Molecular Evolution

<u>Attribution</u>: The development of the logical arguments building up the Jukes and Cantor and the Kimura models, and measuring sequence divergence with those models, have been taken in large part from

<u>Fundamentals of Molecular Evolution</u>, Grauer,D and Li,W-H

Sinauer Associates, Mass 2000

Questions for Today

- What is the probability of finding any specific nucleotide at any given site?

- How many mutations have occurred since sequences diverged

- What is the rate of mutation

Bioinformatics *vis à vis* Biology

– We can use theory to estimate the number of mutations

– We need paleontological evidence to determine the passage of time

Question 1

- Examine some models of mutation
- Show how to determine ultimate probabilities of specific nucleotides appearing, given mutation rate(s)
  - Show that specific nucleotide probabilities are stationary in the limit
    - Analytic approach: By calculus to understand the underlying theory
    - Numeric Approach: By Markov Chains to process simple numerical examples
    - Linear Algebraic approach: By transition matrix eigenstruture to handle complex models and complicated matrices

RECALL

- A (purine) binds (2 H bonds) to T (pyrimidine)
- C (pyrimidine) binds (3 H bonds) to G (purine)

Possible mutations:

purine →purine   A→G, G→A (transition)

pyrimidine →pyrimidine C→T , T→C (transition)

purine → pyrimidine   A→T, A→C, G→T, G→C (transversion)

pyrimidine →purine T→A, C→A, T→G, C→G (transversion)

Note that there are twice as many transversions as transitions possible

The First Question, again:

What is the probability of a specific nucleotide in a specific position?

A Calculus Solution

Nucleotide Substitution
A Simple Model: Jukes and Cantor

Let α = the probability of substitution from one nucleotide to another in one time click. By hypothesis, whether it is a purine-purine, purine -pyrimidine, or pyrimidine-pyrimidine substitution, is not considered in this model; α is the same for all mutations in the Jukes and Cantor model.

If we choose some fixed nucleotide position in the DNA, and begin with a given nucleotide, WLOG, Adenine (A), then

- at the next time click, the probability of substituting to C,G,or T is α each, or 3 α total.

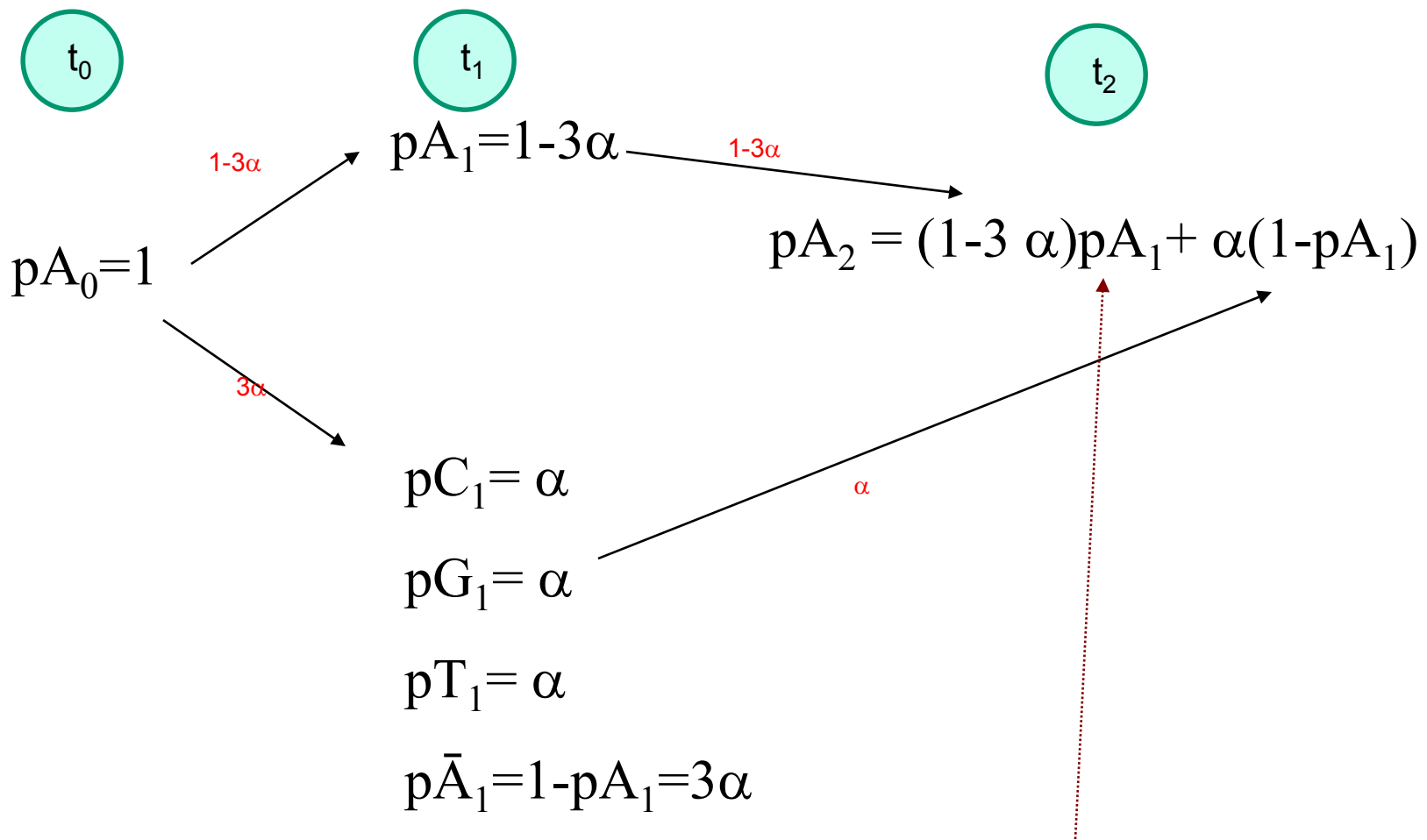- The probability of remaining in A is 1-3α.

In the time click that follows (click 2), there are 2 ways to produce A:

- A Remains as A, where it had been in click 1

OR

- C,T,G (extant in time click 1) mutates to A

t_0

t_1

t_2

$pA_0=1$

$1-3\alpha$

$pA_1=1-3\alpha$

$1-3\alpha$

$pA_2 = (1-3\alpha)pA_1 + \alpha(1-pA_1)$

3α

$pC_1 = \alpha$

α

$pG_1 = \alpha$

$pT_1 = \alpha$

$p\bar{A}_1 = 1-pA_1 = 3\alpha$

$(1-3\alpha)pA_1 = (1-3\alpha)^2$ but we leave it in the expanded form so that we can write a recursion

THE ITERATION

$$pA_{t+1} = (1 - 3\,\alpha)pA_t + \alpha(1 - pA_t)$$

Moving and re-arranging terms, we derive a difference equation

$$pA_{t+1} = (1\text{-}3\ \alpha)pA_t + \alpha(1\text{-}pA_t)$$

$$pA_{t+1} - pA_t = \text{-}3\ \alpha\ pA_t + \alpha(1\text{-}pA_t)$$

$$\Delta pA_t = \text{-}3\ \alpha\ pA_t + \alpha(1\text{-}pA_t)$$

$$\Delta pA_t = \text{-}4\ \alpha\ pA_t + \alpha$$

Making a difference equation into a differential equation

$$\Delta pA_t = -4\ \alpha\ pA_t + \alpha$$

$$\downarrow$$

$$\frac{d\ pA_t}{dt} = -4\alpha\, pA_t + \alpha$$

Solving the 1st order linear differential equation

$$pA_t = \int \left(\frac{d\ pA_t}{dt} \right) = \int \left(-4\alpha\, pA_t + \alpha \right)$$

$$pA_t = \frac{1}{4} + \left(pA_0 - \frac{1}{4} \right) e^{-4\alpha t}$$

We now have an expression for the probability of being in state A at time t

If we start in A (*i.e.* $pA_0=1$):

$$pA_t = \frac{1}{4} + \left(1 - \frac{1}{4}\right)e^{-4\alpha t} = \frac{1}{4} + \left(\frac{3}{4}\right)e^{-4\alpha t}$$

If we start in C,G,T (*i.e.* $pA_0=0$)

$$pA_t = \frac{1}{4} + \left(0 - \frac{1}{4}\right)e^{-4\alpha t} = \frac{1}{4} + \left(-\frac{1}{4}\right)e^{-4\alpha t}$$

Equilibrium

It is easy to see that the probability of ending up in state A asymptotically approaches an equilibrium probability (1/4) after a very large number of clicks. Likewise, the equilibrium probability is approached from a not-A initial condition.

$$\text{since} \quad \lim_{t \to \infty} \left[\left(\frac{3}{4} \right) e^{-4\alpha t} \right] = 0$$

$$\text{then} \quad \lim_{t \to \infty} \left[\frac{1}{4} + \left(\frac{3}{4} \right) e^{-4\alpha t} \right] = \frac{1}{4} \qquad \text{and, likewise,} \quad \lim_{t \to \infty} \left[\frac{1}{4} + \left(-\frac{1}{4} \right) e^{-4\alpha t} \right] = \frac{1}{4}$$

Stationarity Property

- In the language of stochastic processes, this asymptotically derived probability, p=0.25, is called the stationary probability.

2 key stochastic properties to be aware of are stationarity and ergodicity.

- A stationary process has the same moments regardless of which realization is sampled.
- An ergodic process has the same moments regardless of when a single realization is sampled.

A more sophisticated model
(Kimura)

We know that _transitions_

purine $\rightarrow$ purine   A$\rightarrow$G, G$\rightarrow$A

or pyrimidine $\rightarrow$ pyrimidine C$\rightarrow$T , T$\rightarrow$C

are more thermodynamically favorable (hence more probable) than _transversions_

purine $\rightarrow$ pyrimidine   A$\rightarrow$T, A$\rightarrow$C, G$\rightarrow$T, G$\rightarrow$C,

or pyrimidine $\rightarrow$ purine T$\rightarrow$A, C$\rightarrow$A, T$\rightarrow$G, C$\rightarrow$G

The Kimura model captures that notion by assigning separate probabilities to each

Kimura

- In the Kimura model, we will have a mutational rate α for a transition

- There will be a separate transversion rate β

where presumably $\alpha > \beta$

and where $\alpha + 2\beta$ <u>must be</u> < 1 (mathematical requirement)

and where $\alpha + 2\beta$ <u>should be</u> <<< 1 (evolutionary fact)
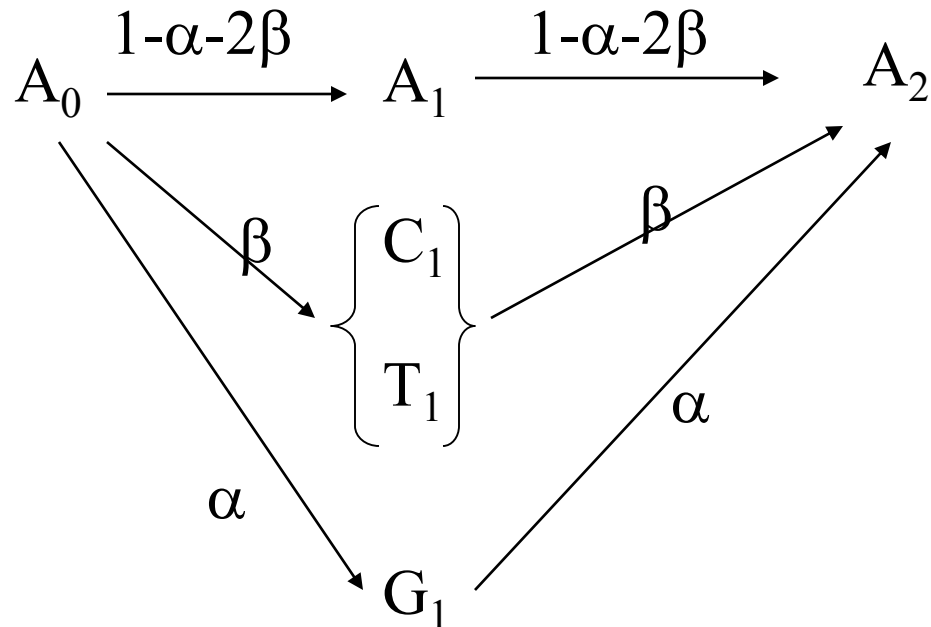
$$pAA_1 = 1 - \alpha - 2\beta$$

The probability of remaining in state A at click 2 is the sum of all the possible journeys

$$pAA_2 = (1 - \alpha - 2\beta)pAA_1$$
$$+ \beta pTA_1$$
$$+ \beta pCA_1$$
$$+ \alpha pGA_1$$

Again, generalizing

$$pAA_{t+1} = (1 - \alpha - 2\beta)pAA_t + \beta pTA_t + \beta pCA_t + \alpha pGA_t$$

Then moving thru the difference equation to the DE

$$\frac{d\ pAA_{t+1}}{dt} = -(\alpha + 2\beta)pAA_t + \beta pTA_t + \beta pCA_t + \alpha pGA_t$$

Solution

$$pAA_t = \frac{1}{4} + \frac{1}{4}e^{-\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}$$

Equilibrium

Just as in the Jukes-Cantor model, in the limit, this much more complicated Kimura model is also stationary with asymptotic probability of 0.25

Another Viewpoint: A Numeric Solution

We can show this same convergence to stationary probabilities by:

- Creating a Transition Matrix (a Markov Matrix)

- Evolve the Transition Matrix to a vector of stationary probabilities using matrix arithmetic

DEFINITIONS

Markov property

*A next-state stochastic process variable, dependent only on the current state, with the property of "forgetting" all states before, has the **Markov property.***

Andrei Markov
1856-1922

Markov process

A <u>continuous</u> stochastic process with the Markov property is called a Markov process. The probability of change from one state to the subsequent state is governed by a Markov propagator.

Markov chain

*A <u>discrete</u> process with the Markov property is called a Markov chain. The probability of change from one state to the subsequent state is called the **transition probability.***

Equilibrium

- We can represent nucleotide change probabilities as transition probabilities in a Markov chain
- Specifically, we write a transition matrix for the Jukes and Cantor model, with each element representing a transition probability from one nucleotide to the next

Properties of Markov Chains

- Absorbing: Possible to get trapped

- Non-absorbing

  - Ergodic (irreducible) Visits everywhere at least once (′εργον ‛οδοσ)

  - Regular: There is some power at which all transitions are (positive) nonzero

Every regular chain is ergodic, but not *vice versa*

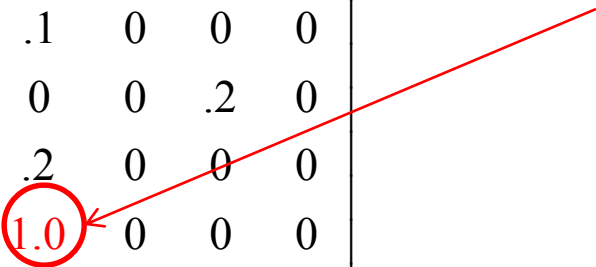# Markov chain in state A with transition probabilities
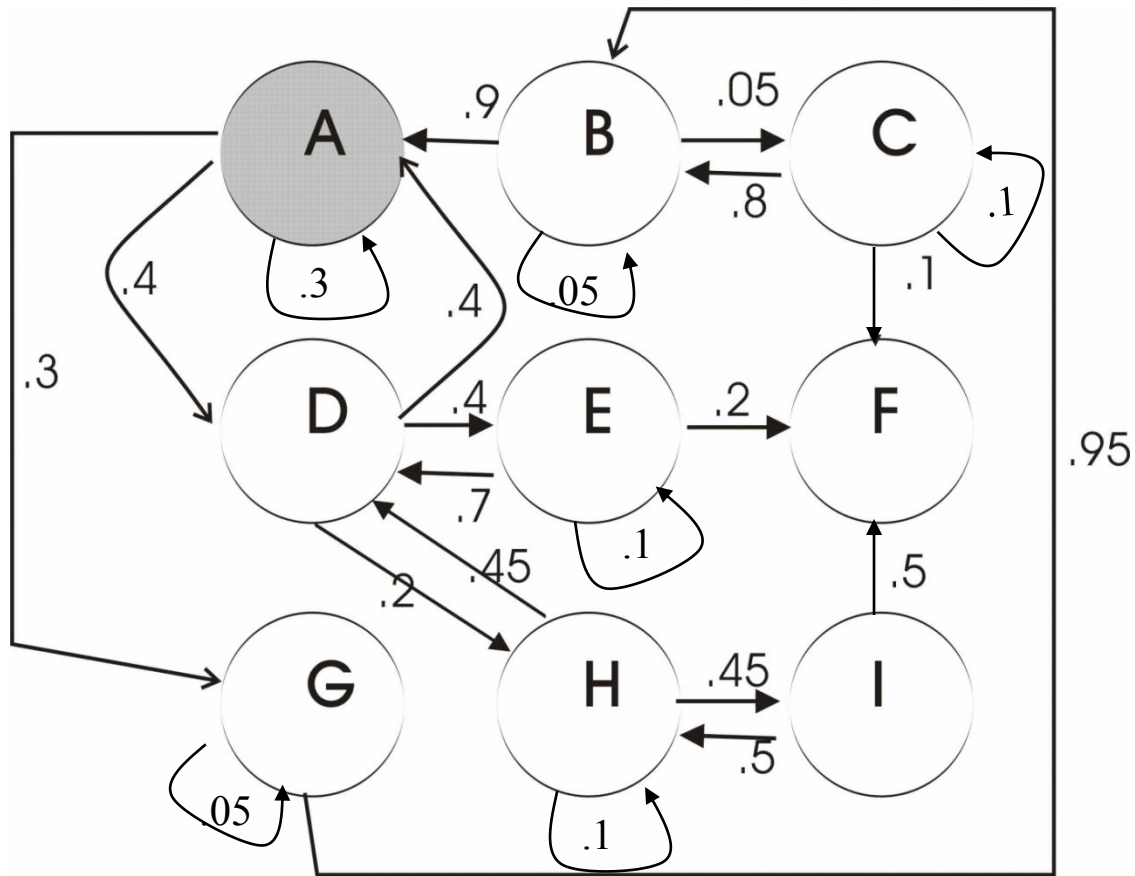
Note that all transition probabilities out of state F are nil.

This is an absorbing Markov chain; transition out of <u>any</u> state will ultimately reach F and become absorbed

Here is the transition matrix for the Markov Chain illustrated in the previous slide

	A	B	C	D	E	F	G	H	I
A	.3	0	0	.4	0	0	.3	0	0
B	.9	.05	.05	0	0	0	0	0	0
C	0	.8	.1	0	0	.1	0	0	0
D	.4	0	0	0	.4	0	0	.2	0
E	0	0	0	.7	.1	.2	0	0	0
F	0	0	0	0	0	1.0	0	0	0
G	0	.95	0	0	0	0	.05	0	0
H	0	0	0	.45	0	0	0	.1	.45
I	0	0	0	0	0	.5	0	.5	0

Note the absorbing state

Imagine one click of arbitrary time. We note the transition probabilities. In the next click, we operate on the state that is current, oblivious to any previous states (Markov property), applying the transition probabilities that exist in this moment.

 In other words, we multiply the current transition matrix by itself to get the 'new' transition probabilities. (Keep in mind that this multiplication is matrix, multiplication, not element-by-element ordinary multiplication.)

After only 20 time clicks, one can begin to observe the probability of staying
in state F (transition F →F) approaching 1.0

	A	B	C	D	E	F	G	H	I
A	0.1448	0.0528	0.0031	0.1190	0.0556	0.5207	0.0478	0.0383	0.0180
B	0.1513	0.0552	0.0032	0.1245	0.0580	0.4989	0.0500	0.0399	0.0188
C	0.1412	0.0516	0.0030	0.1160	0.0542	0.5325	0.0467	0.0373	0.0175
D	0.1190	0.0434	0.0025	0.0981	0.0456	0.6057	0.0394	0.0314	0.0149
E	0.0972	0.0355	0.0021	0.0798	0.0374	0.6781	0.0321	0.0258	0.0120
F	0	0	0	0	0	1.0000	0	0	0
G	0.1585	0.0579	0.0034	0.1302	0.0608	0.4754	0.0524	0.0419	0.0196
H	0.0861	0.0314	0.0018	0.0707	0.0332	0.7149	0.0284	0.0229	0.0107
I	0.0449	0.0164	0.0010	0.0371	0.0172	0.8511	0.0149	0.0119	0.0056

After 100 time clicks, a steady state is reached, and the process is nearly all absorbed into
state F

	A	B	C	D	E	F	G	H	I
A	0.0045	0.0016	0.0001	0.0037	0.0017	0.9852	0.0015	0.0012	0.0006
B	0.0047	0.0017	0.0001	0.0038	0.0018	0.9846	0.0015	0.0012	0.0006
C	0.0043	0.0016	0.0001	0.0036	0.0017	0.9856	0.0014	0.0011	0.0005
D	0.0037	0.0013	0.0001	0.0030	0.0014	0.9879	0.0012	0.0010	0.0005
E	0.0030	0.0011	0.0001	0.0025	0.0011	0.9901	0.0010	0.0008	0.0004
F	0	0	0	0	0	1.0000	0	0	0
G	0.0049	0.0018	0.0001	0.0040	0.0019	0.9838	0.0016	0.0013	0.0006
H	0.0026	0.0010	0.0001	0.0022	0.0010	0.9912	0.0009	0.0007	0.0003
I	0.0014	0.0005	0.0000	0.0011	0.0005	0.9954	0.0005	0.0004	0.0002

The Probability Matrix

- We have given an example, and if you look closely, note that the row values add up to 1.0, as would be the case for a Markov Transition Matrix:

  – Row entries sum to 1.0

- Additionally, for our evolutionary task, we require

  – Non-absorbing

  – Ergodic

  – Regular

Frequently called a 'Stochastic Matrix'

Suppose we allowed transition out of F to C, E, or I, each with probability 0.01.
Then the chain {A, B, C, D, E, F, G, H, I} is **no longer absorbing**. Further, it is **ergodic**
(<u>irreducible</u>) because every element is eventually reachable. Yet further, the underlying
Markov Chain is **regular**, since there is a power where all elements are positive (non-negative
and non-zero) , although this is not obvious at this power of 1).

Here is the transition matrix:

$$R^1 =$$

	A	B	C	D	E	F	G	H	I
A	.3	0	0	.4	0	0	.3	0	0
B	.90	.05	.05	0	0	0	0	0	0
C	0	.8	.1	0	0	.1	0	0	0
D	.4	0	0	0	.4	0	0	.2	0
E	0	0	0	.7	.1	.2	0	0	0
F	0	0	.01	0	.01	.97	0	0	.01
G	0	.95	0	0	0	0	.05	0	0
H	0	0	0	.45	0	0	0	.1	.45
I	0	0	0	0	0	.5	0	.5	0

Markov Transition Matrix

Note that this matrix mimics the transition rules of a Non-Deterministic Finite Automaton

- The Markov property is key

- The requirement for regularity is removed from the NDFA (although it would be bizarre machine design to have non-positive transition probabilities in an NDFA)

- The requirement for ergodicity is also removed in a NDFA but required for the Markov Transition Matrix

Fundamental Limit Theorem for Regular Markov Chains

If P is a regular transition matrix then

$$\lim_{n \to \infty} P^n = U$$

where U is a probability matrix with each row a fixed probability vector and all rows equal

Fundamental Limit Theorem for Markov Chains at Work

Let's put the theorem to work!

matrix. Since it is already

- Ergodic
- Regular
- non absorbing

just exponentiate it (with matrix multiplication, of course)

The corresponding transition matrix is **converging** to the stationary row vector.

After 157 powers it is stationary up to >4 decimal places.

$$R^{157} = \begin{array}{c} \\ \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \end{array} \begin{array}{ccccccccc} A & B & C & D & E & F & G & H & I \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \\ .0980 & .0383 & .0094 & .0823 & .0438 & .6508 & .0304 & .0292 & .0197 \end{array}$$

NOTICE THAT THE STEADY STATE ROWS ARE IDENTICAL AND THE COLUMNS ARE PROBABILITIES. In looking back at the graph, it is not surprising that, in the limit, A is more likely to reach F than any other state, as is borne out by the steady state probability of A→F.

Stationary Probability: An Algebraic Approach to the Eigenstructure of the Transition Matrix

The Perron-Frobenius Theorem motivates the following development:

The stationary transition probabilities are given by the probability eigenvector corresponding to the eigenvalue of 1

Linear Algebra Mini Review

Let A be an n × n matrix.

λ *is an eigenvalue of A if* $\exists$ *non − zero vector X in* $\mathfrak{R}^n$ *such that*

$$AX = \lambda X$$

If X is a non-zero vector satisfying the above, X is an **eigenvector**. There can be more than one solution; there are as many solutions (eigenvectors) as the order (n) of the matrix, although the solutions may not be distinct.

A and B are linear transformation matrices

$BX_2 = \lambda X_2$

X_2

AX_1

X_1
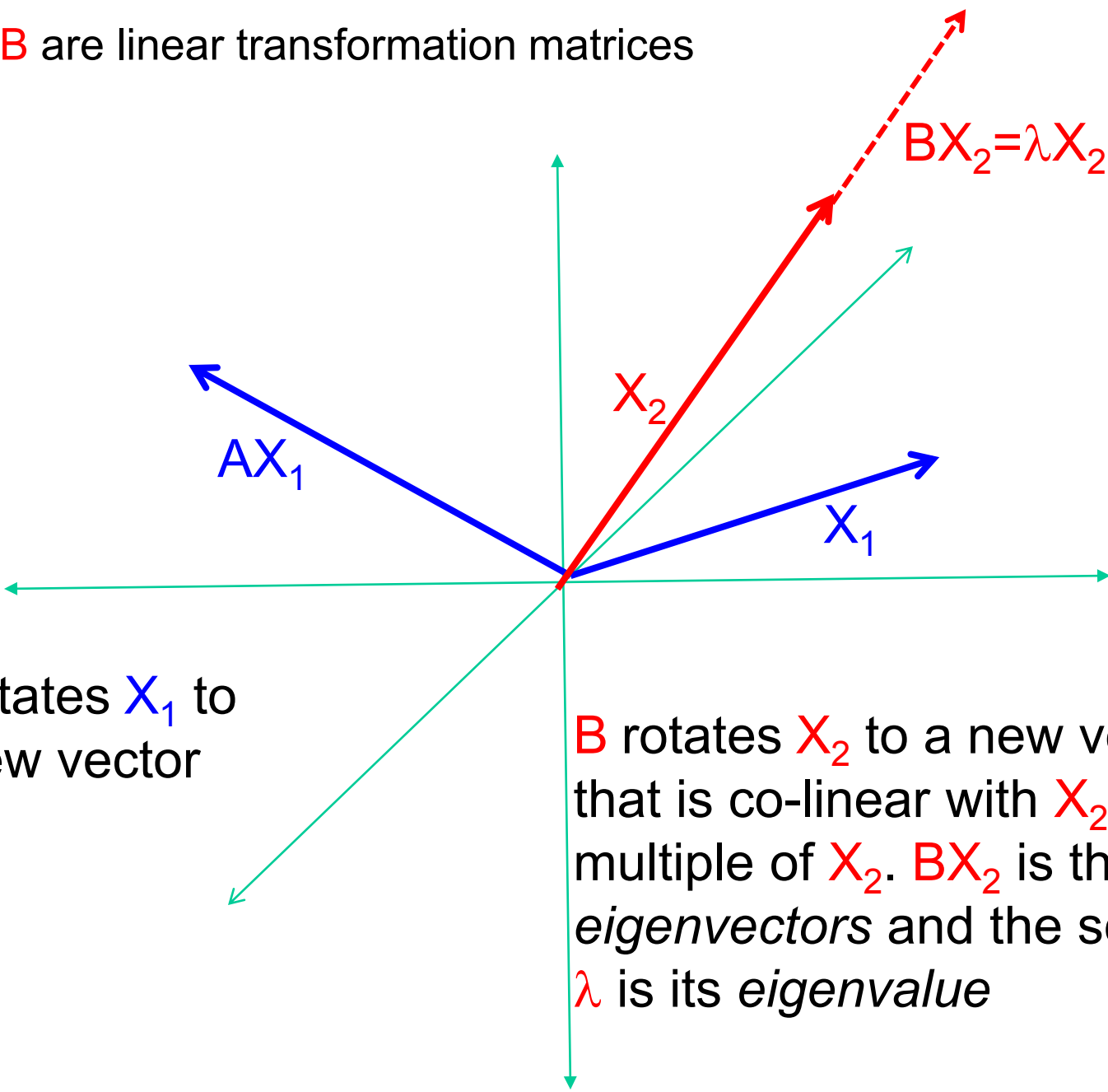
A rotates X_1 to a new vector AX_1

B rotates X_2 to a new vector BX_2 that is co-linear with X_2; it is a scalar multiple of X_2. BX_2 is the one of the *eigenvectors* and the scalar multiple λ is its *eigenvalue*

When you do this rotation, you need to know two things about where you end up:
1. What vectors emerge from the rotation (the eigenvectors)

2. What scale factors emerge from the rotation (the eigenvalues)

Given **A**X=λX, which represents a linear system of equations.

Then, re-writing this:
AX-λ**I**X=0 where **I** is the identity matrix

For this linear system to have a non-trivial solution, its determinant must be zero.

So, $\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = 0$ for a system of rank 2

Solving, we get $(a_{11} - \lambda)(a_{22} - \lambda) - (a_{12})(a_{21}) = 0$
resulting in a quadratic (in a matrix of rank 2) equation in λ. This produces two
values for λ in a system of rank 2, λ_1 and λ_2.

These λ are the characteristic roots, or *eigenvalues*, of the system. There will be an
eigenvector corresponding to each eigenvalue

Linear Algebra and Markov Transition Matrices

Review: There are three prerequisites in order to make classify a transition matrix as a probability matrix and to make it consistent with our biology

- The rows of the transition matrix are <u>probability vectors</u> (column entries in each row sum to 1.0)
- The Matrix is <u>regular</u> (there is at least one power of the matrix which has only positive entries
- The chain is ergodic (<u>irreducible)</u> and <u>aperiodic</u>

Back to the Question:

What is the probability of a specific nucleotide at any given site?

Stationary Probability

If V is the steady state vector, then it stands to reason that **MV**=**V** where **M** is the transition matrix

This is the same thing as **MV**=λ**V** where λ=1

Then it further stands to reason that the steady state vector of **M** is that specific eigenvector **V** corresponding to the eigenvalue λ=1 that is a probability vector (remember that, while there is an infinitude of parallel eigenvectors corresponding to λ=1, only one of them is a probability vector)

The Jukes-Cantor Model of Nucleotide Substitution Probabiities

• Assumes that all nucleotides are equally likely to substitute for each other

• α is a probability reflecting how often a nucleotide substitutes in 1 generation

$$
\begin{array}{cccc}
 & A & C & G & T \\
A & 1-3\alpha & \alpha & \alpha & \alpha \\
C & \alpha & 1-3\alpha & \alpha & \alpha \\
G & \alpha & \alpha & 1-3\alpha & \alpha \\
T & \alpha & \alpha & \alpha & 1-3\alpha
\end{array}
$$

Jukes-Cantor Model

The eigenvalues are $1, (1-4\alpha), (1-4\alpha), (1-4\alpha)$

The corresponding eigenvectors are

$$\begin{pmatrix} r \\ r \\ r \\ r \end{pmatrix}, \begin{pmatrix} -r \\ 0 \\ 0 \\ r \end{pmatrix}, \begin{pmatrix} -r \\ 0 \\ r \\ 0 \end{pmatrix}, \begin{pmatrix} -r \\ r \\ 0 \\ 0 \end{pmatrix}$$

Choose the eigenvector corresponding to the eigenvalue 1

$$\begin{pmatrix} r \\ r \\ r \\ r \end{pmatrix}$$

Jukes-Cantor Model

There are an infinitude of these vectors. But because this is a probability vector, the coefficients must sum to 1. Because the elements are equal, the coefficients must be .25 each.

Thus, the specific eigenvector we require must be:

$$\begin{pmatrix} .25 \\ .25 \\ .25 \\ .25 \end{pmatrix}$$

IMPLICATION:

When our Jukes_Cantor model reaches <u>equilibrium</u>, the probabilities of transitions from each of the nucleotides are <u>equal.</u>

More Complicated Models Kimura

transitions (purine→purine, pyrimidine →pyrimidine)
[A→G , C→T   with frequency α]

   more likely than

transversions (purine→ pyrimidine, pyrimidine → purine)

[A→C , G→T with frequency β]

NOTE simpler versions assume
freq A →G = freq G →A *etc*

Kimura Transition Matrix

$$
\begin{array}{c c c c c}
 & A & C & G & T \\
A & \left(\begin{array}{cccc} 1-\alpha-2\beta & \beta & \alpha & \beta \\
C & \beta & 1-\alpha-2\beta & \beta & \alpha \\
G & \alpha & \beta & 1-\alpha-2\beta & \beta \\
T & \beta & \alpha & \beta & 1-\alpha-2\beta \end{array}\right)
\end{array}
$$

Here, again, the eigenvector corresponding to the eigenvalue of 1 would represent the stationary transition probabilities. It is left as an exercise to evaluate the stationary transition probabilities in this model.