MutScan, A Fourier Based Method For the Identification of Key Amino Acids in Protein Sequences

HUITELEC Emmanuelle[1], CHARTON Philippe[2], BEYLOT Bénédicte[1], LOCATE Salma[1], ROUCH Claude[1] and OFFMANN Bernard[1]

1. Laboratoire de Biochimie et Génétique Moléculaire, UR
2. Institut de Recherche en Mathématiques et Informatique Appliquée, UR

(UR : Université de la Réunion, BP 7151, 97715 Saint Denis Messag Cedex 09, La Réunion)
corresponding author : bernard.offmann@univ-reunion.fr

Keywords : mutagenesis, key amino acids, protein function, coded sequence, Fourier spectra

Introduction

The identification of amino acids that play a key role in the function of a protein is a difficult task to achieve and up to now, it has mainly relied on laboratory experimental evidences such as those from site-directed mutagenesis experiments. Because of the increasing number of protein sequences that are being referenced in public databases, computational methods are being developed in an attempt to predict these key amino acids and to associate parent sequences with function. These methods are mainly based on amino acid content analysis, molecular modelling, homology based algorithms or on the analysis of periodic features in sequences.

We developed a novel method based on the Fourier analysis of numerically coded protein sequences in order to predict, without any prior information, the amino acids that would play a key role in the function of any protein.

Methods

The methodology proceeds into three main steps that are all implemented in our software called BISAR [1]. In the first step, mutated sequences are generated : the original sequence is mutated sequentially from first amino acid position to the last amino acid so that *n* mutated sequences are generated, each differs from the original by only one mutation (figure 1). Each position is arbitrarily mutated into an Alanine except for Ala which is mutated into a Val.

```
Original Sequence :          PEVWTDFWKLFEEYLGKK (length = n)
n mutated sequences:         AEVWTDFWKLFEEYLGKK
                             PAVWTDFWKLFEEYLGKK
                             PEAWTDFWKLFEEYLGKK
                             .
                             .
                             PEVWTDFWKLFEEYLGAK
                             PEVWTDFWKLFEEYLGKA
```

Figure 1. *n* mutated sequences are generated from the original sequence of *n* amino acids.

In the second step, a Fast Fourier Transform (FFT) is performed on each of the sequences that were numerically coded using the EIIP index thus generating spectral representations of the sequences. This index was mainly used in the Resonant Recognition Model (RRM) [2] for that it is supposed to be correlated to function. In the third step, the Fourier spectrum of each mutated sequences is compared to the spectrum of the original sequence at the key frequencies of the functional family as defined in RRM. The difference of the amplitude values at these characteristic frequencies is plotted against mutated position. Those mutated positions that decreases most the amplitude values at the key frequencies are hypothesized to be the key amino acids.

Results

We have predicted key amino acids in glucagon and cytochrome C. The characteristic Fourier frequencies of the functional families to which these sequences belong are f=0.0859 and f=0.473 respectively. We have hence investigated how mutations all along these sequences alter the amplitude values at these frequencies.

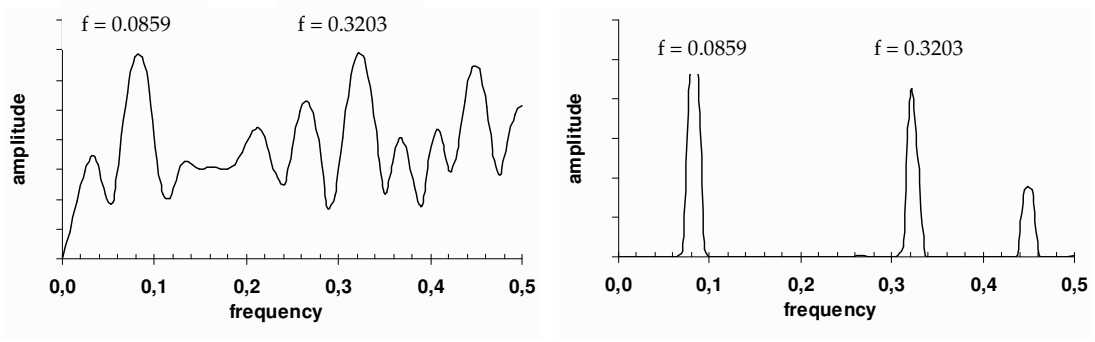

Figure 2a & 2b. Consensus Fourier spectra of 18 glucagon protein sequences. 2a : Mean spectrum and 2b : cross-multiplied spectrum [2].

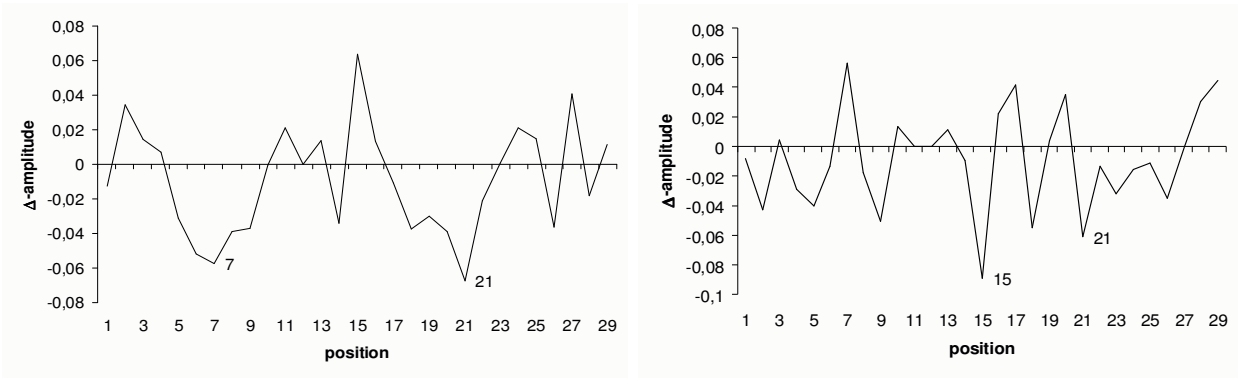

Figure 3a & 3b. Effect of mutation on amplitude values of human glucagon (gluc_human) spectrum. Shown are variations of amplitude values at characteristic frequencies as a function of mutated position. 3a : variation at f = 0.0859 and 3b: variation at f = 0.3203.

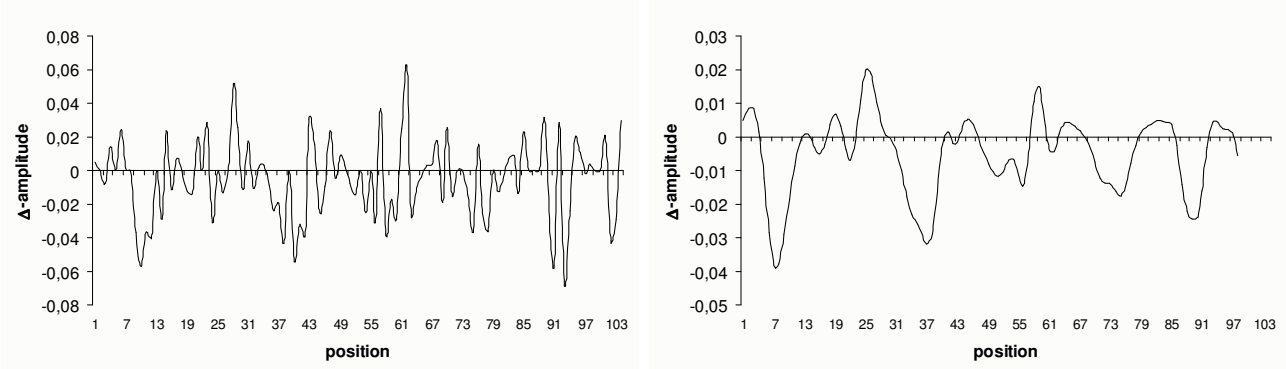

Figure 4a & 4b. Effect of mutation on amplitude values of human cytochrome C spectrum at f=0.473. In 4a, variations as a function of mutated positions. In 4b, curve in 4a was smoothed using a sliding window of 6.

In our model, we postulate that those mutations that decreases most the amplitude value at the RRM [2] characteristic frequency of the functional family are those that contribute more to the function of the protein. Hence, key amino acids can be identified along any sequence. In glucagon, key predicted positions are 7, 15 and 21. When long sequences are processed, it can be convenient to smooth the plot using a sliding window so that key regions are identified rather that particular amino acids. In cytochrome C (figures 4a & 4b), out of 5 predicted regions, a total of 4 regions do possess key amino acids sequence as previously described [3].

References

[1] 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.

[2] I. Cosic, The Resonant Recognition Model of Macromolecular Bioactivity, Theory and Applications. In BioMethods Vol.8, Ed Birkhäuser-Verlag, Berlin, 1997.

[3] D.M. Hampsey, G. Das, F. Sherman, Amino Acid Replacement in Yeast Iso-1-cytochrome *c*. Comparison with the phylogenetic series and the tertiary structure of related cytochromes *c*. J. Biol. Chem. 261(7):3259-3271, 1986.