Journal of Theoretical Biology 304 (2012) 211–218

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Looking for a sequence based allostery definition: A statistical journey at different resolution scales

Saritha Namboodiri [a], Alessandro Giuliani [b,*], Achuthsankar S. Nair [a], Pawan K Dhar [c]

[a] State Inter University Centre of Excellence in Bioinformatics, University of Kerala, Kariyavattom Campus, Thiruvananthapuram, Kerala, India
[b] Environment and Health Dept. Istituto Superiore di Sanità, Roma, Italy
[c] Centre for Systems and Synthetic Biology, University of Kerala, Kariyavattom Campus, Thiruvananthapuram, Kerala, India

ARTICLE INFO

Article history:
Received 6 December 2011
Received in revised form
29 February 2012
Accepted 3 March 2012
Available online 30 March 2012

Keywords:
Recurrence Quantification Analysis
Hydrophobicity
Free-energy-transfer
Allosteric drugs
Co-evolution

ABSTRACT

The aim of this work was to detect allosteric hotspots signatures characterizing protein regions acting as the 'key drivers' of global allosteric conformational change. We computationally estimated the relative strength of intra-molecular interaction in allosteric proteins between two putative allostery-susceptible sites using a co-evolution model based upon the optimization of the cross-correlation in terms of free-energy-transfer hydrophobicity scale (Tanford scale) distribution along the chain. Cross-Recurrence Quantification Analysis (Cross-RQA) applied on the sequences of allostery susceptible sites showed evidence of strong interaction amongst allosteric susceptible sites. This could be due to transient weak molecular bonds between allostery susceptible patches enabling regions far-apart to come together. Further, using a large protein dataset, by comparing allosteric protein set with a randomly generated sequence population as well as a generic protein set, we reconfirmed our earlier findings that hydrophobicity patterning (as formalized by Recurrence Quantification Analysis (RQA) descriptors) may serve as determinant of allostery and its relevance in the transmission of allosteric conformational change. We applied RQA to free-energy-transfer hydrophobicity-transformed amino acid sequences of the allostery dataset to extract allostery specific global sequence features. These free-energy-transfer hydrophobicity-based RQA markers proved to be representative of allosteric signatures and not related to the differences between randomly generated and real proteins. These free-energy-transfer hydrophobicity-based RQA markers when evaluated by pattern recognition tools could distinguish allosteric proteins with 92% accuracy.

1. Introduction

Inter- and intra-molecular interactions are essential elements of cellular function. A particularly well studied kind of intermolecular interaction is the reversible specific binding of a protein with a ligand which may be a protein or a small molecule, at specific sites on its structure called the active sites. Binding of ligand (allosteric effector) at particular sites (allosteric sites) which are distant from active sites are found to either activate or inhibit the binding of ligand at main active sites bringing about a structural change in the protein resulting into a functional modification. This is called allosteric regulation or 'allostery' (Monod et al., 1963) to remark the nature of 'action at distance' of the phenomenon.

Allostery, in terms of configuration change, can also be induced by non-ligand sources such as point mutation, multiple mutations, chemical changes, changes in the pH or molecular crowding. Thus allostery in a sense maybe intended as a property common to all proteins (Gunasekaran et al., 2004).

Allostery involves transmission of signals among sites that are far off in the sequence and structural space of a protein. The effect of ligand binding, mutation, covalent modification over physical distances is well documented (Gerstein and Krebs, 1998; Wang and Kemp, 2001; Lim, 2002; Falcon and Matthews, 2001).

However, the key unanswered questions are: how do two distant protein sites communicate? What are the physical, structural, thermodynamical principles responsible for allostery? Are there some hotspots more involved in protein allostery than others?

To understand the structural basis of allostery, various experimental and computational approaches have been developed in the past (Bradley, 2009; Voorhees et al., 2010; Amaro et al., 2007; Masterson et al., 2008). Ota and coworkers (2005) used kinetic energy to understand the physical mechanism of intra molecular signal transduction process (Ota and Agard, 2005). Subsequently, another promising computational approach to derive the potential intra- and

* Corresponding author.
E-mail address: alessandro.giuliani@iss.it (A. Giuliani).

0022-5193/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2012.03.005

intermolecular pathways of signal transduction based on Markov process was developed (Chennubhotla and Bahar, 2006). Del Sol et al. identified residues crucial in allosteric transmission using graph theory with residues as nodes and between residues van der Waals interactions as edges (Del Sol et al., 2006). Using elastic network model, residues involved in allosteric pathways in proteins like myosin, helicases and DNA and RNA polymerase were predicted (Zheng and Brooks, 2005; Zheng et al., 2007). Further, COREX, a structure-based calculation of the equilibrium folding pathway of proteins was used to find the network of 'allosteric' residues (Hilser and Freire, 1996; Whitten et al., 2005; Liu et al., 2006).

Daily and coworkers studied tensed (T) and relaxed (R) states of 51 allosteric proteins structures in order to characterize local structural and functional differences between them (Daily and Gray, 2007). They could identify residues involved in allosteric transmissions by adopting a set of flexibility descriptors. These descriptors allowed the authors (despite the continuous character of allostery due to the natural consequence of protein dynamics in solution) to set a threshold discriminating allosteric and non-allosteric proteins as well as allosteric and non-allosteric sites within a given protein. Further, mutual dependence of amino acid residues reflecting functional coupling and intra-molecular interactions was observed in allosteric proteins (Lockless and Ranganathan, 1999; Dima and Thirumalai, 2006; Namboodiri et al., 2010).

Nussinov and coworkers are of the opinion that mechanistic understandings of molecular structure, conformational dynamics and their associations in the cellular network have to be integrated in order to provide a better insight into the allosteric behavior. They posit that allosteric signal propagation does not stop at the end of a protein but may be dynamically transmitted across the cell thus providing an initial stimulus of going from drug–receptor interaction to biological effect (Nussinov et al., 2011).

Allosteric 'hotspots' residues of allosteric proteins are residues that are more directly involved in signal transmission. Mutations of these residues perturb allostery and are called allosteric hotspot residues (Demerdash et al., 2009). Such residues are found to form clusters and are called allosteric hotspot regions or allosteric susceptible regions.

In this work we tried and set some 'general rules' of allostery on a pure sequence basis, in doing so we started considering specific tracts of amino acid sequence of 'ras' protein, more involved in allostery i.e., the allosteric hotspot regions, as described in Amaro et al. (2007), Daily and Gray (2007). In 'ras' protein, these residues are found to cluster together in two regions forming allosteric susceptible regions (ap1 and ap2). A non-allosteric region containing residues not-so involved in allostery in 'ras' protein was also considered (nap) for comparison purposes. Our approach builds upon the principle described in Del Sol et al. (2006) stating that residues crucial for allostery are arranged into paths allowing network communication mediating signaling. The establishment of such 'communication lines' implies some sort of interaction between intervening residues. Consequently, the estimated 'interaction strength' between ap1 and ap2 should be significantly greater than the estimated interaction between ap and nap sites.

In order to evaluate the interaction between allosterically susceptible regions (ap1–ap2) and each of their interaction with a non-allosterically susceptible region (nap), we applied Cross-RQA to ap1–ap2, ap1–nap and ap2–nap couples to highlight the possible presence of interacting motifs. Cross-RQA is a sensitive indicator of the probability of peptide interaction and has been successfully experimented in identifying interacting domains in Hepatitis C Virus E1 proteins (Bruni et al., 2009) as well as in other protein systems (Giuliani and Tomasi, 2002; Giuliani et al., 2003). The intra-molecular interaction study based on Cross-RQA on the E1 protein of Hepatitis C Virus revealed prominent cross-correlation indicative of highly interacting pairs (Bruni et al., 2009). The authors also demonstrated that interacting peptides along the protein were much more cross-correlated than what observed in a population of sequence-shuffled simulated peptides with the similar composition. This departure from randomness was not detected in the non-interacting pairs. This 'cross-correlation' optimization in the light of the co-evolution has been demonstrated for interacting proteins (Goh et al., 2000). The above studies demonstrated an increased cross-correlation of intra- and inter-molecular interactions.

One could hold the view that applying the above co-evolution theory on allostery may result in more marked cross-correlation, in terms of free-transfer-energy hydrophobicity scale, amongst the 'ap' pairs indicating the need for coordinated atomic rearrangements in these flexible regions (Giuliani and Tomasi, 2002; Goh et al., 2000; Zbilut et al., 1998).

Further, we used a large dataset of 107 single chain allosteric and random protein sets to understand the mechanistic basis of the phenomenon. We generated randomly shuffled sequences coming from allosteric native structures as non-allosteric internal probe.

By applying Recurrence Quantification Analysis (Eckmann et al., 1987) on free-transfer-energy hydrophobicity-coded sequences, we extracted hydrophobicity patterning features which, when analyzed by pattern recognition tool, discriminated the allosteric from the non-allosteric set with 92% accuracy (83% sensitivity, 100% specificity). We could prove that these free-transfer-energy hydrophobicity-based RQA markers are representative of allosteric signatures and that they do not arise due to the difference between real and random protein. This was proved by considering two independent statistical experiments contrasting a specifically curated allosteric protein data set with another protein data set not specifically selected on allostery considerations. We then demonstrated the possibility to discriminate this general protein data set from its shuffled counterpart thus indirectly proving the departure of pure randomness of protein sequences.

2. Materials and methods

Recurrence Quantification Analysis (RQA) and Cross-RQA were the two main data analysis techniques we used in this study.

2.1. Recurrence Quantification Analysis (RQA)

RQA is a computational approach whose versatility in non-linear system modeling is widely appreciated (Charles, 2009). The technique is well suited for short and non-linear signals and has been extensively used in proteomics (Porrello et al., 2004; Colafranceschi et al., 2005; Zbilut et al., 1998; Zbilut et al., 2004; Giuliani et al., 2002). RQA is based on the basic mathematical concept of recurrence, i.e., the repetition of a specific event at different temporal (or spatial) delay.

Proteins are ordered biopolymers and in this domain we equate the temporal dimension to the order of amino acid residues along the chain. We convert amino acid residues sequence into numeric 'time series' by replacing time with the position of amino acid and attaching to each amino acid a suitable chemico-physical property, in this case, we made use of free-energy-transfer. Free-energy-transfer, measured by Tanford scale, can be intended as an hydrophobicity score, and it was demonstrated to be particularity efficient in detecting interacting sites (http://www.drgutman.org/ORIGINAL_PAPERS/%2314.pdf; http://onlinelibrary.wiley.com/doi/10.1002/bip.20607/full).

We then project this series into higher dimension space to construct an embedding matrix by using time delay approach (Packard et al., 1980; Takens, 1981). Each column of the embedding matrix is the time (sequence order) delayed copy of the original series. The distance in terms of the free-energy-transfer between two rows (epochs of the series which, in our case, are amino acid residues patches) of the embedding matrix is computed and put into the distance matrix. When the distance between two epochs is less than a predefined threshold radius, the two rows (states in dynamical terms) are near to each other and are called neighbors.

This is an instance of recurrence and is represented by a value of 1 at the intersection of the two rows between epochs in the recurrence matrix. This is visualized using recurrence plot which graphical visualizes the distance between the epochs of the corresponding embedding matrix with darkened pixel for 1's, indicating the presence of recurrence, and white pixel for zero, indicating the absence of recurrence. This gives rise to patterns specific to the dynamics of the system under study. Fig. 1 report the free-energy-transfer hydrophobicity based recurrence plot of 'ras' protein.

Further, we used the Cross-RQA to study the correlation among two distinct amino acid sequences representing two different proteins. Basically, the Cross-RQA mode is used to study correlations taken at all intervals whereby the abscissa and ordinate of the recurrence plot represent two different time series. In this case, the computation of the RQA descriptors remains the same, but they represent the correlation of two distinct series of numbers (cross correlation) rather than the internal correlations within the same series (auto correlations).

The dynamical features of a system represented as recurrence plot are quantified into statistical variables (Marwan et al., 2007) as follows: **Recurrence**: represents the ratio of the number of recurrent points to the total number of points. **Determinism**: is the ratio of the number of recurrent points occurring in diagonals (and thus consecutive along the chain) to the total number of recurrent points. **Lmax**: the maximal length of subsequent recurrent points along a diagonal different from the principal one. **Laminarity**: is the ratio between recurrent points occurring in vertical/horizontal lines to the total number of recurrent points. **Information entropy**: is a measure of the Shannon information entropy frequency distribution of the lengths of the diagonal lines of recurrent points. **Trap Time, T1** and **T2**: measure the average length of vertical line indicating the mean time that the system remains in a particular state, the average distance between a point and its recurrence and average distance between a point and its recurrence excluding distance of one unit, respectively.

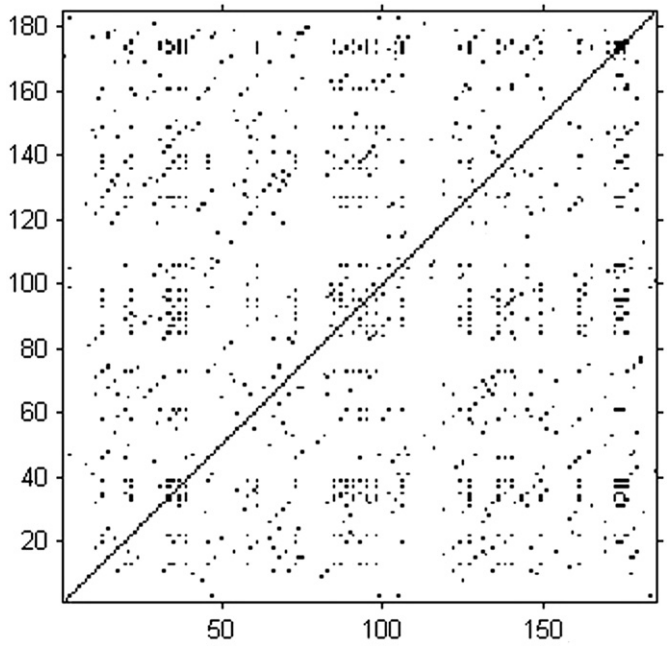

Fig. 1. Recurrence plot of free-energy-transfer hydrophobicity encoded amino acid sequences of 'ras' protein.

2.2. *Specific signatures of allosterically susceptible tracts*

The first part of our work deals with the characterization of 'specific allosteric signatures' that could be considered as the possible 'drivers' of the global conformational change. Experimental and computational approaches have led us to identify residues more deeply involved into allosteric regulations (Amaro et al., 2007; Daily and Gray, 2007). These residues have greater mobility and are identified as those extremely flexible regions driving the global configuration change (Daily and Gray, 2007).

We considered 'ras' (PDB id 4Q21) as an example protein for our study (Daily and Gray, 2007) and applied a sort of 'evolutive paradigm' based on the fact that allostery is a crucial property of protein systems and is plausible 'optimized' to certain extent. We took the residues 24–40 and 60–76 of 'ras' protein as allosteric susceptible regions, ap1 and ap2, respectively and residues 154–170 as non-allosteric patch, nap (Daily and Gray, 2007; Brunger et al. (1990)). The sequence of 'ras' protein with ap1 and ap2 highlighted in bold while nap is in italics is as follows:

MTEYKLVVVGAGGVGKSALTIQL**IQNHFVDEYDPTIEDSY**RKQVVI
DGETCLLDILDT**AGQEEYSAMRDQYMRTG**EGFLCVFAINNTKSFEDIH
QYREQIKRVKDSDDVPMVLVGNK;CDLAARTVESRQAQDLARSYG
IPYIETSAKTRQG*VEDAFYTLVREIRQHKL*RKLNPPDESGPGCMSCKCVLS

Fig. 2 shows the PDB structure of 'ras' protein (in relaxed as well as tensed state) with the two allosteric patches ap1 and ap2 highlighted in red and blue.

These regions are far apart in sequence as well as in structure. Yet they must cooperate to accomplish allosteric regulatory functions. In order to computationally prove their interaction and also find the specific allosteric signatures of these regions, we made use of Cross-RQA approach that has been tried and tested in various studies (Bruni et al., 2009; Giuliani and Tomasi, 2002; Goh et al., 2000; Zbilut et al., 1998; Marwan et al., 2007). The sequences of the three regions were initially transformed into numeric series with their relative free-energy-transfer hydrophobicity scale. The free-energy-transfer hydrophobicity values used for each of the twenty amino acid are as given: Ala: 0.620 Arg: −2.530 Asn: −0.780 Asp: −0.090 Cys: 0.290 Gln: −0.850 Glu: −0.740 Gly: 0.480 His: −0.400 Ile: 1.380 Leu: 1.530 Lys: −1.500 Met: 0.640 Phe: 1.190 Pro: 0.120 Ser: −0.180 Thr: −0.050 Trp: 0.810 Tyr: 0.260 Val: 1.800. We used Cross-RQA which generates a cross-recurrence plot from the two different sequences to investigate for mutual correlation between interacting regions in the peptide under study. An increase in cross-correlation linked to the physical interaction between two proteins has already been demonstrated both by Cross-RQA (Giuliani and Tomasi, 2002) and other computational methods (Selz et al., 2006). Fig. 1 reports the free-energy-transfer hydrophobicity based recurrence plot of 'ras' protein, where the diagonal and vertical/horizontal lines point to deterministic and laminar regions, respectively.

In the case of effectively interacting proteins, it was demonstrated that native sequences are by far, more cross-correlated than their shuffled counterparts (Lockless and Ranganathan, 1999) thereby verifying the observation of co-evolution of interacting protein systems (Giuliani and Tomasi, 2002; Goh et al., 2000). Here, we make the hypothesis that the co-evolution theory applicable to inter-

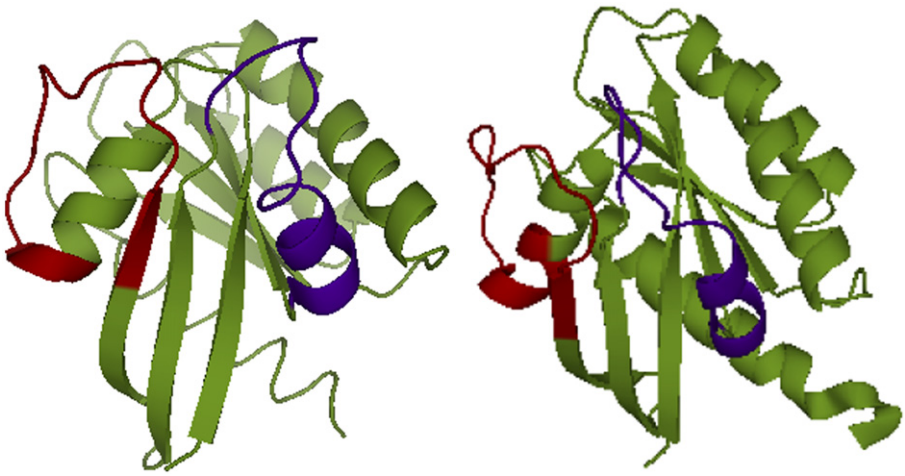

Fig. 2. Structure of 'ras' protein, PDB id 4Q21 (Relaxed state) and PDB id 6Q21 (Tensed state) with allosteric susceptible regions ap1 and ap2 highlighted as red and blue, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

molecular interactions could also be applied to intra-molecular interaction too. In order to verify this hypothesis, we shuffled the sequences of each of the three regions and generate 30 different random sequences for each of the three regions. Determinism and Laminarity parameters (see Section 2.1 for the RQA descriptors) are the order dependent descriptors considered in the analysis. We computed the Cross-RQA and obtained the Cross-determinant and Cross-laminarity of the shuffled sequences of the three possible couples made of different patches. (These are attached as supplementary file.) Mean, standard error and confidence interval of the 30 shuffled copies set were computed so to check the statistical significance of the comparison of the native sequences Cross-RQA with shuffled counterparts.

2.3. Finding global signatures of allostery

The second part of our work was to validate the global signatures of allosteric proteins as obtained from our previous study on a larger dataset and to go in depth into the tenability of the hypothesis of the presence of specific allostery signatures on a population basis. We relied on the dataset derived from AlloSteric Database (ASD), a central resource for allosteric molecules that currently contains 336 allosteric proteins from 101 species (Huang et al., 2011). We selected 107 single chain allosteric proteins for our study. This choice was dictated by the fact that RQA is only able to give an unbiased representation for single chain proteins. We also generated a 107 random sequences set using the RandSeq, a tool to generate random protein sequence by Expasy, a molecular server of Swiss Institute of Bioinformatics. We randomly divided the allosteric and random dataset into a training dataset containing 71 allosteric and random series and a test dataset having 36 allosteric and random proteins. We transformed each sequence into numeric series by substituting each amino acid by its free-transfer-energy hydrophobicity value. We choose hydrophobicity (in terms of Tanford scale of free-energy transfer) as it plays a significant role in folding and dynamics of a protein. Each of these sequences was subjected to Recurrence Quantification Analysis. In our study, each of the free-transfer-energy hydrophobicity transformed sequence was fired one by one onto RQA. After the RQA application each protein was represented as a 10 dimension feature vector comprising of 8 dynamic RQA variables and 2 static (order independent) variables, namely mean and standard deviation of the hydrophobicity values of amino acids in the given sequence.

2.4. Pattern recognition: Do allosteric proteins have a typical hydrophobicity patterning?

Computational experiments were carried out on the so obtained RQA feature vectors using various classifiers available in WEKA (Hall et al., 2009). We initially trained the classifiers using the training dataset and then performed the classification on the test dataset.

After obtaining a very significant discrimination between allosteric and random data sets, we analyzed whether the discrimination of the allosteric and randomly generated protein set was due to the presence of a 'specific allosteric signature' or was it due to the fact that we were comparing real proteins and randomly generated ones i.e., are we extracting 'real protein signature' or 'allosteric sequence signature'?

From an epistemological point of view, this is a very tricky problem because although the allosteric dataset was made up of protein whose kinetics in vitro had the classical allosteric behavior as detected by inhibition experiments, we know that all proteins are, to a certain degree, allosteric.

An indirect appreciation that allostery could probably play a major role in classification could be obtained by performing two related pattern recognition experiments:

- Discriminating a set of random sequences and generic single chain protein set with no imposed bias toward allosteric systems;
- Discriminating an allosteric set and a generic single chain real protein set.

In order to prove that, a possible increase in between sequences similarity inside the real protein data set with respect to the random set, has no role in the classification, we computed the average of pairwise distance between the sequences for the random set as well as for generic single chain real protein set. This was done with the help of MEGA (Molecular Evolutionary Genetics Analysis) software (Tamura et al., 2007) which gave rise to an average pairwise distance of 2.511 and 2.114 for real and random data sets, respectively thus eliminating any possible bias due to the possible increased sequence similarity between real proteins with respect to random sequences. The average between allosteric sequences distance was found to be 2.875, lying in the same order of magnitude of the internal diversity as the two other data sets for real and random proteins This implies that any

significant discrimination obtained cannot be traced back to the presence of an 'homology bias' between proteins.

3. Results

Table 1 reports mean, standard error and confidence interval of a set of 30 shuffled peptide couples from ap1–ap2 and ap1–nap and ap2–nap pairs of the Cross-laminarity and Cross-determinism with the corresponding native pair value.

It is worth noting the departure of ap1–ap2 native Cross-RQA from shuffled counterparts reaching a very high statistical significance, especially for laminarity which we have already detected as one of the main determinants of allostery in the previous work (Namboodiri et al., 2010) thus pointing to an evolutionary 'optimization' of the allosteric pairing. Analyzing the other two comparisons we observed (Table 1) that ap1–nap showed a statistically significant higher Cross-RQA with respect to shuffled set, while ap2–nap had no marked departure from randomness.

While a sort of 'baseline correlation' higher than the shuffled counterpart was expected due to the fact that all the considered patches pertain to the same protein system (and thus they are expected to show a co-evolution signal due to general structural constraints), ap1–ap2 cross-correlation seems to be definitively higher than the ap1–nap and ap2–nap cross-correlation thereby indicating a 'special relation' between the two patches Fig. 3.

We plot the Cross-determinism against Cross-laminarity of ap1–ap2, ap1–nap and ap2–nap of shuffled sequences along with the native values. Fig. 4(a)–(c) depict the cross-correlations with shuffled sequences as black dots and the native sequence as a white dot. The departure from randomness of the ap1–ap2 pair with respect to the other couples is markedly evident by the fact the native pairing (white dot) is at the very extreme of the distribution (Fig. 4(a)).

The result of the global population analysis of allosteric signatures is reported in Table 2 where the performances of the different classifiers for the three experiments are tabulated.

All the three experiments scored a very high classification accuracy thereby giving a global demonstration of the existence of an 'allosteric signature' different from the 'general protein' signature in terms of hydrophobicity patterning along the chain.

It is immediate to observe that both the experiments involving a comparison between real proteins and randomly generated sequences gave rise to classification accuracy higher than the comparison between allosteric and generic real proteins thereby giving a proof-of-concept to two very relevant statements:

- Proteins cannot be considered equivalent to random sequences of residues.
- A specific sequence signature of allostery is apparent.

Table 1
Mean, standard error and confidence interval of a set of 30 shuffled peptide couples from ap1 – ap2 and ap1–nap and ap2–nap pairs of the Cross-laminarity and Cross-determinism with the corresponding native pair value.

	Mean	Standard error	Confidence interval	Native sequence
ap1–ap2				
Cross-laminarity	46.52	3.03	6.17	**85.0**
Cross-determinism	39.72	1.81	3.68	**67.5**
ap1–nap				
Cross-laminarity	43.02	2.72	5.54	**58.33**
Cross-determinism	38.03	1.63	3.32	**69.70**
ap2–nap				
Cross-laminarity	46.53	2.91	5.73	**48.35**
Cross-determinism	47.61	1.76	3.58	**51.61**

4. Discussion

Our search for specific allosteric signatures, in the first part of our study, led us to look into specific regions in allosteric proteins which are more involved in allosteric motions and interactions. Cross-RQA of native ap1–ap2, ap1–nap and ap2–nap (Table 1) revealed higher cross-correlation in the free-energy-transfer for ap–ap2 pair than ap1–nap and ap2–nap regions. This is indicative of greater interaction amongst the allosteric susceptible regions in terms of free-energy-transfer. The hydrophobic effect can be quantified by measuring the partition coefficients of non-polar molecules between water and non-polar solvents. The partition coefficients can be transformed to free-energy-transfer which includes enthalpy and entropy components, $\Delta G = \Delta H - T\Delta S$. These results provide evidence that allosteric interactions are governed

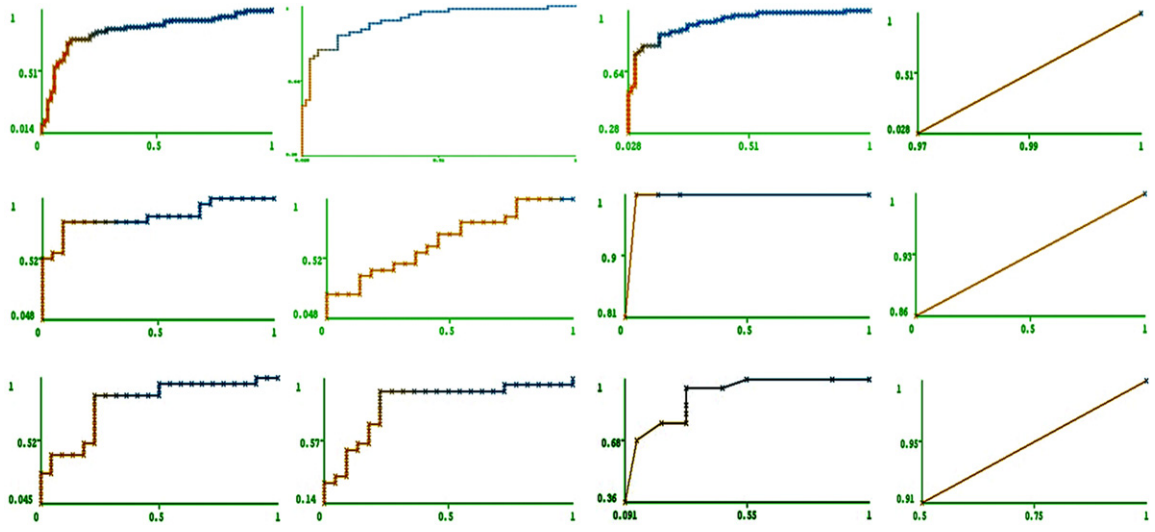

Fig. 3. Depicting the ROC curve of the classifiers Mulitlayer Perceptron, Naïve Bayes, Random forest and LibSVM, in order, on the three datasets namely allosteric proteins vs. randomly generated sequences (row 1), generic proteins vs. randomly generated sequences (row 2) and allosteric proteins vs. generic proteins (row 3).

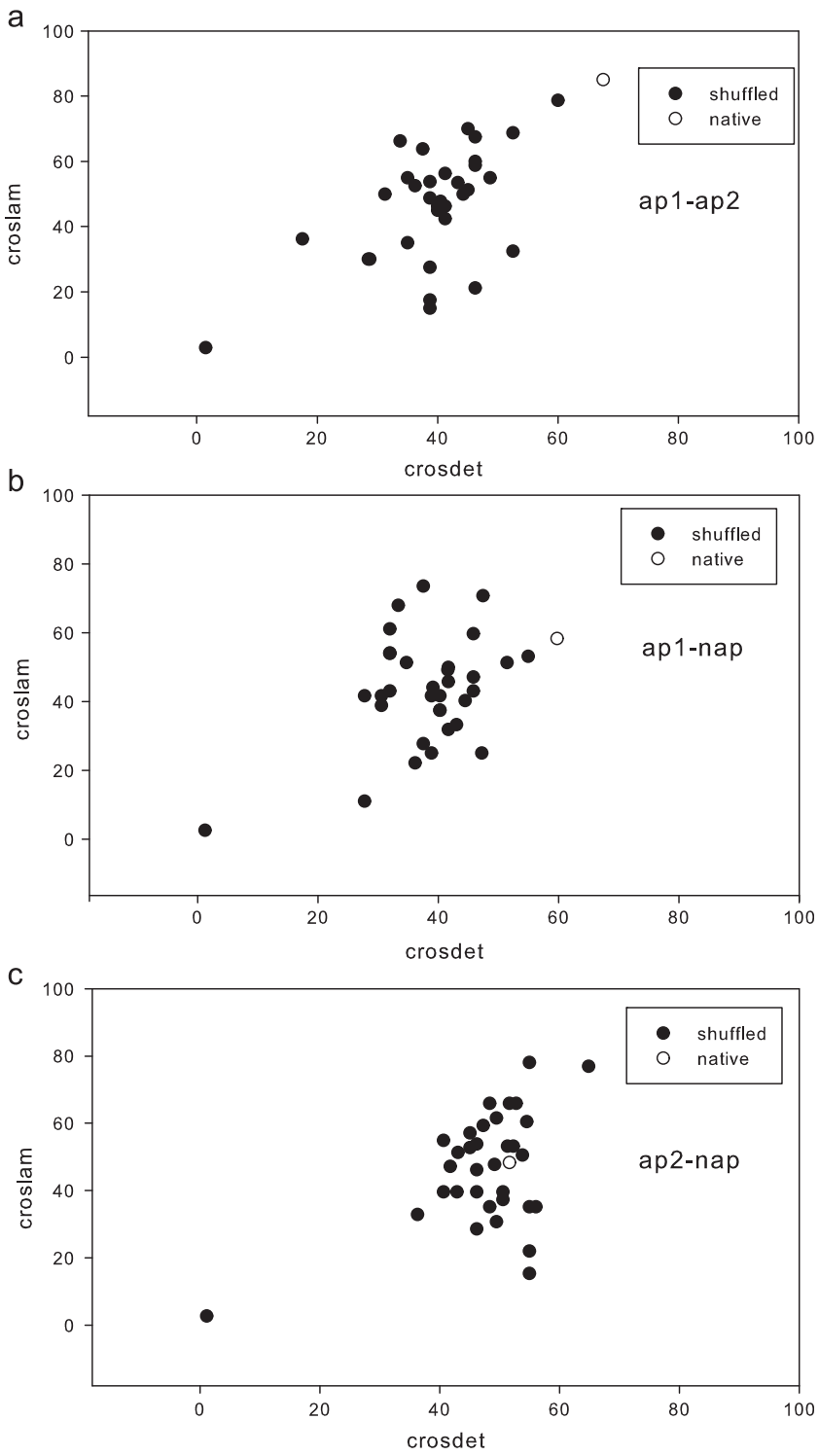

Fig. 4. (a)–(c) Cross-determinism (crosdet) vs. Cross laminarity (croslam) of ap1–ap2, ap1–nap and ap2–nap of 'ras' protein, respectively. Here white dot represent the native sequence and black dots represent the 30 shuffled sequences.

by thermodynamic phenomena (Chung-Jung Tsai and Nussinov, 2008).

The Cross-RQA study between 30 shuffled sequences of allosteric susceptible regions ap1–ap2, and between allosteric susceptible and non-allosteric region of 'ras' proteins ap1–nap and ap2–nap also revealed that ap1–ap2 native pair were much more distinct and optimized in terms of cross-correlation than its shuffled counterparts.

The presence of a marked departure of native and shuffled cross-correlation discloses that a specific selective force is experienced by the ap1–ap2 pair with compared to ap–nap pairs. We found that native ap1–nap pair is to some extent more correlated than its shuffled counterparts whereas the native ap2–nap pairs are not. Fig. 4(a) shows the high correlation between the two allosteric susceptible patches ap1–ap2 in the native sequence than all 30 shuffled allosteric susceptible patches.

Table 2
Results of various WEKA classifiers using test dataset with Sen (Sensitivity), Spec (Specificity), Acc (Accuracy) and AUC (Area under curve) measures.

Dataset	Classifiers	Sensitivity (%)	Specificity (%)	Accuracy (%)	Area under curve (AUC)
Allosteric proteins vs. randomly generated sequences	Mulitlayer perceptron	94.4	86.1	90.3	0.95
	Naïve Bayes	83.3	100	91.7	0.96
	Random forest	88.9	95	81.9	0.90
	LibSVM	27.8	94.4	61.1	0.611
Generic proteins vs. randomly generated sequences	Mulitlayer perceptron	76.1	85.2	85.7	0.909
	Naïve Bayes	76.1	100	88	0.90
	Random forest	100	95.5	97.7	0.996
	LibSVM	85.7	100	93	0.995
Allosteric proteins vs. generic proteins	Mulitlayer perceptron	86.4	72.7	79.5	0.808
	Naïve Bayes	90.9	68.2	79.5	0.82
	Random forest	86.4	68.2	77.3	0.85
	LibSVM	90	50	70.5	0.705

However, the cross-correlation of the pairs constituted by an allosteric and a non-allosteric patch (ap1–nap, ap2–nap) were found to be correlated to some extent when compared to random sequence. This may be due to the global character of allostery together with the presence of 'general co-evolution' of the protein 'as a whole'. The 'extreme' character of allosteric interactions ap1–ap2 pair are evident from Fig. 4(a)–(c). Here we plotted the native pair (as white dot) together with shuffled peptide pairs (as black dots) in the Cross-determinism/Cross-laminarity plane.

From this study, we infer that nature exerts selective force on the sequence order of the allosterically susceptible regions for allosteric regulation. It is worth noting the fact that only ap1–ap2 native pair has the 'most extreme' Cross-laminarity (bold text in Table 2) with respect to the shuffled counterparts. This highlights a sort of 'maximum of fitness' driven by the need to optimize the interaction efficiency of the allosteric hotspots. It also confirms the relevance of laminarity in the allosteric motion (Namboodiri et al., 2010). Laminar regions correspond to the repetition of a common (short) pattern of amino acids, i.e., to low-complexity regions that were already demonstrated to be linked to particular flexible and natively unfolded regions (Porrello et al., 2004; Colafranceschi et al., 2005; Zbilut et al., 1998; Zbilut et al., 2004; Giuliani et al., 2002).

The RQA analyzed enhanced allosteric dataset resulted in parameters which when fed into pattern recognition tools available in WEKA as tabulated in Table 2. We could observe from the results that the Multilayer Perceptron, Naive Baye's and Random forest classifiers produced significant classification. Amongst these, the Naive Bayes classifier and Multilayer Perceptron classifier classified the test dataset with 94% sensitivity, 86% specificity, 90% accuracy and 83% sensitivity, 100% specificity and 92% accuracy, respectively whereas the Random forest classifier classified with 89% sensitivity, 75% specificity and 82% accuracy thus giving a proof-of-concept of the relevance of RQA descriptors on both single regions and whole protein scales.

Experiment with a set of randomly generated protein set and generic single chain protein set with no imposed bias toward allosteric systems demonstrated very strong discrimination between real protein and random sequences with specificity and accuracy ranging from 95–100% to 85–90%, respectively. Albeit preliminary, this result is of utmost importance as it proves that the widespread notion of proteins as 'random sequences' is no more tenable.

Experiment with an allosteric set and generic single chain real protein set gave rise to a discrimination power with specificity and accuracy ranging between 63–68% and 72–79%, respectively using recurrence, mean and standard deviation as major attributes. The relatively low discrimination accuracy indirectly confirmed the notion that any protein has a residual allosteric character. However, we could infer that the allosteric character has a gradation and that our allosteric set was representative of greater 'allostery' than a generic protein data set.

By combining the results of our initial (allosteric set vs. randomly generated protein counterpart) and confirmative (allosteric vs. generic sets) we could infer that RQA 'markers' of allostery are indeed able to discriminate allosteric proteins from non-allosteric proteins. These results assert that allosteric proteins have a global signature in terms of hydrophobicity patterning which was earlier observed in our previous studies on a smaller dataset.

5. Conclusion

Our work suggests that hydrophobicity patterning has a significant role in determining allostery. It is important to note that we have compared 'real allosteric protein' vs. 'randomly generated protein set.' that have with no relation to their allostery propensity. We could demonstrate that the specifically selected 'allosteric' character of the set formed the basis of the observed separation. Our pattern recognition method was based on the subtle protein sequence-structure-dynamics relations of which allostery formed a very prominent feature (Keefe and Szostak, 2001). We detected specific hydrophobicity signatures in proteins in which the allostery is 'particularly intense' compared to randomly generated protein set.

Going into 'microscopic scale' and looking at the specific peptides more involved in allosteric effect in 'ras' protein, we were able to demonstrate that different allosteric 'drivers' experience a kind of evolutionary 'optimization'. This helped us infer that transmission of the allosteric signal could be mediated by protein dynamics bringing two initially far-apart allosteric-prone sites functionally near to each other. In other words the 'transmission line' of allosteric effect could not reside in the peptide bonds but on the establishment of transient weak molecular bonds between patches.

One of the potential applications of 'allosteric signature' could be in protein engineering in terms of designing structures that retain this key property while changing other variables in artificial constructs. The hypothesis set forth by Nussinov et al. (2011) greatly enhance the application range of our results. They put forth that allosteric communication is not restricted to single isolated proteins but extends to different interacting proteins which communicate in the signaling pathway. Thus, a drug designed as 'allosteric drug' (i.e., a molecule that instead of antagonizing the binding to the active site of the natural

substrate, exerts an allosteric action on its receptor) may have a therapeutically relevant effect on the entire pathway its target protein is embedded into. In this realm, having a reliable characterization of allosteric sites in a protein could prove beneficial to drive drug design.

On a more methodological ground this work demonstrates the possibility to integrate knowledge and methods coming from very different fields like systems biology, nonlinear dynamics, evolution biology and multidimensional statistics into a consistent strategy of analysis.

References

Amaro, R.E., Sethi, A., Myers, R.S., Davisson, V.J., Luthey-Schulten, Z.A., 2007. A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. Biochemistry 46, 2156–2173.

Bradley, Michael John, Ph.D, Computational and Experimental Investigation of Allosteric Communication in the Transcriptional Regulator NikR., Washington University in St. Louis, 203 (2009) 3387550.

Bruni, R., Costantino, A., Tritarelli, E., Marcantonio, C., Ciccozzi, M., Rapicetta, M., El Sawaf, G., Giuliani, A., Ciccaglione, A.R., 2009. A computational approach identifies two regions of Hepatitis C Virus E1 protein as interacting domains involved in viral fusion process. BMC Struct. Biol. 9, 48.

Brunger, Axel T., Michael, V., Milburn, Tong, Liang, Abraham, M., Devos, Jancarik, Jarmila, Yamaizumi, Ziro, Nishimura, Susumu, Ohtsuka, Eiko, Kim, Sung-Hou, 1990. Crystal structure of an active form of RAS protein, a complex of a GTP analog and the HRAS p21 catalytic domain. Proc. Nat. Acad. Sci. USA. Biochem. 87, 4849–4853.

Chennubhotla, C., Bahar, I., 2006. Markov propagation of allosteric effects in biomolecular systems: application to GroEL–GroES. Mol. Syst. Biol. 2, 1–13.

Charles Jr., L.Webber, Marwan, Norbert, Facchini, Angelo, Giuliani, Alessandro, 2009. Simpler methods do it better: success of Recurrence Quantification Analysis as a general purpose data analysis tool. Phys. Lett. A 373, 3753–3756.

Colafranceschi, M., Colosimo, A., Zbilut, J.P., Uversky, V.N., Giuliani, A., 2005. Structure-related statistical singularities along protein sequences: a correlation study. J. Chem. Inf. Model. 45, 183–189.

Chung-Jung Tsai, Antonio Del Sol, Nussinov, Ruth, 2008. Allostery: absence of a change in shape does not imply that allostery is not at play. J. Mol. Biol. 378 (1), 1–11.

Del Sol, A., Fujihashi, H., Amoros, D., Nussinov, R., 2006. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol. Syst. Biol. 2, 0019.

Daily, M.D., Gray, J.J., 2007. Local motions in a benchmark of allosteric proteins. Proteins 67, 385–399.

Dima, Ruxandra I., Thirumalai, 2006. Determination of network of residues that regulate allostery in protein families using sequence analysis. Protein Sci. 15, 258–268.

Demerdash, Omar N.A., Daily, Michael D., Mitchell, Julie C., 2009. Structure-based predictive models for allosteric hot spots. PLoS Comput. Biol. 5 (10), e1000531.

Eckmann, J.P., Oliffson Kamphorst, S., Ruelle, D., 1987. Recurrence plots of dynamical systems. Europhys. Lett. 91, 973–977.

Falcon, C.M., Matthews, K.S., 2001. Engineered disulfide linking the hinge regions within lactose repressor dimer increases operator affinity decreases sequence selectivity and alters allostery. Biochemistry 40, 15650–15659.

Gunasekaran, K., Ma, B., Nussinov, R., 2004. Is allostery an intrinsic property of all dynamic proteins? Proteins: structure. Funct. Bioinf. 57, 433–443.

Gerstein, M., Krebs, W.G., 1998. A database of macromolecular motions. Nucleic Acids Res. 26, 4280–4290.

Giuliani, A., Tomasi, M., 2002. Recurrence quantification analysis reveals interaction patterns in paramyxoviridae envelope glycoproteins. Proteins 46, 171–176.

Giuliani, A., Benigni, R., Colafranceschi, M., Chandrashekar, I., Cowsik, S.M., 2003. Large contact surface interaction between proteins detected by time series analysis methods: case study on C-phycocyanins. Proteins 51, 299–310.

Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., Cohen, F., 2000. Co-evolution of proteins with their interaction partners. J. Mol. Biol. 299 (2), 283–293.

Giuliani, A., Benigni, R., Zbilut, J.P., Webber, C.L., Sirabella, P., Colosimo, A., 2002. Nonlinear signal analysis methods in the elucidation of protein sequence/structure relationships. Chem. Rev. 102, 1471–1491.

Hilser, V.J., Freire, E., 1996. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. J. Mol. Biol. 262, 756–772.

Huang, Zhimin, Zhu, Liang, Cao, Yan, Wu, Geng, Liu, Xinyi, Chen, Yingyi, Wang, Qi, Shi, Ting, Zhao, Yaxue, Wang, Yuefei, Li, Weihua, Li, Yixue, Chen2, Haifeng, Chen, Guoqiang, Zhang, Jian, 2011. ASD: a comprehensive database of allosteric proteins and modulators. Nucleic Acids Res. 39 (Database issue).

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, Witten, Ian H., 2009. The WEKA data mining software: an update. SIGKDD Explorat. 11 (Issue 1).

〈http://www.drgutman.org/ORIGINAL_PAPERS/%2314.pdf〉.

〈http://onlinelibrary.wiley.com/doi/10.1002/bip.20607/full〉.

Keefe, A.D., Szostak, J.V., 2001. Functional proteins from a random-sequence library. Nature 410, 715–718.

Lim, W.A., 2002. The modular logic of signaling proteins: building allosteric switches from simple binding domains. Curr. Opin. Struct. Biol. 12, 61–68.

Liu, T., Whitten, S.T., Hilser, V.J., 2006. Ensemble-based signatures of energy propagation in proteins: a new view of an old phenomenon. Proteins: Struct. Funct. Bioinf. 62, 728–738.

Lockless, S.W., Ranganathan, R., 1999. Evolutionary conserved pathways of energetic connectivity in protein families. Science 286, 295–299.

Monod, J., Changeux, J.P., Jacob, F., 1963. Allosteric proteins and cellular control systems. J. Mol. Biol. 20, 306–329.

Masterson, Larry R., Mascioni, Alessandro, Traaseth, Nathaniel J., Taylor, Susan S., Veglia, Gianluigi, 2008. Allosteric cooperativity in protein kinase. A. PNAS 105 (2), 511.

Marwan, N., Carmen, R.M., Thiel, M., Kurths, J., 2007. Recurrence plots for the analysis of complex systems. Phys. Rep. 438, 237–329.

Namboodiri, S., Verma, C., Dhar, P.K., Giuliani, A., Nair, A.S., 2010. Sequence signatures of allosteric proteins: toward rational design. Syst. Synth. Biol. 4, 271–280.

Nussinov, R., Tsai, C.J., Csermely, P., 2011. Allo-network drugs: harnessing allostery in cellular networks. Trends Pharmacol. Sci. 32 (12), 686–693.

Ota, N., Agard, D.A., 2005. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. J. Mol. Biol. 351, 345–354.

Porrello, A., Soddu, S., Zbilut, J.P., Crescenzi, M., Giuliani, A., 2004. Discrimination of single amino acid mutations of the p53 protein by means of deterministic singularities of Recurrence Quantification Analysis. Proteins: Struct. Funct. Bioinf. 55, 743–755.

Packard, N.H., Crutchfield, J.P., Farmer, J.D., Shaw, R.S., 1980. Geometry from a time series. Phys. Rev. Lett. 45 (9), 712–716.

Selz, Karen A., Samoylova, Tatiana I., Samoylov, Alexandre M., Vodyanoy, Vitaly J., Mandell, Arnold J., 2006. Designing allosteric peptide ligands targeting a globular protein. Biopolymers 85, 38–59.

Takens, F., 1981. Detecting Strange Attractors in Turbulence Lecture Notes in Math.. Springer, New York 898.

Tamura K., Dudley J., Nei M., Kumar S., MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24 (2007) 1596–1599. (Publication PDF at 〈http://www.kumarlab.net/publications〉).

Voorhees, Rebecca M., Martin Schmeing, T., Kelley, Ann C., Ramakrishnan, V., 2010. The mechanism for activation of GTP hydrolysis on the ribsome. Science 330, 835–838.

Wang, X., Kemp, R.G., 2001. Reaction path of phosphofructo-1-kinase is altered by mutagenesis and alternative substrates. Biochemistry 40, 3938–3942.

Whitten, S.T., Garcia-Moreno, B., Hilser, V.J., 2005. Local conformational fluctuations can modulate the coupling between proton binding and global structural transitions in proteins. Proc. Nat. Acad. Sci. U.S.A. 102, 4282–4287.

Zheng, W.J., Brooks, B., 2005. Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. J. Mol. Biol. 346, 745–759.

Zheng, W.J., Liao, J.C., Brooks, B.R., Doniach, S., 2007. Toward the mechanism of dynamical couplings and translocation in Hepatitis C Virus NS3 helicase using elastic network model. Proteins: Struct. Funct. Bioinf. 67, 886–896.

Zbilut, J.P., Giuliani, A., Webber, C.L., 1998. Detecting deterministic signals in exceptionally noisy environments using Cross-Recurrence Quantification. Phys. Lett. A 246, 122–128.

Zbilut, J.P., Giuliani, A., Webber, C.L., 1998. Detecting deterministic signals in exceptionally noisy environments using Cross-Recurrence Quantification. Phys. Lett. A 246, 122–128.

Zbilut, J.P., Giuliani, A., Colosimo, A., Mitchell, J.C., Colafrancesch, M., Marwan, N., Webber, C.L., Uversky, V., 2004. Charge and hydrophobicity patterning along the sequence predicts the folding mechanism and aggregation of proteins: a computational approach. J. Proteome Res. 3, 1243–1253.