# Microfluidic Large-Scale Integration and its Application to Systems Biology

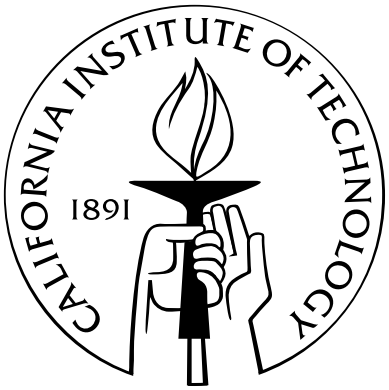
Thesis by

Sebastian Josef Maerkl

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy




California Institute of Technology

Pasadena, California

2007

(Defended June 1, 2007)

# Acknowledgements

Many people contributed to my professional career and personal development over the years. As it is next to impossible to pinpoint any specific moment in time when contributions become significant I might as well start at the beginning. Here I must obviously thank my parents, Robert and Anneliese Maerkl, who made me (to be taken primarily in the platonic sense). I can't thank them enough in words, but I hope that the person I have become lets them know that they were, and still are, great parents. Obviously my three siblings, Lisa, Thomas, and Matthias must not be forgotten either.

Academically, Prof. Quake undoubtedly had the biggest influence on my development as a scientist. Being in his lab for 6 years was an invaluable experience, which I thoroughly enjoyed. Steve allowed me to develop my own research agendum and pursue it independently, preparing me for a career in academia. Another person, with whom I had the pleasure to interact many times over the years, and who was extremely helpful in a wide range of topics, is Prof. Arnold. Prof. Wold got me started on transcription factors, a class of proteins which opened my eyes to a large and interesting field of biology. Caltech as an institution deserves thanks, not only because of all the great people I met during my tenure there (for the great people,

see below and above), but also because it provided endless opportunities to grow as a scientist. Professors that I fondly remember from before my time at Caltech are Prof. Hixenbaugh, Prof. Carlin, Prof. Middleton, and Prof. Perry of the biology department and Prof. Baylouny, Prof. Strange, Prof. Boyer, and Prof. Fordham of the chemistry department at Fairleigh Dickinson University. They fostered a terrific learning environment and were always personally available for the many inquiries I subjected them to.

Spending six years in a confined environment (whether this is deemed to be the laboratory or Caltech, I leave to the reader) certainly provided ample opportunities for making friends. I am glad to have had lab mates such as Carl Hansen, Megan Anderson, Rafael Sjoberg, Heun Jin Lee, Robert Bao, Jerrod Schwartz, Josh Marcus, and Alejandra Torres (by courtesy), who over the many years have not only been fantastic people to work with, but have become fast friends to me. Interestingly these folks can be neatly subdivided into two groups according to the kind of beverage that is consumed during meetings outside of lab. Other friends I made while at Caltech include Peter Meinhold, who also happened to be my next door neighbor together with Carl. Aside from spending a summer on a fantastic road trip, and several other trips, we also spent quite some time having gemuetliche Biergaerten. Jean Huang I would like to thank for essentially getting me started with rock-climbing as well as always giving me an excuse to leave lab to converse with a friend. Last but not least I would like to thank Karen Kapur, who over the last two years has brightened up my days!

# Abstract

The goal of biology is to understand how complex systems such as cells and entire organisms function. Systems Biology attempts to quantitatively characterize all components comprising these systems. A considerable task. Microfluidics provides a powerful tool for undertaking this endeavor. This thesis describes the development of Microfluidic Large-Scale Integration (MLSI) using devices fabricated by Multilayer Soft Lithography (MSL). MLSI and fluidic components, such as multiplexers and free-standing membranes, serve as the infrastructure for performing large-scale biophysical measurements of biological systems. Transcription factor binding energy landscapes were determined using MLSI and MITOMI, a novel method for measuring molecular interactions. The biophysical characterization of transcription factors described herein were the first comprehensive measurements of their kind, and answered basic questions regarding how transcription factors recognize DNA. Furthermore, it was possible to predict the *in vivo* function of transcription factors using only the measured binding topographies and a genome sequence, indicating that biological processes can be predicted with high accuracy. More generally, the methods described in this thesis are generally applicable to understanding the properties of any biological system and should find broad usage in the field of Systems Biology.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Overview

Biology has progressively moved away from characterizing isolated systems and moved towards understanding systems in their entirety, giving rise to the field of Systems Biology. Just as with Genomics and Proteomics before, Systems Biology is primarily a technology-limited field. Pertinent questions to ask are relatively easy to identify, but techniques applicable to these problems generally do not exist and must be specifically designed.

Microfluidics presents an extremely useful technique for asking large-scale questions, because it allows assays to be both scaled down, saving precious biological samples, and highly integrated. Microfluidic Large-Scale Integration (MLSI) was developed to achieve this scaling and parallelization, taking the standard micromechanical valve fabricated by Multilayer Soft Lithography (MSL) and integrating it several thousands of times, generating microfluidic devices that can perform large numbers of complex biological assays.

Besides MLSI, additional components needed to be developed, such as fluid control elements and novel detection methods. These components are described in detail in Chapter 2. Section 2.3.1 describes the design and application of a microfluidic

multiplexer, which allows the addressing of an exponentially large number of fluidic channels with a small number of control elements. Use of a multiplexer is essential to reducing the number of interconnects between the microfluidic device and the lab. The optimal design of a multiplexer was also discovered, which, unlike the commonly used binary design, is indeed a ternary one. A fluidic input tree is described in Section 2.3.2, which allows the introduction of a large number of fluids onto a device without cross-contamination. A third broadly applicable component, or design feature, described in Section 2.3.3 involves the use of a control line cascade, allowing multiple levels of control lines to be staggered. Cascading of control lines achieves even higher-degrees of integration and complexity than would otherwise be possible; cascading can also be used to change the logic functions of existing components such as the multiplexer. A final component, described in Section 2.3.4, is a free-standing membrane, which, unlike a standard valve, does not shunt liquid but rather contacts the flow channel surface, physically protecting it from the surrounding solution. This physical surface contact can be used for surface derivatization as well as detection of molecular interactions. Most of these fluidic components were integrated into devices designed for the high-throughput screening of enzyme libraries generated by directed evolution. These devices and methods are described in Chapter 3.

To study problems in Systems Biology it was not enough to only develop novel microfluidic components; the biology needed to be addressed as well. *In vitro* analysis of biological systems always requires the acquisition of the components to be studied, most commonly proteins and DNA. DNA is relatively easy to obtain as it

can be readily synthesized. Proteins, on the other hand, are difficult to produce and may not always be in a correctly folded state. In Systems Biology obtaining protein is particularly problematic, as not only one or a small number of proteins are needed, but often entire proteomes with thousands of members. Chapter 4 addresses this issue and describes the development of a rapid PCR-based approach for generating linear expression templates for use in *in vitro* transcription/translation systems (ITT). ITT together with the PCR-based approach for template generation can in principle rapidly generate large-scale, functional protein libraries, especially when used in conjunction with microfluidics. Coupling these large libraries of linear expression templates to microfluidics required the development of novel methods for device programming. This issue of interfacing microfluidic devices with the rest of the laboratory is known in the field as the 'world-to-chip interface problem' and is non-trivial. Chapter 6 presents solutions to this problem, including a powerful approach of using microarrays for device programming. Here, a single microarray can program a device with thousands of unique solutions. Once a device is programmed using either microarrrays or standard surface chemistry (Chapter 5), protein can be synthesized by ITT. Microfluidics has advantages over classical bench-top approaches, not only because reagent consumption is minimal, but also because protein may be synthesized in batch, semi-continuous, and continuous mode; the latter two approaches having the potential of drastically increasing the final yields.

Completing the integration of high-throughput biology with microfluidics required the adoption of existing detection methods for molecular interactions and the devel-

opment of entirely novel methods. Chapter 7 describes the adoption of standard methods such as antibody-based pull-downs and enzyme methods useful for signal amplification. A novel detection method, based on the mechanically induced trapping of molecular interactions (MITOMI), was developed to address the need to detect transient and low-affinity biological interactions. All high-throughput methods — including yeast two-hybrid, mass spectrometry, and protein arrays — only detect relatively high-affinity interactions. But many biologically interesting interactions are rather transient in nature. Transcription factor – DNA interactions being one example, and the list can be extended to signalling cascades such as the MAPK kinases and g-protein coupled receptors. MITOMI physically traps interactions at equilibrium, and renders their dissociation rates irrelevant. Due to the simplicity of MITOMI it should prove to be widely applicable, potentially having a large impact on the life sciences.

The technological advances described in Chapters 2–7 were applied to the analysis of transcription factor binding energy landscapes. Transcription factors play an intricate role in most, if not all, cellular processes, as they regulate which genes are expressed at any given time. Due to their importance, transcription factors have been studied for several decades, elucidating not only the structural basis of DNA recognition for the various families of transcription factors, but also the topology of transcriptional regulatory networks. One problem that has thus far been intractable is measuring the thermodynamics of transcription factor – DNA binding. Unlike enzymes, which have been described in great detail, transcription factors bind to not

only a single target but can bind to a large number of possible DNA sequences. To understand the biophysical properties of a transcription factor, its affinity to hundreds if not thousands of possible sequences needs to be measured. Exacerbating this already logistically challenging problem is the fact that measuring the affinity of a single interaction requires several measurements taken over various concentrations of one of the two components. From these binding isotherms, the equilibrium dissociation constant can then be established. This could only be accomplished by using all technological elements described in Chapters 2–7, including device programming and, particularly, MITOMI. Binding energy landscapes were established for 4 bHLH transcription factors, a bZIP transcription factor, and a Zn finger transcription factor. Two of the bHLH transcription factor binding energy landscapes were then used to predict *in vivo* function using solely the biophysical characterization and the genome sequence. The *in silico* prediction of transcription factor function worked exceedingly well, despite the limited amount of information used in the algorithm, raising the question of whether cells may one day be simulated with sufficient information about the system. Not only was it possible to use the binding energy landscapes to predict function, but a basic question of whether base contacts are dependent or independent could also be answered. The binding energy landscapes for all transcription factors determined here indicate that base contacts are indeed dependent, and can't be satisfactorily described with weblogos or position weight matrices (PWMs), which are universally used today.

The final two chapters, 9 and 10, quickly summarize efforts to understand the

structure and function of the 19s regulatory particle of the proteasome and the use of yeast cell arrays, respectively. Chapter 10 provides insights into future research directions, centered on understanding protein dynamics on the single-cell level with high temporal and spatial resolution. In short, life yeast cell arrays are generated by microarraying yeast libraries. These spotted arrays are then transformed into life yeast cell arrays by growing the cells on-chip. Additionally these yeast cell arrays may also be transformed into yeast protein arrays presenting an alternative method for generating protein arrays, as described in Chapter 6.

To summarize, this document describes the development of MLSI and novel components for the high-throughput characterization of proteins using microfluidic devices. Specifically, transcription factors were characterized in unprecedented detail by using MITOMI to measure the binding energy landscape topographies. These topographies could be used to predict the function of the measured transcription factor, as well as answer questions regarding the mode of DNA recognition. More generally, the methods described here are broadly applicable and should increase the knowledge of large biological systems, including transcriptional regulatory networks, protein-protein networks, signaling networks, and so forth.

# Chapter 2

# Microfluidics

## 2.1 Introduction

Microfluidics promises to become a ubiquitous tool in the biological sciences. Reasons for this are numerous and include unprecedented throughput and economy of scale, enabling biologists in the era of '-omics' and Systems Biology research. Economy of scale is accomplished by carrying out reactions on devices with length scales ranging from tens to hundreds of microns, resulting in nano- to picoliter reaction volumes, which are up to six orders of magnitude smaller than current standards. Because of these small volumes and length scales, microfluidic assays can be highly parallelized so that many hundreds to thousands of reactions may be performed simultaneously on a single device with a footprint of just a few square centimeters. Finally, it is possible to perform complex fluidic manipulations, such as mixing and metering, making it possible to scale down most existing bench-top assays, as well as design novel experiments enabled by smaller length scales, volumes, and fluid physics encountered on microfluidic devices. Therefore, microfluidics with its economy of scale, high degree of integration, and fluid control is becoming an integral component of life science

research, especially, Systems Biology.

Common materials for the fabrication of microfluidic devices include poly-dimethyl siloxane (PDMS) and silicon. Devices are manufactured using common semi-conductor industry techniques such as photolithography and etching. The Quake laboratory focuses on PDMS device fabrication using a method called multilayer soft lithography (MSL)[1]. Shortly, MSL produces microfluidic devices with at least two distinct fluidic layers stacked on top of one another. A thin deflectable membrane of PDMS forms the interface between the two layers, which in turn may function as a valve if it is either pneumatically or hydraulically actuated. With a functional valve in hand more complex fluidic structures may be built, such as pumps [1], rotary pumps [2], sieve valves [3, 4], multiplexers [5, 6], and freestanding membranes [7]. One of the main strengths of the monolithic PDMS valve is its simplicity, both in its basic structure as well as in its manufacturing. This simplicity makes it possible to design and manufacture highly integrated devices containing thousands of valves robustly and with minimal effort [5].

Assays performed on microfluidic devices to date can roughly be subdivided into cell- and molecular-based assays. Multi- as well as single-cell assays have been performed on various cell types including *E.coli* [8], *S.cerevisiae*, and eukaryotic cells [9]. Multicellular organisms such as drosophila larvae have also been investigated [10]. Microfluidics excels at cellular assays, primarily because the length scales of the object under study and the device used for interrogation are matched. For mammalian cell assays microfluidics is well suited since the channel dimensions roughly match those of

the cells, which have diameters on the order of 5–25 $\mu$m. Length scales are still off by one to two orders of magnitude for prokaryotic cells, but even here single-cell assays can be performed with relative ease. A steady-state bacterial culture can reach up to $10^{12}$ cells per liter, which translates into 1 cell per picoliter. Thus using chambers on the order of hundreds of picoliters and a hundred fold dilution of the initial stock solution will result in a distribution of single cells. Scales are not completely matched but they are considerably closer than standard bench-top approaches, allowing for more refined and sophisticated experiments at the single-cell level.

By now almost all standard molecular assays in modern biology have been ported to microfluidic devices, including PCR [11], enzyme-linked immunosorbent assays (ELISA) [12], *in vitro* transcription/translation (ITT) [13], separations and purifications of both DNA [14] and protein, as well as protein crystallization [15, 16, 6, 17]. Some of these assays have been performed on the single-molecule level, allowing the investigation of non-ensemble behavior of individual molecules, kinetics, and biophysical properties. The unrivaled economy of scale of microfluidic devices is particularly important for molecular assays, where most assay components are exceedingly expensive or difficult to obtain. Furthermore, the small length scales enable fast reaction times, even without convective mixing.

Microfluidic devices thus serve as a powerful Swiss Army knife for the molecular and cellular biologist. The design modularity of PDMS devices fabricated by MSL is only limited by the creativity of the researcher, leaving plenty of room for improvements on basic device characteristics, components, and applications. The next few

sections describe new fluidic components especially pertinent to studying large-scale molecular interactions, such as the mutliplexer, fluid input trees, and a freestanding membrane used for surface derivatization and molecular detection.

## 2.2 Microfluidic Large-Scale Integration

In 2000 the Quake lab reported a device with a total of 36 micromechanical valves fabricated by MSL [1]. In 2001 work was in progress on a device designed by Todd Thorsen, called serpentine enrichment chip (SEC) containing $\sim$ 1400 valves on a footprint of 6 $cm^2$ or about 233 valves/$cm^2$. Even though the total number of valves was quite high, about 75% of them were part of the central grid and essentially formed a single valve system. A second device in development in the lab during that time was a device for the screening of protein crystallization conditions [15] which contained a total of 480 valves on a $\sim$ 24.75 $cm^2$ footprint yielding a density of about 19 valves/$cm^2$. Over the next few years much more complex fluidic devices were designed with thousands of micro-mechanical valves on a footprint the size of a postage stamp [5]. The first device designed had a valve density of 330 valves/$cm^2$ and was called the microfluidic comparator or Microfluidic High-Throughput Screening Chip ($\mu$MHTSC). The second device designed by Todd Thorsen was a microfluidic memory with a valve density of about 600 valves/$cm^2$. Current DTPA devices contain above 7000 valves and functional elements with densities of around 720 valves/$cm^2$.

Microfluidic Large-Scale Integration [5] was an important milestone in the development of microfluidics. MLSI showed that it was possible to robustly fabricate

devices with thousands of functional elements. Compared to previous devices, which contained either a few passive fluidic channels or a few active elements, the increase to thousands of elements was drastic. Just as in the semi-conductor industry, where the transistor replaced vacuum tubes and paved the way to the integrated circuit, the microfluidic MSL valve replaced more complicated valve designs, and, because of its simplicity, could be highly integrated. The number of floating point operations per second (FLOPS) directly correlate with the number of transistors per CPU, so does the functional density of microfluidic devices correlate with the number of valves or other functional elements on a single device. One can thus expect to perform more complex assays with devices containing a larger number of valves. Finally, increase in valve density has been outpacing Moore's law, which describes the exponential increase of transistor density. MLSI is therefore expected to have a large functional increase in the future and be able to tackle more and more complex problems in Systems Biology.

## 2.3 Components

The principal component of microfluidic devices is a monolithic micromechanical valve with a standard footprint of roughly 100 $\mu$m x 100 $\mu$m. The valve itself consists of a PDMS membrane located between two fluidic channels. This membrane may be deflected either up or down, depending on which channel is pressurized. Channels with depths of around 10 $\mu$m may be closed if the membrane is deflected downwards and channels around 50 $\mu$m in height may be closed if the membrane is deflected

upward. These dimensions depend predominantly on the material properties of PDMS as well as the geometry of the structure created. Therefore, either modification of the elastic properties of PDMS or the use of a different compound altogether will result in the ability to create valves with smaller or larger footprints as well as channel heights. Likewise, changing the membrane thickness will also allow for changes in these parameters.

From this principal component more complex fluidic structures may be built. Peristaltic pumps can easily be generated by placing three or more valves in series. Over the years the Quake lab has found many applications for the valve and new functional uses thereof. The following sections describe new components, including a microfluidic multiplexer used to address a large number of flow channels but requiring only a minimal number of control channels (Section 2.3.1). A fluidic input tree is described in Section 2.3.2 which allows the introduction of multiple solutions onto a microfluidic device without cross-contamination issues. As a variation of the standard actuation scheme, where a control line shunts liquid in a flow line, a control line cascade is described in section 2.3.3. In a cascade a control line is used to control a second set of control lines which then in turn manipulate flow lines. This cascading of control lines allows for drastic improvements in multiplexer efficiency and can be used to invert the logic function of a multiplexer. Finally, a free-standing deflectable membrane not acting as a valve but rather as a way to protect a surface from the surrounding liquid is described in Section 2.3.4. The free-standing membrane is particularly useful for measuring molecular interactions, an application described in more detail in Section

7.4.

### 2.3.1 Multiplexer

Controlling dense arrays of fluidic channels or storage modules is a considerable challenge in the field of microfluidics. The electronics industry solved this problem by inventing an integrated circuit called the multiplexer. A multiplexer encodes a series of wires or channels, into a single output. A de-multiplexer functions in the exact opposite fashion, routing a single input into any one of a number of outputs. Multiplexors are commonly used in random access memory (RAM), where a row and column decoder addresses each memory element so that a bit can be written to that location and read at a later time. In the microfluidics case a multiplexer (or de-multiplexer, depending on the direction of fluid flow) consists of a large number of flow channels, specifically addressed by a smaller number of control channels (Figure 2.1).

A multiplexer allows the addressing of a large number of elements with a logarithmically small number of control elements required to perform the addressing. This is accomplished by forming binary trees of wires or fluidic channels. Each binary tree divides the elements to be addressed by two, allowing each half to be specifically addressed. Using consecutive stages of binary trees ultimately allows the addressing of individual elements. The formula governing the multiplexer efficacy is:

$$b * log_b(n) = m \tag{2.1}$$

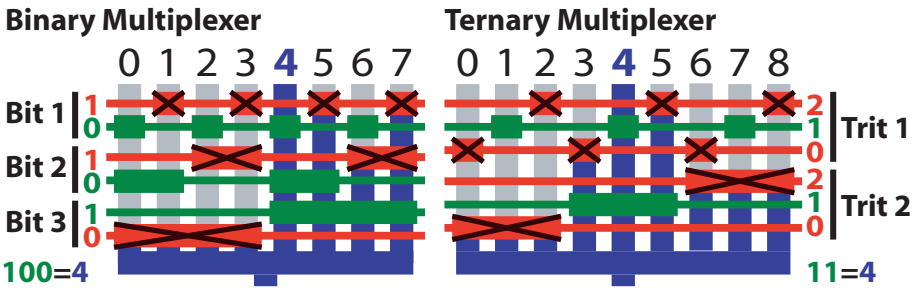



Figure 2.1: Schematics of a binary and ternary fluidic multiplexer. Control channels are indicated by red (actuated) and green (not actuated) colors. Crosses indicate that valves are closed. Each stage of the multiplexers is indicated next to each multiplexer design.

$$2log_2(n) = m \tag{2.2}$$

$$2^{m/2} = n \tag{2.3}$$

where m is the number of control channels controlling n numbers of flow channels and b is the multiplexer base, described in more detail below. The multiplexer was adopted and used in microfluidic devices to address large numbers of fluidic channels with the smallest number of control channels possible. To do so one can simply copy the existing electronics multiplexer and transform it into a fluidic multiplexer (Figure 2.1). But since the microfluidic system is not based solely on binary information it was possible to enhance on the existing multiplexer design by using ternary stages (Figure 2.1) instead of binary ones. So that formula 2.1 becomes

$$3log_3(n) = m \tag{2.4}$$

$$3^{m/3} = n \tag{2.5}$$

This plexer design is the most efficient design possible since the stage integer of 3 is the closest integer to $e$ (2.7). Mathematical proof that $e$ is the function minimum follows:.

$$m = b\,log_b\,n \tag{2.6}$$

$$m = \frac{b\,log\,n}{log\,b} \tag{2.7}$$

$$m = \frac{b}{log\,b} \tag{2.8}$$

$$m' = \frac{ln\,b - 1}{(ln\,b)^2} \tag{2.9}$$

$$0 = \frac{ln\,b - 1}{(ln\,b)^2} \tag{2.10}$$

$$0 = ln\,b - 1 \tag{2.11}$$

$$1 = ln\,b \tag{2.12}$$

$$e = b \tag{2.13}$$

In the physical world only integers are permitted for m and thus either 2 or 3 must be chosen. Indeed it is possible to use any integer for m such as 2, 3, 4, 5, etc. but efficiency drops as the number is deviates from 3. It is nonetheless a useful feature, which can be exploited by mixing stages with different integers for m. For example, if it is necessary to address 54 fluid channels one may either use a binary plexer requiring 6 stages or 12 control lines, a ternary plexer with 4 stages using 12 control lines, or a mix of the two where one stage is binary and the remaining three are ternary and using only 11 control lines instead.

One of the disadvantages of the multiplexer lies in its mode of action: it is designed

to address any one element at a time, but not any number of elements in any combination simultaneously. It is possible to simultaneously address certain symmetry elements, but these are restricted to certain patterns and generally not immediately useful for fluidic applications. It is also not possible to address all but one element at the same time using a multiplexer. This problem was solved by inverting the multiplexer functionality by using a control line cascade described in more detail in Section 2.3.3. A final shortcoming of a standard multiplexer design is its large dead-volumes, which, depending on the type of function the multiplexer has to fulfil, can become prohibitive. This last issue is addressed and solved in Section 2.3.2.

### 2.3.2 Fluid Input Trees

One necessary aspect in the field of microfluidics is interfacing devices with the rest of the laboratory, also known as 'the world-to-chip interface problem'. There are two approaches that will be discussed in this document, one based on programming devices with spotted micro arrays which will be described in detail in Section 6.2.2. The other method introduces liquids via an input hole that reaches the fluidic layer. Even though it is an easy approach for the introduction of liquids, the input hole approach has problems, such as excessive dead volumes leading to contamination when consecutive liquids are introduced through the same input port. The obvious solution to this problem is to have a dedicated input for each solution that is to be introduced. For a small number of liquids this can be accomplished without the need of a multiplexer to address the inputs. Therefore by smartly arranging the

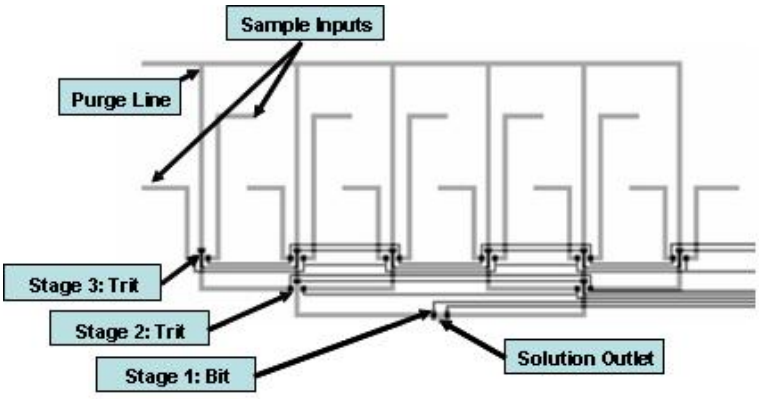



Figure 2.2: Example of a fluid input tree controlled by a multiplexer and every third line sacrificed for buffer input

input tree and valve structure one can avoid any on-chip dead-volumes and thus cross-contamination from the different inputs.

When the number of inputs becomes relatively large, a dozen or more, then the use of a multiplexer becomes necessary. But, as mentioned in Section 2.3.1, the standard multiplexer has large dead volumes associated with it. In order to reduce dead volumes, the valve structure can be changed such that the lines to be addressed are arranged in a tree-like fashion, with the control valves being situated directly at each branch point. In order to completely avoid downstream cross-contamination one can sacrifice every second or every third line of a binary or ternary multiplexer for flushing with a buffer solution. These multiplexer-based fluid input trees are only limited by the large footprint required for the fluid interconnects, can ultimately support up to hundreds of unique solutions, and are particularly well suited for introducing between 12–40 unique solutions. If hundreds or even thousands of unique solutions are to be introduced, a different approach based on spotted microarrays is more suitable (Section 6.2.2).

### 2.3.3 Control Line Cascades

It is possible to control fluid lines, which in turn control another level of fluid lines; this is termed a control line cascade. A control line cascade can be applied to a multi-level multiplexer. Here the first level of fluid lines consists of a multiplexer, which in turn controls a second level of lines that also form a multiplexer. This second stage multiplexer then in turn controls a third and final stage of fluid lines.

Figure 2.3 shows an example of a 3-level control line cascade. The first level of control lines (black) consists of a 2-stage ternary multiplexer which controls 9 lines (blue) of a 3-stage ternary multiplexer. The black multiplexer can be used to sequentially address the blue multiplexer. Thus if the blue multiplexer is initially in an off state, individual lines may be first addressed and then pressurized. The pressure is stored for a certain amount of time, depending on the permeability of PDMS and the solution or gas used to pressurize the system. Now that the blue multiplexer has been addressed it specifically addresses the third stage of green fluid channels. Cascading achieves a further reduction in the number of control inputs required. In this basic example, driving the blue multiplexer directly would have required 9 control lines. By using a control line cascade, the same functionality was achieved with 7 control lines. Efficiency scales favorably with the number of lines, so that a first-stage multiplexer consisting of 5 ternary stages controls 243 second-stage control lines, which in turn control $\sim 4 * 10^{38}$ lines. Likewise, a 4 stage ternary multiplexer controls 81 second-stage lines controlling $\sim$7 million lines. This would nominally require 243 and 81 inputs, respectively, but with a control line cascade is instead accomplished with 16

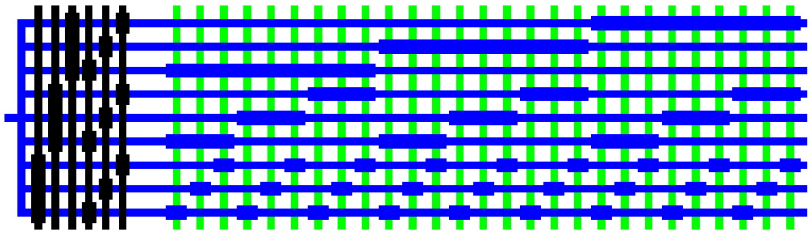


Figure 2.3: An example of a control line cascade in which the black lines control the blue lines, which in turn then address the green flow lines.

and 13 lines respectively. A second advantage of a control line cascade is that it can invert a multiplexer. This is accomplished by using the control line cascade described above, but having each blue line be a single valve controlling a respective green fluid line directly. Now if the black multiplexer is addressed it in turns selects any one of the nine blue control lines, which closes one green fluid channel, effectively inverting the functionality of the the original multiplexer, which can only open a fluid line while having all others closed. Inversion of the multiplexer was necessary for a microfluidic memory device descibed in Section 3.2.3.

### 2.3.4 Free-standing Membranes

In order to generate areas of defined surface chemistry as well as geometry, a new use of the microfluidic valve structure was employed. Here, instead of using a deflectable membrane in the channel structure to obstruct fluid flow by complete closure of the flow channel, the membrane simply makes contact with the channel surface, but is otherwise freestanding and not contiguous with the channel walls. It should be noted that the first membrane designed was indeed not free-standing, but rather mimicked

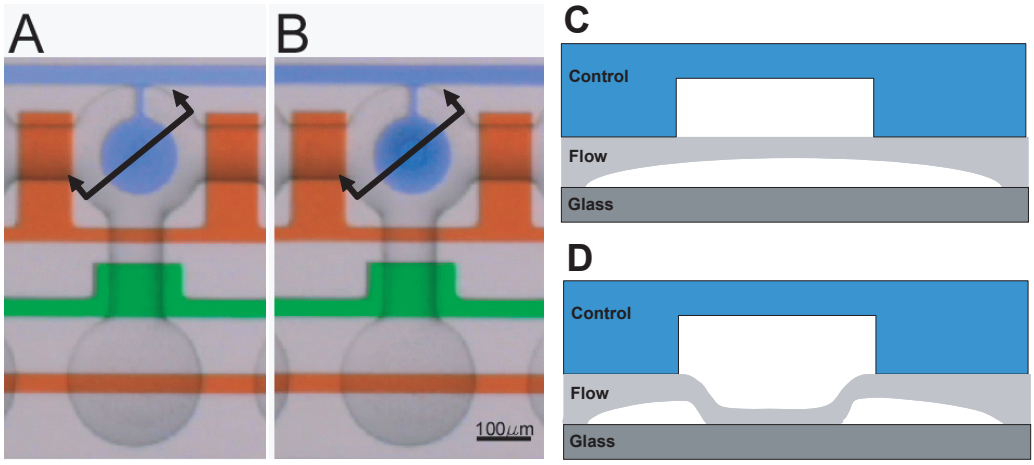



Figure 2.4: Close-up view of a D2PA unit cell with control lines filled with colored food dyes and an empty flow layer. The blue control line creates the freestanding membrane that is in an open (panel A and C) and closed configuration (panel B and D). Panels C and D show a schematic of a cross section through the correlating image along the arrow-demarcated black lines.

Table 2.1: Spot size dependence and uniformity

| Membrane diameter ($\mu$m) | Initial closing pressure (psi) | Spot diameter at 15 psi ($\mu$m) | Std. dev. ($\mu$m) |
|---|---|---|---|
| 180 | 6.5 | 102 | 5 |
| 160 | 6.5 | 80 | 5 |
| 140 | 6.5 | 54 | 5 |
| 120 | 7.5 | 33 | 4 |
| 100 | 19 | - | - |
| 80 | 19 | - | - |

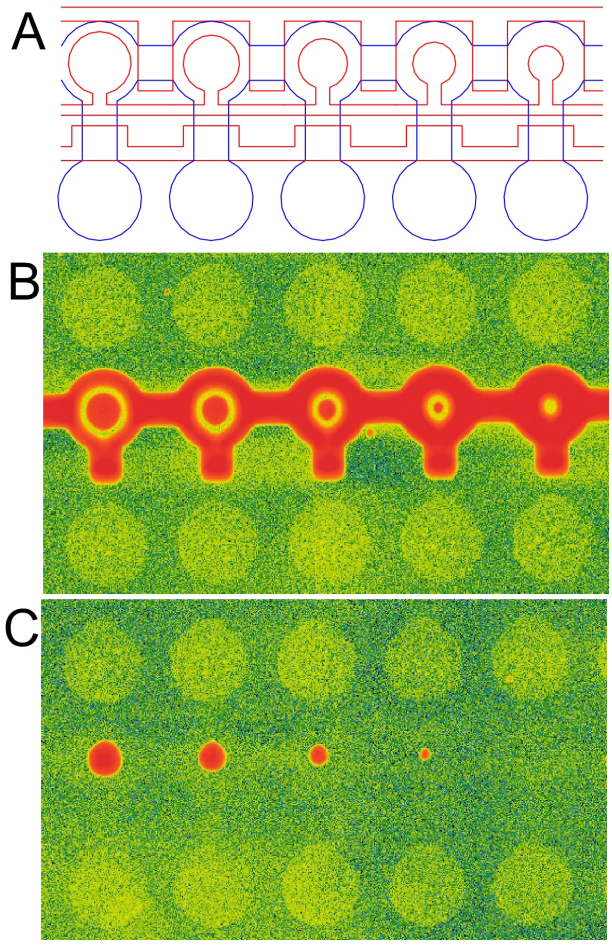



Figure 2.5: Testing the dependence of the actual spot size on membrane diameter. Panel A shows an Autocad diagram of a section of the actual device. Here 5 unit cells with different membrane diameters are shown (red indicates control lines and blue flow lines). Membrane diameters are 180 $\mu$m on the far left, decreasing to 100 $\mu$m in 20 $\mu$m steps. The actual device has one additional unit cell with a 80 $\mu$m membrane. Panel B shows the fluorescence of Cy5 labeled DNA templates filled in the flow channel. The membranes have been closed trapping DNA bound by a surface bound transcription factor. Note the halo of low intensity around the spots, indicative of low non-specific binding of templates due to the membrane action. Panel C shows the same area of the device as Panel B after flushing the flow channel with PBS with the membranes remaining closed to prevent loss of bound material.

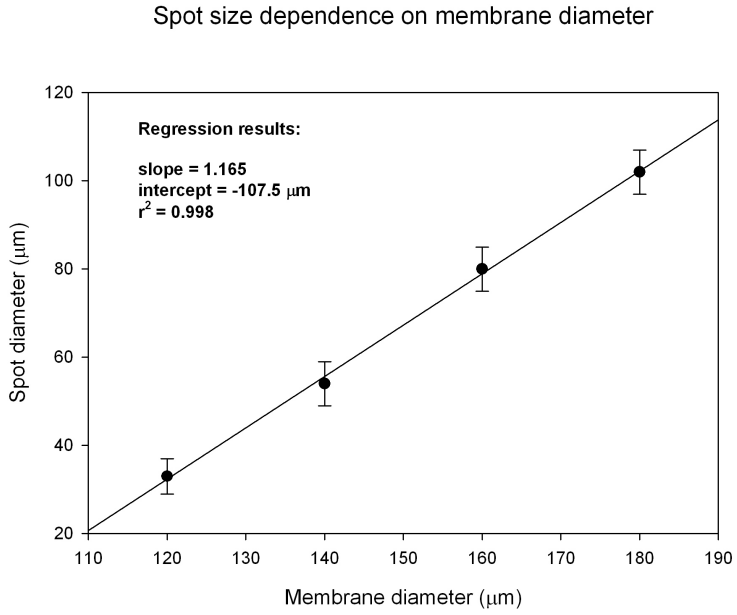



Figure 2.6: Spot size dependence on membrane diameter

a standard valve, surrounded by bus channels, to allow for fluid exchange while the membrane is actuate (see chip design in Figure C.4). This protects the channel surface while the membrane is in the closed state. Figure 2.4 illustrates the fluidic layout and functionality of the freestanding membrane structure. A round membrane is created by the blue control channel. The membrane is situated above a larger circular flow channel. Panel A of Figure 2.4 shows the membrane in an open state, which upon hydraulic actuation makes contact with the surface (Panel B). Membrane closure occurs centrally and extends radially outward allowing the membrane to close without trapping any fluid between it and the surface, an important feature for applications discussed in Section 7.4.

The membrane makes circular contacts of defined diameter, which may be modulated by varying the membrane diameter, actuation pressure, membrane thickness,

and flow channel depth. A chip was designed to address the effect of varying the membrane diameter (Figure 2.5) by systematically varying the membrane diameter in 20 $\mu$m steps ranging from 80–180 $\mu$m. For the given flow channel height and membrane thickness the various membranes closed at pressures ranging from 6.5 psi to 19 psi (see Table 2.1). Figure 2.6 shows the spot size dependance on membrane diameter at a constant closing pressure of 15 psi. The slope is slightly steeper than 1 (1.165) with an intercept of -107.5 $\mu$m. A change in actuation pressure, membrane thickness, and flow channel height will likely only affect the intercept value, but otherwise leave the slope unchanged, making it easy to design membranes that will generate a desired spot diameter. The smallest spot size generated in this experiment was on the order of 30 $\mu$m, or about 1/2 to 1/3 the size of spots achieved with commercially available quill pen spotting technology.

The free-standing membrane has two principal applications. First, the membrane may be used to protect a patch of surface from surface treatments allowing specific molecules to be deposited to generate a circular area with defined chemistry. Chapter 5 describes how these surface chemistries are built and how they are used to detect molecular interactions. Using the free-standing membrane it is possible to generate circular features with variable diameters reaching as low as 30 $\mu$m and possibly lower. Circular features are compatible with most software packages written for the analysis of DNA arrays and thus facilitate data mining. Furthermore, by creating defined areas with a small footprint, mass transfer issues that occur when precious samples with low molarities, such as antibodies, are used for the derivatization are alleviated.

The second, and more important application of the free-standing membrane allows, trapping of molecular interactions occurring on the surface. This enables the detection of interactions with low affinity constants, especially those arising due to high off-rates where flushing or washing to remove unbound molecules is prohibitive due to rapid loss of bound material. Closing the membrane also allows for the exchange of the surrounding solution, useful when enzymatic assays are required for detection (Section 7.3). These assays generally require the introduction of an enzyme and its substrate, and may now be accomplished without loss of bound material, resulting in the highest signal achievable.

## 2.4 Readout Systems

It is a necessity to measure the results of on-chip experiments *in situ*. Many biological analyses are based on optical information such as light microscopy, fluorescence, bio/chemiluminescence, etc. All these methods are applicable to devices fabricated from PDMS since the material is optically transparent. It is thus possible to interrogate samples directly through the device, which is important when flow channels are entirely made up of PDMS. In most devices described herein, the flow channel walls and ceilings consist of PDMS, but the bottom of the channel consists of glass, in which case optical interrogation is trivial.

The most commonly used instrument for optical interrogation is the light microscope, which is ubiquitously used in the laboratory and its features and drawbacks are described in Section 2.4.1. The possibility of using a home-built fluorescent-based

scanning system for the interrogation of microfluidic devices was also investigated, leading to the use of commercially available systems described in Section 2.4.2.

### 2.4.1 Microscope

The light microscope is a standard tool for interrogating biological systems. It has several advantages over other standard optical instruments in that it has high temporal as well as spatial resolution and is a flexible system. The high spatial resolution, limited only by the diffraction limit of light or roughly $\lambda/2$, allows for analysis of cellular fine structures otherwise not accessible. The high temporal resolution allows for the measurement of molecular kinetics for example. In early experiments and for kinetic analysis (Section 8.5) an Olympus IX50 inverted fluorescent microscope equipped with either a ST7-XME (SBIG Astronomical Instruments) cooled CCD camera or a E717-21 PMT (Hamamatsu) was used.

### 2.4.2 DNA Array Scanners

The first array scanner used was a home-built version used to prove the applicability to the interrogation of microfluidic devices. The home-built scanner was based on a FACS setup built by Todd Thorsen, to which a motorized x,y stage was added with a z micrometer for focus adjustment. The FACS setup consisted of an Intelite solid-state diode laser ($\lambda$=473 nm) for illumination and two emission filters centered at 580 nm/30 nm and 535 nm/40 nm. Using the stage, a chip could be panned beneath a 40x or 60x objective through which the laser was focused on the sample and emitted light

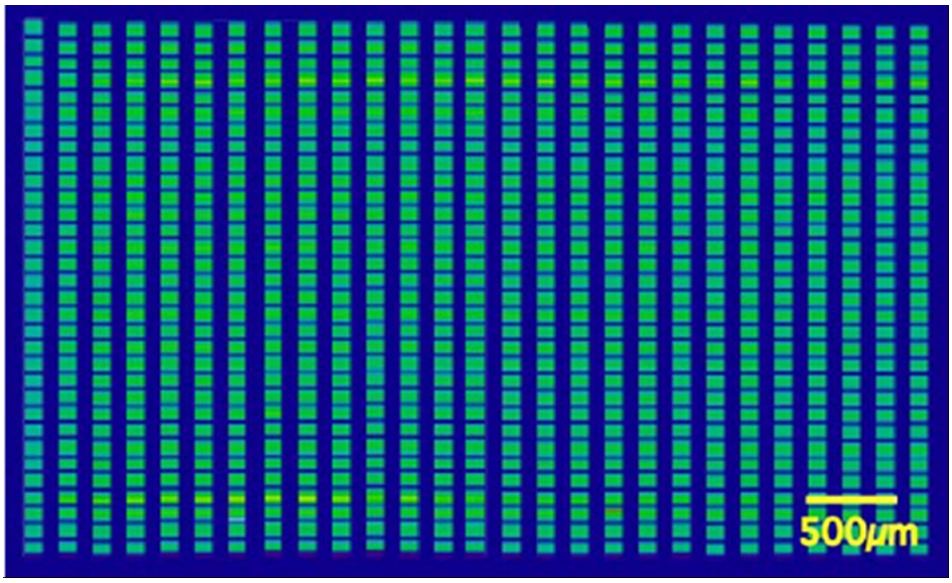



Figure 2.7: Image of a scan taken with a home-built array scanner of a SEC chip filled with 1 mM fluorescein in PBS (pH7.2). The raw data was analyzed in Labview and the resulting image generated using Mathcad.

collected and routed through one of the emission filters. Fluorescence was measured in one of two channels using PMTs for detection and analog voltage output was read by a PCI1200 card (National Instruments). The stages were driven by a DCX-PC100 motion control card (Precision MicroControl) and the whole setup was controlled by a Labview program. Figure 2.7 was acquired using the home-built device. The statistical variation of all 1024 chambers imaged was 5–6.25% standard deviations, dependent on initial data binning, and the time required to scan the entire device was under 5 minutes. Even though it was possible to prove the usefulness of DNA array scanners for reading out microfluidic devices, the home-built version was not engineered well enough to yield consistent and high-quality data. One of the major problems with the home-built setup was that the laser line had to be pulled across the columns or rows of the device, requiring aligning of the chip axes to the axes of the

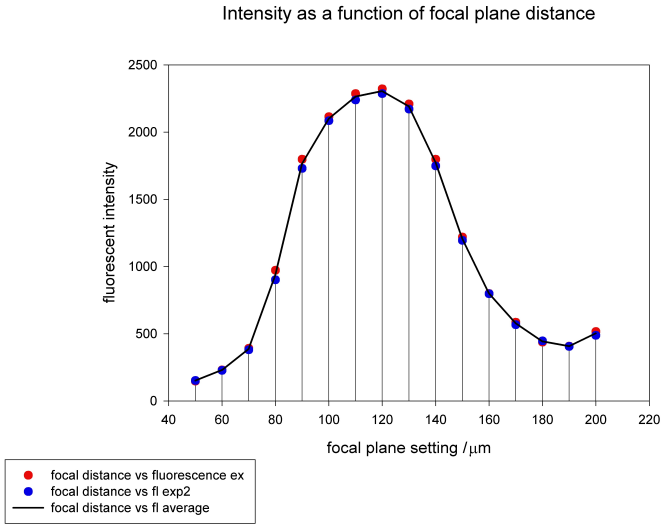



Figure 2.8: Experimental results of signal intensities as a function of focal depth. The results show that the GenePix4000b scanner is confocal and intensities drop off radically as focus shifts

x,y stages, a non-trivial task. Furthermore the data is at best 1-dimensional giving point values of intensities of each chamber. A full 2-dimensional scan of a device is desirable, but again requires more sophisticated engineering.

The first commercial instrument acquired was the GenePix4000b DNA array scanner (Axon Instruments). The scanner has two laser lines for illumination, at 532 nm and at 635 nm. These lines efficiently excite Cy3, Cy5, and similar dyes. The resolution of the system is 5 $\mu$m and has a dynamic range over 3–4 orders of magnitude. This scanner was used for all high-throughput screening experiments described in Chapter 3 and for all scans of standard DNA micro-arrays. Even though a good instrument for scanning standard planar micro-arrays, it was difficult to obtain high-

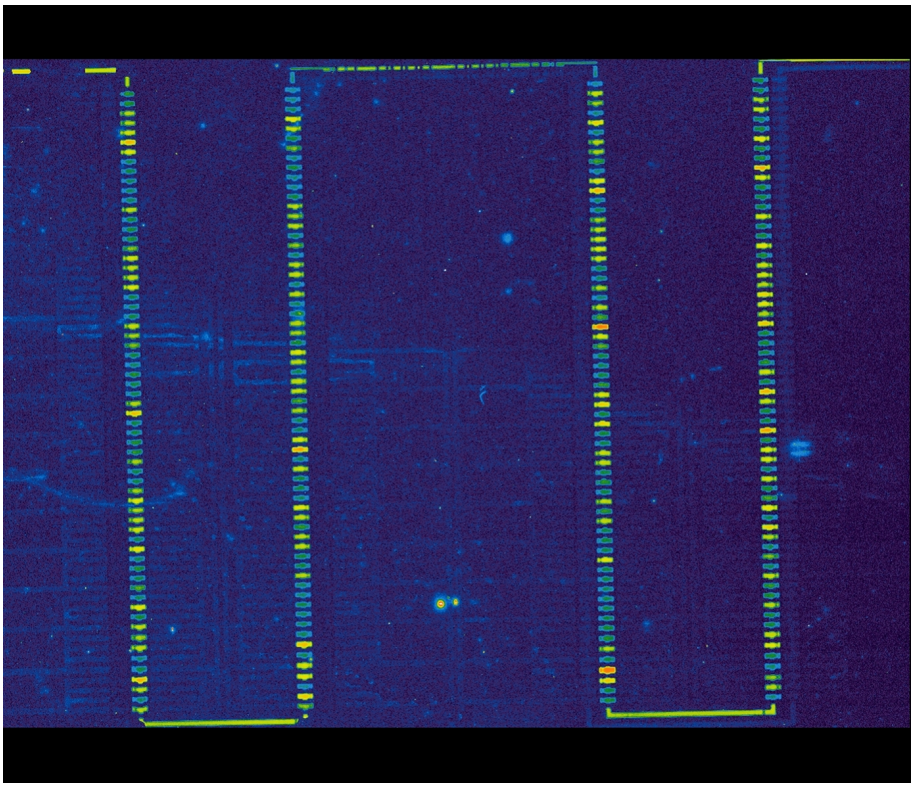


Figure 2.9: Scan obtained of a $\mu$MHTSC device running single-molecule enzyme assays. Note the uniformity of the background and signals across the entire area of the image.

quality device scans because of the shallowness of the depth of focus of the system (Figure 2.8), making it difficult to obtain a focused image over a large area. Compounding this problem were three factors: first, the stage of the array scanner is a tripod design consisting of three sapphire posts; second, due to the limited working distance of the scanner the devices had to be bonded to number 1 or number 0 coverslips; and lastly, the device had to remain connected to pressure lines to control the device while it was being scanned. These three factors made it difficult to level a device consistently. Several level images were nonetheless scanned (Figure 2.9) and it was possible to obtain quality data from these.

Realizing that the GenePix4000b was not suited for the routine analysis of microfluidic devices, a second commercially available scanner had to be found. The instrument selected was the arrayWoRx e (Applied Precision). Several features of this instrument made it more suitable for microfluidic applications. The instrument uses a metal-halide lamp for illumination rather than lasers, which allows for dialing of any excitation from 350–700 nm using bandpass filters, and, more importantly, removes the confocal nature of the dependence of the focal depth. Four filter sets can be housed in the system at any given time and may be easily exchanged, extending its usefulness into the FITC band commonly used in biological assays. Emission is collected via a CCD camera. Extending the focal depth of the system, combined with a custom-designed chip holder for both 1 by 3 inch glass slides as well as 2 by 3 inch slides made this system very useful in the routine analysis of microfluidic devices. All other specifications are on a par with the GenePix4000b, with a slightly better resolution at 3.25 $\mu$m but a lower overall sensitivity and scan speed.

# Chapter 3

# High-Throughput Screening Applications

## 3.1 Introduction

Protein engineering is one of the major tasks of today's field of bioengineering [18]. Proteins provide moldable platforms of thousands of enzymes and structural building blocks that may be engineered at leisure by restructuring their amino acid sequence. This restructuring may yield beneficial changes in the secondary, tertiary, and even quaternary structure of the protein. Beneficial attributes may include increased stability at various abnormal environmental conditions, increase in enzymatic turnover rates and numbers, recognition and processing of unnatural substrates, and changing affinities to binding partners.

To achieve these engineering feats two general approaches may be taken: rational design and random mutagenesis. Both approaches require the novel protein to be screened to ascertain that it possesses the correct function to be engineered. The rational design path generally yields a low number of potential candidates, numbering at most in the hundreds. The random mutagenesis approach on the other hand

yields libraries with member counts reaching astronomical numbers. Historically the approach for screening random mutagenesis libraries, generated by error-prone PCR, consisted of cloning a library of mutant linear PCR products into a vector, which then was transformed into an appropriate host, generally *E. coli*. The transformation step insures that clonal libraries are generated, which may be screened in later steps for the appropriate function. Physical separation of clones is achieved by plating the library, followed by manual or automated picking. Per library, approximately 2,000–20,000 clones may be screened using these methods. To avoid the time-consuming and labor-intensive step of plating and picking of clones, followed by the expressing and screening of each, microfluidic chips were devised which directly take the initial library as input and perform the necessary screening step on single clonal cells. Furthermore the devices allow for the recovery of the screened content, which is a necessary requirement in any high-throughput screening application. Three devices were designed with various features and capabilities and are described in detail in Sections 3.2.1, 3.2.2, and 3.2.3.

## 3.2 HTS Devices and Assays

### 3.2.1 Serpentine Enrichment Chip

The serpentine enrichment chip (SEC) (Figure 3.1) was the first chip designed for the high-throughput screening of bacterial libraries. The design goal for this device was to screen heterogeneous libraries directly without the need for long and tedious

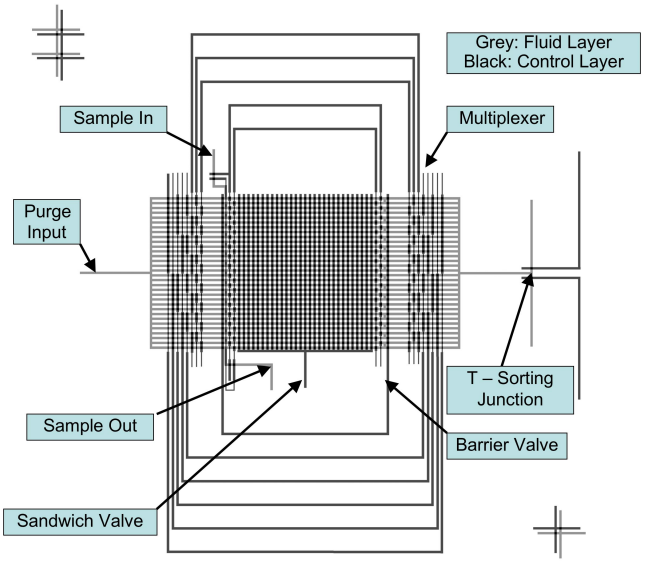



Figure 3.1: Photoshop design drawing of the SEC device. Dark grey lines indicate control lines and light grey lines are flow lines.

separation and screening steps, and is accomplished by dilution and segregation of the heterogeneous mixture into many smaller homogeneous aliquots. On-chip this is done by introducing the mixture into the input port of the device, which then branches out into 32 columns. The liquid introduced into those 32 columns is then segregated into 1024 chambers with a volume of roughly 100 pL each. The input cell density can be chosen to yield on average 1 cell per chamber, according to Poisson statistics. Now, since each chamber contains a single cell, and thus a single member of the clone library, it is possible to screen the function of that single mutant. Of course, due to the statistical nature of the distribution it is possible to obtain 2 or more cells per chamber as well. Signals derived from chambers with 2 or more cells may be deconvoluted from single-cell signals in two ways. One method is to simply detect the signal and assume that two cells will roughly yield twice the signal. This

assumption may be flawed, since many mutants may show strongly reduced activity as compared to wild type, such that a chamber with two inefficient mutants may appear to only contain a single mutant. Using direct counting of cells, as the second method, circumvents that problem. Once a chamber containing a single mutant of interest has been identified, the segregating grid is opened, and the column containing the clone of interest specifically addressed and purged. At the outlet port a T-sorting junction could specifically sort out all negative mutants from the positive clone of interest. Unfortunately this sorting strategy failed due to flow properties in microfluidic devices. Even though the index of each cell was known before the purge was started, once flow occurred index could shift, since some bacteria would reside on the wall whereas others would be located more towards the center of the channel. Due to laminar flow, those bacteria on the sidewalls would move slower since flow velocities drop off towards zero near channel walls. Hence, a bacterium located in the central part of the channel would pass several bacteria located near the walls. This prohibited the recovery of a single mutant directly from the chip in a single pass. Solutions to this problem included recovering a whole column and plating all recovered mutants thus achieving a 32-fold enrichment over the initial chip input. Extending on the above idea it was in principle possible to take that initial column fraction and re-introduce it into the same or a new device. Thus achieving another 32-fold enrichment. This could be repeated until the mutant of interest would be contained in a column all by itself and thus could be recovered from the chip without any contaminating mutants. A chip was designed that linked the out port back to the in port. Since such a

circular geometry does not allow for pressure driven or electrostatic flow, a pump was introduced into the circle to drive the flow. An additional fluid input was also needed so that the single column output that was to be re-introduced into the chip could be diluted accordingly to again fill 32 columns. Sound in principle, the chip was nonetheless never extensively tested. From this initial design it was obvious though that a device that addressed chambers directly and individually was a much more straightforward approach to HTS, leading to the design of the microfluidic memory and $\mu$MHTSC devices.

### 3.2.2 $\mu$MHTSC

In order to overcome the limitations encountered when testing the SEC device, a new generation of HTS devices were designed and tested. The next-generation devices were still based on the idea of segregating a complex mixture into small compartments where the contents could be tested. Additionally each compartment had to be individually addressable, allowing the recovery of a chamber's contents without contamination. This required a much more complex fluidic infrastructure based on large multiplexer arrays. Furthermore the use of more advanced assays was to be enabled on the platform requiring the mixing of two solutions *in situ* so that enzymatic assays could be started on-chip. Thus the basic layout of the $\mu$MHTSC consisted of 256 flow channels controlled by an 8-stage binary multiplexer (Figure 3.3). For sample introduction the flow channels may be closed perpendicularly by barrier valves, which generates two serpentines. Once the two sample fluids have been introduced

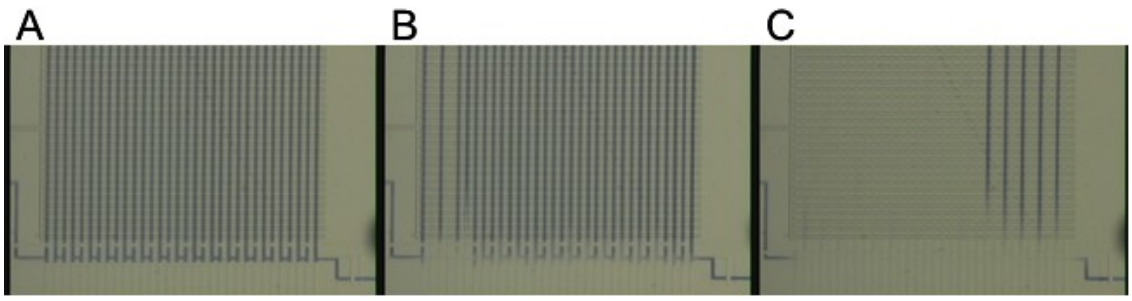



Figure 3.2: Panel A shows a SEC device with its serpentine filled with a saturated solution of bromophenol blue. In panel B columns 2 and 4 (from the left) have been selectively purged and column 6 is in the process of being purged. Panel C shows most of the columns purged with only 5 remaining.

into the device they may be segregated by closing the sandwich barrier valve creating 512 chambers of 375 pL. These chambers may then be mixed pairwise by opening the mixer barrier, creating 256 chambers each of 750 pL, allowing reactions to take place synchronized with a defined null time. Upon readout of the device and identification of a chamber of interest the multiplexer is used to address the specific flow channels and the contents are purged towards one of 4 collection ports (Figure 3.4). In order to allow more than one sample to be recovered, each device contained 4 columns each with 64 chambers from the original string of 256, accomplished by the 4 barrier control lines, which may be seen as an additional multiplexer stage. Each column has a dedicated sample collection port so that no contamination can occur between individual samples. The next step after successful testing of the microfluidics was to run single enzyme assays. As a model system *E.coli* expressing the enzyme cytochrome c peroxidase (CCP) was used. In order to accurately determine enzyme activity of individual clones it is necessary to normalize substrate turnover to the number of cells present in a chamber. It is possible to count bacteria by eye using an 60x oil

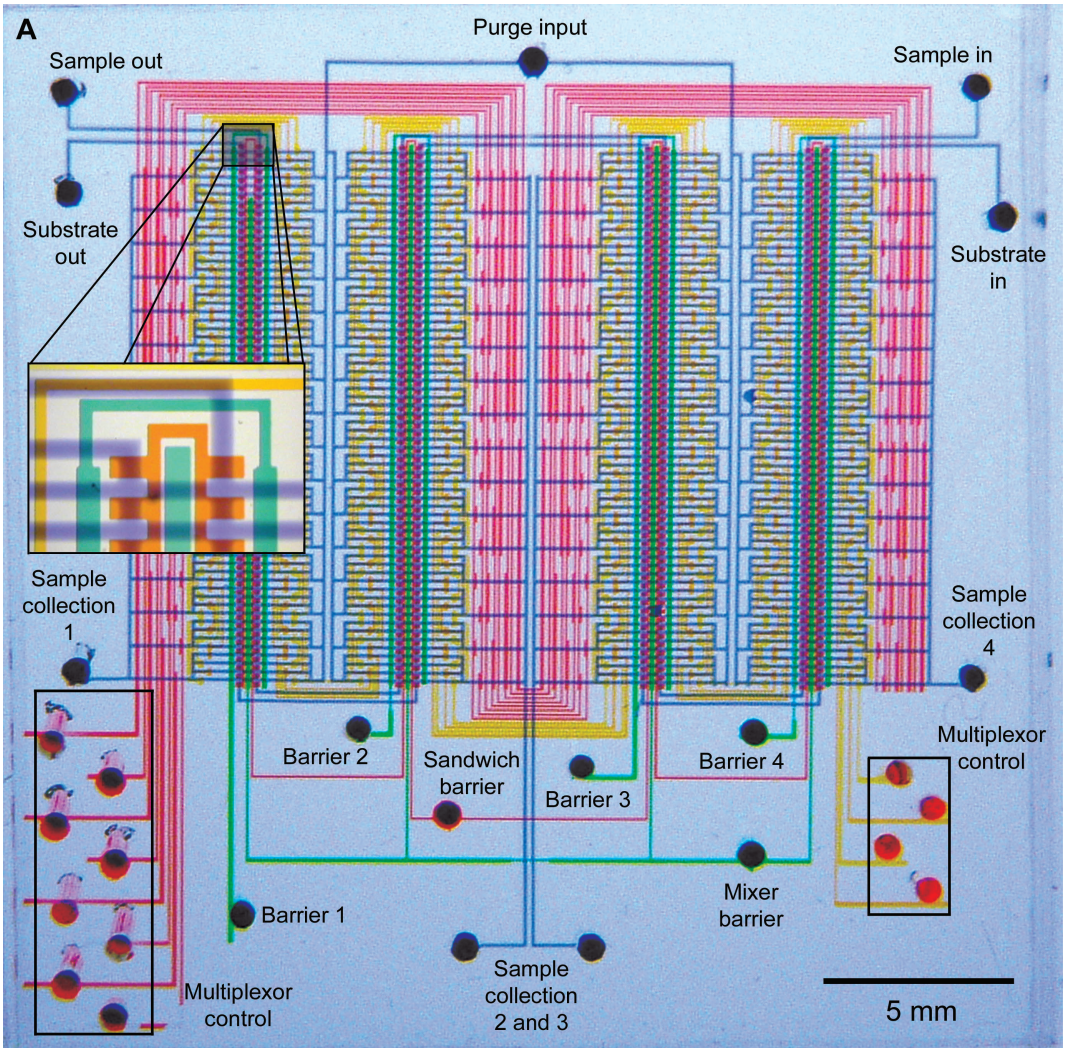



Figure 3.3: Optical micrograph of the $\mu$MHTSC. For visualization, control lines have been loaded with green, red, and yellow food dyes and flow lines with blue dye.

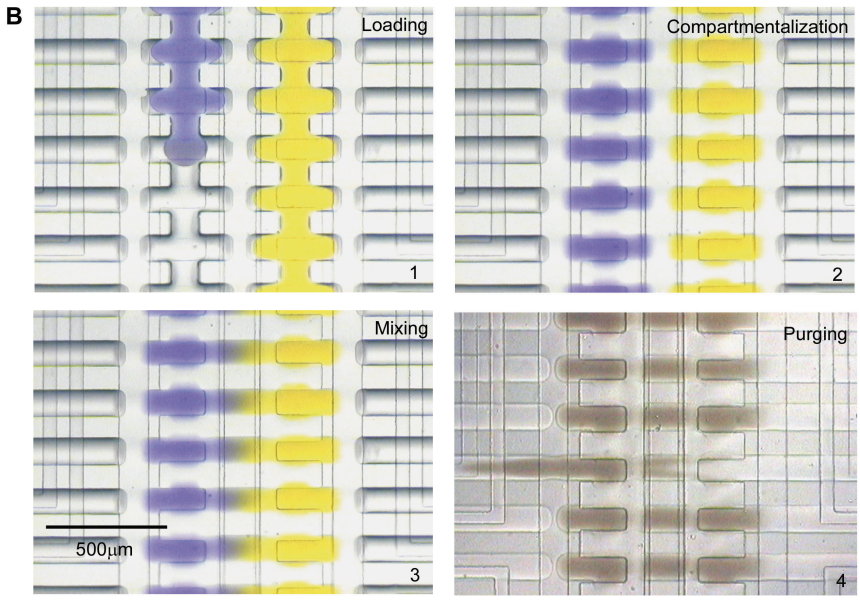



Figure 3.4: These captures of a section of the $\mu$MHTSC show how the device can be loaded with two distinct samples. In this case a bromophenol blue and orange G solution, blue and yellow, respectively (Panel 1). The solutions are introduced into the two serpentines created by closing all barrier valves. Once the serpentines are filled the fluids may be segregated by closing the sandwich barrier (Panel 2). Once the samples have been segregated they may be mixed pairwise (Panel 3) and one chamber can be specifically addressed and its contents purged towards one of the 4 sample collection ports (Panel 4).

immersion objective manually panning and searching all 256 chambers. As manual counting is a very time-consuming and strenuous task, a fluorescent reporter dye in conjunction with the GenePix 4000b scanner was used to count cells more efficiently. One dye suitable for this purpose was SYTO 62, commonly used in cell viability assays. SYTO 62 is membrane permeable and will stain both live and dead cells by intercalating into DNA and RNA, and thus is potentially harmful to the cell. But HTS methods don't require the recovery of live cells. All that needs to be recovered are the plasmid copies coding for the mutant protein of interest harbored in the cell, which then may be PCR amplified off chip and sequenced.

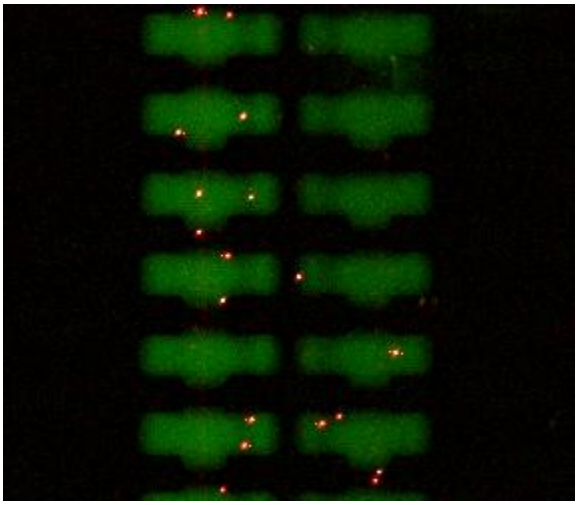


Figure 3.5: GenePix 4000b scan of a solution of SYTO 62-stained bacteria. Individual bacteria light up a small cluster of pixels, indicated by a red to white color.

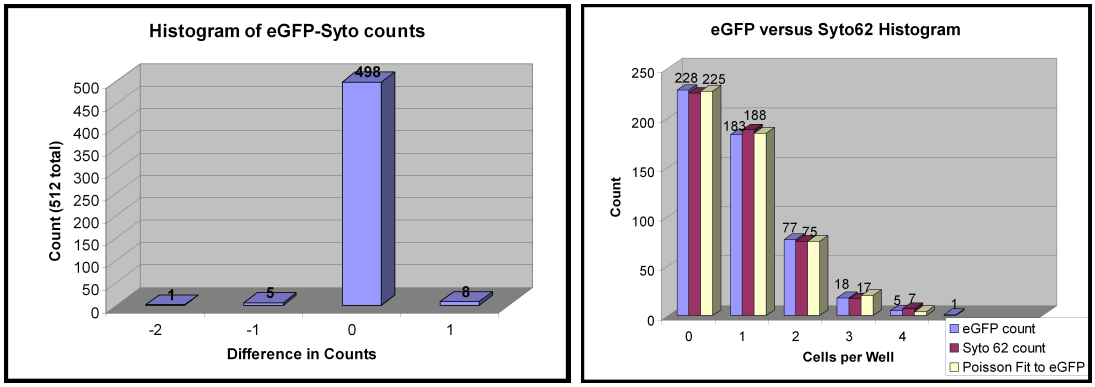



Figure 3.6: The left graph shows the distribution of wrong counts between Syto 62 and the eGFP hand count. At most the difference was -2 per chamber and only 14 out of a total of 512 chambers were counted wrong. The right graph shows the distribution of the two counts as compared to theoretical values derived from a poisson distribution.

Using SYTO 62 as the reporter dye visualized individual cells when scanned with the GenePix array scanner (Figure 3.5). Unfortunately the spatial resolution of that instrument is 5 $\mu$m. Thus it is possible that two cells could cluster together and light up a single pixel. As a control experiment, a SYTO 62-stained culture of cells expressing eGFP was loaded and counted by hand as well as by using the scanner. Comparing the two counts showed that over 97% of all chambers were counted correctly using SYTO 62 in reference to the manual GFP count. Both counts were also compared to a hypothetical poisson distribution and the distribution encountered on-chip followed the expected distribution (Figure 3.6).

Using the above-described method for determining cells per chamber enabled the single-cell enzyme analysis. *E.coli* expressing CCP were induced with isopropylthio-beta-D-galactosidase (IPTG) for 7 hours. A 1 mL confluent culture was spun down and re-suspended in PBS to 1/100 of the initial concentration. This suspension was further diluted 1/10 into Amplex Red reaction mixture containing final concentrations of 100 $\mu$M Amplex Red and 880 $\mu$M $H_2O_2$. This solution was then loaded into one of the serpentines of the $\mu$MHTSC and compartmentalized, followed by a 1-hour-long incubation at room temperature. The chip was then scanned (Figure 3.7) and the resulting image analyzed with Genepix3.0 to determine signal intensities per chamber.

The results indicate that single-cell enzyme assays are possible (Figure 3.8). Chambers containing no cells remained at background signal levels. If a cell is present in a chamber the distribution broadens and the peak shifts to higher fluorescent values, indicating that substrate is being turned over. Distribution broadening is likely due

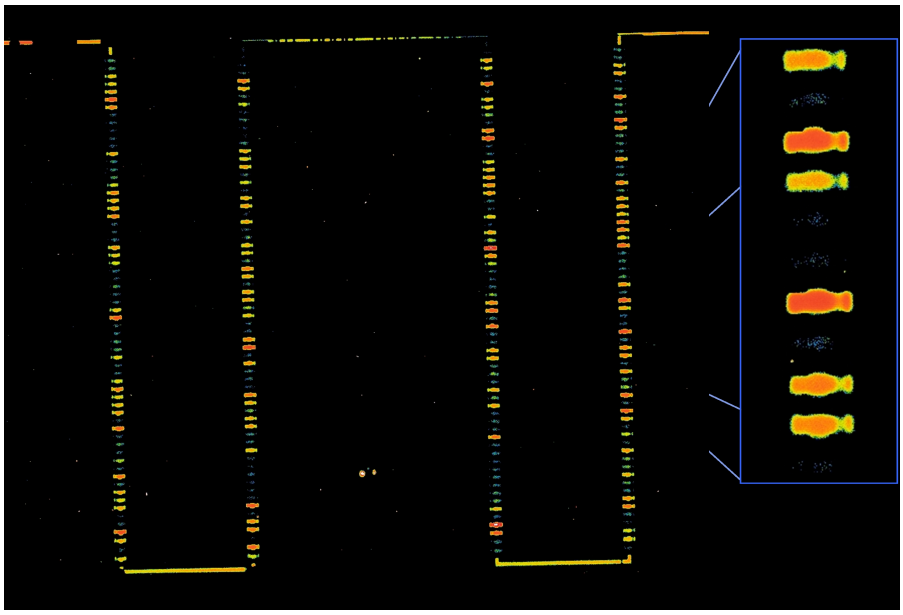


Figure 3.7: GenePix4000b scan of a $\mu$MHTSC device containing a distribution of cells expressing CCP. Chambers containing no cells remain dark, whereas cells containing 1 or more cells expressing CCP turn over Amplex Red and give rise to a fluorescent signal, as indicated by yellow to orange colors. The inset shows a detailed view of a section of the serpentine containing cells.

to intrinsic variation in CCP expression from cell to cell; a cell may contain different numbers of functional CCP molecules depending on its cell cycle and viability. Single-cell variability is an interesting aspect of single-cell measurements. In this case it complicated the goal of determining enzyme fitness on the single-cell level, so that mutants of varying fitness cannot be distinguished from one another since they would likely fall within the standard deviation of the distribution exhibited solely from single-cell variation.

In order to ascertain that the distributions result from variances in single cells, the signals in each category of cells per chamber were averaged. The result is as expected, such that the median of the distribution follows a linear relationship, giving rise to

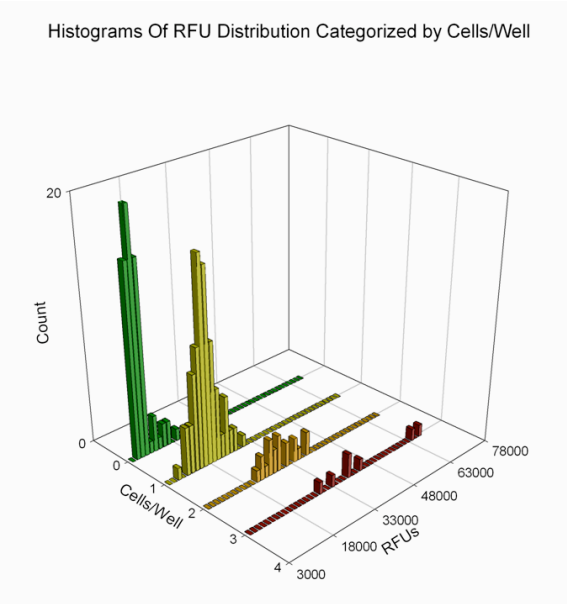



Figure 3.8: Analysis of the single-cell enzyme assay. Chamber intensities are plotted as relative fluorescent units. Categories on the x axis are cells per well, determined by a manual count and are plotted against the RFU of the chamber. No cells in a chamber give rise to a very sharp background distribution of intensities. As the number of cells increases per chamber the distribution broadens due to intrinsic variances in single-cell expression.

quantal increases in fluorescent intensity solely depending on the number of cells per chamber (Figure 3.9).

Seeing such a large distribution in signal derived from single cells stemming from variances in copy numbers of enzyme per cell proved to be a problem for the determination of fit mutants in a library. Possible ways to detect fit mutants in a sea of variance would include higher-resolution measurements of cell morphology. By using cell morphology it should be possible to derive enzyme concentrations more reliably. A simple measurement of cell size could shrink the distribution to levels that are amenable to detecting fit mutants. Unfortunately none of the commercially available DNA scanners have optical resolutions high enough for this task. A second option

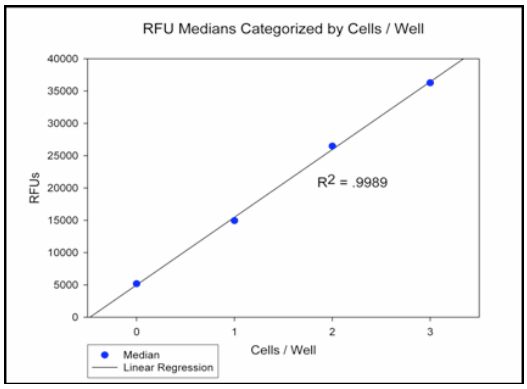



Figure 3.9: Averaging the results from Figure 3.9 should give rise to population averages and quantal increases in fluorescence per chamber dependent on the number of cells contained within.

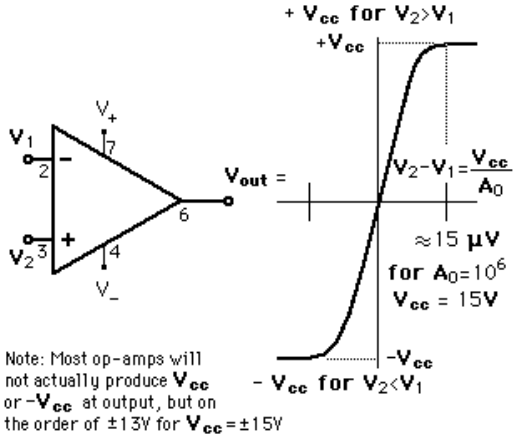



Figure 3.10: Schematic of a comparator

would be to fuse the enzyme of interest with a reporter protein such as GFP so that actual concentrations of enzyme present can be determined directly. Normalizing to actual enzyme concentrations removes any single cell variance present and thus allows one to detect variance in enzyme fitness directly.

To take advantage of the $\mu$MHTSC's mixing capability an experiment was devised to emulate a basic electronics component, namely the comparator. A comparator is an integrated circuit which is a derivative of an operational amplifier. The comparator

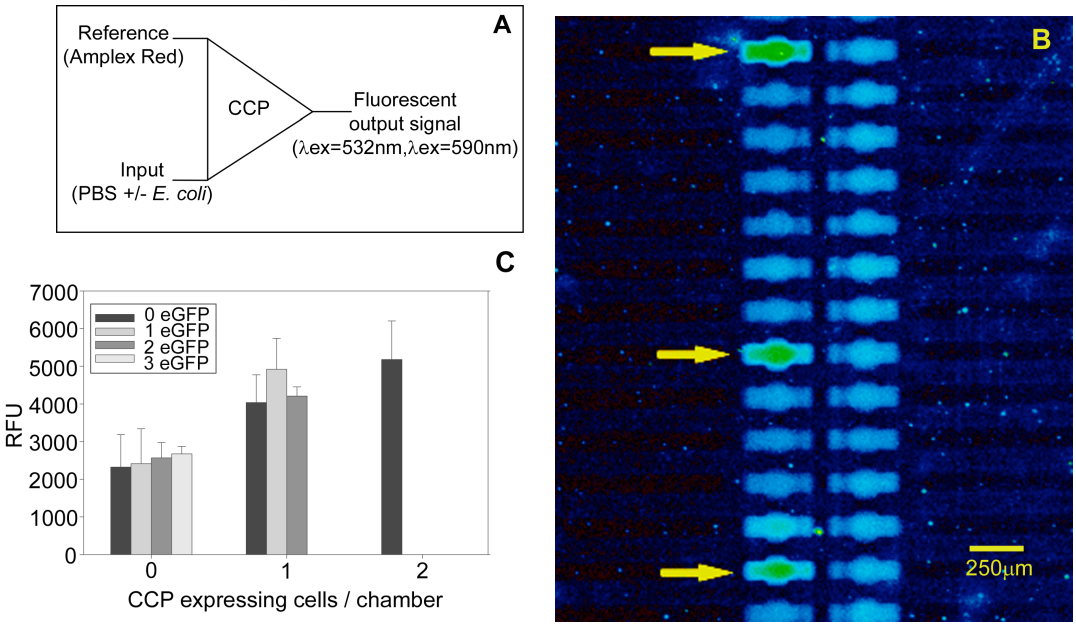



Figure 3.11: (A) Schematic diagram of the microfluidic comparator logic using an enzyme and fluorogenic substrate. When an input signal chamber contains cells expressing the enzyme CCP, nonfluorescent Amplex Red is converted to the fluorescent product, resorufin. In the absence of CCP, the output signal remains low. (B) Scanned fluorescence image of the chip in comparator mode. Left side: Dilute solution of CCP-expressing *E. coli* in sterile PBS (137 mM NaCl, 2.68 mM KCl, 10.1 mM $Na_2HPO_4$, and 1.76 mM $KH_2PO_4$ pH 7.4) after mixing reaction with Amplex Red. Arrows indicate chambers containing single cells. Chambers without cells show low fluorescence. The converted product (resorufin) is clearly visible as green signal. Right side: Uncatalyzed Amplex Red substrate. (C) A micro high-throughput screening comparator: Effect of heterogeneous mixture of eGFP-expressing control cells and CCP-expressing cells on output signal. The resorufin fluorescence measurement ($\lambda_{ex}$=532 nm, $\lambda_{em}$=590 nm) was made in individual comparator chambers containing *E. coli* cells expressing either eGFP or CCP. There is a strong increase in signal only when CCP-expressing cells are present, with little effect on the signal from eGFP-expressing cells. The vertical axis is relative fluorescence units (RFU); error bars represent one standard deviation from the median RFU.

has two inputs and a single output. The input voltages are compared to one another and the output is switched to the comparator's high gain, either to $+V_{cc}$ for $V_2 > V_1$ or $-V_{cc}$ for $V_2 < V_1$ (Figure 3.10).

To realize this experimentally, the mixing capabilities of the $\mu$MHTSC were used to introduce Amplex Red into chambers containing various numbers of cells ranging from 0 to 3. The cells were a heterogeneous mixture of two clones, one expressing eGFP, used as a negative control, and the other expressing CCP. The cells were each diluted to 1/2000 in the same test tube giving rise to a roughly 1:1 ratio of CCP to eGFP cells. This solution was introduced into the sample input port of the $\mu$MHTSC. Next an Amplex Red reaction mixture containing 10 $\mu$M Amplex Red, 88 $\mu$M $H_2O_2$ in PBS was introduced into the substrate input and kept separate from the sample channel containing the cells. After compartmentalization the barrier valve was opened and mixing allowed to proceed for 10 minutes. The chip was then scanned after 45 minutes and the results are summarized in Figure 3.11.

It was shown that the $\mu$MHTSC can be used as 256 parallel biological comparators, each chamber pair constituting one comparator element. The data showed that the presence of *E. coli* expressing CCP switches the output signal into the high state, whereas the absence of a cell or the presence of a cell expressing eGFP caused the output signal to remain at baseline; the reference side containing the Amplex Red remained at baseline levels as well, indicating that either the signal originated from CCP contained in the periplasmic space, or that the mixing time was too fast compared to the diffusion time of CCP from the left into the right chamber.

Finally, single cells may be recovered from a chamber and grown up as a colony on a Luria-Bertani (LB) agar plate. This was accomplished by using polyetheretherketone (PEEK) tubing with an inner diameter (ID) of 125 $\mu$m pushed into the collection well of the $\mu$MHTSC. This type of tubing is commonly used in high-performance liquid chromatography (HPLC) applications due to its bio-inertness and its lack of non-specific binding. A chamber containing a single cell was addressed and its contents purged to the output well, where the eGFP fluorescing cell could be seen entering the PEEK tubing. Purging was continued and a few drops were collected on an agar plate. The plate was then incubated at 37°C overnight and a colony grew where the drop containing the single cell was added to the plate. Negative control drops were also collected from chambers containing no cells and produced no colonies. Three layer devices were used to make all these experiments work. The third layer simply constituted a thin layer of 5:1 PDMS spun onto a RCA cleaned cover slip. This helped attain higher purge pressures and prevented auto-hydrolysis of Amplex Red and sticking of *E.coli* cells to glass. All-PDMS devices, due to these reasons, are generally preferred unless glass chemistry is required.

### 3.2.3 Microfluidic Memory

The microfluidic memory device (Figure 3.12) was designed by Todd Thorsen and intended to accomplish the same as the $\mu$MHTSC, but without mixing capability. Eliminating mixing and arranging the addressable chambers to be addressed in a matrix allowed the device to be very densely integrated. The memory device contains a

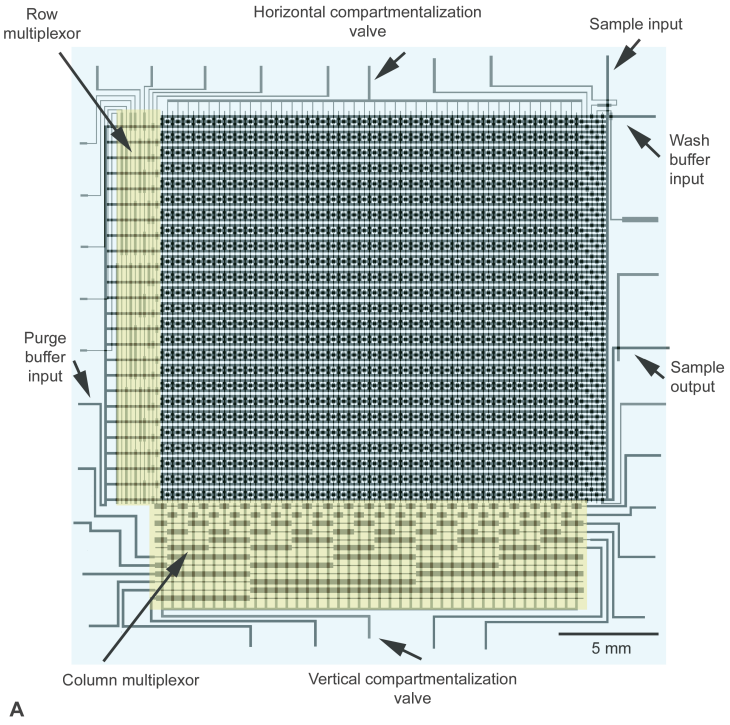



Figure 3.12: AutoCad design drawing of the microfluidic memory.

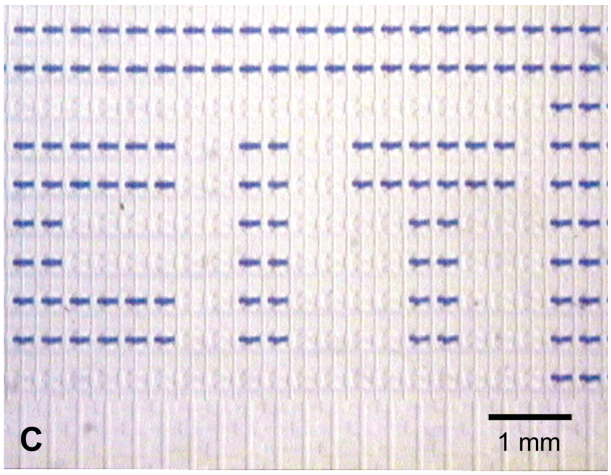



Figure 3.13: Proof of function showing the microfluidic memory acting as a display showing the letters CIT.

total of 1000 individually addressable chambers controlled by 3574 valves. Additional advances included buslines paralleling the chambers so that individual chambers could be addressed and expelled without interfering with contents of neighboring chambers. The buslines in the design allowed the matrix to remain planar, as opposed to having to address each chamber perpendicularly in the Z direction instead. The microfluidic memory device is also the first device in which advantage was taken of the control line cascade (Chapter 2.3.3), inverting the column multiplexer used to address the array.

For a proof of principle experiment, Todd addressed a subsection of the memory device and spelled out 'CIT', showing that chambers could indeed be individually addressed (Figure 3.13). Furthermore it illustrated a potential application as a low-power liquid display; useful mainly for static images, since the frame rate of the device would probably be somewhere around 1 frame per minute at best. The design could potentially be further optimized to achieve reasonable frame rates.

# Chapter 4

# *In vitro* Protein Synthesis

## 4.1 Introduction

Cell-free protein synthesis or *in vitro* transcription/translation (ITT) has been used by biologists for protein synthesis since the early 1960s [19] leading to advances in literally all aspects of molecular biology. ITT provides a convenient and rapid method for the synthesis of proteins, and to-date remains the only scalable and efficient method, since no chemical synthesis (as is available for DNA) exists for proteins. Only short peptide segments can be generated by chemical synthesis, and it has been shown that in some cases individual peptide segments may be linked by chemical ligation to form full-length functional proteins [20, 21]. Yet this method remains labor intensive, is limited to mostly small proteins on the order of 30 kDa or less, and is not immediately applicable to proteomic scale methods. ITT offers many unique features not available in either cell-based expression or chemical ligation. First, ITT uses DNA as the template from which a protein is synthesized, thus one can make use of the many methods developed for the manipulation of DNA, allowing genomic- and thus proteomic-scale application. Second, since ITT is not based on live organisms, it

allows for the synthesis of proteins otherwise toxic to cells. Site or residue-specific incorporation of non-natural amino acids [22, 23] via non-sense suppression [24, 25], spiking of modified natural tRNAs [26], and addition of engineered tRNA synthetases using non-natural amino acids as substrates for the aminoacylation of their cognate tRNA [27] are also easier to achieve in ITT systems than cells. ITT takes advantage of the features of cell based synthesis, as well as chemical ligation, and provides a transparent platform that may, with relative ease, be adjusted to the requirements of the protein being synthesized.

*In vitro* translation systems can be subclassified into various categories, depending on whether they are coupled or uncoupled and what cell source was used to produce the lysate. ITT systems are inherently coupled synthesis systems in which DNA serves as the template for synthesis and is transcribed into mRNA, followed by translation in the same reaction vessel. Uncoupled *in vitro* systems rely on mRNA produced in a separate step, which then serves as the starting material for translation. Coupled systems generally use highly processive phage RNA polymerases such as T7, T3, and SP6 for mRNA synthesis. Common lysate sources include *Escherichia coli*, wheat germ, and rabbit reticulocytes. The latter two are eukaryotic systems and generally provide better post-translational modifications, where rabbit reticulocyte lysates have a tendency to be more aggressive than wheat germ lysates. Yields generally are highest for *E. coli*-based systems, reaching milligram quantities per milliliter of reaction, followed by wheat germ and rabbit reticulocytes. The latter two generally yield submicrograms to micrograms per milliliter of reaction, with wheat germ roughly being

one to two orders of magnitude higher. A notable exception to the standard method of preparing cell-free translation systems was developed by Shimizu et. al [28]. Here the authors reconstituted a working extract system from over 100 individual components, allowing complete control over the reaction conditions. Furthermore all components are labeled with a hexa-histidine tag allowing any product to be 'reverse' purified by simply removing all translational components from the mixture rather then the product itself.

In recent years all three systems have been applied to proteome research, as well as other novel applications such as running genetic circuits in a cell-free system [29], and converting plasmid DNA arrays into protein arrays by *in situ* transcription/translation and diffusion-limited capture of product [30]. A commercially available *E. coli* lysate has successfully produced protein in 100 nL small nanowells [31]. Wheat-germ-based systems have been extensively engineered and optimized to yield milligram quantities of protein, and have been made amenable to high-throughput applications [32, 32, 33, 34]. Only rabbit-reticulocyte-based lysates have not seen rapid development, but remain commonly used in many *in vitro* based functional studies of proteins.

The following chapters describe initial experiments using cell-free systems bench-top, as well as the design and application of a novel PCR-based method for the rapid production of linear expression templates for both wheat germ and rabbit reticulocyte systems using any open reading frame as starting material. Commonly used methods such as epitope tag-based detection, purification, and post-translational biotinylation

of product, as well as enzyme-based assays, are also described.

## 4.2 ITT Systems

This section describes the various *in vitro* transcription/translation systems used. All are commercially available and are either based on *E. coli*, wheat germ, or rabbit reticulocytes. Each systems has its own characteristics, advantages, and disadvantages, which where alluded to in the previous section and will be discussed in more detail here.

### 4.2.1 Prokaryotic-Based Systems

The only prokaryotic ITT systems commercially available are based on *E. coli*. The first kit used was the RTS 100 HY kit (Roche Applied Science) with a reported protein yield of 400 $\mu$g ml$^{-1}$ and a 4–6 hour incubation period at around 30°C. RTS 100 HY is a coupled system with a T7 driven transcription step, which accepts both PCR as well as plasmid templates as input. Required 5'UTRs elements include a T7 promoter, a ribosome binding site (RBS), and a start codon. The only required element in the 3'UTR is a stop codon, though the presence of a poly(A)$_{30}$ tail and a T7 terminator enhances mRNA stability and transcription. Likewise, addition of a g10 sequence or other phage-derived leader sequences are known to also enhance protein yield. Aforementioned 5' and 3'UTR requirements hold true for most if not all coupled *E. coli* ITT systems. Initial results with the RTS 100 HY system were quite promising and led to the continued use of the kit for most studies using *E.*

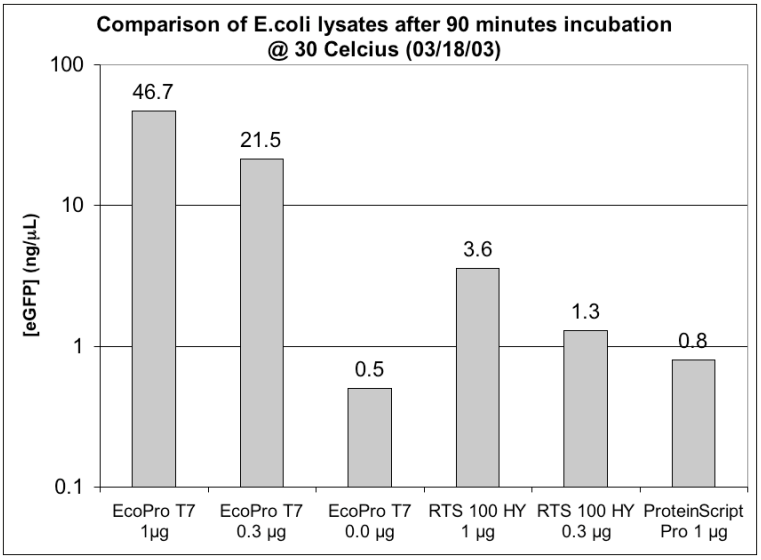



Figure 4.1: Comparison of various commercially available *E. coli* lysates. Three lysates were tested with various input concentrations of a PCR-derived linear template encoding a N-6xHis tagged version of eGFP. The most efficient lysate was EcoPro T7 (Novagen), followed by RTS 100 HY. Protein concentration was determined by measuring eGFP fluorescence.

*coli*-based systems.

A limited comparison of three kits from various vendors was performed in order to determine whether large quantitative differences exist. The test group consisted of EcoPro T7 (Novagen), RTS 100 HY (Roche Applied Science), and ProteinScript Pro (Ambion). All three reactions were spiked with varying amounts of a linear PCR template amplified from a pT7Blue vector housing a linear expression template encoding eGFP n-terminally tagged with a 6xHistidine epitope tag. The original template was generated using pEGFP (Clontech) and Roche's RTS *E coli* $His_6$ linear template kit. The resulting template was screened for expression and cloned into pT7Blue via blunt end ligation. In this case the vector serves as a host for the linear template from which future PCR amplifications may be performed. The results are

shown in Figure 4.1. eGFP concentration was measured on a fluorimeter after 90 minutes of incubation at 30°C. The results indicate that EcoPro T7 is the most efficient of the three lysates. It should be noted though that the RTS 100 HY kit continues synthesis for 4–6 hours. In all cases the final yield is one to two orders of magnitude lower than the expected yield, which is probably due to mRNA secondary structure formation in the template, discussed in more detail in Section 4.2.1.1. RTS 100 HY was selected as the standard kit used in most if not all bench-top as well as on-chip prokaryotic-based ITT reactions, mainly due to higher final yields with optimal templates. EcoPro T7 should be seen as a potential backup ITT mix in scenarios where the RTS 100 HY kit fails to perform, since EcoPro T7 seems to be less sensitive to mRNA secondary structure and produces protein more rapidly than the RTS 100 HY kit.

| | 5'UTR | start | extension | | | | | 6x Histidine | | | | | | linker | | | start eGFP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **eGFP N+3** | ACTTTAAGAAGGAGATATACC | ATG | ACC | ATG | **TCT** | **GGT** | **TCT** | | | | | | | | | | GTG | AGC | AAG | GGC | GAG |
| | | Met | Thr | Met | **Ser** | **Gly** | **Ser** | | | | | | | | | | Val | Ser | Lys | Gly | Glu |
| **eGFP N+3 modified** | ACTTTAAGAAGGAGATATACC | ATG | ACC | ATG | **ACA** | **ACA** | **ACA** | | | | | | | | | | GTG | AGC | AAG | GGC | GAG |
| | | Met | Thr | Met | **Thr** | **Thr** | **Thr** | | | | | | | | | | Val | Ser | Lys | Gly | Glu |
| **eGFP N+9** | ACTTTAAGAAGGAGATATACC | ATG | ACC | ATG | **TCT** | **GGT** | **TCT** | CAT | CAT | CAT | CAT | CAT | CAT | | | | GTG | AGC | AAG | GGC | GAG |
| | | Met | Thr | Met | **Ser** | **Gly** | **Ser** | His | His | His | His | His | His | | | | Val | Ser | Lys | Gly | Glu |
| **eGFP N+9 modified** | ACTTTAAGAAGGAGATATACC | ATG | ACC | ATG | **ACA** | **ACA** | **ACA** | CAT | CAT | CAT | CAT | CAT | CAT | | | | GTG | AGC | AAG | GGC | GAG |
| | | Met | Thr | Met | **Thr** | **Thr** | **Thr** | His | His | His | His | His | His | | | | Val | Ser | Lys | Gly | Glu |
| **eGFP N +12** | ACTTTAAGAAGGAGATATACC | ATG | ACC | ATG | **TCT** | **GGT** | **TCT** | CAT | CAT | CAT | CAT | CAT | CAT | AGC | AGC | GGC | GTG | AGC | AAG | GGC | GAG |
| | | Met | Thr | Met | **Ser** | **Gly** | **Ser** | His | His | His | His | His | His | Ser | Ser | Gly | Val | Ser | Lys | Gly | Glu |

Table 4.1: N-terminal sequence variants of eGFP. The above sequences were generated to assess which lead to efficient expression by *E. coli*-based lysates. Shown is the immediate 5'UTR including the RBS, followed by a start codon and the actual coding sequence. Sequences vary by sequence inserts as well as in the 5' extension, where differences are marked by a bold typeface.

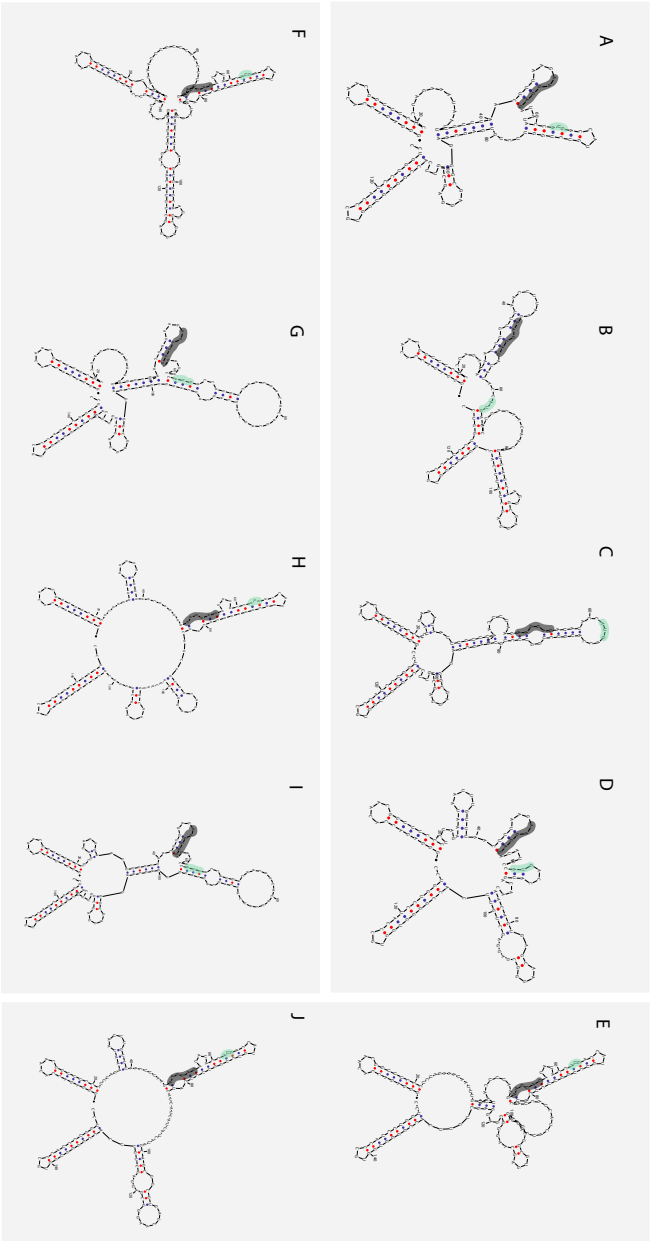



Figure 4.2: mRNA secondary structures for expression of eGFP. The structures were determined using the mFOLD server version 3.2 and 3.1 for 37°C and 30°C, respectively. Panels A and B show eGFP N+3 and eGFP N+3 modified, respectively, folded at 37°C. Panels C and D are the same as A and B but folded at 30°C. Panels F and G are the sequences of eGFP N+9 and N+9 modified, respectively, folded at 37°C and Panels H and I are the same as F and G but folded at 30°C. Panel E and J are the sequence of eGFP N+12 folded at 37°C and 30°C, respectively. The ribosome binding sites are shaded grey and the two possible start codons are green.

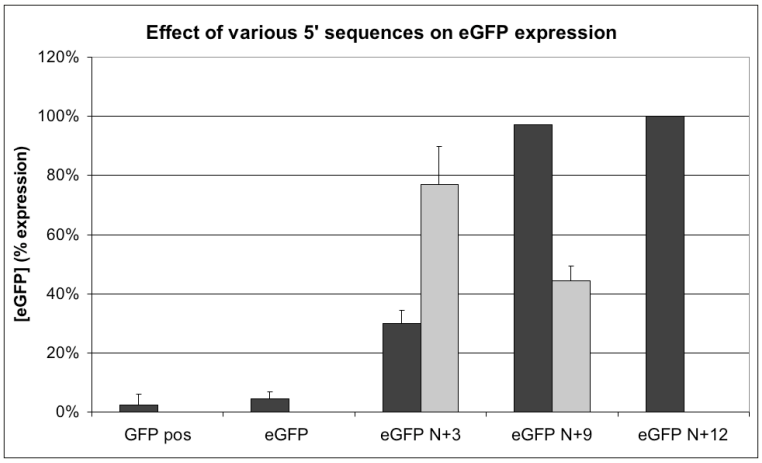



Figure 4.3: Various linear expression templates coding for eGFP with various N-termini were expressed in a RTS 100 HY ITT reaction at 30°C for 3.5 hours, and eGFP concentration was determined on a fluorimeter. The experiment was run in duplicate and the expression levels of both experiments were adjusted to the expression maximum of eGFP N+12, which were 172 ng$\mu$L$^{-1}$ and 351 ng$\mu$L$^{-1}$ in the first and second experiment, respectively. Error bars indicate one standard deviation, where applicable. Black bars are the standard sequences and the grey bars represent values obtained from the modified sequences.

#### 4.2.1.1 5' UTR mRNA Secondary Structure Optimization

The first protein expressed in this work using an *E. coli*-based ITT reaction was a N-terminally 6xHistidine tagged version of eGFP, which expressed well, yielding close to the expected yield of roughly 400 ng $\mu$L$^{-1}$. Upon removal of the tag sequence, expression dropped to around 10 - 20 ng $\mu$L$^{-1}$. The reason for this unexpected drop in expression efficiency was not immediately apparent, but ultimately traced to mRNA secondary structure differences upon testing of various N-terminal sequence additions.

To further define why and how the secondary structure of the mRNA template affects expression, a series of N-terminal deletion mutants were generated. The generic N-terminal 6x Histidine fusion is shown in Table 4.1 as sequence eGFP N+12.

Here the N-terminal sequence consists of 6x Histidine extended by 5 amino acids N-terminally and linked to the generic eGFP sequence by a 3 amino acid linker. To generate modified versions and to assess which parts of this sequence are required to bestow high expressibility on the template, the linker and the linker plus the histidine tag were removed, yielding sequences eGFP N+9 and eGFP N+12 respectively (Table 4.1). Using the mFOLD web-server [35] these three sequences were folded with version 3.2 at 37°C and version 3.1 using 30°C. The resulting sequences are shown in Figure 4.2, Panels A, C, F, H, E, and J. Furthermore in order to study the effect of specific sequence interactions rather than sequence deletions, the last 9 bases of the extension segment were mutated from TCT GGT TCT to $(ACA)_3$, and the resulting secondary RNA structures are shown in Figure 4.2, Panels, B,D,G, and I. The structures that most likely affect the expression efficiency are the ribosome binding site and the start codon, as these are the sites that need to be recognized by the ribosome in order for translation to proceed. In Figure 4.2 these structures are shaded grey and green, respectively.

Linear templates of all 5 sequence variants were generated and expressed in a RTS 100 HY ITT reaction in two separate experiments. Sequence eGFP N+12 yielded the highest quantities of eGFP of 172 ng$\mu$L$^{-}$1 and 351 ng$\mu$L$^{-}$1 in the first and second experiment respectively. These values were normalized to an expression level of 100% and all other expression levels for the remaining 4 sequences were adjusted accordingly in order to allow comparison of both experiments (see Figure 4.3). eGFP N+9 performed almost as well as the original full-length version, indicating that the 3

amino acid linker is not responsible for the drastic drop in expression. Upon removal of the 6xHisitdine in sequence eGFP N+3 the expression drops to a level of roughly 30%, and upon complete deletion of the entire N-terminal addition expression drops to 5%. Notably sequences for eGFP N+12 and N+9 all show the exact same secondary structure, regardless whether folded at 37°C or 30°C in the region starting with the ribosome binding site, forming a stable hairpin with an interior loop at the ribosome binding site and a one-base pair bulge centering on the start codon (Figure 4.2 Panels E, F, H, and J). Existence of this structure apparently confers high expressibility on the mRNA transcript. When the extension region is modified to eGFP N+9 modified, the ribosome binding site forms its own hairpin and the start codon is in a base paired region. This change in structure causes a decrease in expression from 97% to 44%. Interestingly a similar ribosome binding site structure is found in eGFP N+3, which also expresses poorly. In the eGFP N+3 structure the central thymine is bulged similar to the high-expressing sequences. It is thus likely that the structure of the ribosome binding site rather than the start codon is of primary importance. Upon modifying the eGFP N+3 sequence to eGFP N+3 modified in the extension segment expression actually increases from 30% to 77%. Looking at the structure of eGFP N+3 modified it is apparent that the ribosome binding site is no longer in a hairpin conformation but rather again contains an interior loop segment similar to the one found in the other high-expressing sequences of eGFP N+12 and eGFP N+9.

This study indicates that certain RBS secondary structures prohibit binding of the ribosome holoenzyme and thus lead to extremely low protein yields. An internal

loop of the RBS leads to efficient translation of the target sequence, whereas when the RBS is contained in a stable hairpin, protein synthesis is low. The secondary structure environment of the start codon does not seem to play as crucial a role as the RBS itself but may contribute to expression level differences. It is likely that once the ribosome holoenzyme successfully binds to the mRNA strand it progresses to perform a 2-dimensional diffusional scan along the RNA strand, effectively disrupting any secondary structures present. The only structures not consistent with this picture are Panels C and D of Figure 4.2, where the RBS secondary structures are actually reversed. It is possible that these structures have not been correctly predicted by the mFOLD server or that structures with similar energy exist that exhibit the above-described secondary structure characteristics, allowing the ribosome to bind.

Since the mRNA sequence does play a significant role in expression efficiency, every template to be expressed by *E. coli*-based ITT reactions should be checked for adverse structures using a tool such as the mFOLD server, which is easy to use and returns results instantaneously. Despite the fact the the actual coding structure was modified here in order to address expression efficiencies, it should be possible to modify the 5'UTR region instead. This would be preferred over the approach taken here, since it won't cause a change in the product to be synthesized. Likewise, silent mutations can be introduced that will disrupt mRNA secondary structure elements but won't lead to a phenotypic change.

### 4.2.2 Eukaryotic-Based Systems

Eukaryotic expression systems come in two flavors: wheat germ and rabbit reticulocyte lysates. These two systems have many common features, such as an expected yield of roughly 1-10 ng$\mu$L$^{-1}$ seen in commercially available kits (Promega). Recently wheat-germ-based systems have been optimized for increased reaction times and yields of as much as 0.1 to 2.3mg mL$^{-1}$ in a 36-hour reaction [32]. In order to extend the reaction time to 36 hours the authors had to use dialysis in order to continuously remove toxic by-products and supplement the reaction with an ATP regenerating mix. Even though dialysis is extremely useful for bench-top *in vitro* synthesis of protein, it is hard to implement on microfluidic devices, mainly due to the need of a dialysis membrane. It would be possible to implement dialysis on microfluidic devices with a planar membrane situated between two PDMS slabs similar to the devices fabricated by Ismagilov et al. [36], potentially extending protein synthesis by several hours. Extending synthesis to a day or more is harder to implement since it also requires re-introduction of new template DNA due to the degradation of DNA by DNAses present in the system. Microfluidic approaches to extending synthesis times based on discontinuous and continuous synthesis are described in Sections 6.4 and 6.5, respectively.

Eukaryotic expression systems appear to have a higher propensity for successful initiation of translation, unlike *E. coli*-based systems, which were shown to be sensitive to 5'UTR mRNA secondary structure (Section 4.2.1.1). Therefore eukaryotic systems are the prime choice for large-scale applications expressing hundreds to thou-

sands of proteins, where individual optimization of the 5' structure is prohibitive. Expression of eukaryotic sources such as human-derived cDNA clones is also to be expected to work more efficiently in homologous expression systems rather than heterologous ones.

A final characteristic of eukaryotic expression systems is that post-translational modifications are possible to certain extents. Here, rabbit reticulocyte lysates seem to be more aggressive than wheat germ lysates. Those observations are purely empirical and based on differences between apparent binding affinities of the herein-studied family of bHLH transcription factors (Chapter 8).

## 4.3 Template Generation

Generating expression-ready templates is a necessary first step in any application synthesizing protein. Cell-free systems may use either common plasmid-based templates or linear expression templates derived directly with PCR-based methods. Plasmids have the advantage of providing a more robust template, which will not be degraded in the reaction mixture as rapidly as linear templates. The pitfall of plasmids is the requirement for upstream cloning to generate the plasmid itself, which generally requires a PCR step, followed by the digestion with restriction exonucleases of both the plasmid as well as the PCR product to be cloned, and consequent ligation and transformation. This process, if optimized, requires at least two days, not including necessary sequencing controls to ascertain the correctness of the insert. These pitfalls make plasmid-based templates unfeasible for any proteomic-scale application where

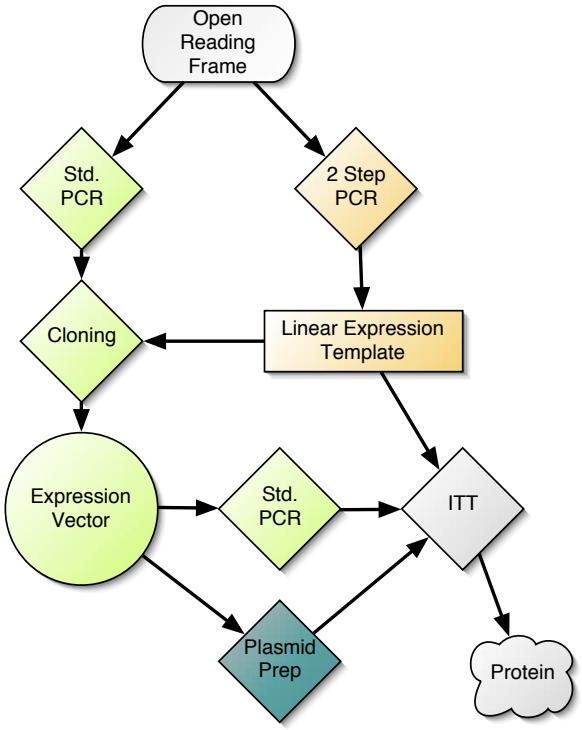



Figure 4.4: Shown here is an outline of the different paths, and their interconnectedness, which may be taken in order to arrive at a DNA template viable for ITT. All approaches require an open reading frame as starting material. The two most commonly taken approaches are shaded green and yellow. The green path is the standard cloning approach which leads through an expression plasmid and requires about 2–3 days from ORF to protein. The second, yellow path runs through the linear expression template and requires the least number of steps and amount of time, on the order of 1 day or less, to arrive at protein. One crossover between the paths occurs, where the linear expression template may be cloned into a vector and from there follow the green plasmid-based path.

hundreds to thousands of targets have to be prepared for rapid synthesis. Plasmids do lend themselves for the routine or large-scale production of a handful to a few dozen targets (see Section 4.3.1). Linear expression templates are the perfect approach for any large-scale application and will be described in detail in Section 4.3.2. Figure 4.4 summarizes in a flowchart the various paths that may be taken in order to generate viable templates for ITT reactions.

### 4.3.1 Cloning

Molecular cloning is the conventional approach for generating expression-ready templates for ITT. Most companies offer their own plasmid for optimal expression with their system. Many cloning strategies in principal start with an initial primer extension PCR on the open reading frame of choice, adding the required endonuclease restriction sites, as well as moving the ORF into the correct expression frame. A second cloning approach is based on blunt-end cloning of PCR templates. This second approach is simpler and has resulted in better yields than approaches based on sticky-end ligation. Except in the case of blunt-end cloning, restriction digests are required to linearize the plasmid as well as generate sticky ends on the linear template. Optimally the linearized plasmid should be de-phosphorylated in order to prevent re-ligation without insert. Once the linear template and the plasmid have been prepared, they are ligated and transformed into a suitable host, generally *E. coli*. The resulting plasmids should always be sequence verified, ascertaining the correctness of the insert, as well as verified for expression. As can be gleaned, the cloning approach is quite tedious and requires some expertise in the method. The resulting plasmid is quite useful for high-yield expression of templates and lends itself as a source for the facile generation of linear expression template using a simple 30-cycle TAQ-based PCR (refer to Figure 4.4). Enhanced GFP (Clontech) was cloned into the vectors pIVEX 2.3d (Roche) and pETBLue-2 (Novagen) by using standard cloning methodology. GFP was PCR amplified using primer extension PCR, adding a NcoI and SmaI restriction site on the 5' and 3' end of the GFP ORF. The resulting products

were purified using a PCR purification kit (Qiagen), followed by a double digest using NcoI and XmaCI. Note that XmaCI is an isoschizomer of SmaI, creating a sticky end rather than a blunt end. The pIVEX 2.3d vector needed to be linearized with the same two restriction enzymes, whereas pETBlue-2 is already in a linear form accepting blunt-end PCR templates. Upon complete digestion of the vector and insert, the linearized and cut products were separated on an agarose gel and selected bands were gel purified (Gel Purification kit, Qiagen). Next the inserts were ligated into their respective vectors using Novagen's Blunt End Ligatin kit and the resulting vectors were transformed into chemically competent cells, strain NovaBlue (Novagen), and spread plated. Resulting colonies were confirmed for the correct size by PCR screening. Colonies containing vectors of correct size were picked and grown overnight in LB and the vectors were prepared for a control digest. The above-described ligation procedure performed well compared to diffculties with using the recommended endonuclease SmaI. Use of Novagen's blunt-end cloning kit to perform the ligation further enhanced the yield and streamlined the process.

In general blunt-end cloning is to be preferred over sticky-end cloning since it does not require processing of the PCR product nor the vector. A variety of expression-ready clones were generated using pT7Blue as a backbone. Constructs include eGFP variants carrying different N- and C-terminal epitope tags, the histones H2A, H2B, H3, and H4, and the caspases 4, 6, and 9.

For eukaryotic-based expression using a vector, pTNT (Promega) was modified to contain a blunt-end cloning site for Pmi I, an N-terminal extension, and a C-terminal

AviTag, creating vector pTNT+.

Blunt-end cloning is not universally applicable, in which case it is necessary to use a sticky-end cloning strategy similar to the one described above. It is quite useful to combine blunt-end cloning with the PCR-based method described in Section 4.3.2, providing a convenient method for generating a stable template source with minimal amount of effort, in addition to being completely modular, as entailed by the PCR approach.

### 4.3.2 PCR-Based Approach

PCR provides the quickest route to viable expression templates. Roche commercialized a system based on a multi-step PCR generating linear templates for most of its ITT kits that it sells. Sawasaki et al. developed a PCR-based approach for wheat-germ-based expression [32]. PCR-based approaches are not only extremely fast, yielding product within a day, but are also highly modular and can accept a variety of input templates ranging from cDNA clones and other plasmid-based ORFs to genomic DNA. One of the main drawbacks to PCR-based approaches is the formation of primer dimers, which is a considerable problem in the Roche based system. If the PCR is not optimized, additional purification steps, such as tedious and time consuming gel purification steps, are required.

A novel PCR-based approach was designed in order to address the following points: 1) only commercially available single-stranded oligomers should be required, limiting individual oligo lengths to 100–135 bp, 2) the PCR should be highly modular, 3) a

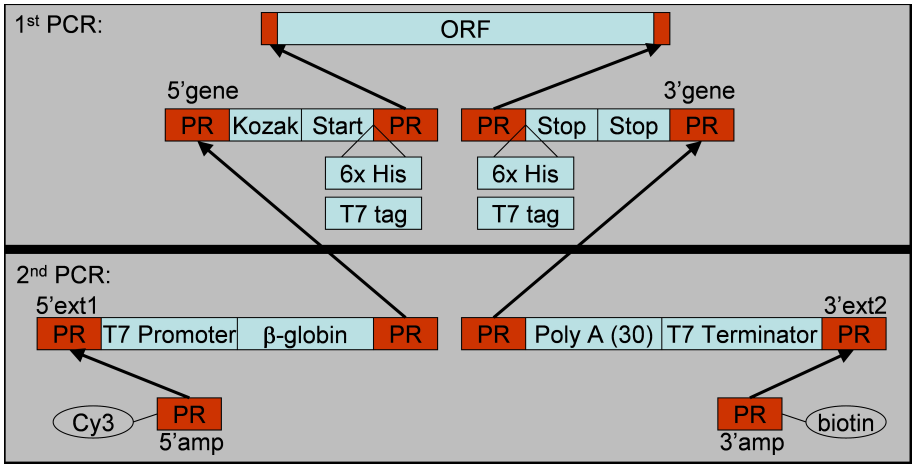



Figure 4.5: 2 step PCR method for generating linear expression templates

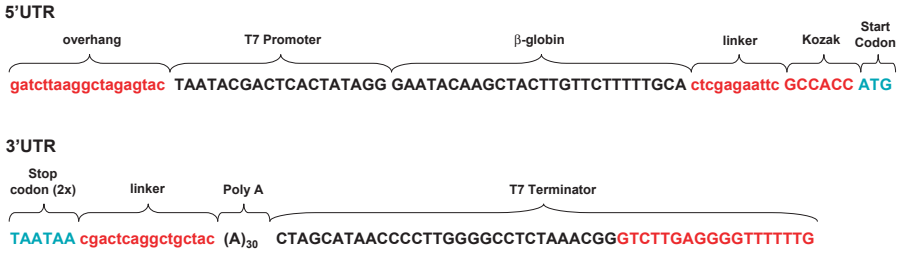



Figure 4.6: 5' and 3' UTR sequences added by the 2-step PCR method. All regions are annotated and all priming sequences are in red. The start and stop codons are colored green. The entire 5' and 3' UTRs are added by the 5'extension and 3'extension primers, respectively, except for the start and stop codons.

minimal number of spin column based purifications should be required, 4) it should be possible to tag the resulting linear template with moieties such as fluorophores and biotin, and 5) if possible be a single-tube, single-step reaction. The resulting design is illustrated in Figure 4.5 and was intended to be used in either wheat-germ- or rabbit-reticulocyte-based coupled ITT systems. It was originally designed as a one-tube, one-step reaction, with the primer melting temperatures staggered so that the melting temperature could be ramped down as the PCR progressed, in order to induce priming of sequences at various times in the reaction and thus sequentially stepping through the reaction scheme. In praxis the one-tube reaction did return an unacceptable amount of incorrectly primed by-products and would have required considerable optimization of melting temperatures and primer concentrations. The final and currently used approach is a two-step approach. In the first step two ORF-specific primers are used, which contain overhang sequences allowing downstream priming with the primers 5'extension 1 and 3'extension 2. It is also possible to conveniently add N- and C-terminal epitope tags to the ORF using the gene-specific primers. One simply adds the desired sequence directly downstream or upstream of the start and stop codon, respectively. Any ORF may be used as template in this first PCR step, including yeast or bacterial genomic DNA, bacterial colonies, or purified plasmids. The first PCR step may be run for 10 to 30 cycles and the resulting product is generally purified using a spin column (PCR purification kit, Qiagen); the product may then be stored at -20°C. Several dozen second-step PCRs may be run on the quantity obtained in the first step, making it a very efficient and convenient method.

The second PCR step consists of two sub-steps, which take place in the same reaction vessel. Here, in the first reaction 10 cycles are run in the presence of the 5' extension 1 and 3'extension 2 primers, which prime the overhang regions added to the ORF in the first reaction. These primers add the entire 5' and 3' UTRs (see Figure 4.6 for sequence details). The products of this sub-step in principal are functional linear expression templates. In order to reduce the chance of primer dimer formation and to add the ability to easily label the linear expression templates with moieties on the 5' and 3' ends, a second set of short primers is used. These function as amplification primers, priming the very ends of the linear expression templates. These primers are also much cheaper to label with fluorophores and other moieties, and may be used at will, depending on which moieties are needed.

The two-step PCR method was also modified, by redesigning the 5' extension primer, to be used in *E.coli*-based ITT systems. Two new extension primers of slightly varying sequence were designed and tested with Gateway Entry clones (Invitrogen) harboring ORFs from *S. pneumoniae* (TIGR Pathogen Functional Genomics Resource Center). The extension primers were based on sequences obtained from Kim (Schwartz Lab Stanford U., private communication) and sequences used in pIVEX vectors (Roche) (refer to Appendix B for sequence details). Preliminary results indicated that the sequences were equally efficient in producing protein.

One concern with using PCR-based approaches for template generation is the accumulation of point mutations due to the DNA polymerase error rate. This is a concern as the cycle number for the PCR may be as high as 65 cycles. In order

to assess the impact of high cycle numbers on the viability of the protein product, a 45-cycle- and a 65-cycle-derived linear template were transcribed and translated. The synthesized transcription factor showed the same affinity to its target sequence regardless of the number of cycles used to generate the template. Point mutations undoubtedly accumulate in the resulting pool of linear templates, but only a subset will lead to phenotypic mutations, and only a subset of those will have detrimental effects on protein function.

To further asses the fidelity of the PCR approach, 7 linear templates generated by a 70-cycle, two-step PCR were bulk sequenced. The resulting sequences (Appendix D) indicate that a 70 cycle PCR using High-Fidelity Polymerase (Expand High Fidelity PCR, Roche) does not have adverse affects on the fidelity of the resulting bulk sequence. This was established by performing a BLAST search (discontiguous megablast) with the obtained sequences. The alignments show that 6 out of the 7 sequenced ORFs align perfectly with their match in the first 300–400 bases of the sequence. Below 400 bps mismatches and gaps are visible, but are due to degradation of the quality of the sequencing run. The only case that showed discrepancies between the sequenced results and the BLAST subject was for C-Myc, which returned v-Myc as the resulting subject, and is the reason for the observed differences. In the case of C-Myc $\Delta$N249, the sequence again perfectly aligned. This strongly suggests that there are no contaminating carryovers of misread sequences in the bulk PCR product. Therefore, since the bulk product seems to be identical to the original sequence, the resulting bulk protein should also resemble the intended product. This method of

bulk sequencing of course does not rule out the possibility that many random point mutations accumulate, but it shows that none of these mutations dominate the PCR after 70 cycles. Again, in order to establish the exact distribution and occurrence of these point mutations it is necessary to sequence individual clones.

# Chapter 5

# Surface Chemistries

## 5.1 Introduction

Surface chemistry is a central component of all on-chip biochemistry, not only due to the high surface-to-volume ratio intrinsic to microfluidic devices, but also because it is necessary for performing pull-down-based molecular assay. In order to perform on-chip protein synthesis from linear expression templates, described in more detail in Chapter 6, and to perform interaction assays, it is necessary to generate specific surface chemistries by derivatizing the surface with molecules that specifically bind other molecules such as streptavidin or antibodies. In order to prevent non-specific binding, surfaces may be treated with molecules such as BSA and PEG.

The two available surfaces to work with are glass and PDMS. Glass was chosen as the preferred surface simply because it has been used for many years as a starting point of various chemistries. Glass surfaces derivatized with functional groups such as amine, aldehyde, and epoxy are also commercially available. Additionally, glass has been widely used as a substrate for DNA arrays, which are being used here as a way of introducing molecular information onto microfluidic devices (Chapter 6). Glass

may also be coated with metals by sputtering or vapor deposition. PDMS may serve as a starting point for most if not all above-mentioned methods as well, but generally has not been optimized and perfected to the point that glass chemistries have.

Glass chemistries have been widely applied to biological molecules, making it a necessity that they be performed at near-neutral pH, close to room temperature, and in aqueous phase in order to prevent macromolecules from irreversibly denaturing and thus loosing their function; yet another beneficial aspect to choosing glass over PDMS chemistry.

The following chapter describes various methods used to arrive at functional surfaces, either on- or off-chip, as well as more detailed information on how the surfaces are generated. The pros and cons of each method and their integration with the rest of the microfluidic device are also discussed.

## 5.2 Building Surfaces

In this study, in order to build specific surface chemistries, only a handful of moieties and reactive groups were used, which nonetheless could be combined in a number of ways. Amine reactive chemistry is the predominant method used to generate covalent linkages. Primary amine groups react by nucleophilic attack with a variety of groups such as epoxy and succinimidyl esters. These reactions occur at slightly basic pH, required to protonate the primary amine, and at room temperature. Most primary derivatization steps are based on an amine reactive group. The second most commonly used chemistry is binding of biotin by the avidin family of proteins such as

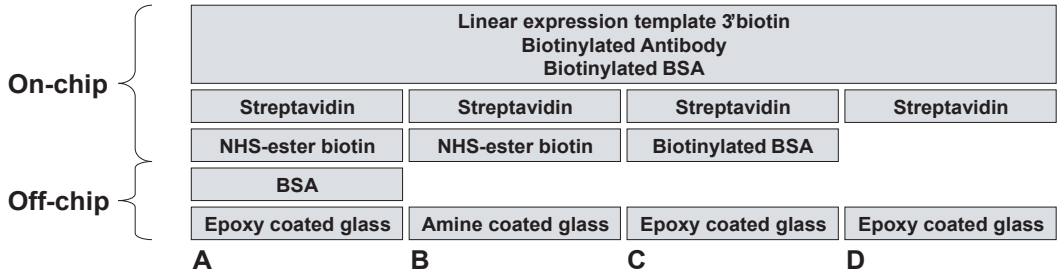



Figure 5.1: Shown here are four commonly used surface chemistries ultimately converging on a streptavidin monolayer which can bind a variety of other functional biotinylated molecules. Panel A, C, and D are chemistries based on epoxy- and Panel B is based on amine-coated slides. In Panel A the first derivatization step is taking place off-chip. All other steps are performed on-chip using continuous flow of reagents over the surface.

avidin, streptavidin, and neutravidin. Biotin binding to these molecules is incredibly persistent with an affinity constant on the order of $10^{15}$ $M^{-1}$ and a half-life of almost a day. Not only do streptavidin and neutravidin have high affinity for biotin, but they also contain four binding sites, allowing streptavidin to bind up to four ligands. This is commonly taken advantage of by using streptavidin as a linker between two or more different ligands, effectively linking streptavidin to a surface-bound ligand followed by binding a second ligand to streptavidin.

The general strategy for generating surface chemistries useful for protein pull-down and interaction studies relies on activating a glass surface with a biotin moiety via amine reactive chemistry. Figure 5.1 summarizes approaches taken here to arrive at a streptavidin monolayer to which any biotinylated molecule may be attached. Panels A, C, and D show epoxy-based approaches. Here the first step in all cases consists of attaching proteins via their primary amine groups to the surface epoxy groups, creating a protein monolayer. This can either be done off-chip (Panel A)

in a PBS buffered solution (Appendix E.4.1), or on-chip (Panels C and D). In the case of off-chip derivatization it is necessary to bond the PDMS device to the protein monolayer. This, even when done at 80°C overnight, is not a strong bond and thus does not permit flow pressures much higher than 5 psi. Additionally it is unlikely that the protein remains in a native state, due to the high temperature and lack of bulk water, and thus it should not be considered folded for further downstream steps. Therefore, use of streptavidin or other functional molecules is not applicable, unless it is possible to refold the proteins on-chip. In the case of BSA it is not necessary for the protein to be in its native conformation, since it only serves as a passivization agent and the second derivatization step is based on a succinimidyl ester reaction with primary amines on BSA. Coating the entire epoxy glass slide in a BSA bath has the advantage that the entire surface is being passivated, which is useful to prevent other molecules such as DNA and protein from binding to the glass surface. Once the device has been bonded to the BSA-coated glass surface it may be activated on-chip using a biotin-succinimidyl ester functional group, covalently linking biotin to all available amine groups on the surface-bound BSA. Biotin in turn may bind streptavidin, allowing further molecules to be added at will. Instead of passivating the entire epoxy surface with BSA, it is possible to bond a PDMS device directly to the surface. This bonding takes place at room temperature, but works better at 40°C. This bond is acceptably strong, consistently holding 5–6 psi on the flow layer. More importantly, it is the only bond that forms at room temperature and thus is extremely useful when DNA, protein, or cell arrays are to be bonded to a PDMS device. No other

bonding method works at room temperature, except for plasma bonding, which is not applicable to arrays for two reasons: it destroys any molecules on the array if the array itself needs to be treated (which is likely the case), and it is prohibitive to obtain exact alignment since bonding is instantaneous (preventing realignment and thus making the process a one-shot deal). Once the chip has been bonded to the epoxy slide, which can be accomplished in 1–2 hours at 40°C, further chemistry can be performed by attaching molecules containing primary amines to the epoxy groups. The first approach (Panel C) parallels the approach taken before by attaching a BSA protein to the surface. In this instance it is possible to attach a biotinylated BSA protein directly to the surface, streamlining the approach by circumventing the biotinylation step. It was not possible to directly derivatize the entire epoxy slide directly with biotinylated BSA in the previous approach since the required 80°C bonding step has detrimental effects on the biotin moiety either destroying it or destroying the linkage to the carrier molecule. The third epoxy-based approach simply attaches streptavidin, or any other protein, directly to the epoxy surface. This approach, even though quick and straightforward, is not optimal, since background binding is higher due to the lack of passivation agents such as BSA. Secondly the streptavidin or antibody molecules will attach in random orientations to the surface, rendering a fraction of the active sites inaccessible and reducing functional surface density.

A fourth approach is based on amine-coated slides rather than epoxy substrates. Here the chip is bonded to the glass slide prior to any surface chemistry treatment. The first on-chip derivatization step consists of covalently linking a biotin moiety to

the amine groups via a succinimidyl ester. The biotin monolayer is then coated with streptavidin. This approach works consistently, but has the same drawbacks as the method described in Panel B, in that there is no passivation layer protecting the surface, causing increased non-specific binding. Furthermore any un-reacted amine groups will exhibit a positive charge, increasing non-specific binding of net negatively charged molecules such as DNA.

Once a streptavidin monolayer has been established, it is quite easy to attach any molecule of interest to this surface. Here, either biotinylated antibodies or biotinylated dsDNA molecules are generally used. Biotinylated antibodies are commercially available (Qiagen, AbCam) and biotinylated dsDNA molecules can easily be synthesized in-house using commercially available biotinylated oligos and PCR (Section 4.3.2). Additionally, by using the freestanding membrane described in Section 2.3.4, it is possible to generate defined spots consisting of one type of molecule surrounded by a second type of molecule achieved by stepwise derivatization.

Figure 5.2 describes one approach to generating a surface that contains linear expression templates from which protein is synthesized *in situ* and a defined circular region of an anti-penta-histidine antibody used to pull down the synthesized protein. Panel A shows a schematic of the final surface chemistry used in the experiment. The initial layers are exactly the same as described in Panel A of Figure 5.1. The functional molecules linked to the streptavidin monolayer in this case are biotinylated linear expression templates coding for a bHLH transcription factor N-terminally tagged with a 6x histidine epitope tag. A biotinylated antibody is also deposited on the device.

The linear expression templates and antibodies are deposited in two consecutive steps. First the linear template is deposited while the freestanding membrane is closed (the membrane-protected area indicates which section of the device is protected. Refer to Section 2.3.4 for more detail on the freestanding membrane and Section 7.4 for MITOMI). Once the non-protected area is covered with linear expression template, a second solution of biotinylated BSA is flown over the same area with the freestanding membrane remaining closed. This step ensures that all still-available streptavidin binding sites are passivated. The biotinylated BSA is then cleared out during a PBS wash step and the membrane is opened, de-protecting the as of yet underivatized areas. Now a solution of biotinylated antibody is flown over the surface causing the antibody to deposit at the previously protected sites. Once the surface chemistry is set, protein is synthesized *in situ* and pulled down via an epitope tag by the surface-linked antibody. Any secondary molecule, in this case short dsDNA carrying a transcription factor recognition site, is co-pulled down (Panel B). Now MITOMI 2.3.4 may be performed by once again bringing the freestanding membrane into contact with the surface, effectively trapping any surface-bound molecules (Panel C). Unbound material can now be washed away without loss of bound material (Panel D) allowing for the sensitive detection of the bound material.

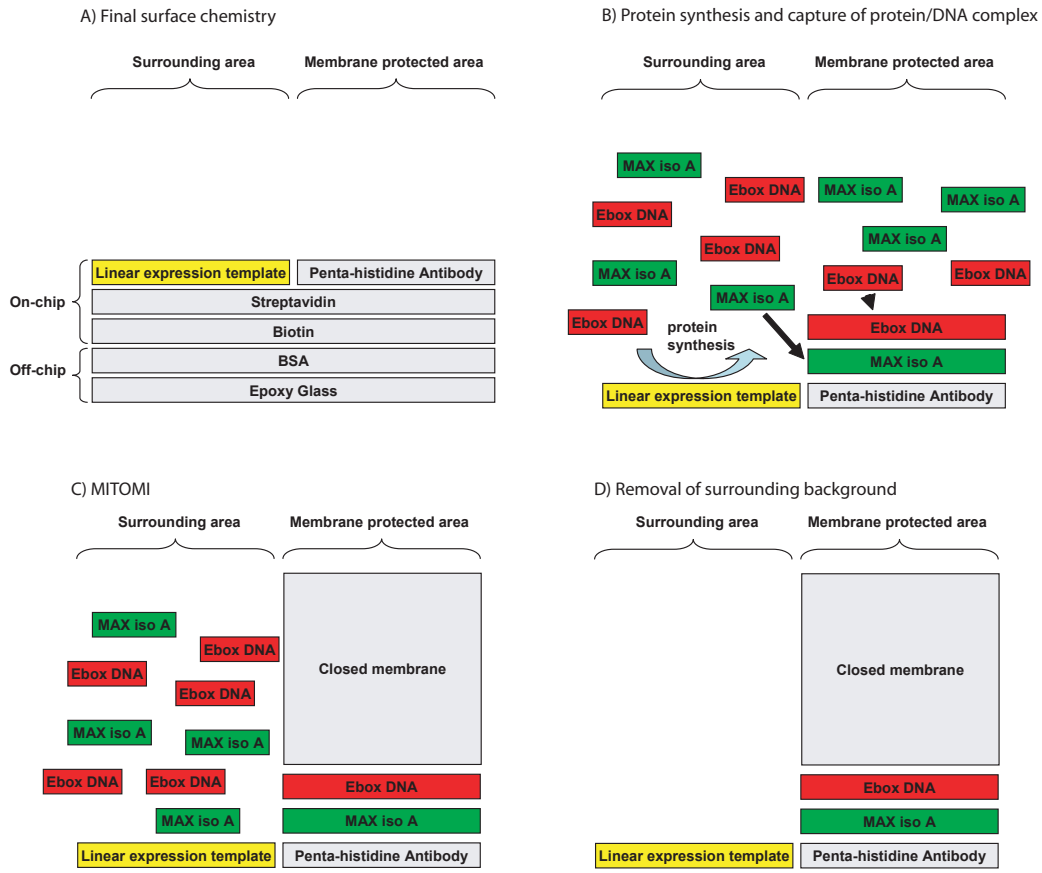



Figure 5.2: This figure shows a schematic of the surface chemistry that was generated on the device as well as the process of protein synthesis, capture and MITOMI. Colored boxes indicate fluorescently labeled molecules: green = fluorescein, yellow=Cy3, and red=Cy5. Panel A shows the final surface chemistry just prior to introduction of the *in vitro* transcription/translation reagents. Each grey block represents a monolayer consisting of the indicated molecule. Panel B describes the process of protein synthesis using the deposited linear expression templates. The synthesized MAX iso A protein diffuses to the antibody-coated surface and is pulled down via its N-terminal 6xHistidine tag. The free Ebox DNA molecules, introduced with the ITT mix, are recognized by MAX iso A and likewise pulled down to the surface. In Panel C MITOMI is performed by closure of the free-standing membrane, trapping any bound material and expunging any unbound material (corresponding image: Figure 2.5 Panel B). Panel D shows the final state of the device after the last PBS wash removes any unbound material from the adjacent material (corresponding image: Figure 2.5 Panel C).

# Chapter 6

# On-chip Protein Synthesis

## 6.1 Introduction

The previous chapter described in detail the use and applicability of ITT to large-scale protein biochemistry in providing a fast and facile path towards protein. ITT has the potential of outperforming classical approaches to generating protein by homologous or heterologous expression in various cell lines by circumventing all necessary cloning and culturing steps. But in order for ITT to be applicable to proteomics, two necessary requirements have to be met: first many thousands of proteins will have to be synthesized in parallel and tested in binary fashion against one another; and second, the first requirement has to be met on scales small enough to be economically viable, as commercially available ITT reactions remain rather expensive (0.2 \$/$\mu$L versus 0.05 \$/$\mu$L for a PCR reaction). Microfluidics presents a solution to both of the above requirements in that it allows for massive parallelization of extremely small reactions with volumes on the order of nano- to picoliters.

Nonetheless, novel methods needed to be developed to address the world-to-chip interface problem, as a single device must be programmed with thousands of expres-

sion templates coding for the proteins to be synthesized. Programming of a microfluidic device with thousands of unique solutions is a non-trivial task. Current methods allow for the introduction of tens to hundreds of solutions onto a single device, which is still one to two orders of magnitude below the projected goal of thousands of unique proteins (required for most prokaryotic proteomes). Figure 6.2 presents two solutions to the programming dilemma. The first approach is more classical and based on the flow deposition of linear expression templates. The second, more powerful and modular, approach is based on using spotted microarrays for device programming. The final sections of the chapter describe how protein can be synthesized *in situ* on devices using either batch (Section 6.3), discontinuous (Section 6.4), or continuous synthesis (Section 6.5).

## 6.2 Programming Devices with DNA

### 6.2.1 Flow Deposition

Flow deposition generally entails derivatizing at least one surface of a microfluidic channel with a linear expression template. This is most easily accomplished using a glass/streptavidin-based surface chemistry (Figure 5.2) in combination with a 3' biotinylated linear expression template generated via PCR (Figures 4.5 and 4.6). The linear expression templates are flown over the streptavidin surface, where they are captured and accumulate until the surface is saturated. Due to the slow dissociation of biotin from streptavidin, these surfaces stay viable for an extended period of time,

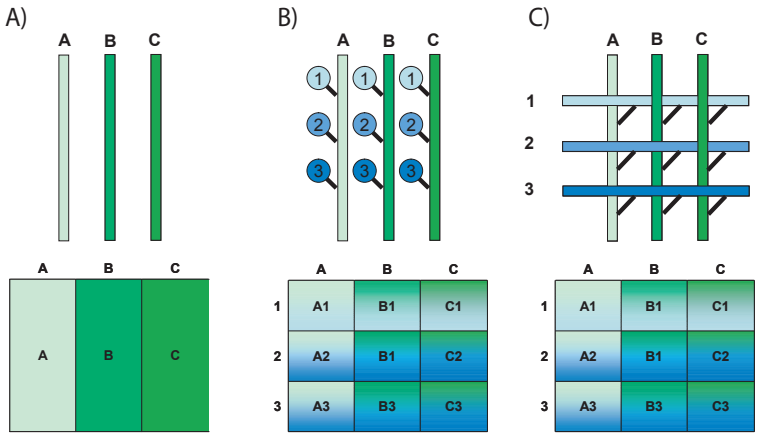



Figure 6.1: Shown are three approaches for programming devices by surface deposition. (A) The simplest of the three approaches involves programming of a flow with a unique template. (B) Hybrid approaches are also possible that combine flow deposition with programming using microarrays. (C) A true combinatoric array can be established by intersecting the vertical flow channels, as in (A), with horizontal channels programmed with a second set of templates.

which is an important aspect for discontinuous as well as continuous ITT. Figure 6.1 describes three possible flow deposition schemes.

The first scheme, depicted in Panel A, is the most straightforward method, in which parallel microfluidic channels are each derivatized with a linear expression template coding for different proteins. The second approach, shown in Panel B, is actually a hybrid between flow deposition and spot deposition described in 6.2.2. The third and final scheme, shown in Panel C consists of two sets of parallel channels running perpendicular to one another. Each set of parallel channels is derivatized with its own unique set of linear expression templates (n[A,B,C] and m[1,2,3]). Crossing the channels allows one to test a complete interaction matrix of all possible combinations of n x m. A chip design successfully fabricated and tested is shown in C.3. This device can test 12 x 12, or 144 different protein combinations. This combinatoric

approach is especially useful when a limited set of proteins are to be tested in all possible combinations. A possible application includes testing molecular complexes, which generally consist of tens of proteins. The reason that the approach is limited to rather small numbers for n and m is that each linear expression template requires a dedicated fluidic input port, each of which has a rather large footprint and is tedious to interconnect. A strength of the approach lies in the fact that the linear expression templates are surface immobilized, allowing for complete buffer exchange with minimal template loss. The immediate benefit of this lies in the ability to run consecutive protein synthesis reactions. This is useful as ITT reactions are generally short lived, on the order of 1–2 hours, and therefore produce only limited amounts of protein. By being able to repeat the synthesis x number of times (where x is only limited by the retention of functional linear expression templates on the surface) one can accumulate large quantities of protein (described in detail in Section 6.4 and Section 6.5). Another technical limitation to flow deposition is based on mass transfer from solution to the surface. Mass transfer can be in a limiting regime so that in long channels with high resistance and low flow velocities the time for complete derivatization can become rather long. Or, in a limited amount of time it might only be possible to saturate the beginning of the channel while the exit remains underivatized. The problem of mass transfer is of course also primarily dependent on the concentration of the material to be deposited; this affects mainly linear expression templates as other molecules used in derivatization such as streptavidin and BSA-biotin, can be obtained in large quantities and at high concentrations.

A solution to the mass transfer problem is shown in the layout of device DTPAx8 (Figure C.17) where rows are addressed in parallel rather than serially. Additionally, a resistance equalizer up and downstream of the parallel flow channels assures that the flow velocities through each row of channels is of the same magnitude. Taken together such a design allows for the rapid and even derivatization of the entire surface of the device, even with material that is present in limited concentrations.

### 6.2.2 Microarrays

Spotted microarrays were first described in 1995 by Brown et al. [37]. Microarrays are fabricated using small dimension quill pens, which pick up liquid from a reservoir, generally a multiwell plate, and deposit drops with diameters ranging from tens to hundreds of microns. The resulting microarrays, with up to tens of thousands of individual spots, are uniquely suited for integration with microfluidic devices as the length scales of the spot diameters and microfluidic channels are well matched.

To integrate a spotted microarray with a microfluidic device, the spots on the array need to be aligned to features on the microfluidic device. This is easily accomplished and requires about the same degree of dexterity as fabricating the microfluidic devices themselves. One important aspect is that the spots need to be visible, which generally is the case since most solutions that are spotted will evaporate leaving a crystalline residue on the surface. If this is not the case, as for some low-concentration DNA solutions in water or TE buffer, salts or BSA may be added to the solution (the choice depending on downstream compatibility).

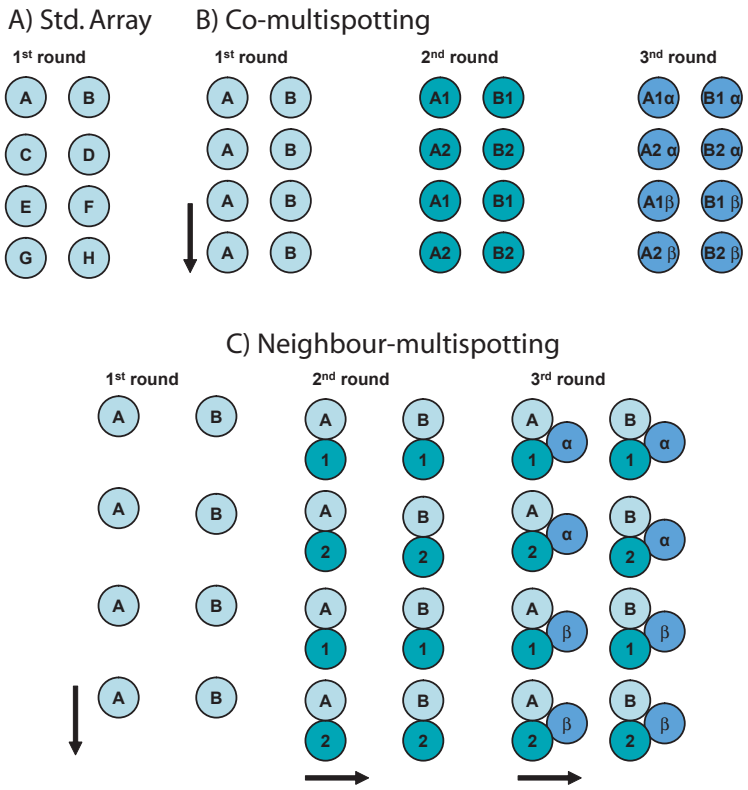



Figure 6.2: Schematic representations of 3 possible approaches to generating spotted microarrays. Panel A shows the standard approach where each spot contains a unique solution (here A–H). Elaborating on this scheme it is possible to co-spot (Panel B) or neighbor spot (Panel C) additional solutions (here A–B, 1–2, and $\alpha - \beta$) either directly on top of (co-spotting) or in proximity (neighbor spotting) of an existing spot.

In all current chip designs using microarrays for device programming, each individual microarray spot is aligned to a chamber on the microfluidic device. These chambers make alignment easy and protect the spot from solvating and suffering diffusional loss during on-chip processing. Since spots are aligned to a chamber, which confines the spot, it becomes possible to co-spot multiple solutions onto the same microarray spot and have them mix on-chip, allowing for complex combinatoric experiments. Figure 6.2 summarizes the three most common and useful approaches to generating spotted microarrays for use with microfluidic devices. The first approach

(Panel A) is a standard spotted microarray where each spot is unique and homogeneous. More useful approaches include co-multispotting (Panel B) and neighbor-multispotting (Panel C). Both of the latter approaches generate spots or clusters of spots containing more than one solution. The co-multispotting approach is the more space efficient of the two, as all the solutions are spotted on top of one another. Co-multispotting also allows for concentration of samples, as repeated deposition of the same solution will result in accumulation of material on the surface. One problem with co-multispotting is the possibility of contamination, and care must be taken to clean the quill pen between spots, which drastically adds to the time required for generating an array. The neighbor-multispotting approach solves the problem of contamination since no spot comes into contact with any other spot. But neighbor spotting does require a considerably larger footprint, limiting the number of reactions that may be run per device.

The use of spotted microarrays is extremely modular, as any soluble substance may be spotted and thus used to program the device. The only requirement is that the spotted solution is compatible with the PDMS device itself. Furthermore, colloidal suspensions, as well as pelleted cellular material, may also be spotted, as demonstrated in Chapter 10, where a library of yeast strains is arrayed.

## 6.3 Batch Synthesis

The simplest approach to protein synthesis is based on porting the standard bench-top ITT reaction onto the microfluidic device. Generally, the expression template is

present on the device from which protein is being synthesized. The linear expression template may also be added directly to the ITT mixture before introduction onto the device. This approach is used in later chapters for transcription factor–DNA binding energy landscape determination, where the device is programmed with target DNA rather than expression templates. Premixing the expression templates into the ITT reaction assures a homogeneous distribution of template, and thus protein expression.

The next two sections describe how protein can be synthesized in batch format either from flow-deposited or spotted DNA. Batch synthesis implies that only one synthesis reaction is run from start to finish. The protein yield is therefore limited by the efficiency of the ITT systems, which in batch mode generally can synthesize protein for up to 1–2 hours.

### 6.3.1 Deposited DNA

Flow-deposited DNA is well suited for high-yield on-chip protein synthesis. One reason for this is that flow-deposited DNA is considerably more concentrated than can be achieved for solution phase expression templates. This considerable increase in DNA concentration is due to the high surface-to-volume ratio in microfluidic devices. In most cases the ratio is generally 1000 $\mu m^2$ for every 100 pL. With such a ratio, expression template concentrations of 1 $\mu$M can easily be achieved. This compares favorably to normal solution phase concentrations, which are in the nanomolar range at best.

A second advantage of flow-deposited expression templates lies in the fact that

discontinuous batch synthesis as well as continuous synthesis are possible, since the expression templates are solid phase, making buffer exchanges possible. The capability to run several batches or continuous synthesis with constant ITT exchange results in drastic increases in protein yield and, coupled to an appropriate protein sink, should find broad application in the large-scale protein synthesis using on-chip ITT.

### 6.3.2 Spotted DNA

Batch synthesis of spotted expression templates is the method of choice for generating small quantities of a large number of different proteins. Current devices can synthesize thousands of different proteins from an appropriate ORF library. As an early proof-of-principle experiment, eGFP was expressed from spotted expression templates using *E.coli* ITT. A 400-chamber device (Figure 6.3) was programmed with 100 spots of expression templates coding for eGFP. Only 1/4 of the device was programmed to ascertain that there is no cross-contamination of adjacent chambers. The device was then loaded with *E.coli* ITT mix and incubated for 7 hours with fluorescence scans being taken every few hours. The results show that eGFP synthesized well in most if not all chambers that were programmed, while all negative chambers remained dark. The eGFP concentration was so high that its fluorescence could be observed on a microscope with simultaneous bright field and fluorescence illumination (Figure 6.3B). The time trace indicates that synthesis seizes after roughly two hours and that the synthesized protein is stable over several hours. Furthermore, synthesis is

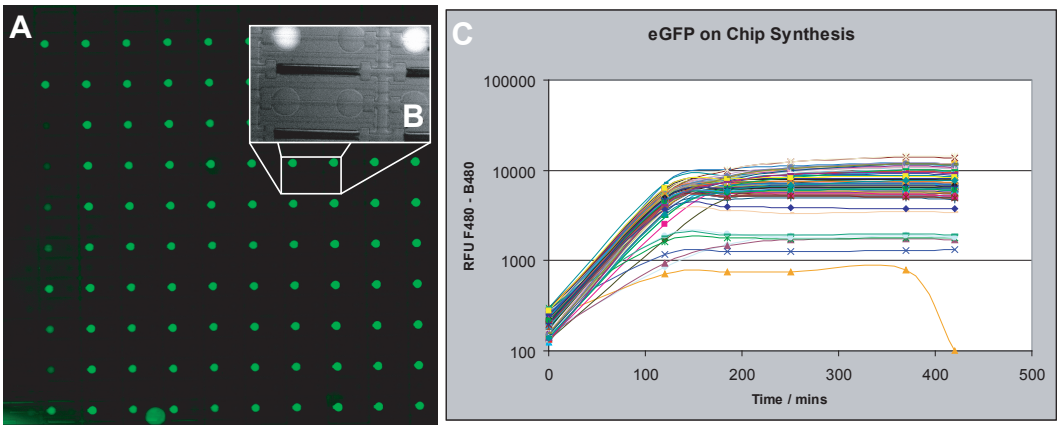



Figure 6.3: eGFP was synthesized from spotted expression templates. (A) Each template was spotted in every 4th chamber and incubated for several hours with *E.coli* ITT. Fluorescence was measured on a microarray scanner. (B) The resulting protein concentration was so high that fluorescence could be observed while in brightfield mode on an epifluorescent microscope. (C) Time trace of eGFP fluorescence

quite uniform across the device, with slight variation possibly arising from different quantities of expression template being deposited in each well.

## 6.4 Discontinuous Synthesis

Discontinuous batch synthesis can be accomplished from a flow-deposited expression template. Running several consecutive batches of synthesis allows for a steady accumulation of protein. Discontinuous synthesis should be coupled with a protein sink, such as an off-chip affinity column that selectively filters out the synthesized protein and thus concentrates it. eGFP was synthesized by discontinuous batch synthesis from flow-deposited linear expression templates and the resulting fluorescence was measured over time (Figure 6.4). Three consecutive batches were run, with the first run yielding the highest amount of protein and the following two batches yielding

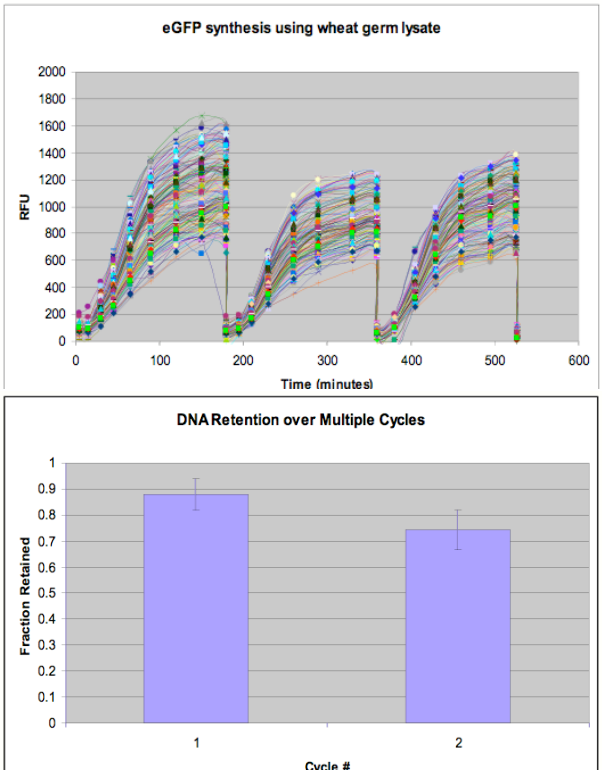



Figure 6.4: eGFP was synthesized in three consecutive batches from flow-deposited linear expression templates, and fluorescence intensity was observed over time on a microarray scanner. The remaining surface-bound linear template was quantified after each cycle and normalized to the starting amount.

comparable amounts. In all three cases, synthesis stopped after about 2 hours. To optimize the system one can optimize the run time of individual batches to under two hours, increasing cycle frequency and protein yield. The only limitation to the number of cycles that can be run depends on loss of template from the surface, either due to dissociation or nuclease activity. The rate of loss of functional template from the surface was determined by labeling the 3' end of the linear template with Cy3. In this experiment, after the first and second cycle roughly 90% and 75% of fluorescence remained. Interestingly, even though only 75% of the initial signal remained after the third cycle, protein synthesis seemed to be unaffected. It thus seems possible that the

signal loss may be due to non-specifically adsorbed molecules or bleaching, neither of which accurately reflects the number of functional linear expression templates present on the surface. As the protein synthesis yield seems unaffected after three cycles, it should be possible to run many more additional cycles with acceptable protein yield.

## 6.5 Continuous Synthesis

In the previous section, discontinuous synthesis has been shown to be easily realized, yielding good repetitive yields over several consecutive batches. Running continuous synthesis has not been put into praxis as of yet, but should be easily accomplished. The simplest instantiation of the method would involve continuously flowing ITT mix through a channel or a set of channels, with a reservoir at the outlets for collection of product. It might prove necessary to chill the ITT mix while it is being introduced onto the device in order to extend its lifetime. Likewise, setting up several ITT batches is possible, but also more labor intensive. The length of each ITT continuous batch run will depend not only on the retention of the expression template on the surface but also on the lifetime of the ITT mix itself. But, judging from previous experience, the overall lifetime should be long enough to result in increased synthesis efficiency and product yield.

# Chapter 7

# Detection of Molecular Interactions

## 7.1 Introduction

To understand molecular interactions one must not only be able to synthesize the required components, described in Chapter 6, but also be able to detect whether two molecules interact. Commonly used methods include: ELISA, yeast two hybrid, mass spectrometry, and surface plasmon resonance. The ELISA principle was adopted for use in microfluidics and two methods are described in Sections 7.2 and 7.3. An entirely novel method, based on the mechanically induced trapping of molecular interactions (MITOMI), is described in Section 7.4. MITOMI takes advantage of the ability to generate free-standing membranes in MSL devices (Section 2.3.4), which are used to mechanically trap molecular interactions taking place on a surface. This method of mechanically trapping interactions is widely applicable, due to its simplicity, and is, in fact, compatible with methods described in Sections 7.2 and 7.3. The advantages and disadvantages of each method are described in detail in their respective sections.

## 7.2 Antibody-Based Detection

Antibody-based detection is the most direct adoption of ELISA, but instead of using substrate turnover by an enzyme for signal generation, a more direct approach is taken by labeling the primary or secondary antibody with fluorophores. Generally two primary antibodies are required, one for immobilizing the bait to the surface, and the second for detecting the prey. Both antibodies recognize common antigens such as epitope tags (6xHis, S-tag, T7, Myc-tag, etc.) or proteins (GFP, GST, etc.), which can be engineered into the bait and prey molecules.

Antibodies vary widely in specificity and affinity for their respective antigens. Specificity to the antigen is generally not as important, as it is generally high for antibodies recognizing short epitope tags. Specificity is more important in the context of non-specific binding of the antibody to the surface. Even though seemingly identical, there is a difference between these two modes of non-specificity, namely whether the Fab region recognizes other epitopes non-specifically or whether any part of the antibody may bind non-specifically. An important metric of antibody performance is its specific affinity, which should be at least in the high pM regime with slow off-rates. A high affinity ensures that most of the synthesized antigen is surface localized and a slow off-rate ensures that the antigens stay there. Antibody clonality can also play a role, particularly in surface immobilization. A polyclonal antibody to GFP recognizes multiple epitopes on GFP and thus may pull the GFP-fusion protein down in various orientations, increasing the likelihood that a physiological interaction may take place without steric hindrance.

In most protein array methods, antibody selection becomes a limiting factor, especially when each antigen requires its own antibody pair. On-chip approaches require a maximum of a handful of matched antibodies, and then only if multiplexing is required. Generally only two antibodies are needed to perform interaction assays: one for surface immobilization and the second for detection of the bait protein. One very useful antibody, used in essentially all experiments, is a anti-penta-Histidine Antibody (Qiagen) labeled either with biotin for surface immobilization, or Cy5 for fluorescent detection. Other useful antibodies include anti-T7 as well as anti-GFP antibodies, also either labeled with biotin or fluorophores. A non-antibody option for pull-down is described in Section 7.3.

In summary, antibody-based methods are easily implemented and give good results in on-chip applications. The antibodies used here provide limited signal amplification on the order of five- to tenfold, due to the multiple labeling of each antibody with several fluorophores. More sensitive approaches include enzyme-based methods (see Section 7.3), where a signal may be amplified several thousandfold. But these latter methods also exhibit greater sensitivity to non-specific binding. Overall, antibody-based methods with direct detection coupled to MITOMI (Section 7.4) are sufficiently sensitive to detect most molecular interactions in high-throughput uses.

## 7.3 S-Tag Assay

A potentially useful method for both purification and detection of proteins is ribonuclease S, commercially available from Novagen. The ribonuclease S protein has been

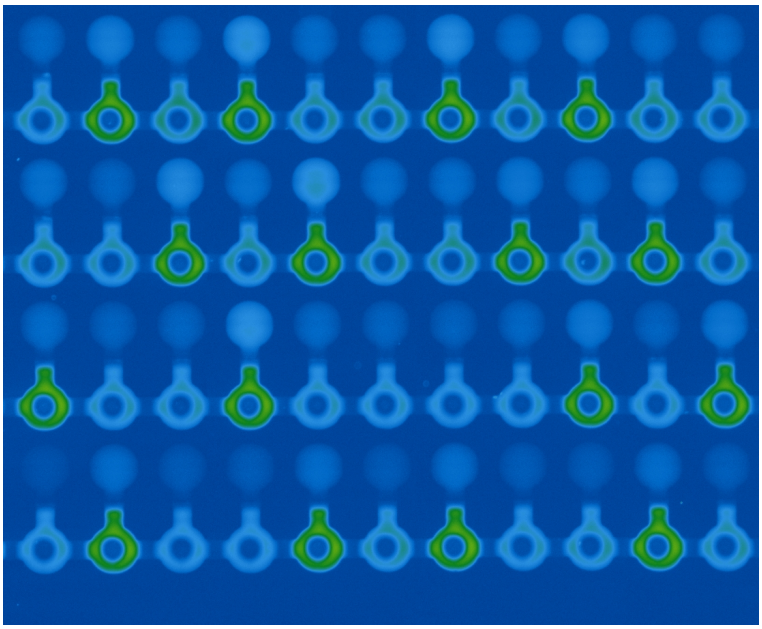


Figure 7.1: A fluorescent scan of an on-chip S-tag assay. Green intensities indicate high substrate hydrolysis and therefore presence of a S-tagged protein. A pattern of high and low intensities is observed correlating to the identity of expressed protein in each unit cell; proteins with and without s-tag tags were expressed and pulled down.

engineered to yield an inactive S-protein and a short S-Tag peptide of 15 amino acids in length. The S-protein can be used to pull-down and specifically bind to the S-tag epitope with an affinity on the order of $10^{-9}M$ [38], which is comparable to weak- to medium-affinity antibodies. Additionally, and more importantly, binding of the S-tag peptide to the S-tag protein reconstitutes its ribonuclease activity and thus can be used as a sensitive assay for protein interactions in lieu of standard antibod-based ELISA assays. The immediate advantage of the S-tag approach is the fact that the S-protein is inactive if not bound to the S-tag. In the case of ELISA approaches, the antibody-enzyme fusions are always functional and therefore very sensitive to non-specific binding. If the S-protein were to non-specifically bind somewhere it would remain inactive and thus won't contribute to background signal. Kelemen et al. [39]

developed a FRET-based assay for ribonucleases relying on the cleavage of a chimeric oligonucleotide dual labeled with a 5' fluorescein and a 3' rhodamine as quencher. The optimal sequence was a tetranucleotide of dArUdAdA, which exhibited the highest ratio of product to substrate and a high turnover rate. To adjust this substrate for use in high-throughput protein-protein interaction assays the FITC and TAMRA fluorophores were replaced with Cy5 and Iowa Black RQ$^{TM}$-SP, a black whole quencher, respectively. This moved the detection band of the substrate into the Cy5 region, which is more sensitive and has a higher signal to noise ratio on the ArrayWorxE than any blue-shifted band. To test whether the S-tag FRET assay could be used on-chip for detection of protein-protein interaction, a proof of principle experiment was set up to assay the interaction between yeast proteins fused with an N-terminal 6xHis tag and those with a C-terminal S-tag. The proteins were expressed on-chip from spotted linear expression templates coding for the yeast proteasomal subunits Rpt1-6. Once synthesized, the proteins bound to the surface localized antibodies via the 6xHis-tag interaction. At this point MITOMI was performed to trap the interactions in place, and, while the button was closed, the S-tag FRET assay mixture (S-protein + 5 $\mu$M substrate) was introduced into the device. The buttons were then opened, allowing the S-protein to bind to the C-terminal S-tag of the pulled down proteins, activating it. The active S-protein then hydrolyzed the substrates, giving rise to signal. The buttons were then closed once more to stop the reaction, and the resulting signal intensities were measured using the ArrayworxE scanner (Figure 7.1). The pattern of high (green) and low (blue) signals seen in Figure 7.1 arises from

positive and negative controls, the latter consisting of no-S-tag as well as no-His-tag versions of the proteins being expressed.

All available substrate in each unit cell was turned over by the S-protein (about 4000 functional S-proteins will turn over all available substrate in a single unit cell in 5 minutes). The results also indicated that the S-tag FRET assay is exceedingly sensitive as non-specific binding of proteins lacking a His tag also gave rise to signal. One partial solution to the non-specific binding problem was to use harsh wash steps with 6M Guanidine HCl and/or NaOH prior to the assay step. This treatment is possible because the button physically protects the detection area from these harsh chemicals. But even though it solves the problem of non-specific sticking of proteins surrounding the detection area, it cannot solve the problem of non-specific binding or trapping of molecules in the detection area. One possible solution is to make the S-tag FRET assay quantitative, which would be beneficial in any case. If the S-tag assay is quantitative, it should be possible to detect the absolute number of bound proteins in the detection area, which should vary as a function of affinity. Since non-specific binding generally occurs in the low-affinity regime, these signals could be removed from true positives by a simple signal intensity cutoff, selecting only for the higher signals. Likewise, one could include a brief buffer wash prior to analysis, which would ensure that only interactions with larger off-rates are retained. Unfortunately any of these methods has the intrinsic problem that they would loose potentially interesting low-affinity interactions. Other possible modifications to optimize the assay lie in adjusting the substrate concentration as well as the substrate sequence

itself; a substrate of the sequence dArGdGdA was tested and showed considerably slower hydrolysis rates.

In summary, the S-tag assay is a viable and easily modified method for the ultra-high sensitivity detection of molecular interactions. The modularity of the tags and the substrate allows easy modification of parameters, such as the optimal detection band, the affinities of individual components, and the substrate turnover rate. Therefore, in the case where direct detection with a fluorescently labeled antibody is not sufficient, the S-tag approach is a good follow-up candidate.

## 7.4 Mechanically Induced Trapping of Molecular Interactions

Protein array platforms have two major shortcomings preventing them from being broadly applied in Systems Biology. The first problem with protein arrays is the fact that each individual protein spot has to be generated using bench-top expression and purification approaches. Recently Ramachandran et al. [30] reported a method to

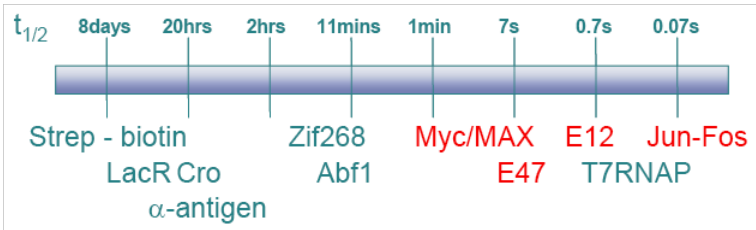



Figure 7.2: Plotted are off-rates for representative biological interactions spanning the entire spectrum from very stable interactions such as streptavidin-biotin to extremely transient ones such as Jun-Fos.

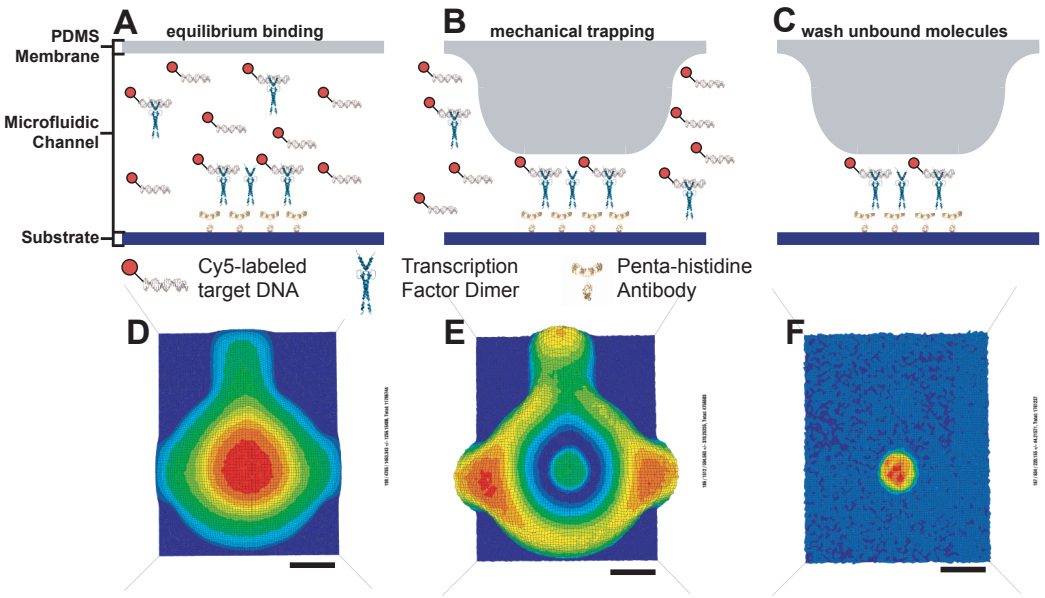



Figure 7.3: (A–C) schematic of the process of MITOMI. The gray structure at the top of each panel represents the deflectable button membrane that may be brought into contact with the glass surface (blue). (A) $His_5$ tagged TFs are localized to the surface and TF-DNA binding is in equilibrium. (B) The button membrane is brought into contact with the surface, expelling any solution phase molecules, while trapping surface-bound material. (C) Unbound material not physically protected is washed away and the remaining molecules are quantified. (D–F) Fluorescent intensity maps of target DNA concentration from an actual device. Panels D–F correspond to the top down perspective of Panels A–C.

rapidly synthesize protein arrays *in situ*. Improved methods based on similar ideas are described in Chapters 4 and 6.

The second problem intrinsic to protein arrays is the issue of transience of molecular interactions. For DNA array platforms this does not present a problem since once DNA templates anneal to the array probes these interactions are extremely stable. Protein arrays, on the other hand, have to detect interactions amongst a wide variety of proteins, whose affinities and kinetics vary over orders of magnitude (see Figure 7.2). Stable interactions are much more readily detected than transient ones,

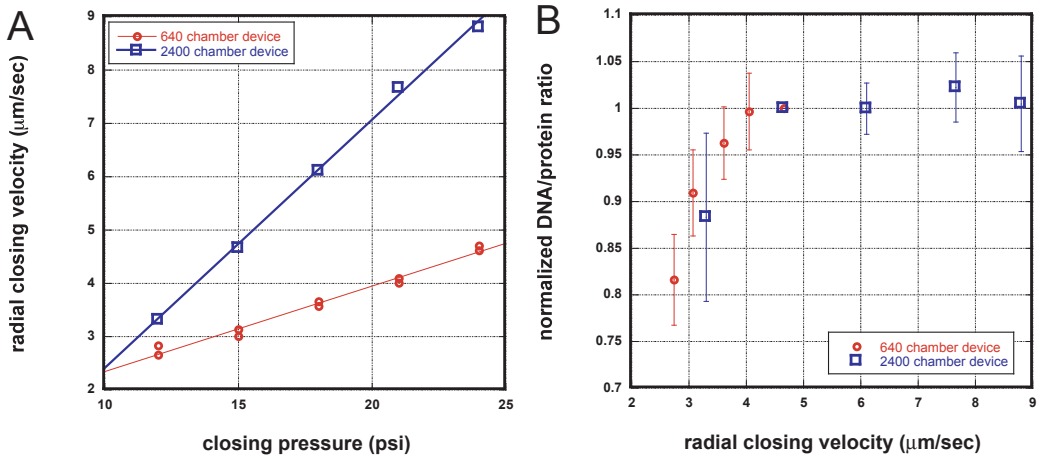



Figure 7.4: Effect of button closing velocity on trapping efficiency. (A) Radial closing velocities were measured at various pressures on two different devices, with slightly different architectures. (B) The effect of the above closing velocities on trapping efficiencies. Below $4\mu m/sec$ trapping is strongly dependent on button closing velocity. This dependence disappears above $4\mu m/sec$ where the response plateaus off.

as transient interactions will rapidly dissociate during wash steps and thus cause considerable loss of signal before an array can be analyzed.

To overcome the problem of transience, a method based on the mechanically induced trapping of molecular interactions (MITOMI) was developed. MITOMI physically traps molecules between two surfaces, essentially freezing surface interactions in place and preventing bound material from diffusing out of the detection area (Figure 7.3).

When trapping the interacting molecules, MITOMI preserves the equilibrium distribution of the molecules. This was tested by determining that the radial closing velocity of the free-standing membrane has no effect on trapping efficiency at velocities above 4 $\mu$m/sec (Figure 7.4). At slower velocities surface-bound molecules are able to dissociate and diffuse out of the detection area as indicated by a lower ratio

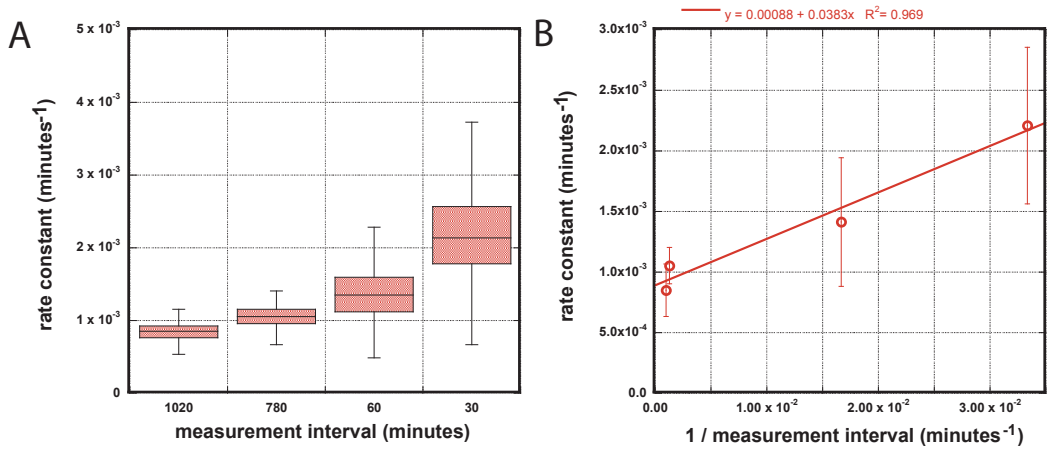



Figure 7.5: Measurements illustrating the effectiveness of the mechanical trapping of molecules. (A) Rate constants determined at four measurement intervals. (B) The same rate constant averages as in (A) plotted as a function of inverse measurement interval. The dependence on the measurement interval indicates that bleaching contributes to the measured rate constants. The intercept represents the actual mass loss rate (without the contribution of bleaching) with a value of 0.0009 minutes$^{-1}$.

of bound material. To further assure that the measured affinities, reflect actual equilibrium affinities previously reported affinities determined by EMSA were positively compared to values determined by MITOMI ([7]).

A second measure of the effectiveness of MITOMI is how well it traps molecules in place. To measure the effective rate loss, a molecular interaction consisting of a transcription factor bound to labeled DNA was trapped with MITOMI and the loss of bound DNA measured over time (Figure 7.5). Signal loss was dominated not by actual mass loss of DNA, but by loss of fluorescence due to bleaching, as indicated by the apparent dependence of the measured rate constant on the measurement interval. The rate constant solely due to mass loss could be determined by plotting the observed rate constants as a function of the inverse of the measurement interval and extrapolating to 0, since $1/\infty = 0$. The rate constant for mass loss measured in this system was

determined to be $9 * 10^{-4}$ minutes or $1.5 * 10^{-5}$ seconds, which translates into a half-life of 13 hours and is 4–5 orders of magnitude larger than the observed half-life without trapping (refer to Figure 7.2). MITOMI is thus extremely effective in trapping surface-localized material, giving ample time to image the device without losing signal in the interim.

Not only does MITOMI preserve the equilibrium concentrations and effectively trap interactions, it is also compatible with a wide variety of systems and detection methods. Literally any molecular interaction can be tested as long as one of the components can be surface localized. Possible applications for MITOMI include detecting interactions between: protein-protein, protein–DNA/RNA, protein–small molecule, DNA/RNA–small molecule, and so on. Trapping itself is also not limited to the free-standing membranes described above, but may be achieved with any two surfaces exhibiting similar properties as the PDMS-glass interface used here. Finally, one specific example of how MITOMI can drastically enhance sensitivity of existing detection methods lies in combining MITOMI with the S-tag-based enzymatic assay. Here MITOMI can be used to initially trap the prey molecules in place followed by a buffer exchange, introducing the S-protein as well as the substrate without loss of bound material. The enzyme reaction may then be started uniformly across the device by opening the button, allowing S-protein to bind S-tags present on the prey molecules. At this stage, prey will dissociate from the detection area, but remain localized to the unit cell, therefore causing no net decrease on signal. This sequence of steps ensures that the highest possible signal is obtained in each unit cell.

Another intriguing possibility lies in determining the kinetic parameters of binary interactions using MITOMI. Determining off-rates is particularly straightforward. Standard valves can be actuated at frequencies of up to 100 Hz [6]. Therefore one actuation cycle is on the order of 10 ms. Interactions can therefore be trapped in place and allowed to dissociate with 10 ms time resolution by rapidly opening the trapping membrane. The half-life for an interaction with an off-rate of 1 $sec^{-1}$ is 700 ms. Therefore a 100 Hz actuation frequency of the membrane will collect 70 samples during a half-life, and even at a 10 Hz actuation frequency 7 samples can be collected. The actuation frequency is thus sufficient to sample even extremely fast dissociations. What is more important is the fact that MITOMI allows the sampling of a large number of off-rates at high time resolution in parallel, as all unit cells are sampled in parallel at the frequency determined by the actuation rate. The on-rate can be deduced from knowing the equilibrium dissociation constant as well as the off-rate of the system, but it could also be directly measured. Measuring the on-rate relies on the same mechanism as measuring the off-rate. Instead of allowing the bound molecules to disassociate, they are allowed to associate. What complicates the on-rate measurement is the fact that the on-rate is concentration dependent. The concentration of the prey molecules must therefore be known beforehand or determined *in situ*.

MITOMI presents a highly integrated approach for quantitatively characterizing not only the equilibrium constants of binary interactions, but also their kinetic parameters. As the topologies of biological networks are becoming well characterized

the underlying parameters governing each interaction, and thus the entire network, have been barely touched, and therefore have been described in binary terms. It is becoming more and more apparent that these parameters need to be measured on proteomic scales in order to understand biological networks in greater depth. Parameters have been lacking thus far due to the considerable technical difficulties in obtaining these measurements. MITOMI provides one generically applicable answer to this problem and should prove useful in a wide variety of applications attempting to discern the properties of molecular interactions.

# Chapter 8

# Helix-Loop-Helix Transcription Factors

## 8.1 Introduction

Helix-loop-helix transcription factors form a structural family of important transcriptional regulators mainly found in eukaryotes. A current InterPro search for the helix-loop-helix domain returned 206 hits in human, 196 in mouse, 102 in the fruit fly, and 12 in yeast, indicating its broad conservation across species. HLH transcription factors are known to play important roles in differentiation, lineage commitment, and sex determination. In yeast HLH transcription factors have been implicated in phosphate regulation and phospholipid biosynthesis [40, 41].

The first HLH transcription factors identified were the E2A gene products E12 and E47 [42]. These proteins were shown to bind to a DNA sequence called $\kappa$E2, a canonical hexamer of CACGTG generically called E-box, found in the immunoglobulin kappa chain enhancer. Other known E-box sites are located in the IgH gene enhancer: $\mu$E1, $\mu$E2, $\mu$E3, $\mu$E4, and in the Ig kappa enhancer: $\kappa$E1, $\kappa$E2, $\kappa$E3. Both E12 and E47 were predicted to contain two amphipathic alpha helices required for

dimer-formation and DNA binding. These helical regions aligned well with Daughterless, MyoD, and Myc resulting in identity or near identity of the hydrophobic residues of the helices, supposedly of importance for dimerization.

Shortly after the discovery of this new class of transcription factors, other members such as c-Myc were interrogated for DNA binding specificities. Blackwell et al. used a newly developed technique: selected and amplified binding-sequence (SAAB) imprinting [43] (SELEX is essentially the same as SAAB) to determine the sequence specificity of C-Myc. In SAAB imprinting a suspected DNA binding protein is subjected to a library of sequences. Any bound material is subjected to rounds of PCR amplification and further selection. The final material is then sequenced to establish the identity of the strongest binders. Using SAAB imprinting, it was determined that MyoD and the E2A products bind a common consensus sequence, CANNTG, but show little specificity for the two central and flanking bases. Similar results were obtained for Myc and MAX [44, 45], which were shown to form functional homo- and heterodimers both *in vitro* and *in vivo* [46, 47]. MAX was previously found to be a binding partner of c-Myc using a comprehensive cDNA screen by Blackwood and Eisenman [48]. Halazonetis et al. [49] screened a small library of E-box sequences all containing the same E-box sequence flanked by various trinucleotides. The results showed that *in vitro* translated c-Myc and TFEB do indeed discriminate between sequences containing various flanking bases.

It was also established that phosphorylation of Max affects its dimer formation and DNA binding preference [44, 50]. Both isoform A and B (p21/22 respectively)

were shown to be specifically phosphorylated by casein kinase II (CKII) [47, 50]. Phosphorylation of the amino terminus of Max causes a reduction in DNA binding affinity if bound as a homodimer, but had no effect on Max/Myc heterodimer affinity to DNA. Interestingly, incubation with rabbit reticulocyte lysate causes phosphorylation of Max [44]; many early experiments were performed on *in vitro* transcribed protein from rabbit reticulocyte [48], and thus were using phosphorylated versions of Max and potentially Myc, which also contains CKII sites. A difference between the two isoforms of Max was also detected; isoform A is more susceptible to affinity changes caused by phosphorylation as compared to its shorter isoform B, which can be explained by an additional phosphorylation site near the basic region present in isoform A but not B. In yeast, phosphorylation also plays an important role in the regulation of Pho4p. Pho4p is phosphorylated by a cyclin-CDK complex, Pho80-Pho85, at four distinct sites. Differential phosphorylation of Pho4p affects its cellular localization and in turn causes a change in gene expression patterns [51, 52]. In the hyper-phosphorylated state, Pho4p is shuttled out of the nucleus and into the cytoplasm by Msn5p [53] causing the de-activation of PHO5. Turning PHO5 off causes inorganic phosphate concentrations to fall and Pho4 is de-phosphorylated, transported back into the nucleus by Pse1p, and once again capable of turning on PHO5. The underlying code specifying nuclear import and export is governed by six serine-proline dipeptides SP-1-6, with SP2 and 3 being necessary for export and SP4 for import out and into the nucleus [54], respectively. Phosphorylation of HLH transcription factors thus seems to be an intimate part of their regulation *in vivo*.

Helix-loop-helix proteins may be classified into various groups or families of proteins [55, 56] (Table 8.1). Murre et al. [56] established 6 classes of HLH transcription factors, whereas Atchley et al. [55] arranged the factors into 5 groups according to E-box specificity, with each group containing several protein families. Murre's grouping into classes puts the E proteins, E12, E47 into Class I. Class I proteins may form homo- as well as heterodimers and are expressed over a wide variety of tissues [56]. Class II includes MyoD and myogenin. Members of this class may only form homodimers, mainly with members of Class I, and are expressed in tissue-restricted patterns. Classes III and IV bHLH transcription factors contain a leucine zipper domain and may dimerize with one another or with themselves. Class V proteins are negative regulators of Class I and Class II transcription factors, in that they may form heterodimers with members of those classes, but since Class V proteins lack a basic region, the resulting heterodimer has no DNA affinity. Finally, Class VI and Class VII are characterized by containing a proline in their basic region and a bHLH-PAS domain, respectively [56].

Extensive structural information is available on HLH dimerization and DNA binding interfaces obtained from crystal structures of MAX [57], E47 [58], Pho4p [59], Myc-Max and Mad-Max heterodimers [60], and NMR solution structures of MAX homodimers [61]. The crystal structures affirmed the proposed helical structure as well as the importance of hydrophobic residues for dimer formation. A dimer consists of a four-helix bundle with a basic region bundle making contact with the E-box sequence. Each monomer contains the following regions starting at the amino termi-

Table 8.1: Helix-Loop-Helix transcription factors

| Protein Families | Included Proteins | Groupings | | | Function |
|---|---|---|---|---|---|
| | | E-box | Murre et al.[56] | LZ | |
| AC-S | ac, sc, ase, l'sc, mash, ash | A | II | | Neurogenesis, determination of neuronal precursors |
| dHAND | dhand, ehand, hxt, hed | A | | | Rardiac morphogenesis, trophoblast cell development |
| E12/Da | e12, e47, itf, pan, G12, me2, da | A | I | | Neurogenesis, sex determination, regulation of myogenesis |
| MyoD | myod1, myogenin, myf5, myf6 | A | II | | Myogenesis |
| TWIST | twist, ec2, paraxis, scleraxis, dermo | A | | | Specification of mesoderm lineages |
| CBF | cbf-1 | B | | | Rentromeric binding and chromosomal segregation |
| HAIRY | hlhm, hairy, hes, deadpan, e(spl) | B | VI | | Neurogenesis, segmentation |
| MAD | mad, mxi1 | B | IV | Yes | Regulation of cell differentiation |
| MYC | c-myc, n-myc, l-myc, max | B | III | Yes | Cell proliferation, differentiation; oncogenesis |
| PHO4 | pho4, nuc1 | B | | | Phosphate regulation in yeast |
| TFE | tfe3, tfeb, mi | B | III | Yes | Transcription activation in immunoglobulin heavy chain enhancer |
| ID | id, heira, emc, hlh462 | D | V | | Negative inhibition of DNA binding; myogenesis, neurogenesis |

*Table adopted from [55].*

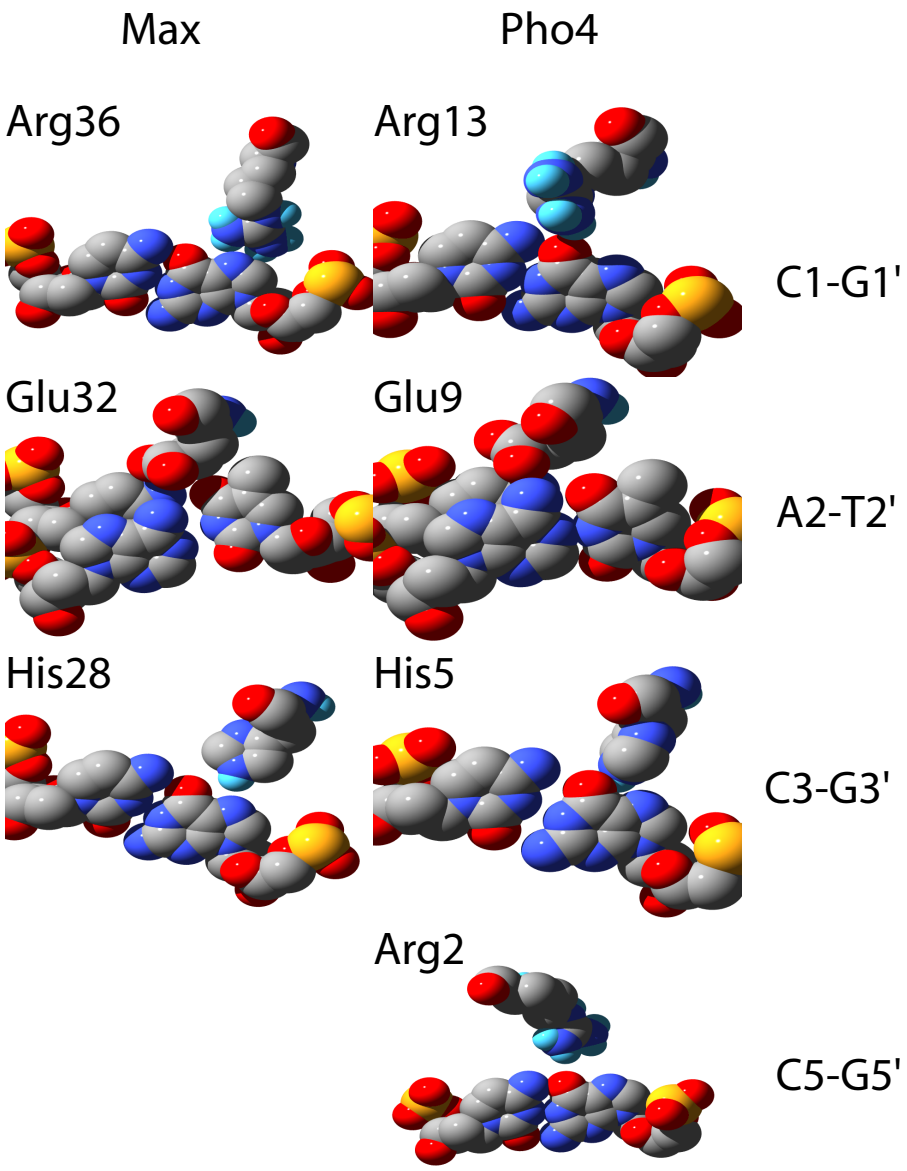



Figure 8.1: Sequence-specific DNA protein contacts of Max and Pho4

nus: basic region-helix1-loop-helix2. Max additionally extends the second helix into a leucine zipper, which drives dimer formation by van der Waals interactions of hydrophobic residues. These residues are mainly leucines and isoleucines, are conserved across the HLH family, and can be found in both helical regions of the HLH domain. Differentiation of various binding pairs, be they homo- or heterodimer, seems to take place in the second helical regions and zippers, if present. In MAX homodimers the residues Asn92 and Gln91 form a stable tetrad, causing a bulge of the zippers and non-optimal packing. This conflict is resolved in MAX/Myc and MAX/MAD heterodimers [60]. E-box base recognition is accomplished by a conserved glutamic acid residue, an arginine residue, and a histidine residue (Figure 8.1). Glu makes contact with $C_3$ and $A_2$, Arg contacts $G_{1'}$, and His makes a contact with $G_{3'}$. An additional contact with $G_{4'}$ or $G_{5'}$ is only seen in PHO4 [59]. Bases are named according to the scheme: $C_3A_2C_1{\bullet}G_{1'}T_{2'}G_{3'}$, where $\bullet$ denotes the dyad symmetry axis (Note that $N_3N_2N_1{\bullet}N_{1'}N_{2'}N_{3'} = N_{-3}N_{-2}N_{-1}N_1N_2N_3$, a notation scheme used below). These contacts are universally seen in structures of MAX [57] as well as Pho4p [59], and, because of their conservation, are likely present in a large number of HLH transcription factors. Additional contacts with DNA bases are seen in the Pho4p structure, and an extensive network of contacts with the phosphate backbone is present in both structures.

More recently, members of the HLH family of transcription factors have been mapped to potential genomic binding sites. Orian et al. mapped the *in vivo* binding sites of the Myc/MAX/MAD network using a DAM methylation approach in

*Drosophila* [62]. Cawley et al. mapped the binding sites for c-Myc and two other non HLH transcription factors, p53 and Sp1, using chromatin immunoprecipitation (ChIP) along chromosomes 21 and 22 [63]. The results are surprising in that c-Myc seems to have up to 25,000 binding sites with roughly 18% of those lying in a defined 5'Exon, which is a 3.7-fold enrichment of what would be expected at random [63]. Another 24% of sites fall into known CpG islands, but over half of the binding sites are located immediately within 3' of known genes and strongly correlate with non-coding RNA [63]. Because c-Myc was chosen as the target for immunoprecipitation, the large number of sequences bound most likely stem from a combination of many possible dimerization states of c-Myc with partners such as MAX, etc., making it hard to extract exact binding characteristics of c-Myc alone. Similar results have been obtained by Orian et al., who tested binding of the MAX network including c-Myc and Mad/Mnt. Here binding sites were determined by DAM methylation followed by hybridization of methylated sites to a *Drosophila* cDNA array. Because a cDNA array was used, only sites binding in coding regions could be determined, as opposed to the comprehensive screen performed by Cawley et al. [63]. Again it was found that the tested HLH transcription factors interacted with about 15% of transcribed regions of the genome. Furthermore it was found that modulating the relative expression level of MAX changes the binding pattern for c-Myc, suggesting that gene control is dependent on relative abundances of individual members of the network [62]. Both approaches yield low-resolution information on potential binding sites, but exact binding motifs are hard to extract. Orian et al. used the REDUCE

algorithm to extract binding sites, which did indeed return the E-box motif for c-Myc and Mnt, but not for MAX. Additional motifs found included CG-repeats, TATC-GATA and GGTCACACT. Whether these results stem from a limit in resolution or from potential tertiary interactions with other factors is unknown.

## 8.2 Energetics of DNA Recognition

DNA recognition by proteins, particularly transcription factors, lies at the heart of gene regulation. Intricate topologies of transcription factor–DNA interactions, or gene regulatory networks, arise from genes having multiple transcription factors as their cis-regulatory input. Likewise one transcription factor can regulate a large number of target genes. Deciphering these networks presents a giant leap towards understanding cellular function; as literally every single cellular pathway converges onto the regulation of genes, and therefore transcriptional regulatory networks.

In order to discern their target genes, transcription factors must physically read out the genetic code base by base. Every DNA binding protein has an amino acid motif that specifically binds or recognizes certain sequences but not others. The major transcription factor families more or less use the same mechanism to achieve this readout (Figure 8.2); the use of an alpha-helix, which fits into the major groove of DNA, is the primary structural element for reading out DNA sequences. The alpha helical pitch also allows every $4^{th}$ amino acid side chain to be positioned in proximity to the exposed nucleotide bases, while other side chains make secondary, non-specific contacts with the phosphate backbone, increasing the overall stability of the complex.

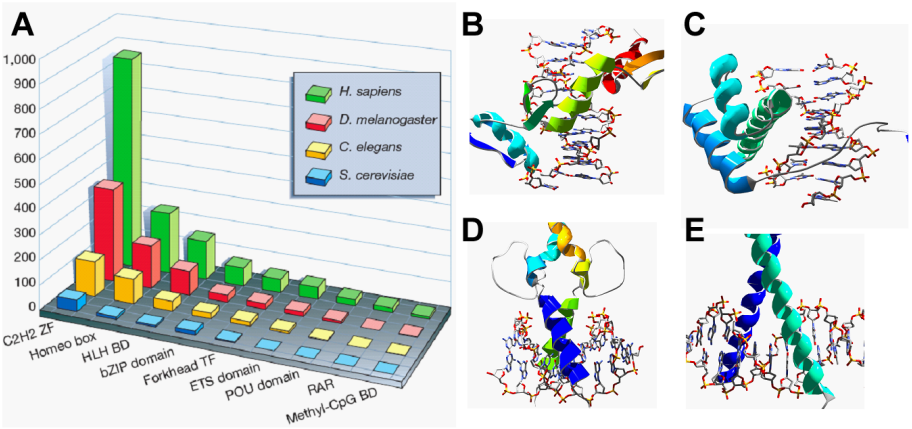



Figure 8.2: Panel A shows the number of transcription factors of each major family found in a genomic screen of 4 eukaryotic genomes. Panel B–D are representative structures of the 4 major transcription factor families. Shown are a Zinc Finger, homeobox, bHLH, and bZIp in B–D, respectively.

But it is the base-specific contacts that impart a transcription factor with its sequence specificity.

Understanding how transcription factors function is of primary importance in understanding transcriptional regulatory networks. A multitude of approaches have been developed over the years based on physical *in vitro* and *in vivo* measurements, as well as bioinformatic approaches. One of the earliest measurements on transcription factor–DNA interaction, and an approach still widely used today, is the electrophoretic mobility shift assay (EMSA). Here a transcription factor is mixed with a specific target sequence. The mixture is then electrophoresed to separate bound complexes from unbound components and the relative concentration of each is measured. A second method is termed SELEX (for systematic evolution of ligands by exponential enrichment). SELEX, as the name implies, is based on screening a large pool of sequences by passing them over a column preloaded with the transcription factor of

interest. The immobilized transcription factor binds its target sequences, which are consequently eluted from the column and amplified. This process is repeated until the highest-affinity sequences are recovered. SELEX is primarily used for discovering a transcription factor's consensus sequence, but otherwise lacks quantitation due to the large number of selection loops required and the exponential amplification of the selected molecules between trials. Another and more recent method relies on dsDNA microarrays, to which a labeled transcription factor is allowed to bind. This method is known as protein-binding micro-arrays (PBMs). These micro-arrays may contain tens of thousands to millions of features, and thus can cover a very large sequence space. After a series of wash steps, the micro-array is imaged and the bound transcription factor quantified. This method is extraordinarily useful for determining consensus binding motifs. One of its major pitfalls is the need for wash steps, causing considerable loss of signal due to the high dissociation rate of transcription factors. It is nonetheless a powerful and semi-quantitative approach for understanding the physical basis of transcription factor DNA recognition. An *in vivo* approach to discovering what genomic regions are bound by transcription factors is ChIP-chip. ChIP-chip is a combination of chromatin immunoprecipitation followed by a DNA micro-array for identification of the pulled-down DNA sequences. Specifically, cells are permealized and all contents, including transcription factors bound to genomic DNA, are covalently cross-linked. Then genomic DNA is either physically sheared or enzymatically digested. The short DNA segments that have transcription factor bound to them are immunoprecipitated using a transcription-factor-specific antibody. After several

clean-up steps, bound DNA is released, amplified, labeled, and eventually detected on a DNA micro-array. ChIP-chip is an extremely powerful approach for determining the *in vivo* function of transcription factors. Several large-scale datasets have been collected, including a comprehensive dataset of all yeast transcription factors and high-resolution mapping of several eukaryotic transcription factors. ChIP-chip, principally a powerful approach, falls short of the promised goals by returning rather noisy datasets, making data analysis difficult at best and generally requiring the recursion to bioinformatic methods for motif discovery. Furthermore the resolution for localizing the actual sites a transcription factor is bound to is disappointingly low, especially when one takes into account that transcription factors only bind short segments of DNA spanning at most tens of bases.

In a similar spirit to the enzyme hunters of the earlier years of molecular biology such as Arthur Kornberg, it should prove extremely useful to completely characterize transcription factor function *in vitro*. But unlike enzymes, which bind to generally only a single substrate and hydrolyze it, characterization of a transcription factor involves understanding its binding affinity to all possible target DNA sequences. Thus, instead of having to characterize a single interaction, several hundreds to millions of possible targets need to be measured. In the case of bHLH transcription factors this sequence space is rather small. The homodimer symmetry essentially reduces the length of the recognized motif by half, such that only the CAC portion of the E-box motif needs to be understood, instead of the entire CACGTG or 6mer space. The ability to reduce the search space considerably—since the number of sequences grows

exponentially with the motif length—made bHLH transcription factors an appealing proof of principle system.

Finally, each individual sequence needs to be studied in sufficient detail to provide a binding affinity. This generally entails generating saturation binding isotherms by varying either the transcription factor or DNA concentration. Varying the DNA concentration is generally the easier of the two and thus the preferred method. In certain circumstances if PBMs are used it does become necessary to vary the protein concentration instead.

### 8.2.1 E-box Libraries

As mentioned in the previous chapter, rather large libraries of dsDNA sequences are required to comprehensively study transcription factor function. These target sequences also need to be stoichiometrically labeled with a fluorescent dye so that concentrations may be determined. Synthetic DNA is readily available from commercial sources in the form of ssDNA oligomers reaching lengths of up to 150 bases. It is financially not viable to order both the Watson and Crick strand of each sequence to be tested and have one of the two strands labeled with a dye. Instead a generic primer, labeled with Cy5, was used to extend a ssDNA target library of primers by isothermal PCR using Klenow 3'-5' exo$^-$ as the polymerase. This reaction achieves both the synthesis of the complementary strand, as well as stoichiometric incorporation of a fluorescent label. Rather high concentrations of dsDNA can be achieved, which is necessary to reach the high mM concentration ranges required for gener-

| | 5' | | | | E-Box | | | | | | | | | 3' | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Name | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Members |
| NNCAC | T | T | N | N | C | A | C | G | T | G | T | T | T | T | 16 |
| NNNCAC | T | N | N | N | C | A | C | G | T | G | T | T | T | T | 64 |
| NNNNGTG | T | T | G | N | N | N | N | G | T | G | G | G | T | G | 256 |
| CACNNN | T | T | G | T | C | A | C | N | N | N | A | C | T | T | 64 |
| GTGNNN | T | T | T | T | C | A | C | G | T | G | N | N | N | T | 64 |
| NNCACGTGXX | T | T | N | N | C | A | C | G | T | G | N | N | T | G | 240 |
| Matrix Library | T | T | G | G | N | N | N | N | N | N | G | G | T | G | 64 |
| NNNx2 | T | T | G | G | N | N | N | X | X | X | G | G | T | G | 64 |
| NNNx2v2 | T | T | T | T | N | N | N | X | X | X | T | T | T | T | 64 |
| NNNx2v3 | C | C | C | C | N | N | N | X | X | X | C | C | C | C | 64 |
| CANXTG | T | T | A | A | C | A | N | X | T | G | G | T | T | G | 16 |
| SSS CAN | C | C | C | C | C | A | N | X | T | G | C | C | C | C | 4 |
| SSS CNC | C | C | C | C | C | N | C | G | X | G | C | C | C | C | 4 |
| SSS NAC | C | C | C | C | N | A | C | G | T | X | C | C | C | C | 4 |
| SSS NCAC | C | C | C | N | C | A | C | G | T | G | X | C | C | C | 4 |
| SSS NCAC | C | C | C | N | C | A | C | G | T | G | X | G | C | C | 4 |
| SSS NCCAC | C | C | N | C | C | A | C | G | T | G | G | X | C | C | 4 |
| | | | | | | | | | | | | | Sum | | 1000 |

Table 8.2: Table of target DNA libraries used for determining bHLH binding energy landscapes. N indicates any nucleotide and X stands for the complementary base in symmetric sequences.

ating complete binding isotherms. No further purification steps were necessary and the PCR reactions could be spotted directly after adding a 1% solution of BSA in $dH_2O$ to prevent covalent attachment of the synthesized DNA to the epoxy substrate. Necessary dilutions were carried out bench-top in 384 well plates prior to spotting.

Figure 8.2 shows the libraries generated for characterizing the bHLH transcription factor family. All libraries are centered around the E-box consensus motif. The first set of libraries, such as NNNNGTG, simply consist of all linear combinations of the 4 bases indicated by N, where N is any base. Other libraries such as NNNx2v1 consist of a symmetric 3mer sequence space NNN, with the symmetric bases indicated by XXX. A few specific members of this library are CATAGT and GCATGC for example. Simpler versions of the above library are the single-base symmetric substitution libraries, such as SSS NAC. The resulting binding energy landscapes that were obtained with these libraries are discussed in the following sections. Libraries can and

should be adjusted to the particular transcription factor or transcription factor family under study. Other libraries that were generated thus far include a set of libraries covering the Gli motif, as well as one CREB library covering the CREB consensus, both discussed in Section 8.6.

### 8.2.2 Transcription Factor Binding Energy Landscapes

Binding energy landscapes were determined for MAX isoform A and isoform B, both human bHLH transcription factors, as well as Pho4p and Cbf1p from yeast. All four of these transcription factors were tested against a comprehensive set of target sequences, obtaining precise binding affinities for each. $K_d s$ are obtained from the measured binding isotherms (Figure 8.3) by fitting a function of $Y = \frac{Bmax[X]}{K_d + [X]}$, where Bmax is the maximal binding and [X] is the concentration of free DNA required for half-maximal binding. For obtaining precise affinities it is necessary to have at least one sequence that reaches saturation, for which Bmax can be determined. The rest of the data can then be globally fit using that parameter. As is obvious from Figure 8.3, most sequences that fall in the low-affinity regime have a concentration-dependent response in the linear regime.

Several datasets were cross-compared to learn the systems' measurement error (Figure 8.4). In Panel A, $\Delta\Delta$G values for the NNNNGTG library of the two MAX isoforms were compared to one another, including N- and C-terminally tagged versions. The isoforms themselves are identical except for a short 9-amino-acid insertion in MAX isoform A just N-terminally of the basic region. Thus, sequence specificity

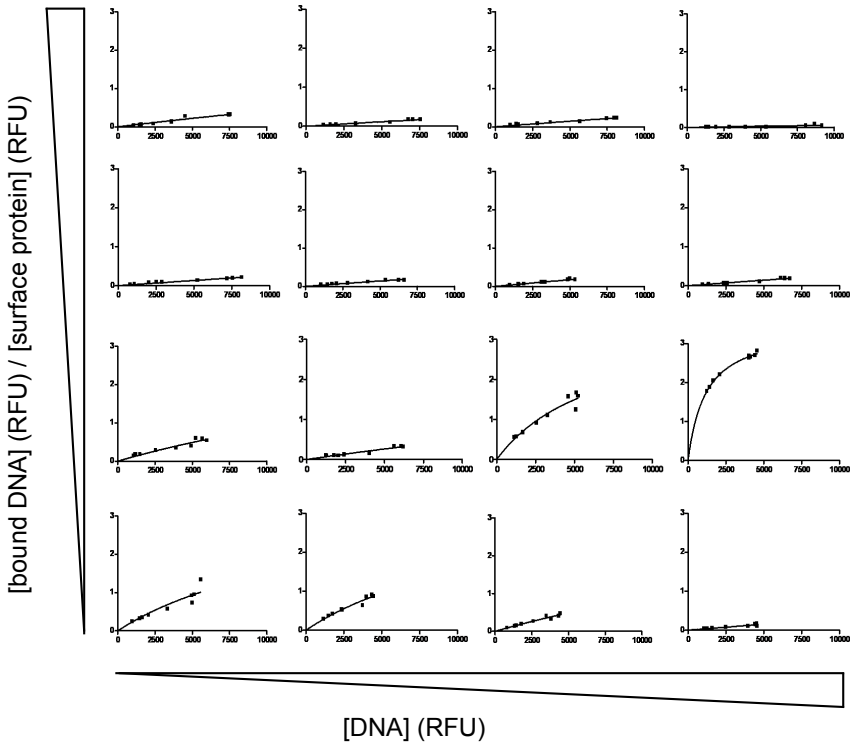



Figure 8.3: Representative binding isotherms from MAX isoform A. Shown are the first 16 sequences of a NNNGTG library. DNA concentrations are plotted on the x-axis with the corresponding response shown on the y-axis.

is expected to remain unchanged, and in fact MAX iso A and iso B compare well to one another. A comparison of the same isoforms across the two possible epitope locations showed a similar error (Figure 8.4, Panel B). Interestingly, the C-terminally tagged versions show enhanced binding affinities over all sequences, exhibited by a slope of 1.33 and 1.25 for MAX iso A and iso B, respectively. Finally, comparing measurements of Pho4p and Cbf1p for a 3-mer CACNNN library again shows a similar experimental repeatability. To estimate a global measurement error all NNNNGTG datasets of MAX iso A, iso B, Pho4p, and Cbf1p were compared using N- and C-terminally tagged variants. The resulting data is shown in Panel D of Figure 8.4. Here the $K_d$ values are plotted, rather their respective $\Delta\Delta$G values. The error was

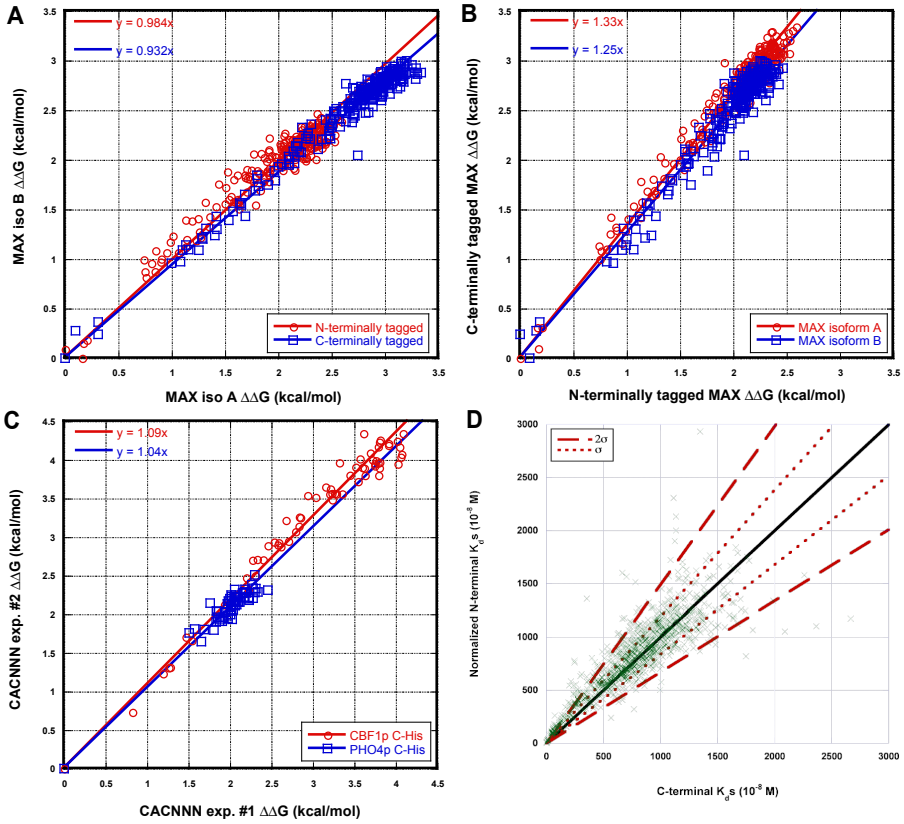



Figure 8.4: Dataset comparisons to determine measurement reproducibility and error. Panel A compares datasets across the two MAX isoforms. Panel B shows the same data as in Panel A compared across epitope tag location. Panel C shows a direct comparison of equivalent datasets for Pho4p and Cbf1p taken on different days. Panel D is a global error estimate comparing $K_d$ values for all NNNNCAC datasets for MAX iso A, iso B, Pho4p, and Cbf1p. Dotted and dashed lines show the 19% and 49% at one and two $\sigma$, respectively.

determined to be 19% at $1\sigma$ and 49% for $2\sigma$, respectively. This percent error transforms to a constant error of 0.17 kcal/mol and 0.40 kcal/mol when applied to $\Delta\Delta$G values.

Figure 8.5 shows the binding energy landscapes for MAX iso A, iso B, Pho4p, and Cbf1p covering 256 sequences of the 5' end of the E-box motif or NNNNGTG. It is immediately obvious that NCAC is the preferred sequence for all 4 transcription factors. It can also be seen that there is diversity in base preference for the first

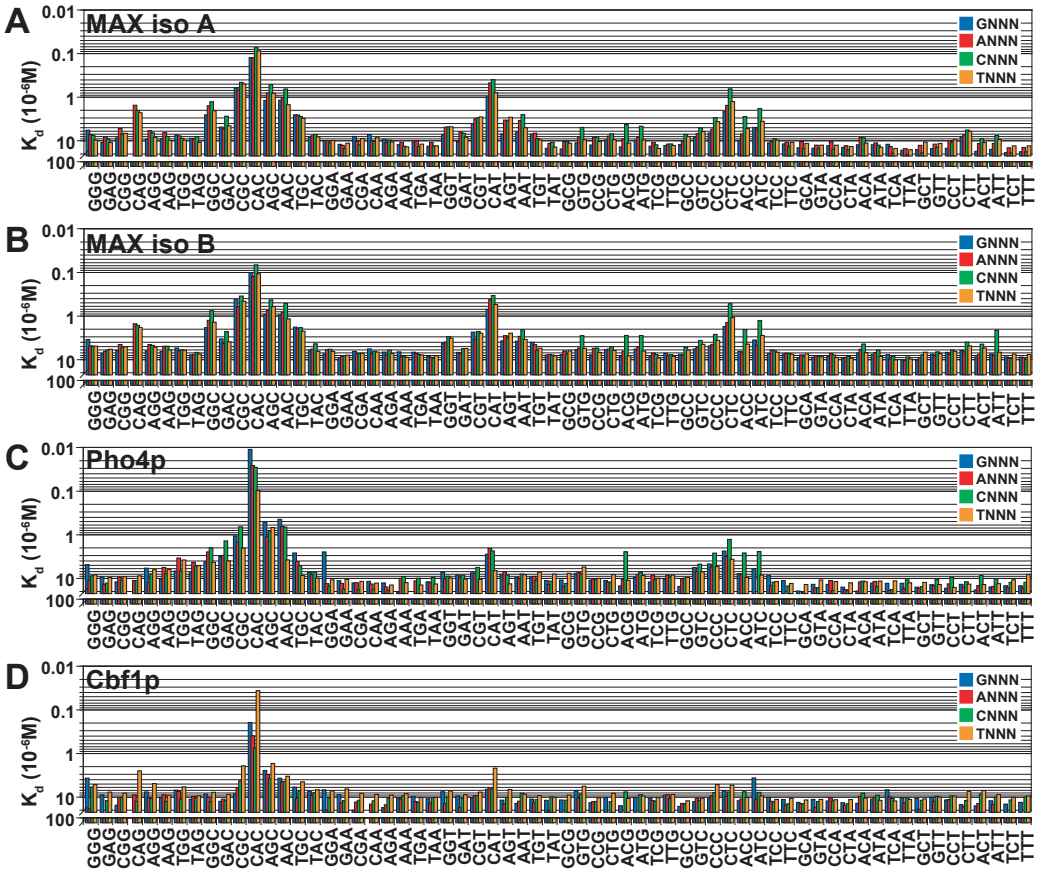



Figure 8.5: Binding energy landscapes for transcription factors MAX iso A, iso B, Pho4p, and Cbf1p. $K_d$s in $\mu$M are shown for the sequences NNNNGTG. Sequences NNNGTG are shown in the x-axis with four values each representing the fourth flanking base, as indicated in the legend.

flanking position. Here MAX prefers a cytosine, while Pho4p and Cbf1p prefer a guanine and thymine, respectively; already hinting at considerable differences in the recognition profile of the flanking bases. The overall topographies of the transcription factors also vary. MAX shows the most rugged landscape, with secondary and tertiary peaks near the E-box sequence neighbors CAT and CTC, as well as CAG. Pho4p and Cbf1p, on the other hand, show much flatter responses than MAX with no considerable secondary peaks. Furthermore, MAX shows affinity spikes for sequences

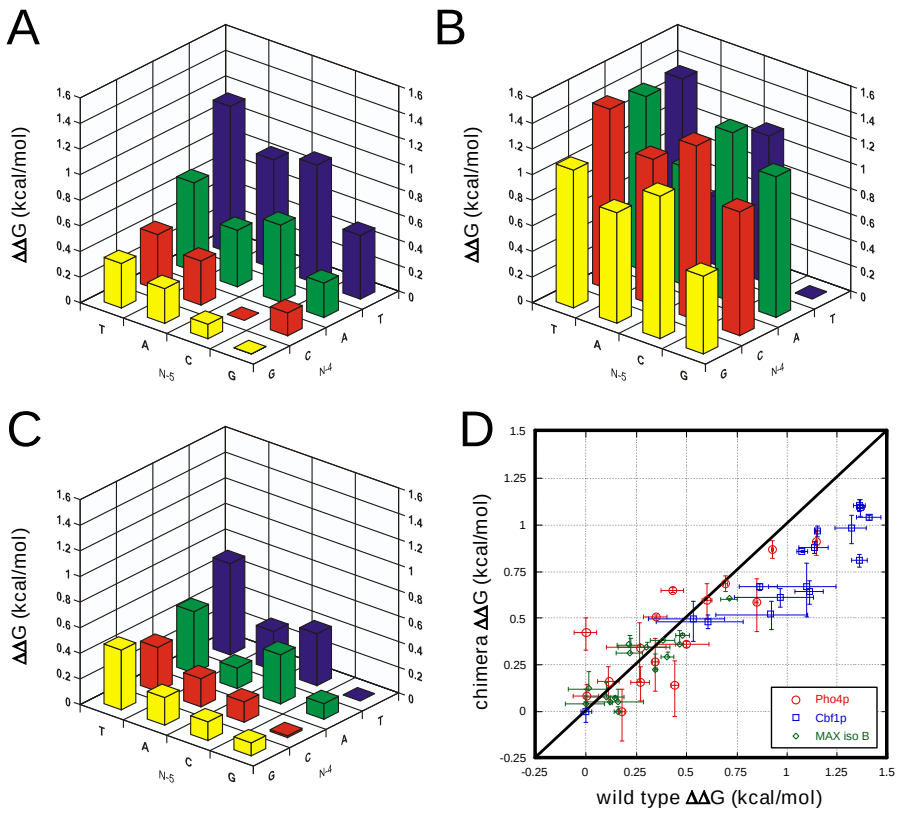



Figure 8.6: The flanking base specificity for bases $N_{-5}N_{-4}$CACGTG. Panels A–C show the landscapes for Pho4p, Cbf1p, and MAX iso B, respectively. Panel D shows a comparison of wild-type transcription factors to MAX iso B chimeras with the indicated basic region substituted for the wild-type sequences.

CACG, CATG, CACC, and CATC, indicating plasticity in the E-box motif previously not found using other methods. In other words, MAX can recognize a split E-box motif with a central single base insertion. Whether this is accomplished through structural changes in the loop domain, for example, or due to diffusional modes is currently not known, but could be interrogated by using bZip transcription factors which lack the loop domain. Also it should be noted that MAX and Pho4p are known to have conserved amino acid residues that make base-specific contacts with DNA. It is thus interesting to see considerable differences in the overall topographies

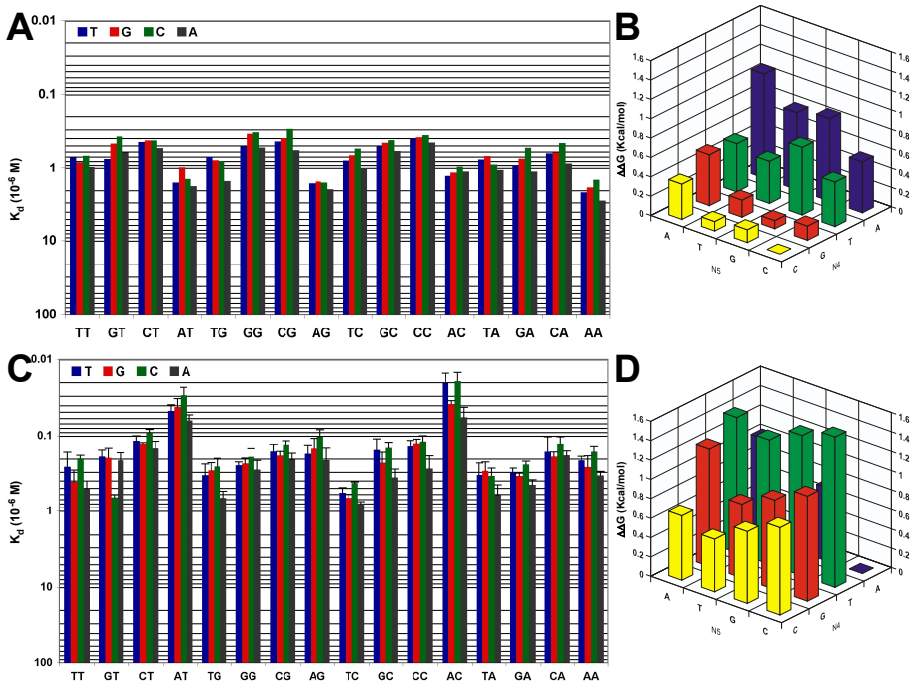



Figure 8.7: Comprehensive 3-mer flanking base landscapes for Pho4p and Cbf1p, Panel A and C respectively. Panel B and D represent a 2-mer subset extracted from the 3-mer datasets for comparison to Figure 8.6.

of these two transcription factors, with MAX showing a more spiked response and Pho4p a very specific response to only the E-box. Here again the reasons for the differences are not known, but could potentially originate from higher-order structures in the transcription factors. Because of the symmetry of the E-box structure and the homodimeric nature of the transcription factors studied, the 4-mer landscape shown in Figure 8.5 is a comprehensive measurement of the known sequence recognition profile of these transcription factors.

In the case for Pho4p and Cbf1p it was interesting to see almost identical energy topographies, which could not explain the differing *in vivo* function of the two transcription factors. Pho4p is known to function in regulating phosphate metabolism

[64, 65], whereas Cbf1p is supposed to regulate methionine synthesis, as well as chromosome structure, by binding to kinetochores [66, 67, 68]. To understand the extent to which these transcription factors recognize flanking bases, DNA libraries covering all possible 2-mer (NNCACGTG) and 3-mer flanking bases (NNNCACGTG and CACGTGNNN) were tested. The observed differences in sequence recognition over a 2-mer space are shown in Figure 8.6. Here all 16 sequences of positions $N_{-5}$ and $N_{-4}$ ($N_{-5}N_{-5}$CACGTG) and their corresponding affinity values (in $\Delta\Delta$Gs) are shown for the transcription factors Pho4p (Panel A), Cbf1p (Panel B), and MAX isoform B (Panel C). Marked differences in flanking base recognition are observable between Pho4p and Cbf1p. The 3' 3-mer library (Figure 8.7) showed the same trend as the 5' 2-mer library. These flanking base measurements extended the consensus sequences for Pho4p and Cbf1p to CCCACGTGGG and [A/G]GTCACGTGAC[T/C], respectively, effectively doubling the known motif in the case of Cbf1p.

To understand whether the basic region amino acid sequence was primarily responsible for these various recognition profiles, basic region chimeras were synthesized. These chimeras consisted of the MAX iso B backbone and the basic regions of Pho4p, Cbf1p, as well as MAX iso B as a positive control. The recognition profiles of the chimeras to the 5' 2-mer libraries were compared to the wild-type topographies (Figure 8.6 D). Overall the chimeras recovered similar sequence-recognition profiles as their wild-type counterparts, particularly Phop4 and, of course, the positive MAX iso B control. Interestingly Cbf1 chimeras showed a considerably lower specificity than their wild-type counterparts, again possibly indicating contributions of other

domains of the protein to the binding, as the sequence topographies are not changed but the elevations are changed proportionally. This experiment indicated that sequence specificities are defined by the amino acid sequence of the basic region, and that the affinities can be modulated not only by the primary basic region sequence, but also by the overall structure of the protein.

The biophysical properties for 4 bHLH transcription factors were comprehensively determined in unprecedented definition. Absolute binding affinities were obtained for hundreds of DNA sequences comprehensively covering the E-box motif, as well as the bases flanking it. These affinities form a basic description of transcription factor function, and present what sequences these transcription factors can bind. Not only were the affinities determined for sequences covering the known E-box motif, but the motif itself could be extended for all transcription factors tested. Indeed, for Pho4p and Cbf1p, the flanking bases proved to be the most important feature of the consensus motif, as the bases covering the central E-box showed no difference in their topography.

### 8.2.3 Binding Site Prediction

### 8.2.4 Yeast Genomic Binding Site Prediction

Having obtained complete biophysical descriptions for the transcription factors Pho4p and Cbf1p, it needed to be established whether these measurements could be used to predict what target genes each transcription factor regulates *in vivo*. A very simple model was used, based solely on the measured binding energy landscapes and the

yeast genomic DNA sequence. Instead of using the affinity of the transcription factor to a specific target sequence directly, probabilities of binding ($P_i$) were established (Equation 8.4). $P_i$s depend on $K_{d,i}$, the affinity of transcription factor X to sequence i as well as on [X], the concentration of the transcription factor. As the *in vivo* concentration of X is exceedingly difficult to establish, it is generally set equal to $K_{d,ref}$, the affinity to the consensus sequence (Equation 8.5), yielding a $P_{ref}$ of 0.5. It should be noted that this assumption lies within an order of magnitude of the known concentration for Cbf1p. Cbf1p is reported to have a cellular concentration of 6890 molecules [69] or 5.5 nM, assuming a cell size of 2pL, which compares well to the measured $K_{d,ref}$ of 16.63 nM. Equation 8.5 can be rewritten in terms of the $\Delta\Delta G_i$ (Equation 8.7) using the equivalents in Equations 8.1–8.3.

$$K_{d,ref} = e^{-\Delta G_{ref}/RT} \tag{8.1}$$

$$\Delta G_i = \Delta G_{ref} - \Delta\Delta G_i \tag{8.2}$$

$$K_{d,i} = e^{-(\Delta G_{ref} - \Delta\Delta G_i)/RT} \tag{8.3}$$

$$P_i = \frac{[X]}{K_{d,i} + [X]} \tag{8.4}$$

$$P_i = \frac{K_{d,ref}}{K_{d,i} + K_{d,ref}} \tag{8.5}$$

$$= \frac{e^{-\Delta G_{ref}/RT}}{e^{-(\Delta G_{ref} - \Delta\Delta G_i)/RT} + e^{-\Delta G_{ref}/RT}} \tag{8.6}$$

$$= \frac{1}{e^{\Delta\Delta G_i/RT} + 1} \tag{8.7}$$

To calculate a probability of binding to a regulatory region consisting of several

hundred bases, individual probabilities of binding to each sequence window are calculated and then integrated to yield a probability of occupancy, $P_{occ}$ (Equation 8.8-10).

$$P_{occ} = 1 - \prod_{i=1}^{windows} (1 - P_i) \tag{8.8}$$

$$1 - P_i = 1 - (\frac{1}{e^{\Delta\Delta G_i/RT} + 1}) = \frac{e^{\Delta\Delta G_i/RT}}{e^{\Delta\Delta G_i/RT} + 1} = \frac{1}{1 + e^{-\Delta\Delta G_i/RT}} \tag{8.9}$$

$$P_{occ} = 1 - \prod_{i=1}^{windows} (\frac{1}{1 + e^{-\Delta\Delta G_i/RT}}) \tag{8.10}$$

$P_{occ}$s were calculated for each regulatory region of 5814 yeast ORFs, using binding energy landscapes determined for Pho4 and Cbf1p and Equation 8.10. Initially the method was tested on known target genes of Pho4p and Cbf1p by calculating $P_i$s over a a range of -800 bps to the start codon of each ORF. For Pho4p 29 target genes, including the PHO and VTC family, were chosen. For Cbf1p a lax gene set included 70 genes and a strict set consisted of 17 genes. For each of these gene sets, binding sites with probabilities above 0.1 were plotted as histograms to visualize the prevalence of binding predicted sites to occur in specific regions of the regulatory sequence (Figure 8.8). For Cbf1p, binding sites can be found extending up to 800 bps away from the ORF start codon, but most sites fall within -600 bps of the start codon. In Pho4p all sites are much more localized, and fall between -150 bps and -350 bps. This is consistent with reported nucleosome-free regions [70] and a bona fide transcription factor binding to these regions. Cbf1p's binding sites are more diffuse and span a broader region, which can be explained by the fact that Cbf1p has

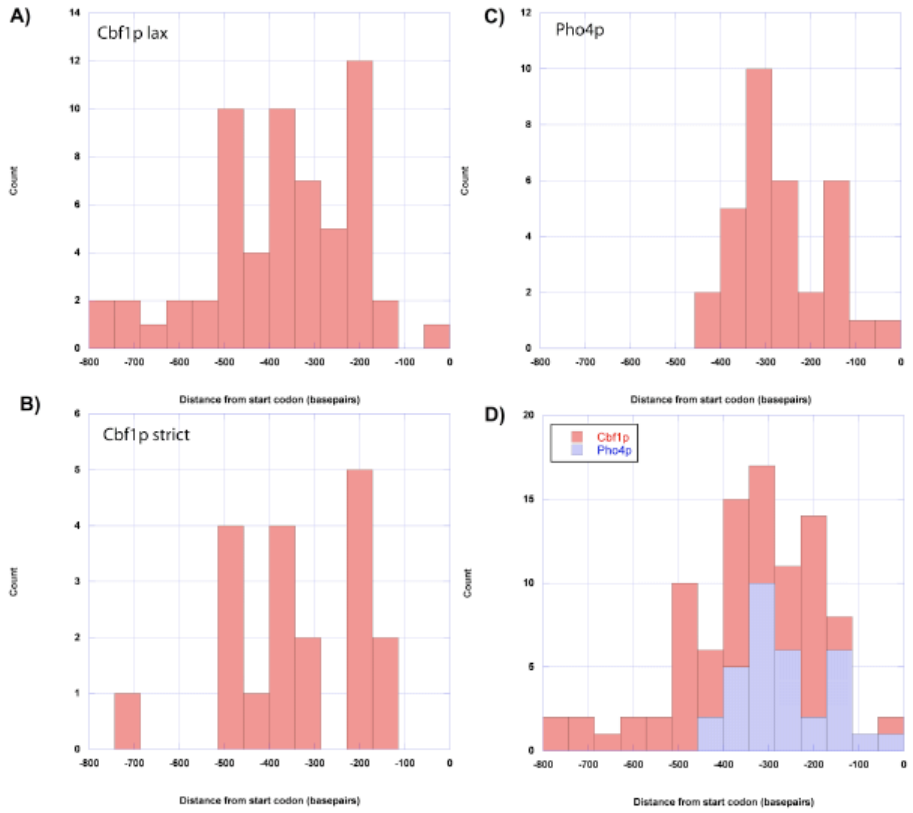



Figure 8.8: Binding site distributions for genes likely to be regulated by Pho4p or Cbf1p. The Cbf1p lax gene set consisted of 70 target gene,s whereas the strict set contained 19 genes. For Pho4p the set consisted of 29 genes. Shown are distributions of binding sites with a $P_i$ of 0.1 or higher. Panel D is a summed histogram for Cbf1p and Pho4p from Panels A and C. The x-axis on all histograms indicates the distance to the start codon of each gene tested.

been reported to function in chromatin remodeling and thus may not be required to bind to regions already cleared of nucleosomes [66]. Therefore, $P_{occ}$s were calculated for Pho4p and Cbf1p on a genomic scale, ranging from the start codon of each ORF to -600 bps and -800 bps, respectively. Using a cutoff criterion of a $P_{occ}$ value of at least 0.2, target genes were chosen for Pho4p and Cbf1p (Figure 8.9). For Pho4p and Cbf1p, 38 and 24 genes were found, respectively. The first indication that these datasets are of high quality is the fact that the Pho4p and Cbf1p gene sets have zero

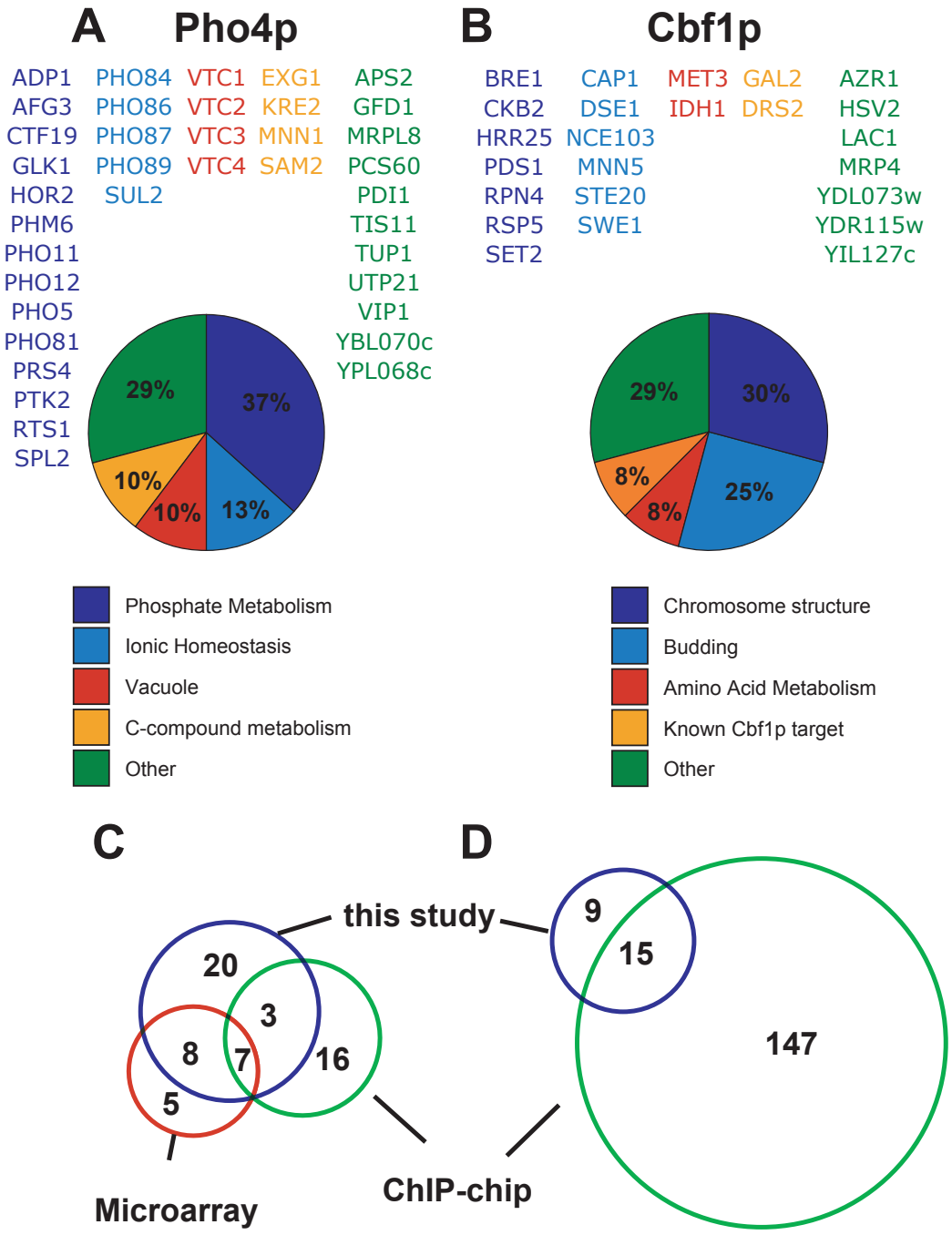



Figure 8.9: *In vivo* function prediction for Pho4p and Cbf1p. (A–B) Genes with regulatory sequences determined to be bound by our *in silico* method. All genes shown here have a Pocc of above 0.2 and a sensu stricto conservation score of 25% or above. Pie charts show the functional distribution of the gene sets. (C–D) Venn diagrams comparing our predicted gene sets to gene sets determined using expression micro-arrays and ChIP-chip.

Pho4p

| FUNCTIONAL CATEGORY | P-VALUE |
|---|---|
| 34.01.03 homeostasis of anions | 1.25E-08 |
| 34.01.03.03 homeostasis of phosphate | 1.06E-07 |
| 20.01.01.07 anion transport (Cl, SO4, PO4, etc.) | 2.63E-07 |
| 20.01.01.07.07 phosphate transport | 2.76E-07 |
| 01.04 phosphate metabolism | 1.72E-06 |
| 01.04.01 phosphate utilization | 4.91E-05 |
| 42.25 vacuole or lysosome | 9.95E-05 |
| 01.05.01 C-compound and carbohydrate utilization | 2.33E-04 |
| 34.01 ionic homeostasis | 4.82E-04 |
| 01 METABOLISM | 6.35E-04 |
| 34 INTERACTION WITH THE CELLULAR ENVIRONMENT | 8.56E-04 |
| 20.01 transported compounds (substrates) | 1.28E-03 |
| 20.01.01 ion transport | 1.34E-03 |
| 01.05 C-compound and carbohydrate metabolism | 1.63E-03 |
| 14.07.02.01 O-directed glycosylation | 3.11E-03 |
| 20 CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES | 3.34E-03 |

Cbf1p

| FUNCTIONAL CATEGORY | P-VALUE |
|---|---|
| 40 CELL FATE | 3.25E-05 |
| 40.01 cell growth / morphogenesis | 1.47E-04 |
| 42 BIOGENESIS OF CELLULAR COMPONENTS | 4.15E-04 |
| 43.01.03.05 budding, cell polarity and filament formation | 6.56E-04 |
| 10.03.01.01.09 G2/M transition of mitotic cell cycle | 6.98E-04 |
| 43 CELL TYPE DIFFERENTIATION | 7.62E-04 |
| 43.01 fungal/microorganismic cell type differentiation | 7.62E-04 |
| 43.01.03 fungal and other eukaryotic cell type differentiation | 7.62E-04 |
| 14.07 protein modification | 9.14E-04 |
| 10.03.01.01 mitotic cell cycle | 2.51E-03 |
| 01.02.01.14 conjunction of sulfate | 3.56E-03 |
| 14.07.09 posttranslational modification of amino acids (e.g. hydroxylation, methylation) | 3.75E-03 |
| 14.07.03 modification by phosphorylation, dephosphorylation, autophosphorylation | 3.88E-03 |
| 14 PROTEIN FATE (folding, modification, destination) | 4.01E-03 |

Figure 8.10: Functional enrichments of the gene sets shown in Panels A and B of Figure 8.9 for Pho4p and Cbf1p. Shown are the significant enrichments returned by the MIPS FunCat server. Blue entries refer to functions consistent with Pho4p and red entries stand for Cbf1p-relevant functions.

overlap. In other words, the measured binding energy landscapes, particularly the landscapes describing flanking base specifities, were sufficient to separate the function of the otherwise very similar transcription factors. Furthermore, when looking at the functional enrichment of these gene sets using Munich's information center for protein sequences (MIPS) yeast functional catalogue (FunCat), the results show that each transcription factor regulates a very defined subset of target genes with similar function (Figure 8.10). For Pho4p, genes involved in phosphate metabolism predominate the predicted gene set–particularly the PHO and VTC familes, with the latter involved in vacuole regulation. Other enriched functional categories are ionic homeostasis and C-compound metabolism. For Cbf1p, genes involved in chromosome

structure and budding predominate. Two genes involved in methionine synthesis, MET3 and IDH1, are also found. Finally, the datasets were compared to existing experiments based on gene expression arrays [64, 65] and large-scale ChIP-chip experiments [71]. For Pho4p these two datasets only find 7 genes in common, whereas the gene sets determined here cover those 7 genes as well as agree with 8 additional genes in the gene expression analysis and 3 additional genes in the ChIP-chip set. For Cfb1p only the ChIP-chip set was available and here overlap between the datasets is minimal, with the ChIP-chip dataset finding a large number of probably erroneous targets. The approach described here thus presents an accurate method for unbiased determination of target genes for transcription factors. It should be mentioned that the gene expression analysis and ChIP-chip methods are inherently different approaches in that they rely on data obtained from *in vivo* measurements. It should be of interest that the ChIP-chip gene sets predicted by Harbison et al. [71] were obtained by generating position weight matrices (PWMs) from the observed DNA fragments pulled down by the transcription factor. These PWMs where then used in a very similar fashion to the binding energy landscapes determined here. It is thus obvious that PWMs obtained from ChIP-chip data are not as efficient in predicting transcription factor function as binding energy landscapes based on physical binding affinities, a point that will be discussed in more detail in Section 8.2.5.

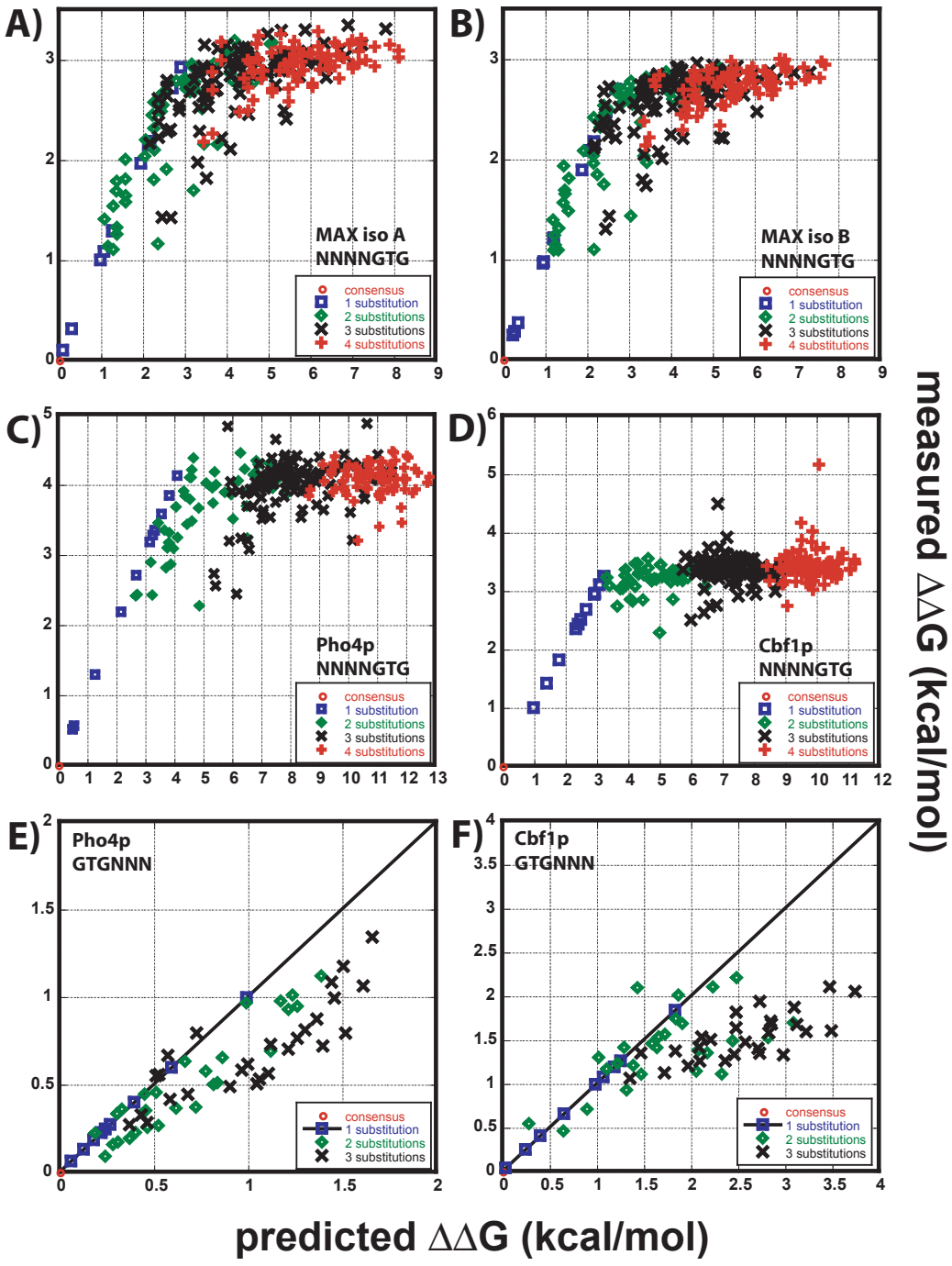



Figure 8.11: Panel A–D show comparisons of PWM predictions versus measured values for NNNNGTG sequences and transcription factors MAX iso A, iso B, Pho4p, and Cbf1p, respectively. Panels E and F show flanking bases sequences GTGNNN measured for Pho4p and Cbf1p, respectively. The number of substitutions of each sequence respective to the consensus sequence is indicated, where the single substitutions represent the PWM training set.

### 8.2.5 Position Weight Matrices Versus Binding Energy Landscapes

There are several approaches to describing how a transcription factor recognizes and binds DNA. The first-order approach is to argue that a transcription factor binds to one sequence, namely its consensus sequence. Finding consensus sequences of transcription factors was and still is a primary goal in understanding the biophysical properties of transcription factors. The consensus sequence can be defined as the DNA sequence to which a transcription factor binds most strongly. But transcription factors can also bind to consensus sequence neighbors with considerable affinity. Takeda and Sarai studied Cro and $\lambda$ repressor of *E.coli*, measuring the affinities to various substitutions of the consensus sequence [72, 73]. They also found that a simple addition of $\Delta\Delta$Gs for each substitution was predictive of the experimentally determined value of the double substitution. This gave rise to the assumption that position weight matrices (PWMs) [74, 75, 76, 77] or WebLogos [78, 79] could be used to describe how a transcription factor binds to all possible sequences. A PWM consists of four entries describing the base preference of a transcription factor at each position of its binding site. To calculate the affinity to any sequence, the entries describing the affinity of the specific base of the new sequence at each position are added together. More recently, with the advent of higher throughput technologies, the assumption that bases are independent and thus can be described as a simple two-dimensional matrix has been contested [80, 81, 82].

Having measured a complete 4-mer space for MAX iso A, iso B, Pho4p, and Cbf1p,

it was possible to test whether bases are recognized dependently or independently by the transcription factors. Since a complete 4-mer space contains all possible single base substitutions, a PWM containing actual affinity values could be extracted from the datasets. These PWMs were then used to calculate the affinity of every possible sequence in the 4-mer set. These values, calculated based on the assumption that bases are independent of one another, were then compared to the actual biophysical affinities of these sequences (Figure 8.11). Panels A–D of Figure 8.11 show the results for both MAX isoforms, Pho4p, and Cbf1p for the sequence library NNNNGTG. In all four cases only a fraction of predicted values actually agree with the values determined experimentally. One major region of the graph lies in the low-affinity domain of about 10 $\mu$M or $\Delta\Delta$G of 2.5–4.5 kcal/mol. The plateau seen in Figure 8.11 relates to the non-specific binding regime observed in the binding energy landscapes. In other words, a transcription factor retains a certain nominal affinity to DNA dominated by non-specific electrostatic interactions with the DNA backbone that is not predicted by PWMs. Most if not all predictions of Cbf1p fall into this regime since there were no additional secondary peaks observed in the binding energy landscape. In the MAX datasets, and to a lesser extent in the Pho4p dataset, a second regime of affinities that lie perpendicular to the diagonal and in intermediate affinity ranges is apparent. These cases are the most interesting, as here the PWMs predict a considerably lower affinity (up to 2–3kcal/mol) than is experimentally observed, indicating that transcription factors can adjust their modes of binding to certain families of sequences and as a result bind them more strongly than predicted. The same trend

is observed in the flanking base 3-mer dataset for Pho4p and Cbf1p, only here the erroneous predictions are observed starting in the high-affinity regimes. Additionally, and supporting the idea that a transcription factor can find additional modes of binding, is the fact that PWMs always over-predict the loss in affinity. In other words, the observed physiological affinity of a transcription factor is always equal to or higher than the predicted affinity assuming independence.

## 8.3 The Basic Region

The basic region is the structural section of a bHLH transcription factor making sequence-specific contact with DNA. The basic region structure is an alpha-helix inserted into the major groove of DNA, a common method used by proteins to read out the genetic code of DNA (see Figure 8.2). An alpha-helical turn is roughly 3.6 amino acids long, therefore every fourth to third amino acid sidechain points in the same direction. In the case of the basic region it is these residues that make sequence-specific contact with DNA. It should thus be possible to learn the DNA recognition code of bHLH transcription factors by an exhaustive screen of the effects of amino acid substitutions over a reasonably large DNA sequence space. As there are only approximately 4–5 amino acids that make base-specific contact, the number of single amino acid mutations is in the range of 80–100 mutants. To first understand what basic regions are naturally occurring, all known bHLH transcription factor sequences were collected from databases, the basic regions extracted, and a sequence alignment performed (Section 8.3.1). To understand the function of each position in the basic

region and the consequence of each possible single residue mutation, an exhaustive experimental screen was run on the basic region mutants (Section 8.3.2).

### 8.3.1 Bioinformatic Sequence Alignment

To understand the naturally occurring basic region diversity, all 206 catalogued human bHLH transcription factor amino acid sequences were obtained and their basic regions extracted. These basic regions were then aligned according to sequence similarity (8.12). This initial analysis contained a number of duplicate basic region entries sometimes due to actual duplicate transcription factor entries, but often because two transcription factors have identical basic regions but different dimerization or activation domains.

To then further distill the dataset, all duplicate basic regions were removed and the remaining basic regions re-aligned, again according to sequence similarity. The resulting basic-region-sequence similarity tree is shown in Figure 8.13. In this tree, nine major branches of similar basic regions are apparent. A table shows the basic regions of each branch and the observed residue conservation amongst these sequences. Conserved residues are shown in yellow in each table. Blue and green colors indicate partially conserved residues. To further reduce the information content of this basic region alignment tree, each branch is summarized in a table (Table 8.3) showing the branch number, the consensus sequence, and the residues that likely make base specific contact (denoted by a 'C'). Well-known members are also shown for each branch.

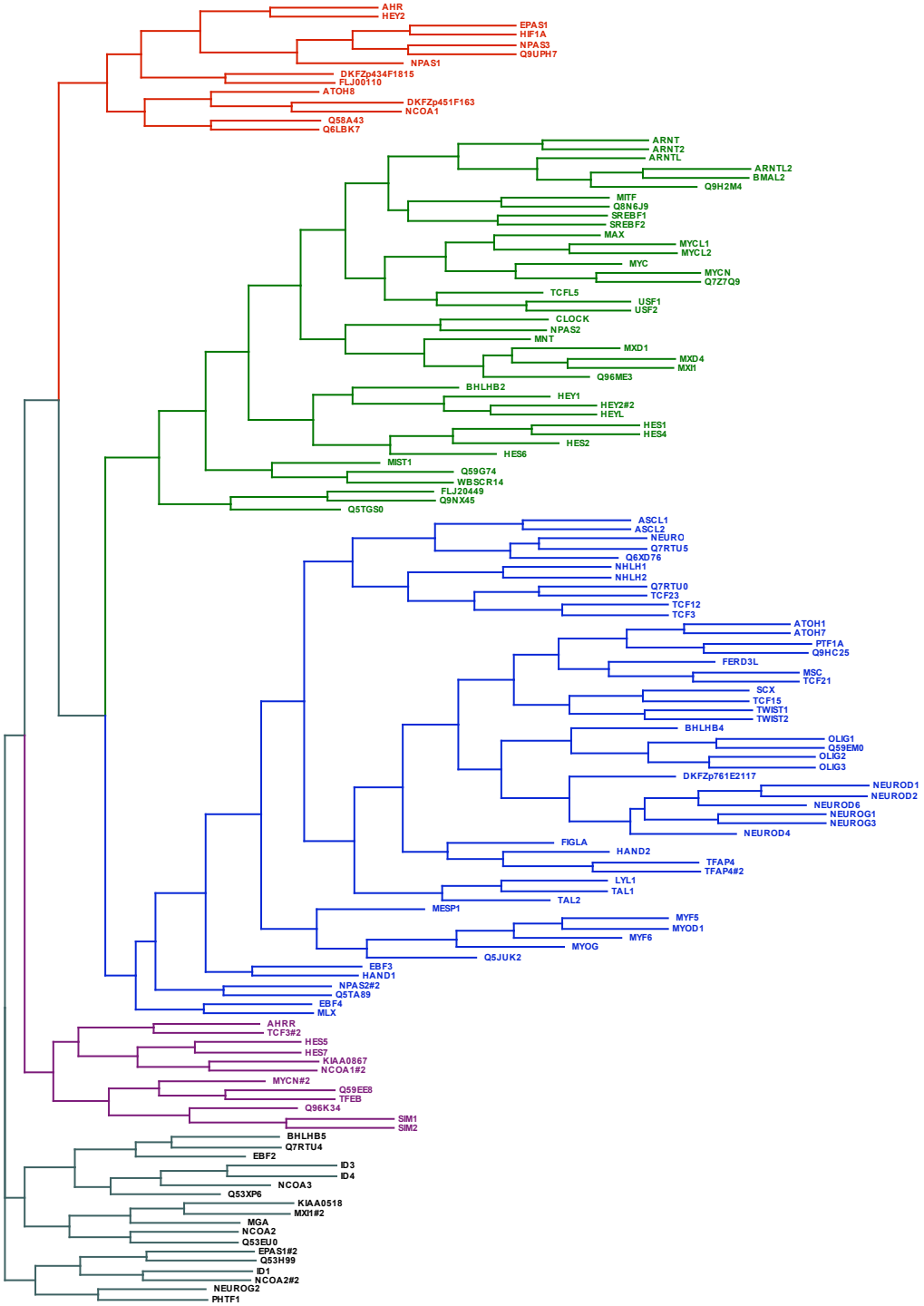



Figure 8.12: Phylogenetic alignment of all obtained bHLH basic regions

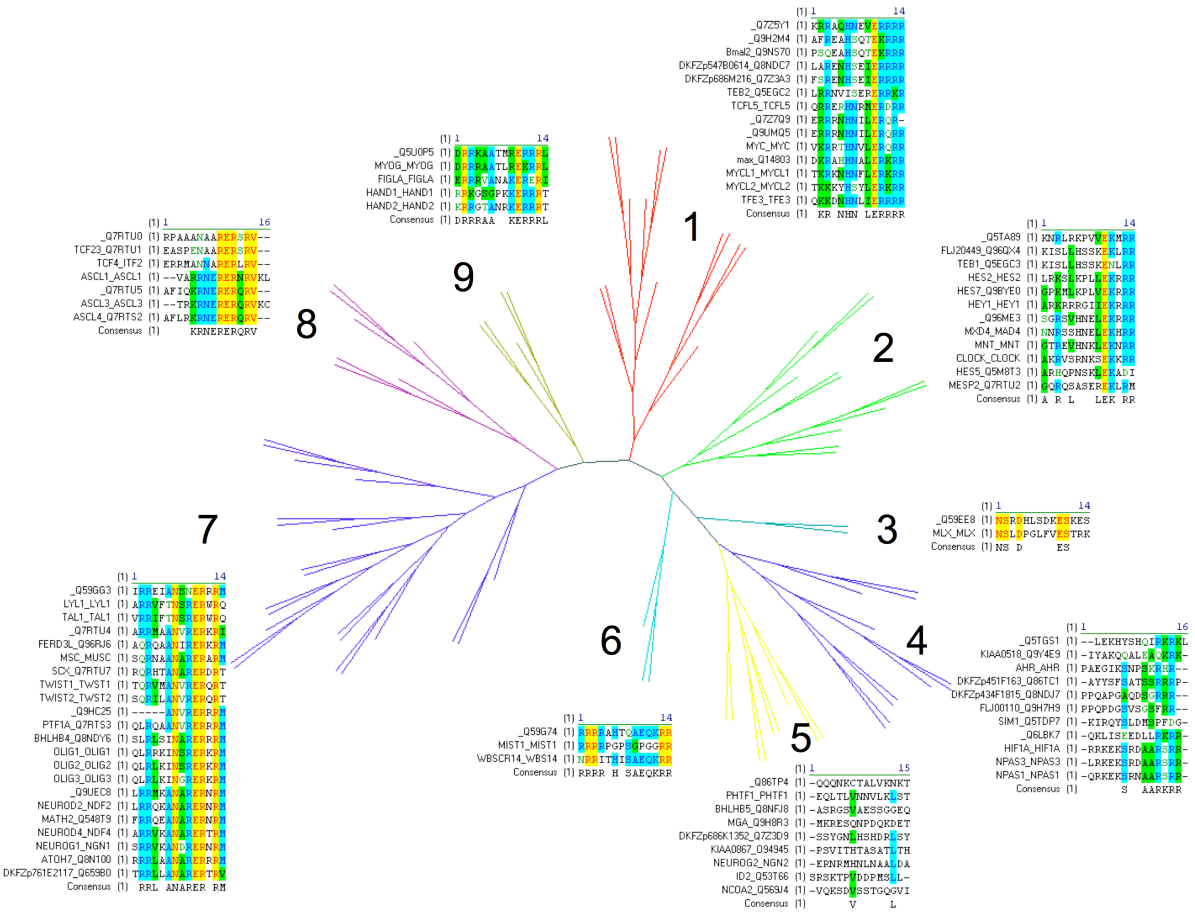



Figure 8.13: Phylogenetic representation of all tested basic regions. Here all redundant names as well as basic region sequences were removed from the alignment

A glutamate in position 10 is conserved in all branches except for branches 4 and 5. These two branches contain basic regions of bHLH transcription factors that do not bind to DNA, such as the Id family. As these transcription factors lack a basic region, the sequence alignment returned only noise, with no residue conservation in either branche. For all other branches the basic region sequence is expected to bind DNA. The presence of a glutamate in position 10 is indicative of this, as E10 must be present in order for the basic region to bind DNA. In branch #1, residues contacting DNA are: R3, H6, E10, and R14. MAX and Pho4p fall into this category, and Cbf1p

| Branch # | 1 Position 14 | | | | | | | | Members |
|---|---|---|---|---|---|---|---|---|---|
| | C | C | C | C | C | C | C | C | |
| 1 | -K\|R\|-N\|H\|N-L\|E\|RRR\|R\| | | | | R | H | E | R | MAX, MYC, TFE3 |
| 2 | A-\|R\|-L\|-\|--L\|E\|K-R\|R\| | | | | R | - | E | R | HES2, MNT, TEB1 |
| 3 | NS\|-\|D-\|-\|---\|E\|S--\|-\| | | | | - | - | E | - | MLX, NRC3 |
| 4 | --\|-\|--\|S\|--A\|A\|RKR\|R\| | | | | - | S | A | R | AHR, NPAS1, SIM1 |
| 5 | --\|-\|-V\|-\|---\|-\|L--\|-\| | | | | - | - | - | - | ID2, NEUROG2,PHTF1 |
| 6 | RR\|R\|R-\|H\|-SA\|E\|QKR\|R\| | | | | R | H | E | R | MIST1, WBS14, TFP4 |
| 7 | -R\|R\|L-\|A\|NAR\|E\|R-R\|M\| | | | | R | A | E | M | TWIST1, NEUROD1, MSC |
| 8 | --\|-\|-K\|R\|NER\|E\|RQR\|V\| | | | | - | R | E | V | TCF4, ASCL1, ASCL3 |
| 9 | DR\|R\|RA\|A\|--K\|E\|RRR\|L\| | | | | R | A | E | L | MYOG, HAND1, FIGLA |

Table 8.3: Table showing the consensus sequences of the 9 branches of the above tree. All residues making specific contact with the E-box are denoted by C on top of the column.

is a close sequence neighbor with a K3 instead of a R3. It should be noted that branch #6 is essentially the same as branch #1 in respect to residues contacting bases, except for MIST1, which caused the alignment to be more distant than it should have been. Branch #2 is similar to branch #1 but shows more variation in positions 3 and 6. Branch #7 shows conserved residues in positions 3 and 10, but allows a variety of residues in positions 6 and 14. Interestingly branch 8 shows a conserved valine and arginine in positions 14 and 6 instead of a arginine and histidine. The basic region sequence is also quite unstructured anterior to position 6. And finally branch 9 has residue E10 shifted by one position in respect to the conserved arginine residues in positions 2 and 13, which should be 3 and 14.

This bioinformatic analysis showed that the glutamate in position 10 is ultra conserved and thus must be absolutely required for DNA binding by bHLH transcription factors. Aside from position 10 the other 3 positions expected to make base spe-

cific contact are not as restricted and several amino acids are used, possibly changing which sequence is recognized, as well as varying the affinity to any given sequence. To understand this recognition code an experimental approach was required to elucidate the function of each position and is described in Section 8.3.2.

### 8.3.2 Mutagenesis Screen

To understand the DNA recognition code of the basic region two approaches were envisioned. Both approaches were based on using the MAX isoform B backbone, which readily homo dimerizes and expresses well in wheat germ extract. Furthermore the basic region in MAX iso B is immediately N-terminal and could thus be easily exchanged for any sequence using an overhang extension PCR. The first approach attempted to use all non-redundant basic regions shown in Figure 8.13 and to substitute these for the wild-type MAX iso B basic region. The synthesis of the chimeras was successful, but it became apparent that it was exceedingly difficult to determine whether the pitch of the inserted basic region was correct, and with an incorrect pitch the function of each basic region could not be guaranteed. In retrospect, each basic region should have been centered by using the conserved glutamate at position 10, which should ensure that the inserted basic regions have the correct helical pitch. The second approach taken was to specifically perform a saturation mutagenesis of positions 2, 3, 6, 10, and 14 by substituting all possible 19 non-wild-type amino acids and testing these novel basic region mutants for function. Each approach has its own intrinsic advantages. The use of naturally occurring basic regions would have

given direct insight into the function of each transcription factor, assuming that only the basic region participates in DNA recognition. The mutagenesis approach, on the other hand, represented a more controlled study of the function of each position with an exhaustive coverage of amino acids. It however does not take into account the identity of the unchanged residues, which, even though they do not contact DNA specifically, may still contribute to the overall binding characteristic.

In a pilot experiment, each of the 100 basic region mutants was screened against the 4 possible base substitutions in the position where the modified amino acid is most likely to make base-specific contact (Figure 8.14). Therefore each set of 20 mutations per position was tested against 4 possible target DNA sequences. To obtain high-quality data each of these 400 possible combinations was tested at 8 target DNA concentrations. The resulting data could then be fit with a linear regression, as the target DNA concentrations were chosen to fall in the regime where the binding response is linear. The data shows that each position in the basic region that is predicted to make contact performs a unique function. Position 14 is responsible for dialing in the specific sequence recognized by the basic region. The wild-type amino acid arginine recognizes a cytosine most specifically. Changing this residue to either leucine, asparagine, glutamine, tryptophan, or tyrosine changes the recognition from a cytosine to a thymine. This list includes the three largest hydrophilic-neutral amino acids (N,Q, and Y). Furthermore, when comparing tryptophan with histidine, a similar trend in specificity can be observed, albeit at different relative affinities. Guanine is only recognized by one amino acid, namely methionine. Alanine also

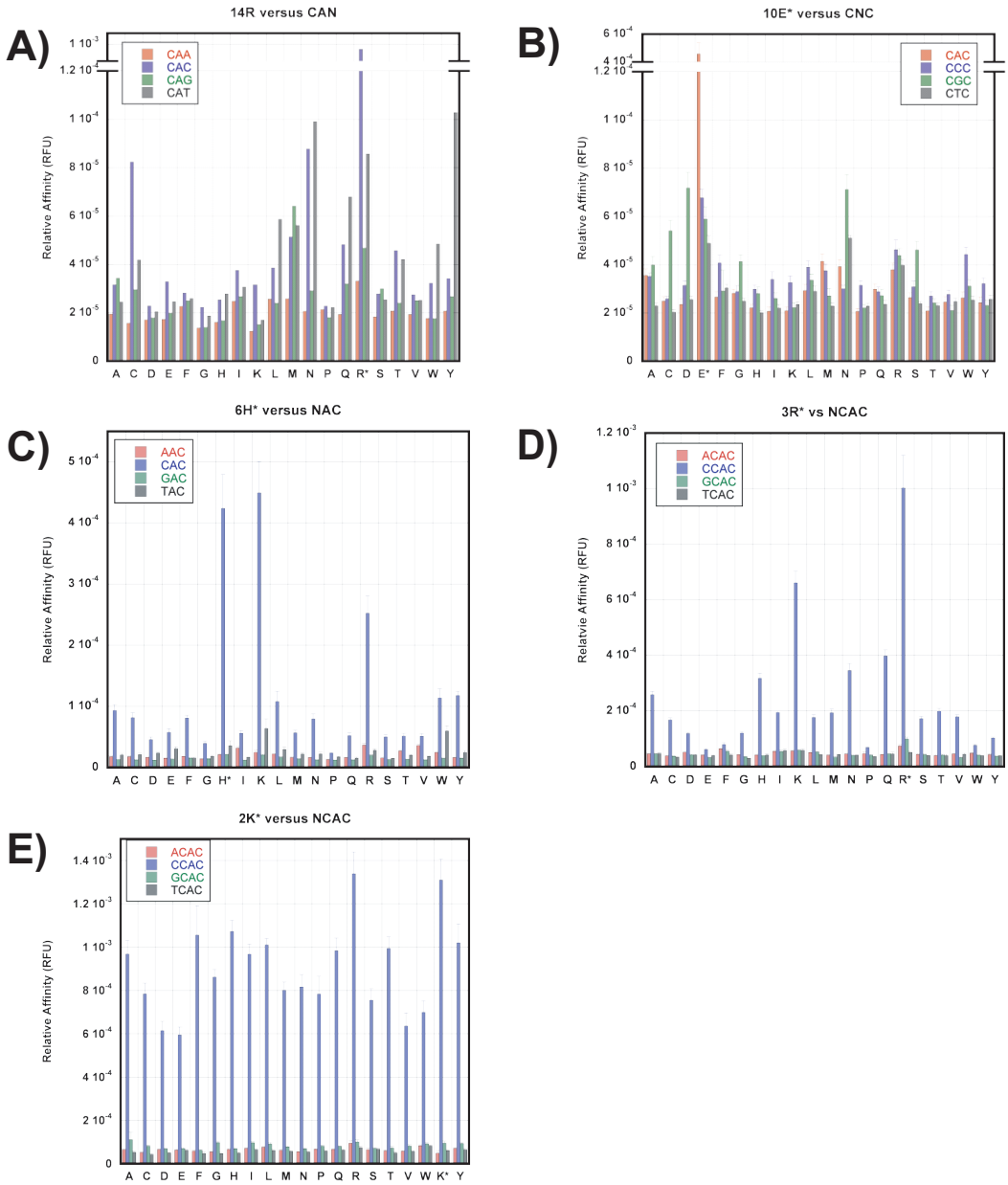



Figure 8.14: Shown are the base specificities of all 20 amino acids for the 5 positions in the basic region predicted to make base-specific contact. The basic region position and target base are indicated in the title of each graph. An asteriks denotes the wild type amino acid in each position.

primarily recognizes guanine but at very low relative affinities. Finally, thymine could not specifically be recognized by any of the 20 possible amino acids, essentially limiting the recognizable sequence space for the central nucleotide to CA[C/A/G].

The binding response of amino acid substitutions in position 10 was markedly different than the response observed for position 14. Here essentially all mutations caused a considerable drop in binding affinity, making E10 an essential component for basic region binding. Several substitutions, including cytosine, aspartate, glycine, asparagine, and serine, caused a recognition change from alanine to guanine. Here small amino acids such as A, S, C, and G dominate, with N and D being slightly larger.

Starting with position 6 no amino acid substitution was able to change the sequence recognition of the basic region. The mutants did exhibit varying degrees of affinity to the wild-type E-box sequence. In position 6, wild-type histidine and lysine showed similar affinities, with arginine being the third-strongest binder. All other amino acids were roughly equal in their affinity, with proline being the least stable. Proline does serve as a negative control as it can not attain the backbone conformation required for helix formation.

Arginine in position 3 shows a similar trend in affinities as position 10, but with different amino acids causing affinity modulation. Here the wild-type residue has the largest affinity, followed by lysine. Alanine, histidine, asparagine, and glutamine show intermediate affinities. Moving out one more position to lysine, affinities change considerably, with all amino acids, including proline, having high affinity to the con-

sensus. On the one hand this indicates that position 2 does not contribute considerably to the overall binding affinity of the basic region. But certain residues such as lysine and arginine are still preferred, most likely because they can form non-specific interactions with the negatively charged DNA backbone. Interestingly, for positions 3 and 2, changing the base in position $N_{-4}$ from cytosine to any of the other three bases caused a complete loss of affinity. This could be due to the fact that position $N_{-4}$ is in fact recognized by a residue other than those in position 3 and 2. It is also possible that the presence of an arginine and a lysine in position 3 and 2 respectively can recover the sequence recognition profile observed for Cbf1p.

Summarizing the above data it became apparent that position 14 functions as the only sequence sensor in the basic region. Here 3 possible bases can be recognized including cytosine, guanine, and adenine. Glutamate in position 10 appears to be essential to basic region function, with a few amino acids seemingly being able to change the base recognized from an adenine to a guanine. In positions 6, 3, and 2 it was not possible to modulate sequence recognition, but affinity could readily be changed by substituting certain amino acid residues. These measurements provided a glimpse into the functions of the various positions on the basic region known to make base-specific contact with DNA. It should be noted that a larger DNA sequence space has to be tested to ascertain that no sequences are missed that might recover binding, as it is most likely that amino acid side chains may contact more than one base at a time. The next phase of the experimental series therefore constitutes testing each mutant against a complete 3-mer DNA library. Up to 6 basic region mutants may

be screened on a single 2400-unit cell device against all 64 possible DNA sequences at 6 concentrations. It will be interesting to see whether this extended screen will reconstitute the intriguing sequence recognition difference observed between wild-type Pho4p and Cbf1p (Figure 8.6). Here Pho4p and Cbf1p prefer CC and GT in positions $N_{-5}N_{-4}$, respectively. The basic region sequences are remarkably similar and the difference in recognition can be essentially limited to residues 2–4 which are KRE and RKD in Pho4p and Cbf1p, respectively. Position 4 is most likely not involved in base recognition since the helical position is phase shifted by roughly 180°, leaving the switch of lysine and arginine in positions 2 and 3. It remains to be seen whether changing one of these positions, but measuring it against a larger sequence space, will affect the switch in sequence recognition, or whether a double mutation is required. If the latter possibility should hold true (as it well might) then the necessary sequence space to be explored will explode exponentially, making the problem experimentally intractable. Nonetheless, testing of each mutant against a larger DNA sequence space will provide a good picture of how bHLH transcription factors, and by extension bZIP transcription factors, recognize DNA. Finally similar experimental approaches can be used with the two other main families of transcription factors, namely the Zinc finger and Homeobox family.

## 8.4 bHLH Heterodimers

An intricate aspect of bHLH transcription factor function is their ability to form not only homodimers, but heterodimers as well. Heterodimer formation not only allows

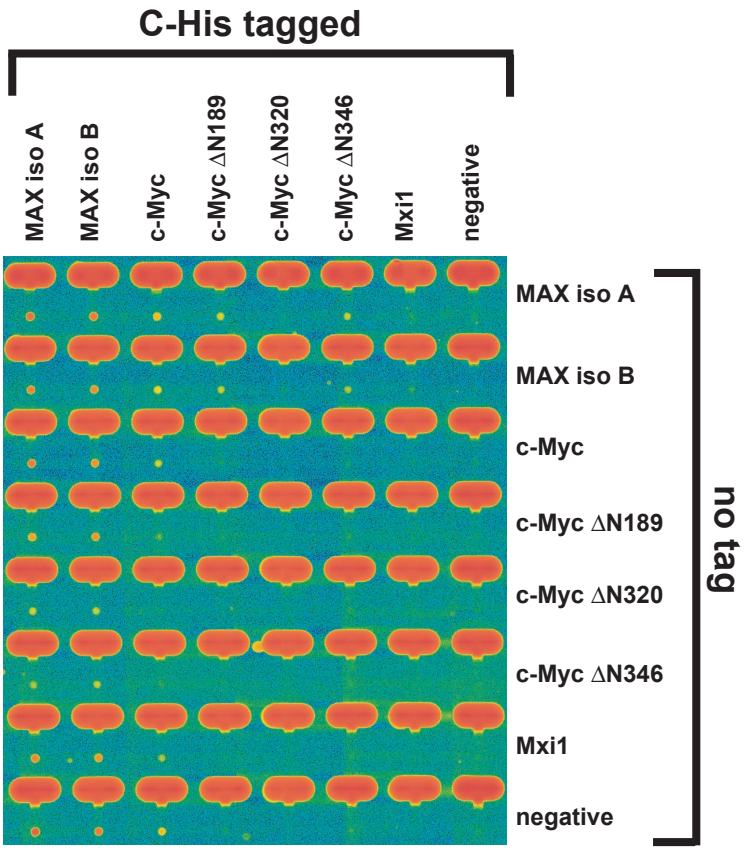



Figure 8.15: Fluorescent scan of a subsection of a DTPAx10 device programmed combinatorially with linear expression templates coding for the transcription factors according to the column and row labels. The fluorescence intensity shows the presence of the a target dsDNA probe labeled with Cy5 pulled down by functional homo- and heterodimers (green-red represent low-high concentrations, respectively).

for a combinatoric sequence space to be explored, but also allows transcription factors to recruit various activation domains. Heterodimer formation was tested on-chip by co-spotting linear expression templates coding for the two transcription factors to be tested, with one of them lacking an epitope tag so that it can only localize to the surface through specific interactions with the second transcription factor, which carries a 6xHis tag. Heterodimers can be detected by directly visualizing the prey transcription factor with a second epitope tag. It is also possible to measure het-

erodimer formation through target DNA pull-down. This second approach is possible as many bHLH transcription factors fail to form homodimers and only bind to DNA if they heterodimerize with a partner bHLH transcription factor. C-Myc for example predominantly forms heterodimers with MAX as well as other members of the family. To test the ability of c-Myc to heterodimerize with MAX, Mxi1, and itself, a combinatoric array of the transcription factors MAX iso A, MAX iso B, c-Myc, C-MycΔN189, C-MycΔN320, C-MycΔN346, and Mxi1 iso B was spotted. C-terminally tagged versions of each transcription factor were co-spotted in all possible combinations with un-tagged versions. In this instance a chip design (Figure C.21) was used that allowed for neighbor spotting rather than having to co-spot each template. The transcription factors were then synthesized using wheat germ ITT in the presence of a target ds-DNA oligo labeled with Cy5. Every c-terminally tagged transcription factor localized to the surface and could interact either with itself or with the second transcription factor. If a functional dimer was generated, it pulled down the target DNA, which could be detected on the ArrayworxE (Figure 8.15). It can be seen that both MAX isoforms are able to homodimerize and functionally bind target DNA. C-Myc, on the other hand, shows the highest DNA pull-down in the presence of its heterodimer partners MAX iso A and iso B. This is particularly obvious for the truncated c-Myc versions ΔN189 and ΔN346, where only the presence of MAX iso A and MAX iso B gave rise to signal. Interestingly, the intermediate truncation ΔN320 showed no pull-down across the board. Whether this is due to a structural reason, or simply an experimental artifact due to low expression, could not be ascertained. It should

be noted that the bHLH domain of c-Myc resides in the c-terminal portion of the transcription and remained unchanged by all three truncations. Mxi1, surprisingly, failed to be functional in any transcription factor combination, despite expectation that it would be functional in the presence of c-Myc. Upon closer inspection it was realized that the Mxi1 version used in this experiment was isoform B, which lacks a basic region resembling the Id bHLH transcription factors and thus is indeed expected to be non-functional.

On-chip co-expression of proteins can therefore be used to study binary protein-protein interactions. The experimental approach described here uses the binding of a target dsDNA sequence as a detection mechanism. A more generalizable approach would have been to detect the binding of the prey transcription factor directly using a second epitope tag such as the S-tag approach described in Section 7.3. Another intriguing possibility lies in multiplexing the dsDNA target by introducing a set of closely related sequences varying CANGTG, for example, and using a different fluorophore for each possible sequence. Then, the presence of a functional dimer could be measured as well as its sequence preference. Since it was shown in Section 8.3.2 that the sequence space recognized by the basic region is considerably restricted, only a library consisting of the sequences comprising CANNTG has to be interrogated.

In order to perform a full combinatoric interaction screen of the roughly 200 human bHLH transcription factors, a good cDNA clone source is needed. Many clones are available from OpenBiosystems, but it became apparent that the annotation of those clones is lacking–the inaccurate annotation of Mxi1 being one example. A second

problem with the OpenBiosystems source is the fact that the clones are harbored in a variety of vectors, making design of a standard PCR method non-trivial. Additionally, each vector has resistance to either Amp or Cam, which also complicates the high-throughput handling of these libraries. A second possible source is available from Invitrogen, where the above problems are non-existent. Unfortunately at a cost of about $800 per clone the Invitrogen clones are also prohibitively expensive.

## 8.5 bHLH Kinetics

Kinetic experiments were run to determine the off-rate of the MAX iso A–target DNA interaction. Briefly, MAX was localized to the chip surface and allowed to bind target DNA containing an E-box sequence. Once DNA was bound by the transcription factor, buffer was exchanged and the dissociation of bound DNA was observed using an inverted fluorescent microscope equipped with a PMT (E717-21, Hamamatsu). The PMT output was measured roughly every 50 ms. Controls included measuring the buffer exchange rate, or flush rate, and the bleach rate, yielding 1.614 $\text{sec}^{-1}$ and $1.4 * 10^{-2} sec^{-1}$, respectively (Figure 8.16). A flush rate of 1.614 $\text{sec}^{-1}$ is reasonably fast for most transient interactions, but could be improved upon. The bleach rate is less important, as it is slow enough and simply adds to the observed loss of stable interactions. Using these two rates provides a function of the form $y = A * e^{-k_d * x} + e^{-k_{bleach} * x} + e^{-k_{flush} * x} + B$, where the function decays to a plateau B from a height of A and with a time constant of $k_d$. With this method, duplicate off-rates of $2.2 * 10^{-1} \text{sec}^{-1}$ and $2.6 * 10^{-1} \text{sec}^{-1}$ were measured for MAX iso A and E-box DNA. Reported off-rates

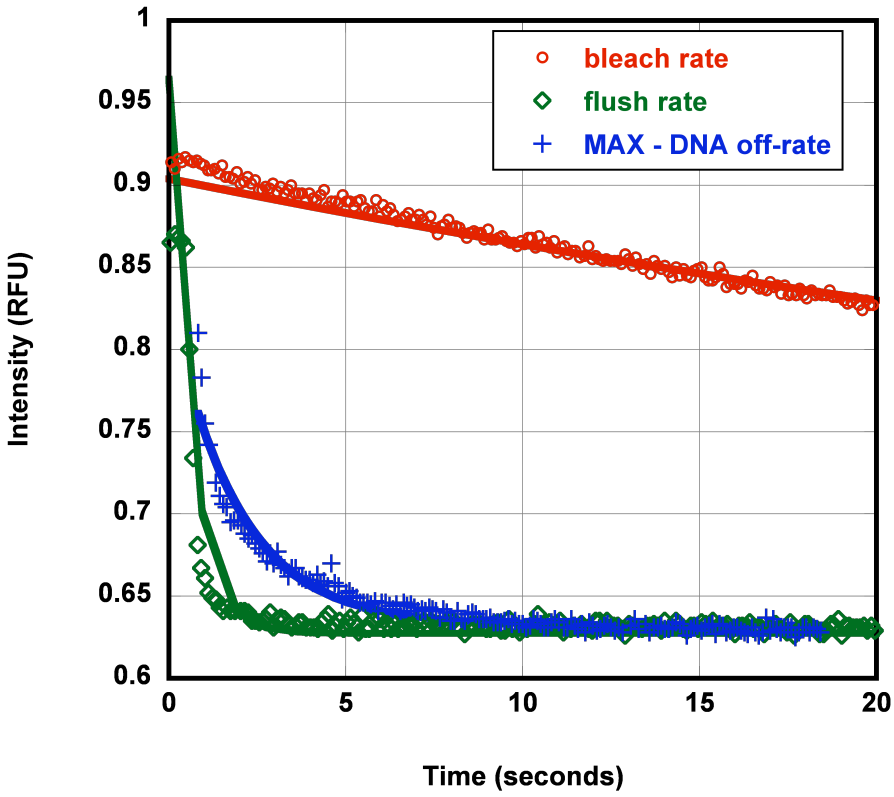



Figure 8.16: The fluorescent intensity of target DNA was measured in real-time using a PMT, so that the relative fluorescent units have the unit of Volts. Two control measurements were taken. First a bleach rate of the Cy5 fluorophore was determined without active flushing. A flush rate was determined by washing away unbound solution phase target DNA. An actual off-rate measurement of MAX - DNA is shown by the blue trace, where DNA was allowed to bind to surface localized MAX transcription factors. All traces were fit with exponential decays (solid lines) as described in the text.

for bHLH transcription factors and DNA vary widely. Spinner et al. report a $k_{off}$ of 0.875–12.3 $\text{sec}^{-1}$ for E12 [83], Grinberg et al. $1.4 * 10^{-3}\text{sec}^{-1}$ and $1.1 * 10^{-2}\text{sec}^{-1}$ for TFE3 and E47 [84], respectively, and Park et al. use off-rates of $3.2\text{–}3.4 * 10^{-2}\text{sec}^{-1}$ for Myc/MAX-DNA and MAX/MAX-DNA [85].

The above described real-time approach to measuring off-rates is easily implemented on a microfluidic device. It would be possible to automate the interrogation

of the device and thus increase the throughput to dozens of interactions. Using MITOMI for measuring off-rates, as described in Section 7.4, would be more appropriate to understand larger numbers of interactions, particularly if individual interactions are long-lived. For permanent interactions, long interrogation durations are necessary, which considerably slow down serial approaches. Interestingly MITOMI-based measurements may also be able to capture faster dissociations, as the flush rate can be decoupled from the actual measurement. The above-measured flush rate of $1.6\text{sec}^{-1}$ lies within the expected range for non-consensus target DNA sequences. Overall the microfluidic methods described are well suited for kinetic measurements of bHLH–DNA interactions and can be adjusted to the specific requirements of the interaction to be investigated.

## 8.6 Other Transcription Factors

### 8.6.1 CREB

The cAMP response element-binding protein (CREB) is a member of the bZIP family of transcription factors (see Figure 8.2) and is implicated in a wide variety of cellular functions, as cAMP is a major transducer of signals received by surface receptors. CREB1 was chosen not only because of its general importance in cellular function, but also because it extends the applicability of MITOMI to a second major family of transcription factors. CREB1 is known to bind a palindromic sequence of TGACGTCA [86, 87]. Interestingly here $N_{-3}$ is a guanine, not a cytosine, indicating that the basic

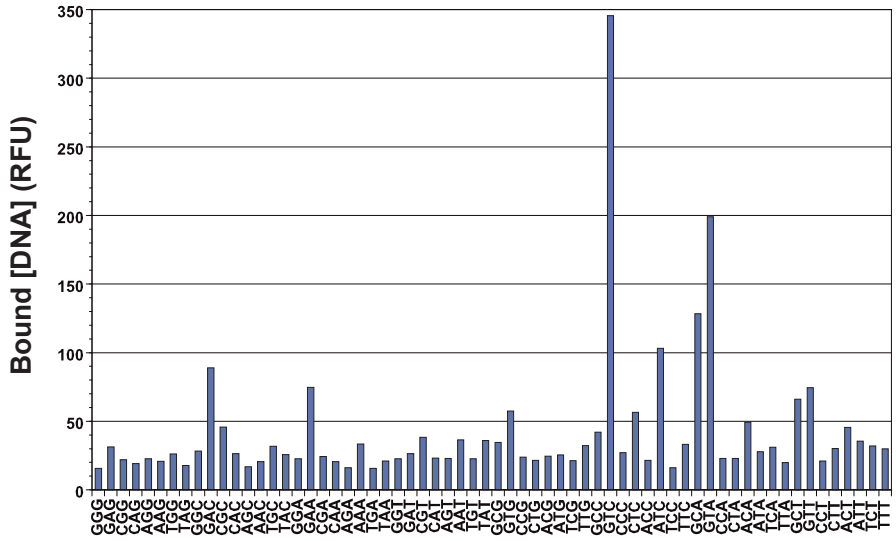



Figure 8.17: The halfsite specificity of CREB was tested using a GACNNN library. Here the surface-bound DNA concentration is plotted, instead of actual affinity constants.

region conserved in bZIPs was able to find a sequence specificity that could not be recovered using the bHLH basic region mutants (see Section 8.3.2). Whether this is due to the slightly different overall structure of the transcription factor (lack of a loop) or is due to different conserved amino acid residues in the basic region needs to be determined. Additionally, CREB1 was an interesting candidate as an unbiased genome-wide location study was available [88]. Here the authors fused ChIP with SAGE, creating an alternative to ChIp-chip. Results obtained with MITOMI could therefore be compared to existing datasets.

To test whether CREB1 could be characterized with MITOMI, a small 64-member library was generated covering the bases TGACNNNA. Instead of spotting concentration gradients, this library was spotted at a uniform concentration. From this a relative affinity could easily be determined, which is a direct function of surface-bound

target DNA. The results are shown in Figure 8.17. The highest affinity sequence was GAC, indicating that the experiment was successful–meaning that not only bHLH transcription factors but bZIPs as well can be measured with MITOMI, covering two of the four largest families. Furthermore, the obtained landscapes differ considerably from those obtained for the bHLH transcription factors. Together this short experiment showed that MITOMI can be applied to other transcription factors and that landscapes can be obtained that differ from those observed previously.

### 8.6.2 Gli Transcription Factors

The hedgehog (Hh) signalling pathway controls many aspects of development, as well as maintains stem-cell populations in adults [90]. On a molecular basis, Hh signalling converges onto the Zinc finger transcription factors Gli1, Gli2, and Gli3 in vertebrates, and their homologue Cubitus interruptus (Ci) in the fly *D. melanogaster*. These Zn finger transcription factors in turn control the expression of decapentaplegic (dpp), patched (ptc1), engrailed (en), collier (col), and iroquois (iro). Ptc1 is a 12-transmembrane protein and serves as the receptor for Hh. Hh-binding inactivates Ptc1, which in turn activates smoothed (Smo), another membrane protein. This signal then is transduced to Ci. In flies Ci can serve both as a transcriptional activator, CiA, and as a repressor, denoted by CiR. Which function Ci fulfills is dependent on its cleavage state. Hh signalling causes the translocation of CiA into the nucleus where it associates with CREB-binding protein (CBP), which is also known to activate CREB (see previous Section 8.6.1), and together Ci and CBP activate target gene expression.

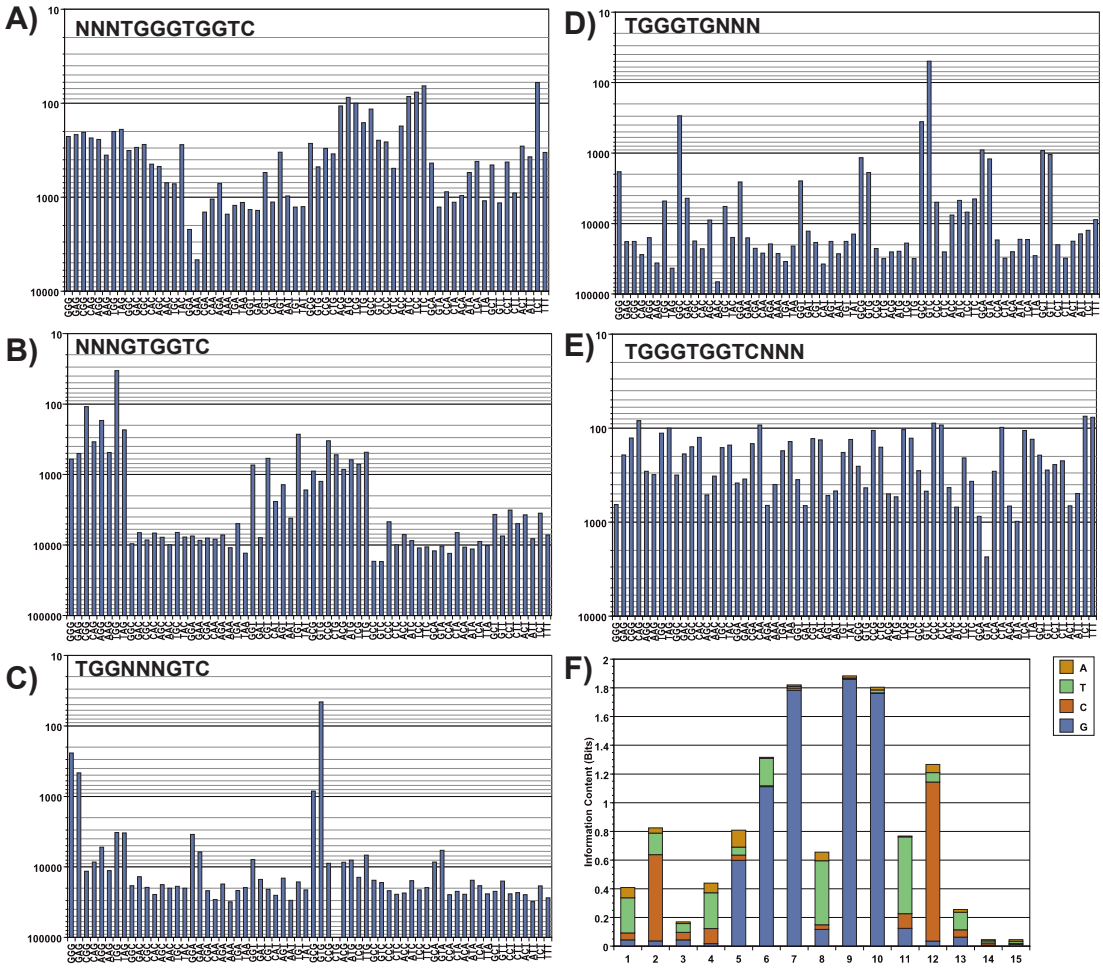



Figure 8.18: The Gli3 motif was tiled into 3-mer sections (A–E) and the measured affinities are displayed in nM. Panel F shows a Weblogo calculated from data displayed in Panels A–E.

In vertebrates the function of Ci has been split into three components: GLI1, GLI2, and GLI3. GLI1 functions as transcriptional activator and GLI3 as the repressor. GLI2 may function both as repressor and activator, but performs mainly the latter function.

Hallikas et al. semi-quantitatively measured the PWM for Gli1–3 and Ci using a pull-down-based competition assay [89]. They then used the established PWMs

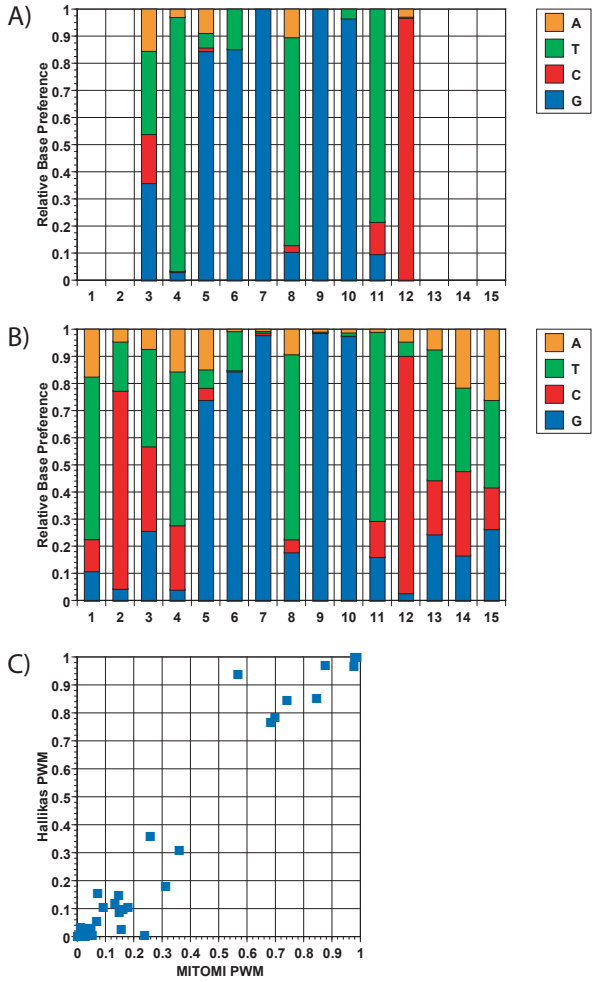



Figure 8.19: The PWMs reported by Hallikas et al. [89] and measured here in Panels A and B, respectively. All values of the PWMs were also cross-compared (Panel C).

to predict genomic binding of these and other transcription factors with a novel algorithm they term 'enhancer element locator' (EEL). EEL takes into account the relative affinity of a transcription factor to a target sequence, target site clustering, as well as conservation. Using EEL the authors were able to find known GLI targets such as Ptch1 and Gli1 out of a total of 42 elements that met selection criteria.

To understand the sequence recognition profile of the GLI transcription factors, they were tested against a library of target DNA sequences permuting 3-mer sections

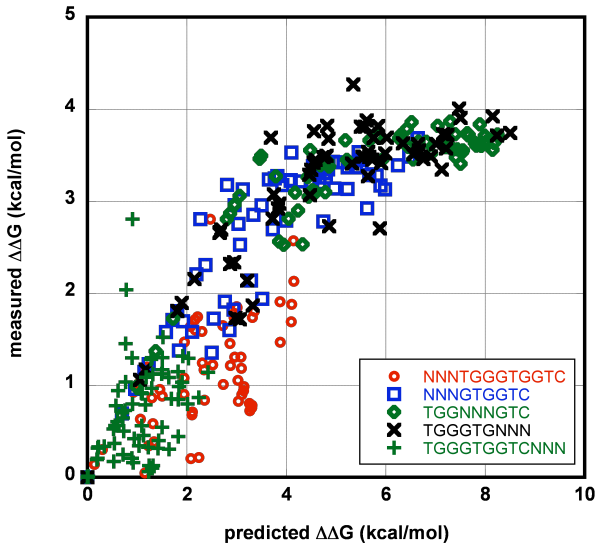



Figure 8.20: As with the bHLH transcription factors in Section 8.2.5 a PWM was used to predict the measured values. A picture very similar to Figure 8.11 can be seen.

tiled over the known GLI motif of TGGGTGGTC. These three core libraries consisting of NNNGTGGTC, TGGNNNGTC, and TGGGTGNNN were supplemented with libraries covering the flanking bases, including NNNCCCTGGGTGGTC, NNNTGGGTGGTC, and TGGGTGGTCNNN. Gli1–3 as well as Ci were obtained and cloned by Tyler Hillman (Scott Lab, Stanford CA). Instead of relying on *in situ* synthesis of the transcription factors, they were expressed bench-top in 25–50 $\mu$L rabbit ITT reactions. The ITT reaction was then loaded onto the device to flow deposit the transcription factor. This step produced good surface coverage, despite possible sub-optimal synthesis yields. The first target tested was GLI3 and the results for all but one library are shown in Figure 8.18. The data confirms the known consensus motif while increasing the depth of information which can be used in *in silico* prediction of target genes. To more closely compare the absolute affinities determined with

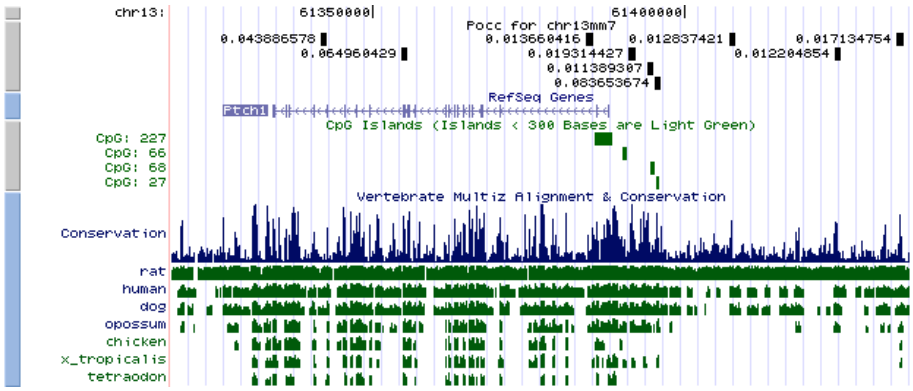



Figure 8.21: A snapshot of the USGS genome browser output for the mouse chromosome build mm7 centered on Ptch1, a known GLI target. A custom track was added showing the calculated high probability binding sites of Gli3 as black peaks in the first track of the figure. The $P_{occ}$ values are shown to the left of each peak.

MITOMI to the PWM values obtained by Hallikas, a PWM was calculated using the single substitution data obtained with MITOMI (Figure 8.18, Panel F). The weblogo was calculated following Schneider and Stephens [78] by calculating the difference between the maximum entropy and observed entropy at each base position:

$$R_{seq}(l) = log_2 B - (-\sum_{b=a}^{B} f(b,l) log_2 f(b,l)) \tag{8.11}$$

where $R_{seq}(l)$ is the information content at position l, b is any of the 4 possible bases (A,C,G, or T), B the total number of possible bases, and $f(b,l)$ is the frequency of base b at position l. The frequencies can be obtained from the experimental data by calculating a probability of observing a base b at position l from the measured affinity $\Delta\Delta G(b,l)$.

$$P(b,l) = \frac{1}{e^{\Delta\Delta G_{b,l}/RT} + 1} \tag{8.12}$$

$$f(b,l) = \frac{P(b,l)}{\sum_{b=a}^{B} P(b,l)} \tag{8.13}$$

The frequency $f(b,l)$ is then taken as the adjusted probability, as the sum of the probabilities at a position has to be 1. $R_{seq}(l)$ defines the total information content in each position l. The contribution of each base to the total information content is the product of the frequency of the base and $R_{seq}(l)$:

$$contribution(b,l) = f(b,l)R_{seq}(l) \tag{8.14}$$

.

As a control, the calculated PWM was compared to the PWM obtained by Hallikas et al. [89]. Hallikas et al. reported their PWM adjusted to a relative value of one (Figure 8.19, Panel A), therefore the PWM from Figure 8.18, Panel F was adjusted as well (Figure 8.19, Panel B). These two PWMs are essentially identical, more easily seen by a direct comparison of all PWM values (Figure 8.19, Panel C). MITOMI therefore successfully measured the relative single base preference of GLI3.

Not only were the single base substitutions measured but complete 3-mer libraries. To again understand whether individual base contacts were non-independent, as was the case with the bHLH transcription factors (Section 8.2.5), measured affinities were

again compared to values calculated from single base substitution data (Figure 8.20). Again, most of the predicted values don't agree with the experimentally determined affinities. Indeed the prediction accuracy was worse for Gli3 than for the bHLH transcription factors, with the flanking 3-mer libraries being particularly erroneous.

Similar to the approach taken by Hallikas et al., the binding energy landscapes determined by MITOMI were used to predict likely genomic target sites [89]. GLI3 measured here was obtained from the mouse, therefore all calculations were performed on the mouse genome builds mm7 and mm8. The *in silico* approach taken here is essentially the same as used for determining $P_{occ}$s in yeast (Section 8.2.4). But instead of calculating $P_{occ}$s for a certain region upstream of every ORF, $P_{occ}$ windows of 500 or 1000 bps in length covering the entire genome were calculated. To assure that no binding sites were missed, the windows had a 15 bp overlap. The resulting $P_{occ}$ windows could then also be centered onto the highest individual 15 bp transcription factor binding site. Figure 8.21 shows calculated high-affinity binding sites near Ptch1 on chromosome 13, a known GLI target. Interestingly, CpG island locations seem to coincide with all three predicted GLI binding sites, just upstream of Ptch1.

# Chapter 9

# Proteasome

The proteasome is a large multi-unit complex involved in the non-lysosomal ATP-dependent proteolysis of proteins in eukaryotic cells. Substrate proteins for the proteasome are abundant and derived from all types of cellular processes. Proteins are targeted to the proteasome via introduction of polyubiquitin chains on lysine residues by the enzymes E1, E2, and E3.

Structurally, the proteasome is a 26s complex (2000 kDa) consisting of 2 major subcomplexes–the 20s proteolytic core complex and the 19s regulatory particle. The proteolytic core consists of a homodimer of two $7\alpha7\beta$ symmetric subcomplexes arranged in a barrel shape (Figure 9.1). The 19s subcomplex, on the other hand, is less structured and various subunits have been associated with it. There are 6 non-ATPase and about 11 ATPase dependent subunits in the 19s regulatory particle (Figure 9.1).

A high-resolution crystal structure is available for the yeast 20s particle [93]. Neither the 19s regulatory particle nor the full 26s complex could be solved with x-ray diffraction as of yet, likely due to the fact that growing crystals of the unstructured 19s particle is difficult. Walz et al. did report a low-resolution electron microscopy

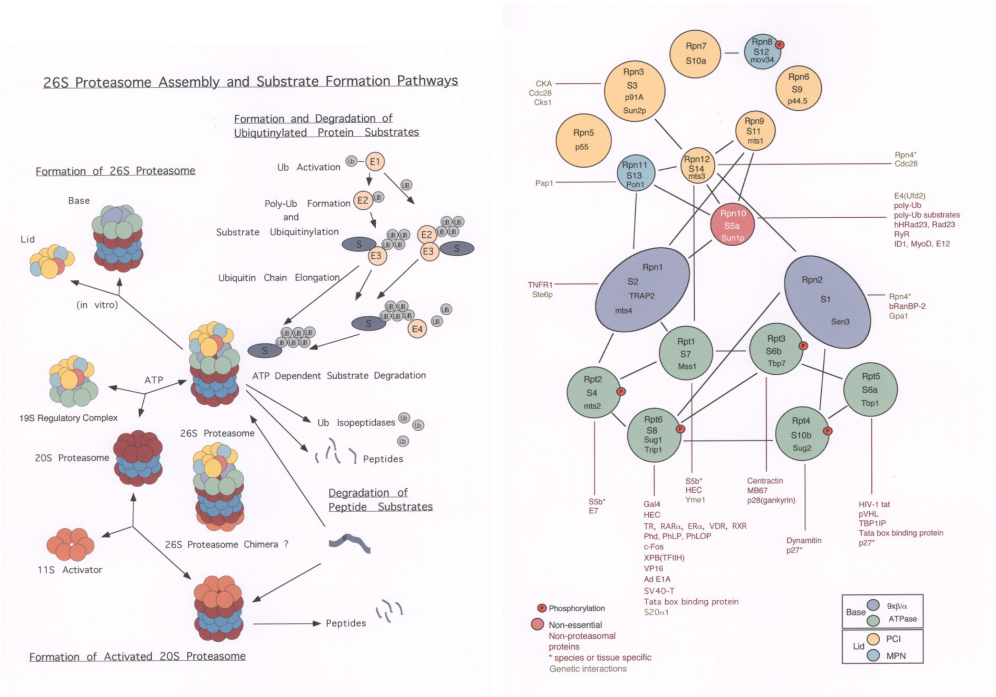



Figure 9.1: The left diagram depicts the overall structure of and function of the 26s proteasome holoenzyme. The right diagram shows the specific proposed interactions taking place in the 19s regulatory particle. (Taken from [91])

structure of the entire 26s proteasome complex [92] (Figure 9.2).

The exact location of the individual 19s subunits is unknown, due to the low resolution of the EM structure. It might be possible to build a map of subunits by determining the interconnectedness of the subunits and fitting the established network into the EM structure. Knowing the position of each subunit would give insight into the possible function of the individual subunits, as well as of the particle at large. Establishing the 19s subunit network requires testing of all possible binary interactions of the subunits, of which there are 18 (Rpt1–6 and Rpn1–12). Linear expression templates were designed for the on-chip expression of the subunits. Several linear expression templates were generated for each subunit, including versions carrying N

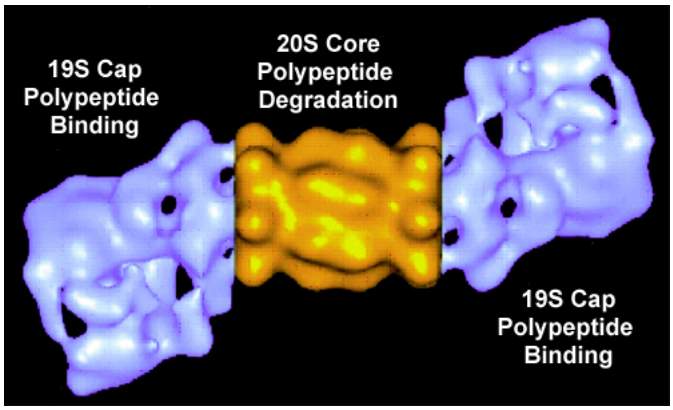



Figure 9.2: A EM structure of the 26s proteasome obtained by cryo-EM (Taken from [92])

and C-terminal 6xHis, T7, and S-tags, as well as versions carrying no epitope tag. All synthesized linear expression templates are listed in Table 9.1. Genomic DNA from yeast *S.cerevisiae* strain S288C (Invitrogen) served as template for all PCR reactions. Bait and prey experiments can be set up with any combination of the various tagged and un-tagged subunits. Generally the 6xHis tagged versions are used for surface localizing the bait. The prey can be detected using the T7 or S-tag (Section 7.3) using fluorescent or enzymatically labeled anti-T7 antibodies or S-protein, respectively.

His tagged versions of a few subunits were expressed on-chip from spotted linear expression templates, using residue-specific incorporation of a $tRNA_{Lys-bodipy-fl}$ for detection. The resulting intensities, after localizing the synthesized proteins to the detection area, are shown in Figure 9.3. Yeast proteins generally express well in wheat-germ-based ITT systems, including the proteasome subunits studied here. It should be noted that most subunit sizes fall within 30–50 kDa, with the exception of Rpn1–2 which have weights above 100 kDa and thus lower expression yields. Even though the subunits were easily expressed on-chip, only limited time and effort was

| | | No Tag | N-His | C-His | N-T7 | C-T7 | N-S-tag | C-S-tag |
|---|---|---|---|---|---|---|---|---|
| 19s | Rpt1 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | |
| | Rpt2 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | |
| | Rpt3 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | |
| | Rpt4 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | |
| | Rpt5 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | |
| | Rpt6 | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | |
| | Rpn1 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn2 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn3 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn4 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn5 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn6 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn7 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn8 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn9 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn10 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn11 | ✔ | ✔ | ✔ | | ✔ | | |
| | Rpn12 | ✔ | ✔ | ✔ | | ✔ | | |
| 20s | alpha-1 | ✔ | ✔ | ✔ | | | | |
| | alpha-2 | ✔ | ✔ | ✔ | | | | |
| | alpha-3 | ✔ | ✔ | ✔ | | | | |
| | alpha-4 | ✔ | ✔ | ✔ | | | | |
| | alpha-5 | ✔ | ✔ | ✔ | | | | |
| | alpha-6 | ✔ | ✔ | ✔ | | | | |
| | alpha-7 | ✔ | ✔ | ✔ | | | | |

Table 9.1: Inventory of linear expression templates coding for the subunits of the proteasome and their respective epitope tags

expended on determining binary interactions. Initially the interactions were tested using standard protein array methodology without the use of MITOMI. Not only was the detection mechanism flawed, it also became apparent that even though the entire complex, as well as the base and lid portion of the 19s regulatory particle, were stable in solution, this did not necessarily extend to individual binary interactions. Using MITOMI in combination with an optimized detection method should allow for the determination of the correct binary interactions. A second possible solution would have been to attempt to measure ternary interactions instead, which might result in more stable complexes if the correct 3 subunits were investigated. Of course

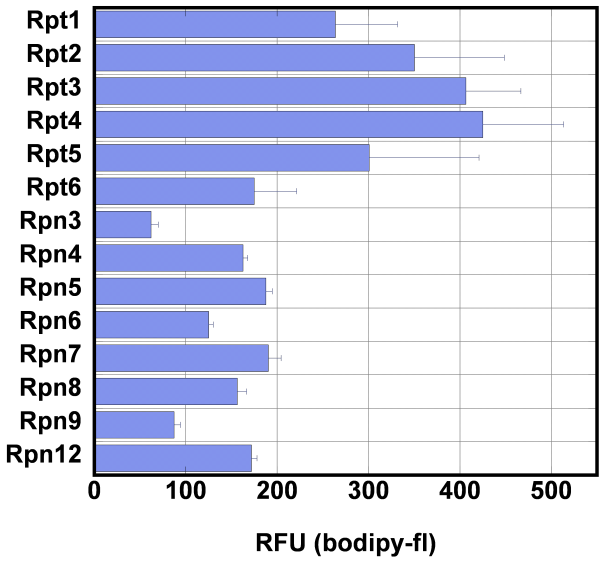



Figure 9.3: Examples of expression efficiencies of proteasomal subunits on-chip

testing ternary interactions would require measuring a total of 5832 possible subunit combinations instead of 324 binary combinations.

Another intriguing possibility presented itself in the fact that with MITOMI it may have been possible to measure not only relative affinities, but absolute affinities, as well as dynamics. Section 7.4 describes in detail the approaches that may be taken to measure the dynamics of molecular interactions, including on- and off-rate measurements of two body systems. Knowledge of the dynamics of individual binary interactions should allow one to simulate complex assembly *in silico*, which in turn should give insight into how complexes assemble and function. The 19s regulatory particle is particularly interesting, as it is a rather complex, asymmetric structure with large possible structural deformations taking place.

# Chapter 10

# Cell Arrays

## 10.1 Introduction

Proteomic and Systems Biology efforts have resulted in the creation of large libraries of ORF clones, as well as knockout strains. With these libraries it is possible to dissect large biological networks by measuring or perturbing one element at a time. These libraries generally cover the entire proteome and thus have sizes ranging in the thousands for yeast ORF clonal libraries. A variety of libraries are available for yeast, including knockout strains [94], genomic TAP fusion libraries [69], plasmid-based ORF libraries [95], and genomic GFP fusions [96]. One problem with these libraries is the fact that they are difficult to interrogate due to their sheer size. Efforts to date include brute force approaches based on classical bench-top techniques [69, 96, 95], as well as more technologically advanced experiments based on FACS sorting [97].

Yet these libraries can be readily used in conjunction with microfluidics and spotted micro arrays for programming. Libraries harbored in resilient hosts, such as yeast and bacteria, may be spotted just as any other solution, followed by alignment to a device. The resulting microfluidic device may house the entire library with thousands

of individual elements, and complex fluidic manipulations may be performed, such as fluid exchange and generation of molecular gradients. Furthermore, the cells are still viable and may be grown *in situ*. Use of a microscope then allows single cells to be followed with high time resolution, so that the response of protein concentrations and localization in the cell as a function of some perturbation may be characterized. This provides unprecedented control over the most important factors, such as the ability to control the introduction of the perturbing agent, as well as a high temporal and spatial resolution down to minutes and single cells. It is also possible to lyse the cells rather than attempt to grow them. When the cells are lysed the fusion protein expressed intracellularly may be purified in an adjacent chamber, providing one of the most rapid and facile methods for generating large-scale protein arrays.

## 10.2 Live Yeast Cell Arrays

As mentioned in the previous chapter, proteome-wide ORF and knock-out libraries are available for many model organisms, such as *E.coli* and yeast. Studying these libraries on a proteomic scale has been challenging though. Micro-arraying these clonal libraries, followed by *in situ* growth and investigation on microfluidic devices, provides a unique opportunity to rapidly screen these libraries. The first instantiation of this approach was tested with the yeast genomic GFP fusion library generated by Huh et al. [96]. The first 96 clones of the library (Invitrogen) were grown in 96-well plates using YPD as the growth medium. After overnight growth at 30°C and agitated at 300–400 rpm the cells are allowed to settle. Optionally the cells may

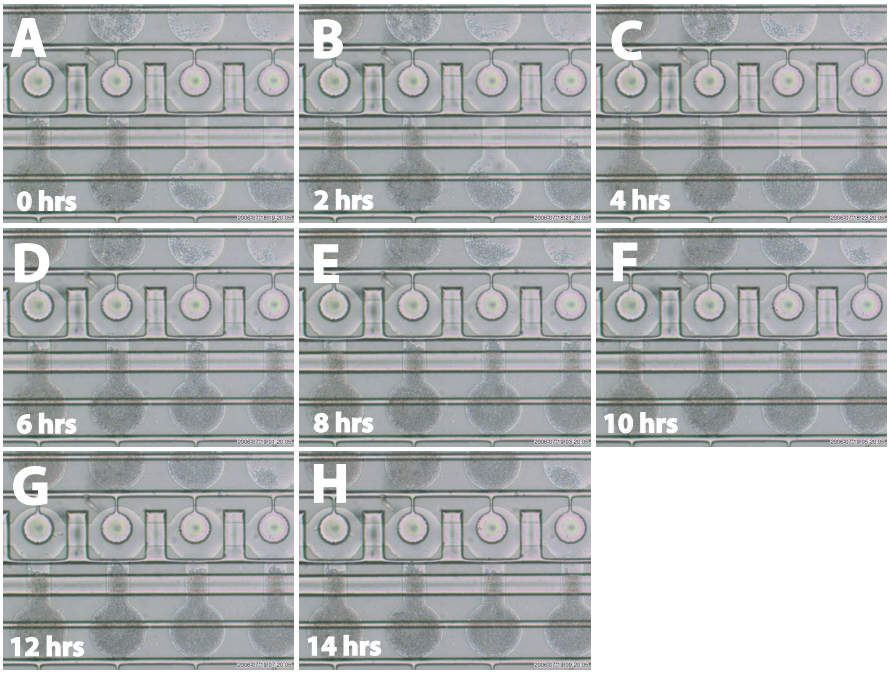



Figure 10.1: Time series of a yeast life cell array. Device chambers were programmed with 96 different yeast–GFP fusion strains from a spotted micro-array. Images were taken every 2 hours, while the growth chambers were continuously perfused with media via the horizontal flow channel. Cell density can be seen increasing in several growth chambers during the 14-hour period.

also be pelleted by centrifugation. The pelleted cells are then micro-arrayed using a standard dip pen method. Here use of v-shaped 96-well plates is advantageous, due to better pellet formation. To ascertain that there is no cellular carryover between spots, a test array was spotted in which every other spot was a blank. This test array showed no cross-contamination between spots due to spotting. A cell array is then aligned to a microfluidic device and bonded to the epoxy substrate at 40°C for 30 to 120 minutes. The duration of bonding is intentionally kept short to keep cell viability as high as possible. Once the device is bonded it can be operated and cells infused with media to re-start growth of the clones. First the entire device is dead-end filled with media followed by continuous flow of medium through the parallel flow channels. The cells

are then grown either in YPD or SC-His, with the chip placed on a hot plate set at 30°C. SC-His was chosen to insure the integrity of the GFP fusion for the duration of the experiment. SC-His does show retarded growth performance over YPD, a problem that may need to be addressed in future experiments. The cells, regardless of medium used, show a day or more of lag before they start growing. This long lag is likely due to the rather stressful spotting procedure and possibly due to the intrinsically low cell numbers in each chamber. The time it takes for the cells to completely suffuse a unit cell is also, and not surprisingly, strongly dependent on the seeding density, as well as the clonal construct (Figure 10.2.) Once the cells are growing they can be kept in an exponential growth phase by continuous removal of overflowing cells at the shear interface. It was also shown that no contamination occurs between unit cells despite use of a passive device where all unit cells are interconnected at all times. The above-mentioned control array was used, in which every other unit cell was lacking a seed spot. Even after a day or more of growth, the negative unit cells remained free of cells. Care must be taken to avoid clogging of cells at any point in the device, as this stops the flow in one of the parallel channels, at which point the entire channel will become overgrown with cells.

Once cells are growing steadily, experimental conditions may be changed and the response observed. Observations may include optical interrogation of growth rate, protein abundance, and location, as well as other accessible parameters. In order to obtain optimal interrogation conditions changes to the current chip design are required. Mainly the flow channel height should be dropped to about 2–4 $\mu$m to

restrict the growing yeast cells and have them grow in a monolayer, allowing single-cell studies to be performed with greater ease. Studying cellular dynamics is one primary application for this method, which requires relatively high time resolution on the order of minutes (compared to cell division times of 20–40 minutes.) The measurement rate depends both on the camera sensitivity, as well as stage velocity; with current technologies a reasonable rate is about 4 Hz. On the current 2400-chamber device a time resolution of 600 seconds or 10 minutes can thus be achieved. A time resolution of 10 minutes is reasonable and will provide good response curves. Higher resolutions of about 1 minute can be achieved by interrogating only part of an array at any given time. A total of 240 chambers, for example, may be interrogated with a time resolution of one minute.

Experimental conditions can easily be changed by addition of one or more components to the culture medium to affect certain aspects of the cellular environment. Likewise, concentration gradients of these components can also be easily set up [98]. Other interesting but technically more challenging manipulations include genetic engineering or the use of RNAi to introduce new genes or specifically perturb genes already present. Finally, and mentioned in greater detail in the next section, the growing cells may be lysed and their contents purified *in situ* to be analyzed on a molecular level.

Specifically, one experiment that can be run on this platform would be an investigation of the dynamic response of protein levels and location on a single-cell level as a function of a DNA modifying drug such as the methylating agent methylmethane

sulfonate (MMS), ionizing radiation [99], or histone depletion [100]. But essentially any method thus far applied to obtain differential gene expression experiments is applicable here.

## 10.3 Yeast Protein Arrays

It was shown in the previous chapter that spotted yeast cells are viable and can be cultured *in situ* on a microfluidic device. This same approach can also be used for generating protein arrays directly from spotted yeast clones by lysis of the cells and consequent purification of the cellular contents. The process for generating a protein array is quite simple. Yeast cells such as the yeast–GFP clones are grown off-chip using 96- or 384-well plates. If other cell types are used carrying ORFs under the control of inducible promoters for example, protein expression can be induced just prior to spotting. The cells are spotted directly from the plate from which they were grown or may be washed with buffer to reduce contaminations contained in the media. Once the cells are spotted, a device is aligned to the array and bonded for a minimum of 2 hours at 40°C. As mentioned in the previous chapter, the cells may then be suspended in media to allow cultivation. Growing each clone on-chip has the advantage that cells multiply and thus will produce more protein to be purified. A second advantage of on-chip growth of the culture before purification is that the cells may be induced *in situ* rather than off-chip, also potentially resulting in better yields. A final advantage is that the cells are shown to be viable and thus concerns about the state of the protein to be purified is minimal.

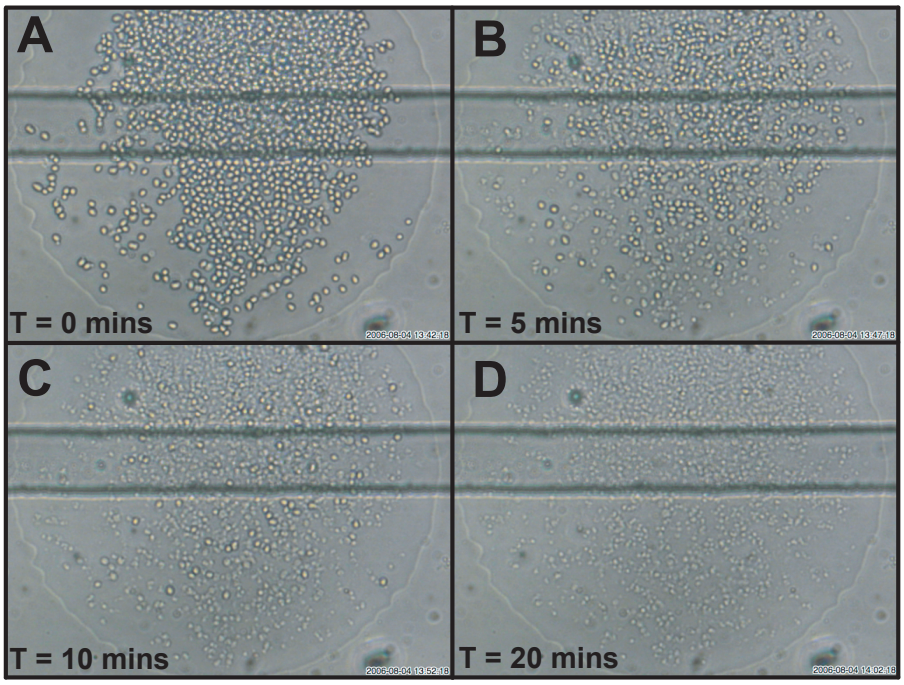



Figure 10.2: Time series of yeast spheroplast formation. Zymolyase was introduced into the chamber by passive diffusion, causing the disintegration of the yeast cell wall over a period of 10 to 20 minutes. The resulting spheroplasts could then be lysed using standard detergent-based methods.

Thus far, growth and purification were not combined and all data taken on protein purification is based on lysing yeast–GFP cells immediately, without additional growth periods. One reason for this is that extensive growth may foul the surface chemistry. The button may be used to prevent surface fouling by physically protecting the surface during culturing. Figure 10.3 shows yeast–GFP cells contained in a unit cell and suspended in a solution of 10 mg/mL Zymolyase, 1% BME, in 0.1 M Tris HCl. This first step breaks down the yeast cell wall, generating yeast spheroblasts which can then easily be lysed using a detergent solution such as Y-PER (Pierce). Similarly these two steps can be combined by dissolving 10 mg/ml Zymolyase, 1% BME in Y-PER directly.

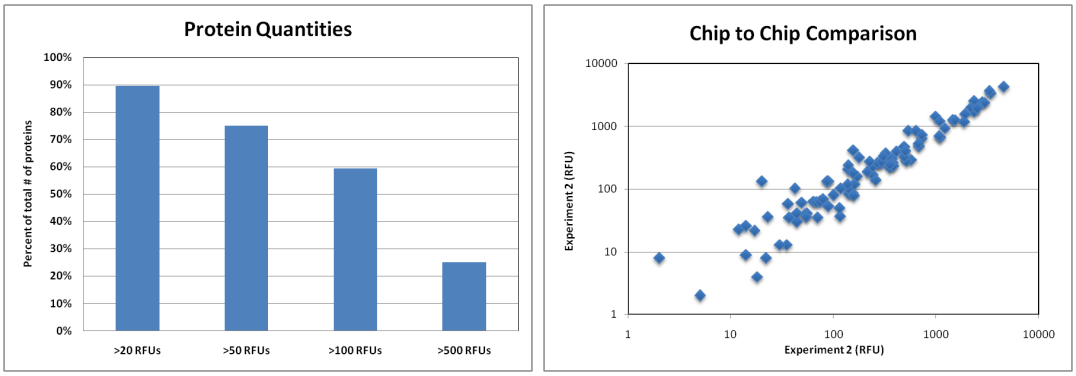



Figure 10.3: The left graph shows the pull-down efficiency of the first 96 yeast–GFP fusion proteins extracted on-chip from yeast cells. The percentage of clones giving a minimal RFU value of 20, 50, 100, and 500 is plotted, with about 75% of the clones giving a signal of 50 or above. The graph on the right shows the reproducibility of the pull-down between on-chip experiments, hinting that the method could be used to quantitatively measure intracellular protein concentrations.

The cellular contents, including the GFP-tagged protein specific to the clone, are now free to diffuse. Surface deposited anti-GFP antibodies capture and localize the GFP-tagged protein, generating a homogeneous spot for interrogation. This of course takes place in every unit cell on the device, generating a unique protein spot for each yeast–GFP clone deposited.

Once capture is complete, all non-specific cellular material is washed away and the resulting GFP fluorescence on each spot may be measured. This fluorescent intensity is directly proportional to the protein concentration of the tagged protein. There is a large amount of intrinsic variability in protein expression of these GFP-tagged proteins, since they have been genomically tagged with GFP and thus are still under the control of their wild-type promoters. It was determined that over 90% of all proteins tested expressed amounts corresponding to at least 20 RFUs, and that about 75% of all clones showed at least 50 RFUs (Figure 10.3, Panel A). The latter concentration

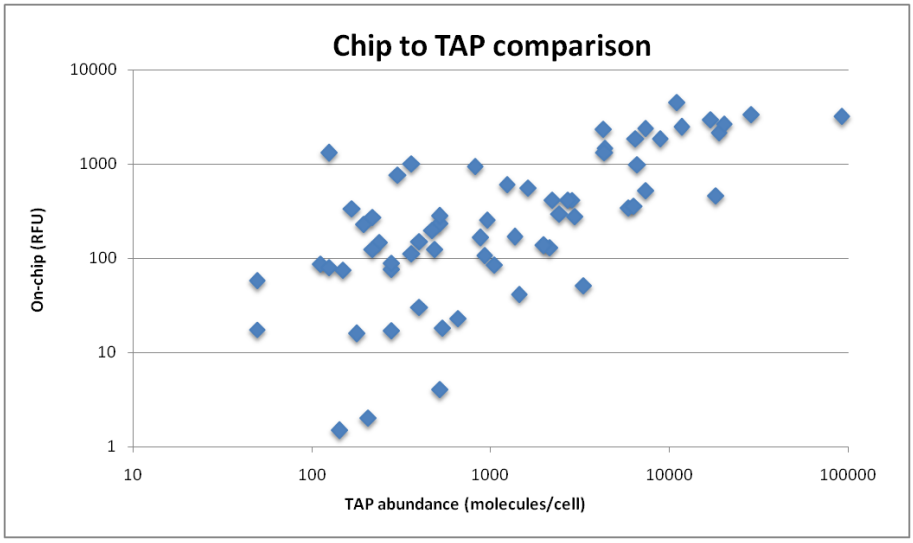



Figure 10.4: Yeast on-chip protein values were compared to values determined by Ghaemmaghami et al. [69] by Western blotting TAP-tagged fusions. A reasonable correlation can be observed, despite the fact that the on-chip data was not adjusted to the number of cells giving rise to the signal.

should be sufficient to perform basic protein–protein pull down assays using these protein arrays. Pull-down is also considerably robust across on-chip experiments as shown in Figure 10.3, Panel B and a dynamic range in protein concentrations of almost five orders is observed. Furthermore the amounts of pulled-down protein from these cultures correlated well with the measured *in vitro* concentration using Western blots on TAP-tagged clones. This is encouraging, as the protein amounts detected on-chip were not normalized in any form. A sensible normalization, such as adjusting the detected on-chip protein intensity by the number of cells giving rise to it, should further increase the correlation between the on-chip determined values and the bench-top measurements. It thus seems possible to perform accurate determinations of *in vivo* protein content on a very small number of cells in a high-throughput format.

# Appendix A

# cDNA clone library

| Common Name | NCBI Accession | Clone ID | Item | Vector | Resistance | Host |
|---|---|---|---|---|---|---|
| MAX | BC036092 | 5299106 | EHS1001-18380 | pBluesscriptR | Amp | DH10B |
| TFE3 | BC026027 | 4576858 | EHS1001-5058351 | pOTB7 | CAM | DH10B (phage-resistant) |
| TFE3 | BC026027 | 4576858 | EHS1001-44637 | pOTB7 | CAM | DH10B(phage-resistant) |
| TFEB | BC032448 | 5180066 | EHS1001-24335 | pCMV-SPORT6 | Amp | DH10B |
| TFEC | BC029891 | 5179088 | EHS1001-19568 | pCMV-SPORT6 | Amp | DH10B |
| C-Myc | BC000917 | 3048750 | EHS1001-953 | pCMV-SPORT6 | Amp | DH10B |
| Mxi 1 | BC012907 | 3882557 | EHS1001-4364050 | pCMV-SPORT6 | Amp | DH10B(phage-resistant) |
| ID1 | BC000613 | 3346009 | EHS1001-3884718 | pOTB7 | CAM | DH10B(phage-resistant) |
| ID2 | BC030639 | 4820416 | EHS1001-12292 | pBluescriptR | Amp | DH10B |
| ID3 | BC003107 | 3543936 | EHS1001-4082645 | pOTB7 | CAM | DH10B(phage-resistant) |
| ID4 | BC014941 | 4552357 | EHS1001-5033850 | pOTB7 | CAM | DH10B(phage-resistant) |
| Myogenin (Myf4) | BC053899 | 6170028 | EHS1001-6519041 | pCMV-SPORT6 | Amp | DH10B(phage-resistant) |
| MyoD (Myf3) | BF219762 | 2961494 | MHS1011-58728 | pOTB7 | CAM | DH10B(phage-resistant) |
| TWIST H1 | BC036704 | 4125830 | EHS1001-35100 | pOTB7 | CAM | DH10B(phage-resistant) |
| HEB | BC050556 | 5767579 | EHS1001-6128112 | pCMV-SPORT6 | Amp | DH10B |
| Myf6 | BC017834 | 4288735 | EHS1001-42241 | pDNR-LIB | Cam | DH10B(T1 phage-resistant) |
| MyoD (Myf3) | BC064493 | 5022419 | EHS1001-5503912 | pOTB7 | CAM | DH10B(phage-resistant) |
| Myf5 | CB856774 | 5793748 | EHS1001-6154281 | pAMP1 | Amp | DH10B |
| MAX isoform A | BC004516 | 3937573 | EHS1001-4419066 | pOTB7 | CAM | DH10B(phage-resistant) |
| MAX isoform B | BC003525 | 3607261 | EHS1001-4145970 | pOTB7 | CAM | DH10B(phage-resistant) |
| Mxi 1 isoform B | BC035128 | 5263647 | EHS1001-27831 | pBluescriptR | Amp | DH10B |
| E12 | BC011665 | 4110737 | EHS1001-38433 | pOTB7 | CAM | DH10B(phage-resistant) |
| E47 | AA876062 | 1338894 | EHS1001-1971203 | pT7T3D-PacI | Amp | DH10B |
| E47 | BE514178 | 3635031 | EHS1001-4173740 | pOTB7 | CAM | DH10B(phage-resistant) |
| MAD 3 | BC000745 | 2821596 | MHS1011-58805 | pOTB7 | CAM | DH10B(phage-resistant) |
| MAD | BC069377 | 7262252 | MHS1768-9143888 | pPCR-Script Amp SK(+) | Amp | XL10 Gold |
| MAD 4 | BC068060 | 6538192 | EHS1001-8947018 | pOTB7 | CAM | DH10B(phage-resistant) |
| | | | | | | |
| EBF | BC041178 | 6045572 | MHS1010-9204744 | pCMV-SPORT6 | Amp | DH10B (phage-resistant) |
| EBF3 | BC011557 | 4561018 | MHS1011-75946 | pOTB7 | CAM | DH10B (phage-resistant) |

Table A.1: cDNA clone inventory. All clones were obtained from OpenBiosystems.

# Appendix B

# MPEP primer library

| | | |
|---|---|---|
| Eukaroyotic | 5'ext | gatcttaaggctagagtac TAATACGACT CACTA TAGGGAAT ACAAG CTACT TGTTC TTTTT GCActcgagaattcgccacc |
| | 3'ext1 | GTAGCAGCCTGAGTCGACTCTAGATTATTAATGATGA TGATGATGATGGCCGCTGCTGCCTTGGAAGTAGAGG TTCTCGGCGGCGGTCTTGAGGCT |
| | 3'ext2 | CAAAAAACCCCTCAAGACCCGTTTAGAGGCCCCAAG GGGTTATGCTAGTTTTTTTTTTTTTTTTTTTTTTTTTTT TTTGTAGCAGCCTGAGTCG |
| | 5'final | gatcttaaggctagagtac |
| | 3'final | CAAAAAACCCCTCAAGAC |
| Prokaryotic | 5'ext_kim | gatcttaaggctagagtacATTAAT ACGACTCACT ATA G GGAGAC CACAACGGTT TCCCTCTAGA GATCATTTTGTTTAACTTTA AGA AGGAGA T ATAGAT |
| | 5'ext_roche | gatcttaaggctagagtacT AATACGACTC ACTATAGGGA GACCACAACG GTTTCCCTCT AGAAATAATT TTGTTTAACT TTAAGAAGGA GATATACC |
| MAX basic region switch | 5'ext | gatcttaaggctagagtac TAATACGACT CACTA TAGGGAAT ACAAG CTACT TGTTC TTTTT GCActcgagaattcgccacc atgagcgataacgatgacatcgaggtggagagcgacgct |

Table B.1: Primer inventory for various 2 step PCR approaches

# Appendix C

# Chip design gallery

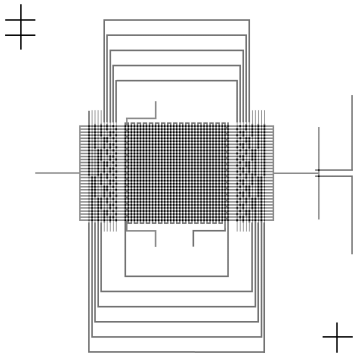


Figure C.1: Serpentine Enrichment Chip

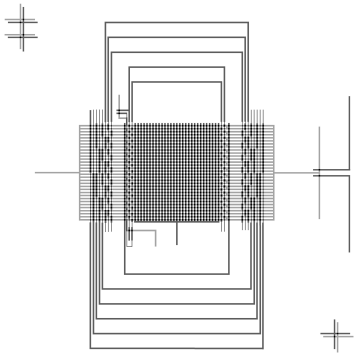


Figure C.2: Serpentine Enrichment Chip G4

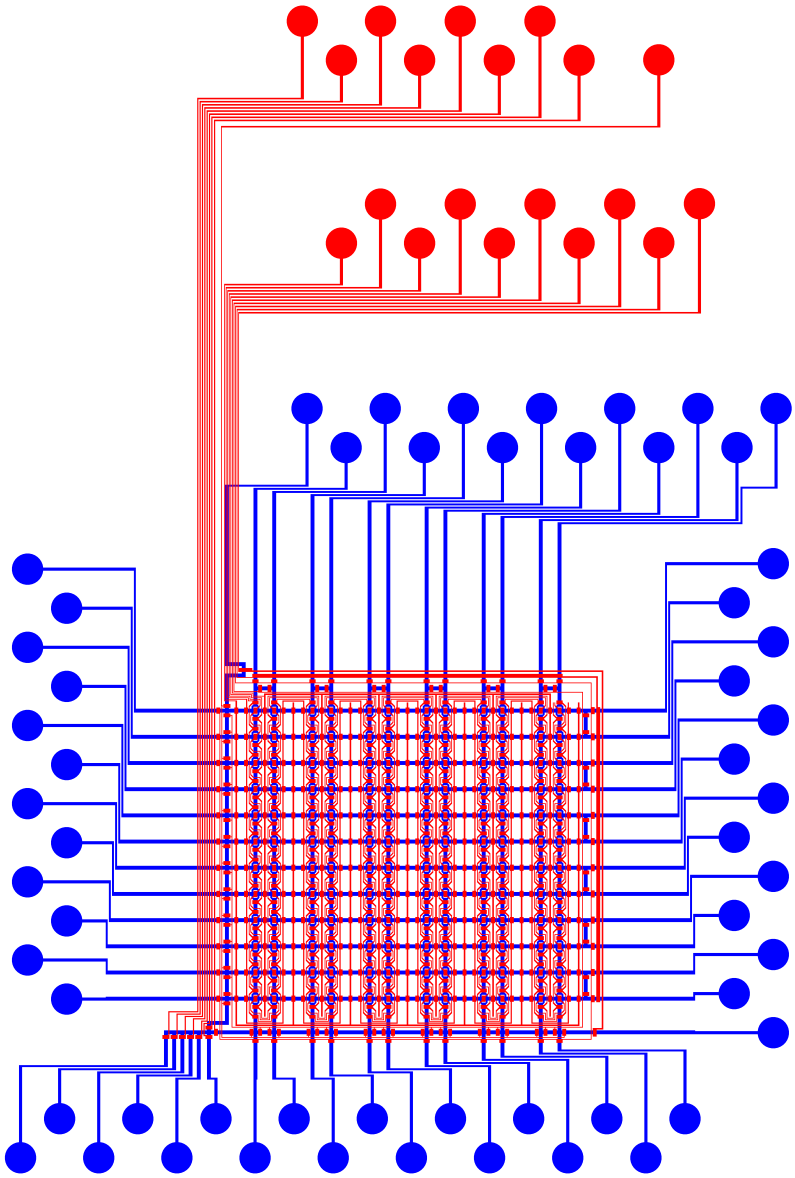


Figure C.3: Binary Interaction Chip v2
Unit Cells: 124
Area: 2.5 x 3.5 $cm^2$
Number of Valves: 826
Valve density: 94 valves/$cm^2$
Notes: first device with a membrane to be used for MITOMI

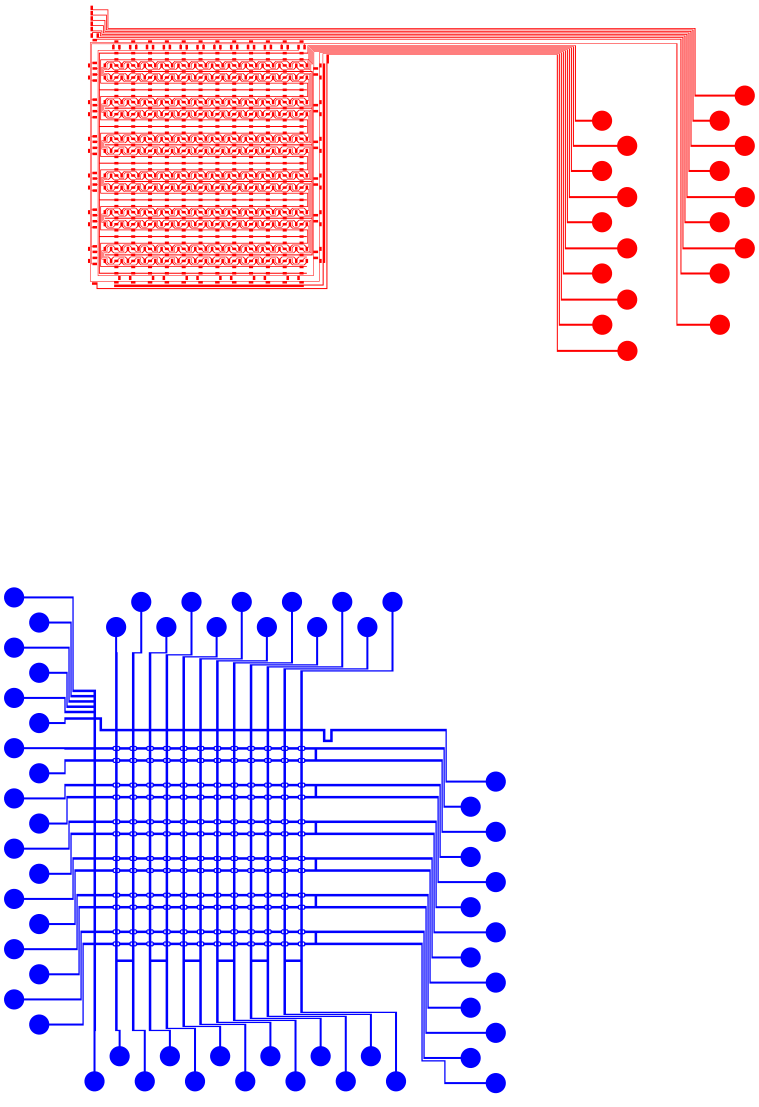


Figure C.4: Binary Interaction Chip v2

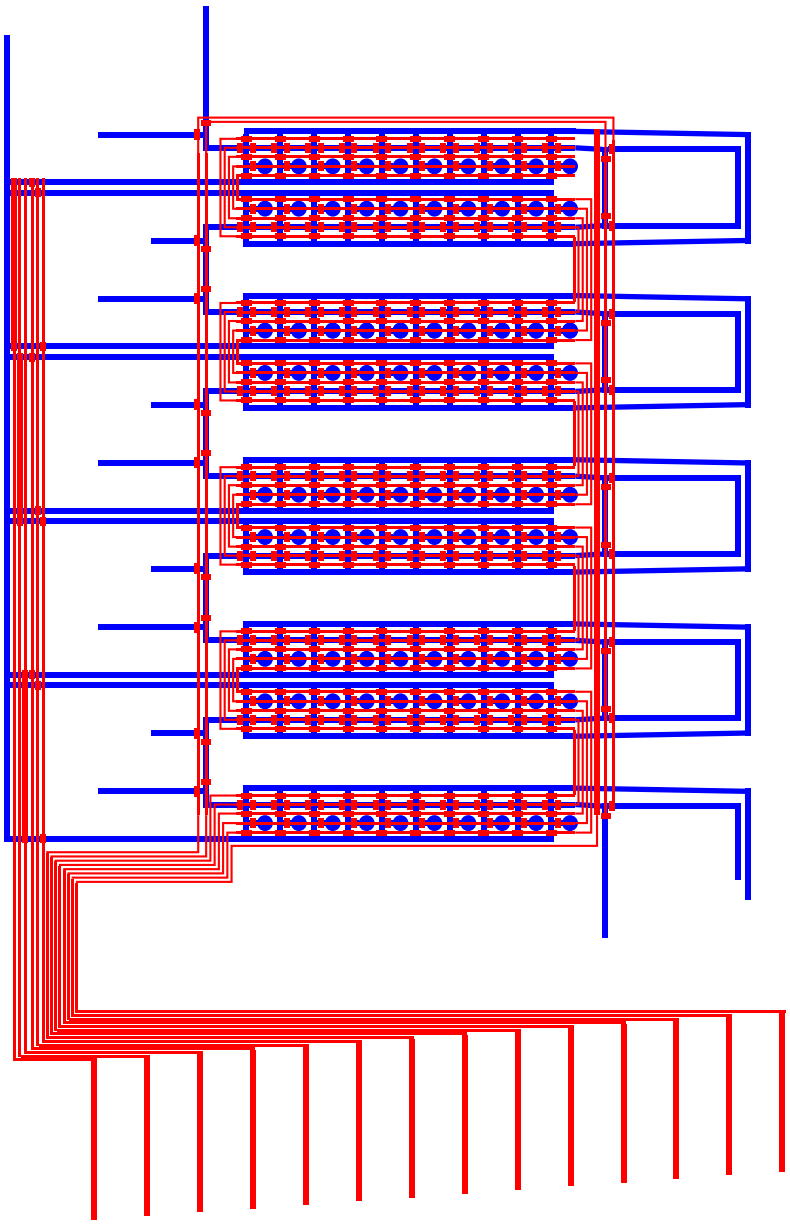


Figure C.5: DNA to Protein Array optimized
Unit Cells: 90
Area: 2.5 x 2.5 $cm^2$
Number of Valves: 612
Valve density: 98 valves/$cm^2$
Notes: spotting and surface chemistry test device

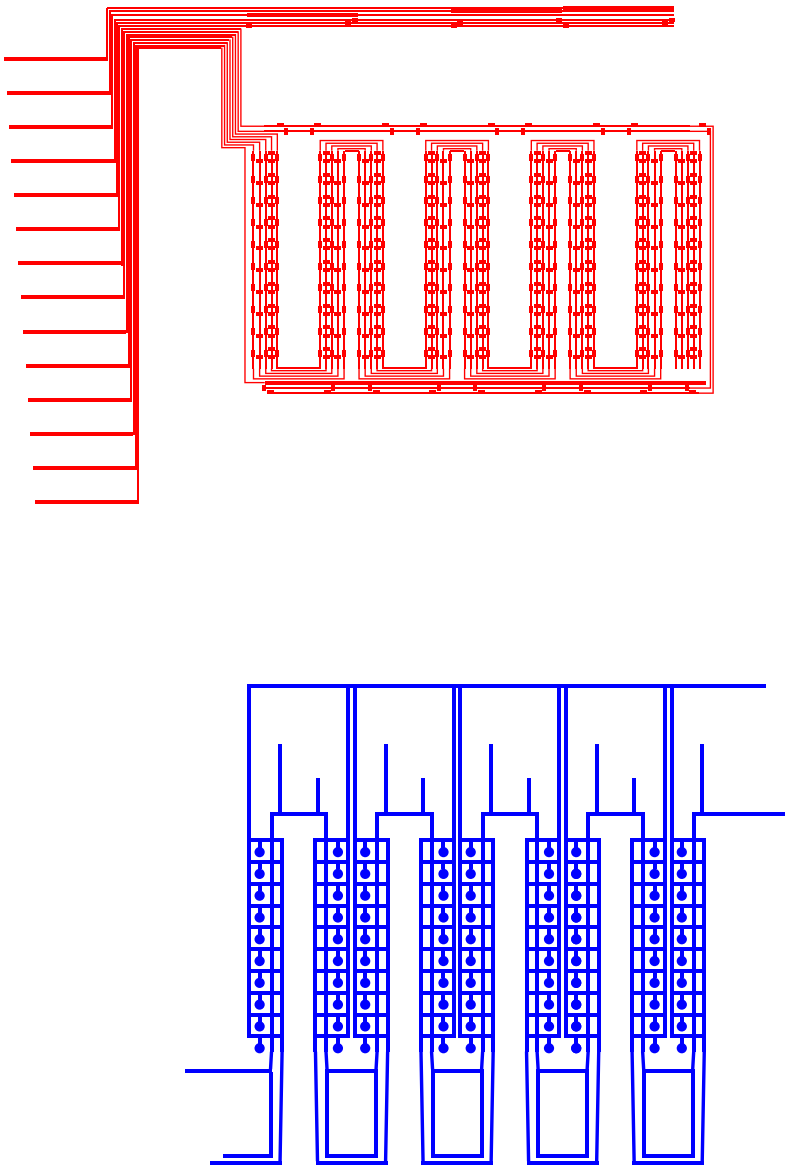


Figure C.6: DNA to Protein Array optimized

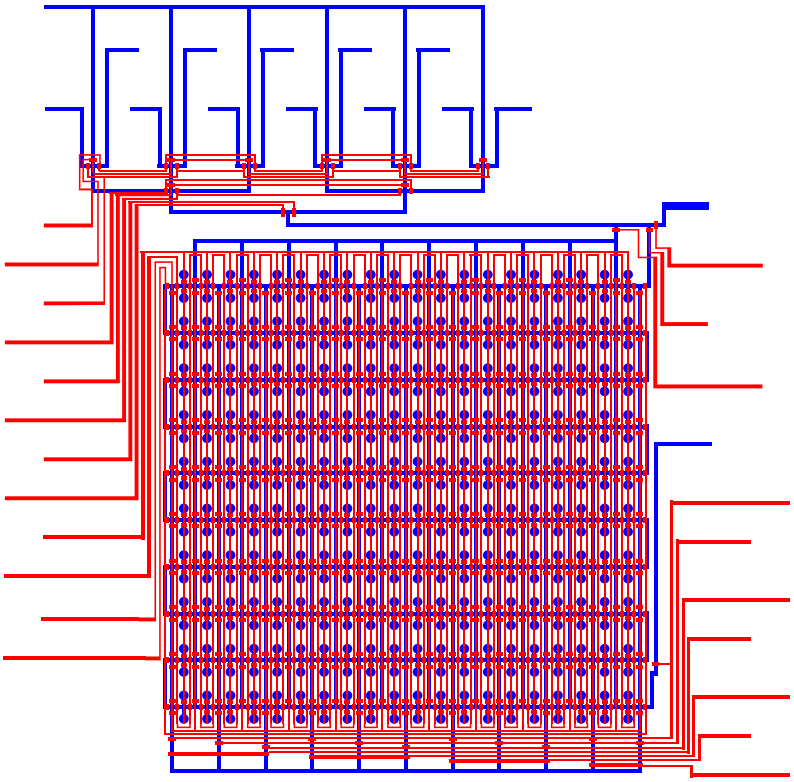


Figure C.7: DNA to Protein Array x2
Unit Cells: 200
Area: 2.5 x 2.5 $cm^2$
Number of Valves: 1832
Valve density: 293 valves/$cm^2$
Notes: device for testing binary protein interactions by coupling two chambers to one detection area

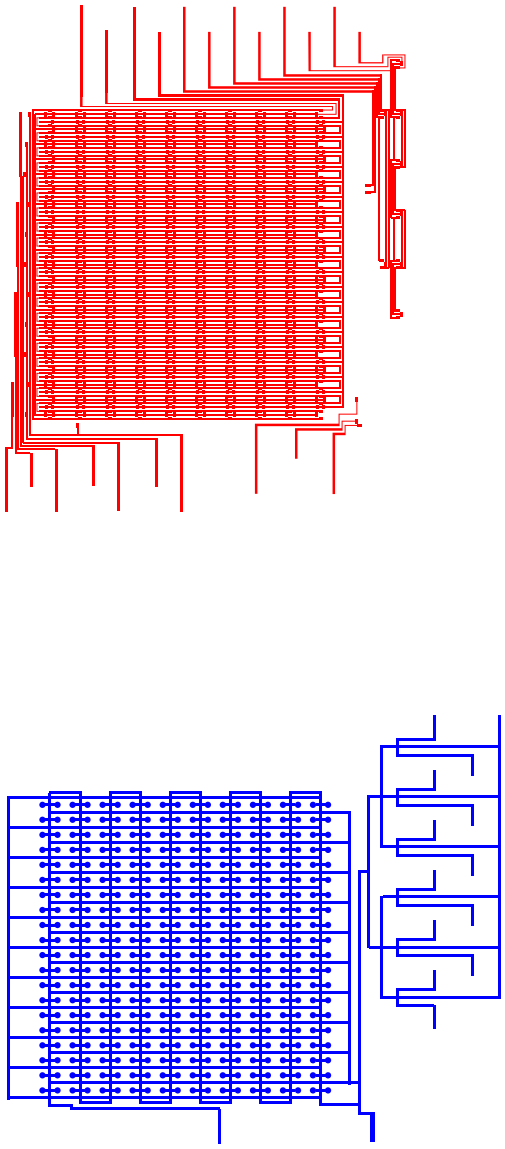


Figure C.8: DNA to Protein Array x2

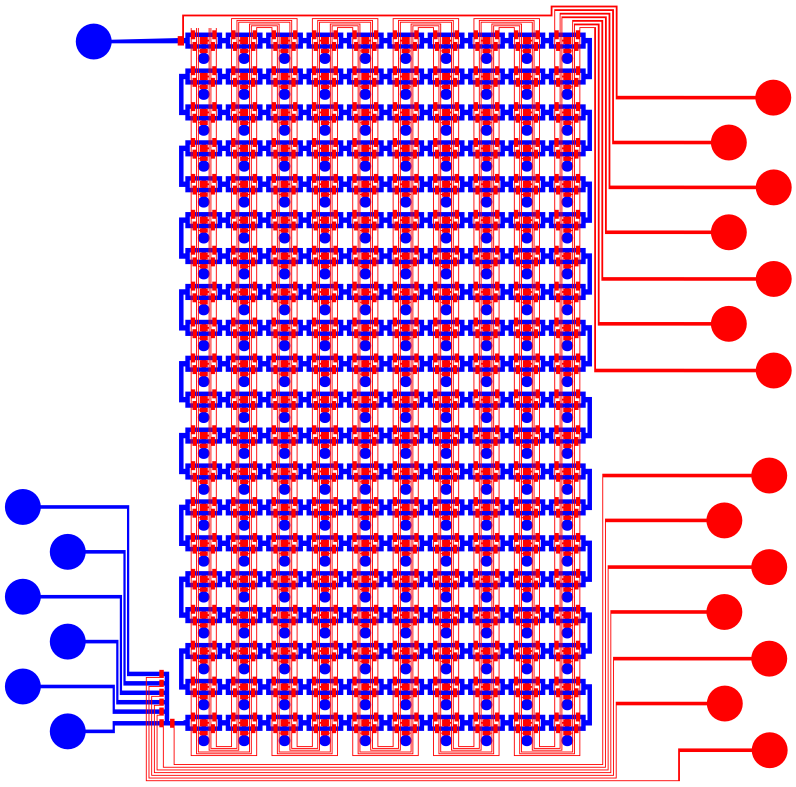


Figure C.9: DNA to Protein Array x4
Unit Cells: 200
Area: 2 x 2 $cm^2$
Number of Valves: 1208
Valve density: 302 valves/$cm^2$
Notes: first test design for MITOMI and specific surface chemistry

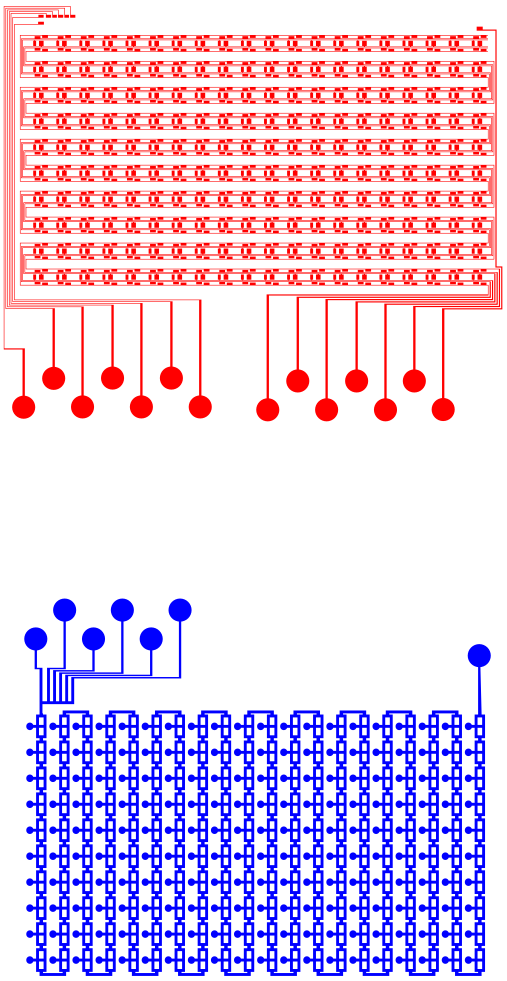


Figure C.10: DNA to Protein Array x4

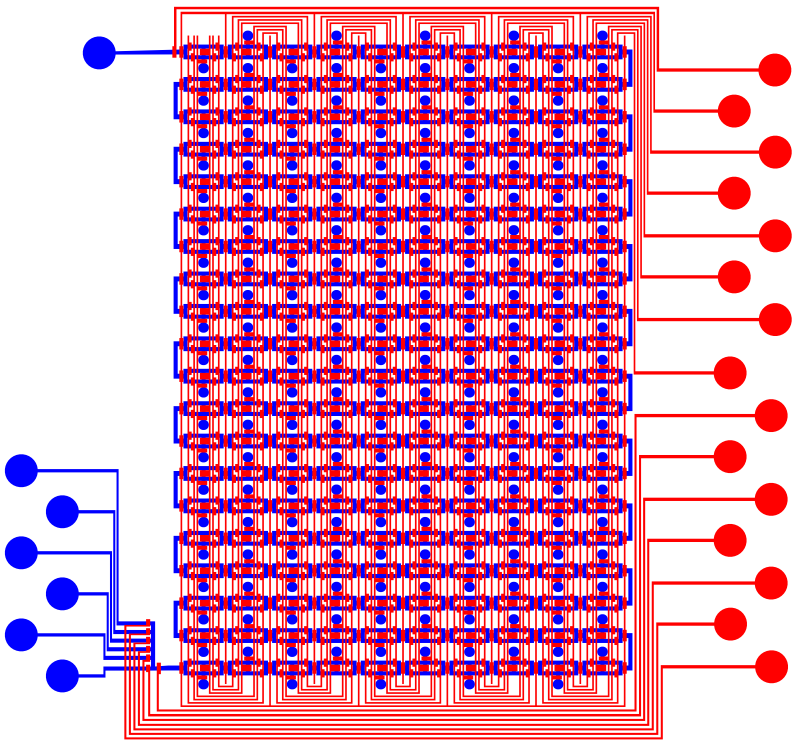


Figure C.11: DNA to Protein Array x5
Unit Cells: 200
Area: 2 x 2 $cm^2$
Number of Valves: 1428
Valve density: 357 valves/$cm^2$
Notes: as DTPAx4 with additional valves for segregating unit cells

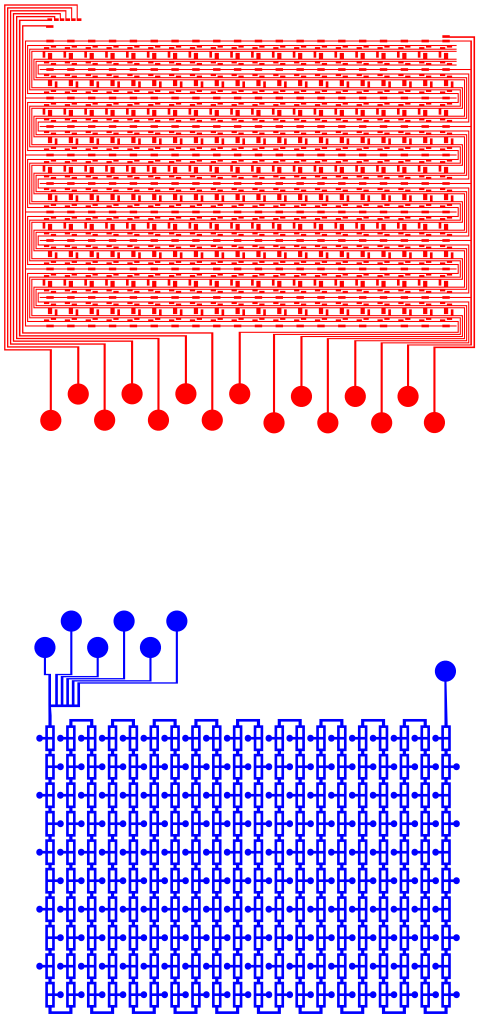


Figure C.12: DNA to Protein Array x5

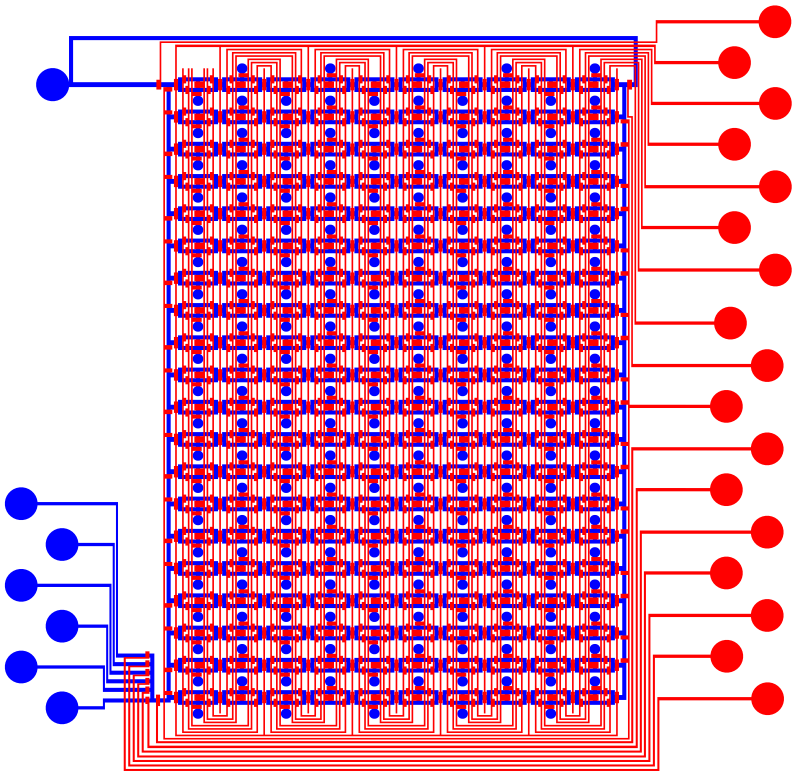


Figure C.13: DNA to Protein Array x6
Unit Cells: 200
Area: 2 x 2 $cm^2$
Number of Valves: 1468
Valve density: 367 valves/$cm^2$
Notes: as DTPAx5 but allowing for parallel rather then serpentine based flow

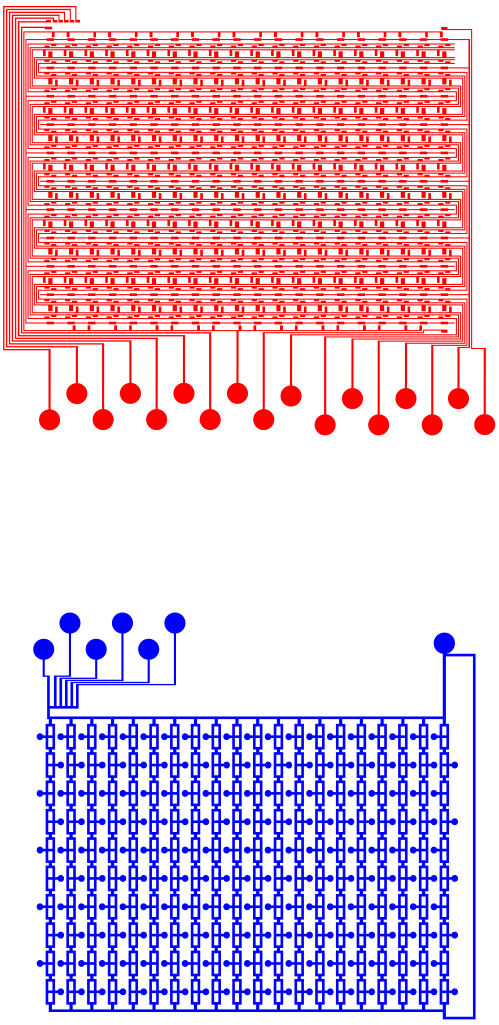


Figure C.14: DNA to Protein Array x6

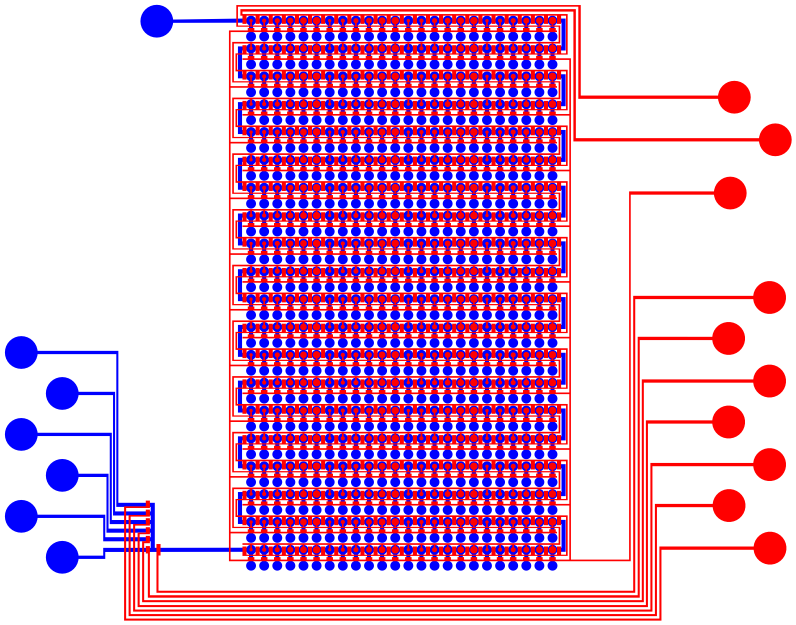


Figure C.15: DNA to Protein Array x7
Unit Cells: 200
Area: 2 x 2 $cm^2$
Number of Valves: 627
Valve density: 157 valves/$cm^2$
Notes: test device for round free-standing MITOMI membrane

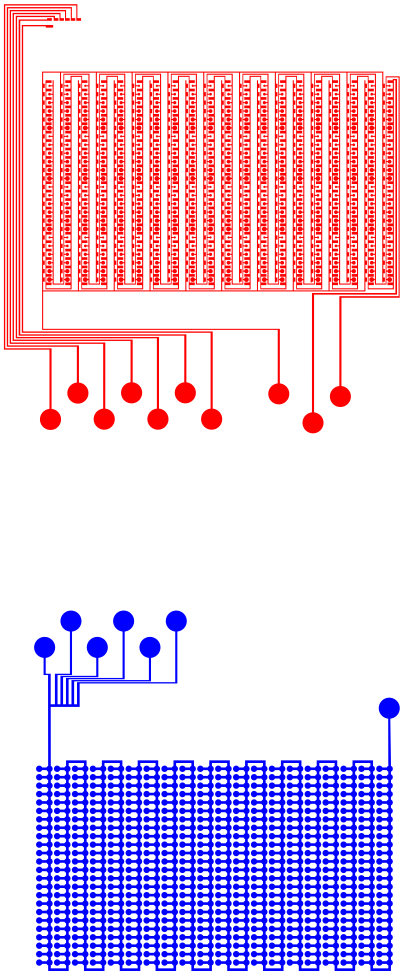


Figure C.16: DNA to Protein Array x7

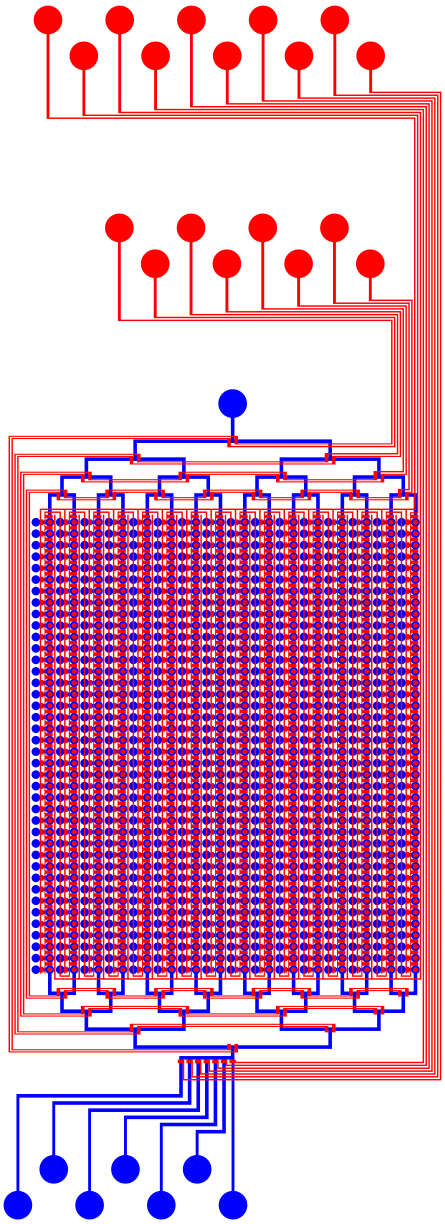


Figure C.17: DNA to Protein Array x8
Unit Cells: 640
Area: 1.5 x 4 $cm^2$
Number of Valves: 1987
Valve density: 331 valves/$cm^2$
Notes: device employs round MITOMI membrane and a multiplexer for addressing rows, parallel adjusted flow design

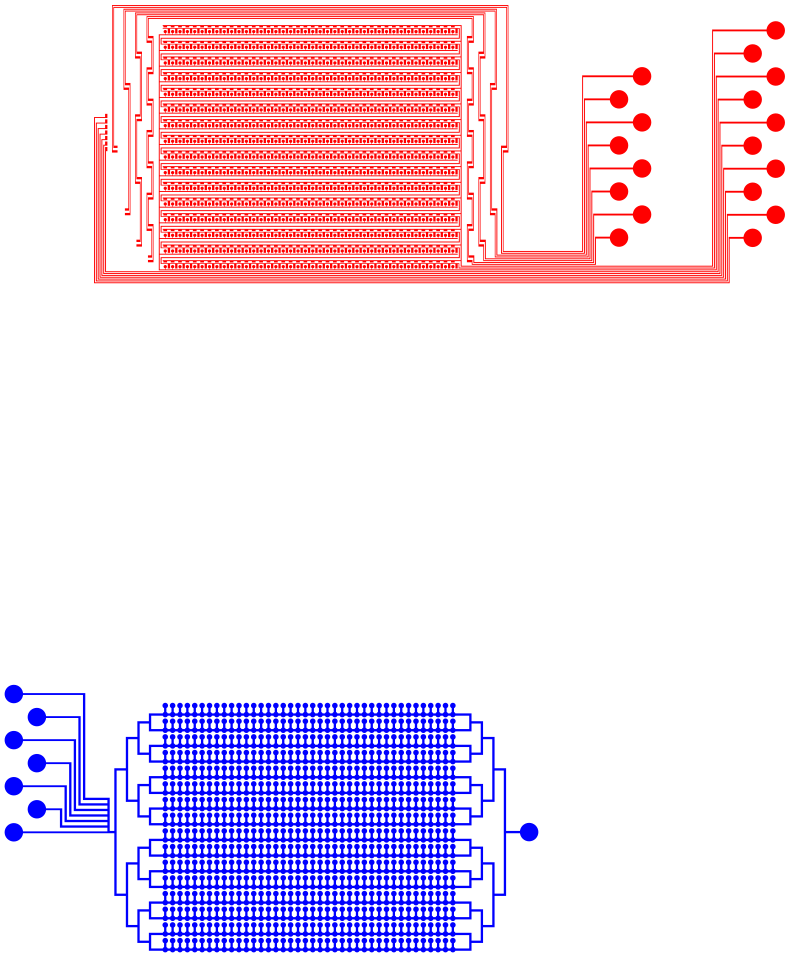


Figure C.18: DNA to Protein Array x8

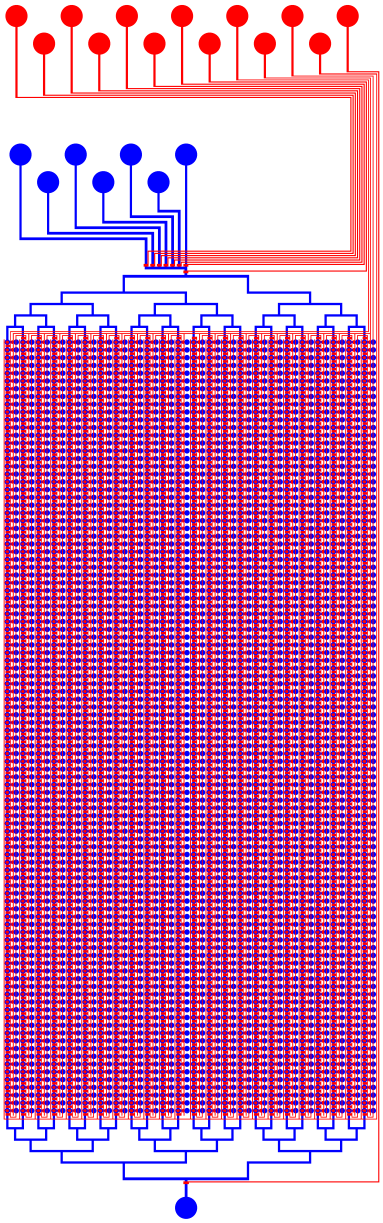


Figure C.19: DNA to Protein Array x9v1
Unit Cells: 2400
Area: 2 x 5 $cm^2$
Number of Valves: 7233
Valve density: 723 valves/$cm^2$
Notes:

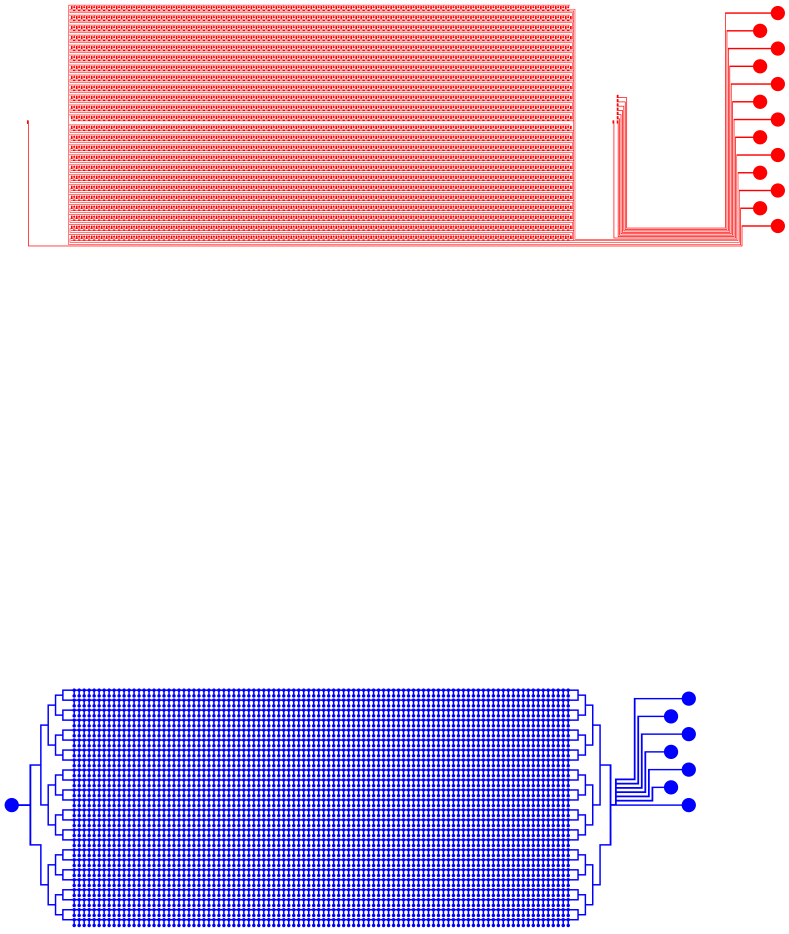


Figure C.20: DNA to Protein Array x9v1

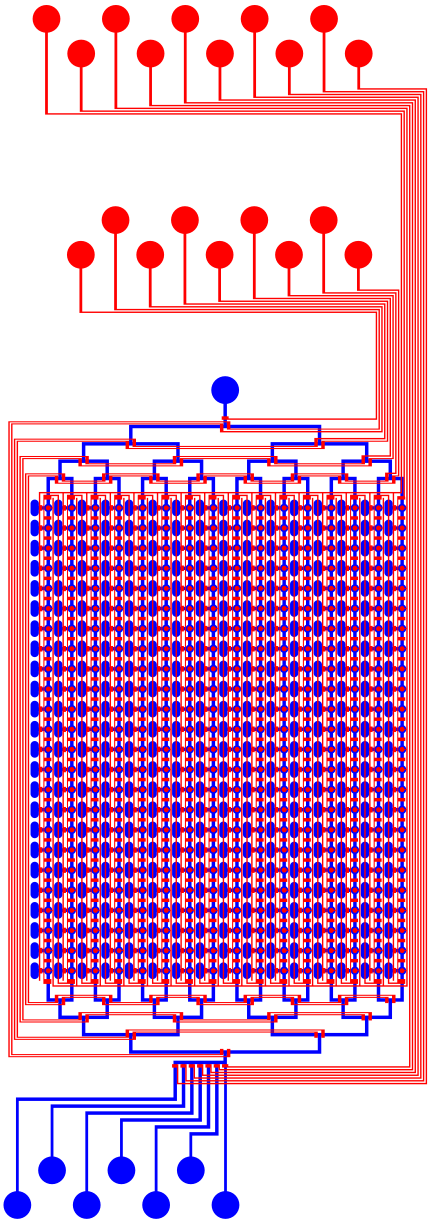


Figure C.21: DNA to Protein Array x10
Unit Cells: 320
Area: 1.5 x 4 $cm^2$
Number of Valves: 1044
Valve density: 174 valves/$cm^2$
Notes:

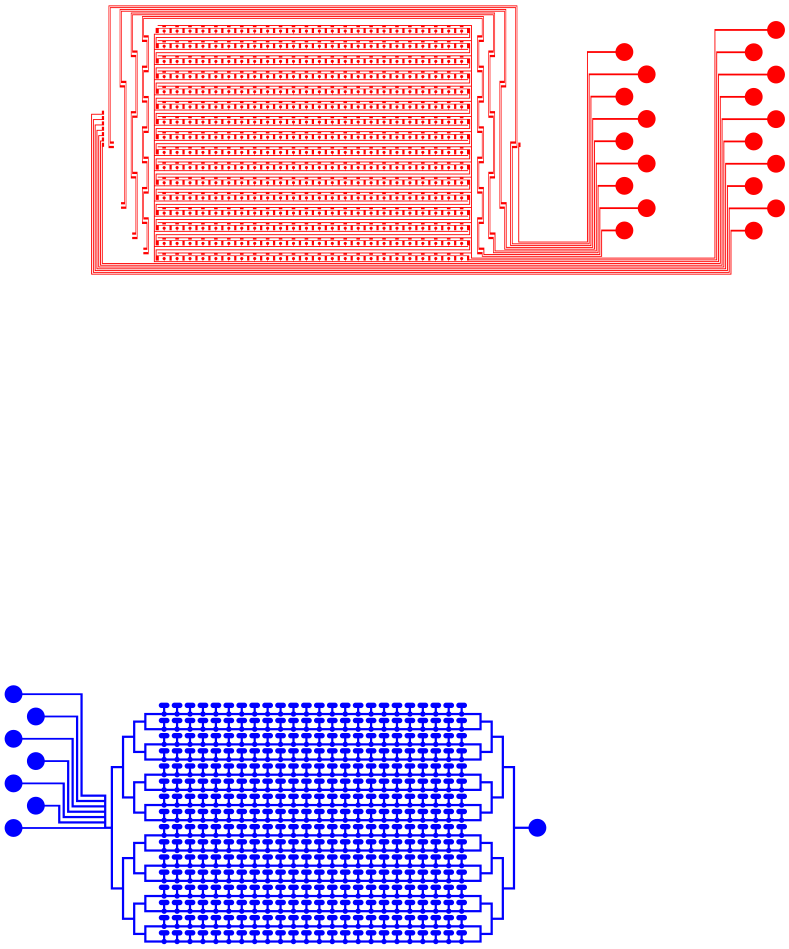


Figure C.22: DNA to Protein Array x10

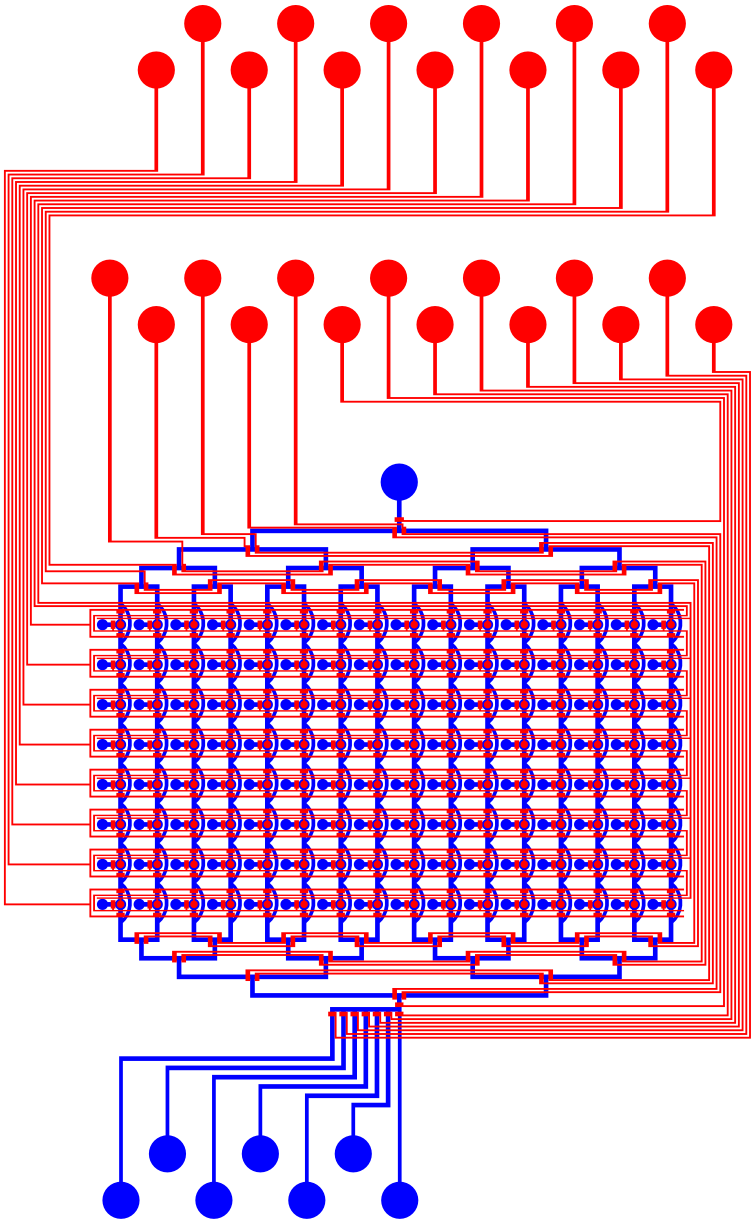


Figure C.23: DNA to Protein Array x11
Unit Cells: 128
Area: 1.5 x 1.5 $cm^2$
Number of Valves: 453
Valve density: 201 valves/$cm^2$
Notes:

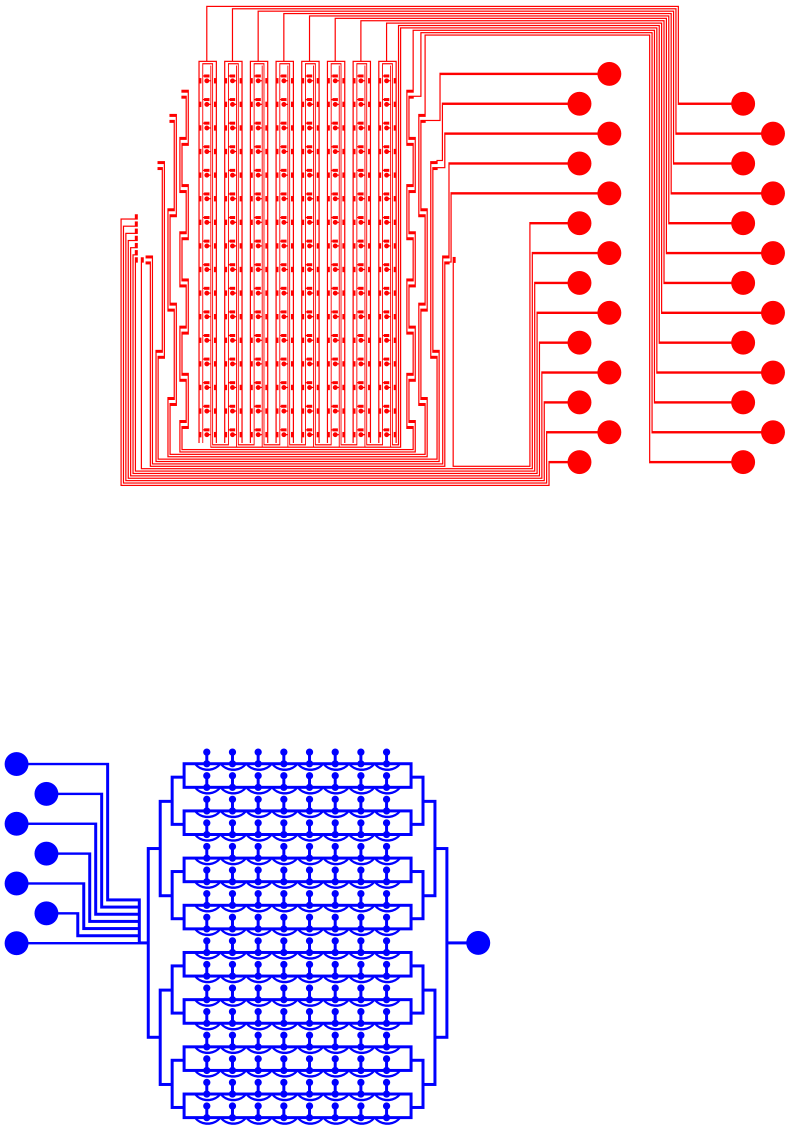


Figure C.24: DNA to Protein Array x11

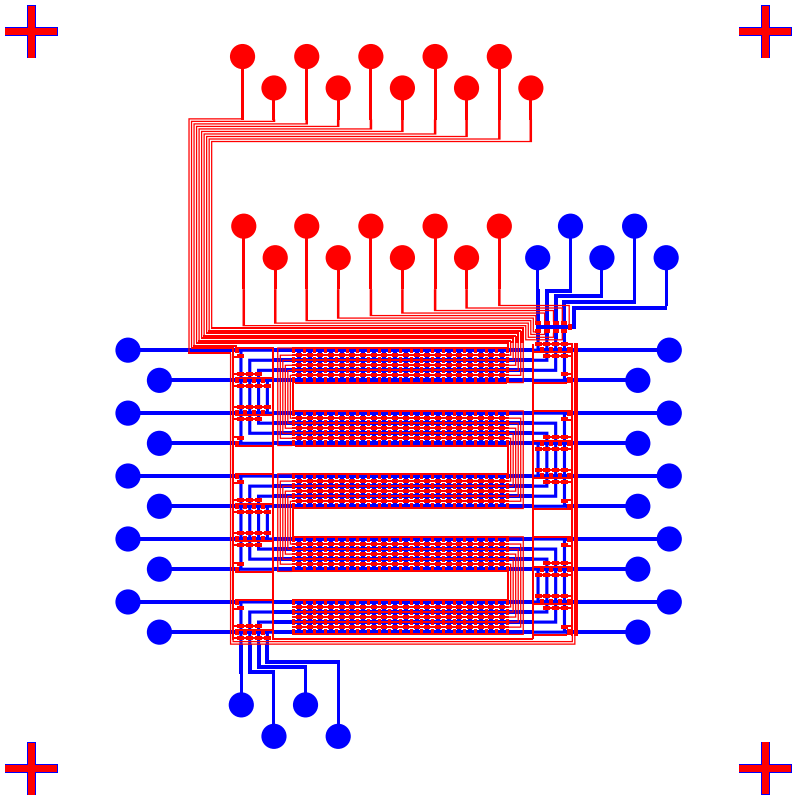


Figure C.25: Wheat Germ revised
Unit Cells: 200
Area: 2 x 2.5 $cm^2$
Number of Valves: 524
Valve density: 105 valves/$cm^2$
Notes:

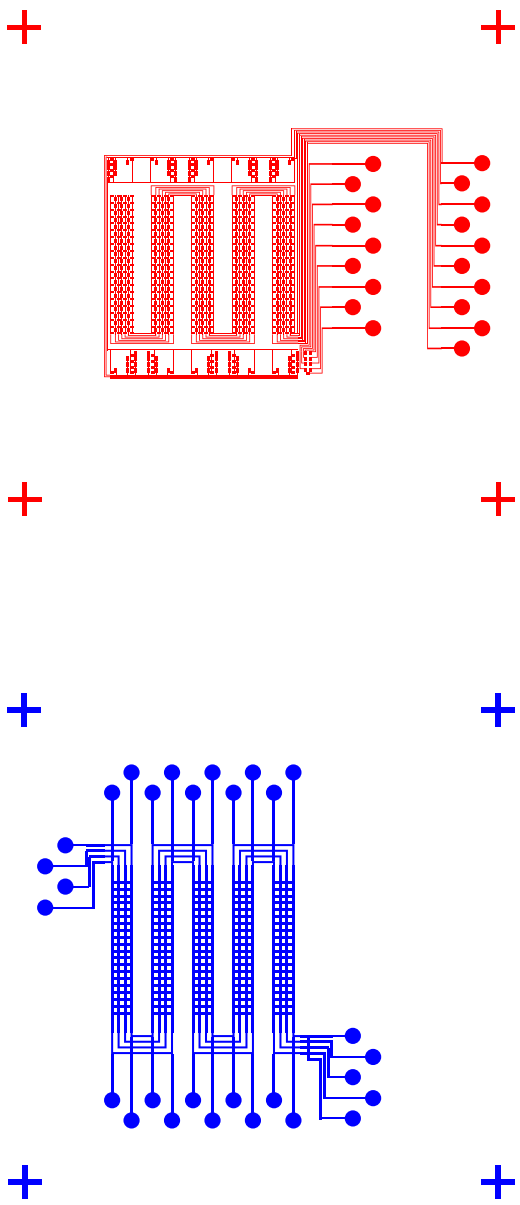


Figure C.26: Wheat Germ revised

# Appendix D

# Sequencing Results

A representative group of PCR derived linear expression templates was bulk sequenced. Both the human and yeast bHLH LTs underwent a total of 70 cycles of PCR, 30 during the 1st step followed by 40 cycles in the second step. The human first step synthesis was performed on 04/14/2005 and the yeast bHLHs were amplified on 07/26/05. The second step was performed on 08/15/2005 and finally purified and sequenced on 12/13/2005.

The resulting sequences were then entered into a discontiguous megablast search and the top matches are shown below.

```
Query = C-Myc NoTag LT
>gi|12803004|gb|BC000917.2|
Homo sapiens v-myc myelocytomatosis viral
oncogene homolog (avian), mRNA
(cDNA clone MGC:5184 IMAGE:3048750), complete cds
 Length=2047

 Score =  564 bits (293),  Expect = 9e-158
 Identities = 303/311 (97%), Gaps = 0/311 (0%)
 Strand=Plus/Plus


C-Myc deltaN249 C-His LT
>gi|12803004|gb|BC000917.2|
Homo sapiens v-myc myelocytomatosis viral
```

```
oncogene homolog (avian), mRNA
(cDNA clone MGC:5184 IMAGE:3048750), complete cds
 Length=2047

 Score =  604 bits (314),  Expect = 7e-170
 Identities = 315/316 (99%), Gaps = 0/316 (0%)
 Strand=Plus/Plus


Query = Cbf1 N-His LT
Subject = YJR060w (Cbf1)
  Score = 1148 bits (597),  Expect = 0.0
 Identities = 623/631 (98%), Gaps = 2/631 (0%)
 Strand=Plus/Plus


Query = MAX iso A NoTag LT
>gi|33871245|gb|BC004516.2|
Homo sapiens MYC associated factor X,
transcript variant 1, mRNA
(cDNA clone MGC:11225 IMAGE:3937573), complete cds
Length=1566

 Score =  923 bits (480),  Expect = 0.0
 Identities = 482/483 (99%), Gaps = 0/483 (0%)
 Strand=Plus/Plus


MAX iso B C-His LT
>gi|13097617|gb|BC003525.1|
Homo sapiens MYC associated factor X,
transcript variant 2, mRNA
(cDNA clone MGC:10775 IMAGE:3607261), complete cds
Length=1963

 Score =  860 bits (447),  Expect = 0.0
 Identities = 451/453 (99%), Gaps = 0/453 (0%)
 Strand=Plus/Plus


Query = Mxi 1 iso B NoTag LT
Alignments
>gi|23272476|gb|BC035128.1|
Homo sapiens cDNA clone IMAGE:5263647, partial cds
```

```
Length=2756

 Score =  621 bits (323),  Expect = 4e-175
 Identities = 330/331 (99%), Gaps = 1/331 (0%)
 Strand=Plus/Plus


>gi|57242780|ref|NM_130439.3|
Homo sapiens MAX interactor 1 (MXI1),
transcript variant 2, mRNA
Length=3470

 Score =  621 bits (323),  Expect = 4e-175
 Identities = 330/331 (99%), Gaps = 1/331 (0%)
 Strand=Plus/Plus


Query = Pho4 C-His LT
Subject = YFR034C (PHO4)
Score = 1044 bits (543),  Expect = 0.0
Identities = 584/592 (98%), Gaps = 7/592 (1%)
Strand=Plus/Plus
```

# Appendix E

# Protocols

## E.1 Photolithography

**Purpose:**Generate 5740 molds for chip fabrication with a rough elevation of 6-10$\mu$m to be used for both the flow as well as the control layer.

**Photoresist:** Shipley 5740

**Developer:** Shipley 2401

**Substrate:** 3' silicon wafers

**Method:**

- Vapor deposit hydroxy-methyl disilane (HMDS) on wafers for 2-4 minutes.
- Spin coat wafers with 5740 in two steps:

| Step | Speed (RPM) | Time (seconds) |
|---|---|---|
| 1 | 500 | 10 |
| 2 | 3000 | 60 |

- Bake wafer at 105°C for 90 seconds.
- Expose on UV source for 40 seconds.

- Develop in 3:1 ($dH_2O$:Developer) until background disappeared.
- Quench reaction in a distilled water bath and rinse with dist. water.
- Anneal flow layer at 180°C for 30 minutes.
- Check the flow layer for a correct height of 8$\mu$m $\pm$2$\mu$m.
- Check the control layer for any interconnects between individual control lines.

# E.2 Chip Fabrication

## E.2.1 Standard 2-Layer PDMS Push-Down Device

**Purpose:** Fabricate a standard two layer push down device from PDMS.

**PDMS:** Sylgard

**Method:**

- Clean both flow and control layer mold of any residual polymerized PDMS.
- Clean both mixing cups of any residual polymerized PDMS.
- Place the control layer mold in a petri dish covered with aluminum foil to facilitate easy removal.
- Vapor deposit TMCS onto both the control and flow mold for 1-2 minutes.
- Prepare 36g of 5:1 Sylgard (30g Part A: 6g Part B) for the control layer and mix for 1 minute followed by degassing for 2 minutes.
- Prepare 15.75g of 20:1 Sylgard (15g Part A: 0.75g Part B) for the flow layer and mix for 1 minute followed by degassing for 2 minutes.

- Pour the mixed 5:1 mixture onto the control layer mold and start degassing.
- Spin coat the 20:1 mixture onto the flow layer with a 15 second ramp, 30 second spin at 3000rpm.
- Remove control layer from vacuum chamber and destroy any residual surface bubbles by blowing on top of the PDMS layer. Carefully remove any visible particles on top of the control channel grid using a toothpick.
- Cure both layers for 30minutes at 80°C.
- Remove both layers from the oven and dice the control layer with a scalpel.
- Punch control input holes.
- Thoroughly clean the channel side of the control layer with tape.
- Align control layer to flow layer.
- Bond devices for 90 minutes at 80°C.
- Remove flow layer from the oven and cut flow layer around each individual chip using a scalpel.
- Peel chips off flow layer and punch flow layer input holes.
- Clean flow channel side thoroughly with tape before bonding to glass substrate.

# E.3 PCR Methods

## E.3.1 Linear Template Generation

The following two subsections describe protocols for the generation of linear expression ready templates for use in wheat germ or rabbit reticulocyte based ITT. Each protocol can be roughly subdivided into the $1^{st}$ and $2^{nd}$ PCR step. The initial PCR step amplifies the target and adds epitope tags which is used as a source for the second generic PCR step. Thus the product of the initial PCR step may seed up to 50 consequent second PCRs.

### E.3.1.1 cDNA Source

**Purpose:** Generate linear expression ready templates for wheat germ or rabbit reticulocyte based *in vitro* transcription/translation from cDNA clones harbored in *E.coli*.

**Template source:** *E.coli* colonie/s

**Polymerase:** Expand High Fidelity Polymerase (Roche)

**Primers:** Gene specific primers at 50$\mu$M (possible to dilute to 200nM in $dH_2O$)

**Method:**

- Pick colonies and suspend in 2.5$\mu$L of Lyse'n Go (Pierce) buffer.
- Heat suspension to 95°C for 7 minutes on thermal cycler then cool to 4°C.
- Add 46.5$\mu$L PCR mix (see below) and 1$\mu$L of the correct gene specific primer pair to each suspension and cycle.
- Purify each reaction using Qiaquick PCR spin columns or on the 6s manifold.

**PCR Mix**

| | |
|---|---|
| 1$\mu$L | 10mM dNTP |
| 5$\mu$L | 10x Buffer + $MgCl_2$ |
| 0.75$\mu$L | HiFi Polymerase |
| 39.75$\mu$L | $dH_2O$ |
| 46.5$\mu$ | total |

**Cycle Program:**

| Step | Temperature (°C) | Time (minutes) | Cycles |
|---|---|---|---|
| 1 | 94 | 4:00 | 1x |
| 2 | 94 | 0:30 | 10x |
| 3 | 55 | 1:00 | |
| 4 | 72 | 1:30 | |
| 5 | 72 | 7:00 | |
| 6 | 4 | $\infty$ | |

- Elute with 50-100$\mu$L of EB.
- Samples may be stored at -20°C and are used as stock solutions for the next two PCR steps.

**Purpose:** Setup reaction to generate final PCR product to be used in ITT.

**Template source**: Purified PCR products from above.

**Polymerase:** Expand High Fidelity Polymerase (Roche).

**Primers:** 5'ext1 + 3'ext2 diluted 1/200 in $dH_2O$ to a final concentration of 250nM each primer.

**Method:**

- Setup a PCR reaction according to table below and cycle.
- Add 1$\mu$L of 5'final + 3'final primers diluted 1/10 in $dH_2O$ to a final concentration of 5$\mu$M each. Note: the 5'final and 3'final primer may be labeled with Cy3

**PCR Mix**

| | |
|---|---|
| 1$\mu$L | 10mM dNTP |
| 5$\mu$L | 10x Buffer + $MgCl_2$ |
| 1$\mu$L | primers |
| 2$\mu$L | template |
| 0.75$\mu$L | HiFi Polymerase |
| 40.25$\mu$L | $dH_2O$ |
| 50$\mu$ | total |

*N*ote: it is possible to double all quantities and run a 100$\mu$L large reaction if larger yields are desired. The below cycle times remain the same.

**Cycle Program:**

| Step | Temperature (°C) | Time (minutes) | Cycles |
|---|---|---|---|
| 1 | 94 | 4:00 | 1x |
| 2 | 94 | 0:30 | 10x |
| 3 | 53 | 1:00 | |
| 4 | 72 | 1:30 | |
| 5 | 72 | 7:00 | |
| 6 | 4 | $\infty$ | |

and biotin respectively, if detection and pull-down of the template is required.

- Continue cycling with below parameters.

**Cycle Program:**

| Step | Temperature (°C) | Time (minutes) | Cycles |
|---|---|---|---|
| 1 | 94 | 4:00 | 1x |
| 2 | 94 | 0:30 | 30x |
| 3 | 50 | 1:00 | |
| 4 | 72 | 1:30 | |
| 5 | 72 | 7:00 | |
| 6 | 4 | $\infty$ | |

- Check 0.5$\mu$L of each sample on a 1% agarose gel.

- Purify each reaction on a Qiaquick spin column or 6s manifold eluting in 50-100$\mu$L 1% BSA $dH_2O$ solution if used for spotting or 50-100$\mu$L EB for flow

deposition.

#### E.3.1.2 Genomic Source

**Purpose:** Generate linear expression ready templates for wheat germ or rabbit reticulocyte based *in vitro* transcription/translation using yeast genomic DNA as the source.

**Template source**: yeast genomic DNA at ∼200ng/$\mu$L (SeeGene).

**Polymerase:** Expand High Fidelity Polymerase (Roche).

**Primers:** Gene specific primers at 50$\mu$M.

**Method:**

- Prepare PCR mix and cycle according to tables below.

**PCR Mix**

| | |
|---|---|
| 1$\mu$L | 10mM dNTP |
| 5$\mu$L | 10x Buffer + $MgCl_2$ |
| 1$\mu$L | primers |
| 5$\mu$L | template |
| 0.75$\mu$L | HiFi Polymerase |
| 37.25$\mu$L | $dH_2O$ |
| 50$\mu$ | total |

**Cycle Program:**

| Step | Temperature (°C) | Time (minutes) | Cycles |
|---|---|---|---|
| 1 | 94 | 4:00 | 1x |
| 2 | 94 | 0:30 | 30x |
| 3 | 53 | 1:00 | |
| 4 | 72 | 1:30 | |
| 5 | 72 | 7:00 | |
| 6 | 4 | $\infty$ | |

*N*ote: it is possible to drop the cycles to 10 instead of 30 in order to reduce accumulation of PCR induced point mutations.

- Optionally check 0.5$\mu$L of each sample on a 1% agarose gel.

- Purify each reaction on a Qiaquick PCR spin column or 6s manifold eluting in 50-100$\mu$L EB.

**Purpose:** Second step of the PCR generating the final expression ready linear templates.

**Template source:** Purified PCR products from above.

**Polymerase:** Expand High Fidelity Polymerase (Roche).

**Primers:** 5'ext1 + 3'ext2 diluted 1/200 in $dH_2O$ to a final concentration of 250nM each.

**Method:**

- Setup a PCR reaction and cycle according to the tables below.

| | **PCR Mix** |
|---|---|
| 1$\mu$L | 10mM dNTP |
| 5$\mu$L | 10x Buffer + $MgCl_2$ |
| 1$\mu$L | primers |
| 1$\mu$L | template |
| 0.75$\mu$L | HiFi Polymerase |
| 41.25$\mu$L | $dH_2O$ |
| 50$\mu$ | total |

*N*ote: it is possible to double all quantities and run a 100$\mu$L large reaction if larger yields are desired. The below cycle times remain the same.

- Add 1$\mu$L (add 2$\mu$L to 100$\mu$L reactions) of 5'final + 3'final primers diluted 1/10 in $dH_2O$ to a final concentration of 5$\mu$M each. Note: the 5'final and 3'final primer may be labeled with Cy3 and biotin respectively, if detection and pull-down of the template is required.

**Cycle Program:**

| Step | Temperature (°C) | Time (minutes) | Cycles |
|---|---|---|---|
| 1 | 94 | 4:00 | 1x |
| 2 | 94 | 0:30 | 10x |
| 3 | 56.5 | 1:00 | |
| 4 | 72 | 1:30 | |
| 5 | 72 | 7:00 | |
| 6 | 4 | $\infty$ | |

- Continue cycling according to below program.

**Cycle Program:**

| Step | Temperature (°C) | Time (minutes) | Cycles |
|---|---|---|---|
| 1 | 94 | 4:00 | 1x |
| 2 | 94 | 0:30 | 30x |
| 3 | 50 | 1:00 | |
| 4 | 72 | 1:30 | |
| 5 | 72 | 7:00 | |
| 6 | 4 | $\infty$ | |

- Check 0.5$\mu$L of each sample on a 1% agarose gel.

- Purify each reaction on a Qiaquick spin column or 6s manifold eluting in 50-100$\mu$L 1% BSA $dH_2O$ solution if used for spotting or 50-100$\mu$L EB for flow deposition.

## E.3.2 E-box Library Generation

### E.3.2.1 PCR

Purpose: Generate double stranded Ebox sequences by primer extension.

Primer sources: all primers should be suspended to 50$\mu$M in TE buffer.

Method:

- prepare a PCR mix as indicated below for each reaction.

**PCR Mix**

| | |
|---|---|
| 1$\mu$L | 10mM dNTP |
| 1$\mu$L | 10x Buffer + $MgCl_2$ |
| 1$\mu$L | 5' Complement + Cy5/Cy3 primer |
| 1.2$\mu$L | unique primer |
| 0.15$\mu$L | HiFi Polymerase |
| 5.65$\mu$L | $dH_2O$ |
| 10$\mu$ | total |

*N*ote: it is possible to increase volumes according to the final quantity of product required.

- Cycle according to the protocol below.

**Cycle Program:**

| Step | Temperature (°C) | Time (minutes) | Cycles |
|---|---|---|---|
| 1 | 94 | 4:00 | 1x |
| 2 | 94 | 0:30 | 10x |
| 3 | 50 | 1:00 | |
| 4 | 72 | 1:30 | |
| 5 | 72 | 7:00 | |
| 6 | 4 | $\infty$ | |

- Purify each reaction using Qiagen's nucleotide removal kit. Purification can be performed on a 6s vacuum manifold if necessary.

# E.4 Miscellaneous

## E.4.1 Coating Epoxy Slides with BSA

Purpose: Generate a BSA monolayer on epoxy slides for use as a substrate for spotted arrays.

Method:

- Prepare 500mL of a 1%BSA PBS solution

- Submerge slides in solution and incubate for 2-3 hours at room temperature
- Rinse each slide with 18MΩ water and dry with a stream of nitrogen or air

Storage: Slides may be stored at room temperature for several months without noticeable loss of activity.

# Bibliography

[1] M. A. Unger, H. P. Chou, T. Thorsen, A. Scherer, S. R. Quake, *Science* **288**, 113 (2000). 0036-8075 Journal Article.

[2] J. Liu, M. Enzelberger, S. Quake, *Electrophoresis* **23**, 1531 (2002). 0173-0835 Journal Article.

[3] J. S. Marcus, W. F. Anderson, S. R. Quake, *Anal Chem* **78**, 3084 (2006). 0003-2700 (Print) Journal Article Research Support, N.I.H., Extramural.

[4] C. C. Lee, *et al.*, *Science* **310**, 1793 (2005). 1095-9203 (Electronic) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

[5] T. Thorsen, S. J. Maerkl, S. R. Quake, *Science* **298**, 580 (2002). 1095-9203 Journal Article.

[6] C. L. Hansen, M. O. Sommer, S. R. Quake, *Proc Natl Acad Sci U S A* **101**, 14431 (2004). 0027-8424 Journal Article.

[7] S. J. Maerkl, S. R. Quake, *Science* **315**, 233 (2007). 1095-9203 (Electronic) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

[8] S. Park, *et al.*, *Proc Natl Acad Sci U S A* **100**, 13910 (2003). 0027-8424 Journal Article.

[9] P. J. Hung, P. J. Lee, P. Sabounchi, R. Lin, L. P. Lee, *Biotechnol Bioeng* **89**, 1 (2005). 0006-3592 Journal Article.

[10] E. M. Lucchetta, J. H. Lee, L. A. Fu, N. H. Patel, R. F. Ismagilov, *Nature* **434**, 1134 (2005). 1476-4687 Journal Article.

[11] J. Liu, C. Hansen, S. R. Quake, *Anal Chem* **75**, 4718 (2003). 0003-2700 Journal Article.

[12] E. Delamarche, A. Bernard, H. Schmid, B. Michel, H. Biebuyck, *Science* **276**, 779 (1997). 0036-8075 Journal Article.

[13] T. Kinpara, *et al.*, *J Biochem (Tokyo)* **136**, 149 (2004). 0021-924x Journal Article.

[14] L. R. Huang, *et al.*, *Nat Biotechnol* **20**, 1048 (2002). 1087-0156 Technical Report.

[15] C. L. Hansen, E. Skordalakes, J. M. Berger, S. R. Quake, *Proc Natl Acad Sci U S A* **99**, 16531 (2002). 0027-8424 Journal Article.

[16] B. Zheng, J. D. Tice, L. S. Roach, R. F. Ismagilov, *Angew Chem Int Ed Engl* **43**, 2508 (2004). 0570-0833 Journal Article.

[17] B. Zheng, R. F. Ismagilov, *Angew Chem Int Ed Engl* **44**, 2520 (2005). 0570-0833 Journal Article.

[18] F. H. Arnold, *Nature* **409**, 253 (2001). 0028-0836 Journal Article Review Review, Tutorial.

[19] M. W. Nirenberg, J. H. Matthaei, *Proc Natl Acad Sci U S A* **47**, 1588 (1961). 0027-8424 (Print) Journal Article.

[20] D. Clayton, *et al.*, *Proc Natl Acad Sci U S A* **101**, 4764 (2004). 0027-8424 (Print) Journal Article Research Support, U.S. Gov't, P.H.S.

[21] P. E. Dawson, S. B. Kent, *Annu Rev Biochem* **69**, 923 (2000). 0066-4154 (Print) In Vitro Journal Article Review.

[22] T. L. Hendrickson, V. de Crecy-Lagard, P. Schimmel, *Annual Review of Biochemistry* **73**, 147 (2004).

[23] D. A. Dougherty, *Curr Opin Chem Biol* **4**, 645 (2000). 1367-5931 (Print) Journal Article Research Support, U.S. Gov't, P.H.S. Review.

[24] M. Taki, T. Hohsaka, H. Murakami, K. Taira, M. Sisido, *FEBS Lett* **507**, 35 (2001). 0014-5793 (Print) Journal Article Research Support, Non-U.S. Gov't.

[25] T. Hohsaka, Y. Ashizuka, H. Taira, H. Murakami, M. Sisido, *Biochemistry* **40**, 11060 (2001). 0006-2960 (Print) Journal Article Research Support, Non-U.S. Gov't.

[26] S. Gite, S. Mamaev, J. Olejnik, K. Rothschild, *Anal Biochem* **279**, 218 (2000). 0003-2697 (Print) Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[27] D. Kiga, *et al.*, *Proc Natl Acad Sci U S A* **99**, 9715 (2002). 0027-8424 (Print) Journal Article.

[28] Y. Shimizu, *et al.*, *Nat Biotechnol* **19**, 751 (2001). 1087-0156 (Print) Journal Article Research Support, Non-U.S. Gov't.

[29] V. Noireaux, R. Bar-Ziv, A. Libchaber, *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12672 (2003).

[30] N. Ramachandran, *et al.*, *Science* **305**, 86 (2004).

[31] P. Angenendt, *et al.*, *Anal Chem* **76**, 1844 (2004). 0003-2700 Journal Article.

[32] T. Sawasaki, T. Ogasawara, R. Morishita, Y. Endo, *Proceedings of the National Academy of Sciences of the United States of America* **99**, 14652 (2002).

[33] T. Kigawa, *et al.*, *FEBS Lett* **442**, 15 (1999). 0014-5793 (Print) Journal Article Research Support, Non-U.S. Gov't.

[34] K. Madin, T. Sawasaki, T. Ogasawara, Y. Endo, *Proc Natl Acad Sci U S A* **97**, 559 (2000). 0027-8424 (Print) Journal Article Research Support, Non-U.S. Gov't.

[35] M. Zuker, *Nucleic Acids Res* **31**, 3406 (2003). 1362-4962 (Electronic) Journal Article Research Support, U.S. Gov't, P.H.S.

[36] R. F. Ismagilov, J. M. Ng, P. J. Kenis, G. M. Whitesides, *Anal Chem* **73**, 5207 (2001). 0003-2700 (Print) Journal Article.

[37] M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995). 0036-8075 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[38] J. S. Kim, R. T. Raines, *Protein Sci* **2**, 348 (1993). 0961-8368 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[39] B. R. Kelemen, *et al.*, *Nucleic Acids Res* **27**, 3696 (1999). 1362-4962 (Electronic) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[40] M. E. Massari, C. Murre, *Molecular & Cellular Biology* **20**, 429 (2000).

[41] K. A. Robinson, J. I. Koepke, M. Kharodawala, J. M. Lopes, *Nucleic Acids Res* **28**, 4460 (2000). 1362-4962 Journal Article.

[42] C. Murre, P. S. McCaw, D. Baltimore, *Cell* **56**, 777 (1989).

[43] T. K. Blackwell, L. Kretzner, E. M. Blackwood, R. N. Eisenman, H. Weintraub, *Science* **250**, 1149 (1990).

[44] E. V. Prochownik, M. E. VanAntwerp, *Proceedings of the National Academy of Sciences of the United States of America* **90**, 960 (1993).

[45] T. K. Blackwell, *et al.*, *Molecular & Cellular Biology* **13**, 5216 (1993).

[46] L. Kretzner, E. M. Blackwood, R. N. Eisenman, *Nature* **359**, 426 (1992).

[47] E. M. Blackwood, B. Luscher, R. N. Eisenman, *Genes & Development* **6**, 71 (1992).

[48] E. M. Blackwood, R. N. Eisenman, *Science* **251**, 1211 (1991).

[49] T. D. Halazonetis, A. N. Kandil, *Proc Natl Acad Sci U S A* **88**, 6162 (1991). 0027-8424 Journal Article.

[50] S. J. Berberich, M. D. Cole, *Genes Dev* **6**, 166 (1992). 0890-9369 Journal Article.

[51] M. Springer, D. D. Wykoff, N. Miller, E. K. O'Shea, *PLoS Biol* **1**, E28 (2003). 1545-7885 Journal Article.

[52] M. Byrne, N. Miller, M. Springer, E. K. O'Shea, *J Mol Biol* **335**, 57 (2004). 0022-2836 Journal Article.

[53] A. Kaffman, N. M. Rank, E. M. O'Neill, L. S. Huang, E. K. O'Shea, *Nature* **396**, 482 (1998). 0028-0836 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[54] A. Komeili, E. K. O'Shea, *Science* **284**, 977 (1999). 0036-8075 (Print) Journal Article Research Support, Non-U.S. Gov't.

[55] W. R. Atchley, W. M. Fitch, *Proc Natl Acad Sci U S A* **94**, 5172 (1997). 0027-8424 Journal Article.

[56] C. Murre, *et al.*, *Biochim Biophys Acta* **1218**, 129 (1994). 0006-3002 Journal Article Review Review, Tutorial.

[57] A. R. Ferre-D'Amare, G. C. Prendergast, E. B. Ziff, S. K. Burley, *Nature* **363**, 38 (1993). 0028-0836 Journal Article.

[58] T. Ellenberger, D. Fass, M. Arnaud, S. C. Harrison, *Genes Dev* **8**, 970 (1994). 0890-9369 Journal Article.

[59] T. Shimizu, *et al.*, *Embo J* **16**, 4689 (1997). 0261-4189 Journal Article.

[60] S. K. Nair, S. K. Burley, *Cell* **112**, 193 (2003). 0092-8674 Journal Article.

[61] S. Sauve, L. Tremblay, P. Lavigne, *J Mol Biol* **342**, 813 (2004). 0022-2836 Journal Article.

[62] A. Orian, *et al.*, *Genes Dev* **17**, 1101 (2003). 0890-9369 Journal Article.

[63] S. Cawley, *et al.*, *Cell* **116**, 499 (2004). 0092-8674 Journal Article.

[64] A. S. Carroll, A. C. Bishop, J. L. DeRisi, K. M. Shokat, E. K. O'Shea, *Proc Natl Acad Sci U S A* **98**, 12578 (2001). 0027-8424 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[65] N. Ogawa, J. DeRisi, P. O. Brown, *Mol Biol Cell* **11**, 4309 (2000). 1059-1524 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[66] N. A. Kent, S. M. Eibert, J. Mellor, *J Biol Chem* **279**, 27116 (2004). 0021-9258 (Print) Journal Article Research Support, Non-U.S. Gov't.

[67] M. Cai, R. W. Davis, *Cell* **61**, 437 (1990). 0092-8674 (Print) Comparative Study Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[68] R. E. Baker, D. C. Masison, *Mol Cell Biol* **10**, 2458 (1990). 0270-7306 (Print) Journal Article Research Support, U.S. Gov't, P.H.S.

[69] S. Ghaemmaghami, *et al.*, *Nature* **425**, 737 (2003). 1476-4687 Journal Article.

[70] G. C. Yuan, *et al.*, *Science* **309**, 626 (2005). 1095-9203 (Electronic) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[71] C. T. Harbison, *et al.*, *Nature* **431**, 99 (2004). 1476-4687 (Electronic) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[72] Y. Takeda, A. Sarai, V. M. Rivera, *Proc Natl Acad Sci U S A* **86**, 439 (1989). 0027-8424 (Print) Journal Article Research Support, U.S. Gov't, P.H.S.

[73] A. Sarai, Y. Takeda, *Proc Natl Acad Sci U S A* **86**, 6513 (1989). 0027-8424 (Print) Journal Article Research Support, U.S. Gov't, P.H.S.

[74] K. D. MacIsaac, E. Fraenkel, *PLoS Comput Biol* **2**, e36 (2006). 1553-7358 (Electronic) Journal Article Research Support, Non-U.S. Gov't.

[75] G. D. Stormo, *J Theor Biol* **195**, 135 (1998). 0022-5193 (Print) Letter.

[76] G. D. Stormo, D. S. Fields, *Trends Biochem Sci* **23**, 109 (1998). 0968-0004 (Print) Journal Article Review.

[77] M. L. Bulyk, *Genome Biol* **5**, 201 (2003). 1465-6914 (Electronic) Journal Article Research Support, Non-U.S. Gov't Review.

[78] T. D. Schneider, R. M. Stephens, *Nucleic Acids Res* **18**, 6097 (1990). 0305-1048 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[79] G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, *Genome Res* **14**, 1188 (2004). 1088-9051 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[80] T. K. Man, G. D. Stormo, *Nucleic Acids Res* **29**, 2471 (2001). 1362-4962 (Electronic) Journal Article Research Support, U.S. Gov't, P.H.S.

[81] P. V. Benos, M. L. Bulyk, G. D. Stormo, *Nucleic Acids Res* **30**, 4442 (2002). 1362-4962 (Electronic) Comparative Study Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[82] M. L. Bulyk, P. L. Johnson, G. M. Church, *Nucleic Acids Res* **30**, 1255 (2002). 1362-4962 (Electronic) Journal Article Research Support, U.S. Gov't, Non-P.H.S.

[83] D. S. Spinner, S. Liu, S. W. Wang, J. Schmidt, *Journal of Molecular Biology* **317**, 431 (2002).

[84] A. V. Grinberg, T. Kerppola, *J Biol Chem* **278**, 11227 (2003). 0021-9258 (Print) Journal Article.

[85] S. Park, *et al.*, *Biochim Biophys Acta* **1670**, 217 (2004). 0006-3002 Journal Article.

[86] Y. Fujii, T. Shimizu, T. Toda, M. Yanagida, T. Hakoshima, *Nat Struct Biol* **7**, 889 (2000). 1072-8368 (Print) Journal Article Research Support, Non-U.S. Gov't.

[87] M. A. Schumacher, R. H. Goodman, R. G. Brennan, *J Biol Chem* **275**, 35242 (2000). 0021-9258 (Print) Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[88] S. Impey, *et al.*, *Cell* **119**, 1041 (2004). 0092-8674 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.

[89] O. Hallikas, *et al.*, *Cell* **124**, 47 (2006). 0092-8674 (Print) Journal Article Research Support, Non-U.S. Gov't.

[90] J. E. Hooper, M. P. Scott, *Nat Rev Mol Cell Biol* **6**, 306 (2005). 1471-0072 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review.

[91] K. Ferrell, C. R. Wilkinson, W. Dubiel, C. Gordon, *Trends in Biochemical Sciences* **25**, 83 (2000).

[92] J. Walz, *et al.*, *J Struct Biol* **121**, 19 (1998). 1047-8477 Journal Article.

[93] M. Groll, *et al.*, *Nature* **386**, 463 (1997). 0028-0836 Journal Article.

[94] E. A. Winzeler, *et al.*, *Science* **285**, 901 (1999). 0036-8075 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[95] D. M. Gelperin, *et al.*, *Genes Dev* **19**, 2816 (2005). 0890-9369 (Print) Journal Article Research Support, N.I.H., Extramural.

[96] W. K. Huh, *et al.*, *Nature* **425**, 686 (2003). 1476-4687 Journal Article.

[97] J. R. Newman, *et al.*, *Nature* **441**, 840 (2006). 1476-4687 (Electronic) Journal Article Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.

[98] S. Paliwal, *et al.*, *Nature* **446**, 46 (2007). 1476-4687 (Electronic) Journal Article Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S.

[99] A. P. Gasch, *et al.*, *Mol Biol Cell* **12**, 2987 (2001). 1059-1524 (Print) Comparative Study Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.

[100] J. J. Wyrick, *et al.*, *Nature* **402**, 418 (1999). 0028-0836 (Print) Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S.