

Chapter 3: Phylogenetics

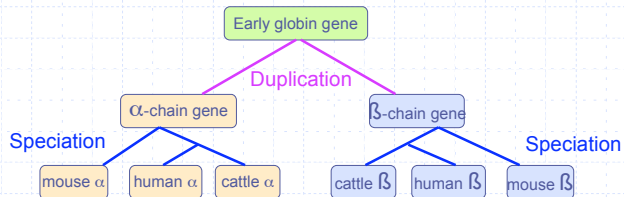
3.3 Markovian Modeling of Evolution

Prof. Yechiam Yemini (YY)
Computer Science Department
Columbia University

COMSW4761--2007

Modeling Evolutionary Changes

- Evolution: replication & speciation
- Speciation: changes + selection
 - Changes: substitutions, indels, reversals...
- Let $C(X,Y)$ = # changes in two aligned sequences X,Y
- How do we compute:
 - “evolutionary distance” $D(X,Y)$ from the # of changes $C(X,Y)$?
 - likelihood of a given evolutionary pathway?
- Need a statistical model of evolutionary changes



COMSW4761--2007

2

Statistical Modeling of Evolution

- Assume
 - Changes are random
 - Change statistics is independent of time (stationary)
 - Change statistics is independent of sites (homogeneous)
- Construct and apply Markovian model of site-evolution
- But first, how good are these assumptions?

COMSW4761--2007

3

Substitution Statistics Can Vary With Codon

<http://www.cs.ucsb.edu/~ambuj/Courses/bioinformatics/lectures.html>

- Mutations operate on DNA, selection operates on proteins
 - Synonymous mutations do not impact protein
- GGG }
GGC }
GGU }
GGA }

Glycine }
Synonymous substitutions }

GAU }
GAC }
GAA }
GAG }

Asp }
Glu }
- Codon positions do not share same substitution statistics
 - Synonymous substitutions are not subject to selection filters
 - Rate of synonymous substitutions best approximates mutations rate
 - Substitutions can cycle back: $G \Rightarrow A \Rightarrow G$; observed rate provides underestimate

Region	#of sites (bp)	# of Changes	Substitution Rate (sub/ site/10 ⁹ year)
Nondegenerate	302	17	0.56
Twofold degenerate	60	10	1.67
Fourfold degenerate	85	20	2.35

COMSW4761--2007

4

Synonymous vs. Non-Synonymous Sites

TABLE 3.3 Ratios of synonymous differences per synonymous site (K_s) and nonsynonymous differences per nonsynonymous site (K_a) for a variety of mammalian genes.

Gene	Codons (in human)	Human/mouse		Human/cow		Human/rabbit		Mouse/cow		Mouse/rabbit		Cow/rabbit		Averages	
		K_s	K_a	K_s	K_a	K_s	K_a	K_s	K_a	K_s	K_a	K_s	K_a	K_s	K_a
Erythropoietin	194	0.481	0.063	0.242	0.068	0.394	0.070	0.495	0.076	0.480	0.058	0.342	0.071	0.406	0.068
Growth hormone	217	0.321	0.100	0.236	0.106	0.220	0.113	0.380	0.046	0.396	0.027	0.244	0.048	0.299	0.073
Prolactin receptor	621	0.304	0.082	0.249	0.122	0.321	0.072	0.358	0.124	0.413	0.088	0.300	0.114	0.324	0.100
Prolactin	226	0.364	0.098	0.368	0.085	0.395	0.064	0.382	0.112	0.307	0.131	0.521	0.064	0.390	0.092
Serum albumin	610	0.528	0.062	0.329	0.067	0.324	0.075	0.477	0.065	0.500	0.065	0.327	0.067	0.414	0.067
Alpha globin	143	0.584	0.022	0.236	0.025	0.204	0.038	0.505	0.025	0.539	0.041	0.242	0.048	0.385	0.033
Beta globin	148	0.324	0.033	0.271	0.046	0.294	0.015	0.263	0.062	0.392	0.039	0.333	0.059	0.313	0.042
Prothrombin	608	0.033	0.687	0.033	1.040	0.075	1.602	0.196	0.887	0.037	1.442	0.078	0.318	0.075	0.996
Apolipoprotein E	317	0.199	0.148	0.132	0.117	0.108	0.114	0.187	0.160	0.165	0.144	0.125	0.126	0.153	0.135
Carbonic anhydrase I	336	0.255	0.159	0.203	0.149	0.207	0.138	0.338	0.113	0.284	0.115	0.187	0.117	0.246	0.132
PS3	392	0.372	0.059	0.351	0.061	0.382	0.045	0.457	0.067	0.412	0.054	0.378	0.056	0.392	0.057
Histone 2A	115	0.967	0.057	1.110	0.057	0.174	0.034	0.298	0.006	1.176	0.025	1.192	0.025	0.820	0.033
Column averages		0.394	0.131	0.313	0.162	0.258	0.198	0.361	0.145	0.425	0.186	0.356	0.093	0.351	0.152

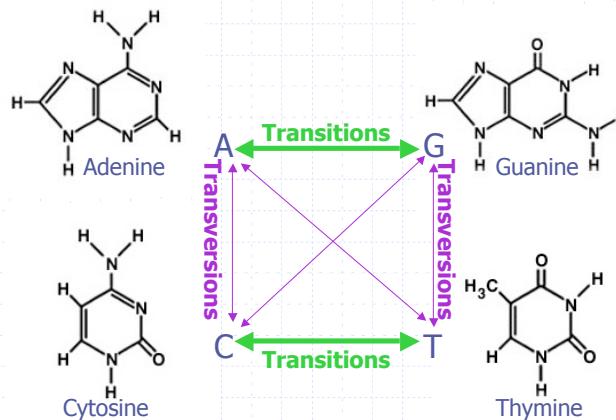
COMSW4761--2007

5

Substitution Statistics Depends On Nucleotides

Two types of substitutions:

- Transitions: $A \leftrightarrow G$ (purine), $C \leftrightarrow T$ (pyrimidine)
- Transversions: purines {A,G} \leftrightarrow pyrimidines {C,T}
- Transitions are more likely than transversions, which need to change ring structure

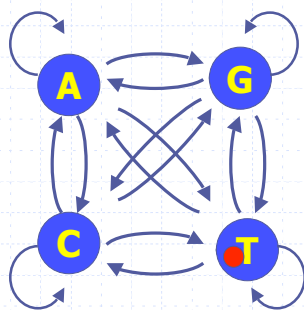


COMSW4761--2007

6

Markov Chains Review

- How do we model statistical sequence changes?
- Start with a finite state machine (FSM) (regular grammar)
- An FSM path generates unique sequence of changes
- Example: $S(t)$ = state at time t



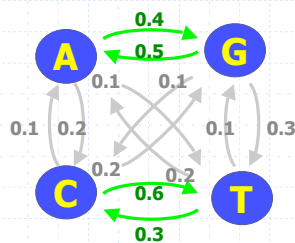
$S(0)=T$
 $S(1)=A$
 $S(2)=G$
 $S(3)=C$
 $S(4)=G$

COMSW4761--2007

7

Markov Chain: FSM + Transition Probabilities

- Finite, homogeneous Markov Chain assumptions:
 - Markovian: transition probabilities depend on current state only
 $P[S_{t+1}=j | S_t=i, S_{t-1}=i_{t-1}, \dots, S_0=i_0] = P[S_{t+1}=j | S_t=i]$
 - Homogeneity: transition probabilities are time independent
 $P[S_{t+1}=j | S_t=i] = a(i,j)$
- Notes:
 - Transition matrix is stochastic: $a(i,j) \geq 0, \sum_j a(i,j) = 1$ ($A \geq 0, A\mathbf{1} = \mathbf{1}$)
 - Diagonal elements of A are given by: $a(i,i) = 1 - \sum_{k \neq i} a(i,k)$



Transition Matrix

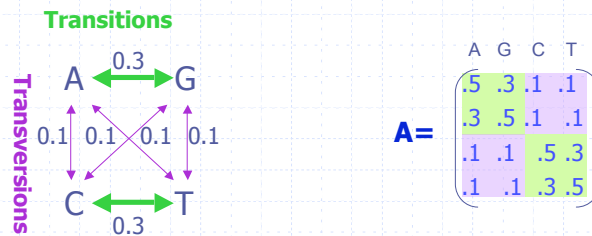
	A	G	C	T
A	0.3	0.4	0.2	0.1
G	0.5	0.1	0.1	0.3
C	0.1	0.2	0.1	0.6
T	0.2	0.1	0.3	0.5

COMSW4761--2007

8

Notes

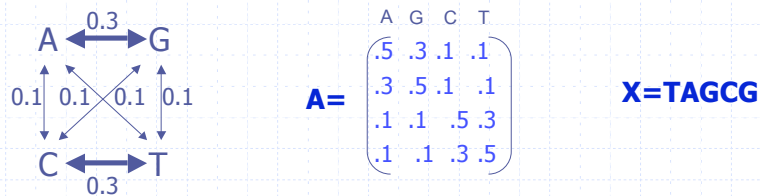
- Additional assumptions:
 - A is symmetric: identical transition rates $X \leftrightarrow Y$
 - This implies that A is doubly stochastic: $A\mathbf{1}=\mathbf{1}^T A=\mathbf{1}$
 - Transversion/transition rates are independent of specific nucleotides
- Do these assumptions match nature's mechanisms?



COMSW4761--2007

9

Computing Probability Of Evolutionary Pathways



- How to compute $p(X)=p(\text{TAGCG})$
- Bayes formula: $p(A \cap B) = p(A)p(B|A)$
 - $p(\text{TAGCG}) = p(\text{TAGC})p(G|\text{TAGC})$
 - $p(X) = p(T)p(A|T)p(G|TA)p(C|\text{TAG})p(G|\text{TAGC})$
- Computing pathway probability
 - $p(X) = p(T)p(A|T)p(G|A)p(C|G)p(G|C) = P(T)(0.1^{\#\text{transversions}})(0.3^{\#\text{transitions}})$
 - $p(X) = p(T)*0.1*0.3*0.1*0.1 = p(T)*0.3*0.1^3$

COMSW4761--2007

10

Evolution of Markov Chains

- Let $\pi_k(t) = P[S(t)=k]$ -- the probability of reaching k at time t
 - Define $\underline{\pi}(t) = (\pi_1(t), \pi_2(t), \dots, \pi_n(t))$ -- the distribution of states at time t

- Evolution equation:

$$\pi(j, t+1) = p(S_{t+1}=j) = \sum_k \pi(k, t) a(k, j) \quad \rightarrow \quad \underline{\pi}(t) = \underline{\pi}(0) \mathbf{A}^t$$

$$\underline{\pi}(t+1) = \underline{\pi}(t) \mathbf{A}$$

- Steady state probabilities: $\underline{\pi} = \lim_{t \rightarrow \infty} \underline{\pi}(t) \rightarrow \underline{\pi} = \underline{\pi} \mathbf{A}$,

- E.g., for the matrix below the steady state is $\underline{\pi} = (.25, .25, .25, .25)$
- Is this a good model for nature?
- Note: conditions for existence & uniqueness of the limit can be established
- Exercise: show that a doubly stochastic matrix has a uniform steady state distribution $\underline{\pi} = (1/n, 1/n, \dots, 1/n)$

$$\mathbf{A} = \begin{pmatrix} .5 & .3 & .1 & .1 \\ .3 & .5 & .1 & .1 \\ .1 & .1 & .5 & .3 \\ .1 & .1 & .3 & .5 \end{pmatrix}$$

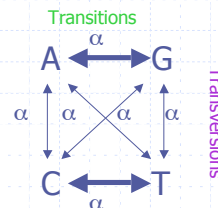
COMSW4761--2007

11

Evolutionary Markov Chain Models

- Simplest model: all substitutions have same rate
- The Jukes-Cantor transition model (69)

$$A = \begin{bmatrix} 1 - 3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1 - 3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1 - 3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1 - 3\alpha \end{bmatrix}$$



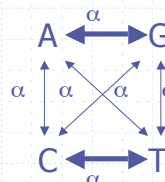
- Need to model time between changes
 - Discrete time vs. continuous time
 - From Markov chain to a Markov process

COMSW4761--2007

12

Temporal Evolution

- Mutations occur at intervals of random durations
 - The Markov Chain model assumes deterministic durations
 - How do we model random transition durations?
- Need a continuous-time model: Markov Process
 - We use a simplified introduction to Markov Processes
- Consider $a(i,j)$ as the transition rates
 - $p[S(t+\Delta t)=j|S(t)=i]=a(i,j)\Delta t$
 - Stationarity: $a(i,j)$ is time-independent

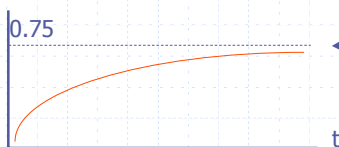


COMSW4761--2007

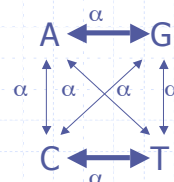
13

Evolution of The Jukes-Cantor Model

- Goal: compute substitution probabilities over time
 - Let $\pi(t)$ be the probability of no substitution during period t
- We track $\pi(t)$; start with $\pi(0)=1$
 - $\pi(t+\Delta t)=(1-3\alpha\Delta t)\pi(t)+\alpha\Delta t[1-\pi(t)] \Rightarrow \pi(t+\Delta t)-\pi(t)=\alpha\Delta t-4\alpha\Delta t\pi(t)$
 - This yields the equation: $\pi'(t)=\alpha-4\alpha\pi(t)$ $\pi(0)=1$
- The solution: $\pi(t)=1/4+(3/4)e^{-4\alpha t}$
- The probability of substitution: $s(t)=1-\pi(t)=(3/4)[1-e^{-4\alpha t}]$



In the limit all substitutions are equally likely



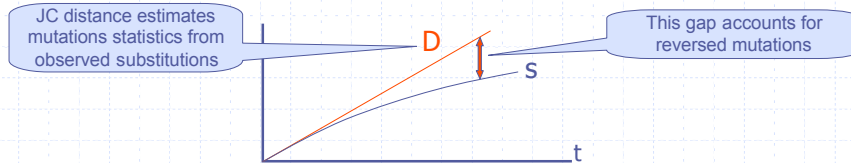
COMSW4761--2007

14

Jukes-Cantor Distance

- Consider two sequences with s sub/site: $s(X,Y) \sim (3/4)[1 - e^{-4\alpha t}]$

- The JC distance: $D(X,Y) = 3\alpha t = -(3/4)\ln[1 - (4/3)s]$



- E.g., consider X,Y below.... $s = 20/50 = 0.4$ $D(X,Y) = 0.572$

X: CCTCGACGGCTTAGATCTGATCTGACCTAATGCTGCAATCGGTCCAAAGT
 Y: CGACCACGAGTAAGAGTTGGTCCGACTTAGTCTCGGATCAAATGATAAT

- JC assumes that transversions are twice as likely as transitions
 is this correct for these sequences?

COMSW4761--2007

15

Notes

- For proteins, JC distance: $D(x,y) = -(19/20)\ln[1 - (20/19)f(x,y)]$
 - Example: two globins (human & bovine) align without indels at 145 sites differing at 23 sites, what is the JC distance?
 - $D(x,y) = -(19/20)\ln[1 - (20/19)(23/145)] \sim 17.3 \times 10^{-2}$
 - Using PAM units (multiply by 100) $D(x,y) \sim 17.3$
- JC is limited in modeling transitions/transversions variation

- Solution: Kimura (80)
- Model Transitions/Transversions at different rates

$$A = \begin{bmatrix} 1 - \alpha - 2\beta & \alpha & \beta & \beta \\ \alpha & 1 - \alpha - 2\beta & \beta & \beta \\ \beta & \beta & 1 - \alpha - 2\beta & \alpha \\ \beta & \beta & \alpha & 1 - \alpha - 2\beta \end{bmatrix}$$

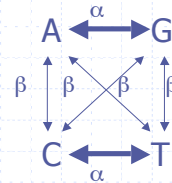
COMSW4761--2007

16

The Kimura Model (80)

- Repeat the derivations as per JC

- $p(t), q(t), r(t)$ are respectively probabilities of no change, transition, or transversion
- $p(t+\Delta t) = (1-\alpha\Delta t - 2\beta\Delta t)p(t) + (\alpha\Delta t)q(t) + (2\beta\Delta t)r(t) \dots q(t+\Delta t) = \dots r(t+\Delta t) = \dots$
- Solve linear differential equations to get: $p(t) = (1/4)[1 + e^{-4\beta t} + 2e^{-2(\alpha+\beta)t}] \dots$
- Check that for $\alpha = \beta$ Kimura reduces to JC



- Distance: $D(x,y) = 2\alpha t + 4\beta t$

- Estimate $D(x,y) = -(1/2)\ln[1-2F-G] - (1/4)\ln[1-2G]$

- $F(x,y)$ = transitions per site; $G(x,y)$ = transversions per site
 - Example: compute $D(X,Y)$ for the sequences below:
 - $F(X,Y) = 9/50 = 0.18$ $G(X,Y) = 11/50 = 0.22$ $D_{\text{Kimura}}(X,Y) = 0.579$ while $D_{\text{JC}}(X,Y) = 0.572$
- X: C C C G A C G G C T A G A T C T G A T C T G A C C T A A T G C T G C A A T C G G T C C A A A G T
Y: C G A C C A C G A G T A A G A G T T G G T C C G A C T T A G T C C T G C G A T C A A A T G A T A A T

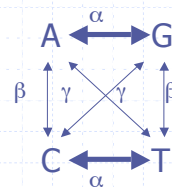
COMSW4761--2007

17

More Models

- Kimura 3-parameters

$$A = \begin{bmatrix} 1-\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & 1-\alpha-\beta-\gamma & \gamma & \beta \\ \beta & \gamma & 1-\alpha-\beta-\gamma & \alpha \\ \gamma & \beta & \alpha & 1-\alpha-\beta-\gamma \end{bmatrix}$$



- The Felsenstein model

For all symmetric models the steady state distribution is: $\pi = (.25, .25, .25, .25)$

[substitution probability of nucleotide proportional to its steady-state probability]

$$A = \begin{bmatrix} 1-u+\omega\pi_a & \omega\pi_g & \omega\pi_c & \omega\pi_t \\ \omega\pi_a & 1-u+\omega\pi_g & \omega\pi_c & \omega\pi_t \\ \omega\pi_a & \omega\pi_g & 1-u+\omega\pi_c & \omega\pi_t \\ \omega\pi_a & \omega\pi_g & \omega\pi_c & 1-u+\omega\pi_t \end{bmatrix}$$

COMSW4761--2007

18

But The Evolution Story Is More Complex

- Rates statistics has complexities beyond the Markovian models
 - Can vary among species
 - Can vary among genes
 - Can vary within a gene (e.g., coding vs. non-coding)
 -

Average Number of Substitutions Per Site			
Type	Human/Rodent	Human/Ref	Rodent/Ref
Non-degenerate	0.17	0.15	0.17
Two-fold degenerate			
Non-synonymous substitutions	0.08	0.06	0.08
Synonymous substitutions	0.32	0.18	0.31
Four-fold degenerate	0.66	0.41	0.6

WV, C.-I., and W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA 82: 1741-1745.

COMSW4761--2007

19

Concluding Notes

- Markovian models can provide useful metrics of evolution
- Thus enable construction of distance-based phylogenies
- But these distance measures are crude and must be carefully applied
- There is a need for new techniques that can reflect more accurate sequence evolution knowledge as it becomes available

COMSW4761--2007

20