

5/30/2005

Computation and statistics (2 hours)

Closed book.

Short questions

1. (a) What does it mean to 'normalize' a database?
Store data in a systematic way that minimizes redundancy. There are a number of "normal forms" that you can normalize to defining different degrees of constraint in reducing redundancy.
 - (b) What is a 'foreign key'?
A field/attribute in a table which references a primary key in another table.
 - (c) Give two reasons that Java or C++ is superior to Perl for large projects.
 - i) Java and C++ are strongly typed for easier debugging, whereas Perl is not
 - ii) Java has better builtin documentation (Javadoc)
 - iii) Java and C++ allow better OOP, (syntax in Perl is ugly)
 - (d) define $O(f(n))$ as used in problem 2.
 $O(f(n))$ is the Big-O computational complexity for a function where $f(n)$ is an asymptotic upper bound. Specifically:
$$O(g(n)) = \{f(n) : \exists c, n_0 > 0 \text{ s.t. } 0 \leq f(n) \leq cg(n) \forall n \geq n_0\}$$
2. Suppose $0 < f(n) < g(n)$ for all $n > 1000$. Is each of the following true or false? If true, give a very brief justification; if false, give a very brief counterexample.
 - (a) $f(n)$ is $O(g(n))$
True, b/c $g(n) > f(n)$ for $n > 1000$
 - (b) $f(n) + g(n)$ is $O(g(n))$
True, b/c we can choose c s.t. $cg(n) \geq f(n) + g(n)$ for $n > 1000$
 - (c) $g(n)$ is **not** $O(f(n))$
False, b/c if $f(n) = n$, $g(n) = 2n$, $O(f(n)) = n = O(g(n))$
 3. Outline in general terms what null hypothesis is being tested in a one-way ANOVA, and describe how the concept of an "analysis of variance" is used to actually carry out the test.
In one way ANOVA, the null hypothesis is that the mean of all groups being tested is identical. In essence, the procedure is carried out by separating the total variance into its components of BGSS (Between Groups Sum of Squares) and WGSS (Within Group Sum of Squares). In essence, BGSS is a signal that is significant only with respect to the noise of WGSS, so taking their ratio (ie BGSS/WGSS) amounts to a test of significance with the appropriate degrees of freedom – the so called F statistic.

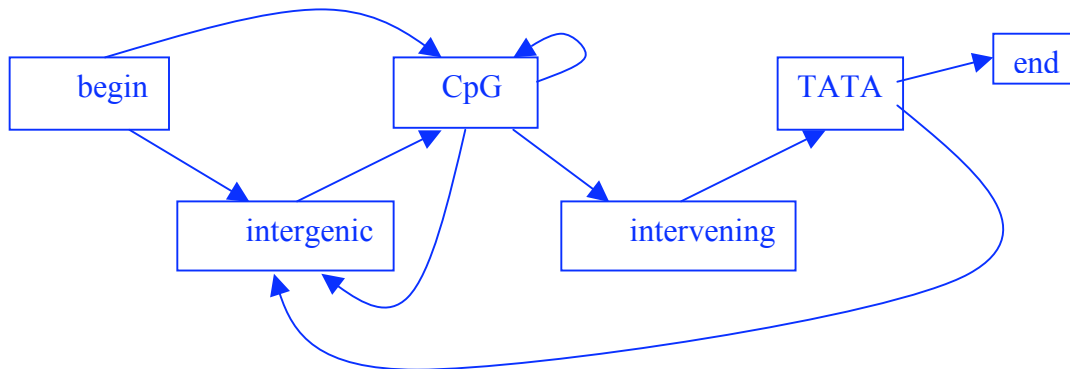
4. Suppose I claim that a given phylogenetic tree is the maximum likelihood estimate derived from some aligned set of data D. State what this means in terms of conditional probabilities.

The likelihood principle states that given D, data, and M, a model, $L(M|D) = f(D|M)$. Wrt phylogeny D is an alignment of root taxa and M is a tree topology, and calculating $f(D|M)$ is straightforward under certain assumptions. A maximum likelihood estimate of a tree is the most likely tree given the data. This translates into the tree for which the data is most probable, ie $f(D|M)$ is maximal. This is implemented by iterating a number of tree topologies, and returning the one that gives the maximum of $f(D|M)$.

Longer questions

5. CpG islands are genomic regions rich in both G+C content as well as CG dinucleotide. TATA-box is a specific DNA sequence motif often found near transcription start sites of genes. We can represent a TATA-box by a position specific probability matrix with 8 columns and 4 rows, where the columns specify positions and rows specify probability of the i th nucleotide. For example, p_{ij} element of this matrix denotes the probability of the i th nucleotide in the j th position of the TATA-box. Suppose we define a promoter as a CpG island followed by a TATA-box with some intervening sequences. Using the following assumptions, design a Hidden Markov Model to recognize promoters in genomic DNA. Clearly state any additional assumptions and define all required parameters.

1. CpG island length is geometrically distributed
2. The four different nucleotides have same marginal probability in genomic DNA
3. The spacing between CpG island and the TATA-box is provided as a probability look-up table.



Shown above is a general state diagram for the HMM. Arrows represent transition probabilities between states. So the arrow from the CpG state back to itself models the geometrically distributed length of the CpG islands. The length of the CpG island can be tweaked by altering this self-transition probability. The intergenic and CpG states have emission probabilities corresponding to nucleotide frequencies associated with the given state. For example, given we are in a CpG state, the probability of emitting a C or a G

will be higher than emitting an A or a T. The “intervening” state models the spacing between the CpG island and the TATA box. The TATA box is modeled by a PWM as described in the question. Our initial constraint is that we start in the begin state.

We could improve sensitivity by modeling the CpG state more finely. Instead of a single state, we could model each base in the CpG island separately (see Durbin p49 for a figure). All bases would be able to transition to all other bases, but certain transitions will be more likely given the sequence structure of CpG islands. For example $P(G|C)$ and $P(C|G)$ would be weighted more highly than other transitions. In so doing a sequence like “GCGCGCGC” would be more probable than a sequence like “GGGGCCCC”. Similarly, the TATA box itself might be modeled more concisely if you allowed for 8 separate states (with the emission and transition probabilities at each state defined from a set of training examples).

It is a possibility (albeit remote) that you might find a "promoter-like" sequence (i.e. a sequence generated from your HMM with high probability) that doesn't actually precede a gene-like sequence. So one might consider placing another state after the TATA states that represents something like "between TATA and 5' UTR" or something. This is probably too fine tuned and makes the model too complicated. So I wouldn't necessarily include it in the diagram, but might mention it in my explanation.

We can start modeling by assuming a first order HMM, where the probability of a state, given all prior states is just the probability of a state given the previous state. Additionally, this model assumes there are no exons in the supplied sequence, as they are not modeled explicitly. Exons in real data would probably be absorbed into the intergenic state. Additionally, we assume that CpG islands can be punctuated by intergenic sequence as shown with the transition between their states. However, once we find a TATA box, we can claim we found a promoter because the only path leading to TATA is defined by a CpG island followed by intervening sequence. Additionally, the model could identify numerous promoters because of transitions from TATA back to the intergenic state are present.

For each state, the emission and transition probabilities are learned from a set of training examples. Once trained on labeled data, we could identify promoters by decoding a novel sequence to find the most likely path given the data. Once we have the path, promoters will be around the labeled TATA box.

6. In a study of lung cancer, the investigators found:

Among 100 cigarette smokers, exactly 15% had lung cancer.

Among 100 non-smokers, exactly 5% had lung cancer.

Answer the following questions:

a. What hypothesis would you test with these data?

Smoking has no effect on cancer incidence

b. Describe how you would use the chi-square test to test the hypothesis, showing the necessary numbers and calculations. (If you need any specific numbers, e.g. for a distribution, just state what you need.)

First set up an observed and expected (if H_0 was true) matrix:

Observed(expected)	Smoker	Non-Smoker	Sum
Cancer	15 (20*100/200=10)	5 (20*100/200=10)	20
No Cancer	85 (180*100/200=90)	95 (180*100/200=90)	180
Sum	100	100	200

Then calculate a χ^2 statistic as $\sum(O-E)^2/E$ over the 4 classes = 5.56.

c. Reach a conclusion about your hypothesis for the data provided, and justify the conclusion.

By looking up a Chi-square distribution with 1 degree of freedom we see a χ^2 of 3.8 represents a P-value of 0.05 whereas a χ^2 of 6.635 represents a P-value of 0.10. If we choose significance, $\alpha = .05$ we would reject H_0 . On the other hand, if choose significance of $\alpha = .10$, we cannot reject H_0 .

7. Consider the various gap penalty schemes for pair-wise alignments. (i) Choose three different classes of penalty functions and sketch out a biological rationale for each. (ii) In coding regions, one may align the nucleotides or the translated amino-acids. Discuss the pros and cons of each and especially in terms of how it might affect the gaps. (iii) In the context of different possible gap penalty functions discuss how one might construct an algorithm to align genomic sequences that contain a mixture of intergenic regions, exonic regions, and intronic regions.

i) A linear gap penalty is proportional to the length of the gap. An affine gap penalty has a constant gap opening penalty, and then a linear gap extension penalty. A constant gap penalty is independent of gap length. The choice of gap penalty depends on what you want to align. For example, orthologous coding regions are more identical than non-coding regions. As such, we expect aligned coding sequences to be less “gappy” (in both number of gaps and length of gaps) in coding regions. Therefore, a linear or affine penalty would be appropriate for aligning coding regions rather than a constant one. Because of the propensity for duplication events to take place, it makes biological sense to penalize the presence of a gap more so than the length of a gap. As such, an affine penalty with a high opening cost, and more gradual gap extension penalty is probably ideal for aligning coding regions. However, for aligning non-coding sequence, we expect a much more “gappy” alignment. In a biological sense, non-coding regions are under reduced selective pressure, as protein structure does not depend on them. They are more prone to tolerate base pair mutations and insertions. As such, a constant gap penalty would probably be better for such alignments.

ii) Because of the redundancy of the genetic code, aligning amino acids is probably preferable to aligning nucleotides in coding regions. Because different codons code for the same amino acid, the same nucleic acid composed of different, but synonymous

codons, would score less than the corresponding amino acid (its alignment would be perfect). As such, aligning nucleotides would probably introduce more gaps than aligning amino acids. Furthermore, aligning nucleotides allows you to weight amino acid similarity. The substitution of similarly functional amino acids (eg a hydrophilic for another hydrophilic) should be less penalized than the substitution of functionally different amino acids (eg a hydrophilic for a hydrophobic) because the latter is less likely to be biologically functional. This type of weighting of functional similarity does not exist for nucleotide alignments. Finally you can always reverse-translate an amino acid alignment back to a nucleotide alignment.

iii) A pair HMM could be implemented that, explicitly models intergenic, exonic and intronic regions, while emitting aligned pairs of nucleotides. In so doing, given the most probable state, you could choose an appropriate gap penalty to use.

8. You are given a binary tree, with String data in each node of the binary tree, and you need to make an exact copy of the binary tree with all its data. (I.e., you want to produce a new binary tree with exactly the same branch structure and contents.) Describe the simplest way to do this.

By doing a post order traversal of the tree (DFS in a sense), you can reconstruct the tree from the leaves up. Using a stack to implement, recursively push nodes starting from the root node on to the stack. If the node is not a leaf, push its children onto the stack. If a node is a leaf, mark it as such before pushing onto the stack. Once all nodes have been stacked, sequentially pop nodes, recreating internal vertices. Every time a non-leaf is popped (ie an internal vertex), its left and right sub trees should already be known.

Alternatively, and probably simpler (courtesy Praveen), we could simply traverse the tree according to some recursive traversal procedure and directly build the new tree. Below is some code that uses post-order traversal to copy a tree:

```
public BinaryTree copy(BinaryNode bt) {  
    if(bt == null) return null;  
    BinaryTree left = copy(bt.left);  
    BinaryTree right = copy(bt.right);  
    return new BinaryTree(bt.value, left, right); }  
}
```

Of course, this code assumes that your current tree to-be-copied is implemented using a class called BinaryTree and that the function copy takes in as a parameter the root node of that tree.

9. (a) What is a stack, and what operations can one perform on it?

A data structure that satisfies the LIFO (Last In First Out) property. You can push data onto the stack and pop data from it.

(b) Give an algorithm to reverse the contents of a stack using only the stack operations. You are allowed to use additional stacks. You may not assume anything about the underlying implementation, except that it provides all the stack operations.

```
Function reverse (stack A)
  Define stack B
  While stack A, is not empty:
    Pop the element out of A
    Push the element onto B
  Let A = B
  return A
```

BONUS QUESTION – ONLY IF YOU HAVE SPARE TIME

10. Define ‘genetic code’ as a grouping of codons such that all codons in a group codes for the same amino acid (without necessarily knowing the identity of amino acids). In an unlikely progression of science, assume we do not know the genetic code but we

1. know of mRNA
2. we can purify and sequence mRNA
3. we have large set of multiply aligned orthologous mRNA

Design a computational approach to predict the genetic code.

Biology – 2 hours

Closed book.

Short questions

1. What are the major protein families involved in the segregation of chromosomes during mitosis. Name at least three major types of proteins and describe how they function.

Separase is responsible for cleaving the **cohesion** complex allowing sister chromatids to separate. The **mitotic spindle** is composed of **microtubules** that are anchored to the **centrioles** and previously attached to the **kinetochore** of chromatids at prophase. Driven by their motor proteins, **Kinesin** and **dyenein**, kinetochore microtubules shorten thereby pulling the chromatids to their respective poles. Furthermore, motor proteins cause astral microtubules to lengthen by pushing overlapping microtubules apart, thereby separating the poles further. The **contractile ring** is an important cytoskeletal structure comprised of **actin filaments** and their **myosin** motor proteins. It forms just beneath the plasma membrane in a plane perpendicular to the axis of the **spindle** apparatus. As the ring contracts, it pulls the dividing cell membrane thereby separating the cell into 2 daughter cells.

2. In eukaryotes, genes are transcribed in the nucleus, mRNAs are translated in the cytoplasm, and the protein products of these mRNAs can end up in many different cellular compartments (nucleus, cytoplasm, plasma membrane, organelles etc). What determines cellular localization of a protein? Provide three examples.

Localization signals determine the cellular localization of a number of proteins. For example, translocation of a secretory protein into the ER is determined by a particular AA sequence. During translation, the signal recognition peptide (SRP) binds the signal sequence if present pausing translation. This complex then binds the SRP receptor in the ER membrane, causing additionally translated polypeptide to be threaded through the aqueous translocation channel (Sec61) into the ER lumen.

Additional localization signals are used to prevent resident ER proteins (such as glycosyl transferases for carbohydrate modifications) from going to the golgi to be secreted. ER resident proteins contain a protein localization signal (KDEL) that is recognized by a receptor in membrane derived from ER. Thus, any ER resident proteins that make their way to the VTC are bound by the KDEL receptors; the surrounding membrane buds into COPI coated vesicles and are returned to the ER via the retrieval pathway.

Besides ER localization signals, there are other amino acid motifs that direct proteins. These signals are recognized by a specific receptors (traditionally clathrin adaptors associated with the clathrin coat of vesicles). These are responsible for processes such as internalization from the plasma membrane and endocytosis.

In vesicular trafficking, specificity is determined by the coat proteins that line the vesicle membrane. Surface markers on the vesicles determine where they may dock. For example, SNARES are transmembrane proteins that exist in as a complementary set of v-

SNAREs (on the vesicle membrane) and t-SNAREs on the target. Complementary SNAREs bind to form a trans-SNARE complex, causing the two membrane surfaces to dock.

In addition, there are non-sequence based manners in which proteins are localized or directed. The ubiquitination system marks proteins for degradation using a covalently bonded ubiquitin tag that directs marked proteins to the proteasome pathway for degradation.

3. Briefly describe various stages and factors affecting the regulation of active proteins in a cell, starting with chromatinized DNA. Also indicate the technologies appropriate to detect the regulation at these various stages.

The nucleosome is the fundamental subunit of chromatinized DNA, and contains ≈ 200 bp of DNA wrapped around a core histone octamer. The characteristics of each histone are modulated via their N-terminal tails by covalent modifications such as acetylation, methylation, and phosphorylation. Patterns of these modifications compose a “histone code”, with the results of loosening the histone-DNA bonds, unravelling DNA, translocating a histone along the DNA, completely preventing other proteins from binding to DNA. Enzymes known as histone acetyl transferases (HATs) de-acetylases (HDACs) and histone methylases and de-methylases catalyze these modifications. Such enzymes are recruited to specific regions of a chromosome by histone remodeling complexes, led by two proteins, Swi and Snf; there are a variety of such proteins. These proteins feature protein binding domains to which adaptor proteins bind, to which enzymes like HATs and methylases bind. In addition there exist variants of histone subunits, which are exchanged to alter the properties of a nucleosome, such as centromere structure, DNA repair, and naturally transcriptional regulation. The result of all of this is the orderly opening and closing of particular areas of DNA, commonly where promoters are found, so that particular genes may be transcribed when needed.

Technologies critical to the detection of such regulation include chromatin immunoprecipitation (ChIP) which essentially takes a static snapshot of the molecular environment of DNA, “pulls down” any proteins bound to it. Additionally, assays such as indirect end labeling are used to study gene activation and chromatin accessibility in which the DNA footprint is compared between an active and inactive gene, thereby revealing potential protein (histone) binding sites on the naked DNA. Also, histone modification assays, are *in-vitro* systems that look at what enzymes modify histones. A gel, enriched with proteins of interest (ie histones), is first crosslinked to immobilize proteins. The gel is run along with appropriate modification enzymes (HAT, histone methyltransferases, etc) and labeled transfer substrate (ie labeled acetyl-coa, SAM, etc). Proteins will light up and can be compared to a control with no substrate proteins.

4. Consider a diploid organism with two pairs of chromosomes.

a. In meiosis, cell division occurs in two successive steps, meiosis I and meiosis II. Describe (words, pictures, or both) what takes place with regard to the chromosomes at each of these steps. Include crossing over in the answer.

b. How does the process of mitosis differ from that of meiosis?

Mitosis involves only one cell division (similar to meiosis II) whereas meiosis involves 2 cell divisions. Sister chromatids do not pair up, but line up at the metaphase plate of meiosis I. At the end of meiosis I, daughter cells have N complement, and another subdivision of the 2 daughter cells at the end of meiosis II gives us 4 haploid daughter cells. So mitosis produces 2 diploid daughter cells, and meiosis produces 4 haploid daughters.

Longer questions

5. In a forward genetic screen, you have identified a new single-gene mutation in the mouse that increases locomotor activity by three fold.

a) How would you identify the gene causing the mutant phenotype?

Using linkage analysis. First we will look for linkage with other loci on the genetic map looking for chromosomal traits that co-segregate with the mutant phenotype. Once we determine the chromosome, we will use a number of molecular markers to perform "chromosome walking" toward the mutant phenotype. Methods will involve rapid PCR-based methods to identify closely linked STS, and automated sequencing to establish tightly linked SNPs and other polymorphic markers (microsatellite, VNTRs, etc). Once we have narrowed down the target space to around 100k or so, we can look on public databases for candidate genes. If annotation is lacking, we may have to look at sequenced ESTs for potential novel open reading frames.

b) How would you prove that the identified DNA alteration is responsible for the observed mutant phenotype?

Assuming an autosomal recessive phenotype, use a balancer chromosome to perform mutational analysis on different stock lines. The balancer suppresses recombination so we can maintain lines already with the mutation. Once we have mutant stock established, we would want to do directed mutagenesis with cre/lox system to try to reproduce the mutant phenotype.

c) Suppose your efforts (points a and b) result in the identification of a mutant gene encoding novel transcription factor. How would you, on a large scale, identify genes that are regulated (downregulated or upregulated) by this transcription factor?

ChIP on chip microarray analysis. First construct array probing with a sampling of genome-wide regulatory regions (promoters and 3' UTRs). Perform ChIP using anti-TF antibody, which will pull down TF binding sites. Uncross link DNA from antibody, and use it to target array (along with genomic reference). Regulatory regions of candidate genes will show up.

6. A classmate of yours has studied two SNPs that are thought to be related in function, or location, or both. The study was done in a group of volunteers in San Francisco, where several ethnic groups are represented in the population. At SNP "A" there are two alleles, A1 and A2; at SNP "B" the two alleles are B1 and B2. It turns out that both loci map to the X chromosome, but it is not known whether the locations are close together.

Among 500 men genotyped by your classmate, she found the following numbers of individuals with each allelic combination: 45 A1B1, 55 A1B2, 255 A2B1, and 145 A2B2.

i. Are the alleles at the A and B loci associated? Explain.

Yes they are associated otherwise you would expect equal proportions of genotypes. To quantify how associated they are use a Chi-square test to come up with a P-value that they are not associated using equal expectations for each genotype with $(2-1)(2-1)=1$ degree of freedom.

ii. Can you tell whether the A and B loci are closely linked? If so, how sure are you? If linked, how closely? Explain.

You need to know the parental genotypes (and one parent has to be homozygous and the other heterozygous at each locus) in order to distinguish recombinant (non-parental) from non-recombinant (parental) before you could start quantifying linkage. However, because Y chromosomes do not recombine, we may be able to make inferences in related men. In this case we have sex linked loci in men, but we have no reason to believe subjects are related, so no you can't tell from the given sequencing data alone.

iii. If the A and B loci were in linkage disequilibrium, how would you quantify the extent of LD from the data provided?

To quantify LD, we need to calculate the correlation between alleles and show that they are more associated than they would be by chance alone. We would need to show that the haplotype frequencies are not equal to those predicted by allele frequencies.

Allele frequencies:

$$A1 = (45 + 55)/500 = 0.2$$

$$A2 = (255 + 145)/500 = 0.8$$

$$B1 = (45 + 255)/500 = 0.6$$

$$B2 = (55 + 145)/500 = 0.4$$

Expected genotype by allele frequencies (expected if no LD):

$$A1B1 = 0.2 * 0.6 = .0012$$

$$A1B2 = 0.2*0.4 = .0008$$

$$A2B1 = 0.8*0.6 = .0048$$

$$A2B2 = 0.8*0.4 = .0032$$

Observed genotype frequencies:

$$A1B1 = 45/500 = 0.09$$

$$A1B2 = 55/500 = 0.11$$

$$A2B1 = 255/500 = 0.51$$

$$A2B2 = 145/500 = 0.29$$

Compare observed vs expected using chi squared statistic with $(2-1)(2-1)=1$ degree of freedom for significance.

iv. If the findings were obtained in a random-mating population, would the relationship between alleles at the A and B loci change in succeeding generations? If yes, what can you predict about the change (direction, rate, etc.)?

LD can be strong, but theoretically speaking, over a significant time period, the LD would be less and less due to recombination (albeit infrequent) in the region. As such over succeeding generations, the observed genotype frequencies will be more similar to the predicted ones,

7. Your dream has come true and the Giraffe Genome Project has finally been funded. Your job is to lead a research team whose task is to complete the genome sequence of *Giraffa camelopardalis* in 4 years. What are the most difficult challenges you expect to encounter? List separately with *brief* explanation (1 sentence for each).

Explain how you will address/overcome the challenges, and complete the genome sequence. In your explanation, use at least 15 of the terms listed (but not in any particular order). Where you consider necessary, define the term in a brief phrase or sentence.

Whole genome shotgun sequencing strategy

An approach to genome sequencing where the whole genome is shared into sequencable fragments, and computationally assembled. All sequencing is done ahead of time using PCR products, to form shotgun libraries of sequence reads.

Clone-by-clone sequencing strategy

An alternative to WGS where a divide and conquer approach is utilized. First, create genomic libraries of clones immortalized in vectors such as BACs. Ideally you want 5-10x redundancy of genomic coverage in your libraries. Then form a tiling path by end sequencing clones and aligning overlapping fragments. In so doing, you will be able to quantify gaps where clones lack coverage. You will sequence individual clones along the tiling path and assemble contigs spanning the genome. Finally work on finishing sequence and plugging gaps.

Hybrid sequencing strategies

A combination of clone by clone and WGS which was used for the mouse and chicken genome projects. Such a compartmentalized shotgun, could for example break the genome up into chromosomes, and then do shotgun sequencing on each chromosome. Probably the best of both worlds, as many genome projects are now adopting a combined approach.

Draft Sequence

Sequence with an error rate of $10^{-3} \rightarrow q=30$

Finished Sequence

Sequence with an error rate of $10^{-4} \rightarrow q=40$

Interspersed Repeat Sequences

Satellite Repeat sequences

Segmental Duplications

Q-value

$-10\log(p)$ where p = the error rate (or probability of an error)

Mate-pair sequences

A pair of sequences derived from the two ends of a single clone. An essential component of shot gun sequencing as the distance between the pairs gives spatial information and assists in resolving repeats.

BAC end sequences

Used to establish mate pairs and construct the tiling path in clone by clone sequencing.

mRNA sequences

Messenger RNA –Eukaryotic transcribed sequences that have been processed (ie spliced and exported out of the nucleus)

EST sequences

Expressed Sequence Tags – a sequenced piece of cDNA, however may not span the whole cDNA transcript. cDNA library generation uses primers to the poly a tail of the mRNA transcript, and a single sequencing trace is usually performed toward the 5' portion of the gene (all this is done on the complement strand).

STS

Sequence Tagged Site – any sequenced fragment of DNA derived from a library of clones that is placed on the physical map of the genome. Each STS is unique and primers, PCR conditions, and product size are immediately quantifiable and storable in a database. Fundamental to the HGP.

Microsatellites

Stretch of repetitive DNA made up of a variable number of several to one hundred or more tandem repeats of a small number of nucleotides. Ex $(AG)_n$ or $(CAG)_n$. Highly polymorphic (in n at least) and heterozygous, and occur around several per hundred kilobases in higher eukaryotes.

SNPs

Single Nucleotide Polymorphisms. Useful for mapping phenotype to gene. Highest resolution of polymorphic markers 1/kb

Meiotic Linkage Maps

Linkage maps based on natural meiotic breaks from homologous recombination.

Radiation Hybrid Maps

Linkage maps based on induced chromosomal breaks from X-ray irradiation.

Fragmented chromosomes are then exposed to hamster cell lines and fragments become either incorporated into the hamster chromosomes (via homologous recombination), or segregate as mini chromosomes.

Cytogenetics

Cytogenetics is the study of chromosomes and the related disease states caused by numerical and structural chromosome abnormalities. FISH is especially used in cytogenetics

FISH

Flourescence Insitu Hybridization. Hybridize fluorescent DNA probe on mitotic chromosome at metaphase. Used in “chromosome painting” where one species chromosomes are labeled and synteny with another species is sought.

BACs

Bacterial Artificial Chromosomes. A system to clone approx 100kb of DNA into bacteria.

Clone-based Physical Maps

Assembled genomic sequence base on hierarchical sequencing of clone libraries

Contig alignment to chromosomes

Euchromatin

Open active DNA with genes being actively transcribed. Classically associated with acetylation of histones and HATs

Heterochromatin

Closed inactive DNA, tightly coiled and not actively transcribes. Classically associated with methylation of and methyl transferases

Inactive

Centromeres

Structures of eukaryotic chromosomes that serve as the attachment for the spindle apparatus during mitosis. Highly repetitive, and separates long arm for short arm in human chromosomes.

Telomeres

Sequences toward the end of chromosomes that contain mainly simple repeats and duplicates. They prevent chromosomes from fusing with each other by forming tertiary structures that protect termini. They are interestingly not replicated by polIII, but rather their own telomerase.

Assessment of assembly quality

Q score

Assessment of coverage of genome in assembly

Redundancy