Noise in gene expression: origins, experimental observations and consequences

Raúl Guantes

http://www.uam.es/raul.guantes
Outline:

- Cell-to-cell variability: Fundamental origins.

- Physicochemical principles underlying noise in gene expression.


- Experimental measurements of variability in gene expression: from single genes in individual cells to genome-wide single molecule quantification.

- Phenotypic consequences of noise in gene expression.
Cells continuously change their internal state to adapt to environmental changes, look for nutrients or coordinate with other cells.

All cellular processes must be tightly regulated and coordinated to ensure proper function and viability. But...do they always appear as continuous and ‘programmed’ events?

Chemotaxis

Competence/sporulation in *B. subtilis*
Do always cells in identical conditions produce the same response (output) after an stimulus (input)?

When the response is measured in single cells, we see that no two cells are identical, even in a clonal population under the same conditions (pH, T, nutrients etc.)
Gene expression in clonal cells is variable, heterogeneous, noisy or stochastic

Operational definition of biological noise: cell-to-cell variability (of a protein) in a homogenous population, quantified as

$$\text{Coefficient of variation (CV)} = \frac{\text{Width (standard deviation)}}{\text{Peak (mean value)}}$$

But we don’t need to do sophisticated single-cell assays to infer variability in gene expression!

A. Novick and M. Weiner, Enzyme induction as an all or none phenomenon *PNAS* 43, 553 (1957).
Benzer (1953): Induction kinetics at high [TMG]

Individual kinetics = population kinetics

Novick and Weiner: Induction kinetics at low [TMG]

Individual kinetics ≠ population kinetics

Instead of

“the transition from uninduced to induced is the consequence of a single random event”
Origins of noise in gene expression

Physicochemical principles underlying variability in gene expression

The cell as a (bio)chemical reactor

Consider the simplest gene expression system (birth/death protein dynamics)

\[ \phi \xrightarrow{k} P \quad k: \text{Translation efficiency/rate} \]

\[ P \xrightarrow{\delta} \phi \quad \delta: \text{Degradation rate} \]

How does the concentration of protein P changes with time?
**Mass action law:** total reaction rates are proportional to the products of the reactant concentrations

\[ \frac{d[P]}{dt} = k - \delta[P] \]

**Kinetic or reaction rate equation**

\[ [P](t) = \frac{k}{\delta} \left(1 - e^{-\delta t}\right) \]

**Equilibrium:** production = degradation

\[ [P]_{eq} = \frac{k}{\delta} \]

\[ k = 100 \quad ([\text{concent}] / \text{time}) \]

\[ \delta = 2 \quad (1/\text{time}) \]
But let’s take the numbers seriously

*Escherichia Coli:*

2 μm long, 1 μm diameter

\[ V \approx 10^{-15} \text{ l} \]

1nM ~ 1 molecule

\[ 10^{-9} \text{ mol/l} \times 10^{-15} \text{ l} \times 6 \times 10^{23} \text{ molecules/mol} \]

[RNA polymerase] ~ 100 nM ~ 100 molecules

*E. Coli* cell cycle period ~ 20 mins.

Degradation of stable proteins ~ ln 2/cell cycle period ~ 2 h\(^{-1}\)

\[ k = 100 \quad \text{nM/h} \]

\[ \delta = 2 \quad \text{h}^{-1} \]

in the previous example
Many molecules that take part in gene expression (including DNA, transcription factors and the enzyme polymerase) act at low intracellular concentrations.

What is the consequence of these low copy numbers?

Biochemical reactions can not be considered a continuous (deterministic) process, but a random (stochastic) process.

1.- In stochastic systems we can not predict the state of the system (number of molecules) at a given time given its state at the current time. We must deal with reaction probabilities and particle densities (probabilities for number of molecules).

2.- A reaction occurs when two chemical species approach and collide in the appropriate way (random microscopic events). Then molecules are created or destroyed in discrete steps with random reaction times.
Stochastic description of chemical reactions.

\[
\phi \xrightarrow{k} P \quad \text{Probability for protein ‘birth’ (production):} \quad \Pr(P_{t+dt} = n+1|P_t = n) \equiv a(n+1|n)dt \propto k \, dt
\]

\[
P \xrightarrow{\delta} \phi \quad \text{Probability for protein ‘death’ (degradation):} \quad \Pr(P_{t+dt} = n-1|P_t = n) \equiv a(n-1|n)dt \propto n \, \delta \, dt
\]

Reaction probabilities instead of reaction rates, but note that the microscopic reaction probabilities are proportional to the macroscopic reaction rates (mass action law).

How does the probability of having \( n \) molecules of protein, \( p(n) \), changes with time (dynamic evolution)?
The appropriate description is not a kinetic or reaction rate equation as previously but a **Master equation**.

This is a gain/loss equation (**total gain probability** – **total loss probability**) as in the deterministic case (production – degradation).

Consider all possible steps involving production/degradation of one molecule

![Diagram](https://via.placeholder.com/150)

Recall:

\[
\begin{align*}
    a(n-1 \mid n) dt &= n\delta dt \\
    a(n \mid n-1) dt &= k dt \\
    a(n \mid n+1) dt &= (n+1) \delta dt \\
    a(n+1 \mid n) dt &= k dt
\end{align*}
\]

We have 1 species (protein P), 2 reactions (translation/degradation) and 4 propensities \( a \) for gain (to \( p(n) \)) or loss (from \( p(n) \)) of a protein molecule.
Taking the limit $dt \to 0$:

$$\frac{dp(n)}{dt} = a(n \mid n-1)p(n-1) + a(n \mid n+1)p(n+1) - a(n-1 \mid n)p(n) - a(n+1 \mid n)p(n)$$

\[\text{gain for } p(n)\qquad\text{loss for } p(n)\]

A summary for stochastic description of chemical reactions:

1.- We must know the number of reactions for each molecular species.

2.- We have to know the reaction probabilities (propensities, related to
macroscopic reaction rates).

3.- We add all reactions contributing to a gain of $n$ molecules of each species, and subtract all reactions producing loss of $n$ molecules.
To solve the Master Equation given some initial values for each molecular species, all we need is to answer these two questions:

1) When will the next reaction occur?
2) What kind of reaction will it be?

This can be numerically obtained using Gillespie’s algorithm.


[DIZZY](http://magnet.systemsbiology.net/software/Dizzy/)

There are also approximate methods for stochastic chemical reactions:

**Langevin equations:**

\[
\frac{d}{dt} P = k - \delta P + \xi(t)
\]

where \(\xi(t)\) is a random or noise term with specified statistics.
Now let us have a look at some ‘real’ protein dynamics.

\[ k = 100 \text{ molecules/h} \]

The mean of the probability \( p(n) \) is the deterministic value

Reaction rate equations give the evolution of the mean of molecule numbers
Let’s decrease the number of molecules by decreasing the translation rate

\[ k = 10 \text{ molecules/ h} \]

The lower the number of molecules involved, the larger the variability (noise)
Now let’s increase 10 times the translation rate with respect to first example

\[ k = 1000 \text{ molecules/h} \]

The deterministic limit (blue line) is recovered for very large number of molecules

Protein noise scales with protein abundance as \( 1/\sqrt{n} \)
Summary of stochastic gene expression. Part I.

1.- There is inherent **variability** in gene expression under the same conditions (can be quantified with the **coefficient of variation**).

2.- A fundamental source of variability is the **random** nature of biochemical reactions involving **low** copy numbers.

3.- The proper description is the **Master equation** for the evolution of probabilities of having $n$ molecules of a given species at a given time. The stochastic evolution for the number of molecules can be numerically obtained using Gillespie’s algorithm.

4.- The reaction rate or kinetic equations for the concentrations give the evolution of **mean** number of molecules. They are only a proper description at very **large** copy numbers.
Stochastic gene expression. Part II: different sources of noise in prokaryotic and eukaryotic gene expression

We have seen that low copy numbers are a source of variability in gene expression (finite size noise). But, what are other fundamental sources of variability?

Consider the extended (but still simplified) gene expression system:

\[ k_{on} \text{ = transition rate from repressed to active promoter} \quad s_p \text{ = translation rate} \]
\[ k_{off} \text{ = transition rate from active to repressed promoter} \quad \delta_M \text{ = mRNA degradation rate} \]
\[ s_A \text{ = transcription rate from active promoter} \quad \delta_p \text{ = protein degradation rate} \]
\[ s_R \text{ = transcription rate from repressed promoter} \]

Now we have 4 molecular species (R,A,M,P) and 7 reactions. We can write deterministic equations applying mass action law:

\[
\frac{dA}{dt} = k_{on} R - k_{off} A \quad \text{Active promoter dynamics}
\]

\[
\frac{dR}{dt} = k_{off} A - k_{on} R \quad \text{Repressed promoter dynamics}
\]

\[
\frac{dM}{dt} = s_A A + s_R R - \delta_R M \quad \text{mRNA dynamics}
\]

\[
\frac{dP}{dt} = s_p M - \delta_P P \quad \text{protein dynamics}
\]
The problem can be further simplified if we consider the time scales for each reaction

transitions between repressed and active promoter are fast (usually in the scale of seconds) while the rest of reactions are slower (in the scale of minutes)

Then we can consider that A and R promoter states are in equilibrium with respect to mRNA and protein changes

\[
\frac{dR}{dt} = \frac{dA}{dt} \approx 0 \quad \Rightarrow \quad \frac{k_{on}}{k_{off}} = \frac{A}{R}
\]

Also

\[
A + R = 1
\]

(coppy number of the gene \( p \) is constant and equal to 1)

Then we can eliminate the variables A and R from the equations
\[
\frac{dM}{dt} = \frac{k_{on}}{k_{on} + k_{off}} s_A + \frac{k_{off}}{k_{on} + k_{off}} s_R - \delta_R M \quad \text{mRNA dynamics}
\]

Fraction of time the promoter is \textit{active} \quad \text{Fraction of time the promoter is \textit{repressed}}

\[
\frac{dP}{dt} = s_p M - \delta_p P \quad \text{protein dynamics}
\]

The equation for protein dynamics is the same as in the birth/death protein system, except that \textit{‘total’ translation rate } s_p M \textit{ is not constant}, but depends on mRNA dynamics

We can change both the number of mRNA molecules and protein molecules \textit{(finite size effect of mRNA?)}
\[ s_A = 50, \quad s_R = 5 \]
\[ s_P = 0.2 \]
\[ k_{on} = k_{off} = 10 \]
units per min

**Transcription rate**

**Translation rate**

**Activation/deact rate**

protein + mRNA noise

\[ s_A = 0.5, \quad s_R = 0.05 \]

**Transcription rate decreased**
only mRNA noise

\[
\begin{align*}
& s_A = 0.5, \quad s_R = 0.05 \quad \text{Transcription rate decreased} \\
& s_p = 20 \quad \text{Translation rate increased}
\end{align*}
\]

Average number of proteins is high, but fluctuations are very large!!!

Noise is propagated (transmitted) from mRNA to protein

‘Translational bursting’ mechanism: the larger the number of proteins produced per transcript (burst parameter or translational efficiency, \( b = s_p/\delta_M \)), the larger the fluctuations.

Variability is more strongly associated to changes in the number of mRNAs
All these mechanisms assume that promoter kinetics (activation/deactivation rates) is fast. What happens for slow promoter kinetics?

\[ k_{on} = k_{off} = 0.7 \]

\[ k_{on} = k_{off} = 0.01 \]

**Promoter state half-life ~ 1 h**

**Bimodality**: Protein alternates between high (active promoter) and low (repressed promoter) expression states

‘Transcriptional bursting’ mechanism: the longer the time in the active state, the more mRNAs are synthesized in rapid succession (noise increases with transcriptional efficiency, \( s_A/k_{off} \)).
In summary, for a simple activation/transcription/translation gene expression model we have different noise sources:

1. **Finite size noise**: low copy numbers of protein and mRNA (prevalence of the last one).

2. **Translational bursting**: proteins are produced in ‘bursts’ (few mRNAs with high translational efficiency).

3. **Transcriptional bursting**: mRNAs are produced in ‘bursts’ (episodes of long promoter activation with high transcriptional efficiency).

Fine, but: what’s going on in real cells?.

What kind of experiments can we design to determine where is gene expression noise coming from in cells?
Experiments or **real life**: Single cell, one gene.

Noise in prokaryotic gene expression


We want to ‘see’ the the proteins (use **GFP**)

And we want to ‘count’ them (to plot the probability distributions) → **Flow cytometer**

How do we vary the gene ‘parameters’

(translation, **transcription**)?
Transcription can be varied using an inducible promoter (for instance, using IPTG) upstream of the \textit{gfp} gene (reporter).

Translation can be varied by point mutations in ribosome binding sites (RBS) and initiation codon of \textit{gfp}.

Control: Transcription is also varied with point mutations in the promoter.
A useful trick to experimentally distinguish between two different noise sources

- Translational efficiency varied
- Transcription rate varied

Quantify noise by **noise strength** or **Fano factor**=
\[
\text{Fano factor} = \frac{\text{Std}^2}{\text{Mean}}
\]

It distinguishes between both mechanisms when only protein number is changing, which is what we usually measure in experiments.
This can be shown from an approximate theory for single gene expression:

\[ \frac{\sigma_p^2}{\langle p \rangle} \simeq 1 + b \]

\[ b = \text{Translation rate/mRNA degradation rate} \]
Then count the number of GFPs in a bacterial population, plot the histograms and calculate the noise strength under different gene ‘parameters’

Recall!
Translational bursting is a source of noise in \textit{B. subtilis} gene expression.

Varying translational efficiency in all strains and mutants has a larger effect on noise than varying transcriptional efficiency.
Is it the same in all cell types?: noise in eukaryotic gene expression


Same experimental strategy with the yeast *GAL1* promoter
Translational efficiency is varied using GFP variants with different translation rates, owing to differences in synonymous codon usage (codon-specific tRNAs have different abundances).

One can also look at the evolution of protein number probabilities (remember the Master equation).

Different modes of transcriptional control give very different transient responses.
Let us analyze with the simple mathematical model of gene expression what is the influence of different modes of transcription induction on noise strength.

- Transcription rate $s_A$ varied, with fast active/repressed promoter transitions (prokaryotic case)
- Transcription rate $s_A$ varied, with slower promoter activation/repression rates
- Promoter activation rate $k_{on}$ varied
- Promoter deactivation rate $k_{off}$ varied

Experiments:

Remember E. coli
What about varying translation?:

Eukaryotic gene expression:

Although translational bursting may be an important source of noise in yeast, transcriptional bursting due to pulsatile mRNA production (resulting from transcriptional reinitiation) seems to be the most important source of noise, in contrast to prokaryotic gene expression.
In summary:

1.- When one takes into account promoter and mRNA dynamics (activation/inactivation rates, transcription) there are different sources of noise in gene expression (translational bursting, transcriptional bursting).

2.- We can have a large number of protein molecules, but a high variability, due to propagated noise from other species (mRNA).

3.- Clever single cell experiments help to understand the source of variability.

4.- In E.coli, noise in gene expression seems to be due to a translational bursting mechanisms (few mRNA synthesize many proteins).

5.- In yeast there can be a substantial contribution from slow promoter dynamics (many mRNA are synthesized when promoter is ON)

6.- Mathematical models, although simplified, may help to guide and understand experimental results.
Single cell, two genes: different noise sources.

**Intrinsic versus extrinsic noise**

Up to now we have only analyzed possible origins of noise based on the discrete nature of biochemical reactions involved in gene expression (promoter activation, transcription, translation). This is called *intrinsic* noise.

Imagine other possible sources of noise:

1. Differences in the internal states of population of cells (cell cycle..).
2. Subtle environmental differences.
3. Changes in local concentrations and states of other regulatory proteins, polymerases, etc.

All these are *extrinsic* sources of noise.
How can we distinguish experimentally between intrinsic and extrinsic noise?


**Intrinsic noise:** Difference in gene expression between two identical copies of a gene expressed under precisely the same conditions.

**Two colour experiments**

Two equivalent Lac-repressible fluorescent reporter genes inserted in the *E. coli* chromosome controlled by the same promoter, and on opposite sites and equidistant from the replication origin.
Then we must follow the activity of the two genes inside a given cell along time (stochastic trajectories) to see their differences in gene expression: time lapse fluorescence microscopy (Elowitz lab)

What would one expect if only extrinsic noise is acting?
What would one expect if only intrinsic noise is acting?

Promoters repressed by wild type repressor (lacI) gene

- low transcription (low number of mRNA molecules)
- high intrinsic noise
Promoters unrepressed due to addition of IPTG

high transcription (high number of mRNA molecules)

low intrinsic noise
Single cell, many genes: genomewide properties of noise


**Large scale studies:** They determined mean protein abundance and noise for 43 different GFP fusion proteins and 11 different experimental conditions in yeast.

\[ CV^2 = \frac{\sigma_p^2}{\langle p \rangle^2} \propto \frac{1200}{\langle p \rangle} \]

Cells expressing GFP-fused version of each protein

Green points: before stimulus is applied
This dependence of protein noise with the inverse of the abundance (remember simple birth/death protein model) is general irrespective of the protein fusion, promoter and experimental conditions.
High throughput flow cytometry: Protein abundance measurements of >2,500 proteins in rich and minimal media.

They use the two-colour strategy (double fluorescent protein fusions) to separate intrinsic from extrinsic noise.

Again, for low/medium protein abundance intrinsic noise dominates as seen from the inverse linear scaling.
Theory predicts that mRNA noise is a possible global noise source. But both translational bursting and transcriptional bursting give the same scaling with $1/\langle p \rangle$

Transl. burst.

$$\frac{\sigma_m^2}{\langle m \rangle} \propto \frac{1}{\langle m \rangle}$$

Transcrip. burst.

$$\frac{\sigma_m^2}{\langle m \rangle} \propto \frac{1}{\langle m \rangle} \frac{s_d}{(\delta_u + b)}$$

The scaling is the same, only the proportionality constants change

So we should try to ‘see’ how single proteins and mRNAs are produced in cells (single-molecule techniques)
Single molecule, one gene: probing mRNA and protein dynamics in real time.

Single molecules of GFP are generally undetectable in vivo due to background fluorescence: need for different experimental strategies:


I Golding et al., *Real time kinetics of gene activity in individual bacteria*, *Cell* 123, 1025 (2005)

Count the number of mRNA transcripts coating mRNA with 50-100 binding sites for GFP → they directly measure transcriptional bursts in *E. coli*
Single molecule mRNA experiments show that transcription bursts are a general mechanism of noise in *E. coli*

Novick and Weiner revisited: following stochastic events in phenotype switching.


**Detection by localization**: reducing diffusion of GFP increases autofluorescence over the background.

A **stochastic dissociation** of the Lac repressor from its binding sites is the *single random event* triggering the switch from uninduced to induced state in the cell.
Single molecule, many genes: proteome and transcriptome abundance at single molecule resolution.


Combined highly sensitive fluorescence microscopy with microfluidics
Inverse linear scaling (intrinsic noise) is the main contribution for proteins with low copies (< 10).

Individual time traces and two-colour measurements indicate that extrinsic noise contribution is dominant at high expression levels.
A brief remark on the consequences of noisy gene expression: is noise always a nuisance?

Noise produces phenotypic variation, which can be advantageous!
Examples in single cells: Latent to infectious switch in HIV virus


Noise in key regulators of the latency/active state is sufficient to drive transitions from active to latent.
Adaptation to changing environments: bacterial persistence

**Stochastic cell fate commitment:** Cells stochastically choose among two different fates with fixed proportions

Entry into competence by *B. Subtilis* (20/80 %)

*Drosophila* eye photoreceptor development (30/70 %)

The Waddington’s landscape picture of cellular differentiation: Cells are driven to different stable expression states (valleys), and the driving force is many times stochastic


**In summary:**

Noise and variability in gene expression is not always a fundamental limitation that a cell has to control.

Noise can be beneficial in some circumstances, and plays diverse functional roles (fast adaptation to fluctuating environments, helping development by favouring decisions between different cell types, etc.)…