Genetic distances and nucleotide substitution models

Comparative Genomic Analyses

Rui Borges
Vetmeduni Vienna

Genetic distances

Any two sequences derived from a common ancestor that evolve independently of each other eventually diverge

```
         Ancestral sequence
            AACCTGTGCA
                 |
    ---------------------------
    |                         |
    |                         |
Sequence 1               Sequence 2
AATCTGTGCC               AACCTATGCC
  *      *                    *   *
```

- measures of sequence divergence are genetic distances
- genetic distances provide the basis to infer evolutionary trees

Observed and expected distances

The simplest approach to measure divergence is to count the number of sites where they differ

```
                *  *
Sequence 1    AATCTGTGCC
Sequence 2    AACCTATGCC
```

- the proportion of different homologous sites is called observed distance or *p*-**distance**
- express the number of nucleotide differences per site

Observed and expected distances

The *p*-distance is a very intuitive measure but has some shortcomings

```
Sequence 1    AACCTATGCC
                  |
               *       *
Sequence 2    ACCCTATGCT
                  |
               *       *
Sequence 3    ATCCTATGCC
```

- multiple same-site mutations are counted as a single difference
- observed distances are blind to back substitutions

Observed and expected distances

Exercise
A sequence B of 20 nucleotides diverges from sequence A. These sequences are slightly different; in particular, one knows that the following replacements occurred:

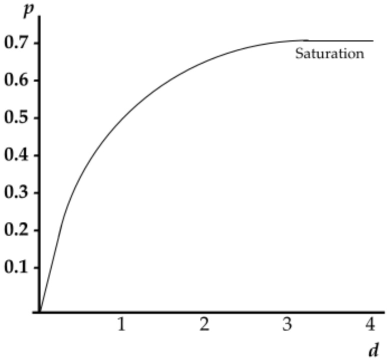
- a substitution at sites 3 and 16
- two consecutive substitutions at sites 5, 11 and 19
- a back substitution at sites 1 and 13

Calculate the *p*-distance between these sequences? How does it differ from the actual distance?

Observed and expected distances

Observed distances underestimate true **d-distances**: the actual number of substitutions per site that occurred

- observed distances are not informative about the number of substitutions that actually occurred if the degree of divergence is high
- sequences become random or reach substitution saturation



Observed and expected distances

Exercise

A sequence of 20 nucleotides evolved for 50 units of time and sampled at each five time units. The actual genetic distance *d* was traced until the present time.

Time	Sequences	d
0	ATCTA TTTAC TATCA TTTTA	0
5	ATCGA TTTAC TATCA TTTGA	3
10	ATCGC TTTAC CATCC CTTGA	7
15	ATCGT TTTAG CAACC GTTGC	12
20	AACGT TTGAG CAACC GTAGC	16
25	CACGT TTGAG CAAAA GTAGT	20
30	TACGT TTGAG CAAAA ATACT	23
35	TACGT CTTAG TAAAA ATACT	26
40	TACGT CTTAC GACAA ATACC	30
45	TCGGT CTTAC GTCAA ATACC	33
50	TGCGT CTTAC GTCAA ATACG	37

Plot the *p*-distance as a function of the true distance *d*.

Interpret the obtained plot.

Modeling substitutions with a Markov process

The substitution of nucleotides in a sequence is usually modeled as a random event

- described as a so-called continuous-time **Markov processes**.
- are fully defined by a Q matrix: specifies the relative rates of change of each of the nucleotides

$$Q = \{q_{ij}\} = \begin{array}{c|cccc} & A & C & G & T \\ A & \cdot & r_{AC} & r_{AG} & r_{AT} \\ C & r_{CA} & \cdot & r_{CG} & r_{CT} \\ G & r_{GA} & r_{GC} & \cdot & r_{GT} \\ T & r_{TA} & r_{TC} & r_{TG} & \cdot \end{array}$$

Modeling substitutions with a Markov process

The model of evolution Q allows calculating the probabilities of change from any base to any other during the evolutionary time t:

$$P(t) = e^{Qt}$$

- compute the expected genetic distances between sequences
- estimate parameters (like phylogenetic tree and branch lengths)
- test hypotheses regarding the evolutionary process

Models of sequence evolution

The simplest possible nucleotide substitution model: JC69

- introduced by Jukes and Cantor in 1969
- equilibrium frequencies of the four nucleotides are 1/4
- rates of change between any two nucleotides are the same during evolution

$$Q^{JC69} = \begin{bmatrix} . & 1 & 1 & 1 \\ 1 & . & 1 & 1 \\ 1 & 1 & . & 1 \\ 1 & 1 & 1 & . \end{bmatrix}$$

Models of sequence evolution

The probability matrix P has a formal solution for the JC69 model

$$P^{JC69}(t) = e^{Q^{JC69}t} = \begin{cases} \frac{1}{4} - \frac{1}{4}e^{-t} & \text{if } i \neq j \\ \frac{1}{4} + \frac{3}{4}e^{-t} & \text{if } i = j \end{cases}$$

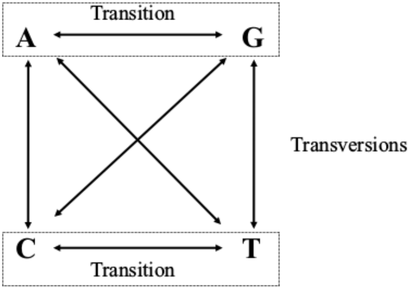
Exercise
Calculate $\lim_{t\to\infty} P_{ij}$ when $i = j$ and when $i \neq j$.
Interpret the obtained limits.

The JC69 model hardly describes the evolution of real sequences, and their assumptions are very restrictive.

Models of sequence evolution

Nucleotide replacements all fall into two major groups

- substitutions where a purine is exchanged by a pyrimidine or vice versa are called **transversions**
- all the other substitutions are **transitions**



Models of sequence evolution

Kimura 1980 model (K80)

- often referred to as Kimura's two parameter model (K2P)
- two exchangeable classes α and β between transversions and transitions, respectively

$$Q^{K80} = \begin{bmatrix} . & \alpha & \beta & \alpha \\ \alpha & . & \alpha & \beta \\ \beta & \alpha & . & \alpha \\ \alpha & \beta & \alpha & . \end{bmatrix}$$

Models of sequence evolution

Felsenstein 1981 model (F81)

- base frequencies are allowed to vary from 0.25

$$Q^{F81} = \begin{bmatrix} . & \pi_C & \pi_G & \pi_T \\ \pi_A & . & \pi_G & \pi_T \\ \pi_A & \pi_C & . & \pi_T \\ \pi_A & \pi_C & \pi_G & . \end{bmatrix}$$

Models of sequence evolution

Hasegawa-Kishino-Yano 1985 model (HKY85)

- unequal base composition
- different rates for transitions and transversions

$$Q^{HKY85} = \begin{bmatrix} . & \alpha\pi_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & . & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & . & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & \alpha\pi_G & . \end{bmatrix}$$

Models of sequence evolution

A very important class of general time reversible models (GTR or REV)

- ρ: relative substitution rate of each nucleotide to any other
- π: nucleotide frequencies at equilibrium or the stationary distribution

$$Q^{GTR} = \begin{array}{c c} & \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{array}{c} A \\ C \\ G \\ T \end{array} & \left[\begin{array}{cccc} \cdot & \pi_C\rho_{AC} & \pi_G\rho_{AG} & \pi_T\rho_{AT} \\ \pi_A\rho_{AC} & \cdot & \pi_G\rho_{CG} & \pi_T\rho_{CT} \\ \pi_A\rho_{AG} & \pi_C\rho_{CG} & \cdot & \pi_T\rho_{GT} \\ \pi_A\rho_{AT} & \pi_C\rho_{CT} & \pi_G\rho_{GT} & \cdot \end{array}\right] \end{array}$$

- Reversibility: the rate of change from nucleotide i to j is always the same than from j to i or alternatively $\rho_{ij} = \rho_{ji}$.

Models of sequence evolution

Exercise

Match the models in the first column with their sequence evolution features on the second column.

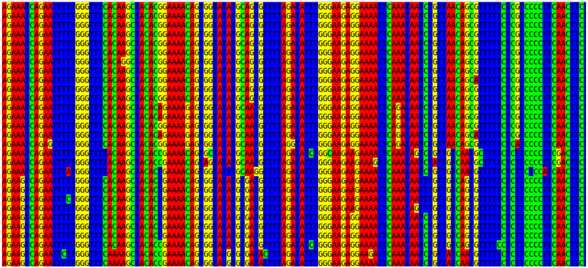
- JC69
- K80
- HKY
- GTR

- Homogeneous base composition
- Heterogeneous base composition
- A single exchangeability
- Heterogeneous exchangeabilities
- Reversibility

Rate heterogeneity among sites

The rate of nucleotide substitution can vary substantially for different positions in a sequence

- in protein-coding genes, third codon positions substitutes faster than first positions
- presence of different structural and functional constrains along the sequence
- the resulting models are represented by a suffix Γ: GTR+Γ.

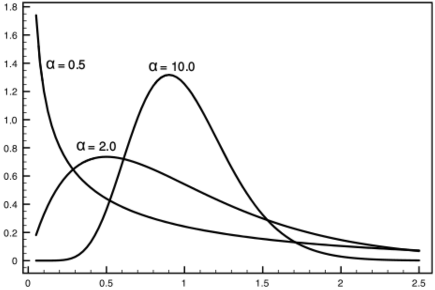


Rate heterogeneity among sites

The common approach is to use a gamma distribution with expectation 1.0 and variance $1/\alpha$, α controls form different degrees of rate heterogeneity

$$f(r) = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} e^{-\alpha} r^{\alpha-1}$$

- $\alpha > 1$: the distribution is bell-shaped, modelling weak rate heterogeneity over sites
- $\alpha < 1$: the distribution if L-shaped, describing the situation of strong rate heterogeneity



Rate heterogeneity among sites

Exercise
G-protein-coupled receptors are the largest and most diverse group of membrane receptors in eukaryotes. These cell surface receptors act like an inbox for messages in the form of light energy, peptides, lipids, sugars, and proteins. Structurally, these proteins are integral membrane proteins that possess seven membrane-spanning domains or transmembrane helices.

Does it make sense to model across site variation if we are to perform phylogenetic analyses with the GPCRs gene family? If yes, what is the expectation for the value or α?

Literature

The Phylogenetic Handbook by Lemey, Salemi and Vandamme (2009)
Cambridge University Press

- Chapter 4: sections 4.2, 4.3, 4.4 and 4.6