Modeling correlated sequence mutations

Peter Arndt, Terence Hwa (UCSD)  
Chris Burge (MIT)  

Work supported by: DFG, NSF
Outline

Uncorrelated and correlated mutation

Processes

Modeling

- Computing the stationary state
- Dynamic behavior

Application

- Mutations in Human Alu repeats
Point Mutations

One distinguishes the following point mutation processes:

Purines

\[ \begin{align*}
A & \quad q \quad G \\
p & \quad p & \quad p \\
\end{align*} \]

Pyrimidines

\[ \begin{align*}
C & \quad q \quad T \\
p & \quad p & \quad p \\
\end{align*} \]

Transversions:
Rate: \( p \sim 10^{-9} \text{/(bp year)} \)
(for Humans)

Transitions:
Rate: \( q \approx 4p \)

[Walker et al. 1999, Kapitonov 1995]
The stationary state

... is quite simple

- no interactions between neighboring bases
- state space is 4-dimensional

\[
\begin{align*}
C(f_c) & \xrightarrow{p} C \\
G(f_G) & \xrightarrow{q} A(f_A) \xrightarrow{p} C \\
T(f_T) & \xrightarrow{p} T
\end{align*}
\]

\[
0 = \frac{\partial}{\partial t} \begin{pmatrix} f_A \\ f_G \\ f_C \\ f_T \end{pmatrix} = \begin{pmatrix} d & q & p & p \\ q & d & p & p \\ p & p & d & q \\ p & p & q & d \end{pmatrix} \begin{pmatrix} f_A \\ f_G \\ f_C \\ f_T \end{pmatrix}
\]

\[d = -2p - q\]

- \( f_A = f_C = f_G = f_T = 0.25 \) (for all \( p \) and \( q \))
- CG content \( f_C + f_G = 50\% \)
- Pair-correlations are simply: \( f_{ab} = f_a f_b \)
### But: Two-point Correlations

Two-point correlations are found in intergenic DNA.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\rho_{ab})</td>
<td>(\rho_{ab})</td>
<td>(\rho_{ab})</td>
<td>(\rho_{ab})</td>
<td>(\rho_{ab})</td>
</tr>
<tr>
<td>(a=A)</td>
<td>1.09</td>
<td>0.86</td>
<td>1.11</td>
<td>0.91</td>
</tr>
<tr>
<td>(b=A)</td>
<td></td>
<td>1.11</td>
<td>0.20</td>
<td>1.11</td>
</tr>
<tr>
<td>(C)</td>
<td>1.20</td>
<td>1.21</td>
<td>0.98</td>
<td>0.86</td>
</tr>
<tr>
<td>(G)</td>
<td>1.04</td>
<td>1.22</td>
<td>0.79</td>
<td>0.98</td>
</tr>
<tr>
<td>(T)</td>
<td>0.98</td>
<td>1.20</td>
<td>1.20</td>
<td>1.10</td>
</tr>
</tbody>
</table>

#### Odds Ratios:

\[
\rho_{ab} = \frac{f_{ab}}{f_a f_b}
\]

Without any two-point correlations the odds ratios would be 1.
Correlated Mutations

CpG → TpG or CpA
rate ≈ 25 x Transversion rate

Methylation
Deamination
Repair successful
Repair unsuccessful

CH₃
The Model

CGATACAT...

The Model

Correlated Pair Mutations (unidirectional)

Transitions, Transversions,
Computing the stationary state

is not equal to a enlargement of the alphabet

A T (f_{AT})
A G (f_{AG})
C A (f_{CA})
G A (f_{GA})
C G A (f_{CGA})
A A (f_{AA})
C A T (f_{CT})

no detailed balance, since there is back-reaction for the correlated mutation process
two point correlations depend on three-point correlations
non-equilibrium dynamics
Rate Equations

The corresponding rate equation for the above process:

\[ \frac{\partial}{\partial t} f_{AA} = +pf_{CA} + qf_{GA} + pf_{TA} - (2p + q)f_{AA} \]
\[ + pf_{AC} + qf_{AG} + pf_{AT} - (2p + q)f_{AA} \]
\[ + rf_{CGA} \]

Gives 16 Eq. for \( f_{ab}(t) \) + 64 Eq. for \( f_{abc}(t) \) + ...  
To truncate this hierarchy we use a Cluster Approximation:

\[ f_{abc} \approx f_{ab} f_{bc} / f_b \]

Solve the 16 non-linear differential eq. for the steady state:

\[ \frac{\partial}{\partial t} f_{AA} = G(f_{AA}, f_{AC}, ..., f_{TT}) = 0 \]
Monte Carlo Simulations

... show that the correlation length is small

\[ f_a \cdots b (r) - f_a f_b \]

→ Cluster Approximation justified
Solution of Cluster Approx.

The frequencies are given by:

\[
f_A = f_T = \frac{1}{4} + \frac{\Delta}{2} \quad f_C = f_G = \frac{1}{4} - \frac{\Delta}{2}
\]

\[
\Delta = \frac{(3p + q)r}{16(p + q)(3p + q) + 4(7p + 3q)r}
\]

Connected correlation functions:

\[
\hat{f}_{CA} = \frac{(1 + \Delta)\Delta}{4} \quad \hat{f}_{CG} = -\frac{r(1 - 2\Delta)^2 - 16(p + q)\Delta}{16r}
\]

\[
\hat{f}_{CC} = -\frac{(2\Delta - 1)(4r\Delta^2 + 8(2p + 2q + r)\Delta - r)}{32(\Delta - 1)}
\]

\[
f_{AC} = f_{AT} = f_{GC} = f_{GT} = 0 \quad \hat{f}_{ab} = f_{ab} - f_a f_b
\]
**Pair Correlations**

... calculated by the Cluster Approx.

<table>
<thead>
<tr>
<th></th>
<th>$f_a$</th>
<th>$\rho_{ab}$</th>
<th>$b=A$</th>
<th>$C$</th>
<th>$G$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a=A$</td>
<td>0.29</td>
<td>0.91</td>
<td>1</td>
<td>1.12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0.21</td>
<td>1.31</td>
<td>1.11</td>
<td>0.31</td>
<td>1.12</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.21</td>
<td>0.91</td>
<td>1</td>
<td>1.11</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.29</td>
<td>0.92</td>
<td>0.91</td>
<td>1.30</td>
<td>0.91</td>
<td></td>
</tr>
</tbody>
</table>

- Agree with Monte-Carlo results within 0.1%
- CG content is not 50%
- as expected: non-trivial pair correlations:
  - fewer CG pairs, more CA and TG pairs
- but also: changes for other dinucleotides
- we may include other processes

[http://bioinfo.ucsd.edu/dinucleotides]
Dynamic Behavior
... by Monte-Carlo Simulations

Stationary state is reached over some tens of Myr.
Initial conditions & dynamics matter on smaller time-scales
Relaxation of Dinucleotide Corr.

\[ f_{ab}(t) - f_{ab}^{\infty} \]

\[ \exp(-t / \tau_{CG}) \]

\[ \exp(-t / \tau'_{CG}) \]

\[ \exp(-t / \tau_{AG}) \]

\[ \exp(-t / \tau_{TA}) \]

\[ \tau_{AG} = 70 \text{ MYr} \]

\[ \tau_{CG} = 4.2 \text{ MYr} \]

- Many different timescales \(\rightarrow\) different clocks
- Time-scales (solid lines) calculated by cluster approx.
Alu Repeats

- Retrotransposon
- 280 bp long
- the master sequence is well conserved and CpG rich
- the oldest Alu’s are about 60 Myr old, we may use the relaxation of the CG dinucleotide
- about $10^6$ Alu sequences in the Human Genome ($\approx 10\%$)
Changes on CpG and non-CpG sites

Master sequence:

Evolved sequence:

[CpG site]
[Britten et al., 1988]
changed CpG, non-CpG site

We count the number of changes on CpG and non-CpG positions and compare them with expectations from the model.
Changes on CpG and non-CpG sites

Monte-Carlo Simulations of the model

Difference on CpG sites in %

r=25

r=0
(no correlated mutations)

Difference on non-CpG sites in %

Alu data taken from Britten et al., 1988
Summary

• The pattern of dinucleotide correlations let us deduce the underlying mutation processes
• Different correlations relax with different rates → different clocks

Outlook

• Incorporation of the model into DNA Sequence Evolution
  – Useful for comparative Genomics approach to gene finding, motif finding, ...