Modelling Protein Evolution

Olivier Gascuel
LIRMM-CNRS, Montpellier, France
www.lirmm.fr/~gascuel

Joint work with Quang Le Si

Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

Modelling Protein Evolution (Substitutions)

- **Evolutionary forces**
  - Genetic code
  - Physics and chemistry
  - Structure and function
- **The standard model**
- **More complex models**
- **Simpler models**
- **Discussion**

The data

Man	M	A	E	I	G	R	L	I	E	F	S	A	M	V	D	F	W	Q	N	R	C
Frog	M	A	E	I	G	R	L	V	E	Y	S	A	M	V	D	F	W	Q	N	R	C
Zebrafish	M	A	D	L	G	K	L	I	D	Y	S	A	L	V	D	F	W	Q	N	R	C
Fly	M	S	D	I	G	K	L	V	E	F	S	P	M	V	E	F	W	Q	Q	K	C
Yeast	M	S	E	I	G	R	L	V	E	F	-	-	-	-	-	F	W	Q	N	R	C
Amoeba	L	S	E																		
Paramecium	L	A	E																		
Blue algae	L	S	D																		

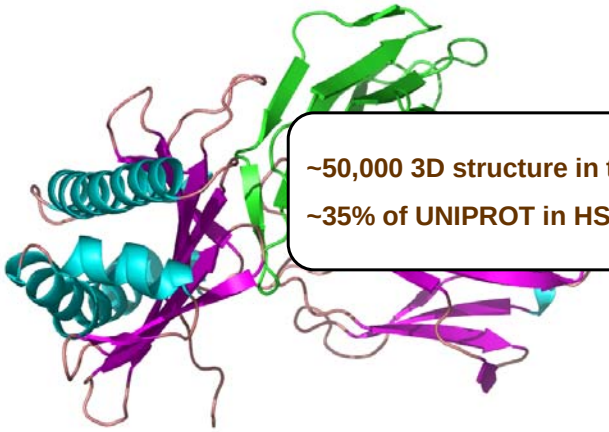
Inferring the phylogeny

Understanding the evolutionary process

Reconstructing ancestral sequences

Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

The data



~50,000 3D structure in the PDB

~35% of UNIPROT in HSSP

The data

```
AAACTTAAAGTTGTAGGACAGGACTCCAATGAGATCCACTTCCGAGTGAAGATGACC
K  L  K  V  V  G  Q  D  S  N  E  I  H  F  R  V  K  M  T

ACACAGATGGGCAAGTTAAAGAAGTCATACAGTGAGCGGGTGGGAGTCCCTGTAGCA
T  Q  M  G  K  L  K  K  S  Y  S  E  R  V  G  V  P  V  A

TCACTGCGTTTCCTCTTCGATGGACGACGCATTAACGACGAAGAAACGCCCAAAGCT
S  L  R  F  L  F  D  G  R  R  I  N  D  E  E  T  P  K  A

CTGGAAATGGAGAATGATGATGTAATTGAAGTGTACCAGGAGCAGACCGGCGGCCAT
L  E  M  E  N  D  D  V  I  E  V  Y  Q  E  Q  T  G  G  H
```


Dogma

- **independent and random mutations at the DNA level**
- **substitution rates involving several nucleotide changes should be zero**

First Position	Codon: C	Codon: A	Codon: G	
		Tyr [Y]	TGT Cys [C]	T
		Tyr [Y]	TGC Cys [C]	C
		Ter [end]	TGA *Ter* [end]	A
		Ter [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

First Position	Second Position of Codon: T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

First Position	Second Position of Codon: T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

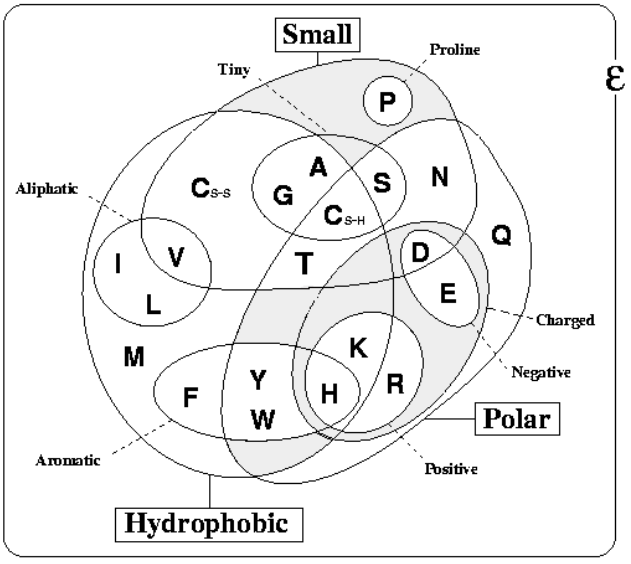
First Position	Second Position of Codon: T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

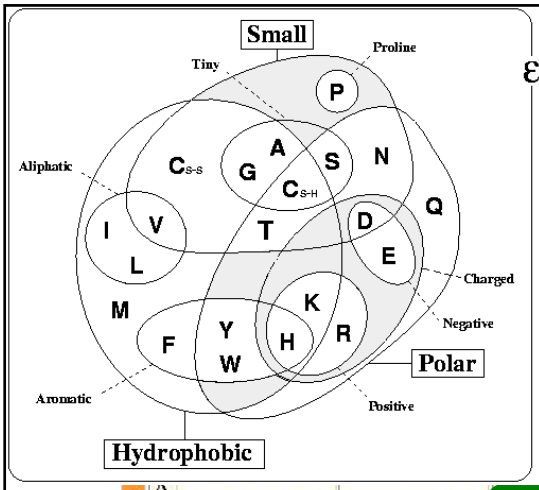
Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

First Position	Second Position of Codon: T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

First Position	Second Position of Codon: T	C	A	G	Third Position
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

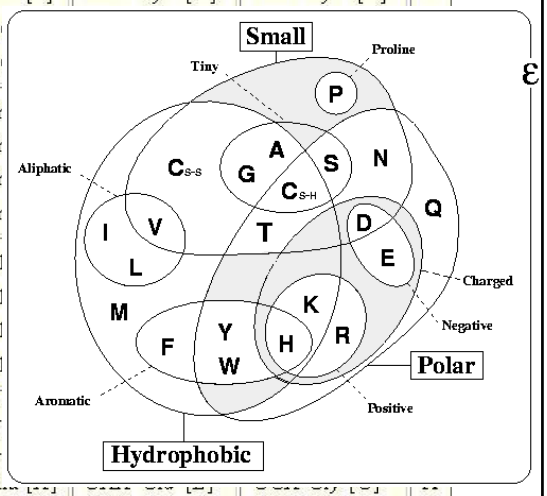
Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011



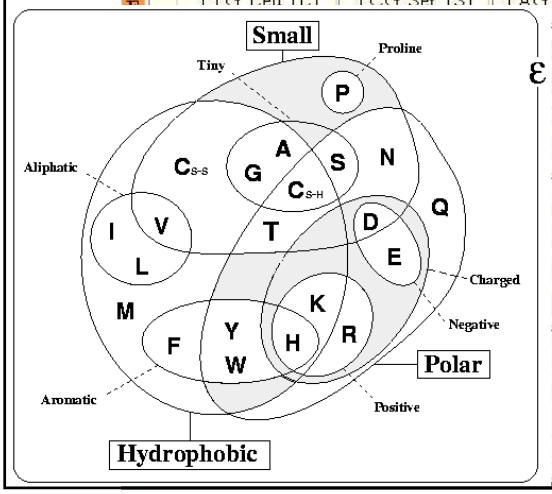


			f Codon		
		A		G	
		Tyr [Y]	TGT	Cys [C]	T
		Tyr [Y]	TGC	Cys [C]	C
		Ter [end]	TGA	*Ter* [end]	A
		Ter [end]	TGG	Trp [W]	G
		His [H]	CGT	Arg [R]	T
		His [H]	CGC	Arg [R]	C
		Gln [Q]	CGA	Arg [R]	A
		Gln [Q]	CGG	Arg [R]	G
		Asn [N]	AGT	Ser [S]	T
		Asn [N]	AGC	Ser [S]	C
t i o n	A	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
		ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
	G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
		GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
		GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

		Second Position of Codon				
		T	C	A	G	
F i r s t P o s i t i o n	T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
		TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
		TTA Leu [L]	TCA S			
		TTG Leu [L]	TCG S			
	C	CTT Leu [L]	CCT Pr			
		CTC Leu [L]	CCC Pr			
		CTA Leu [L]	CCA Pr			
		CTG Leu [L]	CCG Pr			
	A	ATT Ile [I]	ACT T			
		ATC Ile [I]	ACC T			
		ATA Ile [I]	ACA T			
		ATG Met [M]	ACG T			
	G	GTT Val [V]	GCT A			
		GTC Val [V]	GCC A			
		GTA Val [V]	GCA A			
		GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

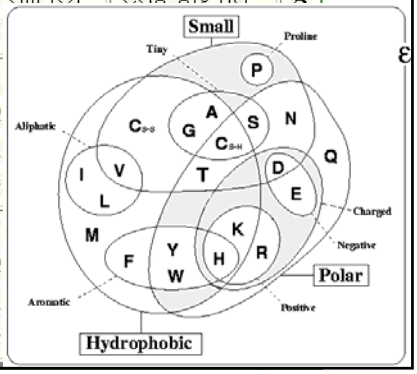


First Position	Second Position of Codon: T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

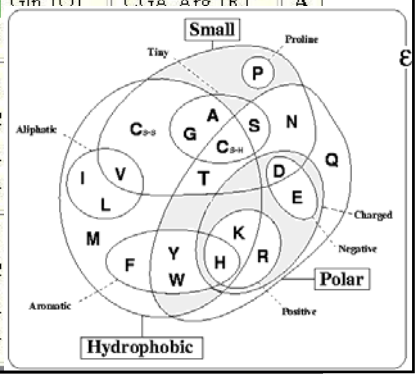


Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

First Position	Second Position of Codon: T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

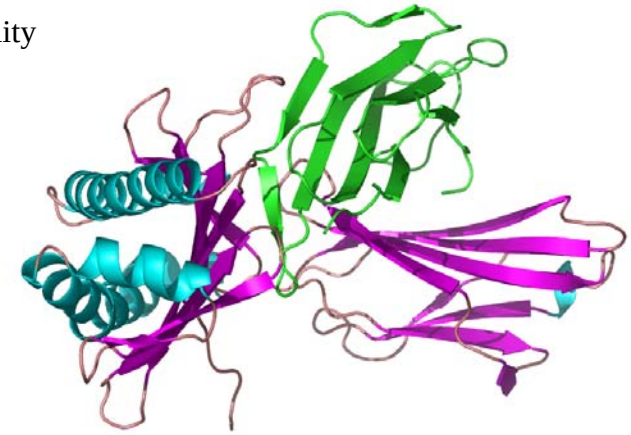


First Position	Second Position of Codon T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	? His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG		
A	ATT Ile [I]	ACT Thr [T]	AAT		
	ATC Ile [I]	ACC Thr [T]	AAC		
	ATA Ile [I]	ACA Thr [T]	AAA		
	ATG Met [M]	ACG Thr [T]	AAG		
G	GTT Val [V]	GCT Ala [A]	GAT		
	GTC Val [V]	GCC Ala [A]	GAC		
	GTA Val [V]	GCA Ala [A]	GAA		
	GTG Val [V]	GCG Ala [A]	GAG		



The structure (and function)

- Highly conserved
- Secondary structure: helix, beta sheet or extended and coil
- Solvent accessibility



The structure (and function)

- Highly conserved
- Secondary structure: helix, beta sheet or extended and coil
- Solvent accessibility
- Obtained by homology
- ~35% of proteins

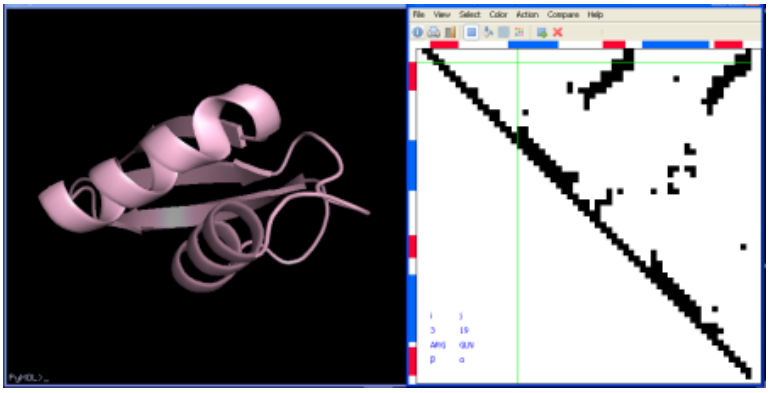
```
QACQIQMSDPAYNINISLPSYYPDQKSLENYIAQTR
QACLIQMSDPAYNTNISLPSYYPDQKSLENYIAQTR
EVCKIQMTEPGYTVDISLPSNYPDGKSLESFVRETG
QSCQVHAAAAEYTVEFSFPAGYPDEQAVSAYLTQTR
QICHVHASGPKYMLDMTFPVDYPDQQALTDYITQNR
QMCHIHATGSTYTLDITFPQDYPDQQALTDYITLNR
QTCHVHAENSTYRLDYTFPVDYPDQQALAAYLTQTR
QICQVHTANATYRLDYTFSTTYPDQEAVVGYLSQTR
DMCLVQVENPTYRLDYSFPADYPDQQALTAYLTQTR
QMCRLRAAGPNYTINMVFPADYPDEQALTDYITQNR

S.EEEEEEETTEEEEEE..S..TTTHHHHHHHHHHH
81010555483220201012703203300530260 1
```

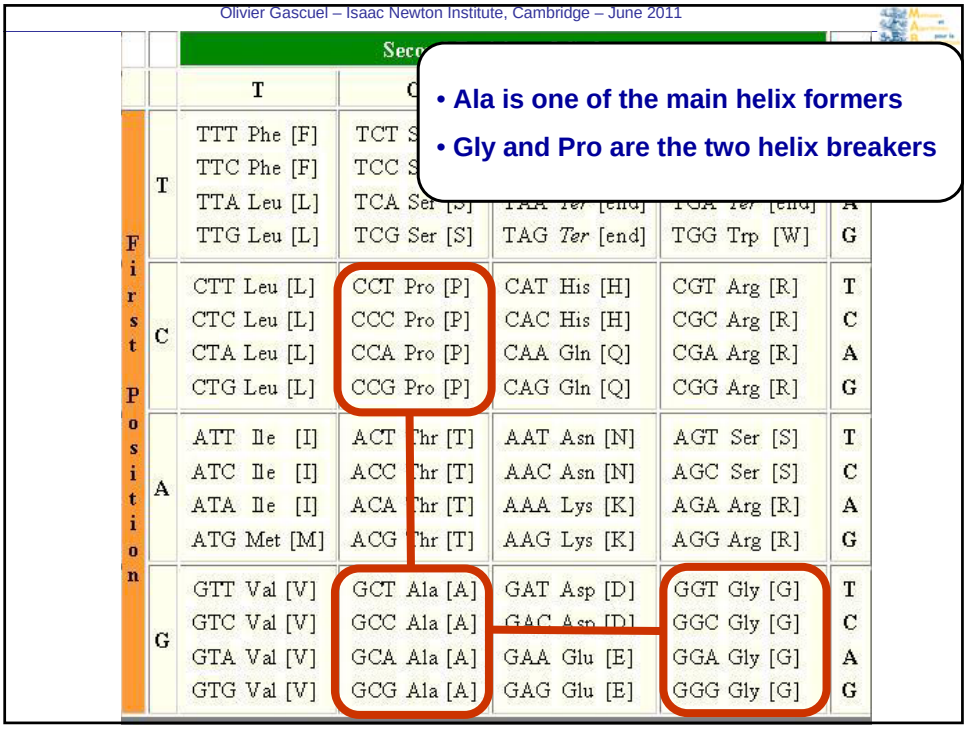
Secondary structure
Solvent accessibility

Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

The structure (and function)

- Highly conserved
- Contact maps





- **Ala is one of the main helix formers**
- **Gly and Pro are the two helix breakers**

First Position	T	C	A	G	
T	TTT Phe [F]	TCT S			T
	TTC Phe [F]	TCC S			C
	TTA Leu [L]	TCA Ser [S]	*Ter* [end]	*Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G
G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

The standard Markov model

Markovian modelling is a poor approach to model evolution; the only thing that can honestly be said in its favor is that it is about eight times as good as any other model phylogenetics people have ever tried

Democracy is a poor system of government at best; the only thing that can honestly be said in its favor is that it is about eight times as good as any other method the human race has ever tried (W. Churchill)

The standard Markov model

- A replacement rate matrix $\mathbf{M} = \left(M_{x \to y}\right)$
- $M_{x \to y}$ is the instantaneous rate of change from x to y
- $\mathbf{P}(t) = \left(P_{x \to y}(t)\right)$ the matrix of probability changes when evolutionary time is t

$$P_{x \to y}(dt) = M_{x \to y} dt$$

$$\mathbf{P}(t) = e^{\mathbf{M}t}$$

Time-reversible homogeneous replacement matrices

- **M** is defined by: $M_{x \to y} = \pi_y R_{x \leftrightarrow y}$

Equilibrium frequency

Exchangeability $\mathbf{R} = \left(R_{x \leftrightarrow y}\right)$

Tree likelihood and discrete gamma distribution

- Tree likelihood $L(T,\mathbf{M};D)=\prod_i L(T,\mathbf{M};D_i)$

- With a discrete gamma distribution of site rates

$$L(T,\mathbf{M};D)=\prod_i \sum_\rho \frac{1}{R} L(T,\rho\mathbf{M};D_i)$$

R equally weighted rate categories with rate ρ

that is, a mixture, based on a unique replacement matrix **M**

Estimating replacement matrices

- Counting approach of Dayhoff et al. (1972), using pairwise alignments of closely related proteins (PAM, JTT, …).

- Logarithmic (Gonnet et al 1992) and resolvent (Muller et al 2000) counting approaches to deal with pairs of remote proteins

- A strong tendency is to estimate different matrices for different protein groups (mitochondrial, membrane, viral, arthropoda …).

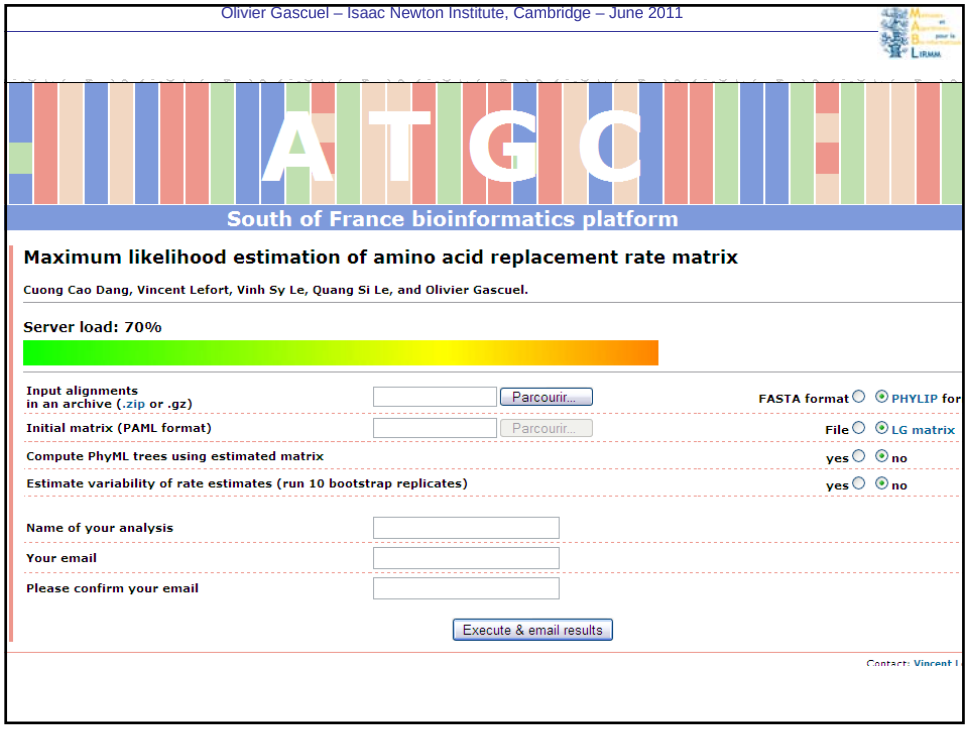
- But general matrices (e.g., JTT, WAG, LG) are still widely used

ML estimation of replacement matrices

- Counting methods are not able to deal with multiple alignments, which contain much more information than protein pairs
- ML methods exploit multiple alignments and phylogenies
- Adachi&Hasegawa (1996) proposed a ML method and estimated a replacement matrix from a unique concatenated alignments of mitochondrial genes (~3350 sites, 20 taxa)
- Whelan and Goldman (2001) estimated by ML their WAG matrix from a number of multiple alignment and a much larger data set (186 alignments, ~51.000 sites, ~900.000 AAs)

Le & OG (INI 2007, MBE 2008) estimation procedure

- We refined the Whelan and Goldman method by accounting for the gamma distribution of site rates within the matrix estimation procedure (intensive use of PhyML and XRATE)
- Our LG matrix was estimated from a very large data set extracted from Pfam (3,912 alignments, ~600,000 sites, ~6,5 millions AAs)
- The same procedure was applied to estimate a replacement matrix for Flu (Dang et al. 2010)
- This procedure is available from a web server to estimate replacement matrices for specific protein groups or species (Dang et al. 2011, in press).



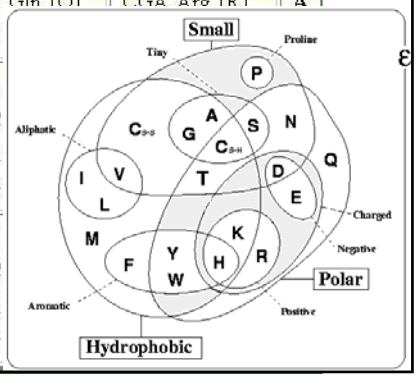
Flu (virus) replacement matrices

- Proteins are highly conserved

- We then expect that almost no hidden substitutions occurred, which makes easier the estimation procedure and should result in matrices closer to the genetic code

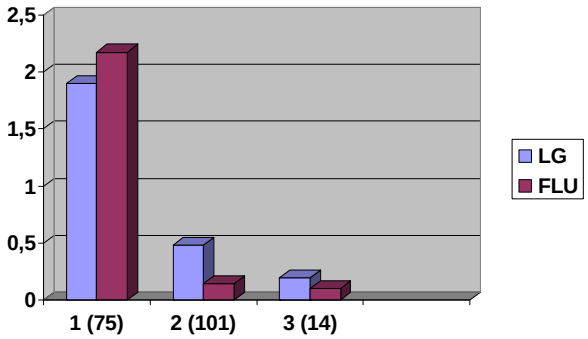
```
CLGHHAVPNGTLVKTITNDQIEVTNATELVQSSPTGRICDSPHRILDGKNCTLIDALLGD
CLGHHAVPNGTLVKTITNDQIEVTNATELVQSSSTGRICDSPHQILDGENCTLIDALLGD
----------------------VTNATELVQSSSTGRICDSPHQILDGENCTLIDALLGD
CLGHHAVPNGTIVKTITDDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLGD
CLGHHAVPNGTIVKTITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLGD
CLGHHAVPNGTIVKTITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLGD
CLGHHAVPNGTIVKTITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLGD
CLGHHAVPNGTIVKTITNDQIEVTNATELVQSSSTGGICDSPHQILDGENCTLIDALLGD
CLGHHAVANGTKVNTLTEKGIEVVNATETVETTNIKKICTQGKRPTDLGQCGLLGTLIGP
CLGHHAVSNGTKVNTLTERGVEVVNATETVERTNVPRICSKGKRTVDLGQCGLLGTITGP
```

First Position	Second Position of Codon: T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA *Ter* [end]	TGA *Ter* [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG *Ter* [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG		
A	ATT Ile [I]	ACT Thr [T]	AAT		
	ATC Ile [I]	ACC Thr [T]	AAC		
	ATA Ile [I]	ACA Thr [T]	AAA		
	ATG Met [M]	ACG Thr [T]	AAG		
G	GTT Val [V]	GCT Ala [A]	GAT		
	GTC Val [V]	GCC Ala [A]	GAC		
	GTA Val [V]	GCA Ala [A]	GAA		
	GTG Val [V]	GCG Ala [A]	GAG		



Flu and virus replacement matrices

- Proteins are highly conserved
- We then expect that almost no hidden substitutions occurred, which makes easier the estimation procedure and should result in matrices closer to the genetic code



More complex models
Accounting for exposure and secondary structure
Syst. Biol. 2010

Accounting for exposure and secondary structure

- Substitutions clearly depend on secondary structure and exposure

- Overington et al.1990; Lüthy et al. 1991; Topham et al. 1993; Wako and Blundell 1994; Goldman et al. 1996 (to infer both the structure and the phylogeny).

- Not (or rarely) used today in phylogenetics, though the structure of dozens of thousands of proteins is now available (~ 35% UNIPROT)

- We revisited the question thanks to
  (1) our improved replacement matrix estimation procedure,
  (2) the huge, current databases (HSSP)
  (3) the structural knowledge that is available for many proteins

Learning and testing data

- We extracted from HSSP (homology-derived structures of proteins) 1,771 non-redundant (sub)alignments.
- ~27 millions AAs in total.
- Secondary structure (Helix, Sheet, Coil) and exposure (Exposed, Buried) are available for all the sites, but not fully reliable (80-90% of conservation).
- We randomly selected 300 alignments as a test set, leaving 1,471 alignments to learn substitution matrices for various site categories (E, B; α, β, χ; E&α, E&β, E&χ …).

2 matrices EX | 3 matrices EHO | 6 matrices EX_EHO

Tree likelihood using site partition (e.g. codon positions)

Each category is associated to a replacement matrix; the category and corresponding matrix $\mathbf{M}_i$ are known for every site i

$$L\left(T,\left(\mathbf{M}_i\right),\Theta\middle|D\right)=\prod_i L\left(T,\mathbf{M}_i,\Theta\middle|D_i\right)$$

Extra parameters: gamma, proportion of invariant sites, etc.

No extra parameter, compared to single-matrix models ; same CPU

Mixture model (e.g. discrete gamma distribution of site rates)

Site category is unknown. We have a set of replacement matrices $\{\mathbf{M}\}$ corresponding to various categories with probabilities $\pi_{\mathbf{M}}$

$$L\left(T,\{\mathbf{M}\},\Theta\middle|D\right)=\prod_i\left[\sum_{\mathbf{M}}\pi_{\mathbf{M}}L\left(T,\mathbf{M},\Theta\middle|D_i\right)\right]$$

$|\{\mathbf{M}\}|-1$ extra parameters, regarding single-matrix models, or none when the $\pi_{\mathbf{M}}$ are known ; $|\{\mathbf{M}\}|$ x CPU

Confidence-based model (CONF/LG)
(Le & OG, Syst Biol 2010)

Site category is "known", but not fully reliable

$$L\left(T,(\mathbf{M}_i),\Theta\middle|D\right)=\prod_i\left[\begin{array}{l}c\,L\left(T,\mathbf{M}_i,\Theta\middle|D_i\right)+\\(1-c)L\left(T,\mathbf{LG},\Theta\middle|D_i\right)\end{array}\right]$$

Confidence coefficient, estimated separately for each alignment;

$c\sim 1$: useful site assignments,

$c\sim 0$: useless site assignments

One extra parameter, compared to single-matrix models ; 2 x CPU

Confidence-based model (CONF/MIX)

Site category is "known", but not fully reliable

$$L\left(T,(\mathbf{M}_i),\Theta\middle|D\right)=\prod_i\left[\begin{array}{l} c\,L\left(T,\mathbf{M}_i,\Theta\middle|D_i\right)+\\ (1-c)\sum_{\mathbf{M}}\pi_{\mathbf{M}}L\left(T,\mathbf{M},\Theta\middle|D_i\right)\end{array}\right]$$

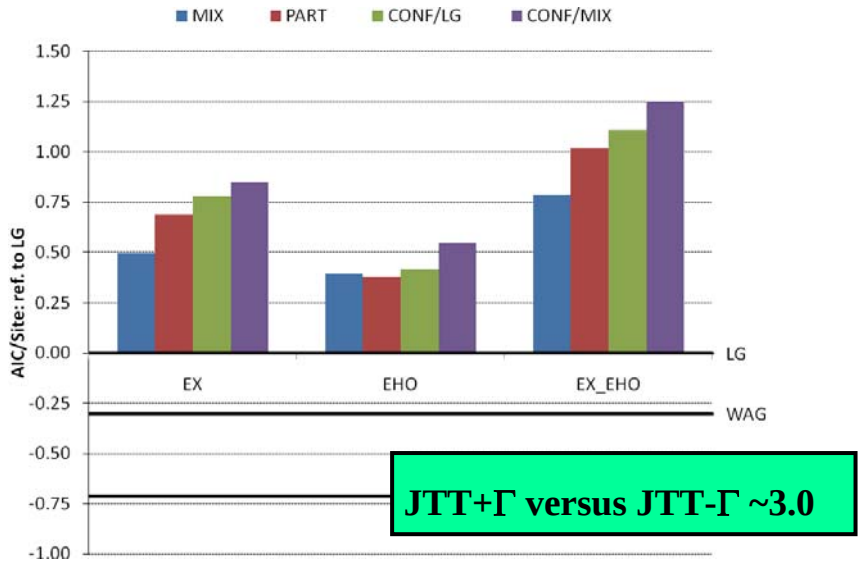
Confidence coefficient, estimated separately for each alignment;

$c \sim 1$: useful site assignments,

$c \sim 0$: useless site assignments

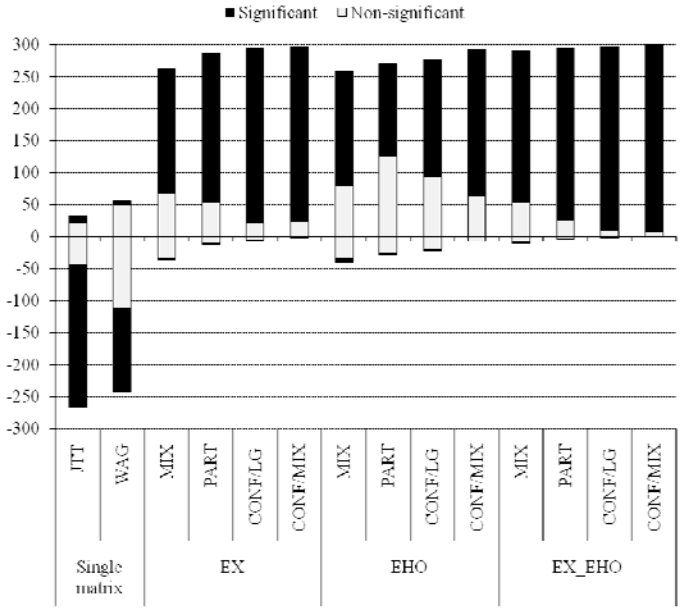
One more parameter than mixture; $\left|\{\mathbf{M}\}\right|$ xCPU

AIC/site with 300 HSSP test alignments

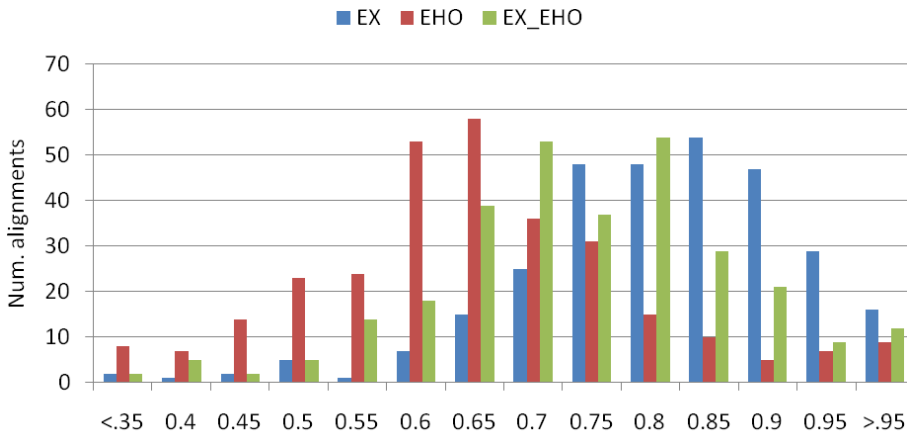


JTT+Γ versus JTT-Γ ~3.0

Results per data set (300 test alignments)



Confidence coefficient *c*



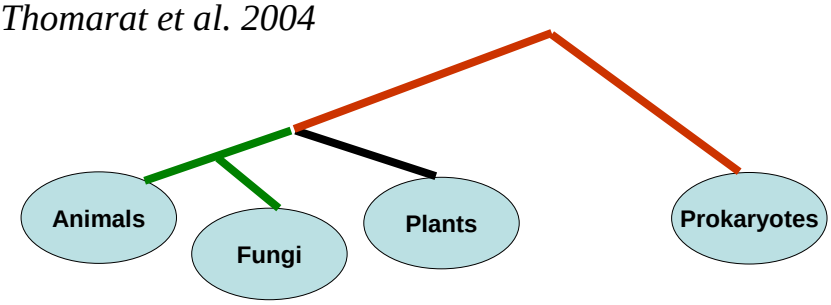
Topological impact regarding bootstrap supports

- 2 trees T1, T2, with bootstrap supports bp1, bp2
- one branch in T1 is not supported in T2 if:
  bp1>bp2+0.5, having bp2=0 when that branch is not in T2
- LG versus EX_EXO_CONF/MIX: 171 trees, 338 branches
- JTT versus EX_EXO_CONF/MIX: 213 trees, 509 branches
- JTT+Γ versus JTT-Γ: 209 trees, 510 branches

Again, we find an impact similar to that of the gamma distribution

Microsporidia case study (68 alignments)

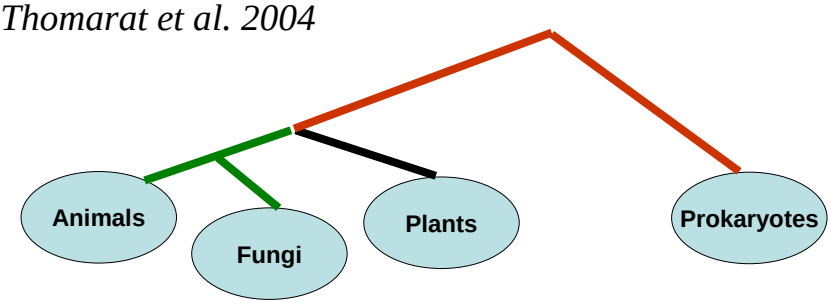
Thomarat et al. 2004



	NJ	ML JTT -Γ	ML JTT +Γ
Animals - Fungi	9	11	13
Basal Eukaryotes	56	54	51
Plants	3	3	4

Microsporidia case study (68 alignments)

Thomarat et al. 2004

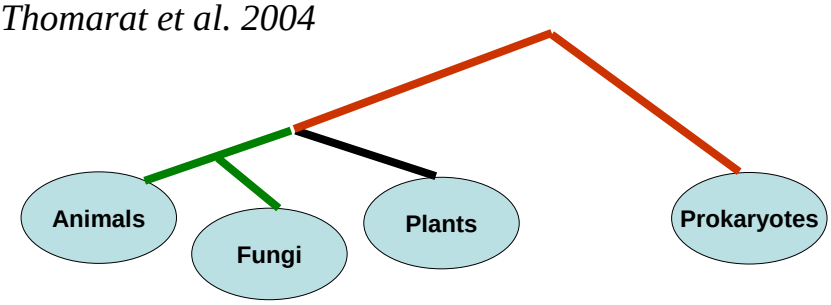


	NJ	ML JTT -Γ	ML JTT +Γ	ML LG +Γ
Animals - Fungi	9	11	13	17
Basal Eukaryotes	56	54	51	48
Plants	3	3	4	3

Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

Microsporidia case study (68 alignments)

Thomarat et al. 2004



	NJ	ML JTT -Γ	ML JTT +Γ	ML LG +Γ	EX_EHO CONF/MIX
Animals - Fungi	9	11	13	17	19
Basal Eukaryotes	56	54	51	48	44
Plants	3	3	4	3	5

Simpler models
One replacement matrix per site rate category
Submitted…

Simplified mixture models

- The structure is not available for 65% of the proteins
- Standard two-level mixture models are time and memory consuming

$$L(T,\mathbf{M};D)=\prod_i\left[\sum_{c=1}^{C}\pi_c\sum_{\rho}\frac{1}{R}L(T,\rho M_c;D_i)\right]$$

- For example with EX_EHO and 4 gamma categories, we have a mixture with 24 categories, and the computing time and memory consumption are proportional to this product (*RC*).

Simpler mixture models

- We then explored simpler one-level mixture models

$$L\left(T,\mathbf{R},\{\mathbf{M}_c\};D\right)=\prod_i\left[\sum_{c=1}^{C}\pi_c L\left(T,\rho_c\mathbf{M}_c;D_i\right)\right]$$

- **LG4M:** discrete gamma rate model **R** with 4 categories (no additional parameter compared to standard +Γ option)**:**

$$\pi_c=\frac{1}{4},\ \rho_c \text{ is gamma distributed}$$

- **LG4X:** distribution free scheme, the rates and probabilities are estimated from the data set being analyzed (i.e. 6 free parameters).

Simplified mixture model (LG4X)

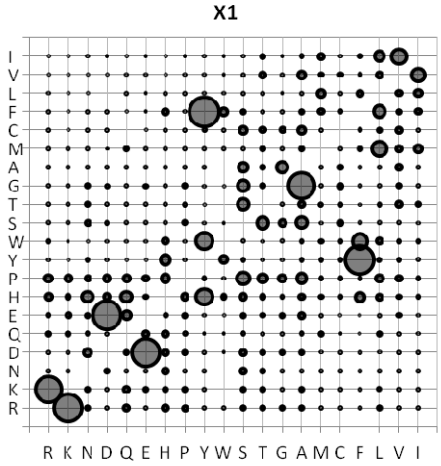
To learn the four LG4X matrices, we proceeded in a semi-supervised way:

1. We first categorized the sites using the posteriors of the 4 gamma rate categories

2. We learned with XRATE a different matrix for each category

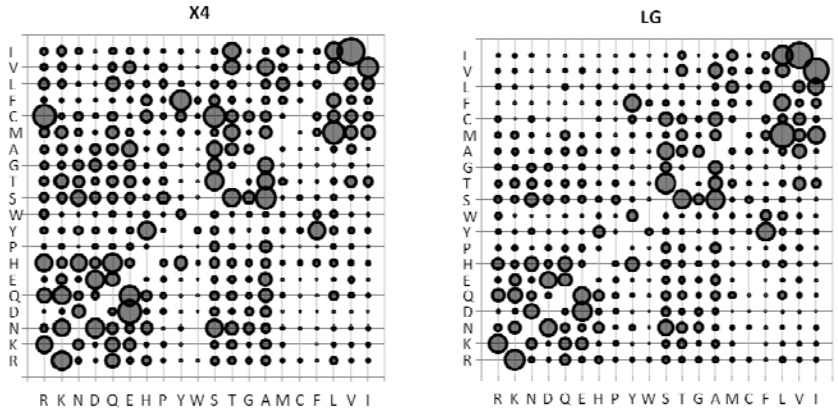
3. We categorized the site based on the posteriors and iterated the procedure until convergence (8 iterations)

Very slow matrix X1



Close to the genetic code!

Fast matrix X4



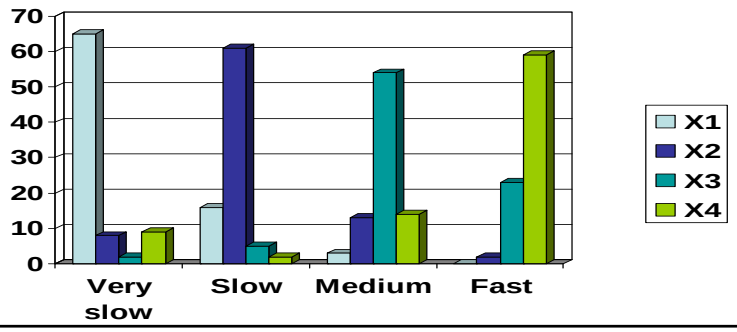
Tolerant, relatively close to LG

Clear correlation with solvent accessibility

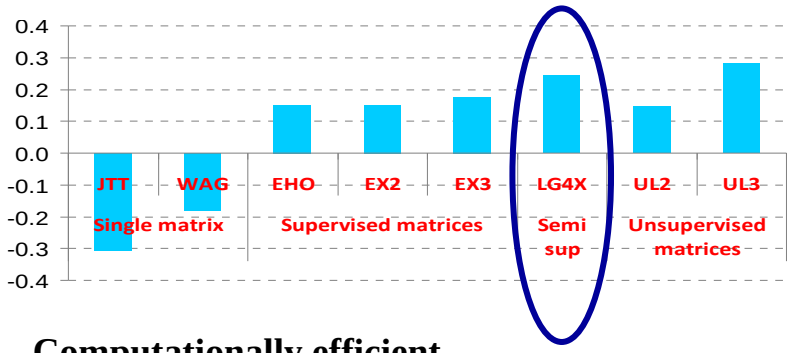
	Very Slow	Slow	Medium	Fast
ClosestMatrix	Buried (0.880)	Buried (0.885)	Intermediate (0.946)	Exposed (0.987)
Hydropathy	0.934	0.325	-0.816	-1.815

Quite flexible model

- The rates ρ are estimated independently for each data set analyzed, without any constraint.
- The "Fast" X4 matrix is sometimes the slowest…
- This illustrates the need for flexible models!



Important AIC gains



Computationally efficient

- Likelihood calculation require the same amount of time and memory as the standard model
- Some additionnal cost comes from optimizing the (8) rate model parameters

Olivier Gascuel – Isaac Newton Institute, Cambridge – June 2011

Discussion

Continuation of previous works by

- Goldstein, Thorne, Goldman (and many others) regarding solvent accessibility and secondary structure
- Yang, Lartillot, Philippe, Roger (and many others) regarding mixture models

Discussion

These studies further illustrate

- The advantage of multiple matrices, flexible models, specific models, ….
- The clear benefit of explicitly using structural information, when it is available

There is still room for improvements along these lines

Challenges for mathematicians

- **Identifiability of multiple-matrix mixture models?**
- **Convergence of ancestral states reconstructions?**

Challenge for biologists?

- **Explaining multiple simultaneous changes at the DNA level**

Big challenge for all of them together

- **Finding better models (explanatory virtue, computationally efficient…)**