Universiteit Leiden

Optimizing BERT-based reference mining from patents

Zahra Abbasiantaeb, Suzan Verberne, Jian Wang

Universiteit Leiden

Outline

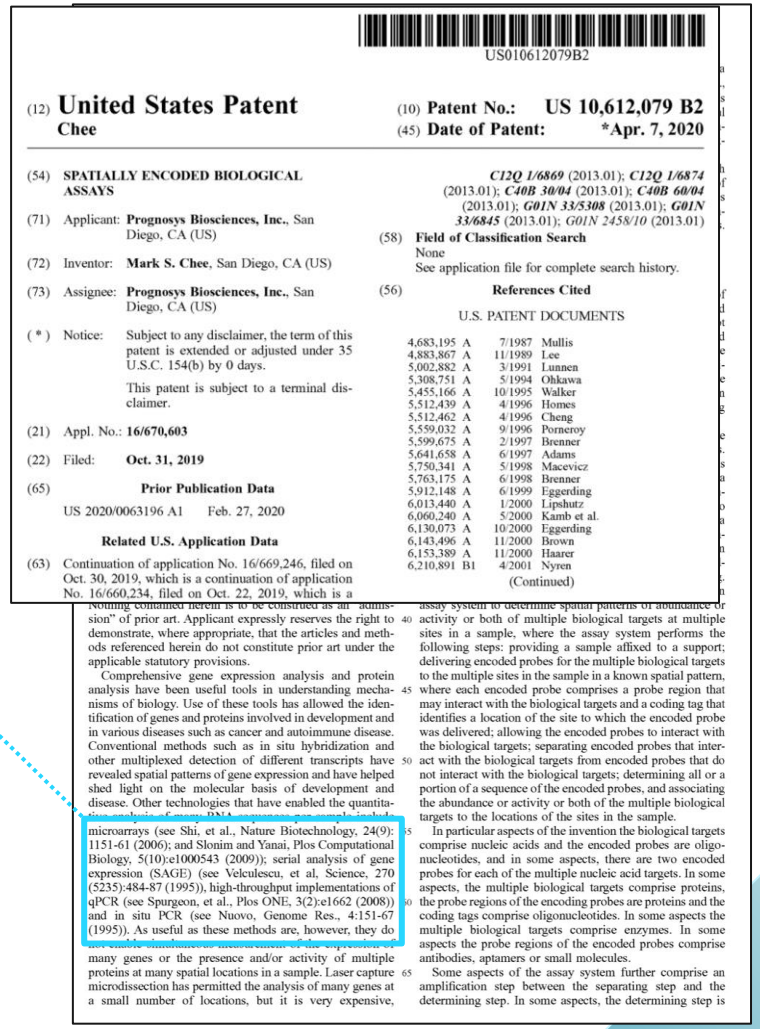
Universiteit Leiden

01

Introduction

Introduction

microarrays (see Shi, et al., Nature Biotechnology, 24(9): 1151-61 (2006); and Slonim and Yanai, Plos Computational Biology, 5(10):e1000543 (2009)); serial analysis of gene expression (SAGE) (see Velculescu, et al, Science, 270 (5235):484-87 (1995)), high-throughput implementations of qPCR (see Spurgeon, et al., Plos ONE, 3(2):e1662 (2008)) and in situ PCR (see Nuovo, Genome Res., 4:151-67 (1995)). As useful as these methods are, however, they do



US010612079B2

(12) **United States Patent**
Chee

(10) **Patent No.: US 10,612,079 B2**
(45) **Date of Patent: *Apr. 7, 2020**

(54) **SPATIALLY ENCODED BIOLOGICAL ASSAYS**

(71) Applicant: **Prognosys Biosciences, Inc.**, San Diego, CA (US)

(72) Inventor: **Mark S. Chee**, San Diego, CA (US)

(73) Assignee: **Prognosys Biosciences, Inc.**, San Diego, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **16/670,603**

(22) Filed: **Oct. 31, 2019**

(65) **Prior Publication Data**

US 2020/0063196 A1 Feb. 27, 2020

Related U.S. Application Data

(63) Continuation of application No. 16/669,246, filed on Oct. 30, 2019, which is a continuation of application No. 16/660,234, filed on Oct. 22, 2019, which is a

C12Q 1/6869 (2013.01); *C12Q 1/6874* (2013.01); *C40B 30/04* (2013.01); *C40B 60/04* (2013.01); *G01N 33/5308* (2013.01); *G01N 33/6845* (2013.01); *G01N 2458/10* (2013.01)

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,683,195	A	7/1987	Mullis
4,883,867	A	11/1989	Lee
5,002,882	A	3/1991	Lunnen
5,308,751	A	5/1994	Ohkawa
5,455,166	A	10/1995	Walker
5,512,439	A	4/1996	Homes
5,512,462	A	4/1996	Cheng
5,559,032	A	9/1996	Porneroy
5,599,675	A	2/1997	Brenner
5,641,658	A	6/1997	Adams
5,750,341	A	5/1998	Macevicz
5,763,175	A	6/1998	Brenner
5,912,148	A	6/1999	Eggerding
6,013,440	A	1/2000	Lipshutz
6,060,240	A	5/2000	Kamb et al.
6,130,073	A	10/2000	Eggerding
6,143,496	A	11/2000	Brown
6,153,389	A	11/2000	Haarer
6,210,891	B1	4/2001	Nyren

(Continued)

Nothing contained herein is to be construed as an admission" of prior art. Applicant expressly reserves the right to demonstrate, where appropriate, that the articles and methods referenced herein do not constitute prior art under the applicable statutory provisions.

Comprehensive gene expression analysis and protein analysis have been useful tools in understanding mechanisms of biology. Use of these tools has allowed the identification of genes and proteins involved in development and in various diseases such as cancer and autoimmune disease. Conventional methods such as in situ hybridization and other multiplexed detection of different transcripts have revealed spatial patterns of gene expression and have helped shed light on the molecular basis of development and disease. Other technologies that have enabled the quantita-

microarrays (see Shi, et al., Nature Biotechnology, 24(9): 1151-61 (2006); and Slonim and Yanai, Plos Computational Biology, 5(10):e1000543 (2009)); serial analysis of gene expression (SAGE) (see Velculescu, et al, Science, 270 (5235):484-87 (1995)), high-throughput implementations of qPCR (see Spurgeon, et al., Plos ONE, 3(2):e1662 (2008)) and in situ PCR (see Nuovo, Genome Res., 4:151-67 (1995)). As useful as these methods are, however, they do

many genes or the presence and/or activity of multiple proteins at many spatial locations in a sample. Laser capture microdissection has permitted the analysis of many genes at a small number of locations, but it is very expensive,

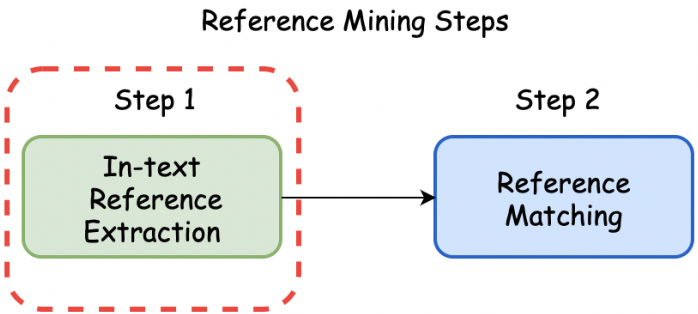
assay system to determine spatial patterns of abundance or activity or both of multiple biological targets at multiple sites in a sample, where the assay system performs the following steps: providing a sample affixed to a support; delivering encoded probes for the multiple biological targets to the multiple sites in the sample in a known spatial pattern, where each encoded probe comprises a probe region that may interact with the biological targets and a coding tag that identifies a location of the site to which the encoded probe was delivered; allowing the encoded probes to interact with the biological targets; separating encoded probes that interact with the biological targets from encoded probes that do not interact with the biological targets; determining all or a portion of a sequence of the encoded probes, and associating the abundance or activity or both of the multiple biological targets to the locations of the sites in the sample.

In particular aspects of the invention the biological targets comprise nucleic acids and the encoded probes are oligonucleotides, and in some aspects, there are two encoded probes for each of the multiple nucleic acid targets. In some aspects, the multiple biological targets comprise proteins, the probe regions of the encoding probes are proteins and the coding tags comprise oligonucleotides. In some aspects the multiple biological targets comprise enzymes. In some aspects the probe regions of the encoded probes comprise antibodies, aptamers or small molecules.

Some aspects of the assay system further comprise an amplification step between the separating step and the determining step. In some aspects, the determining step is

Introduction

- Patent mining: Identify science references inside the patents
- Impact of science on technological advances



- Our focus: <u>In-text Reference Extraction</u>

Problem Formulation

- No standard style of referencing
- Sequence Labeling Approach
  - BIO labels
  - Pre-trained BERT models

Tratschin et al . , Mol . Cell . Biol . 5 , 3251 - 3260 (1985)
B I

Thoraval et al . , 1995 , Transgenic Res . 4 : 369 - 377
B I

Schiest and Petes (Proc . Nat . Acad . Sci . U . S . A . 88 , 7585 - 7589 (1991))
B I

Gilbert , " Egg albumen and its formation " in Physiology and Biochemistry of the Domestic Fowl , Bell
B I
and Freeman , eds . , Academic Press , London , New York , pp . 1291 - 1329

Previous Works

- Verberne and Chios 2019[1]
  - **Method:** CRF and Flair models for reference extraction
  - **Dataset:** 22 patents from Google Patents

- Voskuli and Verberne 2021[2]
  - **Method:** BERT model for reference extraction
  - **Dataset:** Improved the quality of previous dataset

[1] Verberne, Suzan, Ioannis Chios and Jian Wang. "Extracting and Matching Patent In-text References to Scientific Publications." *BIRNDL@SIGIR* (2019).
[2] Suzan Verberne, Ioannis Chios, Jian Wang (2019). "Extracting and Matching Patent In-text References to Scientific Publications" . In the Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019), co-located with SIGIR 2019, pp. 35-42

In-text Reference Extraction

Improving the reference extraction component

1) Multiple pre-trained BERT models, including patent-specific models

- PatentBERT (The claims parts of the USPTO patents)
- Bert for Patents (Complete text of patents, BERT-Large)
- BioBERT
- SciBERT (Scientific)
- BERT

In-text Reference Extraction

2) A more effective method for sequence splitting

$$\{t_1, t_2, t_3, \dots, t_n\}: \max(n) \mid \{|t_1| + |t_2| + |t_3| + \dots + |t_n| <= 512\}$$

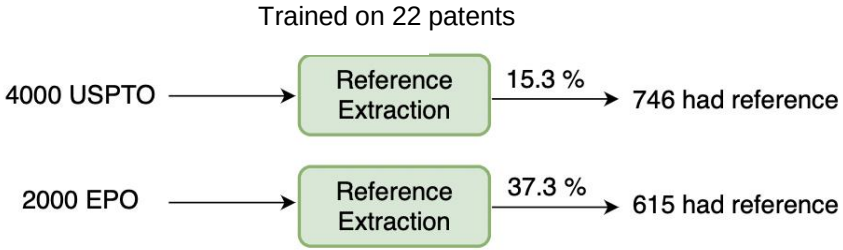
3) Down sampling to cope with the class imbalance

- 14,270 sequences where 8,530 of them have no ‘B’ or ‘I’ labels
- Imbalance between B/I labels and O label → A biased model
- Eliminate the sequences with no B/I labels

In-text Reference Extraction

4) Collecting a larger dataset from EPO and USPTO (On-going).

- Consider utility patents after 1990 for sampling
- A random sample of 4000 USPTO and 2000 EPO
- Hired 8 students for annotation
- Manually annotating at least 600 of USPTO and 600 of EPO patents



Universiteit Leiden

02

Experiments

Dataset and Evaluation

Data: 22 patents dataset collected by Chios and Verberne (2019) [1]

- Google Patents
- Domain of Biotech

Evaluation: Leave-One-Out Cross-Validation

[1] Suzan Verberne, Ioannis Chios, Jian Wang (2019). "Extracting and Matching Patent In-text References to Scientific Publications". In the Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019), co-located with SIGIR 2019, pp. 35-42

Effect of Down Sampling

- Model trained more smoothly
- Sequences with no reference are less informative for the model

Down Sampling	Label	Precision	Recall
True	B	0.884	0.922
	I	0.967	0.966
False	B	0.880	0.919
	I	0.959	0.964

Comparing BERT Models

- PatentBERT: Fine-tuned only on claims of patents, Uncased
- BERT for patents: Uncased
- **Recall is more important**
- Small scale and single-domain

Model	Label	Precision	Recall
BERT-base	B	0.884	0.922
	I	0.967	0.966
SciBERT (base)	B	0.954	**0.968**
	I	0.980	**0.986**
BERT for patents (large)	B	**0.961**	0.965
	I	0.983	0.978
Patent-BERT (base)	B	0.945	0.963
	I	0.979	0.970
BioBERT (base)	B	0.952	0.962
	I	0.984	0.980
Our Baseline (SciBERT)	B	0.947	0.954
	I	0.986	0.976

Universiteit Leiden

03

Future work

Future work

Reference Extraction

- Use new dataset (larger and more diverse)
- Add other labels to dataset
- Further pre-train the BERT cased model on patents

Reference Matching

- Extracted references → Scientific publications (Web of Science (WOS) database)
- Text matching techniques for ambiguous matching

Universiteit Leiden

Thanks!

Any Questions?

z.abbasiantaeb@leidenuniv.nl