Computational Systems Biology

. . . Biology X – Lecture 3 . . .

Bud Mishra

Professor of Computer Science, Mathematics, & Cell Biology

Some Biology

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:
    - $\Sigma = \{A, T, C, G\}$

Complementarity

- DNA is a double-stranded polymer
  - should be thought of as a pair of sequences over Σ.
- A relation of **complementarity**
  - $A \Leftrightarrow T, C \Leftrightarrow G$
  - If there is an A (resp., T, C, G) on one sequence at a particular position then the other sequence must have a T (resp., A, G, C) at the same position.
- The sequence length
  - Is measured in terms of base pairs (bp): Human (H. sapiens) DNA is 3.3×10^9 bp, about 6 ft of DNA polymer completely stretched out!

Genome Size

- **The genomes vary widely in size:**
  - 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^{3}bp.
  - Mycoplasmas (a unicellular organism) has the smallest cellular genome: 5×10^{5}bp.
  - *C. elegans* (nematode worm, a primitive multicellular organism) has a genome of size ~ 10^{8}bp.

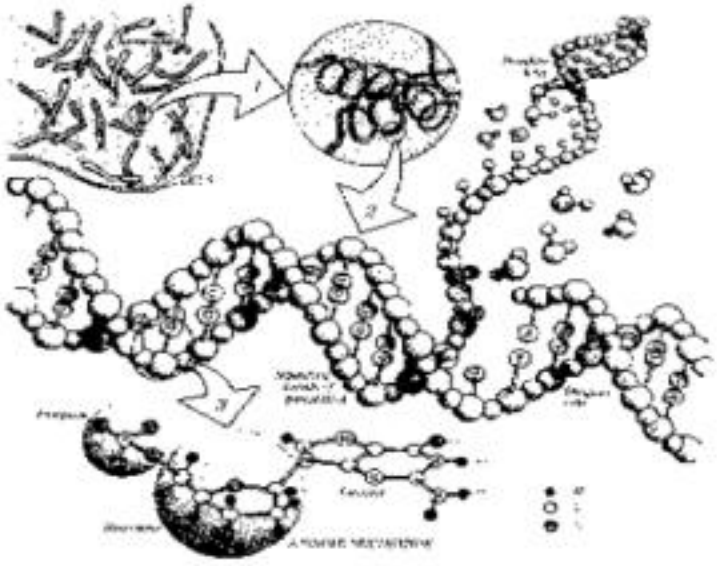
Species	Haploid Genome Size	Chromosome Numer
E. Coli	4.64×10^{6}	1
S.cerevisae	1.205×10^{7}	16
C. elegans	10^{8}	11/12
D. melanogaster	1.7×10^{8}	4
M. musculus	3×10^{9}	20
H. sapiens	3×10^{9}	23
A. Cepa (Onion)	1.5×10^{10}	8

DNA ⇒ Structure and Components

- Double helix
  - 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 slightly different spacing.
  - The "backbone" in each strand is made of alternating sugar molecules (Deoxyribose residues: $C_5 O_4 H_{10}$) and phosphate ($(P O_4)^{-3}$) molecules.
- Each of the four bases, an almost planar nitrogenic organic compound, is connected to the sugar molecule.
  - The bases are:

    Adenine ⇒ A; Thymine ⇒ T; Cytosine ⇒ C; Guanine ⇒ G

Genome in Detail



The Human Genome at Four Levels of Detail.

Apart from reproductive cells (gametes) and mature red blood cells, every cell in the human body contains 23 pairs of chromosomes, each a packet of compressed and entwined DNA (1, 2).

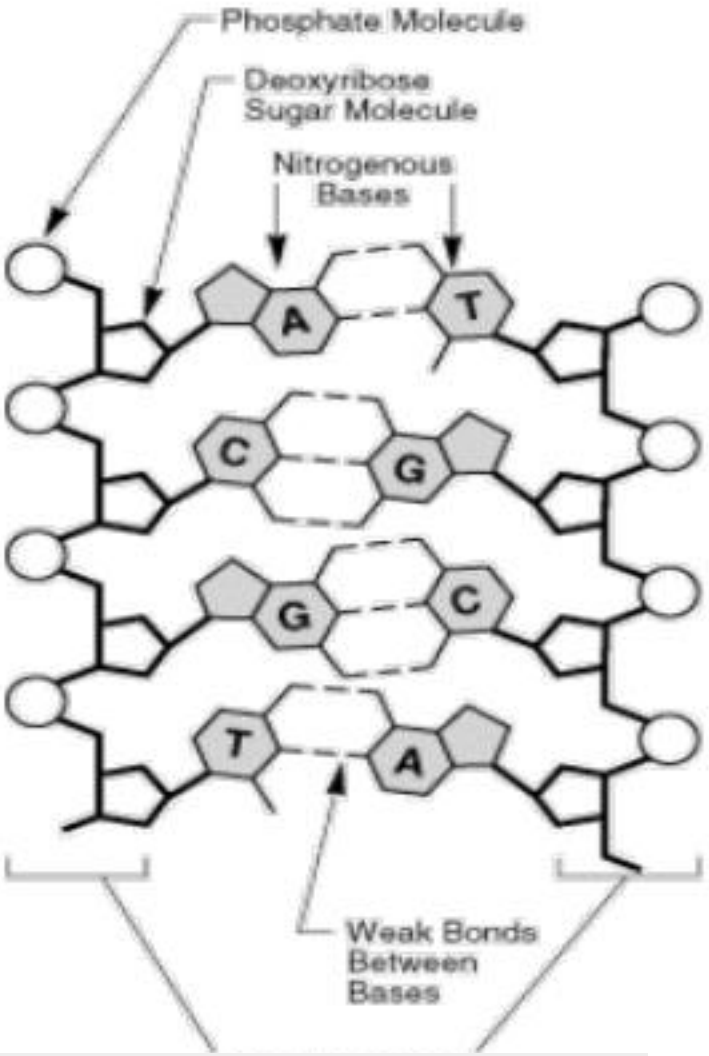
DNA ⇒ Structure and Components

- Complementary base pairs
  - (A-T and C-G) are connected by hydrogen bonds and the base-pair forms a coplanar "rung"
    - 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.

DNA ⇒ Structure and Components

- Chemically inert and mechanically rigid and stable
  - 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.
- Most uninteresting molecule:
  - "DNA, on its own, does nothing," smirked Natalie Angier recently. "It can't divide, it can't keep itself clean or sit up properly — proteins that surround it do all those tasks. Stripped of context within the body's cells ... DNA is helpless, speechless — DOA."

DNA Structure.



- The four nitrogenous bases of DNA are arranged along the sugar-phosphate backbone in a particular order (the DNA sequence), encoding all genetic instructions for an organism. Adenine (A) pairs with thymine (T), while cytosine (C) pairs with guanine (G). The two DNA strands are held together by weak bonds between the bases.

DNA ⇒ Structure and Components

- 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.

DNA ⇒ Structure and Components

- 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.

The Central Dogma

- 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 and Transcription

- 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).

RNA, Genes and Promoters

- 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

- 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;

Transcriptional Regulation

- The enhancer and silencer sequences
  - 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

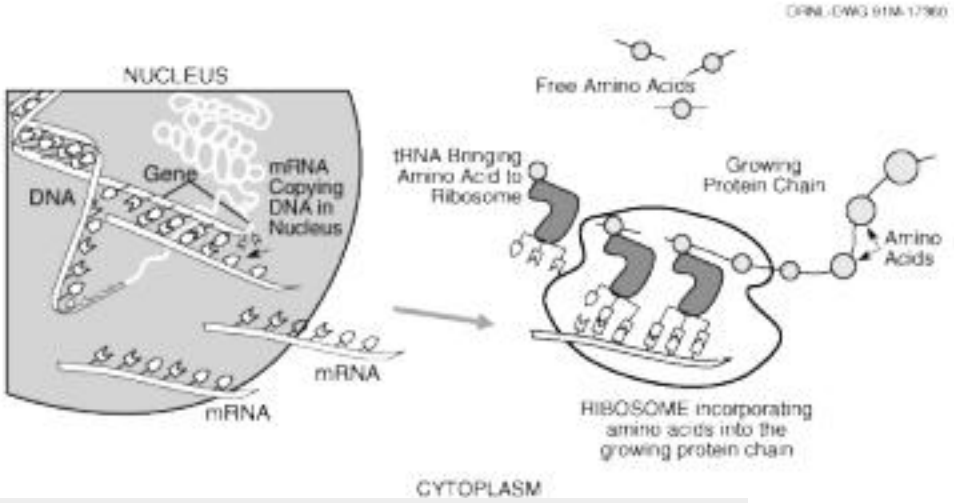
- The transcription of DNA in to mRNA
  - is performed with a single strand of DNA (the sense strand) around a gene.
  - This newly synthesized mRNA are capped by attaching special nucleotide sequences to the 5′ and 3′ ends.
- This molecule is called a pre-mRNA.

Transcription

- 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.

Gene Expression

- When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA, mRNA.
- The mRNAs leave the cell nucleus and enter the cytoplasm, where triplets of bases (codons) forming the genetic code specify the particular amino acids that make up an individual protein.
- This process, called translation, is accomplished by ribosomes (cellular components composed of proteins and another class of RNA) that read the genetic code from the mRNA, and transfer RNAs (tRNAs) that transport amino acids to the ribosomes for attachment to the growing protein.



Interrupted Genes

- 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.

Exons and Introns

- 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 (usually AUG) and is mediated by the transfer RNA (tRNA), made up of a group of small RNA molecules, each with specificity 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.

Protein and Translation

- The translation stops
  - when one of the three trinucleotides UAA, UAG or UGA is encountered.
- Codon
  - 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

- Redundancy in Codons:
  - 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.
- Open reading frame (ORF)
  - The line of nucleotides between and including the start and stop codons.

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

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

$\rightarrow \{A, R, D, N, C, E, Q, G, H,$

$I, L, K, M, F, P, S, T, W, Y, V\}$

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	Gla	glatamic 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	aspargining	AA(U+C)
P	Pro	proline	CC(U+A+C+G)
Q	Gln	glatamine	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)

The Cell

- Small coalition of a set of genes
  - held together in a set of chromosomes (and even perhaps unrelated extrachromosomal elements).
- Set of machinery
  - made of proteins, enzymes, lipids and organelles taking part in a dynamic process of information processing.

The Cell

- In eukaryotic cells
  - the genetic materials are enclosed in the cell nucleus separated from the other organelles in the cytoplasm by a membrane.
- In prokaryotic cells
  - the genetic materials are distributed homogeneously as it does not have a nucleus.
  - Example of prokaryotic cells are bacteria with a considerably simple genome.

Organelles

- The organelles common to eukaryotic plant and animal cells include
  - Mitochondria in animal cells and chloroplasts in plant cells (responsible for energy production);
  - A Golgi apparatus (responsible for modifying, sorting and packaging various macromolecules for distribution within and outside the cell);
  - Endpolastic reticulum (responsible for synthesizing protein); and
  - Nucleus (responsible for holding the DNA as chromosomes and replication and transcription).

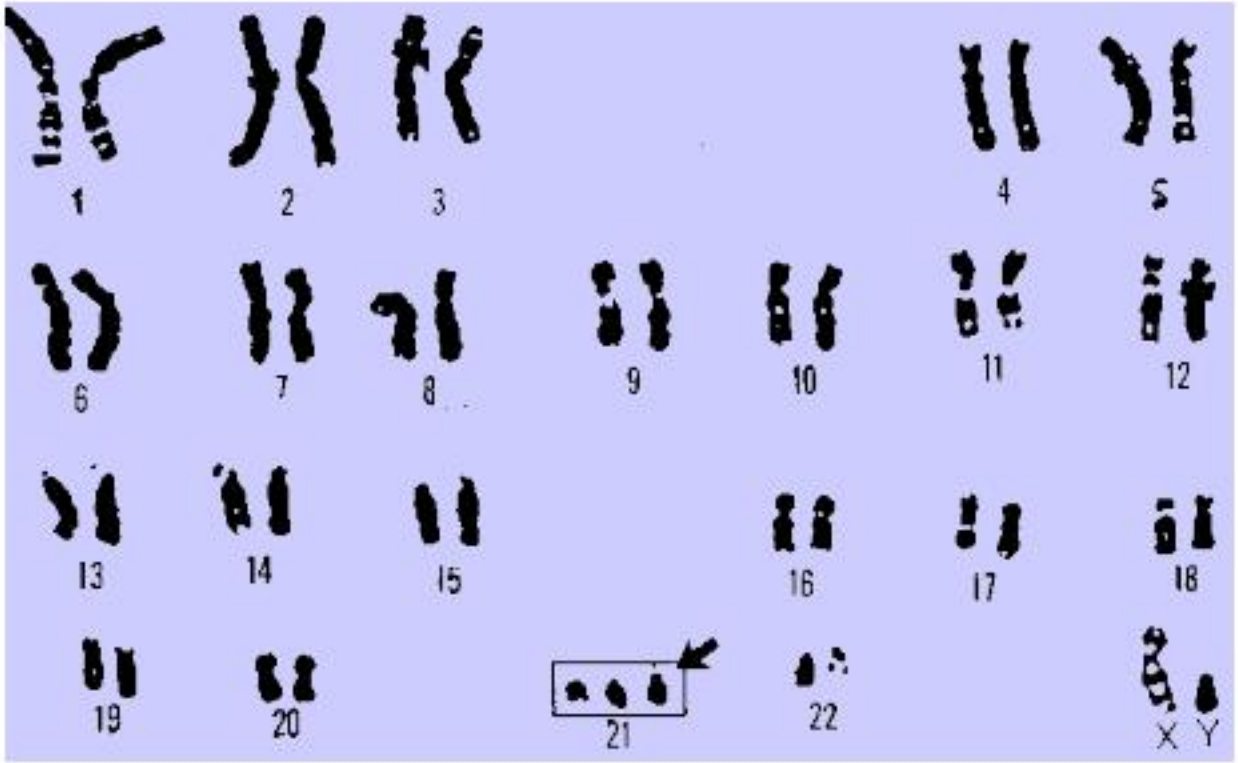
Chromosomes

- The entire cell
  - is contained in a sack made of plasma membrane.
  - In plant cells, they are further surrounded by a cellulose cell wall.
- The nucleus of the eukaryotic cells
  - contain its genome in several chromosomes, where each chromosome is simply a single molecule of DNA as well as some proteins (primarily histones).

Chromosomes

- The chromosomes
  - can be a **circular** or **linear**, in which case the ends are capped with special sequence of telomeres.
- The protein
  - in the nucleus binds to the DNA and effects the compaction of the very long DNA molecules.
- Ploidy
  - In somatic cells of most eukaryotic organisms, the chromosomes occur in homologous pairs,
  - Exceptions: X and Y –sex chromosomes.

Chromosomes



- Karyotype.
- Microscopic examination of chromosome size and banding patterns identifies 24 different chromosomes in a karyotype, which is used for diagnosis of genetic diseases.
- The extra copy of chromosome 21 (trisomy) in this karyotype implies Down's syndrome.

Ploidy

- Gametes contain only unpaired chromosomes;
  - the egg cell contains only X chromosome and the sperm cell either an X or an Y chromosome. The male has X and Y chromosomes; the female, 2 X's.
  - Cells with single unpaired chromosomes are called haploid;
  - Cells with homologous pairs, diploid;
  - Cells with homologous triplet, quadruplet, etc., chromosomes are called polyploid—many plant cells are polyploid.

Chromosomal Aberrations

- Point mutations
- Breakage
- Translocation (Among non-homologous chromosomes.)
- Formation of acentric and dicentric chromosomes.
- Gene Conversions
- Amplification and deletions
- Jumping genes a Transposition of DNA segments
- Programmed rearrangements a E.g., antibody responses.

Point Mutations

- In exon:
  - Can change the protein,
    - if it is transcriptional factor, it can affect many other genes
  - Can terminate the mRNA too early by changing a non-stop codon into a stop codon
    - (NMD: Nonsense Mediated Degradation)
  - Small indel can cause frame-shift, thus changing the entire protein

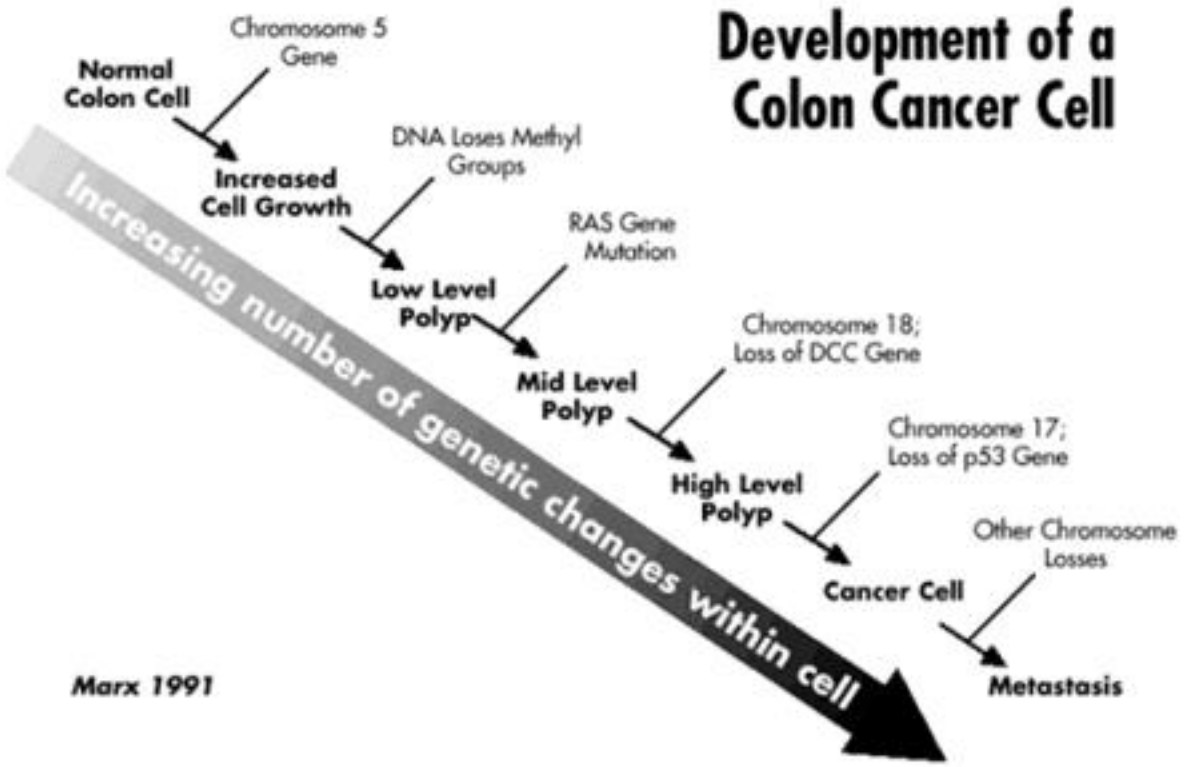
Point Mutations

- In promoter:
  - Can change its regulation: Over express, under express or silence
- In intron:
  - Can change the splicing patterns

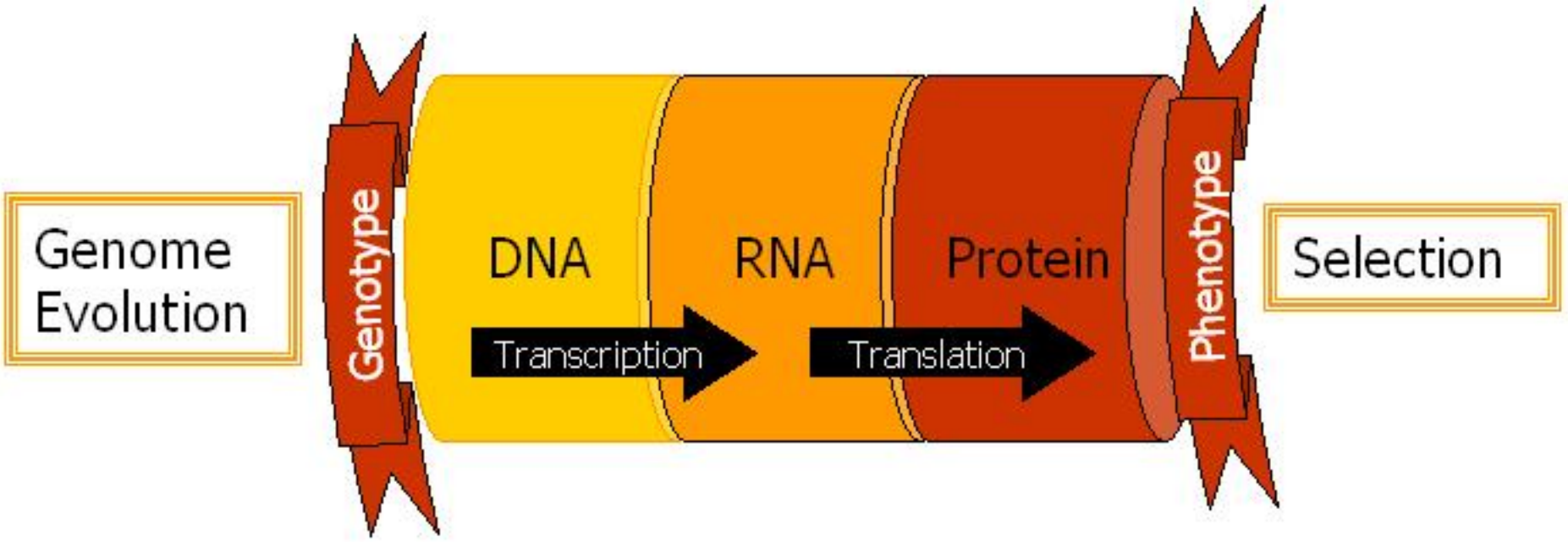
Loss or gain

- Translocation:
  - Can fuse two genes
  - Can activate a silent gene by placing it near some active regulatory region
- Amplification:
  - Over expression; Highly active gene
- Deletion
  - Under expression: Inactive gene

Amplifications & Deletions



The New Synthesis



Part-lists, Annotation, Ontologies

Cancer Initiation and Progression

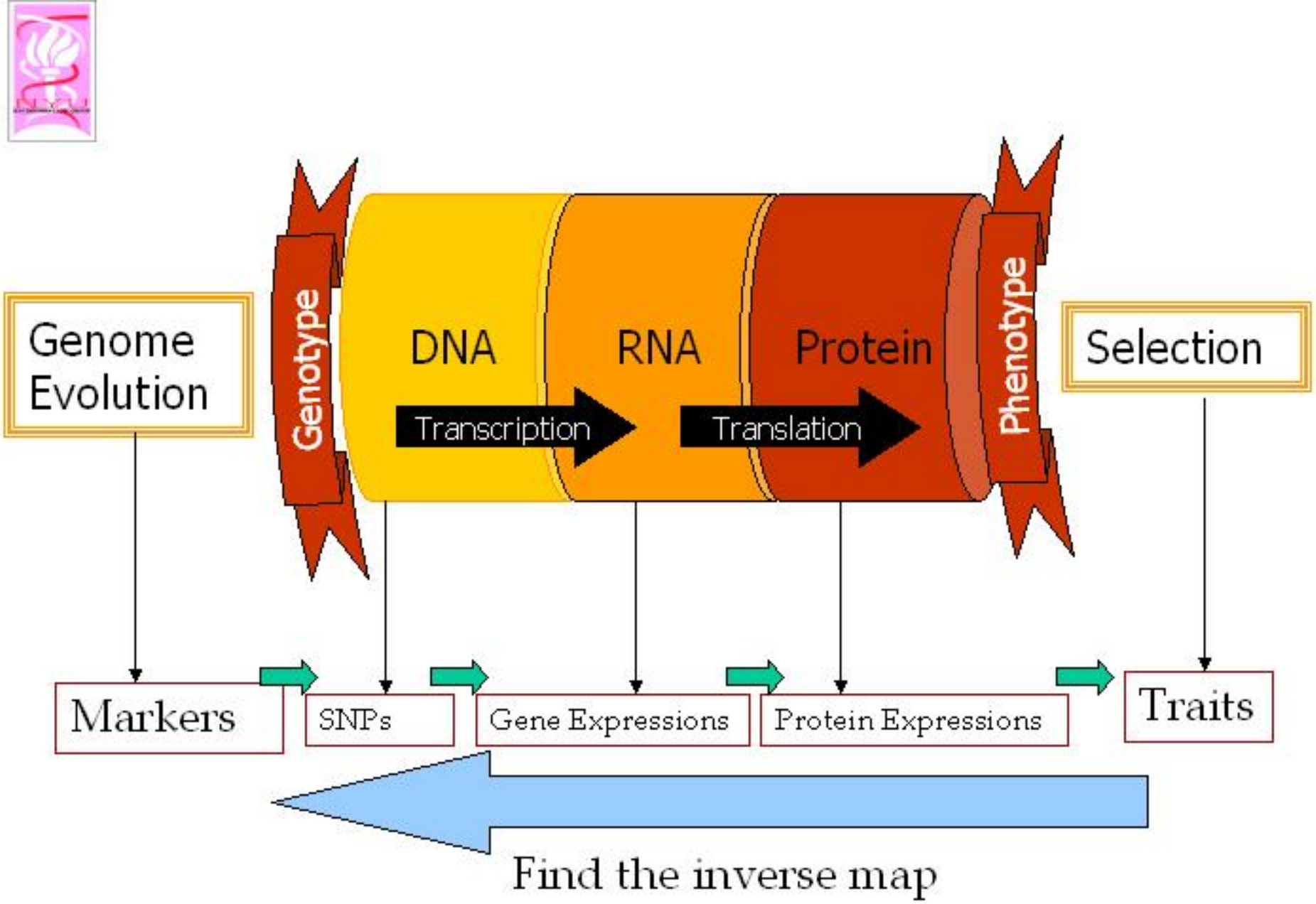
Mutations, Translocations, Amplifications, Deletions

Epigenomics (Hyper & Hypo-Methylation)

Alternate Splicing

Cancer Initiation and Progression

Proliferation, Motility, Immortality, Metastasis, Signaling

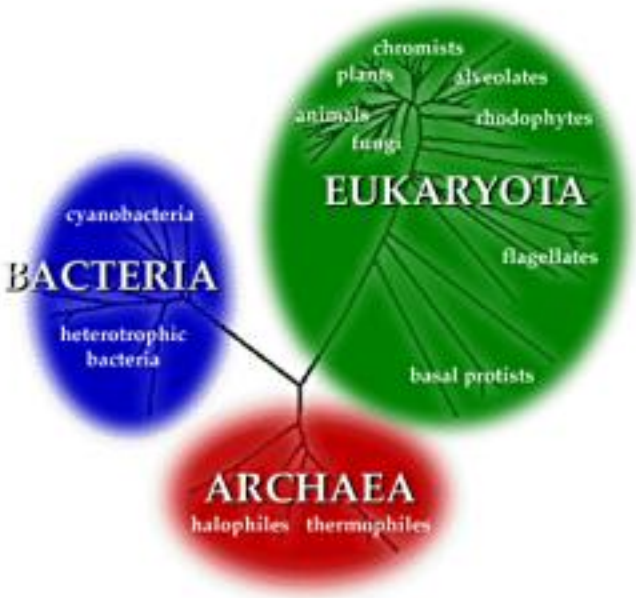


Molecular Evolution

Bio-Diversity

- Life is ubiquitous and old. (3.7 billion years old!)
- Living organisms on the Earth have diversified and adapted to almost every environment.
- All living organisms can replicate, and the replicator molecule is DNA.
- The information stored in DNA is converted into products used to build similar cellular machinery.
- Comparative study of the DNA can shed light on its function in the cell and the process of evolution.

Tree of Life



- All living organisms are divided into five kingdoms:
  1. Protista,
  2. Fungi,
  3. Monera (bacteria),
  4. Plantae, and
  5. Animalia.
- A different scheme:
  1. Prokaryotae (bacteria, etc.)
     1. Bacteria
     2. Archea
  2. Eukaryotae (animals, plants, fungi, and protists).
- No one of these groups is ancestral to the others.
- A fourth group of biological entities, the viruses, are not organisms...

Human Evolution

- Two Models:
  - Multiregional Model
  - Out of Africa Model
    - Evolution of a tree of hominids originating in Africa. Left Africa about 1 million years ago. Two waves of migration are speculated.
- African human population has the most diversity.
- Australopithecus (3.5million years old), Homo habilis (2 million yrs), Home erectus (1 million yrs), Homo sapiens (60,000-100,000 yrs)
  - Cro Magnon Man (Our immediate H. sapien ancestor)
  - Neanderthal Man (Became extinct ~30,000 yrs ago.)
- Two distinct species; supported by DNA amplification and sequence alignment (S. Paabo)

Mitochondria and Phylogeny

- **Mitochondrial DNA (mtDNA):** Extra-nuclear DNA, transmitted through maternal lineage. Mitochondria are inherited in a growing mammalian zygote only from the egg.
- 16.5 Kb, contains genes: coding for 13 proteins, 22 tRNA genes, 2 rRNA genes.
- mtDNA has a pointwise mutation substitution rate 10 times faster than nuclear DNA.
- Phylogeny based on human mtDNA can give us molecular (hence accurate?) information about human evolution.

African Eve

- Statistical analysis of mtDNA extracted from placental tissue of 147 women of different races and regions. (Cann, Stoneking, & Wilson, 87).
- Phylogenetic tree (assuming a constant molecular clock) was constructed by Wilson.
- A single rooted tree with the root being closest to the modern African woman.
- **Conclusion**: Modern man emerged from Africa 200,000 years ago. Race differences arose 50,000 years ago. "Mitochondrial Eve Hypothesis"

Mitochondrial Eve's Africanness

- A simple reordering of the data could result in 100 distinct trees al at most 2 steps away---all supporting non-African hypothesis. (Templeton)
- Assuming a non-constant molecular clock results in a least universal common ancestor (Luca) 105 to 106 years old.
- In general, mathematical descriptions and algorithms that may lead to "historically correct phylogenetic tree" remain to be developed.

Taxon

- **Taxon (Taxonomical Unit)**: is an entity whose similarity (or dissimilarity) can be numerically measured. E.g., Species, Populations, Genera, Amino Acid Sequences, Nucleotide Sequences, Languages.
- Phylogeny is an organization of the taxa in a rooted tree, with distances assigned to the edges in a such manner that the "tree-distance" between a pair of taxa equals the numerical value measuring their dissimilarity.
- The dissimilarity and the edge lengths of the phylogenic trees can be related to the rate of evolution (perhaps determined by a molecular

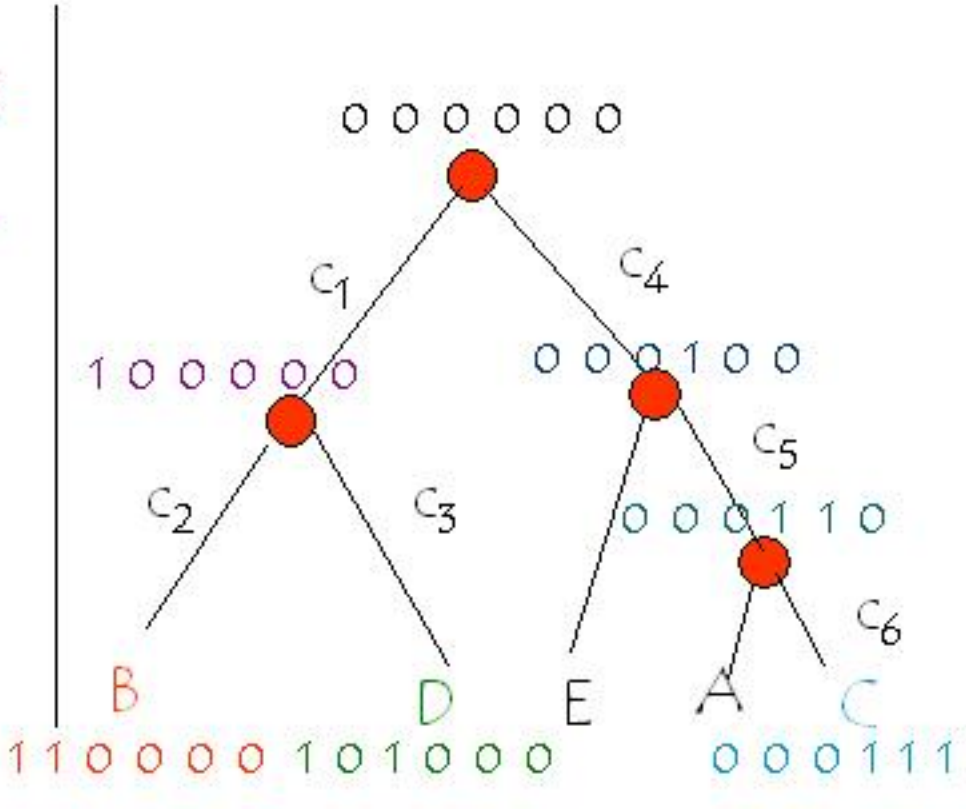
Comparing a Pair of Taxa

- *Discrete Characters*: Each taxon possesses a collection of characters and each character can be in one of finite number of states. One can describe an n taxa with characters by an $n \times m$ matrix over the state space. Character State Matrix.
- *Comparative Numerical Data*: A *distance* is assigned between every pair of taxa. One can describe the distances between n taxa by an $n \times n$ matrix over R_+. Distance Matrix.

Examples

A character state matrix

Taxon	c_1	c_2	c_3	c_4	c_5	c_6
A	0	0	0	1	1	0
B	1	1	0	0	0	0
C	0	0	0	1	1	1
D	1	0	1	0	0	0
E	0	0	0	1	0	0



Edges where state transition takes place is given by an associated character.

Character States

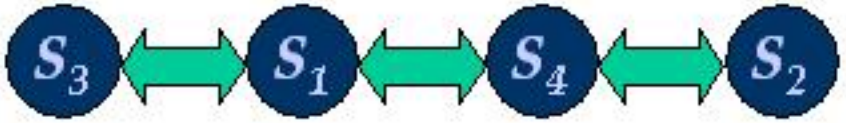
- Some Assumptions:
  - The characters are inherited independently from one another.
  - Observed states of a character have evolved from one "original state" of the nearest common ancestor of a taxon.
  - Convergence or parallel evolution are rare. That is the same state of a character rarely evolve in two independent manners.
  - Reversal of a character to an ancestral state is rare.

Classifying Characters

- Characters:
  - **Unordered / Qualitative Character**: All state transitions are possible.
  - **Ordered / Cladistic Character**: Specific rules regarding state transition are assumed.

    – *Linear Ordering*

    – *Partial Ordering* (with a derivation tree).



Perfect Phylogeny

- A **phylogenic tree** T (with edges labeled by state transitions) is called **perfect**, if it does not allow *reversal* or *convergence*--that is, with respect to any character c and any pair of states w and s at most one edge is labeled

$$w \rightarrow s \text{ or } s \rightarrow w.$$

- **Example**: Binary characters with two states {0=ancestral, and 1=derived}: any character c_i labels at most one edge and implies a transition from

$$0 \rightarrow 1 \text{ in the } i^{th} \text{ position.}$$

- *Perfect Phylogeny Problem:*
  - **Given**: A set O with n taxa, a set C of m characters, each character having at most r states.
  - **Decide**: If O admits a perfect phylogeny.
- A set of defining characters are **compatible**, if a set of objects defined by a character set matrix admits a perfect phylogeny.

Compatibility Criteria

- Allow reversal and convergence properties in the models of evolution.
- *Parsimony Criteria*: Minimize the occurrences of reversal and convergence events in the reconstructed phylogeny tree.
  - **Dollo Parsimony Criterion**: Minimize reversal while forbidding convergence.
  - **Camin-Sokal Parsimony Criterion**: Minimize convergence while forbidding reversal.
- *Compatibility Criteria*: Exclude minimal number of characters under consideration so that the reconstructed phylogeny tree is perfect and does not admit any occurrence of reversal or convergence.

Computational Infeasibility

- Perfect Phylogeny Problem for arbitrary (>2) number of unordered characters and arbitrary (> 2) number of states is NP-complete.
- Optimal Phylogeny Problem under compatibility criteria is NP-complete.
- Optimal Phylogeny Problem either under Dollo or Camin-Sokal parsimony criteria is NP-complete.

Binary Character Set

- Each character has two states = {0, 1}
- If a character is ordered then $0 \rightarrow 1$ (0=ancestral and 1=derived), or converse.
- For binary characters (ordered or unordered), perfect phylogeny problem can be solved efficiently
  - Poly time, for n taxa and m characters, Time = $O(nm)$.
- A two phase algorithm:
  1. Perfect Phylogeny Decision Problem
  2. Perfect Phylogeny Reconstruction Problem

Compatibility Condition

- T = Perfect Phylogeny for M iff

$$(\forall\ c_i = \text{character})(\exists!\ e = \text{tree-edge})\ \text{label}(e) = \{c_i, 0 \rightarrow 1\}$$

$$\text{root}(T) = (0, 0, 0, \ldots, 0)$$

- A path from root to a taxon t is labeled $(c_{i1}, c_{i2}, \ldots, c_{ij})$

$\Rightarrow$ t has 1's in positions $i_1, i_2, \ldots, i_j$.

- **Perfect Phylogeny Condition**
  - $M = n \times m$ Character State Matrix, $j \in \{1..m\}$
  - $O_j = \{i = \text{taxon} : M_{ij} = 1\}$
  - $O_j^c = \{i = \text{taxon} : M_{ij} = 0\}$

Key Lemma

- **Lemma**: *A binary matrix M admits a perfect phylogeny iff*

  $(\forall\, i, j \in \{1, m\})\ O_i \cap O_j = \emptyset$ or $O_i \subseteq O_j$ or $O_i \supseteq O_j$

- **Proof**: ($\Rightarrow$) T_i = subtree containing O_i, T_j = subtree containing O_j

  r_i = root(T_i) and r_j = root(T_j)

  r_i is neither an ancestor nor descendant of $r_j \Rightarrow O_i \cap O_j = \emptyset$

  r_i is a descendant of $r_j \Rightarrow O_i \subseteq O_j$

  r_i is an ancestor of $r_j \Rightarrow O_i \supseteq O_j$

- ($\Leftarrow$) By induction, Base case m=1 is trivial. Induction case, $m = k+1$:

  T_k = Tree for k characters. O_{k+1} is contained in a subtree with minimal # taxa rooted at r.

  r must be a leaf node. Either an edge needs to be labeled or the subtree rooted at r has to be split. □

Simple Algorithm based on the Lemma

- Compare every pair of columns for the intersection and inclusion properties. Total of $O(m^2)$ pairs, each comparison can be done in $O(n)$ time.
- Total Time Complexity = $O(nm^2)$
- Can be improved to $O(nm)$ time.

Rate of Evolutionary Changes

- Taxa of nucleotide or amino acid sequences.
- Given two taxa s_i and s_j, measure their distance
  - Distance(s_i, s_j), d_{ij} = Edit distance based on pairwise sequence alignment.
- Assumptions about the Molecular Clock (governing rate of evolutionary change):
  - Only independent substitutions
  - No back or parallel mutations
  - Neglect selection pressure.

Amino Acid Sequences

- λ = Amino Acid substitution rate per site per year.
- λ varies between organisms and protein classes
- Example:
  - λ for guinea pig insulin $\approx 5.3 \times 10^{-9}$
  - λ for other organisms $\approx 0.33 \times 10^{-9}$
- Other Examples of λ:
  - Fibrinopeptide $\approx 9 \times 10^{-9}$
  - Histone $\approx 1 \times 10^{-11}$

Estimating λ

- X & Y = homologous proteins of same length n
- n_d = Number of differences between homologous amino acid sites.
- X and Y are isolated from two distantly related species that diverged t time ago.
- $p \approx n_d / n$ = Probability of an amino acid substitution occurring at a given site of either X or Y.

Estimating λ (Contd.)

- $q = 1 - p = 1 - n_d/n$ = Pr[# mutations at site $X_i = 0$]

  $\times$ Pr[# mutations at site $Y_i = 0$]

- Z = Random variable counting the number of mutations over time t at a fixed site for an amino acid sequence with substitution rate λ per site per year ~ Poisson(λ t)

$$\Pr[Z = k] = \exp\{-\lambda t\} (\lambda t)^k / k!$$

$$q = e^{-2\lambda t}$$

$$\lambda = \ln(1/q)/2t$$

Example: Histone H4

- X & Y = Hisones from cow and pea.
- $n = 105$, $n_d = 2$, $q = 1 - n_d/n = 103/105$
- $t = 10^9$; Plants and animals diverged about a billion years ago.

$$\lambda = (1/2t)(-\ln(1 - n_d/n))$$

$$\approx (n_d/n)/(2t)$$

$$\approx (2 \times 10^{-2})(2 \times 10^9) \approx 10^{-11}$$

To be continued…

…