

Classification of Fourier spectra of short protein sequences compared to their corresponding structural classification.

TYAGI Manoj¹, RALAMBONDRAINY Henri², CADET Frédéric¹,
CHARTON Philippe², OFFMANN Bernard¹

1. Laboratoire de Biochimie et Génétique Moléculaire, UR
2. Institut de Recherche en Mathématiques et Informatique Appliquées, UR
(UR: Université de la Réunion, BP 7151, 97715 Saint Denis Messag Cedex 09, La Réunion)
bernard.offmann@univ-reunion.fr

Keywords : protein structure, classification, secondary structure, protein blocks, Fourier transform

Introduction

The linkage between sequence and structure in proteins is a major issue in bio-informatics. Since structure is determining in biological function, *ab-initio* prediction of folding patterns of sequences into secondary and tertiary structural motifs are being constantly attempted. Many methods and algorithms have been developed to address this issue. These include Neural Networks, Bayesian approaches and other grammatical or statistical methodologies. More recently, DeBrevin et al [1] has introduced the Protein Blocks (PBs) concept that defines a structural alphabet composed of 16 small protein fragments of 5 consecutive C α . These PBs allow an efficient approximation of the 3D backbone structure. In an attempt to develop a complementary approach for the prediction of protein structural motifs, we have investigated how classification of Fourier spectra of coded biological sequences is correlated to structural classification.

Method

The general framework of the analysis consists in three main steps. First, from a non redundant PDB subset of 1667 structures from the PDB, sequence segments were extracted and were either grouped according to their corresponding secondary structure (alpha-helix or beta-strand) or were grouped according to their corresponding Protein Blocks (PBs). Affection to one of the 16 PBs was performed by mapping extracted vectors of ϕ and φ angles of 5 consecutive C α onto a trained neural network as described previously [1]. Second, in our software named BISAR [2], these sequences were numerically coded using the EIIP index scale which has been used to develop the Resonant Recognition Model [3] and the resulting numerical signals were processed using Fast Fourier Transform (FFT) after aligning the lengths of all signals to 32 (for PBs) or 64 (for secondary structure) using the zero-padding technique. Fourier coefficients were used to calculate power spectra also called Fourier spectra. Third, classification of these spectral representations into alpha or beta or into any of the 16 PBs was attempted using either Linear Discriminant Analysis (LDA) or non parametric Decision Tree method (DNP) using the Kolmogorov-Smirnov distance [4,5].

Secondary Structure Analysis

Classification of EIIP-based Fourier spectra using LDA or Decision Tree methods into alpha or beta groups are given in table 1. Prediction rate of beta-strands were highest (76.4-77.2%) using LDA and sequences associated with helix structures were successfully classified at a rate of 56.9%-65.1%. When decision tree algorithm is used, alpha-helices were best predicted (77.6%) while only 57.7% of beta-strands were correctly classified. Our results show that overall prediction rates for beta-strands and alpha-helices average 77%. Hence, our method that implements Fourier transform of numerically coded sequences and that do not take into account any prior information on the sequence (like homology), can be used for the affection to a secondary structure class with a relatively good accuracy.

Table 1. Classification of Fourier spectra into secondary structures using LDA and DNP. The spectra in the calibration and validation sets were generated from respectively 686 and 294 randomly selected sequences from all secondary structures in PDB Select. Validation set results are given in between parenthesis

Original structure	Linear Discriminant Analysis		Decision Tree Method	
	Predicted structure		Predicted structure	
	Alpha-helix	Beta-sheet	Alpha-helix	Beta-sheet
Alpha-helix	230 (78)	123 (59)	274 (91)	79 (46)
Beta-sheet	76 (37)	257 (120)	141 (67)	192 (90)
Prediction rate	65.1% (56.9%)	77.2% (76.4%)	77.6% (66.4%)	57.7% (57.3%)
Overall prediction rate	71% (67.4%)		67.6% (61.8%)	

Protein Blocks Analysis

The classification of Fourier spectra of sequences of 5 amino acids extracted from PDB Select and for which each mapped onto one of the DeBrevern's 16 PBs is given in table 2. Overall, individual prediction rates (for each PB) were very low and ranged between 7% and 40%. However, two distinct groups of protein blocks could be distinguished in the tree. The first group included the PBs 00 to 08. PB 00 to 05 would correspond to sheets-related PBs *a* to *f* in DeBrevern's paper while PB06 to 08 to coil-related PBs *g* to *i*. The second group are PBs 09 to 15. PBs 10 to 15 would correspond to helix-related structures PBs *k* to *p* while PB 09 is associated to coil-related structure *j*. None of the sequences belonging to PBs 00 to 08 was classified in PBs 09 to 15 and reciprocally. This result provide evidence for the usefulness of Fourier spectra for that short EIIP-encoded sequences of 5 aa can be classified unambiguously (>99%) into two distinct groups of PBs (*a* to *i* and *j* to *p*). Further application of Bayesian probabilistic classification *a-posteriori* of DNP classification significantly improves the prediction rate in terms of PBs when compared to De Brevern's et al paper [1]. Overall simple Bayesian prediction rate was improved to 49.3% (vs 34.0%). When the first four predicted PBs are taken into account, the overall prediction rate reached 85% (vs 77%).

Table 2. Classification of Fourier spectra into Protein Blocks using Decision Tree method. The spectra in the calibration and validation sets were respectively generated from 3200 and 800 randomly selected sequences of 5 consecutive C α extracted from PDB files in the non redundant PDB Select database. A total of 250 sequences per PB was selected. Validation set results are given in between parenthesis

		original protein blocks																Total
		pb00	pb01	pb02	pb03	pb04	pb05	pb06	pb07	pb08	pb09	pb10	pb11	pb12	pb13	pb14	pb15	
predicted protein blocks	pb00	75 (10)	36 (16)	45 (15)	42 (14)	65 (16)	41 (12)	52 (11)	38 (16)	33 (7)	0 (0)	0 (0)	1 (0)	0 (1)	0 (1)	0 (1)	0 (0)	428 (120)
	pb01	14 (4)	26 (7)	14 (3)	12 (1)	12 (4)	16 (4)	12 (3)	12 (3)	11 (6)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	129 (35)
	pb02	15 (3)	29 (7)	38 (5)	30 (3)	29 (3)	19 (7)	24 (7)	20 (2)	24 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	228 (41)
	pb03	12 (7)	13 (5)	9 (2)	30 (8)	7 (3)	17 (4)	13 (5)	13 (4)	15 (4)	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	2 (0)	132 (43)
	pb04	14 (2)	11 (0)	16 (4)	11 (2)	24 (4)	15 (3)	8 (1)	9 (2)	6 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	114 (19)
	pb05	7 (1)	9 (1)	9 (0)	14 (0)	11 (2)	16 (5)	7 (0)	6 (5)	3 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	82 (18)
	pb06	7 (2)	14 (2)	8 (1)	8 (2)	2 (4)	14 (3)	15 (1)	8 (1)	14 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	90 (16)
	pb07	19 (3)	10 (5)	24 (9)	8 (8)	17 (4)	19 (7)	27 (7)	38 (5)	11 (8)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	173 (56)
	pb08	42 (13)	48 (11)	35 (13)	47 (10)	39 (4)	39 (9)	47 (10)	56 (12)	89 (10)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)	1 (0)	0 (1)	444 (93)
	pb09	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	23 (3)	17 (5)	11 (5)	11 (3)	5 (3)	8 (4)	11 (3)	86 (26)
	pb10	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	40 (8)	57 (11)	41 (10)	29 (9)	29 (12)	32 (15)	33 (10)	261 (75)
	pb11	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	46 (8)	23 (9)	61 (16)	32 (13)	34 (16)	37 (7)	41 (11)	274 (80)
	pb12	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	39 (6)	41 (11)	27 (6)	66 (4)	29 (6)	34 (10)	34 (9)	270 (52)
	pb13	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	41 (12)	39 (11)	28 (9)	37 (7)	77 (13)	39 (11)	38 (9)	299 (72)
	pb14	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	13 (4)	8 (4)	13 (4)	12 (5)	14 (4)	33 (6)	11 (3)	104 (30)
	pb15	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (5)	11 (1)	14 (4)	16 (4)	6 (1)	10 (2)	27 (7)	86 (24)
Total.	205 (45)	196 (54)	198 (52)	202 (48)	206 (44)	196 (54)	205 (45)	200 (50)	206 (44)	204 (46)	197 (53)	196 (54)	204 (46)	194 (56)	194 (56)	197 (53)	3200 (800)	
Prediction rate (%)	36.6 (22.2)	13.3 (13.0)	19.2 (9.6)	14.9 (16.7)	11.7 (9.0)	8.2 (9.3)	7.3 (2.2)	19.0 (10.0)	43.2 (22.7)	11.2 (6.5)	28.9 (20.8)	31.0 (29.6)	32.4 (8.7)	39.7 (23.2)	17.0 (10.7)	13.7 (13.2)		
	PBs <i>a</i> to <i>f</i> : 55.9% (53.8%)						PBs <i>g</i> to <i>j</i> : 48.9% (53.2%)				PBs <i>k</i> to <i>p</i> : 99.6% (99.7%)							

In conclusion, the classification of Fourier spectra of numerically coded short protein sequences are correlated to structural classification in terms of secondary structure and in terms of Protein Blocks. In particular, our results reinforces the pertinence of pre-processing using the protein blocks concept before sophisticated molecular modelling.

References

- [1] A.G. De Brevern, C. Etchebest and S. Hazout, Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41:271-287, 2000.
- [2] B. Offmann, J.-F. Fontaine and F. Cadet, BISAR, a java tool for the Fourier analysis of biological sequences. Journées Ouvertes Biologie Informatique Mathématiques (JOBIM), Saint-Malo, 10-12 juin 2002.
- [3] I. Cosic, The Resonant Recognition Model of Macromolecular Bioactivity, Theory and Applications. In *BioMethods Vol.8*, Ed Birkhäuser-Verlag, Berlin, 1997.
- [4] J.H Friedmann, A recursive partitioning decision rule for non parametric classification. *IEEE Trans. Comp.* 26 :404-408, 1977.
- [5] G. Celeux, E. Diday, Y. Lechevallier, H. Ralambondrainy, Classification automatique des données, environnement statistique et informatique. Ed Dunod, Paris, 1989.