Selection of amino acid parameters for Fourier transform-based analysis of proteins

Jadranka Lazović

Abstract

Fourier analysis of the parametric profile of a sequence for the detection and localization of the structural motifs that are characteristic for biologically related proteins has been proposed. In order to select parameters that are most appropriate for this analysis, the informational capacity of 226 physicochemical, thermodynamic, structural and statistical amino acid parameters was analyzed. Based on the results, obtained for the four functionally unrelated protein model groups (lysozyme c, HIV-1 gp120, tubulin and tau proteins, and steroid hormone receptors), the electron-ion interaction potential has been selected as the unique amino acid property that can be used in Fourier transform-based analysis of proteins, independently of their biological function.

Introduction

With the rapid expansion of protein databases, the identification of the biological function of newly sequenced proteins, or the determination of their relationship with defined functional families, becomes a very real problem. This problem is additionally complicated if the primary structure of the new protein is remarkably different from those of proteins with already known biological function. Therefore, the introduction of additional information concerning the relationship between amino acids within the protein sequence would be helpful. One of the most efficient approaches proposed in this direction is Fourier analysis of parametric profiles representing primary structure of proteins (McLachlan and Stewart, 1976; Veljković et al., 1985; McLachlan, 1993). The main problem in this approach represents the proper choice of the amino acid parameter(s) that is relevant for the biological function of the analyzed proteins. For now, only two amino acid properties have been frequently used in this analysis: electron–ion interaction potential (EIIP) (Veljković et al., 1985, 1992; Ćosić et al., 1986, 1989, 1991, 1994; Veljković and Ćosić, 1987; Ćosić and Nešić, 1988; Skerl and Pavlović, 1988; Veljković and Metlaš, 1988, Veljković and Lalović, 1990; Ćosić and Hearn, 1991, 1992) and hydrophobicity (Eisenberg et al., 1984; DeLisi and Berzofsky, 1985; Mitaku et al., 1985; Yanagihara et al., 1989; Suwa et al., 1993). This leaves open the question about the suitability of other amino acid properties in determining the parametric profiles of proteins that are subjected to Fourier analysis.

To answer this question, we compare the informational capacity of 226 physicochemical, thermodynamic, structural and statistical amino acid parameters. The obtained results suggest that, among the analyzed parameters, EIIP represents the unique amino acid property that could be successfully used in structure/function analyses of proteins performed by Fourier analysis, independently of their biological function.

Materials and methods

Fourier analysis of parametric profiles

There are two main problems in protein structure/function analysis: (i) the determination of the unknown biological function of the protein and (ii) the identification of mutually interacting pairs of proteins (such as ligand and receptor). In the first case, the analyzed protein and members of the particular protein functional family may have: (1) common biological function and similar primary structures; (2) common biological function, but different primary structures; (3) similar primary structures, but different biological functions; or (4) both different primary structures and biological functions. These problems, except points (1) in (i), can not be successfully treated with conventional approaches in structure/function analysis. One of the most efficient approaches that could help the solution of these problems is protein analysis based on Fourier transform (Veljković et al., 1980; Ćosić and Hearn, 1991).

Detailed descriptions of protein analyses performed by Fourier analysis of their parametric profiles have been published elsewhere (McLachlan and Stewart, 1976; Veljković et al., 1985). According to this approach, a protein sequence is transformed into a signal (parametric profile) by assigning the corresponding numerical value to each amino acid. These 20 values can represent any
selected physical, chemical, thermodynamic, structural or statistical property of the amino acids. It is assumed that the points in the signal are equidistant with distance \( d = 1 \).

Before computing the discrete Fourier transform, the signal average is subtracted from the original signal. The obtained signal is further decomposed by discrete Fourier transform into a sum of periodical functions. The result is a series of frequencies and corresponding amplitudes and phases. The whole information, primarily defined by the amino acid sequence, is thus presented in the spectral form.

The informational spectrum (IS) of a sequence is defined by Veljković and co-workers (Veljković et al., 1985) as the energy density spectrum of the discrete Fourier transform:

\[
S(n) = X(n)X^*(n) = |X(n)|^2, \quad n = 1, 2, \ldots, N/2
\]

where \( N \) is the total number of points in the signal, \( X(n) \) are discrete Fourier transformation coefficients multiplied by \( N \), and \( S(n) \) are amplitudes in IS at frequency \( n/N \). The maximal frequency in IS is \( F = 1/2d = 0.5 \). The frequency range is independent of the sequence length, yet the sequence length influences the resolution of the spectrum which is \( 1/N \) for an \( N \)-point signal. The obtained amplitudes and corresponding frequencies describe periodic properties in the parametric profile of the protein sequence. To determine the common frequency components for \( K \) protein sequences, their cross-spectrum is calculated as follows:

\[
C(j) = \prod_{i=1}^{K} S(i,j)
\]

where \( S(i,j) \) is the \( j \)th frequency component in IS of sequence \( i \) and \( C(j) \) is the \( j \)th frequency component in the cross-spectrum. In general, peak frequencies in the cross-spectrum of a group of sequences correspond to the frequency components common in all their individual ISs, while frequency components not present in all IS are eliminated in the cross-spectrum. Specifically, the cross-spectrum of a group of sequences with a common biological activity is called the consensus informational spectrum (CIS).

The significance of a particular peak in a spectrum is measured by its amplitude and/or by its signal-to-noise ratio (S/N). The S/N is defined as the ratio between the corresponding amplitude and the mean amplitude of the whole spectrum.

The four steps of comparing two or more sequences are represented in Figure 1 for two acyl carrier proteins (ACP) that operate in different systems—the former one in the fatty acid synthetase (FS) system in *Escherichia coli* and the latter one in the citrate lyase (CL) system in *Klebsiella aerogenes* (Doolittle, 1981). In their optimized alignment, obtained by Fastdb (Brutlag et al., 1990), these two sequences display only 24% residue identity.

Methods based on Fourier transform can be used for comparison of proteins with non-similar primary structures because these methods do not require any sequence alignment.

**Parameters representing amino acids**

In this study, 226 amino acid parameters are compared. The first 222 of them are from the Amino Acid Index Database (Nakai et al., 1988) and the last four (223—the valence number, 224—the mean valence number, 225—the Fermi energy and 226—the ion-ion interaction potential) from unpublished results. Prior to the analysis, each parameter \( p_k \) (\( k = 1, 2, \ldots, 226 \)) was normalized using the following formula:

\[
p'_{ki} = (p_{ki} - \bar{p}_k)\sqrt{\frac{\sum_{j=1}^{20}(p_{kj} - \bar{p}_k)^2}{20}}
\]

where \( p_{ki} \) is the value of parameter \( k \) for amino acid of type \( i \), \( \bar{p}_k \) is the mean value of parameter \( k \) and \( p'_{ki} \) is the value of the normalized parameter \( k \) for amino acid of type \( i \). Parameters normalized in this way, presented as vectors \( \bar{p}'_k = (p'_{k1}, p'_{k2}, \ldots, p'_{k20}) \), have a unitary length, i.e. \( |\bar{p}'_k| = 1 \). Later in the text, by parameters we will consider the normalized parameters.

**Selection of model groups**

For the comparison of amino acid parameters, it is first necessary to determine a model group, i.e. a learning set of \( N \) proteins that, according to the above-defined problems (i) and (ii), must either (1) belong to a particular functional group of proteins, i.e. share the same biological activity, or (2) represent pairs of mutually interacting proteins. Some additional groups of sequences must be also used: the positive control with proteins sharing the same biological activity as proteins from the corresponding model group and the negative control comprising the same number of proteins with unrelated biological activity. Here a substantially larger set (always 128) of randomly selected proteins is also used. These proteins are retrieved from the Swiss-Prot protein database (Bairoch and Boeckmann, 1992) using a combination of a pseudo-random number generator and an additional condition to correspond by length (± 10%) to the mean length of the proteins from the corresponding model group. The proteins used here both in the positive and in the negative control also correspond by their length to the proteins from the model group.

Theoretically, the analysis for a given model group can be extended to some specific proteins related in their
Selection of amino acid parameters

\[\text{ACP-FS (Escherichia coli)}\]
STIEERVKII1GEQLGKVQEEVTDNASFVE
DLGADSLDTVELMAEEMEFDEEAE
KITTQQAIDYINGQA

\[\text{ACP-CL (Klebsiella aerogenes)}\]
MEMKIDALAGTLESSDVIMRGPAQPGIQ
LEIDSIVKQFGQAIIQVYSETLQAQLGVKE
CDVQLARVQAALRWQQ

\[\text{Fig. 1. An example of the four steps of comparing two or more sequences in a protein sequence analysis method based on Fourier transform, in this case two acyl carrier proteins ACP-FS and ACP-CL: (a) the protein sequences; (b) their parametric (EIIP) profiles; (c) their individual ISs; (d) their cross-spectrum. The prominent peaks denote common frequency components. The abscissa represents frequencies and the ordinate amplitudes in the spectrum.}\]
activity to members from the model group (proteins of type $x$) and to proteins unrelated, in that particular manner, to the model group (proteins of type $y$). We considered two biologically interesting cases. In the first one, protein of type $x$ shares the same biological activity as proteins from the model group, while protein of type $y$ has a different activity, although being more similar in its primary structure to proteins from the model group than the former one. In the other case, proteins of type $x$ and $y$ are mutually functionally related, but only the first one is involved in the biological activity defined by the model group by interacting with the members of the model group.

This analysis is performed on four functionally distinct model groups: (a) lysozyme $c$, as an example of an enzyme; (b) HIV-1 gp120, as a viral hypervariable envelope protein; (c) tubulin and tau proteins, as mutually interacting structural proteins forming the cytoskeleton; (d) different steroid hormone receptors, all DNA-binding proteins from the same superfamily. Table I summarizes the database identifications of all protein sequences used in this study, except of those belonging to the four groups of randomly selected proteins. The comparison of each model group with the corresponding negative control and randomly selected sequences, performed using Fastdb (Brutlag et al., 1990) to align the consensus sequence of the model group with individual members of these other groups, showed that each model group has a mean residue identity with the sequences from these other groups that is $< 20\%$. The consensus sequences for the model groups were obtained using the multiple aligning program Genalign (Martinez, 1988).

### Table I. Selection of protein sequences

<table>
<thead>
<tr>
<th>Group</th>
<th>Protein SEquences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(a) Lysozyme c</strong></td>
<td>**</td>
</tr>
<tr>
<td>MG</td>
<td>LYC_BOVIN, LYC_CHRAM, LYC_COLLI, LYC_HORSE, LYC_HUMAN, LYC_LOPCA, LYC_LOPLE, LYC_NUMME, LYC_ORTE, LYC_PAPAN, LYC_RABIT, LYC_RAT, LYC_SHEEP,</td>
</tr>
<tr>
<td>pc</td>
<td>LYC_AXIAK, LYC_CHICK, LYC_COLVI, LYC_MELGA, LYCM_MOUSE, LYCP_MOUSE, LYC_PHACO, LYC_PREEN, LYS_SYRRE,</td>
</tr>
<tr>
<td>nc</td>
<td>ATPE_ARATH, CYB5_CHICK, DNBI_ADE07, FLGB_SALTY, GSHX_HUMAN, H31_SCHPO, MYP2_HUMAN, RS6_ECOLI, RUB1_PSEOL;</td>
</tr>
<tr>
<td>x</td>
<td>LYG_STRCA;</td>
</tr>
<tr>
<td>y</td>
<td>LCA_SHEEP;</td>
</tr>
<tr>
<td><strong>(b) HIV-1 gp120</strong></td>
<td>**</td>
</tr>
<tr>
<td>MG*</td>
<td>ENVELOPEADA, ENV6ALA1, ENVELOPEBRA, ENVCHAN, ENV6HAR83, ENV6JARF1, ENVELOPEAL, ENV6MFA, ENV6NDA, ENV6OPU43, ENV6QY1, ENV6PESC;</td>
</tr>
<tr>
<td>pc*</td>
<td>ENVELOPEBRU, ENVELOPEBX2, ENVELOPEMN, ENVELOPEFS2;</td>
</tr>
<tr>
<td>nc</td>
<td>BENA_ACICA, CYS4_BRANA, NAHH_PSEPU, TF_MOUSE;</td>
</tr>
<tr>
<td>x</td>
<td>CD4_HUMAN;</td>
</tr>
<tr>
<td>y</td>
<td>CD8A_HUMAN;</td>
</tr>
<tr>
<td><strong>(c) Tubulin and tau proteins</strong></td>
<td>**</td>
</tr>
<tr>
<td>MG</td>
<td>TBA_PIG, TBA1_RAT, TBA2_MOUSE, TBA4_HUMAN, TBA5_CHICK, TBB_PIG, TBB1_HUMAN, TBB1_RAT, TBB2_MOUSE, TBB3_CHICK, TA12_BOVIN, TA34_BOVIN, TAU1_HUMAN, TAU2_MOUSE, TAU3_MOUSE;</td>
</tr>
<tr>
<td>pc</td>
<td>TBA1_CHICK, TBA1_HUMAN, TBA6_MOUSE;</td>
</tr>
<tr>
<td>nc</td>
<td>TBB5_HUMAN, TBB5_MOUSE, TBB6_CHICK, TA35_BOVIN, TAU1_RAT, TAU2_HUMAN;</td>
</tr>
<tr>
<td>x</td>
<td>ADH_FRAAN, BENA_ACICA, CPC5_RABIT, CYS4_BRANA, DHA_BACSH, KITH_HSTVT, LACG_LACCA, NAHH_PSEPU, TF_MOUSE;</td>
</tr>
<tr>
<td>y</td>
<td>CD8A_HUMAN;</td>
</tr>
<tr>
<td><strong>(d) Steroid hormone receptors</strong></td>
<td>**</td>
</tr>
<tr>
<td>MG</td>
<td>ANDR_RAT, ANDR_MOUSE, ESTR_RAT, ESTR_CHICK, GCRA_HUMAN, GCR_MOUSE, GCR_RAT, PRGR_CHICK (type a), PRGR_HUMAN;</td>
</tr>
<tr>
<td>pc</td>
<td>ANDR_HUMAN, ESTR_MOUSE, GCRB_HUMAN, PRGR_CHICK (type b);</td>
</tr>
<tr>
<td>nc</td>
<td>CLAT_PIG, KFPS_AVISP, NFM_RAT, RRP3_IAWIL;</td>
</tr>
</tbody>
</table>

*The sequences of HIV-1 gp120 were taken from the Human retroviruses and AIDS database (Myers et al., 1992) and all other sequences from Swiss-Prot 22 (Barroch and Boeckmann, 1992).

**The average percentage of identical residues within the sequence groups was calculated from the pairwise optimized alignments generated by the program Fastdb (Brutlag et al., 1990).
In the case of the lysozyme c model group (Table I), lysozyme g (LYG_STRCA) was considered as a protein of type x as it shares the same catalytic activity as lysozyme c (Weaver et al., 1985) yet has an average of 21% identical residues with the members of the model group, while α-lactalbumin (LCA_SHEEP) was considered as a protein of type y as it has a different function than lysozyme c (Nitta and Sugai, 1989), although having an average of 38% identical residues with members of the model group.

For the HIV-1 gp120 model group, the receptor for gp120 (CD4_HUMAN) was used as a protein of type x, while a similar membrane protein which does not interact with gp120 (CD8A_HUMAN) was used as a protein of type y.

The Fourier transform-based procedure and criteria for the selection of amino acid parameters

A parameter is appropriate for protein structure/function analyses performed by Fourier analysis only if the repetitivity of its distribution along a protein sequence, i.e. along the parametric profile, correlates with the biological activity of proteins. This parameter will generate in CIS of functionally related proteins one dominant peak corresponding to their common biological activity (Veljkovic et al., 1985).

The procedure used for the selection of amino acid parameters involves the following steps:

1. The calculation of CIS$^i$ of N proteins from the selected model group for all 226 parameters ($i = 1, \ldots, 226$), and the determination of the maximal amplitude ($A^{\text{max}}_i$) and the corresponding signal-to-noise ratio ($S/N^{\text{max}}_i$) and the corresponding frequency $f^{\text{max}}_i$ in each of these 226 CIS$^i$.

2. The calculation of cross-spectra between each CIS$^i$ and IS$^p_j$ of each of the M proteins from the positive control, and the calculation of the mean amplitude ($A^{p \text{ average}}_j$) and the mean signal-to-noise ratio ($S/N^{p \text{ average}}_j$) for frequency $f^{\text{max}}_j$ from all M obtained cross-spectra.

3. The calculation of cross-spectra between each CIS$^i$ and IS$^{n \text{ negative}}_j$ of each of the M proteins from the negative control, and the calculation of the mean amplitude ($A^{n \text{ negative}}_j$) and the mean signal-to-noise ratio ($S/N^{n \text{ negative}}_j$) for frequency $f^{\text{max}}_j$ from all M obtained cross-spectra.

4. The calculation of cross-spectra between each CIS$^i$ and IS$^{c \text{ control}}_j$ of each of the 128 randomly selected proteins ($j = 1, \ldots, 128$), and the calculation of the mean amplitude ($A^{c \text{ control}}_j$) and the mean signal-to-noise ratio ($S/N^{c \text{ control}}_j$) for frequency $f^{\text{max}}_j$ from all 128 obtained cross-spectra.

When proteins of type x and type y are also considered, the procedure contains two additional steps that represent additional controls in the selection of the parameters appropriate for the particular model group:

5. The calculation of cross-spectra between each CIS$^i$ and IS$^{y \text{ y}}_j$ of each of the n$_1$ proteins of type x ($j = 1, \ldots, n_1$), and the determination of the amplitude ($A^{y \text{ y}}_j$) and the signal-to-noise ratio ($S/N^{y \text{ y}}_j$) for frequency $f^{\text{max}}_j$ in all n$_1$ obtained cross-spectra.

6. The calculation of cross-spectra between each CIS$^i$ and IS$^{y \text{ y}}_j$ of each of the n$_2$ proteins of the type y ($j = 1, \ldots, n_2$), and the determination of the amplitude ($A^{y \text{ y}}_j$) and the signal-to-noise ratio ($S/N^{y \text{ y}}_j$) for frequency $f^{\text{max}}_j$ in all n$_2$ obtained cross-spectra.

The selection of parameters was performed according to the following two groups of criteria A and B:

A) The criteria which compare the amplitude values derived from the cross-spectra obtained using the $i$th amino acid parameter ($i = 1, \ldots, 226$):

- $A_1: A^{p \text{ average}}_j - A^{n \text{ negative}}_j > 0$
- $A_2: A^{c \text{ control}}_j - A^{n \text{ negative}}_j > 0$

and, if steps 5 and 6 exist:

- $A_3: A^{y \text{ y}}_j - A_j > 0$ where $j = 1, \ldots, n_1$
- $A_4: A^{y \text{ y}}_j - A_j > 0$ where $j = 1, \ldots, n_1$

and $k = 1, \ldots, n_2$.

Criteria A1 and A2 for parameter $i$ request that, on average, proteins sharing the same biological activity as those from the model group (positive controls) must have in their spectra a more expressed frequency component characteristic for the model group than both randomly selected proteins and functionally unrelated proteins (negative controls). Criteria A3 and A4 request for parameter $i$ that each protein that either has the same biological activity as in the model group or is involved in that activity by interacting with the members of the model group (protein of type x) must have in its IS a more expressed frequency component characteristic for the model group than randomly selected proteins on average, and than each functionally unrelated protein of type y.

B) The criteria which compare the $S/N$ values derived from the cross-spectra obtained using the $i$th amino acid parameter ($i = 1, \ldots, 226$):

- $B_1: S/N^{p \text{ average}}_j - S/N^{n \text{ negative}}_j > 0$
- $B_2: S/N^{c \text{ control}}_j - S/N^{n \text{ negative}}_j > 0$
- $B_3: S/N^{y \text{ y}}_j - S/N^{n \text{ negative}}_j > 0$

and, if steps 5 and 6 exist:

- $B_4: S/N^{y \text{ y}}_j - S/N^{n \text{ negative}}_j > 0$ where $j = 1, \ldots, n_1$
- $B_5: S/N^{y \text{ y}}_j - S/N^{n \text{ negative}}_j > 0$ where $k = 1, \ldots, n_2$. 
Fig. 2. Graphic presentation of the amino acid parameters chosen as applicable in Fourier transform-based structure/function analyses of the corresponding groups of proteins. The abscissa represents the serial number of the parameter in the database (Nakai et al., 1988). The ordinates represent values of log (C1) and C2.
In other words, criteria B for parameter $i$ request that, on average, both positive controls and proteins of type $x$ increase the S/N corresponding to the characteristic frequency of the model group, while randomly selected, functionally unrelated proteins (negative controls) and proteins of type $y$ decrease it. According to criteria A and B, a corresponding subset of parameters is determined (selected) for each model group separately.

In spectra in which all amplitudes but one are close to zero, which is the case in CIS of the chosen model groups (with nine or more proteins with a common biological property), the S/N of the only prominent peak tends to be close to the total number of frequencies in the spectrum, determined by the resolution. Therefore, S/N is not sensitive enough for mutual comparison of amino acid parameters. As this is not the case for the amplitudes, and also as all amino acid parameters were normalized prior to the Fourier transform, the efficiency of each parameter selected by criteria A and B is further estimated by two measures based on the amplitudes. These are $C_1 = A_{\max}$, measuring the 'strength', and $C_2 = \tilde{A}/\tilde{A}'$, measuring the 'selectivity' of the parameter for the given model group.

The described Fourier transform-based procedure for one protein model group and a set of amino acid parameters was implemented in a Borland Turbo Pascal 5.5 program available on request from the author (elazovij@ubbg.etf.bg.ac.yu).

**Results and discussion**

The aim of this study was to select, among 226 parameters representing different physicochemical, thermodynamic, structural and statistical amino acid properties, those that are most suitable for structure/function analyses of proteins, performed by Fourier analysis of their parametric profiles. The comparison of parameters was performed in two steps: (1) by selecting the parameters whose repetitiveness along the sequences of functionally related proteins correlates with their common biological activity (i.e. those satisfying criteria A and B); and (2) by their sorting according to the expressed 'selectivity' (i.e. according to the value $C_2$). The results obtained for the four functionally unrelated model groups of proteins are summarized in Figure 2 and in Table II. It should be emphasized that, for a same set of protein sequences, common motifs (characteristics) in their parametric profiles, as well as their distribution along the sequences, depend on the values assigned to the amino acids. In the spectral representation obtained using the selected parameters, this means that the frequency corresponding to the maximal amplitude in CIS ($f_{\max}$), representing the common biological activity of the model group, is not the same for different parameters (Table II).

The results obtained reveal that, for lysozyme $c$, only 10 of all 226 parameters satisfy criteria A and B (parameters 25, 32, 46, 52, 85, 111, 115, 163, 178 and 204), meaning that only these parameters might be successfully used in Fourier analyses of this group of proteins (Figure 2a). However, these 10 parameters are not equally efficient in their ability to distinguish proteins functionally related to lysozyme $c$ from other proteins. As shown in Table Ii, according to the value $C_2$, the most 'selective' among them is parameter 204. According to the value $C_1$, parameters 52 and 204 are 'stronger' than the other eight selected parameters.

For the second model group, composed of 13 HIV-1 gp120 that all bind the CD4 molecule, 18 parameters were found to satisfy criteria A and B (3, 10, 20, 22, 29, 41, 46, 53, 87, 96, 135, 144, 181, 182, 194, 204, 205 and 209). However, for this model group parameter 182 might be too general as it gives in CIS the maximal amplitude on the first point, revealing that common repetitive motifs with lengths below the length of the analyzed sequences are absent in the corresponding parametric profiles. Generally, the presence of $f_{\max}$ at the first position in CIS of any group of proteins suggests that the corresponding parameter is not sufficiently 'sensitive' for analyzing these proteins. These results suggest that only the other 17 selected parameters are applicable in Fourier analyses of HIV-1 gp120 (Figure 2b). When sorted according to 'selectivity' ($C_2$), parameters 209 and 204 appear to be the best among them, with parameters 181, 10 and 41 ensuing, respectively, as listed in Table Iib. According to 'strength' ($C_1$), the most distinguished are parameters 194, 209, 204, 3, and 181, respectively. These results suggest that, among all 226 compared amino acid parameters, 209, 204 and 181 are the most suitable for Fourier transform-based structure/function analyses of HIV-1 gp120.

Compared with the two previous model groups, the procedure for the estimation of amino acid parameters has, in the case of the model group of tubulin and tau proteins, only the first four steps. As the positive control covers three groups of proteins (tubulin $\alpha$, tubulin $\beta$ and tau protein), we required that criteria A1, A2 and B1 must be satisfied not only on the level of the whole positive control, but also on the level of each of these three groups. In this way a total of eight parameters has been selected (parameters 25, 33, 100, 135, 137, 138, 204 and 219). For this model group, parameter 219 gives in CIS the maximal amplitude on the first point, suggesting that only the other seven selected parameters are suitable for Fourier analyses of tubulin and tau proteins (Figure 2c). For this model group (Table Iic), parameter 138 is the best according to its 'strength', while parameter 204 is the best according to its 'selectivity'.
Table II. The amino acid parameters found to be applicable in Fourier transform-based analysis of the four protein model groups

<table>
<thead>
<tr>
<th>No.*</th>
<th>Amino acid parameter</th>
<th>$f^{\text{max}}$</th>
<th>$S/N^{\text{max}}$</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>204</td>
<td>Electron–ion interaction potential (Veljković et al., 1985)</td>
<td>0.328</td>
<td>127.6</td>
<td>$3.3 \times 10^{18}$</td>
<td>9.5 ± 0.3</td>
</tr>
<tr>
<td>115</td>
<td>Normalized frequency of $C$ (Maxfield and Scheraga, 1976)</td>
<td>0.148</td>
<td>108.9</td>
<td>$5.3 \times 10^{19}$</td>
<td>8.8 ± 0.3</td>
</tr>
<tr>
<td>178</td>
<td>Information measure for extended without H-bond (Robson and Suzuki, 1976)</td>
<td>0.336</td>
<td>116.9</td>
<td>$5.1 \times 10^{17}$</td>
<td>6.0 ± 0.2</td>
</tr>
<tr>
<td>88</td>
<td>Flexibility parameter for no rigid neighbors (Karplus and Schulz, 1985)</td>
<td>0.281</td>
<td>124.5</td>
<td>$6.4 \times 10^{17}$</td>
<td>4.7 ± 0.2</td>
</tr>
<tr>
<td>32</td>
<td>Frequency of the 3rd residue in turn (Chou and Fasman, 1978)</td>
<td>0.148</td>
<td>87.8</td>
<td>$1.7 \times 10^{17}$</td>
<td>4.3 ± 0.2</td>
</tr>
<tr>
<td>46</td>
<td>Molecular weight (Fasman, 1976)</td>
<td>0.285</td>
<td>90.8</td>
<td>$8.5 \times 10^{16}$</td>
<td>4.0 ± 0.2</td>
</tr>
<tr>
<td>163</td>
<td>Average relative fractional occurrence in E$_{(i-1)}$ (Rackovsky and Scheraga, 1982)</td>
<td>0.035</td>
<td>96.4</td>
<td>$7.7 \times 10^{15}$</td>
<td>3.8 ± 0.2</td>
</tr>
<tr>
<td>111</td>
<td>Conformational preference for antiparallel β-strands (Lifson and Sander, 1979)</td>
<td>0.199</td>
<td>94.1</td>
<td>$5.9 \times 10^{15}$</td>
<td>3.4 ± 0.2</td>
</tr>
<tr>
<td>26</td>
<td>Normalized frequency of C-terminal non-helical region (Chou and Fasman, 1978)</td>
<td>0.039</td>
<td>105.7</td>
<td>$2.2 \times 10^{14}$</td>
<td>3.3 ± 0.2</td>
</tr>
<tr>
<td>52</td>
<td>Helix–coil equilibrium constant (Finkelstein and Pitsyn, 1977)</td>
<td>0.039</td>
<td>127.6</td>
<td>$2.0 \times 10^{14}$</td>
<td>2.7 ± 0.2</td>
</tr>
</tbody>
</table>

The above results point out that, among all 226 tested amino acid parameters, the EIP (Veljković and Slavić, 1972; Veljković, 1973; Veljković et al., 1985) represents the unique property that may be successfully used in Fourier analysis of parametric profiles of the proteins from the four analyzed model groups. The EIP values for the 20 amino acids are listed in Table III. Assuming that each parameter can be chosen with the same probability, the probability of selecting one and the same parameter (i.e. EIP in this case) in four independently chosen subsets

For the model group of steroid receptors (SR), a total of 36 parameters satisfy criteria A1, A2 and B1–B3 (parameters 8, 20, 32, 35, 37, 53, 61, 63, 64, 72, 80, 82, 85, 92, 94, 96, 97, 109, 111, 116, 124, 132, 137, 139, 140, 149, 151, 153, 158, 176, 177, 190, 194, 204, 212 and 223), yet only six of them (61, 72, 94, 116, 204 and 212) do not have in the corresponding CIS the maximal amplitude at the first point (Figure 2d and Table IIc), suggesting that only these six parameters are suitable for Fourier transform-based structure/function analyses of SR.

*The serial number of the parameter in the database (Nakai et al., 1988).

The parameters are sorted according to their 'selectivity' determined by the value of C2 (see the text).
Table III. Values of EIIP for the amino acids

<table>
<thead>
<tr>
<th>Amino acid residue</th>
<th>EIIP (Ry)*</th>
<th>EIIP normalized**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>0.0373</td>
<td>-0.0667</td>
</tr>
<tr>
<td>Arg</td>
<td>0.0959</td>
<td>0.2674</td>
</tr>
<tr>
<td>Asn</td>
<td>0.0036</td>
<td>-0.2589</td>
</tr>
<tr>
<td>Asp</td>
<td>0.1263</td>
<td>0.4408</td>
</tr>
<tr>
<td>Cys</td>
<td>0.0829</td>
<td>0.1933</td>
</tr>
<tr>
<td>Gln</td>
<td>0.0761</td>
<td>0.1545</td>
</tr>
<tr>
<td>Glu</td>
<td>0.0058</td>
<td>-0.2463</td>
</tr>
<tr>
<td>Gly</td>
<td>0.0050</td>
<td>-0.2509</td>
</tr>
<tr>
<td>His</td>
<td>0.0242</td>
<td>-0.1414</td>
</tr>
<tr>
<td>Ile</td>
<td>0.0000</td>
<td>-0.2794</td>
</tr>
<tr>
<td>Leu</td>
<td>0.0000</td>
<td>-0.2794</td>
</tr>
<tr>
<td>Lys</td>
<td>0.0371</td>
<td>-0.0679</td>
</tr>
<tr>
<td>Met</td>
<td>0.0823</td>
<td>0.1899</td>
</tr>
<tr>
<td>Phe</td>
<td>0.0946</td>
<td>0.2600</td>
</tr>
<tr>
<td>Pro</td>
<td>0.0198</td>
<td>-0.1665</td>
</tr>
<tr>
<td>Ser</td>
<td>0.0829</td>
<td>0.1933</td>
</tr>
<tr>
<td>Thr</td>
<td>0.0941</td>
<td>0.2572</td>
</tr>
<tr>
<td>Trp</td>
<td>0.0548</td>
<td>0.0331</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.0516</td>
<td>0.0148</td>
</tr>
<tr>
<td>Val</td>
<td>0.0057</td>
<td>-0.2469</td>
</tr>
</tbody>
</table>

The EIIP values are given in Rydbergs (Veljković et al., 1985).

**Normalized according to equation (3).

The results obtained suggest that EIIP is the most suitable known amino acid property that can be used in structure/function Fourier-based analyses of proteins generally, i.e. independently of their biological function. Contrary to the parameter-independent techniques that depend on learning sets of biologically related proteins (Viari et al., 1990), Fourier-based protein structure/function analyses based on a particular and universally applicable parameter also allow the study of proteins that are not classified in any of the known functional families.

This study also points out EIIP as the property that could help a better understanding of the physical characteristics of proteins that are responsible for their biological activity; namely, the most important step in biological processes based on highly selective interactions between macromolecules (enzyme–substrate, ligand–receptor, antibody–antigen, etc.) is the efficient recognition which precedes the formation of any covalent bond and which takes place when the molecules are still at a relatively long distance. Such a recognition could be accomplished through electric forces determined by the electrostatic potential around a molecule, which depends on the distribution and energy states of the valence electrons, that are best described by EIIP (Veljkovic, 1980; Chapeville and Haenni, 1980; Politzer, 1988). It is important to note that strong correlations between EIIP and different biological properties of small organic molecules (such as carcinogenicity, toxicity, antibiotic activity or cytostatic activity) have been already reported (Veljkovic, 1980). Therefore, in future protein structure/function analyses, EIIP should be considered with due attention.

Acknowledgements

The authors thank V. Veljkovic, R. Metlaš and V. Skerl for useful discussions and the referees for their comments on the previous version of the manuscript. This work was supported by the Fund for Scientific Research of Serbian Republic (Yugoslavia).

References


Received on March 26, 1996; revised on June 6, 1996; accepted on June 18, 1996.