Modeling evolution of transcription factor binding sites

Saket Choudhary

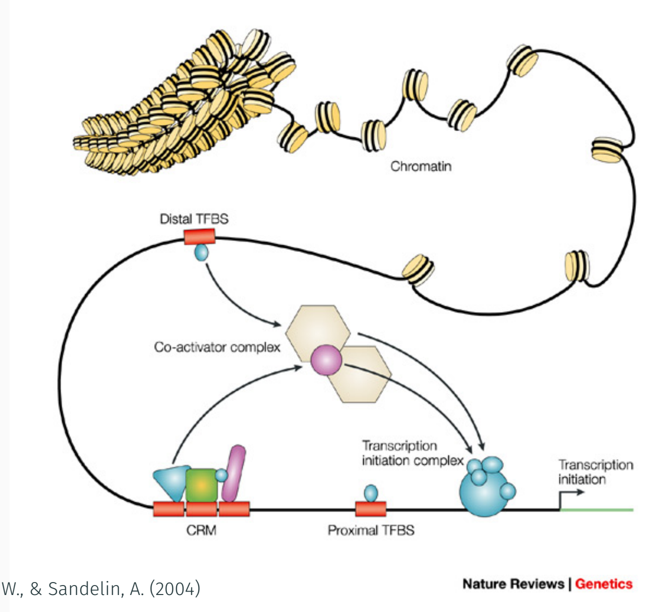
September 25, 2016

Table of contents

1. Introduction
2. Separation of mutability and selection
3. Site Level Selection
4. Functional Turnover
5. (Lineage/Specie) specific models

Introduction

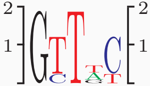
TFs bind to specific sets of short sequences



Wasserman, W. W., & Sandelin, A. (2004)

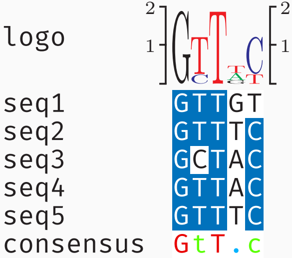
TFBS: Properties

- Short sequences (5-25bp)

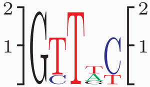
logo	
seq1	GTTGT
seq2	GTTTC
seq3	GCTAC
seq4	GTTAC
seq5	GTTTC
consensus	GtT.c

TFBS: Properties

- Short sequences (5-25bp)
- Proximity to TSS (100-1000bp)



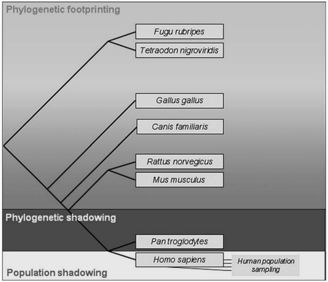
- Short sequences (5-25bp)
- Proximity to TSS (100-1000bp)
- Degeneracy

logo 

seq1	GTTGT
seq2	GTTTC
seq3	GCTAC
seq4	GTTAC
seq5	GTTTC
consensus	GtT.c

Separation of mutability and selection

Phylogenetic footprinting for identifying regulatory elements



- Selective pressure causes slower evolution of regulatory elements
- Phylogenetic footprinting – Identifying highly consered sequences in evolutionary diverse species
- Need to explicitly model phylogenetic relationship over simple conservation based approaches

Tagle et al. (1988)

Substitution Models

- Evolution can be modeled as a continuous time markov chain. Transition Matrix $P(t) = \{P_{\alpha\beta}\}$
- Rate matrix $Q = \begin{pmatrix} * & \mu_{AC} & \mu_{AG} & \mu_{AT} \\ & & \dots & \end{pmatrix}$
- $p_\alpha(t + \delta t) = p_\alpha(t) + \sum_{\beta \neq \alpha} \mu_{\beta\alpha} p_\beta(t) - \sum_{\beta \neq \alpha} \mu_{\alpha\beta} p_\alpha(t)$
- $P(t) = \exp(Qt)$
- Simple models
    - Jukes Cantor (JC69): Equal base frequencies and equal mutation rates
    - Kimura (K80): Distinguishes between transition and transversion ratios
    - Felenstein (F81): Allows different base frequencies
    - HKY: Kimura+Felenstein

Halpern Bruno Model : Accounting for position specific selection

- Substitution v/s Mutation : Different things
- JC/K80/F81: Do not explicitly differentiate mutation from selection
- HB Model:

$$\underbrace{r^i_{\alpha\beta}}_{\text{Substitution rate}} = \underbrace{\mu_{\alpha\beta}}_{\text{Probability of mutation(inst.)}} \times \overbrace{f^i_{\alpha\beta}}^{\text{Probability of fixation}}$$

  - 'Position-specific selection aware' substitution model, originally formulated for amino acids
  - All positions in the binding site evolve independently at equal rates
  - Covariation structure between different species are ignored

Halpern Bruno Model : Estimating fixation probability

$$r^i_{\alpha\beta} = \mu_{\alpha\beta} \times f^i_{\alpha\beta}$$

- Selection coefficient(s) – Relative reduction in contribution of β over α to fitness

$$F(\alpha) = 1; F(\beta) = 1 + s$$

Halpern Bruno Model : Estimating fixation probability

$$r^i_{\alpha\beta} = \mu_{\alpha\beta} \times f^i_{\alpha\beta}$$

- Selection coefficient(s) – Relative reduction in contribution of β over α to fitness

$$F(\alpha) = 1; F(\beta) = 1 + s$$

- Kimura's fixation probability: $f_{\alpha\beta} = \frac{1-e^{-2(F(\beta)-F(\alpha))}}{1-e^{-2N(F(\beta)-F(\alpha))}} = \frac{1-e^{-2s}}{1-e^{-2Ns}}$

Halpern Bruno Model : Estimating fixation probability

$$r^i_{\alpha\beta} = \mu_{\alpha\beta} \times f^i_{\alpha\beta}$$

- Selection coefficient(s) – Relative reduction in contribution of β over α to fitness
$$F(\alpha) = 1; F(\beta) = 1 + s$$
- Kimura's fixation probability: $f_{\alpha\beta} = \frac{1-e^{-2(F(\beta)-F(\alpha))}}{1-e^{-2N(F(\beta)-F(\alpha))}} = \frac{1-e^{-2s}}{1-e^{-2Ns}}$
- Weak-mutation approximation(s << 1): $f_{\alpha\beta} \approx \frac{2s}{1-e^{-2Ns}}$, $f_{\beta\alpha} \approx \frac{-2s}{1-e^{2Ns}}$

Halpern Bruno Model : Estimating fixation probability

$$r^i_{\alpha\beta} = \mu_{\alpha\beta} \times f^i_{\alpha\beta}$$

- Selection coefficient(s) – Relative reduction in contribution of β over α to fitness

$$F(\alpha) = 1; F(\beta) = 1 + s$$

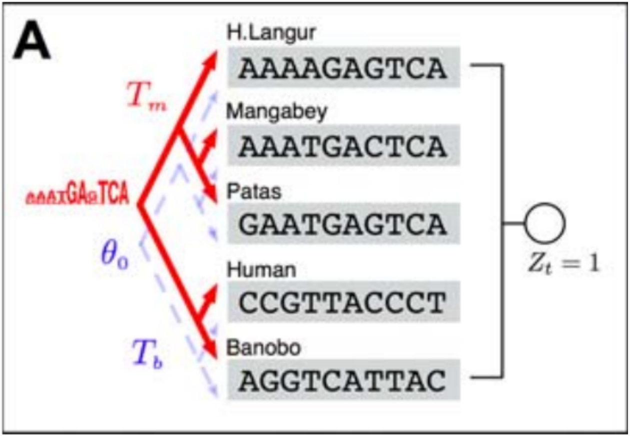
- Kimura's fixation probability: $f_{\alpha\beta} = \frac{1-e^{-2(F(\beta)-F(\alpha))}}{1-e^{-2N(F(\beta)-F(\alpha))}} = \frac{1-e^{-2s}}{1-e^{-2Ns}}$
- Weak-mutation approximation($s << 1$): $f_{\alpha\beta} \approx \frac{2s}{1-e^{-2Ns}}$, $f_{\beta\alpha} \approx \frac{-2s}{1-e^{2Ns}}$
- Reversibility condition: $\pi_\alpha \mu_{\alpha\beta} f_{\alpha\beta} = \pi_\beta \mu_{\beta\alpha} f_{\beta\alpha} \implies \frac{\pi_\beta \mu_{\beta\alpha}}{\pi_\alpha \mu_{\alpha\beta}} = \frac{f_{\alpha\beta}}{f_{\beta\alpha}} = e^{2Ns}$

Halpern Bruno Model : Estimating fixation probability

$$r^{i}_{\alpha\beta} = \mu_{\alpha\beta} \times f^{i}_{\alpha\beta}$$

- Selection coefficient(s) – Relative reduction in contribution of β over α to fitness
$$F(\alpha) = 1; F(\beta) = 1 + s$$
- Kimura's fixation probability: $f_{\alpha\beta} = \frac{1-e^{-2(F(\beta)-F(\alpha))}}{1-e^{-2N(F(\beta)-F(\alpha))}} = \frac{1-e^{-2s}}{1-e^{-2Ns}}$
- Weak-mutation approximation($s << 1$): $f_{\alpha\beta} \approx \frac{2s}{1-e^{-2Ns}}$, $f_{\beta\alpha} \approx \frac{-2s}{1-e^{2Ns}}$
- Reversibility condition: $\pi_\alpha \mu_{\alpha\beta} f_{\alpha\beta} = \pi_\beta \mu_{\beta\alpha} f_{\beta\alpha} \implies \frac{\pi_\beta \mu_{\beta\alpha}}{\pi_\alpha \mu_{\alpha\beta}} = \frac{f_{\alpha\beta}}{f_{\beta\alpha}} = e^{2Ns}$
- $f_{\alpha\beta} \propto \frac{ln \frac{\pi_\beta \mu_{\beta\alpha}}{\pi_\alpha \mu_{\alpha\beta}}}{1-\frac{\pi_\alpha \mu_{\alpha\beta}}{\pi_\beta \mu_{\beta\alpha}}} \implies r_{\alpha\beta} = \mu_{\alpha\beta} \frac{\ln \frac{\pi_\beta \mu_{\beta\alpha}}{\pi_\alpha \mu_{\alpha\beta}}}{1-\frac{\pi_\alpha \mu_{\alpha\beta}}{\pi_\beta \mu_{\beta\alpha}}}$

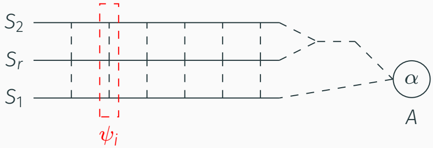
TFBS Prediction: Using HB Model over background



Modeling full phylogeny as one component: HB or JC/F81/HKY.

$F(x|\theta) = \frac{\log P(S|HB)}{\log P(S|JC)}$

HB model: Example with aligned sequences



MSA of Orthologous Sequences

$$
\begin{aligned}
P(\psi_i) &= \sum_{\alpha} P(\psi_i, A_i = \alpha|\theta) \\
&= \sum_{\alpha} P(A_i = \alpha) P(\psi_i|A_i = \alpha, \theta) \\
&= \sum_{\alpha} P(A_i = \alpha) \prod_{s_i} P(s_i|A_i = \alpha, \theta)
\end{aligned}
$$

$$S = \{\psi_1, \psi_2, \ldots, \psi_L\};$$

$$\psi_i = \{s_1^i, s_2^i, \ldots, s_N^i\}$$

$A =$ Unobserved ancestral sequence

Site Level Selection

Selection acting on whole TFBS as a unit

- Substitution rates are position specific in TFBS but independence assumption does not necessarily hold

Selection acting on whole TFBS as a unit

- Substitution rates are position specific in TFBS but independence assumption does not necessarily hold
- Intuition: A TFBS will retain functionality if it is close enough to optimality even if a crucial nucleotide undergoes substitution (and eventually getting fixed)

Selection acting on whole TFBS as a unit

- Substitution rates are position specific in TFBS but independence assumption does not necessarily hold
- Intuition: A TFBS will retain functionality if it is close enough to optimality even if a crucial nucleotide undergoes substitution (and eventually getting fixed)
- The same substitution in a far less optimal site might lead to a functional loss

Selection acting on whole TFBS as a unit

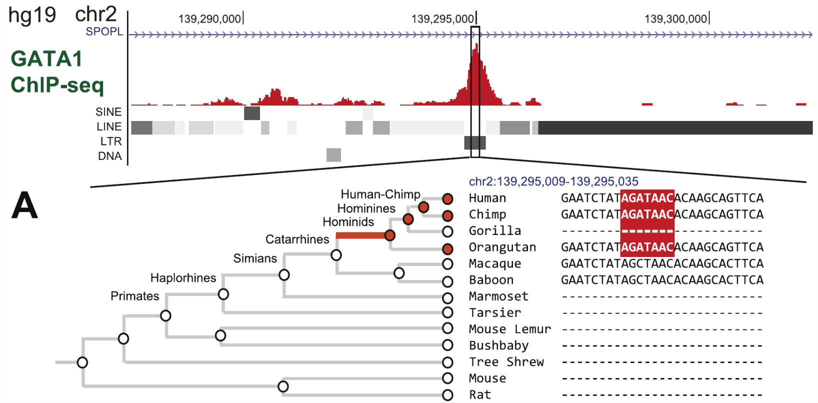
- Substitution rates are position specific in TFBS but independence assumption does not necessarily hold
- Intuition: A TFBS will retain functionality if it is close enough to optimality even if a crucial nucleotide undergoes substitution (and eventually getting fixed)
- The same substitution in a far less optimal site might lead to a functional loss
- A better model would be to account for substitution of entire site i.e. site-level selection treating binding sites as evolutionary units

Selection acting on whole TFBS as a unit

- Substitution rates are position specific in TFBS but independence assumption does not necessarily hold
- Intuition: A TFBS will retain functionality if it is close enough to optimality even if a crucial nucleotide undergoes substitution (and eventually getting fixed)
- The same substitution in a far less optimal site might lead to a functional loss
- A better model would be to account for substitution of entire site i.e. site-level selection treating binding sites as evolutionary units
- How: Reformulate the previous problem for two sites a, b instead of bases

Functional Turnover

TFBS Turnover



Functional turnover: TFBS can be gained or lost during evolution

Functional turnover: Birth & Death Process

Aim: Detect lineage-specific rates of TFBS evolution and the branch of origin of individual TFBS

- Binding sites are known to show turnover: TFBS can be gained/lost during speciation events
- Estimate rate of birth α and death β from orthologous sequences
- Infer ancestral states; branch of origin

Functional turnover: Birth & Death Process

$$
\begin{aligned}
w(t) &= \text{Probability that TFBS exists at time } t \\
\alpha, \beta &= \text{Birth, death rate respectively} \\
w(t+1) &= \alpha(1 - w(t)) + (1 - \beta)w(t) \\
w'(t) &= \alpha - (\alpha + \beta)w(t)
\end{aligned}
$$

We formulate two type of solutions, $u(t), v(t)$ such that: $u(t)$ represents those class of motifs present at $t = 0$ and $v(t)$ represents class of motifs that did not exist at $t = 0$.

Functional turnover: Birth & Death Process

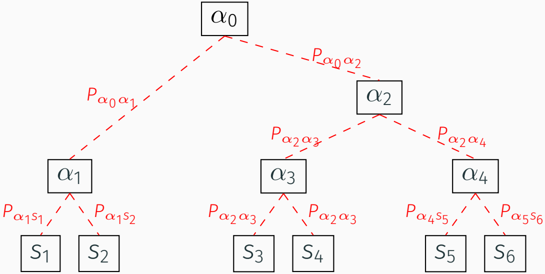
Let $p_{ij}(t)$ represent the probability of observing j motif occurrences after t , initial i

$$u(t) = \frac{1}{\alpha + \beta}(\alpha + \beta e^{-(\alpha+\beta)t})$$
$$v(t) = \frac{\alpha}{\alpha + \beta}(1 - e^{-(\alpha+\beta)t})$$

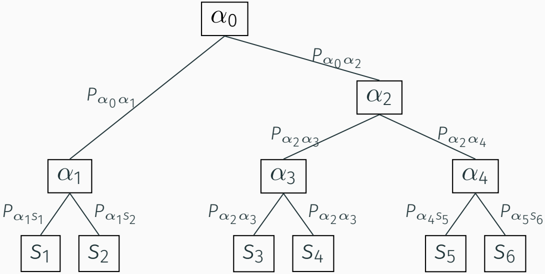
- At each node calculate the likelihood of observing daughter nodes given α, β
- Determine most likely ancestral state using MLE
- Infer branch of origin

(Lineage/Specie) specific models

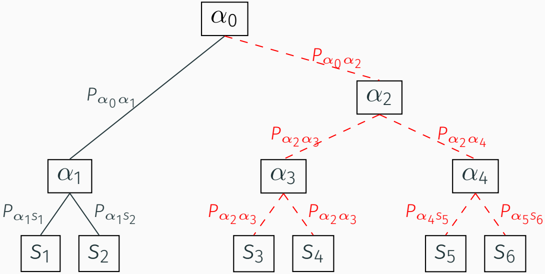
Full phylogeny evolving following motif model



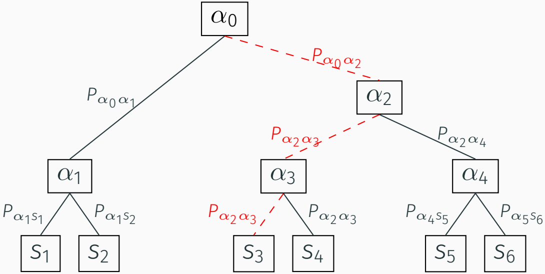
Full phylogeny evolving following background model

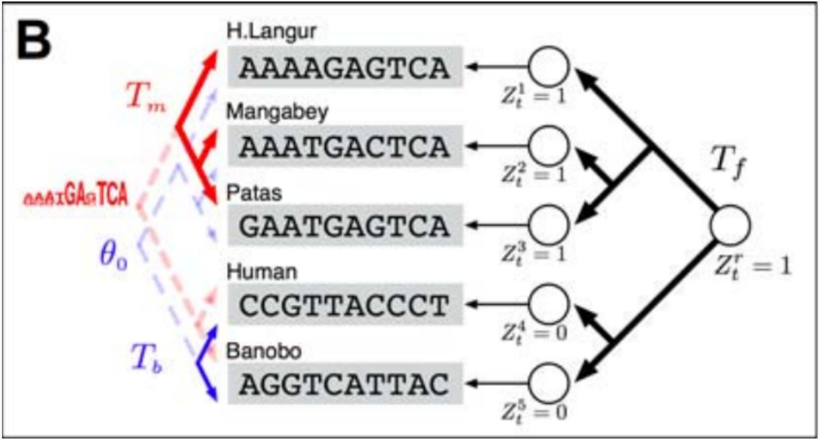


Lineage Specific Evolution



Specie Specific Evolution





Lineage specific model

Lineage Specific Evolution: Model

- Explicitly model functional turnover long T_f as a JC substitution process

$$P_f = \begin{pmatrix} \frac{1}{2} + \frac{1}{2}e^{-2\beta} & \frac{1}{2} - \frac{1}{2}e^{-2\beta} \\ \frac{1}{2} - \frac{1}{2}e^{-2\beta} & \frac{1}{2} + \frac{1}{2}e^{-2\beta} \end{pmatrix}$$

  $\beta =$ branch length

- Conditioning on TFBS functionality to model nucleotide substitution
- Capture function-specific evolution in every lineage

Summary

- HB model accounts for selection in TFBS evolution
- HB model can be extended to allow TFBS as a unit of evolution
- Turnovers can be treated in birth-death framework
- More general models can account for turnover and functional dependency across lineages

Questions?

Ornstein-Uhlenbeck Model I

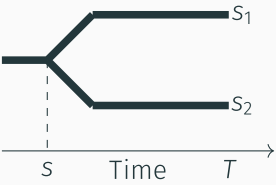
- HB models neglects lineage or specie specific selection
- OU models this gap by accounting for lineage/specie specific selection by requiring regime specific optima to be obtained
- OU models can model evolution by defining a quantitative trait as a score attached to the TFBS: $X(t)$
- Motivation: Account for the optima in the phylogeny regime assuming the change in optima coincide with phylogenetic branch points
- $X(t)$ evolves by two components one deterministic(selection), other stochastic (mutation)

Ornstein-Uhlenbeck Model II

$$
\begin{aligned}
dX(t) &= \alpha(\theta - X(t)) + \sigma dB(t) \\
\alpha &= \text{Strength of selection} \\
\theta - X(t) &= \text{Distance from optimum value} \\
\sigma &= \text{strength of random drift} \\
dB(t) &= \text{random white noise}
\end{aligned}
$$

Farther the TFBS from 'optimum' $\implies$ higher the selection force

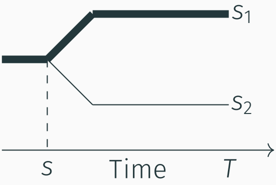
Ornstein-Uhlenbeck Model: Multivariate normal



$$E[\mathrm{X}(t)] = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$$

$$\Sigma = \sigma^2 \begin{pmatrix} T & s \\ s & T \end{pmatrix}$$

s_1, s_2 – BM



$$\begin{aligned} E[X_1(T)] &= \theta_0 e^{-\alpha T} + \theta_1(1 - e^{-\alpha T}) \\ E[X_2(T)] &= \theta_0 e^{-\alpha T} + \theta_1 e^{-\alpha(T-s)}(1 - e^{-\alpha s}) \\ &\quad + \theta_2(1 - e^{-\alpha(T-s)}) \end{aligned}$$

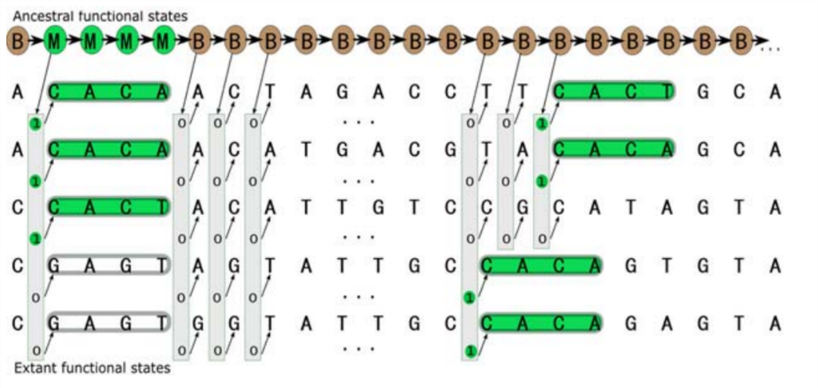
s_2 – new optimum regime, s_1 – ancestral

Jukes Cantor

$$Q = \begin{pmatrix} -\frac{3\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & -\frac{3\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & -\frac{3\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & -\frac{3\mu}{4} \end{pmatrix}$$

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{3}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

Lineage Specific Evolution



Ancestor = background $\implies$ evolution independent

Ancestor = motif $\implies$ TFBS evolves as unit