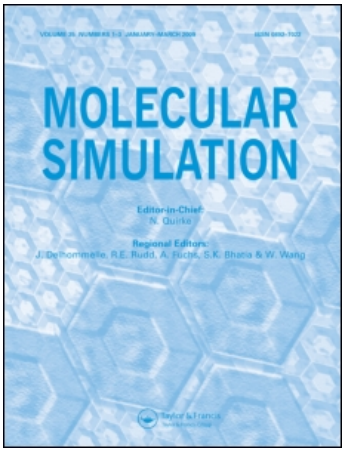


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

The use of ionisation constants of amino acids for protein signal analysis within the resonant recognition model--application to proteases

Elena Pirogova[a]; Irena Cosic[a]
[a] Bioelectronics Group, Department of Electrical and Computer Systems Engineering, Monash University, Victoria, Australia

Online publication date: 26 October 2010

To cite this Article Pirogova, Elena and Cosic, Irena(2002) 'The use of ionisation constants of amino acids for protein signal analysis within the resonant recognition model--application to proteases', Molecular Simulation, 28: 8, 853 — 863
To link to this Article: DOI: 10.1080/0892702021000002539
URL: http://dx.doi.org/10.1080/0892702021000002539

Molecular Simulation, 2002 Vol. 28 (8–9), pp. 853–863

THE USE OF IONISATION CONSTANTS OF AMINO ACIDS FOR PROTEIN SIGNAL ANALYSIS WITHIN THE RESONANT RECOGNITION MODEL—APPLICATION TO PROTEASES

ELENA PIROGOVA and IRENA COSIC*

Bioelectronics Group, Department of Electrical and Computer Systems Engineering, PO Box 35, Monash University, Victoria 3800, Australia

(Received May 2001; in final form November 2001)

Biological functions of proteins and their active 3D structures are determined by the linear sequences of amino acids. The resonant recognition model (RRM) is a physico-mathematical model developed for structure/function analysis of protein and DNA sequences. Here, we are comparing results of the RRM analysis [1,2] of protease proteins using the electron–ion interaction potential (EIIP) and ionisation constant (IC) of amino acids. The results obtained reveal that the IC parameter can be successfully used to determine the characteristic patterns of different functional protease subgroups.

Keywords: Digital signal processing; Molecular modelling; Ionisation constant; Amino acids; Protein function

INTRODUCTION

The resonant recognition model (RRM) is based on the finding that there is a significant correlation between spectra of the numerical presentation of amino acids and their biological activity [1]. The RRM interprets the protein primary structure, i.e. amino acid sequence, using digital signal analysis methods [1,2]. It has been shown that certain periodicities (frequencies) within the distribution of energies of delocalised electrons along the protein molecule are critical for

*Corresponding author. E-mail: irena.cosic@eng.monash.edu.au

ISSN 0892-7022 © 2002 Taylor & Francis Ltd
DOI: 10.1080/0892702021000002539

protein biological function (i.e. interaction with its target). Once the RRM characteristic frequency for a particular biological function or interaction has been determined, it is possible then to identify the individual amino acid's so called "hot spots", or domains that contribute mostly to the characteristic frequency and thus to the protein's biological function as well. Originally, the amino acid sequences are transformed into numerical series using the electron–ion interaction potentials (EIIP) for each amino acid in the sequence [1,2]. Here, we present the ionisation constant parameter (IC) of amino acids, which could also be suitable to identify common characteristics for the whole protease protein family.

METHOD AND BACKGROUND

The RRM has been employed in this study for the determination of the biological profile of protein groups under examination and for investigation of the possible usage of the IC parameter in further RRM analysis instead of the EIIP. The RRM comprises two stages. The first involves the transformation of the amino acid sequence into a numerical sequence. Each amino acid is represented by the value of the EIIP [3], which describes the average energy states of all valence electrons in particular amino acid. The EIIP values for each amino acid were calculated using the general pseudopotential model [3] as follows:

$$\langle k+q|w|k\rangle = 0.25\,Z\sin(\pi 1.04\,Z)/(2\pi) \tag{1}$$

where q is a change of momentum of the delocalised electron in the interaction with potential w (EIIP), while:

$$Z = \left(\sum Zi\right)/N \tag{2}$$

where Zi is the number of valence electrons of the i-th component of each amino acid and N is the total number of atoms in the amino acid. Each amino acid or nucleotide, irrespectively of its position in a sequence, can thus be presented by a unique number. Numerical series obtained this way are then analysed by digital signal analysis methods in order to extract information pertinent to the biological function. The original numerical sequence is transformed to the frequency domain using the discrete Fourier transform (DFT). As the average distance between amino acid residues in a polypeptide chain 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 $d = 1$. Then the maximum frequency in the

spectrum is $F = 1/2d = 0.5$. The total number of points in the sequence influences the resolution of the spectrum only. Thus for an N-point sequence the resolution in the spectrum is equal to $1/N$. The n-th point in the spectral function corresponds to the frequency $f = n/N$. In order to extract common spectral characteristics of sequences having the same or similar biological function, the following cross-spectral function was used:

$$Sn = XnYn^* \quad n = 1, 2, \ldots N/2 \tag{3}$$

where Xn are the DFT coefficients of the series $x(m)$ and Yn^* are complex conjugate DFT coefficients of the series $y(m)$. Peak frequencies in the amplitude cross-spectral function denote common frequency components of the two sequences analysed.

To determine the common frequency components for a group of protein sequences, we have calculated the absolute values of multiple cross-spectral function coefficients M, which are defined as follows:

$$|Mn| = |X1n| \cdot |X2n| \ldots |XMn| \quad n = 1, 2, \ldots, N/2 \tag{4}$$

Peak frequencies in such a multiple cross-spectral function denote common frequency components for all sequences analysed.

The signal-to-noise ratio (S/N) for each peak is defined as a measure of similarity between sequences analysed. S/N is calculated as the ratio between signal intensity at the particular peak frequency and the mean value over the whole spectrum. The extensive experience gained from previous research [1,2] suggests that a S/N ratio of at least 20 can be considered as significant. The multiple cross-spectra function for a large group of sequences with the same biological function has been named the "consensus spectrum". The presence of a peak frequency with significant S/N in a consensus spectrum implies that all of the analysed sequences within the group have one frequency component in common. This frequency is related to the biological function provided the following RRM criteria are met:

- One peak only exists for a group of protein sequences sharing the same biological function.
- No significant peak exists for biologically unrelated protein sequences.
- Peak frequencies are different for different biological functions.

Through an extensive study, the RRM has reached a fundamental conclusion: *one RRM characteristic frequency characterises one particular biological function or interaction* [1,2].

In our previous studies, the above criteria were tested with over 1000 proteins from 25 functional groups [1,2]. The following fundamental conclusion was drawn from those studies: each specific biological function of protein or regulatory DNA sequence(s) is characterised by a single frequency. Once the characteristic frequency for a particular protein function/interaction is defined, it is possible then to utilise the RRM approach to predict the amino acids in the sequences, which predominantly contribute to this frequency and are likely to be crucial for the observed function as well as to design peptides having desired periodicities [4,5]. Such designed peptides expressed the desired biological function, as already shown in preliminary examples of FGF peptidic antagonists [4] and HIV agonist [5]. The assignment of a particular number for each amino acid is a crucial step in all RRM calculations. This set of numbers should have a physical meaning related to the protein's biological function. Although amino acid indices [6,7] have been found to correlate in some ways with the biological activity of the whole protein, our investigations [1,2] have shown that optimum correlation can be achieved with parameters, which are related to the energy of the delocalised electrons of each amino acid. These findings can be explained by the fact that the electrons delocalised in the particular amino acid have the strongest impact on the electronic distribution of the whole protein. In our

TABLE I EIIP and ionisation constant values

Amino acid	*EIIP*	*IC*
L	0	2.40
I	0	2.40
N	0.0036	2.20
G	0.0050	2.46
V	0.0057	2.35
E	0.0058	2.30
P	0.0198	2.00
H	0.0242	2.30
K	0.0371	2.20
A	0.0373	2.30
Y	0.0516	2.20
W	0.0548	2.37
Q	0.0761	2.06
M	0.0823	2.17
S	0.0829	2.10
C	0.0829	1.96
T	0.0941	2.09
F	0.0946	1.98
R	0.0959	1.82
D	0.1263	1.88
Correlation coefficient		−0.794

previous studies it has been shown that the choice of the proper parameter is critical and that the majority of them are not suitable for the RRM approach [8,9].

Despite the successful usage the EIIP values in previous studies [1,2,4,5], there was a number of criticisms regarding the approximate nature of the EIIP. Thus, in our recent investigations we have tried to apply the IC as the parameter instead of the EIIP to represent each amino acid in the sequence [10–12]. IC is an experimentally measurable physical property [13,14] and is more appropriate than the EIIP to use since the later was obtained from an approximate theoretical calculation involving many assumptions.

The experimental values of the IC are measured at 25°C. The method for the determination of the IC in most cases is direct potentiometric titration, while in some cases are optical methods [13,14]. Titration measurements at high ionic strength predominantly reflect only electrostatic interactions between nearby charged groups. For intact proteins, a total potentiometric titration is a moderately simple experiment when sufficient quantities of proteins are available. However, the interpretation of the results is difficult because it is often not clear to which residue to assign a particular region of the titration curve. The corresponding values of ionisation constant at room temperature are taken from references [15] and presented in Table I (column 3). The correlation calculations were carried out sequentially between IC parameter values and the referential EIIP values in order to test for any linear relationship between them. Parameters having correlation coefficients in excess of $+/-$ 0.5 were deemed to be the most strongly correlated [16]. The calculated correlation coefficient is -0.794 showing that there is a high correlation between EIIP and IC parameter values.

Here, each amino acid in the protein sequence is then represented by the corresponding value of the IC parameter instead of by the EIIP value as was done previously. Numerical series obtained in this way are analysed as previously by transforming them into frequency domain using DFT in order to extract information pertinent to the biological function.

RESULTS

We have investigated in this study seven different protein families.

These protease functional subgroups are:

- cysteine protease (23 sequences),
- metallo-protease (17 sequences),
- serine protease (38 sequences),
- aspartic protease (47 sequences),

- trypsin/chymotrypsin (15 sequences),
- AIDS related protease (25 sequences)
- and protease inhibitors (26 sequences).

Sequences from all selected functional groups were tested and a multiple cross-spectral analysis was performed for each protein group, using EIIP and IC parameter values consequently (Table I). As a result protein characteristic frequencies for each protein functional group were obtained. The peak frequency and signal-to-noise values for each analysed protein group are shown in Table II.

DISCUSSION AND CONCLUSION

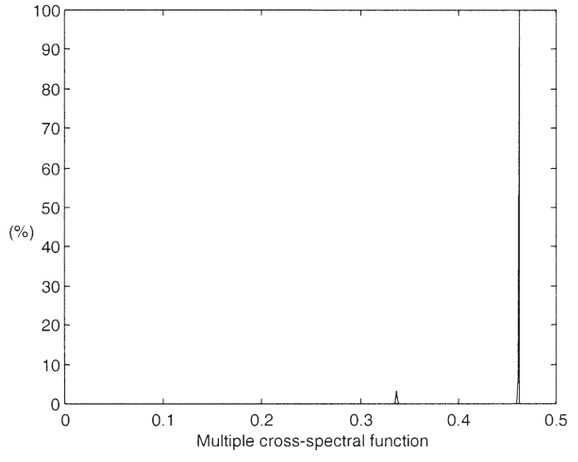
The assumption used here is that sequences with the same biological function share the same characteristic periodic component in the distribution of energy of delocalised electrons.

It could be observed that in both cases (using EIIP and IC values) each specific biological function is characterised by a single frequency. Results obtained have shown that the analysed IC parameter generates in the consensus spectrum one dominant peak (Figs. 1–3) corresponding to common biological activity of the studied proteins and is satisfied all RRM criteria listed previously.

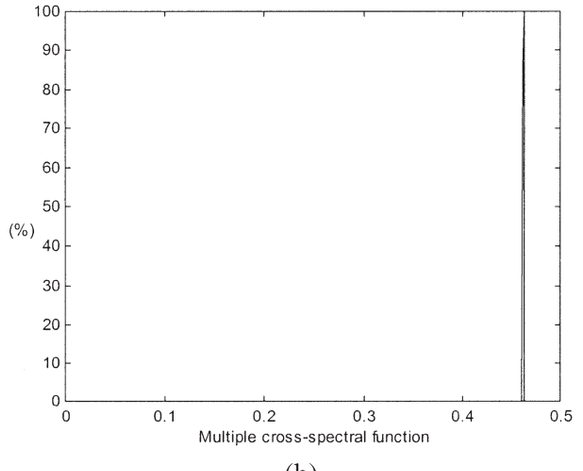
It could be observed from Table II that for serine proteases and protease inhibitors the same characteristic frequencies (within the calculation error) were obtained using both EIIP and IC parameters. This similarity is to be expected, as there is a strong correlation -0.794 between these two parameters. The analogy of the peak frequencies implies that in this particular frequency the analysed protein group reveals the same specific biological function. Based on our previous investigations we conclude that the significance of the correlation between selected parameters is then reflected by the analogy of the spectra of the

TABLE II Peak frequency and signal-to-noise ratio values for protein groups

Parameter	*EIIP*		*IC*	
Protein Group	*Frequency*	*S/N*	*Frequency*	*S/N*
Cysteine	0.0596	197.9	0.4902	248.1
Metallo	0.2520	268.4	0.1699	165.9
Serine	0.4619	463.9	0.4619	274.5
Aspartic	0.1816	429.9	0.2178	167.4
Trypsin/Chymotrypsin	0.3457	178.3	0.0332	174.7
AIDS proteins	0.3447	481.0	0.0127	489.5
Inhibitors	0.3564	179.9	0.3584	244.0

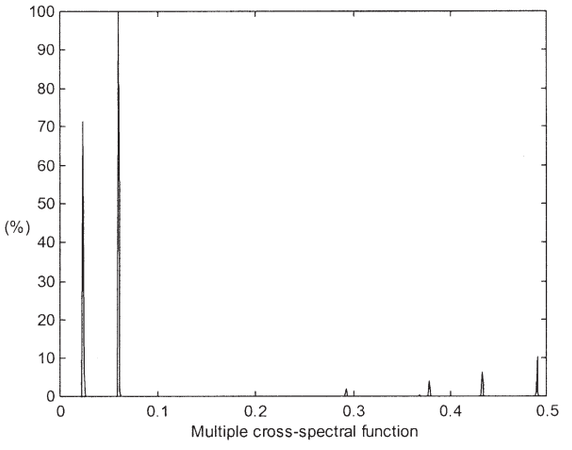


(a)

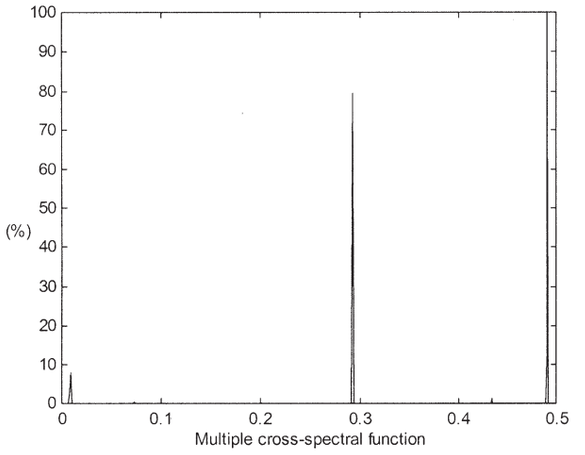


(b)

FIGURE 1 Multiple cross-spectral function of serine proteases using: (a) EIIP parameter, (b) IC parameter. The prominent peaks denote common frequency components. The abscissa represents RRM frequencies, and the ordinate is the normalised intensity.

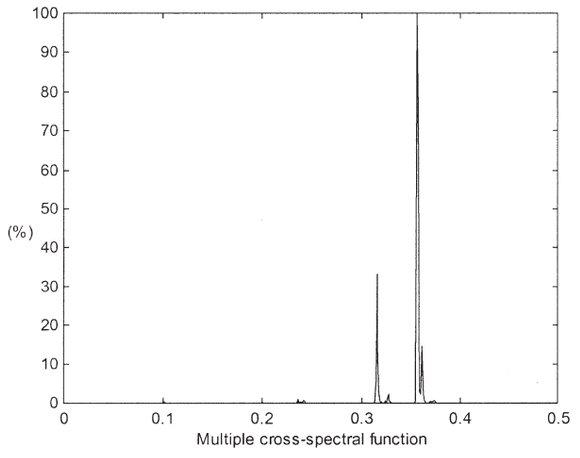


(a)

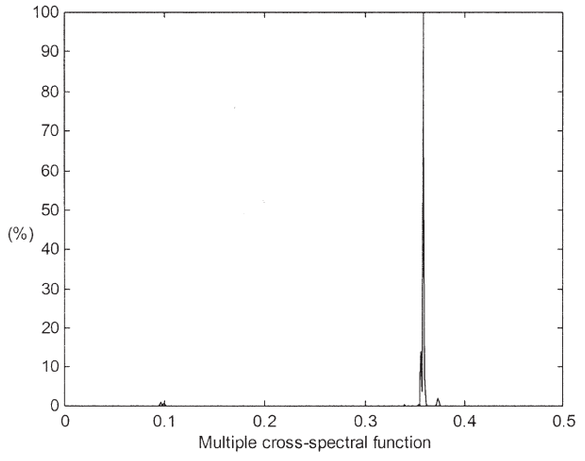


(b)

FIGURE 2 Multiple cross-spectral function of cysteine proteases using: (a) EIIP parameter, (b) IC parameter. The prominent peaks denote common frequency components. The abscissa represents RRM frequencies, and the ordinate is the normalised intensity.



(a)



(b)

FIGURE 3 Multiple cross-spectral function of protease inhibitors using: (a) EIIP parameter, (b) IC parameter. The prominent peaks denote common frequency components. The abscissa represents RRM frequencies, and the ordinate is the normalised intensity.

proteins analysed. Furthermore, this analogy implies that we can then detect that particular characteristic frequency which is relevant to common biological behaviour of the whole functional protein group. Therefore, the correlation coefficient value of the studied parameter could be considered as one of the major factors influenced on the selection of the analysed parameter for its further usage for structure/function analysis of unrelated proteins within the RRM.

Of interest is the fact that for the rest of the analysed functional groups: cysteine, metallo, aspartic, trypsin/chymotrypsin proteases and AIDS related proteases the characteristic frequencies are different. It could be explained by the fact that frequencies are different for different biological functions. Although the frequencies are different using EIIP and IC parameters, the results obtained are satisfied the RRM criteria mentioned above and therefore we conclude that the introduced IC parameter could be also used in the RRM analysis. However, as the IC parameter is a measurable physical property, the use of the IC within the RRM would be significant in better understanding of the physical nature of the protein's biological function.

Finally, we have compared results of the RRM analysis of protease proteins using both EIIP and IC parameters. The results obtained have shown that the IC parameter can successfully identify common characteristic frequency for all selected protease functional subgroups. However, although in some cases the IC is more suitable for structure/function analysis of different protein families it is still difficult to conclude which one of two comparable parameters solely would be the best parameter to use for the determination of biological profiles of different functional protein groups.

References

[1] Cosic, I. (1994) "Macromolecular bioactivity: is it resonant interaction between molecules?—Theory and applications", *IEEE Trans. Biomed. Eng.* **41**, 1101–1114.
[2] Cosic, I. (1997) "The resonant recognition model of macromolecular bioactivity", *BioMethods* (Birkhauser, Basel) **Vol. 8**.
[3] Veljkovic, I. and Slavic, M. (1972) "General model of pseudopotentials", *Phys. Rev. Lett.* **29**, 105–108.
[4] Cosic, I., Drummond, A.E., Underwood, J.R. and Hearn, M.T.W. (1994) "*In vitro* inhibition of the actions of basic FGF by a novel 16 amino acid peptide", *Mol. Cell. Biochem.* **130**, 1–9.
[5] Krsmanovic, V., Biquard, J.-M., Sikorska-Walker, M., Cosic, I., Desgranges, C., Trabaud, M.-A., Whitfield, J.F., Durkin, J.P., Achor, A. and Hearn, M.T.W. (1998) "Investigations into the cross-reactivity of rabbit antibodies raised against nonhomologous pairs of synthetic peptides derived from HIV-1 gp 120 proteins", *J. Pept. Res.* **52**, 410–420.
[6] Kanehisa, M. (1988) "A multivariate analysis method for discriminating protein secondary structural segments", *J. Protein Eng.* **2**, 87–92.
[7] Kawashima, S. and Kanehisa, M. (2000) "Aaindex: amino acid index database", *Nucleic Acids Res.* **28**, 374.

Pirogova, E., Fang, Q., Lazoura, E. and Cosic, I. (1998) "Analysis of amino acid parameters in the resonant recognition model", *Proceedings of the 2nd International Conference on Bioelectromagnetism*, 71–72.

Pirogova, E. and Cosic, I. (2001) "Examination of amino acid indices within the resonant recognition model", *Proceedings of the 2nd Conference of the Victorian Chapter of the IEEE EMBS*, 124–127.

Cosic, I. and Pirogova, E. (1998) "Applications of ionisation constants of amino acids for protein signal analysis within the resonant recognition model", *Proceedings of 20th Annual International Conference of the IEEE EMBS* **20**(2), 1072–1075.

[11] Cosic, I., Fang, Q. and Pirogova, E. (1998) "Modification of the RRM model using wavelet transform and ionisation constants to predict protein active sites", *IEEE EMBS* **21**(2), 1215–1217.

[12] Cosic, I. and Pirogova, E. (2000) "Usage of ionisation constant of amino acids for protein signal analysis within the RRM—application to Oncogene", *Proceeding of the IEEE-EMBS Asia-Pacific Conference on Biomedical Engineering*, 413–414.

[13] Perrin, D.D., Dempsey, B. and Serjeant, E.P. (1981) *pKa Prediction for Organic Acids and Bases* (Chapman and Hall, London).

[14] Albert, A. and Serjeant, E.P. (1971) *The Determination of Ionisation Constants: A Laboratory Manual* (Chapman and Hall, London).

[15] Serjeant, E.P. and Dempsey, B. (1979) *Ionisation Constants of Organic Acids in Aqueous Solutions, IUPAC* (Pergamon Press, New York).

[16] Munro, B.H. (1997) *Statistical Methods for Health Care Research* (Lippincott, Philadelphia).