Journal of Mathematical Chemistry, Vol. 43, No. 2, February 2008 (© 2007)
DOI: 10.1007/s10910-007-9229-7

On a geometry-based approach to protein sequence alignment

Milan Randić*

National Institute of Chemistry, Ljubljana, Slovenia
E-mail: mrandic@msn.com

Received 17 December 2006; Revised 19 January 2007

We consider a novel numerical representation of proteins obtained by assigning to individual amino acids the polar coordinate on a unit circle. As a result one can represent protein sequence as one-dimensional numerical sequence, the entries of which when subtracted facilitates search for alignment between pairs of proteins of interest. The alignment is sought by shifting one sequence relative to another by several sequence units to the left or to the right. The novel approach is illustrated on two yeast proteins having 174 and 171 amino acids.

KEY WORDS: graphical representation of proteins, protein signature, protein alignment

1. Introduction

In view that there is no rigorous solution to the protein structure alignment this challenging problem continues to attract attention and leads to innovations. Existing algorithms for protein structure alignment involve computationally intensive schemes like the conventional techniques of dynamic programming [1], the probabilistic Monte Carlo approach [2], the sequence alignment genetic algorithm [3], and alternative schemes such as three-dimensional clustering [4, 5], combinatorial extension of an alignment path [6] as well as some graph theoretical approaches [7, 8]. Computer-oriented approaches consider various penalties for deletion, substitution and permutation of protein labels (i.e., amino acids), associated with the metrics of Levenshtein [9], also known as the "edit distance." While weighting of different operations differently is legitimate and plausible, selection of the weighting factors nevertheless maintains an element or arbitrary decision-making. Is it possible to develop a methodology for protein sequence alignment that will bypass such, to a degree arbitrary, decisions?

*Visiting Emeritus from the Department of Mathematics & Computer Science Drake University, Des Moines, Iowa.
Permanant Address: Milan Randić, 3225 Kingman Rd, Ames, IA 50014, USA.

We will outline in this paper one such approach, which in a search for the optimal protein alignment, appears to be free of "human" intervention. The approach is based on a particular graphical representation of proteins and subsequent geometry-based numerical representation of the graphical image of proteins. By shifting the relative positions of the leading amino acids of the two sequences one is seeking for the optimal alignment of amino acids. This approach in its present form applies to the "alignment" in its narrow meaning, being synonymous with "identity" search in which two sequences are compared step-by-step. We will start with a brief review of graphical representations of proteins. As will be seen the novel approach to the problems of protein alignment is unusually simple and one may say elegant (in view that it does not holds cumbersome mathematical manipulations). It is, however, confined to one particular graphical representation of proteins in which proteins is depicted as a spectrum-like pattern. The alignment process is straightforward that at the first sight it may appear surprising that it has not been considered earlier. However, we should mention that this progress towards arriving at graphical alignment, which involves a geometrically based algorithm for protein sequence representation was possible only after the appearance of the first graphical representations of proteins, which have been initiated very recently – the topic of graphical representation of proteins is barely one year old!

The approach to graphical alignment of proteins, to be outlined here, is based on one particular graphical representation, which involves one particular ordering of 20 natural amino acids. Thus we have avoided problems involving factorial number of alternative assignments of amino acids when considering graphical representations of proteins. We will briefly comment at the end of this article on the metrics associated with search for the optimal alignment.

2. On numerical representation of protein and DNA sequences

Graphical representations of DNA were initiated in late 1980s and early 1990s by the pioneering work of Hamori [10], Jeffrey [11] and others [12–19]. By prescribing certain rules for assigning the four nucleic bases A, C, G and T (standing for adenine, cytosine, guanine and thymine, respectively) to particular elements of underlying geometrical template used for graphical representation one can "transform" given DNA sequence into a geometrical object that allows visual inspection of DNA. It took another dozen years until it has been realized that several existing graphical representations of DNA can lead to numerical characterization of DNA. In this way hitherto qualitative comparisons of DNA sequences could be up-graded into quantitative, yielding thus useful numerical characterization on DNA primary sequences [20].

Although it does not take imagination to realize that a similar (that is graphical) treatment of proteins may yield similar advantages, graphical repre-

sentations of proteins emerged only very recently [21–26]. The delay has been apparently caused by the combinatorial explosion associated with arrangement of 20 amino acids among 20 equivalent (or non-equivalent) geometrical sites of potential geometrical objects to be used as a template for graphical representation of proteins. Here two routes are possible: (1) One adopts one out of zillion possible assignments for the 20 amino acids as the standard and applies it to the analysis, assuming that while affecting individual protein representation it will not significantly affect when one consider differences between protein sequences; and (2) One searches for protein representations which will be label-insensitive. Recently it has been realized that one can bypass the combinatorial problem associated with 20! (or about $2.43\,10^{18}$) alternative assignments of 20 objects by taking an advantage of the well-known characteristic flexibility in assigning arbitrarily labels to graphs or even using unlabeled graphs to arrive at pictorial representation of proteins. This approach appears to offer promising novel direction for exploring graphical representations of protein [24, 25]. However, more important for our approach to the problem of protein sequence alignment considered in this article is the recognition that the "alphabetic" representations of bio-sequences even if based on an arbitrary selected assignment of amino acids, which will affect individual graphical representation will not necessarily significantly impair comparative study, where the dominant role will be played by examination of differences among graphical representation of two or more proteins. The essential novelty of the present approach to protein alignment is transformation of "alphabetic" protein sequences into "numerical" sequences of the same bio-sequences. At the first look this may appear as a trivial modification, but the significance of representing proteins (and DNA sequences) by numerical sequences is that this allows one to perform arithmetic manipulations on sequences, something that has hitherto not been possible.

3. 2-D Graphical representations of proteins

We starts by placing 20 amino acids on the periphery of a unit circle [26], as shown in figure 1, which is one of 20! possible ways of arranging the 20 natural amino acids. This simple geometrical "device" allows two alternative routes for graphical representation of proteins. One way, which has been outlined in ref. [26], assigns to amino acids a pair of (x, y) coordinates in the interior of the unit circle. In the alternative approach one locates amino acids on the circumference of the unit circle and continue to represent the successive amino acids of a protein sequence as spots over the circle circumference, which is a one-dimensional geometrical object. By placing amino acids on the circumference of the unit circle one can associate with each of the 20 natural amino acids a *single* numerical value, the corresponding polar angle, which are listed in table 1 (which are expressed in radians). The unit circle with assigned 20 amino acids has been

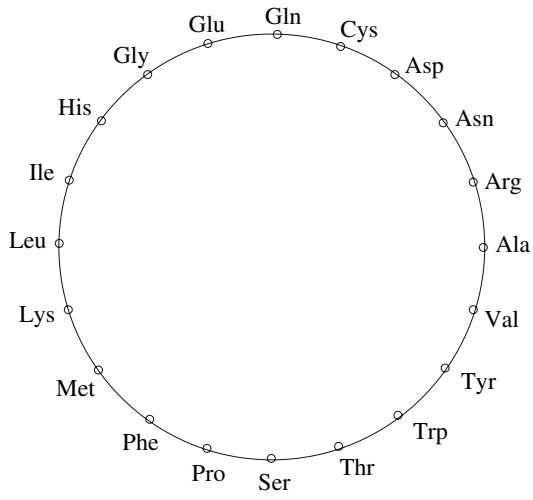

Figure 1. Twenty amino acids uniformly placed on the circumference of unit circle.

Table 1
The angular coordinates of 20 natural amino acids placed on the circumference of a unit circle.

Amino Acid	3-Letter Code	1-Letter Code	Radians
Alanine	Ala	A	0
Arginine	Arg	R	0.314159
Aspargine	Asn	N	0.628319
Aspartic acid	Asp	D	0.942478
Cysteine	Cys	C	1.256637
Glutamine	Gln	Q	1.570796
Glutamic acid	Glu	E	1.884956
Glycine	Gly	G	2.199115
Isoleucine	His	I	2.513274
Histidine	Ile	H	2.827433
Leucine	Leu	L	3.141593
Lysine	Lys	K	3.455752
Methionine	Met	M	3.769911
Phenylalanine	Phe	F	4.084070
Proline	Pro	P	4.398230
Serine	Ser	S	4.712389
Threonine	Thr	T	5.026548
Tryptophan	Trp	W	5.340707
Tyrosine	Tyr	Y	5.654867
Valine	Val	V	5.969026

recently proposed by Randić et al. [26], for one of novel graphical representation of proteins. The former approach is conceptually related to the algorithm of Jeffrey [11], designed for highly compact graphical representation of DNA. The difference between using this approach for graphical representation of DNA and extending the approach to proteins is that instead using the interior of a square,

the corners of which have labels of the four nucleic bases, here one uses circle and its periphery. The ingenious algorithm of Jeffrey [11] designed for graphical representation of DNA, represents a modification of the mathematical "chaos game" of Barnsley and Rising [27]. The difference is that Barnsley considers N-gon and starts depicting a sequence of random numbers by starting at an arbitrary point in the interior of the N-gon. He moves then from that arbitrary point half a way towards another vertex of N-gon selected at random. He continues the "game" indefinitely (that is till some large number of points have been consumed). Jeffrey modified the algorithm of Barnsley by selecting as polygon a square, the corners of which have been labeled by the four nucleic acids A, C, G and T. Graphical representation of DNA of Jeffrey is obtained by starting at the center and plotting DNA bases within the interior of a square following analogous rule to that of the chaos game, except that vertices are not selected at random. Hence, Jeffrey's DNA "game" is deterministic, not random.

In the approach of Randić et al. [26] instead of a square with four its corners labeled by A, C, G and T one considers a circle labeled by 20 amino acids. The 2-D graphical representation of proteins is obtained by using the *interior* of the unit circle with 20 amino acids. Protein of any size can in this way be graphically represented by N points *within* the interior of the unit circle, where N is the number of amino acids of the protein considered. Although so constructed graphical representations of two related proteins may not offer immediately novel insights, when the *difference* between the corresponding points in two sequences are plotted one can observe matching of amino acids in the sequences. This is illustrated in figure 2 for the rather short stretches of polypeptide sequences (from ref. [26]):

Protein I W F V E S Q N D P K N S P V V L W L N G G P G C S S L D G L

Protein II W F V E S Q S N P S T D P V V L W L T G G P G C S G L S A L

The above two proteins have been used to illustrate the Needlemen–Wunsch algorithm [28] and Smith–Waterman algorithm [29] in the Molecular Modeling book of A. R. Leach [30].

While the approach of Randić, Butina and Zupan has been successful in indicating the degree of matching between the two short protein sequences the problem remains how to generalize such an approach so that it can apply for searching for protein sequence matching when the two sequences have not been fixed one relative to the other. At that time there was no solution in sight for this more demanding and more realistic problem, which remained for a while an important "unsolved issue." In particular it remained unclear how the algorithm will manage with gaps and inserts of amino acids. Considerations of these issues takes time and one should not expect methodology that rest on novel approach overnight to produce a "product" that can be comparable with methods that

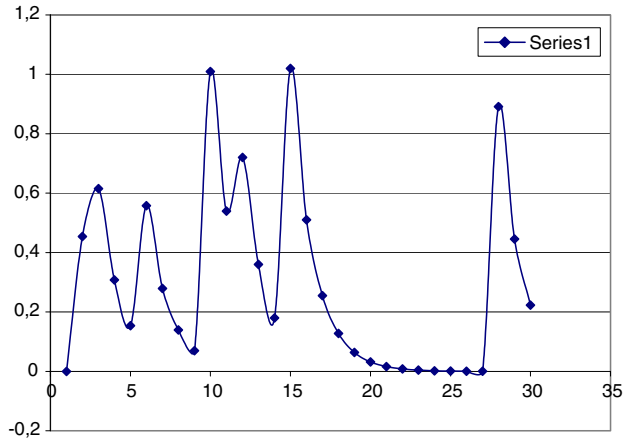

Figure 2. Difference in coordinates of corresponding amino acids in two proteins showing locations of mismatch in protein sequences.

have been available for over 25 or 35 years (see the dates of refs. [27, 28] respectively). One may recall a question posed to Michael Faraday on his presentation at the Royal Society Lectures in London on the eve of his discovery of the magnetic induction: "Sir, for what use is going to be this new discovery?" To this Faraday replied: "For what use is a new born baby?" In our view it is clearly premature to discuss the present approach to protein alignment in relation to currently used algorithms, just as it would be inappropriate to make comparison of 25 and 35 old athletes with child that just has been born.

Another question may be raised that relates to non-zero amplitudes in the present approach and their potential interpretation. These amplitudes definitely depend on adopted ordering of amino acids on the periphery of the unit circle, and thus will vary if different standards for ordering of amino acids are selected. Here we also think that, while this question is in place, it may be nevertheless somewhat premature. While amino acids are ordered at random (which includes the alphabetic ordering) it is not likely that useful information will be associated with non-zero difference amplitudes. However, it may be that there is a way to extract some useful information from "non-zero" amplitudes in the difference "spectra" if amino acids can be ordered in some meaningful way, such as based on their physico-chemical properties. In recent work on the alignment of DNA sequences based on comparison of 64 codons [31, 32], it appears that by suitable ordering of codons on the circumference of the unit circle, which gives preference to the first two bases of each triplet, the "height" of non-zero amplitudes when differences of DNA sequences are considered may offer useful information. Similarly, in a preliminary work on alignment of protein sequences of four natural peptides (having 33–34 amino acids) three of which have antimicrobial and hemolytic activity, while the fourth peptide has only antimicrobial activity

it was possible to take an advantage of the results other than zero by locating of the presence of "undesirable" amino acid at certain critical places [33]. In this particular application of protein comparisons the problem has been to find a model that can discriminate among otherwise similar peptides and classify them with respect to desired properties. The problem was approached by using geometry-based graphical alignment essentially as presented in this work, which was followed by close examination of parts of peptides in which they display lack of alignment. In particular attention was given to the "non-zero" "spikes" in the difference graph, which means amino acids that are different and which appear as peaks of different height. It may be premature to speculate that such an approach will solve numerous problems of protein alignment, but even if applied to few situations it may deserve further attention.

4. Representation of proteins on the circumference of the unit circle

The similarity between the graphical alignment approach introduced in this article and 2-D representation of proteins in which the *interior* of the unit circle was used for depicting individual amino acids of a protein is precisely in that here we use *circumference* of the circle (which is a one-dimensional object) and not its 2-D interior. The significance of this is that each amino acid in the novel graphical representation of proteins is assigned a *single* and constant coordinate (e.g., polar angle) and not two coordinates x, y, which specify amino acid within the circle interior, which incidentally continually vary for each amino acid as one moves along the protein sequence. The advantage of the periphery of the circle is obvious: There is no need to calculate coordinates for each individual base and there is no need to calculate distances between pairs of coordinates at each step of graphical representation of protein, because *the same* coordinates are used again and again when the same amino acid appears in the sequence. Hence, the difference between locations of two bases is given by the difference of a single coordinate. If we add 2π each time we circle around the periphery and pass the zero angle to arrive at the next amino acid we obtain unique coordinates for each amino acid in considered protein. However, there is no loss of information if only the list of values in the interval $0 \leqslant \theta \leqslant 2\pi$ is kept for representation of a protein. The list of angular position of successive amino acids gives a one-dimensional numerical representation of proteins, which can be displayed for visual inspection as a bar graph or spectrum-like graph of connecting points shifted horizontally by unit length.

In table 2 we listed the entries of the one-dimensional protein representations in the adjacent numerical columns for the first dozen amino acids of carboxypeptidase Y from *Saccharomyces cerevisiae* and of mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region also from *Saccharomyces cerevisiae*. The full list of amino acids of the two proteins is shown in table 3,

Table 2
The initial two dozen amino acids of carboxypeptidase Y from *Saccharomyces cerevisiae* and of mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region also from *Saccharomyces cerevisiae*.

K	3.455752	3.455752	P	4.39823	4.39823
I	2.513274	8.796459	S	4.712389	4.712389
L	3.141593	9.424778	K	3.455752	9.738937
G	2.199115	14.76548	L	3.141593	15.70796
I	2.513274	15.07964	G	2.199115	21.04867
D	0.942478	19.79203	I	2.513274	21.36283
P	4.39823	23.24778	D	0.942478	26.07522
N	0.628319	25.76106	T	5.026548	30.15929
V	5.969026	31.10177	V	5.969026	31.10177
T	5.026548	36.44247	K	3.455752	34.87168
Q	1.570796	39.26991	Q	1.570796	39.26991
Y	5.654867	43.35398	W	5.340707	43.03982
T	5.026548	49.00884	S	5.026548	49.00884
G	2.199115	52.46459	G	2.199115	52.46459
Y	5.654867	55.92035	Y	5.654867	55.92035
L	3.141593	59.69026	M	3.769911	60.31858
D	0.942478	63.77433	D	0.942478	63.77433
V	5.969026	68.80088	Y	5.654867	68.17256
E	1.884956	70.99999	K	3.455752	72.57079
D	0.942478	76.3407	D	0.942478	76.3407
E	1.884956	77.28318	S	4.712389	80.11061
D	0.942478	82.62388	K	3.455752	85.13716
K	3.455752	85.13716	H	2.827433	90.79202
H	2.827433	90.79202	F	4.08407	92.04866

which we will examine in the next section when searching for protein sequence alignment. The two sequences of length 174 and 171, respectively, have been mentioned in an article by von Homeyer [34] in order to illustrate the program SAGA for protein alignment [3], which uses the genetic algorithm for multiple sequence alignment. We will, however, consider only pair-wise alignment of these two protein sequences.

Before continuing let us point out that the essential step in the present approach is "transformation" of an *alphabetic* sequence to *numerical* sequence. However, instead of the numerical sequence of polar coordinates (expressed in radians) listed in table 2:

$$3.455, 2.513, 3.141, 2.199, 2.513, 0.942, 4.398, 0.628, 5.969, 5.026\ldots$$

one can consider the sequence of integers, in which integer numbers indicate the multiplets of the polar angle $2\pi/20$ (or 18°):

$$11, 8, 10, 7, 8, 3, 14, 2, 19, 16, \ldots$$

Table 3
Complete sequence of amino acids of carboxypeptidase Y from *Saccharomyces cerevisiae* (Protein 1) and of mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region also from *Saccharomyces cerevisiae* (Protein 2).

Protein 1:

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

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

G P D L K P I G N P Y S W N S N A T V I F L D Q P V N V G F S Y S

G S S G V S N T V A A G K D V Y N F L E L F F D Q F P E Y V N K G

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

S V L I G N G L T

Protein 2:

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

D P A N D P I I L W L N G G P G C S S F T G L L F E L G P S S I G

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

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

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

I G N G I T

Integer sequences may be simpler to handle, but they do not introduce any novelty. Thus the only change in figure 3 in which we depicted the above sequence, would be in the scale on the y-axis, while the form of the spectrum-like graphs remains unchanged.

In figure 3 we have illustrated the 1-D graphical representation of one of the two proteins of table 3, in which on the x-axis is the sequential order of amino acids and on the y-axis (in the range from 0 to 2π) is the polar angle (in radians) giving the location of the corresponding amino acid on the periphery of the unit circle of figure 1. Such graphical representations are not very informative, except to indicate at a close look that there are on y-axes 20 different magnitudes ("heights"), which result in the depicted spectrum-like diagram or the "signature" of the protein. As mentioned if one wishes one could replace the 20 polar coordinates by numbers 1–20, without changing the graphical representation dramatically. We may mention here that application of the here outlined graphical alignment approach to DNA, which has been recently published [35] does use integer coordinates (multiplets of $\pi/2$ or 90°) for numerical representation of the four nucleic acids.

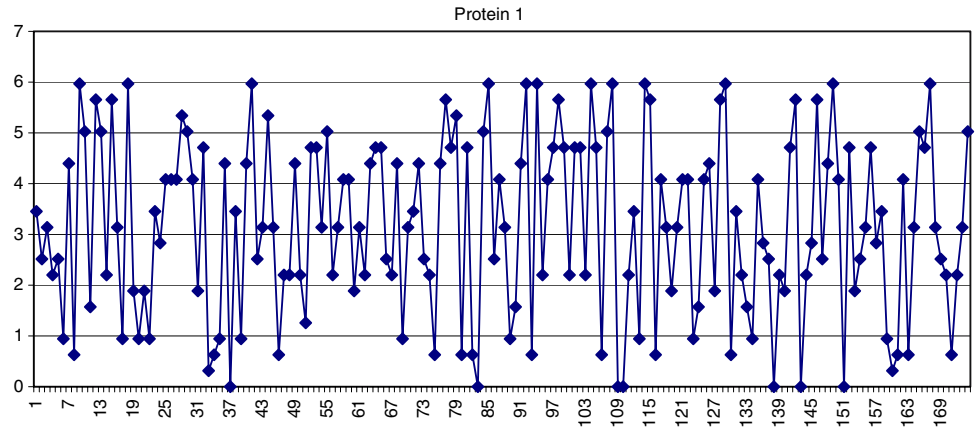

Figure 3. The plot of angular coordinates of amino acids of the unit circle for the 174 amino acids of carboxypeptidase Y from *Saccharomyces cerevisiae*.

5. Search for protein sequence alignment

Rather than visually inspecting 1-D representations of proteins we will now examine them numerically (using Excel for all computations). Let us refer to the two protein sequences to be compared as A and B, the leading entries of which are:

$$\mathbf{A}: A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, \ldots$$

$$\mathbf{B}: \quad B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8, \ldots$$

Rather then examining the 1-D protein "signatures" illustrated in figure 3 we will instead construct and consider the *difference* plot of the two 1-D signatures:

$$\mathbf{A_n - B_n}: A_1 - B_1, A_2 - B_2, A_3 - B_3, A_4 - B_4, \ldots$$

which is illustrated in figure 4 (top). Observe that the differences are positive and negative, all they all lie in the range $(-2\pi, 2\pi)$. Should there be many adjacent zero points in various segments of the signature that would indicate coincidental fragments of protein sequence. However, as we see from figure 4 that appears not to be the case. In order to better see the zero values in the region 1–20 in figure 5 (top) we illustrate this portion of the difference plot at higher resolution. Observe that no alignment involving more than two amino acids, which is not significant, has been detected in this region.

Next then one shifts the two sequences by 1 place, that is, one computes the sequences

$$\mathbf{A_n - B_{n+1}}: A_1 - B_2, A_2 - B_3, A_3 - B_4, A_4 - B_5, \ldots$$

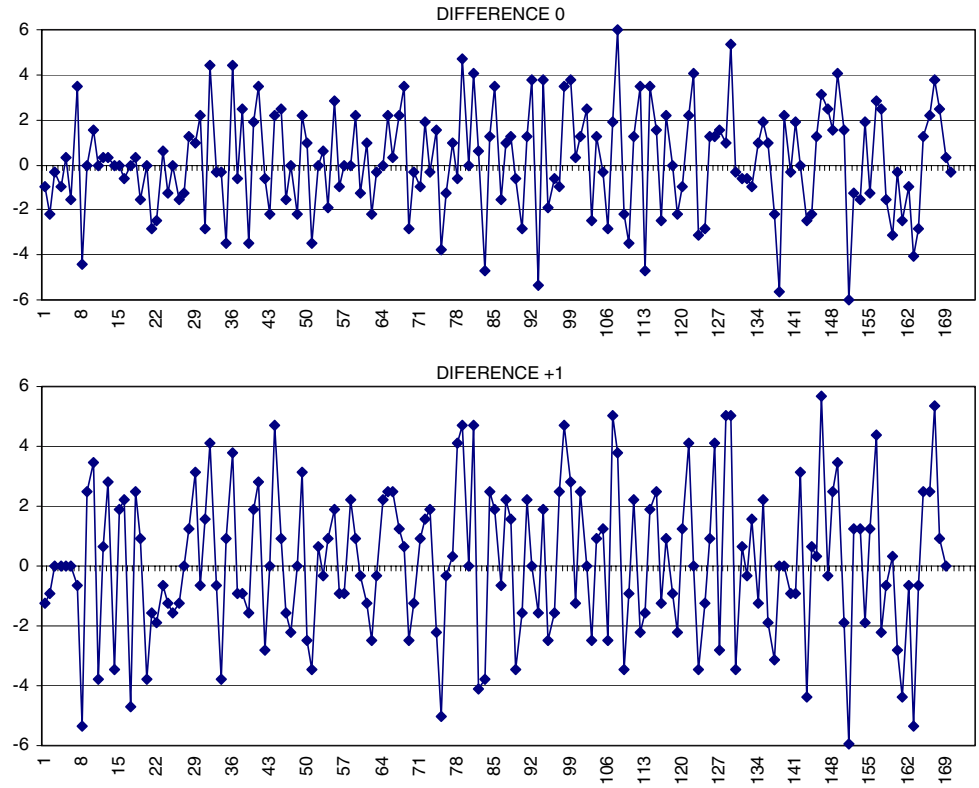

Figure 4. The difference in the radian coordinates of the corresponding amino acids of amino acids of carboxypeptidase Y from *Saccharomyces cerevisiae* (top) and amino acids of mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region also from *Saccharomyces cerevisiae*.

and again plots the computed difference between the coordinates, which we display in figure 4 (at the bottom). Again one finds no significant overlap of amino acid, except for four amino acids in the initial part of the signature, which are also illustrated in figure 5 (bottom) at a higher resolution, which may and need not be significant. By shifting the calculation of the differences of the 1-D radial coordinates by two places, that is by depicting the sequence:

$$\mathbf{A_n - B_{n+2}} : A_1 - B_3, A_2 - B_4, A_3 - B_5, A_4 - B_6, \ldots$$

no overlap in sequences of the two proteins was again observed. If one continues with few more steps in the same direction one does not find any novel fragment of identical portions in the two sequences. In view of that we turned to examine shifts of the two protein sequences in the opposite direction, that is, we decided to consider the sequences:

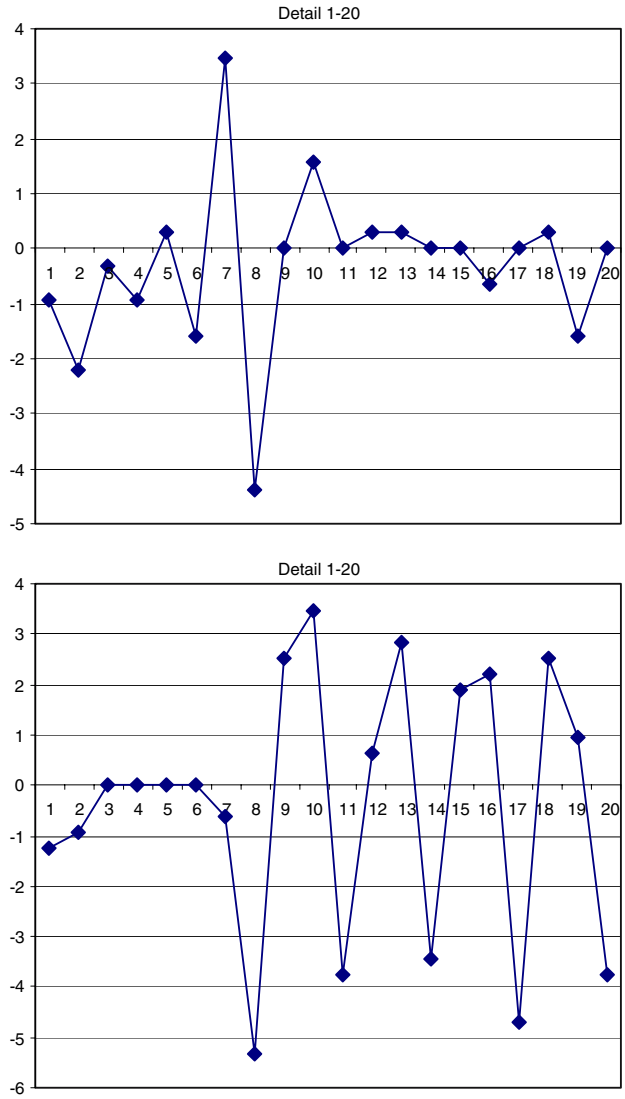

Figure 5. Details of the alignments in the region 1–15 for the two difference plots of figure 4 (top) and (bottom), respectively.

$$\mathbf{A_{n+1}} - \mathbf{B_n} : A_2 - B_1, A_3 - B_2, A_4 - B_3, A_5 - B_4, \ldots$$
$$\mathbf{A_{n+2}} - \mathbf{B_n} : A_3 - B_1, A_4 - B_2, A_5 - B_3, A_6 - B_4, \ldots$$
$$\mathbf{A_{n+3}} - \mathbf{B_n} : A_4 - B_1, A_5 - B_2, A_6 - B_3, A_7 - B_4, \ldots$$
$$\mathbf{A_{n+4}} - \mathbf{B_n} : A_5 - B_1, A_6 - B_2, A_7 - B_3, A_8 - B_4, \ldots$$

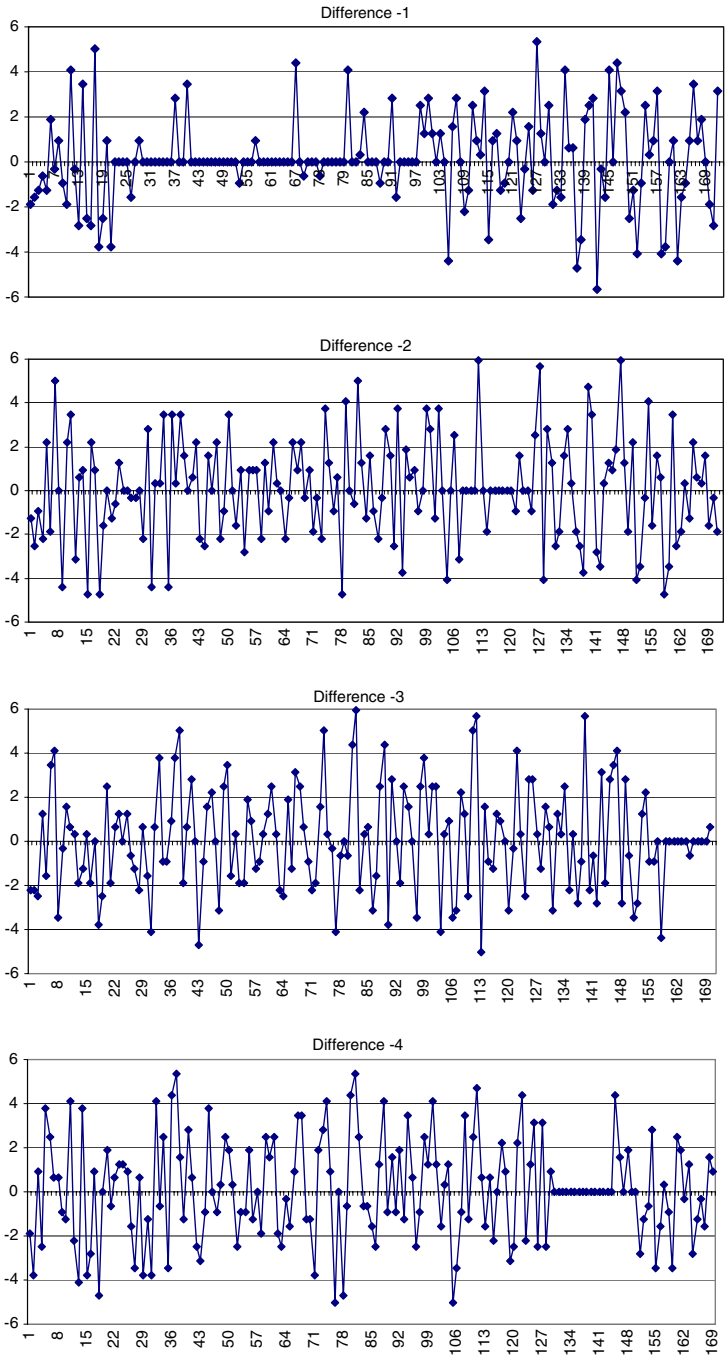

Figure 6. The difference in the radian coordinates of amino acids of carboxypeptidase Y from *Saccharomyces cerevisiae* (top) and amino acids of mature putative serine carboxypeptidase in ESR1-IRA1 intergenic region also from *Saccharomyces cerevisiae* shifted to the left by one to four places.

A comparison of the above sequences leads to the graphical differences shown in figure 6a–d, in which one can detect considerable portions in the sequences showing matching of amino acids in various segments. In figure 6a (top), with a shift by one place between the two sequences, a lengthy matching occurs of amino acids between sites 20–100, while in figure 6b there is an alignment of the two sequences in the region around 110–120. By further increasing the shift of the two sequences we arrive at figure 6c which shows alignment of the two sequences in the region around 160–170, and finally by making an additional shift between the two protein sequences one can observe alignment in the region around 130–145 in figure 6d (bottom). By continuing to consider the difference

$$\mathbf{A_{n+5}} - \mathbf{B_n} : A_6 - B_1, A_7 - B_2, A_8 - B_3, A_9 - B_4, \ldots$$

no alignment of additional significant region of the two protein was detected. The corresponding plots are of similar form as those of figure 4, the most part of which display typical form of a "noise" diagram.

6. Alignment pattern

The information of figure 6 can be combined into a single table in which two proteins are depicted one above the other such that after shifting the two sequences the maximal alignment results. Close look at figure 6 allows one to identify the positions in the two sequences at which gaps are to be inserted. figure 4 at best show some limited matching in the region 10–15, but even this region is interrupted by mismatches, thus not leading to a conclusive establishment for the presence of longer matches. While one can view an occurrence of a segment of three and four consecutive amino acids as statistically significant, not being likely to be caused by the chance probability, when isolated they are of limited interest. However, when the two proteins have been shifted by one place immediately we saw in the region 22–99 an impressive portion of the two proteins being aligned. The similar situation occurs when the two sequences are shifted by two places, which results in a dozen matches in the region 108–120 of the corresponding amino acids. Another 10 matches in the region 159–169 occurs when the two sequences are shifted by three place, and finally by shifting the two sequences by four places 15 matches are found in the region 130–145. One can combine the information from different sections of figure 6 by inserting gaps at the sites that precede the matches and in this way arrive at table 4 that essentially displays the same information in an alternative (standard) format. There are additional pairs of adjacent identical amino acids in isolation along the sequence of the two proteins (e.g., the pair VS around location 105 and the pair EI around 153), but we decided not to count isolated pairs as statistically

Table 4
Reconstruction of the alignment of protein 1 and protein 2.

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


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


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

significant (unless additional arguments, such as multiple alignment, would tell differently).

We have covered a large section of the two proteins, totaling 117 out of 169 sites. It is possible that there are additional regions, in which there is alignment of few amino acids, and if one wish to be sure one should continue shifting the two sequences till one exhaust all relative shifts of two sequences. However, one need not to consider all $2(n-1)$ possible pairwise graphs, where n is the protein's (equal) length, as would be the case in the general case, which for the examples given in table 3 would be over 300 alignments, because as we have seen for over 2/3 bases alignment has been completed! At the present there is no formal metric applied to decide whether alignment is significant or not, except that alignment of two amino acids in isolation is not considered as being of interest. Already alignment of 3–4 (or more amino acids) appears to be statistically significant, and is not likely to occur at random, having the probability around $1/20^3$–$1/20^4$ or from 1.25×10^{-4} about 6.2×10^{-6} to occur at random.

Clearly there are a number of unsettled questions concerning here outlined graphical alignment of proteins. However, one should not overlooked, that the same may hold also with applications of currently adopted computer programs for protein alignment. For our approach, for example, the question is can one detect in some instances a significant drift in primary sequence may take place for functionally related proteins. Similarly, the present approach need not detect fragments of proteins with a reversed order of amino acids. It remain to be investigated how broad applicability of the present approach in such situations may be and whether it will recognize "alignment" in proteins deemed homologous by other available method. While we still have to await for such applications and such answers for sure one can expect that the present approach can always be used as auxiliary methodology that, at least at the present time, can be combined with other frequently used method. It cannot only offer independent route to the same problems and thus check the degree in which independent approaches agree, but can also facilitate and hopefully increase the efficiency of other pro-

cedure by eliminating considerations of a number of pairwise protein primary sequence comparisons, relating to sequences for which alignment is found, as being redundant.

7. Concluding remarks

Let us indicate the significance of the novel numerical approach, which is characterized by unusual simplicity, elegance and negligible computational time. All the calculations and plots here performed were using Excel. In fact the "bottleneck" of current use of geometry-based graphical alignment is typing the input information on protein sequences into Excel! Observe that in contrast to computer-oriented searches for protein sequence alignment, which may involve some arbitrary decisions with respect to the relative magnitude of penalty for shits or substitution in this approach there are no similar arbitrary decision. What is arbitrary in this approach, and in fact is one out of immense number of 20! possibilities, is the singular (alphabetic) arrangements of 20 amino acids on the periphery of the unit circle. However, while this will influence the relative magnitudes of the radian coordinates of individual amino acids (those listed in tables 1 and 2) and will produce different "signatures" for the same protein sequences, this will not influence the search for graphical alignment of proteins, which is given by the *differences* in the coordinates of selected pairs of amino acids in two proteins. We believe that the current approach can be combined with some of the available computer-based algorithms for further possible fine tuning of such programs that could increase their efficiency.

Acknowledgments

The author wishes to express thanks to Professor Jure Zupan and Dr. Marjana Novič of the Laboratory for Chemometrics of the National Institute of Chemistry, Ljubljana, Slovenia for the hospitality and to the Ministry of Science of Slovenia for the financial support within the project P1-017: Modeling of relationship between chemical structure and properties QSAR–QSPR.

References

[1] C.A. Orengo, N.P. Brown and W.T. Taylor, Proteins Struct. Funct. Gen. 14 (1992) 139.
[2] L. Holm and C. Sander, J. Mol. Biol. 233 (1993) 123.
[3] C. Notredame, L. Holm and D.G.A. Higgins, Bioinformatics 14 (1998) 407.
[4] G. Vriend and C. Sander, Proteins 11 (1991) 52.
[5] D. Fisher, O. Bachar, R. Nussinov and H. Wolfson, J. Biomol. Struct. Dyn. 9 (1992) 769.
[6] I.N. Shindyalov and P.E. Bourne, Protein Eng. 11 (1998) 739.
[7] N.N. Alexandrov, Protein Eng. 9 (1996) 727.
[8] N.N. Alexandrov and D. Fisher, Proteins 25 (1996) 354.

[9] V.I. Levenshtein, Doklady Akademii Nauk SSSR 163 (1965) 845 also Sov. Phys. Doklady 10 (1966) 707.
[10] E. Hamori, BioTechniques 7 (1989) 710.
[11] H.I. Jeffrey, Nucleic Acid Res. 18 (1990) 2163.
[12] A. Nandy, Curr. Sci. 66 (1994) 309.
[13] X. Guo, M. Randić and S.C. Basak, Chem. Phys. Lett. 350 (2002) 106.
[14] M. Randić, N. Lerš and D. Plavšić, Chem. Phys. Lett. 368 (2003) 1.
[15] M. Randić and J. Zupan, SAR QSAR Environ. Res. 15 (2004) 147.
[16] M. Randić, Chem. Phys. Lett. 386 (2004) 468.
[17] J. Zupan and M. Randić, J. Chem. Inf. Model. 45 (2005) 309.
[18] M. Randić, Period. Biol. 107 (2005) 415.
[19] M. Randić, D. Vikić-Topić, A. Graovac, N. Lerš and D. Plavšić, Period. Biol. 107 (2005) 437.
[20] M. Randić, M. Vračko, A. Nandy and S.C. Basak, J. Chem. Inf. Comput. Sci. 40 (2000) 1235.
[21] M. Randić, J. Zupan and A.T. Balaban, Chem. Phys. Lett. 397 (2004) 247.
[22] M. Randić, SAR QSAR Environ. Res. 15 (2004) 191.
[23] M. Randić, A.T. Balaban, M. Novič, A. Založnik and T. Pisanski, Period. Biol. 107 (2005) 403.
[24] M. Randić, J. Zupan, and D. Vikić-Topić, Reported at the 2nd Meeting of Int. Acad. Math. Chem. June 15–17, Dubrovnik, Croatia; J. Mol. Graphics Model. (in press).
[25] M. Novič and M. Randić, Reported at the 2nd Meeting of Int. Acad. Math. Chem. June 15–17, Dubrovnik, Croatia; J. Math. Chem. (submitted).
[26] M. Randić, D. Butina and J. Zupan, Chem. Phys. Lett. 419 (2006) 528.
[27] M.F. Barnsley and H. Rising, *Fractals Everywhere*, 2nd ed. (Academic Press, Boston, MA, 1993).
[28] S. Needleman and C.D. Wunsch, J. Mol. Biol. 48 (1970) 443.
[29] T.F. Smith and M.S. Waterman, J. Mol. Biol. 147 (1981) 195.
[30] A.R. Leach, *Molecular Modelling – Principles and Applications*, 2nd ed. (Pearson Educational Limited, Harlow, England, 2001).
[31] M. Randić, Acta Chim. Slovenica 53 (2006) 477.
[32] M. Randić, M. Novič, D. Vikić-Topić and D. Plavšić, Graphical and numerical representation of DNA based on codons.
[33] M. Randić and D. Juretić, Anti-alignment as a tool for differentiation of peptides (to be published).
[34] A. von Homeyer, in: *Handbook of Chemoinformatics – from Data to Knowledge*, Vol.3 ed. J. Gasteiger (Wiley-VCH, Weinheim, Germany, 2003), pp. 1239–1280.
[35] M. Randić, J. Zupan, D. Vikić-Topić and D. Plavšić, Chem. Phys. Lett. 431 (2006) 375.