University of BRISTOL

Fluctuations, backtracking and proofreading in Transcription

Tanniemola B Liverpool
Department of Mathematics
University of Bristol

KITP, May 2011

University of BRISTOL

Acknowledgements

In collaboration with

M Voliotis

(Mathematics, Bristol)

N Cohen and C Molina-París

(Applied Maths & Computing, Leeds)

MRC Medical Research Council

EPSRC Engineering and Physical Sciences Research Council

THE ROYAL SOCIETY

Plan

1. Fluctuations, pauses and backtracking in transcription
    1. Motivation - stochastic gene expression
    2. **RNAP**: microscopics of transcription
    3. Pauses and bactracking $\Rightarrow$ single molecule experiments
    4. Distribution of elongation times
    5. Integrated model of transcription $\Rightarrow$ RNA population dynamics
    6. A model for proofreading in transcription
2. Conclusions and outlook

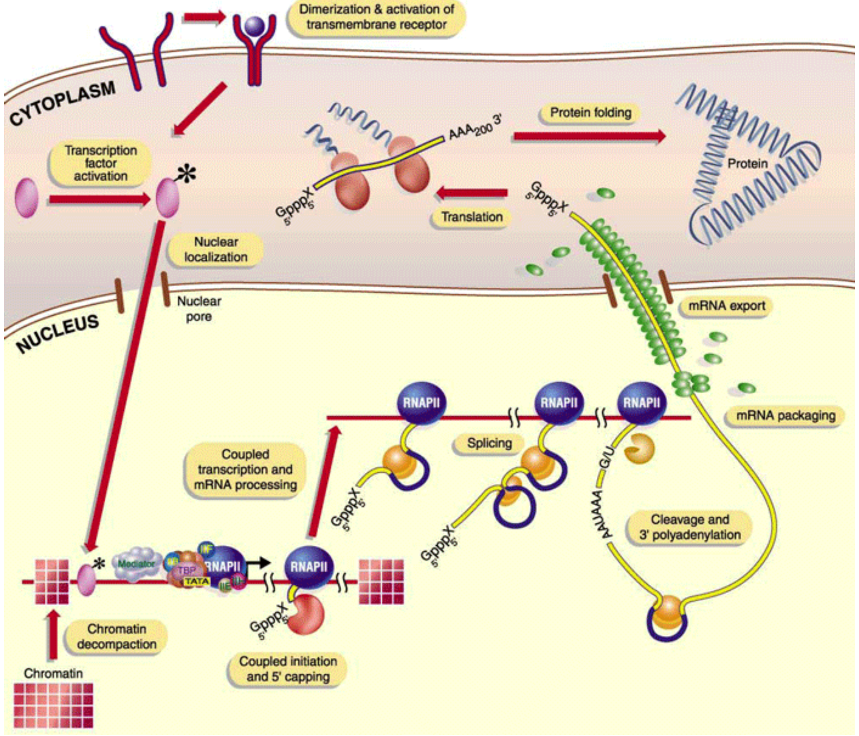
Gene expression

Central Dogma

DNA ⇒ RNA ⇒ protein
Transcription Translation

Gene expression = set of reactions that control quantities of gene products (proteins)

Differences in gene expression is thought to control most aspects of cellular behaviour ⇒ phenotype



However even within related cells of the same phenotype there are variations in the levels of expressed genes

Variations in gene expression

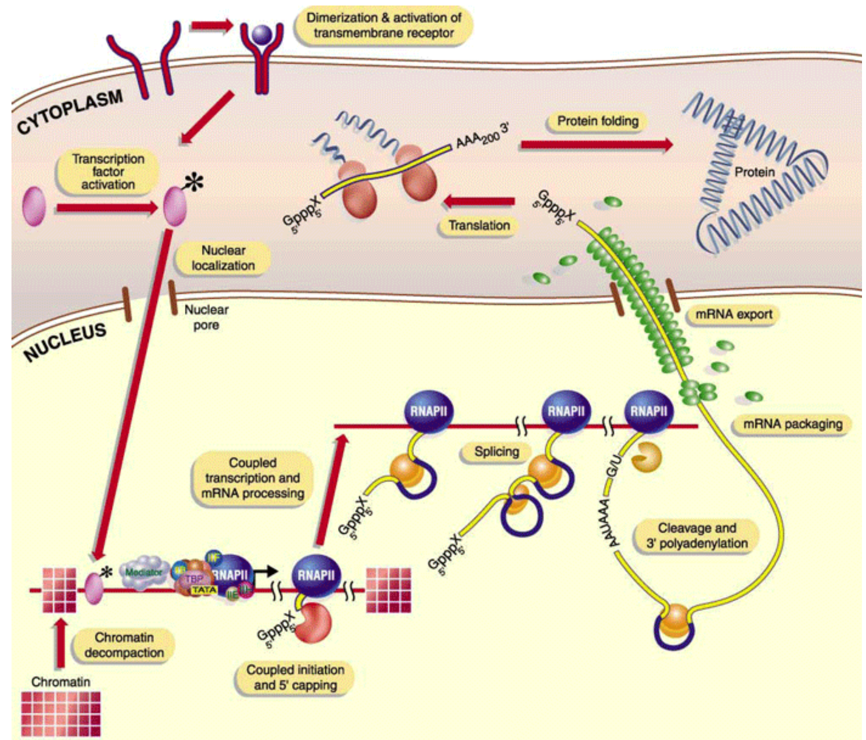
Standard ‘systems’ view

Microarray, Northern blot, RT-PCR experiments

Bulk RNA/protein levels from ‘homogeneous’ population extracts

Gives impression of gene expression as continuous smooth process

BUT highly irregular ⇒ periods of activity and inactivity



Fluctuations in space and time

Important : differences in transcription/translation between related cells ⇒ differentiation, disease, …

University of BRISTOL

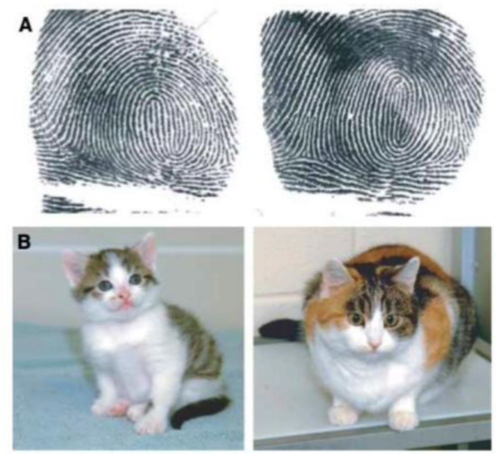
Stochastic Gene expression

Fluctuations ⇒ quantify variability of cellular behaviour

Fluctuations can be due to intrinsic or extrinsic factors (somewhat vague operational definition)

New observational techniques

Raser & O'Shea (2005)



Twins

Texas A&M

Copy Cat

Genetic Savings and Clone inc.

Sources of fluctuations

Macroscopic fluctuations in environment

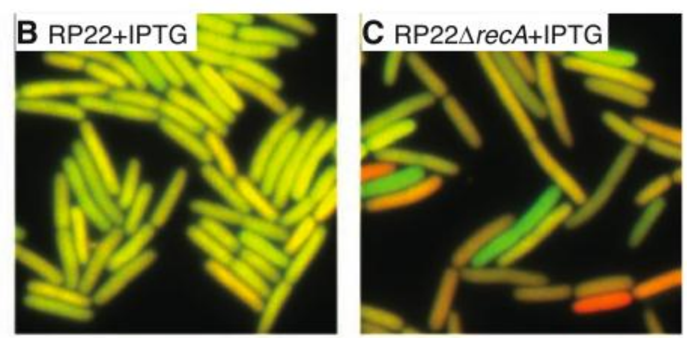
Variability of internal state of cell

Genetic mutation

Small numbers of macromolecules involved in gene regulation/expression

Stochastic nature of production and degradation of RNA transcription products

E-Coli



Elowitz et al (2002)

Modelling Fluctuations

Chemical Master equations

Discrete reactants , probabilistic chemical reactions

Transcription and translation as one-stage processes

Poisson population statistics $\frac{\sigma^2_{\mathrm{mRNA}}}{\mu_{\mathrm{mRNA}}} = 1$

- True for some experiments

Zenklusen et al, Nature Struc. Mol. Biol. (2008)

However production/degradation of proteins/mRNA are multi-stage processes

- BUT recent expts. tracking RNA expression levels in single cells see non-Poissonian fluctuations

$$\frac{\sigma^2_{\mathrm{mRNA}}}{\mu_{\mathrm{mRNA}}} > 1$$

Golding et al , Cell (2005)

Raj et al , PLOS Biology (2006)

Transcription ⇒ 3 main stages

Initiation

Elongation

Termination

How consistent is this with simple exponential birth/death Markov processes which give Poisson statistics?

- Alternative processes ?

Chubb and Liverpool, *Current Opinion in Genetics and Dev. (2010)*

Chemical master equations

Prob of *n* molecules **Production rate** **Degradation rate**

$$\partial_t P_n = \lambda_+ P_{n-1} - \lambda_- \, n \, P_n - \lambda_+ P_n + \lambda_- (n+1) \, P_{n+1}$$

- Single timescales
- Steady state distribution is Poisson

$$P_n = \frac{1}{n!} \mu^n e^{-\mu} \qquad \lambda_- = \lambda_+ \, \mu \, e^{-\mu}$$

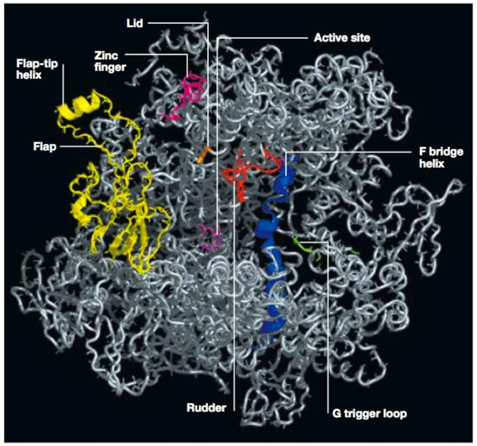
$$\langle n \rangle = \mu \quad ; \quad \sigma^2 \equiv \langle n^2 - \langle n \rangle^2 \rangle = \mu$$

- COMPLEX

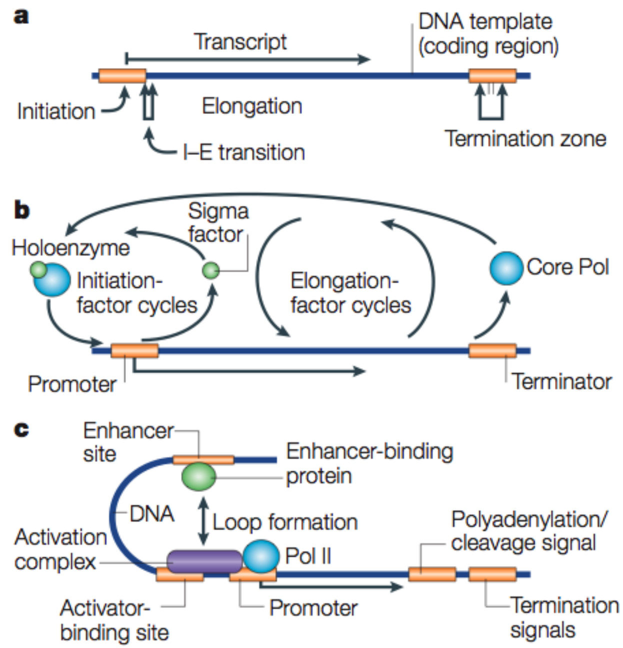
$$\lambda_i (\text{everything else})$$

Transcription: DNA ⇒ mRNA

RNA polymerase (~ 150 KDa)



Benoit Coulombe (Montreal)

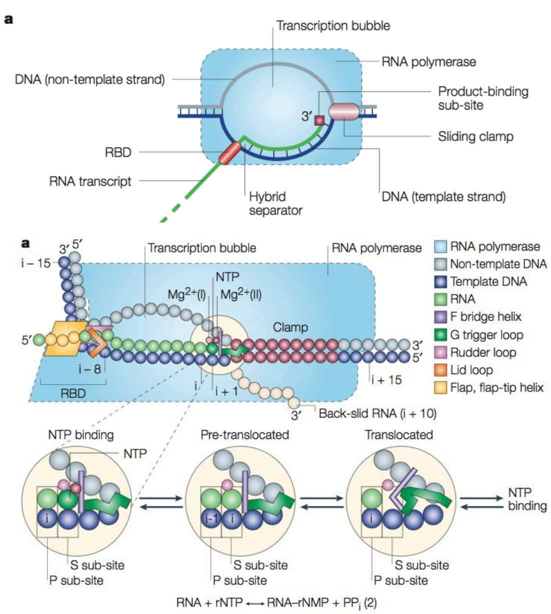


Initiation

Elongation

Termination

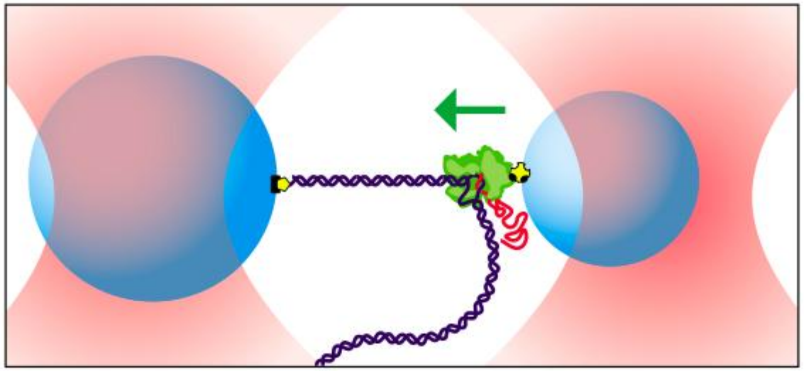
Greive & von Hippel, (2005)



- Nobel prize for Medicine 1965, F. Jacob, J. Monod and A. Lwoff (prokaryotic)
- Nobel prize for Chemistry 2006, Roger Kornberg (eukaryotic RNAP)

Single molecule experiments

Elongation phase

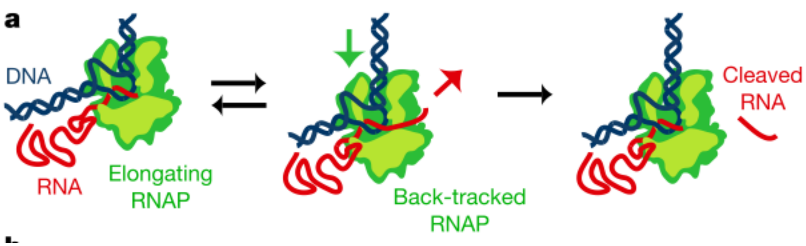


Optical tweezer experiments

Pauses of E-Coli RNAP observed in-vitro
Operationally experiments classified into 'short' (< 20s) and 'long' pauses
Backtracking of E-Coli RNAP observed during long pauses.

Shaevitz et al, Nature, **426**, 684 (2003)

Backtracking ⇒ rearwards motion of RNAP along DNA template in direction opposite to normal elongation



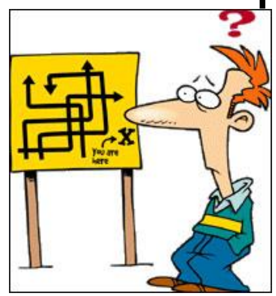
Shift of transcription bubble, however DNA-RNA hybrid remains in register, 3' end of RNA moves away from active site

RNAP transcribes with remarkably low error rate in-vivo
Has recently been suggested that back-tracking is involved in proofreading

Broad non-exponential temporal distribution of 'long' pauses

Pipe dreams …

Dynamical question - To understand how the **complex** behaviour of cells controlled by the expression of their genes **emerges** from their components.



????????????????

Some much smaller goals - using simple physical models …

•Can we understand *something* about the origin of intrinsic fluctuations from the **bottom up** ?

• Can we link **single molecule** behaviour to gene expression experiments at the **cellular** level ?

A theoretical model of transcription

Based on biochemical mechanism proposed by Yager and von Hippel, *Biochemistry*, 30, 1097 (1991)

✓ Initiation - Poisson process

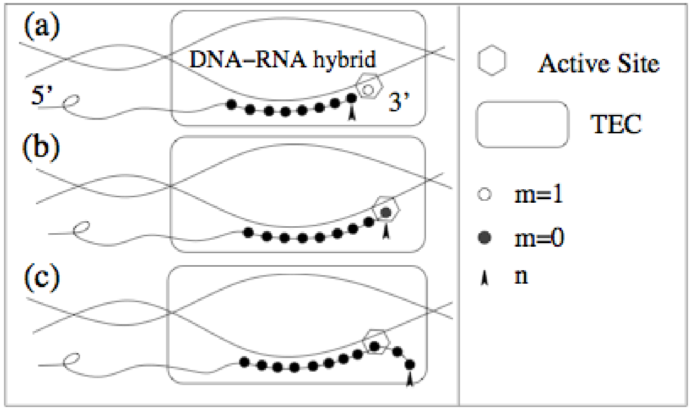
Elongation phase

- Position of last transcribed nucleotide $\Rightarrow n$ (size of transcribed mRNA)

$$0 < n < N$$

- Position of polymerase active site relative to $n \Rightarrow m$

$$-n < m < 1$$



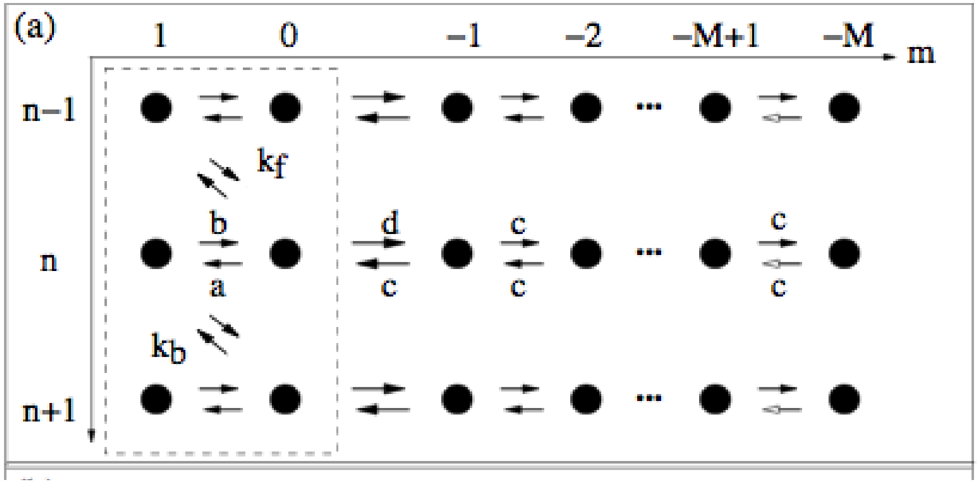
$m=0$ pre-translocated
$m=1$ post-translocated
$m<0$ backtracked

- Polymerisation (+ new nucleotide) only from post-translocated state
- Depolymerisation (- new nucleotide) only from pre-translocated state

$$(n, m = 0) \rightleftharpoons (n, m = 1)$$
$$(n, m = 1) \rightleftharpoons (n + 1, m = 0)$$

- Initiation $(n=0)$ and termination $(n=N)$

Model of transcription elongation



Schematic of state transitions

Backtracking restricted to *-M >-n* (hairpins,cleavage,…)

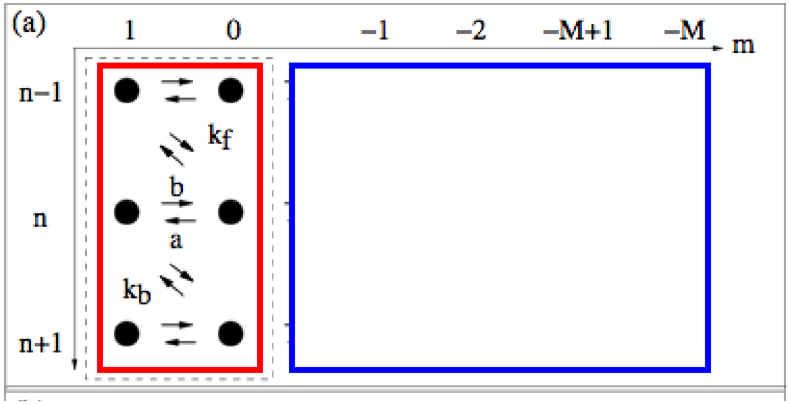
We want to find the statistics of the elongation time

(time to get from *n=0* to *n=N*)

First passage problem

Voliotis et al, Biophys. J, **94**, 334 (2008)

Model A: no backtracking



Schematic of state transitions

Backtracking restricted to $-M > -n$ (hairpins,cleavage,…)

We want to find the statistics of the elongation time (time to get from $n=0$ to $n=N$)

Model A :translocation limited polymerisation

$$\frac{\partial P_{n,0}}{\partial t} = k_f P_{n-1,1} + b\, P_{n,1} - (k_b + a)\, P_{n,0}$$

$$\frac{\partial P_{n,1}}{\partial t} = k_b P_{n+1,0} + a\, P_{n,0} - (k_f + b)\, P_{n,1}$$

polymerisation $\Rightarrow k_f$

depolymerisation $\Rightarrow k_b$

forward translocation $\Rightarrow a$

backward translocation $\Rightarrow b$

Reflecting BC $(n=0)$

Absorbing BC $(n=N)$

Mean field approximation: $k_f, k_b \ll a, b$

$$\frac{\partial}{\partial t}\mathcal{P}_n = p_-\mathcal{P}_{n+1} + p_+\mathcal{P}_{n-1} - (p_+ + p_-)\mathcal{P}_n \qquad p_+ \approx \frac{k_f a}{a+b} \qquad p_- \approx \frac{k_b b}{a+b}$$

biased random walk

Voliotis et al, Biophys. J, **94**, 334 (2008)

First passage problem easy

University of BRISTOL

Model A: no backtracking

We want to find the statistics of the elongation **time** (**time** to get from *n=0* to *n=N*)

- Under normal conditions, polymerisation overwhelmingly favoured over depolymerisation

$$K = \frac{p_-}{p_+} \ll 1$$

- Mean elongation **time**, μ_t and variance σ_t^2

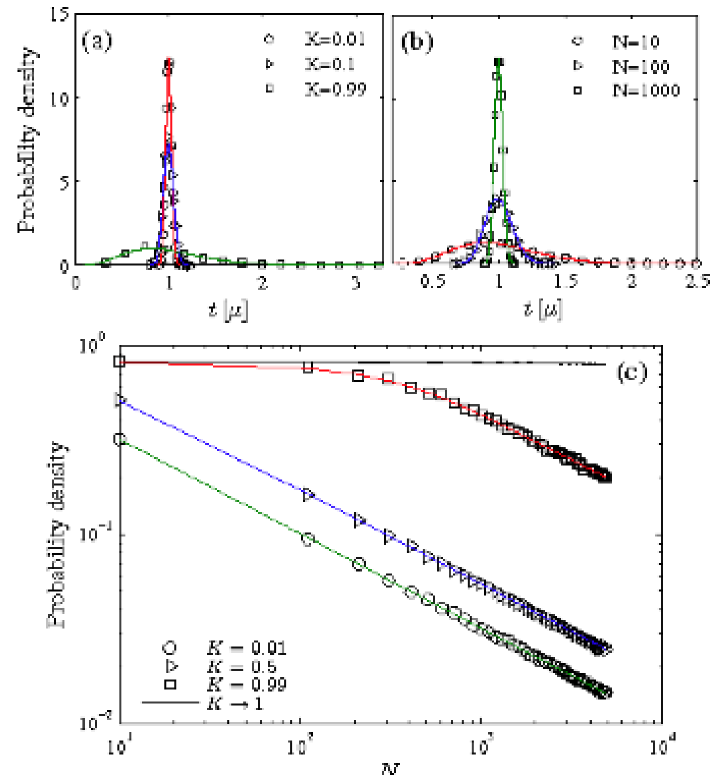
$$\mu_t = \langle t \rangle = \frac{N}{p_+} + K\frac{(N-1)}{p_+} + \mathcal{O}(K^2),$$

$$\sigma_t^2 = \langle t^2 \rangle - \langle t \rangle^2 = \frac{N}{p_+^2} + K\frac{(4N-4)}{p_+^2} + \mathcal{O}(K^2).$$

- For large N, approaches a Gaussian
- $\Rightarrow$ Fluctuations smaller than exponential process ($\Rightarrow$ mRNA **Population** sub-Poisson)

$$\sigma_t^2/\langle t \rangle^2 = 1/N \to 0$$

- Compare with Monte Carlo simulations of full model



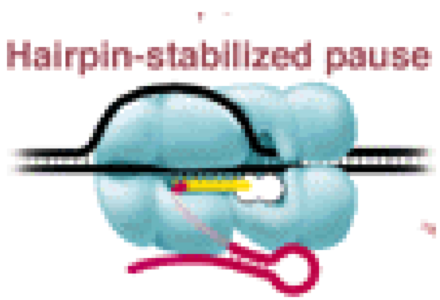
M Voliotis et al, Biophys. J, **94**, 334 (2008)

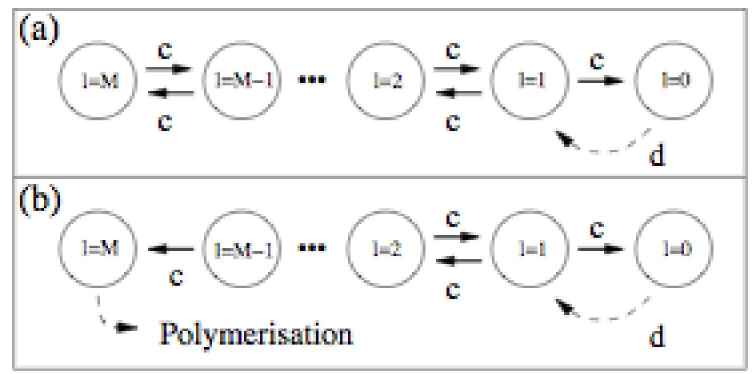
A model for backtracking pauses

Before we include pauses in our dynamics we need a model for backtracking pauses themselves.

Can we explain the broad distribution of pause durations?

- A pause starts when the TEC enters the state $m=-1$ from the state $m=0$
- From $m=-1$, TEC hops across backtracked states with hopping rate c
- Because of hairpins, RNA-DNA interactions backtracking is restricted and proceeds up to $m=-M$ for $M < n$ and up to $m=-n$ for $n > M$





Dynamics of backtracking pauses

Use new variable $l=-m$ where $1 \le l \le M$

Probability of finding polymerase starting at $l=1$ at $t=0$ at position l at time t is $P(l,t)$

$$\frac{\partial}{\partial t} P(l,t) = c\, P(l+1,t) + c\, P(l-1,t) - 2c\, P(l,t)$$

- Obtain the time to terminate a pause by returning to state $l=0$
- First passage problem in a finite domain with reflecting BC's

$$\begin{aligned} c\, P(M,t) &= c\, P(M+1,t) && \text{(reflecting)} \\ P(0,t) &= 0 && \text{(absorbing)} \end{aligned}$$

- Probability flux to the state $l=0$ $\Leftrightarrow$ probability of exiting the pause at time t

$$F(0,t) = cP(1,t)$$

- Can do a similar calculation when there is transcript arrest (absorbing BC's both sides)

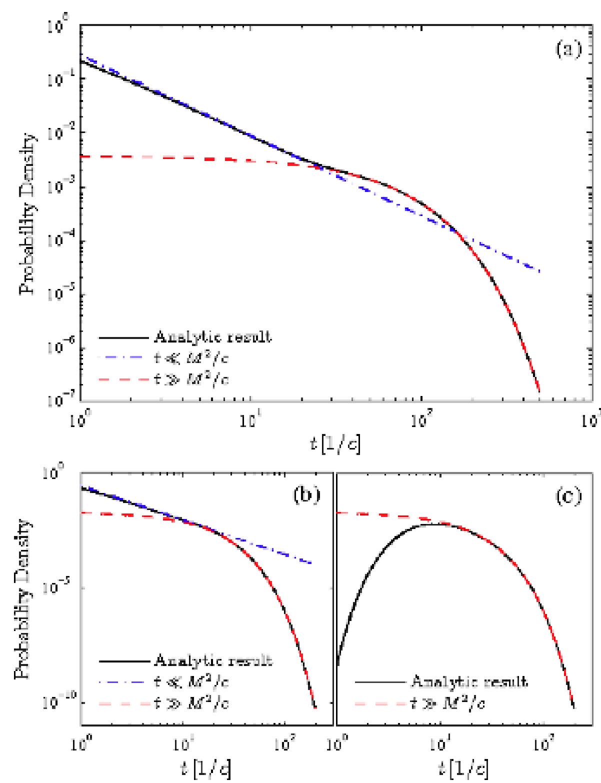
University of BRISTOL

Dynamics of backtracking

Series & asymptotic expansion of θ_1 function used to obtain limiting behaviour.

$$\mathcal{P}(t) = \frac{(1+M)}{\sqrt{\pi}\sqrt{c}\,t^{3/2}} \sum_{n=-\infty}^{n=+\infty} (-1)^n e^{-\frac{(1+M)^2}{ct}\left(n-\frac{1}{2M}\right)^2} \left(n - \frac{1}{2M}\right)$$

$$\mathcal{P}(t) \approx \begin{cases} \dfrac{1}{2\sqrt{\pi}\sqrt{c}t^{3/2}} & , \dfrac{1}{c} \ll t \ll \dfrac{M^2}{c}, \\ \dfrac{\pi c \sin\left(\dfrac{\pi}{2(M+1)}\right)}{(1+M)^2} e^{-\frac{c\pi^2}{4(1+M)^2}t} & , t \gg \dfrac{M^2}{c}. \end{cases}$$



Mean pause duration $\langle t \rangle = \dfrac{M}{c}$

Power law behaviour for $t << M^2/c$
⇔ heavy-tailed distribution
observed by Shaevitz et al.

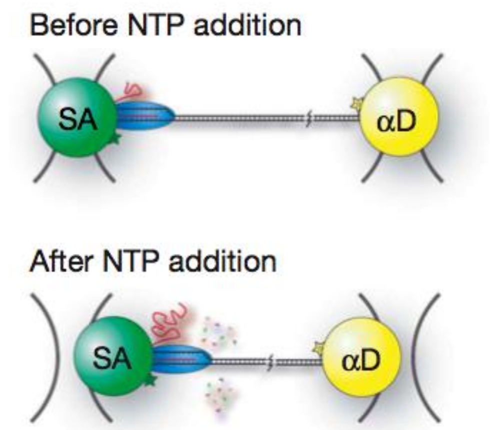
M Voliotis et al, Biophys. J, **94**, 334 (2008)

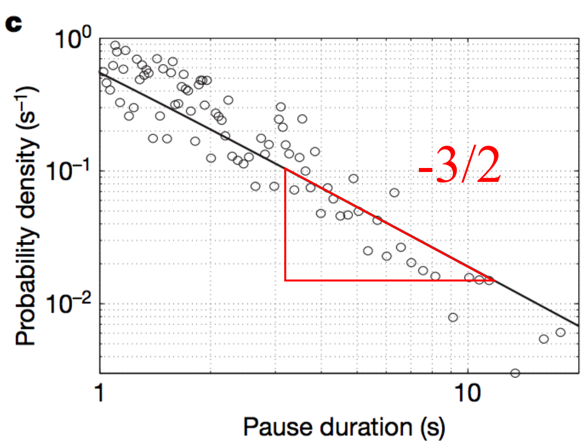
University of BRISTOL

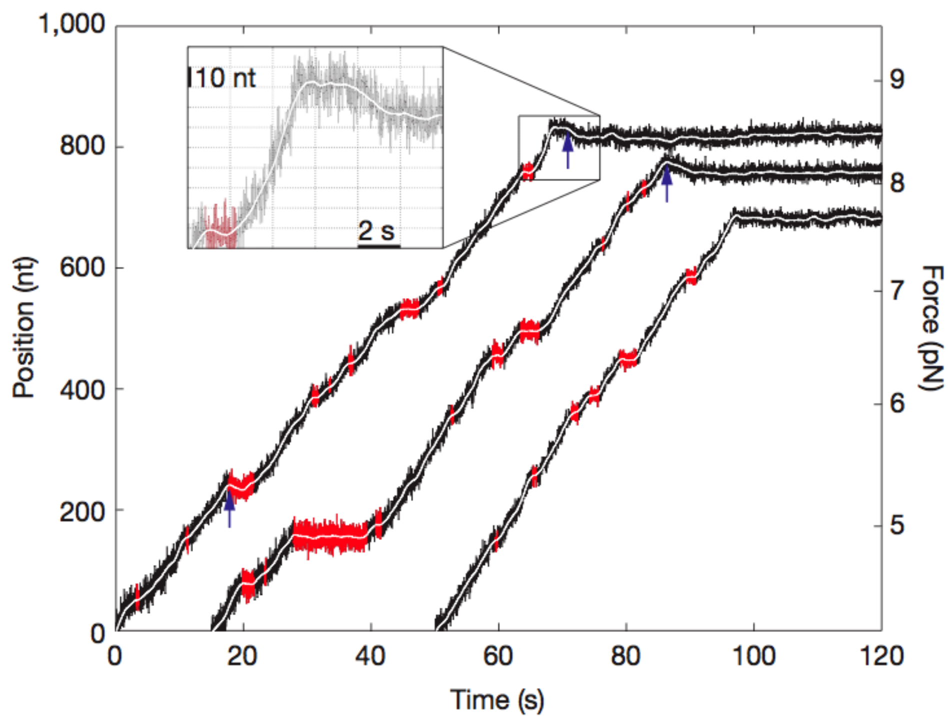
Single molecule experiments

Recent quantitative experiments of pause distributions of eukaryotic RNAP II

Galburt et al, Nature (2007)

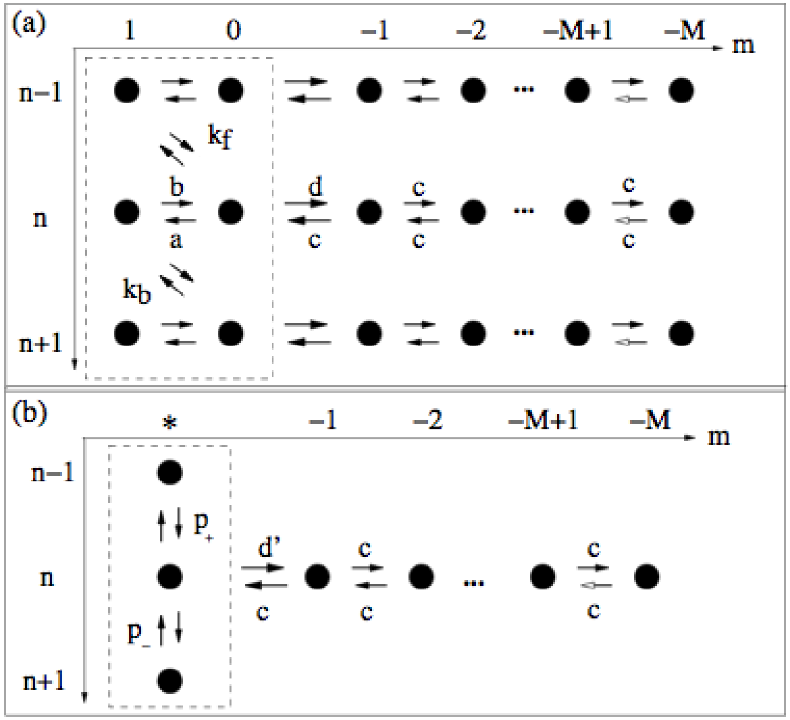






Power law with $t^{-3/2}$?

Model B: transcription with pauses



Now we are in the position to study a model for elongation with pauses

Macroscopic observables

- o Number of pauses δ over a DNA template of length N
- o Sum of lifetime of pauses relative to time spent on active polymerisation

$$\frac{N}{p_+} \gg \delta \frac{M}{c}$$

Pauses negligible - model A

$$\frac{\delta}{N} = \frac{d\frac{a}{a+b}}{d\frac{a}{a+b} + p_+ + p_-} = \frac{d'}{d' + p_+ + p_-}$$

$$\frac{N}{p_+} \ll \delta \frac{M}{c}$$

Pauses dominate elongation time

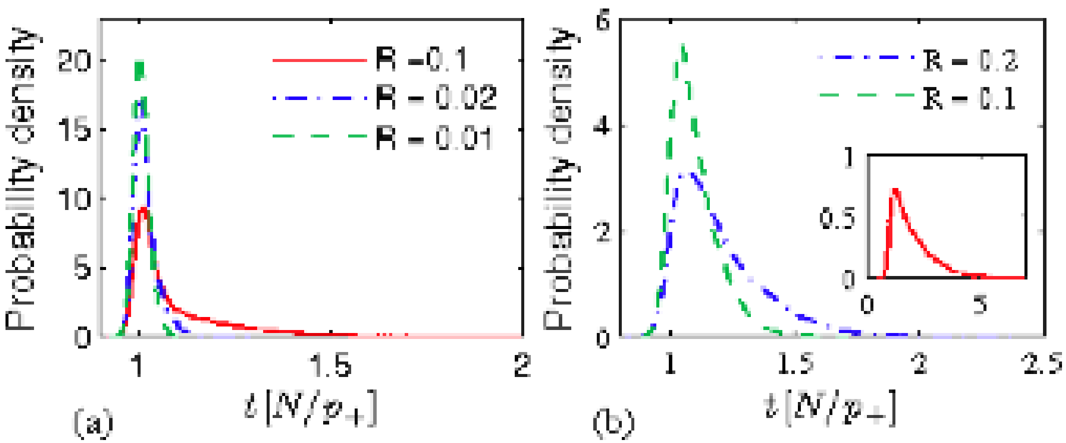
Model B: transcription with pauses

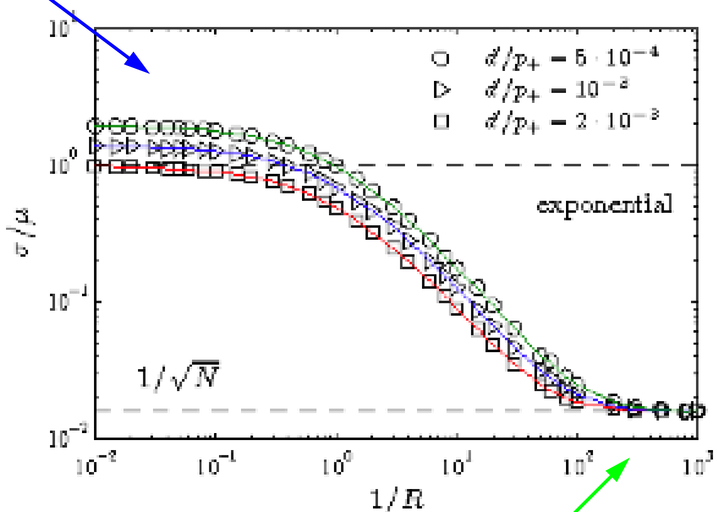
When translocation not the rate-limiting step $p_+ \gg d$

If polymerisation favoured $p_+ \gg p_-$

Define $R = d' \frac{M}{c}$

$R \gg 1$ Pauses dominate



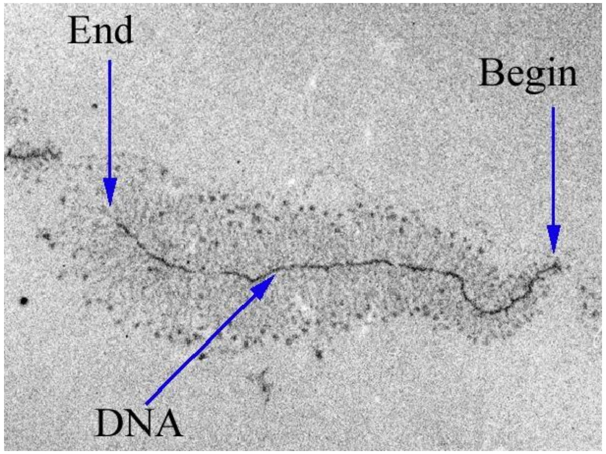


As R ↑, distribution becomes broader

$R \ll 1$ Pauses negligible

Integrated model of mRNA production

We are really interested in mRNA levels in the cell ⇒ Need to consider production (transcription) and degradation



Initiation + model B + termination(fast) + degradation

k_i k_d

Many initiations can occur in the time to produce a single RNA ⇒ multiple occupation of DNA template by TECs moving in tandem

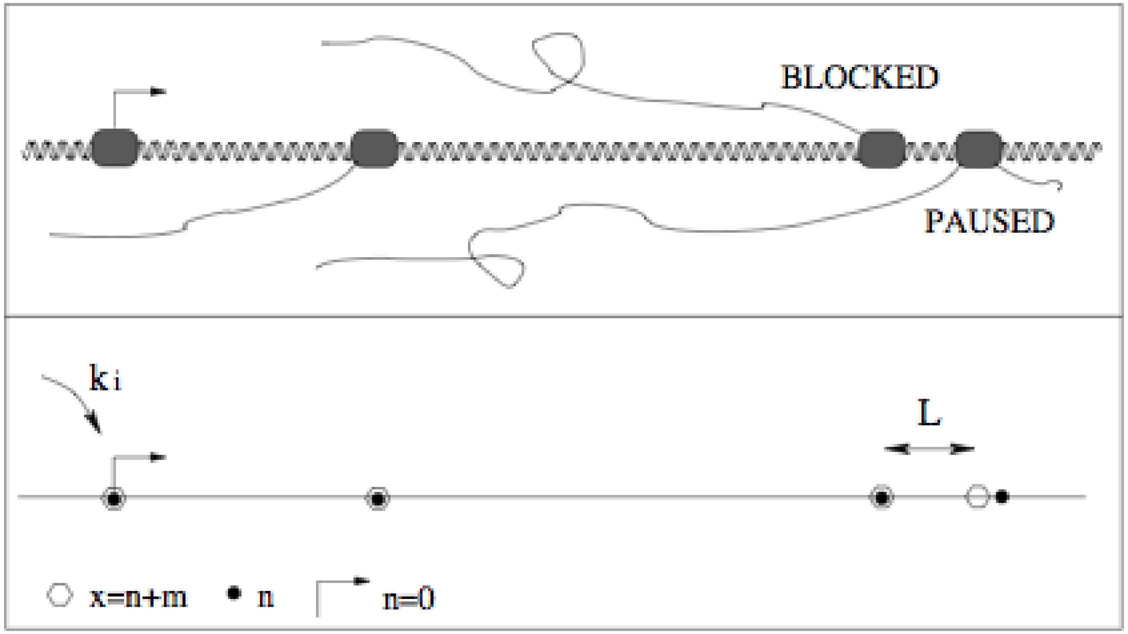
S.L. Gotta et al, *J. Bacteriol.*, **173**, 6647 (1991)

Each polymerase synthesizing a nascent mRNA

'Christmas Tree'

Polymerases cannot get too close because of additional work to deform the DNA helix

Model of mRNA production



Minimum (exclusion) distance between TEC's

$$L \ll N$$

Active site of a TEC with *(n,m)* is at position

$$x = n + m$$

$$|x_1 - x_2| < L$$

Relevant timescales

Time for transcription initiation $\tau_1 = 1/k_i$

Time needed by the TEC to transcribe *L* nucleotides $\tau_2 \approx L/p_+$

The mean time of a backtracking pause $\tau_2 = M/c$

Study numerically using MC simulations

University of BRISTOL

Model of mRNA production

When initiation is the rate limiting step

$\tau_1 \gg \tau_2, \tau_3$ (approx Poisson)

$$\mu_{\mathrm{mRNA}} = \sigma^2_{\mathrm{mRNA}}$$

Polymerisation is the rate limiting step

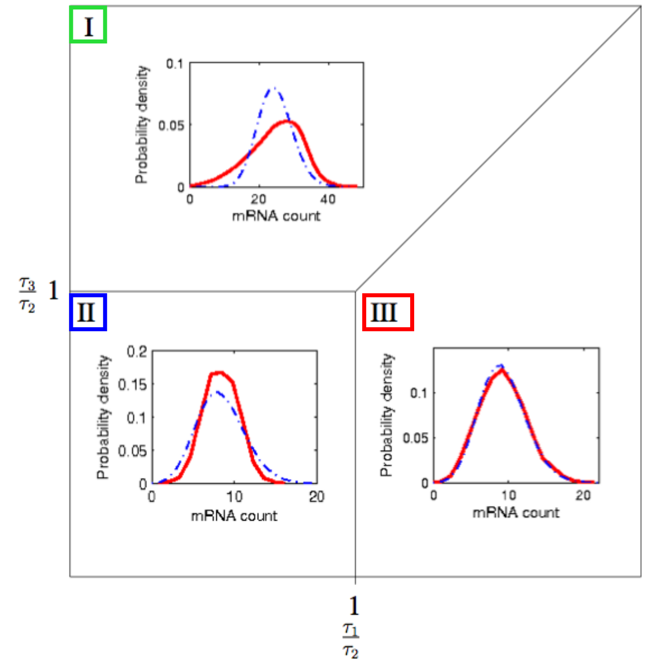
$\tau_2 \gg \tau_1, \tau_3$ (Sub-Poisson)

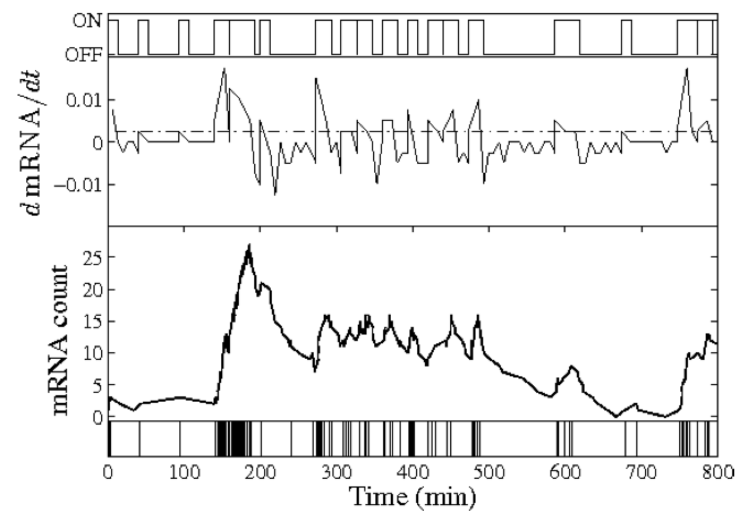
$$\mu_{\mathrm{mRNA}} > \sigma^2_{\mathrm{mRNA}}$$

Long pauses dominate transcription

$\tau_3 \gg \tau_1, \tau_2$ (Super-Poisson)

$$\mu_{\mathrm{mRNA}} < \sigma^2_{\mathrm{mRNA}}$$





Bursts of RNA production due to rare and long-lived pauses of TECs acting as congestion points

Switches between states of high and low mRNA production

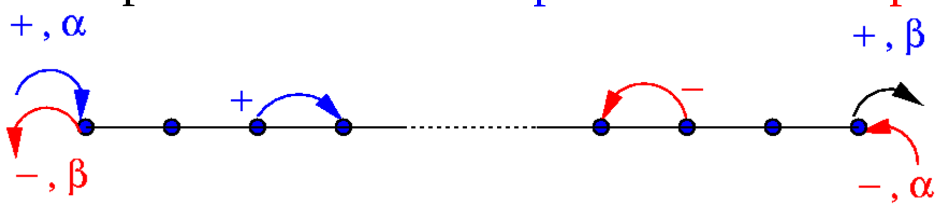
Relation to ASEP?

The Asymmetric Simple Exclusion Process

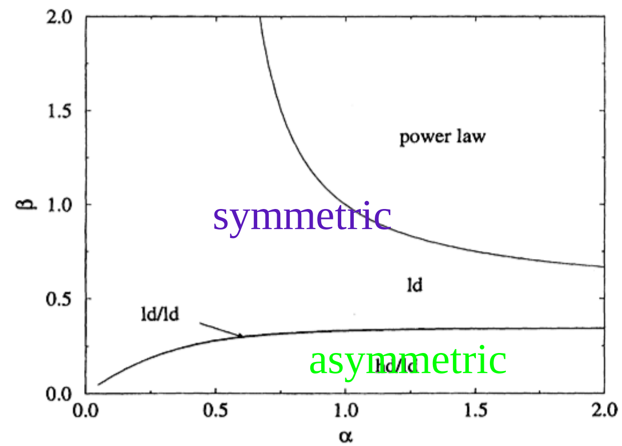
B. Derrida, *Phys. Rep.*, **301**, 65 (1998).

1-d non-equilibrium model been subject of much study

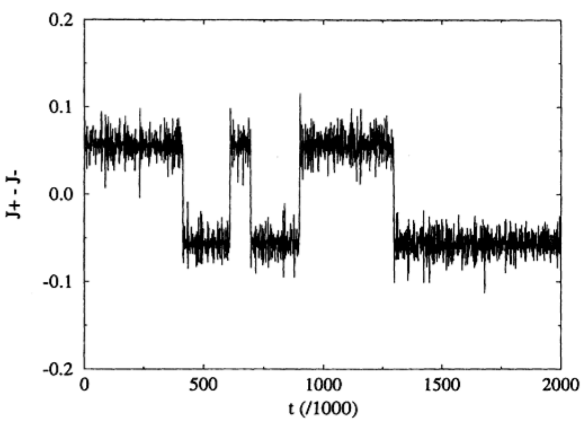
Two species model + particles ⇒ - particles ⇐



Symmetry breaking transition



Switching between left and right moving states



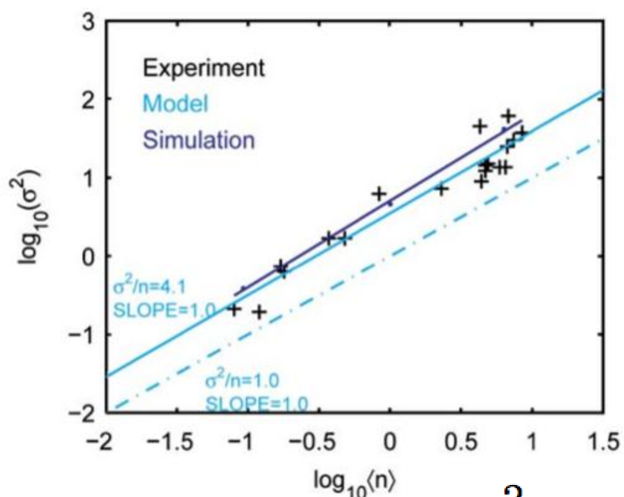
M.R. Evans et al, *PRL*, **74**, 208 (1995)

mRNA populations in E-Coli

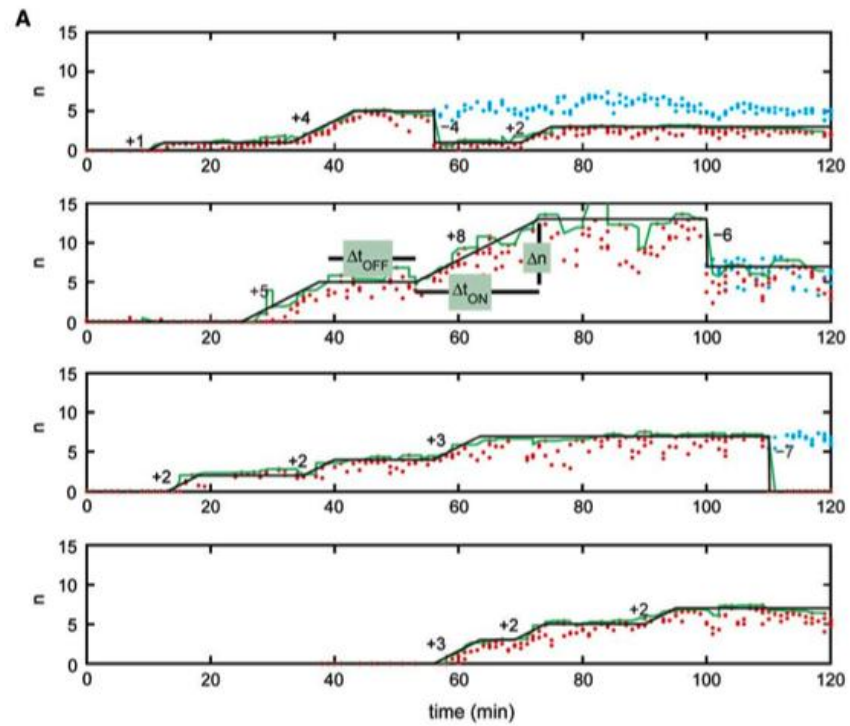
Direct measurement of mRNA populations

mRNAs with 96 MS2 binding site

MS2-GFP fusion protein



$$\frac{\sigma^2_{\rm mRNA}}{\mu_{\rm mRNA}} \approx 4.1$$



$$\Delta t_{\rm OFF} \approx 37min$$
$$\Delta t_{\rm ON} \approx 6min$$

Golding et al, Cell, **123**, 1025 (2005)

Numbers

Polymerisation rate: 25bp/s-50 bp/s Shaevitz et al (2005)

Depolymerisation rate : two orders magnitude lower

mRNA degradation rate: 0.014/min Golding et al (2005)

Rate of entering back-tracking state
= rate of NTP misincorporation: 1 error/kbp

Rate of backtracking diffusion c: 0.1-0.2 bp/s

Initiation rate $\approx$ 0.1/s $\Rightarrow$ bursting .

$$t_{\mathrm{ON}} \approx \mathrm{min}$$
$$t_{\mathrm{OFF}} \approx 10\mathrm{min}$$

But what about stochastic dynamics of transcription factor binding ?

Backtracking & error correction …

Thermodynamic fraction of misincorporated nucleotides $\sim 10^{-2} - 10^{-3}$

Observed transcriptional error fraction $\sim 10^{-5}$

Kinetic proofreading : there must exist an active mechanism of error correction

Hopfield, Proc. Natl. Acad. Sci. (1974)
Ninio, Biochimie 57, 587 (1975)

$$E + S_c \underset{k_c}{\overset{k'_c}{\rightleftharpoons}} ES_c \overset{\alpha_c}{\rightarrow} E + P_c$$

Requires a branching process

$$E + S_w \underset{k_w}{\overset{k'_w}{\rightleftharpoons}} ES_w \overset{\alpha_w}{\rightarrow} E + P_w$$

Possible mechanism - circumstantial evidence …

Backtracking
mRNA cleavage (Gre, TFIIS)

Limiting **error fraction** $\mathcal{E}_0 = \frac{k_c}{k_w} = e^{-\beta \Delta G}$

Backtracking & error correction …

Thermodynamic fraction of misincorporated nucleotides $\sim 10^{-2} - 10^{-3}$

Observed transcriptional error fraction $\sim 10^{-5}$

Kinetic proofreading : there must exist an active mechanism of error correction

Hopfield, Proc. Natl. Acad. Sci. (1974)
Ninio, Biochimie 57, 587 (1975)

$$E + S_c \underset{k_c}{\overset{k'_c}{\rightleftharpoons}} ES_c \overset{\alpha_c}{\rightarrow} ES^*_c \overset{\beta_c}{\rightarrow} E + P_c$$
$$\downarrow l_c$$
$$E + S_c$$

$$E + S_w \underset{k_w}{\overset{k'_w}{\rightleftharpoons}} ES_w \overset{\alpha_w}{\rightarrow} ES^*_w \overset{\beta_w}{\rightarrow} E + P_w$$
$$\downarrow l_w$$
$$E + S_w$$

$$\mathcal{E} = \frac{k_c}{k_w}\frac{l_c}{l_w} = \mathcal{E}_0^2 \qquad \alpha_j, \beta_i \ll k_j, l_i$$

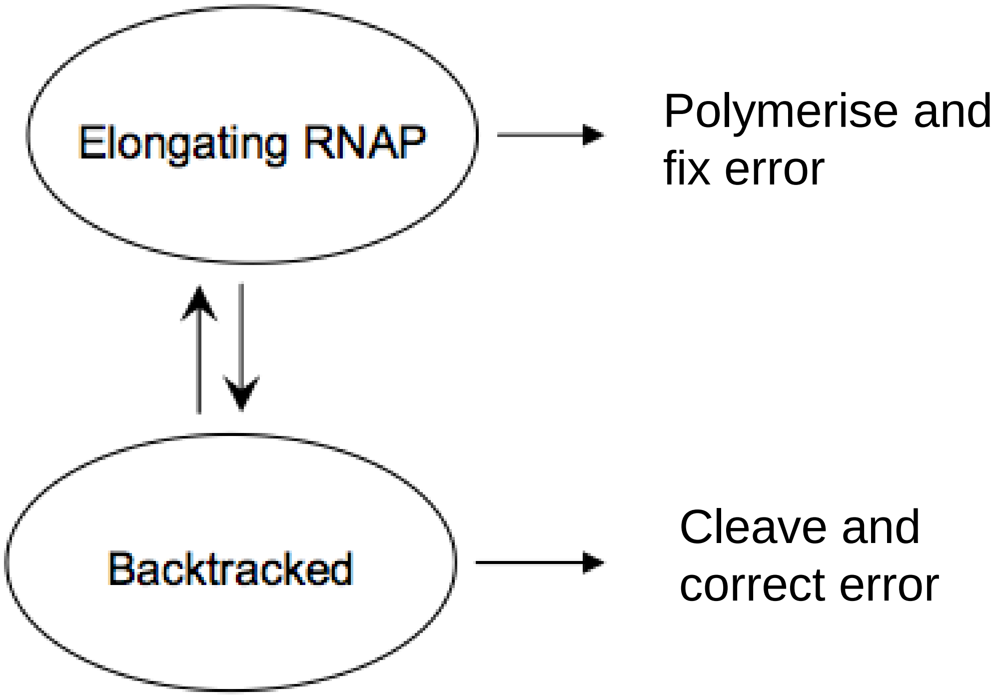
Requires a branching process

Possible mechanism - circumstantial evidence …

Backtracking
mRNA cleavage (Gre, TFIIS)

University of BRISTOL

Backtracking & error correction

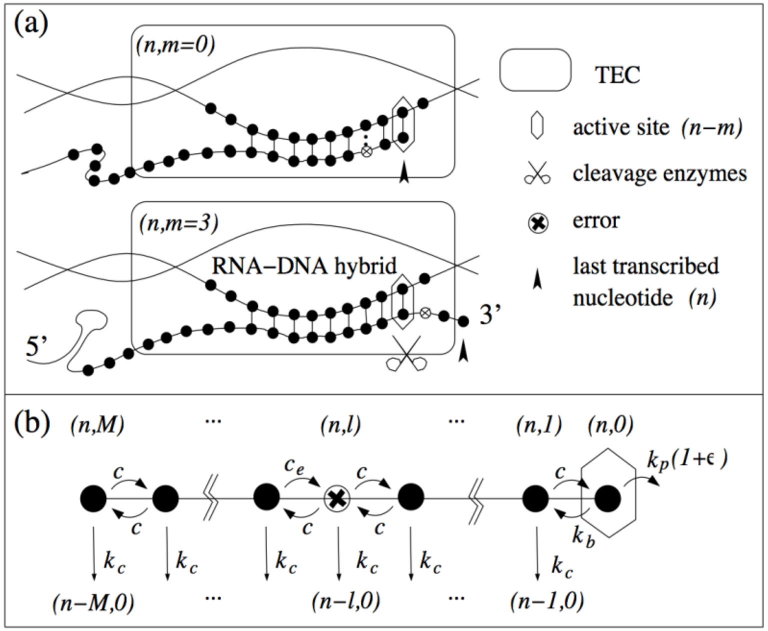


⇒ **Renormalized Error fraction :** ratio of wrong to correct nucleotides

Voliotis et al, PRL, **102**, 258101 (2009)

University of BRISTOL

A model of error correction



State of TEC given by (n,m)

$n \in [0, N]$ - last transcribed nucleotide

$m \in [0, M]$ - position of active site relative to n

At each nucleotide position $(n,0)$:

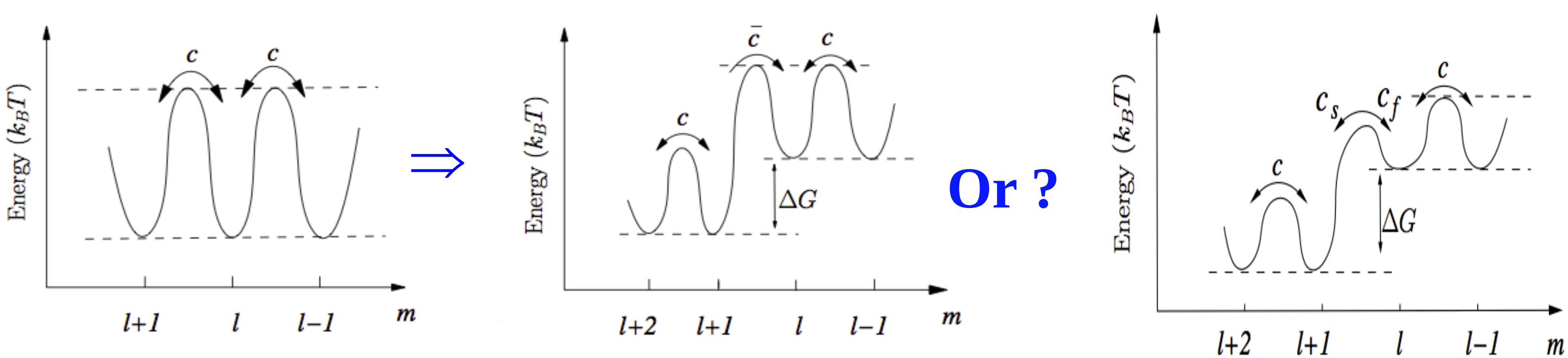
correct nucleotide k_p

wrong nucleotide $\bar{k}_p$

backtrack $k_b, \bar{k}_b$

Cleave at rate k_c

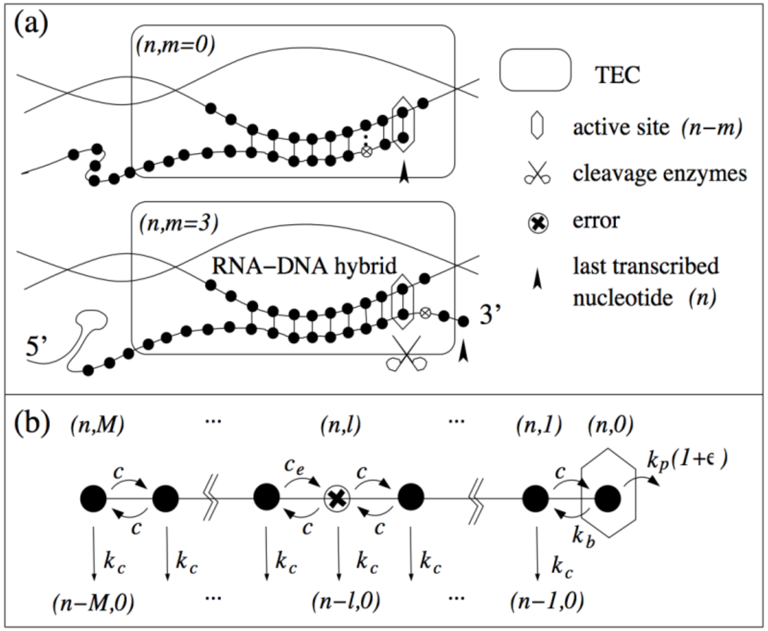
Error at position l



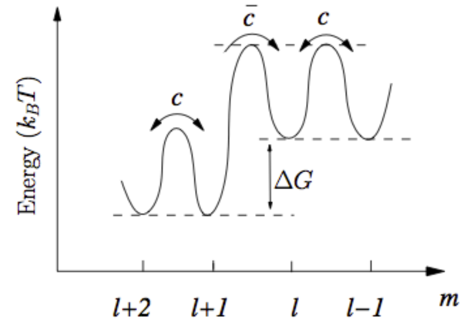
Voliotis et al, PRL, **102**, 258101 (2009)

University of BRISTOL

A model of error correction



Many possible energy landscapes, take for example …



$$\bar{c} = c_e$$

$$\bar{k}_b = k_b$$

$$c_e/c \simeq \epsilon \simeq e^{-\Delta G/k_B T}$$

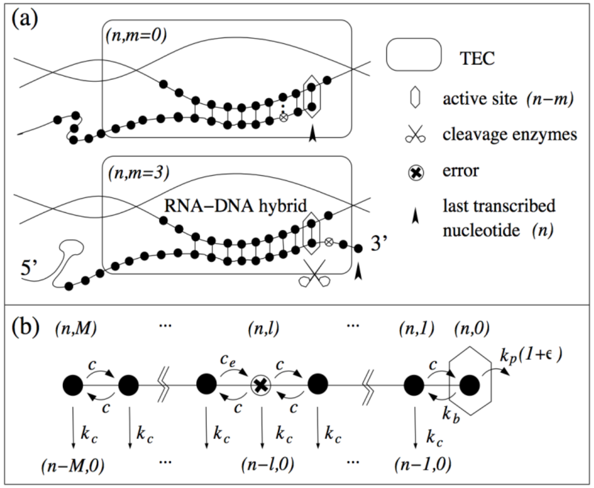
$$\frac{\bar{k}_p}{k_p} = e^{-\beta \Delta G} = \mathcal{E}_0$$

Quantities of interest that characterise the different competing processes

$$\begin{aligned} \alpha_1 &= k_c/c \\ \alpha_2 &= k_c/\bar{c} = \alpha_1/\epsilon \\ K &= k_p/k_b \end{aligned}$$

Voliotis et al, PRL, **102**, 258101 (2009)

A model of error correction



Dynamics described using Master eqn.

$$\frac{d\mathbf{P}}{dt} = \mathbf{W}^{(s)} \cdot \mathbf{P} \qquad \mathbf{P}(t) = [P(0,t), \ldots, P(M,t)]$$

Where *s* is a binary string that keeps track of correct and wrong nucleotides along the nascent mRNA,

- $s_i = 0$ - wrong nucleotide at position *i*
- $s_i = 1$ - correct nucleotide at position *i*

$\mathbf{W}^{(s)}$ - denotes the dependence of the rates on the sequence of correct and wrong nucleotides.

$$s = \underbrace{\{0, 1, \ldots, 0\}}_{n \text{ elements}}$$

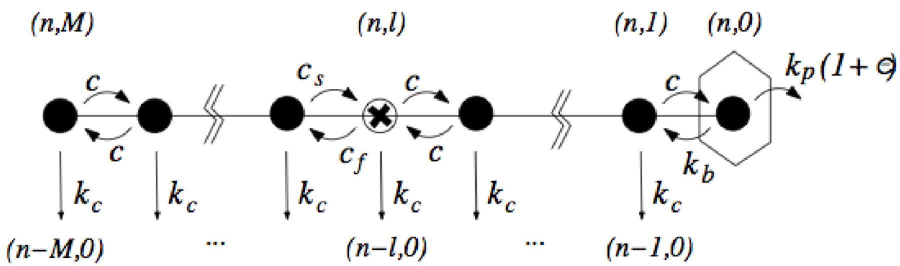
- $M+2$ possible outcomes
  - Polymerise correct nucleotide $(n,0) \rightarrow (n+1^c,0)$
  - Polymerise wrong nucleotide $(n,0) \rightarrow (n+1^w,0)$
  - Cleave from any backtracked state $(n,m=l>0) \rightarrow (n-l,0)$
- *Using Laplace transform techniques we can obtain the splitting probabilities* p_i *for each of the* $m+2$ *outcomes as well as their conditional mean exit time* τ_i

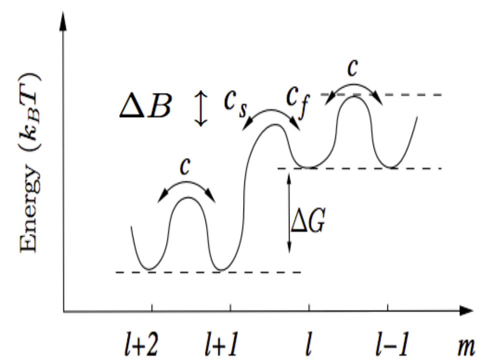
A model of error correction

For $K = k_b/k_p \gg 1$ **Error fraction** $\mathcal{E} \sim \dfrac{\epsilon^{M+1} M^M}{M!}$

boundary M $\Rightarrow M$ attempts to correct an error

Unrealistic to think that backtracking rate unchanged by error and $K \gg 1$





$$c_s/c_f = e^{-\beta\Delta G} = \epsilon$$

$$c_f/c = \bar{k}_b/k_b = e^{\beta\Delta B} \qquad \Delta B = f\Delta G$$

Presence of error leads to 'renormalization' of K $\Rightarrow$ $K^* \simeq K\,\bar{k}_b/k_b$

$\Rightarrow$ Can have effective $K^* \gg 1$ even if $K < 1$ when no error present

Numbers

Spontaneous error fraction : $10^{-2} - 10^{-3}$ ($\Delta G \approx 4 - 7 k_B T$) [1].

Blank et al, *Biochemistry* (1986)

The cleavage rate : $0.1 - 1 s^{-1}$ for bacterial RNAP in the presence of saturating concentrations of accessory cleavage factors

Sosunova et al, PNAS (2003)

Hopping rate : $1 - 10\ s^{-1}$ [3,4].
Relative backtracking rate, K ~ 0.1

Galburt et al, Nature (2007); Shaevitz et al, Nature (2003)

Conclusions

- single step birth/death models not sufficient to model fluctuations in gene transcription
- a model of back-tracking pauses show that they can play a significant role in determining fluctuations
- Single step models valid if initiation is rate-limiting step
- Inclusion of pausing dynamics with multiple RNAs on DNA template leads to bursting dynamics
- Backtracking can be used as model for proofreading in transcription

Perspectives

elongation under force …

transcript arrest in integrated model …

sequence dependence …

translation ..

bursting dynamics in gene expression …