



Mathematics of Bioinformatics

---Theory, Practice, and Applications (Part I)

Matthew He, Ph.D.

Professor/Director

Division of Math, Science, and Technology

Nova Southeastern University, Florida, USA

December 18-21, 2010, Hong Kong, China

BIBM 2010



OUTLINE

- ❖ INTRODUCTION: FUNDAMENTAL QUESTIONS
- ❖ PART I: GENETIC CODES, BIOLOGICAL SEQUENCES, DNA AND PROTEIN STRUCTURES
- ❖ PART II: BIOLOGICAL FUNCTIONS, NETWORKS, SYSTEMS BIOLOGY AND COGNITIVE INFORMATICS



TABLE OF TOPICS: PART I

I. Bioinformatics and Mathematics

- 1.1 Introduction
- 1.2 Genetic Code and Mathematics
- 1.3 Mathematical Background
- 1.4 Converting Data to Knowledge
- 1.5 Big Picture: Informatics
- 1.6 Challenges and Perspectives

II. Genetic Codes, Matrices, and Symmetrical Techniques

- 2.1 Introduction
- 2.2 Matrix Theory and Symmetry Preliminaries
- 2.3 Genetic Codes and Matrices
- 2.4 Challenges and Perspectives

III. Biological Sequences, Sequence Alignment, and Statistics

- 3.1 Introduction
- 3.2 Mathematical Sequences
- 3.3 Sequence Alignment
- 3.4 Sequence Analysis/Further Discussions
- 3.5 Challenges and Perspectives



TABLE OF TOPICS: PART I

IV. Structures of DNA and Knot Theory

- 4.1 Introduction**
- 4.2 Knot Theory Preliminaries**
- 4.3 DNA Knots and Links**
- 4.4 Challenges and Perspectives**

V. Protein Structures, Geometry, and Topology

- 5.1 Introduction**
- 5.2 Computational Geometry and Topology**
- 5.3 Protein Structures and Prediction**
- 5.4 Statistical Approach and Discussions**
- 5.5 Challenges and Perspectives**



TABLE OF TOPICS: PART II

VI. Biological Networks and Graph Theory

- 6.1 Introduction
- 6.2 Graph Theory and Network Topology
- 6.3 Models of Biological Networks
- 6.4 Challenges and Perspectives

VII. Biological Systems, Fractals, and Systems Biology

- 7.1 Introduction
- 7.2 Fractal Geometry Preliminaries
- 7.3 Fractal Geometry in Biological Systems
- 7.4 Systems Biology and Perspectives
- 7.5 Challenges and Perspectives

VIII. Matrix Genetics, Hadamard Matrix, and Algebraic Biology

- 8.1 Introduction
- 8.2 Degeneracy of the Genetic Code
- 8.3 The Genetic Code and Hadamard Matrices
- 8.4 Genetic Yin-Yang Algebras
- 8.5 Challenges and Perspectives



TABLE OF TOPICS: PART II

IX. Bioinformatics, Living Systems and Cognitive Informatics

- 9.1 Introduction
- 9.2 Emerging Pattern, Dissipative Structure, and Evolving Cognition
- 9.3 Denotational Mathematics and Cognitive Computing
- 9.4 Challenges and Perspectives

X. The Evolutionary Trends and Central Dogma of Informatics

- 10.1 Introduction
- 10.2 Evolutionary Trends of Information Sciences
- 10.3 Central Dogma of Informatics
- 10.4 Challenges and Perspectives

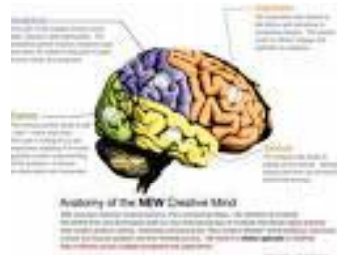
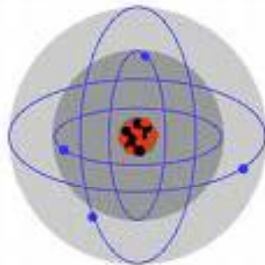
INTRODUCTION: FUNDAMENTAL QUESTIONS

What is matter? → Physical Sciences

What is life? → Biological Sciences

What is mind? → New Science of Mind

What is information? → Informatics





WORDS ON BIOLOGY AND MATHEMATICS...

- *There's millions and millions of unsolved problems. Biology is so digital, and incredibly complicated, but incredibly useful. Biology easily has 500 years of exciting problems to work on, it's at that level.*

Don Knuth

- *Where the telescope ends, the microscope begins. Which of the two has the grander view?*

Victor Hugo

- *Mathematics if Biology's Next Microscope, only better; Biology is Mathematics' Next Physics, Only Better.*

Joel Cohen



FROM GENETIC CODE TO LIFE

- Life is founded on **mathematical pattern** of the physical world. Genetics exploits and organized these patterns. **Mathematical regularities** are exploited by the organic world at every level of **form, structure, pattern, behavior, interaction, and evolution**. (Ian Stewart, Life's other secret)
- The **Natural Technology** of genetic coding is major and most effective technology ensuring life on our planet. And acquirement of this technology, occurring in modern time, is major movement in evolution of mankind. The biological evolution can be interpreted as process of deployment and duplicating of the certain forms of **ORDERING**. (Sergey Petokhov, The Biperiodic table of genetic code and number of protons)

CENTRAL DOGMA OF MOLECULAR BIOLOGY

- Genetics

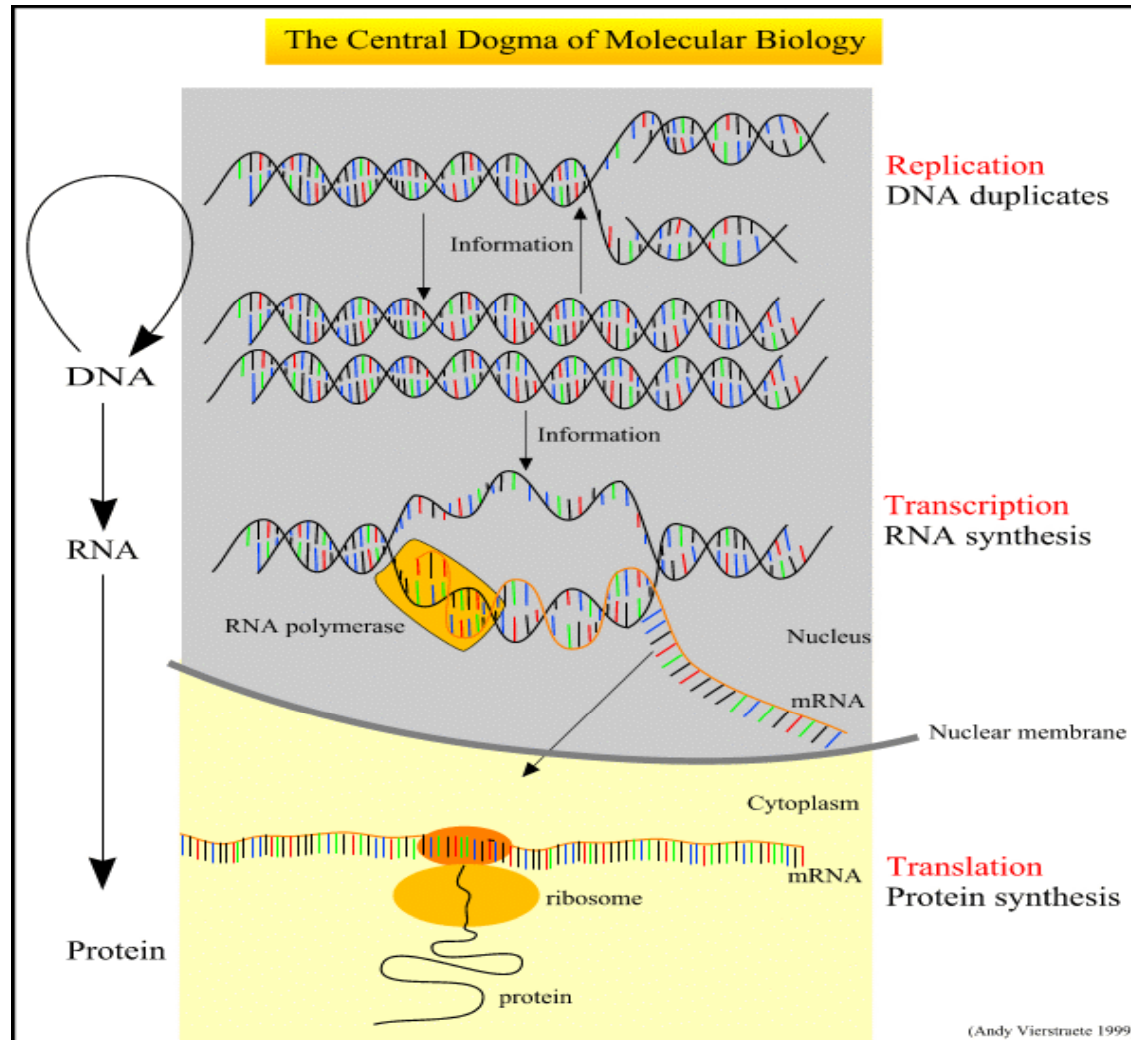
- ☐ DNA: A, C, G & T
- ☐ (RNA): A, C, G & U

- Codons (Triplets)

- ☐ A string containing three of the above characters
- ☐ Ex. AUG, ACU, GAC, UAA/ UAG /UGA ...



CENTRAL DOGMA OF MOLECULAR BIOLOGY





INFORMATION BUILDING BLOCKS

Monomer	Polymer
Amino Acids: <ul style="list-style-type: none">• Cysteine (Cys)• Alanine (Ala)• Proline (Pro)	Protein: Met-Cys-Gly-Pro-Pro-Arg...
Nucleotides: <ul style="list-style-type: none">• Adenine (A)• Cytosine (C)• Thymine (T)	DNA: ACTGGTAGCCTTAGA...
Letters: A, B, C...	Words: CAT, GO, FRIEND...
Symbols: 0, 1	Binary Code: 1001011100101...
Monomial: 1, x, x²,...	Polynomial: P(x),...



THE LANDSCAPES OF BIOLOGICAL SCIENCES

Six Fundamental Questions

How is it Built?	How does it work?	What goes wrong?
How is it fixed?	How it begin?	What is it for?

Nine Key Domains of Biological Sciences

Molecules	Cells	Tissues
Organs	Individuals	Population
Communities	Ecosystems	Biosphere

Two Important Dimensions

Time scales	Photosynthesis	B. Years of Evolution
Spatial scales	Molecular	Cosmic



THE LANDSCAPES OF APPLIED MATHEMATICS

Domains	Functions/Purpose
Data structures	Ways to organize data
Algorithms	Procedures for manipulating symbols
Theories	Used to Analyze both data and ideas
Models	Used to Analyze both data and ideas
Computers/Software:	Implementation and computation



THE LANDSCAPES OF RESEARCH IN BIOLOGY AND MATHEMATICS

Combinations of

One or more Biological questions, domains, time scales and spatial scales

With

One or more data structures, algorithms, theories or models, and means of computation



MATHEMATICS OF BIOINFORMATICS

9 KEY DOMAINS

Genetic Matrices	Biological Sequences	DNA Structures
Protein Structures	Biological Networks	Systems Biology
Algebraic Biology	Cognitive Informatics	Universal Evolution



MATHEMATICS, COMPUTER SCIENCE, AND BIOLOGY

- $\int \int \int [M(\text{xyz}) + C(0\&1) + B(\text{dna})] d(\text{Info})$
- Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications (*Sergey Petoukhov and Matthew He, IGI Global, 2009*)
- Mathematics of Bioinformatics: Theory, Practice, and Applications (*Matthew He and Sergey Petoukhov, in press, Wiley-Interscience, 2011*)



Part I Genetic Codes, Biological Sequences, DNA and Protein Structures

1. Bioinformatics and Mathematics

- ❖ Introduction
- ❖ Genetic Code and Mathematics
- ❖ Mathematical Background
- ❖ Converting Data to Knowledge
- ❖ Big Picture: Informatics
- ❖ Challenges and Perspectives



1.1 INTRODUCTION

Mathematics and biological data have a synergistic relationship:

- Biological information creates interesting problems.
- Mathematical theory and methods provides models to understand them.
- Biology validates the mathematical models.

A model is a representation of a real system.

- Real systems are too complicated, and observation may change the real system.
- A good system model should be simple, yet powerful enough to capture the behaviors of the real system.
- Models are especially useful in bioinformatics.



Historical Background

- **Mendel's genetic experiments and laws of heredity:** The discovery of genetic inheritance by Gregory Mendel back in 1865 was considered as the start of bioinformatics history.
 - **The Law of Segregation**
 - **The Law of Independent Assortment**
 - **The Law of Dominance**
- **Origin of species:** Charles Darwin published “On the Origin of Species” by Means of Natural Selection (Darwin, 1859) or The Preservation of Favored Races in the Struggle for Life" in 1895.
- **First genetic map:** In 1910, after the rediscovery of Mendel's work, Thomas Hunt Morgan did crossing experiments with the fruit fly (*Drosophila Melanogaster*) at Columbia University. He proved that the genes responsible for the appearance of a specific phenotype were located on chromosomes.



Historical Background

- **Transposable genetic elements:** In 1944 Barbara McClintock discovered that genes can move on a chromosome. Genes can jump from one chromosome to another.
- **DNA double helix:** In 1953, James Watson and Francis Crick proposed a double helix model of DNA. They suggested that genetic information flows only in one direction, from DNA to messenger RNA to protein, the central concept of the central dogma.
- **Genetic code:** The genetic code was finally "cracked" in 1966. Marshall Nirenberg, Heinrich Mathaei and Severo Ochoa demonstrated that a sequence of three nucleotide bases, a codon or triplet, determines each of the 20 amino acids found in nature.



Historical Background

- **First recombinant DNA molecules:** In 1972, Paul Berg of Stanford University (USA) created the first recombinant DNA molecules by combining the DNA of two different organisms.
- **DNA sequencing and database:** In early 1974, Frederick Sanger from the U.K. Medical Research Council was first to invent DNA sequencing techniques. During his experiments to uncover the amino acids in bovine insulin, he developed the basics of modern sequencing methods.
- **Human Genome Project:** In 1990, the U.S. Human Genome Project started as a 15-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the expected completion date to 2003.



Historical Background

HG Project goals were to:

- identify all the genes in human DNA,
 - determine the sequences of the 3 billion chemical base pairs that make up human DNA, store this information in databases,
 - improve tools for data analysis,
 - transfer related technologies to the private sector, and
 - address the ethical, legal, and social issues that may arise from the project.
-
- The draft human genome sequence was published on February 15th 2001, in the journals Nature and Science.



1.2 GENETIC CODE AND MATHEMATICS

The secrets of life are more complex than DNA and the genetic code:

- One secret of life is the self-assembly of the first cell with a genetic blueprint that allowed it to grow and divide.
- Another secret of life may be the mathematical control of life as we know it and the logical organization of the genetic code and the use of math in understanding life.
- All knowledge is intrinsically unified and lies in a small number of natural laws.



1.2 GENETIC CODE AND MATHEMATICS

- Math can be used to understand life from the molecular to the biosphere level
 - the origin and evolution of organisms,
 - the nature of the genomic blueprints
 - the universal genetic code
 - ecological relationships.
- Math helps us look for trends, patterns and relationships that may or may not be obvious to scientists.
- Math allows us to describe the dimensions of genes, sizes of organelles, cells, organs and whole organisms.



1.3 MATHEMATICAL BACKGROUND

- **ALGEBRA:** Algebra is the study of structure, relation and quantity through symbolic operations for the systematic solution of equations and inequalities. In addition to working directly with numbers, algebra works with symbols, variables, and set elements.
- **ABSTRACT ALGEBRA:** Abstract algebra extends the familiar concepts from basic algebra to more general concepts. Abstract algebra deals with the more general concept of *sets*: a collection of all objects selected by property, specific for the set under binary operations. Binary operations are the keystone of algebraic structures studied in abstract algebra: they form part of groups, rings, fields and more.



1.3 MATHEMATICAL BACKGROUND

- **PROBABILITY:** Probability is the language of uncertainty. It is the likelihood or chance that something is the case or will happen. Probability theory is used extensively in areas such as statistics, mathematics, science, philosophy, psychology, and in the financial markets to draw conclusions about the likelihood of potential events and the underlying mechanics of complex systems. An impossible event has a probability of 0, and a certain event has a probability of 1.
- **STATISTICS:** Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. Probability and statistics have been successfully used to investigate sequence analysis, alignments, profile searches and phylogenetic trees and many problems in bioinformatics.



1.3 MATHEMATICAL BACKGROUND

- **DIFFERENTIAL GEOMETRY:** Differential geometry is a mathematical discipline that uses the methods of differential and integral calculus to study problems in geometry. In biological and medical sciences, differential geometry has been used to study protein confirmation and elasticity of non-rigid objects such as human hearts and human faces.
- **TOPOLOGY:** Topology is the mathematical study of the properties that are preserved through deformations, twistings, and stretchings of objects. DNA topology and protein topology are active research areas.
- **KNOT THEORY:** Knot theory is the mathematical branch of topology that studies mathematical knots, which are defined as embeddings of a circle in 3-dimensional Euclidean space, \mathbf{R}^3 . Chemists and biologists use knot theory to understand, for example, chirality of molecules and the actions of enzymes on DNA.



1.3 MATHEMATICAL BACKGROUND

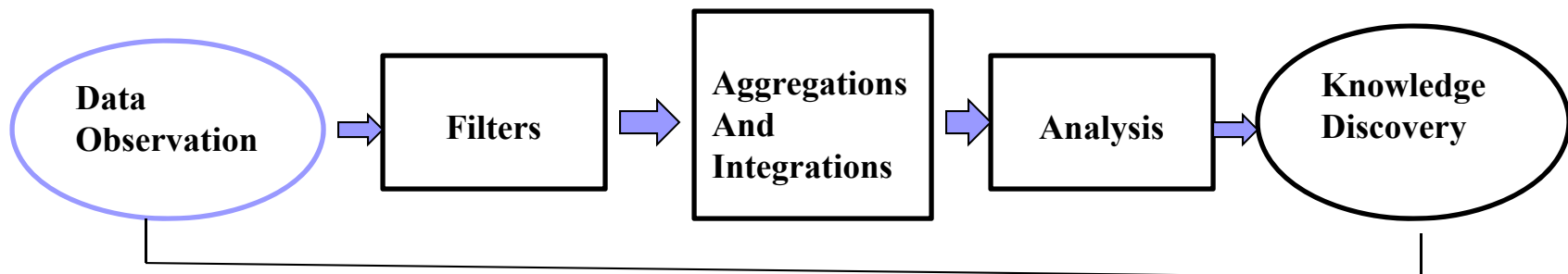
- **GRAPH THEORY:** Graph theory is the study of *graphs*. Graphs are mathematical structures used to model pairwise relations between objects from a certain collection. Many applications of graph theory exist in the form of network analysis.
- **FRACTALS:** A fractal is generally "a rough or fragmented geometric shape that can be split into parts, each of which is (at least approximately) a reduced-size copy of the whole," a property called self-similarity. Because they appear similar at all levels of magnification, fractals are often considered to be infinitely complex (in informal terms). Approximate fractals are easily found in nature. These objects display self-similar structure over an extended, but finite, scale range. Examples include clouds, snow flakes, crystals, mountain ranges, lightning, river networks, cauliflower or broccoli, and systems of blood vessels and pulmonary vessels.

1.4 CONVERTING DATA TO KNOWLEDGE

The biological information we gain allows us to learn

- About ourselves,
- About our origins,
- About our place in the world.

The process of converting data to knowledge:





1.5 BIG PICTURES: INFORMATICS

- Structure, behaviors, and interactions of natural and artificial computational systems
- Representation, processing, and communication of information in natural and artificial systems
- Computational, cognitive and social aspects
- The central notion is the transformation of information-whether by computation or communication, whether by organisms or artifacts.



COMPUTATIONAL SYSTEMS

- Natural
 - Artificial
 - Internal structure, behavior, and interaction with the environment.
 - Construct (or reconstruct) computational systems
 - Analytical, experimental and engineering methodologies
-
- The computer language systems and their interfaces with various data types are illustrated below.



COMPUTATIONAL SYSTEMS

Communications Between Computer Languages and Data Types

Computer Languages	Design Goals
FORTRAN	Numerical analysis
LISP	Symbolic computation
C	System programming
C++	Objects, speed, compatibility with C
Java	Objects, internet
Perl	System administration
Python	General programming



1.6 CHALLENGES AND PERSPECTIVES

- Integration: How do we incorporate variation among individual units in nonlinear systems and biological systems?
- Scaling: How do we explain the interactions among phenomena that occur on a wide range of scales and molecular levels, of space, time, and organizational complexity?
- Pattern Discovery: What is the relation between pattern and process both in mathematical and biological systems?



Part I Genetic Codes, Biological Sequences, DNA and Protein Structures

2. Genetic Codes, Matrices, and Symmetrical Techniques

- ❖ Introduction
- ❖ Matrix Theory and Symmetry Preliminaries
- ❖ Genetic Codes and Matrices
- ❖ Challenges and Perspectives



2.1 INTRODUCTION

- All living organisms are unified by nature. All of them have identical molecular bases of the system of genetic coding.
- The set of four letters (A, C, G, T/U) forms the complementary pairs C-G and A-U (or A-T).
- The complementary letters C and G are connected by three hydrogen bonds.
- The complementary letters A and U (or A and T) are connected by two hydrogen bonds.
- The genetic code is named “the degeneracy code” because its 64 encode 20 amino acids and different amino acids are encoded by different quantities of triplets.



2.2 MATRIX THEORY AND SYMMETRY PRELIMINARIES

Matrix

- A rectangular table of *elements* (or *entries*), which may be numbers or, more generally, any abstract quantities that can be added and multiplied.
- Matrices are used to describe linear equations, keep track of the coefficients of linear transformations and to record data that depend on multiple parameters.
- Matrix Operations



2.2 MATRIX THEORY AND SYMMETRY PRELIMINARIES

Operation	Definition
Addition	<p>Given m-by-n matrices \mathbf{A} and \mathbf{B}, their <i>sum</i> $\mathbf{A}+\mathbf{B}$ is calculated entrywise, i.e.</p> $(\mathbf{A} + \mathbf{B})_{ij} = \mathbf{A}_{ij} + \mathbf{B}_{ij}, \text{ where } 1 \leq i \leq m \text{ and } 1 \leq j \leq n.$
Scalar multiplication	<p>Given a matrix \mathbf{A} and a number (also called a scalar in the parlance of abstract algebra) c, the <i>scalar multiplication</i> $c\mathbf{A}$ is given by multiplying every entry of \mathbf{A} by c:</p> $(c\mathbf{A})_{ij} = c \cdot \mathbf{A}_{ij}.$
Transpose	<p>The <i>transpose</i> of an m-by-n matrix \mathbf{A} is the n-by-m matrix \mathbf{A}^T (also denoted by \mathbf{A}^{tr} or ${}^t\mathbf{A}$) formed by turning rows into columns and columns into rows:</p> $(\mathbf{A}^T)_{ij} = \mathbf{A}_{j,i}.$
Kronecker (or tensor) multiplication	<p>Given m-by-m matrix $\mathbf{A}=(a_{ij})$ and n-by-n matrix $\mathbf{B}=(b_{ij})$, their Kronecker multiplication is mn-by-mn matrix $\mathbf{A} \otimes \mathbf{B}$:</p> $\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1m}\mathbf{B} \\ \dots & \dots & \dots & \dots \\ a_{1m}\mathbf{B} & a_{2m}\mathbf{B} & \dots & a_{mm}\mathbf{B} \end{bmatrix}$



2.2 MATRIX THEORY AND SYMMETRY PRELIMINARIES

Symmetry

- An object is *symmetric* with respect to a given mathematical operation, if, when applied to the object, this operation does not change the object or its appearance.
- In 2D geometry the main kinds of symmetry of interest are with respect to the basic Euclidean plane isometries: translations, rotations, reflections, and glide reflections.
- Many structural features of molecules are governed by consideration of symmetry.
- Symmetries may also be found in living organisms including humans and other animals.

2.3 GENETIC CODE AND MATRICES

A, C, G, T



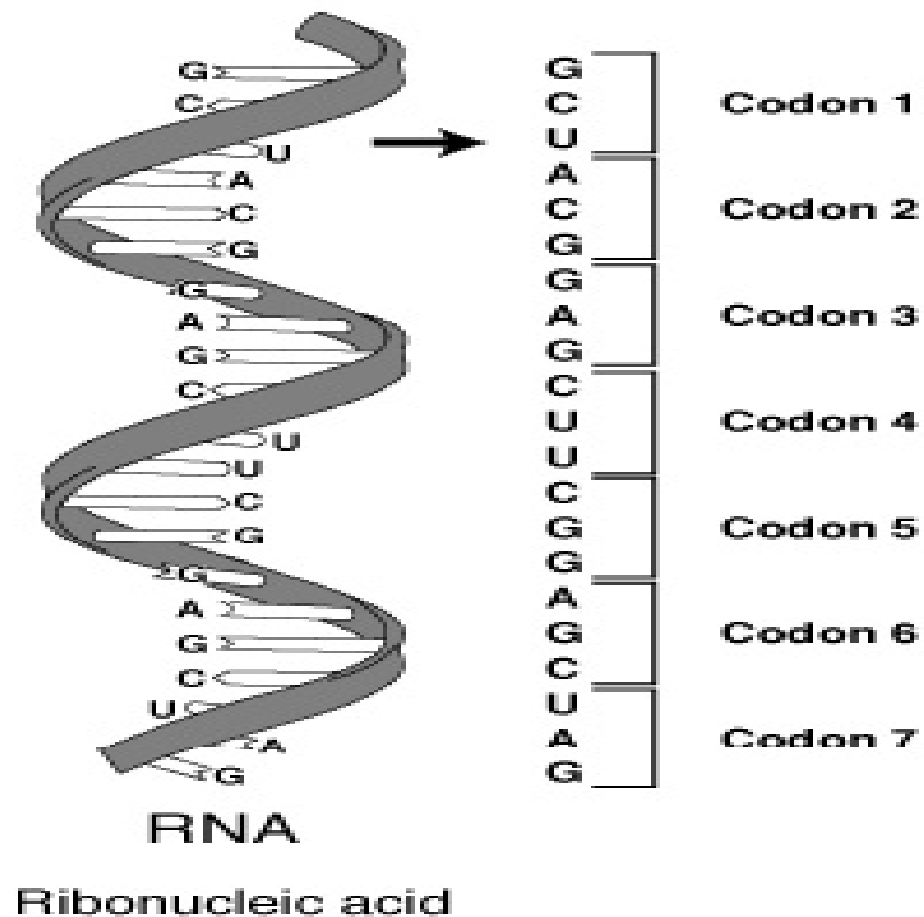
$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$



STANDARD GENETIC CODE (64! ARRANGEMENTS)

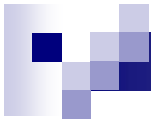
CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG

BUILDING BLOCKS OF PROTEINS: AMINO ACID/CODONS



BINARY REPRESENTATION OF STANDARD GENETIC CODE

G_3	000	001	011	010	110	111	101	100
000	000 000	001 000	011 000	010 000	110 000	111 000	101 000	100 000
001	000 001	001 001	011 001	010 001	110 001	111 001	101 001	100 001
011	000 011	001 011	011 011	010 011	110 011	111 011	101 011	100 011
010	000 010	001 010	011 010	010 010	110 010	111 010	101 010	100 010
110	000 110	001 110	011 110	010 110	110 110	111 110	101 110	100 110
111	000 111	001 111	011 111	010 111	110 111	111 111	101 111	100 111
101	000 101	001 101	011 101	010 101	110 101	111 101	101 101	100 101
100	000 100	001 100	011 100	010 100	110 100	111 100	101 100	100 100



HAMMING DISTANCE OF STANDARD GENETIC CODE

0	1	2	1	2	3	2	1
1	0	1	2	3	2	1	2
2	1	0	1	2	1	2	3
1	2	1	0	1	2	3	2
2	3	2	1	0	1	2	1
3	2	1	2	1	0	1	2
2	1	2	3	2	1	0	1
1	2	3	2	1	2	1	0

GENETIC CODE EQUIVALENCE

- **First kind of equivalence:** Two pairs of equivalent letters, where $A = C$ and $U = G$, are formed according to an attribute, **A and C have a property of amino-mutating** of two nitrogenous bases – A and C - in RNA under action of nitrous acid HNO_2 . The other two bases **U and G do not have the property of amino-mutating** and do not have such a located amino-group; so they are equivalent from viewpoint of absence of this attribute. This was classified by **Wittmann in 1961**. Here we have **$G = U$ and $A = C$** .
- **Second kind of equivalence:** Second kind of pairs of equivalent letters is formed on the basis of the attribute of complementation of these nitrogenous bases in molecules of nucleic acids: **$C = G$ (they form complementary pair with three hydrogen bonds between them)** and **$A = U$ (they form complementary pair not with three, but with two hydrogen bonds)**. This equivalence relation is denoted by $C = G$ and $A = U$.



ATTRIBUTIVE MAPPINGS

- We'll use these attributes equivalence to assign RNA bases A, C, G, U values of 0, 1, 2, and 3 for each pair of equivalence. The following lists these assignments:
- Case 1: $G = U = 0$, $A = C = 1$, amino- mutating absence/present (0, 1)-combination,
- Case 2: $C = U = 1$, $A = G = 2$, pyrimidines /purines ring-based (1, 2)-combination,
- **Case 3:** $A = U = 2$, $C = G = 3$, hydrogen bonds-based (2,3)-combination.



ATTRIBUTIVE MAPPINGS

- Based on these three attributes equivalences and assignments, three mapping relations from $\mathbf{R} = \{A, C, G, U\}$ to $\mathbf{N} = \{0, 1, 2, 3\}$ were defined as follows (onto and subjective):
- $\alpha: \{A, C, G, U\} \rightarrow \{0, 1\}$ with $\alpha(G) = \alpha(U) = 0$, $\alpha(A) = \alpha(C) = 1$,
- $\beta: \{A, C, G, U\} \rightarrow \{1, 2\}$ with $\beta(C) = \beta(U) = 1$, $\beta(A) = \beta(G) = 2$,
- $\gamma: \{A, C, G, U\} \rightarrow \{2, 3\}$ with $\gamma(A) = \gamma(U) = 2$, $\gamma(C) = \gamma(G) = 3$.



Matrix $G\beta$ [1,2] ($C = U = 1$, $A = G = 2$, β with addition and total c/r sums)

3	4	4	5	4	5	5	6	36
3	4	4	5	4	5	5	6	36
3	4	4	5	4	5	5	6	36
3	4	4	5	4	5	5	6	36
3	4	4	5	4	5	5	6	36
3	4	4	5	4	5	5	6	36
3	4	4	5	4	5	5	6	36
3	4	4	5	4	5	5	6	36
24	32	32	40	32	40	40	48	288



Basic properties of G_γ [2,3]

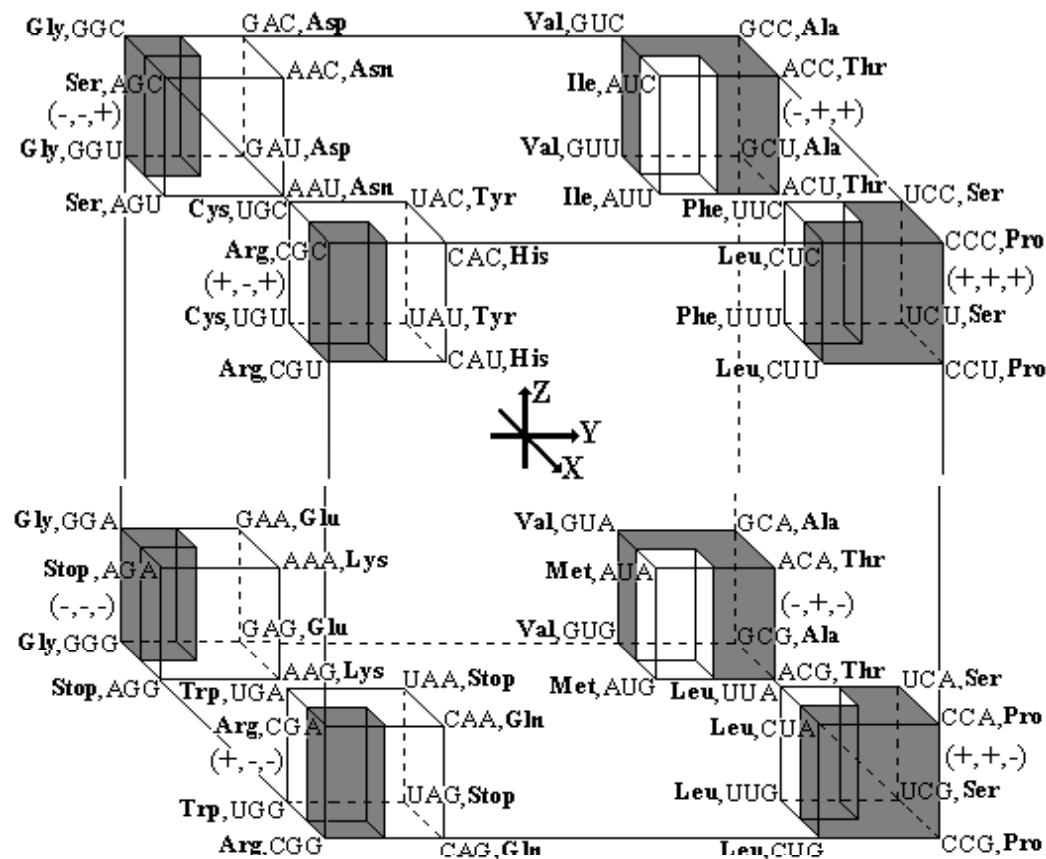
- The matrix $G(i,j)$ is symmetric since $G(i,j) = G(i,j)^T$.
- The matrix $G(i,j)$ is singular since $\text{Det}(G(i,j)) = 0$.
- The eigenvalues of $G(i,j)$ is $\{\lambda_1, \lambda_2, \dots, \lambda_8\} = \{0, 0, 0, 0, -4, -4, -4, 60\}$.
- The eigenvectors are $\{2, -1, -1, 0, -1, 0, 0, 1\}$, $\{1, 0, -1, 0, -1, 0, 1, 0\}$, $\{1, -1, 0, 0, -1, 1, 0, 0\}$, $\{1, -1, -1, 1, 0, 0, 0, 0\}$, $\{-1, 0, 0, 1, -1, 0, 0, 1\}$, $\{0, -1, 0, -1, 1, 1, 0, 1, 0\}$, $\{0, 0, -1, -1, 1, 1, 0, 0\}$, $\{1, 1, 1, 1, 1, 1, 1, 1\}$.



Basic properties of $G_\gamma [2,3]$

- These 8 vectors are linearly independent. They form a basis for a vector space of dimension of 8.
- The power of matrix is stochastic and the limit is a stochastic matrix with constant entries.
- $G_\gamma[2,3]=9P_1+8(P_2+P_3+P_4)+7(P_5+P_6+P_7)+6P_8$
- These 8 vectors are linearly independent. They form a basis for a vector space of dimension of 8.
- The power of matrix is stochastic and the limit is a stochastic matrix with constant entries.
- $G_\gamma[2,3]=9P_1+8(P_2+P_3+P_4)+7(P_5+P_6+P_7)+6P_8$

Hypercube of Standard Genetic Code



- Hypercube Representation
- Codon
 - Amino Acids
 - Each Vertex has 8 codons associated with a 8x8 permutation matrix



2.4 CHALLENGES AND PERSPECTIVES

- Why the genetic alphabet consists of four letters?
- Why does the genetic code encode 20 amino acids?
- How is the system structure of the molecular genetic code connected with known principles of quantum mechanics, which were developed to explain phenomena on atomic and molecular levels?
- Why has nature chosen the special code conformity between 64 genetic triplets and 20 amino acids?
- What kind of mathematical approach should be chosen among many possible approaches to represent and model structuralized ensembles of molecules of the genetic code?



Part I Genetic Codes, Biological Sequences, DNA and Protein Structures

3. Biological Sequences, Sequence Alignment, and Statistics

- ❖ Introduction
- ❖ Mathematical Sequences
- ❖ Sequence Alignment
- ❖ Sequence Analysis and Further Discussions
- ❖ Challenges and Perspectives



3.1 INTRODUCTION

Biological sequences

- DNA sequences (also called genetic sequences or nucleotide sequences).
- A succession of letters representing the primary structure of a real or hypothetical DNA molecule or strand, with the capacity to carry information. The possible letters are *A*, *C*, *G*, and *T*, representing the four nucleotide subunits of a DNA strand - adenine, cytosine, guanine, thymine bases covalently linked to a phospho-backbone. In the typical case, the sequences are printed abutting one another without gaps, as in the sequence AAAGTCTGAC, going from 5' to 3' from left to right.
- DNA sequences instruct the formation of amino acid sequences and determine the expression and regulation of genes. They determine the main aspects of the life process.



3.1 INTRODUCTION

Biological sequences

- Amino acid sequences (also called peptide sequences or protein sequences).
- Amino Acid Alphabets: A, R, ...V
- Amino acid sequences determine the structures and functions of proteins. The abundant biological sequence data provide us with the most important information of life.



STANDARD AMINO ACID ABBREVIATIONS

Amino Acid	3-Letter Abbreviation	1-Letter Abbreviation	Amino Acid	3-Letter Abbreviation	1-Letter Abbreviation
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic acid	Asp	D	Phenylalanine	Phe	F
Cysteine	Cys	C	Proline	Pro	P
Glutamic acid	Glu	E	Serine	Ser	S
Glutamine	Gln	Q	Threonine	Thr	T
Glycine	Gly	G	Tryptophan	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V



3.2 MATHEMATICAL SEQUENCES

Mathematical Sequence

- An ordered list of objects (or events). It contains members (also called *elements* or *terms*).
- The number of members (possibly infinite) is called the *length* of the sequence.
- Unlike a set, order matters, and the exact same elements can appear multiple times at different positions in the sequence.



3.2 MATHEMATICAL SEQUENCES

- In the language of monoids, a finite set is called an alphabet denoted by Σ . For example,
 - $\Sigma = \{0, 1\}$ is an alphabet of binary numbers: Binary sequences
 - $\Sigma = \{A, C, G, T\}$ is an alphabet of DNA basis. genetic or DNA sequences are sequences over the alphabet of nucleotides
 - Amino acid sequences are sequences over the alphabet of amino acids
- A subsequence of a given sequence is a sequence formed from the given sequence by deleting some of the elements without disturbing the relative positions of the remaining elements.



3.2 MATHEMATICAL SEQUENCES

- An infinite binary sequence can represent a formal language (a set of strings) by setting the n -th bit of the sequence to 1 if and only if the n th string is in the language. Therefore, the study of complexity classes, which are sets of languages, may be regarded as the study of sets of infinite sequences.
- An infinite sequence drawn from the alphabet $\{0, 1, \dots, b-1\}$ may also represent a real number expressed in the base- b positional number system. This equivalence is often used to bring the techniques of real analysis to bear on complexity classes.



3.3. SEQUENCE ALIGNMENT

- The foundation of sequence alignment and analysis is based on the fact that biological sequences develop from pre-existing sequences instead of being invented by nature from the beginning.
- The sequence of a gene can be altered in a number of ways. Three kinds of changes can occur at any given position within a sequence:
 - **Point mutations**, often caused by chemicals or malfunction of DNA replication, exchange of a single nucleotide for another. Most common is the transition that exchanges a purine for a purine ($A \leftrightarrow G$) or a pyrimidine for a pyrimidine, ($C \leftrightarrow T$).



3.3. SEQUENCE ALIGNMENT

- **Insertions** add one or more extra nucleotides into the DNA. They are usually caused by transposable elements or errors during replication of repeating elements (e.g., AT repeats). Insertions in the coding region of a gene may alter splicing of the mRNA (splice site mutation), or cause a shift in the reading frame (frame shift), both of which can significantly alter the gene product. Insertions can be reverted by excision of the transposable element.
- **Deletions** remove one or more nucleotides from the DNA. Like insertions, these mutations can alter the reading frame of the gene. Note that a deletion is not the exact opposite of an insertion: the former is quite random while the latter consists of a specific sequence inserting at locations that are not entirely random or even quite narrowly defined.



3.3 SEQUENCE ALIGNMENT

- An alignment between two (or more) sequences is a pairwise (multiple) comparison between the characters of each sequence
- The basic sequence analysis is to ask if two or more sequences are related.
- A true alignment of biological sequences is one that reflects the evolutionary relationship between two or more homology which are the sequences that share a common ancestor.



3.3 SEQUENCE ALIGNMENT

- The key issues to sequence alignments are
 - What sorts of alignment should be considered:
 - The scoring system used to rank alignments;
 - The algorithm used to find optimal (or good) scoring alignments;
 - The statistical methods used to evaluate the significance of an alignment score.
- Biological sequence alignment is a difficult problem (The Number of Alignments!)

THE NUMBER OF ALIGNMENTS

- Let $\mathbf{a} = a_1 a_2 \dots a_m$ and $\mathbf{b} = b_1 b_2 \dots b_n$ be two sequences over the alphabet Σ of length, n and m . An alignment of the sequences \mathbf{a} and \mathbf{b} is a pair of sequences $\mathbf{a}^* = a \ a \dots a$ and $\mathbf{b}^* = b \ b \dots b$ of equal length of L defined by inserting blanks to the sequences \mathbf{a} and \mathbf{b} over the extended alphabet $\Sigma^* = \Sigma \{-\}$. The alignment of \mathbf{a}^* and \mathbf{b}^* is represented in a tabular form:

$$a_1 \ a_2 \ \dots \ a_m$$

$$b_1 \ b_2 \ \dots \ b_n$$

where $\max\{m, n\} \leq L \leq m + n$. When $L = m + n$, the alignment is given by

$$\begin{array}{ccccccc} a_1 & a_2 & \dots & a_m & - & - & \dots & - \\ - & - & \dots & - & b_1 & b_2 & \dots & b_n \end{array}$$



THE NUMBER OF ALIGNMENTS

- A column that contains two identical characters is called a match,
- A column that contains two different nonblank characters is called mismatch,
- A column that contains a blank is called an indel (**i**nsertion/**d**eletion).

The total number of alignments $f(m, n)$ satisfies following recurrence relation:

$$f(m, n) = f(m-1, n) + f(m-1, n-1) + f(m, n-1)$$

- This recurrence relation was derived by Waterman and it was demonstrated that this number increases rapidly. For example, two sequences of length 1000 have

alignments. $f(1000, 1000) \approx 10^{767.4...}$



PAIRWISE SEQUENCE ALIGNMENT

- Pairwise sequence alignment methods are used to find the best-matching piecewise (local) or global alignments of two query sequences.
- Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high homology to a query).
- Three primary methods of producing pairwise alignments are
 - Global alignment,
 - Local alignment,
 - Global-local alignment.



GLOBAL ALIGNMENT

- Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size.
- It provides the common means to measure the degree of overall similarity between two sequences. FASTA (FAST ALL) developed by Pearson and Lipman (Pearson and Lipman, 1988) is a heuristic algorithm for global sequence alignment. It's widely used to align a query sequence against all sequences of a database.



GLOBAL ALIGNMENT

- Here is a commonly used algorithm for optimal global alignment. We point out that the optimal alignments depend on the input sequences and the algorithm parameters. The algorithm parameters assigned to matches, mismatches and indels are determined by experience.
- Optimal sequence alignment is closely related to the problem of finding the optimal edit distance in binary code. This is an old problem in coding theory introduced by Levenshtein (Levenshtein, 1966). The theory of semigroups and monoids provides the mathematical background for the manipulation of words over a finite alphabet.



SCORE FUNCTION/SIMILARITY SCORES

- Let $\mathbf{a} = a_1 a_2 \dots a_m$ and $\mathbf{b} = b_1 b_2 \dots b_n$ be two sequences over the alphabet Σ^* of approximately the same length. We define the similarity scores $s(a, b)$ over the alphabet Σ^* as follows:
 - $s(a, a) > 0$ for all a
 - $s(a, b) < 0$ for some (a, b) pairs
 - $s(a, -) = s(-, a) = -g(a)$ ($-g(a)$ is the indel penalty associated with a .)
- The global pairwise similarity alignment problem is to find the maximum similarity between the two sequences.

$$S(\mathbf{a}, \mathbf{b}) = \max \sum_{i=1}^L s(a_i^*, b_i^*)$$



SCORE FUNCTION/SIMILARITY SCORES

- where the maximum is over all alignments. Here the individual score $s(x, y)$ may be defined as

$$s(x, y) = \log \frac{p_{x, y}}{q_x q_y}$$

- where $p_{x,y}$ is the probability of the characters x and y to occur as an aligned column pair in a pairwise alignment of the match model defined as

$$P(a, b | M) = \prod p_{x, y}$$

- And q_x is the relative frequency of the character x to occur in the sequences a and b in the random model R defined as

$$P(a, b | R) = \prod q_x \prod q_y$$



SCORE FUNCTION/DISTANCE MEASURES

- The distance measure can be defined for the global pairwise distance alignment. Let $d(a, b)$ be the distance over the alphabet Σ^* as below:
 - $d(a, a) = 0$ for all a
 - $d(a, b) = d(b, a)$, cost of a mutation of a into b
 - $d(a, -) = d(-, a) = g(a)$, positive cost of inserting or deleting of the character a .

Define

$$D(\mathbf{a}, \mathbf{b}) = \min \sum_{i=1}^L d(a_i^*, b_i^*)$$

where the minimum is over all alignments of \mathbf{a} with \mathbf{b} .

- The main results on global pairwise alignment are stated below.



OPTIMAL GLOBAL SIMILARITY ALIGNMENT

THEOREM 3.1 (Optimal Global Similarity Alignment): Let $\mathbf{a} = a_1 a_2 \dots a_m$ and $\mathbf{b} = b_1 b_2 \dots b_n$ be two sequences over the alphabet Σ , define

$$S(i, j) = S(a_1 a_2 \dots a_i, b_1 b_2 \dots b_j)$$

And set

$$S(0, 0) = 0, S(0, j) = \sum_{k=1}^j s(-, b_k), S(i, 0) = \sum_{k=1}^i s(a_k, -)$$

Then

$$S(i, j) = \max \{S(i-1, j) + s(a_i, -), S(i-1, j-1) + s(a_i, b_j), S(i, j-1) + s(-, b_j)\}.$$

In particular,

$$S(\mathbf{a}, \mathbf{b}) = S(m, n).$$



OPTIMAL GLOBAL DISTANCE ALIGNMENT

THEOREM 3.1 (Optimal Global Distance Alignment): Let $\mathbf{a} = a_1 a_2 \dots a_m$ and $\mathbf{b} = b_1 b_2 \dots b_n$ be two sequences over the alphabet Σ , define

$$D(i, j) = D(a_1 a_2 \dots a_i, b_1 b_2 \dots b_j)$$

And set

$$D(0, 0) = 0, D(0, j) = \sum_{k=1}^j d(-, b_k), D(i, 0) = \sum_{k=1}^i d(a_k, -)$$

Then

$$D(i, j) = \min \{D(i-1, j) + d(a_i, -), D(i-1, j-1) + d(a_i, b_j), D(i, j-1) + d(-, b_j)\}.$$

In particular,

$$D(\mathbf{a}, \mathbf{b}) = D(m, n).$$



LOCAL ALIGNMENT

- Biological sequences often contain similar subsequences that are preserved during the course of evolution.
- Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context.
- The problem of finding highly related subsequences of two sequences is accomplished by local alignment.
- The Smith-Waterman algorithm is a general local alignment method also based on dynamic programming. With sufficiently similar sequences, there is no difference between local and global alignments.



LOCAL ALIGNMENT

- The BLAST (Basic Local Alignment Sequence Tool) is a fast heuristic algorithm for local alignment developed by Altschult et al., in 1990. BLAST finds regions of similarity.
- Here we consider only the subsequences of consecutive elements. Any subsequence of a sequence $a_1 a_2 \dots a_m$ has the form $a_i a_{i+1} \dots a_{m+k}$ for some $1 \leq i \leq m$ and $k \leq m-i$. We present the optimal local alignment developed by Smith-Waterman algorithm (Smith and Waterman, 1981). Let $\mathbf{a} = a_1 a_2 \dots a_m$ and $\mathbf{b} = b_1 b_2 \dots b_n$ be two sequences over the alphabet Σ .

Define

$$S(ij, kl) = S(a_i \dots a_j, b_k \dots b_l).$$

- What is the maximum similarity between subsequences of \mathbf{a} and \mathbf{b} ? That is, find

$$L(\mathbf{a}, \mathbf{b}) = \max \{S(ij, kl) = S(a_i \dots a_j, b_k \dots b_l) \mid 1 \leq i \leq j \leq m, 1 \leq k \leq l \leq n\}.$$



OPTIMAL LOCAL ALIGNMENT

THEOREM 3.3 (Optimal Local Alignment): Let $\mathbf{a} = a_1 a_2 \dots a_m$ and $\mathbf{b} = b_1 b_2 \dots b_n$ be two sequences over the alphabet Σ . Define

$$L(i,0) = 0, \quad 0 \leq i \leq m, \quad L(0,j) = 0, \quad 0 \leq j \leq n,$$

and

$$L(i,j) = \max \{ 0, L(i-1,j-1) + s(a_i, b_j), L(i-1,j) + s(a_i, -), L(i,j-1) + s(-, b_j) \} \\ 1 \leq i \leq m, \quad 1 \leq j \leq n,$$

where $s(x, y) \geq 0$ if x and y match; $s(x, y) \leq 0$ if x and y do not match or one of them is a blank.

Then

$$L(j, l) = \max \{ 0, S(a_i \dots a_j, b_k \dots b_l) \mid 1 \leq i \leq j \leq m, 1 \leq k \leq l \leq n \}.$$

Each maximal entry $L(j^*, l^*)$ of the array L corresponds to an optimal local alignment of the sequences \mathbf{a} and \mathbf{b} .



GLOBAL-LOCAL ALIGNMENT

- Global-local alignment (hybrid alignment) compares a sequence with the subsequences of another sequence.
- This can be especially useful when the downstream part of one sequence overlaps with the upstream part of the other sequence. In this case, neither global nor local alignment is entirely appropriate: a global alignment would attempt to force the alignment to extend beyond the region of overlap, while a local alignment might not fully cover the region of overlap by Lipman, et al., in 1984.
- Let $\mathbf{a} = a_1 a_2 \dots a_m$ and $\mathbf{b} = b_1 b_2 \dots b_n$ be two sequences of different length over the alphabet Σ . Here we let $m \leq n$. The problem is to find the maximum matching of the shorter sequence with the longer one. That is, find

$$H(\mathbf{a}, \mathbf{b}) = \max \{ S(\mathbf{a}, b_k \dots b_l) \mid 1 \leq k \leq l \leq n \}.$$



GLOBAL-LOCAL ALIGNMENT

THEOREM 3.4 (Optimal Global-Local Alignment): Let $\mathbf{a} = a_1 a_2 \dots a_m$ and $\mathbf{b} = b_1 b_2 \dots b_n$ be two sequences over the alphabet Σ . Define

$$H(0, j) = 0, \quad 0 \leq j \leq m, \quad H(i, 0) = 0, \quad 0 \leq i \leq n,$$

And

$$H(i, j) = \max \{ H(i-1, j-1) + s(a_i, b_j), H(i-1, j) + s(a_i, -), H(i, j-1) + s(-, b_j) \} \\ 1 \leq i \leq m, \quad 1 \leq j \leq n,$$

where $s(x, y) \geq 0$ if x and y match; $s(x, y) \leq 0$ if x and y do not match or one of them is a blank.

Then

$$H(i, j) = \max \{ S(a_i \dots a_i, b_k \dots b_j) \mid 1 \leq i \leq m, \quad 1 \leq k \leq j \leq n \}.$$

In particular,

$$H(\mathbf{a}, \mathbf{b}) = \max \{ H(m, j) \mid 1 \leq j \leq n \}.$$



MULTIPLE SEQUENCE ALIGNMENT

- Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences at a time.
- The multiple sequence alignment is simultaneously aligning a number of sequences to determine common features among the collection of sequences.
- Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related.
- To identify the common features, one needs to determine an optimal alignment for the entire collection of sequences. Multiple sequence alignments are computationally difficult to produce and most formulations of the problem lead to NP-complete combinatorial optimization problems.

MULTIPLE SEQUENCE ALIGNMENT

➤ Let $\Omega = (\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_k)$ be a family of sequences over the alphabet Σ ,

$$\begin{aligned} a_1 &= a_{11} \cdots a_{1n_1} \\ &\vdots \\ a_k &= a_{k1} \cdots a_{kn_k} \end{aligned}$$

and

$\Sigma^* = (\mathbf{a} \mathbf{a} \dots \mathbf{a})$ be a corresponding family of sequences with equal length l over the extended alphabet $\Sigma^* = \Sigma \{-\}$,

$$\begin{aligned} a_1^* &= a_{11}^* \cdots a_{1l}^* \\ &\vdots \\ a_k^* &= a_{k1}^* \cdots a_{kl}^* \end{aligned}$$

by inserting blanks. Where $\max\{n_1, n_2, \dots, n_k\} \leq l \leq n_1 + n_2 + \dots + n_k$.



MULTIPLE SEQUENCE ALIGNMENT

- The optimal global alignment is to find the maximum similarity between these sequences Ω in terms of a scoring function $s(\Omega^*)$, that is

$$S(\Omega) = \max \{ s(\Omega^*) \mid \Omega^* \text{ is a multiple alignment of } \Omega \},$$

where

$$s(\Omega^*) = \sum_{i=1}^l s(a_{1i}^*, \dots, a_{ki}^*)$$

- is the sum of scores of the columns. Here it is assumed that the columns of the alignment are statistically independent. We are now in a position to state the optimal multiple sequence alignment result.

OPTIMAL GLOBAL MULTIPLE SEQUENCE ALIGNMENT

THEOREM 3.5 (Optimal Global Multiple Sequence Alignment): Let $\Omega = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k)$ be a family of sequences over the alphabet Σ ,

$$\begin{aligned} \mathbf{a}_1 &= a_{11} \cdots a_{1n_1} \\ &\vdots \\ \mathbf{a}_k &= a_{k1} \cdots a_{kn_k} \end{aligned}$$

and $\mathbf{B} = (b_1, \dots, b_k)$ be binary vector over $\{0, 1\}$ and define $b * x = x$ if $b=1$ and $b * x = -$ if $b = 0$. For all index vectors (i_1, \dots, i_k) , define

$$S(i_1, \dots, i_k) = \max \{S(i_1 - b_1, \dots, i_k - b_k) + s(b_1 * a_{1i_1}, \dots, b_k * a_{ki_k})\}$$

where the maximum is taken over all nonzero binary vectors \mathbf{B} . Also we set

$$S(0, \dots, 0) = 0$$

Then

$$S(i_1, \dots, i_k) = S(a_{1i_1}, \dots, a_{1i_1}, \dots, a_{ki_k}, \dots, a_{ki_k})$$

In particular,

$$S(\Omega) = S(n_1, \dots, n_k).$$



MULTIPLE SEQUENCE ALIGNMENT-EXAMPLE

- **EXAMPLE** Here we display the representation of a protein multiple sequence alignment produced with ClustalW (Chenna, et al., 2003). The sequences are instances of the acidic ribosomal protein P0 homolog (L10E) encoded by the *Rplp0* gene from multiple organisms. The protein sequences were obtained from SwissProt searching with the gene name. This is generated by Miguel Andrade February 2006 (UTC).
- **TABLE 3.2** Only the first 90 positions of the alignment are displayed. The colours represent the amino acid conservation according to the properties and distribution of amino acid frequencies in each column. Note the two completely conserved residues arginine (R) and lysine (K) marked with an asterisk at the top of the alignment.



```

*      :      *      :      :
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MVRENKAAWKAQYFIKVVELFDEFKCFIVGADNVGSKOMQIIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN
-----MSGAG-SKRKKLFIEKATKLFYTDKMIVAEADFGSSQLQKIRKSIIRGI-GAVLMGKNTMIRKIVIRDLADSI
-----MSGAG-SKRKNVFIEKATKLFYTDKMIVAEADFGSSQLQKIRKSIIRGI-GAVLMGKNTMIRKIVIRDLADSI
-----MAKLSKQKKQMYIEKLSSLIQQYSKILIVHVDNVGSKOMQIIRMSLRGK-ATILMSGKNTMIRKIVIRDLADSI
-----MIGLAVTTTKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG---
-----MRIMAVITQERKIAKWKIEEVKELEOKLREYHTIIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG---
-----MKRLALALKQQRKVASWKEEVKELTELKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFGIAAKNAG---
MSVVS LVGQMYKREKPIPEWKTLMLELEELFSKHRVFLFADLTGPTTFVVRVVRKKLWKK-YPMMAVAKKRIILRAMKAAGLE-
-MMLAIGKRRYVRTQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRY-GVIKIIKPTLFGIAAKNAG---
-----MAEERHHTTEHIPQWKDEIENIKELIQSHKVFVGMVIEGILATKMKIRRDLDKV-AVLKVSNTLTERALNQLG---
-----MAEERHHTTEHIPQWKDEIENIKELIQSHKVFVGMVIEGILATKMKIRRDLDKV-AVLKVSNTLTERALNQLG---
-----MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVFAGOMOKIRREFRGK-AEIKVVKNTLLERALDAG---
MAVKAKGQPPSGYE PKVAEWKRREVKELKELMDEYENVGLVDLEGIAPQLQEIIRAKLRERDTIIRMSRNTLMRIAEEKLEDEI
-----MAHVAEWKKKEVEELANLIKSYPIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAQEI
-----MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIIRDKIR-GTMTLKMSRNTLIERAIKEVAEEI
-----MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPARQLQEIIRDKIR-DQMTLKMSRNTLIERAIKEVAEEI
-----METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIIRDKIR-DKVKLRMSRNTLIERAIKEVAEEI
-----MAHVAEWKKKEVEELANLIKSYPIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAQEI
-----MAHVAEWKKKEVEELAKLIKSYPIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAQEI
-----MAHVAEWKKKEVEELANLIKSYPIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAQEI
-----MAHVAEWKKKEVEELANLIKSYPIALVDVAGVPAYPLSKMRDKIR-GKALLRVSRNTLIELAIKKAQEI
-----MSAESERKTETIPEWKQEEVDIVEMIESYESVGVVNIAGIPSRQLQDMRDLHGT-AELRVSRNTLIERALDDVD---
-----MSESEVRQTEVIPQWKREVDVDFIESYESVGVVNIAGIPSRQLQDMRDLHGT-AELRVSRNTLIERALDDVD---
-----MSAEQRTTEVIPQWKREVDVDFIESYESVGVVNIAGIPSRQLQDMRDLHGT-AELRVSRNTLIERALDDVD---
-----MKEVSQKKKELVNEITRIKASRSVAIVDTAGIRTRQIDIRGKNRGK-INLKVVKNTLLFKALENLGD---
-----MRKINPKKKEIVSELAODITKSKAVAIVDIKGVRTROMQDIRAKNRDK-VKIKVVKNTLLFKALENLGD---
-----MTEPAQWKIDFVKNLENEINSRKVAIVSIKGLRNNEFKIRNSIRDK-ARIKVSRRARLLRLAIENIGK---
1.....10.....20.....30.....40.....50.....60.....70.....80.....

```

Only the first 90 positions of the alignment are displayed

MULTIPLE SEQUENCE ALIGNMENT-EXAMPLE



PROFILE AND SEQUENCE ALIGNMENT

- Profile analysis has long been a useful tool in finding and aligning distantly related sequences and in identifying known sequence domains in new sequences.
- A profile is a description of the consensus of a multiple sequence alignment. It represents the common characteristics of a family of similar sequences where any single sequence is just one realization of the family's characteristics.
- The optimal profile-sequence alignment is to find the maximum similarity between the profile \mathbf{P} and the sequence \mathbf{a} , that is

$$S(\mathbf{P}, \mathbf{a}) = \max \{ s(\mathbf{P}^*, \mathbf{a}^*) \mid (\mathbf{P}^*, \mathbf{a}^*) \text{ is an alignment of } (\mathbf{P}, \mathbf{a}) \},$$

where $s(\mathbf{P}^*, \mathbf{a}^*)$ is a score function that may be defined as

$$s(\mathbf{P}^*, \mathbf{a}^*) = \sum_{i=1}^l \sum_{x \in \Omega^*} s(a_i^*, x) p_x$$

- with individual similarity score $s(a, x)$ on the alphabet Σ^* and the score between probability distribution $\mathbf{p} = (p_x)$ on the alphabet Ω^* and character x in Σ^* .



PROFILE AND SEQUENCE ALIGNMENT

Theorem 3.6 (Optimal Profile-Sequence Alignment): Let $\mathbf{P} = \mathbf{p}_1 \mathbf{p}_2 \dots \mathbf{p}_n$ be the profile of a multiple sequence alignment and $\mathbf{a} = a_1 a_2 \dots a_n$ be a sequence over the alphabet Σ^* , define

$$S(i, j) = S(\mathbf{p}_1 \mathbf{p}_2 \dots \mathbf{p}_i, a_1 a_2 \dots a_j), \quad 1 \leq i \leq m, \quad 1 \leq j \leq n$$

and set

$$S(0, 0) = 0, \quad S(i, 0) = \sum_{k=1}^i s(p_k, -), \quad S(0, j) = \sum_{k=1}^j s(-p, a_k) .$$

Then

$$S(i, j) = \max \{ S(i-1, j) + s(\mathbf{p}_i, -), S(i-1, j-1) + s(\mathbf{p}_i, a_j), S(i, j-1) + s(-p, a_j) \} .$$

In particular,

$$S(\mathbf{P}, \mathbf{a}) = S(m, n) .$$



OPTIMAL PROFILE-PROFILE ALIGNMENT

THEOREM 3.7 (Optimal Profile-Profile Alignment): Let $\mathbf{P} = \mathbf{p}_1 \mathbf{p}_2 \dots \mathbf{p}_m$ be the profile of a multiple sequence alignment and $\mathbf{Q} = \mathbf{q}_1 \mathbf{q}_2 \dots \mathbf{q}_n$ be the second profile of a multiple sequence alignment over the alphabet Σ^* , then define

$D(\mathbf{P}, \mathbf{Q}) = \min \{d(\mathbf{P}^*, \mathbf{Q}^*) = |(\mathbf{P}^*, \mathbf{Q}^*) \text{ is an alignment of } (\mathbf{P}, \mathbf{Q})\}$
as the minimum distance between the profiles \mathbf{P} and \mathbf{Q} . Let

$$D(i, j) = D(\mathbf{p}_1 \mathbf{p}_2 \dots \mathbf{p}_i, \mathbf{q}_1 \mathbf{q}_2 \dots \mathbf{q}_j), 1 \leq i \leq m, 1 \leq j \leq n$$

and set

$$D(0, 0) = 0, D(i, 0) = \sum_{k=1}^i d(\mathbf{p}_k, -_p), D(0, j) = \sum_{k=1}^j d(-_p, \mathbf{q}_k).$$

Then

$$D(i, j) = \min \{D(i-1, j) + d(\mathbf{p}_i, -_p), D(i-1, j-1) + s(\mathbf{p}_i, \mathbf{q}_j), S(i, j-1) + d(-_p, \mathbf{q}_j)\}.$$

In particular,

$$D(\mathbf{P}, \mathbf{Q}) = D(m, n).$$



3.4. SEQUENCE ANALYSIS /FURTHER DISCUSSIONS

- A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, musical score following, partial discharges and bioinformatics.
- Pattern Discovery: Given a sequence of data such as a DNA or amino acid sequence, a motif or a pattern is a repeating subsequence. Such repeated subsequences often have important biological significance and hence discovering such motifs in various biological databases turns out to be a very important problem in computational biology. Of course, in biological applications the various occurrences of a pattern in the given sequence may not be exact and hence it is important to be able to discover motifs even in the presence of small errors. Various tools are now available for carrying out automatic pattern discovery. This is usually the first step towards a more sophisticated task such as gene finding in DNA or secondary structure prediction in protein sequences at system level.



3.4. SEQUENCE ANALYSIS/FURTHER DISCUSSIONS

- Scoring functions: The choice of a scoring function that reflects biological or statistical observations about known sequences is important to producing good alignments.
- Structural alignments, which are usually specific to protein and sometimes RNA sequences, use information about the secondary and tertiary structure of the protein or RNA molecule to aid in aligning the sequences. These methods can be used for two or more sequences and typically produce local alignments; however, because they depend on the availability of structural information, they can only be used for sequences whose corresponding structures are known (usually through X-ray crystallography or NMR spectroscopy).



3.5 CHALLENGES AND PERSPECTIVES

The issues need to be addressed may include:

- Architecture of Data and Knowledge Repositories
- Databases: Flat, Relational and Object-Oriented; what is most appropriate?
- The imminent need for Ontologies in biology
- The Middle Layer: How to design it?
- Applications and integration of applications into the middle layer
- Reduction and Analysis of Data: the largest challenge!
- How to integrate legacy knowledge with data?
- User Interfaces: web browser and beyond



Part I Genetic Codes, Biological Sequences, DNA and Protein Structures

4. Structures of DNA and Knot Theory

- ❖ Introduction
- ❖ Knot Theory Preliminaries
- ❖ DNA Knots and Links
- ❖ Challenges and Perspectives



4.1 INTRODUCTION

- DNA is the genetic material of all cells, containing coded information about cellular molecules and processes.
- DNA is tightly packed into genes and chromosomes.
- In order for replication or transcription to take place, DNA must first unpack itself so that it can interact with enzymes.
- Replication and transcription are much easier to accomplish if the DNA is neatly arranged rather than tangled up in knots.
- Enzymes are essential to unpacking DNA. Enzymes act to slice through individual knots and reconnect strands in a more orderly way.
- Enzymes maintain the proper geometry and topology during the transformation and also ``cut" the DNA strands and recombine the loose ends.



DNA STRUCTURES

- B-DNA: Fully hydrated DNA, the most common encountered in vivo. Owing to the location of the helical axis in the center of the base pairs, the edges of the base pairs are about equally deep in the interior.
- A-DNA: When B-DNA is dehydrated, there is a reversible structural change to A-DNA
- Z-DNA: Unlike B-DNA and A-DNA, Z-DNA is a left-handed helix. The conformational change from B-DNA to Z-DNA is one mechanism for relief of the torsional strain found in B-DNA in vivo, and may serve as a switch mechanism to regulate gene expression.

The three structural variations of these grooves ("A", "B" and "Z" DNA), which differ in the relationship between the bases and the helical axis, offer one mechanism by which reactivity of DNA is modulated.



FORMS OF DNA

- **Supercoiled (or "knotted"):** Double stranded circular (or linear) DNA can have tertiary or higher order structure. Superhelicity is therefore sometimes referred to as DNA's tertiary structure. Supercoils refer to the DNA structure in which double-stranded circular DNA twists around each other. Supercoiling can be:
 - **negative** (right-handed): Supercoils formed by deficit in link are called negative supercoils.
 - **positive** (left-handed): Supercoils formed by an increase in link are called positive supercoils.



FORMS OF DNA

- **Relaxed:** Circular DNA without any superhelical twist is known as a relaxed molecule. DNA in its relaxed (ideal) state usually assumes the B configuration. In a relaxed double-helical segment of DNA, the two strands twist around the helical axis once every 10.6 base pairs of sequence. The following structures are consistent with the relaxed state:
 - Linear DNA (either straight or curved)
 - Closed circular DNA, provided its axis lies in a plane or on the surface of a sphere



FORMS OF DNA

Supercoiling is vital to two major functions




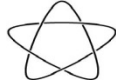


- It helps pack large circular rings of DNA into a small space by making the rings highly compact.
- It also helps in the unwinding of DNA required for its replication and transcription.
- Supercoiled DNA is thus the biological active form. The normal biological functioning of DNA occurs only if it is in the proper topological state.



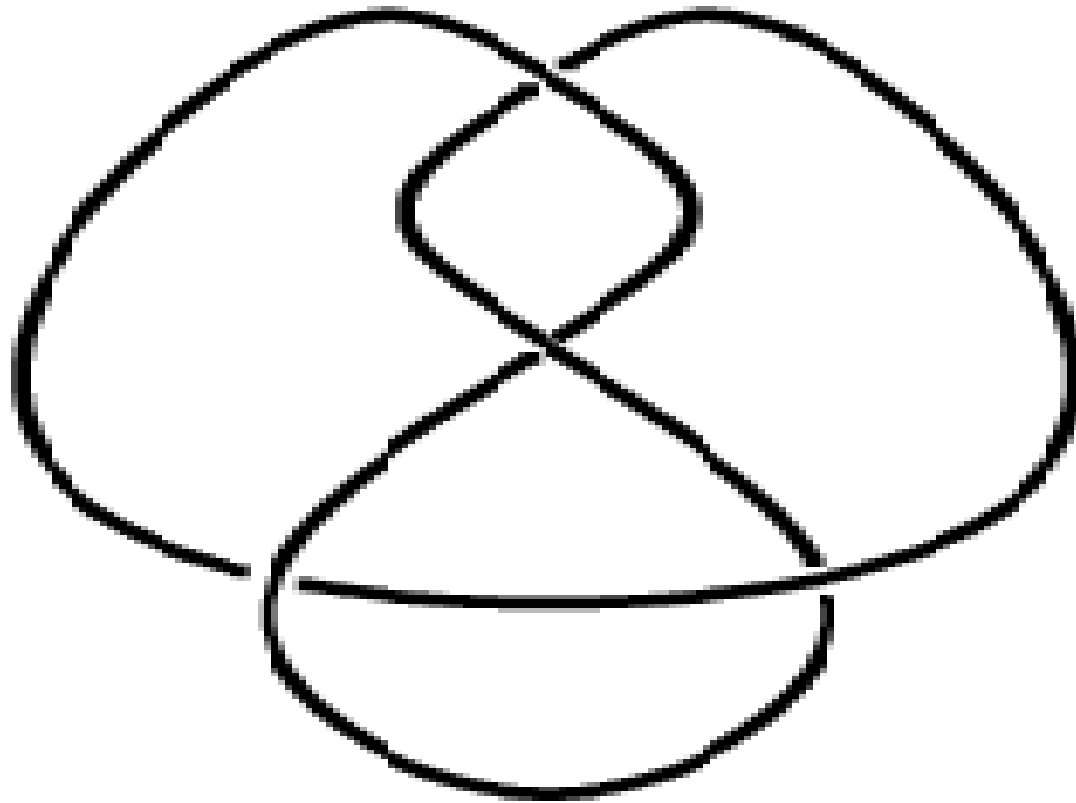
4.2. KNOT THEORY PRELIMINARIES

- A knot is a closed continuous curve in space that does not intersect itself anywhere.
- When a knot is deformed (i.e. stretched, compressed, bent, or twisted), but not cut or torn, all the deformed curves will be considered to be the same as the original closed knotted curve.
- The simplest knot of all is the unknotted circle, which we call the unknot or the trivial knot denoted by C . The next simplest knot is called a trefoil knot



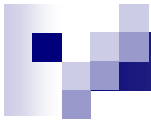
knot symbol	prime knot	Knot projection
0_1	unknot	
3_1	trefoil knot	
4_1	figure eight knot	
5_1	Solomon's seal knot	
6_1	stevedore's knot	
6_2	Miller Institute knot	

Primary Knots



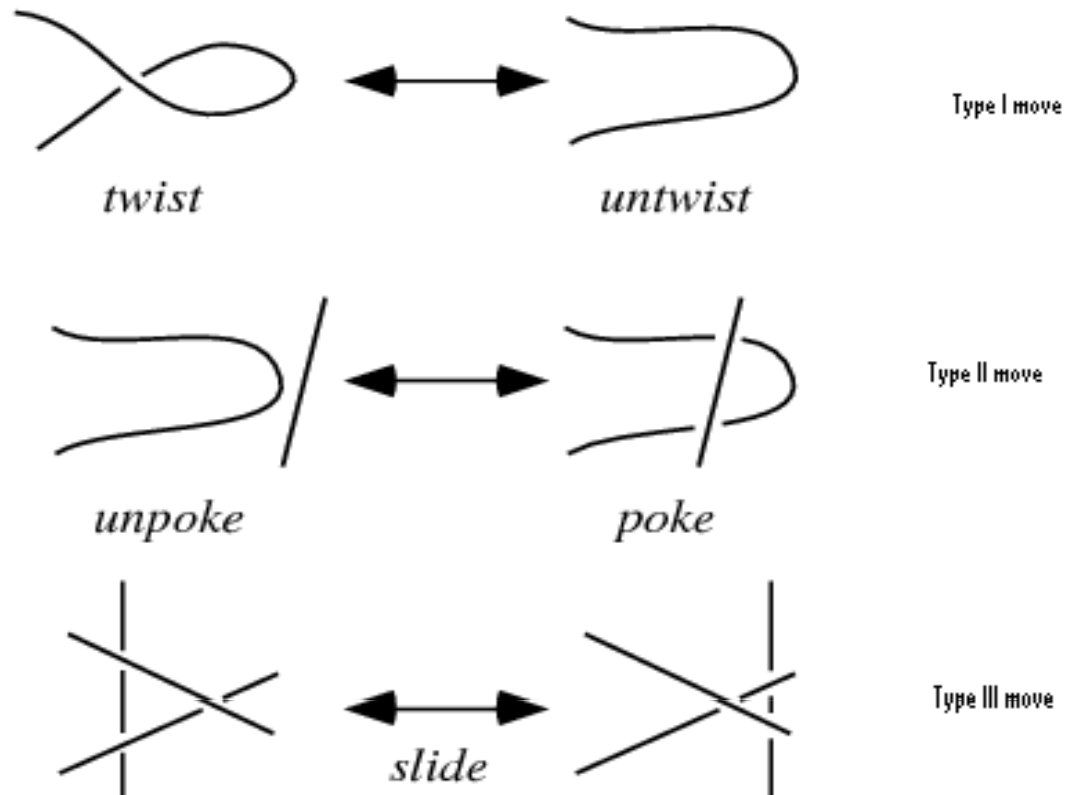
Crossing Number

The crossing number of a knot K denoted by $c(K)$, is the least number of crossings that occur in any projection of the knot. If a knot is nontrivial, then it has more than one crossing in a projection. The figure above called the figure-eight knot has four crossings.



COMPOSITION OF KNOTS

- Given two projections of knots and assuming the two projections do not overlap, one can compose a new knot by deleting a small arc from each knot projection and then connecting the four ending points by two new arcs. The resulting knot is called the composition (or knot sum) of the two knots, denoted by $K_1 \# K_2$ (or $K_1 + K_2$).
- Knot Moves



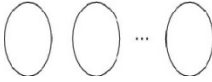



Reidemeister moves



LINKS

- A link is the union of a finite number of disjoint knots in three dimensional space.
- A knot will be considered a link of one component.
- Four common links, known as trivial link (or unlink), the Hopf link, the Whitehead link, and the Borromean links listed in Figure 4.6. The notation and ordering follows that of Rolfsen (1976), where c_k^r denotes the k th r -component link with crossing number c .
- Two links are considered to be the same if we can deform the one link to the other link without ever having any one of the knots intersect itself or any of the other loops in the process, That is, two links are considered equal if they are isotopic.



Link number	Link name	Link Diagram
0_1^2	Trivial link	
2_1^2	Hopf link	
5_1^2	Whitehead link	
6_2^3	Borromean link (rings)	

Link numbers

Trivial link, Hopf link, Whitehead link, and Borromean link

LINKING NUMBER

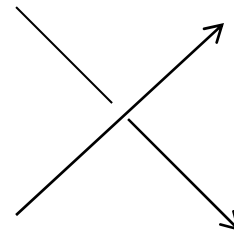
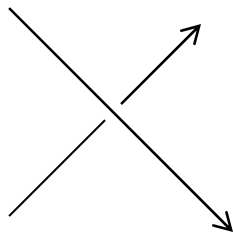
- Formally, a linking number is defined as the sum of +1 crossings and -1 crossing over all crossings between the two links divided by 2 calculated by the following formula:

$$L(K_1, K_2) = \frac{1}{2} \sum_{p \in \alpha \cap \beta} \varepsilon(p),$$

- where $\alpha \cap \beta$ is the set of crossings of α with β , and $\varepsilon(p)$ is the sign of the crossing.

Computing Linking Number:

- Let K_1 and K_2 be two components in a link L , and choose an orientation on each component. Then at each crossing between the two components, we count a +1 for each crossing of the first type, and a -1 for each crossing of the second type.





COMPUTING LINK NUMBER

- In other words, to each of these crossings is associated an index number of +1 or -1, according to the direction in which the tangent vector to the top curve must be rotated to coincide with the tangent vector to the bottom curve. If the rotation is clockwise, the index number is -1, and if it is counterclockwise, the index number is +1. Adding all the indices associated to all the crossings and dividing by 2 gives the link number of two knots denoted by $L(K_1, K_2)$.



PROPERTIES OF LINKING NUMBERS

- The linking number $L(K_1, K_2)$ is a property of the curves in space and is independent of the planar projection.
- The linking number $L(K_1, K_2)$ is unchanged if either of the curves is deformed continuously provided no breaks are made in either curve. Moreover the Reidemeister moves don't affect linking number.
- The linking number $L(K_1, K_2)$ changes sign if the direction of one of the curves is reversed.
- The linking number $L(K_1, K_2)$ changes sign if a pair of curves is reflected in a plane.



PROPERTIES OF LINKING NUMBERS

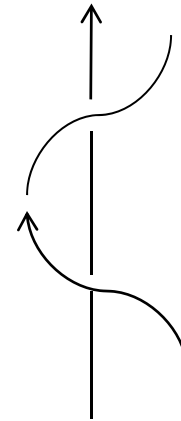
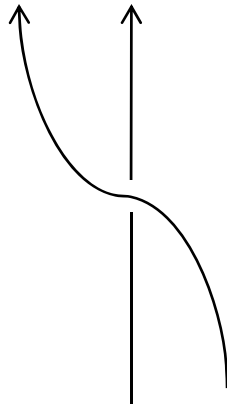
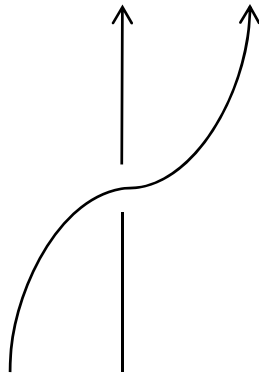
- Two oriented curves K_1 and K_2 bound a ribbon-like surface, the linking number $L(K_1, K_2)$ is the sum of two geometric quantities: twist $T(K_1, K_2)$, and writhe $W(K_1)$.

$$L(K_1, K_2) = T(K_1, K_2) + W(K_1)$$

- This important characteristic together with the invariance of linking number have been applied to the study of circular DNA structure by Adams, 1994.

TWIST $T(K_1, K_2)$

- The twist $T(K_1, K_2)$ of one curve K_1 about another curve K_2 measures the magnitude of the spinning of K_1 around K_2 . The twist of helices about a linear axis is the number of times the helix (K_1) resolves about the axis (K_2). This number $T(K_1, K_2) > 0$ if the helix K_1 is right-handed and $T(K_1, K_2) < 0$ if the helix K_1 is left-handed ($T(K_1, K_2) = 1/2; -1/2; -1$)





TWIST $T(K_1, K_2)$

- For the more general cases in which K_2 is not linear, or planar, the definition of the twist is much more complex for the concept is no longer geometrically obvious. The twist of K_1 around K_2 is defined to be the measure of the total change of V in the direction of $T \times V$ as x moves along the entire curve K_2 . This is given by the line integral (normalized in turns) over the curve K_2 :

$$T(K_1, K_2) = \frac{1}{2\pi} \int_{K_2} (T \times V) \cdot dV$$

- This integral is not necessarily an integer. It changes under deformations of either the curve K_2 or the corresponding surface.
- Since the cross-product operation is not commutative, the twist depends on the ordering of the curves. The twist of K_1 about K_2 is not necessarily the twist of K_2 about K_1 .



WRITHE $W(K_1)$

- The writhing number of a curve K_1 , denoted by $W(K_1)$, is a knot property defined as the sum of crossings p of a curve K_1 ,

$$W(K_1) = \sum_{p \in C(K_1)} \varepsilon(p)$$

- where $\varepsilon(p)$ is defined to be ± 1 if the overpass slants from top left to bottom right or bottom left to top right and $C(K_1)$ is the set of crossings of an oriented curve
- The linking number $L(K_1, K_2)$ is a topological invariant. However the twist number $T(K_1, K_2)$ and writhing number $W(K_2)$ are not, and in fact, vary under deformation. Therefore, while the twist and a change in writhing could increase or increase linking, the linking number is invariant under deformation.



4.3. DNA KNOTS AND LINKS

- Geneticists have discovered that DNA can form knots and links which can be described mathematically.
- By understanding knot theory more completely, scientists are becoming more able to comprehend the massive complexity involved in the life and reproduction of the cell.
- The particular fascination in this process for geneticists is the fact that chemical changes occur in the DNA strand as a result of this process.
- Changes in the DNA structure due to the actions of these enzymes have required geneticists to use very advanced mathematical topology (which includes knot theory) and geometry in their study of molecular biology.



DESCRIPTIVE PROPERTIES ASSOCIATED WITH SUPERCOILING

- **"Supercoiling"** is an abstract mathematical property and represents the sum of what are termed "twist" and "writhe". "Supercoil" is the combination of twists and writhes that impart the supercoiling, and these occur in response to a change in the linking number.
- **Writhing:** The writhing number describes the supertwisting or supercoiling of the helix in space. It is the number of turns that the duplex axis makes about the superhelix axis. Writhe describes the coiling of the DNA coil. It is a measure of the DNA's superhelicity (supercoiling) and can be positive or negative. When a molecule is relaxed and contains no supercoils, the linking number = the twist number since $W = 0$. The linking number of relaxed DNA is $L = N/10.5$, where N is the number of base pairs in the DNA fragment.



DESCRIPTIVE PROPERTIES ASSOCIATED WITH SUPERCOILING

- **Twisting:** Twist is the number of helical turns in the DNA, i.e., the complete revolutions that one polynucleotide strand makes about the duplex axis in the particular conformation under consideration. Twist is normally the number of base pairs divided by 10.5. Twist is altered by deformation and is a local phenomenon. The total twist is the sum of all of the local twists. Twist is a measure of deformation due to a twisting motion.
- **Linking number:** This is a topological property that determines the degree of supercoiling. It defines the number of times a strand of DNA winds in the right-handed direction around the helix axis when the axis is constrained to lie in a plane. Topology theory indicates that the sum of T and W equals the linking number: $L = T + W$. If both strands are covalently intact, the linking number cannot change. Link is thus a topological invariant, remaining unaltered even if the two curves are deformed in space -- as long as neither is cut.



DESCRIPTIVE PROPERTIES ASSOCIATED WITH SUPERCOILING

- For example, in the circular DNA of 5400 base pairs, the linking number is $5400/10 = 540$.
- When a molecule is relaxed and contains no supercoils, the linking number = the twist number since $W = 0$. Thus if there is no supercoiling, then $W = 0$, $L = T + W = 540$.
- If there is positive supercoiling, $W = +20$, $T = L - W = 520$.



4.4 CHALLENGES AND PERSPECTIVES

In the area of DNA structure, several subareas are particularly amenable to mathematical analysis:

- A complete analysis of the packaging of DNA in chromatin. Only the first order coiling into core nucleosomes is understood. By far the largest compaction of DNA comes from higher order folding.
- Presentation of the topological invariants that describe the structure of DNA and its enzymatic transformations. The goal is to be able to predict the structure of interstate or products from enzymatic mechanisms and in turn to predict mechanisms from structure.
- An analysis of the reciprocal interaction between secondary and higher order structures. This includes the phenomena of bending, looping, and phasing.



4.4 CHALLENGES AND PERSPECTIVES

Many doubts and suspicions exist in understanding of the genetic language.

- How was life information accumulated and evolved in the DNA sequence?
- How can we understand the possible function of the large amount of nongenic DNA in the genome and extract life information from DNA sequence under the background of strong noises?
- What is the principle that governs the functional networks in a genome?
- How can we predict the molecular structure from its sequence information?



Part I Genetic Codes, Biological Sequences, DNA and Protein Structures

5. Protein Structures, Geometry, and Topology

- ❖ Introduction
- ❖ Computational Geometry and Topology
- ❖ Protein Structures and Prediction
- ❖ Statistical Approach and Discussions
- ❖ Challenges and Perspectives



5.1 INTRODUCTION

Proteins play crucial roles in almost every biological process:

- Responsible in one form or another for a variety of physiological functions,
- Function as catalysts,
- Transport and store other molecules such as oxygen,
- Provide mechanical support and immune protection,
- Generate movement,
- Transmit nerve impulses,
- Control growth and differentiation.



5.1 INTRODUCTION

- They perform many vital functions, e.g.:
 - Catalysis of reactions
 - Transport of molecules
 - Building blocks of muscles
 - Storage of energy
 - Defense against intruders
- They are large molecules—containing 100s to 1000s atoms.
- They are made of *amino acids*.
 - There are 20 different types of amino acids.



5.2 COMPUTATIONAL GEOMETRY AND TOPOLOGY

Computational Geometry

- The study of efficient algorithms to solve geometric problems, such as given N points in a plane, what is the fastest way to find the nearest neighbor of a point? Given N straight lines, find the lines which intersect with each other.
- Many questions in molecular modeling can be understood geometrically in terms of arrangements of spheres in three dimensions.



5.3 COMPUTATIONAL GEOMETRY AND TOPOLOGY PRELIMINARIES

Computational Geometry

- Problems include computing properties of such arrangements such as their volume and topology, testing intersections and collisions between molecules, finding offset surfaces, data structures for computing inter-atomic forces and performing molecular dynamics simulations, and computer graphics algorithms for rendering molecular models accurately and efficiently.
- Computational geometry can be also used as a tool for studying topology and architecture of macromolecules and macromolecular complexes.



FUNDAMENTAL GEOMETRIC OBJECTS

- **Polygons:** A polygon is a collection of line segments, forming a cycle, and not crossing each other. A polygon can be represented as a sequence of points.
- **Convex Hull:** The convex hull of a set of points S in n dimensions is the intersection of all convex sets containing S .
 - Finding the convex hull of a set of points is *the* most elementarily interesting problem in computational geometry, just as the minimum spanning tree is the most elementarily interesting problem in graph algorithms.
 - Novel patterns based on convex hull representation are firstly extracted from a protein structure, then the classification system is constructed and machine learning methods such as neural networks and Hidden Markov Models (HMM) have been applied.



FUNDAMENTAL GEOMETRIC OBJECTS

- **Triangulation:** Triangulation is the division of a surface or plane polygon into a set of triangles, usually with the restriction that each triangle side is entirely shared by two adjacent triangles.
 - Triangulation is a fundamental problem in computational geometry, because the first step in working with complicated geometric objects is to break them into simple geometric objects.
 - The simplest geometric objects are triangles in two dimensions, and tetrahedra in three.
 - Classical applications of triangulation include finite element analysis and computer graphics. Recently, triangulation has been applied to the computation of molecular surface by Ryu, et al in 2007 and 2009).
 - Molecular surface is used for both the visualization of the molecule and the computation of various molecular properties such as the area and volume of a protein, which are important for studying problems such as protein docking and folding.



FUNDAMENTAL GEOMETRIC OBJECTS

- **Nearest-neighbor search:** Nearest-neighbor search (or similarity search) is a search to quickly find the nearest neighbor to a query point; that is, given a set S of n points in d dimensions, and a query point q , which point in S is closest to q ?
 - The nearest-neighbor search has been used to approximate the protein structure by Lotan and Schwarzer, 2004.
- **Shape similarity:** Shape similarity is a problem that underlies much of pattern recognition. Given two polygonal shapes, P_1 and P_2 , how similar are P_1 and P_2 ? Definition of similarity is application dependent.
 - The shape similarity measures are widely used in the protein structure comparison and prediction by Lotan and Schwarzer, 2004; Sael et al, 2008.

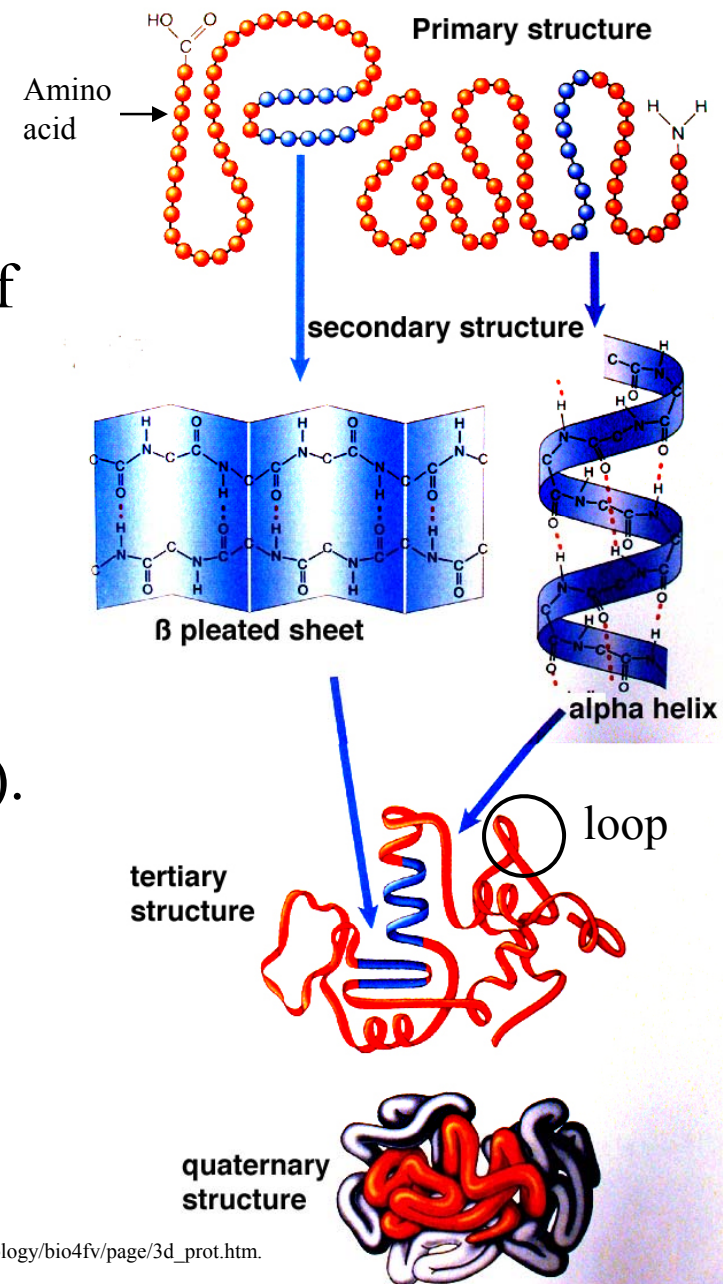


FUNDAMENTAL GEOMETRIC OBJECTS

- **Topology** is a branch of mathematics. It can be defined as "the study of qualitative properties of certain objects (called topological spaces) that are invariant under certain kind of transformations (called continuous maps), especially those properties that are invariant under a certain kind of equivalence (called homeomorphism)." The mathematical definition of topology is briefly described here.
- Let \mathbf{X} be any set and let \mathbf{T} be a family of subsets of \mathbf{X} . Then \mathbf{T} is a topology on \mathbf{X} if
 - Both the empty set and \mathbf{X} are elements of \mathbf{T} .
 - Any union of arbitrarily many elements of \mathbf{T} is an element of \mathbf{T} .
 - Any intersection of finitely many elements of \mathbf{T} is an element of \mathbf{T} .
 - If \mathbf{T} is a topology on \mathbf{X} , then \mathbf{X} together with \mathbf{T} is called a topological space.
- DNA topology and protein topology are active research areas.

5.3 PROTEIN STRUCTURES AND PREDICTION

- A protein has one or a few chains of amino acids.
- A chain of amino acids folds into a 3D structure.
 - Some substructures are regular helix shape (*alpha helix*), or instant noodle shape (*beta sheet*).
 - The rest are irregular shape, called *loops*.
- Chains aggregate together into a bigger 3D structure.





SECONDARY STRUCTURE PREDICTION

Information Theoretic Approach:

- The prediction of protein secondary structure from its amino acid sequence can be considered as the problem of finding the correlation between the two objects. It can be studied in the framework of information theory.
- The amino acid sequence can be regarded as an information source. The corresponding secondary structure can be considered as an information receiver. For an amino acid sequence of length N one can construct a secondary structure sequence of the same length written by three letters α , β , and c following the one-to-one correspondence between residue and secondary structure.

SECONDARY STRUCTURE PREDICTION

- Let $p(a_i)$ be the probability of structure a_i in the secondary structure sequence ($a_i = \alpha, \beta, c$) and let $p(s_i)$ be the probability of amino acid s_i in the protein ($j = 1, 2, \dots, 20$). Define average mutual information

$$I(X;Y) = H(X) - H(X|Y) = -\sum_i p(a_i) \log p(a_i) + \sum_i \sum_j p(s_i) p(a_i | s_i) \log p(a_i | s_i)$$

Similarly, we can also define

$$I(Y;X) = H(Y) - H(Y|X) = -\sum_j p(s_j) \log p(s_j) + \sum_i \sum_j p(a_i) p(s_i | a_i) \log p(s_i | a_i)$$

It is easy to prove that

$$I(X;Y) = I(Y;X)$$

- The maximum of $H(X|Y)$ is $H(X)$ which corresponds to no correlation between X and Y. So the correlation between secondary structure (X) and amino acid (Y) is defined by

$$r_1 = \frac{I(X;Y)}{H(X)}, (a_i = \alpha, \beta, c; s_j = A, C, \dots, W, Y)$$

where r_1 takes values between 0 and 1:



INFORMATION THEORETIC APPROACH

- $r_1=0$ means no correlation;
- $r_1=1$ means the full determination of secondary structure by amino acid, this occurs in the case of $p(a_i|s_j)=0$ or 1 for all a_i and s_j .
- The single peptide-structure correspondence can be easily extended to di-peptide (tri-peptide)-structure correspondence through residue numeration by shifting a window of width 2 (3). The above equations can be generalized in these cases. For the case of di-peptide-structure correspondence a_i takes 9 confirmations, that is

$$\alpha\alpha, \alpha\beta, \alpha\gamma, \beta\alpha, \beta\beta, \beta\gamma, \gamma\alpha, \gamma\beta, \gamma\gamma.$$

s_j takes 400 di-peptides in the above equations, that is,

$$AA, AC, \dots, WY, YY.$$



INFORMATION THEORETIC APPROACH

- The correlation between secondary structure and neighboring di-peptide can be defined by

$$r_2 = \frac{I(X;Y)}{H(X)}$$

- The correlation between secondary structure and tri-peptide can be defined by

$$r_3 = \frac{I(X;Y)}{H(X)}, (a_i = \alpha\alpha\alpha, \alpha\alpha\beta, \dots ccc; s_j = AAA, AAC, \dots WYY, YYYY)$$

- It can be demonstrated that the correlation of protein secondary structure with di-peptide frequency is much stronger than that with single peptide and the correlation with tri-peptide frequency is much stronger than that with di-peptide. Therefore, the prediction of protein secondary structure from di-peptide and tri-peptide distribution is a better approach than single peptide prediction. Thus, the information theoretic approach provides a method to estimate the efficiency of a structural prediction. The averaged mutual information $I(X:Y)$ is a useful quantity for the estimate.



TERTIARY STRUCTURE PREDICTION: POTENTIAL ENERGY SURFACE DEFINED BY FORCE FIELDS

Molecular Mechanics :

- Consider a molecule with N atoms. The position of the i -th atom is denoted by the vector x_i .
- Describe the potential energy surface of a protein by molecular mechanics.
- Molecular mechanics states that the potential energy of a protein can be approximated by the potential energy of the nuclei. Therefore, the energy contribution of the electrons is neglected.
- This approximation allows one to write the potential energy of a protein as a function of the nuclear coordinates.



TERTIARY STRUCTURE PREDICTION: POTENTIAL ENERGY SURFACE DEFINED BY FORCE FIELDS

Molecular Mechanics :

- A typical molecular modeling force field contains five types of potentials. These potentials correspond to deformation of
 - Covalent bond length
 - Bond angles,
 - Torsional motion associated with rotation about bonds,
 - Electrostatic interaction,
 - van der Waals interaction.

$$V(x) = V_{length} + V_{angle} + V_{torsion} + V_{electrostatic} + V_{weak}$$

- The potential energy $V=V(x)$ is a function of the atomic coordinate x of the molecule. The distance is measured in Ångstrom (Å), energy in kcal/mol, and mass in atomic mass unit (Dalton).



TERTIARY STRUCTURE PREDICTION: POTENTIAL ENERGY SURFACE DEFINED BY FORCE FIELDS

- The bond length potential is given by

$$V_{lengh} = \sum_{\substack{i,j \\ bonds}} k_0 (r_{ij} - r_0)^2$$

- Where $r_{ij} = ||x_i - x_j||$ is the bond length, r_0 is the reference bond length, and k_0 is a force constant. Reference bond lengths and force constants depend on the bond type. The bond potential corresponds to covalent bond deformation. The bond length deformations are sufficiently small at ordinary temperatures and in the absence of chemical reactions. The bond deformation energy between the i-th and j-th atom is given by a harmonic potential

$$k_0 (r_{ij} - r_0)^2$$



TERTIARY STRUCTURE PREDICTION: POTENTIAL ENERGY SURFACE DEFINED BY FORCE FIELDS

- The bond angle potential is given by

$$V_{angle} = \sum_{\theta_{angle}} k_0 (\theta - \theta_0)^2$$

- Where θ_0 is the reference bond angle and k_0 is a force constant. Reference bond angle and force constant depend on the type of atom involved. The angle θ between the bonds $\mathbf{p} = \mathbf{x}_j - \mathbf{x}_i$ and $\mathbf{r} = \mathbf{x}_k - \mathbf{x}_j$ is given by

$$\cos(\theta) = \frac{\mathbf{p} \cdot \mathbf{r}}{\|\mathbf{p}\| \|\mathbf{r}\|}, \theta \in [0, \pi].$$

- The bond angle potential corresponds to angle deformation. Bond angle deformations are sufficiently small at ordinary temperatures and in the absence of chemical reactions.



TERTIARY STRUCTURE PREDICTION: POTENTIAL ENERGY SURFACE DEFINED BY FORCE FIELDS

- The potentials for bond length and bond angle deformation are considered as the hard degrees of freedom in a molecular system in the sense that considerable energy is necessary to cause significant deformation from their reference values. The most variation in structure and relative energy comes from the remaining potential energy terms.
- The torsion potential corresponds to the barriers of bond rotation which involves the dihedral angles of the rotatable bonds. The barriers of torsion can be expressed as a series of cosine functions. The mathematical expression for the torsion potential is given by

$$V_{torsion} = \sum_{\theta::dihedral} |k_0| - k_0 \cos(n_0 \theta)^2$$

where n_0 is the multiplicity of the angle and k_0 is a force constant. Both multiplicity and force constants depend on the type of atoms involved. The dihedral angle θ can be obtained from



TERTIARY STRUCTURE PREDICTION: POTENTIAL ENERGY SURFACE DEFINED BY FORCE FIELDS

where n_θ is the multiplicity of the angle and k_θ is a force constant. Both multiplicity and force constants depend on the type of atoms involved. The dihedral angle θ can be obtained from

$$\cos(\theta) = \frac{|(p \times r) \cdot (r \times q)|}{\|p \times r\| \|r \times q\|}, \theta \in [-\pi, \pi]$$

where

$$p = x_j - x_i, r = x_k - x_j, q = x_l - x_k$$

and the sign of the angle θ is given by the sign of the inner product $(p \times q) \cdot r$.
The complementary angle $\pi - \theta$ is the torsion angle of the bond $x_j - x_k$.



TERTIARY STRUCTURE PREDICTION: POTENTIAL ENERGY SURFACE DEFINED BY FORCE FIELDS

- The electrostatic potential corresponds to the nonbounded interaction between the charged atoms in a molecule. The interaction is attractive when the charges have opposite sign and repulsive when the charges have the same sign. The electrostatic potential of a molecule is given by

$$V_{electrostatic} = \sum_{\substack{i < j \\ \text{atoms}}} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

where q_i is the point charge of the i -th atom and ϵ_0 is the dielectric constant of vacuum, and r_{ij} is the distance between i -th and j -th atoms.

TERTIARY STRUCTURE PREDICTION: POTENTIAL ENERGY SURFACE DEFINED BY FORCE FIELDS

- The van der Waals potential corresponds to the interaction between nonbounded atoms in a molecule. This interaction comes from attractive and repulsive forces. The van der Waals potential is given by

$$V_{weak} = \sum_{\substack{i < j \\ atoms}} \left[\left(\frac{A_{ij}}{r_{ij}^{12}} \right) - \left(\frac{B_{ij}}{r_{ij}^6} \right) \right]$$

Where A_{ij} and B_{ij} are given by

$$A_{ij} = \frac{1}{2} B_{ij} (R_i + R_j)^6$$

$$B_{ij} = \frac{3}{2} \frac{1}{\sqrt{4\pi\delta_0}} \frac{1}{\sqrt{m_e}} \frac{e\hbar\alpha_i\alpha_j}{\sqrt{\alpha_i/N_i} + \sqrt{\alpha_j/N_j}}$$

where e is the electron charge, \hbar is the reduced Planck constant, m_e is the electron mass, α_i is the polarizability of the i th atom, N_i is the effective number of outer shell electrons in the i th atom. R_i is the van der Waals radius of the i -th atom.



CONFORMATIONAL SEARCH METHODS

- The objective of conformational search is to find all preferred conformations of a molecule.
- The conformational search of the global minimum energy surface of a protein from the amino acid sequence is one of the challenging problems in bioinformatics.
- In recent years, several optimization approaches to solve this problem have appeared in the literature. The most common approach is to model the protein surface by using a force field.
- The general scheme is to define a smooth operator that is linear and each term of the potentials can be separately smoothed.



THE PROCESS OF SMOOTHING THE TORSION POTENTIAL OF A PROTEIN

- Express the dihedral angles by distances. We assume that bond lengths and bond angles are fixed to their reference values. Then the cosine of a dihedral angle θ can be expressed by the distance $r = \|x_l - x_i\|$ of the first and last of the involved atoms:

$$\cos(\theta) = \alpha + \beta r^2$$

- where α and β are constants depending on the reference bond lengths and reference bond angles. In general $\cos(n\theta)$ of a multiple dihedral angle can be represented as a Chebyshev polynomial in $\cos(\theta)$, which is a polynomial in r^2 .

Let $x = \cos(\theta)$, then the Chebyshev polynomials can be written as

$$T_n(x) = \cos(n\theta) = \cos(n \arccos(x))$$



THE PROCESS OF SMOOTHING THE TORSION POTENTIAL OF A PROTEIN

- Furthermore, we have

$$T_n(x) = \cos(n\theta) = T_n(\alpha + \beta r^2)$$

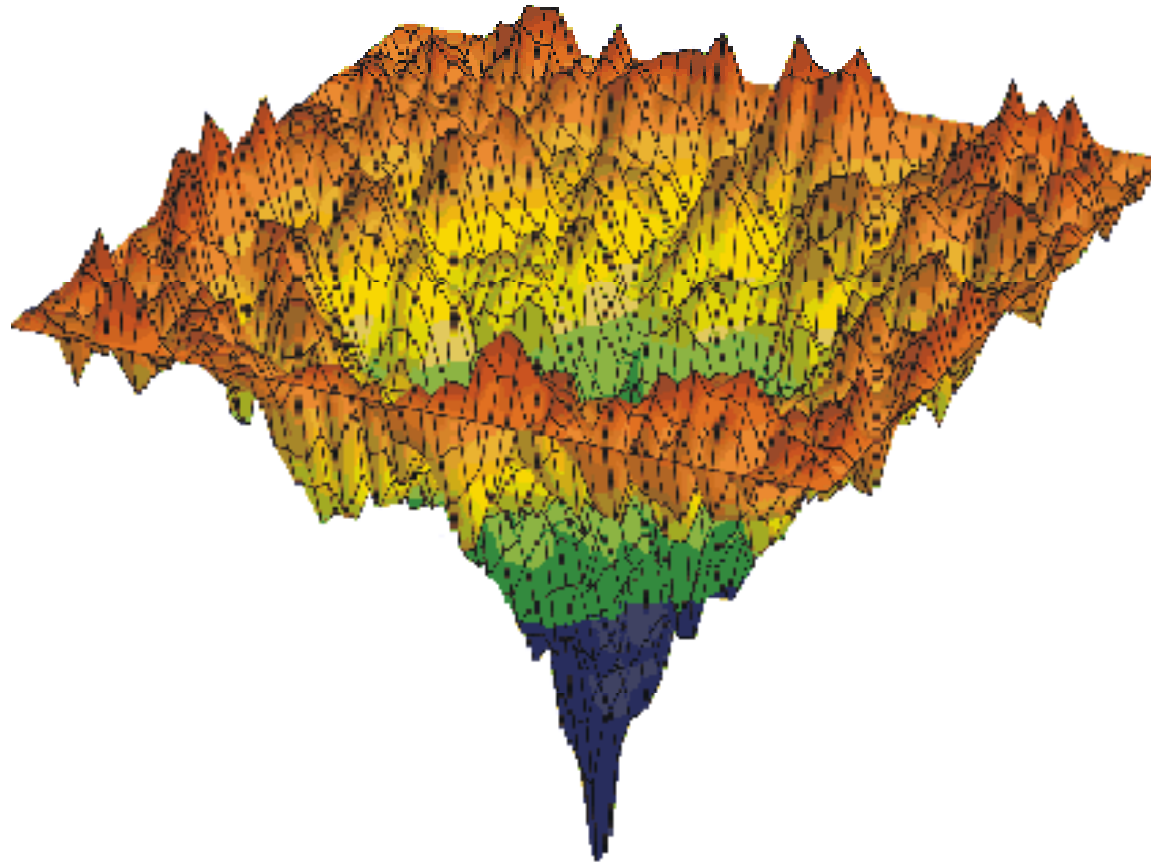
- Consequentially, the torsion potential can be expressed as a linear combination of Chebyshev polynomials

$$V_{torsion} = \sum_{\theta: \text{dihedral}} k_0 | -k_0 T_n(\alpha + \beta r^2)$$

- Each term is a polynomial in r^2 and so the torsion potential $V_{torsion}(x)$ can be smoothed by the linear operator Ψ_t ,

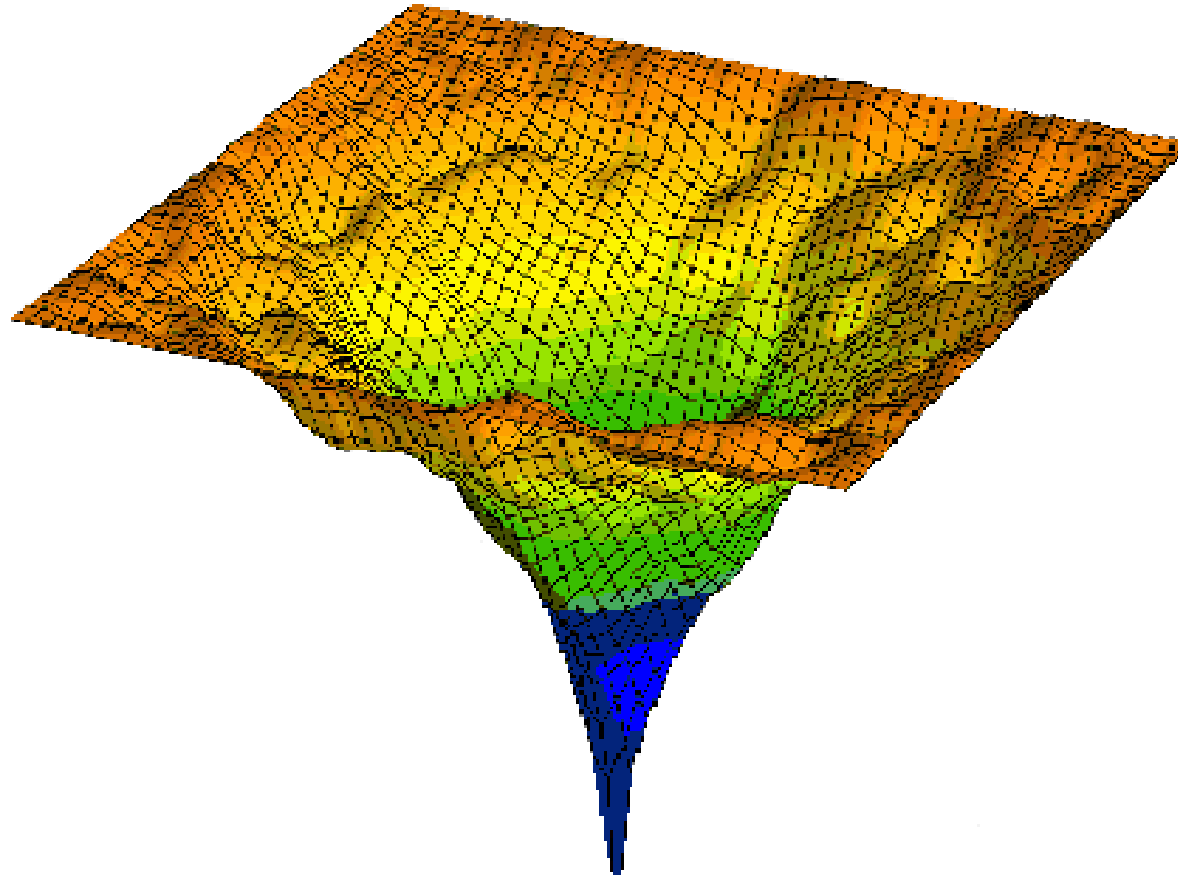
$$\tilde{V}_{torsion}(x, t) = \Psi_t V_{torsion}(x)$$

- The potential energy surface of a protein and smoothed potential energy surface of protein are illustrated below



Potential energy surface of protein

The process of smoothing the torsion potential of a protein



Smoothed potential energy surface of protein
The process of smoothing the torsion potential of a protein



5.4 STATISTICAL APPROACH AND DISCUSSIONS

Fold Recognition

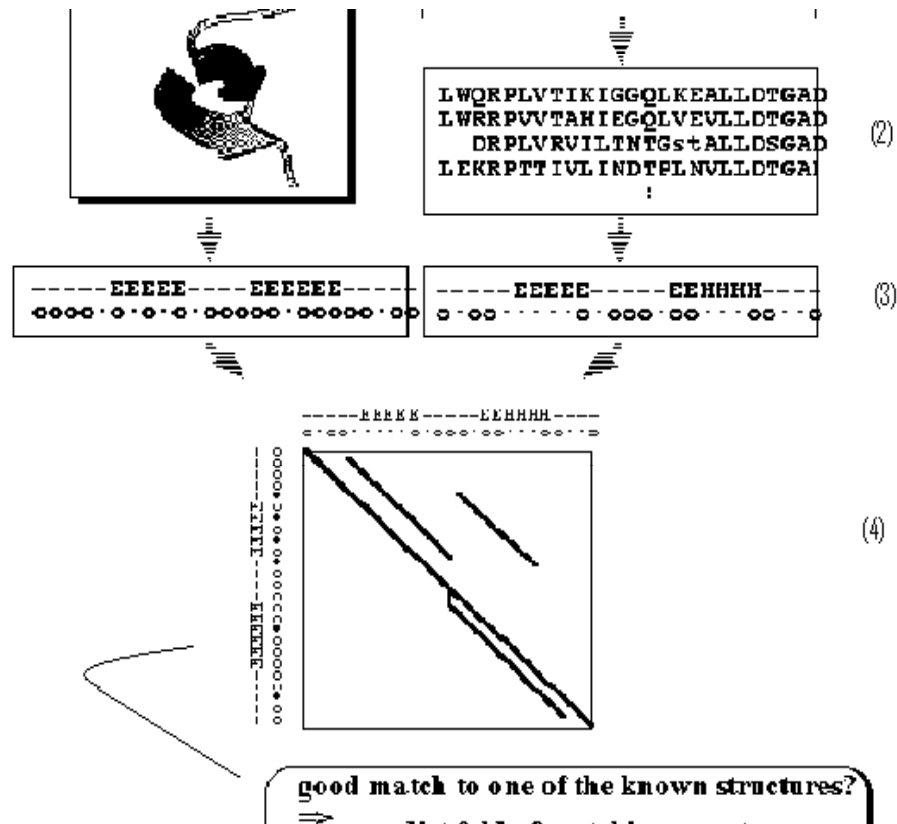
- Proteins may have similar tertiary structures even if their primary structures are not sufficiently similar or different.
- This observation has led to the hypothesis that there are only a limited number of significantly distinct tertiary structures.
- The main goal of fold recognition is to predict the tertiary structure of a protein from its amino acid sequence by finding the best match between the amino acid sequence and some tertiary structure in a protein database.
- A basic approach to fold recognition is comparative modeling.



5.4 STATISTICAL APPROACH AND DISCUSSIONS

Fold Recognition

- Let **A** be the amino acid sequence of a protein with unknown tertiary structure, align the sequence **A** to the primary structures of all proteins in the database of tertiary protein structures. Suppose the sequence **A** best aligns to the primary structure of **B**. This sequence alignment can be used to infer the structural alignment. For example, if the residue a_i of **A** aligns with the residue b_j of **B**, then the position of the residue a_i in the unknown tertiary structure is defined as the position of the residue b_j in the tertiary structure in the database. Subsequences of the sequence of **A** aligned with a series of blanks of the sequence of **B** are modeled as coil region.



Fold Recognition

Threading predicted 1D structure profiles into known 3D structures: (1) Input sequence; (2) Generate sequence alignment; (3) Predict 1D structure; (4) Align predicted and known structure(s)



3D PROFILE-SEQUENCE ALIGNMENT

- A more sophisticated approach to fold recognition makes use of the method of 3D profile-sequence alignment. For this, we make use of both sequence database and protein database.
- Let **A** be a sequence of amino acid and **P** be the 3D profile of a protein. We align **A** to **P**.
- Let $\sigma(P, A)$ be the corresponding alignment score. To estimate the significance of these alignment scores, we align the protein with 3D profile **P** against all amino acid sequences of a sequence database.
- The Z score for aligning the amino acid sequence **A** to the protein with 3D profile **P** is given by

$$Z(P, A) = \frac{\sigma(P, A) - \mu(P)}{\sigma(P)}$$



3D PROFILE-SEQUENCE ALIGNMENT

where $\mu(P)$ is the mean score of alignment scores given by

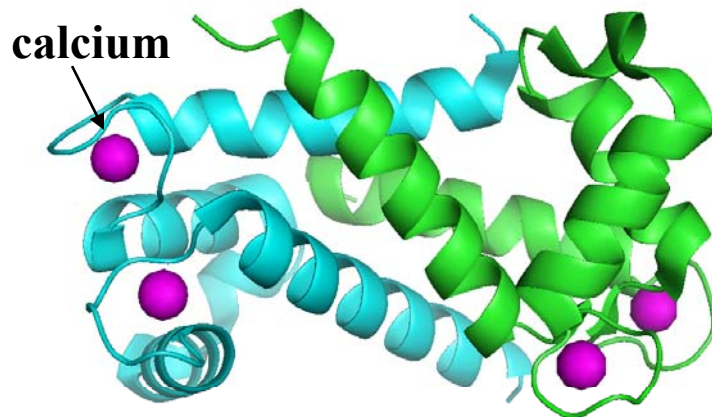
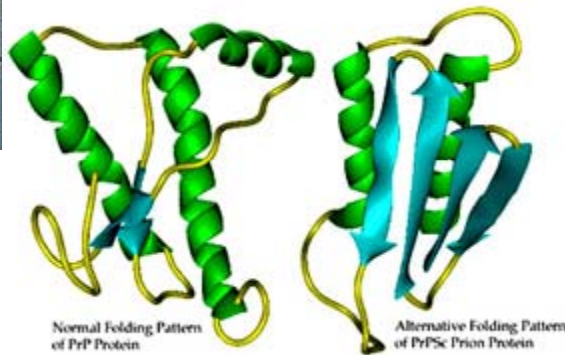
$$\mu(P) = \frac{1}{M} \sum_A \sigma(P, A)$$

with M as the number of sequences in the sequence database, and $\sigma(P)$ is the standard deviation of the scores given by

$$\sigma(P) = \sqrt{\frac{1}{M} \sum_A (\sigma(P, a) - \mu(P))^2}$$

A high Z score $Z(P, Z)$ may indicate that amino acid sequence A has similar tertiary structure as the protein with the 3D profile P.

5.5. CHALLENGES AND PERSPECTIVES



- The sequences of similar structures in PDB, how to identify the correct templates and how to refine the template structure closer to the native.
- The sequences without appropriate templates, how to build models of correct topology from scratch.
- Protein function is closely related to its 3D structure and applications.
 - E.g., mad-cow disease is due to PrP misfolding.
 - E.g., calcium atoms bind to good-shape-loops.