Models for the evolution of gene families

Jiahao Diao[1]

with Tristan L. Stark[2], David A. Liberles[2], Małgorzata M. O'Reilly[1,3],
Barbara R. Holland[1]

13-15 February 2019

The Tenth International Conference on Matrix-Analytic Methods in Stochastic Models (MAM10)

[1] School of Natural Sciences, University of Tasmania.

[2] Department of Biology, Temple University.

[3] ARC Centre of Excellence for Mathematical and Statistical Frontiers.

Outline

1. Introduction
2. Binary matrix model
3. Four-dimensional model
4. Future work

Motivation

- To model the evolution of gene families;
- We consider a two-dimensional model described in [Teufel, A. I., Zhao, J., O'Reilly, M., Liu, L., & Liberles, D. A. , 2014];
- We construct a binary matrix Markovian model to record full information;
- We construct a less complex, four-dimensional model, in which we approximate the transition rate.

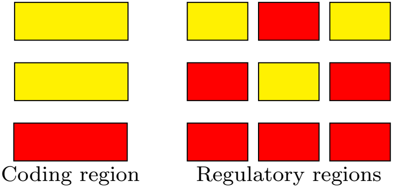
Four types of events

1. The family loses a gene.
2. A gene gains a new function.
3. One of the genes duplicates itself.
4. One of the genes loses a function.

Assumption:

Functions are protected by selective pressure.

Gene structure



- Regions hit by null mutation are coloured red;
- Regions which are protected by selective pressure are coloured yellow.

Two-dimensional model

CTMC $\{X_t : t \geq 0\}$ with state space

$$\mathcal{S} = \{(n, m) : n = 1, 2, \ldots ; m = 0, 1 \ldots, n\}$$

- n, the number of genes;
- m, the number of redundant genes.

Redundant genes are not protected by selective pressure.

Transition rates

Transition rate in two-dimensional model

1. c, duplication rate, per copy of a gene;
2. a, loss rate, per redundant copy of a gene;
3. b, loss rate, per non-redundant copy of a gene;
4. g, neofunctionalisation rate, per copy of a gene;
5. $h(t)$, subfunctionalization rate, per copy of a gene.

Here a, b, c, g are Poisson rate and function $h(t)$ can be modelled using a gamma distribution $\Gamma(k, \theta)$, as example.

Transition types

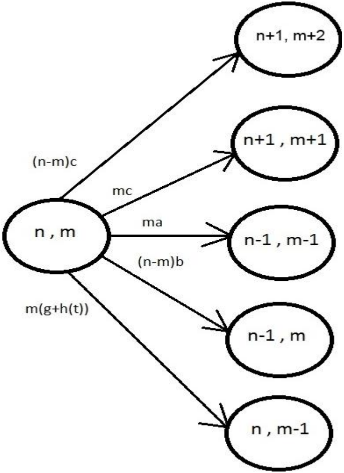


Figure: From
[Teufel, A. I., Zhao, J., O'Reilly, M., Liu, L., & Liberles, D. A. , 2014, Section 10]

The binary matrix model

CTMC $\{Y_t : t \geq 0\}$ with state space

$$\mathcal{S} = \{\mathbf{A} = [A_{i,j}] : A_{i,j} \in \{0,1\}, i = 1, \ldots, n; j = 1, \ldots, z; n, z = 1, 2, \ldots\}$$

1. n, the number of genes in the family;
2. z, the number of functions in the regulatory regions of the genes in the family;
3. $A_{i,j} = 1$ means that gene i has function j ($A_{i,j} = 0$ if gene i does not have function j).

Example

Suppose

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Here

- $n = 3$ (we have 3 genes);
- $m = 2$ (gene 2 is protected by selective pressure);
- column 3 is referred to as a pivot column.

Model setting

We assume,

1. u_c, Poisson rate of losing a row in matrix $\mathbf{A}$ [Loss of a gene];
2. u_f, Poisson rate of gaining a pivot column [A gene gains a new function];
3. u_d, Poisson rate of gaining a copy of a row in matrix $\mathbf{A}$ [Gene duplication];
4. u_r, Poisson rate of $1 \rightarrow 0$ in some entry $A_{i,j}$ [Loss of a function].

Four possible transition types (1)

A loses row *i* (family loses gene i)

(a) $(n, m) \to (n-1, m-\ell)$, $\ell = 1, \ldots, m$.

Transition type 1(a),

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ \color{red}{1} & \color{red}{0} & \color{red}{0} \end{bmatrix} \to \mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

$$(3, 2) \to (2, 0)$$

Four possible transition types (2)

$0 \rightarrow 1$ in $A_{i,z+1}$ (gene i gains function $z+1$), (E_2)

(a) $(n, m) \rightarrow (n, m-1)$;

(b) $(n, m) \rightarrow (n, m)$.

Transition type 2(b),

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \rightarrow \mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

$$(3, 2) \rightarrow (3, 2)$$

Four possible transition types (3)

Row *i* is duplicated (gene i is duplicated)

(a) $(n, m) \to (n+1, m+1)$;

(b) $(n, m) \to (n+1, m+2)$.

Transition type 3(b),

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \to \mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

$$(3, 2) \to (4, 4)$$

Four possible transition types (4)

$1 \rightarrow 0$ in $A_{i,j}$ entry (gene i loses function j), (E_4)

(a) $(n, m) \rightarrow (n, m)$;

(b) $(n, m) \rightarrow (n, m-1)$;

(c) $(n, m) \rightarrow (n-1, m-1)$;

(d) $(n, m) \rightarrow (n-1, m-2)$.

Transition type 4(a),

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \rightarrow \mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

$$(3, 2) \rightarrow (3, 2)$$

Transition rate $T_{(n,m)\to(n,m)}$

Obtain the transition rate $T_{(n,m)\to(n,m)}$,

$$T_{(n,m)\to(n,m)} = P\big((n,m)\to(n,m)\big)\lambda_{(n,m)},$$

where $P\big((n,m)\to(n,m)\big)$ is calculated as below:

Type 2 (b):

$$P\big((n,m)\to(n,m) \mid 1\to 0 \text{ in } A_{i,z+1}\big),$$

Type 4 (a):

$$P\big((n,m)\to(n,m) \mid 1\to 0 \text{ in } A_{i,j}\big).$$

Transition rate $\lambda_{(n,m)}$

The total transition rate of leaving the current state (n, m) given matrix $\mathbf{A}$ can be described as

$$\lambda_{(n,m)} = m \times u_c + n \times u_f + u_d + \left(\mathbf{1}^T \mathbf{A} \mathbf{1} - n_{piv}\right) \times u_r,$$

where n_{piv} is the number of pivot columns in matrix $\mathbf{A}$.
Here we denote

1. $p = P(A_{i,j} = 1)$, the probability of the entry $A_{i,j}$ is equal to 1;
2. E_2, we observe $1 \to 0$ in $A_{i,z+1}$ in the CTMC $\{Y_t : t \geq 0\}$;
3. E_4, we observe $1 \to 0$ in $A_{i,j}$ in the CTMC $\{Y_t : t \geq 0\}$.

Expression for $T_{(n,m)\to(n,m)}$

After the calculation, we obtain

$$\begin{aligned}
T_{(n,m)\to(n,m)} =& \Big[P\big((n,m)\to(n,m) \mid E_2\big) P\big(E_2\big) + \\
& P\big((n,m)\to(n,m) \mid E_4\big) P\big(E_4\big)\Big] \lambda_{(n,m)} \\
=& \Big[\big(1-(1-p)^{n-1}-(n-1)p(1-p)^{n-2}\big) \\
& \times \big(1-(1-p)^{z-1}\big) + (n-1)p(1-p)^{n-2} \times \frac{n-m}{n} \Big]^2 \\
& \times \Big(\mathbf{1}^T \mathbf{A} \mathbf{1} - n_{piv}\Big) \times u_r + \frac{(n-m)^2 \times u_f}{n}.
\end{aligned}$$

Remark

Calculations require the value of

1. n_{piv}, given current state,
2. p, given current state,
3. $\mathbf{1}^T\mathbf{A}\mathbf{1}$, the total number of 1s.

These can not be calculated using (n, m) only.

So the two-dimensional model $\{X_t : t \geq 0\}$ is not suitable.

State space

Consider a CTMC $\{Z_t : t \geq 0\}$ with four-dimensional state space

$$\mathcal{S} = \{(n, m, z, c) : n = 1, \ldots; m = \max\{0, n - z\}, \ldots, n; \\ z = 1, \ldots; c = z, \ldots, n \times z\}.$$

- n, the number of genes;
- m, the number of redundant genes;
- z, the number of functions in gene family;
- $c = \mathbf{1}^T \mathbf{A} \mathbf{1}$ is the total number of 1s in $\mathbf{A}$.

Possible transition

1. **A** loses row i (family loses gene i)
   (a) $(n, m, z, c) \to (n-1, m-\ell, z, c - \sum_k A_{i,k})$.
2. $0 \to 1$ in $A_{i,z+1}$ (gene i gains function $z+1$)
   (a) $(n, m, z, c) \to (n, m-1, z+1, c+1)$;
   (b) $(n, m, z, c) \to (n, m, z+1, c+1)$.
3. Row i is duplicated (gene i is duplicated)
   (a) $(n, m, z, c) \to (n+1, m+1, z, c + \sum_k A_{i,k})$;
   (b) $(n, m, z, c) \to (n+1, m+2, z, c + \sum_k A_{i,k})$.
4. $1 \to 0$ in $A_{i,j}$ entry (gene i loses function j)
   (a) $(n, m, z, c) \to (n, m, z, c-1)$;
   (b) $(n, m, z, c) \to (n, m-1, z, c-1)$;
   (c) $(n, m, z, c) \to (n-1, m-1, z, c-1)$
   (d) $(n, m, z, c) \to (n-1, m-2, z, c-1)$.

Estimating p given (n, m, c, z)

We estimate the probability $p = P(A_{i,j} = 1)$ using

$$p = \frac{c}{n \times z}.$$

Assumption

Observing $A_{i,j} = 1$ is modelled using Bernoulli trials.

Estimating n_{piv} given (n, m, c, z)

The number of pivot columns n_{piv} can be calculated as

$$n_{piv} = \begin{cases} z & \text{if } m = 0, \\ (n - m) + K & \text{if } m \geq 0. \end{cases}$$

$K = 0, 1, \ldots, z - (n - m)$, is the number of additional pivot columns.

Then we consider the expectation of n_{piv} as

$$\mathbb{E}(n_{piv}) = (n - m) + \mathbb{E}(K \,|\mathbf{A} \text{ exists}).$$

Reordered matrix **A**

$$\mathbf{A}' = \begin{array}{cc}
 & \begin{array}{cc} \overbrace{\hphantom{1 \quad 0 \quad \cdots \quad 0}}^{v} & \overbrace{\hphantom{A'_{1,v+1} \quad \cdots \quad A'_{1,z}}}^{z-v} \end{array} \\
\begin{array}{r} v\left\{\begin{array}{c} \\ \\ \\ \\ \end{array}\right. \\ m\left\{\begin{array}{c} \\ \\ \\ \end{array}\right. \end{array} &
\left.\left[\begin{array}{cccc|ccc}
1 & 0 & \cdots & 0 & A'_{1,v+1} & \cdots & A'_{1,z} \\
0 & 1 & \cdots & 0 & A'_{2,v+1} & \cdots & A'_{2,z} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\
0 & 0 & \cdots & 1 & A'_{v,v+1} & \cdots & A'_{v,z} \\
\hline
0 & 0 & \cdots & 0 & A'_{v+1,v+1} & \cdots & A'_{v+1,z} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\
0 & 0 & \cdots & 0 & A'_{n,v+1} & \cdots & A'_{n,z}
\end{array}\right]\right\} n = v + m
\end{array}$$

- $v = n - m$, is the number of non-redundant genes.

Condition for existence of $\mathbf{A}'$

Three conditions need to be considered

1. Each redundant gene has to have at least one function, $\sum_{j=v+1}^{z} A'_{i,j} \geq 1$ with $i = v+1, v+2, \ldots, n$.
2. Each function is protected by selection, $\sum_{i=1}^{n} A'_{i,j} \geq 1$ with $j = v+1, v+2, \ldots, z$.
3. Rows $\ell = v+1, ..., n$ correspond to redundant genes, there exists at least one column $\ell = v+1, ..., z$ with at least two ones in it (which is not a pivot column), $\sum_{i=1}^{n} A'_{i,\ell} \geq 2$ for some $\ell = v+1, \ldots, z$.

Unconditional distribution of K

Let N_j be the number of 1s in th column j. Then we have

$$P(K=k)=\binom{z-v}{k}P(N_{v+1}=1,N_{v+2}=1,\ldots,N_{v+k}=1,\\ N_{v+k+1}\geq 2,\ldots,N_z\geq 2).$$

It leads to

$$P(K=k)=\binom{z-v}{k}\sum_{\substack{\ell_1,\ldots,\ell_{z-v-k}\geq 2;\\ \ell_1+\ldots+\ell_{z-v-k}=c-v-k}} P(N_{v+1}=1,\ldots,N_{v+k}=1,N_{v+k+1}=\ell_1,\ldots,N_z=\ell_{z-v-k})$$

Further work

1. Complete mathematical analysis of the four-dimensional models;
2. Simulation of the binary model to understand the performance of the proposed models;
3. Fit the parameter of the model to the real data, such as TAED (the adaptive evolution database) `https://liberles.cst.temple.edu/TAED/index.html`.

References

- Stark, T. L., Liberles, D. A., Holland, B. R., & O'Reilly, M. M. , (2017)
  Analysis of a mechanistic Markov model for gene duplicates evolving under subfunctionalization.
  BMC evolutionary biology, 17(1), 38.

- Teufel, A. I., Zhao, J., O'Reilly, M., Liu, L., & Liberles, D. A. , (2014)
  On mechanistic modeling of gene content evolution: birth-death models and mechanisms of gene birth and gene retention.
  Computation, 2(3), pp. 112-130.

- Feller, W., (2008)
  An introduction to probability theory and its applications (Vol. 2).
  John Wiley & Sons.

Acknowledgement

Australian Government

Australian Research Council

This research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP180100352)

ASHMORES

Thank you!