Scientific texts analysis and classification by using machine learning methods

Nicolas LAFITTE

September 10, 2019

Abstract

The project is part of the Artificial Intelligence (AI) and Machine Learning (ML) training course held at Sorbonne University (Paris, France) during 2018-2019 academic year. As final project, a work on the analysis and classification of scientific texts have been chosen.
The initial motivations were to discover and learn mathematical algorithms and methods used in AI and ML in order to evaluate any application for the Fluigent European project HoliFAB. It aims at adapting an existing pilot line for the production of microfluidic instruments, and develops hardware and software strategies for the optimization of the production and system. Some mathematical developments at the beginning of the project led to the implementation of regression functions that auto-place and auto-wire a system layout.

However the motivation for the current project raises from the frustration as a researcher to not be able to read and study all papers relevant of a field. Making a bibliography on a specific topic is common and easy to perform thanks to different database and search engine available on the internet. However, when it comes to study a broad scientific field for a global understanding or new research investigation or market analysis, the task often lacks an intensive study of the domain.

Text mining, which is the task of extracting meaningful information from text, is developed here on a database of scientific papers. More precisely Latent Dirichlet Algorithms (LDA) are studied in order statistically categorize text and extract topics. The tool helps to find and name topics and applied on different database of different years check for evolution in the scientific research and next innovation that will probably lead the market.

Keywords: mining, classification, clustering, information extraction, topic extraction, information retrieval, Latent Dirichlet Algorithm (LDA)

Contents

Chapter 1

Introduction

Disclaimer: This report is widely inspired by the review of Allahyari et al., A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques [Allahyari et al., 2017].

In a very recent paper [Tshitoyan et al., 2019], researchers at Lawrence Berkeley National Laboratory have developed an artificial intelligence (AI) that paves the way to predict discoveries in science. The scientists gathered 3.3 million articles on materials science from 1,000 different journals published between 1922 and 2018, and trained the renowned Word2vec algorithm in order to build statistical connections between words that are in the same context (cf. words embedding). On the one hand, the program was able to classify well known thermoelectric materials explicitly mentioned in the scientific abstracts alongside the word "thermoelectric" or associated words like 'ZT', 'zT', 'seebeck', 'thermoelectric', 'thermoelectrics', 'thermoelectrical', 'thermoelectricity', 'thermoelectrically' or 'thermopower'. However by mathematical projection of all the materials, it is also indicating a relationship that is not explicitly written in the text Figure 1.1. Particularly Figure 1.1-c demonstrates how words that are chemical formula, ie. totally new character strings, are associated through concept words expressing applications (electronic, optoelectronic, photovoltaic), physical parameters (bandgap, heusler compound) or others known thermoelectric material (PbTe, Cu_2Te, Cu_5Te_3).
On the other hand, articles from 2000 to 2018 were removed and 18 new predictive models were trained in order to predict discoveries the following years and evaluate prediction abilities (Figures 1.2 and 1.3). Each of these models ranked materials according to their similarity to the word "thermoelectric" (or "ferroelectric", "photovoltaic", "topological insulator"), and took the top 50 that were not studied as thermoelectrics as of that year. It turns out, many of these materials were subsequently reported as thermoelectrics in future years.

This work is perfect example of the target of the current work, ie. study future research and markets on a specific field that is here microfluidics.

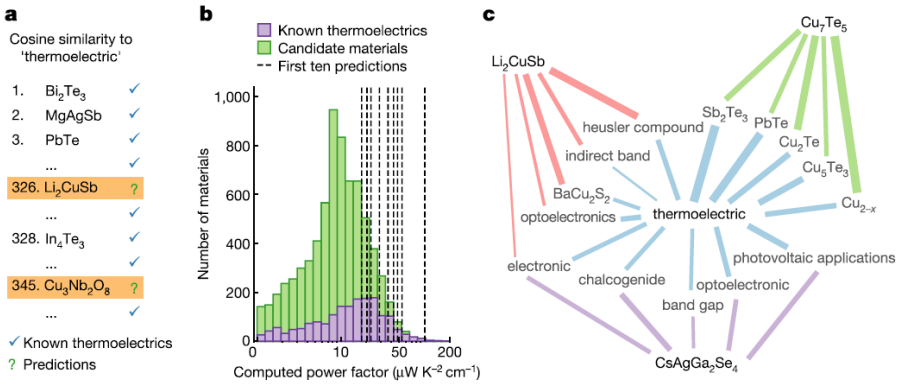


Figure 1.1: Prediction of new thermoelectric materials based on 3.3 million scientific articles from 1,000 different journals published between 1922 and 2018 [Tshitoyan et al., 2019].
a. Ranking table of thermoelectric materials. Materials with a check symbol are found in the context with thermoelectric terms (ie. 'ZT', 'zT', 'seebeck', 'thermoelectric', 'thermoelectrics', 'thermoelectrical', 'thermoelectricity', 'thermoelectrically' or 'thermopower'). However materials with a interrogation point are not explicitly studied as thermoelectric but are "mathematically" close and potential predictions that can be tested in the future.
b. Distributions of the power factors (computed with specific chemistry calculations) for 1,820 known thermoelectrics in the literature (purple) and 7,663 candidate materials not yet studied as thermoelectric (green). Dashed lines show the 10 first predictions of table **a**: Li_2CuSb ...
c. Graph showing how the context words of materials predicted to be thermoelectrics connect to the word thermoelectric. The materials are the first (Li_2CuSb), third ($CsAgGa_2Se_4$) and fourth (Cu_7Te_5) predictions of table **a**. Examination of the context words demonstrates that the algorithm is making associations on the basis of crystal structure, co-mentions with other materials for the same application, between different applications and key phrases that describe the material's known properties.

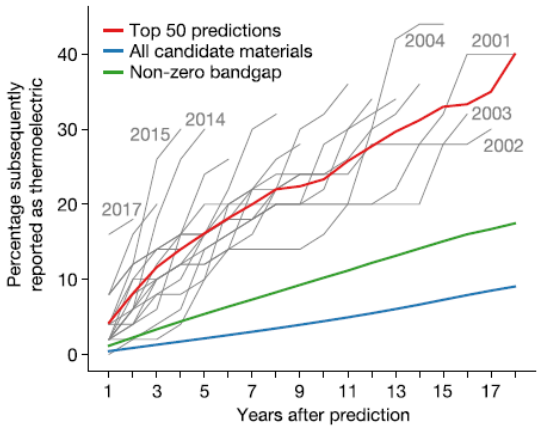


Figure 1.2: Predictions of thermoelectric materials based on previously published papers [Tshitoyan et al., 2019].
For example, predictions for 2001 are performed using abstracts from 2000 and earlier, and the grey lines plot the cumulative percentage of predicted materials subsequently reported as thermoelectrics in the years following their predictions. The results are averaged (red curve) and are 8 times more likely to have been studied as thermoelectrics within the next five years as compared to a randomly chosen unstudied material from our corpus at that time (blue curve) or three times more likely than a random material with a non-zero band gap (green curve).

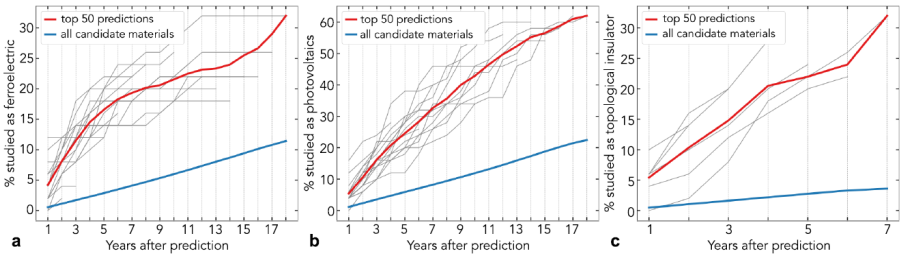


Figure 1.3: Same prediction methods as Figure 1.2, but with words related to ferroelectric (graph **a**), photovoltaic (graph **b**) and topological insulator (graph **c**). Supplementary material of [Tshitoyan et al., 2019].

1.1 Text mining

Text Mining or knowledge discovery from text refers to the process of extracting high quality of information from structured database (i.e. RDBMS), semi-structured (ie. XML and JSON files), or unstructured text resources (ie. .pdf documents). It widely covers a large set of related topics and algorithms for analyzing text, spanning various communities, including information retrieval, natural language processing, data mining, machine learning many application domains web and biomedical sciences. Allahyari et al. describe the notions of knowledge discovery, data mining, information retrieval, information extraction, text summarization or natural language processing in [Allahyari et al., 2017].

1.1.1 Existing approaches

Unsupervised learning methods

Unsupervised learning methods try to find hidden structure out of unlabeled data. Clustering and topic modeling are the two commonly used unsupervised learning algorithms used in the context of text data.
Clustering is the task of segmenting a collection of documents into partitions where documents in the same group (cluster) are more similar to each other than those in other clusters.
In topic modeling a probabilistic model is used to determine a soft clustering, in which every document has a probability distribution over all the clusters as opposed to hard clustering of documents. In topic models each topic can be represented as a probability distributions over words and each documents is expressed as probability distribution over topics. Thus, a topic is akin to a cluster and the membership of a document to a topic is probabilistic

Supervised learning methods

As aforementioned in the beginning of the introduction, supervised learning methods are machine learning techniques pertaining to infer a function or learn a classifier from the training data in order to perform predictions on unseen data. There is a broad range of supervised methods such as nearest neighbor classifiers, decision trees, rule-based classifiers and probabilistic classifiers.

1.1.2 Some current applications

Text processing

Natural Language Processing (NLP)

Text summarization

Text streams and social media mining

Opinion mining and sentiment analysis

1.2 Current work objective, database and approach

1.2.1 Objective

The work targets to study scientific papers specialized in microfluidics and extract hot topics.

On the one hand, amount of accessible texts has been increasing rapidly, and potentially contain a great wealth of knowledge. On the other hand, analyzing huge amounts of textual data requires a tremendous amount of work in reading all of the text and organizing the content. Thus, the increase in accessible textual data has caused an in- formation flood in spite of hope of becoming knowledgeable about various topics.

1.2.2 Text database

First samples are from a conference proceedings specialized in microfluidics: MicroTAS or µTAS (Appendice A, http://microtas2019.org/). As undermentioned in following Chapter, the text pre-processing is complex because of the text encoding of the .pdf files first and the heterogeneous paragraph sectioning by the authors.

The second step is to analyze more microfluidics papers for scientific journals (ie. RSC Lab-on-a-Chip, Springer Microfluidics and Nanofluidics . . .), and microfludics applications journals (ie. à remplir . . .)

In both cases, the access to the full text is quite expensive. However, in the case of journals, title, keywords and abstract are accessible on the internet and potentially in semi-structured files such as XML.

1.2.3 Selected approach

Probabilistic Methods for Text Mining: There are various probabilistic techniques including unsupervised topic models such as probabilistic Latent semantic analysis (pLSA) and Latent Dirichlet Allocation (LDA), and supervised learning methods such as conditional random fields that can be used regularly in the context of text mining.

Chapter 2

Implementation

The main source of machine-interpretable data for the materials research community has come from structured property databases. Beyond property values, publications contain valuable knowledge regarding the connections and relationships between data items as interpreted by the authors. To improve the identification and use of this knowledge, several studies have focused on the retrieval of information from scientific literature using supervised natural language processing, which requires large hand-labelled datasets for training.

2.1 Pipeline

2.2 Data extraction

2.2.1 Sources

2.2.2 Converting pdf to dataframe

2.3 Text transformation

The texts need some transformation to be relevant for algorithms or models applied afterwards. Typically models sensitive to term appearance frequency to determine importance of the words will be biased by recurrent term like in English: *I, and, or ...* that do not carry much meaning.
A transformer is an abstraction that includes feature transformers and learned models: ie. A transformer implements a method, which converts one dataframe into another, generally by appending a new column.

2.3.1 Tokenization

Tokenization is the process of taking the text (such a sentence) and breaking it into individual terms (usually words).

2.3.2 Stop words

Stop words process takes as input a sequence of strings (e.g. the output of the tokenization) and drops all the stop words.
Stop words are words which should be excluded because the words appears frequently and carry as much meaning.

2.3.3 Stemming

2.4 Algorithm

2.4.1 Latent Dirichlet Algorithm

2.4.2 Others

Chapter 3

Results

Chapter 4

Conclusion and perspectives

Bibliography

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Appendices

A MicroTAS proceeding papers

MicroTAS or μTAS (standing for micro total analysis systems) is annual conference series that are the premier forum for reporting research results in microfluidics, microfabrication, nanotechnology, integration, materials and surfaces, analysis and synthesis, and detection technologies for life science and chemistry. The Conference offers plenary talks as well as contributed oral presentations and posters selected from submitted abstracts. This year, the Twenty Third μTAS 2019 will be held at the Congress Center Basel, in Basel, Switzerland from 27-31 October, 2019.

A.1 Proceeding paper feature

Every year about 700 abstracts are accepted. Proceedings papers are 4 pages paper with title, authors names and affiliations, keywords and regular sections such as abstract, main, results, conclusions and references. However these section labels are sometimes not respected by the authors and consequently complicated for text extraction.

A.2 MicoTAS proceeding paper example (from 2010)

AN OPEN MICROFLUIDIC DEVICE WITH ACTIVE VALVES FOR ACCURATE TRAPPING OF DNA BY SILICON NANOTWEEZERS

N. Lafitte[1*], M. Kumemura[1], M. Nagai[2], L. Jalabert[1], D. Collard[1,2] and H. Fujita[2]
[1]LIMMS-CNRS/IIS, UMI2820, The University of Tokyo, JAPAN and
[2]Institute of Industrial Science, The University of Tokyo, JAPAN

ABSTRACT

This paper demonstrates the real-time monitoring of λ-DNA molecule trapping by silicon nanotweezers in an open microfluidic chamber. An active microfluidic device has been developed aiming to allow the insertion of MEMS tweezers and to control the biological solution inlets for an accurate sensing of bioreactions through the tweezer mechanical frequency response.

KEYWORDS: Nanotweezers, DNA trapping, Open microfluidic, Active valves

INTRODUCTION

During the past years, direct manipulation of DNA molecules has expanded our understanding on molecular biology. New tools for basic investigations on DNA mechanical properties have enabled single-molecule assays of enzyme mechanisms, clearing the nature of interactions between DNA and proteins and the forces within which the cellular machinery operates [1-4]. Hence a huge interest exists on systematic and real-time molecular analysis, we proposed a micromechanical-electro system (MEMS) for the manipulation of biomolecules and the characterization of interactions between DNA and enzymes [5,6].

As the characterization of biological phenomena with MEMS tweezers is based on parameter tracking, obtaining reliable biological environment for high sensitivity experiment presents a real challenge for relevant experiments. In this new development, a complementary microfluidic device has been designed and fabricated in order to achieve control of the solution inlet/outlet sequences and volumes (Figure 1). Here, we performed biological experiments in reliable conditions settling reaction time and preventing evaporation.

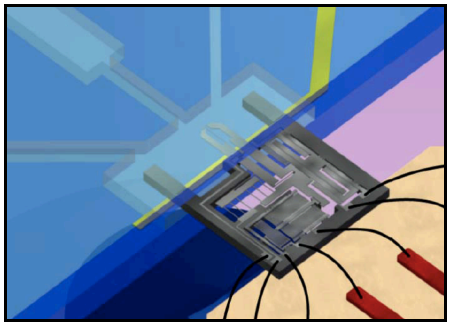


Figure 1: 3-D view of the experimental setup combining nanotweezers and a specific microfluidic device

METHODS

The MEMS tweezers consist of a pair of opposing nanotips. The distance between the tips can be adjusted with nanometer accuracy by a comb drive actuator and measured by a capacitive sensor. With tweezers, bundles of DNA can be repeatedly trapped between their tips in a droplet, and the temporal development of bioreactions on λ-DNA bundle was measured with HindIII restriction enzymes [6]. On the other hand, as passive microfluidic devices do not allow a proper filling of an open chamber and remote control microfluidic system cannot provide enough quick response to compensate evaporation, techniques of soft lithography have been used to make monolithic valves from polydimethylsiloxane (PDMS) [7].

We fabricated our valves using crossed-channel architecture. The device is fabricated by replica molding from two masters and sealing the layers together. A thin layer is produced for controllable flow channels when a thick layer is produced for the control channel and the implementation of an open reaction chamber. The valve membranes are formed where the control and the flow channels intersect orthogonally. The control layer is bonded underneath the flow layer forming push-up valves, and the flow layer is sealed with a glass slide as top layer for optical convenience (Figure 2).

Flow channel wafer is patterned with AZ-4903 photoresist and the resist is reflowed by thermal heating to form rounded-shape. The shape of the flow channel is consequential for proper actuation and hermetic closing of the valve. The 10-µm thin elastomeric membrane is created above the patterns by adjusting the PDMS spin coating speed. Control channel wafer is patterned with SU-8 photoresist. Lastly a 200-µm chamber is implemented to insert the 25-µm-thick of tweezer probes.

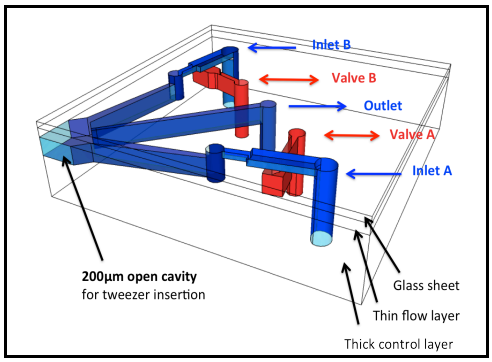


Figure 2: 3-D schematic of the open microfluidic device: Flow channels and the open reaction chamber (in blue) and control channels (in red).

At the cross-section, channels are 600-µm wide, making the active area 600 µm by 600 µm and determining the valve actuation pressure (100 kPa). When pressure is applied to the lower channel, the membrane deflects upward and closes the upper channel stopping the flow (Figure 3-b). Via was implemented to allow liquid transitions between the flow channels and the open chamber (Figure 3-f). Finally, combined to a convenient solution pressure (<10 kPa), the response time of the device is fast enough to precisely fill the reaction chamber making it suitable for the control of the biological solutions.

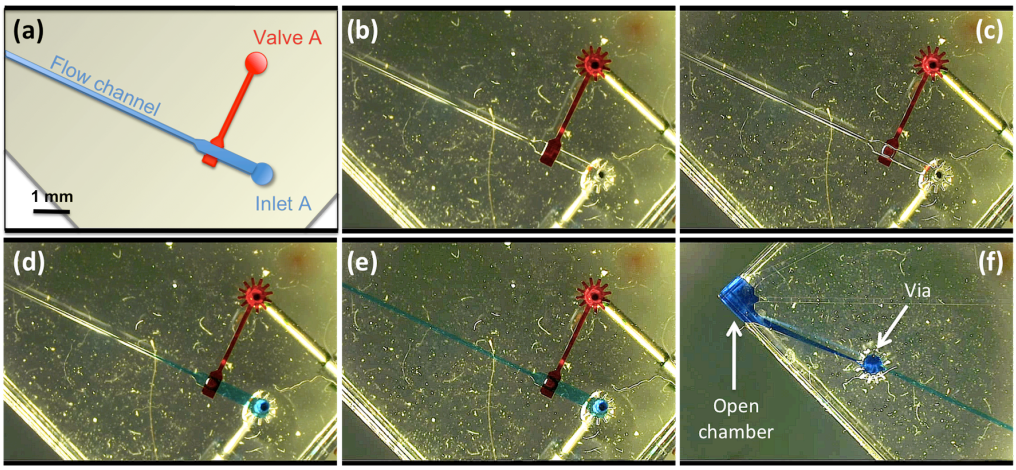


Figure 3: Video sequences of the controlled filling of the reaction cavity. (a) Red channel is the membrane-valve control while blue channel is the biological solution channel. (b) Red control channel is under pressure (100 kPa) closing the valve. (c) Control pressure is released; the valve is open. (d) The pressure released; blue "biological" solution flow crossed the valve. (e) The blue "biological" solution is reaching the reaction cavity. (f) The blue "biological" solution is properly filling the open reaction cavity.

EXPERIMENTAL

After have properly filled the reaction chamber, the probes of tweezers are introduced into from the open side (Figure 4a). An AC voltage (1 MHz, 20 Vpp) is applied between the tips, as by dielectrophoresis, DNA molecules elongate along the most intensive line of the resulting electric field [5]. During the trapping, the frequency response of the tweezers is continuously recorded following the phase rotation at the resonance frequency. At the end, a microscopic visualization confirmed that a DNA bundle is trapped between the two tips (Figure 4b).

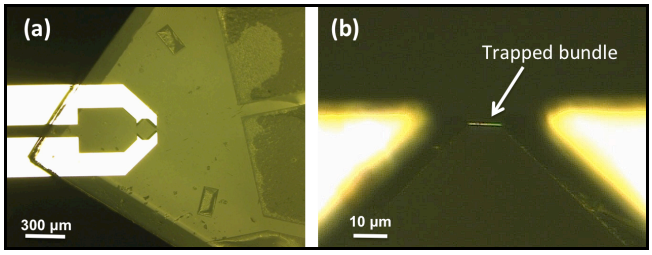


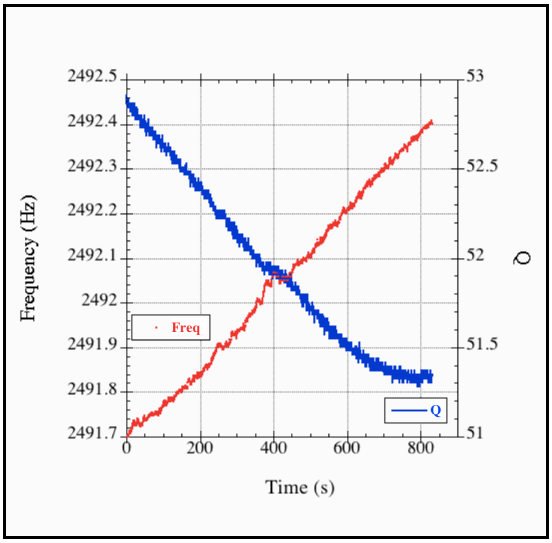
Figure 4: Tweezers in experience: (a) tweezers inserted in the open reaction chamber, (b) a trapped λ-DNA bundle bridges tweezer gap.

RESULTS AND DISCUSSION

As trapping proceeds, F increases due to the addition of DNA bundle stiffness k_{bundle} (Equation 1, Figure 5). At the same time, Q tends to decrease as the viscous losses v_{bundle} increases with the bundle formation (Equation 2, Figure 5). Precise frequency measurements allow the sensing of 5.10^{-3} Hz shift, around the typical resonance frequency (2.5 kHz), corresponding to 10 λ-DNA molecules stiffness. From the equations 1 and 2 and knowing the single molecule rigidity ($k_{\lambda\text{-DNA}}=3.10^{-5}$ N/m, [1]), the time evolutions of the bundle rigidity and viscosity can be deduced (Figure 6). Focusing on the first 300 seconds, the trapping rate was 0.9 molecule/second.

$$F = \frac{1}{2\pi}\sqrt{\frac{k_{TW} + k_{bundle}(t)}{M_{TW}}} \quad \ldots\ldots\ldots\ldots \quad (1)$$

$$Q = \frac{\sqrt{\left(k_{TW} + k_{bundle}(t)\right)M_{TW}}}{v_{TW} + v_{bundle}(t)} \quad \ldots\ldots\ldots\ldots \quad (2)$$



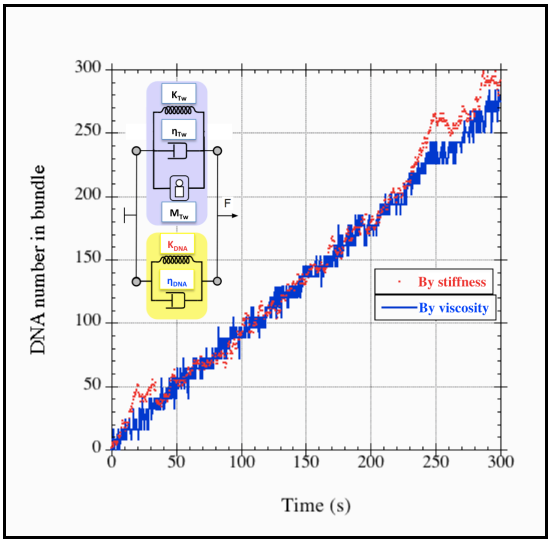


Figure 5: (On left) Tweezers + λ-DNA bundle resonance frequency and quality factor vs. DEP time.
Figure 6: (On right) Evolution of trapped λ-DNA molecules vs. DEP time. Number of trapped molecules is deduced from bundle rheological model and single molecule rigidity ($k_{\lambda\text{-DNA}}=3.10^{-5}$ N/m [1]).

CONCLUSION

The sensing of DNA molecule trapping with MEMS tweezers was demonstrated with high sensitivity. An active microfluidic device allowed dynamic control of the biological solution inlets. Moreover in order to sense slow enzyme kinetics with HindIII (until 1 hour), the device makes possible the compensation of drawbacks resulting from the solution evaporation. This new method shows the possibility to control the experiment conditions for accurate and systematic biological tests on filamentary molecules with the MEMS tool and electronic read-out.

REFERENCES

[1] C. Bustamante, Z. Bryant and S.B. Smith, Tens years of tension: single-molecule DNA mechanism, Nature, Vol. 421, pp. 423-427, (2003).

[2] G.J. Gemmen, R. Millin, and D.E. Smith, DNA looping by two-site restriction endonucleases: heterogeneous probability distributions for loop size and unbinding force", Nucleic Acids Research Vol. 34, pp. 2864-2877, (2006).

[3] D. Normanno, F. Vanzi, and F.S. Pavone, "Single-molecule manipulation reveals supercoiling-dependent modulation of lac repressor-mediated DNA looping", Nucleic Acids Research, Vol. 36, pp. 2505-2513, (2008).

[4] D. Anselmetti, F.W. Bartels, A. Becker, B. Decker, R. Eckel, M. McIntosh, J. Mattay, P. Plattner, R. Ros, C. Schäfer, and N. Sewald, "Reverse engineering of an affinity-switchable molecular interaction characterized by atomic force microscopy single-molecular force spectroscopy", Langmuir, Vol. 24, pp. 1365-1370, (2008).

[5] C. Yamahata, D. Collard, B. Legrand, T. Takekawa, M. Kumemura, G. Hashiguchi and H. Fujita, Silicon nanotweezers with subnanometer resolution for the micromanipulation of biomolecules, J. of Microelectromechanical Systems, Vol. 17, pp. 623-631, (2008).

[6] M. Kumemura, D. Collard, S. Yoshizawa, D. Fourmy, N. Lafitte, S. Takeuchi, T. Fujii, L. Jalabert, and H. Fujita, Direct bio-mechanical sensing of enzymatilc reaction on DNA by silicon nanotweezers. Proc. MEMS 2010, Hong-Kong, pp. 915-918, (2010).

[7] M.A. Unger, H.P. Chou, T. Thorsen, A. Scherer and S.R. Quake, Monolithic microfabricated valves and pumps by multilayer soft lithography, Science, Vol. 288, pp. 113-116, (2000).

CONTACT

*N. Lafitte, LIMMS-CNRS/IIS, UMI2820, Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, 153-8505, Tokyo, Japan, Tel: +81-3-5452-6088; Fax: +81-3-5452-6037; E-mail: lafitte@iis.u-tokyo.ac.jp

B Text processing into semi-structured file