Vision-Based Layout Detection from Scientific Literature using Recurrent Convolutional Neural Networks

Huichen Yang, William H. Hsu
Laboratory for Knowledge Discovery in Databases
http://www.kddresearch.org
Department of Computer Science, Kansas State University, USA

MOTIVATION

- **Unstructured scientific literature contains huge valuable information**
- **Rate of published scientific literature is growing rapidly into huge dataset**
- **Lack of metadata information extraction for existing tools, e.g., OCR**
- **Scientific literature layout detection presents a solution of automatic construction of large corpus for downstream tasks of NLP**

CONTRIBUTIONS

- **Consider scientific literature layout detection (SLLD) as object detection task of computer vision**
- **Purpose an end-to-end learning framework based on two-stage object detection framework Faster R-CNN**
- **Present a novel approach to detect the main regions of scientific articles, and output blocks and their corresponding labels**
- **Create synthesis dataset for scientific literature documents layout detection (SLLD) task**

Ground truth

ACS APPLIED MATERIALS & INTERFACES — Research Article — www.acsami.org

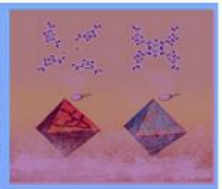
Enhanced Hydrothermal Stability and Catalytic Performance of HKUST-1 by Incorporating Carboxyl-Functionalized Attapulgite

Bo Yuan, Xiao-Qian Yin, Xiao-Qin Liu,* Xing-Yang Li, and Lin-Bing Sun*

Jiangsu National Synergetic Innovation Center for Advanced Materials (SICAM), State Key Laboratory of Materials-Oriented Chemical Engineering, College of Chemistry and Chemical Engineering, Nanjing Tech University, Nanjing 210009, China

Supporting Information

ABSTRACT: Much attention has been paid to metal−organic frameworks (MOFs) due to their large surface areas, tunable functionality, and diverse structure. Nevertheless, most reported MOFs show poor hydrothermal stability, which seriously hinders their applications. Here a strategy is adopted to tailor the properties of MOFs by means of incorporating carboxyl-functionalized natural clay attapulgite (ATP) into HKUST-1, a well-known MOF. A new type of hybrid material was thus fabricated from the hybridization of HKUST-1 and ATP. Our results indicated that the hydrothermal stability of the MOFs as well as the catalytic performance was apparently improved. The frameworks of HKUST-1 were severely destroyed after hydrothermal treatment (hot water vapor, 60 °C), while that of the hybrid materials was maintained. For the hybrid materials containing 8.4 wt % of ATP, the surface area reached 1302 $m^2 \cdot g^{-1}$ and was even higher than that of pristine HKUST-1 (1245 $m^2 \cdot g^{-1}$). In the ring-opening of styrene oxide, the conversion reached 98.9% at only 20 min under catalysis from the hybrid material, which was obviously higher than that over pristine HKUST-1 (80.9%). Moreover, the hybrid materials showed excellent reusability and the catalytic activity was recoverable without loss after six cycles. Our materials provide promising candidates for heterogeneous catalysis owing to the good catalytic activity and reusability.



KEYWORDS: carboxyl functionalization, natural clay, hydrothermal stability, ring-opening reaction, reusability

INTRODUCTION

MOFs are highly porous and crystalline solids whose framework is constructed from metal ions and organic ligands.[1−6] Owing to their high surface area, tunable functionality, and diverse structure, MOFs show interesting properties for gas storage[7−9] and separation,[10−33] as well as use as catalysts[14−17] and sensors.[18−20] However, MOFs are formed by metal−organic coordination bonds, and their hydrothermal stability is poor. Taking MOF-5 as an example, the collapse of structure occurs irreversibly when exposed to an atmospheric environment for 10 min.[21] Water molecules can attack the coordination bonds in the frameworks of MOFs, resulting in the collapse of structure.[22−24] In addition, the structure is unrecoverable even if the water is taken away.[23,25]

Since the discovery of MOFs, many methods have been attempted to overcome the drawback of weak hydrothermal stability. One is the modification of ligands,[27−29] in which hydrophobic ligands are introduced to prevent the coordination bonds from being attacked by water molecules. Another method is the hybridization of MOFs with some materials including silica,[30−32] graphite oxide (GO),[33−36] and carbon nanotubes (CNT).[37−39] Hybridization can improve the properties such as adsorption capacity or hydrothermal stability. However, these silica and carbon hybrids require artificial synthesis, and the use of low-cost and eco-friendly natural materials is highly desired.

Attapulgite is a kind of natural clay and belongs to hydrous magnesium−aluminum silicate minerals. The theoretical formula of attapulgite is $Al_2Mg_2Si_8O_{20}(OH)_2(OH_2)_4 \cdot 4H_2O$.[40−42] There are significant attapulgite reserves in various countries such as China, America, and Spain. In recent years, much attention has been given to the application of attapulgite, owing to its eco-friendly nature, low-cost, and particular morphologies. The structure of attapulgite is shown in Figure S1 in Supporting Information, in which two bands of silica tetrahedra are linked by aluminum ions in octahedral coordination. Additionally, there are abundant active OH groups on the surface that can be used for modification.[43−45] Moreover, it has one-dimensional fibrous morphology, and each fiber has a length ranging from several hundred nanometers to several micrometers, as displayed in Figures S2 and S3, Supporting Information.[43,46] Meanwhile, MOFs have a three-dimensional network structure with a crystal size that varies from nanometers to micrometers and even to millimeters.[47−49] Taking HKUST-1 as an example, it is a typical member of the MOF family and composed of 1,3,5-benzenetricarboxylate (H_3BTC) organic linkers bound by a dicopper tetracarboxylate paddlewheel secondary building unit (SBU). HKUST-1 has an octahedral shape with a width of 15−

Received: April 7, 2016
Accepted: June 7, 2016
Published: June 7, 2016

DOI: 10.1021/acsami.6b04127
ACS Appl. Mater. Interfaces 2016, 8, 16457−16464

Contents lists available at ScienceDirect

Journal of Alloys and Compounds

journal homepage: www.elsevier.com/locate/jalcom

Effects of growth temperatures on the characteristics of n-GaN nanorods–graphene hybrid structures

San Kang [a], Arjun Mandal [a], Ji-Hyeon Park [a], Dae-Young Um [a], Jae Hwan Chu [b], Soon-Yong Kwon [b], Cheul-Ro Lee [a,*]

[a] Semiconductor Materials Process Laboratory, School of Advanced Materials Engineering, Engineering College, Research Center for Advanced Materials Development (RCAMD), Chonbuk National University, Baekje-daero 567, Jeonju 561-756, Republic of Korea
[b] School of Materials Science and Engineering, Low Dimensional Carbon Materials Center, Ulsan National Institute of Science and Technology, UNIST-gil 50, Ulsan 689-798, Republic of Korea

ARTICLE INFO

Article history:
Received 20 April 2015
Received in revised form 8 May 2015
Accepted 11 May 2015
Available online 16 May 2015

Keywords:
Ultraviolet photoconductive devices
Hybrid structures
GaN nanorods
Graphene
Metal organic chemical vapor deposition
Growth temperature

ABSTRACT

The effects of different growth temperatures of n-GaN nanorods (NRs) on the material and electrical properties of n-GaN NRs–graphene hybrid device structures are being demonstrated for the first time. A high quality graphene transfer method was applied for transferring the graphene layer on Si (1 1 1) substrate and n-GaN NRs were synthesized on the graphene layer on Si using a metal organic chemical vapor deposition (MOCVD) process of high V/III ratio. No metal-catalyst or droplet seeds were formed when growing n-GaN NRs. The growth temperature of the n-GaN NRs was varied from 860 °C to 900 °C. Raman spectroscopy confirmed the prominent existence of an undamaged graphene layer in all of the highly-matched hybrid device structures under study. Improvement in the structural, crystalline and material properties was established from FE-SEM, XRD and PL studies for the hybrid structure where n-GaN NRs were grown at 890 °C. The same hybrid structure also showed a ten-fold enhancement in photocurrent along with increased sensitivity and photoresponsivity. Therefore, it can be concluded that a suitable growth temperature of n-GaN NRs is the most important factor for the fabrication of high quality n-GaN NRs–graphene hybrid structures.

1. Introduction

Nowadays, semiconductor–graphene hybrid structure [1–3] is an area of research which is being studied extensively. As we know, graphene [4–6] has excellent electrical conductivity [7] and can act as a highly efficient transport medium of charge carriers [2,8,9], and is therefore ideal to be used in electronic devices. Thus, graphene integrated with direct band gap semiconductors may offer photoconductive devices with enhanced performances and these hybrid structures have the potential for low-cost mass production. For example, in the case of solar cells, high charge carrier mobility in the graphene layer can enhance the rate of collection of photo-generated carriers and so its efficiency [10–12].

Undoubtedly, semiconductor nanorods (NRs) [13–15] are among those diverse quantum structures which are being studied actively. In this article, the photoconductive behavior of GaN NRs integrated with graphene is explored. Over the last decade, GaN-based nano-scale semiconductor devices have drawn much attention from researchers mainly because of their outstanding properties like wide direct band gap (3.4 eV), high thermal conductivity (1.3 W cm^{-1} K^{-1}) and high saturated electron velocity [16]. To be cost-effective, the vapor–liquid–solid (VLS) technique [17] using metal organic chemical vapor deposition (MOCVD) [18] is ideal for growing GaN nanorods. The MOCVD technique also offers precise growth control and mass production.

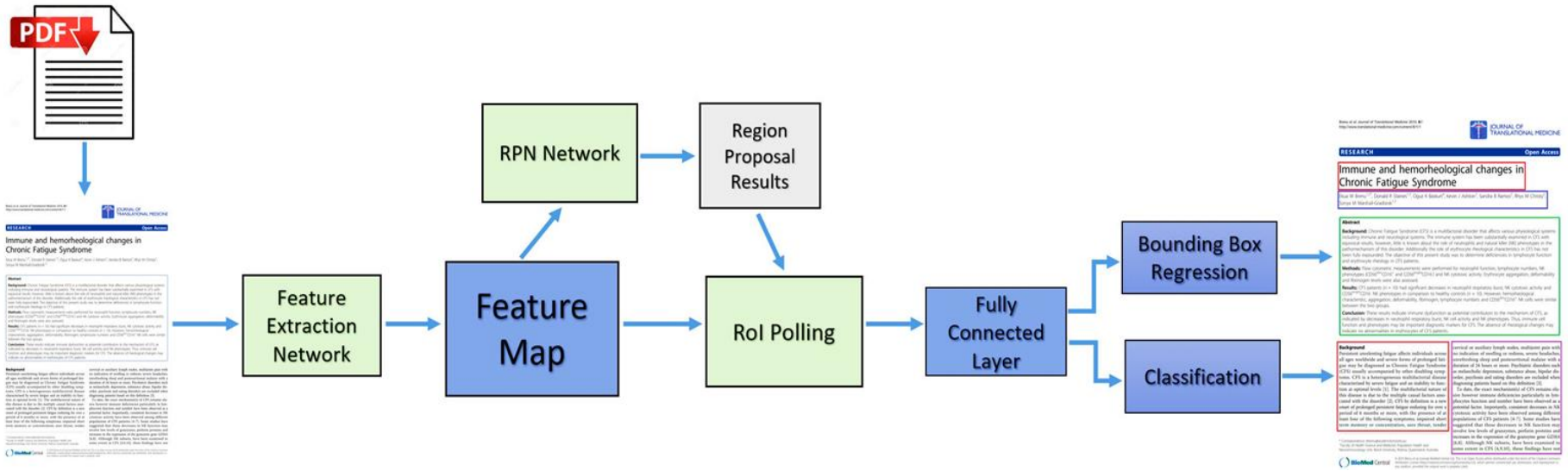
The assurance of the prominent presence of the underlying graphene layer in the NRs–graphene hybrid structure is a must for efficient photoconduction in the device. In most cases, upon completion of nanorod growth on graphene, the graphene layer is found to be impaired or the semiconductor properties of the NRs are compromised. Our present study deals with this challenge and succeeds in presenting an n-GaN NRs–graphene hybrid structure on Si substrate keeping the original properties of GaN NRs and graphene intact. n-GaN NRs were synthesized on graphene using the MOCVD technique and the effects of variation of growth temperature of n-GaN NRs were studied.

The hybrid device structures under study are ideal for developing photoconductive devices of high efficiency. Moreover, to avoid contamination, no metal-catalyst was used to grow n-GaN NRs

* Corresponding author. Tel.: +82 63 270 2304; fax: +82 63 270 2305.
E-mail address: crlee7@jbnu.ac.kr (C.-R. Lee).

http://dx.doi.org/10.1016/j.jallcom.2015.05.098
0925-8388/

SCIENTIFIC LITERATURE LAYOUT DETECTION FRAMEWORK

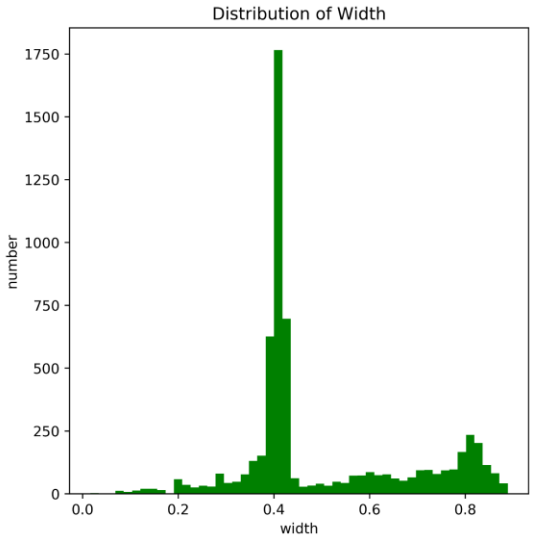


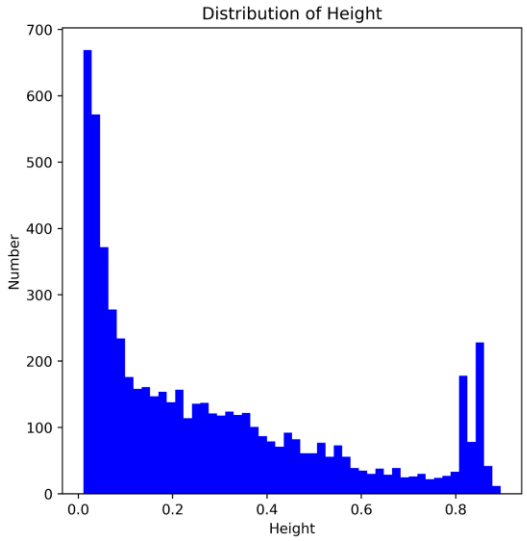
METHODOLOGY
BACKBONE

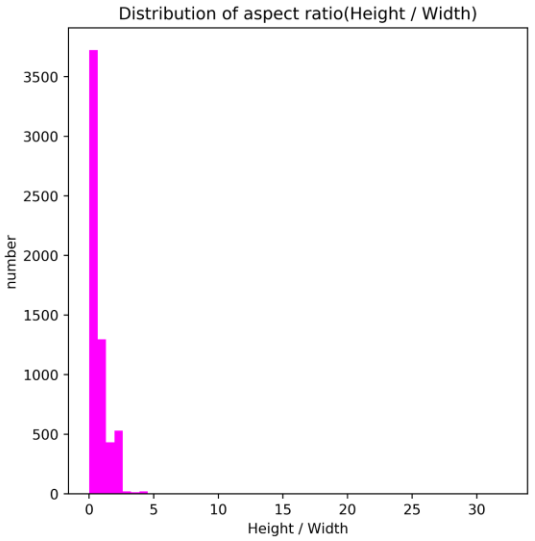
- **Use pre-trained VoVnet-v2 [Huang et al., 2017] on MS COCO dataset as backbone for feature extraction**
  - **Better performance – aggregate concatenation feature only once in last feature map**
  - **Residual connection enables to train deeper networks**
  - **Squeeze-and-Excitation (eSE) attention module improves feature extractor performance**

Methodology
Aspect ratio selection

- **Anchors aspect ratio selection**
  - **Analyze distribution bounding box sizes-based ground truth synthesis dataset**
  - **Use K-means cluster anchor box selection to get aspect ratios for different blocks ranging from 0.1 to 4.0**







Experiment dataset

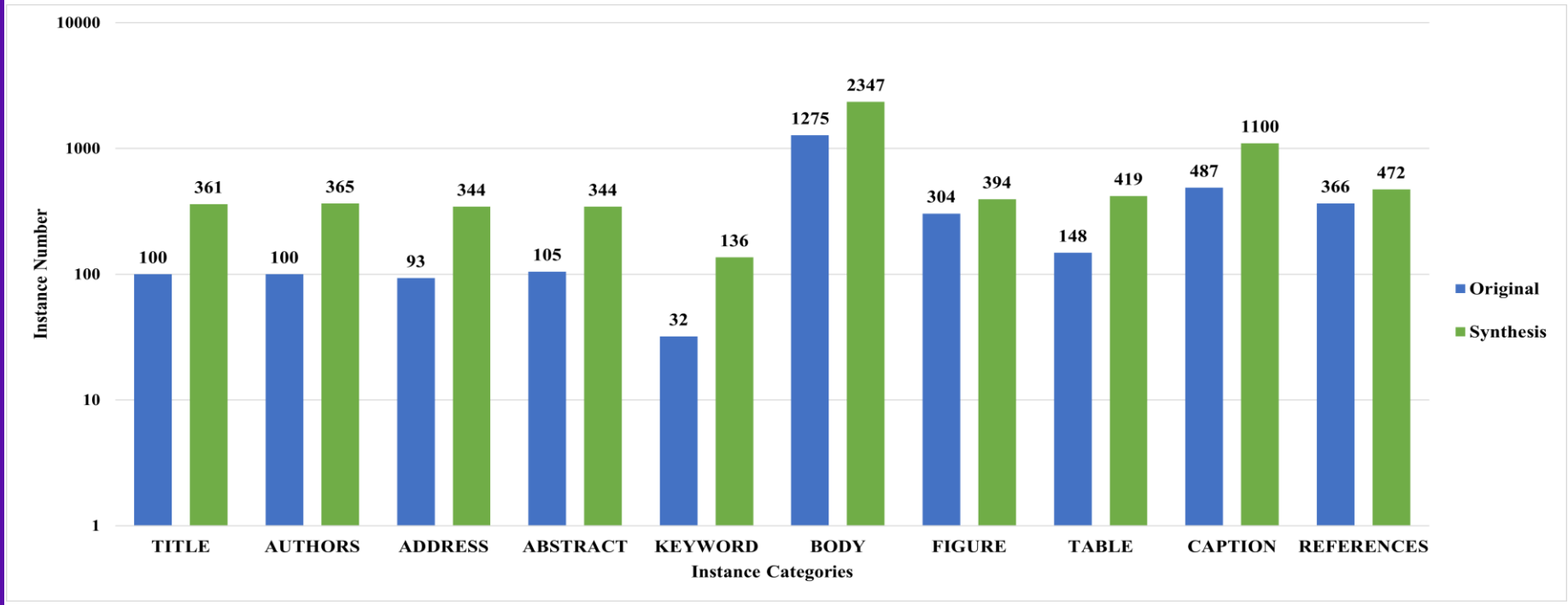
- **Propose synthesis dataset to relieve imbalance issue**
  - **Region annotations dataset – 822 images from 100 PDF scientific literature [Soto and Yoo, 2019]**
  - **ICDAR-2013 – 150 table images from 76 PDF documents [Gobel et al., 2013]**
  - **GROTOAP- 113 annotated PDF scientific literature [Traczyk et al., 2012]**
- **Final synthesis dataset**
  - **1550 images from 363 PDF documents**
  - **Convert images to fixed size – 612 x 729 at 200 dpi**

Experiment
dataset labels

- **Use 10 labels to classify major regions from scientific literature**
  - ***Title*: the title and subtitles**
  - ***Authors*: the author names**
  - ***Address*: the affiliation information of authors, including authors' address, email, etc.**
  - ***Abstract*: an abstract section**
  - ***Keyword*: the selected keywords**
  - ***Body*: the main block of articles**
  - ***Figure*: all figures but excluding logos or icons from publishers.**
  - ***Table*: the tabular contents**
  - ***Caption*: the captions for both figures and tables**
  - ***Reference*: the bibliography information, excluding post-references notes**

Dataset comparison

Instances comparison between region annotations dataset and our dataset by labels



EVALUATION

- **Detector performance is evaluated with IoU (intersection of unit)**
  - **Data set1(D1): original data set [2] - 600 image for training, 222 images for testing**
  - **Data set2(D2): synthesis data set - 1225 images for training, 325 images for testing**

Detector	Backbone	Data Set	mAP	AP50	AP75	APs	APm	APl	AR
Soto et al.(30 epochs) [28]	ResNet101	D1	-	70.30	-	-	-	-	-
Faster R-CNN (**baseline**)	ResNet50_FPN	D1	69.76	87.46	76.49	-	51.65	77.41	62.70
Faster R-CNN	ResNet50_FPN	D2	77.48	92.39	84.42	35.00	63.32	77.65	69.50
Mask R-CNN	ResNet50_FPN	D1	70.68	87.60	82.90	-	52.05	75.05	65.50
Mask R-CNN	ResNet50_FPN	D2	77.66	91.79	85.80	40.00	64.378	75.604	69.50
YOLOV3 (49 epochs) [28]	-	D1	-	68.90	-	-	-	-	-
YOLOV3	DarkNet53	D2	45.90	66.50	57.10	-	-	-	46.33
Faster R-CNN	VoVNetV2-39	D1	67.12	89.01	72.84	-	47.66	73.56	60.50
Faster R-CNN (**ours**)	VoVNetV2-39	D2	76.39	**95.02**	**86.46**	**75.00**	62.25	74.22	68.80

Detection results

Brügger-Andersen *et al. Thrombosis Journal* 2010, **8**:1
http://www.thrombosisjournal.com/content/8/1/1

THROMBOSIS JOURNAL

ORIGINAL CLINICAL INVESTIGATION Open Access

title 100%

The activity of pregnancy-associated plasma protein A (PAPP-A) as expressed by immunohistochemistry in atherothrombotic plaques obtained by aspiration thrombectomy in patients presenting with a ST-elevation myocardial infarction: a brief communication

authors 100%

Trygve Brügger-Andersen[1,2*], Leif Bostad[4,5], Dagny Ann Sandnes[3], Alf Inge Larsen[1,2], Vernon VS Bonarjee[1,2], Ståle Barvik[1,2], Tor Melberg[1,2], Dennis WT Nilsen[1,2]

abstract 100%

Background: The expression of pregnancy-associated plasma protein A (PAPP-A) was identified by immunohistochemistry (IHC) in culprit atherothrombotic plaque specimens harvested from patients admitted with ST-segment elevation myocardial infarction (STEMI).

Methods: The atherothrombotic samples were collected from a consecutive cohort consisting of 20 individuals admitted with STEMI to Stavanger University Hospital, Norway, from 2005-2006, presenting angiographically with an acute thrombotic occlusion of a coronary artery characterized by TIMI flow 0. The atherothrombotic plaques were obtained by aspiration thrombectomy during percutaneous coronary intervention within 12 hours from the onset of symptoms and prepared for IHC analysis.

Results: In the IHC analysis staining for PAPP-A occurred in the extracellular matrix of the plaques and no evidence of staining for PAPP-A was found in the thrombi.

Conclusion: Our results indicate that in vivo PAPP-A is strongly expressed in atherothrombotic plaques harvested from patients admitted with STEMI, as documented by IHC.

Trial registrationbiobankregisteret@fhi.no1846

body 100%

Background

Pregnancy-associated plasma protein A (PAPP-A) is a zinc-binding matrix metalloproteinase that can be detected in the blood of patients with acute coronary syndromes (ACS) [1,2]. There is histological evidence, using specific monoclonal antibodies, that PAPP-A is abundantly expressed in both eroded and ruptured coronary plaques, but not in stable plaques, in patients who have died suddenly of cardiac causes. Furthermore, accumulating evidence suggests that PAPP-A may play a pivotal role in the development of atherosclerosis and subsequent plaque instability in ACS patients [1].

In a prior study we have assessed the immediate effects of coronary reperfusion procedures on the plasma concentrations of PAPP-A in patients admitted with ACS and ST-elevation myocardial infarction (STEMI) [3]. However, existing data does not allow us to define the exact role of PAPP-A in plaque disruption. Although, this metalloproteinase has been shown in earlier studies to be expressed in ruptured plaques, these results were limited by the fact that the histological samples were collected postmortem [1]. Therefore, we wanted to identify the expression of PAPP-A by immunohistochemistry (IHC) in culprit

author block 100%

*Correspondence: hagerman@online.no
[1]Institute of Medicine, University of Bergen, 5021 Bergen, Norway

body 100%

garding future questionnaires and research. ACT participants were drawn from a cohort of individuals usually tested because of a family member with AAT deficiency. Individuals identified through the Alpha-1 Foundation Registry were drawn from a group of 3090 registry participants with severe AAT deficiency or the carrier state. Since individuals with symptoms are more likely to be tested for AAT deficiency, the participants of the ACT Study and the Research Registry do not represent population-based cohorts. Therefore, to approximate the same biases in a control population, each individual with the PiSS genotype was randomly age, sex, race, and US state matched to an ACT participant with the PiMM genotype as a control.

Study participants were telephone interviewed by a certified genetic counselor, from November 2007 through February 2009, with a standardized questionnaire to identify demographic features, cigarette smoking history, medication history, past and present clinical diseases, and surgeries. A detailed pedigree was obtained, including race/ethnicity and a history of fetal demise or stillbirth. Social history including vocation and artistic experience was collected.

Statistics were performed by JMP (SAS Institute, Cary, NC). A two-tailed Student's *t*-test for continuous variables and Pearson's Chi-Squared test for categorical variables were used for comparative statistics when comparators had more than five observations. A *P*-value <.05 was accepted as significant. No correction for multiple comparisons was made.

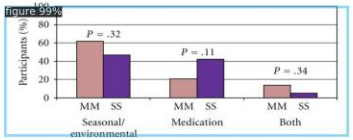
The Hardy-Weinberg statistics were applied to the genotypes detected in the ACT study to calculate the S allele frequency in that study population. The total number of S alleles detected in the ACT study (812) was divided by the total number of alleles (15320) thus determining the S allele frequency (5.3%). The S allele frequency was squared and then multiplied by the total number of participants in the ACT study (7660) to determine the predicted number of individuals with the PiSS genotype (21).

A comprehensive literature review was conducted by two investigators (DM, CS). A recent meta-analysis of chronic obstructive pulmonary disease (COPD) risk associated with PiS alleles [4] was the starting point for our evaluation since the 119 articles reviewed in detail reported only seven articles that mentioned cases of the PiSS genotype. These seven articles were reviewed for any other clinical information in addition to COPD. In addition, testing studies that reported outcomes on more than 100 patients were reviewed from the dataset of de Serres [5] to ascertain if information was included on individuals that were PiSS. References of articles that discuss cases of PiSS were indexed and translations of articles in other languages from English were obtained. Tables of general population and disease specific testing were constructed.

3. Results

The ACT study and the Alpha-1 Foundation Research Registry enrollments by genotype as of January 1, 2008 are shown in Table 1. The ACT study identified 34 individuals

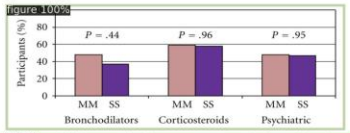
figure 99%



body 99%

FIGURE 1: Percentage of seasonal/environmental allergies, medication allergies, or both seasonal/environmental and medication allergies in the study cohort ($N = 19$) compared to the controls ($N = 29$).

figure 100%



caption 99%

FIGURE 2: Lifetime medication use percentage in the study cohort ($N = 19$) compared to the controls ($N = 29$).

body 100%

with the PiSS genotype. Eight of these individuals chose to not continue in the longitudinal outcome study after receiving their results per study protocol [6], and contact information was not available. Sixteen of the remaining 26 individuals were interviewed. One of the 26 identified individuals died prior to being interviewed. The remaining nine individuals were not interviewed due to incorrect contact information ($N = 6$) or did not return three telephone calls ($N = 3$).

Five individuals in the Registry were identified with the PiSS genotype. One of these five individuals was also in the ACT study. Therefore, four additional individuals were identified for the current study. Three individuals were interviewed. The remaining person did not return the researcher's phone calls on three occasions but reported asthma in the Registry database. Therefore, 19 individuals comprise the current PiSS study cohort.

3.1. Participants of Current Study. Table 2 shows patient demographics for the cohort of the PiSS genotype and the control participants with the PiMM genotype. Eighteen participants were non-Hispanic Caucasian. One ACT study participant was Hispanic. There was no difference in smoking status, smoking duration, secondhand smoking incidence, or secondhand smoking duration between participants with the PiSS genotype and the PiMM genotype. Table 3 lists physician-reported diagnoses for COPD, asthma, and hepatic steatosis (fatty liver) and self-reported jaundice, frequent pneumonia, and bronchitis. Lung disease symptoms, present in 11 of 19 (57.9%) participants from the

REFERENCE

- [Huang et al., 2017] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).
- [Soto and Yoo, 2019] Soto, C., & Yoo, S. (2019, November). Visual Detection with Context for Document Layout Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP)* (pp. 3455-3461).
- [Gobel et al., 2013] Gobel, M., Hassan, T., Oro, E., & Orsi, G. (2013, August). ICDAR 2013 table competition. In *2013 12th International Conference on Document Analysis and Recognition* (pp. 1449-1453). IEEE.
- [Traczyk et al., 2012] Tkaczyk, D., Czeczko, A., Rusek, K., Bolikowski, L., & Bogacewicz, R. (2012, June). GROTOAP: ground truth for open access publications. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (pp. 381-382).

Thanks

Contact: KSU KDD Lab

http://www.kddresearch.org