Mathematics and the real world

R. Rajaraman

School of Physical Sciences, Jawaharlal Nehru University, New Delhi 110 067, India

'... the enormous success of mathematics in the natural sciences is something bordering on the mysterious and ... there is no natural explanation for it.'

—Eugene Wigner

ONE of the most fascinating features of the intellectual development of the human species is the role that mathematics has played in it, not only as an academic discipline but also as a powerful instrument for understanding the external world and coping with it. In today's popular imagination, this role is captured by images of some absent minded Einstein-like professor whose indecipherable scribblings on the blackboard lead, mysteriously, to astounding practical consequences. But the role of mathematics in human thought is much more pervasive than that. It is not limited just to instances of sophisticated mathematics being used to discuss modern physics, design computers, plan satellite orbits or study the structure of economic models. Down-to-earth race course bookies, far removed from the world of science, rely on elementary probability theory in offering their bets, as do hard headed insurance companies in making their billion dollar profits. In fact practically everyone finds the need to use some mathematics, at least at the level of elementary arithmetic, such as addition, subtraction and multiplication, which are essential for day-to day life. Conversely, illiterate labourers in our country are continually being exploited by employers and shopkeepers because of their inability to perform these simplest of mathematical manipulations.

Thus the need for mathematics is really widespread, all the way from arithmetic in day-to-day life to the most esoteric concepts of topology and group theory employed in physics. But what remains intriguing even after more than two millennia of mathematical applications at different levels is the underlying issue of why mathematics 'works' in the real world and the way in which it works. Indeed, it is not a priori obvious that mathematics should work at all in the physical world, for reasons we will explain shortly. Not only is there no comprehensive or universally accepted theory explaining the success of mathematics in science, but any attempt to unravel this question in some logical fashion is fraught with difficulty. One can get tangled up in issues as varied as the evolution and functioning of the human brain, the role of the environment in developing it, the distinction between the 'inner' and the 'outer' world and so on. Many people have worried about this question, including some great minds, but there has been no unique and established answer. An example of such an attempt is the essay (which contains the quote given at the beginning of this article) entitled 'The unreasonable effectiveness of mathematics in the Natural Sciences' by Eugene Wigner, one of the most distinguished theoretical physicists of the 20th century[1].

Clearly, we cannot attempt here to give definitive answers to this riddle that has been tackled by so many people without reaching conclusive answers. Rather we will share some of our own thoughts on this question. We will also try to give those who have not had the opportunity to ponder on this issue some flavour of why the applicability of mathematics to the real world is indeed puzzling and merits explanation.

Is mathematics a science?

Perhaps one should begin with that old adage that 'Mathematics is the queen of all sciences'. Queen she certainly is. Mathematics represents one of most profound constructs of the human mind and fully deserves royal status. The author views mathematics with respect bordering on awe. But is she the queen of sciences? Is mathematics even a science? Science, whether it be physics, chemistry or biology, is supposed to study what is called the physical or external world 'out there', the entire spectrum varying from celestial bodies to chemicals, metals, gases, living organisms, plants, animals, molecules and sub-atomic particles. The observations and experiments which form the basis of science are expected to have an objective reality of their own, independent of the human observer and his psychological predispositions. (This criterion holds in a fairly straightforward sense for most of science. For systems in which quantum principles play a major role, it still holds in a statistical sense when appropriately generalized to ensembles.)

If science deals with the natural world 'out there', mathematics by contrast seems, at least at the working level, to be a construct solely of human minds. You could put a group of mathematicians in a closed room denied of all contact with the external world except for their pencil and writing pads (many of them do not even need that!) and they can still make progress in their subject and create new mathematics. Correctness or incorrectness of a mathematical statement, given a set of hypotheses, is decided by the fraternity of mathematicians and not by the physical world outside. To quote Keith Devlin, a mathematician who has done a lot of work to popularize the subject, '... for all that

e-mail: cloug0700@mail.jnu.ac.in

mathematical research feels like discovery, I firmly believe that mathematics does not exist outside of humans. It is something we, as a species, invent. I don't see what else it could be . . .'[2].

How does it happen that a subject like mathematics, seemingly constructed and policed entirely by the 'inner world' of human minds, ends up being such a successful tool in describing and indeed harnessing the external physical world? Is it that the physical world has some intrinsic 'mathematical order', which then instilled in human brains the basic concepts of mathematics and logic through the evolutionary process? In other words, *did the human mind learn about mathematics from the external world rather than the other way around?*

The miracle of mathematical consensus

To pursue this further, consider the closely related issue of how mathematics, even by itself, manages to exist as a human discipline. Let us elaborate on this a little. The very existence of mathematics seems to rely not so much on the external physical world, but on the presence of a common mental understanding among human beings on what is 'logically correct' and what is not, on what 'follows' from a set of hypothesis and what does not. Without such a understanding and universal agreement on logical reasonableness, mathematics cannot exist as a subject. Most times such agreement on the logical validity or otherwise of a proof is immediate among practising mathematicians. Other times, in exceedingly complex mathematical work, the correctness of a step (or series of steps) may not be easy to determine. For instance in the initial 1993 version of Andrew Wiles' proof of Pierre de Fermat's conjecture, there were apparently some gaps in argument discovered by colleagues and reviewers. Of course once this was pointed out Wiles agreed and tentatively withdrew his claim of having proved the historic conjecture. Subsequently Wiles himself was able, with the benefit of related work and suggestions by others, to find ways of proving the missing elements and offered the complete proof. This has since then withstood the scrutiny of the community. But that whole process took about a year![3] (To digress for a moment, this story is one more illustration of the nobility of mathematicians as a community. At least as far as outsiders could tell, it supported Wiles in completing the missing links in his proof, rather than indulge in tasteless quarrels of priority and credit through a whole year of uncertainty about this coveted proof.) From the point of view of our article, this story tells us something about the nature of mathematics – that even in a mental construct as complex as Wiles' proof, a large number of human beings (in fact all mathematicians with sufficient expertise to scrutinize the proof) agreed with one another, first that the original proof carried some extremely delicate flaw and subsequently that the new version was complete.

The interesting question here, and indeed with all mathematical literature and applications is: what is responsible for such agreement between all these people of diverse nationalities and background on such complex issues? If I may be permitted some caricature, what made them all so united in shaking their heads in criticism when Wiles' first proof was scrutinized, and then nodding their heads in agreement when the revised proof came? The same set of people would never show such unanimity in other matters, such as their political views, or their choice of what is the best piece of art or music.

The key players in the Wiles–Fermat theorem came from such diverse origins as England, Europe, the US and Japan. Their early lives and backgrounds were undoubtedly very different. They were not of the same detailed racial material, except in the larger sense of being humans. It is not as if they had some specific life-experience in common, some incident that happened to all of them, which predisposed them all to such agreement. In any case mathematics at the level of Wiles' proof is so abstract and far removed from day-to-day experiences that lifestyles can hardly matter. From where, then, comes such unanimity about which arguments are logically valid and which are not?

In this context it is useful to compare mathematics with music, one of the other great creations of the human mind. As with mathematics, appreciation of music again involves a commonality of taste and sensibility on the part of large numbers of people. But in some sense it is less universal. Music at the highest and most sophisticated level, whether it be Indian or Western classical music, was developed as an abstraction of various folk tunes of the region, religious songs, hymns and other sounds of the local environment. Therefore its appeal can be more localized. In my own experience, I have known many serious scholars and aficionados of Indian music to be quite immune to the greatness of Bach or Beethoven. Conversely I have also known both performers and music professors in the West (particularly prior to the 'sixties before likes of Ravi Shankar and the Beatles led to some 'globalization' of Indian music) to be unable to appreciate Indian music, in part because of its sliding notes, unfamiliar scales and its non-insistence on absolute pitch. Also, as acts of creativity musical compositions are ultimately personal. To quote Devlin again, 'If Beethoven had not lived, we would never have heard the piece we call his Ninth Symphony. If Shakespeare had not lived, we would never have seen Hamlet. But if, say, Newton had not lived, the world would have gotten calculus sooner or later, and it would have been exactly the same!'[4]

Thus even compared to music and other forms of art, all of which are profound human mental constructs and require a commonality of sensibility among vast numbers of people, there is a greater universality about mathematical logic. It is almost as if it has some 'external objective truth' which may take great human cleverness to uncover but whose validity is independent of that individual. The approach towards the proof of a result and its style and notations may be characteristic of that person, but not its core content.

These questions about the very existence of mathematics give us a possible clue on the relation between mathematics and the sciences. This very fact that the same basics of mathematical logic inhabit human minds of all nationalities and cultures may indicate that it is some sort of an inheritance of the collective human experience already at a very early and primitive level. If that is true, it may also make plausible that mathematical ideas, abstracted by human beings from nature in the first place, are then able to help in understanding nature.

Incorrigibility of mathematical results

A completely different family of explanations for why mathematics is so relevant to the physical world could go roughly as follows. In these scenarios, the external world is not necessarily orderly or mathematical but the human mind chooses to study, as quantitative science, only those aspects of the external world that are amenable to man-made mathematical laws, ignoring the rest as ‘non-science’. Or alternately one may argue that mathematics is used as a way of classifying physical phenomena and extracting idealizations of them amenable to quantitative analysis. In pursuing this class of explanations, it may be useful to recall Douglas Gasking’s essay ‘Mathematics and the world’[5,6]. In discussing the fundamental differences between mathematics on the one hand and empirical sciences on the other, Gasking argues that while mathematical propositions are ‘incorrigible’ by our experience, scientific propositions are corrigible. That is, while scientific laws are constantly open to correction in the light of newer observations, no empirical observation of the physical world can alter any statement considered a truism by mathematicians. The correctness or otherwise of the latter is decided entirely by the ‘internal’ rules of mathematical logic. Gasking’s assertion may sound too strong, and we may need to think long and deep to satisfy ourselves whether it is always true. But simple examples in mathematics already illustrate his assertion.

Consider, for instance, the theorem from Euclidean geometry that the sum of the angles of a triangle add up to 180°. Suppose you want to verify this theorem empirically by drawing a triangle on some surface. The theorem will never be vindicated to 100% accuracy by your measurement. For one thing, all real life measurements will unavoidably be subject to observational errors. Besides, even if you take pains to keep measurement errors very small you may still find that in some situations the sum of the angles of your triangle do not *even approximately* add up to 180°. When this happens a mathematician will not discard or correct Euclid’s theorem. Rather, he will point out that the theorem holds only for triangles drawn on planes and that it is your fault that the surface you have used is not planar. (For instance, you might have drawn your triangle on the surface of a football.) The same holds for the familiar Pythagorus theorem which states that the sides of a right angled triangle obey the relation $a^2 + b^2 = c^2$. It too has never been verified with 100% accuracy for the same reasons, viz. that measurements will always have instrumental errors and that the surface on which the triangle is drawn will, in real life, never be a perfect plane. Such empirical non-verifiability of a mathematical theorem to perfect accuracy never leads us to doubt the correctness of the theorem. The theorem is considered correct provided some conditions (hypotheses) are met, conditions which are mentally idealized, but not strictly available anywhere in nature!

This is true even for absurdly simple examples from basic arithmetic involving only integers, so that fractional errors in observation do not come into play. We have the mathematical statement that $20 + 10 = 30$. Now, suppose a boy’s mother places a plate of 20 hot freshly fried samosas on the dining table for some guests who are expected and a few minutes later brings from the kitchen another batch of 10 more samosas. As per the mathematical result mentioned, one expects altogether 30 samosas on the table. But what if the mother, upon seeing the boy standing nearby with a guilty look, decides to double-check and finds only 28 samosas? Any attempt on the boy’s part to wriggle out of the situation by modifying the mathematical proposition to read $20 + 10 = 28$ will not work. His mother will consider the original mathematical result $20 + 10 = 30$ to be sacrosanct and there may follow a painful investigation, right in front of the arriving guests, into where the ‘missing’ samosas went.

The above example may appear far too trivial in the context of our serious discussion. After all everybody would agree that there were indeed 30 samosas in existence altogether, as required by the addition rule, and the missing two can be ‘accounted for’. That is because we are so accustomed, in day-to-day life, to the total number of objects being conserved. But this need not always be the case. If we were dealing with, say, π^0 mesons instead of samosas, their numbers can fail to add up even without any guilty parties gobbling them up, since the number of π^0 mesons is not conserved (these elementary particles can be created or destroyed through their mutual interaction). If you placed 20 high energy π^0 mesons in an empty box, added 10 more and looked at the system a little later, you may well find 28 or 32 π^0 mesons in the box. But as a mathematical result, $20 + 10$ always equals 30. It is in this sense that mathematical propositions are incorrigible no matter what you observe in the physical world. Any deviation from the prediction of a mathematical statement will be attributed to the deficiencies of the physical system (instrumental errors, curvature of the surface, the boy’s inability to resist gobbling up samosas, etc.). We have deliberately chosen very elementary examples from basic arithmetic and geometry to illustrate this point in the simplest possible context, but the same holds also for possible physical realizations of any mathematical theorem, however complex.

If mathematical results have a sanctity of their own regardless of whether or not they are accurately realized in physical applications, then how does mathematics end up being so valuable in the real world? The above examples point to one important way in which mathematical results tell us something about a physical system. The failure of an empirical observation to agree with some mathematical prediction can indicate the extent to which the physical system did not accord with the hypotheses that went into the mathematical result. This can be made quite quantitative. For instance, if you draw a very small triangle the difference between 180° and the measured value of the sum of its angles can be used, after correcting for measurement errors, to get a quantitative measure of the local curvature of the surface on which it is drawn.

This also tells us something about the nature of mathematical predictions for the real world. If you had to predict the sum of the angles of a triangle (let us denote it by S) drawn on some unknown surface somewhere in the universe, mathematics would have no prediction at all for it without further information. Those angles may add up to 180° or they may not! It all would depend, among other things, on the curvature of that surface. Thus, a mathematical prediction about any system in the physical world requires other input information about the system. In the above example you need to know the curvature of the surface at every point on it. Now, one way to specify the curvature is to give the measured value of S for all different triangles drawn on the surface! If you do that, then that particular theorem has no predictive content but is only a definition, i.e. a classification index for the curvature of surfaces. In particular, Euclid's result the $S = 180°$ has no predictive value at all unless you know that the surface is a plane. But once a surface is established as a plane to some high accuracy, then all the other theorems of Euclidean geometry do contain many other predictions about triangles, circles and other figures drawn on that surface.

From this one sees that the nature of mathematical applications to any physical system involves correlating different properties of that system with one another through logical connections. Is this giving us some new information about those systems or is it just a case of some hidden tautology? For instance it will really take an infinite number of measurements to ascertain that any finite region of a surface is truly a plane. You would have to check the curvature at every infinitesimal (tiny) sub-region in that region, say, by drawing tiny triangles and measuring the value of S around each point. Therefore to be sure that the numerous theorems of plane geometry are applicable to a given surface you really have to first provide the infinitely many pieces of information that go into ascertaining that it is indeed a plane. Similarly, if you wish to make completely certain predictions with 100% accuracy about the properties of photons and electrons using quantum electrodynamics (QED), you first need to know that the system of electrons and photons does obey the Lagrangian of QED at every space time point, which again amounts to giving as input an infinite amount of information!

In actual practice of course that is not the way we use mathematics for real systems. We use a combination of approximations and 'modelling'. We *assume* that a given physical surface *is* a plane, apply some plane geometric theorems and see if the answers agree with measurements to some acceptable accuracy. Similarly we assume that electrons and photons obey the QED Lagrangian and other postulates of quantum field theory, calculate scattering cross sections using mathematical techniques, and then look for agreement with observation. Clearly this requires the ingenuity of the scientist in finding a good model for a given system, being conversant with the mathematical techniques needed for deriving the consequences of that model and having a large number of samples and observations to verify various predictions.

This also raises the reverse possibility that even discoveries in pure mathematics, at least in its early days, were made possible by the availability of physical systems that were fairly accurate manifestations of the postulates underlying that mathematics. For instance, the theorems of Plane Geometry were discovered by Euclid because the lines, triangles and circles that people dealt with in his time were drawn mostly on planar (flat) surfaces such as table tops. Important early applications of geometry were in land surveys and architecture which again involved plane vertical walls, pillars and land (the earth is flat to a high accuracy within the size-scale of farms and towns). Thus the various theorems of plane geometry were vindicated by numerous empirical observations (within the accuracy of measurements). Once the conclusions of a theorem are empirically obeyed in many cases, then it becomes useful even in those cases where there is some disagreement, as discussed earlier. A piece of land on which Euclidean theorems are not obeyed is presumably not planar, and one could then try to rectify this by land filling and so on.

(But if Euclid had been a creature living on highly curved membranes, soap bubbles and rumpled bed sheets, it is doubtful if he would have discovered his geometrical rules. Of course the theorem that the sum of the angles of a planar triangle add up to 180° would still have been a truism. But, for creatures living on highly curved surfaces, this truism may only eventually emerge as a special case of some more complicated theorems they discover. Similarly if the universe had consisted solely of (highly intelligent) π^0 mesons or photons, whose numbers are not conserved, it is doubtful if they would have discovered our elementary rules of addition and subtraction. For that matter, even the concept of numbers may not have been discovered in such a world where the number of objects has no sanctity!)

Such considerations tend to support the possibility raised earlier, that the content of mathematics (or at least its early simple branches) was developed in response to the physical environment of human beings. Indeed, as we trace the growth of modern physics we can see the parallel symbiotic growth in mathematical methods and results.

The physics–mathematics symbiosis

From its beginning 'modern' physics (i.e. of the seventeenth century and beyond) has been very closely related to mathematics. In fact the organizational separation that prevails between the two subjects in today's academic community was simply not there in the early days of science. Both subjects were viewed as part of the larger pursuit of natural philosophy. Often the same people worked on both areas even at the highest levels. Newton, in order to make precise his laws of motion and obtain exact consequences from them had perforce to also to learn and invent aspects of differential calculus. Similarly in order to calculate the gravitational force of large bodies, starting from the force between point-masses required the development of integral calculus. (Incidentally the controversy between Newton and Leibnitz on credit for discoveries in calculus is well known. But as Herbert W. Turnbull, the British algebrist, points out, long before Newton, another member of the physics pantheon, Kepler, in effect must have used some techniques of infinitesimals in calculating areas bounded by curved orbits. Recall his Law on equal areas swept by planets. He apparently developed these methods in part to estimate the volume of wine-casks![7]) Returning to Newton, although he is generally listed among the great physicists, he was of course also a great mathematician of his time, and the title of his magnum opus was in fact *Philosophiae Naturalis Principia Mathematica* (Mathematical Principles of Natural Philosophy). Similarly, as an undergraduate I first encountered the name of Gauss in connection with physics – the Gauss theorem as used in electrostatics. It is only later in life that I found mathematicians claiming this great man as their own, and as it turns out, with justice, given the huge contributions he had made to pure mathematics. In the mind of the common man too, the distinction between mathematicians and theoretical physicists is blurred. Was Einstein a physicist or a mathematician? To me he was undoubtedly a physicist, but in popular parlance his name is synonymous with mathematical genius!

The further advancement of physics beyond classical mechanics to cover electricity, magnetism, sound and light waves were all made possible by corresponding developments in the theory of ordinary and partial differential equations. The derivations of the results were greatly simplified by the use of complex variable theory. For a long time after that there were continuing instances of physical developments motivating new areas and results in mathematics, such as symmetries in physics spurring the growth of group theory, Brownian motion leading to functional integrals and the Weiner measure (which in turn facilitated Feynman's formulation of quantum mechanics) or most recently the use of non-abelian gauge theory combined with supersymmetry in particle physics set the stage for important results in modern mathematics through the work of Witten and others.

Such a symbiotic mutually supportive growth in our understanding of the physical world on the one hand and our discoveries in mathematics on the other, renders less mysterious the successful application of the latter in understanding and coping with the world. But as mathematics has grown more and more abstract in the past 50 years and farther removed from human experience, instances of such synergy between mathematics and physics become rarer. Correspondingly most of modern mathematics was developed, not from physical examples, but from exerting the power of logical thinking inherent in the mathematician's brain. As we have remarked earlier the basic framework of logic on which all mathematical results rest must be a very general attribute assimilated by the human brain fairly early in human history – or else it could not enjoy the kind of universality it does. It seems incredible that the machinery for constructing the excruciatingly abstract areas of today's modern mathematics could have evolved from primitive human experience of the external world. Besides, even the broad contours of when and how, in biological terms, the 'software' of mathematics (or is it actually the hardware in terms of neural connections?) was loaded into human brains in the evolutionary process are not available. At which stages of human history did the rules of logic and axiomatic deduction, common to all branches of mathematical thinking, get imbedded in human minds? Alternately, if the development of mathematics in humans had nothing to do with the environment, why then is it so successful in describing aspects of the external world?

So we are back to questions with which we had started. We have done no more than elaborate on the issues without arriving at firm conclusions. But, as we said in the beginning, established and universally accepted answers to these questions are not available. We still do not fully understand, to use Wigner's phrase, the 'unreasonable effectiveness' of mathematics.

Mathematics and good taste

We would like to conclude with a somewhat judgmental comment on the use of mathematics in other branches of knowledge. A crucial ingredient in such use in any science or social science is a sense of taste and proportion in the choice of the mathematics to be used. Such taste has not always been displayed and its importance has not been emphasized as much as it should have been. Historically, mathematical methods were first used to any substantial extent only in physics and astronomy. By the end of the nineteenth century, the theoretical underpinning of all of physics and physical engineering was mathematics based. The precision of analysis and accuracy of prediction that this brought to physics were impressive. Gradually but inevitably, this led to a trend among other fields of knowledge as well, of emulating physics in the practice of using mathematical formulations. Today chemical, biological and environmental sciences include many significant sub-fields that employ mathematical techniques. In social sciences, quantitative empirical data and their statistical analysis are used to substantiate and augment qualitative and intuitive theories. In economics, advanced ideas of game theory and even topology come into play. Management theorists employ complicated optimization techniques.

By and large this trend had desirable consequences. More and more subjects were driven by this development towards including quantitative forms of analysis as part of the field. The use of mathematical formulations and equations also induced further precision of thought. But, for all the advantages it offers, there is also a negative side to such widespread use of mathematical methods and 'models' in more and more fields. It can give rise to a misguided impression that a piece of work heavy with mathematical equations necessarily contains results commensurately useful or relevant to the system it has set out to study. Conversely, it fosters a feeling that people who do not explicitly use mathematical methods and symbols are less precise or rigorous in their thinking. Not infrequently this leads to some snobbery associated with the use of mathematical methods. These methods with their mysterious symbols and equations are used to overwhelm others not fluent in them. Such trends can be injurious to the healthy development of a field. They can distort priorities, deflect attention away from the really important issues relevant to the subject matter and should be curbed. It must be remembered that most pioneers and deep thinkers in any field are blessed with intrinsic powers of precision and analysis regardless of the formal level of mathematics they employ. Neither Freud nor Karl Marx nor Darwin employed any mathematical techniques in their gigantic path-breaking work. (One could even argue that it is your 'average good' scientist for whom the mathematical language is more important. It keeps him on the straight and narrow path of logic and prevents him from wandering into vague or internally contradictory statements. Mathematical equations expose such follies.)

As a corollary to this, even in those areas where some mathematics is truly needed and useful, a sense of good taste has to prevail to avoid excess. There is a level of mathematical formulation appropriate to any given problem. Using a more sophisticated version would not only be a case of cracking peanuts with a sledgehammer, but can often obfuscate the real issues. Of course there are many topics, particularly in physics, which genuinely and unavoidably require advanced mathematical machinery. When Dirac first formally codified quantum theory he had to employ the canvas of infinite dimensional vector spaces and operators. There is no significantly simpler, less complicated, mathematics that can comprehensively describe the broad array of new physical and philosophical ideas that quantum theory contains. (The alternate but equivalent formulation by Feynman using path integrals – a method whose core idea was suggested by Dirac himself – is in its own way equally complicated, with basic quantum results like the Heisenberg uncertainty principle requiring very gingerly treatment of the jagged 'paths'). Einstein's General Theory of Relativity again had to unavoidably use the differential geometry of curved 4-dimensional space-time. Similarly, the theory of elementary particles and their strong and electro-weak interactions has to unavoidably use quantum non-abelian gauge fields. All this is fine, as long as the criteria of minimality and unavoidability characterize the choice of the mathematical apparatus used to study a problem. Anything less will not do the job and anything more will be wasteful, if not detrimental.

1. *Symmetries and Reflections – Scientific Essays of Eugene P. Wigner,* Indiana University Press, Bloomington, USA, 1967.
2. Keith Devlin, MAA ONLINE July/August 2001, The Mathematical Association of America, Copyright © 2004.
3. For a short and readable history of the Fermat conjecture see the article by O'Connor, J. J. and Robertson, E. F., available at the website http://www-gap.dcs.st-and.ac.uk/~history/HistTopics/Fermat's_last_theorem.html.
4. Keith Devlin, *op cit.*
5. Flew, A. G. N. (ed.), Mathematics and the World by Douglas Gasking. In *Logic and Language*, 2nd Series, Blackwell and Mott Ltd.
6. Gasking's essay has been reprinted in *The World of Mathematics* (ed. James R. Newman), Simon and Schuster, New York, 1956, vol. 3, p. 1708. Newman's book is a rich source of essays on mathematics by several distinguished authors.
7. The Great Mathematicians by Prof. Herbert W. Turnbull, Methuen & Co Ltd, reprinted in Newman, *op cit,* pp. 75–168.

ACKNOWLEDGEMENTS. I thank my colleague Prof. Deepak Kumar for drawing my attention to several source books on this subject and letting me borrow them. I am also grateful to him and Prof. Rajesh Gopakumar for helpful discussions.

The reciprocal interaction between mathematics and natural law

Sunil Mukhi

Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005, India

In this article it is argued that the relationship between mathematics and physics has undergone a qualitative change in the last two decades. We have entered a new era in which research in theoretical physics is providing a stream of new mathematical ideas and relationships, some of which are yet to be understood using conventional mathematical tools. Some examples of this new paradigm are provided.

IN ancient times the distinction among mathematics, physics and other natural sciences was not clearly articulated, nor was there necessarily a distinction among the practitioners of these sciences. Many mathematicians were also physicists and vice versa, Newton being only the most famous example. But over the last century or so, it is harder to find examples of individuals straddling both fields, and the modern university typically has distinct mathematics and physics departments in pursuit of distinct goals.

In the era of 'modern physics', starting from the early twentieth century, physicists have seen mathematics primarily as a language in which their observations can be codified. In his famous article 'The unreasonable effectiveness of mathematics in the natural sciences', Wigner put it as follows: 'The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift....' Many new developments have taken place in the four-and-a-half decades since Wigner made this observation. The miracle to which he was referring remains alive and well today, but there are also indications that major changes are on the way.

As Wigner indicated, the new physical notions that came up during the twentieth century readily found a setting in established areas of mathematics. One such example is the theory of continuous groups, known as 'Lie groups', and their associated algebras. It was realized that rotations in space form a group, the orthogonal group in three dimensions, O(3), whose representations play an especially important role in quantum mechanics. In special relativity, the rotation group gets subsumed into the Lorentz group, O(3, 1), which contains besides rotations, the 'boosts' between space and time. In the 1960s it became necessary to invoke a new class of groups, the unitary groups, to classify fundamental particles. The experimentally observed particles conveniently fell into representations, or multiplets, of these unitary groups.

e-mail: mukhi@theory.tifr.res.in

It was just as well that orthogonal and unitary groups had already been discovered and classified by the time they were needed in physics. Otherwise, for example, the discoverers of particle spectra would have been at a loss to deal with the regularities they were observing, which could have found no explanation without an understanding of unitary symmetry.

Once the concept of Lie groups entered physics there was considerable work done by physicists to further develop the mathematics. However, it is fair to say that from the mathematicians' point of view, this work did not add significant conceptual material to what was already present in the mathematics literature. To be sure, physicists did discover new and important facts about Lie groups, but they were typically not the sort of facts, and did not have the degree of beauty and depth, that would appeal to mathematicians. Hence it is not surprising that the enthusiasm of physicists to learn mathematics has usually been greater than that of mathematicians to learn physics, the latter being less confident that they would have something to gain.

Many other twentieth-century developments, such as the introduction of differential geometry in the study of gravity, of infinite-dimensional vector spaces in quantum mechanics, and of fibre bundles in gauge theory, exemplify the same story. The role of mathematics in physics in all these cases was limited in two ways. First, the core mathematical results needed had already been obtained, and second, despite considerable mathematical work by physicists, mathematicians did not learn much of significance to themselves from the physicists.

Of course this is a value judgement, and heavily depends on what we mean by 'of significance'. Wigner himself wrote a textbook entitled *Group theory*. But then, the subtitle of this book was *and its Application to the Quantum Mechanics of Atomic Spectra* and the primary audience for the book is physicists. Taking other similar cases into account, this value judgement does not seem far off course.

However, during the 1980s, something about this situation changed quite drastically. The change is hard to understand at present, because we are still in the middle of it, or maybe close to the beginning. It will be the focus of this article. My main observation will be that the interface between mathematics and physics has turned into a two-way relationship of a unique and unprecedented nature. Not only does mathematics provide physics with a formulation it can use, but physics has started to provide mathematics with methods and results that it needs, desires and appreciates on its own terms. It would be tempting to describe this new situation in terms of 'The unreasonable effectiveness of physics in mathematics', which would have made an appropriate title for this article. But an internet search reveals that this very title has been used in articles and colloquia by distinguished mathematicians and physicists like Michael Atiyah, Robbert Dijkgraaf and Arthur Jaffe, just over the last two years. So there appears to be a growing consensus that the role reversal which will be the subject of our discussion is really taking place.

Consistency and experiment

Before turning to the new paradigms thrown up in the 1980s, I would like to briefly touch upon an extra ingredient in the mathematics–physics relationship – the remarkable role played by consistency. In the twentieth century it happened, more than once, that mathematics did not merely act as a codifying tool, but appeared to have predictive power over nature. Merely by requiring mathematical consistency, one was sometimes forced to believe in new, as yet undiscovered, physical phenomena. The phenomena in question were then later confirmed by experiment.

The classic example was Dirac's equation for the electron, which he postulated in 1928. It was soon realized that while the equation worked well for electrons, it also predicted the existence of another particle with the same mass as an electron but opposite electric charge. This prediction arose because for a charged particle of spin one-half, a Lorentz-invariant equation necessarily has at least two kinds of solutions. So it was mathematical consistency that required the new particle and no way could be found to get rid of it from the equation. Instead, in 1932 the new particle, the 'positron', was experimentally detected.

As is widely known, Dirac later commented: 'it is more important to have beauty in one's equations that to have them fit experiment', a statement that comes across as bold, controversial, even downright unscientific. What would Dirac, or anyone else, have done if the positron had not been discovered? Or was he saying that the positron had no choice but to be discovered?

The rest of Dirac's comment, which is less widely publicized, gives a better picture of what he had in mind. 'If there is no complete agreement between the results of one's work and experiment, one should not allow oneself to be too discouraged, because the discrepancy may well be due to minor features that are not properly taken into account and that will get cleared up with further development of the theory.' He was only saying that mathematics is a more reliable guide to nature than one might expect, but not that it is an ultimate arbiter of natural law.

In 1931, Dirac proposed the existence of magnetic monopoles from similar considerations of mathematical consistency. He argued that monopoles were consistent with all known facts about quantum mechanics, therefore they ought to exist. Fifty years later, during which period no monopole had been discovered, he retracted his proposal in the most abjectly empiricist language. He wrote to Abdus Salam in 1981: 'I am inclined now to believe that monopoles do not exist. So many years have gone by without any encouragement from the experimental side'.

Taken as a composite, the quotations above give a more accurate picture of the situation. Dirac's views about the role of mathematics were far from extremist. He simply believed that beauty in the equations was a more reliable guide than most people had hitherto believed. I take this as a slight strengthening of Wigner's observations on the effectiveness of mathematics in the natural sciences, but no more than that.

Mathematics and quantum field theory

The major influence in creating a different paradigm for the mathematics–physics relationship is undoubtedly Edward Witten at Princeton, though others, notably Witten's former student Cumrun Vafa at Harvard, have been extremely active at the forefront.

Despite being a physicist, Witten received the highest honour in mathematics, the Fields Medal, in 1990 for his contributions to mathematics using tools of physics. In the citation for the award, Ludwig Faddeev said that 'Physics was always a source of stimulus and inspiration for Mathematics… In classical time(s) its connection with mathematics was mostly via Analysis, in particular through Partial Differential Equations. However, (the) quantum era gradually brought a new life. Now Algebra, Geometry and Topology, Complex Analysis and Algebraic Geometry enter naturally into Mathematical Physics and get new insights from it.'

Putting it as simply as possible, Witten's strategy was to rephrase mathematical concepts in the language of Quantum Field Theory, a framework developed for the description of fundamental particles and their interactions. He would then use insights and tools native to quantum field theory, such as symmetries and invariances, and the (notoriously hard to define) Feynman–Kac path integral. This would lead to a 'natural setting' for the mathematics problem and often provide not only a solution of it, but an enormous variety of generalizations. The solutions that Witten provided were not rigorous, but in many cases were rendered rigorous after some additional work. In an article that was intended to be Witten's Fields citation but which he was unable to actually deliver at the award ceremony, Michael Atiyah notes 11 papers of Witten that justify his being awarded this prestigious prize. To give the reader a flavour of what was done, I will briefly touch upon a few of them.

Knot theory

The study of knots in three-dimensional space was initiated in the 19th century by the physicist Lord Kelvin. He had a theory of 'vortex atoms', according to which the different atoms that occur in nature are made of the same 'substance' but knotted in different ways. According to this theory, the classification of atoms could be reduced to the classification of knots, and Kelvin accordingly embarked on the latter problem.

Over the ensuing decades, considerable progress on classifying knots was made by mathematicians. Two knots are distinct if they cannot be continuously deformed into each other. But for complicated configurations, it is hard to establish whether or not this is possible just by looking at pictures of the knots. Two pictures can look quite different from each other, but there might be some 'moves' that smoothly deform one of them to the other – then they would describe the same knot. The idea then emerged of associating an 'invariant' to a knot. This would be a number that is calculable starting from any picture of the knot, and that remains invariant under smooth deformations of the knot. Now if from two pictures we get two different invariants, then we can be sure the knots are distinct. The converse is not in general true, but a powerful knot invariant will be able to distinguish more and more subtly differing knots from each other.

Knot invariants were extensively developed and the most powerful ones (some of which were found by the mathematician Vaughn Jones) turned out to be polynomials in one or more parameters. Witten sought a relationship between these polynomials and a certain gauge field theory in three spatial dimensions called 'Chern
In this theory, the natural observable is a 'Wilson line' (a line of generalized electric flux). The flux can run along an arbitrary closed contour, the same thing as a knot. Being an observable, the Wilson line may have a quantum mechanical 'expectation value' in the vacuum of the theory. When computed using techniques of quantum field theory, this expectation value turned out to be a polynomial in some parameters related to the coupling constant of the theory and the representation (the type of charges) flowing around the Wilson line. In fact, the expectation value was none other than the Jones polynomial.

For knot theory, this was a dazzling illumination: it gave an interpretation to the polynomials, it explained (via a relation to conformal field theory) some of their properties and relations to the theory of quantum groups, it related the parameters in the knot polynomials to an expansion parameter in a perturbations series, and it provided an easy way to generalize the polynomial, by changing the gauge group of the gauge theory from SU(2) to an arbitrary Lie group. Moreover, as Atiyah points out, Witten's field-theory interpretation 'is the only intrinsically three-dimensional interpretation of the Jones invariants; all previous definitions employ a presentation of a knot by a plane diagram or by a braid.'

Now the Chern–Simons gauge theory at the root of this brilliant discovery is a cousin of the Yang–Mills theory, that describes three of the four fundamental interactions in nature. Yang–Mills theory, in turn, was a generalization of Maxwell's theory of electromagnetism, a relativistic field theory with a gauge symmetry. So without the whole circle of ideas and experiments relating to relativity and electromagnetism at the start of the twentieth century, we (more precisely, Witten) would probably not have found a new way to describe knot invariants, and a major mathematical illumination could never have taken place.

Morse theory

Morse theory studies the topology of differentiable manifolds. It relates two different quantities: on the one hand the critical points of some suitable function defined on them, and on the other hand their homology, or the kinds of nontrivial 'cycles' that can be embedded in them. One can gain insight into the topology of a manifold by knowing about its cycles and, via Morse theory, one gets this by studying critical points. Witten provided a proof of certain inequalities, the 'Morse inequalities' by setting up an equivalence between the homology of the manifold and a toy physical system called supersymmetric quantum mechanics.

For the lay person, this is harder to appreciate as compared to knot polynomials. But here too, one can see where the physics input came from. Starting around 1974, theoretical physicists made a breakthrough in their search for a symmetry that unifies bosons and fermions, the two different kinds of particles that occur in nature. The new symmetry was dubbed 'supersymmetry'. There are compelling reasons to believe that, although it is hidden from view in everyday life, nature makes use of it in some subtle way. So compelling are these reasons, that the Large Hadron Collider (LHC) in Geneva will start looking for evidence of supersymmetry from 2007 and the chances of finding it are considered rather high. Witten's supersymmetric quantum mechanics is a highly simplified version of the physical theories whose experimental confirmation is being sought at LHC, but uses the same basic structure. By mapping it onto the problem of Morse theory, Witten was able to use some properties of the supersymmetric path integral to make headway into pure mathematics. Again, without the experimental fact of bosons and fermions and the physicist's desire to unify them, this mathematical insight into Morse inequalities would not have been possible.

The Moduli spaces of Riemann surfaces

To a lay person, a Riemann surface is just a two dimensional surface, examples being a sphere, a torus (like a bicycle tyre) and multi-handled generalizations. They are more naturally described in the language of complex analysis. Mathematicians would like to know how these surfaces can be smoothly varied and in what way their mathematical properties change as we vary them. The 'moduli space' of such a surface is the space of parameters that label this variation. Sophisticated results due to Mumford and Morita and others, had given considerable insight into the global structure (topology) of these moduli spaces.

Witten was inspired to address this by considering a problem in gravitation. However, his theory of gravity was not the one discovered by Einstein, but a simpler version of it where spacetime is two-dimensional. From a physical point of view the theory has very little dynamics in this setup, but Witten analysed the operators that made up the observables of this theory, and asked what were their correlation functions – which measure how quantum fluctuations are interlinked in a field theory. What he discovered was that these correlation functions were identical to the Mumford–Morita classes on the moduli space of Riemann surfaces. As usual, this approach also gave rise to important generalizations of the known mathematical results.

This time the physical inspiration came from gravity, and its theoretical structure as elucidated by Einstein in his theory of General Relativity. Though Witten's gravity theory was very different, it is clear that if General Relativity had remained undiscovered, or if there were simply no gravitational force in nature (and assuming we still existed), this insightful approach to moduli spaces would never have been found. And therefore some profound results about the abstract manifolds called Riemann surfaces would not have been known to mathematicians.

Atiyah once made a comment about Witten's contribution to mathematics on the following lines. Usually, physics provides an intuitive notion, or a way of thinking about things, and then mathematicians prove a corresponding rigorous result. But in the case of Witten's work, rigorous results in mathematics already existed and Witten then provided the intuitive explanation for them. In this sense, the mathematicians were ahead of the game, though to say this in no way undermines the contribution of Witten's work. As already indicated, his physical re-interpretations led to generalizations of the mathematics which would have been almost impossible to guess if one only knew the rigorous methods. And by providing a new setting for old ideas, they provided a number of pointers towards new mathematical directions. We will look at one of those directions now.

Mathematics and string theory

The subject of this section will be an area of research in which not only is physics in the process of influencing mathematics, but, for the moment at least, the physics approach seems to be ahead of the game.

We will not need to know much string theory for the purpose of this discussion. It is enough to mention here that string theory is the most serious candidate known for a consistent quantum theory of gravity. It has not received direct experimental confirmation, but it is built on many experimental facts: the number of dimensions we live in, the presence in nature of both fermions and bosons, and the existence of gauge symmetry as in Yang–Mills theory and of general coordinate invariance as in Einstein's theory of General Relativity.

Einstein taught us that gravity is the geometry of spacetime. But he was talking about classical gravity, not the quantum version, for the simple reason that in his lifetime, no quantum version was known. The classical field theory of gravity used a great deal of mathematics from the realm of Riemannian geometry. Spacetime is a (pseudo)-Riemannian manifold. A quantum theory should be a theory of fluctuating manifolds, consistent with the basic principles of quantum mechanics but such a theory proved very elusive in Einstein's own lifetime.

Even in Einstein's time, there were some theories of gravity that required not only a physically observable spacetime, but also an extra hidden space that was too small to observe experimentally. This hidden space affected our world only indirectly. The notion of such hidden spaces and their promising role in the unification of forces was proposed by Kaluza and Klein early in the twentieth century. The hidden space was to be a Riemannian manifold and the possibilities for what it could be were governed by standard geometric considerations, well-known to mathematicians.

Now in discovering string theory, it appears that we have finally discovered a quantum theory of gravity. Like the older Kaluza–Klein theories, string theory too requires a hidden space in addition to the observable spacetime. But since the new theory is a quantum theory, the observable and hidden spaces can no longer be thought of as purely classical objects. They need not correspond to conventional Riemannian geometry except in a limit where quantum effects are negligible. To what sort of geometry do they correspond in general?

The answer is not completely known, but there is by now considerable evidence that there is a vast geometrical structure, which we will call 'stringy quantum geometry' for lack of a better word, that generalizes the usual mathematical notions of geometry in an exceedingly strange way.

Here I will just highlight two (related) properties of the quantum geometry associated to string theory. One is called target–space duality and the other is called mirror symmetry.

Target–space duality

In conventional Riemannian geometry, one has the notion of a metric which defines the distance between two points. This is a very appealing notion to a physicist. It plays a key role in general relativity where the metric of spacetime is itself the fundamental dynamical variable. In conventional geometry, a distance can take any value from zero to infinity.

Now let us consider an internal spatial direction that is compactified into a circle. Classically this circle has a radius, given by the minimum distance one traverses in making a complete tour of the circle and returning to the starting point. In the limit of large radius, we would call the direction 'non-compact' and roughly that is what the three familiar spatial dimensions in the real world look like.

Suppose now that we consider a finite value of the radius, and then vary it so that it becomes smaller and smaller. In classical geometry this process never ends, and the circle simply continues to shrink. If we place a quantum mechanical particle on this space, it will have difficulty exciting itself into a mode that can propagate on the small circle. The reason is that a well-defined wave on a tiny circle needs to be rapidly varying and therefore must have a high wave number, or momentum. Therefore to probe this circle, the particle must have a high energy. This is precisely what we mean by saying that a compact internal dimension is physically unobservable when its radius is small.

Now replace the particle with a string. Physically it is clear that at low energies a string behaves very much like a point particle, and it will have the same difficulty entering the small internal space unless we give it a large energy. However, the string can do something else that also probes presence of this space. It can wind itself around this circular direction. This is a classical configuration and it is easy to see that it carries an energy proportional to the length of the direction. So the energy required to excite this 'stringy' mode actually decreases as the circle shrinks. Therefore the string has a spectrum of heavy momentum modes as well as light winding modes.

Now suppose we replace the circle of a given radius by one of the inverse radius (in appropriate units). In this way a small-radius circle gets mapped to a large-radius one. In conventional geometry this is a major change. But a string on this space again has a spectrum of light modes (but now they are the modes of a string propagating on the large circle) and a spectrum of heavy modes (in which the strings winds over the large circle). In other words, the spectrum of a string remains invariant on replacing a tiny direction with a huge one. Going beyond the intuitive picture presented here, inversion of the radius can be shown to be an exact symmetry in string theory. This symmetry goes by the name of target–space duality or 'T-duality'.

What is the consequence of this for mathematics? Geometry, the study of shapes of objects, clearly originated with something or someone being able to actually probe an object. That probe is the macroscopic-sized human being, or perhaps one of the point particles out of which we are made. But with the introduction of a new probe, the string, the observed geometry is very different. In this new geometry, there would be a minimum length scale and the geometry would be invariant under inversion of a compact direction.

It is too early to say what are the implications of this result within mathematics. Perhaps one day, schoolchildren will be taught that a circle of a given radius is the same as a circle of the inverse radius, even though pointlike objects would not be the appropriate probes to demonstrate this fact. Perhaps it will enter the lore in a different way. Whatever the case, stringy quantum geometry as illustrated by this example promises to be as new and as strange as Riemannian geometry was when compared to the older Euclidean geometry of flat space.

Mirror symmetry

In string theory we believe that the internal space hidden from our view is a 6-dimensional geometrical manifold. The number 6 arises because strings like to propagate in 10 dimensions, while we live in only 4. This means the remaining 6 dimensions must be invisible, so they are assumed to be compact and small. String theory requires them to be special manifolds called 'Calabi Yau' spaces, with a definite set of properties.

Now in the 1980s some groups of researchers discovered that such 6-dimensional manifolds come in pairs, with each member of the pair having little or no resemblance to the other one in the conventional geometric sense. However, strings propagate in precisely the same way on both members of the pair, and this is again known to be an exact symmetry of string theory. In other words, in the domain of stringy quantum geometry these manifolds would be completely indistinguishable from each other. They are called 'mirror

Here is a regularity that was totally unexpected at the outset. On encountering it, the first thing a physicist would do is to enquire if mathematicians understand this regularity. And here is the remarkable surprise: they do not.

Just like a small circle and a large circle, two Calabi–Yau spaces that form a mirror pair appear to be completely different from each other, but string theory recognizes them as being the same. In fact, the underlying feature of string theory responsible for this symmetry is the same in both cases, target–space duality and mirror symmetry. The latter is a more complicated manifestation of the same phenomenon.

In summary, we believe there is a new branch of mathematics, 'stringy quantum geometry', within which the symmetries discussed above are manifest. But today all that we know about this new branch of mathematics is what string theory tells us. As mentioned earlier, in this case the physics technique is ahead and the mathematicians have yet to catch up, though many are working hard on it at present.

I would guess that the newly two-sided relationship between mathematics and physics is likely to occupy a part of centre-stage in both fields for the next few decades.

Role of mathematics in chemistry

Debashis Mukherjee

Indian Association for the Cultivation of Science, Jadavpur, Kolkata 700 032, India

ALTHOUGH chemistry was practised from the dawn of civilization as the discipline to create materials, including the extraction of metals, it was more of a craft of artisans than a subject of enquiry into the fundamentals. Physics, in contrast, evolved very rapidly as an 'exact' science, in particular in the hands of Archimedes, Galileo and Newton, where the fundamental laws of motion got quantified and predictions were precise enough to distinguish between the rival conceptual frameworks. The major reason why chemistry developed into an exact science relatively late is that the underlying laws of binding and transformations of chemical substances have their basis in the quantum behaviour of the constituents of matter. The behaviour of chemical substances, as isolated species or in bulk – which dominates our world of senses – are, however, only indirectly related to their microscopic constitution and this has remained a problematic ontological issue which deterred an intellectually satisfying and integrated quantitative conceptual framework for chemistry. Moreover, chemistry as a discipline enjoys a degree of autonomy in the sense that the desirable goals of a chemist (control of emergent chemical behaviour, designing molecules with specific chemistry, monitoring chemical transformations into well-defined channels) are determined by questions and aesthetics of chemical nature. In this sense, chemistry is more 'complex' than physics.

I look upon the less complex science as one which engenders the fundamental basis of a science which lies one tier higher in complexity. To borrow the terminology from biology we may say that the fundamental basis of a science are genotypes, while the emergent properties arising out of the genotypical laws are phenotypes. I want to argue that chemistry is the simplest science of complexity since the fundamental physical laws are its genotypes and the emergent chemical expressions are the phenotypes. Chemistry is thus compatible with physical laws but not reducible to them. The really interesting problems in chemistry seem to remain fully unresolved in terms of understanding from physical principles because scientists have not come to grips in discerning the pattern, structure and interconversions displayed by molecules from the fundamentals of subatomic physics. This is despite the fact that we understand the quantum and statistical mechanical laws of physics well enough but it is neither unique nor trivial to pose questions of chemical nature in terms of physical laws. The complexity of chemistry has even an underlying extra degree of freedom in the sense that the superstructure of chemical functions is to some extent insensitive to the physical laws underpinning them. Results from a more quantitative formulation from a more fundamental basis often lead to qualitatively similar but quantitatively different conclusions, so that certain empirical generalizations can well describe chemistry and even lead to an illuminating understanding, quite independent of the underlying laws of the substratum. Obviously, by 'understanding' we mean assessing the relative importance of the various processes reflected in some conceptual constructs which act together to shape the phenomena of interest. Models emerge when we tie up understanding and quantitative descriptions of the conceptual constructs and weave a story out of it. Stories are complete or convincing to various extents depending on the mix of understanding and quantification. Another appealing simile is sculpting. Much is removed but much remains also for the pattern to emerge. Recognizing what to remove yet emphasizing the essence in all its splendour is an orphic endeavour of sorts, involving inspiration, metaphor, symbolic representation and innovative analogy. The role of mathematics in chemistry must satisfy this polysemiotic and polymimetic richness.

We thus distinguish quantum molecular physics as somewhat distinct from theoretical chemistry when we want to discuss the role of mathematics in chemistry. An appreciation of this difference is often not made, leading either to a perception that brute force computation or even empirical quantitative simulation would lead to understanding chemical significance or to the dismissive attitude of the experimental chemists that theoretical chemistry fails to provide predictive answers of chemical significance, when in fact they are probably pointing out the limitations of computational molecular physics. When I talk about the role of mathematics in chemistry, I have in mind an evolving, conceptually integrated and many layered theoretical framework of molecular science, which subsumes both molecular/materials physics and chemistry and chemical biology as subdisciplines. This is a never ending saga but the stories become more and more complex as we begin to see more intricate patterns and can relate them more and more to the physical laws underpinning them.

The genotypes of chemistry are embedded in quantum mechanics, equilibrium and non-equilibrium statistical mechanics, and diffusion behaviour in fluids. The phenotypes are the molecules displaying myriad chemical properties in isolation and in transformation. The chemical concepts like bonds, lone pair, aromaticity, electronegetivity, reso-

e-mail: pcdm@iacs.res.in

nance, functional groups, etc. evolved from the attempts of chemists to grapple with the diverse behaviour via these empirical constructs to systematize chemical behaviour.

The unreasonable effectiveness of mathematics in the natural sciences, in particular physics – as noted by Wigner, emphasizes the striking synergy of mathematics and physics. Examples abound spanning several centuries which tell us how old mathematics have been found to be tailor-made to promote progress in physics, how new mathematics got evolved to quantify new-found laws and how demands of the physical world have led to creation of new mathematics. Starting from the theory of curves, the development of calculus in quantifying laws of motions, and the use of non-Euclidean geometry to understand gravitation to the more recent examples of modular and elliptic curves and complex manifolds, and Calabi-Yau spaces are some striking examples of this symbiosis. Another versatile tool which has been finding more and more use in physics is compression of information using the idea of algorithmic complexity. On an immediate level this leads to optimal representation of a model with a given ontological content; as a long-term perspective it offers the possibility of modelling complex sciences through the irreducible genotype components – augmenting them with the initial and boundary conditions that are extraneous to the fundamental laws and thus of leap-frogging from a less complex science to a more complex form.

As of now, mathematics of classical era of physics has permeated the whole of chemistry as a quantifying tool of simpler chemical processes and for providing a microscopic understanding of chemical transformation from first principles. This has several levels of sophistication: (a) simple quantification of empirical data (empirical models) (b) statistical chemical data analysis, quantitative analytical chemistry, emerging into cheminformatics and chemical data analysis (c) molecular similarity described in terms of target phenotypes leading to quantitative structure–activity correlation (d) quantitative simulation models: reducing *ab-initio* theories to computational chemistry via parametrization and (e) simulation via modellization on a quantum mechanical scenario (path-integral molecular dynamics). Evolution of powerful computers has been providing the chemists with quantitative answers for predicting reactive intermediates, branching ratios of various products or to understand the relative stability in their ground, excited, or ionized states. Methods of molecular electronic structure theory are being routinely used to predict theories via black-box programmes. Studying the properties of metal alloys, complex fluids and materials of various types have also matured as robust computational discipline. These are the triumphs of quantum and statistical molecular physics. On a simpler substrata, simulations based on empirical fits or extracting quantum mechanical many-body potentials have led to successful predictions of reaction processes in solvated species, of the dynamics of the conformational changes and to ion migration through membranes. These developments, along with the emergence of cheminformatics have resulted in a discipline which is turning out to be very important as fundamental inputs to structural biology and bioinformatics. Control of spatio-temporal patterns via nonlinear systems equations, theoretical electrochemistry on random and ultrametric surfaces, behaviour of solvated species in critical and sub-critical conditions are also some of the important quantitative developments where mathematical modellings have led to fundamental insight into reacting diffusive systems, electrode processes and behaviour of solvated matter under phase transitions. They embody a vast corpus of chemical activity at a quantitative level, which lies at the interface of molecular physics and theoretical chemistry. This part of the story-telling entails first principles formulations leading to phenotypes from genotypes but only for the simplest of the chemical reality. Of far greater importance are of course the desiderata transcending the border of physics into the autonomous theoretical constructs of chemistry – in short, generating chemical theories.

The emerging frontier of theoretical chemistry already encompasses the use of algebraic topology to discern patterns in structure–activity correlation. The various notions of theoretical linguistics are also finding their place in the axiomatic formulation of chemistry, although they have not succeeded as yet in making useful predictions. The concept of virial fragments in identifying functional groups separated by surfaces of zero density gradient in a molecule is a very fruitful innovation where theoretical chemistry could morph a chemical concept out of physical laws which has a credible autonomous validity and predictability. The use of concept of homotopy and manifolds has led to a very concise understanding of the various classes of the excited states of potential hyperenergy surfaces and to enumerate all possible topologically distinct reaction pathways. Use of artificial intelligence has proved a powerful tool to prune stray pathways in predicting chemical reactions and this, coupled with the homotopy theory, will prove more and more useful in classifying and discerning patterns of chemical transformations. The methods of control theory have also been innovatively transcribed into the field of laser-control of chemical reactions and fundamental insight has already been obtained in fine tuning of bond-breaking and bond-making process and also for identifying transition states. Although we are still far away from monitoring and controlling channel-specific reactions, the activity in this field promises to be very significant and control and systems theory will enter more and more in reaction dynamics which will redefine the boundaries of molecular science.

The complexity of chemistry precludes having a single ‘correct’ analysis of the chemical entities expressed in a single adequate language. The world of chemistry is just too rich and diverse, requiring multifaceted representations which capture the various layers of chemical reality. In contrast to physics, chemistry requires and has naturally evolved a symbolic representation (a language?) that is expressed in notations that are at once symbolic and iconic as in Chinese ideograms. In empirical chemistry, a chemist, like an artist, moves subconsciously between the phenotypes and genotypes, between the levels of potentiality and actuality of transformations or even between different levels of reality. A chemist may enquire about a molecular structure and function generated by the primary, the secondary and the tertiary conformations. On one hand a chemist may try to understand the reactivity of a molecule (which is a property of propensity) and may try to relate it to a real reaction (actuality) which demands the linguistic dichotomy of the symbolic and iconic components embodying the propensity and actuality. The effect of environment generating the concepts of the directionality of the bonds and lone pairs reflects another level of dichotomy where standard quantum mechanics, in both time-independent and dependent forms, seems ineffective and theories of subdynamics describing open systems appear as more natural tools. Chemistry as a craftsmanship has always been aware of the looseness and multilayered nature of the concepts of bonds and lone pairs, and has creatively used octet rules, double quartet rules and various other stereoselectivity criteria to design rational synthesis and to understand chirality of molecules and the powerful use of chiral induction and chiral catalysis. A quantitative phrasing of these concepts in terms of subdynamics, where the environment leads to the emergence of the molecular shapes, including emergence of chirality from achiral interactions via symmetry breaking, is yet to emerge in an unambiguous manner but the theory of subdynamics has already reached a state of development where more quantitative and predictive formulations are anticipated in the foreseeable future. The effect of relativity in molecules containing heavy atoms has already been quantitatively understood for simple molecules, although the qualitative concepts embodying these quantitative formulations have been rather slow to emerge. Such innovations are essential for understanding molecular aggregates with heavy atoms, clusters with heavy atoms and their fruitful use in molecular electronics and for molecular assemblages tailor-made for molecular collective computation.

Matter in extreme conditions such as plasma has already been quantitatively understood in terms of gas and ion reactions and more quantitative models for ion beam imaging and ion lithography are rapidly emerging. Another intriguing frontier is evolving where very cold atoms appear to generate aggregates in Bose–Einstein condensation, with the possibility of a crossover between Bose–Einstein condensation of atoms (which are bosons due to intra-atomic coupling of electronic and nuclear spins) to the BCS type of pairing condensates via magnetic monitoring. These are giant aggregates (condensates) which are not molecules in the conventional sense, thus creating new forms of matter. This opens up the exotic field of ultra-cold chemical aggregates.

It seems eminently possible that mathematical tools like pattern recognition, artificial intelligence, collective computing, algebraic topology and fuzzy logic will make major forays into chemistry and create an integrated platform where these tools will enmesh fruitfully with quantum mechanics and statistical mechanics. The entire field of molecular, materials and biomolecular science is in a state of flux – an ideal melting pot to brew new ideas. The tantalizing transition from quantum molecular physics to theoretical chemistry will happen sooner than later.

Chemistry – The middle kingdom

Gautam R. Desiraju

School of Chemistry, University of Hyderabad, Hyderabad 500 046, India

Chemistry occupies a unique middle position between physics and mathematics on the one side and biology, ecology, sociology and economics on the other. It is said that chemistry is reducible into physics and finally mathematics. However, in moving from the covalent to the non-covalent world we obtain a new chemistry, one that is a starting point for the emergence of the soft sciences. This article argues that this new chemistry represents a paradigm shift in the way in which chemists think about their subject today. Biology may be considered as emerging out of this new chemistry, which in itself cannot be reduced into physics and mathematics as was the case for chemistry thus far practised. This dualistic nature of chemistry, reducible and irreducible, is a new development but nevertheless one that ensures that the subject will remain alive and well in the foreseeable future.

'. . . so kann Chymie nichts mehr als systematische Kunst, oder Experimentallehre, niemals aber eigentliche Wissenschaft werden, weil die Principien derselben ¼ der Anwendung der Mathematik unfähig sind.'

— Immanuel Kant
Metaphysische Anfangsgründe der Naturwissenschaft
Riga, 1786

'. . . so chemistry can be no more than systematic art or experimental teachings, indeed never real science, because its principles . . . do not lend themselves to the application of mathematics.'

— Immanuel Kant

THE position of mathematics on the scientific grandstand is indisputable and indeed nearly axiomatic to all practitioners of science[1]. One accepts, without any argument, that mathematics provides a template for rational thought and for the logical development of scientific discourse. The subject defines order and discipline, furnishing protocols to establish relationships between cause and effect. It is impossible to conceive of any science without the mathematical underpinning. I have yet to come across a good scientist who disliked mathematics. I will not dwell further on the primary role of mathematics in the natural sciences, namely, as an aid to organised thought. It is the second role of mathematics, as a subject into which all other scientific disciplines may be reduced, that is far more alluring to a chemist.

According to reductionist thinking, all science can ultimately be reduced into mathematics. Reductionism would have it that biology is reducible into chemistry, chemistry into physics, and ultimately, physics into mathematics. This 'unreasonable effectiveness' of mathematics[2] in explaining natural phenomena confer upon it almost mystical qualities. In keeping with these qualities, and also because mathematics has been termed a language, one may draw analogies between it and Sanskrit, the language of the Gods. Both these languages are precise and accurate, and yet remain aloof. They seem to describe the reality that surrounds them only too well, and yet they remain tantalizingly apart from this very reality. Wigner said that the appropriateness of mathematics for the formulation of laws that govern physical phenomena is 'a wonderful gift which we neither understand nor deserve'[2]. Might I add that the same could be said about the appropriateness of Sanskrit to an understanding of the workings of the mind?

The origins of chemistry

During antiquity and medieval times, Western science was based on the holistic thinking of Aristotle. Modern science, with its emphasis on reductionism, came into being with the Renaissance[3], and astronomy and physics were the first sciences to feel the impact of mathematics. Chemistry, however, was curiously resistant to these developments for nearly two centuries. With its origins in alchemy and the black arts and with the frenzied attempts of its practitioners to transmute base metals into gold, chemistry retained its qualitative character. It is curious, even amusing, to note today that the great Newton was a closet alchemist who felt at the time of his death that his work in alchemy would eventually be recognized as prominently as his contributions to mathematics, astronomy and physics. The first winds of change came from the work of Boyle who sensed the concept of the modern chemical elements and demolished the Aristotlean concept of 'four elements and three principles'. The real break with the past, however, came with Lavoisier who emphasized the importance of quantitative experimentation. He and, independently, Dalton provided the first framework of atomic theory and were the earliest of the great chemists. Still, it is worthwhile to ponder a little on why chemistry resisted quantification for so long. The subject is deliciously qualitative even today and this dichotomy of character between the quantitative and the qualitative, the reducible and the irreducible, is what I wish to highlight in this article.

e-mail: gautam_desiraju@yahoo.com

The nineteenth century

Friedrich Wöhler's synthesis of urea from ammonium cyanate in 1828 triggered two important developments. Until that time, urea was only obtainable from animal matter, and yet ammonium cyanate is a salt of indisputably inanimate origin. Wöhler's experiment signalled the beginning of the end of vitalism as a scientific dogma[4].

Secondly, organic chemistry emerged as a separate subject within the chemical domain. The philosophy of vitalism went back to 1600 with its roots in the distinction that was perceived to exist between organic and inorganic compounds. This distinction had to do with the behaviour of these compounds upon heating. Inorganic compounds could be recovered upon removing the heat source, or so this argument went, whereas organic compounds appeared to undergo mysterious and irreversible transformations when heated. This led to the thought that organic compounds were imbued with a special vital force and in turn to the belief that while organic compounds might obey the same physical and chemical laws as inorganic compounds, life could not be governed by just these laws. The synthesis of urea, an organic compound, from ammonium cyanate, an inorganic compound, sounded the death knell of the vital force theory, and the letters between Wöhler and his teacher Berzelius, a staunch advocate of vitalism, make for fascinating reading even to this day.

From then onwards, chemistry in the nineteenth century was one unbridled run of synergistic analysis and synthesis. A vast amount of empirical data mostly on organic compounds were painstakingly accumulated, especially in Germany, and in the hands of grandmasters like Kekulé, Liebig, Baeyer and Willstätter, rational and reductionist thought assumed a nearly art form in chemistry. For, reductionism is nothing other than analysis and synthesis coupled together, when cause can be used to predict effect or when, and with equal validity, effect may be used to decipher cause. The object of study of these German organic chemists was the isolated molecule, and gradually there arose a considerable body of work in support of the notion that all the physical and chemical properties of a substance are characteristic of, or contained within, its molecular structure. This dogma was to persist for more than a hundred years. In my view, though, the high noon of reductionism in nineteenth century chemistry did not belong to organic chemistry but to Mendeleev and his periodic table of elements[5]. Even today the appeal of this table to students of chemistry is palpable. Which novitiate has not marvelled at the fact that the properties of bromine are nearly the mean of the corresponding properties of its congeners chlorine and iodine? When Mendeleev asked Lecoq to check the specific gravity of the newly discovered gallium once again because it was lower than what he had predicted, and when this value was revised upwards from 4.8 to the predicted 5.9 after careful purification of the sample, we have before us one of the most impressive examples of the success of reductionism.

All this synchronized well with other developments in the natural and social sciences. The late nineteenth century saw the zenith of the Industrial Revolution, the emergence of capitalism and colonialism as economic doctrines and the importance given to the individual in relation to the group. Aristotle's holistic thinking finally gave way to the reductionism of Darwin[6]. Even in chemistry, it was recognized that there were areas of the subject that were even more amenable to reduction into physics and mathematics than organic chemistry, however systematized the latter had become, and the work of Ostwald, van't Hoff and Arrhenius led to the demarcation of physical chemistry as a separate field of study[7]. This was an exciting and new subject in the late 19th century, the molecular biology of its time. Studies of aqueous solutions and chemical thermodynamics transformed scientific knowledge of chemical affinity. This emergence of a new discipline at the boundaries of physics and chemistry wrought deep-seated changes throughout chemistry. In turn, physical chemistry was eclipsed by its own offspring, quantum chemistry and for this we need to consider the contributions of the most outstanding chemist of all time, Pauling.

The Al(l)chemist

Linus Pauling was of the greatest significance to the growth and development of chemistry as a subject because it was he who showed conclusively the distinction between chemistry and physics. From wave mechanics to quantum chemistry is but a subtle step but the consequences are fundamental and deep-seated. Pauling's essential contribution, the concept of the covalent bond, meant that chemistry did not need physics any longer in its day-to-day functioning and operation. This articulation of chemistry as an independent subject was the handiwork of this great scientist[8].

Pauling's contributions were important and varied, and extended across disciplines like crystallography, mineralogy, biology, medicine, anaesthesia, immunology and above all, structural chemistry. The basic theme that runs through his work is that one can explain the structures and properties of molecules with an understanding of the chemical bond, especially the covalent bond[9]. His influence may be assessed by the fact that at the time of the first edition of *The Nature of the Chemical Bond* in 1939, less than 0.01% of today's structural information was available and yet the generalizations and conclusions he drew then on molecular structure are largely valid even today. His impact on inorganic chemistry was immediate in that he could explain the magnetic properties of transition metal coordination compounds. Curiously, however, Pauling was silent about organic chemistry even after enunciating its most basic feature, namely, that a sharing out of electrons evenly among equivalent energy states, or what chemists call hybridization of bond orbitals, leads to an explanation for the tetrahedral valences of the saturated carbon atom. This silence has been ascribed to various reasons of a non-scientific type, but I feel that his reluctance to come to terms with organic chemistry arose from his realization that his reductionist approach could only be taken so far in this most qualitative branch of the subject. Pauling's ideas apply well to structure, reactivity and analysis but not as easily to dynamics and synthesis.

Supramolecular chemistry – beyond Pauling

Pauling's work elevated the molecule to the high altar and it was taken as the delimiter of all the important physical and chemical properties of a substance, to the extent that there was no world outside it. His formidable influence on chemistry in general might have accounted for the relatively late take-off of the subject of supramolecular chemistry, nearly seventy years after Emil Fischer enunciated his famous lock-and-key principle of enzyme action[10]. The scope and possibilities of this new subject were clearly enunciated by Jean-Marie Lehn[11]. Supramolecular chemistry literally means chemistry beyond the molecule and the main idea here is that if molecules such as **A** and **B** were to form an aggregate of the type [**A**.**B**] using weak non-covalent interactions, the properties and more significantly the functions of the aggregate need not be readily derivable from the individual properties of **A** and **B**.

This type of thinking is especially appropriate to biological systems because some of the most important biological phenomena do not involve the making and breaking of covalent bonds – the linkages that connect atoms to form molecules. Instead, biological structures are usually made from loose aggregates that are held together by weak, non-covalent interactions. Because of their dynamic nature, these interactions are responsible for most of the processes occurring in living systems. Chemists were slow to recognize the enormous variety – in terms of structure, properties and functions – offered by this more relaxed approach to making chemical compounds[12].

Fischer's lock-and-key mechanism proposed that an enzyme interacts with its substrate as a key does with its lock[13]. This elegant mechanism contains the two main principles of supramolecular chemistry – molecular recognition and supramolecular function. The idea of molecular recognition is that it takes place provided there is compatibility between the interacting partners **A** and **B** with respect to both the geometry and the non-covalent interactions. In turn, specific recognition leads to useful and specific supramolecular functions. For example, it is important that an enzyme works only on the appropriate substrate and not on any other compound. A key without its own lock or a lock without its own key is quite useless. **A** without **B**, or **B** without **A**, is meaningless in a functional context.

The implications are profound as far as the reductionist approach to chemistry is concerned. Reductionism in chemistry, that is the explanation of chemical phenomena in physical and mathematical terms, began with Wöhler and progressed through Pauling until the present time. But with the arrival of supramolecular chemistry, chemists looked more closely at the reduction of biology into chemistry. Can biology be really reduced into chemistry? If so, what are the implications? How do life processes work at a molecular level? How does one differentiate life from non-life? The fantastic levels of specificity achieved by biological machines may still, in principle, be reduced to the chemistry of weak interactions.[14] Yet, a reductionist approach is simplistic beyond the extreme. One may apply reductionist arguments in going from biology to chemistry but one would lose so much detail that it would be impossible to reconstruct the original from the reduction. Living and non-living matter differ not in content but rather in organizational complexity – and our understanding of this theme may well turn out to be the biggest breakthrough in modern chemistry.

Complex and complicated – emergence

Supramolecular chemistry provides a convenient introduction to chemists about the notion of complexity. At the outset, it is necessary to distinguish between the terms *complex* and *complicated*. A complicated system, like a high precision Swiss chronometer, consists of many components each of which is well understood in isolation. The functioning of such a system is also fully understood and derivable from the functions of the individual components. A characteristic of a complicated system is that if one of the components stops working, the whole system can quickly grind to a halt. Therefore when one designs a complicated system, one builds in redundancy. A complex system is, however, quite different. Complexity is well illustrated by the continuous flow of traffic through an intersection of many roads, such as is seen in large American cities. There may be as many as ten roads approaching the intersection and any vehicle may approach from any road and proceed onwards through the intersection onto any other road, all this taking place without any vehicle ever stopping at the intersection. The functioning of a complex system is not easily understood from the functioning of its individual components. For example, a traffic intersection of ten roads is not easily designed or derived from an intersection of say, four roads. Returning to chemistry, a 20-step synthesis of a natural product with several stereocentres is an example of a complicated system. A supramolecular synthesis as exemplified by the crystallization of a small organic molecule or the folding of a protein[15], or the spread of cancer in living tissues[16] are examples of complexity. If functional complicated systems need to incorporate redundancy (many doctoral students working on the natural product synthesis and all doing essentially the same things), complex systems are characterized by adaptability (crystal polymorphism[17], biological signalling pathways[18] in the spread of cancer).

Unlike a complicated system, a complex system does not necessarily break down because one of the components is not present or working. It merely modifies or mutates. If one of the roads leading to the busy traffic intersection were to be blocked, traffic on the other roads would continue to move largely unimpeded. If one were to change the solvent in a crystallization experiment, one might obtain another polymorph[19], and when one tries to fight cancer with a new drug, the disease adapts itself so that it attacks the cell using another pathway.

Closely allied to the notion of complexity is the idea of *emergence*[20]. Emergent phenomena are structures, behaviours, events or patterns that arise only when a large number of individual agents (molecules, cells, water droplets, musical notes, ants, birds, people, stars) somehow aggregate. Unless a critical number of agents act together, the phenomenon does not occur. An emergent property is created when something becomes more than the sum of its parts. The whole is difficult to predict from the properties of individual parts and it is no surprise then that supramolecular chemistry is full of emergent phenomena. Crystallization, for example, is a process of complex pattern formation arising from cooperative behaviour between components, and is still very hard to predict[21].

Emergence and reductionism are nearly antithetical. Reductionism implies the ease of understanding one level in terms of another. Emergent properties are, however, more easily understood in their own right than in terms of lower level properties. This suggests that emergence is a psychological property and not a metaphysical absolute. A property is classified as emergent based at least in part on the difficulty of an observer deducing the higher level property from the lower level property. Conversely, an increase of knowledge about the way certain effects are obtained in a system may reveal that they are decomposable into the effects contributed by the subcomponents of that system. In the mid-19th century, the reaction of sodium hydroxide (NaOH) and hydrochloric acid (HCl) to give NaCl and H_2O was quoted as an emergent property, as it was held that the properties of NaCl and H_2O are not understandable from the aggregate of the properties of NaOH and HCl (for instance NaOH and HCl are both corrosive while NaCl and H_2O are harmless)[22]. After the electronic structure of atoms was known, the above reaction became easily understandable. In the end, complexity is a temporal attribute. What is complex today might become merely complicated tomorrow, or even trivial, like the acid-base neutralization reaction given above.

A useful way of looking at mathematics and its relationship to the physical and natural sciences is in terms of emergence. Rather than say that biology can be reduced into chemistry, which can then be reduced into physics and finally into mathematics, one could say that biology emerges out of chemistry, which emerges out of physics, which emerges out of mathematics, which emerges out of the mind contemplating the Absolute, like Sankara's doctrines of *advaita*. We note that each level of investigation (mathematics, physics, chemistry, biology) has its own explanatory relationships, and yet if we check carefully there is no 'added extra' coming in from anywhere. There are no mysterious ingredients added as we proceed from a lower level to a higher level. The only place from which these value additions can emerge is the mind. Hence one concludes again that emergent phenomena are psychological in nature.

An analogy from the world of music is appropriate here. From the twelve notes in geometric progression that are used in the well-tempered scale of Western music, one progresses to the 22 microtones or *srutis* within an octave in Indian music[23]. *Ragas* or musical forms/moods, are characterized by the use of particular microtones that occur within smaller frequency ranges located around the twelve notes[24]. But if this were all, a raga would be reducible into srutis. This is clearly not the case. In the Carnatic system, one obtains pairs of ragas like Darbar and Nayaki, Aarabhi and Devagandhari, or Surati and Kedaragaula wherein the microtones are practically identical but their structuring, scaffolding and emphasis (*raga svarupa*) are so different that even a non-expert can distinguish between the ragas in any pair. In the Hindustani system, one has the raga trio of Puriya, Marwa and Sohini where again one perceives a similar effect. So rather than say that Darbar and Nayaki can be reduced to Kharaharapriya (the parent scale of microtones which is one of 72 possibilities called *melakartas*)[24], one could more constructively say that Darbar and Nayaki emerge out of Kharaharapriya. The value addition again arises from within the mind, and ragas then surely emerge out of srutis.

Emergent properties

In supramolecular chemistry, one makes higher level aggregates (supermolecules) from lower level entities (molecules) using weak intermolecular interactions as a glue.

Crystal structure prediction. Crystallization is the ultimate supramolecular reaction, just as the crystal is the ultimate supermolecule. A molecule may be said to consist of several functionalities or functional groups $[F_1, F_2, \ldots, F_n]$ and during crystallization, these functionalities come together through a process of molecular recognition and utilizing weak interactions to generate supramolecular synthons $[S_1, S_2, \ldots, S_N]$[25]. The conjunction of particular supramolecular synthons uniquely defines a crystal structure. However, there are two serious problems that arise when one attempts to predict the outcome of crystallization.

(i) The number of possible supramolecular synthons is large because the intermolecular interactions are weak.

$S_1 = F_1 \ldots F_1$ $\qquad$ $S_4 = F_1 \ldots F_4 \ldots F_5$

$S_2 = F_1 \ldots F_2$ $\quad$ $S_5 = F_2 \ldots F_3$

$S_3 = F_1 \ldots F_2$ (with F_3 linked to both F_1 and F_2: \ /, F_3) $\quad$ $S_6 = F_2 \ldots F_3 \ldots F_4$ (with F_2 linked to F_3: |, F_2) and so on …

From this it is obvious that the number of possible supramolecular synthons quickly becomes very large, even though the appearance of some supramolecular synthons will preclude the formation of others.

(ii) Interference from remote functionalities may be unpredictable. One notes that instead of say, S_1 appearing from the association of F_1 and F_1, it may associate preferentially with F_2 to give S_2 because of the presence (or absence) of some F_i in another location of the molecule, with F_i seemingly unrelated or unconnected with either F_1 or F_2.

Both these problems are endemic and it is for good reason that the prediction of a crystal structure of a small organic molecule (higher level property) from the structural formula (lower level property) has been deemed to be one of the most challenging scientific problems of the 21st century[21,26].

Hydrogen bridges – water aggregates. The hydrogen bridge, or hydrogen bond, is an interaction X–H…A wherein an electropositive H-atom acts as a bridge between two electronegative atoms X and A. There are many varieties of hydrogen bond and the energies associated with X–H and H…A may be widely different to nearly the same[27,28]. Although the phenomenon has been studied extensively for a century, it is surprising that there are no rules that allow the chemist to estimate the geometry and energetics of the hydrogen bridge from the formulas of the interacting molecules. This indicates that chemists have so far not been able to understand the hydrogen bridge phenomenon in all its complexity[29]. While this is, in general, true of all intermolecular interactions, the hydrogen bridge is the most important interaction in molecular recognition, supramolecular chemistry and biology and therefore it merits special mention as an emergent property.

In this connection, hydrogen bond arrangements that involve water are most fascinating. Water is almost a philosophical abstraction. Hardly a molecule in the usual sense, its surface is composed entirely of strong hydrogen bond donor and acceptor regions. It is small but supramolecularly very potent. Therefore it plays a crucial role in molecular association and aggregation – and indeed for life itself. Water is found associated with other molecules, and with itself, in many ways. The study of liquid water is fascinating and there are 'different' types of water molecules present[30]. These are characterized by different geometrical coordinations (supramolecular synthons), by different dynamical properties (slow, fast) and different locations (surface, bulk). In crystals water occurs in a myriad environments, bound to itself or to other molecules. The amazing feature of these patterns is that new ones are constantly being discovered[31]. There seem to be no limit to the variety of water…water association patterns. Clearly this is an important example of an emergent property of fundamental importance.

Fluorinated compounds. The supramolecular behaviour of the halogens (fluorine, chlorine, bromine, iodine) is still very difficult to understand[32]. Reference has already been made to the periodic law in connection with Cl, Br and I. However, this law breaks down more or less regularly in supramolecular chemistry and the properties of Br are not the mean of the corresponding properties of Cl and I. In supramolecular chemistry, Br behaves more or less like Cl, or more or less like I, *depending on the system under consideration.* Accordingly, Br-atom interactions are emergent properties.

Fluorine is even more complex. If one takes a hydrocarbon and successively replaces the H-atoms by F-atoms, the boiling point rises (as it is expected to) for a while but after the extent of fluorination crosses a critical value, the boiling points begin to fall. The boiling point of the perfluoro derivative may even be lower than that of the original fully hydrogenated compound. For example, the boiling point of methane and its fluorinated derivatives are as follows: CH_4 (–161.5°C), CH_3F (–78.4°C), CH_2F_2 (–51.7°C), CHF_3 (–82.2°C), CF_4 (–128.0°C). Such behaviour is not seen in the other halogens. For example, the boiling points of the corresponding chloromethanes are: CH_3Cl (–24.2°C); CH_2Cl_2 (39.5°C); $CHCl_3$ (61.2°C); CCl_4 (76.0°C). No one has been able to explain such anomalies satisfactorily. Fluorine is also unusual in that the so-called 'fluorous' compounds with many C–F bonds (say, teflon) are neither hydrophilic or hydrophobic. But a fluorous molecule is not simply a fluorine-rich molecule – perfluorohexane is fluorous but hexafluorobenzene is not. So a new concept of fluorophilicity/fluorophobicity is invoked,[33] but no one has really been able to quantify this. Many drugs that are in active clinical use today contain fluorine but no one knows just why the F-atom is so ubiquitous in medicinal chemistry. As a sampling, one might mention Allergan, Cifran, Clinoril, Dalmane, Diflucan Haldol, Lescol, Orap, Prozac and Uftoral (all registered trademarks). Again, the C–F group is unable to accept hydrogen bonds like the C–O and C–N groups, although fluorine is more electronegative than oxygen and nitrogen. Truly, fluorine chemistry is one of the last frontiers in chemical research and emergence is more or less rampant[34].

To summarise, universality in the behaviour of complex systems often reveals itself in forms that are essentially independent of the details of microscopic dynamics. A representative paradigm of complex behavior in nature is cooperative evolution, seen in structural and supramolecular chemistry as self-assembly and crystallization (chemical sociology). The interaction of individuals gives rise to a wide variety of collective phenomena that strongly differ from individual dynamics such as demographic evolution, cultural and technological development, and economic activity. Each human is part of a family of six billion members.

The middle kingdom

Chemistry is poised midway among the sciences, straddling the space between physics and mathematics on the one side, and biology, ecology, sociology, economics and the higher sciences on the other. The history of chemistry from the early 19th to the late 20th centuries represents the consolidation of reductionist and Paulingesque thought, the triumph of inductive and deductive logic in synthesis as seen in the work of Woodward, Corey and others[35], and defines a tightly knit body of work that, in the end, is more or less reducible into physics and finally mathematics[36]. Much of this chemistry was built with concepts and models like acidity/basicity, electronegativity/electropositivity, oxidation/reduction, hardness/softness, enthalpy/entropy, kinetics/thermodynamics, reactivity/selectivity, electrophilicity/nucleophilicity, hydrophobicity/hydrophilicity and chirality/achirality. Although most of these models cannot be rigorously derived from physics and mathematics, they still constitute a continuum with these more exact sciences. This continuity was established through physical chemistry, which enjoyed a dominant role in the development of chemistry as a whole during the 20th century. Chemistry is oceanic with respect to factual information, but it has always been contained with respect to the number of concepts and models that were required to understand all these facts.

In contrast, the reduction of biology to chemistry has always been problematic[37]. The appearance of supramolecular chemistry on the scene in the late 20th century stimulated new thinking about the relationship of chemistry with biology. This new chemistry is less about structure and more about organization, less about reactivity and more about dynamics, less about synthesis and more about association. All this represents a paradigm shift in the way in which chemists think about their subject today[38]. The argument is that biology and the other higher sciences may be considered as emerging out of this new chemistry, *which in itself cannot be reduced into physics and mathematics* as was the case for chemistry as it has been practiced thus far. An entirely new set of properties can emerge from the interplay of macrosystems that are not related directly to their component atoms and molecules[39,40]. The idea of emergence is being linked to biological pathways and this approach is being used to explain the evolution of complex self-organizing systems in a way that opens up a huge discontinuity from physics and 20th century chemistry. Living systems are viewed as autonomous self-reproducing entities that operate upon information, that originates at the molecular level by covalent chemistry, transferred and processed through non-covalent chemistry, expanded in complexity at the system level and are ultimately changed through reproduction and natural selection[41].

This new chemistry then promises to be the language of biology in the same way that mathematics is the language of physics and the older chemistry[1]. As a language, the new chemistry shares much with mathematics. While biology cannot exist without chemistry, (supramolecular) chemistry seems to develop well without biology[42,43]. When chemistry is used in biology, it is only a small fraction of the totality of chemistry that is so used. In other words, there is a large surplus of chemistry that is not even relevant to biology. This is a characteristic of a language. To paraphrase Jean-Marie Lehn, chemistry is all about diversity[11] but biology does not need all this diversity. Biology is about complexity and in the process of emerging from (supramolecular) chemistry, complexity builds itself around the chemical core.

Chemistry then occupies a unique middle position in the scientific arena. Its development thus far may be traced as an emergence from the harder sciences, physics and mathematics. In moving from the covalent to the non-covalent world, however, it enters totally different territory, a domain that is a starting point for the development of the softer sciences. This dualistic nature of chemistry is a new development but one that ensures that the subject will remain robust in the foreseeable future. It has been bemoaned that most of the important problems in chemistry have been solved and that all that remains now is to fill in the details. Nothing could be further from the truth. Almost imperceptibly, and in silent revolution, the subject has evolved so dramatically that future possibilities for the middle kingdom appear almost limitless.

1. Sarukkai, S., Science studies. *Economic and Political Wkly*, 30 August 2003, pp. 3647–3702.
2. Wigner, E. P., *Commun. Pure Appl. Math.*, 1960, **13**.
3. Leicester, H. M., *The Historical Background of Chemistry*, Wiley, New York, 1956.
4. Gupta, S., *Nature*, 2000, **407**, 677.
5. Strathern, P., *Mendeleev's Dream. The Quest for the Elements*, Penguin, London, 2000.
6. Davies, K. G., *Nature*, 2000, **407**, 135.
7. Laidler, K. J., *The World of Physical Chemistry*, OUP, Oxford, 1993.
8. Desiraju, G. R., *Nature*, 2000, **408**, 407.
9. Pauling, L., *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, Cornell University Press, Ithaca, 1939.
10. Fischer, E., *Ber. Dtsch. Chem. Ges.*, 1894, **27**, 2985.
11. Lehn, J.-M., *Supramolecular Chemistry. Concepts and Perspectives*, Verlag Chemie, Weinheim, 1995.
12. Desiraju, G. R., *Nature*, 2001, **408**, 407.
13. Behr, J.-P. (ed.), *The Lock and Key Principle*, Perspectives in Supramolecular Chemistry, Wiley, Chichester, 1994, vol. 1.
14. Stoddart, J. F. (ed.), Special issue of *Acc. Chem. Res.* on molecular machines, 2001, **34(6)**.
15. Dunitz, J. D. and Scheraga, H. A., *Proc. Natl. Acad. Sci.*, 2004, **101**, 14309.
16. *TIME* magazine Special issue on cancer, 28 May 2001.
17. Bernstein, J., *Polymorphism in Molecular Crystals*, Clarendon, Oxford, 2002.

18. Bhalla, U. S. and Iyengar, R., *Science*, 1999, **283**, 381.
19. Thallapally, P. K. *et al.*, *Angew. Chem., Int. Ed. Engl.*, 2004, **43**, 1149.
20. Rothman, T. and Sudarshan, G. (eds), *Doubt and Certainty*, Scientia, New Delhi, 2001, pp. 126–152.
21. Desiraju, G. R., *Nature Mater.*, 2002, **1**, 77.
22. Mill, J. S., *System of Logic*, Longmans, Green, Reader and Dyer, London, 1843.
23. Sambamoorthy, P., *Karnataka Sangeetham* (in Tamil), 3 volumes, The Indian Music Publishing House, Chennai, 1980.
24. Ayyangar, R. R., *History of South Indian (Carnatic) Music from Vedic Times to the Present*, Aryabhushan Press, Pune, 1972.
25. Desiraju, G. R., *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 2311.
26. Dunitz, J. D., *Chem. Commun.*, 2003, 545.
27. Jeffrey, G. A., *An Introduction to Hydrogen Bonding*, OUP, Oxford, 1997.
28. Desiraju, G. R. and Steiner, T., *The Weak Hydrogen Bond in Structural Chemistry and Biology*, OUP, Oxford, 1999.
29. Gilli, G. and Gilli, P., *J. Mol. Struct.*, 2000, **552**, 1.
30. Bhattacharyya, K., *Acc. Chem. Res.*, 2003, **36**, 95.
31. Selected recent references include: Moorthy, J. N., Natarajan, R. and Venugopalan, P., *Angew. Chem., Int. Ed. Engl.*, 2002, **41**, 3417; Infantes, L., Chisholm, J. and Motherwell, S., *Cryst. Eng. Commun.*, 2003, 480; Sarkhel, S. and Desiraju, G. R., *Proteins*, 2004, **54**, 247; Ball, P., *Nature*, 2004, **427**, 19; Jayaram, B. and Jain, T., *Annu. Rev. Biophys. Biomol. Struct.*, 2004, **33**, 343; Mukherjee, A., Saha, M. K., Nethaji, M., Chakravarty, A. R., *Chem. Commun.*, 2004, 716. See also, Biradha, K., Edwards, R. E., Foulds, G. J., Robinson, W. T. and Desiraju, G. R., *J. Chem. Soc., Chem. Commun.*, 1995, 1705; Gerstein, M. and Chothia, C., *Proc. Natl. Acad. Sci.*, 1996, **93**, 10167; Yu, B., Blaber, M., Gronenborn, A. M., Clore, G. M. and Caspar, D. L. D., *Proc. Natl. Acad. Sci.*, 1999, **96**, 103.
32. Metrangolo, P. and Resnati, G., *Chem. Eur. J.*, 2001, **7**, 2511.
33. Gladysz, J. and Curran, D. P., *Tetrahedron*, 2002, **38**, 3823.
34. Selected recent references include: Hof, F. and Diederich, F., *Chem. Commun.*, 2004, 484; Enright, G. D., Udachin, K. A. and Ripmeester, J. A., *Chem. Commun.*, 2004, 1360; Choudhury, A. R. and Row, T. N. G., *Crystallogr. Growth Des.*, 2004, **4**, 47.
35. Nicolaou, K. C., Vourloumis, D., Winssinger, N. and Baran, P. S., *Angew. Chem., Int. Ed. Engl.*, 2000, **39**, 44.
36. Rücker, C., Rücker, G. and Bertz, S. H., *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 378.
37. Nanjundiah, V., in *Science Studies* (ed. Sarukkai, S.), Economic and Political Wkly, 30 August 2003, pp. 3671–3677.
38. Whitesides, G. M., *Angew. Chem., Int. Ed. Engl.*, 2004, **43**, 3632.
39. Hernandez, R., Tseng, H.-R., Wong, J. W., Stoddart, J. F. and Zink, J. I., *J. Am. Chem. Soc.*, 2004, **126**, 3370.
40. Reinhoudt, D. N. (ed.), *Supramolecular Materials and Technologies*, Perspectives in Supramolecular Chemistry, Wiley, Chichester, 1999, vol. 4.
41. Lee, D. H., Severin, K. and Ghadiri, M. R., *Curr. Opin. Chem. Biol.*, 1997, **1**, 491.
42. Steed, J. W. and Atwood, J. L., *Supramolecular Chemistry. A Concise Introduction*, Wiley, Chichester, 2000.
43. Atwood, J. L. and Steed, J. (eds), *Encyclopedia of Supramolecular Chemistry*, Marcel Dekker, New York, 2004.

ACKNOWLEDGEMENTS. I thank the Institute of Advanced Study, University of Bologna (Prof. R. Scazzieri, Prof. D. Braga) for the award of a Senior Visiting Fellowship during which tenure this article took shape.

The natural effectiveness of mathematics in the biological sciences

R. Ramaswamy

School of Physical Sciences, Jawaharlal Nehru University, New Delhi 110 067, India

An increasingly quantitative approach within the biological sciences has been accompanied by a greater degree of mathematical sophistication. However, there is a need for new paradigms within which to treat an array of biological phenomena such as life, development, evolution or cognition. Topics such as game theory, chaos theory and complexity studies are now commonly used in biology, if not yet as analytic tools, as frameworks within which some biological processes can be understood. In addition, there have been great advances in unravelling the mechanism of biological processes from the fundamental cellular level upwards that have also required the input of very advanced methods of mathematical analysis. These range from the combinatorics needed in genome sequencing, to the complex transforms needed for image reconstruction in tomography. In this article, I discuss some of these applications, and also whether there is any framework other than mathematics within which the human mind can comprehend natural phenomena.

IT is a commonplace that in recent years the biological sciences have gradually become more quantitative. Far from being the last refuge of the nonmathematical but scientifically inclined, the modern biological sciences require familiarity with a barrage of sophisticated mathematical[1] and statistical[2] techniques.

By now the role of statistics in biology is traditional, and has been historically derived from the need to systematize a large body of variable data. The relation has been two-sided: biological systems have provided a wealth of information for statisticians and have driven the development of many measures, particularly for determining significance, as in the c^2 or Student's-*t* tests. Indeed, Galton's biometrical laboratory was instrumental in collecting and tabulating a plethora of biological measurements, and these and similar data formed the testing ground for a number of statistical theories.

The role of mathematics in biology is more recent. The phenomenal developments in experimental techniques that have helped to make biology more quantitative have necessitated the applications of a number of different mathematical tools. There have been unexpected and frequently serendipitious applications of techniques developed earlier and in a different context. The widespread use of dynamic programming techniques in computational biology, of stochastic context-free grammars in RNA folding, hidden Markov models for biological feature recognition in DNA sequence analysis, or the theory of games for evolutionary studies are some instances of existing methods finding new arenas for their application. There have also been the mathematics and the mathematical techniques that derive inspiration from biology. The logistic mapping, the discrete dynamical system that is so central to chaos theory, arose first in a model of population dynamics. Attempts to model the human mind have led to the burgeoning field of artificial neural networks, while the theory of evolution finds a direct application in the genetic algorithm for optimization.

Mathematics is about identifying patterns and learning from them. Much of biology is still most easily described as phenomena. The underlying patterns that appear are nebulous, so extracting a set of rules or laws from the huge body of observations has not always been easy. Or always possible since some experiments (like evolution) are unrepeatable, and separating the essential from the inessential can be very difficult. Detail is somewhat more important in the life sciences: often it has been said that the only law in biology is that to every 'law', there is an exception. This makes generalizations difficult: biological systems are more like unhappy families[3]. With the exception of natural selection, there are no clearly established universal laws in biology.

This is, of course, in sharp contrast to the more quantitative physical sciences where the unreasonable effectiveness of mathematics has often been commented upon[4,5]. It might be held that these observations, coming as they do in the twentieth century, comment on a science that has already had about three centuries of development. The earlier stages of the fields that we now call physics or chemistry were also very poorly described by mathematics—there was no general picture beyond a set of apparently unrelated observations, and it required the genius of a Mendeleev, of a Faraday or Maxwell or Einstein to identify the underlying patterns and expose the mathematical structure that lay under some aspects of these fields. This structure made much of the modern physical sciences possible, and led to some of the most accurate verifications of the laws of physics. As predictive theories, relativity and quantum electrodynamics are unparalleled and have achieved astonishing accuracy. In a more complex setting, the seemingly infinite

e-mail: r.ramaswamy@mail.jnu.ac.in

possibilities of organic chemical reactions have found organizational structure in the Woodward–Hoffman rules that combine an elementary quantum mechanics with notions of graph theory to make precise, semiquantitative predictions of the outcome of a large class of chemical reactions. What will it take to similarly systematize biology? Or to rephrase the question, what will the analogous grand theories in biology be?

The *inevitable* applications of mathematics are those that are a carry over from the more quantitative physical sciences. As in the other natural sciences, more refined experiments have spearheaded some of these changes. The ability to probe phenomena at finer and finer scales reduces some aspects of biology to chemistry and physics, which makes it necessary to borrow the mathematics that applies there, often without modification. For instance, tomographic techniques rely on a complicated set of mathematical transforms for image reconstruction. These may be largely unknown to the working biologist who uses NMR imaging, but are a crucial component of the methodology, nonetheless. Similarly, the genome revolution was catalysed by the shotgun sequencing strategy which itself relied on sophisticated mathematics and probability theory to ensure that it would work. Several of the problems in computational biology arose (or at least were made more immediate, and their resolution more pressing) by the very rapid increase in experimental power.

The other sort of application of mathematics is, for want of a better descriptor, *systemic*. Namely that which is not predicated by the reductionist approach to biology but instead by a need to describe the behaviour of a biological entity *in toto*.

Even the simplest living organisms appear to be complex, in way that is currently poorly described and poorly understood, and much as one would like, it is not possible to describe in all totality the behaviour of a living organism in the same way as one can the behaviour of, say, a complex material[6]. The promise that there could be mathematical models that capture the essence of this complexity has been held out in the past few decades by several developments, including that of inexpensive computational power which has made possible the study of more realistic models of biological systems. Theoretical developments—cellular automata, chaos theory, neural networks, self-organization—have provided simple mathematical models that seem to capture one or the other aspect of what we understand as ‘complexity’, which itself is an imprecise term. There is one class of applications of mathematical or physical models to biology which attempt to adapt an existing technique to a problem, while another aims to develop the methods that a given problem needs. Each of these approaches have their own value and appeal. In the next sections of this article, I discuss some of the ways in which they have found application in the study of biological systems.

The resonance of the title with those of the well-known essays by Wigner[4] and Hamming[5] is deliberate, as is the dissonance. There are applications in the physical sciences where knowledge of the underlying mathematics can provide very accurate predictions. Comparable situations in the biological sciences may not arise, in part because it may be unnecessary, and in part because biological systems are inherently unpredictable[6] since they are so fundamentally complex. The demands, as it were, that are made of mathematics in the life and physical sciences are very distinct, and therefore, it is very reasonable that the mathematics that finds application in the two areas can also be very different.

Is there any framework other than mathematics within which we can systematize *any* knowledge? Recent advances in cognitive studies, as well as information that is now coming from the analysis of genomes and genes, suggest that several aspects of human behaviour is instinctual (or ‘hardwired’). That mathematical reasoning is an instinct that we are endowed with is a distinct possibility, and therefore, it may not be given to us (as a species) to comprehend our world in any other manner. This point of view, that it is very natural that we should use mathematics to understand any science, is explored in the final sections of this article.

Complexity

In the last few years there has been a veritable explosion in the discipline, if it can be termed such, of complex systems studies. The concept of complexity is itself poorly defined (‘the more complex something is, the more you can talk about it’[7]), and as pointed out by Vicsek[8], ‘If a concept is not well-defined, it can be abused.’ Nevertheless, there is some unity in what studies of complexity aim to uncover[9].

A common feature of many complex systems is that they are composed of many interconnected and interacting subunits. Many systems, natural as well as constructed, are, in this sense complex. Examples that are frequently cited apart from those involving living organisms such as ecologies or societies, are the human brain, turbulent flows, market economies or the traffic. A second feature of complex systems is that they are capable of adaptation and organization, and these properties are a consequence of the interconnection and interactions of the subunits. The mathematics of complex systems would thus appear a natural candidate for application to biology. The drawback is that there is, at present, no unifying framework for the study of complex systems although there are some promising leads offered by studies of dynamical systems, cellular automata and random networks.

That the description of phenomena at one level may be inadequate or irrelevant at another has been noted for a long time. Thus the electronic structure of atoms can be understood quite adequately without reference to quarks, and is itself irrelevant, for the most part, when dealing with the thermodynamics of the material of which the atoms are constituents. Schrödinger, in a chapter of his very influential book[10] entitled ‘ *Is Life Based on the Laws of Physics*’, observed that with regard to ‘the structure of living matter, that we must be prepared to finding it working in a manner that cannot be reduced to the ordinary laws of physics’. He further contrasted the laws of physics and chemistry, most of which apply in a statistical sense, to biological phenomena, which, even though they involve large numbers of atoms and molecules, nevertheless have nothing of the uncertainty associated with individual properties of the constituent atoms. Indeed, given a radioactive atom, he says, ‘its probable lifetime is much less certain than that of a healthy sparrow’.

But even at a given level, it frequently happens that the properties of a system cannot be simply inferred from those of its constituents. The feature of *emergence*, namely the existence of properties that are characteristic of the entire system but which are not those of the units, is a common feature of systems that are termed complex.

Distinction should be drawn between the complex and the complicated, though this boundary is itself poorly defined. For instance, it is not clear whether or not in order to be deemed complex, a system requires an involved algorithm (or set of instructions). The algorithmic complexity, defined in terms of the length of the (abstract) program that is required is of limited utility in characterizing most systems[11].

Attempts to decode the principles that govern the manner in which new properties emerge—for example the creation of a thought or an idea, from the firing of millions of neurons in the brain, or the cause of a crash in the stock market from the exit poll predictions in distant electoral constituencies—require new approaches. The principles themselves need not necessarily be profound. A simple example of this is provided by a study of flocking behaviour in bird flight[12]. A purely ‘local’ rule: each bird adopts the average direction and speed of all its neighbours within distance R, say, is enough to ensure that an entire group adopts a common velocity and moves in unison. This behaviour depends on the density of birds as well as the size of R relative to the size of a bird in flight. If R is the size of a bird, then each bird flies on its own path, regardless of its neighbours: there is no flock. However, as R increases to a few times that of the bird, depending on the density, there can be a phase transition, an abrupt change from a random state to one of ordered, coherent, flight[13]. And such a system can adapt rapidly: we have all seen flocks navigate effortlessly through cities, avoiding tall buildings, and weaving their way through the urban landscape at high speed.

But there are other aspects of complexity. A (western) orchestra, for instance, consisting as it does of several musicians, requires an elaborate set of rules so that the output is the music that the composer intended: a set of music sheets with the detailed score, a proper setting wherein the orchestra can perform, a specific placement of the different musical instruments, and above all, strict obedience to the conductor who controls what is played and when. To term this a complex system would not surprise anyone, but there is a sense in which such a system is not: it cannot *adapt*. Should the audience demand another piece of music, or music of another genre, an orchestra which has not prepared for it would be helpless and could not perform. Although the procedure for creating the orchestra is undoubtedly complicated, the result is tuned to a single output (or limited set of outputs). There is, of course, emergence: a single tuba could hardly carry a tune, but in concert, the entire orchestra creates the symphony.

Models like this illustrate some of the features that complex systems studies aim to capture: adaptability, emergence and self-organization, all from a set of elementary rules. The emphasis on elementary is deliberate. Most phenomena we see as complex have no obvious underlying conductor, no watchmaker, blind or not[14] who has implemented this as part of a grand design. Therefore, in the past few decades, considerable effort has gone into understanding ‘simple’ systems that give rise to complex behaviour.

‘Simple mathematical models with very complicated dynamics’, a review article published[15] in 1976 was responsible in great measure for the phenomenal growth in the study of chaotic dynamics. In this article—which remains one of the most accessible introductions to chaos theory—May showed that the simplest nonlinear iterative dynamical systems could have orbits that were as unpredictable as a coin-toss experiment. Indeed, the logistic mapping,

$$x_{n+1} = a\, x_n(1 - x_n), \qquad (1)$$

where x_n is the value of the dependent variable x at time n with $a = 4$ was used as the first pseudo-random number generator by von Neumann and Ulam[15]. The quadratic nonlinearity in the logistic map is the simplest imaginable, and the unpredictable nature of the solutions, as a function of the parameter a have formed the focus of innumerable studies since then.

When studying dynamical systems, some of the features of interest are the nature of the motion as a function of the time, namely the dependence of x_n on n, and perhaps the long time behaviour, namely what happens as $n \to \infty$, or how does the motion depend on initial conditions x_0. The important feature of chaotic dynamics is *sensitivity to initial conditions*. When the equations of motion are linear, a change in x_0 leads to a comparable change in x_n, while for nonlinear systems, the smallest change in initial position will inevitably lead to an exponentially large change in x_n, and thus give dynamics which is completely distinct. The logistic map, eq. (1), has been a paradigm system for such studies, since upon varying a, different behaviour results: for a between 1 and 3, the motion eventually comes to rest on a fixed point ($\lim_{n\to\infty} x_n \to 1 - 1/a$ for almost any initial x_0), while for other a, one can have periodic motion of different periodicities, and also (mathematically provable!) chaotic motion, for $a = 4$.

Considerable theoretical and experimental (numerical) work in the 1970s and ’80s established that chaotic systems come in so-called universality classes (an example of Tolstoy’s happy families) which are characterized by the same qualitative and quantitative features. Regardless of the details of a given system, within each class the manner in which chaotic dynamics is created as well as its characteristics would be the same. One such route is period-doubling: as a parameter (such as a) is varied, the dynamics changes abruptly at a specific (bifurcation) value from a given periodicity to double the period. This happens repeatedly, and the successive bifurcation values of the parameter are geometrically spaced, with spacing dependent on the nature of the nonlinearity. The veracity of this picture was rapidly established in experiments as diverse as Rayleigh–Bénard flow, the oscillatory Belusov–Zhabotinsky reaction, cellular rhythms, mechanical systems, forced pendulums, dripping faucets . . . Not all aperiodic dynamics is created via period-doubling; there are other known routes to chaotic dynamics, but all of them share the feature that they apply in a specific set of circumstances, and that they are sufficiently general[16].

The gradual evolution of regular, nonchaotic, dynamics-to-chaotic motion as a parameter is varied can be likened to a phase transition. And like phase transitions, the most interesting behaviour occurs at the transition point itself, which here is 'at the edge of chaos', between the regular and the chaotic[17].

For complexity, one neither wants regular dynamics which is repetitive (and dull), nor chaotic dynamics which is sensitive to small changes (and hence too unstable). When the motion is neither truly chaotic nor truly regular, it is capable of novel behaviour which involve correlations that are temporally long-ranged. Studies of cellular automata, dynamical systems where both time and space are discrete rather than continuous, have demonstrated that systems which are at this edge can show very complex patterns[18]. The idea that there can be dynamical models that mimic life go back to von Neumann who had showed that there could be cellular automata that were self-reproducing. More recent work in the area of artificial life[9] has shown that there are simple cellular automaton-like models that can capture a variety of features of living systems; these systems are on the edge of chaos.

To summarize, the thrust of much work in the past few decades has been to establish that complex temporal behaviour can result from simple nonlinear dynamical models. Likewise, complex spatial organization can result from relatively simple sets of local rules. Taken together, this would suggest that it might be possible to obtain relatively simple mathematical models that can capture the complex spatio-temporal behaviour of biological systems. In small measure such programs are slowly becoming possible.

Application and applicability

With the large amount of genomic data that has rapidly become available through sequencing projects, the time and effort required for biochemical or biological analysis of genomic function is neither available nor does the endeavour appear to be cost-effective. Isolating and studying a single protein can be arduous, and the smallest genomes, of which today there are over 200 already sequenced completely, have several hundred genes and a similar number of proteins. Characterizing even one genome is therefore a mammoth undertaking. One can make the case for studying the human genome in all its detail, given its importance to our species, but the case for other genomes is less compelling. Much of the analysis of DNA sequences (and to some extent, the other important macromolecules such as RNA and proteins) has now become computational and mathematical studies of symbolic strings.

The identification of a similarity or a pattern is a common way in which biological function is inferred, even if erroneously. That two things that appear the same probably have a similar function is a mode of reasoning which we frequently employ, based on the experience that this is often true. (Of course, it is also an experience that different things can be used for the same biological function, so avenues for learning must be kept open!) Comparing two biological sequences has become the major technique for learning whether or not there is some possibility that the two sequences have a similar function and/or a common evolutionary history.

The nature of the demands made of mathematical tools in a biological context is best typified in this area of sequence comparison and alignment. Briefly, the problem is as follows. It is required that two (or more) strings, of lengths n and m, say, composed of symbols in a finite alphabet (4 letter alphabets for DNA or RNA molecules, 20 letter alphabet for proteins), are to be aligned (written one below the other, say) to display the similarity between the strings, keeping the order of symbols in each string unchanged, but introducing additional symbols such as '–' or gaps in either string as required. A placement of symbol a_i over b_j is an indication that the two are similar, and effort is made to get the optimal alignment, namely that alignment that displays the most similarities between the strings. This similarity is quantified via a 'score', which is itself based on the statistics of presently available information. The total number of possible alignments is of the order of the binomial coefficient ${}^{n+m}C_n$, so that enumeration is not possible for any but the shortest strings, and for modest biologically interesting problems, n and $m > 1000$. An exact method can give the best alignment, but the need to scan large databases (where m exceeds several billion today) necessitates heuristics. Judging the significance of the alignment score requires an appreciation of the statistics of extreme values. The final analysis, therefore, is based on a combination of mathematics, statistics, probability, and, at present, a set of biological facts and intuition.

Useful as they are—the single most widely used mathematical tool in biology today is the BLAST set of programs for sequence alignment—the contrast with the typical use of mathematics in the physical sciences could not be more striking.

Some of the newer applications of mathematics in biology which include the use of graph theory in its modern manifestations to analyse networks and clustering methods to classify the huge amount of data generated in microarray experiments, in addition to a host of algorithms and techniques for the analysis of genomic information, derived from computer science. Other branches of mathematics that also find biological application are game theory, in understanding genetic competition in ecosystems and even within genomes, and dynamical systems theory in understanding the complex temporal behaviour of the huge number of coupled chemical reactions consequent upon the genetic network of entire organisms.

The mathematical theory of games, developed from the 1920s onward, deals with models of conflict and cooperation among interacting interdependent agents. Initial applications were to sociological problems, but quickly game theory found a natural context in evolutionary biology, and today has been extensively applied to problems ranging from why the sex-ratios are as observed in different populations and different times, to the evolution of language itself. In a recent application, the 'prisoner's dilemma'[19] has been *experimentally* studied in the context of viral evolution, while game-theoretic analyses of ecosystems can permit very sophisticated inference regarding species fitness and evolutionary strategies[20].

A number of recent ambitious programs intend to study cellular dynamics, metabolism and pathways in totality entirely *in silico*[21]. Since the elementary biochemical processes are, by and large, well-understood from a chemical kinetics viewpoint, and in some cases the details of metabolic pathways have also been explored, entire genomes have been sequenced and the genes are known, at least for simple organisms, the attempt is to integrate all this information to have a working computational model of a cell. By including ideas from network theory and chemical kinetics, the global organization of the metabolic pathway in *E. coli* has been studied computationally[22]. This required the analysis of 739 chemical reactions involving 537 metabolites and was possible for so well-studied an organism, and the model was also able to make predictions that could be experimentally tested. The sheer size of the dynamical system is indicative of the type of complexity that even the simplest biological organisms possess; that it is even possible for us to contemplate and carry out studies of this magnitude is indicative of the analytic tools that we are in a position to deploy to understand this complexity. More powerful computational methods could be necessary and will become possible in the future.

The biology of mathematics

In recent years, there has been considerable debate, and an emerging viewpoint, that the human species has an *instinct* for language. Champions of this school of thought are Chomsky, and most notably, Steven Pinker who has written extensively and accessibly on the issue[23].

The argument is elaborate but compelling. It is difficult to summarize the entire line of reasoning that was presented in *The Language Instinct*[23], but one of the key features is that language is not a cultural invention of our species (like democracy, say), but is hard-wired into our genome. Like the elephant's trunk or the giraffe's neck, language is a biological adaptation to communicate information and is unique to our species.

Humans are born endowed with the ability for language, and this ability enables us to learn any specific language, or indeed to create one if needed. Starting with the work of Chomsky in the 1950s, linguists and cognitive scientists have done much to understand the universal mental grammar that we all possess. (The use of stochastic context-free grammars in addressing the problem of RNA folding is one instance of the remarkable applicability of mathematics in biology.) At the same time, however, our thought processes are not language dependent: we do not think in English or Tamil or Hindi, but in some separate and distinct language of thought termed'mentalese'[23].

Language facilitates (and greatly enriches) communication between humans. Many other species do have sophisticated communication abilities—dolphins use sonar, bees dance to guide their hivemates to nectar sources, all birds and animals call to alarm and to attract, ants use pheromones to keep their nestmates in line, etc.—and all species need to have some communication between individuals, at least for propagation. However, none of these alternate instances matches anything like the communication provided by human language.

It is not easy to separate nature from nurture, as endless debates have confirmed, but one method for determining whether or not some aspect of human behaviour is innate is to study cultures that are widely spaced geographically, and at different stages of social development. Such cross-cultural studies can help to identify those aspects of our behaviour that are a consequence of environment, and those that are a consequence of heredity. The anthropologist Donald Brown has attempted to identify human 'universals'[24], a set of behavioural traits that are common to *all* tribes on the planet.

All of us share several traits beyond possessing language. As a species we have innumerable taboos relating to sex. Some of these, like incest avoidance, appear as innate genetic wisdom, but there are other common traits that are more surprising. Every culture, from the Inuit to the Jarawa, indulges in baby talk. And everybody dreams. Every tribe however 'primitive', has a sense of metaphor, a sense of time, and a world view. Language is only one (although perhaps the most striking) of human universals. Other universals that appear on the extensive list in his book, and which are more germane to the argument I make below, are conjectural reasoning, ordering as cognitive pattern (continua), logical notions, numerals (counting; at least 'one', 'two' and 'many') and interpolation.

The last few mentioned human universals all relate to a set of essentially mathematical abilities. The basic nature of enumeration, of counting, of having a sense of numbers is central to a sense of mathematics and brings to mind Kronecker's assertion, 'God made the integers, all else is the work of man'[25]. The ability to interpolate, to have a sense of a continuum (more on this below), also contribute to a sense of mathematics, and lead to the question: Analogous to language, do humans possess a *mathematics instinct?*

Writing a century ago, Poincaré had an inkling that this might be the case. '…we possess the capacity to construct a physical and mathematical continuum; and this capacity exists in us *before any experience* because, without it, experience properly speaking would be impossible and would be reduced to brute sensations, unsuitable for any *organization*; …' The added emphasis is mine; the observations are from the concluding paragraph of his essay, 'Why space has three dimensions'[26].

If mathematics is an instinct, then it could have evolved like any other trait. Indeed, it could have co-evolved with language, and that is an argument that Keith Devlin has made recently[27].

At some level, mathematics is about finding patterns and generalizing them and about perceiving structures and extending them. Devlin suggests that the ability for mathematics resides in our ability for language. Similar abstractions are necessary in both contexts. The concept of the number three, for example, is unrelated either to the written or spoken word three, or the symbol 3 or even the more suggestive alternate, III. Mathematical thought proceeds in its version of mentalese[27].

An innate mathematical sense need not translate into universal mathematical sophistication, just as an innate language sense does not translate into universal poetic ability. But the thesis that we have it in the genes begs the question of whether mathematical ability confers evolutionary advantage, namely, is the human race selected by a sense for mathematics?

To know the answer to this requires more information and knowledge than we have at present. Our understanding of what constitutes human nature in all its complexity is at the most basic level. The sociobiologist E. O. Wilson has been at the vanguard of a multidisciplinary effort[28] toward consilience, gathering a coherent and holistic view of current knowledge which is not subdivided in subdisciplinary approaches. This may eventually be one of the grand theories in biology, but its resolution is well in the future. We need to learn more about ourselves.

Summary

Traditionally, any sense of understanding physical phenomena has been based on having the requisite mathematical substructure, and this tradition traces backward from the present, via Einstein, Maxwell and Newton, to Archimedes and surely beyond.

Such practice has not, in large measure, been the case in biology. The view that I have put forth above ascribes this in part to the stage of development that the discipline finds itself in at this point in time, and in part, to the manner in which biological knowledge integrates mathematical analysis. The complexity of most biological systems, the competing effects that give rise to organization, and the dynamical instabilities that underlie essentially all processes make the system fundamentally unpredictable, all require that the role played by mathematics in the biological sciences is of necessity very different from that in the physical sciences.

Serendipity can only occasionally provide a ready-made solution to an existing problem whereby one or the other already developed mathematical method can find application in biology. Just as, for example, the research of Poincaré in the area of dynamical systems gave birth to topology, the study of complex biological systems may require the creation of new mathematical tools, techniques, and possibly new disciplines.

Our instincts for language and mathematics, consequences of our particular evolutionary history, are unique endowments. While they have greatly facilitated human development, it is also worth considering that there are modes of thought that may be denied to us, as Hamming has observed[5], similar to our inability to perceive some wavelengths of light or to taste certain flavours. 'Evolution, so far, may possibly have blocked us from being able to think in some directions; there could be unthinkable thoughts.' In this sense, it is impossible for us to think nonmathematically, and therefore there is no framework other than mathematics that can confer us with a sense of understanding of any area of inquiry.

In biology, as Dobzhansky's famous statement goes, nothing makes sense except in the light of evolution. To adapt this aphorism, even in biology nothing can really make sense to us except in the light of mathematics. The required mathematics, though, may not all be uncovered yet.

1. Maynard Smith, J., *Mathematical Ideas in Biology*, Cambridge University Press, Cambridge, 1968.
2. Batschelet, E., *Circular Statistics in Biology*, Academic Press, New York, 1981.
3. All happy families resemble one another, each unhappy family is unhappy in its own way, Tolstoy, L., the opening words of *Anna Karenina*.
4. Wigner, E. P., The unreasonable effectiveness of mathematics in the natural sciences. *Commun. Pure Appl. Math*., 1960, **13**.
5. Hamming, R. W., The unreasonable effectiveness of mathematics, *Am. Math. Monthly*, 1980, **87**.
6. Ramaswamy, R., Genes, brains, and unpredictability: Developments in the sciences and reflections on what it means to be alive. *Curr. Sci.*, 2001, **80**, 1381.
7. Parisi, G., attributed by M. Virasoro, TIFR lecture, 1996.
8. Vicsek, T., Complexity: The bigger picture. *Nature*, 2002, **418**, 131.
9. Waldrop, M. M., *Complexity*, Simon and Schuster, New York, 1992.
10. Schrödinger, *What is Life?*, Cambridge University Press, Cambridge, 1967.
11. The infinite string 01010101010… has a lower algorithmic complexity than the string 0111001010110010101001111… since the program for the first string is simple (repeat '01' endlessly) while

that for the second, where there is no discernable pattern, is of the order of the length of the string (write 0, write 1, etc.). It is not easy to devise a measure of algorithmic complexity which is useful in most biological situations; see Badii, R. and Politi, A., *Complexity*, Cambridge University Press, Cambridge, 2000.

12. Reynolds, C., http://www.red3d.com/cwr/boids/
13. Thus there need be no leader issuing the command 'Flock!'; depending on the species (and the range of vision, size, speed, etc.) there is an effective R (relative to bird-size) which determines flocking. As can be easily appreciated, very large R gives rise to very coherent flocks. Larger birds (smaller R) would also make poorer flockers.
14. Dawkins, R., *The Blind Watchmaker*, Norton, New York, 1996.
15. May, R. M., Simple mathematical models with very complicated dynamics. *Nature*, 1976, **261**, 459.
16. Serra, R., Andretta, M., Zanarini, G. and Compiani, M., *Introduction to the Physics of Complex Systems*, Pergamon Press, Oxford, 1986.
17. Langton, C., Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D*, 1990, **42**, 12.
18. Wolfram, S., *A New kind of Science*, Wolfram Media Inc, 2002.
19. This is a standard game—theoretic game; two prisoners are independently given a choice of either 'cooperating' or 'defecting', with the returns being dependent on what the other prisoner chooses. If a prisoner cooperates, his reward is a if the other prisoner also cooperates, and b if the other defects, while if he defects, his reward is c if the other cooperates, and d if the other also defects. This reward is usually represented in terms of a 'fitness': a cooperator's fitness is, say, 1 if the other also cooperates, $(1 - s)$ if the other defects, while a defector has 9 fitness $(1 + t)$ if the other cooperates, and $(1 - c)$ if the other also defects. When this game is played in a population, if $(1 - c)$ exceeds $(1 - s)$, defection can sweep through the population, even though it would have benefited all to cooperate since that would give, on average, greater fitness. When to choose cooperation over defection constitutes the prisoner's dilemma. See Turner, P. E. and Chao, L., *Nature*, 1999, **398**, 441.
20. Sinervo, B., Svensson, E. and Comendant, T., *Nature*, 2000, **406**, 985.
21. eCell, A multiple algorithm, multiple timescale simulation software environment, http://www.e-cell.org.
22. Almas, E., Kovács, B., Vicsek, T., Oltvai, Z. N. and Barabási, A.-L., Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, 2004, **427**, 839.
23. Pinker, S., *The Language Instinct*, Morrow, New York, 1994.
24. Brown, D., *Human Universals*, Temple University Press, Philadelphia, 1991.
25. There are various translations of the original quotation, 'Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk'.
26. Poincaré, H., *Dernières Pensées*, Flammarion, Paris, 1913; Translated by Bolduc, J. W., as *Mathematics and Science: Last Essays*, Dover, New York, 1963.
27. Devlin, K., *The Maths Gene*, Wiedenfeld and Nicolson, London, 2000.
28. Wilson, E. O., *Consilience: The Unity of Knowledge*, A. A. Knopf, New York, 1998.

ACKNOWLEDGEMENTS. My research in the area of computational biology has been supported by the Department of BioTechnology, and in the area of nonlinear science by the Department of Science and Technology.

Mathematics and biology*

Vidyanand Nanjundiah

Indian Institute of Science and Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore 560 012, India

From the well-known saying, 'All chemistry is physics and all biology is chemistry', one might be tempted to conclude that because physical laws are most compactly expressed in the language of mathematics, biology too can, or should, be 'mathematized'. On the other hand, in popular perception, mathematics and biology are as far apart as is possible for any two fields of enquiry to be. Mathematics is believed to be analytical and deductive whereas biology is thought of as largely descriptive and based on induction. Admittedly, this oversimplifies a complex issue. Nevertheless, the fact stands out, that on the whole the practice of biology – as reflected in teaching programs and research publications – is free of mathematical language and symbols to an extent that would be unthinkable in contemporary physics or chemistry. The reasons for this are many. An important one is that biological systems confirm to the laws of physics but are not *necessary* consequences of the laws. Mathematics can illuminate aspects of the structure or function of a living system and can also be used to make sense of the statistical consequences of dealing with large aggregates of genes or cells or organisms. All the same, it is difficult to conceive of a living creature, taken as a whole, as being reducible to mathematical formalisms such as those that embody the laws of physics. The difficulty stems from the manner in which evolution has shaped the history of life on earth. Historical contingencies, the random nature of mutations and evolutionary opportunism all make it difficult to encompass biology within a mathematical framework.

IN a famous article, the physicist Eugene Wigner expanded on what he called 'The unreasonable effectiveness of mathematics in the natural sciences'. According to Wigner, it was extraordinary that mathematics – a set of formal rules for manipulating abstract symbols and a creation of the human mind – could explain the messy and complicated real world[1]. Others have drawn attention to the astonishing fact that a simple algebraic equation that a schoolchild can understand, the inverse-square law of gravitation, holds within it the key to 'the motions of the planets, the comets, the moon, and the sea' (Newton, from the first edition of the *Principia*). The present essay aims to make the point that the term 'natural sciences' in Wigner's title is inappropriate in the context of biology, almost certainly so in the case of biological entities that are at least as complex as a cell[2].

Reductionism and determinism

In what follows, by *reduction* I mean explaining something in terms something else, the 'something else' usually consisting of entities at a lower level of organization than what is sought to be explained. An outcome is said to be *determined* if, given a certain initial condition, it is inevitable. Inevitability implies that it is the consequence of a 'law', rule, or set of rules. A rule may be expressed in ordinary language or it may be put tersely in the form of a mathematical expression. Roughly speaking, reductionism is equivalent to the assertion that the whole is explainable in terms of its parts, and determinism is equivalent to saying that a specified initial condition leads unambiguously to a particular end result. A system that is determined (in this sense) is often, but not always, predictable. Chaotic systems are well-known exceptions: in their case, seemingly insignificant differences in initial conditions can lead to major differences in the outcome. In principle, reductionism is always valid. If an object is composed of structurally or functionally distinct parts, the object must be fully explicable in terms of the properties of those parts and their interactions. But reduction may or may not be useful; consider for example the case of a painting.

One can think of three roles for mathematics in biology. The first is an extension of its role in physics or chemistry. It would follow as a consequence of the reduction of biology to either of those two sciences, were such a thing possible. The second role is a modern one. It originates in the expectation, fuelled originally by the success of Mendel's principles and more recently by the discovery of the genetic code[3], that living creatures are essentially informational entities. The underlying hypothesis is that biological information – the information required for making a plant or an animal – is encoded in terms of a set of rules, sometimes referred to as a program. The program is said to determine the organism. A reliable set of procedures, an algorithm, leads one from the program to the organism. For example, the development of a multicellular organism from a fertilized egg has been compared to an algorithmic process. The third role for mathematics, including of course computational mathematics, is more conventional. Mathematics is an aid to organized thought and a bridge between empiri-

*This article is adapted from 'The role of mathematics in biology', (*Economic and Political Weekly*, XXXVIII (35), 3671–3677; Aug 30–Sep 05, 2003).

e-mail: vidya@ces.iisc.ernet.in

cal knowledge and theoretical models. Similarly, the computer is an invaluable tool when it comes to handling large masses of data or carrying out many intricate logical operations at great speed. The first two roles are the more interesting ones, and their validity depends on whether it is useful to think of biology in deterministic or reductionist terms. An old-style biologist (say a biologist of the 19th century or earlier) would have questioned both, most likely on the grounds that organisms exhibit emergent properties that cannot be reduced to the properties of their constituent parts. However, the biology of the latter half of the 20th century and since has been characterized by an approach, known as molecular biology, in which the themes of reductionism and determinism are deeply embedded.

Molecular biology

The announcement of the double helical structure of DNA set the seal on a notion that was already gaining acceptance, namely, that heredity had a chemical basis: Mendel's particulate units of inheritance (genes) turned out to be molecules of DNA[4]. Today we know that the proteins that characterize living cells are synthesized on the basis of certain rules of correspondence relating sequential subsets of DNA (base triplets) to sequential subsets of proteins (amino acids). In parallel, we have discovered that development, the transformation of an egg into an adult, is accompanied by changes in the patterns of activity of DNA. The patterns differ from one cell to another and are altered when the DNA undergoes heritable changes. Thus the same chemical, DNA, is central to both heredity and development.

The interesting thing about the chemistry involved in making proteins from DNA is that it depends solely on interpreting the information contained in DNA; the DNA itself remains unchanged. In consequence, a DNA sequence can be thought of as an encoded or symbolic representation of a particular protein. This realization unleashed a burst of research in biology and set in motion an approach to understanding living forms, still going strong, known as molecular biology. Fuelled by the rate at which new facts came to be accumulated – on the whole, facts that were entirely unexpected, the field acquired unprecedented prestige and an ethos of its very own[5]. At its heart, molecular biology was, and even now to some extent is, driven by an assumption and a hope. The assumption is that the (known) relationship between one DNA molecule and the protein it encodes offers a clue to the (unknown) relationship between all the DNA molecules in an organism, the genome, and the organism itself. The hope is that this assumption can be shown to be correct – in other words, that the genome can be demonstrated to be an implicit, symbolic representation of the organism. Because both computers and molecular biology began to take hold of peoples' imagination at about the same time, it became popular to compare the genome to a computer program written in chemical language. The program was believed to reside in sequences of DNA and the organism was thought of as the output of the program. Salvador Luria, one of the pioneers of molecular biology, has put it pithily: 'Like the computer, the programmed organism has a tape containing the instructions (genes, DNA) and the machinery to implement the instructions on the tape'[6].

In short, molecular biology contains within itself both reductionist and determinist aspirations. These are, respectively, that life can be explained in terms of its molecular constituents, and that the properties of an organism are deducible from its DNA. The reductionist hypothesis remains unquestioned in principle. But in practice, and this is true of reductionism in general, the hypothesis is not useful beyond a particular level of organization. Once one starts looking at an entity sufficiently removed from genes and proteins, say the cell (let alone the organism), new attributes may emerge and entirely new explanatory concepts may be called for. To take a simple example, water is made up of hydrogen and oxygen atoms and water 'finds its own level'. But it would be a waste of time to try and explain the latter observation in terms of the former. Similarly, it is not so much that a reductionist approach to biology is wrong. Rather, and especially if the object of interest is the whole organism, it provides little insight. What about the second possibility? Increasingly, the prospects for genetic determinism too look dubious[7]. In both cases the reason has to do with the nature of life, and in particular, with the role played by evolution in shaping living forms.

The nature of life

It is notoriously difficult to give a crisp definition of life. Taken together, however, living creatures exhibit features that stand out from those of non-living entities. One might summarize them by saying that being alive is a property possessed by temporary, open, organized forms of matter that can store and transmit information and evolve by natural selection. Briefly, what this means is as follows.

Living matter exists far from equilibrium

To sustain life, organisms must constantly renew themselves. Both organisms and their constituent parts can maintain themselves only temporarily. Cells are re-modelled so rapidly that hardly any of the molecular components of our physical body is more than a few weeks old (which makes it interesting that we retain a continuous and unbroken sense of ourselves)[8].

Even temporary upkeep requires an input of energy and information

This process, known as metabolism, involves the chemical transformation of the air we breathe and the food we eat.

Metabolism creates new structures in the body and repairs or discards old ones. Besides exchanging matter, the organism continually scans the environment and extracts useful information from it. Both literally and metaphorically, we feed on our surroundings.

Life involves organization at various levels in a hierarchy

To begin with there is molecular organization. DNA, RNA, proteins, lipids and sugars are gigantic molecules made up of anywhere from a hundred to a million atoms. At the next level, the cells that are built from them are organized into distinct compartments that subserve (in part) different functions. Organized networks of production, distribution and disposal coordinate metabolism. Cells form bodies made up of specialized tissues and are arranged in a recognizable fashion. Lastly, the reception, storage and retrieval of information, both between cells and between the body and the external world, exhibits an intricate organization.

Life continually evolves

Species change through a process whose underlying basis is haphazard and undirected. Despite the randomness that lies at the core of evolution, living forms give the impression of having been designed to fit an end. But the impression is misleading: order results from randomness because randomness is filtered by evolutionary opportunism[9,10].

The implication is that an organism is best viewed as a special-purpose device that has been shaped by a series of opportunistic responses to the conditions encountered by it in its evolutionary history. It is this ‘special- spect of organisms that leads to serious difficulties when we try to reduce biology to physics and chemistry.

Evolution and reductionism

Opportunism is inherent in the explanation of evolution by natural selection, first put forward by Darwin and Wallace almost 150 years ago. Natural selection begins with a random change, a mutation, in a DNA sequence. The change can spread through a population if, on the average, it leads to an improvement in the individual in whom it occurred. Here ‘improvement’ means one thing and one thing only. It means that the individual is more likely to have offspring (who inherit the changed DNA molecule) than the average individual in the population[11].

What one might call a unit evolutionary episode begins with a mutation, continues through the fate of the altered DNA sequence over generations and, on occasion, concludes with the establishment or elimination of the mutant gene. A mutant or variant version of a gene becomes established or ‘fixed’ if it replaces the pre-existing variety. The enormous number of independent mutations that can occur in the genome of a species within a single generation makes the occurrence of any one of them at a particular time unpredictable in advance. Besides, the likelihood of a mutation is independent of whether it leads to a particular outcome or not. A mutation can be judged to be beneficial or harmful only *post facto*. The judgement depends on the consequence of the mutation, which in turn depends on a host of circumstances. A mutation that happens to be successful builds on an earlier mutation that happened to be successful. The complex organism that results is put together via unforeseen steps and has not come about as the result of planning. Thus, the course of evolution is indeterminate and therefore unpredictable[12].

As if that were not bad enough, life on earth has been affected by catastrophic accidents that disrupted and re-set the course of natural selection. The accidents led to major changes in the composition of living forms. During each change the relative abundance of forms was drastically altered; a large number even went extinct. As a result, it is impossible to account for what happened after any of the catastrophic episodes as a logical extension of what was going on before. Taken together with the basic randomness that is inherent in evolutionary change, this means that nothing makes it *necessary* that a particular species of plant or animal should have existed in the past, or exists at present (including our own species, *Homo sapiens*).

To sum up, living creatures are entities in their own right and, as such, obey the principles of physics and chemistry. But because they have an evolutionary past, they are also products of history. They have been shaped by a series of random steps that resulted in successful responses to contingencies that were faced by their ancestors; therefore they are endowed with a substantial element of arbitrariness in the way they are put together. This makes it difficult, if not impossible, to think of them solely as products of physics and chemistry. Physics and chemistry set limits within which evolution can take place but cannot specify in advance either its course or its end.

Evolution and determinism

A cell contains encoded information within itself in the form of genes: every protein is the decoded version of some gene. Might not a living creature also be a decoded entity, the encoding being carried out by the genome as a whole? If the answer is yes, one could say that genes determine organisms. Then it would make sense to look for a theoretical underpinning of biology in the set of rules, the genetic program, hidden behind the organism. The programmatic view of development is that the genome has an algorithmic structure, that it is put together logically, as we would put together any set of instructions designed to ensure a desired end, for instance a cookbook recipe, a play or a computer program. However, many findings point to the lack of a rule-based, logical structure to the genome as a whole[7,13].

Algorithmic operations generally lead to unique outcomes, whereas the output of genomes can be highly flexible

The same genome, or very similar genomes, can be consistent with creatures that look drastically different. Sometimes this is because of environmental influences, sometimes because of differences in life stages and sometimes because the link between the genome and the organism has an intrinsically indeterminate component to it. Turtle and crocodile eggs develop into males or females depending on the temperature at which they are incubated; the caterpillar and the butterfly, in appearance so different, are genetically identical; bacteria that are members of the same clone and are raised in the same environment can differ in their ability to metabolize sugar.

Conversely, very different genotypes can be consistent with very similar organisms

South and North American mammals, or molluscan and vertebrate eyes, are textbook examples. Thus the relationship between genomes and organisms is not one-to-one. Chimpanzees and humans, different in so many respects, are said to share 98% of their DNA sequences. 'Sibling spe
frogs share a much smaller fraction of DNA but require an expert to tell apart.

The same gene or DNA sequence can contribute to seemingly unrelated traits

Genes that mediate sexual differentiation in the fly also function in the development of their nervous system. In albino cats, a gene that encodes an enzyme in the pathway responsible for the synthesis of a skin pigment, melanin, also affects the manner in which the eye and brain are connected.

In short, genomes are organized very differently from plays, recipes or computer programs. DNA sequences and their functioning reflect the opportunistic fashion in which organisms have evolved. If there are rules that lie behind the ways in which plants and animals are built, the rules are not logical consequences of the ways in which genes are copied or gene activity is regulated (no more, for instance, than the functioning of human societies is a logical consequence of human biology).

The role of mathematics in biology

Undoubtedly it should be possible, and indeed is possible, to treat a living system as one does any other physical system, so long as one does so within a suitably restricted framework. The narrower the focus, the likelier it is that a physical or chemical, and therefore mathematical, approach will be useful[14]. Many aspects of the functioning of cells and tissues can be subject to mathematical treatment quite successfully. Metabolic networks can be modelled as linked cascades of chemical reactions, signalling in the nervous system is similar to electrical conduction through a network of wires and the flow of blood depends on hydrodynamic principles. Population genetics, a highly mathematical branch of biology, deals with the spread and distribution of genes in populations. Population geneticists aim (among other things) to show that measurable changes in the distribution are consistent with known or assumed evolutionary forces. None of these cases constitute counter-examples to the point I have been trying to make. As an illustration of precisely why not, let me take the case of something that has long been a popular subject among mathematically minded biologists.

Fifty years ago, in a publication that has been accorded the status of a classic, the logician and computer scientist Alan Turing constructed a mathematical model for the spontaneous origin of biological form. He began with what he thought was a formless structure such as a newly fertilized egg[15]. Turing's model was 'global': he looked at the embryo as a whole, and assumed that it could be compared to a bag with chemicals that were transported by diffusion and reacted with one another. The model was elegant and plausible. From it, Turing showed that depending on the chemical reactions and rates of diffusion, thermodynamic fluctuations alone could lead to a variety of long-term outcomes. Under some conditions there was a high concentration of chemicals at one place in the embryo and relatively low concentrations elsewhere. Under other conditions the chemicals became distributed in spot-like or stripe-like regularities, which mimicked the pattern of tentacles in a *Hydra* or arms in a starfish. Under yet other conditions the concentrations oscillated in a clocklike, and sometimes wavelike, manner.

Turing thought that his scheme could exemplify global or system-wide mechanisms for the genesis of developmental patterns. After an initially cool reception, the model began to evoke a great deal of interest. Later, stripe formation was shown to be ubiquitous in early embryogenesis. For example, even a fly embryo exhibits patterns of gene activity that resemble a zebra's stripes; one could think that they were Turing patterns. Unfortunately, subsequent experiments showed otherwise. Indeed, it appears that there is no global rule behind stripe formation in the fly embryo. Rather, the manner in which the fly makes stripes is bizarre and confounds expectations. The stripes turn out to depend on distinct genetic regulatory interactions that lead separately to the appearance of each stripe, and not even in serial order at that[16].

How can one account for the existence of such a non-intuitive stripe-forming system? A possible answer comes from the hypothesis that global schemes such as the Turing model tell us something about the evolutionary antecedents of present-day patterns – antecedents dating from a period when evolutionary embellishments were minimal and the link between genes and development was not as intimate as it is today[17]. According to this proposal, global systems for patterning, based on physics and chemistry, may have existed in the past. Inevitably, their outcomes would have been 'noisy', meaning subject to large variations, and so unreliable. The gene-based patterning mechanisms that we see today could have come about in the course of evolution because they buffered the variations and ensured that the patterns were produced reliably[18].

Stripe formation illustrates a case of a theory not standing up. More glaring is the absence of any theory. To realize this, one has only to see that whereas a series of findings have extended our knowledge of how genes, cells and organisms function, the bulk of the new knowledge has been in the form of one surprise after another. Consider some striking examples: the discovery that not all DNA codes for proteins; the fact that genes can consist of discontinuous segments of DNA; the observation that when mutated, some genes seem to leave the organism unchanged; and the absence of any correlation between genome size and gene number. Testifying to the absence of a theory of living systems, let alone a logical theory, none of these major findings were anticipated.

Summing up

Unlike physical objects, which can be accounted for as the necessary consequences of the operation of natural laws, living entities are products of an essentially ad hoc process known as evolution. They have been moulded by natural selection in a manner that has preserved a succession of minor, randomly caused changes that turned out to be successful. The properties that they exhibit demand evolutionary explanations. An answer to the question of just what constitutes an evolutionary explanation falls outside the scope of this article, but there are at least two respects in which it is very different from what would be called an explanation in physics or chemistry. First, in evolution the environment, including the ancestral environment, plays a central role in defining the organism. Second, the basic unit of change in evolution is not the individual organism. Instead, it is a collection of entities known as the species. Because of evolution, living creatures are products of history. They make sense, that is, are capable of being understood, *only in the context* of their history. Evolutionary explanations would be out of place in the case of purely physical objects. For example, no one would think of saying that to understand the hydrogen atom, all the hydrogen atoms in the universe had to be studied, in addition to how they got that way. Of course observation, description, experimentation, logical analysis and the construction of testable hypotheses – what is sometimes called the method of science – are as much a part of biology as of the other natural sciences. But there is something that makes biology special, and that is the history of change undergone by living matter and the manner in which the change has come about. It is because of this that one must doubt whether biology can ever have a mathematical structure in the way that theoretical physics does. The evolutionary biologist Theodosius Dobzhansky put it in the form of a maxim: 'Nothing in biology makes sense except in the light of evolution'. Coming back to Wigner, mathematics is unlikely to be 'effective' in biology, let alone 'unreasonably effective', because of evolution[19].

Notes and references

1. Wigner, E. P., *Communications in Pure and Applied Mathematics*, New York: John Wiley & Sons, Inc, 1960, vol. 13. One might paraphrase Wigner by saying that a game invented by human beings for their pleasure and amusement happens, just by chance, to explain features of the real world. Wigner acknowledges later that he is referring to the physical sciences alone ('The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics …'), but this qualification tends to be ignored. In *Conversations on Mind, Matter, and Mathematics* (edited and translated by M. B. DeBevoise; Princeton University Press, 1998), Jean-Pierre Changeux and Alain Connes argue whether mathematics is a creation of the mind or reflects the external world. It is worth noting that biology has influenced mathematics in the past, and the influence is likely to get stronger in the future. But the present article does not deal with this issue.
2. A second aim is to question, if only indirectly, the implicit assumption that a product of the mind can be analysed as if it need not carry an imprint of the forces that shaped that mind. From the viewpoint of evolutionary biology, a theory of knowledge must be embedded in the evolutionary history of the mind. The ethologist Konrad Lorenz gave the clearest explanation of evolutionary epistemology. See Donald T. Campbell, *Evolutionary Epistemology* In *The philosophy of Karl R. Popper*, edited by P. A. Schilpp, 1974, pp. 412–463. LaSalle, IL: Open Court.
3. With some exceptions that we can ignore, a gene is something that formally is defined by the following properties: it is a molecular polymer belonging to the general class known as DNA; it is found inside the nucleus of living cells; it can be passed *in toto* from a father or mother to a child; and it can change, in which case some traits are also changed (thereby making it appear that traits are passed down from parents to offspring whereas actually only the genes are). The link between a gene and a trait is via a protein. The units that make up a DNA polymer, taken in sequence, specify – via certain rules of correspondence – the sequence of units that make up a protein polymer. The rules are known as the genetic code. Taken together, the set of DNA molecules inside a cell constitute the genome. Because not all DNA molecules encode proteins, the genome has both genes and non-coding DNA. For a nuanced discussion of the philosophical complexities lurking behind these definitions, see *Genetics and Reductionism* by Sahotra Sarkar (Cambridge University Press, 1998). In *A New Biology for a New Century* (*Microbiol. Mol. Biol. Rev.*, 2004, **68**, 173–186), C. R. Woese distinguishes between 'empirical reductionism', a methodological tool, and 'fundamentalist reductionism', which is exemplified by the reductionism of 19th century classical physics.
4. J. D. Watson gives a racy account of the discovery in *The Double Helix* (Athenaeum, New York, 1968). *The Path to the Double Helix* by R. Olby (University of Washington Press, Seattle, 1974) is the more scholarly history. Curiously, even though molecular biology benefited enormously by the entry of physicists, and was built on the experimental tools devised by physicists, physics as such has played only a minor role in molecular biology (a prominent exception being the role of X-ray crystallography). Rather, what physicists contributed to the field was a 'physics way' of looking at problems, of getting down to essentials. Initially they were inspired by the hope raised by Erwin Schrödinger that 'we must be prepared to find a new type of physical law prevailing in it [living

matter]' (*What is Life?*, Cambridge University Press, 1944), but that hope gradually faded away.

5. Morange, M., *A History of Molecular Biology*, Harvard University Press, Cambridge, 2000.
6. See page 4 of *36 Lectures in Biology* by S. L. Luria (MIT Press, Cambridge, 1975). *DNA: The Secret of Life* by J. D. Watson and A. W. Berry (Knopf, 2003) adopts a similar viewpoint. A forceful rebuttal can be found in *Unravelling the Secret of Life* by Barry Commoner (*GeneWatch*, 2003, **16**, 1–8).
7. Nijhout, H. F., Metaphors and the Role of Genes in Development. *BioEssays*, 1990, **12**, 441–446. The phrase 'phenotypic plasticity' expresses the observation that genotypes and phenotypes do not have a one-to-one relationship. *Morphogenesis and Pattern Formation in Biological Systems* (eds Sekimura, T. *et al*.), Springer Verlag, Tokyo, 2003 and *Origination of Organismal Form* (eds Müller, G. B. and Newman, S. A.), MIT Press, 2003 contain more on the same theme. *Visual pathways in albinos* by R. W. Guillery (*Sci. Am.*, 1974, **230**, 44–54) contains a striking example of the complex relationship between genes and traits. Also see Marks, J., *What it means to be 98% Chimpanzee*, University of California Press, 2002.
8. Once life departs, decay can be extraordinarily rapid. It has been pointed out that after ten years a discarded automobile can still be found where it was left; an animal that dies in a forest vanishes in a matter of months if not weeks or days.
9. The contention of evolutionary theory is that inanimate nature and blind chance are sufficient to engineer outcomes that seem to be designed with a prior purpose (Dawkins, R., *The Blind Watchmaker*, Penguin, London, 1989).
10. Excellent accounts of evolutionary opportunism can be found in *The meaning of evolution* by G. G. Simpson (Yale University Press, New Haven, 1967) and *The Possible and the Actual* by F. Jacob (University of Washington Press, Seattle, 1982).
11. Natural selection is the only inanimate agency that we know of that is capable of mimicking conscious design. But all evolutionary change need not be because of natural selection. See for example *Ever Since Darwin* by S. J. Gould (W. W. Norton, New York, 1977).
12. Thereby resembling, albeit for entirely different reasons, what happens in systems that exhibit chaos. A mathematical equation may be deterministic in the sense that the outcome is exactly predictable once the starting situation is precisely specified. But if infinitesimal variations in initial conditions lead to large variations in the outcome, for all practical purposes the outcome is unpredictable.
13. For a contrary view of pattern formation in biology, see *A New Kind of Science* (Wolfram, S., 2002, Wolfram Media). Following the tradition established by D'Arcy Thompson. Wolfram is explicitly anti-selectionist: 'What I have come to believe is that many of the most obvious examples of complexity in biological systems actually have very little to do with adaptation or natural selection'. Apart from telling us how to go about creating an interesting but imaginary world, the relevance of his automata-based approach for understanding biological form is unclear.
14. This is true of fields ranging from literary studies to sports. There is of course a purely utilitarian role that mathematical or computer-based approaches serve. No one questions the usefulness of statistical analysis whenever quantitative measurements are carried out. But we are asking whether a mathematical or algorithmic structure might explain biology in the way that it seems to explain the physical universe.
15. Turing, A. M., In *The Chemical Basis of Morphogenesis. Philos. Trans. Roy. Soc. London*, 1952, **B52**, 14–152. For a discussion of the history of Turing models see *Alan Turing and the Chemical Basis of Morphogenesis*, Nanjundiah, V., In *Morphogenesis and Pattern Formation in Biological Systems* (eds Sekimura, T. *et al*.), Springer Verlag, Tokyo, 2003, pp. 33–45.
16. Akam, M., Making stripes inelegantly. *Nature*, 1989, **341**, 282–283. The point is not that a Turing mechanism is impossible; nor is it that no such mechanism exists in any organism. The point is that the same outcome can be reached by what is presumably an ad-hoc evolutionary process.
17. See Newman, S. A., Is segmentation generic?. *BioEssays*, 1993, **15**, 277–283. Multicellular development, including pattern formation, has to do with the behaviour of groups of cells. It demands, therefore, a 'sociological' approach. The same point has been made with regard to aberrant development in *The Society of Cells: Cancer and control of cell proliferation* (C. Sonnenschein and A. M. Soto, Springer-Verlag, New York, 1999).
18. C. H. Waddington made the concept of reliability or buffering the defining element his way of looking at development and coined the word 'canalisation' to describe it.
19. This calls into question the attitude that an area of study can be called a science only if it supports a mathematical framework. The mathematics-in-biology story has many other angles, among them sociological ones, that are not considered here. The intellectual prestige associated with mathematics, and at one step removed, of theoretical physics, has sometimes tended to overawe biologists. In *The Growth of Biological Thought* (Harvard University Press, 1985) Ernst Mayr offers an acid comment on this tendency: 'Physics envy is the curse of biology'. As to human and other reasons why mathematical biology has not been welcomed by many biologists, see Evelyn Fox Keller in *Making Sense of Life* (Harvard University Press), 2002.

Computation – The new mathematics for sciences

N. D. Hari Dass

Institute of Mathematical Sciences, CIT Campus, Taramani, Chennai 600 113, India

COMPUTATION is actually old mathematics, in fact very old and basic, consisting of the basic operations of arithmetic, viz. addition, subtraction, multiplication and division. What is new is the tremendous speed with which these basic operations can be carried out on modern computers and supercomputers. The fastest supercomputer today, the BlueGene, can perform 70 thousand billion such operations in a second (70 Teraflops). The supercomputer KABRU, self-assembled at the Institute of Mathematical Sciences, Chennai, can perform at a Teraflop level.

Strictly speaking, the word computation can mean any logical operation even if it is very abstract. Here we are using it in the sense of an operation on a computer. Computers themselves can be of radically different types. The most familiar are the digital and analogue computers. The latter make use of the fact that certain physical quantities obey certain mathematical equations and by directly measuring the physical quantity in question, one obtains the solution for the relevant mathematical equation. In other words, one has 'computed' the solution. An illustrative example is the current in an electrical circuit made up of resistors, capacitors and the so-called inductors; the current is a solution to a differential equation whose parameters are determined by the capacitance, resistance and inductance of the circuit. By varying these, one has a whole family of differential equations whose solutions, represented by the current, can simply be measured directly or even graphically displayed on an oscilloscope.

In the case of the digital computer all mathematical manipulations are broken down to the basic operations of arithmetic, which are then carried out by specific pieces of hardware. There are other types of computers that have been thought about lately like the DNA-computer and the chaotic computer. In fact a computer has been built out of water pipes and junctions! Basically any physical phenomenon that 'simulates' the basic operations of arithmetic can be used to build the 'gates' of a computer. But the range of computations that can be done speedily is rather limited on analogue computers, at least for the moment.

Before discussing the significance of the tremendous speeds of computers that have jettisoned the computational methods to the forefront, it is worthwhile commenting on the 'reliability' of computers. After all, every form of computer mentioned before is made up of physical components that are liable to have fluctuations in their responses or even liable to breakdowns. How then can one rely on the results obtained on a computer? To drive the point home, in the case of the analogue computer, the resistance, for example, can have small fluctuations or could be affected by the heating of the elements due to the flowing current. This would certainly introduce errors in the computations. As the celebrated mathematician-physicist and the founder of modern computers, John von Neumann, put it, 'Can we build a reliable computer out of unreliable components?' This leads to the notion of fault tolerance which is achieved by either choosing components that have the desired level of reliability or even better, through very ingenious means of 'error correction'. These error-correcting techniques are in fact what allow us to have a clear phone conversation over a noisy channel. The modern supercomputers can perform hundred billion billion operations without a single error!

So what is the significance of such tremendous speeds for scientific investigations? To appreciate this it is instructive to say a few words about the alternative to numerical computations on a computer, namely, the so-called 'analytical' calculations. Here the aim is to derive formulae for various quantities of interest even though such formulae involve lot of approximations and even questionable metamorphosis of the basis framework (theory, model, etc.) itself. This is often considered more satisfactory than a numerical evaluation because of the belief that 'a single formula is worth more than a million numbers' somewhat similar to the adage that 'a picture is worth more than a thousand But like all adages and beliefs this too hides more than what it reveals. Not all pictures are worth more than a thousand words nor all formulae worth more than a million numbers. What is of prime importance is, of course, the quality of numbers versus the quality of the formula. Besides, as Hamming has aptly said, 'the objective of computations is insight and not numbers'.

Many times, even when one has a formula for a quantity of interest, it is difficult to extract the desired information in a straightforward manner. For example, if it is known that the quantity of interest is the lowest zero of a theta function, no formula may be available for it explicitly despite one having an 'analytical result'. In such cases one can use computers for a numerical evaluation of the result.

But this is not where the real impact of the great strides made in computation manifests itself. That occurs in what

e-mail: dass@imsc.res.in

are called numerical simulations. Here one tries to actually simulate a system or some controlled approximation to it directly on powerful computers. To give an example that should be familiar to all, we can consider an accurate simulation of the motion of the planets of the solar system. While the motion of a single planet around the Sun as given by Newton's and Kepler's laws can be solved analytically in terms of simple formulae, the mere inclusion of a third body, say, Jupiter, makes an exact solution intractable and there are reasons to believe that this problem cannot be solved at all. Adding all the other planets and their moons aggravates the problem that much more. So one need not go to black holes or string theory to encounter mathematically very challenging problems.

However this problem can be simulated on a computer if we are willing to replace the continuum of time by grid of discrete points with the simulations becoming more and more reliable as the grids are made finer and finer. The grids can be made finer only at the cost of vastly increased computer resources both in terms of speed and of memory. The necessity of discretizing time arises as we can only deal with a finite, though very very large, number of quantities on a computer. The great speeds in computation allow, for example, to simulate the motion of all planets over several billion years in a very short time.

Today the repertoire of the class of problems that can be numerically simulated is really vast. It includes problems like turbulence in fluid flow (still an unsolved problem in the analytical sense), large scale weather forecast, behaviour of biomolecules, *ab initio* study of chemical reactions, black hole dynamics, problem of quark confinement (yet another outstanding problem of theoretical physics), quark-gluon plasma which was the state of matter in early universe, and many more.

While the simulation of planetary motion involves transcribing Newton's laws for the planets to discrete grids, simulation of the problem of quark confinement is made possible by exploiting some remarkable structures of quantum field theory whereby a quantum mechanical problem can be simulated to desired accuracy on a classical computer. Another important distinction between the planetary motion simulation and simulation of quantum field theories is that the former is deterministic while the latter is probabilistic. This means the simulation algorithm produces results according to some probability law and this is just what the laws of statistical mechanics are all about. For the simulations to be reliable one needs to understand subtle issues like ergodicity and ensure that the simulations are consistent with these principles.

The stochastic or probabilistic simulations heavily use techniques of statistical mechanics. Even though the list alluded to earlier includes areas that are vastly different, what is remarkable about them is the surprising similarity of computational techniques used. For example, the techniques of Monte Carlo or molecular dynamics are used for studying nonabelian gauge theories of particle physics as well as to study dynamics of polymers. In fact the most striking aspect of computations in science is the universality of computing techniques for very diverse applications. This helps scientists to overcome traditional barriers and boundaries and to realize science as one undivided path to understanding nature.

Be it elementary particle physics, hydrodynamics, biological systems or any of the many areas where computational approach has made a big impact, a formidable challenge is to understand phenomena over vastly differing scales in a mutually consistent unified manner. For example, knowing that all matter is made up of atoms, how do we understand the bulk properties of matter? Or knowing that all neutrons and protons consist of quarks how do we understand the forces that bind nuclei? Knowing that cells constitute living beings, how do we understand the texture of muscle tissues? This problem of relating phenomena at many different scales is a recurring theme in all sciences. On the analytical side, it is very difficult to make much progress on this front in a quantitatively precise way though physics is full of instances where ingenious insights have paved the way, as in Feynman's theory of superfluid helium or in the BCS theory of superconductivity, etc.

It is this multi-scale facet that can be most efficiently studied using numerical simulations. One way to do this is to distil from simulations at a particular scale, the relevant degrees of freedom at that scale and use these as the atomic constituents for simulations at a larger scale and iterate this procedure. This is not an approach that can be applied in a purely mechanical manner as it needs ingenuity in identifying the so-called relevant degrees of freedom at each scale. It is however possible to do the so-called multiscale or multigrid simulations whereby one can study phenomena at several scales simultaneously. In such a multiscale simulation, variables at many different scales are simultaneously manipulated which allows one a grand view of the problem across many scales.

I would like to illustrate the nature of this multiscale phenomenon with an example. Let us consider water and look at it at different scales. At one level we can understand it as a molecule of two hydrogen and one oxygen atoms. At this level we understand its structure to be that of a tetrahedron. But if we pack a lot of such water molecules together, they start forming long chains (polymerization is the technical description) and the behaviour of water is now governed by the string-like constitution. Depending on the temperature this assembly behaves like a liquid or like a complex solid called ice with varied crystalline structure. Inside a living cell, in the vicinity of biomolecules like DNA and proteins, the very structure and properties of water get modified which in turn has profound implications for the behaviour and functionality of the biomolecules. Thus the same entity, water, has vastly different properties at different scales. It makes more sense to view these properties as 'emergent', arising out of the nature of assembly, rather than as a property encoded in the microscopic constituents.

Another important reason why simulations are becoming indispensable is that the parameter space available to real experiments is restricted while in simulations one could probe arbitrary regions of the parameter space. To illustrate this we could go back to our example of the planetary system. There are peculiar phenomena like the moon always showing the same face to the earth which arise out of special confluence of parameters. To get a proper understanding of this would require studying what would happen when the parameters take more general values. But that is beyond the realm of actual experiments but is easily accessible to simulations. There are even more complex systems where this aspect of simulations becomes crucial.

Here are some of the major results that have been obtained through simulations that were not obtained through other mathematical techniques. I give only a very partial list here: (i) evidence that the planet Pluto's motion is chaotic and therefore unpredictable in the long run, (ii) the demonstration that the nature of the vacuum state of QCD is such that it leads to permanent quark confinement, (iii) synthesis of new materials, (iv) *ab initio* study of chemical kinetics, (v) vortices in Bose–Einstein condensates (first established in simulations and later verified through laboratory experiments) and many more.

The remarks made so far should not be misconstrued as saying that the numerical approach is a substitute for analytical work. That would indeed be preposterous. What is being emphasized is that for many the analytical approach is the only way of applying mathematics in sciences. We hold that view to be extremely myopic and counterproductive. The class of problems that are amenable to reliable analytical methods is fast shrinking while those that can be tackled with reliable numerical techniques are exploding. An efficient numerical algorithm always makes use of the most powerful insights afforded by the analytical route but the reverse happens seldom except in some happy exceptions like the four colour problem or some investigations in number theory.

To paraphrase one of my friends, just as the analytical techniques of mathematics made possible, the transition from premodern to modern science, computational methods will mark the transition from the modern to the future paradigms of science. I would go a step further and assert that physics and science in general will become more and more computational in character.

In summary the great speeds available to us on modern supercomputers has made possible many investigations unthinkable till now. The computational paradigms transcending the artificial boundaries between different disciplines should herald a new age of science. Many of the techniques and algorithms in numerical simulations are independent of the disciplines. This arises partly due to similarities in basic equations from unrelated disciplines if one transforms the variables adequately. Problems about DNA can be structurally very similar to those arising in cosmic strings; hydrodynamic flow equations in certain regimes look very much like the equations of general relativity; simulation algorithms for protein-folding are very much like well-known algorithms in statistical mechanics. Of course, there are important differences between disciplines but the commonality of the simulation methodology can bring sciences together so that they can appreciate and be tolerant of these differences more than was ever possible. The message of different paradigms and different notions of reality at different scales is also philosophically appealing as against a theory of everything for all scales which appears distinctly pre-Copernican in spirit.

Mathematics and computer science: The interplay

C. E. Veni Madhavan

Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560 012, India

Mathematics has been an important intellectual pre-occupation of man for a long time. Computer science as a formal discipline is about seven decades young. However, one thing in common between *all* users and producers of mathematical thought is the almost involuntary use of *computing*. In this article, we bring to fore the many close connections and parallels between the two sciences of mathematics and computing. We show that, unlike in the other branches of human inquiry where mathematics is merely utilized or applied, computer science also returns additional value to mathematics by introducing certain new computational paradigms and methodologies and also by posing new foundational questions. We emphasize the strong interplay and interactions by looking at some exciting contemporary results from number theory and combinatorial mathematics and algorithms of computer science.

Nature of the sciences

MATHEMATICS has been an important intellectual pre-occupation of man for a long time. It is said that the *mathematics of the common man* or the *mathematics of the millions* does not possess the esoteric abstractions and expressions of the *mathematics of the professional mathematician* or the *millennium mathematics problems*. However, one thing in common between *all* users and producers of mathematical thought is the almost involuntary use of *computing*. The nature, intensity and extent of the actual computing may differ quite widely among these different forms of engagements with mathematics and computing. The basic processes of empirical verification, development of abstractions from concrete instances inherent in mathematical thinking has much in common with algorithmic or computational thinking.

Mathematics is as old as humanity, while computer science is a young discipline, about seven decades old. The strong interplay between mathematics and computer science is at its peak today.

Much has been written by philosophers on the nature of human inquiry. Yet, it is quite difficult to define a term such as mathematics in a comprehensive and succinct manner. A whimsical, circular definition states that mathematics is what is practiced by mathematicians. It is not possible to attempt even such an indicative definition of the discipline of computer science. Computer science is perceived as an inter-disciplinary field with clear tracks of contributions from electrical and electronic engineering way of thinking and a clear track of contributions from the mathematical way of thinking. Further, these two tracks ramify into many different sub-tracks and rejoin after a few complex interactions. At a broad level computer science thus straddles simultaneously scientific, engineering and technological principles.

The term computing science seems to refer to the act of computing whereas the term computer science seems to refer to the principles of architecting and engineering a computer. Indeed, it is true as seen from the previous sentence, computer science sometimes refers to all of these and more. We will use somewhat synonymously the terms, computer science and computing science. The other terms computational science, science of computing, scientific computing, mathematics of computation, computational mathematics all have slightly different specialized meanings and occasionally allow some overlap. We will be conscious that we sometimes broaden our boundaries. The purpose of this article is not to delineate and distinguish clear demarcations. On the other hand the intent is to bring out the beautiful connections, similarities and parallels between the two disciplines of *mathematics* and *computer science*.

In our work we consider only the facet of theoretical underpinnings of computer science leaning on the science of algorithms. We do not consider the meta-physical, cognitive connections and hence do not make forays into topics of formal logics, computational linguistics, artificial intelligence, machine learning and other related fields. Also we do not discuss the exciting new possibilities of computing machines based on the quantum mechanical or biological processes, that may render as irrelevant some of the deep theoretical and practical issues we study here.

The early formalisms in computer science have made clear the connections among the three entities of machines, languages and computation. They also established these connections by formalisms based on various forms of automata, the corresponding formal language classes and the generative mechanisms of grammars. Machines are synonymous with algorithms in a formal sense. From these theoretical underpinnings arose the formal specifications of algorithms and programs. The evolution of the theoretical foundations of computing science took place along these two clear tracks – algorithms and programs.

e-mail: cevm@csa.iisc.ernet.in

The older form of the word algorithm is *algorism*. This term refers to the process of doing arithmetic using Arabic numerals. Mathematical historians trace the etymology of the word *algorithm* to the word algorism. The term comes from the name of a Persian author, Abu Musa al-Khowarizmi (c. 825 AD)[1], who wrote a celebrated book, *Kitab al jabr w'al-muqabala* (Rules of Restoration and Reduction). The word algebra stems from the title of this book. The ancient town of Khowarizm is the modern town of Khiva in Russia.

An *algorithm* is a compact, precise description of the steps to solve a problem. It is a finite, definite, effective procedure taking an input and producing an output. A *program* is a precise set of statements expressed using the syntactic, semantic and linguistic constructs of a programming language, that embodies the steps of an algorithm. The remarkable evolution and applications of computers is due to the immense range of developments in the science of algorithms and programs. The two main issues in the study of algorithms and programs are *correctness* and *computational complexity*. Many theoretical and pragmatic ideas along these directions have led to remarkable advances.

We discuss the notion of proofs in mathematics and the role of computers and computing in this context. We discuss the issues of correctness and computational complexity in the context of design and analysis of algorithms. It is here that many fascinating connections between mathematics and computing science appear in many surprising ways. These connections have led to very exciting developments in both fields of inquiry.

We do not discuss in this paper, the formal methods for reasoning about programs and their intrinsic relationship with certain abstract structures of mathematical logic and algebraic structures. We merely cite the excellent works and treatises of the pioneers Dijkstra[2] on the science of programming, of Wirth on the notion of data structures which together with algorithms leads to the realization of programs and of Manna[4] on the logics used to reason about the correctness of programs.

Algorithms and computational complexity

A lot has been said and written on the topic of design and analysis of algorithms. There are a large number of great text books, advanced monographs and high quality specialized journals and conferences in this area. They cover a range of issues related to design strategies (such as greedy, divide and conquer, branch and bound, binary doubling, etc.), implementation (choice of data structures to store and process information), and analysis (model of computation, asymptotics, lower bounds, structures in complexity, etc.). We focus only on the larger macro level features and their relationships with mathematics.

A beautiful summary of the common features shared by algorithmic thinking and mathematical thinking is made by Knuth[5]. The two forms of thinking are characterized by features such as formula manipulation, representation of reality, reduction to simpler problems, abstract reasoning. The notable differences in features are the way of handling the uncountable continuum in mathematics and the way of handling the notion of size of proof (computational complexity) in computer science.

The algorithm of Euclid (*Elements, Book 7*) for calculating the greatest common divisor (gcd) of two integers is a fine example of the modern notion of algorithm. It displays all the properties and features related to establishing the correctness and computational complexity. One measures the computational complexity of an algorithm by expressing the number of basic steps taken by the algorithm to solve a problem, in terms of the sizes of the input (and the sizes of all the intermediate results). In general, the maximum number of steps taken over all possible inputs over asymptotically large-sized inputs is used as an upper bound and serves to characterize the behaviour of the algorithm. This measure, expressed as a function of the input size is known as the *worst case asymptotic* (*time*) *complexity* of the algorithm.

The size of an element is usually the number of binary bits needed to encode the element. Let the sizes of the two input integers x, y be n bits. The Euclidean gcd algorithm to obtain the gcd(x, y) requires number of steps (or equivalently time) proportional to $O(n^3)$. Thus the computational complexity of the Euclidean gcd algorithm is said to be cubic in input size. An algorithm whose complexity is polynomial in the input size is considered *good*. A problem that admits a polynomial time algorithm is said to be *tractable* or *easy*.

On the other hand there are a large number of problems, that do not seem to be *easy* (i.e. they do not seem to be solvable by means of a polynomial time algorithm). Consider the mathematical structure *graph*. Let $G = (V, E)$ be a graph, where V is a set of n *vertices*, and E is a collection of *edges* (a subset of the set of $n(n-1)/2$ pairs of vertices). The size of the input in the case of a graph is proportional to the number of vertices and edges and is bounded by $O(n^2)$. For example, the problem of determining whether a *graph* G has a *Hamiltonian cycle* (a cyclic traversal of the edges of the graph visiting every vertex exactly once), does not seem to be *easy* in the above technical sense. Clearly, by listing all possible $n!$ permutations of the vertices and checking whether the graph can be traversed along the edges as per the vertex permutation, one can solve the problem. It is obvious that this algorithm, adopts a *brute-force* approach of exhaustive enumeration of all possible solutions. The computational complexity is *exponential* in the number of vertices since the function $n!$ grows approximately as $O(n^{n+1/2}\exp^{-n})$.

Proofs in mathematics and computer science

In the previous section, we discussed how we assess the computational complexity of an algorithm. There is the other aspect of *checking* the answer produced. How do we know that the answer produced by an algorithm or the realization of the algorithm is *correct*? Is there a simple mechanism, a formula, an expression into which we can plug in the answer and check the answer? For a simple example, consider the problem of subtraction of integer b from integer a. We immediately recognize that we could add the result of subtraction to b and check for match with a. Now, what is the complexity of carrying out this verification? In this case, addition and subtraction are both *linear* in the length of the operands a and b and hence the verification can be carried out in *polynomial* time and hence we say that the verification is *easy*. Here we verified the correctness of an algorithm by *using* another algorithm. Of course we assumed that the other algorithm produced correct answers and that the result of that can be verified easily.

The answers to problems that have polynomial time algorithms for producing the answers seem to be verifiable in polynomial time (see the subtraction problem) by running another complementary algorithm. Another pair of simple, complementary algorithms are the multiplication and division algorithms with quadratic running time. This situation prevails in a non-obvious way in many situations. The non-obvious nature comes from the role of the *mathematical theorems* that characterize the situation. For example, in the case of the greatest common divisor of integers, there is a theorem due to Bezout, that states that if $\gcd(x, y) = g$, then there exist two integers u, v such that $ux + vy = g$. The Euclidean algorithm can be *extended* to determine the constants u, v besides g, given x, y. While the Bezout relation can be used to check if the gcd g is correct, assuming the quantities u, v are correct. The catch is that u, v are also obtained by the same algorithm. However, there is another way. From the fundamental theorem of arithmetic and the consequent unique factorization theorem the $\gcd(x, y)$ is immediate from the prime factorizations of x and y. Now we have a different catch. Factorization of an integer is not known to be polynomial time solvable. Much of our current research is centered around this problem.

In the next section, we discuss the formal structures that cope with these kinds of situations. The main message here is that *independent characterizations* enable the determination and verification of an *answer*. It is not at all clear whether two or more independent characterizations would be available for a problem. It is not true that the different characterizations will lead to the development of polynomial time algorithms to produce the answer. Such situations are dealt with in problems of optimization, by means of a variety of duality, reciprocity and min-max theorems. We discuss these in a subsequent section. Although the initial formal notions of algorithm and computational complexity figured most prominently in the computer science literature during the1970s, the notion made its first appearance in a classic paper by Edmonds[6] on maximum matchings in graphs.

On the other hand, many problems have the peculiar characteristic that the answers could be verified in polynomial time; but they do not seem to admit polynomial time algorithms to find the answer (see the Hamiltonian cycle problem). It is clear that the *Yes* answer to the question of Hamiltonicity of a graph can be verified in linear time once a Hamiltonian cycle is given. This situation prevails in the case of many problems of different nature (decision, search, optimization, etc.), arising in different domains (sets, integers, graphs, etc.) and from many different application areas (scheduling, information retrieval, operations research, etc.).

Computer scientists realized quite early, that there is a philosophic difference between the computational complexity of an algorithm used to solve a problem and the computational complexity of verification of the answer. They also put on a firm formal footing these notions of algorithmic production of a solution and that of verification. They also set up, in the spirit of mathematical abstraction, equivalence classes of problems categorized according to their computational complexity. The equivalence class of problems whose answers could be verified in polynomial time was called **NP** and the former class **P**. Clearly **P** is contained in **NP**. An important result was the postulate of an equivalence class of *hard* problems within **NP** called the **NP-complete** class. An intriguing situation in the realm of computational complexity is that there does not seem to be easily verifiable *succinct certificates* to the **No** answers to decision versions of many problems. For example, it is not at all known how one can provide a short polynomial time verifiable proof to the answer that a given graph does *not* have a Hamiltonian cycle.

Computer scientists have built many interesting sub-structures within **NP**. The fundamental meta-scientific question is whether **P** is equal to **NP**. This question has engaged computer scientists for the last three decades, in developing formalisms and a lot of technical hair-splitting. We will not go into further details of this fascinating question from the foundations of theoretical computer science. We refer the reader to the classic book on this subject by Garey and Johnson[7]. We just make a few comments on the current perspectives on this subject which borders on meta mathematics.

I would borrow the quaint little phrase *Meta-magical Themas* used by Douglas Hofstadter as the title for his column in the journal *Scientific American*, to describe these questions. Interestingly, this phrase was coined by Hofstadter, a computer scientist, as an anagram of the title *Mathematical Games* used by Martin Gardner for many years as the title of his column in *Scientific American*.

The Clay Mathematics Institute of Massachusetts, USA, named seven Prize problems, each carrying a cash award of one million dollars, in May 2000. These include the celebrated Riemann hypothesis, and a few other outstanding conjectures of twentieth century mathematics. These problems are toasted to be on the same class of problems stated by D. Hilbert in the year 1900. The eminent French mathematician, J. P. Serre, commenting on these problems, said that the choice by the Clay Institute was very apt. He hoped that the **P** = **NP** question would not turn out to be *undecidable*. This question perhaps, is a meta question and is on the same footing as certain axiomatic questions in the foundations of set theory upon which the entire edifice of modern mathematics rests.

Thus the notion of quickly verifiable *succinct proof* or a short certificate to the correctness of the solution determined by an algorithm, is a cornerstone contribution of computer science to mathematical thought. In the following sections we look at the notion proofs in mathematics and some close connections with computational thought.

Many proofs and many algorithms

In mathematics the notion of proof is a central ingredient of all results. No mathematical statement is made without a conscious attempt to establish or prove the statement. Mathematicians are taught and trained in the course of their development from a student to a professional practitioner, the science of proving statements. Mathematical statements consist of a set of axioms, hypotheses, antecedents and consequents. The statements are established by a sequence of logical consequences from the hypotheses and antecedents to the consequents by a chain of logical deductions. Of course, mathematics is not to be construed as a mere set of dry, mechanical deductions. If that were so, such a human endeavour would not have survived. There is a certain innate beauty in many mathematical statements and theorems and very often in the many twists and turns of the proofs. One wonders at times if this innate quality of mathematics is akin to the surrealistic patterns of human inquiry in fields such as the fine arts, literature, music, painting and sculpture. Indeed, the famous mathematician Littlewood once described eloquently certain results of Srinivasa Ramanujan as evoking the sense of wonder produced by looking at a certain ecclesiastical architectural masterpiece.

Many mathematicians and philosophers have pondered over this *innate* quality of mathematical thought. This innate quality transcends the physical world. Scientists in many disciplines have come across a similar phenomenon in explaining their science. Mathematical statements and proofs have an existence of their own, as pure figments of imagination. However, there is evident in many of these forms of expressions of thought beautiful insight into the structure of the world of mathematical objects. In general, these objects are used to model the physical world and the statements are about the way these objects interact to produce more complex structures with real world consequences.

In modern times, mathematicians and computer scientists have contributed many significant ideas that have all the features discussed above. In the last few decades computer science has emerged as a mature foundational science, with its own corpus of axioms, statements and body of results. It is here that the science of algorithms, or sometimes referred to as algorithmics (particularly by the French research community), stands out like a beacon proclaiming its status on par with mathematics. The fraternity of mathematicians and computer scientists have come to recognize and acknowledge with mutual respect certain exquisite results from their respective disciplines. Of course, mathematics being a much older science has many more theorems and results of great beauty and significance than the algorithms of computer science. Increasingly, computer science is enriching mathematics in many ways. This phenomenon is most visible in the area of *discrete mathematics*. We look at this phenomenon more closely in the next section.

We discussed above the conceptual parallels between the nature of proofs in mathematics and the nature of algorithms in computer science. We take this further to identify certain other similarities in our paper. We identify two typical characteristics. (i) *many alternative proofs* (ii) *short*, *succinct* or *elegant proofs*. Both characteristics are present in abundance in the sciences of discrete mathematics and algorithms. In fact, these are precisely the reasons that make these two subjects so fascinating. It is tempting to cite and annotate the several examples to illustrate these two features from the two disciplines. I shall restrict to just a couple of illustrations to make the point.

Many proofs

The *law of quadratic reciprocity* of Gauss is lauded as the gem of arithmetic by the mathematical historian, E. T. Bell. This law expresses the quadratic residuosity of a prime integer p modulo another prime q, in terms of the quadratic residuosity of q modulo p in the form of a surprisingly simple expression in terms of the parities of p and q. As a mathematical statement this is indeed a gem. It is beautiful, simple and so natural to think up. What is more, this law occupied the centre stage of number theoretic research during a large part of the 19th century. It is not surprising that Gauss himself discovered eight proofs, and a 152nd proof of the law was published[8] in 1976. The hunt for a generalization of this law, meaning higher order reciprocity connecting solutions of higher power congruences modulo prime integers, occupied the legendary mathematicians Gauss, Kummer, Jacobi, Dirichlet. Finally, Eisenstein obtained a proof of the law in 1865. What fascinates me besides these interesting *many* proofs by *many* mathematicians is the *many* connections to many other *problems* that have led to surprising new developments. Indeed, Gauss, Eisenstein, Kummer, Dirichlet and others were all seeking proofs of higher power reciprocity laws (see Edwards[9] and Lemmermeyer[8]) while attempting to prove Fermat's last theorem. In the course of this quest, Kummer had a big success in proving the Fermat's last theorem for the class of *irregular prime* exponents and paved the way for the development of modern class field theory.

Mathematicians devise ingenious *different* proofs by applying their intuition together with techniques and principles from diverse branches of mathematics in surprising ways.

For example, the different proofs of the law quadratic reciprocity cited above draw upon principles from elementary enumerative combinatorial arguments to sophisticated use of algebraic structures and complex analytic tools. To a specialist mathematician, the hundreds of proofs mentioned above would all fall into a few categories. Nevertheless, these few categories encompass surprisingly different principles. That speaks for what is known as scientific serendipity without which much of the science of mathematics would lose its charm.

The famous *prime number theorem* states that the number of primes less than an integer x is about $x/\log(x)$. This simple statement about the lore of integers has fascinated number theorists for a long time. The proof of this theorem by Hadamard and de la Vallee Poussin is a classic example in the world of mathematics of a less well-known mathematician's name being associated with a great result. Surprisingly, this and many other similar, related results in the fascinating branch of mathematics called number theory[10,11], invoke simultaneously techniques for dealing with the discrete and the continuous. At a deeper level, this is not too surprising. As Kronecker, put it *God made natural numbers, rest is the work of man*. A natural mathematical approach is to begin modelling and analysing a finite, discrete world and then entering the continuum by adopting an asymptotic limiting process. I shall not go into further explanations of this metaphor.

I end with the comment about the alternative proof of the prime number theorem by Erdos and Selberg. This proof used combinatorial arguments entirely, and did not use any complex analysis techniques. The authors called their proof an elementary proof. This proof is an illustration of what we term, elementary but not easy. It is elementary in the sense that it does not use advanced techniques available only to the professionally trained mathematician. It is not easy in the sense that the arguments are involved and intricate. However, this proof is *longer* than the analytic proof. We use the adjective longer here, to indicate the degree of difficulty in wading through the verification of several minute details. Mathematicians, like artists, sometimes refer to the quality of a proof by the term *elegance*. We will consider this notion of *mathematical elegance* in the next sections.

Many algorithms

Mathematicians are coming around to the viewpoint that *algorithms* as studied in computer science could be considered very close to the object they study, called *theorems*. The basis for this acceptance has been discussed above. We further add here that in the last two decades, many prestigious fora and awards that recognize significant contemporary mathematical results have welcomed the fresh breath of exciting mathematical results from computer science.

Computer scientists have devised many interesting algorithms for the same problem. In doing this, they have often blended great insights from the structures in the problem domain, the structures inherent in the input data and certain insightful generic algorithmic strategies. I give some sophisticated examples later. Here we mention a few basic problems for which computer scientists have come up with surprisingly different ways of solving them. These are: (i) sorting a set of elements[12], (ii) multiplication of polynomials, integers and matrices[10], (iii) finding shortest paths in graphs[13], (iv) rapid exponentiation in groups[10].

Some of the brilliant, and far reaching ideas from the world of algorithmics, that have enriched the mathematical world come from topics such as polynomial time interior point-based algorithms for the linear optimization problem, equivalence of combinatorial optimization problems in terms of computational complexity; the applicability of lattice basis reduction for designing algorithms for optimization problems; the notions of approximation, randomized algorithms; algorithmic combinatorial geometry; polynomial time primality testing algorithm for integers; modular exponentiation-based public key cryptographic algorithms, syndrome decoding algorithm in error correcting codes.

What I set out to illustrate in this section is the situation of *many* algorithms for solving a problem, like the situation of many proofs for the same theorem in mathematics. These *different* algorithms for the same problem provide a vast repertoire of illustrative pedagogic, research principles and techniques that enliven the science of algorithms. They provide the continual challenges to the scientists to come up with *better, elegant, efficient and practical* algorithms.

Elegant proofs and elegant algorithms

The legendary twentieth century mathematician Paul Erdos liked to talk about *The Book* in which God maintains the most elegant or perfect proofs. The Hungarian mathematician was known as the most prolific mathematician of all time, comparable to the other prolific mathematician, Euler of an earlier era. Erdos maintained that even if one did not believe in the existence of God, one should believe in the existence of The Book. Aigner and Ziegler[14], made a selection of 30 items in a collection published in 1998. It is an unenviable task to assemble such a collection. A certain amount of mathematical chauvinism is inevitable and so there are possibilities of omission in the selection. However, it is irrefutable that what figures in such a selection is indeed elegant.

I cite a few from this collection.

- For any positive integer n there is a prime between n and $2n$.
- A natural number n is a sum of two squares if and only if every prime factor p of the form $4m + 3$ appears to an even power in n.
- p, p^2, $\exp^r$ for rational r, are all irrational.

- If G is a connected, planar graph, then $n - e + f = 2$. Here n, e, f are the number of vertices, edges and faces in G.
- No more than 12 spheres can simultaneously touch a sphere of the same size.
- There are n^{n-2} different labeled trees on n vertices.
- A partial Latin square of order n with at most $n - 1$ filled cells can be completed to a Latin square of the same order.
- Every planar map can be five coloured.
- For any museum with n walls $n/3$ guards suffice.
- A triangle-free graph on n vertices has at most $n^2/4$ edges (generalizations exist for p-clique-free graphs for all p).

The main intent in giving the list above is to display the simplicity of the statements. Other interesting aspects are that none of the proofs run for more than a few pages; and are accessible to anyone with a good high school level of mathematical training. That is the nature of *succinctness* or *elegance* of mathematical thought.

It would be extremely fascinating to think up *The Book* of algorithms. Computer science is perhaps, nearly ready to offer candidates. I could make the following random selection, without much trepidation.

- The Euclidean algorithm for greatest common divisor.
- The quick-sort algorithm.
- Algorithm to test the planarity of a graph.
- Finding the maximum matching in a graph.
- The randomized algorithm for the roots of a polynomial over a finite field.
- The Fast Fourier transform algorithm.
- The deterministic polynomial time primality testing algorithm.
- Determination of the convex hull of a set of points in 3 dimensions.
- Run length compression of binary strings.
- Set union, search algorithm.

The interplay

In the above sections we discussed the principal features of *theorems* of mathematics and *algorithms* of computer science to bring out the close parallels, connections and also some differences. The intent was to display the nature of thinking: *mathematical* and *algorithmic*. In this section, we develop this idea and show how these two disciplines have enriched each other. Significant developments in the field of combinatorial mathematics have happened in the last few decades alongside *related* developments in computer science. Often questions are posed across the disciplines and the results obtained have a bearing on each other.

In combinatorial mathematics one begins with a *ground set* and then studies certain objects built from the elements of the ground set, called *configurations*. For example, permutations, combinations are simple configurations on discrete sets. More sophisticated configurations are partitions and collections with specific properties. Various ground sets made of elements of algebraic structures (groups, residue classes modulo prime integers, elements of finite fields, vectors in a vector space), geometric structures (points, lines, hyperplanes, regions, polyhedra), arithmetical structures (integers, rationals, Diophantine equations, equations over finite fields) are all of great interest.

In the field of combinatorial mathematics, the questions posed fall into three classes: *existential*, *enumerative* and *constructive*. Existential combinatorics is concerned with questions of proof of existence or non-existence of certain specific types of configurations. They also seek to obtain mathematical characterizations or necessary and sufficiency conditions for the existence.

For example, the famous condition by Kuratowski states that a graph is planar if and only if it does *not* have a subgraph *homeomorphic* to the complete graph on 5 vertices, denoted a K_5, or a complete bipartite graph on 6 vertices, denoted a $K_{3,3}$. The proof of this theorem does not lend itself to an algorithm for checking the condition. Thus the proof is purely existential and not effective. However, the proof can be modified to a version that suits an algorithmic determination of planarity, by actually constructing a planar embedding or reporting impossibility if the graph is *not* planar. Further, the famous planarity testing algorithm of Hopcroft and Tarjan works in *linear time*. It uses a special data structure. Contrast this situation with that of the computer based proof of the *Four Colour Theorem*. We mentioned in the list in the previous section, that there is a nice theorem with a short proof to the fact that every planar graph can be coloured with 5 colours. The proof by Appel and Haken[3] of the 4 Colour Theorem, required a voluminous inspection of various sub-cases, using a computer to eliminate *unavoidable configurations* after formally, reducing the general situation to this large set. Much debate has taken place about the *nature* and hence the acceptability of this kind of proof. As a scientific work, this proof is as rigorous as any in mathematics. The dissatisfaction about this kind of proof stems from the fact that it seems to have to consider *many cases*. The purist likes to see a *short*, *elegant*, *symbolic* proof. The algorithmician (my contrived word to describe a person who is interested in design and analysis of algorithms), likes to see a *neat constructive* proof, that he can convert to an *efficient* algorithm.

The proofs of many great combinatorial theorems, particularly in graph theory, have this stigma that the proofs take recourse to case analysis. This situation itself seems to be unavoidable, when it comes to reckoning with collections of configurations made from finite sets. Another striking result, proved recently is the *strong perfect graph conjecture*, denoted SPGC. A graph $G = (V, E)$ is said to be *perfect* if the size of the maximum independent set (clique) is equal to the minimum number of cliques (independent sets) to cover for every induced subgraph of G. The SPGC states that a graph is *perfect* if and only it does not contain subgraphs homeomorphic to an odd cycle or an odd anti-cycle. This proof is not as cumbersome as the planarity proof and did not need the computer in its act. However, the proof is quite intricate and does involve case analysis arguments. This is typical of many theorems of graph theory which are based on what is known as a *forbidden subgraph* characterization.

I will not go into the details of this beautiful general theorem. I emphasize that a large number of interesting classes of perfect graphs arise in many practical applications. There exist polynomial time algorithms for many interesting graph theoretic optimization problems, such as colouring, covering, finding independence number, etc., on many subclasses of graphs, while these problems on general graphs are in general *hard* problems. In fact, this area is a fertile source of imaginatively solved problems for both mathematics and algorithmics.

Once again, I am fascinated by this dichotomy in metaphors. A theorem is beautiful, but its proof is not so elegant. However, the theorem and its implications lend themselves to polynomial time algorithms for many problems which are otherwise hard in general. Occasionally a theorem and its proof are both beautiful and elegant. Ironically, the proof may not be *effective* or the consequent algorithm may not be *efficient*.

A departure from this kind of bipolar situation has been the recent brilliant result of Manindra Aggarwal and his students at IIT, Kanpur, who showed that testing the primality of an integer can be carried out in deterministic polynomial time (see the authors' web page at IIT Kanpur and also[15]). This beautiful result uses a simple machinery of testing certain polynomial congruences. The work involved in checking these congruences is about the sixth power of the length of the integer to be tested. Of course the irony of algorithmic results is present in this situation too. There are certain probabilistic algorithms for testing primality of an integer which are used in practical applications, that are more efficient than this algorithm at the present.

The second class of questions in combinatorial mathematics is the *enumerative* kind. The questions are about the number of configurations. There are a whole lot of techniques based on combinatorial counting principles, generating functions, recurrence equations, Polya's theory of counting for inequivalent configurations etc. (see van Lint[16]). Suffice to say, that this body of knowledge is intimately connected with the science of algorithms in the area of analysis of algorithms. One is interested in the numbers of configurations to determine the average running time of an algorithm over all possible configurations.

Finally, the third class of questions is purely *constructive*. One is interested in *generating* the entire set or sub-sets of configurations. For example, collections of spanning trees, shortest paths, have further algebraic properties of theoretical and practical interest. The algorithmic computer scientist is interested in developing efficient algorithms to determine these configurations. It is here that the coming together of the two types of thinking, *mathematical* and *algorithmic* brings to a resounding success new ideas, thoughts, and results that enrich the beautiful corpus of results in both mathematics and computer science.

Many combinatorial proofs are also constructive in nature. They are indeed ingenious and sometimes surprisingly simple. I cite two particularly charming results from two different fields of mathematics, the first from graph theory and the second from number theory. (i) The necessary and sufficient condition for the maximality of a matching M in a bi-partite graph $G = (V_1, V_2, E)$ is the non-existence of any alternating path for M beginning and ending in unmatched vertices of M. (ii) The number $M_p = 2^p - 1$ is prime for a prime exponent p if and only if $v_{p-1} = 0$ in the sequence of integers $(v_1 = 4; \ldots v_i^2 = v_{i-1}^2 - 2; \ldots\}$, modulo M_p. The proofs (see van Lint[16] for the first and Knuth[10] for the second), of both these theorems are constructive in nature and both proofs lead to elegant, efficient polynomial time algorithms.

There are many combinatorial configurations which arise as *duals* of each other and often the counts of such configurations get related by what are known as extremal results. These results are typically of the form of maximum number of one type of configurations being equal to the minimum number of the other type of configurations. Such theorems are also known as *min-max* theorems. I indicate a few such theorems from the beautiful lore of these combinatorial gems. I refer the reader to (ref. 17) for a collection of beautiful results in combinatorics by the original authors and the recent book[18] on extremal combinatorics for more details. In fact the following theorems are also mutually related.

In a bi-partite graph (which we encountered above), the cardinality of a maximum matching is equal to the size of a minimum vertex cover. This theorem, named after *Konig*[19] and *Hall*[20] has many *avatars*. They are: the theorem of *Menger* that the maximum number of edge disjoint paths is equal to the minimum size of a vertex cut; of *Ford and Fulkerson* that the size of the maximum flow is equal to the size of the minimum cut, of *Konig*[19] on 0–1 matrices, that the maximum number of *disjoint* 1 elements in the matrix is equal to the minimum number of lines required to cover the 1 elements. The theorem of Dilworth[21], that in a partially ordered set, the maximum size of an antichain is equal to the minimum number of chains required to cover the poset, is also similar.

Coincidentally, there exist *elegant* polynomial time algorithms to determine all these combinatorial invariants. In fact almost all of these have constructive proofs that can be converted to algorithms. It is believed by researchers in this field, that combinatorial duality, min-max theorems and polynomial time solvability are intimately related from a philosophic point of view.

There is another metaphor in combinatorics known as a bijective proof. This typically arises in enumerative combinatorics where the equinumerous nature of two types (duals) of configurations is to be established. For example, many integer partition identities, lattice point counting paradigms, correspondences in tableau admit exquisite bijective proofs. These proofs have a close resemblance to the principles of design of efficient algorithms. Thus we see that mathematical and algorithmic ideas get entwined in many combinatorial contexts in resplendent ways.

Conclusions

Many applications draw heavily upon existing body of mathematical results and occasionally demand new mathematics. Contemporary computing science provides a new form of engendering new mathematical results. It provides new ways of looking at classical results.

I just mention a few exciting contemporary developments in both mathematics and computer science that have many connections based on deep mathematical and algorithmic thinking – Ramanujan's modular functions and expander graphs, computation of trillions of digits of constants like ***p***; Ramanujan's asymptotic methods in combinatory analysis[22] and their implications in analysis of algorithms; the theory of elliptic curves (one of the three ingredients in the proof of Fermat's Last theorem by A. Wiles) and its role in modern public key cryptography[15,23]; new algebraic and number theoretic questions arising out of cryptography and coding theory[24] such as the use of number fields in integer factoring, divisors of algebraic curves in cryptosystems and codes[15]; lower bounds on computational complexity in algebraic models of computation; pseudo-randomness in algorithms and mathematics; many combinatorial optimization questions stemming out of computer algorithms for applications in the computer world of networks, circuit design; many new algorithmic questions in linear algebra with new applications.

Thus we live in an era when the two disciplines of mathematics and computer science have set up many strong interactions. These interactions and their interplay are leading to the enrichment of both disciplines. Together they may provide the right force multiplier effect to study and answer some of the deepest questions bothering mankind – of cognition, self, and thought.

1. Knuth, D. E., *The Art of Computer Programming Vol. 1 Fundamental Algorithms*, Addison Wesley, New York, 1973.
2. Dijkstra, E. W., *A Short Introduction to the Art of Programming*, Computer Society of India Press, Mumbai, 1977.
3. Appel, K. and Haken, W., Every planar map is four colourable. *Illinois J. Math*., 1977, **21**, 429–567.
4. Manna, Z., *Logics of Programs*, Addison Wesley, 1987.
5. Knuth, D. E., Algorithmic thinking and mathematical thinking. *Am. Math. Monthly*, 1986, **81**, 322–343.
6. Edmonds, J., Paths, trees and flowers. *Can. J. Math*., 1965, **17**, 449–467.
7. Garey, M. R. and Johnson, D. S., *Computers and Intractability – A Guide to the Theory of NP-completeness*, W. H. Freeman, San Francisco, 1979.
8. Lemmermeyer, F., *Reciprocity Laws: From Euler to Eisenstein*, Springer, New York, 2000.
9. Edwards, H. M., *Fermat's Last Theorem*, Springer, New York, 1977.
10. Knuth, D. E., *The Art of Computer Programming Vol. 2 Seminumerical algorithms*, Addison Wesley, New York, 1981.
11. Yan, S. Y., *Number Theory for Computing*, Springer, New York, 2000.
12. Knuth, D. E., *The Art of Computer Programming Vol. 3 Sorting and Searching*, Addison Wesley, New York, 1981.
13. Aho, V., Hopcroft, R. E. and Ullman, J. D., *Design and Analysis of Algorithms*, Addison Wesley, New York, 1982.
14. Aigner, M. and Ziegler, G. M., *Proofs from The Book*, Springer, New York, 1998.
15. Abhijit Das and Veni Madhavan, C. E., *Public Key Cryptography: Theory and Practice*, Manuscript of a forthcoming book, 2005.
16. van Lint, J. H. and Wilson, R. M., *A Course in Combinatorics*, Cambridge University Press, London, 1992.
17. Gessel, I. and Gian-Carlo Rota (eds), *Classic Papers in Combinatorics*, Birkhauser, 1987.
18. Jukna, S., *Extremal Combinatorics with Applications to Computer Science*, Springer, New York, 2001.
19. Konig, D., *Math. Ann*., 1916, **77**, 453–465.
20. Hall, P., On representatives of subsets. *J. London Math. Soc*., 1935, **10**, 26–30.
21. Dilworth, R. P., A decomposition theorem for partially ordered sets. *Ann. Math*., 1950, **51**, 161–166.
22. Hardy, G. H. and Ramanujan, S., Asymptotic formulae in combinatory analysis. *Proc. London Math. Soc*., 1918, **17**, 75–115.
23. Koblitz, N., *A Course in Number Theory and Cryptography*, Springer, New York, 1996.
24. MacWilliam, F. J. and Sloane, N. J. A., *The Theory of Error Correcting Codes*, North Holland, New York, 1977.

Mathematics in engineering

Anindya Chatterjee

Department of Mechanical Engineering, Indian Institute of Science, Bangalore 560 012, India

I try to convey some of the variety and excitement involved in the application of mathematics to engineering problems; to provide a taste of some actual mathematical calculations that engineers do; and finally, to make clear the distinctions between the applied subject of engineering and its purer parents, which include mathematics and the physical sciences. Two main points of this article are that in engineering it is approximation, and not truth, that reigns; and that an engineer carries a burden of responsibility that mathematicians and scientists are spared.

What is engineering?

ENGINEERS have done a poor job of defining who they are, at least in India.

Most people are aware of the excessively many engineering colleges of our country, producing engineers yearly in lakhs for an economy needing only a fraction thereof.

However, the employed and capable engineers who design and build rockets, satellites, cameras and missiles for organizations like ISRO or DRDO are called scientists. Engineers who work on industrial R&D for companies like, say, Tata Motors (which recently changed its name from one that included the word 'engineering') are called managers. Engineers who develop brand new products and sell them successfully are called entrepreneurs. Engineers who program computers are called IT professionals. Any engineer who achieves something risks being called by another name!

Yet, engineering is exciting and worthwhile. Its broad sweep encompasses physics, chemistry, biology, mathematics, economics, psychology and more. It is the name for activity geared towards the purposeful exploitation of the laws, forces and resources of nature, not merely towards uncovering further esoteric truths but towards a direct improvement of the human condition.

In an idle moment, the reader may find it amusing to look on the web for a definition of engineering. I suggest a Google search for 'what is engineering'. Here are three I found:

1. Derived from the Latin ingenium, engineering means something like brilliant idea, flash of genius. The word was created in the 16th century and originally described a profession that we would probably call an artistic inventor.
2. A general term which refers to the systematic analysis and development of measurable physical data, using applied mathematical, scientific, and technical principles, to yield tangible end products which can be made, produced, and constructed.
3. The use of scientific knowledge and trial-and-error to design systems.

Of these the first is, irrationally, gratifying (as if the Latin source makes all engineers ingenious). The second is reasonable but wordy. The third points to the tremendous role of trial and error in design; the multitude of ideas that are tested, with most subsequently rejected; and the systematic use of ideas from proven, successful, designs. I mention this here because the rest of this article is focused on the 'use of scientific knowledge'.

And science, and mathematics?

The physical sciences are concerned with the truths of nature, and the laws that govern the world. Mathematics is a beautiful subject that, though inspired by the study of the world, does not depend on it: it can exist by and grow within itself. These subjects are pure. Engineering, in contrast, is *not* pure.

What thickness of reinforced concrete is sufficient for the roof of a 3 m wide tunnel through a mountain? What is a safe wall thickness for a lead container of given diameter used to carry radioactive waste so that it can, say, sustain a drop from a given height? How many tons of airconditioning are needed to maintain a temperature of 20°C in an office of 200 square meters floor area, 42 people, 35 personal computers, 26 windows and 6 doors, if situated in (say) Hyderabad? How much rocket fuel is needed to transport one kilogram of gold to the moon?

These are *technological* questions, as opposed to scientific or mathematical questions. They are faced by engineers, as opposed to scientists and mathematicians. They neither seek fundamental truth about nature, nor require some pure standard of intellectual rigour in their answers. They also supply incomplete information and leave room for the engineer to make simplifying assumptions, develop models, carry out calculations, draw on prior experience, and use safety factors where applicable, to obtain reasonable answers with reasonable effort. And there is a human price to be paid for a wrong answer, be it the loss of life, health, comfort or gold.

The process of training an engineer to answer such questions requires a study of engineering models and the mathematical techniques used to analyse them. Those models,

e-mail: anindya@mecheng.iisc.ernet.in

though approximate, require correspondence with reality in their conception, and precision in their description. And those mathematical techniques, like all mathematical techniques, require practice, sophistication and rigour. In this way, the technological world of an engineer builds up from the purer disciplines of mathematics and the sciences, but is not contained in them.

Engineering mathematics: everywhere, everyday

From stress analysis of machine components (using finite element packages), to numerical descriptions of the artist-drawn shapes of new gadgets (using CAD packages), to the use of numbers associated with the mundane jobs of production, inspection, and statistical quality assurance (using statistical packages), to the economically critical planning problem of what material to buy in what amount from where (using optimization packages), and so on, applied mathematics is everywhere in the everyday world of software applications in routine engineering.

From calculations of heat and mass flow in steam power plants and car radiators, to calculations of air flow in cooling fans, to calculations of molten metal flowing and mixing in weld pools, applied mathematics turns the wheels of engineering analysis and design.

From reliability in electrical power system grids to traffic in networks (both tar roads and optical fibres), mathematics crosses boundaries in a way no other technical subject can.

The applications mentioned above are the subjects of many books. Yet, they collectively fail to convey the excitement that engineering applications of mathematics can have. There is more to the story than a list of applications.

The workhorses of computational analyses

Any serious discussion of mathematics in modern engineering must involve the role of computers. Computers are good at moving numbers around and doing arithmetic with them; and they excel at doing these things with large *arrays* of numbers. Coincidentally or consequently, almost all big mathematical problems in engineering are somehow reduced to manipulation of, and arithmetic with, large arrays. In the engineering mathematics context, these arrays are called *matrices*.

The rest of this section concentrates on matrix calculations. For the nonmathematical reader, let me say this section discusses the three important problems of matrix-based calculations. These are called the linear homogeneous system, the standard linear system, and the linear eigenvalue problem. Of these three, the first one is important mostly in understanding the other two problems. The second is central: a large number of applied problems with apparently nothing in common are considered solved when reduced to the standard linear system. The third is independent, and almost as important: it is crucial in understanding vibrations, stability, and more generally sets of solutions that are peculiarly special for the system under study. For a concrete example, consider an engineer designing a bridge: the linear homogeneous system plays a role in deciding whether the bridge can bear loads at all; the standard linear system is used to calculate the deformation or deflection in the bridge when carrying, say, a bus; and the linear eigenvalue problem summarizes vibrations of the bridge.

What follows in this section is somewhat technical. Nontechnical readers may skim or skip as needed.

Three problems

There is much applied work that can be fruitfully done through an understanding of the following three equations (see note 1).

$$Ax = 0, \tag{1}$$

$$Ax = b, \tag{2}$$

$$Ax = lx. \tag{3}$$

In the above, A is a matrix, x and b are column matrices, and l is a number (possibly complex). Of these, x and l are usually unknown, and have to be found when the others are given. No serious computational work in the traditional areas of engineering involving the physical sciences is possible without a good understanding of at least one of these equations. In many advanced problems, one has to use a sequence of steps, each of which has something to do with one of these three equations.

The linear homogeneous system (1)

The equation $Ax = 0$, with A an $m \times n$ matrix (m rows, n columns), is usually associated with the question of whether x can be nontrivial (nonzero). If $n \leq m$, then the existence of nonzero x implies that the columns of A are not linearly independent, and A is said to be rank deficient: in particular, if $n = m$, then A is singular.

Equation (1) is relevant to eqs (2) and (3) above. If $Ax = 0$ has nonzero solutions, then eq. (2) does not have unique solutions (note 2). Also, in eq. (3), we write $(A - lI)x = 0$ to obtain eq. (1) (here I is the identity matrix).

Equation (1) is needed for understanding when a system is controllable and when it is observable (there are formal definitions of these terms[1]). In these applications, the relevant coefficient matrices A should be of full rank, i.e. there should exist no nontrivial solutions for x.

With numerical roundoff or measurement error

As suggested above, an important property of a matrix is its rank, which equals the number of linearly independent columns it has. For example, if

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 0 & 2 \\ 0 & 4 & 4 \end{bmatrix},$$

then its rank is 2 (because the third column is the sum of the first two, and so linearly dependent on them).

In numerical work with large matrices that are nearly rank-deficient, the border between invertible and rank-deficient matrices becomes blurred. An important tool here is the singular value decomposition (SVD), which is a characterization of a matrix in terms of some non-negative numbers called its singular values, and some vector directions called its singular vectors. The number of strictly positive singular values equals the rank of the matrix[2,3].

For a numerical example, consider

$$A = \begin{bmatrix} 1 & -1 & 0.02 \\ 2 & 0 & 1.98 \\ 0 & 4 & 4.01 \end{bmatrix}.$$

In exact arithmetic, the rank of A is 3. However, the third column is clearly almost equal to the sum of the first two columns. So the rank is somehow close to 2 (a hazy notion). More concretely, the singular values of A are (from Matlab) {5.8962; 2.6898; 0.0164}. The third singular value is much smaller than the second.

Suppose that, in an experiment, an engineer is trying to characterize the vibrations of a platform (see Figure 1). She has placed several accelerometers on the platform, and measured their output while some machinery on it was running. She may now ask, is the platform effectively rigid? If not, how many independent types of vibrational motions does it have? These questions, which represent noise-polluted versions of the linear homogeneous system, can be tackled using the SVD.

The standard linear system (2)

The equation $Ax = b$ is the backbone of engineering calculations. $Ax = 0$ impinges on it largely to the extent of understanding or eliminating nonuniqueness of solutions.

The commonest situation involves A square and invertible, in which case there is a unique and exact solution which can be found using accurate algorithms that are guaranteed to work. Much ingenuity in engineering has been expended in casting important problems into this form (and, for large systems, storing the matrices and solving the equations iteratively).

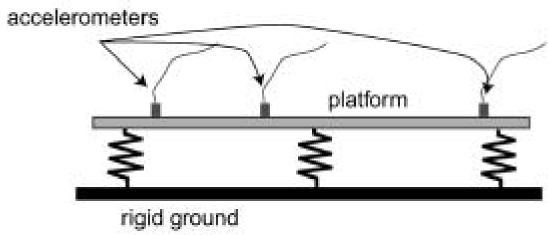

Figure 1. A vibrating platform.

The somewhat less common but also important case where A is $m \times n$ with $n < m$ is called an overdetermined system. It is not solvable exactly unless b satisfies special conditions, but can be solved approximately in a *least squares* sense by solving the *normal* equations

$$A^T Ax = A^T b,$$

where the T-superscript denotes transpose and $A^T A$ is $n \times n$, i.e. square. In applications, sometimes A is deliberately made overdetermined in order to get a better overall fit for some inexact model, and the resulting equations are solved in a least squares sense as above. In numerical work with roundoff errors, there are nominally equivalent methods that in fact keep accumulating roundoff errors under tighter control[2].

Let us briefly consider an overdetermined system. Consider two numbers p and q. We are told:

1. the sum of the numbers is 6,
2. the second number is twice the first one, and
3. the second number is 3 more than the first one.

The first two conditions above imply $p = 2$, $q = 4$ (provided we ignore the third condition). Similarly, the second and third conditions (ignoring the first) inconsistently imply $p = 3$, $q = 6$. And the first and third condition imply $p = 3/2$, $q = 9/2$. There is no choice of p and q that satisfies all three conditions. Any two of these conditions uniquely determine p and q; and the extra inconsistent condition makes the system overdetermined. To connect with the matrix algebra, we can write the three conditions as

$$p + q = 6; \qquad q = 2p, \quad \text{and} \quad q = p + 3,$$

which in matrix notation is

$$\begin{bmatrix} 1 & 1 \\ -2 & 1 \\ -1 & 1 \end{bmatrix} \begin{Bmatrix} p \\ q \end{Bmatrix} = \begin{Bmatrix} 6 \\ 0 \\ 3 \end{Bmatrix}.$$

Since the above matrix equation has no solution (being overdetermined), we might use the normal equations

$$\begin{bmatrix} 6 & -2 \\ -2 & 3 \end{bmatrix} \begin{Bmatrix} p \\ q \end{Bmatrix} = \begin{Bmatrix} 3 \\ 9 \end{Bmatrix},$$

whose solution is $p = 1.93$, $q = 4.29$. The reader may check for approximate satisfaction of all three conditions.

We now drop overdetermined systems and focus on square A.

A boundary value problem

Boundary value problems are very common in engineering. They involve differential equations along with boundary conditions. The simplest ones involve second order ordinary differential equations with two boundary points. Here is one such:

$$\frac{d^2 y}{dx^2} + xy = 1, \qquad x \in (0,1) \qquad y(0) = y(1) = 0. \tag{4}$$

A simple way to solve this problem computationally is to choose a large positive integer N, and then use a uniform mesh of points $x_k = kh$, with $k = 0, 1, 2, \ldots, N$ and $h = 1/N$. Let $y_k = y(x_k)$.

Then we simply write the following equations, where the derivative is approximated using sums and differences so that we eventually get the standard linear system:

$$y_0 = 0 \quad \text{from boundary conditions,} \tag{5}$$

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + x_k y_k = 1 \text{ for } k = 1, 2, \ldots, N-1, \tag{6}$$

$$y_N = 0 \quad \text{from boundary conditions.} \tag{7}$$

In the above equations, all quantities except the y's are known; and so they can be assembled to form eq. (2). This *finite difference* method can be extended to complex problems in 2 and 3 dimensions. It is conceptually simple but in higher dimensional problems requires care near tilted and/or curved boundaries.

An alternative solution technique is to assume an approximate solution of the form

$$y \approx \sum_{i=1}^{N} a_i f_i(x) = (\text{say}) \sum_{i=1}^{N} a_i \sin ip\, x.$$

Here, as indicated above, the f's are shape functions we choose in advance; and the a's are unknown coefficients that we will find. The specific choice of sines for the f's in this case respects the boundary conditions (this is important). We then proceed by substituting the approximation into

$$\frac{d^2 y}{dx^2} + xy - 1$$

to obtain what is called a residual, say $r(x; a_1, a_2, \ldots, a_N)$. The next step is called a *Galerkin* projection or the method of *weighted residuals*[4], and involves multiplying by each of the shape functions, integrating over the domain, and setting the result to zero, i.e.

$$\int_0^1 r(x; a_1, a_2, \ldots, a_N) \sin kp\, x dx = 0, \quad \text{for } k = 1, 2, \ldots, N.$$

This again results in a system of the form of eq. (2). (These calculations can be conveniently done using symbolic algebra packages like Maple or Mathematica.)

Some numerical results for both the above methods are given in Figure 2, for $N = 6$ and $N = 10$. Both methods perform well. The finite difference results are interpolated using broken lines for better visibility alone; only the nodal values (at the corners) are to be used for comparison.

The method of weighted residuals can also be used with shape functions f that are zero everywhere except inside some small regions; this leads to the powerful and versatile *finite element* method, well suited for complex geometries, but too specialized for this article.

Signs of trouble

Let us now consider the similar looking boundary value problem

$$\frac{d^2 y}{dx^2} + 19xy = 1, \qquad x \in (0,1) \qquad y(0) = y(1) = 0. \tag{8}$$

Numerically solving this using the finite difference method as described above, we obtain the results shown in Figure 3. Now the solution converges sluggishly, and fairly large N is needed before a reliable solution is obtained. Why? A partial answer lies in the singular values of the coefficient matrix A for each N. These values are plotted for

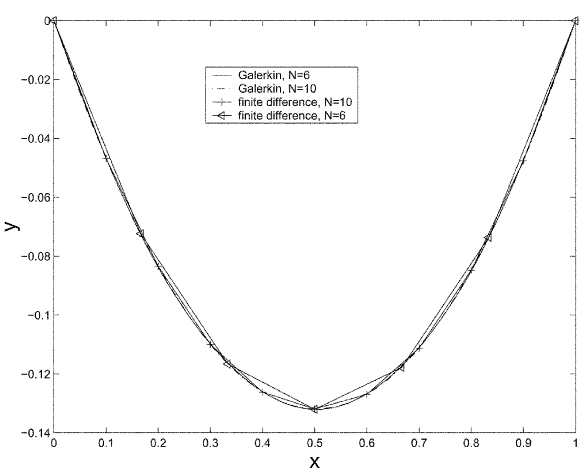

Figure 2. Solution of eq. (4).

$N = 80$, 160 and 320 in Figure 4. In each case, it is seen that the smallest few singular values drop off rapidly in magnitude, with the smallest one in each case being much smaller than the second smallest. Although the smallest singular value is nonzero for each N, and therefore A is invertible, the numerical shadow of singular A in eq. (1) has fallen on the calculation! We now proceed to eq. (3).

The linear eigenvalue problem (3)

In studying uniqueness of solutions for equations like eq. (8), we seek nonzero solutions of

$$\frac{d^2 y}{dx^2} + mxy = 1, \qquad x \in (0,1) \qquad y(0) = y(1) = 0. \tag{9}$$

Here, m is a parameter that is as yet unknown.

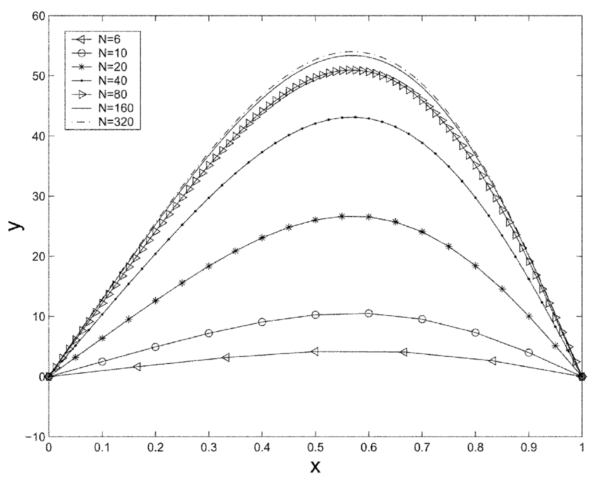

Figure 3. Solution of eq. (8).

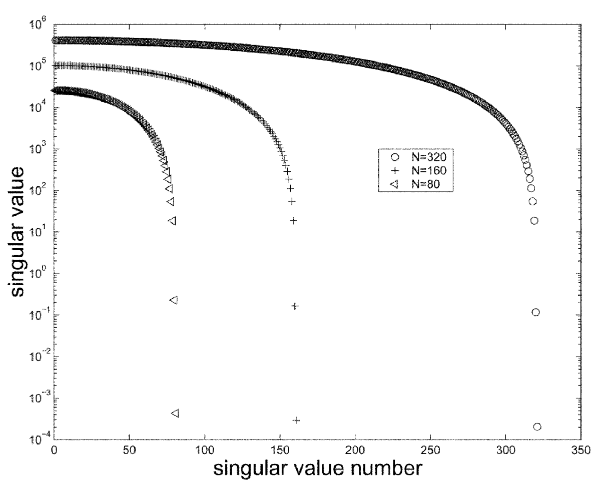

Figure 4. Singular values of A from eq. (8) for different N.

The finite difference equations can now be written in the form

$$Ay = mBy$$

which can in turn be rearranged to

$$A^{-1}By = \frac{1}{m} y = \text{(say)}\ l\ y,$$

which matches eq. (3). The numerically obtained values of l (from Matlab) for the case of $N = 10$ are

0, 0, 4.19E – 4, 9.17E – 4, 1.42E – 3, 1.92E – 3, 2.50E – 3, 3.52E – 3, 5.83E – 3, 1.27E – 2, and 5.33E – 2,

where 'E – 4' denotes '$\times 10^{-4}$' and so on. The estimate of m corresponding to $1/(5.33 \times 10^{-2}) = 18.8$, which should be compared with the coefficient in eq. (8). The actual eigenvalue of eq. (9) is even closer to 19, as may be found using larger N.

Eigenvalue problems arise in many places: in linear vibrations, buckling problems, systems of linear differential equations, and stability calculations in a variety of situations, to name a few.

In discussion of the above three equations (1–3) I have not touched upon the many strategies needed and developed to tackle very large and/or sparse systems, as well as systems with special properties such as A being symmetric and positive definite; these topics, too specialized for this article, are important in applications.

More mathematics

There is more to mathematics in engineering than the matrix calculations of mentioned in the previous section. For flavour, I will discuss two topics: one computational or numerical, and the other analytical or symbolic.

Nonlinearity

Nonlinearity is everywhere. Yet much of undergraduate engineering education in India concentrates exclusively on linear problems. The problems discussed in the previous section were all linear as well.

We now consider a nonlinear problem, and carry out some calculations. The trivial example discussed below hardly matches the many difficult nonlinear problems that engineers, scientists and mathematicians solve, but highlights some key points: local approximation by a simpler problem, small steps, and iteration or sequential refinement.

Consider

$$Ax + (x^T x)Bx = b,$$

where A and B are $n \times n$ matrices, b is $n \times 1$, and the unknown x is $n \times 1$ as well. The above problem is clearly a departure from eq. (2), and is chosen for ease of presentation: it represents no particular physical application.

One way to solve this problem is iterative, using Newton's method. Say at some stage our estimate of the solution is x_k; our next and improved estimate $x_{k+1} = x_k + \Delta x$ (say), should ideally satisfy the original equation

$$A(x_k + \Delta x) + (x_k^T x_k + 2x_k^T \Delta x + \Delta x^T \Delta x)B(x_k + \Delta x) = b.$$

Simplifying matters by ignoring terms proportional to $||\Delta x||^2$ and smaller, we obtain

$$x_{k+1} = x_k - (A + 2Bx_k x_k^T + x_k^T x_k B)^{-1}\{(A + x_k^T x_k B)x_k - b\}. \tag{10}$$

The above iterative procedure can be begun with an initial guess x_0; and if x_0 is sufficiently close to the correct solution, then the iteration will converge (under some technical conditions which usually hold).

As a particular example, taking

$$A = \begin{bmatrix} 1 & 2 \\ -3 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 3 & 1 \\ -1 & 2 \end{bmatrix} \quad \text{and} \quad b = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix},$$

and starting with

$$x_0 = \{1, 0\}^T,$$

we find the solution converges rapidly to

$$x = \{0.1502, 0.3623\}^T.$$

In general and large nonlinear problems, there may be multiple or no solutions; many guesses for x_0 may fail to produce convergence; and there may be severe difficulties with estimating or even repeatedly using the appropriate linearized update scheme. However, methods based on linearization (exact or approximate) and the above iterative scheme (possibly with simplifying approximations) are powerful and widely used.

Sometimes a good guess for x_0 might be so hard to find that we are reduced to using *continuation methods*. These involve solving a simpler problem; using its solution as the initial guess for a slightly changed problem; and using *its* solution for a further changed problem; and so on, until the simpler problem is changed completely to the original problem. These and other tricks, many of them tailored to suit specific applications, are an important part of the overall attack on nonlinear problems.

Asymptotic approximations

Asymptotic approximations or perturbation approximations (see, e.g. refs 5, 6) can be useful for certain engineering problems. I will give an example of such approximations, both to display them in their own right as well as enable comparison with some non-asymptotic approximations that follow.

Consider the equation

$$ex^5 + x^2 -$$

where $0 < \ \ll 1$ is small and known, and we seek x. The smallness of e allows approximations. Noting that for $e = 0$ there is the solution $x = 1$, we can pose a series of the form

$$x = 1 + \sum_{k=1}^{\infty} a_k e^k.$$

By a routine procedure involving collecting and equating terms, we obtain

$$x = 1 - \frac{1}{2}e + \frac{9}{8}e^2 - \frac{7}{2}e^3 + \frac{1615}{128}e^4 + \cdots.$$

If we fix the number of terms up to those shown and take e smaller and smaller, the error will eventually be proportional to e^5. For the particular value $e = 0.05$, the actual root is 0.977441 while the asymptotic approximation agrees well, at 0.977454. This sort of perturbation expansion is called regular, as opposed to singular (more interesting for specialists).

Asymptotic techniques have an uncomfortable place in modern, computer-empowered engineering. Yet, many good engineers, even ones not mathematically inclined in the usual sense, think in a way that keeps track of the relative orders of magnitude of physical features and effects, using these in an intuitive way that loosely resembles asymptotic approximation.

Clever insights

Clever insights abound in any mature subject where complex phenomena are viewed in simplified forms for better understanding (i.e. where truth is traded for understanding). These insights help us to both construct approximate solutions to seemingly difficult problems, as well as to suitably interpret numerical results from, say, a finite element code, and sometimes even to simply find a useful approximate solution to a mathematical problem which might otherwise be tedious. I will present here an example of each. These examples are drawn from within mechanical engineering due to my own limitations. They are, unavoidably, somewhat technical; however, at least the third example is free from equations.

Impact between elastic spheres

The first example involves the low velocity impact of two elastic spheres, for which the Hertzian theory of contact was developed (see ref. 7; and references therein). The Hertzian contact analysis shows how a seemingly formidable problem can, by clever insights, be brought into a tractable yet meaningful form. In contrast, rigorous mathematical treatment of this problem raises many difficult issues.

Hertz's approach proceeds by noting that the size of the contact region is small compared to the size of the sphere; and that the impact occurs over a time that is long compared to the time period of vibrations in the sphere. Thus, impact proceeds as if contact between two point masses is mediated by a nonlinear spring. The behaviour of this nonlinear spring is given by

$$F = Cd^{3/2}, \tag{11}$$

where F is the force, d is the compression, and C is a known constant depending on sphere material and size.

For those not interested in the finer details of Hertzian contact, here is an alternative and simple way to determine (or at least understand) eq. (11). In Figure 5, an elastic sphere of radius R is pressed into a rigid surface by an amount d (the undeformed sphere is drawn, to emphasize the amount of compression). The resulting contact patch has a radius of approximately r which satisfies, for small d,

$$r = \sqrt{2Rd}\,.$$

Now, assume that deformations are localized in a cylinder of diameter and height both equal to $2r$. The area is πr^2; and the strain is $d/(2r)$. F is E times strain times area,

$$F = E \times \frac{d}{2r} \times p\, r^2 = \frac{p}{\sqrt{2}} E \sqrt{R} d^{3/2},$$

which is off from the correct C of eq. (11) by less than a factor of 2.

Heated strip on a rigid substrate

Here is another example of nice insight into an engineering problem. Consider the system in Figure 6 (top). The figure shows the cross section of a rectangular strip perfectly bonded to a rigid substrate. The strip is heated, so that it tends to expand but is restrained by the non-expanding substrate. We are interested in the interface forces and stresses generated. This idealized problem is also relevant to piezoelectric actuators used to control vibrations in structures.

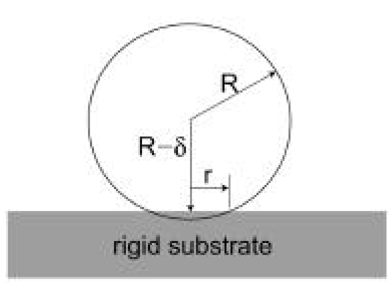

Figure 5. Contact between an elastic sphere and a rigid surface.

Those acquainted with the theory of elasticity will anticipate stress singularities near the edges of the bond. An approximate analysis based on simplifying insights can proceed as follows.

Since the substrate is rigid, the width of the strip is basically unchanged. The thickness of the strip is not important. For the strip to retain its lateral dimensions while staying flat, the net force from the substrate should be as sketched in the figure (middle), for some compressive force (per unit length in the direction into the page) that exactly counteracts the tendency to expand due to heating. Since the problem is actually three dimensional, there are similar forces at all the edges of the strip.

Taking the material's Young's modulus to be E, Poisson's ratio to be n, coefficient of thermal expansion to be a, the change in temperature to be ΔT, and the thickness of the strip to be h, we are led to conclude that

$$F = \frac{a\,Eh}{1-n}\Delta T.$$

We note, next, that the force F at the midplane of the strip (see figure (middle)), actually comes from the interface. The correct interface forces must therefore be the equivalent system shown in the figure (bottom), with

$$M = \frac{Eh}{2}.$$

For finer details, finer analysis is needed. With the above insight, from FEM simulations, we expect strong stress concentrations near the edges of the strip.

Testing a sprung bicycle

This example involves bicycles. For rider comfort, some bicycles frames are designed to flex. For example, two

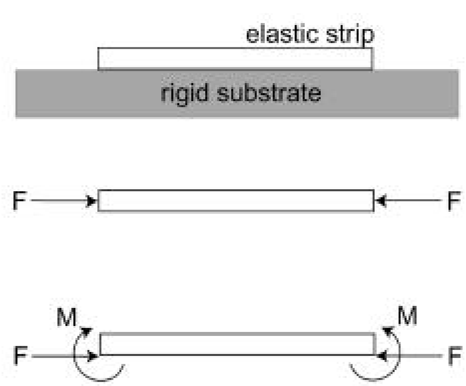

Figure 6. Heated strip on a rigid substrate.

rigid parts of the frame might be connected at a pivot, along with a spring to resist relative rotation there. Such a sprung bicycle is sketched in Figure 7. The pivot is at *P*.

A potential buyer may wish to check two design criteria. When the bicycle goes over a bump, the frame should flex. Yet, when the rider pedals hard, the frame should *not* flex (the rider's power should go into acceleration, not deforming the spring even while on a smooth road). A quick and approximate evaluation of the second aspect might proceed as follows.

Consider an accelerating bicycle. The rider exerts forces to cause this acceleration. Assuming the rider is significantly heavier than the bicycle, and using D'Alembert's principle[4] (see note 3), we need merely consider a static problem provided there is the correct horizontal backwards force acting approximately at the rider's center of mass. The rider ties a rope from his waist to a pillar and then, with rope taut, pushes on the pedal exactly as if accelerating hard. If the frame does not flex, the second criterion above is met and the design is good. Needless to say, a full analysis of the bicycle requires much additional work.

Modelling and approximation

I once read on some engineer's door the following jocular lines:

Engineers think theory approximates reality.
Physicists think reality approximates theory.
Mathematicians never make the connection.

Mathematicians can, and often do, work in a world that needs no contact with any physical reality. Physicists speak ardently about the fundamental and universal truths of nature. But in engineering, there is no special place for absolute truth. The flat-earth theory works if you want to build a bridge, and the point-mass-earth theory is good if you want to calculate the trajectory of the moon.

I am a mechanical engineer, and enjoy the subject of vibrations. So I will use vibrations to discuss a key idea. To that end, here is a crash course on vibration theory.

Figure 7. A sprung bicycle.

Mechanical vibrations

See Figure 8 *a*. A block slides on a frictionless surface. It is restrained by a spring. If disturbed from its equilibrium, the block oscillates or vibrates, to and fro. It is intuitively clear that stiffening the spring will make the block oscillate faster (with a higher frequency, or smaller time period), while increasing the mass of the block will make it oscillate slower (lower frequency, larger time period). Now consider Figure 8 *b*, where two such blocks are held in place by three springs. A multitude of apparently complex motions are possible for this system, but there are two special motions, corresponding to *normal modes* of vibration. In one of these, the two blocks vibrate in synchrony (*in phase*), and the spring in the middle plays no role: the system in this mode is merely two copies of the first system. In the second mode, the blocks vibrate exactly in opposition (*out of phase*), which really is another face of synchrony. These two normal modes are indicated schematically in Figure 8 *c* and *d*. The frequency of vibration in the first mode is lower than that in the second mode, where the middle spring gets called into play as well.

These ideas are broadly applicable. If the two masses or the springs are not identical then there are still two normal modes, only they are not so easy to sketch from intuition and must be found through equations (some of that comes in the next subsection). If there are more blocks (i.e. more masses), then there are more normal modes. One mode per block. If we consider longitudinal vibrations of a rod, where mass and stiffness are merged into a continuum of material, then there are infinitely many normal modes associated with as many, and steadily increasing, frequencies; of these, in engineering, only the first several (2, 10, 50, or more, depending on the application) may be considered important.

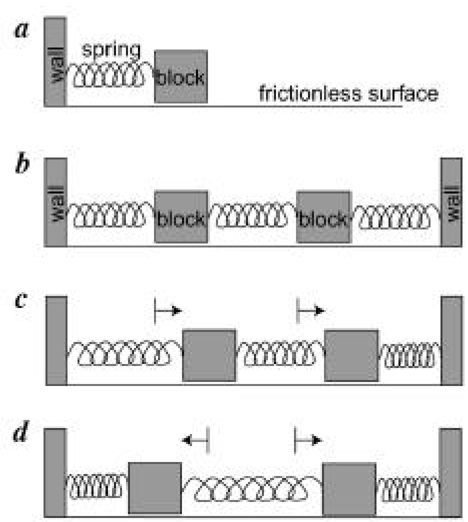

Figure 8. Vibration theory: normal modes.

An important idea remains: general motions are *linear combinations* of motions in the normal modes. See Figure 9. Let the first graph represent the displacement of the first mass in Figure 8 *c*, plotted against time. And let the second graph, showing more direction reversals within the same time (hence, higher frequency), represent the displacement of the first mass in Figure 8 *d*. A general motion is the sum of two such motions, each of arbitrary amplitude or size, as sketched in the third graph of Figure 9. From this third graph, it is not immediately obvious that there are motions at two frequencies simply added up; that insight comes from mathematical abstraction, though it soon becomes part of the physical world of every vibration engineer.

Much remains: energy dissipation through friction, effects of forcing, resonance, nonlinearity, design issues, and practical troubleshooting. But the heart of vibration theory lies in its first step: linear combinations of normal modes.

Modelling

Now, some actual calculations. Consider the system sketched in Figure 10. A steel rod of diameter $d = 0.01$ m, length $L = 1$ m, density $r = 7800\ \mathrm{kg/m^3}$, and Young's modulus (material stiffness) $E = 210$ GPa is built into a strong wall at one end and has a steel ball of diameter $D = 0.1$ m at the other. What are the time periods for the normal modes of *lateral* (sideways) vibrations in this system?

The *first* normal mode (and hence its time period) can be approximated by ignoring the mass of the rod, and treating the sphere as a point mass. In this engineering approximation, the rod simply acts as a spring. By a routine calculation, we obtain the time period as

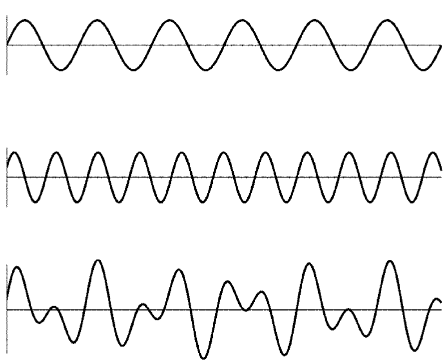

Figure 9. Vibration theory: general motions.

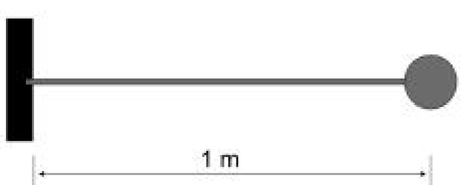

Figure 10. A cantilever beam with an end mass.

$$T = \frac{2p}{\sqrt{\dfrac{3EI}{mL^3}}} = 0.7221\,\mathrm{s},$$

where m is the mass at the end, and I is something called the second moment of area of the cross section (for details, see ref. 8).

We might approximately incorporate the beam's mass by adding an effective inertia to the point mass at the end. The mass m_r of the rod is (ignoring the overlap between the rod and the sphere)

$$m_r = p\,\frac{d^2}{4}\,L\,r = 0.6126\ \mathrm{kg}.$$

Not all of this mass moves equally: it is only near its right end that the rod moves as much as the sphere; so (skipping a more complex calculation) we might add on about $m_r/3$, giving

$$T = 0.7221\,\sqrt{1+\frac{m_r}{3\,m}}\ \mathrm{s} = 0.7399\ \mathrm{s}.$$

A more careful, and presumably accurate, answer can be found by treating the rod as a cantilevered Euler–Bernoulli beam[8] with a point mass at the end. Then the governing PDE is

$$EIu_{xxxx} + \frac{m_r}{L}\,u_{tt} = 0,$$

with boundary conditions

$$u(0, t) = 0; \quad u_x(0, t) = 0 \quad \text{(cantilever end)}$$

along with

$$u_{xx}(L, t) = 0 \quad \text{(zero end moment)};$$

and

$$EIu_{xxx}(L, t) = mu_{tt}(L, t) \quad \text{(force from end mass)}.$$

Now we are no longer restricted to the first normal mode, and can seek 'all' of them (within the approximations of *this* model, that is). The solution for the above, straightforward though tedious, is omitted here. The net result for the first time period is

$$T = 0.7347\ \mathrm{s},$$

a reasonable match with foregoing estimates.

More accurate estimates might be obtained using laborious calculations that incorporate the rotary inertia of the sphere, and then shear deformations and rotary inertia in the beam itself; and, eventually running out of simple theories, using the finite element method.

Approximation

The above sequence of increasingly accurate models notwithstanding, errors in understanding of boundary conditions, specification of geometry, measurement of material constants and incorporation of effects like damping will always limit the numerical accuracy to which the time period can be found. And models of more complex systems with, e.g. plasticity, fracture, frictional contact and impact will be less accurate still.

In engineering there is no absolute truth, no perfect model, and no grand unified theory. Ever finer FE meshes do not take us ever closer to something infinitely pure and true. But there is no grief in this. The world is infinite and our minds are limited. Truth and understanding are in conflict: a little truth lost is the price for a little understanding gained.

The burden of truth

Statements of results in technical subjects must bear what may be called their subject's burden of truth.

In mathematics, truth is nothing less than absolute proof. This can be very hard. However, the leeway allowed to mathematicians is that difficult problems can be attacked, by any and all, over a long period of time[9] for the story of a 350-year problem). In the interim, offshoots that lead to other truths, as well as partial results, are all acceptable contributions to the steady march of knowledge.

In physics, truth has a high place as well. But here, it is interpreted as consistency with every known relevant experiment of the past, pending possible invalidation by just one experiment, as yet unforeseen, that throws up an inconsistency. The more determined, exhaustive, and dramatic the search for that invalidating experiment is, the more satisfying and reassuring is the failure to find it. Limitations of time, space or money do not impinge on this philosophy.

In comparison, the engineer's burden of truth may seem lighter. The ceiling of that tunnel through the mountain should just remain intact as a million cars pass under; that radioactive material should just safely reach its resting place at the bottom of the sea; that office in Hyderabad should just remain comfortable even on crowded days in the middle of summer; and that gold should just reach its intended place on the moon. Once the design succeeds, and the system functions as needed, the burden of truth is met. The issue is not reopened (nor design fee refunded) when someone else does it better, faster or cheaper. The unique charm of mathematics in engineering lies in the many levels and forms in which it is invoked, revoked, used, abused, developed, implemented, interpreted and ultimately *put back in the box of tools*, before the final engineering decision, made within the allotted resources of time, space and money, is given to the end user.

Mathematicians are not blamed for failing to prove theorems. And physicists are not blamed for failing to disprove theories. In their subjects, truth has divine status; and its burden is borne by the gods themselves. But the engineer's burden of truth, though lighter in comparison, rests on human shoulders.

So, pause for a moment to consider this tidbit I read somewhere: in ancient Roman times, the builder of a bridge stood *under* it when the first chariot crossed over. And pause, again, for these lines from the *Hymn of Breaking Strain* by Rudyard Kipling:

The careful text-books measure
(Let all who build beware!)
The load, the shock, the pressure
Material can bear.
So, when the buckled girder
Lets down the grinding span,
The blame of loss, or murder,
Is laid upon the man.

Notes and references

Note 1. I have left out some other important problems. One certainly is the standard linear program, encountered in optimization. Another, perhaps, is the general nonlinear first order system of ordinary differential equations.

Note 2. Let $Ax_1 = 0$ with $x_1 \neq 0$, and let $Ax_2 = b$. Then $A(x_1 + x_2) = b$ as well, showing nonuniqueness. Nonunique solutions can disturb, say, an engineer trying to calculate the deflection of a bridge under the weight of a bus.

Note 3. Add $-m\mathbf{a}$ to each point mass m, where $\mathbf{a}$ is its acceleration; and then treat the system as static, not dynamic.

1. Franklin, G. F., Powell, J. D. and Emami-Naeini, A., *Feedback Control of Dynamic Systems*, Pearson Education, Singapore; Indian Branch: Delhi, 2002, 4th edn.
2. Golub, G. H. and van Loan, C. F., *Matrix Computations*, Johns Hopkins University Press, 1989, 2nd edn.
3. Chatterjee, A., An introduction to the proper orthogonal decomposition. *Curr. Sci.*, 2000, **78**, 808–817.
4. Finlayson, B. A., *The Method of Weighted Residuals and Variational Principles*, Academic Press, New York, 1972.
5. Hinch, E. J., *Perturbation Methods*, Cambridge University Press, Cambridge, 1991.
6. Bender, C. M. and Orszag, S. A., *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, London, 1978.
7. Goldsmith, W., *Impact: The Theory and Physical Behaviour of Colliding Solids*, Edward Arnold, 1960.
8. Clough, R. W. and Penzien, J., *Dynamics of Structures*, McGraw-Hill, London, 1975.
9. Singh, S., *Fermat's Last Theorem*, 1997, Fourth Estate, Paperback edn, 2002.

ACKNOWLEDGEMENTS. The first section of this article was prompted by conversations with B. Gurumoorthy. The Hertz contact approximation was shown to me by J. Papadopoulos. The heated strip analysis is close to something H. D. Conway showed me. The bicycle example is from J. Papadopoulos via A. Ruina, who has also much influenced the way I think about approximations in engineering. R. N. Goverdhan, D. Chatterjee and A. Ruina read the manuscript and helped me improve it. Kipling's *Hymn of Breaking Strain* is a good poem, but the rest of it is not about engineering.

Revisiting the 'unreasonable effectiveness' of mathematics

Sundar Sarukkai

National Institute of Advanced Studies, Indian Institute of Science Campus, Bangalore 560 012, India

Although the phrase 'unreasonable effectiveness of mathematics' is widely used, it is not clear what it means. To understand this phrase critically, we first need to understand the meaning of mathematics and what it means to use it in the sciences. This paper begins by considering the different views on the nature of mathematics, the diversity of which points to the difficulty in understanding what mathematics really is, a difficulty which adds to the mysteriousness of the applicability of mathematics. It is also not clear as to what is applied when we apply mathematics. What is clear however is that mathematics cannot be applied to the world but only to some descriptions of the world. This description occurs through the medium of language and models, thus leading us to consider the role of mathematics as language. The use of a language like English to describe the world is itself 'unreasonably effective' and the puzzle with mathematics is just one reflection of this larger mystery of the relation between language and the world. The concluding parts of this paper argue how the view of mathematics as language can help us understand the mechanisms for its effective applicability.

SOME words and phrases are destined to capture the imagination and in so doing get widely used. As a consequence, they are also open to serious misunderstanding. 'Unreasonable effectiveness of mathematics' is one such phrase which is often invoked but little analysed or understood. This phrase was made famous by Eugene Wigner in the Richard Courant Lecture in Mathematical Sciences at New York University in 1959, which was subsequently published in the *Communications in Pure and Applied Mathematics* in 1960. I will begin by summarizing Wigner's arguments in order to understand exactly what he meant when he used the phrase 'unreasonable effectiveness', after which I will analyse what mathematics and mathematization means, and then conclude with one explanation for the effectiveness of mathematics.

Wigner begins with a story of two friends, one of whom, a statistician, was working on population distribution[1]. When the statistician explained the symbol ð occurring in a particular distribution, the friend, who presumably was not a mathematician, thought it was a joke and said, 'surely the population has nothing to do with the circumference of a circle'. Wigner learns a lesson or two from this story. He first notes that mathematical concepts turn up unexpectedly thereby providing close descriptions of some phenomena. Secondly, he believes that because we do not know the reason why mathematics is so unexpectedly useful we will not be able to say with certainty whether a theory we hold true is uniquely appropriate to a phenomenon or not. With this as his starting point he analyses the usefulness of mathematics in the natural sciences and comments that this usefulness is mysterious and has 'no rational explanation' for it.

The significant use of mathematics in the sciences owes a great debt to the belief that the laws of nature are written in the language of mathematics, a statement attributed to Galileo and one which has been echoed for centuries after by figures such as Newton, Einstein and Feynman. Wigner too joins this chorus and begins by correctly noting that only some mathematical concepts are used in the formulation of laws of nature and these concepts are not chosen arbitrarily. One of the elements contributing to the mystery of mathematics lies in the physicist stumbling upon a mathematical concept that best describes a phenomenon only to find that the mathematician has already developed that concept independently. As examples, Wigner cites complex numbers and functions, the appropriateness of which is especially manifested in the formulation of the complex Hilbert space which is so essential to quantum mechanics. The surprising (to the common sense) and necessary role of complex numbers and functions along with the idea of analytic functions is one example of the 'miracle' of mathematization.

The important argument here is that mathematical concepts are not accidentally useful but are *necessary* in the sense that they are the 'correct language' of nature. Wigner offers three examples to illustrate this necessary relation. The first is that of Newton's law. Not only was this law based on 'scanty observations', it also contained the physically non-intuitive idea of the second derivative and yet exhibited an extremely high sense of accuracy. The second example is the matrix formulation of quantum mechanics. The miracle in this case, according to Wigner, lay in the fact that one could apply these matrix methods even in cases where Heisenberg's initial rules did not apply, as illustrated in the calculation of the lowest energy level of helium.

e-mail: sarukkai@nias.iisc.ernet.in

The third example is that of quantum electrodynamics, particularly the theory of Lamb shift, a theory which again showed extremely high accuracy with experiment. From this, Wigner concludes that mathematical concepts, 'chosen for their manipulability', are not only appropriate but are also accurate formulation of the laws of nature. For him, these laws together with the laws of invariance are the foundation of the mathematical method in sciences. Finally, he considers the uniqueness of theories in physics and asks whether mathematics alone can help adjudicate which theories are essentially right. The problem here is that some theories which are known to be false also give 'amazingly accurate results'. The examples he gives of these 'false' theories are Bohr's early model of the atom, Ptolemy's epicycles and the free-electron theory.

Wigner concludes by saying that the 'miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve'. I have no comments on whether we *deserve* this 'gift' or not but as for understanding it, we can at least make an honest try – and this many philosophers have done.

But what is mathematics?

Much of what Wigner says must perforce depend on what he means by mathematics. Wigner says little about what mathematics is but what he says is suggestive. Wigner writes, 'mathematics is the science of skillful operations with concepts and rules invented just for this purpose. The principal emphasis is on the invention of concepts'. This ability to create concepts takes the mathematician into unchartered realms to the point of being imaginatively 'reckless'. Further, there is a notion of generality, simplicity and beauty inherent in this creation.

Wigner, like many scientists, blissfully ignores some of the seminal contributions from philosophy to the understanding of mathematics. His view of mathematics, emphasizing the importance of rules and the human creative element in creating concepts and rules, runs counter to some dominant views on mathematics. Although Wigner does not explicitly push this point further, it is clear that his understanding of mathematics as being rule-driven makes the effectiveness of it a much greater mystery. Namely, how is an activity of humans, driven as it is by rules we create and with human-centred ideas such as beauty, so well matched with the natural world? Wigner was right in his characterization of mathematics even though his analysis of the question is incomplete.

It will be useful to briefly discuss the dominant views of mathematics before we consider the question of applicability. I will summarize five different views on the nature of mathematics. The divergence of these positions clearly suggests that the mysteriousness of applicability has its origins in the 'mysteriousness' of mathematics itself.

Platonism

Let me first consider the realist view of mathematics. Realists about mathematics believe that mathematical entities exist independently of humans just as trees and tables do. Platonism about mathematical entities is the dominant realist tradition. Platonists believe that mathematical entities have an existence independent of human minds. These entities inhabit a special world, the Platonic world. Platonism thus believes not only in the independent existence of mathematical objects and relations but also believes that the 'reality' of that world explains the universal nature of mathematical truth. However, Platonism, although popular among mathematicians and scientists, runs into serious problems when confronted with the applicability of mathematics. In this case, the basic problem is to understand how these Platonic entities, which do not have spatial or temporal characteristics, can get in touch with our physical world, which is defined by spatio-temporal extension. In other words, how do we as humans access these Platonic objects? And how do these objects link up with our real world?

Logicism

One dominant view of mathematics relates it intrinsically to logic. Logic elucidates the structure and validity of arguments. Reduction of mathematics to logic, in particular deductive logic, meant that the complete domain of mathematical activity was a logical one. Echoing this, the influential logician and philosopher Frege argued that 'mathematics was nothing but the systematic construction of complex deductive arguments', a view which has been dubbed the logicist view of mathematics[2]. Russell attempted to show that all mathematical concepts could be redefined in terms of purely logical concepts. The reduction of mathematics to logic would then imply, for Russell, that all of mathematics, including its axioms and postulates, could be derived entirely from logical laws. However, as Dummett notes, there are various problems in this reduction of mathematics to logic, including Zermelo's axiom of choice and the axiom of infinity[3]. Moreover, there was a serious problem even with a fundamental mathematical entity, the set. If logicism is right, then a set should be a logical concept. However, it was clear that a set was not a logical concept – one reason being that there are many incompatible axiomatizations of set theory.

Formalism

Another view of mathematics, influential in its own way, is called formalism[4]. This school was largely associated with the German school of mathematics and most notably with the illustrious mathematician David Hilbert. The basic idea in the formalist view of mathematics is that mathematics is nothing but a set of rules and formal manipulations of mathematical symbols and terms according to these rules. For formalists there are no meanings attached to mathematical objects, equations or operations over and beyond these meaningless formal manipulations, whether in proof or applications. An analogy that has often been made is that mathematics is like a chess game, which has its objects such as pawns, queen, king and so on, and rules of movement for each of these pieces. The formalist view of mathematics argues that there is no meaning to mathematics over and beyond the game which is played with these mathematical objects according to some given rules. Not only was Hilbert a strong proponent of this view but so was G. H. Hardy who believed that mathematics was just like chess. Moreover, Hardy describes even formal mathematical proof in terms of the structure of chess: ‘The *axioms* correspond to the given position of the pieces; the *process of proof* to the rules for moving them; and the *demonstrable formulae* to all possible positions which can occur in the game’[5]. The basic problem with formalism is that it seems difficult to accept mathematics as just a game; in particular its applicability to the sciences then seems totally arbitrary and forces us to ask, why is not chess applicable to the world like mathematics is? In fact, Frege believed that it is the applicability of mathematics alone that makes mathematics more than just a game. On the other hand, for Hardy, the very idea of applying mathematics was distasteful and he writes that mathematics which has practical uses is ‘on the whole, rather dull’ and has ‘least aesthetic va [6].

Intuitionism

In contrast to formalism is intuitionism[7]. The contrast is also illustrated in the nationalities associated with these two views. Intuitionism was predominantly influenced by the French while formalism was developed by German mathematicians. The father of intuitionism was the French mathematician Brouwer who (ironically?) drew upon the German philosopher Immanuel Kant’s ideas of intuition and *a priori* truth of mathematics. Intuitionism accepts the ‘obviousness’ of mathematical entities and places them on par with objects such as chairs and tables. It is in this sense that Godel says that we can perceive mathematical objects like sets in a manner similar to our perception of objects in our world. Godel suggests that ‘we do have something like a perception of the objects of set theory, as is seen from the fact that the axioms force themselves upon us as being true’[8]. This is an intriguing way of understanding perception; namely, perception of something is not the reason for it being true but recognizing the truth of something actually suggests its perceptability. However, mathematical intuitionism seems counter-intuitive, at least with regard to our common understanding of perception. The intuitionists find the idea of infinity problematic and Brouwer argues that the formalists’ approach to infinity and transfinite set theory is ‘meaningless’ since these are beyond the limits of mathematical intuitions. For the intuitionists, mathematics is something to be created and not discovered, and the role of a creator is best exhibited when the mathematician has to exhibit proof for all existential mathematical assertions.

Mathematics as language

Finally, let me consider the view that mathematics is a product of human imagination, is grounded in our experience with the world and functions like a language. First is the obvious fact that mathematics is a product of humans and is created through our interaction with the world. This implies that the world catalyses mathematical ideas, including the kinds of mathematical entities such as numbers, sets, functions and so on. For example, the mathematical principle of linearity illustrates the physical principle of superposition[9]. If we think of mathematics as beginning with numbers along with some operations like addition we can find an immediate link between human experience (including the activity of counting and aggregating), the structure of the world around us and mathematics. While this does not mean that every mathematical entity or operation is somehow connected to our activity in this world it suggests that the distance between mathematics and our world is not that far removed in the first place. And this relation between the world, humans and mathematics can be analysed in different ways. Steiner argues that anthropocentrism pervades applied mathematics[10]. Rotman argues that mathematics is not divorced from the world and mathematical concepts and objects arise first from the world[11]. Sarukkai has shown how various discursive strategies actually help to create mathematics as we know it now[12]. This view of mathematics offers a canonical answer to the puzzle of unreasonable effectiveness of mathematics. Part of the puzzle lies in the mysteriousness of the relation between two different kinds of worlds – the physical and the mathematical. But if we question the proposition that these are different worlds and argue that mathematics actually ‘arises’ from the world then the unnatural connection is no longer there, thereby diluting the puzzle as far as this relation is concerned. I will discuss this view of mathematics in much more detail towards the end.

So we find that there is no simple answer to what mathematics really is. This ambiguity about the scope and depth of mathematics gets transferred to the question of applicability. The mysteriousness that is enshrined in the phrase ‘unreasonable effectiveness’ of mathematics reflects as much a confusion about the mysteriousness of what mathematics is as much as its applicability. Moreover, there are many different types of applicability and different meanings to applicability. I will briefly discuss this issue in a later section.

Mathematization in modern science: lessons from the early days

How exactly did mathematization of the sciences begin? Historians trace the origin of the modern sensibility of mathematization from Galileo onwards although he was not the first person to use mathematics to describe the world. The

Greeks placed mathematics on the highest pedestal; as is well known, the golden section was one of the most privileged concepts in Greek art, architecture, ethics and science. Indian astronomy made extensive use of mathematics, as did Ptolemy. But what was special to Galileo was that he combined mathematics with experimentation, thereby justifying his being called the father of modern science.

Although Galileo radically changed some fundamental presuppositions, his effort nevertheless was built on work by others. For example, there were mathematical-philosophers predating Galileo who, among other things, had analysed the idea of motion in great detail. What Galileo did was to relook at the phenomenon of motion in order to describe it as faithfully as possible with the help of mathematics, corresponding to his belief that physical events were *describable* correctly by mathematics. Again what differentiated Galileo from other natural philosophers who used mathematics was his insistence that experiments were necessary to test and verify the mathematical description. This 'harmony between the world of experience and the mathematical form of knowledge, to be attained through experiment and critical observation'[13] was the unique contribution of Galileo.

Let me analyse one particular component of Galileo's method to illustrate why the method of mathematization seems to work so effectively. Galileo's mathematics was not calculus but number sequences. He discovered by his experiment on motion that the distance of free fall of an object is proportional to the square of the interval of time. How does mathematics manifest itself in this case? Let us assume that we have the necessary apparatus to do this experiment. We drop a ball and find the distance it travels after one second, two seconds, three seconds and so on. Just by noting the distance travelled, we can see a pattern, which is that the distance fallen is in multiples of 4, 9 and so on. Without needing to know any physical laws or calculus we can conjecture that distance varies as the square of the distance[14].

The basic point is this: a pattern about free fall motion is discernible by a particular kind of observation that measures some parameter, in this case distance. Neither the act of measurement nor the use of numbers constitutes mathematization of this problem. But what they do is to illustrate a pattern about motion which is not otherwise discernable. That the distance varies as time squared is of profound importance – this observation plays an important role in helping Newton postulate the gravitational force law as an inverse square law.

Suppose somebody claimed that we could as well have described the fall of the object in English instead of mathematics. So when asked to describe this free fall, this person could say that the object falls fast, faster and … Note that in using English we do not have the capacity to specify the relation between fast and faster. Mathematics, as a language, has this capacity to tell us something about relations. It can tell us that the distance fallen after two seconds is not only greater than the distance fallen after one second but that the distance is four times more. So the use of numbers gives us more information about the distances compared to the use of phrases such as 'greater than'.

But this still does not explain the mystery of mathematization. Suppose we had numbers but did not have multiplication or the concept of proportion. Then we can conceivably give values for distance fallen but we will find no proportional relation between them. Say an object falls 16 ft after the first second and 64 ft after the second second. Let us suppose (however improbable it may seem!) that our mathematics has no concept of multiplication and division but only addition. Then looking at these numbers we cannot find the law that distance varies as time squared. So just having numbers is not enough but we also need an appropriate set of operations. The question therefore is: if we discover new operations and new kinds of numbers would we be able to have a 'better' description of nature? But how do we know what operations are needed? Can nature tell us that? Or does mathematics first offer us this? Also, note that there are already prior physical concepts in use even in this simple problem. Even in a simple mathematical description there are many physical concepts which makes possible the mathematization. For example, before Newton's law can be written down in a mathematical form the physical ideas of force, mass and acceleration need to be present. Describing acceleration as a second-derivative comes after the physical intuition of acceleration as a 'property' of the moving object. Thus, the miracle is not in the use of second derivative as Wigner has it but in the discovery of acceleration as an essential physical concept. Even in the case of Newton's equations, Newton himself notes in his *Principia* that Galileo had known the first two laws of motion – this without the use of the second derivative!

Descartes, one of the most influential mathematicians and philosophers of all time, believed that physics is a branch of mathematics as well exemplified by his statement that 'no other principles are required in physics than are used in Geometry or Abstract Mathematics, nor should any be desired, for all natural phenomena are explained by them'[15]. However, his view on mass is an instructive example about the pitfalls of ignoring the differences in the ideas of the physical and the mathematical. Consider two ways of characterizing mass: mass as extensional and as point-like. In one sense, mass as extensional reflects a brute facticity whereas mass as point-like seems to be counter-intuitive to the common sense. Descartes, for all his belief that physics is a branch of mathematics, conceptualized mass as being extensional. Newton, on the other hand, believed that the essence of mass was to be point-like, a move which allowed him to formulate his physics. Although Descartes had formulated the principle of inertia which was 'formally equivalent' to that of Newton, he did not discover Newtonian physics partly because of his belief in the essence of matter as being extensional. Descartes' belief that physics was a branch of mathematics came in the way of his acknowledging the importance of experiments in physics. Although he had formulated his rule of inertial motion and set of rules of impact, they were incorrect because he did not consider the vector nature of momentum. Cohen notes that Descartes could have easily discovered his mistake by simple experiments[16].

As a final example, I will briefly consider a seminal contribution of Newton to the process of mathematization. This particular method of mathematization which he initiated continues to influence the way mathematics is used in modern science. Cohen isolates one aspect of Newton's use of mathematics, what he calls 'Newton style', as illustrated in Newton's derivation of Kepler's law. First, Newton considers a purely mathematical system, nothing to do with how the world is but dictated by the concerns of pure mathematics. Here a 'single mass-point moves about a centre of force'[17]. Mathematically, if the centre of force is stationary and if the force is always directed towards the centre then Kepler's law of areas can be derived. This is a mathematical problem and treated as such. From this model he goes on to derive the other laws of Kepler, under appropriate conditions. After doing this, Newton compares this imaginary world with the real one. This immediately necessitates him to deal with two-particle motion since the centre of force is also a massive object. Then he develops the mathematics of this system. Next, when he compares his model with the real world, he finds that he has to take into account a much more complex world which has more than two bodies in the solar system. This dynamic interplay between mathematical ideas and comparison with the physical world made Newton realize that laws are absolutely correct only as mathematical laws but in physics they are only approximations (what he called 'hypothes
of mathematization is a continuation of this method (that is, the creation of ideal models) along with the concomitant realization of the important role of the notion of approximation[18].

What exactly is being mathematized and applied?

Let us begin with a catalogue of the furniture of mathematics – there are objects such as numbers, sets, functions and matrices; operators such as the ones used in arithmetic and calculus; rules of operation which make possible calculation; the equality sign (and associated with it appropriate inequalities) and a host of concepts such as continuous, analytical, differentiable and so on. So the first point to note is that applying mathematics could mean applying any or all of the above elements that belong to mathematics.

What is mathematics being applied *to*? Even in the simple example of applying a number we notice an interesting facet of application. For example, let us say that we first start with a statement (in English): 'there are some apples on the table', then apply the concept of number to this and get, say, the statement 'there are ten apples on the table'. This is a proto application of mathematics. But what is getting applied to what? Here, the idea of a number is being 'applied' to a sentence in English. The concept of number is not being applied to the real apples in the world but to a particular description of the world which is first expressed in English with the help of the word 'some'. The lesson from this simple example is one that is central to the process of mathematization: mathematics is first and foremost applied not to phenomena in themselves but to descriptions of phenomena. The two common modes of describing phenomena are through language and through idealized models (it can be argued that models themselves are one kind of linguistic description). Here I would like to focus attention on language and in particular to considering the possibility that the first defining characteristic of mathematical application is not the application of mathematics to the world as such but to other language(s). One way of understanding this is as follows: models and languages mediate between mathematics and the physical world.

Consider the example given by Wigner. Wigner was surprised that a second-derivative, which stands for acceleration, was integral to the mathematical formulation yet had no common-sensical correlate. But did Newton really write his equation this way? It is well known that in *Principia* Newton states his law as follows: 'The change of motion is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed'[19]. What exactly has been mathematized in this case? Where is the mathematics in Newton's law expressed in the modern form which says force is equal to mass times acceleration? In writing force as F, there has really been no mathematics done. In the usual form of the equation $F = ma$, we only have a *symbolic shorthand* for a longer sentence and this simple strategy is an important element of mathematization. Moreover, force, mass and acceleration are *not* 'mathematical' concepts. They are physical ideas and the genius of Newton lay in formulating the *appropriate* physical concepts first. (For Galileo, appropriate for science implies that they can be measured.) The mystery for Wigner would lie in the fact that the physical idea of acceleration can actually be correctly described by a second derivative.

Mathematics is therefore applied not to the world but to language and this application can even be at the level of creating appropriate symbolizations. It may also be argued that mathematization is actually an application of mathematics to (idealized) models. In the case of planetary motion, for example, one applies mathematics to idealized pictures and models of the planets. The use of mathematics in order to create new descriptions of pictures and models is also closely related to the mechanism of applying mathematics to language. In the next section I will discuss a little more about this particular process of application of mathematics.

Finally, the common belief that there is a miraculous correspondence between mathematical entities and physical concepts might suggest that recognition of this correspondence is instantaneous. But this is hardly the case.

Physical concepts like mass or force get refined over centuries. During this process, they come to be associated with various physical and mathematical meanings till they settle down to some stable mode of description. The real mystery might occur when mathematics itself begins to supply physical concepts[20].

The problem is compounded when we consider the following: the space of mathematics is much larger than that which is applied or perhaps even applicable. There is a surplus of mathematics and only a part of it finds use in the sciences. And more problematically, the same mathematics can be used to model and describe worlds which are not only very different but also contradictory to our world. That is, as far as the truths of our world are concerned, mathematics is quite indifferent to them. And if we believe that science correctly describes our world then this indifference of mathematics to the 'truths' of our world is a potential embarrassment for science if we want to claim that mathematics is essential to it.

Explaining the obvious: the unreasonable effectiveness of language[21]

Here is one possible way of analysing the usefulness of mathematics in the sciences. First, mathematics constitutes a particular kind of description of the world. Description is an activity of language. Languages describe the world around us. Different languages offer descriptions that are unique to that language. The unique elements of a language include the kinds of concepts the language possesses, its grammatical structure and its larger vocabulary and meaning. A same phenomenon can in principle be described in different ways by using different languages.

The capacity of language to 'correctly' describe our world is already mysterious. The problem is simple. Assume that the world is given to us. The world, distinct from language, is nothing but a collection of objects and events. Language arises in learning to talk about the objects and events of the world. Language not only seems to give us a 'proper' description of the world but also allows us to negotiate and intervene with the world in various ways. For simplicity, in what follows let me consider English as an example of a language. Describing the world with the help of English seems to capture some important facets of the world. Consider this simple example. Say we are seeing two objects in front of us and we describe our perception by saying 'one object is to the left of the other'. The capacity of English to create a word called 'left', which describes not an object in itself but a relation, is itself surprising but what is more amazing is that the linguistic statement 'one object is to the left of the other' seems to correctly match with our perception. But we somehow seem to take it for granted that there is no mysteriousness in the capacity of English to describe the world. We do not think that the use of English suggests an 'unreasonable effectiveness' just as the use of mathematics does. What could possibly be the reason for this lack of surprise at the role of language?

One possible reason is this: a natural language like English seems to largely arise out of our interaction with the world. The word 'tree' denotes an object tree – suggesting that we create a word in our language to say something about an object that we already have in front of us. We can name by pointing to things and children often learn the association of a word to thing through the act of pointing. In this naïve sense, words in a language seem to be derivative to the real world around us and arise in response to the given world. Objects and events surround us and we use language to talk about them leading us to the commonly held view that the world comes first and language follows the dictates of the world. This in a way reduces the mysteriousness in the act of using language to talk about the world, because it is *expected* that a natural language like English, since it arises from the wellspring of the world, should well describe the world.

And this is exactly where mathematics is seen to differ from English. Mathematical objects are not seen as those that belong to the natural world. Many mathematicians and scientists in fact believe quite the opposite – namely, mathematical entities belong to a Platonic world. However, mathematics functions in a way similar to natural language in the sense that the mathematical language is also a language, one which describes the mathematical world. For example, it gives names (such as 'sets') to mathematical objects (namely, sets), presumably existing in the world of mathematics. Therefore, the surprise is all the more exaggerated when it is found that mathematical objects, which presumably exist independently of our physical world, are very apt in describing our physical world. The surprise arises in finding that mathematics is doing a work which it supposedly should not be doing. And ironically, it seems to be doing it 'better' than natural language. In what sense is it doing a better job?

That mathematics does a better job than natural languages is perhaps most forcefully explained by the predictive success of the sciences based on mathematics. It is the predictive success of the sciences, based on mathematics, which gives the most important validation of mathematics. The mysteriousness of the effectiveness of mathematics is enhanced when a scientist stumbles upon a mathematical term which is then found to be the best fit to a particular physical description, like in the case of groups and symmetry or gauge theory and fibre bundles. Echoing this sentiment, Weinberg says that it is 'positively spooky how the physicist finds the mathematician has been there before him or her'[22].

However, both these descriptions of the character of English and mathematics are only partly right. English, although arising from a response to our natural world, also has the capacity to generate words which stand for physically non-existent objects. Abstract nouns, for example, refer to an abstract entity. Even the very act of having a word 'number', referring to a mathematical entity number, shows the capacity of natural language to refer to things which are beyond our physical world. Further, English generates a large amount of words which have nothing to do with physical objects. And the flip side of this is also that mathematics is not to be understood as being totally concerned with a Platonic world. So, both English and mathematics share some important, common features of languages, including the capacity to use both of them in different kinds of predictions. Mathematical description seems to be far more suited to certain types of description, typically quantitative, whereas a description in English may have superior qualitative expressions.

There are also some important differences between mathematics and English that we need to note[23]. In the context of applicability, I believe that the most important distinction which we need to focus upon is the observation that mathematics is not 'one' language like English. It is actually a collection of sub-languages each of which has some common links. Geometry, algebra, topology, etc. are sub-languages of a larger entity called mathematics. These are sub-languages in the sense that they function like a separate language in terms of the concepts they possess, the methodologies they use, their aesthetics and so on, yet share a common world with each other. Each discipline of mathematics is actually like a sub-language and in talking about mathematics as a language, as something homogenous, we overlook this important diversity and difference of its many sub-languages.

This diverse character of mathematics is very important and actually offers an explanation of why mathematics is so unreasonably effective. The different sub-languages that constitute mathematics make the descriptive enterprise of mathematics very interesting. Languages, when they are used to describe, explain, define, argue and so on, have specific narrative structures. Languages create narratives. A description is one kind of narration. The nature and the effectiveness of the description depend on the narrative structure of a language. In mathematics, the narrative structure is composed of the different elements of its different sub-languages, thereby expanding the scope of its narrative capability. Therefore, description in mathematics consists of much larger and more complex narratives than description restricted to only English. Let me give a simple illustration of how this is done[24].

Consider light reflecting from a mirror. How can we invoke mathematics here? What kinds of descriptions can we develop about this event with the use of mathematics? First, as is commonly done, we can give a pictorial representation of this process. The mirror is represented by a straight line and the incoming ray and the outgoing ray by two straight lines. Drawing the normal, we have the angle between the rays and the normal. This pictorial representation is very useful for science in that it allows us to do what we want with an idealized system. Mathematics comes into play on this idealized picture when we 'name' angles and use properties of terms such as momentum. So from a picture of the process we move into geometry (a sub-language of mathematics) of the system. This allows us to define and describe components of the momenta, forces and so on. At this stage we begin to do geometry – in the particular case of the reflection of light we do geometry on a plane. The results of these calculations will depend on some results that belong to the domain of this geometry and not the domain of the real phenomenon.

So typically this is what happens in the process of mathematization. The event in the world is first represented pictorially, for example, which can then be expressed in another sub-language, say geometry, and then in algebra and so on. Each one of these steps takes the real world event into different narrative domains. For example, light bouncing off a mirror has no velocity component in the real world but a mathematical description talks as if the components of momentum are real. So the shift into pictures and other sub-languages succeeds in adding new descriptions of the original event. It is important to note that these descriptions are unique to the different sub-languages. Description in the pictorial form is very different when compared to the ones derived from the geometrical narrative, which is itself very different from the one derived from using algebra. For example, once we enter the descriptive space of algebra we have a new vocabulary that is available to us to describe the process: continuity, rate of change, equations of motion and so on. This vocabulary, which was not present in the earlier sub-languages of pictorial representation or geometry, succeeds in expanding the narrative possibilities of this process. In the realm of algebra, the vocabulary allows us to talk of motion in higher dimensions, the possibility of transformations of co-ordinates, even the physically non-intuitive idea of transforming momenta into co-ordinates and so on. The important sub-language of calculus along with algebra allows us to develop extremely rich narratives about a simple process such as a ray of light bouncing off a mirror.

Thus, we see that the process of mathematization using the many sub-languages of mathematics enlarges the possible descriptions one can have of a process. There are literally no conceivable limits to what sub-languages we can use for this description. If, for example, someone finds the vocabulary and grammar of topology useful in the description of a bouncing ball then it becomes part of the larger mathematical description of this process.

So, first and foremost, using mathematics to describe the physical world is a means of finding ways to create multiple descriptions of a physical object or event. We can see that a language like English will only create limited narratives about a phenomenon because it does not have the rich sub-languages that mathematics has. When we use mathematics as a language to describe a process we first of all create a rich storehouse of possible narratives. What among them will fit the world is an issue that mathematics is unconcerned about. The job of mathematics in sciences is essentially to proliferate narratives and the more number of narrative descriptions are possible the better *probability* that there will be a fit somewhere, sometime.

Compounding the problem – the pictorial role of mathematics

A great deal of creative mathematics, both pure and applied, depends on analogy. I want to illustrate a few cases of analogies that have to do with formal patterns of the mathematical symbols. It is remarkable that mimicking the patterns of written mathematical terms yields profound new ideas in physics. The history of mathematics and science is replete with this strategy. The following elementary example is primarily to illustrate a seemingly arbitrary method used in mathematization. A detailed analysis of the role of form in mathematical discourse is extremely illuminating but here I will only briefly touch upon this issue[25].

The basic strategy is this: by looking at the way in which mathematical expressions are written and arise in the course of calculation we are able to identify some new information. A simple example is that of a term which looks like $1/2ab^2$. When we see such an expression in the context of some calculation it seems natural to identify a with a mass term and b with a velocity term since this expression *looks* like a kinetic energy term. Identifying and discovering such terms can be very important steps in theoretical research in science. Consider the following example from Landau and Lifshitz's *Mechanics*. Consider two particles with masses m_1, m_2 and velocities v_1 and v_2 in an interactive potential field. The total kinetic energy of the system is the sum of the kinetic energies of the two particles. In this system, we can rewrite the total kinetic energy as one term which looks like $1/2ab^2$. Now, looking at this we interpret a as the mass term (the reduced mass) and b as the velocity term. Further, the authors claim that because the expression of two-particle kinetic energy terms reduces to that of one kinetic energy term the two-particle motion is *equivalent* to the motion of one particle[26].

This strategy of 'discovering' mass is one that is practiced right across the many disciplines of science. In physics, it is extensively used in areas ranging from classical physics to particle physics. In fact, the identification of mass terms in quantum field theory follows similar 'pattern recognition' of symbolic terms. The importance of mathematical form should not be underestimated. Mathematical form is not about doing mathematics alone; it is also about writing mathematics in some specific ways, the underlying belief being that physical terms are expressed by unique mathematical forms. Thus, classical kinetic energy will be of the form $1/2mv^2$ or in terms of momentum as $p^2/2m$. When we move from classical to quantum physics, the identification of 'kinetic energy' continues to be $p^2/2m$ at the formal level although the physical meaning of kinetic energy for a wave is very different from that of a particle. In quantum theory, we replace p by an operator but the form of the term remains the same, as seen in the Schrödinger's equation.

Interestingly, Steiner points out that there is another formal analogy in the Schrödinger equation, which is that this *formally* identical to the equation for a *monochromatic* light wave in a *nonhomogeneous* medium'[27]. Similarly, the various meanings ascribed to the mass term in classical, electromagnetic, relativistic and quantum mechanical theories were significantly dependent on formal symbolic identification[28]. There are innumerable examples of the importance of symbolic manipulation based on formal similarity. In fact, I would go to the extent of saying that the effectiveness of mathematization significantly depends on the power of symbols to act like *pictures* of ideas, concepts and events. The role of mathematics in the sciences seems to be essentially dependent on the possibility of using mathematical symbols as 'pictures'. For example, we could look upon $1/2ab^2$ as the 'picture' of kinetic energy. So even in contexts that are very different we can still recognize that picture and identify it with the kinetic energy of that object or system. Similarly, the generation of 'alphabets' in mathematics is itself a very creative process and these alphabets many times visually suggest the kinds of things that can be done with them[29].

The above discussion indicates the complexity involved in the process of mathematization of the world. The great challenge to science will lie not only in the creation of new mathematics but also in the possibility of creating new modes of expressions and new languages in the unending scientific search for mapping the universe.

Notes and references

1. Wigner, E. P., The unreasonable effectiveness of mathematics in the natural sciences. *Commun. Pure Appl. Math.*, 1960, **13**.
2. For example, see *Philosophy of Mathematics* (ed. Jacquette, D.), Blackwell, Oxford, 2002.
3. Dummett, M., What is mathematics about? In *Philosophy of Mathematics: An Anthology* (ed. Jacquette, D.), Blackwell, Oxford, 2002, pp. 19–29.
4. See Brouwer, L. E. J., Intuitionism and formalism. In *Philosophy of Mathematics: An Anthology* (ed. Jacquette, D.), Blackwell, Oxford, 2002, pp. 269–276.
5. Hardy, G. H., Mathematical proof. In *Philosophy of Mathematics: An Anthology* (ed. Jacquette, D.), Blackwell, Oxford, 2002, p. 180.
6. Kanigel, R., *The Man Who Knew Infinity*, Scribners, London, 1991, p. 147.
7. See Brouwer, L. E. J., Intuitionism and formalism. In *Philosophy of Mathematics: An Anthology* (ed. Jacquette, D.), Blackwell, Oxford, 2002, pp. 269–276.
8. Quoted in Parsons, C., Mathematical intuition. In *Philosophy of Mathematics: An Anthology* (ed. Jacquette, D.), Blackwell, Oxford, 2002, p. 277.
9. Steiner, M., *The Applicability of Mathematics as a Philosophical Problem*, Harvard University Press, Cambridge, 1998, p. 30.
10. Steiner, M., *The Applicability of Mathematics as a Philosophical Problem*, Harvard University Press, Cambridge, 1998.
11. Rotman, B., *Ad Infinitum: The Ghost in Turing's Machine*, Stanford University Press, Stanford, 1993.
12. Sarukkai, S., *Translating the World: Science and Language*, University Press of America, Lanham, 2002.
13. Cohen, I. B., *Revolutions in Science*, Harvard University Press, Belknap, 1985, p. 139.
14. For Galileo it was much more difficult since he had to extrapolate in a complicated way from observations on an inclined plane.

15. Cohen, B., *Revolutions in Science*, Harvard University Press, Belknap, p. 156.
16. Cohen, B., *Revolutions in Science*, Harvard University Press, Belknap, p. 155.
17. Cohen, B., *Revolutions in Science*, Harvard University Press, Belknap, p. 166.
18. For a discussion on approximation in this context, see Sarukkai, S., *Translating the World: Science and Language*, University Press of America, London, pp. 114–118.
19. Chandrasekhar, S., *Newton's Principia for the Common Reader*, Clarendon Press, Oxford, 1995, p. 23.
20. See Sunil Mukhi's contribution in this collection.
21. For a more detailed discussion on some of these aspects of language and the unreasonable effectiveness of English, see Sarukkai, S., Applying mathematics: The paradoxical relation between mathematics, language and reality. *Econ. Pol. Weekly*, 2003, **XXXVIII**, **35**, 3662–3670.
22. Weinberg, S., Lecture on the applicability of mathematics, *Not. Am. Math. Soc.*, 33:5, 1986. However, we should also note that stumbling upon the right mathematical terms is quite similar to a musician stumbling upon the right notes or a poet stumbling upon the right word or phrase. What special value can be added to a 'right term' just because one stumbles upon it?
23. I do not subscribe to the opinion that the degree of precision and non-ambiguity distinguishes natural language and mathematics. The association of precise meanings with mathematical terms and ambiguity with English words does not represent the true picture. For more on semantic plurality and metaphorical use of mathematical terms, see Sarukkai, S., *Translating the World*. A related issue is whether mathematics can be totally divorced from its connection with natural language and whether mathematics needs natural language in an essential sense. For more on this see, Sarukkai, S., Mathematics, language and translation. *META*, 2001, **46**, 664–674. (The full text of the article is available at http://www.erudit.org/revue/meta/2001/v46/n4/.)
24. For more detailed discussion, see Sarukkai, S., *Translating the World: Science and Language*, University Press of America, London, 2002, pp. 60–70.
25. For more details, see Sarukkai, S., *Translating the World*, University Press of America, London, 2002, Part One.
26. Landau, L. D. and Lifshitz, L. M., *Mechanics*, Pergamon Press, Oxford, 1976, p. 29.
27. Steiner, M., *The Applicability of Mathematics as a Philosophical Problem*, Harvard University Press, Cambridge, 1998, p. 79.
28. See Jammer, M., *Concepts of Mass*, Harvard University Press, Cambridge, 1961.
29. For examples of alphabetisation in mathematics and in general the role of discursive strategies in mathematics, see Sarukkai, S., *Translating the World*, University Press of America, Lanham, 2002.

ACKNOWLEDGEMENTS. S. K. Rangarajan was instrumental in clarifying many ideas on mathematics and I am indebted to him for his patient attempt to understand my views even though it might have grated his sensibilities about mathematics! I also thank Dhanu Nayak for comments on various points about language.