Appendix

Table of Contents

BRCA1 and *BRCA2* Background Information

BRCA1, BRCA2 and the Risk of Cancer

The names *BRCA1* and *BRCA2* stand for **br**east **ca**ncer susceptibility gene **1** and **br**east **ca**ncer susceptibility gene **2**, respectively. The BRCA1 (sometimes pronounced BRA-kah 1) and BRCA2 (sometimes pronounced BRA-kah 2) proteins play vital roles in genomic stability and can act as tumor suppressors in both men and women. A tumor suppressor is a gene that normally prevents cancer. Mutations in these genes can lead to cancer when the normal function is lost. Together, mutations in these genes account for 5-10% of all breast cancer cases and approximately 45% of all familial [inherited] breast cancer.

By convention, the names of genes are usually italicized, while the names of their encoded proteins are not. Therefore, the gene is written as *BRCA1*, while the protein is simply "BRCA1." Mutations in *BRCA1* or *BRCA2* can result in amino acid changes or changes in the mRNA reading frame that lead to shorter proteins. Some of these changes dramatically increase the risks of breast and ovarian cancer. The lifetime risk of breast cancer for the average woman is 12%, and the lifetime risk of ovarian cancer is 2%. Many factors such as excess weight, lack of exercise, having a first period at a young age, and not having children can increase the risk of breast cancer in all women. Increased weight and lack of exercise are associated with increased estrogen, which can promote cancer. Menstruation is associated with physiological changes in the breast that are conducive to the development of cancer, and beginning menstruation at a young age allows more total time for cancer to develop; some of these changes are mitigated by pregnancy and breastfeeding. Certain mutations in *BRCA1* or *BRCA2* can increase these risks to 36-85% for breast cancer and 20-60% for ovarian cancer. It is important to note that *BRCA1* and *BRCA2* mutations also confer increased risk of breast and prostate cancer in men. All carriers are also at increased risk for other types of cancer including pancreatic, laryngeal, and stomach cancers, as well as melanoma.

The risk of breast and ovarian cancer associated with *BRCA1* and *BRCA2* alleles containing cancer-causing mutations is inherited in an autosomal dominant fashion, because only a single defective copy must be passed from parent to offspring for the offspring to inherit the cancer risk. However, both copies of the gene must be mutated for cancer to develop, making this BRCA-associated cancer autosomal recessive. See the next *Appendix* section, "*BRCA1*: Is it Dominant or Recessive?" for more information. Cancer is thought to develop In a person with one functional copy of a BRCA gene and one mutant (non-functional) copy when a new mutation occurs that deactivates the functional copy of the gene. The need for a second mutation could explain why some people don't get breast or ovarian cancer, even when they have mutant copies of the *BRCA1* or *BRCA2* genes. This phenomenon is known as "incomplete penetrance." In some cases, the functional copy of the gene is lost when a tumor develops, while the germline-encoded mutant copy is retained. Inheritance of two mutated copies of *BRCA2* is associated with another kind of genetic disease called "Fanconi anemia." Inheritance of two mutated copies of *BRCA1* is lethal to an embryo.

Cellular Functions of BRCA1 and BRCA2

Both BRCA1 and BRCA2 play crucial roles in genomic stability – making sure that DNA remains intact. In particular, they participate in a biochemical pathway for repairing breaks in double-stranded DNA. They also have a number of other functions. They act as transcriptional coactivators and ubiquitin-protein ligases, and they bind to zinc and tubulin. Many of these seemingly unrelated abilities may contribute to their role in repairing DNA.

If DNA is damaged, the cell must repair the DNA before proceeding through the cell cycle. BRCA1 and BRCA2 both interact with other proteins that respond to DNA damage and participate in homologous recombination. Cells respond to damaged DNA by making proteins to repair the damage and by stopping cell division. BRCA1 might put the brakes on cell division by binding to other proteins that act as tumor suppressors, DNA damage sensors, or signaling proteins. BRCA1 also binds to RNA polymerase II and stops transcription by interacting with a protein that chemically modifies histones. Transcription often involves modifying histones so that the chromosomes decondense or "loosen up" and allow RNA polymerase to access the target genes. The BRCA2 protein contains several copies of a 70 amino acid sequence called the BRC motif, which functions by binding to a DNA repair protein called RAD51.

BRCA1

Gene and Protein Structures

BRCA1 is encoded by 24 exons and is located on the long arm of chromosome 17. Alternative splicing occurs frequently with *BRCA1*-encoded mRNA and many forms of alternatively spliced mRNAs have been isolated. Some alternatively-spliced forms are associated with disease-causing mutations, including frameshifts and premature truncations. A related pseudogene, which is also located on chromosome 17, has been identified. (Pseudogenes are genes that do not encode functional proteins. This is usually because they contain stop codons that prevent the translation of a functional protein.)

The BRCA1 protein contains three types of domains: a RING finger domain near the beginning of the protein (the N-terminus), Nuclear Localization Signals in the middle of the protein, and two BRCT domains at the end of the protein (see ***Figure 1***). BCRT stands for "BRCA1 C-Terminal." The "RING" in RING finger stands for "Really Interesting New Gene." These protein domains are characterized by an amino acid sequence motif containing cysteines and histidines (Cys3HisCys4) that binds to DNA, RNA, and protein or lipid substrates through interactions with zinc cations. The BRCA1 RING finger is thought to facilitate binding of the protein to DNA. The Nuclear Localization Signals are kind of like an address on a piece of mail. They tell the cell to send the BRCA1 protein to the nucleus. BRCT domains consist of repeated sequences of 90-100 amino acids each that bind to other proteins, including other molecules of BRCA1. The BRCT domains are found in multiple proteins that participate in cell cycle regulation and DNA repair, including DNA ligase III. BRCT domains adopt a characteristic parallel four-stranded beta sheet, with two to three alpha helices packed against one face and a single alpha helix packed against the opposite face of the sheet. In BRCA1, the two BRCT repeats interact in a head-to-tail manner and facilitate the interaction of BRCA1 with protein partners such as p53.

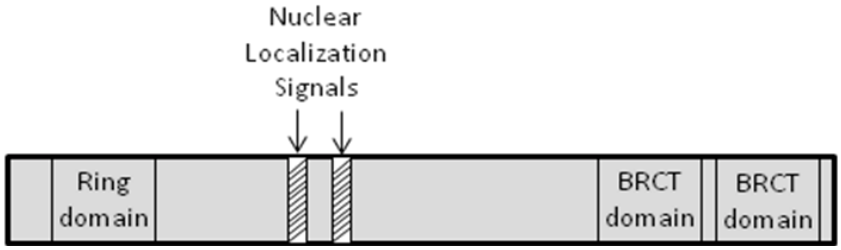


Figure 1: BRCA1 Protein Domains.

BRCA2

BRCA2 is encoded by 27 exons and is located on the long arm of chromosome 13.Common *BRCA2* mutations associated with cancer include small insertions and deletions, which can result in frameshifts that create defective proteins. The N-terminal region (exon 3) of BRCA2 has been shown to function in transcription of genes involved in DNA repair, the cell cycle, and programmed cell death (apoptosis) via RNA polymerase II. As with BRCA1, Nuclear Localization Signals in BRCA2 target the protein to the nucleus, where it interacts with RAD51 via a series of BRC repeat domains (see ***Figure 2***).

Figure 2: BRCA2 Protein Domains.

Options for Those Who Test Positive for Cancer-Associated Mutations

There are three main options for those who test positive for cancer-associated mutations in either *BRCA1* or *BRCA2*.

The first option is surgery. A prophylactic oophrectomy – removal of both ovaries before cancer can strike – can reduce the risk of breast cancer by 50% due to the subsequent reduction in estrogen production, and reduces the risk of ovarian cancer by 95%. Because minute amounts of tissue may remain, the risk of cancer is not completely eliminated. Female carriers of cancer-associated mutations in *BRCA1* or *BRCA2* are encouraged to have an oophrectomy by age 40. A prophylactic mastectomy – removal of both breasts before cancer can strike – reduces the risk of breast cancer by 90%. As with the oophrectomy, there is a possibility that some residual tissue or cancer cells might be left behind in the chest wall. As a consequence, the risk of breast cancer can never be completely eliminated.

The second option is chemoprevention (treatment with drugs). Because estrogen can promote breast cancer, drugs that block estrogen like tamoxifen or raloxifene can help reduce the risk of cancer. They also cause temporary, reversible menopause, with all its side effects, including hot flashes and disrupted ovulation.

The third option is to test frequently so that breast cancer can be found at an early stage. This includes increasing the frequency of mammograms to twice yearly; increased breast exams, breast MRIs, and blood tests; and at least yearly Pap smears, pelvic exams, and vaginal ultrasounds to screen for ovarian cancer. Unfortunately, screening for ovarian cancer is not very effective, and mortality rates for this form of cancer are quite high: the average 5-year survival rate is only 46%.

For more information about *BRCA1*, *BRCA2* and genetic testing, see the "National Cancer Institute Fact Sheet: *BRCA1* and *BRCA2*: Cancer Risk and Genetic Testing" at http://www.cancer.gov/cancertopics/factsheet/risk/brca and see the *Sources* below.

Sources

American Cancer Society (2009, May 13). *Detailed Guide: Breast Cancer in Men*. Retrieved August 6, 2009 from: http://www.cancer.org.

Couch, F.J., DeShano, M.L., Blackwood, M.A., et al. (1997). *BRCA1* mutations in women attending clinics that evaluate the risk of breast cancer [see comments]. *N Engl J Med*., 336: 1409–15.

Easton, D.F., Steele, L., Fields, P., et al. (1997). Cancer risks in two large breast cancer families linked to *BRCA2* on chromosome 13q12–13. *Am J Hum Genet*., 61: 120–8.

Ford, D., Easton, D.F., Stratton, M., et al. (1998). Genetic heterogeneity and penetrance analysis of the *BRCA1* and *BRCA2* genes in breast cancer families. *Am J Hum Genet*., 62: 676–89.

Harmon, A. (2007, September 16). Cancer free at 33, but weighing a mastectomy. *The New York Times*. Retrieved from: http://www.nytimes.com/2007/09/16/health/16gene.html.

National Cancer Institute. (2009, June 29). *Genetics of Breast and Ovarian Cancer*. Retrieved August 4, 2009 from: http://www.cancer.gov/cancertopics/pdq/genetics/breast-and-ovarian.

National Cancer Institute. (2006, July 27). *National Cancer Institute Fact Sheet: Preventive Mastectomy*. Retrieved August 4, 2009 from: http://www.cancer.gov/cancertopics/factsheet/Therapy/preventive-mastectomy/.

Shiozaki, E.N. et al. (2004). *Mol. Cell*., 14(3), 405-412. PDB: 1T29.

Struewing, J.P., Hartge, P., Wacholder, S., et al. (1997). The risk of cancer associated with specific mutations of *BRCA1* and *BRCA2* among Ashkenazi Jews. *N. Engl. J. Med*., 336: 1401–8.

BRCA1: Is it Dominant or Recessive?

Both BRCA1 and BRCA2 proteins are involved in DNA synthesis and repair of DNA breaks, helping to maintain the integrity of the genome. But is the inheritance of cancer risk associated with mutations in the *BRCA1* or *BRCA2* genes dominant or recessive?

A genetic trait is considered **dominant** if it is expressed in a person who has only one copy of that allele.

A genetic trait is considered **recessive** if it is expressed only when two copies of the allele are present.

Unfortunately, the classical Mendelian terms of "dominant" and "recessive" don't apply well in the case of BRCA1 and BRCA2.

Many *BRCA1* and *BRCA2* resources say that cancer-predisposing alleles of the *BRCA1* and *BRCA2* genes are ***dominant***. If an individual inherits a single copy of a mutated *BRCA1* or *BRCA2* gene, that individual has inherited an ***increased risk of cancer***.

However, *BRCA1* and *BRCA2* function as tumor suppressors, making them functionally ***recessive***. Both copies must be mutated in a cell and made non-functional for cancer to develop. However, inheritance of two mutated copies of *BRCA1* or *BRCA2* is lethal to an embryo. In a pedigree, *BRCA1* and *BRCA2* ***appear to be autosomal dominant***, because only one mutated allele is inherited. The second copy is mutated later in life.

This is often understood in terms of Knudson's Hypothesis or the "Two-Hit Hypothesis."

Let's consider that case of *BRCA1*.

A "normal cell" (i.e., non-cancerous) contains two functional copies of the *BRCA1* gene. Neither allele contains mutations associated with cancer. Functional BRCA1 proteins are made from both alleles, and help repair DNA damage as it occurs.

An individual who has inherited a germline *BRCA1* mutation in one allele (i.e., one "hit") has a much higher lifetime risk of developing cancer than an individual born with two functional copies of the gene. Every cell in the body contains this mutated allele. Because the non-mutated *BRCA1* allele still encodes a functional protein, the cell is able to repair DNA damage and remains non-cancerous.

If DNA damage or an error in DNA replication causes the second copy of the *BRCA1* gene to become mutated in any given cell (i.e., a second "hit"), that cell can no longer make any functional BRCA1 protein. Researchers believe that a defective or missing BRCA1 protein is unable to help repair damaged DNA or fix mutations that occur in other genes. As these defects accumulate, they can allow cells to grow and divide uncontrollably and form a tumor.

Objective

Clear up any confusion about whether *BRCA1* and *BRCA2* mutations are dominant or recessive.

In genetic terms, a **dominant trait** is one that is phenotypically expressed in heterozygotes.

In genetic terms, a **recessive trait** is one that is phenotypically expressed only in homozygotes.

It is possible that a mutation in one copy of the *BRCA1* (or *BRCA2*) gene makes it more likely that an individual will eventually develop a mutation in the second copy of the gene.

Over 1600 different mutations have been identified in *BRCA1* and over 1800 have been found in *BRCA2*. Many families have their own type of mutation that stays within the family. There are differences in cancer risk associated with different mutations, but we don't know enough about these genes yet to fully understand why this is. About a third of mutations identified in either the B*RCA1* or *BRCA2* gene so far are of uncertain clinical significance.

Additional Resources:

The Bio-ITEST Program has developed a two-part animation to highlight the normal function of *BRCA1* in the cell, and how mutations in the *BRCA1* gene can lead to cancer. The animation can be found under the Resources tab on the Bio-ITEST Genetic Testing web page at: http://www.nwabr.org/curriculum/introductory-bioinformatics-genetic-testing.

National Cancer Institute Fact Sheet, "*BRCA1* and *BRCA2*: Cancer Risk and Genetic Testing." http://www.cancer.gov/cancertopics/factsheet/risk/brca.

Genetics Home Reference: *BRCA1*. http://ghr.nlm.nih.gov/gene=brca1.

The Breast Cancer Information Core: An Open Access Online Breast Cancer Mutation Database through the National Human Genome Research Institute: http://research.nhgri.nih.gov/bic/.

GINA

The Genetic Information Nondiscrimination Act of 2008
Information for Researchers and Health Care Professionals
April 6, 2009

The information presented in this fact sheet is intended for general informational purposes only. While this fact sheet does not cover all of the specifics of GINA, it does provide an explanation of the statute to assist those involved in clinical research to understand the law and its prohibitions related to discrimination in health coverage and employment based on genetic information. The information should not be considered legal advice. In addition, some of the provisions discussed involve issues for which the rules have not yet been finalized, and this information is subject to revision based on publication of regulations.

What is GINA?

The Genetic Information Nondiscrimination Act of 2008 (P.L. 110-233, 122 Stat. 881)[1], also referred to as GINA, is a new Federal law that prohibits discrimination in health coverage and employment based on genetic information. The President signed the act into law on May 21, 2008. The section of the law relating to health coverage (Title I) generally will take effect between May 22, 2009, and May 21, 2010.[2] The sections relating to employment (Title II) will take effect on November 21, 2009. GINA requires regulations pertaining to both titles to be completed by May 2009.

[1] http://frwebgate.access.gpo.gov/cgibin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ233.110.pdf.

[2] The effective date of the insurance provisions is not the same in all cases because for group health plans, Title I will take effect at the start of the "plan year" beginning one year after GINA's enactment. Because some health plans do not designate their "plan years" to correspond to a calendar year, there will be variation among plans as to when Title I takes effect for the plans. However, for individual health insurers, GINA will take effect May 22, 2009.

How does the Federal law affect state laws?

GINA provides a baseline level of protection against genetic discrimination for all Americans. Many states already have laws that protect against genetic discrimination in health insurance and employment situations. However, the degree of protection they provide varies widely, and while most provisions are less protective than GINA, some are more protective. All entities that are subject to GINA must, at a minimum, comply with all applicable GINA requirements, and may also need to comply with more protective State laws.

What will GINA do?

GINA generally will prohibit discrimination in health coverage and employment on the basis of genetic information. GINA, together with already existing nondiscrimination provisions of the Health Insurance Portability and Accountability Act, generally prohibits health insurers or health plan administrators from requesting or requiring genetic information of an individual or the individual's family members, or using it for decisions regarding coverage, rates, or preexisting conditions. The law also prohibits most employers from using genetic information for hiring, firing, or promotion decisions, and for any decisions regarding terms of employment.

The statute defines 'genetic information' as information about:

- an individual's genetic tests (including genetic tests done as part of a research study);
- genetic tests of the individual's family members (defined as dependents and up to and including 4th degree relatives);
- genetic tests of any fetus of an individual or family member who is a pregnant woman, and genetic tests of any embryo legally held by an individual or family member utilizing assisted reproductive technology;
- the manifestation of a disease or disorder in family members (family history);
- any request for, or receipt of, genetic services or participation in clinical research that includes genetic services (genetic testing, counseling, or education) by an individual or family member.

Genetic information does not include information about the sex or age of any individual.

The statute defines 'genetic test' as an analysis of human DNA, RNA, chromosomes, proteins, or metabolites that detects genotypes, mutations, or chromosomal changes. The results of routine tests that do not measure DNA, RNA, or chromosomal changes, such as complete blood counts, cholesterol tests, and liver-function tests, are not protected under GINA. Also, under GINA, genetic tests do not include analyses of proteins or metabolites that are directly related to a manifested disease, disorder, or pathological condition that could reasonably be detected by a health care professional with appropriate training and expertise in the field of medicine involved.

How will the law be enforced and what are the penalties for violation of the law?

The law will be enforced by various Federal agencies. The Department of Labor, the Department of the Treasury, and the Department of Health and Human Services are responsible for Title I of GINA, and the Equal Employment Opportunity Commission (EEOC) is responsible for Title II of GINA. Remedies for violations include corrective action and monetary penalties. Under Title II of GINA, individuals may also have the right to pursue private litigation.

What won't GINA do?

- GINA's health coverage non-discrimination protections do not extend to life insurance, disability insurance and long-term care insurance.
- GINA does not mandate coverage for any particular test or treatment.
- GINA's employment provisions generally do not apply to employers with fewer than 15 employees.
- For health coverage provided by a health insurer to individuals, GINA does not prohibit the health insurer from determining eligibility or premium rates for an individual based on the manifestation of a disease or disorder in that individual. For employment-based coverage provided by group health plans, GINA permits the overall premium rate for an employer to be increased because of the manifestation of a disease or disorder of an individual enrolled in the plan, but the manifested disease

or disorder of one individual cannot be used as genetic information about other group members to further increase the premium.

- GINA does not prohibit health insurers or health plan administrators from obtaining and using genetic test results in making health insurance payment determinations.

What is the status of regulations to implement GINA?

The law requires regulations by May 2009. The Department of Health and Human Services (Centers for Medicare & Medicaid Services (CMS) and the Office for Civil Rights), the Department of Labor, the Department of the Treasury (the Internal Revenue Service), and the EEOC are currently working on the regulations. The Department of Labor, the Department of the Treasury, and CMS put forth a Request for Information about issues relevant to some of the health coverage provisions in Title I on October 10, 2008, which closed on December 9, 2008.[3]

[3] http://www.dol.gov/federalregister/PdfDisplay.aspx?DocId=21604.

Is GINA retroactive?

GINA will not be retroactive, i.e., it cannot apply to acts or omissions that occurred prior to GINA's effective dates. However, once GINA takes effect, it will prohibit certain uses of genetic information in connection with health coverage and employment, no matter when the information was collected. For example, a health insurer that has been collecting or using genetic information for underwriting would need to change its business practices once GINA takes effect. Likewise, certain employers requiring genetic tests or family history information from employees or prospective employees will no longer be able to do so after GINA takes effect and will be prohibited from discriminating based on any genetic information that they had already collected.

Does GINA have specific research provisions?

Yes. GINA's prohibitions apply to 'genetic information' which is defined as including receipt of genetic services (genetic tests, genetic counseling, or genetic education) by an individual or family member participating in clinical research. There is, however, a research exception.

GINA provides a specific "research exception" to allow health insurers or group health plans engaged in research to request (but not require) that an individual undergo a genetic test. This exception permits the request to be made but imposes the following requirements:

1. the request must be made pursuant to research that complies with HHS regulations at 45 CFR part 46, or equivalent Federal regulations, and any applicable state or local laws for the protection of human subjects in research;
2. there must be clear indication that participation is voluntary and that non-compliance has no effect on enrollment or premiums or contribution amounts;
3. no genetic information collected or acquired as part of the research may be used for underwriting purposes;

4. the health insurer or group health plan must notify the Federal government in writing that it is conducting activities pursuant to this research exception and provide a description of the activities conducted; and
5. the health insurer or group health plan must comply with any future conditions that the Federal government may require for activities conducted under this research exception.

What information about GINA should be communicated as part of the informed consent process to individuals participating in a research study or those considering study participation?

Although GINA has not yet taken effect, there may currently be situations where it is appropriate for researchers to discuss the provisions of the law with individuals participating in a research study or those considering study participation. For more information, see the following guidance document prepared by the Office for Human Research Protections: http://www.hhs.gov/ohrp/humansubjects/guidance/gina.html (URL), http://www.hhs.gov/ohrp/humansubjects/guidance/gina.pdf (PDF).

Courtesy: National Human Genome Research Institute. http://www.genome.gov.

This document is available from the National Human Genome Research Institute at: http://www.genome.gov/Pages/PolicyEthics/GeneticDiscrimination/GINAInfoDoc.pdf .

Ethics Background

Principles: Respect, Maximize Benefits/Minimize Harms, and Justice

Summary

The focus of this perspective is on the four ***principles*** supported by or compromised by the question or issue at hand.

Philosophers Tom Beauchamp and Jim Childress identify four principles that form a commonly held set of pillars for moral life:

Respect for Persons/Autonomy	Value the worth and dignity of each individual. Acknowledge a person's right to make choices, to hold views, and to take actions based on personal values and beliefs.
Maximize Benefits	Provide benefits to persons and contribute to their welfare. Refers to an action done for the benefit of others. Also known as ***beneficence***.
Minimize Harms	Obligation not to inflict harm intentionally; in medical ethics, the physician's guiding maxim is "First, do no harm." Also known as ***nonmaleficence***.
Justice	Treat others equitably, distribute benefits/burdens fairly.

Contributions

- Draws on principles or pillars that are a part of American life—familiar to most people, although not by their philosophical terms.
- Compatible with both outcome-based and duty-based theories (respect for persons and justice are duty-based, while minimizing harms and maximizing benefits are outcome-based).
- Provides useful and fairly specific action guidelines.
- Offers an approach that is appropriate for general bioethics and clinical ethics.
- Requires weighing and balance—flexible, responsive to particular situations.

Challenges

- Lacks a unifying moral theory that ties the principles together to provide guidelines.
- Principles can conflict and the theory provides no decision-making procedure to resolve these conflicts.
- Difficult to weigh and balance various principles.
- Autonomy in some cultures refers to individual autonomy, while in others it refers to group/family/community autonomy.

Additional Information

Additional information about ethical theories and perspectives can be found in *An Ethics Primer: Lesson Ideas and Ethics Background* by Jeanne Ting Chowning and Paula Fraser, produced through the Northwest Association for Biomedical Research. The complete Ethics Primer is available free for download from http://www.NWABR.org.

Creating Discussion Ground Rules

Introduction

The study of ethics involves consideration of conflicting moral choices and dilemmas about which reasonable people may disagree. Since a wide range of positions is likely to be found among students in most classrooms, it is especially important to foster a safe classroom atmosphere by creating some discussion ground rules. These ground rules are often referred to as "norms." An agreed-upon set of ground rules should be in place before beginning the *Using Bioinformatics* curriculum.

Procedure

Ask the students, "What can we do to make this a safe and comfortable group for discussing issues that might be controversial or difficult? What ground rules should we set up?" Allow students some quiet reflection time, and then gather ideas from the group in a brainstorming session. One method is to ask that students generate a list of ground rules in small groups and then ask each group to share one rule until all have been listed. Clarify and consolidate the ground rules as necessary.

Post norms where they can be seen by all and revisit them often. If a discussion gets overly contentious at any time, it is helpful to stop and refer to the ground rules as a class to assess whether they have been upheld.

Some possible student ground rules/norms could include:

- A bioethics discussion is not a competition or a debate with a winner and a loser.
- Everyone will respect the different viewpoints expressed.
- If conflicts arise during discussion, they must be resolved in a manner that retains everyone's dignity.
- Everyone has an equal voice.
- Interruptions are not allowed and no one person is allowed to dominate the discussion.
- All are responsible for following and enforcing the rules.
- Critique ideas, not people.
- Assume good intent.

Objective

Students will be able to:

- Create and agree to classroom discussion norms

Amino Acid
Abbreviations and Chemistry Resources

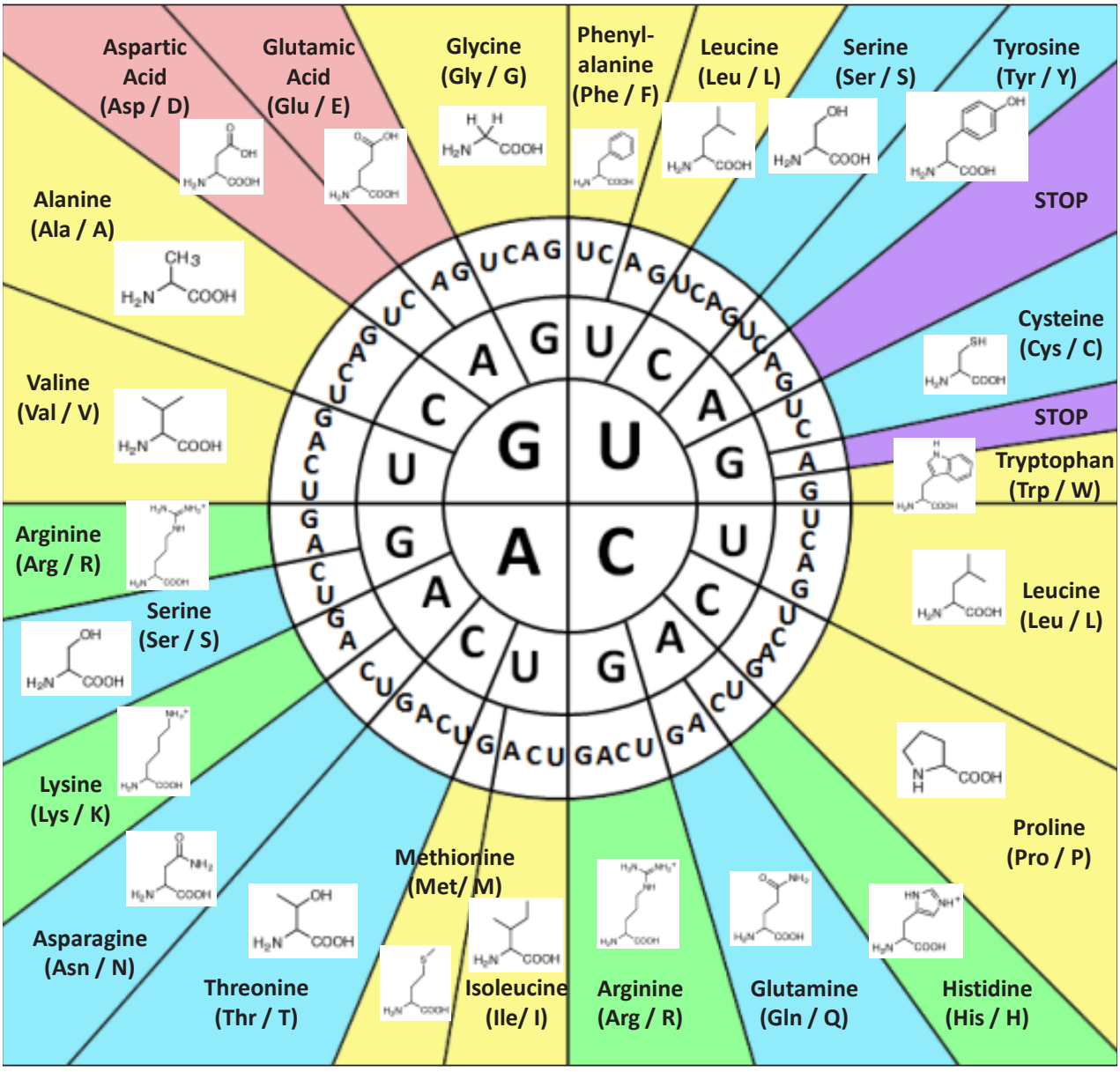
Single-letter Amino Acid Abbreviations

A – Alanine
C – Cysteine
D – Aspartic Acid
E – Glutamic Acid
F – Phenylalanine
G – Glycine
H – Histidine
I – Isoleucine
K – Lysine
L – Leucine
M – Methionine
N – Asparagine
P – Proline
Q – Glutamine
R – Arginine
S – Serine
T – Threonine
V – Valine
W – Tryptophan
Y – Tyrosine

Amino Acid Abbreviations and Categorization by Chemistry

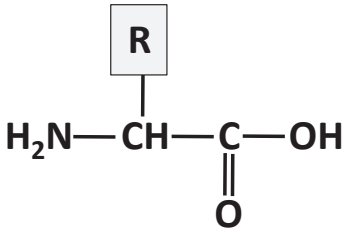
	Uncharged	Positive	Negative
Hydrophilic	Asparagine (Asn – N) Cysteine (Cys – C) Glutamine (Gln – Q) Serine (Ser – S) Threonine (Thr – T)	Arginine (Arg – R) Histidine (His – H) Lysine (Lys – K)	Aspartic Acid (Asp – D) Glutamic Acid (Glu – E)
Hydrophobic	Alanine (Ala – A) Glycine (Gly – G) Isoleucine (Iso – I) Leucine (Leu – L) Methionine (Met – M) Phenylalanine (Phe – F) Proline (Pro – P) Tryptophan (Trp – W) Tyrosine (Tyr – Y) Valine (Val – V)		

Codons and Amino Acid Chemistry



Amino Acid **Side Chain (R-Group)**:

Amino Acid **Backbone**:



Side Chain (R-Group) Chemistry:

- Hydrophobic / Nonpolar
- Hydrophilic/ Polar
- Acidic / Negative
- Basic / Positive

Behind the Scenes with the NCBI Databases and the *Entrez* Search Engine

We have already discussed the similarity between the NCBI databases and iTunes®. Now, we're going to go a little bit farther and consider what happens when data are submitted to NCBI and when we use Entrez to do a database search.

When researchers submit data to the NCBI, they do so by filling in a form from the NCBI website. The sections in the form where information gets entered are called "fields." Different data types have different kinds of fields. For example, the nucleotide database (GenBank) has fields for the gene name, organism, sequence length, and other information related to DNA or RNA sequences. The taxonomy database entry form includes fields for information about the common name, the scientific name, and the rank. Field names are used to help organize and find information.

Entrez is the software system that searches NCBI databases. When you type terms into the NCBI search box, Entrez takes those terms and searches all the fields, in all the database records, to see if those terms can be found. Sometimes this can lead to some puzzling results. For example, searching the nucleotide database with the word "lion" returns several records that come from *Sus scrofa*. *Sus scrofa* is the scientific name for "pig." While some lions might act like pigs, their DNA sequences should be different.

To solve this mystery, we can select the link to one of the *Sus scrofa* records and look at the results. If we search the record for the word "lion," we see that the journal is published from an address at "Lion Mountain 1" street.

What if we were searching for something from lions but instead found thousands of records from pigs? What could we do to improve our results?

We can get ideas by looking at the way Entrez did the search. Selecting the "Details" tab from our search results shows us that Entrez searched the "Organism" field with the scientific name for lion (*Panthera leo*) and Entrez searched all the fields with the word "lion."

Figure 1: Select the "Details" tab.

Consequently, our results included all the records where *Panthera leo* could be found in the organism field plus all the records that included "lion" anywhere in the record. We can use this information to help guide our quest for more specific results.

Discussion: What do you think would happen if you used "*Panthera leo*" [Organism] as a query instead of "lion?"

You can do the experiment and find out.

Understanding BLAST

BLAST stands for Basic Local Alignment Search Tool. An alignment is a way of lining sequences up in rows so they're easier to compare. A local alignment is one where short regions of the sequence are aligned preferentially over long regions.

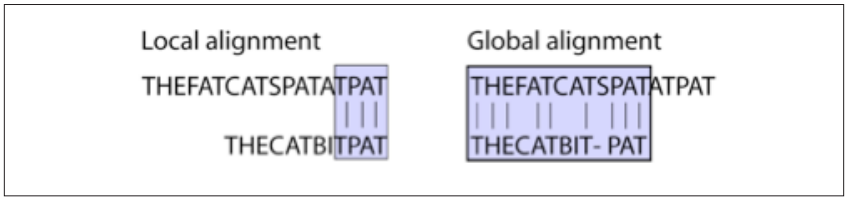


Figure 1: Local vs. global alignments.

Although the name "BLAST" sounds like one program, BLAST is really a family of programs that are widely used by biologists all over the world to compare sequences from DNA, RNA, and proteins. Nucleotide blast (blastn) is used to compare nucleotide sequences. Protein blast (blastp) is used to compare protein sequences. Other kinds of blast programs add a step where nucleotide sequences are translated to protein sequences before searching. Blastx, for example, translates a nucleotide query in all six reading frames and compares the predicted amino acid sequences to a protein database. Tblast compares a protein query to a translated nucleotide database; and tblastx translates both a nucleotide query sequence and the nucleotide database sequences before doing a comparison.

How does BLAST work?

BLAST begins the process of comparing sequences and aligning matching regions by breaking sequences into shorter strings of text, called "words." A typical "word" might be 11 bases or amino acids long with each base or amino acid represented by a single letter in the word. BLAST creates words for both the query sequence (the one we're testing) and all the sequences in a database. Then, every word from the query sequence is compared to every word in the database until words are found that match perfectly.

Once BLAST has found a word from the query that matches a database word, the program evaluates the letters at the end of each word to determine whether the matching region can be extended. This process continues until the end of the sequence is reached or the sequences no longer match.

BLAST scores and statistics

When the BLAST programs were first written in 1990, their major function was to determine whether two sequences were similar enough to make it likely they evolved from a common ancestor. Since the original goal for BLAST was to find matching sequences and measure the significance of the match, BLAST provides many statistics for each search and assigns different scores that can be used to evaluate the results.

BLAST scores from protein comparisons are based on evolution. If a mutation occurs in a nucleic acid sequence that changes an amino acid, the altered protein experiences natural selection. If the change has a beneficial or neutral effect, the change can persist and be inherited. If an amino acid change is harmful, negative selection will make it less likely to persist in a population. In general, amino acid replacements are tolerated better when the new amino acid is either chemically similar or located in a less important part of a protein.

When researchers wrote the scoring system for BLAST, they looked at all the changes that took place between amino acid sequences from the same protein in different organisms and used that data to calculate probability values for each possible amino acid replacement. A BLAST score for a pair of two protein sequences is calculated by looking at each position, finding the likelihood for each position that one amino acid will be replaced by another, and adding those values together. For example, say we had one protein sequence like "ELVIS" and another like "ELVES." We would look at the BLAST scoring table to find the probability of E replacing E is 5, the value for L replacing L is 4, for V replacing V is 4, for E replacing I is -3, and for S replacing S is 4. We add these together: 5 + 4 + 4 + -3 + 4 and get a BLAST score of 14. For nucleotide sequences, BLAST calculates a score based on identity. BLAST assigns two points for each position where a pair of nucleotides matches and subtracts points for each position where they do not. Once BLAST has calculated a score, the program applies corrections based on the size of the database and the length of the sequence to arrive at a value called the "E" or "Expect" value. The E value corresponds to the number of sequences that one would expect to find, with an equivalent number of matching residues, in a database of certain size, containing random sequences. If a BLAST result has an E value of 5, it means we would expect to find five sequences in a random set. If a BLAST result has an E value so low that BLAST rounds it off to zero, we would not expect to find a match that good in a random set.

Other applications where BLAST is used

Although BLAST was written with the goal of finding homologous sequences, scientists use BLAST for many other tasks. BLAST can be used to determine where sequences with matching regions are positioned relative to one another, to view the relationship between mRNA and genomic DNA, to design and test PCR primers, to distinguish between different species, and to identify genetic variation and mutation sites. The NCBI even uses BLAST as a step in producing phylogenetic trees. Over the years, BLAST has become one of the most commonly used programs in biology.

Finding Structures
in the NCBI Structure Database

1. Go to the NCBI website (http://www.ncbi.nih.gov).
2. Enter the name of a protein or gene in the text box and click the "Search" button (see ***Figure 1***).

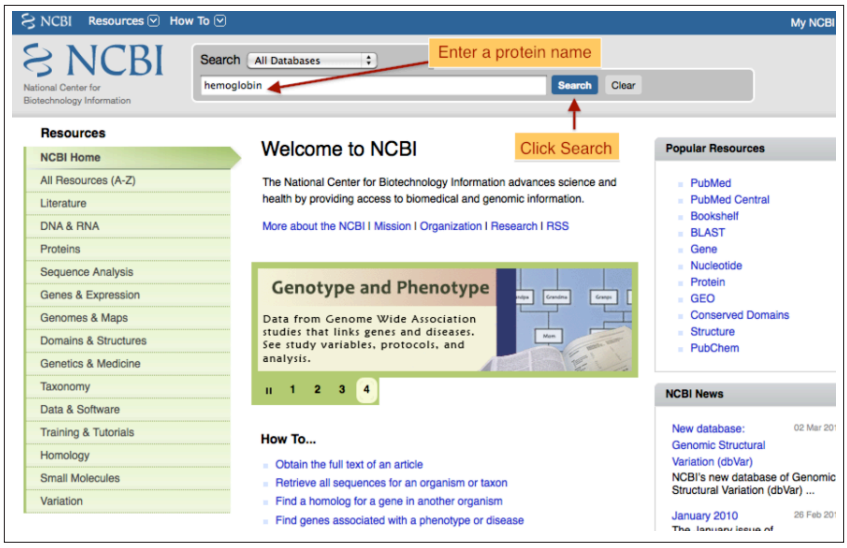


Figure 1: Enter the protein or gene name and click "Search." Credit: NCBI MMDB.

Searches that begin at the NCBI home page scan the contents of all the NCBI databases and provide the results on a page like the one below. The number of matching records appears next to the name for each database.

3. Click "Structure" to obtain the search results from the structure database. (See ***Figure 2***).

Figure 2: Choose the "Structure" database. Credit: NCBI MMDB.

The search results consist of a list of records from the database. Each record has a unique number (an accession number) that can be used to access the record.

If you find too many search results, you may narrow the search by using the word "AND" in combination with other search terms. For example, if you wish to find structures for human proteins, you may wish to search with the terms "hemoglobin AND homo." (See ***Figure 3***).

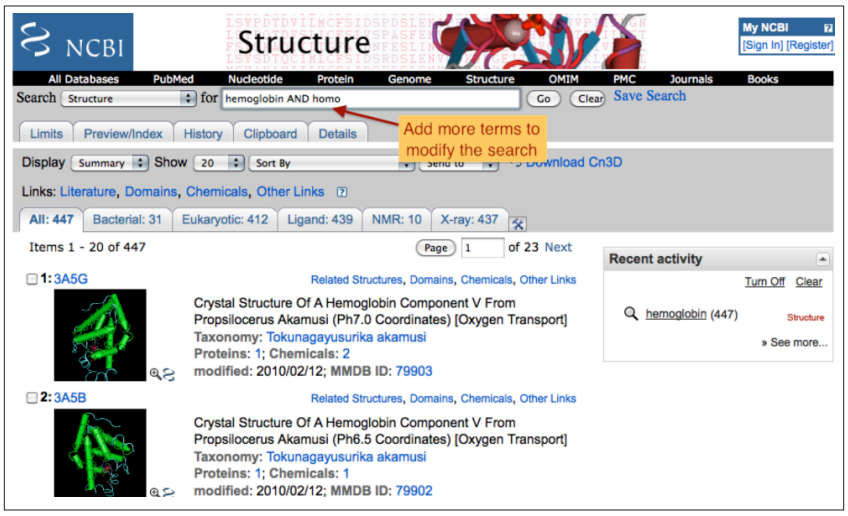


Figure 3: Narrow your search by using a combination of search terms. Credit: NCBI MMDB.

4. Click the accession number (see ***Figure 4***) for a record to view the complete information and access the download link for an individual structure.

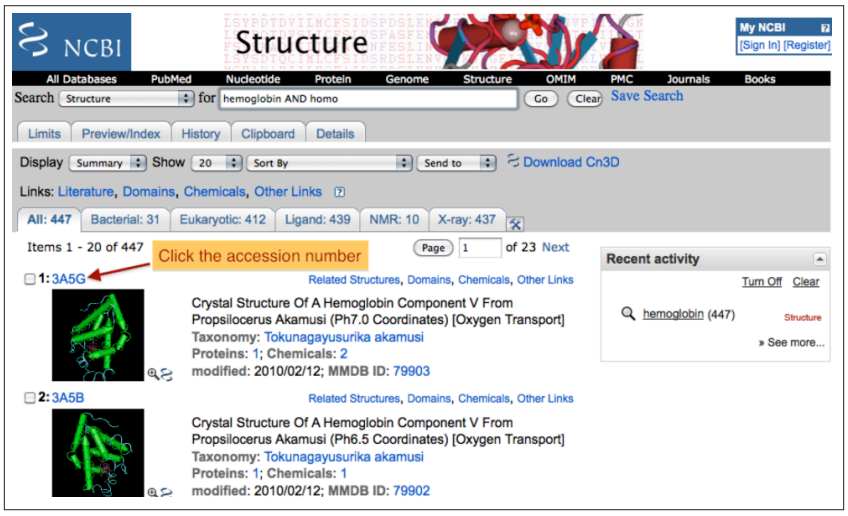
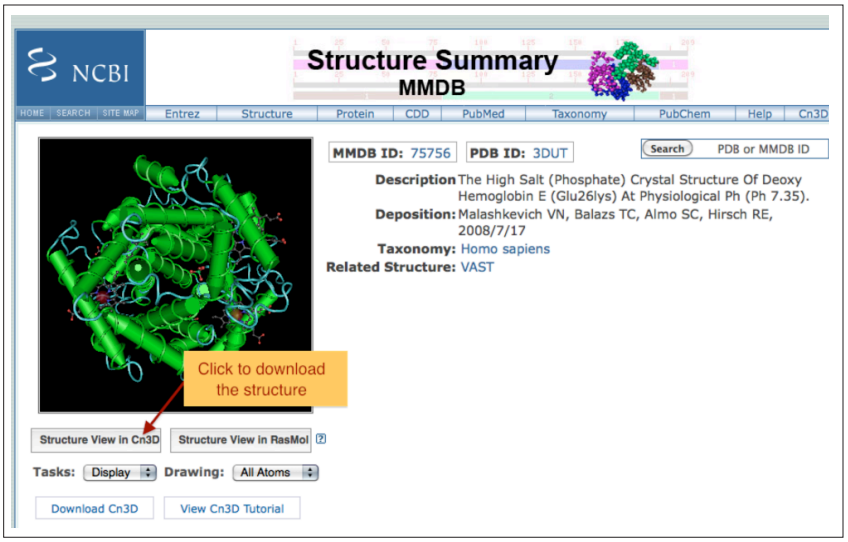


Figure 4: Click the accession number. Credit: NCBI MMDB.

5. Click the "Structure View" in Cn3D box to download the structure to your computer. (See ***Figure 5***).

Figure 5: Click "Structure View" to download the structure. Credit: NCBI MMDB.



6. Save the file on your computer and open it in Cn3D.