

## Lecture 5

# Biological background review

MCB 416A/516A

Statistical Bioinformatics and Genomic Analysis

Prof. Lingling An

Univ of Arizona

# Outline

---

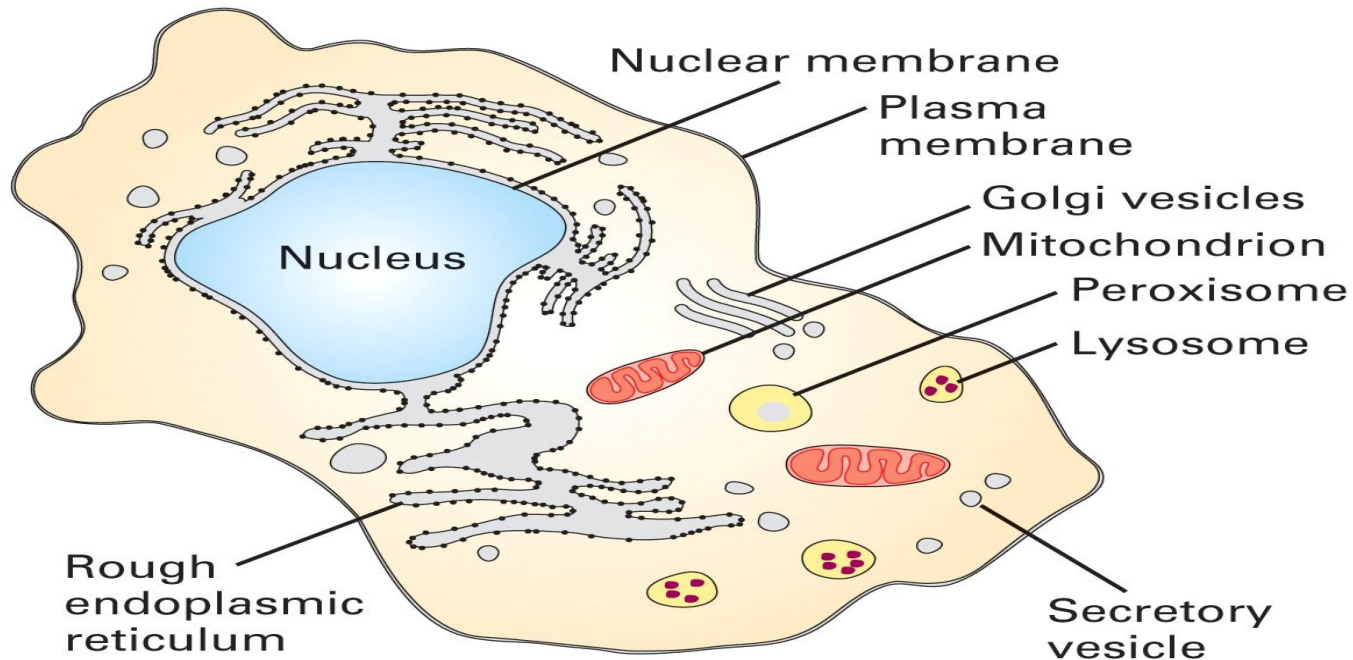
- Introduction to molecular biology
  - What is life made of?
  - DNA
  - RNA
  - Protein
  - How do individuals of a species differ?
  - Why Bioinformatics?

# What is Life made of?

---

- All living things are made of Cells
  - *Prokaryotic, Eukaryotic*
- Cells:
  - Chemical composition-by weight
    - ◆ 70% water
    - ◆ 7% small molecules
      - e.g., salts, amino acids, nucleotides
    - ◆ 23% macromolecules
      - e.g., Proteins
  - biochemical (metabolic) pathways
  - translation of mRNA into proteins

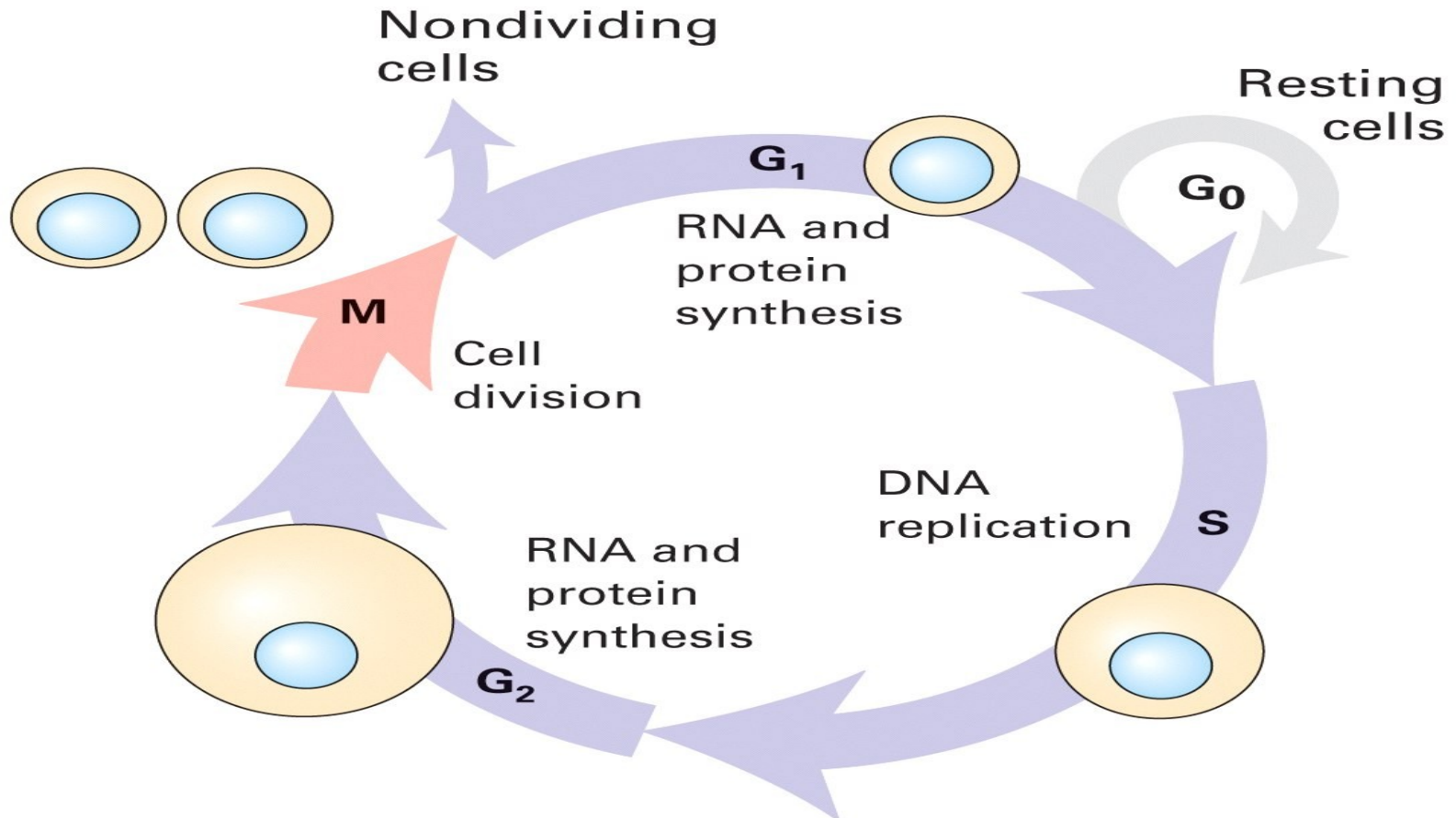
# Life begins with Cell



- A cell is a smallest structural unit of an organism that is capable of independent functioning
- All eukaryotic cells have some common features

# All Cells have common Cycles

---



- Born, eat, replicate, and die

# Overview of organizations of life

---

- **Nucleus = library**
- **Chromosomes = bookshelves**
- **Genes = books**
- Almost every cell in an organism contains the same libraries and the same sets of books.
- Books represent all the information (DNA) that every cell in the body needs so it can grow and carry out its various functions.

# Some Terminology

---

- **Gene**: a discrete units of hereditary information located on the chromosomes and consisting of DNA.
- **Genome**: an organism's genetic material
- **Genotype**: The genetic makeup of an organism
- **Phenotype**: the physical expressed traits of an organism
- **Nucleic acid**: Biological molecules(RNA and DNA) that allow organisms to reproduce;

# More Terminology

---

- The **genome** is an organism's complete set of DNA.
  - a bacteria contains about 600,000 DNA base pairs
  - human and mouse genomes have some 3 billion.
  - Each chromosome contains many **genes**.
- **Gene**
  - basic physical and functional units of heredity.
  - specific sequences of DNA bases that encode instructions on how to make **proteins**.
- **Proteins**
  - Make up the cellular structure
  - large, complex molecules made up of smaller subunits called **amino acids**.



# All Life depends on 3 critical molecules

---

- DNAs

- Hold information on how cell works

- RNAs

- Act to transfer short pieces of information to different parts of cell

- Provide templates to synthesize into protein

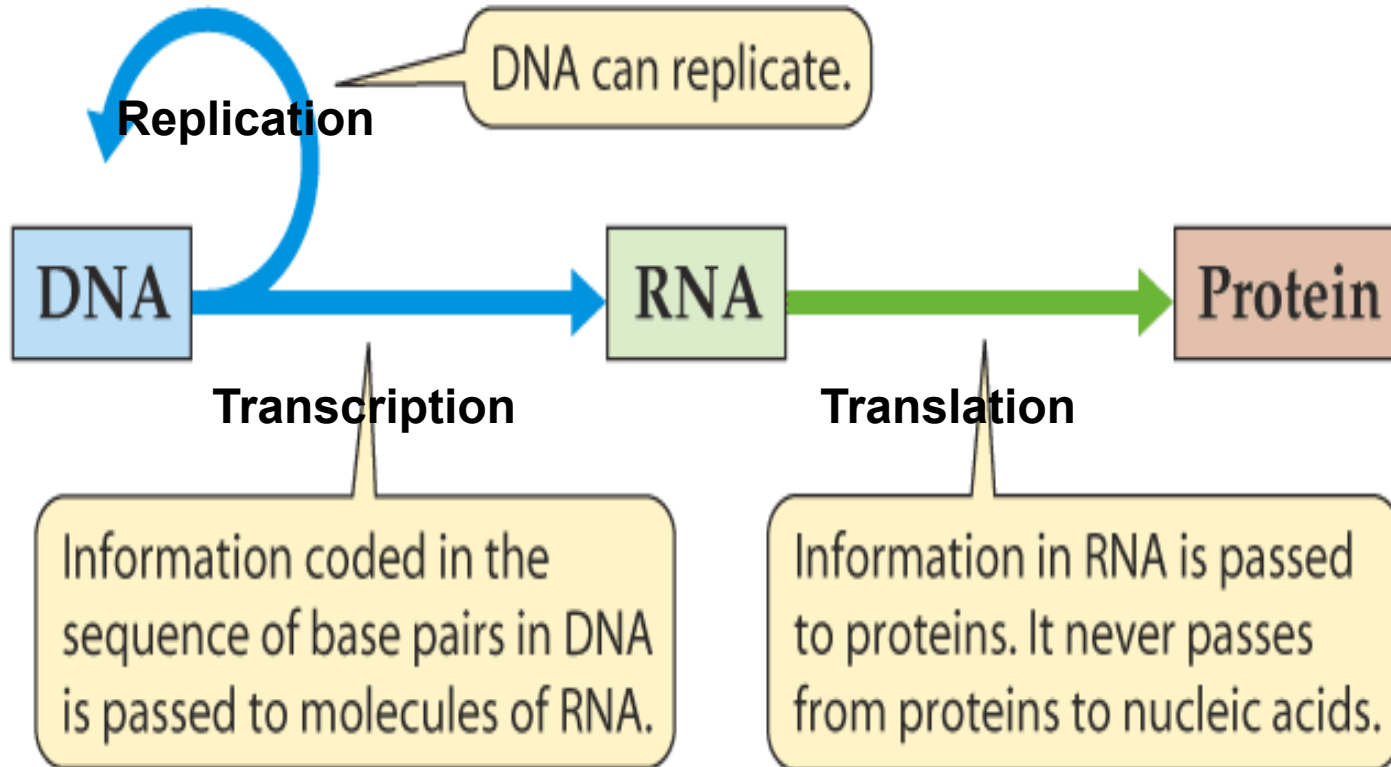
- Proteins

- Form enzymes that send signals to other cells and regulate gene activity

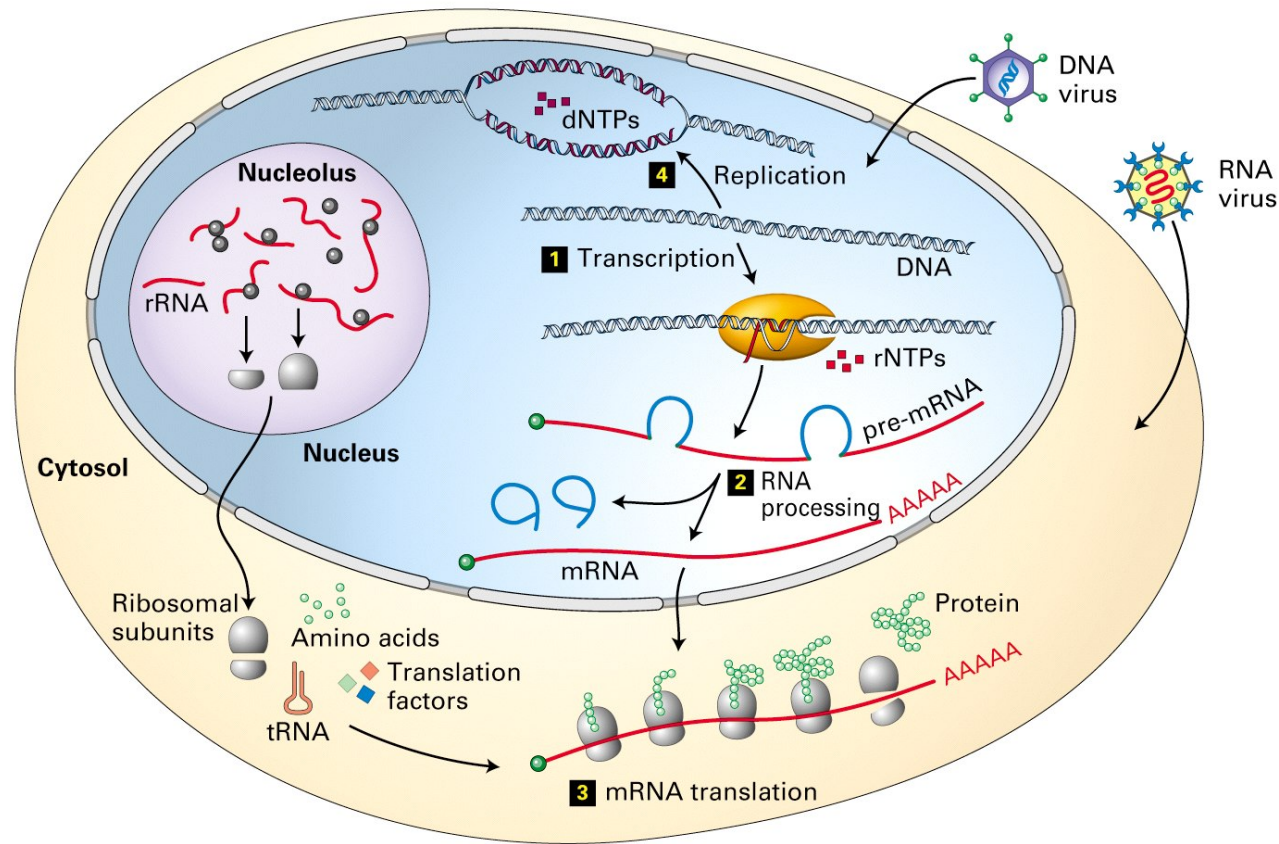
- Form body's major components (e.g. hair, skin, etc.)

# Central Dogma

---



# Overview of DNA to RNA to Protein



## ■ A gene is expressed in two steps

- 1) Transcription: RNA synthesis
- 2) Translation: Protein synthesis

# Discovery of DNA

## ■ DNA Sequences

— Chargaff and Vischer, 1949

- ◆ DNA consisting of A, T, G, C  
Adenine, Guanine, Cytosine, Thymine

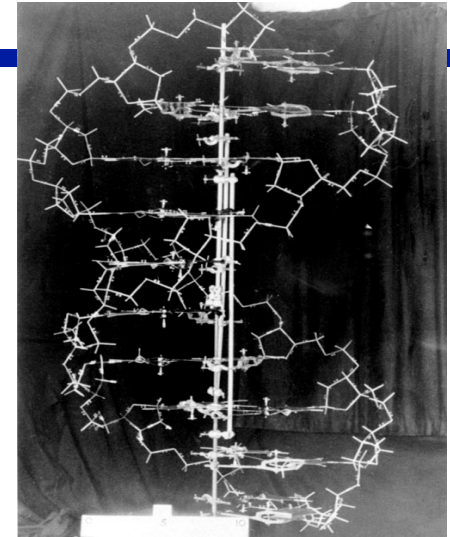
— Chargaff Rule

- ◆ Noticing #A - #T and #G - #C  
A “strange but possibly meaningless”  
phenomenon.

## ■ Wow!! A Double Helix

— Watson and Crick, *Nature*, April 25, 1953

—  
1 Biologist  
1 Physics Ph.D. Student  
+ 900 words  
= Nobel Prize



Original DNA demonstration model (scale gives distance in Angstroms)  
Cold Spring Harbor Laboratory Archives



Watson and Crick walk along the Beach  
Cold Spring Harbor Laboratory Archives

Crick

Watson



No. 4356 April 25, 1953

NATURE

737

738

NATURE

April 25, 1953 VOL. 171

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

<sup>1</sup>Young, T. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **46**, 140 (1929).

<sup>2</sup>Longest-Higgins, M. S., *Mon. Not. Roy. Astr. Soc., Geophys. Supp.*, **4**, 255 (1949).

<sup>3</sup>Von Aix, W. S., *Woods Hole Papers in Phys. Oceanogr. Meteor.*, **11** (5) (1950).

<sup>4</sup>Kronn, V. W., *Astr. J. Mat. Astron. Faint. (Stockholm)*, **2** (11) (1938).

## MOLECULAR STRUCTURE OF NUCLEIC ACIDS

### A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey<sup>1</sup>. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.



This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining 5'-deoxyribofuranose residues with 3',5' linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furbert's<sup>2</sup> model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furbert's 'standard configuration', the sugar being roughly perpendicular to the attached base. There

is a residue on each chain every 3.4 Å. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 Å. The distance of a phosphorus atom from the fibre axis is 10 Å. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally<sup>3,4</sup> that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data<sup>5,6</sup> on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on inter-atomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

King's College, London. One of us (J. D. W.) has been aided by a fellowship from the National Foundation for Infantile Paralysis.

J. D. WATSON  
F. H. C. CRICK

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge, April 2.

<sup>1</sup>Pauling, L., and Corey, R. B., *Nature*, **171**, 346 (1952); *Proc. U.S. Nat. Acad. Sci.*, **38**, 81 (1952).

<sup>2</sup>Furbert, S., *Acta Chem. Scand.*, **4**, 694 (1952).

<sup>3</sup>Chargaff, E., for references see *Kazanahof, S., Trautman, G., and Chappell, R., Biochim. et Biophys. Acta*, **9**, 402 (1952).

<sup>4</sup>Wyatt, G. R., *J. Gen. Physiol.*, **36**, 201 (1952).

<sup>5</sup>Arthur, W. T., *Symp. Soc. Exp. Biol.*, **1**, Nucleic Acid, 69 (Cavendish Univ. Press, 1952).

<sup>6</sup>Wilkins, M. H. F., and Randall, J. T., *Biochim. et Biophys. Acta*, **10**, 192 (1953).

## Molecular Structure of Deoxypentose Nucleic Acids

WHILE the biological properties of deoxypentose nucleic acid suggest a molecular structure containing great complexity, X-ray diffraction studies described here (cf. Arthur<sup>1</sup>) show the basic molecular configuration has great simplicity. The purpose of this communication is to describe, in a preliminary way, some of the experimental evidence for the polynucleotide chain configuration being helical, and existing in this form when in the natural state. A further account of the work will be published shortly.

The structure of deoxypentose nucleic acid is the same in all species (although the nitrogen base ratios alter considerably) in nucleoprotein, extracted or in cells, and in purified nucleate. The same linear group of polynucleotide chains may pack together parallel in different ways to give crystalline<sup>1-3</sup>, semi-crystalline or paracrystalline material. In all cases the X-ray diffraction photograph consists of two regions, one determined largely by the regular spacing of nucleotides along the chain, and the other by the longer spacings of the chain configuration. The sequence of different nitrogen bases along the chain is not made visible.

Oriented paracrystalline deoxypentose nucleic acid ('structure B' in the following communication by Franklin and Gosling) gives a fibre diagram as shown in Fig. 1 (cf. ref. 4). Arthur suggested that the strong 3.4-Å. reflexion corresponded to the inter-nucleotide repeat along the fibre axis. The ~34 Å. layer lines, however, are not due to a repeat of a polynucleotide composition, but to the chain configuration repeat, which causes strong diffraction as the nucleotide chains have higher density than the interstitial water. The absence of reflexions on or near the meridian immediately suggests a helical structure with axis parallel to fibre length.

### Diffraction by Helices

It may be shown<sup>5</sup> (also Stokes, unpublished) that the intensity distribution in the diffraction pattern of a series of points equally spaced along a helix is given by the squares of Bessel functions. A uniform continuous helix gives a series of layer lines of spacing corresponding to the helix pitch, the intensity distribution along the *n*th layer line being proportional to the square of *J<sub>n</sub>*, the *n*th order Bessel function. A straight line may be drawn approximately through

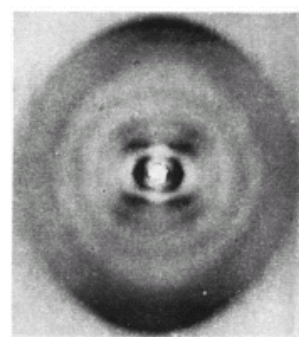


Fig. 1. Fibre diagram of deoxypentose nucleic acid from *B. coli*. Fibre axis vertical.

the innermost maxima of each Bessel function along the origin. The angle this line makes with the equator is roughly equal to the angle between an element of the helix and the helix axis. If a unit repeats *n* times along the helix there will be a meridional reflexion (*J<sub>0</sub>*) on the *n*th layer line. The helical configuration produces side-bands on this fundamental frequency, the effect being to reproduce the intensity distribution about the origin around the new origin, on the *n*th layer line, corresponding to *C* in Fig. 2.

We will now briefly analyse in physical terms some of the effects of the shape and size of the repeat unit or nucleotide on the diffraction pattern. First, if the nucleotide consists of a unit having circular symmetry about an axis parallel to the helix axis, the whole diffraction pattern is modified by the form factor of the nucleotide. Second, if the nucleotide consists of a series of points on a radius at right-angles to the helix axis, the phases of radiation scattered by the helices of different diameter passing through each point are the same. Summation of the corresponding Bessel functions gives reinforcement for the inner-

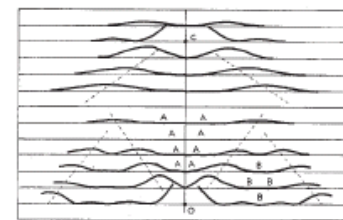
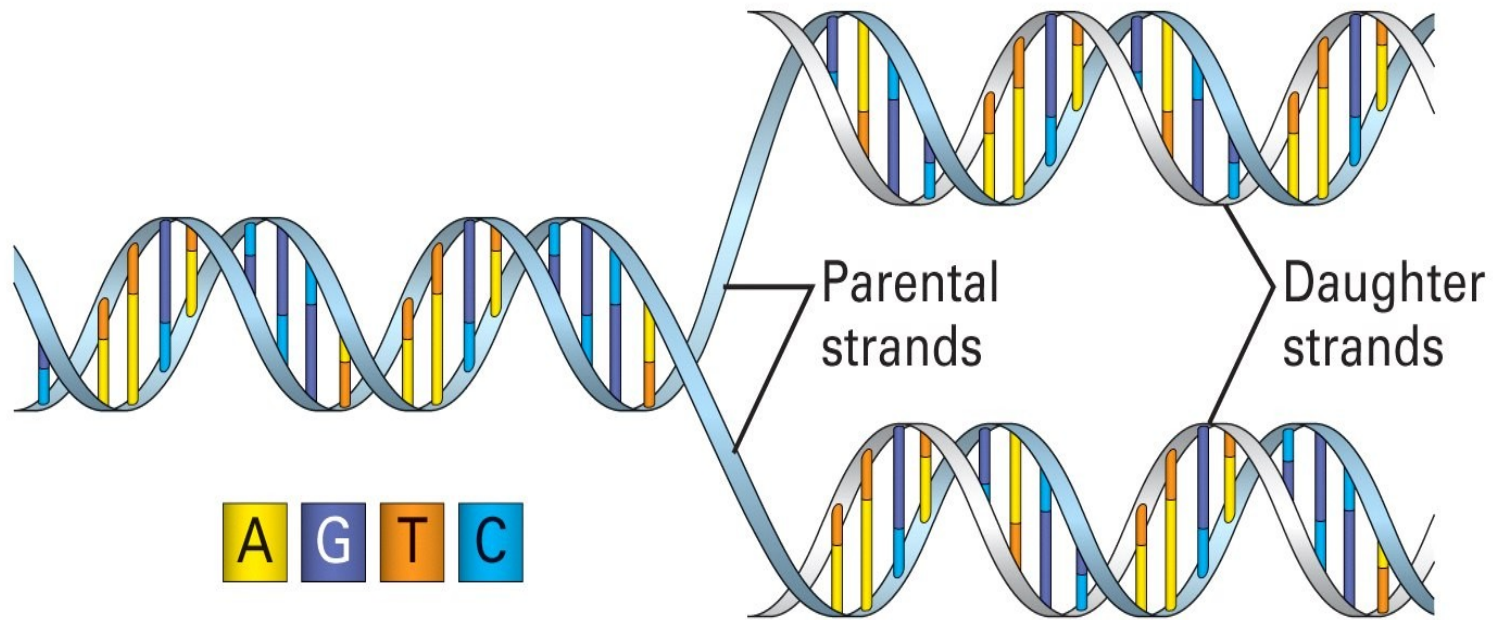


Fig. 2. Diffraction pattern of system of helices corresponding to structure of deoxypentose nucleic acid. The squares of Bessel functions are plotted along *C* on the equator and on the first, second, third and fifth layer lines for half of the nucleotide mass at 25 Å. diameter and remainder distributed along a radius, the mass at a given radius being proportional to the radius. About *C* on the tenth layer line smaller functions are plotted for an outer diameter of 15 Å.

# DNA: The Code of Life

---



- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.

# DNA, continued

---

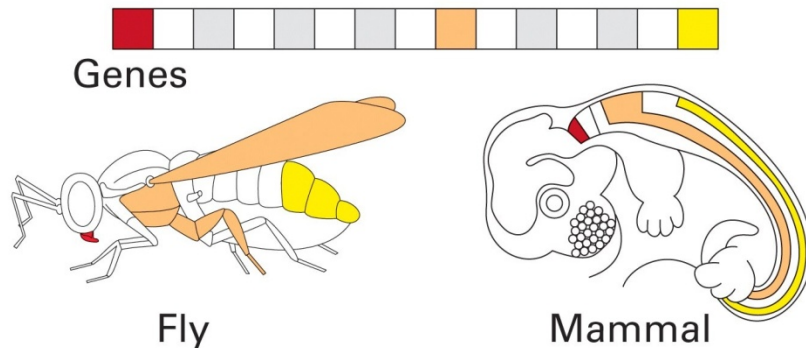
- DNA has a double helix structure. However, it is not symmetric. It has a “forward” and “backward” direction. The ends are labeled 5’ and 3’ after the Carbon atoms in the sugar component.

5’ AATCGCAAT 3’

3’ TTAGCGTTA 5’

DNA always reads 5’ to 3’ for transcription and replication

# DNA the Genetics Makeup



- Genes are inherited and are expressed
  - **genotype** (genetic makeup)
  - **phenotype** (physical expression)
- On the left, is the eye's phenotypes of green and black eye genes.





# RNA

---

- RNA is similar to DNA chemically but T(hyamine) is replaced by U(racil)
- It is usually only a single strand and shorter and less stable than DNA.
- Some forms of RNA can form secondary structures by “pairing up” with itself. This can change its properties dramatically.

# RNA, continued

---

Several types exist, classified by function

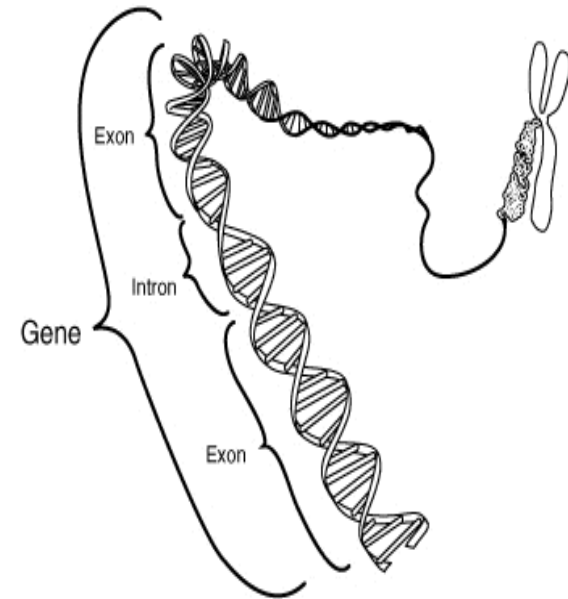
- **mRNA** – this is what is usually being referred to when a Bioinformatician says “RNA”. This is used to carry a gene’s *message* out of the nucleus.
- **tRNA** – *transfers* genetic information from mRNA to an amino acid sequence
- **rRNA** – *ribosomal* RNA. Part of the ribosome which is involved in translation.

- 
- Central dogma video

<http://www.youtube.com/watch?v=ZNcFTRX9i0Y>

# Definition of a Gene

- A gene is a portion of DNA that contains both "coding" sequences that determine what the gene does, and "non-coding" sequences that determine when the gene is active (expressed).
- Exons: coding sequence  
1 to 178 exons per gene (mean 8.8)  
8 bp to 17 kb per exon (mean 145bp)
- Introns: noncoding seq, junk DNA  
average 1 kb – 50 kb per intron
- Gene size: Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.
- After transcription and splicing, mRNA contains exons only



# Uncovering the code

---

- Scientists conjectured that proteins came from DNA; but how did DNA code for proteins?
- If one nucleotide codes for one amino acid, then there'd be  $4^1$  amino acids
- However, there are 20 amino acids, so at least 3 bases codes for one amino acid, since  $4^2 = 16$  and  $4^3 = 64$ 
  - This triplet of bases is called a “codon”
  - 64 different codons and only 20 amino acids means that the coding is degenerate: more than one codon sequence code for the same amino acid

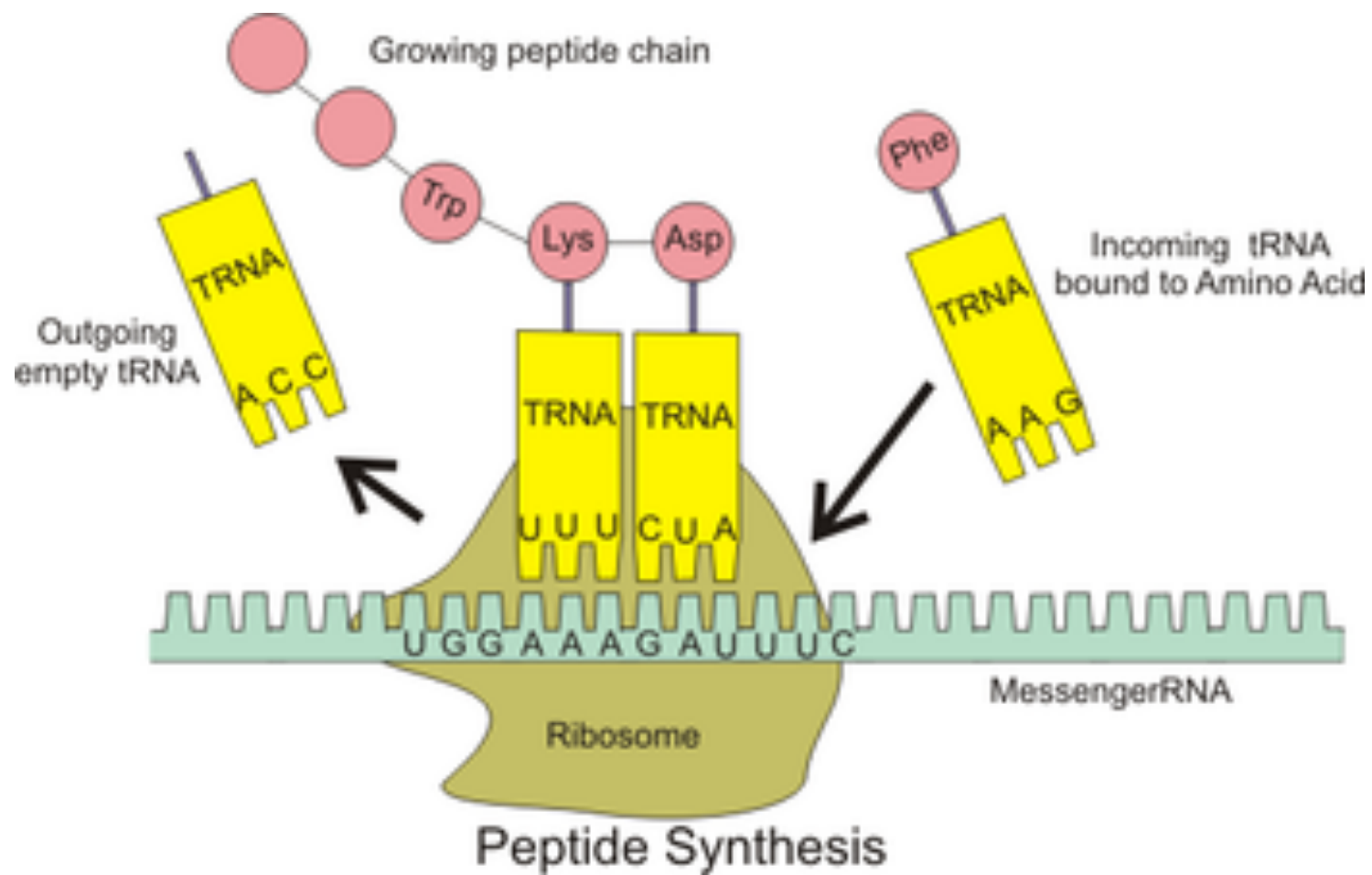
# Translation

- Ribosomes and *transfer-RNAs* (tRNA) run along the length of the newly synthesized mRNA, decoding one codon at a time to build a growing chain of amino acids (“peptide”)
- Always starts with Methionine and ends with a stop codon

Three base pairs of RNA (called a codon) correspond to one amino acid based on a fixed table.

		SECOND POSITION				THIRD POSITION	
FIRST POSITION		U	C	A	G		
	U	phenyl-alanine	serine	tyrosine	cysteine		U
							C
		leucine		stop	stop		A
				stop	tryptophan		G
	C	leucine	proline	histidine	arginine		U
				glutamine		A	
						G	
	A	isoleucine	threonine	asparagine	serine	U	
						C	
		* methionine		lysine	arginine	A	
						G	
G	valine	alanine	aspartic acid	glycine	U		
						C	
			glutamic acid		A		
					G		

\* and start



# Proteins

---

- Complex organic molecules made up of amino acid subunits
- 20\* different kinds of amino acids. Each has a 1 and 3 letter abbreviation.
- Proteins are often enzymes that catalyze reactions.
- Also called “poly-peptides”

\*Some other amino acids exist but not in humans.



# Proteins: Workhorses of the Cell

---

- fold up into specific three-dimensional structures that define their particular functions in the cell
- Proteins do all essential work for the cell
  - build cellular structures
  - digest nutrients
  - execute metabolic functions
  - Mediate information flow within a cell and among cellular communities.
- Proteins work together with other proteins or nucleic acids as "molecular machines"
  - structures that fit together and function in highly specific, lock-and-key ways.

# How Do Individuals of Species Differ?

---

- Physical Variation and Diversity
- Genetic Variation

# How Do Individuals of Species Differ?

---

- Genetic makeup of an individual is manifested in traits, which are caused by variations in genes
- While 99.9% of the 3 billion nucleotides in the human genome are the same, small variations can have a large range of phenotypic expressions
- These traits make some more or less susceptible to disease, and the demystification of these mutations will hopefully reveal the truth behind several genetic diseases

# The Diversity of Life

---

- Not only do different species have different genomes, but also different individuals of the same species have different genomes.
- No two individuals of a species are quite the same – this is clear in humans but is also true in every other sexually reproducing species.
- Imagine the difficulty of biologists – sequencing and studying only one genome is not enough because every individual is genetically different!

# Physical Traits and Variances

---

- **Individual variation** among a species occurs in populations of all sexually reproducing organisms.
- Individual variations range from hair and eye color to less subtle traits such as susceptibility to malaria.
- Physical variation is the reason we can pick out our friends in a crowd, however most physical traits and variation can only be seen at a cellular and molecular level.



# Sources of Physical Variation

---

- Physical Variation and the manifestation of traits are caused by **variations in the genes** and **differences in environmental influences**.
  - An example is height, which is dependent on genes as well as the nutrition of the individual.
- Not all variation is inheritable – only genetic variation can be passed to offspring.
- Biologists usually focus on genetic variation instead of physical variation because it is a better representation of the species.

# Genetic Variation

---

- Despite the wide range of physical variation, genetic variation between individuals is quite small.
- Out of 3 billion nucleotides, only roughly 3 million base pairs (0.1%) are different between individual genomes of humans.
- Although there is a finite number of possible variations, the number is so high ( $4^{3,000,000}$ ) that we can assume no two individual people have the same genome.
- What is the cause of this genetic variation?

# Sources of Genetic Variation

---

- **Mutations** are rare errors in the DNA replication process that occur at random.
- When mutations occur, they affect the genetic sequence and create genetic variation between individuals.
- Most mutations do not create beneficial changes and actually kill the individual.
- Although mutations are the source of all new genes in a population, they are so rare that there must be another process at work to account for the large amount of diversity.



# Sources of Genetic Variation

---

- **Recombination** is the shuffling of genes that occurs through sexual mating and is the main source of genetic variation.
  - Recombination occurs via a process called **crossing over** in which genes switch positions with other genes during meiosis.
  - Recombination means that new generations inherit random combinations of genes from both parents.
  - The recombination of genes creates a seemingly endless supply of genetic variation within a species.

# Why Bioinformatics?

---

- Bioinformatics is the combination of biology and computing.
- DNA sequencing technologies have created massive amounts of information that can only be efficiently analyzed with computers.
- So far hundreds of eukaryotic species sequenced
  - Human, rat, chimpanzee, chicken, and many others.
- As the information becomes ever so larger and more complex, more computational tools are needed to sort through the data.
  - Bioinformatics to the rescue!!!

# What is Bioinformatics?

---

- Bioinformatics is generally defined as the analysis, prediction, and modeling of biological data with the help of computers



# Bio-Information

---

- Since discovering how DNA acts as the instructional blueprints behind life, biology has become an information science
- Now that many different organisms have been sequenced, we are able to find meaning in DNA through *comparative genomics*, not unlike comparative linguistics.
- Slowly, we are learning the syntax of DNA

# Biological Databases

---

- Vast biological and sequence data is freely available through online databases
- Use computational algorithms to efficiently store large amounts of biological data

## Examples

- **NCBI GeneBank** <http://ncbi.nih.gov>  
Huge collection of databases, the most prominent being the nucleotide sequence database
- **Protein Data Bank** <http://www.pdb.org>  
Database of protein tertiary structures
- **SWISSPROT** <http://www.expasy.org/sprot/>  
Database of annotated protein sequences
- **PROSITE** <http://kr.expasy.org/prosite>  
Database of protein active site motifs

# It is Sequenced, What's Next?

---

- Tracing Phylogeny
  - Finding family relationships between species by tracking similarities between species.
- Gene Annotation (cooperative genomics)
  - Comparison of similar species.
- Determining Regulatory Networks
  - The variables that determine how the body reacts to certain stimuli.
- Proteomics
  - From DNA sequence to a folded protein.
- Systems Biology
- ....

# Modeling

---

- Modeling biological processes tells us if we understand a given process
- Because of the large number of variables that exist in biological problems, powerful computers are needed to analyze certain biological questions

# Topics in Bioinformatics

---

- Sequence analysis
- Protein folding, interactions and modelling (structural genomics)
- High-throughput experimental data analysis: Microarray; Mass Spectrometry; next generation sequencing
- Comparative genomics
- Regulatory network modeling; Systems Biology
- Database exploration and management
- ...



# The future...

---

- Bioinformatics is a dynamic and evolving field.
- Much is still to be learned about how proteins can manipulate a sequence of base pairs in such a peculiar way that results in a fully functional organism.
- How can we then use this information to benefit humanity without abusing it?

# R is a free software environment for statistical computing and graphics ([www.r-project.org](http://www.r-project.org)).


The R Project for Statistical Computing - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

Address <http://www.r-project.org/> Go Links

## The R Project for Statistical Computing



About R  
[What is R?](#)  
[Contributors](#)  
[Screenshots](#)  
[What's new?](#)

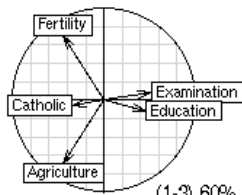
Download  
[CRAN](#)

R Project  
[Foundation](#)  
[Members & Donors](#)  
[Mailing Lists](#)  
[Bug Tracking](#)  
[Developer Page](#)  
[Search](#)

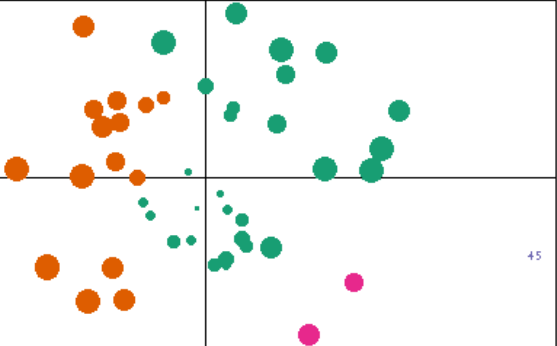
Documentation  
[Manuals](#)  
[FAQs](#)  
[Newsletter](#)  
[Books](#)  
[Other](#)

Misc

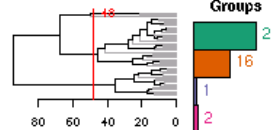
PCA 5 vars  
princomp(x = data, cor = cor)



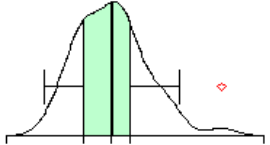
(1-3) 60%



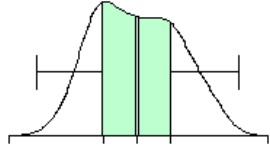
Clustering 4 groups



Factor 1 [41%]



Factor 3 [19%]



### Getting Started:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred [CRAN mirror](#).

Microsoft PowerPoint - [060106\_W1\_Intro\_Bio.ppt]

Start | Inbox - Microsoft Outl... | 060106\_W1\_Intro\_bio | The R Project for Stat... | Microsoft PowerPoint ... | 上午 01:54

- 
- Review slides of basic molecular biology if you are not familiar with it.
  - Download and install R software. Follow the tutorial from the course website and practice basic operation in R.