INTRODUCTION

The closing decades of the second millennium have brought forth a cascade of technology and knowledge that promises to bring forth far greater changes than have been brought at any other time in history. Biological knowledge and biotechnology have perhaps lagged other cornerstones of change such as the microchip revolution, but may in the long run have the greatest impact on the evolution of human life and society.

Biological study is increasingly a study of complex systems. It has always been, and it is becoming more so. Our capacity as scientists to analyze and understand complex systems is rapidly maturing. An organism, such as a human being, is a complex system itself composed of multiple complex systems. One of these systems is the genome. It is a complex task to understand the genome, and such understanding is but a small step towards understanding the organism.

However, progress is made of small steps. In this dissertation, I will discuss a few small steps towards understanding the genome. Scientific progress is often driven by a marriage of new technologies with a cutting edge problem. I will present and discuss two strategic technologies for genome study. I will then apply genomic knowledge to the analysis of vertebrate trypsinogen evolution, which is itself a model for the complexity of genome evolution.

The technologies for genome study I present here are *strategies* for genomic analysis. The emphasis underscores a worldwide change in the paradigm for genetic experiments. To date, most genetics have been done from a “bottom-up” perspective. Isolated problems in genetics were selected and analyzed in detail. Global genomic information was not sought. Any global information obtained was added piecemeal and without coordination to the body of scientific knowledge.

Now, increasing numbers of genomes are being sequenced in their entirety, with several large genome sequencing projects underway (Rowen et al., 1997). Emphasis is on a “top-down” perspective, with global information sought primarily.[1] Intent is that content not only be applied to the study of isolated systems but also to systems that arise from complex interactions between multiple genes and proteins (e.g., DeRisi et al., 1997). An expectation is that a comprehensive global effort will not only be more thorough, but will also be more efficient than multiple “bottom-up” approaches.

Therefore, efficient strategies for genome sequencing are needed. The scale of genome projects is such that even minor improvements in strategy can effect major changes in cost, effort, and even feasibility.

In Chapter 1, I focus attention on random subcloning, which is a simple strategy for genome analysis. Until now, no adequate mathematical analysis of random subcloning has been available. It has been impossible to predict, other than empirically, project outcomes or costs. It has been impossible to compare complex strategies with basic standards. Here, I provide some fundamentals for addressing these issues.

In Chapter 2, I present pairwise end sequencing as an example of a more complex genome mapping and sequencing strategy. A robust mathematical analysis of pairwise end sequencing continues to elude researchers, so I analyze the strategy with the aid of computer simulations. Such simulations are a powerful way to answer the practical questions of project design and evaluation, even in the absence of a complete mathematical analysis.[2]

In Chapter 3, I depart from the purely theoretical themes of the first two chapters and address the evolution of the genome. I focus on the evolution of the trypsinogens, which are present in vertebrate genomes as a multicopy family tightly linked to the T-cell receptor β locus. These genes display a panoply of evolutionary modalities, including striking examples of coincidental evolution. Coincidental evolution is the tendency of genes present in the same

---

[1] Searching for a needle in a haystack is analogy for a “bottom-up” approach to genome analysis. The needle represents a sought-after gene such as that for Huntington’s Disease (i.e., HD) or breast cancer (e.g., BRCA1). The haystack is the genome. The search for each of these genes can be viewed as having been an independent search through the same haystack for different needles. A “top-down” approach to haystack analysis would be to sort through the entire haystack, one straw at a time, categorizing everything found in the haystack. Such an approach might be inefficient if just one needle is sought, but would become more efficient the more needles there were in the haystack. Identifying the approximately 75,000 genes in the human genome would require 75,000 independent bottom-up searches. Alternatively, one comprehensive search could be conducted. Providing this comprehensive search is one goal of the Human Genome Project. An excellent summary of the present state and future potential of genomic study has been provided by McKusick (1997). Estimates of the number of genes in the human genome have been tabulated by Fields et al. (1994).

[2] For background on the power, utility, and limitations of computer simulations, the reader may wish to consider Galper and Brutlag (1993). The need for computer simulations in the Human Genome Project is highlighted by Koonin (1998).

genome to evolve in a covariant manner. I also point out several difficulties that complicate evolutionary analysis of multigene families.

Strategies for genomic analysis are often analyzed with simple assumptions about the nature of multigene families and other repeated elements. These assumptions are often used to call into question the utility of proposed genomic strategies. Understanding the nature of repeats, including multigene families, and how they are formed is thus an important adjuvant to a discussion of structural genomics. With an understanding of the nature of evolution, we can begin to predict how similar repeats are likely to be, and this prediction, in turn, affects parameterizations of strategies and overall cost.

From a functional genomics point of view, the nature of repeat families tells us much not only about evolution, but also of the nature of complex biological systems interaction. Complex modern vertebrates are positioned at a pinnacle of billions of years of biological evolution on Earth.[3] An important enabling feature of this evolution has been the ability of the evolutionary process to reuse and readapt previously constructed building blocks (Henikoff et al., 1997). Since the dawn of vertebrate evolution 600 million years ago, it may be that very few truly new genes have evolved (Holland and Garcia-Fernàndez, 1996). Rather, nature has adapted previously existing genes, sometimes in novel combinations, to new functions. Original genes are usually left untouched, with novel adaptation operating on duplicate copies of original gene family members. The resulting multiple isoforms of genes permit the establishment of complex systems with multiple similar but subtly different components. Evolutionary operations on gene families are likely to have been a major mode of evolution in the vertebrate subphylum.

The serine proteases, of which trypsin is a member, are one of the largest and most diverse gene superfamilies (Barrett and Rawlings, 1993). Their origins lie buried in the earliest stages of evolution – trypsin genes are present in eubacterial genomes (Rypniewski et al., 1994). Modern vertebrate serine proteases include chymotrypsin, elastase, the recently-evolved prostate specific antigen, blood clotting enzymes, several granzymes, and many other genes involved in immunological defense. The serine protease gene superfamily rivals the immunoglobulin gene superfamily in terms of its importance to both vertebrate evolution and

[3] All currently living species are positioned on their own respective “pinnacles of evolution.” The height of each pinnacle is the amount of evolution that has occurred to a species since the origin of life. If evolution is measured by time, all current pinnacles are equally high. The pinnacles that were reached by extinct species are lower. An introduction to evolutionary concepts can be found in Dawkins (1990 and 1996).

vertebrate immune defense (Smyth et al., 1996; Hunkapiller et al., 1989). It is fascinating to ponder the close genetic linkage of the trypsin and T-cell receptor β gene loci.

The vertebrates almost certainly possess the most complex immune system of all living organisms. This has been made possible by the repeated use of duplications in multicopy gene families (see, for example, Ohno, 1978; Hunkapiller et al., 1989; Hood and Hunkapiller, 1991; Raport et al., 1996). These families include the immunoglobulins the serine proteases, and many others. There are, in fact, few genes of importance to immunology that are not members of multigene families. To study multigene families is to study the fabric of immunological complexity.