Chapter 7
Applications of Supercomputers in Population Genetics

Gerard G. Dumancas
Oklahoma Medical Research Foundation, USA

ABSTRACT

Population genetics is the study of the frequency and interaction of alleles and genes in population and how this allele frequency distribution changes over time as a result of evolutionary processes such as natural selection, genetic drift, and mutation. This field has become essential in the foundation of modern evolutionary synthesis. Traditionally regarded as a highly mathematical discipline, its modern approach comprises more than the theoretical, lab, and fieldwork. Supercomputers play a critical role in the success of this field and are discussed in this chapter.

INTRODUCTION

The general goals of population genetic studies are to characterize the extent of genetic variation within species and account for this variation (Weir, 1996). The amount of genetic variation can be determined by the frequency of genes and the forces that affect such frequencies such as mutation, migration, selection, and genetic drift (Gall, 1987). Throughout the years, population genetics has evolved to develop theoretical models to explain changes of allele and genotype frequencies in natural populations of different organisms through time. For example, with these built models, it is possible to determine the length of time for a given allele to be fixed given a certain selective force for it. With the determination of the genetic variations within individuals, it is possible to find the genetic causes that give rise to and maintain variation both within species and between species. For example, where do the genes of the Europeans come from? The main way to gain insight into past population processes is to analyze and interpret current patterns of genetic variation (Sokal, 1991) (von Haeseler et al., 1996).

Over the past decades, the field of population genetics has undergone remarkable changes. This has been due to the development of sophisticated DNA sequencing technologies, which makes it possible to generate large quantities of the most direct kind of genetic data easy and affordable (Wakeley, 2004). Current and future challenges in this field in both computational methodology and in analytical theory are to develop models

DOI: 10.4018/978-1-4666-7461-5.ch007

and techniques to extract the most information possible from multilocus DNA datasets. To solve this problem, technical improvements, such as the use of robotics would streamline the gathering of relatively large genetic datasets (Wakeley, 2004). Further, the use of supercomputers has tremendously improved the computational methodology in this field. Thus, the primary objective of this chapter is to discuss the applications of supercomputers in population genetics.

BACKGROUND

The beginnings of genetics and of population genetics are one. Both started with Mendel; and were unrecognized until the rediscovery in 1900. That is in his simple genetics experiment wherein he considered the consequences of repeated self-fertilization; he showed that heterozygosity is reduced by half each generation and gave formulas for genotype frequencies in successive generations. This consequently led to the foundation of population genetics. The "golden age" of population genetics was the period when Haldane, Fisher, and Wright were producing their great work. That is, they reconciled biometry with genetics, quantified the approach to evolution, and created a totally new science. Thus, it is arguably the most successful mathematical theory in biology (Crow, 1987).

Essential to the understanding of population genetics as a specialized field, different techniques currently used or used in the recent past by various scientists are necessary to generate data for population genetic models. Protein electrophoresis, restriction endonucleases (restriction fragment length polymorphisms of purified DNA), and DNA polymorphisms (randomly amplified polymorphic DNAs, DNA resequencing) are some of the most commonly used methods for generating population genetic data (Templeton, 2006). In protein electrophoresis, proteins are separated on the basis of their physical properties such as net charge, size, and shape. In this procedure, the separated proteins are stained to reveal specific classes of enzymes. Genetic diversity is identified when the stained isozyme bands show migration at different rates (Conkle, 1972). Restriction endonucleases procedure involve the use of restriction enzymes that break DNA bonds for splicing different DNA together (Sherlock et al., 2002) while DNA polymorphisms experiments involve the use of advance DNA sequencing technologies to identify DNA variations between individuals within populations. Basically these experiments are all performed with the aim of assaying and identifying genetic variations within individuals and within populations. For example, Eanes and Koehn studied the genetics of different Monarch butterfly populations in the early 1970s using electrophoresis that examines the same proteins in different individual butterflies. They found that Monarchs have allele frequencies that sort out into groups somewhat in the summer and become uniform again during migration. These results showed that Monarchs divide into slightly isolated populations during the summer but mix together during migration (Eanes et al., 1978). With the advent of these sophisticated techniques as well as the emergence of new, polymorphic, genetic markers, a number of questions regarding populations and the interactions of the individuals within them may be addressed using genetic data.

Genetic variation simply refers to the variation in alleles of genes, occurring both within and among populations. In other words, it describes naturally occurring differences among individuals of the same species. This variation allows flexibility and survival of a population in the face of changing environmental factors (2013). Genetic diversity at the population level of a species plays an important role in the interaction of the species with the environment. These interactions will consequently structure the ecosystem, so that the spatial and temporal partitioning of genetic diversity will occur (Medline et al., 2000). Further, elucidating the inherited basis of genetic variation in human health and disease is considered to be