Inference, Control and Driving in Natural Systems: Cellular Chemosensing

Thomas Ouldridge

Imperial College

t.ouldridge@imperial.ac.uk

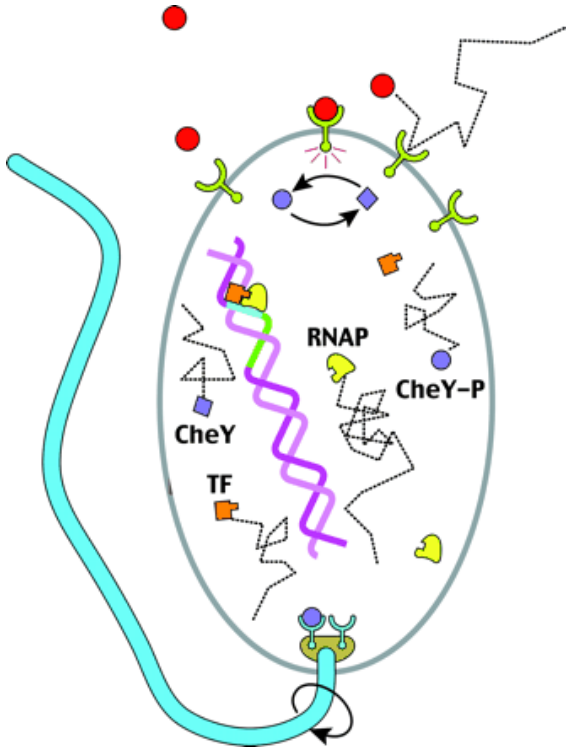
January 13, 2016

Overview

References

[1] K. Kaizu et al., Biophys. J. 106:976-985 (2014).
[2] H. C. Berg and E. M. Purcell, Biophys. J. 20:193-219 (1977).
[3] R. G. Endres and N. S. Wingreen, Phys. Rev. Lett. 103:158101 (2009).
[4] A. H. Lang et al., Phys. Rev. Lett. 113:148103 (2014).
[5] P. Nelson, Biological Physics: Energy, Information, Life.
[6] http://www.info612.ece.mcgill.ca/lecture_02.pdf
[7] C. C. Govern and P. R. ten Wolde. Phys. Rev. Lett. 113:258102 (2014).
[8] C. C. Govern and P. R. ten Wolde. Proc. Nat. Acad Sci USA 111:17486-91 (2014).

Biological Sensing of Chemicals



- We are interested in the problem of sensing the value of a constant external concentration of chemicals.
- Our canonical motif will be a receptor that binds to (and unbinds from) a ligand.
- Other situations and motifs also occur.

Taken from Bialek and Setayeshgar, Proc. Nat. Acad. Sci. USA 102:10040-10045 (2005).

A Single Receptor

Assume the cell attempts to estimate the concentration of an external ligand through the instantaneous occupancy of a single receptor.

- Ligands bind to the receptor at a rate $k_+ C_L$.
- Ligands unbind at a rate k_- (see Ref. [1] for a discussion of subtleties).
- Average receptor occupancy $\bar{r} = \frac{k_+ C_L}{k_+ C_L + k_-}$.
- At any instant, receptor state is $r = 0$ with $p = 1 - \bar{r}$ and $r = 1$ with $p = \bar{r}$.

We have a problem: clearly the instantaneous value of r tells us very little about $\bar{r}$ and hence C_L; an instantaneous guess $c_L = k_- r/(k_+(1-r))$ of C_L has an infinite variance.

Many Receptors

This problem can be alleviated if we have many receptors. With many receptors, the fraction with ligands bound (R) will typically show small deviations about $\bar{R} = p$.

- With N_R independent receptors, $\sigma_R^2 = \frac{p(1-p)}{N_R}$.
- Standard error propagation (see eg. Wikipedia) gives

$$\left(\frac{\sigma_{c_L}}{C_L}\right)^2 = \left(\frac{1}{C_L}\right)^2 \left(\frac{\mathrm{d}c_L}{\mathrm{d}R}\right)^2_{R=\bar{R}} \sigma_R^2.$$

$$= \frac{1}{p^2(1-p)^2}\sigma_R^2 = \frac{1}{N_R p(1-p)}. \quad (1)$$

- Note that this result is inconsistent with infinite variance for $N_R = 1$. *Identify the cause of the inconsistency, and verify that the variance for $N_R = 1$ is indeed infinite.*
- *Are receptors on a cell independent? Is it worth covering 100% of a cell surface with receptors? See Ref. [2].*

Berg-Purcell Limit I

A good Bayesian would consider the history of the receptor. In a classic paper (Ref. [2]) Berg and Purcell suggested that a cell might average over the state of a receptor over some period T. We will initially ignore the question of how this might be done, and simply estimate the variance of such an estimate. For a single receptor,

$$r_T = \frac{1}{T}\int_{t_1}^{t_1+T} \mathrm{d}t\, r(t). \tag{2}$$

The variance in concentration estimate is given by

$$\left(\frac{\sigma_{c_L}}{C_L}\right)^2 = \left(\frac{1}{C_L}\right)^2 \left(\frac{\mathrm{d}c_L}{\mathrm{d}r_T}\right)^2_{r_T=\bar{r}} \sigma^2_{r_T}. \tag{3}$$

Berg-Purcell Limit II

Using $c_L = k_- r_T/(k_+(1-r_T))$,

$$\left(\frac{\sigma_{c_L}}{C_L}\right)^2 = \frac{1}{p^2(1-p)^2}\sigma^2_{r_T}, \tag{4}$$

Note that this is the same as of the error from the instantaneous value of R (Eq. 1), but with $\sigma^2_{r_T}$ instead of σ^2_R. This is because $\bar{r}_T = \bar{R} = p$, and so c_L is the same function of r_T and R.

Our task is to evaluate $\sigma^2_{r_T}$.

Berg-Purcell Limit III

$$\sigma^2_{r_T} = \frac{1}{T^2}\int_{t_1}^{t_1+T}\int_{t_1}^{t_1+T} \mathrm{d}s\mathrm{d}t\langle r(t)r(s)\rangle - p^2 \qquad (5)$$

r can take two values, 0 or 1. $P_1(t) = P(r=1,t)$ is governed by the master equation

$$\frac{\mathrm{d}P_1(t)}{\mathrm{d}t} = k_+ C_L(1-P_1(t)) - k_- P_1(t). \qquad (6)$$

The solution is trivial, and allows us to extract $P(r=1,t+\tau|r=1,t)$. This is all you need to show

$$\langle r(t)r(t+\tau)\rangle = p(1-p)\exp(-(k_+C_L+k_-)|\tau|) + p^2. \qquad (7)$$

Derive this relationship. Don't forget that p, the average occupancy of a ligand, is given by $p = k_+C_L/(k_+C_L+k_-)$.

Berg-Purcell Limit IV

Thus

$$\sigma_{r_T}^2 = \frac{p(1-p)}{T^2} \int_{t_1}^{t_1+T} \int_{t_1-t}^{t_1+T-t} \mathrm{d}\tau \mathrm{d}t \exp(-|\tau|/\tau_r), \qquad (8)$$

with $\tau_r = 1/(k_+ C_L + k_-)$. The integral is straight-forward, provided care is taken to split the integral over τ into two parts ($t_1 - t < \tau < 0$ and $0 < \tau < t_1 + T - t$). The result is

$$\sigma_{r_T}^2 = \frac{2p(1-p)\tau_r}{T}\left(1 - \frac{\tau_r}{T}(1 - \exp(-T/\tau_r))\right). \qquad (9)$$

Verify this result. Eqs. 4 and Eq. 9 imply (for $T \gg \tau_r$)

$$\left(\frac{\sigma_{c_L}}{C_L}\right)^2 = \frac{2\tau_r}{Tp(1-p)} = \frac{2}{Tk_+ C_L(1-p)} = \frac{2}{\bar{n}}, \qquad (10)$$

in which $\bar{n}$ is the average number of binding events in time T.

Berg-Purcell Limit VI

The Berg-Purcell result can be extended to N_R independent receptors; in this case

$$\left(\frac{\sigma_{c_L}}{C_L}\right)^2 = \frac{2}{\bar{n} N_R}. \tag{11}$$

I always have the following nagging doubts with the Berg-Purcell result:

- Although we're doing better than before and synthesizing data from multiple time points, is taking the average R_T the best a cell can do?
- How would a cell take an average over some time period anyway?

Maximum Likelihood Estimation I

We (and presumably the cell) have a model of the physics with a single unknown parameter, the ligand concentration. We can therefore calculate the likelihood of a series of binding and unbinding events as a function of c_L [3].

Assume we start with a single receptor in an unbound state. The likelihood of surviving for a time t and then binding to a ligand is exponentially distributed with an average of $(k_+ c_L)^{-1}$:

$$P_{\mathrm{u}}(t) = k_+ c_L \exp(-k_+ c_l t). \tag{12}$$

Similarly, the probability of remaining ligand-bound for some time s is

$$P_{\mathrm{b}}(s) = k_- \exp(-k_- s). \tag{13}$$

Maximum Likelihood Estimation II

The likelihood of observing a series of n binding and unbinding events, with waiting times $t_i^+, t_i^-, i = 1..n$, given a ligand concentration c_L is then

$$P(\{t_i^+, t_i^-\}|c_L) = (k_- k_+ c_L)^n \exp\left(-\sum_{i=1}^{n} k_- t_i^- + k_+ c_L t_i^+\right)$$

$$= (k_- k_+ c_L)^n \exp(-k_- t_\mathrm{b} - k_+ c_L t_\mathrm{u}). \quad (14)$$

Here, $t_\mathrm{b} = \sum_{i=1}^{n} t_i^-$ and $t_\mathrm{u} = \sum_{i=1}^{n} t_i^+$ are the total time that the receptor spends in bound and unbound states respectively.

Maximum Likelihood Estimation III

$$P(\{t_i, s_i\}|c_L) = (k_- k_+ c_L)^n \exp(-k_- t_\mathrm{b} - k_+ c_L t_\mathrm{u}). \quad (15)$$

In principle, we (or the cell) could combine this likelihood with prior beliefs about c_L to respond optimally to a series of binding and unbinding events.

Main point: t_b only appears in an overall normalisation factor – it does not make certain values of c_L more likely than others. Any optimal statement about the external ligand concentration should not include t_b. For example, with a flat prior, the c_L that maximises the a posteriori likelihood is

$$c_L = \frac{n}{t_\mathrm{u} k_+}. \quad (16)$$

Maximum Likelihood Estimation IV

How does this estimate compare to that obtained by Berg and Purcell? The variance in c_L estimated from n binding cycles is

$$\left(\frac{\sigma_{c_L}}{C_L}\right)^2 = \left(\frac{1}{C_L}\right)^2 \left(\frac{\mathrm{d}c_L}{\mathrm{d}t_\mathrm{u}}\right)^2_{t_\mathrm{u}=\bar{t_\mathrm{u}}} \sigma^2_{t_\mathrm{u}} = \frac{\sigma^2_{t_\mathrm{u}}}{\bar{t_\mathrm{u}}^2}. \tag{17}$$

Further,

$$\sigma_{t_\mathrm{u}}{}^2 = n\sigma^2_{t^+} = n(\bar{t}^+)^2 = \bar{t_\mathrm{u}}^2/n \tag{18}$$

So the variance in our estimate obtained from n cycles is $1/n$, compared to a variance of $2/\bar{n}$ obtained by Burg-Purcell in a time T. In the limit of long times, therefore, the MLE is twice as precise.

Maximum Likelihood Estimation V

Clearly we can also combine estimates of the average waiting time for binding from N_R independent receptors, reducing the variance of our estimate. But what is the real physical content of the MLE?

In the Berg-Purcell case, the time intervals in which the ligand is bound contribute to our estimate. Thus the variance associated with these time intervals contributes to our uncertainty. But, as highlighted by the MLE, these time intervals actually tell us nothing about C_L: ligand unbinding is characterised by a C_L-independent rate k_-.

Instead, we should just measure the average waiting time for binding.

MLE in cells I

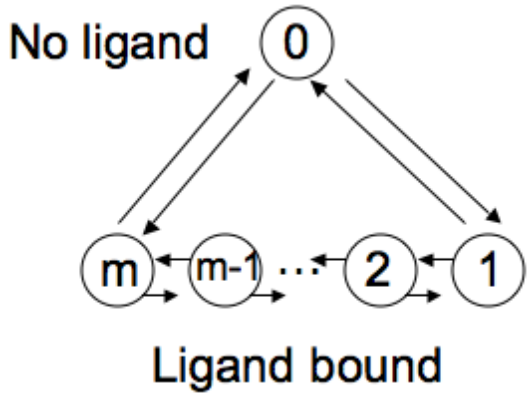
It is easy for us to look at a time trace and implement the MLE. It is much harder to see how a cell would do it.

It seems like the downstream circuitry needed to calculate an average unbinding time would be even more complicated than that needed to integrate R for the Berg-Purcell approach.

- Is it possible to redesign the receptor so that we reduce the variance associated with dissociation times?
- Does this help?

MLE in cells II

So far we have modelled detachment as a Poisson process. Such processes have a a variance equal to the square of their mean; the are broad. In a markov process in which the ligand-bound configuration is a single, discrete state, this is unavoidable.



A process that requires several steps can have a much lower variance in completion time compared to its mean [4]. If a process involves $m > 1$ irreversible exponentially-distributed steps of mean τ, the overall mean is $m\tau$ and variance $m\tau^2 < m^2\tau^2$.

MLE in cells III

To reduce variance, a systematic tendency to step in one direction is vital (*see practical*). In terms of our receptors, we need ligands to systematically bind in state 1 and unbind from state m. This requires *driving*: an input of fuel (eg. ATP) [4,5].

Detailed balance [5]: Consider two states i and f. The rate constants for transitions, $q_{i\to f}$ and $q_{f\to i}$, are related by

$$\frac{q_{i\to f}}{q_{f\to i}} = \frac{\pi_f}{\pi_i} = \exp((G_i - G_f)/kT), \tag{19}$$

in which π_i and π_f are the probabilities of being in states i and f in thermodynamic equilibrium. G_i and G_f are free energies, defined by $\pi_x = \exp(-G_x/kT)$.

MLE in cells IV

Given

$$\frac{q_{i\to f}}{q_{f\to i}} = exp((G_i - G_f)/kT), \tag{20}$$

then in a loop $0 \to 1 \to ... \to m \to 0$, the total free energy change is

$$\Delta G_{\rm loop} = -kT \ln \prod_{i=0}^{m} \frac{q_{i\to i+1}}{q_{i+1\to i}} < 0, \tag{21}$$

where I have added '$<$' because we are considering a system that tends to go in one direction around the loop. But G_0 is the probability of being in state 0 in equilibrium. It should therefore be single-valued - so how can $\Delta G_{\rm loop}$ be non-zero?

This can only be achieved if something we have coarse-grained away has *actually changed* during the cycle; i.e., the state of a fuel molecule. *We will explore these questions further in the practical.*

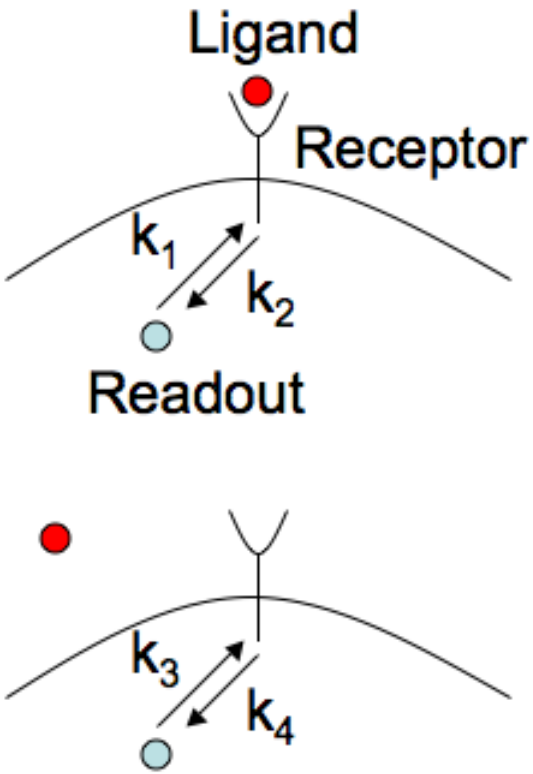
Cells must perform readouts

Cells must have a molecular mechanism for reading out from receptors.

- The signal must be conveyed to the interior of the cell.
- We would like to do some form of integration over the receptor history.

A passive readout mechanism

Can we achieve these two goals with a readout that binds and unbinds from the receptor in a way that is influenced by the binding of a ligand? This is a passive readout (it does not consume fuel).



Such a passive readout mechanism is shown in the Figure. The fraction of free readout molecules, X, will depend on the fraction of ligand-bound receptors R (provided $k_3/k_4 \neq k_1/k_2$). Clearly the readout can convey the signal to the cell's interior, but we will show that it can't integrate receptor history.

Mutual information I

Our discussion will make use of mutual information. We could consider the external ligand concentration C_L, the fraction of receptors ligand bound R and the fraction of readouts bound to receptors X as related random variables. The uncertainty of a variable Z is quantified by its entropy $H(Z)$

$$H(Z) = -\sum_z p(z) \log p(z). \tag{22}$$

We can also quantify the uncertainty in Z given knowledge of a second variable Y through the conditional entropy

$$H(Z|Y) = -\sum_{z,y} p(y)p(z|y) \log p(z|y). \tag{23}$$

Mutual information II

The mutual information is the difference between the two

$$I(Z,Y) = H(Z) - H(Z|Y) = \sum_{z,y} p(y,z) \log \left(\frac{p(y,z)}{p(y)p(z)} \right). \tag{24}$$

$I(Z,Y)$ quantifies the reduction in uncertainty about Z provided by knowing about Y. An effective cellular readout network X will have a high $I(X, C_L)$.

- The mutual information is symmetric.
- $I(Z,Y) \geq 0$ (*prove this – see [6] for guidance*).

Data Processing Inequality I

Let's say we have three variables V, Y, Z. Let us assume that V depends on Z only via Y;

$$P(V|Y,Z) = P(V|Y). \tag{25}$$

Then we can show

$$I(Z,V) \leq I(Z,Y). \tag{26}$$

This is the data processing inequality.

Data Processing Inequality II

$$I(Z,V) \leq I(Z,Y). \tag{27}$$

Prove this. Hint:

- First prove that the mutual information between Z and the combined variable (V,Y) can be expanded as

$$I(Z,(V,Y)) = I(Z,V) + I(Z,Y|V) = I(Z,Y) + I(Z,V|Y), \tag{28}$$

where

$$I(Z,V|Y) = \sum_{v,y,z} p(y)p(v,z|y) \log \frac{p(v,z|y)}{p(v|y)p(z|y)}. \tag{29}$$

- Show that $I(A,B|C) \geq 0$ in general.
- Show that $I(Z,V|Y) = 0$ in our special case.

Passive Readouts Cannot Integrate [7]

Why is all this relevant? In a steady state, our passive sensing system must be in thermodynamic equilibrium.

Thermodynamic equilibrium implies a particular distribution of free readout fraction X for a given fraction of receptors in the ligand-bound state R. Therefore the probability distribution of X at any given time is fully determined by R regardless of the actual value of C_L.

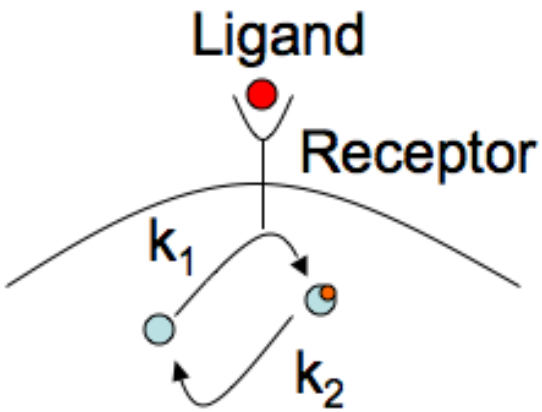
$$P(X|R, C_L) = P(X|R). \tag{30}$$

Thus $I(C_L, X) \leq I(C_L, R)$ and we can't do any better than the instantaneous state of the receptors by looking at passive readouts.

Active Readouts Can Integrate

To break this restriction, we need to push the system out of equilibrium. One way to achieve this is to use the receptors to catalyse a change of state of the readouts [8].

Readouts can be phosphorylated/methylated by receptors, and then decay by a separate pathway. This implies a net cycle in the system, which must therefore be out of equilibrium.



Such a cycle can only be maintained by using up chemical fuel. To read more about how such a process can effectively integrate the receptor signal, and to understand the trade-offs involved, see Ref. [8].

Summary

- Cell surface receptors are a key ingredient of chemosensing.
- The instantaneous signal from a single receptor is of limited use.
- To improve, we need multiple receptors or to integrate a single receptor's state over time.
- Simply averaging a receptor's occupancy over time is not the optimal way to extract information on ligand concentration from a trajectory - unbinding times offer no information, but introduce variance into the signal.
- Variance in ligand unbinding times can be reduced if the cell is willing to burn fuel.
- Internal readout molecules can be used to integrate the receptor's signal over time. To do so, however, the readout process must itself use chemical fuel (it cannot be passive).

Problems I

1. Please complete the tasks highlighted in the notes.
2. Consider a 4-state loop with transition rate matrix M (q is an arbitrary rate constant with dimensions of inverse time):

$$\begin{pmatrix} 0 & q(2-\epsilon)/100 & 0 & \epsilon q/100 \\ \epsilon q & 0 & q(2-\epsilon) & 0 \\ 0 & \epsilon q & 0 & q(2-\epsilon) \\ q(2-\epsilon) & 0 & \epsilon q & 0 \end{pmatrix} \quad (31)$$

M_{ij} is the rate of stepping from i to j. State 1 is a receptor without a ligand; 2, 3 and 4 are ligand-bound states. What does ϵ quantify? How much free energy is dissipated in a loop $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 1$ (see Eq. 21)?

Problems II

3. Use the function "gillespie.m" to write a code for simulating a markov model. Your code needs to take an initial state and repeatedly apply "gillespie.m" to generate a trajectory.
   - "gillespie.m" will generate a transition and a transition time with the appropriate probability. It will return a vector [new state, transition time].
   - "gillespie.m" requires an input of the form [transition rate matrix, old state].
4. Estimate (through simulation) the mean and variance of the time bound in a single binding event (time spent in states 2,3,4 before reaching 1) as a function of $0 \leq \epsilon \leq 1$.
5. Estimate (through simulation) the mean and variance of fractional receptor occupancy during a certain number of binding cycles or over some time period T as a function of $0 \leq \epsilon \leq 1$.

Tasks

1. Identify why the error propagation formula fails on p6 for $N_R = 1$, and calculate the true variance. This simply requires some thought about the approximations inherent to the formula. Output: 1 sentence and a couple of lines of maths.
2. Are cell surface receptors independent? A helpful discussion can be found in Ref [1]. Output: a short paragraph.
3. Complete the gaps in the Berg-Purcell derivation on p9 & p10. Output: about 1 or 2 sides of maths; no special tricks are needed.
4. Prove that the mutual information is non-negative (use Ref. [2] for guidance). Output: a few lines of maths.
5. Prove the data processing inequality. Use the hints on p26. Output: a few lines of maths.

References

[1] H. C. Berg and E. M. Purcell, Biophys. J. 20:193-219 (1977).

[2] http://www.info612.ece.mcgill.ca/lecture_02.pdf

Calculating the error propagation formulae

January 13, 2016

Many of the derivations in the notes are based on error propagation formulae – ie., Eq. 1, 4, 10 and 17. Here I outline an example calculation. The essence of these formulae is always the same: we would like to estimate a quantity y (for us, the external concentration C_L) from a random variable x (something to do with receptor state, eg. R or r_T), and we know that the average of x is related to y by $\bar{x} = f(y)$. Then, following standard error propagation, the relative variance in our basic estimate of y, $\tilde{y} = f^{-1}(x)$ (c_L is our equivalent of $\tilde{y}$), from a single value of x is

$$\left(\frac{\sigma_{\tilde{y}}}{y}\right)^2 = \left(\frac{1}{y}\right)^2 \left(\frac{\mathrm{d}\tilde{y}}{\mathrm{d}x}\right)^2_{x=\bar{x}} \sigma_x^2. \tag{1}$$

For both an instantaneous inference from multiple receptors (Eq. 1) and the Berg-Purcell estimate (Eq. 4), the function f is of the form

$$\bar{x} = f(y) = \frac{ay}{ay+b}. \tag{2}$$

Thus

$$\tilde{y} = \frac{bx}{a(1-x)}. \tag{3}$$

Hence

$$\left(\frac{1}{y}\right)^2 \left(\frac{\mathrm{d}\tilde{y}}{\mathrm{d}x}\right)^2_{x=\bar{x}} = \left(\frac{b\bar{x}}{a(1-\bar{x})}\right)^{-2} \frac{b^2}{a^2(1-\bar{x})^4} = \frac{1}{\bar{x}^2(1-\bar{x})^2}. \tag{4}$$

Generally, we know the average value of x (for Eq. 1 and 4, it's just the receptor occupancy probability p). We then combine this result with our calculation of σ_x^2, which is easy for the instantaneous state of multiple receptors and more involved for time integration, to get the overall error.

In the case of the maximum likelihood expression, $\tilde{y} = c/x$ (Eq. 16), so its even easier to calculate the error propagation term in Eq. 17. Here, σ_x^2 follows from considering the variance of n independent poission processes.

*Note that, formally speaking, there is always a finite possibility of all receptors being occupied, which would give $\tilde{y} \to \infty$. So formally speaking, the naive estimate of $\tilde{y}$ variance and mean are infinite. But in practice, with many receptors, large fluctuations from the average so rare that they don't really contribute to the performance of the cell, and the simple error propagation formula (which ignores extreme fluctuations) is appropriate for understanding the generic behaviour.