CROATICA CHEMICA ACTA CCACAA **75** (4) 899–908 (2002)

ISSN-0011-1643
CCA-2840

Original Scientific Paper

Resonant Recognition Model Defines the Secondary Structure of Bioactive Proteins

Nikola Štambuk,[a,*] *Paško Konjevoda,*[a] *Biserka Pokrić,*[a] *Igor Barišić,*[b] *Roko Martinić,*[b] *Vladimir Mrljak,*[c] *and Pero Ramadan*[c]

[a] *Rudjer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia*

[b] *Clinical Hospital Split, Šoltanska 1, 21000 Split, Croatia*

[c] *Faculty of Veterinary Medicine, University of Zagreb, Heinzelova 55, 10000 Zagreb, Croatia*

Received February 21, 2001; revised August 21, 2001; accepted September 5, 2001

The Resonant Recognition Model (RRM) of protein bioactivity is applied to the protein secondary structure prediction. The method is based on the physical and mathematical model of the electron-ion interaction pseudopotential (EIIP) and uses signal analysis to interpret linear information contained in a macromolecular sequence. The method of analysis is based on a two-step procedure. Protein sequence is first transformed into a numerical series by means of the individual EIIP amino acid values. The second step of the model involves the Fourier spectral analysis of the obtained numerical series. Ćosić *et al.*[1–8] have shown that single frequency peaks of the spectrum define characteristic positions of the amino acids, *i.e., hot spots*, correlated to the biological function of the protein. We have analysed the secondary protein structure by comparing the patterns of 20 most prominent frequency peaks of the single-series Fourier RRM periodogram. The patterns within 140 nonhomologous α- and β-protein folds obtained from the Jpred and SCOP databases were analysed by means of the classification tree in order to obtain the algorithm for the α- and β-fold classification. This quick and simple procedure of the secondary fold prediction showed high accuracy of 98.55%. The stability of the tree algorithm solution was confirmed by jack-knife testing of the tree algorithm (mean error 2.6). This method of the secondary structure predic-

* Author to whom correspondence should be addressed. (E-mail: stambuk@rudjer.irb.hr)

tion is presented in more detail on a subset of 30 different cytokines, hormones, enzymes and viral proteins. Our results indicate that resonant spectral analysis of the protein primary amino acid sequence may be used to extract information about its secondary structure.

Key words: resonance, recognition, model, protein folding, secondary structure, prediction, bioactive macromolecules.

INTRODUCTION

The Resonant Recognition Model (RRM) of protein bioactivity is a physical and mathematical model that uses signal analysis to interpret linear information contained in the macromolecular sequence.[1–8] The method of analysis is based on a two-stage procedure.[1–4] The first step involves the transformation of the protein amino acid sequence into a numerical series, which is called the Information Spectrum Method (ISM). Each of the amino acid elements is described by means of the electron-ion interaction pseudopotential value (EIIP).[1–8] This amino acid pseudopotential model represents the average energy states of all valence electrons.[1–8] The second step of the model involves the Fourier spectral analysis of the obtained numerical series.[1–8]

Ćosić *et al.* showed that prominent frequency peaks obtained by means of the spectral analysis denote characteristic positions of the amino acids, *i.e.*, *hot spots*, correlated to the biolological function of the protein.[1–8] The model was confirmed by the successful prediction of a) macromolecular receptor binding; b) enzyme and oncogene activity; c) protein-DNA interactions; d) bioactive parts of cytokines, hormones, viral proteins and antibodies.[1–8]

Despite successful RRM applications in the prediction of particular protein *hot spots* of the protein secondary (and tertiary) structure, the problem of defining its basic structural classes was not solved by the single characteristic frequency approach. Since the defining of the protein secondary structure is essential for defining its structure and function,[1,9,10] we have adopted a slightly different approach to the spectral analysis of the protein RRM sequence. Instead of searching for a single common characteristic frequency peak of different folding types (which is often not found due to sequence diversity irrespective of the secondary structure), we have analysed a limited number of the most prominent frequency peaks in the protein spectrum of different folds.

METHODS

Protein Sequences

The test set of a total of 140 nonhomologous α- and β-protein folds was retrieved from Jpred and SCOP databases for the protein secondary structure prediction.[9–12] The proteins had the structure defined by means of X-ray or NMR. The lengths of the folds ranged from 24 aa to 414 aa, and a set consisted of 70 α-helices and 70 β-sheets. The sequences are listed in the Appendix.

Information Spectrum Method

Protein sequences of the test set were transformed into numerical sequences of the RRM by assigning the electron-ion interaction pseudopotential value in Ry to each amino acid of the macromolecule.[1–8] The values of EIIP for 20 amino acids are given in Table I.

TABLE I

Electron-ion interaction pseudopotential (EIIP)[1] of amino acids

Amino acid	EIIP/Ry
D (Aspartic acid)	0.1263
R (Arginine)	0.0959
F (Phenylalanine)	0.0946
T (Threonine)	0.0941
C (Cysteine)	0.0829
S (Serine)	0.0829
M (Methionine)	0.0823
Q (Glutamine)	0.0761
W (Tryptophan)	0.0548
Y (Tyrosine)	0.0516
A (Alanine)	0.0373
K (Lysine)	0.0371
H (Histidine)	0.0242
P (Proline)	0.0198
E (Glutamic acid)	0.0058
V (Valine)	0.0057
G (Glycine)	0.0050
N (Asparagine)	0.0036
I (Isoleucine)	0.0000
L (Leucine)	0.0000

Spectral Analysis

Molecular resonant analysis of the string spectra was performed by means of a single-series Fourier analysis with the software STATISTICA® for Windows version 5.0 (www.StatSoft.com). Twenty frequency peaks of the largest periodogram values were obtained for each protein sequence, and a new database was constructed. Frequency peak patterns of different α- and β-protein folding spectra were analysed by means of the classification trees, with the software R version 1.1.1 (The R Development Core Team, 2000).

RESULTS AND DISCUSSION

Protein sequences can be easily converted into a numerical sequence by assigning EIIP value (Table I) to each amino acid of the macromolecule.[1–8] An example of this procedure for the basic Fibroblast Growth Factor is given in Figure 1. Following the transformation of the protein amino acid sequence into its information spectrum, the Fourier spectral analysis of the obtained numerical series was performed.

Spectral analysis is concerned with the exploration of the cyclical patterns of data. The purpose of the procedure is to decompose a complex time series with cyclical components into a few underlying sinusoidal (sine and

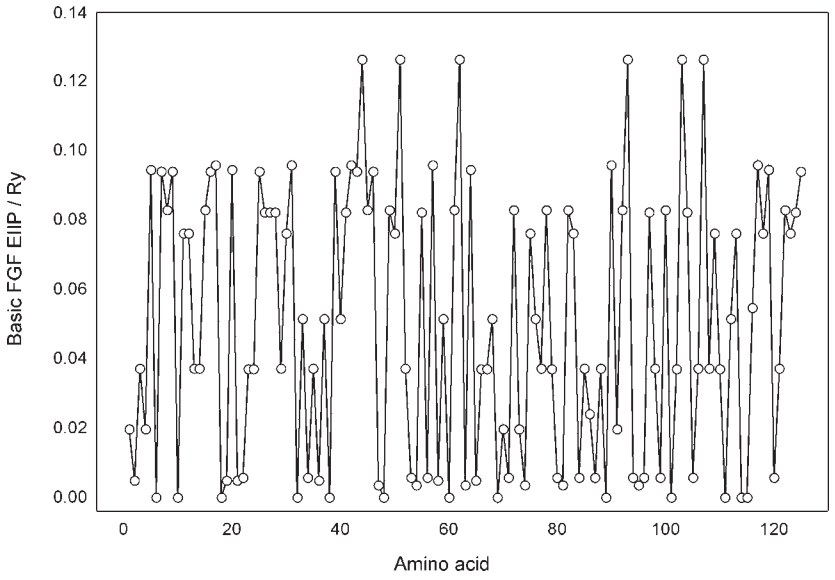

Figure 1. Protein sequence of the basic Fibroblast Growth Factor is transformed into a numerical sequence, *i.e.*, information spectrum, by assigning an electron-ion interaction pseudopotential value (Table I) to each amino acid.

cosine) functions of particular wavelengths. Numerical series representing protein EIIP spectrum function were transformed into the frequency domain using the Discrete Fourier Transform (DFT), *i.e.*, the Fast Fourier Transform (FFT).[1–8,13]

Implementation of the FFT algorithm in the Time Series analysis by means of the STATISTICA® software allows the user to take full advantage of the savings afforded by this algorithm on most standard computers. For the analysis of protein strings, which are of relatively small size (< 1000), the Time Series module of the software uses the simple explicit computational formulas, and the number of computations can be performed in a relatively short amount of time, *i.e.*, the analysis of each string takes only a few seconds.

Frequencies of the 20 largest periodogram values of 140 α- and β-protein folds were compared with respect to the similarities in their patterns (Table II, Figure 2). Twenty frequency peaks were selected by analogy to 20 amino acid elements that constitute the spectrum (Table I). The similarity of the frequency patterns of the α- and β-protein folds was subsequently analysed by means of the classification tree (Figure 3).

TABLE II

Classification of 140 secondary protein folds from the JPred database. Decision tree based pattern analysis was done for 20 largest single series Fourier periodogram values of the frequency parameter. Protein series were obtained by transforming amino acids into the corresponding EIIP values of the Resonant Recognition Model (Table I).

Classification of 70 α- and 70 β-protein folds	
Correctly classified	138
Missclassified	2
Missclassification Error rate	0.0145
Residual mean deviance	0.0482
Jack-knife mean	2.629
Jack-knife SE	12.71

The decision making classification tree is a useful procedure for encapsulating and structuring the knowledge by selecting the variables that enable the best prediction possible.[14] Each terminal node of the tree gives a predicted class and the resulting tree represents the decision making algorithm. The tree based classification model in Figure 3 enabled correct classification of 98.55% of nonhomologous α- and β-protein folds, from their spec-

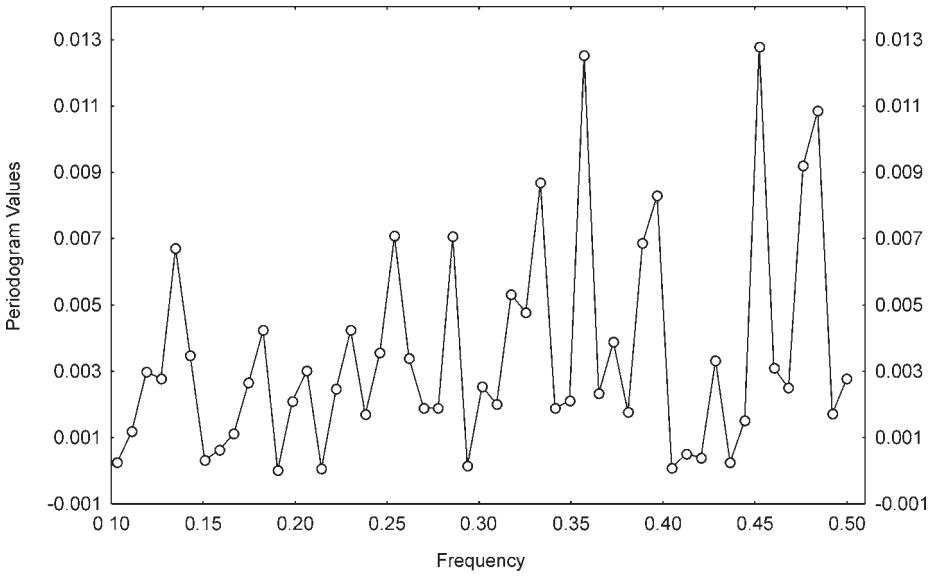

Figure 2. Resonant Recognition Model of the basic Fibroblast Growth Factor (Figure 1, Table I) defined by the frequency parameter of the single series Fourier periodogram.

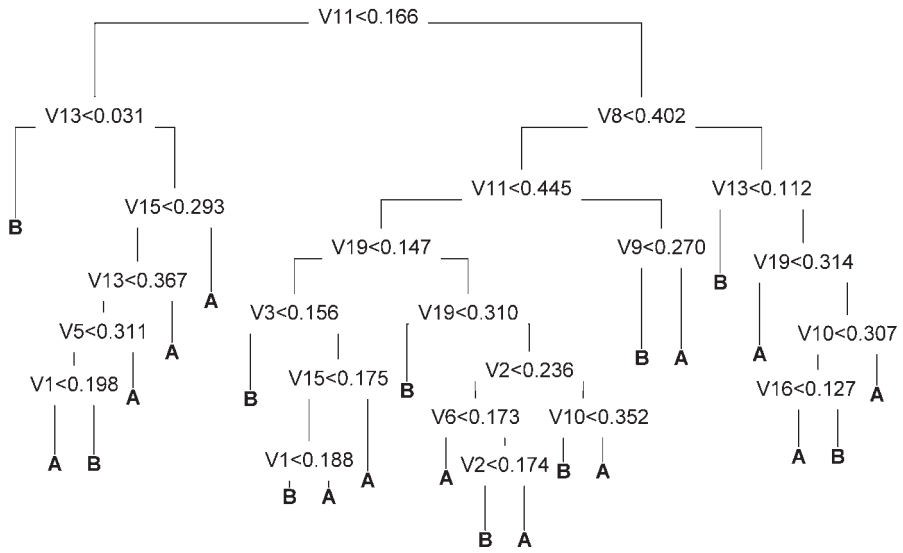

Figure 3. Classification tree for 20 largest Fourier periodogram frequency peaks (V) of the α- (A) and β-protein folds (B). Protein spectrum patterns were obtained by means of the Resonant Recognition Model (Figures 1 and 2, Table I).

tral frequency patterns (Table II). Cross-validation of the procedure by the jack-knife testing[9,14] confirmed the stability and validity of the classification algorithm (Table II). Accurate classification of 30 different α- and β-protein folds of the test set is presented in Tables III and IV. The folds belong to different hormones, cytokines, growth factors, signal trandsduction factors, enzymes and viral proteins. The classification algorithm is stable regardless of the species and protein function (Tables II-IV, Appendix).

TABLE III

Prediction of α-protein folds by means of the classification tree for the analysis of protein RRM frequency patterns (Figure 3)

Class	Predicted Class	Protein	Organism	JPred Code
α-fold	α-fold	Phospholipid binding protein	Homo sapiens	1avhb3
α-fold	α-fold	Phospholipid binding protein	Homo sapiens	1avhb3
α-fold	α-fold	Gag polyprotein	HIV type 1	1hiws-1-AS
α-fold	α-fold	Interleukin 10	Homo sapiens	1lik-1-AS
α-fold	α-fold	Interleukin 10	Homo sapiens	1lik-2-AS
α-fold	α-fold	Leukemia inhib. factor	Homo sapiens	1lki-1-AS
α-fold	α-fold	Apolipoprotein E	Homo sapiens	1lpe-1-DOMAK
α-fold	α-fold	P-26 Ca-binding protein	Bos taurus	1rec1
α-fold	α-fold	P-26 Ca-binding protein	Bos taurus	1rec2
α-fold	α-fold	Colony stimulating factor	Homo sapiens	1rhgc-1-DOMAK
α-fold	α-fold	Coat protein	Tobacco mo. virus	2tmvp
α-fold	α-fold	Vit.-D Ca-binding protein	Bos taurus	3icb
α-fold	α-fold	Interleukin 2	Homo sapiens	3inkd-1-DOMAK
α-fold	α-fold	Phospholipase A2	Bos taurus	4bp2

Our results indicate that the pattern recognition of protein EIIP frequency periodogram obtained by the spectral Fourier analysis enables accurate classification of the secondary protein folds. Searching for the »common characteristic frequency« peak of different folding types[1,2] is not the most appropriate method for defining protein folding. The latter is probably due to the low information content of a single periodogram peak, unable to provide a discriminating parameter for the complex structure, which is highly diverse, and consequently often spectrally deviant, with respect to the species and function. The decision tree based pattern analysis of a large number of Fourier periodogram frequency peaks seems to provide a useful alter-

TABLE IV

Prediction of β-protein folds by means of the classification tree for the analysis of protein RRM frequency patterns (Figure 3)

Class	Predicted Class	Protein	Organism	JPred Code
β-fold	β-fold	Basic FGF	Homo sapiens	1bfg-1-DOMAK
β-fold	β-fold	PI3-kinase	Homo sapiens	1pht-1-AUTO.1
β-fold	β-fold	Coat protein	Bean mottle virus	1bmv1
β-fold	β-fold	Coat protein	Bean mottle virus	1bmv2
β-fold	β-fold	DNA binding protein	Coliphage T2	1gpc-1-AS
β-fold	β-fold	Capsid protein	Human rhinovirus	1r092
β-fold	β-fold	P53 suppressor	Homo sapiens	1tupc-1-AUTO
β-fold	β-fold	VCAM-1	Homo sapiens	1vcab-1-AUTO.1
β-fold	β-fold	VCAM-1	Homo sapiens	1vcab-2-AUTO.1
β-fold	β-fold	VMO I	Gallus gallus	1vmob-1-AS
β-fold	β-fold	Capsid protein	Avian myelobl. virus	2rspa
β-fold	β-fold	Coat protein	Satel. tobacco necr. virus	2stv
β-fold	β-fold	CD4	Homo sapiens	3cd4
β-fold	β-fold	Polyprotein	Human rhinovirus	4rhv1
β-fold	β-fold	Polyprotein	Human rhinovirus	4rhv3
β-fold	β-fold	Polyprotein	Human rhinovirus	4rhv4

native to the »common characteristic frequency« determination, since it enables accurate recognition of α- and β-protein folding types.

Acknowledgement. – The support by the Ministry of Science and Technology of the Republic of Croatia is highly appreciated (research grant No. 00981108).

REFERENCES

1. I. Ćosić, *The Resonant Model of Macromolecular Bioactivity*, BioMethods Vol. 8, Birkhäuser, Basel, 1997, pp. 1–139.
2. V. Krsmanović, J.-M. Biquard, M. Sikorska-Walker, I. Ćosić, C. Desgranges, M.-A. Trabaud, J. F. Whitfield, J. P. Durkin, A. Achour, and M. T. W. Hearn, *J. Peptide. Res.* **52** (1998) 410–420.
3. I. Ćosić, *IEEE Trans. Biomed. Eng.* **41** (1994) 1101–1114.
4. I. Ćosić and S. Birch, *Proc. IEEE EMBS* **16** (1994) 256–266.
5. I. Ćosić, A. E. Drummond, J. R. Underwood, and M. T. W. Hearn, *Mol. Cell. Biochem.* **130** (1993) 1–9.

6. I. Ćosić and M. T. W. Hearn, *Eur. J. Biochem.* **205** (1992) 613–619.
7. I. Ćosić, A. Hodder, M. Aguilar, and M. T. W. Hearn, *Eur. J. Biochem.* **198** (1991) 113–119.
8. I. Ćosić and D. Nesic, *Eur. J. Biochem.* **170** (1987) 247–252.
9. K. C. Chou and G. M. Maggiora, *Protein Engineering* **11** (1998) 523–538.
10. J. A. Cuff and G. J. Barton, *PROTEINS: Structure, Functions and Genetics* **34** (1999) 508–519.
11. J. A. Cuff, M. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton, *Bioinformatics* **14** (1998) 892–893.
12. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, *J. Mol. Biol.* **247** (1995) 536–540.
13. R. N. Bracewell, *The Fourier Transform and Its Applications,* McGrow-Hill, Singapore, 2000, pp. 258–292.
14. B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996, Chapter 7, pp. 230–235.

Appendix

α-protein folds:
1adeb-2-AUTO.1; 1aorb-3-AS; 1avhb-3-AS; 1avhb-4-AS; 1cc5; 1ceo-2-AUTO.1; 1clc-2-AS.1; 1csmb-1-auto.1; 1dsbb-2-AUTO.1; 1ecl-4-AS; 1fc2c; 1gal-2-AS; 1gdj; 1gln-3-AS; 1gln-4-AS; 1grj-1-AS; 1hcra-1-DOMAK; 1hiws-1-AS; 1hup-1-AS; 1hyp-1-DOMAK; 1lik-1-as; 1lik-2-AS; 1isab-1-GJB; 1lis-1-DOMAK; 1lki-1-AS; 1lmb3; 1lpe-1-DOMAK; 1mmoh-1-AS; 1pdnc-2-AS; 1poc-1-DOMAK; 1rec-1-DOMAK; 1rec-2-DOMAK; 1rhgc-1-DOMAK; 1rpo-1-AUTO.1; 1sra-1-AS; 1tndb-2-DOMAK; 256ba; 2abk-2-AS; 2asr-1-DOMAK; 2bltb-2-AUTO.1; 2cyp; 2end-1-DOMAK; 2mtac-1-AS; 2pgd-2-AUTO.1; 2utga; 2wrpr; 3icb; 3inkd-1-DOMAK; 3mddb-1-AS; 3mddb-3-AS; 4bp2; 4fisb-1-DOMAK; 6cpp; 1erc; 1aca; 1vas; 1lyn; 1hsm; 1rpr; 1pou; 1phb; 1tro; 1rhg; 2tct; 1boc; 1ctz; 1fip; 1hdd; 1dpr; 1tnt.

β-protein folds:
1amg-2-AS; 1aozb-1-AS; 1aozb-2-AS; 1aozb-3-AS; 1azu; 1bbpa; 1bcx-1-DOMAK; 1bfg-1-DOMAK; 1bmv1; 1bmv2; 1bncb-4-AS; 1bovb-1-DOMAK; 1cgu-2-GJB; 1cgu-3-GJB; 1cgu-4-GJB; 1clc-1-AS; 1cfb-1-AS; 1ctm-2-DOMAK; 1ctn-1-AS; 1eft-3-DOMAK; 1epbb-1-DOMAK; 1fnd; 1gog-1-AS.1; 1gog-2-AS.1; 1gog-3-AS.1; 1gp2g-2-AS; 1gpc-1-AS; 1hplb-2-AS; 1hxn-1-AS; 1krcb-1-AS; 1lib-1-DOMAK; 1mdta-3-AS; 1mjc-1-DOMAK; 1mspb-1-AS; 1paz; 1pht-1-AUTO.1; 1r092; 1smpi-1-AS; 1srja-1-DOMAK; 1tssb-2-DOMAK; 1tupc-1-AUTO; 1vcab-1-AUTO.1; 1vcab-2-AUTO.1; 1vmob-1-AS; 1vjs-3-GJB; 1wapv-1-AUTO.1; 2aaib-2-DOMAK; 2afnc-1-AUTO.1; 2alp; 2bat-1-GJB; 2cab; 2gn5; 2hft1-AS; 2hft-2-AS; 2ltna; 2ltnb; 2mev4; 2rspa; 2sil-1-AS; 2sns; 2sodb; 2stv; 3ait; 3cd4; 3mddb-2-AS; 3hmga; 4rhv1; 4rhv3; 4rhv4; 8adh.

SAŽETAK

Model rezonantnog prepoznavanja definira sekundarnu strukturu bioaktivnih proteina

Nikola Štambuk, Paško Konjevoda, Biserka Pokrić, Igor Barišić, Roko Martinić, Vladimir Mrljak i Pero Ramadan

Model rezonantnog prepoznavanja (RRM) bioaktivnosti proteina primijenjen je za predviđanje sekundarne proteinske strukture. Metoda se temelji na fizikalnom i matematičkom modelu elektronsko-ionskog interakcijskog pseudopotencijala (EIIP), te s pomoću analize signala interpretira linearnim nizom predočenu informaciju dobivenu iz odsječka makromolekule. Proteinska sekvencija prvo se transformira u niz pojedinačnih EIIP-vrijednosti aminokiselina. U drugom koraku model rabi spektralnu Fourier-ovu analizu signala dobivenih brojevnih nizova. Ćosić i sur.[1-8] su pokazali da signali pojedinačnih frekvencija spektra određuju karakteristične položaje aminokiselina, odnosno *aktivna mjesta*, povezana s biološkom ulogom proteina. U radu su analizirani proteini definirane sekundarne strukture, uspoređujući sheme 20 najizrazitijih signala pojedinačnih frekvencija Fourier-ova periodograma. Sheme 140 nehomolognih α- i β-proteinskih sekundarnih struktura iz baza podataka Jpred i SCOP, analizirane su s pomoću klasifikacijskog stabla kako bi se dobio algoritam za razlučivanje α- i β-strukture. Točnost ove brze i jednostavne metode za predviđanje sekundarne proteinske strukture jest 98.55%. Stabilnost algoritma potvrđena je pomoću »jack-knife« testiranja dobivenog stabla (srednja pogreška = 2.629, SE = 12.71). Opisana metoda za predviđanje sekundarne strukture proteina opširnije je prikazana na uzorku 30 različitih citokina, hormona, enzima i virusnih proteina. Naši rezultati pokazuju da se rezonantnom spektralnom analizom proteinske sekvence može izdvojiti informacija o sekundarnoj proteinskoj strukturi.