Syst Synth Biol (2010) 4:271–280
DOI 10.1007/s11693-011-9072-9

RESEARCH PAPER

Sequence signatures of allosteric proteins towards rational design

Saritha Namboodiri · Chandra Verma ·
Pawan K. Dhar · Alessandro Giuliani ·
Achuthsankar S. Nair

Received: 1 February 2011 / Accepted: 3 February 2011 / Published online: 22 February 2011

Abstract Allostery is the phenomenon of changes in the structure and activity of proteins that appear as a consequence of ligand binding at sites other than the active site. Studying mechanistic basis of allostery leading to protein design with predetermined functional endpoints is an important unmet need of synthetic biology. Here, we screened the amino acid sequence landscape in search of sequence-signatures of allostery using Recurrence Quantitative Analysis (RQA) method. A characteristic vector, comprised of 10 features extracted from RQA was defined for amino acid sequences. Using Principal Component Analysis, four factors were found to be important determinants of allosteric behavior. Our sequence–based predictor method shows 82.6% accuracy, 85.7% sensitivity and 77.9% specificity with the current dataset. Further, we show that *Laminarity-Mean-hydrophobicity* representing repeated hydrophobic patches is the most crucial indicator of allostery. To our best knowledge this is the first report that describes sequence determinants of allostery based on hydrophobicity. As an outcome of these findings, we plan to explore possibility of inducing allostery in proteins.

Keywords Allostery · Recurrence Quantitative Analysis · Sequence-based predictor · Hydrophobicity

Introduction

Binding of a ligand at particular sites of certain proteins produces unique structural changes in the protein. This effect known as allostery, accompanies conformational changes associated with altering interactions at the active sites (which may be far away from the causal binding site) (Monod et al. 1963). Allosteric behavior affects the functional property of the protein and is therefore, important in understanding the construction of proteins and designing novel interactions. Figure 1 illustrate the allosteric behavior of a protein.

Further, understanding allostery can be used to study metabolic regulation, cell signaling and provide pointers to new classes of drugs that function by allosteric mechanisms (Berg et al. 2002). This raises key sequence-centric questions. Can allostery be diagnosed/predicted from sequence information alone? If so, how and to what level of accuracy? Bio-informatics approaches to answer such questions begin with enquiry into sequence signatures at the amino acid residue level. Lockless et al. (2003) identified evolutionary conserved networks of residues in three structurally and functionally distinct protein families, G protein–coupled receptors, the chymotrypsin class of serine proteases and hemoglobins, using multiple sequence alignment. Subsequently, residues involved in allosteric interactions were identified from sequence conservation in PDZ (Post synaptic density protein (PSD95), Drosophila disc large

S. Namboodiri · A. S. Nair (✉)
State Inter University Centre of Excellence in Bioinformatics,
University of Kerala, Kariyavattom Campus,
Thiruvananthapuram, Kerala, India
e-mail: sankar.achuth@gmail.com

S. Namboodiri
e-mail: saritha16.namboodiri@gmail.com

C. Verma
Bioinformatics Institute (BII), Buona Vista, Singapore

P. K. Dhar
Centre for Systems and Synthetic Biology, University of Kerala,
Kariyavattom Campus, Thiruvananthapuram, Kerala, India

A. Giuliani
Environment and Health Deptartment, Istituto Superiore di
Sanità, Rome, Italy

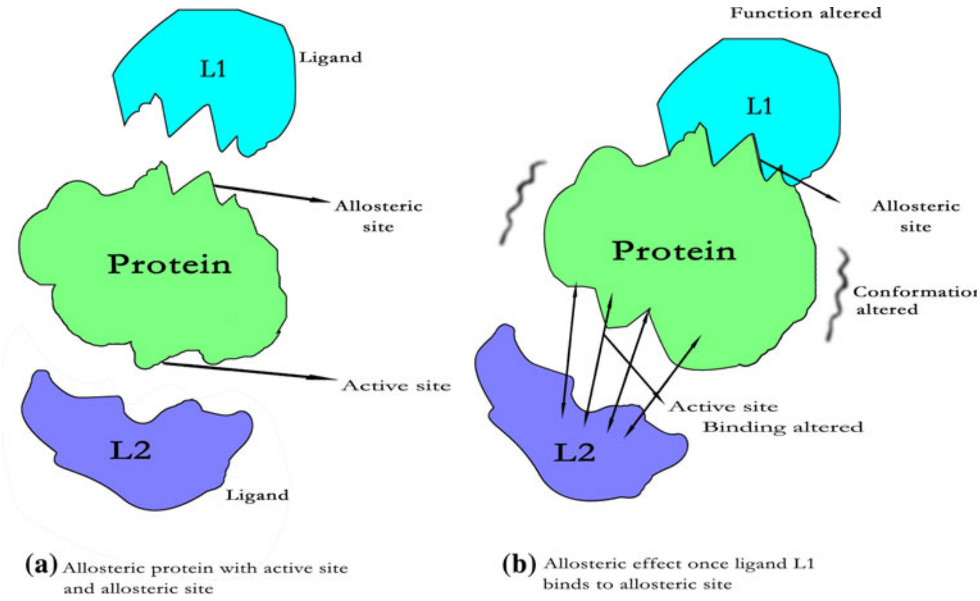

Fig. 1 Allosteric behavior of a protein

tumor suppressor (DlgA), and Zonula occludens-1 protein (zo-1)) domain family, G protein-coupled receptors and Lectins using multiple sequence alignment (Dima and Thirumalai 2006). We did not come across any study aimed at extracting protein sequence features and connecting with the higher level allostery phenomena. Our goal was not a symbolic analysis of the sequence information, but a holistic feature extraction obtained through a transformation of sequence information into a global representation.

Recurrence Quantification Analysis (RQA) is a non-linear dynamic approach based on Recurrence plot, a graphical tool, developed by Eckmann et al. (1987) to study recurrence phenomena. It was later quantified by Zbilut and Webber (1992) and popularised as Recurrence Quantification Analysis. RQA has been found to be suitable in finding patterns in short, non-stationary numerical sequences and has been successfully applied to the analysis of amino acid sequences of proteins (Porrello et al. 2004; Zbilut et al. 2004; Colafranceschi et al. 2005; Bruni et al. 2009).

Recurrence Quantification Analysis was used for the comparison of different signals that in our study are amino acid sequences (Porrello et al. 2004; Zbilut et al. 2004; Colafranceschi et al. 2005; Bruni et al. 2009). In this work we demonstrate how the sequence representation, based on the dynamical signature of residue hydrophobicity distribution along the chain, allows for the prediction of allosteric character of proteins.

Materials and methods

Data set

We used available dataset of allosteric protein consisting of 51 allosteric proteins and 21 non-allosteric proteins from the previously compiled database (Daily and Gray 2007). This database describes six local motions of residues that define allostery and is currently the best known dataset to our knowledge. To meet good data quality standards, we considered only single chain proteins for an unbiased representation in terms of RQA of their sequences. This in turn implies that our model focuses on the conformational changes at the tertiary level of the structures, which controls the basic functions of the protein. Changes in the quaternary structure are outside the scope of the current proposed approach. Single chain proteins of varying lengths were taken from different species of bacteria, fungi and mammals including *R. norvegicus, H. sapiens, O. mykiss B. Taurus and P. catodon*. These fall under the classes of allosteric and signaling proteins such as DNA-binding proteins, G proteins, protein kinases and bacterial response regulators. The allostery protein dataset comprised of 14 allosteric and 9 non-allosteric single chain proteins of varying lengths and these are shown in Table 1.

Method

Recurrence quantification analysis as applied to the protein data set

Recurrence Quantification Analysis was applied to transform allostery protein dataset into quantified statistical variables. Among several options available, we chose hydrophobicity as a key parameter to effect this transformation since it significantly impacts protein folding and interaction dynamics. Based on the extracted signature features of allosteric behavior, we developed a tool for predicting allostery from sequence information alone.

Table 1 Data set of allosteric and non-allosteric proteins used in this study (Daily and Gray 2007)

Sl. No.	Id	Description
Allosteric		
1	1FTN	Human RhoA-GDP complex
2	1IRK	Tyrosine kinase domain of the human insulin receptor.
3	1KAO	Small G protein Rap2A in complex with its substrate GTP, with GDP and with GTPgammaS
4	2TRT	Tet repressor-tetracycline complex and regulation of antibiotic resistance
5	1XTQ	GTPase RHEB
6	1E0S	Guanine nucleotide exchange factors
7	1NH8	ATP phosphoribosyltransferase from Mycobacterium tuberculosis
8	4Q21	Protooncogenic ras proteins.
9	1L5Z	Sinorhizobium muliloti dctd
10	1T48	Protein tyrosine phosphatase 1B.
11	1TAG	Alpha-subunit of a heterotrimeric G protein
12	3CHY	Escherichia coli CheY
13	1CD5	Glucosamine-6-phosphate deaminase
14	1ERK	MAP kinase ERK2
Non-allosteric		
15	1A6 M	Myoglobin-ligand complexes
16	1DG3	Human guanylate-binding protein 1 representing a unique class of GTP-binding proteins
17	1D2 K	Chitinase from the pathogenic fungus Coccidioides immitis.
18	1S0Q	Pancratic bovine Trypsin native and inhibited with Benzamidine from synchotron data.
19	1J9Y	Mannanase 26A from Pseudomonas cellulose
20	1D2 K	Chitinase from the pathogenic fungus Coccidioides immitis.
21	1LMN	Lysozyme from the rainbow trout (Oncorhynchus mykiss).
22	1KHI	Woronin body function in Neurospora crassa.
23	1GQZ	Haemophilus influenzae diaminopimelic acid epimerase (DapF)

To obtain a recurrence plot for each of the hydrophobicity mapped amino acid sequences of allostery protein dataset, we first produced an embedding matrix which projects the sequence information into a higher dimensional space. Distance matrix was computed from this embedding matrix by finding the Euclidean distance among all the rows of the matrix. From the distance matrix, values that fell below a predefined radius were considered as recurrences and set to 1 and the rest of the values were set to 0 to form the recurrence matrix. Recurrence plot is a visualization of a recurrence matrix where all the 1s and 0s of the recurrence matrix are replaced by black and white dots respectively.

We demonstrate our method through a toy example. Consider a hypothetical protein having a length of 22 amino acids: MTRMTRMTRGHHHHHELHHHHH. This sequence was mapped by replacing all amino acids with their corresponding Miyazawa Jernigan hydrophobicity index (Jernigan 1985) (this index has been shown to be satisfactory from other studies).

8.95 4.49 4.10 8.95 4.49 4.10 8.95 4.49 4.10 4.48 5.10
5.10 5.10 5.10 5.10 3.65 8.47
5.10 5.10 5.10 5.10 5.10

This series was mapped into an embedding matrix. Each column of the embedding matrix is a delayed copy of the original series. The number of column elements represents the embedding dimension. The embedding matrix for our toy example with embedding dimension = 3 and time delay = 2 are as given below:

$$\mathrm{EM} = \begin{bmatrix} 8.95 & 4.10 & 4.49 \\ 4.49 & 8.95 & 4.10 \\ 4.10 & 4.49 & 8.95 \\ 8.95 & 4.10 & 4.49 \\ 4.49 & 8.95 & 4.10 \\ 4.10 & 4.49 & 4.48 \\ 8.95 & 4.10 & 5.10 \\ 4.49 & 4.48 & 5.10 \\ 4.10 & 5.10 & 5.10 \\ 4.48 & 5.10 & 5.10 \\ 5.10 & 5.10 & 5.10 \\ 5.10 & 5.10 & 3.65 \\ 5.10 & 5.10 & 8.47 \\ 5.10 & 3.65 & 5.10 \\ 5.10 & 8.47 & 5.10 \\ 3.65 & 5.10 & 5.10 \\ 8.47 & 5.10 & 5.10 \\ 5.10 & 5.10 & 5.10 \end{bmatrix}$$

A distance matrix was constructed by computing the pairwise Euclidean norm of all the rows of the embedding matrix as follows:

$$\mathrm{D}(\mathrm{P},\mathrm{Q}) = \sqrt{(\mathrm{P}_1 - \mathrm{Q}_1)^2 + (\mathrm{P}_2 - \mathrm{Q}_2)^2 + \ldots (\mathrm{P}_\mathrm{n} - \mathrm{Q}_\mathrm{n})^2}$$

where (P_1, P_2, P_3 … P_n) and (Q_1, Q_2, Q3 … Q_n) are the two rows in consideration. The distance matrix of our example is computed and given below:

$$D = \begin{bmatrix}
0.00 & 6.60 & 6.60 & 0.00 & 6.60 & 4.87 & 0.61 & 4.52 & 4.99 & 4.62 & 4.02 & 4.07 & 5.63 & 3.92 & 5.86 & 5.43 & 1.27 & 4.02 \\
6.60 & 0.00 & 6.60 & 6.60 & 0.00 & 4.49 & 6.66 & 4.58 & 4.00 & 3.98 & 4.02 & 3.92 & 5.86 & 5.43 & 1.27 & 4.07 & 5.63 & 4.02 \\
6.60 & 6.60 & 0.00 & 6.60 & 6.60 & 4.47 & 6.20 & 3.87 & 3.90 & 3.92 & 4.02 & 5.43 & 1.27 & 4.07 & 5.63 & 3.92 & 5.86 & 4.02 \\
0.00 & 6.60 & 6.60 & 0.00 & 6.60 & 4.87 & 0.61 & 4.52 & 4.99 & 4.62 & 4.02 & 4.07 & 5.63 & 3.92 & 5.86 & 5.43 & 1.27 & 4.02 \\
6.60 & 0.00 & 6.60 & 6.60 & 0.00 & 4.49 & 6.66 & 4.58 & 4.00 & 3.98 & 4.02 & 3.92 & 5.86 & 5.43 & 1.27 & 4.07 & 5.63 & 4.02 \\
4.87 & 4.49 & 4.47 & 4.87 & 4.49 & 0.00 & 4.91 & 0.73 & 0.87 & 0.95 & 1.33 & 1.44 & 4.16 & 1.45 & 4.15 & 0.98 & 4.46 & 1.33 \\
0.61 & 6.66 & 6.20 & 0.61 & 6.66 & 4.91 & 0.00 & 4.48 & 4.95 & 4.58 & 3.98 & 4.23 & 5.21 & 3.88 & 5.82 & 5.39 & 1.11 & 3.98 \\
4.52 & 4.58 & 3.87 & 4.52 & 4.58 & 0.73 & 4.48 & 0.00 & 0.73 & 0.62 & 0.87 & 1.69 & 3.48 & 1.03 & 4.04 & 1.04 & 4.03 & 0.87 \\
4.99 & 4.00 & 3.90 & 4.99 & 4.00 & 0.87 & 4.95 & 0.73 & 0.00 & 0.38 & 1.00 & 1.76 & 3.52 & 1.76 & 3.52 & 0.45 & 4.37 & 1.00 \\
4.62 & 3.98 & 3.92 & 4.62 & 3.98 & 0.95 & 4.58 & 0.62 & 0.38 & 0.00 & 0.62 & 1.58 & 3.43 & 1.58 & 3.43 & 0.83 & 3.99 & 0.62 \\
4.02 & 4.02 & 4.02 & 4.02 & 4.02 & 1.33 & 3.98 & 0.87 & 1.00 & 0.62 & 0.00 & 1.45 & 3.37 & 1.45 & 3.37 & 1.45 & 3.37 & 0.00 \\
4.07 & 3.92 & 5.43 & 4.07 & 3.92 & 1.44 & 4.23 & 1.69 & 1.76 & 1.58 & 1.45 & 0.00 & 4.82 & 2.05 & 3.67 & 2.05 & 3.67 & 1.45 \\
5.63 & 5.86 & 1.27 & 5.63 & 5.86 & 4.16 & 5.21 & 3.48 & 3.52 & 3.43 & 3.37 & 4.82 & 0.00 & 3.67 & 4.77 & 3.67 & 4.77 & 3.37 \\
3.92 & 5.43 & 4.07 & 3.92 & 5.43 & 1.45 & 3.88 & 1.03 & 1.76 & 1.58 & 1.45 & 2.05 & 3.67 & 0.00 & 4.82 & 2.05 & 3.67 & 1.45 \\
5.86 & 1.27 & 5.63 & 5.86 & 1.27 & 4.15 & 5.82 & 4.04 & 3.52 & 3.43 & 3.37 & 3.67 & 4.77 & 4.82 & 0.00 & 3.67 & 4.77 & 3.37 \\
5.43 & 4.07 & 3.92 & 5.43 & 4.07 & 0.98 & 5.39 & 1.04 & 0.45 & 0.83 & 1.45 & 2.05 & 3.67 & 2.05 & 3.67 & 0.00 & 4.82 & 1.45 \\
1.27 & 5.63 & 5.86 & 1.27 & 5.63 & 4.46 & 1.11 & 4.03 & 4.37 & 3.99 & 3.37 & 3.67 & 4.77 & 3.67 & 4.77 & 4.82 & 0.00 & 3.37 \\
4.02 & 4.02 & 4.02 & 4.02 & 4.02 & 1.33 & 3.98 & 0.87 & 1.00 & 0.62 & 0.00 & 1.45 & 3.37 & 1.45 & 3.37 & 1.45 & 3.37 & 0.00
\end{bmatrix}$$

Distance matrix is a symmetric matrix having $n \times n$ elements (where 'n' is the number of rows in embedding matrix) with an all-zero main diagonal. This distance matrix was transformed into the recurrence matrix by implementing a cut-off limit (radius) to capture the coarse information from the matrix. All the elements in the distance matrix with distance equal to or below the selected radius were set to 1 and all other elements to 0. The recurrence matrix, RM for our example with radius $r = 0.5$ is given below:

Figure 2 illustrates the recurrence plot of a hypothetical protein MTRMTRMTRGHHHHHELHHHHH with embedding dimension = 3, radius = 0.5 and time delay = 2. Here 0 is represented as a white dot and 1 as a black dot.

Given that fact that for quantitative studies, visual representation is insufficient, a meaningful quantification was derived out of recurrence plots using ten such quantifications.

$$\mathrm{RM} = \begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

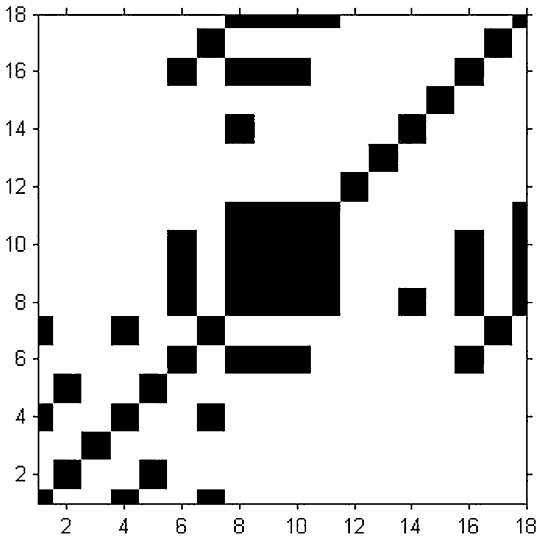

Fig. 2 Recurrence plot of 'ras' (PDB id – 42Q1) protein. The highlighted square indicates a hydrophobicity patch

R1: Recurrence measure is a measure of the proportion of recurrence phenomena derived from the recurrence plot. This is indicated by the 1s in the recurrence matrix. R1 is defined as *the number of non zero elements in the recurrence matrix (excluding the main diagonal elements)/total number of elements in the recurrence matrix*. In our example this is 0.15 (i.e., 42/324).

R2: Deterministic measure represents a measure of the predictability (determinism) of the system under consideration. A periodic signals results in very long diagonal lines whereas chaotic and stochastic signals give short diagonal lines or no diagonal lines at all. We considered the proportion of the recurrent points forming diagonal lines having a length of at least 2. Diagonal lines orthogonal to the main diagonal line were not considered. R2 was defined as the number of points in the diagonal lines divided by the total number of recurrence points excluding the elements of the main diagonal. The value of deterministic measure for our example turned out to be 0.43 (i.e., 26/61).

R3: Lmax is the length of the longest diagonal line excluding the main diagonal line. As diagonal lines are related to determinism, if R3 is small (≤ 2) it indicates chaotic system. If the value is large (>2), it indicates a deterministic system. In the case of our toy example the length of the diagonal line other than the main diagonal is 2.5.

R4: Information entropy is a measure of the Shannon information entropy computed over the frequency distribution of the lengths of the diagonal lines of recurrent points. It measures the variety of diagonal lines and reflects the complexity of the Recurrence Plot. For periodic systems its value is rather small (≤ 2) indicating low complexity. For our example R4 is 0.69.

R5: Laminarity measure is a measure of the proportion of recurrence points forming vertical/horizontal lines having at least a minimum length of 2. R5 is defined as the number of points in such vertical/horizontal lines divided by the total number of recurrence points. In our example the laminarity measure was found to be 0.43 (i.e., 26/61).

R6: Trap Time measures the average length of vertical lines indicating the mean time that the system will remain in a specific state. R6 is defined as the number of recurrence points forming vertical/horizontal lines divided by the length of the vertical/horizontal structure. In our example the trap time was found to be 2.88 (26/ 2 + 3 + 4).

R7: T1 (Mean Recurrence Time—Type I) is the average time distance between a point and its recurrence in the embedding matrix. In the case of our example this was found to be 2.69.

R8: *T2 (Mean Recurrence Time—Type II)* is the average time distance between a point and its recurrence in the embedding matrix excluding time distance of one unit. It is a measure of time distance with laminarity measure removed. In our example this was found to be 4.23.

In addition to the above order-dependant quantifications, two more order independent quantifications were added for analysis. These were R9, the hydrophobicity values of all amino acids in the given sequence and R10, the standard deviation of the hydrophobicity value of all the amino acids in the sequence. In the example these corresponded to 5.48 and 1.67 respectively.

The 10 element feature vector for our toy example equals following values:

$$R = \begin{array}{cccccccccc} \mathrm{R1} & \mathrm{R2} & \mathrm{R3} & \mathrm{R4} & \mathrm{R5} & \mathrm{R6} & \mathrm{R7} & \mathrm{R8} & \mathrm{R9} & \mathrm{R10} \\ [0.15 & 0.43 & 2.0 & 0.69 & 0.43 & 2.86 & 2.69 & 4.23 & 5.48 & 1.67] \end{array}$$

Principal component analysis of RQA descriptors

The 10 element feature vector was extracted for each protein in the allostery dataset using Recurrence

Quantitative Analysis. The characteristic matrix was derived as follows:

$$A = \begin{bmatrix}
0.17 & 0.77 & 3.04 & 1.41 & 0.57 & 3.28 & 5.67 & 9.50 & 5.21 & 1.96 \\
0.15 & 0.76 & 2.95 & 1.35 & 0.47 & 3.17 & 6.54 & 9.68 & 5.45 & 2.04 \\
0.15 & 0.75 & 2.98 & 1.36 & 0.48 & 3.05 & 6.11 & 9.17 & 5.25 & 1.96 \\
0.13 & 0.74 & 2.83 & 1.25 & 0.40 & 2.79 & 6.78 & 9.14 & 5.38 & 1.90 \\
0.17 & 0.81 & 3.18 & 1.50 & 0.56 & 3.55 & 5.45 & 9.23 & 5.25 & 2.00 \\
0.16 & 0.77 & 3.18 & 1.50 & 0.54 & 3.59 & 5.93 & 9.85 & 5.30 & 1.96 \\
0.14 & 0.73 & 2.88 & 1.30 & 0.38 & 2.94 & 6.83 & 9.23 & 5.37 & 1.92 \\
0.15 & 0.76 & 2.88 & 1.29 & 0.58 & 3.04 & 5.83 & 9.67 & 5.18 & 1.86 \\
0.14 & 0.80 & 3.15 & 1.47 & 0.33 & 3.40 & 6.56 & 8.66 & 5.42 & 2.07 \\
0.14 & 0.72 & 2.83 & 1.25 & 0.47 & 2.89 & 6.58 & 9.56 & 5.17 & 1.86 \\
0.14 & 0.75 & 2.93 & 1.33 & 0.44 & 3.06 & 6.64 & 9.50 & 5.37 & 1.95 \\
0.16 & 0.77 & 3.19 & 1.50 & 0.55 & 3.54 & 5.62 & 9.54 & 5.19 & 1.89 \\
0.16 & 0.75 & 3.05 & 1.42 & 0.48 & 3.49 & 6.03 & 9.25 & 5.28 & 1.92 \\
0.17 & 0.78 & 3.01 & 1.39 & 0.59 & 3.23 & 5.55 & 9.49 & 5.29 & 1.80 \\
0.15 & 0.75 & 2.99 & 1.36 & 0.47 & 2.88 & 6.54 & 9.57 & 5.21 & 1.93 \\
0.17 & 0.78 & 3.17 & 1.49 & 0.57 & 3.56 & 5.54 & 9.50 & 5.10 & 1.88 \\
0.19 & 0.80 & 3.22 & 1.53 & 0.59 & 3.79 & 5.27 & 9.42 & 5.18 & 2.00 \\
0.15 & 0.73 & 2.91 & 1.32 & 0.47 & 2.85 & 6.37 & 9.25 & 5.14 & 1.74 \\
0.11 & 0.73 & 3.11 & 1.44 & 0.44 & 3.81 & 7.19 & 11.02 & 5.28 & 1.78 \\
0.17 & 0.78 & 3.25 & 1.52 & 0.63 & 3.67 & 5.03 & 9.59 & 5.31 & 1.82 \\
0.15 & 0.75 & 3.01 & 1.38 & 0.39 & 3.06 & 6.25 & 8.48 & 5.36 & 1.94 \\
0.18 & 0.77 & 3.15 & 1.46 & 0.42 & 3.08 & 5.22 & 7.42 & 5.16 & 2.01 \\
0.14 & 0.77 & 3.12 & 1.46 & 0.55 & 3.67 & 6.66 & 11.18 & 5.21 & 1.81
\end{bmatrix}$$

The most prominent variables of this feature vector were extracted using Principal Component Analysis. We first processed the matrix A by subtracting the mean value of each of the columns from each element of the column and obtain the standardized matrix B. Then a standardized correlation matrix C was determined using the formula:

$$C = \frac{1}{n-1} BB^{T}$$

The correlation matrix (Table 2), C represents the interrelationship between the standardized variables and how they co-vary. This standardized correlation matrix, C is used in the characteristic equation $|C - \lambda\, I| = 0$, where I is the identity matrix, λ represents the eigen-value. Solving this equation gives rise to eigen-values and eigenvectors. Each eigen-value is the amount of the variance and the corresponding eigenvector represents the direction of variance, each eigenvector is a factor. We found ten factors that are orthogonal to each other and whose eigen-values are a linear combination of the variance of the ten variables. Table 3 shows the factors, eigen-values, the proportion of variance accounted for by the factor in percentage and the cumulative variance in percentage. The eigen-values were arranged in descending order to highlight the important eigen-values (Table 3).

Initially factors 1–6 (shown in bold) were retained which cover up to 98.9% of the total variance. From the factors retained, key factors were selected. All factor loadings with values of ±3.0 or greater were considered significant. The amount of variance of the original variables on each of these factors is represented in Table 4. The most significant factor loadings are shown in bold.

Prediction of allostery based on derived factors

The allostery dataset (Table 1) was sorted into training and test datasets. The training dataset was used to develop a model to predict the allosteric behavior of proteins and the test dataset was used to assess the goodness of the model. The training and test datasets are presented in Table 5.

First, a Linear Discriminant Classification Analysis was conducted on the training dataset. This took four crucial factors retrieved by the Principal Component Analysis as input and resulted in classification functions for allosteric and non-allosteric group of proteins as given below:

Table 2 Correlation matrix C of standardized characteristic matrix B

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
R1	1.00	0.68	0.55	0.57	0.62	0.31	−0.94	−0.40	−0.35	0.27
R2	0.68	1.00	0.75	0.76	0.48	0.59	−0.66	−0.12	−0.03	0.44
R3	0.55	0.75	1.00	0.99	0.44	0.86	−0.60	0.04	−0.20	0.18
R4	0.57	0.76	0.99	1.00	0.45	0.87	−0.59	0.05	−0.19	0.19
R5	0.62	0.48	0.44	0.45	1.00	0.51	−0.69	0.39	−0.50	−0.30
R6	0.31	0.59	0.86	0.87	0.51	1.00	−0.35	0.44	−0.12	−0.01
R7	0.94	−0.66	−0.60	−0.59	−0.69	−0.35	1.00	0.37	0.43	−0.13
R8	−0.40	−0.12	0.04	0.05	0.39	0.44	0.37	1.00	−0.06	−0.52
R9	0.35	−0.03	−0.20	−0.19	−0.50	−0.12	0.43	−0.06	1.00	0.39
R10	0.27	0.44	0.18	0.19	−0.30	−0.01	−0.13	−0.52	0.39	1.00

Table 3 Eigen-value distribution of the characteristic matrix A

Factors	Eigen values (or variance)	Proportion of variance	
		In %	In Cumulative %
1	**4.95**	**49.51**	**49.51**
2	**2.17**	**21.71**	**71.22**
3	**1.63**	**16.29**	**87.50**
4	**0.62**	**6.20**	**93.70**
5	**0.34**	**3.44**	**97.14**
6	**0.18**	**1.77**	**98.91**
7	0.05	0.55	99.46
8	0.05	0.51	99.97
9	0.00	0.02	99.99
10	0.00	0.01	100.00

Table 4 Factor loading of variables (R1 to R10)

Original variables (Elements of feature vector)	Factor loadings					
	f1	*f2*	*f3*	*f4*	*f5*	*f6*
R1	**0.82**	−0.30	−0.41	0.15	−0.05	0.13
R2	**0.85**	−0.26	0.20	0.22	0.11	−0.34
R3	**0.90**	0.03	0.32	−0.27	−0.08	0.00
R4	**0.91**	0.03	−0.33	−0.24	−0.05	0.05
R5	**0.71**	0.47	**0.49**	**0.40**	−0.04	0.06
R6	**0.76**	0.36	**0.45**	−0.08	−0.02	0.12
R7	**−0.85**	0.20	0.36	−0.08	0.14	−0.03
R8	−0.03	**0.87**	0.36	−0.26	0.18	0.04
R9	−0.35	−0.40	**0.65**	**0.41**	−0.35	0.05
R10	0.18	**−0.83**	0.31	0.10	0.38	0.15

$$\mathrm{S}_{\mathrm{testA}} = 1751.47 \times f1 - 244.46 \times f2 + 3930.75 \times f3 - \mathbf{2178.92} \times f4 - 3866.07$$

$$\mathrm{S}_{\mathrm{testNA}} = 1708.82 \times f1 - 242.18 \times f2 + 3844.75 \times f3 - \mathbf{2083.54} \times f4 - 3706.48$$

From the classification functions so obtained, we could predict the probability of a protein in the test dataset (to be allosteric or non-allosteric) using the Bayesian based Linear Discrimination Analysis approach. Likewise, we subjected the entire allostery dataset to a Linear Discriminant Classification Analysis and derived the following classification functions:

Table 5 Test and training datasets

Test dataset	Allosteric:1FTN, 1IRK, 1KAO, 2TRT, 1XTQ, 1E0S, 1NH8
	Non-allosteric:1A6 M, 1DG3, 1S0Q, 1J9Y
Train dataset	Allosteric:4Q21, 1L5Z, 1T48, 1TAG, 3CHY, 1CD5, 1ERK
	Non-allosteric: D2 K, 1LMN, 1KHI, 1GQZ, 1KHI

$$\mathrm{S}_{\mathrm{entireA}} = -3.0 \times f1 - 0.37 \times f2 + 0.09 \times f3 + \mathbf{0.70} \times f4 - 0.22$$

$$\mathrm{S}_{\mathrm{entireNA}} = 0.47 \times f1 + 0.58 \times f2 - -0.13 \times f3 - \mathbf{1.09} \times f4 - 0.52$$

Root Mean Square Deviation (RMSD), a measure of the general flexibility measure of a protein, was computed for all the proteins in the allostery dataset. Root Mean Square Deviation is a general feature of the between configuration comparison, computed over the entire protein 3D structure. In addition, a correlation between this global flexibility and allosteric flexibility (probability estimate of allostery) was also observed.

Results

Recurrence plots extracted from comparative data analysis showed single black dots, diagonal lines and vertical/horizontal lines. Single dots emerged when the states occurred

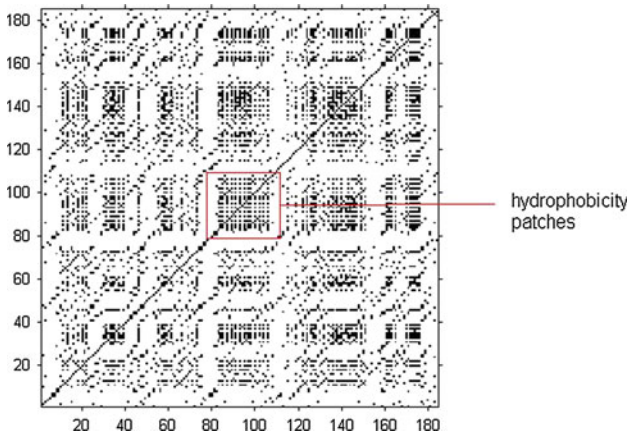

Fig. 3 Recurrence plot of the sequence MTRMTRMTRGHHHH HELHHHHH. (Embedding dimension = 3, radius = 0.5 and time delay = 2); 0 is in white and 1 is in black

sporadically. Diagonal lines running parallel to the main diagonal indicate recurrences consecutively occurring in the sequence. The diagonal lines perpendicular to the main diagonal are formed when the same states are revisited but in the reverse temporal order. Vertical/horizontal lines are formed when states are not altered for a successive period of time, i.e. when the same value is repeated. Such patches are found repeating horizontally and vertically. Figure 3 shows the recurrence plot of 'ras', an allosteric protein (PDB id–42Q1). The square box indicates a region with typical repeated hydrophobicity patches.

Six factors (Table 4) that emerged, after applying Principal Component Analysis on the characteristic vector obtained upon subjecting the allostery dataset to Recurrence Quantitative Analysis, were interpreted as follows: We observed that factor *f1* had high correlation amongst all RQA variables R1 to R7 except R8. These variables represented hydrophobicity patterns extracted by correlating a protein with itself. Thus factor *f1* was interpreted as the relative amount of hydrophobicity autocorrelation of the protein. We found that the most loaded variable on factor *f2* was R8 and was negatively correlated with R9. Thus *f2* implies that the variability in amino acid composition tends to lower the possibility of nearby recurrence. The third factor, *f3,* had R9, which stood for mean, as the leading loaded variable. This implies that high values of *f3* go together with high average hydrophobicity of the sequences. R9 (Mean) and R5 (Laminarity measure) were found to be positively correlated in factor *f4*. We represented *f4* as *Laminarity-Mean-hydrophobicity* measure. *Laminarity-Mean-hydrophobicity* measure points to the presence of repetitive elevated hydrophobicity patches along the sequence. This feature is important since laminarity measure was observed to correlate with high motion and flexible sequence patches (Zbilut et al. 2004). Thus we could interpret that the presence of relatively long repetitive patches of hydrophobic residues corresponds to more flexible parts of the protein structure. We ignored *f5* and *f6* as they did not have any meaningful correlation. Thus factors *f1, f2, f3* and *f4* were found to be crucial and were retained for further study. This step ends with a general picture of our data set proteins in terms of the hydrophobicity distribution along the chains. This picture is represented by mutually independent dimensions (factors) that can in turn be studied for its predictive power.

Linear Discriminant Analysis on both training and test dataset as well as on the entire allostery dataset resulted in classification functions in which the weights assigned to factor *f4* has maximal difference between allosteric and non-allosteric coefficients. From this marked difference we inferred that *f4* plays a major role in the classification of allosteric from non-allosteric groups. This classification proved to be highly significant at Fisher exact test ($P < 0.007$). On the test dataset, our prediction resulted in 82.6% accuracy, 85.7% sensitivity and 77.8% specificity whereas 82.6% accuracy, 85.7% sensitivity and 77.9% specificity was observed on the entire allostery dataset. Figure 4 reports the classification of the entire allostery dataset using the probability estimate of allostery, P(A). This probability estimate is found to separate allosteric and non-allosteric groups of proteins with a mixed phase in between. Allostery is an intrinsic property of any protein (Gunasekaran et al. 2004). Our method of discriminating allosteric and non-allosteric structures makes sense only along a continuous line. As a consequence we could observe the presence of an intermediate 'grey zone' between the 'extremely allosteric' and 'weakly allosteric' sub-sets (Ferreiro et al. 2011). We observed a statistically significant correlation between the probability estimate of

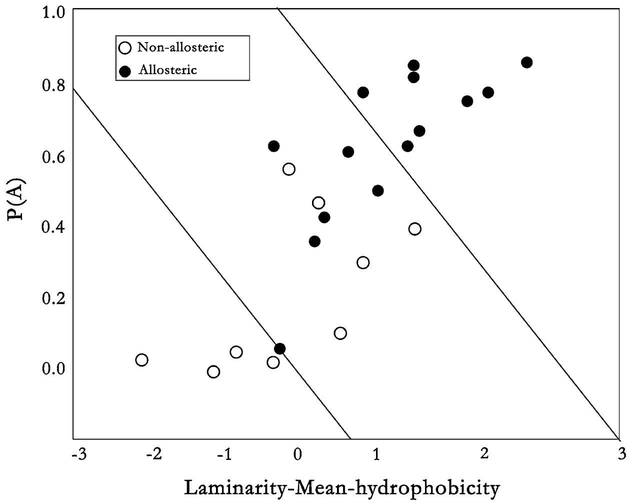

Fig. 4 Classification of allosteric and non-allosteric proteins of the entire allostery dataset using probability estimate

Table 6 Root Mean Square Deviation and probability estimate of allostery of proteins in the allostery dataset

PDB id (A-allosteric, NA-non-allosteric)	Probability estimate of allostery P(A)	Root Mean Square Deviation RMSD
1IRK (A)	0.99	0.09
1KAO (A)	0.72	0.78
1L5Z (A)	0.97	0.63
1T48 (A)	0.74	0.79
1TAG (A)	0.59	0.65
2TRT (A)	0.9	0.73
4Q21 (A)	0.88	0.85
3CHY (A)	0.73	1.01
1XTQ (A)	0.51	0.96
1CD5 (A)	0.94	0.98
1E0S (A)	0.10	1.02
1ERK (A)	0.43	0.09
1NH8 (A)	0.9	1.02
1A6 M (NA)	0.55	0.15
1D2 K (NA)	0.06	0.44
1DG3 (NA)	0.37	1.10
1S0Q (NA)	0.09	0.23
1LMN (NA)	0.03	0.12
1KHI (NA)	0.45	0.30
1GQZ (NA)	0.67	1.04
1IFC (NA)	0.07	0.45
1J9Y (NA)	0.15	0.22

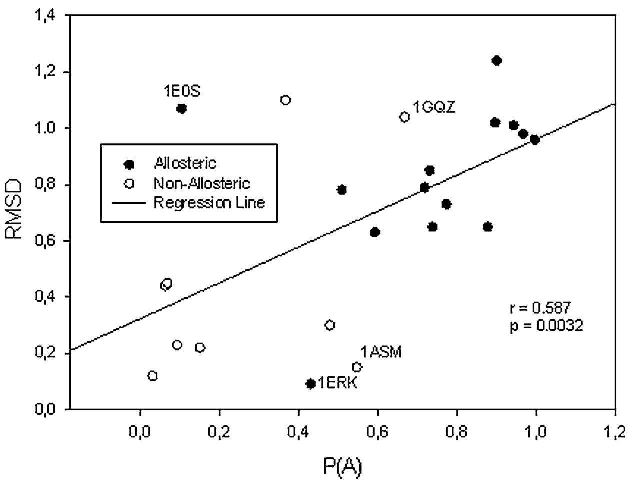

Fig. 5 Graph relating estimated probability of allostery of proteins in the allostery dataset and Root Mean Square Deviation

allostey computed using Linear Discriminant Analysis with Root Mean Square Deviation, as shown in Table 6.

Figure 5 depicts the correlation between the probability estimate of allostery, P(A) and Root Mean Square Deviation. It shows four proteins that were incorrectly predicted by Linear Regression Analysis. They were found to be relatively distant from the regression line between probability estimate of allostery, P(A) and Root Mean Square Deviation, confirming that the model is driven by estimated probability of allostery of proteins.

Discussion

Allostery is an important biological regulatory mechanism in organisms. Majority of the findings related to allostery are found along the classical structure–function paradigm. This has been complemented at the sequence level by identifying residues that participate in allosteric interactions and the evolutionarily conserved residues in allosteric proteins (Lockless et al. 2002). To our best knowledge, the current report is the first study that attempts to extract allosteric behavior from amino acid sequence information alone. Our definition of allostery is based on specific residues that are more involved than others in the allosteric configuration switch. In this study, we used a reductionist approach whereby the primary sequence data of the protein was considered for Recurrence Quantitative Analysis, with hydrophobicity as the key driver of the conformational change exhibited by such proteins. The hydrophobicity index describing the various amino acids was quantified into variables using Recurrence Quantitative Analysis. Combining these models with multivariate statistical analysis produced a model which could significantly distinguish allosteric and non-allosteric proteins. Positive correlations between two Recurrence Quantitative Analysis variables, hydrophobicity mean and hydrophobicity laminarity emerged as the most significant factor in determining allostery. This implies that when the variability is high, the possibility of finding repeated hydrophobic patches (laminarity measure) also increases. Laminarity measure was observed to be correlated to high motion and flexible regions of the proteins (Zbilut et al. 2003). Allostery is intrinsically related to flexibility and allosteric proteins undergo greater conformational changes, thus displaying greater flexibility and regions with high mobility (Gunasekaran et al. 2004). From these findings it was inferred that laminarity is relatively greater in allosteric proteins i.e. the possibility of finding repeated hydrophobic patches is high. Our approach though promising does come with some limitations. Our method is applicable only to single chain proteins (it is not possible to get a unique sequence based representation chemico-physical interactions. Nevertheless, the method shows potential to predict allostery based purely on sequence information. This has application towards synthetic design of proteins. Creating artificial protein switches is an important goal of synthetic biology due to its application towards developing novel biosensors. Our study offers interesting pointers in the direction of designing novel proteins with allosteric properties. However, more work needs to done in order to find linear correlates of allostery at the sequence level.

Acknowledgments This work was funded by Kerala Government grant for the Inter University Centre for Excellence. The authors would like to thank Prof. Anders Liljas for his valuable comments.

References

Berg JM, Tymoczko JL, Stryer L (2002) Biochemistry, 5th edn. WH Freeman, New York

Bruni R, Costantino A, Tritarelli E, Marcantonio C, Ciccozzi Rapicetta M, El Sawaf G, Giuliani A, Ciccaglione AR (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

Colafranceschi M, Colosimo A, Zbilut JP, Uversky VN, Giuliani A (2005) Structure-related statistical singularities along protein sequences: A correlation study. J Chem Inf Model 45:183–189

Daily MD, Gray JJ (2007) Local motions in a benchmark of allosteric proteins. Bioinform Proteins 67:385–399

Dima RI, Thirumalai D (2006) Determination of network of residues that regulate allostery in protein families using sequence analysis. Protein Sci 15:258–268

Eckmann JP, Oliffson Kamphorst S, Ruelle D (1987) Recurrence plots of dynamical systems. Europhys Lett 91:973–977

Ferreiro DU, Hegler JA, Komives EA, Wolynes PG (2011) On the role of frustration in the energy landscapes of allosteric proteins. PNAS published ahead of print January 27, 2011

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

Jernigan MR (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552

Lockless SW, Wall MA, Ranganathan R (2002) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol 10:59–68

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

Porrello A, Soddu S, Zbilut JP, 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 Bioinform 55:743–755

Zbilut JP, Webber CL Jr (1992) Embeddings and delays as derived from quantification of recurrence plots. Phys Lett A 171: 199–203

Zbilut JP, Colosimo A, Conti F, Colafranceschi M, Manetti C, Valerio MC, Webber CL Jr, Giuliani A (2003) Proteiin aggregation/folding: the role of deterministic singularities of sequence hydrophobicity as determined by nonlinear signal analysis of acylphosphatase and Aβ(1–40). Biophysical J 85: 3544–3557

Zbilut JP, Giuliani A, Colosimo A, Mitchell JC, Colafrancesch M, Marwan N, Webber CL, 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