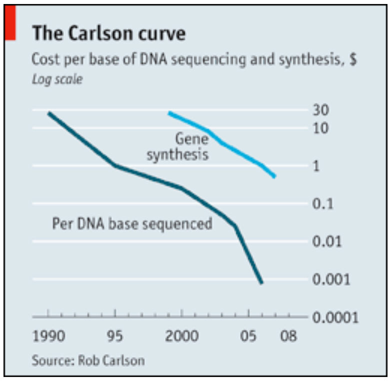
SBM notes

course page

October 5, 2009

In the early days of molecular biology the analogy between electronic-based logic and transcriptional regulation is already commonplace - there is the celebrated lac-operon example (a tiny transcriptional circuit used by E. Coli to wake up suitable enzymes when nutrients change). Today chemists and bioinformaticians have made sequencing and DNA synthesis cheap - see in Fig. **??**).


The Carlson curve

Cost per base of DNA sequencing and synthesis, $

Log scale

Source: Rob Carlson

Figure 1: From The Economist: Hacking goes squishy

This means that there is an opportunity to turn this powerful original metaphor into an actual engineering discipline. One would design parts (eg promoters, coding sequences for sensors, reporters and other useful proteins, see below) and assemble them to obtain new logics or rewire/modify old ones. Which parts, logics, and which modelling methodologies to accompany the design process -in particular how to characterise parts- this is what this course is about.

1 Thermodynamic models of transcription

The following is from a review paper on thermodynamics models of gene regulation [**?**, **?**]. The paper looks at a series of elementary cis-regulatory configurations and develops the associated probability of the $RNAp$ being bounded to the promoter site, from first principles. This is simple but very widely applicable.

The question behind is do TF-based gates and circuits have predictible models? modulo due calibration etc ...

1.1 The idea

The idea is that we have a piece of DNA, say $D(p, s_1, \dots, s_n)$, with a promoter p for the $RNAp$ to bind and some other sites (operators) s_1, ..., s_n for various TFs to bind (contact map).

We want to find an expression for the promoter activity (or the transcription rate) as a function of the concentration of the various factors.

Each bond formed in the complex where D sits, including between bindees, provides a certain additive contribution to the total energy of the D complex. (This preliminary explanation is simplified in two ways: other refined energies are possible, eg for DNA loops -more in a later class; and also we are not considering the background state, ie that of the rest of the system - more in this class).

Write I for the set of such complexes, also called occupancy states of D. Each $i \in I$ has a certain *transcriptional activity* γ_i which reflects the various combination of activators and repressors, and each i also has an energy ϵ_i which is the sum of the energies of its bonds.

Eg $\gamma(D(p^{RNAp}, rep^R))/\gamma(D(p^{RNAp}, rep)) \ll 1$ means that the activity is much smaller when the repressor R is present - which is quantitative way to say that R is a repressor.

We are going to assume that:
- the system is described by a reversible continuous time Markov chain with equilibrium distribution p_i (see below for a general definition), [which means that the associated reactions of binding/unbinding are fast compared to the initiation of transcription];
- and the activity of D (roughly the transcription rate) is the mean activity $\gamma := \sum_i p_i \gamma_i$.

So that all we have to do to determine the rate of transcription is to compute the equilibrium distribution p_i which is given by the Boltzmann law:

$$p_i/p_j = e^{-\beta(\epsilon_i - \epsilon_j)} \tag{1}$$

To have a tractable expression for p_i and γ, we will make some rather strong approximations -see below.

But first let us review some of the basics of continuous time Markov chains (on a finite space).

1.2 CTMC reminder

Suppose a (finite) state space I, and rates $q_{ij} \geq 0$ to go from j to i, for $i \neq j$.

This means that the probability that the chain jumps from j to i within dt is by definition $q_{ij}dt$.

Be careful of the inversion $q_{ij} = q_{i \leftarrow j}$.

So one can write the (linear) differential equation for $p(i)$ the time-dependent probability to be in state i at time t:

$$d/dt\, p(i) = -(\sum_{j \neq i} q_{ji})p(i) + \sum_{j \neq i} q_{ij}p(j) = \sum_{j \neq i}(q_{ij}p(j) - q_{ji}p(i)) \tag{2}$$

We write Q for the matrix such that $d/dt\, p = Qp$ (p is seen as a column vector of size $n \times 1$).

There is always a unique solution to the steady state equation $Qp = 0$ (when Q is strongly connected?), for $i \in I$, it is called the *steady state distribution.*

Sometimes the stationary probability is an *equilibrium* meaning for $i < j \in I$:

$$q_{ij}p(j) = q_{ji}p(i) \tag{3}$$

In words Eq. (**??**) says that at equilibrium the probability to see a jump from j to i equals that of seeing a jump from i to j. Eq. (**??**) is also called *detailed balance.* (By contrast, at steady state one has a weaker property: the probability to see a jump to i equals that of seeing a jump to i)

The steady state probability associated to Q is an equilibrium iff:
- *[Reversibility]* $q_{ij} = 0 \Rightarrow q_{ji} = 0$,
- *[Wegscheider]* for every cycle C in the *undirected* transition graph (defined as $i, j \in E$ iff $q_{ij} > 0$) one has $\prod_C q_{ij}/q_{ji} = 1$.

The only if part is easy.

Given a real-valued *energy* function ϵ defined on I, and a $\beta \geq 0$, one can define a unique probability on I such that $p_\epsilon(i)/p_\epsilon(j) = e^{-\beta(\epsilon_i - \epsilon_j)}$ (also depends on β).

Equivalently $p_\epsilon(i) = e^{-\beta\epsilon_i}/\sum_j e^{-\beta\epsilon_j}$ *[Boltzmann].*

Note that $e^{-\beta(\epsilon_i - \epsilon_j)} \geq 1$ iff $\epsilon_i \geq \epsilon_j$ that is to say the energy decreases when jumping from j to i; therefore $p_\epsilon(i) \geq p_\epsilon(j)$ iff $\epsilon_i \geq \epsilon_j$. Note also that at $\beta = 0$ (infinite temperature), p_ϵ is the uniform distribution.

Now p_ϵ is the equilibrium of Q iff $q_{ij}/q_{ji} = e^{-\beta(\epsilon_i - \epsilon_j)}$. This only determines q_{ij}/q_{ji}, hence Q, up to a scalar.

In conclusion it is equivalent to say that Q has an equilibrium, and to say that it has an energy function - in which case the equilibrium is given by the Boltzmann expression.

1.3 Trivial case (no TFs)

To return to our problem, consider $D(p)$, and suppose one has n_p (free) copies of $RNAp$:
- $I = \{\varnothing, i\}$, for $i < n_p$ the bound $RNAp$,
- $q_{i\varnothing}/q_{\varnothing i} = k^+/k^- = e^{-\beta\Delta\epsilon}$, where $\Delta\epsilon = \epsilon_b - \epsilon_f$ is the energy gained (usually negative) from binding one $RNAp$ to p, $k^\pm$ are the on- and off-rates for this bond.

So writing p_{busy} for the probability that any $RNAp$ is bound:

$$\begin{aligned} p_{busy}/p_{free} &= n_p e^{-(p-1)\beta\epsilon_f} e^{-\beta\epsilon_b}/e^{-p\beta\epsilon_f} = n_p e^{-\beta\Delta\epsilon} \\ p_{busy} &= 1/(1 + 1/(n_p e^{-\beta\Delta\epsilon})) \end{aligned}$$

One sees that p_{busy} is an increasing function of n_p, the more transcriptional machines are standing by, the more busy is the promoter; also, it is a decreasing function of $\Delta\epsilon$, the less favourable is the binding energetically, the less busy the promoter.

We can do the same computation under a more realistic assumption by introducing N competing non-specific sites, with bond energy ϵ_{ns} -and supposing there are *no* free $RNAp$s.

Then I becomes $[n_p; N+1]$ where we write $[p; n]$ for the set of injections from p to n.

One has $[p; n] = n!/(n-p)!$, and so $[p+1; n] = [p; n](n-p)$.

Indeed $[0; n] = 1$, $[1; n] = n$, $[2; n] = n(n-1)$, ...

The states can be partitioned between the free and busy states as in the first example:
- free states: energy $n_p\epsilon_{ns}$, number $[n_p; N]$
- busy states: energy $n_p\epsilon_{ns} + \epsilon_b - \epsilon_{ns}$, number $n_p[n_p - 1; N] = n_p[n_p; N]/(N - n_p + 1)$.

So the respective weights (defined up to a scalar) of the free and busy class are:
- $[n_p; N]e^{-\beta n_p \epsilon_{ns}}$
- $n_p[n_p; N]/(N - n_p + 1)e^{-\beta n_p\epsilon_{ns}} e^{-\beta(\epsilon_b - \epsilon_{ns})}$

Therefore:

$$\begin{aligned} p_{busy}/p_{free} &= n_p e^{-\beta(\epsilon_b - \epsilon_{ns})}/(N - n_p + 1) \simeq (n_p/N)e^{-\beta\Delta\epsilon} \\ p_{busy} &= 1/(1 + 1/(\tfrac{n_p}{N} e^{-\beta\Delta\epsilon})) \end{aligned}$$

If we compare with the first expression, we see a competitive factor N appearing; this means that the promoter activity will decrease if competition -ie N- increases.

(TD: compare this computation with the one where one applies pre-symmetries on the state space; do the same thing for a general model where $RNAp$s can be free and bound to non specific sites.)

We wish to derive more general expressions of the form

$$p_{busy} = 1/(1 + 1/(F\frac{n_p}{N}e^{-\beta\Delta\epsilon}))$$

where F is the *regulation factor*, $F > 1$ if the cis-regulation is globally favourable to transcription.

1.4 Derivation for a simple case

Let us see if we can derive similar expressions for the case of an activator.

Consider a $D(a, p)$ where:
- *[activator]* CRP binds a with energy $-\Delta\epsilon_a$, has n_c copies,
- *[RNA polymerase]* $RNAp$ binds p with energy $-\Delta\epsilon_r$, has n_r copies,
- *[cooperation]* and an additional CRP, $RNAp$ cooperation term ϵ_{ar},
- *[background]* N non specific binding sites with energy ϵ.

Set $\binom{N}{n,m} := \binom{N}{n+m}\binom{n+m}{n}$ the number of unordered pairs of disjoint subsets of respective size n, and m in a set of size N (symmetric in n, m). This is the number of ways in which our CRPs and $RNAp$s can bind to non specific sites.

There are four local occupancy states -supposing that when both CRP and $RNAp$ are present, they bind instantly.

Their different weights are:
- *[empty]* $\binom{N}{n_c,n_r}e^{-\beta(n_c+n_r)\epsilon}$
- *[CRP]* $\binom{N}{n_c-1,n_r}e^{-\beta(n_c+n_r)\epsilon}e^{-\beta\Delta\epsilon_a}$
- *[RNAp]* $\binom{N}{n_c,n_r-1}e^{-\beta(n_c+n_r)\epsilon}e^{-\beta\Delta\epsilon_r}$
- *[both]* $\binom{N}{n_c-1,n_r-1}e^{-\beta(n_c+n_r)\epsilon}e^{-\beta(\Delta\epsilon_a+\Delta\epsilon_r+\epsilon_{ar})}$

If N is large, $\binom{N}{n,m-1}/\binom{N}{n,m} = m/(N-n-m+1) \simeq m/N$. Similarly $\binom{N}{x-1}/\binom{N}{x} \simeq x/N$.

So for large Ns, the above can be simplified (and normalised by the free weight):
- *[empty]* 1
- *[CRP]* $\frac{n_c}{N}e^{-\beta\Delta\epsilon_a}$
- *[RNAp]* $\frac{n_r}{N}e^{-\beta\Delta\epsilon_r}$
- *[both]* $\frac{n_c}{N}\frac{n_r}{N}e^{-\beta(\Delta\epsilon_a+\Delta\epsilon_r+\epsilon_{ar})}$

Now the probability that $RNAp$ is bound is:

$$\frac{\frac{n_r}{N}e^{-\beta\Delta\epsilon_r}(1+\frac{n_c}{N}e^{-\beta(\Delta\epsilon_a+\epsilon_{ar})})}{1+\frac{n_c}{N}e^{-\beta\Delta\epsilon_a}+\frac{n_r}{N}e^{-\beta\Delta\epsilon_r}+\frac{n_c}{N}\frac{n_r}{N}e^{-\beta(\Delta\epsilon_a+\Delta\epsilon_r+\epsilon_{ar})}}$$

1.5 Parenthesis on reaction classes

The situation will be the same when one goes from reactions to rules. Suppose $\epsilon(x)$ is defined as the number of edges in x, where x is a site graph (perhaps ϵ is weighted per type of edge). This coarse energy is compatible with any association/dissociation pair.

Now consider r_1 which dissociates $A(c, b^1), B(a^1, c)$ with eq dissociation constant K_1, and r_2 which dissociates the same edge in a different context $A(c^2, b^1), B(a^1, c^3), C(a^2, b^3)$ with K_2. The energy difference $\epsilon(r_i \cdot x) - \epsilon(x) = \epsilon(a, b)$ is the same for both rules, hence one must have $K_1 = K_2$ for p_ϵ to be the equilibrium of the rule set r_1, r_2.

2 Rule-based modeling of standard biological parts

Two early papers proved the synbio concept: reversible switch, oscillator (Nature 2000).

However, people felt the need for standard parts, well described; with generic conventions to concatenate sequences; iGEM competition and Bio-Bricks was born (2001?).

Today, we see that biological parts resist modularity in many ways. Sometimes there are endogenous interferences between various devices. Eg pairs of transcriptional regulators and promoters, that individually seems to work perfectly, might interact in unwanted ways when put together. The USTC entry in the iGEM'07 competition tried to address this problem by designing non interfering transcriptional wires. Sometimes there are exogenous problems, eg the host will recognise elements of the construct and degrade them - in BioBricks speech the host is not always a benevolent chassis.

This situation where parts refuse to be modular (familiar from the field of evolvable hardware [**?**]) clearly stands in the way of the rational engineering of biological systems. One of the fundamental challenges to synthetic biology is to engineer simple parts with a well-understood and well-described behaviour in relation to other parts -a description which constitutes the part data sheet [**?**,**?**]. This means one needs high-resolution measurement protocols in controlled environments and associated model calibration methods. But one also needs -and this point often receives less attention- a modeling language for the formalisation of the interactions that the data sheet is attempting to furnish parameters for in the first place.

We propose to use rule-based modeling [**?**, **?**, **?**, **?**, **?**] to do exactly that. That is to say we propose that a data sheet should be interpreted -with a resolution that can be tuned depending on the particulars- as a set of rules describing the ways in which the part interacts with other parts present in the design, as well as with the services offered by the host (possibly a cell-free environement). This way one makes completely clear what the part is supposed to be doing.

Fig1 illustrates this concept with the familiar example of the transcription of a ribosome binding site coding sequence. We choose to represent DNA parts by agents with:
- two lateral sites to bind other DNA parts upstream and downstream,
- one to hold the type of the agent (materialised here by its reference number in the BioBrick registry)
- and one to bind proteins interacting with the agent (typically transcription factors and the RNA polymerase).

The transformation depicted in Fig1 is called a rule, or sometimes also a reaction class [?], and not a reaction because not all sites are shown. This rule expresses the fact that to a good approximation (as endorsed by the BioBricks foundation), the RNA polymerase does not care what lies downstream.

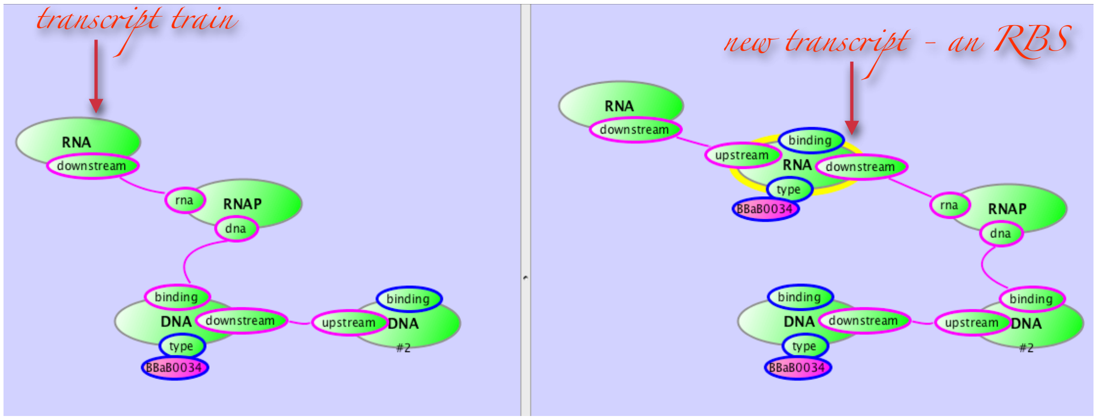


Figure 2: *A possible rule-based description of transcription (using the Kappa language). The RNAp agent representing the RNA polymerase binds a genetic element -BBa0034 (representing a part of the BioBrick registry derived from Elowitz' repressilator)- and can therefore transcribe it (see the new transcript in the right hand side of the rule).*

Fig2 presents another example which illustrates the flexibility afforded by rules. This particular rule expresses the fact the RNA polymerase might fall off the DNA -with a certain likelihood which is to be determined experimentally. This happens instead of continuing transcription in situations where a polymerase is blocked due to a protein or another polymerase. Likewise, detailed and realistic rules can be written for combinatorial promoters [?].

In addition, for a successful modeling approach to *mammalian* synthetic biology one needs models of various forms of control among which features prominently the combinatorial control of mammalian promoters and the epigenetic control of transcription. The explicit representation of domain to domain binding used in Kappa -the specific rule-based language used here [?]- meshes well with the representation of such processes. In fact, rule-based techniques have been used recently to formulate a simple model of epigenetic information repair which should be a suitable starting point [?]. There are other means of affecting protein stability and/or activity such as ubiquitination, sumoylation, etc. the modelling of which needs to be explored in a similar way.

One key property of rules beyond their concision, ease of use and tuneable grainedness is that they allow a compositional approach. Just as parts can be put together, their associated rule sets can, and will generate implicitly the dynamical system of interest. This has been done already in the context of ODE modelling [?]. A related advantage is that one never needs to enumerate the species that a set of rules might produce, which can be useful when designing even moderatly combinatorial circuits. The underlying notion of dynamics is inherently stochastic (a continuous-time Markov chain) and as such particularly suited to the modeling of the noisy transcriptional devices one uses in synthetic biology.

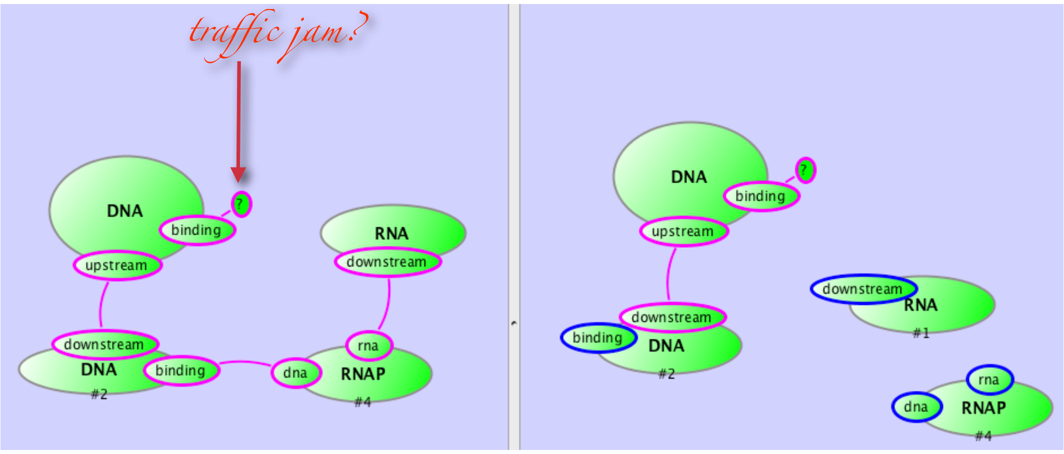


Figure 3: *The RNAp agent (representing the RNA polymerase) might fall off DNA if the next DNA element downstream is already bound (typically by a repressor or another RNAp agent). This causes the mRNA to be liberated from the RNAP to be degraded later (see right hand side) and prevents the buildup of RNAPs on DNA.*

One can imagine using rule-based modeling to increase the realism and hence the yield of modeling during the design cycle of a synthetic biological construct. To do so one needs to:
- 1) conceive and implement a quick and agile way to map a BioBrick-like design - including generic bio-sensors, transcriptional parts, and epigenetic ones- to a stochastic model,
- 2) adapt and implement existing numerical strategies to calibrate such models,
- 3) develop and implement scaleable methods for their stochastic simulation,
- 4) conceive and implement an open access on-line registry of such virtual parts for everyone to use, remodel and calibrate.

3 Next classes (tentative)

- methode de Stelling pour les ODE

- model-driven engineering/data-driven modelling de Jim Collins

- Chris French: BioBricks and IGEM concepts

- Alistair Elfick: measurement, characterisation of parts for synthetic biology

- Kim de Mora: how to choose a synthetic biology project with case studies looking at good/bad igem projects

- Scott Cockroft: chemical aspects of synthetic biology, DNA synthesis

- Jane Calvert: societal discussion topics

- exercise in Kappa a la Ty Thomson: reconstruct the promoter repressed by LacI and activated by CI

- Elaine Murphy: lecture on refined RBM techniques for bio-bricks modelling

- Lev: MD and computational design of modified receptors (eg)

- Ricardo: artificial mamallian chromosome?

References

[1] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev*, 15(2):125–35, Apr 2005.

[2] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15(2):116–24, Apr 2005.

[3] B. Canton, A. Labno, and D. Endy. Refinement and standardization of synthetic biological parts and devices. *Nat Biotechnol*, 26(7):787–793, Jul 2008.

[4] P.-L. Curien, V. Danos, J. Krivine, and M. Zhang. Computational self-assembly. *Theoretical Computer Science*, 404(1–2):61–75, Sep 2008.

[5] V. Danos. Agile Modelling of Cellular Signalling. *Computation in Modern Science and Engineering, Volume 2, Part A*, 963:611–614, Sep 2007.

[6] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling of cellular signalling. In L. Caires and V. Vasconcelos, editors, *Proceedings of the 18th International Conference on Concurrency Theory (CONCUR'07)*, volume 4703 of *LNCS*, pages 17–41, Sep 2007.

[7] V. Danos, J. Feret, W. Fontana, R. Harmer, and J. Krivine. Rule-based modelling, symmetries, refinements. In Springer, editor, *FMSB 2008*, volume 5054 of *LNBI*, pages 103–122, Jun 2008.

[8] V. Danos, J. Feret, W. Fontana, and J. Krivine. Abstract interpretation of cellular signalling networks. In F. L. et al., editor, *VMCAI'08*, volume 4905 of *LNCS*, pages pp. 83–97. Springer, Jan 2008.

[9] J. Feret, V. Danos, R. Harmer, J. Krivine, and W. Fontana. Internal coarse-graining of molecular systems. *PNAS*, 106(16):6453–8, Apr 2009.

[10] D. Forger and C. Peskin. A detailed predictive model of the mammalian circadian clock. *Proceedings of the National Academy of Sciences*, 100(25):14806–14811, 2003.

[11] N. J. Guido, X. Wang, D. Adalsteinsson, D. Mcmillen, J. Hasty, C. R. Cantor, T. C. Elston, and J. J. Collins. A bottom-up approach to gene regulation. *Nature*, 439(7078):856–860, Feb 2006.

[12] J. Krivine, V. Danos, and A. Benecke. Modelling epigenetic information maintenance - a kappa tutorial. In *Proceedings of CAV'09*, Jul 2009.

[13] M. Marchisio and J. Stelling. Computational design of synthetic gene circuits with composable parts. *Bioinformatics*, Jan 2008.

[14] A. Thompson. An evolved circuit, intrinsic in silicon, entwined with physics, 1996.