

# Research Statement

**Haris Vikalo, Ph.D.**

Associate Scientist

California Institute of Technology  
1200 E. California Blvd, MC 136-93  
Pasadena, CA 91125, USA

hvikalo@caltech.edu

My current research is in the general area of biosensors and genomic signal processing. I am particularly interested in the development and applications of new approaches to the data acquisition and information processing in biosensors, the modeling of biosensors, and understanding fundamental limits of their performance. The ultimate technological goal of my research is the design of high performance sensors for biomedical applications such as molecular diagnostics, pathogen detection, and high-throughput screening. The interdisciplinary nature of the field poses many challenges and requires understanding of molecular biology, physical chemistry, and statistical signal processing. It also presents a tremendous research opportunity due to the growing demands for improvement of the performance and increase of the throughput of the existing technologies, as well as due to the emergence of new applications.

High-throughput and high performance biosensors have become essential research tools in genomics, and are a rapidly growing segment of the life sciences industry. Among the most prominent such technologies are DNA microarrays and quantitative (real-time) polymerase chain reaction (QPCR) systems. DNA microarrays have attracted much interest due to the large scale, parallel nature of their experiments, and the abundance of the information that they provide in high-throughput screening applications. Data acquisition in DNA microarrays is based on hybridization, the process in which complementary single strands of nucleotides bind to each other. Unfortunately, the probabilistic nature of the hybridization, as well as the inherent randomness of the preparation and conducting of microarray experiments, causes a high level of measurement uncertainties. This in turn makes the reliability of the information inferred from the microarray experiments suspect. On the other hand, QPCR is a technology which potentially enables precise quantification of a single DNA target, an important task in applications such as molecular diagnostics. However, simultaneous quantification of multiple targets of interest appears to be rather challenging.

My previous studies of microarray and QPCR systems, both theoretical and experimental in nature, include probabilistic modeling and optimal estimation for DNA microarrays; optimal estimation of the initial number of molecules in QPCR based DNA amplification schemes; and finding explicit analytical (so-called Cramer-Rao) bounds on the mean-square errors of the optimal estimation algorithms in both microarray- and QPCR-based schemes. A major result of our work is the development of the *real-time microarray* (RT- $\mu$ Array) platform, a fluorescent-based biosensor assay capable of quantifying kinetics of the hybridization process. This stands in contrast to the conventional microarrays, where the single measurement is taken only after the hybridization reached a steady-state. By processing the large amount of acquired data, the RT- $\mu$ Array systems achieve higher signal-to-noise ratio, smaller estimation error, and broader detection dynamic range compared to the conventional microarrays. Moreover, the RT- $\mu$ Array systems may enable implementation of the multiplex QPCR in a single well and become an efficient alternative to the multiplex QPCR which is facilitated in multiple wells and requires microfluidic solutions. Finally, it is important to note that the RT- $\mu$ Array data acquisition principles extend to other types of arrays, e.g., protein arrays. In what follows, I will give an outline of the future research projects.

## 1 Real-time microarrays

Recently, we have developed a novel methodology for detection of targets in microarrays (and other affinity-based biosensors), which we refer to as the real-time microarray (RT- $\mu$ Array) system. Here I give a brief background on the RT- $\mu$ Array systems, and outline future projects.

A DNA microarray is an affinity-based biosensor where the binding is based on hybridization, a chemical processes in which single DNA strands specifically bind to each other creating structures in a lower energy state. In a fluorescent-based DNA microarray, the measured signal emanates from the fluorescently labeled target molecules which have hybridized to the probes at the surface of the microarray. Typically, the detection of the captured targets is carried out by scanning and/or various other imaging techniques *after* the hybridization step is completed. The reason for this is simple: a large concentration of floating (unbounded) labeled targets in the hybridization solution may overwhelm the specific signal emanating from the captured targets. Hence, the conventional microarrays typically do not allow the presence of the solution during the fluorescent and reporter intensity measurements. So, after the hybridization, the solution is washed away from the array surface; this often causes washing artifacts that can make the data analysis difficult.

On the other hand, the RT- $\mu$ Array system evaluates the abundance of multiple targets in a sample by performing the real-time detection of the target-probe binding events. This system samples fluorescent signals emanating from the probes capturing quencher-labeled targets in the solution and thus does not require any washing step. The RT- $\mu$ Array systems may employ various time averaging schemes to suppress the Poisson noise and fluctuation of the target bindings. Due to all these advantages, the RT- $\mu$ Array systems achieve higher signal-to-noise ratio, smaller estimation error, and broader detection dynamic range compared to the conventional microarrays. Furthermore, unlike the typical conventional microarrays, the RT- $\mu$ Array systems can measure the probe density variations prior to hybridization, thus providing much needed quality control.

The RT- $\mu$ Array systems require further developments and refinements. Specific future projects include the development of optimal algorithms for detection and suppression of cross-hybridization, finding the limits of the RT- $\mu$ Array performance, and performing additional assay optimization.

The paradigm shift in data acquisition, from measuring single steady-state data point in the conventional microarrays to obtaining full hybridization kinetics in the RT- $\mu$ Array systems, requires the development of novel detection algorithms. Quantification of targets in the RT- $\mu$ Array systems can be performed by means of estimating the parameters of the hybridization kinetics (in particular, the binding rate). Since we need not wait that hybridization enters steady-state, the target quantification can be performed potentially much faster than in the conventional microarrays; for the same reason, saturation (the scenario where the number of target molecules is much higher than the number of probe molecules) becomes a non-issue in the RT- $\mu$ Array systems. On another note, in all affinity-based biosensors, there will always be a number of cross-hybridizations (non-specific bindings) of targets to their partial complements on the array. Moreover, there is an inherent randomness of the biochemical processes in microarrays which is inevitable since it originates from the probabilistic and quantum mechanical nature of molecular interactions present therein. This randomness is manifested by the probabilistic nature of both hybridization and cross-hybridization. Therefore, to find the optimal algorithms for detection of targets and suppression of cross-hybridization, we clearly need well-developed statistical models of the RT- $\mu$ Array systems. Finally, to understand their capabilities, we need to investigate the limits of the performance of the RT- $\mu$ Array systems, e.g., find the bounds akin to the Cramer-Rao bounds on the minimum mean-square error of estimation algorithms that we earlier computed for conventional microarrays. Such results will likely have implications on the design of the RT- $\mu$ Array systems.

## 2 Performance and throughput: A fundamental tradeoff in biosensors

Microarray systems are capable of detecting and quantifying many targets in a single experiment. However, their performance is affected by the inherent biochemical noise, and their precision and dynamic range are ultimately limited by the interference (i.e., the cross-hybridization between non-specific target-probe pairs). On the other hand, the quantitative (real-time) polymerase chain reaction (QPCR) is a technology which enables precise quantification of a single DNA target, an important task in applications such as molecular diagnostics. However, simultaneous quantification of multiple targets of interest appears to be rather challenging. A promising technology that may enable the multiplex QPCR requires use of microfluidic devices which distribute the biological sample of interest into separate wells so that an independent QPCR may be performed in each of those wells. However, beside the expensive equipment, this technology may require complicated experiment design; furthermore, because it needs to be divided among many wells, the biological sample clearly needs to be quite larger than in the single-well QPCR.

On the other hand, the RT- $\mu$ Array systems may enable implementation of multiplex QPCR in a single well and become an alternative to the microfluidic-aided multiplex QPCR. This project requires long-term commitment and involves the assay design and optimization; statistical modeling of the novel platform; development of the optimal algorithms for detection of multiple targets; and finding the limits of performance of the proposed system. The new system will rely on the property that the biochemical reactions in the RT- $\mu$ Arrays can be temperature-controlled and thus are reversible; therefore, the polymerase chain reaction can be implemented on the RT- $\mu$ Array. Essentially, the burden of resolving multiple amplicons can be migrated from the microfluidic devices to the signal processing algorithms. By combining the microarray and QPCR technologies, we would build a system that allows for an explicit tradeoff between performance and throughput.

## 3 Applications to system biology, drug discovery, and molecular diagnostics

Microarrays are one of the essential tools in the system biology research. As an example, the data acquired in microarray experiments is often grouped in clusters in order to explore the correlation between various genes and thus help deduce the structure of genetic regulatory networks. Improving performance of microarray systems clearly has positive impact on the accuracy of the deduced gene network models. Such an improvement of the model accuracy can be obtained by the RT- $\mu$ Array systems, and I would like to establish collaborative efforts exploring this topic. On a related note, the microarray statistical models which we developed may help lead to novel clustering strategies.

Study of the gene regulatory networks is a promising approach to speeding up and decreasing the cost of drug discovery. There, it is important to discover the biomarkers relevant to the studied disease. The RT- $\mu$ Arrays are well suited for studies of a moderate number of genes, which is a required step in the process of narrowing down the genes that reveal some information about the disease. Of course, understanding the gene networks is only the first step towards controlling them.

On another note, the multiplex QPCR enabled by the RT- $\mu$ Array platform is well suited for diagnostics applications. The objective there is to achieve robust and accurate detection platform, while minimizing false positive/false negative errors. In parallel, I intend to continue the work on analytically finding the limits of performance of the QPCR systems, thus helping us set realistic expectations from these devices in various applications.

Finally, since the biochemical processes observed by RT- $\mu$ Arrays are temperature-regulated and reversible, we can use feedback to establish a closed-loop control system. I believe that this may lead to the use of RT- $\mu$ Arrays in other applications, e.g., the *in vitro* studies of biochemical circuits.

## 4 Prior work: Algorithms, signal processing, and communications

My graduate school research was in the general area of signal processing and communications. It was focused on the development and analysis of algorithms for solving global optimization problems in the statistical setting commonly encountered in various communication problems. Solving the global optimization problems of combinatorial nature often requires an exhaustive search over the space of possible solutions. However, in a randomized setting, it may be possible to exploit statistics of the problem in order to design efficient algorithms and to analyze the complexity of such algorithms. This is the case for the integer least-squares problem in the communication context, which geometrically corresponds to finding the closest lattice point to a given point.

The most meaningful statistical measure of the performance of a communication system is the probability of transmission error which, for equally likely codewords, is minimized by the maximum-likelihood (ML) decoding. Hence, ML decoding naturally arises as the target criterion for the decoder design. However, it was quickly realized that ML decoding is often computationally infeasible in practice. In many communication systems, ML decoding reduces to an integer least-squares problem of finding the solution to the system of linear equations where the unknown vector is comprised of integers, while the given vector and coefficient matrix are comprised of real entries. Geometrically, this can be interpreted as a search for the closest point in a given lattice. The closest point search in a lattice is well-known to be an NP-hard problem and thus designers often turn to suboptimal criteria with computationally efficient solutions. Indeed, complexity of the optimal signal processing in high-dimensional communication systems (e.g., the multi-antenna systems) is a main obstacle to delivering the large data-rates that are promised by theory.

While the former is all true for the general integer least-squares problems, our main observation has been that in communications applications, the given vector is not arbitrary. Instead, it is an unknown point in a finite lattice that has been perturbed by an additive noise vector of known statistical properties. In communication problems both the noise and the channel-generated lattice are random. This leads to a probabilistic, rather than the more traditional deterministic, notion of complexity. Thus, a natural question to ask is what are the statistics of the algorithms that solve integer least-squares problems. A major result of my thesis answers this question for the so-called sphere decoding algorithm, for which we found closed-form analytic expressions for the expected complexity and complexity variance. We showed that for a wide range of signal-to-noise ratios, the expected complexity of sphere decoding is comparable to the complexity of the best heuristics, and in fact often sub-cubic. This suggests that ML detection, which was previously thought to be computationally intractable, can in fact be implemented with complexity similar to that of heuristic methods, but with significant performance gains – a result with many practical implications.

I extended this work in several important directions. Examples include the development of near-capacity multi-antenna systems with efficient soft detection schemes which are based on the sphere decoding idea; the development and analysis of algorithms for the joint ML detection and decoding of linear block codes; combining the sphere constrained search idea of sphere decoding with the dynamic programming idea of the Viterbi algorithm to develop an efficient algorithm for the ML detection on channels with memory; and the development and analysis of algorithms for the nearest codeword search in a Hamming distance sense, i.e., for efficient decoding of error-correcting codes beyond the minimum distance. Moreover, using convex optimization techniques, we developed several computationally efficient modifications of the original sphere decoding algorithm.

An important message of this work is that, for problems where there is an underlying statistical model, the complexity of any algorithm is best viewed as a random variable. Therefore, this approach could prove valuable in studies of other problems and algorithms in stochastic settings.