PROCEEDINGS OF THE ROYAL SOCIETY A

*Proc. R. Soc. A* (2011) **467**, 1052–1072
doi:10.1098/rspa.2010.0382
*Published online* 20 October 2010

# Hybrid approach to analysis of β-sheet structures based on signal processing and statistical consideration

By V. Vojisavljevic[1,*], E. Pirogova[1], D. M. Davidovic[2] and I. Cosic[1]

[1]*Health Innovation Research Institute, School of Electrical and Computer Engineering, RMIT University, PO Box 2476, Melbourne, Victoria, Australia*

[2]*The Institute of Nuclear Sciences 'Vinca', 11001 Belgrade, PO Box 522, Serbia*

A number of biotechnology applications are based on protein design. For this design, the relationship between a protein's primary structure and its conformation is of vital importance. A β-sheet is a common feature of a protein's two-dimensional structure; therefore, elucidating the principles governing β-sheet structure and its stability is critical for understanding the protein-folding process. In the three-dimensional representation of protein molecules, $C^{\alpha}$ carbon coordinates (carbon atom immediately adjacent to the carboxylate group) have often been employed instead of the complete set of coordinates for the corresponding residues. Using the $C^{\alpha}$ carbon coordinates, we showed that particular amino acids are not randomly distributed within a β-sheet structure. On the basis of a new statistical approach for the analysis of a spatial distribution of amino acids in a protein, presented by their physico-chemical parameters, the electron–ion interaction potential (EIIP) and hydrophobicity, are described here. The relationship between amino acid positions inside the β-sheet and the EIIP and hydrophobicity parameters was established. The correlation between amino acid propensities related to the β-sheet was examined using multiple cross-spectra analysis. We also applied the continuous wavelet transform for the analysis of selected β-sheet structures using the EIIP and hydrophobicity parameters. The findings provide new insight into conformational propensities of amino acids for the adaption of β-sheet structures.

**Keywords: β-sheet conformation; electron-ion interaction potential; hydrophobicity; amino acid; signal processing; wavelet transform**

## 1. Introduction

Peptides and proteins are used in the design of new drugs and other pharmaceutical products. They also present the desirable components for different bio-devices that become key elements in different biotechnological applications. In this study, we have analysed β-sheet structures that are common and functionally important spatial motifs in proteins. In the β-sheet, the polypeptide chains are

*Author for correspondence (vuk.vojisavljevic@rmit.edu.au).

*Received* 21 July 2010
*Accepted* 17 September 2010

usually spatially arranged side-by-side to form a structure that resembles a series of pleats where the hydrogen bonds are formed between adjacent segments of a polypeptide chain. The R groups of the adjacent amino acids protrude from the zigzag structure in the opposite directions and create the alternating pattern. The adjacent polypeptide chains in a β-sheet can be either parallel or anti-parallel having the same or opposite amino-to-carboxyl orientations, respectively (Nelson & Cox 2007). Apart from being an important part of the protein-folding process in many proteins, the β-sheet plays a very important functional role. Recent investigations revealed that the β-sheet is a key element for understanding the rules governing protein–DNA, protein–RNA and protein–protein interactions (Fadeev *et al.* 2009; Campagne *et al.* 2010). In comparison to the α-helix propensity of amino acids, the β-sheet propensity is more difficult to analyse because of the strong influence of the amino acids surrounding the β-sheet conformation (Smith & Regan 1997). It is known that different amino acids are characterized by significantly different propensities of the β-sheet conformation (Williams *et al.* 1987; Smith & Regan 1997). This study was aimed at addressing the following two critical issues:

— Do the spatial locations of amino acids appear completely spatially random inside the β-sheet structure?
— Are there any physico-chemical parameters that can characterize the tendency of amino acids to be in a particular location within the β-sheet structure?

To solve these two problems, we have analysed a spatial distribution of amino acids in a protein molecule, which were presented by their physico-chemical properties, the electron-ion interaction potential (EIIP) and hydrophobicity. The EIIP parameter describes the average energy states of all valence electrons in a particular amino acid. The EIIP values for each amino acid were calculated from the general model of pseudopotentials (Veljkovic & Slavic 1972; Veljkovic 1980):

$$\langle k+q|w|k\rangle = \frac{0.25Z\sin(1.04\pi Z)}{2\pi},$$

where $q$ is a change of momentum $k$ of the delocalized electron in the interaction with potential $w$, while

$$Z = \frac{\sum_i Z_i}{N},$$

where $Z_i$ is the number of valence electrons of the $i$th atom of each amino acid and $N$ is the total number of atoms in the amino acid. The EIIP and hydrophobicity values of 20 amino acids are shown in table 1.

Hydrophobic interactions are the most important non-covalent forces that are responsible for protein structure stabilization, binding of enzymes to substrates and folding of proteins. The hydrophobic character of amino acids is recognized to be the main driving force in the protein-folding process. The hydrophobic effect is believed to play a major role in organizing the self-assembly of water-soluble, globular proteins (Smialowski *et al.* 2007). There are several different scales of the hydrophobicity parameter for amino acids. Such scales can be classified as solution measurements, empirical calculations or some combination of the two.

Table 1. The electron–ion interaction potential and hydrophobicity parameters of amino acids in three-letter representation.

| amino acid | EIIP | hydrophobicity | amino acid | EIIP | hydrophobicity |
|---|---|---|---|---|---|
| Ala | 0.0373 | 0.61 | Leu | 0 | 1.53 |
| Arg | 0.0959 | 0.6 | Lys | 0.0371 | 1.15 |
| Asn | 0.0036 | 0.06 | Met | 0.0823 | 1.18 |
| Asp | 0.1263 | 0.46 | Phe | 0.0946 | 2.02 |
| Cys | 0.0829 | 1.07 | Pro | 0.0198 | 1.95 |
| Gln | 0.0761 | 0 | Ser | 0.0829 | 0.05 |
| Glu | 0.0058 | 0.47 | Thr | 0.0941 | 0.05 |
| Gly | 0.005 | 0.07 | Trp | 0.0548 | 2.65 |
| His | 0.0242 | 0.61 | Tyr | 0.0516 | 1.88 |
| Ile | 0 | 2.22 | Val | 0.0057 | 1.32 |

Solution scales are based on the distribution coefficients between an aqueous phase and a suitably chosen organic phase, while empirical scales are based on partitioning between the solvent-accessible surface and the buried interior in proteins of known structure.

Supposing that $C^{\alpha}$ atoms are each located randomly in the β-sheet, then the assumption of spatial randomness is that the locations of these $C^{\alpha}$ atoms have no influence on one another. Hence, under spatial randomness, the random variables are assumed to be statistically independent. This assumption defines the fundamental hypothesis of complete spatial randomness (CSR), which is referred to as the CSR hypothesis. In this study, the normalized average hydrophobicity scale (Cid *et al.* 1992) was selected for investigation as a physico-chemical property of a given amino acid. The use of the EIIP and hydrophobicity parameters for determination of the spatial position of amino acids inside the β-sheet structures has been evaluated statistically. The results obtained show that the particular amino acids do not follow a pattern under the CSR.

Signal-processing methods have been successfully used in structure–function analysis of proteins and DNA (Cosic *et al.* 1989; Cosic 1994, 1995, 1997; Pirogova *et al.* 2002, 2008). Here, we applied signal-processing techniques for the retrieval of informational content about protein biological activity contained in their β-sheet structures. By applying the multiple cross-spectral analysis method, the correlation between amino acid parameters related to β-sheet conformation has also been examined. Wavelet transform (WT), a signal-processing tool for multi-resolution analysis and local feature extraction of non-stationary signals, has been used successfully for analysis of biological signals. The WT replaces the Fourier transform (FT) by a family of functions generated by translations and dilations of a window called a wavelet. In wavelet function, there are two arguments: time and scale. In our previous studies, continuous wavelet transform (CWT) was applied to determine the functional active sites of different protein families (Pirogova *et al.* 2003). In this study, we applied the CWT for analysis of the selected β-sheet structures and prediction of functionally important regions within selected polypeptide chains. The EIIP and hydrophobicity were used as representative amino acid properties for this analysis.

## 2. Materials and methods

### (*a*) *Beta sheet*

Amino acid sequences with $C^{\alpha}$ atoms coordinates, that form a particular β-sheet, were derived from the PDB (http://www.rcsb.org/pdb). In our analysis, we selected the structures according to the following criteria: (i) only the X-ray crystallographic structures with a resolution higher than 1.6 Å are included; (ii) protein structures contained residues with unknown amino acids and incomplete coordinate data are excluded; (iii) only proteins that contain a single chain are included, aiming to remove the disturbance between the chains; and (iv) only β-sheets with more than three strands and at least six amino acids in each strand are included.

### (*b*) *Geometrical centre of protein sequence and statistical analysis*

The term 'spatial location' of an amino acid is used here to specify a position relative to the geometrical centre (GC) of the $C^{\alpha}$ atoms for the particular amino acid. The GC of an object represents a point in space characterized by the minimal sum of distances between that point to all other points that belong to the analysed object. The coordinates of the GC for a particular β-sheet are calculated as follows:

$$x_C = \frac{\sum x_i}{N}, \quad Y_C = \frac{\sum y_i}{N} \quad \text{and} \quad Z_C = \frac{\sum z_i}{N}$$

where $N$ is the total number of amino acids that belong to the particular β-sheet and $x_i, y_i, z_i$ are coordinates of the $C^{\alpha}$ atoms of the $i$th amino acid, where $i = 1, \ldots, N$. The distance from the GC to the $C^{\alpha}$ atoms corresponding to the $i$th amino acid is computed as:

$$d_i = \sqrt{(x_i - x_C)^2 + (y_i - y_C)^2 + (z_i - z_C)^2}.$$

In order to analyse the randomness of the spatial distribution of amino acids, the following assumptions have been made:

— Each β-sheet represents one realization of the random stochastic process.
— For each amino acid, we summarize the statistics of the distances from $C^{\alpha}$ atoms of the particular amino acid to the GC.
— For a particular β-sheet and $k$ amino acid, we select the amino acid that is at the minimal distance to the GC. This distance ($d^k_{\min}$) represents a single event of the stochastic process.
— Finally, the statistics of $d^k_{\min}$ are summarized over a set of 70 β-sheet structures.

To test the CSR for a particular amino acid, we summarize statistics of distances $d^k_{\min}$ for the whole assemble of 70 β-sheets, where $k$ represents the

analysed amino acid. The mean nearest distance from the GC has been calculated for the $k$th amino acid using the equation:

$$\bar{d}^k_{\min} = \sum_{l}^{n} \frac{d^k_{\min,l}}{n},$$

where $n$ is the number of realizations of a stochastic process (the number of analysed β-sheets), and index $l$ represents the $l$th β-sheet.

Moreover, for each amino acid, a null hypothesis $H_0$, where the particular amino acid falls in the pattern under the CSR, can be used. Generally, the acceptance of the null hypothesis means that a pattern of the particular amino acid is not significantly different from the CSR. The rejection of the null hypothesis implies that the actual pattern is significantly different from the pattern estimated using the CSR. Hence, we use Clark and Evans' approach (Clark & Evans 1954), where the expected value of the average nearest neighbouring distance is:

$$E(d_i) = 0.5\sqrt{\frac{A}{N}}, \tag{2.1}$$

where $A$ is the total tested area, and $N$ is the total number of points in the area $A$. The observed and expected values now can be compared using normally distributed $\boldsymbol{z}$ statistics of the form.

$$z = \frac{[\bar{d}^k_{\min} - E(\bar{d}^k_{\min})]}{\sqrt{\mathrm{var}(\bar{d}^k_{\min})}}, \tag{2.2}$$

where $\mathrm{var}(\bar{d}^k_{\min}) = 0.0683(A/N^2)$.

### (c) *EIIP*($n$) *and* $H$($n$) *of amino acids*

It has been suggested that understanding the correlation between the positions of amino acids and their physico-chemical parameters can enhance our knowledge of the protein-folding process and improve protein design. In order to test how the EIIP and hydrophobicity parameters can influence the positions of particular amino acids in the β-sheet structure, we calculated the average EIIP and average hydrophobicity values as a function of distance from the GC:

$$\mathrm{EIIP}_{\mathrm{ave}}(n) = \frac{1}{|N(n)|} \sum_{N(n)} \mathrm{EIIP}_k;\ \mathrm{EIIP}_k \text{ is the EIIP of } k\text{th amino acid} \tag{2.3}$$

and

$$H_{\mathrm{ave}}(n) = \frac{1}{|N(n)|} \sum_{N(n)} H_k; \quad H_k \text{ is the hydrophobicity of } k\text{th amino acid}, \tag{2.4}$$

Table 2. The selected β-propensity and hydrophobicity amino acid parameters used for analysis. The parameters are taken from Kanehisa (1988).

| | | | |
|---|---|---|---|
| B020 | B106 | B172 | B257 |
| B028 | B120 | B187 | B275 |
| B039 | B139 | B218 | B276 |
| B045 | B141 | B221 | B277 |
| B046 | B161 | B225 | B278 |
| B061 | B164 | B226 | B279 |
| B101 | B167 | B232 | H35 |
| B102 | B168 | B234 | H36 |
| B103 | B169 | B251 | H58 |

Table 3. Values of spatial criterion $r$ for calculating amino acid positions to the GC in the β-sheet structure.

| $N(n)$ | $N$th concentric sphere |
|---|---|
| $N(1)$ | $r < 0.35$ nm |
| $N(2)$ | $0.35\,\text{nm} < r < 0.7\,\text{nm}$ |
| $N(3)$ | $0.7\,\text{nm} < r < 1.05\,\text{nm}$ |
| $N(4)$ | $1.05\,\text{nm} < r < 1.4\,\text{nm}$ |

where $N(n)$ is a concentric sphere used for summation, and $|N(n)|$ is the number of amino acids inside the sphere. The parameters of $N(n)$ as a function of $n$ are shown in table 3.

### (*d*) *Multiple cross-spectral analysis of amino acid properties*

Multiple cross-spectral analysis was applied here to determine the correlation between the amino acid parameters representing a β-sheet structure as follows:

— The original amino acid sequence is transformed into a numerical sequence by assigning each amino acid a particular value for the physical parameter relevant to the β-sheet structure (β-propensities, normalized average hydrophobicity scales and the EIIP. The parameter values are shown in table 2).
— Two additional parameters representing the expected minimal distance ($d^k_{\text{min}}$) from the geometrical centre for each of the 20 amino acids are shown in table 4. These are the new parameters introduced by the authors of this paper.
— The numerical sequences obtained are analysed using FTs in order to extract information pertinent to the two-dimensional structure of a given protein. As the average distance between amino acid residues in a protein sequence is about 3.8 Å, it can be assumed that the points in the numerical sequence derived are equidistant. For further numerical analysis, the distance between points in these numerical sequences is set at an arbitrary value of $d = 1$. Peak frequencies in the amplitude cross-spectral function define common frequency components (periodicity) of the two or

Table 4. The expected values of $d^k_{\min}$ calculated for the whole β-sheet and for the part of the β-sheet inside the space ($d_{\min} < 0.75$ nm) along with the results calculated from PDB coordinate files. The ratio $R = d^k_{\min}/E(d^k_{\min})$ is a measure of the degree to which the observed distribution approaches or departs from random expectation. For random distribution $R = 1$.

| amino acid | frequency of amino acids for the whole β-sheet calculated from PDB coordinate files | frequency of amino acids ($d < 7.5$ A) calculated from PDB coordinate files | expected values for the whole β-sheet | expected values of $d^k_{\min}$ for ($d < 7.5$ A) | values of $d^k_{\min}$ calculated from PDB coordinate files | $R$ for the whole β-sheet | $R$ for ($d < 7.5$ A) |
|---|---|---|---|---|---|---|---|
| Ala | 0.0403 | 0.051 | 8.33 | 6.53 | 7.38262 | 1.06 | 1.2 |
| Arg | 0.0452 | 0.058 | 7.87 | 6.1 | 7.60239 | 1.04 | 1.24 |
| Asn | 0.0432 | 0.029 | 8.04 | 8.63 | 8.66092 | 0.9 | 1.03 |
| Asp | 0.0262 | 0.025 | 10.33 | 9.23 | 9.85152 | 1.04 | 1.08 |
| Cys | 0.0578 | 0.018 | 6.96 | 10.92 | 9.53537 | 0.71 | 0.89 |
| Gln | 0.0534 | 0.055 | 7.23 | 6.3 | 6.94654 | 1.09 | 1.05 |
| Glu | 0.0179 | 0.051 | 12.47 | 6.53 | 7.52529 | 1.57 | 1.22 |
| Gly | 0.0364 | 0.025 | 8.76 | 9.23 | 8.85041 | 0.89 | 1.07 |
| His | 0.0422 | 0.025 | 8.13 | 9.23 | 8.70438 | 1.01 | 0.87 |
| Ile | 0.0777 | 0.091 | 6 | 4.88 | 5.84019 | 1.01 | 1.22 |
| Leu | 0.0631 | 0.069 | 6.65 | 5.6 | 6.61608 | 0.95 | 1.24 |
| Lys | 0.0359 | 0.047 | 8.82 | 6.77 | 7.69853 | 1.09 | 1.19 |
| Met | 0.051 | 0.029 | 7.4 | 8.63 | 7.91685 | 0.96 | 0.9 |
| Phe | 0.067 | 0.069 | 6.46 | 5.6 | 6.40478 | 0.97 | 1.19 |
| Pro | 0.0267 | 0.004 | 10.23 | 24.42 | 11.6685 | 0.88 | 0.48 |
| Ser | 0.0364 | 0.08 | 8.76 | 5.21 | 7.16893 | 1.22 | 1.38 |
| Thr | 0.0578 | 0.069 | 6.96 | 5.6 | 6.77616 | 0.95 | 1.31 |
| Trp | 0.0666 | 0.036 | 6.48 | 7.72 | 6.92008 | 0.86 | 0.97 |
| Tyr | 0.0714 | 0.051 | 6.26 | 6.53 | 7.102 | 0.86 | 1.12 |
| Val | 0.0826 | 0.116 | 5.82 | 4.32 | 5.09636 | 1.14 | 1.18 |

more numerical sequences analysed. To determine the common frequency components for a group of sequences, we have calculated the values of multiple cross-spectral function coefficients $M_n$, which are defined as follows (Cosic 1994):

$$M_n = \prod_{k=1}^{L} |X_{k,n}|, \quad n = 1, \ldots, \frac{N}{2}, \tag{2.5}$$

where $L$ is the number of cross-correlated sequences, $M_n$ represents the $n$th spectral component in the cross-spectral function, and $X_{k,n}$ is the $n$th spectral component of the $k$th numerical sequence. Peak frequencies in such a multiple cross-spectral function denote common frequency components for all numerical sequences analysed.

### (*e*) *The continuous wavelet transform*

The CWT is a relatively new signal-processing tool efficient for multi-resolution analysis and local feature extraction of non-stationary signals (Pirogova *et al.* 2002). The WT can be viewed as an inner product operation that measures the similarity or cross-correlation between signals and wavelets. The continuous version of the WT of signal $(t)$ is defined as:

$$\mathrm{CWT}_{\omega}(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} X(t) \Psi \left( \frac{t-a}{b} \right) \mathrm{d}t, \tag{2.6}$$

where $b$ is the shift factor (the translation factor of the wavelet function along the time axis) and $a$ is the scale factor (it scales a function by compressing or stretching it). CWT is one of the time or space-frequency representations of a signal. A time (space) frequency representation of a signal provides information about how the spectral content of the signal evolves with time (space), thus providing an ideal tool to dissect, analyse and interpret signals with transients or localized events. This is performed by mapping a one-dimensional signal in the time (space) domain into two-dimensional time (space)–frequency representation of the signal. Because CWT provides the same time–space resolution for each scale and thus, CWT can be chosen to localize individual events, such as the active site identification. In our previous study (Pirogova *et al.* 2002), we compared the performance of different wavelet functions, including Morlet, Meyer, Daubechies, Simlets, Coiflets and Mexican Hat, for the detection of the active sites of the oncogene protein with the previously determined characteristics. We showed that Morlet wavelet was the most suitable for the identification of active sites of different oncogene proteins, as by using the Morlet we achieved the best correlation between the predicted and experimentally determined protein active sites. Here, the Morlet function was employed for the identification of critical amino acids in the selected β-sheet structures, which presents a locally periodic wave-train:

$$\omega(t) = Ce^{(-(t^2/2)+j\omega t)}, \tag{2.7}$$

where $\omega_0 = 5.33$ and $C$ is the constant used for normalization.

From equation (2.6), it can be seen that the Morlet wavelet is a complex sine wave modulated by a Gaussian function. The time–frequency version of CWT can be achieved by making the substitution, $a = f_0/f$:

$$\mathrm{CWT}(t,f) = \int s(\tau) \sqrt{\frac{f}{f_0}} \Psi \left( \frac{f}{f_0}(t-\tau) \right) \mathrm{d}\tau, \tag{2.8}$$

in which the analysing wavelet becomes essentially a prototype band-pass filter with centre time $t = 0$ and centre frequency $f_0$. The centre frequency and frequency bandwidth of the CWT vary with scale. However, their ratio remains fixed. It is the constant property of the wavelet. The underlying property of wavelets is that they are pretty well-localized in both time and frequency (Pirogova *et al.* 2002). A product of the uncertainties of both time and frequency is bound by the Heisenberg uncertainty principle; no filter can have a width product smaller than $1/\pi$. The Gaussian filters (Morlet wavelet) attain this theoretical limit. Strictly speaking, the CWT provides a time-scale representation rather

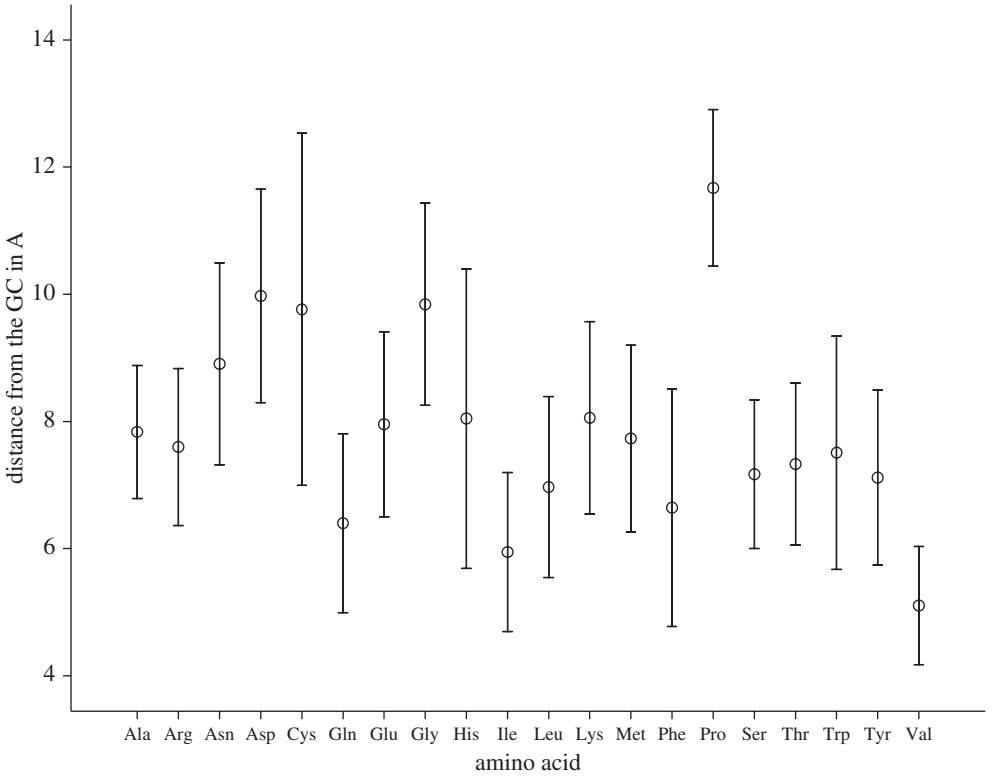


Figure 1. The statistics of the minimal distances ($d^k_{\min}$) for each amino acid shown as a 95% error box diagram.

than a time-frequency representation. However, the scale factor of the CWT is closely related to the frequency and this makes the mapping from time-scale representation to time-frequency representation possible. β-Sheet structure is usually built up of domain(s) within the protein molecule. Applying the WT, we can observe a whole frequency-spatial distribution and thus, are able identify the domain(s) of high energy of a particular frequency along the protein sequence. These regions can be determined by studying the set of local extrema of the moduli in the WT domain. Those energy-concentrated local extrema are the locations of sharp variation points of the amino acid property selected for analysis (in our case the EIIP and hydrophobicity), and are proposed as the most critical locations for functionally important regions within the β-sheet structure.

## 3. Results

### (*a*) *Clustering of amino acids*

To evaluate the CSR for the particular amino acid, we summarized statistics of distances $d^k_{\min}$ for the whole assemble of 70 β-sheet structures, where $k$ represents the analysed amino acid. The results obtained present the distribution of $d^k_{\min}$ that is shown in a 95 per cent confidence box diagram (figure 1). The different distributions of amino acids obtained denote the possibility of clustering these

Table 5. Classification of the amino acids based on the average value of the parameter $d^k_{\min}$ (Tukey algorithm).

| group | amino acid | | | | | | |
|---|---|---|---|---|---|---|---|
| I | Val | 5.09636 | | | | | |
| II | Ile | | 5.84019 | 5.84019 | | | |
| | Phe | | 6.40478 | 6.40478 | | | |
| III | Leu | | 6.61608 | 6.61608 | 6.61608 | | |
| | Thr | | 6.77616 | 6.77616 | 6.77616 | | |
| IV | Trp | | 6.92008 | 6.92008 | 6.92008 | 6.92008 | |
| | Gln | | 6.94654 | 6.94654 | 6.94654 | 6.94654 | |
| | Tyr | | 7.102 | 7.102 | 7.102 | 7.102 | |
| | Ser | | 7.16893 | 7.16893 | 7.16893 | 7.16893 | |
| | Ala | | 7.38262 | 7.38262 | 7.38262 | 7.38262 | |
| | Glu | | 7.52529 | 7.52529 | 7.52529 | 7.52529 | |
| | Arg | | 7.60239 | 7.60239 | 7.60239 | 7.60239 | |
| | Lys | | 7.69853 | 7.69853 | 7.69853 | 7.69853 | |
| | Met | | 7.91685 | 7.91685 | 7.91685 | 7.91685 | |
| V | Asn | | | 8.66092 | 8.66092 | 8.66092 | |
| | His | | | 8.70438 | 8.70438 | 8.70438 | |
| | Gly | | | 8.85041 | 8.85041 | 8.85041 | |
| VI | Cys | | | | 9.53537 | 9.53537 | 9.53537 |
| | Asp | | | | | 9.85152 | 9.85152 |
| VII | Pro | | | | | | 11.6685 |

amino acids into particular groups. By using the Tukey algorithm (Tukey 1997), the amino acids were classified into nine groups which are presented in table 5. The distances between the distributions of amino acids were calculated for each pair of amino acids using a *t*-test analysis. From table 6, 1-$\alpha$ levels can be seen that allow accepting the hypothesis that two amino acids have the same distribution parameters. Furthermore, several distinctive findings could be emphasized:

— Val, Ile, Leu are over-represented at distances smaller than 0.4 nm from the GC (32%, 26%, 20%, respectively, compared with the expected value of 6%).
— For Glu, Thr, Trp, Tyr and Phe amino acids, the difference between expected and actual values is highest at the distance $0.35\,\text{nm} < r < 0.7\,\text{nm}$ (42%, 41%, 35%, 46%, 38%, respectively, compared with the expected value of 13%).
— Pro and Asp are under-represented in the central part of the corresponding β-sheet structure, and over-represented at distance far from the GC, where $r < 1.4\,\text{nm}$ (figure 1 and table 5).
— The amino acids Arg, Asn, Gln, His, Met and Ser are randomly distributed within the β-sheet structure.

Interestingly, Pro is well known as a secondary structure breaker as well as a very conservative amino acid. Pro has a tendency to be positioned towards the ends of the β-sheet structures (Chou & Fasman 1978). Owing to

Table 6. The 1-$\alpha$ values for rejecting hypothesis that pair of amino acid has the same distribution. The italic entries represent amino acids with similar $d^k_{\min}$ distributions.

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 1.00 | | | | | | | | | | | | | | | | | | | |
| Arg | *0.80* | 1.00 | | | | | | | | | | | | | | | | | | |
| Asn | 0.13 | 0.20 | 1.00 | | | | | | | | | | | | | | | | | |
| Asp | 0.00 | 0.01 | 0.15 | 1.00 | | | | | | | | | | | | | | | | |
| Cys | 0.04 | 0.06 | 0.39 | *0.76* | 1.00 | | | | | | | | | | | | | | | |
| Gln | *0.63* | 0.46 | 0.05 | 0.00 | 0.02 | 1.00 | | | | | | | | | | | | | | |
| Glu | *0.87* | 0.93 | 0.16 | 0.01 | 0.05 | *0.51* | 1.00 | | | | | | | | | | | | | |
| Gly | 0.08 | 0.13 | *0.81* | 0.23 | *0.50* | 0.03 | 0.10 | 1.00 | | | | | | | | | | | | |
| His | 0.17 | 0.24 | *0.96* | 0.22 | 0.45 | 0.07 | 0.20 | *0.87* | 1.00 | | | | | | | | | | | |
| Ile | 0.05 | 0.02 | 0.00 | 0.00 | 0.00 | 0.18 | 0.03 | 0.00 | 0.00 | 1.00 | | | | | | | | | | |
| Leu | 0.33 | 0.21 | 0.01 | 0.00 | 0.00 | 0.69 | 0.23 | 0.00 | 0.02 | 0.27 | 1.00 | | | | | | | | | |
| Lys | *0.71* | 0.91 | 0.23 | 0.01 | 0.07 | 0.39 | *0.83* | 0.15 | 0.27 | 0.01 | 0.15 | 1.00 | | | | | | | | |
| Met | *0.60* | *0.75* | 0.45 | 0.05 | 0.16 | 0.35 | *0.69* | 0.34 | 0.46 | 0.03 | 0.17 | *0.82* | 1.00 | | | | | | | |
| Phe | 0.24 | 0.14 | 0.00 | 0.00 | 0.00 | 0.53 | 0.16 | 0.00 | 0.01 | 0.44 | *0.78* | 0.10 | 0.12 | 1.00 | | | | | | |
| Pro | 0.00 | 0.00 | 0.00 | 0.04 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | | | | | |
| Ser | *0.79* | 0.59 | 0.06 | 0.00 | 0.02 | *0.79* | *0.65* | 0.03 | 0.09 | 0.07 | 0.45 | 0.49 | 0.43 | 0.32 | 0.00 | 1.00 | | | | |
| Thr | 0.45 | 0.29 | 0.01 | 0.00 | 0.00 | *0.84* | 0.32 | 0.01 | 0.03 | 0.18 | *0.82* | 0.22 | 0.22 | 0.62 | 0.00 | *0.59* | 1.00 | | | |
| Trp | 0.63 | 0.47 | 0.06 | 0.00 | 0.02 | *0.98* | 0.52 | 0.04 | 0.09 | 0.23 | *0.73* | 0.40 | 0.36 | *0.58* | 0.00 | *0.78* | *0.87* | 1.00 | | |
| Tyr | *0.74* | *0.54* | 0.05 | 0.00 | 0.02 | *0.86* | *0.60* | 0.03 | 0.08 | 0.09 | *0.52* | 0.45 | 0.40 | 0.37 | 0.00 | *0.93* | 0.66 | *0.84* | 1.00 | |
| Val | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.28 | 0.03 | 0.00 | 0.00 | 0.07 | 0.00 | 0.00 | 0.01 | 0.04 | 0.01 | 1.00 |
| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |

the small size of the Asn molecule, it is expected that this amino acid will be positioned inside the β-sheet structure. However, our calculations revealed that in fact Asn is uniformly distributed in the β-sheet. The hydrophobic residues Val, Ile, Phe, Leu have high hydrophobicity and low EIIP values. Similar clustering was reported by Noar *et al.* (1996) using the amino acid 'pair interchange' approach.

### (*b*) *Amino acids and complete spatial randomness*

A question is raised: do the spatial locations of amino acids appear completely spatially random or not? A statistical analysis, used for CSR testing, is usually based on a random sample of $n$ points. Distribution theory for the test is based on the independence of $n$ sample points. In our statistical experiment, each GC is regarded as a sample point. If the position of the amino acids in the β-sheet structure is regarded as an event, then the point-to-event distances can be summarized and computed. These calculated distances will not depend on the real coordinates of the amino acids. To proceed further, we calculated the ratio $A/N_k$ using the simple relationship:

$$\frac{A}{N} = \frac{A}{N} \cdot \frac{N}{N_k},$$

where $N$ represents the total number of amino acids in some area. We found that the $A/N$ ratio is practically constant and has a value of $1.66 \times 10^{-19}\ \mathrm{m}^2$. The value $N/N_k$ is practically the reciprocal of the propensity or frequency of a given amino acid. It is assumed that $N/N_k$ corresponds to the statistics covering the whole area of β-sheets, and $H_0$ is the mean nearest distance from the GC for the $k$th amino acid, that is the result of the CSR. For $\alpha = 0.05$, a value of $z$ from equation (2.2) is derived from the normal distribution $N(\bar{d}_i, \mathrm{var}(\bar{d}_i))$ and calculated for the $k$th amino acid. The results are summarized and shown in table 4.

### (*c*) *Electron–ion interaction potential and β-sheet structure*

In this study, the relationships between the spatial distribution of amino acids and two physico-chemical parameters, the EIIP and hydrophobicity, have been investigated. Insight into the distribution of amino acids within the β-sheet conformation can provide additional information for better understanding of the protein-folding process. To analyse the use of these selected parameters for allocation of amino acid position in the β-sheet structure, we introduced a new statistically based approach. The values of $\mathrm{EIIP}_{\mathrm{ave}}(n)$ and $H_{\mathrm{ave}}(n)$ have been calculated using the formulas (2.3) and (2.4), and are shown in figures 2 and 3. The amino acids characterized by the low values of EIIP have a tendency to cluster close to the GC ($r < 0.25\,\mathrm{nm}$). The possible explanation for this is the high concentration of amino acid valine (Val) in the area close to the GC. Val does not form polar bonds with other amino acids (Nelson & Cox 2007), and leucin (Leu) expresses similar behaviour. Therefore, we suggest that amino acids located near the GC, can represent a core of the β-sheet. On the other hand, the value of the $\mathrm{EIIP}_{\mathrm{ave}}(2)$ (figure 2) is much higher.

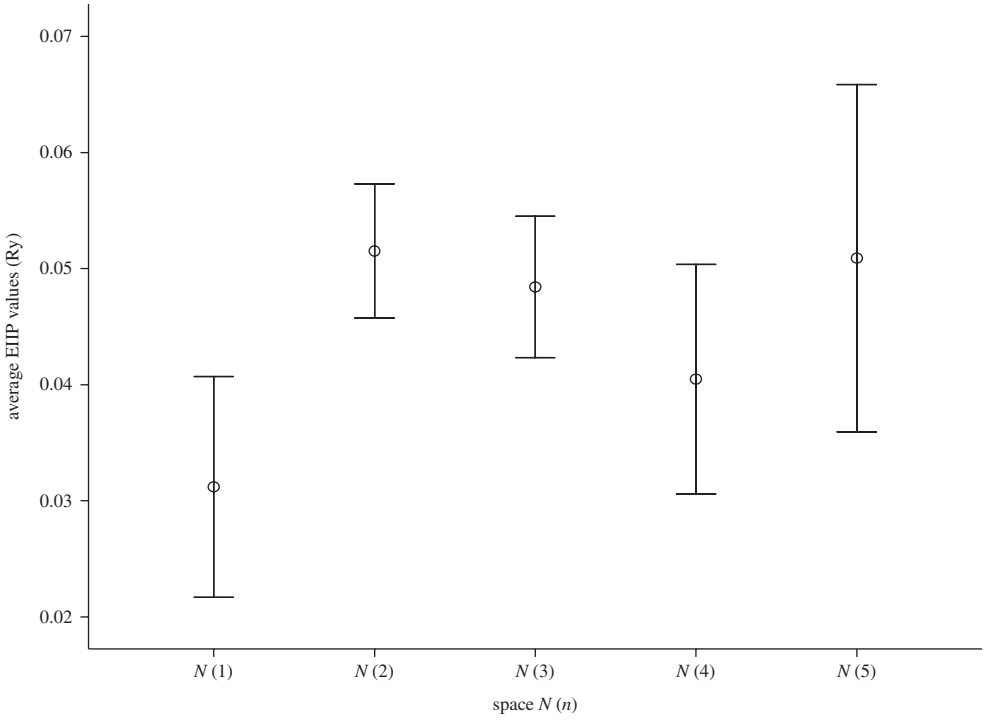


Figure 2. The 95% error box plot representing values of the $\mathrm{EIIP}_{\mathrm{ave}}(n)$ for the first $n$ regions.

The hydrophobic and aromatic amino acids Glu, Thr, Trp, Tyr are prevalent in this volume. These amino acids have a large volume, and they are mainly hydrophobic. Therefore, we suggest that they play a role in diminishing the influence of amino acids, which do not belong to the β-sheet, on its stability. For $n > 2$, $\mathrm{EIIP}_{\mathrm{ave}}(n)$ has a value close to the expected values under the CSR hypothesis.

### (*d*) *Hydrophobicity and β-sheet structure*

From figure 3, we can observe that $H_{\mathrm{ave}}(n)$ is a decreasing function that has its maximum in the volume around the GC, and also it has a significant drop between $n = 1$ and $n = 2$; $H_{\mathrm{ave}}/H_{\mathrm{ave}}(1) = 1.2$ comparing with $\mathrm{EIIP}_{\mathrm{ave}}(2)/\mathrm{EIIP}_{\mathrm{ave}}(1) = 1.7$. Results revealed that the distribution of the amino acids presented by the hydrophobicity parameter is a non-uniform and a nonlinear (figure 3) process. This implies that an amino acid positioned close to the GC ($r < 0.35\,\mathrm{nm}$) as well as far from the GC ($r > 1.4\,\mathrm{nm}$) are significantly more hydrophobic than amino acids located at a distance of $0.35\,\mathrm{nm} < r < 1.4\,\mathrm{nm}$ to the GC.

In contradiction, by using the EIIP parameter, we can cluster the amino acids that are closely positioned to the GC ($r < 0.35\,\mathrm{nm}$). These defined amino acids Val, Ile and Leu (table 5) are important for the stability of the whole β-sheet, and thus this knowledge can contribute to a better understanding of the protein-folding process and improvement of protein design.

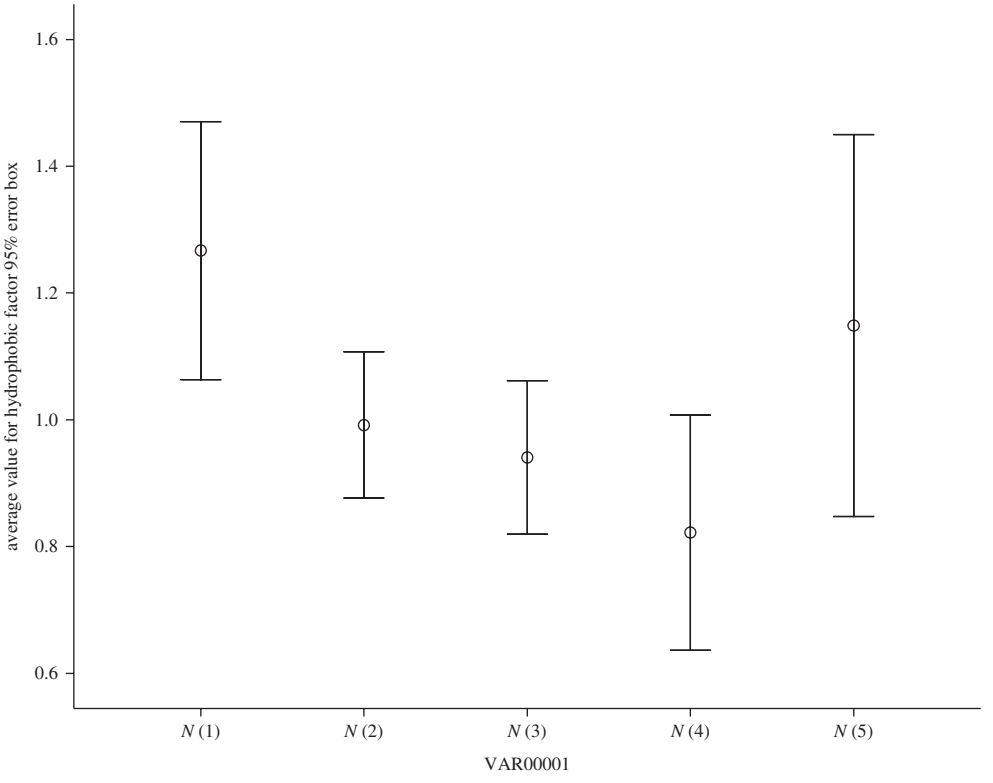


Figure 3. The 95% error box plot representing values of the $H_{\mathrm{ave}}(n)$ for the first $n$ regions.

### (*e*) *Cross-spectral analysis of amino acid properties*

A number of amino acid parameters were derived to represent amino acids in a protein molecule. In general, they are separated into groups: α-propensities, β-propensities, hydrophobicity and physico-chemical properties of amino acids (Nakai *et al.* 1988). To examine the correlation between the amino acid parameters related to β-sheet conformation, we selected the following indexes: β-propensities, normalized average hydrophobicity scales, EIIP (table 1) and minimal distances from the geometrical centre (new data introduced by the authors, table 4). Thus, the amino acids in the analysed β-sheet were replaced with the corresponding numerical values of the selected parameters. The resulting numerical sequences were then analysed using the multiple cross-spectra to determine the correlation between these parameters. Hence, 38 amino acid properties related to the β-sheet structure were analysed. Multiple cross-spectral analysis was performed for all these parameters and the resulting cross-spectral function is shown in figure 4. From figure 4, we can observe a peak at the frequency $f = 0.265$ regarded as a common frequency component for all analysed numerical sequences. The studied parameters characterize the β-sheet conformation and thus it can be postulated that the prominent peak obtained at $f = 0.265$ is a characteristic feature of a β-sheet. The existence of this prominent peak in

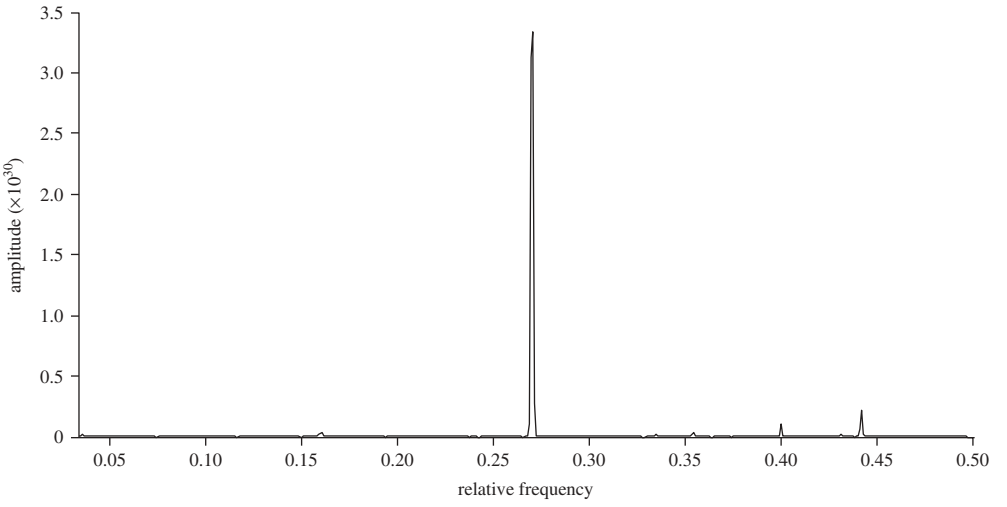


Figure 4. Multiple cross-spectral function of 38 amino acid properties (β-propensities, hydrophobicity, EIIP and four parameters from table 4) calculated for uricase 2yze polypeptide chain.

the cross-spectra of 38 amino acid properties revealed a high cross-correlation between these analysed parameters, which is an important indicator in prediction of β-sheet structures.

### (*f*) *Continuous wavelet transform for analysis of selected protein β-sheet structure*

Owing to limitation of the space delocalization in the FT used in the multiple cross-spectral analysis outlined above, we applied the CWT transformation to study the position/locations of the patterns in the analysed numerical sequences at the characteristic frequency, $f = 0.265$. The EIIP and hydrophobicity parameter values were assigned to each amino acid in the analysed polypeptide chain to convert the original amino acid sequences into the numerical sequences. For calculation of the Wavelet coefficients, the standard function 'CWT' from computer software package Matlab Inc., v. 7, was used. To estimate the wavelet scale coefficient corresponding to the relative frequency $f = 0.265$, we calculated the CWT for a purely periodic signal. This procedure provided us with the approximate scale coefficient around 6.5 for the scale of 1–10, which corresponds to the β-sheet structure. The CWT distributions of two selected proteins: uricase protein (2yze) polypeptide chain (figures 4 and 5) for EIIP and hydrophobicity as parameters, respectively, and 4-hydroxyphenylpyruvate dioxygenase (1sqd) polypeptide chain (figures 6 and 7) for hydrophobicity and EIIP as parameters. Applying the WT leads to the possibility of observing a whole frequency/spatial distribution along the sequence and thus, identifying domains of high energy for a particular frequency for the analysed numerical sequence. The results obtained show that the Morlet wavelet function used for analysis of two selected polypeptide chains is appropriate for the detection of biologically important regions of these β-sheet structures as can be seen

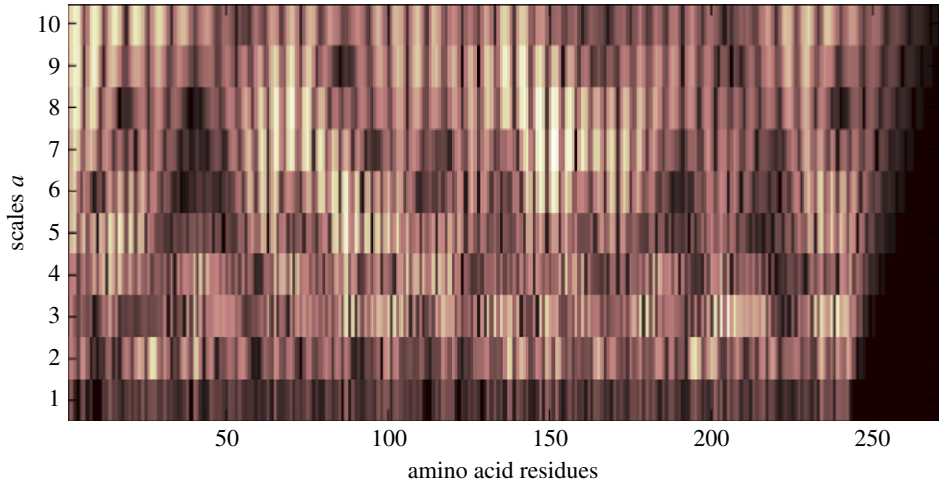


Figure 5. Morlet wavelet coefficient distribution calculated for uricase 2yze chain using the EIIP. (Online version in colour.)

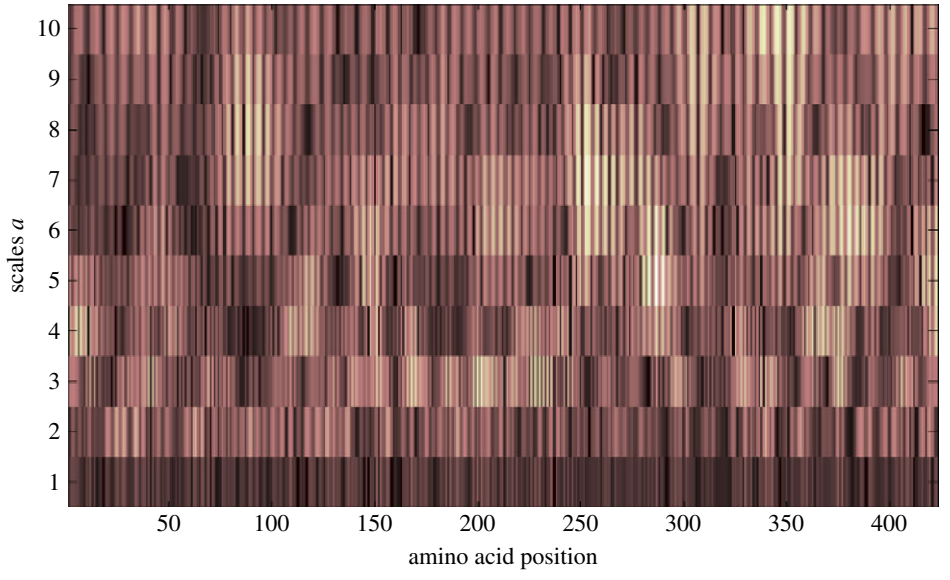


Figure 6. Morlet wavelet coefficient distribution calculated for 4-hydroxyphenylpyruvate dioxygenase (1sqd) using the EIIP. (Online version in colour.)

from the continuous scalograms of uricase protein (2yze) polypeptide chain and hydroxyphenylpyruvate dioxygenase (1sqd) polypeptide chain (figures 4–7). The real form of the Morlet wavelet function is $\omega(t) = e^{(-t^2/2)} \cdot \cos(5t)$. There are two constants in this function, 2 and 5. As the constant 2 determines the waveform amplitude modulation degree and 5 determines the centre frequency, they are named here as the *amplitude factor* and the *frequency factor*, respectively.

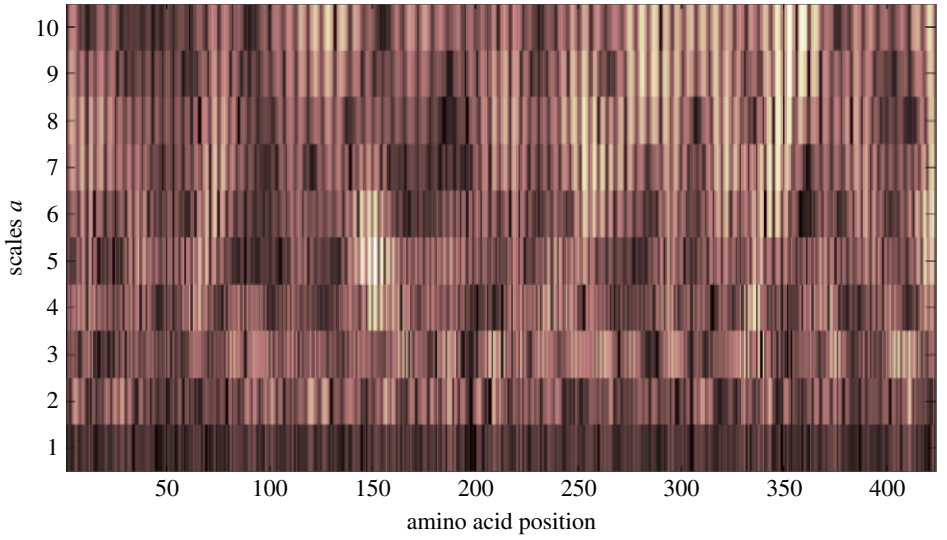


Figure 7. Morlet wavelet coefficient distribution calculated for 4-hydroxyphenylpyruvate dioxygenase (1sqd) using the hydrophobicity. (Online version in colour.)

To find out if we can optimize the function for use with protein sequences, we have modified both these factors to produce wavelets having similar shape but different centre frequencies and modulation degrees. All the scalograms generated here have maximum scale at 10.

## 4. Conclusion

Up to date research studies have been focused on investigating different hydrophobicities of the β-sheet configuration. The attempts have been undertaken to either analyse the whole β-sheet structure or its inner part separately from the border part (Chou & Fasman 1978). In aiming to perform an efficient protein design, it is necessary to consider more specific statistical data that can characterize the spatial positions of each amino acid within the β-sheet structure. In this study, we introduced a new hybrid approach, based on spatial statistic and signal-processing methods, for the analysis of amino acid distributions in the β-sheet structure. The statistical analysis of amino acids that form β-sheet configuration with respect to their positions along the specific strands was performed. The statistical approach presented here is based on a key feature of this study; the so-called GC. By investigating the hypothesis of the CSR, we found that for the majority of amino acids, this process cannot be regarded as random in close proximity to the GC. Moreover, with increasing distance from the GC, the process can be regarded as the CSR.

In this study, we also analysed the efficacy of amino acid parameters, $EIIP_{ave}(n)$ and $H_{\mathrm{ave}}(n)$ to describe amino acid distributions within a β-sheet. The results obtained demonstrate the distinctive patterns in the spatial distribution of amino acids within close proximity (central region of β-sheet) to GC

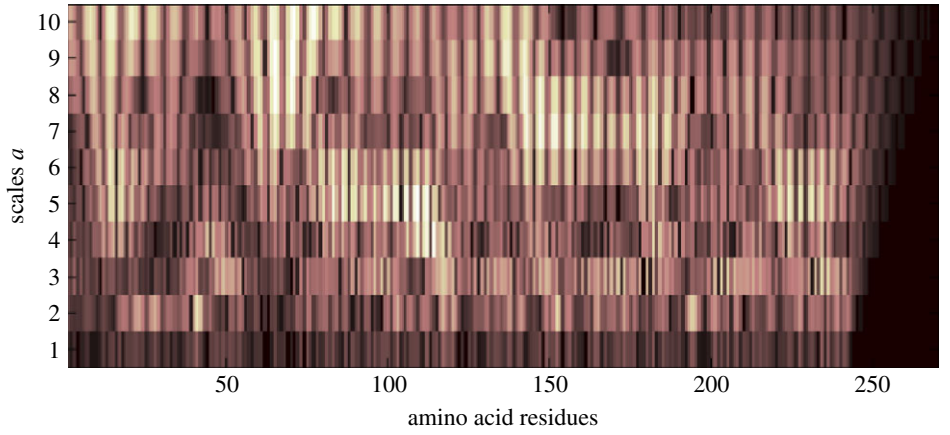


Figure 8. Morlet wavelet coefficient distribution calculated for uricase 2yze chain using the hydrophobicity. (Online version in colour.)

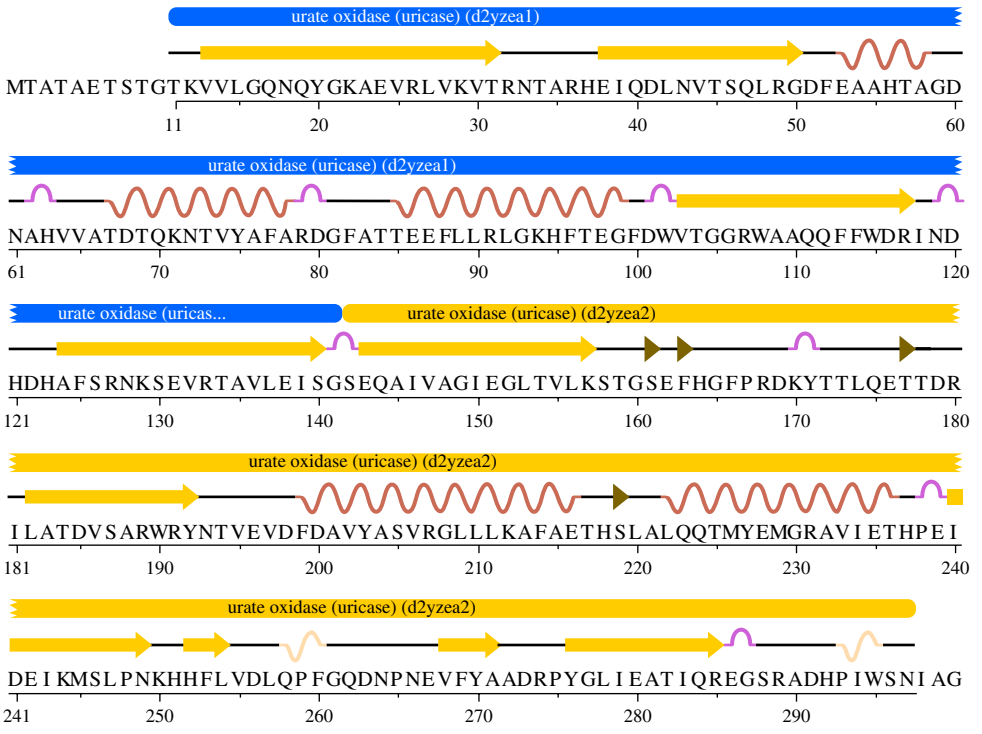


Figure 9. Uricase protein 2yze polypeptide chain. The structure of 2yze has eight chains in total. (Online version in colour.)

(figures 2 and 3) for both $\mathrm{EIIP}_{\mathrm{ave}}(n)$ and $H_{\mathrm{ave}}(n)$ parameters. These findings could be explained by the fact of the existence of spatially conserved structures or a core in the area of the β-sheet structure at a short distance from the GC. Therefore, we can raise a question with regard to the evolution of the β-sheet by considering the tendency for the conserved characteristics of the β-sheet area

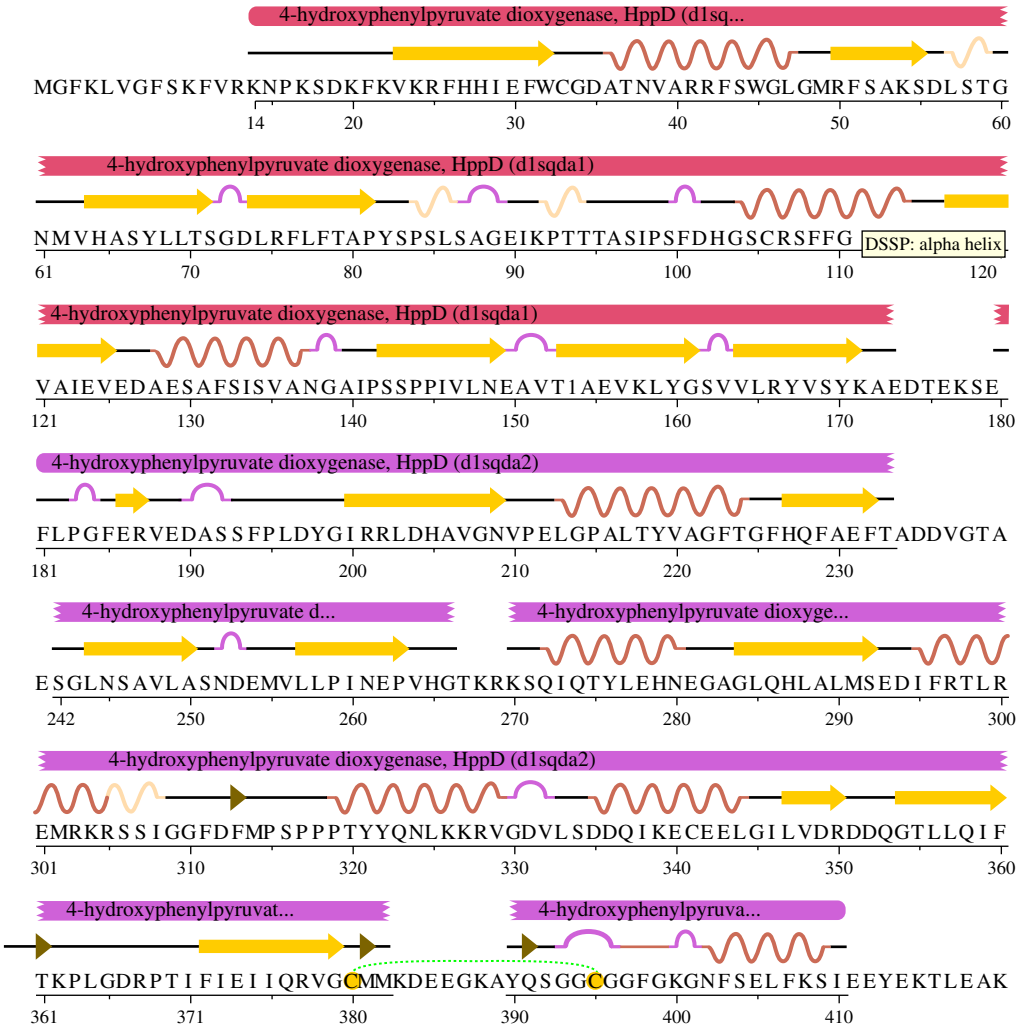


Figure 10. 4-Hydroxyphenylpyruvate dioxygenase (1sqd) polypeptide chain. (Online version in colour.)

close to the GC. It should be noted that almost all amino acids located at a small distance to the GC are hydrophobic. It is known that hydrophobic amino acids build up the inner part of the β-sheet (Chou & Fasman 1978). Thus, the continuous decrease in the value of $H_{\mathrm{ave}}(n)$ as shown in figure 3 is expected. More intriguing results were obtained for the $\mathrm{EIIP}_{\mathrm{ave}}(n)$ parameter. It was shown that the $\mathrm{EIIP}_{\mathrm{ave}}(n)$parameter is efficient in predicting the locations of amino acids Ile, Val and Leu in close proximity to the GC ($r < 0.35$ nm, table 1). Using the $\mathrm{EIIP}_{\mathrm{ave}}(n)$, we have also determined the locations of Phe, Thr, Trp, Tyr and Glu amino acids within the β-sheet to be at a distance of $0.35\,\mathrm{nm} < r < 0.7\,\mathrm{nm}$. These data are important for more accurate prediction of two-dimensional protein structures and thus, can be useful for protein design. In addition, multiple cross-spectral analysis was performed to determine the relationship

between 38 amino acid properties related to the β-sheet structure. The results obtained showed a high cross-correlation between the analysed amino acid parameters, which is useful information for β-sheet structure predictions. This paper presents the results of the application of Morlet wavelet function for the identification of functionally important regions in the analysed β-sheet structures. The domains identified at the 12–31; 38–50;102–118; 123–158; 181–191 amino acids within the CWT correspond to the functionally important regions defined experimentally for uricase protein (2yze) polypeptide chain; and areas 22–32; 62–81; 140–172; 200–209; 242–250 were predicted for the second polypeptide chain 4-hydroxyphenylpyruvate dioxygenase (1sqd), respectively (figures 5–8). These sites predicted by the CWT correspond to the site(s) determined experimentally by other authors (figures 9 and 10). This is largely owing to the advantageous properties of the space–frequency analysis pertinent to the CWT.

There are many different bioinformatics approaches used for the prediction of β-sheet structures; however, these methods are based on energy-minimization algorithms (Summa & Levitt 2007; Levy-Moonshine *et al.* 2009). Here, we presented and discussed the hybrid approach that incorporates statistical and signal-processing techniques. This novel approach enables us to gain additional insights into the spatial distribution of amino acids in the β-sheet structure that cannot be retrieved using other bioinformatics tools. The results showed that particular amino acids are not randomly distributed within a given β-sheet. This finding can be used as a restrictive factor to be incorporated into the currently used algorithms for β-sheet structure predictions. Using this restrictive factor, it becomes possible to increase the computational efficiency of the algorithm by reducing its processing time.

## References

Campagne, S., Saurel, O., Gervais, V. & Milon, A. 2010 Structural determinants of specific DNA-recognition by the THAP zinc finger. *Nucleic Acids Res.* **38**, 3466–3476. (doi:10.1093/nar/gkq053)

Chou, P. Y. & Fasman, G. D. 1978 Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47**, 251–276. (doi:10.1146/annurev.bi.47.070178.001343)

Cid, H., Bunster, M., Canales, M. & Gazitua, F. 1992 Hydrophobicity and structural classes in proteins. *Prot. Eng.* **5**, 373–375. (doi:10.1093/protein/5.5.373)

Clark, P. J. & Evans, F. C. 1954 Distance to the nearest neighbour a measure of spatial relationship in populations. *Ecology* **35**, 445–453. (doi:10.2307/1931034)

Cosic, I. 1994 Macromolecular bioactivity: is it resonant interaction between molecules?—theory and applications. *IEEE Trans. Biomed. Eng.* **41**, 1101–1114. (doi:10.1109/10.335859)

Cosic, I. 1995 Virtual spectroscopy for fun and profit. *Bio-technology.* **13**, 236–238. (doi:10.1038/nbt0395-236)

Cosic, I. 1997 *The resonant recognition model of macromolecular bioactivity: theory and application.* Basel, Switzerland: Birkhauser.

Cosic, I., Vojisavljevic, V. & Pavlovic, M. 1989 Prediction of 'hot spots' in interleukin-2 based on informational spectrum characteristics of growth regulating factors. *Biochimie* **71**, 333–342. (doi:10.1016/0300-9084(89)90005-9)

Fadeev, E. A., Sam, M. D. & Clubb, R. T. 2009 NMR structure of the amino-terminal domain of the lambda integrase protein in complex with DNA: immobilization of a flexible tail facilitates β-sheet recognition of the major groove. *J. Mol. Biol.* **388**, 682–690. (doi:10.1016/j.jmb.2009.03.041)

Kanehisa, M. 1988 A multivariate analysis method for discriminating protein secondary structural segments. *Protein Eng. Des. Sel.* **2**, 87–92. (doi:10.1093/protein/2.2.87)

Levy-Moonshine, A., Amir, M. & Keasar, C. 2009 Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. *Bioinformatics* **25**, 2639–2645. (doi:10.1093/bioinformatics/btp449)

Nakai, K., Kidera, A. & Kanehisa, M. 1988 Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* **2**, 93–100. (doi:10.1093/protein/2.2.93)

Nelson, D. & Cox, M. 2007 *Lehninger principles of biochemistry*, 5th edn. New York, NY: W. H. Freeman and Company.

Noar, D., Fischer, D., Jernigan R. L., Wolfson, H. & Nussinov, R. 1996 Amino acid pear interchanges at spatially conserved locations. *J. Mol. Biol.* **256**, 924–938. (doi:10.1006/jmbi.1996.0138)

Pirogova, E., Fang, Q., Akay, M. & Cosic, I. 2002 Investigation of the structure and function relationships of Oncogene proteins. *Proc. IEEE* **90**, 1859–1867. (doi:10.1109/JPROC.2002.805305)

Pirogova, E., Simon, G. P. & Cosic, I. 2003 Investigation of the applicability of dielectric relaxation properties of amino acid solutions within the resonant recognition model. *IEEE Trans. Nanobiosci.* **2**, 63–69. (doi:10.1109/TNB.2003.813936)

Pirogova, E., Akay, M. & Cosic, I. 2008 Investigating the interaction between oncogene and tumor suppressor protein. *IEEE Trans. Inform. Technol. Biomed.* **13**, 10–15. (doi:10.1109/TITB.2008.2003338)

Smialowski, P., Martin-Galiano, J. A., Mikolajka, A., Girschick, T., Holak, T. A. & Frishman, D. 2007 Protein solubility: sequence based prediction and experimental verification. *Bioinformatics* **23**, 2536–2542. (doi:10.1093/bioinformatics/btl623)

Smith, C. K. & Regan, L. 1997 Construction and design of $\beta$-sheets. *Acc. Chem. Res.* **30**, 153–161. (doi:10.1021/ar9601048)

Summa, C. M. & Levitt, M. 2007 Near-native structure refinement using in vacuo energy minimization. *Proc. Natl Acad. Sci.* **104**, 3181. (doi:10.1073/pnas.0611593104)

Tukey, J. W. 1997 *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Veljkovic, V. 1980 *A theoretical approach to preselection of carcinogens and chemical carcinogenesis*. NewYork, NY: Gordon & Breach.

Veljkovic, V. & Slavic, I. 1972 On the general model of pseudopotentials. *Phys. Rev. Lett.* **290**, 105–108. (doi:10.1103/PhysRevLett.29.105)

Williams, R. W., Chang, A., Juretic, D. & Loughran, S. 1987 Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta* **916**, 200–204. (doi:10.1016/0167-4838(87)90109-9)