Molecular and Cellular Computing

Lecture series at Universidad Politécnica de Madrid

Martyn Amos
Department of Computing and Mathematics
Manchester Metropolitan University
United Kingdom

http://www.martynamos.com

Day 3: Biological Engineering
1. Biological Background

Some slides from *An Introduction to Bioinformatics Algorithms*, Neil C. Jones and Pavel A. Pevzner (MIT Press), http://www.bioalgorithms.info

Basics of Life

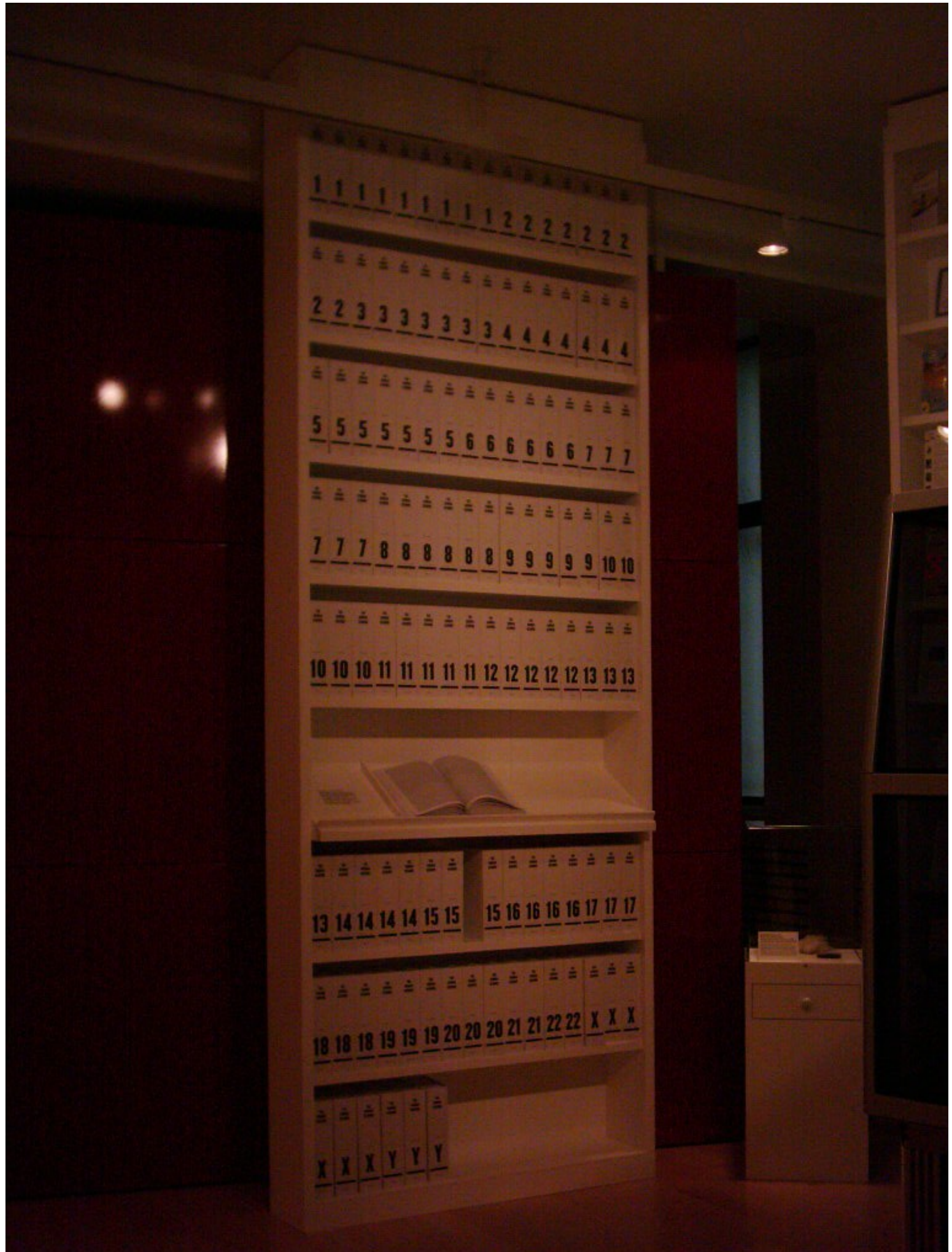
- Cells store all information to replicate themselves
  - Human genome is around 3 billions base pairs long
  - Almost every cell in human body contains same set of genes
  - But not all genes are used or expressed by those cells
- Machinery:
  - Collect and manufacture components
  - Carry out replication
  - Kick-start its new offspring

(A cell is like a car factory)

Basics 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

Manchester Metropolitan University



5m x 2m,
120 volumes

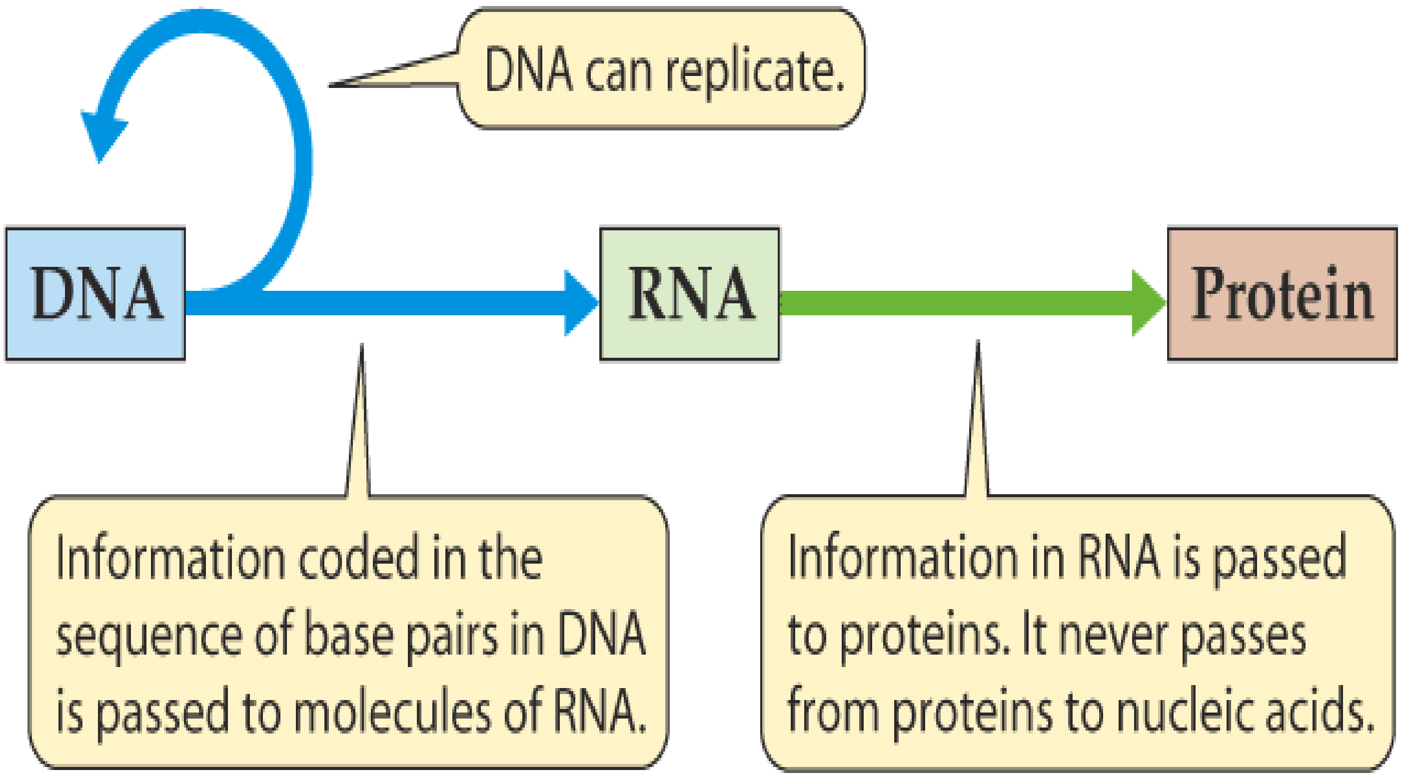
Terminology

- The *genome* is an organism's complete set of DNA.
  - A bacterium contains about 600,000 DNA base pairs
  - Human and mouse genomes have some 3 billion
- Human genome has 24 distinct chromosomes (1-22, X and Y)
  - 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*

Three Critical Molecules

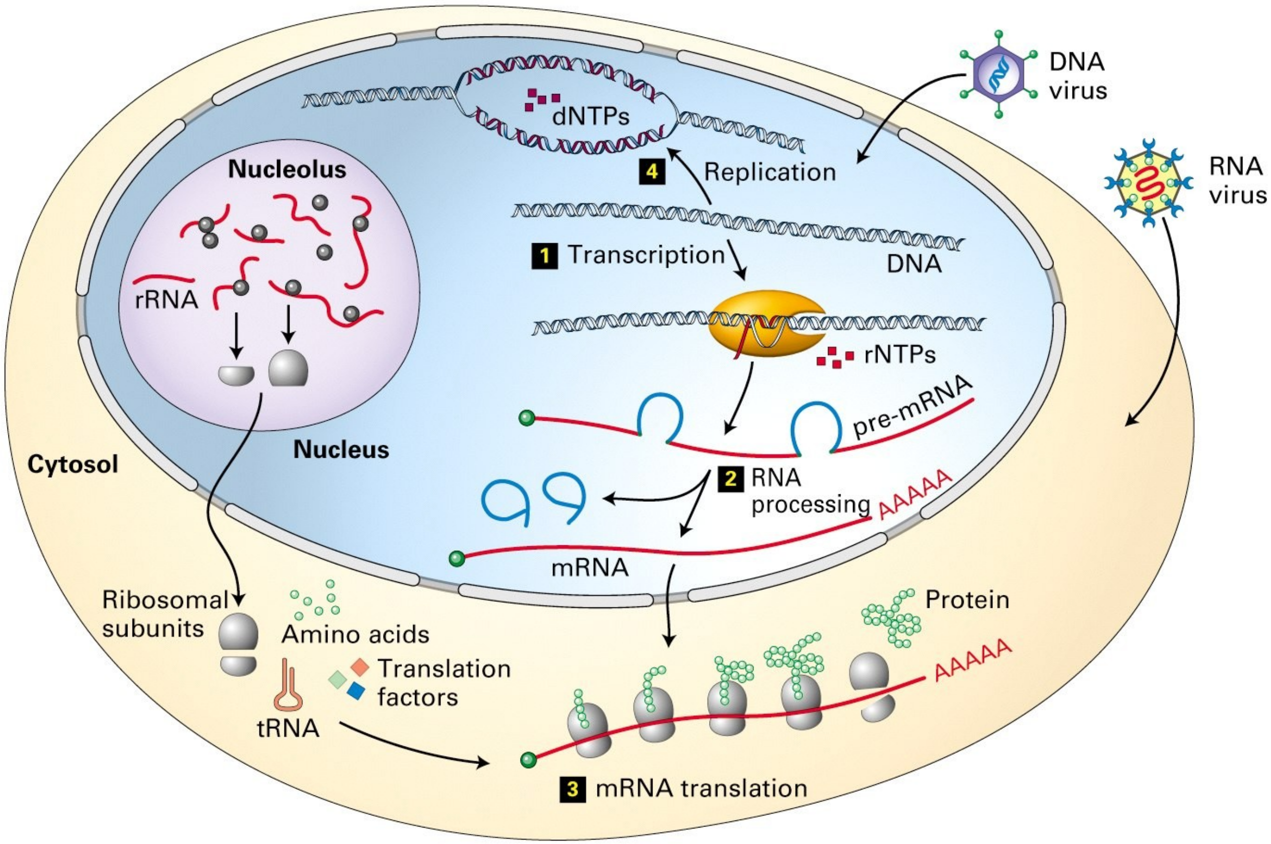
- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components (e.g. hair, skin, etc.)

DNA, RNA and Information Flow



Manchester Metropolitan University

Overview of DNA to RNA to Protein



- A gene is expressed in two steps
  1) *Transcription*: RNA synthesis
  2) *Translation*: Protein synthesis

Proteins: Workhorses of the Cell

- 20 different *amino acids*
  - Different chemical properties cause the protein chains to 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

Manchester Metropolitan University

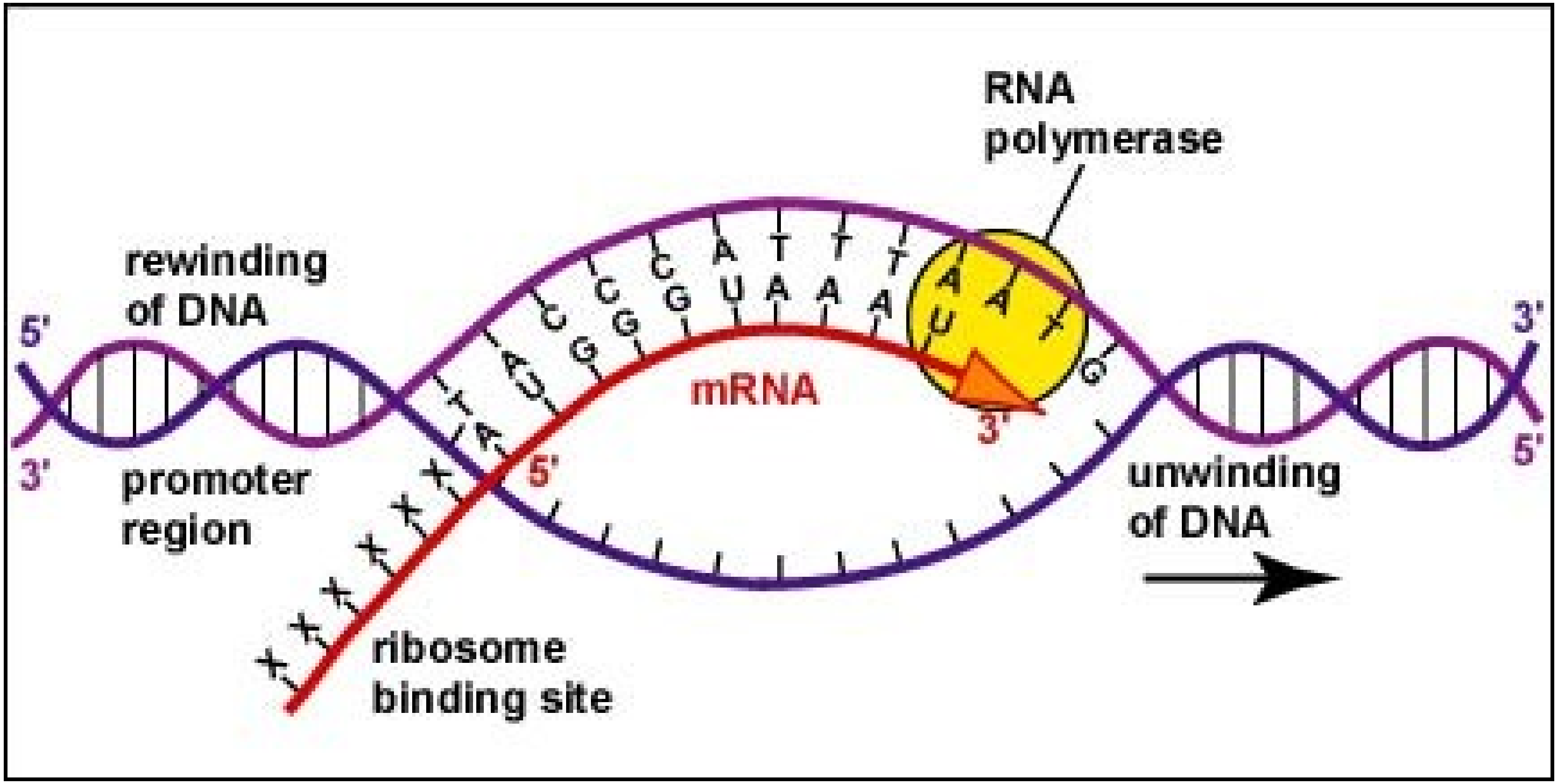
DNA to RNA: Transcription

- DNA gets transcribed by a protein known as *RNA-polymerase*
- This process builds a chain of bases that will become mRNA
- RNA and DNA are similar, except that RNA is single stranded and thus less stable than DNA
  - Also, in RNA, the base uracil (U) is used instead of thymine (T), the DNA counterpart

Transcription

- To begin transcription requires a *promoter*, a small specific sequence of DNA to which polymerase can bind (~40 base pairs “upstream” of gene)
- There can also be *repressors* and *inhibitors* acting in various ways to stop transcription
- This makes regulation of gene transcription *complex* to understand (as we shall see)

Transcription



The Genetic Code

- Scientists conjectured that proteins came from DNA; but how did DNA code for proteins?
- If one nucleotide coded for one amino acid, then there'd be 4^1 amino acids
- However, there are 20 amino acids, so at least 3 bases code 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

Manchester Metropolitan University

RNA to Protein: 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”)
- Three base pairs of RNA (called a codon) correspond to one amino acid based on a fixed table

Second Position

First Position	U		C		A		G		Third Position
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	*Stop*	UGA	*Stop*	A
	UUG		UCG		UAG	*Stop*	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Protein Folding

- Proteins are not linear structures, though they are built that way
- The amino acids have very different chemical properties; they interact with each other after the protein is built
  - This causes the protein to start fold and adopting it’s functional structure
  - Proteins may fold in reaction to some ions, and several separate chains of peptides may join together through their hydrophobic and hydrophilic amino acids to form a polymer

Protein Folding

- The structure that a protein adopts is vital to its chemistry
- Its structure determines which of its amino acids are exposed carry out the protein's function
- Its structure also determines what substrates it can react with
- Protein structure prediction is one of the "Holy Grails" of modern biology

Genetic Regulation

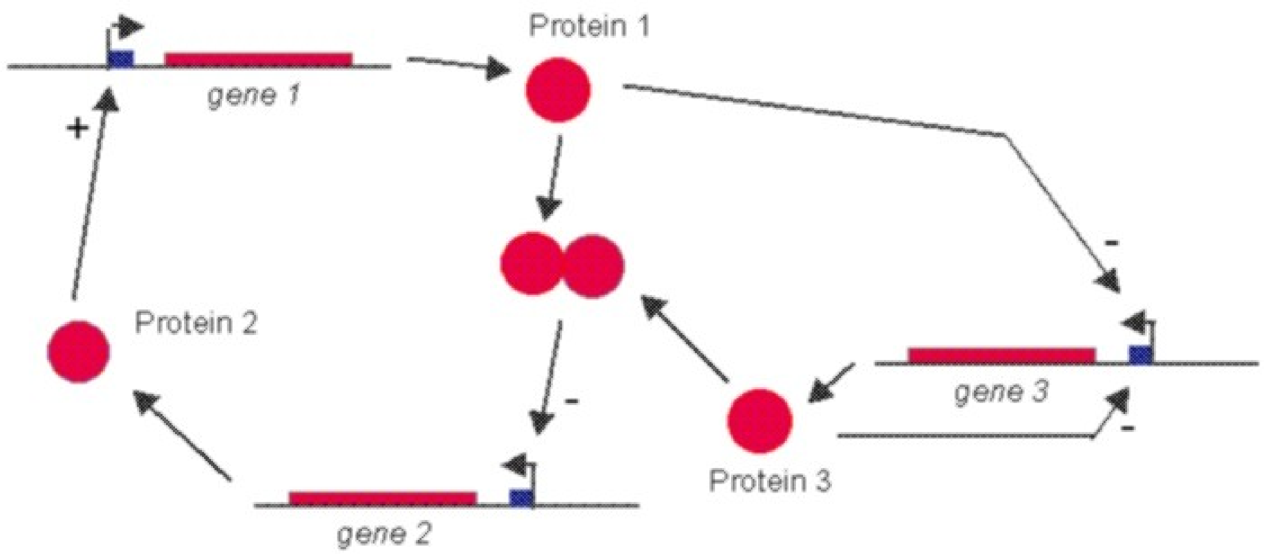
"The punched tape running along the inner seam of the double helix is much more than a repository of enzyme stencils. It packs itself with regulators, suppressors, promoters, case-statements, if-thens".

Richard Powers, *The Gold Bug Variations*, p. 365, HarperPerennial

Genetic Regulation

- DNA contains the potential coding information for a vast range of possible proteins
- Gene expression is not a linear process
- Genes may require the product(s) of other gene(s) to in order to be expressed
- The product of one gene may turn off the expression of another gene
- The product of a gene can even effect its *own* expression (feedback)

Example



Gene 2 codes for a protein that activates the transcription of gene 1, while gene 1 and gene 3 code for proteins that form a complex inhibiting the transcription of gene 2. Activation and inhibition of gene expression are indicated by + and -, respectively.

The *lac* Operon

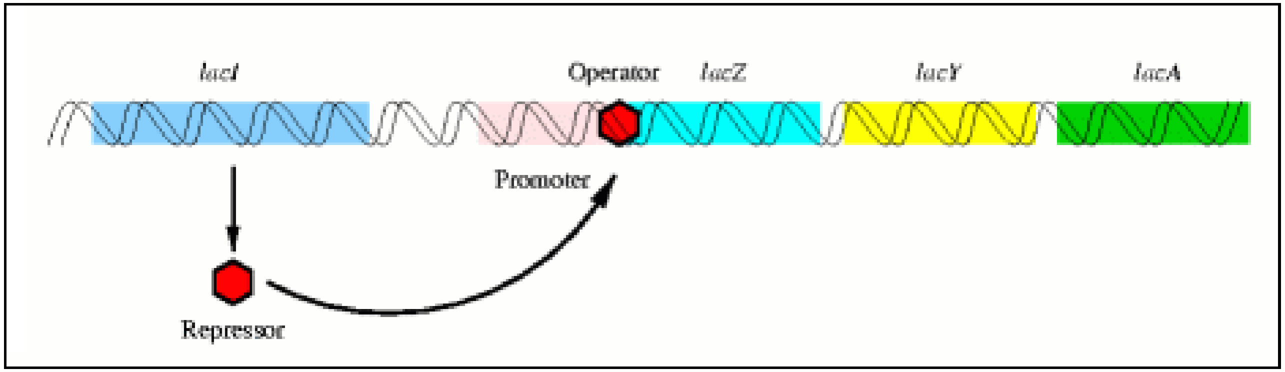
- *Operon* - set of functionally related genes with common promoter
- *lac* operon contains three structural genes that allow *E. coli* to utilise lactose
- When bacteria grown in glucose, product of *lacI* gene *represses* transcription of *lac*
- When grown in glucose *and* lactose, lactose by-product inhibits repressor, and the genes are expressed
- *lac* operon controlled by two sugars (inputs)

Jacob F; Monod J (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*. **3**: 318-56

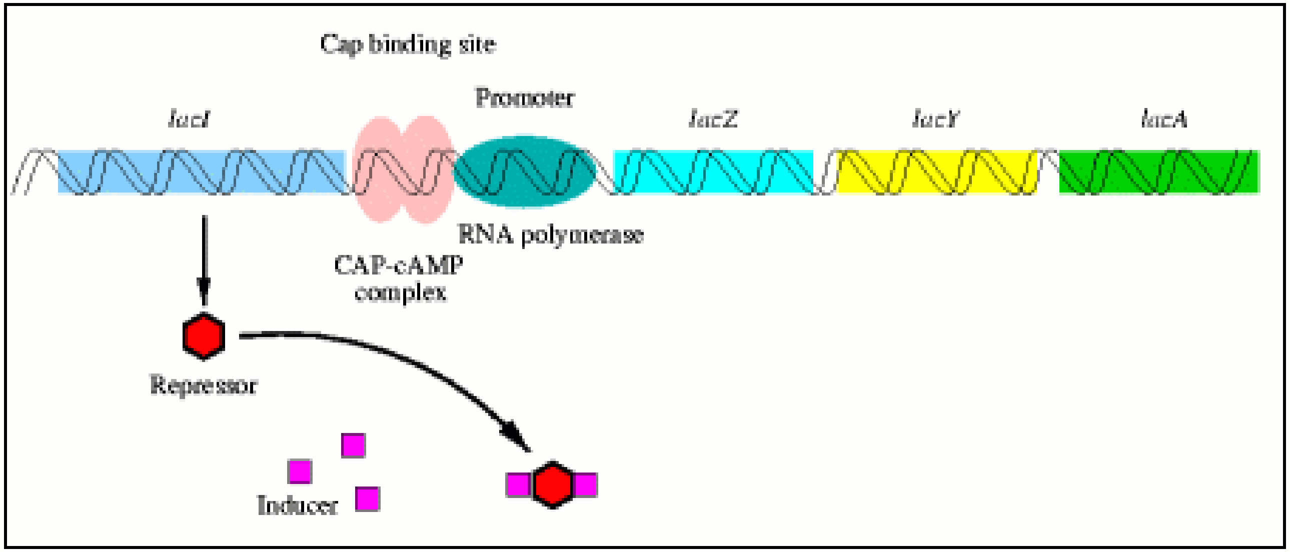
The *lac* operon

- The *lac* genes are so-called because, in the E. coli bacterium, they combine to produce a variety of proteins that allow the cell to metabolise the sugar lactose
- For reasons of efficiency, these proteins should only be produced (i.e., the genes be turned on) when lactose is present in the cell's environment
- Making these proteins when lactose is absent would be a waste of the cell's resources
- However, a different sugar – glucose – will always be preferable to lactose, if the cell can get it, since glucose is an "easier" form of sugar to metabolise

Repression and Inhibition



Just glucose

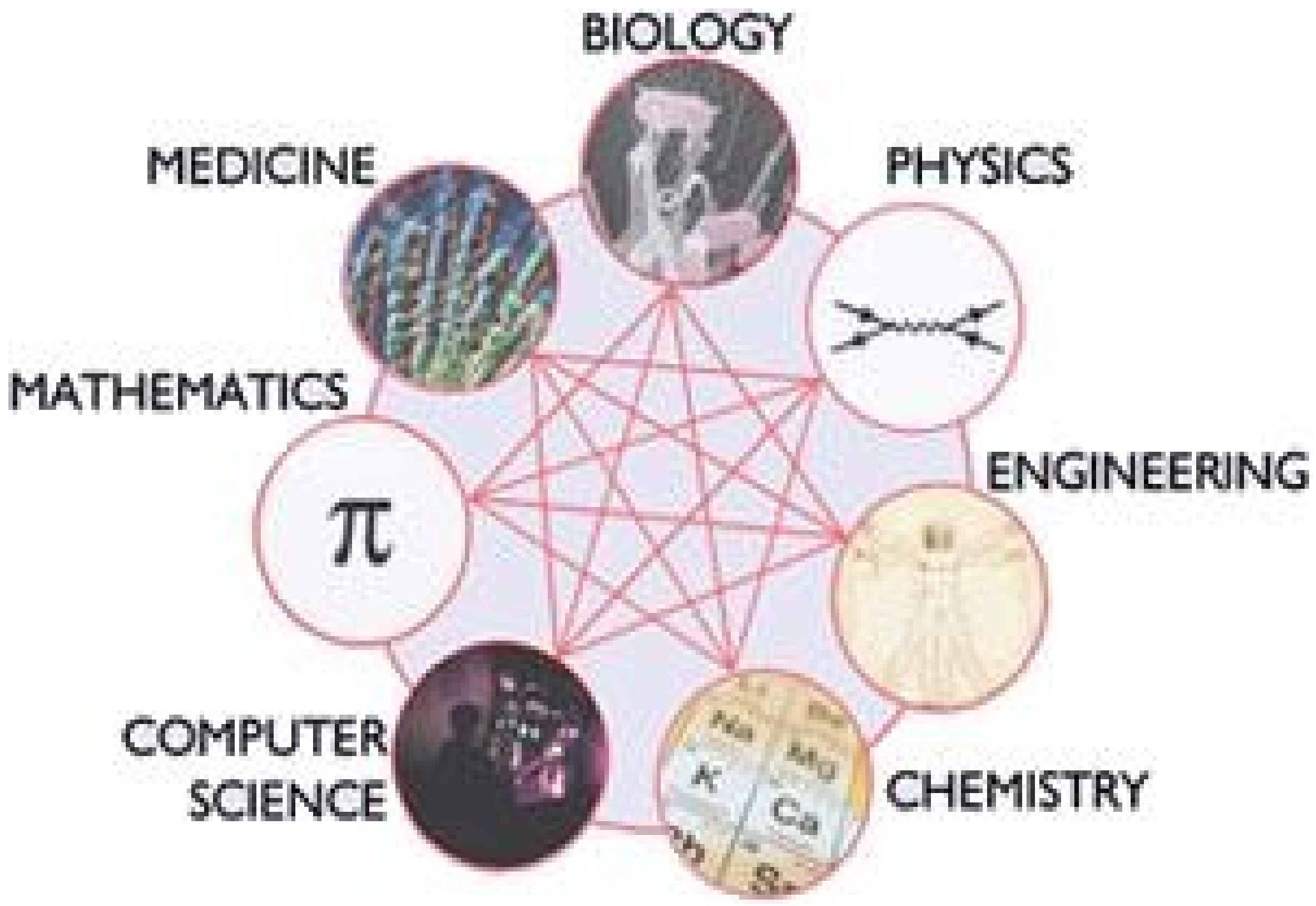


Glucose and lactose

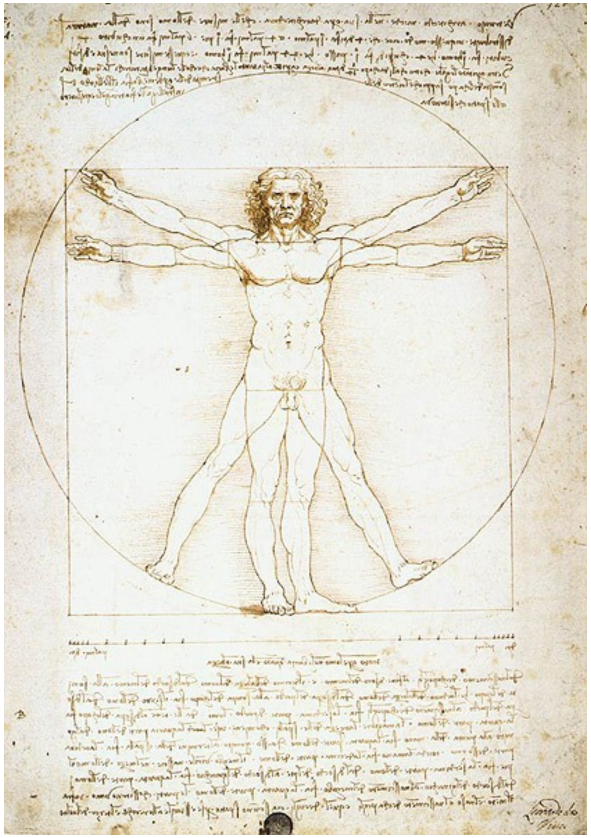
Sugar Logic

- "The logic of biological regulatory systems abides...like the workings of computers, by the propositional logic of George Boole." F. Jacob
- *lac* operon may be thought of as a genetic circuit that computes the function
  - L AND (NOT G)
- Car seat belt circuit
- By showing how one gene could affect the expression of another, Jacob and Monod laid the foundations for a *systems* level understanding of the genome

Systems Biology



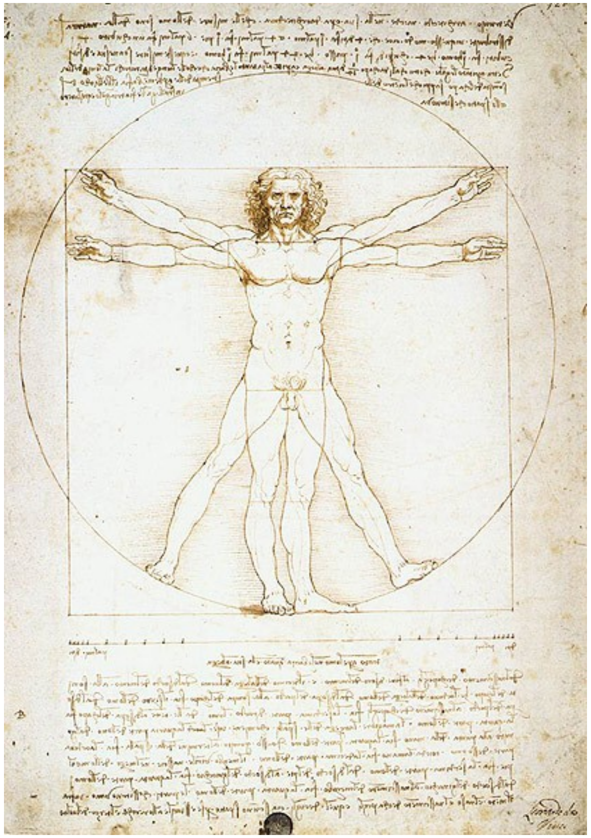
Systems Insight



30,000 genes (approx.)

Manchester Metropolitan University

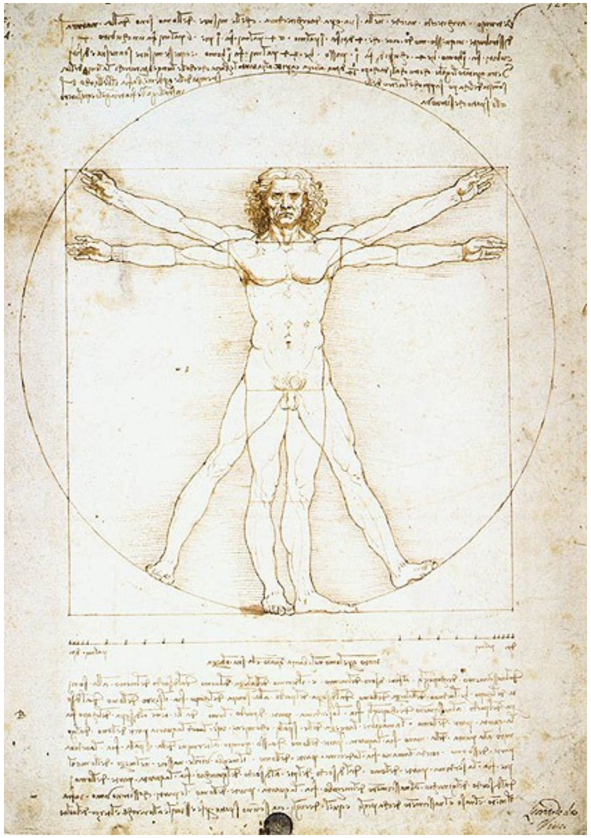
Systems Insight



30,000 genes (approx.)

Manchester Metropolitan University

Systems Insight



30,000 genes (approx.)

20,000 genes!

Manchester Metropolitan University

$2^{10,000}$ =

19950631168807583848837421626835850838234968318861924548520089498529
43883022194663191996168403619459789933112942320912427155649134941378
11175937859320963239578557300467937945267652465512660598955205500869
18193311542508608460618104685509074866089624888090489894838009253941
63325785062156830947390255691238806522509664387444104675987162698545
32228685381616943157756296407628368807607322285350916414761839563814
58969463899410840960536267821064621427333394036525565649530603142680
23496940033593431665145929777327966577560617258203140799419817960737
82456837622800373028854872519008344645814546505579296014148339216157
34588139257095379769119277800826957735674444123062018757836325502728
32378927071037380286639303142813324140162419567169057406141965434232
46388012488561473052074319922596117962501309928602417083408076059323
20161268492288496255841312844061536738951487114256315111089745514203
31382020293164095759646475601040584584156607204496286701651506192063
10041864222759086709005746064178569519114560550682512504060075198422
61898059237118054444788072906395242548339221982707404473162376760846
61303377870603980341319713349365462270056316993745550824178097281098
32913144035718775247685098572769379264332215993998768866608083688378
38027643282775172273657572744784112294389733810861607423253291974813
12019760417828196569747589816453125843413595986278413012818540628347
66490886905210475808826158239619857701224070443305830758690393196046
03404973156583208672105913300903752823415539745394397715257455290510
21231094732161075347482574077527398634829849834075693795564663862187
45694992790165721037013644331358172143117913982229838458473344402709
64182851005072927748364550578634501100852987812389473928699540834346
15880704395911898581514577917714361969872813145948378320208147498217
18580113890712282509058268174362205774759214176537156877256149045829
04992461028630081535583308130101987675856234343538955409175623400844
88752616264356864883351946372037729324009445624692325435040067802727
38377553764067268986362410374914109667185570507590981002467898801782
71925953381282421954028302759408448955014676668389697996886241636313
37639390337345580140763674187771105538422573949911018646821969658165
14851304942223699477147630691554682176828762003627772577237813653316
11196811280792669481887201298643660768551639860534602297871557517947
38524636944692308789426594821700805112032236549628816903573912136833
83935917564187338505109702716139154395909915981546544173363116569360
31122249937969999226781732358023111862644575299135758175008199839236
28461524988108896023224436217377161808635701546848405862232979285387
56234865564405369626220189635710288123615675125433383032700290976686
50568557157505516727518899194129711337690149916181315171544007728650
57318955745092033018530484711381831540732405331903846208403642176370
39115506397890007428536721962809034779745333204683687958685802379522
18629120080742819551317948157624448298518461509704888027274721574688
13159475040973211508049819045580341682694978714131606321068639151168
17743047925965709376