Correct problem statements in biomedical data processing

by N. I. MOISEEVA, M. YU. SIMONOV and V. M. SYSUEV
Academy of the Medical Sciences of the USSR
Leningrad, USSR

ABSTRACT

The correct formulation of any task appears to be not only the first step in its solution but sometimes the solution itself. Biomedical data processing is unlike other problems encountered by biologists and they often use mathematical methods without taking into consideration both the possibilities and limitations of the given methodology and the properties of the data to be processed. As a result the approach to the solution of the problem is often inadequate.

Moreover, biologists encounter difficulties with the choice of a method for data conversion into the form suitable for mathematical processing. Often this leads to the incomplete utilization of the power of a given mathematical method. Sometimes only the correct data representation is sufficient to draw final conclusions. For example, data given in the form of an interval distribution histogram or mean frequency function allow the investigator to obtain information on time-dependent features of the process.

Science “creates and supports conditions whereby functional foundations become the field of controversy, resulting in competitive but different ways of doing the same thing. In other words the number of alternatives constantly increases due to science.”

In fact the problem of choice always faces the investigator not only in relation to the optimal method of data processing, but also in relation to optimally configuring electronic devices to be used in the implementation of these methods.

The correctly formulated requirements for biomedical data processing could be and should be used as a basis for this choice. Thus there arises the special problem of stating correctly the processing task itself. From our point of view the correct approach to its solution must take into account all a priori information. We must systematically look at the final aims of our research (in the sense that it is a biological data processing task) and at existing methods of analysis, their scope and limitations.

PECULIARITIES IN THE STATEMENT OF BIOMEDICAL TASKS

The complications arising in the study of biological objects derive above all from the fact that life itself, from the point of view of physics, “is too intricate and does not lend itself to mathematical interpretation.” Living organisms possess a large number of possible stable states; every concrete subject during its whole life uses only some of them, and different subjects prefer different approaches (if it is possible to use this word in reference to processes in the vegetative nervous system). Moreover, in the same organism even stable states are different as they depend, for example, on age and functional states. Thus the problem of selecting appropriate statistical tools in biology becomes quite intricate. In fact, even the idea of “normality” is not clearly defined in biology so far.

Transmission of signals in living organisms is carried on both with the aid of discrete impulses (e.g., in the nervous system) and analog-type information, like the constant potential of membranes and tissue, humoral regulation of functions and so on. Under these conditions signals propagate in two ways—along a specific channel to a specific organ or group of cells and “in a general way,” without any defined “address.” The latter signals could act upon any part of the organism and could, in fact, be perceived either by any part of the organism (i.e., general nonspecific reactions like arousal) or by the tissues and organs specially prepared for the reception of this particular signal or group of signals. For example, the hormones of endocrinous glands excreted in blood are intended for specific “target-tissue.”

The next complication arises from the fact that there are no really stable functions in a living organism, every one of them representing oscillation processes.

The relative stability of an organism is the result of the rhythmic activity of the regulating systems. Some of these rhythms are endogenous: That is, they originate in the organism itself. Others reflect external events like the change of day and night, seasons of the year, and other influences from the environment.
As a specific property of this oscillating system the irregularity of growth and decay of different parts should be mentioned. For example, the growth of bones and muscles in human beings practically ceases as we attain adulthood, but some internal organs and the cardiovascular system continue their development during the whole life. Organisms develop and degrade in an irregular manner.

All this leads to the idea that biological analysis must take into account structural (synchronous) as well as temporal (diachronic, functional) aspects of living objects.

SUGGESTED CLASSIFICATION OF BIOMEDICAL DATA PROCESSING TASKS.

In the literature on this problem we were able to find only one attempt to systematize biomedical tasks. We are here not concerned with the classification based on the coefficient of utilization of computer memory. Rather we refer to the paper by Dixon in which the author discusses both the so-called "source of information" (in fact, dealing with the modelling of functions, selecting interconnections, estimating the state of health etc.) and the desired processes. The taxonomy is rather incidental and classification indicators are not selected.

We made an attempt to distinguish two groups of biomedical processing tasks after having selected as a basic indicator the relation to time: those tasks where some variables and their interconnections are analyzed without regard to time (synchronous analysis) and those tasks where the processes develop in time (diachronic analysis).

Within each of these groups the tasks are classified according to final aims (description and identification or classification) on one side and to the object of the investigation (elementary events, processes or interrelations) on the other side.

In Tables I and II the systematization of biomedical processing tasks is shown both for the synchronous and diachronic analysis.

It is well-known that all the curiosity as to events in our universe can be expressed with the help of only six questions: who (what); what kind of; where; when; how; why (what for). All entries in Table I are in response to questions about the kind of objects and the nature of their interactions; or to questions whether certain objects belong to certain classes. Entries in Table II address the questions "what kind of" and "when" (analysis of elementary events and their interactions within the time continuum), the questions "what kind of", "when" and "where" (analysis of processes and their interconnections), or the questions "what" and "when" (classification of phenomena and processes). The questions "how" and "why" are not within the scope of our analysis.

Definition of causes, aims and mechanisms is derived from the combination of analytical and synthetical processes in the human brain. Without the latter component the results of analysis bring to life only the question "so what?".

SUGGESTED CLASSIFICATION OF BIOMEDICAL DATA ANALYSIS METHODS

Many books on mathematics for biologists exist, of course, but there are no specific methodologies (and, of course, branches of mathematics) specially tailored to deal with biomedical problems. Biologists in their investigations simply use more or less successfully "standard" mathematics. We were unable to find in the literature any guidelines for biologists to aid their understanding of the relations between different methods or giving advice on the use of various kinds of processing techniques.

Biologists and non-biologists are becoming aware of the need to systematize methods of analysis. Isolated attempts to create particular schemes of analysis have been made, as shown in Tables III and IV (from the

| TABLE I—Levels of Structural (Synchronous) Analysis of Biological Objects |
|---|---|
| Analysis of elementary events (who) | Analysis of the interconnections between the biological objects (why) |
| Description of the objects (where) | Identification of the causes (when) |
| Description of the elementary objects (what) | Identification of the interconnections (how) |

| TABLE II—Levels of the Diachronic Analysis of Biological Objects |
|---|---|
| Analysis of processes (who) | Analysis of the objects (why) |
| Analysis of the interconnections (where) | Analysis of the elementary processes (how) |
| Identification of the causes (when) | Identification of the elementary objects (what) |

From the collection of the Computer History Museum (www.computerhistory.org)
TABLE III—General Procedure for Analyzing Individual Sample Records

book by Bendat and Piersol) for the complex of realization and Table V for time series used for the rhythmometric investigations.

Since neither general schemes nor a classification of methods exist, we made an attempt to classify the methods of analysis on the basis of the above taxonomy of the data processing tasks.

As time and spatial coordinates are processed identically in calculating procedures we combined the methods of spatial structural analysis and methods of time series analysis.

The basis for our classification is the ability of a given method to answer one or more questions from the above mentioned set. Thus we systematized ways of processing and separated the methods which could be used as:

—methods of qualitative analysis (indicative methods) answering questions on the general features of the object, event or a process (from a mathematical point of view).
—quantitative methods of investigation for estimation of the numerical characteristics of processes or objects.
—methods for qualitative and/or numerical description of interconnections.
—methods for analysis of external (related to the environment) features of objects and processes.

TABLE IV—General Procedure for Analyzing a Collection of Sample Records

It turned out that this classification satisfactorily covered all biological methods. Still, it was necessary to make an additional subdivision within every class of methods depending on the level of analysis and to select accordingly: (1) the methods for detection of signals on a noisy background (D); (2) methods of phenomenological description (P); (3) analytical methods (A); and (4) methods of systemic analysis (S).

As an indicative method of qualitative analysis answering the question "what kind of", that is, to provide information on the general properties of an object or process, the following methods can be considered:

The methods of detection of signal in the presence of noise or separation of an object

As a result of applying these methods the following can be discovered.

—pure noise
—a signal or a complex of signals

The signal can be separated by the following means:

(1) Filtering—the detection of a signal with known parameters or elimination of noise with known frequency spectrum.
(2) Different ways of smoothing, like least squares method and special type filtering.
(3) Accumulation of signal (synchronous detection, correlation methods).
(4) Identification of signal presence by applying test stimuli.

Methods for testing of stationarity of the processes or, in general, testing for homogeneity within a group of objects

This method implies the reproducibility of characteristics of the process or object under given conditions.
and in some cases the ergodicity—the stationarity both in time and over the ensemble of realizations.

The following processes can be identified by comparing sequential fragments of the process or by sequentially measuring parameters of an object and also by comparing sequential fragments and separate realizations within the ensemble of realizations:

- nonstationary
- quasistationary
- stationary
- ergodic

Stationarity can be tested as follows:

1. Calculation of the variance of the parameters as a measure of stationarity.
2. Comparison of sequential fragments of the processes or sequentially measured parameters of an object, with the aid of statistical criteria and an analysis of the matrix of criterion values.
3. Calculation of sequential or general correlation coefficients and their comparison in order to detect any dependence on the point of sampling.
4. Plotting of the sequential spectra (Walsh, Fourier or other basis) to detect any dependence on the point of sampling.
5. Calculation of the transition probability matrix and estimation of its numerical parameters.
6. Calculation of reproducibility matrices—the matrices of a criterion value for the sequential fragments of the process.
7. Selection of the optimal value of intervals of stationarity tested by different methods (primarily, variance analysis).

By applying these methods we learn whether we can use numerical parameters to describe the whole realization (as for stationary or ergodic processes) or only intervals of stationarity (in case of a quasistationary process).

In the absence of stationary fragments we must find other informative parameters, for instance the degree of stationarity.

**Methods for analyzing the periodicity of the processes**

As a result of their application we can determine whether the signal is:

- periodic with constant period
- quasiperiodic (with changing period)
- aperiodic

The detection of periodicity is possible with the help of the following methods:

1. Methods of approximation.
2. Correlation methods.
5. Analysis of extremal values.
6. Cosinor method.
7. The calculation of a matrix of reproducibility—the matrix of criterion values for sequential fragments of the process.

**Methods for investigating the internal coherence of the process**

As a result of this investigation the following can be obtained:

- Markov processes of the 1st, 2nd and higher orders.
- Coherent processes.

The test of coherence requires:

1. Calculation of the transitional probability matrix and the estimation of the order of the Markov process.
2. Estimation of the interval of coherence using the correlation function.
3. Calculation of matrices of reproducibility.

In case a high degree of coherence is discovered filtering of data can be used to reduce their volume without the loss of information.

A high degree of coherence of a process also indicates high inertia of the studied object; that is of practical importance for the study of transient processes.

**METHODS FOR EVALUATING NUMERICAL PARAMETERS OF OBJECTS, EVENTS AND PROCESSES**

To answer the question “what kind of” in the narrow sense of the word are the following:

**Different ways of estimating signal-to-noise ratio:**

1. Smoothing of numerical data sequences.
2. Optimal linear and nonlinear filtering.
3. Elimination of noise with known frequency parameters.

**Methods for evaluating integral parameters of a process**

They are different for periodic and aperiodic processes. For periodic processes the following methods are applicable:

1. Analysis of the probability density distribution:
   a. calculation of statistical moments (mean, variance, coefficients of asymmetry and
excess, and higher moments) and their comparison by means of parametric criteria.

(b) Calculation of distribution functions and their comparison by means of nonparametric criteria (rank correlation, Spearman, Mann-Witney, Kendall, Kolmogorov-Smirnov criteria and others).

(2) Calculation of reproducibility matrices.

(3) Evaluation of time series of values of the periods (spectrum of periods) by means of calculation of intervals between successive extremal values; crossing the zero line; and analysis of statistical properties of these time series.

(4) Calculation of the spectral density function using different bases.

(5) Investigation of the phase structure of processes analyzing the dependence between the accumulated phase shift and time.

(6) The “Cosinor” method.

(7) Calculation of the transfer function by means of operators method.

(8) Method of test stimuli.

(9) Plotting of the amplitude-frequency characteristics. For aperiodic processes we may apply:

(10) Analysis of amplitude distribution and calculation of moments.

(11) Analysis of power spectrum.

(12) Investigation of overshoots and their statistical properties.

(13) Calculation of the sequence of intervals between events.

(14) Investigation of patterns and their reproducibility.

Methods for detection of coherence of the process

Calculation of the transitional probability matrix and its numerical parameters (trace, determinant etc.).

ANALYSIS OF THE INTERCONNECTIONS BETWEEN ELEMENTS, PHENOMENA AND PROCESSES

As a separate parameter to be investigated (in response to the question “what kind of connection is this?”), we primarily selected the following features of interconnections:

a. Type of interconnection—linear or nonlinear. Within the nonlinear group we study those which can be represented by polynomials of higher order, harmonic functions, exponential, hyperbolic interconnections and so on.

b. Degree of linearity.

c. Direction of interconnection (vector of interconnection).

Tensor of interconnections can be used accordingly to describe completely the interconnections within the group of elements.

The first routine stage in the analysis of interconnection is the calculation of the correlation ratio indicating the presence or absence of any connection. If this parameter exceeds certain thresholds special tests concerning linearity of interconnection can be carried out. For that purpose the correlation coefficient reflecting the strength of linear (or approximately linear) interconnections is calculated.

The direction of interconnection can be obtained either by using the time shift of the principal maximum of the crosscorrelation function or by calculating the regression lines of each process relating to the other ones.

Nonlinear interconnections can be successfully approximated with linear functions within short intervals, principally for monotonic increasing or decreasing functions. In case such an approximation is not accurate enough and for periodical interconnections, we choose empirically approximating functions from routine methods (least squares, maximum likelihood, minimization of objective functions, nonlinear programming). The direction of interconnection can be specified by means of regression or factor analysis.

To investigate the interconnections on different structural levels and between the objects of different complexity we must take the following steps:

(1) In case the parameters of a single object are to be analyzed:

(a) Calculation of value and sign of correlation ratio.

(b) Calculation of correlation coefficient to estimate the degree of linearity of the interconnection.

(2) To analyze the interconnections between elements and elementary events, calculate vector of interconnections.

(3) When analyzing the interconnections within the system calculate tensor of interconnections.

METHODS OF INVESTIGATING THE EXTERNAL PROPERTIES OF PROCESSES AND OBJECTS

These are relating to the environment and answering the questions “what kind of,” “when,” “where” and “what”.

Methods of calculating the unobservable

Observation relates here to the field of metrology space-time coordinates. Such calculations are carried out with the aid of:

(1) Crosscorrelation analysis.
(2) Description of objects' motion in space-time coordinates by means of differential equations.
(3) Calculation of dynamic transfer functions (related to relaxation processes) by routine methods.
(4) Evaluation of accumulated phase shift as a function of time.

Methods of sorting and distributing objects according to certain indicators

(1) Calculation of statistical moments.
(2) Calculation of distribution function.
(3) Calculation of the multidimensional joint probability density functions and estimation of their parameters.

Methods of analysis of motion and functioning of an object in a space-time coordinate system

(1) Calculation of matrices of reproducibility.
(2) Construction and solution of pertinent differential equations.

Methods of identifying and classifying objects, events and processes

As a possible way to answer the question “what” we determine whether one observed phenomenon is similar to another. The problem is to identify an object as belonging to a certain class—it is pattern recognition. In such cases the following methods could be applied:

(1) Search for the complete or partial precedent.
(2) Diagnostic procedures in multidimensional space on the basis of Bayesian or other criteria.
(3) Calculation of discriminant functions of the first and higher orders.

Cases where there is no a priori information on classes

Their images can be constructed using:

(1) Factor analysis—principal components in particular.
(2) Cluster-analysis.
(3) Adaptive classification (with learning).

The enumeration of these methods shows that any one method can belong to more than one group (like correlation analysis which can be applied to signal detection, to testing the periodicity and for estimating the numerical parameters of the process). From our point of view, this does not make the suggested classification less convenient because of the different interpretations of the results in each case.

The complete classification is given in Table VI where the arrows denote what questions can be answered by means of a particular method. Of course, not all the possible mathematical methods are exhibited but only those widely used in the solution of biomedical processing tasks according to the literature.

Not being mathematical specialists we have undoubtedly missed a few “tools” which may appear more or less important to others, but the open-ended structure of the classification scheme allows us to add anything necessary. There is also no doubt that from the pure mathematician’s point of view this classification is rather eclectic. However, our aim was not mathematical harmony but the distribution of analytical tools over the “shelves” to make it convenient for the user to take the needed one from there according to his/her requirements.

FEASIBLE FORMULATION OF BIOMEDICAL PROCESSING TASKS

The difficulty of choosing an adequate method of analysis derives not only from the above mentioned large number of possible methods and the complexity of the problems themselves. This was only the tip of the iceberg. Complications arise when optimal time periods must be selected (also the correct digitizing interval) for a given method in order to find the desired parameters. However, it is impossible to discuss these problems here in detail.

The way to overcome these difficulties seems to be formalization of the approach to the statement of biomedical data processing tasks, thus guaranteeing the correct approach.

As a first step along this line we suggest the use of a standard form with the following items:

(1) Parameter (or parameters) to be investigated.
(2) Intervals of parameter values and the given precision of the measurement of each one.
(3) Duration of observations (adding criteria of choice if there are any).
(4) Digitizing interval (number of samples per observation).
(5) Object of analysis (elements, elementary events, processes, interconnections).
(6) The real aim of the analysis (to answer the questions “What kind of,” “Where,” “When,” “What”).
(7) The final aim of the analysis (biomedical problem).

These seven pieces of information are supplied by the research worker; the people who deal with the data processing tasks should add one more:

(8) Results of preliminary analysis of data by means of indicative methods (general information about the properties of the object or process).

Here the following are useful:

(a) signal-to-noise ratio.
(b) whether the process is stationary and the in-
TABLE VI—Quantitative Methods of Investigation of Internal Properties of Objects, Events and Processes

(c) statistical parameters, primarily the type of distribution.
(d) whether the process is periodical.
(e) results of Markov property test.

Analysis of data thus obtained in combination with the suggested classification schema will facilitate the choice or the optimal method of data processing. In case there are several possible methods one should remember the Okkam razor and choose the simplest one so as not to increase unnecessarily the number of essential descriptors.

In conclusion we would like to stress that the choice of data processing methods must be based upon their applicability to the data at hand. A bad choice will lead to the data having to be “reprepared” in a perhaps less useful format.

REFERENCES