Artykuły przeglądowe

Mathematical methods in molecular biology

*Przemysław Waliszewski**
Department of General
and Gastrointestinal Surgery
University Medical School
Poznań

Traditionally, mathematics is considered to be a language of physics. However, some of the mathematical methods have been developed or improved as a result of questions addressed by biological experiments. It may seem surprising, but biology has had an impact on the development of several mathematical theories including the theory of ordinary and partial nonlinear differential equations, the theory of dynamical systems, differential topology, differential geometry, graph theory, stochastic methods and statistics. Also, processing of the large number of biological data has stimulated the development of more powerful computers and computational methods. Of course, mathematical theories have a long lifetime in comparison with the biological ones. They can be developed independently on any real system. However, they allow a more complete analysis and aid in the design of future experiments if applied to the description of biological phenomena. Sometimes,

* on leave from the Department of Cancer Biology NNI-06, The Cleveland Clinic Foundation, Cleveland, Ohio, USA.

a mathematical approach helps to discover a new aspect of the phenomenon which would not be noticed by any other method. For example, it is possible to construct equations describing molecular interactions. A computer simulation of these equations enables to investigate the dynamics of the interactions at very short intervals of time e.g. nanoseconds or close to the points of stability. At this moment, it is important to realize that mathematical description does not explain the real molecular events. It only suggests a mechanism. If possible, the finding should be confirmed experimentally.

The complexity of biological phenomena may, however, require improved methods or completely new theoretical approaches. These situations influence the development of mathematics and revitalize it. For example, this happened in the case of catastrophe theory, developed by Rene Thom. He had been inspired by Waddington's concept of an epigenetic landscape. As a result, he developed singularity theory and investigated topological features of dynamical systems (1). This elegant theory is an important tool to study the stability of such systems. A similar situation has occurred in the case of complexity theory. This theory is being developed by Haken, Rosen and a group of theoretical physicists working in the science center in Santa Fe, New Mexico, USA. The inspiration for this study was structural complexity of biological systems. One of the most important questions this theory is trying to answer is the self-organization, development and future of such systems (2).

From the mathematical point of view, a mathematical model should be a simplified, but realistic representation of the biological phenomenon. It may neglect some aspects of the complex phenomenon in order to reveal its essential components. Assumptions should agree with experimental data. The model solution should be compatible with questions, preferably quantitative. Biological interpretation of mathematical results should be carried out in terms of insights and predictions because the application of mathematical methods depends on the biology of the phenomenon and not vice-versa. Sometimes, mathematical model can be applied as a method of indirect argumentation.

A classic example of the routine application of mathematical methods in molecular biology is the theory of ordinary and partial nonlinear differential equations. These equations are functions defined over the real numbers and contain derivatives as variables e.g. $F(t, y(t), y'(t)) = 0$. The equation is considered to be ordinary if there is a derivative for only one variable. Otherwise, it is called a partial differential equation e.g. $F(t, x, y(t,x), dy/dt, dy/dx) = 0$. The equation is called a linear differential equation if the solution manifold is a linear space. Otherwise, it is called nonlinear. Only the nonlinear differential equations, usually partial, are useful in describing a complex biological reality. They have been applied successfully to the analysis of both enzyme-substrate and ligand-receptor interactions (3).

Differential geometry is the discipline which applies differential calculus to determine the differential invariants of manifolds. The manifold is defined as a mathematical construct which contains a subset of points of the given topological space and a homeomorphism, a projecting function, which creates an image of the manifold in the n-dimensional space of the real numbers. Geometry investigates rigid transformations of the structure. Topology investigates flexible transformations. If there is a function which projects one object into another, and both functions, f and f^{-1} are continuous and possess derivatives then these objects are considered to be topologically diffeomorphic. Briefly speaking, they are topologically equivalent. Topological methods investigate invariants of shape. They enable the discovery of topologically identical objects. For example, these methods are useful in describing a variety of DNA chain configurations, driven from one to another by thermal motions, ion concentration, enzyme activity, etc. It is possible to detect mathematically unstable conformations or barriers between them which can be crossed only by the transient breaking and reforming of the chemical bonds which define macromolecule. Also, topologically identical configurations can be identified (4,5).

Graphs are algebraical structures which are defined as a collection G = (X, Y, u), where X is a set of vertices, Y is a set of links between them i.e. a set of edges, arcs or loops. Edges, arcs and loops code relationships between vertices. Vertices can simulate any real objects. u is a mapping (projection) u: X x X → Y, which associates with each pair (v,w), for which it is defined, an element z = u(v,w) of the set Y — a link which connects the vertices v and w in such a way that z does not connect any other pair of vertices.

Theorems of graph theory enabled to construct algorithms for the problem of transportation and conduction in the neuronal network. A computer image analysis of three dimensional objects is based upon them. In molecular biology, they have been applied to the construction of the evolutionary trees. Using this approach, a computer analysis of gene homology between species have been possible (6).

Recently, topological properties of a new class of graphs are investigated. This class of graphs is called embedded graphs to refer to graphs with a fixed configuration in three dimensional space. These graphs are important tools in studying of the forbidden stereochemical configurations of macromolecules (7,8). Embeddings of the flexible circular graphs in three dimensional space are domain of the Knot theory. This theory is helpful in the understanding of topological features of circular DNA and its transformations under the different physicochemical conditions.

Increasing numbers of sequencing and crystallographic data resulted in the need to set up special databases. Protein sequences are stored in the Protein Identification Resource (PIR). DNA sequences are collected in the GenBank or in the Nucleic Acid Database (NDB). Also, there exists a bank of crystallographic data. All these databases are available to European scientists. They can be searched from Europe through the Internet* computer

* cc: ..... is the Internet email address. The complete list of the meeting participants and abstracts can be obtained from Sylvia J. Spengler, Ph.D., University of California at Berkeley, Program in Mathematics and Molecular Biology, 103 Donner Laboratory, Berkeley, CA 94720, USA, email: sylviaj@violet.berkeley.edu

system using anonymous ftp access and Gopher program. For example, subscription to the NDB can be sent through the Internet to: ndblib@helix.rutgers.edu, subject: subscribe. The NDB library can also be accessed by anonymous FTP using the following instructions: start the FTP program by typing: ftp helix.rutgers.edu and log in by typing: user anonymous. FTP will request a password, to which user's name should be typed. Then type: cd pub. Then issue a command of the following form to choose the directory with the files to be transferred: cd<directory>. <directory> is replaced by the name of one of the following directories: <newsletter>, <reports_ascii>, <reports_ps>, <coordinates>, <torsions>. To obtain a description of the contents of this directory, type: get index. This will create a file called index in your current working directory. This file can be examined after having quit the FTP program. To obtain any of the files, type: get<file> where <file> is replaced by the name of one of the files, as described in the index. Questions about the file server should be addressed to: ndbadmin@helix.rutgers.edu

When a new DNA sequence is determined, GenBank is searched for approximate similarities with the new sequence. A special computer program translates the DNA sequence into the corresponding amino acid sequence which is used to search the protein database. Another computer program identifies nucleotide positions sensitive to digestion with restriction enzymes and plots a restriction map.

Sequence alignment is one of the most important tasks carried out by molecular biologists. However, at the same time, it is one of the most difficult mathematical problems which involves the advanced methods of statistics, graph theory and computer programming. At present, GenBank contains about 129,968,355 bases what equals to 111,911 sequences (9). In order to search this number of data, dynamic programming methods have been devised (10,11). These methods are based on the family of programs called FASTA, FASTN etc. The idea of alignment is to locate diagonals in the program and to align similar sequence elements with positive scores and dissimilar elements with negative scores. The program aligns two sequences into the maximum scoring alignments. However, alignment of more than two sequences requires extended computation time. Different methods have been developed to reduce this time and so far up to ten short sequences can be aligned together.

At the meeting in Santa Fe, Martin Vingron, from the Department of Mathematics at the University of Southern California, cc: mvingron@hto.usc.edu* presented an intertesting approach based upon the attributing weights to sequences, where each nucleotide had its own numerical value written into the Kronberg's matrix 4 x 1. Then, the matrix of similarities between pairs of sequences was set up and values of the given sequence were compared to the values of an average sequence within the superfamily. A better alignment of the new sequence results from the inclusion of more sequences recognized as members of the superfamily. However, this approach is effective only if the sequence belongs to the known gene superfamily, or if sequences of the same gene in different species are compared.

The problem of sequence comparison implies the problem of estimating the significance levels for the alignment scores. This problem has been solved by approximating the distribution of sums of dependent indicator random variables, by using the Poisson distribution. Dependent random variables are the set of possible alignment scores from the two compared sequences (12).

DNA sequence data can be used to reconstruct phylogenetic trees and analyze genetically complex traits. During the same meeting, Joseph Felsenstein, from the Department of Genetics at the University of Washington in Seattle,cc: joe@genetics.washington.edu, discussed the application of maximum likelihood methods to multisequence data with sequences from different species, and from members of the same species. With aligned sequences, maximum likelihood phylogenies (evolutionary trees) can be estimated under a model with equal rates of change at all sites. Dr Felsenstein generated an extension to unequal rates using Hidden Markov Chain techniques and presented a solution with the Markov Chain Monte Carlo computational method.

The meeting in Santa Fe was dominated by the problems of structural analysis and molecular dynamics simulation of supercoiled DNA. A mathematical approach to these problems is based on methods of differential geometry and differential topology. Applications of these two branches of mathematics appear to be particularly useful in the analysis of DNA secondary structure with geometric concepts of tilt, roll, shear and twist (reviewed in Dickerson, R.E. (1989) (13)). Also, DNA interaction with ligands and proteins was studied with these methods (14). The list of problems solved by these methods is long and comprises the analysis of closed DNA supercoiling (15), enzymatic influence on the topology of the DNA chain (16, 17), estimation of the extent of winding in nucleosomes (18), determination of free energy associated with supercoiling (19) and determination of the helical repeat of DNA in solution and DNA wrapped on protein surfaces (20).

Molecular dynamics simulation is a mathematical and computational method which starts with a three-dimensional structure of a molecule. Then, the total force, acting on each atom, is calculated. Subsequently, according to Newton's second law of motion, acceleration of each atom is determined. Integration of these functions gives a numerical trajectory of each atom, describing its position and velocity as a function of time. This approach allows a molecule to be seen in different dynamical states. However, due to computer limitations, the number of the analyzed atoms can not be higher than a million and simulation time can not be longer than a nanosecond. Hopefully, the progress in computer technology will enable us to simulate behaviour of more complex molecules. The same methods can be applied to solve problems connected with drug design. A special computer graphic program, called docking technique, enables the study of structural and functional features of macromolecules if their crystallographic structure is known (21, 22). Recently, this technique was used successfully to investigate enzyme active sites and to design monoclonal antibodies with a catalytic capability.

It is expected that this technique will help to design more specific monoclonal antibodies against antigens that have well-recognized X-ray crystallographic structure. Also, it can improve targeting of gene carriers used in gene transfer.

Arthur J. Olson, from the Scripps Institute in La Jolla, cc: olson@scripps.edu, discussed a method of approximating, characterizing and visualizing the surfaces and properties of macromolecules. This method uses expansions of spherical harmonic functions. His approach relies on a topological mapping. Each point on the surface of the molecule has its own unique point representation on the unit sphere. Then, spherical harmonic expansion coefficients are calculated for the new surface. This approach enables the calculation of vector and scalar shape properties such as electrostatics, hydrophobicity or hydrogen bonding potential. His method is a general tool for characterizing and visualizing molecular surfaces and for constructing techniques for the automated docking of macromolecules.

Tamara Schlick from the Department of Chemistry at New York University, cc: schlick@acfclu.nyu.edu, constructed energy functions for supercoiled DNA. Subsequently, she investigated the minimum energy of supercoiled configurations and fluctuations about these states. She applied both global stochastic optimization by the Metropolis-Monte Carlo algorithm and local deterministic optimization by the truncated Newton minimization algorithm. In her model, shape and motions of the supercoiled DNA changed depending on the number of linkage between two closed and oriented DNA strands, and on chain length, sequence and ionic environment.

Yang Yang, from the Department of Chemistry of the Rutgers University in New Jersey, cc: yyang@jove.rutgers.edu, examined the influence of various stressed conditions on intrinsically straight DNA (6400 bp) with a few highly bent segments (200 bp). He calculated three dimensional equillibrium configurations and elastic free energies and demonstrated that both the magnitude of the intrinsic curvature and the positions of the curved segments influenced the topology of this DNA. Also, these intrinsically curved segments were able to orient the knotting and supercoiling of DNA.

Robert Tan, from the Department of Biochemistry at the University of Alabama, cc: tan@neptune.cmc.uab.edu, modeled small closed circles which may develop on the linear DNA. He found, that depending upon the number of linking deficit, some topoisomers had unusually compact structure. He suggested that this was a reason for the poor reactivity with DNA topoisomeras I which he observed experimentally using a curved 217 basepair sequence derived from *Crithidia fosciculata.*

Assuming that the interaction between estradiol, estrogen receptors and estrogen responsive elements is of a kinetic nature, the author, cc: Walisz-P@ccsmtp.ccf.org, analyzed a system of ten nonlinear differential equations with strongly interrelated variables that described this interaction. They had been constructed according to the Law of Mass Action. The computer analysis, however, revealed that the solution dynamics was independent of the variables, the amounts of interacting molecules. Instead, it was dependent on the parameters, particularly the dissociation constants of both estradiol/estrogen receptor complex and the estrogen receptor/estrogen responsive element complex. These constants determine the slowest steps of the interaction. Thus, the mathematical model suggests a catalytic mechanism of the interaction. Also, these results suggest that the structure of the interacting macromolecules is more important for the regulation of gene expression than their amounts.

There are some limitations in the mathematical approach to molecular biology. One of the unanswered question is how far a mathematical description can approximate the biological reality limited by the same principles as quantum physics? Also, the problem of nonlinear dynamics with variations between individuals and the problem of space, time and hierarchical complexity of different phenomena are true difficulties in the mathematical approach to biological systems.

All the mentioned techniques are highly advanced mathematical tools. They require from both mathematicians and biologists an appropriate training before they can benefit biological research. This can not happen until mathematical biology will be taught at the appropriately advanced level at the Universities and at the Departments of Biology.

To summarize, for the mathematicians, biology offers new challenging problems that may stimulate development of new mathematical theories. For the biologists, mathematical methods, if applied with the knowledge of their limitations, are research tools that can support experimental work. Mathematical biology is a field of study which flourishes and benefits only when developed in cooperation between representatives of both disciplines.

I acknowledge D. Benbrook, Ph.D. from the University of Oklahoma for discussion, suggestions and minor editing. Also, I acknowledge K.P. Hadeler, Ph.D. from Tubingen University for reading of the manuscript and valuable suggestions.

References

1. Thom R., (1975), *Structural stability and morphogenesis*, W.A., Benjamin, Reading.
2. Rosen R., (1985), *Theoretical Biology and complexity*, Academic Press, Orlando.
3. Murray J.D., (1989), *Mathematical Biology. Biomathematics* 19. Springer Verlag, Berlin.
4. Sumners D.W., (1987), J. Math. Chem., 1, 1 – 14.
5. Sumners D.W., (1987), *The role of knot theory in DNA research. Geometry and topology*, Eds. McCrory C., Schifrin T., Marcel Dekker, 297 – 318.
6. Mirkin B.G., Rodin S.N., (1984), *Graphs and genes*, Springer-Verlag, Berlin.
7. Flapan E., (1987), Pacific Journal of Math., 129(1), 57 – 66.
8. Walba D., (1983), *Stereochemical topology. Chemical applications of topology and graph theory*, Ed. King R.B., Elsevier, Amsterdam, 17 – 32.
9. NCBIGen Bank Database, (1993), Release 76.0 (April 15th).
10. Lipman D.J., Pearson W.R., (1985), Science, 227, 1435 – 1441.
11. Smith T.F., Waterman M.S., (1981), J. Mol. Biol., 147, 195 – 197.

12. Arratia R., Goldstein L., Gordon L., (1989), Ann. Probab., 17, 9 - 25.
13. Dickerson R.E., (1989), Nucleic Acid Research, 17(5), 1797 - 1803.
14. Wang J.C., Peck L.F., Becherer K., (1983), Quant. Biol., 47, 85 - 91.
15. Bauer W.R., (1978), Ann. Rev. Biophys. Bioeng., 7, 287 - 313.
16. Cozzarelli N.R., (1980), Science, 207, 953 - 960.
17. Wasserman S.A., Cozzarelli N.R., (1986), Science, 232, 951 - 960.
18. Travers A.A., Klug A., (1987), Philos. Trans. R. Soc. Lond., B317, 537 - 561.
19. Depew R.E., Wang J.C., (1975), Proc. Natl. Acad. Sci. USA, 72, 4275 - 4279.
20. White J.H., Cozzarelli N.R., Bauer W.R., (1988), Science, 241, 323 - 327.
21. Duncan B.S., Olson A.J., (1993), Biopolymers, 33.
22. Goodsell D.S., Lauble H., Stout C.D., Olson A.J., (1993), Proteins.

Metody matematyczne w biologii molekularnej

Streszczenie

Powszechnie uważa się, że matematyka jest językiem fizyki. Tymczasem rozwój niektórych współczesnych metod matematycznych takich jak: teoria zwyczajnych i cząstkowych nieliniowych równań różniczkowych, teoria systemów dynamicznych, topologia różniczkowa, geometria różniczkowa, teoria grafów oraz metody stochastyczne inspirowane były wynikami eksperymentów biologicznych. Metody te stały się podstawą modelowania matematycznego i molekularnego struktury i funkcji makromolekuł. Złożoność struktur molekularnych wymusiła konieczność konstrukcji odpowiednich programów komputerowych, komputerów zdolnych do przetwarzania dużej liczby danych eksperymentalnych oraz współpracy w sieci.

Metody matematyczne, chociaż niezauważone w codziennej pracy laboratoryjnej, będą odgrywały coraz większą rolę w biologii molekularnej. Ich zastosowanie uzależnione jest od współpracy, prezentacji problemów językiem zrozumiałym dla przedstawicieli obu dyscyplin oraz wiedzy o zakresie możliwości.

W artykule tym dokonano przeglądu metod matematycznych, osiągnięć oraz perspektyw interdyscyplinarnego podejścia do problemów biologii molekularnej. Omówiono także niektóre prace przedstawione na III Konferencji „Mathematics and molecular Biology. Computational Approaches to Nucleic Acid Structure and Function", 7 - 11 listopada 1992, Santa Fe, New Mexico, USA.

Key words:
computer graphics, differential equations, graphs, macromolecules, topology.

Adres dla korespondencji:
Przemysław Waliszewski, Department of General and Gastrointestinal Surgery, University Medical School, ul. Łąkowa 1/2, 61-878 Poznań.