NAME_____TA_____ SEC____

7.012 Problem Set 7 FRIDAY December 3, 2004

Not due unless you have had a prior agreement with Claudette Gardel
Solutions will be posted on the web.

Question 1

Leukemia is type of cancer characterized by the uncontrolled proliferation of white blood cells (leukocytes). Chronic Myelogenous Leukemia (CML) is a type of leukemia primarily caused by a translocation that relocates an oncogene, called *abl*, from the long arm of chromosome 9 to the long arm of chromosome 22 in the *bcr* region (breakpoint cluster region). The resulting *bcr-abl* fusion protein encodes a constitutively active tyrosine kinase, which when expressed, leads to the CML phenotype.

a) Gleevec is an effective drug treatment of CML. What design principles were applied to the development of Gleevec? How is this approach different from that used to develop conventional cancer therapies?

Rational drug design was applied to the development of Gleevec. The idea is to design a drug that specifically targets the underlying molecular defect, rather than nonspecifically target all dividing cells, as do many conventional cancer therapies

b) How does Gleevec work at the molecular level?

Gleevec works by binding to the ATP binding site of bcr-abl thus inhibiting its tyrosine kinase activity.

c) How many protein targets does Gleevec have? Name each target.

Gleevec does bind to two other tyrosine kinases: kit and the PDGF receptor.

d) Besides target specificity, what other characteristics of a drug should you consider when designing a therapy?

Toxicity, absorption efficiency, is it metabolized, possible reactivity of metabolites, etc.

In contrast to drug treatments, gene therapy attempts to correct the defect by introducing a functional copy of the malfunctioning gene that is responsible for the disease phenotype. The functional gene copy can be introduced directly into the diseased organ or can be used to genetically modify isolated tissue that is later re-introduced into the patient.

e) What do you need to know about the target disease in order to apply gene therapy?

You need to know the gene involved in the disease. In particular, the disease should be caused by a defect in a single gene.

f) How could you deliver the functional gene into the diseased cells?

A viral vector could be used (example: adenovirus).

g) What do you think are some of the challenges facing gene therapy?

Challenges include efficient uptake of the functional gene, stability of the gene, long term expression, side effects of viral vector.

h) Based on your understanding of oncogenes, why would *bcr-abl* be a particularly challenging target for gene therapy?

You would have to replace the fused bcr-abl gene with a functional copy of the abl tyrosine kinase, not simply introduce the functional copy into the cell.

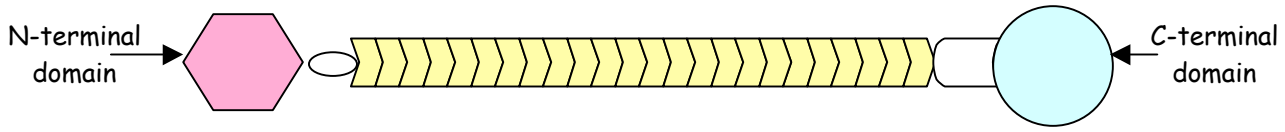
Question 2

Duchenne muscular dystrophy (DMD) is an X-linked recessive disorder caused by mutations in the gene encoding dystrophin, a protein involved in maintaining membrane integrity in muscle cells. The dystrophin gene spans roughly 2.5 Mb and is spliced to form a 14 kb mRNA transcript consisting of 79 exons.

a) Is dystrophin is a typical human gene in terms of its size and exon count?
Dystrophin is encoded by one of the largest human genes. Although there is great variation both in gene size and exon count in the human genome, the average human gene length is roughly 10-15 kb and average exon count is approximately 4-8 exons per gene.

b) Why is DMD much more common in boys than in girls?
The gene is on the X chromosome and confers a recessive phenotype. Females have 2 X chromosomes
c) The DMD phenotype results from mutations that disrupt the reading frame of the dystrophin mRNA. What is the impact of such mutations on the dystrophin protein?
A frameshift mutation in the gene would result in a truncated protein product

d) Based on what you know of dystrophin's normal function described above, what do you think are characteristics of the DMD disease phenotype?
Muscle cells would be severely damaged in response to the normal pressures exerted on them. Symptoms of DMD include severe muscle damage and muscle weakness.
The dystrophin protein contains several functionally distinct regions (domains). The N-terminal domain interacts with proteins inside the muscle cell, while the C-terminal domain enables dystrophin to bind to membrane-associated proteins. The long rod-like central part of the protein is composed of 24 identical rod domains (internal repeats). As a consequence of the modular structure of dystrophin, proteins missing some of the internal repeats can be fully functional or at least partly active.



One therapeutic strategy involves excluding, or skipping, an internal exon during splicing. In particular, skipping the exon that contains the frame-shift mutation in the DMD individual can result in a shorter than wild type but functional dystrophin protein. This exon-skipping therapy involves introducing a small RNA molecule that is complementary to the exon to be skipped. Binding of the RNA to the pre-mRNA prevents proper recognition of the exon by the splicing machinery and results in specific exon skipping.

e) In order for this therapy to work, what properties must this small RNA have? (Think about cellular localization.)
The therapeutic RNA would need to localize to the site of splicing to cause efficient exon skipping. Not only does the RNA need to enter the nucleus, but it must bind to its target exon on the pre-mRNA. In fact, the small RNA can be engineered to bind to a specific protein that localizes to the spliceosome! It is also important the RNA be stable and resistant to rapid degradation in the cell.
f) How would you deliver the RNA molecule into the target cell?
You could use a viral vector to introduce a gene encoding the therapeutic RNA into the cell (allows for efficient uptake and expression).

g) How is this therapeutic approach different from gene therapy?
This approach is different from gene therapy in that it modifies the spliced mRNA rather than just counteracting the genomic defect by introducing a functional copy of the whole gene.

h) What challenges do gene therapy and exon-skipping therapy have in common?
Efficient uptake, stability and expression of the foreign molecule, side effects due to viral vector, etc.

Question 3

a) What percentage of genomic nucleotides do you expect two randomly chosen people to have in common?
99.9 % DNA sequence identity
Genome sequencing has revealed that the average genome nucleotide difference between two randomly selected chimpanzees is roughly four times greater than between two humans.

Images removed due to copyright reasons.

Based on population genetic theory, levels of genetic variation within species should correlate positively with population size. However, the human population numbers in the billions and the population size of chimpanzees is fewer than a hundred thousand.

b) How can you explain the comparatively little variation between human individuals?
The human population likely experienced a severe, but short-lived, population bottleneck, where the population was likely reduced to a few thousand, which in itself may have reduced variation. In such a reduced population size, genetic drift can act as a strong force to reduce genetic variation.

c) What is genetic drift?
Genetic drift refers to the changes in allele frequency in a population as a result of the role of chance in the production of offspring (genes in offspring are not a perfectly representative sampling of parental genes). Genetic drift has a stronger effect on genetic diversity in small populations.

A few years ago, an international consortium was formed to uncover the locations of genetic variation in the human genome. In particular, the consortium worked to identify single nucleotide polymorphisms (SNPs) within the human population.

d) Is the genomic variation between individuals randomly distributed across the genome or does such variation occur at common sites?

Variation tends to occur at common sites

e) What is a haplotype?

SNPs that are close together tend to be inherited together. A set of associated SNP alleles in a region of a chromosome is called a haplotype. Most chromosome regions have only a few common haplotypes that account for most of the variation between people. While a chromosome region may contain many SNPs, only a few SNPs need to be recorded to capture most of the information about the genetic variation in that region.

f) Why is an understanding of genomic variation useful for the study human health?

The 0.1 % that is different between individuals is important because it contains the genetic variants that may play a role how people differ in their risk of disease or response to drugs. Many common diseases, such as diabetes, are the result of many affected genes. Comparing the variation between a group of diabetic individuals and a group of non-diabetic individuals may reveal regions of the genome that are different between these groups and may help to reveal the many genetic changes underlying the disease.

Question 4

Now that the first draft of the complete* genomic sequence of the Chicken is available on the web (http://www.genome.ucsc.edu/cgi-bin/hgGateway?org=Chicken) you decide to start analyzing the sequence.

* except for the sex chromosomes! (We're waiting on Winston for this.)

Luckily, there are a lot of resources available. After all, Chicken is an important model organism for the study of viruses, cancer, and developmental biology (and it's finger lickin' good…)

a) Genbank (available from http://www.ncbi.nlm.nih.gov) lists 559394 mRNA sequences derived from Chicken. Explain how you could use these sequences to find genes in the Chicken genome.

Align mRNA sequences to the Genomic sequences, Exons (in the mRNA) will align to the genome at genes.

b) Does your method also provide information about the structure of genes (i.e. intron/exon boundaries, splice sites)? What are the possible limitations of your method?

Yes, introns will appear as gaps in the alignment where there is no sequence in the mRNA corresponding to the genomic sequence. The method is limited to those mRNAs found in the database, many alternatively spliced exons may be missed.

c) Now that you know the locations of all of the autosomal genes in the Chicken genome, you would like to start trying to predict regulatory sequences. Unfortunately, Chicken is the only bird that has been sequenced. Why might this be a problem? How have regulatory sequences been found in other organisms, such as yeast and mammals?

The alignment of genomes of multiple organisms that are closely related allows the prediction of regulatory elements. Non-genic sequences 5' to genes that align across several species are likely regulatory sequences. Because no other bird sequences are available, only the most conserved elements will show up in alignments with other species.

d) You talk to Eric about your problem, and he offers to sequence 3 bird species to help you out. He's been thinking about sequencing some birds anyway, so he offers to let you help him decide which ones to pick. He is considering sequencing the ostrich, finch, quail, turkey, condor, pheasant, and guinea fowl. Which ones do you pick and why? (It is okay to use physical characteristics as an indicator of relatedness.)

Turkey, Pheasant, Quail, and Guinea Fowl are all very closely related to chicken. You should pick the Ostrich, finch, and condor, since they are the birds on the list most distantly related to chicken. This will maximize the signal to noise ratio while trying to detect possible regulatory sequences.

Images removed due to copyright reasons.

Question 5

There are several subspecies of the mouse, *Mus musculus*, living throughout the world. You decide to study their spread from an ancestral population using sequences from mitochondria and Y chromosomes.

a) Why are mitochondria and Y chromosomes often used for this purpose?

Mitochondria and Y chromosomes are inherited clonally, without recombination, from one generation to the next. Mitochondria are inherited through the maternal lineage, Y chromosomes through the paternal lineage. Recombination scrambles up polymorphisms between chromosomes. Mitochondria evolve quickly enough to date divergences within a species.

You isolate mitochondrial DNA from 10 mice from geographically diverse locations, and then sequence an essential mitochondrial gene. Dashes indicate bases that are identical to the top sequence.

```
England          AAAGCAGAAGAAATAGATGCAAAGGCAGAAGAAGAGTTCAA
Japan            --GCAT-----------C---------------AATT
South Africa     ------C----------------------TA----------
India            GG-CG--C----------GGG-----GG-------CGGG
China            TTT---T----------TT--------T-----GGTT
United States    ------C----------T-------------------
Philippines      TTT---T----------TC--------T-----GGTT
New Zealand      ------C----------C-----TA----------
Russia           --GCAT-----------CC--------------A-TT
Yemen            -------------------------C-------
```

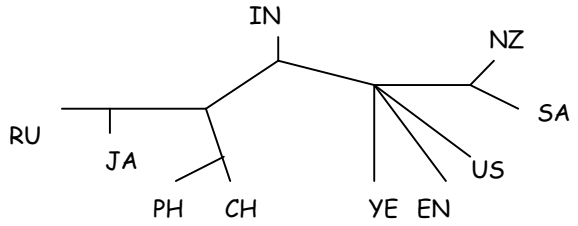
b) Circle the region above that is most likely to be important for the function of the mitochondrial gene.

Should circle the region with the most dashes

c) Construct a tree based on these mitochondrial DNA sequences. How many major sub-species of mice do you suppose that there are?

((YE,(EN,(US,(SA, NZ)))),IN,((RU,JA),(CH,PH)))

Any number between 3 and 10 is acceptable, but we're hoping for three.



Next you decide to look at markers on the Y chromosome in mice. You study a highly polymorphic locus called Mick-Y (for Many Interesting Changes Known on the Y). You find that populations of mice in different locales have different alleles.

Population	Alleles of Mick-Y
England	A,B,C
Japan	L,M
South Africa	B
India	G,H,I
China	J,K,L
United States	C
Philippines	K,L
New Zealand	B
Russia	L,M,N,O
Yemen	A,B,C,D,E,F

d) Based on this information (and a little bit of knowledge about history) what do you think is the origin of the mouse populations in the United States, New Zealand, and South Africa?

Mice probably came along with European colonists, each descended from a small founder population of English mice.