Some Protein Sequences

I looked up the *Insulin-like growth factor binding protein 3* sequence for several species:

```
>IBP3_BOVIN
--MLRAPPRLWAAALTALTLLRGPPAARAGAGTMGAGPVVRCEPCDARAVAQCAPPPPSPPCAELVRDAG
CGCCLTCALREGQPCGVYTERCGSGLRCQPPPGDPRPLQALLDGRGLCANASAVGRLRPYLLPS--ASGN
GSES-----EEDHSMGSTENQAGPSTHRVPVSKFHPIHTKMDVIKKGHAKDSQRYKVDYESQSTDTQNFS
SESKRETEYGPCRREMEDTLNHLKFLNMLSPRGIHIPNCDKKGFYKKKQCRPSKGRKRGFCWCVDKYGQP
LPGFDVKGKGDVHCYSMESK--------------

>IBP3_PIG
-----------------------------GSGAVGTGPVVRCEPCDARALAQCAPPPAAPPCAELVREPG
CGCCLTCALREGQACGVYTERCGAGLRCQPPPGEPRPLQALLDGRGICANASAAGRLRAYLLPAPPAPGN
GSES-----EEDRSVDSMENQALPSTHRVPDSKLHSVHTKMDVIKKGHAKDSQRYKVDYESQSTDTQNFS
SESKRETEYGPCRREMEDTLNHLKFLNMLSPRGIHIPNCDKKGFYKKKQCRPSKGRKRGFCWCVDKYGQP
LPGFDVKGKGDVHCYSMESK--------------

>IBP3_RAT
--MHPARPALWAAALTALTLLRGPPVARAGAGAVGAGPVVRCEPCDARALAQCAPPPTAPACTELVREPG
CGCCLTCALREGDACGVYTERCGTGLRCQPRPAEQYPLKALLNGRGFCANASAASNLSAY-LPSQPSPGN
TTES-----EEDHNAGSVESQVVPSTHRVTDSKFHPLHSKMEVIIKGQARDSQRYKVDYESQSTDTQNFS
SESKRETEYGPCRREMEDTLNHLKFLNVLSPRGVHIPNCDKKGFYKKKQCRPSKGRKRGFCWCVDKYGQP
LPGYDTKGKDDVHCLSVQSQ--------------

>IBP3_MOUSE
--MHPARPALWAAALTALTLLRGPPVAELAAGAVG-GPVVRCEPCDARAVSQCAPPPTAPACTELVREPG
CGCCLTCALREGDACGVYTERCGTGLRCQPRPAEQYPLRALLNGRGFCANASAAGSLSTY-LPSQPAPGN
ISES-----EEEHNAGSVESQVVPSTHRVTDSKFHPLHAKMDVIKKGHARDSQRYKVDYESQSTDTQNFS
SESKRETEYGPCRREMEDTLNHLKFLNVLSPRGVHIPNCDKKGFYKKKRCRPSKGRKQSFCWCVDKYGQR
LPGYDTKGKDDVHCLSVQSQ--------------

>IBP3_HUMAN
--MQRARPTLWAAALTLLVLLRGPPVARAGASSGGLGPVVRCEPCDARALAQCAPPP--AVCAELVREPG
CGCCLTCALSEGQPCGIYTERCGSGLRCQPSPDEARPLQALLDGRGLCVNASAVSRLRAYLLPAPPAPGN
ASES-----EEDRSAGSVESPSVSSTHRVSDPKFHPLHSKIIIIKKGHAKDSQRYKVDYESQSTDTQNFS
SESKRETEYGPCRREMEDTLNHLKFLNVLSPRGVHIPNCDKKGFYKKKQCRPSKGRKRGFCWCVDKYGQP
LPGYTTKGKEDVHCYSMQSK--------------
```

Some Protein Sequences

```
>IBP2_CHICK
MALGGVGRGGAARAAWPRLLLAALAPALALAGPALPEVLFRCPPCTAERLAACSP-AARPPCPELVREPG
CGCCPVCARLEDEACGVYTPRCAAGLRCYPDPGAELPPQALVQGQGTCARPPDTDEYGASTEPPADNGDD
RSESILAENHVDSTGGMMSGASSRKPLKTGMKEMPVMREKVNEQQRQMGKVGKAHHNHEDSKKSRMPTGR
TPCQQELDQVLERISTMRLPDERGPLEHLYS--LHIPNCDKHGLYNLKQCKMSVNGQRGECWCVDPIHGK
VIQGAPTIRGDPECHLFYTAHEQEDRGAHALRSQ

>IBP2_BRARE
-MLSYVSCG---------LLLALVT----FHGTARSEMVFRCPSCTAERQAAC-P-MLTETCGEIVREPG
CGCCPVCARQEGEQCGVYTPRCSSGLRCYPKPDSELPLELLVQGLGRCGRKVDTEPTG-SAEPREVSG--
-----------EVQDPLDIGLTEVPPIRKPTKDSP-WKESAVLQHRQQLKSKMKYHKVEDPKAPHAKQ--
SQCQQELDQVLERISKITFKDNRTPLEDLYS--LHIPNCDKRGQYNLKQCKMSVNGYRGECWCVNPHTGR
PMPTSPLIRGDPNCNQYLDGQE-MDPSVDPPN--

>IBP4_SHEEP
-----------------------------------DEAIHCPPCSEEKLARCRP-PVG--CEELVREPG
CGCCATCALGKGMPCGVYTPDCGSGLRCHPPRGVEKPLHTLVHGQGVCMELAEIEAIQESLQPSD-----
-------------KDEGDHPNNSFSPCSAHDRK---CLQKHLAKIRDRSTSGGKMKVIGAPREEVRPVPQ
GSCQSELHRALERLAAS----QSRTHEDLYI--IPIPNCDRNGNFHPKQCHPALDGQRGKCWCVDRKTGV
KLPGGLEPKGELDCHQLADSFRE-----------
```

These proteins are orthologous, meaning that they are presumed to derive from a common ancestor.

But how???

A Protein and its Sequence

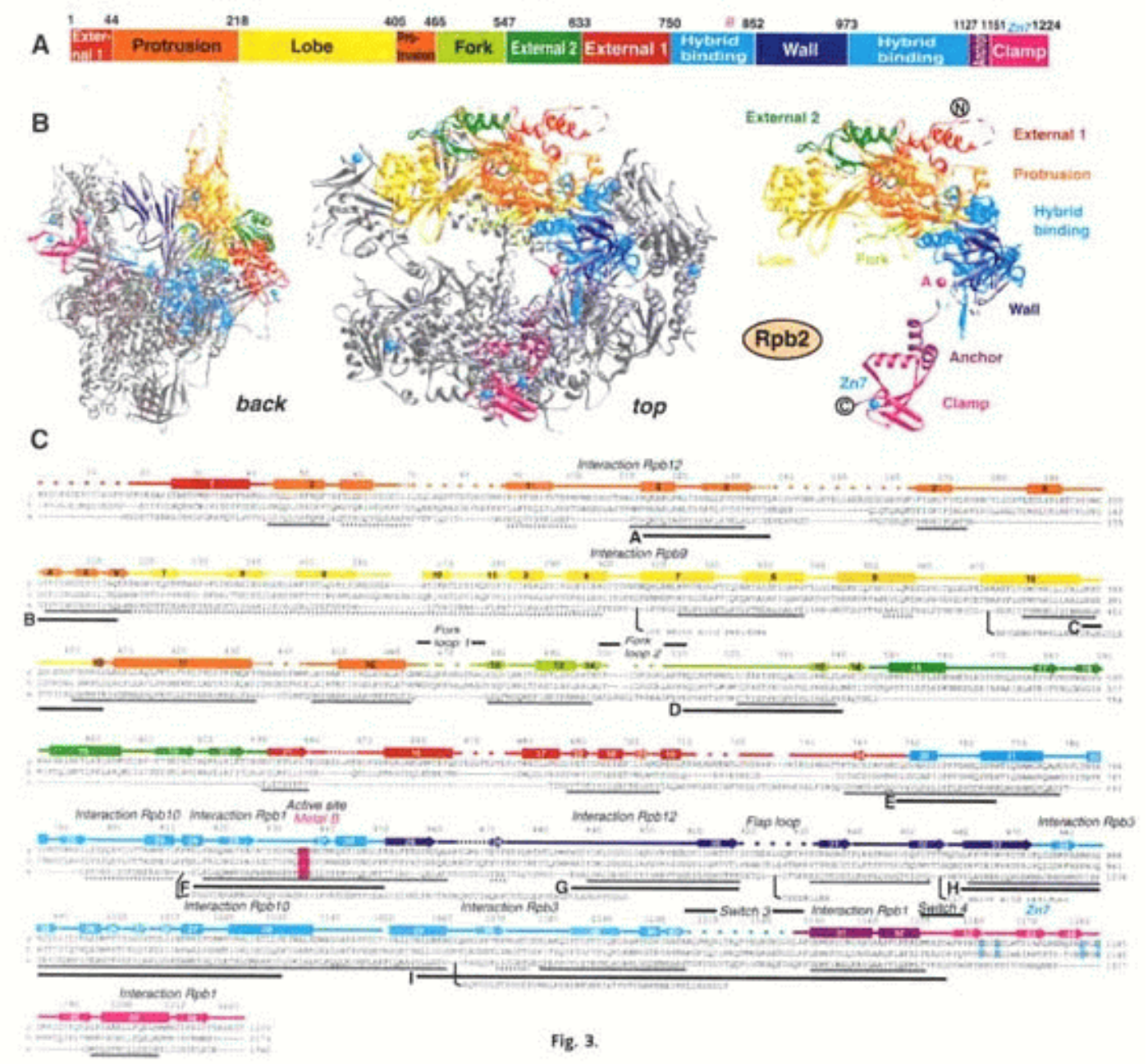


Fig. 3.

Source: Science, June 8, 2001 v292 i5523 p1863 Structural Basis of Transcription: RNA Polymerase II at 2.8 Angstrom Resolution. (Statistical Data Included) Patrick Cramer; David A. Bushnell; Roger D. Kornberg.

Reconstructing Phylogenies

Given orthologous DNA or protein sequences, how can we infer the evolutionary history of the species involved?

Technically, we can't. All we can infer is the history of that particular gene or protein.

Consider how speciation occurs: First enough genetic variation occurs within the population to force speciation, but then at least one population still has the variations, implying that some members of the population will be more similar to members of the new species than to members of its own species, according to some genes.

(Little help from the biologists, please...)

But if we do this for enough genes or proteins we may be able to arrive at some consensus tree.

Reconstructing Phylogenies

So let's build a tree anyway. What is the first thing you would suggest we do with those eight sequences?

Align!

```
IBP3_BOVIN  --MLRAPPRL WAAALTALTL LRGPPAARAG AGTMGAGPVV RCEPCDARAV
  IBP3_PIG  ---------- ---------- ---------G SGAVGTGPVV RCEPCDARAL
  IBP3_RAT  --MHPARPAL WAAALTALTL LRGPPVARAG AGAVGAGPVV RCEPCDARAL
IBP3_MOUSE  --MHPARPAL WAAALTALTL LRGPPVAELA AGAVG-GPVV RCEPCDARAV
IBP3_HUMAN  --MQRARPTL WAAALTLLVL LRGPPVARAG ASSGGLGPVV RCEPCDARAL
IBP2_CHICK  MALGGVGRGG AARAAWPRLL LAALAPALAL AGPALPEVLF RCPPCTAERL
IBP2_BRARE  -MLSYVSCG- --------LL LALVT----F HGTARSEMVF RCPSCTAERQ
IBP4_SHEEP  ---------- ---------- ---------- ------DEAI HCPPCSEEKL

IBP3_BOVIN  AQCAPPPPSP PCAELVRDAG CGCCLTCALR EGQPCGVYTE RCGSGLRCQP
  IBP3_PIG  AQCAPPPAAP PCAELVREPG CGCCLTCALR EGQACGVYTE RCGAGLRCQP
  IBP3_RAT  AQCAPPPTAP ACTELVREPG CGCCLTCALR EGDACGVYTE RCGTGLRCQP
IBP3_MOUSE  SQCAPPPTAP ACTELVREPG CGCCLTCALR EGDACGVYTE RCGTGLRCQP
IBP3_HUMAN  AQCAPPP--A VCAELVREPG CGCCLTCALS EGQPCGIYTE RCGSGLRCQP
IBP2_CHICK  AACSP-AARP PCPELVREPG CGCCPVCARL EDEACGVYTP RCAAGLRCYP
IBP2_BRARE  AAC-P-MLTE TCGEIVREPG CGCCPVCARQ EGEQCGVYTP RCSSGLRCYP
IBP4_SHEEP  ARCRP-PVG- -CEELVREPG CGCCATCALG KGMPCGVYTP DCGSGLRCHP

IBP3_BOVIN  PPGDPRPLQA LLDGRGLCAN ASAVGRLRPY LLPS--ASGN GSES-----E
  IBP3_PIG  PPGEPRPLQA LLDGRGICAN ASAAGRLRAY LLPAPPAPGN GSES-----E
  IBP3_RAT  RPAEQYPLKA LLNGRGFCAN ASAASNLSAY -LPSQPSPGN TTES-----E
IBP3_MOUSE  RPAEQYPLRA LLNGRGFCAN ASAAGSLSTY -LPSQPAPGN ISES-----E
IBP3_HUMAN  SPDEARPLQA LLDGRGLCVN ASAVSRLRAY LLPAPPAPGN ASES-----E
IBP2_CHICK  DPGAELPPQA LVQGQGTCAR PPDTDEYGAS TEPPADNGDD RSESILAENH
IBP2_BRARE  KPDSELPLEL LVQGLGRCGR KVDTEPTG-S AEPREVSG-- ----------
IBP4_SHEEP  PRGVEKPLHT LVHGQGVCME LAEIEAIQES LQPSD----- ----------

IBP3_BOVIN  EDHSMGSTEN QAGPSTHRVP VSKFHPIHTK MDVIKKGHAK DSQRYKVDYE
  IBP3_PIG  EDRSVDSMEN QALPSTHRVP DSKLHSVHTK MDVIKKGHAK DSQRYKVDYE
  IBP3_RAT  EDHNAGSVES QVVPSTHRVT DSKFHPLHSK MEVIIKGQAR DSQRYKVDYE
IBP3_MOUSE  EEHNAGSVES QVVPSTHRVT DSKFHPLHAK MDVIKKGHAR DSQRYKVDYE
IBP3_HUMAN  EDRSAGSVES PSVSSTHRVS DPKFHPLHSK IIIIKKGHAK DSQRYKVDYE
IBP2_CHICK  VDSTGGMMSG ASSRKPLKTG MKEMPVMREK VNEQQRQMGK VGKAHHNHED
IBP2_BRARE  -EVQDPLDIG LTEVPPIRKP TKDSP-WKES AVLQHRQQLK SKMKYHKVED
IBP4_SHEEP  ---KDEGDHP NNSFSPCSAH DRK---CLQK HLAKIRDRST SGGKMKVIGA
```

Reconstructing Phylogenies

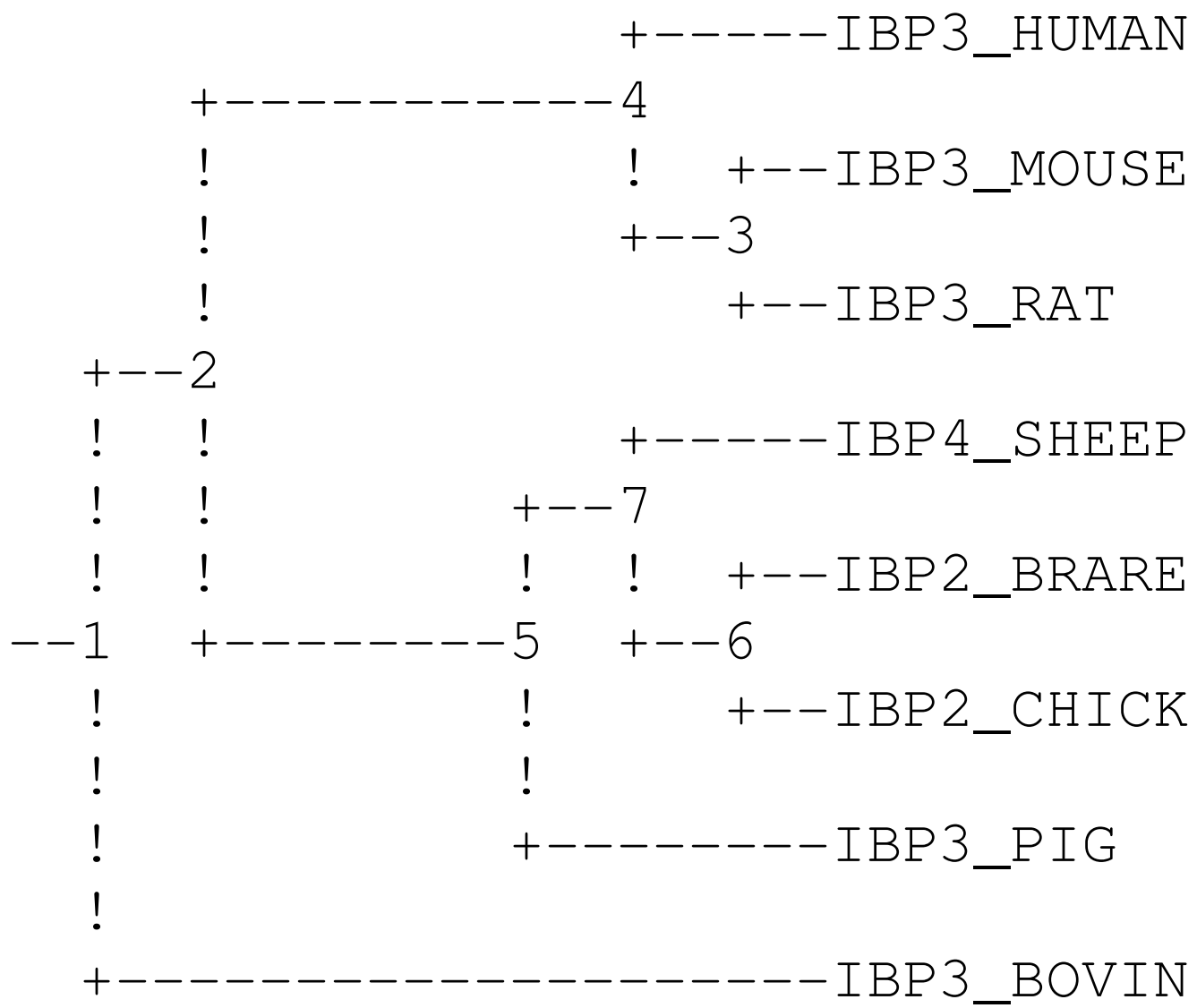
```
IBP3_BOVIN  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNMLS PRGIHIPNCD
  IBP3_PIG  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNMLS PRGIHIPNCD
  IBP3_RAT  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNVLS PRGVHIPNCD
IBP3_MOUSE  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNVLS PRGVHIPNCD
IBP3_HUMAN  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNVLS PRGVHIPNCD
IBP2_CHICK  SKKSRMPTGR TPCQQELDQV LERISTMRLP DERGPLEHLY S--LHIPNCD
IBP2_BRARE  PKAPHAKQ-- SQCQQELDQV LERISKITFK DNRTPLEDLY S--LHIPNCD
IBP4_SHEEP  PREEVRPVPQ GSCQSELHRA LERLAAS--- -QSRTHEDLY I--IPIPNCD

IBP3_BOVIN  KKGFYKKKQC RPSKGRKRGF CWCVDKYGQP LPGFDVKGKG DVHCYSMESK
  IBP3_PIG  KKGFYKKKQC RPSKGRKRGF CWCVDKYGQP LPGFDVKGKG DVHCYSMESK
  IBP3_RAT  KKGFYKKKQC RPSKGRKRGF CWCVDKYGQP LPGYDTKGKD DVHCLSVQSQ
IBP3_MOUSE  KKGFYKKKRC RPSKGRKQSF CWCVDKYGQR LPGYDTKGKD DVHCLSVQSQ
IBP3_HUMAN  KKGFYKKKQC RPSKGRKRGF CWCVDKYGQP LPGYTTKGKE DVHCYSMQSK
IBP2_CHICK  KHGLYNLKQC KMSVNGQRGE CWCVDPIHGK VIQGAPTIRG DPECHLFYTA
IBP2_BRARE  KRGQYNLKQC KMSVNGYRGE CWCVNPHTGR PMPTSPLIRG DPNCNQYLDG
IBP4_SHEEP  RNGNFHPKQC HPALDGQRGK CWCVDRKTGV KLPGGLEPKG ELDCHQLADS

IBP3_BOVIN  ---------- ----
  IBP3_PIG  ---------- ----
  IBP3_RAT  ---------- ----
IBP3_MOUSE  ---------- ----
IBP3_HUMAN  ---------- ----
IBP2_CHICK  HEQEDRGAHA LRSQ
IBP2_BRARE  QE-MDPSVDP PN--
IBP4_SHEEP  FRE------- ----
```

Optimally Parsimonious Tree

This tree is the uniquely optimally parsimonious reconstruction.



(Complements of PHYLIP, via the Biology Workbench)

Distance Matrices

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
IBP3_BOVIN	(1)	0.00	0.09	0.19	0.20	0.17	0.69	0.70	0.67
IBP3_PIG	(2)	0.09	0.00	0.18	0.18	0.16	0.66	0.70	0.67
IBP3_RAT	(3)	0.19	0.18	0.00	0.07	0.17	0.69	0.71	0.68
IBP3_MOUSE	(4)	0.20	0.18	0.07	0.00	0.20	0.71	0.72	0.70
IBP3_HUMAN	(5)	0.17	0.16	0.17	0.20	0.00	0.69	0.71	0.68
IBP2_CHICK	(6)	0.69	0.66	0.69	0.71	0.69	0.00	0.47	0.64
IBP2_BRARE	(7)	0.70	0.70	0.71	0.72	0.71	0.47	0.00	0.62
IBP4_SHEEP	(8)	0.67	0.67	0.68	0.70	0.68	0.64	0.62	0.00

This is the distance matrix generated by PHYLIP. It is a reflection of how “evolutionarily close” two proteins are, and can be reflected in the branch lengths of the tree.

Sequence Distance

Distances between protein sequences are measures of how *dissimilar* two sequences are.

Distances range from 0, meaning that the sequences are identical, to 1, meaning that they are completely dissimilar.

We will not be concerned with how distances are derived, but here's one example. Let *similarity*

$$S = (S_{real} - S_{rand}) / (S_{ident} - S_{rand})$$

S_{real} = actual alignment score of two sequences
S_{rand} = average alignment of the two sequences after many random shuffles of the sequences
S_{ident} = average of the two scores obtained by aligning the sequences with themselves

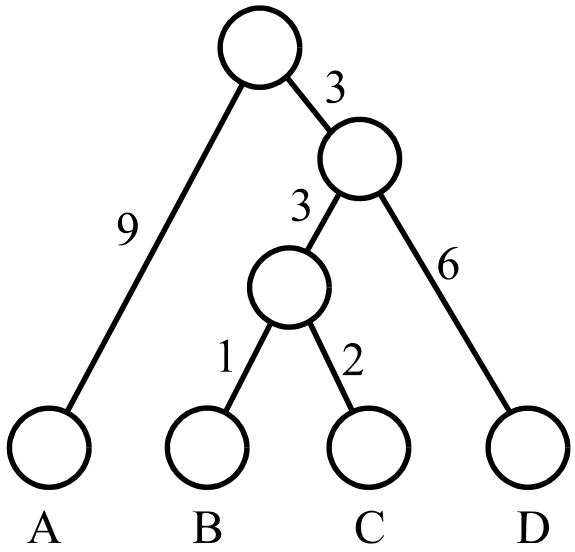
Then distance:

$$D = -\log(S)$$

Treelike Distance Matrices

A distance matrix is called *treelike* if the distances in the matrix correspond to actual distances in a weighted tree between its leaves.

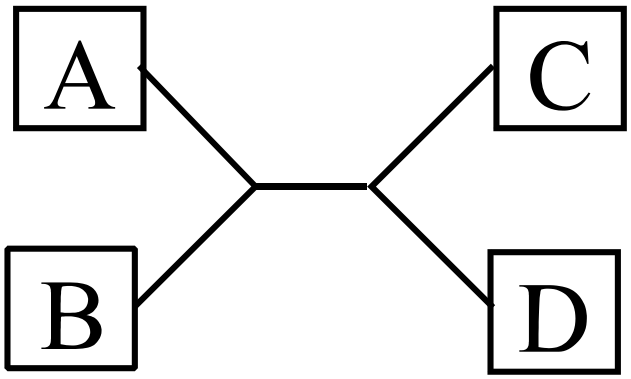
For example:



	A	B	C	D
A				
B				
C				
D				

Neighborliness Method of Sattath and Tversky

Consider four species in some phylogenetic tree:



With distances on the branches of the tree, one of these values will be smallest:

$$d(A, B) + d(C, D)$$
$$d(A, C) + d(B, D)$$
$$d(A, D) + d(B, C)$$

Neighborliness Method of Sattath and Tversky

Start with a distance matrix.

Given a pair of species X and Y, we consider every other pair S and T, and compute how often d(X, Y) + d(S, T) is the smallest of the three sums:

$$d(X, Y) + d(S, T)$$
$$d(X, S) + d(Y, T)$$
$$d(X, T) + d(S,Y)$$

The number of times this happens, is called the *neighborliness of X and Y.*

At each step of the Sattath-Tversky method, we select the most neighborly pair, join them in the tree we are building, replace their entries in the matrix with a single entry, and set the distances to this node from the other nodes to the average of the old distances.

Distance Methods for Tree Inference

UPGMA — Unweighted Pair Group Method with Arithmetic Mean.

Distance methods start with some matrix of distances between the species and iteratively pair together nearest neighbors, replacing their two matrix entries with a new, combined entry representing an internal node of our tree.

For example:

	Bovin	Pig	Rat	Mouse	Human
Bovin	0.00	0.10	0.20	0.20	0.17
Pig	0.10	0.00	0.18	0.19	0.16
Rat	0.20	0.18	0.00	0.07	0.18
Mouse	0.20	0.19	0.07	0.00	0.20
Human	0.17	0.16	0.18	0.20	0.00

UPGMA

1. Begin with a distance matrix and an edgeless tree

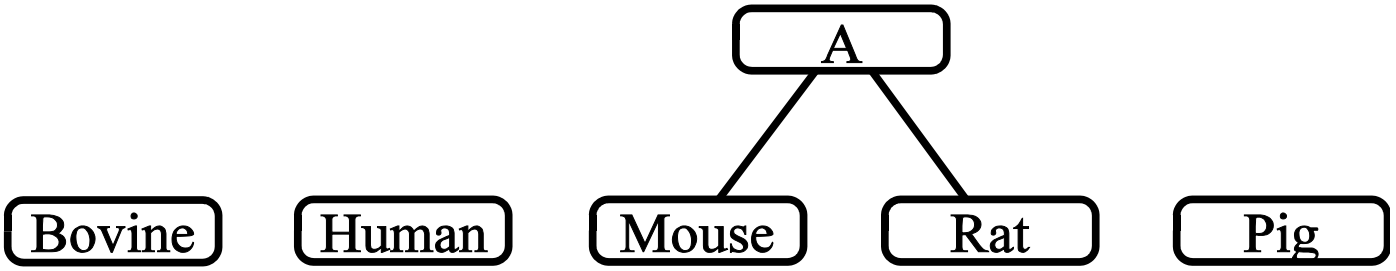
	Bovin	Pig	Rat	Mouse	Human
Bovin	0	0.098	0.197	0.204	0.173
Pig	0.098	0	0.181	0.189	0.163
Rat	0.197	0.181	0	0.072	0.176
Mouse	0.204	0.189	0.072	0	0.204
Human	0.173	0.163	0.176	0.204	0

2. Identify the closest pair

	Bovin	Pig	Rat	Mouse	Human
Bovin	0	0.098	0.197	0.204	0.173
Pig	0.098	0	0.181	0.189	0.163
Rat	0.197	0.181	0	0.072	0.176
Mouse	0.204	0.189	0.072	0	0.204
Human	0.173	0.163	0.176	0.204	0

UPGMA

3. Join them in the tree with an internal node



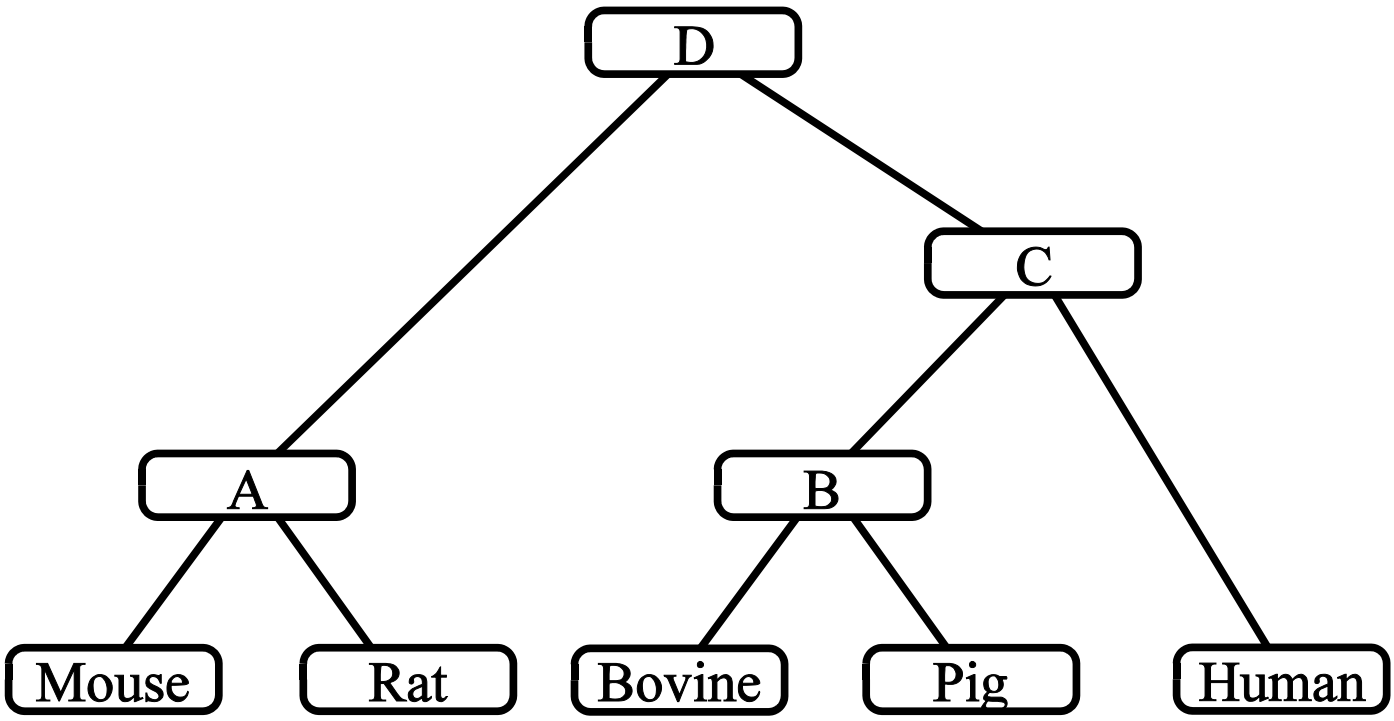
4. Replace them in the matrix with a single, new entry corresponding to that internal node. The new values are the averages of the old ones.

	Bovin	Pig	A	Human
Bovin	0	0.098	0.201	0.173
Pig	0.098	0	0.185	0.163
A	0.201	0.185	0	0.19
Human	0.173	0.163	0.19	0

5. Iterate until all species are connected.

Well... What are you waiting for???

The Tree



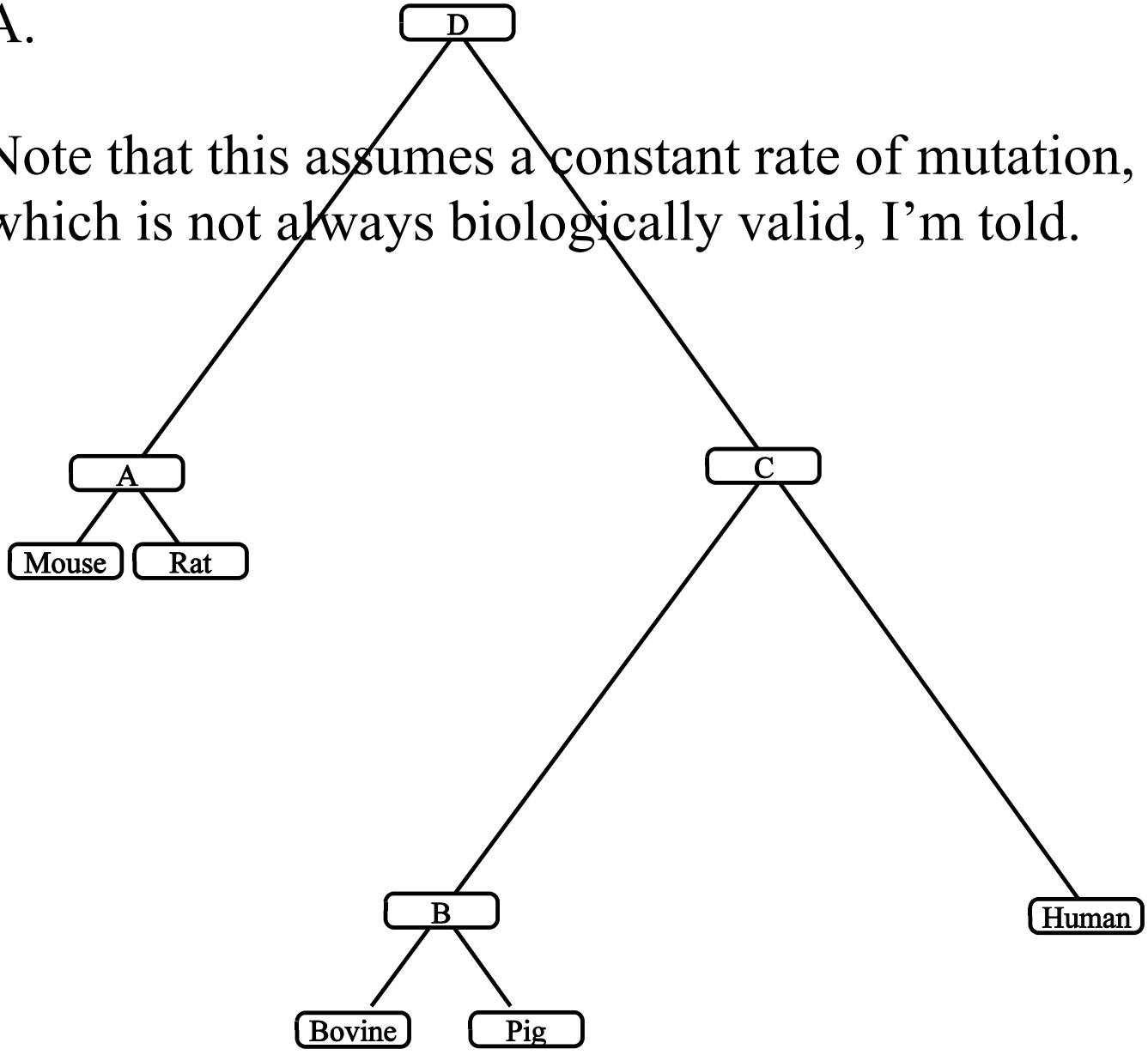
Notice that this tree agrees with the parsimony tree from a few slides ago (these are five of those eight species).

Picturing Divergence with Branch Lengths

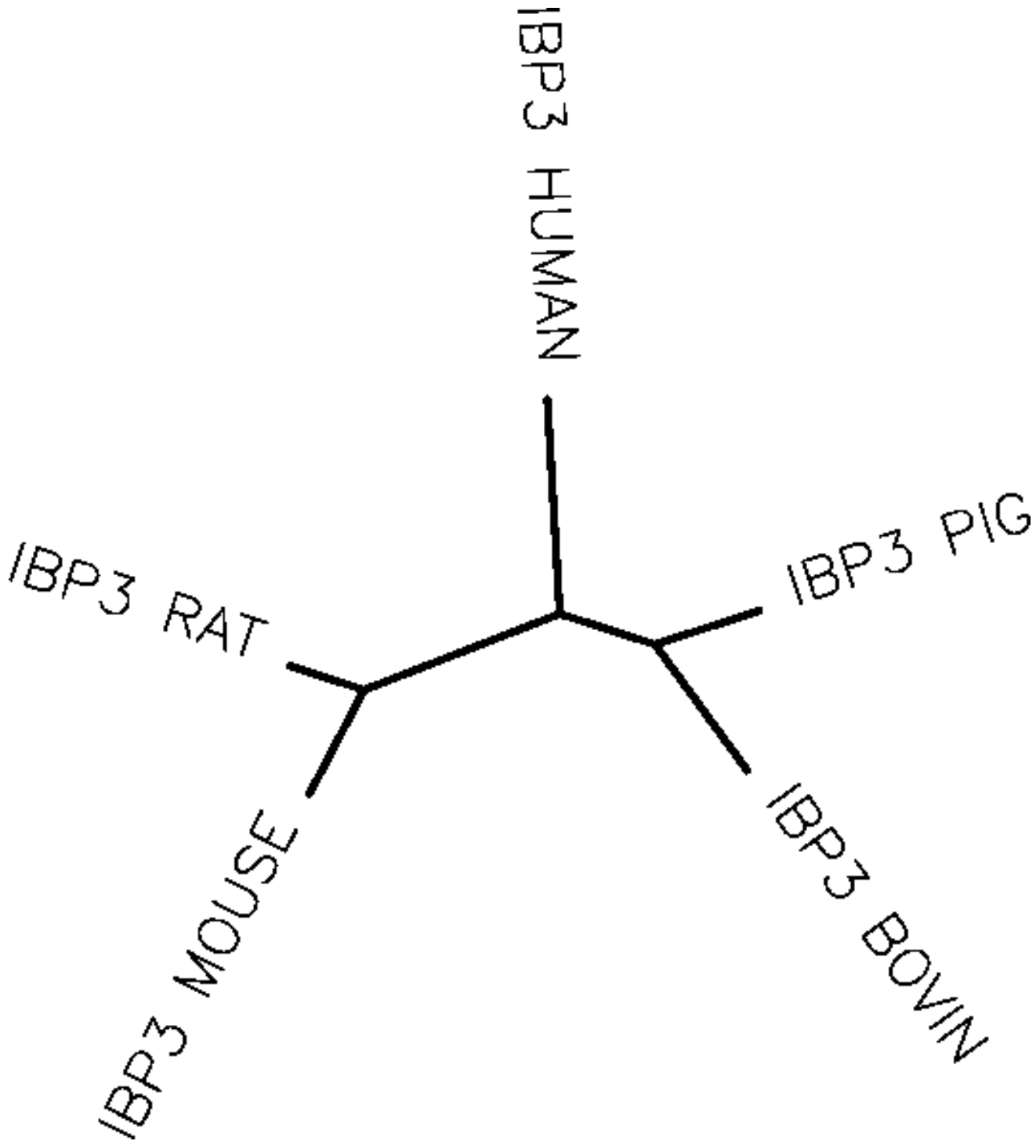
We could have drawn the edges at each step with a length proportional to their distance in the matrix.

Thus the edge from Mouse to A would have had length 0.036, as would have the edge from Rat to A.

Note that this assumes a constant rate of mutation, which is not always biologically valid, I'm told.



PHYLIP Tree



Neighbor Joining
(The method of Studier and Keppler)

If a distance matrix is *treelike*, then their algorithm can quickly recover the tree.

1. Compute $S_{i,j}$ for each pair i, j of species

$$S_{i,j} = (N-2)D_{i,j} - \sum D_{i,k} - \sum D_{j,k}$$

2. Select the pair that gives the least quantity and join them in the tree, and replace them in the matrix with a new entry u

3. Fill the matrix with new values $D_{k,u}$ where

$$D_{k,u} = \frac{1}{2}(D_{i,k} + D_{j,k} - D_{i,j})$$

4. Iterate until the tree is complete

Neighbor Joining

Here is the distance matrix, with marginal sums:

	Bovin	Pig	Rat	Mouse	Human	
Bovin	0	0.098	0.197	0.204	0.173	0.672
Pig	0.098	0	0.181	0.189	0.163	0.631
Rat	0.197	0.181	0	0.072	0.176	0.626
Mouse	0.204	0.189	0.072	0	0.204	0.669
Human	0.173	0.163	0.176	0.204	0	0.716
	0.672	0.631	0.626	0.669	0.716	

1. Compute $S_{i,j}$ for each pair i, j of species

$$S_{i,j} = (N-2)D_{i,j} - \sum D_{i,k} - \sum D_{j,k}$$

   The sums are taken over *all* entries in the row (or column).

2. Select the pair that gives the least quantity and join them in the tree, and replace them in the matrix with a new entry u

Neighbor Joining

3. Fill the matrix with new values $D_{k,u}$ where

$$D_{k,u} = \frac{1}{2}(D_{i,k} + D_{j,k} - D_{i,j})$$

and set the branch lengths

$$L_{i,u} = \frac{1}{2(N-2)}[(N-2)D_{i,j} + \sum_k D_{i,k} - \sum_k D_{j,k}]$$

$$L_{j,u} = \frac{1}{2(N-2)}[(N-2)D_{i,j} + \sum_k D_{j,k} - \sum_k D_{i,k}]$$

(The homework used *D* for these, which was confusing.)

4. Iterate until the tree is complete

Neighbor Joining

Here is our example worked out:

$D_{i,j}$	Bovin	Pig	Rat	Mouse	Human	
Bovin	0	0.098	0.197	0.204	0.173	0.672
Pig	0.098	0	0.181	0.189	0.163	0.631
Rat	0.197	0.181	0	0.072	0.176	0.626
Mouse	0.204	0.189	0.072	0	0.204	0.669
Human	0.173	0.163	0.176	0.204	0	0.716
	0.672	0.631	0.626	0.669	0.716	

Compute the matrix of $S_{i,j}$ values.

$S_{i,j}$	Bovin	Pig	Rat	Mouse	Human
Bovin		-1.009	-0.707	-0.729	-0.869
Pig			-0.714	-0.733	-0.858
Rat				-1.079	-0.814
Mouse					-0.773
Human					

Select the least value. This is between rat and mouse

Branch Lengths:

$L_{R,RM} = (1/6)(0.072*3 + 0.626 - 0.669) = 0.0288$

$L_{M,RM} = (1/6)(0.072*3 - 0.626 + 0.669) = 0.0432$

Neighbor Joining

Compute a new distance matrix:

$D_{i,j}$	Bovin	Pig	RM	Human	
Bovin	0	0.098	0.1645	0.173	0.4355
Pig	0.098	0	0.149	0.163	0.41
RM	0.1645	0.149	0	0.154	0.4675
Human	0.173	0.163	0.154	0	0.49
	0.4355	0.41	0.4675	0.49	

Compute the matrix of $S_{i,j}$ values.

$S_{i,j}$	Bovin	Pig	RM	Human
Bovin		-0.6495	-0.574	-0.5795
Pig			-0.5795	-0.574
RM				-0.6495
Human				

Select the least value. There is a tie (can you see why there has to be?). We'll join Bovin with Pig.

$L_{B,BP} = (1/4)(0.098*2 + 0.4355 - 0.41) = 0.0554$
$L_{P,BP} = (1/4)(0.098*2 - 0.4355 + 0.41) = 0.0426$

Neighbor Joining

Compute a new distance matrix:

$D_{i,j}$	BP	RM	Human	
BP	0	0.1082	0.119	0.2272
RM	0.1082	0	0.154	0.2622
Human	0.119	0.154	0	0.273
	0.2272	0.2622	0.273	

Compute the matrix of $S_{i,j}$ values.

$S_{i,j}$	BP	RM	Human
Bovin		-0.3812	-0.3812
RM			-0.3812
Human			

Select the least value. There is a tie (Because there is just one unrooted tree on 3 vertices). Let's join RM with Human, adding a new internal vertex A.

$L_{A,RM} = (1/2)(0.154 + 0.2622 - 0.273) = 0.0716$

$L_{A,Human} = (1/2)(0.154 - 0.2622 + 0.273) = 0.0824$

and

$L_{BP,A} = (1/2)(0.1082 + 0.119 - 0.154) = 0.0366$

Handout #1 — Some Aligned Insulin Growth Binding Protein Sequences

```
IBP3_BOVIN  --MLRAPPRL WAAALTALTL LRGPPAARAG AGTMGAGPVV RCEPCDARAV
  IBP3_PIG  ---------- ---------- ---------G SGAVGTGPVV RCEPCDARAL
  IBP3_RAT  --MHPARPAL WAAALTALTL LRGPPVARAG AGAVGAGPVV RCEPCDARAL
IBP3_MOUSE  --MHPARPAL WAAALTALTL LRGPPVAELA AGAVG-GPVV RCEPCDARAV
IBP3_HUMAN  --MQRARPTL WAAALTLLVL LRGPPVARAG ASSGGLGPVV RCEPCDARAL
IBP2_CHICK  MALGGVGRGG AARAAWPRLL LAALAPALAL AGPALPEVLF RCPPCTAERL
IBP2_BRARE  -MLSYVSCG- --------LL LALVT----F HGTARSEMVF RCPSCTAERQ
IBP4_SHEEP  ---------- ---------- ---------- ------DEAI HCPPCSEEKL

IBP3_BOVIN  AQCAPPPPSP PCAELVRDAG CGCCLTCALR EGQPCGVYTE RCGSGLRCQP
  IBP3_PIG  AQCAPPPAAP PCAELVREPG CGCCLTCALR EGQACGVYTE RCGAGLRCQP
  IBP3_RAT  AQCAPPPTAP ACTELVREPG CGCCLTCALR EGDACGVYTE RCGTGLRCQP
IBP3_MOUSE  SQCAPPPTAP ACTELVREPG CGCCLTCALR EGDACGVYTE RCGTGLRCQP
IBP3_HUMAN  AQCAPPP--A VCAELVREPG CGCCLTCALS EGQPCGIYTE RCGSGLRCQP
IBP2_CHICK  AACSP-AARP PCPELVREPG CGCCPVCARL EDEACGVYTP RCAAGLRCYP
IBP2_BRARE  AAC-P-MLTE TCGEIVREPG CGCCPVCARQ EGEQCGVYTP RCSSGLRCYP
IBP4_SHEEP  ARCRP-PVG- -CEELVREPG CGCCATCALG KGMPCGVYTP DCGSGLRCHP

IBP3_BOVIN  PPGDPRPLQA LLDGRGLCAN ASAVGRLRPY LLPS--ASGN GSES-----E
  IBP3_PIG  PPGEPRPLQA LLDGRGICAN ASAAGRLRAY LLPAPPAPGN GSES-----E
  IBP3_RAT  RPAEQYPLKA LLNGRGFCAN ASAASNLSAY -LPSQPSPGN TTES-----E
IBP3_MOUSE  RPAEQYPLRA LLNGRGFCAN ASAAGSLSTY -LPSQPAPGN ISES-----E
IBP3_HUMAN  SPDEARPLQA LLDGRGLCVN ASAVSRLRAY LLPAPPAPGN ASES-----E
IBP2_CHICK  DPGAELPPQA LVQGQGTCAR PPDTDEYGAS TEPPADNGDD RSESILAENH
IBP2_BRARE  KPDSELPLEL LVQGLGRCGR KVDTEPTG-S AEPREVSG-- ----------
IBP4_SHEEP  PRGVEKPLHT LVHGQGVCME LAEIEAIQES LQPSD----- ----------

IBP3_BOVIN  EDHSMGSTEN QAGPSTHRVP VSKFHPIHTK MDVIKKGHAK DSQRYKVDYE
  IBP3_PIG  EDRSVDSMEN QALPSTHRVP DSKLHSVHTK MDVIKKGHAK DSQRYKVDYE
  IBP3_RAT  EDHNAGSVES QVVPSTHRVT DSKFHPLHSK MEVIIKGQAR DSQRYKVDYE
IBP3_MOUSE  EEHNAGSVES QVVPSTHRVT DSKFHPLHAK MDVIKKGHAR DSQRYKVDYE
IBP3_HUMAN  EDRSAGSVES PSVSSTHRVS DPKFHPLHSK IIIIKKGHAK DSQRYKVDYE
IBP2_CHICK  VDSTGGMMSG ASSRKPLKTG MKEMPVMREK VNEQQRQMGK VGKAHHNHED
IBP2_BRARE  -EVQDPLDIG LTEVPPIRKP TKDSP-WKES AVLQHRQQLK SKMKYHKVED
IBP4_SHEEP  ---KDEGDHP NNSFSPCSAH DRK---CLQK HLAKIRDRST SGGKMKVIGA

IBP3_BOVIN  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNMLS PRGIHIPNCD
  IBP3_PIG  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNMLS PRGIHIPNCD
  IBP3_RAT  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNVLS PRGVHIPNCD
IBP3_MOUSE  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNVLS PRGVHIPNCD
IBP3_HUMAN  SQSTDTQNFS SESKRETEYG PCRREMEDTL NHLKFLNVLS PRGVHIPNCD
IBP2_CHICK  SKKSRMPTGR TPCQQELDQV LERISTMRLP DERGPLEHLY S--LHIPNCD
IBP2_BRARE  PKAPHAKQ-- SQCQQELDQV LERISKITFK DNRTPLEDLY S--LHIPNCD
IBP4_SHEEP  PREEVRPVPQ GSCQSELHRA LERLAAS--- -QSRTHEDLY I--IPIPNCD

IBP3_BOVIN  KKGFYKKKQC RPSKGRKRGF CWCVDKYGQP LPGFDVKGKG DVHCYSMESK
  IBP3_PIG  KKGFYKKKQC RPSKGRKRGF CWCVDKYGQP LPGFDVKGKG DVHCYSMESK
  IBP3_RAT  KKGFYKKKQC RPSKGRKRGF CWCVDKYGQP LPGYDTKGKD DVHCLSVQSQ
IBP3_MOUSE  KKGFYKKKRC RPSKGRKQSF CWCVDKYGQR LPGYDTKGKD DVHCLSVQSQ
IBP3_HUMAN  KKGFYKKKQC RPSKGRKRGF CWCVDKYGQP LPGYTTKGKE DVHCYSMQSK
IBP2_CHICK  KHGLYNLKQC KMSVNGQRGE CWCVDPIHGK VIQGAPTIRG DPECHLFYTA
IBP2_BRARE  KRGQYNLKQC KMSVNGYRGE CWCVNPHTGR PMPTSPLIRG DPNCNQYLDG
IBP4_SHEEP  RNGNFHPKQC HPALDGQRGK CWCVDRKTGV KLPGGLEPKG ELDCHQLADS
```

Handout #2 — UPGMA method of phylogenetic tree construction

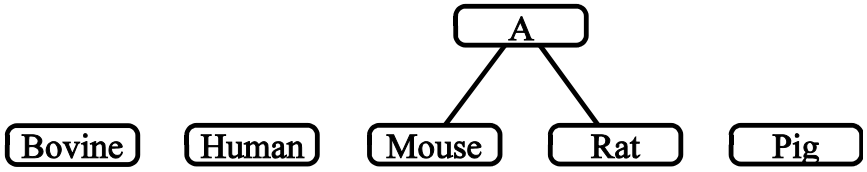
1. Begin with a distance matrix and an edgeless tree

	Bovin	Pig	Rat	Mouse	Human
Bovin	0	0.098	0.197	0.204	0.173
Pig	0.098	0	0.181	0.189	0.163
Rat	0.197	0.181	0	0.072	0.176
Mouse	0.204	0.189	0.072	0	0.204
Human	0.173	0.163	0.176	0.204	0

2. Identify the closest pair

	Bovin	Pig	Rat	Mouse	Human
Bovin	0	0.098	0.197	0.204	0.173
Pig	0.098	0	0.181	0.189	0.163
Rat	0.197	0.181	0	0.072	0.176
Mouse	0.204	0.189	0.072	0	0.204
Human	0.173	0.163	0.176	0.204	0

3. Join them in the tree with an internal node



4. Replace them in the matrix with a single, new entry corresponding to that internal node. The new values are the averages of the old ones.

	Bovin	Pig	A	Human
Bovin	0	0.098	0.201	0.173
Pig	0.098	0	0.185	0.163
A	0.201	0.185	0	0.19
Human	0.173	0.163	0.19	0

5. Iterate until all species are connected.

Handout #3 — The Neighbor Joining Method of Studier and Keppler

1. Compute $S_{i,j}$ for each pair i, j of species

$$S_{i,j} = (N-2)D_{i,j} - \sum D_{i,k} - \sum D_{j,k}$$

2. Select the pair that gives the least quantity and join them in the tree, and replace them in the matrix with a new entry u

3. Fill the matrix with new values $D_{k,u}$ where

$$D_{k,u} = \frac{1}{2}(D_{i,k} + D_{j,k} - D_{i,j})$$

4. Iterate until the tree is complete

	Bovin	Pig	Rat	Mouse	Human
Bovin	0	0.098	0.197	0.204	0.173
Pig	0.098	0	0.181	0.189	0.163
Rat	0.197	0.181	0	0.072	0.176
Mouse	0.204	0.189	0.072	0	0.204
Human	0.173	0.163	0.176	0.204	0

Handout #4 — Neighbor Joining Example

Here is our example worked out:

$D_{i,j}$	Bovin	Pig	Rat	Mouse	Human	
Bovin	0	0.098	0.197	0.204	0.173	0.672
Pig	0.098	0	0.181	0.189	0.163	0.631
Rat	0.197	0.181	0	0.072	0.176	0.626
Mouse	0.204	0.189	0.072	0	0.204	0.669
Human	0.173	0.163	0.176	0.204	0	0.716
	0.672	0.631	0.626	0.669	0.716	

Compute the matrix of $S_{i,j}$ values.

$S_{i,j}$	Bovin	Pig	Rat	Mouse	Human
Bovin		-1.009	-0.707	-0.729	-0.869
Pig			-0.714	-0.733	-0.858
Rat				-1.079	-0.814
Mouse					-0.773
Human					

Select the least value. This is between rat and mouse

Branch Lengths:

$L_{R,RM} = (1/6)(0.072*3 + 0.626 - 0.669) = 0.0288$

$L_{M,RM} = (1/6)(0.072*3 - 0.626 + 0.669) = 0.0432$

Compute a new distance matrix:

$D_{i,j}$	Bovin	Pig	RM	Human	
Bovin	0	0.098	0.1645	0.173	0.4355
Pig	0.098	0	0.149	0.163	0.41
RM	0.1645	0.149	0	0.154	0.4675
Human	0.173	0.163	0.154	0	0.49
	0.4355	0.41	0.4675	0.49	

Compute the matrix of $S_{i,j}$ values.

$S_{i,j}$	Bovin	Pig	RM	Human
Bovin		-0.6495	-0.574	-0.5795
Pig			-0.5795	-0.574
RM				-0.6495
Human				

Select the least value. There is a tie (can you see why there has to be?). We'll join Bovin with Pig.

$L_{B,BP} = (1/4)(0.098*2 + 0.4355 - 0.41) = 0.0554$

$L_{P,BP} = (1/4)(0.098*2 - 0.4355 + 0.41) = 0.0426$

Compute a new distance matrix:

$D_{i,j}$	BP	RM	Human	
BP	0	0.1082	0.119	0.2272
RM	0.1082	0	0.154	0.2622
Human	0.119	0.154	0	0.273
	0.2272	0.2622	0.273	

Compute the matrix of $S_{i,j}$ values.

$S_{i,j}$	BP	RM	Human
Bovin		-0.3812	-0.3812
RM			-0.3812
Human			

Select the least value. There is a tie (Because there is just one unrooted tree on 3 vertices). Let's join RM with Human, adding a new internal vertex A.

$L_{A,RM} = (1/2)(0.154 + 0.2622 - 0.273) = 0.0716$

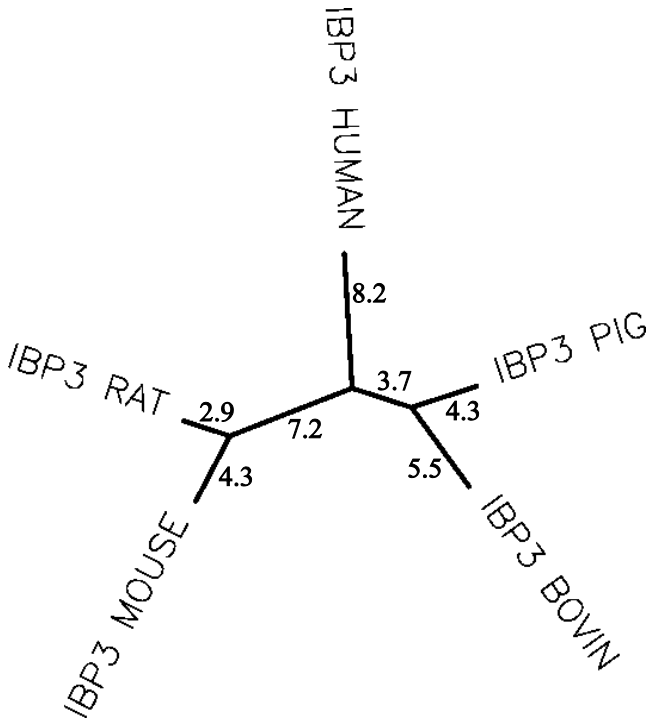
$L_{A,Human} = (1/2)(0.154 - 0.2622 + 0.273) = 0.0824$

and

$L_{BP,A} = (1/2)(0.1082 + 0.119 - 0.154) = 0.0366$

And we are done.

The Resulting Tree



Exercises — Phylogeny

Quick Concepts:

1. Use the UPGMA method to discern the phylogeny tree on the following six species, where the distances have been scaled and rounded to be nice integers. Draw your answer as a rooted tree.

	A	B	C	D	E	F
A	0	4	11	7	9	10
B		0	3	12	4	6
C			0	9	7	7
D				0	2	13
E					0	8
F						0

2. Suppose you are given a treelike matrix and you are told that all the weights on the edges are "1." Does that make it any easier to reconstruct the tree? Try it with the following treelike matrix, which does correspond to such an (unrooted) tree:

	A	B	C	D	E	F
A	0	3	4	3	4	3
B		0	5	2	5	4
C			0	5	2	3
D				0	5	4
E					0	3
F						0

3. Please make sure you have an account on Biology Student Workbench. Create a new session called "Homework 5."

4. Recall from the workshop that one type of similarity is that given by:

$$S = (S_{real} - S_{rand}) / (S_{ident} - S_{rand})$$

where

S_{real} = actual alignment score of two sequences
S_{rand} = average alignment of the two sequences after many random shuffles of the sequences
S_{ident} = average of the two scores obtained by aligning the sequences with themselves

Explain why S will usually be between 0 and 1, that S will be close to 1 when the sequences are very closely related, and S will be close to 0 when the sequences are not related.

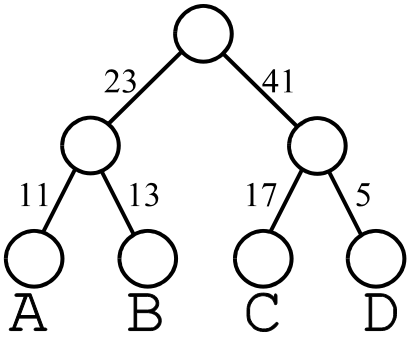
Presentation Problems:

5. Go to bsw-uiuc.net, and enther the portal to the Student Interface to Biology Workbench.
   a. In your session "Homework 5"(which you created in the previous problem) go to protein tools and use "Ndjinn" to find the human P53 protein in the SwissProt protein sequence database. Import that sequence into your session, along with the P53 protein sequences for at least 9 other species, including Chinese hamster and the Common tree shrew.
   b. Back in protein tools, select your sequences (at the bottom of the protein tools page) and use CLUSTALW to align these protein sequences. (Note that this will perform a *multiple sequence alignment*, which is not something we have talked about.)
   c. When the alignment is done, scan down the page to see the alignment.
      1. What do the blue letters mean?
      2. What do the asterisks, colons and periods at the bottom of the alignment mean?
      3. Back near the top of that alignment page, click the "Import Alignment" button to import this alignment into your session. Note that alignments are treated as objects, and can be manipulated in the "Alignment Tools" portion of the Workbench.
   d. In the Alignment Tools portion of the Workbench (you are probably already there if you just imported your alignment) select the alignment (near the bottom of the page) and use BOXSHADE to see a nifty view of this alignment. What is your favorite feature of the BOXSHADE routine? What does "consensus" mean on that page?
   e. Return to Alignment Tools. Select your alignment, and use DRAWTREE to see the PHYLIP tree that Prof. Felsenstein's program creates. What inferences do you draw about the evolutionary history of the species in your tree. (Be aware that some of the branches of the tree might be quite small...
   f. Is this a rooted or unrooted tree?
   g. Repeat steps e. and f. for the DRAWGRAM function, instead of DRAWTREE.
   h. Select another protein and find that protein in as many of the species you used in this problem as possible. (Should be at least five.) Create the DRAWGRAM for those species, and compare it to the DRAWGRAM created for the P53 in *that set* of species. (This will require creating a new alignment for just that set of species and their P53 proteins.)
      1. Should it be the same tree?
      2. Must it be the same tree?
      3. Is it the same tree?

6. Below is the distance matrix for four species. Use UPGMA to draw the phylogeny tree of these species with edge lengths proportional to the distances between the joined pairs. To get the distances between the joined pairs, each time you add a new node, let it split the distance between the two joined species, as given in the matrix.

	A	B	C	D
A	0	32	14	22
B		0	14	19
C			0	25
D				0

7. What treelike matrix arises from the weighted tree below?



8. Recall that a "treelike" distance matrix is one which arose from a weighted binary tree by setting the distance between each pair of vertices to the sum of the weights on the edges on the path between them. The matrices below are treelike. In each case, obtain and draw the corresponding (unrooted) phylogeny tree using whatever method you wish.

	A	B
A	0	5
B		0

	A	B	C
A	0	3	4
B		0	5
C			0

	A	B	C	D
A	0	10	4	12
B		0	12	6
C			0	14
D				0

9. Perform one (or more, if you really have nothing better to do) iteration of the algorithm of Studier and Keppler on the distance matrix below. That means you should find the closest pair, join them and replace their entries in the matrix with a new entry whose distances to the remaining entries are given by the formula on Handout #3.

	A	B	C	D	E
A	0	7	3	3	11
B		0	9	1	21
C			0	13	2
D				0	5
E					0

10. The branch lengths connecting the new node to the two joined leaves are given by the formulas below. What are these lengths? Show that the sum of these lengths is equal to what it should be.

$$L_{i,u} = \frac{1}{2(N-2)}[(N-2)D_{i,j} + \sum_k D_{i,k} - \sum_k D_{j,k}]$$

$$L_{j,u} = \frac{1}{2(N-2)}[(N-2)D_{i,j} + \sum_k D_{j,k} - \sum_k D_{i,k}]$$