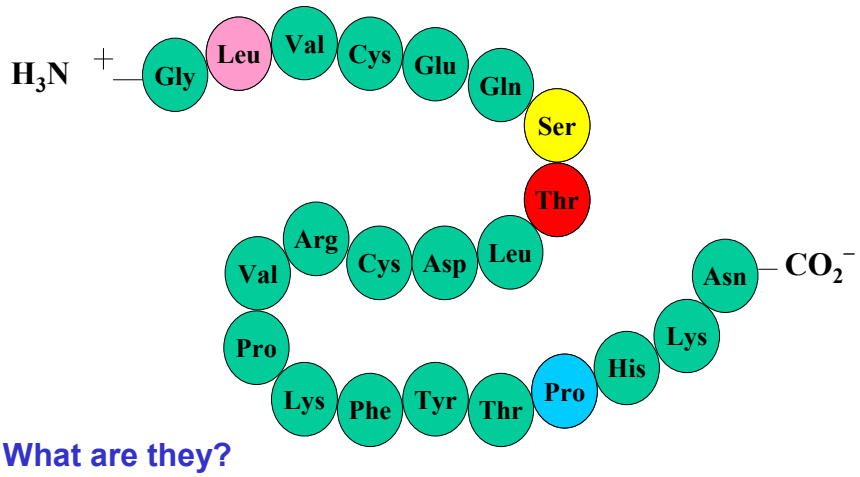
EMBL-EBI
European Bioinformatics Institute

Codon models

Carolin Kosiol
kosiol@ebi.ac.uk

EMBL-EBI

Substitution models



What are they?

Markov model: Each amino acid evolves independent of other sites' evolution and of its past history.
Most widely known example: PAM (Dayhoff et al.,1978).

Codon models. Why bother?

- Give insights into the pressures and processes of evolution.
- We will investigate protein evolution through an empirical codon model showing:
  i. Doublet and triplet nucleotide changes occur.
  ii. Physicochemical properties dominate observed substitution patterns.
  iii. Alignment procedure can lead to different substitution patterns for different evolutionary distances.
- Could be used in alignment programs, database searches, phylogeny, whole genome analysis …

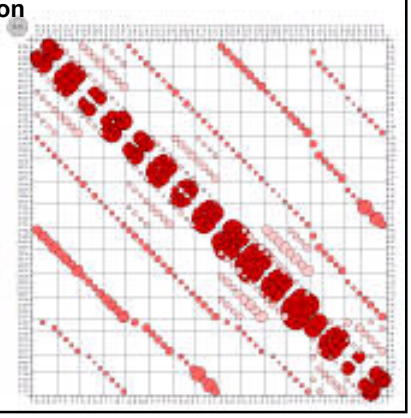
Two types of models

- Empirical models summarise the substitution patterns from large quantities of data
- Lots of successful amino acid models: PAM, JTT, WAG etc.
- Not much work on empirical codon models so far

- Mechanistic models use parameters to model factors of protein evolution
- AA models consider fitness of the amino acid as function of their physico-chemical properties (Koshi, Mindell, Goldstein, 1999)
- Codon models to detect selection consider synonymous –nonsynonymous amino acid substitution bias (Yang and Goldman, 1994)

The mechanistic codonmodel M0

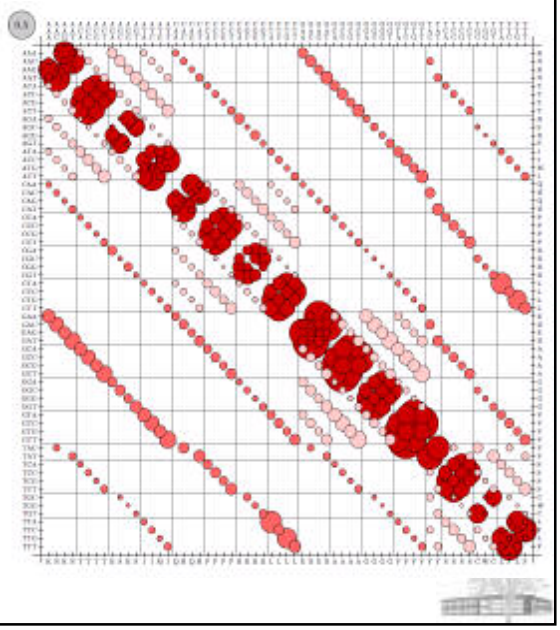
$$Q_{ij} = \begin{cases} 0 & \text{if } i \text{ -> } j \text{ is > 1 nucleotide substitution or j is a stop codon} \\ \pi_j & \text{if } i \rightarrow j \text{ synonymous transversion} \\ \pi_j \kappa & \text{if } i \rightarrow j \text{ synonymous transition} \\ \pi_j \omega & \text{if } i \rightarrow j \text{ nonsynonymous transversion} \\ \pi_j \kappa \omega & \text{if } i \rightarrow j \text{ nonsynonymous transition} \end{cases}$$



where
κ : transition/transversionrate ratio
ω : nonsynonymous/synonymous rate ratio
π_j : equilibrium frequency of codon *j*

(Goldman &Yang 1994,Yang et al. , 2000)

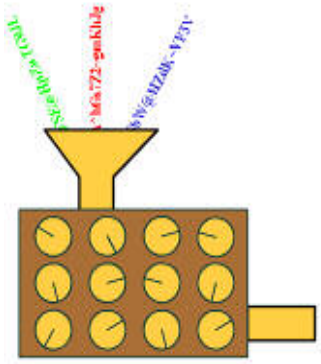
M0



How to get an empirical model?

Data:

•Translate cDNA to codon sequences
•Multiple sequence alignments
•Phylogenetic trees
e.g. Pandit 12.0 (http://www.ebi.ac.uk/goldman-srv/pandit/)



Estimation with Dart:
C++ implementation of an expectation maximization (EM) algorithm to estimate substitution matrices (Holmes & Rubin, 2003).

Result:
Instantaneous rate matrix

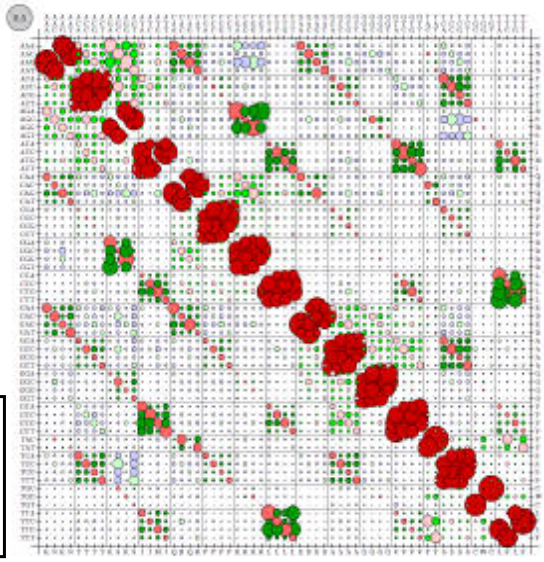
Pandit: Doublet and Triplet changes

- Multiple nucleotide changes

Single = 65.9%

Double = 24.7%

Triple = 9.4%



	positions
1 nt change	○ ○ ○
2 nt change	○ ○ ○
3 nt change	○

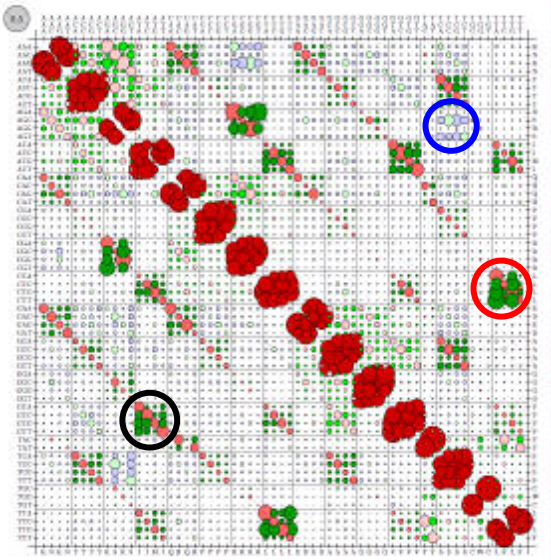
Pandit: Importance of the genetic code

Serine S ↔ S:
Direct changes require multiple nt changes (e.g. AGC ↔ TCT)

Leucine L ↔ L:
Multiple nt or successive single changes via another L (e.g. TTG ↔ CTG ↔ CTC)

Valine V ↔ Isoleucine I:
Multiple nt changes between physicochemical related amino acids (e.g. GTT ↔ ATA)

	positions
1 nt change	
2 nt change	
3 nt change	



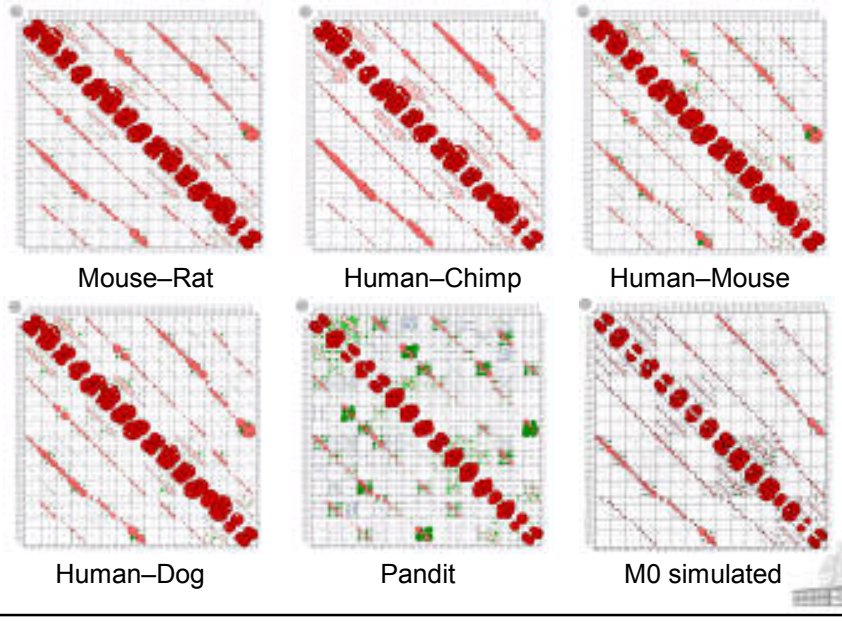
Alignment artifacts?

More data sets

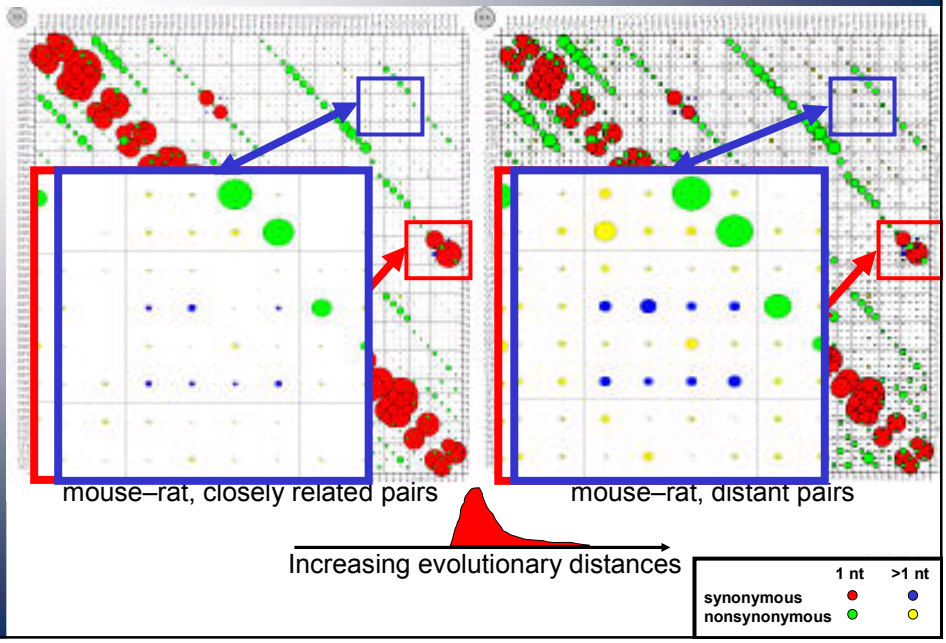
Sequences:
Pairwise alignments of the cDNA of orthologous genes were taken from vertebrate genome database Ensembl (http://www.ensembl.org)

Trees:
Evolutionary distances were estimated using PAML.

Lots of bubble plots



Mouse-Rat



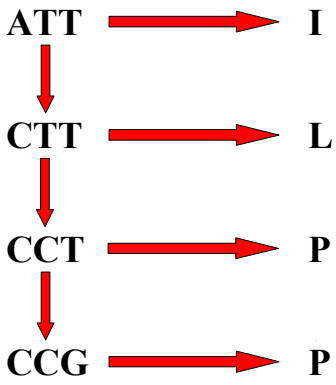
Different substitution patterns at different evolutionary distances?

- Non-evolutionary model:
  BLOSUM matrices (Henikoff et al., 1992).
- Cannot identify one process that could generate the separate matrices for protein sequences at different divergence times (Mitchison and Durbin, 1996).
- Different patterns of mutations
  Early stages: genetic code influences the probability of changes.
  Advanced stages: chemical properties dominate (Benner et al. 1994).

Question:

But how can an amino acid in a protein know where it is in time?

On the codon and the amino acid level



Aggregated Markov Process (AMP)

Continuous Markov process {X(t), t ≥ 0} on state space $\mathcal{S}$ = {AAA, AAC, …, TTT}, with equilibrium distribution π and probability matrix $P(t) = e^{tQ}$.

Genetic code:

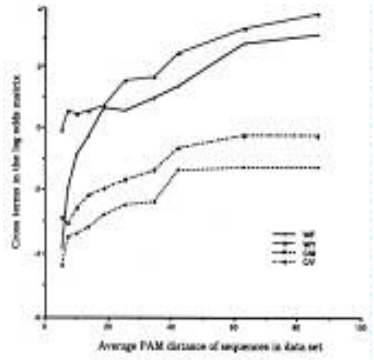
Deterministic function of Markov process Y(t) = f(X(t)) which maps the state space to aggregated set $\mathcal{A}$={A, R, N, …,V}.

The aggregated Markov process {Y(t), t≥0} on amino acids can be observed (Larget, B., 1998).

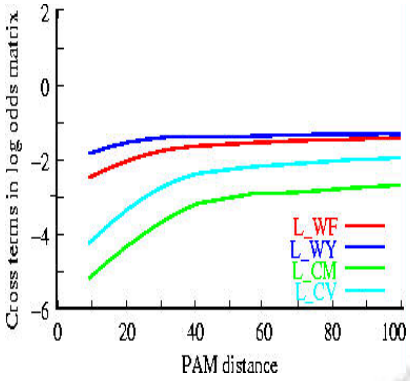
Aggregated Markov processes are a subclass of hidden Markov models (HMMs).

Similar in chemical properties but distant in the genetic code

(Benner et al. 1994)

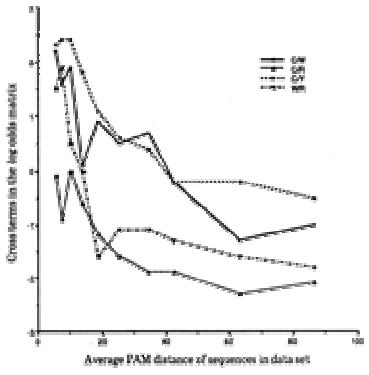


AMP simulation

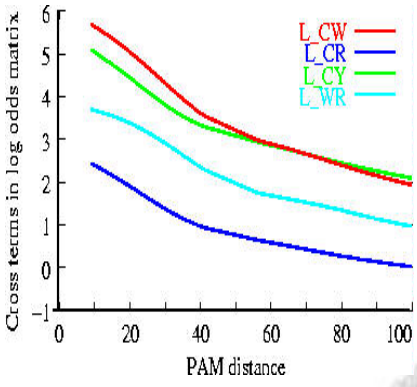


Different in chemical properties but close in the genetic code

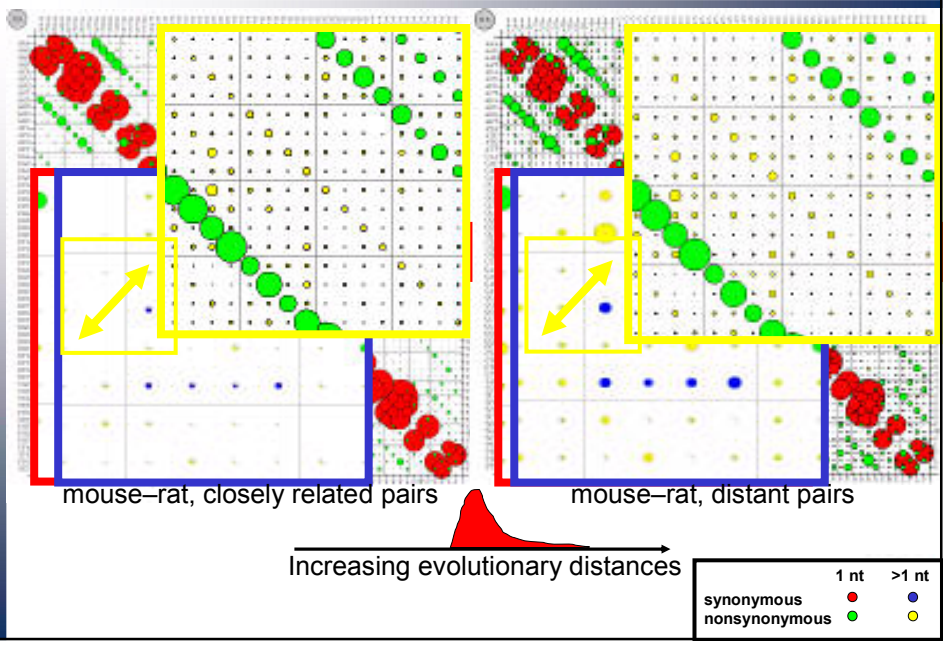
(Benner et al. 1994)



AMP simulation



Mouse-Rat



Summary

- The empirical codon model is different from standard view because we observe doublet and triplet nucleotide changes.
- Physicochemical properties dominate substitution patterns.
- Alignment artefacts can explain why we observe different substitution patterns for different evolutionary times.
- Some of the patterns cannot be explained by the alignment procedure.

Future work

- Assessment of performance of empirical codon model by likelihood comparison.
- Combine mechanistic and empirical codon models: Reintroduce mechanistic codon parameters (κ for transition-transversion bias and ω for selection).
- Investigate consequences for selection models.

Thanks

- Goldman group at EBI:
  Nick Goldman, Simon Whelan, Ari Loytynoja Irmtraud Meyer, Lee Bofkin, Fabio Pardi, Nicolas Rodriguez
- Ian Holmes (Berkeley)
- Jessica Severin and Abel Ureta-Vidal (Ensembl at EBI)
  Rasmus Nielsen (Copenhagen Univ.), Andrew Clark (Cornell)
- EMBL
- Wellcome Trust