CROATICA CHEMICA ACTA CCACAA **70** (3) 847–861 (1997)

ISSN–0011-1643
CCA–2459

*Original Scientific Paper*

# Sequence Analysis by Set-Spectrum Algorithms

*Lech Schulz*

*Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznan, Poland*

Received October 18, 1996; revised February 27, 1997; accepted April 9, 1997

An extended viewpoint of biochemical structure has been applied in the structural comparison beyond anticipating visualization and alignment. The self-organization features of biopolymers, prespecified by the underlying molecular components, were reflected in set-theoretical representations (called set spectra) to enrich the structure recognition. Then, the combinations of factors of similarity and dissimilarity involved in special metrics were effectively used in practical algorithms for a resultant, quantitative comparison of the varieties of properties. The nonstatistical, alternative algorithms used allow one to assess precisely the relationship between the distinguished properties of sequences and physical phenomena. In the work, the 5*S* rRNA families of sequences were tested with respect to their nonvisualized (in general), mathematically determined properties, and then correlations with biological systems were elicited. The method used has many advantages as an alternative approach to the research on the mechanisms of life, diseases, mutations, genetic code problems, *etc.*

## INTRODUCTION

In biochemistry, there is a common approach to the structures of biopolymers: simple visualization, *e.g.* by effective computer graphics techniques. This simplest method has been characterized in Ref. 1 as »just look at it«. In this way, the essential information encoded in the self-organization structure can be lost since it is well known in mathematics that only a small portion of structures can be geometrically visualized. For example, the spatial double helix model is regarded in Ref. 2 as the result of »self-organization through inherent properties«. Such a model is shaped by mutual distances

and angles of the component species. On the other hand, the features grasped from some representations of the labeled sequences[3] cannot be exhibited or concluded in this way. But, in spite of their abstract character which is far beyond the observational biochemistry, they can be hypothesized to influence biological phenomena. A glance at the existing methods of biopolymer comparison was helpful in the choice of tools to elucidate this intriguing relationship in a fully quantitative way. One can simply consider whole sequences with their ordering of component units. Also, single or repeating fragments such as those included in the Gnomic Dictionary[4] are of importance for the study of biopolymers. In Ref. 5, the dinucleotide, trinucleotide and codon usages are employed for sequence analysis. Comparing sequences means to compare their properties »derived« from basic structures. The above-mentioned ones pertain to the primary structures of biopolymers which prespecify (to some extent) the so called secondary and tertiary structures. Listing the desired properties allows first step estimations of similarity relationships between two or more sequences. A more sophisticated way consists in calculating the so called evolutionary distance,[1,5–7] which is determined at the minimal cost of altering one sequence into another by making insertions, deletions or replacements. The cost of alterations depends on associated alignments and on weighting single or block indels.[1,7] In these cases, the idea of evolutionary distance generally underlies all the more sophisticated concepts of sequence comparison. Also, probabilistic and statistical tools are usually involved in such considerations.[8–10] In Ref. 11, the sequences are compared »by totalling the number of matching paired characters under every possible alignment«. Statistical methods to complete the research are additionally used in this work. The »distance« between two sequences is estimated by the use of a complex formula.[12] Statistical trends are observed in Refs. 13 and 14 to perform linguistic analysis of nucleotide sequences. Except for the theoretically sophisticated methods, a simple estimation by the so called percentage of homology is usually used by biochemists. The fraction of the same nucleotides (or other units) for identical positions in the sequences compared is calculated in this case. There is an interesting problem addressed in Ref. 5 as to whether the nucleotide order of two sequences contributes essentially to their similarity or whether it can be explained by dinucleotide or codon usages. Moreover, Beyer's call[15] for assessments via metrics with clearly defined relationships (from the viewpoint of interpretations) between metric values and behaviour of their arguments can be noted. The latter two ideas have consistently been involved in the theoretical alternative solution of structure comparison in biochemistry. To illustrate some applications, the diversification of plants[16] was taken into account as one of the most astonishing facts of life. This is, in principle, regarded as a non-random phenomenon as well as natural biopolymer structural features.[17] But, the non-random means »subordinate to the laws«. Hence, using mathematical methods to search for structural and taxonomic relevances can be of great scientific interest.

## THEORETICAL ASPECTS

There is an underlying set $\mathscr{X}$ of biopolymers which is partitioned in two ways by $F$ and $P$ (*i.e.*, a partitioning $P$ is a set of disjoint subsets of $\mathscr{X}$ such that the union of the elements in $P$ is $\mathscr{X}$); $r_F$ and $r_P$ are the corresponding equivalence relations viewed as subsets of the Cartesian product $\mathscr{X} \times \mathscr{X}$ (*i.e.*, $r_P$ is the set of all $(x,y) \in \mathscr{X} \times \mathscr{X}$ such that there is a member of $P$ which contains $\{x,y\}$ as a subset).

In order to compare two finite partitions of natural biopolymers $F$ and $P$ it is enough to know the relevant factors of similarity and dissimilarity,[3,18] *i.e.*,

$$FS(r_F, r_P) = |r_F \cap r_P| \tag{1}$$

$$FD(r_F, r_P) = |r_F \Delta r_P| = |r_F| + |r_P| - 2\,FS(r_F, r_P).$$

The fraction of dissimilarities is then

$$\rho(r_F, r_P) = \frac{FD(r_F, r_P)}{FS(r_F, r_P) + FD(r_F, r_P)}. \tag{2}$$

One of the partitions (say $P$) is assumed to be a given pattern (*e.g.* a standard taxonomical classification) and the second ($F$) originates from structural data. Since pattern $P$ is well known, partition $F$ remains to be constructed. To this aim, the necessary set spectra (defined in Ref. 3) of labeled sequences have been listed here in a way suitable for programming. They are the following:

1. Length. The simplest set spectrum closely corresponds to the number of component units in a sequence. For example, a biopolymer comprising $n$ component units has a set to compare, *i.e.* its set spectrum called length consists of numbers 0, ..., $n$–1.

2. Component species. In this case, species of component units in the sequence are considered. Different kinds of aminoacids or nucleotides can be taken into account. For example, the sequence AATCCA has its set spectrum of this kind consisting exactly of A,T,C elements.

3. Common pairs. This is a special case of 1-subsequences described in section 4.

4. $m$-Subsequences. In this item, all the subsequences of $m$ consecutive component units along with their positions (marked by natural numbers) are considered. In particular, the set spectrum for 1-subsequences with respect to sequence ACTTAG consists of pairs (1,A), (2,C), (3,T),..., for 2-subsequences of ((1,A),(2,C)), ((2,C),(3,T)), ((3,T),(4,T)), ... .

5. Cumulative $m$-subsequences. This set spectrum includes all the possible $m$-subsequences that can be distinguished for a given sequence. Thus, the parameter $m$ varies from $m = 1$ to $m = L$ where $L$ is the length of the sequence.

6. Frequencies of $m$-subsequences. If subsequences are distinguished for a fixed $m$, then their frequencies in a given sequence are of interest. For example, in the oligonucleotide TTACATTAC the 3-subsequences TTA and TAC occur twice, others only once. The set to compare consists of pairs (TTA,0), (TTA,1), (TAC,0), (TAC,1), (ACA,0), ... . It can be observed that this set spectrum, though relevant, is not yet useful to grasp the more sophisticated problems of synonymic codons and »codon dialects«.[19]

7. Frequency cumulations of $m$-subsequences. The set to compare includes all the frequencies of $m$-subsequences for m varying between 1 and $L$.

8. Disjoint $m$-subsequences. The $m$-subsequences mentioned above are not of interest now. For example, the sequence ATCGGTGACC has the set of disjoint 3-subsequences consisting of ATC, GGT, GAC triples which are under consideration in this item. Such set spectra have been taken into account for m up to the length of the given sequence.

9. Frequencies of disjoint $m$-subsequences. Like in the case of $m$-subsequences in question, the number of appearences of disjoint $m$-subsequences distinguished can be considered. Such a set spectrum involves data on the frequencies of subsequences selected in the manner shown in section 6.

10. Set-spectra cumulations. This option cumulates the set spectra chosen in a standard way.

The objects compared (*e.g.* biopolymers) are taken as arguments of the described functions. The function values are special sets (called set spectra for short) which are the same if arguments are isomorphic. However, this property does not exhaust the sense of the notion of set spectrum mapping since the intention was not only to reflect set-theoretical properties of the objects considered but also to construct sets to compare. For example, a real number relevant to an object is not a set to compare in general. But the partitions mentioned or a finite Von-Neumann ordinal expressing the quantity of elements in a finite set can be used for comparison. The similarity set of two finite, nonempty sets to compare is taken to be the common part of these sets. Their so-called symmetric difference constitutes a dissimilarity set. Thus, properties of sequences expressed by set spectra are not at once involved in a function estimating a similarity relationship. In general, if a and b are finite and nonempty sets, then the cardinality of similarity set is denoted *FS(a,b)* and is called the factor of similarity. The cardinality of the dissimilarity set is denoted *FD(a,b)* and is called the factor of dissimilarity. The concept of mathematical similarity and dissimilarity can be enriched by the use of factors with tolerances quoted at the end of this paper. Owing to the use of factors, properties of sequences reflected in set spectra are not immediately involved in a function, which could result in a loss of clarity in the similarity-dissimilarity estimations. It could be noted that neither of the factors considered separately gives sufficient characterization of the object

proximity. In human perception, a resultant estimation (called resemblance here) does occur. It is contributed both by the properties constituting similarities and properties responsible for differences. To obtain such estimations, any real function with two kinds of factors as arguments can be used. However, the estimating resemblance function has been constructed, for some reasons, to fulfil mathematically suitable metric axioms. For example, the ratio expressed by the fraction of dissimilarities for two finite, nonempty sets $a,b$ (*i.e.*, $FD(a,b)/(FD(a,b) + FS(a,b))$ has the metric properties required.[3] Assessments of resemblance by the use of metrics (or pseudometrics) are called distances here. It is always of interest (*cf.* Ref. 18) to relate similarities and dissimilarities to the distances calculated. If the above-mentioned fraction is applied, then the behaviour of distances calculated depending on the arguments and *vice versa* is clear. By analogy, cumulating various set spectra in a standard way[3,18] has been used. Since the factors have the property of additivity under disjoint set spectra, this is done through summation of the component factors corresponding to the respective set spectra. The combinations of factors considered are determined on sequences and only via set spectra are they substituted into the respective metrics. Therefore, one deals with pseudometrics (*i.e.*, the equality $\rho(fx,fy) = 0$ does not imply $x = y$, where $x$ and $y$ are biopolymers to be compared and $fx$, $fy$ are their set spectra).

An essential step is finding an organizing principle to arrange the underlying objects, *e.g.*, biopolymers. For example, the so called property of nondecreasing distances has been applied in Ref. 18 to array chemical molecules into sequences. An organizing principle for the underlying biopolymers and the given set-spectrum mapping can be determined by means of any relation or function with the factors of similarity and dissimilarity involved. However, firstly, metrics have been employed because of their potential power to constitute interesting arrangements. In Ref. 5, the problem of orderings and random influences is approached. This goal is broadened here by making efforts to show quantitatively a hierarchy of significance of the set spectra considered with respect to the phenomena supposed to be constituted by the underlying objects and processing, *e.g.*, biopolymers. To approach this problem, a suitable organizing principle is to be looked for, namely to result in the partition $F$ characterized above. It is easily seen that for any relation $r$ determined in a family of biopolymers, there exists the smallest equivalence relation $s$ which includes $r$. This equivalence relation divides the family of investigated objects into disjoint classes. For example, $r$ can be determined as used in the computational application section: $x$ is $r$-related to $y$ if there is no element of $\mathscr{X}\backslash\{x\}$ nearer to $x$ than $y$. As a result, a partition (under $s$) of the family $\mathscr{X}$ is obtained in which equivalence classes are blocks of sequences connected by pathways determined by $s$. So, each family of sequences can be classified according to the set-spectrum mapping

fixed and some relation $r$ chosen. It is obvious that alterations in set-spectrum mappings will yield alterations in the partitions associated. Thus, some structures or in other words arrangements of the underlying biopolymers have emerged to depend on their structural properties. They are called arrangements given *via* set spectra and an organizing principle. Thus, the required partition $F$ has been constructed.

Besides such arrangements based on theoretical premises, one can consider arrangements given *via* phenomena. They can be constituted, for example, by the classifications considered below. Their characteristic property is an involvement of experimental or experience data to arrange the underlying objects. For example, carboxylic acids have monotonicly been arrayed in Ref. 18 according to the experimentally established dissociation constants. In this work, the taxonomical classifications, mainly based on some abilities of our mind to distinguish species of perceived living organisms, are used. In pattern $P$, the biopolymers can be classified according to the group of the organism from which they originate. Thus, two different kinds of arrangements: *via* phenomena, and *via* set spectra accompanied by an organizing principle can be taken into account. Hence, a natural task to compare two kinds of classifications is to be imposed. This goal is achieved considering equivalence relations which correspond to the respective partitions of a family of biopolymers being classified. If $r_P$ is an equivalence relation generated by a partition given *via* phenomena and $r_F$ is an equivalence relation corresponding to the partition given *via* set spectrum, then the distance between them can be estimated by the use of factors of similarity and dissimilarity (1) between $r_F$ and $r_P$ and by calculating the resultant estimation, *e.g.*, by the use of the fraction of dissimilarities Eq. (2). Thus, substituting various set spectra, a family of corresponding partitions is generated and sorted according to the distances from the partition given *via* phenomena. It is supposed that the significance of the properties reflected in set spectra is very precisely measured in this way. A practical application of the algorithms described and some biochemical tests are given in the next section.

## COMPUTATIONAL APPLICATION

The family of 5*S* rRNA's to be tested in this work has been reported in Ref. 20. Evolutionary aspects of 5*S* rRNA sequences discussed in Ref. 20 suggest some bearings on the structural features of sequences and the hierarchy of organisms. Hence, sharp quantitative relationships between sequence stucture and taxonomy can be looked for. The research performed here is illustrative with respect to the method abilities. Firstly, the sequences chosen have been classified (*cf.* Ref. 20) according to the morphology of organisms of their origin. In this way, a taxonomical classification given

*via* phenomena has been applied to the groups of the sequences considered. The classification is as follows:

TABLE I

Taxonomical classification (I)

| PATTERN I | | Group $\mathcal{P}$I |
|---|---|---|
| Division | Class | Species (Abbreviation) |
| *Bryophyta* | *Anthocerotae* | *Anthoceros punctatus* (Antpun) |
| | *Hepaticae* | *Marchantia polymorpha* (Marpol) |
| | | *Lophocolea heterophylla* (Lophet) |
| | *Musci* | *Plagiomnium trichomanes* (Platri) |
| *Pteridophyta* | | *Psilotum nudum* (Psinud) |
| | *Lycopsida* | *Lycopodium clavatum* (Lyccla) |
| | *Sphenopsida* | *Equisetum arvense* (Equarv) |
| | *Filicinae* | *Dryopteris acuminata* (Dryacu) |
| *Spermatophyta* | *Cycadinae* | *Cycas revoluta* (Cycrev) |
| *(Gymnospermae)* | *Ginkgoinae* | *Gingko biloba* (Ginbil) |
| | *Coniferae* | *Pinus silvestris* (Pinsil) |
| | | *Metasequoia glyptostroboides* (Metgly) |

Distinguishing groups of sequences within particular taxons, we obtain partitions suitable to be compared with »artificial« classifications generated by set spectra. For example, the sequences of group $\mathcal{P}$I (Table I) classified on the level of division, as shown above, have been divided, consistently with the set spectra listed:

1. Length (0.5588)
   Antpun, Lophet, Pinsil, Marpol.
   Dryacu, Lyccla, Psinud.
   Cycrev, Platri, Ginbil, Equarv, Metgly.
2. Component species (0.6667). The entire group
3. Common pairs (0.2500)
   Antpun, Metgly, Pinsil, Cycrev, Ginbil.
   Lyccla, Dryacu, Equarv, Psinud.
   Lophet, Marpol, Platri.
4. 3-subsequences (0.1667)
   Antpun, Lophet, Marpol, Platri.
   Cycrev, Ginbil.
   Metgly, Pinsil.
   Dryacu, Equarv, Lyccla, Psinud.
5. Cumulative $m$-subsequences (0.6667)
   The entire group.

6. Frequencies of 3-subsequences (0.2500)
   Lyccla, Antpun, Marpol, Lophet, Platri.
   Cycrev, Pinsil, Ginbil, Metgly.
   Dryacu, Equarv, Psinud.
7. Frequency cumulations of $m$-subsequences (0.2500).
   Antpun, Psinud, Marpol, Lophet, Platri.
   Ginbil, Cycrev, Pinsil, Metgly.
   Dryacu, Equarv, Lyccla.
8. Disjoint 3-subsequences (0.6667).
   The entire group.
9. Frequencies of disjoint 3-subsequences (0).
   Antpun, Lophet, Platri, Marpol.
   Cycrev, Pinsil, Ginbil, Metgly.
   Dryacu, Equarv Lyccla, Psinud.
10. Cumulations (0).
    The same as in point 9.

Distances between the pattern partition into divisions (Table I) *via* phenomena of the group considered and partitions given *via* set spectra are shown in the parentheses. Hence, the classifications corresponding to the respective set spectra can be arrayed monotonicly according to increasing distances.

Pattern I – set sp. 9, 10 (0) – set sp. 4 (0.1667) – set sp. 3,6,7 (0.2500) – set sp. 1 (0.5588) – set sp. 2,5,8 (0.6667)

Another group of sequences has been taken into account with respect to the lower rank taxon: the family (Table II; group $\mathscr{Z}$II). The pattern classification based on the morphological properties is given in Table II.

Set-spectra classifications are of the following form in this case:

1. Length. (0.6611)
   Seccer, Spiole, Betvul, Branap, Helann, Triaes.
   Zeamay, Alfalf, Lupang, Nictab, Lemmin, Lycesc, Phavul, Vicfab, Luplut.
2. Component species (0.8000) The entire group.
3. Common pairs (0.2857)
   Seccer, Triaes, Zeamay.
   Luplut, Lupang, Vicfab, Alfalf, Phavul.
   Betvul, Branap, Lemmin, Spiole, Helann.
   Nictab, Lycesc.
4. 3-subsequences (0.4762).
   Seccer, Triaes, Zeamay.
   Lemmin, Spiole, Betvul, Branap, Helann.
   Luplut, Lupang.
   Vicfab, Alfalf, Phavul.
   Nictab, Lycesc.

TABLE II

Taxonomical classification (II)

| | | Pattern II | Group ℵII |
|---|---|---|---|
| Division | Class | Family | Species (Abbreviation) |
| *Spermatophyta (Angiospermae)* | *Monocotyledones* | *Graminae* | *Secale cereale* (Seccer)<br>*Triticum aestivum* (Triaes)<br>*Zea mays* (Zeamay) |
| | | *Lemnaceae* | *Lemna minor* (Lemmin) |
| | *Dicotyledones* | *Papilionaceae* | *Lupinus Luteus* (Luplut)<br>*Lupinus angustifolius* (Lupang)<br>*Vicia faba* (Vicfab)<br>*Phaseolus vulgaris* (Phavul)<br>*Medicago sativa* (Alfalf) |
| | | *Chenopodiaceae* | *Spinacia oleracea* (Spiole)<br>*Beta vulgaris* (Betvul) |
| | | *Compositae* | *Helianthus annus* (Helann) |
| | | *Solanaceae* | *Nicotiana tabacum* (Nictab)<br>*Lycopersicon esculentum* (Lycesc) |
| | | *Cruciferae* | *Brassica napus* (Branap) |

5. Cumulative 3-subsequences (0.4407).
   Seccer, Triaes, Zeamay.
   Lemmin, Spiole, Betvul, Branap.
   Nictab, Lycesc.
   Luplut, Lupang, Helann.
   Vicfab, Alfalf, Phavul.
6. Frequencies of 3-subsequences (0.4762).
   Seccer, Triaes, Zeamay. Lemmin, Spiole, Helann, Betvul, Branap.
   Nictab, Lycesc.
   Luplut, Lupang.
   Vicfab, Phavul, Alfalf.
7. Frequency cumulations of 3-subsequences (0.4407).
   The same as in point 5.
8. Disjoint 3-subsequences (0.4407).
   The same as in points 5 and 7.
9. Frequencies of disjoint 3-subsequences (0.5352).
   Seccer, Triaes, Zeamay.
   Lemmin, Betvul, Spiole, Branap.
   Lycesc, Helann, Nictab, Luplut, Lupang.
   Vicfab, Alfalf, Phavul.

10. Cumulations (0.4407). Tha same as in points 5, 7, 8.

The monotonic array of the partitions obtained is the following:

Pattern II – set sp. 3 (0.2857) – set sp. 5, 7, 8, 10 (0.4407) – set sp. 4, 6 (0.4762) – set sp. 9 (0.5352) – set sp. 1 (0.6611) – set sp. 2 (0.8000).

The common pairs »3«, 3-subsequences »4«, cumulative $m$-subsequences »5«, disjoint 3-subsequences »8« are considered as positional ($p$) set spectra. On the other hand, the frequencies of 3-subsequences »6«, frequency cumulations of $m$-subsequences »7«, frequencies of disjoint 3-subsequences »9« are classified as frequency ($f$) set spectra. The average distances from two sets of organisms ($\mathfrak{A}$I, $\mathfrak{A}$II) classified on the level of division and family, respectively, are the following:

TABLE III

Average distances

| Taxon | (frequency) $f$ | (positional) $p$ |
|---|---|---|
| Division, $\mathfrak{A}$I | 0.1667 | 0.4375 |
| Family, $\mathfrak{A}$II | 0.4840 | 0.4108 |

As it has already been indicated, it is possible to compute distances for a family of set spectra chosen in a cumulative manner, *i.e.*, simply by summation of the component factors of similarity and dissimilarity and substituting the resultant factors to the formula on the fraction of dissimilarities. Such a cumulation is entirely nonstatistical but the result closely corresponds to that obtained by averages. The cumulative distances with respect to pattern groups are as follows:

TABLE IV

Cumulative distances

| Taxon | (frequency) $f$ | (positional) $p$ |
|---|---|---|
| Division, $\mathfrak{A}$I | 0.1750 | 0.5459 |
| Family, $\mathfrak{A}$II | 0.4870 | 0.4098 |

One can easily observe a property of opposite monotonicity which clearly shows the different role of positional and frequency set spectra for the taxons and groups of organisms considered. In symbols:

$$\mathscr{X}\mathrm{I}_f < \mathscr{X}\mathrm{I}_p,\ \mathscr{X}\mathrm{I}_f < \mathscr{X}\mathrm{II}_f$$
$$\mathscr{X}\mathrm{II}_f > \mathscr{X}\mathrm{II}_p,\ \mathscr{X}\mathrm{I}_p > \mathscr{X}\mathrm{II}_p$$

In general, the results indicate that frequency set spectra have a better relevance regarding diversification of simpler organisms classified on the level of division as opposed to the positional ones which have the smallest average and cumulative distances for higher organisms diversified on the level of family.

## DISCUSSION

Another view on the alternative ways can shed additional light on the approval of the set-spectrum method used. For example, the calculation of distances based on the costs of sequence alterations with the possible use of weighting and similarity functions is at one disposal.[7] However, such estimations involve arbitrary creations such as weights and alignments on which the valuation of similarity depends. Besides, the evolutionary distance is not sensitive to the same insertions made in the matched fragments of the two sequences compared. In the case of the frequencies considered, *e.g.* dinucleotide or codon usages, the mean permutation distance does involve further statistical disguisings of the concrete, individual relationships. An issue of the situation is to construct indices which can reflect the sequence ordering used, *e.g.*, in Ref. 12. However, the use of auxiliary parameters, the immediate substitution of sequence data into estimating function and further statistical treatment induce again uncertainty into assessments of finer properties, make unclear the correspondence between distances and alterating sequence data, introduce limitations in the comparison of possible structural features. An immediate use of sequence parameters as arguments of indices can give rise to constructing many estimating functions and the determination of ambiguous distances between the biopolymers fixed. Statistical characteristics, however, constitute an autonomic and inevitable tool, *e.g.*, when the general distributions like linguistic texts or specific regions of biosequences are investigated.[21]

Insisting on a nonprobabilistic approach in this work does not imply rejecting statistical methods in the analysis. It is supposed, however, that detailed, fine features of biopolymers can be effectively grasped by means of the mathematical tools sensitive to any alterations. The problem of sensitivity and constructing proper sets to compare leads to some generalizations of factors. In this concept, pairs of similar elements are arbitrarily selected by the use of the relation of the tolerance t defined, as usual, to be reflexive and symmetric. Such a relation can be involved in the factors in the following way.

$$FS(a,b) \overset{\text{df}}{=} |((a \times b) \cup (b \times a)) \cap t|$$

$$FD(a,b) \overset{\text{df}}{=} |a \setminus \text{dom}((a \times b) \cap t| + |b \setminus \text{im}((a \times b) \cap t)| \quad (3)$$

where dom and im denote the domain and image of the respective relations and $t$ selects pairs of similar elements from the cartesian product of $a$ and $b$. As a consequence, the similarity set grows larger and the dissimilarity one gets more »narrow« as a rule. When the tolerance comprises only pairs of identical elements, the factors generalized have the previous form of the common part and the symmetric difference of two sets. For the factors generalized, the additivity after standard cumulations is the case as well, *i.e.* the resultant factors can be derived to be summations of the components with tolerances. The scope of possible discoveries of organizing principles and varying set spectra within them is very wide and the above illustrative investigation is intended to show some alternative or parallel paths to the sophisticated problems of the structure comparison presented, *e.g.*, in Refs. 22 and 23, or important problems of mutations.[24] There is a growing interest in comparing chemical structures on the molecular level,[25–27] which essentially enriches the field of similarity and dissimilarity analyses (*cf.* the related concept of »chemical distance« in Ref. 28).

## APPENDIX

For illustration of the algorithms used, a simulation to calculate distances, partitions and distances between partitions has been performed. For example, let us take into account a collection of sequences:

$S_1$ = ACAACU, $S_2$ = ACAGACA, $S_3$ = AACCC, $S_4$ = UGGAUG, $S_5$ = CCAG, $S_6$= UCUGUCGUC, $S_7$ = UCGG.

The distances determined for the selected set spectra have been gathered in the Tables V–VII.

The desire to get partitions can be started by selecting the nearest »companion« for the sequence $S_1$ within the data relevant to the set spectrum chosen, then one can repeat the procedure for $S_2$ and so on. Thus, there are some partitions found which correspond to the respective set spectra. They are given below.

| | |
|---|---|
| »length« | $F = \{\{S_1, S_2, S_3, S_4, S_6\}, \{S_5, S_7\}\}$ |
| »component species« | $F = \{\{S_1, S_2, S_3, S_4, S_5, S_6, S_7\}\}$ |
| »common pairs« | $F = \{\{S_1, S_2, S_3, S_4, S_5, S_6, S_7\}\}$ |
| »2-subsequences« | $F = \{\{S_1, S_2, S_3, S_4, S_5, S_6, S_7\}\}$ |

Unfortunately, the collection of example sequences considered did not turn out sufficiently sensitive to the last three cases of set spectra to result in proper partitions. Note the results after standard cumulations. For ex-

TABLE V

Upper diagonal: »length«
Lower diagonal: »component species«

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|---|
| $S_1$ | 0 | 0.14286 | 0.16667 | 0 | 0.33333 | 0.33333 | 0.33333 |
| $S_2$ | 0.50000 | 0 | 0.28571 | 0.14286 | 0.42857 | 0.30000 | 0.42857 |
| $S_3$ | 0.33333 | 0.33333 | 0 | 0.16667 | 0.20000 | 0.44444 | 0.20000 |
| $S_4$ | 0.50000 | 0.50000 | 0.75000 | 0 | 0.33333 | 0.33333 | 0.33333 |
| $S_5$ | 0.50000 | 0 | 0.33333 | 0.50000 | 0 | 0.55556 | 0 |
| $S_6$ | 0.50000 | 0.50000 | 0.75000 | 0.50000 | 0.50000 | 0 | 0.55556 |
| $S_7$ | 0.50000 | 0.50000 | 0.75000 | 0.50000 | 0.50000 | 0 | 0 |

TABLE VI

Upper diagonal: »common pairs«
Lower diagonal: »2-subsequences«

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|---|
| $S_1$ | 0 | 0.70000 | 0.77778 | 0.90910 | 0.75000 | 0.92857 | 0.88889 |
| $S_2$ | 0.77778 | 0 | 0.90910 | 1 | 0.62500 | 0.76923 | 0.77778 |
| $S_3$ | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| $S_4$ | 1 | 1 | 1 | 0 | 1 | 0.84615 | 0.75000 |
| $S_5$ | 0.85714 | 0.71429 | 1 | 1 | 0 | 0.81818 | 0.66667 |
| $S_6$ | 1 | 1 | 1 | 1 | 1 | 0 | 0.70000 |
| $S_7$ | 1 | 1 | 1 | 1 | 1 | 0. 90000 | 0 |

TABLE VII

Upper diagonal: »common pairs« + »2-subsequences«
Lower diagonal: »length« + »common pairs«

| | $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ |
|---|---|---|---|---|---|---|---|
| $S_1$ | 0 | 0.73684 | 0.88889 | 0.95238 | 0.80000 | 0.96296 | 0.94118 |
| $S_2$ | 0.47059 | 0 | 0.95238 | 1 | 0.66667 | 0.88889 | 0.88889 |
| $S_3$ | 0.53333 | 0.66667 | 0 | 1 | 1 | 1 | 1 |
| $S_4$ | 0.58824 | 0.70000 | 0.70588 | 0 | 1 | 0.92308 | 0.87500 |
| $S_5$ | 0.57143 | 0.53333 | 0.71429 | 0.75000 | 0 | 0.90909 | 0.83333 |
| $S_6$ | 0.69566 | 0.56522 | 0.78261 | 0.63636 | 0.70000 | 0 | 0.80000 |
| $S_7$ | 0.66667 | 0.62500 | 0.71429 | 0.57143 | 0.40000 | 0.63158 | 0 |

ample, the factors of similarity and dissimilarity between $S_1$ and $S_2$ are 3;7 and 2;7 for »common pairs« and »2-subsequences«, respectively. The cumulated set spectrum has added the corresponding factors and the distances are determined in the usual way (*cf.* Table VII). Let us now divide the family $\{S_1, S_2, S_3, S_4, S_5, S_6, S_7\}$ using the cumulation just considered. One gets an astonishing result $F = \{\{S_1, S_2, S_3, S_5\},\{S_4, S_6, S_7\}\}$. An opposite effect of cumulation is possible. For example (*cf.* Table VII), if »length« and »common pairs« are cumulated, then no proper partition is obtained although the sequences given were divided by »length« into two classes. In general, cumulations can serve to find resultant distances even for incommensurable but influencing quantities.

*Acknowledgements.* – The following persons are gratefully acknowledged: Prof. M. Wiewiorowski, J. Barciszewski and M. Nalaskowska for biochemical consulting, and M. Popenda for the heuristic, computer -aided determination of primary classifications.

## REFERENCES

1. M. S.Waterman, *Bull. Math. Biol.* **16** (1984) 473.
2. F. Cramer, *Chaos and Order, VCH*, Weinheim, 1993.
3. L. Schulz, *Appl. Math. Comput.* **41** (1991) 1.
4. E. N. Trifonov and V. Brendel, *A Dictionary of Genetic Codes*, Balaban Publishers, Rehovat, Philadelphia, 1986.
5. S. F. Altschul and B. W. Erickson, *Mol. Biol. Evol.* **2** (1985) 526.
6. M. S. Waterman, *Meth. in Enzym.* **164** (1988) 765.
7. W. Miller and E. W. Myers, *Bull. Math. Biol.* **50** (1988) 97.
8. S. F. Altschul and B. W. Erickson, *Bull. Math. Biol.* **48** (1986) 617.
9. S. F. Altschul and B. W. Erickson, *Bull. Math. Biol.* **48** (1986) 633.
10. S. F. Altschul, B. W. Erickson, *Bull. Math. Biol.* **50** (1988), 77.
11. D. C. Benson and *Nucl. Acid Res.* **18** (1990) 3001.
12. A. A.Mironov and N. N. Alexandrov, *Nucl. Acid Res.* **16** (1988) 6169.
13. P. A. Pevzner, M. Y. Borodovsky, and A. A.Mironov, *J. Biomol. Struct. Dynam.* **6** (1989) 1027.
14. P. A. Pevzner, M. Y. Borodovsky, and A. A. Mironov., *J. Biomol. Struct. Dynam.* **5** (1988) 1013.
15. W. A. Beyer, C. Burks, and W. B. Goad, *Los Alamos Sci.* **9** (1983) 62.
16. L. Bush, W. B. Lacy, J. Burkhardt, D. Hemken, J. Moraya-Rojel, T. Koponen, and H. do Souza Silva, *Making Nature, Shaping Culture: Plant Biodiversity in Global Context,* Univ. Nebrasca Press, 1996.
17. R. J. Nussinov, *Theor. Biol.* **125** (1987) 219.
18. L. Schulz, *J. Mol. Struct. (Theochem)* **231** (1991) 367.
19. S. M. Gusein-Zade and M. Y. Borodovsky, *J. Biomol. Struct. & Dynam.* **7** (1990) 1185.
20. T. D.Mashkova, M. Z. Barciszewska, A. Joachimiak, M. Nalaskowska, and J. Barciszewski, *J. Biol. Macromol.* **12** (1990) 247.
21. P. V. Kostetsky and R. R. Vladimirova, *J. Biomol. Struct. Dynam.* **9** (1992) 1061.

22. P. Bamborough, C. JR Hedgecock, and W.G. Graham, *Structure* **2** (1994) 839.
23. Y. Takahashi, *Topics in Current Chemistry* **174** (1995) 105.
24. N. F. Cariello, L. Cui, C. Beroud, and T. Soussi, *Cancer Research* **54** (1994) 4454.
25. R. E. Carhart, D. H.Smith, and R. Ventkataraghavan, *J. Chem. Inf. Comput. Sci.* **25** (1985) 64.
26. S. K. Kearsley, S. Sallamack, E. M. Fluder, J. D. Andose, R. T. Mosley, and R. P. Sheridan, *J. Chem. Inf. Comput. Sci.* **36** (1996) 118.
27. Ch. Cheng, G. Maggiora, M. Lajiness, and M. Johnson, *J. Chem. Inf. Comput. Sci.* **36** (1996) 909.
28. V. Kvasnicka and J. Pospichal, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1109.

## SAŽETAK

### Sekvencijska analiza s pomoću algoritma spektra skupova

*Lech Schulz*

U uspoređivanju struktura primijenjen je širi pogled na biokemijsku strukturu koji ide dalje od vizualizacije i uspoređivanja. Samoorganizacijska svojstva biopolimera, predodređena građevnim molekulskim komponentama, prikazana su s pomoću reprezentacija teorije skupova (poznate kao spektri skupova) kako bi se obogatilo prepoznavanje strukture. Zatim su kombinacije faktora sličnosti i različitosti uključenih u posebnu metriku, efikasno upotrebljene u praktičnim algoritmima za rezultantnu, kvantitativnu usporedbu različitih svojstava. Upotrijebljeni alternativni nestatistički algoritmi omogućuju precizno otkrivanje odnosa između svojstava sekvencija i fizičkih fenomena. U ovom radu, porodice sekvencija 5*S* rRNA testirane su s obzirom na njihova općenito nepredočiva, matematički određena svojstva, te su zatim postavljene korelacije s biološkim sustavima. Primijenjena metoda ima brojne prednosti kao alternativni pristup istraživanju mehanizama života, bolesti, mutacija, problema vezanih za genetički kod, itd.