS and MUTATION POSITION we build a relationship named "mClassmPos" and connect data cells with it, which is illustrated in Fig. 2.

For a horizontal table we need to change the order of the reading path: cell c_{00} → cell c_{01} → cell c_{10} → cell c_{11} → cell c_{02} → cell c_{12} → ... cell c_{0n} → cell c_{1n}. Considering the horizontal table in Fig. 1, the reading path for row 3 is: Patients → 1 → Size of deletion → 70 Mb → 2 → 30 Mb → 3 → 47 Mb → 4 → 28 Mb → 5 → 34 Mb → 6 → 30 Mb → 7 → 10 Mb → 8 → 20 Mb → 9 → 18 Mb Now we can populate the ontology as we did for Fig. 4.

To solve the problem of missing column headings, we take the caption of the table as a source of the appropriate heading. To confirm this choice we check whether the values of that column fall under this concept. In the future, this will be performed automatically, but here, these two steps were performed manually.

We have observed column (or row) heading that have values like “Classes of Mutation” or “Position of Mutation”, instead of directly matching the name of the concept in our ontology (Mutation Class and Mutation Postion). In these cases we check the possible word variations and further confirm our choice by observing the often highly stylized form of the data associated with that heading, making certain that it corresponds to the ontology concept that we have chosen.

2 Evaluation

The system correctly populates the ontology with the information contained in the development set of 107 genotype-phenotype tables.[1] The proposed ontology and population method are further verified by populating the ontology with the data from 31 previously unseen vertical tables, curated from 17 papers using the same keyword search, comprising 150 columns in total.[1] Column headings should map to concepts in the ontology. To calculate the accuracy of our system we consider the number of columns that successfully map into the table ontology, success being marked by the software finding a concept to map to and correctness of the concept being verified by human judgement. According to this criteria, the 120 correctly mapped columns gives the following accuracy: $\frac{120}{150} * 100\% = 80\%$. The causes of missed or incorrect interpretations for the 30 columns by our current system are summarized below.

Headings in 18 columns representing 10 distinct concepts do not map to a concept in our current ontology. In 5 cases the column header refers to a concept, but the values in the column belong to an aspect or property of the concept or a different concept. For 5 cases the column heading synonym list is inadequate for the mapping to the correct concept that exists in the ontology. In two other cases we found one column missing a column heading and one column heading which is actually a combination of two concepts joined by “and”.

The first problem, finding concepts in tables that are not in our ontology, has been anticipated: the ontology is meant to evolve, especially in its gestation period. Most of the mapping problems encountered by our system will be overcome with appropriate updates to the ontology, which include having a good base of linguistic synonyms that map to the same ontological concept. We are currently investigating automatic and semi-automatic methods for adding concepts to the ontology. The second problem is much rarer. We already have procedures in place to confirm the mapping of column/row headers using the data values in the column/row (for instance, we do this for the missing column header problem).

[1] `http://www.csd.uwo.ca/~mercer/PhenGenTable-corpus-bibliography` provides a bibliography of the 67 papers and `http://www.csd.uwo.ca/~mercer/PhenGenTable-corpus`, the corpus of 138 tables.

3 Related Work

Wong *et al.* [4] provides an automated system to extract information about mutations (gene, exon, mutation, codon and related statistics) from tables. They classify the table data to map the column/row values to a relevant entity, and then extract mutation information from these data. Mulwad *et al.* [3] introduces a domain independent framework for the intended semantics of tables. Column headers are mapped to class labels from an ontology; relationships between columns are discovered; cell values are linked to Linked Open Data entities and appropriate linked data.

In comparison, our work provides a domain-based ontology to store not only the data from the table but also the relationships that hold among the data cells. Furthermore, we are interested in a broader range of concepts for our ontology than the first work: mutation, gene, exon, phenotype, genotype, disease, syndrome.

4 Conclusions and Future Work

This paper reports on a table ontology designed to represent the tabular information in phenotype-genotype tables in scholarly biomedical papers. We extend the reading path concept to make it functional for our concept of table orientation, to populate the ontology with data and to preserve the various relationships among the table data. The populated ontology represents the semantics of each piece of information and preserves the relationship among cells in the table.

For future work, adding concepts automatically or semi-automatically to the incomplete ontology needs investigation. Unanticipated complications encountered during evaluation need to be addressed. As well, two issues arose that address aspects of the design and population of ontologies in a more general way. Firstly, we discovered in our evaluation phase, one table that referred to *the lack* of a mutation. Secondly, our biomedical expert has pointed out that knowledge changes over time and this is reflected in how the data is reported (e.g. epigenetic changes are not understood as well as sequence-based or structural chromosomal changes; and uncertainty in interpretation will be communicated in inconsistent ways). Our ontology will have to address this temporal aspect to record information of these types.

References

1. Hunter, L., Cohen, K.B.: Biomedical language processing: What's beyond PubMed? Molecular Cell 21(5), 589–594 (2006)
2. Hurst, M.F.: The interpretation of tables in texts. PhD thesis, University of Edinburgh (2000)
3. Mulwad, V., Finin, T., Joshi, A.: A domain independent framework for extracting linked semantic data from tables. In: Ceri, S., Brambilla, M. (eds.) Search Computing. LNCS, vol. 7538, pp. 16–33. Springer, Heidelberg (2012)
4. Wong, W., Martinez, D., Cavedon, L.: Extraction of named entities from tables in gene mutation literature. In: Workshop on BioNLP, pp. 46–54 (2009)