화학 구조 문서 합성 데이터셋 제안 및 Mask R-CNN 기반의 화학 구조 인식

윤정환, 조남익
서울대학교 전기정보공학부
hamil951753@snu.ac.kr, nicho@snu.ac.kr

Synthetic Chemical Structure Documentation Dataset Proposal and Mask R-CNN Based Chemical Structure Segmentation

Jeong Hwan Yoon, Nam Ik Cho
Department of ECE, Seoul National University

요 약

최근 인공지능 신경망에 대한 활발한 연구를 바탕으로 다양한 분야에서의 적용에 대해 많은 시도들이 이루어지고 있다. 이러한 흐름에 맞추어 화학 문서에서 화학 구조를 인식하는 문제 또한 딥러닝을 이용하여 해결하려는 시도들이 생겨나고 있다. 본 논문에서는 화학 문서에서 화학 구조를 인식하는 모델을 학습시키기 위한 합성 데이터셋을 제안하였다. 문서의 구조를 이용하여 정교하게 화학 구조들을 문서에 합성하여 데이터셋을 생성하였고, 이를 최신 딥러닝 모델 중 하나인 Mask R-CNN[7]에 학습시켜 제안한 데이터셋을 이용하여 문서에서 화학 구조를 인식할 수 있음을 보였다.

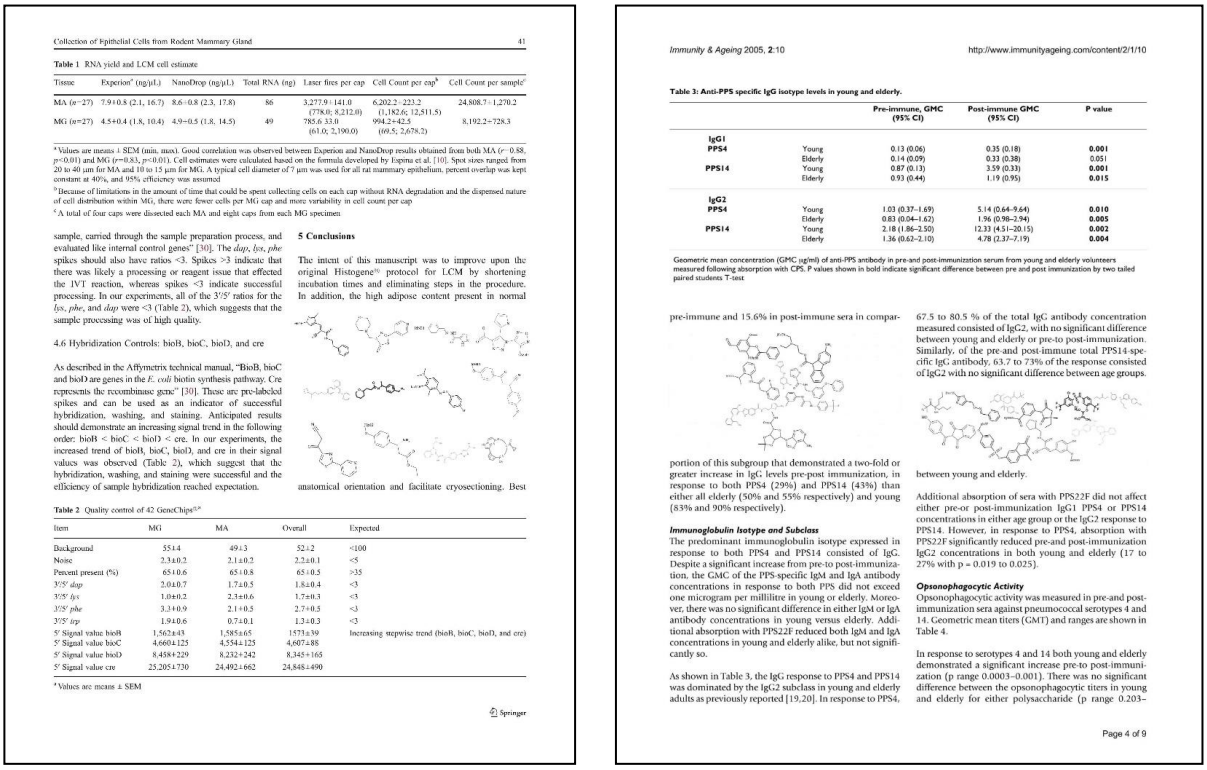
1. 서론

화학 문서에서 화학 구조를 자동으로 인식하는 것은 화학 문서에 존재하는 화학 구조식들을 데이터베이스화 하기 위하여 선행되어야 하는 필수적인 단계이다 [1]. 이러한 화학 구조식들을 화학 문서로부터 인식하고자 하는 수 많은 연구들이 진행되어 왔고, 그 중 일부는 소프트웨어 시스템으로 개발되어 공개되었다 [8,9,10].

하지만 기존의 연구들은 대부분 사람이 직접 설계한 복잡한 규칙을 기반으로 만들어졌기 때문에 유지, 보수에 높은 수준의 전문 지식이 필요했다. 예를 들어 낮은 품질의 이미지나 새로운 도메인의 화학구조가 입력으로 들어오는 경우 인식이 제대로 되지 않아 알고리즘을 개선할 필요가 있을 때 쉽게 대처하기 어렵다는 단점이 있었다.

최근 인공지능 신경망에 대한 활발한 연구를 바탕으로 다양한 분야에서의 적용에 대해 많은 시도들이 이루어지고 있다. 문서 내의 화학 구조 인식에 대해서도 학습 기반의 딥러닝 모델이 기존의 룰베이스드(rule-based) 방식의 대안으로 제시되고 있다 [1,2,6]. 이를 위하여 본 논문에서는 딥러닝을 이용하여 문서에서 화학 구조를 인식하기 위한 합성 데이터셋을 제안하였다. 문서 layout 분석에 사용되는 데이터셋의 레이블을 이용하여 보다 정교하게 문서에 분자 구조들을 합성하였다. 제안한 데이터셋을 이용하여 segmentation 모델을 학습시키고 정량적, 정성적 성능을 평가하여 제안한 데이터셋을 이용하여 학습한 모델이 화학 구조 인식에 효과적임을 보였다.

2.1 합성 데이터셋의 생성



Collection of Epithelial Cells from Rodent Mammary Gland 41

Table 1 RNA yield and LCM cell estimate

Tissue	Experion[a] (ng/μL)	NanoDrop (ng/μL)	Total RNA (ng)	Laser fires per cap	Cell Count per cap[b]	Cell Count per sample[c]
MA (n=27)	7.9±0.8 (2.1, 16.7)	8.6±0.8 (2.3, 17.8)	86	3,277.9±141.0 (778.0; 8,212.0)	6,202.2–223.2 (1,182.6; 12,511.5)	24,808.7±1,270.2
MG (n=27)	4.5±0.4 (1.8, 10.4)	4.9±0.5 (1.8, 14.5)	49	785.6 33.0 (61.0; 2,190.0)	994.2±42.5 (69.5; 2,678.2)	8,192.2±728.3

[a] Values are means ± SEM (min, max). Good correlation was observed between Experion and NanoDrop results obtained from both MA (r^2 = 0.88, p<0.01) and MG (r^2=0.83, p<0.01). Cell estimates were calculated based on the formula developed by Espina et al. [10]. Spot sizes ranged from 20 to 40 μm for MA and 10 to 15 μm for MG. A typical cell diameter of 7 μm was used for all rat mammary epithelium, percent overlap was kept constant at 40%, and 95% efficiency was assumed

[b] Because of limitations in the amount of time that could be spent collecting cells on each cap without RNA degradation and the dispersed nature of cell distribution within MG, there were fewer cells per MG cap and more variability in cell count per cap

[c] A total of four caps were dissected each MA and eight caps from each MG specimen

sample, carried through the sample preparation process, and evaluated like internal control genes" [30]. The *dap*, *lys*, *phe* spikes should also have ratios <3. Spikes >3 indicate that there was likely a processing or reagent issue that effected the IVT reaction, whereas spikes <3 indicate successful processing. In our experiments, all of the 3′/5′ ratios for the *lys*, *phe*, and *dap* were <3 (Table 2), which suggests that the sample processing was of high quality.

4.6 Hybridization Controls: bioB, bioC, bioD, and cre

As described in the Affymetrix technical manual, "BioB, bioC and bioD are genes in the *E. coli* biotin synthesis pathway. Cre represents the recombinase gene" [30]. These are pre-labeled spikes and can be used as an indicator of successful hybridization, washing, and staining. Anticipated results should demonstrate an increasing signal trend in the following order: bioB < bioC < bioD < cre. In our experiments, the increased trend of bioB, bioC, bioD, and cre in their signal values was observed (Table 2), which suggest that the hybridization, washing, and staining were successful and the efficiency of sample hybridization reached expectation.

5 Conclusions

The intent of this manuscript was to improve upon the original Histogene[tm] protocol for LCM by shortening incubation times and eliminating steps in the procedure. In addition, the high adipose content present in normal

anatomical orientation and facilitate cryosectioning. Best

Table 2 Quality control of 42 GeneChips®[a]

Item	MG	MA	Overall	Expected
Background	55±4	49±3	52±2	<100
Noise	2.3±0.2	2.1±0.2	2.2±0.1	<5
Percent present (%)	65±0.6	65±0.8	65±0.5	>35
3′/5′ *dap*	2.0±0.7	1.7±0.5	1.8±0.4	<3
3′/5′ *lys*	1.0±0.2	2.3±0.6	1.7±0.3	<3
3′/5′ *phe*	3.3±0.9	2.1±0.5	2.7±0.5	<3
3′/5′ *trp*	1.9±0.6	0.7±0.1	1.3±0.3	<3
5′ Signal value bioB	1,562±43	1,585±65	1571±39	Increasing stepwise trend (bioB, bioC, bioD, and cre)
5′ Signal value bioC	4,660±125	4,554±125	4,607±88	
5′ Signal value bioD	8,458±229	8,232±242	8,345±165	
5′ Signal value cre	25,205±730	24,492±662	24,848±490	

[a] Values are means ± SEM

Springer

Immunity & Ageing 2005, **2**:10 http://www.immunityageing.com/content/2/1/10

Table 3: Anti-PPS specific IgG isotype levels in young and elderly.

		Pre-immune, GMC (95% CI)	Post-immune GMC (95% CI)	P value
IgG1				
PPS4	Young	0.13 (0.06)	0.35 (0.18)	**0.001**
	Elderly	0.14 (0.09)	0.33 (0.38)	0.051
PPS14	Young	0.87 (0.13)	3.59 (0.33)	**0.001**
	Elderly	0.93 (0.44)	1.19 (0.95)	**0.015**
IgG2				
PPS4	Young	1.03 (0.37–1.69)	5.14 (0.64–9.64)	**0.010**
	Elderly	0.83 (0.04–1.62)	1.96 (0.98–2.94)	**0.005**
PPS14	Young	2.18 (1.86–2.50)	12.33 (4.51–20.15)	**0.002**
	Elderly	1.36 (0.62–2.10)	4.78 (2.37–7.19)	**0.004**

Geometric mean concentration (GMC μg/ml) of anti-PPS antibody in pre-and post-immunization serum from young and elderly volunteers measured following absorption with CPS. P values shown in bold indicate significant difference between pre and post immunization by two tailed paired students T-test

pre-immune and 15.6% in post-immune sera in compar-

portion of this subgroup that demonstrated a two-fold or greater increase in IgG levels pre-post immunization, in response to both PPS4 (29%) and PPS14 (43%) than either all elderly (50% and 55% respectively) and young (83% and 90% respectively).

Immunoglobulin Isotype and Subclass

The predominant immunoglobulin isotype expressed in response to both PPS4 and PPS14 consisted of IgG. Despite a significant increase from pre-to post-immunization, the GMC of the PPS-specific IgM and IgA antibody concentrations in response to both PPS did not exceed one microgram per millilitre in young or elderly. Moreover, there was no significant difference in either IgM or IgA antibody concentrations in young versus elderly. Additional absorption with PPS22F reduced both IgM and IgA concentrations in young and elderly alike, but not significantly so.

As shown in Table 3, the IgG response to PPS4 and PPS14 was dominated by the IgG2 subclass in young and elderly adults as previously reported [19,20]. In response to PPS4, 67.5 to 80.5 % of the total IgG antibody concentration measured consisted of IgG2, with no significant difference between young and elderly or pre-to post-immunization. Similarly, of the pre-and post-immune total PPS14-specific IgG antibody, 63.7 to 73% of the response consisted of IgG2 with no significant difference between age groups.

between young and elderly.

Additional absorption of sera with PPS22F did not affect either pre-or post-immunization IgG1 PPS4 or PPS14 concentrations in either age group or the IgG2 response to PPS14. However, in response to PPS4, absorption with PPS22F significantly reduced pre-and post-immunization IgG2 concentrations in both young and elderly (17 to 27% with p = 0.019 to 0.025).

Opsonophagocytic Activity

Opsonophagocytic activity was measured in pre-and post-immunization sera against pneumococcal serotypes 4 and 14. Geometric mean titers (GMT) and ranges are shown in Table 4.

In response to serotypes 4 and 14 both young and elderly demonstrated a significant increase pre-to post-immunization (p range 0.0003–0.001). There was no significant difference between the opsonophagocytic titers in young and elderly for either polysaccharide (p range 0.203–

Page 4 of 9

그림 1 화학 구조 합성 데이터셋 예시. 격자 구조로 합성한 경우(좌)와 겹치지 않도록 가까이 합성한 경우(우).

본 논문에서는 [2]로부터 영감을 받아 화학 구조가 존재하지않는 문서에 분자 구조를 합성하고 이를 딥러닝 모델의 학습에 사용하여 화학 구조를 segmentation 하는 모델을 만든다. 화학 구조 이미지들은 공개된 PubChem database 에서 접근 가능한 분자 구조들을 이용하였다. 수집된 화학 구조들을 Rdkit 을 이용해 분자 구조의 결합 선 두께, 폰트, 전체적인 크기와 회전 등의 방법으로 데이터를 증대시켜 총 10 만장의 분자 구조 이미지를 생성하였다.

추가적으로 조금 더 정교한 데이터셋 생성을 위해 Publaynet[3]의 레이블을 이용하였다. 이는 PMC Open Access Subset 에서 백만 개 이상의 PDF 문서를 XML 표현과 매칭시켜 자동으로 생성한 36 만장의 문서 데이터셋으로, 주로 문서 layout 분석에 사용되고 있다. 세부적으로는 문서의 layout 을 텍스트 단락, 그림, 표, 리스트, 제목 등 다섯 가지로 분류했고, 각각의 bounding box 와 다각형의 segmentation 표현을 포함하고 있다. 본 논문에서는 Publaynet[3]의 레이블 된 문서 중 대부분이 그림 혹은 표로 이루어져 있는 등 화학 구조를 합성하기에 부적합한 문서들을 제외하고, 분자 구조 이미지의 수를 고려하여 후보 문서 이미지 2 만장을 선택하였다.

Publaynet 의 레이블을 참고하여 임의의 텍스트 단락들을 제거하고 그 위치에 복수의 화학 구조 이미지들을 합성하였다. 합성하는 화학 구조 이미지들은 임의로 선택하였고, 격자 구조 혹은 각 분자 구조의 마스크가 겹치지 않도록 두 가지 방법을 이용하여 배치하였다. 그림 1 은 분자 구조를 합성한 화학 문서의 예시이다. 이를 통하여 최종적으로 2 만 장의 합성 문서 데이터셋을 생성하였고, 학습 18,000 장, 검증 1,000 장, 테스트 1,000 장으로 분류하였다.

2.2 모델 학습 및 결과

본 논문에서 제안한 데이터셋을 이용하여 Mask R-CNN 모델을 학습 시켜 화학 구조 인식을 진행하였다. 현재 다른 문서에서 화학 구조를 segmentation 하는 데이터셋이 공개되지 않았기 때문에 제안한 데이터셋의 테스트셋을 이용하여 모델의 정량적인 성능을 측정하였다. 표 1 은 테스트셋에 대한 정밀도, 재현율, F1 score 를 나타낸 것이다.

또한 데이터셋의 유효성을 검증하기 위해 실제 특허 문서를 이용하여 모델의 성능을 측정하였다. 그림 2 는 실제 특허 문서들에 대한 화학 구조 인식 결과를 나타낸 것이다. 그림 2 의 좌측과 같이 분명하게 나뉘어져 있는 분자 구조들에 대해서는 제대로 인식이 되는 반면, 우측과 같이 분자 구조들의 간격이 좁아 인식하기 어려운 경우들에 대해서는 상대적으로 낮은 성능을 보였다.

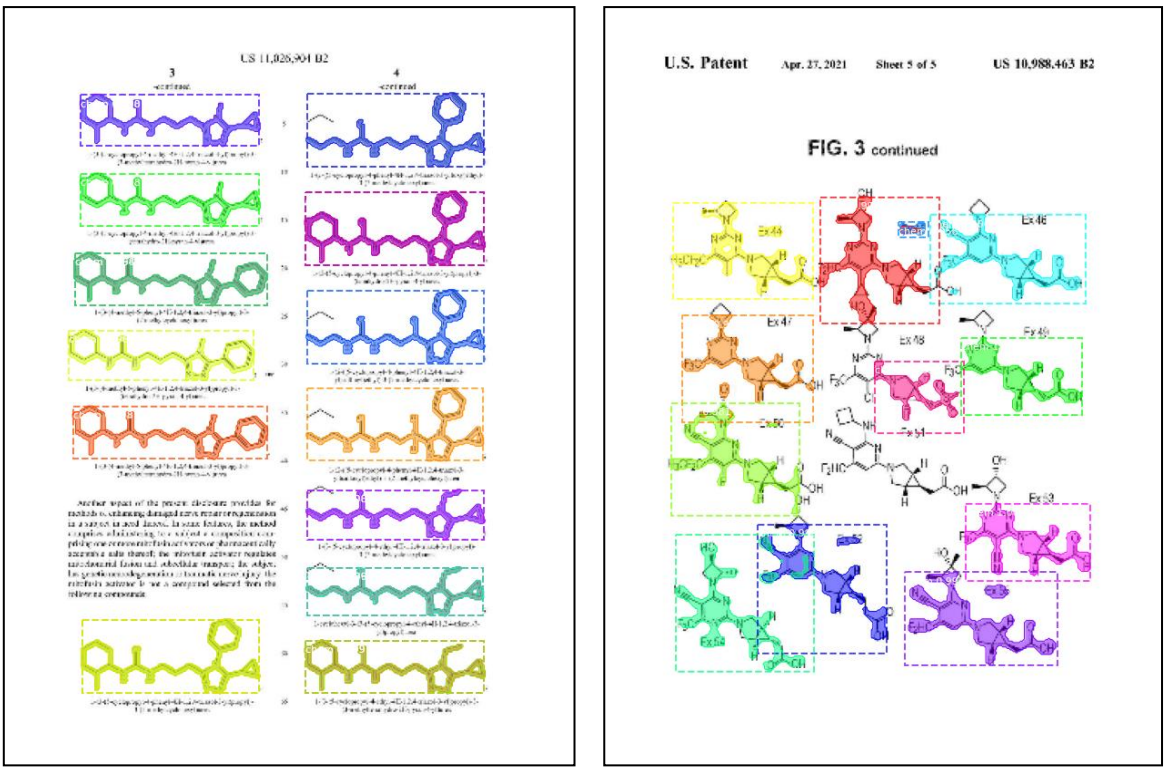


그림 2 특허 문서에 대한 화학 구조 인식 결과. 화학 구조를 잘 인식한 경우(좌)[4] 와 상대적으로 잘 인식하지 못한 경우(우)[5].

표 1 Mask R-CNN[7]을 제안한 합성데이터셋의 테스트셋에 실험한 결과

	정밀도	재현율	F1-score
Mask R-CNN	0.938	0.925	0.931

3. 결론

본 논문에서는 기존에 존재하던 문서 layout 분석 데이터셋[3]을 이용하여 보다 정교한 화학 문서 합성 segmentation 데이터셋을 제안하였다. 최신 segmentation 모델을 제안한 데이터셋을 이용하여 학습시켜 화학 구조를 인식하는 모델을 만들었다. 추후 이를 활용하여 화학 구조를 인식해 분류하는 모델까지 발전시킬 수 있으리라 기대한다.

감사의 글

이 논문은 2022 년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의하여 지원되었음. 그리고 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (2021R1A2C2007220).

참고문헌

[1] RAJAN, Kohulan; ZIELESNY, Achim; STEINBECK, Christoph. DECIMER: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 2020, 12.1: 1-9.

[2] STAKER, Joshua, et al. Molecular structure extraction from documents using deep learning. *Journal of chemical information and modeling*, 2019, 59.3: 1017-1029.

[3] ZHONG, Xu; TANG, Jianbin; YEPES, Antonio Jimeno. Publaynet: largest dataset ever for document layout analysis. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019. p. 1015-1022.

[4] DORN, II Gerald W. *Mitofusin activators and methods of use thereof*. U.S. Patent No 11,026,904, 2021.

[5] DOWLING, Matthew, et al. *Substituted 3-azabicyclo [3.1. 0] hexanes as ketohexokinase inhibitors*. U.S. Patent No 10,988,463, 2021.

[6] OLDENHOF, Martijn, et al. ChemGrapher: optical graph recognition of chemical compounds by deep learning. *Journal of chemical information and modeling*, 2020, 60.10: 4506-4517.

[7] HE, Kaiming, et al. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 2961-2969.

[8] SMOLOV, Viktor; ZENTSEV, Fedor; RYBALKIN, Mikhail. Imago: Open-Source Toolkit for 2D Chemical Structure Image Recognition. In: TREC. 2011.

[9] VALKO, Aniko T.; JOHNSON, A. Peter. CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. Journal of chemical information and modeling, 2009, 49.4: 780-787.

[10] FUJIYOSHI, Akio; NAKAGAWA, Koji; SUZUKI, Masakazu. Robust method of segmentation and recognition of chemical structure images in cheminfty. In: Pre-proceedings of the 9th IAPR international workshop on graphics recognition, GREC. 2011.