Introduction to Bioinformatics for Pharmacogenomics
for PHR 396F (Spring 2009 version)
by David Rigney, GENETWORKS Inc. (drigney@genetworks.com)

List of Terms and Topics:
Rationale for Pharmacogenomics
Bioinformatics
SNPs
NCBI and its web site
Entrez
Submitting DNA sequences to GenBank
BLAST for pairwise sequence comparison and alignment
clustering algorithms and multiple-alignment
Unigene
Entrez Gene
PubMed
OMIM
Pfam
SNP database
Hidden Markov Models
ESTs
Microarrays
The SNP Consortium & The International HapMap Project
Hierarchical and Non-Hierarchical Clustering
Gene Ontology
KEGG
GEO
Test your knowledge

Pharmacogenomics basic definition:

Pharmacogenomics is the analysis of the effect of genetic variation (polymorphisms) on drug response, with the practical aim of selecting optimal drugs and doses based on the genetic makeup of patients.

This subject has a long history. For example, Roger G. Williams, Professor at the University of Texas and the Clayton Foundation Biochemical Institute (best know for discovering the B-vitamin pantothenic acid) wrote a book on the subject in 1956, entitled Biochemical Individuality, with chapters "Genetic Basis of Biochemical Individuality" and "Pharmacological Manifestations."

A person's racial background is already well-known to be one predictor (but not a determinant) of what drug concentrations will work best, for certain drugs:
Wood AJ, Zhou HH. Ethnic differences in drug disposition and effectiveness. Clin Pharmacokinet. 1991; 20:350-373; Burroughs VJ, Maxey RW, Levy RA. Racial and ethnic differences in response to medicines: Towards individualized pharmaceutical treatments. J. National Med. Assoc. 94: (suppl., Oct 2002) 1-26.

Nevertheless, use of race (rather than an individual's family history or actual genotype) is controversial, as evidenced by recent public discussion surrounding the following drugs:
- BiDil, a patented heart failure drug that combines two generic drugs (isorbide dinitrate and hydralazine HCl, which is the first drug to receive FDA approval for indication based on race (African-Americans)

- Iressa (gefitinib), a cancer drug (tyrosine kinase inhibitor) that failed to show a benefit in the general population, but that initially appeared to extend survival in Japanese populations
- Crestor, a cholesterol-lowering drug that has a recommended initial dose that is lower for "North Americans of Asian descent" (5mg) than other North Americans, and even lower in Japan (2.5mg)
To understand the controversy about basing medical decisions partly on self-identified racial type, see: Genetic Variation, Classification and 'Race' by L.B. Jorde and S.P. Wooding. Nature Genetics Supplement (2004) 36:S28-S33.

So what's new about pharmacogenomics?
The field of pharmacogenomics really developed only in the last 15 years because of new technology for the large-scale and high-speed collection of data, for measuring **DNA sequence variations**, especially single nucleotide polymorphisms (SNPs), and **drug responses**, especially mRNA expression analysis using microarrays. Coincidentally, the rise of the internet made it possible for many people to access these data collections.

Bioinformatics (the subject of this lecture) is concerned with the construction of computer databases for these large-scale data collections, as well as computer methods for accessing, viewing, and analyzing those databases. I will demonstrate the subject with bioinformatics tools and databases at the NIH web site, called the National Center for Biotechnology Information (NCBI). The main web site is http://www.ncbi.nlm.nih.gov/

What is a single nucleotide polymorphism (SNP)?
It is variability in a single nucleotide (A, C, G, or T) within an otherwise fixed DNA sequence, where a certain fraction of the population have one nucleotide (say C) and another fraction of the population have another nucleotide (say T).

For example, on chromosome 17 there is a gene called BRCA1 that is associated with breast and ovarian cancer. Within the DNA sequence of that gene, there is a part that looks like this:

```
TTCATGGGCA TTAATTGCAT GAATGTGGTT AGATTAAAAG GTGTTCAGCT AGAACTTGTA
GTTCCATACT AGGTGATTTC AATTCCTGTG CTAAAATTAA TTTGTATGAT ATATTTTCAT
TTAATGGAAA GCTTCTCAAA GTATTTCATT TTCTTGGTGC CATTTATCGT TTTTGAAGCA
GAGGGATACC ATGCAACATA ACCTGATAAA GCTCCAGCAG GAAATGGCTG AACTAGAAGC
TGTGTTAGAA CAGCATGGGA GCCAGCCTTC TAACAGCTAC CCTTCCATCA TAAGTGACTC
Y
TCTGCCCTTG AGGACCTGCG AAATCCAGAA CAAAGCACAT CAGAAAAAGG TGTGTATTGT
TGGCCAAACA CTGATATCTT AAGCAAAATT CTTTCCTTCC CCTTTATCTC CTTCTGAAGA
GTAAGGACCT AGCTCCAACA TTTTATGATC CTTGCTCAGC ACATGGGTAA TTATGGAGCC
TTGGTTCTTG TCCCTGCTCA CAACTAATAT ACCAGTCAGA GGGACCCAAG GCAGTCATTC
ATGTTGTCAT CTGAGTACCT ACAACAAGTA GATGCTATGG GGAGCCCATG GAAGATACAT
```

For 68% of individuals in a global population, the base shown as Y is actually T
For 32% of the individuals, the base shown as Y is actually C.
(In more detail, on their two chromosome 17s, 16% are C/C, 32% are C/T, and 52% are T/T).

How do I know this? By searching the NIH databases, as follows:
Go to http://www.ncbi.nlm.nih.gov/ Click on Entrez Home
In the box "Search across databases for" enter BRCA1, then click "go". The results then show how many entries there are in various NIH databases for BRCA1. Click on the SNP database icon, then

click on the "Human" tab, then in the "Sort by" box select SNP_ID. The data given above are for SNP "rs1060915", found by clicking the "Next Page" button.(Click on "rs1060915" to see the data).

So how are SNPs discovered? By sequence comparisons.
If you sequence a portion of someone's DNA, you can deposit it in a database like GenBank. For example, there is a shared resource at UT with a machine for sequencing DNA. See http://www.icmb.utexas.edu/core/DNA/ . You can then deposit your DNA sequence data into a public repository of sequence information (**GenBank**) using a software tool like Sequin or Bankit. See http://www.ncbi.nlm.nih.gov/Genbank/submit.html . After the sequence is deposited, the database gives it a unique identifier name (**accession number**), such as U14680.

To see if someone has already submitted a DNA sequence that is very similar to the one that you submitted, you can use a search tool like **BLAST**, as follows:
Go to http://www.ncbi.nlm.nih.gov/, then click on the word "BLAST", which is in the line above the search boxes. Then click on "nucleotide blast".
In the "Enter Query Sequence" box, copy/paste the DNA sequence for the SNP shown above. Use the options to for "Human genomic + transcript" and "highly similar sequences". Then click the BLAST button. The list of results describes each individual result with a label, such as one containing NM_007294.2 Find that one and click on it. You will find there that "This record has been curated by NCBI staff. The reference sequence was derived from BC072418.1, U14680.1 and BU617173.1." Click on the link to U14680.1 as an example of an original Genbank deposit that matches your BLAST query. If you had used "others (nr, etc.)" for your BLAST search instead of "Human genomic + transcript", the search results would include actual Genbank submissions like U14680.1 instead of just consensus reference sequences.

The BLAST bioinformatics tool works by searching the GenBank database to find other DNA sequences that can line up pair-wise with the sequence that you specified for the search. Note from the results of this search example that many sequences are found to differ only by a single nucleotide, namely, the SNP mentioned above. So, SNPs are discovered by simply comparing sequences from different individuals, using a computer program like BLAST.

Digression 1: Unigene and Multiple-Sequence Comparisons
Because many people deposit DNA sequences into GenBank independently of one another, there are many instances in which different sequences are very similar to one another. For example, different laboratories may discover the same gene independently of one another and deposit their results into GenBank independently. So, when you perform a sequence search you may be interested in knowing, for example, whether there is a group of sequences that correspond to alleles of the same gene, in Homo sapiens or in other species. Because this is such a common application, the NIH has already done the grouping of **multiple sequences** throughout GenBank, using what are called **clustering algorithms** (which will be described later). For example, Go to http://www.ncbi.nlm.nih.gov/ and click on Entrez Home.
In the box "Search across databases" enter U14680 (which is the accession number for an early BRCA1 sequence), then click "go". Then, click on "Unigene". Then click on the results link (Hs. 194143) which is the "UniGene" group of related sequences into which the BRCA1 gene is assigned.

Digression 2: Links to other databases with information about the gene:
After reaching the Unigene entry for a gene, you can find a lot of functional information about the gene by clicking on the "Links" hyperlink, located in the upper-right part of the web page:
Entrez "Gene" – a database giving a summary of each known gene, along with literature references about specific functions of the gene (GeneRIF). Click on "Gene" to see the information. Note that it has a section called Gene Ontology with sections describing function, process, and component. Gene Ontology will be discussed later in more detail.
PubMed – a database of literature citations. If you click on the PubMed links in GeneRIF, you can read the associated literature citation.
OMIM (Online Mendelian Inheritance in Man) – summarizes the medical significance of each known gene. After going to this link, click on the one numbered 113705: BREAST CANCER 1 GENE; BRCA1 BREAST CANCER, TYPE 1, INCLUDED
SNP - the single nucleotide polymorphism database (dbSNP). For example, you can also reach the rs1060915 data (mentioned above) through this link to UniGene.
Conserved Domains -- shows short domains within the protein that are thought to have functional significance. These functional domains were obtained using the Pfam (Protein family) database and related tools. To see more about what these functional motifs are, click on them or go to http://pfam.sanger.ac.uk/, select "Keyword Search", and enter "BRCA1" in the keyword search box.

Digression 3: Hidden Markov Models (HMMs)
Pfam (and many other databases and tools for finding patterns such as intron/exon splice sites), make use of mathematical models known as Hidden Markov Models (HMMs). Models of this type are derived from early work done for automatic speech recognition (a kind of artificial intelligence) that represents speech as a stream of sounds, just like DNA and protein sequences are streams of nucleotides and amino acids. All such models represent the changing state of a system in a probabilistic way. They are called Markov models because changes from one state to another are independent of previous states (Example: flipping a coin, which produces a sequence of heads and tails HTTHTH...). A DNA sequence is a similar string (aacatgg...), where the next base (a, g, c, or t) is described in the model by a table of transition probabilities (Paa for a→a, Pat for a→t, Pgt for g→t, etc.). These probability parameters are fit using a training set of known examples. However, HMMs are termed "Hidden" Markov Models because the true state of the HMM cannot in general be known by looking at the observed data. For example, the transition probabilities may depend on whether the next base to be predicted lies within an intron or an exon, which is not generally known in advance, and the intron/exon status becomes part of the state-variables representing the state of the system. Once the parameters of the model have been fit using a training set, the model may be applied to test data, for example, to predict whether a base is within an intron or exon (along with an estimate of the probability of error). In your handout "The Babel of Bioinformatics", it was noted that such predictive pattern recognition models sometimes have a mediocre success rate. This may be because the model does a poor job of representing a system, or because the training set is not big enough to provide reliable estimates for the model's parameters.

Digression 4: ESTs and the IMAGE Consortium cDNA libraries
After reaching the "Gene" and "UniGene" web pages for a gene, you can find a variety of gene sequences related to the entry. Some of these are not real sequences, but are instead “consensus” sequences among many genes in the Unigene cluster. Others correspond to publicly available clones in the cDNA libraries of the “**IMAGE consortium**” (see http://image.hudsonalpha.org/). A cDNA library is made by (1) extracting the mRNA from some tissue; (2) making cDNA from the mRNA

using reverse transcriptase, which ordinarily does not correspond to the full length of the mRNA; and (3) inserting the cDNA into a bacterial vector. After growing single clones of the bacteria, then extracting the bacterial vector, then cutting out the cDNA insert, you can obtain useful quantities of the cDNA derived from a particular mRNA species. An **EST** (expressed sequence tag) refers to the cDNA in a particular clone of the cDNA library. Until the cDNA is sequenced, it is not know which mRNA species corresponds to the clone. So, to identify the gene corresponding to the mRNA, partial sequences were obtained for one or both ends of each of the IMAGE Consortium clones. Often, by comparing the sequence of the EST with other sequences in GenBank, it is possible to assign the EST to a particular UniGene. However, many UniGene entries have no known function, and many UniGene entries consist of only of a few similar ESTs.

Digression 5: Microarrays

About 15 years ago, technology was developed that made it possible to deposit micro-drops of cDNA solution in a grid pattern onto a hybridization filter. The grid pattern is also called an array, so the filter with these deposited microdrops is called a microarray. Each of the drops may contain cDNA from a different clone in the IMAGE Consortium library. When a test DNA solution is then hybridized to the microarray, the hybridization will occur on the microarray only where there is sequence homology between DNA in the test solution and particular spots in the microarray. The hybridized spots can be visualized by autoradiography (or phosphorimagers) if the test DNA solution is first made radioactive by a labeling procedure. So, hybridization of the test DNA solution to the microarray can be used to quantify the presence, in the test solution of DNA, of DNA that matches the cDNA of the various clones on the microarray. If the DNA test solution was itself cDNA made from the mRNA of some specimen, the hybridization can be used to measure the presence of mRNA in the specimen of genes that are represented as spots on the microarray, i.e., measure a gene expression profile of the test specimen in terms of the genes represented on the microarray.

Different types of microarrays have been developed since the first ones described above. They have been developed to use oligonucleotide spots rather than cDNA; they have been made on glass microscope slides rather than nylon hybridization filters; and they have been developed for use with fluorescence labeling rather than radioactive labeling. But despite these technical variations, all such microarrays are basically used for the same types of hybridization experiments. (See http://www.gene-chips.com/ and http://en.wikipedia.org/wiki/DNA_microarray). The University of Texas has a microarray core facility that enables you to design and make microarrays for your own use (See http://cssb.icmb.utexas.edu/UTMCF/Home.html). More recently, microarray technology has also been combined with chromatin immuno-precipitation methods to locate the DNA binding sites of proteins of interest (transcription factors, replication-related proteins, histones, etc.). See http://en.wikipedia.org/wiki/ChIP-on-chip.

Pharmacogenomics and the SNP Consortium

In 1999, a consortium of pharmaceutical companies began to systematically identify and publish SNPs. By 2001, the consortium essentially finished its SNP discovery phase with the identification of 1.4 million SNPs, by comparing the DNA of 24 anonymous individuals from around the world. A subsequent consortium, The International HapMap Project, analyzes the DNA from 270 people from around the world (See http://snp.cshl.org/) It is now thought that about 10 million SNPs exist in human populations, where the rarer SNP allele has a frequency of at least 1%. Alleles of 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". It is estimated that there are 300-600 thousand haplotypes.

When the SNP discovery phase ended, the emphasis shifted to studying SNPs in populations, using selected SNPs that have already been discovered. Special experimental methods and equipment have been developed to do this, because the aim is to measure the population variations in the single variable base in a SNP (or haplotype), not sequence large sections of the genome. For example, one such device and method (developed by Affymetrix) makes use of a microarray of 1494 SNP DNA spots. The amount of DNA needed from a patient is 120 ng, which is used for hybridization to the microarray. Only if the patient's DNA contains a perfect match to the particular SNP allele represented in a microarray spot, will a signal be produced at that spot. See: PY Kwok. Methods for genotyping single nucleotide polymorphisms. Ann. Rev. Genetics Human Genet. 2001; 2:235-258. However, the advantages of these special methods may eventually be lost if nanopore sequencing makes it possible to sequence an entire genome for roughly $1000. See Branton et al (Nat Biotechnol. 2008 26(10):1146-53 at http://www.nature.com/nbt/journal/v26/n10/pdf/nbt.1495.pdf.

Ultimately, the intended application is – Perform a clinical trial to study the efficacy and toxicity of a drug; collect DNA from the subjects in the trial; look at SNPs in their DNA; and correlate drug response with the SNPs. To date, there have been some practical results:
Example: cytochrome P450 (CYP) enzymes break down 60% of all marketed drugs. Some people metabolize too quickly and others metabolize too slowly, depending on CYP genetic variation.
Example: thiopurines are used to treat childhood leukemia. The enzyme thiopurine methyltransferase (TPMT) breaks down the thiopurines, but there are genetic variations. A genetic test for TPMT may be used to guide selection of thiopurine doses.
Example: The FDA approved warfarin prescribing information based on CYP2C9 and VKORC1 variation. See http://www.fda.gov/bbs/topics/NEWS/2007/NEW01684.html
Examples of commercial tests: CodeLink P450 from GE Healthcare (formerly Amersham), xMAP technology from Luminex (an Austin TX company), DrugMEt from Jurilab, Roche AmpliChip CYP450. Fifteen companies make these tests available to the public, mostly through physicians, but some may make the tests available directly to patients (DNA Direct -$250 to $630). According to Genelex (http://www.healthanddna.com/drug-safety-dna-testing/dna-drug-reaction-testing.html), medical insurance will ordinarily pay for the tests if they are not for general screening purposes (e.g., for warfarin evaluation, $260 for CYP2C9 test, $550 includes additional vitamin K test, VKORC1). See "A Case Study of Personalized Medicine" by S.H. Katsanis, G. Javitt, and K. Hudson. Science (2008) 320:53-54, which argues that some such testing may be unjustified, based on a government-sponsored evaluation group known as EGAPP (http://www.egappreviews.org/).

However, pharmacogenomics is still largely a research enterprise, with results that are not generally ready for clinical application. For the status of publicly reported pharmacogenomics research see the web site http://www.pharmgkb.org/

A change in direction, away from SNP analysis and towards expression profiling:
Pharmaceutical companies began to lose their initial enthusiasm in SNP tests because:
* fewer success stories than anticipated
* common conditions like asthma, diabetes, hypertension, schizophrenia and heart disease involve many genes and environmental factors, so use of genetic testing is difficult
* if only one or two approved drugs are available for a treatment, why measure genotype?
* it became difficult to patent SNPs because they were public information
* companies realized that personalized medicine might segment their markets
* physicians do not want to know every genetic variation associated with drug responses
* fear that the FDA will mandate genotype-specific tests

Now, pharmacogenomics is more likely to be used by pharmaceutical companies to weed out potentially dangerous drugs in their development pipeline (not individualized medicine). A typical objective is – screen for patterns of toxicity response with microarray gene expression profiling. In the clinic, gene expression profiling is most often used to evaluate breast cancer treatment options using the Oncotype DX or MammaPrint tests. See http://www.ahrq.gov/clinic/tp/brcgenetp.htm. This involves the processing of a fresh or fixed tumor specimen to measure the expression of selected genes, and correlating that expression profile with a database of responses to different drugs/doses.

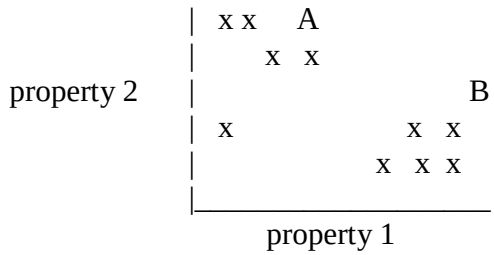
Gene Expression Profiling with MicroArrays
In a gene expression profiling experiment, mRNA is extracted from some tissue and used to prepare cDNA using reverse transcriptase. The cDNA may be made radioactive or fluorescent, by making it with radioactive or fluorescent bases. The sample cDNA is then hybridized to a microarray, containing spots corresponding to as many as several thousand genes. The microarray is then imaged (using a phosphorimager or fluorescence imaging equipment). Image processing software is then used to measure the hybridization intensity at each spot.
D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer and J.M. Trent. Expression profiling using cDNA arrays. Nature Genetics 21, Suppl 1: 10-14, 1999.

Example of commercial microarray: Affymetrix Rat Toxicology U34 Array
Example from GENETWORKS Inc. research and development in association with UT:
Go to http://www.biomedcentral.com/1471-2164/3/35
Fig.5 -- image processing to extract spot intensities
Fig.1 -- example of microarray images after background subtraction
Fig.2 -- example of scatterplot of data to find interesting microarray spots

In a typical pharmacology/toxicology experiment, the objective is to compare the gene expression response of a drug vs. placebo.
Example: work at UT for which GENETWORKS Inc. did consulting:
In this example, the herbal medicine ginseng is administered to lean and obese Zucker rats, which serve as an animal model for Type 2 diabetes.
Go to http://www.biomedcentral.com/imedia/1928369199122223_article.pdf
Fig. 8 – lean placebo vs. obese placebo scatterplot;
Fig. 7 -- obese placebo vs. obese ginseng scatterplot;
Fig. 6 -- example of clustering of the microarray data

Clustering:
The purpose of clustering is to organize a collection of objects systematically. Clustering problems arise in many fields, such as the classification of natural objects (rocks, stars, plant specimens, DNA sequences, etc.) There are two basic approaches to clustering, hierarchical and non-hierarchical. Hierarchical clustering first organizes pairs of objects in the collection as being most similar to one another. Then, the pairs are compared with other pairs, in order to organize pairs of pairs as being most similar to one another. Then the pairs of pairs are compared with other pairs of pairs, etc. until the entire collection is organized into a hierarchy. A final display of the organization of the collection looks like an evolutionary tree. The other approach to clustering (non-hierarchical clustering) aims to divide the object into sets of separate groups, where the number of groups and the properties of the groups are generally not known in advance.

The clustering methods are implemented as computer algorithms that produce their results automatically. Many clustering algorithms have been described (hundreds of them). There is no best clustering method, unless judged by some additional, application-specific criterion. Often, investigators judge the clustering results subjectively. For example, if two properties are measured for each object in the collection, and a scatter-plot is prepared showing those values for each object, the collection may look something like this:



By eye, it appears that the objects cluster into two groups (A and B) with one outlier that is hard to classify. Different clustering algorithms may insist that the outlier belongs to group A or that it should be a group by itself.
A good general reference on clustering methods is L. Kaufman and P.J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley (1990)
For an early discussion of clustering of microarray gene expression data see:
M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95: 14863-14868 1998.

Another example of clustering:
Y chromosome ancestry tests measure traits in Y-chromosome DNA called short tandem repeats (STRs). The measurement at each STR consists of counting the number of times that certain DNA sequences are repeated. Background information about the STR traits may be found at the following web site: http://www.cstl.nist.gov/div831/strbase//y_strs.htm. The tests are commercially available for a fee of about $100. For example, National Geographic will do the test as part of its Genographic Project (see https://www3.nationalgeographic.com/genographic/), by mailing in a cheek swab.

The Genographic Project measures 12 such traits on an individual's Y chromosome, where the typical (reference) number of repeats at each locus, the range in number, and the number observed in a particular individual are as follows:

Trait (Locus)	Reference count	Range	An individual's result
DYS393	12	9-17	14
DYS 19	15	10-19	15
DYS 391	11	6-14	10
DYS 439	13	9-14	11
DYS 389-1	12	9-17	13
DYS 389-2	29	24-34	18
DYS 388	12	10-18	13
DYS 390	24	17-28	23
DYS 426	12	10-12	11
DYS 385a	11	7-28	15
DYS 385b	11	7-28	15
DYS 392	13	6-17	12

Y chromosome data have already been clustered for many individuals, using methods adapted since the 1960s for phylogenetic analysis by Cavalli-Sforza and colleagues. In genetics, the clusters are called "haplogroups" and have names like E1b1a, G2c, J1, N, and R1a. The haplogroups and their names may change as more data are accumulated from more individuals.

An individual who has recent test results may find out the cluster (haplogroup) to which he belongs by entering his test data at the following web site: http://www.hprg.com/hapest5/index.html. At that web site, he would click on either of the links for "21-Haplogroup Program" and enter his data to find out which cluster his data most closely resembles. For example, if the data above are entered, it is calculated that the data belong to the I2b1 haplogroup with a 100% probability.

Interpreting the gene expression results:

When gene expression data are plotted as a scatterplot, certain genes stand out in the comparison of control vs. intervention data. For many investigators, the identification of such differential gene expression is the main objective of the experiment. The most common statistical test for whether the differential expression is significant is known as "Significance Analysis of Microarrays (SAM)". See http://www-stat.stanford.edu/~tibs/SAM/.

Interpreting the fact that those genes have been differentially expressed consists of (1) determining whether those genes have some functions in common, as determined by a review of the literature and other databases; and (2) seeing whether anyone else has previously obtained a similar result.

Review of the relevant literature is facilitated by the Entrez Gene, PubMed, OMIM, and other NCBI databases mentioned previously. Example: differentially expressed genes in the ginseng experiment above, for example, the gene AMPK (Adenosine 5'monophosphate-activated protein kinase).

Other than literature databases, the database that is most often referred to, when seeking common functions among genes in a list of differentially expressed genes, is **Gene Ontology** (see http://www.geneontology.org/). "Ontology" is a term used by workers in computer science to mean a formal description of the central data types and concepts in a domain of knowledge. It is like a thesaurus, except that it involves hierarchical definitions of increasing generality as well as branches in the definitions. (See WordNet at http://wordnetweb.princeton.edu/perl/webwn for an ontology of common English words -- type in a word like "pharmacy", and look under its inherited hypernym). The Gene Ontology contains terms about genes that involve cellular components, biological processes and molecular functions (see http://www.geneontology.org/GO.doc.shtml). You can manually browse Gene Ontology with AmiGO (go to http://amigo.geneontology.org/cgi-bin/amigo/go.cgi, type in a term like BRCA1 under "genes or proteins", click on associations for Homo sapiens, then study the process, function, and component ontologies).

If you are searching for common functions among genes in a microarray experiment, you can use any one of several dozen software tools that will automatically go through Gene Ontology (or similar database) and attempt to discover whether a list of genes is significantly rich in association for some cellular component, process, or function. The following paper reviews 68 such software tools: Da Wei Huang, Brad T. Sherman and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research (2009)37:1–13. The review concludes that the currently available software tools are useful as

exploratory data-mining, but do not produce results that can be believed without the use of independent judgment and further analysis on the part of the user. As an example of such software, the DAVID tool is available at this web site: http://david.abcc.ncifcrf.gov/home.jsp.

Another useful database is **KEGG**, which describes known pathways.
Go to http://www.genome.ad.jp/kegg
Then, under Table of Contents, click on the "Pathway" link
Then, click on any one of the links there for an example, say, “Pentose Phosphate Pathway”.
Software tools for deciding whether differentially expressed genes lie preferentially along some known pathway (in a database like KEGG or the one at http://www.reactome.org/) are discussed in Werner,Thomas. Bioinformatics applications for pathway analysis of microarray data.Current Opinion in Biotechnology. Volume 19, Issue 1 (2008) pp. 50-54. As described there, usually only a few differentially expressed genes hit pathways in the database, because regulatory networks (not just pathways) are usually involved. The networks are not fixed entities, but change according to their developmental or pathophysiological context, and they cross multiple pathways at relatively few connection points within each pathway.

If the gene in question is an EST without any functional description, it is not possible to interpret the results based on the literature. In fact, your experiment may be the first indication of a function of the gene. To corroborate your data, you can search a database of gene expression, which is a repository mostly of microarray experiments -- **GEO** (Gene Expression Omnibus).
Go to http://www.ncbi.nlm.nih.gov/
Click on Entrez Home
You may be interested in expression of a particular gene. So, in the box “Search across databases” enter the name of the gene then click “go”
Click on the GEO Profiles icon, then view the results for each experiment.
A tool that searches GEO profiles for you is available at http://seq.mc.vanderbilt.edu/exalt/.
Databases similar to GEO are available at http://www.oncomine.org/, http://cibex.nig.ac.jp/index.jsp, http://www.ebi.ac.uk/microarray-as/ae/ and http://www.broad.mit.edu/cmap/.

When groups of genes have similar patterns of gene expression, it is thought that the genes may all have a similar mechanism of regulation. For example, they might all have the same transcription factors. Clustering therefore provides a way to suggest which genes are transcriptionally co-regulated. After the co-regulated genes are identified, the focus of analysis may become the cluster as a whole and its relation to other clusters.
Examples: My patents (U.S. 7,289,911 and 7,400,981) entitled "System, Methods, and Computer Program Product for Analyzing Microarray Data", which automatically goes into the scientific literature about genes in each of the clusters and finds keywords and phrases that distinguish each cluster from all of the other clusters (a search engine). The invention also assesses the extent to which the word/phrase theme is statistically significant, in order to evaluate different methods of clustering. See http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=7289911 and http://patft1.uspto.gov/netacgi/nph-Parser?patentnumber=7400981.

Test your knowledge:

You were given as a handout an article from the journal Nature that describes a project to investigate the genetics of depression. The article describes developing microarrays with spots for ~800 genes.

1a. What type of data will they get from this study?

The study will produce datasets like the one in the GEO database, in which the levels of expression of the 800 genes are measured as a function of genotype and intervention.

1b. How will it be integrated into useful information?

Genes that are differentially expressed as a function of the normal vs. depressed state are potential targets for drug development. Analysis of the pathways along which these genes lie, as well as finding out which genes are co-regulated, can result in a better general understanding of the molecular physiology of the depressed state.

1c. How will therapeutically useful drugs come out of this?

The drugs that cause gene expression patterns of depressed animals to more resemble the gene expression patterns of normal animals are suitable candidates for further investigation. Drugs that cause gene expression patterns characteristic of toxic reactions may be eliminated at an early stage of investigation.

2. What are the tools of bioinformatics and what are their strengths and weaknesses?.

If you are not a molecular biologist doing laboratory work, then the tools that are likely to be useful to you are web-based tools such as those available at the NCBI web site.

If you are a molecular biologist doing laboratory work, you will need some specialized bioinformatics software tools that were not discussed here (for example, tools that help you select restriction enzymes for cloning experiments, tools for designing oligonucleotides, or tools for displaying the three-dimensional structure of proteins). For those applications, investigators may purchase software for desktop computers, such as MacVector. (See http://www.macvector.com/Products/macvector.html). If your research involves microarrays and related technologies, you may need to use software tools like the ones listed at: http://www.g6g-softwaredirectory.com/ListingsByApplication.html. For some applications, you may need custom consulting support that may be provided by an in-house bioinformatics specialist or contracted externally, for example, through GENETWORKS® research and development services (For such consulting support, you can contact me at drigney@genetworks.com).

In general, bioinformatics tools work well when dealing with straightforward questions like finding sequence matches. Their weakness is that for some tasks, they cannot be used without a great deal of expert human judgment. The more the tools need to use artificial intelligence, such as in finding protein motifs or interpreting whole microarray experiments, the less likely they are to give correct or satisfying results.

3. Given a particular sequence, how do you find what it does?

The simplest thing to do is to search GenBank using BLAST to find out if a similar sequence is already there, and if so, find out whether it corresponds to a UniGene entry. If there is a UniGene entry, database links will take you to more information about what the gene does. It is difficult to infer the function of the gene from the sequence alone. It is sometimes possible to locate functional motifs within the sequence as indicated in Pfam or Entrez “Conserved Domains” web pages, for example, whether it is likely to be a kinase or a DNA-binding protein. But this information is usually not very specific about the function. If a related sequence has been used for a microarray spot, it might be possible to search the GEO database for the circumstances under which the gene is up- or down-regulated, and from this information you might be able to infer something about its function.