Chapter 1
Optimal DNA Codes for Computing and Self-Assembly

Max H. Garzon
The University of Memphis, USA

Vinhthuy Phan
The University of Memphis, USA

Andrew Neel
The University of Memphis, USA

ABSTRACT
DNA has been re-discovered and explored in the last decade as a “smart glue” for self-assembly from the “bottom-up” at nanoscales through mesoscales to micro- and macro-scales. These applications require an unprecedented degree of precision in placing atom-scale components. Finding large sets of probes to serve as anchors for such applications has been thus explored in the last few years through several methods. We describe results of a tour de force to conduct an exhaustive search to produce large codes that are (nearly) maximal sets while guaranteeing high quality, as measured by the minimum Gibbs energy between any pair of code words, and other criteria. We also present a quantitative characterization of the sets for sizes up to 20-mers and show how critical building blocks can be extracted to produce codes of very high quality for larger lengths by probabilistic combinations, for which an exhaustive search is out of reach.

INTRODUCTION
Until very recently, DNA has been considered just as primary storage of instructions for the makeup of living organisms. Over a decade ago, (Adleman, 1994) gave the successful demonstration of a DNA’s potential use for different purposes, more specifically, a solution to the Hamiltonian Path Problem (HPP), a problem considered beyond reach for feasible solution on conventional computers. A computational equivalent of the well known traveling salesman problem (TSP), the Hamiltonian Path problem (HPP) calls for deciding the existence of a Hamiltonian path to traverse the edges a directed graph with two singled out as source and destination vertices. The problem calls for a Boolean decision whether there exists a Hamiltonian path joining the source
to the destination (i.e., a path passing through every vertex exactly once). Later in the 1990s, several authors (Winfree et al., 1998; Seaman, 1999; Benenson et al., 2001) realized a second, perhaps more important use of DNA, as a “smart glue” for self-assembly from the “bottom-up” of atomic and molecular components to form much larger and complex structures in the order of mesoscales and microscales. Even more recently, (Neel and Garzon, 2006; Garzon et al., 2003) have suggested yet a third use of DNA as a structure capable of indexing large data repositories in terascale memories searchable in feasible times, for example in the form of DNA microarrays or DNA chips (Draghici, 2003).

All these applications require an unprecedented degree of precision in placing and operating on atomic-scale components, unachievable by traditional “top-down” assembly methods, such as microlithography. The specificity and selectivity of hybridization between (nearly) complementary small DNA oligonucleotides confers upon DNA a well recognized advantage to serve as a material to self-assemble such components. Several recent results have demonstrated the exquisite degree of control that DNA would offer on structures built on DNA scaffolds. Methods to find large sets of probes (or code sets, as will be referred to below) to serve as building blocks for such operations have thus been explored in the last few years. Although the problem of finding optimal sets has been shown to be computationally difficulty (NP-complete), even for very specific and very coarse measures of hybridization affinity (Phan and Garzon, 2008), a variety of methods have been tried to find such DNA codeword sets, as well as of techniques to measure their quality. These methods have produced good DNA codes that have sufficed for prototypes of the applications mentioned above. For example, (Phan and Garzon, 2008) used a new construction called shuffling that theoretically builds good codes of size nearly optimal (within a constant factor of the optimal set) from seed DNA codes of shorter length. The size of the codes is, however, sensitively dependent on the quality and size of the seed codes. The search for such codes is thus not only of theoretical importance, but also of practical significance since they are likely to be required for the full realization of self-assembly applications in practical settings. The gigantic size of the search spaces involved even for small lengths (exp(2,4^16) for 16-mers) is a primary road block for progress in this area, especially when nearly optimal performance is expected or desirable.

In this paper, we describe results of a tour de force to conduct an exhaustive search to produce code sets that are large enough to be nearly maximal sets while guaranteeing high quality. The quality criteria and a summary of prior methods are described in Section 2. Section 3 shows how to solve the problem of actually producing the actual composition of the sets, unlike its counterpart in vitro. Consequently, we also present a quantitative characterization of these codes through an analysis of occurrence of shorter k-mer blocks. In Section 4 we summarize some of the implications of these analyses for the application of these code sets. Finally, we also discuss some of the remaining problems to these sets and mention two interesting problems for further research. We assume that the reader is familiar with basic facts about molecular biology; see (Watson et al., 2003; Wetmur, 1997) for background details. Some of the results presented here have been announced in preliminary or incomplete form in (Bobba et al., 2006) and (Garzon et al., 2006). Full results and complete analyses are presented here.

THE SEARCH FOR DNA CODES

The smart part in the “smart glue” property of DNA is hybridization affinity (Watson et al., 2003). The fundamental difficulty with hybridization affinity is that it is an uncertain property determined not only by the specific sequence of oligonucleotides making up the strands, but also