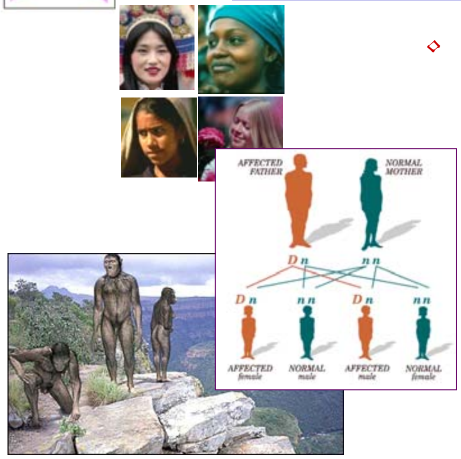
Computational Biology

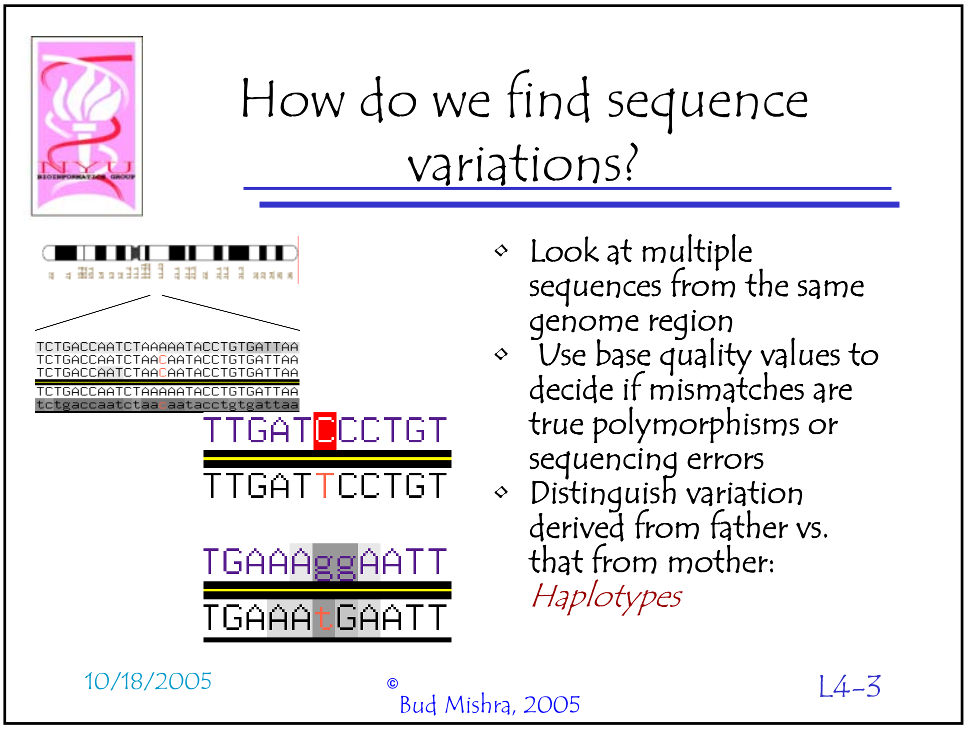
Lecture #5: Haplotypes

Bud Mishra
Professor of Computer Science, Mathematics, & Cell Biology
Oct 17 2005

Polymorphisms in Population



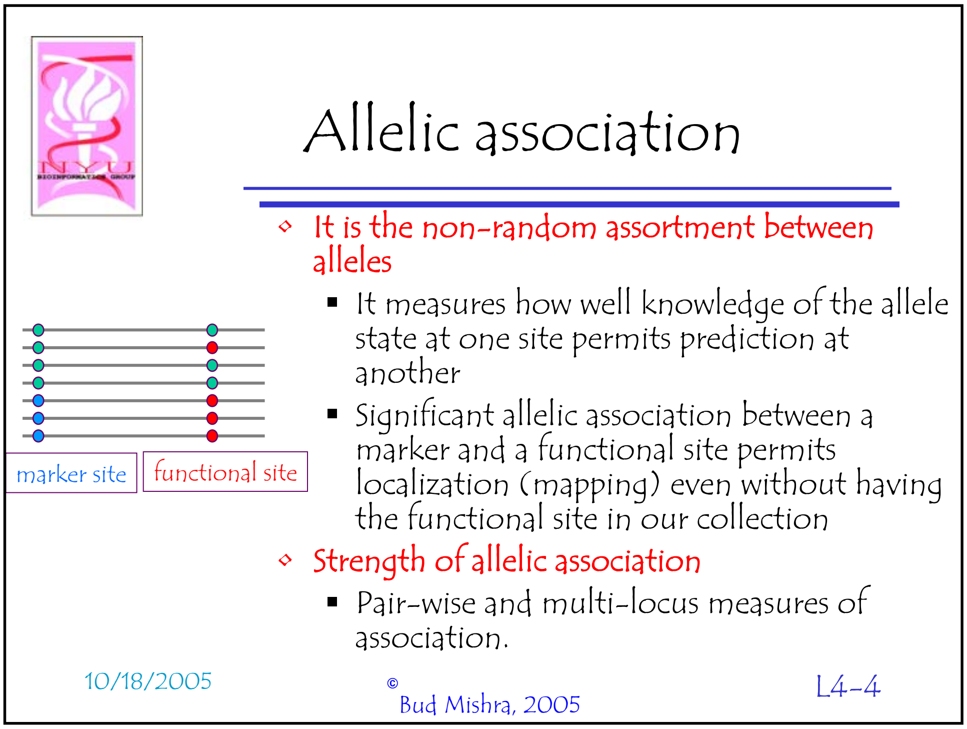
- Why do we care about variations?
  - Underlie phenotypic differences
  - Cause inherited diseases
  - Allow tracking ancestral human history


TCTGACCAATCTAAAAATACCTGTGATTAA
TCTGACCAATCTAACAATACCTGTGATTAA
TCTGACCAATCTAACAATACCTGTGATTAA
TCTGACCAATCTAAAAATACCTGTGATTAA
tctgaccaatctaacaatacctgtgattaa
TTGATCCCTGT
TTGATTCCTGT
TGAAAggAATT
TGAAAtGAATT

How do we find sequence variations?

- Look at multiple sequences from the same genome region
- Use base quality values to decide if mismatches are true polymorphisms or sequencing errors
- Distinguish variation derived from father vs. that from mother: *Haplotypes*

10/18/2005 © Bud Mishra, 2005 L4-3



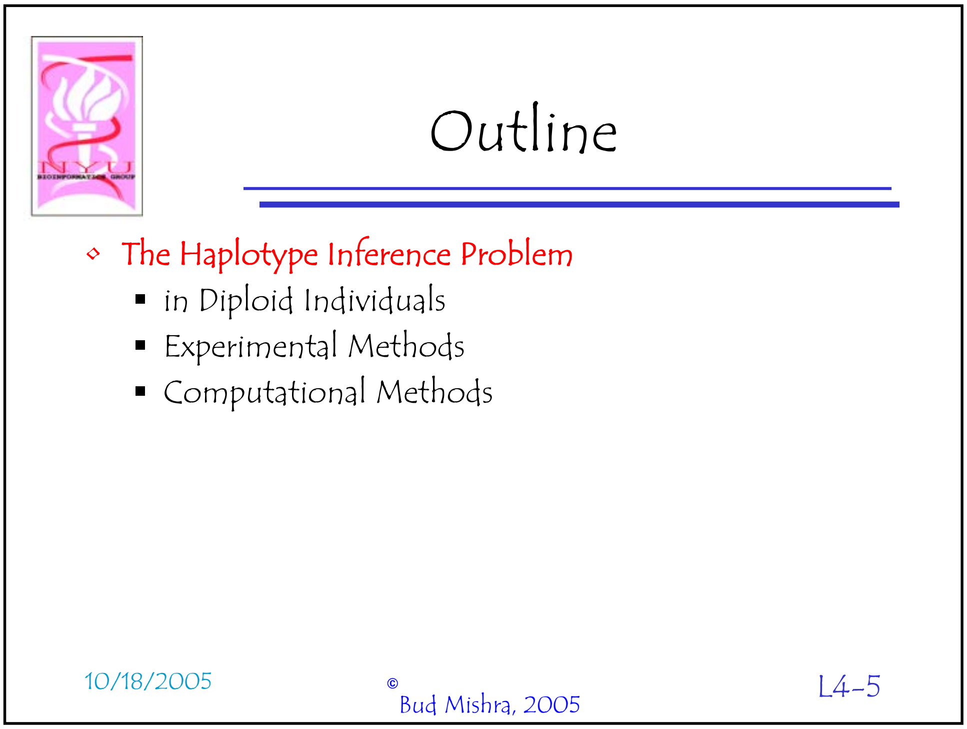
Allelic association

- It is the non-random assortment between alleles
  - It measures how well knowledge of the allele state at one site permits prediction at another
  - Significant allelic association between a marker and a functional site permits localization (mapping) even without having the functional site in our collection
- Strength of allelic association
  - Pair-wise and multi-locus measures of association.

marker site

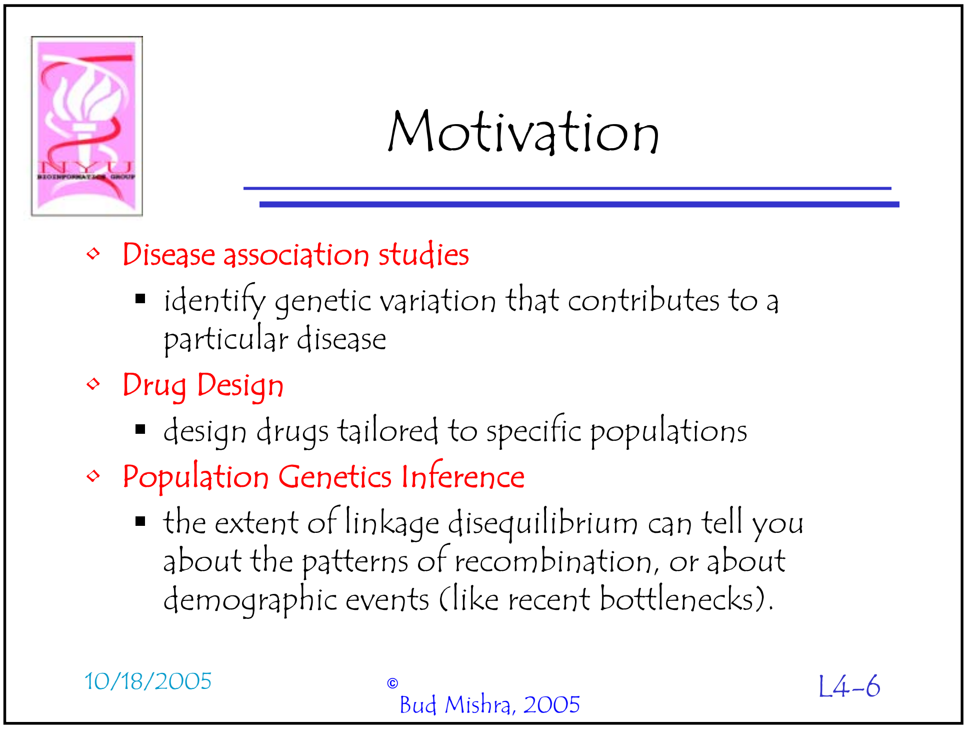
functional site

10/18/2005 © Bud Mishra, 2005 L4-4



Outline

- The Haplotype Inference Problem
  - in Diploid Individuals
  - Experimental Methods
  - Computational Methods



Motivation

- Disease association studies
  - identify genetic variation that contributes to a particular disease
- Drug Design
  - design drugs tailored to specific populations
- Population Genetics Inference
  - the extent of linkage disequilibrium can tell you about the patterns of recombination, or about demographic events (like recent bottlenecks).

Inferring Population Genetics

- The limited diversity in the European population as compared to the African population
  - It may be indicative of the founder effect.
  - It supports the out-of-Africa theory.
- IBM-National Geographic project:
  - ❖GENOGRAPHIC
  - https://www9.nationalgeographic.com/genographic/index.html

Genographic Project

The human journey
National Geographic teams with IBM in major study of human origins.
THE GENOGRAPHIC PROJECT

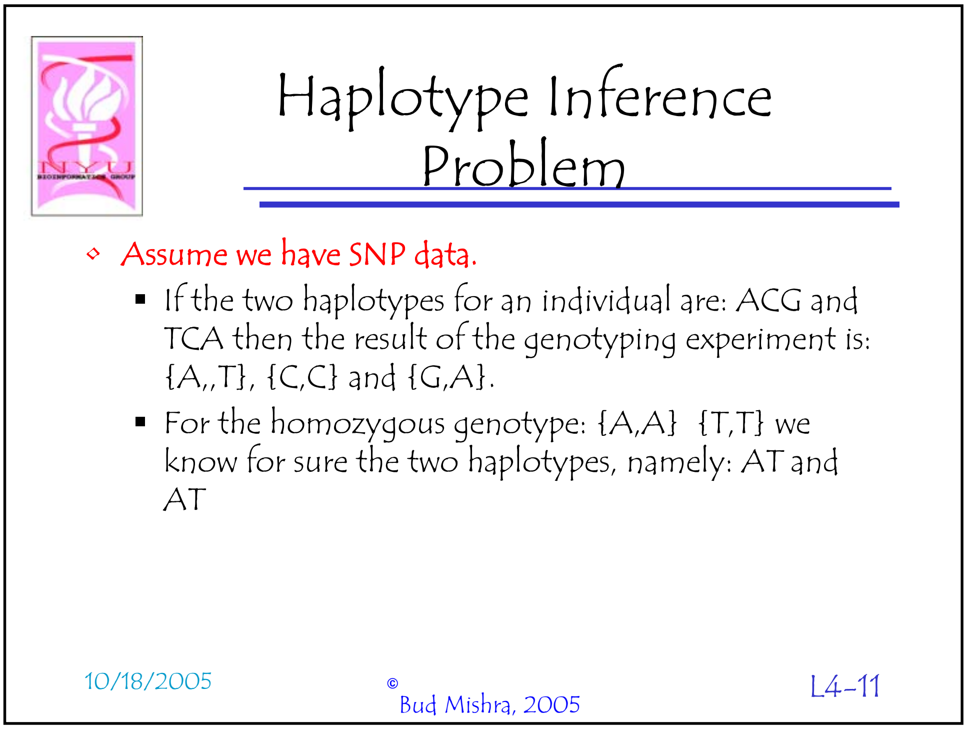
- What is expected:
- Public database of anthropological genetic information
- Virtual museum of human history
  - Online at nationalgeographic.com/genographic,
  - Information about genetics, migration, linguistics, indigenous populations and the threats facing them, anthropology, archaeology, and more.
  - Public participation
- New information on genetic anthropology
- Improved global awareness of indigenous populations

The Haplotype Inference Problem in Diploid Individuals

- Diploid individuals have two copies of their genetic material. (Two homologous chromosomes). The genetic material on a single chromosome is called HAPLOTYPES.
  - Current high-throughput genotyping methods can only determine which two alleles are present at a locus, but lose information as to which of the two chromosomes each allele belongs to (ambiguous phase)
  - This causes problems if the individual is heterozygous for more than one locus.

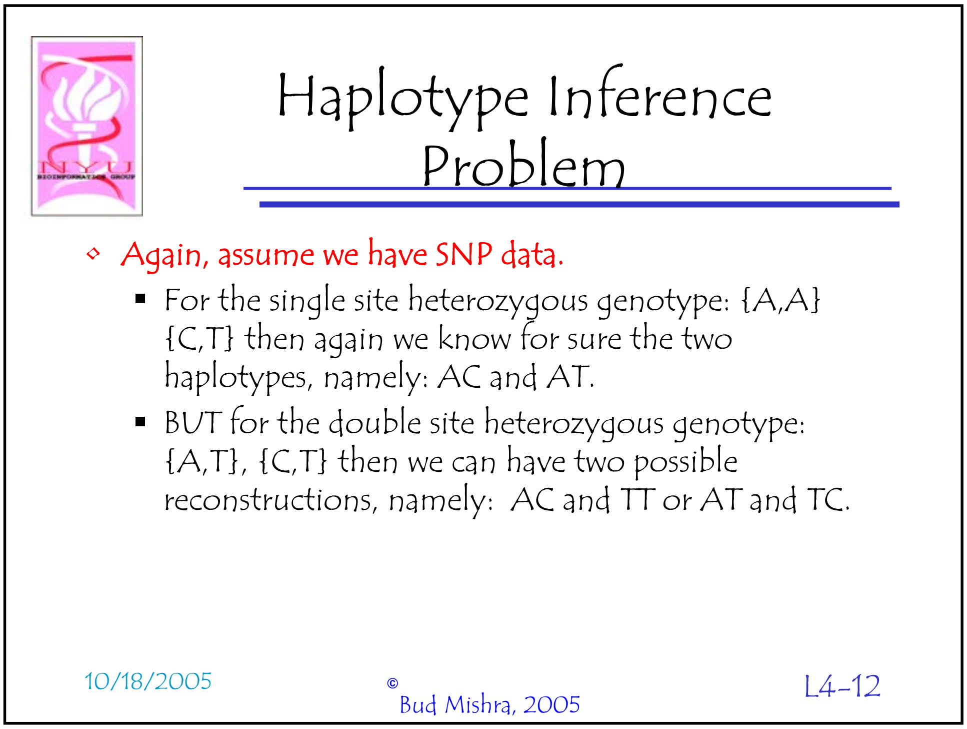
Example: SNP

- SNP: Single Nucleotide Polymorphisms:
  - One genomic location varies in its single base pair composition across a population.
  - One in about seven hundred base pairs.



Haplotype Inference Problem

- Assume we have SNP data.
  - If the two haplotypes for an individual are: ACG and TCA then the result of the genotyping experiment is: {A,,T}, {C,C} and {G,A}.
  - For the homozygous genotype: {A,A} {T,T} we know for sure the two haplotypes, namely: AT and AT



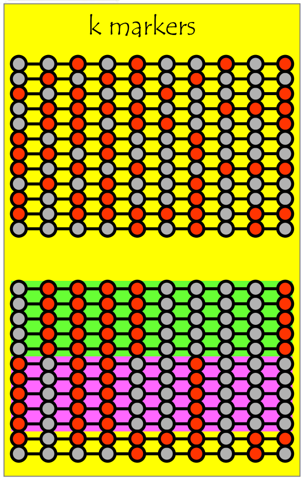
Haplotype Inference Problem

- Again, assume we have SNP data.
  - For the single site heterozygous genotype: {A,A} {C,T} then again we know for sure the two haplotypes, namely: AC and AT.
  - BUT for the double site heterozygous genotype: {A,T}, {C,T} then we can have two possible reconstructions, namely: AC and TT or AT and TC.

Haplotype Inference

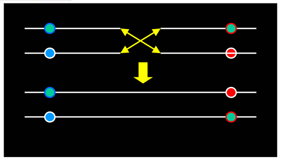
- The problem that we try to solve is this:
  - Given a pool of genotypes, we wish to estimate the haplotypes of each individual in the pool and also their frequencies.

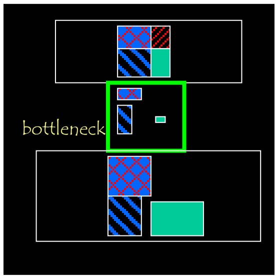
Haplotype diversity



- The most useful multi-marker measures of associations are related to haplotype diversity
- If a genotype is heterozygous for k positions then 2^{k-1} possible haplotype pairs exist, but we want the unique, true pair.
  - Random assortment of alleles at different sites
  - Strong association: few common haplotypes (Reduced haplotype diversity)

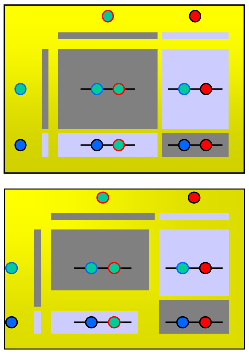
The determinants of allelic association





- Recombination:
  - breaks down allelic association by "randomizing" allele combinations
- Demographic history of effective population size:
  - bottlenecks increase allelic association by non-uniform re-sampling of allele combinations (haplotypes)

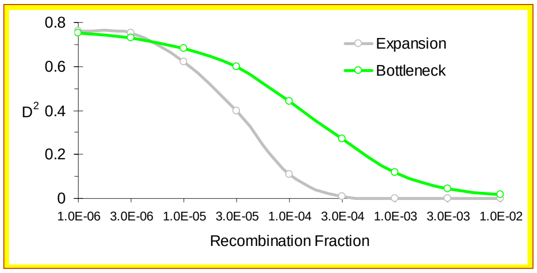
Linkage disequilibrium



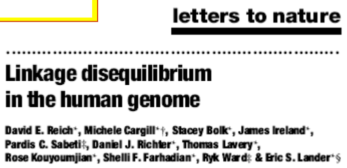
D=f(●●) – f(●) x f(●)

- LD measures the deviation from random assortment of the alleles at a pair of polymorphic sites
- Other measures of LD are derived from D, by e.g. normalizing according to allele frequencies (r^2)

Strength of LD in the human genome



- LD is stronger, extends longer than previously thought

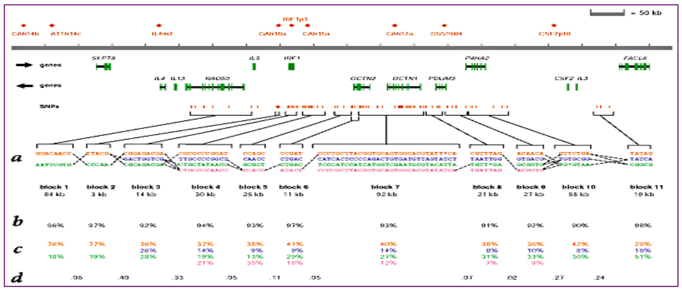

letters to nature

Linkage disequilibrium in the human genome

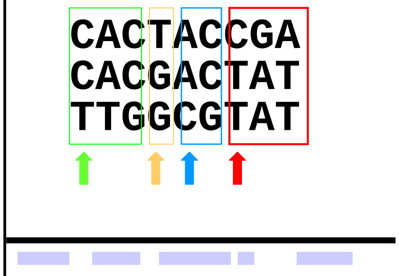
David E. Reich*, Michele Cargill*†, Stacey Bolk*, James Ireland*, Pardis C. Sabeti‡, Daniel J. Richter*, Thomas Lavery*, Rose Kouyoumjian*, Shelli F. Farhadian*, Ryk Ward‡ & Eric S. Lander*§

Haplotype blocks

- Experimental evidence for reduced haplotype diversity (mainly in European samples): Daly et al, *Nature Genetics* 2001

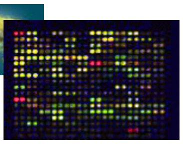


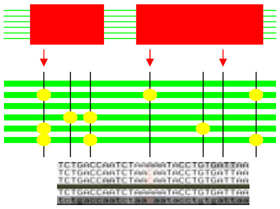
Medical Genetics



- Within blocks a small number of SNPs are sufficient to distinguish the few common haplotypes ↦ significant marker reduction is possible
  - If the block structure is a general feature of human variation data, whole-genome association studies will be possible at a reduced genotyping cost. (Gibbs et al. Nature 2003)
- This motivated the HapMap project.

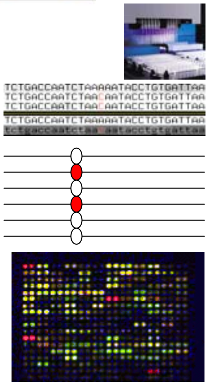
The HapMap initiative





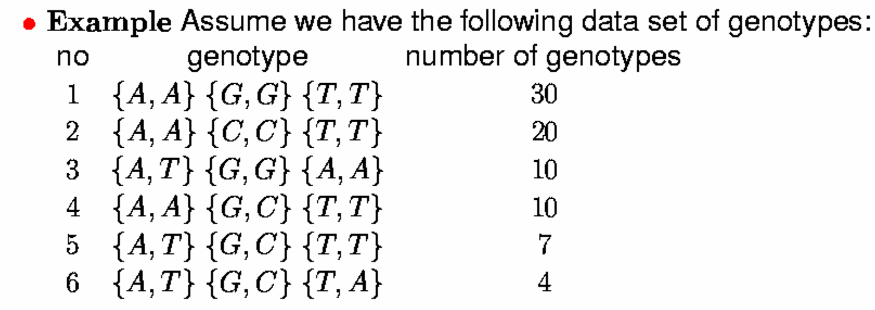
- Goal:
  - to map out human allele and association structure of at the kilobase scale
- Deliverables:
  - a set of physical and informational reagents

HapMap physical reagents



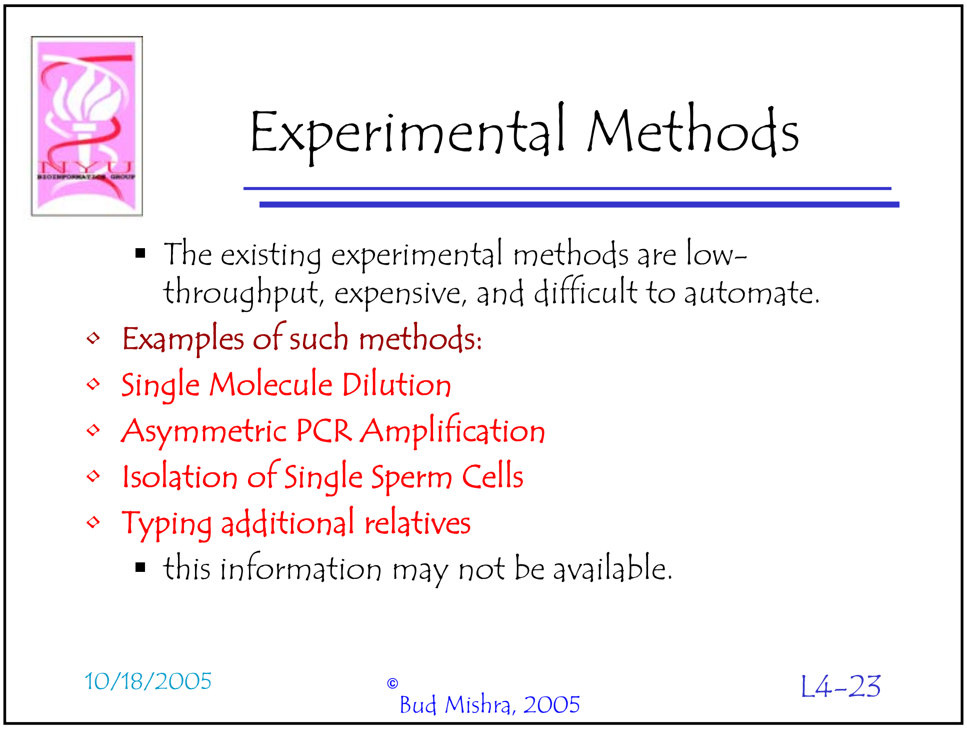
- Reference samples:
  - 4 world populations, ~100 independent chromosomes from each
- SNPs:
  - computational candidates where both alleles were seen in multiple chromosomes
- Genotypes:
  - high-accuracy assays from various platforms; fast public data release

Genotype Data



- **Example** Assume we have the following data set of genotypes:

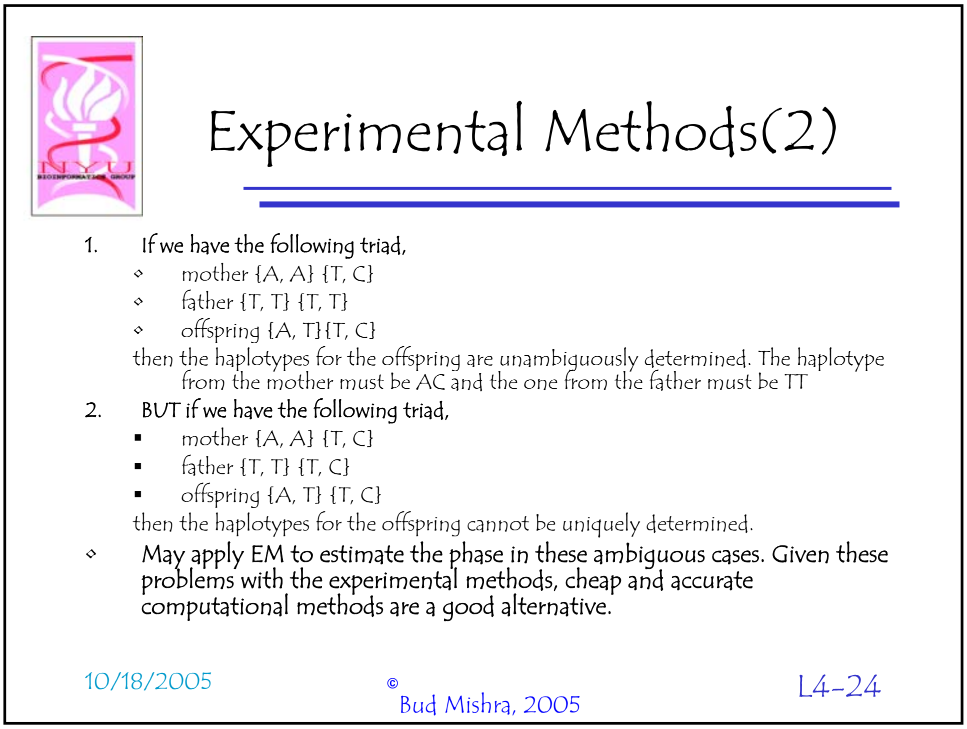
no	genotype	number of genotypes
1	$\{A,A\}\ \{G,G\}\ \{T,T\}$	30
2	$\{A,A\}\ \{C,C\}\ \{T,T\}$	20
3	$\{A,T\}\ \{G,G\}\ \{A,A\}$	10
4	$\{A,A\}\ \{G,C\}\ \{T,T\}$	10
5	$\{A,T\}\ \{G,C\}\ \{T,T\}$	7
6	$\{A,T\}\ \{G,C\}\ \{T,A\}$	4



Experimental Methods

- The existing experimental methods are low-throughput, expensive, and difficult to automate.

- Examples of such methods:
- Single Molecule Dilution
- Asymmetric PCR Amplification
- Isolation of Single Sperm Cells
- Typing additional relatives
  - this information may not be available.



Experimental Methods(2)

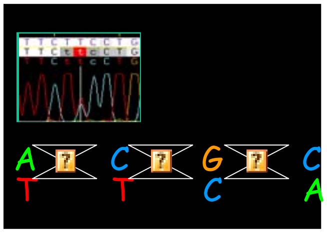
1. If we have the following triad,
   - mother {A, A} {T, C}
   - father {T, T} {T, T}
   - offspring {A, T}{T, C}

   then the haplotypes for the offspring are unambiguously determined. The haplotype from the mother must be AC and the one from the father must be TT
2. BUT if we have the following triad,
   - mother {A, A} {T, C}
   - father {T, T} {T, C}
   - offspring {A, T} {T, C}

   then the haplotypes for the offspring cannot be uniquely determined.

- May apply EM to estimate the phase in these ambiguous cases. Given these problems with the experimental methods, cheap and accurate computational methods are a good alternative.

Informational reagents



- The problem:
  - the substrate for genotyping is diploid, genomic DNA; phasing of alleles at multiple loci is in general not possible with certainty
- Experimental methods are expensive
  - (single-chromosome isolation followed by whole-genome PCR amplification, radiation hybrids, somatic cell hybrids)

Computational Methods

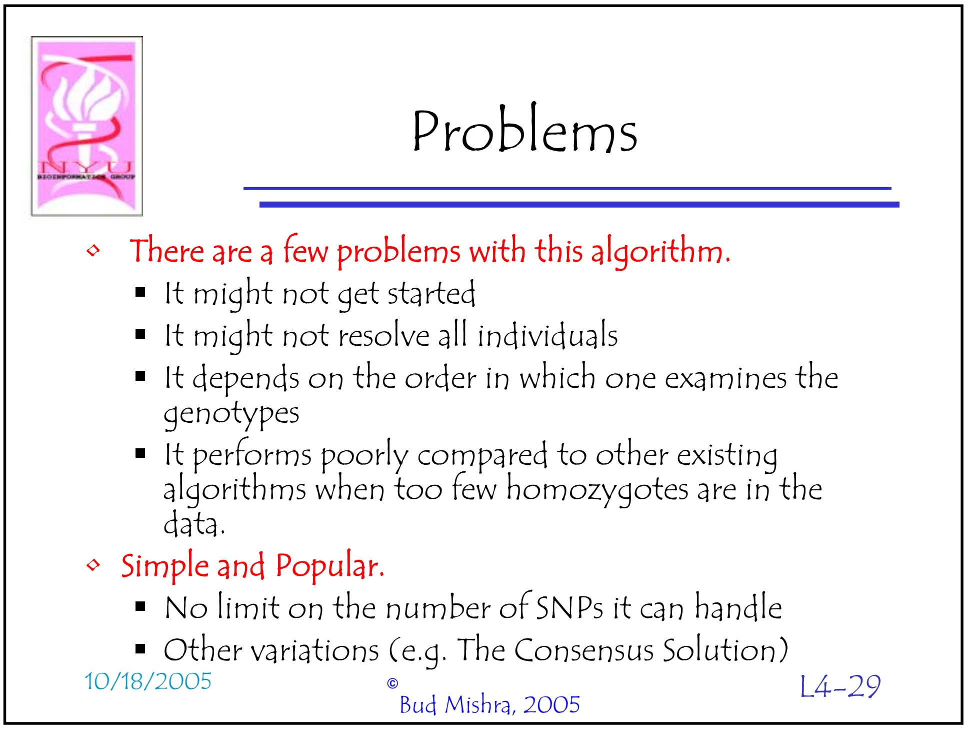
- Clark's Algorithm
- Perfect Phylogeny Solutions
- Statistical Solutions

Clark's Algorithm -1990

- This is a parsimony approach
  - It tries to solve the genotypes in the data set with as few haplotypes as possible.
  - It starts with the list of haplotypes that can by unambiguously inferred from the genotype data, i.e. the ones coming from homozygous or single-site heterozygous individuals.
  - It then tries to solve the phase ambiguous individuals by using these already determined haplotypes.

Example

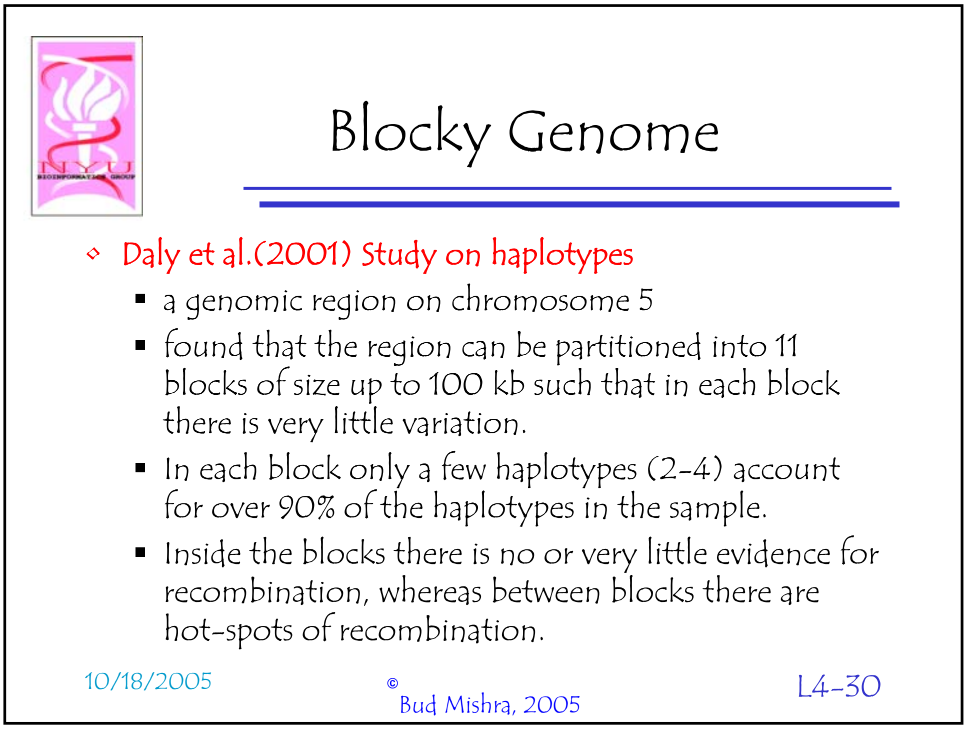
- For the data set that we have, we know that the following haplotypes are present in the population: {AGT, ACT, AGA, TGA}
- Now, for each known haplotype we traverse the list of ambiguous individuals and ask whether each individual can be solved by that haplotype: e.g. {A,T} {G, C} {T, T}, can be solved as AGT and TCT.
- By doing this we also acquired a new haplotype (TCT) that we add to the end of the list. We do this process until either all individuals are resolved or we can't find any more solutions.



Problems

- There are a few problems with this algorithm.
  - It might not get started
  - It might not resolve all individuals
  - It depends on the order in which one examines the genotypes
  - It performs poorly compared to other existing algorithms when too few homozygotes are in the data.
- Simple and Popular.
  - No limit on the number of SNPs it can handle
  - Other variations (e.g. The Consensus Solution)

10/18/2005 © Bud Mishra, 2005 L4-29



Blocky Genome

- Daly et al.(2001) Study on haplotypes
  - a genomic region on chromosome 5
  - found that the region can be partitioned into 11 blocks of size up to 100 kb such that in each block there is very little variation.
  - In each block only a few haplotypes (2-4) account for over 90% of the haplotypes in the sample.
  - Inside the blocks there is no or very little evidence for recombination, whereas between blocks there are hot-spots of recombination.

10/18/2005 © Bud Mishra, 2005 L4-30

Blocky Genome

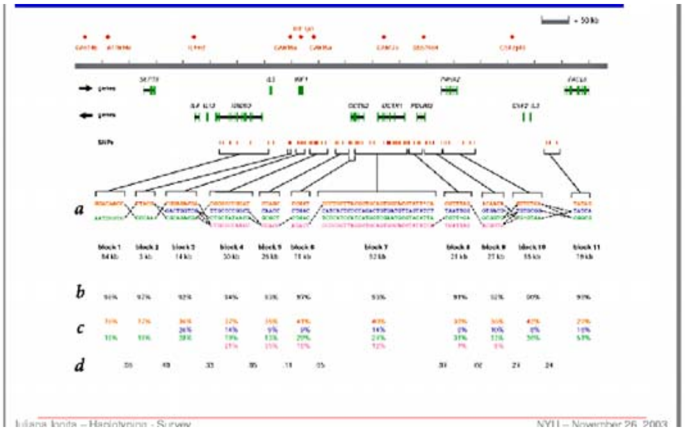
- While more studies are necessary to confirm that this blocky structure of the genome is general across the genome, other studies (e.g. Rioux et al.) agree with these findings.

Blocky Genome

- Reducing the complexity of the genome.
  - Having such an extended LD is important because it means that only few sites encode the information present in the entire region. (Knowing the information at these sites gives you the entire haplotype).
  - So no need to genotype all sites.
- Motivated by these findings, several deterministic algorithms that work specifically on these blocks of limited diversity have been designed.

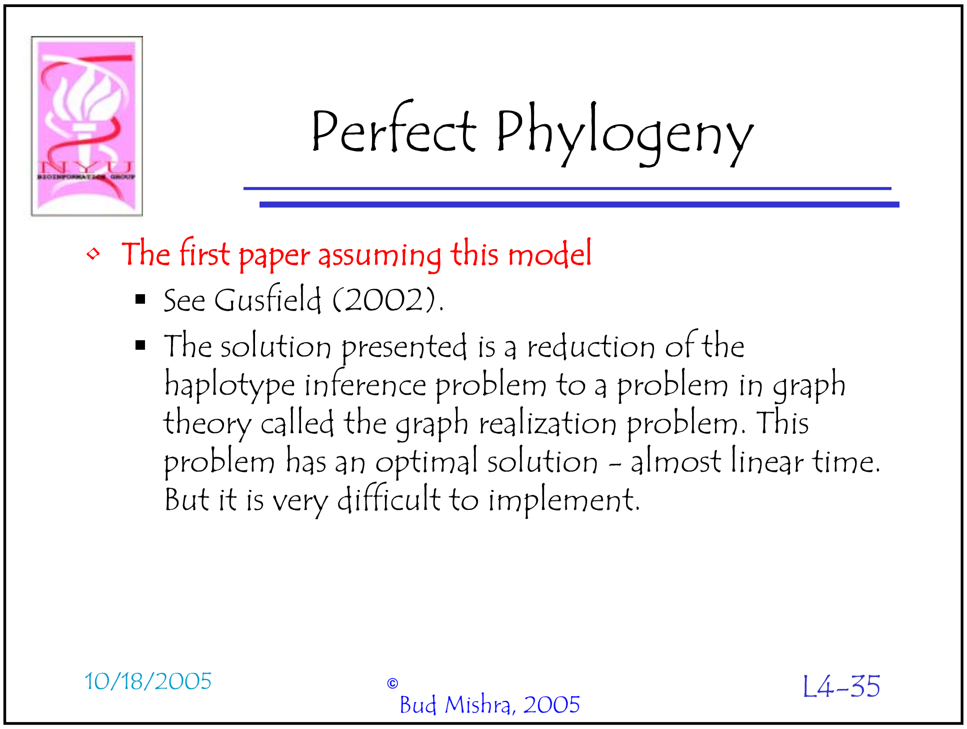
Blocky Genome

– Picture Daly et al.



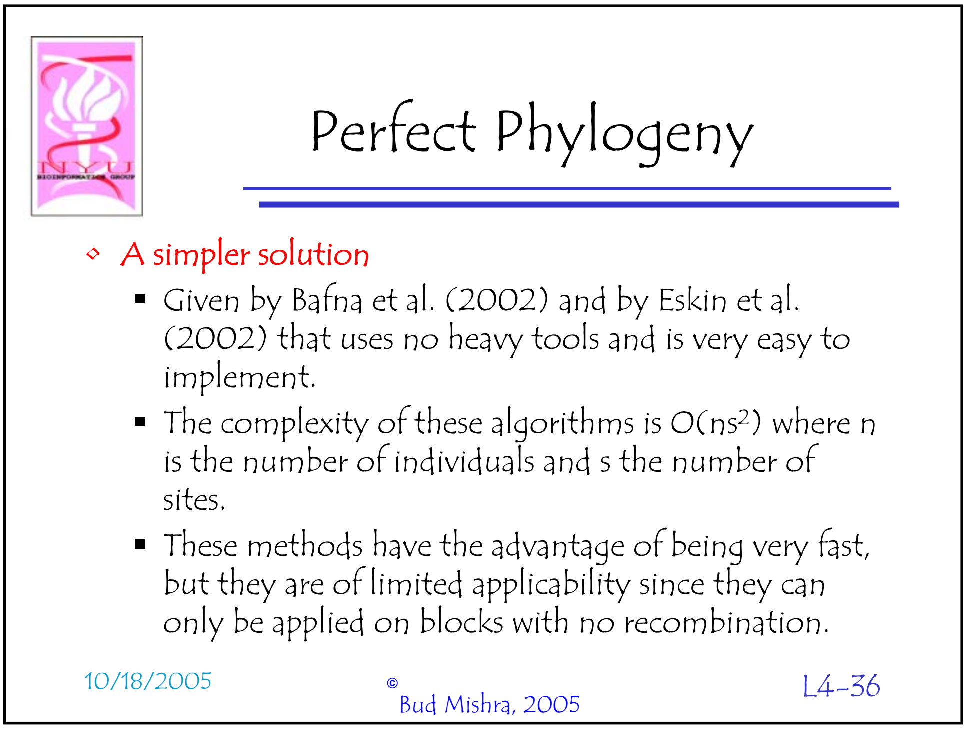
Perfect Phylogeny 2001-03

- The Perfect Phylogeny model of haplotype evolution
  - It assumes that there is no recombination and the usual infinite-site mutation model of population genetics applies.
  - Given the existence of these blocks, the PP model seems a reasonable model when working with SNP data.



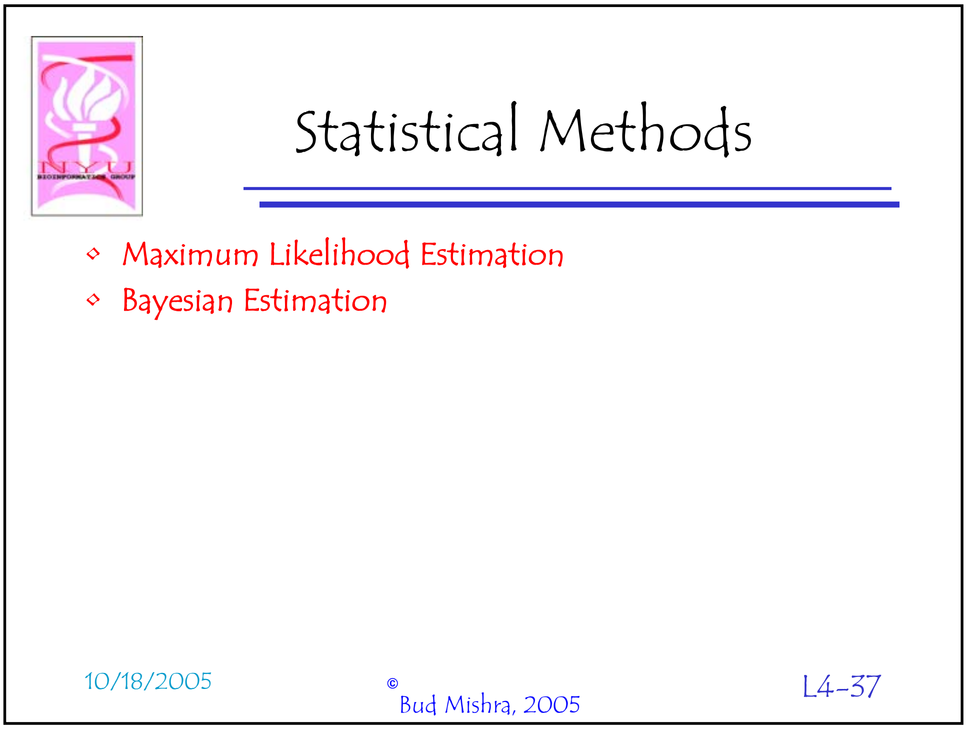
Perfect Phylogeny

- The first paper assuming this model
  - See Gusfield (2002).
  - The solution presented is a reduction of the haplotype inference problem to a problem in graph theory called the graph realization problem. This problem has an optimal solution – almost linear time. But it is very difficult to implement.



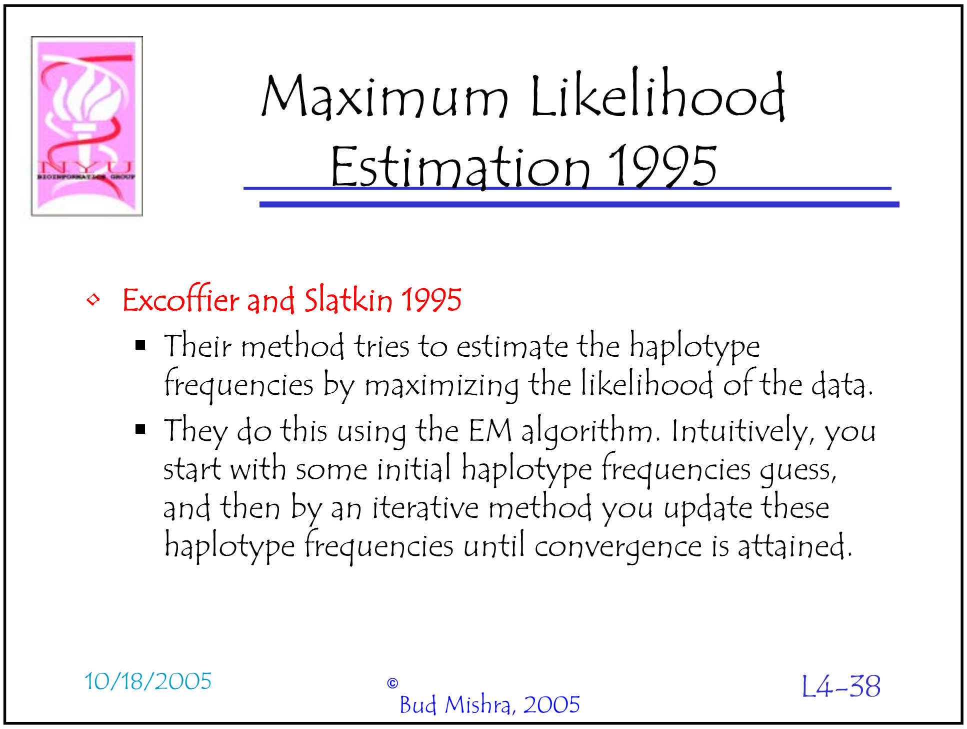
Perfect Phylogeny

- A simpler solution
  - Given by Bafna et al. (2002) and by Eskin et al. (2002) that uses no heavy tools and is very easy to implement.
  - The complexity of these algorithms is $O(ns^2)$ where n is the number of individuals and s the number of sites.
  - These methods have the advantage of being very fast, but they are of limited applicability since they can only be applied on blocks with no recombination.



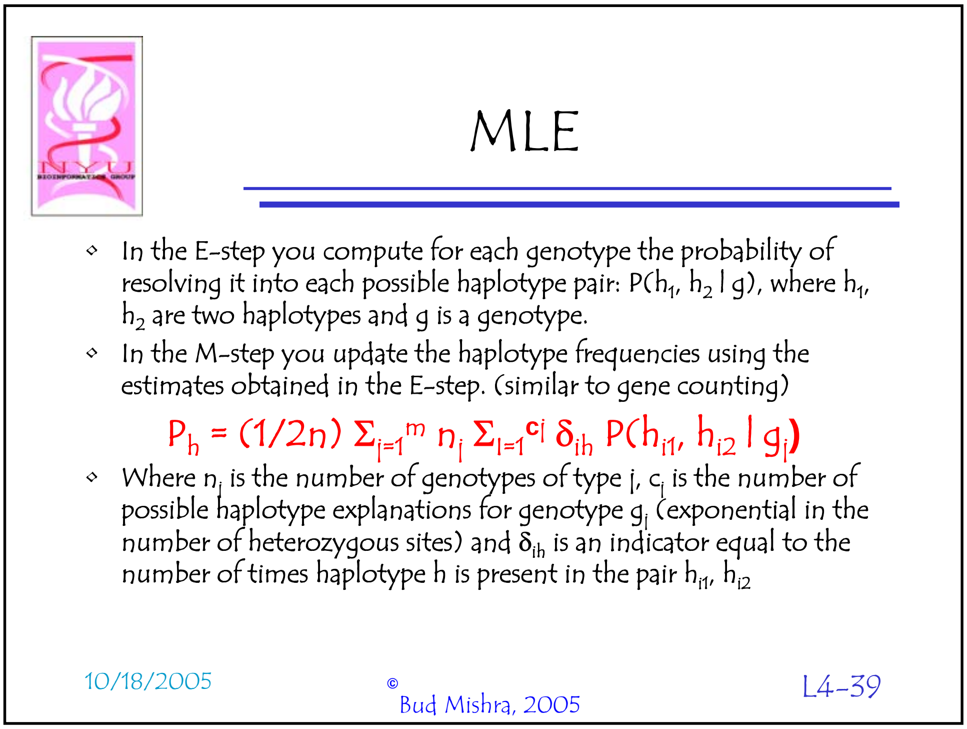
Statistical Methods

- Maximum Likelihood Estimation
- Bayesian Estimation



Maximum Likelihood Estimation 1995

- Excoffier and Slatkin 1995
  - Their method tries to estimate the haplotype frequencies by maximizing the likelihood of the data.
  - They do this using the EM algorithm. Intuitively, you start with some initial haplotype frequencies guess, and then by an iterative method you update these haplotype frequencies until convergence is attained.

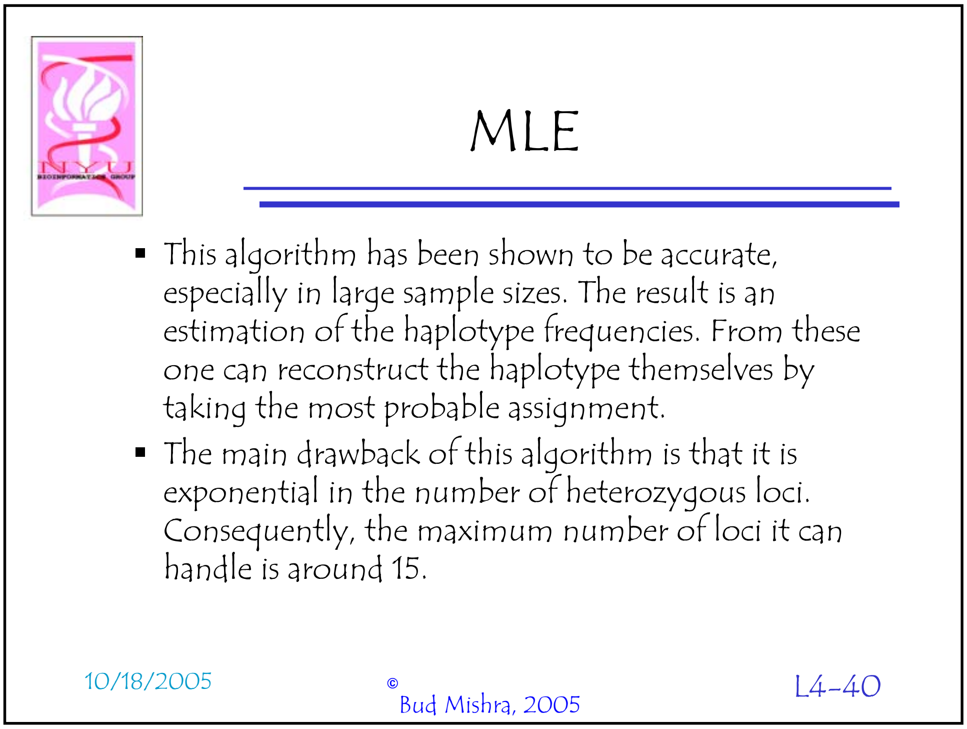


MLE

- In the E-step you compute for each genotype the probability of resolving it into each possible haplotype pair: $P(h_1, h_2 \mid g)$, where h_1, h_2 are two haplotypes and g is a genotype.
- In the M-step you update the haplotype frequencies using the estimates obtained in the E-step. (similar to gene counting)

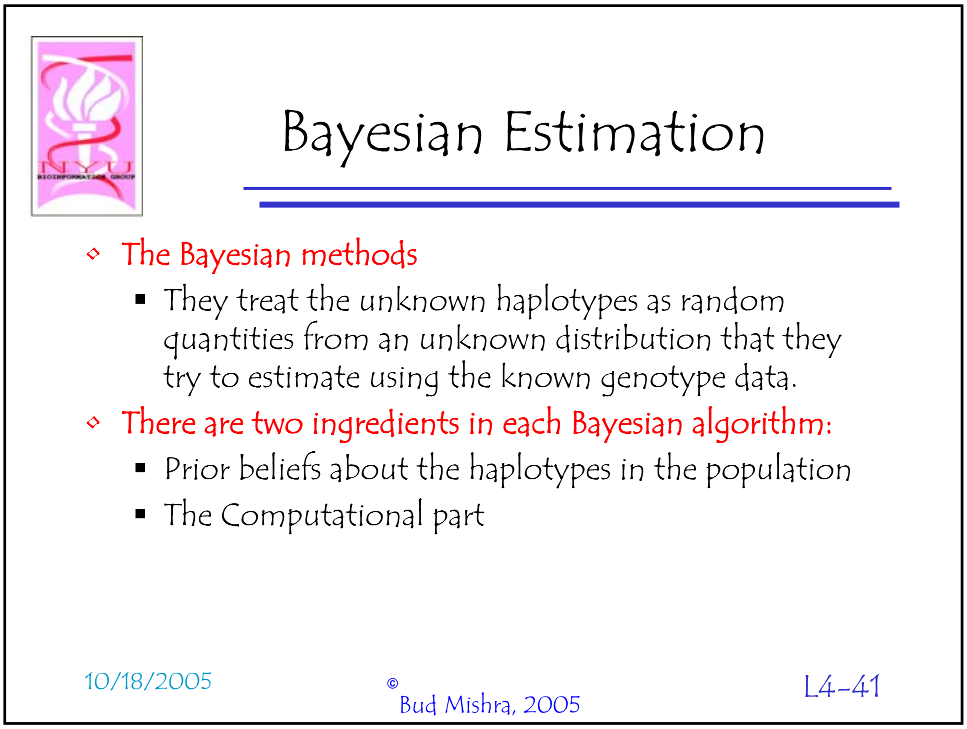
$$P_h = (1/2n)\ \Sigma_{j=1}^{m}\ n_j\ \Sigma_{l=1}^{c_j}\ \delta_{ih}\ P(h_{i1}, h_{i2} \mid g_j)$$

- Where n_j is the number of genotypes of type j, c_j is the number of possible haplotype explanations for genotype g_j (exponential in the number of heterozygous sites) and δ_{ih} is an indicator equal to the number of times haplotype h is present in the pair h_{i1}, h_{i2}



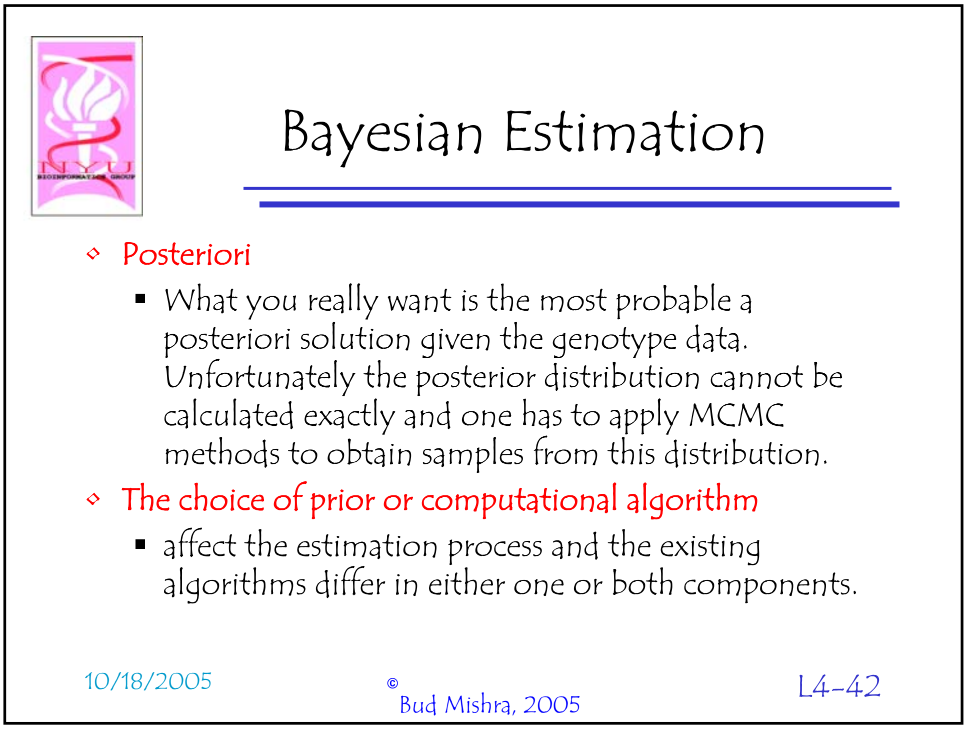
MLE

- This algorithm has been shown to be accurate, especially in large sample sizes. The result is an estimation of the haplotype frequencies. From these one can reconstruct the haplotype themselves by taking the most probable assignment.
- The main drawback of this algorithm is that it is exponential in the number of heterozygous loci. Consequently, the maximum number of loci it can handle is around 15.



Bayesian Estimation

- The Bayesian methods
  - They treat the unknown haplotypes as random quantities from an unknown distribution that they try to estimate using the known genotype data.
- There are two ingredients in each Bayesian algorithm:
  - Prior beliefs about the haplotypes in the population
  - The Computational part



Bayesian Estimation

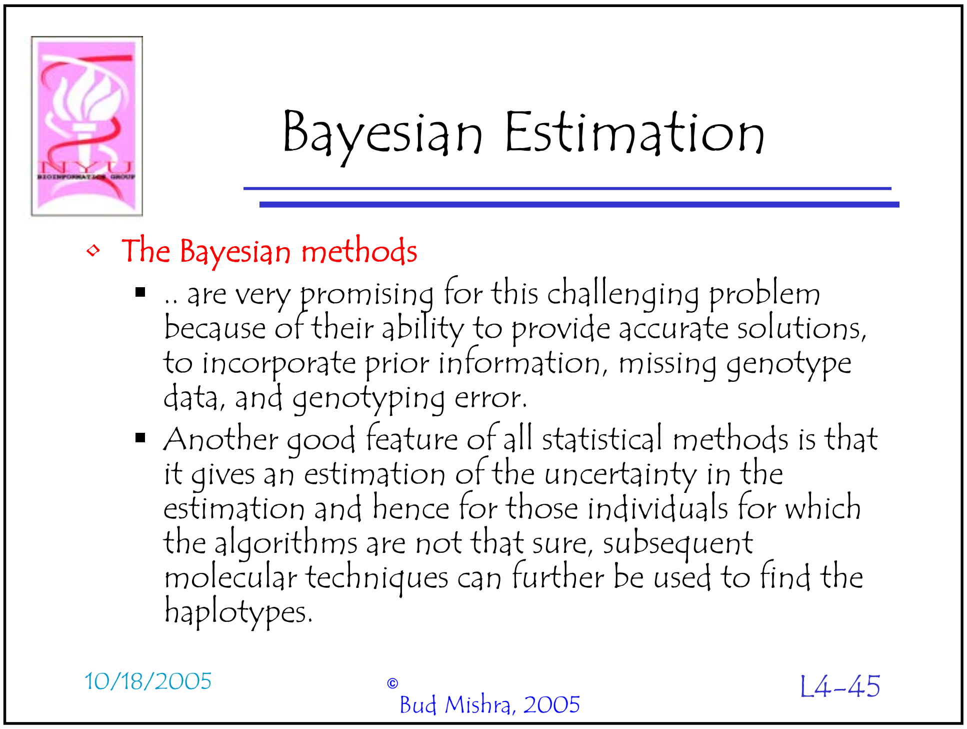
- Posteriori
  - What you really want is the most probable a posteriori solution given the genotype data. Unfortunately the posterior distribution cannot be calculated exactly and one has to apply MCMC methods to obtain samples from this distribution.
- The choice of prior or computational algorithm
  - affect the estimation process and the existing algorithms differ in either one or both components.

Bayesian Estimation

- Stephens et al.
  - Two Bayesian algorithms were proposed by Stephens et al. Both use a Gibbs sampler, but different priors.
  - The Gibbs sampler is an MCMC algorithm that constructs a MC whose stationary distribution is P(H|G).

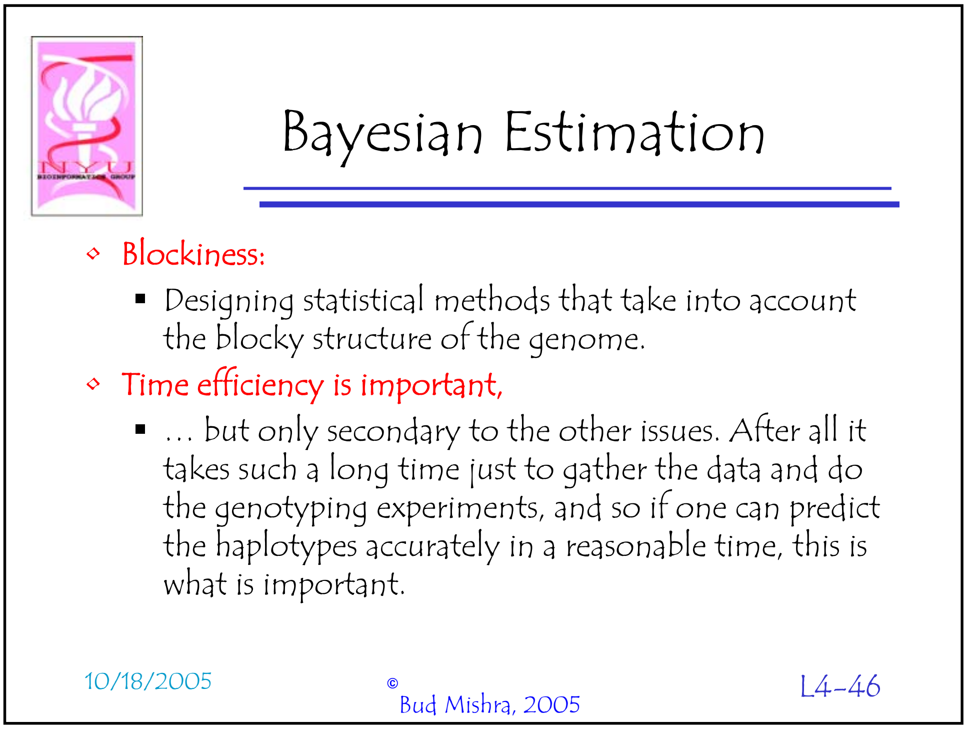
Bayesian Estimation

- It starts with an initial guess of haplotypes H^0 and then repeatedly chooses an individual at random from the ambiguous individuals and estimates its haplotypes given the haplotypes of the other individuals:
  - Sample (h_{i1}, h_{i2}) from $P((h_1, h_2) | G, H_{-i})$ where H_{-i} are the estimated haplotypes for the other individuals.
  - Repeat this process until convergence.
- These conditional distributions are influenced by the priors assumed. The first one assumes a Dirichlet prior on the haplotype frequencies, while the second one assumes a better prior based on the coalescent



Bayesian Estimation

- The Bayesian methods
  - .. are very promising for this challenging problem because of their ability to provide accurate solutions, to incorporate prior information, missing genotype data, and genotyping error.
  - Another good feature of all statistical methods is that it gives an estimation of the uncertainty in the estimation and hence for those individuals for which the algorithms are not that sure, subsequent molecular techniques can further be used to find the haplotypes.



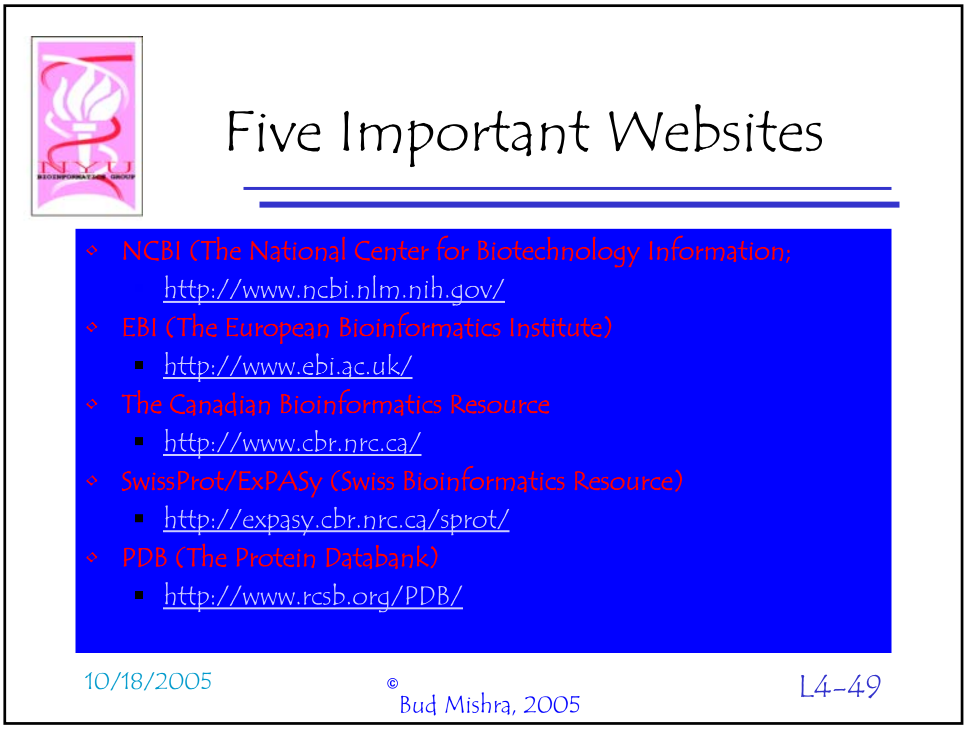
Bayesian Estimation

- Blockiness:
  - Designing statistical methods that take into account the blocky structure of the genome.
- Time efficiency is important,
  - … but only secondary to the other issues. After all it takes such a long time just to gather the data and do the genotyping experiments, and so if one can predict the haplotypes accurately in a reasonable time, this is what is important.

Bioinformatics Databases of Interest

Bioinformatics DataSources

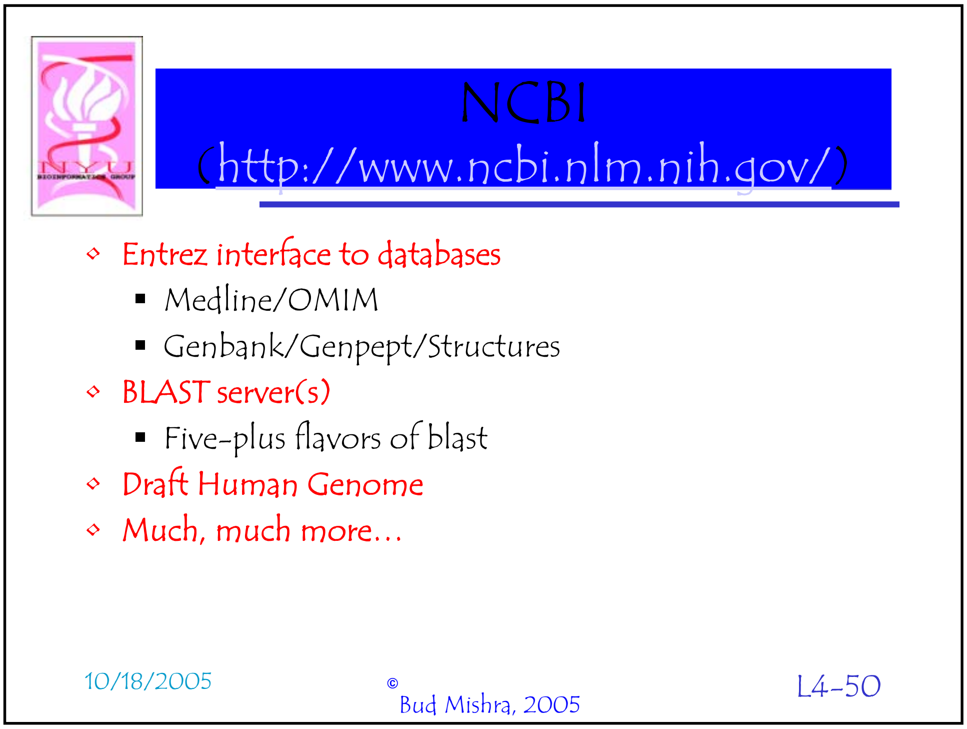
- Database interfaces
  - Genbank/EMBL/DDBJ, Medline, SwissProt, PDB, …
- Sequence alignment
  - BLAST, FASTA
- Multiple sequence alignment
  - Clustal, MultAlin, DiAlign
- Gene finding
  - Genscan, GenomeScan, GeneMark, GRAIL
- Protein Domain analysis and identification
  - pfam, BLOCKS, ProDom,
- Pattern Identification/
- Characterization
  - Gibbs Sampler, AlignACE, MEME
- Protein Folding prediction
  - PredictProtein, SwissModeler



Five Important Websites

- NCBI (The National Center for Biotechnology Information;
  - http://www.ncbi.nlm.nih.gov/
- EBI (The European Bioinformatics Institute)
  - http://www.ebi.ac.uk/
- The Canadian Bioinformatics Resource
  - http://www.cbr.nrc.ca/
- SwissProt/ExPASy (Swiss Bioinformatics Resource)
  - http://expasy.cbr.nrc.ca/sprot/
- PDB (The Protein Databank)
  - http://www.rcsb.org/PDB/

10/18/2005 © Bud Mishra, 2005 L4-49



NCBI (http://www.ncbi.nlm.nih.gov/)

- Entrez interface to databases
  - Medline/OMIM
  - Genbank/Genpept/Structures
- BLAST server(s)
  - Five-plus flavors of blast
- Draft Human Genome
- Much, much more…

10/18/2005 © Bud Mishra, 2005 L4-50

EBI (http://www.ebi.ac.uk/)

- SRS database interface
  - EMBL, SwissProt, and many more
- Many server-based tools
  - ClustalW, DALI, …

SwissProt (http://expasy.cbr.nrc.ca/sprot/)

- Curation…
  - Error rate in the information is greatly reduced in comparison to most other databases.
- Extensive cross-linking to other data sources
- SwissProt is the 'gold-standard' by which other databases can be measured, and is the best place to start if you have a specific protein to investigate

A few more resources

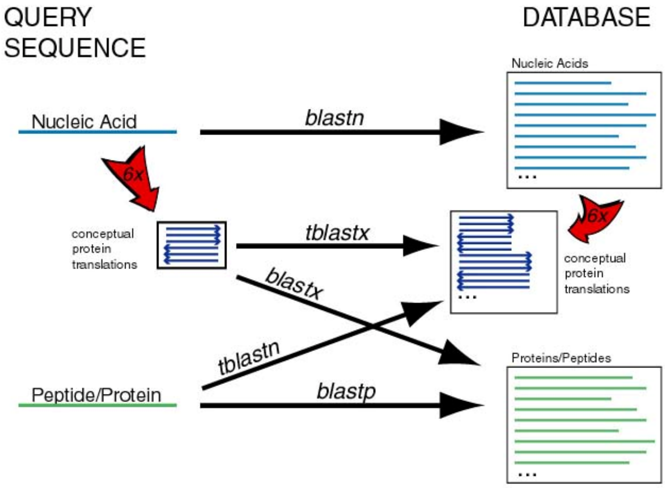
- Human Genome Working Draft
  http://genome.ucsc.edu/
- TIGR (The Institute for Genomics Research)
  http://www.tigr.org/
- Celera
  http://www.celera.com/
- (Model) Organism specific information:
  - Yeast: http://genome-www.stanford.edu/Saccharomyces/
  - Arabidopis: http://www.tair.org/
  - Mouse: http://www.jax.org/
  - Fruitfly: http://www.fruitfly.org/
  - Nematode: http://www.wormbase.org/
- Nucleic Acids Research Database Issue
  http://nar.oupjournals.org/

Example 1:

- Searching a new genome for a specific protein
- Specific problem:
  - We want to find the closest match in *C. elegans* of *D. melanogaster* protein NTF1, a transcription factor

- First- understanding the different forms of blast

The different versions of BLAST



Some possible methods

- If the domain is a known domain:
- SwissProt
  - text search capabilities
  - good annotation of known domains
  - crosslinks to other databases (domains)
- Databases of known domains:
  - BLOCKS (http://blocks.fhcrc.org/)
  - Pfam (http://pfam.wustl.edu/)
  - Others (ProDom, ProSite, DOMO,…)

Nature of conservation in a domain

- For new domains, multiple alignment is your best option
  - Global: clustalw
  - Local: DiAlign
  - Hidden Markov Model: HMMER
- For known domains, this work has largely been done for you
  - BLOCKS
  - Pfam

Protein Tools

- Search/Analysis tools
  - Pfam
  - BLOCKS
  - PredictProtein (http://cubic.bioc.columbia.edu/predictprotein/predictprotein.html)

Different representations of conserved domains

- BLOCKS
  - Gapless regions
  - Often multiple blocks for one domain
- PFAM
  - Statistical model, based on HMM
  - Since gaps are allowed, most domains have only one pfam model

To be continued…

. . .