Harmonic Analysis of Protein Sequences

Meeta Rani Chanchal K. Mitra
ckmsl@uohyd.ernet.in

Department of Biochemistry, School of Life Sciences
University of Hyderabad, Hyderabad 500 046, India

1 Introduction

Proteins are informational molecules and their primary sequence holds the key to their tertiary structure and hence to their biological activity. The tertiary structures are the folded and compact. Protein folding can bring into vicinity some of amino acid residues which may be physically distant in the primary sequence. Such correlations are usually important for retaining the folded form of the tertiary structure and are generally preserved in evolution. We are interested in eliciting such positional correlations (in terms of the participating amino acid pairs) by harmonic analysis of the protein sequences. With the availability of large protein sequences databanks and good computing facilities this has become less difficult and more interesting.

Here, we report the technique and the results of the harmonic analysis of the primary sequences of proteins from the SWISS-PROT protein sequence databank carried out by us, with the main objective of eliciting regularities, periodicities and other general patterns in the protein primary sequences. For carrying out harmonic analysis, we have treated the primary sequences as time-series. Time-series are useful in studying the relationship of values from one term to next, in serial correlation along the series and for constructing a simple system of a mathematical kind which can be used to describe the behaviour of the series in a concise manner. It can provide insights into the underlying causation and allow making projections into future more accurate. The ideas (of the time-series) have been extended to a spatial situation by us in case of protein sequences. Protein sequences may be considered as time-series due to the fact that during protein synthesis, the successive temporal addition of an amino acid residue to the elongating polypeptide chain is reflected in its primary sequence. Hence in case of proteins, on the x-axis, we plot the positions (of the amino acid). The positional distribution of residues in proteins as time-series are then subjected to harmonic analysis.

Since the positional distribution (up to 200 positions only, for sake of computational reasons) of each amino acid residue is treated as a single time-series, we have twenty series of positional distributions one corresponding to each amino acid. To detect positional correlation between the amino acid within the series, we have carried out the correlation analysis of the 20 series. Since proteins consist of 20 types of amino acids, we also need to understand the position correlation between same as well as different kinds of residues at different positions in the protein primary sequences. Hence we need to calculate both autocorrelation as well as cross correlation functions. While, the auto- correlations give intra-dependencies within a time-series and the cross-correlations give the inter-dependencies between the various time-series at various time-lags (here, position lags).

2 The methodology of correlation analysis

The positional distribution of each amino acid upto 200 positions in the protein sequences is calculated separately and it constitutes an individual time-series. Hence we obtain 20 such time-series, one for each amino acid distribution. Each of the 20 series are treated as follows:

1. The mean and variance of each of the time-series (positional distribution of a residue) is calculated.
2. The autocovariances and corresponding autocorrelations up to order 99 are calculated.
3. The cross-covariances and cross-correlations are calculated mostly in the same way. except for some difference arising due to the multiple time-series in this case.
4. The fourier transformation of the autocorrelations yield the spectrum for each series. The cross-correlations, on fourier analysis yield spectra having real as well as imaginary term and they are dealt with differently. However, the analysis of the spectra obtained in this case is just the same as for those obtained from autocorrelations.
5. The spectrum for each pair (there are 400 pairs) is analyzed and positional correlation between these amino acids are elicited. All the strong amino acid pair correlations and corresponding inter-residue distances are recorded.
6. The strong pair correlations and inter-residue distances are recorded and these values are cross-checked by directly scanning the protein sequences in the data base to verify that they are real and not artifacts.

Based on the results, we have constructed a knowledge base of the preferred inter-residue distances between members of various pair (there are 400 possible pairs with the 20 amino acids) within the protein sequences. On cross-checking these periodicities in the real data base (and also that of Monte Carlo simulated random data base of sequences), we find that our results are very meaningful and we also attempt to provide explanations for these position correlations based on higher level structures seen in proteins. The results indicate the existence of order and regularities in the protein sequences. For example, our results indicate that Cys-Cys are highly preferred pairs. This is explained by the requirement for disulfide bond formations. The results also indicate a high occurrence of Gly-X-Pro which can be explained by their excessive occurrence in some kinds of structural proteins. The results can be useful to guide protein engineers for manufacturing synthetic peptides for various purposes. Though to the best of our knowledge we are the first to carry out harmonic analysis of the protein sequences in the manner we have done [1, 2], we believe that our methods can be further refined.

We suggest that similar analyses can be carried on the genomic sequences also.

References

[1] Meeta Rani and Chanchal K Mitra, "Periodicities in Protein Sequences, " *Journal of Biosciences*, Vol. 19, pp 255-266, 1994.

[2] Meeta Rani and Chanchal K Mitra, "Correlation Analysis of the distribution of amino acids in protein sequences," *Journal of Biosciences*, Vol. 20, pp 7-16, 1995.