Fall 2005 Genomics Exam #1
Genomic Sequences

There is no time limit on this test, though I don’t want you to spend too much time on it. I work hard to design challenging tests that continue your learning and hopefully will stimulate you too. You do not need to read any additional papers other than the ones I send to you. There are 4 pages, including this cover sheet, and 5 questions for this test. You are not allowed discuss the test with anyone until all exams are turned in at 11:30 am on Friday September 30. EXAMS ARE DUE AT CLASS TIME ON FRIDAY SEPTEMBER 30. You may use a calculator, a ruler, your notes, the book, and the internet. You may take it in as many blocks of time as you want. Submit your paper and electronic version before 11:30 am (eastern time zone:-). The answers to the questions must be typed in a Word file and emailed to me as an attachment. Be sure to backup your test answers just in case (I suggest a thumb drive or other removable medium). You will need to capture screen images as a part of your answers which you may do without seeking permission since your test answers will not be in the public domain. Remember to explain your thoughts in your own words and use screen shots to support your answers. Screen shots without your words are worth very few points.

You may want to use some of the resources on this page <http://bioinformatics.org/sms/> but you may not need to. Just wanted to supply everyone with a common suite of tools.

DO NOT READ or DOWNLOAD ANY NEW PAPERS FOR THIS EXAM. RELY ONLY ON THE FIGURES PROVIDED IN THIS EXAM, YOUR EXPERIENCE, AND YOUR SKILLS.

-3 pts if you do not follow this direction.

Please do not write or type your name on any page other than this cover page.

Staple all your pages (INCLUDING THE TEST PAGES) together when finished with the exam.

Name (please print):

Write out the full pledge and sign (by typing a second time and signing paper version):

How long did this exam take you to complete (excluding typing)?
20 pts.

1) Access the GRAMENE Genome Browser at this web site ([http://www.gramene.org/](http://www.gramene.org/)).

a. Consider the synteny of rice and maize and make an evaluation of the large-scale genomic changes that have occurred when comparing these two species. When you think you have a good sense of what has happened, use a screen shot of the most extreme example of what you observed to support your conclusions.

From this figure, you can see many maize chromosomes have DNA that map to a single rice chromosome. This is reminiscent of the puffer fish to human synteny comparison.

This is a good example to illustrate the probable duplication that took place in the maize genome.

Based on size alone, it is clear that maize must have undergone at least one duplication:

460 Mb genome = rice
3,000 Mb genome = maize

b. Now search for Histone H3. Tell me where this gene is encoded in the Rice Japonica genome. Use a screen shot to support your findings.

**LOC_Os01g64640** 37493499 - 37494192 bp (37.5 Mb) on chromosome 1
**LOC_Os02g25940** 15170251 - 15170925 bp (15.2 Mb) on chromosome 2
**LOC_Os03g27310** 15633410 - 15636041 bp (15.6 Mb) on chromosome 3
c. Use data to convince me whether histone h3 is highly transcribed or not.
Highly is a relative term, but based on the number of ESTs, it looks like Histone H3 is highly transcribed.

d. Is this gene similarly expressed in other grain plants? How do you know?
Yes, it appears fairly constant across the different species (see the number of ESTs).
e. What strand and which reading frame encodes the h3 protein you have selected? Explain how you found your answer and provide a screen shot of the DNA and protein sequences as shown in GRAMENE that validates your findings.

**LOC_Os01g64640** 37493499 - 37494192 bp (37.5 Mb) on chromosome 1

<table>
<thead>
<tr>
<th>CDS</th>
<th>&gt;11667.m06477</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGGCCCGCACAGCAGACGGCGAGGAAGTCCACCGGCGGCAAGGCGCCGAGGAAGCAG</td>
<td></td>
</tr>
<tr>
<td>CTGGCGACGAAGGCGGCGCGCAAGTCGGCCCCGGCCACCGGCGGCGTGAAGAAGCCCCC</td>
<td></td>
</tr>
<tr>
<td>ACCGACCTCGGCAAGCTCCCGCGGCGCCGTGAGAGGCGGCGAGCTGAGTACAG</td>
<td></td>
</tr>
<tr>
<td>CGTGCTGAGTACGGTCCCCAGCCCGGCGCGTGCGGGAGATCGCGCAGGACTTCAAG</td>
<td></td>
</tr>
<tr>
<td>ACCGACCTCGGCAAGCTCCCGCGGCGCCGTGAGAGGCGGCGAGCTGAGTACAG</td>
<td></td>
</tr>
<tr>
<td>CGTGCTGAGTACGGTCCCCAGCCCGGCGCGTGCGGGAGATCGCGCAGGACTTCAAG</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein</th>
<th>&gt;11667.m06477</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARTKQTARKSTGGKAPRKQLATKAARDSAPATGGVVKPHRFRPTGVALREIRΥΚΚ</td>
<td>see the amino acid sequence MARTK…. just above the cursor arrow. The top strand has the ATG which means the coding strand is on the bottom strand. The reading frame closes to the DNA sequence is the appropriate reading from to start the CDS.</td>
</tr>
</tbody>
</table>

20 pts.

2) Start with this sequence and answer the following questions:

| CCTAGTCCCTCCCTCCCTCCCTGCTGCTGCTACATGTACCGGCACAAGGGCACGTACCACACCAATGAGGCCAAGG |
| GCACGGAGTTTGCTGAGAATGACAGATGCAGACCTGCAGGGAGACCTGCCCTCCAAGATGCTGGTGATAGCAGCAAGG |
| AAGGAGATATTATTATGAGGACAACAGACTTCTCCTGCTGATACCACCAGACAGAGAGAGAGCACTTGATTCTTCCCGAGATAGCCACCTGGA |
| AACACTAGGTGCCTGCCCAGGGAGGAACGGAGGAGGACTCGCGCTACAAGAG |

See the amino acid sequence MARTK…. just above the cursor arrow. The top strand has the ATG which means the coding strand is on the bottom strand. The reading frame closes to the DNA sequence is the appropriate reading from to start the CDS.
a. What is this?  
**Homo sapiens glycophorin C (Gerbich blood group) (GYPC)**

b. Provide me with the protein sequence.  
MWSTRSPNSTAWPLSLEPDPGMSGWPDGRMETSTPTIMDIVVIAGVIAAVALVSSLFVMLRYMYRHKGTYHTNEAKGTEFASADAALQGDPALQDAGDSSRKEFYI

c. List the proteins functions and your source.  
From Entrez Gene  
**GYPC glycophorin C (Gerbich blood group)**  
*Homo sapiens*  
GeneID: 2995  
Locus tag: HGNC:4704; MIM: 110750  
Glycophorin C (GYPC) is an integral membrane glycoprotein. It is a minor species carried by human erythrocytes, but plays an important role in regulating the mechanical stability of red cells.

GeneOntology  
**Process**  
organ morphogenesis  
protein amino acid N-linked glycosylation  
protein amino acid O-linked glycosylation  

**Component**  
integral to plasma membrane  
plasma membrane  

Evidence  
TAS PubMed  
TAS PubMed  
TAS PubMed  
TAS PubMed  
NAS PubMed

From OMIM:  
It is a putative receptor for the merozoites of *Plasmodium falciparum* (Pasvol et al., 1984).

d. Are there any STS markers for this gene?  
Support your answer with data.  
Yes, UniSTS found 7 for humans.

e. Are there any splice variants? Support your answer with data.  
Yes, there are two.
f. Find the percent identity in the protein sequence of this gene with its chimp ortholog. Show data to support your findings.

Human NP_058131
MWSTRSPNSTAWPLSLEPDGMASASTTMTHTTTIAEPDPMGSPGWPDGRMETSTPTIMDIVVIAGVIAAVAILSVLLFVMLRYMYRKGTKYHTNEAKGTEFESAADAALQGDPALQDAGDSSRKEYFI

Pan ENSF00000008433
GLYCOPHORIN C PAS 2' GLYCOPROTEIN BETA GLPC GLYCOCONNECTIN SIALOGLYCOPROTEIN D GLYCOPHORIN D GDP
http://www.ensembl.org/Pan_troglodytes/geneview?gene=ENSPTRG00000012422
XPDPGMASASTTMTHTTTIAEPDPMGAGGVDWMEASTPTIIDIVIIIAGVIAAVAILISLLFVMLRYMYRKGTKYHTNEAKGTEFESAADAALQGDPALQDAGDSSRKEYFI

BLAST2

74% identical

---

<table>
<thead>
<tr>
<th>Sequence 1</th>
<th>lcl!Chimp</th>
<th>Length</th>
<th>107 (1..107)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence 2</td>
<td>lcl!Human</td>
<td>Length</td>
<td>107 (1..107)</td>
</tr>
</tbody>
</table>

Score: 140 hits (352), Expect = 2e-32
Identities = 71/95 (74%), Positives = 72/95 (75%)

Query: 13 IAEPDPGMAGGVDWMEASTPTIIDIVIIIAGVIAAVAILISLLFVMLRYMYRKGTKYHT 72
IAEPDPGMAGGVDWMEASTPTIIDIVIIIAGVIAAVAILISLLFVMLRYMYRKGTKYHT 72

Subject: 13 IAEPDPGMAGGVDWMEASTPTIIDIVIIIAGVIAAVAILISLLFVMLRYMYRKGTKYHT 72
IAEPDPGMAGGVDWMEASTPTIIDIVIIIAGVIAAVAILISLLFVMLRYMYRKGTKYHT 72

Query: 73 NEAKGTEFESAADAALQGDPALQDAGDSSRKEYFI 107
NEAKGTEFESAADAALQGDPALQDAGDSSRKEYFI 107

Subject: 73 NEAKGTEFESAADAALQGDPALQDAGDSSRKEYFI 107
NEAKGTEFESAADAALQGDPALQDAGDSSRKEYFI 107

---

g. Tell me the percent identity for human and chimp proteins granulysin, protamine, and semenogelin. Use a table to show your results.

Using ENSEMBL databases

<table>
<thead>
<tr>
<th>Protein</th>
<th>Identities</th>
<th>Percentage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Granulysin</td>
<td>117/126</td>
<td>92%</td>
<td>(immunity)</td>
</tr>
<tr>
<td>Protamine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semenogelin</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
protamine Identities = 27/47 (57%) (sperm protein)

semenogelin Identities = 54/97 (55%) (sperm protein)

BLAST2 alignment of human →
and chimp semenogelin

h. Human and chimp have 29% of their orthologous proteins 100% conserved and the average number of amino acid changes is 2 per protein. Propose an hypothesis to explain these genome-wide numbers with the numbers you found in f. and g. above. These 4 proteins appear to have evolved faster than the average protein. Two are related to sexual reproduction, one to immunity and the first one to malaria infection. There appears to be a strong selection pressure for these to evolve faster than most proteins.

20 pts.
3) Open the attached file called worksheet.doc.
a. Find the largest open reading frame. In which frame is the largest ORF? How many amino acids? What is the predicted molecular weight? What conserved domains are there, if any? Show data for each of these answers.

First reading frame.
1480 amino acids
The protein weighs **168.14** kilodaltons
(The Sequence Manipulation Suite)

A list of conserved domains is shown below, in graphic and text formats.
cd00267, ABC_ATPase, ABC (ATP-binding cassette) transporter
pfam00664, ABC_membrane, ABC transporter transmembrane region...
pfam00005, ABC_tran, ABC transporter. ABC transporters for a ...
COG1132, MdlB, ABC-type multidrug transport system, ATPase and...
COG2274, SunT, ABC-type bacteriocin/antibiotic exporters, con...
COG4988, CydD, ABC-type transport system involved in cytochrom...
COG4987, CydC, ABC-type transport system involved in ...
COG4618, ArpD, ABC-type protease/lipase transport system, ATPa...
COG1131, CcmA, ABC-type multidrug transport system, ATPase com...
COG5265, ATM1, ABC-type transport system involved in Fe-S clus...
COG1120, FepC, ABC-type cobalamin/Fe3+-siderophores transport ...
COG1122, CbiO, ABC-type cobalt transport system, ATPase compon...

b. Is this the wt reference sequence or a mutant DNA sequence? Show data to support your answer.

When I performed a BLAST, I got this:

\textbf{Identities = 6125/6129 (99\%)}

Therefore, the query appears to be mutated.

c. Given this allele of the gene/cDNA, explain how it could lead to a disease in humans.

Use a Genome Browser and consider the setting “Stanf Meth” to support your hypothesis with data.

Perhaps the 4 base changes alter the methylation status and thus its expression. Methylation tends to silence genes.
4) Find out some information about genome for the bacterium *Helicobacter pylori* J99 to answer these questions:


a. How big is the genome?
1,643,831 or 1,667,866 base pairs accepted

b. What is the coding density?
91% (also found in paper freely available)

c. What is the overall %GC?
39% (not the same thing as frequency of G followed by C)

d. Find a gene with a significantly different %GC. Tell me the gene, its %GC, and provide me with the DNA sequence.
CGACCCTTTGAAAAAGATTGTACCTCCGGTCCACCGTGAAAGGGTGGATCCTTGGACCTAGATGAAA
GGGTC    tRNA-Glu from *Helicobacter pylori* 26695
%GC = 49%

e. Identify the origin of replication and support your answer with two independent types of supporting data.
First, I searched for *DnaA* and found this image

chromosomal replication
initiator protein (*dnaA*)
f. What percentage of \textit{H. pylori}'s genes are considered to be essential? Provide me with the numbers you used to calculate your answer and how you obtained these numbers.

DEG = 326

Total genes = 1587 (this number varied by source)

20% are essential.

20 pts.

5) I have just read an interesting paper about RNAi in mouse. See if you can make sense of these clues and data.

a. Find the DNA sequences for \textit{NM\_0021476 Mus musculus cysteinyl leukotriene receptor 1 (Cysltr1), mRNA}.

\texttt{>gi|31542448:394-1452 Mus musculus cysteinyl leukotriene receptor 1 (Cysltr1), mRNA}

\begin{verbatim}
ATGTACCTCCAAGGCACCAAGCAGACATTCCTGGAGAACATGAATGGAACTGAAAATCTGACGACATCTC
TCATTAATAACACGTGTCATGACACAATTGATGAATTCCGAAATCAAGTATACTCCACTATGTATTCTGT
GATCTCTGTTGTGGGTTTCTTTGGCAATAGCTTTGTGCTCTATGTCCTCATAAAAACATACCATGAGAAA
TCAGCCTTCCAAGTATACATGATTAATCTAGCCATAGCAGATCTACTCTGTGTATGTACATTGCCTCTCC
GTGTGGTCTATTATGTTCACAAAGGCAAGTGGCTCTTTGGTGACTTTTTGTGCCGCCTCACCACCTATGC
CTTGTACGTTAATCCCTATATGAGCATATTCTTCTTATGAGCAGCCTAGCTTTTTCCGGTGTTGCAATT
GTCTTTCCAGTCCAGTGGTTAATTTTGTTACACAGAAAAAAGCCAGTTCTTGGCAATTTTCAGAATTTGG
TTTTTGGTATTTTGGCAAGATTCTTCCCTTTTTTAAATATTTTGAGCATTTGTGCAATTGATTTTGC
TTTTTGGTATTTTGGCAAGATTCTTCCCTTTTTTAAATATTTTGAGCATTTGTGCAATTGATTTTGC
TTTTTGGTATTTTGGCAAGATTCTTCCCTTTTTTAAATATTTTGAGCATTTGTGCAATTGATTTTGC
TTTTTGGTATTTTGGCAAGATTCTTCCCTTTTTTAAATATTTTGAGCATTTGTGCAATTGATTTTGC
TTTTTGGTATTTTGGCAAGATTCTTCCCTTTTTTAAATATTTTGAGCATTTGTGCAATTGATTTTGC
TTTTTGGTATTTTGGCAAGATTCTTCCCTTTTTTAAATATTTTGAGCATTTGTGCAATTGATTTTGC
TTTTTGGTATTTTGGCAAGATTCTTCCCTTTTTTAAATATTTTGAGCATTTGTGCAATTGATTTTGC
\end{verbatim}

It is not universal, but many genes point their transcription away from DnaA.
D530007L20
Mus musculus 13 days embryo stomach cDNA, RIKEN full-length enriched library, clone:D530007L20 product:hypothetical Rhodopsin-like GPCR superfamily/G-protein coupled receptors family 1 profile containing protein, full insert sequence.

b. Find their chromosomal positions.

Used sequence accession number to find map location
>gi|74210910|

What we see from this view is that NM_021476 and D530007L20 overlap on opposite strands of the same DNA. You can see that D530007L20 is cDNA due to the broken segments that match the BLAST results.

c. Look at Figure1.
d. Use NCBI tools to determine if the sequences you retrieved in “a” are in fact sense antisense sequences. Show data to support your findings.

The figure above demonstrates they are antisense sequences in that they come from opposite strands, but the distributions do not overlap with the exons. So the mRNAs would not be sense and antisense mRNA. Pre-mRNA transcripts would be, however.

e. Do these two coding segments have the same codon bias? Support your answer with data and interpret the implication of your findings.

<table>
<thead>
<tr>
<th>The Sequence Manipulation Suite: Codon Usage</th>
<th>The Sequence Manipulation Suite: Codon Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Results for 1059 residue sequence &quot;gi</td>
<td>31542468:394-1451 receptor 1 (Cysitril), mRNA&quot; starting &quot;atgatctccc&quot;.</td>
</tr>
<tr>
<td>AmAcid</td>
<td>Codon</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Gly</td>
<td>GGC</td>
</tr>
<tr>
<td>Gly</td>
<td>GGA</td>
</tr>
<tr>
<td>Gly</td>
<td>GGT</td>
</tr>
<tr>
<td>Gly</td>
<td>GCC</td>
</tr>
<tr>
<td>Glu</td>
<td>GAG</td>
</tr>
<tr>
<td>Glu</td>
<td>GAA</td>
</tr>
<tr>
<td>Asp</td>
<td>GAT</td>
</tr>
<tr>
<td>Asp</td>
<td>GAC</td>
</tr>
<tr>
<td>Val</td>
<td>GTG</td>
</tr>
<tr>
<td>Val</td>
<td>GCA</td>
</tr>
<tr>
<td>Val</td>
<td>GCT</td>
</tr>
<tr>
<td>Val</td>
<td>GTC</td>
</tr>
<tr>
<td>Ala</td>
<td>GCG</td>
</tr>
<tr>
<td>Ala</td>
<td>GCA</td>
</tr>
<tr>
<td>Ala</td>
<td>GCT</td>
</tr>
<tr>
<td>Ala</td>
<td>GCC</td>
</tr>
<tr>
<td>Arg</td>
<td>AGG</td>
</tr>
<tr>
<td>Arg</td>
<td>AGA</td>
</tr>
<tr>
<td>Ser</td>
<td>AGT</td>
</tr>
<tr>
<td>Ser</td>
<td>AGC</td>
</tr>
<tr>
<td>Lys</td>
<td>AAG</td>
</tr>
<tr>
<td>Lys</td>
<td>AAA</td>
</tr>
<tr>
<td>Arg</td>
<td>AAT</td>
</tr>
<tr>
<td>Arg</td>
<td>AAC</td>
</tr>
<tr>
<td>Met</td>
<td>ATG</td>
</tr>
<tr>
<td>Ile</td>
<td>ATA</td>
</tr>
<tr>
<td>Ile</td>
<td>ATT</td>
</tr>
<tr>
<td>Ile</td>
<td>ATC</td>
</tr>
<tr>
<td>Thr</td>
<td>ACG</td>
</tr>
<tr>
<td>Thr</td>
<td>ACA</td>
</tr>
<tr>
<td>Thr</td>
<td>ACT</td>
</tr>
<tr>
<td>Thr</td>
<td>ACC</td>
</tr>
</tbody>
</table>
Based on the 3 boxed areas, it appears these two genes do not share identical codon bias.

f. Now look at the file called “SOM Fig5.pdf” and try to make sense of all the data you have collected so far. Summarize your conclusions based on all the data you have, both from your own research and from this publication.

The key points are these:
1) When NM_021476 is induced, D530007L20 is repressed and vice versa.
2) These two genes overlap on opposite strands.
3) Their exons do not appear to overlap at all.
4) They have different codon bias.
5) The question indicated it was related to RNAi.

The implication is that two genes may act as RNAi for each other. D530007L20 is only a hypothetical protein, but its mRNA is real. When NM_021476 is repressed, D530007L20 is induced. Perhaps D530007L20 acts as RNAi to completely silence NM_021476 residual mRNA. However, we have a problem that RNAi appears to work on mRNA but these two “genes” do not have codons in common. Therefore, it appears that RNAi may work inside the nucleus or perhaps non-spliced RNA is able to reach the cytoplasm and serve as RNAi raw material.