

*Understanding the role of germline
selection on patterns of linkage in
HapMaP III data*

Dibyoy Roy

DIMACS-REU Program

6/11/10

Natural Selection

- The phenomenon in which organisms that are best suited to survive in their environment due to certain traits are more likely to proliferate and transmit their genes, while the other organisms tend to be eliminated.
- Mechanism by which Darwin's Theory of Evolution occurs
- The observable traits are called phenotypes and the actual gene causing that expression is the genotype.
- Those organisms with a desirable phenotype have a higher genetic fitness (X) level than those that have a less desirable one.

Natural Selection (cont.)

- Individuals that possess the advantageous form of a gene (an allele) will help to transmit their version of the gene to further generations until it becomes commonplace amongst the population.
- Ex: Peppered Moth in UK
- Also new alleles can be introduced into a population by mutation or recombination, which in rare cases can create an even more advantageous phenotype for the species.

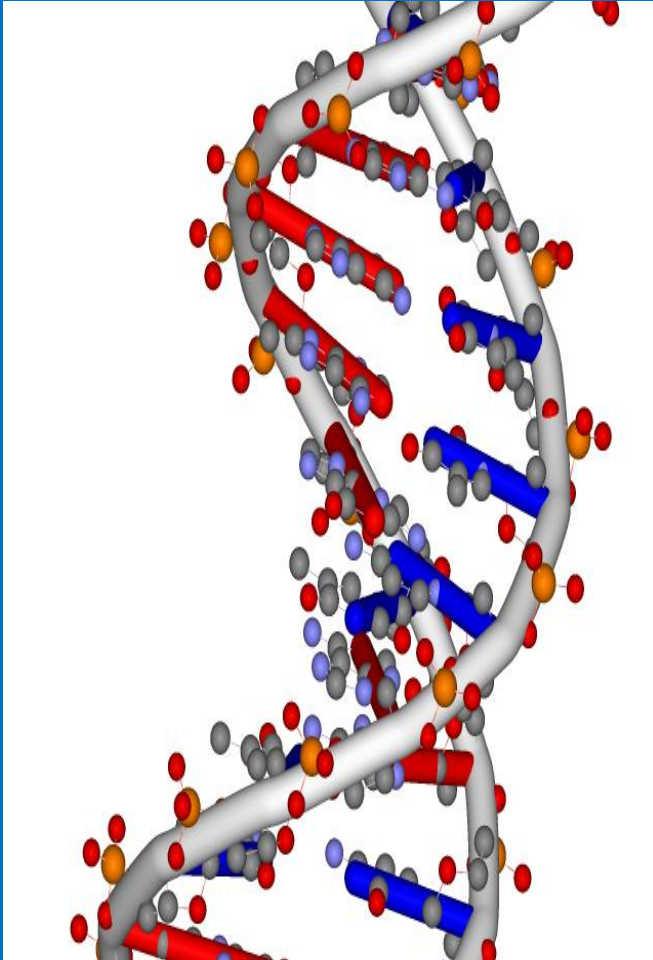


http://economicdarwinism.files.wordpress.com/2009/03/darwin_bicentennial.png?w=550&h=453

We know what everyone else is doing...

- The HapMap Project is the development of a haplotype map that will show common patterns in human genetic variation.
- Since all humans share about 99.5% of their DNA sequence, the project focuses on the 0.5% of SNPs that are different in at least 1% of the world population. Each SNP is known as an allele.
- Usually SNPs are correlated with other SNPs nearby in the sequence. These consecutive alleles are known as a haplotype.
- Disease can then be analyzed for which SNPs and haplotypes are major factors behind the emergence of the disease.
- The 1000 Genomes Project is a similar project, but where the full genome of 1000 individuals will be determined.

..... but, what are you doing?



http://geoffhutchison.net/gallery/main.php?g2_view=core.DownloadItem&g2_itemId=372&g2_serialNumber=2

- Using Matlab and other software, I will be artificially creating “populations” of about 100 to 1,000 “individuals” represented by sequences of 100 to 500 “nucleotides”.
- These “individuals” will then be “evolved” over many generations with or without selection pressure from “environmental factors” induced arbitrarily.

And why?



<http://bioinfo.au.tsinghua.edu.cn/member/xwang/image/HapMap.jpg>

- The goal of the project is to determine algorithms to assess the fraction of loci under selection and their possible locations from the sequences that remain after many generations.
- The project will demonstrate a basic model for how to identify selection from the large amounts of sequence data which is expected to become available from “HaPMaP” and the “1000 genomes project”.

Let's Go Into Detail

- DNA has 4 nucleotides (Adenine, Guanine- Purines; Cytosine, Thymine- Pyrimidines) that comprise its code
- Most mutations (95%) are transition mutations (changing purine to other purine, pyrimidine to other pyrimidine), and the effect is not severe compared to transversions (changing purine to pyrimidine or vice versa)
- As a result, it will be adequate to use a binary system in the created sequences, in place of the four nucleotide system.

A

T

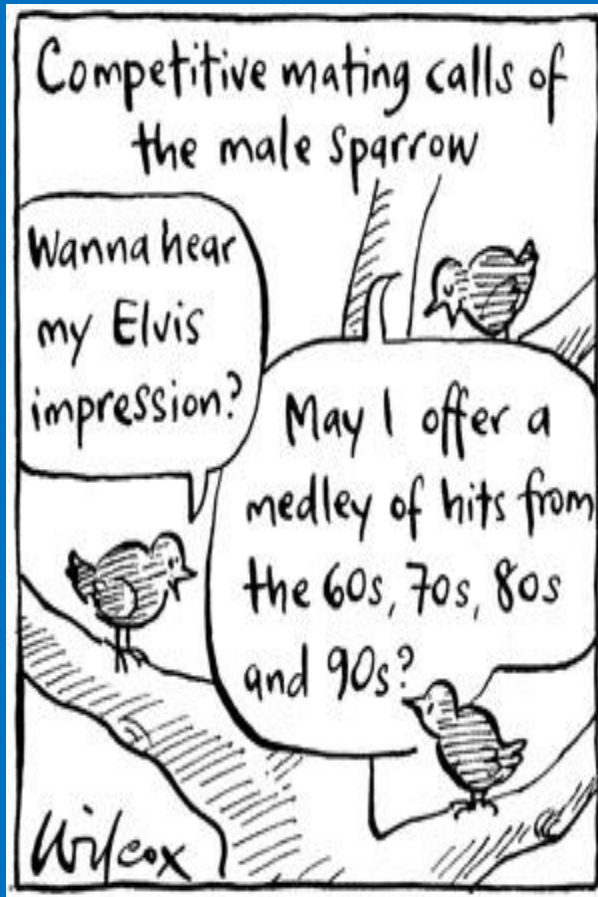
C

G

Let's Go Into Detail (cont.)

- A reference sequence will then be created (a string of zeros) and at initial time ($t=0$) in the simulation, this reference sequence will represent the “optimal fitness” level for surviving in its “environment”.
- The main focus of the project will then be to model the evolution of this population under selection pressure due to changes in the “environmental condition”. Such changes can be modeled by altering the “optimum fitness” string.

Let's Go Into Detail (cont.)



http://blogs.smh.com.au/sit/stay_cartoon.jpg

- The initial population consists of individuals who deviate from optimality by some random amount, represented by a random variable X .
- X (fitness) will help determine which sequences will be allowed to "mate" with each other.

Variables

- N- Number of individuals
- M- Number of units in sequence of individual
- The initial population will be set up so that there will be variation amongst the individuals. In each individual, a given locus will have a probability **p** of differing from the reference sequence.
 - This will create a “population” with individual of differing fitness levels like most real-world populations.



<http://www.coolopticalillusions.com/illusions/abc123.gif>

Calculating Fitness (X)

- Thus, if $A(j,k,t)$ is a binary variable (it only takes values 0 or 1) representing the nucleotide at position k ($k=1,2,...M$) in individual j ($j = 1,2,...N$) at time t , and if $A_0(k, t)$ ($k = 1,2,...M$) is the “optimum sequence” at time t , then,

- $$X(j,t) = 1 - \frac{[\sum_k \text{XOR} (A(j,k,t), A_0(k,t))]}{M} \quad (1)$$

- Where XOR is the Exclusive OR function, represented by the rule:
 $\text{XOR}(0,0) = \text{XOR}(1,1) = 0$ and $\text{XOR}(0,1) = \text{XOR}(1,0) = 1 \quad (2)$

- Note that as defined, X takes values in $(0,1)$.

More Variables

- The reference sequence will also be changed at **f** number of sites, which represents the “environmental changes” creating a new definition of “optimal fitness” for the “population”.
 - This is representative of which genotype would be most advantageous in a population due to selective pressure.
- The “offspring” will have a mutation rate, **u**, that will randomly change some of the “nucleotides” in their sequence from 0 to 1 or vice versa.
 - These mutations may increase or decrease fitness depending on how the variable **f** changes.

Procedure

- Given the population at time t , the population at time $t+1$ is generated as follows:
 1. Compute the fitness X for each individual at time t .
 2. Select individuals from the population such that the probability to be selected is proportional to fitness.
 3. A pair of individuals so selected at t will produce a pairs of individuals at $t+1$. This will be done by swapping a random piece of the genome between these individuals.
 4. Induce mutations in the newly created individuals with probability u per nucleotide for a $0 \leftrightarrow 1$ change in their sequence.
 5. Repeat for further generations.

Endgame

- *The overall task of this project is to develop a simulation using Matlab which models the time course of the adaptation of the population to this change as a function of p , u , f , N and M .*
- After a number, t , of generations, phylogenetic trees will be constructed (utilizing Phylip software) using the sequences generated throughout the project and subsequently analyzed for patterns showing which loci were central for the evolution of the population.
- In the end, a similar model will have been constructed to that of the one which researchers will utilize when analyzing Hapmap data for their own research.

Sources

1. <http://cplus.about.com/od/glossar1/g/xor.htm>
2. http://evolution.berkeley.edu/evolibrary/article/0_0_0/evo_25
3. <http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html.en>
4. <http://www.1000genomes.org/page.php?page=about>
5. http://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html
6. <http://members.logical.net/~marshall/AbramsRefEnv.pdf>