BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 23 2008, pages 2760–2766
doi:10.1093/bioinformatics/btn502

Databases and ontologies

A bioinformatics analysis of the cell line nomenclature

Sirarat Sarntivijai[1], Alexander S. Ade[1], Brian D. Athey[1,2] and David J. States[1,3,*]

[1]National Center for Integrative Biomedical Informatics and the Center for Computational Medicine and Biology, [2]Department of Psychiatry and [3]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA

Received on March 26, 2008; revised on August 10, 2008; accepted on September 19, 2008

Advance Access publication October 10, 2008

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Cell lines are used extensively in biomedical research, but the nomenclature describing cell lines has not been standardized. The problems are both linguistic and experimental. Many ambiguous cell line names appear in the published literature. Users of the same cell line may refer to it in different ways, and cell lines may mutate or become contaminated without the knowledge of the user. As a first step towards rationalizing this nomenclature, we created a cell line knowledgebase (CLKB) with a well-structured collection of names and descriptive data for cell lines cultured *in vitro*. The objectives of this work are: (i) to assist users in extracting useful information from biomedical text and (ii) to highlight the importance of standardizing cell line names in biomedical research. This CLKB contains a broad collection of cell line names compiled from ATCC, Hyper CLDB and MeSH. In addition to names, the knowledgebase specifies relationships between cell lines. We analyze the use of cell line names in biomedical text. Issues include ambiguous names, polymorphisms in the use of names and the fact that some cell line names are also common English words. Linguistic patterns associated with the occurrence of cell line names are analyzed. Applying these patterns to find additional cell line names in the literature identifies only a small number of additional names. Annotation of microarray gene expression studies is used as a test case. The CLKB facilitates data exploration and comparison of different cell lines in support of clinical and experimental research.

Availability: The web ontology file for this cell line collection can be downloaded at http://www.stateslab.org/data/celllineOntology/cellline.zip.

Contact: dstates@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Cells cultured *in vitro* are powerful and convenient model systems that are widely used in biomedical research. In the past year alone, some 30 000 papers were published on work using cell lines. Often, a great deal of biological information is associated with the particular cell line used for analysis, and knowledge of this context is important to fully understand the implication of a publication. The rapid growth of biotechnology and biomedical research is generating massive collections of free text data that would benefit from improved data organization and management. In this article, we describe a simple ontology structure that underlies this knowledgebase for cell lines and its application to machine-assisted analysis of biomedical literature.

The names of cell lines grown in culture are frequently generated by the laboratory of origin and have not been subject to systematic organization. As a result, the nomenclature is inconsistent and sometimes ambiguous. For instance, there are numerous examples where the same name is applied to at least two different cell lines and many cases where different names are applied to a single cell line. Improved organization of cell line names would be beneficial to the clinical and experimental research communities. Traditionally, a researcher decides on which cell line he/she will use in addressing the question at hand based on historical knowledge ('what is known to work under the given condition?'). However, in scientific research, the original plan may not work out and different approaches need to be considered. Thus, researchers often need to explore alternative cell lines and to access information about these cell lines in planning their experiments. Repositories of cell lines exist, and names used by these repositories provide a useful starting point for a nomenclature. Although repositories typically retain the name provided by the original developer of a cell line, they often add catalog numbers or other *de facto* synonyms, further complicating matters. The nature of how cell lines are used in biomedical research leads us to the classic informatics question 'How can we transform text data with cell line names into well-structured information about these cell lines?' Another issue in information management is quality assurance and quality control, 'How do we know if we are looking at an accurate set of data?' Even though there exists an ontology for related information of cell types and other biomedical artifacts (Bard *et al*., 2005; Shulz *et al*., 2006), we are still missing another important component of cell line information; and thus, our attempt to create a cell line knowledgebase (CLKB). During the construction of CLKB, we have encountered issues that need to be addressed in the community in order to create structured knowledge about cell lines to answer questions in information-seeking domains. The structure of GENIA ontology also suggests that cell line information from our CLKB can be integrated to create an enhanced network of information (Rinaldi *et al*., 2006; Shulz *et al*., 2006).

2 METHODS

The CLKB is an initial attempt to normalize the cell line nomenclature. We draw data from Hyper Cell Line Data Base (HyperCLDB) version 4.200201 (http://www.biotech.ist. unige.it/interlab/cldb.html) (Manniello *et al*., 1996; Parodi *et al*., 1993; Romano *et al*., 1993), and the American

*To whom correspondence should be addressed.

Table 1. Examples of synonymous cell lines

CellLineName	ATCC No.	HyperCLDB html
2HX-2		cl51.html
2Hx-2	HB-8117	
34-5-8 S		cl5138.html
34-5-8S	HB-102	
3T3 L1		cl71.html
3T3-L1		cl72.html
3T3 Swiss Albino		cl83.html
3T3-Swiss albino	CCL-92	
Swiss-3T3		cl4451.html
3T6		cl86.html
3T6 Swiss Albino		cl89.html
3T6-Swiss albino	CCL-96	
4/4 R.M.-4	CCL-216	
4/4 RM-4		cl91.html
72 A1		cl5143.html
72A1	HB-168	
7C subscript(2) C subscript(5) C subscript(12)	HB-8678	
7C2C5C12		cl5111.html
BALB 3T3 clone A31		cl386.html
BALB/3T3 clone A31	CCL-163	

A complete table of synonymous cell lines is also available in Supplementary Material.

Type Culture Collection (ATCC) cell line catalog (available online at http://www.atcc.org/common/documents/pdf/CellCatalog/CellIndex.pdf as of November 2006), and link names derived from these collections to the National Library of Medicine Medical Subject Headings (NLM MeSH, 2007). The focus of this project lies in permanent cell lines, not primary cells. HyperCLDB html files were downloaded and stored in a local storage for data processing. A python script was written to parse out cell line names and their corresponding organism, tissue, pathology and tumor information, and store the information on our database server. There were 6609 cell line records including duplicates (cell lines with the same label deposited by multiple laboratories) on HyperCLDB. The collection was then processed to eliminate duplicates keeping only one record for a specific cell line name with reference pointers to the name duplicates. The unique-name version of HyperCLDB contained 5888 distinct cell lines.

All data were processed in a case-sensitive manner; in the non-standardized cell line annotation across different laboratories, capitalization may signify different representations for different cell lines. Thus, we could not normalize capitalization for natural language processing. On the other hand, different capitalization of the same spelling often does not indicate different cell lines. Such variants are treated as synonyms of the cell line name as shown in Table 1. An advanced algorithm for machine learning would be required in this case. The capitalization inconsistency was manually examined and curated at a later stage, as the machine could not easily distinguish the differences without human intervention (e.g. the case of Hep-2 and HEp-2). ATCC cell lines were converted from PDF format to Microsoft Excel file format. We manipulated the data obtained from ATCC catalog on Microsoft Excel utilizing VBA macro rather than other programming scripts, as some of the ATCC cell line names contained Unicode characters. Implementing text processing using Python or Perl scripting does not usually handle Unicode characters, and converting Unicode encrypting to another workable format adds complexity to calculation (see Supplementary Material). A total of 3488 cell lines were extracted from ATCC catalog.

The attributes for each ATCC cell line were cell line name, ATCC number, species, source/application, morphology and growth mode. The information stored in association with each ATCC cell line was then processed to a format compatible with the HyperCLDB dataset. After text processing, each cell line record has the attributes of CellLineID, Organism, Tissue, Pathology, Growth Mode, Repository Source, ATCC Number, HyperCLDB html and MeshID. For each attribute used in an individual record, standard nomenclature is applied to normalize textual content for a better organization of the knowledgebase (i.e. we try to use a controlled list of vocabulary to describe definition of terms being used in the knowledgebase; e.g. using NCBI taxonomy to describe organism, using NCBI MeSH terms where applicable, etc.)

The 8914 cell line names from both sources were combined in alphanumeric order. Cell lines that appear to share similar names (but different punctuation marks, and/or space) are examined; if they share the same characteristics (same CellLineID, Organism, Tissue and Pathology), the entries are merged to one primary entry and its cross-reference identifiers are stored in our data table. If a group of names appear to share only the similar names but their characteristics differ, the entries are kept as is. Table 1 provides examples of merged names. Repository source is the name of primary source that the cell line is taken from. If related names for a single cell line exist in both repositories, we use ATCC as the primary repository source because ATCC is a widely used source of cell lines and their database is dynamically maintained and curated. Note that the construction of this CLKB aims to rationalize the nomenclature so that it represents the actual cell lines. At least one of the identifier attributes of the primary repository source (ATCC number, or HyperCLDB html) must be present for each cell line entry. Some cell line names that exist in multiple sources contain the identifier(s) for each source as cross-references. ATCC number is the ATCC catalog number. HyperCLDB identifiers are the original html file names as taken from the base URL (http://www.biotech.ist.unige.it/cldb/*.html). A few cell lines are also listed in NCBI MeSH collection. These cell lines also have the corresponding value in their MeSH identifier attributes.

3 RESULTS

The structure of this CLKB was constructed using Protégé (Noy *et al*., 2000, 2003). After manual review, we were left with 8740 unique-name entries. A Java script was implemented to automate the instance creation in a W3C Web Ontology Language format (OWL-DL file extension).

3.1 Cell line knowledgebase

The primary aim of building the CLKB is to facilitate research using cell cultures. Therefore, in addition to the simple underlying ontology structure, CLKB contains information on cell line culture, availability and biological characteristics. We have constructed an online web interface that queries the CLKB. The URL for this knowledgebase is http://clkb.ncibi.org/. This CLKB is a public data warehouse for searching cell line data extracted from ATCC and HyperCLDB as described previously. Constructing this knowledgebase also leads us to the next application of this CLKB, ontology mapping. Often, information can be linked from other sources. For example, the mapping by string literals of attribute 'Tissue' in cell line to the Cell Type Ontology (Bard *et al*., 2005). There are 8473 cell lines (out of the total 8740 instances in this knowledgebase), each of which has the attribute 'Tissue' that can be mapped to cell-type name in Cell Type Ontology using exact-string alignment method. There are 265 cell line instances that do not have tissue information (Tissue = NULL). This leaves us with only two cell line instances where their tissue attribute cannot be automatically mapped to a cell-type name (SVEC4-10EE2 and SVEC4-10EHR1). Even these two remaining cases could be mapped when extraneous blanks were removed from their tissue attribute values. This preliminary CLKB—Cell Type Ontology mapping

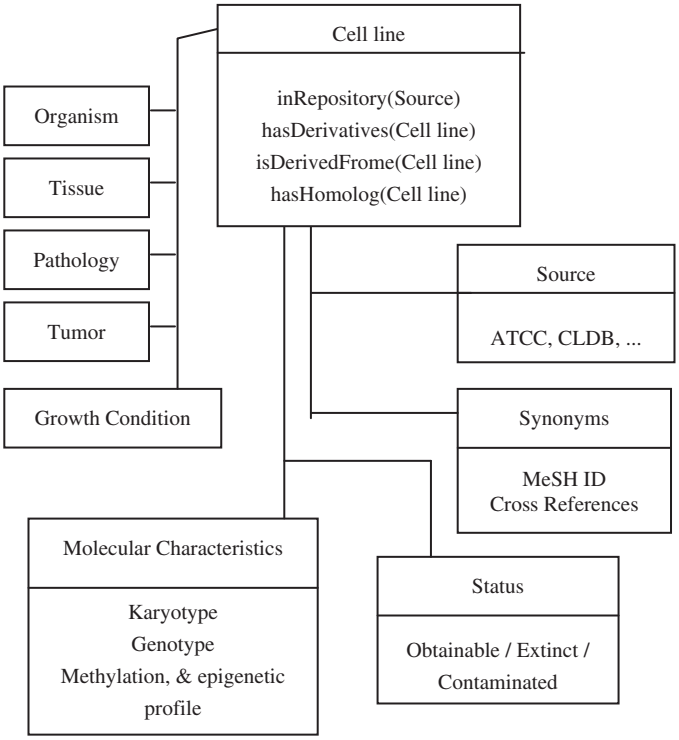

Fig. 1. Diagram describing the entities and relationships of the CLKB.

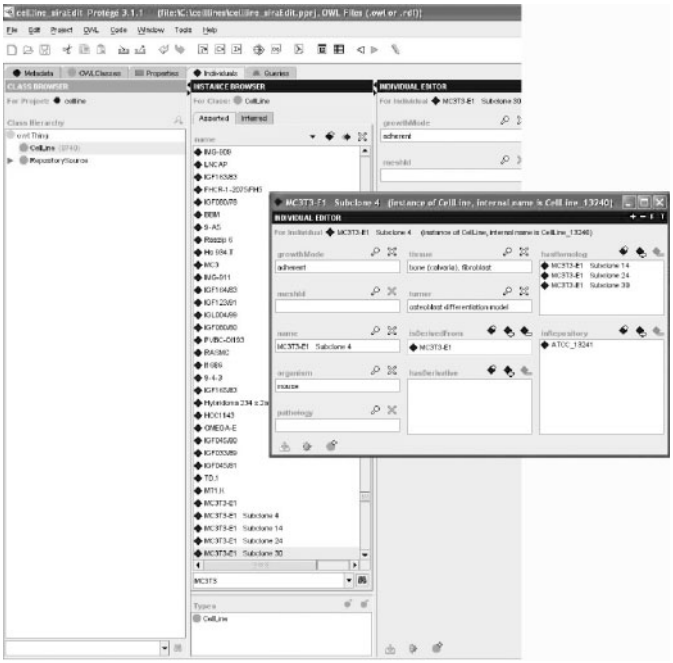

Fig. 2. Shown in this figure is a screenshot of the Protégé ontology editor.

clearly demonstrates a promising solution and is a step towards building the 'biomedical knowledge network'.

3.2 Data structure

In the CLKB, there are two distinct entity types, 'CellLine' and 'RepositorySource'. A knowledgebase contains both classes and instances. A cell line is defined as a class. An instance of a cell line is a culture of that class of cells maintained by a particular vendor or laboratory. A cell line entity contains the attributes as described. A repository entry contains the information of Repository association class, and the URL in which the cell line refers to (html indicator for HyperCLDB instances, and ATCC catalog PDF for ATCC instances). The CLKB was developed with these processed data as outlined in Figure 1. The majority of cell lines are stand-alone entities, and we include a link to their source of origin.

A number of cell lines are derivatives of some common cell lines. These cell lines contain the 'isDerivedFrom' relationship to the parent cell line, and the parent cell line could have one or more 'hasDerivatives' relationships as well. Derivatives at the same level are considered 'homologs' (for example, multiple clones derived from a common parent cell line, thus there must first be a common parent in the dataset). Determination of derivatives was based on explicit information in the cell line name field (e.g. the word 'subclone'). Although there may have been indicators other than this explicit statement for cell line derivatives, we intentionally did not include these rules in our automated script as the common substring approach could be misleading. We do not want to identify a cell line as a derivative for another cell line when it was not a true derivative (e.g. a substring in a cell line name may be just a manufacturer's abbreviation and a system-generated number). To avoid this foreseen issue, we decided to leave out other clonal implications that are not explicitly stated in the cell line names.

The CLKB can be viewed in most ontology editors that are capable of processing W3C Web Ontology Language file format (.OWL extension). A screenshot from Protégé ontology editor of this knowledgebase is also shown in Figure 2. There are likely going to be more novel cell line information available from various sources, we encourage the use of an ontology editor in merging these cell lines into our current version (for example, Protégé as we have successfully worked with in this project—http://protege.stanford.edu). The ontology mapping should be relatively easy to accomplish as our CLKB structure was created in such a way that supports an effective and easy-to-understand downstream activities.

4 DISCUSSION

We find many ways in which the cell line nomenclature could be made more useful to the community. These include formalizing non-standardized formatting of textual data, the use of special characters or too common English words as cell line name in research. A serious issue is inconsistent name of a cell line caused by cross-species contamination or propagation method.

4.1 Cell line ontology

As a first step in developing the knowledgebase, we defined this cell line ontology. An ontology described and defined in one context may or may not be a consensus to all users. Depending on the perspective of the user, there may be multiple 'correct' ontologies that capture the ontology's user needs. Because the primary purpose for this type of ontology is to rationalize the cell line nomenclature in experimental and clinical research, we have chosen to simplify the structure of the

Table 2. Top 10 HyperCLDB cell line names that appear in non-biomedical corpus ranked by number of occurrences

Cell line name	Count
M4	48
35	44
Aa	11
EC	11
P1	11
380	9
BT	8
L	7
CAR	4
OK	4

cell line ontology as described in the previous sections. This core cell line ontology structure can be extended in multiple ways.

4.2 Cell line nomenclature

The compilation of this knowledgebase reveals a number of cell line names that can potentially be problematic (Tables 1 and 2 and Supplementary Material). First, there exist cell lines with identical cell line names, which are, however, distinct cell lines that should be listed separately. For example, the 15C6 cell line is used in more than one distinct ATCC catalog numbers, CRL-2431 and HB-326, one referring to a mouse hybridoma line and the other to a rat/mouse line (Supplementary Table B for the complete list). Second, we have observed inconsistency in the use of cell line annotations between different laboratories when referring to the same cell line. For example, cell lines LLC-MK2, LLCMK2, $LLCMK_2$, all appear to be the cell lines with identical attribute values. Other similar cases are of NIH/3T3 and NIH-3T3 cells, or BxPC3, BXPC-3, BxPC-3.

Another source of inconsistencies is formatting of text. For example, capitalization, dashes, slashes and other punctuation marks are frequently used inconsistently between different public repositories or even within one repository if the cell line is obtained from different depositing laboratories. An NLP approach may be a suitable method to identify these occurrences. Table 1 gives some example of one cell line being described with different names. As it appears, many cell line records obtained from HyperCLDB contain different capitalization and punctuation even though they refer back to the same cell line in ATCC. The uncontrolled use of text formatting adds synonyms to the nomenclature and leads to more confusion. Symbols and Greek letters are another source of inconsistency. For example, Ψ may be shown as a Greek character in one instance and spelled out as 'psi' in another. Unicode presents another issue. When annotating a cell line with Unicode characters (e.g. the use of α, γ, δ or ψ in some cases) different information systems may handle the data differently, in some cases even ignoring such characters altogether! Further, there are visually similar Unicode characters with distinct Unicode encodings. 'Ψ' and 'ψ' are both renderings of Psi in upper and lower case, but with distinct Unicode encoding. Although they may appear similar to the human eye, an automated script will fail to match them. Fourth, some cell line names cannot be recognized or distinguished easily when appearing in a different context or even in the same biomedical domain. For example, there exist cell line names that are too common to be easily tagged.

Table 3. HyperCLDB cell line names that appear in NCBI GEO Microarray sample description, ranked by number of occurrences

Cell line names	Count	False positives
M9	122	All
F2	120	All
Bm	72	All
IMR-90	41	
MS	38	All
35	35	All
A549	31	
C2C12	30	
FRT	29	All
E2	28	All
NMU	24	All
CCRF-CEM	19	
HCT-8	18	
MDA-MB-231	17	
SC	16	All
HL-60	14	
P4	14	All
Aa	12	All
NIH 3T3	11	
N1E-115	9	
A673	9	
E3	8	All
MCF-7	8	
F1	8	All
D2	8	All
MCF7	8	
697	7	

The later update of local microarray database may result in different numbers of occurrences, however is insignificant to the false positives indicator.

Examples include HORSE, OK, WISH, 35 and 81.3. We have identified cell line names that are also common English words by constructing a dictionary of cell line names from HyperCLDB collection and searching against a moderate-size corpus of Wall Street Journal text that is unlikely to include references to cell lines. Table 2 shows the top 10 HyperCLDB cell line names that occur in a non-biomedical context. Furthermore, we have also investigated the use of cell line names in NCBI GEO microarray description fields. We parsed out 26 109 microarray sample descriptions and tagged the named entities of cell line names based on the HyperCLDB cell line name dictionary that we constructed in the Wall Street Journal corpus experiment as described previously. Some of the occurrences of cell line '35' were false positive as one may have already expected. 'Bm' may as well be an acronym for bone marrow, and not a cell line name at all. Table 3 demonstrates the list of tagged cell line names in microarray sample descriptions.

Note that, only a few of the known cell line names appear in the sample description, and some are more frequently used than the others. This may be due to the fact that many GEO samples contain null description field, or very short phrases. Further investigation reveals that *tissueOrCelllineName* field in GEO sample attributes has been left blank (NULL) in the majority of the data deposited in GEO database. Some GEO samples have information of tissue of origin, and only a small part contain the cell line information.

Table 4. Ten most common cell line names in GEO microarray sample description demonstrating tradeoff of recall and specificity

Cell line name	# with rules	# without rules
L	0	5689
G	0	4213
FR	0	2968
ST	0	2821
U	0	2459
FO	0	2312
NE	0	2133
TE	0	1914
ME	0	1768
MO	0	1243

Table 5. Ten most common cell line names in GEO microarray series description demonstrating tradeoff of recall and specificity

Cell line name	# with rules	# without rules
L	0	855
G	0	763
NE	0	504
ST	0	394
U	0	381
FR	0	352
FO	0	331
ME	0	323
TE	0	247
MO	0	226

Furthermore, very few values in *sampleNameInExperiment* from the GEO sample attributes are useful pointers to cell line information of the sample used in that experiment. However, some GEO series descriptions contain information that a rule-based NLP implementation may be able to extract useful cell line information from them. A script implementing NLP was written to demonstrate that a simple rule-based NLP approach could help eliminate some common false hits and gain some information of cell line name in text scanning. We tagged cell line names that appeared in 8091 GEO sample descriptions, 1059 GEO series descriptions and 5 187 422 PubMed sentences containing the word 'cell' or 'cells' [University of Illinois at Urbana Champaign sentence splitter (UIUC, 2004) was used to create this PubMed sentence table].

Comparisons of the common names of cell line names in GEO sample descriptions, series description and PubMed sentences are given in Tables 4–6. The complete tables are given in Supplementary Materials. The rule-based procedure tagged only cell line names (taken from the existing cell line name dictionary) that preceded 'cell', or 'cells' tokens, or followed 'cell line' token. Also, we used the spaced-tagging strategy ('_% cells|cell_') to avoid false hits in tagging short cell line names (two characters long or shorter). An obvious example in this case is a single-letter cell line name like 'L cells'; a non-spaced tagging will result in false positives in phrases like '… small cells', or '… epithelial cells'. The non-spaced tagging ('% cells|cell') was used for cell line names that are longer than 2 characters as cell line names can appear at the beginning of a sentence. Another important point regarding rule-based tagging concerns the use of textual qualifiers. There may be other qualifiers for such tagging that researchers use in free text. Further, authors sometimes drop the qualifiers and use only the cell line name token in other sections of a document. However, we have discovered from this experiment that, in the case of wildly used cell lines, authors seem to conform well to our rules/patterns of 'cell' or 'cells' token. HeLa cell line was used as an example. Our scripts recalled 22 431 documents that contained 'HeLa' (and its spelling variants) tokens. In these documents, there are 43 255 sentences containing the word 'HeLa' and its variants. Out of these 43 255 sentences, 34 255 sentences are the sentences where the cell line tokens were tagged in the context of '% hela cell%'. Therefore, 79.19% of sentences tagged with the cell line token were correctly identified as true positives when using rule-based method. Furthermore, when we looked at documents that contained multiple sentences mentioning HeLa and compared with the documents in which only a single sentence was found tagged with HeLa, there are smaller number of documents in the multiple-sentence set (8558 documents) than the single-sentence set (10 026 documents). Among documents with multiple sentences that co-occur with 'hela', looking for an instance that matches the rule '% hela cell%' in one sentence and then assuming that all instances of 'hela' in that document were references to an actual cell line that found additional 3196 sentences in 1215 documents (these 1215 documents are not a subset of the 8558 multiple-sentence documents). It should also be noted that while we gain better precision with this rule-based strategy, there is also a tradeoff in losing the overall recall. Even though this rule-based NLP approach remains effective in achieving high recall at larger scale text scanning, further study of finding an optimal adjustment may still be required for a smaller dataset.

Table 6. Twenty most common cell line names in PubMed sentences containing 'cell' or 'cells' tokens demonstrating tradeoff of recalls when increasing specificity

Cell line name	# with rules	# without rules
L	10401	5187367
U	177	4865827
G	1492	4393413
TE	600	4020437
ST	354	3056536
NE	1352	2815441
EC	3313	2539329
LI	29	2537483
ME	205	2208233
MO	102	1893124
FO	49	1482950
LAT	10	1244902
LT	37	1082924
FR	41	920858
RS	968	895430
YT	491	802212
SC	376	755576
FER	2	751252
TUR	103	652539
HEL	2263	479636

With a large dataset, increasing specificity leads to optimal recalls.

Table 7. Examples of cross-contaminated cell lines

Cell line (Cell type)	Described as	Reference
HeLa (cervical adenocarcinoma)	2563, MAC-21 (lung lymphoma)	Nelson-Rees *et al*. (1981)
	ADLC-5M2 (lung carsinoma)	MacLeod *et al*. (1999)
	AO (amnion)	Nelson-Rees and Flandermeyer (1976)
	BCC1/KMC (basal cell carcinoma)	MacLeod *et al*. (1999)
	BrCa 5 (breast carcinoma)	Nelson-rees *et al*. (1981)
	CaOV (ovarian carcinoma)	Nelson-Rees and Flandermeyer (1976)
	Chang liver (liver)	Nelson-Rees and Flandermeyer (1976)
	Wong-Kilbourne (conjuntiva)	Nelson-Rees and Flandermeyer (1976)
T-24 (bladder carcinoma)	ECV-304 (normal endothelium)	Dirks *et al*. (1999)
	GHV (astrocytoma)	MacLeod *et al*. (1999)
	HAG (adenomatoid goiter)	MacLeod *et al*. (1999)
	RAMAK-1 (muscle synovium)	MacLeod *et al*. (1999)
TE-2, TE-3, TE-7, TE-12, and TE-13 (esophageal squamous cell carcinoma)	TE-7 (esophageal adenocarcinoma)	Boonstra *et al*. (2007)
SK-NEP-1 (Ewing sarcoma)	SK-NEP-1 (Wilms tumor)	Smith *et al*. (2008)
EW8(Rh1) (Ewing sarcoma)	Rh1 (rhabdomyosarcoma)	Smith *et al*. (2008)

4.3 Discovery of uncataloged cell line names and synonyms

As there are often newly created cell lines in research laboratories, one may wish to utilize a set of existing cell line names for machine learning of biomedical terms using a similar approach to the construction of GENIA ontology (Liu *et al*., 2006) to recognize novel cell line names that have not been submitted to a repository. This leads to a question that whether or not one can derive a contextual format in which a cell line name may occur. A BayesNet classifier model using Weka (Frank *et al*., 2004) was introduced to this project in attempt to identify and discover the potentials token of novel cell line names in literature (here, we conducted our experiment with PubMed and NCBI GEO sample description sentences). One generalized observation is that tokens 'cell' or 'cell line' are often found in co-occurrence with an existing cell line name. The classifier could successfully identify the named entities to be of either cell-type or cell line classes based on the information from MeSH identifiers under A11 Tissue sub-tree, and the known cell line names in our dictionary. BayesNet classification turned out to be a powerful method to build this classifier. However, over-generalized names and the other natural language processing issues (as described previously) introduce a real obstacle for such discovery because they generate a large number of false matches in text scanning.

We also took another approach to distinguish cell line names from other named entities in an attempt to identify novel cell line names. As standard nomenclature, we know that capitalization signifies differentiation between DNA name and protein name. We have also observed the cell line names in parentheses, as many cell line names are acronyms. For example, papers describing Chinese Hamster Ovary cell culture often have token 'CHO' in parentheses. Focusing on short tokens in parentheses, we can narrow down the search space of potential cell line named entities. Further, we could assume that a token containing only digits and uppercase letters may potentially be a novel cell line name. '293T' token comes up when using this rule-based NLP named entity tagging [there is not a cell line named '293T' in either ATCC catalog (the closest name is 293T/17), or in HyperCLDB listing (the closest name is 293)]. Out of 46 976 tokens found in the context of '% cell%' ranked by number of occurrences, '293T' comes up at 47th place with 2809 counts, 2797 occurrences with '293T' spelling and 12 occurrences with '293t' spelling. When checking with tokens that appear in parentheses, out of 51 216 tagged tokens ranked by frequency of occurrences, '293T' comes up at 174th place with 2811 counts, 2799 with '293T' spelling and 12 with '293t' spelling. As confirmed by experimental biologists, 293T is a bona fide cell line. Our ability to recognize this as a cell line name based on lexical context suggests a promising direction for further investigation of novel cell line name discovery in free text.

4.4 Cross-contaminated cell lines

Cross-contamination is a very serious source of confusion in cell line nomenclature. Cell lines may be contaminated at many steps during propagation and the maintainer of that tissue culture may not be aware of such contamination. Misattribution of cross-contaminated cell lines causes extensive confusion about what a cell line truly is (examples given in Table 7). The widespread contamination with HeLa cells was first recognized by Walter Nelson-Rees using banded marker chromosomes as indicators of intra-species cellular contamination (Nelson-Rees *et al*., 1974). Cross-contamination of cell cultures has been an ongoing issue since (Nelson-Rees, 2001). Despite 30 years of effort, cross-contamination remains a problem. A study of contamination in leukemia–lymphoma cell lines has also shown that ∼15% of these cell lines are not the true representation of what they actually are, or are assumed to be (Drexler *et al*., 2002b, 2003) .

Moreover, even if a cell line is not contaminated, it may not be stable and its character may change from passage to passage. This may cause the birth of a 'new' cell line, sometimes without the knowledge of the users. DNA profiling and cytogenetic analysis are robust methods to identify different cell lines (Drexler *et al*., 2002a; MacLeod *et al*., 1997). These assays are becoming less expensive, and we hope will become a standard practice.

5 CONCLUSIONS

A recent study of dictionary-based named entity tagging of protein name (Liu *et al*., 2006) reveals that, despite the standardized HUGO Gene Nomenclature Committee (HGNC), issues of ambiguity remain in biomedical applications of automated NLP. Our study shows that there are many other categories of information in the biomedical domain where there is also a need to eliminate ambiguity wherever possible. This includes not only that the agreement on

a standard nomenclature with unique and distinctive names would greatly facilitate text interpretation, but also the elimination of cross-contamination in cell lines. As a first step toward standardizing the nomenclature for cell lines, we present the results of this research in Supplementary Material—cell line token analysis. We also propose that a standard protocol of the minimal set of information including molecular characteristics of cell line should be complied at the development and at repository level as well.

The mapping of cell line names to word tokens in GEO sample descriptions demonstrates that this ontology can be very useful in data quality assurance and control. One may wish to cross check, for example, whether or not the organism labeled for each GEO sample really matches with its corresponding organism in the cell line of that sample that was grown in. With the issues described here and how much automated NLP applications can potentially accomplish, one cannot stress enough the importance of standardization of biomedical terms. Furthermore, it is to our surprise to find that ever so often people do not cite where they obtain cell cultures in their research. Without an explicit explanation from the author, it is not practical to assume the source of cell cultures. The current lack of standardized nomenclature in many areas will be an obstacle to the development of effective natural language processing for the interpretation and analysis of free-text biomedical information.

The issue of cross-contaminated cell lines remains serious and has been stressed with the NIH notice regarding authentication of cultured cell lines (NIH, 2007) and an open letter to the US Department of Health and Human Services regarding the issue of cell line cross-contamination (Nardone *et al*., 2007).

ACKNOWLEDGEMENTS

The authors thank Mark Musen and Natasha Noy at Stanford BMI for Protégé support, and Chachrist Srisuwanrat for VBA technical support.

Funding: National Institutes of Health (grant U54 DA021519 for the National Center for Integrative Biomedical Informatics and R01 LM008106).

Conflict of Interest: none declared.

REFERENCES

Bard,J. *et al*. (2005) An ontology for cell types. *Genome Biol.*, **6**, R21.

Boonstra,J.J. *et al*. (2007) Mistaken identity of widely used esophageal adenocarcinoma cell line TE-7. *Cancer Res.*, **67**, 7996–8001.

Dirks,W.G. *et al*. (1999) ECV304 (endothelial) is really T24 (bladder carcinoma): cell line cross contamination at source. *In Vitro Cell. Dev. Biol.*, **35**, 558–559.

Drexler,H.G. *et al*. (2002a) DNA profiling and cytogenetic analysis of cell line WSU-CLL reveal cross-contamination with cell line REH (pre B-ALL). *Leukemia*, **16**, 1868–1870.

Drexler,H.G. *et al*. (2002b) Mix-ups and mycoplasm: the enemies within. *Leukemia Res.*, **26**, 329–333.

Drexler,H.G. *et al*. (2003) False leukemia-lymphoma cell lines: an update on over 500 cell lines.*Leukemia*, **17**, 416–426.

Frank,E. *et al*. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.

Lee,K.J. *et al*. (2004) Biomedical named entity recognition using two-phase model based on SVMs. *J. Biomed. Inform.*, **37**, 436–447.

Liu,H. *et al*. (2006) Quantitative assessment of dictionary-based protein named entity tagging. *J. Am. Med. Inform. Assoc.*, **13**, 497–507.

MacLeod,R.A.F. *et al*. (1997) Identity of original and late passage Dami megakaryocytes with HEL erythroleukemia cells shown by combined cytogenetics and DNA fingerprinting. *Leukemia*, **11**, 2032–2038.

MacLeod,R.A.F. *et al*. (1999) Widespread intraspecies cross-contamination of human tumor cell lines arising at source. *Int. J. Cancer*, **83**, 555–563.

Manniello,A. and Ruzzon,T. (1996) Cell Line Data Base and HyperCLDB. *Biotech. Knowl. Source.*, **9**, 3.

Nardone,R.M. (2007) An open letter regarding the misidentification and cross-contamination of cell lines: significance and recommendations for correction. Addressed to U.S. Department of Health and Human Services, July 11th, 2007. http://www.hpacultures.org.uk/media/A0D/E3/Open_Letter_Final_7-11-07.pdf

Nelson-Rees,W.A. (2001) Responsibility for truth in research. *Phil. Trans. R. Soc. Lond.* B, **356**, 849–851.

Nelson-Rees,W.A. and Flandermeyer,R.R. (1976) HeLa cultures defined. *Science*, **191**, 96–98.

Nelson-Rees,W.A. *et al*. (1974) Banded marker chromosomes as indicators of intraspecies cellular contamination. *Science*, **184**, 1093–1096.

Nelson-Rees,W.A. *et al*. (1981) Cross-contamination of cells in culture.*Science*, **212**, 446–452.

NLM (2007) Medical subject headings. Available at http://www.nlm.nih.gov/mesh/meshhome.html.

NIH (2007) Notice Number: NOT-OD-08-017 Notice Regarding Authentication of Cultured Cell Lines, November 28, 2007

Noy,N.F. *et al*. (2000) Creating semantic web contents with Protégé-2000 Intelligent Systems. *IEEE Intelligent Systems*, **16**, 60–71.

Noy,N.F. *et al*. (2003) Protégé-2000: an open source ontology-development and knowledge-acquisition environment. *AMIA Ann. Symp. Proc. 2003*, **2003**, 953.

Parodi,B. *et al*. (1993) *Human and Animal Cell Lines Catalogue*. Editrice abc – Officine Grafiche, Genova.

Rinaldi,F. *et al*. (2006) An environment for relation mining over richly annotated corpora: the case of GENIA. *BMC Bioinformatics*, **7** (Suppl. 3), S3.

Romano,P. *et al*. (1993) Interlab Project Databases: an effort towards the needs of a wider body of unskilled users. *Binary*, **5**, 66–72.

Shulz,S. *et al*. (2006) Towards an upper level ontology for molecular biology. *AMIA Ann. Symp. Proc.*, **2006**, 694–698.

Smith,M.A. *et al*. (2008) SK-NEP-1 and Rh1 are ewing family tumor lines. *Pediatr. Blood Cancer*, **50**, 703–706.

University of Illinois at Urbana Champaign (2004) Sentence segmentation tool. Available at http://l2r.cs.uiuc.edu/~cogcomp/atool.php?tkey=SS. (last accessed date April 13, 2008).