Lecture 23 - Functional Genomics and Proteomics (AMG text pp. 244-250)

Accessing Molecular Genetic Information through the Internet
The Internet has become a new type of electronic tool for molecular genetic research. The rapid growth in the number of internet users has been due to the introduction of a powerful graphical interface called the world wide web (WWW). The essential WWW software programs developed by the CERN computer scientists are named HTTP (HyperText Transfer Protocol) and HTML (HyperText Markup Language).

The Internet explosion occurred in the early 1990s when two individuals, Tim Berners-Lee at CERN in Switzerland, and Mark Andreessen at the University of Illinois, developed the necessary software and protocols to create what is now called the World Wide Web. Berners-Lee developed HTTP and HTML for physicists to better standardized internet files, whereas, Andreessen developed the first Internet Browser application he called Mosaic. Andreessen later went on to found Netscape Communications.

Internet users can connect to the WWW through a telephone link with an internet service provider (ISP) or through an institutional local area network (LAN).

Dial-up connections to the internet require telephone lines and modem devices to link users with an ISP which provides regional access to the internet through a high speed host computer. Dial-up users require PPP (Point to Point Protocol) software for proper modem communication.

LAN connections use on-site hardware servers to funnel users through ethernet links to organizational internet host computers (or to an ISP). All internet connections require a special communication software called TCP/IP (Transmission Control Protocol/Internet Protocol).

The primary components are high speed servers that are linked to fiber optic cables which make up the backbone of the Internet. We access this backbone through local networks that can interface with personal computers. At the University of Arizona we use campus-wide ethernet connections to computers in the Center for Computing and Information Technology (CCIT), whereas, from home, we connect to these servers through telephone lines and modem devices. The way we connect to the Internet is changing rapidly as the desire for high-speed connectivity increases.

Aside from the use of e-mail and on-line journal publications, the primary research-related activities are;

1) DNA sequence analysis using sequence alignment programs,

2) Searching integrated databases to locate biological resources and reagents, and

3) Interactive communication with specialized biological sciences user groups.
DNA sequence analysis programs and databases
One of the most convenient methods to search public domain databases is to use an internet connection to search GenBank with the Basic Local Alignment Search Tool (BLAST). GenBank is maintained by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine which is located on the NIH campus in Bethesda, Maryland. Information about BLAST is readily available through the NCBI web site.

GenBank is a member of the International Nucleotide Sequence Database Collaboration, an organization that includes the European Molecular Biology Laboratory (EMBL) and the DNA DataBank of Japan (DDBJ). The GenBank, EMBL and DDBJ databases freely exchange DNA sequence files on a daily basis to maintain a comprehensive set of all known sequences.

Try a GenBank search and nucleotide sequence comparison on-line.

1. Highlight and copy the DNA sequence below to use as a test sequence.
2. Go to the GenBank Blast page and paste the sequence into the query window.
3. Search the nr database using the blastn algorithm by clicking on the Submit button.
4. On the new page posted by NCBI, click on the Format Results button and wait for the results.

Try a GenBank search using a protein sequence comparison on-line.

1. Highlight and copy the amino acid sequence called "AMG Test Protein."
2. Go to the GenBank Blast page and paste the sequence into the query window.
3. Search the nr database using the tblastn algorithm by clicking on the Submit button.
4. On the new page posted by NCBI, click on the Format Results button and wait for the results.

The results of BLAST search queries are used to design laboratory experiments that test the biological relevance of DNA and protein sequence homologies. For example, a BLAST search using coding sequences from the hypothetical mouse AMG gene could be done to determine if there are any homologous or orthologous (evolutionarily conserved) gene sequences in GenBank.
Three possible outcomes are illustrated;

(a) AMG identity with a previously cloned mouse cDNA.

(b) Limited, but highly significant homology, with several mouse and human genes.

(c) Yeast gene is found that shares the same region of homology as the mouse and human homologues.

Integrated biological resource databases
One of the most exciting molecular genetic applications of the WWW has been the ability to exploit hypertext file linking functions using strategies that create internet entry points for integrated resource databases. The basic idea of these linked web sites has been to enable researchers to place DNA sequence information into a proper biological context.

One example of an integrated database resource is the GenomeNet project in Japan, which is operated jointly by Kyoto University and the University of Tokyo. GenomeNet is a WWW-accessible network of computational services and databases that were initially developed as a research tool for molecular and cell biological studies in Japan.

Bionet newsgroups for molecular genetic research
So far we have only described ways that molecular genetic researchers use the internet to obtain information, but it is also an important means of communication between individuals (e-mail), and between groups of users that have a common interest (newsgroups or forums). Internet newsgroups function like electronic bulletin boards and have been around since the inception of the internet network.

One of my favorite protein structures is the glucocorticoid receptor bound to DNA. You can also investigate glucocorticoid receptor structure and function using a steroid receptor integrated database put together by researchers at George Washington University.

An example of an Internet company started by graduate students from Stanford University to facilitate lab method reviews (and earn money for themselves through advertising revenue) is BioWire.com which allows free access to a user-updated database. A often cited not-for-profit web site resource run out of Iowas State is Pedro's Biomolecular Research Tools page.

In addition to the ubiquitous use of e-mail to communicate with fellow scientists, researchers also make use of BIOSCI newsgroups called BioNet which can provide a rich resource of information in a specific area.

What do you conclude about the possible function of the "AMG Test Protein" based on the GenBank query results?
What is the primary difference between the "nr" and "dbest" databases? Is the yeast genome sequence included in the "nr" database, or do you have to search the "yeast" database specifically to find orthologous genes?

What is the difference between a "blastn" search and a "tblastx" search? Why can’t you run the "tblastx" algorithm through the NCBI web site interface, how could you run this search?

Which integrated Internet database is used to query protein structure information? Where would you go on the Internet to find text abstracts from published research articles?

What is meant by the team "database mining" in the context of applied molecular genetics?

**Functional Genomics and Proteomics using Mass Spectrometry**

The availability of the entire DNA sequence of an organism presents many opportunities for gene analysis. The term "functional genomics" refers to the study of genome expression, i.e., assigning function to every gene product encoded by the genome. Since RNA represents an intermediate "messenger" in genomic expression, it has limited usefulness - one of which is high throughput "profiling" of genomic expression by microarrays. However, it is really the proteins in a cell that carry out the business of biochemical life. Indeed, gene expression in the functional sense, really refers to protein synthesis and biological control.

The study of an organism's "proteome" is made possible by a new approach to biochemical analysis called "proteomics." Proteomics is the large scale analysis of gene functions at the protein level. The major tool in proteomic research is Mass Spectrometry. The birth of proteomics can be traced back to at least the 1970s when Pat O'Farrell developed the use of two-dimensional protein gels to analyze thousands of proteins at a time in cell extracts. Modern day proteomics combines this "proteome" approach with sensitive instrumentation in the form of mass spectrometry. The key to proteomics is the use of powerful computer algorithms that search genomic databases to match protein mass data with theoretical protein masses. This strategy of using genomic databases as a tool to understand gene function is a major component of bioinformatics.

Mass spectrometry measures the mass of particles by several different methods based on determining the mass to charge ratio. Mass spectrometry requires charged gaseous molecules for analysis. Up until the 1980s mass spectrometry had few applications in the life sciences because native proteins and peptides are large polar molecules that are not easily transferred to gas phase and ionized. Two techniques developed in the late 1980s changed all of this by providing a way to ionize proteins that could then be analyzed by mass spectrometry.

The first method was developed by Karas and Hillenkamp and is called MALDI for matrix-assisted laser desorption ionization. MALDI is based on coprecipitating peptides with matrix material that is then dried onto a metal surface and irradiated with nanosecond laser pulses to ionize the peptides and create a gas phase. The precise
mechanism of MALDI-based ionization is not known. MALDI is often coupled to another
technique called time of flight (TOF) analysis which is how the charge to mass ratio is
determined (MALDI-TOF).

The second method is called electrospray mass spectrometry or electrospray ionization
(ESI) which was first described in 1989 by Fenn and coworkers. ESI utilizes a mixture of
peptides in acid (MeOH or acetonitrile) that is pumped through a hypodermic needle at
high voltage to electrostatically disperse small droplets which evaporate and impart
charge onto the peptide. This gaseous material contains charged peptides that are suitable
for mass spectrometry analysis.

Mass spectrometry has three primary roles in proteomic applications
- Quality control for protein production in biotechnology
- Protein identification in biochemical research
- Analysis of post-translational modification of proteins

The use of analytical tools such as mass spectrometry to investigate proteome dynamics
has benefited greatly by the availability of the complete sequence of organismal genomes.
The reason for this is that bioinformatics can then predict the entire set of proteins (or at
least hypothetical proteins) that an organism is capable of expressing, and this set is used
to interpret mass spectrometry data.

Two main applications of proteomics to functional genomics are:

Expression Proteomics - identifying proteins that are up- or down-regulated (2D gels).
Functional Proteomics - identifying proteins that are part of large complexes (Co-IP expts.)

Let's look at one of the most powerful applications of mass spectrometry in proteomics,
the identification of co-immunoprecipitated proteins using epitope-tagged "bait" proteins, a
strategy that supersedes the yeast two-hybrid approach. The ability to identify proteins of
very low abundance from protein gel slices (5-50ng of a 50 kDa protein or 0.1 - 1 pmol)
makes the co-immunoprecipitation method more reliable for in vivo interactions since it
does not depend on fusion protein functions in yeast cells.

Cleavage of proteins into peptides using trypsin is the first step in protein identification by
mass spectrometry. Trypsin cleaves polypeptides on the carboxy terminal side of Lysine
(K) or Arginine (R) residues. Therefore, analysis of peptides by mass spectrometry
includes the tentative assignment of the carboxyl terminal residue as a lysine or arginine
with a known mass.

In tandem mass spectrometry (MS/MS), a mixture of peptides is separated out into
individual peptides that are energized and susceptible to breakage at peptide bonds. The
mass of the resulting peptide fragments are then determined and the protein sequence is
deduced by calculating the masses of fragments differing by one amino acid. Tryptic
fragments derived from the C-terminal end contain a Lys or Arg residue and ar called the
"y" ions, fragments with a common N-terminal are called the "b" ions. Some amino acids
cannot be resolved because their masses are very similar, for example, the mass of isoleucine is exactly the same as the mass of leucine.

Importantly, protein sequence determination by mass spectrometry requires comparison of the determined masses to predicted peptides in genomic databases. Three different ways to identify proteins by mass spectrometry using Genomic Database queries:

What is a proteome? What is the meant by the term proteomics? What are two key applications of proteomics to investigating genomic functions?

What has contributed to the application of mass spectrometry to proteomics. i.e., what are the major breakthroughs that permitted the field of proteomics to expand so rapidly?

Explain how random fragmentation at peptide bonds by tandem mass spectrometry can be used to predict the sequence of a peptide? How can the spectral peaks originating from b ions and y ions be distinguished, is this known a priori?

What is the "molecular genetic" strategy that makes it possible to use proteomic methods to identify proteins that associate with each other inside cells? Why might this be better than the yeast two hybrid method?

Describe an experiment in which proteomic methods could be used to identify proteins that are differentially expressed in response to hormone stimulation of a liver cell line. How would the proteins be resolved and how would differential patterns be identified?

Explain how proteomic methods could be used to identify proteins that are phosphorylated only in the presence of hormone? What are the limitations of this "whole proteome" approach with regard to level of sensitivity?