

Some Protein Sequences

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

>IBP3_BOVIN

```
--MLRAPPRLLWAAALTALTLLRGPPAARAGAGTMGAGPVVRCEPCDARAVAQCAPPPPSPPCAELVRDAG
CGCCLTCALREGQPCGVYTERCGSGLRCQPPPGDPRPLQALLDGRGLCANASAVGRLRPYLLPS--ASGN
GSES-----EEDHSMGSTENQAGPSTHRVPVSKFHPIHTKMDVIKKGHAKDSQRYKVDYESQSTDTQNF
SESKRETEYGPCRREMEDTLNHLKFLNMLSPRGIHIPNCDKKGFIKKKQCRPSKGRKRGFCWCVDKYGQP
LPGFDVKGKGDVHCYSMESK-----
```

>IBP3_PIG

```
-----GSGAVGTGPVVRCEPCDARALAQCAPPPAAPPCAELVREPG
CGCCLTCALREGQACGVYTERCGAGLRCQPPPGEPRLQALLDGRGICANASAAGRLRAYLLPAPPAPGN
GSES-----EEDRSVDSMENQALPSTHRVPDSKLHSVHTKMDVIKKGHAKDSQRYKVDYESQSTDTQNF
SESKRETEYGPCRREMEDTLNHLKFLNMLSPRGIHIPNCDKKGFIKKKQCRPSKGRKRGFCWCVDKYGQP
LPGFDVKGKGDVHCYSMESK-----
```

>IBP3_RAT

```
--MHPARPALWAAALTALTLLRGPPVARAGAGAVGAGPVVRCEPCDARALAQCAPPPTAPACTELVREPG
CGCCLTCALREGDACGVYTERCGTGLRCQPRPAEQYPLKALLNNGRGFCANASAASNLSAY-LPSQPSPGN
TTES-----EEDHNAGSVESQVVPSTHRVTDKSFHPLHLSKMEV I IKGQARDSQRYKVDYESQSTDTQNF
SESKRETEYGPCRREMEDTLNHLKFLNVLSPRGVHIPNCDKKGFIKKKQCRPSKGRKRGFCWCVDKYGQP
LPGYDTKGKDDVHCLSVQSQ-----
```

>IBP3_MOUSE

```
--MHPARPALWAAALTALTLLRGPPVAELAAGAVG-GPVVRCEPCDARAVSQCAPPPPTAPACTELVREPG
CGCCLTCALREGDACGVYTERCGTGLRCQPRPAEQYPLRALNNGRGFCANASAAGSLSTY-LPSQPAPGN
ISES-----EEEHNAGSVESQVVPSTHRVTDKSFHPLHAKMDVIKKGHARDSQRYKVDYESQSTDTQNF
SESKRETEYGPCRREMEDTLNHLKFLNVLSPRGVHIPNCDKKGFIKKKRCRPSKGRKQSFWCVDKYGQR
LPGYDTKGKDDVHCLSVQSQ-----
```

>IBP3_HUMAN

```
--MQRARPTLWAAALTLLVLLRGPPVARAGASSGGLGPVVRCEPCDARALAQCAPP--AVCAELVREPG
CGCCLTCALSEGQPCGIYTERCGSGLRCQPSPEARPLQALLDGRGLCVNASAVSRLRAYLLPAPPAPGN
ASES-----EEDRSAGSVESPSVSSTHRVSDPKFHPLHSKII I IKKGHAKDSQRYKVDYESQSTDTQNF
SESKRETEYGPCRREMEDTLNHLKFLNVLSPRGVHIPNCDKKGFIKKKQCRPSKGRKRGFCWCVDKYGQP
LPGYTTKGKEDVHCYSMQSK-----
```

Some Protein Sequences

```
>IBP2_CHICK
MALGGVGRGGAARAAPRLLLLAALAPALALAGPALPEVLFRCPPTAERLAACSP-AARPPCPELVREPG
CGCCPVCARLEDEACGVYTPRCAAGLRCPDPGAELPPQALVQGQGTARPPDTDEYGASTEPPADNGDD
RSESI LAENHVDSTGGMMMSGASSRKPLKTGMKEMPVMREKVNEQQRQMGKVGKAHHNHEDSKKSRMPTGR
TPCQQELDQVLERISTMRLPDERGPLEHLYS--LHIPNCDKHGLYNLQCKMSVNGQRGECWCVDPIHGK
VIQGAPTIRGDPECHLFYTAHEQEDRGAAHALRSQ
```

```
>IBP2_BRARE
-MLSIVSCG-----LLLALVT----FHGTARSEMVFRCPSCTAERQAAC-P-MLTETTCGEIVREPG
CGCCPVCARQEGEQCGVYTPRCSSGLRCYPKPDSELPLELLVQGLGRCGRKVDTEPTG-SAEPREVSG--
-----EVQDPLDIGLTEVPPIRKPTKDSP-WKESAVLQHRQQLKSKMKYHKVEDPKAPHAKQ--
SQCQQELDQVLERISKITFKDNRTPLEDLYS--LHIPNCDKRGQYNLQCKMSVNGYRGECWCVNPHTGR
PMPTSPILIRGDPNCNQYLDGQE-MDPSVDPPN--
```

```
>IBP4_SHEEP
-----DEAIHCPPCSEEKLARCRP-PVG--CEELVREPG
CGCCATCALGKGMPGCVYTPDCGSLRCHPPRGVEKPLHTLVHGQVCMELAEIEAIQESLQPSD-----
-----KDEGDHPNNSFSPCSAHDRK---CLQKHLAKIRDRSTSGGKMKVIGAPREEVRPVPQ
GSCQSELHRALELAAS----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

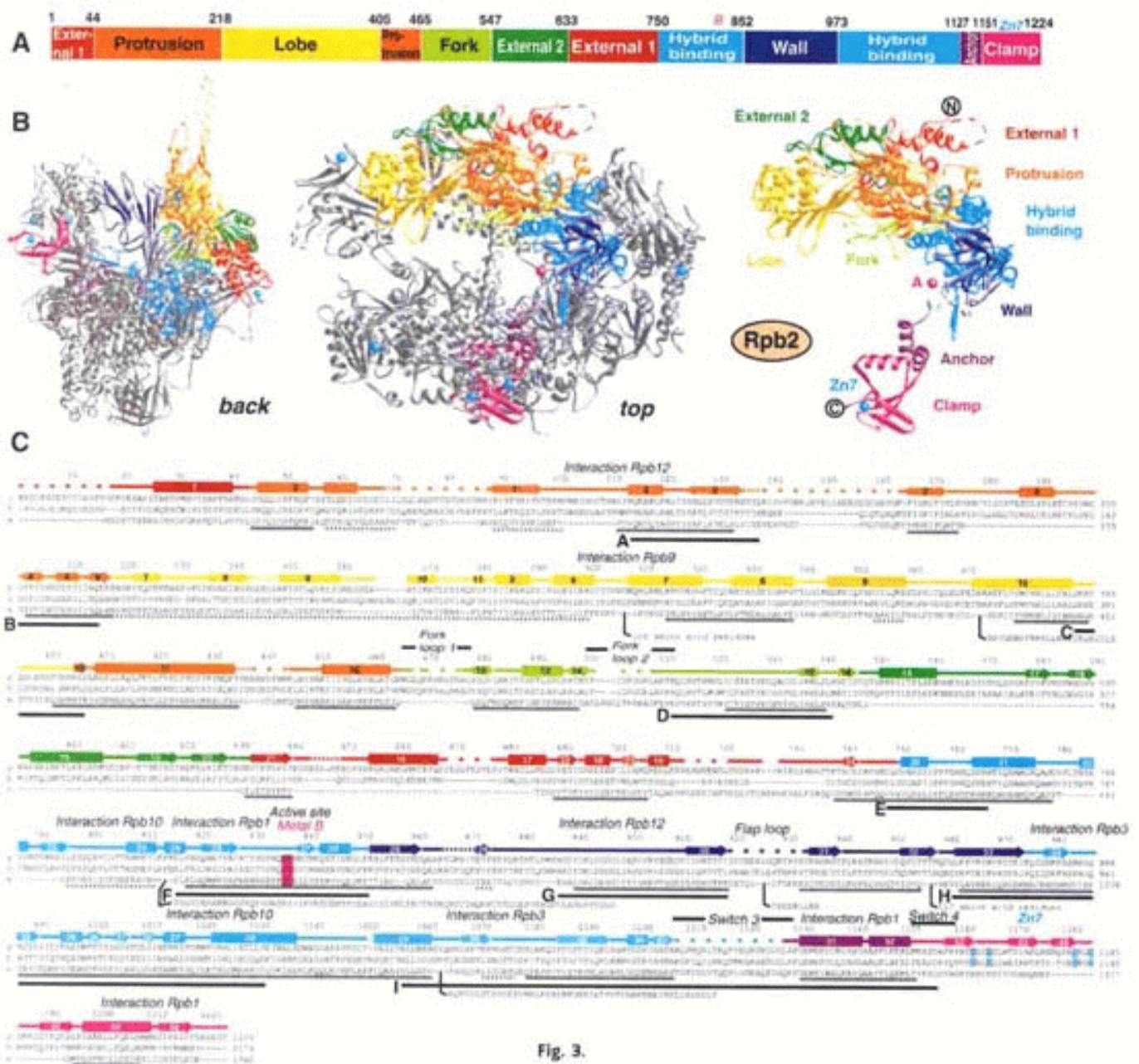


Fig. 3.

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	WAAALTALT	LRGPPAARAG	AGTMGAGPVV	RCEPCDARAV
IBP3_PIG	-----	-----	-----G	SGAVGTGPVV	RCEPCDARAL
IBP3_RAT	--MHPARPAL	WAAALTALT	LRGPPVARAG	AGAVGAGPVV	RCEPCDARAL
IBP3_MOUSE	--MHPARPAL	WAAALTALT	LRGPPVAELA	AGAVG-GPVV	RCEPCDARAV
IBP3_HUMAN	--MQRARPTL	WAAALTLLVL	LRGPPVARAG	ASSGGLGPVV	RCEPCDARAL
IBP2_CHICK	MALGGVGRGG	AARAAWPRL	LAALAPALAL	AGPALPEVLF	RCPPCTAERL
IBP2_BRARE	-MLSIVSCG-	-----LL	LALVT----F	HGTARSEMV	RCPSCAERQ
IBP4_SHEEP	-----	-----	-----	-----DEAI	HCPPCSEEL
IBP3_BOVIN	AQCAPPSP	PCAEVLVDAG	CGCCLTCALR	EGQPCGVYTE	RCGSGLRQCQ
IBP3_PIG	AQCAPPAP	PCAEVLVREP	CGCCLTCALR	EGQACGVYTE	RCGAGLRQCQ
IBP3_RAT	AQCAPPPTAP	ACTELVREP	CGCCLTCALR	EGDACGVYTE	RCGTGLRCQ
IBP3_MOUSE	SQCAPPPTAP	ACTELVREP	CGCCLTCALR	EGDACGVYTE	RCGTGLRCQ
IBP3_HUMAN	AQCAPP--A	VCAELVREP	CGCCLTCALS	EGQPCGIYTE	RCGSGLRQCQ
IBP2_CHICK	AACSP-AARP	PCPELVREP	CGCCPVCARL	EDEACGVYTP	RCAAGLRCCY
IBP2_BRARE	AAC-P-MLTE	TCGEIVREP	CGCCPVCARQ	EGEQCGVYTP	RCSSGLRCY
IBP4_SHEEP	ARCRP-PVG-	-CEELVREP	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	LVQGGGTCTAR	PPDTDEYAS	TEPPADNGDD	RSESILAENH
IBP2_BRARE	KPDSELPLEL	LVQGLGRCGR	KVDTEPTG-S	AEPREVSG--	-----
IBP4_SHEEP	PRGVEKPLHT	LVHGQVCME	LAEIEAIQES	LQPSD-----	-----
IBP3_BOVIN	EDHSMGSTEN	QAGPSTHRVP	VSKFHPIHTK	MDVIKKGHAK	DSQRYKVDYE
IBP3_PIG	EDRSVDSMEN	QALPSTHRVP	DSKLHSHVHTK	MDVIKKGHAK	DSQRYKVDYE
IBP3_RAT	EDHNAGSVES	QVVPSTHRVT	DSKFHPLHSK	MEVILKQAR	DSQRYKVDYE
IBP3_MOUSE	EEHNAGSVES	QVVPSTHRVT	DSKFHPLHAK	MDVIKKGHAR	DSQRYKVDYE
IBP3_HUMAN	EDRSAGSVES	PSVSSTHRVS	DPKFHPLHSK	IIIIKKGHAK	DSQRYKVDYE
IBP2_CHICK	VDSTGGMSG	ASSRKPLKTG	MKEMPVMREK	VNEQQRQMGK	VGAHHNHED
IBP2_BRARE	-EVQDPLDIG	LTEVPPIRKP	TKDSP-WKES	AVLQHRQQLK	SKMKYHKVED
IBP4_SHEEP	---KDEGDHP	NNSFSPCSAH	DRK---CLQK	HLAKIRDRST	SGGKMKVIGA

Reconstructing Phylogenies

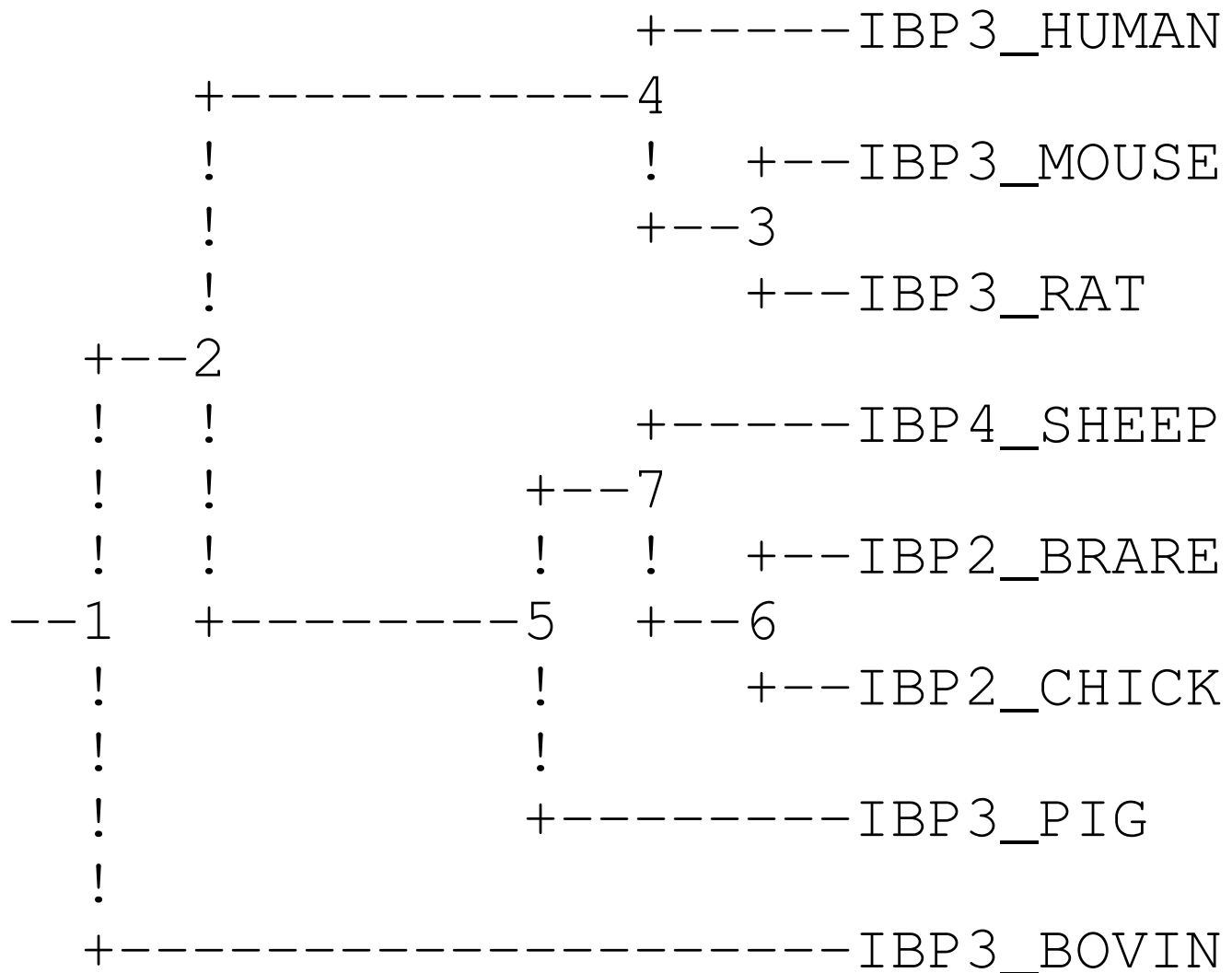
IBP3_BOVIN	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNMLS	PRGIHIPNCD
IBP3_PIG	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNMLS	PRGIHIPNCD
IBP3_RAT	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNVLS	PRGVHIPNCD
IBP3_MOUSE	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNVLS	PRGVHIPNCD
IBP3_HUMAN	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNVLS	PRGVHIPNCD
IBP2_CHICK	SKKSRMPTGR	TPCQQELDQV	LERISTMRLP	DERGPLEHLY	S--LHIPNCD
IBP2_BRARE	PKAPHAKQ--	SQCQQELDQV	LERISKITFK	DNRTPLEDLY	S--LHIPNCD
IBP4_SHEEP	PREEVRFPVQ	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	PMPTSPILRG	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-MDP	SVDP PN--
IBP4_SHEEP	FRE-----	----

Optimally Parsimonious Tree

This tree is the uniquely optimally parsimonious reconstruction.



(Complements of PHYLIP, via the Biology Workbench)

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

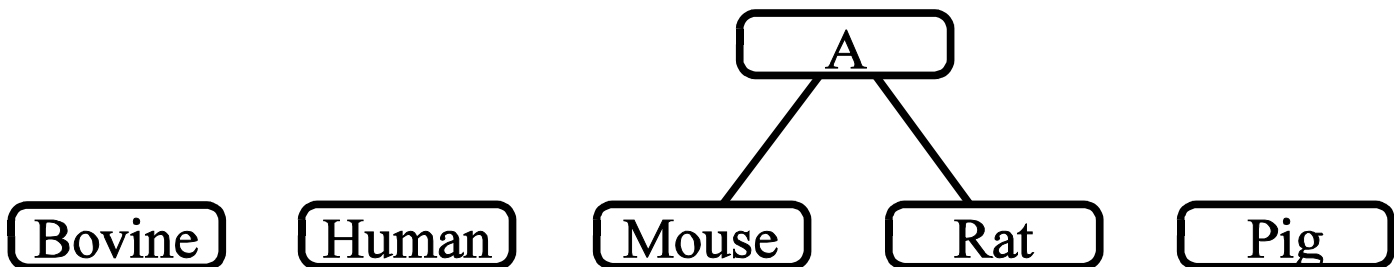
Bovine
Human
Mouse
Pig
Rat

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



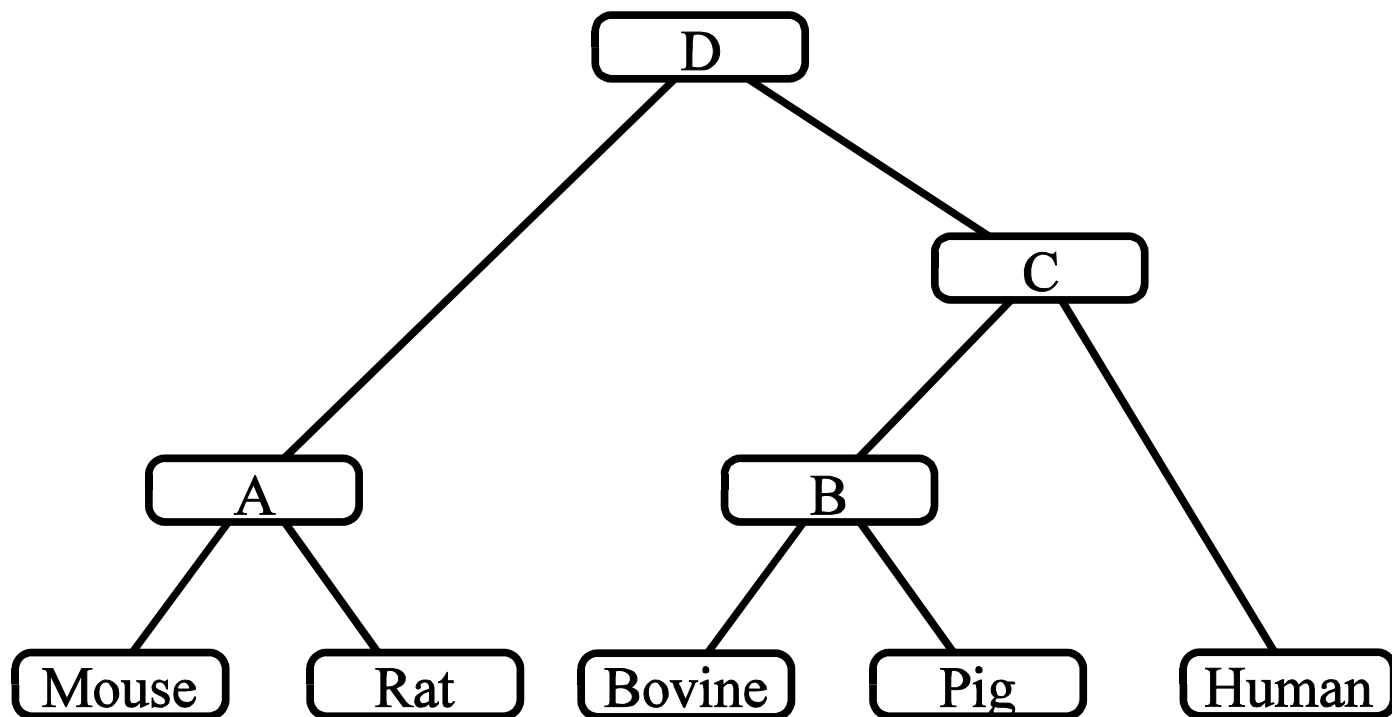
4. Replace them in the matrix with a single, new entry corresponding to that internal node. The new values are the averages of the old ones.

	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



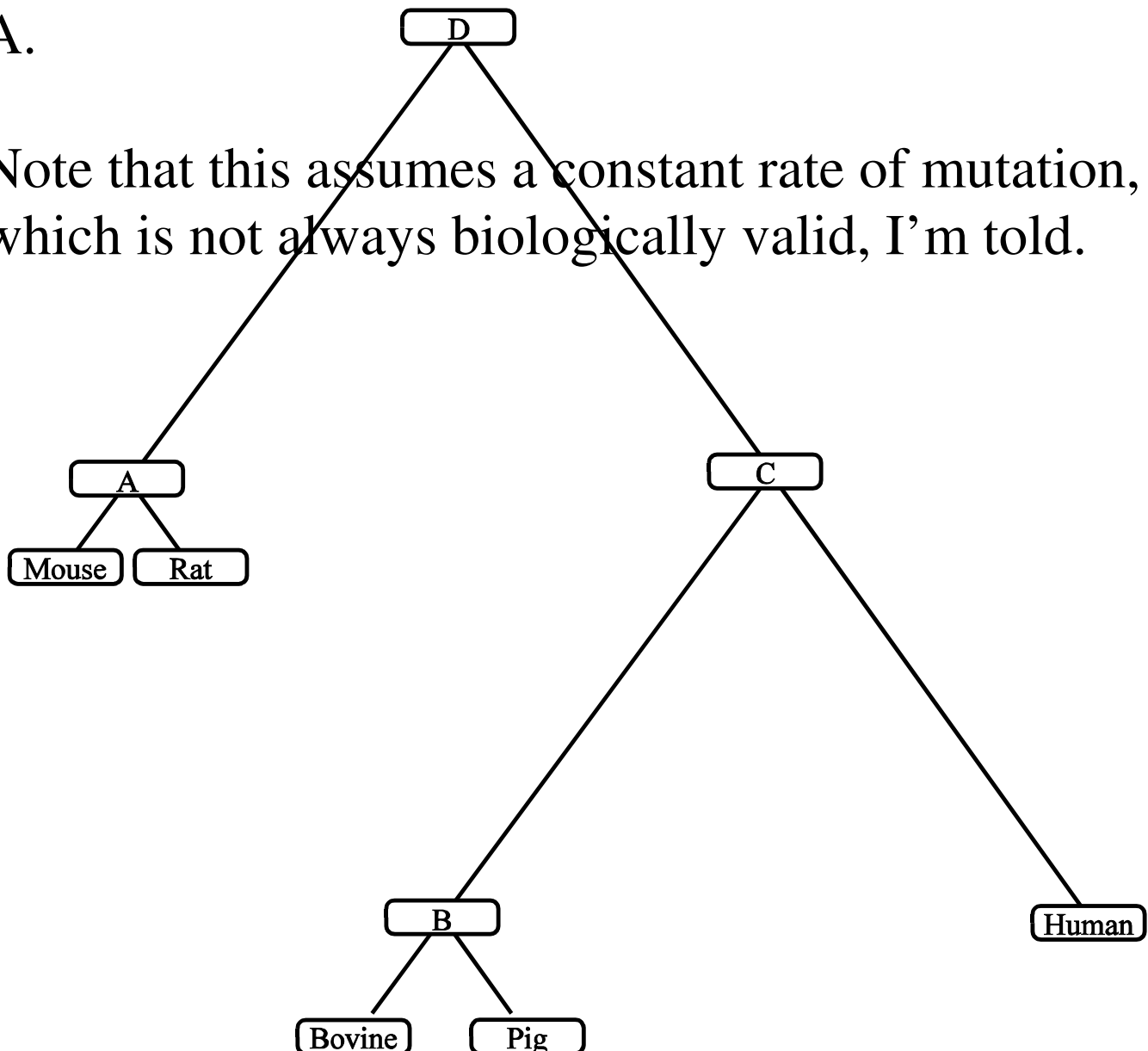
Notice that this tree agrees with the parsimony tree from a few slides ago (these are five of those eight species).

Picturing Divergence with Branch Lengths

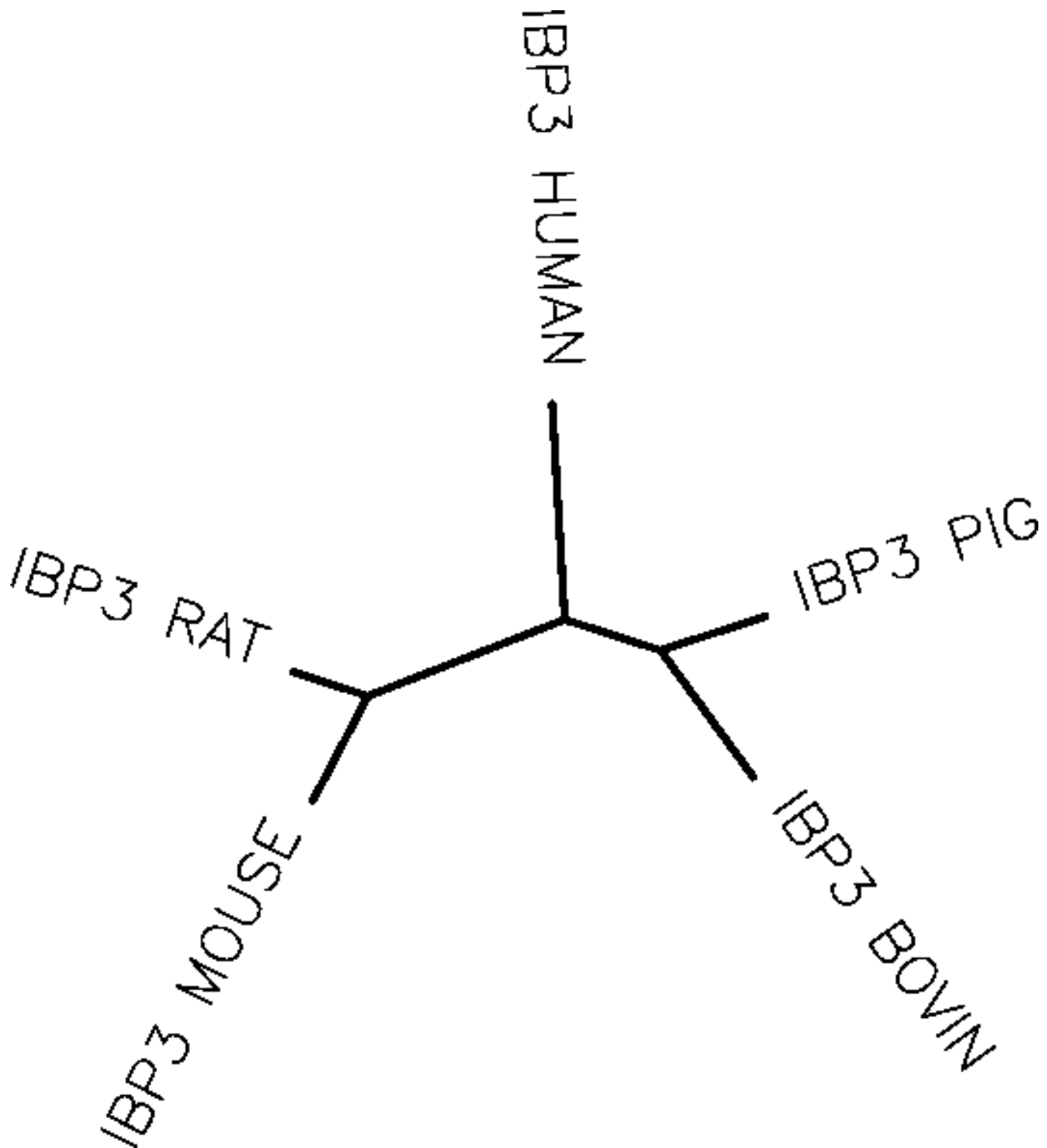
We could have drawn the edges at each step with a length proportional to their distance in the matrix.

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.



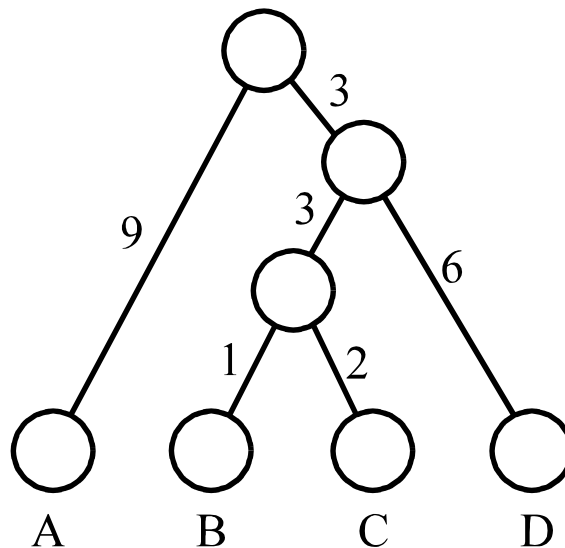
PHYLIP Tree



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				

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

	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

	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

	B	A	Human
B	0	0.193	0.168
A	0.193	0	0.19
Human	0.168	0.19	0

	A	C
A	0	0.192
C	0.18	0

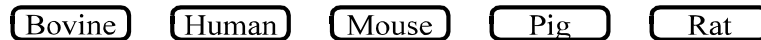
Handout #1 — Some Aligned Insulin Growth Binding Protein Sequences

IBP3_BOVIN	--MLRAPPRL	WAAALTALT	LRGPPAARAG	AGTMGAGPVV	RCEPCDARAV
IBP3_PIG	-----	-----	-----G	SGAVGTGPVV	RCEPCDARAL
IBP3_RAT	--MHPARPAL	WAAALTALT	LRGPPVARAG	AGAVGAGPVV	RCEPCDARAL
IBP3_MOUSE	--MHPARPAL	WAAALTALT	LRGPPVAELA	AGAVG-GPVV	RCEPCDARAV
IBP3_HUMAN	--MQRARPTL	WAAALTLLVL	LRGPPVARAG	ASSGGLGPVV	RCEPCDARAL
IBP2_CHICK	MALGGVGRGG	AARAAWPRL	LAALAPALAL	AGPALPEVLF	RCPPCTAERL
IBP2_BRARE	-MLSIVSCG-	-----LL	LALVT----F	HGTARSEMVF	RCP SCTAERQ
IBP4_SHEEP	-----	-----	-----	-----DEAI	HCPPCSEEKL
IBP3_BOVIN	AQCAPPSP	PCAEIVRDAG	CGCCLTCALR	EGQPCGVYTE	RCGSGLRQCQ
IBP3_PIG	AQCAPPAP	PCAEIVREPG	CGCCLTCALR	EGQACGVYTE	RCGAGLRQCQ
IBP3_RAT	AQCAPPPTAP	ACTELVREPG	CGCCLTCALR	EGDACGVYTE	RCGTGLRQCQ
IBP3_MOUSE	SQCAPPPTAP	ACTELVREPG	CGCCLTCALR	EGDACGVYTE	RCGTGLRQCQ
IBP3_HUMAN	AQCAPP--A	VCAELVREPG	CGCCLTCALS	EGQPCGIYTE	RCGSGLRQCQ
IBP2_CHICK	AACSP-AARP	PCPELVREPG	CGCCPVCARL	EDEACGVYTP	RCAAGLRCCY
IBP2_BRARE	AAC-P-MLTE	TCGEIVREPG	CGCCPVCARQ	EGEQCGVYTP	RCSSGLRCCY
IBP4_SHEEP	ARCRP-PVG-	-CEELVREPG	CGCCATCALG	KGMPCGVYTP	DCGSGLRCHP
IBP3_BOVIN	PPGDRPLQA	LLDGRGLCAN	ASAVGRLRPY	LLPS--ASGN	GSES-----E
IBP3_PIG	PPGEPRPLQA	LLDGRGLCAN	ASAAGRLRAY	LLPAPPAPGN	GSES-----E
IBP3_RAT	RPAEQYPLKA	LLNGRGFCAN	ASAASNLSAY	-LPSQSPGN	TSES-----E
IBP3_MOUSE	RPAEQYPLRA	LLNGRGFCAN	ASAAGSLSTY	-LPSQPAPGN	ISES-----E
IBP3_HUMAN	SPDEARPLQA	LLDGRGLCVN	ASAVSRLRAY	LLPAPPAPGN	ASES-----E
IBP2_CHICK	DPGAELPPQA	LVQGGTCTAR	PPDTDEYGAS	TEPPADNGDD	RSESILAENH
IBP2_BRARE	KPDSELPLEL	LVQGLGRCGR	KVDTEPTG-S	AEPREVSG--	-----
IBP4_SHEEP	PRGVEKPLHT	LVHGQGVCM	LAEIEAIQES	LQPSD-----	-----
IBP3_BOVIN	EDHSMGSTEN	QAGPSTHRVP	VSKFHPHTK	MDVIKKGHAK	DSQRYKVDYE
IBP3_PIG	EDRSVDSMEN	QALPSTHRVP	DSKLHSHVTK	MDVIKKGHAK	DSQRYKVDYE
IBP3_RAT	EDHNAGSVES	QVVPSTHRVT	DSKFHPLHSK	MEVIKQAR	DSQRYKVDYE
IBP3_MOUSE	EEHNAGSVES	QVVPSTHRVT	DSKFHPLHAK	MDVIKKGHAR	DSQRYKVDYE
IBP3_HUMAN	EDRSAGSVES	PSVSSTHRVS	DPKFHPLHSK	IIIIKKGHAK	DSQRYKVDYE
IBP2_CHICK	VDSTGGMSG	ASSRKPLKTG	MKEMPVMREK	VNEQQRQMGK	VGKAHHNHED
IBP2_BRARE	-EVQDPLDIG	LTEVPPIRKP	TKDSP-WKES	AVLQHRQQLK	SKMKYHKVED
IBP4_SHEEP	---KDEGDHP	NNSFSPCSAH	DRK---CLQK	HLAKIRDRST	SGGKMKVIGA
IBP3_BOVIN	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNMLS	PRGIHIPNCD
IBP3_PIG	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNMLS	PRGIHIPNCD
IBP3_RAT	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNVLS	PRGVHIPNCD
IBP3_MOUSE	SQSTDTQNF	SESKRETEYG	PCRREMEDTL	NHLKFLNVLS	PRGVHIPNCD
IBP3_HUMAN	SQSTDTQNF	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--IPNCD
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	PMPTSPILRG	DPNCNQYLDG
IBP4_SHEEP	RNGNFHPKQC	HPALDQGQGRK	CWCVDRKTGV	KLPGGLEPKG	ELDCHQLADS

Handout #2 — UPGMA method of phylogenetic tree construction

1. Begin with a distance matrix and an edgeless tree

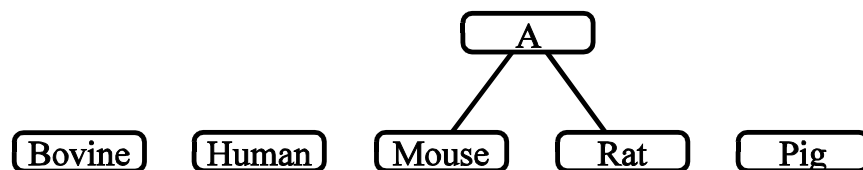
	Bovin	Pig	Rat	Mouse	Human
Bovin	0	0.098	0.197	0.204	0.173
Pig	0.098	0	0.181	0.189	0.163
Rat	0.197	0.181	0	0.072	0.176
Mouse	0.204	0.189	0.072	0	0.204
Human	0.173	0.163	0.176	0.204	0



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_{ij} 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

Exercises — Phylogeny

Presentation Problems:

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	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

2. Below is the distance matrix for four species. Use our modification of UPGMA to draw the phylogeny tree of these species with edge lengths proportional to the distances between the joined pairs.

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

3. 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 a 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

4. 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 phylogeny tree using whatever method you wish.

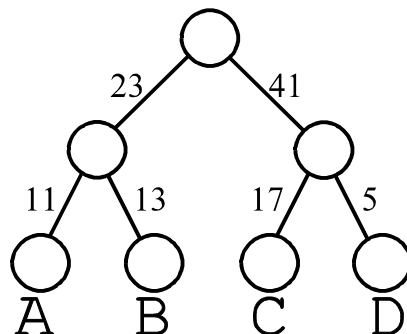
	A	B
A	0	5
B		0

Okay. If you did this first 2×2 matrix problem you already had to deal with the issue of whether or not to put a root, and where to put the "5" if there is a root. Let's not put in any roots for the three parts of this problem, but use unrooted trees instead.

	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

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



6. 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

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

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

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