Graphical Representation of Proteins†

Milan Randić,*,†,‡ Jure Zupan,† Alexandru T. Balaban,§ Dražen Vikić-Topić,‡ and Dejan Plavšić‡

National Institute of Chemistry, P.O. Box 3430, 1001 Ljubljana, Slovenia; NMR Center, Ruđer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia; and Texas A&M University at Galveston, Galveston, Texas 77553

Received February 29, 2008

Contents

1. Prologue

This review article is, in a way, different from traditional review articles in that, besides being informative and reporting on the past published work in the area of graphical representation of complex biological data, it has an additional purpose in being provocative with respect to potentials of novelty in the future, rather than dwelling excessively on the accomplishments of the past. The review is factual; it is introductory and trying to be comprehensive with respect to its central topic—the graphical representation of proteins. Because it covers a topic, which may be viewed to be out of the mainstream avenue of

† "The most rewarding work is usually to explore a hitherto untouched field. These are not easy to find today. However, every once in a while some new theory or new experimental method or apparatus makes it possible to enter a new domain. Sometimes it is obvious to all that this opportunity has arisen, but in other cases recognition of the opportunity requires more imagination." E. B. Wilson (*Introduction to Scientific Research*; McGraw-Hill: New York, 1952.)

* To whom correspondence should be addressed. E-mail: mrandic@msn.com.
† National Institute of Chemistry.
‡ Ruđer Bošković Institute.
§ Texas A&M University at Galveston.

10.1021/cr800198j
Published on Web 10/12/2010

Milan Randić is a native of Croatia and a citizen of the United States and Croatia, born in 1930. He studied Theoretical Physics at the University of Zagreb during 1949−1953 and studied for the Ph.D degree at the University of Cambridge, England (1954−1958). From 1960 to 1970, he was at the Rudjer Bošković research institute in Zagreb, Croatia, where he founded the Theoretical Chemistry Group. During 1971−1980, he was visiting various universities in the United States, including Johns Hopkins, MIT, Harvard, Tufts, and Cornell. After 1973, his research became oriented towards applications of Discrete Mathematics and Graph Theory, in particular for characterization of molecules and biomolecules. During 1980 to 2000, he was professor in the Department of Mathematics and Computer Science at Drake University, Des Moines, Iowa. During the past 15 years, he has spent three months each year at the National Institute of Chemistry, Ljubljana, Slovenia, collaborating with scientists from the Laboratory for Chemometrics. He is a member of the Croatian Academy of Sciences and founder of the International Academy of Mathematical Chemistry, the seat of which is in Dubrovnik, Croatia. Since the year 2000, his research has been mostly shifted towards Bioinformatics, with particular emphasis on graphical representation and numerical characterization of DNA, proteins, and proteome, though his fascination with Kekulé valence structures and aromaticity remains undiminished. In 1996, he obtained the Herman Skolnik Award of the Division of Chemical Information and Computer Science of the American Chemical Society. He is Honorary Member of The Croatian Chemical Society, The International Academy of Mathematical Chemistry, and The National Institute of Chemistry, Ljubljana, Slovenia. His other interests include development of a universal pictographic language Nobel.

Jure Zupan was born in Ljubljana, Slovenia, in 1943. He studied Physics at the University of Ljubljana and graduated in 1966. He obtained his Ph.D. in Chemistry in 1972. He did his first research on the magnetic properties of solids at the Josef Stefan Institute, Ljubljana (1963−1973). In 1974, he joined the National Institute of Chemistry in Ljubljana to work on Computerized Databases, Chemometrics, and Artificial Intelligence. He did his postdoctoral research at ETH Zürich (1975) and at NIH, Bethesda, MD (1978). Since 1985, he is a Full Professor at the University of Ljubljana. He was Visiting Professor at the Arizona State University in Tempe, Arizona (1982), at the Vrije Universiteit Brussels, Belgium (1988), for 3 consecutive years (each year for three months) at the Technical University Munich, Germany (1990−1992), and at the University Rovira i Virgili, Tarragona, Spain (1995). After 1988, his research focused on the field of Artificial Neural Networks. He wrote with Johnny Gasteiger a very successful book *Neural Networks for Chemists*, which received more than 500 citations and was nominated the *book of the month in 1993*. The book was translated into Chinese and Japanese and was expanded with more practical applications as a second edition under the title *Neural Networks and Drug Design*. He is now mostly interested in the multidimensional data representation and context extraction from large assemblies of multidimensional data. He is a member of the European Academy of Science (Salzburg) and member of the Engineering Academy of Slovenia. He is author and editor of 10 books and monographs and has written more than 200 articles in different journals and books.

comparative protein studies and an outline of events along a path not frequently trodden, we included sufficient introductory material on closely related topics of graphical representations of DNA and the secondary structure of RNA. It is not possible to study and understand proteins without studying and understanding DNA and RNA, so inclusion of material on DNA and RNA has been almost mandatory.

Graphical representation of biosequences was a novelty 25 years ago as it is also today, as the subject expanded from DNA, RNA to proteins and proteomics and as it grew from qualitative and pictorial representations to quantitative and numerical characterizations of biosequences. In science, whenever one enters a novel territory, one comes across novel challenges and old animosities. Graphical representation of proteins is not an exception, even though by its significance it cannot be compared to Pleiades of outstanding discoveries that were met by hostile reception. Recall how Van't Hoff was greeted by Kolbe when he advanced the notion of three-dimensional tetrahedral structure of carbon atoms;[2] recall the hesitation of the physical chemistry community toward the novelty of quantum mechanics and the difficulty of Pauling and others to publish their work in *Journal of Physical Chemistry*, which led in 1933 to the initiation of the *Journal of Chemical Physics*;[3] recall the skepticism toward the pioneering work of Samuel Francis Boys[4−9] on quantum chemical ab initio calculations based on Gaussian orbitals (for which all molecular integrals could be exactly computed, which was not the case with Slater-type orbitals that were exclusively used at that time); recall in turn the distrust of the Gaussian ab initio quantum chemists toward DFT, the density-functional theory,[10−12] which is today accepted, no doubt mostly due to the Nobel Prize awarded to W. Kohn, "*for his development of the density-functional theory*", who shared the 1998 Nobel Prize for chemistry with J. A. Pople, "*for his development of computational methods in quantum chemistry*."

Some skepticism is welcome in science, but skepticism combined with a lack of imagination, which illustrates all the above-mentioned cases of hostility toward novelty in science and numerous other fields is, to say the least, unproductive. Hostility mostly reflects a lack of understanding of the subject combined with fear as if the novelty necessarily threatens the existing status quo. The novelty of graphical representations of biosequences is not in that they are very abstract and difficult to comprehend but that they appear very arbitrary! This may be sufficient for some to raise objections without realizing that arbitrary approaches in science are not whimsical, but are done with a reason and according to some plan. Hence, such models, while arbitrary conceptually, differ little from use of *nonobservable* quantities in science, which are widespread in modeling, with chemistry in particular, as can be seen from Table 1, where we have listed several such tools merely to remind readers of their common occurrence.

It is not essential for a model to be based on observable quantities but to provide *novel insights* into structure−property

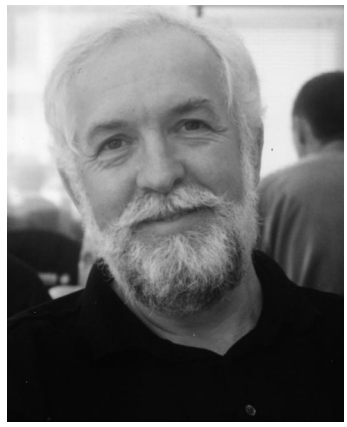
Alexandru T. Balaban was born in Timisoara, Romania, in 1931 and is a citizen of the United States and Romania. He studied in 1949−1953 at the Faculty of Industrial Chemistry, Bucharest Polytechnic University, and graduated as a chemical engineer; his subsequent Ph.D. thesis described a new synthesis of pyrylium salts (the Balaban−Nenitzescu−Praill "name reaction"). Between 1956 and 1999, he taught organic chemistry at the same university except for 1967−1970, when he was a senior research officer at the International Atomic Energy Agency in Vienna, working with radiopharmaceuticals. In the new millennium, he became a professor of chemistry at the Texas A&M University at Galveston. His chemical research led to new syntheses of oxazoles and indolizines, to new boron or arsenic chelate compounds, to persistent nitrogen-centered free radicals stabilized by push−pull electronic effects, and to the discovery of the catalytic *automerization* of ^{13}C-labeled phenanthrene. He pioneered Chemical Applications of Graph Theory and edited the first monograph with this title in 1976, which was translated into Chinese. Such applications involve valence isomers of annulenes (he coauthored a three-volume book on this topic in 1987), theoretical aspects of aromaticity including *dualists* of polycyclic aromatic hydrocarbons, *fibonacenic* or *isoarithmic* benzenoids (italics denote terms he introduced). The first of the three smallest trivalent graphs with girth 10, and the unique such graph with girth 11, are known as Balaban 10-cage and Balaban 11-cage, respectively. He described several topological indices: the average distance−sum connectivity is known as the Balaban index *J*, and is used successfully in quantitative structure−activity or structure−property relationships. The Division of Chemical Information of the American Chemical Society honored him with the 1994 Herman Skolnik Award. In 1963, he was elected as a corresponding member of the Romanian Academy, but had to wait until 1990 for being promoted as titular member (academician). In 2001, he was elected as honorary member of the Hungarian Academy of Sciences. He was the first president of the International Academy of Mathematical Chemistry (2005−2007).

D. Vikić-Topić was born in Zagreb, Croatia, in 1954. He received his bachelor's degree in 1978 in inorganic chemistry, working on intermetallic compounds at the University of Zagreb, Croatia, under the guidance of Prof. Z. Ban. He began his research career in 1979 at the Ruđer Bošković Institute in Zagreb, in NMR and molecular spectroscopy. In 1986, he obtained his M.Sc. degree in theoretical and physical-organic chemistry, and in 1988, he obtained his Ph.D. in NMR spectroscopy, both under the supervision of Prof. Z. Meić, at the University of Zagreb, working on fluorine effects, deuterium isotope effects, and long-range interactions in NMR spectroscopy. He did postdoctoral research in 1989 at National Institute of Chemistry, Ljubljana, Slovenia, with Prof. J. Kidrič and in 1990 at the Indian Institute of Science, Bangalore, with Prof. C. L. Khetrapal. From 1991 to 1993, he was Visiting Associate with Dr. E. D. Becker at National Institutes of Health, Bethesda, MD, investigating different bioorganic molecules by NMR spectroscopy and working on the experimental determination and the theoretical concept of long-range deuterium isotope effects on carbon-13 chemical shifts. In 1994, he was Senior Research Associate with Prof. S. Macura at Mayo Clinic and Foundation, Rochester, MN, working on mechanisms of the transfer of magnetization and theoretical calculations of quantitative NOE. From 1995 until now, he is studying organometallic compounds, especially with mercury and cadmium, by NMR and other spectroscopy methods. He spent 1999 as a Visiting Professor at National Institute of Chemistry, Ljubljana, Slovenia, working on NMR of organomercurials with Prof. J. Plavec. In recent years, he is more involved in mathematical chemistry and application of its methods in investigations of biopolymers.

relationships, whether they pertain to individual molecules, physicochemical molecular properties, bulk properties, QSAR (quantitative structure−activity relationship), or to biomolecules and their properties, including their interactions, their folding, and their alignment. So the ultimate judgment on whether a novelty is desirable or not will be decided by those who tried such models, not those who read about them! Recall the proverb *the proof of the pudding is in the eating*? This is the main reason that we have outlined enough details of historical developments of graphical representations of DNA, RNA, and proteins and enough computational details so that any interested reader can verify results and duplicate computations, if warranted.

It is up to interested individuals to evaluate sections of this review and judge on the novel insights they offer to traditional bioinformatics, but before discarding the novelty as being redundant in the sense that the "available methods have served us well", as one skeptic who reviewed some of our work has lamented, let us recall one aspect and one result that "traditional" approaches could not do or failed to do. First, in contrast to computer manipulations of DNA and protein sequences, which always consider at least *two* or more sequences at a time, analyses based on graphical approaches to DNA and proteins apply also to a *single* DNA or protein. This makes it possible to compile a catalogue on individual DNA and protein molecules, which will have more than sequential lists of nucleotides or amino acids, analogous to similar catalogues on molecules, which besides formulas have lists of selected molecular properties (e.g., the Merck Index). Second, graphical representations of DNA stimulated work on extending some of these graph-theoretical approaches to quantitative characterization of proteomics maps. One such application resulted in a nonlinear quantitative characterization of dose−response of rat liver cell proteome (the collection of cellular proteins) to variation in concentration of peroxisome proliferators. Such J-shaped dose−response of drugs is known as hormesis (the term for generally favorable biological responses to low exposures to toxins). Hormesis has been reported for whole organisms for quite some time. It is interesting that, already in the 16th century, Paracelsus had recognized that the efficacy of medical use of small amounts of toxic chemicals depended principally on the dose.[13] For the first time, however, hormesis has been reported on the cellular level very recently.[14]

For the past two decades, we have witnessed well-documented hostile opinions of some theoretical chemists toward chemical graph theory[15] and occasional flares of hostility toward extensions of graph theory to bioinformatics. So in the spirit of preventive medicine, this article also has



Dejan Plavšić was born in Zagreb, the Republic of Croatia, where he received his primary and secondary education. After obtaining his B.Sc. degree in Chemistry from the Faculty of Science, University of Zagreb, he joined the Department of Physical Chemistry at the Ruđer Bošković Institute in Zagreb. He earned the M.Sc. degree in Chemistry from the University of Zagreb in 1980, and received his Ph.D. degree in Chemistry in 1984 from the same university under the supervision of Professor Jaroslav Koutecký, Institut für Physikalische Chemie, Freie Universität Berlin, Germany, and Professor Leo Klasinc, Ruđer Bošković Institute. In 2005, Dr. Plavšić joined the NMR Center at the Ruđer Bošković Institute, where he is a senior research scientist. His research interests are in the areas of mathematical chemistry, chemical graph theory and its applications, metal clusters, organometallic compounds, and catalysis. Since 2003, he became interested in extending graphical and numerical characterizations to DNA, proteins, and proteome maps using graph theoretical tools, entering thus into Bioinformatics. Two of his papers in this area have received recognition as being among 50 Most Cited Papers in Chemical Physics Letters during the 2003−2007 period. In 2005, when the International Academy of Mathematical Chemistry was founded, he was one of the initial founding members of the Academy. In 2006, Dejan Plavšić was the first recipient of the International Award Latium between Europe and the Mediterranean for Medicine, Physics or Chemistry presented in Rome, Italy.

Table 1. Illustrations of Nonobservables Quantities and Concepts of Physics and Chemistry

nonobservable	discipline
atomic orbitals	quantum theory
molecular orbitals	quantum chemistry
hydridization	quantum chemistry
potential curve	spectroscopy
aromaticity	chemistry
bond dipole moments	physical chemistry
ring currents	NMR
reaction coordinate	chemistry
molecular descriptors	QSAR

a mission to disperse doubts about the role of graph theory in bioinformatics. That aspect of this review article is what can be characterized as its "provocative" element, which ought to be differentiated from "propaganda", which one occasionally notices in opinions of some of the critics of novelty in science. Apparently most critics of mathematical novelty in chemistry and biology prefer to remain anonymous as reviewers of articles submitted for publication. Most such critics are also unaware that graph theory is synonymous to combinatorial and topological explorations to molecular structure, DNA, RNA, proteins, and proteomics, and while graphs may be simple, graph theory is not. Recall the definition of "propaganda" as given by J. W. Cornforth,[16,17] who shared the 1975 Nobel Prize in Chemistry with V. Prelog:

"*Propaganda is that sort of art of lying that consists in deceiving your friends without deceiving your enemies*."

Let us continue by mentioning several additional "historical blunders" in science (to use the characterization of Calabrese,[18,19] which reflects on the fact that novelty in science is not well received by opinioned Establishment. We could start with the historical blunder in pharmacology, related to the nonlinear nature of dose−response relationship, which Calabrese considered. Most so-called controversies in science originate with scientists expressing their opinions without realizing that opinions are often undocumented views. Lyn Margulis, the distinguished biologist, commented on "opinions" in science by stating:[20]

"*So I don't see how people can have strong opinions. ... Let me put this way: opinions aren't science. There is no scientific basis! It is just opinion!*"

She knows what she is talking about because one of her own papers, a paper on the fundamentally new "theory of the origin of eukaryotic cells (that is, cells that divide by classical mitosis)" was rejected by about 15 scientific journals, to be finally accepted and published by another journal to which she submitted her work. This some 15 times rejected paper is today considered as a paper that made significant progress in her field of research and is recognized "as a landmark and a key to the understanding of the genesis of organelles."[21] Recall also the hostility toward David Baltimore, the leading molecular biologist, who was accused of fraud and even targeted by a congressional investigation of scientific fraud, because some results made in his laboratory could not be reproduced. Some went so far as to request that his Nobel Prize be rescinded and to have him expelled from the National Academy of Sciences, but about 10 years later it turned out that David Baltimore, who stood behind the results of his younger collaborator, was found to have been right all the time.

As Max Planck has said:[22]

"*... a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it ....*"

Thus, it is not our intention here to convince those who are satisfied with the "tool" they already have (such as FASTA,[23] BLAST,[24] and other computer software) about potentially useful novelties of the graphical approach to the characterization of proteins but rather to alert open-minded scientists and newcomers about the arrival of a new "tool", which may or may not be of interest to them for dealing with their problems. Being also interested in comparative studies of DNA, RNA, proteins, and even proteome, we are not in *competition* but in *cooperation* with the existing methodologies. There is no doubt that FASTA, BLAST, and other computer software will continue to be useful; nobody questions that. The question is whether there is room for novelty, a question that sometimes falls on deaf ears. As is well known, it is generally premature to speculate on the potential of novel schemes at their early stages of development, as was the answer of Faraday to the question about the usefulness of electromagnetic induction, so we leave the matter and will see what the future will bring. This is also our answer to those impatient scientists who have a limited background and are overlooking the fact that "*every once in a while some new theory or new experimental method or apparatus makes it possible to enter a new domain.*" We are not in the business of predicting the future nor selling our products, and the only guidance that we can advance to readers is to alert them to the presence of false prophets, who have been well characterized already 2500 years ago by Confucius, when saying:

Who does not know, and knows that he does not know
Is a child, teach such;
Who knows, and does not know that he knows
Sleeps, awake such;
Who does not know, and does not know that he does not know
Is dangerous, avoid such
Who knows, and knows that he knows
Is wise, follow such.

So before one accepts opinions of others, it is advisable to check on the background of opinioned persons, to see if they know what they think they know. This should include making a check on their formal education; recall a statement by Frederic C. Bartlett (1886−1969), a British psychologist, made over 75 years ago:[25]

"*Educated subjects are likely to understand and remember astonishingly little of any scientific subject concerning which they have been given no specialized training.*"

Hence, views of those without formal education in Discrete Mathematics and Graph Theory (who are mostly physicists and chemists, but not computer scientists) on applications of graph theory to molecular structure are irrelevant. They may be at best amusing (e.g., "*It frankly reminds me of the geometrical puzzles my 10-year old does*", as an anonymous referee of the distinguished *Journal of Chemical Physics* characterized one of publications in chemical graph theory).[26,27] They may be a product of wishful thinking, but most of the time they reflect on the incompetence of speakers, "people who do not know that they do not know." In the case of Chemical Graph Theory, the problem is accentuated by the fact that the subject is deceivingly "simple," but as V. Prelog warned chemists:[28]

"*Pictorial representation of graphs are so easily intelligible that chemists are often satisfied with inspecting and discussing them without paying too much attention to their algebraic aspects, but it is evident that some familiarity with the theory of graphs is necessary for deeper understanding of their properties.*"

However, it is not difficult to test one's own familiarity with the subject, by trying to answer the questions made available for such a purpose.[29] Clearly all these words of caution are not addressed to newcomers, except to alert them about the noise coming from old-timers lacking tolerance to novelty.

2. Historical Introduction

Graphical approaches have been used in science not only to visualize complex relationships but also to represent functional relationships, outcomes of complicated processes, and interactions, as well as to simplify scientific notation. In Table 2, we have collected illustrations of graphical tools for mathematics, physics, chemistry, biochemistry, biology, and computer science, which includes Feynman diagrams,[30] Kekulé structures,[31] phylogenetic trees,[32] and flowcharts.[33] With the emergence of genomics and proteomics, graphical representations of lengthy biosequences expanded into bioinformatics.

Graph Theory,[34−38] a branch of Discrete Mathematics that has facilitated the development of Chemical Graph Theory,[39,40] has introduced mathematical descriptors for use in structure−property studies and their extension, structure−activity studies, which then have led to mathematical modeling of drug design and drug−receptor relationship, thus broadening the traditional QSARs. More recently, the methodology of Graph Theory has been extended toward the study of biomolecules, such as characterizations of DNA and proteins, which are of interest for comparative studies of DNA, RNA, and proteins, including also the problem of DNA and protein sequence alignment. Finally, the methodology of Chemical Graph Theory has reached toward the characterization of the proteome, the collection of proteins inside cellular systems.

Table 2. Illustrations of Graphical tools of Mathematics, Physics, Chemistry, Biochemistry, Biology, and Computer Science

graphical tool	discipline
Young diagrams	mathematics
Feynman diagrams	physics
Kekulé structures	chemistry
resonance graphs	chemistry
degenerate rearrangement graphs	chemistry
retrosynthesis	chemistry
Krebs cycle	biochemistry
metabolic pathways	biochemistry
phylogenetic trees	biology
graphical representation of biosequences	bioinformatics
flowcharts	computer science
network of word meaning	linguistics

In this review, we will concentrate on the development of graphical schemes for graphical and nongraphical representations of proteins and subsequent numerical characterizations of proteins. This topic, even though it has emerged only very recently, appears promising in offering novel quantitative characterizations of similarities and differences among proteins.

We start by outlining the basic elements of selected graphical approaches for characterization of biosequences and will follow with illustrations and discussion of various graphical representations of proteins. We then continue to outline the development of graphical representations of DNA, though we will confine our attention only to a fraction of DNA graphical representations, because our interest here is mostly on proteins and their graphical representations. Besides, there is a recent review on graphical representations of DNA,[41] where readers can find more detailed accounts of various graphical representation of DNA. We have selected, therefore, to describe only those few graphical representations of DNA that have been extended to, or are related to, graphical representations of proteins.

The first graphical representations of DNA have been proposed over 25 years ago by Hamori.[42−46] In contrast, the first graphical representation of proteins emerged only very recently, about five years ago. The first graphical representation of proteins assumes a unique correspondence between one selected collection of 20 nucleotide triplets, the so-called Virtual Genetic Code, and the 20 amino acids, which they represent.[47] The Virtual Genetic Code converts a protein sequence into a hypothetical DNA sequence, and allows one to use available graphical representations of DNA to generate a graphical representation for proteins. More recently, novel graphical approaches were developed for graphical representation of proteins that allow a direct representation of proteins as 2D mathematical objects, without a need to know which codons of RNA that encode amino acids were employed in the biosynthesis of proteins.

We will describe the virtues and limitations of available graphical models for proteins and will express our views on the potential and the promise of graphical representations of proteins for characterizing proteins. It is hoped that

graphical representations of proteins, just as was the situation with graphical representations of DNA, can be used as a tool for seeking a better understanding of the ever-increasing complexity of biological systems. It is important, here at the beginning, to emphasize that graphical representations of DNA and proteins are virtual mathematical objects (usually viewed as 2D geometrical structures and occasionally 3D geometrical structures) that are devoid of any relationship with the factual geometries of DNA or protein structures. Thus, they merely represent alternative *fictitious* "images" of complex systems but are expected to reflect faithfully on similarities and differences of the underlying sequences.

A parallelism may be drawn here with the early molecular models of quantum chemistry, namely, the valence bond (VB) and the molecular orbital (MO) theories. For example, in VB modeling of aromatic compounds, molecules have been typically represented by a set of Kekulé valence structures, while in MO computations, molecules have been typically described by a set of atomic orbitals. As a result of MO calculations, one obtains a set of molecular orbitals for representation of molecular structures. Both the Kekulé valence structures and the set of molecular orbitals are *fictitious* constructions, just as are graphical images of DNA and proteins. There is an additional distinction between graphical representations of DNA, RNA, and proteins on one hand, and quantum chemical models of molecules on the other hand, in that the Kekulé valence structures and various molecular orbitals relate to the actual molecular geometry, which is not the case with graphical representations of DNA and proteins. In that respect, graphical representations of DNA, RNA, and proteins play a role similar to diagrams depicting *nonexisting* objects, such as are diagrams representing the partial ordering, the genealogical (family) trees, or the Tree of Life. These geometrical constructions allow one to visualize relationships between different objects such as, in the case of the Tree of Life, the phylogeny of organisms, that is, the history of lineages as organisms, including animals and plants, as they changed through time.

The motive for the introduction of graphical representations of DNA was to facilitate visual inspection of similarities and differences among lengthy DNA sequences, which is almost impossible without some kind of preprocessing. The pioneers on graphical representation of DNA were Eugene Hamori and John Ruskin,[42] who introduced a geometrical illustration for biological four-letter sequences. The first step in their approach to transform a DNA sequence into a graphical mode was to arrive at a sequence of numbers; specifically, they proposed to use a set of five entries for each nucleic acid. The first four of the five entries stood for the four nucleotide bases: A for adenine, C for cytosine, G for guanine, and T for thymine and the fifth entry represented the sequential position of each base. For example, the sequence ATGGTG, with the following convention,

A = (1, 0, 0, 0)
C = (0, 1, 0, 0)
G = (0, 0, 1, 0)
T = (0, 0, 0, 1)

results in the following numerical sequence

A (1, 0, 0, 0, 1); T (1, 0, 0, 1, 2); G (1, 0, 1, 1, 3); G (1, 0, 2, 1, 4); T (1, 0, 2, 2, 5); G (1, 0, 3, 2, 6);

when subsequent entries are cumulatively added. The last coordinate is the running index indicating the sequential position of each base. Each member of such sequences represents a point in 5-dimensional space, which of course cannot be visualized. To obtain a visual representation of DNA, Hamori and Ruskin placed the four bases at corners of a square, assigning them coordinates A = (1, −1); C = (−1, 1); G = (−1, −1); T = (1, 1,), and added as the third coordinate the running index. In this way, they obtained a 3D graphical representation of DNA, which is referred to by Hamori and Ruskin as an H-curve. The initial 6-base fragment DNA listed above becomes then the sequence (0, 0, 0); (1, −1, 1); (2, 0, 2); (1, −1, 3); (0, −2, 4); (1, −1, 5); (0, −2, 6), to which, as the beginning the origin, (0, 0, 0) was added.

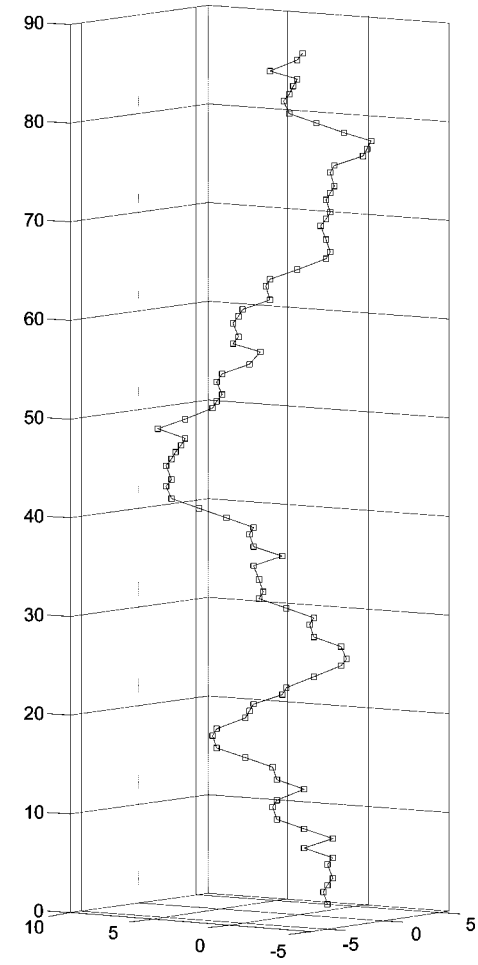


Figure 1. 3D graphical representation of the complete coding sequence of the first exon of β-globin gene of humans, according to the algorithm of Hamori and Ruskin.

The 6-base fragment ATGGTG, taken for illustration of the approach of Hamori and Ruskin, is the beginning of the first exon of human β-globin gene, the graphical representation of which is fully illustrated in Figure 1. The complete first exon of human β-globin gene, having 92 bases, is shown as a zigzag curve placed in a box of size 18 × 13 × 92. Such figures offer visual inspection of similarity of different DNA sequences, but it is clear that, with the increase of the length of the DNA sequence, one is going to lose visual representations of finer details of the H-curve.

The work of Hamori and Ruskin stimulated Gates,[48] Leong and Morgenthaler,[49] and Nandy[50] to follow with construction of 2D graphical representations for DNA instead of 3D representations. These 2D graphical representations of DNA had two limitations: (1) In the case of lengthy DNA sequences, they required considerable space if they were to show enough details of DNA. (2) The graphical representa-

tions of DNA were accompanied with some loss of information. The first limitation, which is a lesser disadvantage, was shared with the 3D graphical representation of Hamori and Ruskin. After all, even the simple listing of lengthy DNA sequences, such as those involving 10 000−100 000 bases, requires space. A more serious limitation is that of accompanying loss of information. That meant not only that 2D graphical representations need not necessarily be unique but also that the depicted DNA did not allow a reconstruction of the initial sequence.

We will see later how both limitations, of which the second is theoretically a more serious limitation, can be removed. However, from the practical point of view, it is unlikely that, for DNA sequences of interest, one will arrive at identical 2D graphical representations of two different DNA sequences, particularly for DNA sequences of moderate size. Moreover, even if such a case would occur, this will not lower the value of the 2D DNA graphical representations of Gates, Leong and Morgenthaler, and Nandy but perhaps may make them even more interesting. This is because graphical representations were introduced to facilitate searches for similar DNA sequences, and DNA sequences having the same graphical representation may, undoubtedly, have other inherent similarities.

3. Graphical Representations of DNA

We start by emphasizing that, whenever one talks about DNA, a simple change of T into U (thymine into uracil) will allow the representation of RNA instead of DNA. However, in the case of RNA, it is generally of more interest to consider representations of the RNA secondary structure, a topic that we will address briefly in this review at the end of this section. The work of Hamori and Ruskin on a graphical representation of DNA was followed by Gates,[48] who viewed a DNA sequence as a "path" in the (x, y) plane. The "path" was obtained by taking the four directions associated with the positive and the negative coordinate axes as directions representing the four nucleotides A, C, G, and T. The 2D representation of DNA is obtained by following the sequential appearance of the bases in a DNA primary sequence and for each base moving one step along the corresponding direction. The resulting path gives the overall graphical representation of DNA, though the information on the individual steps has been lost. The graphical representations of DNA considered by Leong and Morgenthaler[49] and Nandy[50] differ from that of Gates only in selecting alternative assignments for the four directions along the coordinate axes assigned to the four bases. In Figure 2, we have illustrated the graphical representation of the coding DNA sequences in the first exon of β-globin genes of human, lemur, and opossum following Nandy, who assigned A and G along the negative and positive x-axis and T and C along the negative and positive y-axis, respectively.

A close look at Figure 2 reveals advantages and limitations of this particular 2D graphical representation of DNA, which is sometimes referred to as a random-walk plot. The plot, of course, is not random but follows the sequence history of DNA. While the plot may allow the visualization of certain periodic patterns, at the same time it obscures the presence of any repeating patterns in which the path is traversed along already shown steps. This deficiency has been later removed by modifying the four directions for A, C, G and T nucleotides so that they do not oppose one another. Such a modification of the path prevents it from overlapping itself and allows the reconstruction of DNA from its graphical path.[51−56]

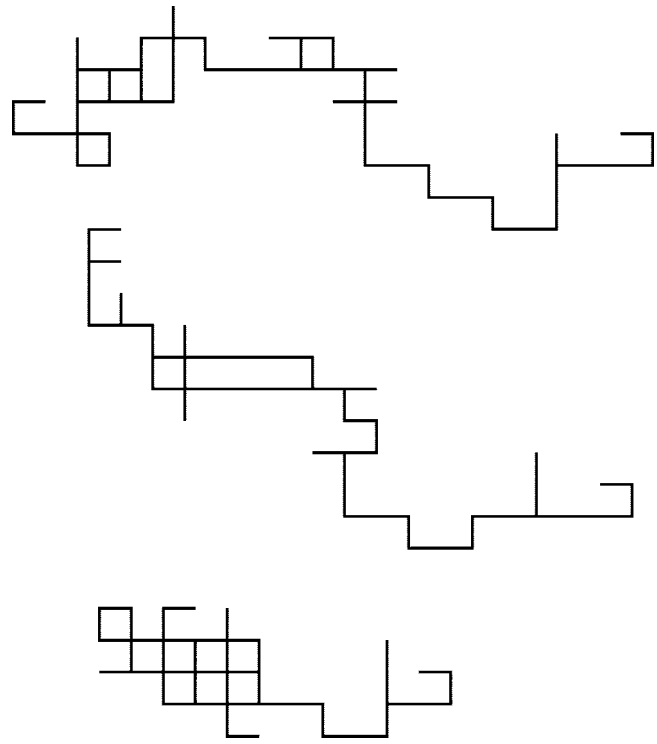


Figure 2. Graphical representation of the coding sequences of the first exon of β-globin gene of human, lemur, and opossum.

3.1. Numerical Characterization of DNA

An important novelty in graphical representations of DNA was introduced only relatively recently, in the year 2000, when it was demonstrated that *graphical* representations could be accompanied by *numerical* characterizations of the considered DNA sequence.[57] This represents an important advancement, which made it possible not only to upgrade the qualitative representations of DNA into quantitative representations but also to delegate comparative studies of DNA to computer-driven analysis. The numerical characterization was based on the approach initially used for characterizing the degree of bending of smaller-chain molecules (conjugated polyene rotamers) that have a fixed geometry in 2D or 3D space.[58] In this approach, which in the past was used for characterization of bending or folding of molecular chain-like structures, one associates with each molecule a special distance matrix, the so-called distance−distance (**D/D**) matrix. The element (i, j) of the **D/D** matrix is defined as the quotient of two distances: the Euclidean distance between atoms i and j though space and the "topological" (graph theoretical) distance between the same two atoms measured along the chain. It turns out that the leading eigenvalue of the **D/D** matrix is a measure of the degree of bending or folding of a chain structure,[58] which is why this matrix remains of interest in structural chemistry.

At first it may appear surprising and conflicting that Graph Theory, and then also Chemical Graph Theory, occasionally deals with graphs as objects of fixed geometry, which also allows (in addition to the graph-theoretical distance based on the count of edges between vertices) the measurement of Euclidean distances between vertices. However, geometrical objects that can be embedded in a regular periodic grid, such

as the square or hexagonal grid in a plane or the cubic or diamond lattices in 3D space, share equal distances and the same angles, and thus, often their properties are of combinatorial and topological, rather than geometrical, origin. Therefore, it should not be surprising to find that carbon skeletal forms for *n*-alkanes, which are chain structures superimposable on the diamond grid, can be specified by *three-letter codes*, because when walking a path along the diamond grid at each vertex, one can select one of *three* routes.[59] For more on the graph−geometry connection, consult articles in the book *From Chemical Topology to Three-Dimensional Geometry*.[60]

A digression to benzenoid polycyclic hydrocarbons may be instructive as it also reflects on the innate relationship between molecular graphs and molecular geometrical structure—the two opposing structural concepts: one of rigid geometry and one of undefined or fluxional geometry. For brevity, benzenoid polycyclic hydrocarbons will be called henceforth benzenoids. They are of two kinds: (i) either flat portions of the "honeycomb" graphene planar net sharing CC bonds or, more generally, (ii) polycyclic aromatic hydrocarbons containing benzenoid rings that share CC bonds, for there is no restriction about planarity so that even three-dimensional benzenoids such as the very stable helicenes or nanotubes are covered by the second definition.

From a different viewpoint, benzenoids may be classified as *cata-condensed (catafusenes)* if they have no carbon atom common to three rings, as in the case of naphthalene, anthracene, and phenanthrene, or as *peri-condensed (perifusenes)* if they have carbon atoms common to three rings, as in the case of perylene or pyrene. Benzenoids can be represented in a simplified form by the so-called "dualist" graphs, which can be constructed by replacing benzene rings with vertices at their center that are connected by edges if they share a CC bond. This leads to an alternative definition for *catafusenes* as benzenoids with acyclic dualists, which may be branched or unbranched graphs, and *perifusenes* as benzenoids whose dualists have three-membered rings. In addition, *coronafusenes* are defined as benzenoids with dualists having larger rings.[61,62]

Observe that, because of the geometrical regularity of the hexagonal planar grid, the dualist graphs for benzenoids of the first kind have a fixed geometry. Because of this, it was possible to devise a simple code for catafusenes by using labels 0, 1, and 2.[62] One starts from one end of the longest string and denotes by zero the angle of 180° between two successive edges, or by 1 and 2 the angles of 120° and 240°, respectively. Once a "left" or "right" turn has been indicated by digit 1 or 2, then this digit is conserved for the dualist. A canonical notation for the catafusene dualist results when, among all possible codes, one selects the code containing the smallest number formed by all digits. Branching is indicated by a pair of brackets containing the digits corresponding to the string of benzenoids forming the branch, with the convention that a branch consisting of a single benzenoid (such as in triphenylene) is denoted by a dot. The code is enclosed in square brackets. Thus, anthracene, phenanthrene, benzanthracene, and triphenylene have canonical codes [0], [1], [01], and [1(.)], respectively.[61,62]

The numbers of resonance structures (Kekulé valence structures) of benzenoids that have no anthracenic subgraph (therefore no digit 0 in their code) are Fibonacci numbers, and therefore, such benzenoids have been called fibonacenes.[63,64] Helicenes whose codes are sequences of digit 1, and zigzag catafusenes whose codes are alternating sequences of digits 1 and 2, are examples of fibonacenes. All fibonacenes with up to eight benzenoid rings have been enumerated in a paper published in 1989.[65]

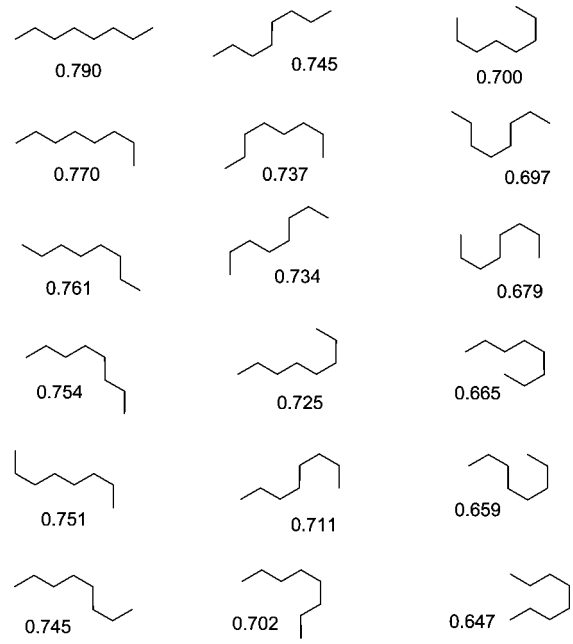


Figure 3. Eighteen skeletons of 1,3,5,7-octatetraene conformers and diastereomers obtained by superposition of CC bonds over graphite lattice and the leading eigenvalue of the corresponding **D/D** matrix.

In Figure 3 are illustrated molecular skeletons of all 18 possible planar rotational isomers (conformers and diastereomers) of an eight-carbon polyene chain, when superimposed on a graphite lattice without overlapping edges/vertices. Next to each structure, we show the leading eigenvalue of the corresponding distance/distance (**D/D**) matrix, the elements of which are defined as the quotients of the Euclidean distance and the distance along the edges between pairs of vertices. As we move from the *all-trans* conformer, at the top left corner of Figure 3, to the highly bent conformer, at the bottom right end, it is apparent from Figure 3 that the leading eigenvalue of the **D/D** matrix gradually decreases. It is even not difficult to see why this has to happen. As molecular skeletons become more bent, the "through spaces", Euclidean distances between distant pairs of atoms, are getting smaller and smaller, while the graph-theoretical distances (measured along the molecular skeleton) remain the same. This reduces the corresponding matrix elements given by the distance−distance quotients as the degree of bending increases and the corresponding matrix row sums become smaller and smaller. As the matrix elements decrease, so do the largest row sums and the smallest row sums of the **D/D** matrices decrease, which, as is known from matrix algebra, determines the upper and lower bounds for the leading eigenvalue of symmetric matrices. So, as molecules get more bent, their leading eigenvalues become smaller.

Some may be surprised to find that Graph Theory, which is concerned with combinatorial and topological properties of graphs, discrete finite systems defined by the connectivity, not their geometry, can be extended to characterize mathematical objects of fixed 2D and 3D geometry. Rigid structures are also objects of interest to Graph Theory.[66,67] One such illustration is the topic of the "unit-distance

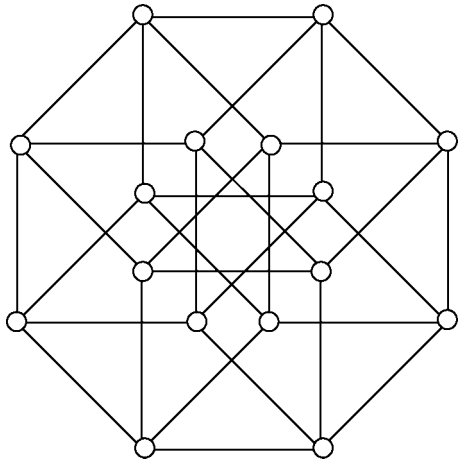


Figure 4. Unit-distance representation of four-dimensional cube.

graphs", which are graphs that can be depicted in the Euclidean plane with all edges having the same length.[68] Already in 1946, Erdös posed the problem of estimating the number of points in a plane that could be placed at a unit distance from each other.[69] In graph-theoretic terms, this problem is related to the question "how dense can be a unit-distance graph?" In Figure 4, we show the hypercube Q_4, a four-dimensional cube, which is one such graph. The simple unit-distance graphs include the cycle graphs C_n (that is, for $n = 3$, triangle; for $n = 4$, square; for $n = 6$, hexagon; etc.), the star graphs S_n (which have one central vertex and $(n - 1)$ adjacent nearest neighbors), and the generalized star graphs, which have one central vertex from which radiate linear branches. As we will see later, generalized star graphs have been used for graphical representations of proteins.[70]

For structures, including graphs, that have a fixed geometry, one can always construct the distance/distance (**D/D**) matrix. The leading eigenvalue, and if necessary, the set of all eigenvalues of the **D/D** matrix may offer useful characterization of such systems, as has been demonstrated on graphs,[71−73] molecular models,[74−76] model proteins,[77,78] proteins,[79] and even proteomics maps.[80−82] However, in contrast to its use for characterization of the degree of bending or folding in smaller molecules and proteins in the case of graphical representations of DNA, RNA, proteins, and numerical characterization of proteomics maps, one is not interested in characterization of the geometry of underlying models but in characterization of the system as a whole. As we will see, in the case of DNA sequences, the **D/D** matrix not only offers additional matrix invariants for their characterization, which may serve as DNA descriptors for a comparative study of DNA, but also allows the recovery of the lost information on DNA primary sequences, which accompanies some of the 2D graphical representations of DNA.

Let us illustrate the construction of **D/D** matrix on a short segment of DNA: ATGGTGCACCTG, which contains 12 of the leading bases on the coding sequence of the first exon of β-globin gene that we have already illustrated in Figures 1 and 2. In Figure 5, we have reproduced the initial twelve steps of the graphical representation of Nandy (top diagram of Figure 2), to which we have added numerical labels 1−12, which indicate the sequential sites of the bases in the considered fragment of DNA. It is this step (putting the sequential labels to the points of the "path" representing DNA) that constitutes the *recovery* of the lost information.

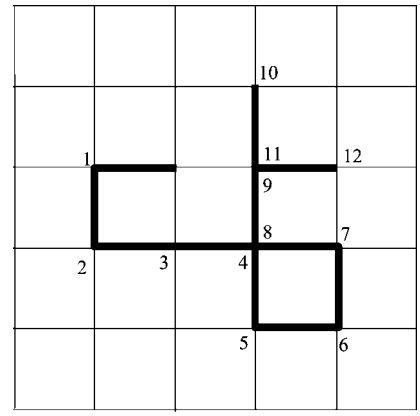


Figure 5. Numbering of the first 12 steps of graphical representation of Nandy of β-globin gene.

As is known from analytical geometry, adding labels is analogous to parametrical representations of functions $f(x)$, and more generally curves $f(x, y) = 0$, which recovers information on the motion, which is fully expressed by the pair of equations: $x = x(t)$; $y = y(t)$.

We have illustrated in Table 3 the initial portion of the **D/D** matrix belonging to the fragment of DNA depicted in Figure 5. As one can see, all matrix elements for the bases that are adjacent are equal to one, because for them the Euclidean and the graph-theoretical distance are equal. The same also holds for the repeating bases in a DNA sequence, because they form linear fragments in their graphical representation, and for linear segments again the Euclidean and the graph-theoretical distances are equal. For other matrix elements, one has to evaluate the Euclidean distance and divide it by the distance along the path, which is given by the number of separating links (in view that all steps are of the same length). Thus, the elements in the first row are: (1, 2) = 1; (1, 3) = $\sqrt{2}/2$; (1, 4) = $\sqrt{5}/3$; (1, 5) = $2\sqrt{2}/4$; (1, 6) = $\sqrt{13}/5$; (1, 7) = $\sqrt{10}/6$; (1, 8) = $\sqrt{5}/7$, and so on. Observe that although the Euclidean distance between the first and the fourth base and the first and the eighth base are the same (equal to $\sqrt{5}$), the matrix elements (1, 4) and (1, 8) are different, because the number of steps that separate bases 1 and 3 and 1 and 7 are different, being three and seven links, respectively. On the other hand, the matrix elements (1, 3) and (1, 5), which involve different distances, are the same because of the simple proportionality.

The graphical representation of Figure 2 of Nandy (and the same is true for the corresponding 2D graphical representations of Gates, and Leong and Morgenthaler) are not only associated with loss of information, which as mentioned above has now been recovered, but are also accompanied with an arbitrary assignment of the four directions of the coordinate axes for the four bases. The same feature, an "arbitrary assignment" among equivalent geometrical alternatives, is true for a number of other 2D graphical representations of DNA. This already was true with H-curves of Hamori and Ruskin, and is also true with the adaptation of the "chaos game" representation of DNA, introduced by Jeffrey[83,84] and based on the mathematical model of the representation of lengthy sequences of random numbers, called the "Chaos Game" by Barnsley and Rising.[85] Incidentally, the highly condensed representation of DNA of Jeffrey, which allows visual representation of DNA sequences having huge numbers of bases, as we will see later, does not suffer from loss of information. The same equally holds for modified 2D graphical representations of DNA of

Table 3. Initial Portion of the D/D Matrix of 2D Graphical Representation of DNA from Figure 5

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	$\sqrt{2}/2$	$\sqrt{5}/3$	$\sqrt{8}/4$	$\sqrt{13}/5$	$\sqrt{10}/6$	$\sqrt{5}/7$	2/8	$\sqrt{5}/9$	2/10	3/11
2		0	1	2/2	$\sqrt{5}/3$	$\sqrt{10}/4$	3/5	2/6	$\sqrt{5}/7$	$\sqrt{8}/8$	$\sqrt{5}/9$	$\sqrt{10}/10$
3			0	1	$\sqrt{2}/2$	$\sqrt{5}/3$	2/4	1/5	$\sqrt{2}/6$	$\sqrt{5}/7$	$\sqrt{2}/8$	$\sqrt{5}/9$
4				0	1	$\sqrt{2}/2$	1/3	0	1/5	2/6	1/7	$\sqrt{2}/8$
5					0	1	$\sqrt{2}/2$	1/3	2/4	3/5	2/6	$\sqrt{5}/7$
6						0	1	$\sqrt{2}/2$	$\sqrt{5}/3$	$\sqrt{10}/4$	$\sqrt{5}/5$	2/6
7							0	1	$\sqrt{2}/2$	$\sqrt{5}/3$	$\sqrt{2}/4$	1/5
8								0	1	2/2	1/3	$\sqrt{2}/4$
9									0	1	0	1/3
10										0	1	$\sqrt{2}/2$
11											0	1
12												0

Table 4. Initial Portion of the D/D Matrix of 3D Graphical Representation of DNA

	1	2	3	4	5	6	7	8	9
1	0	1	$2/2\sqrt{3}$	$\sqrt{11}/3\sqrt{3}$	$4/4\sqrt{3}$	$\sqrt{27}/5\sqrt{3}$	$\sqrt{8}/6\sqrt{3}$	$\sqrt{11}/7\sqrt{3}$	$\sqrt{8}/8\sqrt{3}$
2		0	1	$\sqrt{12}/2\sqrt{3}$	$\sqrt{11}/3\sqrt{3}$	$\sqrt{24}/4\sqrt{3}$	$\sqrt{19}/5\sqrt{3}$	$\sqrt{12}/6\sqrt{3}$	$\sqrt{11}/7\sqrt{3}$
3			0	1	$2/2\sqrt{3}$	$\sqrt{11}/3\sqrt{3}$	$\sqrt{8}/4\sqrt{3}$	$\sqrt{3}/5\sqrt{3}$	$2/6\sqrt{3}$
4				0	1	$2/2\sqrt{3}$	$\sqrt{3}/3\sqrt{3}$	0	$\sqrt{3}/5\sqrt{3}$
5					0	1	$2/2\sqrt{3}$	$\sqrt{3}/3\sqrt{3}$	$\sqrt{8}/4\sqrt{3}$
6						0	1	$2/2\sqrt{3}$	$\sqrt{11}/3\sqrt{3}$
7							0	1	$2/2\sqrt{3}$
8								0	1
9									0

low degeneracy.[51−56] However, in all these cases, alternative assignments of the four bases are possible, which would produce different geometrical representations.

The arbitrary assignment of the A, C, G, T bases to four corners of a square can be eliminated by replacing the four corners of a square by the four corners of a tetrahedron, which, of course, transforms a 2D graphical representation of DNA into 3D. In the case of a tetrahedron, once a single vertex is labeled (by one base), the remaining three corners of the tetrahedron still remain fully equivalent (except for chirality, which is here of no interest). For instance, without loss of generality, one can place the vertices of the tetrahedron at the following (x, y, z) coordinates:[57]

- A $(+1, -1, -1)$
- G $(-1, +1, -1)$
- C $(-1, -1, +1)$
- T $(+1, +1, +1)$.

Because all four tetrahedral directions are equivalent, it does not matter which vertex is assigned to which nucleotide base. In contrast, while the four corners of a square are equivalent, once one of them has been labeled, of the remaining three corners, two are adjacent to one already labeled and the third is not adjacent; thus, they are not equivalent. In Table 4, we show a small portion of the 3D **D/D** matrix of the coding sequence of the first exon of human β-globin gene (as reported in ref 57).

Before closing this section, let us point out that **D/D** matrices,[57] illustrated here on 2D and 3D graphical representations of DNA, have a wider application beyond characterization of linear real or virtual systems. They have been used for the numerical characterization of spectral[86,87] and zigzag representation of DNA,[88] as well as spectral[89,90] and zigzag representations of proteins[91] and proteomics maps.[80−82,92−94] Such wide applicability of **D/D** matrices justifies the above introduction and brief illustration of **D/D** matrices. The **D/D** matrices also offer numerical characterization of the H-curves of Hamori (Figure 1) and the "Chaos Game" representations of DNA, to be considered in the next section, which, as the name indicates, appear "chaotic" but, as we already mentioned, are not associated with loss of information and fully allow the reconstruction of DNA sequences from their graphical diagrams.

3.2. New Metric for DNA Similarity

In this section, we focus our attention on the use of codons as basic elements in comparative studies of DNA and will consider only the coding part of exons of DNA genes, rather than DNA as a whole (thus excluding introns and noncoding introductory or ending parts of exons). This limits applications but, as will be seen, this approach has some advantages with respect to several alternative graphical representations of DNA. A spectral representation of DNA based on codons has been reported[86,87] in which the $4^3 = 64$ codons have been arranged alphabetically (which is as good as a random ordering) on the circumference of a unit circle, thus assigning multiples of $2\pi/64$ radians to individual codons. Besides offering user-friendly depictions of DNA, spectral representations of DNA have important advantages over other 2D graphical representations. The most important advantage of spectral representations is that they allow *arithmetic manipulations* of spectral diagrams, which not only facilitate visual inspection of similarity between DNA sequences but also allow one to search for DNA alignments.[86] Let us now outline this novel graphical representation of DNA based on codons, which has been specifically designed for comparative study of CDS (coding sequences of DNA). As is well known, the correspondence between codons and amino acids is not unique, except for methionine (AUG) and tryptophan (UGG). On the other hand, codons that encode the same amino acid often have the same initial two RNA bases, like for instance, the four codons that encode for proline start with CC, the four codons that encode for threonine start with AC, the four codons that encode for valine start with GU, and the four codons that encode for glycine start with GG. In all above-mentioned cases, the third base obviously does not play a role. However, in the case of aspargine and lysine, or aspartic acid and glutamic acid, both pairs of which are encoded with the same two initial bases, AA and GA, respectively, the third base, which is different, defines the resulting amino acid. In the case of

Table 5. Comparison of the First Dozen Codons of the Coding Sequences of the β-globin DNA Genes of Human and Lemur or Bovine and Goat, Respectively

codon	human	lemur	penalties	codon	bovine	goat	penalties
1	ATG	ATG	0	1	ATG	ATG	0
2	GTG	ACT	14	2	CTG	CTG	0
3	CAT	TTG	14	3	ACT	ACT	0
4	CTG	CTG	0	4	GCT	GCT	0
5	ACT	AGT	4	5	GAG	GAG	0
6	CCT	GCT	9	6	GAG	GAG	0
7	GAG	GAG	0	7	AAG	AAG	0
8	GAG	GAG	0	8	GCT	GCT	0
9	AAG	AAT	1	9	GCC	GCC	0
10	TCT	GCT	9	10	GTC	GTC	0
11	GCC	CAT	14	11	ACC	ACC	0
12	GTT	GTC	1	12	GCC	GCC	0

leucine and arginine, there are two sets of initial bases, e.g., CU and UU for leucine, CG and AG for arginine, both of which have the same second base. Finally, in the case of serine, the codes have the initial two bases completely different: UC and AG. In view of all the above, when comparing codons it is of interest to assign different weights or penalties to first, second, and third base. The weights can be selected in such a way that the dominant base has the larger weight.

The simple assignment of the weights 3, 2, and 1 to first, second, and third base has a disadvantage in that it gives the same penalty for different situations. Thus the penalty (of $3 = 2 + 1$ units) for the case when two codons differ only in the first nucleotide is the same as the penalty when two codons have the same first base and differ in the second and the third base. However, the assignment of the weights: 9, 4, 1, that is: 3^2, 2^2, and 1^2, to codons that differ in the first base, the second, and the third base, respectively, when combined differ for the eight possible cases, as shown below. Here the sign ($=$) signifies the same base in two codons, while the sign ($\neq$) means that the bases are different.

base	1	2	3	4	5	6	7	8
1st	$=$	$=$	$=$	$=$	$\neq$	$\neq$	$\neq$	$\neq$
2nd	$=$	$=$	$\neq$	$\neq$	$=$	$=$	$\neq$	$\neq$
3rd	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$	$=$	$\neq$
penalty	0	1	4	5	9	10	13	14

Compare for illustration the first exon of human β-globin gene with that of lemur and the first exon of bovine with that of goat β-globin gene, the first dozen codons of which are listed in Table 5. As one can see from Table 5, among the first dozen codons only 4 out of 12 are the same for human and lemur, while for the pair of bovine and goat, all the first dozen codons are equal. In Figure 6, we have illustrated the difference of spectral representations for the whole coding sequence of β-globin gene of the human and lemur. As one can see from Figure 6 among the first 30 codons, almost half are identical, indicating a limited similarity for this portion of DNA sequences. However, even though at the beginning of the β-globin genes of human and lemur there is a limited similarity, when the remaining 120 codons are viewed one does not find any agreements, except for possible accidental identity of a few additional codons. Thus, from Figure 6 one has to conclude that the β-globin gene of human and lemur are fairly dissimilar. In contrast, when one compares the bovine and goat β-globin gene (Figure 7), one finds great similarity between the two DNA sequences: they differ only at 16 sites out of 147. The dramatic difference between the β-globin gene of human and lemur, two similar species, is unexpected and suggests that the human/lemur DNA have not been properly aligned.

When aligning coding parts of exons, one should maintain the open reading frame (ORF). An open reading frame is a portion of an organism's genome that contains a sequence of nucleotide bases that encode a protein. So when we shift human−lemur sequences of the β-globin genes by one codon (three nucleotides), we obtain the plot shown in Figure 8, which clearly shows considerable similarity between the two β-globin genes. It is instructive to compare Figure 8 with Figure 6, which shows a lack of similarity between human and lemur when DNA sequences are not aligned. The two figures appear as complementary: in Figure 6 there is similarity between the first 30 codons, while in Figure 8 the similarity is between the remaining 115 codons. By combining the results of Figures 6 and 8, one finds that there is considerable similarity between the coding portion of the human and lemur β-globin gene. The new metric for comparison of coding portions of DNA, as was illustrated, gives satisfactory results when shifting of sequences is considered. The approach is clearly superior to schemes in which DNA sequences of different species are compared solely on their composition of nucleotides, disregarding the possibility

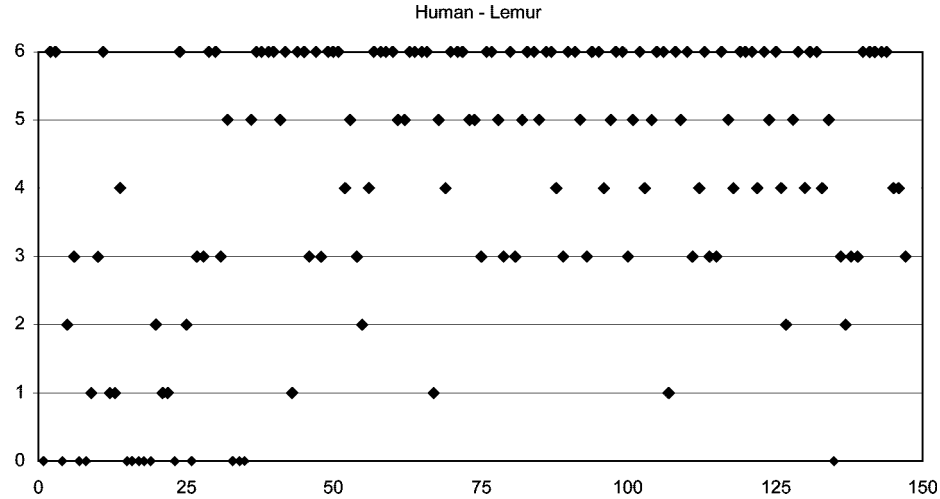


Figure 6. Plot of the weighted difference between codons of the β-globin gene of human and lemur. Spots on the x-axis indicate identical codons while the spot on the $y = 6$ line indicates codons that differ in all three bases.

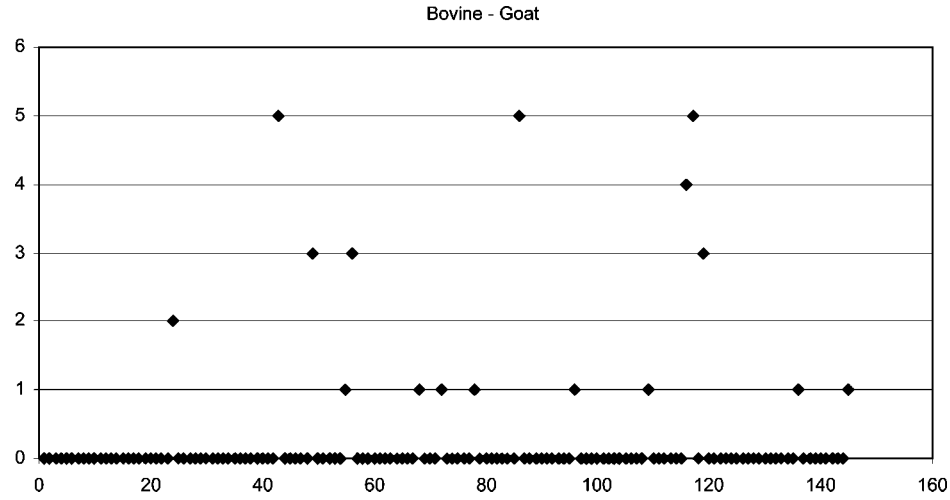


Figure 7. Comparison of bovine−goat β-globin gene.

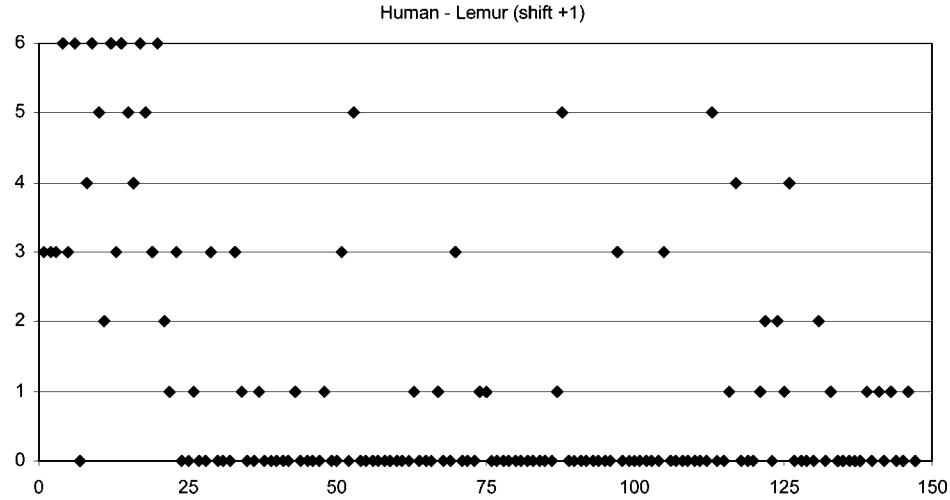


Figure 8. Comparison of human−lemur β-globin gene when the two sequences have been shifted for one codon.

of codon shifts that preserve the open reading frame of the coding part of DNA.

3.3. The Magic Square

The graphical representation of DNA that Jeffrey[83] introduced in 1990 has two important properties: It is highly *compact* and is not accompanied by *loss of information*. The representation has a form of a 2D "map" in which individual bases are represented as spots forming the map. The high compactness of this representation means that, regardless of the length of any DNA sequence, its graphical representation is confined to the interior of a square that can fit on a single sheet of a paper. The first step in the construction of Jeffrey's graphical representation of DNA is to assign to the four corners of a square the four labels A, C, G and T for adenine, cytosine, guanine, and thymine, respectively, as depicted in Figure 9. One depicts individual bases of a DNA sequence as spots in the interior of the square, considering the bases sequentially, one at the time. One starts from the center of the square and moves from the center halfway toward the corner belonging to the first base. From this point, one continues to move halfway toward the corner indicated by the second base and so on.

In Figure 9, we have depicted the locations of the ten bases ATGGTGCACC, which are the initial bases of the first exon of human β-globin gene. We have added arrows

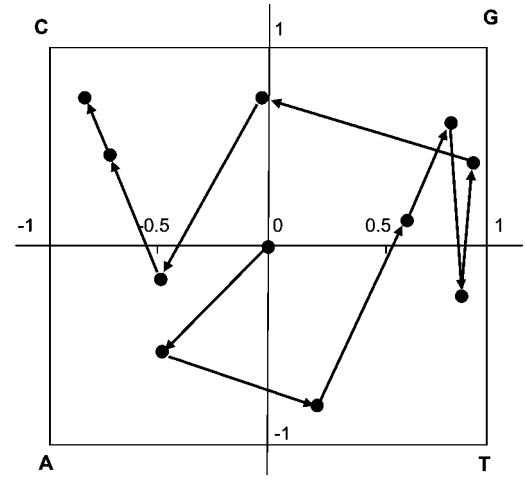


Figure 9. The construction of highly condensed graphical representation of DNA according to Jeffrey.

to illustrate the construction, but graphical representations of DNA of Jeffrey depict only the collection of points, as is illustrated in Figure 10, where is shown the graphical representation of the complete first exon of human β-globin gene (listed below), which has 92 bases (30 codons):

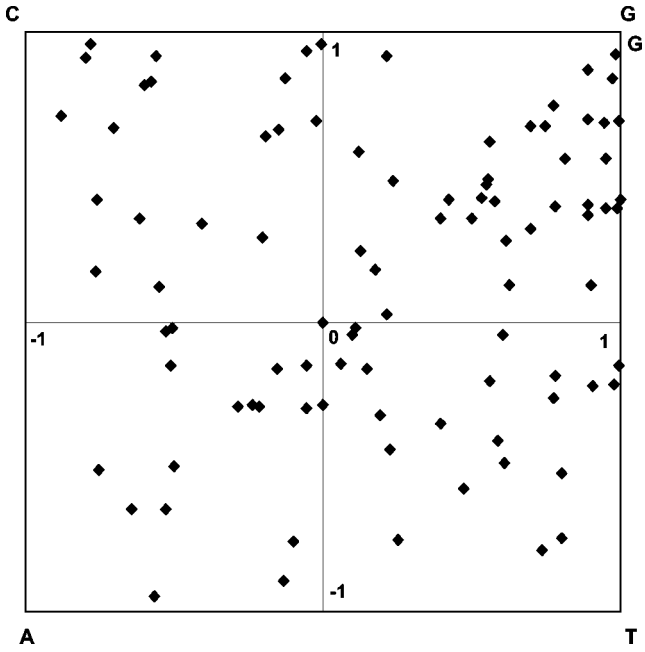


Figure 10. Graphical representation of the first exon of human β-globin gene.

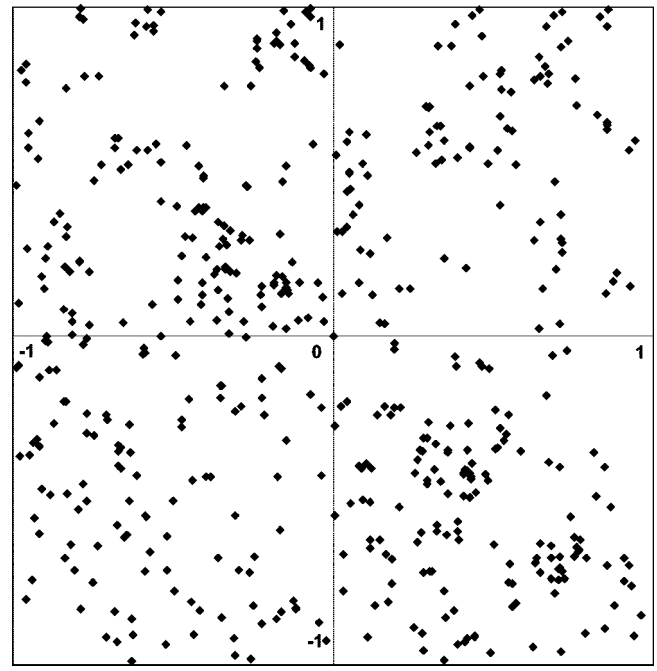


Figure 11. Graphical representation of a proteins having 170 amino acids by 510 spots (three spots coding each amino acid).

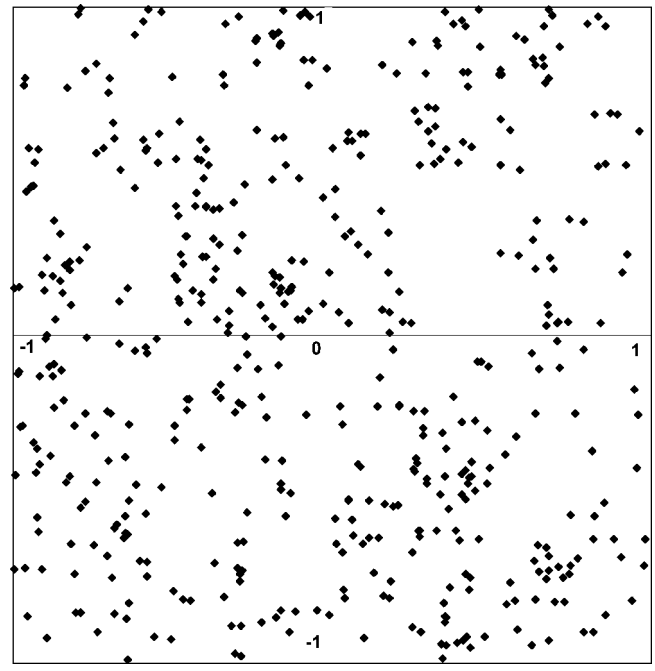


Figure 12. Graphical representation of a proteins having 169 amino acids by 507 spots (three spots coding each amino acid).

ATG GTG CAC CTG ACT CCT GAG GAG AAG TCT GCC GTT ACT GCC CTG TGG GGC AAG GTG AAC GTG GAT GAA GTT GGT GGT GAG GCC CTG GGC AG

The bases have been grouped in codons, which immediately give a hint for the possibility to extend such graphical constructions from DNA to proteins (as will be outlined later).

We should mention that Jeffrey, as well as the mathematician Barnsley who "invented" the Chaos Game in 1988,[85] and others who followed,[96−101] were interested in graphical representations of very long sequences. In particular, Jeffrey considered graphical representations for DNA sequences having 100 000 bases and more. The present authors and their collaborators were the first to consider the chaos game algorithm applied to *short* and *very short* DNA sequences, not only those having a few hundred or at most a few thousand of bases, but also codons, that is, sequences having three bases. In order to appreciate the distinction between the two "camps," in Figures 11 and 12 we show graphical representations of two proteins (to be revisited later), each depicted by some 500 spots, which may suffice to show the apparent "random" nature of such graphical representations and justify the label chaos game for this kind of graphical diagrams. It may not be so difficult to imagine what such diagrams would look like when one would consider 100 times more "spots" than the 2D maps of Figures 11 and 12.

Jeffrey's highly condensed representation of DNA, ingenious in a way, represents a modification of the mathematical Chaos Game of Barnsley. Barnsley considered graphical representations of lengthy sequences of random numbers and introduced the construction in which one chooses first a regular polygon having n vertices, then one selects at random a point in its interior. After selecting at random a vertex of a polygon, Barnsley finds the point that is half way between the first point and the vertex chosen. In the next step, the process continues by selecting another corner of a polygon at random and moving half way toward it, and so on. As a result, various fractal-like pictorial patterns have been generated,[102,103] some irregular, like fern fractals, and some very regular, like the Sierpiński sieve, also known as Sierpiński gasket or Sierpiński triangle,[104,105] construction of which is illustrated in Figure 13. Incidentally the pattern of the Sierpiński triangle is "hidden" in the famous Pascal triangle, which gives a geometrical arrangement of the binomial coefficients that was also already known to early Indian and Chinese mathematicians.[106] To see the hidden Sierpiński triangle, one has to replace all odd entries in the triangle with black dots and all even numbers by blank space, or more technically one has to transcribe the Pascal triangle into binary (mod 2) form.[107,108]

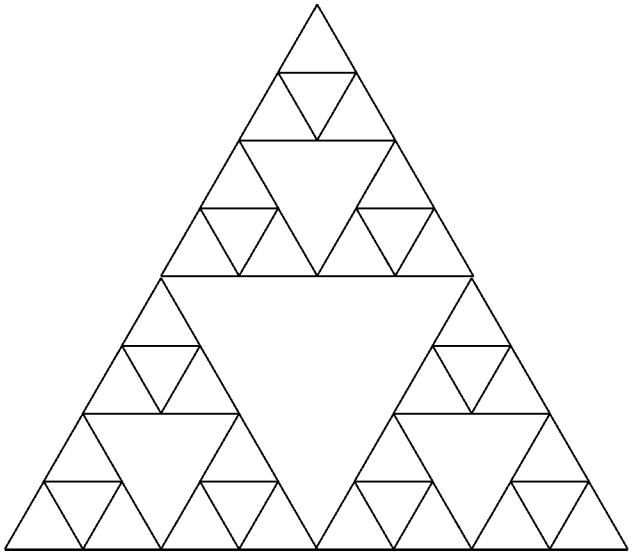


Figure 13. Initial phase in the construction of Sierpiński gasket or Sierpiński triangle.

There appears to be no end to the fascinations with seemingly unrelated mathematical objects. Thus, Gardner[109] has shown that the number of the sides for constructible polygons (those that can be constructed by using only a compass and an unmarked ruler), which are 1, 3, 5, 15, 17, 51, 85, etc., are given as the rows of the binary Pascal−Sierpiński triangle, when binary rows are interpreted in the standard digits, base 10. As if this is not enough, we may add that not only Catalan numbers,[110] which among others give the number of triangles in triangulated polygons, are "hidden" in the Pascal's triangle but also Fibonacci numbers,[111] 1, 1, 2, 3, 5, 8, 13, 21, ..., can be extracted from the Pascal triangle. Fibonacci numbers, which were introduced to count the fast multiplications of rabbits, occur also in chemistry: among other things, they count the number of Kekulé structures in angularly fused benzenoids—naphthalene, phenanthrene, chrysene, fulminene, etc.[112−115] Finally, the Pascal triangle also has been used to construct the "higher-order" Fibonacci numbers, "higher" analogs of Fibonacci numbers.[116]

Jeffrey has modified the Chaos Game constructions of Barnsley by (1) selecting a square as the polygon to be used for graphical representation of DNA, (2) assigning to the four corners of the square the four bases A, T, G and C, and (3) starting at the center and choosing the corners for constructing a new spot in the order in which bases appear in a DNA. Although the graphical representations of DNA based on the "magic square" (a colloquial name that we introduced for the approach of Jeffrey) produce apparently "chaotic" patterns (see Figures 8−12), it is highly significant that there is no loss of information accompanying Jeffrey's graphical representation.

In order to illustrate the reconstruction, we will examine Figure 9 more closely. Let us first point out some of the properties of Jeffrey's construction. Observe that, when the interior of the magic square is divided into four quadrants, all spots in the first quadrant belong to base G, all spots in the second quadrant belong to base C, all spots in the third quadrant belong to base A, and all spots in the fourth quadrant belong to base T. If that would not be the case, then the point (nucleotide base) that preceded it would have to be outside the magic square, which is a contradiction. So if one knows which is the last point (corresponding to the last base), one would start with that point and move in the opposite direction from the *nearest* corner of the square at twice the distance of the spot from the corner to find its predecessor point. From this point, the reversed process of construction of the compact 2D representation of DNA continues, again at twice the distance of the just-found point from its nearest corner is to be found the point that is its predecessor, and the process of reconstruction continues until all spots have been identified. If the last spot is not known, one selects *any* point and initiates the outlined reconstruction by searching for the point at the double distance from the corner in the quadrant in which is the selected point. The reverse-construction process will be exhausted when the spot corresponding to the first base has been arrived at. One then continues with the reconstruction by selecting one of the spots not yet identified, which will lead to points already identified. By repeating this procedure, eventually one will identify the last base and have arrived at complete reconstruction.

We will later return to the construction of condensed DNA representations of Jeffrey; however, as will be seen, we will continue to confine our interest not only to relatively short DNA sequences that can be related to smaller proteins but also to "extremely" short DNA triplets, which depict codons, and will see how this leads to alternative graphical representations of proteins.

3.4. Spectrum-like Representation of DNA

The early DNA graphical representations were followed by graphical representations of DNA based on alternative geometrical motives. One such representation merits special attention despite its simple, and some may even say simplistic, appearance. This is a spectrum-like graphical DNA representation based on four horizontal lines, to which the four bases A, C, G, and T are assigned, introduced by two of the present authors.[117,118] To obtain a graphical representation of DNA, one moves through a DNA sequence and assigns to each base a spot at the corresponding horizontal line at uniform horizontal separations. The thus-distributed spots on four lines are connected successively, forming a spectrum-like representation, which is depicted in Figure 14 for the coding sequence of the first exon of human β-globin gene.

In order to obtain a numerical characterization of such spectrum-like representations, one considers the construction of a **D/D** matrix associated with the line segments of the "spectrum". At first sight, as already mentioned, such graphical representations of DNA may appear "primitive", but they have hidden sophistication, which has led unexpectedly to the use of such spectrum-like DNA representations in the problem of alignment of DNA sequences and, in general, comparative studies of DNA. Namely, by replacing

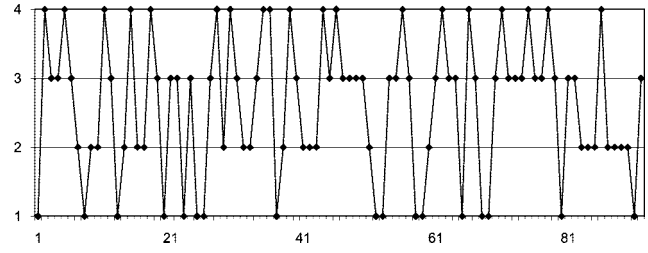


Figure 14. Spectrum-like graphical representation of the coding sequence of the first exon of human β-globin gene.

Table 6. (Top) Alignment of Two Test DNA Sequences Obtained by an Alignment Program;[48] (Middle and Bottom) Better Alignments Having One Additional Identity

Sequence	Alignment
seq1	GGCCTCTGCCTAATCACACAGAT-CTAACAGGATTATTTC
	||||||||||| || |||| || ||||||| |||||
seq2	GGCCTCTGCCTTATTACACAAATCTTAACAGGACTATTTC
seq1	GGCCTCTGCCTAATCACACAGATC-TAACAGGATTATTTC
	||||||||||| || |||| ||| ||||||| |||||
seq2	GGCCTCTGCCTTATTACACAAATCTTAACAGGACTATTTC
seq1	GGCCTCTGCCTAATCACACAGATCT-AACAGGATTATTTC
	||||||||||| || |||| |||| ||||||| |||||
seq2	GGCCTCTGCCTTATTACACAAATCTTAACAGGACTATTTC

the letters A, C, G, T, which signify labels for the four lines by the numbers 1, 2, 3, 4, one transforms an *alphabetic* sequence into a *numerical* sequence. The advantage of such a "trivial" substitution is that one can perform on numerical sequences arithmetic operations that can also be presented graphically and visualized. As briefly outlined by the present authors,[119] it takes just another step from here to arrive at graphical alignments of DNA, by considering additional difference plots for two sequences, obtained after the sequences have been shifted relatively to one another by one or more steps. Similarly from spectral-like graphical representations of proteins, one can construct an analogous graphical approach to the problem of protein alignment,[119,120] which we will review more closely in a later section of this article.

3.5. On Graphical Alignment of DNA

Chronologically, the recognition that one can numerically manipulate DNA sequences and in this way arrive at a "graphical alignment of DNA",[119] which we will shortly outline, came after such operations have been first recognized for proteins, even though the article on graphical alignment of proteins was published almost two years later.[120] In the case of proteins, instead of four horizontal lines, one has to consider 20 horizontal lines, to which 20 different numerical values ought to be assigned. Otherwise there is no essential distinction between the two approaches, one for DNA and the other for proteins. For proteins one also performs arithmetic operations on sequences and views the difference plots for locations in which the two sequences completely cancel one another, which is easily detected by considering points in the difference diagram that lie on the *x*-axis. To find the full alignment, the two sequences are shifted gradually relatively to one another.

The traditional approach to the problem of alignment of two DNA sequences, S_1 and S_2, requires one to consider insertion of gaps, either within sequences or at their ends, and then placing the thus-modified resulting sequences one over the other so that every letter in either sequence corresponds to the same letter or a gap in the other sequence. At the top of Table 6 in the top row, we show the output of an alignment program applied on two test DNA sequences consisting of 39 and 40 bases, respectively, displayed on the web page of University of Wisconsin, where the program and the output have been presented.[121] A close look at this alignment reveals that a better alignment is possible if one shifts the first gap in S_1 one or two places to the right, as shown in the middle and the bottom rows of the table, respectively, which have one additional alignment. Hence, in this case, this particular alignment program did not produce the best result.

In Figure 15, we show the 4-line graphs of the two DNA sequences S_1 and S_2 listed in Table 6. A close look at the graphical "signatures" ("spectra") of the two DNA sequences

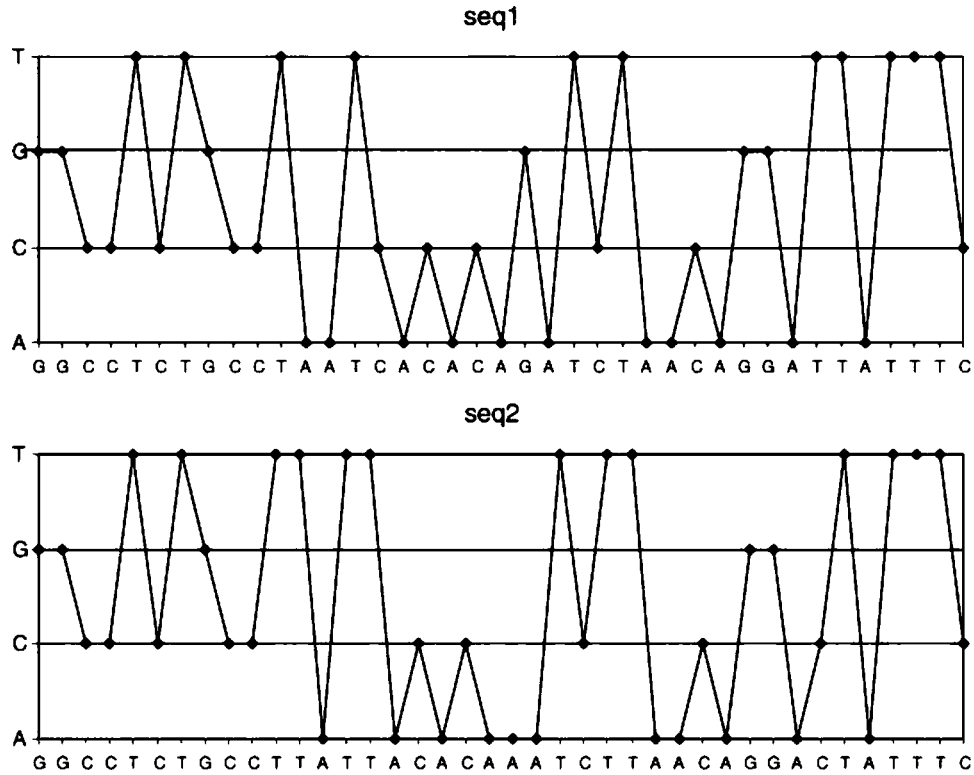


Figure 15. Graphical "signatures" of S_1 and S_2 consisting of 39 and 40 bases, respectively. Reproduced by kind permission from ref 119. Copyright 2006 Elsevier Publ.

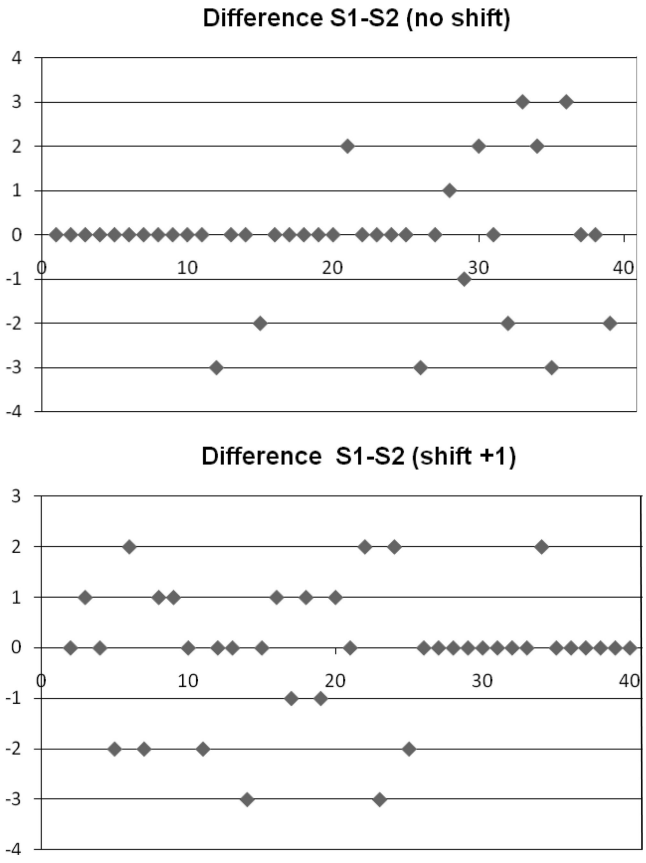


Figure 16. Differences between S_1 and S_2, and the difference between S_1 and S_2 when shifted by one place.

clearly reveals a considerable overall similarity between them. The considerable portion of the beginning, the central part, and the end part of the two spectra show the same aspects. On the other hand, the same spectrum-like graphical "signatures" also show the presence of local differences, particularly around sites 12−15, 25, 30, and 35. To arrive at graphical alignment of two DNA sequences, one first replaces the *alphabetic* sequences by the corresponding *numerical* sequence. Thus, for example, the sequence A T G G T G C A C C T G A C T C C T G A, which represents the beginning of the coding sequence of the first exon of human β-globin gene, is transformed into the integer sequence: 1 4 3 3 4 3 2 1 2 2 4 3 1 2 4 2 2 4 3 1. The values 1−4 represent the *y*-coordinates ("amplitudes") of the spectrum-like 4-line graphical representation of DNA, while the *x*-coordinates of the bases are their sequential index.

Let us now return to the two sequences of Table 6. In order to determine the similarities and the differences between the two sequences S_1 and S_2, one subtracts the two numerical sequences and one plots their difference. When the corresponding bases in two sequences are equal, they will have the same *y*-coordinate and the numerical difference between the two is zero. In Figure 16 (the top part), we show the plot of the difference ($S_1 - S_2$), and in the same figure at the bottom, we show the difference between two sequences when the second sequence is shifted one place to the right. From the left side of the top diagram showing the difference plot, one can see that the corresponding bases are equal, except in the positions 12, 15, and 21. Hence, only in three locations do the two sequences not match, which is also visible from Table 6 when one looks at the matching over the first 25 sites—there are missing three vertical lines, signifying matching. Beyond position 25, the two sequences differ considerably, being only occasionally coincidental (and possibly accidentally, the same corresponding bases that produce the zero difference).

The bottom part of Figure 16 is obtained when the two sequences were shifted by one place and then the difference plot was constructed. As one can see now, not surprisingly, the first part of S_1 and shifted S_2 lack any matching. That was to be expected, in view that the very good matching of this half of the DNA sequences has now been disturbed. But what is important is that the shift produced an almost complete matching in the second part of the sequences S_1 and S_2, which have hitherto shown no matching. Observe that now, beyond site 26, there is but a single place (at location 34) where the two sequences do not fully overlap. By combining the information from the top and bottom parts of Figure 16, one can immediately construct the alignment diagrams for the sequences S_1 and S_2 shown in the middle and at the bottom part of Table 6. However, already from Figure 16, one can observe that, in all, there will be four mismatches (three at the left side of site 25 of the two sequences and one at the right side of site 26) plus one gap (at either site 25 or 26). The alignment illustrated in the first row of Table 6 has five mismatches and one gap.

Here it is not so important that the outlined graphical alignment produced a *better* alignment than a particular computer program, because this particular computer program is likely to be modified and give better results, while other computer programs may have no similar fault. What is important is that the graphical alignment produced the *correct* answer, the same as would be obtained by sophisticated available computer programs, resulting in an answer that one can visualize. Clearly the top part of Figure 16 shows the very good local alignment of S_1 and S_2 from the position 1 up to 25 (22 identities and 3 mismatches), while from the bottom part of Figure 16 one can see that insertion of a gap into S_1 before position 26 creates the very good local alignment from position 26 up to 40 (14 identities and 1 mismatch). Hence, a gap has to be inserted either in the position 25 or 26 in order to obtain the best global alignment of S_1 and S_2. The graphical alignment outlined above, which is neither conceptually nor computationally involved, enables one to quickly find the two best global alignments of the sequences considered. All calculations (finding difference before and after shifting two sequences by one or more sites) can easily be performed using Microsoft Excel.

4. Graphical Representation of RNA Secondary Structure

4.1. Approximate Graphical Representation of RNA Secondary Structure

We will refer to all the publications dealing with the secondary RNA structure published prior to mid-2009, which use eight symbols to represent RNA as approximate graphical representations in view that all of them are accompanied with some loss of information. Of this we have become aware only very recently, with the work of Liao et al.[122] These authors have shown on a very short RNA having 17 bases that discrimination between bases A, C, G, U and their hydrogen-bonded counterparts A′, C′, G′, U′ is not sufficient for the reconstruction of RNA. Here is their illustration for RNA structures depicted in Figure 17:

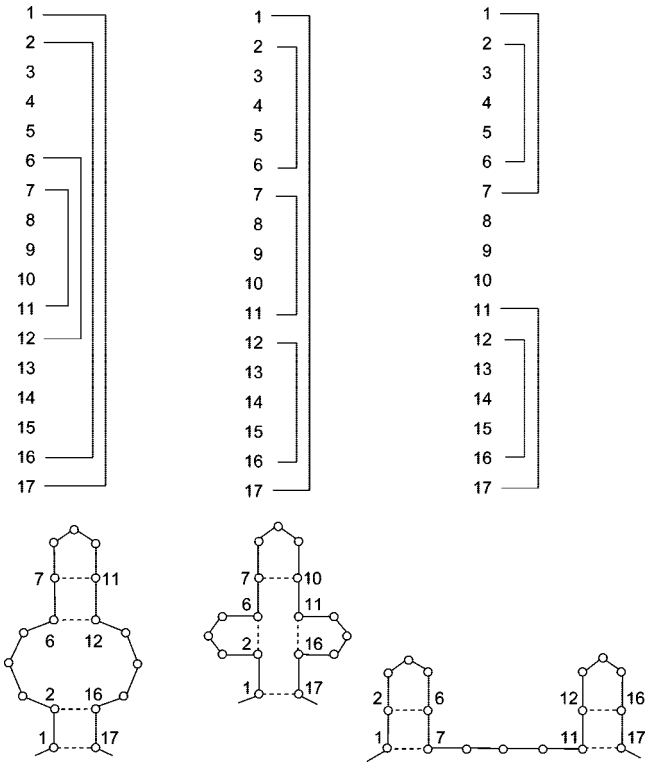


Figure 17. Three different secondary structures for the same primary structure.

G′ G′ A U U C′ C′ C A U G′ G′ C G A C′ C′
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

The same sequence allows at least two alternative hydrogen-bond pairings:

- Case (i) 1−17; 2−16; 6−12; 7−11;
- Case (ii) 1−17; 2−6; 7−11; 12−16,

illustrated in Figure 17. Hence, the two different secondary structures are represented by the same eight-symbol sequence. In fact these are not the only secondary RNA structures represented by the same string of eight symbols. One such additional case (illustrated at right in Figure 17) is

- Case (iii) 1−7; 2−6; 11−17; 12−16.

We will outline an approach in the next section that allows recovery of the lost information; it suffices now just to be aware of the limitations of the current approaches, which analyze the secondary structure of RNA using eight-symbol codes.

For characterization of the secondary structure of RNA, as mentioned above, one needs, besides the sequence of the four bases A, C, G, U (U standing for uracil, which replaces thymine present in DNA), information on hydrogen-bond pairing of A−U and C−G. All hitherto published graphical studies of the secondary RNA structure[123−132] were based on discriminating the coding of the unpaired bases A, C, G, U and paired bases A′, C′, G′, U′, which transforms a 2D secondary representation of RNA to sequences of eight symbols. To graphically represent such RNAs one can, for example, use eight horizontal lines and assign four horizontal lines to the unpaired bases A, C, G, U and four additional horizontal lines to paired bases A′, C′, G′, U′.[133] In this way, the sequence of 49 bases for viral RNA (TSV-3 of Figure 18) is transformed into the sequence

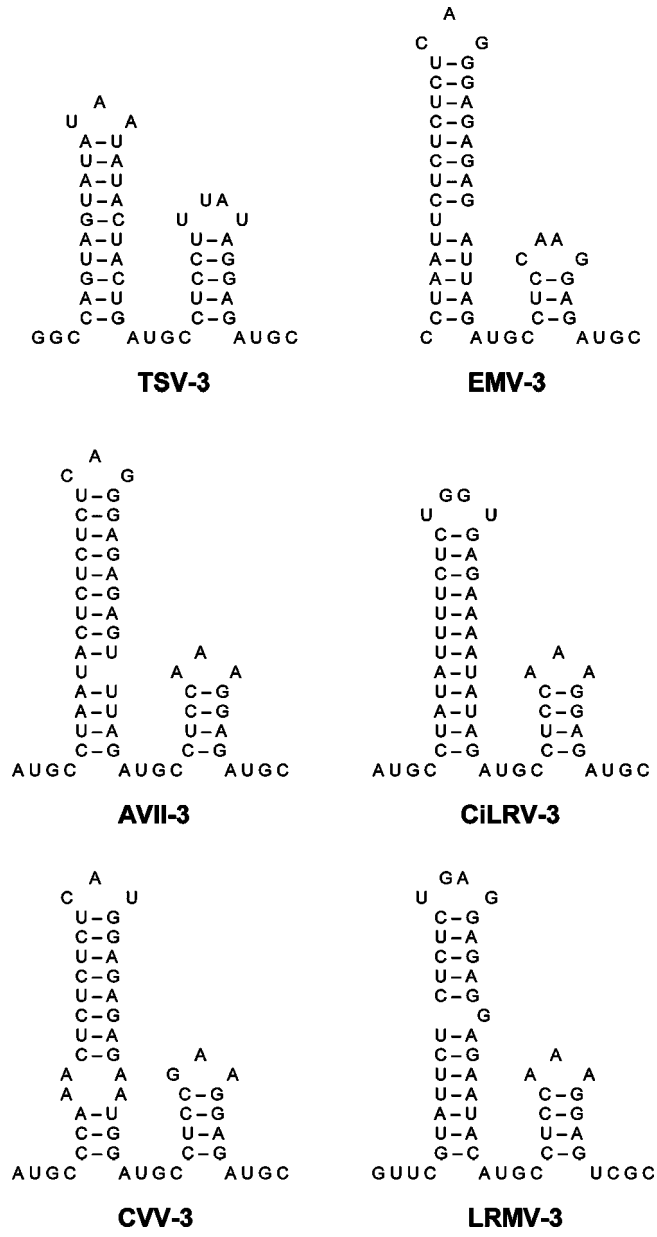


Figure 18. Six RNA sequences selected to illustrate spectrum-like graphical representation of RNA.

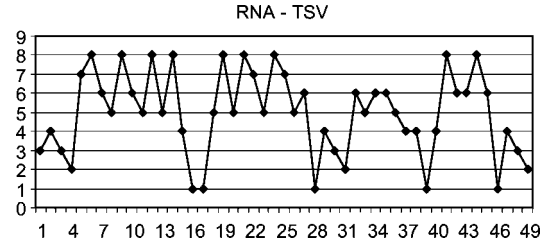


Figure 19. Spectrum-like graphical representation of the secondary structure of the TSV-3 RNA.

G U G C C′ A′ G′ U′ A′ G′ U′ A′ U′ A′ U A A U′ A′ U′
A′ C′ U′ A′ C′ U′ G′ A U G C C′ U′ C′ C′ U′ U U A U A′
G′ G′ A′ G′ A U G C

which is depicted in Figure 19 as a spectrum-like representation. The eight nucleotides have been assigned the following *y*-coordinates: A = 1; C = 2; G = 3; U = 4; U′ = 5; G′ = 6; C′ = 7; and A′ = 8, while the *x*-coordinate of each base takes the running values from 1 to *N* (*N* is the number of RNA bases). Observe that the horizontal lines from *y* = 1

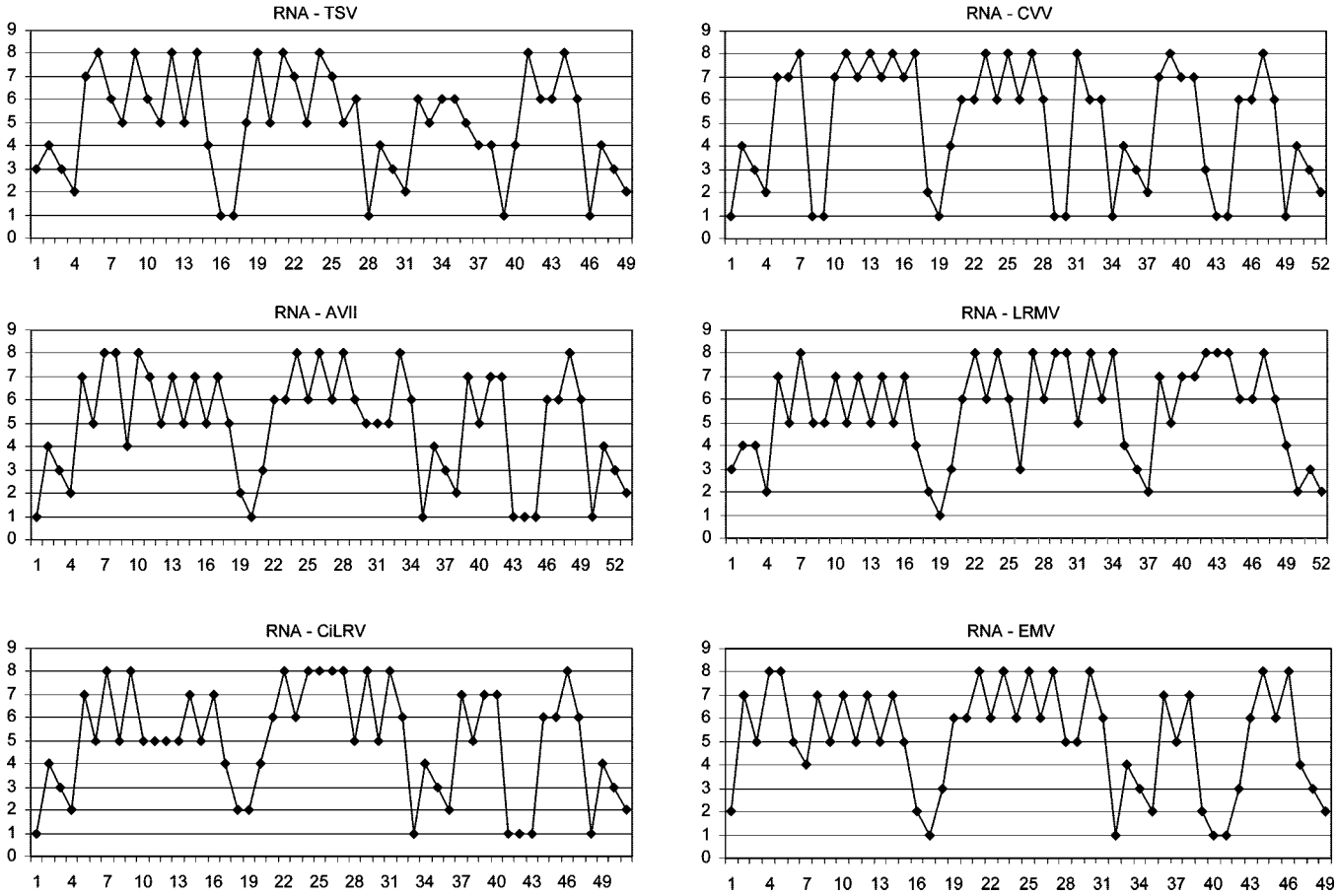


Figure 20. Spectrum-like graphical representation of the secondary structure of the six RNA sequences.

to $y = 4$ belong to unpaired bases, while the lines from $y = 5$ to $y = 8$ belong to paired bases. A close examination of Figure 19 reflects the complementarities of the pairing of the bases in RNA, which can be seen by reflection of corresponding spectral fragments for $y = 5$ to $y = 8$.

In Figure 20 are illustrated graphical representations for six of the nine RNA viruses studied by Reusken and Bol.[134] Reusken and Bol were interested in predicting RNA secondary structures and were investigating the bonding of coat protein to AUGC motifs, which are required to initiate infection by these viruses. The abbreviations below stand for structures at the 3′ terminus of RNA; references to the original articles for the six viruses have also been included: TSV = Tobacco streak virus;[135,136] EMV = Elm mottle virus[137] AVII = Asparagus virus II;[134] CiLRV = Citrus leaf rugose virus;[138,139] CVV = Citrus variegation virus;[138,139] and LRMV = Lilac ring mottle virus.[138,139]

A close look at Figure 20 is informative. By looking at the "shape" of the RNA spectra in Figure 20, one can easily notice considerable similarity for RNA of AVII and EMV viruses, and AVII with LRMV, possibly CiLRV with LRMV, and perhaps of RNA of TSV and LRMV viruses. The spectra of CVV and CiLRV appear more individualistic, except that they may show similarity with AVII, particularly in the region of 30−50 nucleotides. These qualitative estimates need to be followed by quantitative evaluations. By contrast, visual inspection of Figure 18 is more tedious and less instructive, because while it is easy to view the overall shapes of the secondary structures, the length of individual hairpin sections of RNA tells nothing about which bases are involved. If one wants to discern the degree of similarity between different RNA secondary structures, then one needs to compare the corresponding sites for agreement in respective nucleotides, which becomes quite tedious.

4.2. Numerical Characterization of RNA Secondary Structures

Numerical characterization of RNA sequences requires a search for RNA sequence invariants. A way to obtain them is to use geometrical properties of graphical representations of RNA for construction of various distance matrices, such as the distance−distance **D/D** matrix,[58] the Euclidean distance matrix,[140] the **S/S** matrix based on sequential labels of nucleotides,[141] and the **L/L** matrix.[117] In **D/D** and **L/L** matrices, the element (i, j) is defined as the quotient of the "through-space" distance (Euclidean distance) and the "through-bond" distance (the distance along the curve). The difference between **D/D** and **L/L** matrices is that, in the former, the through-bond distance is reduced to the count of bonds between vertices i and j, while in the latter, actual geometrical "through-bond" distances are used. The eigenvalues or the average row sums of **D/D** or **L/L** matrix can serve as sequence descriptors.

An alternative approach has been described by Feng and Wang,[126] who constructed graphical representations of RNA secondary structure based on chaos game. They first classified nucleotides as weak and strong hydrogen-bond pairs (A−U; G−C); as having amino and keto (A, C vs G, U) groups; and as belonging to purine and pyrimidine (A, G vs C, U) types, and then they used the leading eigenvalue of **L/L** matrices for the characterization of RNA. They found the following pairs as the most similar:

AVII−LRMV; EMV−LRMV; AVII−CVV; CVV−EMV; CVV−LRMV; and AVII−EMV. Because of false-positive results (the possibility that two secondary structures of RNA are found as similar, when they are not), it is of interest to compare results based on *different* characterizations of RNA sequences. Namely, *small* numerical entries in the similarity/dissimilarity tables are *necessary* but not *sufficient* conditions for two sequences to be similar. In contrast, *large* numerical entries are *sufficient* to point out that two of RNA are not similar. Clearly models that could reduce false-positive results are desirable.

4.3. RNA Nucleotide Adjacency-Count Matrix

For each RNA, one can construct an 8 × 8 matrix, the rows and columns of which correspond to the eight nucleotides: A, C, G, U, U′, G′, C′, A′. The (*i*, *j*) entry of such matrices counts the occurrence of nucleotide X followed by nucleotide Y when the RNA sequence is read from left to right. In Table 7, we illustrate the 8 × 8 adjacency-count matrix for the TSV-3 virus. Observe that adjacency-count matrices are nonsymmetrical and that their transposed matrices correspond to reading the RNA sequence from right to left. The 64 matrix entries of RNA nucleotide adjacency matrix can be viewed as 64 components of a vector that describes the RNA secondary structure. Two RNA sequences that are similar will have similar entries in the corresponding cells of the 8 × 8 matrices, but at the same time, as already mentioned, the converse need not be true: two matrices that have similar matrix entries may be quite different. This is because the nucleotide adjacency matrix records only immediate neighborhoods of bases, which may give the same count of adjacencies, but nucleotides may be distributed quite differently in different RNA sequences. Adjacency-count matrices have been introduced only recently for characterization of proteins.[142,143] An entry (*i*, *j*) in such a matrix counts how many times amino acid *i* is followed by amino acid *j* when the protein sequence is read from left to right, which results in a 20 × 20 matrix for any protein, regardless of its size.

From the 8 × 8 RNA matrices, one can construct several 8-component vectors, such as the vector whose entries represent the abundances of the eight bases; the repetition of the bases; and the difference in the number of adjacencies when the RNA sequence is read from left to right and from right to left. Thus, for TSV-3 (shown in Table 7), one obtains the following:

- Abundance vector: (5, 3, 4, 7, 9, 6, 6, 9)
- Repeating bases vector: (1, 0, 0, 1, 0, 1, 1, 0)
- Left−right difference: (4, 5, 5, 4, 10, 8, 7, 9).

The abundance vector can be constructed from the row sums corrected for the last base (or from the column sums corrected for the first base) of the 8 × 8 RNA matrix. The vector based on the count of repeating nucleotide bases is given by diagonal entries of 8 × 8 adjacency-count matrix. The left−right difference vector indicates the asymmetry in adjacency of nucleotides of DNA sequence when the sequence is read from left to right or vice versa. There is some loss of information that accompanies the reduction of a 64-component matrix to vectors having eight components, which one compensates for, in part, by considering several such vectors. In Table 8, we collected information on 8-component vectors for the six viral RNAs considered.

Table 7. RNA 8 × 8 Adjacency Matrix for TSV-3 Virus

	A	C	G	U	U′	G′	C′	A′
A	1	0	0	3	1	0	0	0
C	0	0	0	0	0	0	2	0
G	0	3	0	1	0	0	0	0
U	2	0	3	1	0	0	0	1
U′	0	0	0	1	0	1	1	6
G′	2	0	0	0	2	1	0	1
C′	0	0	0	0	4	0	1	1
A′	0	0	0	1	2	4	2	0

Table 9. Similarity/Dissimilarity among the Six RNAs of Table 8 Based on 64-Component Vectors (Above the Main Diagonal) and Based on 8-Component Abundance Vectors (Below the Main Diagonal)

	TSV	AVII	CiLRV	CVV	LRMV	EMV
TSV		36	36	41	42	42
AVII	12		30	27	26	24
CiLRV	7	12		39	36	38
CVV	23	11	23		43	39
LRMV	10	8	10	16		34
EMV	12	6	13	16	9	

4.4. Similarity/Dissimilarity among RNA

One can evaluate the similarity/dissimilarity between the six RNAs of Figure 18 using (i) the complete 8 × 8 matrices that record the adjacency of nucleotides in RNA, allowing the construction of 64-component vectors, obtained by listing all matrix elements (illustrated in Table 7 for one of the six viral RNAs) as a single sequence; and (ii) the three 8-component vectors listing the abundance, the repeated bases, and the left−right difference, as indicated in Table 8. In Table 9 above the main diagonal is shown the resulting similarity/dissimilarity table between the six RNA viruses based on 64-components of the 8 × 8 matrices. The higher degrees of similarities correspond to small entries, while the higher degrees of dissimilarities correspond to large entries. The three smallest entries above the main diagonal suggest as the most similar the following pairs of RNA:

- AVII−EMV
- AVII−LRMV
- AVII−CVV

Observe that only two of the three pairs listed above are among the five pairs listed earlier as apparently similar based on visual inspection of the spectrum-like representation of RNA. This indicates limitations of qualitative visual com-

Table 8. Abundance, Repeated Bases, And the Difference in Reading RNA is Read from Left to Right for Six RNA of Figure 18

	TSV	AVII	CiLRV
abundance	5, 3, 4, 7, 9, 6, 6, 9	7, 4, 4, 4, 9, 9, 8, 8	6, 3, 5, 5, 10, 6, 6, 10
repeated bases	1, 0, 0, 1, 0, 1, 1, 0	2, 0, 0, 0, 2, 2, 1, 1	2, 0, 1, 0, 3, 1, 1, 3
left−right	4, 5, 5, 4, 10, 8, 7, 9	7, 7, 8, 6, 6, 4, 6, 4	5, 5, 6, 8, 0, 2, 4, 0
	CVV	LRMV	EMV
abundance	10, 4, 4, 4, 6, 11, 9, 5	5, 4, 5, 6, 8, 7, 9, 8	5, 4, 4, 3, 9, 9, 7, 8
repeated bases	3, 0, 0, 0, 0, 3, 2, 0	2, 0, 0, 1, 1, 1, 1, 1	0, 0, 0, 0, 1, 1, 0, 1
left−right	9, 7, 8, 8, 4, 4, 5, 2	4, 5, 4, 6, 0, 2, 6, 2	9, 6, 7, 6, 5, 5, 6, 0

parisons of relatively complex biological structures. The three least-similar RNAs found are as follows:

- CVV−LRMV
- TSV−LRMV
- TSV−EMV

Observe that numerical characterization of the similarity between TSV−LRMV and Ci−LRMV, both of which appeared visually estimated as similar, does not suggest these to be similar.

The entries below the main diagonal in Table 9 indicate the degree of similarity among the six RNAs using the 8-component vectors based on the abundance of the eight bases. Clearly these reduced representations may lead to identification of similar pairs of RNA.

There is no guarantee that pairs identified as similar using the 64-component vectors are necessarily similar. Hence, similarity/dissimilarity tables should be viewed only as pointing to possible or potential similarity. In order to identify and remove "false" similarity, it is important that one uses two or more alternative representations of biological sequences, such as characterization of RNA based on nucleotide adjacency matrices, chaos game representation,[84] element-contact graphs,[144] classification of nucleotide bases,[145] 2D graphical representation,[132] 3D graphical representation,[146] and 4-dimensional representation of DNA/RNA.[147]

Another aspect of similarity/dissimilarity that may have been often overlooked and which is now more apparent is the fact that, at best, one can only relate to the most similar structures, be it RNA, DNA, or proteins. Relationships for more distant structures are likely to be speculative, because of lack of consensus among different approaches. Thus, presentation of results of the similarity/dissimilarity in a dendrogram form may be reliable only for closely related species and may be premature concerning less-similar species, unless different approaches to the same classification show convergence and until newer RNA, DNA, and protein descriptors emerge that are less subjective (depending on specific geometrical aspects of sequence representations) and more objective (depending on inherent sequence invariants). Meanwhile, for establishing the degree of similarity of biosequences, it is essential to consider consensus between similarity/dissimilarity tables from different sources, whose combined use will allow one to filter out "false" similarities.

5. Graphical Representation of RNA Secondary Structure without Loss of Information

The problem with eight-symbol representations of the secondary structure of RNA is, as a close look at Figure 18 may reveal, that there is no discrimination for hydrogen bonds A−U and U−A, and similar hydrogen bonds C−G and G−C. However, as Randić and Plavšić have shown,[148] if one discriminates between the first and the second base of a hydrogen bond X−Y, which amounts to differentiating between X−Y and Y−X base pairs of RNA involved in hydrogen bonding, it will be possible to discriminate between the three RNA secondary structures of Figure 17. In other words, the direction in which one reads RNA structure is important and symbols X_1-Y_2 and Y_1-X_2 can stand for the two possibilities, where X_1 and Y_1 designate bases in the primary sequence that are closer to the 5′ end of RNA and X_2 and Y_2 designate bases in the primary sequence that are closer to the 3′ end of RNA sequence. Hence, each base can be in one of three different situations: free (X) or hydrogen bonded as the first (X′) or the second base (X″) when RNA is read from left to right. Thus, we arrive at a need for 12-symbol representation of the secondary structure of RNA. With a convention of using symbols X for free and X′ and X″ for hydrogen-bonded bases, the 17-base RNA secondary structures of Figure 17 become the following:

Case (i): G′G′A U U C′C′C A U G″G″C G A C″C″
Case (ii): G′G′A U U C″C′C A U G″G′C G A C″C″
Case (iii): G′G′A U U C″C″C A U G′G′C G A C″C″

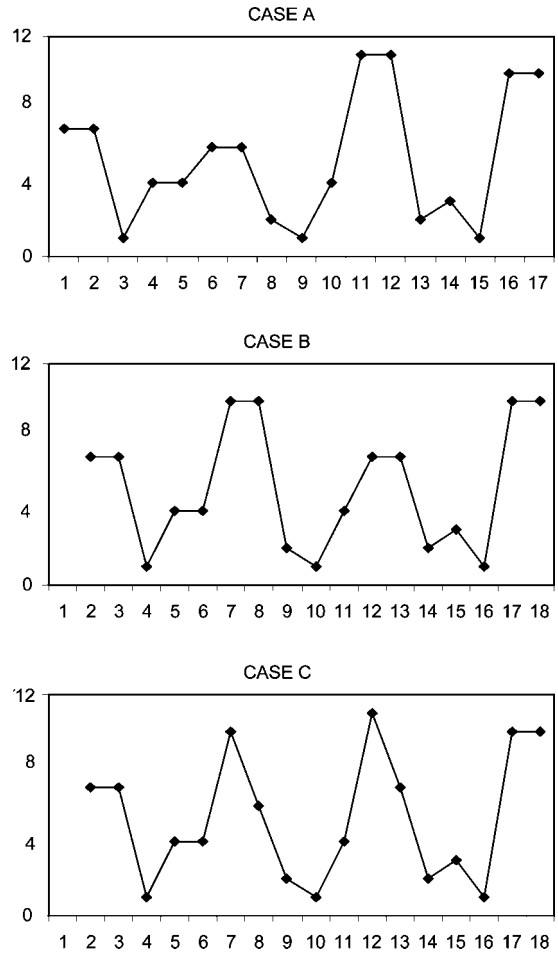


Figure 21. Spectral representation of hypothetical RNA with three possible different secondary structures.

In Figure 21 are depicted spectral graphical representations for the three distinct secondary structures based on 12 horizontal lines using the following assignment of *y*-coordinates:

$$\begin{array}{lll} A = 1 & A' = 5 & A'' = 9 \\ C = 2 & C' = 6 & C'' = 10 \\ G = 3 & G' = 7 & G'' = 11 \\ U = 4 & U' = 8 & U'' = 12 \end{array}$$

In Table 10, we show the corresponding 12 × 12 adjacency-count matrix for the first of the three secondary RNA structures of Figure 18. One of the first purposes for constructing graphical representations of biosequences is that they allow the construction of accompanying matrices using geometrical information for calculating their entries, which then lead to constructing a set of matrix invariants as sequence descriptors. In the case of adjacency-counting matrices, be it for RNA or protein sequences, it is not difficult to extract the following three vectors: the 12-component abundance vector (A), the 12-component diagonal entries

Table 10. RNA 12 × 12 Adjacency Matrix for TSV-3 Virus

	A	C	G	U	A′	C′	G′	U′	A″	C″	G″	U″
A	1	0	0	3	0	0	0	0	0	0	0	0
C	0	0	0	0	0	1	1	0	0	0	0	0
G	0	2	2	0	0	0	0	0	0	0	0	0
U	2	0	2	1	0	0	0	0	1	0	0	0
A′	0	0	0	1	0	0	2	1	0	0	0	0
C′	0	0	0	0	1	1	0	2	0	0	0	0
G′	0	0	0	0	0	0	0	2	0	0	0	0
U′	0	0	0	1	3	1	0	0	0	0	0	0
A″	0	0	0	0	0	0	0	0	0	2	2	1
C″	0	0	0	0	0	0	0	0	0	0	0	2
G″	2	0	0	0	0	0	0	0	1	0	1	0
U″	0	0	0	0	0	0	0	0	3	0	1	0

vector (B) that gives information on repeating bases, and the 12-component sequence asymmetry vector (C) which gives information on adjacency count when the sequence is read from left to right and from right to the left. For the three secondary structures of Figure 18, one obtains:

12-component abundance vector:
A(i): (3, 2, 1, 3, 0, 2, 2, 0, 0, 2, 2, 0)
A(ii): (3, 2, 1, 3, 0, 0, 4, 0, 0, 4, 0, 0)
A(iii): (3, 2, 1, 3, 0, 1, 3, 0, 0, 3, 1, 0)
12-component diagonal entries vector:
B(i): (0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0)
B(ii): (0, 0, 0, 1, 0, 0, 2, 0, 0, 2, 0, 0)
B(iii): (0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0)
12-component sequence asymmetry vector:
C(i): (6, 4, 1, 4, 0, 2, 2, 0, 0, 1, 2, 0)
C(ii): (6, 4, 2, 4, 0, 0, 3, 0, 0, 34, 0, 0)
C(iii): (6, 4, 2, 4, 0, 1, 4, 0, 0, 3, 2, 0)

The abundance vector can be constructed from row sums and column sums when the entries in them are different (because of not counting the first or the last entry in the sequence) by taking the larger value.

5.1. An Illustration

We will illustrate the novel 12-symbol representation of the secondary structure of RNAs by considering the RNA of APMV-3 apple mosaic virus and PDV-3 prune dwarf ilavirus illustrated in Figure 22.

The secondary structures of these two viruses correspond to the following two 12-symbol sequences, respectively,

AUGCC′C′A′C′A′A′C′GUGAAG″U″U″G″U″G″G″AU GCC′C′C′GUUAG″G″G″AAGC

AUGCC′C′U′C′A′C′C′GUGAAG″G″U″G″A″G″G″AUGCC′C′C′U′ UAAA″G″G″G″AUGC

which translates into the following numerical sequences:

1, 4, 3, 2, 6, 6, 5, 6, 5, 5, 6, 3, 4, 3, 1, 1, 11, 12, 12, 11, 12, 11, 11, 1, 4, 3, 2, 6, 6, 6, 3, 4, 4, 1, 11, 11, 11, 1, 1, 3, 2

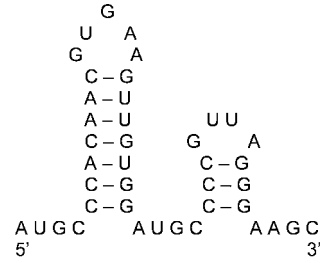
and

1, 4, 3, 2, 6, 6, 8, 6, 5, 6, 6, 3, 4, 3, 1, 1, 11, 11, 12, 11, 9, 11, 11, 1, 4, 3, 2, 6, 6, 6, 8, 4, 1, 1, 9, 11, 11, 11, 1, 4, 3, 2

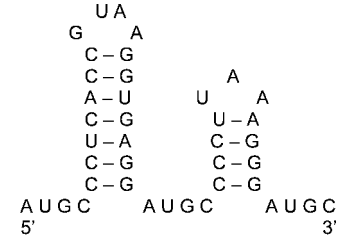
In Figure 23, we have illustrated the above representations of APMV-3 and PDV-3 in a form of spectral plots, where at equal intervals along the *x*-axis are plotted the *z*-coordinate values. The two horizontal lines across Figure 23 at $y = 4$ and $y = 8$ separate ACGU from A′C′G′U′ and A′C′G′U′ from A″C″G″U″. Observe also complementarities of spectral fragments belonging to C′C′A′C′A′A′C′ and G″U″U″G″ U″G″G‴, that is, a fragment $x = 5$ to $x = 11$ in the range 4−8 is *inverted* and *reverse* of the corresponding fragment $x = 17$ to $x = 23$ in the range 8−12. Whenever this happens, one is assured that one deals with the regular hydrogen-bond base pairing A−U and C−G.

5.2. Alignment of RNA

An important advantage of spectral representations of biosequences, whether involving DNA, RNA, or proteins, is that such representations can be arithmetically manipulated, thus facilitating their alignment. Neither from the alphabetic codes of RNA nor from their spectral representations (such as Figure 23) is it clear to what degree two such sequences overlap and what the degree of similarity is between them. By contrast, from the top of Figure 24, which shows the arithmetic difference between the spectral representations of APMV-3 and PDV-3, one can immediately see that the first three-quarters of the sequences show considerable overlap, which gives as difference the $y = 0$ value visible as points along the *x*-axis. When the two sequences are shifted by one place, an additional overlap between the two RNA structures can be detected. Counting zeros along the *x*-axis, one finds 26 coincidences in the interval 1−31 and an additional 6 coincidences in the interval 35−41, a total of 32 matches out of 41 bases. By counting all mismatches when one combines the information of both parts of Figure 24, one



APMV-3



PDV-3

Figure 22. Secondary structure of RNA of apple mosaic virus AMPV-3 and prune dwarf ilavirus PDV-3. Reproduced by kind permission from ref 148. Copyright 2009 Elsevier Publ.

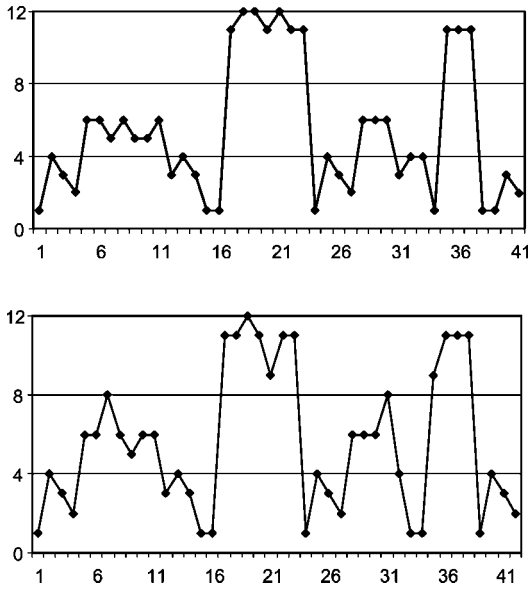


Figure 23. Spectral representation of APMV-3 (top) and PDV-3 (bottom) secondary RNA structures. The horizontal lines separate ACGU, A′C′G′U′, and A″C″G″U″, respectively. Reproduced in part by kind permission from ref 148. Copyright 2009 Elsevier Publ. The lower part of the figure is missing in the original paper.

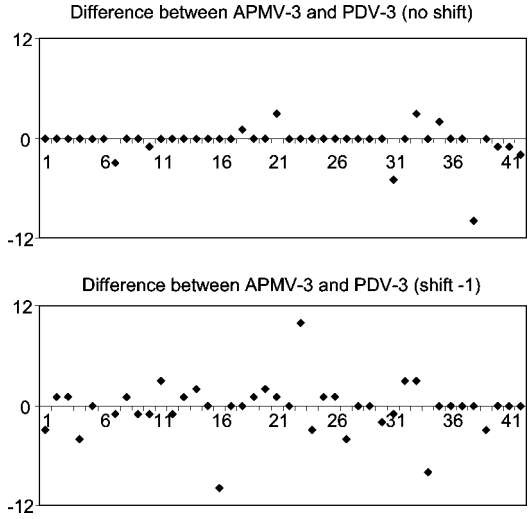


Figure 24. RNA alignment. The top diagram shows direct subtraction of the two numerically encoded RNAs, while the bottom diagram shows subtraction when the second RNA has been shifted by one site to the right.

finds 7 mismatches, which gives 14 displacements. Liao et al.[122] reported on 15 gaps, but it is not difficult to see that their results can be slightly improved in the segment 12−15.

6. Graphical Representation of Proteins

The emergence of graphical representations of proteins, in comparison with the emergence of graphical representations of DNA, was delayed for a considerable time because of the combinatorial complexity and the combinatorial explosion accompanying 20 factorial possible assignments for the 20 natural amino acids when considering alternative geometrical sites to be used for constructing graphical representations of proteins. Constructions involving a four-letter alphabet bring at most four factorial (24 alternatives) that one may need to consider, which is not excessive, but 20! is a prohibitive number, which demands exploration of new directions and new considerations.

One may conveniently classify graphical and nongraphical approaches for representation of proteins into the following types:

(1) Graphical representations based on modifications of existing graphical representations of DNA;

(2) Graphical representations based on (arbitrary) selection of one among 20! orderings of amino acids;

(3) Graphical and nongraphical representations based on constructions that bypass the critical step of selecting one of 20! possible assignments of the 20 amino acids among geometrically equivalent elements of underlying mathematical objects.

The listing of the above classes also represents the chronological development of graphical representations of proteins that we will review here.

6.1. Codon-Based Representations

When triplet codons that code for individual amino acids of a protein are known, it is straightforward to arrive at graphical representations of proteins by simply applying the available graphical representation of DNA directly to proteins by grouping the amino acid into triplets of bases that form codons, and then, with their graphical representation, one can associate the corresponding amino acid. This has been illustrated in ref 88 by adopting Jeffrey's "magic square" and combining three spots representing individual codons into a triangle, the center of which determines the location of a "spot" that represents the corresponding amino acid. If the three spots for the three bases have the coordinates (x_1, y_1), (x_2, y_2), (x_3, y_3), the accompanying amino acid will be at location $(x_1 + x_2 + x_3)/3$, $(y_1 + y_2 + y_3)/3$. It may happen that the triangle (x_1, y_1), (x_2, y_2), (x_3, y_3) collapses into a straight-line segment, but that does not change the algorithm.

One could consider, instead of the center of a triangle as the location of an amino acid associated with a codon, all three bases as the locations representing a single amino acid. In this case, the graphical construction of proteins (of known DNA history) is expanded to three times the number of spots than corresponding to the number of amino acids. An alternative would be to consider instead of the center of a triangle for the location of amino acids associated with a codon one of the three nucleotides (the first, the middle, or the last). In this case, the graphical construction of proteins reduces to considering every third spot of the graphical representation of corresponding DNA. In either case, graphical constructions and the accompanying computations are simple (and easily performed using Microsoft Excel).

6.2. Virtual Genetic Code

The simple procedure outlined above solves the problem of graphical representation of proteins when their underlying RNA sequence is known. Unfortunately, for most proteins of interest, the list of codons responsible for the synthesis is not known, which makes impossible the straightforward graphical construction of proteins as an extension of graphical constructions of DNA. A way out of this quagmire is the notion of the "Virtual Genetic Code",[47] which ascribes (in a somewhat arbitrary manner, one must add) to each amino acid a *single* codon uniquely. The virtual genetic code then allows one to "transcribe" any protein sequence into a corresponding *virtual* DNA sequence, which then allows one to adopt any of the available 2D or 3D graphical representations for depicting proteins of interest. The form of this 2D

Table 11. Virtual Genetic Code (in DNA T replaces U)

position in Figure 34	virtual code	amino acid	3-letter code	1-letter code	binary code
0	GCG	alanine	Ala	A	011001
1	CGG	arginine	Arg	R	100101
2	AAC	aspargine	Asn	N	000010
3	GAU	aspartic acid	Asp	D	010011
4	UGC	cysteine	Cys	C	110110
5	CAG	glutamine	Gln	Q	100011
6	GAA	glutamic acid	Glu	E	100001
7	GGC	glycine	Gly	G	010000
8	CAU	histidine	His	H	010110
9	AUC	isoleucine	Ile	I	001110
10	CUC	leucine	Leu	L	101110
11	AAG	lysine	Lys	K	000001
12	AUG	methionine	Met	M	001101
13	UUC	phenylalanine	Phe	F	111110
14	CCA	proline	Pro	P	101000
15	UCU	serine	Ser	S	111011
16	ACA	threonine	Thr	T	001000
17	UGG	tryptophan	Trp	W	110101
18	UAU	tyrosine	Tyr	Y	110011
19	GUU	valine	Val	V	011111

or 3D graphical representation for proteins will be influenced not only by artifacts associated with the particular graphical representation of DNA but also by artifacts accompanying the adoption of the particular virtual genetic code. Therefore, this may appear to some people as "fog on fog" modeling, the phrase used by Hermann Weyl with reference to the mathematician Georg Cantor, the originator of today's very respected branch of mathematics, the set theory. Cantor was criticized by the leading mathematical authority of his time, Leopold Kronecker, Professor in Berlin, for his notion of classification of "infinities". However, just like Cantor's infinities, so "tons" of arbitrariness do not devalue such models, which remain legitimate and potentially useful—because in many applications one is interested in *differences* between proteins, and they will be less affected by details of underlying mathematical models.

In Table 11, we illustrate the virtual genetic code proposed in ref 47. The task was to select 20 codons out of 61 available triplets (excluding the three "stop" codons from the 4^3 triplets) in such a way that each triplet of bases represents one of the possible triplets for cases when an amino acid can be encoded in more than one way. This excludes methionine (M) and tryptophan (W), which are the only natural amino acids encoded by single triplets: ATG and TGG, respectively. In the construction of the virtual genetic code, triplet codons were selected so that the four bases, A, C, G and T, appear with the same frequency, each 15 times. In Figure 25, we show a spectral representation of human ND6 proteins of Table 12 obtained by replacing each amino acid by its "virtual codon". Thus, the 174 amino acids are represented as $3 \times 174 = 522$ nucleotides A, C, G, T, the sequence of which is illustrated on four horizontal lines to which the nucleotides A, C, G, T have been assigned.

The pattern displayed in Figure 25, which may be viewed as a kind of protein "fingerprinting", offers a not very economical representation of proteins and not necessarily user-friendly representation, in view of giving a very detailed figure, which will be of limited use in direct visual comparison of different proteins. Nevertheless, the representation is unique and allows

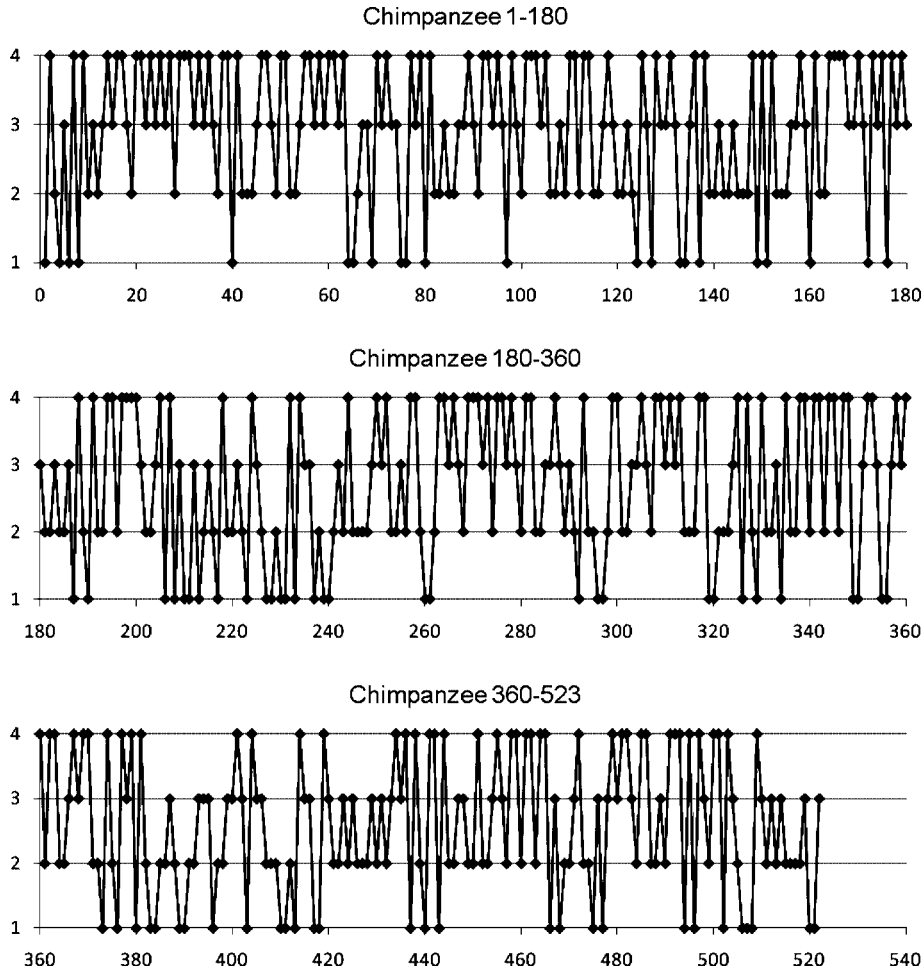


Figure 25. Spectral representation of chimpanzee ND6 proteins transformed via virtual genetic code into hypothetical RNA, depicted on four horizontal lines.

Table 12. ND6 Proteins for Human and Gorilla Having 174 Amino Acids (AA), Which Are Grouped in Tens for Easier Reading

		Human 174 AAs		
MMYALFLLSV	GLVMGFVGFS	SKPSPIYGGL	VLIVSGVVGC	VIILNFGGGY
MGLMVFLIYL	GGMMVVFGYT	TAMAIEEYPE	AWGSGVEVLV	SVLVGLAMEV
GFVLWVKEYD	GVVVVVNFNS	VGSWMIYEGE	GSGFIREDPI	GAGALYDYGR
WLVVVTGWPL	FVGVYIVIEI	ARGN		
		Gorilla 174 AAs		
MTYVLFLLSV	GLVMGFVGFS	SKPSPIYGGL	VLIVSGVVGC	AIILNCGGGY
MGLMVFLIYL	GGMMVVFGYT	TAMAIGEYPE	AWGSGVEVLV	SVLVGLAMEV
GLVLWVKEYD	GVVVVVNFNN	VGSWMIYEGE	GSGLIREDPI	GAGALYDYGR
WLVVVTGWTL	FVGVYIVIEI	ARGN		

the full reconstruction (i.e., identification) of the protein that it represents. However, as we will see soon, spectral graphical representations (representing either DNA/RNA or proteins) allow one to combine them arithmetically and thus visually inspect their similarities and even approach the problem of DNA and protein alignment graphically.

6.3. Justification of the Virtual Genetic Code

The Virtual Genetic Code is a mathematical construction that allows one to arrive at graphical representations of proteins when the underlying DNA/RNA is not known. Not long ago, an anonymous referee expressed reservations about the notion of virtual genetic code, stating that "*The idea of Virtual Genetic Code lacks any biological meaning. It is not good to transform a complex 64−20 mapping into a 1−1 mapping, which disturbs the authentic coding from DNA to protein*." This message is confusing *mathematical tools* with *biological reality*. Nobody is claiming that a 64−20 mapping is not factual; all that the virtual genetic code accomplishes is that, when the 64−20 mapping is *not known*, the hypothetical 1−1 mapping offers graphical representations of proteins, which may lead to useful results. In order to prove this, we will show that a representation of a protein with known RNA and a representation of a protein based on virtual genetic code produce the same result.

We will prove the above statement by considering two ND6 proteins of Table 12. The virtual genetic code transforms the 174 amino acids of human ND6 proteins into a hypothetical RNA of length 522, the beginning of which is as follows:

AUG AUG UAU GCG CUC UUC CUC CUC UCU GUU GGC CUC GUU AUG GGC UUC GUU GGC UUC UCU UCU AAG CCA UCU CCA AUC UAU GGC GGC CUC GUU...

For illustration we show chimpanzee ND6 protein transformed via virtual genetic code.

In order to compare the ND6 proteins of human and gorilla, both of which have 174 amino acids and 522 virtual nucleotides, one combines their spectrum-like representations arithmetically and constructs the differences between them. In Figure 26 (top), we have illustrated the "difference graph", which results in peaks in the y interval $(-3, +3)$. Despite the lengthy and visually "complicated" spectral representations of proteins (illustrated for human ND6 in Figure 25), one can immediately conclude from Figure 26 (top) that the human and gorilla proteins are very similar. As one sees, there are only nine amino acids out of 174 in which the two proteins differ. They are located at the sites: 2, 4, 41, 46, 76, 102, 120, 134, and 159. We should add that the heights of the "peaks" in Figures 25 and 26 are not significant, because they depend on the assignment of the horizontal lines to the four nucleotides, which is arbitrary. Because the spectral amplitudes have no useful meaning, one could transform Figure 26 (top) into the plot illustrated in Figure 26 (bottom) in which nonzero amplitudes are all set to be equal to $+1$.

In Figure 27, we display the graphical representation of the same hypothetical RNA of chimpanzee ND6 protein using a spiral as template. The construction starts at the center with each vertical segment of the spiral representing the binary code for the four bases, which we selected as follows:[150]

A = 0 0, G = 0 1, C = 1 0, U(T) = 1 1.

An apparent advantage of the 2D protein map representation of Figure 27 over the corresponding spectral representations is its compactness, while a disadvantage is the loss of visual qualities. This, however, is true of all graphical representations based on the chaos game model. Hence, such graphical representations have to be processed, as has been outlined in the literature by Zupan and Randić,[86] transforming a map into a spectral representation. The illustrated ND6 protein in Figure 27, when combined with a similar 2D representation of ND6 protein of gorilla, leads to the difference graph of Figure 28 with $2 \times 522 = 1044$ entries, in view that each nucleotide is represented by two binary digits. Binary codes have been previously used for representation of DNA sequences.[147−149]

There is considerable diversity between the graphical representations of proteins based on 20 horizontal lines (by direct use of 20 amino acids), based on 2D chaos game maps, based on 8×8 Table of Codons, based on four-color maps[150,151] and on 2D spiral representation.[152] They generally complement each other by involving different structural motifs, which helps in identifying artifacts of such graphical representations. The binary code transforms a DNA or RNA sequence of length N into a binary sequence of length $2N$ and a protein sequence of length N into a binary sequence of length $6N$, because each codon is represented as a six-digit binary label (see the last column of Table 11).

Now we will consider spectral representations of human and gorilla ND6 proteins directly by using their known sequence of 174 amino acids. We have ordered the 20 amino acids alphabetically

Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val

and assigned them values 1−20. This ordering of amino acids, and the form of the resulting graphical representations, is as arbitrary as any other, including random orderings, except for graphical representation of proteins that are based on use of physicochemical properties of amino acids.[153]

In Figure 29 (first two plots), we show spectral representations of human and gorilla ND6 based on 20 horizontal lines, with the top line corresponding to valine. Already a glimpse at the two spectra shows that ND6 proteins are rich in valine ($y = 20$) and glycine ($y = 8$) while, for instance, poor in cysteine ($y = 5$). At the bottom in Figure 29, we show the

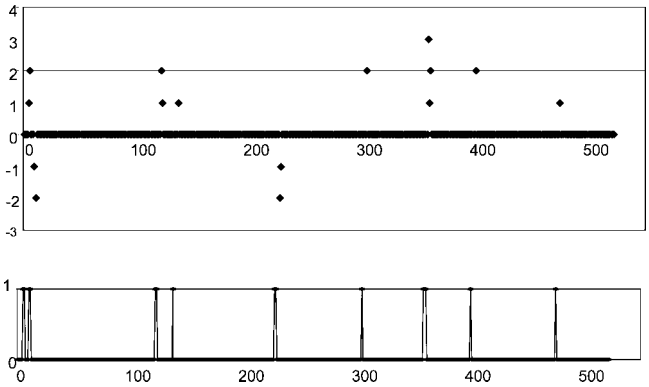


Figure 26. (Top) Difference plot between the spectral representations of human−gorilla ND6 proteins based on the virtual genetic codes. (Bottom) Corresponding graph depicting the binary difference for the same diagram.

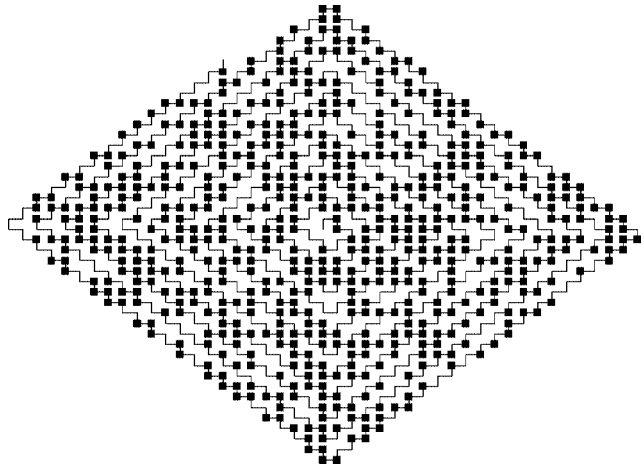


Figure 27. Graphical representation of chimpanzee ND6 proteins transformed via virtual genetic code into hypothetical RNA, which is shown on a "worm" spiral using binary codes to depict the four bases.

plot of the absolute difference between the numerical spectral values for human and gorilla ND6 proteins. This difference plot should be compared with Figures 26 and 28, when by using the virtual genetic code the ND6 proteins were first converted into a sequence of four nucleotides, and then converted into binary sequence by using binary codes for A, C, G and U. Observe that again we obtained the difference plot that shows the same nine sites in which the two proteins differ (except that in Figure 26 the individual spectral peaks may consist of 1−3 bases). The differences in the heights of the peaks in the difference plots, as we know, are not important. Thus, we demonstrated the validity of the virtual generic code for graphical representations of proteins, because we obtained the same results using two conceptually completely unrelated graphical representations. Moreover, one can view the six-digit binary codes of Table 11, which led to Figure 28, as abstract codes, ignoring thus the underlying graphical representation of Figure 27. Hence, one can argue that graphical representations of proteins are not essential for determining the legitimacy of the virtual genetic code. They may merely be viewed as a convenience, which offers visual representation of proteins.

It may be of interest to mention that it was George Gamow who was first to suggest that the genetic code should involve three nucleotide bases, based on the fact that 4^3 is the smallest power of four that exceeds 20. Crick et al.[154] were first to demonstrate experimentally that indeed three DNA bases are responsible for encoding a single amino acid by being able to show that induced mutations of one, two, or four nucleotides produced nonfunctional strains but mutations of three nucleotides resulted in mutant strains that are functional. Clearly for the four nucleotides, one needs four two-digit binary numbers, and thus for each codon, one needs six binary digits, two for each base.

The six-digit binary codes for each amino acid transform a protein into a binary sequence, which in the case of human and gorilla ND6 proteins start as MMYALFLLSV... and MTYVLFLLSV..., respectively, which produce the following:

001101001101110011011001101110111110101110100011010011011100110110011011101111101011101000110100110111001101100110111011111010111010

1110111011011111...(human)
1110111011011111...(gorilla)

The bottom part of Figure 29 is a graphical representation of the difference of the above two binary sequences (of length 1044). It is clear that any numerical substitution of the 20-letter symbols of the 20 natural amino acids will yield the same difference plot as the one shown in Figure 29, with the virtual genetic code being one such specific selection. It also ought to be clear that alternative selections of 20 codes would produce the same results. For example, instead of the 20 codons used for constructing of the virtual genetic code of Table 11, one may choose among all synonymous codons (i.e., cases when different nucleotide triplets codes for the same amino acid), those codons that are most frequently coding the corresponding amino acids.[155,156]

6.4. Search for Alternative Representations of Proteins

The main reason for construction of alternative graphical representations for proteins is the same that holds for continuing interest in alternative graphical representations of RNA/DNA, which ultimately has been the main reason for construction of an ever-increasing number of topological indices in QSAR.[157−164] Alternative graphical representations of molecules, RNA/DNA, and proteins will lead to *novel* mathematical and structural invariants that can serve as additional molecular descriptors for such systems. Whereas in general the number of mathematical invariants of molecular systems is unlimited, it is essential, when constructing

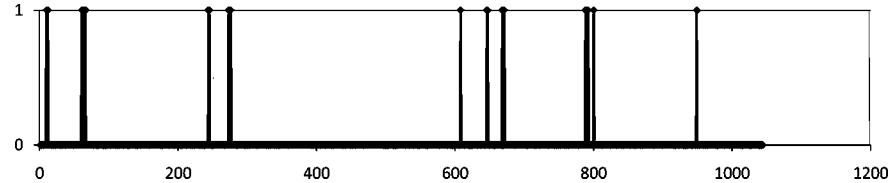


Figure 28. Difference between human and chimpanzee ND6 proteins when amino acids are coded by six-digit binary codes of Table 11.

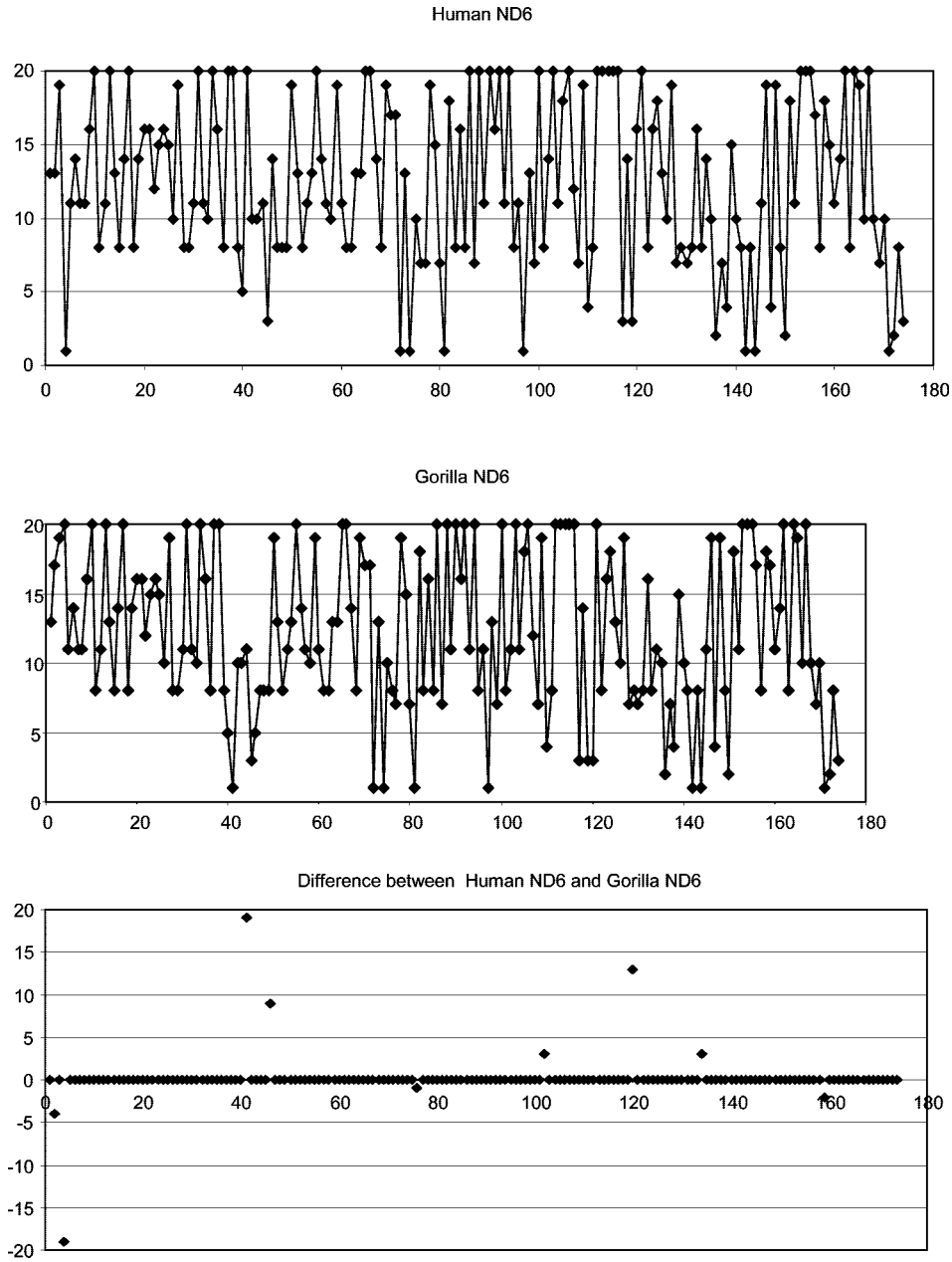


Figure 29. Spectral representation of ND6 proteins of human and gorilla based on assignment of values $y = 1$ to $y = 20$ to the 20 amino acids. The bottom graph depicts the difference between the spectral representations of the two proteins.

additional molecular descriptors, to keep in mind that their *raison d'être* is that they must be *useful* if they are to be taken seriously. Mathematics may afford the luxury to be "pure" and "useless",[165] but the importance of Applied Mathematics, which is much closer to the needs of mechanics, physics, chemistry, and, more recently, biology and medicine, is proportional to its relevance in dealing with problems of natural sciences. Mathematical descriptors, whether for molecules, for RNA/DNA, for proteins, or even for proteome, have to parallel some properties of molecules, proteins, or proteome, otherwise they will remain irrelevant.

A search for mathematical descriptors for complex systems and their construction may be even viewed in some instances as an "art", but in contrast to classical arts, such as painting, sculpture, and music, where artists are free to experiment, here the burden of "experiment" is on the "artists" themselves, that is on authors, who have to demonstrate that the proposed innovations, novel descriptors and novel approaches, are *useful* and that they may lead to *novel insights* into the complex systems considered. A "proof" of usefulness of mathematical descriptors is in demonstrating that they offer either novel structural insights or simpler characterizations of complex molecular systems, which include DNA, RNA secondary structures, proteins and protein folding, and proteomics. Here, we will focus on proteins, for which in very recent times several novel graphical representations have been advanced, the essential feature of which will be briefly outlined in the next sections.

6.5. 8 × 8 Tables of Codons

We have outlined Jeffrey's graphical representation of DNA, introduced in Figure 9, and have illustrated it on the coding sequence of the first exon of human β-globin gene in Figure 10, as well as on two proteins represented by some 500 "spots" in Figures 11 and 12. In the past, nobody applied the algorithm of Jeffrey to very short DNA segments such as the triplets of the nucleotide bases forming codons. In

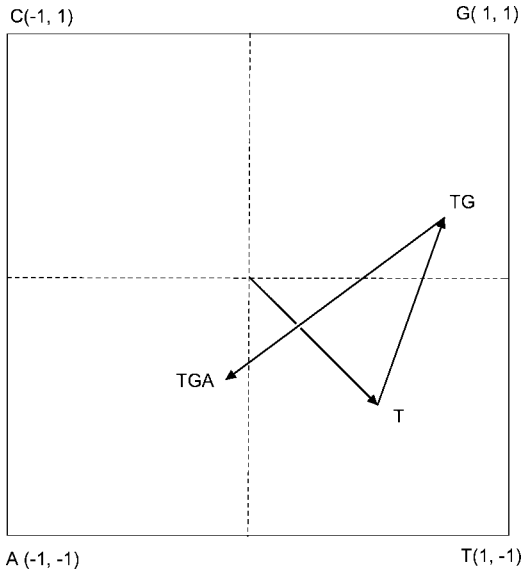


Figure 30. Graphical construction of the codon TGA.

Figure 30, we show a graphical representation of codon TGA by the chaos game algorithm of Jeffrey.

One starts at the center of the square and moves first toward the corner T, ending in the fourth quadrant, then continues toward the corner G, ending in the first quadrant labeled with G at the point indicated as TG, and finally one arrives in the third quadrant belonging to A, at the site TGA. Almost trivial, but what is far from trivial, is to consider that the same construction is repeated for the remaining 63 codons on *the same* "magic square". One then obtains an additional 63 spots, which, as one can see from Table 13, are distributed uniformly inside the interior of the magic square. For better visibility, the interior of the square is divided into 64 square cells, and the center of each cell is the "spot" for individual codons.

The result of the above simple "exercise" is, in a way, astonishing. All 64 codons, when superimposed on the same map, yield a chesslike 8 × 8 table with unique assignment of individual codons to all available 64 locations.[166] Of course, we should add that the distribution of codons depends on the particular assignment of the four bases to the four corners of the magic square.

The significance of the 8 × 8 "Table of Codons" is that it is a precursor to "Table of an 8 × 8 Table of Natural Amino Acids" (AAs), in which all 20 AAs and the three "stop" signals are assigned to the 64 cells of the table. Because of the degeneracy of the genetic code, all amino acids, with the exception of Met and Trp, appear in several locations in the 8 × 8 table. Thus, for instance, arginine (R), leucine (L), and serine (S) appear six times, with each of these three amino acids being coded by six triplets; alanine (A), glycine (G), proline (P), threonine (T), and valine (V) appear four times in the 8 × 8 table, with each of them being coded by four triplets; isoleucine (I) appears three times, being coded by three codons; nine amino acids, aspargine (N), aspartic acid (D), cysteine (C), glutamic acid (E), glutamine (Q), histidine (H), lysine (K), phenylalanine (F), and tyrosine (Y), are each coded by two triplets; and methionine (M) and tryptophan (W) are coded each by a single triplet. The three triplets, TAA, TAG and TCA, represent the "stop" codons.

Observe that, in Table 13, all triplets in the first quadrant end on G, all triplets in the second quadrant end of C, all in the third quadrant end on A, and all in the fourth quadrant end on T. Hence, the quadrant in which amino acid is to be found is determined by the *third base* of the triplet. However, it is well-known that only the first two bases play the dominant role, and that the third base is the least important for the genetic code, in view that several amino acids are fully determined by the first two bases of the code. A consequence of this is that the codons that encode the same amino acids are scattered all over the 8 × 8 board in cells belonging to different quadrants. It would be desirable, if possible, to distribute the codons in an 8 × 8 table so that the same amino acids are grouped in close locations, rather than being scattered.

It turned out that it is possible to construct such a Table of Codons by reformulating the algorithm of Jeffrey for graphical representation of DNA so that the *first base* determines the quadrant in which the codons will appear. This could not have been accomplished by merely changing the assignment of the four bases to the four corners of the square and still using Jeffrey's algorithm. In such cases, one obtains a different distribution of triplets, but again the locations of triplets will be determined by the third, the least important base, which determines the quadrants in which codons will be located. A way to arrive at a Table of Codons in which related triplets would be grouped together as much as possible is to design an algorithm in which the dominant base, the first base, determines the quadrants and the structure of the table.

An approach that leads to such an arrangements of codons has been outlined recently, and the corresponding algorithm is illustrated in Figure 31, on the triplet ATG (different from TGA of Figure 30). At the left side of Figure 31 is shown the graphical construction for the triplet ATG using Jeffrey's algorithm, and on the right is shown the construction of the novel site for the same triplet based on a modified algorithm that gives priority to the first two bases. From the comparison of the two diagrams, we see that the first step of both algorithms is the same; however, the new algorithm confines the second and third steps of the construction to the interior of the same quadrant. Some parallelism between the two

Table 13. Distribution of the 64 Codons (in Locations in the Center of Individual Cells) Obtained by Applying the Algorithm of Jeffrey When Superimposed on the Same Magic Square

CCC	GCC	CGC	GGC	CCG	GCG	CGG	GGG
ACC	TCC	AGC	TGC	ACG	TCG	AGG	TGG
CAC	GAC	CTC	GTC	CAG	GAG	CTG	GTG
AAC	TAC	ATC	TTC	AAG	TAG	ATG	TTG
CCA	GCA	CGA	GGA	CCT	GCT	CGT	GGT
ACA	TCA	AGA	TGA	ACT	TCT	AGT	TGT
CAA	GAA	CTA	GTA	CAT	GAT	CTT	GTT
AAA	TAA	ATA	TTA	AAT	TAT	ATT	TTT

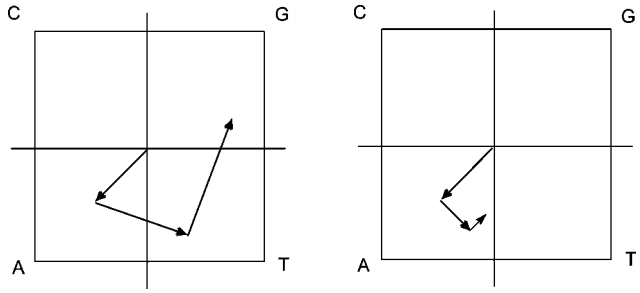


Figure 31. Illustration of the original Jeffrey's algorithm and the modified Jeffrey's algorithm, which gives priority to the first two nucleotide bases in codon ATG.

algorithms is apparent. The second step in the new algorithm is obtained from the first step by focusing on the quadrant belonging to the first base and ignoring the remaining quadrants. The selected quadrant is now considered as the smaller "magic square", to the four corners of which the labels A, T, G, and C are assigned afresh, in the same relative pattern as assigned initially to the magic square. Observe that the first "move" brought us to the center of the smaller magic square, so from the center of this quadrant (which is the place where the first base brought us) one moves halfway toward the corner of the second base, the corner T, of that quadrant. One thus arrives at a center of a still smaller square, to the four corners of which one again assigns the labels A, T, G, and C. By moving to the corner belonging to the third base, one completes the construction of the graphical representation of the codon considered.

In Table 14, we show the revised 8 × 8 Table of Codons. Now, as can be seen, all codons in the first quadrant *start* with guanine, all codons in the second quadrant *start* with cytosine, all codons in the third quadrant *start* with adenine, and all codons in the fourth quadrant *start* with thymine. Thus, with this Table of Codons, we have accomplished the task of grouping codons having the same first nucleotide in close neighborhood. It is interesting to compare more closely Table 13, based on Jeffrey's algorithm, and Table 14, based on the modified algorithm. Observe that the 16 palindrome codons (which are the same whether read from left to right or vice versa, e.g., CCC or AGA) are in the same locations in both tables. Moreover, the comparison of the two tables shows that Table 14 is the same as Table 13 if each entry in Table 13 is read from right to left, instead of from left to right! The two algorithms are essentially different if applied to longer DNA sequences. If one would continue this process, using the modified algorithm just outlined, for any DNA sequence one would end with a location in the quadrant determined by the first base regardless of how long the DNA sequence is considered. The new algorithm thus not only leads to a novel distribution of codons dominated by the first base but hints on the possibility of constructing an atlas in which every DNA and every protein can be represented by a *single* point, corresponding to the end base, which is located in the quadrant of the first base. Such a map can be viewed as an atlas of the whole "DNA World" or "Protein World," respectively. Of course, the construction of maps for larger DNA and larger proteins would require higher and higher precision, because at each step the scale of the square cell used for continuing the construction is reduced by a factor of 2.

We should add that the element of arbitrariness associated with the assignment of the four bases to the four corners of the square can be eliminated by considering a spatial alternative to the square—a tetrahedron. Randić and Zupan proposed a 3D variant of Jeffrey's approach in which the interior of a regular tetrahedron, rather than interior of a square, is used to depict bases of DNA. This results in a unique 3D spatial graphical representation of DNA, except for the chirality of such representations associated with left−right reflections, which here are immaterial, and which will not be reflected in the numerical characterizations of such 3D graphical patterns. In such representations, each base is associated with triplet coordinates (x, y, z) in the interior of a tetrahedron.[64,86] In a tetrahedron, even after the first corner is selected, the remaining three corners of the tetrahedron (but not of a square) remain fully equivalent (except for the chirality, which is of no interest here).

Observe that, in Table 14, most of the codons belonging to the same amino acids are grouped together. For example, the codons CCC, CCT, CCA, and CCG, which all code for proline, are in the left upper corner of the table; the codons GGG, GGT, GGA, and GGC, which all code for glycine, are in the right upper corner of the table; and similarly codons that code for alanine (A), tryptophan (T), and valine (V) are grouped together. Similarly, amino acids coded by two codons are close to each other, even if not adjacent, and the same is true of the three stop codons TAA, TAG, and TGA. However, the six triplets that code for arginine (R), lysine (L), and serine (S) are separated into two blocks each. This table would almost suffice. However, it is desirable that the center of gravity of all the triplets that code for the same amino acid should be different for each amino acid. This is not the case with

Table 14. Distribution of the 64 Codons (In Locations at the Center of Individual Cells) Obtained by Applying the Modified Algorithm of Jeffrey and the Location of Corresponding Amino Acids

CCC	CCG	CGC	CGG	GCC	GCG	GGC	GGG
P	**P**	**R**	**R**	**A**	**A**	**G**	**G**
CCA	CCT	CGA	CGT	GCA	GCT	GGA	GGT
P	**P**	**R**	**R**	**A**	**A**	**G**	**G**
CAC	CAG	CTC	CTG	GAC	GAG	GTC	GTG
H	**Q**	**L**	**L**	**D**	**E**	**V**	**V**
CAA	CAT	CTA	CTT	GAA	GAT	GTA	GTT
Q	**H**	**L**	**L**	**E**	**D**	**V**	**V**
ACC	ACG	AGC	AGG	TCC	TCG	TGC	TGG
T	**T**	**S**	**R**	**S**	**S**	**C**	**W**
ACA	ACT	AGA	AGT	TCA	TCT	TGA	TGT
T	**T**	**R**	**S**	**S**	**S**	**stop**	**C**
AAC	AAG	ATC	ATG	TAC	TAG	TTC	TTG
N	**K**	**I**	**M**	**Y**	**stop**	**F**	**L**
AAA	AAT	ATA	ATT	TAA	TAT	TTA	TTT
K	**N**	**I**	**I**	**stop**	**Y**	**L**	**F**

Table 15. 8 × 8 Table of Codons Used for Construction of Polar Coordinates for 64 Codons after Modified Construction Rule, Which Groups All Triplets Having the Same First Base in the Same Quadrant

CCC	CCT	CTC	CTT	TCC	TCT	TTC	TTT
CCA	CCG	CTA	CTG	TCA	TCG	TTA	TTG
CAC	CAT	CGC	CGT	TAC	TAT	TGC	TGT
CAA	CAG	CGA	CGG	TAA	TAG	TGA	TGG
ACC	ACT	ATC	ATT	GCC	GCT	GTC	GTT
ACA	ACG	ATA	ATG	GCA	GCG	GTA	GTG
AAC	AAT	AGC	AGT	GAC	GAT	GGC	GGT
AAA	AAG	AGA	AGG	GAA	GAG	GGA	GGG

Table 16. Table of Amino Acids Based on the Graphical Representation of Codons of Table 15

Pro	Pro	Leu	Leu	Ser	Ser	Phe	Phe
Pro	Pro	Leu	Leu	Ser	Ser	Leu	Leu
His	His	Arg	Arg	Tyr	Tyr	Cys	Cys
Gln	Gln	Arg	Arg	Stop	Stop	Stop	Try
Thr	Thr	Ile	Ile	Ala	Ala	Val	Val
Thr	Thr	Ile	Met	Ala	Ala	Val	Val
Asn	Asn	Ser	Ser	Asp	Asp	Gly	Gly
Lys	Lys	Arg	Arg	Glu	Glu	Gly	Gly

Table 14, because several pairs of amino acids, placed diagonally in adjacent cells, such as the pairs (K, N) in the bottom left part of the table, (H, Q) in the left bottom corner of the second quadrant, (D, E) in the bottom left part of the first quadrant, and (F, L) at the bottom right part of the table, have common centers of mass.

Table 15 is a further modified Table of Codons, which differs from Table 14 by swapping the labels G and T of the magic square. The accompanying Table 16 shows the locations of the corresponding amino acids. As one can see from Table 16, the pairs of amino acids that have been in Table 14 at diagonal positions are now adjacent and, thus, have different centers of mass. From Table 16, it is not difficult to obtain the coordinates that belong to the individual AAs, by averaging the centers of the location of the corresponding codons. For example, the four codons CCC, CCT, CCA, and CCG, at the top left corner of the table, which all code for proline, have the (x, y) coordinates $(-7/2, +7/2)$, $(-5/2, +7/2)$, $(-7/2, +5/2)$, and $(-5/2, +6/2)$, respectively, the average of which is $(-3, +3)$. As one can see from Table 17, where are listed the coordinates of center of mass for all 20 amino acids, in this way one obtains *unique* coordinates for all 20 AAs.[156,168]

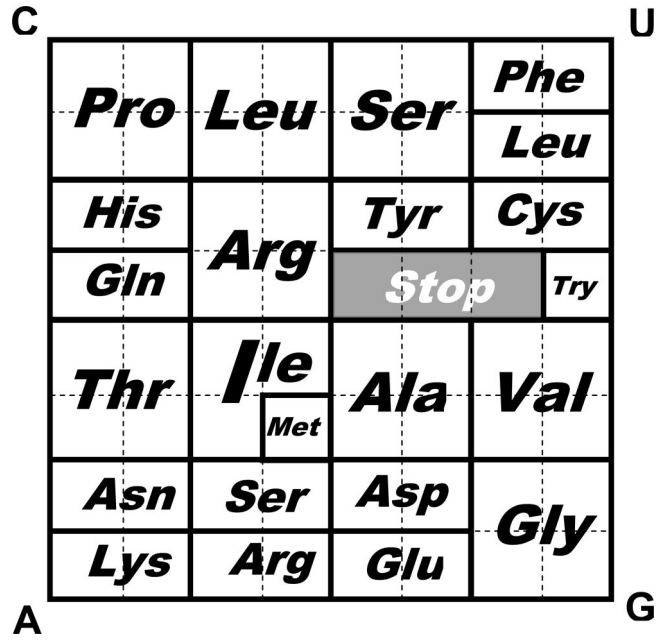


Figure 32. Assignment of amino acids to the 8 × 8 Table of Codons, which groups triplets having the same first and the second base in the same quadrant.

The unique coordinates of amino acid can serve for construction of novel geometrical patterns to represent proteins graphically. Because the novel coordinates are "fixed", the construction of accompanying matrices is relatively simple in comparison with similar constructions based on the chaos game approach, where the same amino acid each time has different coordinates. In Figure 32, we illustrate Table 16 by superimposing three-letter amino acid codes over adjacent cells of codons that encode the same amino acid, while in Figure 33, we have illustrated the locations of the centers of mass for all 20 amino acids as derived from Table 16 or Figure 32.[156]

6.6. Representation of Proteins via the "Magic Circle"

We have referred to the graphical approach of Jeffrey for DNA representations as based on the "magic square",

Table 17. Centers of Mass for Amino Acids (*X*, *Y*) (Based on Table 10),[83] the (*x*, *y*) Coordinates of AA, and the Angular Coordinates (Radians) of the 20 Natural Amino Acids Placed on the Circumference of a Unit Circle (Figure 17) (In Ref 84, the Positions of Ile and His Are in Error and Are Here Exchanged)

virtual code	3-letter code	AA	*X*	*Y*	*x*	*y*	radians
GCG	Ala	A	1/4	−1/4	1	0	0
CGG	Arg	R	−1/4	−1/8	0.951057	0.309017	0.314159
AAC	Asn	N	−3/4	−5/8	0.809017	0.587785	0.628319
GAU	Asp	D	1/4	−5/8	0.587785	0.809017	0.942478
UGC	Cys	C	3/4	3/8	0.309017	0.951057	1.256637
CAG	Gln	Q	−3/4	1/8	0	1	1.570796
GAA	Glu	E	1/4	−7/8	−0.309017	0.951057	1.884956
GGC	Gly	G	3/4	−3/4	−0.587785	0.809017	2.199115
CAU	His	H	−3/4	3/8	−0.809017	0.587785	2.513274
AUC	Ile	I	−7/24	−5/24	−0.951057	0.309017	2.827433
CUC	Leu	L	1/6	17/24	−1	0	3.141593
AAG	Lys	K	−3/4	−7/8	−0.951057	−0.309017	3.455752
AUG	Met	M	−1/8	−3/8	−0.809017	−0.587785	3.769911
UUC	Phe	F	3/4	7/8	−0.587785	−0.809017	4.084070
CCA	Pro	P	−3/4	3/4	−0.309017	−0.951057	4.398230
UCU	Ser	S	1/12	7/24	0	−1	4.712389
ACA	Thr	T	−3/4	−1/4	0.309017	−0.951057	5.026548
UGG	Trp	W	7/8	1/8	0.587785	−0.809017	5.340707
UAU	Tyr	Y	1/4	3/8	0.809017	−0.587785	5.654867
GUU	Val	V	3/4	−1/4	0.951057	−0.309017	5.969026

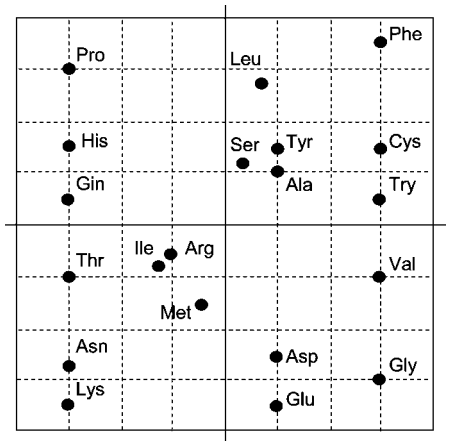


Figure 33. Centers of gravity for amino acids on the basis of averaging the coordinates of multiple codons.

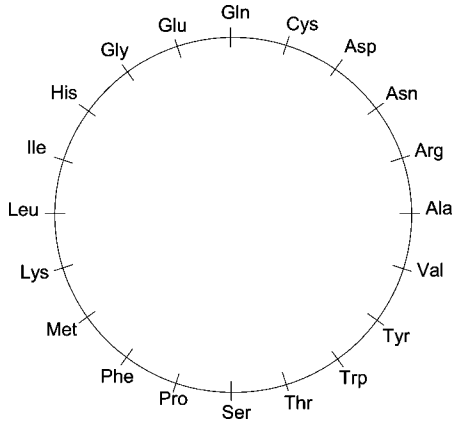


Figure 34. Unit circle with a uniform distribution of 20 natural amino acids on its circumference.

because the interior of a square has been used for the representation of DNA. Randić, Butina, and Zupan[169] generalized the approach of Jeffrey by placing 20 natural amino acids on the periphery of the unit circle, thus replacing a square by a 20-sided polygon. The 20-sided polygon appears, for our needs, as practically being a circle, and therefore, we will refer to this extension of the magic square approach of Jeffrey for DNA to proteins as the "magic circle" approach. On the periphery of the magic circle are listed amino acids in the 3-letter alphabetic order, starting with alanine (Ala), which was placed on the x-axis, and moving counterclockwise around the circle. All 20 amino acids are separated by the constant angle of $2\pi/20$ radians (Figure 34). In this representation, the 3-letter alphabetical order starts with alanine (no. 0) and ends with valine (no. 19).

Using this labeled circle, one can construct a graphical representation of proteins that bears parallelism with the graphical representation of DNA of Jeffrey, except that instead of a scatter of the points representing nucleotides in the interior of the magic square, now one has a scatter of points representing amino acids in the interior of the magic circle. Observe that there is no loss of information with the resulting graphical representation of proteins, just as there had not been any loss of information with the graphical representation of DNA in the interior of the square, even though the reconstruction of the protein from the pattern of scattered points is going to be more tedious.

In Figure 35, we have illustrated graphical representations of two short protein fragments using the approach of the magic circle, and for a better visibility of the construction, we have left the connecting lines that join adjacent AAs in the protein sequence. Leaving the connecting lines is an optional step in the graphical construction of proteins and may be of interest especially for smaller proteins. The zigzag curve may help visualization of similarities and differences among proteins and may facilitate construction of **D/D** matrices for the proteins considered. Not surprisingly, the constructed graphical "images" of shorter proteins (either having the connecting lines or without them) do not offer much insight to visual inspection, just as the construction for DNA sequences also resulted in "chaotic" patterns.

However, as will be now described, further processing of such 2D graphical representation, which may be numerical and/or graphical, may offer useful insights into similarities of proteins. The construction of global and local **D/D** matrices[170] is one such "processing", though it is numerical, leading to numerical matrices, which then may lead to potentially useful descriptors for proteins. Local **D/D** matrices, which consider a subset of amino acids in proximity, have not been much used so far, though they have an important numerical advantage in not being computationally intensive. The entries in local **D/D** matrices, in contrast to the standard global **D/D** matrix, refer to local domains of a protein sequence, and thus result in *sparse* matrices, i.e., matrices that have numerous zero entries. In contrast, the global **D/D** matrix is *dense*, with typically zero entries only on the main diagonal. For a brief discussion on sparse and dense graphs and matrices, see ref 171. Let us also point to an additional advantage of using "fixed" (x, y) coordinates for amino acids, besides the obvious one that fixed coordinates do not require continuous updating when calculating matrix elements of **D/D** matrices. By using a fixed set of coordinates in the construction of **D/D** matrices, although the **D/D** matrix will not be sparse, it will nevertheless have numerous zeros, which will occur whenever the Euclidean distances between the same amino acid (AA) are considered.

In contrast to the "numerical" constructions accompanying the zigzag representations of proteins, which may result in less laborious processing, a "graphical" processing obtained by plotting the *differences* between the (x, y) coordinates of the corresponding amino acids in two sequences, as we will see, may lead to "instant gratification", by offering useful alternative visual patterns that may throw more light on similarities and differences among the proteins considered. In the case of the two protein fragments of Figure 35, in such an analysis, one obtains Figure 36, in which on the x-coordinate is the running index i of AAs in a protein and on the y-axis are plotted the Euclidean distances between the "spots" representing amino acids at the same sequential position, which may be the same amino acid or different. The Euclidean distances are given by the expression: $\sqrt{\{(x_{Ai} - x_{Bi})^2 + (y_{Ai} - y_{Bi})^2\}}$, where the subscripts A and B indicate amino acids in protein I and protein II.

Figure 36 clearly shows four sites at which there is a sudden "jump" in the Euclidean distances of the corresponding amino acids. A close look at the two sequences reveals that these four sites are the places at which the amino acids in two proteins differ. After each of the sudden "jumps", as one can see from Figure 36, there is a gradual decrease in the corresponding distances for the following amino acids. The "decay" parts of the "peaks" in Figure 36 show that the

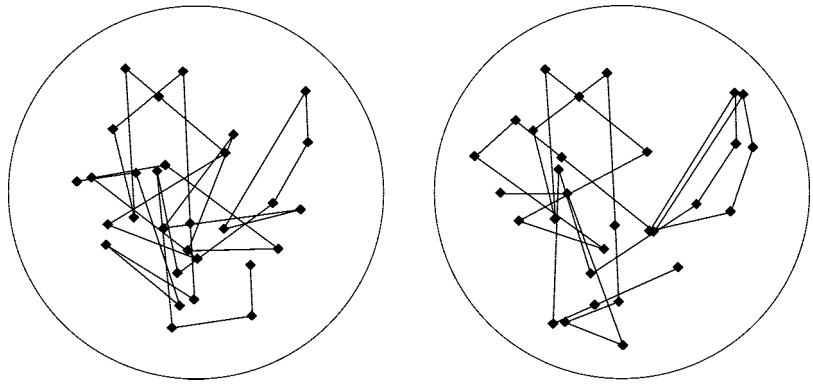


Figure 35. Graphical representation of the proteins fragments I and II of yeast *Saccharomyces cerevisiae*, having 30 amino acids. Reproduced by kind permission from ref 169. Copyright 2006 Elsevier Publ.

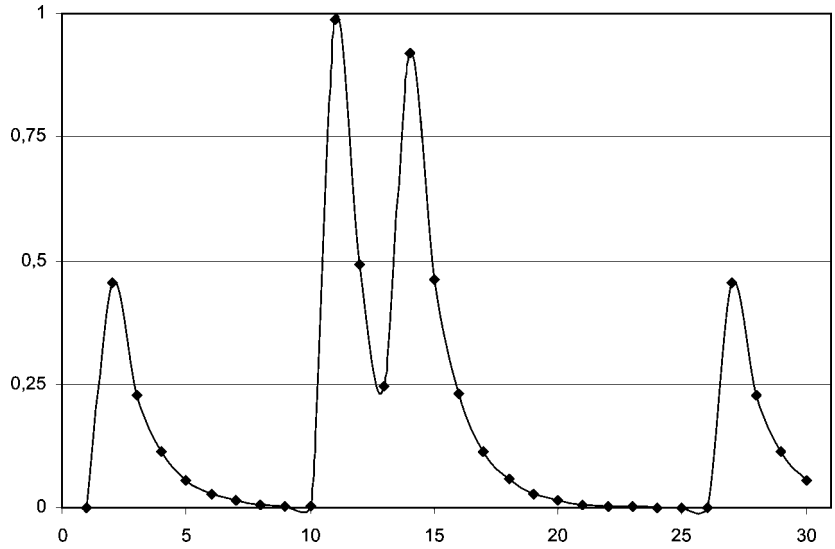


Figure 36. Difference in coordinates of corresponding amino acids in two proteins showing locations of mismatch in protein sequences. Reproduced by kind permission from ref 169. Copyright 2006 Elsevier Publ.

same fragment in different proteins, regardless of the site of "difference", slowly converges to similar patterns of points, until the next "perturbation" (the occurrence of different amino acids in the two proteins) appears.

Figure 36 represents a significant result—it shows that there may be a "graphical" route to one of the central problems of comparative study of proteins, the problem of protein alignment. It represents the beginning of a potentially important novel application of graphical representations of proteins, and by extension, of graphical representations of DNA for the problem of DNA alignment. Thus, we can compare the graphical representation of the first exon of the lemur β-globin gene, shown in the middle of Figure 2, with Figure 37 in which is shown the graphical representation of the first exon of the lemur β-globin gene. Again, without further processing, no useful insight can be obtained by direct comparisons of such graphical representations. However, if one follows the same recipe and constructs the difference in (x, y) coordinates for nucleotides at the same sequential position for human and lemur, then one obtains the plot shown in Figure 38, which immediately shows that the two exons exhibit considerable similarity.

Figure 38 has an appearance of a "spectrum-like" diagram and, as such, offers great visual advantages, though as we have seen, it belongs to a *pair* of proteins, not a *single* protein. Because of potential visual advantages of spectrum-like diagrams, it is of considerable interest to develop spectrum-like graphical representations of proteins, just as was possible for DNA sequences. Not only do such graphical representations facilitate the visual inspection of the similarities and differences among protein sequences, but they may

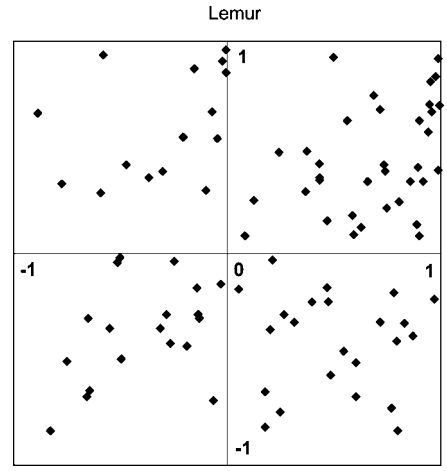


Figure 37. First exon of lemur β-globin gene displayed using Jeffrey's algorithm for graphical representation of DNA. Reproduced by kind permission from Randić, M. Another Look at the Chaos-Game Representation of DNA. *Chem. Phys. Lett*. **2008**, *456*, 84−88. Copyright 2008 Elsevier Publ.

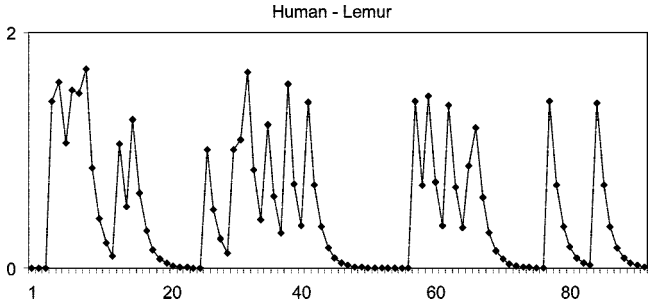


Figure 38. Difference plot between the graphical representations of the first exon of human and lemur β-globin gene. Reproduced by kind permission from Randić, M. Another Look at the Chaos-Game Representation of DNA. *Chem. Phys. Lett.* **2008**, *456*, 84−88. Copyright 2008 Elsevier Publ.

be further manipulated graphically by considering their differences, and also differences when two protein sequences are shifted one relative to the other. This would lead to a graphical, hence, *visual*, approach to the protein alignment problem, a topic to be considered shortly.

6.7. Representation of Proteins Based on Starlike Graphs

Common to all hitherto mentioned graphical representations of proteins and DNA, with the exception of the 3D graphical representation of DNA within the interior of a regular tetrahedron, is the arbitrary assignment of amino acids (or bases in the case of DNA) to alternative geometrical elements of structure used for their graphical representation. The selected assignments of the four nucleotides or the 20 amino acids represent but one particular assignment among a multitude of possible alternative assignments. This is particularly critical in the case of proteins because of the insurmountable 20 factorial possibilities! However, as we will shortly show, it is possible to construct graphical representations of proteins that bypass difficulties and dilemmas of choosing a single option among many alternative possibilities for assigning amino acids to various equivalent elements of the underlying geometrical template used in the construction of graphical representations. The first such advancement in graphical representations of proteins has been introduced recently[70] and is based on using starlike graphs as models for proteins. Starlike graphs have a single central vertex and numerous branches consisting of "rays" in which can appear only vertices of degree 2 or 1. In Figure 39, we have depicted starlike graphs, which have 11 branches and 21 vertices besides the central vertex in three different forms. Observe now, and this is the crux of the approach, that although the three graphs of Figure 39 look different, all three forms represent the same graph. Neither the lengths of edges (which are all the same in Figure 39) nor their relative positioning is important, as long as the connectivity of the graph is maintained correctly. That the three copies of Figure 39 represent the same graph is not difficult to verify: all three graphs have a single central vertex, three rays of length 1, seven rays of length 2, and one ray of length 4. The length of rays is given by the number of edges.

In general, it is very difficult to establish if two graphs are identical or different. This problem, which is known as "graph isomorphism problem", has received considerable attention, both in the mathematical[173,174] and the chemical literature.[175−177] The first chemical reaction graphs were published by Balaban et al. in 1966,[178] and two graphs in particular had striking symmetries. One of these has 20 points and is known as the Desargues-Levi graph, representing the rearrangements of trigonal bipyramidal structures with five different ligands attached to a central atom (taking enantiomerism into account), or the rearrangements of an ethyl−carbenium cation (isotopically labeled at one carbon atom of the ethyl group). When enantiomerism is ignored or when there is no isotopic labeling, graph vertices become pairwise identical and the 20-vertex graph reduces to a graph known as the Petersen graph[38,179−181] having 10-vertices.

In Figure 40, we illustrate the Petersen graph, a cubic (trivalent) graph on ten vertices, in two different forms, each exhibiting distinct symmetry properties of this graph. The Petersen graph is the 5-cage (smallest cubic graph with minimal circuits of five vertices). It is also of interest in chemistry because it represents the degenerate rearrangement of the trigonal bipyramidal complex XY_5, when axial ligands are exchanged with equatorial ligands. It was studied by Dunitz and Prelog in 1968,[182] and later its symmetry properties were investigated.[183] At first sight, and to those not initiated in Graph Theory, such questions may appear simple, if not simplistic, but as Prelog warns:[184]

"*Pictorial representations of graphs are so easily intelligible that chemists are often satisfied with inspecting and discussing them without paying too much attention to their algebraic aspects, but it is evident that some familiarity with the theory of graphs is necessary for deeper understanding of their properties*."

A remarkable property of the Petersen graph is that it is vertex- and edge-transitive, and also transitive for paths of lengths 2 and 3! It is also called *the 5-cage* (smallest graph with girth, or smallest circuit, equal to 5). Its automorphism group is $10 \times 3 \times 2^2 = 120$.

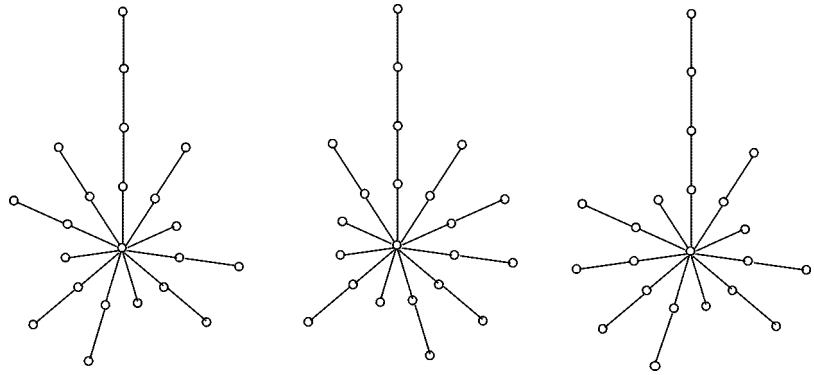


Figure 39. Same starlike graph represented in three different forms. Reproduced by kind permission from ref 70. Copyright 2007 Elsevier Publ.

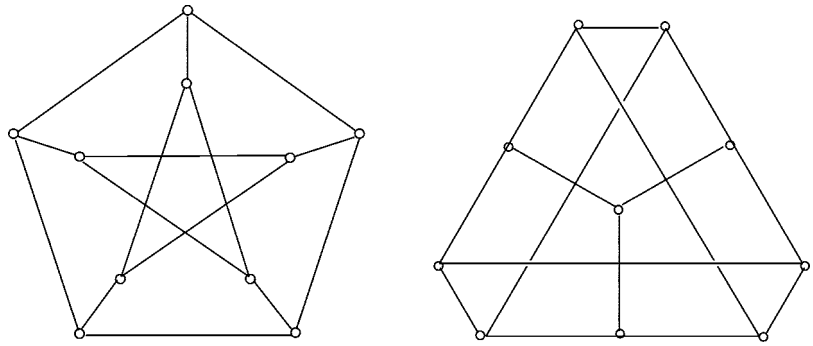


Figure 40. Petersen graph drawn in two alternative forms.

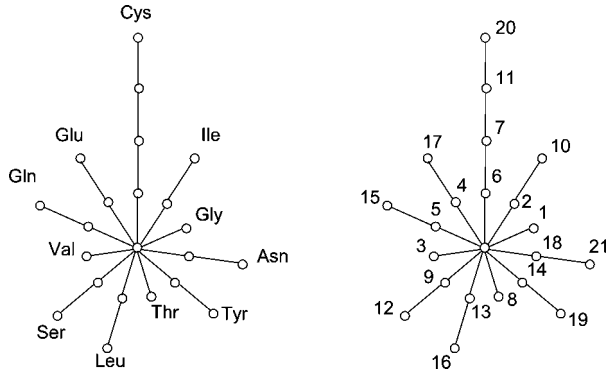


Figure 41. Human insulin strand A represented by starlike graph. Reproduced by kind permission from ref 70. Copyright 2007 Elsevier Publ.

The graph of Figure 39 in fact depicts a small protein: the A strand of human insulin, which has 21 amino acids of 11 kinds. In Figure 41, at the left, we have labeled the 11 branches of the graphs with the 11 amino acids present in insulin, and in the same figure at the right, the vertices of this labeled graph are given sequential numbers in the order as they appear in the protein sequence. The graph at the left clearly has less information than the accompanying graph at the right, which when combined with the information at the left allows the full reconstruction of the protein. In the general case, starlike graphs representing proteins will have 20 branches, with each branch indicating the presence of 1 of the 20 natural amino acids.

As is well-known from Graph Theory, the same graph can be depicted in many different ways, which is generally viewed as a disadvantage, because this calls for isomorphism testing when the results are not obvious. Here, in contrast, we are taking advantage of this disadvantage, exploiting this particular feature of graph representation to our benefit—avoiding the arbitrary assignment of amino acids to alternative geometrical elements of templates used for constructing graphical representations of proteins. One should keep in mind that, regardless of how a particular graph is graphically depicted, or how one labels its vertices, the graph will always have the same set of graph *invariants*. The graphs will appear different, the adjacency matrix and the distance matrix will look different, but any of the computed graph *invariants*, which can serve as protein descriptors, will be the same. Because the properties of graphs (even when displayed differently and using different labels) do not depend on the choice of labels, one can assign the branches of starlike graphs arbitrarily to the 20 amino acids. Hence, one may therefore adopt, for instance, the alphabetical order of amino acids in circling around the central vertex following the scheme already described for labeling the 20 vertices of the "magic circle."

Table 18. Distance Matrix between the Terminal Vertices of the Starlike Graph of Insulin (Fig. 22)

	1	2	3	4	5	6	7	8	9	10	11
1	0	3	4	6	4	4	3	4	4	3	4
2		0	3	5	3	3	2	3	3	2	3
3			0	6	4	4	3	4	4	3	4
4				0	6	6	5	6	6	5	6
5					0	4	3	4	4	3	4
6						0	3	4	4	3	4
7							0	3	3	2	3
8								0	4	3	4
9									0	3	4
10										0	3
11											0

There is another interesting property of graphs, that may not be so well-known, which is of interest here: in the case of trees (i.e., acyclic graphs), which include the starlike graphs adopted for depicting proteins, there is no loss of information if the graph is represented by the distance matrix that involves only the distances between *terminal vertices*.[185] In the case of the graph of Figure 39, this will be an 11 × 11 matrix, shown in Table 18, whose elements are given by the number of edges separating vertices.

In a general case, the terminal-distance matrix for starlike representations of proteins will be at most a 20 × 20 matrix, because proteins are built from 20 natural amino acids (ignoring the exceptional cases when additional post-translation amino acids may occur). If a protein has less than 20 amino acids (as in the case of insulin, which was selected here for illustration), one can augment the terminal-distance matrix of such graphs by insertion of distances for nonexisting amino acids, which are assigned to the central (core) vertex. In this way for any protein, one can obtain a "uniform" 20 × 20 matrix representation, regardless of its size. This has some advantages, particularly when comparing proteins that may differ in their size.

Generally, a direct comparison of proteins (and DNA) of different sizes may introduce some difficulties in that certain invariants, such as eigenvalues of matrices, will result in vectors of different dimensions (or sequences of invariants having different lengths). Truncation or normalization in such cases may introduce additional arbitrary steps that, if possible, should be avoided. When using eigenvalues of matrices involving Euclidean distances as a source for characterizing proteins and DNA, there is an interesting, and not well-known, procedure, which is of interest here. Matrices, the elements of which are defined by higher powers of Euclidean distances, have the property that, regardless of their size, they have m^2 nonzero eigenvalues, where m is the degree of potentiation.[186] Thus, regardless of their size, matrices $\mathbf{D}^2$,

$\mathbf{D}^3$, $\mathbf{D}^4$, $\mathbf{D}^5$, etc., the elements of which represent distances in 2D space, will have 4, 9, 16, 25, etc. nonzero eigenvalues, a fact that allows direct comparison of their spectra without additional restrictions.

One should not overlook the fact that there could be considerable loss of information on proteins represented by starlike graphs if the sequential labels (shown on the right for the graph of human insulin in Figure 41) are omitted. Hence, one has an option to build uniform 20×20 matrices that have only the information on the abundance of amino acids in a protein (based on the distances between the terminal vertices of starlike graphs) and to construct the complete $n \times n$ distance matrices, where n is the number of amino acids in a protein. The former involves limited computations and offers fast screening of a large pool of proteins, while the latter involves information on the sequential occurrence of amino acids in a protein, and can then be applied to a preselected smaller group of proteins for more detailed analysis.

6.8. Protein Profiles

We have mentioned the leading eigenvalue of the **D/D** matrix (and other matrices) as a descriptor for DNA and protein sequences—but in each case, that is just a *single* descriptor. As additional descriptors, one may add the average row sums of matrices (or alternatively the average matrix element), but for complex systems, a set of ad hoc descriptors may still not offer sufficient discrimination between similar sequences. It is desirable to have a set of *ordered* descriptors that can simulate a "basis" for describing such systems. In the case of molecules, such a basis can involve the paths of different lengths,[187,188] the set of extended paths,[189] the set of the connectivity indices,[190,191] the valence connectivity indices,[192−195] and the "molecular profiles"[196−199] for structure−property/activity studies for molecules.

Molecular profiles are constructed from the distance matrix, or related matrices associated with a molecule, by considering the "higher"-order matrices obtained from the initial matrix by raising the individual matrix elements to higher powers. These matrices have been first considered by Randić and Razinger[200] in an attempt to characterize molecular shapes. Raising the individual matrix elements of matrices to higher powers is known in Matrix Algebra as the Hadamard matrix product[201] (also known as the Schur product[202] and the entrywise product, which is also a submatrix of Kronecker's product of matrices[203]) and is one of the matrix operations available in MATLAB.[204]

In Figure 42, we have illustrated protein "profiles" based on suitably normalized leading eigenvalues of the terminal distance matrix and the complete distance matrix for the A strand of human insulin. The entries 1−15 on the abscissa correspond to matrices of order 1−15, in which individual entries of the matrix are raised to powers from 1 to 15. The thus-constructed diagrams have been referred to as "protein profiles," in analogy with the similar molecular diagrams, which have been referred to as "molecular profiles." The word "profile" is taken in its broad sense as "any outline or contour",[205] thus not implying the outline of actual protein form. Protein profiles are sensitive to the amino acid composition of proteins. Thus, they may offer a useful visual comparison between proteins and, when necessary, can be supplemented by additional characterizations based on considerations of additional accompanying data such as those encoded by sequential labels on the amino acids that are present.

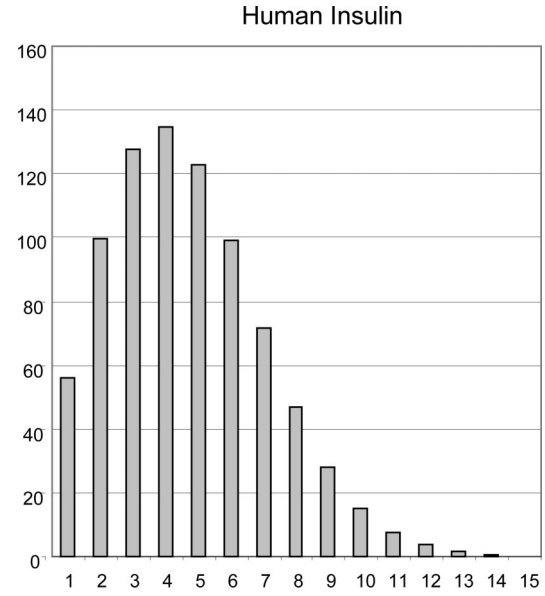


Figure 42. Profile of strand A of human insulin on the basis of reduced distance matrix. Reproduced by kind permission from ref 70. Copyright 2007 Elsevier Publ.

The representation of proteins by starlike graphs has offered, besides the graphical visualization of proteins, an important feature to the representation of proteins that should be kept in mind. By virtue of being independent of any adopted assignment of amino acids to individual graph branches, the accompanying numerical descriptors will be free of artifacts that accompany *other* graphical representations of proteins. So while other graphical representations may involve, to a lesser or greater extent, spurious features specific to the particular graphical representations under consideration, which are influenced by the adopted assignments of amino acids (such as the use of alphabetical order of amino acids for the magic circle), the invariants based on starlike representations of proteins will reflect only *genuine* inherent properties of protein sequences.

One disadvantage of the starlike graphical representation is that, in the case of large proteins, depicting such proteins may be not only sizable but also uneconomical regarding the used space, because the space between adjacent branches of the star graphs is "wasted". This aspect of starlike representation was recently addressed, and a solution was proposed in which one combines the "good feature" of Jeffrey's algorithm, the compactness of the representation, which is illustrated in Figure 43. Simply, one first draws the unit circle in the center of which is placed the "core" vertex of the star graph, while the 20 directions follow around the periphery of the unit circle. One starts then in the middle and moves 1/2 toward the site of the first amino acid. Similarly, one continues with successive amino acids, each time moving from the center of the circle in the direction of the corresponding amino acid. When the same amino acid appears again, one moves from the site of its first occurrence and moves only at the distance of 1/2 from the periphery. In this way, all points representing N amino acids of a protein of length N will be within the unit circle. The end result will have some similarity with paintings inspired by the hyperbolic plane.

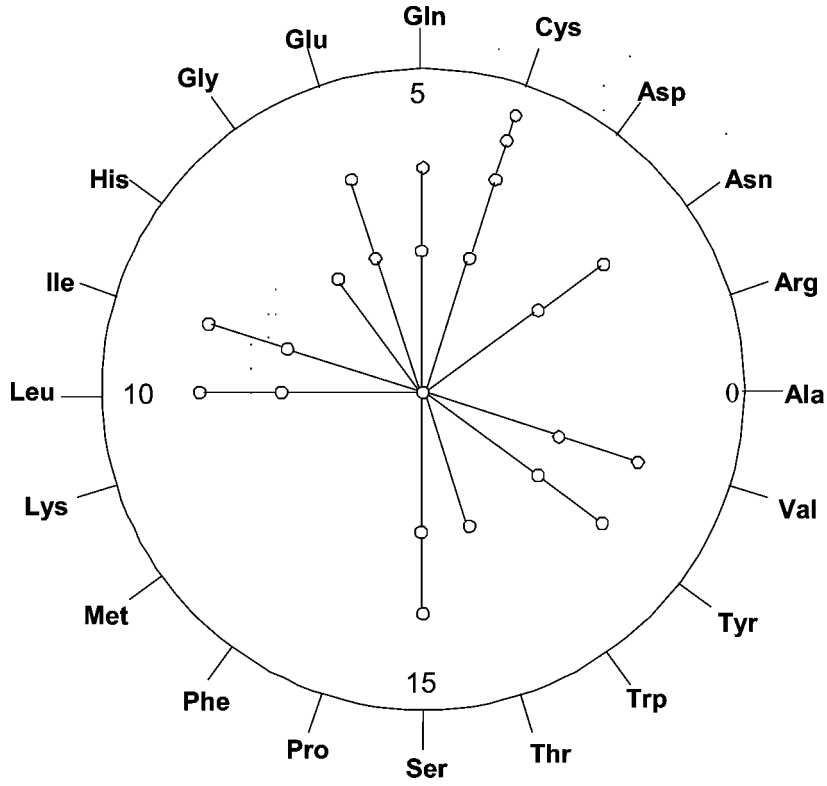


Figure 43. Graphical representation of a smaller protein confined to the interior of the magic circle.

7. Spectrum-like Representation of Proteins

There is one aspect of graphical representations of DNA, codons, and proteins on which we have not commented, except briefly, and this involves the dimensionality of the representation. Clearly 2D graphical representations offer the most for visual inspection, while 3D ones may have added graphical complexity, which occasionally can contribute to one's advantage by viewing such representations as pairs of stereodiagrams, when available. On the other hand, 1-dimensional representations have limitations when literally displayed along a line. However, one can display 1-dimensional representations of mathematical objects (which are defined by a *single* parameter or coordinate) by listing the *single parameter* as the variable on the *y*-axis at *equal* intervals on the *x*-axis. This formally transforms the 1-dimensional representation into a 2-dimensional object, which differs from genuine 2-dimensional mathematical objects, the points of which are defined by a pair of coordinates (*x*, *y*). Representations having a single degree of freedom associated with them can be displayed as histograms (of equal width) or frequency distributions (at equal intervals). We will refer to such figures as "spectrum-like" representations, which differ from "2D spectra" in that, in the latter, the spectral points are characterized by a pair of coordinates, that is, the points need not be at equidistant separations along the *x*-axis.

Already in Figures 14 and 15 we have shown spectrum-like 1-dimensional representations of DNA based on the use of four horizontal lines.[117,118] In contrast, the spectral-like graphical representations of DNA reported by Zupan and Randić[86] (and illustrated in Figure 44 on the first exon of β-globin gene of 10 species) were obtained by applying Jeffrey's algorithm to individual DNA bases using a (−1, +1) line interval for determining the "amplitudes" of points, instead of placing them inside a 2D square. The pair of bases A and C were assigned value −1, and the pair of bases T and G were assigned the value +1. Again, one starts from the origin (the middle of the segment) and moves halfway to either the terminal point −1 or +1, depending on the base, and from that point again halfway toward terminal points −1 or +1, depending on the next base, and so on. In fact, the spectra thus-constructed represent the projections of a 2D representation of the same DNA based on Jeffrey's magic

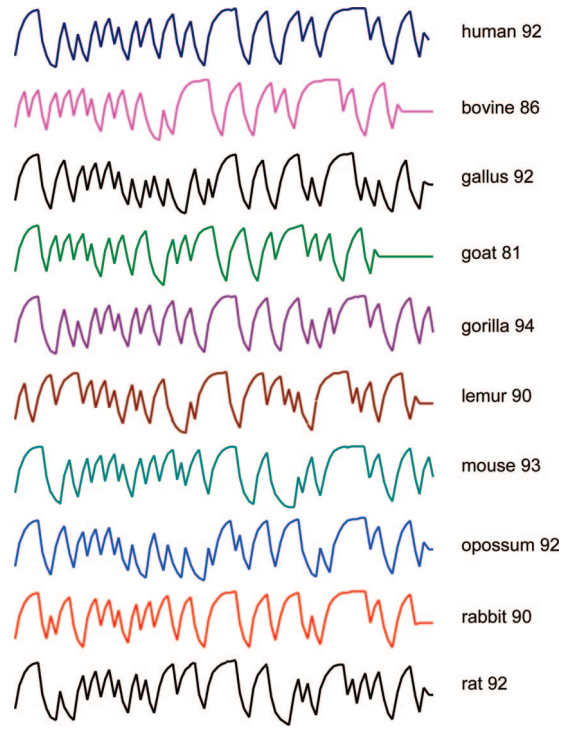


Figure 44. Spectrum-like graphical representation of the first exon of the β-globin gene of 10 species. Reproduced with permission from Zupan, J.; Randić, M. Algorithm for Coding DNA Sequences into "Spectrum-like" and "Zigzag" Representations.[95] Copyright 2005 American Chemical Society.

Table 19. Reconstruction of the Alignment of Two Carboxypeptidase Proteins

```
-KILGIDPNVTQYTGYLDVEDEDKHFFFWTFESRNDPAKDPVILWLNGGPGCSSLTGLFFELGPS
   ||||       ||       |||| | |||||||| || ||||||||||| ||| ||||||
PSKLGIDTVKQWS-GYMDYKDS-KHFFYWFFESRNDPANDPIILWLNGGPGCSSFTGLLFELGPS

SIGPDLKPIGNPYSWNSNATVIFLDQPVNVGFSYSGSSGVSNTVAAGKDVYNFLELFFDQFPEYV
||| | ||| |||||| ||  ||| ||  |||||           |||| | |||| |  ||
SIGADMKPIHNPYSWNNNASMIFLEQPLGVGFSYGDEKVSSTKL-AGKDAYIFLELEFEAFPHLR

NKGQDFHIAGESYAGHYIPVFASEILSHKD-RNFNLTSVLIGNGLT
    ||||||||||||||           | |||||| |||| |
SN--DFHIAGESYAGHYIPQIAHEIVVKNPERTFNLTSVMIGNGIT
```

square on the base of the square and depicting the projections as points on the *y*-coordinate at uniform distances along the *x*-axis. The difference between the two approaches is that, on using of the values 1−4 for the four bases, each of the four nucleotide bases always receives the same numerical value for its amplitude, while Jeffrey's algorithm requires updating of the coordinate value for successive bases in the DNA sequence.

In order to obtain a 1-dimensional graphical representation of proteins that would be analogous to the four-line DNA representation, one could assign to each of the 20 amino acids a numerical value, such as entries from 1 to 20. Another possibility is to use the angular polar coordinate of amino acids, which are uniformly arranged on the circumference of the unit circle. In a similar manner, one could arrange the 64 codons uniformly on the periphery of the unit circle and assign to them polar angles (multiples of $2\pi/64$ radians) and arrive at a 1-dimensional "spectrum-like" representation of DNA based on codons. Observe that, in these 1-dimensional representations of DNA (based on the four nucleotides or codons), or of protein sequences, one transforms the *alphabetic* sequences into *numerical* sequences. Numerical sequences allow one to perform simple numerical operations on the elements of the sequence, such as subtracting the corresponding members of two sequences or subtracting sequences that have been shifted one relative to the other. We will see in the next section that this is the essential step for arriving at graphical solutions to the problem of DNA and protein alignment.

8. Graphical Approach to Protein Alignment

There is no rigorous solution to the problem of protein−protein structure alignment; thus, this challenging problem continues to attract the attention of researchers in bioinformatics. The existing algorithms for protein alignments involve computationally intensive techniques, which include dynamic programming,[206] probabilistic Monte Carlo approach,[207] the sequence alignment genetic algorithm,[208] and alternative schemes such as three-dimensional clustering,[209,210] the combinatorial extension of an alignment path,[211] and graph-theoretical approaches.[212,213] The computer-oriented approaches consider various penalties for deletion, substitution, and permutation of sequence labels (i.e., amino acids), which are associated with the metrics of Levenshtein,[214] also known as the "edit distance." While it is plausible and legitimate to weigh different operations differently, the choice of the weighting factors nevertheless involves elements of arbitrary decision making.

In contrast to these available computer-based programs for protein alignment, which search for an optimal alignment of proteins when various penalties for deletions, substitutions, and gaps are assumed, we will outline a graphical approach for the direct comparison of two protein sequences. Admittedly, this represents but one aspect of the topic of protein alignment, but hopefully, with time, other areas of the problem of protein alignment may be similarly treated. The recently outlined graphical approach to protein alignment identifies the same amino acids in two protein sequences by locating the zeros on the plot of the difference between two numerical representations of two proteins. To arrive at the complete analysis, however, one has to consider the differences between sequences of proteins when shifted by one or more positions relative to the other, both to the left and to the right.

We will illustrate the graphical alignment of proteins on two proteins of Table 19 that have about 170 amino acids.[215] The first protein pertains to carboxypeptidase Y from *Saccharomyces cerevisiae* (baker's yeast), whereas the second belongs to the mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region, also from *Saccharomyces cerevisiae*. The 1D graphical representations of the two proteins are illustrated in Figure 45. A close look at Figure 45 would reveal 20 different amplitudes, all within the range $0-2\pi$ radians, where values having the same "height" correspond to the same amino acid. Thus, the spots at the top (valine) and the bottom of the spectra (alanine) immediately indicate that protein 1 has 12 valine ($y = 6$) and 7 alanine AAs ($y = 0$), whereas protein 2 has 5 valine and 9 alanine AAs. In fact, it is not difficult to count the number of spots on each (invisible) horizontal line, by simply gradually sliding a sheet of paper (or using a computer), which gives the abundance count for proteins. Thus, in the case of protein 1 and protein 2 of Figure 45, one obtains for the 20 amino acids, ordered alphabetically according to their three-letter codes:

(A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V)

the following 20 abundance values, respectively:

(7, 2, 13, 11, 1, 4, 7, 18, 10, 4, 15, 2, 0, 15, 11, 16, 8, 3, 8, 12) for protein 1 and

(9, 3, 11, 9, 1, 3, 9, 17, 13, 6, 13, 9, 4, 14, 11, 15, 7, 4, 8, 5) for protein 2.

An additional look at Figure 45 also shows that, from the spectrum-like representation, one can immediately identify repeated occurrences of the same amino acid by simply looking at the spectra for locations of adjacent "spots" on

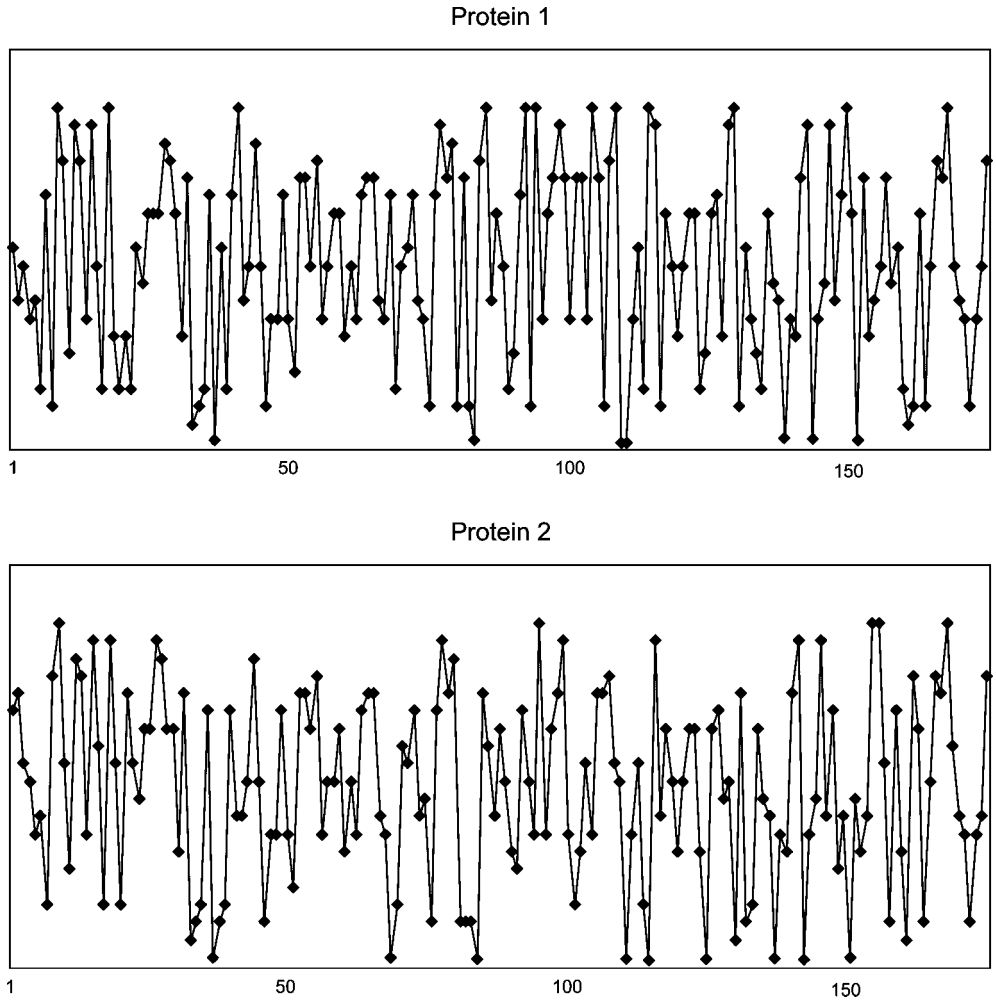


Figure 45. Plot of angular coordinates of amino acids of the unit circle for the 174 amino acids of carboxypeptidase Y from *Saccharomyces cerevisiae* (top) and of 171 amino acids of mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region also from *Saccharomyces cerevisiae* (bottom).

the same horizontal line. Thus, in protein 1, one finds AA, GG, FFF, and FF twice and SS thrice, whereas in protein 2, one finds NNN, GG, HH, LL, and FF three times and SS twice.

The 20-component "abundance" vectors allow a fast preliminary screening of proteins for their similarities or lack thereof. Clearly there is considerable loss of information in such vectors, which tell nothing about the distributions of amino acids, but on the other hand, the similarity of these vectors is a necessary, even if not sufficient, condition for similarity among proteins. Already a glance at the two vectors suggests that protein 1 and protein 2 have an appreciable degree of similarity. A plot of the two 20-component vectors representing the abundance of the two proteins one against the other is shown in Figure 46, which shows a fair correlation, with three outliers: alanine (A), methionine (M), and leucine (L).

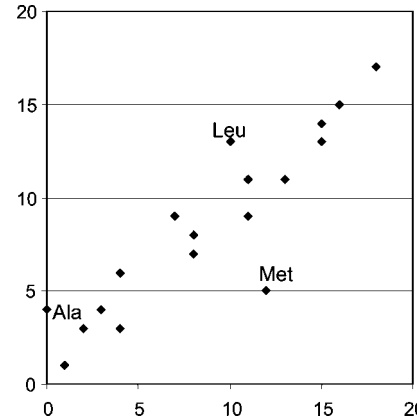


Figure 46. Plot of the abundance of amino acids in two carboxypeptidase proteins.

If we were to plot the difference between the spectral representations of protein 1 and protein 2, one would find that the resulting plot would, for the most part, oscillate above and below the *x*-axis, which is because there are no segments of the two proteins in which amino acids overlap and which would result in differences equal to zero except for a few accidental cases. However, when the two sequences are shifted by one and two locations, one obtains the difference diagrams shown in Figure 47, which clearly show that *significant* portions of the two proteins display alignment for amino acids. The top diagram corresponds to the shift of two sequences by one position and gives alignments of amino acid in the region 22−99. The shift of the two protein sequences by two positions shows additional alignment between the two proteins in the region 108−120. If one continues to shift the two sequences further, by three and four sites, one would find additional local alignments of amino acids in the regions 159−169 and 130−145, respectively.[89]

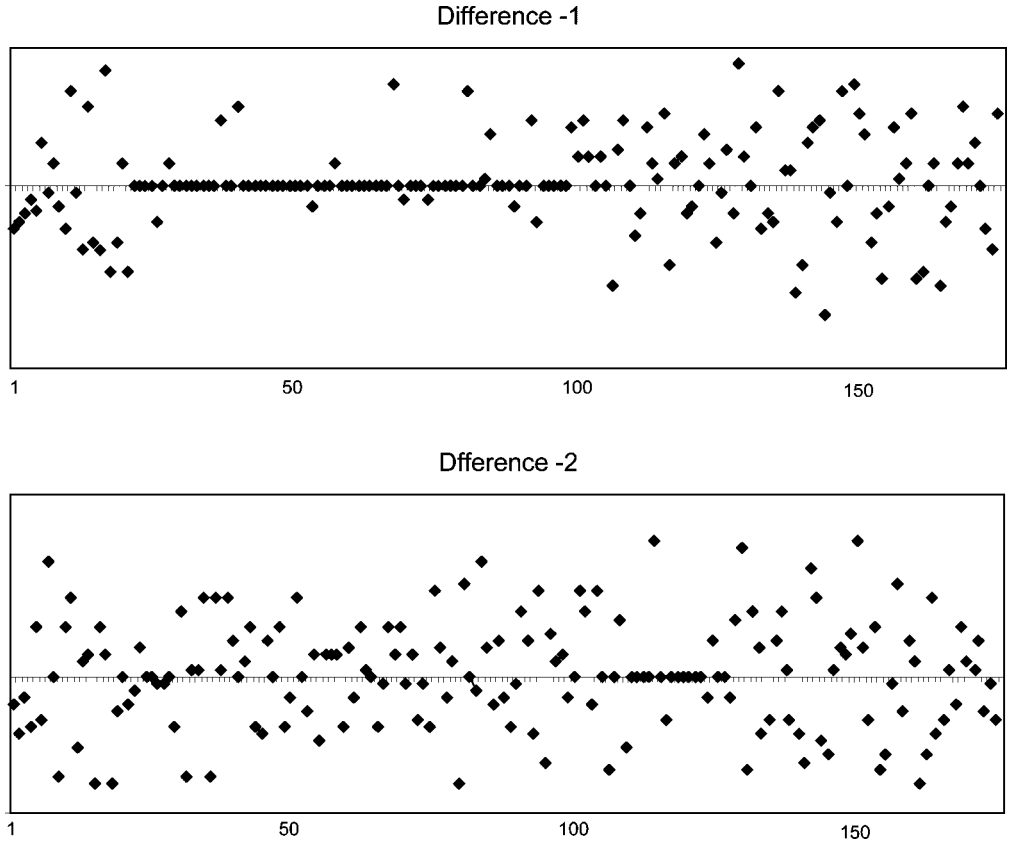


Figure 47. Difference in the radian coordinates of the corresponding amino acids of amino acids of carboxypeptidase Y from *Saccharomyces cerevisiae* (top) and amino acids of mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region also from *Saccharomyces cerevisiae* shifted to the left for one and two places, respectively.

Figure 47 represents the essence of a novel graphical approach to a protein-alignment problem. By combining the information obtained by considering the difference spectra of the four shifts of proteins 1 and protein 2 (only two of which have been illustrated in Figure 47), one can construct the alignment pattern for the two proteins, which was shown in Table 19. It is possible that there are additional local alignments of shorter lengths that could be detected if the search was continued, but for the outline of the novel graphical approach for protein alignment, this is not essential. As one can see, the four shifts of spectrum-like (20-line) representations of proteins achieved an overall matching in 117 sites out of 169.

We should emphasize the salient feature of the graphical alignment approach: the result was obtained without considering penalties for various gaps. The graphical alignment approach, just outlined, represents an alternative search route for alignments, which is conceptually and computationally simple and in this rudimentary form does not involve considerations of potential weighting schemes to expedite the search. But even at this early stage of its development, it is possible to conceive further improvements. For example, one can replace the alphabetical ordering of amino acids along the periphery of the unit circle by grouping amino acids of similar type, and even vary the spacing between amino acids that belong to different classes along the periphery on the unit circle. Such simple modifications will not affect the distributions of the "zeros" along the *x*-axis in the difference plots but will affect the nonzero amplitudes, and will make the differences (the heights of the "peaks") between amino acids belonging to the same class visibly smaller, which will point to sites with different amino acids but having similar properties. Thus, small "peaks" would offer additional information, a greater similarity among proteins than the apparent mismatch of amino acids may indicate. Some work along these lines is briefly described in the last section of this review. Finally, we should add that all graphical displays and all the computations discussed in this review can easily be performed with Microsoft Excel, which appears to be particularly suitable for such work.

Some readers may view this route to protein alignment as having limited potential, not competitive to the currently available computer packages such as BLAST[23,216] and FASTA.[24,217] That may be so now and for the immediate future, but the graphical approach to protein alignment has just emerged, whereas many of the available computer-graphic packages have been around for quite a while (20 years and more). It is likely that, with time, the graphical alignment approach will further evolve, but even now it offers supplemental and alternative viewing on the "classical" problem of protein alignment. Time will prove or disprove advantages and limitations of the graphical approach to the protein alignment outlined here; meanwhile, recall the question that was raised to Michael Faraday after his lecture on the eve of discovery of magnetic induction, mentioned in the Prologue. That discovery of Faraday was later exploited on a grand industrial-scale worldwide and culminated in the inventions of the genius of multiphase currents, Nikola Tesla.[218]

So in the same spirit, it is not fair, as some critics do,[219] to compare the potential of "grown-ups" with those of "newborn babies". Nevertheless, we should add that, in the case of a DNA alignment, described in ref 119, that followed the route outlined here for graphical alignments of proteins,[89]

not only could it successfully reproduce the computer-based result but it has also shown that in that case there are *better* solutions, obviously not detected by the particular computer program.

9. Nongraphical Representation of Proteins

The advantage of graphical representations is obvious; what is not so obvious is that nongraphical representations, such as the representation of DNA in 4-dimensional space,[151] or the representation of proteins in 20-dimensional space (vide infra), can lead to *pictorial* representations of the results, even though calculations were based on higher-dimensional representations. We have already at the very beginning of the review mentioned Hamori and Ruskin's construction of H-curves, which were initially defined for 5-dimensional space and then reduced to a 3-dimensional representation of DNA. We will illustrate in the next section one additional such case, in which one first extracts from an *N*-dimensional representation of a protein a set of invariants, and then one tries to illustrate the derived numerical quantities as a plot, as a bar graph, or by other visual modes. Thus, sometimes the distinction between graphical and nongraphical is more academic than pragmatic.

9.1. Representation of Proteins as Walks in Space

We have selected to illustrate this particular representation of proteins, in which proteins are viewed as "mathematical" objects in 20D space,[220] because this approach also bypassed the problem of selecting one among 20 factorial alternative orderings of amino acids. The starting point for this particular nongraphical representation of proteins is the binary representation of amino acids, which is illustrated in Table 20. To each of the 20 natural amino acids, which we have ordered alphabetically (based on their three-letter codes), we have assigned one unit vector for each of 20 possible directions along the coordinate axes of the 20-dimensional space. Clearly all 20 directions are fully equivalent, just as are the 3 directions of the 3D space taken along the *x*, *y*, and *z* coordinate axes. Every protein is then represented as a *walk* in the 20-dimensional vector space by moving away from the origin in the directions determined by the sequence of amino acids.

Table 20. Assignment of Unit Vectors for the 20 Directions in the 20D Space to the 20 Natural Amino Acids Ordered Alphabetically According to the Three-Letter Abbreviations

1	Ala	A	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
2	Arg	R	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
3	Asn	N	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
4	Asp	D	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
5	Cys	C	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
6	Gln	Q	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
7	Glu	E	0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
8	Gly	G	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
9	His	H	0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
10	Ile	I	0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
11	Leu	L	0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
12	Lys	K	0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
13	Met	M	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
14	Phe	F	0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
15	Pro	P	0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
16	Ser	S	0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
17	Thr	T	0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
18	Trp	W	0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
19	Tyr	Y	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
20	Val	V	0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
21	Leu	L	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

In Table 21, we have illustrated on the A chain of human insulin consisting of 21 amino acids the construction of the protein path

GIVEQCCTSICSLYQLENYCN

The process is fully analogous to the representation of DNA in 4-dimensional space in which the four unit vectors—1 0 0 0; 0 1 0 0; 0 0 1 0; and 0 0 0 1—have been used to represent the four bases A, C, G, and T.[151] The difference is only in the somewhat increased computational complexity arising from the presence of 20 amino acids in proteins in contrast to the presence of only 4 nucleotides in DNA sequences, which has thus increased considerably the dimensionality of the accompanying vector space.

Let us consider the initial steps in representing the A chain of human insulin: one starts with glycine (Gly) in the direction 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 and one moves in the direction of isoleucine (Ile) 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0, which brings one to point 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 of the 20D space (obtained by simply adding the corresponding unit vectors). In Table 21 are listed all 21 steps associated with the A chain of human insulin. For comparison, in Table 22 is shown the corresponding walk for humanin, a closely related new polypeptide compound having 24 amino acids, recently discovered in Japan.[221] Humanin appears to have promising use for protection against forms of Alzheimer's disease.

It is of interest to examine the columns of Table 21 more closely. The first column gives the running index of alphabetically ordered amino acids, and the second and third columns provide the three-letter and one-letter codes, respectively. In the next column are shown consecutive coordinates of the path in 20-dimensional space obtained by adding the unit vectors of each successive amino acid (listed in Table 20). The next column, which counts the cumulative frequency for each of the 20 amino acids as one moves from the left to the right along the protein sequence, which is in the center of the table, is also central for construction of the remaining three columns. Observe that, what is very important, this column can be constructed directly by using the information given by the protein sequence, *without* construction of the coordinates of the points of the path representing the protein in 20-dimensional space. The next column shows the increments in the squared distances of adjacent steps in the path, while the next column shows its cumulative sum. The last column is obtained by taking the square root of the preceding column.

Table 22 gives the same information on humanin as Table 21 on strand A of human insulin, except that we do not show the coordinates of amino acids in 20D space. They have been removed to emphasize that the construction of Table 22 can be obtained without recourse to the information on the protein path in the 20D space, and that any reference to 20D space is solely for the purpose of interpretation of the resulting numbers in the last column of such tables. The table can be easy to construct, as outlined above, from the information on the frequency of occurrence of individual amino acids, shown in the first numerical column of Table 22. Hence, there is no need to calculate or record the coordinates in 20D space, corresponding to individual steps of the path, and this simplifies and makes almost trivial the evaluation of the required distances (shown in the last column) that describe different proteins.

The numerical entries in the last column of Tables 21 and 22, for the A strand of insulin and humanin, respectively,

Table 21. Walk in 20D Space Corresponding to 21 Amino Acids of Strand A of Human Insulin, Also Showing the Squared Increments (inc.) between Adjacent Steps and the Distance of the Walk Measured from the Origin

	AA		coordinates	freq.	inc.	sum	dist.
1	Gly	G	0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0	1	1	1	1
2	Ile	I	0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0	1	1	2	1.41421
3	Val	V	1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0	1	1	3	1.73205
4	Glu	E	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0	1	1	4	2.00000
5	Gln	Q	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0	1	1	5	2.23607
6	Cys	C	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 0 0 0 0	1	1	6	2.44949
7	Cys	C	1 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 2 0 0 0 0	2	3	9	3.00000
8	Thr	T	1 0 0 1 0 0 0 0 0 0 0 1 0 1 1 1 2 0 0 0 0	1	1	10	3.16228
9	Ser	S	1 0 0 1 1 0 0 0 0 0 0 1 0 1 1 1 2 0 0 0 0	1	1	11	3.31662
10	Ile	I	1 0 0 1 1 0 0 0 0 0 2 0 1 1 1 2 0 0 0 0	2	3	14	3.74166
11	Cys	C	1 0 0 1 1 0 0 0 0 0 2 0 1 1 1 3 0 0 0 0	3	5	19	4.35890
12	Ser	S	1 0 0 1 2 0 0 0 0 0 2 0 2 1 1 3 0 0 0 0	2	3	22	4.69042
13	Leu	L	1 0 0 1 2 0 0 0 0 1 2 0 2 1 1 3 0 0 0 0	1	1	23	4.79583
14	Tyr	Y	1 1 0 1 2 0 0 0 0 1 2 0 2 1 1 3 0 0 0 0	1	1	24	4.89897
15	Gln	Q	1 1 0 1 2 0 0 0 0 1 2 0 2 1 2 3 0 0 0 0	2	3	27	5.19615
16	Leu	L	1 1 0 1 2 0 0 0 0 2 2 0 2 1 2 3 0 0 0 0	2	3	30	5.47723
17	Glu	E	1 1 0 1 2 0 0 0 0 2 2 0 2 2 2 3 0 0 0 0	2	3	33	5.74456
18	Asn	N	1 1 0 1 2 0 0 0 0 2 2 0 2 2 2 3 0 1 0 0	1	1	34	5.83095
19	Tyr	Y	1 2 0 1 2 0 0 0 0 2 2 0 2 2 2 3 0 1 0 0	2	3	37	6.08276
20	Cys	C	1 2 0 1 2 0 0 0 0 2 2 0 2 2 2 4 0 1 0 0	4	7	44	6.63325
21	Asn	N	1 2 0 1 2 0 0 0 0 2 2 0 2 2 2 4 0 2 0 0	2	3	47	6.85565

Table 22. Construction of the Distances of Vertices Forming the Path for Humanin in 20D Space from the Origin, Solely from the Information on the Cumulative Abundance of Individual Amino Acids in the Protein

AA	freq.	inc.	sum	distance
Met	1	1	1	1
Ala	1	1	2	1.41421
Pro	1	1	3	1.73205
Arg	1	1	4	2.00000
Gly	1	1	5	2.23607
Phe	1	1	6	2.44949
Ser	1	1	7	2.64575
Cys	1	1	8	2.82843
Leu	1	1	9	3.00000
Leu	2	3	12	3.46410
Leu	3	5	17	4.12311
Leu	4	7	24	4.89898
Thr	1	1	25	5.00000
Ser	2	3	28	5.29150
Glu	1	1	29	5.38516
Ile	1	1	30	5.47723
Asp	1	1	31	5.56776
Leu	5	9	40	6.32456
Pro	2	3	43	6.55744
Val	1	1	44	6.63325
Lys	1	1	45	6.70820
Arg	2	3	48	6.92820
Arg	3	5	53	7.28011
Ala	2	3	56	7.48331

are of a similar magnitude, which is a necessary but not sufficient condition that two protein sequences are similar. Observe that, for the first six steps, there are in both tables the same numerical entries in the last column, yet the first six amino acids of A strand of human insulin and humanin are different! To fully establish the degree of similarity between pairs of proteins, one would need to examine the actual path in 20D space, or alternatively, to consider additional protein invariants that for similar systems remain numerically similar but for less-similar systems may introduce detectable variations in relative magnitudes, which would not be difficult to observe.

9.2. Protein Descriptors

The columns of Table 21 suggest the construction of several protein descriptors. For instance, the column "increments" allows one to construct a novel matrix, called the "line-distance" matrix,[222−224] the eigenvalues of which provide additional descriptors for characterizing biological sequences. The line-distance matrix has an interesting property in that the entries in its first row completely define the matrix. A similar property, that the whole matrix is defined by the matrix elements of its first row, one finds also for the Topliz[225] and Hankel[226] matrices.[227−231] Given n points on a line, the line will be partitioned in $(n - 1)$ segments. Matrix elements of the first row of the line-distance matrix are defined by the distances between the first point of the line to the remaining points on a line. Matrix elements in the second row are defined by distances between the second point on the line to the remaining points on a line, which can be obtained by subtracting the difference of the first two points from the distances listed in the first row of the matrix, and so on.

For the construction of novel descriptors, instead of using the distances, one could use the squared distances, because then one works with integers when the points along the line (as is the case here) have integer coordinates. Alternatively, one can consider the "increments" in the path, which measure the departure of points on the path from the origin, and one considers their higher powers as outlined earlier. From these, one can construct the "profile" of a protein, which is given by normalized sequence increments. We will illustrate this construction for amino acids of strand A of human insulin. One starts with the summation of the entries of Table 21 in the column "increments":

$$1 + 1 + 1 + 1 + 1 + 1 + 3 + 1 + 1 + 3 + 5 + 3 + 1 + 1 + 3 + 3 + 3 + 1 + 3 + 7 + 3 = 47$$

In the next step, one considers the sum of the squares of the same entries:

$$1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 3^2 + 1^2 + 1^2 + 3^2 + 5^2 + 3^2 + 1^2 + 1^2 + 3^2 + 3^2 + 3^2 + 1^2 + 3^2 + 7^2 + 3^2 = 157$$

The process continues by constructing sums of the increasing higher powers:

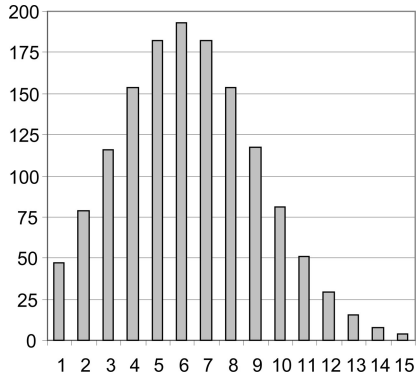


Figure 48. Profile of strand A of human insulin based on the entries of the line-distance matrix.

$$1^k + 1^k + 1^k + 1^k + 1^k + 1^k + 3^k + 1^k + 1^k + 3^k + 5^k + 3^k + 1^k + 1^k + 3^k + 3^k + 3^k + 1^k + 3^k + 7^k + 3^k$$

In view that the sums increase with the exponent k, one has to introduce a normalization factor $1/k!$, which will ensure the convergence of the process. Thus, instead of 157 for the second sum, one has 157/2, or 78.5. In this way for strand A of human insulin, one obtains the sequence:

47, 78.5, 115.8, 153.5, 182.4, 193.2, 182.4, 154.0, 117.0, 80.7, 50.8, 29.4, 15.8, 7.9, 3.7. This sequence, which is illustrated Figure 48 as a bar graph, is referred to as "protein profile". As one can see, it is of similar shape as the profile of A strand of human insulin illustrated earlier in Figure 42.

There is some loss of information that accompanies the construction of protein profiles, so it could happen that different proteins have the same profile. However, because the theoretical number of proteins having N amino acids is 20^N, which is so huge for already relatively small N, the occurrence of proteins having the same profiles is unlikely. If this happens, there are additional protein invariants that will differentiate such cases. What is more important is that the profiles may facilitate identification of similar proteins.

10. Most Recent Results

10.1. Amino Acid Adjacency-Count Matrix

One can associate with any protein a 20 × 20 matrix, the rows and columns of which belong to the 20 natural amino acids (AAs).[232,233] The matrix element (i, j) counts the number of times amino acid (i) is followed by amino acid (j) in the protein sequence when it is read from left to right. This novel matrix is nonsymmetrical, and whenever the same amino acid repeats itself in the sequence on adjacent sites, it increases the nonzero entries on the main diagonal. The row sums of the 20 × 20 AA matrix give the abundance of individual amino acids, except for the first amino acid, which is not included in the corresponding row. Similarly, the column sums give the abundance of amino acids except for the last amino acid, which is not included in the column sum. While the entries in the rows indicate the adjacencies when the protein sequence is read from left to right, the entries in the columns indicate the adjacencies when the protein sequence is read from right to left.

If one is interested in the construction of a symmetrical adjacency matrix for amino acids, this can be obtained by adding the adjacency matrix A with its transpose A^T. We will illustrate the construction of the AA adjacency matrix on a smaller protein having 30 AAs,

WTFESRNDPAKDPVILWLNGGPGCSSLTGL

belonging to yeast *Saccharomyces cerevisiae*, which already has been used to illustrate the protein sequence alignment.[89] The rows and columns of the 20 × 20 matrix have been labeled with one-letter amino acid codes, though the amino acids have been ordered alphabetically, based on their three-letter codes.

The AA adjacency matrix offers a novel graphical representations of proteins, which one obtains by connecting the adjacent sites in the AA adjacency matrix of the protein using the centers of the cells of the corresponding matrix elements (i, j). The resulting zigzag curve is illustrated in Figure 49. An alternative graphical representation of a protein can be obtained by using the unit circle of Figure 34, but instead of applying Jeffrey's construction of DNA by depicting AAs as spots within the "magic circle", as already illustrated in Figure 43, one can use the points representing amino acids, which are on the circle periphery (and thus have constant coordinates). The resulting zigzag curve is illustrated in Figure 50. There is loss of information in this kind of graphical representation because any repetition of the same

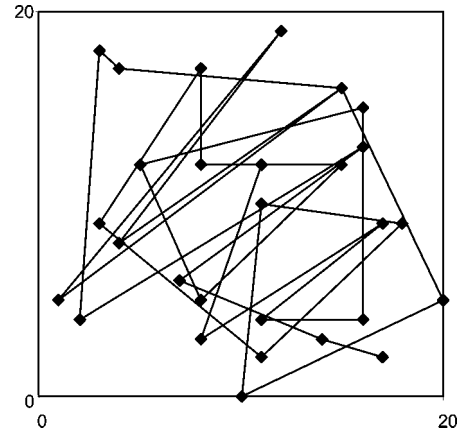


Figure 49. Graphical zigzag representation of a protein of *Saccharomyces cerevisiae* on the basis of the 20 × 20 AA matrix.

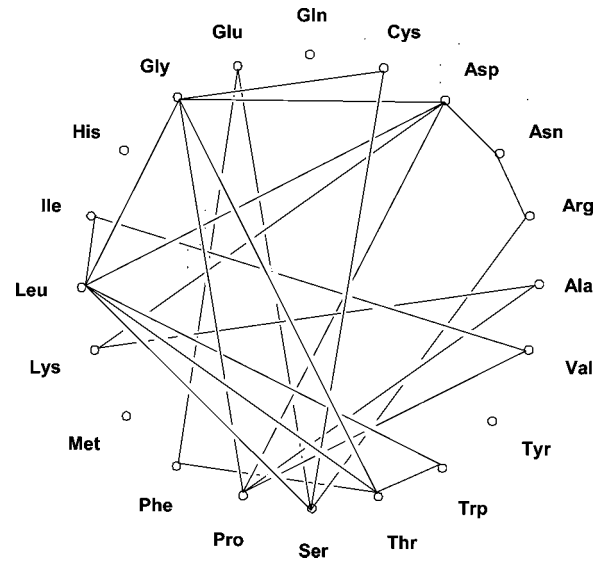


Figure 50. Graphical zigzag representation of a protein of *Saccharomyces cerevisiae* on the basis of connecting amino acids located at the periphery of the unit circle matrix.

Table 23. ND6 Proteins (Having between 168 and 175 AAs) for Six Additional Species (Groups of 10 Amino Acids Are Set Together for Easier Reading)

Opossum 168 AAs

MKMMTIYIIS LLLMIGFVAF ASKPSPIYGG LSLVVSGGLG CGMVVSLEDV
FLGLVVFLVY LGGMLVVFGY TTAMATEEYP ETWVGNVVAF IMLLFVLLLQ
VGWYFMSKLV YIIMAIKLFD FVETSLVGQD YNGVSQLYYC GGWALALLGW
ILFMTIYVVL EVVRERSY

Harbor seal 175 AAs

MMTYIVFILS IIFVVSFVGF SSKPSPIYGG LVLIISGAVG CGIVLSFGGS
FLGLMVFLIY LGGMLVVFGY TTAMAIEQYP EVWVSNKAVL GAFVMGLLSE
LLLACYILKD DEVDVVFEFN GMGDWVIYDT GDSGFFSEEA MGIAALYSYG
TWLVIVTGWS LLTGVLVIME VTRGN

Grey seal 175 AAs

MMTYIVFILS IIFVISFVGF SSKPSPIYGG LVLIISGAVG CGIVLSFGGS
FLGLMVFLIY LGGMLVVFGY TTAMATEQYP EVWVSNKAVL GAFVMGLLSE
LLLACYILKD DEVDAVFEFN GMGDWVIYDT GDSGFFSEEA MGIAALYSYG
TWLVIVTGWS LFIGVLVIME VTRGN

Rat 172 AAs

MTNYMFILSL LFLTGCLGLA LKPSPIYGGF GLIVSGCIGC LMVLGFGGSF
LGLMVFLIYL GGMLVVFGYT TAMATEEYPE TWGSNWFIFS FFVLGLFMEL
VVFYLFSLNN KVELVDFDSL GDWLMYEIDD VGVMLEGGIG VAAIYSCATW
MMVVAGWSLF AGIFIIIEIT RD

Mouse 172 AAs

MNNYIFVLSS LFLVGCLGLA LKPSPIYGGL GLIVSGFVGC LMVLGFGGSF
LGLMVFLIYL GGMLVVFGYT TAMATEEYPE TWGSNWLILG FLVLGVIMEV
FLICVLNYYD EVGVINLDGL GDWLMYEVDD VGVMLEGGIG VAAMYSCATW
MMVVAGWSLF AGIFIIIEIT RD

Chimpanzee 174 AAs

MTYALFLLSV SLVMGFVGFS SKPSPIYGGL VLIVSGVVGC AIILNYGGGY
MGLMVFLIYL GGMMVVFGYT TAMAIEEYPE AWGSGVEVLV SVLVGLAMEV
GLVLWVKGYD GMVVVVNFNS VGSWMIYEGE GPGLIREDPI GAGALYDYGR
WLVVVTGWTL FVGVYIVIEI ARGN

pairs of amino acids and any reversal of adjacent amino acids are not visibly recorded. However, all this information is recovered in the process of constructing the **D/D** matrix, the construction of which is simplified because one uses fixed coordinates for amino acids. The approach depicted in Figure 50 will be more suited for smaller proteins and for screening of similar proteins, which will result in similar graphical representations, regardless of the accompanying loss of information. Even though such graphical representations need not offer visual insights on protein sequences, the graphs depicting the *difference* between the coordinates of the corresponding AA in proteins are often likely to offer useful visual information on similarity among proteins.

The AA adjacency matrices can be used for detecting similarities among proteins, as well as for fast preliminary filtering of large numbers of protein sequence data. For an illustration, we will consider ND6 protein sequences of eight species, listed in Table 12 (for human and gorilla) and Table 23 (for opossum, harbor seal, gray seal, rat, mouse, and chimpanzee), all having lengths around 170 amino acids.[233] When considering similarity/dissimilarity among eight species, one is interested in locating, among the 28 possible pairs of proteins, one-half dozen pairs of proteins as the most similar. In general, when comparing large proteins, or large numbers of proteins, particularly if one is screening for very similar proteins, if we base such comparison on the AA adjacency matrices, it would be useful to have a computationally less-intensive procedure than to compare 400 matrix elements each time (of 20 × 20 matrices). If proteins are similar, then the corresponding fragments will be similar—and hence, one may therefore restrict attention to fragments of proteins at a fraction of the time required for the full analysis.

One possibility is to consider AA abundances, which can be obtained from the row or the column sums of the AA adjacency matrix. The resulting 20-component vectors (shown in Table 24 for the ND6 proteins of the eight species of Tables 12 and 23) offer such a simplified representation of proteins. In Table 25 is shown the similarity/dissimilarity between the eight species based on the 20-component abundance vectors viewed as points in 20D vector space and taking the Euclidean distance among the points as a measure of the degree of similarity. As one can see, just by confining attention to the abundance of amino acids in the ND6 protein,

Table 24. Vectors Indicating the Abundance of Amino Acids for the ND6 Protein of Eight Species

species	AA abundance
	A R N D C Q E G H I L K M F P S T W Y V
human 174	(8, 3, 4, 3, 1, 0, 10, 29, 0, 12, 17, 2, 10, 10, 5, 10, 3, 5, 11, 31)
opossum 168	(8, 2, 2, 3, 2, 3, 7, 18, 0, 11, 26, 4, 11, 11, 3, 10, 7, 4, 12, 24)
harbor seal 175	(9, 1, 3, 6, 2, 1, 8, 23, 0, 15, 21, 3, 9, 13, 3, 14, 8, 4, 9, 23)
gray seal 175	(10, 1, 3, 6, 2, 1, 8, 23, 0, 16, 20, 3, 9, 14, 3, 14, 8, 4, 9, 21)
rat 172	(8, 1, 4, 6, 4, 0, 8, 23, 0, 14, 25, 2, 11, 17, 3, 10, 8, 5, 8, 15)
mouse 172	(8, 1, 5, 6, 4, 0, 8, 26, 0, 14, 26, 1, 11, 11, 3, 8, 6, 5, 9, 20)
gorilla 174	(8, 3, 5, 3, 2, 0, 9, 30, 0, 12, 19, 2, 9, 7, 4, 9, 5, 5, 11, 31)
chimpanzee 174	(9, 3, 4, 3, 1, 0, 9, 29, 0, 12, 19, 2, 10, 7, 5, 10, 5, 5, 11, 31)

Table 25. Similarity/Dissimilarity of the Eight Species Based on 20-Component Vectors of Table 24 Representing AA Abundances in ND6 Proteins (The Smallest Entries Are Shown in Bold)

	human	opossum	h. Seal	g. Seal	rat	mouse	gorilla	chimp
human	0	25.7	29.4	30.4	26.0	25.7	**11.6**	**6.6**
opossum		0	23.2	25.1	24.1	22.0	25.9	26.4
h. Seal			0	**5.4**	18.7	20.6	27.0	27.6
g. Seal				0	19.1	21.3	27.8	29.0
rat					0	**13.2**	27.1	26.2
mouse						0	23.8	26.2
gorilla							0	**7.0**
chimp								0

a comparison of the eight proteins already gives a very satisfactory result. The five smallest entries in Table 25 are as follows:

harbor seal−gray seal, 5.4;
human−chimpanzee, 6.6;
gorilla−chimpanzee, 7.0;
human−gorilla, 11.6;
rat−mouse, 13.2.

The differences between other pairs, not listed above, are visibly larger. The above numerical values appear plausible, because even the degree of difference between these most similar pairs of species agree with the expectations: the two seal species are found to be the most similar species, while the two rodent species show the largest difference, differing in about 10% of DNA, in comparison with the difference in overall DNA of primates, which differ in 1% of DNA and are characterized by intermediate similarity.

Because proteins having the same or very similar abundance can still be very different, it is desirable to have additional simplified characterizations of proteins, to be available for fast comparative studies of a larger pool of proteins. Instead of using the 20 AA abundance values in proteins, one can consider fewer matrix elements of the 20 $\times$ 20 AA adjacency matrices instead of using the total of 400 matrix elements. Let us review a few such possibilities. For instance, one can take the 20 diagonal elements of AA adjacency-count matrices, expecting that similar proteins are likely to have similar diagonal elements, while dissimilar proteins are likely to have dissimilar diagonal elements. In Table 26 are listed the diagonal entries of the AA adjacency matrices for the eight species.

Inspection of this table, even without performing any calculations, already indicates qualitatively that again the two seal ND6 proteins are similar, the rat and mouse proteins are similar, and the proteins of human, gorilla, and chimpanzee are similar. However, by viewing the diagonal entries of AA adjacency-count matrices again as components of 20D vectors, one can calculate the degree of similarity between all 28 pairs using the Euclidean distance between the vectors in 20D space as the similarity measure. The results of such

Table 26. Vectors Indicating the Count of the Repetition of the Same Amino Acid in Protein AA Sequences

species	A R N D C Q E G H I L K M F P S T W Y V
human	(0, 0, 0, 0, 0, 0, 1, 4, 0, 1, 1, 0, 2, 0, 0, 0, 1, 0, 0, 8)
opossum	(0, 0, 0, 0, 0, 0, 1, 4, 0, 3, 6, 0, 1, 0, 0, 0, 1, 0, 0, 7)
harbor seal	(1, 0, 0, 1, 0, 0, 1, 4, 0, 2, 5, 0, 1, 1, 0, 1, 2, 0, 0, 3)
gray seal	(1, 0, 0, 1, 0, 0, 1, 3, 0, 2, 3, 0, 1, 1, 0, 1, 1, 0, 0, 1)
rat	(1, 0, 1, 1, 0, 0, 1, 4, 0, 2, 1, 0, 1, 1, 0, 0, 0, 0, 0, 3)
mouse	(1, 0, 1, 1, 0, 0, 1, 4, 0, 2, 0, 0, 1, 0, 0, 1, 0, 0, 1, 3)
gorilla	(0, 0, 1, 0, 0, 0, 0, 4, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 8)
chimpanzee	(0, 0, 0, 0, 0, 0, 1, 4, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 7)

Table 27. Similarity/Dissimilarity of Eight Species Based on 20-Component Vectors of Table 26 on the Basis of the Count of Repeating AAs in ND6 Proteins

	human	opossum	h. seal	g. seal	rat	mouse	gorilla	chimp
human	0	$\sqrt{31}$	$\sqrt{48}$	$\sqrt{60}$	$\sqrt{32}$	$\sqrt{34}$	$\sqrt{5}$	$\sqrt{3}$
opossum		0	$\sqrt{23}$	$\sqrt{51}$	$\sqrt{47}$	$\sqrt{59}$	$\sqrt{34}$	$\sqrt{30}$
h. seal			0	$\sqrt{10}$	$\sqrt{22}$	$\sqrt{32}$	$\sqrt{51}$	$\sqrt{37}$
g. seal				0	$\sqrt{12}$	$\sqrt{18}$	$\sqrt{61}$	$\sqrt{45}$
rat					0	$\sqrt{4}$	$\sqrt{31}$	$\sqrt{23}$
mouse						0	$\sqrt{31}$	$\sqrt{23}$
gorilla							0	$\sqrt{4}$
chimp								0

calculations are shown in Table 27, the entries of which fully confirm the expectations.

Observe that both Tables 25 and 27, when one focuses attention on the smallest entries, support one another. We used the values 13.3 and $\sqrt{10}$ as cutoffs for similarity classification in Tables 25 and 27, respectively. However, even if one would use the value $\sqrt{12}$ as the threshold for Table 27, it follows from Table 25 that gray seal and rat are not similar enough, because the corresponding value in Table 25 (19.1) is too large. It suffices to find a single large entry in similarity tables for the same species based on different descriptors, to exclude a pair of species as similar.

In Figure 51 are plotted the similarities obtained using 20 entries of Table 25 against the corresponding entries obtained when all 400 of the elements of the AA matrices are taken into account. Figure 51 visually confirms the assertion that the partial information on proteins may suffice to locate the most similar protein pairs. Observe in the left lower part of Figure 51 a group of five "spots" associated with the smallest entries of Table 25, which correspond to the following species:

5.4, harbor seal and gray seal
6.6, human and chimpanzee
7.0, chimpanzee and gorilla
11.6, human and gorilla
13.2, rat and mouse.

An additional reduced characterization of proteins follows when one considers the difference in the count of the AA adjacencies when the protein sequence is read from left to right and vice versa.[233] While the list of abundances has no

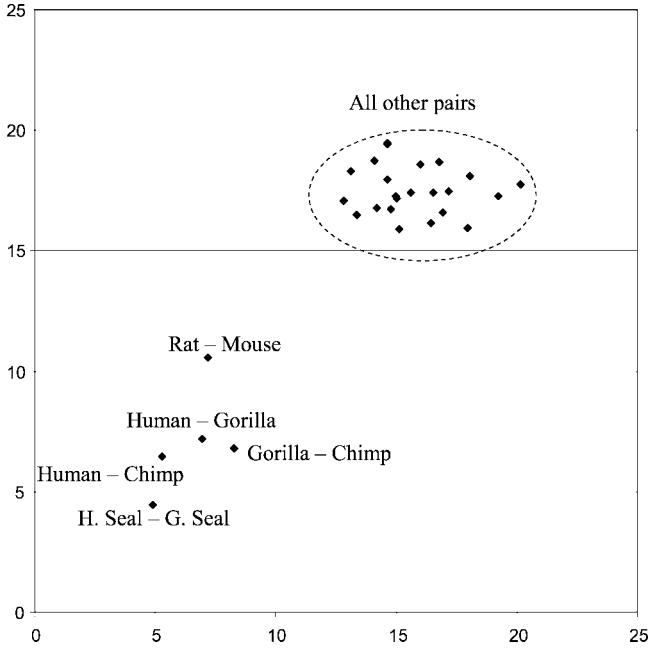


Figure 51. Plot of the similarity obtained by using 20 entries of Table 25 against the corresponding entries obtained when all 400 elements of the AA matrices are taken into account. Most similar pairs (shown at the low left corner of the figure) have been identified. The remaining pairs of low similarity have been encircled.

information on the *distribution* of AAs within a protein sequence, the difference in the adjacencies when the protein sequence is read in the opposite direction offers information on the *nearest neighbors* in the sequence. Information on the nearest amino acids has been occasionally presented in tabular form, or simply by listing all AA pairs, and has also been used in modeling protein folding.[234−236] Such lists merely serve as documentation, while by contrast, when the same entries are arranged as elements of a matrix, they form a novel mathematical object, which allows the same information to be considered in totality. Matrices satisfy the rules of Matrix Algebra (matrix transformation properties) and offer novel invariants such as eigenvalues, etc., which can be used for characterizing proteins.

Let us continue with illustrating the characterization of proteins by 20-component vectors, the components of which are constructed from absolute *differences* between the corresponding entries in rows and columns of the 20 × 20 AA adjacency matrix. By adding such differences for each row, one obtains 20 quantities, which offer an alternative novel simplified representation of proteins. For example, the first row and the first column of the matrix of Table 28 and their difference (written below as 20-component vectors) are:

Row 1: 0, 0, 0, 0, 1, 0, 1, 2, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0
Column 1: 0, 1, 0, 0, 0, 0, 0, 1, 0, 2, 2, 0, 2, 0, 0, 0, 0, 1, 0, 0
Difference: 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0

The absolute difference for the above two vectors adds to 10. When the same is done for the remaining 19 rows and columns of Table 28, one obtains the following 20 component vector

(10, 4, 5, 4, 2, 0, 8, 16, 0, 12, 12, 4, 9, 2, 6, 4, 6, 6, 8, 14)

which offers a novel sequence descriptor in proteins. Observe that, in this case, all but one component of the 20-component vector for the reduced representation of the protein are different from zero. This considerably decreases the chance that different proteins will have identical vectors. In Table 29, we have collected vectors for the eight species of Tables 12 and 23, the above vector being listed again in the last row of Table 29.

The construction of the novel 20-component vectors may at first sight appear somewhat artificial, but let us be reminded that in nature there is no preference for reading protein sequences from right to left or from left to right, though in the case of DNA the sequence direction is essential for identifying the open reading frame that encodes for protein coding. In Table 30 we show the similarity/dissimilarity matrix for the eight ND6 proteins.

Again, already a superficial glance at the table allows one to identify the five entries in the table that are visibly smaller than the rest of the entries. They correspond to the same five species listed before:

4.90, harbor seal−gray seal;
5.29, gorilla−chimpanzee;
6.92, human−chimpanzee;
7.21, rat−mouse;
8.25, human−gorilla.

The next smallest entry is considerably larger (being 12.21, which corresponds to the pair human−harbor seal), which

Table 28. AA Adjacency Matrix for The ND6 Protein of Chimpanzee (The Last Protein of Table 23 with 174 AAs); Rows Indicate the Left Amino Acid and Columns Indicate the Right Amino Acids for Each Pair

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	0	0	0	1	0	1	2	0	1	1	0	1	0	0	0	1	0	1	0
R	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1
D	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0
C	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	1	0	0	0	0	1	1	0	2	0	0	1	0	1	0	0	0	1	1
G	1	1	0	1	0	0	2	4	0	1	1	1	2	1	1	2	1	1	3	6
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	2	0	0	0	0	0	1	0	0	1	3	0	1	0	2	0	0	0	1	1
L	2	0	0	0	0	0	0	5	0	1	1	0	0	2	0	1	1	1	1	4
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
M	2	0	0	0	0	0	0	2	0	0	1	0	1	0	0	0	0	1	1	1
F	0	0	1	0	0	0	0	2	0	0	2	0	0	0	0	0	0	0	0	2
P	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
S	0	0	1	0	0	0	0	2	0	0	1	0	0	1	1	1	0	0	0	3
T	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1
W	1	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
Y	0	0	1	1	0	0	1	3	0	3	1	0	0	0	0	0	1	0	0	1
V	0	0	0	0	0	0	2	3	0	2	6	0	3	2	0	3	0	1	0	7

Table 29. 20-Component Vectors Based on Adjacency of AAs When Reading the Protein Sequences of Tables 12 and 23 in Opposite Directions

species	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
human	8	4	5	4	2	0	10	16	0	10	10	4	7	6	8	6	4	6	8	14
opossum	8	2	2	4	2	4	6	12	0	12	16	6	11	6	4	16	8	8	11	18
harbor seal	10	2	3	8	2	2	12	14	0	6	8	6	5	6	4	12	10	6	6	12
gray seal	10	2	3	10	2	2	12	12	0	8	6	6	5	6	4	12	8	6	6	14
rat	8	2	6	5	8	0	6	14	0	12	18	4	7	8	4	10	10	8	10	12
mouse	8	2	6	5	8	0	8	12	0	12	20	2	11	8	4	8	6	8	10	12
gorilla	8	4	3	4	2	0	8	16	0	12	12	4	9	2	6	4	6	6	6	10
chimpanzee	10	4	5	4	2	0	8	16	0	12	12	4	9	2	6	4	6	6	8	14

Table 30. Similarity/Dissimilarity among the Eight Species on the Basis of 20-Component Vectors That Show the Difference in Adjacencies When Protein Sequence Is Read from Left to Right and Right to Left; The Smallest Entries Are Shown in Bold

	human	opossum	h. seal	g. seal	rat	mouse	gorilla	chimp
human	0	16.79	12.81	13.11	14.63	15.03	**8.25**	**6.92**
opossum		0	16.91	17.15	13.34	14.63	18.06	16.55
h. seal			0	**4.90**	16.44	19.24	14.97	15.10
g. seal				0	17.94	20.15	16.00	15.62
rat					0	**7.21**	14.63	14.07
mouse						0	14.76	14.21
gorilla							0	**5.29**
chimp								0

clearly indicates a relatively distant relationship from the evolutionary point of view. As one sees, the novel 20-component vectors have passed the utility test with high marks, even though in this case rat−mouse appears as more similar than human−gorilla. Such differences should alert one that, though quantitative, the similarity/dissimilarity analysis should be viewed with some caution, and only if several different approaches give the same results can such results be taken as reliable.

10.2. Higher-Order AA Adjacency-Count Matrices

It is desirable to include additional information on proteins than merely count the occurrence of adjacent amino acids. If one is to include information on three successive amino acids, it leads to the construction of 400 × 400 matrices; inclusion of information on four successive amino acids leads to construction of 8 000 × 8 000 matrices; inclusion of information on five successive amino acids leads to construction of 160 000 × 160 000 matrices, with each step increasing the size of the matrices exponentially as 20^n. An alternative way is to consider the "higher-order AA matrices", the matrix elements of which count the occurrence of *k*th nearest-neighbor amino acid when reading the sequence from left to right. The case $k = 1$ corresponds to the already considered AA matrix that counts the occurrence of adjacency for all pairs of amino acids.

It is not difficult to construct the corresponding higher-order matrices, but it may suffice only to construct a selection of various 20-component vectors from such matrices, which will be used in comparative studies of proteins. For illustration, we listed below five 20-component vectors based on the higher-order matrices of order one to five, the elements of which are the differences in *k*th neighborhoods when reading the human ND6 protein from left to right and from right to left. The ordering of amino acids is again

A R N D C Q E G H I L K M F P S T W Y V

which is the alphabetic ordering of the 20 natural amino acids based on their three-letter codes: Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val.

First-order vector:

(8, 4, 5, 4, 2, 0, 10, 16, 0, 10, 10, 4, 7, 6, 8, 6, 4, 6, 8, 14)

Second-order vector:

(8, 6, 3, 6, 2, 0, 12, 19, 0, 16, 12, 2, 16, 16, 6, 8, 6, 8, 12, 22)

Third-order vector:

(12, 5, 7, 6, 2, 0, 14, 25, 0, 14, 8, 4, 12, 6, 10, 10, 6, 4, 7, 18)

Fourth-order vector:

(8, 3, 5, 4, 2, 0, 12, 25, 0, 14, 8, 2, 12, 14, 10, 14, 4, 6, 15, 22)

Fifth-order vector:

(6, 1, 7, 4, 2, 0, 10, 17, 0, 9, 9, 2, 10, 12, 2, 14, 6, 8, 13, 24)

All the above information can be extracted from five 20 × 20 matrices. However, the same information can also be extracted directly from the protein sequence, bypassing the construction of the corresponding matrices. By contrast, a matrix that would consider the totality of information on five consecutive amino acids would have over three million rows and columns.

10.3. Simplified Graphical Representations of Proteins

There are two possible routes to arrive at a graphical representation of proteins that would bypass the combinatorial complexities of 20! ways of choosing (ordering) apparently similar objects, such as the 20 natural amino acids (AAs) that constitute proteins: (i) use for the template mathematical objects or models that are not label-sensitive and (ii) use simplified protein models that do not discriminate between AAs that have similar properties. The former approach has taken advantage of the fact that labeling of vertices in graphs, which does not affect the calculation of graph invariants, is in general arbitrary, which has led to representation of proteins by starlike graphs.[70] Similarly, because all directions in *n*-dimensional space are equivalent, it is possible to represent proteins as a path in 20D space and thus to avoid considering any ordering of amino acids.[220]

The use of simplified protein models has been initiated by Dill and co-workers over 20 years ago:[237−239] the 20 natural amino acids are grouped in hydrophobic or nonpolar (H) and hydrophilic or polar (P). These models have been considered in the literature for quite some time.[240−246] Simplified protein models are of two kinds: (i) developing a general strategy based on graph-theoretical methods in order to find sequences of amino acids that fold into protein conformations of the desired native state and (ii) modeling simplified representations of proteins, elaborating on the HP model of Dill, by reducing the set of 20 natural amino acids to two representative classes that will probably show similar properties to the original set of amino acids.

In Figure 52, we are illustrating the lattice protein of Šali, Shahknovich, and Karplus[247] based on a 3 × 3 × 3 cubic

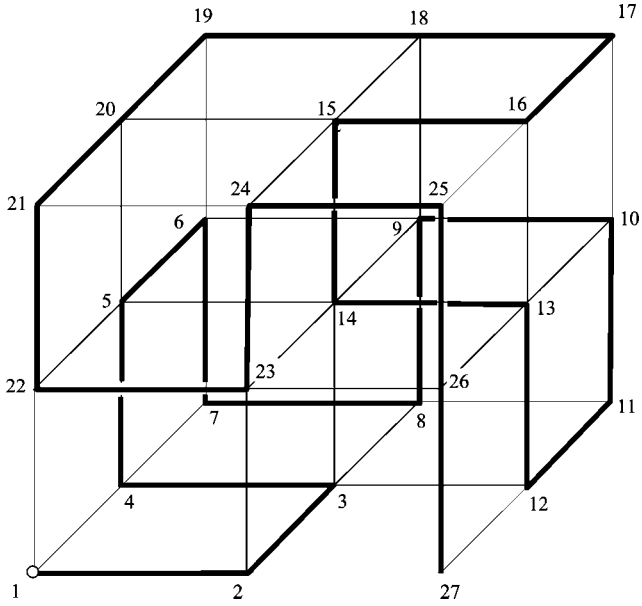


Figure 52. Lattice protein of Šali, Shahknovich, and Karplus shown as self-avoiding chain based on 3 × 3 × 3 cubic lattice.

lattice in which amino acids are placed inside a cube that depicts a protein as a compact, self-avoiding chain structure, a model that belongs to the first kind. For such lattice proteins, as outlined by Šali et al., one can calculate the total energy for the conformation, which is given as the sum of the contact energies B_{ij} between nonbonded adjacent amino acids on the lattice:

$$E = \sum_{i<j} \Delta(r_i, r_j) B_{ij}$$

The $\Delta(r_i, r_j)$ value is equal to one if amino acids are in contact (nonbonded but adjacent) and zero otherwise. In this model, two amino acids are in contact if they are not adjacent in the protein sequence and are at unit distance from each other in the lattice. It is interesting to observe that, when one assumes all B_{ij} to be equal (which corresponds to homopolymer rather than a heteropolymer model), the above model is transformed into a pure graph-theoretical model that would be equivalent to the HMO model of π-electron hydrocarbon calculations of Erich Hückel,[248−250] that is, the model of "hard ball" potential of Bloch[251] used for calculations of electron mobility in metals.[252]

We will continue with the examination of one particular less-simplified model in which 20 amino acids are grouped into five classes. In such a case, the graphical representation of proteins, instead of facing the problem of selection of one among 20 factorial possibilities, is reduced to a problem of manageable complexity. We will follow the grouping of amino acids into five classes by Riddle et al.[253] Their classification follows from the experimental explorations of searching for a subset of the natural amino acids that produces proteins consisting of β-sheets. Wang and Wang[254] studied the validity of the reduced protein models using statistical, thermodynamic, and kinetic characteristics and concluded that the reduced model captures the most essential features of proteins. The following are the five classes of AAs on which we will base the graphical representation of proteins:

Class 1: cysteine (C), isoleucine (I), leucine (L), methionine (M), phenylalanine (F), tryptophan (T), tyrosine (W), and valine (V);

Class 2: alanine (A), histidine (H), and threonine (T);

Class 3: glycine (G) and proline (P);

Class 4: aspartic acid (D) and glutamic acid (E);

Class 5: arginine (R), aspargine (N), glutamine (Q), lysine (K), and serine (S).

10.4. Five-Line Representations of Proteins

The classification of AAs into five groups allows one to transcribe the 20-letter protein sequence into a 5-letter sequence. The 5-letter protein sequences allow one to construct a simplified graphical representation of proteins based on *five* horizontal lines, which are analogous to "four-line" graphical models for DNA.[117] After replacing the 5 letters with the numbers 1−5, one can consider arithmetic manipulations of derived graphical representations of protein sequences,[255] similar to the manipulations of the spectrum-like four-line representations of DNA illustrated earlier.

We will illustrate the novel five-line spectrum-like representation of proteins on the eight ND6 proteins listed in Tables 12 and 23. In Table 31, we show the reduced protein sequences for human ND6 and opossum ND6, which represent a pair of rather dissimilar species. A direct comparison of the two protein sequences that have limited similarity is, to say the least, rather strenuous. On the other hand, when the same sequences are shown as the five-line graphical representations (Figure 53), the proteins can be visually examined, offering a more convenient view on the similarities and dissimilarities among them.

A more careful look at Figure 53 points to a considerable lack of similarity among the two proteins. It is hard to spot any regions, even a short one, that appear to have the same spectral pattern. The dissimilarity among the two proteins can be easily seen when one plots the difference between the two spectra, because spots that lie above the zero line indicate disagreements in amino acid types, while spots on the zero line show the sites at which two proteins agree. In Figure 54, we show graphically the difference of ND6 proteins of human and opossum, from which one can immediately see by visual inspection the lack of parallelism between the two proteins. From Figure 54, one can locate

Table 31. Simplified Human ND6 and Opossum ND6 Proteins (For Better Visibility, Each 10 Amino Acids Are Separated)

		Human ND6, length 174 AAs		
BBBOBBBBZB	UBBBUBBUBZ	ZZUZUBBUUB	BBBBZUBBUB	BBBBZBUUUB
BUBBBBBBBB	UUBBBBBUBO	OOBOXXBUX	OBUZUBXBBB	ZBBBUBOBXB
UBBBBBZXBX	UBBBBBZBZZ	BUZBBBBXUX	UZUBBZXXUB	UOUOBBXBUZ
BBBBBOUBUB	BBUBBBBBXB	OZUZ		
		Opossum ND6, length 168 AAs		
BZBBOBBBBZ	BBBBBUBBOB	OZZUZUBBUU	BZBBBZUUBU	BUBBBZBXXB
BBUBBBBBBB	BUUBBBBBUB	OOOBOOXXBU	XOBBUZBBOB	BBBBBBBBBZ
BUBBBBZZBB	BBBBOBZBBX	BBXOZBBUZX	BZUBZZBBBB	UUBOBOBBUB
BBBBOBBBBB	XBBZXZB			

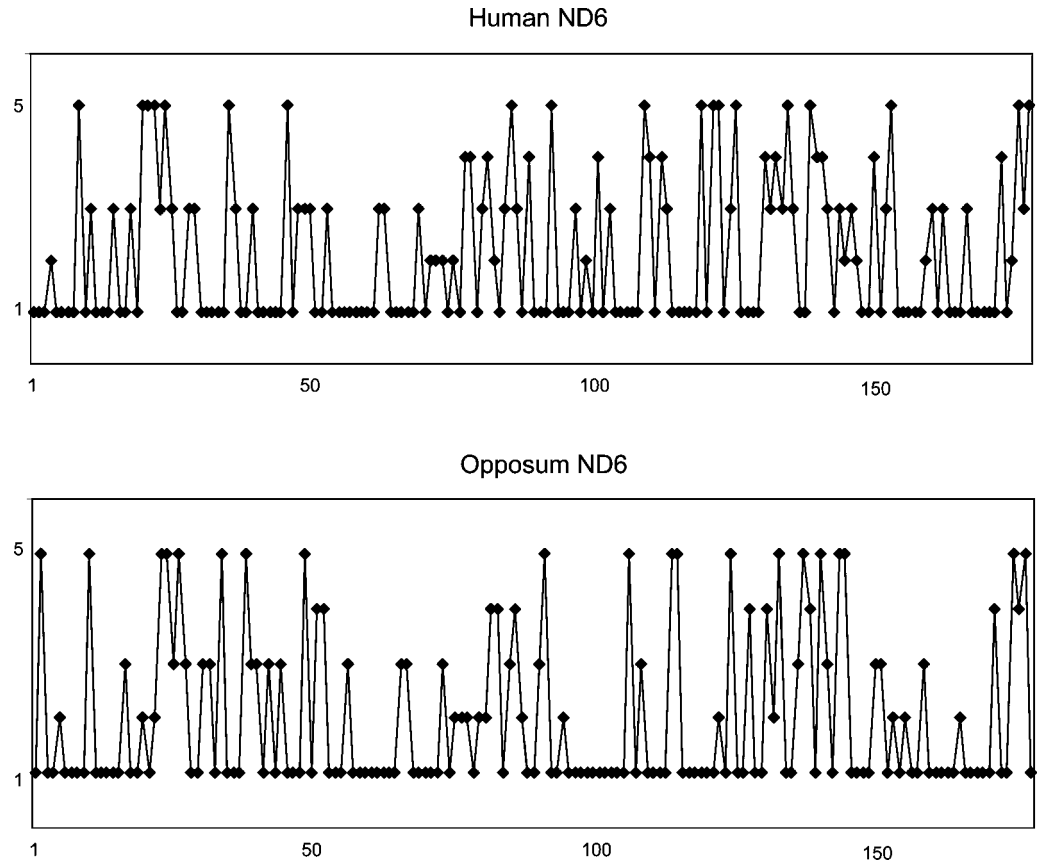


Figure 53. Five-line representation of ND6 proteins of human and opossum, two dissimilar species.

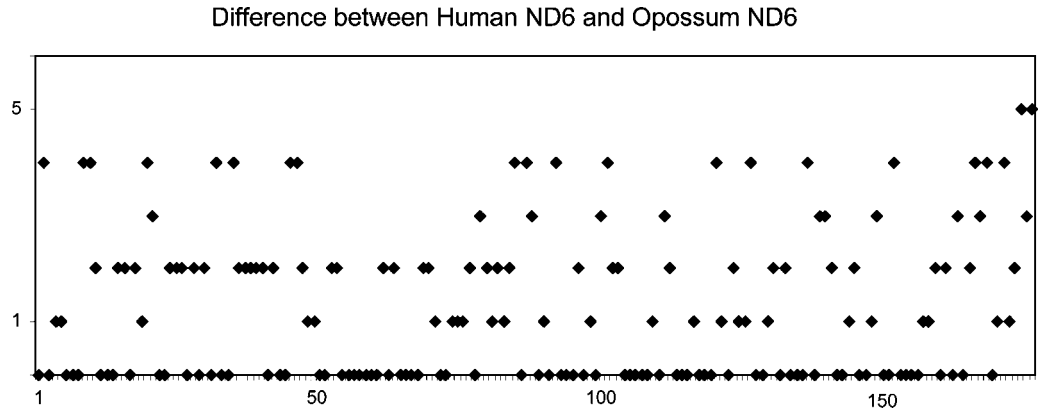


Figure 54. Graphical illustration of the difference in ND6 proteins showing considerable dissimilarities: human−opossum.

sites at which the AA types of the two proteins agree—the sites at which the difference plot gives zero amplitudes, 77 places out of 174, which is too few.

In contrast in Figure 55, we show the five-line spectra for ND6 proteins of gorilla and chimpanzee, which clearly shows great similarity. Observe that, except for the initial part of the spectra, almost the whole remaining region is almost identical. Thus, it is much easier to spot the similarities in protein sequences when they abound, such as is the case with ND6 proteins of gorilla and chimpanzee, than when they are limited, as was the case of ND6 proteins of human and opossum.

However, even in such situations, the five-line graphical representation of proteins may facilitate comparisons. In Figure 56, we show the difference plots for the ND6 proteins of primates, all of which are highly similar, showing between five and seven different AA types. The same high degree of similarity is expected and is found between the two species of seals, and between rat and mouse, as illustrated in Figure 57. It is interesting that again one finds rat and mouse, although highly similar, showing a somewhat lesser degree of mutual similarity. Figures 54, 56, and 57 illustrate the important advantages of graphical representations of proteins based on the five-line spectra, which appear to be particularly useful when one is comparing species that are relatively close on the Tree of Life (the graphical representation of evolution).[256]

There have been additional and alternative simplified models for protein, such as, for example, use of a six-letter model[257,258] and of a variable number of letters.[259] A study for the validity of possibly simplified representations has also received attention.[260,261]

Before outlining another additional graphical representation of proteins, let us clarify the rationale for continuing the development of graphical and numerical approaches to DNA, RNA, and proteins. Dozens of graphical algorithms

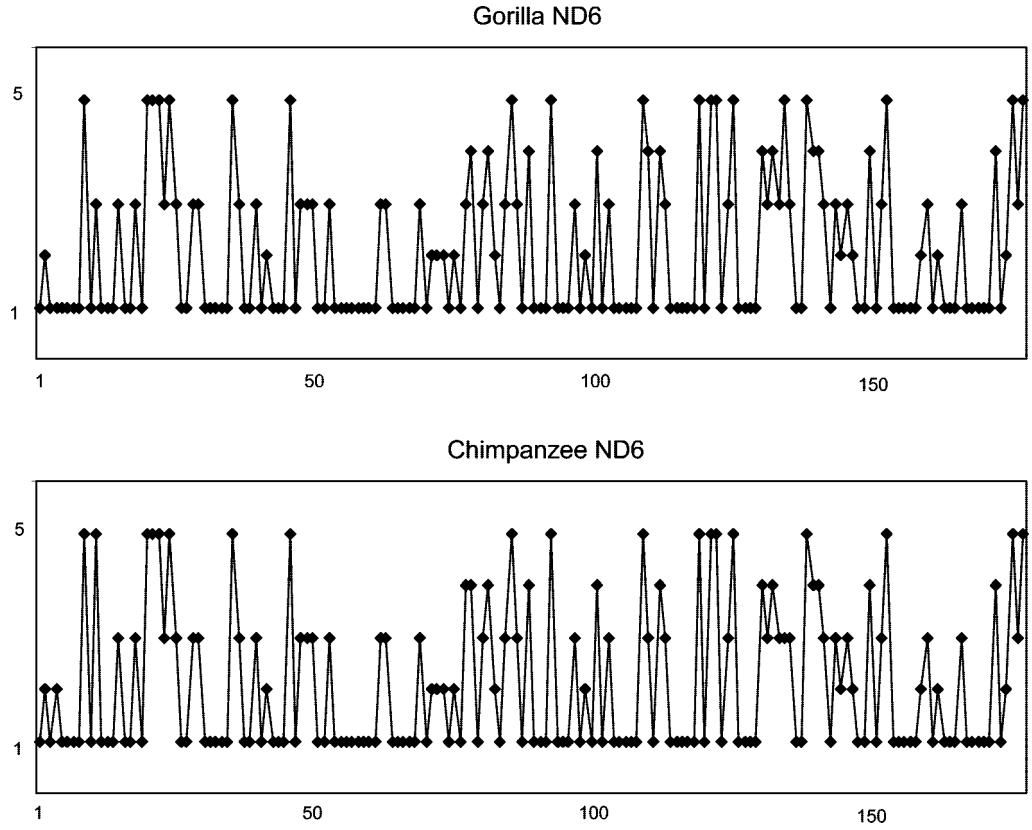


Figure 55. Five-line representation of ND6 proteins of gorilla and chimpanzee, two similar species.

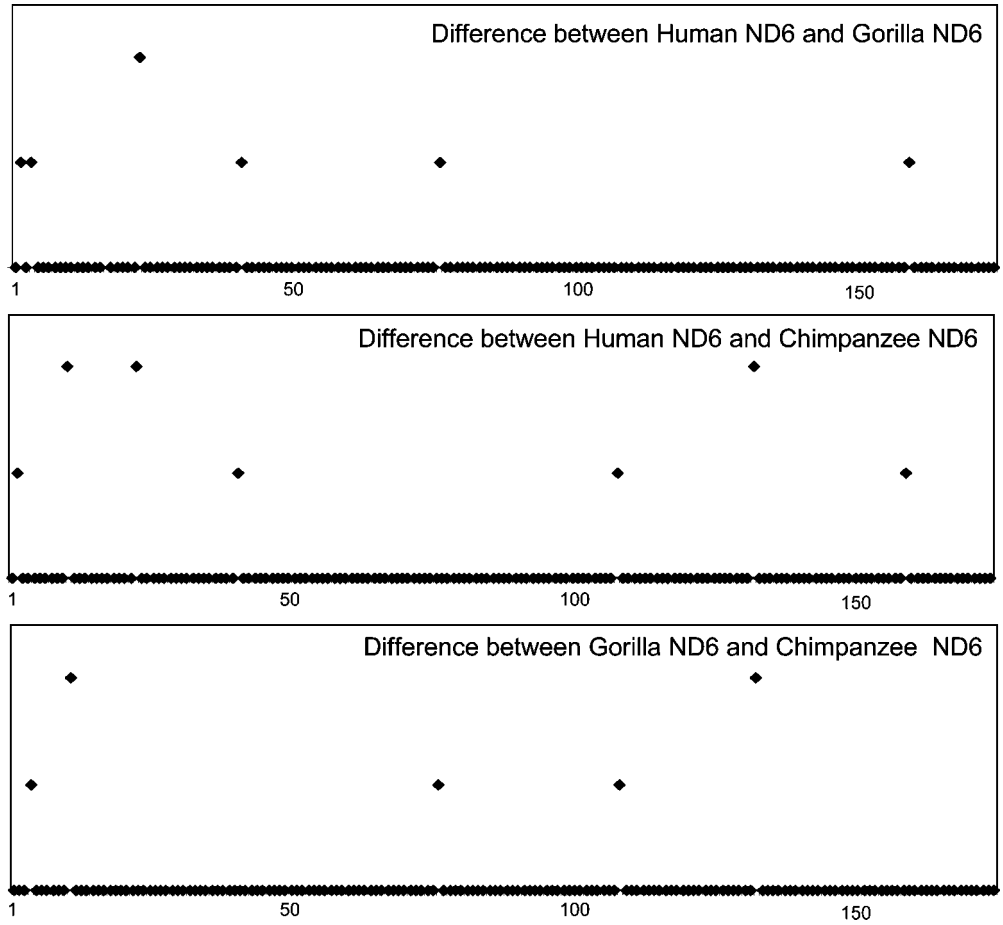


Figure 56. Differences in ND6 proteins of primates.

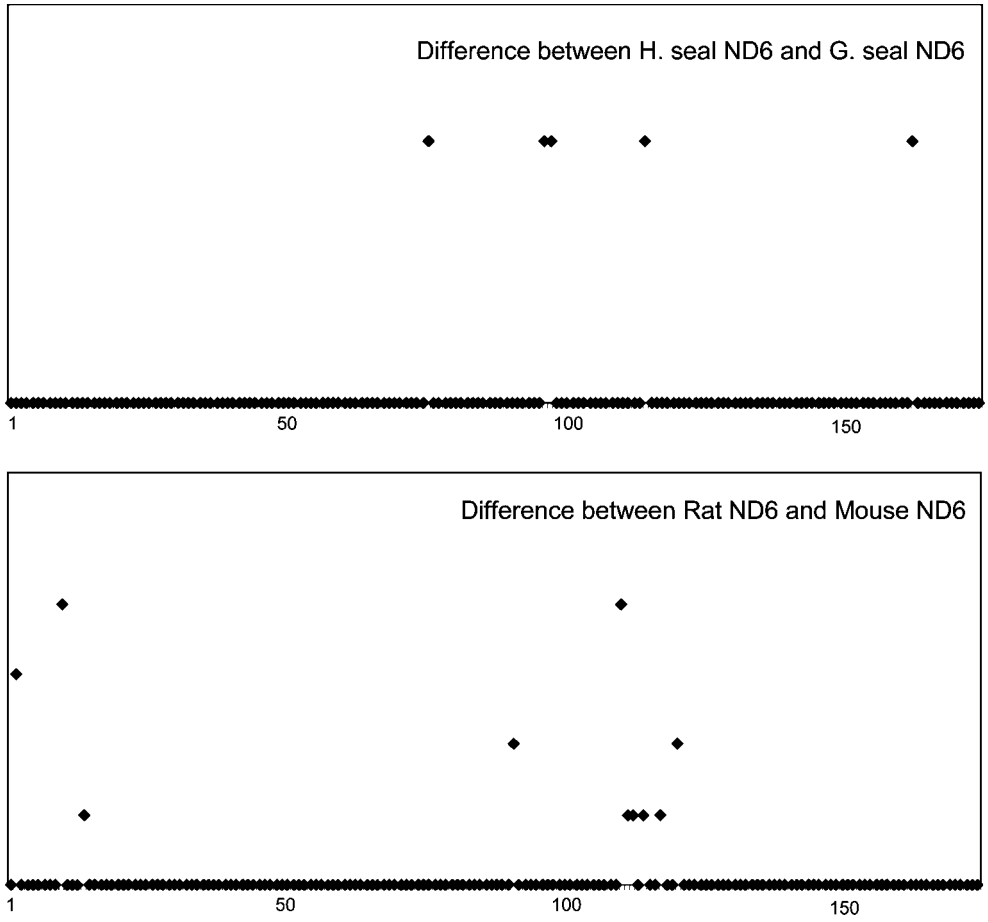


Figure 57. Differences between proteins of similar species: the two seals (top) and the two rodents (bottom).

have emerged, particularly for DNA, during the past 10 years, which to some may appear as a proliferation of theoretical methods, a kind of "academic exercise", which is remote from biology.[262] We feel this to be a mischaracterization, despite the fact that many of the past applications have focused on the same sets of DNA and may, therefore, have limited biological content. However, the purpose of such work is not to solve pending problems of biology but to *develop tools for solving* pending problems in biology! Incorporation of meaningful biological content to bioinformatics remains on the shoulders of biologists, while the task of applying Discrete Mathematics to complex systems aims to supply biologists with a suitable, effective, and adequate computational machinery for solving problems of biology.

10.5. Amino Acid Maps

It stands to reason that ordering amino acids based on their physicochemical properties may offer better insights in comparative studies of proteins than representations of proteins based on alphabetical ordering of amino acids, which is essentially equivalent to random ordering. The problem with the implementation of such schemes is that many reported properties of amino acids vary from laboratory to laboratory, a fact which introduces difficulties when selecting such data. For example, there are at least four different scales for hydrophobicity of the 20 amino acids reported in the literature.[263−266] Which one to choose? What to do when two amino acids are reported to have the same hydrophobicity as occurs in the work of Kyte and Doolitle?[265] In their scale, the hydrophobicity values vary from 0 (for glutamine) to 2.65 (for tryptophan), but two pairs of amino acids happen to have the same hydrophobicity: serine and threonine (0.05), and alanine and histidine (0.61). The presence of duplicate values does not allow a unique ordering of the 20 natural amino acids. In addition, in general, when dealing with experimentally measured quantities, one may question the significance of minor differences of the reported hydrophobicity values.

Table 32. Novel Strategy for Construction of Graphical Representation and Numerical Characterization for Protein Sequences

physicochemical properties
↓
two complementary properties
↓
partial order
↓
amino acids map
↓
matrix representations
↓
matrix invariants

In view of such problems, an alternative strategy is presented as a chart in Table 32. One starts with a pool of physicochemical properties of amino acids and selects *two complementary* properties as the basis for construction of a graphical representation of proteins. Here, "complementary properties" are taken conceptually, which means structurally or functionally related properties, not merely a pair of randomly selected properties. Such would be properties like molecular surface and molecular volume, or solubility in different media. In the next stage, one orders the 20 AAs with respect to the selected pair of properties, which leads to a *partial ordering* diagram. This diagram can be viewed

Table 33. Acid Dissociation Constants of Side Chains of Amino Acids

amino acid	3-letter code	1-letter code	$pK_a(COO^-)$	$pK_a(NH_3^+)$
		small hydrophilic		
glycine	Gly	G	2.34	9.60
alanine	Ala	A	2.34	9.69
threonine	Thr	T	2.63	10.43
serine	Ser	S	2.21	9.15
proline	Pro	P	1.99	10.60
		small hydrophobic		
valine	Val	V	2.32	9.62
leucine	Leu	L	2.36	9.60
isoleucine	Ile	I	2.36	9.68
methionine	Met	M	2.28	9.21
		aromatic		
phenylalanine	Phe	F	1.83	9.13
tyrosine	Tyr	Y	2.20	9.11
tryptophan	Trp	W	2.38	9.39
		acids and their amides		
aspartic acid	Asp	D	2.09	9.82
glutamic acid	Glu	E	2.19	9.67
aspargine	Asp	N	2.02	8.80
glutamine	Gln	Q	2.17	9.13
		bases		
lysine	Lys	K	2.18	8.95
arginine	Arg	R	2.17	9.04
histidine	His	H	1.82	9.17
		sulfyldryl		
cysteine	Cys	C	1.71	10.78

as a *2D map of amino acids*, to serve as a template for graphical representation of proteins. Such 2D maps are analogous to the 2D map for amino acids based on the 8 × 8 Table of Codons,[166] except that now the (x, y) coordinates of the 20 natural amino acids reflect some physicochemical properties of amino acids, rather than being based solely on theoretical models (such as are 2D protein maps based on modifications of Jeffrey's magic square[83] or the magic circle[169]). An important computational advantage of such an approach over 2D maps based on the magic square or the magic circle is that individual AAs of a protein always have the same pair of (x, y) coordinates and, hence, need not be repeatedly calculated, as must be the case with representations based on the magic square and the magic circle.

After finding locations for the 20 amino acids on the (x, y) plane, proteins are graphically represented by zigzag curves over the 2D map of amino acids. As already mentioned or implied several times, zigzag curves allow the construction of structural matrices, such as **D/D** matrix[58] and line-distance matrix.[222−224]

The two complementary properties of amino acids selected to illustrate this particular novel graphical representation of proteins are the pK_a values ($pK_a = -\log_{10} K_a$) for the terminal amino acid groups $-NH_3^+$ and $-COO^-$. They determine the activity of enzymes and are therefore of major importance in biochemistry. In proteins at physiological pH, amino acid and carboxy groups are charged (NH_3^+ and COO^-), but pK_a values are listed for protonated groups (NH_3^+ and COOH). The adopted pK_a values listed in Table 33 are taken from the internet.[267] In Figure 58, we show the 2D map for amino acids resulting from ordering the amino acids along the x-axis with respect to pK_a(COOH) and along the y-axis with respect to $pK_a(NH_3^+)$. The numerical values in the last two columns of Table 33 are taken as the x and y coordinates of the 2D map of Figure 58.

To illustrate the approach, we will again consider the two short segments of yeast *Saccharomyces cerevisiae* protein,

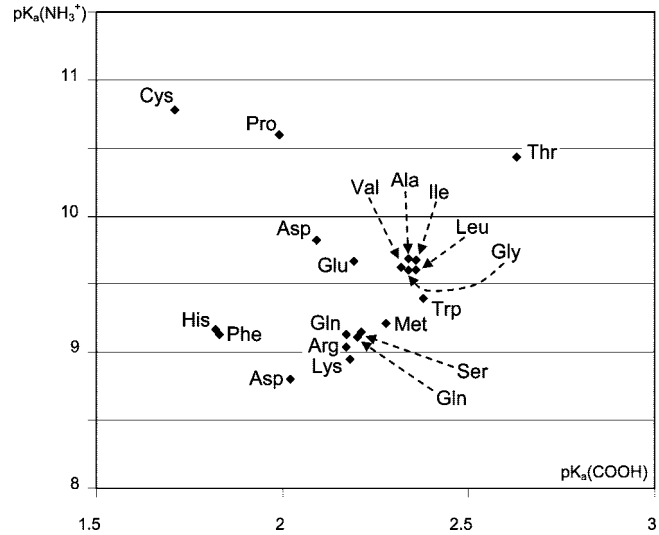


Figure 58. Plot of the pK_a values for the NH_3^+ end of proteins against the corresponding values for the COO^- end of proteins.

taken from the *Handbook of Chemoinformatics*.[215] In Figure 59, we have illustrated for two segments of the protein the zigzag curves, obtained by connecting adjacent amino acids using their sites (listed in Table 34) sequentially. The corresponding proteins are as follows:

Protein I: **WTFESRNDPAKDPVILWLNGGPGCSS-LTGL**

Protein II: **WFFESRNDPANDPIILWLNGGPGCSS-FTGL**

Figure 59 appears uninformative, just as has been the case with Figures 35, 49, and 50. Their main purpose of such figures with zigzag lines is to allow construction of **D/D** matrices, the invariants of which could serve as protein descriptors.

When the *differences* between the coordinates (listed in the last column of Table 34) of the corresponding amino acids of the two proteins are illustrated (Figure 60), one immediately observes the degree of alignment between the two proteins, which have different amino acids only at sites 2, 11, 14, and 27. Figure 60 can be compared with the similar Figure 36, which illustrates the alignment of the same two protein fragments using the variable (x, y), which provides nonconstant coordinates for individual amino acids. There, just as here, the spectral "peaks" show mismatches, but the matches of amino acids were there shown as points along the slopes following the peaks. Here, the matches of amino acids are shown as points along the x-axis.

Moreover, in Figure 60, the amplitudes of the mismatches have some meaning and are not differences of random numbers, as was the case with the "peaks" of Figure 36. Thus, the peak at site 14, where valine in protein I is substituted by isoleucine in protein II, is a mismatch, but because it has a small amplitude, it relates to two amino acids of the same class. Indeed, both valine and isoleucine have been classified as "small hydrophobic" amino acids. Therefore their replacement need not cause significant differences in protein properties, as would be the case when amino acids belonging to different classes are swapped. Hence, when such representations are implemented on graphical alignments of proteins that are based on ordering of amino acids with respect to some of their physicochemical properties, this approach may offer additional insights, indicating that there may be a difference between different

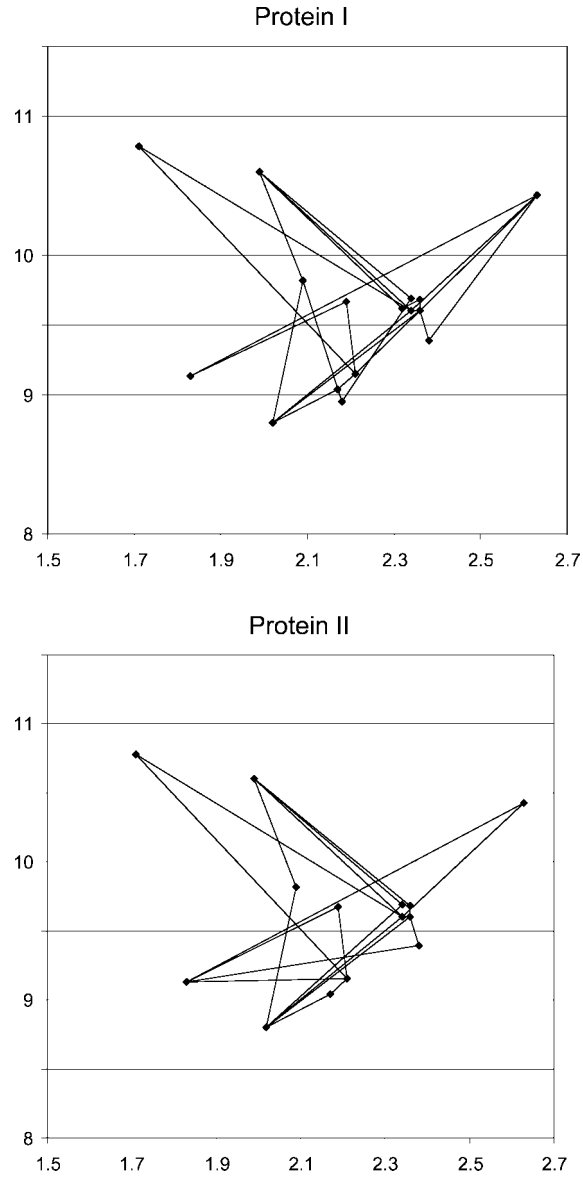


Figure 59. Zigzag representation of the two protein fragments of yeast *Saccharomyces cerevisiae*.

Table 34. Coordinates of Graphical Representations of Proteins I and II and Their Differences (Shown for the Initial 15 Amino Acids)

protein I	protein II	*x*	*y*	*x*	*y*	difference
W	W	2.38	9.39	2.38	9.39	0
T	F	2.63	10.43	1.83	9.13	1.5264
F	F	1.83	9.13	1.83	9.13	0
E	E	2.19	9.67	2.19	9.67	0
S	S	2.21	9.15	2.21	9.15	0
R	R	2.17	9.04	2.17	9.04	0
N	N	2.02	8.8	2.02	8.8	0
D	D	2.09	9.82	2.09	9.82	0
P	P	1.99	10.6	1.99	10.6	0
A	A	2.34	9.69	2.34	9.69	0
K	N	2.18	8.95	2.02	8.8	0.2193
D	D	2.09	9.82	2.09	9.82	0
P	P	1.99	10.6	1.99	10.6	0
V	I	2.32	9.62	2.36	9.68	0.0721
I	I	2.36	9.68	2.36	9.68	0

mismatches. Mismatches that are depicted with small amplitudes will be associated with swapping of amino acids that belong to the same class, and such exchanges are less likely to change drastically the properties of proteins. In conclusion, proteins could be more similar than the mere count of mismatches would suggest. This appears to be an advantage of 2D amino acid maps based on physicochemical properties of amino acids as the basis for graphical representations of proteins as compared with corresponding maps based on arbitrary (including alphabetical) ordering of amino acids.

10.6. Lattice Representations of DNA

The graphical representation of DNA was initiated over 25 years ago by Hamori and Ruskin[42] and, as outlined in the introductory part of this review, was followed by contributions of a dozen researchers in the period of mid-1995 to mid-2005.[48−57,86−95] Most graphical representations of DNA can be grouped into four types:

(1) Path representations;
(2) Map representations;
(3) Spectral representations;
(4) Lattice representations.

As we have seen, one obtains a path representation of DNA by assigning to successive nucleotide bases A, C, G, and T coordinates (x, y) or (x, y, z) in a Cartesian system, which when connected represents a zigzag curve in 2D or 3D space, respectively, which to a degree overlaps itself.[48−51] In map representations of DNA, individual nucleotide bases are represented either by disjoint points or by fused adjacent cells belonging to the same bases.[154,155] Spectral representations of DNA are obtained by assigning to the four bases integers 1−4, which transform four-letter sequences into numerical sequences displaying the pattern of spectra.[86,87,117] Lattice representations may be viewed as a special case of path representations in which the (x, y) and (x, y, z) coordinates are limited to integer values associated with coordinate grids.[48−50] As we have seen from Figure 2, which depicted the 2D graphical representation of the first exon of human, lemur, and opossum β-globin gene, such representations are accompanied with loss of information because individual steps can be repetitive and the curve can have cyclic parts with loss of direction on continuing the walk representing DNA. When such a path was represented by a **D/D** matrix, the lost information was recovered,[58] but it is also possible to construct lattice representations of DNA that fully determine the DNA sequence.

Graphical representations of DNA of Nandy and others,[48−50] besides offering visually useful representations, have an additional quality: they have integer coordinates (x, y) for all nucleotides. This simplifies computational manipulations with such representations and allows one to refer to them as "lattice representations". However, let us mention first that representations of DNA by Gates, Nandy, and Leong and Morgenthaler are not the only possible lattice representations of DNA over the Cartesian coordinate system. In Figure 61, we illustrate an alternative lattice representation of the first exon of human β-globin gene, depicted earlier in Figure 1, which is based on assigning the following four vectors to the four nucleotides:

- A → (+1, −2)
- T → (+2, −1)
- G → (+2, +1)
- C → (+1, +2)

Again when plotting DNA one obtains a zigzag curve in a plane, but the new plot has an important advantage over its couterpart in Figure 2. The zigzag curve of Figure 61 is not accompanied with any loss of information because it

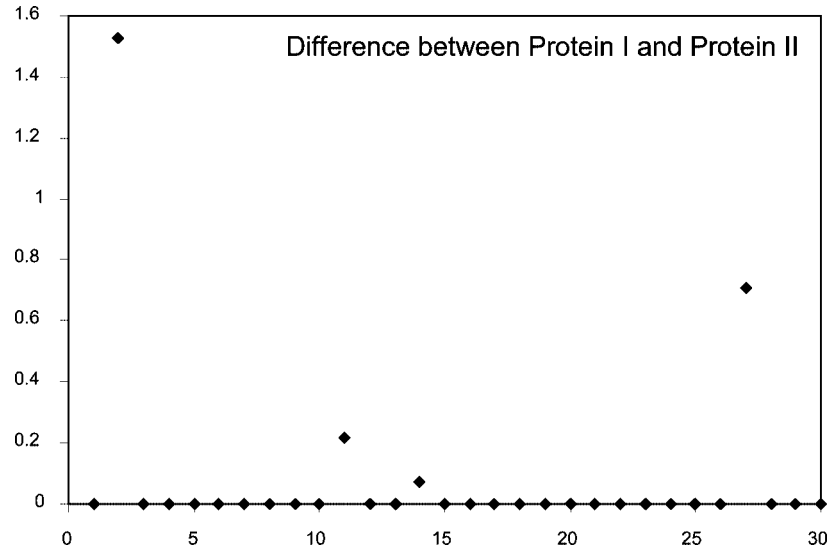


Figure 60. Differences in (x, y) Coordinates of the Two Proteins of Figure 42 (*Saccharomyces cerevisiae*).

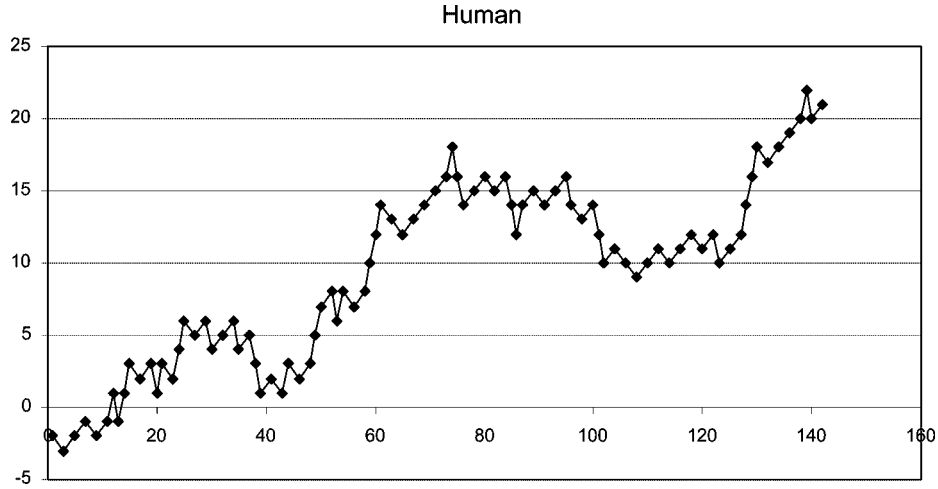


Figure 61. Novel lattice representation of DNA with no loss of information; the first exon of human β-globin gene using vectors A → (+1, −2,); T → (+2, −1); G → (+2, +1); C → (+1, +2).

represents DNA as a walk, rather than a path. The distinction is that walk representations, in which one knows which points follow which, allow one to reconstruct the DNA sequence. This was possible because each of the four nucleotide bases has a positive x-component, and thus, the path cannot retract, make cyclic routes, or overlap itself.

10.7. Lattice Representations of Proteins

Alternative types of DNA representations possess specific advantages. For example, path representations, and by extension lattice representations, which are special cases of path representations, offer plots that can be visually inspected. This is an important property of graphical representations of biological sequences. By contrast, spectral representations allow one to search graphically for sequence alignments by considering spectral differences for sequences that have been shifted relative to one another. Map representations, and in particular those based on chaos game algorithms, offer very compact representations of DNA and proteins. Thus, we see that it is desirable to have diverse representations of DNA, and it would be equally desirable to have diverse graphical representations of proteins.

Condensed zigzag line representations, spectral representations, and map representations of proteins have received some attention in the literature. In addition, proteins have been graphically represented as starlike graphs.[70] However, no lattice representation of proteins has been considered, if one excludes the representation of proteins as paths in 20D space,[220] which does not allow a simple graphical visualization. What is needed is a user-friendly 2D graphical representation of proteins analogous to the 2D graphical representations of Nandy on the Cartesian lattice[50] shown in Figure 2, but with as little self-crossing as possible, which would offer visual inspection of proteins and, at the same time, would assign to all amino acids integer coordinates. Even though graphical representations of proteins are of relatively recent time, they also have been shown to be of different types. However, the path-like and lattice-like 2D graphical representations of proteins, which would be, from the visual point of view, the most attractive, have been conspicuously missing until now. It is needless to elaborate on the desirability for such 2D representations of proteins, which would facilitate detecting proteins that are more similar and thus will increase the efficiency in comparative studies of proteins by visually identifying collections of less similar proteins, which may not deserve further attention in particular studies.

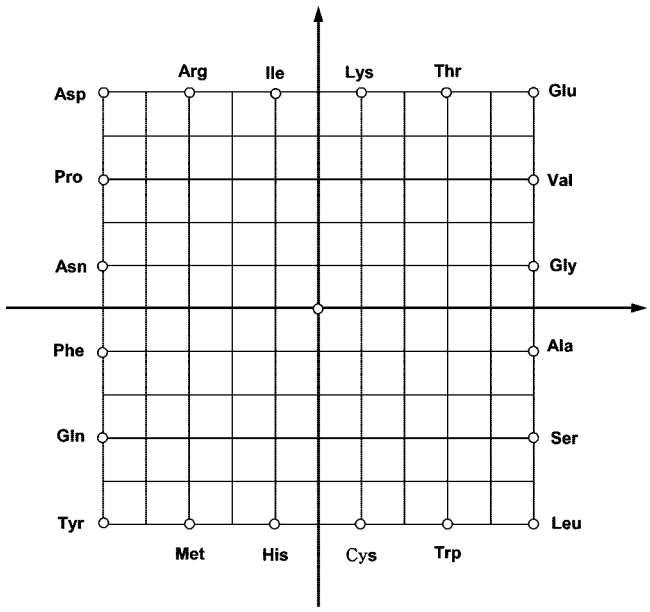


Figure 62. Twenty amino acids placed anticlockwise in decreasing order of abundance along the periphery of a 10 × 10 square starting from leucine at the bottom right corner.

We have started this review with user-friendly graphical representations of DNA illustrated in Figure 2 and will end with one such lattice-like user-friendly graphical representation of proteins reported most recently.

The reason for the delay of visually friendly graphical representations of proteins is that, in contrast to DNA, where there are only four directions that lead to graphical representation of DNA, in the case of proteins there are 20 directions to be selected to represent 20 amino acids. There are two further "complications" with graphical representations of proteins as compared to DNA, which may expain why it took 20 years for the emergence of graphical representations of proteins after the publication of the first graphical reperesentation of DNA. These "complications" concern: (1) how to select 20 directions or 20 alternative geometrical elements to serve as a template for graphical representations of proteins and (2) how to order the 20 amino acids, once the geometrical basis of the graphical representation for the problem has been selected.

A way to arrive at 20 points with integer coordinates in (x, y) is illustrated in Figure 62. One starts with a 10 × 10 square centered in the origin of the Cartesian coordinate system and considers 20 points on its periphery to which are assigned the 20 natural amino acids. The ordering of amino acids illustrated in Figure 62 is based on relative abundances of amino acids in proteins (listed in Table 35). It starts with leucine, the most abundant amino acid, which is placed at the bottom right corner of the square. Leucine is followed by serine, the second most abundant amino acid, and continues anticlockwise ending with tryptophan, the least abundant natural amino acid in proteins. In the last column of Table 35 are listed the corresponding Cartesian coordinates for the 20 natural amino acids. This particular ordering ensures that intersections of the path with itself are considerably reduced because amino acids at the right side and the top side of the 10 × 10 square, which in longer proteins will be more numerous than amino acids at the left and the bottom edge of the square, will pull the graph depicting proteins to the right and up and thus avoid most of the crossings.

Table 35. Twenty Amino Acids Placed Anticlockwise along the Periphery of a 10 × 10 Square Starting from Leucine in Decreasing Order of Relative Abundance

amino acid	3-letter code	1-letter code	relative abundance	(x, y)
leucine	Leu	L	9.36	(+5, −5)
serine	Ser	S	7.38	(+5, −3)
alanine	Ala	A	7.34	(+5, −1)
glycine	Gly	G	6.89	(+5, +1)
valine	Val	V	6.48	(+5, +3)
glutamic acid	Glu	E	6.22	(+5, +5)
threonine	Thr	T	5.85	(+3, +5)
lysine	Lys	K	5.81	(+1, +5)
isoleucine	Iso	I	5.76	(−1, +5)
arginine	Arg	R	5.20	(−3, +5)
aspartic acid	Asp	D	5.12	(−5, +5)
proline	Pro	P	5.00	(−5, +3)
aspargine	Asn	N	4.57	(−5, +1)
phenylalanine	Phe	F	4.12	(−5, −1)
glutamine	Gln	Q	3.96	(−5, −3)
tyrosine	Tyr	Y	3.25	(−5, −5)
methionine	Met	M	2.32	(−3, −5)
histidine	His	H	2.26	(−1, −5)
cysteine	Cys	C	1.76	(+1, −5)
tryptophan	Trp	W	1.34	(+3, −5)

In Figure 63 is illustrated the lattice-like graphical representation of human ND6 protein, whose (x, y) coordinates of initial amino acids are listed in the middle part of Table 36. At the right-hand side of Table 36 are listed cumulative coordinates obtained by adding coordinates of all preceding amino acids, which when plotted produce the graphical representation of ND6 human protein shown in Figure 63. Observe a relatively compact 2D graphical representation of the protein with rather limited self-crossing of the zigzag curve in Figure 63. Just as was the case with DNA, self-crossing of the zigzag curve can be completely eliminated by assigning to all amino acids positive x-coordinates, which would result in less compact representations. We have illustrated in Figure 64 one such construction, which is based on the amino acid coordinates listed in Table 37. We should also add that if one is interested in a section of the zigzag curve, one can plot it at greater resolution, as has been illustrated in Figure 65 for amino acids 50−280.

Before showing an illustration of similar user-friendly lattice representations of other proteins of Tables 12 and 23, let us briefly comment on the numerical characterization of zigzag curves illustrated in Figures 64 and 65. Of course, one can always construct the complete **D/D** matrices for such curves, but distance matrices are dense and computations with large numbers of such matrices take a long time, making them not suitable for fast screening of data. Often all fine details of zigzag curves need not be so important for a comparative study as is an overall view of such curves. We illustrate in Figure 66 the zigzag curve of Figure 63, but displaying locations of only every 4th (top), every 8th (middle), and every 16th (bottom) amino acid. Observe how these reductions, including even the very drastic reduction shown at the bottom of Figure 66, still carry enough of the visible information on the overall shape of the zigzag curve of Figure 63. This suggests that reduced matrices accompanying reduced curves of Figure 66 may have sufficient information for comparative study of such curves, which would then speed up preliminary examination of such data.

In Figure 67, we have collected novel user-friendly graphical representations for the eight ND6 proteins con-

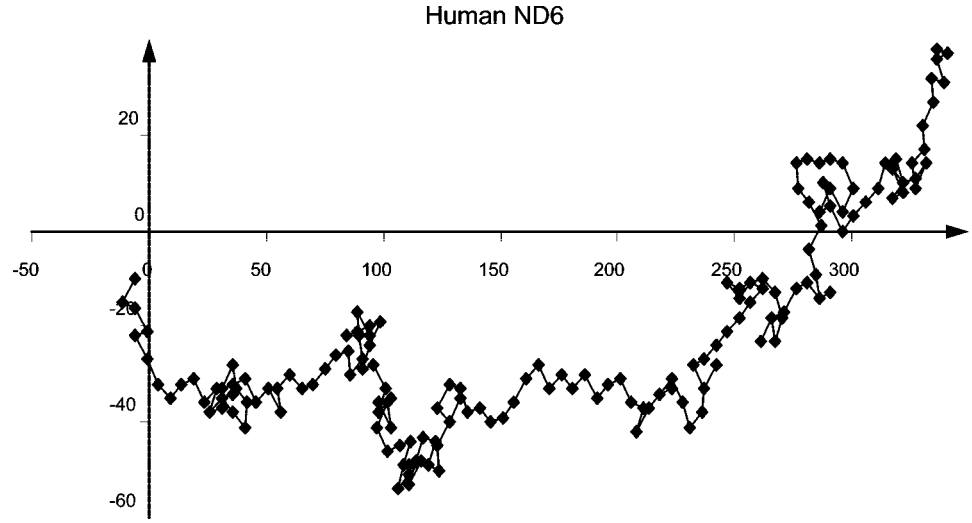


Figure 63. Illustration of human ND6 protein as a lattice graph.

Table 36. Initial First 20 Steps for the Lattice Representation of Human ND6 Protein of Figure 63

amino acid	(*x*, *y*)	cumulative
M Met	(−3, −5)	(−3, −5)
M Met	(−3, −5)	(−6, −10)
Y Tyr	(−5, −5)	(−11, −15)
A Ala	(+5, −1)	(−6, −16)
L Leu	(+5, −5)	(−1, −21)
F Phe	(−5, −1)	(−6, −22)
L Leu	(+5, −5)	(−1, −27)
L Leu	(+5, −5)	(+4, −30)
S Ser	(+5, −3)	(+9, −33)
V Val	(+5, +3)	(+14, −30)
G Gly	(+5, +1)	(+19, −29)
L Leu	(+5, −5)	(+24, −34)
V Val	(+5, +3)	(+29, −27)
M Met	(−3, −5)	(+26, −32)
G Gly	(+5, +1)	(+31, −31)
F Phe	(−5, −1)	(+26, −32)
V Val	(+5, +3)	(+31, −29)
G Gly	(+5, +1)	(+36, −28)
F Phe	(−5, −1)	(+31, −29)
S Ser	(+5, −3)	(+36, −32)

sidered for illustration of various aspects of graphical representations of proteins. Just a glance at this figure immediately points to the advantages of such graphical representations of proteins. It is evident from Figure 67 that ND6 proteins of all primates (human, gorilla, and chmpanzee) are very similar, and even more so for the ND6 proteins of the two seals, whereas graphical displays of ND6 for

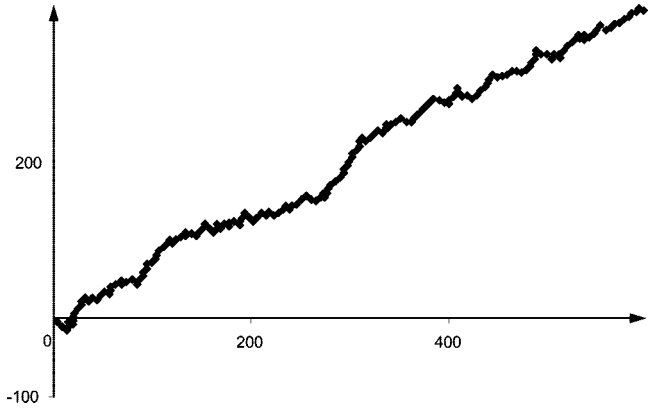


Figure 64. Illustration of the same ND6 protein as a lattice graph but using the *x*, *y* coordinates listed in Table 37.

Table 37. 20 Amino Acids Placed on 20 Points of the Positive Half-Plane Starting from Leucine with Coordinates (0, 5) and Moving Clockwise in Decreasing Order of Relative Abundance

3-letter code	1-letter code	(*x*, *y*)
Leu	L	(0, +5)
Ser	S	(+1, +5)
Ala	A	(+2, +5)
Gly	G	(+3, +5)
Val	V	(+4, +5)
Glu	E	(+5, +5)
Thr	T	(+5, +4)
Lys	K	(+5, +3)
Iso	I	(+5, +2)
Arg	R	(+5, +1)
Asp	D	(+5, 0)
Pro	P	(+5, −1)
Asn	N	(+5, −2)
Phe	F	(+5, −3)
Gln	Q	(+5, −4)
Tyr	Y	(+5, −5)
Met	M	(+4, −5)
His	H	(+3, −5)
Cys	C	(+2, −5)
Trp	W	(+1, −5)

mouse and rat, while showing considerable similarities, also show some differences, mostly in the initial region (1−50) and the middle section. Opossum ND6 is visibly different from other ND6 proteins considered.

11. Simplified Representations

11.1. Protein 20 × 20 Distance Matrix

One possible mathematical characterization of the lattice representation of proteins that has just been outlined is to use reduced matrices, which are based on graphical representations of reduced lattice graphs. Such a characterization is clearly of an approximate character. However, there is an alternative approach based on information of all amino acids in proteins that can be compacted into a 20 × 20 distance matrix for all proteins regardless of their size. In order to arrive at the 20 × 20 distance matrix, one considers each of the 20 amino acids present in a protein separately and calculates for each amino acid the corresponding average (*x*, *y*) coordinates. For example, methionine appears 10 times in human ND6 protein, at positions 1, 2, 14, 51, 63, 64, 73, 98, and 125, the (*x*, *y*) coordinates of which are (−3, −5),

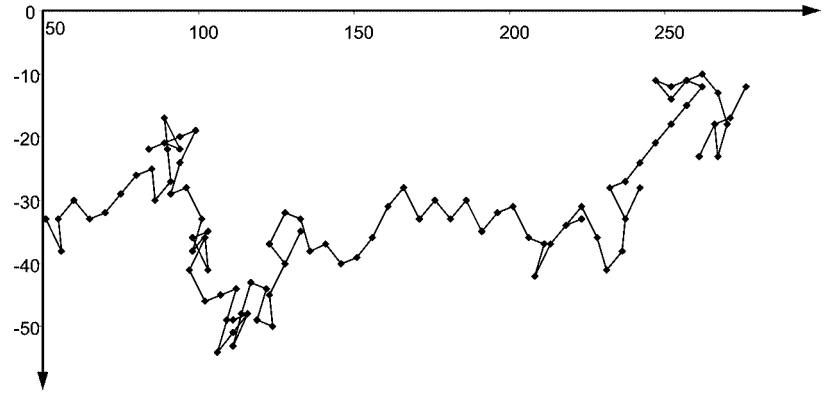


Figure 65. Detail of the lattice representation of human ND6 protein (amino acids 31−130).

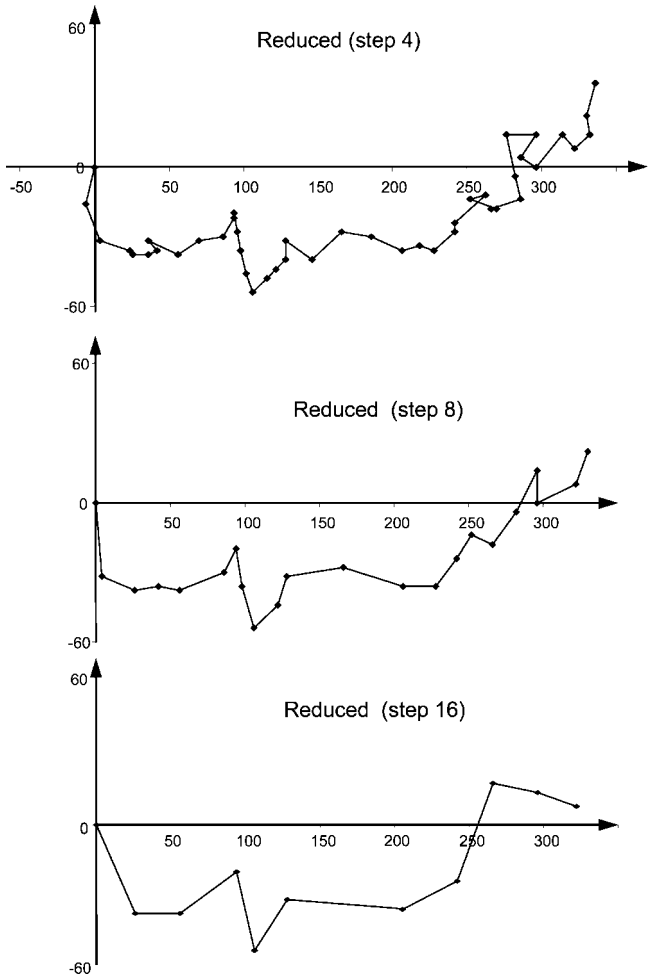


Figure 66. Simplified zigzag curve of Figure 63 showing every 4th, 8th, and 16th amino acid only.

(−6, −10), (26, −38), (91, −29), (98, −38), (109, −49), (106, −54), (119, −49), (208, −42), and (267, −23), respectively.

From these, it is not difficult to find that the average methionine coordinates are (101.5, −33.7). When the same is repeated for the remaining amino acids, one obtains for the average coordinates for the 18 amino acids present in human ND6 protein the values listed in Table 38. These average (x, y) coordinates of the 18 amino acids of human DN6 protein are depicted in Figure 68. In Figure 69, we have added as the 19th point the center of all amino acids, which is at (185.7348, −16.9803), and have connected all amino acids to the central point, thus converting the small map into a small graph.

Table 38. Average Cartesian Coordinates for Amino Acids of Human ND6 Protein

amino acid	x	y
A	197.71	−8.57
R	309.00	16.00
N	232.25	−1.25
D	268.33	−4.33
C	86.00	−30.00
E	269.83	−11.17
G	166.00	−19.79
I	192.67	−10.50
L	131.94	−26.76
K	153.50	−32.50
M	101.50	−33.70
F	141.70	−25.50
P	158.20	−16.20
S	132.50	−29.70
T	181.67	−25.67
W	278.50	−11.00
Y	157.10	−24.90
V	184.82	−10.11

Observe that the representations of ND6 proteins illustrated in Figures 68 and 69 are *geometrical* rather than *graphical*, because all points have fixed (x, y) coordinates. Because of this, we can immediately construct an 18 × 18 distance matrix for the protein considered, which is missing two amino acids: glutamine and histidine. In the general case in this way, one will arrive at a 20 × 20 distance matrix, which for the case of human ND6 proteins is shown in Table 39 with the rows and columns belonging to glutamine and histidine shown as "empty" (∅).

The significance of the novel protein distance matrix is that, regardless of the protein size, this approach results in uniform-size matrices for all proteins. This is the first time that such uniform matrices have been constructed for proteins−a novelty that is bound to not only have useful applications in comparative studies of proteins but also may play an auxiliary role in protein nomenclature, as illustrated in the following section.

12. Protein Nomenclature

12.1. Uniform Auxiliary Nomenclature for Proteins

The condensed graphical representation of ND6 protein in Figure 69 offers a novel approach to protein nomenclature. In Table 40 are listed the average coordinates for amino acids

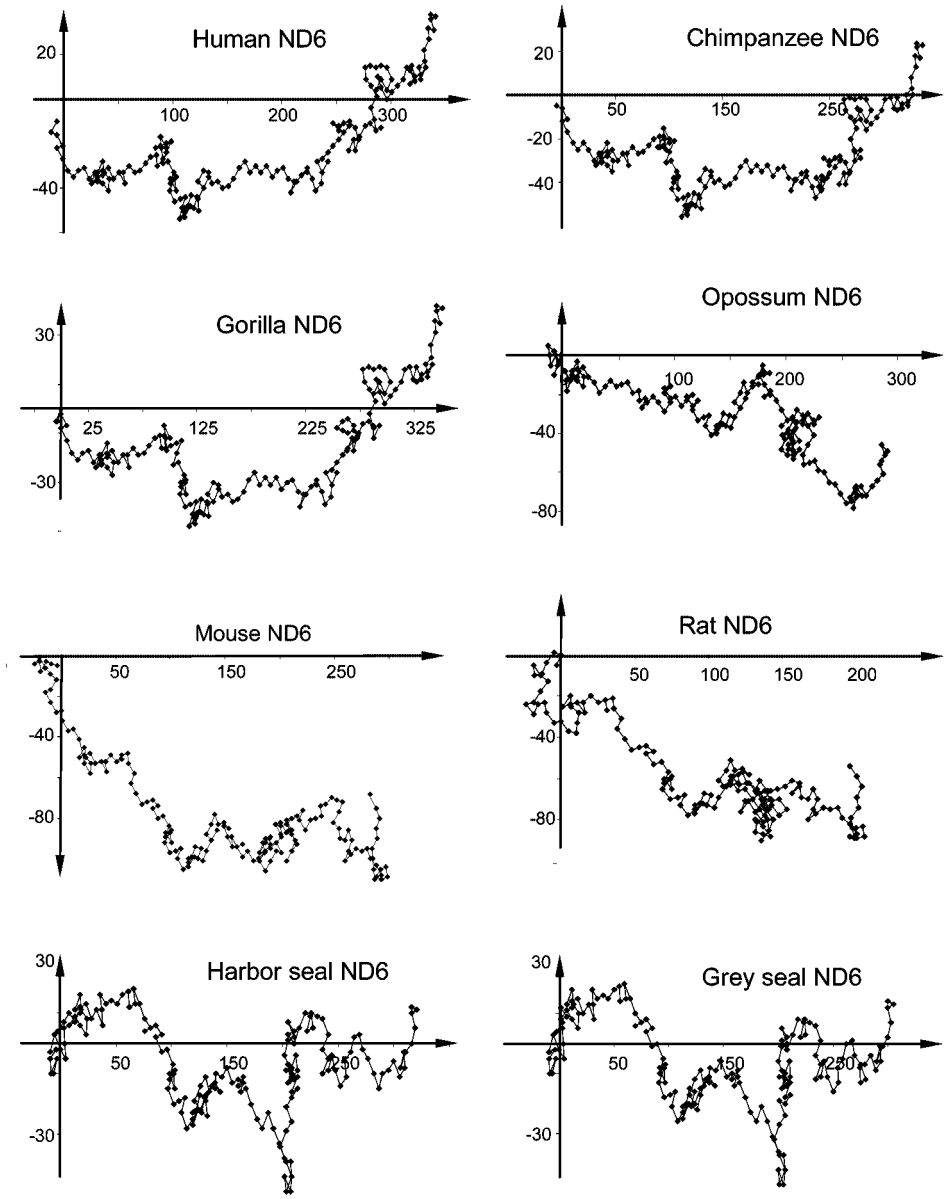


Figure 67. Lattice representation of ND6 proteins of eight species.

of the human ND6 protein relative to the coordinates of the center of the lattice representation (x_0 = 185.7348, y_0 = −16.9803), and the corresponding polar coordinates with respect to the overall center for the average for all amino acids.

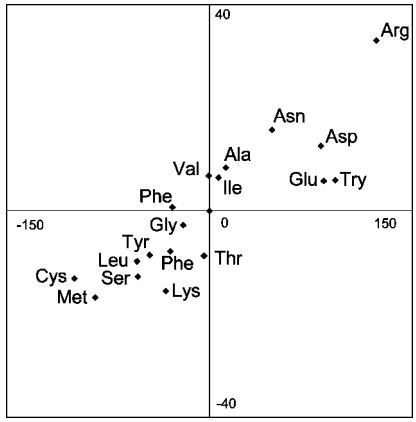


Figure 68. Amino acid map of human ND6 protein and corresponding embedded starlike graph of fixed geometry. Observe the use of different scales for the *x*- and *y*-coordinates.

These coordinates determine the quadrant in which are located individual amino acids relative to the center (x_0, y_0), which allows one to determine the polar angle. Thus, A, R, N, D, E, I, W are in the first quadrant; P, V are in the second quadrant; and C, G, L, K, M, F, S, T, Y are in the third quadrant, which allows amino acids in each quadrant to be ordered in increasing magnitude of the polar angle φ. As a result, the 18 amino acids of human ND6 protein are ordered as

W E D R N A I P V C G L F M S Y K T

By adding at the end in the alphabetical order the nonexistent amino acids one obtains

W E D R N A I P V C G L F M S Y K T H Q

Different proteins are expected to show different orderings of amino acids, although the possibility that related proteins may show the same ordering cannot be discarded, despite the fact that there are 20! possible arrangements of 20 letters in a sequence. When this happens, one can distinguish such by associating with each protein also a 20-component numerical vector that lists the radial distances of average amino acid centers from the center of the lattice representa-

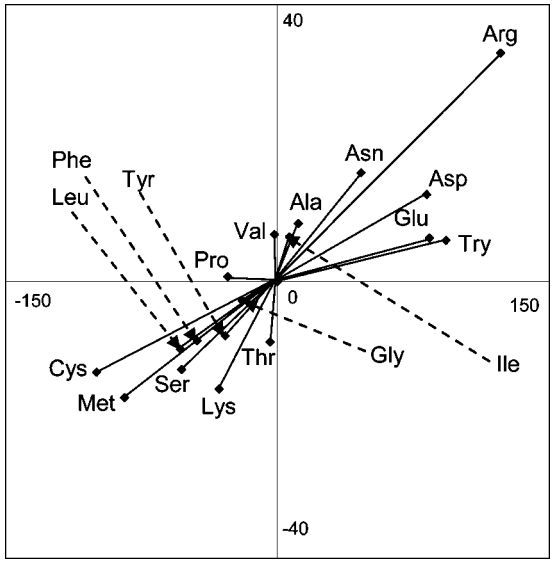


Figure 69. Graph of human ND6 protein as embedded starlike graph of fixed geometry.

tion of the protein. For the above case of human ND6 protein, if one truncates the radial magnitudes to integer parts only, one obtains

(92, 84, 83, 127, 49, 14, 9, 27, 6, 100, 19, 54, 44, 85, 54, 29, 35, 9, 0, 0).

It is very unlikely that there will be two proteins that will have identical orderings both for the alphabetic and the numerical sequence. In the case of such an unlikely overlap in numerical sequences, one can always include the decimal part of radial distances to discriminate such cases. It seems, therefore, that the alphabetic and numerical sequences as described above for human ND6 can serve as an auxiliary notation of proteins. The proposed dual auxiliary notation has an important advantage: a scientist in any laboratory can easily construct the "dual" notation for proteins of his/her interest, and if such "dual" labels are used by the rest of the protein community, this will facilitate the search of relevant protein data.

Table 40. Average Coordinates for Amino Acids of Human ND6 Protein Relative to the Coordinates of the Center of the Lattice Representation ($x_0 = 185.7348$, $y_0 = -16.9803$) and the Corresponding Polar Coordinates with Respect to the Overall Center for the Average for All Amino Acids

	$x - x_0$	$y - y_0$	R	φ
A	11.9795	8.40887	14.6362	0.612026
R	123.2652	32.9803	127.6010	0.261432
N	46.5152	15.7303	49.1030	0.326102
D	82.5985	12.647	83.5611	0.151934
C	−99.7348	−13.0197	100.5810	0.129809
E	84.0985	5.8136	84.2992	0.069019
G	−19.7348	−2.8054	19.9332	0.141209
I	6.9319	6.4803	9.48923	0.751740
L	−53.7936	−9.7844	54.6762	0.179921
K	−32.2348	−15.5197	35.7763	0.448704
M	−84.2348	−16.7197	85.8781	0.195942
F	−44.0348	−8.5197	44.8514	0.191115
P	−27.5348	0.7803	27.5459	−0.028330
S	−53.2348	−12.7197	54.7333	0.234539
T	−4.0681	−8.6864	9.59182	1.132804
W	92.7652	5.9803	92.9578	0.064378
Y	−28.6348	−7.9197	29.7098	0.269831
V	−0.9134	6.8732	6.93363	−1.438680

13. Condensed Graphical Representation of Lengthy DNA

To illustrate this novel graphical representation of biosequences, we will again return to the first exon of human β-globin gene, even though this 92-base DNA can hardly be considered lengthy. Let us count for all four nucleotides the occurrence of adjacencies for pairs of bases XY, where X, Y are A, C, G, and T. In all, one can have 16 pair combinations when one discriminates between XY and YX. In Table 41, we show the 16 × 16 DNA pair adjacency-count matrix, in which the ordering of the labels (alphabetical) is immaterial, as long as it is kept constant for all DNA sequences considered. As a result, we have an asymmetrical matrix as an additional representation of DNA. The novelty consists in viewing this matrix as a map, when one assigns to element i,j the coordinates x_i and y_j and the magnitude a_{ij}. Binary matrices, in which the matrix entries are either one or zero, can also be viewed as maps and can be simply represented graphically by spots, analogous to graphical representations of DNA based on Chaos Game, such as

Table 39. New (20 × 20 Symmetrical) Distance Matrix of Human ND6 Protein; Amino Acids Are Ordered Alphabetically Based on Their Three-Letter Codes

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	0	113	35	70	113	Ø	72	33	Ø	5	68	50	23	44	50	68	23	80	43	12
Arg		0	78	45	227	Ø	47	147	Ø	119	182	162	213	172	154	182	133	40	157	126
Asn			0	36	149	Ø	38	68	Ø	40	103	84	134	93	75	103	56	47	78	48
Asp				0	184	Ø	6	103	Ø	75	138	118	169	128	110	138	89	12	113	83
Cys					0	Ø	184	80	Ø	108	46	67	15	55	73	46	95	193	71	100
Gln						0	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
Glu							0	104	Ø	77	138	169	169	128	111	138	89	8	113	85
Gly								0	Ø	28	34	17	65	24	8	34	16	112	10	21
His									0	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
Ile										0	62	44	94	53	34	63	18	85	38	7
Leu											0	22	31	9	28	2	49	147	25	55
Lys												0	52	13	16	21	28	126	8	38
Met													0	41	59	31	80	178	56	86
Phe														0	18	10	39	137	15	45
Pro															0	29	25	120	8	27
Ser																0	49	147	25	55
Thr																	0	97	24	15
Trp																		0	122	93
Tyr																			0	31
Val																				0

Table 41. Part of the 16 × 16 Matrix Counting Adjacencies among Pairs of Nuclotides in the First Exon of Human β-globin Gene; Observe That, Even though Matrix Is Asymmetrical, Row Sum and Column Sum Are the Same for the Same Pair of Nucleotides

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	0	0	0	0	0	0	1	0	0	0	1	2	0	0	0	0
AC	0	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0
AG	1	0	0	0	0	0	0	0	2	1	0	1	0	1	0	1
AT	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CA	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
CC	0	0	0	0	0	0	0	2	0	0	0	1	0	0	4	0
CG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
CT	0	0	0	0	0	1	0	0	2	2	1	1	0	0	0	0
GA	0	1	2	0	0	0	0	1	1	0	3	0	0	0	0	0
GC	1	1	1	0	0	2	1	0	0	0	0	0	0	0	0	0
GG	1	0	2	1	1	1	0	0	0	2	1	0	0	0	4	0
GT	0	0	0	0	0	0	0	1	2	1	3	0	1	0	1	0
TA	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
TC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
TG	1	1	2	0	1	2	0	0	1	0	3	2	0	0	1	0
TT	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 10, where the first exon of human β-globin gene has been depicted, except for the fact that x, y coordinates in Figure 10 are real numbers, while for maps based on matrices they will be integers.

More generally, a nonbinary matrix can be depicted as a map by using a bubble representation, in which the element i, j, having coordinates x_i and y_j, is assigned a circle with a radius of magnitude a_{ij}. Bubble graphs have been used, for example, to represent data extracted from proteomics maps.[81,91,307] In Figure 70, we have illustrated Table 41, which counts pair adjacencies for the first exon of the human β-globin gene, as a bubble diagram. However, regardless of the length of DNA, such maps can have at most 256 spots, though they can be of different size.

Bubble diagrams, similarly to spectral representations of DNA and proteins, make it possible to consider differences among diagrams by direct subtraction of such diagrams, thus allowing visual inspection of similarities among sequences.

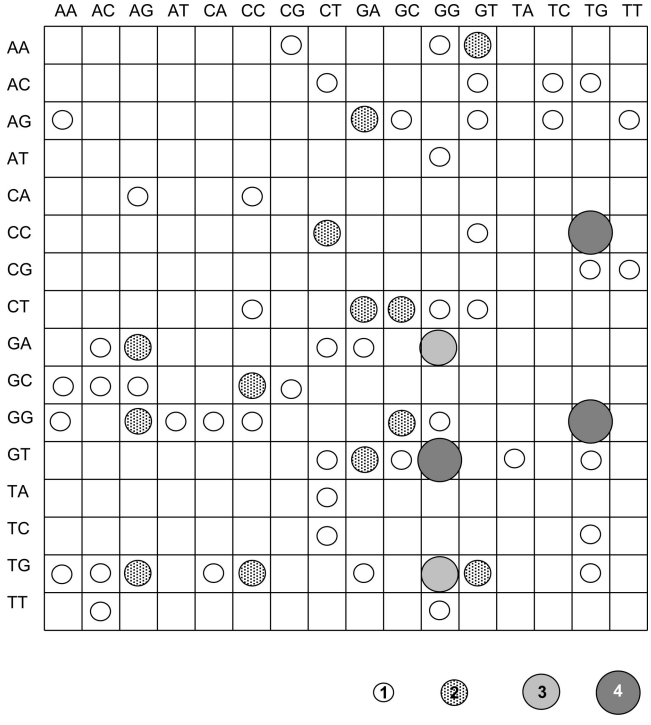


Figure 70. 16 × 16 matrix counting adjacencies among pairs of nucleotides in the first exon of human β-globin gene.

14. Miscellaneous Approaches to Graphical and Numerical Representation of Proteins

Most readers could have noticed as we are approaching the end of the review that this review was focused on our own research. Well, the central theme of this review is proteins, so in the introductory parts we limited the discussion of DNA and RNA mostly to our work, as it serves as an introduction to our own approaches to graphical and numerical characterization of proteins. An overview of graphical approaches to DNA and RNA would have to include numerous contributions from other scientists; particularly to be noticed are contributions from China, but DNA and RNA are outside the scope of the present review. We have mentioned in the introductory part the recent review of Nandy, Harle, and Basak,[41] which in part cover DNA "territory".

However, with respect to proteins, in the past few years there have been a number of valuable contributions, with innovative approaches, some of which will be briefly mentioned. Thus, A. Nandy, Gosh, and P. Nandy[268] proposed a new method to compare sequences through a 20D representation and constructed phylogenetic trees without needing multiple alignments. The approach was then tested by applying it to the issue of evolutionary relationships of rat and human voltage-gated sodium channel α-subunits. Liu and Wang[269] developed a graphical representation for protein secondary structure in which one discriminates among amino acids if they are located in α-helix, β-sheet, or a coil, referred to as helix residue (H), sheet residue (E), or coil residue (C), respectively. They then constructed an equilateral triangle to vertices of which are assigned labels H, E, C and followed with suitably modified Chaos Game algorithm, reminiscent of Jeffrey's graphical representation of DNA. Yu, Anh, and Lau[270] applied the Chaos Game representation to the primary sequences of amino acids in proteins by following the grouping of the 20 natural amino acids of Brown[271] into four classes: nonpolar AAs (alanine, isoleucine, leucine, methionine, phenylalanine, proline, tryptophan, and valine); negative polar AAs (aspartic acid and glutamic acid); uncharged polar class (asparagine, cysteine, glutamine, glycine, serine, threonine, and tyrosine); and positive polar class (arginine, histidine, and lysine).

To these approaches, we should add "generalizations" of DNA graphical representations to proteins, one of which is represented by using codons (either known or based on the

Virtual Genetic code) to identify corresponding amino acids, which was already outlined. Bai and Wang[272] plotted codons and used their coordinates for numerical analysis of proteins. They started with a 2D representation of DNA nucleotide triplets and introduced $z(n)$ to denote a triplet (amino acid) as follows,

$$z(n) = h_0 y(n) + h_1 y(n-1) + h_2 y(n-2)$$

where $y(n)$, $y(n-1)$, and $y(n-2)$ denote three adjacent nucleotides (of the coding region) of DNA and h_0, h_1, and h_2 are real numbers (e.g., $h_0 = 1$, $h_1 = 1/2$, $h_2 = 1/4$), to be combined with y(A), y(T), y(G), and y(C) values, with A and T defined as conjugate ($b \pm di$), and G and C as conjugate ($d \pm bi$), with $b = 1/2$ and $d = \sqrt{3}/2$.

In another study, Bai and Wang[273] considered a regular dodecahedron in 3D space and assigned to its 20 vertices the 20 natural amino acids, which led to a graphical representation of proteins as a walk in 3D space over the vertices of a dodecahedron.

Yang and Wang[274] considered 3D graphical representations of proteins by assigning amino acids to the 20 vertices of a regular dodecahedron proposed by Bai and Wang:

Alanine A = (0, 0, 1);
Cysteine C = (2/3, 0, $\sqrt{5}/3$);
Aspartate D = ($-1/3$, $\sqrt{3}/3$, $\sqrt{5}/3$);
Glutamate E = ($-1/3$, $-\sqrt{3}/3$, $-\sqrt{5}/3$).

and so on, where the entries in parentheses are the x, y, z coordinates of a regular dodecahedron.

The following 3×20 matrix,

$$\begin{matrix} 0 & 2/3 & -1/3 & -1/3 & \ldots \\ 0 & 0 & \sqrt{3}/3 & -\sqrt{3}/3 & \ldots \\ 0 & \sqrt{5}/3 & \sqrt{5}/3 & -\sqrt{5}/3 & \ldots \end{matrix}$$

when multiplied by a column vector (A_i, C_i, D_i, E_i,.. .), where A_i, C_i, D_i, E_i, ..., are the cumulative occurrence numbers of amino acids A, C, D, E, ..., converts a protein sequence into a series of vectors P_0, P_1, P_2, P_3, ..., P_n that gives a graphical representation of a protein as a path in 3D space.

Guo and Wang[275] revisited the concise description of proteins, known as TOPS (topology of protein structures), initiated over 30 years ago by Sternberg and Thornton,[276] in which one considers in protein secondary structures the presence of strands and helices, neglecting the length and structure of loops. These secondary structure elements are considered to have a direction "up" or "down" (that is, out of the plane of the diagram or in plane). The secondary structure of a protein is thus reduced to a four-letter alphabet with conventions attributable to Yuan, Liao, and Wang[277]

$(-1, 0, i) \rightarrow e$ "down" strand;
$(1, 0, i) \rightarrow E$ "up" strand;
$(0, -1, i) \rightarrow h$ "down" helix;
$(0, 1, i) \rightarrow H$ "up" helix,

where i is the running index of the amino acid. Besides offering a visual representation, this simplified TOPS notation immediately allows one to identify proteins belonging to *eE*, and *hH* groups, that is, "strand only" and "helix only", which are clearly going to be grouped in the corresponding branches of phylogenetic code, as was illustrated by Guo and Wang on a set of 34 proteins of Chew and Kedem.[278]

Finally, we may mention in this section the work of Gonzales-Dias and his research group from Santiago de Compostela, Spain, whose interest covers the use of topological indices in Medicinal Chemistry and Bioinformatics, including proteomics.[279−286]

15. Characterization of Proteomics Maps by Matrix Invariants

15.1. Introduction About Proteomics

Proteomics maps report on separation of cellular proteins by mass and by charge using electrophoresis and gel chromatography. The experimental results are obtained as two-dimensional gel plates, which allow one to list the x, y coordinates for the proteins that are well resolved. Proteins always appear at the same location in the 2D gel, having thus always the same (x, y) coordinates in all proteomics maps (following the same experimental protocol). Thus, proteomics maps of the same organ cells will differ in the abundance of individual proteins (seen on a plate as spots of different size), and this can vary considerably in an unpredicted manner. Often proteomics maps are visually inspected, whereas the differences in abundances of individual proteins are measured.

In order to make quantitative comparisons of different maps, one needs to characterize numerically maps M(x_i, y_i, z_i), where the suffix i refers to the ith protein ($i = 1, ..., N$), with N being the number of proteins considered. Hence, the problem is a characterization of the proteomics map as a whole, which is tantamount to considering the list of (x, y, z) coordinates (where x and y indicate the mass and the charge of proteins, and z measures the abundance) in totality. A way to do this is to select a suitable geometrical object for a map M(x_i, y_i, z_i), which will then lead to a mathematical representation of 2D proteomics maps.

Work on numerical characterization of proteomics maps was initiated in 2003,[287,288] and the following geometrical objects were considered for such a numerical analysis of proteomics maps:

(1) Zigzag curve connecting N most abundant points;[287−290]
(2) Graph of partial ordering of spots with respect to mass and charge;[291−293]
(3) Cluster graph connecting all spots with distances smaller than a selected threshold value;[294]
(4) Graph connecting the nearest pairs of spots among the set of N spots selected;[295]
(5) Graph connecting the nearest pairs of spots sequentially, restricting attention only to spots having larger abundance;[296] and, most recently,
(6) Adjacency matrix for Voronoi regions constructed for N proteins of a map.[293]

The above schemes offer sets of invariants, which can be used as map descriptors that often are of different structural origin. We already know from structure−property−activity studies that it is useful to have alternative approaches and different sets of descriptors, a fact which has been demonstrated by the use of numerous so-called topological indices as molecular descriptors.[297−304] Use of different descriptors in SAR and QSAR (structure−activity relationship and quantitative structure−activity relationship, respectively) facilitates arriving at better regression models. Similarly, one may expect that the use of different map descriptors will facilitate comparative studies of proteomics map construction by leading to better numerical characterization of proteomics maps. This will help to recognize and eliminate false positives, that is, to identify as dissimilar maps that would accidentally be characterized by similar sets of descriptors.

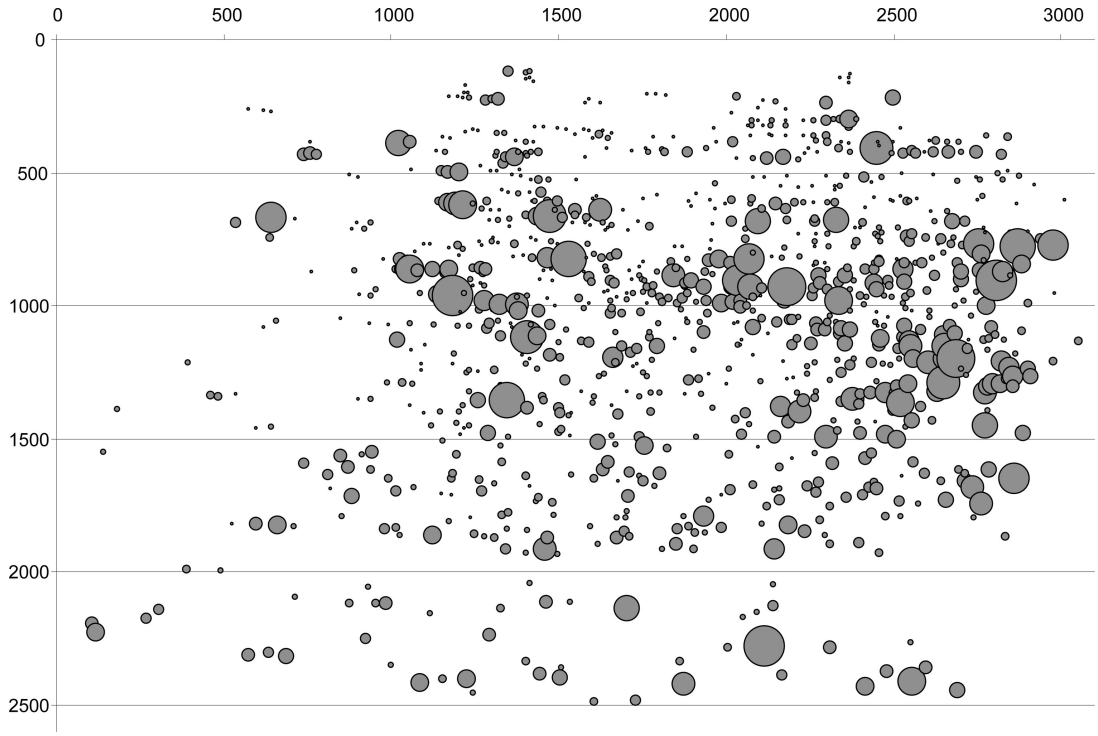


Figure 71. Bubble diagram summarizing positions and abundances of proteins in a proteomics map. Reproduced with permission from ref 91. Copyright 2001 American Chemical Society.

Table 42. (*x*, *y*) Coordinates of 20 Most Abundant Protein Spots in the Control Group and the Corresponding Abundances When Four Different Proliferators Were Added to the Feed of Rats (From the Work of Anderson et al.[308])

	x	*y*	**control**	PFOA	PFDA	clofibrate	DEHP
1	2111.7	2278.6	**144357**	108713	95028	147081	165886
2	2804.3	903.6	**143630**	155565	188582	159898	155055
3	1183.9	959.6	**136653**	113859	150253	163645	8111
4	2182.2	928.8	**127195**	99160	73071	76642	112096
5	2685.6	1196.1	**118581**	112790	49769	109856	138795
6	1527.9	825.5	**114929**	192437	221567	166080	180590
7	1346.0	1352.5	**112251**	58669	38915	73159	77075
8	2868.5	778.0	**108883**	26105	50735	45923	116849
9	1406.3	1118.1	**98224**	91147	82963	84196	92942
10	2450.2	409.2	**93601**	83172	62934	79870	109381
11	1474.0	665.1	**90004**	129340	112361	112655	119402
12	2974.9	772.8	**86730**	70746	78691	105760	116281
13	2068.4	823.1	**84842**	73814	45482	71911	97444
14	642.2	669.8	**82492**	73974	74466	84703	88545
15	2860.7	1649.9	**81965**	16137	16501	60077	148992
16	2032.7	902.8	**80015**	77314	80072	76027	100836
17	2752.7	765.6	**79847**	20782	13103	38816	53830
18	2334.2	982.2	**72791**	76369	52749	55599	77432
19	1053.6	864.3	**72173**	77982	60376	46808	78121
20	2519.5	1365.9	**69452**	37838	16129	57167	71274

As is well known, any characterization of mathematical objects by a finite set of invariants may be accompanied by loss of information. Combined use of descriptors arising from different representations may in part recover some of the lost information. This suffices to encourage the search for alternative graphical and numerical representations of proteomics maps and the construction of additional map invariants. We may add that some of the above-mentioned approaches (1−6) are computationally more involved, either requiring more time for construction of geometrical templates for maps or leading at the end to *dense* matrices (matrices with few zeros, except on the main diagonal[305]). This is illustrated by the construction of the distance/distance matrices,[58] whether when considering zigzag curves (when one has to compute all $N(N-1)/2$ distances between N points) or when one considers the construction of matrices for cluster graphs (which requires finding the shortest paths in a graph having numerous alternative cyclic routes between pairs of spots). As is known, the search for shortest paths in cyclic graphs, the algorithm for which was reported by Dijkstra,[306] is time-consuming. By contrast, graphs resulting from partial ordering of proteins in a map with respect to mass and charge, and graphs depicting the nearest neighbors, lead to *sparse* matrices, which would be therefore more suitable for characterization of proteomics maps when one considers large numbers of maps or includes in the analysis most or many proteins in a map. Clearly, representations of 2D proteomics maps by *sparse* matrices are desirable.

We will now briefly outline the above-listed approaches for a quantitative characterization of proteomics maps. It may be of interest to recall from earlier sections of this review that we outlined several graphical representations of DNA

and protein sequences, which are 1-dimensional objects, as 2D maps. This was the case with Jeffrey's chaos-game representation of DNA, with the spiral representations of DNA using a "wormlike" spiral or resulting in a four-color map for DNA, and chaos-game representation of proteins using either virtual code or the interior of the "magic circle". With proteomics maps, we may have the opposite situation, as illustrated already by the first numerical characterization of proteomics maps, which was converted to a 1-dimensional object (a sequence).

All the approaches listed initially try to summarize collective information on proteome contained in a 2D gel. Experimental results consist of a list of (x, y) coordinates of protein spots and their abundances. This information on a 2D gel can be visually presented as a "bubble" diagram (Figure 71)[81,91,307] in which spots are shown as circles with the radius of a circle indicating the relative abundance. For us, the "bubble" diagrams of a proteomics map represent an *input* information on a map for theoretical considerations. In order to delegate a comparative study of proteomics maps to computers, one needs map *invariants* (map *descriptors*). Even though (in general) from a set of invariants one cannot reconstruct a molecule, DNA, protein, or a map, nevertheless, invariants may allow *comparisons* of maps and finding degrees of *similarity* among maps. The "art" in construction of map invariants, just as has been the case with construction of molecular descriptors,[298−304] is in finding invariants that capture important structural features leading to *useful* data reduction. The underlying assumption here is that one can extend Emil Fisher's dogma that *similar compounds* will show *similar properties* from the structure−property−activity studies to the structure−proteome perturbation studies.

We will illustrate characterizations of proteomics maps by considering data of Table 42, where are listed (in descending order) the (x, y) coordinates for the most abundant 20 protein spots of rat liver cells for the control group and four peroxisome proliferators as reported in ref 308. Observe a somewhat *chaotic* behavior of the cell proteome under different situations, showing no apparent regularity in changes in the abundance of different proteins. A protein that has increased the abundance in one case may have decreased the abundance dramatically when a different agent is used, it may stay unchanged, or it may have increased its presence. The data of Table 42 leads then to simplified "theoretical" proteomics maps having only 20 spots, illustrated in Figure 72, with the five proteomics maps differing only in abundance of corresponding proteins.

Useful descriptors for characterization of proteomics maps include the *eigenvalues* of various matrices **M** that one can associate with the proteomics map. In particular, the *leading eigenvalues* of such matrices have been frequently selected. Formally, this means solving the equation: Det $| \mathbf{M} - \lambda \mathbf{I} | = 0$. Here Det is the determinant of a matrix, **I** is an $N \times N$ unit matrix (having all elements zero except on the main diagonal, which are equal to one, with N being the number of vertices), **M** is the selected graph matrix, and λ are the eigenvalues sought. There are several computer programs that calculate matrix eigenvalues, including MATLAB (an abbreviation for Matrix Laboratory),[309] which we found suitable for our needs.

15.2. Zigzag Curve

The shape of the zigzag curve depends on ordering of points selected from a proteomics map considered. In order to *facilitate* comparisons of different 2D maps, we use the following rules:

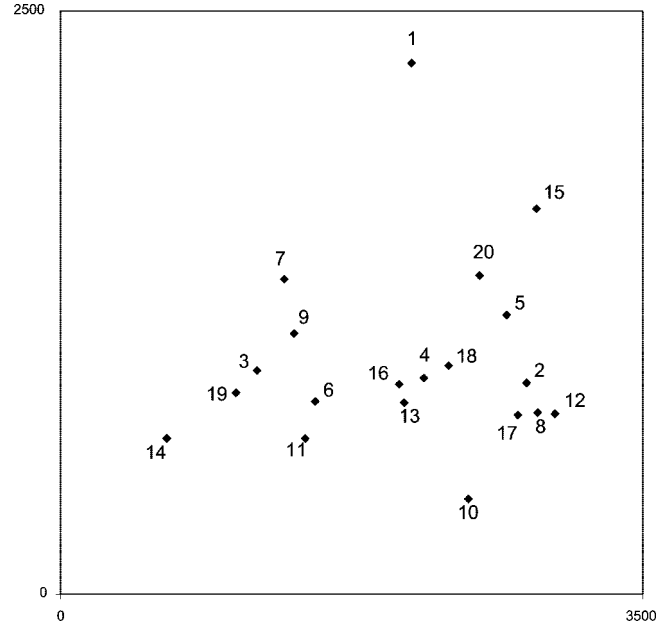


Figure 72. Simplified proteomics map showing 20 most abundant proteins.

Rule 1: First one selects any N protein spots, where there are abundance variations that one wishes to examine more closely.

Rule 2: The set of selected N spots are ordered according to some adopted criterion (such as the relative abundance).

Rule 3: Proteins in all maps to be compared are ordered in the same way. For example, according to the order chosen for the control group (e.g., healthy animals, or tissue).

In Figure 73 we displayed the zigzag curve based on ordering of proteins according to relative abundance in the control group of Table 42. An advantage of the zigzag curve is the simplicity of the construction of the curve and Euclidean and distance matrices.

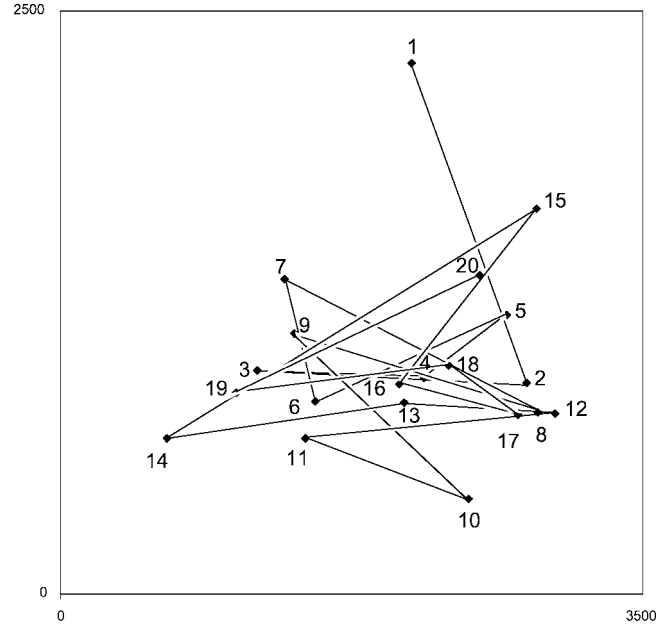


Figure 73. Zigzag curve connecting spots in proteomics maps of Table 42 in order of decreasing abundance.

15.3. Graph of Partial Ordering

Partial order consists of ordering spots with respect to their masses and their charges *simultaneously*. Let us again consider data on the control group of Table 42. As one can see, in both columns x and y, protein 1 is ahead of protein 7, which is ahead of protein 3, which is ahead of protein 19, which is finally ahead of protein 14. Hence, the relative order $1 > 7 > 3 > 19 > 14$ is satisfied for *both* x and y, for mass and charge as protein properties. Thus, we have established a *dominance* of protein 1 over protein 7 over protein 3, etc. A list of all such dominance sequences represents the partial ordering. Graphically this is illustrated in Figure 74, where dominance is indicated by connecting the corresponding protein spots by lines. The above case, $1 > 7 > 3 > 19 > 14$, is depicted as a *path* starting at vertex 1 and ending at vertex 14, which is the most left path in Figure 74. The numbering of spots is shown in Figure 72.

Construction of the partial ordering graph over a map is not difficult if one observes a simple rule for such a construction: (1) all the connecting lines have a positive slope and (2) no connecting lines are needed for intermediate members in an ordering sequence. Thus, in the above case, it suffices to connect 1 to 7 and 7 to 3, but there is no need to connect 1 and 3, since their dominance already has been indirectly established. As one can see from Figure 74, the number of lines (edges) between points has increased approximately by a factor of 2 in comparison with the number of line segments in the zigzag curve (Figure 73).

An advantage of the graph of partial ordering is that it can be constructed regardless of the data on the control group. This is important in a case where there is no control group or no obvious control group. For example, proteomics maps of tumor cells (leukemia, melanoma, etc.) or proteomics of swollen tissues (e.g., synovial fluid), of different bacteria, etc. have no such obvious control group. The partial ordering graphs are not very dense, even if the number of spots has increased considerably; thus, it is not difficult to find the shortest distances of interest for the construction of distance matrices.

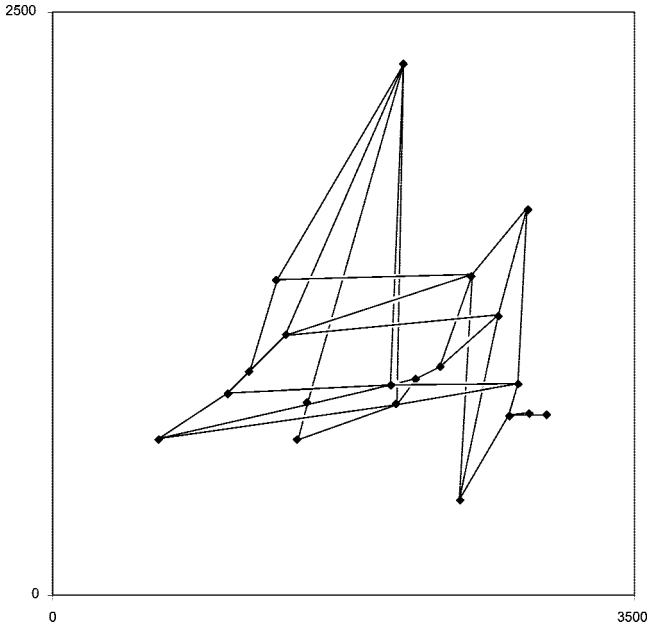


Figure 74. Diagram of partial ordering superimposed on the proteomics maps having 20 proteins spots.

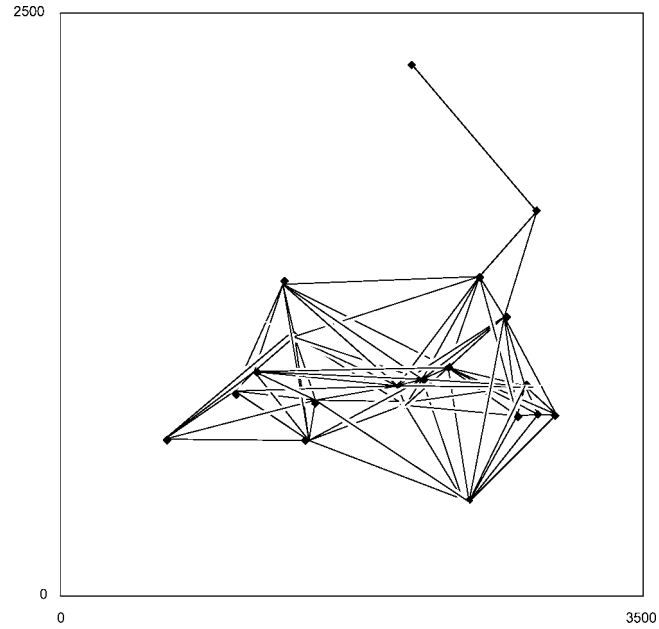


Figure 75. Diagram of partial ordering superimposed on the proteomics maps having 20 protein spots.

15.4. Graph of Clustering of Spots

In Figure 75, we show a graph obtained by connecting all spots that are within a selected prescribed distance. A cluster graph can be based on 2D Euclidean distances between spots or on 3D Euclidean distances between spots, in which case one takes the relative abundance of spots as the third coordinate. As we see from Figure 75, the number of lines (edges) between points has increased by an order of magnitude in comparison with the graph of partial ordering (Figure 74). Matrices constructed for cluster graphs have more information on the underlying maps, which may be an advantage and allow variations of the critical distance, thus adding some variability to this approach. A disadvantage is that the search for shortest paths could be computer intensive and represents the slow step in the construction of the distance-adjacency matrix, **E/D** matrix, which is the adjacency matrix weighted by Euclidean distances between the adjacent vertices. The use of Dijkstra′s algorithm for finding the shortest paths in the case of modeling proteomics maps by 100−200 spots may take a few seconds of CPU.[294]

15.5. Nearest-Neighbor Graphs

There are two kinds of the nearest-neighbor graphs: (1) For selected N protein spots, one lists k nearest neighbors ($k = 1, 2,...$) and connects them. (2) For selected N protein spots, one sequentially connects vertex i to the nearest vertex i' among vertices $i' < i$. The case $k = 2$ of nearest neighbors for 20 proteins of Table 42 is illustrated in Figure 76, while for the same 20 spots, the graph of the sequential nearest neighbors is illustrated in Figure 77. The advantage of the approach using k nearest neighbors is the added flexibility, which allows one to vary k and thus use more information of protein spots. The disadvantage of this approach is that, if one decides to include additional proteins from the proteomics maps, then one has to recalculate all distances and construct a novel graph, for which the part that corresponds to the initial N spots may change. The advantage of the sequential-neighbor graphs is that a successive increase

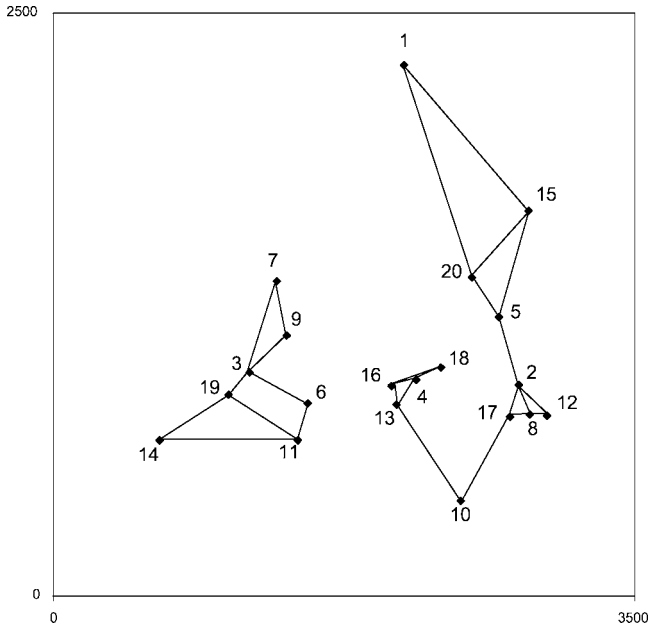


Figure 76. Graph of nearest neighbors connecting $k = 2$ spot in proteomics maps of Table 41 that are the closest.

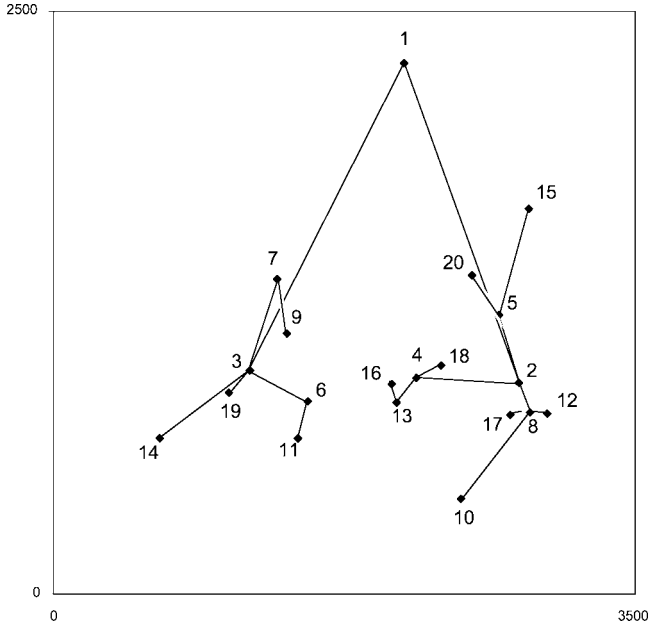


Figure 77. Graph of sequential nearest neighbors connecting spot in proteomics maps of Table 41.

of the number of proteins to be considered does not alter the results already obtained but rather incorporates them.

15.6. Graph Based on Adjacency of Voronoi Regions

This most recently developed approach,[293] which is based on the construction of the adjacency matrix of Voronoi regions, is outlined at the end of the review. Voronoi regions partition the plane having N discrete points into N regions such that all (x_i, y_i) points in the plane that are closest to spot N_i belong to the same region. Two advantages of this approach are, like the case of the partial ordering diagram, that (1) it is not highly sensitive to the precise position of proteins spots in a 2D gel and (2) it does not need a reference map (like control-group proteomics maps) for processing proteomics maps, which as already mentioned is important when one considers data for which there are no obvious control-group cells.

16. Map Invariants

Once one constructs matrices for quantitative comparative study of proteomics maps, in order to increase the number of descriptors, one can consider the leading eigenvalue of matrices $\mathbf{M}^{(n)}$ and $\mathbf{M}^{[n]}$, the first representing the standard multiplication of a matrix by itself and the second representing the Hadamard multiplication of matrices by itself. In the limiting case where $n \rightarrow \infty$, the leading eigenvalue of the matrix $\mathbf{M}^{[\infty]}$ contains only the contributions from adjacent vertices while all others disappear.

Having constructed the adjacency matrix with elements weighted by the Euclidean distance between vertices i and j, which we denoted as **ED** matrix, one proceeds to construct the quotient matrix $\mathbf{Q}_{ij} = (\mathbf{ED}_{ij})/(\mathbf{PD}_{ij})$, where $\mathbf{PD}_{ij}$ is a path-distance matrix, the elements of which are defined as the length of the shortest path between vertices i and j. The advantage of a quotient matrix is that its elements are nondimensional, that is, they do not depend on the units chosen to measure distances. Moreover, all the elements are <1, except for the spots that are adjacent (or accidentally collinear), in which case the matrix elements are equal to 1. Hence, when constructing the higher-order matrices by using the Hadamard product, there is no need for any additional normalization, because the elements of the higher-order matrices will steadily decrease in magnitude. However, when one considers the higher-order matrices obtained by the standard matrix multiplication, one has to introduce a normalization factor (such as $1/n!$) in order to secure the convergence of leading eigenvalues of the corresponding matrices.

16.1. 3D Zigzag Curve

There are two ways of incorporating the information on the relative abundance of protein spots in numerical analysis of proteomics maps: (1) one can include the information on relative abundance in the adjacency or the distance matrix of a map, or the diagonal zero entries by suitably normalized relative abundances, or (2) one can view the abundance values as the third coordinate and arrive at a 3D spatial zigzag curve as a representation of the proteomics map. In Figure 78, we show the 3D zigzag line connecting the first 20 most intensive spots of Tables 42, and in Table 43, we have illustrated a 5×5 fragment of the **E/D** 20×20 matrix. As one can see, the elements on the main diagonal are zero whereas the adjacent elements are equal to 1, because the Euclidean distance and the distance along the connecting line are the same for directly connected spots. All other elements are necessarily <1, unless it happens (which is possible but not likely) that three points are collinear. In Table 44, we have listed the leading eigenvalues of 20×20 **E/D** matrices for a selection of exponents n. As the exponent n increases to ∞, the leading eigenvalues converge to the limiting value $\lambda = 1.977\,661\,652\,450\,26$, which is the leading eigenvalue of a path of length 19 (in general the leading eigenvalue of a path of length $N - 1$ for $N \times N$ matrix.

16.2. Cluster Graphs

The cluster graph illustrated in Figure 75 is based on the 3D distance between spot no. 1 and spot no. 15 (of Table 7) as the critical distance. This distance is equal to

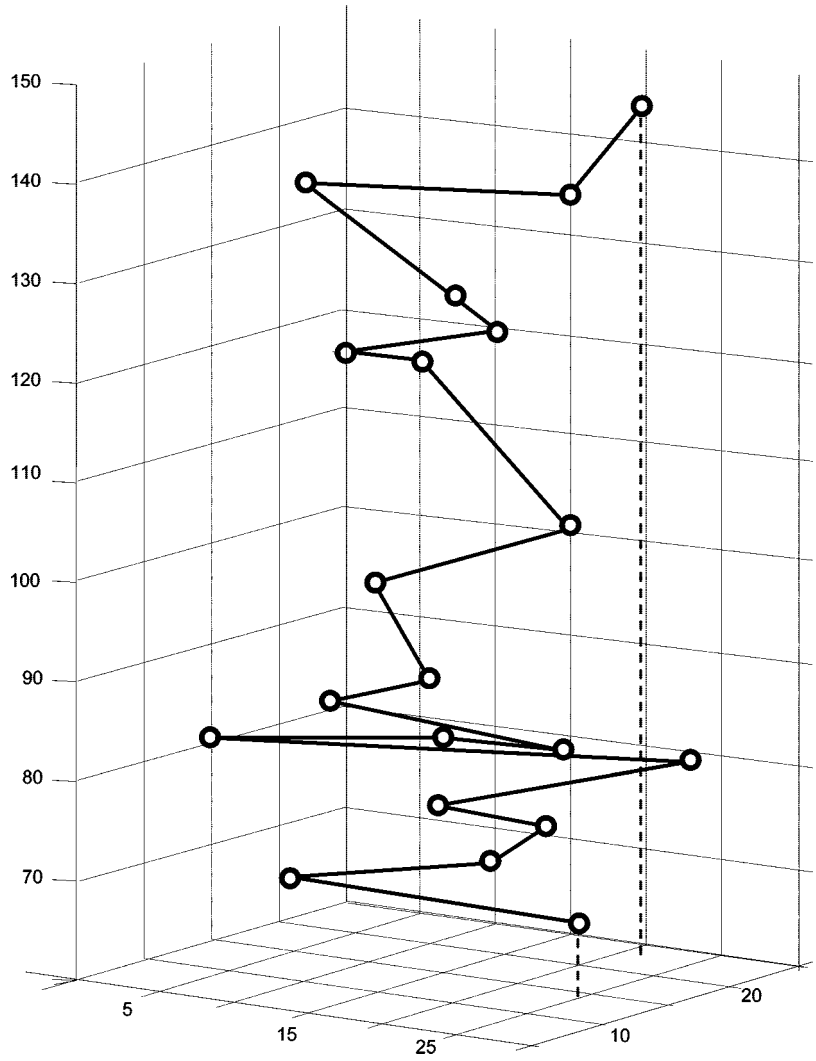


Figure 78. 3D zigzag line connecting the first 20 most intensive spots of Table 42.

Table 43. 5 × 5 Fragment of the 20 × 20 Matrix E/D (the Matrix Elements of Which Are Given As the Quotient of Euclidean Distance between Vertices *i*, *j* and the Distance along the Shortest Path)

	1	2	3	4	5
1	0	1	0.5309	0.4728	0.5002
2	1	0	1	0.5582	0.6047
3	0.5309	1	0	1	0.9741
4	0.4728	0.5582	1	0	1
5	0.5002	0.6047	0.9741	1	0

0.454 47 and is the maximal smallest distance between any pair of spots. In this way, the cluster graph remains connected, and if we were to select a smaller distance, this would result in a disconnected graph. Besides the zero diagonal elements, the zero entries in the distance-weighted adjacency matrix appear for protein spots not connected in Figure 75; otherwise, the matrix elements are given as the Euclidean distance between the points. In the construction of the shortest-path matrix elements, one has to find for every pair of vertices the shortest paths among numerous possible paths.

16.3. Characterization of Changes in Cell Proteome

A good way to test map descriptors is on proteome variations when the same toxic agent is administered to animals at different concentrations. Here we reexamined the data of Anderson et al.,[308] who reported data for mouse liver cells under variations of dose of peroxisome proliferator LY711883. Using the zigzag approach for the 20 most abundant proteins of the control group for the proteomics maps having concentrations 0.003, 0.01, 0.03, 0.1, 0.3, and 0.6, one constructs the **D/D** matrix after rescaling the *x*, *y* coordinates and the relative abundances in the interval (−1, +1) following the recommendation of Kowalski and Bender.[314] The unscaled and scaled coordinates and the relative abundances for the first 5 of the 99 spots under consideration are listed in Table 45.

From the constructed **D/D** matrices, one obtains the leading eigenvalues for each concentration:

concentration	0.003	0.01	0.03	0.1	0.3	0.6
leading eigenvalue	9.691	9.745	9.764	9.718	9.668	9.642

The above entries can be found as the first line in Table 4 of ref 290 and are plotted in Figure 79. This figure could have been reported in ref 290 but in fact has been constructed for the first time for this review. The shape of the curve in this figure is very interesting. It is of a typical J-shape, which one can find in some plots of the dose−response curve.

Interestingly, for the same data of Anderson et al.,[308] Randić and Estrada have shown in 2005,[14] that is, a few years

Table 44. Leading Eigenvalues of 20 × 20 E/D Matrices for a Selection of Exponents *n* (As the Exponent *n* Increases to ∞, The Leading Eigenvalues Converge to the Limiting Value, Which Is the Leading Eigenvalue of a Path of Length 19)

n	zigzag
1	8.080 0
2	4.610 7
3	3.340 3
4	2.818 2
5	2.578 1
10	2.381 1
20	2.373 4
50	2.358 7
100	2.336 1
250	2.276 0
500	2.197 2
1 000	2.096 6
2 500	1.998 1
5 000	1.979 5
10 000	1.977 6
25 000	1.977 661 652
50 000	1.977 661 652 450 26
limit	1.977 661 652 450 26

Table 45. Unscaled and Scaled Coordinates and the Relative Abundances for the First 5 of the 99 Spots Considered (In the Interval (−1, +1) Following the Recommendation of Kowalski and Bender[314])

spot	*x*	*y*	0	0.003	0.01	0.03	0.1	0.3	0.6
					unscaled				
1	78	147	24.6	24.6	32.0	34.4	41.8	46.7	56.6
2	103	161	15.3	18.4	18.4	16.8	16.8	15.3	16.8
3	78	128	54.1	54.1	59.5	59.5	59.5	70.3	75.7
4	84	144	17.6	21.1	21.1	19.4	19.4	19.4	21.1
5	120	130	17.2	25.8	22.4	24.1	37.8	49.9	55.0
					scaled				
1	−0.103	+0.118	0.072	0.054	0.099	0.127	0.168	0.208	0.295
2	−0.037	+0.159	0.015	0.022	0.026	0.023	0.023	0.015	0.023
3	−0.103	+0.063	0.251	0.203	0.248	0.275	0.272	0.353	0.424
4	−0.087	+0.110	0.029	0.036	0.040	0.038	0.036	0.040	0.054
5	+0.008	+0.069	0.027	0.060	0.047	0.066	0.145	0.227	0.284

later, for the first time *hormesis*, a nonlinear dose−response, at the cellular level, illustrated in Figure 80.

Randić and Estrada calculated the differences in the abundance for 100 spots of proteomics map for the control group and when peroxisome proliferator LY 711883 was administrated into animal diet. The differences in abundances

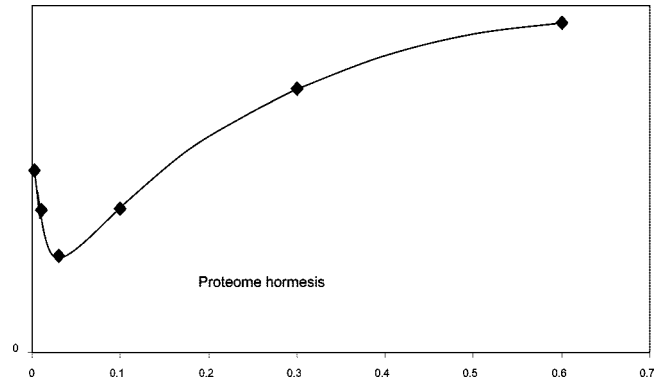


Figure 80. Nonlinear dose−response curve for variation of concentrations of peroxisome proliferator LY 171883. Reproduced with permission from ref 14. Copyright 2005 American Chemical Society.

between the corresponding proteins in the control map and variable concentrations maps are illustrated in Figure 81, in which along the *x*-axis is the running index of spots and along the *y*-axis is the departure in the abundance of proteins relative to the abundance of the control group.

Already visual inspection of Figure 81 shows that, as the concentration of LY 711883 increases from 0.003 to 0.01 and 0.03, the variations in the abundance of proteins from those in the control group (indicated by the heights of the peaks in Figure 80) decrease. However, after having a minimum at 0.03 with a further increase of the concentration of LY 711883, the differences in the abundance increase. An alternative graphical representation of the same is shown in Figure 82, as the correspondence of the protein abundance at concentrations of LY 711883 against the protein abundance in the control, where also the numerical magnitudes for the overall changes have been added.

Figure 79 thus leads to the following conclusions. It shows (1) that one can detect hormesis also directly from the proteomics maps; (2) that the quotient of the leading eigenvalues of the **ED** matrix of a map and the control maps can be interpreted as an index of the overall variation of proteome in cells, more specifically, at least this is the case for the liver cells of experimental animals and peroxisome proliferator LY 711883; and (3) that one can arrive at the correct shape for the dose−response curve by considering a rather limited number

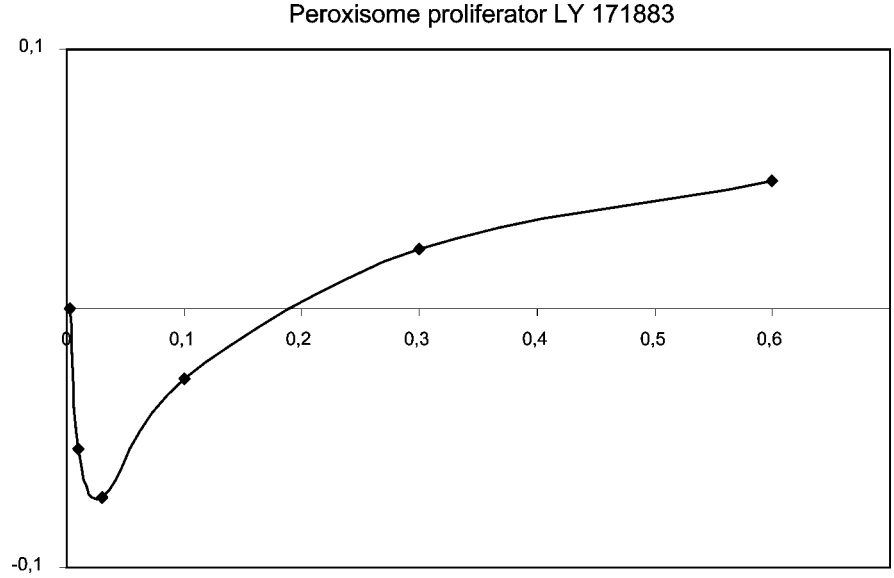


Figure 79. Plot of the leading eigenvalue of **D/D** matrices for various concentrations of peroxisome proliferator LY 171883.

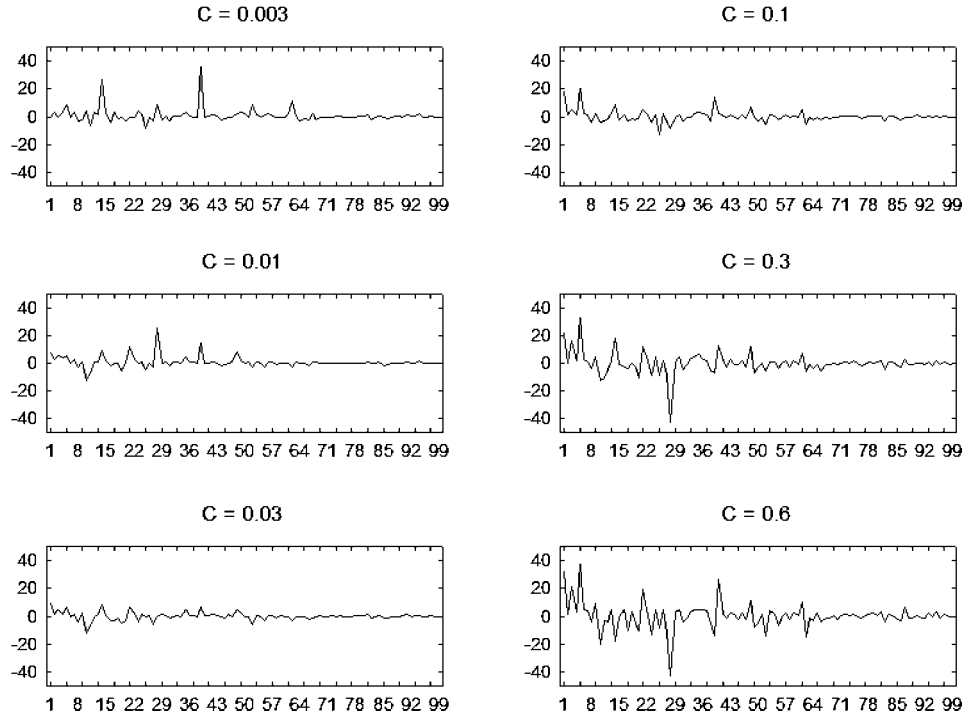


Figure 81. Variation of protein abundance for 99 proteins with different doses of the peroxisome proliferator LY 171883. Reproduced with permission from ref 14. Copyright 2005 American Chemical Society.

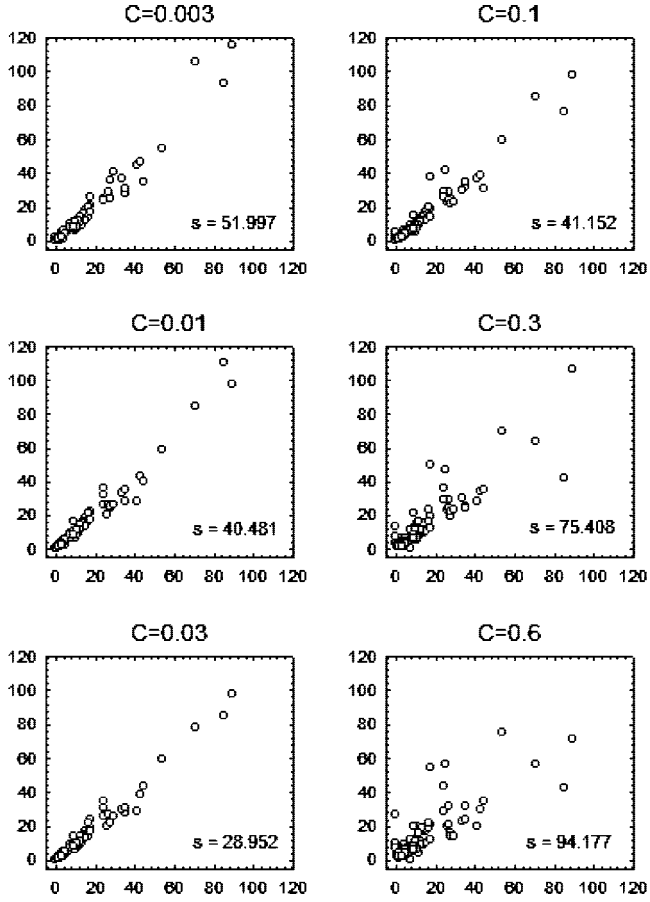


Figure 82. Alternative graphical representation of variation of protein abundance for 99 proteins with different doses of the peroxisome proliferator LY 171883 plotted against the protein abundance in the control. Reproduced with permission from ref 14. Copyright 2005 American Chemical Society.

of most abundant proteins, as was the case here by considering only the 20 most abundant proteins.

16.4. Adjacency of Voronoi Regions

We will end the section of proteomics maps and this review by outlining a promising quantitative characterization of 2D proteomics maps using Voronoi regions[310,311] as the starting step in analyzing proteomics data. For a given set of N points in a plane, Voronoi regions are defined so that the area closest to a point N_i forms a region. In Figure 83, we illustrate Voronoi regions for the 20 points with the coordinates (x, y) listed in Table 42.

The boundaries of Voronoi regions for some cells at the periphery need not be obvious but can be easily established from the Delaunay triangulation of the Voronoi diagram, illustrated in Figure 84.

The Delaunay triangulation of the Voronoi diagram is the *dual* of the Voronoi diagram, and is constructed by connecting the points that define Voronoi regions if they have a common boundary.[312,313] The starting point for numerical characterization of proteomics maps is the adjacency matrix accompanying the Voronoi partition of proteomics map, which is shown in Table 46.

We have illustrated the approach on a simplified map based on the 20 most abundant proteins spots of proteomics maps of Anderson and co-workers,[308] who studied the effects of peroxisome proliferators on proteins in mouse liver. The x, y coordinates for the 20 most abundant proteins in the control group are those already listed in Table 42, to which Figures 83 and 84 belong. Before the quantitative analysis of data, we should draw the attention of readers to the fact that the three variables (x, y, z), where z gives the relative abundance of proteins, each use a different scale to measure experimental quantities (mass, charge, abundance). Because we will combine the x, y, z triplets into an augmented adjacency matrix, the role of different scales has to be considered. Kowalski and Bender[314] were addressing some aspects of this problem, and they recommended a rescaling of different quantities all to a [0, 1] interval. Because the x and y coordinates for the 20 proteins considered are of the same

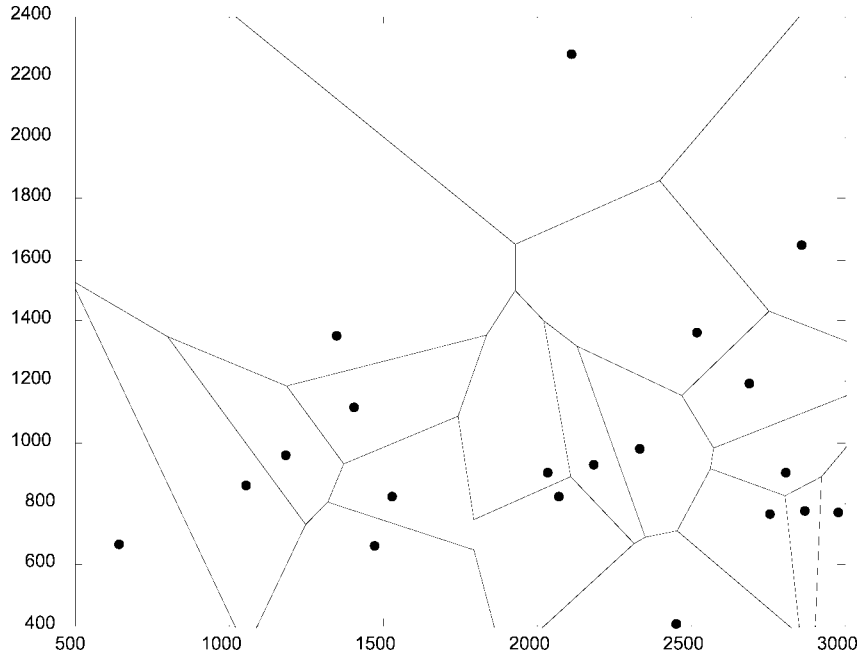


Figure 83. Voronoi regions for the proteomic maps considered in this contribution.

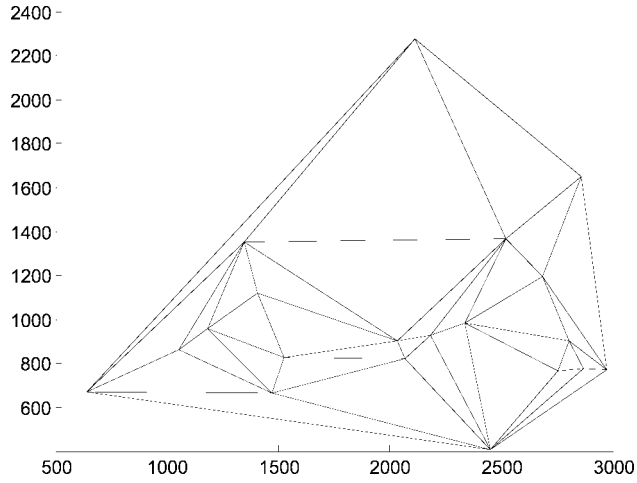


Figure 84. Accompanying Delaunay triangulation of the Voronoi regions of Figure 83.

magnitude, there is no urgency for their rescale. However, because the relative abundance of proteins will enter as diagonal entries into adjacency matrices, one has to scale the abundance values of proteins, which poses a problem beyond the approach of Bender and Kowalski that does not take the size of the adjacency matrix into consideration, which is here important. If one wants to maintain the relative role of the diagonal and the off-diagonal entries of a matrix, one should rescale differently the diagonal entries of a 20 × 20 adjacency matrix and the diagonal entries of a 40 × 40 adjacency matrix. A way to do this is to normalize diagonal entries so that the matrix trace (the sum of diagonal entries) equals the sum of off-diagonal entries in the matrix.

For example, the Laplace matrix,[315] which is of interest also in Chemical Graph Theory,[316,317] satisfies this criterion. When this approach is adopted for the adjacency matrix entries of Table 46, one obtains the relative abundances listed in Table 47. The scaling factor used is 102/1 998 615 = 1/1 959.26, where 102 is the sum of off-diagonal entries in

Table 46. Adjacency Matrix for Voronoi Regions of Figure 83 (Off-diagonal Zero elements Are Not Shown)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0						1							1	1					1
2		0			1			1				1					1	1		
3			0			1	1		1		1								1	
4				0						1			1			1		1		1
5		1			0							1			1			1		1
6			1			0			1		1		1			1				
7	1		1				0		1					1		1			1	1
8		1						0		1		1					1			
9			1			1	1		0							1				
10				1				1		0	1	1	1	1			1	1		
11			1			1				1	0		1	1					1	
12		1			1			1		1		0			1					
13				1		1				1	1		0			1				
14	1						1			1	1			0					1	
15	1				1							1			0					1
16				1		1	1		1				1			0				1
17		1						1		1							0	1		
18		1		1	1					1							1	0		1
19			1				1				1			1					0	
20	1			1	1		1								1	1		1		0

Table 47. Scaled Abundances of Augmented Adjacency Matrix

control	PFOA	PFDA	clofribrate	DEHP
7.3673	5.5482	4.8498	7.5063	8.4661
7.3302	7.9393	9.6243	8.1605[a]	7.9133
6.9741	5.8108	7.6682	8.3517	0.4140
6.4914	5.0607	3.7292	3.9115	5.7209
6.0518	5.7563	2.5400	5.6065	7.0835
5.8654	9.8211	11.3078	8.4760	9.2165
5.7288	2.9942	1.9860	3.7337	3.9336
5.5569	1.3323	2.5893	2.3437	5.9634
5.0129	4.6517	4.2340	4.2970	4.7433
4.7770	4.2447	3.2119	4.0762	5.5823
4.5934	6.6009	5.7344	5.7494	6.0937
4.4263	3.6105	4.0160	5.3975	5.9344
4.3299	3.7671	2.3212	3.6700	4.9731
4.2100	3.7753	3.8004	4.3228	4.5189
4.1831	0.8236	0.8421	3.0661	7.6039
4.0836	3.9457	4.0865	3.8801	5.1462
4.0750	1.0606	0.6687	1.9810	2.7472
3.7149	3.8975	2.6921	2.8375	3.9518
3.6834	3.9798	3.0813	2.3889	3. 9869
3.5445	1.9311	0.8231	2. 9175	3.6375
102	86.55	79.81	92.67	107.62

[a] Underlined digit 5 represents rounding of smaller digit that follows to 5, and if fewer digits are used it should not be rounded to higher number.

the adjacency matrix of Table 46 and 1 998 615 is the sum of abundances of all 20 most abundant protein for the control group (listed in the "control" column in Table 42). The sum of the abundance indicates the overall increase or decrease of proteins in cells exposed to different conditions. As is visible from the last row of Table 47, the total abundance of proteins for polyfluorooctanic acid (PFOA), polyfluorodecanoic acid (PFDA), and clofibrate (for the first 20 proteins of the control group) has decreased, whereas for bis(2-ethylhexyl) phthalate (DEHP) it has somewhat increased.

16.5. Novel Map Descriptors

The leading eigenvalues have been used as molecular descriptors in view of having useful structural interpretation.[88,318,319] The accompanying leading eigenvector, which has no nodes (separate positive and negative regions), has hardly received attention when looked at alone. As is well known from MO calculations in Quantum Chemistry, the coefficients of populated eigenvectors were of interest for construction of atom charges and bond orders, but here we continue to confine attention only to the first eigenvector. The first eigenvectors of the augmented adjacency matrix of Voronoi regions for the five proteomics maps are labeled as Ψ_1 in Table 48.

A close look at Table 48 is instructive. First to observe is that all entries in the first column (belonging to the control group) are of similar magnitude, between 0.3262 and 0.1378. The other four columns of Table 48 show considerable variations of the coefficients of the first eigenvector. Thus, for PFOA, PFDA, clofibrate, and DEHP, the extreme values are (0.0171; 0.7827); (0.0061; 0.8768); (0.037; 0.5679); and (0.0603; 0.379), respectively. More interesting, however, is the *parallelism* between the respective columns of Tables 47 and 48, which we have illustrated for PFOA in Figure 85. This shows that the first eigenvector is a good descriptor for the abundance of proteins in a map and may be used for numerical estimates of the degree of similarity between proteomics maps.

In Figure 86 are plotted for visual inspection the relative (experimental) abundances for the five proteomics maps under consideration. From this figure, one sees that the effects

Table 48. First Eigenvector of Augmented Adjacency Matrices for the Five Proteomics Maps Considered

Ψ_1(A*)	Ψ_1(B*)	Ψ_1(C*)	Ψ_1(D*)	Ψ_1(E*)
0.2676	0.0506	0.0215	0.1379	0.3646
0.2717	0.0489	0.0196	0.0971	0.3790
0.3045[a]	0.2790	0.2973	0.5412	0.0444
0.2670	0.1004	0.0453	0.0937	0.1684
0.2209	0.0376	0.0084	0.0694	0.3721
0.2365	0.7827	0.8768	0.5679	0.0469
0.3162	0.1196	0.0766	0.2013	0.1502
0.1664	0.0225	0.0095	0.0389	0.2136
0.1915	0.2051	0.1721	0.2227	0.0602
0.2639	0.1211	0.0541	0.1354	0.2550
0.2210	0.3337	0.2214	0.3217	0.1417
0.1710	0.2018	0.0116	0.0688	0.3174
0.1916	0.0318	0.1343	0.1777	0.1242
0.1895	0.0970	0.0513	0.1394	0.1519
0.1379	0.0171	0.0061	0.0471	0.3813
0.2188	0.1979	0.1598	0.1904	0.1342
0.1378	0.0237	0.0085	0.0371	0.1199
0.2011	0.0252	0.0169	0.0649	0.1391
0.1484	0.1121	0.0692	0.1392	0.0698
0.2299	0.0610	0.0288	0.0991	0.2329

[a] Underlined digit 5 represents rounding of smaller digit that follows to 5, and if fewer digits are used it should not be rounded to higher number.

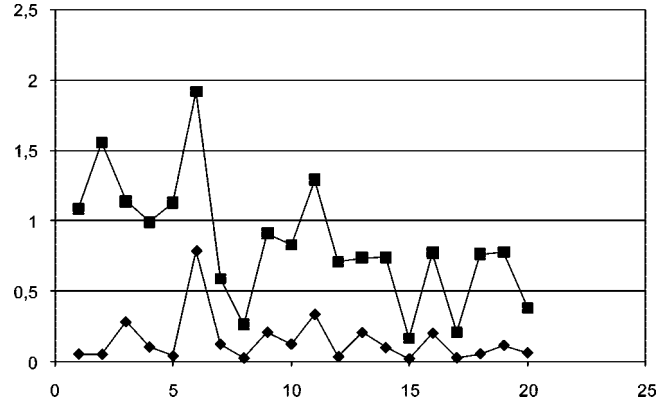


Figure 85. Plot of the first eigenvector for PFOA (bottom) and the experimental scaled abundances for 20 most abundant proteins of the control group.

of PFOA and PFDA on liver cell proteome are comparable, with the effects of clofibrate following, whereas the effect of DEHP appears to be most dissimilar. Such qualitative observations are confirmed by using the coefficients of the first eigenvector as map descriptors. The results are shown in Table 49, from which one sees that indeed the most similar proteomics maps are those belonging to PFOA and PFDA, with some similarity between these two proteomics maps and those of clofibrate.

It may be premature to speculate on the potential of this approach, but the results obtained clearly warrant further investigation along the lines outlined here.

17. Concluding Remarks

17.1. Challenges Ahead

Before we end this brief outline of currently available numerical procedures for obtaining quantitative characterizations of proteomics maps, we should point out that the same procedures are equally good for characterizing other kinds of 2D maps, such as are, for instance, 2D NMR maps. Recently, for example, this has been demonstrated by a chemometric approach to the quantification of structural identity/similarity of proteins in various G-CSF (granu-

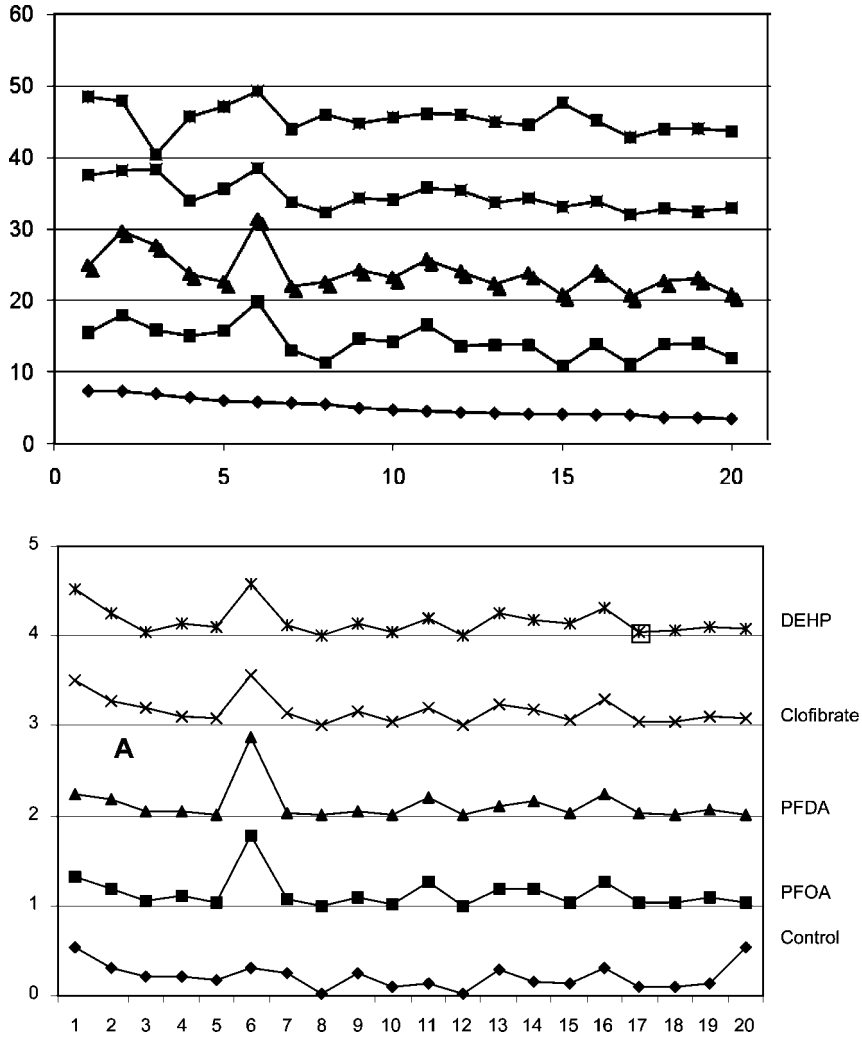


Figure 86. Experimentally reported relative abundance for 20 proteins of the five proteomic maps considered for illustration of a novel scheme for numerical characterization of proteomic maps.

Table 49. Similarity/Dissimilarity Matrix for the Five Proteomic Maps Considered

	control	PFOA	PFDA	clofibrate	DEHP
control	0	2.822	3.345	2.362	2.077
PFOA		0	0.832	1.025	3.952
PFDA			0	1.720	4.198
clofibrate				0	3.657
DEHP					0

locyte colony-stimulating factor) isolates.[320] The G-CSF isolates share the same amino acid sequence, but their preparation was carried out by different approaches. The comparison of 3D structures was made on the basis of 2D NMR nuclear Overhauser enhancement spectroscopy (NOESY). The analysis starts by constructing, for a collection of 200 peaks in 2D NMR, a graph obtained by linking peaks based on sequential nearest neighborhood. The analysis indicated that the pairwise similarity of new isolates prepared in the laboratory was larger than the similarity of any of the new isolates with the commercially available drugs, which was expected. This work illustrates an application of map invariants on the control of the material to be used in sensitive experiments.

Among the challenges lying ahead, we may mention the systematic search for alternative map invariants for the existing methodologies as well as the search for additional mathematical objects to be associated with maps. Recall that 25 years ago there were not many topological indices for characterization of molecular structure while today we have several hundreds, and their number continues to grow. It stands to reason that one can anticipate that a similar situation will happen with sequence invariants (used for characterization of DNA sequences, RNA secondary structures, and protein sequences) and map invariants to be used in proteomics studies. At the same time, it remains to better explore the limitations of current schemes, which incorporate a number of arbitrary choices, including the selection of proteins spots to be included in numerical analyses. One should continue to investigate how sensitive are characterizations of proteomics maps to the number of spots included and the number of lines incorporated in geometrical objects used, such as partial ordering or cluster graph. We can summarize the present situation in quantitative proteomics with the very same words that Winston Churchill (1874−1965) uttered in November 1942, halfway through WWII: "*Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning*."

So hopefully already in the near future one can expect a doubling of map invariants and doubling of the new methodologies—all designed to facilitate the use of computers in digesting the abundance of data on the abundance of proteins in proteomes.

17.2. Conclusions and Perspectives

In contrast to various graphical representations of molecules, which try to depict certain features of molecules, graphical representations of DNA and proteins may at first sight appear far from reality as they do not relate to the actual geometry associated with atomic groups and molecular fragments but are purely virtual. However, one could then also say that primary sequences of DNA or proteins are also devoid of reality, in that they tell nothing of the geometry of such systems. Despite their virtual nature, such artificial constructions, nevertheless, as already has been demonstrated with various graphical representations of DNA, allow one to search and quantify similarities and differences among biological sequences. We have shown in this review that the brief history of graphical representations of proteins appears to be equally promising and attractive in offering not only a visual alternative for displaying proteins but also numerical quantitative characterizations, and finally, graphical manipulations that offer novel routes to protein alignment. Finally, we should add that graphical approaches to DNA and proteins allow one to characterize a *single* DNA and a *single* protein sequence. This allows one to construct a library of DNA sequences, or protein sequences, which can be updated not only by inclusion of new biosequences but also by incorporating novel sequence invariants, as they become available with time. Thus, one would be able at will also to compare sequences and revise such comparisons if necessary. In contrast, computer programs, such as BLAST and FASTA, are offering insights on similarities and differences among DNA and proteins, two or more at the time, but do not relate to the characterization of a single biological sequence. This may be what interests most researchers, but graphical representations allow one also to catalogue and build libraries of DNA and proteins similar to molecular libraries that compile information on single molecules. In addition, graphical characterization and similarity analysis based on graphical approaches can be, and have been already, applied to secondary structures of RNA.

18. Acknowledgments

This work has been supported in part by the Ministry of Higher Education, Science and Technology of Republic of Slovenia, under Research Grant P1-017, and the Ministry of Science, Education and Sports of the Republic of Croatia through Project 098-0982929-2917. We also thank Dr. Marjan Vračko for construction of figures 27 and 28 and Rok Orel for construction of figures 83 and 84.

19. References

(1) Wilson, E. B. *Introduction to Scientific Research*; McGraw-Hill: New York, 1952.
(2) Kolbe's criticism of van't Hoff: http://www.chem.yale.edu/~chem125/125/history99/6Stereochemistry/Kolbe.html.
(3) From a letter of R. G. Parr to M. Randić: "... .. *In the very early 1930s the leading young American physical chemists got very upset by the disdain for quantum mechanics, molecular spectroscopy, and the like, shown toward such subjects by the editor of the then Journal of Physical and Colloid Chemistry (later JPC), a chap named Bancroft, and the fact that the chemical leadership in the country would not force a change. They {Harold Urey, Joe Mayer, John Kirkwood, Robert Mulliken, others} went to the American Physical Society, and there soon followed the founding of JCP. Take a look at issue 1, volume 1, of JCP, and you will [be] amazed by the high quality. It took forty years or so for JPC to recover (perhaps one should say "be born again")... ..*"
(4) Boys, S. F. *Proc. R. Soc. [London]* **1950**, *A200*, 542.
(5) Boys, S. F. *Proc. R. Soc. [London]* **1950**, *A201*, 125.
(6) Boys, S. F. *Proc. R. Soc. [London]* **1951**, *A206*, 489.
(7) Boys, S. F. *Proc. R. Soc. [London]* **1951**, *A207*, 181.
(8) Boys, S. F. *Proc. R. Soc. [London]* **1951**, *A207*, 197.
(9) Boys, S. F. *Philos. Trans. R. Soc.* **1952**, *A245*, 95.
(10) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *B136*, 864.
(11) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *A140*, 1133.
(12) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev.* **1988**, *B37*, 785.
(13) German alchemist and physician Theophrastus Bombastus von Hohenheim (1493−1541), known as Paracelsus, has recognized medical use of small amounts of toxic chemicals. For more on the history of hormesis, see: Stebbing, A. R. D. BELLE Newsletter 1997, *6*; No. 2, September.
(14) Randić, M.; Estrada, E. *J. Proteome Res.* **2005**, *4*, 2133.
(15) Randić, M. *MATCH Commun. Math. Comput. Chem.* **2008**, *59*, 5.
(16) Randić, M. *Indian J. Math. Teach.* **2000**, *26*, 11 (Unfortunately, galley proofs were not sent to the author, leaving the text with several confusing printing errors left uncorrected).
(17) Calabrese, E. J.; Baldwin, L. A. *Annu. Rev. Pharmacol. Toxicol.* **2003**, *43*, 175.
(18) Calabrese, E. J. *Cell. Mol. Biol.* **2005**, *51*, 643.
(19) Calabrese, E. J.; Baldwin, L. A. *Nature* **2003**, *421*, 691.
(20) As cited in John Horgan's *End of Science*: *Facing the Limits of Knowledge in the Twilight of the Scientific Age*; Addison Wesley: New York, 1996.
(21) Gernert, D. *J. Sci. Explor.* **2008**, *22*, 233.
(22) Planck, M. *Survey of Physical Theory*; Dover Publications: New York, 1993.
(23) Lipman, D. J.; Pearson, W. R. *Science* **1985**, *227*, 1435.
(24) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403.
(25) Bartlett, F. C. *Remembering: A Study in Experimental and Social Psychology*; Cambridge University Press: England, 1932; p 168.
(26) See Randić, M. *MATCH Commun. Math. Comput. Chem.* **2008**, *59*, 46.
(27) After being rejected by *J. Chem. Phys*. the "childish" paper: "Symmetry properties of chemical graphs. VI. Isomerization of octahedral complexes" was published in collaboration with M. I. Davis in *Int. J. Quantum Chem* 1984, *26*, 69.
(28) Part of the foreword in *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976.
(29) Randić, M. *Indian J. Chem.* **2003**, *42A*, 1207.
(30) Veltman, M. *Diagrammatica: The Path to Feynman Diagrams (Cambridge Lecture Notes in Physics)*; Cambridge University Press: Cambridge, U.K., 1995.
(31) Pauling, L. *The Nature of the Chemical Bond*, 3rd ed.; Cornell University Press: New York, 1972.
(32) Lecointre, G.; Le Guyader, H. *The Tree of Life: A Phylogenetic Classification*; Harvard University Press Reference Library: Cambridge, MA, 2006.
(33) Goldstine, H. H. *The Computer from Pascal to von Neuman*; Princeton University Press: Princeton, NJ, 1993; pp 266−267.
(34) König, D. *Theorie der endlichen und unendlichen Graphen*; Akademische Verlagsgesellschaft: Leipzig, 1936. English translation with commentary by W. T. Tutte, Birkhäuser: Basel, Switzerland, 1990.
(35) Harary, F. *Graph Theory*; Addison-Wesley: Reading, MA, 1969.
(36) Wilson, R. J. *Introduction to Graph Theory*; Oliver & Boyd: Edinburgh, U.K., 1972.
(37) Bondy, J. A.; Murty, U. S. R. *Graph Theory with Applications*; American Elsevier: New York, 1976.
(38) Pisanski, T.; Randić, M. In *Geometry at Work, Papers in Applied Geometry*; Gorini, C. A., Ed.; Math. Assoc. America No. 53: Washington, DC, 2000; pp 174−194.
(39) *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976.
(40) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
(41) Nandy, A.; Harle, M.; Basak, S. C. *ARKIVOC* **2006**, *9*, 211.
(42) Hamori, E.; Ruskin, J. *J. Biol. Chem.* **1983**, *258*, 1318.
(43) Hamori, E. *Nature* **1985**, *314*, 585.
(44) Hamori, E. *Biotechniques* **1989**, *7*, 710.
(45) Hamori, E.; Varga, G.; LaGuardia, J. J. *Comput. Appl. Biosci.* **1989**, *5*, 263.
(46) Hamori, E. In *Frontiers of Scientific Visualization*; Pickover, C. A., Tewksbury, S. K., Eds.; Wiley-Interscience: New York, 1994; pp 90−121.
(47) Randić, M. *SAR QSAR Environ. Res.* **2004**, *15*, 147.
(48) Gates, M. A. *J. Theor. Biol.* **1986**, *119*, 319.
(49) Leong, P. M.; Morgenthaler, S. *Comput. Appl. Biosci.* **1995**, *11*, 503.
(50) Nandy, A. *Curr. Sci.* **1994**, *66*, 309.
(51) Guo, X.; Randić, M.; Basak, S. C. *Chem. Phys. Lett.* **2001**, *350*, 106.
(52) Guo, X.; Nandy, A. *Chem. Phys. Lett.* **2003**, *369*, 361.
(53) Zhang, C.-T.; Zhang, R.; Ou, H.-Y. *Bioinformatics* **2003**, *19*, 593.
(54) Randić, M. *Chem. Phys. Lett.* **2004**, *386*, 468.
(55) Liao, B.; Tan, M.; Ding, K. *Chem. Phys. Lett.* **2005**, *402*, 380.
(56) Yau, S. S-T.; Wang, J.; Niknejad, A.; Lu, C.; Jin, N.; Ho, Y.-K. *Nucleic Acids Res.* **2003**, *31*, 3078.
(57) Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235.

(58) Randić, M.; Kleiner, A. F.; De Alba, L. M. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 277.
(59) Randić, M. *Int. J. Quantum Chem.* **1980**, *18* (S7), 187.
(60) *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997.
(61) Balaban, A. T.; Harary, F. *Tetrahedron* **1968**, *24*, 2505.
(62) Balaban, A. T. *Tetrahedron* **1969**, *25*, 2949.
(63) Anderson, P. G. In *Fibonacci Numbers and Their Applications*; Philippou, A. N., Bergum, P. G., Horadam, A. F., Eds.; D. Reidel Publishing Company: Dordrecht, The Netherlands, 1986; p 2.
(64) See p 3553 in Randić, M. *Chem. Rev.* **2003**, *103*, 3449.
(65) Balaban, A. T. *MATCH Commun. Math. Comput. Chem.* **1989**, *24*, 29.
(66) Zhu, H.; Klein, D. J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1067.
(67) Ivanciuc, O.; Ivanciuc, T.; Klein, D. J. *MATCH Commun. Math. Comput. Chem.* **2001**, *44*, 251.
(68) Spencer, J.; Szemerédy, E.; Trotter, W. T., Jr. In *Graph Theory and Combinatorics 1983: Cambridge Combinatorial Proceedings in Honour of Paul Erdos*; Bollobás, B., Ed.; Academic Press: London, 1984; pp 293−303.
(69) Erdös, P. *Am. Math. Mon.* **1946**, *53*, 248.
(70) Randić, M.; Zupan, J.; Vikić-Topić, D. *J. Mol. Graphics Model.* **2007**, *26*, 290.
(71) Randić, M.; Razinger, M. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1996; pp 159−236.
(72) Randić, M.; Vračko, M.; Novič, M.; Basak, S. C. *MATCH Commun. Math. Comput. Chem.* **2000**, *42*, 181.
(73) Bytautas, L.; Klein, D. J.; Randić, M.; Pisanski, T. *DIMACS Series Discovery Math. Theor. Comput. Sci.* **2000**, *51*, 39.
(74) Randić, M. *Int. J. Quantum Chem.* **1995**, *56* (S22), 61.
(75) Randić, M.; Krilov, G. *Int. J. Quantum Chem.* **1996**, *60* (S23), 1851.
(76) Randić, M.; Krilov, G. *Int. J. Quantum Chem.* **1997**, *65*, 1065.
(77) Randić, M.; Krilov, G. *Chem. Phys. Lett.* **1997**, *272*, 115.
(78) Randić, M.; Krilov, G. *Int. J. Quantum Chem.* **1999**, *75*, 1017.
(79) Krilov, G.; Randić, M. *New J. Chem.* **2004**, *28*, 1608.
(80) Randić, M.; Zupan, J.; Novič, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1339.
(81) Randić, M.; Witzmann, F.; Vračko, M.; Basak, S. C. *Med. Chem. Res.* **2001**, *10*, 456.
(82) Randić, M. In *Handbook of Proteomic Methods*; Conn, P. M., Ed.; Humana Press: Totowa, NJ, 2003; pp 429−450.
(83) Jeffrey, H. J. *Nucleic Acids Res.* **1990**, *18*, 2163.
(84) Jeffrey, H. J. *Comput. Graph.* **1992**, *16*, 25.
(85) Barnsley, M. F.; Rising, H. *Fractals Everywhere*, 2nd ed.; Academic Press: Boston, MA, 1993.
(86) Zupan, J.; Randić, M. *J. Chem. Inf. Model.* **2005**, *45*, 309.
(87) Randić, M. *Acta Chim. Slov.* **2006**, *53*, 477.
(88) Randić, M.; Zupan, J. *SAR QSAR Environ. Res.* **2004**, *15*, 191.
(89) Randić, M. *J. Math. Chem.* **2008**, *43*, 756.
(90) Randić, M.; Vračko, M.; Plavšić, D. *Int. J. Chem. Model.* **2008**, *1*, 305.
(91) Randić, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1330.
(92) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C. *Croat. Chem. Acta* **2004**, *77*, 345.
(93) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C. *J. Proteome Res.* **2004**, *3*, 778.
(94) Randić, M.; Novič, M.; Vračko, M. *J. Chem. Inf. Model.* **2005**, *45*, 1205.
(95) Zupan, J.; Randić, M. *J. Chem. Inf. Model.* **2005**, *45*, 309.
(96) Fiser, A.; Tusnady, G. E.; Simon, I. *J. Mol. Graph.* **1994**, *12*, 302.
(97) Yu, Z.-G.; Anh, V.; Lau, K.-S. *J. Theor. Biol.* **2004**, *226*, 341.
(98) Basu, S.; Pan, A.; Dutta, C.; Das, J. *J. Mol. Graphics Modell.* **1997**, *15*, 279.
(99) Solovyev, V. V. *Biosystems* **1993**, *30*, 137.
(100) Joseph, J.; Sasikumar, R. *BMC Bioinformatics* **2006**, *7*, 243.
(101) Parthasarathy, S. In *Book of Abstracts of Fifth International Conference on Dynamic Systems and Applications*, Atlanta, GA, May 30−June 2, 2007.
(102) Mandelbrot, B. B. *Fractals and Chaos. The Mandelbrot Set and Beyond*; Springer Verlag: New York, 2004.
(103) Mandelbrot, B. B. *The Fractal Geometry of Nature*; W. H. Freeman: New York, 1983.
(104) Sierpinski, W. *Compt. Rend. Acad. Sci. Paris* **1915**, *160*, 302.
(105) http://mathworld.wolfram.com/SierpinskiSieve.html.
(106) For history, see Pascal's triangle in Wikipedia (a free, web-based, collaborative, multilingual encyclopedia project supported by the nonprofit Wikimedia Foundation), http://en.wikipedia.org/wiki/Pascal's_triangle.
(107) Sloane, N. J. A. *The On-Line Encyclopedia of Integer Sequences*; http://www.research.att.com/~njas/sequences/Seis.html.
(108) See: Sloane, N. J. A. Home Page, http://www.research.att.com/~njas/.
(109) Gardner, M. *Mathematical Carnival*; Vintage Books: New York, 1977; pp 194−207.
(110) *Applications of Discrete Mathematics*, updated ed.; Michaels, J. G., Rosen, K. H., Eds.; McGraw-Hill: New York, 2007.
(111) Bondarenko, B. A. *Fibonacci Triangles and Pyramids, Their Fractals, Graphs, and Applications*; The Fibonacci Association: Santa Clara, CA, 1993.
(112) Gordon, M.; Davison, W. H. T. *J. Chem. Phys.* **1952**, *20*, 428.
(113) Balaban, A. T.; Tomescu, I. *Croat. Chem. Acta* **1984**, *57*, 391.
(114) Balaban, A. T.; Tomescu, I. *MATCH Commun. Math. Comput. Chem.* **1985**, *17*, 91.
(115) See p 3481 and Fig. 31 in Randić, M. *Chem. Rev* **2003**, *103*, 3449.
(116) Randić, M.; Morales, D. A.; Araujo, O. *J. Math. Chem.* **1996**, *20*, 79.
(117) Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. *Chem. Phys. Lett.* **2003**, *368*, 1.
(118) Randić, M.; Vračko, M.; Lerš, N.; Plavšić, D. *Chem. Phys. Lett.* **2003**, *371*, 202.
(119) Randić, M.; Zupan, J.; Vikić-Topić, D.; Plavšić, D. *Chem. Phys. Lett.* **2006**, *431*, 375.
(120) Graphical alignment is not confined only to spectral representations of proteins. It can be extended to chaos-game representations of DNA and proteins, as will be clarified in the section on representation of proteins via the "magic circle". The paper on graphical alignment of proteins was submitted, revised, and accepted for publication in 2006 by *J. Proteome Res.* At the stage of typsetting, objections were raised to citing a question raised by one of the reviewers, with request that this reference to anonymous reviewer be deleted. However, even before responding, an additional report (for an already accepted paper) came with request by the same reviewer who *insisted* on the removal of the figure showing 20 amino acids uniformly placed on the circumference of a unit circle because it was already published before. In addition, the anonymous reviewer wanted to continue the dialogue! Since the author did not want to remove the figure nor was he interested in private dialogue with an anonymous scientist, the manuscript was withrawn and resubmitted to *J. Math. Chem*., where it appeared in *J. Math. Chem.* **2008**, *43*, 756 (unchanged as accepted by *J. Proteome Res*.).
(121) http://www.bioweb.uwlax.edu/GenWeb/Molecular/Seq_Anal/Alignment/alignment.htm.
(122) Liao, B.; Chen, W.; Sun, X.; Zhu, W. *J. Comput. Chem.* **2009**, *30*, 2205.
(123) Liao, B.; Wang, T.-M. *J. Biomol. Struct. Dyn.* **2004**, *21*, 827.
(124) Zhan, J.; Liao, B.; Zhang, Y. Internet Electronic Conference of Molecular Design 2003, November 23−December 6, 2003,http://biochempress.com/Files/IECMD_2004/IECMD_2004_018.pdf.
(125) Bai, F.; Zhu, W.; Wang, T. *Chem. Phys. Lett.* **2005**, *408*, 258.
(126) Feng, J.; Wang, T.-M. *Chem. Phys. Lett.* **2008**, *454*, 355.
(127) Liu, L.; Wang, T. *J. Math. Chem.* **2007**, *42*, 595.
(128) Yao, Y.-H.; Nan, X.-Y.; Wang, T.-M. *J. Comput. Chem.* **2005**, *26*, 1339.
(129) Yao, Y.-H.; Liao, B.; Wang, T.-M. *J. Mol. Struct: THEOCHEM* **2005**, *755*, 131.
(130) Li, C.; Xing, L.; Wang, X. *Chem. Phys. Lett.* **2008**, *458*, 249.
(131) Zhu, W.; Liao, B.; Ding, K. *J. Mol. Struct: THEOCHEM* **2005**, *757*, 193.
(132) Zhang, Y.; Qui, J. Q.; Su, L. Q. *Chem. Phys. Lett.* **2008**, *458*, 180.
(133) Randić, M.; Vračko, M.; Novič, M.; Plavšić, D. *Int. J. Quantum Chem.* **2009**, *109*, 2982.
(134) Reusken, C. B. E. M.; Bol, J. F. *Nucleic Acids Res.* **1996**, *24*, 2660.
(135) Koper-Zwarthoff, E. C.; Bol., J. F. *Nucleic Acids Res.* **1980**, *8*, 3307.
(136) Cornelissen, B. J. C.; Janssen, H.; Zuidema, D.; Bol., J. F. *Nucleic Acids Res.* **1984**, *12*, 2427.
(137) Houser-Scott, F.; Bear, M. L.; Liem, K. F.; Cai, J. M.; Gehrke, L. N. *J. Virol.* **1994**, *68*, 2194.
(138) Scott, S. W.; Ge, X. *J. Gen. Virol.* **1995**, *76*, 957.
(139) Scott, S. W.; Ge, X. *J. Gen. Virol.* **1995**, *76*, 1801.
(140) Balaban, A. T.; Plavšić, D.; Randić, M. *Chem. Phys. Lett.* **2003**, *379*, 147.
(141) Randić, M. *Chem. Phys. Lett.* **2000**, *317*, 29.
(142) Randić, M.; Novič, M.; Vračko, M. *SAR QSAR Environ. Res.* **2008**, *19*, 339.
(143) These matrices are in general nonsymmetric. Nevertheless, one can construct additional 8-component vectors as supplementary characterization of RNA by using amplitudes of complex numbers as vector entries.
(144) Shu, W.; Bo, X.; Zheng, Z.; Wang, S. *BMC Bioinformatics* **2008**, *9*, 188.
(145) Yao, Y.-H.; Nan, X.-Y.; Wang, T.-M. *J. Comput. Chem.* **2005**, *26*, 1339.
(146) Liu, L.; Wang, T. *J. Math. Chem.* **2007**, *42*, 595.
(147) Randić, M.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532.

(148) Randić, M.; Plavšić, D. *Chem. Phys. Lett.* **2009**, *476*, 277.
(149) Randić, M.; Vračko, M.; Novič, M.; Plavšić, D. Uniform representation of proteins by distance matrix. In *Quantuim Frontiers of Atoms and Molecules*; Putz, M. V., Ed.; Nova Publishers: New York, 2009.
(150) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C.; Balaban, A. T. *Chem. Phys. Lett.* **2005**, *407*, 205.
(151) Randić, M.; Mehulić, K.; Vukičević, D.; Pisanski, T.; Vikić-Topić, D.; Plavšić, D. *J. Mol. Graphics Modell.* **2009**, *27*, 637.
(152) Randić, M. *Chem. Phys. Lett.* **2004**, *386*, 468.
(153) Randić, M. *Chem. Phys. Lett.* **2007**, *444*, 176.
(154) Crick, F. H. C.; Barnett, L.; Brenner, S.; Watts-Tobin, R. J. *Nature* **1961**, *192*, 1227.
(155) Fickett, J. W. *Nucleic Acids Res.* **1982**, *10*, 5303.
(156) Balaban, A. T.; Randić, M. In *Proceedings of the International Conference on Biocomputation, Bioinformatics, and Biomedical Technologies*, June 29−July 5, 2008, Bucharest, Romania; pp 76−82. IEEE Trans.
(157) Randić, M. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; pp 3018−3032.
(158) Balaban, A. T. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp 403−453.
(159) Basak, S. C.; Grunwald, G. D.; Niemi, G. J. In *From Chemical Topology to Three-Dimensional Geometry*; Plenum Press: New York, 1977; pp 73−116.
(160) Estrada, E. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp 403−453.
(161) Randić, M.; Zupan, J. In *Topology in Chemistry: Discrete Mathematics of Molecules*; Rouvray, D. H., King, R. B., Eds.; Horwood Publ. Ltd.: Chichester, U.K., 2002; pp 249−291.
(162) Randić, M. *J. Mol. Graphics Modell.* **2001**, *20*, 19.
(163) Randić, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 607.
(164) Randić, M. *New J. Chem.* **1997**, *21*, 1078.
(165) Hardy, G. H. *A Course of Pure Mathematics*, 10th ed.; Cambridge University Press: Cambridge, U.K., 2002.
(166) Randić, M.; Zupan, J.; Balaban, A. T. *Chem. Phys. Lett.* **2004**, *397*, 247.
(167) Randić, M.; Vikić-Topić, D.; Graovac, A.; Lerš, N.; Plavšić, D. *Period. Biol.* **2005**, *107*, 437.
(168) Randić, M.; Balaban, A. T.; Novič, M.; Založnik, A.; Pisanski, T. *Period. Biol.* **2005**, *107*, 403.
(169) Randić, M.; Butina, D.; Zupan, J. *Chem. Phys. Lett.* **2006**, *419*, 528.
(170) Randić, M.; Zupan, J.; Novič, M.; Gute, B. D.; Basak, S. C. *SAR QSAR Environ. Res.* **2002**, *13*, 689.
(171) Randić, M.; DeAlba, L. M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1078.
(172) Randić, M. *Chem. Phys. Lett.* **2008**, *456*, 84.
(173) Read, R. C.; Corneil, D. G. *J. Graph Theory* **1977**, *1*, 339.
(174) Gatti, G. *J. Graph Theory* **1979**, *3*, 95.
(175) Randić, M. *J. Chem. Phys.* **1974**, *60*, 3920. addendum *J. Chem. Phys.* 1975, *62*, 309.
(176) Randić, M. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105.
(177) Balaban, A. T.; Liu, X.; Klein, D. J.; Babić, D.; Schmalz, T. G.; Seitz, W. A.; Randić, M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 396.
(178) Balaban, A. T.; Farcasiu, D.; Bănică, R. *Rev. Roum. Chim.* **1966**, *11*, 1205.
(179) Julius Petersen (1839−1910), Danish mathematician.
(180) Petersen, J. *L'Intermédiaire des Mathématiciens* **1898**, *5*, 225.
(181) Holton, D. A.; Sheehan, J. *The Petersen Graph, Australian Math. Soc. Lecture Series No.7*; Cambridge University Press: Cambridge, U.K., 1993.
(182) Dunitz, J. D.; Prelog, V. *Angew. Chem.* **1968**, *80*, 700.
(183) Randić, M. *Croat. Chem. Acta* **1977**, *49*, 643.
(184) Prelog, V. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976.
(185) Zaretskij, K. A. *Usp. Mat. Nauk* **1965**, *20*, 90.
(186) We would like to emphasize that Platt had a vision and foresight to realize the potential of path numbers as molecular descriptors. Very many topological indices are related to path numbers and represent their variations using various weighting schemes, e.g., the connectivity index[190] and the higher-order connectivity indices.[191]
(187) Platt, J. R. *J. Chem. Phys.* **1947**, *15*, 419.
(188) Platt, J. R. *J. Phys. Chem.* **1952**, *56*, 328.
(189) Randić, M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 57.
(190) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(191) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. *J. Pharm. Sci.* **1976**, *65*, 1226.
(192) Kier, L. B.; Hall, H. L. *J. Pharm. Sci.* **1976**, *65*, 1806.
(193) Hall, H. L.; Kier, L. B.; Murray, W. J. *J. Pharm. Sci.* **1975**, *64*, 1974.
(194) Murray, W. J.; Hall, L. H.; Kier, L. B. *J. Pharm. Sci.* **1975**, *64*, 1978.
(195) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(196) Randić, M.; Razinger, M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 594.
(197) Randić, M. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373.
(198) Randić, M. *New J. Chem.* **1995**, *19*, 781.
(199) Randić, M. *J. Math. Chem.* **1996**, *19*, 375.
(200) Randić, M.; Razinger, M. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1977; pp 159−236.
(201) Jacques Hadamard (1865−1963), French mathematician.
(202) Issai Schur (1875−1941), German/Jewish mathematician.
(203) Leopold Kronecker (1823−1891), German mathematician.
(204) Gilat, A. *MATLAB: An Introduction with Applications*, 3rd ed.; Wiley & Sons: New York, 2008. MATLAB is the acronym for "matrix laboratory". It was invented by Cleve Moler at the Computer Science Department of the University of New Mexico in the late 1970s. For more information, consult Wikipedia.
(205) *Funk & Wagnalls Standard Desk Dictionary*; Harper & Row Publishers: New York, 1984.
(206) Orengo, C. A.; Brown, N. P.; Taylor, W. T. *Proteins* **1992**, *14*, 139.
(207) Holm, L.; Sander, C. *J. Mol. Biol.* **1993**, *233*, 123.
(208) Notredame, C.; Holm, L.; Higgins, D. G. *Bioinformatics* **1998**, *14*, 407.
(209) Vriend, G.; Sander, C. *Proteins* **1991**, *11*, 52.
(210) Fisher, D.; Bachar, O.; Nussinov, R.; Wolfson, H. *J. Biomol. Struct. Dyn.* **1992**, *9*, 769.
(211) Shindyalov, I. N.; Bourne, P. E. *Protein Eng. Des. Sel.* **1998**, *11*, 739.
(212) Alexandrov, N. N. *Protein Eng. Des. Sel.* **1996**, *9*, 727.
(213) Alexandrov, N. N.; Fisher, D. *Proteins* **1996**, *25*, 354.
(214) Levenshtein, V. I. *Sov. Phys. Dokl.* **1966**, *10*, 707.
(215) *Handbook of Chemoinformatics—From Data to Knowledge in 4 Volumes*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 3.
(216) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389.
(217) Pearson, W. R.; Lipman, D. J. *Proc. Natl. Acad. Sci.* **1988**, *85*, 2444.
(218) Martin, T. C. *The Inventions, Researches and Writings of Nikola Tesla*; Barnes & Noble Books: New York, 1995.
(219) Comment by an anonymous reviewer.
(220) Novič, M.; Randić, M. *SAR QSAR Environ. Res.* **2008**, *19*, 317.
(221) Hashimoto, Y.; Niikura, T.; Tajima, H.; Yasukawa, T.; Sudo, H.; Ito, Y.; Kita, Y.; Kawasumi, M.; Kouyama, K.; Doyu, M.; Sobue, G.; Koide, T.; Tsuji, S.; Lang, J.; Kurokawa, K.; Nishimoto, I. *Proc. Natl. Acad. Sci.* **2001**, *98*, 6336.
(222) Jaklič, G.; Pisanski, T.; Randić, M. *J. Comput. Biol.* **2006**, *13*, 1558.
(223) Jaklič, G.; Pisanski, T.; Randić, M. *MATCH Commun. Math. Comput. Chem.* **2007**, *58*, 301.
(224) Randić, M.; Zupan, J.; Pisanski, T. *J. Math. Chem.* **2008**, *43*, 674.
(225) Otto Topliz (1881−1940), German mathematician.
(226) Hermann Hankel (1839−1873), German mathematician.
(227) Brookes, M. *The Matrix Reference Manual*, http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html.
(228) Hladnik, M.; Marušič, D.; Pisanski, T. *Discrete Math.* **2002**, *244*, 137.
(229) Hladnik, M. *Linear Algebra Appl.* **1999**, *286*, 261.
(230) Davis, P. J. *Circulant Matrices*; Wiley: New York, 1979.
(231) Putnam, C. R. *Pac. J. Math.* **1964**, *14*, 651.
(232) The form of the matrix will depend on ordering of amino acids, but because we are interested in matrix invariants, any ordering is acceptable. We will continue to order amino acids alphabetically based on three-letter amino acid codes.
(233) Randić, M.; Novič, M.; Vračko, M. *SAR QSAR Environ. Res.* **2008**, *12*, 339.
(234) Chou, K.-C. *J. Protein Chem.* **1999**, *18*, 473.
(235) Liu, W.-m.; Chou, K.-C. *Protein Eng. Des. Sel.* **1999**, *12*, 1041.
(236) Tsai, H.-H.; Gunasekaran, K.; Nussinov, R. *Structure* **2006**, *14*, 1059.
(237) Dill, K. A. *Biochemistry* **1985**, *24*, 1501.
(238) Chan, H. S.; Dill, K. A. *Macromolecules* **1989**, *22*, 4559.
(239) Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986.
(240) Shih, C. T.; Su, Z. Y.; Gwan, J. F.; Hao, B. L.; Hsieh, C. H.; Lee, H. C. *Phys. Rev. Lett.* **2000**, *84*, 386.
(241) Li, H.; Helling, R.; Tang, C.; Wingreen, N. S. *Science* **1996**, *273*, 666.
(242) Wang, B.; Yu, Z.-G. *J. Chem. Phys.* **2000**, *112*, 6084.
(243) Wolynes, P. G. *Nat. Struct. Biol.* **1997**, *4*, 871.
(244) Li, H.; Tang, C.; Wingreen, N. S. *Phys. Rev. Lett.* **1997**, *79*, 765.
(245) Shakhnovich, E.; Gutin, A. *J. Chem. Phys.* **1990**, *93*, 5967.
(246) Klimov, D. K.; Thirumalai, D. *Phys. Rev. Lett.* **1996**, *76*, 4070.
(247) Sali, A.; Shahknovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.
(248) Hückel, E. *Z. Phys.* **1930**, *60*, 423.
(249) Hückel, E. *Z. Phys.* **1931**, *70*, 204.
(250) Hückel, E. *Z. Phys.* **1932**, *76*, 628.

(251) Bloch, F. *Z. Phys.* **1928**, *52*, 555.
(252) It is not widely known that E. Hückel was a theoretical physicist and his seminal work on HMO applied to benzene was part of his Magister thesis.
(253) Riddle, D. S.; Santiago, J. V.; Bray-Hall, S.; Doshi, N.; Grantcharova, V. P.; Yi, Q.; Baker, D. *Nat. Struct. Biol.* **1997**, *4*, 805.
(254) Wang, J.; Wang, W. *Nat. Struct. Biol.* **1999**, *6*, 1033.
(255) Randić, M.; Vračko, M.; Novič, M.; Plavšić, D. *SAR QSAR Environ. Res.* **2009**, *20*, 415.
(256) According to Wikipedia, "*Tree of life (science), a metaphor used first by Charles Darwin and subsequently to describe the interrelatedness of all living things through evolution.*"
(257) Yu, Z.-G.; Anh, V.; Lau, K.-S. *Physica A* **2004**, *337*, 171.
(258) Yang, J.-Y.; Yu, Z.-G.; Anh, V. *Chaos Solitons Fractals* **2009**, *40*, 607.
(259) Li, T.; Fan, K.; Wang, J.; Wang, W. *Protein Eng. Des. Sel.* **2003**, *16*, 323.
(260) Wang, J.; Wang, W. *Phys. Rev. E.* **2000**, *61*, 6981.
(261) Dokholyan, N. V. *Proteins* **2004**, *54*, 622.
(262) The characterization of such work as "academic exercise" originates with an anonymous referee of one of our manuscripts.
(263) Janin, J. *Nature* **1979**, *277*, 491.
(264) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. *Biochemistry* **1981**, *20*, 849.
(265) Kyte, J.; Doolite, R. F. *J. Mol. Biol.* **1982**, *157*, 105.
(266) Rose, G.; Geselowitz, A.; Lesser, G.; Lee, R.; Zehfus, M. *Science* **1985**, *229*, 834.
(267) http://cds.unina.it/petrilli/aminoac/amiac.htm?htm.
(268) Nandy, A.; Gosh, A.; Nandy, P. *In Silico Biol.* **2009**, *9*, 0008.
(269) Liu, N.; Wang, T. *Chem. Phys. Lett.* **2007**, *435*, 127.
(270) Yu, Z.-G.; Anh, V.; Lau, K.-S. *J. Theor. Biol.* **2004**, *226*, 341.
(271) Brown, T. A. *Genetics, A Molecular Approach*, 3rd ed.; Chapman and Hall: London, 1998.
(272) Bai, F.; Wang, T. *Chem. Phys. Lett.* **2005**, *413*, 458.
(273) Bai, F.; Wang, T. *J. Biomol. Struct. Dyn.* **2006**, *23*, 537.
(274) Yang Y. Wang, T. Private communication.
(275) Guo, Y.; Wang, T.-m. *J. Biomol. Struct. Dyn.* **2008**, *26*, 367.
(276) Sternberg, M. J. E.; Thornton, J. M. *J. Mol. Biol.* **1977**, *110*, 269.
(277) Yuan, C.; Liao, B.; Wang, T.-M. *Chem. Phys. Lett.* **2003**, *379*, 412.
(278) Chew, L. P.; Kedem, K. *Algorithmica* **2003**, *38*, 115.
(279) González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. *Curr. Top. Med. Chem.* **2007**, *7*, 1015.
(280) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Utiarte, E. *Proteomics* **2008**, *8*, 750.
(281) González-Díaz, H.; Prado-Prado, F. J. *J. Comput. Chem.* **2008**, *29*, 656.
(282) González-Díaz, H.; Agüero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Podda, G. *J. Comput. Chem.* **2007**, *28*, 1049.
(283) González-Díaz, H.; Saíz-Urra, L.; Molina, R.; González-Díaz, Y.; Sánchez-González, A. *J. Comput. Chem.* **2007**, *28*, 1042.
(284) González-Díaz, H.; Saíz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. *J. Proteome Res.* **2007**, *6*, 904.
(285) González-Díaz, H.; Peréz-Castillo, Y.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1990.
(286) González-Díaz, H.; Prado-Prado, F.; Ubeira, F. M. *Curr. Top. Med. Chem.* **2008**, *8*, 1676.
(287) Randić, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1330.
(288) Randić, M.; Zupan, J.; Novič, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1339.
(289) Randić, M.; Witzmann, F.; Vračko, M.; Basak, S. C. *Med. Chem. Res.* **2001**, *10*, 456.
(290) Randić, M.; Vračko, M.; Novič, M. *J. Proteome Res.* **2002**, *1*, 217.
(291) Randić, M. *Int. J. Quantum Chem.* **2002**, *90*, 848.
(292) Randić, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 983.
(293) Randić, M.; Novič, M.; Vračko, M.; Plavšić., D. *J. Theor. Biol.* **2010**, *266*, 21.
(294) Bajzer, Ž.; Randić, M.; Plavšić, D.; Basak, S. C. *J. Mol. Graphics Modell.* **2003**, *22*, 1.
(295) Randić, M.; Lerš, N.; Plavšić, D.; Basak, S. C. *J. Proteome Res.* **2004**, *3*, 778.
(296) Randić, M.; Novič, M.; Vračko, M. *J.Chem. Inf. Model.* **2005**, *45*, 1205.
(297) The approaches based on Voronoi regions can be further generalized by considering boundaries that will reflect the relative weights (amplitudes) of the protein spots.
(298) Randić, M.; Pompe, M.; Mills, D.; Basak, S. C. *Molecules* **2004**, *9*, 1177.
(299) Balaban, A. T. In *QSPR/QSAR Studies by Molecular Descriptors*; Diudea, M. V., Ed.; Nova Sci. Publ.: Huntington, NY, 2001; pp 1−30.
(300) Randić, M. *J. Math. Chem.* **1992**, *9*, 97.
(301) Balaban, A. T. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp 403−453.
(302) Katritzky, A. R.; Lobanov, V.; Karelson, M. *CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis*); University of Florida: Gainesville, FL, 1994.
(303) Karelson, M. *Molecular Descriptors*; Wiley: New York, 2000.
(304) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics, Vol. I Alphabetical Listing/Vol. II Appendices, References (Methods and Principles in Medicinal Chemistry*; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH: Weinheim, Germany, 2009; Vol. 41.
(305) Randić, M.; DeAlba, L. M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1078.
(306) Dijkstra, E. W. *Numer. Math.* **1959**, *1*, 269.
(307) With courtesy of Witzmann, F. *Molecular Anatomy Laboratory, Department of Biology*; Indiana University & Purdue University: Columbus, IN 47203.
(308) Anderson, N. L.; Esquer-Blasco, R.; Richardson, F.; Foxworthy, P.; Eacho, P. *Toxicol. Appl. Pharmacol.* **1996**, *137*, 75.
(309) MATLAB (abbreviation for Matrix Laboratory) is a product of The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098.
(310) Georgy Voronoi (1868−1908), Russian mathematician. His first paper (on factorization of polynomials) appeared while he was in high school.
(311) Voronoi, G. *J. Reine Angew. Math.* **1907**, *133*, 97.
(312) Boris Delaunay (1890−1980), Russian mathematician, a student of G. Voronoi. The Delaunay triangulation was introduced in 1934.
(313) Delaunay, B. *Izvestia Akad. Nauk SSSR, Otdel. Mat. Estest. Nauk* **1934**, *7*, 793.
(314) Kowalski, B. R.; Bender, C. F. *J. Am. Chem. Soc.* **1972**, *94*, 5632.
(315) Brualdi, R. C.; Cvetković, D. *A Combinatorial Approach to Matrix Theory and Its Applications*; Chapman & Hall/CRC Press: Boca Raton, FL, 2009.
(316) Babić, D.; Klein, D. J.; Lukovits, I.; Nikolić, S.; Trinajstić, N. *Int. J. Quantum Chem.* **2002**, *90*, 166.
(317) Chan, O.; Lam, T. K.; Merris, R. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 762.
(318) Lovász, L.; Pelikán, J. *Period. Math. Hung.* **1973**, *3*, 175.
(319) Randić, M.; Guo, X. F.; Bobst, S. *DIMACS Series Disc. Math. Theor. Comput. Sci.* **2000**, *51*, 305.
(320) Župerl, Š.; Pristovšek, P.; Menart, V.; Gaberc-Porekar, V.; Novič, M. *J. Chem. Inf. Model.* **2007**, *47*, 737.

CR800198J