Skryté Markovove modely

Andrej Lúčny

Katedra aplikovanej informatiky FMFI UK

a MicroStep-MIS

andy@microstep-mis.com

http://www.microstep-mis.com/~andy

Motivačná úloha

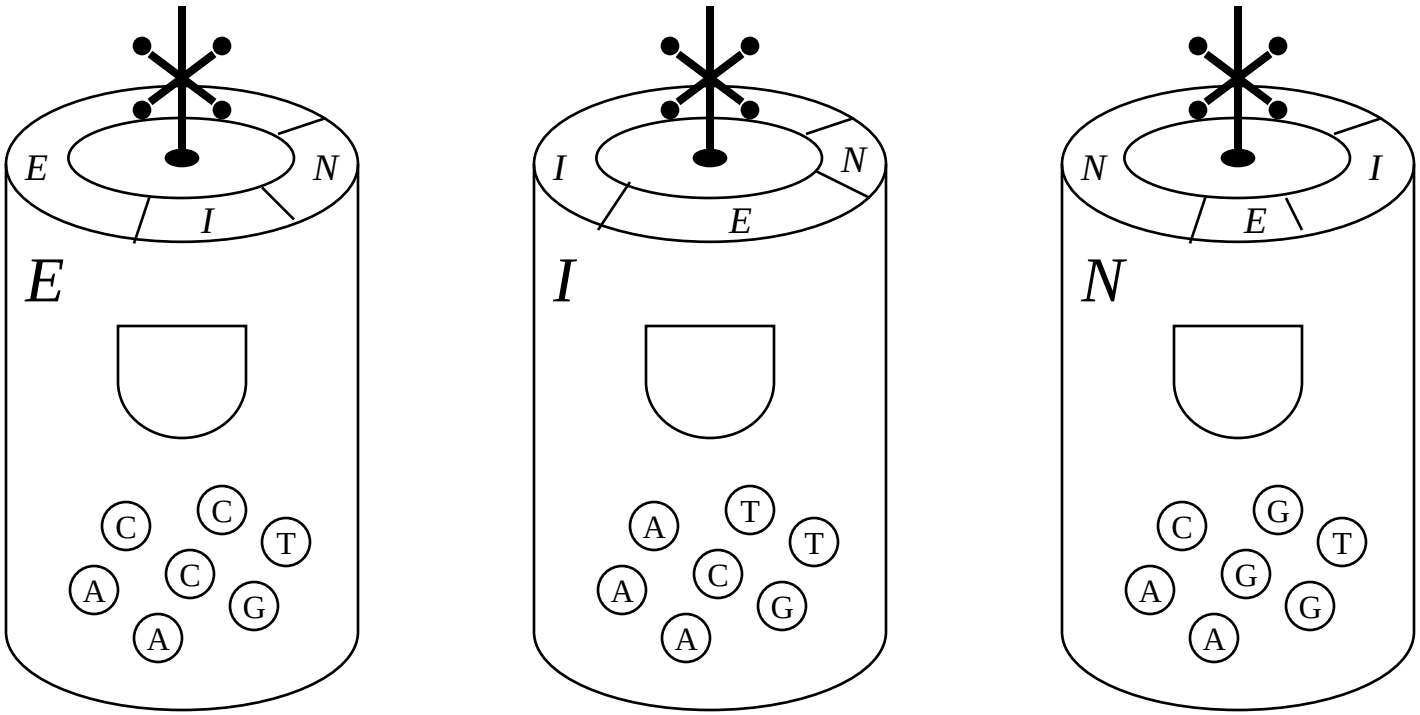
G A G T T G A G G G A A A A A G C G

Máme jedno vlákno DNA, čo je postupnosť báz A, C, G, T

Potrebujeme určiť, ktoré časti vlákna sú exony (E), introny (I) a gény nekódujúca DNA (N)

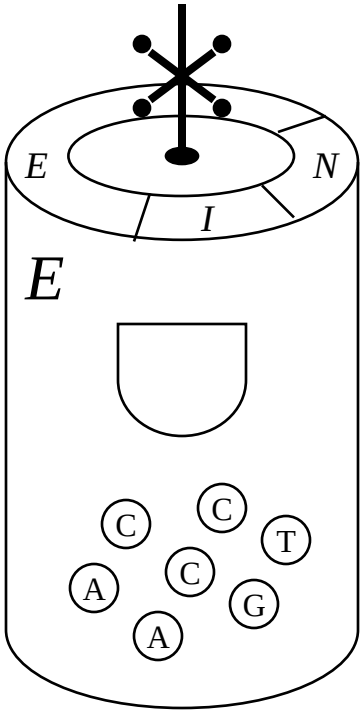
G A G T T G A G G G A A A A A G C G

N N N N N N I I I I I E E E E E E N

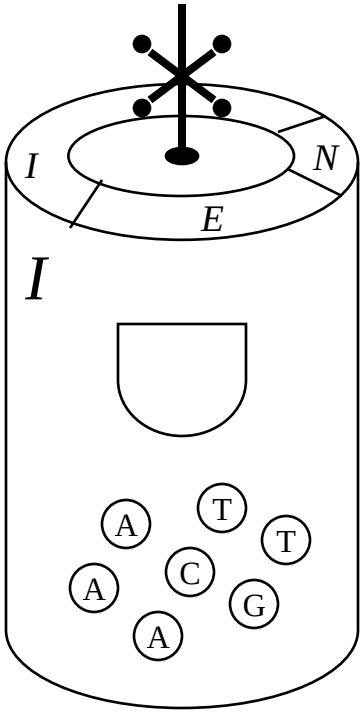


Markovov model DNA:

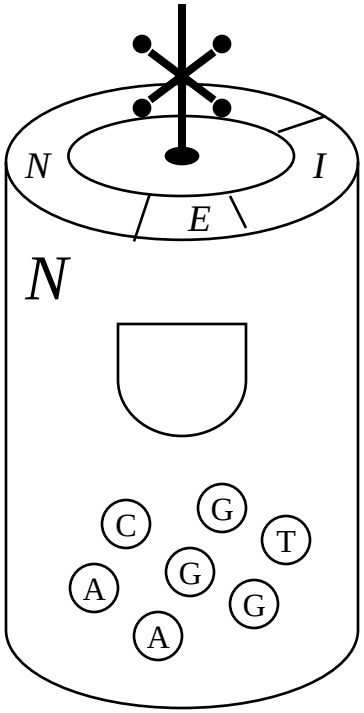
Máme tri urny: E, I a N. V každej sú schované guličky A, C, G a T a na každej je ruleta, ktorej jamky sú podelené na úseky E, I a N



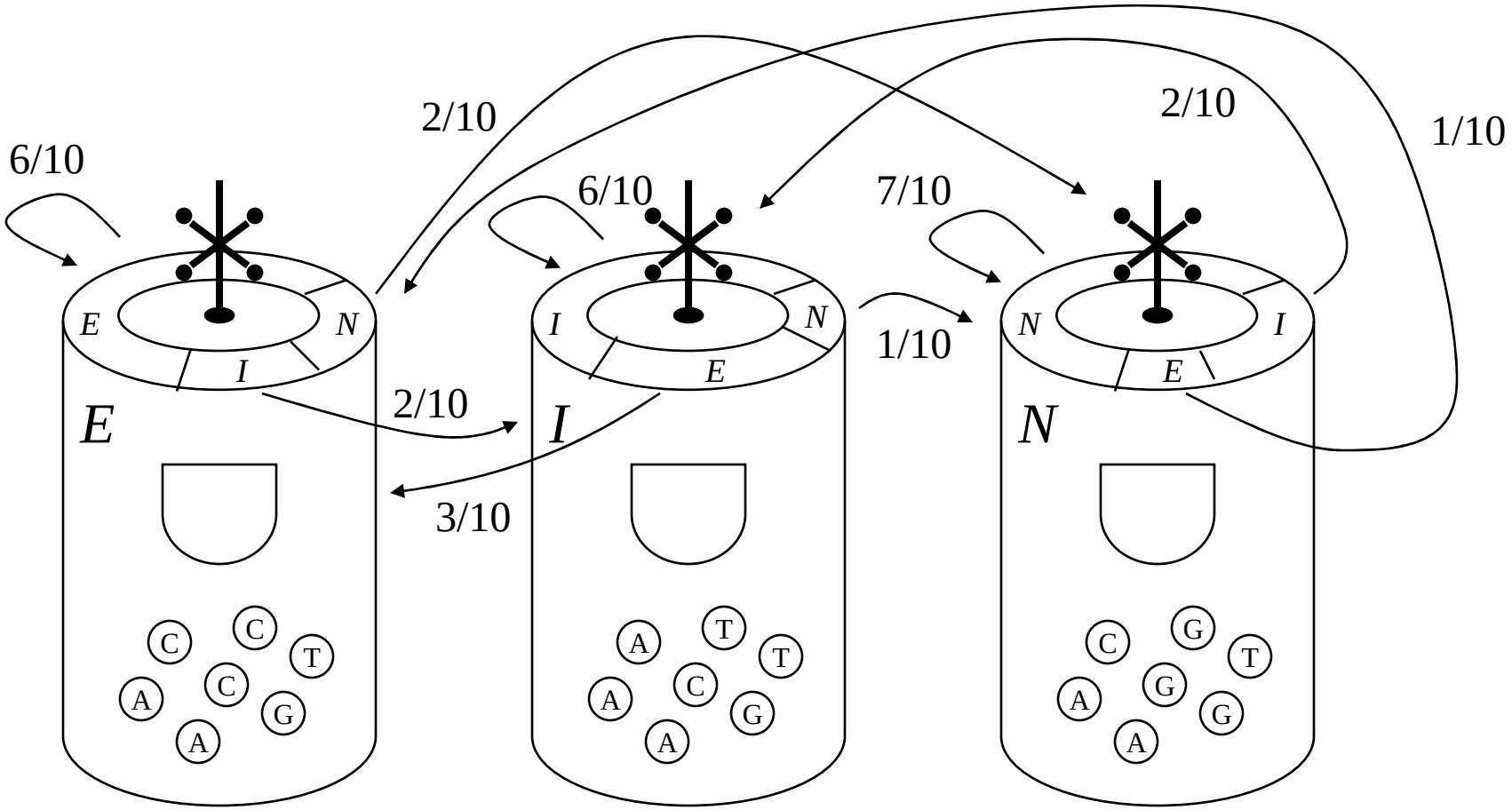
p(A) = 2/7
p(C) = 3/7
p(G) = 1/7
p(T) = 1/7

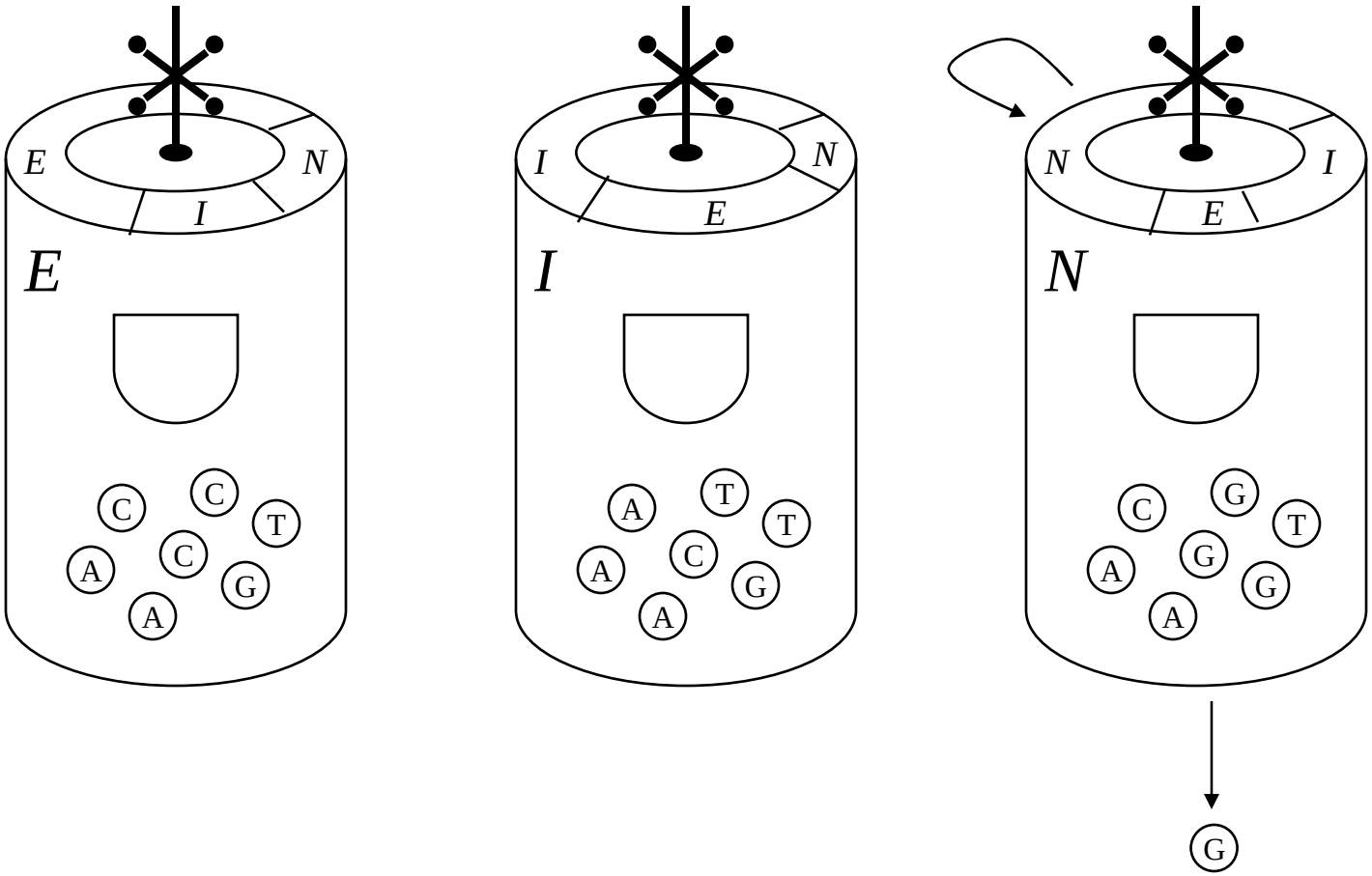


p(A) = 3/7
p(C) = 1/7
p(G) = 1/7
p(T) = 2/7



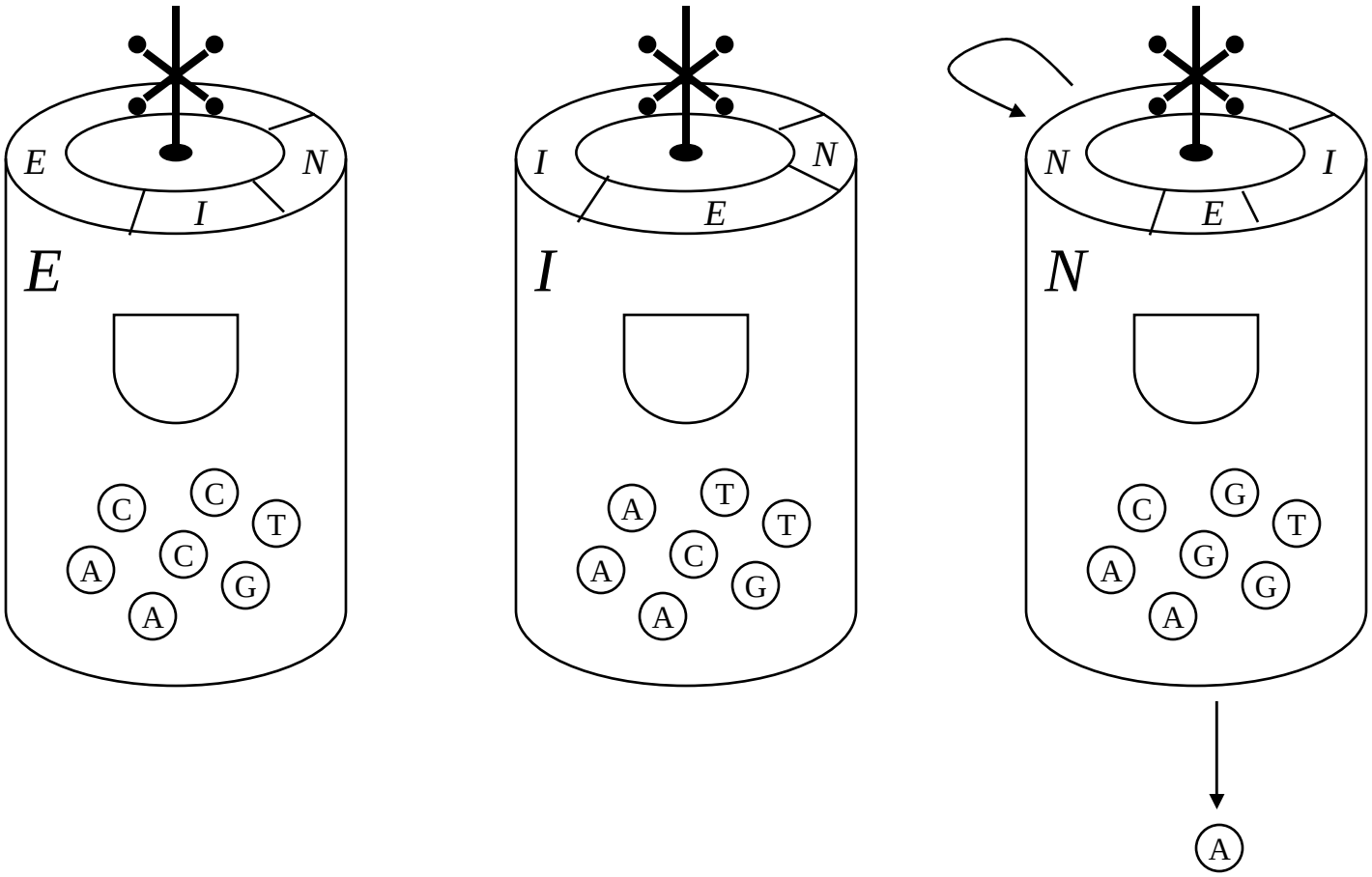
p(A) = 2/7
p(C) = 1/7
p(G) = 3/7
p(T) = 1/7


6/10
2/10
2/10
1/10
6/10
7/10
E
N
I
I
E
N
N
E
I
1/10
2/10
E
I
N
3/10
C
C
T
C
A
G
A
A
T
T
C
A
G
A
C
G
T
G
A
G
G
A



Markovov proces:

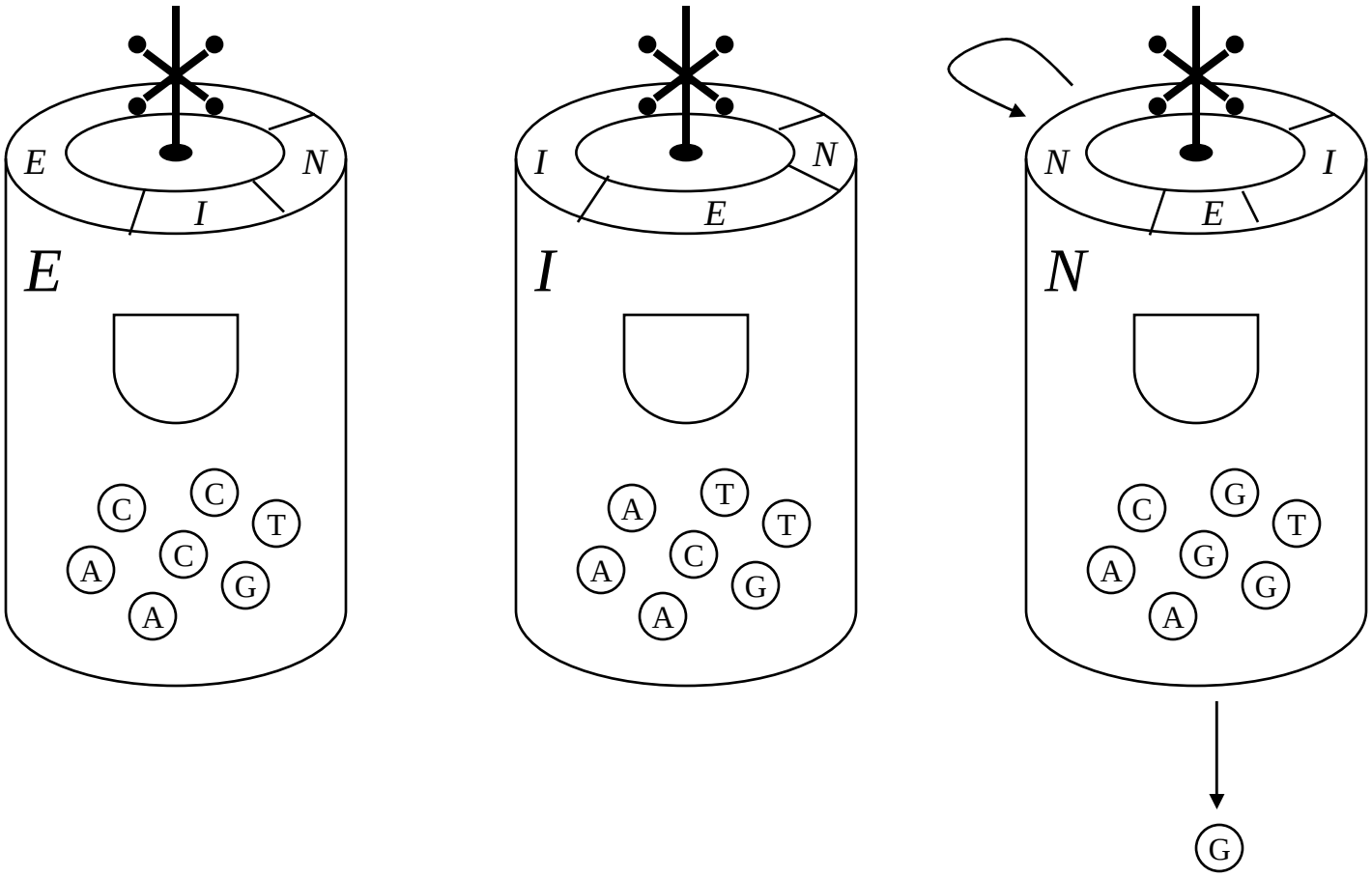
1. Vytiahneme jednu guličku z urny (a vrátime ju späť)
2. Zatočíme ruletou na urne a podľa výsledku sa presunieme



N N

. . . . .

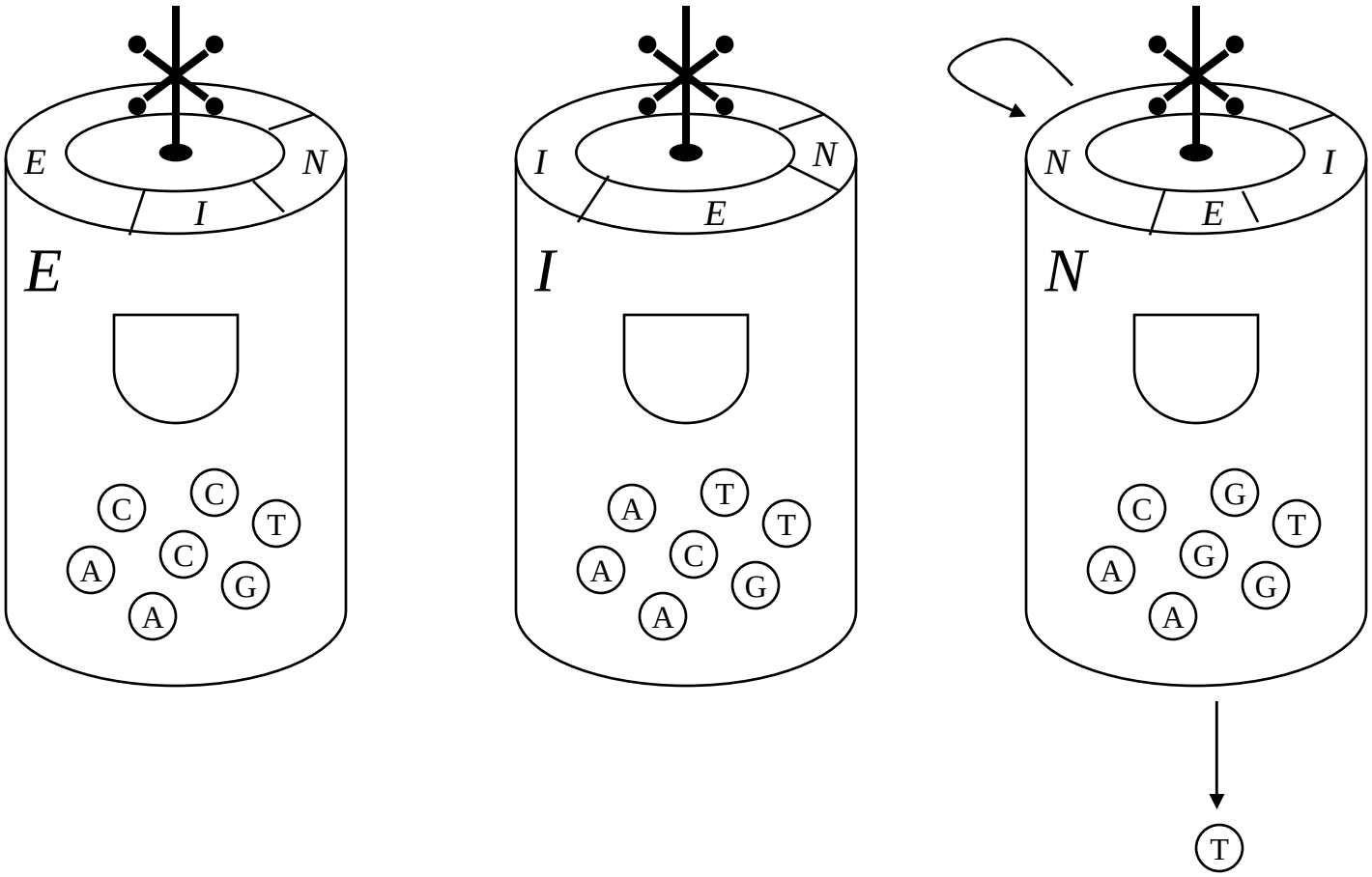
G A



N N N

. . . . .

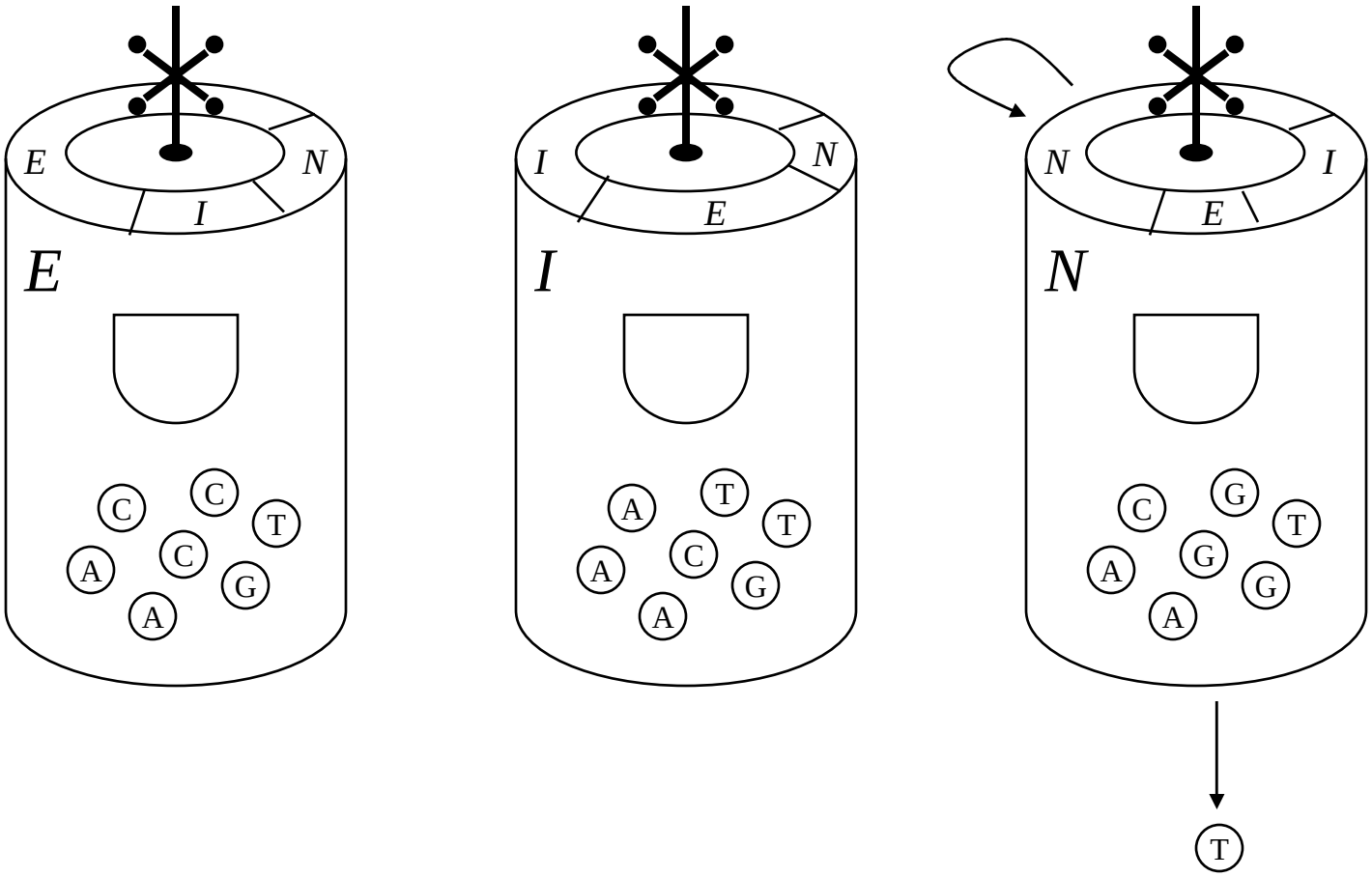
G A G



N N N N

. . . . .

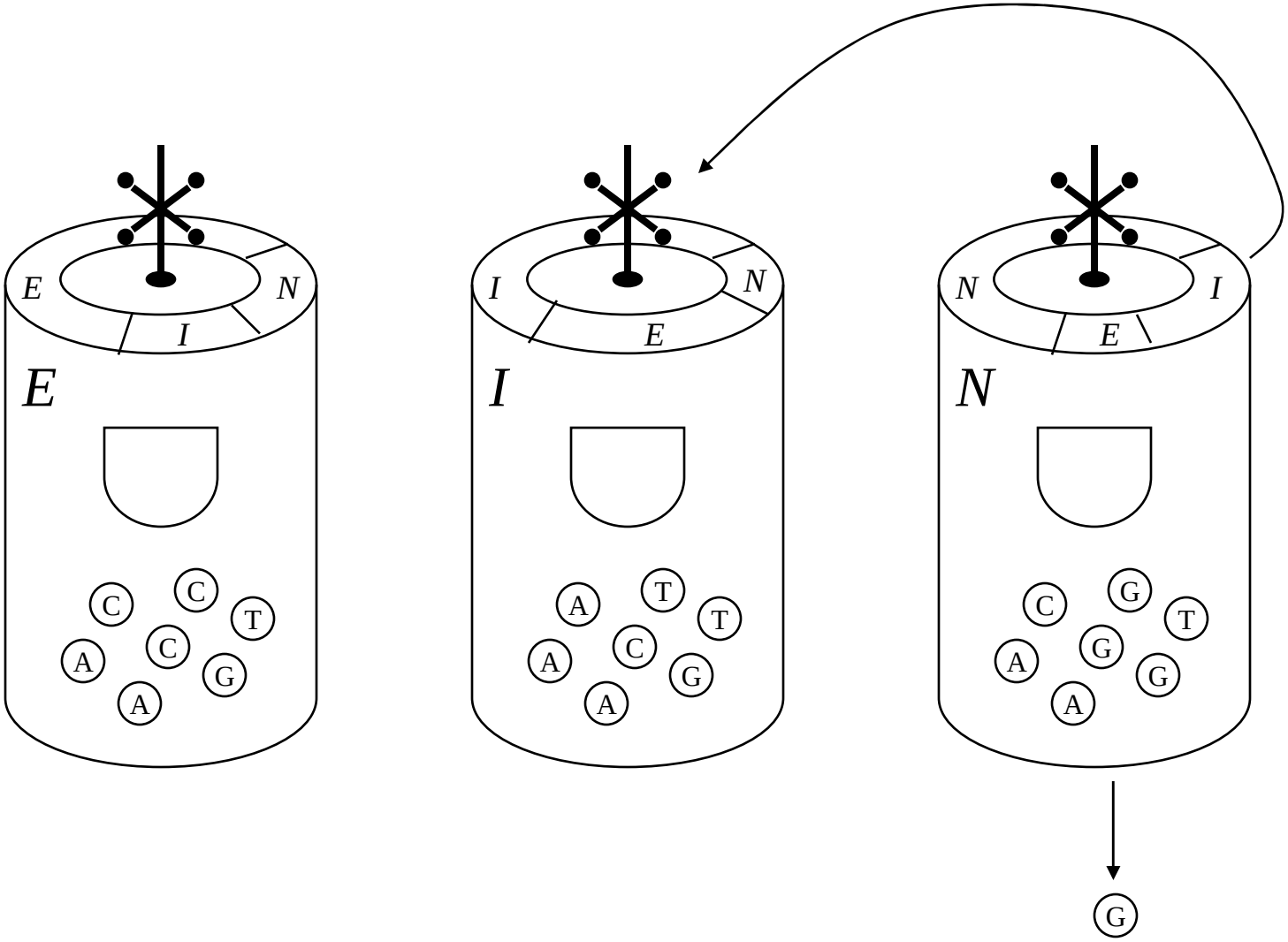
G A G T



N N N N N

G A G T T

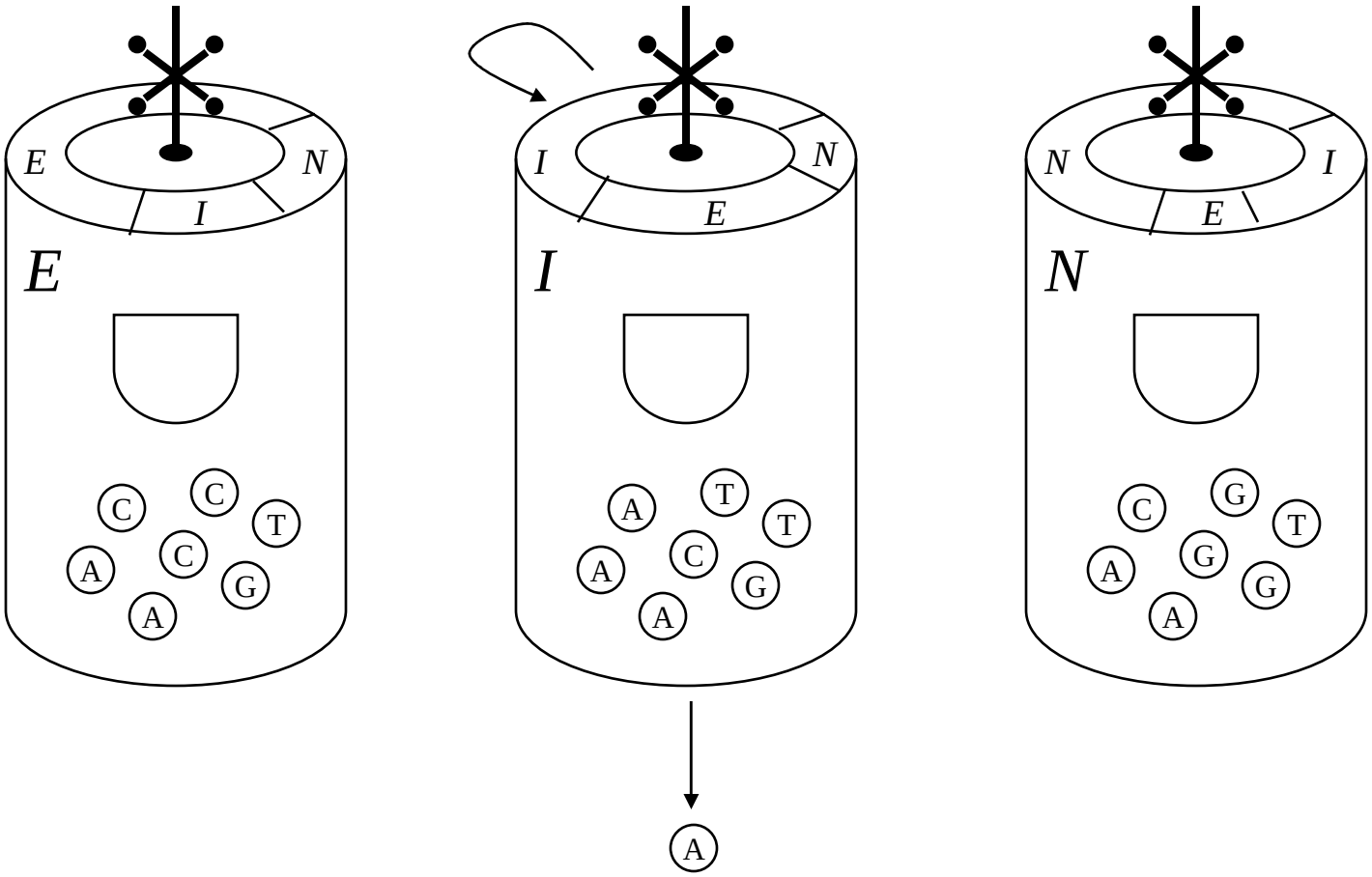
. . . . .



N N N N N N

. . . . .

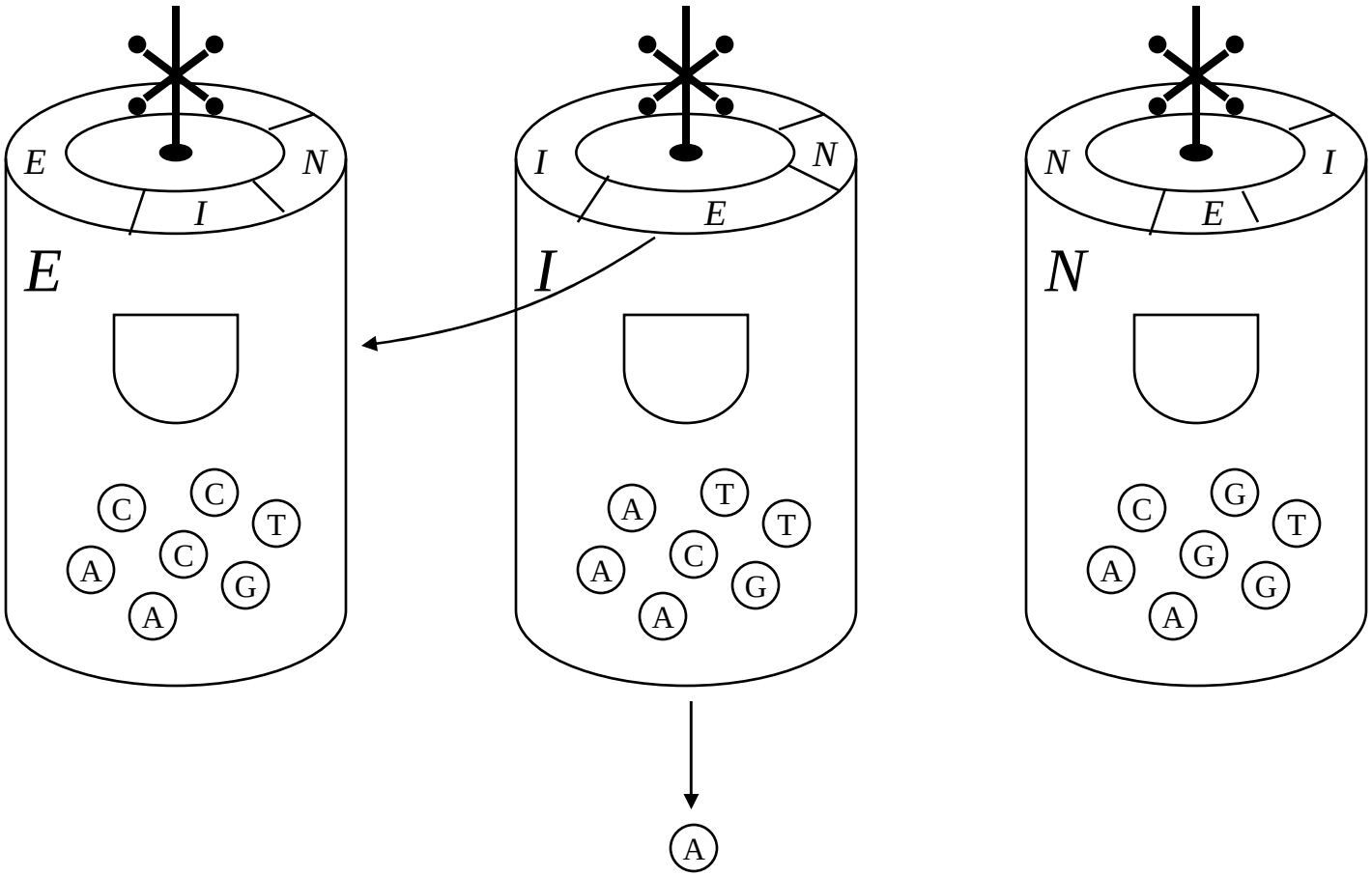
G A G T T G



N N N N N N I

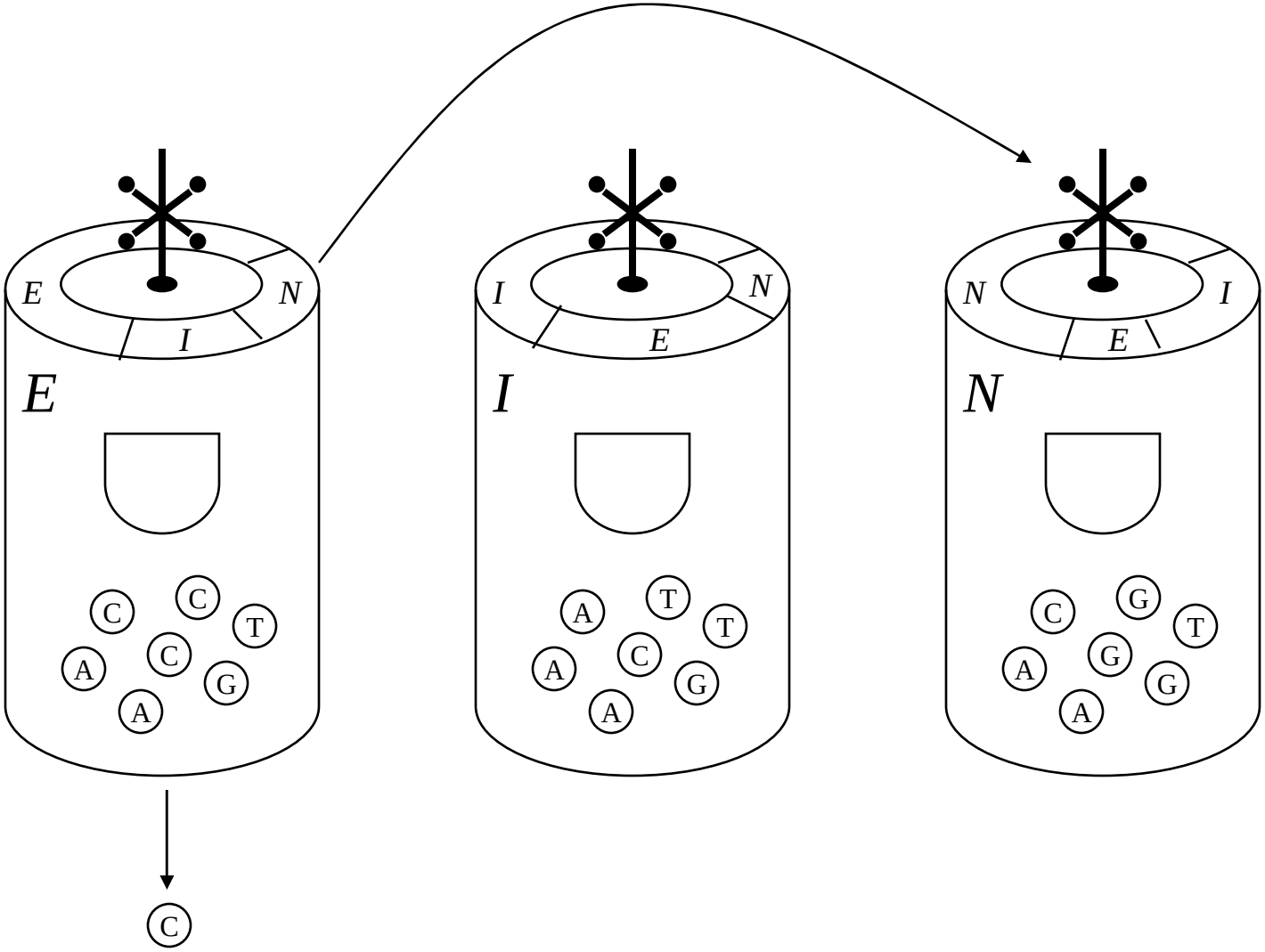
. . . . .

G A G T T G A



N N N N N N I I I I I . . . . .

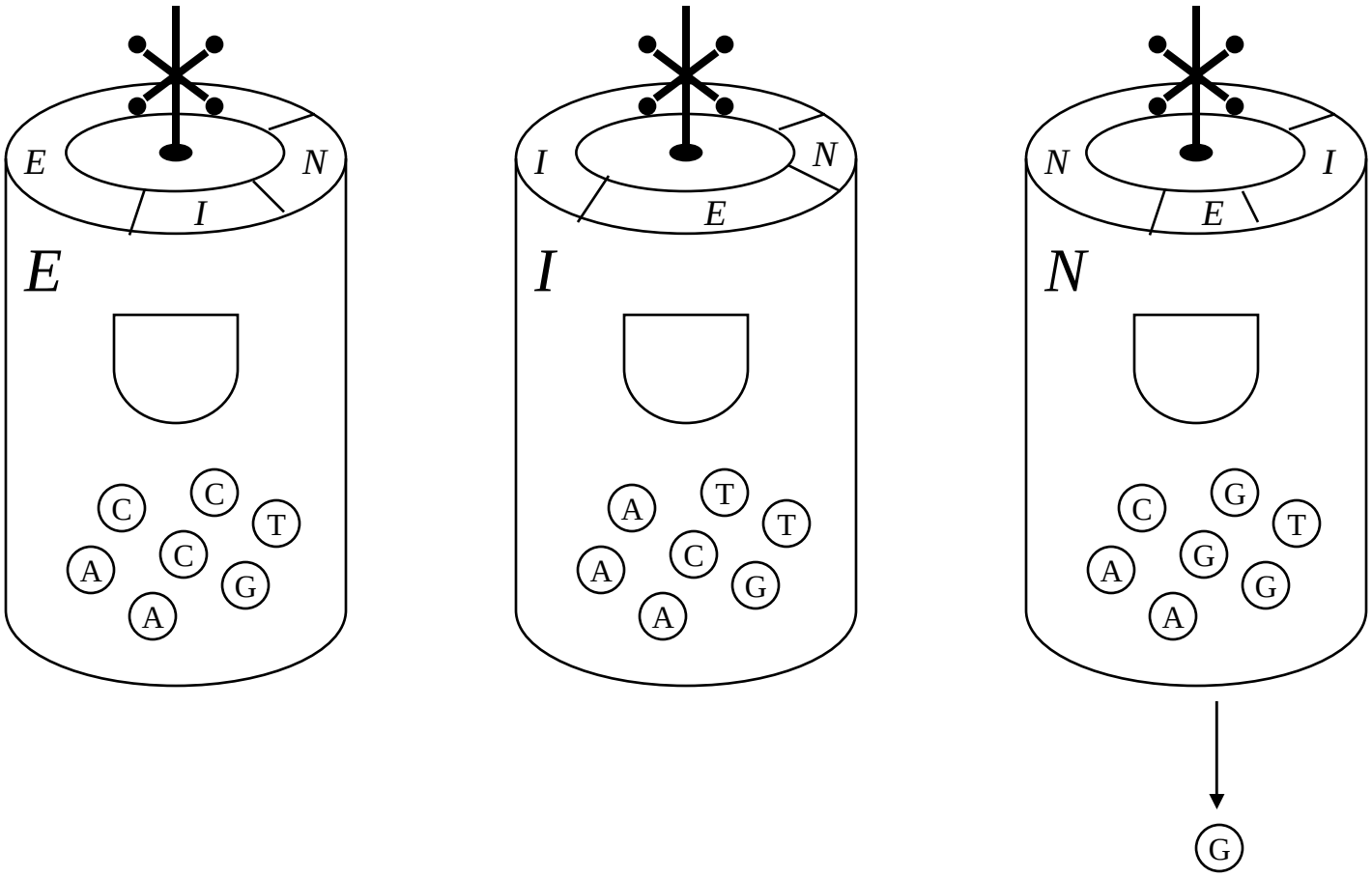
G A G T T G A G G G A



N N N N N N I I I I I E E E E E E

G A G T T G A G G G A A A A A G C . . . . .

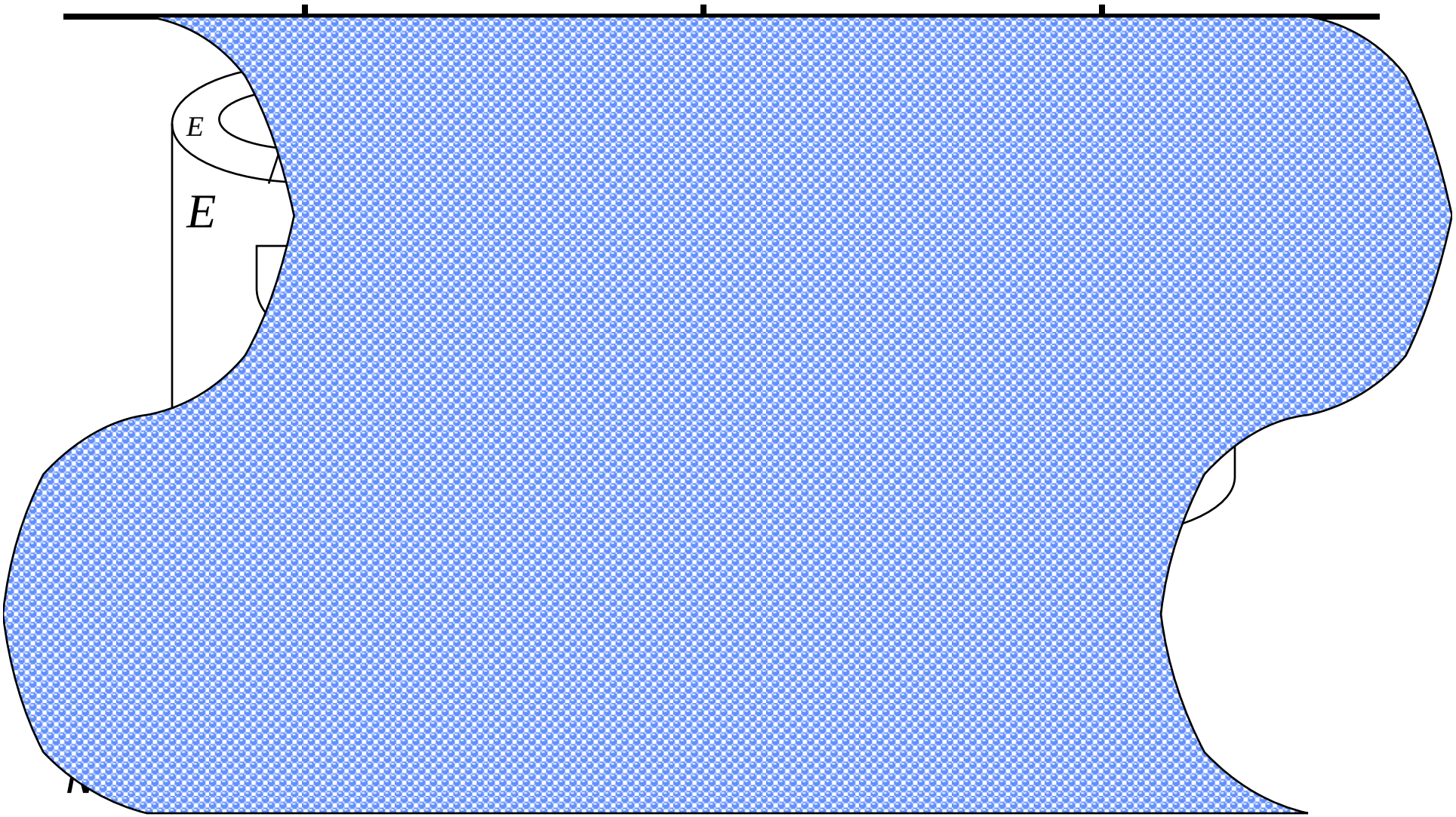
Dostávame výsledok Markovovho procesu:
čo sme kedy z ktorej urny vytiahli



N N N N N N I I I I I E E E E E E N

G A G T T G A G G G A A A A A G C G

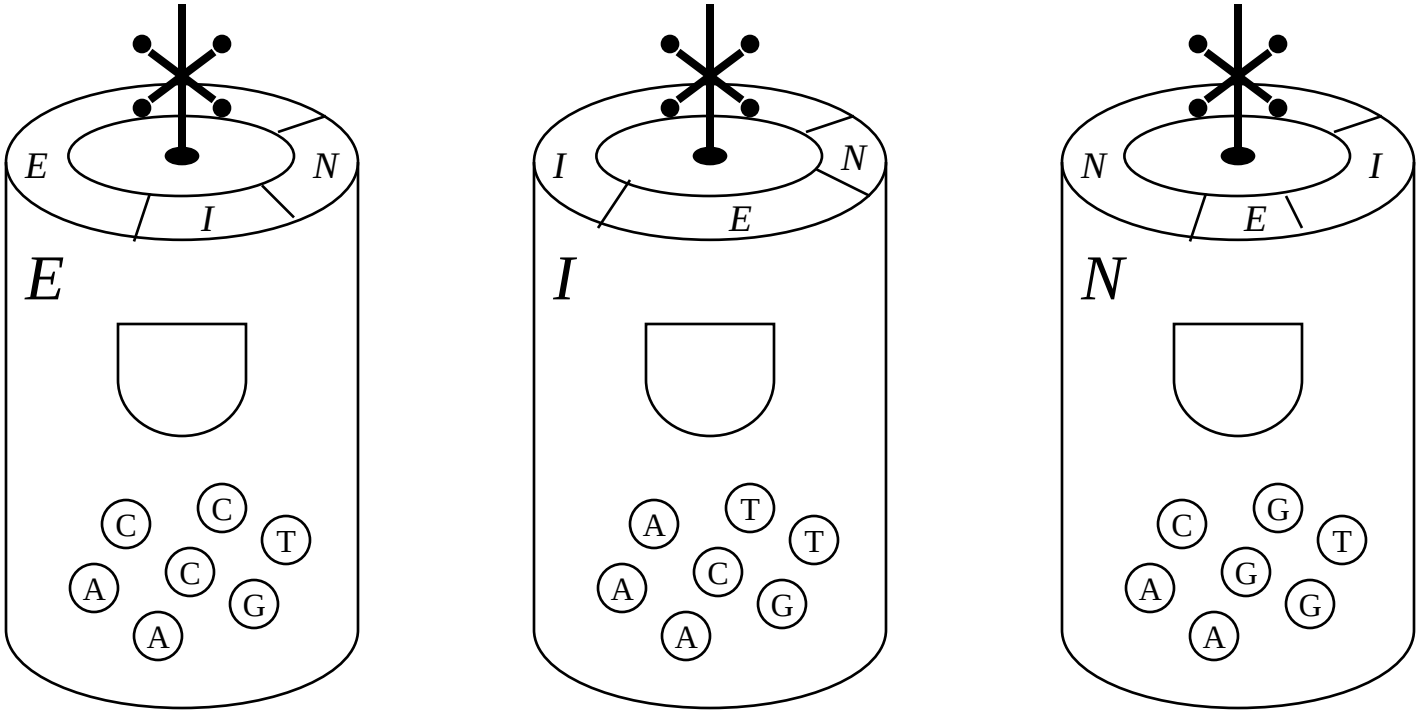
Skrytý Markovov proces: vidíme iba čo vyťahujeme, ale nie z ktorej urny



G A G T T G A G G G A A A A A G C G

Pre danú postupnosť ťahov máme teraz uhádnuť z akých urien boli vytiahnuté

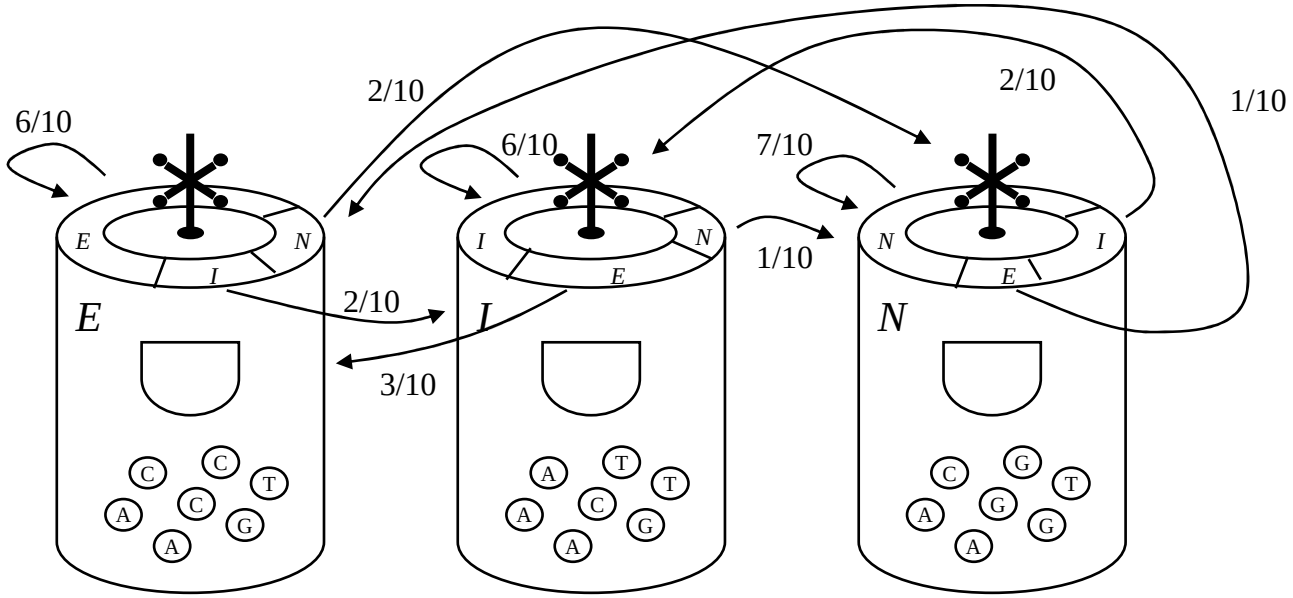
Vieme pre nejakú hypotézu spočítať
nakoľko je pravdepodobná ?



p(*N I I I E E E I I I E E E N N I I N*) = ?

hypotéza

G A G T T G A G G G A A A A A G C G



p(A) = 2/7
p(C) = 3/7
p(G) = 1/7
p(T) = 1/7

p(A) = 3/7
p(C) = 1/7
p(G) = 1/7
p(T) = 2/7

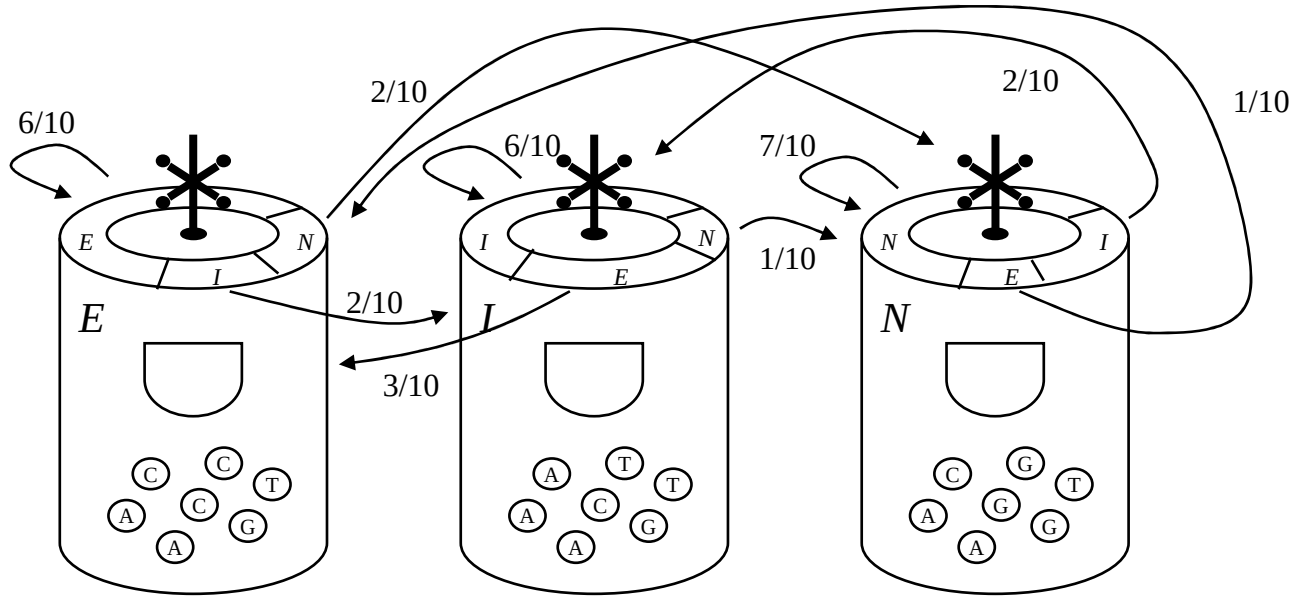
p(A) = 2/7
p(C) = 1/7
p(G) = 3/7
p(T) = 1/7

ľahko to z markovovho modelu spočítame ako súčin pravdepodobností jednotlivých ťahov a presunov

ťah	G	A	G	T	T	G	A
h	N	I	I	I	E	E	E
p(tah)	3/7	3/7	1/7	2/7	1/7	1/7	2/7
p(presun)	2/10	6/10	6/10	3/10	6/10	6/10	

. . . . .

p(*N I I I E E E I I I E E E N N I I N*) = 1.6768E-19



p(A) = 2/7
p(C) = 3/7
p(G) = 1/7
p(T) = 1/7

p(A) = 3/7
p(C) = 1/7
p(G) = 1/7
p(T) = 2/7

p(A) = 2/7
p(C) = 1/7
p(G) = 3/7
p(T) = 1/7

keď zrátame pravdepodobnosť hypotézy rovnajúcej sa skutočnosti, je vyššia než u náhodne zvolenej hypotézy

ťah	G	A	G	T	T	G	A
h	N	N	N	N	N	N	I
p(tah)	3/7	2/7	3/7	1/7	1/7	3/7	3/7
p(presun)	7/10	7/10	7/10	7/10	7/10	2/10	

. . . . .

p(*N N N N N N I I I I I E E E E E E N*) = 8.7350E-16

G A G T T G A G G G A A A A A G C G

⋮

p(*N I I I E E E I I I E E E N N I I N*) = 1.6768E-19

⋮

p(*N N N N N N I I I I I E E E E E E N*) = 8.7350E-16

⋮

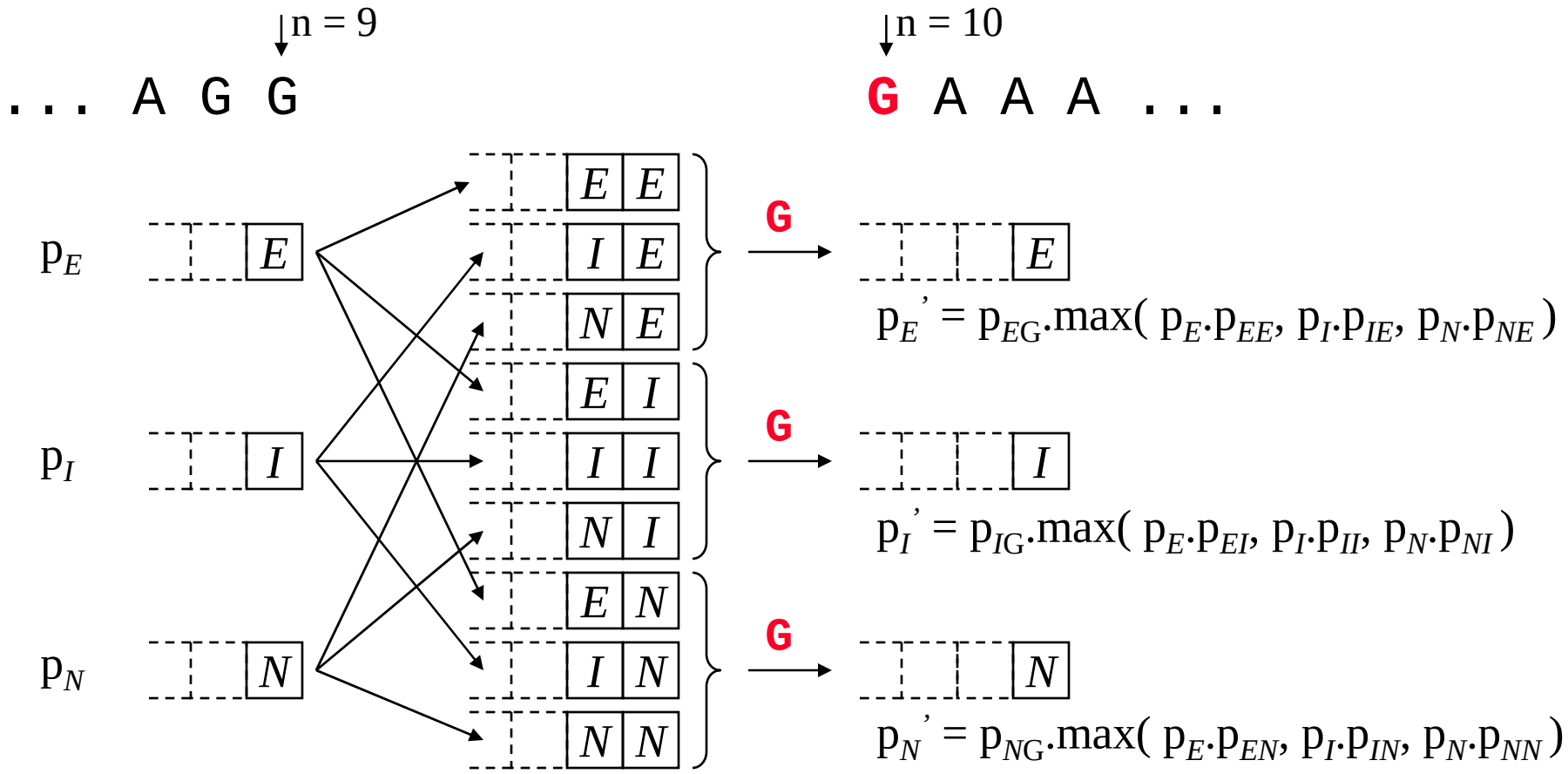
Stačilo by teraz prebehnúť všetky možnosti a nájsť najpravdepodobnejšiu. Asi to nebola presne tá skutočná, ale bola by snáď aspoň dostatočne podobná, aby sme sa z toho niečo dozvedeli. Tých možností je ale šialene veľa. Našťastie celkom jednoduchý (Viterbiho) algoritmus vie najpravdepodobnejšiu možnosť priamo spočítať.

Viterbiho algoritmus

Zamerajme sa na samotnú hodnotu pravdepodobnosti najpravdepodobnejšej postupnosti urien.

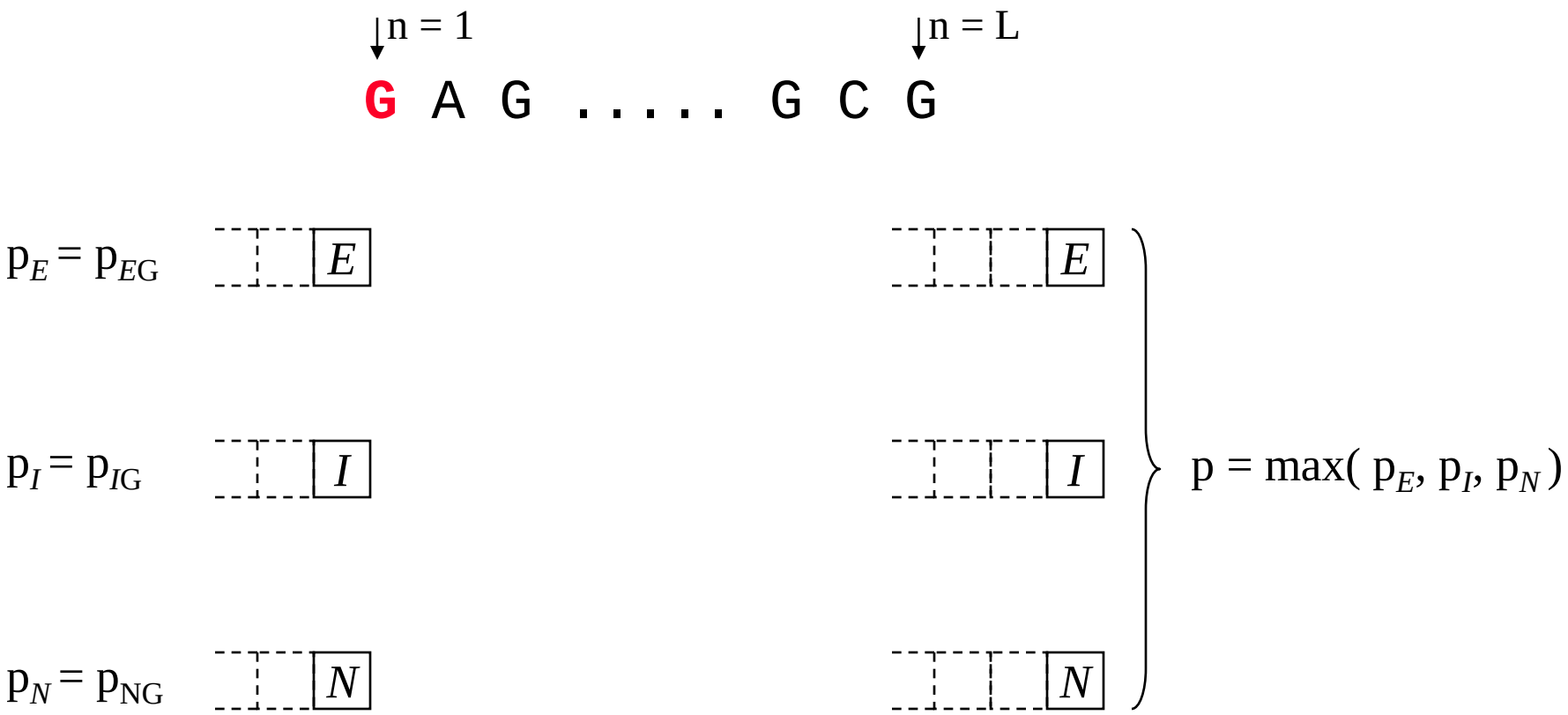
Počítajme pravdepodobnosti čiastočných postupností tvorených prvými n členmi postupnosti urien, ktoré končia urnou E, I a N

Predstavme si, že ich vieme pre určité n. Ako ich získame pre n+1 ?



Zvyšok je aplikovanie indukcie.

Viterbiho algoritmus

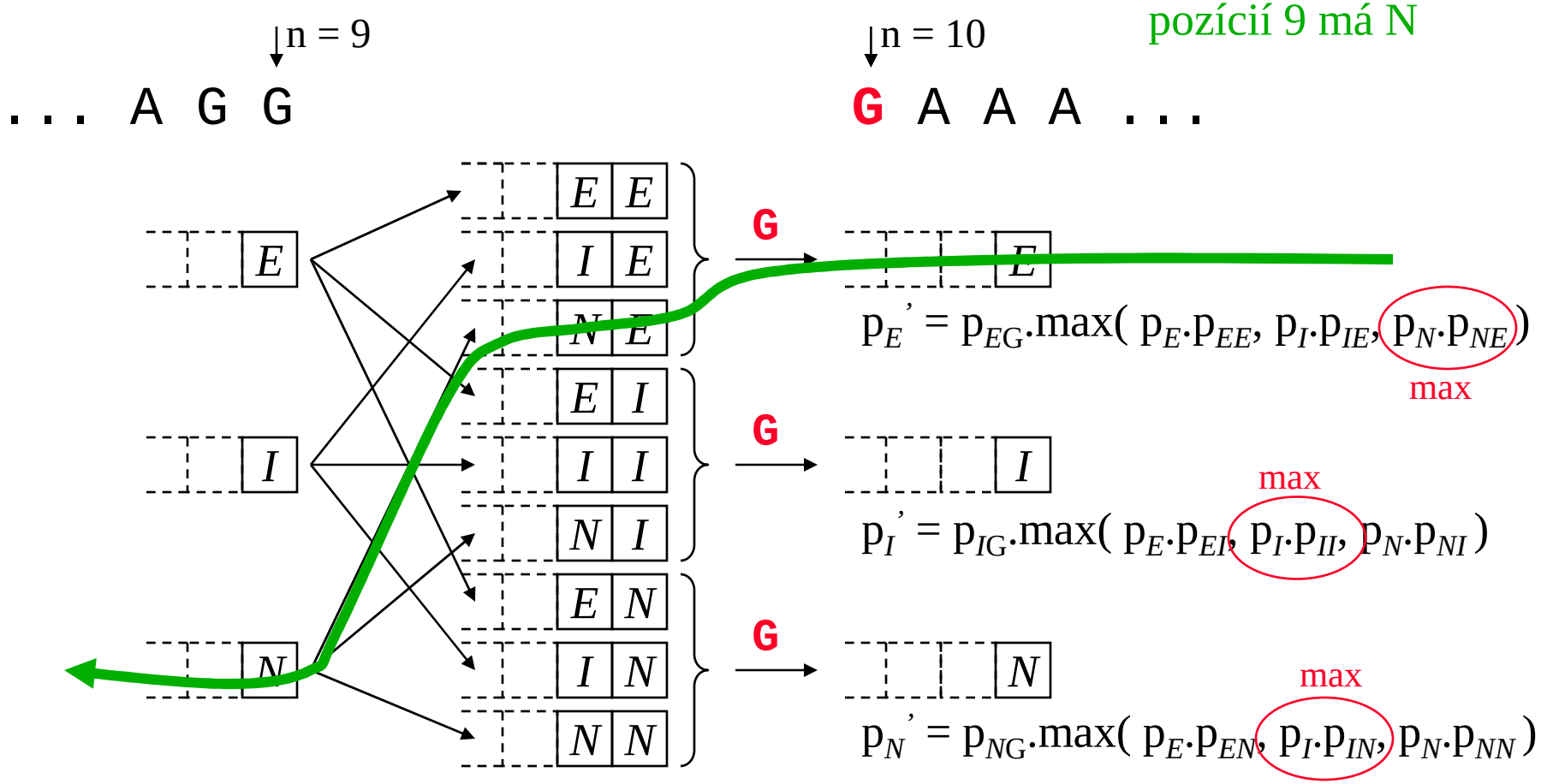


Viterbiho algoritmus

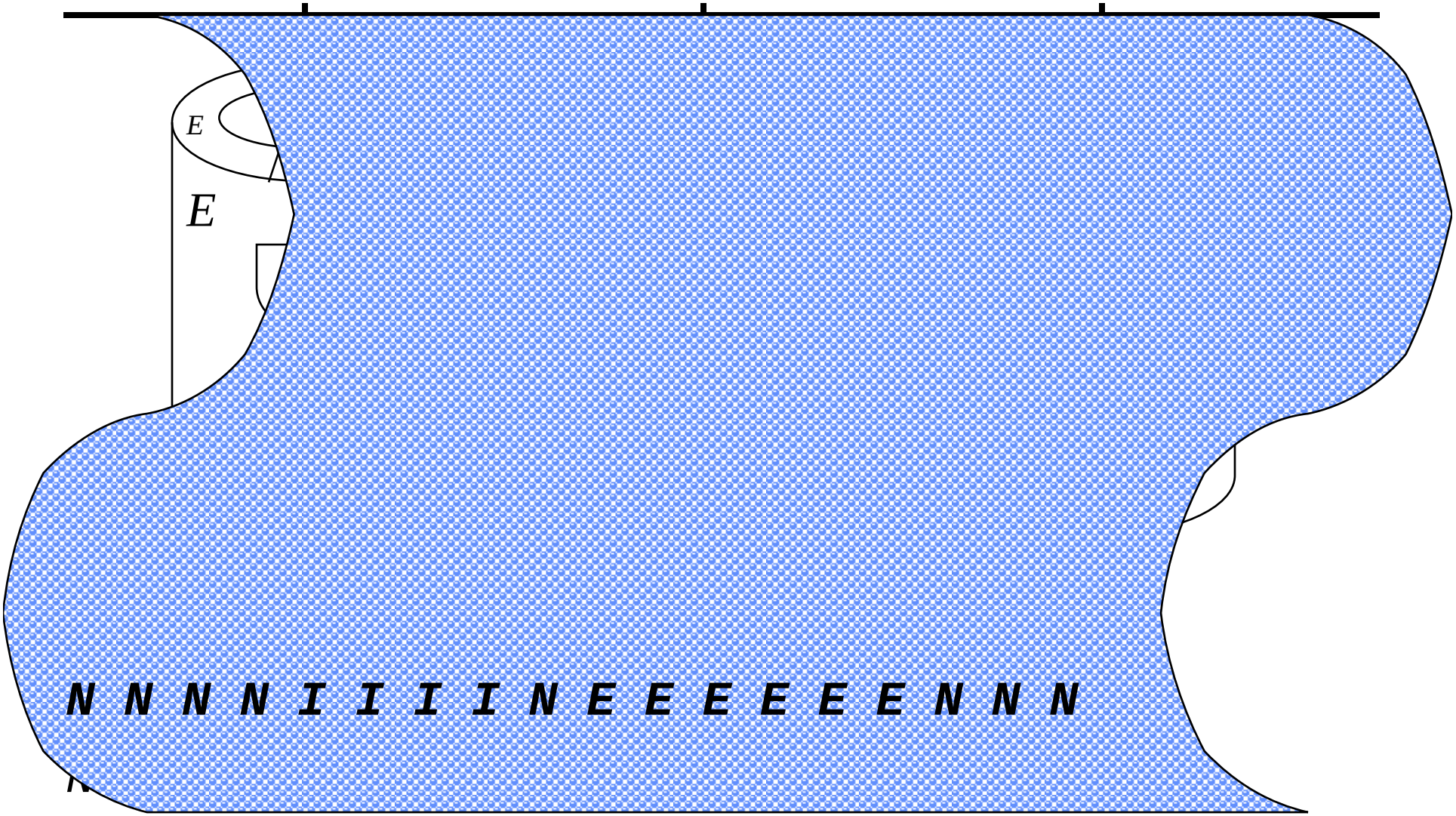
Treba samozrejme vypočítať nielen pravdepodobnosť ale aj postupnosť.

Na to stačí ísť spätne vždy po tej vetve ktorá bola najpravdepodobnejšia

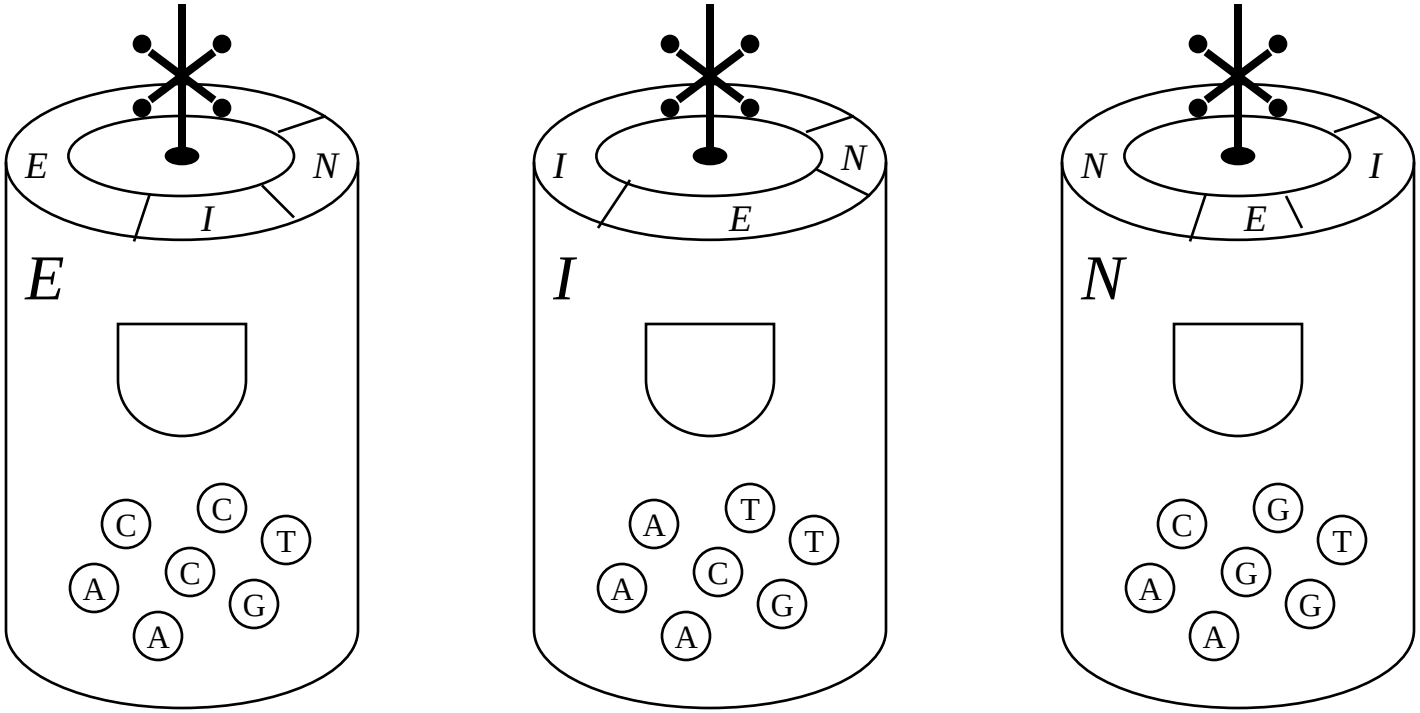
ak hľadaná postupnosť má na pozícii 10 E, potom na pozícii 9 má N



Týmto spôsobom môžeme pre náš prípad odhadnúť, že postupnosť urien bola nasledovná:



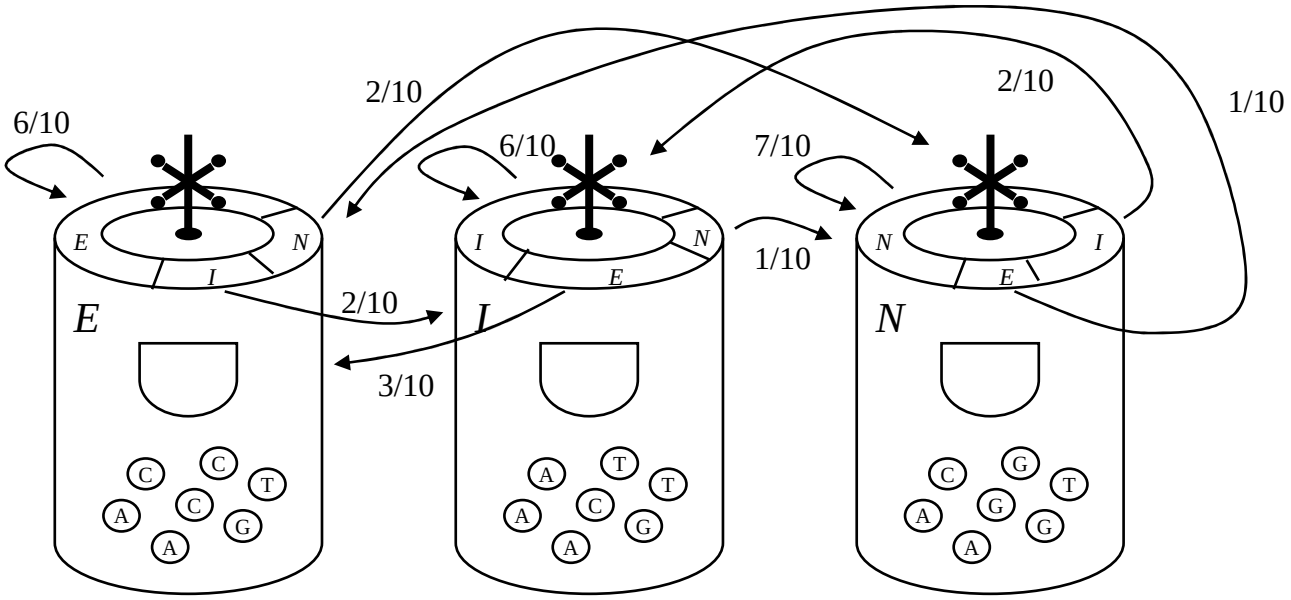
Nájdená postupnosť síce nemusí nezodpovedať skutočnosti, ale je najpravdepodobnejšou možnosťou



N N N N I I I I N E E E E E E N N N

N N N N N N I I I I I E E E E E E N

G A G T T G A G G G A A A A A G C G



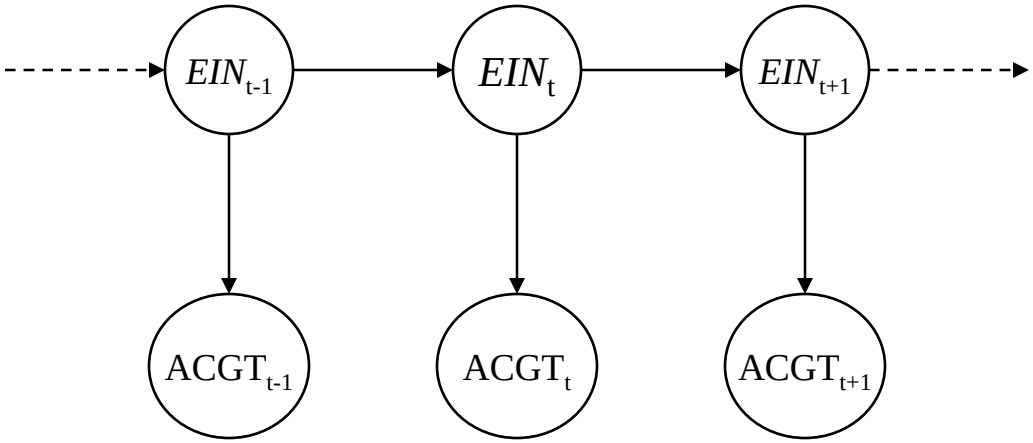
p(A) = 2/7	p(A) = 3/7	p(A) = 2/7
p(C) = 3/7	p(C) = 1/7	p(C) = 1/7
p(G) = 1/7	p(G) = 1/7	p(G) = 3/7
p(T) = 1/7	p(T) = 2/7	p(T) = 1/7

Ale ako vôbec získame Markovov model ?

Nuž z malej vzorky dát a predpokladáme, že je to dobrý odhad

Napr. ručne určíme 1% DNA a zvyšok spočítame

HMM ako Bayesova sieť



HMM je založený na zjednodušení, že viditeľná manifestácia ako i zmena skrytého stavu systému sú závislé iba na jeho aktuálnej hodnote (a nie napr. na predchádzajúcich hodnotách)

HMM rozšírenia

- závislosť od predchádzajúcich N hodnôt
- automatické generovanie vnútorných stavov
- spojité HMM

Ďakujem za pozornosť !

Skryté Markovove modely

Andrej Lúčny

KAI FMFI Bratislava & MicroStep-MIS

andy@microstep-mis.com

http://www.microstep-mis.com/~andy