CHARACTERIZATION AND USE OF STRUCTURE AND COMPLEXITY OF DNA SEQUENCES

by

Hasan H. Otu

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Interdepartmental Area of
Major: Engineering

(Electrical Engineering)

Under the Supervision of Professor Khalid Sayood

Lincoln, Nebraska

August, 2002

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

DISSERTATION TITLE

Characterization and Use of Structure and Complexity of DNA Sequences

BY

Hasan H. Otu

SUPERVISORY COMMITTEE:

	APPROVED	DATE
Signature	Khalid Sayood	8/19/02
Typed Name	Khalid Sayood	
Signature	Sharad Seth	8/19/02
Typed Name	Sharad Seth	
Signature	M. W. Hoffman	8/19/02
Typed Name	Michael W. Hoffman	
Signature	Lance C. Perez	19 August 2002
Typed Name	Lance C. Perez	
Signature		
Typed Name		
Signature		
Typed Name		

GRADUATE COLLEGE
UNIVERSITY OF NEBRASKA

CHARACTERIZATION AND USE OF STRUCTURE AND COMPLEXITY OF DNA SEQUENCES

Hasan H. Otu, Ph.D.

University of Nebraska, 2002

Adviser: Khalid Sayood

In this dissertation we analyze biological sequences using two proposed methods of characterization. The first method uses the Average Mutual Information (AMI) profile of the sequences. This captures the statistical properties of the strings and provides a concise representation. The second method utilizes the notion of "complexity." Using the Lempel-Ziv (LZ) complexity measure we define a distance metric for sequences.

We use AMI profiles to solve the fragment assembly problem which is to reconstruct a target DNA sequence from randomly sampled fragments. Most existing fragment assembly techniques follow the overlap – layout – consensus approach, which requires extensive computation in each phase and becomes inefficient with increasing numbers of fragments. We propose a new algorithm which jointly solves the overlap, layout, and consensus problems. The fragments are clustered with respect to their AMI profiles using the k-means algorithm. This removes the unnecessary requirement that the collection of fragments be considered as a whole. Instead, the orientation and overlap detection are solved efficiently, within the clusters.

We apply the second method of characterization to phylogeny construction. Most existing approaches for phylogenetic inference use multiple alignment of sequences and assume some sort of an evolutionary model. The multiple alignment strategy does not work for all types of data, e.g. whole genome phylogeny, and the evolutionary

models may not always be correct. We propose a new sequence distance measure based on the relative information between the sequences using LZ complexity. The distance matrix thus obtained can be used to construct phylogenetic trees. The proposed approach does not require sequence alignment and is totally automatic.

The proposed methods are not limited to the applications studied in this dissertation. They capture universal properties of the sequences and can be used to tackle other problems posed by computational biology.

ACKNOWLEDGEMENTS

Thank God, it is finally finished! I would like to thank my advisor, Dr. Khalid Sayood, for his patience and guidance throughout this long process. I was lucky to have found an advisor and colleague who always listens to my problems and provides me with solid advice. His technical and editorial suggestions were essential to the completion of this dissertation and have taught me invaluable lessons and insights on the course of performing research. He gave me room to explore and make mistakes, always visualized important long-term goals and never got lost or distracted by bureaucratic details. He was always open to suggestions and very flexible in dealing with roadblocks that unavoidably occurred in completing my graduate studies.

I am fortunate to have the opportunity to meet and work with Dr. Michael Hoffman, who was also a member of my committee. I appreciate his energy, wisdom, encouragement, and advice that helped to finish this dissertation. I also thank my other committee members Dr. Sharad Seth and Dr. Lance Perez for reading previous drafts of this dissertation and providing many valuable comments that improved the contents of this dissertation.

I would like to thank the department for the financial support and all members of the department for contributing to such an inspiring atmosphere. My special thanks go to James Nau for his IT support, and our department secretaries Pat Masek and Jane Craig for all their help.

My thanks also go to my parents who made the trip all the way from Turkey to help me with the final phases of this dissertation. I thank my American parents, Joey and Jerry Vernon, for their love and caring. I also would like to acknowledge the support of my brothers, sisters, nephew, and nieces for always being there for me whenever I needed their help and support. Last, but not least, I would like to thank my dearest friend Dr. Handan Kaplan. Her constant support, patience, love,

and encouragement, especially during the past two years, have become my primary source for help. I benefited from many fruitful discussions about all things in life during the time we have spent together.

Finally, if this dissertation inspires a researcher in any positive way, the purpose of its existence will be fulfilled.

Contents

List of Tables

List of Figures

Glossary

ALIGNMENT of two sequences, obtained by inserting spaces to both of the sequences so that the length of the sequences are equal and no two spaces occur at the same position in the sequences. Each column of an alignment pairs up either the corresponding symbols in the sequences or a space with a symbol. Depending on the symbols that span a column in the alignment, the column can be composed of a match, a mismatch, or a gap.

ALPHA-HELIX A particular helical folding of the polypeptide backbone in protein molecules. Alpha-helices comprise a regular local theme in the secondary structure of proteins.

AMINO ACID A group of organic molecules that bond with each other to build proteins.

ANNOTATION Noting biologically significant features on DNA sequences.

BASE See *nucleotide.*

BETA-SHEET A structure of proteins where the peptide is extended and stabilized by hydrogen bonding. Beta-sheets comprise a regular local theme in the secondary structure of proteins.

CHROMOSOME A discrete strand of DNA which is a part of the cell's genome. Occurs in homologous pairs in diploid organisms.

CLADE Families or subfamilies of organisms that descend from a common ancestor in a phylogenetic tree.

CODON Three bases of DNA or RNA that code for a single amino acid or signals the start or end of protein synthesis.

COMPLEMENTARY STRANDS The two strands of the DNA double helix. Due to specificity in base pairing and the direction of sequencing, the base composition of one strand is the reverse complement of the base composition of the other strand.

CONSENSUS SEQUENCE A description of the multiple alignment of sequences. The i^{th} position in the consensus is the most frequently occurring symbol in the i^{th} column of the multiple alignment.

CpG ISLAND The region on a DNA sequence that is rich in the dinucleotide CG. Should not be confused with the regions of the DNA that are rich both in G and C.

DNA Deoxyribonucleic acid. A double-stranded molecule twisted into a double helix that encodes genetic information. Held together by the bonds between base pairs of nucleotides. There are four bases in DNA: adenine (A), guanine (G), cytosine (C), and thymine (T), where an A always bonds with a T and a G always bonds with a C.

EST Expressed Sequence Tag. Is an STS that occurs on the coding parts of DNA. Used to identify the coding sequences on DNA by its similarity to a known EST.

EUKARYOTES Organisms with cells that have a nuclear envelope. Animals. plants, and fungi, all belong to this group.

EXON Non-overlapping continuous parts on the sequence of the protein coding gene. Exons are concatenated to form the mRNA that gets translated into a protein.

FRAGMENT ASSEMBLY Reconstructing a target DNA from a collection of fragments that are randomly sampled from both strands of the target DNA.

GENE A region of DNA that codes for a polypeptide chain or specifies an RNA molecule. An inheritable trait which in turn has an influence on some characteristic of the organism.

GENE MAPPING Finding the relative positions of genes on a DNA sequence.

GENE STRUCTURE The mapping of genes on a genome and the mapping of exons and introns on a gene.

GENOME The ensemble of DNA in a cell. Also referred to as the collection of genes in an organism.

HISTONE Any of various simple water-soluble proteins that are rich in the basic amino acids lysine and arginine. They bind to about 200 base pairs of DNA to form the repeating structure of chromatin.

HOMOLOGY Denotes the similarity between molecular sequences or the similarity of structure or function of proteins .Usually hints at a common evolutionary origin.

INTRON Non-overlapping continuous parts on the sequence of the protein coding gene. Introns are not transcribed to mRNA that gets translated into a protein.

INVERTEBRATES Animals that have no spinal column.

MUTATION A usually small change on a sequence due to insertion, deletion, and substitution of its symbols.

mRNA Messenger RNA. Transcribed from DNA by copying the protein coding regions. mRNA is single stranded and acts as a template for protein synthesis.

MULTIPLE ALIGNMENT Alignment of more than two sequences. Spaces are inserted in all of the sequences such that their lengths are the same and no given position consists of spaces only.

NUCLEOTIDE Basic building blocks of nucleic acids.

PHYLOGENETIC TREE Also known as evolutionary tree. A tree representing the evolutionary relationship between species. Shows the divergence of species from a common ancestor where the ancestor is represented at the base of a central stem in the tree.

POLYTENE CHROMOSOMES Giant chromosomes produced by the successive replication of homologous pairs of chromosomes. Polytene chromosomes join together without chromosome separation or nuclear division. The best known polytene chromosomes are those of the salivary gland of the larvae *Drosophila melanogaster*.

PROKARYOTES Organisms with cells that do not have a nuclear envelope. All bacteria and archea belong to this group.

PROTEIN A combination of amino acids in peptide linkages. Proteins control almost all biological and chemical processes in the cell. Enzymes, hormones,

etc. are all proteins. Twenty different amino acids are commonly found in proteins.

PURINE A class of heterocyclic organic base found in nucleic acids adenine and guanine. Also found in other biological compounds such as sugar derivatives.

PYRIMIDINE A class of heterocyclic organic base found in nucleic acids cytosine, uracil, and thymine. Also found in other biological compounds such as sugar derivatives.

RETROTRANSPOSON Transposons that jump to a new location via an RNA intermediate.

RIBOSOME A small particulate organelle that translates the mRNA sequence into a protein.

RNA Ribonucleic acid. RNA is found in all living organisms and carry the information from DNA to form proteins. There are four bases in RNA: adenine (A), guanine (G), cytosine (C), and uracil (U).

rRNA Ribosomal RNA. Functions as a structural element in ribosomes.

STS Sequence Tagged Sites. Unique substrings in a DNA strand. Used as landmarks in genome mapping.

TANDEM REPEAT Multiple copies of a DNA subsequence occurring one after another; e.g. ACTACTACTACT.

TRANSPOSON A piece of DNA that can cut itself out of the genome and reinsert at another position. Results in interspersed repeats in the DNA sequence.

TRANSCRIPTION of DNA, being copied to RNA.

TRANSITION Change of a purine base (A or G) into a purine base or change of a pyrimidine base (C or T) into a pyrimidine base.

TRANSLATION of mRNA, being converted into a protein.

TRANSVERSION Change of a purine base (A or G) into a pyrimidine base (C or T) and vice versa.

VERTEBRATES All animals that have a backbone composed of bony vertebrae. The subdivisions or classes of Vertebrates are Mammalia, Aves, Reptilia, Amphibia. Pisces, Marsipobranchia, and Leptocardia.

Chapter 1

Introduction

Biology offers many open and complex questions. As the techniques used in biology improve, most of these questions demand the cooperation of different sciences such as biology, physics, engineering, computer science and chemistry. This cooperation presents itself in understanding the two macromolecules: deoxyribonucleic acid (DNA) and proteins.

1.1 DNA and protein

DNA is a very long, threadlike macromolecule made up of a large number of deoxyribonucleotides each composed of a base, a sugar, and a phosphate group [146]. The bases of DNA carry genetic information, whereas their sugar and phosphate groups perform a structural role. The genetic information of all cells and many viruses are stored in DNA. Some viruses, however, use ribonucleic acid (RNA) as their genetic material. There are four different bases in DNA: *adenine* (A), *cytosine* (C), *guanine* (G) and *thymine* (T). The three-dimensional structure of DNA, discovered in 1953, is a double helix. The most important aspect of the DNA double helix is the specificity of the pairing of bases in which A must pair with T and C must pair with G.

The ensemble of DNA in a cell is called *genome*. The entire genome of an organism can be contained on a single *chromosome* (the entire double helix) as in *E. coli* or on 46 chromosomes as in humans. A striking characteristic of DNA molecules is their length. For example the size of the DNA genome is 48.6 kbp (kilo base pairs) in λ *phage* (a virus) 4,000 kbp in *E. coli* and 2,900.000 kbp in human.

Proteins (derived from the Greek word *proteios*, which means "of the first rank") play crucial roles in virtually all biological processes. The basic structural units of proteins are called amino acids. The amino acid sequence of a protein is specified by messenger RNA (mRNA), which is translated into protein in a process catalyzed by ribosomes. Simply put, mRNAs are the transcribed segments (protein-coding genes) of DNA. This information is later translated into proteins by combinations of three nucleotides (*codons*) coding for amino acids. Even though there are 64 possible combinations, only 20 amino acids exist in nature. One important theme to be considered is that protein-coding genes in bacteria differ from those in eukaryotes. In eukaryotes, the protein coding genes consist of coding (exons) and non-coding (introns) parts. In bacteria, protein coding genes do not contain introns and may be arranged consecutively to form a unit of gene expression (operon). Still, not all of bacterial DNA is used in protein coding; between any two genes lies the so called intergenic region.

1.2 Challenges in Computational Biology

Owing to the advances in sequencing of genetic information, biology has become an increasingly information-intensive discipline, and a relatively new interdisciplinary science called *computational biology*, or *bioinformatics* has come into being. A systematic definition of bioinformatics is given in [105]: