Fall 2004 Genomics Exam #1
Genomic Medicine and Sequencing Tools

There is no time limit on this test, though I have tried to design one that you should be able to complete within 6 hours, except for typing and web searches. There are three pages for this test, including this cover sheet. You are not allowed discuss the test with anyone until all exams are turned in at 11:30 am on Friday October 1. **EXAMS ARE DUE AT CLASS TIME ON FRIDAY OCTOBER 1**. You may use a calculator, a ruler, your notes, the book and the internet. This is a challenging test, so do NOT put it off too long. You may take it in as many blocks of time as you need to.

The **answers to the questions must be typed within this Word file.** If you do not write your answers in the appropriate location, I may not find them. 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. Paste the images within your Word file at the appropriate places. Print one hard copy (B&W or color, either is fine) to turn in no later than Friday at 11:30 am in class. In addition, please email me a copy of your Word file, also due by 11:30 am.

-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:

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

30 Points

1) Start with this partial sequence:

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKT

a) From what protein is this sequence?

α-synuclein from a mammal

b) With what disease or diseases is/are this protein associated?

Dementia, Lewy body

Parkinson disease 4, autosomal dominant Lewy body

Parkinson disease, familial

Alzheimer's

http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=127750

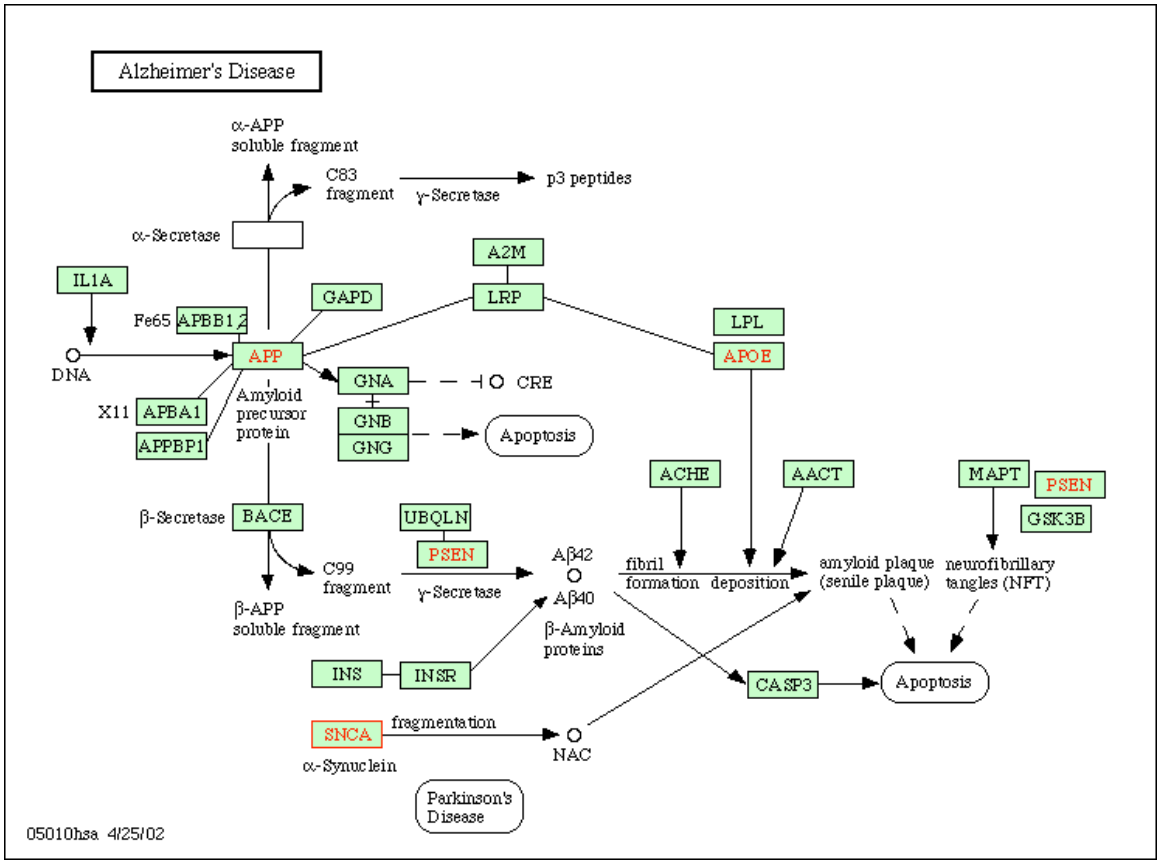
http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=605543

http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=168601

http://www.genome.jp/dbget-bin/show_pathway?hsa05010+6622

c) Show me a picture of its biochemical pathway.

http://www.genome.jp/dbget-bin/show_pathway?hsa05010+6622



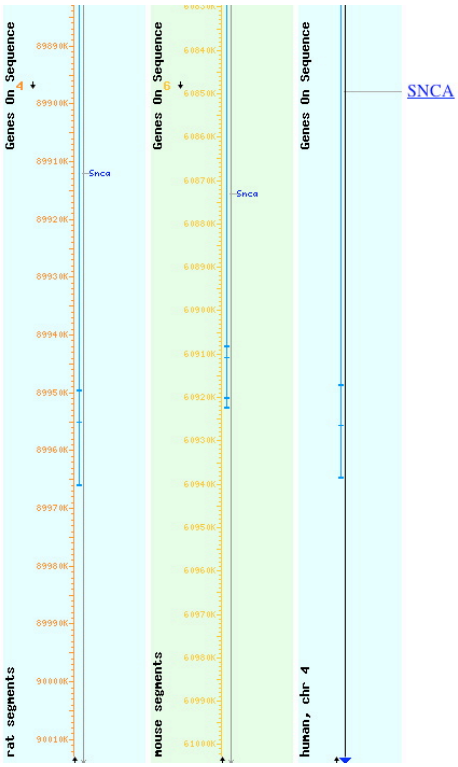
d) Describe the cellular "function" of this protein? Provide the URL(s) for your source(s).

cytoplasm cellular_component
DNA binding molecular_function
pathogenesis biological_process
protein binding molecular_function
regulation of transcription, DNA-dependent biological_process
transcription factor activity molecular_function
http://www.godatabase.org/cgi-bin/amigo/go.cgi?action=query&view=query&session_id=9146b1095793435&query=SNCA&search_constraint=gp

e) Are there any alternative spliced forms of this protein? Support your answer with data.
Yes. http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=643590
LOCUS D31839 1096 bp mRNA linear PRI 07-FEB-1999
DEFINITION Human alternatively spliced mRNA for NACP (precursor of non-A beta component of Alzheimer's disease amyloid), complete cds.
vs.
http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=L08850&dopt=DocSum&dispmax=1000
L08850
Links
Human AD amyloid mRNA, complete cds
gi|437364|gb|L08850.1|HUMAMY[437364]

f) Based on your answer to part b, what can you deduce about this protein and its cellular roles?
Probably multi-facetted. Interacts with more than one protein. Alternative forms probably alter normal function.

g) On what chromosomes are the human, mouse and rat orthologs? Support your answer with a single image.
Human and Rat Chrom. 4
Mouse Chrom. 6



h) What is the Rat Accession Number for the mRNA/cDNA? What is the human accession number?

XM_225768(rat) NM_005460.1 (human) OR

Rat: gi:9507124

Human gi:6806896 and gi:6806897

i) What differences are there between human an rat orthtologs at the amino acid level?

Depends which ones you compared. Here is one result. From cDNA conversion and cut and pasted sequences: Score = 1315 bits (3404), Expect = 0.0 Identities = 683/907 (75%), Positives = 728/907 (79%), Gaps = 8/907 (0%)(from BLAST2)

NP_062042.1 rat

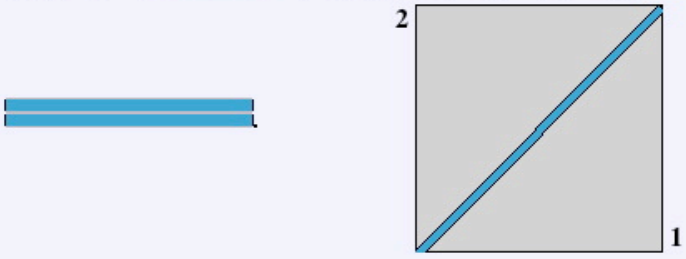
MGAEGSETGMDICAISSELCGVWTCFSIVSQSVILIASPGLRLG
QRLKGVVFPSTRICLWKRRASKQTRAPLAFYDIISYSVTSLKTIPALCRRCDSQNEDR
PVSSSSWNCGVSTLITNPQKPTGIADVYSKFRPVKRVSPLKHQPETLESNESDDQKNN
TVEYQKGGETDQGPQPEELSPEDGVGGLPGKGSEPSQALGELEHYDLDMDEILDVPYI
KSSQQLAPLTKVTSEKRILGLCTTVNGLSAKTCPIVSAETSTPNMAPLCVLSPVKSPH
LRKVPSVLRDQHKLPAEESENSPAPGKCGPAFESENHSKDFLNKVFSDPHSRKAEKSG
PDCKLRPFRLQTSAAGAKPEEQVNGVSWASAQGAEERTEYLQKVRSILNIVNEGQISL
LPHLAADNLDKIHDENGNNLLHVAASKGHAECLQHLTSLMGEDCLNERNAEQLTPAGL
AIKNGQLECVRWMVSETEAIAELSCSKDFPSLIHYAGCYGQEKILLWLLQFMQEQGIS
LDEVDQEGNSAVHVASQHGYLGCIQTLVEYGANVTMQNHAGEKPSQSAERHGHTLCSR
YLVVVETCMSLASQVVKLTKQLKEQTVERVTLQSQLQQLLEAQKSEGKSLPSSPSSPS
SPASRKSQWKILDADDESTGKSKLGTQEGIQVLGNLSSRARTKGKDEDSDKILRQLLG
KEISENVCTQEKLSLEFQDAQVSSRNSKKIPLEKRELKLARLRQLMQRSLSESDTDSN
NSEDPKNTPVKRVDRPRPQPIVESVENMDSAESLHLMIKKHSVASGRRFPFGMKASKS
LDGHSPSPTSESSEPDLDSHCPSLGMTPPTQPSTEATQCSPDSATAQKVATSPKSALK
SPSSKRRTSQNSKLRVTFEEPVVQMEQTSLELNGEKDKERGRAPQRTSESGEQMKRPF
GTFRSIMESLSGNQNNNNNYQPASQLKTCTLPLTSLGRKTADAKGNPVSPASKGKNKA
AMYSSCIHLPSNALVEEHLRDYARSDVSPWSLKTYAFVPETKEHKDLANSLEAERKNA
FQTPRATGNEIINVTADLSCQKCFTLPFYKERKKAGHFS

AAP36433.1 human

MEAPEYLDLDEIDFSDDISYSVTSLKTIPELCRRCDTQNEDRSA
SSSSWNCGISTLITNTQKPTGIADVYSKFRPVKRVSPLKHQPETLENNESDDQKNQKV
VEYQKGGESDLGPQPQELGPGDGVGGPPGKSSEPSTSLGELEHYDLDMDEILDVPYIK
SSQQLASFTKVTSEKRILGLCTTINGLSGKACSTGSSESSSSNMAPFCVLSPVKSPHL
RKASAVIHDQHKLSTEETEISPPLVKCGSAYEPENQSKDFLNKTFSDPHGRKVEKTTP
DCQLRAFHLQSSAAESKPEEQVSGLNRTSSQGPEERSEYLKKVKSILNIVKEGQISLL
PHLAADNLDKIHDENGNNLLHIAASQGHAECLQHLTSLMGEDCLNERNTEKLTPAGLA
IKNGQLECVRWMVSETEAIAELSCSKDFPSLIHYAGCYGQEKILLWLLQFMQEQGISL
DEVDQDGNSAVHVASQHGYLGCIQTLVEYGANVTMQNHAGEKPSQSAERQGHTLCSRY
LVVVETCMSLASQVVKLTKQLKEQTVERVTLQNQLQQFLEAQKSEGKSLPSSPSSPSS
PASRKSQWKSPDADDDSVAKSKPGVQEGIQVLGSLSASSRARPKAKDEDSDKILRQLL
GKEISENVCTQEKLSLEFQDAQASSRNSKKIPLEKRELKLARLRQLMQRSLSESDTDS
NNSEDPKTTPVRKADRPRPQPIVESVESMDSAESLHLMIKKHTLASGGRRFPFSIKAS
KSLDGHSPSPTSESSEPDLESQYPGSGSIPPNQPSGDPQQPSPDSTAAQKVATSPKSA
LKSPSSKRRTSQNLKLRVTFEEPVVQMEQPSLELNGEKDKDKGRTLQRTSTSNESGDQ
LKRPFGAFRSIMETLSGNQNNNNNYQAANQLKTSTLPLTSLGRKTDAKGNPASSASKG
KNKAA

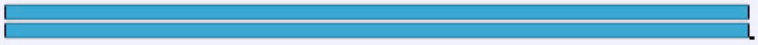
Using Accession numbers above, not cDNA accession numbers.

Sequence 1 gi 9507125 synuclein, alpha [Rattus norvegicus]
Sequence 2 gi 30584369 Homo sapiens synuclein, alpha (non A4 component of amyloid precurs



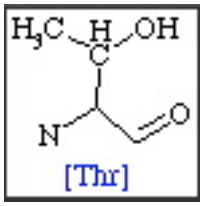
NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score =  201 bits (511), Expect = 4e-51
Identities = 103/140 (73%), Positives = 106/140 (75%)
```

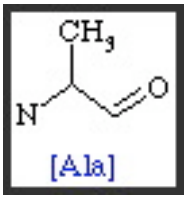


```
Query:     1   MDVFMKGLSXXXXXXXXXXXXXXQGVAEAAGKTKEGVLYVGSKTKEGVVHGVTTVAEKTK 60
               MDVFMKGLS               QGVAEAAGKTKEGVLYVGSKTKEGVVHGV TVAEKTK
Sbjct:     1   MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTK 60
Synuclein 1   ************************************************************
```

j) Look at amino acid 53 in rat v. human, and you will see they are different (from h above). The rat amino acid 53, when found in humans, is associated with one of the diseases from your answer in part b of this question. Using the physical properties of the two amino acids being compared, explain why this difference in the protein could have a functional consequence. Use screen shots to support your answer.



[Thr]



[Ala]

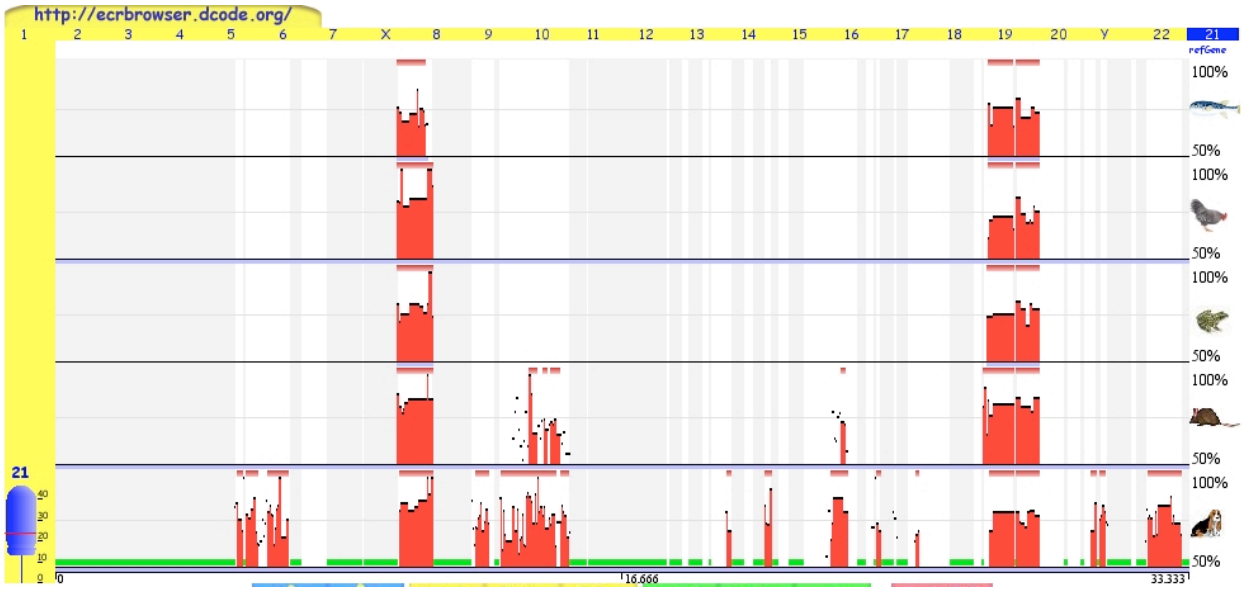
Threonine is hydrophilic and alanine is hydrophobic. These two amino acids are likely to have a significant impact on protein shape due to the amino acids' different interactions with water.

k) How can wt rats have amino acid #53 their way but if we have it we have a disease? Since each has a different wildtype amino acid in this position and the rat one leads to human diseases, there must be compensatory mutations in rats that prevent the one residue from altering the rat's overall physiology. Because the cell web is so complex, other proteins must be involved and thus probably interact fine with this shape.

20 Points

2) Use the ECR Browser (see accompanying paper) to answer the following questions. You will need access to the paper to help you navigate.

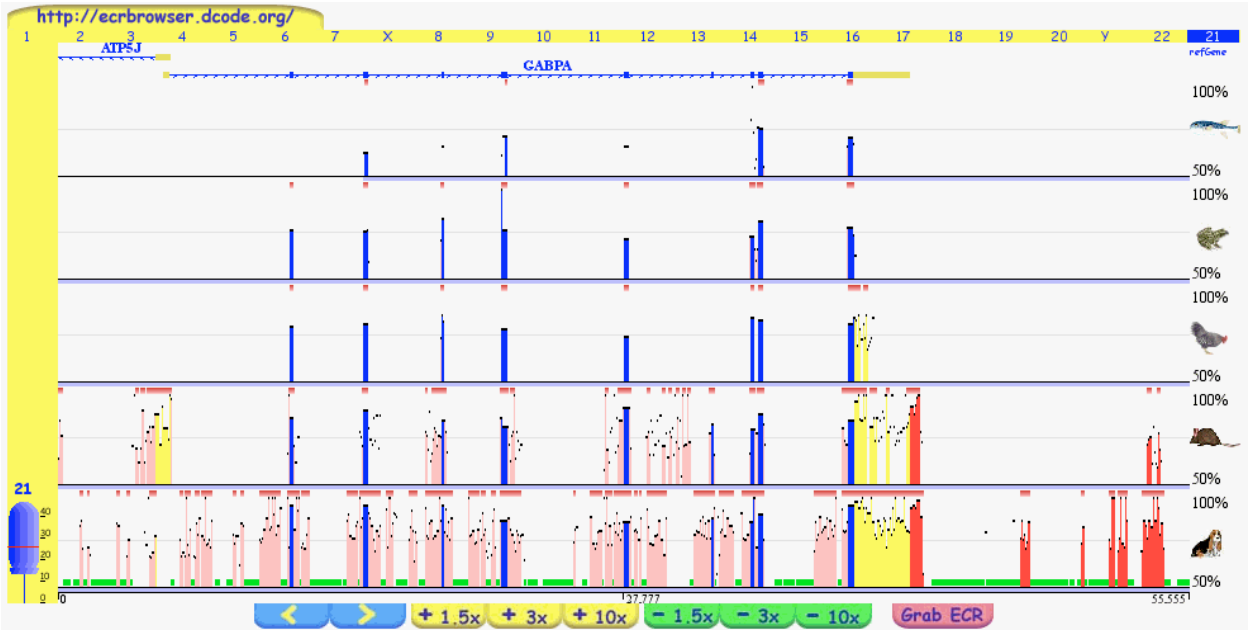
a) Look at chromosome 21 in this region: 23674384-23707716. Take a screen shot and then interpret what you see.



Answer needed to include: This is non-coding DNA and yet it is highly conserved across many species. The two main blocks are large sections of DNA and their conservation is striking.

b) Go to 26024345-26079899 on chromosome 21. Interpret what you see.

GA-BINDING PROTEIN TRANSCRIPTION FACTOR, ALPHA SUBUNIT; GABPA



Answer needed to address that mammalian conservation of exons and non-exons. Non-mammals conserved exons only.

c) Comment on the degree of conservation you found in these two regions and hypothesize on the significance of this conservation.

Key point was that non-coding DNA was at least as well conserved as exons and more conserved than introns and UTR's. This indicates a strong selection pressure on the first section of DNA even though we do not know what functions are in these regions.

20 Points

3) There is another mystery yet to be solved: human protein called TAF1L.

a) What is its function?

http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&CID=522061
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=138474
http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=607798

TAF1-like RNA polymerase II, TATA box binding protein (TBP)-associated factor, 210kDa

Function

DNA binding

Process

regulation of transcription, DNA-dependent transcription initiation

Component

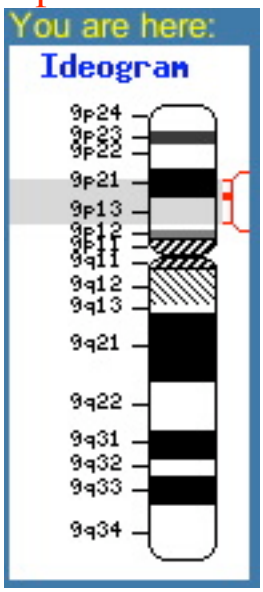
transcription factor TFIID complex

b) Where is it expressed?

Testis

c) Where is this gene located in the genome?

9p21.1 – near the centromere.



d) What have we learned in class that connects parts b and c from above?

We discussed a paper in class that showed the centromere is a region that is accumulating duplicated genes and these centromeric versions are often expressed in the testis. This gene fits the pattern. We looked at two figures in class and analyzed these data.

e) Is there a functional mouse ortholog? Support your answer with data.

No. Only the paralog of 250kDa. The table below shows a mouse protein of about 10% the TAF1L protein size.

UniGene Cluster Hs .522061 *Homo sapiens*

TAF1-like RNA polymerase II, TATA box binding protein (TBP)-associated factor, 210kDa (TAF1L) Links

SELECTED PROTEIN SIMILARITIES

organism, protein and percent identity and length of aligned region		
H. sapiens:	pir:A40262 - A40262 transcription initiation factor IID 250K chain splice form 1 - human	93.39 % / 1814 aa (see ProtEST)
D. melanogaster:	pir:A47371 - A47371 transcription initiation factor IID 230K chain - fruit fly	47.16 % / 1730 aa (see ProtEST)
M. musculus:	sp:Q9JHD2 - GCL2_MOUSE General control of amino acid synthesis protein 5-like 2	34.26 % / 108 aa (see ProtEST)
A. thaliana:	ref:NP_174552.1 - hypothetical protein [Arabidopsis thaliana]	29.29 % / 613 aa (see ProtEST)
S. cerevisiae:	pir:S50237 - S50237 TATA box-binding protein-associated factor chain TAFII145 - yeast	30.59 % / 493 aa (see ProtEST)
C. elegans:	ref:NP_493426.1 - transcription initiation factor TFIID [Caenorhabditis elegans]	38.39 % / 1129 aa (see ProtEST)

MAPPING INFORMATION

f) Any known diseases associated with this locus?

None found in databases. Presumably, this might lead to male sterility.

15 Points

4) a) Interpret figure 1A and B as fully as you can. Do not use information from part b to augment your interpretation for part a.

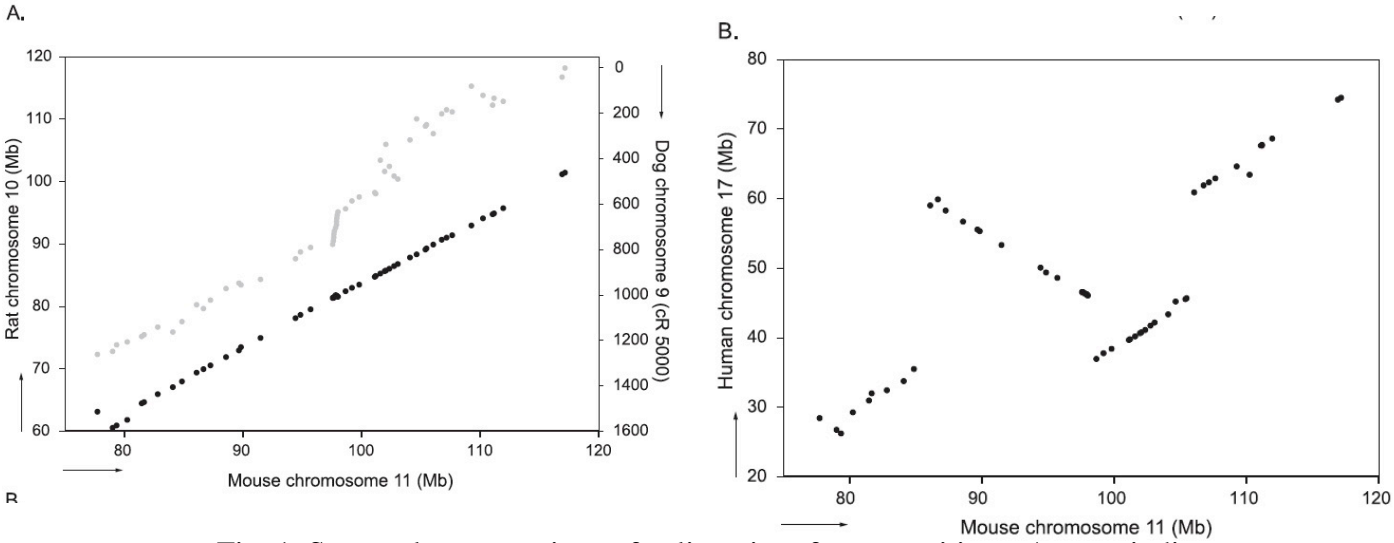
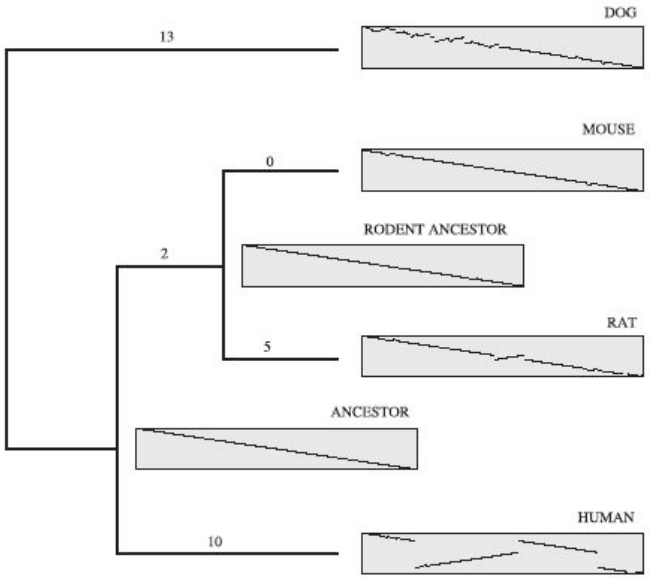


Fig. 1. Scatterplot comparison of colinearity of gene positions. Arrows indicate chromosomal orientation from centromere to telomere. (A) Black-filled circles: murine gene positions in Mb, on the x axis, versus rat gene positions in Mb, left y axis. Gray-filled circles: murine gene positions in Mb, on the x axis, versus dog gene positions in cR5000, right y axis. (B) Murine gene positions in Mb, on the x axis, versus human gene positions in Mb, y axis.

Key points; mouse and rat are highly collinear. Dog less so with an odd spread of the DNA in the top right corner of the graph. Human DNA has undergone at least 4

inversions, with one large one in the middle, a part of which was later re-inverted (the downward shifted piece).

b) Open the Exam_PDF_2.pdf file. Explain figure 3. Figure 2 is intended to help you if you want to see details. You will notice that some of the text has "accidentally" been lost. Do not try to track down this paper. You have all the information you need in the figures and the figure legends.



The ancestor sequence of genes (not DNA) is shown. Rodents have 2 inversions, mouse has no more. Rats have 5 more inversions as shown in the graphic and number values. Because the mouse line is zero inversions, the lines cannot be considered to scale (time or number of changes). Humans have 10 inversions that are easier to see now. Dogs have 13 inversions from the ancester and these are unrelated to either the rodent or human inversions. The dog has many small inversions while the human has a few large ones.

c) Having seen part b, What can you add to your answer for part a?

Two striking features: dog has many small inversions that explain the odd spread of dots in part a; humans have more than just 4 inversions.

15 Points

5)

a) Describe as fully as you can this protein:

```
MTLTTKLSALAIAGIMAVIGAPMVTQSAMASGRAPAPDAATTQPKLVTGDITSTDQSGTHLFFGKNI
VRNAKGAIMKVDRTWPAAVPAPLPDVRADSSTRMLLGPVVDLAVNEHPAGVFYRIPALATASNGDLL
ASYDLRPGSAADAPNPNSIVQRRSRDNGRTRGPQTVIHAGTLGRRKVGYSDPSYLVDPATGHILNFH
VKSYDRGFATSEVGTDPDDRHVLHAEVSTSTDNGHTWTYRDITREITPDPTTRTRFVASGQGIALLH
GPHAGGLIAQMTVRNSVGQQAQSIYSDDHGITWHAGNPVGRMMDENKVVELSDGTLMLNSRDAARSG
RRKVAYSHDGGLTWGPVKLVDDLIDPTNNAQIIRAYPNARAGSAKARILLFTNARNATERVNGTLSV
SCDDGRTWVSHQTYMPGEVGYTTAAVQSDGALGVLWERDGIRYSTIPMGWLNSVCPVAPSGRPTSGE
PTSGTSLPLTATPSGSLHGGASSRPTSLPHTGD
```

Be sure to include is function(s), species of origin, and any other aspects you can discover.

Functions: sialidase precursor that metabolized terminal terminal sialic acid residues from various glycoproteins and extracellular matrix molecules.

Species *Propionibacterium acnes* (causative agent for some acne) See this PubMed entry http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&list_uids=15286373&dopt=Citation

b) There is a problem with this sequence

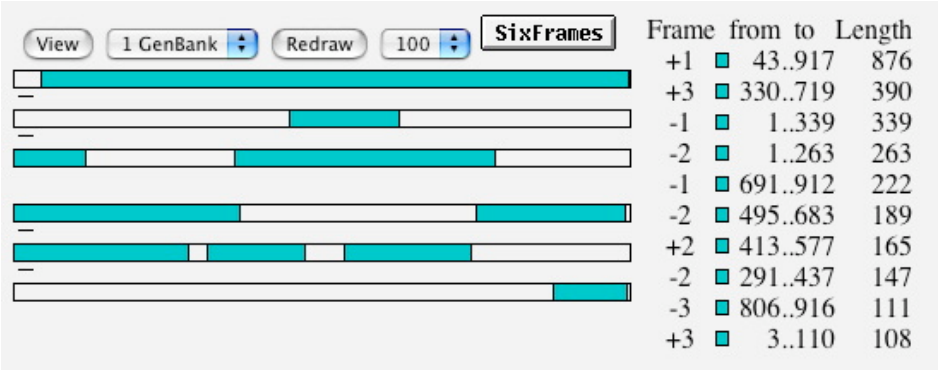
```
GAGTTCTGGTTCGCTGTCATCAAAGTCGTCGCGATTCTTGCGATGATCGTGCTGGGTGTCCTTATCA
TTGCAACTGGCCTGGGTGGTGGCCCTCCGACCGGGATAGGTAACCTGTGGCGACACGGAGGATTCTT
TCCAACCGGCATCAGCGGGATGCTGTGCGGTTTTGTCGTGGTGATGTTCAGCTTTGGAGGGGTCGAG
CTCATCGGGATTACGGCAGGGGAGGCTGACGATCCGCGTCGGTCTATTCCGCGAGCGATCAATCAAG
TCGTGTATCGGATCCTCATTTTCTACATCGGTGCAATTTCGGTCATGTTGTGTCTTTTTCCATGGAA
CCAGATCGGCAAGGCAGGCAGCCCCTTCGTGACGATCTTCGACAAAATCGGAGTCGCAGGTGCGGCG
AATATCCTCAATGTTGTGGTGCTTACCGCTTCCATGTCGGCCTACAACTCGGGCCTATACTCCAACG
GGCGGATGCTTTACAGCTTGGCCGCTCAGCACAACGCTCCCGGGATCTTCTGGAAGACGAATCGGCT
GGGGGCGCCGTGGGTGGGAGTGCTCGCCTCCTCGGTGGTGACGGCAACGGCGGTGCTGCTGACGTAC
TTGATTCCTGGAAAGGTGTTTTTGTACATCATCTCGATCGCCTTGATCTCTGGGGTCATCAATTGGA
CGATGATCATCATCACCAACCTAAAGTTTCGGCGAAGGATCGGTCCTGAAGGTGTCGCAGCGTTGGA
ATTTCGGATGCCGGGTAATCCCGTCACCAGTTACGTGGTGTTGGTTTTTCTGGCGCTCGTGGTGGTC
ATCATGGCGATGATGCCGAGCTACCGAGTGGCACTCGTTGTTGGTCCCGTCTGGTTGGCGTTGCTGT
GGGTGGGTTATGACGTGTCCTGCCTGGTGCGACGCCGTCATGCCTGA
```

a) Is this coding sequence? Support your answer with data. Use screen shots if they help you document your case.

Yes, see below.

```
Sequences producing significant alignments:                    (bits) Value

gi|15559766|gb|BC014236.1|  Homo sapiens cDNA clone IMAGE:45...  1820    0.0
gi|50839098|gb|AE017283.1|  Propionibacterium acnes KPA17120...  1812    0.0
```



b) What is the problem with this sequence and hypothesize how the problem happened?

This probably was a bacterial contaminant of human and mouse cDNA libraries that came from a technician rather than a prokaryote gene. It accidentally got annotated in the human genome but is supposed to be in the prokaryotes only. Because there is only one base difference, this is unlikely to be horizontal transfer.

Oops, accidents happen.