Bioinformatics - Lectures 5 and 6

Louis Wehenkel

Department of Electrical Engineering and Computer Science
University of Liège

Montefiore - Liège - October 23, 2007

Find slides: *http://montefiore.ulg.ac.be/∼lwh/IBIOINFO/*

Chapter 5. Variation within and between species
- Variation in DNA sequences
- Mitochondrial DNA
- Dynamics of genetic variations (for haploids)
- Estimating genetic distance

Chapter 6. Natural selection at the molecular level
- About the HIV virus, and viruses in general
- Quantifying natural selection on DNA/RNA sequences
- Estimating K_A/K_S

Homework for chapters 5 and 6

Variation within and between species

Objective:

- How do DNA sequences change over time ?
- How can we use this information to infer history and functions of different parts of the genome ?

NB. In this chapter we will focus on variation without considering the effect of selection.

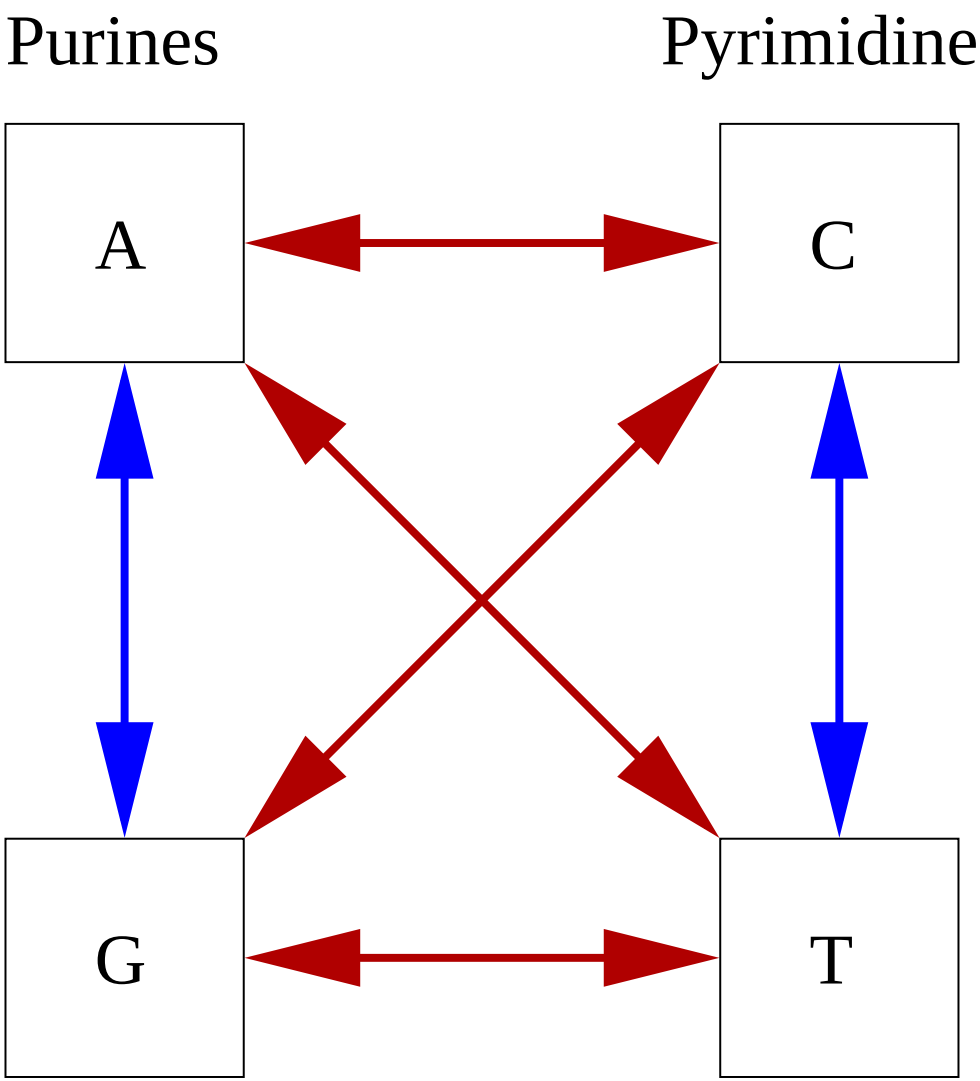
Variation in DNA sequences

- Every individual has a slightly different genome
- Genomic variations are due to
  - mutations (rates are variable, but often very low e.g. 10^{-9})
  - recombinations (in diploid species)
- Analysis of variation can be made by comparing two or more genomic sequences
- Because mutation rates are very low, most of the differences between two individuals are actually inherited from their parents
- Shared mutations are thus indicative of shared ancestry

Mutations and SNPs

- Germline vs other mutations: every (germline) mutation first appears in some individual and some are passed to offsprings
- If we observe differences between individuals (of a species) at a certain location, we call this a polymorphism or a polymorphic site
- Single nucleotide polymorphisms (SNPs)
  - Mutation at a specific site: a base is replaced by another base
  - Observed bases in species are called the alleles of the SNP
  - Human nuclear genome: about 1SNP every 1500 bases
  - Density of SNPs is lower for younger species (e.g. human)
- Other types of mutations/polymorphisms:
  - Short tandem repeats (number of repeats vary): much higher mutation rates (used for forensic DNA based identification)
  - Copy number variations, inversions, duplications, transpositions, Indels ...

Transitions vs transversions



In humans transitions are twice as likely as transversions

Genetic code is more robust to transitions

DNA and AA mutation and substitution matrices

- DNA (4 by 4 mutation rate matrix)
  - Most detailed: all mutations use different rates
  - Most simple: only one rate for all mutations (BLAST default)
  - Compromise: distinguish only between transitions and transversions
- Proteins (20 by 20 substitution rate matrix)
  - PAM: obtained from global alignments of very close protein sequences (e.g. less than 1% divergence for PAM1)
  - BLOSUM: obtained from local alignments of sequences of differing similarity (e.g. at least 62% identical for BLOSUM62)

Mitochondrial DNA: a model for the analysis of variation

- We will focus on mtDNA to study human evolution and compare species
- 16569 bp long circular genome, received from mother only
- Reasons for focusing on mtDNA:
  - Higher mutation rate (one order of magnitude at least)
  - Only inherited from mother without recombination (makes analysis simpler, no haplotype inference problem)
  - Several copies in each cell and small size (makes sequencing easier)
- Genetic code of mitochondrial DNA is slightly different.
- Focus on specific (hypervariable) regions in D-Loop (1,1 kb).

Long term dynamics induced by mutations

- Stochastic model for infinite population size with mutations:
  - Assume one locus with two alleles $A1$ and $A2$.
  - Assume that there is no selective pressure.
  - Assume mutation rates $u = P(A1 \rightarrow A2), v = P(A2 \rightarrow A1)$.
  - Assume random reproduction, and denote by $f(t)$ the proportion of allele $A1$ in population at generation t.
  - We have, $f(t+1) = f(t)(1-u) + (1-f(t))v$.
  - Fixed point: $f(t+1) = f(t) \Leftrightarrow f(t) = v/(u+v)$
  - Steady state distribution $(v/(u+v), u/(u+v))$: proportions of $A1$ and $A2$ such that they are preserved at the next iteration.
  - One can show that independently of initial allele frequencies, the process will converge to this steady state distribution.

Thus, we will have a proportion $f(\infty) = \frac{v}{u+v}$ of alleles $A1$ after sufficiently (in practice very) long time.

Concl. Mutation (even very small rates, e.g. 10^{-6}) maintains genetic diversity in a population.

Random genetic drif in finite populations

- Model for finite population size without mutations:
  - Denote by $f(t)$ the relative number of individuals of genotype $A1$ at generation t in a population of size N.

Then, for generation $t+1$ we will have (binomial distribution) $P(f(t+1) = j/N | f(t) = i/N) = C_N^j f(t)^j (1 - f(t))^{N-j}$.

  - Variable $f(t)$ follows a Markov chain, which has two absorbing states: $f = 0$ and $f = 1$.

Thus, after sufficiently many generations either allele $A1$ or $A2$ will be fixed in the population.

  - These two possibilities occur respectively with probability $f(0)$ and $(1 - f(0))$.
  - If we split the population in two parts at some stage, these two parts will evolve along different trajectories.

Concl. Genetic drift tends to reduce genetic diversity inside a population, but allows for genetic diversity among populations. The drift is faster for smaller populations.

Further complications

- To take into account selection, we need to model how phenotypes induced by genotypes may influence the probability of survival and reproduction.
- In practice, we need also to consider the combination of mutations with the fact that populations are of finite size.
- If we consider several loci at the time, or diploid populations, further complications arise (e.g. recombination).
- These refinements are out of the scope of this course.

Estimating genetic distance

- Going back in time, from two homologous sequences $\mathbf{s}_1$ and $\mathbf{s}_2$, at some stage we will find a common ancestor.
- Genetic distance K of two homologous sequences $\mathbf{s}_1$ and $\mathbf{s}_2$
  - Total number of substitutions per unit of sequence length, that have occurred since common ancestor to produce $\mathbf{s}_1$ and $\mathbf{s}_2$.
  - Not equal to fraction of sites with observed nucleotide differences d between $\mathbf{s}_1$ and $\mathbf{s}_2$, because of the possibility of having multiple changes over time on a single locus.
  - K is thus a hidden variable that can only be estimated.
- Knowing K could allow us to estimate the number of generations since when the sequences have diverged.

NB. If d is very small, then it is generally a good estimate of K

Jukes-Cantor model

- $K = -3/4 \ln(1 - 4d/3)$
  - If $d = 0.5$: $K = 0.82396$;
  - $d = 0.01$: $K = 0.01007$.
- Assumptions of derivation (see book):
  - Independent mutations at sites which obey to Markov chain with $M = \begin{bmatrix} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{bmatrix}$
  - Sequence is long enough, so that $d \approx 3/4 - 3/4(1 - 4\alpha/3)^t$ after t time steps.
  - If t is large enough, we have also $K \approx t\alpha$
  - Since α is very small, we have $\ln(1 - 4\alpha/3) \approx -4\alpha/3$
  - Putting all this together, we obtain
    - $t = -\frac{3}{4\alpha} \ln(1 - 4d/3)$ and
    - $K = -3/4 \ln(1 - 4d/3)$.

Kimura two-parameter model

- Assumes different rates
  - for transitions (α) and
  - for transversions (β).
- Similar analysis to JK yields
  - $K = -\frac{1}{2}\ln(1-2P-Q) - \frac{1}{4}\ln(1-2Q)$ where
  - P is the fraction of nucleotides with observed transitions and
  - Q is the fraction of nucleotides with observed transversions.

NB. It is interesting to observe that the mutations rates do not intervene in the relationship between K and d. They are however necessary to estimate number of generations from d (resp. P and Q).

Natural selection at the molecular level

Objective:

- A first exploration on the effect of natural selection on sequence variation.
- Explain why we need to look at amino-acid sequences to do this properly.
- Highlight applications of such analyses by looking at the evolution of HIV.
- Discuss in this context, the interest of localized sequence variation analysis.

About the HIV virus

- HIV is one of the major sources of death, specially in developping countries
  - 60% of population is infected in some African nations
  - Current treatments are heavy and very expensive; they have to be observed during life-time
- HIV was identified and sequenced in the early eighties; since then it has been intensively studied:
  - HIV is a 9,193 bp RNA retroviros
  - HIV 'attacts' the host immune system
  - It has high mutation rates, and short generation times
  - It is therefore able to evolve quickly enough to resist to the reaction of the immune system and the effect of drugs
  - For each infected individual, the virus evolves over time
  - Need to understand how the virus evolves in order to find better treatments, and possibly vaccine.

Virus genomes and natural selection

- The stability of some parts of a virus genome are fundamental to its operation
- The instability of some other parts are essential for its survival against drugs, and against the action of the immune system
- Theory of evolution (Darwin)
  - Variability among individuals comes from mutations
  - Depending on the environment, some mutations are favorable and some are deleterious
  - The former are positively selected, and the latter are negatively selected

Stable versus unstable parts of the genome and natural selection

- Some mutations have no effect on protein (cf genetic code):
  - they are called synonymous.
- Some other variations do have an effect:
  - they are called non-synonymous.
- By studying the ratio of synonymous/non-synonymous mutations along the genome inside a Virus species,
  - one can identify which variations are favorable to the survival of the virus and which are not;
  - one can infer which regions play which biological role.

Immune system response

- Translation of the RNA of the virus yields proteins, which are specific to the virus (not found in the uninfected organism).
- Some of these proteins, called epitopes, stick out of the cell membrane and may be recognized as 'foreign' proteins by the immune system of the host organism.
- The immune system of the host reacts when such epitopes are recognized by generating 'epitope specific' killer cells.
- The epitope proteins can mutate in order to escape the killer cells. In that case, the cellular machinery is used by the virus to replicate itself and infect other cells.
- The success of either virus or immune system depends on the relative rate of epitope mutation versus killer cell production.

Quantifying natural selection on DNA/RNA sequences

How do we know which mutations are neutral, deleterious, or advantageous?

- Mesure the fitness of mutants: but this is not always possible.
- Compare the rate of non-synonymous and synonymous mutations in a single gene, or genomic area.
  - We will assume that synonymous mutations are selectively neutral.
  - We will assume that non-synonymous ones are not neutral.
  - If in a given protein coding area of the genome, we observe that the former ones are more frequent than the later ones, we will conclude that selective pressure is against variation, and vice-versa.
  - More precisely, we will compare the ratio of both types of mutations to the the expected ratio under some random model that is neutral from the selective point of view.

Estimating K_A/K_S

- Principle of Nei-Gojobori algorithm (see book for details)
  - Distinguish between subsitution rate of synonymous (S) and non-synonymous (A) differences/substitutions/sites.
- Outline
  - Align the two sequences ($\rightarrow$ codon-codon correspondence)
  - Count the number of A and S "sites" (A_c and S_c) for each sequence
  - Compute averages over the two sequences $\hat{A}_c$ and $\hat{S}_c$.
  - Count A_d and S_d, the numbers of actual codon "differences" between the two sequences.
  - Compute $d_a = A_d/\hat{A}_c$ and $d_s = S_d/\hat{S}_c$.
  - Apply JK correction to get K_A and K_S.

Homework for chapters 5 and 6

- Deadline November 20 (no course next week).
- Read chapters 5 and 6 (specially sections 6.5 and 6.6).
- Play with case studies "Jukes-Cantor", "Neanderthal" and "HIV" (book website) to get familiar with notions and tools.
- Carry out exercises of section 6.7 and write report on this.