I will start with some examples from evolutionary biology to demonstrate how distributions of various words in DNA sequences can be used to learn important facts in evolutionary biology. Many of these problems related to evolutionary biology lead to unsupervised or partially supervised learning problems involving high dimensional data. Effective statistical analysis requires appropriate reduction of dimensionality and involves an interesting variable selection problem. This motivated developing certain probabilistic models for replication of DNA sequences which will be discussed.