Makov Chains

Juliana Silva Bernardes

Markov Chains

- **Formal definition**
  - A Markov chain is a sequence of random variables $X_1, X_2, X_3, \ldots$ with the Markov property.

    $P(X_n \mid X_0, X_1, \ldots, X_{n-1}) = P(X_n \mid X_{n-1})$ there is no memory in a Markov process.

  - The possible values of X_i form a countable set S called the **state space** of the chain.
  - $P(X_n \mid X_{n-1})$ values are obtained from P, the **Transition Matrix**

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,j} & \cdots \\ p_{2,1} & p_{2,2} & \cdots & p_{2,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix}.$$

$$\sum_i P_{ij} = 1$$

The probability of transitioning from i to j in two steps is then given by the $(i,j)^{th}$ element of the square of P

$$(P^2)_{ij}$$

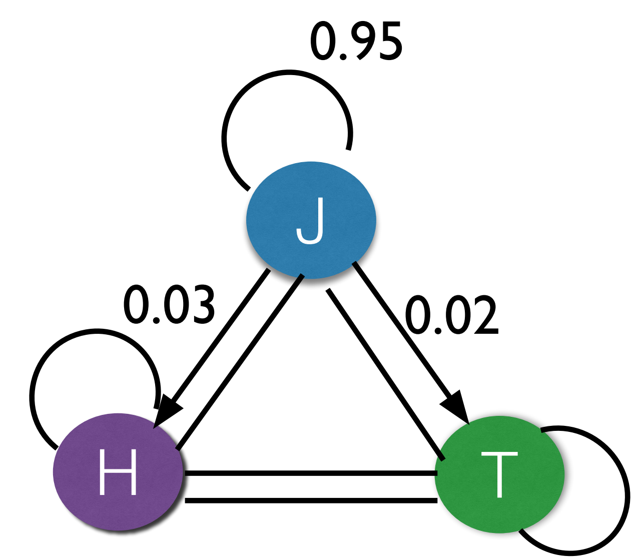
Markov Chains

- Example: Velib
  - Take a bike in a station **i** and give back in a station **j**,

stations:

- Jussieu (J)
- Hotel de ville (H)
- Tour Eiffel (T)

S = {J, H, T}

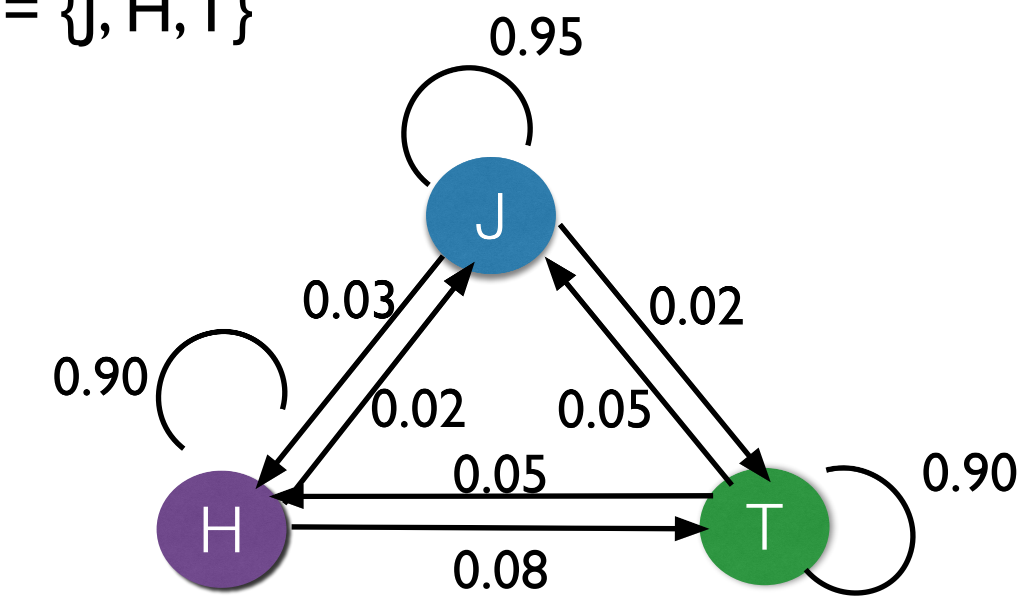


Markov Chains

- Example: Velib
  - Take a bike in a station **i** and give back in a station **j**,

stations:
- Jussieu (J)
- Hotel de ville (H)
- Tour Eiffel (T)

S = {J, H, T}



Give back

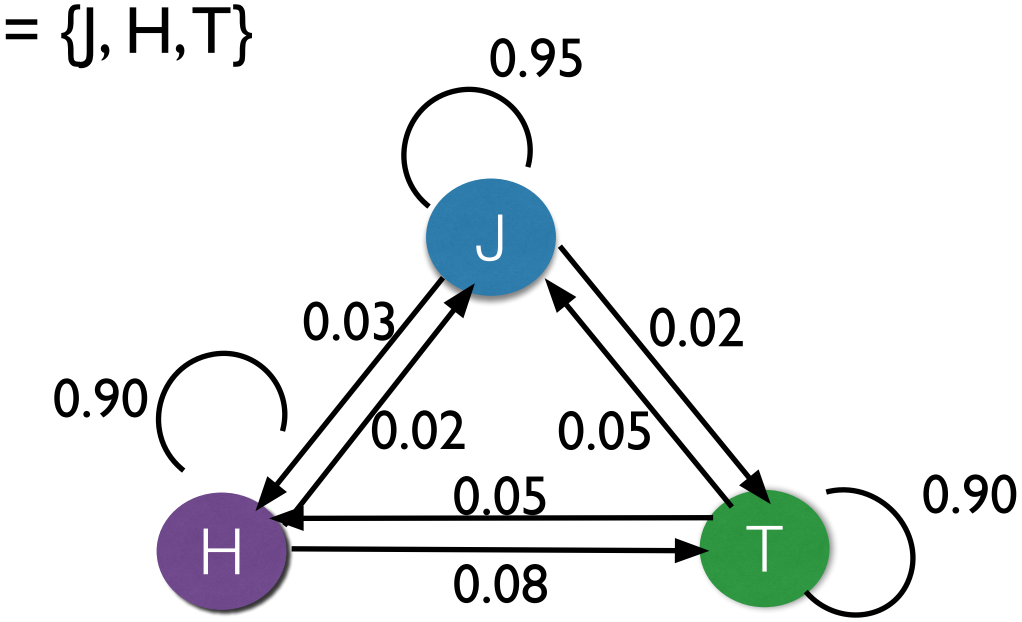
Take		J	H	T
	J	0.95	0.03	0.02
	H	0.02	0.90	0.08
	T	0.05	0.05	0.90

Markov Chains

- Example: Velib
  - Take a bike in a station **i** and give back in a station **j**,

stations:
- Jussieu (J)
- Hotel de ville (H)
- Tour Eiffel (T)

S = {J, H, T}



Give back

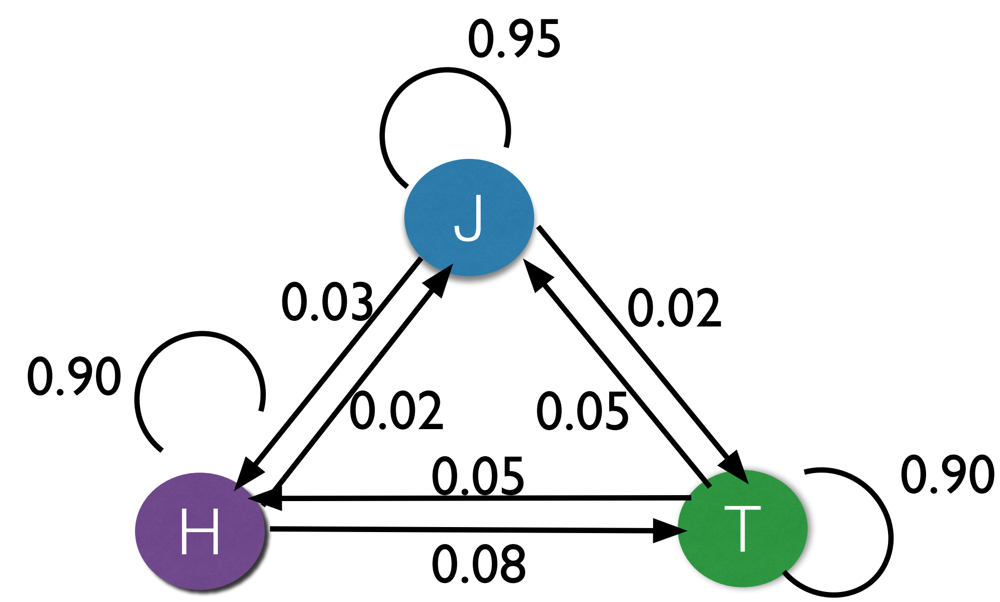
Take		J	H	T	
	J	0.95	0.03	0.02	=1
	H	0.02	0.90	0.08	=1
	T	0.05	0.05	0.90	=1

Transition Matrix P

Markov Chains

- Example: Velib
  - Take a bike in a station i and give back in a station j, stations $S = \{J, H, T\}$

		Give back		
		J	H	T
	J	0.95	0.03	0.02
Take	H	0.02	0.90	0.08
	T	0.05	0.05	0.90



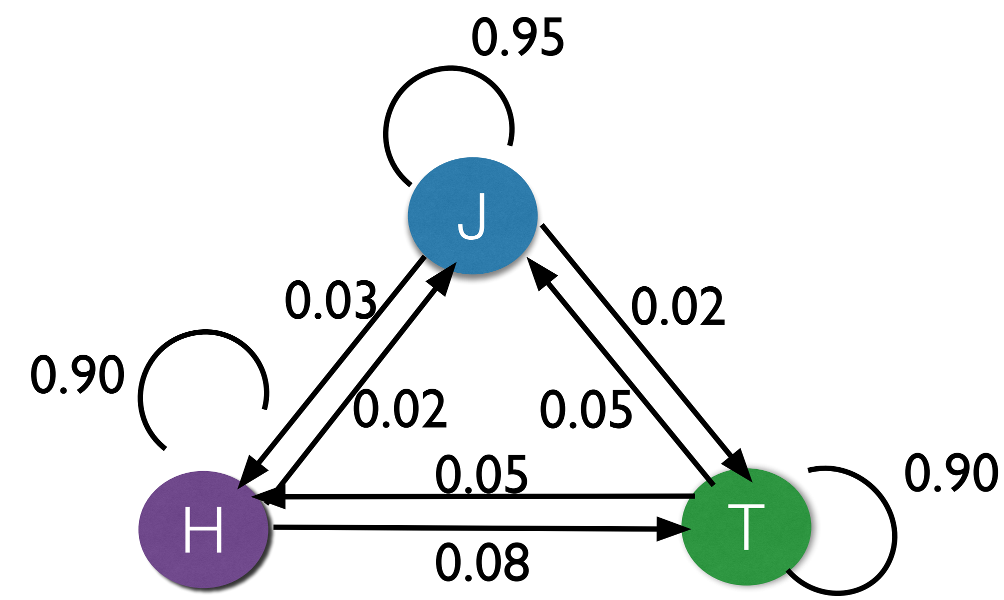
$$\pi_0 \begin{array}{c} \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} \end{array} \begin{array}{c} J \\ H \\ T \end{array} \qquad \pi_1 = P\pi_0 \quad = 0.5 \begin{bmatrix} 0.95 \\ 0.03 \\ 0.02 \end{bmatrix} + 0.3 \begin{bmatrix} 0.02 \\ 0.90 \\ 0.08 \end{bmatrix} + 0.2 \begin{bmatrix} 0.05 \\ 0.05 \\ 0.90 \end{bmatrix} = \begin{bmatrix} 0.491 \\ 0.295 \\ 0.214 \end{bmatrix} \begin{array}{c} J \\ H \\ T \end{array}$$

Initial distribution

Markov Chains

- Jussieu (J)
- Hotel de ville (H)
- Tour Eiffel (T)

		Give back		
		J	H	T
Take	J	0.95	0.03	0.02
	H	0.02	0.90	0.08
	T	0.05	0.05	0.90



$$\pi_0 \; \begin{array}{cc} \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} & \begin{array}{c} J \\ H \\ T \end{array} \end{array}$$

$$\pi_1 = P\pi_0 = 0.5 \begin{bmatrix} 0.95 \\ 0.03 \\ 0.02 \end{bmatrix} + 0.3 \begin{bmatrix} 0.02 \\ 0.90 \\ 0.08 \end{bmatrix} + 0.2 \begin{bmatrix} 0.05 \\ 0.05 \\ 0.90 \end{bmatrix} = \begin{array}{cc} \begin{bmatrix} 0.491 \\ 0.295 \\ 0.214 \end{bmatrix} & \begin{array}{c} J \\ H \\ T \end{array} \end{array}$$

$$\pi_2 = P\pi_1 = PP\pi_0 = P^2\pi_0$$

$$\pi_k = P\pi_{k-1} \quad \text{ou} \quad \boxed{\pi_k = P^k\pi_0}$$

$$\pi_{49} \begin{bmatrix} 0.417 \\ 0.278 \\ 0.305 \end{bmatrix} = \pi_{50} = \pi_{51}$$

$$\pi_{49} \longrightarrow \pi^*$$

Stationary probability vector

Stationary Probabilistic Vector

- Stationary probability vector π^*

  is defined as a vector that does not change under application of the transition matrix;

$$P\pi^* = \pi^*$$

It is defined as the eigenvector of the probability matrix, associated with eigenvalue 1:

How to compute π^* in a way that $P\pi^*=\pi^*$?

How to compute π^* in a way that $P\pi^*=\pi^*$?

- Let u and v be two vectors
- They are scalar multiples of each other, also parallel or collinear, if there is a scalar λ, such that

$v = \lambda u$ v is the eigenvector of u

λ is the eigenvalue of u

Example: $\mathbf{u} = \begin{Bmatrix} 1 \\ 3 \\ 4 \end{Bmatrix}$ and $\mathbf{v} = \begin{Bmatrix} -20 \\ -60 \\ -80 \end{Bmatrix}$. $\mathbf{u} = \lambda \mathbf{v}$.

In this case $\lambda = -1/20$.

- Now consider the linear transformation of n-dimensional vectors defined by an $n \times n$ matrix A, that is, Av=w
- If it occurs that w and v are scalar multiples then

$$Av=\lambda w$$

$A=P$ $v=w=\pi^*$ $\lambda=1$

$$P\pi^* = \pi^*$$

π^* is the eigenvector of the probability matrix P, associated with eigenvalue 1

Markov Chains

- Determining the transition matrix P^n after n iterations to find π^n and π^*

$$\boxed{\pi_k = P^k \pi_0}$$

$$\begin{array}{c} \\ J \\ H \\ T \end{array} \begin{array}{c} \pi \\ \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} \end{array}$$

n = 2

$\pi_2 = P^2 \pi_0$

$$P^2 = \begin{array}{c|ccc} & J & H & T \\ \hline J & 0.9041 & 0.0410 & 0.0935 \\ H & 0.0565 & 0.8146 & 0.0915 \\ T & 0.0394 & 0.1444 & 0.8150 \end{array}$$

π_2

J 0.48305
H 0.29093
T 0.22602

n = 3

$\pi_3 = P^3 \pi_0$

$$P^3 = \begin{array}{c|ccc} & J & H & T \\ \hline J & 0.861995 & 0.062462 & 0.131405 \\ H & 0.079943 & 0.741590 & 0.125905 \\ T & 0.058062 & 0.195948 & 0.742690 \end{array}$$

π_3

J 0.4760171
H 0.2876295
T 0.2363534

Markov Chains

- Determining the transition matrix P^n after n iterations to find π^n and π^*

	0	1	2	3	...	48	49	50
J	0.5	0.491	0.48305	0.4760171	...	0.4173411	0.41728	0.4172245
H	0.3	0.295	0.29093	0.2876295	...	0.2775646	0.2775831	0.2776
T	0.2	0.214	0.22602	0.2363534	...	0.3050944	0.3051369	0.3051755

$$\boxed{\pi_k = P^k \pi_0}$$

n = 49 $\quad \pi_{49} = P^{49}\pi_0$

$$P^{49} = \begin{array}{c|ccc} & J & H & T \\ \hline J & 0.4227516 & 0.4101549 & 0.4142888 \\ H & 0.2757332 & 0.2800620 & 0.2784892 \\ T & 0.3015152 & 0.3097831 & 0.3072220 \end{array}$$

n = 50 $\quad \pi_{50} = P^{50}\pi_0$

$$P^{50} = \begin{array}{c|ccc} & J & H & T \\ \hline J & 0.4222044 & 0.4107375 & 0.4145053 \\ H & 0.2759182 & 0.2798496 & 0.2784300 \\ T & 0.3018773 & 0.3094128 & 0.3070647 \end{array}$$

π_{49}

J 0.417
H 0.278
T 0.305

π_{50}

J 0.417
H 0.278
T 0.305

π^*

Predicting trajectories

$$\pi_0 = \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix}$$

Give back

		J	H	T
P= Take	J	0.95	0.03	0.02
	H	0.02	0.90	0.08
	T	0.05	0.05	0.90

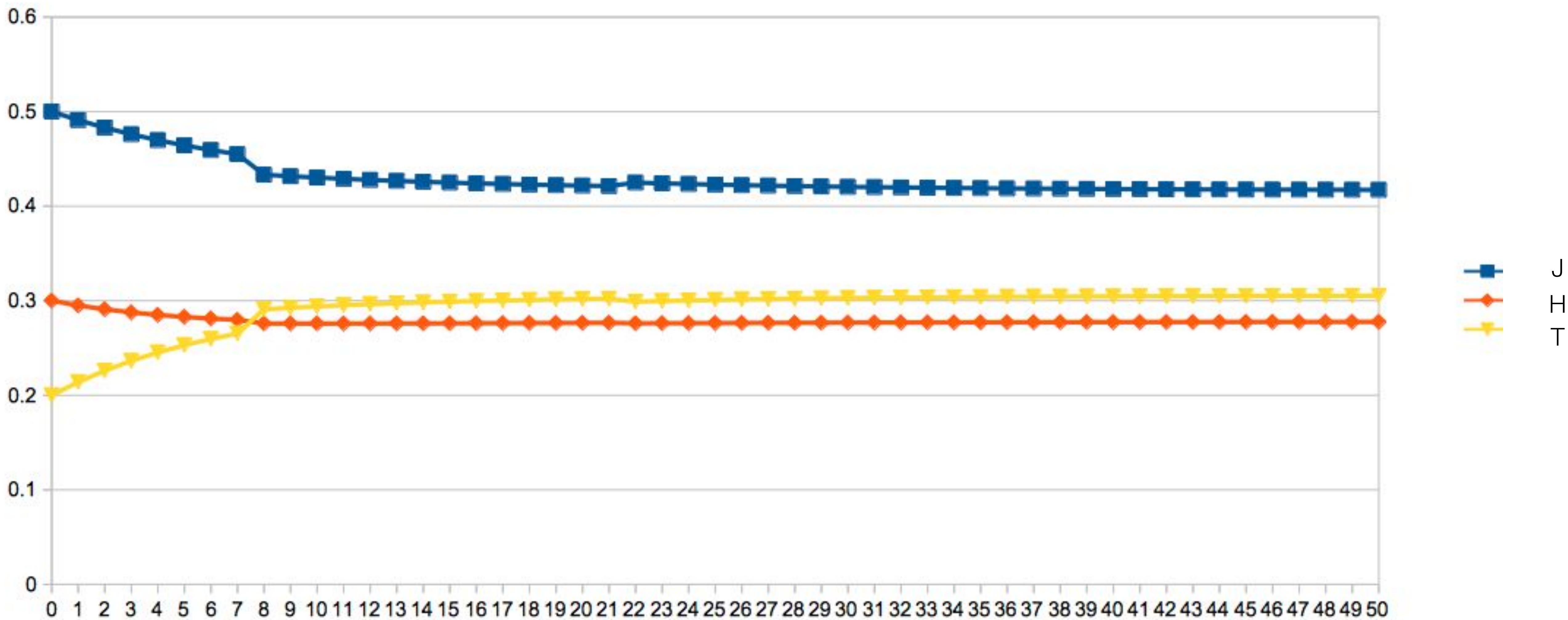
$$\boxed{\pi_k = P\pi_{k-1}}$$

$$\pi_1 = P\pi_0 \qquad \pi_1 = \begin{bmatrix} 0.491 \\ 0.295 \\ 0.214 \end{bmatrix}$$

$$\pi_2 = P\pi_1 \qquad \pi_2 = \begin{bmatrix} 0.483 \\ 0.290 \\ 0.226 \end{bmatrix}$$

	0	1	2	3	...	48	49	50
J	0.5	0.491	0.48305	0.4760171	...	0.4173411	0.41728	0.4172245
H	0.3	0.295	0.29093	0.2876295	...	0.2775646	0.2775831	0.2776
T	0.2	0.214	0.22602	0.2363534	...	0.3050944	0.3051369	0.3051755

Plotting trajectories



Predicting trajectories

Now let's change π_0

$$\pi_0 = \begin{bmatrix} 0.1 \\ 0.6 \\ 0.3 \end{bmatrix}$$

$$\boxed{\pi_k = P\pi_{k-1}}$$

Give back

$$P = \begin{array}{c|ccc} & J & H & T \\ \hline J & 0.95 & 0.03 & 0.02 \\ H & 0.02 & 0.90 & 0.08 \\ T & 0.05 & 0.05 & 0.90 \end{array}$$

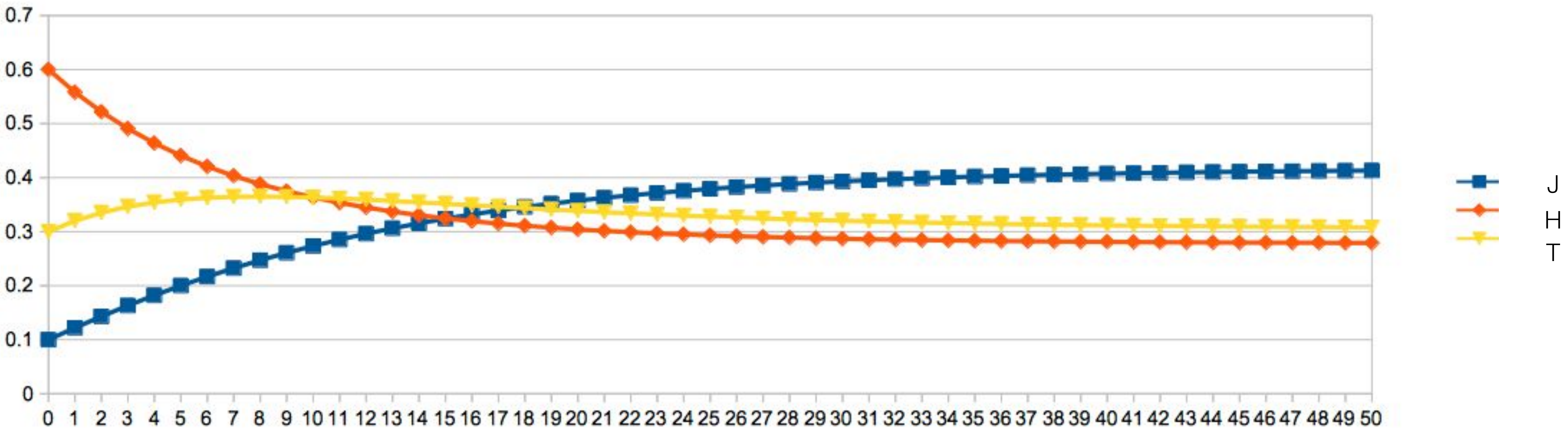
Take

$$\pi_1 = P\pi_0 \qquad \pi_1 = \begin{bmatrix} 0.122 \\ 0.558 \\ 0.32 \end{bmatrix}$$

$$\pi_2 = P\pi_1 \qquad \pi_2 = \begin{bmatrix} 0.143 \\ 0.521 \\ 0.335 \end{bmatrix}$$

0	1	2	3	4	...	48	49	50
0.1	0.122	0.14306	0.1630982	0.1820668	...	0.4122596	0.4126547	0.4130145
0.6	0.558	0.52186	0.4907198	0.4638499	...	0.2792971	0.2791573	0.2790306
0.3	0.32	0.33508	0.346182	0.3540833	...	0.3084433	0.308188	0.3079548

Plotting trajectories



	0	1	2	3	...	48	49	50
J	0.5	0.491	0.48305	0.4760171	...	0.4173411	0.41728	0.4172245
H	0.3	0.295	0.29093	0.2876295	...	0.2775646	0.2775831	0.2776
T	0.2	0.214	0.22602	0.2363534	...	0.3050944	0.3051369	0.3051755

0	1	2	3	4	...	48	49	50
0.1	0.122	0.14306	0.1630982	0.1820668	...	0.4122596	0.4126547	0.4130145
0.6	0.558	0.52186	0.4907198	0.4638499	...	0.2792971	0.2791573	0.2790306
0.3	0.32	0.33508	0.346182	0.3540833	...	0.3084433	0.308188	0.3079548

Simulating trajectories

$X_0, X_1, X_2, X_3, \ldots, X_n$

$\downarrow$

π_0

Give back

		J	H	T
Take	$P_1=$ J	0.95	0.03	0.02
	$P_2=$ H	0.02	0.90	0.08
	$P_3=$ T	0.05	0.05	0.90

$$\pi_0 = \begin{array}{c|c} \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} & \begin{array}{c} J \\ H \\ T \end{array} \end{array}$$

We take a random number y_0 in [0,1]

if $y_0 < 0.2$ then $X_0 = T$

elsif $y_0 >= 0.2$ and $y_0 < 0.5$ then $X_0 = H$

else $X_0 = J$

$\boxed{y_0 = 0.7}$ $\boxed{X_0 = J}$

Simulating trajectories

J

$X_0, X_1, X_2, X_3, \ldots, X_n$

$\downarrow$

P_1

Give back

		J	H	T
$P1=$	J	0.95	0.03	0.02
$P_2=$	H	0.02	0.90	0.08
$P_3=$	T	0.05	0.05	0.90

Take

We take a random number y_1 in [0,1]

if $y_1 < 0.02$ then $X_1 = T$

elsif $y_1 >= 0.02$ and $y_1 < 0.05$ then $X_1 = H$

else $X_1 = J$

$y_1 = 0.4$

$X_1 = J$

$$\pi_0 = \begin{array}{c} \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} \end{array} \begin{array}{c} J \\ H \\ T \end{array}$$

Simulating trajectories

J J

$X_0, X_1, X_2, X_3, \ldots, X_n$

$\downarrow$

P_1

Give back

Take

		J	H	T
$P_1=$	J	0.95	0.03	0.02
$P_2=$	H	0.02	0.90	0.08
$P_3=$	T	0.05	0.05	0.90

$$\pi_0 = \begin{array}{cc} \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} & \begin{array}{c} J \\ H \\ T \end{array} \end{array}$$

We take a random number y_2 in [0,1]

if $y_2 < 0.02$ then $X_2 = T$

elsif $y_2 >= 0.02$ and $y_2 < 0.05$ then $X_2 = H$

else $X_2 = J$

$y_2 = 0.046$

$X_2 = H$

Simulating trajectories

J J H

$X_0, X_1, X_2, X_3, \ldots, X_n$

$\downarrow$

P_2

Give back

Take

		J	H	T
$P1=$	J	0.95	0.03	0.02
$P_2=$	H	$\boxed{0.02\ \ 0.90\ \ 0.08}$		
$P_3=$	T	0.05	0.05	0.90

$$\pi_0 = \begin{array}{c|c} \begin{bmatrix} 0.5 \\ 0.3 \\ 0.2 \end{bmatrix} & \begin{array}{c} J \\ H \\ T \end{array} \end{array}$$

We take a random number y_3 in [0,1]

if $y_3 < 0.02$ then $X_3 = J$

elsif $y_3 >= 0.02$ and $y_3 < 0.1$ then $X_3 = H$

else $X_3 = J$

$y_3 = 0.6$

$X_3 = H$

J J H H … T

$X_0, X_1, X_2, X_3, \ldots, X_t$

Simulating trajectories

- N populations
  - N = 10 and n=17

$X_0, X_1, X_2, X_3, \ldots, X_n$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
2	T	T	T	T	T	T	T	T	T	T	T	T	T	H	T	T	T
3	H	H	H	H	J	J	J	J	J	J	J	J	J	J	J	J	J
4	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
5	H	H	H	H	T	T	T	T	T	T	T	T	T	T	T	T	T
6	T	T	T	T	T	T	T	T	T	T	T	J	J	J	J	J	J
7	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
8	T	T	T	T	T	H	H	H	H	T	T	T	T	T	T	T	T
9	H	H	H	H	H	H	H	H	J	J	J	J	J	J	J	J	J
10	T	T	T	T	T	T	T	T	T	J	J	J	J	J	H	H	T

Trajectoire moyenne

H = 6/10 =0.6

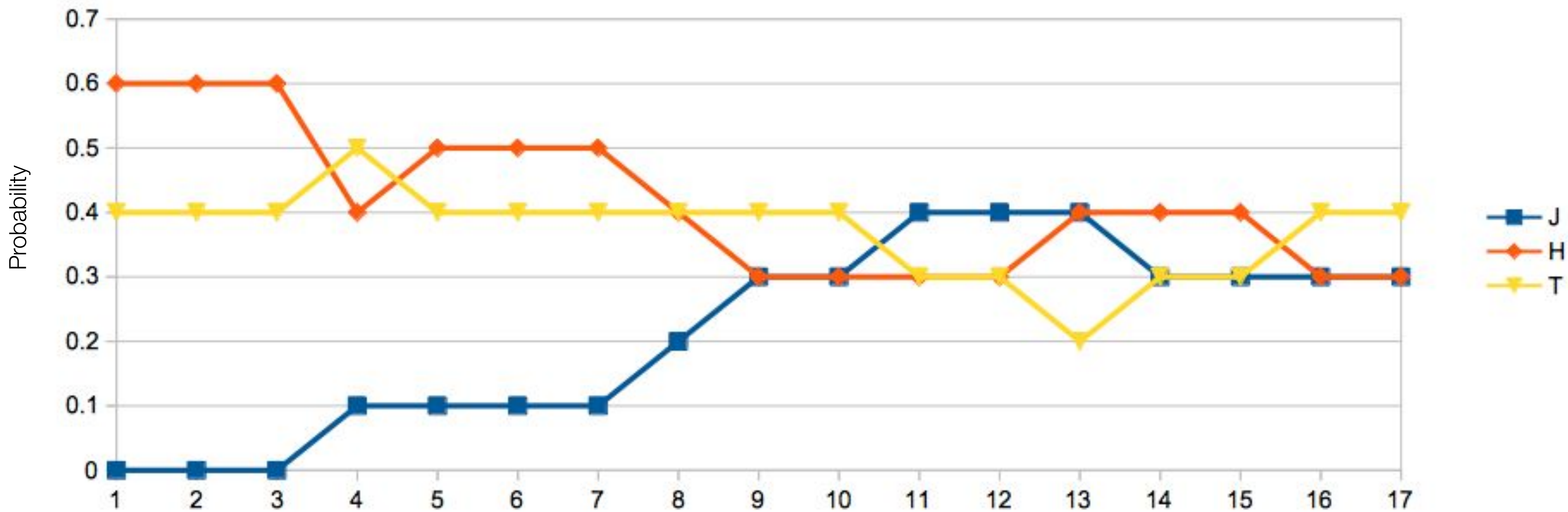
J = 0/10 =0.0

T = 4/10 =0.4

	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
J	0	0	0	0	0.1	0.1	0.1	0.1	0.2	0.3	0.3	0.4	0.4	0.4	0.3	0.3	0.3	0.3
H	0.6	0.6	0.6	0.6	0.4	0.5	0.5	0.5	0.4	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.3	0.3
T	0.4	0.4	0.4	0.4	0.5	0.4	0.4	0.4	0.4	0.4	0.4	0.3	0.3	0.2	0.3	0.3	0.4	0.4

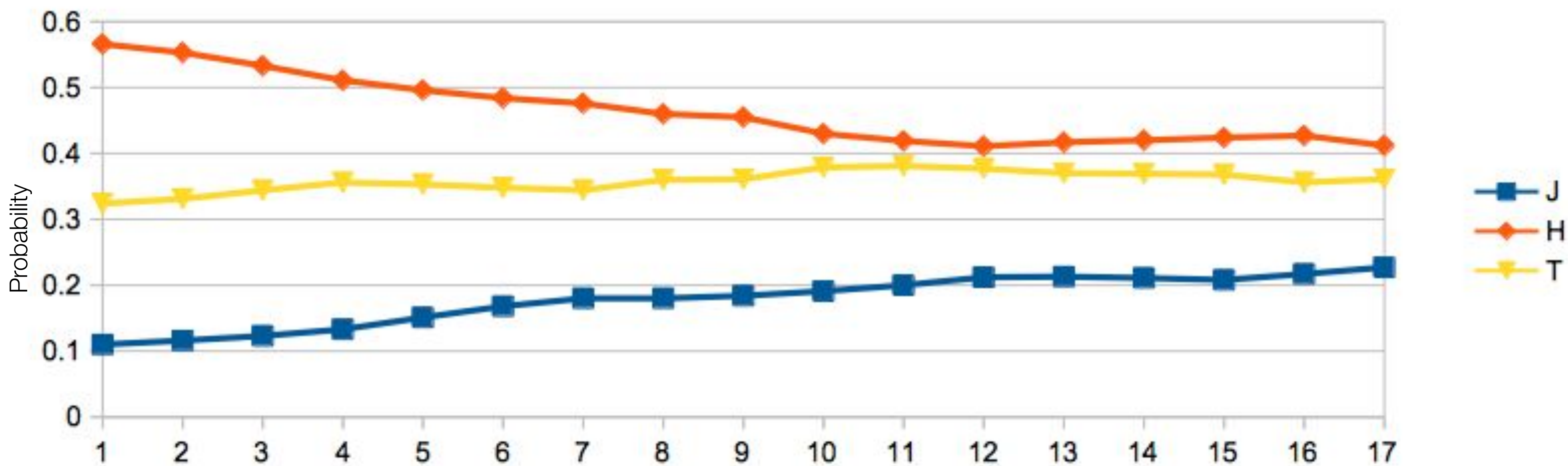
Simulating trajectories

- Simulating trajectories n populations
  - N = 10 and n=17



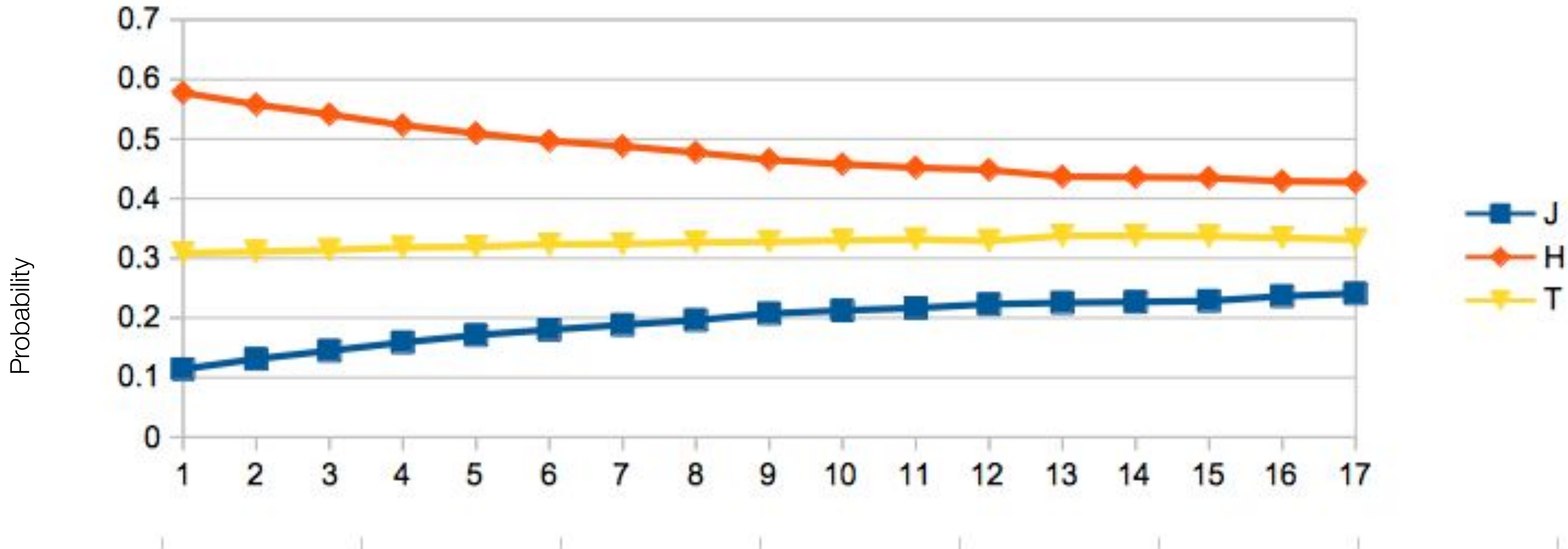
Simulating trajectories

- Simulating trajectories n populations
  - N = 1000 and n=17



Simulating trajectories

- Simulating trajectories n populations
    - N = 10000 and n=17



Computing probability of a trajectory

- Let's suppose we have simulated the trajectory$_1$

trajectory$_1$: HHHJJT

Give back

		J	H	T
Take	J	0.95	0.03	0.02
	H	0.02	0.90	0.08
	T	0.05	0.05	0.90

$$\pi_0 = \begin{array}{c} \left[\begin{array}{c} 0.1 \\ 0.6 \\ 0.3 \end{array}\right] \end{array} \begin{array}{l} J \\ H \\ T \end{array}$$

Prob(trajectory$_1$) = P(H) P(H|H) P(H|H) P(J|H) P(J|J) P(T|J)

log(Prob(trajectory$_1$)) = Log[P(H)] + log[P(H|H)] + log[P(H|H)] + log[P(J|H)] + log[P(J|J)] + log[P(T|J)]

Log[P(H)] = log(0.6) = -0.5108256

Computing probability of a trajectory

- Let's suppose we have simulated the trajectory_1

		Give back		
		J	H	T
Take	J	0.95	0.03	0.02
	H	0.02	0.90	0.08
	T	0.05	0.05	0.90

$$\pi_0 = \begin{array}{c} \begin{bmatrix} 0.1 \\ 0.6 \\ 0.3 \end{bmatrix} \end{array} \begin{array}{c} J \\ H \\ T \end{array}$$

trajectory_1 : HHHJJT

$\text{Prob}(\text{trajectory}_1) = P(H)\, P(H|H)\, P(H|H)\, P(J|H)\, P(J|J)\, P(T|J)$

$\log(\text{Prob}(\text{trajectory}_1)) = \text{Log}[P(H)] + \log[P(H|H)] + \log[P(H|H)] + \log[P(J|H)] + \log[P(J|J)] + \log[P(T|J)]$

Log[P(H)] = log(0.6) = -0.5108256

Log[P(H|H)] = log(0.90) = -0.1053605

Log[P(H|H)] = log(0.90) = -0.1053605

Log[P(J|H)] = log(0.02) = -3.912

Log[P(J|J)] = log(0.95) = -0.05129329

Log[P(T|J)] = log(0.02) = -3.912

$\text{Log}(\text{Prob}(\text{trajectory}_1)) = -8.5968859603254852$

Markov chain Applications

CpG islands

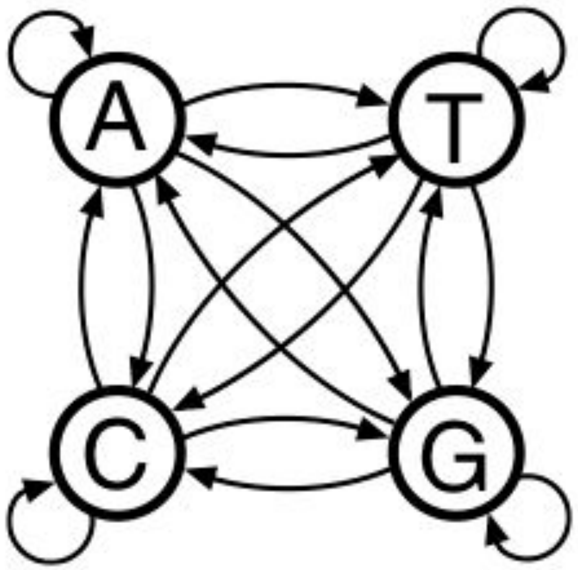
- CpG sites are regions of DNA where a cytosine is followed by a guanine in the linear sequence of bases along its 5' → 3' direction.

- CpG islands (or CG islands) are regions with a high frequency of CpG sites.

- CpG island is a region with at least 200 bp, and a GC percentage that is greater than 50%

- Many genes in mammalian genomes have CpG islands associated with the start of the gene, so it is used in gene border predictions.

Markov Chains for CpG islands



Derive two Markov chain models:

'+' model: from the CpG islands

'-' model: from the remainder of sequence

Transition probabilities for each model:

$$a^{+}_{st} = \frac{c^{+}_{st}}{\sum_{t'} c^{+}_{st'}}$$

c^{+}_{st} is the number of times letter ***t*** followed letter ***s*** inside the CpG islands

$$a^{-}_{st} = \frac{c^{-}_{st}}{\sum_{t'} c^{-}_{st'}}$$

c^{-}_{st} is the number of times letter ***t*** followed letter ***s*** outside the CpG islands

Markov Chains for CpG islands

CpG islands

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

non CpG islands

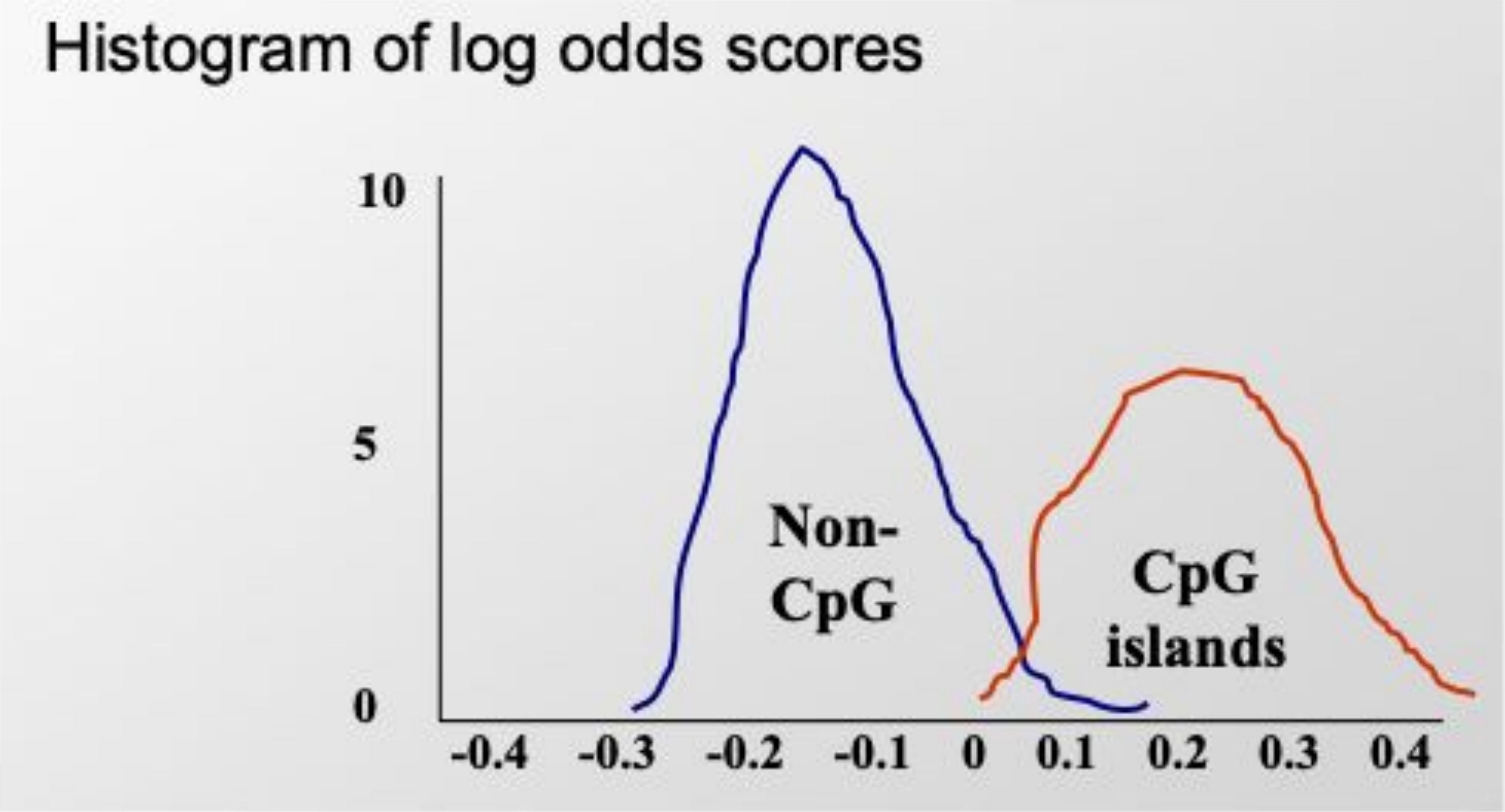
−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Given a short sequence *x*, does it come from CpG island (**Yes-No** question)

- To use these models for discrimination, calculate the log-odds ratio:

$$S(x_L) = \frac{\Pr(x|\ \text{model}+)}{\Pr(x|\ \text{model}-)} \qquad \sum_{i=1}^{L} \log \frac{A^{+}(x_{i-1}, x_i)}{A^{-}(x_{i-1}, x_i)}$$

Markov Chains for CpG islands



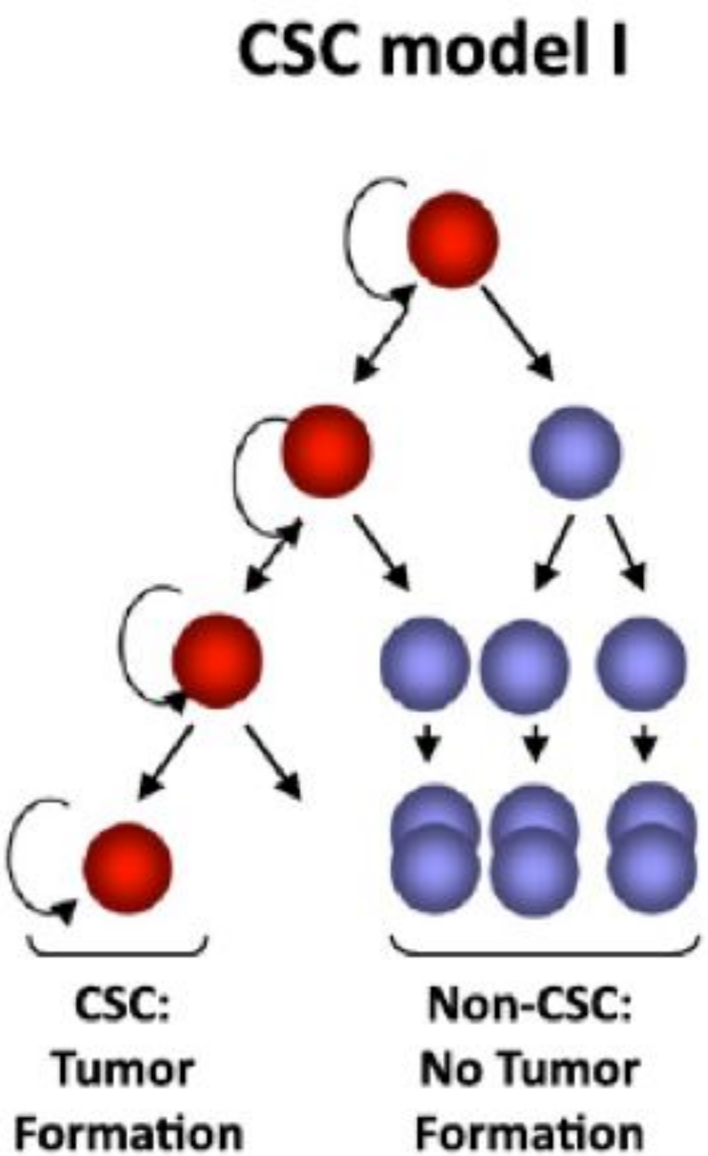
Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells

Gupta et, al.

Cancer cells phenotypic

- Cancer cells within individual tumors often exist in distinct phenotypic states that differ in functional attributes.

The classical Picture



Cancer stem cells (CSC)

- one of phenotypical states
- give rise to non-CSCs during tumor growth
- non-CSCs cannot produce CSCs
- CSC differentiation gives rise to cell-state equilibrium
- only CSC seed tumors

‣ hierarchical cell-lineage structure like in normal tissue development

Cancer cells phenotypic

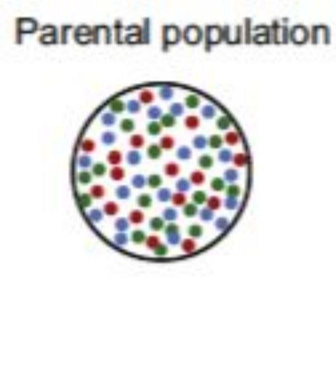
- The goal is to study the dynamics of phenotypic proportions in human breast cancer cell lines : SUM159 and SUM149

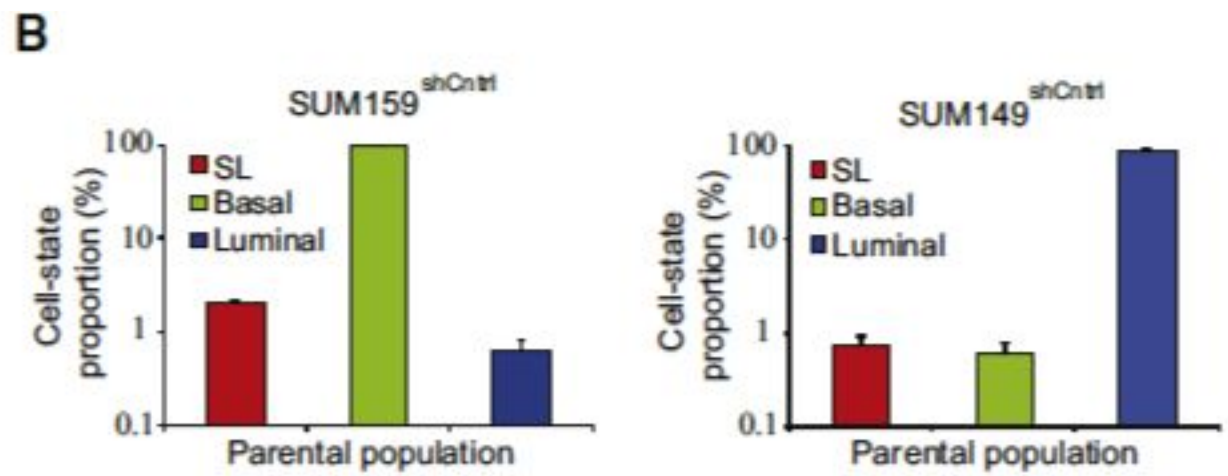
- It is showed that subpopulations of cells purified for a given phenotypic state return towards equilibrium proportions over time.

- This equilibrium can be explained by a Markov model in which cells transition stochastically between states.

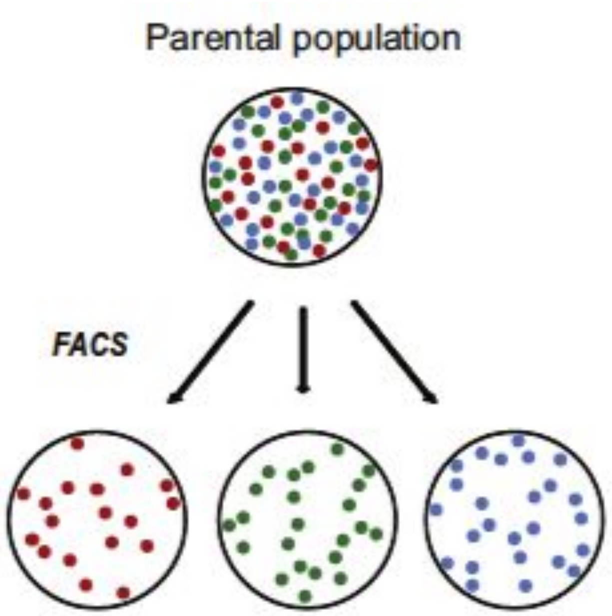
The experiment

- Determination of the proportions of the individual cell-states in the SUM159 and SUM149



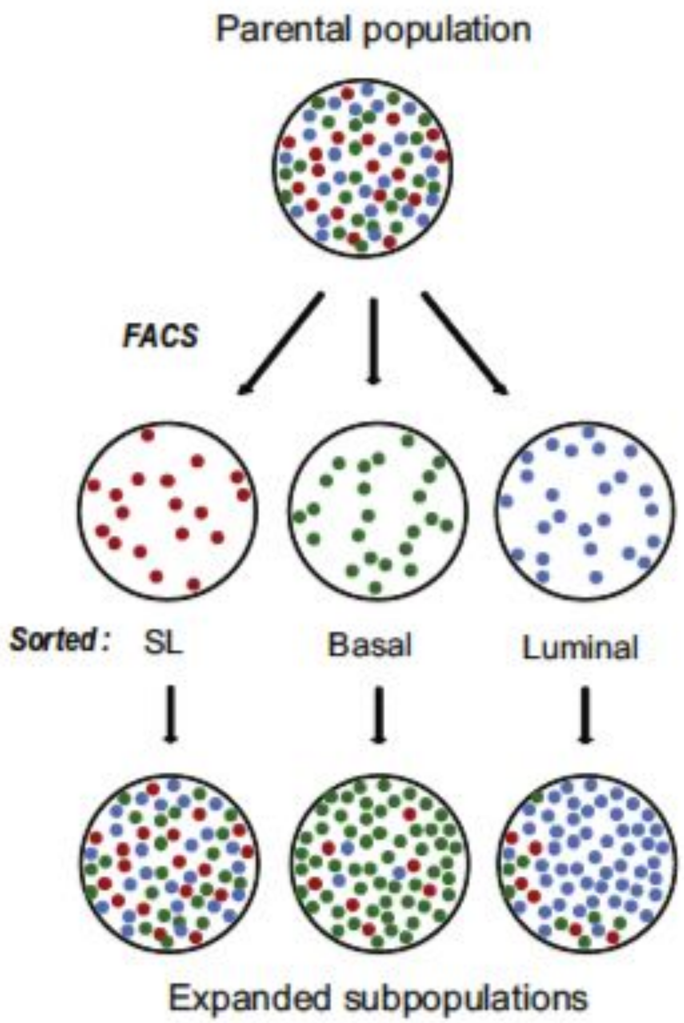


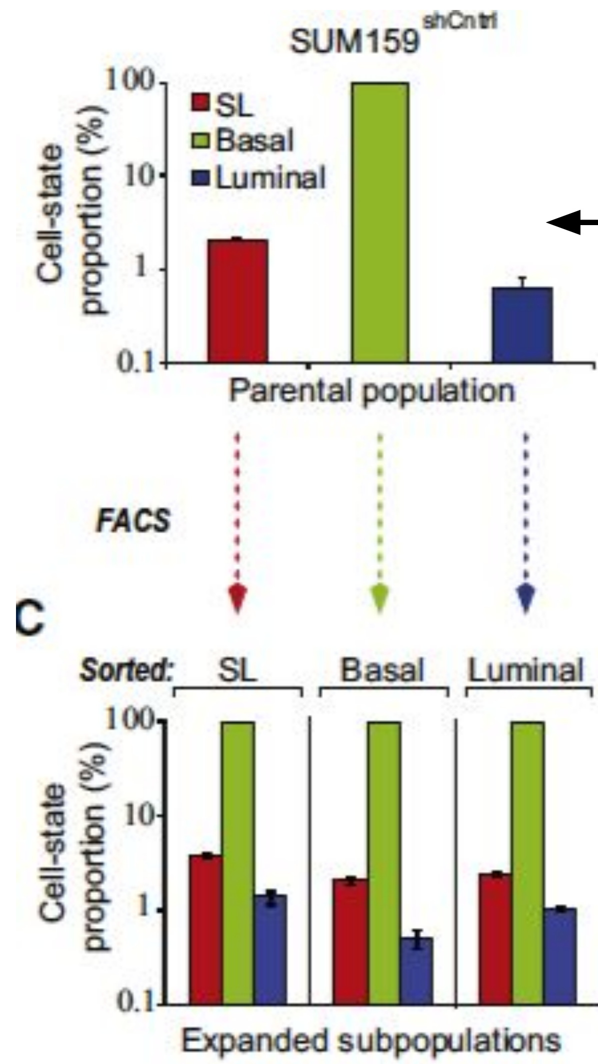
- Fluorescence-activated cell sorting (**FACS**) to isolate three mammary epithelial cell states: stem-like, basal, and luminal



- 99% purification by FACS (fluorescence-activated cell sorting)
- 6 days of cell growth (ca. 6 cell cycles)

Results SUM159





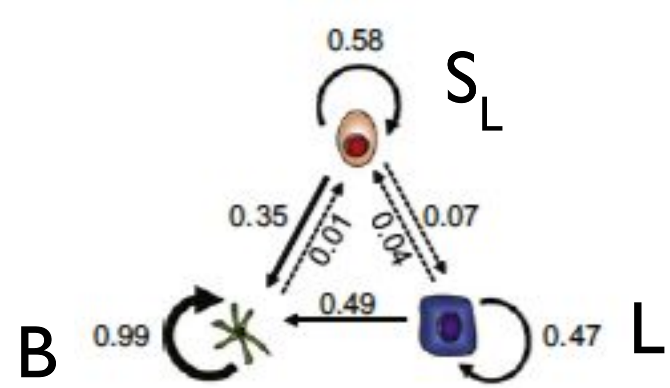
π_0

After 6 days

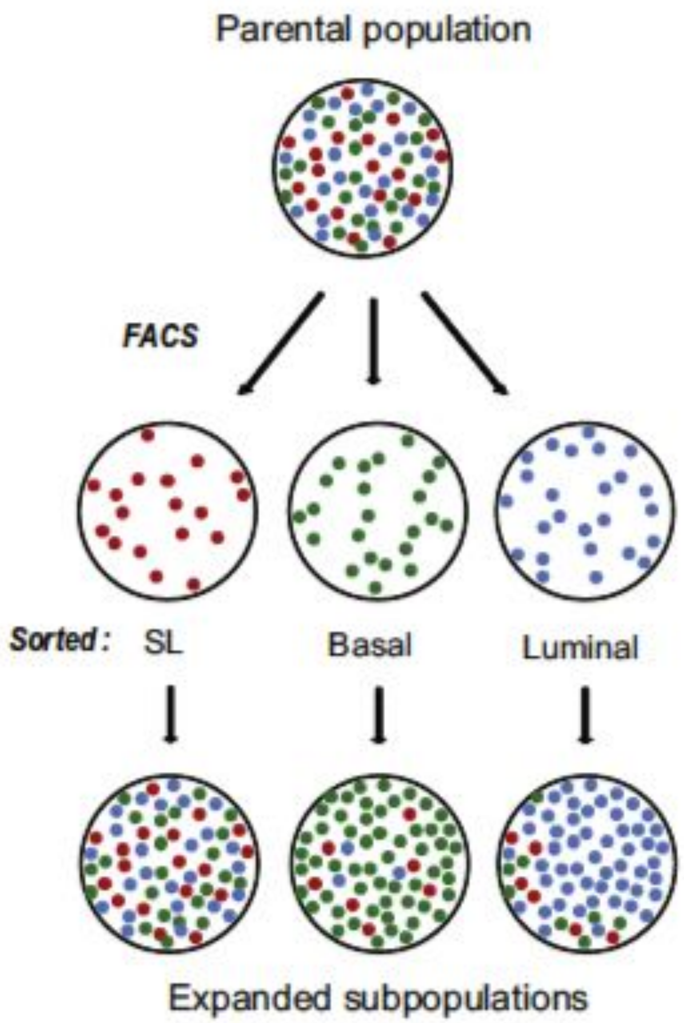
$A = P^6$

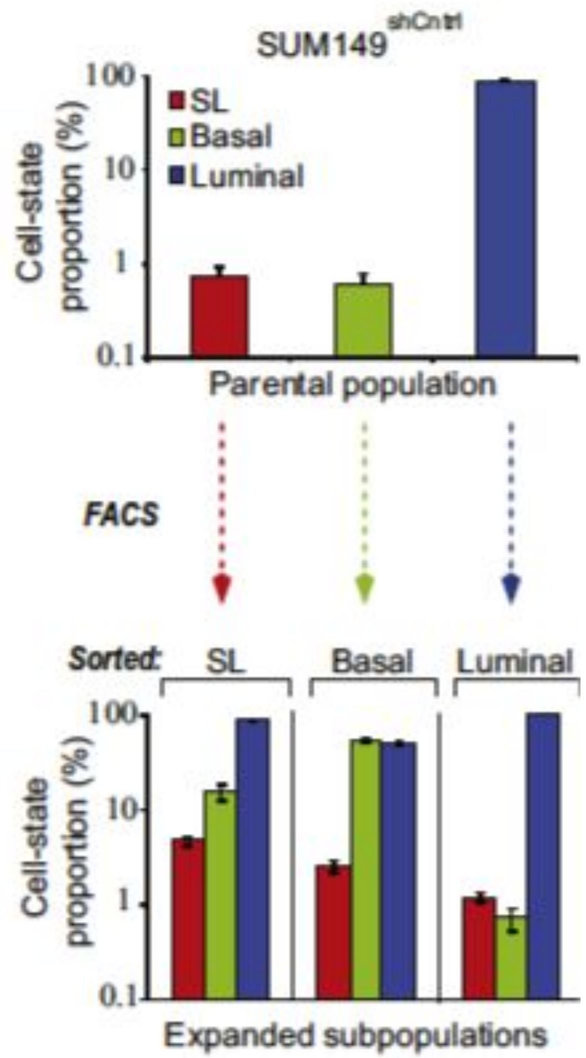
How to obtain P ?

$P = A^{1/6}$



Results SUM149





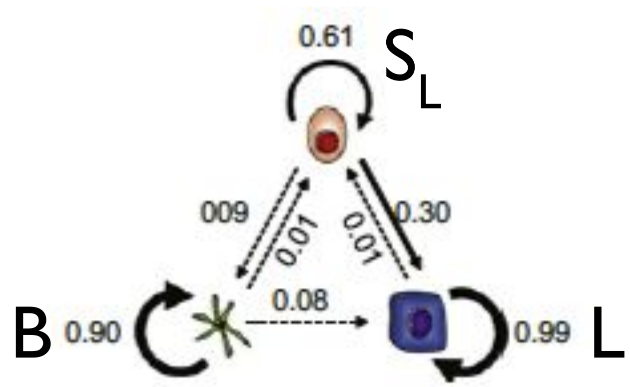
π_0

After 6 days

$A = P^6$

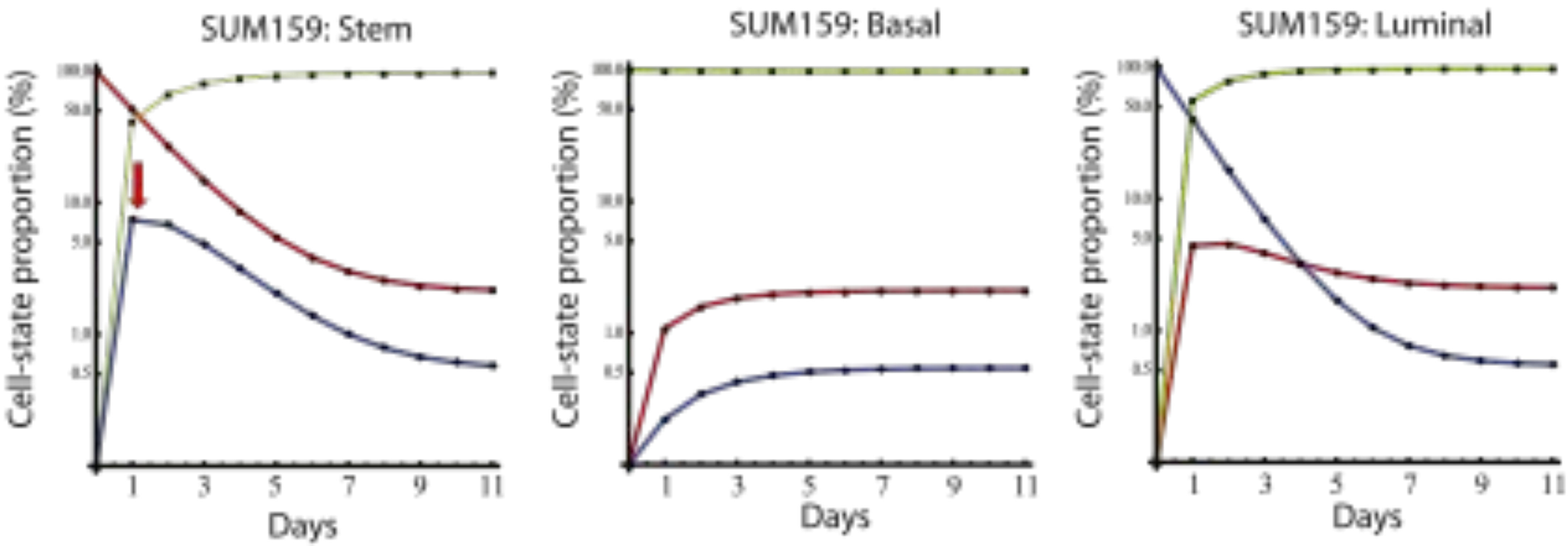
How to obtain P ?

$P = A^{1/6}$



Predicting trajectories

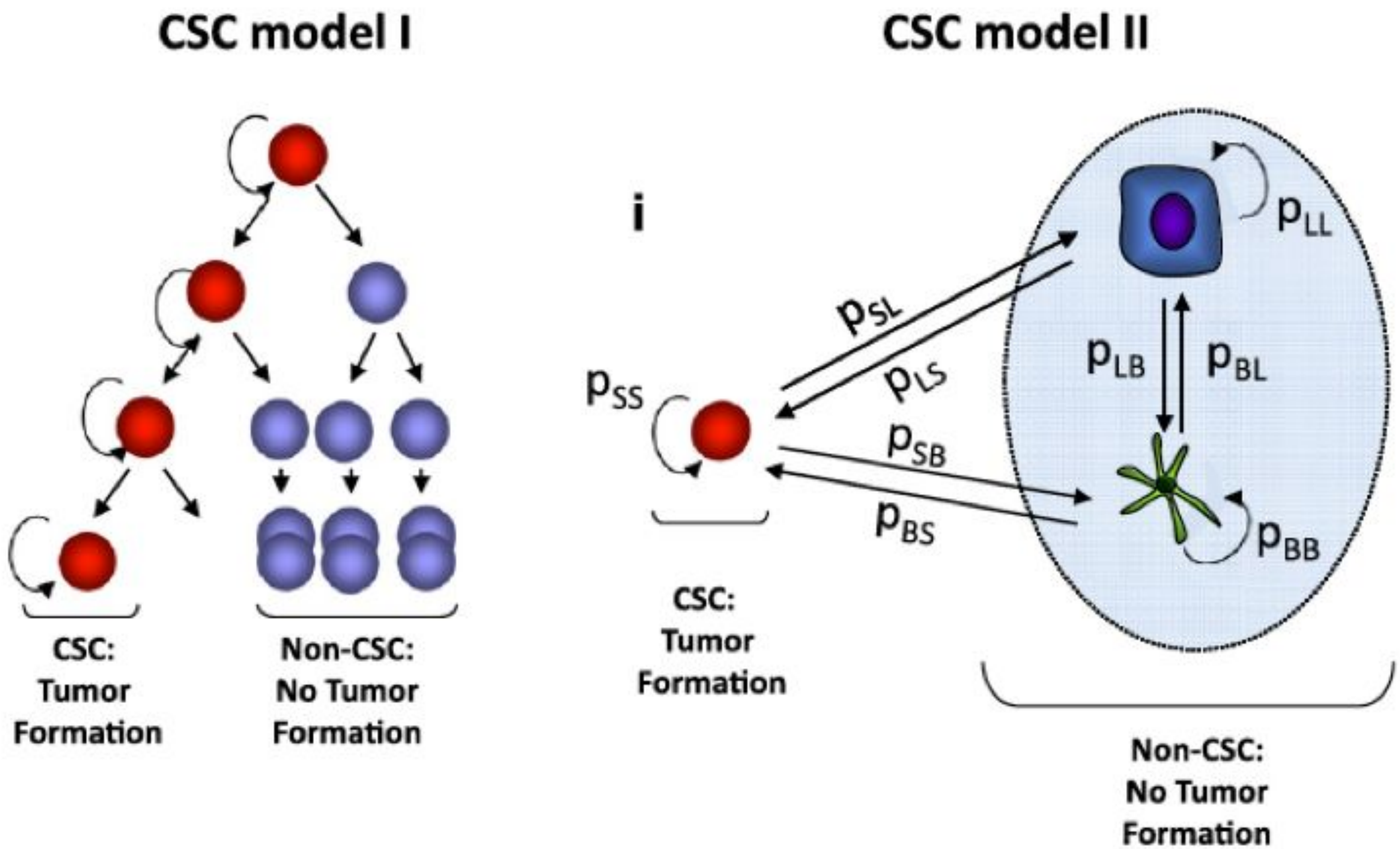
- Predicting trajectories to predict how a population of cells evolves over time given the initial proportion of cell states



S = 100%	S = 0%	S = 0%
B = 0%	B = 100%	B = 0%
L = 0%	L = 0%	L = 100%

Conclusion

A modified picture



Important therapeutic implications!