Special Topics in Computational Biology

Lecture #1: Introduction

Bud Mishra
Professor of Computer Science and Mathematics
1 ¦ 25 ¦ 2001

Syllabus

Introductory Material
- What do we know?
- Biological information
- Biotechnology (e.g. arrays, PCR, hybridization; single molecules; mass spectrometry)
- Some biology (terminology)

Population Genetics
- Diseases
- Linkage analysis
- Kinship analysis

Comparative Genomics
- Phylogeny
- Gene rearrangements between species
- Gene families within specie

Functional Genomics
- Taking cells at different stages of development, what can we infer from gene expression levels data? Can we determine the sequence of gene activation? Tools that allow biologists to try to answer these questions.)
- Genetic Networks
- Clustering algorithms

Proteomics

Cancer Genomics
- (What can be done here)

Introduction to Biology

Genome:

Hereditary information of an organism is encoded in its DNA and enclosed in a cell (unless it is a virus). All the information contained in the DNA of a single organism is its *genome*.

DNA molecule can be thought of as a *very* long sequence of **nucleotides** or **bases**:

= {A, T, C, G}

Complementarity

DNA is a double-stranded polymer and should be thought of as a pair of sequences over . However, there is a relation of **complementarity** between the two sequences:

A , T, C , G

That is if there is an A (respectively, T, C, G) on one sequence at a particular position then the other sequence must have a T (respectively, A, G, C) at the same position.

We will measure the sequence length (or the DNA length) in terms of **base pairs (bp)**: for instance, human (H. sapiens) DNA is 3.3 £ 10^9 bp measuring about 6 ft of DNA polymer completely stretched out!

Genome Size

The genomes vary widely in size: measuring from »

Few thousand base pairs for viruses to 2 » 3 £ 10^{11}bp for certain amphibian and flowering plants.

Coliphage MS2 (a virus) has the smallest genome: only 3.5 £ 10^3bp.

Mycoplasmas (a unicellular organism) has the smallest cellular genome: 5 £ 10^5bp.
C. elegans (nematode worm, a primitive multicellular organism) has a genome of size » 10^8bp.

Goal of a Genome Study
E.g. Human Genome Project

Genetic Maps:
Physical Maps: (For instance, the Human Genome Project [HGP] requires a complete map of the human genome at a resolution of 100 Kb = 10^5bp. That is, the map would consist of "markers" spaced at most 10^5bp apart.)
DNA Sequencing:
Gene Identification: Identify genes (parts of the DNA involved in controlling the metabolic processes through proteins they encode) on physical maps or sequenced DNA.
Informatics: Elucidate the structure of the DNA as encoding of all the relevant information.

Diagnostic and Therapeutic Tools: Necessary for the treatment of genetic diseases.
Phylogenetic Tools: Used in understanding the process and mechanism of evolution.

DNA) Structure and Components

The usual configuration of DNA is in terms of a **double helix** consisting of two **chains** or **strands** coiling around each other with two alternating grooves of slighltly different spacing.
The "backbone" in each strand is made of alternating big sugar molecules (Deoxyribose residues: $C_5\ O_4\ H_{10}$) and small phosphate ($(P\ O_4)^{-3}$) molecules.
Now, one of the four bases (the letters in our alphabet), each one an almost planar nitrogenic organic compound, is connected to the sugar molecule. The bases are:

Adenine) A
Thymine) T
Cytosine) C
Guanine) G

DNA) Structure and Components (contd.)

The sequence of bases defines the information encoded by the DNA.
Complementary base pairs (A-T and C-G) are connected by hydrogen bonds and the base-pair forms a coplanar "rung" connecting the two strands.

Cytosine and **thymine** are smaller (lighter) molecules, called **pyrimidines**
Guanine and **adenine** are bigger (bulkier) molecules, called **purines**.
Adenine and thymine allow only for double hydrogen bonding, while cytosine and guanine allow for triple hydrogen bonding.

Thus the chemical (through hydrogen bonding) and the mechanical (purine to pyrimidine) constraints on the pairing lead to the complementarity and makes the double stranded DNA both **chemically inert and mechanically quite rigid and stable.**

DNA) Structure and Components (contd.)

The building blocks of the DNA molecule are four kinds of **deoxyribonucleotides**,
where each deoxyribonucleotide is made up of **a sugar residue, a phosphate group and a base**.
From these building blocks (or related, **dNTPs** deoxyribonucleoside triphosphates) one can synthesize a strand of DNA.
The sugar molecule in the strand is in the shape of a pentagon (4 carbons and 1 oxygen) in a plane parallel to the helix axis and with the 5th carbon (5' C) sticking out.
The phosphodiester bond (-O-P-O-) between the sugars connects this 5' C to a carbon in the pentagon (3' C) and provides a directionality to each strand.
The strands in a double-stranded DNA molecule are **antiparallel**.
Most of the enzymes moving along the backbone moves in the 5'-3' direction.

The Central Dogma

The intermediate molecule carrying the information out of the nucleus of an eukaryotic cell is RNA, a single stranded polymer.
RNA also controls the translation process in which amino acids are created making up the proteins.
The central dogma(due to Francis Crick in 1958) states that these information flows are all unidirectional:

"The central dogma states that once `information' has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible.

Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein."

RNA and Transcription

The polymer RNA (**ribonucleic acid**) is similar to DNA but differ in several ways:

it's single stranded;

its nucleotide has a ribose sugar (instead of deoxyribose) and

it has the pyrimidine base *uracil*, U, substituting *thymine*, T-- U is complementary to A like thymine.

RNA molecule tends to fold back on itself to make helical twisted and rigid segments.

For instance, if a segment of an RNA is

5' - GGGGAAAACCCC - 3',

then the C's fold back on the G's to make a hairpin structure (with a 4bp **stem** and a 5bp **loop**.

The secondary RNA structure can even be more complicated, for instance, in case of *E. coli*, **Ala tRNA** (transfer RNA) forms a cloverleaf shape.

Prediction of RNA structure is an interesting computational problem.

RNA, Genes and Promoters

A specific region of DNA that determines the synthesis of proteins (through the transcription and translation) is called a **gene**

Originally, a gene meant something more abstract---a unit of hereditary inheritance.

Now a gene has been given a physical molecular existence.

Transcription of a gene to a **messenger RNA**, **mRNA**, is keyed by an RNA polymerase enzyme, which attaches to a **core promoter** (a specific sequence adjacent to the gene).

Regulatory sequences such as **silencers** and **enhancers** control the rate of transcription

by their influence on the RNA polymerase through a feedback control loop involving many large families of **activator** and **repressor** proteins that bind with DNA and

which in turn, transpond the RNA polymerase by **coactivator proteins** and **basal factors**.

Transcriptional Regulation of Gene

The entire structure of transcriptional regulation of gene expression is rather dispersed and fairly complicated:

The enhancer and silencer sequences occur over a wide region spanning many Kb's from the core promoter on either directions;

A gene may have many silencers and enhancers and can be shared among the genes;

They are not unique---different genes may have different combinations;

The proteins involved in control of the RNA polymerase number around 50 and

Different cliques of transcriptional factors operate in different cliques.

Any disorder in their proper operation can lead to cancer, immune disorder, heart disease, etc.

Transcription

The transcription of DNA in to mRNA is performed with a single strand of DNA (**the sense strand**) around a gene.

The double helix

Untwists momentarily to create a **transcriptional bubble** which moves along the DNA in the 3' - 5' direction (of the sense strand)

As the complementary mRNA synthesis progresses adding one RNA nucleotide at a time at the 3' end of the RNA, attaching an U (respectively, A, G and C) for the corresponding DNA base of A (respectively, T, C and G),

Ending when a termination signal (a special sequence) is encountered.

This newly synthesized mRNA are **capped** by attaching special nucleotide sequences to the 5' and 3' ends.

This molecule is called a **pre-mRNA**.

Exons and Introns

In eukaryotic cells, the region of DNA transcribed into a pre-mRNA involves more than just the information needed to synthesize the proteins.

The DNA containing the code for protein are the **exons**, which are interrupted by the **introns**, the non-coding regions.

Thus pre-mRNA contains both exons and introns and is altered to excise all the intronic subsequences in preparation for the translation process---this is done by the **spliceosome**.
The location of splice sites, separating the introns and exons, is dictated by short sequences and simple rules such as

"introns begin with the dinucleotide GT and end with the dinucleotide AG" (**the GT-AG rule**).

Protein and Translation

The translation process begins at a particular location of the mRNA called the translation start sequence (usuall **AUG**) and is mediated by the transfer RNA (tRNA), made up of a group of small RNA molecules, each with specifity for a particular amino acid.
The tRNA's carry the amino acids to the ribosomes, the site of protein synthesis, where they are attached to a growing polypeptide.
The translation stops when one of the three trinucleotides **UAA, UAG or UGA** is encountered.
Each 3 consecutive (nonoverlapping) bases of mRNA (corresponding to a codon codes for a specific amino acid.
There are $4^3 = 64$ possible trinucleotide **codons** belonging to the set

$$\{U, A, G, C\}^3$$

Genetic Codes

The codon AUG is the **start codon** and the codons UAA, UAG and UGA are the **stop codons**.

That leaves 60 codons to code for 20 amino acids with an expected redundancy of 3!
Multiple codons (one to six) are used to code a single amino acid.

The line of nucleotides between and including the start and stop codons is called an **open reading frame (ORF)**
All the information of interest to us resides in the ORF's.
The mapping from the codons to amino acid (and naturally extended to a mapping from ORF's polypeptides by a homomorphism) given by

$$F_P : \{U, A, G, C\}^3 \; ! \; \{A, R, D, N, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$$

Amino Acids with Codes

A	Ala	alanine	GC(U+A+C+G)
C	Cys	cysteine	UG(U+C)
D	Asp	aspertic acid	GA(U+C)
E	Glu	glutamic acid	GA(G+A)
F	Phe	phenylanine	UU(U+C)
G	Gly	glycine	GG(U+A+C+G)
H	His	histine	CA(U+C)
I	Ile	isoleucine	AU(U+A+C)
K	Lys	lysine	AA(A+G)
L	Leu	leucine	(C+U)U(A+G) + CU(U+C)
M	Met	methionine	AUG
N	Asn	asparginine	AA(U+C)
P	Pro	proline	CC(U+A+C+G)
Q	Gln	glutamine	CA(A+G)
R	Arg	arginine	(A+C)G(A+G)+CG(U+C)
S	Ser	serine	(AG+UC)(U+C)+UC(A+G)
T	Thr	threonine	AC(U+A+C+G)
V	Val	valine	GU(U+A+C+G)
W	Trp	tryptophan	UGG
Y	Tyr	tyrosine	UA(U+C)