Symbolic Calculation in the Life Sciences — Some Trends and Prospects

Michael P. BARNETT
Meadow Lakes, Hightstown, NJ 08520
michaelb@princeton.edu

Abstract

I discuss the literature and benefits of symbolic calculation in the life sciences; some ways to develop the field; and a way to model sequential processes.

Introduction

I went to college in 1945, grew up professionally in an age of scientific optimism, and see symbolic calculation in biology as part of the scientific infrastructure that supports the development of products and services to increase and extend the quality of life. The use of symbolic calculation in teaching, research and the practice of the life sciences will have a profound impact on

1. the inter-related issues of public health and environmental protection,
2. the clinical practice of prophylaxis, diagnosis and therapy throughout "people–care" and the protection and treatment of animal and plant life,
3. agricultural, industrial and commercial operations that involve biological materials and generate products which impact biological systems.

Successive sections of this paper discuss

1. the present literature of symbolic calculation in the life sciences,
2. the benefits of symbolic calculation,
3. interdisciplinary communication,
4. curricular issues,
5. a way to model sequential processes with some possible life science uses.

1. Published accounts of life science applications

Well over 300 research papers and several books that report symbolic calculation in biological research have been published already. These manipulate mathematical formulas and perform other non-numerical processes algorithmically. I explain the background of the applications, summarize the topics and gives detailed citations to a large part of this material in the survey [5] which can be accessed online at

`http://www.princeton.edu/~allengrp/ms/annobib.cals`.

The following list of topics maintains the structure and order of presentation in the survey which should be consulted for more details on items that are given without references. Material that postdates the survey is included below with explicit mention of authors and references. Most of this further material either

1. is summarized in a second survey of symbolic calculation in chemistry and related fields [13] or
2. is part of a thematic issue of the International Journal of Quantum Chemistry on Mathematical Methods and Symbolic Calculation in Chemistry and Chemical Biology that I put together [9]. This is in press.

I organized the material in [5] as follows.

1. **Clinical work** on cancer, the cardiovascular system, drug design and testing, EEG scans, food poisoning, genetic counseling, the immune system, medical decision making, medical imaging, mental health, prosthetic devices, and trauma care. This work includes the control of equipment and the analysis of data in diagnosis and treatment. Biostatistics and other probabilistic methods are used.
2. **Public health issues** of risk analysis, survival analysis, drug testing, epidemiology of infectious diseases and other ailments. Statistical methods, Markov chains, sensitivity analysis and quantifier elimination are used.
3. **Bioengineering** of prosthetic robotics, computer vision and ergonomic design. The Denavit-Hartenburg formulation, differential geometry, algebraic invariance, topology, group theory and probability are used.
4. **Population dynamics** that deals with the time variation of abundance and spatial distribution of populations of people, animals and plants. Models cover invasion, diffusion, dispersal, natural and controlled reproduction, crop planting, migration, and death. Differential, difference and integro-difference equations are used. Many of the models are dynamical systems and some involve biological oscillators and waves. Predator-prey models are formalized as Lotka-Volterra systems. Matrix models are used extensively.
5. **Population and evolutionary genetics** that are based, respectively, on classical theories of heredity and modern theories of genomics. Work includes the breeding of plants and animals, some forms of genetic counseling, maximum likelihood estimates of gametic disequilibrium, choice of markers to determine probabilities of genetic diversity, mutation rates, cross pollination, phylogenetic distance and rates of evolution. Statistical methods, systematic enumeration, Markov chains, multivariate polynomial equations, and discrete and computational geometry are used.
6. **Physiology** and related work on the behavior of cellular substructures and complete cells and organs that are determined by the laws of physics and chemistry. Kuchel and Mulquinney use a powerful combination of methods in an ongoing study of erythrocytes [43]. Work of other authors includes studies of transport and diffusion across living cell membranes, electroencephalography, the growth of cells in the geniculate nucleus, and models of carcinogenesis. Compartmental analysis and identifiability in metabolic networks lead to heavy Gröbner basis calculations [2] *et ante*. Much of the experimental work uses nuclear magnetic resonance that is discussed in [5] and a little more fully in [13]. It includes solid state characterization of membrane proteins by Nielsen *et al* [15].
7. **Biochemical kinetics.** This deals with the rates of metabolic processes. Yildirim deals with pseudo-steady state enzyme reactions [62] *et ante*. Cobelli's group use compartmental analysis and identifiability methods that are computationally intense [2] *et ante*. Both lines of work use Gröbner bases. Fraser uses slow manifold theory [24]. Tóth and Rospars discuss the time

evolution of the differential equations of signal transduction and ion signaling [55]. Ratkiewicz and Truong use a graph theoretic approach to enumerate complex reaction networks [50].

8. **Molecular structure — photophysics.** Boens and Ameloot apply compartmental analysis to the energy transfer kinetics of fluorescence spectroscopy for biological assays [16].
9. **Molecular structure — stereochemistry.** Calculations that are based on the linkage representation of organic molecules [29] include the recent work of Coutsias, Seok, Wester and Dill on tripeptide loop closure [19], Emiris, Fritzilas and Manocha on pharmacophore behavior [22], Barnett and Capitani on puckering in small ring molecules [11], Lewis and Bridgett on docking [38], and the textbook study of cyclohexane [28]. The geometrical considerations usually lead to multivariate polynomial equations that are solved using Gröbner bases and resultants.
10. **Molecular structure and dynamics — other methods.** Symbolic calculatons in **nuclear magnetic resonance**, **laser experiments** and **crystallography** are discussed without specific reference to life sciences in [13]. The discussion of solitons in the same survey points to biological examples and to the extensive symbolic calculations of Gao and Tian. More recently, Güntert described his MATHEMATICA package for spin operator analysis of pulsed NMR work on biological materials [31]. Eick and Souvignier [21], Fritzsche [25], Stokes and Hatch [52] and Vail, Aris and Scarlete [56] report their work on crystallography and group theory.
11. **General statistical methods** provide the infrastructure for several topics mentioned above. The recent books by Andrews and Stafford [1], by Rose and Smith [51] with its supporting `mathStatica` software package, and by Pistone, Riccomagno and Wynn [48] are important resources for teaching and research. Many more papers on statistical work of biological interest that use symbolic calculation are described in [5]

The most popular software platforms in this work have been MAPLE and MATHEMATICA. Other systems that have been used in life science research include AXIOM (formerly called SCRATCHPAD), MACSYMA, REDUCE, and the specialized polynomial system SINGULAR [30]. Very intense symbolic calculations are supported by FORM and SACLIB, but I have not found applications of these to biology. Ratkiewicz stresses the power of FORTRAN 95 in [50] mentioned above. The trade-off between special purpose packages that are

1. based on general purpose computer algebra platforms *e.g.* the Canon tensor package [41] and
2. written independently *e.g.* the tensor contraction engine [47]

remains an open question (see *e.g.* [17]). Interfacing different packages provides the benefits of both. The special purpose computer for biological sequencing [53] widens the issue. Recently, Capitani and I wrote a package `igm.m` [12] to postprocess GAUSSIAN 03 calculations of electronic structure in MATHEMATICA. The parser in `igm.m` can handle the output of many other packages, too. Makarov and Meliu have built a genetic programming style on MATHEMATICA [40]. The numerical system MATLAB has an add-on that accesses MAPLE. The MAPLE worksheet and MATHEMATICA notebook resources are used to teach a host of topics. Information on how this activity can stimulate research applications would be useful. Several major non-numerical applications of computers can be handled by systems such as MATHEMATICA and it is timely to bring these within the focus of the symbolic calculation community. These include isomer enumeration, chemical structure searching, computer aided chemical synthesis and combinatorial chemistry, and gene sequencing.

2. Benefits of symbolic calculation

Symbolic calculation software performs analytic differentiation and integration, takes limits, constructs power series expansions, solves some algebraic equations analytically, and simplifies mathematical expressions. The major systems have elaborate graphics capabilities and support file and character string operations. The software also has an extensive mathematical infrastructure [28]. These features provide many benefits.

1. Increased **dependability of mathematical formulas** through:
   (a) **reduced errors** of transcription and manipulation, though the possibility of programming errors must be recognized,
   (b) new ways to **check** the results, which is essential, though seldom mentioned,
   (c) the clarity of **simple** mathematical methods that were precluded by tedium or digital erosion in the past.
2. Increased **dependability** and **precision** of **numerical results** through:
   (a) unrestricted-precision integer, rational, real and complex arithmetic that overcome the loss of accuracy which made many formulas useless for numerical work in the past,
   (b) mutual annihilation of expressions of opposite sign and identical magnitude, that were added in piecewise numerical computations,
   (c) retention of radicals and other irrationals symbolically to late stages of a calculation, and use of series expansions when an expression such as $ae^{px} - be^{qx}$ loses precision when computed directly.
   (d) interval arithmetic that is unrestrictedly precise,
   (e) ability to bypass error bound analyses that are weak, obscure, controversial, wrong or omitted.

   I distinguish
   (a) **dependability** *i.e.* reliability of all the digits in a result, from
   (b) **precision** *i.e.* the actual number of digits.
3. **Improved models** that are based on **enriched** mathematical and algorithmic **repertoires**. These include:
   (a) the **simple mathematics** as just been mentioned, and
   (b) more **eclectic methods** that are not practical without the support of symbolic calculation,
   (c) manipulation of **operators** and **high level mathematical objects** such as actual derivations and proofs in n-th order expressions.
4. **Improved numerical methods**. These include
   (a) powerful convergence accelerators requiring high precision, and
   (b) new numerical analysis algorithms
      i. that are specialized *e.g.* to particular forms of integrand in quadrature formulas [26] and particular grids for PDEs,

ii. that go to higher dimensionality in interpolation,
iii. that go to higher orders in numerical methods for differential equations *e.g.* [57].

5. **Mechanized optimization**. Long sets of alternative computational schemes that are based on a single set of formulas can be enumerated, calibrated and selected mechanically for optimal performance [7].
6. **Collegial computing** in which
   (a) formulas are developed and validated at multiple sites,
   (b) distributed as machine readable input for further computation that
   (c) produce results in print for visual inspection and hard copy publication, including step-by-step display of intermediate results.

Classifying the uses and benefits helps

1. to pinpoint new areas of application and to stimulate work on these,
2. to analyze software needs that software developers need to know,

Looked at another way, symbolic calculation supports

1. The construction of **very long end results** with demands on permanent storage and transmission and a need for abbreviated display.
2. The construction of **very long tables of formulas**, as in the 40,000 hyperspherical harmonics computed by Wang and Kuppermann. These have the same needs as item 1. Using long tables of formulas raises further practical questions — see `DEMO_NOTES.dta` and `DEMOS.mma` in `http://www.princeton.edu/~allengrp/ms/overlap/`.
3. Calculations that involve **very long intermediate results** and short end results, *e.g.* when subexpressions of opposite sign and identically equal magnitude annihilate each other. Short term storage is critical.
4. The construction of **very long proofs** and **derivations** that, at times, involve very large numbers of special cases, and adapting these in mechanized reasoning by analogy.
5. **Combinatorial** and other processes that involve **lengthy iterations** and **recursions**, where computational intensity can limit the models that are studied, and parallelization is important.
6. Producing long sets of **coursework problems** and **worked examples**.
7. Producing long sets of **diagrams** and on-demand-production of individual diagrams that depict *e.g.* phylogenetic trees, spin coupling NMR, metabolic and other reaction pathways, stereochemistry, electron distribution, isopotential surfaces and other properties of molecules and molecular fragments and aggregates — see [5, 13, 18].

Another categorization which bears strongly on the modeling ideas in §5 distinguishes calculations that, respectively,

1. produce numerical results or diagrams,
2. produce formulas that give numerical results or diagrams when numbers replace the constituent parameters,

3. produce general formulas that give specific formulas of the kind just mentioned when numbers replace the constituent parameters,

and so on to increasing levels of abstraction. Also, "algorithm" can replace "formula" in items 2 and 3.

Symbolic calculations use mathematical methods from several disciplines, with varied terminology and approaches. In particular

1. biometrics — dominated by statistics, probability and stochastic theory,
2. computational physics and chemistry — dominated by differential equations and linear algebra,
3. control systems and signal processing — topics similar to those in item 2 often couched in different terminology needing different prior knowledge,
4. structural mechanics and fluidics — dominated by linear algebra and differential equations.

Also, symbolic calculation has stimulated work on

1. Gröbner bases and resultants, matrix and other algebraic methods discussed in [28]. Applications are discussed in [42] and [65].
2. Quantifier elimination, differential and Clifford algebras, discrete and computational geometry. Applications are discussed in [5].
3. Lie algebras in perturbation theory [23], Lie symmetry methods for differential equations [35], homotopy theory [32], Painlevé test methods [59], slow manifold theory [24], nonlinear evolution equations [39], soliton theory *e.g.* [54], and stability analysis [3], the many papers on methods for solving the time dependent Schrödinger equation and other nonlinear PDEs mentioned in [13] and discussed more recently *e.g* by [60], and
4. group theory [21, 25, 52].

3. Improving interdisciplinary communication

Interdisciplinary teaching and research is a major issue in higher education today. New organizational structures and multi-million dollar institutes are appearing on campuses, particularly in the United States. Teaching and research that uses symbolic calculation in the life science can be pursued much less dramatically in ways discussed below. But several communication issues remain. In principle, applications experts can present their work in mathematical terms for mathematicians and computer scientists to solve. It has been a truism since the start of electronic computing that whilst this approach has occasional success, the computational breakthroughs occur when the applications and computer experts learn enough about the opportunities and needs of each other to establish priorities that home in on tractable significant issues, avoiding unnecessary demands and simplifications.

Responsible popular science explanations help considerably, but there is no substitute for collegial enthusiasm. Ironically, the barriers of terminology, formalism and ideas that are taken for granted between different specializations and schools that form the biological sciences, and even within some of these, compound the communication problem. And similar compartmentalization seems to exist within mathematics and computer science. The sheer vastness of activity and literature makes these problems almost inevitable.

Cultural differences A deeper problem, discussed at length in a major report on interdisciplinary cooperation [46] lies in cultural differences between (1) the natural sciences and (2) mathematics and computer science. The extensive use of symbolic calculation by natural scientists leads inevitably to new methodologies as well as to new scientific results. This information has to be exchanged. Fortunately, every major journal run by natural scientists who use mathematics has a tradition of publishing scripts in FORTRAN. With the advent of symbolic calculation, scripts in MAPLE, MATHEMATICA and other languages have been accepted without demur. For some computer science journals, however, this is the kiss of death. These journals have also adopted the preoccupation with algorithms and theorems that the National Research Council report mentions [46]. Symbolic calculation involves a lot more. Exclusivity and intolerance have to be set aside by journals that wish to mediate the symbiotic development of their primary subjects with fields that interact with these.

Language issues A more tractable problem that can wreak havoc until recognized concerns the baggage that the same term carries in the minds of people with different backgrounds. At ISSAC03, I showed an audience of computer scientists what "ring", "field" and "homology" mean to chemists by flourishing a doughnut to suggest a benzene molecule, some refrigerator magnets that produced magnetic fields, and some Leggo towers to demonstrate homology — I really wanted sausage strings of different lengths. A recent paper with Minamair [42] explains Gröbner bases without mentioning the word "ideal". This word stops many chemists dead in their tracks. Mathematicians and computer scientists must recognize the barrier that a set of definitions creates, no matter how familiar they are with the underlying ideas.

Also, natural and computer scientists must be sure they are talking about the same thing when they think they are. My paper on parameterized eigenvalues [14] got off the ground after a one-hour conversational mismatch when I used linear dependence for the plot of y *versus* x looking reasonably close to a straight line and Thomas Decker thought I meant y identically equal to kx.

Starting a dialogue can be the hardest part of the development of joint work. I have found it useful to concoct some tenuous suggestions for the application of a software resource to a field outside my expertize, and then to ask an expert for comment. This usually evokes — "that is not a real problem" or "that would not work" followed by "but ... would be useful". Presenting an abstract "solution in search of a problem" usually evokes "cannot think of anything".

4. Curricular measures

Many higher education systems allow "special projects" courses in which each student, or a team of two or three students, undertakes an individual research exercize. Many systems let one department offer a course that is attended by students from other departments. Within this framework, a natural science student can team up with a mathematics / computer science student to explore the application of symbolic calculation to a problem in the life sciences. Scenarios in several countries are described in [9].

Problems can be suggested by faculty from their immediate interests or by reference to books that include the mathematical biology texts by Edelstein-Keshet [20], Hoppensteadt and Peskin [34], Keener and Sneyd [36], and Murray [44], none of which mention symbolic calculation, and by Yeargers, Shonkwiler and Herod [61] which uses MATHEMATICA pivotally. Edelstein-Keshet discusses discrete and continuous processes: population growth, continuous models in population dynamics, molecular events, limit cycles, oscillations and excitability; spatially distributed processes: population dispersal, microbial motion, traveling waves, transport, pattern formation and morphogenesis; with the primary

mathematical focus on ordinary and partial difference and differential equations. Hoppensteadt and Peskin discuss, successively, heart and lungs, cell behaviour, renal mechanisms, muscle mechanics, neural systems, microbial population dynamics, genetics, epidemiology, and population growth and dispersal. Keener and Snyd progress through discussions of cell physiology: biochemical reactions, cellular homeostasis, membrane ion channels, excitability, calcium dynamics, bursting electrical activity, intercellular communication, passive electric flow in neurons, wave and cardiac propagation, calcium waves and cell regulation; system physiology: cardiac rhythmicity, circulation, blood and respiration, muscle, hormones, kidneys, gastrointestinal system, seeing and hearing. Murray deals exhaustively with differential equations and their biological effects.

The reissue [20] contains a superb list of more than 50 further books that apply mathematics to biology and which focus individually on bioeconomics of resource management, branching processes, cell regulation and evolution, dynamic state models, dynamical systems, ecological complexity, fluidics, genetic and biochemical networks, immunology and virology, oscillations and cellular rhythms, pattern formation, physiological kinetics, population statistics and epidemiology, probability, scaling, and stochastic processes. Some of the books are aimed at mathematics students and some at biologists.

Mathematicians and computer scientists who think they have found a potential application should run a literature search using a catch phrase that identifies the topic in conjunction with "Maple" or "Mathematica" or "symbolic calculation" or "symbolic computation" or "computer algebra" as the key. And above all, they should make sure that an applications expert is at hand or accessible on the web to ensure perspective on what may be achieved.

The papers by Hereman [32] and Hydon [35] illustrate another approach. These deal with homotopy theory and Lie symmetry methods in the solution of differential equations. Both authors gave me a list of named equations that their methods address. I used these names as keys in Chemical Abstracts online searches. The results are summarized in endnotes to [32] and [35]. I got over 1000 hits. Most relate to plasma physics and related topics. But a substantial number relate to topics in chemistry and biology. These include applications of the Korteweg–de Vries (KdV) equation to chemical ecology, food sterilization, porous media, colloid chemistry, surface viscosity, mechanism of adherence, and the physiology of biomechanical behaviour; the Boussinesq equation to ultra short pulsing, and the chemical biology of the neuron; soliton equations for the behaviour of proteins and muscles, surface catalysis, liquid films, Raman processes in molecular liquids, carcinogenic radiation, coherent biosystem excitation, hydrogen bonded molecular chain solids, and polyelectrolyte light scattering; the Thomas equation to ion-exchange processes in purification and separation; the Burgers equation to plasticity, fluid interfaces, and cardiac behavior; the Maxwell-Vlasov equation to cluster growth, and free-electron and ultra-intense lasers; the Fitzhugh-Nagumo equation to physiological excitable systems including cardiac behavior, neurophysiological modeling, cell membranes, chemical and biological oscillation, and amoeba aggregation; and the isotachophoresis equation which is the theoretical basis of electrophoresis and chromatography. I have done a similar search that led from integer programming to a substantial number of papers in the life sciences.

I think that new life science applications of symbolic calculation will be found by literature searches that start from mathematical methods which have been facilitated by symbolic calculation in any field. Also, getting students (and faculty and industrial practitioners) of mathematics and computer science accustomed to the Chemical Abstracts and Science Citation Index and other search tools of the natural sciences, with the help of subject specialist librarians, as well as to internet search engines will be invaluable in the development of strong interdisciplinary work.

Two technical points. To facilitate interdisciplinary work, get "symbolic calculation" and related terms into the thesauri of indexing and abstracting services and into the controlled keyword lists of journals. And try to establish a mechanism to get information about papers on life science applications that computer journals publish into MEDLARS and other life science information resources which do not cover the journals directly.

5. Modeling

5.1. Background

I have used MATHEMATICA to construct thousands of formulas for work on computational chemistry over the past 15 years. Most of these symbolic calculations involved special functions of mathematical physics, specialized integration, limits, recurrence schemes and matrix operations. Many of the formulas comprise machine readable tables. Recent papers that describe this work include [7, 11]. These refer to earlier work that is also listed on the website [10]. I use a mnemonic programming style and a collection of scripts that I call MATHSCAPE. The basic process in the majority of MATHSCAPE sessions is, mathematically, the conversion of X to Y by a postfix operator P, where X and Y denote a pair of equations or identities or algebraic or geometric or logical statements.

$$Y = X \bullet P \tag{1}$$

Additionally, several of my recent calculations involve processes of generalization, reasoning by analogy and optimization that are, mathematically,

$$S = R \bullet \mathcal{F} \tag{2}$$

where R and S denote a pair of operators with the status of P in Eq. (1) and $\mathcal{F}$ is a second order operator. A few calculations have involved corresponding third order operators.

Güntert's paper [31] in the forthcoming thematic issue of the International Journal of Quantum Chemistry uses MATHEMATICA constructions that are very similar to the MATHSCAPE style. These simulate the cumulative effect of some physical processes that occur sequentially in real time. I found the similarity startling, and it led me to a growing realization that similar constructions could model a host of other sequential processes. The invitation to contribute to AB2005 gave this idea impetus. Hence §§5.5.–5.7., that I present in the "kite-flying" spirit described at the end of §3..

5.2. Composition of unary operators

To illustrate the basic ideas of MATHSCAPE without recourse to specialized applications, consider Pólya's discussion of "conjecture E" — Euler's expansion of the sine as an infinite product [49].

$$\sin x = x \lim_{N\to\infty} \prod_{n=1}^{N} \left(1 - \frac{x}{n\pi}\right).\left(1 + \frac{x}{n\pi}\right) \tag{3}$$

I represent this by

```
eqn[conjectureE]  =
  sin[x] ==
     x lim[N->infinity][
      prod[n,1,N][(1 - x/(n pi))(1 + x/(n pi))]]
```
(4)

Curry-style expressions represent the limit and the indexed product. The evaluation of the infinite

product in Eq. (5) is set as exercise 6.23 in [49]. The proof of this result illustrates many features of MATHSCAPE without the need for specialized knowledge of computational chemistry.

$$\left(1-\frac{1}{4}\right)\left(1-\frac{1}{9}\right)\left(1-\frac{1}{16}\right)\ldots = \frac{1}{6} \tag{5}$$

The following MATHSCAPE script converts Eq. (3) to Eq. (5) .

```
eqn[polya[6.23]] =
  eqn[conjectureE]  //
    pipe[
      toTheProduct[leftExpand],
      toTheLimit[moveCoefficientLeft],
      toTheProduct[reindex[n, n-1]],
      transposeFromRhsFactorsFreeOf[lim],
      propagateLimit[x->pi],
      toTheProductand[Apart],
      canonicalizeTheLimit]
```
(6)

The postfix expression x `// pipe[`$f_1, \ldots, f_n$`]` represents the composition of unary functions that is nested as f_n`[...[`f_1`[`x`]]...]`. In my early experiments with MATHEMATICA, I found myself using postfix compositions increasingly, in several *ad hoc* ways. When I found how convenient these are to build and to document, and the reasons for this, I systematized the style. This evolution highlights the extent to which symbolic calculation is an experimental activity that leads to insights about programming languages which could not be found by abstract speculation.

The constructions `toTheProduct`, `toTheLimit`, `toTheProductand` are "targeting" functions that focus action on components of a nested mathematical structure. MATHSCAPE supports a large, open-ended class of these. Their usage is explained in detail in [8]. The following `pipe` expression illustrates two further features of MATHSCAPE targeting. It refers to a hypothetical identity between two algebraic expressions in x, y, z.

```
pipe[
  toTheLhs[
    toTermContaining[x^2]][
      toFactorFreeOf[x]][Expand]]],
  toTheRhs[
     toTheNumerator[
       collectPowersOf[y], factorOut[z]],
     toTheDenominator[
       collectivelyToTermsContaining[z]][Factor]]]]
```
(7)

This shows the nesting of targeting functions and the usage that allows argument lists of unary functions that are composed. Not surprisingly, nested operators are powerful tools for manipulating nested expressions. And, as a consequence, nested operators are amenable to manipulation by higher level nested operators.

5.3. Analogy

Exercise 6.24 in [49] is the evaluation of

$$\left(1-\frac{4}{9}\right)\left(1-\frac{4}{16}\right)\left(1-\frac{4}{25}\right)\ldots \tag{8}$$

The operator that reduced Eq. (5) is retrieved by

```
op[6.23] = betweqn[conjecture E, polya[6.23]]
```

(9)

Inspection suggests

1. moving the two low order elements out of the scope of the $\prod$ operator, instead of just one,
2. reducting the lower limit of n correspondingly by 2 instead of 1, and
3. replacing x by 2π instead of π,

The new composition expression is formed and applied, accordingly, by

```
op[6.24] =
  op[6.23] //
     pipe[leftExpand->leftExpand^2, n-1->n-2, pi->2pi]

eqn[polya[6.24]] = eqn[conjectureE]  // op[6.24]|
```

(10)

This reduces the product Eq. (8) to 1/6. By analogy, the next formula in the style of Eqs. (5) and (8) should be formed by changing 2 to 3 in the composition that was just used. I tried

```
eqn[conjectureE] // (op[6.24] /. {2->3, -2->-3})
```

(11)

It produced

$$\frac{1}{20} = \lim_{N\to\infty} \prod_{n=1}^{N} \left(1 - \frac{9}{(3+n)^2}\right). \tag{12}$$

5.4. Optimization

The combination of Eqs. (6, 9, 10) corresponds to the expression of analogy by the scheme

$$Y_1 = X \bullet \mathcal{F}_1, \quad \mathcal{F}_2 = \mathcal{F}_1 \bullet \mathcal{G}, \quad Y_2 = X \bullet \mathcal{F}_2 \tag{13}$$

Optimization is expressed by

$$Y = X \bullet \mathcal{F}_1, \quad \mathcal{F}_2 = \mathcal{F}_1 \bullet \mathcal{G}, \quad Y = X \bullet \mathcal{F}_2 \tag{14}$$

where the conversion of X to Y using $\mathcal{F}_2$ is faster or causes less swelling than the conversion using $\mathcal{F}_1$. This covers many practical situations with different characteristics. Replacing the builtin `Simplify` function by a simplification function that addressed special features of the target speeded one calculation by an order of magnitude [6] .

The calculation of a table of 10,000 "molecular integrals" described in [7] involves a much more complicated situation that is endemic to numerical and symbolic scientific computing. The end products in [7] — for present purposes call them $A_{n_1,\ldots,n_6}(x,y)$ — depend on 6 integer parameters and 2 real variables. Each A function is computed from a combination of B and C functions that depend on several integer parameters, with ranges that are related to those that characterize the As. And so on down through 8 levels, described by a set of 17 formulas. Some of these involve recurrence.

When a set of function $C_{m_1,\ldots}$ is given by an expression containing functions $D_{\ell_1,\ldots}$ and $E_{k_1,\ldots}$, alternative computational schemes include calculating

1. all the C functions first, followed by all the Ds,
2. each C function and each D function whenever it is needed,

3. subsuming the equation for the Ds, in terms of still lower level functions, into the equation for the Cs,

and obvious variations. Enumerating the alternatives and assessing or measuring their relative merits is the key to an important class of optimization procedures. In [7] I discuss an approach to this problem that expresses an overall computation scheme based on the given set of equations as a nested `pipe` expression that acts on the list of sets of indexes $m_1, \ldots$ that characterize the set of C functions that are needed. The constituent unary functions specify the equations and the ranges of indexes to be used in successive steps. Altering the computing scheme corresponds to coalescing pairs of parent sibling nodes or transposing sibling nodes. This alteration is covered by Eq. (14). The steps that are needed in a full scale enumeration–assessment are discussed in [7].

5.5. NMR and MRI

Güntert has written a MATHEMATICA package POMA that simulates an important class of pulse sequence NMR experiments [31]. The package converts a representation of the sequence of radiofrequency pulses and other actions that comprise an experiment into a Fourier series for the time evolution (dependence) of the "density operator". The coefficients contain certain molecular parameters that depend on the interactions of the nuclear spins of constituent atoms. NMR experiments are used to identify substructures and to measure molecular dynamics in a vast range of biological studies. There are many texts on the subject *e.g.* [37]. The online introduction [64] includes helpful animations. Individual concepts are explained accessibly in [33].

The following MATHEMATICA expression for the 3QF-COSY calculation, in the POMA documentation shows the style of input.

```
spin[1,z] // pulse[ 90,{0,Pi/3,2Pi/3,Pi,4Pi/3,5Pi/3}] //
 delay[t,{{1,2},{1,3}}] // pulse[90,{0,Pi/3,2Pi/3,Pi,4Pi/3,5Pi/3}] //
 pulse[ 90,{x, x, x, x, x, x}] // receiver[{x, -x, x, -x, x, -x}] //
 observable
```

The successive `pulse` and other formal operators correspond to physical actions and happenings. The physical experiment can be described, accordingly, as

```
experiment[3QF-COSY] =
 pulseSequence[
  pulse[90, {0,Pi/3,2 Pi/3,Pi,4 Pi/3,5 Pi/3}], delay[t, {{1,2},{1,3}}],
            ..., observable]
```

where `pulseSequence` is a placeholder, and `pulse[ ...]`, ..., `observable` describe physical happenings. Then the 3QF-COSY calculation is performed by

```
spin[1, z] // (experiment[3QF-COSY] /. pulseSequence -> pipe)
```

The presence of MATHEMATICA procedures with pattern names that match the subexpressions `pulse[...]`, ..., `observable` causes automatic evaluation.

Going from `pulseSequence` to `pipe` makes only a minute change in the appearance of the expression and it requires only minute word processing capabilities. But the impact is dramatic — it moves from the representation of a time sequenced multi–step process in the physical world to an operator whose action is to simulate the cumulative effects. Formally, let D be a description of a member of some class of physical process, let $\mathcal{D}$ denote the operator that simulates D, and let $\mathcal{T}$

denote the mapping of the Ds to the corresponding $\mathcal{D}$. Then the general transformation from the physical world to the operator world is encapsulated by

$$\mathcal{D} = D \bullet \mathcal{T} \tag{15}$$

Of course modeling programs have always converted input representations of physical reality to data for table driven software or to sequences of statements that call appropriate subroutines. Mnemonic descriptions of physical objects, scenarios and engineering systems are accepted by the modeling languages that are described in texts on computer graphics and by languages such as MODELICA [66, 45]. In object-oriented programming "software objects are often used to model real-world objects you find in everyday life" [63]. The MATHSCAPE tactics which have just been described, however, open up innumerable opportunities because of

1. the ease of handling tree structures, that describe the physical system, in MATHEMATICA,
2. targeting capabilities to pinpoint where the system is to be changed,
3. pattern matching that enhances the targeting actions,
4. the ability to use overlapping repertoires of commands to operate on
   (a) representations of physical objects and time-sequenced processes,
   (b) the operators that act on these,
   (c) the `pipe` expressions that do the modeling,
   (d) the operators that act on these,

with further opportunities to work by analogy and to optimize the physical systems using the methods discussed in earlier subsections.

Güntert's work raises the following questions in the specific area of NMR.

1. Can algorithms be devised to convert the `pulseSequence` description of an experiment into computer controls to drive the experiment?
2. Can sets of operators be devised to support the convenient interactive change of the `pulseSequence` description of one experiment into the corresponding description of an experiment with related objectives, by reference to the changes in the target material or conditions or objective?
3. Can algorithms be devised to do this?
4. Can algorithms be devised to construct `pulseSequence` descriptions of experiments directly from the chemical nature of the study for significant classes of investigation, that are more elaborate than tables of numerical parameters?

In private communications, Istvan Pelczer and Malcolm Levitt have suggested material that I should review for possible answers to some of these questions, which I will do. The `pulseSequence` displayed above contains items with relatively little internal structure. The conventions can be extended to include gradients, pulse shapes and further details.

It is a very short leap of thought from modeling NMR to modeling MRI and on to other medical imaging equipment. These are controlled by computer software. Medical staff provide numerical data for protocols that they select. I have started a dialogue with a clinical radiologist about the possible application of the modeling methods in this paper to medical imaging and I hope to pursue this in coming months, too.

5.6. Organic chemistry

Organic chemistry provides fertile ground for modeling based on Eq. (15) that bears on pharmaceuticals and biochemistry. Representations are the key. I use nested MATHEMATICA expressions, *e.g.* ethyl chloride $CH_3.CH_2.Cl$ is represented by `ch[C[H, H, H], sb, C[H, H, Cl]]` where `ch` and `sb` denotes a chain and a single bond, respectively, and `X[A, B,...]` denotes the group consisting of `X` attached to atoms or subsidiary groups `A`, `B`, ... The notations are under development. They are meant to show the kinds of tactics that are possible and the ways these are supported, to encourage other workers to develop notations that meet their individual needs rather than adopting the precise details of my scripts. Representations are non-unique and can be restructured mechanically.

Properties of atoms and bonds are included in extra brackets. The meaning can be implicit, as in `C[13][H, H, H]` for a methyl group containing ^{13}C. Alternatively, the meaning can be made explicit, as in `sb[bondLength[145 pm], bondOrder[1.2]]`. The MATHEMATICA syntax allows a host of other labeling tactics that include `==` and `->` as connectives, as in `sb[dipole -> 1.3]`, but `=` should not be used in this context. The item `bondTo[n]`in a `ch` expression connotes bonding to the non-adjacent atom at position n in the chain, with further conventions for bonding to out of chain atoms, for double and Kekulé bonds and for fused rings and three-dimensional structures. The chemical mark up language ChemML may suggest notations for particular topics [67].

The representation of chemical reactions is illustrated by the dehydration

$$CH_3.CH_2OH \longrightarrow CH_2 : CH_2 \tag{16}$$

This is represented by the notationally similar

```
dehydration[1] =
 ch[C[H, H, H], sb, C[H, H, OH]] -> ch[C[H, H], db, C[H, H]]
```

The representation of a reaction can be modified by use of targeting functions. For example,

```
dehydration[2] = dehydration[1] // toBothSides[
   toAtom[C][1][toSubstituent[1][H -> R1], toSubstituent[2][H -> R2]],
   toAtom[C][2][toSubstituent[1][H -> R3], toSubstituent[2][H -> R4]]]
```

assigns the name `dehydration[2]` to the more general reaction

```
ch[C[R1, R2, H], sb, C[R3, R4, OH]] -> ch[C[R1, R2], db, C[R3, R4]]
```

Elementary pattern transformation applies these representations to individual compounds. MATHSCAPE treats `patternRule[dehydration[2]]` automatically as

```
ch[C[R1_, R2_, H], sb, C[R3_, R4_, OH]] :> ch[C[R1, R2], db, C[R3, R4]]
```

with the standard meaning of the underscore [58]. Hence, for example, the statement

```
ch[C[H, H, H], sb, C[CH3, OH], sb,
     C[CH3, OH], sb, C[H, H, H]] //. patternRule[dehydration[ ]]
```

converts the representation of pinacol to that of 2,3-dimethylbutadiene-1,3.

Chemistry texts often extend Eq. (16) to show the dehydrating agent by

$$CH_3.CH_2OH \xrightarrow{Al_2O_3} CH_2 : CH_2 \tag{17}$$

I represent this kind of extension by, *e.g.*

```
dehydration[1, ] =
 {ch[C[H, H, H], sb, C[H, H, OH]],
 {Al2O3 || kaolin}, tempRange[350,360]}
 {ch[C[H, H], db, C[H, H]]}
```

In general, {{*reactants*}, {*further details*}, {*products*}} allows multiple reactants, conditions and products. Usually, the reactants are generic in the sense that `dehydration[2]` contains R1, R2, R3, R4. The *further details* can have an elaborate conditional structure to meet the needs of different reactants. The `patternRule` formed from a triple of this kind is an operator that acts on a reactant list that matches the pattern formed from *reactants* if *further details* is null or `True` as the result of earlier assignments. The *further details* are omitted when they be taken for granted.

A recent article by Ganesan [27] illustrates heterocyclic combinatorial synthesis by the 3-step production of the thiohydantoin library by (1) alkylating an α-amino acid ester using an aldehyde, (2) treating the product with an isothiocyanate to give a transient thiourea, (3) letting this cyclize automatically. I am using this to test the current conventions and software at the time of writing. The plan is to base the first step on the alkylation of a primary amine $XNH_2 + Y.CHO \rightarrow X.NH.CH_2.Y$ by substituting, in the text processing sense, $R_1.CH_2.CO_2Me$ for X and $CH_2.R_2$ for Y. The second step will be based on the reaction of a secondary amine $NH(X_1)X_2 + CS{:}NX_3 \rightarrow N(X_1)(X_2).CS.NH.X_3$ by substituting R_3 in place of X_3. The pattern rule that is formed for this step should transform the result of the first step to the representation of the thiourea as a tertiary amine

```
N[C[S,N[H,R3]], C[H,H,R2], C[H,R1,C[O],O[Me]]]
```

Then a command to restructure this as a chain beginning with the `O[Me]` and containing the `C[S]` will put this in the form that is cyclized by the pattern rule for the third step.

I assume that my computational experiments so far cannot produce results that match mainstream software for combinatorial synthesis but I think it may lead in new directions. A simple change to the representation of the thiohydantoin synthesis gives the synthesis of thioxodihydropyrimidinones from β-amino acid esters. Formal analogy between different classes of reaction should lead to high order abstractions. The syntax allows further commands that represent separation, purification and experiments to measure properties, and action that is conditionalized on the result of these measurements. This tactic addresses some needs of chemical robotics. Parallel operations on two or more species that are produce during a synthesis can be represented easily. So can the rejoining of these, leading on to the representation of metabolic networks.

5.7. Cell signaling

It is very tempting to speculate on the possible extension of these ideas to cell signaling. Tentatively, I am considering a model that cycles through successive time increments, maintaining data for a list of metabolic processes. The data for each process X includes a list of substances that terminate X immediately and a list of substances that terminate X when a preset tolerance is reached. The level of each of these substances can change from cycle to cycle and the levels are stored, too. The widely used diagrams of cell signaling suggest a style analogous to the `ch` expressions to represent a receptor sequence on a membrane protein. Then the attachment of ligands to receptors, and the further attachment of molecules to the ligands, and interaction between ligands on adjacent receptors would be represented in a style analogous to the reaction sequences discussed above. Some forms of abberant behaviour could be described by changes in the spatial sequence of receptors and the order and time of arrival of signals. A process X that changes a process Y can be regarded as an operator and this idea extends immediately to operators of increasing level. But I would be presumptuous to go further with these speculations now.

I have had a long term interest in biological information processing. But exigencies of time and space preclude more than a passing reference to the account of the PST model in [4]. It depends on string substitutions by molecular assemblages that act as associative memory devices, with latent codes that are activated by the same mechanism. .

6. Acknowledgements

I thank Katsuhisa Horimoto and Hirokazu Anai for inviting me to contribute to AB2005, and L. C. Allen, J. Eng, A. Ganesan, M. Levitt, M. Todd and I. Pelczer for the benefit of helpful comment.

References

[1] D. F. Andrews, J.E. Stafford, *Symbolic computation for statistical inference.* (2001) Oxford University Press, New York.
[2] S. Audoly, G. Bellu, L. D'Angiò, M. P. Saccomani, C. Cobelli, IEEE Trans. Biomed. Eng. 48 (2001) 55.
[3] A. V. Banschikov, L. A. Burlakova, Programming and Computer Software, 23 (1997) 173.
[4] M. P. Barnett, *Molecular Electronic Devices, ed. F. L. Carter, R. E. Siatkowski and H. Wohltjen,* (1988) 229, North Holland, Amsterdam.
[5] M. P. Barnett, SIGSAM Bulletin, 36 (2002) 5.
[6] M. P. Barnett, J. Chem. Inf. Sci. 43 (2003) 1158.
[7] M. P. Barnett, Int. J. Quant. Chem. 95 (2003) 791.
[8] M. P. Barnett, Some simple styles of symbolic calculation. (a) J. Symb. Comp. (in press). (b) http://www.princeton.edu/ allengrp/ms/annobib/scadAll.pdf
[9] M. P. Barnett, Mathematical methods and symbolic calculation in chemistry and chemical biology — the gathering momentum. Int J Quant Chem, (in press).
[10] M. P. Barnett, website www.princeton.edu/ allengrp/ms.
[11] M. P. Barnett, J. F. Capitani, Modular chemical geometry and symbolic calculation. Int. J. Quant. Chem. (in press).
[12] M. P. Barnett, J. F. Capitani, Interfacing GAUSSIAN with MATHEMATICA. (in press).
[13] M. P. Barnett, J. F. Capitani, J. von zur Gathen, J. Gerhard, Int. J. Quant. Chem. 100 (2004) 80.
[14] M. P. Barnett, T. Decker, W. Krandick, J. Chem. Phys. 114 (2001) 10265.
[15] M. Bjerring, T. Vosegaard, A. Malmendal, N. C. Nielsen, Concepts in Magn. Res. A, 18A (2003) 111.
[16] N. Boens, M. Ameloot, Compartmental modeling and identifiability analysis in photo physics: a review. Int J Quant Chem. (in press).
[17] V. A. Brumberg, *Analytical techniques of celestial mechanics.* (1995) Springer, New York.
[18] M. A. Caprio, Comp. Phys. Comm. 171 (2005) 107.
[19] E. A. Coutsias, C. Seok, M. J. Wester, K. A. Dill, Resultants and loop closure. Int J Quant Chem. (in press).
[20] L. Edelstein-Keshet, *Mathematical models in biology.* (2005) SIAM, Philadelphia.
[21] B. Eick, B. Souvignier, Algorithms for crystallographic groups. Int J Quant Chem. (in press).
[22] I. Z. Emiris, E. D. Fritzilas, D. Manocha, Algebraic algorithms for structure determination in biological chemistry. Int J Quant Chem. (in press).

[23] F. M. Fernández, E. A. Castro, *Algebraic methods in quantum chemistry and physics.* (1996) CRC Press, Boca Raton.
[24] S. J. Fraser, Symbolic methods for invariant manifolds in chemical kinetics. Int J Quant Chem. (in press).
[25] S. Fritzsche, Application of point-group symmetries in chemistry and physics: a computer-algebraic approach. Int J Quant Chem. (in press).
[26] H. Fukuda M. Katuya, E.O. Alt and A.V. Matveenko, Comp. Phys. Comm. 167 (2005) 143.
[27] A. Ganesan, Pure App. Chem. 73 (2001) 1033.
[28] J. von zur Gathen, J. Gerhard, *Modern Computer Algebra, 2nd ed.* (2003) Cambridge University Press, New York.
[29] N. Go, H. A. Scheraga, Macromol. 6 (1973) 273.
[30] G.-M. Greuel, G. Pfister, H. Schönemann, Singular 2.0. A computer algebra system for polynomial computations. Centre for Computer Algebra, University of Kaiserslautern, www.singular.uni-kl.de, 2001.
[31] P. Güntert, Symbolic NMR product operator calculations. Int J Quant Chem. (in press).
[32] W. A. Hereman, Symbolic computation of conservation laws of nonlinear partial differential equations in multi-dimensions. Int J Quant Chem. (in press).
[33] S. W. Homans, *A dictionary of concepts in NMR.* (1992) Clarendon Press, Oxford.
[34] F. C. Hoppensteadt, C. S. Peskin, *Modeling and simulation in medicine and the life sciences. 2nd ed.* (2002) Springer, New York.
[35] P. E. Hydon, An introduction to symmetry methods. Int J Quant Chem. (in press).
[36] J. Keener, J. Sneyd, *Mathematical physiology.* (1995) Springer, New York.
[37] M. H. Levitt, *Spin dynamics: basic principles of NMR spectroscopy.* (2001) Wiley, New York.
[38] R. H. Lewis, S. Bridgett, Conic tangency equations and Apollonius problems in biochemistry and pharmacology. (in press).
[39] Z. B. Li, Y. P. Liu, Comp. Phys. Comm. 163 (2004) 191.
[40] D. E. Makarov, H. Meliu, J. Chem. Phys. 108 (1998) 590.
[41] L. R. U. Manssur, R. Portugal, Comp. Phys. Comm. 157 (2004) 173.
[42] M. Minimair, M. P. Barnett, Mol. Phys. 102 (2004) 2521.
[43] P. J. Mulquiney, P. W. Kuchel, *Modelling metabolism with* MATHEMATICA*: incorporating a detailed analysis of human erythrocyte metabolism.* (2003) CRC Press, Boca Raton.
[44] J. D. Murray, *Mathematical biology. 2nd ed.* (1998) Springer, New York.
[45] H. Nilsson, A. Courtney, J. Peterson, *Proc.* ACM *2002* SIGPLAN *Haskell workshop*, (2002) 51.
[46] National Research Council, *Mathematical challenges from theoretical/computational chemistry.* (1995) National Academy Press, Washington D.C.
[47] P. Piecuch, S. Hirata, K. Kowalski, P. D. Fan, T. L Windus, Automated derivation and parallel computer implementation of renormalized and active-space coupled-cluster methods. Int J Quant Chem. (in press).
[48] G. Pistone, E. Riccomagno, H. P. Wynn, *Algebraic statistics: computational commutative algebra in statistics.* (2001) CRC Press, Boca Raton.
[49] G. Pólya, *Mathematics and Plausible Reasoning.* (1954) Princeton University Press.
[50] A. Ratkiewicz, T. N. Truong, Automated mechanism generation — from symbolic calculation to complex chemistry. Int J Quant Chem. (in press).
[51] C. Rose, M. D. Smith, *Mathematical statistics with* MATHEMATICA. (2002) Springer, New York.
[52] H. T. Stokes, D. M. Hatch, Phys. Rev. B65 (2002) 144114.

[53] T. Sugie, T. Ito, T. Ebisuzaki, Comp. Phys. Comm. 162 (2004) 37.
[54] B. Tian, W. Li, Y. T. Gao, Acta Mech. 160 (2003) 235.
[55] J. T'oth, J.-P. Rospars, Biosystems. 79 (2005) 33.
[56] B. Vail, D. Aris, M. Scarlete, Symbolic computation engines and molecular modeling templates: MAPLE-assisted point group analysis of the vibrational activity of molecules Int J Quant Chem. (in press)
[57] Z. Wang, Y. Ge, Y. Dai, D. Zhao, Comp. Phys. Comm. 160 (2004) 23.
[58] Wolfram, S. The Mathematica Book, 4th ed. Cambridge University Press, 1999.
[59] G.-Q. Xu, Z.-B.Li, Comp. Phys. Comm. 161 (2004) 65.
[60] Z. Yan, Phys. Lett. A. 331 (2004) 193.
[61] E. K. Yeargers, R. W. Shonkwiler, J. V. Herod, An introduction to the mathematics of biology: with computer algebra models. Birkhauser Boston Inc. Cambridge, MA, 1996.
[62] N. Yildirim, Use of symbolic and numeric computation techniques in the analysis of biochemical reaction networks. Int J Quant Chem. (in press).
[63] http://java.sun.com/docs/books/tutorial/java/concepts/
[64] http://www.cis.rit.edu/htbooks/nmr.
[65] http://www.claymath.org/programs/cmiworkshops/ascb/schedule.php.
[66] http://www.modelica.org/
[67] http://www.xml-cml.org/information/position.html