Proteome Fingerprinting as a Localization Scheme for Nanobots

Regine Wendt, Florian-Lennert Lau, Lena Unger, Stefan Fischer
{regine.wendt,f.lau,l.unger,stefan.fischer}@uni-luebeck.de
Institute of Telematics, University of Lübeck
Lübeck, Germany

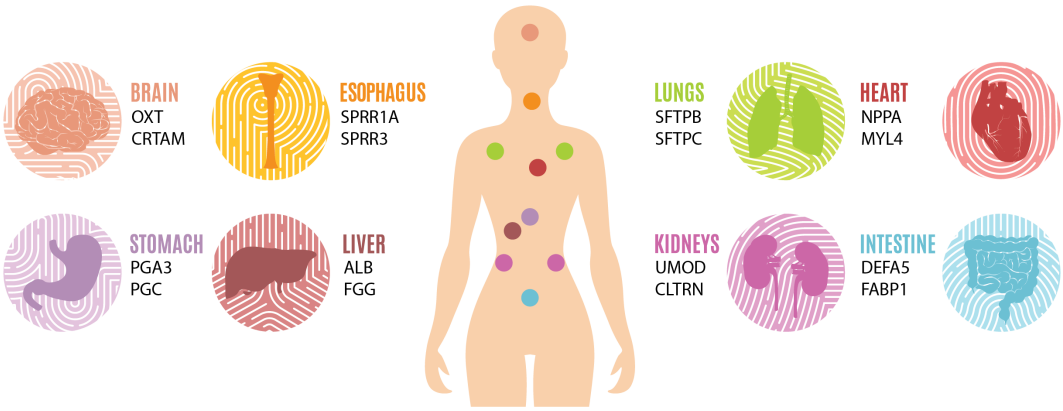


Figure 1: Proteome Fingerprint of eight major organs.

ABSTRACT

The localization of nanobots in the human body is a crucial element to enable diagnostic ability. Current localization schemes for nanobots primarily rely on mathematical principles, but our proposed approach offers a different perspective. In this paper, we present a completely novel idea to locate nanobots within the human body by employing local pattern recognition based on unique fingerprints. We thoroughly investigate and assess various substances in the vicinity of nanobots to develop distinctive fingerprints for all major tissues. Among the candidates, we identify the human proteome as the most suitable option due to its high tissue specificity. Through our research, we determine unique combinations of protein-coding genes, ensuring exclusive localization for each specific body region. Each tissue's optimal fingerprint consists of only two protein-coding genes, which do not intersect with other tissues, further guaranteeing accurate localization. We propose the detection of these fingerprints by using DNA-based nanonetworks, enabling targeted drug delivery and facilitating the precise localization of nanobots or their measurements within the human body.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
NANOCOM '23, September 20–22, 2023, Coventry, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0034-7/23/09…$15.00
https://doi.org/10.1145/3576781.3608728

CCS CONCEPTS

• **Applied computing** → *Life and medical sciences*; • **Computing methodologies** → *Model development and analysis.*

KEYWORDS

Nanonetworks, Medical application, Nano medicine

ACM Reference Format:
Regine Wendt, Florian-Lennert Lau, Lena Unger, Stefan Fischer. 2023. Proteome Fingerprinting as a Localization Scheme for Nanobots. In *The 10th ACM International Conference on Nanoscale Computing and Communication (NANOCOM '23), September 20–22, 2023, Coventry, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3576781.3608728

1 INTRODUCTION TO NANODEVICES IN MEDICAL APPLICATIONS

To explore the cause of physical discomfort, people often seek the assistance of a physician, consult information sources, or talk to friends and family. While successful treatment at a later stage is feasible for some diseases, certain disorders need to be detected as soon as possible. This early diagnosis could be challenging as the patient might not have any symptoms until later. People with chronic diseases are in even greater need of constant monitoring due to possible immunodeficiency. Regular examinations at the physician can reduce the risk, but even daily visits do not ensure continuous surveillance. Not to mention that daily medical appointments are difficult to reconcile with everyday life. The vision of nanonetworks is that *nanoscale devices* (nanodevices or nanobots), for example, patrol the body, take measurements wherever necessary, and send collected data to the outside. Therefore, e.g., a tumor can be detected before it starts metastasizing. These machines may even immediately work on problems they detect within the body, such

as cancer cells, arteriosclerosis, or *human immunodeficiency viruses* (HIV). To correctly detect abnormalities, it is essential to connect the measurements of the nanobots to the corresponding organs or tissues in the human body. That requires a reliable localization of the nanobots in question.

Current localization schemes make use of mathematical algorithms like *Function Centric Nano-Networking* (FCNN) or trilateration [8]. Using biological or chemical values is an entirely new approach outlining a promising possibility to differentiate particular body regions. Fingerprinting aims to determine the position of nanobots via local pattern recognition. An individual fingerprint is assigned to each body region by the properties of the environment in question. By identifying these fingerprints, nanobots can determine their position in the human body and communicate it to the outside.

2 FINGERPRINTING APPROACHES

To the best of our knowledge, no existing method assigns histological fingerprints to different body regions. Therefore, this paper explores various histological values in the human body to identify a reliable fingerprinting method. Promising candidates include systems encompassing the entire body, such as the lymphatic, endocrine, and cardiovascular systems.

The lymphatic system is unsuitable since nanobots are intended to primarily function in the blood circuit [3]. Thus, the values used for fingerprinting should be present in the cardiovascular system.

The endocrine system and hormonal balance offer several parameters that could serve as fingerprinting characteristics. Different hormones are produced in various glands and distributed throughout the endocrine system. However, this distribution makes it challenging to distinctly differentiate body regions, as certain hormones are not exclusive to specific regions.

Three methods exhibit favorable characteristics for fingerprinting and merit further analysis: blood gases, trace elements, and the human proteome.

2.1 Requirements

To be a reliable source for the localization of nanobots, the parameters used for fingerprinting have to fulfill specific requirements. For a distinct differentiation between body regions, either a singular compound only existing in specific regions or a compound with significantly distinguishable concentrations in different regions could be chosen. The existence and the concentration of the compound in question should be stable and not influenced by other factors, e.g., diseases or physical activity, for reliable and continuous monitoring of patients. To make reasonable differentiations inside the human body, research for a sufficient number of organs or tissues has to be available. Explicitly critical organs like the heart, lungs, liver, intestines, and kidneys should be covered.

2.2 Suitability of Fingerprinting Approaches

The initial compounds considered for fingerprinting are the gases in the cardiovascular system, measured through general blood gas analysis of O_2, CO_2, and pH levels. However, this method is unsuitable for the complete localization of body regions as it lacks clear differentiation. Additionally, physical activity can cause fluctuations in gas concentrations, leading to uncertainties in diagnostics. Thus, blood gases are excluded as fingerprinting compounds.

Another possibility is trace elements, such as Cobalt, Copper, Iodine, Iron, Manganese, Molybdenum, Selenium, and Zinc, occurring in minute concentrations [10]. Since the same trace elements are present in different tissues, the concentration of the elements must be considered for the fingerprinting. Trace elements for blood, packed blood cells, urinary excretion, lung tissue, liver tissue, kidney tissue and muscle tissue were compared in [10]. Four of the eight previously mentioned essential trace elements occur in all examined tissues: Copper, Manganese, Selenium, and Zinc. The concentration gradients in different tissues of those four elements can provide insight into the position in the human body. For the remaining elements, the presence in the respective tissues is the relevant factor. However, reported concentrations for trace elements in different human tissues vary significantly, and dietary insufficiencies can introduce further problems [10]. The variability and minuscule concentrations would decrease the reliability of the fingerprint and render the procedure ineffective.

Besides, the data from J. Versieck [10] on blood, packed blood cells, and muscle tissue cannot be used to identify different body regions. Only urinary excretion, lung tissue, liver tissue, and kidney tissue remain, but they are insufficient for valuable differentiation between body parts. Additionally, differentiating concentrations for various ages and sexes poses a problem for fingerprinting. Thus, we rule out trace elements as a candidate for fingerprinting.

The human proteome encompasses all the proteins present in the human body, and it originates from the genetic material encoded in the human genome. The *genotype* refers to the entire set of deoxyribonucleic acid (DNA) within a living organism, serving as the blueprint for all its proteins. However, these observable traits become visible only when the genetic information is interpreted, leading to the *phenotype*. The phenotype is determined by the synthesis of proteins that control the organism's structure and development or act as enzymes facilitating specific metabolic pathways.

Figure 2 illustrates the process of gene expression, where DNA leads to protein formation. In the first step, *transcription*, individual protein-coding genes (DNA) are transcribed into messenger ribonucleic acid (mRNA) molecules, collectively known as the transcriptome [1]. In the subsequent *translation* step, the mRNA sequence is used to assemble a polypeptide chain consisting of amino acids. Each amino acid is coded by three bases from the mRNA molecule. The resulting polypeptide chain then undergoes a series of folding steps, acquiring a three-dimensional structure, and ultimately forming a protein.

While an organism's genotype is generally uniform across all its cells, the same cannot be said for its phenotype [4]. Human tissues and cells exhibit varying gene expressions, resulting in specific proteomes for each cell type and tissue [9]. Analyzing gene expression patterns in different tissues enables the identification of their respective proteomes. The Human Proteom Atlas categorizes human genes based on their specificity to various organs and tissues throughout the body [2, 9]. This valuable information enables the creation of distinct proteome fingerprints for different body regions.

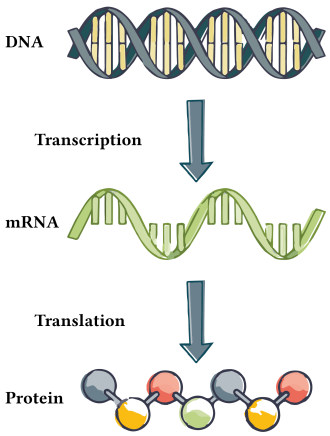
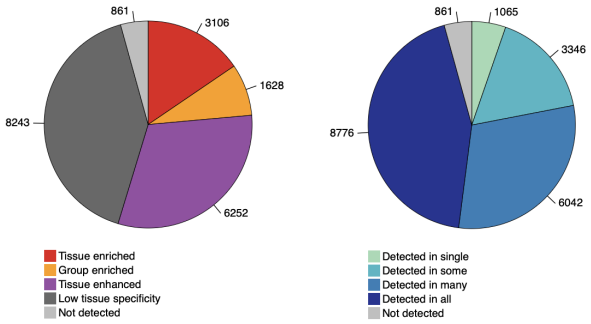


Figure 2: Gene Expression is the process by which a gene is transcribed into mRNA inside the nucleus of a cell. The mRNA then leaves the nucleus and gets translated into a protein.



(a) Specifity: The distribution of all genes across the five categories based on transcript specificity.

(b) Detection: The distribution of all genes across the five categories based on transcript detection (nTPM ≥ 1).

Figure 3: Specificity and distribution of transcribed mRNA molecules across all putative 20090 protein-coding genes in all 37 analyzed tissues. [2]

With the introduction of proteome fingerprints, we consider detecting gene products arising from gene expression, namely mRNA or proteins. In Section 5, we will explore how this detection can be achieved. For the sake of brevity in the remainder of the paper, we will only refer to genes while maintaining the underlying assumption that the detection pertains to the gene products.

3 THE HUMAN PROTEOME AS A FINGERPRINTING SCHEME

The distribution, gene expression rates, and tissue specificity are taken into account for the choice of protein-coding genes for the fingerprints. Uhlén et al. published a database covering more than 90% of the putative protein-coding genes available at the *Human Protein Atlas* (HPA) [2]. Performing immunohistochemistry on 44 human tissues complemented by RNA-sequencing on 37 of them, protein and mRNA expression data were derived [7].

From the approximately 20,000 human protein-coding genes, around 9,000 genes show low tissue specificity and represent the housekeeping proteome [2]. The other 11,000 genes show an elevated expression in particular tissues, subdivided into (i) tissue-enriched genes with mRNA levels in one tissue at least four-fold higher than the maximum of any other tissue, (ii) group-enriched genes with four-fold higher mRNA levels in a small group of tissues, and (iii) tissue enhanced genes with mRNA levels in a particular tissue at least four-fold higher than the average in all other tissues [2]. As for the distribution, genes can be (i) detected in a single tissue, (ii) detected in some tissues (more than one but less than one-third of all), (iii) detected in many tissues (at least one-third but not all), or (iv) detected in all tissues [2]. The distribution visualizes the number of genes showing detectable mRNA levels in different tissues. Figure 3 shows the specificity and distribution of all 20090 protein-coding genes. The detection of active genes is done by the normalization of transcriptomics data at hand from several different research groups [2]. A gene has detectable levels of mRNA molecules if the *normalized expression* (NX value) is above one [2].

3.1 Creating Fingerprints: Selection Algorithm

In the database, 37 tissues are represented. For now, we are focusing on searching for the fingerprints in the 18 tissues that are represented in our cardiovascular simulator BloodVoyagerS (BVS) [3]: adrenal gland, brain, esophagus, gallbladder, heart, intestine, kidney, liver, lung, pancreas, parathyroid gland, pituitary gland, retina, salivary gland, stomach, thyroid gland, tongue and urinary bladder. Genes that are not present in the respective tissues and genes with low tissue specificity are excluded from the sample space used to select genes for the fingerprints. Table 1 gives an overview of the occurrence of genes according to the HPA in the major organs.

Table 1: Occurrence of Genes in Major Organs.

Tissue (Number)	Detected	Elevated Genes	Detected in single	Detected in some
Brain (2)	16,465	2,587	33	685
Esophagus (3)	14,129	429	0	112
Heart (5)	14,409	387	3	123
Intestine (6)	15,609	764	14	337
Kidney (7)	14,823	413	8	182
Liver (8)	14,110	936	39	306
Lung (9)	15,021	239	1	79
Stomach (15)	14,707	159	1	91

The elevated genes can be subdivided into four groups regarding the distribution, as mentioned earlier. Genes only detected in a single tissue are advantageous to ensure as unique fingerprints as possible. Since not all organs feature one gene or more from this category, genes detected in some tissues are also considered. Table 1 shows the fraction of the elevated genes, as well as the number of elevated genes that are detected in a single tissue and some tissues.

A comparison of the gene data of all relevant tissues showed that none of the tissues have a complete overlap of elevated genes. For reliable localization, the number of genes should be as high as necessary but as low as possible to be still dependable and to get detection rates in reasonable time frames. We started the analysis with five genes for each fingerprint and then reduced the number of genes step by step.

The key element for choosing genes for the fingerprints is the expression level, represented in the NX value. Genes with an NX value of one and higher count as detectable. The higher the NX value, the higher the mRNA level of a gene is in the respective tissue. A *tissue specificity score* (TSS) compares the tissue with the highest mRNA level of a gene to the tissue with the second-highest mRNA level. Therefore the TSS gives an idea of how exclusively high the mRNA level is. However, for genes that are detected not only in a single tissue but in some tissues, this value provides information for the group of tissues. As we aim for the detection of single tissues we sort the remaining genes suitable for the fingerprints for their NX values. As a result, some genes that were detected in single tissue but have a low NX value are not eligible for the fingerprint. This will increase the likeliness of fast detection and reduce the likeliness of false positives when genes are present in more than one tissue. The genes with the highest scores are subsequently chosen for the fingerprints. We implemented a Matlab code to read the gene data files and find the best-suited fingerprints for different scenarios. Table 2 shows the gene combination for the 5-fingerprint of the heart with the corresponding distributions, NX value and TSS.

Table 2: Fingerprint of the Heart using 5 genes.

Gene	Tissue Distribution	NX	TSS
NPPA	Detected in some	1267.7	418
MYL4	Detected in some	712.3	68
TNNT2	Detected in some	679.2	241
MYL2	Detected in some	656.7	11
MYH7	Detected in some	644.7	6

4 EVALUATION OF DIFFERENT FINGERPRINT SIZES

To guarantee unique fingerprints, we compared the genes chosen for the fingerprints. When using five genes, there is a maximum overlap of one gene between two fingerprints, ensuring unique gene combinations in general. Now, we want to determine what an ideal number of involved genes is. The fingerprint should be distinctive but as small as possible since the combined probability of detection decreases with each participant. To demonstrate this and find an optimum, we consider four different metrics: dominance, risk, detection potential, and confidence.

Definition 1. A *fingerprint (F)* is a set of n genes that act as a unique identifier of a tissue when detected by a nanobot. *F(i)* is the *i'th* gene, that constitutes the fingerprint.

As a metric to find the best-suited genes for the fingerprint, we considered the NX value. The higher the NX value, the higher the mRNA level of a gene is in the respective tissue. This also means, the higher the NX value is, the more likely the gene gets detected by our nanobot. We, therefore, use NX to search for dominant fingerprint combinations.

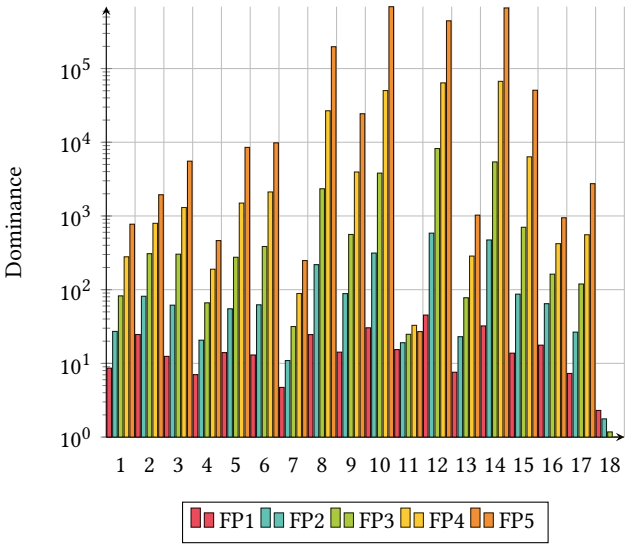


Figure 4: Dominance per Tissue on a logarithmic scale for 1-5-gene-fingerprints.

Definition 2. $NX_{F(i)}$ is the NX value of *F(i)*.

Definition 3. NX_T is the sum of all genes present in *tissue (T)*. $\overline{NX_T}$ is the mean NX value of *tissue (T)*. All detectable genes are taken into account.

The mean NX value of a tissue gives information about how likely it is to detect a gene of this tissue in general. We then aim for a fingerprint with a dominance over one, so that it has an above-average occurrence in said tissue. We multiply the NX values of the participants so that if a fingerprint gene occurs less frequently in the tissue than the average, the overall dominance value decreases. To achieve comparability for different sample sizes, we divide by the number of fingerprint participants.

$$\text{Dominance}_F = \frac{\prod_{i=1}^{n} \frac{\text{NX}_{F(i)}}{\overline{\text{NX}_{T_F}}}}{n} \tag{1}$$

Figure 4 shows the dominance per tissue on a logarithmic scale and for different fingerprint sizes. For most tissues, the dominance increases with more genes used, meaning all included genes have an NX over the tissue average. For tissue 11, the parathyroid gland, the dominance stays nearly the same. Tissue 18, the urinary bladder however even loses dominance with bigger fingerprints. The 4- and 5-fingerprint, have dominance values smaller than one, meaning the fingerprint combination is not very dominant in the tissue. It is less likely to find the fingerprint in that tissue, than a random combination of two genes of that tissue. This is one reason to aim at reducing the fingerprint size. The second reason gets clear when looking at the detection potential.

Since the fingerprints must be detected simultaneously, we need to consider not only their occurrence relative to the average but

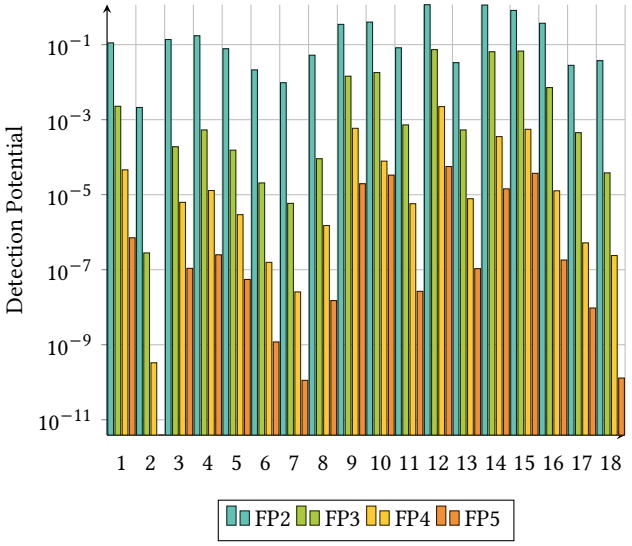


Figure 5: Detection Potential in percent.

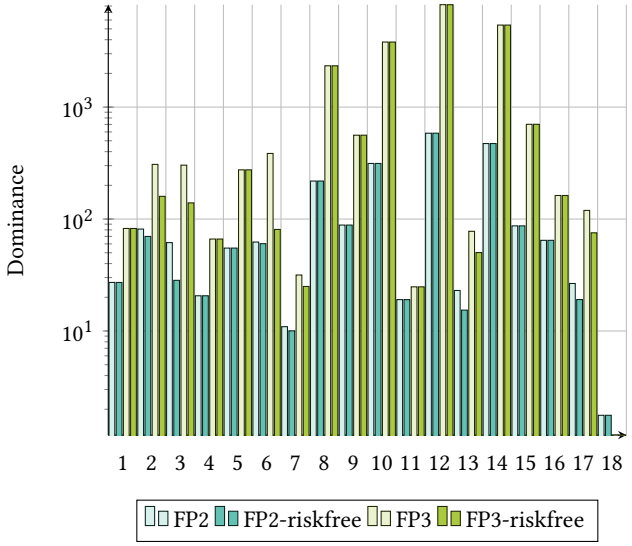


Figure 6: Dominance of 2- and 3-fingerprints before and after risk reduction.

also their occurrence relative to the total volume. This is reflected in the detection potential.

$$\text{Potential}_F = \prod_{i=1}^{n} \frac{\text{NX}_{F(i)}}{\text{NX}T_F} \tag{2}$$

Figure 5 shows the detection potential per tissue on a logarithmic scale and for different fingerprint sizes. Now, one can see that the likelihood of finding larger fingerprints is significantly lower. With the addition of each extra gene, the probability, on average, decreases by two decimal places.

While the dominance indicates the sensitivity of a fingerprint and the potential illustrates the detection probabilities we also need to guarantee that there is no risk of a false positive. We aim for a risk of zero, which means that a combination of the used fingerprint is impossible to find in any other tissue. When we look at Table 1, for the tissues with genes available that are detected in single, one fingerprint gene would be enough for zero risk. However, if the NX of said genes is low, they are still not as suitable as the combination of two (or more) not single but in their combination unique genes. In the fingerprint selection algorithm as described in Section 3.1. there are non unique fingerprints as a result for smaller fingerprint numbers. For the 1-fingerprints there are six tissues with a risk of false identification. For the 2- and 3-fingerprints there are still two fingerprints, that are completely found in another tissue. Since we want to reduce the genes used for the fingerprint to increase detection probabilities we have to reduce the risk in the smaller fingerprints to zero as well.

We improve our fingerprint selection algorithm so that if there is a risk higher than zero, the gene of the fingerprint gets replaced with the next gene in line and the risk gets reevaluated until it is zero for each tissue. To increase the level of uniqueness, the genes of fingerprints with partly overlap get replaced too. The smaller the number of genes used for a fingerprint, the more important the risk reduction feature gets.

Figure 6 shows the dominance compared of fingerprints with two and three genes before and after risk reduction. One can note that risk reduction has no significant influence on dominance. It is therefore a reasonable way to improve specificity with no effect on sensitivity.

To make a conclusive decision on the optimal fingerprint size, we consider both dominance and detection potential of the risk reduced fingerprints together and refer to it as Confidence_F

$$\text{Confidence}_F = \text{Dominance}_F * \text{Potential}_F \tag{3}$$

As shown in Figure 7, it becomes evident that, for all tissues, the fingerprint size of 2 outperforms all others. We do not consider fingerprints of size 1, as they do not allow for a risk reduction to zero. 2-fingerprints can clearly identify the tissue, have a relatively high probability of detection, and are therefore preferable to all other combinations.

5 DETECTION OF FINGERPRINTS

Now, we need a concept to reliably and simultaneously detect the 2 genes of the identified fingerprints. One promising approach to solve this problem is DNA-based nanonetworks [5]. For more information on the underlying principles and definitions, please consult this paper, as they exceed the scope of this publication.

These networks are capable of detecting a multitude of predetermined DNA or RNA sequences. The detected sequences may then be used as inputs to perform the computation of for example an n-bit logical AND operation. In doing so, the DNA-based nanonetwork may ensure that a number of previously specified RNA sequences must be present for a computation to evaluate to "True". It is also possible to compute a threshold operation instead, as presented in [6].

Figure 8 shows an example assembly of DNA molecules called a *message molecule* that computes the aforementioned logical AND

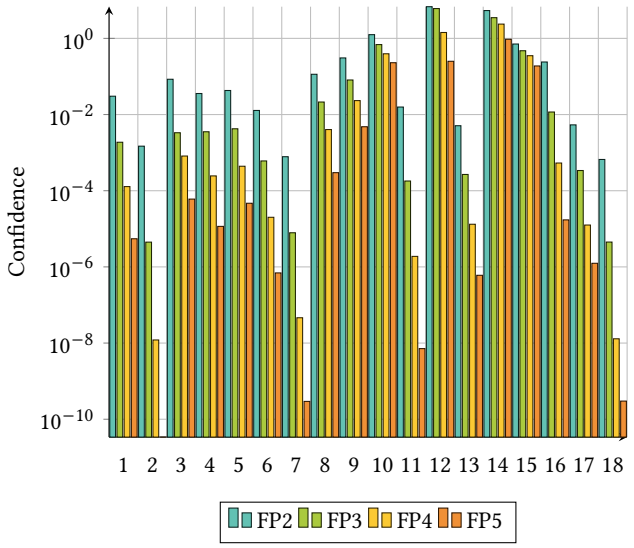


Figure 7: Confidence to find the fingerprint in a reasonable amount of time and with zero risk of false positives.

on two inputs. The message molecule may only completely self-assemble given the presence of the tiles M_1 and M_2. Each of those tiles indicates the presence of a specific RNA sequence and may be conditionally released by nanosensors.

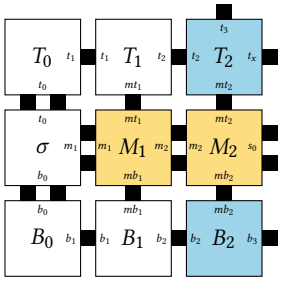


Figure 8: A fully assembled, message molecule that represents a fingerprint. Blue tiles represent the ligands and yellow tiles show the tiles necessary for fingerprinting a tissue type. For a detailed explanation of the tiles see [5].

Once the message molecule is fully self-assembled, it is clear that we identified the tissue we were looking for. The message molecule itself may then be detected by other devices via DNA or RNA bindings. In the use case of targeted drug delivery said message molecule could also be the key opening a drug-loaded DNA tile box, which then would release its payload only in the target tissue.

6 CONCLUSION AND OUTLOOK

We proposed a new approach to locating nanobots in the human body based on local pattern recognition, specifically through the identification of unique fingerprints. We investigated and evaluated different substances in the environment of the nanobots to develop these fingerprints. We identified the human proteome as the most suitable candidate as it shows high tissue specificity. We determined unique combinations, accounting for only one body region at a time. For each tissue we found an optimal fingerprint consisting of two protein-coding genes, showing no intersections with other tissues, ensuring an exclusive localization for each tissue. The detection of these fingerprints can be realized with DNA-based nanonetworks and may be used for targeted drug delivery and as a method of localizing nanobots or their measurements in the human body. The message molecules could also be adapted to require higher concentrations of certain proteins or to allow for bigger as well as smaller individual fingerprints for difficult/simple-to-classify organs. In the future, one could further examine the concentration gradients of the fingerprints, therefore determining the position even over some distance.

The correct binding and measurement of the fingerprints in general rely on the chemical pull between the nanobot and genes. Further research on the biological structure and binding affinity of the fingerprints depending on nanobot type and detection method is necessary. A simulation framework for fingerprint detection must be developed to test different scenarios. Integrating localization into existing simulations, such as BVS [3], would be valuable in determining the practicality of distribution and binding parameters for a fingerprint detection scheme in the blood flow. Proteome fingerprinting is a completely new and promising approach for localization and can be helpful in many other use cases as we will argue in the feature.

7 ACKNOWLEDGMENTS

This work has been supported in part by the German Research Foundation (DFG): Project 419981515, NaBoCom II.

REFERENCES

[1] Terence A. Brown. 2002. *Genomes.* Wiley-Liss, Chapter 3: Transcriptomes and Proteomes. http://www.ncbi.nlm.nih.gov/books/NBK21128/
[2] The Human Protein Atlas Consortium. 2003. *The Human Protein Atlas.* Retrieved april 7, 2023 from https://www.proteinatlas.org
[3] Regine Geyer, Marc Stelzner, Florian Büther, and Sebastian Ebers. 2018. BloodVoyagerS: simulation of the work environment of medical nanobots. In *Proceedings of the 5th ACM International Conference on Nanoscale Computing and Communication.* ACM, 1–6. https://doi.org/10.1145/3233188.3233196
[4] Sul JY, Wu CW, Zeng F, Jochems J, Lee MT, Kim TK, Peritz T, Buckley P, Cappelleri DJ, Maronski M, Kim M, Kumar V, Meaney D, Kim J, and Eberwine J. 2009. Transcriptome transfer produces a predictable cellular phenotype. *Proc Natl Acad Sci U S A* (May 5 2009).
[5] Florian-Lennert Lau, Florian Büther, Regine Geyer, and Stefan Fischer. 2019. Computation of decision problems within messages in DNA-tile-based molecular nanonetworks. *Nano Communication Networks* 21 (Sept. 2019).
[6] Florian-Lennert Adrian Lau, Regine Wendt, and Stefan Fischer. 2021. Efficient in-message computation of prevalent mathematical operations in DNA-based nanonetworks. *Nano Communication Networks* 28 (June 2021), 100348. https://doi.org/10.1016/j.nancom.2021.100348
[7] M. Uhlén et al. 2015. Tissue-based map of the human proteome. *Science* 347 (Jan. 2015), 1260419. Issue 6220. https://doi.org/10.1126/science.1260419
[8] Marc Stelzner, Falko Dressler, and Stefan Fischer. 2017. Function Centric Nano-Networking: Addressing Nano Machines in a Medical Application Scenario. *Nano Communication Networks* 14 (Dec. 2017), 29–39. https://doi.org/10.1016/j.nancom.2017.09.001
[9] Peter J. Thul and Cecilia Lindskog. 2018. The human protein atlas: A spatial map of the human proteome. *Protein Science: A Publication of the Protein Society* 27 (Jan. 2018), 233–244. Issue 1. https://doi.org/10.1002/pro.3307
[10] Jacques Versieck. 1985. Trace elements in human body fluids and tissues. *Critical reviews in clinical laboratory sciences* 22 (1985), 97–184. Issue 2. https://doi.org/10.3109/10408368509165788