COMPUTER ALGEBRA IN SYSTEMS BIOLOGY

REINHARD LAUBENBACHER AND BERND STURMFELS

Abstract. Systems biology focuses on the study of entire biological systems rather than on their individual components. With the emergence of high-throughput data generation technologies for molecular biology and the development of advanced mathematical modeling techniques, this field promises to provide important new insights. At the same time, with the availability of increasingly powerful computers, computer algebra has developed into a useful tool for many applications. This article illustrates the use of computer algebra in systems biology by way of a well-known gene regulatory network, the *Lac Operon* in the bacterium *E. coli*.

1. Systems Biology

Molecular biology has undergone a dramatic revolution during the second half of the twentieth century, beginning with the discovery of the structure of DNA. Since then a series of technological advances has given experimentalists the ability to make ever-more detailed measurements of an increasing number of molecular components of the cell. DNA microarrays, for instance, are small silicon chips spotted with short segments of DNA that can be used to measure the activity levels of thousands of different genes in tissue samples simultaneously. Soon it might be possible to make large-scale quantitative measurements in a single cell. Being able to take such global snapshots of molecular processes has opened up the possibility of studying the changes that are constantly going on in cells as a coherent dynamical system with intricately interacting parts, rather than studying the parts in isolation. Thus, the new field of *systems biology* has emerged [1; 11; 21].

Biological networks tend to be highly complex, with many variables that interact with each other in nonlinear ways, making it difficult to study such systems without the help of sophisticated mathematical tools and concepts. It is even unclear what the right formal language should be for the description of molecular systems [14]. A characteristic feature of systems biology research is its heavy use of mathematical methods. One tool which has been applied recently to biological problems is *computer algebra*, a field of

Date: December 27, 2007.

Key words and phrases. biochemical network, systems biology, computer algebra.

Reinhard Laubenbacher acknowledges support by the National Science Foundation (DMS-051144) and the National Institutes of Health (RO1 GM068947-01). Bernd Sturmfels was supported by the National Science Foundation (DMS-0456960) and the DARPA Program *Fundamental Laws of Biology*. We are grateful to Lior Pachter for his comments.

mathematics that combines the ability of computers to carry out symbolic calculations with concepts from abstract algebra. Computer algebra has been used in the life sciences in a variety of ways, such as the construction of phylogenetic trees encoding the evolutionary relationship between different species [5; 6], or the construction and analysis of models of intracellular biochemical networks [13; 22]. For many more such applications see [3; 17].

2. Computer Algebra

Computer algebra provides tools for computing with symbols rather than with floating point numbers. Software systems for computer algebra include familiar commercial packages, such as `Maple`, `Mathematica`, or `Magma`, as well as a wide range of more specialized systems, many of which are free and often run faster on specialized tasks. One important theme in computer algebra is the solution of non-linear algebraic equations. In the context of systems biology, this problem arises when one wishes to compute the steady states of a dynamic model. As an example, consider the following system of two equations where x and y are the unknowns and k_1 and k_2 are parameters:

$$x^2 + k_1xy - 1 \quad = \quad y^2 + k_2xy - 1 \quad = \quad 0.$$

Using a computer algebra technique known as *Gröbner bases* [8; 17; 19], the two given equations can easily be rewritten in the following equivalent form:

$$(k_1k_2-1)y^4 + (k_2^2-k_1k_2+2)y^2-1 \;=\; k_2x+(1-k_1k_2)y^3+(k_1k_2-k_2^2-1)y \;=\; 0.$$

The first equation involves no x. Using the quadratic formula, we can therefore express y in terms of the parameters k_1, k_2. The second equation gives x in terms of y and k_1, k_2. Further analysis reveals that there are always four real solutions (x, y) if $k_1k_2 < 1$ but only two real solutions if $k_1k_2 > 1$.

For a second example, consider the following equations in five unknowns which are derived from the discrete model for the Lac Operon in Section 4:

$$M = A, B = M, A = A+LB+ALB, L = P+L+LB+LP+LPB, P = M.$$

Here we are seeking solutions whose coordinates are 0 or 1 and where $1 + 1$ is redefined to be 0. Thus, we are working over the field with two elements. A Gröbner basis for the given system consists of the simplified equations

$$B \;=\; A \;=\; M \;=\; P^2 \;=\; LP \;=\; P,$$

and from this we see that there are precisely three solutions:

$$(M, B, A, L, P) \;=\; (0,0,0,0,0),\;\; (0,0,0,1,0),\;\text{ and }\;\; (1,1,1,1,1).$$

Of course, this answer could have been found easily without Gröbner bases, for instance, by plugging in all 32 binary vectors of length 5. However, the types of discrete dynamical systems that are of interest in biology are now much more complex (due to advances in the experimental technologies as argued above). For such systems, naive approaches will not work, and more sophisticated tools, such as computer algebra, are needed for the analysis.

3. The Lac Operon

We illustrate the use of computer algebra in systems biology by way of a gene regulatory network which was discovered by Jacob and Monod [9], who received the 1965 Nobel Prize in Medicine for this discovery. The *E(scherichia) coli Lac(tose) Operon* is one of the earliest and best understood examples of gene expression regulation [12; 15; 16]. Gene regulation in bacteria serves the cells to adjust to changes in the nutritional environment so that their growth and division can be optimized. E. coli can use glucose or lactose as energy and source of carbon. When cells grow in glucose-based medium, the activity of the enzymes involved in the metabolism of lactose is very low, even if lactose is available. However, when glucose is exhausted from the medium and lactose is present, the activity of enzymes involved in lactose metabolism increases. This process is called *induction* [15].

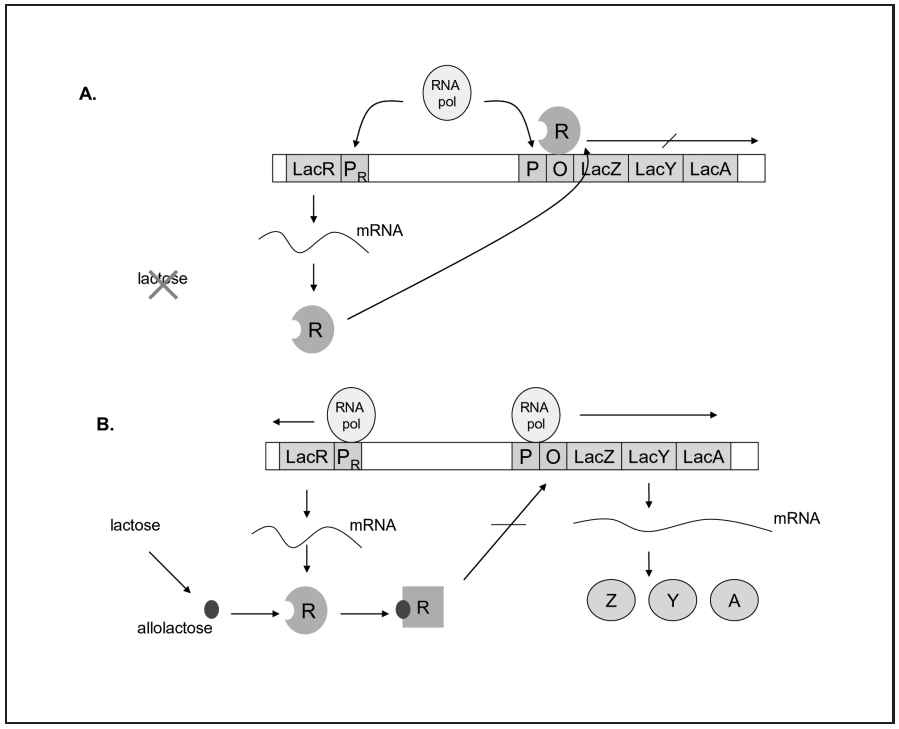

Figure 1. The *Lac operon*.

A group of genes that are regulated by a common promoter and operator is called an *operon*. Such genes are typically organized in a tandem arrangement. Operons also contain control elements – transcription factors – that bind to regulatory elements in the DNA and activate or inhibit the transcription of structural genes. Transcription factors that stimulate transcription are called inducers. They bind to regulatory elements in DNA

called promoters. Repressors, on the other hand, bind to elements in DNA called operators and they are involved in the repression of transcription.

The lac operon (Figure 1) contains structural genes for three enzymes involved in the metabolism of lactose (LacZ, LacY, LacA), one structural gene encoding a repressor protein (LacR), and three control elements involved in the regulation of transcription. The LacY gene encodes lactose permease, which is involved in the transport of lactose into the cell, LacZ encodes β-galactosidase, an enzyme that converts lactose into glucose and galactose, sugars that will be further metabolized by the cell, and LacA encodes thiogalactoside transacetylase, an enzyme of still unknown function.

How is the lac operon regulated? The structural genes LacZ, LacY and LacA are only expressed when lactose is present in the cell. In the absence of lactose, the lac-repressor R binds to the operator region O, and RNA polymerase, bound to the promoter P, is unable to move past this region. Hence, no transcription of LacZ, LacY and LacA occurs (Figure 1.A). When lactose enters the cell, it is converted into a similar molecule (isomer) called allolactose, also by the action of β-galactosidase. Allolactose is the inducer of the lac operon, binding to the lac-repressor R and inducing a conformational change that prevents R from binding to the operator region. The RNA polymerase is able to move along the DNA, transcription of the three genes occurs, and lactose is metabolized (Figure 1.B).

4. A Discrete Model

We first present a discrete model for this gene regulation network, in the form of a *Boolean network*, taken from [18]. Like all models, it is very simplified in its representation of biological details and mechanisms. What we are attempting here is to capture a basic dynamic feature of the *Lac Operon*, namely its bistability. Simply speaking, this means that the operon is either ON or OFF, each resulting in a single steady state of the system.

The model has five variables, representing the concentrations of (1) mRNA for the genes LacZ, LacY, and LacA (M), (2) intracellular allolactose (A), (3) β-galactocidase (encoded by LacZ) (B), (4) intracellular lactose (L), and (5) lactose permease (P). The model is qualitative, in the sense that it uses a very coarse-grained measurement of these concentrations, keeping track only of the absence (0) or presence (1) of these chemical species.

The model assumes that the molecular mechanisms leading from activation of a gene to the production of the corresponding protein (transcription plus translation) happen in one time step, as does mRNA and protein degradation. It also assumes that extracellular lactose is always available. The relationships between the variables are expressed in terms of logical formulas, one for each variable. For instance, the biological mechanisms leading to transcription of the LacZ, LacY, and LacA genes depend on the presence of allolactose, needed to block the action of the repressor gene. That is, the Boolean function controlling the state of the Boolean variable M is $f_M = A$.

Similarly, the structure of the other functions can be derived as follows:

$$f_B = M,\ \ f_A = A \vee (L \wedge B),\ \ f_L = P \vee (L \wedge \neg B),\ \ f_P = M.$$

The function f_A, for instance, indicates that allolactose is present at time $t+1$ if it was present at time t, or if lactose and β-galactosidase were present at time t, which then react to produce allolactose at time $t+1$.

We now show that this simple model, based on very few assumptions, displays a dynamic behavior that captures an essential feature of the lac operon, namely, bistability. Our analysis is tantamount to examining the long term dynamics, or steady states and periodic states, of the model. A state of the system is represented through a binary 5-tuple (M, B, A, L, P), such as $(0, 0, 1, 0, 1)$. This particular 5-tuple indicates a state in which allolactose and lactose permease are present and all other molecular species are absent. We compute the time evolution of the system by applying the five Boolean functions to this state. This results in the terminating trajectory

$$(0, 0, 1, 0, 1) \to (1, 0, 1, 1, 0) \to (1, 1, 1, 1, 1) \to (1, 1, 1, 1, 1).$$

That is, the system reaches the steady state in which all substances are present. We would like to compute all such steady states for the system. This can be done polynomial computation tools provided by computer algebra.

We translate the Boolean functions in the model into polynomials. This uses the binary field $\mathbb{F}_2 = \{0, 1\}$, that is, arithmetic modulo 2. To translate a Boolean function into a polynomial function, we observe first that every Boolean function is expressed using the logical operators $\wedge, \vee$, and $\neg$. These can be translated into polynomial operations by simply observing that the functions $a \wedge b$ and $a \cdot b$ take on the same Boolean values for given values of the variables, that is, both functions take on the value 1 precisely if both a and b take on the value 1, otherwise the functions take on the value 0. Similarly, we see that $a \vee b = a + b + a \cdot b$ and $\neg a = a + 1$. If we apply this dictionary to the Boolean functions in our model we obtain the following:

$$\begin{aligned} f_M &= A, \\ f_B &= M, \\ f_A &= A + LB + ALB, \\ f_L &= P + L + LB + LP + LPB, \\ f_P &= M. \end{aligned}$$

A steady state of the system is one for which the functions do not change the value of the variables. That is, if (M, B, A, L, P) is a steady state, then $f_M(M, B, A, L, P) = M$, and similarly for the other four functions. A steady state is therefore a solution to the system of polynomial equations

$$f_M = M,\ f_B = B,\ f_A = A,\ f_L = L,\ f_P = P.$$

We solved this system in Section 2 and found a total of three steady states:

$$(1, 1, 1, 1, 1),\ \ (0, 0, 0, 0, 0),\ \ (0, 0, 0, 1, 0).$$

The first steady state was observed in our trajectory. The second one also makes sense, but the third one is biologically not meaningful since it would imply that the bacterium does not metabolize the intracellular lactose present. This is an indication that our model is not entirely accurate and needs to be modified. A first step is to compare model dynamics to known biological properties. Another test could be to see if the model fits available experimental data, which is beyond the scope of this paper. Our model is so small that we can depict (in Figure 2) all possible state transitions. Construction of such a diagram is impossible for larger models.

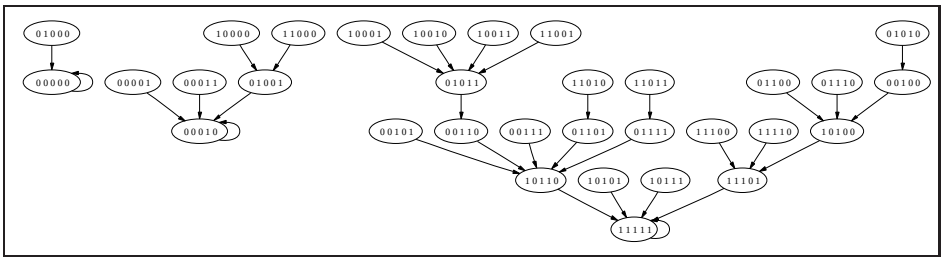

FIGURE 2. The dynamics of the first model.

Studying the lac operon further, we see that one problem with the model is that it does not represent all of the molecular species that influence the dynamics. We leave it to the reader to verify that the following model is biologically more accurate, for instance, using the biological materials posted at the website [16]. The new model, presented in [18], has eight variables measuring the concentration of (1) mRNA for the genes LacZ, LacY, and LacA (M), (2) intracellular allolactose (A), (3) β-galactocidase (encoded by LacZ) (B), (4) the lac repressor (encoded by LacI) (R), (5) intracellular lactose (L), and (6) lactose permease (P). We also need variables for (7) intracellular glucose (G), and (8) extracellular lactose (E).

The logical functions encoding the model structure are as follows:

$$f_M = (\neg R) \wedge (\neg G),\ f_A = E \vee (L \wedge B),\ f_B = M,$$
$$f_R = \neg A,\ f_L = P \wedge E,\ f_P = M,\ f_E = E,\ f_G = G.$$

Note that E and G do not depend on the other variables. External lactose is either present or absent and similarly for internal glucose.

If we now analyze the model dynamics, either through the use of computer algebra methods as before or enumeration of all state transitions (which is possible in this case, using tools like the one at `http://dvd.vbi.vt.edu`) we see that the model is more faithful to the biology of the system. For instance, if external lactose is available ($E = 1$) and internal glucose is absent ($G = 0$), then we obtain the unique steady state

$$(M, B, A, L, P, E, G) = (1, 1, 1, 1, 1, 1, 1, 0),$$

which is what one would expect from biological considerations.

5. A Continuous Model

As the oldest known gene regulation network, the lac operon has been studied extensively and many different mathematical models have been constructed for it [20]. The most common type of model is based on ordinary differential equations. For a model from the recent literature see [22]. We discuss here the very simple dynamical systems model described in Section 5.2 of the undergraduate text book [4]. It consists of three equations, modeling the concentration of R and the rates of change of M and A (with variable meanings as in the last section). The three equations are:

$$\begin{aligned} R &= \frac{1}{1+A^n}, \\ \frac{dM}{dt} &= c_0 + c(1-R) - \gamma M, \\ \frac{dA}{dt} &= ML - \delta A - \frac{vMA}{h+A}. \end{aligned}$$

Here $c_0, c, \gamma, v, \delta, h$ and L are certain model parameters, n is a fixed positive integer, and the concentrations R, M and A are functions of time t.

This model is also based on several assumptions. For instance, we do not distinguish between intracellular and extracellular lactose, and denote both by L. Another assumption is that β-galactosidase is proportional to operon activity M and is not represented explicitly. The concentration of the repressor R is represented by a sigmoid function, a so-called *Hill function*. When extracellular lactose is present and is transported to the intracellular environment by lactose permease, produced by the activity of the operon, the allolactose concentration increases and inhibits the repressor R. The rate of change of the gene transcripts M is composed of a baseline activity represented by the constant c_0, the concentration A of allolactose (which inhibits the repressor R), and a degradation term γM. The concentration of allolactose A increases with the activity M of the operon genes in conjunction with the presence of lactose L. Its degradation (the terms on the right-hand side with minus signs) is represented by a Michaelis-Menten type enzyme substrate reaction composed of two terms. The parameters $c_0, c, \gamma, v, \delta, h$ and L need to be estimated using biological considerations or numerical methods, to ensure that the model is consistent with experimental data.

This model is also quite simplified, both from a biological and a mathematical point of view. But even a simple model can be useful. The purpose of modeling is to identify the essence of a system, that is to identify the components and dynamics that are key to conferring the biological function. This is like identifying that the engine is what pushes the bus forward. The art of constructing mathematical models of biological systems (or any other type of system, for that matter) is to incorporate the most important features and mechanisms and discard the irrelevant ones. Comparing the model to the Boolean network model constructed in the previous section,

we can see certain basic similarities, even though the mathematics is different. We have a time-discrete finite dynamical system on the one hand and a continuous-time system given by differential equations on the other hand.

It is now time to analyze the dynamics of the continuous model, just as we analyzed the discrete model, by computing its steady states. We do this by again phrasing the problem in way that makes it amenable to using algebra, namely by setting the right hand sides of the differential equations to zero:

$$c_0 \,+\, c \cdot (1 - \frac{1}{1+A^n}) \,-\, \gamma \cdot M \;\;=\;\; M \cdot L \,-\, \delta \cdot A \,-\, \frac{vMA}{h+A} \;\;=\;\; 0.$$

This is a system of two algebraic equations in two variables A and M, which depends on the various parameters. Note that the steady state values for the missing variable R are determined by the equation $R = 1/(1+A^n)$.

Following the discussion in [4, Section 5.2], we leave the concentration L of lactose unspecified while the other parameter values are fixed as follows:

$$c = \gamma = v = 1,\; c_0 = \frac{1}{20},\, h = 2,\, m = 5,\, \delta = \frac{1}{5}.$$

We also set $n = 5$. Our algebraic equations now take the form

$$\frac{1}{20} \,+\, \frac{A^5}{1+A^5} - M \;\;=\;\; M \cdot L - \frac{1}{5} \cdot A \,-\, \frac{MA}{2+A} \;\;=\;\; 0.$$

By clearing denominators and eliminating the unknown M, we find that

$$4\,A^7 \,+\, (29 - 21L)\,A^6 \,-\, 42L\,A^5 \,+\, 4\,A^2 \,+\, (9-L)\,A \,-\, 2L \;\;=\;\; 0.$$

This is a polynomial of degree 7 in A. The discriminant of this polynomial in A is a complicated polynomial of degree 12 in the parameter L. This discriminant has precisely two positive roots, which we determine to be

$$L_1 \,=\, 0.68453896581348... \quad \text{and} \quad L_2 \,=\, 1.5105398398447....$$

For all values of L between L_1 and L_2, there are three positive steady states. For example, if $L = 1$ then the steady states (R, M, A) of our system are

$$(0.2272, 0.0506, 0.9994)\,,\;\; (0.6907, 0.1859, 0.8642)\,,\;\; (2.3717, 1.0368, 0.0132).$$

The above expression $4A^7 + (29-21L)A^6 + \cdots$ is the equation of the bifurcation diagram in the (A, L)-plane which is depicted in [4, Figure 5.3(b)]. It describes the steady-state allolactose concentration A as a function of the lactose concentration L. As argued in [4, Section 5.3], the emergence of these three steady states shows that this model correctly captures key features of the lac operon. Computer algebra allows us to vary other parameters and enables us to conduct a very careful analysis of the dynamics of this model. In particular, using computer algebra, we can derive a precise algebraic description of the region in parameter space for which the dynamical system has more than one stationary point, and we can identify parameter values at which interesting phenomena (e.g. Hopf bifurcations [10]) might occur.

6. Discussion

It is generally agreed that modern molecular biology can benefit greatly from the use of new mathematical techniques that allow the construction of system-level sophisticated models of biological networks. Conversely, the problems that arise in today's biological research can provide important stimuli for mathematical research. This is aptly expressed in the title *Mathematics is Biology's Next Microscope, Only Better; Biology is Mathematics' Next Physics, Only Better* of a recent article [7]. We have attempted here to describe through mathematical models of the lac operon how algebra can contribute to a formal description and an analytical understanding of biological phenomena. One goal was to show that different types of mathematical models (discrete and continuous) can provide insight into biological mechanisms. Furthermore, we have demonstrated that computer algebra, not traditionally used in biology, is a powerful tool that can help construct and analyze biological models. Thus, this paper should be viewed as an advertisement for an in-depth study of the relationship between computer algebra in particular, and mathematics in general, and systems biology. The marriage between the two promises to be extremely fruitful for both.

One forum for such interactions is the annual international conference series on *Algebraic Biology* [2] which was started in 2005. Another one is the special program on *Algebraic Methods in Systems Biology and Statistics* which will be held at the Statistical and Applied Mathematical Sciences Institute (SAMSI) in North Carolina during the academic year 2008-09.

References

[1] U. Alon: *An Introduction to Systems Biology: Design Principles of Biological Circuits*, Chapman and Hall, 2006.

[2] H. Anai and K. Horimoto: *Algebraic Biology 2005*, Proceedings of the First International Conference on Algebraic Biology – Computer Algebra in Biology – held on November, 28–30, 2005, in Tokyo, Japan, Universal Academy Press.

[3] M. Barnett: *Computer algebra in the life sciences*, ACM SIGSAM Bulletin, Volume 36, Issue 4 (December 2002), pp. 5–32.

[4] R. D. Boer: *Theoretical Biology*, Undergraduate Course at Utrecht University, book posted at `http://theory.bio.uu.nl/rdb/books/`.

[5] M. Casanellas and J. Fernández-Sánchez: *Performance of a new invariants method on homogeneous and non-homogeneous quartet trees*, Molecular Biology and Evolution **24** (2007) 288–293.

[6] B. Cipra: *Algebraic geometers see ideal approach to biology*, SIAM News **40**, Number 6, July/August 2007.

[7] J. Cohen: *Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better*, PLoS Biology **2** (2004) 2017–2023.

[8] D.A. Cox, J. Little, and D. O'Shea: *Ideals, Varieties and Algorithms*, Springer Verlag, New York, third edition, 2007.
[9] F. Jacob and J. Monod: *Genetic regulatory mechanisms in the synthesis of proteins*, J. Molecular Biology **3** (1961) 318–356.
[10] J. Guckenheimer, M Myers and B. Sturmfels: *Computing Hopf bifurcations*, SIAM J. Numerical Analysis **34** (1997) 1–21.
[11] H. Kitano: *Systems biology: a brief overview*, Science **295** (2002) 1662–1664.
[12] J. Koolman and K.-H. Roehm: *Color Atlas of Biochemistry*, Thieme, Stuttgart, New York, 1996.
[13] R. Laubenbacher and B. Stigler: *A computational algebra approach to the reverse engineering of gene regulatory networks*, Journal of Theoretical Biology, **229** (2004) 523–537.
[14] Y. Lazebnik: *Can a biologist fix a radio? – or what I learned while studying apoptosis*, Cancer Cell **2** (2002) 179–182.
[15] H. Lodish, A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell: *Molecular Cell Biology*, W.H. Freeman and Company, New York, 2000.
[16] A. Martins, P. Vera-Licona, and R. Laubenbacher: *Model your genes the mathematical way – a mathematical biology workshop for secondary school teachers*, Course materials available at `http://polymath.vbi.vt.edu/mathbio2006`.
[17] L. Pachter and B. Sturmfels: *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.
[18] B. Stigler and A. Veliz-Cuba: *A boolean model of the lac operon*, in preparation.
[19] B. Sturmfels: *What is a Gröbner basis?*, Notices of the American Mathematical Society, **52** (2005) 1199–1200.
[20] J.M.G. Vilar, C.C. Guet, and S. Leibler: *Modeling network dynamics: the* lac *operon, a case study*, The Journal of Cell Biology **161** (2003) 471–476.
[21] H. Westerhoff and B. Palsson: *The evolution of molecular biology into systems biology*, Nature Biotech **22** (2004) 1249–1252.
[22] N. Yildirim and M. Mackey: *Feedback regulation in the lactose operon: A mathematical modeling study and comparison with experimental data*, Biophysical Journal **84** (2003) 2831–2851.

Virginia Bioinformatics Institute and Mathematics Department, Virginia Polytechnic Institute and State University
E-mail address, Reinhard Laubenbacher: `reinhard@vbi.vt.edu`

Mathematics Department, University of California, Berkeley
E-mail address, Bernd Sturmfels: `bernd@math.berkeley.edu`