Adding complexity...

Heterogeneous substitution processes, partitioned models, and mixture models



Bob Thomson

Outline for this session

- Process heterogeneity
- Partitioned-data models
- Challenges & Difficulties
- Tutorial and Activity

Heterogeneity

	1	10	20	30	40	50	60	70
Consensus	AGCCAAA	TCAA CCAGC	CT CT T G AA G TT.	AG CTGTATG	TGCTCCTGAT	YATGCTTGGA	AAAGTCTTCCT	TOTOTAA
identity								
1. hg19	C G T CAAA	CAAA CCAGC	CCCTAGATGTT.	A <mark>A</mark> CT <mark>A</mark> T <mark>C</mark> TG C	TATTCTTGAT	CATACTTGG G	A A A <mark>A</mark> T <mark>A</mark> T T <mark>A T</mark> 7	r <mark>a</mark> aa t <mark>a</mark> t C <mark>C</mark>
pytMol0	G CCATA	T <mark>TGC</mark> CC T GC	CT TC TG AA G CT	CATGGAGCT C	TGCT <mark>G</mark> CT <mark>T</mark> AT	C ATGCT <mark>G</mark> GG <mark>G</mark>	A A A G T <mark>A</mark> T T <mark>G</mark> C 1	r ca <mark>g</mark> t <mark>g</mark> t <mark>g</mark> t
3. galGal3	CCCCAGC	TCA <mark>G</mark> CC <mark>G</mark> GC	CCCTCGGAGATC	AG CTGCATG C	TGCTTCTGAT	CATGCTTGGA	AAAGT GG GCC	T <mark>GG</mark> AT <mark>T</mark> TCT
4. anoCarz	GGACACA	TCTACTCTT	CCTCTGAGGTT.	AATGGAGCT	T TG TG CT TAT	CATGCTCGGA	AAAGTCATACT	CAATCTCT.
5. alliviisu	AGCCGAG	TCAA CCCGC	CT CT TG AA G CT.	AG CTGTATG C	TGCTCCTGAT	TATGCTTGGA	AAAGT CT CCC	COATCTCT
7. taeGut1	ACCACA	SCAG CCAGC	CT CCTGAA CTC	AG CT ACATG C	TGGTCCTGAT	ATGCTTGGA	AAAGTCTTCC	PRGATETCT
8. chrPic0	AGCCAAC	TCAALCAAC	CT CT TG AAG CT.	AGCTGTATGC	TGCTCCTGAT	TATGCTTGGA	AAAGT GT CGC	CAATCTCT
						-		
-	1	10	20	30	40	50	60	70
Consensus	CTTCCA	ATGTCCGTG	CAATYGCCTAT	ACTTGGTTC	ACCTTCAGGC	TCGAAAACGC	AAGTACTTTAA	AAAACATG.
Identity								
1. hq19	CTTCCAC	ATGTCCGTG	CAAT	ACTTGGTTCA	ACCT <mark>G</mark> CAGGC	TCGAAAACGC	AAGTACTTTAA	AAAGCATG.
2. pytMol0	CTTCCAC	ATGTCCGTG	CAATGCCTA	ACTTGGTTCA	ACCTTCAGGC	TCGAAAACGC	AAGTACTTTAA	AAAACATG.
2								

3. galGal3 CTTCCGCATGTCCGTGCAAT GCCTATACTTGGTTCAACCTCCAGGCTCGAAAACGCAAGTACTTCAAAAAACATG.
 4. anoCar2 CTTCCACATGTCCGTGCAATGGCCTACACTTGGTTCAACCTTCAGGCTCGAAAACGCAAGTACTTTAAAAAAGCATG.
 5. allMis0 CTTCCACATGTCCGTGCAATGGCCTATACTTGGTTCAACCTTCAGGCTCGAAAACGCAAGTACTTTAAAAAACATG.
 6. ornAna1 CTTCCACATGTCCGTGCAATGGCCTACACTTGGTTCAACCTGCAGGCTCGAAAACGCAAGTACTTTAAAAAACATG.
 7. taeGut1 CTTCCACATGTCCGTGCAATGGCCTATACTTGGTTCAACCTGCACGCTCGAAAACGCAAGTACTTTAAAAAACATG.
 8. chrPic0 CTTCCACATGTCCGTGCAATGGCCTATACTTGGTTCAACCTTCAGGCTCGAAAACGCAAGTACTTTAAAAAACATG.

Different Genes, Different Processes

A key distinction...

• Process heterogeneity is distinct from rate heterogeneity









A key distinction...

• Process heterogeneity is distinct from rate heterogeneity

	1 1	0	20	30	40	50	60	70
Consensus	AGCCAAA TCA	ACCAGCCT	CTTG AAGTT	G CT GT ATG CT	GCTCCTGATY	ATGCTTGGAA	AGTCTTCCT	CAATCTCT
Identity								
1. hg19	CG TCAAA CAA	ACCAGCC	CT AGATGTT A	A CT AT CTG CI	ATTCTTGAT	A T <mark>A</mark> CT TGG <mark>G</mark> A <i>I</i>	A AT AT TAT T	A A A T <mark>A</mark> T C <mark>C</mark>
2. pytMol0	GGCCATATTC	CCCCCCC	TCTGAAGCT	CATGGAGCT CI	GCT <mark>G</mark> CT <mark>T</mark> AT C	ATG CT <mark>G</mark> GG <mark>G</mark> A <i>l</i>	AGT <mark>A</mark> TT <mark>G</mark> CT	CA <mark>G</mark> T <mark>G</mark> TGT
3. galGal3	CCCCAGCTCA	GCCGGCCC	CT GGAGAT CA	AG CT G CATG CI	GCTTCTGAT	A TG CT TGG AA I	AGT GG GCCT	GGATTTCT
4. anoCarz	GGACACATC	ACTCTTC	TCTGAGGTTA	ATCGACCTCI	TGTGCTTAT	ATG CT <mark>G</mark> GGAA <i>I</i>	AGTCATACT	CAATCTCT.
6 ornAnal	AGCCGAGTCA	CCACCT	CTTGAAGOTA	AG CTGTATG CI	GCTCCTGAT	ATG CTTGGAAA	AGT CT CCCT	CATCTCT
7. taeGut1	ACCCA CA TCA	GCCARCCT	CCTGAAATCA	IG CT ACATG CI	GGTCCTGAT	ATG CTTGG AA	AGTCTTCCT	TGA TTTCT
8. chrPic0	AGCCAACTCA	ACCAACCT	CTTGAAGCTA	G CT GT ATG CI	GCTCCTGAT	ATGCTTGGAA	AGT <mark>G</mark> T <mark>CG</mark> CT	CAATCTCT

Consensus	1	10	20	30	40	50	60	70
Identity	CTTCCA	CATOTCCG	T <mark>GCAAT</mark> YGCCT	ATACTTGGT	TCAACCTTCAC	GCTCGAAAAC	G <mark>CAAGTACT</mark>	
1. hg19 2. pytMol0 3. galGal3 4. anoCar2 5. allMis0	CTTCCA CTTCCA CTTCCG CTTCCA CTTCCA	CATGTCCG CATGTCCG CATGTCCG CATGTCCG CATGTCCG	TGCAAT GCCT TGCAAT GCCT TGCAAT GCCT TGCAAT GCCT TGCAAT GCCT	ATACTTGGT ACACTTGGT ATACTTGGT ACACTTGGT ATACTTGGT	FCAACCT <mark>G</mark> CAG FCAACCTTCAG FCAACCT C CAG FCAACCTTCAG	GCTCGAAAAC GCTCGAAAAC GCTCGAAAAC GCTCGAAAAC GCTCGAAAAC	GCAAGTACTT GCAAGTACTT GCAAGTACTT GCAAGTACTT GCAAGTACTT	TAAAAA <mark>G</mark> CATG. TAAAAAACATG. QAAAAAACATG. TAAAAA <mark>G</mark> CATG. TAAAAA <mark>G</mark> CATG.
6. ornAna1	CTTCCA	CATGTCCG	TGCAAT <mark>C</mark> GCCT	ACTTGGT	FCAACCT <mark>G</mark> CAG	GCTCGAAAAC(GCAAGTACTT	TAAAAAACATG.
7. taeGut1	CTTCCA	CATGTCCG	TGCAAT <mark>D</mark> GCCT	ATACTTGGT	FCAA <mark>B</mark> CTTCAG	GCTCGAAAAC(GCAAGTACTT	TAAAAAACATG.
8. chrPic0	CT <mark>A</mark> CCA	CATGTCCG	TGCAAT <mark>C</mark> GCCT	ATACTTGGT	FCAACCTTCAG	GCTCGAAAAC(GCAAGTACTT	TAAAAAACATG.

Different Genes, Different Processes

Heterogeneity

	1		1	0			20			30)			40			50			6	0			70	
Consensus Frame 1	G <mark>A</mark> G(E	G <mark>A</mark> G E	G AA E	G G G (G GG. G	T C	G G G	L L	P C C G	Y	CCC P	C C	C <mark>A</mark> G Q	F	C C	G <mark>A</mark> D	AAG K	S S	TTC F	A DI I	R R	CTG L	AG C S	Y Y	CTT. L
Identity																									
1. hg19 Frame 1	G AG(E	G AG	G A A E	G G G G G	G G C	A C <mark>G</mark> T	G G C C	CTG L	C C <mark>A</mark> P	T A C Y	CC P	TG <mark>C</mark> C	CAG Q	TTC F	TGC C	GAC D	AAG K	TCC S	TTC F	AT I	CGC R	T G L	AG C S	TAC Y	TT <mark>G</mark> i L
 pytMol0 Frame 1 	G AG(E	GAG E	G A A E	G G G (G G G J G	ACT(T	G G C (G	CTG L	C C G P	TAT Y	CCC P	TG <mark>C</mark> C	CAG Q	TTT F	TGC	GAT D	AAG K	TCC S	TTC F	ATT I	CG <mark>G</mark> R	CTG L	AG C S	TAC Y	CTT: L
 galGal3 Frame 1 	GAG(E	G AG	G A A E	G G G (G G G J G	ACT(T	G G C C	CTG L	P	TAT Y	CCC P	TGT C	CAG Q	TTT F	TG C	GAT D	AAG K	TCC S	TTC F	ATT I	CGC R	T G L	AG C S	TAC: Y	CTT: L
4. anoCar2 Frame 1	GAG(E	G AG	G A A E	G G G G G	G G G I G	T T	G G C C	L TG	P	TAT Y	CCC P	TGT C	CAG Q	TTT F	TGC	GAT D	AAG K	TCC S	TTC F	AT I I	CGC R	CTG L	AG C S	TAC: Y	CTT: L
5. allMis0 Frame 1	GAG(E	G AG	G A A E	G G G G G	G G G I G	A CT(T	G G C C	L TG	P P	TAT Y	CCC P	TGT C	CAG Q	TTT	C	GAT D	AAG K	TCC S	TTC F	AT I I	CGC R	CTG	AG C S	TAC: Y	CTT: L
6. ornAna1 Frame 1	GAG(E	GAG E	GAA	G G G (G G G J G	A C C	GGC	L TG	P	T A C Y	CCC P	TG <mark>C</mark> C	CAG Q	TTC F	TGC C	GAC D	AAG K	TC <mark>G</mark> S	TTC F	AT C I	CG <mark>G</mark> R	CTG L	AG C S	TAC: Y	CT Ci L
7. taeGut1 Frame 1	GAG(E	G A G E	G A A E	G G G (G G G I G	A CT(GGC	L TG	P	TAT Y	P	TGT C	CAG Q	TTT	TG' C	GAT D	AAG K	TCC S	TTC F	ATT I	R	CTG	AG C S	TAC: Y	CTT: L
8. chrPic0 Frame 1	G AG(E	GAG E	GAA	G G G G G	G G G I G	ACT(T	GGCO	L TG	P P	TAT Y	CCC P	C	CAG Q	TTT	TGC	GAT D	AAG K	rcc s	TTC F	ATT I	R	CTG	AG C S	TAC: Y	CTT: L

Single Gene, Different Processes

Heterogeneity



Single Gene, Different Processes

Nucleotide Substitution Models

- Recall (from yesterday) CTMC models for nucleotide substitution make several simplifying assumptions:
 - $\circ~$ the rate of the substitution process is constant across sites
 - $\circ~$ the nature of the substitution process is constant cross sites
 - sites are independant

For instance, the GTR:

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix}$$

 μ

Process Heterogeneity

Conconcus	1	10	20	30	40	50	60	70	80	90	100	110	120	130	140	152
Identity	- 1 -							- 1 - 1								
C 1. pytMol0									-							
D 2. anoCar2																
🖙 3. hg19																
🖙 4. ornAna1																
🖙 5. galGal3																
🖙 6. taeGut1																
🖙 7. allMis0																
C 8. chrPic0																

Class 1

Class 2

Process Heterogeneity



Class 1

Class 2

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$

non-partitioned model

Accommodating Heterogeneity



Class 1

Class 2

Partitioned model

Accommodating Heterogeneity



$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$

$$Q = q_{ij} = \begin{pmatrix} - & \pi_c r_{ac} & \pi_g r_{ag} & \pi_t r_{at} \\ \pi_a r_{ac} & - & \pi_g r_{cg} & \pi_t r_{ct} \\ \pi_a r_{ag} & \pi_c r_{cg} & - & \pi_t r_{gt} \\ \pi_a r_{at} & \pi_c r_{ct} & \pi_g r_{gt} & - \end{pmatrix} \mu$$

$$more highly partitioned mode$$

Accommodating Heterogeneity

$$L(\tau, \boldsymbol{\nu}, \boldsymbol{\Phi}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i | \tau, \boldsymbol{\nu}, \boldsymbol{\Phi})$$

- Typically (but not necessarily):
 - $\circ au, oldsymbol{
 u}$ shared among data partitions
 - Φ -independent for each data partition $r = (r_{ac}, r_{ag}, r_{at}, r_{cg}, r_{ct}, r_{gt})$ $\pi = (\pi_a, \pi_c, \pi_g, \pi_t)$ α

How does this influence inference?





Brown & Lemmon 2007



Brown & Lemmon 2007

Underparameterization is often worse than overparameterization

Generating model is GTR+G



Underparameterization is often worse than Overparameterization

Generating model is JC



Huelsenbeck & Rannala 2004

From earlier today...



Number of Parameters



Brown & Lemmon 2007

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors
- this effect is worse in large (more heterogenous) datasets

Data Set	Data Type ^a	Таха	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.	
Anderson_2013	м	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)	
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)	34 [63] 03135615
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)	
Brown_2012	Ν	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)	
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)	
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)	
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)	
Devitt_2013	м	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)	
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)	
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA	
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA	
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)	
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)	
Fong_2012	Ν	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)	
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)	
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)	
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)	
Kang_2013a	Ν	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	
Kang_2013b	м	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)	
Lartillot_2012	Ν	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)	
Leavitt_2013	м	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA	
Li_2008	Ν	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA	
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)	
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)	
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)	
Seago_2011	м	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)	
Sharanowski_2011	Ν	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)	
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)	
Tolley_2013	м	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)	
Unmack_2013	м	139	6,827	25	Melanotaeniidae	Rainbowfish	Unmack et al. (2013b)	Unmack et al. (2013a)	
Wainwright_2012	N	188	8,439	30	Acanthomorpha	Fishes	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	Wainwright et al. (2012)	
Ward_2010	Ν	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA	
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)	Kainer and Lanfear 2015

Data Set	Data Type ^a	Таха	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.	
Anderson_2013	м	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)	
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)	
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)	
Brown_2012	Ν	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)	
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)	
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)	
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)	4 partitioning schemes
Devitt_2013	м	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)	0
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)	
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA	1 linnartitioned
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA	
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)	2 - Drigri by fosturo
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)	
Fong_2012	N	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)	
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)	2 chocon using All
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)	3. CHOSEH USHIR AIC
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)	9
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	1 chacan using DIC
Kang_2013b	м	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)	
Lartillot_2012	Ν	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)	
Leavitt_2013	м	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA	
Li_2008	N	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA	
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)	
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)	
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)	
Seago_2011	м	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)	
Sharanowski_2011	N	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)	
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)	
Tolley_2013	м	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)	
Unmack_2013	м	139	6,827	25	Melanotaeniidae	Rainbowfish	Unmack et al. (2013b)	Unmack et al. (2013a)	
Wainwright_2012	N	188	8,439	30	Acanthomorpha	Fishes	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	Wainwright et al. (2012)	
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA	
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)	Kainer and Lanfear 2015

Data Set	Data Type ^a	Таха	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.	
Anderson_2013	м	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)	
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)	<u>34 (PAI NATASPTS</u>
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)	
Brown_2012	Ν	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)	
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)	
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)	
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)	4 partitioning schemes
Devitt_2013	м	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)	Simple
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)	Simple
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA	1 unnartitioned
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA	
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)	2 a priori by foaturo
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)	
Fong_2012	Ν	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)	
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)	2 chocon using AlC
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)	5. LINSEIT USITIS AIL
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)	J 0
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	1 chocon using DIC
Kang_2013b	м	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)	
Lartillot_2012	Ν	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)	inore complex
Leavitt_2013	м	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA	
Li_2008	N	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA	
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)	
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)	
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)	
Seago_2011	м	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)	
Sharanowski_2011	N	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)	
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)	
Tolley_2013	м	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)	
Unmack_2013	м	139	6,827	25	Melanotaeniidae	Rainbowfish	Unmack et al. (2013b)	Unmack et al. (2013a)	
Wainwright_2012	N	188	8,439	30	Acanthomorpha	Fishes	Wainwright, Smith, Price, Tang, Sparks, Ferry, Kuhn, Eytan, et al. (2012)	Wainwright et al. (2012)	
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA	
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)	Kainer and Lanfear 2015

Data Set	Data Type ^a	Taxa	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.	
Anderson_2013	м	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)	
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)	
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)	Jirediddiddets
Brown_2012	N	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)	
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)	
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)	
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)	4 partitioning schemes
Devitt_2013	м	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)	0
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)	a (')' (
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA	1 linnartitioned
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA	
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)	2 a priori by foaturo
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)	
Fong_2012	N	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)	
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)	2 chocon using MC
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)	3. CHUSEH USHIY AIL
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)	2 0
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	1 chocon using PIC
Kang_2013b	м	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)	0
Lartillot_2012	N	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)	
Leavitt_2013	м	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA	A 1
Li_2008	N	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA	Δηρίντο
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)	Andryze
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)	•
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)	
Seago_2011	M	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)	
Sharanowski_2011	N	139	3,982	11	Braconidae	Wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)	
Siler_2013	M,N	61	2,697	7	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)	
Tolley_2013	M	203	5,054	16	Chamaeleonidae	Chameleons	Tolley et al. (2013b)	Tolley et al. (2013a)	
Unmack_2013	M N	139	6,827	25	A conthe meaning	Kaindownsn Eisbos	Unmack et al. (2013b)	Unmack et al. (2013a)	
wainwright_2012	N	188	8,439	30	Acantnomorpna	FISNES	Sparks, Ferry, Kuhn, Eytan, et al. (2012)	wainwright et al. (2012)	
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA	
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)	Kainer and Lanfear 2015

Data Set	Data Type ^a	Таха	Sites	Data Blocks	Clade (Latin)	Clade (English)	Study Ref.	Data Set Ref.	
Anderson_2013	м	145	3,037	4	Loliginidae	Pencil squids	Anderson et al. (2014)	Anderson et al. (2013)	
Bergsten_2013	M,N	38	2,111	8	Dytiscidae	Diving beetles	Bergsten et al. (2013a)	Bergsten et al. (2013b)	
Broughton_2013	M,N	61	19,997	61	Osteichthyes	Bony fishes	Broughton et al. (2013b)	Broughton et al. (2013a)	
Brown_2012	Ν	41	1,665	7	Ptychozoon	Asian geckos	Brown et al. (2012b)	Brown et al. (2012a)	
Caterino_2001	M,N	37	3,228	9	Papilionidae	Butterflies	Caterino et al. (2001)	Kuo et al. (2001)	
Cognato_2001	M,N	44	1,896	7	Scolytinae	Bark beetles	Cognato and Vogler (2001b)	Cognato and Vogler (2001a)	
Day_2013	M,N	152	3,586	11	Synodontis	African catfish	Day, Peart, Brown, Friel, et al. (2013)	Day, Peart, Brown, Bills, et al. (2013)	4 partitioning schemes
Devitt_2013	м	69	823	4	Ensatina	Salamander	Devitt et al. (2013b)	Devitt et al. (2013a)	0
Dornburg_2012	M,N	44	5,919	21	Holocentridae	Squirrel fishes	Dornburg et al. (2012b)	Dornburg et al. (2012a)	a (')' (
Dsouli_2011	M,N	39	1,635	7	Muscidae	Flies	Dsouli et al. (2011)	NA	hannartitioned
Ekrem_2010	M,N	74	2,701	10	Chironomidae	Midges	Ekrem et al. (2010)	NA	
Elias_2009	M,N	143	4,159	12	Nymphalidae	Butterflies	Elias, Joron, Willmott, Silva-Brandão, et al. (2009)	Elias, Joron, Willmott, Kaiser, et al. (2009)	2 a priori by foaturo
Fishbein_2001	N,C	40	9,005	11	Saxifragales	Core Eudicots	Fishbein et al. (2001b)	Fishbein et al. (2001a)	
Fong_2012	Ν	110	25,919	168	Vertebrata	Vertebrates	Fong et al. (2012b)	Fong et al. (2012a)	
Grande_2013	M,N	65	4,027	12	Paracanthopterygii	Fish	Grande et al. (2013a)	Grande et al. (2013b)	2 chocon using MC
Guschanski_2013	M,C	110	17,092	63	Cercopithecini	Monkeys	Guschanski et al. (2013b)	Guschanski et al. (2013a)	J. LIUSEII USIIIS AIL
Kaffenberger_2011	M,N	54	6,548	26	Gephyromantis	Malagasy frogs	Kaffenberger et al. (2012)	Kaffenberger et al. (2011)	
Kang_2013a	N	28	7,276	15	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	1 chocon using Bl
Kang_2013b	M	28	1,239	6	Xiphophorus	Swordtail fish	Kang et al. (2013)	NA	
Kawahara_2013	M,N	70	2,238	9	Hyposmocoma	Caterpillar	Kawahara and Rubinoff (2013a)	Kawahara and Rubinoff (2013b)	0
Lartillot_2012	N	78	15,117	51	Eutheria	Mammals	Lartillot and Delsuc (2012b)	Lartillot and Delsuc (2012a)	
Leavitt_2013	м	34	15,404	87	Acridoidea	Grasshoppers	Leavitt et al. (2013)	NA	A 1
Li_2008	Ν	56	7,995	30	Actinopterygii	Fishes	Li et al. (2008)	NA	
Murray_2013	M,N	237	3,111	9	Eucharitidae	Wasps	Murray et al. (2013a)	Murray et al. (2013b)	Andryze
Rightmyer_2013	M,N	94	3,692	25	Hymenoptera	Bee	Rightmyer et al. (2013b)	Rightmyer et al. (2013a)	•
Sauquet_2011	N,C	51	5,444	10	Nothofagus	Beeches	Sauquet et al. (2012)	Sauquet et al. (2011)	
Seago_2011	M	116	2,253	7	Coccinellidae	Ladybirds	Seago et al. (2011b)	Seago et al. (2011a)	
Sharanowski_2011	N	139	3,982	11	Braconidae	wasps	Sharanowski et al. (2011b)	Sharanowski et al. (2011a)	Ach what changes
Siler_2013	/VI, N	61	2,697	1	Lycodon	Wolf snakes	Siler, Oliveros, et al. (2013)	Siler, Brown, et al. (2013)	A2K WIIGT CIIGIISG2
Tolley_2013	1/1	203	5,054	16	Chamaeleonidae	Deinhaufeh	Tolley et al. (2013b)	Tolley et al. (2013a)	0
Wainwright 2012	/VI	109	0,827	25	Acanthamamha	Fishes	Wainwright Smith Brice Tang	Weinwright et al. (2013a)	
wannwright_2012	N	100	0,439	50	Acanchomorpha	risiles	Sparks, Ferry, Kuhn, Eytan, et al. (2012)	wannwngnt et al. (2012)	
Ward_2010	N	54	9,173	27	Dolichoderinae	Ants	Ward et al. (2010)	NA	
Welton_2013	M,N	145	4,552	16	Varanus	Lizards	Welton et al. (2013b)	Welton et al. (2013a)	Kainer and Lanfear 2015



Topology changes

Kainer and Lanfear 2015



Kainer and Lanfear 2015



Nodes subtended by short branch lengths and low support disproportionately effected

Kainer and Lanfear 2015

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors
- this effect is worse in large (more heterogenous) datasets

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors, and branch lengths, and topologies
- $\circ\,$ this effect is worse in large (more heterogenous) datasets

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors, and branch lengths, and topologies
- this effect is worse in large (more heterogenous) datasets
- $\circ\,$ this effect is worse for difficult to resolve nodes

- abundant empirical evidence that substitution process varies
- models that ignore heterogeneity can give inaccurate estimates of node posteriors, and branch lengths, and topologies
- this effect is worse in large (more heterogenous) datasets
- $\circ~$ this effect is worse for difficult to resolve nodes

Precise conditions of 'Big Data' studies

How do we select a partitioning model?

- This is a similar problem to what we covered this morning
 - Seek to capture the relevant variation in the data, balancing the bias-variance tradeoff



• We can use similar tools as before

How do we select a partitioning model?



Höhna et al. Partition Tutorial

Consensus	1	10	20	30	40	50	60	70	80	90	100	110	120	130	140	152
Identity					b-ladel			d. all								
D= 1. pytMol0																
2. anoCar2																
🖙 3. hg19																
🖙 4. ornAna1																
🖙 5. galGal3																
🖙 6. taeGut1																
🖙 7. allMis0																
🖙 8. chrPic0																

- Many possible partition models even for 'small data'.
- Simple partitioning by biological features: 2 genes x 3 codon
 positions: # of Classes # of possible models

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	203

- Many possible partition models even for 'small data'.
- Simple partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	203

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models	Two issues∙	1 This counts only
1	1		
2	31		possibilities where
3	90		the partition model is
4	65		the same for all
5	15		
6	1		parameters
total	203		

÷

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

Phylogenetic Model Parameters

# of Classes	# of possible models	
1	1	Topology $ au$
2	31	Branch lengths $ u$
3	90	Exchangeability rates $\boldsymbol{r} = (r_{rr}, r_{rr}, r_{rr}, r_{rr}, r_{rr}, r_{rr}, r_{rr}, r_{rr})$
4	65	= (ac, ag, at, cg, ct, gt)
5	15	$\pi = (\pi_a, \pi_c, \pi_g, \pi_t)$
6	1	ASRV Gamma Shape α
total	203	

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

Phylogenetic Model Parameters

# of Classes	# of possible models		
1	1	Topology $ au$	
2	31	Branch lengths $ u$	Shared across partitions
3	90	Exchangeability rates r	$= (r_{a}, r_{a}, r_{a}, r_{a}, r_{a}, r_{a}, r_{a})$
4	65	Base frequencies	(ac, ag, al, cg, cl, gl)
5	15		$= (\pi_a, \pi_c, \pi_g, \pi_t)$
6	1	ASKV Gamma Snape $ lpha $	
total	203		

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

Phylogenetic Model Parameters

# of Classes	# of possible models		
1	1	Topology \mathcal{T}	
2	31	Branch lengths ν Shared across	partitions
3	90	Exchangeability rates $m{r}=(r_{ab},r_{ab},r_{ab})$	rat rag rat rat)
4	65	Proce frequencies $ (-ac, -ag, -)$	al, cg, cl, gl
5	15	base nequencies $\pi = (\pi_a, \pi_c, \pi_c)$	(π_t,π_t)
6	1	ASRV Gamma Shape $lpha$	
total	203	Independent for each	partition

1

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

Phylogenetic Model P	arameters
----------------------	-----------

# of Classes	# of possible models	
1	1	Topology \mathcal{T}
2	31	Branch lengths $ u$ Shared across partitions
3	90	Exchangeability rates $m{r}=(r_{1},r_{2},r_{3}$
4	65	Base frequencies $\mathbf{\sigma} = (\pi ac, \pi ag, \pi at, \pi cg, \pi ct, \pi gt)$
5	15	$\pi = (\pi_a, \pi_c, \pi_q, \pi_t)$
6	1	ASRV Gamma Shape $lpha$
total	203	Should we consider other partial partition models?

If so, **many** more models are possible.

÷

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

Phylogenetic Model Parameters

# of Classes	# of possible models	
1	1	Topology \mathcal{T}
2	31	Branch lengths $ u$ Shared across partitions
3	90	Exchangeability rates $\boldsymbol{r}=(r, r, r$
4	65	= ('ac, 'ag, 'at, 'cg, 'ct, 'gt)
5	15	base frequencies $\boldsymbol{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$
6	1	ASRV Gamma Shape $lpha$
total	203	Should we consider other partial partition models?

If so, **many** more models are possible.

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models	Τωρ ίς ςμρς.	2 This counts only
1	1		Z. THIS COULTS ONLY
2	31		partition models for
3	90		classes that we
4	65		thought to write
5	15		thought to write
6	1		down a priori.
total	203		•

- Many possible partition models even for 'small data'.
- **Simple** partitioning by biological features: 2 genes x 3 codon positions:

# of Classes	# of possible models
1	1
2	31
3	90
4	65
5	15
6	1
total	203

Like many problems in phylogenetics, this doesn't scale easily

• **Model selection:** select a single 'best fitting' model by comparing many possible alternatives

- **Model selection:** select a single 'best fitting' model by comparing many possible alternatives
 - estimate marginal likelihood and use Bayes Factors to do model selection
 - **AIC** or **BIC** model selection (e.g., Lanfear et al. 2012, Lanfear et al. 2017)

- **Model selection:** select a single 'best fitting' model by comparing many possible alternatives
 - estimate marginal likelihood and use Bayes Factors to do model selection
 - **AIC** or **BIC** model selection (e.g., Lanfear et al. 2012, Lanfear et al. 2017)

General approach: Calculate likelihood of data under a particular model, penalize for model complexity via the prior (marginal likelihood) or some penalty term (AIC or BIC). Compare.

Model Selection and Model Averaging

- **Model Selection:** Choose the model that best fits the data.
 - The data are random variables, the parameters are (or are not) random variables, but **the model is fixed.**

Model Selection and Model Averaging

- **Model Selection:** Choose the model that best fits the data.
 - The data are random variables, the parameters are (or are not) random variables, but **the model is fixed.**
- **Model Averaging:** Consider alternative models in proportion to their probability
 - The **model itself is a random variable** with associated uncertainty, so we account for this.

- Model Averaging: Treat the partition-model itself as a random variable and use MCMC to marginalize over possible partition models (and other parameters of the joint phylogenetic model)
 - Moore et al. 2014 AutoParts
 - Wu et al. 2013 substBMA (addon for BEAST 2)
 - RevBayes

- Model Averaging: Treat the partition-model itself as a random variable and use MCMC to marginalize over possible partition models (and other parameters of the joint phylogenetic model)
 - Moore et al. 2014 AutoParts
 - Wu et al. 2013 substBMA (addon for BEAST 2)
 - RevBayes

General approach: Model the number of partitions and the assignment of sites or classes to those partitions using the Dirichlet Process

Dirichlet Process

• A stochastic process that allows us to describe the prior probability of a mixture model for the **number of partitions** and the **assignment of data to those partitions**

Dirichlet Process

- A stochastic process that allows us to describe the prior probability of a mixture model for the **number of partitions** and the **assignment of data to those partitions**
- More simply: A prior probability distribution for clustering problems
 - How many clusters are there?
 - Which observations belong to which clusters?

Imagine a restaurant with an infinite number of tables.
 Customers walk in one at a time and choose a table to sit down at:



- Imagine a restaurant with an infinite number of tables.
 Customers walk in one at a time and choose a table to sit down at:
 - The first customer always chooses the first table.



- Imagine a restaurant with an infinite number of tables.
 Customers walk in one at a time and choose a table to sit down at:
 - The first customer always chooses the first table.
 - The nth customer chooses the first unoccupied table with probability $c \frac{\alpha}{n-1+\alpha}$, and an occupied table with probability $c \frac{\alpha}{n-1+\alpha}$, where c is the number of people already at that table.



- Imagine a restaurant with an infinite number of tables.
 Customers walk in one at a time and choose a table to sit down at:
 - The first customer always chooses the first table.
 - The nth customer chooses the first unoccupied table with probability $c \frac{\alpha}{n-1+\alpha}$, and an occupied table with probability $c \frac{\alpha}{n-1+\alpha}$, where c is the number of people already at that table.

Large alpha - many tables in use Popular tables stay popular





Sorts customers into tables

http://topicmodels.west.uni-koblenz.de/ckling/tmt/crp.html



Sorts customers into tables Sorts observations into clusters



Sorts customers into tables Sorts observations into clusters Sorts alignment classes into partitions

Dirichlet Process Prior



Consensus Identity	1	10	20	30 4	40 50	60	70	80	90 100	110	120 1:	30 140	152
C 1. pytMol C 2. anoCa C 3. hg19 C 4. ornAna C 5. galGal C 5. galGal C 6. taeGut C 7. allMisC C 8. chrPicC	10 r2 a1 3 t1 0 0							•		•		•	
Consensus	1	50,000	100,000	150,000	200,000	250,000	300,000	350,000) 400,000	450,000	500,000	550,000	613,454
ldentity	ilina na h	la (a la la para	n literation of the second	hard for the state of the state	yldullaardotteed	n-Usupana	and the second second	police (second	legel (malered	urred and for	an (sur front and front	and A gradient	al and a lot
1. hg19 2. galGal3 3. anoCar2 4. allMis0 5. ornAna1 6. taeGut1													
7. chrPic0	This can be extremely challenging for 'big data'.												

Consensus Identity	1	10	20	30 40	50	60	70	80	90 100	110	120 1	30 140	152
C 1. pytMo C 2. anoCa C 3. hg19 C 4. ornAn C 5. galGal C 6. taeGu C 7. allMis(C 8. chrPic	a1 3 13 0 0							•					
Consensus	1	50,000	100,000	150,000	200,000	250,000	300,000	350,000	400,000	450,000	500,000	550,000	613,454
ldentity	Norman Norman	tadik (ki ki k	and all and an and an	watty fallen wat	ulikatering-soluk	allon anar	h _{rlin} mailehi	policialities	ladas findrias ai	urray filling a	indentenne indent	and January ne	J.m.v.la,
1. hg19 2. galGal3 3. anoCar2 4. allMis0													
5. ornAna1 6. taeGut1 7. chrPic0						<u> </u>			• (1	_			
			This	can be	extre	mely a	challe	nging	for big	g data			

e.g. 100 classes:

47585391276764833658790768841387207826363669686825611466616334637559114497892442622672724044217756306953557882560751 partition models

יסי

0

Consensus	1	50,000	100,000	150,000	200,000	250,000	300,000	350,000	400,000	450,000	500,000	550,000	613,454
	We was	haddaladaraa	(NATURAL)	urth hill and h	develop-with-	llaur na argun	Un man hall be had been a start of the	luce ware produced	nyunturya	undy the part	lin have a hard	and have a president	ⁿ lowardty
Identity													
1. hg19 2. galGal3 3. anoCar2													
4. allMis0 5. ornAna1													
7. chrPic0													

Scaling up may also mean that we need to relax this assumption

Phylogenetic Model Parameters

Topology Branch lengths	$ au \ oldsymbol{ u}$	Shared across partitions				
Exchangeability rates $oldsymbol{r}=(r_{ac},r_{ag},r_{at},r_{cg},r_{ct},r_{gt})$						
Base frequencies	π	$= (\pi_a, \pi_c, \pi_g, \pi_t)$				
ASRV Gamma Shape	lpha	` U /				

Independent for each partition



Tutorial

- Partition models in RevBayes
 - **set up** partitioned model
 - run **mcmc**
 - stepping stone integration for marginal likelihood estimation

Exercise

- I have supplied data for four 'loci' (sim_locus1.nex,...sim_locus4.nex)
- Choose (and sign up for) one partition model of the form (12,34 or 1,2,34 or 1,2,3,4)
- GTR + Gamma model for all partitions
- Settings for stepping stone integration:
 - ∘ 50 power posteriors (cats=50)
 - burnin=500, tuning interval=200
 - 2000 generations per power posterior
 - $\circ\,$ add your estimate to the spreadsheet

