Bioinformatics Applications for the Study of Evolution and Heme Proteins

AP Biology Teachers Workshop
Susan Cates, Ph.D.
Courses

**Bios - 301** is the first of two courses in Biochemistry. The first semester will be spent developing the fundamentals of how living organisms carry out the chemistry essential to life. In part this means learning the nomenclature of this chemistry in the form of understanding the chemical structures and properties of the building blocks that constitute proteins and nucleic acids. The spring semester will examine metabolism and molecular physiology.

**Bios - 352** Study of selected aspects of physical chemistry as it relates to the biosciences. Includes thermodynamics, reaction rate theory, quantum mechanics, and atomic and molecular structure. Required for biochemistry majors and graduate students in biochemistry & cell biology.

**AP Biology Teachers Workshop**, Rice University, June 2006

AP Lecture 1, John S. Olson, Ph.D.
AP Lecture 2, John S. Olson, Ph.D.
AP Lecture 3, John S. Olson, Ph.D.
AP Lecture 4, John S. Olson, Ph.D.
Bioinformatics Presentation by Susan Cates, Ph.D.
Syllabus for Bioinformatics Presentation by Susan Cates, Ph.D.
Syllabus for Molecular Biology Demonstration by Angela Hvitved, Predoctoral Candidate
Syllabus for Laser Photolysis Demonstration by George Blouin, Predoctoral Candidate
Courses

Bios - 301 is the first of two courses in Biochemistry. The first semester will be spent developing the fundamentals of how living organisms carry out the chemistry essential to life. In part this means learning the nomenclature of this chemistry in the form of understanding the chemical structures and properties of the building blocks that constitute proteins and nucleic acids. The spring semester will examine metabolism and molecular physiology.

Bios - 352 Study of selected aspects of physical chemistry as it relates to the biosciences. Includes thermodynamics, reaction rate theory, quantum mechanics, and atomic and molecular structure. Required for biochemistry majors and graduate students in biochemistry & cell biology.

AP Biology Teachers Workshop, Rice University, June 2006

AP Lecture 1, John S. Olson, Ph.D.
AP Lecture 2, John S. Olson, Ph.D.
AP Lecture 3, John S. Olson, Ph.D.
AP Lecture 4, John S. Olson, Ph.D.
Bioinformatics Presentation by Susan Cates, Ph.D.
Syllabus for Bioinformatics Presentation by Susan Cates, Ph.D.
Syllabus for Molecular Biology Demonstration by Angela Hvitved, Predoctoral Candidate
Syllabus for Laser Photolysis Demonstration by George Blouin, Predoctoral Candidate
Structural and Computational Biology, Molecular Biophysics, Bioinformatics, Membrane Biology, Genomics, etc.
Structural and Computational Biology, Molecular Biophysics, Bioinformatics, Membrane Biology, Genomics, etc.
Structural and Computational Biology, Molecular Biophysics, Bioinformatics, Membrane Biology, Genomics, etc.
Structural and Computational Biology, Molecular Biophysics, Bioinformatics, Membrane Biology, Genomics, etc.
Structural and Computational Biology, Molecular Biophysics, Bioinformatics, Membrane Biology, Genomics, etc.
Structural and Computational Biology, Molecular Biophysics, Bioinformatics, Membrane Biology, Genomics, etc.
Structural and Computational Biology, Molecular Biophysics, Bioinformatics, Membrane Biology, Genomics, etc.
Background:

Why do we study heme proteins?

Why do we study evolution?
Heme Proteins
Heme Proteins

1. Hemoglobin
   oxygen transport (blood)
2. Cytochrome c oxidase
   cellular respiration, ATP synthesis
3. Peroxidases
   reduction of hydrogen peroxide
4. Cyclooxygenases
   formation of inflammatory prostanoids,
   aspirin and ibuprofen inhibit - pain relief
5. Nitric oxide synthases
   NO - cell signaling in nervous, cardiovascular
   and immune systems
6. Cytochrome P450
   metabolism, often have multiple substrates,
   and can catalyze multiple reactions -
   57 human genes for cytochrome P450s
Heme Proteins

Basic Structure of Single Domain Globins
Heme Proteins

Basic Structure of Single Domain Globins

CD corner

Heme group (red color)
Heme

prosthetic group that consists of an iron atom contained in a porphyrin ring
Study of Heme Proteins
Study of Heme Proteins

1. Blood substitutes

2. Historical Precedent

3. Disease
Study of Heme Proteins
Study of Heme Proteins

1. Blood substitutes

2. Historical Precedent

3. Disease
Blood Substitutes

A bloodless revolution

A growing interest in artificial blood substitutes has resulted in new products that could soon improve transfusion medicine.

Blood transfusion was once regarded as a safe and effective practice to save a patient's life after massive blood loss. But the AIDS epidemic and concerns that donated blood could be contaminated with HIV and other infectious agents, such as the hepatitis C virus, shattered public confidence in blood transfusions. This only added to already existing supply problems. In the USA alone, about 4 million patients require some 10–12 million blood units each year. About two-thirds of these are used to replace blood lost after trauma and surgery, and the remainder is used in chronic blood loss, cancer and anaemia. Because the demand for blood products continues to increase due to more sophisticated surgical procedures and the expanded health needs of an ageing population, there are often seasonal and regional shortages. The increasing demand and the loss of public faith in blood transfusion safety has revitalized the search for artificial blood substitutes. After many failed attempts, researchers in academia and industry are now ready to test the first promising products in laboratory and clinical trials. However, replacing our lifeblood is not as easy as it seems. "Development of an effective and safe oxygen carrier to be used instead of blood transfusion is an extreme challenge," said Robert Winslow, founder of Sangart (San Diego, CA, USA), a company that develops blood substitutes.

Although the idea of using a blood substitute in transfusions was first advanced in the seventeenth century by the British scientist Sir Christopher Wren, real efforts to develop such a product only began in the early 1930s. William Amberson and colleagues at the State College of Medicine in Memphis (TN, USA) showed that bovine haemolysates could transport oxygen in mammals (Amberson et al, 1933), and later found that human haemolysates had the same potential when infused in patients (Amberson et al, 1949). Since then, the development of transfusion medicine has received much of its impetus from the military—not surprising, given that haemorrhage is the leading cause of death on the battlefield. In the 1960s, during the Vietnam War, the US Army began to collaborate with several private biomedical companies and supported research on blood substitutes based on different approaches.

Wartime experience guided developers in delineating the ideal properties of an oxygen carrier for military use: universal compatibility, and thus no time-consuming testing for blood type; no transmittable pathogens or allergens; long-term storage capability, preferably under non-refrigerated conditions; safety and non-toxicity; and superior oxygen delivery capability, resulting in oxygenation of peripheral tissues even after massive blood loss. Intuitively, this describes applications that go far beyond wounded soldiers, ranging from immediate treatment of patients in haemorrhagic shock after traumatic injuries, to relieving shortages and avoiding contaminated supplies, to assuring intervention even when religious objections or rare blood groups prevent normal blood transfusion. Additional uses can be envisaged in a variety of diseases with...
Blood Substitutes

Blood Supply Update

10-09-2006

REGIONAL BLOOD INVENTORY: low

As of 10-09-2006

American Red Cross
Blood Services
Penn-Jersey Region
Blood Substitutes
Bleeding (Exsanguination, Hemorrhagic Shock) is still the number one cause of death on the battlefield.
## Table 3: Testing for the four major TTIs expressed as percentages of total units reported by 38 countries in 2004

<table>
<thead>
<tr>
<th></th>
<th>Tested</th>
<th></th>
<th>Tested</th>
<th></th>
<th>Unknown</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Units</td>
<td>%</td>
<td>Units</td>
<td>%</td>
<td>Units</td>
<td>%</td>
</tr>
<tr>
<td>HIV</td>
<td>2185909</td>
<td>98.52%</td>
<td>3890</td>
<td>0.18%</td>
<td>28777</td>
<td>1.3%</td>
</tr>
<tr>
<td>HBV</td>
<td>2097201</td>
<td>94.53%</td>
<td>16021</td>
<td>0.72%</td>
<td>105354</td>
<td>4.75%</td>
</tr>
<tr>
<td>HCV</td>
<td>1796420</td>
<td>80.97%</td>
<td>157592</td>
<td>7.10%</td>
<td>264564</td>
<td>11.92%</td>
</tr>
<tr>
<td>Syphilis</td>
<td>2135389</td>
<td>96.25%</td>
<td>54410</td>
<td>2.45%</td>
<td>28777</td>
<td>1.30%</td>
</tr>
</tbody>
</table>
Transfused human donor blood carries disease

Table 3: Testing for the four major TTIs expressed as percentages of total units reported by 38 countries in 2004

<table>
<thead>
<tr>
<th>Tested</th>
<th>Units</th>
<th>%</th>
<th>Untested</th>
<th>Units</th>
<th>%</th>
<th>Unknown</th>
<th>Units</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV</td>
<td>2185909</td>
<td>98.52%</td>
<td>3890</td>
<td>0.18%</td>
<td></td>
<td>28777</td>
<td>1.3%</td>
<td></td>
</tr>
<tr>
<td>HBV</td>
<td>2097201</td>
<td>94.53%</td>
<td>16021</td>
<td>0.72%</td>
<td></td>
<td>105354</td>
<td>4.75%</td>
<td></td>
</tr>
<tr>
<td>HCV</td>
<td>1796420</td>
<td>80.97%</td>
<td>157592</td>
<td>7.10%</td>
<td></td>
<td>264564</td>
<td>11.92%</td>
<td></td>
</tr>
<tr>
<td>Syphilis</td>
<td>2135389</td>
<td>96.25%</td>
<td>54410</td>
<td>2.45%</td>
<td></td>
<td>28777</td>
<td>1.30%</td>
<td></td>
</tr>
</tbody>
</table>
Advantages of Blood Substitutes:

1. Unlimited availability
2. Universal donor (no typing)
3. Disease-free
4. Longer shelf life
Study of Heme Proteins
Study of Heme Proteins

1. Blood substitutes

2. Historical Precedent

3. Disease
Historical Precedent
Nobel Prize Ceremonies, December 10, 1962

**Physiology or medicine:** Watson, Crick, and Wilkins for double helical DNA
**Chemistry:** Perutz and Kendrew for the 3-D structures of Proteins
Historical Precedent

The Nobel Prize in Chemistry 1962

"for their studies of the structures of globular proteins"

Max Ferdinand Perutz
1/2 of the prize
United Kingdom

John Cowdery Kendrew
1/2 of the prize
United Kingdom
Historical Precedent

The Nobel Prize in Chemistry 1962

"for their studies of the structures of globular proteins"

Max Ferdinand Perutz
1/2 of the prize
United Kingdom

John Cowdery Kendrew
1/2 of the prize
United Kingdom

Myoglobin (Mb) and Hemoglobin (Hb)
Historical Precedent

The Nobel Prize in Chemistry 1962

"for their studies of the structures of globular proteins"

Myoglobin (Mb) and Hemoglobin (Hb)

Max Ferdinand Perutz
United Kingdom

John Cowdery Kendrew
United Kingdom

1/2 of the prize
1/2 of the prize
First three dimensional structures of proteins (1957-1960): hemoglobin (Hb) and myoglobin (Mb)
Study of Heme Proteins
Study of Heme Proteins

1. Blood substitutes

2. Historical Precedent

3. Disease
Hemoglobin related diseases
Hemoglobin related diseases

a. Thalassemia
b. Sickle Cell Anemia
c. Staph Infection - (*Staphylococcus Aureus*)
Disease

Staph Infection - *Staphylococcus aureus*
Disease

Sickle Cell anemia

Red blood cells containing HbS $\beta$ subunits

Normal red blood cells with no deformations

http://www.sicklecelledisease.org/research/index.phtml
Disease

Sickle Cell anemia

Red blood cells containing HbS \( \beta \) subunits

Normal red blood cells with no deformations

http://www.sicklecelldisease.org/research/index.phtml
Disease

Sickle Cell anemia

Red blood cells containing HbS β subunits

Normal red blood cells with no deformations

http://www.sicklecelldisease.org/research/index.phtml
Disease

Sickle Cell anemia

Sickling cell

Irreversibly sickled cells

Echinocyte

Normal red blood cells with no deformations

Red blood cells containing HbS \( \beta \) subunits

http://www.sicklecelldisease.org/research/index.phtml
Disease

Sickle Cell anemia
Disease

Sickle Cell anemia

Hb AS is benign (only shortens red cell life time to ~80 days)

Chromosome 11

<table>
<thead>
<tr>
<th>Hb</th>
<th>1-6</th>
<th>ε</th>
<th>γA</th>
<th>γG</th>
<th>δ</th>
<th>β</th>
<th>βs</th>
</tr>
</thead>
</table>
Sickle Cell anemia

Hb AS is benign (only shortens red cell life time to ~80 days)

Hb SS leads to severe anemia and early death
Disease

Sickle Cell anemia

Hb AS is benign (only shortens red cell life time to ~80 days)

Chromosome 11

Hb SS leads to severe anemia and early death

Chromosome 11

Hemoglobin S solution
Hemoglobin S polymer

Oxygenated
Deoxygenated

Hemoglobin S cell
Cell heterogeneity

Vaso-occlusion
Sickled cells
Disease

Sickle Cell anemia
Disease

Sickle Cell anemia

Anopheles gambiae,
Sickle Cell anemia

Infection with Plasmodium falciparum

Anopheles gambiae,
Sickle Cell anemia

Infection with *Plasmodium falciparum*

The *Plasmodium* parasite must remain in the red blood cells for more than 90 days to be able to change its cell coat and avoid antibody responses.
Disease

Sickle Cell anemia

Infection with *Plasmodium falciparum*

The *Plasmodium* parasite must remain in the red blood cells for more than 90 days to be able to change its cell coat and avoid antibody responses

*Anopheles gambiae,*
Sickle Cell anemia

Infection with *Plasmodium falciparum*

HbAS cells only last ~80 days and release immature parasite which is destroyed by the immune system.

The *Plasmodium* parasite must remain in the red blood cells for more than 90 days to be able to change its cell coat and avoid antibody responses.
Sickle Cell Anemia

The sickle cell trait is maintained by strong selective pressure because AS individuals are resistance to malaria. It is a classic case of natural selection.
Sickle Cell Anemia

The sickle cell trait is maintained by strong selective pressure because AS individuals are resistance to malaria. It is a classic case of natural selection.
Sickle Cell Anemia

The sickle cell trait is maintained by strong selective pressure because AS individuals are resistance to malaria. It is a classic case of natural selection.
Sickle Cell Anemia

The sickle cell trait is maintained by strong selective pressure because AS individuals are resistance to malaria. It is a classic case of natural selection.
Disease

Sickle Cell Anemia

The sickle cell trait is maintained by strong selective pressure because AS individuals are resistance to malaria. It is a classic case of natural selection.
Disease

Sickle Cell Anemia

The sickle cell trait is maintained by strong selective pressure because AS individuals are resistance to malaria. It is a classic case of natural selection.
Disease

Thalassemia

Alpha thalassemia
(i) Silent carrier state - hematologically normal
(ii) Mild alpha-thalassemia - mild anemia
(iii) Hemoglobin H disease - severe anemia
(iv) Hydrops fetalis - usually fatal in utero

Beta thalassemia
(i) Thalassemia minor, or thalassemia trait - mild anemia
(ii) Thalassemia intermedia - significant anemia
(iii) Thalassemia major - chronic transfusions required
Background:

Why do we study heme proteins?

Why do we study evolution?
Evolution of Species
Phylogenetic Trees show the relatedness of organisms

Phylogenetic Tree of Life

Bacteria
- Spirochetes
- Green Filamentous bacteria
- Proteobacteria
- Gram positives
- Cyanobacteria
- Planctomyces
- Bacteroides
- Cytophaga
- Thermotoga
- Aquifex

Archaea
- Methanosarcina
- Methanobacterium
- Methanococcus
- T. celer
- Thermoproteus
- Pyrodictium
- Entamoebae
- Slime molds
- Animals
- Fungi
- Plants
- Ciliates
- Flagellates
- Trichomonads
- Microsporidia
- Diplomonads

Eucarya
Evolution of Species
Phylogenetic Trees show the relatedness of organisms

Phylogenetic Tree of Life

Bacteria
  - Spirochetes
  - Proteobacteria
    - Cyanobacteria
      - Planctomyces
    - Bacteroides
    - Cytophaga
    - Thermotoga
    - Aquifex
  - Gram positives
    - Green filamentous bacteria

Archaea
  - Methanosarcina
  - Methanobacterium
  - Methanococcus
  - T. celer
  - Thermoproteus
  - Pyrodictium
  - Halophiles
  - Entamoebae
  - Slime molds

Eucarya
  - Animals
    - Fungi
      - Plants
    - Ciliates
    - Flagellates
    - Trichomonads
    - Microsporidia
    - Diplomonads

Common Ancestor (Root of the tree)
Molecular Evolution
Molecular Evolution

Mb

Hb
Molecular Evolution
Molecular Evolution: relatedness of biological molecules

I. Genes
determined primarily by nucleic acid sequence

II. Proteins
determined primarily by amino acid sequence
Multidomain flavohemoglobins (FHbs) & Globin-coupled sensors (GCS)

Globin phylogenetic trees

Vinogradov et al. (2005) Proc. Natl. Acad. Sci., USA 102, 11385-11389

E. coli flavoHb

Cyanobacterial Hb

2/2 or truncated globins (TrHbs)

Bacillus subtilis Hb

Rice non-symbiotic Hb

Sperm whale Mb

Human HbA

3/3 Single domain globins (SDGs or “classical” hemoglobins)
Multidomain flavohemoglobins (FHbs) & Globin-coupled sensors (GCS)

Globin phylogenetic trees

E. coli flavoHb

Cyanobacterial Hb

2/2 or truncated globins (TrHbs)

Bacillus subtilis Hb

Sperm whale Mb

Rice non-symbiotic Hb

3/3 Single domain globins (SDGs or “classical” hemoglobins)

Vinogradov et al. (2005) Proc. Natl. Acad. Sci., USA 102, 11385-11389
Single domain globin
(The classical myoglobin fold)

CD corner

Myoglobin (SDgb)
3/3 AGH/BEF
(C and D are small and variable)
Single domain globin
(The classical myoglobin fold)

CD corner

Myoglobin (SDgb) 3/3 AGH/BEF (C and D are small and variable)
Single domain globin
(The classical myoglobin fold)

CD corner

Myoglobin (SDgb)
3/3 AGH/BEF
(C and D are small and variable)
The major animal globins participate in $O_2$ storage and transport and provide lessons for designing rHb-based blood substitutes.
Multidomain flavohemoglobins (FHbs) & Globin-coupled sensors (GCS)

Globin phylogenetic trees

Vinogradov et al. (2005) Proc. Natl. Acad. Sci., USA 102, 11385-11389
Multidomain flavohemoglobins (FHbs) & Globin-coupled sensors (GCS)

Globin phylogenetic trees

Vinogradov et al. (2005) Proc. Natl. Acad. Sci., USA 102, 11385-11389

- E. coli flavoHb
- Cyanobacterial Hb
- Sperm whale Mb
- Human HbA
- Rice non-symbiotic Hb

2/2 or truncated globins (TrHbs)

3/3 Single domain globins (SDGs or "classical" hemoglobins)

Bacillus subtilis Hb
Single domain globin (truncated fold)

BE corner

Short or missing A helix

large F loop

*Bacillus subtilis* Hb

2/2 **BE/GH**

(F is variable, A is small)
Multidomain flavohemoglobins (FHbs) & Globin-coupled sensors (GCSs)

Globin phylogenetic trees

Vinogradov et al. (2005) Proc. Natl. Acad. Sci., USA 102, 11385-11389

- E. coli flavoHb
- Cyanobacterial Hb
- Sperm whale Mb
- Human HbA
- Rice non-symbiotic Hb
- 3/3 Single domain globins (SDGs or “classical” hemoglobins)
- 2/2 or truncated globins (TrHbs)
Chimeric (bi-domain) globin
3/3 fold
Globin domain

FAD
Heme

Reductase domain

E. coli FlavoHb (Hmp)
Chimeric (bi-domain) globin
3/3 fold
Globin domain

FAD
Heme

Reductase domain
E. coli FlavoHb (Hmp)

Myoglobin (chordate)
Chimeric (bi-domain) globin
3/3 fold
Globin domain

Heme

FAD

Reductase domain

E. coli FlavoHb (Hmp)

VHb (bacterial)
(Vitreoscilla stearcorium)

Myoglobin (chordate)
Flavohemoglobinins (NO dioxygenation)

Pathogenic fungi

Defense mechanisms against host macrophages and other sources of NC
Flavohemoglobinins (NO dioxygenation)

Pathogenic fungi

Defense mechanisms against host macrophages and other sources of NC
Macrophage engulfing bacteria or fungi.

Human macrophages up-regulate inducible nitric oxide synthase (iNOS) and produce increased levels of NO after infection.
Macrophage engulfing bacteria or fungi.

Human macrophages up-regulate inducible nitric oxide synthase (iNOS) and produce increased levels of NO after infection.

Fox, S. I. Human Physiology (7th Edition), Fig. 15.5
NO is potentially very toxic
NO is potentially very toxic

1. NO can inactivate aconitase at levels $\leq 200$ nM; shuts down TCA cycle (Gardner et al., 1997 J. Biol. Chem. 272, 25071-76)
NO is potentially very toxic

1. NO can inactivate aconitase at levels ≤ 200 nM; shuts down TCA cycle (Gardner et al., 1997 J. Biol. Chem. 272, 25071-76)

2. NO can inhibit cytochrome oxidases at levels ≤ 1 μM (Sarti et al. (M. Brunori), 2000 Biochem. Biophys. Res. Com. 274, 183-7; Tania et al. (R. Poole), 2000, J. Biol. Chem. in press)
NO is potentially very toxic

1. NO can inactivate aconitase at levels \( \leq 200 \text{ nM} \); shuts down TCA cycle (Gardner et al., 1997 J. Biol. Chem. 272, 25071-76)

2. NO can inhibit cytochrome oxidases at levels \( \leq 1 \text{ \( \mu \)M} \) (Sarti et al.(M. Brunori), 2000 Biochem. Biophys. Res. Com. 274, 183-7; Tania et al.(R. Poole), 2000, J. Biol. Chem. in press)

3. NO synthesis inhibits respiration in endothelial and smooth muscle cells (Clementi et al. (S. Moncada), 1999, PNAS 96, 1559-1562)
NO is potentially very toxic

1. NO can inactivate aconitase at levels $\leq 200$ nM; shuts down TCA cycle (Gardner et al., 1997 J. Biol. Chem. 272, 25071-76)

2. NO can inhibit cytochrome oxidases at levels $\leq 1$ $\mu$M (Sarti et al.(M. Brunori), 2000 Biochem. Biophys. Res. Com. 274, 183-7; Tania et al.(R. Poole), 2000, J. Biol. Chem. in press)

3. NO synthesis inhibits respiration in endothelial and smooth muscle cells (Clementi et al. (S. Moncada), 1999, PNAS 96, 1559-1562)

NO dioxygenase (NOD) activity detoxifies NO (and $O_2$)

$$2\text{Hb(Fell)}O_2 + 2\text{NO} \rightarrow 2\text{NO}_3^- + 2\text{Hb(Fell)}\text{(non-toxic)}$$

O$_2$

Cytochrome b$_5$ or flavoprotein (FAD)

NAD

(reductase)

NADH
NO is potentially very toxic

1. NO can inactivate aconitase at levels $\leq 200$ nM; shuts down TCA cycle (Gardner et al., 1997 J. Biol. Chem. 272, 25071-76)

2. NO can inhibit cytochrome oxidases at levels $\leq 1$ $\mu$M (Sarti et al. (M. Brunori), 2000 Biochem. Biophys. Res. Com. 274, 183-7; Tania et al. (R. Poole), 2000, J. Biol. Chem. in press)

3. NO synthesis inhibits respiration in endothelial and smooth muscle cells (Clementi et al. (S. Moncada), 1999, PNAS 96, 1559-1562)

NO dioxygenase (NOD) activity detoxifies NO (and $O_2$)

\[
2\text{Hb(Fell)}O_2 + 2\text{NO} \rightarrow 2\text{NO}_3^- + 2\text{Hb(Fell)}\text{(non-toxic)}
\]

FlavoHbs are expressed to detoxify NO, increasing the resistance of pathogenic and symbiotic microorganisms to host defense mechanisms.

(Gardner et al., 1998, PNAS 95, 10378-10383; Stamler, Poole, and others)
FlavoHbs are expressed in response to NO

*Escherichia coli*

*Candida albicans*

*Aspergillus fumigatus*
FlavoHbs are expressed in response to NO

Escherichia coli

Aspergillus fumigatus

Candida albicans
FlavoHbs are expressed in response to NO

Globin domain

Escherichia coli

Aspergillus fumigatus

Candida albicans

NO
FlavoHbs are expressed in response to NO
FlavoHbs are expressed in response to NO

Heme
Globin domain
Reductase domain

Escherichia coli
Aspergillus fumigatus
Candida albicans

NO
NO
NO
Flavohbns are expressed in response to NO
Bioinformatics: Using libraries created by sequencing genes from different species

21st Century Biology
<table>
<thead>
<tr>
<th>Animal</th>
<th>F-helix sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>whelk</td>
<td>AKKLRSRNHTA</td>
</tr>
<tr>
<td>aardvark</td>
<td>IQPLAQSHAT</td>
</tr>
<tr>
<td>pig</td>
<td>LTPLAQSHAT</td>
</tr>
<tr>
<td>mouse</td>
<td>IQPLAQSHAT</td>
</tr>
<tr>
<td>rabbit</td>
<td>IKPLAQSHAT</td>
</tr>
<tr>
<td>sheep</td>
<td>VKHLAESHAN</td>
</tr>
<tr>
<td>bovine</td>
<td>VKHLAESHAN</td>
</tr>
<tr>
<td>horse</td>
<td>LKPDAQSHAT</td>
</tr>
<tr>
<td>elephant</td>
<td>LKPDAQSHAT</td>
</tr>
<tr>
<td>whale</td>
<td>LKPDAQSHAT</td>
</tr>
<tr>
<td>dog</td>
<td>LKPDAQSHAT</td>
</tr>
<tr>
<td>chicken</td>
<td>LKPDAQSHAT</td>
</tr>
<tr>
<td>alligator</td>
<td>LKPDAQSHAT</td>
</tr>
<tr>
<td>tuna</td>
<td>LKPLAKSHAL</td>
</tr>
<tr>
<td>shark</td>
<td>LKPLANSHAT</td>
</tr>
<tr>
<td>sea slug</td>
<td>VKELADTHIN</td>
</tr>
<tr>
<td>red clam</td>
<td>LSQFAKEHVG</td>
</tr>
<tr>
<td>fly larva</td>
<td>VKEFAVNHIT</td>
</tr>
<tr>
<td>roundworm</td>
<td>VNTFVASHKP</td>
</tr>
<tr>
<td></td>
<td>AREIVDPHLR</td>
</tr>
</tbody>
</table>

**Bioinformatics:** Using libraries created by sequencing genes from different species

21st Century Biology
Bioinformatics: Using libraries created by sequencing genes from different species

21st Century Biology

whelk
aardvark
pig
mouse
rabbit
sheep
bovine
horse
elephant
whale
dog
chicken
alligator
tuna
shark
sea slug
red clam
fly larva
roundworm

F-helix sequences

helical position

F1  F4  F8

AKKLSRNHTA
IQPLAQSHAT
LTPLAQSHAT
IQPLAQSHAT
IKPLAQSHAT
VKHLAESHAN
VKHLAESHAN
LKPLAQSHAT
IQPLAQSHAT
LKPLAQSHAT
LKPLAQTHAT
LKPKLAKSHAL
LKPKLANSHTA
VKEIADTHIN
LSQFAKEHVG
VEKFAVNHIT
VNTFVASHKP
AREIVDPHLR
Basic Structure of Single Domain Globins

Required histidine (F8) for coordination to the iron atom
Basic Structure of Single Domain Globins

Heme group (red color)

Required histidine (F8) for coordination to the iron atom

CD corner
Sequence comparisons of globins from bacteria to man (weak homologies except in key structural (3°) regions)

<table>
<thead>
<tr>
<th>Protein</th>
<th>α-Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. catodon SWMb</td>
<td>A1...B10...CD1...D1</td>
</tr>
<tr>
<td>H. sapiens Hb β</td>
<td>VHLTPEEKSATVLTGKV--NVDEVQGAEALRLLVYWTQRFFESFGDLS--TP</td>
</tr>
<tr>
<td>A. aeolicus AaTgb</td>
<td>--MLSEETIRIKSTVPLKKEHGEITARMYELLFLSKYPKTKELEAPAGASEE--</td>
</tr>
<tr>
<td>V. stercoraria Hb</td>
<td>--MLDQQTINIIKATPVKHEVTVTTTFYKNLFAKHPESVPLFDGMGQESLE--</td>
</tr>
<tr>
<td>E. coli Fhb</td>
<td>--MLDAQTITAVKATIPPLLVTGKLTAHFYDRMFTHNPELKEIFNMSNQRN--</td>
</tr>
<tr>
<td>A. eutrophus Fhb</td>
<td>--MTQKTKDIVKAATPVLAEHGYYDIIKCFYQRMEAEHPELKNVNFNMAHQ--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>α-Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. catodon SWMb</td>
<td>E1...E7...E11...E20...F1...F4...F8...G1...</td>
</tr>
<tr>
<td>H. sapiens Hb β</td>
<td>AEMKASEDLKKHGVTULTALGAILKKGK----HEAEIKPLAQSHATKHKIPKYL</td>
</tr>
<tr>
<td>A. aeolicus AaTgb</td>
<td>DAVMNPKVKALGGKVLGAFSDGLHALDN----LKGRFATLSHELHCDKLHVPDENV</td>
</tr>
<tr>
<td>V. stercoraria Hb</td>
<td>-----------OPKPLNANAIAYATYIDREELDINATISTARSHEVRN-VKPEHY</td>
</tr>
<tr>
<td>E. coli Fhb</td>
<td>-----------OPKALAMTVLAAQNENLPAILPPAVKKIAVHCQAG-VAAAY</td>
</tr>
<tr>
<td>A. eutrophus Fhb</td>
<td>-----------GDQREALFNAAAYASNENLPAALLPAVEKIAQHTSFQ-IKPEQY</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>α-Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>P. catodon SWMb</td>
<td>G19...H1...H21...</td>
</tr>
<tr>
<td>H. sapiens Hb β</td>
<td>EFISEAIIHVLHSRHPDGFDAQGAMNKALELFREDKIAYKELGYQ---</td>
</tr>
<tr>
<td>A. aeolicus AaTgb</td>
<td>RLLGNVLVCVLHGFKEFTPPVQAAYQKAVAGVANLAHKYH----</td>
</tr>
<tr>
<td>V. stercoraria Hb</td>
<td>PLVKECCLQAIIEELNP----GEEVNLKAWEEEYDFLASTLITLEKLYSQP--</td>
</tr>
<tr>
<td>E. coli Fhb</td>
<td>PIVGQELLSGAIKEVLGDAATDDIDLWAGKAYGVIADVFIQVEADLYAQAVE-</td>
</tr>
<tr>
<td>A. eutrophus Fhb</td>
<td>NIVGEHLLATLDEMFSP----GQEVLGDAWKGAYGVLANFTHNRAEYIYENASK</td>
</tr>
</tbody>
</table>

Miranda et al. (2006) J. Biol. Chem.280, 3674-35761
Sequence comparisons of globins from bacteria to man (weak homologies except in key structural (3°) regions)

Miranda et al. (2006) J. Biol. Chem. 280, 3674-35761
Applications in Human developmental biology

HbEmbryonic = $\alpha_2\varepsilon_2$  
HbF = $\alpha_2\gamma_2$  
HbA = $\alpha_2\beta_2$

Chromosome 16

$\zeta$  
$\alpha_1$  
$\alpha_2$

Chromosome 11

HS 1-6  
$\varepsilon$  
$\gamma_A$  
$\gamma_G$  
$\delta$  
$\beta$
Applications in Human developmental biology

HbEmbryonic = $\alpha_2\varepsilon_2$  
HbF = $\alpha_2\gamma_2$  
HbA = $\alpha_2\beta_2$

Chromosome 16

$\zeta$  
$\alpha_1$  
$\alpha_2$

Chromosome 11

HS  
1-6  
$\varepsilon$  
$\gamma_A$  
$\gamma_G$  
$\delta$  
$\beta$
Applications in Human developmental biology

$Hb_{Embryonic} = \alpha_2\varepsilon_2$  $HbF = \alpha_2\gamma_2$  $HbA = \alpha_2\beta_2$

Chromosome 16

\[
\begin{array}{c}
\zeta & \zeta \\
\alpha_1 & \alpha_1 \\
\alpha_2 & \alpha_2 \\
\end{array}
\]

Chromosome 11

\[
\begin{array}{c}
\varepsilon & \varepsilon \\
\gamma A & \gamma A \\
\gamma G & \gamma G \\
\delta & \delta \\
\beta & \beta \\
\end{array}
\]

[HS 1-6] [HS 1-6]
α₂ε₂ Hb
embryonic
(High O₂ affinity)
α₂ε₂ Hb
embryonic
(High O₂ affinity)
\( \alpha_2 \epsilon_2 \) Hb embryonic

(High \( O_2 \) affinity)

\( \alpha_2 \gamma_2 \) HbF

(moderate \( O_2 \) affinity)
\( \alpha_2 \varepsilon_2 \text{ Hb embryonic} \)  
(High \( O_2 \) affinity)

\( \alpha_2 \gamma_2 \text{ HbF} \)  
(moderate \( O_2 \) affinity)
\[ \alpha_2 \varepsilon_2 \text{ Hb embryonic} \]  
(High \( O_2 \) affinity)  

\[ \alpha_2 \gamma_2 \text{ HbF} \]  
(moderate \( O_2 \) affinity)  

\[ \alpha_2 \beta_2 \text{ HbA} \]  
(low \( O_2 \) affinity)
Acid, base, and isopropanol tests for hemoglobin stability

Embryonic and Fetal Hbs are much more stable (application for blood substitutes)
Applications in genetic diseases

Sickle Cell Anemia
Applications in genetic diseases

Hb AS is benign (only shortens red cell life time to ~80 days)

Chromosome 11

Sickle Cell Anemia
Applications in genetic diseases

Hb AS is benign (only shortens red cell life time to ~80 days)

Hb SS leads to severe anemia and early death

Sickle Cell Anemia
Applications in genetic diseases

Chromosome 16

ξ
α1
α2

Chromosome 11

HS 1-6
ε
γA
γG
δ
β

Thalassemia
Applications in genetic diseases

Chromosome 16

\[ \zeta \quad \zeta \]

\[ \alpha_1 \quad \alpha_1 \]

\[ \alpha_2 \quad \alpha_2 \]

Chromosome 11

\[ \text{HS 1-6} \]

\[ \varepsilon \]

\[ \gamma_A \]

\[ \gamma_G \]

\[ \delta \]

\[ \beta \]

Thalassemia
Applications in genetic diseases

Chromosome 16
- $\xi$
- $\alpha_1$
- $\alpha_2$

Chromosome 11
- HS 1-6
- $\varepsilon$
- $\gamma_A$
- $\gamma_G$
- $\delta$
- $\beta$

Thalassemia
Applications in genetic diseases

Alpha thalassemia
(i) Silent carrier state - hematologically normal
(ii) Mild alpha-thalassemia - mild anemia
(iii) Hemoglobin H disease - severe anemia
(iv) Hydrops fetalis - usually fatal in utero

Beta thalassemia
(i) Thalassemia minor, or thalassemia trait - mild anemia
(ii) Thalassemia intermedia - significant anemia
(iii) Thalassemia major - chronic transfusions required

Thalassemia
Applications in genetic diseases

Chromosome 16

- ζ
- α1
- α2

Chromosome 11

- HS 1-6
- ε
- γA
- γG
- δ
- β

Alpha thalassemia
(i) Silent carrier state - hematologically normal
(ii) Mild alpha-thalassemia - mild anemia
(iii) Hemoglobin H disease - severe anemia
(iv) Hydrops fetalis - usually fatal in utero

Beta thalassemia
(i) Thalassemia minor, or thalassemia trait - mild anemia
(ii) Thalassemia intermedia - significant anemia
(iii) Thalassemia major - chronic transfusions required
Applications in genetic diseases

Alpha thalassemia
(i) Silent carrier state - hematologically normal
(ii) Mild alpha-thalassemia - mild anemia
(iii) Hemoglobin H disease - severe anemia
(iv) Hydrops fetalis - usually fatal in utero

Beta thalassemia
(i) Thalassemia minor, or thalassemia trait - mild anemia
(ii) Thalassemia intermedia - significant anemia
(iii) Thalassemia major - chronic transfusions required
Applications in genetic diseases

Chromosome 16

- $\zeta$ $\zeta$
- $\alpha_1$ $\alpha_1$
- $\alpha_2$ $\alpha_2$

Alpha thalassemia
(i) Silent carrier state - hematologically normal
(ii) Mild alpha-thalassemia - mild anemia
(iii) Hemoglobin H disease - severe anemia
(iv) Hydrops fetalis - usually fatal in utero

Thalassemia
Applications in genetic diseases

Alpha thalassemia
(i) Silent carrier state - hematologically normal
   1 inactive alpha gene - 3 are normal
(ii) Mild alpha-thalassemia - mild anemia
(iii) Hemoglobin H disease - severe anemia
(iv) Hydrops fetalis - usually fatal in utero

Thalassemia
Applications in genetic diseases

Alpha thalassemia
(i) Silent carrier state - hematologically normal
   1 inactive alpha gene - 3 are normal
(ii) Mild alpha-thalassemia - mild anemia
   2 inactive alpha genes - red cells are smaller
(iii) Hemoglobin H disease - severe anemia
(iv) Hydrops fetalis - usually fatal in utero

Thalassemia
Applications in genetic diseases

Chromosome 16

Alpha thalassemia
(i) Silent carrier state - hematologically normal
   1 inactive alpha gene - 3 are normal
(ii) Mild alpha-thalassemia - mild anemia
    2 inactive alpha genes - red cells are smaller
(iii) Hemoglobin H disease - severe anemia
     3 inactive alpha genes - appear malnourished, enlarged spleens, bony abnormalities
(iv) Hydrops fetalis - usually fatal in utero

Thalassemia
Applications in genetic diseases

Chromosome 16

α2

α1

α1

ξ

ξ

Alpha thalassemia
(i) Silent carrier state - hematologically normal
  1 inactive alpha gene - 3 are normal
(ii) Mild alpha-thalassemia - mild anemia
  2 inactive alpha genes - red cells are smaller
(iii) Hemoglobin H disease - severe anemia
  3 inactive alpha genes - appear malnourished, enlarged spleens, bony abnormalities
(iv) Hydrops fetalis - usually fatal in utero
  4 inactive alpha genes - in utero transfusions followed by life-long, chronic transfusions required to survive

Thalassemia
Applications in genetic diseases

Beta thalassemia
(i) Thalassemia minor, or thalassemia trait - mild anemia
(ii) Thalassemia intermedia - significant anemia
(iii) Thalassemia major - chronic transfusions required

Thalassemia
Applications in genetic diseases

Beta thalassemia
(i) Thalassemia minor, or thalassemia trait - mild anemia
mildly suppressed beta gene - main danger is passing the trait to children with a partner who also has thalassemia trait
(ii) Thalassemia intermedia - significant anemia
(iii) Thalassemia major - chronic transfusions required

Thalassemia
Applications in genetic diseases

Beta thalassemia

(i) Thalassemia minor, or thalassemia trait - mild anemia
mildly suppressed beta gene - main danger is passing the trait to children with a partner who also has thalassemia trait

(ii) Thalassemia intermedia - significant anemia
suppressed or inactive beta gene - small stature, poor weight gain, poor energy levels, and susceptibility to infection

(iii) Thalassemia major - chronic transfusions required
Applications in genetic diseases

Beta thalassemia
(i) Thalassemia minor, or thalassemia trait - mild anemia
   mildly suppressed beta gene - main danger is passing the trait to children with a partner who also has thalassemia trait
(ii) Thalassemia intermedia - significant anemia
   suppressed or inactive beta gene - small stature, poor weight gain, poor energy levels, and susceptibility to infection
(iii) Thalassemia major - chronic transfusions required
   Severely suppressed or inactive beta gene(s) - chronic blood transfusions are needed

Thalassemia
BREAK
Bioinformatics Exercises

Advantages:
1. Most schools have computer labs or library computer access.
2. Exercises only require a browser.
3. Students interested in all areas of bioscience must learn to mine computational databases - unavoidable for future scientists.

Pitfalls to avoid:
1. Databases are HUGE, must give specific, step-by-step instructions or the student will get hopelessly lost.
2. Some sites go down temporarily, design a 2- or 3-part exercise that uses more than one site, or make the due date flexible.
3. Practice the exercise each year, databases and interfaces change rapidly - what worked last semester may not work now.
Databases

1. Choose exercises or background reading that stresses the size and growth rate of biological databases.

2. Convey the necessity of computers in storing and analyzing such a large amount of data.

3. Curated vs. uncurated databases.

4. Discuss search engines and search design: specificity vs. sensitivity
Databases

1. Choose exercises or background reading that stresses the size and growth rate of biological databases. **NCBI**

2. Convey the necessity of computers in storing and analyzing such a large amount of data.

3. Curated vs. uncurated databases.

4. Discuss search engines and search design: specificity vs. sensitivity
Databases

1. Choose exercises or background reading that stresses the size and growth rate of biological databases. NCBI

2. Convey the necessity of computers in storing and analyzing such a large amount of data. Human Genome Project

3. Curated vs. uncurated databases.

4. Discuss search engines and search design: specificity vs. sensitivity
Databases

1. Choose exercises or background reading that stresses the size and growth rate of biological databases.  NCBI

2. Convey the necessity of computers in storing and analyzing such a large amount of data.  Human Genome Project

3. Curated vs. uncurated databases.  Refseq vs. Swissprot

4. Discuss search engines and search design: specificity vs. sensitivity
Databases

1. Choose exercises or background reading that stresses the size and growth rate of biological databases.  NCBI

2. Convey the necessity of computers in storing and analyzing such a large amount of data. Human Genome Project

3. Curated vs. uncurated databases. Refseq vs. Swissprot

4. Discuss search engines and search design: specificity vs. sensitivity Entrez
1. Choose exercises or background reading that stresses the size and growth rate of biological databases.

NCBI

International sequence databases exceed 100 gigabases

In August 2005, the INSDC announced the DNA sequence database exceeded 100 gigabases. GenBank is proud of its contributions toward this milestone. We thank all the scientists who have worked through the submission process at GenBank and made their sequence data available to the world. See the related press release.
2. Convey the necessity of computers in storing and analyzing such a large amount of data.

**Human Genome Project**

Begun formally in 1990, the U.S. Human Genome Project was a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the completion date to 2003. Project goals were to:

- **identify** all the approximately 20,000-25,000 genes in human DNA,
- **determine** the sequences of the 3 billion chemical base pairs that make up human DNA,
- **store** this information in databases,
- **improve** tools for data analysis,
- **transfer** related technologies to the private sector, and
- **address** the ethical, legal, and social issues (ELSI) that may arise from the project.
3. Curated vs. uncurated databases.

**Refseq vs. Swissprot**

**BIOLOGY WORKBENCH:**

**Databases selected:** GENPEPT

Matches (0 to 10) / 187

RESULTS OF (hemoglobin AND gamma)

**Databases selected:** SWISSPROT

Matches (0 to 10) / 118

RESULTS OF (hemoglobin AND gamma)
4. Discuss search engines and search design: specificity vs. sensitivity **Entrez PubMed**

- Verapamil-sensitive Ca(2+) channel regulation of Th1-type proliferation of splenic lymphocytes induced by Walker 256 tumor development in rats.
  - PMID: 16996495 [PubMed - in process]
4. Discuss search engines and search design: specificity vs. sensitivity **Entrez PubMed**
URLs or google searches for bioinformatics students:

The Human Genome Project:
http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

The Human Genome Sequencing Center at Baylor College of Medicine
http://www.hgsc.bcm.tmc.edu/

Cells Alive:
www.cellsalive.com

The Biology Workbench:
http://workbench.sdsc.edu/

National Center for Biotechnology Information:

Expasy (Swiss Institute of Bioinformatics)
http://us.expasy.org/tools/

European Bioinformatics Institute
http://www.ebi.ac.uk/Tools/
How many bases in a genome?
How many bases in a genome?

Human
Rat
Chicken
Fish
Tuberculosis (bacteria)
How many bases in a genome?

Human
Rat
Chicken
Fish
Tuberculosis (bacteria)
How many bases in a genome?

Human
Rat
Chicken
Fish
Tuberculosis
(bacteria)

~ 3 billion
~ 3 billion
How many bases in a genome?

Human \hspace{1cm} \sim 3 \text{ billion}
Rat \hspace{1cm} \sim 3 \text{ billion}
Chicken \hspace{1cm} \sim 1 \text{ billion}
Fish
Tuberculosis \hspace{1cm} \text{(bacteria)}
How many bases in a genome?

- Human: ~3 billion
- Rat: ~3 billion
- Chicken: ~1 billion
- Fish: ~400 million
- Tuberculosis (bacteria)
How many bases in a genome?

Human  ~ 3 billion
Rat    ~ 3 billion
Chicken ~ 1 billion
Fish   ~ 400 million
Tuberculosis (bacteria) ~ 4 million
Emphasize the factors that contribute to the dependence of biological studies on computers

How many bases in a genome?

Human  ~ 3 billion
Rat    ~ 3 billion
Chicken ~ 1 billion
Fish   ~ 400 million
Tuberculosis (bacteria) ~ 4 million
“About the Human Genome Project”

Question 1: How many genes are found in the human genome?

Question 2: How many DNA base pairs make up the human genome?

Question 3: Name 2 project goals that will require the help of computers.
What is the Human Genome Project?

Begun formally in 1990, the U.S. Human Genome Project was a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the completion date to 2003. Project goals were to

- identify all the approximately 20,000-25,000 genes in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA,
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

Question 1: How many genes are found in the human genome?

Question 2: How many DNA base pairs make up the human genome?

Question 3: Name 2 project goals that will require the help of computers.
Question 1: How many genes are found in the human genome?

~ 20,000 - 25,000

Question 2: How many DNA base pairs make up the human genome?

Question 3: Name 2 project goals that will require the help of computers.
Question 1: How many genes are found in the human genome?

~ 20,000 - 25,000

Question 2: How many DNA base pairs make up the human genome?

~ 3 billion

Question 3: Name 2 project goals that will require the help of computers.
What is the Human Genome Project?

Begun formally in 1990, the U.S. Human Genome Project was a 13-year effort coordinated by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances accelerated the completion date to 2003. Project goals were to:

- identify all the approximately 20,000-25,000 genes in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA,
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

Question 1: How many genes are found in the human genome?

~ 20,000 - 25,000

Question 2: How many DNA base pairs make up the human genome?

~ 3 billion

Question 3: Name 2 project goals that will require the help of computers.

1. store this information in databases
2. tools for data analysis
Is the Houston Medical Center involved in genomics?
Is the Houston Medical Center involved in genomics?

http://www.hgsc.bcm.tmc.edu/
Is the Houston Medical Center involved in genomics?

http://www.hgsc.bcm.tmc.edu/
Chimpanzee Genomic Analysis

About the project

The Human Genome Sequencing Center has sequence chimpanzee, *Pan troglodytes*. This data is available for

The [NHGRI](https://www.nhgri.nih.gov) has funded a [chimpanzee genome project](https://www.nhgri.nih.gov) and [WUGSC](https://www.wugsc.org).

[International Genome Consortium Database](https://www.internationalgenome.org)

[Conditions for use](https://www.internationalgenome.org/terms)
• How many genome projects are being sequenced for different organisms at the Human Genome Sequencing Center, Baylor College of Medicine?

• How many primate genome projects are listed?

• Why do you think so many primate genomes are being sequenced?

• Why is it important to humans to learn about bovine genomes?

• Why is it important to humans to learn about microbial genomes?
Red Blood Cells

A human red cell has no nucleus and is responsible for carrying oxygen to the tissues. Read more about red cells and other images in the Gallery.
The student can look up the answers to related cell biology questions at cellsalive.com, example: Where are genes located?

Animal Cell
The student can look up the answers to related cell biology questions at cellsalive.com, example: Where are genes located?

Animal Cell

NUCLEUS (with DNA inside)
The student can look up the answers to related cell biology questions at cellsalive.com, example: Where are genes located?

Animal Cell

Molecular chains of Deoxyribonucleic Acid (DNA) inside each cell encode the organism’s genes. They are the hereditary information that determines what characteristics each cell, and, in a bigger sense, each organism will have.
Sequence comparison

Healthy vs. diseased

One organism vs. another

Unknown function vs. known
Sequence comparison

Healthy vs. diseased
Identify genes involved in diseases

One organism vs. another

Unknown function vs. known
Sequence comparison

Healthy vs. diseased
Identify genes involved in diseases

One organism vs. another
How closely related are two organisms

Unknown function vs. known
Sequence comparison

Healthy vs. diseased
Identify genes involved in diseases

One organism vs. another
How closely related are two organisms

Unknown function vs. known
Lots of genes are not understood
Sequence comparison

One protein within a family vs. another
Sequence comparison

One protein within a family vs. another

Identify mechanisms of disease, identify favorable characteristics (stability, specificity of substrate, affinity for substrate, etc.)
Vocabulary

If the same letter occurs in two aligned sequences then this position has been **conserved** in evolution. If the letters differ it is assumed that the two derive from an ancestral letter (which could be one of the two or neither).

Evolutionary processes in biology can introduce **insertions** or **deletions** in sequences.

In a sequence alignment, a letter or a stretch of letters may be paired up with dashes in the other sequence, called **gaps**, to signify an insertion or deletion.

If a biologist makes the statement that two sequences are related, he means that they are believed to have a **common evolutionary origin**.
The residues in aligned positions of different sequences are implied to have a common evolutionary origin.

<table>
<thead>
<tr>
<th>LGBA_Soybn</th>
<th>VAFTEKQDALVSSSFEEAFKANIPQYSVVFYTSILEKAPAAKDLFLSFLANGVDPTNPKLTLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYG_Phyca</td>
<td>V-ISEGEWQLVAVMPKVEADVAGHGDILIRLFKSHPETLEKDRFKHLKTEAEMKASE</td>
</tr>
<tr>
<td>consensus</td>
<td>Va-tE----LV----f-----A-i----------i-----P-----d-F-------------K-t-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LGBA_Soybn</th>
<th>HAELLFALVRDSAGQLKASQTVVAD-AALGSSVHAQNAVTDFQFV-VKEALIKTIKAA-</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYG_Phyca</td>
<td>DALKHGVTULTALGAILKKKCHHEALEKPLAQSHATKHKIPIKYLEFISEFAIHVLHSHRH</td>
</tr>
<tr>
<td>consensus</td>
<td>----K----V-----G-iLK--G--Adl--Lg--HA--K--------fve-v-EAll--------h</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LGBA_Soybn</th>
<th>VGDKWSDE----LSRAWDEVAYDELAAAIK----KA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYG_Phyca</td>
<td>PGDFGADAQGAMNKALELFRKDIAAKY-KELGYQG</td>
</tr>
<tr>
<td>consensus</td>
<td>-GD-----D-qgal-rA-Ev-----elAA-----Kelgy-a</td>
</tr>
</tbody>
</table>
The residues in aligned positions of different sequences are implied to have a common evolutionary origin.
12 GAPS (yellow)

(an insertion in one sequence or a deletion in the other sequence?)

The residues in aligned positions of different sequences are implied to have a common evolutionary origin.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Alignment</th>
<th>Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGBA_SOYBN</td>
<td>VAFTekQDALVSSSFEEAFKNIPQYSVVFYTSTSILEKAPAAKDLFSLANGVDPTNPKLTLG</td>
<td></td>
</tr>
<tr>
<td>MYG_PHYCA</td>
<td>V-ISEGEWQLVHLHVMAIADVEADVAGHODILIRLFKSHPETLEKRDFKHLKTEAEKMKGSE</td>
<td></td>
</tr>
<tr>
<td>consensus</td>
<td>Va-tE----LV-f---A-i----i----i---P----d-F-------K-t-</td>
<td></td>
</tr>
<tr>
<td>LGBA_SOYBN</td>
<td>HAELLFALVRDSAGQ-LKASGTVVAD-AALGSVHAQNAVTDPOFV-VVK EALLIKTIKAATV</td>
<td></td>
</tr>
<tr>
<td>MYG_PHYCA</td>
<td>DLKKGHTVULTALGAILKKKGHHEAEKLPLAQSHTKHKIPIKYLEFISEAEIIHVHLSRH</td>
<td></td>
</tr>
<tr>
<td>consensus</td>
<td>----K----V----G-iLK--G---Adl--Lg--HA-K-------fve-v-EAll--i----h</td>
<td></td>
</tr>
<tr>
<td>LGBA_SOYBN</td>
<td>VGEKWSDE----LSRAWEVAYDELAAAIAK----KA</td>
<td></td>
</tr>
<tr>
<td>MYG_PHYCA</td>
<td>PGDFGADAQGAMNKELELFKDIAAKYKELGYOG</td>
<td></td>
</tr>
<tr>
<td>consensus</td>
<td>-GD---D-qgal-rA-Ev---elAA--Kelgy-a</td>
<td></td>
</tr>
</tbody>
</table>
12 GAPS (yellow)
(an insertion in one sequence or a deletion in the other sequence?)

The residues in aligned positions of different sequences are implied to have a common evolutionary origin

LGBA_SOYBN  VAFTEKQDALVSSSFTEAFKKNIPQYSVVFYTSILEKAPAADKLFSFLANGVDPTNPKLTG
MYG_PHYCA  V--ISEGEWOLVLHVWAILVEADVAGHODILIRLFKLHPETELKIDRFLKHLKTEAEMKASE
consensus  Va-tE-----LV---f-----A-i-----i-----P----d-F--------K-t-

LGBA_SOYBN  HAEELFALVRDSAGQ-LKASGTVVAD-AALGSVHAOKAVTDPQVF-VKKEALIKTIKAA-
MYG_PHYCA  DLKTHGVTULTALGAILKKGHEAEKLPLAQSAHATKHKIPIKYLEFISEFAIIHLHSRH
consensus  ---K----V-----G-iLK--G---Adl--Lg--HA--K------fve-v-EAll--i----h

LGBA_SOYBN  VGEKSWSDE---LSRAWQEVAYDELAAAIK-----KA
MYG_PHYCA  PGEFGADAQGAMNKALELFKDIAAKYKEELGYQG
consensus  -GD---D-qgal-rA-Ev---elAA--Kelgy-a

29 identities (green)
12 GAPS (yellow) (an insertion in one sequence or a deletion in the other sequence?)

The residues in aligned positions of different sequences are implied to have a common evolutionary origin

29 identities (green)  20 similarities (cyan)
Identity
indicates exact match in two (or more) sequences

Similarity
indicates chemical or structural similarity between unidentical aligned residues in two (or more) sequences

Homology
the source of the similarity between unidentical aligned residues in two (or more) sequences is biological, such as evolutionarily related sequences in different species (same origin and function) or relationship between members of a chromosome pair in diploid organisms (homologous sequences are similar, but similar sequences are not always homologous)
**Specificity**
The ability to reject false relationships, measured by the ratio of the number of true negatives to the sum of false positives and true negatives.

\[
\text{true negatives} \\
(\text{true negatives} + \text{false positives})
\]

**Sensitivity**
The ability to detect all true relationships, measured by the ratio of the number of true positives to the sum of true positives and false negatives.

\[
\text{true positives} \\
(\text{true positives} + \text{false negatives})
\]
Studying distantly related sequences:

1. Use protein sequence.

Studying closely related sequences (identity, homology, paralogy):

1. Nucleotide sequence might be preferred (can see subtle changes that might be invisible in protein sequences)
use protein sequences rather than DNA when possible *(why?)*
use protein sequences rather than DNA when possible (why?)

Higher signal to noise ratio in protein sequences - what are the causes?
use protein sequences rather than DNA when possible (why?)

Higher signal to noise ratio in protein sequences - what are the causes?

I. Mathematical Probability:
   From a strictly mathematical point of view, assuming that there is an equal likelihood of any nucleotide appearing at any point in a sequence (which is generally NOT true biologically), what are the chances that a G in a nucleotide sequence will be randomly matched by a G in the same position in a different sequence?

   From the same point of view, what are the chances that a G in a protein sequence will be randomly matched by a G in the same position in a different sequence?
use protein sequences rather than DNA when possible (why?)

Higher signal to noise ratio in protein sequences - what are the causes?

I. Mathematical Probability:
   From a strictly mathematical point of view, assuming that there is an equal likelihood of any nucleotide appearing at any point in a sequence (which is generally NOT true biologically), what are the chances that a G in a nucleotide sequence will be randomly matched by a G in the same position in a different sequence? 1/4

From the same point of view, what are the chances that a G in a protein sequence will be randomly matched by a G in the same position in a different sequence?
use protein sequences rather than DNA when possible (why?)

Higher signal to noise ratio in protein sequences - what are the causes?

I. Mathematical Probability:
   From a strictly mathematical point of view, assuming that there is an equal likelihood of any nucleotide appearing at any point in a sequence (which is generally NOT true biologically), what are the chances that a G in a nucleotide sequence will be randomly matched by a G in the same position in a different sequence? 1/4

   From the same point of view, what are the chances that a G in a protein sequence will be randomly matched by a G in the same position in a different sequence? 1/20
Higher signal to noise ratio in protein sequences - what are the causes?

II. Degeneracy of the genetic code:

a. 18 of the 20 amino acids are coded for by > one codon - therefore, a single mutation in the DNA code does not necessarily translate into a change in the amino acid code (particularly true of mutations in the 3rd codon)

b. a single change within a triplet codon is often not sufficient to cause a codon to code for an amino acid in a different category (nonpolar, polar, positively charged, negatively charged)
Higher signal to noise ratio in protein sequences - what are the causes?

II. Degeneracy of the genetic code:

a. 18 of the 20 amino acids are coded for by > one codon - therefore, a single mutation in the DNA code does not necessarily translate into a change in the amino acid code (particularly true of mutations in the 3rd codon)

   **UUC to UUU mutation:**

b. a single change within a triplet codon is often not sufficient to cause a codon to code for an amino acid in a different category (nonpolar, polar, positively charged, negatively charged)
Higher signal to noise ratio in protein sequences - what are the causes?

II. Degeneracy of the genetic code:

a. 18 of the 20 amino acids are coded for by > one codon - therefore, a single mutation in the DNA code does not necessarily translate into a change in the amino acid code (particularly true of mutations in the 3rd codon)

   **UUC to UUU mutation:** UUC encodes PHE (F)

b. a single change within a triplet codon is often not sufficient to cause a codon to code for an amino acid in a different category (nonpolar, polar, positively charged, negatively charged)
Higher signal to noise ratio in protein sequences - what are the causes?

II. Degeneracy of the genetic code:

a. 18 of the 20 amino acids are coded for by > one codon - therefore, a single mutation in the DNA code does not necessarily translate into a change in the amino acid code (particularly true of mutations in the 3rd codon)

   **UUC to UUU mutation:**  
   - UUC encodes PHE (F)
   - UUU encodes PHE (F)

b. a single change within a triplet codon is often not sufficient to cause a codon to code for an amino acid in a different category (nonpolar, polar, positively charged, negatively charged)
Higher signal to noise ratio in protein sequences - what are the causes?

II. Degeneracy of the genetic code:

a. 18 of the 20 amino acids are coded for by > one codon - therefore, a single mutation in the DNA code does not necessarily translate into a change in the amino acid code (particularly true of mutations in the 3rd codon)

UUC to UUU mutation: UUC encodes PHE (F)
UUU encodes PHE (F)

b. a single change within a triplet codon is often not sufficient to cause a codon to code for an amino acid in a different category (nonpolar, polar, positively charged, negatively charged)

AAG to AGG mutation:
Higher signal to noise ratio in protein sequences - what are the causes?

II. Degeneracy of the genetic code:

a. 18 of the 20 amino acids are coded for by > one codon - therefore, a single mutation in the DNA code does not necessarily translate into a change in the amino acid code (particularly true of mutations in the 3rd codon)

   **UUC to UUU mutation:**
   - UUC encodes PHE (F)
   - UUU encodes PHE (F)

b. a single change within a triplet codon is often not sufficient to cause a codon to code for an amino acid in a different category (nonpolar, polar, positively charged, negatively charged)

   **AAG to AGG mutation:**
   - AAG encodes LYS (K)
Higher signal to noise ratio in protein sequences - what are the causes?

II. Degeneracy of the genetic code:

a. 18 of the 20 amino acids are coded for by > one codon - therefore, a single mutation in the DNA code does not necessarily translate into a change in the amino acid code (particularly true of mutations in the 3rd codon)

  UUC to UUU mutation: UUC encodes PHE (F)
                      UUU encodes PHE (F)

b. a single change within a triplet codon is often not sufficient to cause a codon to code for an amino acid in a different category (nonpolar, polar, positively charged, negatively charged)

  AAG to AGG mutation: AAG encodes LYS (K)
                      AGG encodes ARG (R)
Higher signal to noise ratio in protein sequences - what are the causes?

Similarity “signals” contribute more information in protein sequences than in nucleotide sequences

a. Many categories, some can be weighted more heavily than others (nonpolar, polar, positively charged, negatively charged, aromatic, structural similarity)

b. Nucleotides -
   - transitions purine to purine, pyrimidine to pyrimidine
   - transversions purine to pyrimidine, pyrimidine to purine

![Diagram of nucleotides]
Sequence comparison

Healthy vs. diseased

Identify genes involved in diseases

One organism vs. another

How closely related are two organisms

Unknown function vs. known

Lots of genes are not understood
Sequence comparison

Healthy vs. diseased

Identify genes involved in diseases

One organism vs. another

How closely related are two organisms

Unknown function vs. known

Lots of genes are not understood
Proteins involved in genetic diseases
Proteins involved in genetic diseases

Lactase - *digests milk sugar.*
Proteins involved in genetic diseases

Lactase - *digests milk sugar.*
Insulin receptor - *mediates the proper response to glucose.*
Proteins involved in genetic diseases

Lactase - *digests* milk sugar.
Insulin receptor - *mediates* the proper response to glucose.
P53 protein - *tumor suppressor*. 
Exercise in Sequence Alignment:
Our example is HbB vs. HbS
Type the following web site into your browser:
Next to the “Search” box, select Protein, to search the NCBI database containing protein sequences.
Exercise in Sequence Alignment:
Our example is HbB vs. HbS
Type the following web site into your browser:
Next to the “Search” box, select Protein, to search the NCBI database containing protein sequences.
Exercise in Sequence Alignment:
Our example is HbB vs. HbS
Type the following web site into your browser:
Next to the “Search” box, select Protein, to search the NCBI database containing protein sequences.
Exercise in Sequence Alignment:
Our example is HbB vs. HbS
Type the following web site into your browser:
Next to the “Search” box, select Protein, to search the NCBI database containing protein sequences.
The record for hemoglobin S should be returned. Hemoglobin is the protein in our blood cells that carries oxygen. Click on the link entitled “1HBSB”.

1HBSB
Chain B, Hemoglobin S (Deoxy)
gil229959|pdb|1HBSlB[229959]
The record for hemoglobin S should be returned. Hemoglobin is the protein in our blood cells that carries oxygen. Click on the link entitled “1HBSB”.

[Image of a protein database search showing results for 1HBSB protein]
Next to the word Display in the grey region at the top of the file, change “GenPept” to “FASTA”.

![NCBI interface screenshot](image)
Next to the word Display in the grey region at the top of the file, change “GenPept” to “FASTA”.
This will display the amino acid sequence for hemoglobin S in FASTA format.

>gi|229959|pdb|1HBS|B Chain B, Hemoglobin S (Deoxy)
VHLTPVEKSAVTALWGKVNDEVGGEALGRLLVYYPWTQRFSESFGDLSSTPDAVMGNPKVKAHGKKVLGA
FSDGLAHLDNLKGFATLSELHCDDLHVDPENFRLLGNVLVCVLAHHFGEPTPPVQAAYQKVAVANALAHKYH
This will display the amino acid sequence for hemoglobin S in FASTA format.

1: 1HBSB. Reports Chain B, Hemoglobin...[gi:229959]

>gi|229959|pdb|1HBS|B Chain B, Hemoglobin S (Deoxy)
VHLTPVEKSAVTALWGKVNDEVGGEALGRLLVYYPWTQRFESFGDLSYTPDAMGNPKVKAHGKKVLGA
FSDGLAHLDNLKGTATLSLHCDKLVDPENRLLGNVLVCVLAHHFGKEFTPVPVQAAYQKVAGVANA
LAHKYH
This will display the amino acid sequence for hemoglobin S in FASTA format.

>gi|229959|pdb|1HBS|B Chain B, Hemoglobin S (Deoxy)
VHLTPVEKSAVTALWGKVNVDEVGGEALGRLLVYYPWTQRFFESFGDLSTPDAVMGNPVKAHGKKVLGA
FDGGLAHLDNLKGTFTALSELHCDELHVDPENFRLGNVLVCLAHHPGKEFTPPVQAAYQKVAGVANA
LAHKYH
Hold down the left mouse button while you move the mouse over the sequence. This should highlight the amino acid sequence in blue. Now choose “Edit:Copy” from the browser window, or hit the buttons “Ctrl” and “C” to copy.

1: IHBSB. Reports Chain B, Hemoglobin...[gi:229959]

>gi|229959|pdb|1HBS|B Chain B, Hemoglobin S (Deoxy)
VHLTPVEKSAVTLWGKVNVDEVGGEALGRLLVYVPWTQRRFESFGDLSTPDAVMGNPKVKAHGKKVLGA
FSGLAMHLDNLKGFATLSHELCDKLHVDPENPRLGNVLCVLHAFGKEFTPVPQAAYQKVAGVANA
LAHKYH
Now, click on the NCBI logo in the upper left corner of the web page to return to the main page.
Now, click on the NCBI logo in the upper left corner of the web page to return to the main page.
In the dark blue menu bar at the top of the page, click on the word “BLAST”.
In the dark blue menu bar at the top of the page, click on the word “BLAST”.
In the box of Protein options, click on the link entitled “Protein-protein BLAST (blastp)”.

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)
In the box of Protein options, click on the link entitled “Protein-protein BLAST (blastp)”. 
Click in the Search box and choose “Edit: paste” from the browser menu or hit the “Ctrl” and “P” keys to paste the sequence into the search box.
Click in the Search box and choose “Edit: paste” from the browser menu or hit the “Ctrl” and “P” keys to paste the sequence into the search box.

Sequence:

```plaintext
>gi|229959|pdb|1HBS| B Chain B, Hemoglobin S (Deoxy)
VHLPVEKSAVTALWGKNVDEVGGEALGRLLVYYPWTQRFSEFGDLSTPDAVMGNPKVKA
FSDGLAHLDNLKGTFATLSHELCDKLHVDPENFRLLGNVLCVLHAHFGKEFTPVPVQAAYQK
LAHYH
```
Change the “nr” database to “swissprot”, then click the BLAST! button.
Change the “nr” database to “swissprot”, then click the BLAST! button.
Change the “nr” database to “swissprot”, then click the BLAST! button.
Click the Format! button.

Putative conserved domains have been detected, click on the image below for detailed results.

The request ID is 1140215936-8025-124841773514.BLASTQ4

A new window will open containing our sequence alignments.
Click the Format! button.

Putative conserved domains have been detected, click on the image below for detailed results.

The request ID is 1140215936-8025-124841773514.BLASTQ4

A new window will open containing our sequence alignments.
Under the graph indicating the length of the top alignments, there will be a list of aligning sequences in order of decreasing alignment scores. Click on the score of the first item in the list, which is the highest scoring alignment. This will take you to the section of the file where you can view the alignment.

Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>gi</th>
<th>56749858</th>
<th>sp</th>
<th>P68873</th>
<th>HBB_PANT</th>
<th>Score (Bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>232230</td>
<td>sp</td>
<td>P02024</td>
<td>HBB_GORGO</td>
<td>300</td>
<td>9e-82</td>
</tr>
<tr>
<td>gi</td>
<td>122528</td>
<td>sp</td>
<td>P18988</td>
<td>HBB2_PANLE</td>
<td>299</td>
<td>2e-81</td>
</tr>
<tr>
<td>gi</td>
<td>122616</td>
<td>sp</td>
<td>P02025</td>
<td>HBB_HYLLA</td>
<td>296</td>
<td>1e-80</td>
</tr>
<tr>
<td>gi</td>
<td>122668</td>
<td>sp</td>
<td>P02032</td>
<td>HBB_PREEN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>122593</td>
<td>sp</td>
<td>P19885</td>
<td>HBB_COLPO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td>122584</td>
<td>sp</td>
<td>P02028</td>
<td>HBB_CERAE</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Under the graph indicating the length of the top alignments, there will be a list of aligning sequences in order of decreasing alignment scores. Click on the score of the first item in the list, which is the highest scoring alignment. This will take you to the section of the file where you can view the alignment.

<table>
<thead>
<tr>
<th>Sequence ID</th>
<th>Score (Bits)</th>
<th>E Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>56749858</td>
<td>sp</td>
</tr>
<tr>
<td>gi</td>
<td>232230</td>
<td>sp</td>
</tr>
<tr>
<td>gi</td>
<td>122528</td>
<td>sp</td>
</tr>
<tr>
<td>gi</td>
<td>122616</td>
<td>sp</td>
</tr>
<tr>
<td>gi</td>
<td>122668</td>
<td>sp</td>
</tr>
<tr>
<td>gi</td>
<td>122593</td>
<td>sp</td>
</tr>
<tr>
<td>gi</td>
<td>122584</td>
<td>sp</td>
</tr>
</tbody>
</table>
Identify the differences in the sequence of Query 1 and Subject 2

> gi|56749858|sp|P68873|HBB_PANTR  Hemoglobin beta subunit (Hemoglobin beta chain)  
> gi|56749857|sp|P68872|HBB_PANPA  Hemoglobin beta subunit (Hemoglobin beta chain)  
> gi|56749856|sp|P68871|HBB_HUMAN  Hemoglobin beta subunit (Hemoglobin beta chain)  
Length=147

Score = 300 bits (768),  Expect = 9e-82  
Identities = 145/146 (99%),  Positives = 145/146 (99%),  Gaps = 0/146 (0%)

<table>
<thead>
<tr>
<th>Query</th>
<th>Sbjct</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>61</td>
<td>62</td>
<td>120</td>
</tr>
<tr>
<td>121</td>
<td>122</td>
<td>146</td>
</tr>
</tbody>
</table>

Sequence differences:

<table>
<thead>
<tr>
<th>Query 1</th>
<th>Sbjct 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>VHLTPVEKSAVTALWGKVNDEVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV</td>
<td>VHLTP EKSAVTALWGKVNDEVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV</td>
</tr>
<tr>
<td>KAHGKKVLGAFSDGLAHLDNLKGTATLSELHCDKLHVDPENFRLLNVLVCVLAHHF</td>
<td>KAHGKKVLGAFSDGLAHLDNLKGTATLSELHCDKLHVDPENFRLLNVLVCVLAHHF</td>
</tr>
<tr>
<td>EFTPPVQAAYQKVVGVANALAHKYH</td>
<td>EFTPPVQAAYQKVVGVANALAHKYH</td>
</tr>
<tr>
<td>EFTPPVQAAYQKVVGVANALAHKYH</td>
<td>EFTPPVQAAYQKVVGVANALAHKYH</td>
</tr>
</tbody>
</table>
Identify the differences in the sequence of Query 1 and Subject 2

> gi|56749857|sp|P68873|HBB_PANTR  Hemoglobin beta subunit (Hemoglobin beta chain)
> gi|56749857|sp|P68872|HBB_PANPA  Hemoglobin beta subunit (Hemoglobin beta chain)
> gi|56749856|sp|P68871|HBB_HUMAN  G  Hemoglobin beta subunit (Hemoglobin beta chain)
Length=147

Score = 300 bits (768), Expect = 9e-82
Identities = 145/146 (99%), Positives = 145/146 (99%), Gaps = 0/146 (0%)

Query 1  VHLTPVEKSAVTALWGKNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKV  60
        VHLP VEKSAVTALWGKNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKV
Sbjct 2  VHLTPVEKSAVTALWGKNVDEVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKV  61

Query 61  KAHGKVLGFSDLGNDNLKGTATLSELHCDKLHVDPEFRLLLGNNVLCVLALHFGK  120
        KAHGKVLGFSDLGNDNLKGTATLSELHCDKLHVDPEFRLLLGNNVLCVLALHFGK
Sbjct 62  KAHGKVLGFSDLGNDNLKGTATLSELHCDKLHVDPEFRLLLGNNVLCVLALHFGK  121

Query 121  EFTPPVQAAYQKVAGVANALAHKYH  146
        EFTPPVQAAYQKVAGVANALAHKYH
Sbjct 122  EFTPPVQAAYQKVAGVANALAHKYH  147
Identify the differences in the sequence of Query 1 and Subject 2

Score = 300 bits (768), Expect = 9e-82
Identities = 145/146 (99%), Positives = 145/146 (99%), Gaps = 0/146 (0%)

Query 1
VHLTPVEKSAVTALWGVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVGMNPKV 60
VHLTP EKSAVTALWGVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVGMNPKV

Subject 2
VHLTPVEKSAVTALWGVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVGMNPKV 61

Query 61
KAHGKKVLGAFSDGLAHLDNLKGTFTATLSELHCDKLHVDPENFRRLLGNVLVCVLAHFFGK 120
KAHGKKVLGAFSDGLAHLDNLKGTFTATLSELHCDKLHVDPENFRRLLGNVLVCVLAHFFGK

Subject 62
KAHGKKVLGAFSDGLAHLDNLKGTFTATLSELHCDKLHVDPENFRRLLGNVLVCVLAHFFGK 121

Query 121
EFTPPVQAAYQKVAGVANALAHKYH 146
EFTPPVQAAYQKVAGVANALAHKYH

Subject 122
EFTPPVQAAYQKVAGVANALAHKYH 147

A dissimilar substitution occurs at amino acid number 6.
Glu6 to Val
The $\beta$ E6V mutation dramatically decreases the solubility of hemoglobin and causes long fibers to form that sickle the red cells. This was the first genetic disease that was identified at the amino acid level.
The β E6V mutation dramatically decreases the solubility of hemoglobin and causes long fibers to form that sickle the red cells. This was the first genetic disease that was identified at the amino acid level.
The $\beta$ E6V mutation dramatically decreases the solubility of hemoglobin and causes long fibers to form that sickle the red cells. This was the first genetic disease that was identified at the amino acid level.

Hb AS is benign (only shortens red cell life time to ~80 days)

<table>
<thead>
<tr>
<th>Chromosome 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS 1-6</td>
</tr>
<tr>
<td>$\epsilon$</td>
</tr>
<tr>
<td>$\gamma_A$</td>
</tr>
<tr>
<td>$\gamma_G$</td>
</tr>
<tr>
<td>$\delta$</td>
</tr>
<tr>
<td>$\beta$</td>
</tr>
<tr>
<td>$\beta_s$</td>
</tr>
</tbody>
</table>

Hb SS leads to severe anemia and early death

<table>
<thead>
<tr>
<th>Chromosome 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS 1-6</td>
</tr>
<tr>
<td>$\epsilon$</td>
</tr>
<tr>
<td>$\gamma_A$</td>
</tr>
<tr>
<td>$\gamma_G$</td>
</tr>
<tr>
<td>$\delta$</td>
</tr>
<tr>
<td>$\beta_s$</td>
</tr>
</tbody>
</table>
The β E6V mutation dramatically decreases the solubility of hemoglobin and causes long fibers to form that sickle the red cells. This was the first genetic disease that was identified at the amino acid level.

Hb AS is benign (only shortens red cell life time to ~80 days)

Chromosome 11

Hb SS leads to severe anemia and early death

Chromosome 11
The sickle cell mutation in Hemoglobin.

Sickle cell anemia is a blood condition seen most commonly in people of African ancestry and in the tribal peoples of India.

The individual must have two copies of the mutant hemoglobin gene to exhibit the sickle-shaped cells indicative of the condition.

The hemoglobin S beta subunit has the amino acid valine at position 6 instead of the glutamic acid that is normally present. This alteration is the basis of all the problems that occur in people with sickle cell disease.
ClustalW is a multiple sequence alignment routine available online at the EBI website: [http://www.ebi.ac.uk/clustalw/](http://www.ebi.ac.uk/clustalw/)
Exercise in
Multiple Sequence Alignment:
Our example is non-alpha versus alpha Hb

The Biology Workbench
http://workbench.sdsc.edu/

The Biology Workbench is one of my favorite teaching tools, because the student can do a complete Bioinformatics project with the Workbench, from retrieving the sequences to performing multiple alignments and creating phylogenetic tree diagrams.

The Biology WorkBench is a web-based tool for biologists. The WorkBench allows biologists to search many popular protein and nucleic acid sequence databases. Database searching is integrated with access to a wide variety of analysis and modeling tools, all within a point and click interface that eliminates file format compatibility problems.

First time users: please register for a free account.

Click to Enter the Biology Workbench 3.2
Exercise in
Multiple Sequence Alignment:
Our example is non-alpha versus alpha Hb

The Biology Workbench
http://workbench.sdsc.edu/

The Biology Workbench is one of my favorite teaching tools, because the student can do a complete Bioinformatics project with the Workbench, from retrieving the sequences to performing multiple alignments and creating phylogenetic tree diagrams.

Biology WorkBench

The Biology WorkBench is a web-based tool for biologists. The WorkBench allows biologists to search many popular protein and nucleic acid sequence databases. Database searching is integrated with access to a wide variety of analysis and modeling tools, all within a point and click interface that eliminates file format compatibility problems.

First time users: please register for a free account.

Click to Enter the Biology Workbench 3.2
1. Retrieve sequences

a. Be careful, there are many databases - too much information - too many results from a query confuses the student
b. GenPept - Genbank gene products - full release
c. SwissProt - manually curated European database
<table>
<thead>
<tr>
<th>AAINDEX_1</th>
<th>AAINDEX_2</th>
<th>AFCS</th>
<th>BLOCKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBCAT</td>
<td>ECDC</td>
<td>EPD</td>
<td>FLYBASE</td>
</tr>
<tr>
<td>GBBCT</td>
<td>GBEST1</td>
<td>GBEST10</td>
<td>GBEST11</td>
</tr>
<tr>
<td>GBEST12</td>
<td>GBEST13</td>
<td>GBEST14</td>
<td>GBEST15</td>
</tr>
<tr>
<td>GBEST16</td>
<td>GBEST17</td>
<td>GBEST18</td>
<td>GBEST2</td>
</tr>
<tr>
<td>GBEST3</td>
<td>GBEST4</td>
<td>GBEST5</td>
<td>GBEST6</td>
</tr>
<tr>
<td>GBEST7</td>
<td>GBEST8</td>
<td>GBEST9</td>
<td>GBHTC</td>
</tr>
<tr>
<td>GBINV</td>
<td>GBMAM</td>
<td>GBNEW</td>
<td>GBPAT</td>
</tr>
<tr>
<td>GBPHG</td>
<td>GBPLN</td>
<td>GBPRI</td>
<td>GBREFFUNGI</td>
</tr>
<tr>
<td>GBREFINV</td>
<td>GBREFMAM</td>
<td>GBREFMICROBIAL</td>
<td>GBREFMITO</td>
</tr>
<tr>
<td>GBREFPLANT</td>
<td>GBREFPLASMID</td>
<td>GBREFPLASTID</td>
<td>GBREFPROTOZOA</td>
</tr>
<tr>
<td>GBREFSEQNEW</td>
<td>GBREFVERT</td>
<td>GBREFVIRAL</td>
<td>GBROD</td>
</tr>
<tr>
<td>GBSTS</td>
<td>GBSYN</td>
<td>GBUNA</td>
<td>GBVRL</td>
</tr>
<tr>
<td>GBVRT</td>
<td>GENPEPT</td>
<td>GENPEPTNEW</td>
<td>KEGG_GENES</td>
</tr>
<tr>
<td>KEGG_ORTHOLOG</td>
<td>KEGG_PATHWAY</td>
<td>LIGAND_COMPOUND</td>
<td>LIGAND_ENZYME</td>
</tr>
<tr>
<td>MHCPEP</td>
<td>NRL3D</td>
<td>OMIM</td>
<td>PDBFINDER</td>
</tr>
<tr>
<td>PDBSEQRES</td>
<td>PFAM</td>
<td>PFAM_B</td>
<td>PRINTS</td>
</tr>
<tr>
<td>PRODOM</td>
<td>PROSITE</td>
<td>PROSITE_DOC</td>
<td>PROSITE_QDOC</td>
</tr>
<tr>
<td>REBASE</td>
<td>REBASE_ABS</td>
<td>REBASE_SRC</td>
<td>SDSCNR</td>
</tr>
<tr>
<td>SEQANALREF</td>
<td>SWISSPROT</td>
<td>TRANSFAC_CELL</td>
<td>TRANSFAC_CLASS</td>
</tr>
</tbody>
</table>
Genpept search of “fetal hemoglobin”: 7 results in over 1 minute
SwissProt search of “fetal hemoglobin”: 59 results in ~ 20 seconds

Databases selected: SWISSPROT

Matches (0 to 10) / 59

RESULTS OF (fetal AND hemoglobin)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th>Matching Database Record</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>60</td>
<td><strong>SWISSPROT:HBG2_HUMAN</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hemoglobin gamma-2 subunit (Hemoglobin gamma-2 chain) (Gamma-2-globin) (Hemoglobin gamma-G chain) (Hb F Ggamma) [Homo sapiens (Human)]</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
<td><strong>SWISSPROT:HBG1_HUMAN</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hemoglobin gamma-1 subunit (Hemoglobin gamma-1 chain) (Gamma-1-globin) (Hemoglobin gamma-A chain) (Hb F Agamma) [Homo sapiens (Human)]</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td><strong>SWISSPROT:HBBF_SHEEP</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hemoglobin beta subunit, fetal (Hemoglobin beta chain, fetal) (Beta-globin, fetal) (Hemoglobin gamma chain) [Ovis aries (Sheep)]</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td><strong>SWISSPROT:HBG_PAPCY</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hemoglobin gamma subunit (Hemoglobin gamma chain) (Gamma-globin) [Papio cynocephalus (Yellow baboon)]</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td><strong>SWISSPROT:HBG_MACMU</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hemoglobin gamma subunit (Hemoglobin gamma chain) (Gamma-globin) [Macaca mulatta (Rhesus macaque)]</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td><strong>SWISSPROT:HBG_MACFU</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hemoglobin gamma subunit (Hemoglobin gamma chain) (Gamma-globin) [Macaca fuscata fuscata (Japanese macaque)]</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td><strong>SWISSPROT:HBBF_CAPHI</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hemoglobin beta subunit, fetal (Hemoglobin beta chain, fetal) (Beta-globin, fetal) (Hemoglobin gamma chain) [Capra hircus (Goat)]</td>
</tr>
</tbody>
</table>
JSO’s slide on globin expression during human development

$\alpha_2 \varepsilon_2$ Hb embryonic

(High $O_2$ affinity)

$\alpha_2 \gamma_2$ HbF

(moderate $O_2$ affinity)

$\alpha_2 \beta_2$ HbA

(low $O_2$ affinity)

$\varepsilon$ subunit

(placenta)

$\gamma$ subunit

(lungs)

$\beta$ subunit
Hb non-alpha subunit alignment

Analysis: HbG1 and HbG2 have an A:G substitution at position 136
HbE has the A at 136, HbB has the G
Why does the HbS sequence have an N-terminal methionine?
Hb alpha and non-alpha subunit alignment

### Sequence alignment

**Consensus key** (see documentation for details)
- * - single, fully conserved residue
- : - conservation of strong groups
- . - conservation of weak groups
- - no consensus

**CLUSTAL W (1.81) multiple sequence alignment**

<table>
<thead>
<tr>
<th></th>
<th>HBG2_HUMAN</th>
<th>HBG1_HUMAN</th>
<th>HBE_HUMAN</th>
<th>HBB_HUMAN</th>
<th>HBS_HUMAN</th>
<th>HBA_HUMAN</th>
<th>HBG2_HUMAN</th>
<th>HBG1_HUMAN</th>
<th>HBE_HUMAN</th>
<th>HBB_HUMAN</th>
<th>HBS_HUMAN</th>
<th>HBA_HUMAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-GHFTEDKATITSLWGKV--NVEDAGGETLGRLLLVYYPWTQRFFDSFGNLSSASAIMGN</td>
<td>-GHFTEDKATITSLWGKV--NVEDAGGETLGRLLLVYYPWTQRFFDSFGNLSSASAIMGN</td>
<td>-VHFTAEKAATVSLWSKM--VVEEAGGGEAALRLLVYYPWTQRFFDSFGNLSSASAIMGN</td>
<td>-VHLTPEEKSAVTALWGKV--NVEVGGEALRGRLVYYPWTQRFFESFGDLSTDPADVMGN</td>
<td>MVHLTVEKSAVTAXWGKV--NVEVGGEALRGRLVYYPWTQRFFESFGDLSTDPADVMGN</td>
<td>--VLSPADKTNVAAWGBKVGAGHEYGAEALERMFLSFPTTKTFPHF-DLS-----HGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>::</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
<td>:::*</td>
</tr>
<tr>
<td></td>
<td>PKVKAHKVKVLTLGDAIKHLDDLGTFQAQLSELHCDKLHVDPENFKLLGNVLVTVAIH</td>
<td>PKVKAHKVKVLTLGDAIKHLDDLGTFQAQLSELHCDKLHVDPENFKLLGNVLVTVAIH</td>
<td>PKVKAHKVKVLTLGDAIKNMNLPFAKSLSELHCDKLHVDPENFKLLGNMVIIALT</td>
<td>PKVKAHKVKVLGAPSGDLHNLKTFATLSLEHCDKLHVDPENFLGNNVLVCLVLAH</td>
<td>PKVKAHKVKVLGAPSGDLHNLKTFATLSLEHCDKLHVDPENFLGNNVLVCLVLAH</td>
<td>PKVKAHKVKVLGAPSGDLHNLKTFATLSLEHCDKLHVDPENFLGNNVLVCLVLAH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AQVKGHKKVDALTNAHVHDDMPNALSALSLDHAKLRVDPVFNFKLSSHCLLVTLAH</td>
<td>AQVKGHKKVDALTNAHVHDDMPNALSALSLDHAKLRVDPVFNFKLSSHCLLVTLAH</td>
<td>AQVKGHKKVDALTNAHVHDDMPNALSALSLDHAKLRVDPVFNFKLSSHCLLVTLAH</td>
<td>AQVKGHKKVDALTNAHVHDDMPNALSALSLDHAKLRVDPVFNFKLSSHCLLVTLAH</td>
<td>AQVKGHKKVDALTNAHVHDDMPNALSALSLDHAKLRVDPVFNFKLSSHCLLVTLAH</td>
<td>AQVKGHKKVDALTNAHVHDDMPNALSALSLDHAKLRVDPVFNFKLSSHCLLVTLAH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FGKEFTPEVQAASWQKMVTGVASALSSRYH</td>
<td>FGKEFTPEVQAASWQKMVTGVASALSSRYH</td>
<td>FGKEFTPEVQAASWQKMVTGVASALSSRYH</td>
<td>FGKEFTPEVQAASWQKMVTGVASALSSRYH</td>
<td>FGKEFTPEVQAASWQKMVTGVASALSSRYH</td>
<td>FGKEFTPEVQAASWQKMVTGVASALSSRYH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **HGS** (highlighted)
Analysis:
Highest scoring pairwise alignments:
1. HbG1 and HbG2
2. HbS and HbB
3. HbE with HbG1, HbG2

Lowest scoring pairwise alignments:
1. HbA and HbE
2. HbA and HbG1, HbG2
3. HbA and HbB, HbS