BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Database

Volume 8(22)

# CellLineMiner: a knowledge portal for human cell lines

**Sigve Nakken[1]*, Morten Johansen[1], Julien Fillebeen[2], Ole Petter Berge[2], Harald Kirkerød[2], Tor-Kristian Jenssen[2] & Eivind Hovig[1,3,4]**

[1]Department of Tumor Biology, Institute for Cancer Research, Norwegian Radium Hospital - Oslo University Hospital, Norway; [2]PubGene AS, Oslo, Norway; [3]Department of Informatics, University of Oslo, Norway; [4]Institute for Medical Informatics, Oslo University Hospital, Norway; Sigve Nakken – Email: sigven@ifi.uio.no; Phone: +47 95 75 30 22; *Corresponding author

Received October 22, 2012; Accepted October 26, 2012; Published November 13, 2012

**Abstract:**
Experimental models of human tissues and disease phenotypes frequently rely upon immortalized cell lines, which are easily accessible and simple to use due to their infinite capability of cell division. For decades, cell lines have been used to investigate cellular mechanisms of disease and the efficacy of drugs, most prominently for human cancers. However, the large body of knowledge with respect to human cell lines exists primarily in an unstructured fashion, that is, as free text in the scientific literature. Here we present *CellLineMiner*, a novel text mining-based web database that provides a comprehensive view of human cell line knowledge. The application offers a simple search in all indexed cell lines, accompanied by a rapid display of all identified literature associations. The *CellLineMiner* is intended to serve as a knowledge resource companion to the cellular model systems used in biomedical research.

**Availability:** *CellLineMiner* is accessible at http: / / dev.pubgene.com / cellmine.

**Background:**
Immortalized cell lines represent powerful model systems in experimental biomedical research – they have an indefinite growth capability, they can be frozen for decades, and shared among scientific laboratories world-wide. Although the functional resemblance between cell lines and their *in vivo* counterparts in the intact tissue is subject to much variation, cell lines still constitute a fundamental tool in drug development studies and in unraveling the cellular mechanisms of many human diseases **[1, 2].**

Most cell lines are available through bioresource providers or cell line banks, such as the American Type Culture Collection (ATCC) and the German Collection of Microorganisms and Cell Cultures (DSMZ). Other important cell line resources include the Cancer Cell Line Project, which aims to systematically characterize simple mutations and large genomic alterations in ~800 cancer cell lines, and the Cell Line Data Base, which is a cross-reference resource for all available cell lines in any given species **[3, 4].**

The scientific literature contains extensive reports about functional relationships between specific cell lines and other biomedical properties, such as drugs, diseases and biological processes. Cell line models have been particularly important for cancer research, highlighted by a recent study that used a panel of >130 cell lines to unravel mechanisms underlying drug sensitivity **[5].** There is however no simple means by which one can query cell line names and retrieve the most relevant biomedical concepts in a structured fashion. In an effort to address this matter, we have used text mining techniques to develop CellLineMiner, a web database that is intended to serve as a knowledge resource companion to the cellular model systems used in biomedical research.

**Methodology:**

***Creation of biomedical dictionaries***

A human cell line dictionary was created by integration of known cell line designations from three primary sources: the American Type Culture Collection (ATCC **[6]**), the German Collection of Microorganisms and Cell Cultures (DSMZ **[7]**),

 

and the Cancer Cell Line Project **[8]**. We manually identified, through inspection of the indexing results (see below) cell line designations that coincided with gene symbols or other general abbreviations. Cell line contexts were added to these designations in order to retrieve true positive hits. For example, the cell line *ABC-1* was found to coincide with an alias for the human ABCA1 gene (ATP-binding cassette, sub-family A), so this designation was replaced with the more specific terms "*cell line ABC-1*" and "*ABC-1 cells*". The final dictionary contained nevertheless designations for a total of 1,622 different human cell lines.

Dictionaries of human diseases, symptoms, anatomies, procedures / treatments, genes / proteins, drugs, chemicals, gene ontology terms, and medical subject headings (MeSH) were indexed in free text in the same manner as human cell line names (more below). See ***Supplementary Material*** for more detailed information about how these dictionaries were created.

***Creation of literature indices***

Abstracts and titles of all records in the MEDLINE database (baseline November 2011; 20,494,848 citations) were used as the source for the creation of a literature index of human cell lines and related concepts. The indexing pipeline included text tokenization, part-of-speech tagging, chunk extraction, and finally mapping of cell line names to the resulting chunks. The implementation of the pipeline was based upon work from a previous study **[9].**

Associations between cell lines and other biomedical concepts were established based upon the citation co-occurrence principle. This is a well-known approach within text mining **[10].** The observed co-occurrence frequency between two terms was compared with the expected random co-occurrence frequency under a binomial distribution. A chi-square goodness-of-fit test was used to rank the term relationships in a statistical manner.

Our NLP pipeline produced MEDLINE hits for approximately 58.8% of the cell lines in our manually curated dictionary (954 out of 1622). These hits corresponded to 182,290 unique MEDLINE records. **Table 1 (see supplementary material)** shows the total number of statistically significant co-occurrence relationships between cell lines and entities in other concept categories.

**Utility to the biological community:**

The web interface to the *CellLineMiner* web database consists of a search suggestion box where a user can type in cell line names and choose among the ones that exist in the *CellLineMiner* dictionary. Next a ranked list of literature associations are displayed for each of the other concepts that have been mined, also allowing the user to further explore the actual MEDLINE records where cell lines and concepts are co-occurring **(Figure 1)**. If available, we also display links to cell line mutation data, as provided by the Cancer Cell Line Project **[3].**

**Caveats & future developments:**

Our attempt to disambiguate cell line names was done in a manual fashion. One may therefore suspect that there could still be some designations in our cell line dictionary that are ambiguous. Future developments will thus include a benchmark of our performance with respect to cell line identification in biomedical text. Moreover, we will update *CellLineMiner* according to MEDLINE updates, and add more cross-references to cell line mutation data as these are identified and made available.

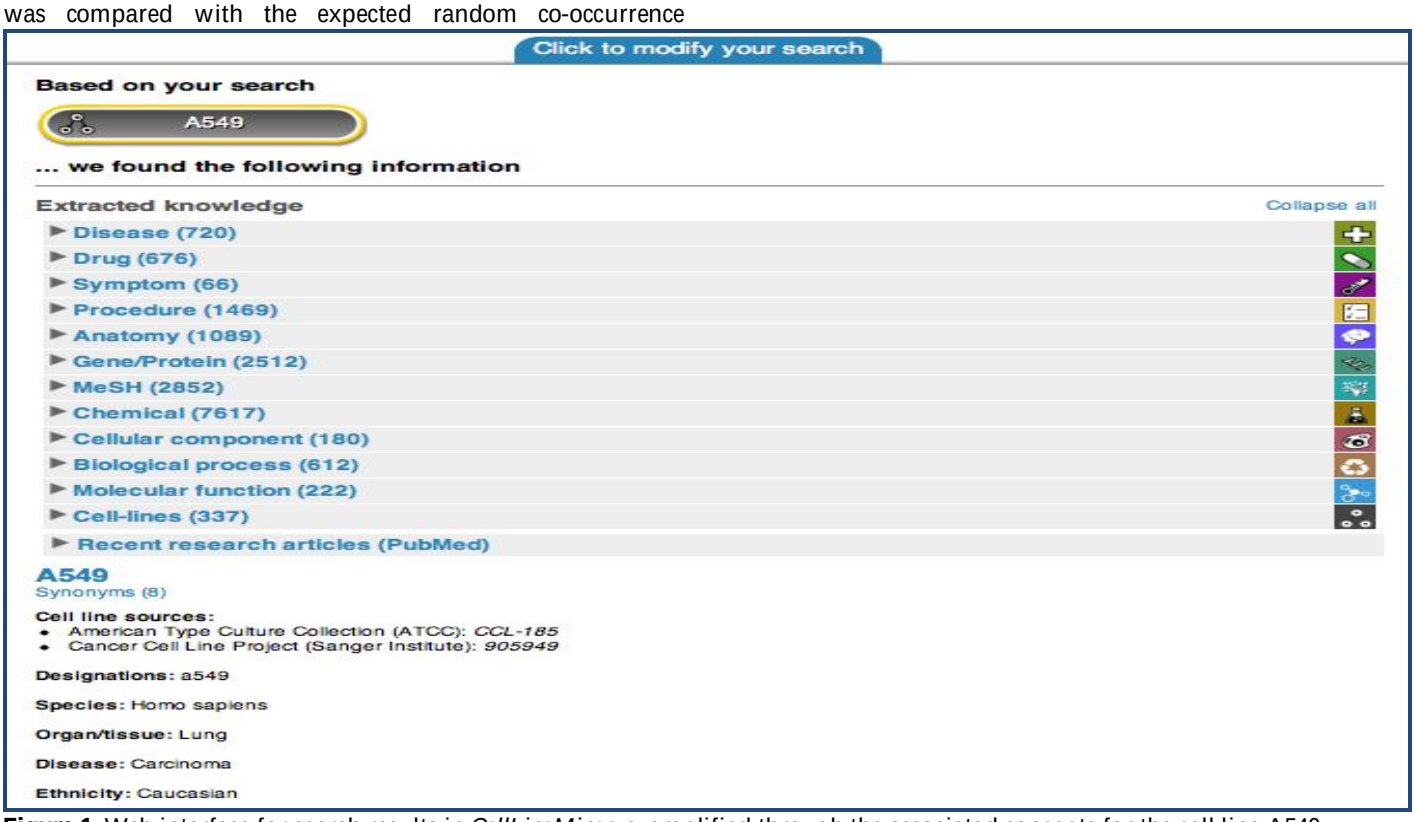


**Figure 1:** Web interface for search results in *CellLineMiner*, exemplified through the associated concepts for the cell line A549.

**Acknowledgement:**
This research received funding from the European Commission (FP7-2008, No 223367-MultiMod).

**References:**

**[1]** Thomson JA *et al. Science.* 1998 **282**: 1145 [PMID: 9804556]
**[2]** Sharma SV *et al. Nat Rev Cancer.* 2010 **10**: 241 [PMID: 22460902]
**[3]** Forbes SA *et al. Nucleic Acids Res.* 2011 **39**: D945 [PMID: 20952405]
**[4]** Romano P *et al. Nucleic Acids Res.* 2009 **37**: D925 [PMID: 18927105]
**[5]** Garnett MJ *et al. Nature.* 2012 **483**: 570 [PMID: 22460902]
**[6]** http://www.atcc.org
**[7]** http://www.dsmz.de
**[8]** http://www.sanger.ac.uk/genetics/CGP/CellLines/
**[9]** Jenssen TK *et al. Nat Genet.* 2001 **28**: 21 [PMID: 11326270]
**[10]** Jensen LJ *et al. Nat Rev Genet.* 2006 **7**: 119 [PMID: 16418747]

**Edited by P Kangueane**
**Citation: Nakken *et al*.** Bioinformation 8(22): 1119-1122 (2012)

BIOINFORMATION open access

## Supplementary material:

**Implementation**

*Biomedical dictionaries:* Standardized terms from the Unified Medical Language System (UMLS) were used to create dictionaries for disease (4,518 entities), anatomy (101,570 entities), procedures / treatments (107,824 entities), and symptom concepts (2,589 entities). Our collection of chemical substances was based on PubChem and Registry Numbers (i.e. RN tags) found within MEDLINE records, providing a total of 1,006,412 different chemical entities. A human gene / protein dictionary was based on data from EntreZ Gene, UniProt and RefSeq and contained 42,129 entities (each with a primary symbol and primary name, and potentially alternative symbols and names). We created one dictionary for each of the three structured vocabularies that are part of the Gene Ontology Project; biological process (21,551 entities), cellular components (2,918 entities), and molecular function (9,149 entities). A total of 10,833 different entities formed the basis for medical subject headings (MeSH). With respect to the MeSH-Medline index, we merged the results from our free-text based index (i.e. NLP) with the manually annotated MeSH tags found within MEDLINE records.

**Table 1**: Number of statistically significant ($p<0.05$) literature relationships between cell lines and other biomedical concepts

| Concept | #Literature relationships with cell lines |
|---|---|
| Gene / protein | 67,183 |
| Disease | 30,086 |
| Cell line | 7,926 |
| Anatomy | 30,757 |
| GO – molecular function | 5,153 |
| GO – cellular component | 6,536 |
| GO – biological process | 21,547 |
| Chemical | 171,816 |
| MeSH | 58,173 |
| Procedure | 43,662 |
| Symptom | 506 |