


RESEARCH

Open Access



Understanding the evolutionary trend of intrinsically structural disorders in cancer relevant proteins as probed by Shannon entropy scoring and structure network analysis

Sagnik Sen^{1*} , Ashmita Dey^{1†}, Sourav Chowdhury^{3†}, Ujjwal Maulik¹ and Krishnananda Chattopadhyay²

From 17th International Conference on Bioinformatics (InCoB 2018)
New Delhi, India. 26-28 September 2018

Abstract

Background: Malignant diseases have become a threat for health care system. A panoply of biological processes is involved as the cause of these diseases. In order to unveil the mechanistic details of these diseased states, we analyzed protein families relevant to these diseases.

Results: Our present study pivots around four apparently unrelated cancer types among which two are commonly occurring viz. Prostate Cancer, Breast Cancer and two relatively less frequent viz. Acute Lymphoblastic Leukemia and Lymphoma. Eight protein families were found to have implications for these cancer types. Our results strikingly reveal that some of the proteins with implications in the cancerous cellular states were showing the structural organization disparate from the signature of the family it constitutes. The sequences were further mapped onto respective structures and compared with the entropic profile. The structures reveal that entropic scores were able to reveal the inherent structural bias of these proteins with quantitative precision, otherwise unseen from other analysis. Subsequently, the betweenness centrality scoring of each residue from the structure network models was resorted to explore the changes in dependencies on residue owing to structural disorder.

Conclusion: These observations help to obtain the mechanistic changes resulting from the structural orchestration of protein structures. Finally, the hydrophobicity indexes were obtained to validate the sequence space observations using Shannon entropy and in-turn establishing the compatibility.

Keywords: Shannon entropy scoring, Multiple sequence alignment, Structure network model, Hydrophobicity index

Background

Biological processes are the results of highly orchestrated interactions among the biological macro-molecules. A majority of these processes occurring within the confines of cell cytosol pivot around the coordinated functions of protein molecules. Encoded by the one-dimensional

code script of DNA, the biological relevance of protein molecules reside on their three-dimensional structures. These three-dimensional protein structures arise after a series of sub-structural interactions primarily dictated by sequence information of the protein molecules. In many cases, proteins do not have a stable three-dimensional structures. These proteins are broadly known as Intrinsically Disordered Proteins (IDPs) [1, 2]. IDPs become interesting for the researchers, due to their diverse biological roles and apparent revocation of traditional structure-function paradigm. Regardless of the

*Correspondence: sagnik.sen2008@gmail.com

†Sagnik Sen, Ashmita Dey and Sourav Chowdhury contributed equally to this work.

¹Department of Computer Science and Engineering, Jadavpur University, 700032 Kolkata, India

Full list of author information is available at the end of the article



lack of three-dimensional structure, different biophysical techniques evidenced that IDPs actively participated in various biological processes like control of cell cycle, transcriptional activation, signaling, and they frequently interacted with or functioned as central hubs in protein interaction networks [3].

Proteins are being folded to perform specific functions. Sometimes acquired ordered globular structure may be accompanied by interaction with other proteins. The folding mechanism can be driven by different changes in protein environment. Since proteins are actively involved in different biological processes, a loss of protein structure and disruption in associated interactions can lead to a series of metabolic disruptions in turn inducing a pathological state [4]. A wide range of diseases are caused due to the misfolding of proteins [5]. Misfolding or misfolding function can develop from point mutation or an exposure to internal or external toxins, impaired post-translational modifications (PTMs) [6], an increased probability of degradation, impaired trafficking, oxidative damage or lost binding partners. These factors can act independently or in associations with one another.

Misfolding may cause numerous neurodegenerative and malignant diseases. Reports suggest [7–9], IDPs have an evolutionary significance and correlation with complexity. More elaborately, connection or changes in proteins from most primitive species to modern species can be analyzed depending on the transition from ordered to disordered state or vice versa. The variation in protein residues in protein sequences is responsible for the structural transition which are directly associated with sequence based complexity of the proteins.

Multiple Sequence Alignment (MSA) of a protein family can provide a consensus sequence of that family which might be considered as family sequence representative with the most evolutionarily conserved set of amino acids. As the consensus sequences consist of evolutionarily conserved amino acid residues, so the consensus sequence of a protein family can represent the structural trait for almost all individual members of that protein family. Hence, the complexity score of a consensus in terms of disorder and order can summarize the structural trend of most of the individual proteins from a protein family.

In this article, four diverse cancer types are considered, among which two are well known and frequent malignant diseases viz., Prostate Cancer [10–12], Breast Cancer [13–15], and two relatively less abundant forms viz. Acute Lymphoblastic Leukemia [16, 17] and Lymphoma [17, 18] respectively, along with the proteins responsible for these diseases. Not only the human protein forms are considered, rather the whole family protein sequences are collected in order to compute the MSA and its corresponding consensus sequences. In order to analyze the evolutionary changes, all the sequences of a protein family are

studied in details. The Shannon entropy is calculated for the consensus sequence of responsible proteins and also for each sequence from those protein families. Depending on the entropic scores, the proteins were classified as order or disorder in nature. In order to understand the Shannon entropic impact, the sequences are mapped in their respective structure. The hydrophobic index is calculated for each member of the protein family in order to compare the sequence complexity in terms of entropic scores. Hence the main motivation of this study is to find the general traits of a protein family in terms of structured and unstructuredness applying complexity scoring.

Methods

In this section, we have discussed the proposed framework. Two different databases were used viz., (1) UniProt [19] and (2) Pfam [20]. Initially, eight proteins which were responsible for selected four diverse cancer types, were selected for this study based on frequent occurrence. Later, the sequences of those protein families were considered for further research. In Fig. 1, the flow of the proposed framework is given. The proposed method is discussed below:

Database information

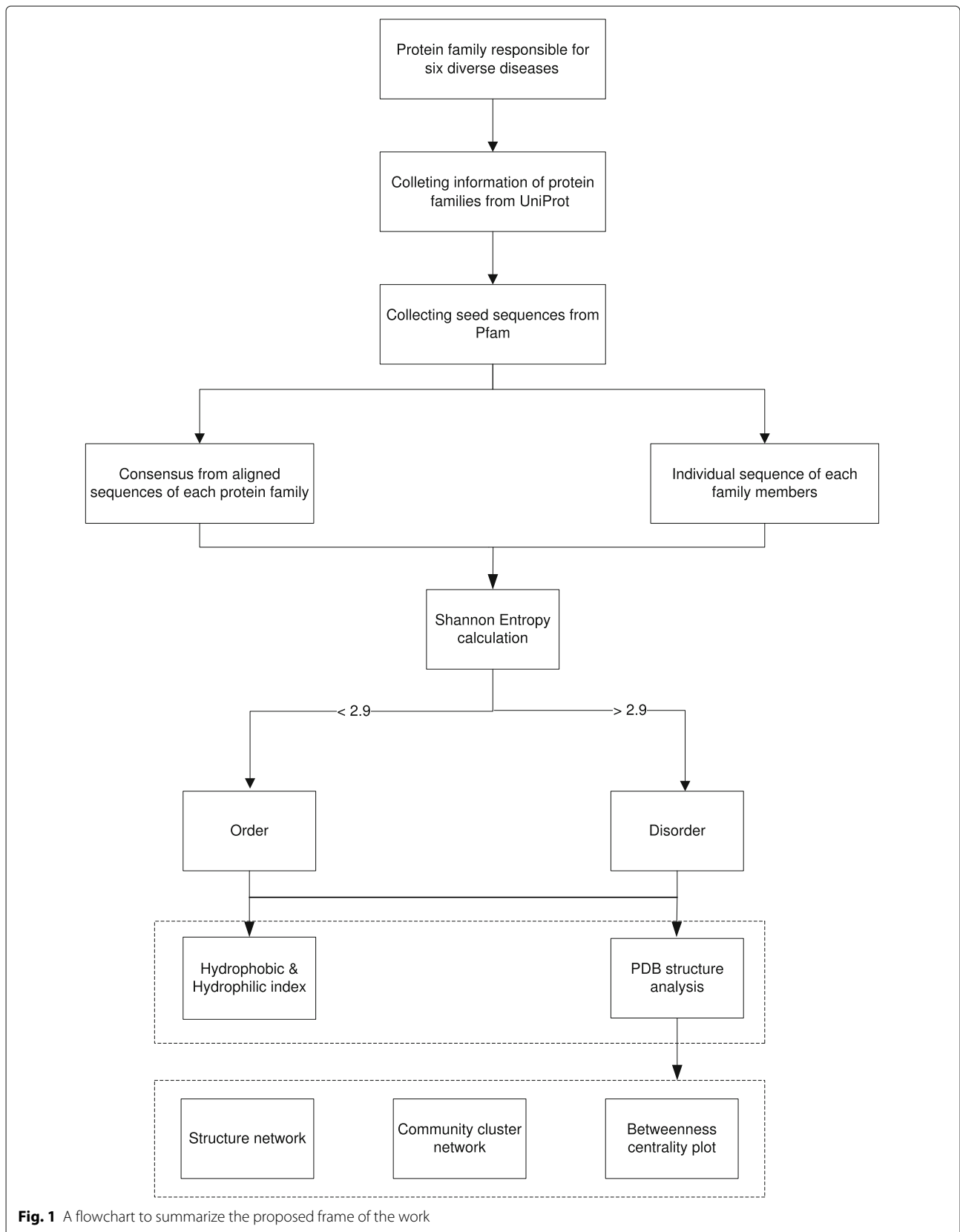
UniProt

UniProt [19] is a cumulative set of sequences and annotated information of these proteins. This database provides around 60 million protein sequences. Since 2014, the database contains around 5631 proteomes. Along with that, the protein family and domain information are described at Uniprot.

Pfam

Pfam [20] is another database, consisting of protein family information, multiple sequence alignment and profile hidden Markov models. More than 3000 protein family information is given.

Four cancer types were selected and their commonly responsible protein families in human, depending on the literature survey such as Heat shock protein beta-1 [21, 22], BAG family molecular [23, 24], Breast cancer type 2 [25], Endophilin-B1 [26, 27], Apoptosis regulator Bcl-2 [28, 29], Calpain-type cysteine [30, 31], Cellular tumor antigen p53 [32, 33] and RNA-binding protein 38 [34, 35] were identified from UniProt database. Heat shock protein beta-1 played a role as a molecular chaperone probably maintaining denatured proteins in a folding-competent state. Similarly, BAG family molecular act as a nucleotide-exchange factor, the breast cancer type 2 susceptibility protein (BRCA2) is a breast tumour suppressor involved in double-strand break repair and/or homologous recombination Endophilin-B1 has been observed to regulate the membrane dynamics of various intracellular



compartments, Apoptosis regulator Bcl-2 regulates cell death by controlling the mitochondrial membrane permeability, Calpain-type cysteine involved in epiderm development, Cellular tumor antigen p53 acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type and RNA-binding protein 38 specifically bind the 3'-UTR of CDKN1A transcripts, leading to maintain the stability of CDKN1A transcripts, thereby acting as a mediator of the p53/TP53 family to regulate CDKN1A. Using the Pfam database the aligned sequences of a particular protein family was identified. It was observed that MSA played an important role in comparative functional and structural analysis of biological sequences. In this regard, seed alignment for FASTA format were selected, which included tree ordering sequences and lower case letters were considered with dashes as a gap characteristics. Furthermore, this also provided a biological insight regarding the relationship between the structural and functional behavior of proteins [36]. Therefore, to analyze the type of protein orchestration (i.e., order and disorder) of protein families aligned sequences were considered. Subsequently, a consensus sequence [37] was constructed from the aligned results using Consensus Maker tool. The MSA of each family were provided to this online tool. This tool computed a consensus using customary parameters. These sequences represented as a logo or signature of that protein family, shown in Additional file 1. In this regard, based on the frequency of amino acid i.e. the highest frequency of amino acid was considered as an entry of building consensus in that particular position. A consensus sequence is a set of amino acids which are evolutionarily conserved in protein family [38]. For the analyses of the protein sequences present in a particular family, it was necessary to understand the structural orchestration of sequences i.e. propensity towards order and disorder. In this regard, the Shannon entropy (SE) score was calculated for each consensus sequence. As it was evidenced that entropy posses an idea of disorder. Entropy was directly proportional to the rate of disorder i.e. if the disorder increases it signifies higher entropy. Shannon entropy was defined as follows:

$$SE(i) = - \sum_{i=1}^N P_i \log_2 P_i \quad (1)$$

where P_i was the probability of given amino acids and N was the number of letters in a sequence. The summation run over the 20 residues that normally were present in a protein sequence. The probability P_i represent the composition of the consensus sequence. So the entropy range lied between 0 and the $\log_2(20) = 4.32$. If the Shannon entropy score of consensus sequences was less than 2.9 then it signify that, particular protein family was

ordered [39], on the other hand, if the Shannon entropy was very high then that protein family was disorder and have an important impact on the cause of diseases. The Shannon entropy of each sequence of each family was also calculated in order to validate the results, reported in Additional files 2-9. Moreover, one sample t-test was performed on each protein of a particular family, in order to understand the sample mean which was statistically different from a known or hypothesized population mean. Statistical significance is determined by looking at the p -value. The p -value gives the probability of observing the test results under the null hypothesis. The lower the p -value, the lower the probability of obtaining a result like the one that was observed if the null hypothesis was true. The One Sample t Test is a parametric test. It is defined as:

$$t = \frac{\bar{a} - \mu}{\sigma} \quad (2)$$

Here, \bar{a} is the sample mean of entropy scores, σ is the standard deviation of list entropy scores of a family. μ is the specified population mean of list of entropy scores of a family. where,

$$\sigma = \sqrt{\frac{S^2}{n}} \quad (3)$$

S^2 is the sample variance, n is the sample size which is total number of proteins from a family. Furthermore, to understand this observation we tried to find the structure of these sequences [40]. These models were further analyzed for structure network analysis. However, complex systems have been analyzed with a help of network models, the interaction between the components of the machines were described through nodes and edges. Generally, secondary structure and folding arrangement mechanism were utilized to understand the protein structures. Another promising method for analysis of the protein structure was through the network [41]. In this network model the amino acid residues represented as nodes and edges which represent the interaction among them, the interaction was established based on the interaction energy or spatial distance. Interactions usually have a weight, which characterized their strength. Depending on this strength the edges were drawn between the two amino acid nodes. The equation was described below.

$$F_{ij} = \left[\frac{x_{ij}}{\sqrt{(X_i * X_j)}} \right] * 100 \geq F_c \quad (4)$$

F_c was the threshold of interaction strength, the default value is 4%. Here, x_{ij} was the number of side chain atom

pairs of residues i and j . X_i and X_j were the normalization factor for residues types i and j [42, 43].

In this paper, depending on the normal mode analysis (NMA) a correlation matrix was obtained in order to perform a cross-correlation matrix. Then by means of correlation network analysis, we generated structure networks [41] of different protein depending on their tertiary structure. The weight of the connection of nodes represented the value of cross-correlation respectively. By means of correlation network analysis, a full residue network was generated and it was split into a highly correlated coarse-grained community cluster network by using Girvan-Newman [44] clustering method where the highly interacting residues were clumped together in the clusters. Here some lower value elements in the raw correlation matrix from NMA were excluded because of being lower than the cutoff value 0.3.

The role of a particular node as a connector between other nodes viz., the importance of a residue to a network in its functioning as a bridging point can be manifested by measuring the number of shortest paths passing through that particular node. Betweenness centrality characterizes the regions of a protein that show differences in coupled motions between networks. Residues having significant contribution to intrinsic dynamics of the protein show high centrality value. Also depending on the centrality scoring, Euclidean distances among the protein mutant types were calculated and the subsequent hierarchical cluster was generated. The betweenness centrality was performed to find the bottlenecks in communication networks and community detection whereas the NMA was performed to generate structure networks of different protein depending on their tertiary structure.

Results

Reports suggested that selected eight proteins (Endophilin-B1 [26], Breast cancer type 2 susceptibility protein [25], Heat shock protein beta-1 [21], BAG family molecular chaperone regulator 1 [23], Apoptosis regulator Bcl-2 [28], Calpain-type cysteine protease DEK1 [30], Cellular tumor antigen p53 and RNA-binding protein 38 [34]) were associated with four malignant diseases. These proteins were found to have intimate connection with metabolic cascades and interaction networks leading to cancer states. We referred to protein families of these selected proteins so as to understand the generic structural propensity of the protein families which these proteins constitute. To understand the generic structural trend of these protein families, we had performed the Shannon entropy of the consensus sequences. Depending on the entropic score of the consensus sequences, the

protein families were being classified as order or disorder. In Table 1, the entropy score of the consensus sequences of each protein families was reported along with the score of t-test [45]. In most of the cases, the protein sequences from a disorder or order class were expected to be disordered or ordered. However, few sequences were reported to be disordered being a part of an ordered protein family in terms of entropic scoring of the consensus protein sequence and vice versa. It is to be noted that, Endophilin-B1 is a responsible protein for breast cancer disease and the entropic score is 2.87. We reported this protein as disorder even though the entropic score is not high but the structural, as well as functional domain, provides an evidence to this. From Protein Data Bank (PDB), we found that the Solution structure of the SH3 domain of Endophilin B1 has higher loops and turns than the number of Beta-coils (PDB Id:1X43). Moreover, the literature studies [26, 27] show the nature of the dynamic functionality of this protein. These evidences supported the disorderedness of this protein as well as our finding.

We observed the transition points of transformation and considered those sequences for further analysis. In Fig. 2, a representative of each protein families was identified where the sudden deviation from the entropic score of consensus sequences occurred and reported their Shannon entropic changes along with the structures. Also, other two sequences of the protein families were mapped in their respective structures. From these structures, the deviation of structural changes along the entropic scores was easily visualized. Figure 2a and b represented the two proteins responsible for breast cancer. Similarly, Fig. 2c-h represented selected proteins responsible for prostate cancer, acute lymphoblastic leukemia and lymphoma respectively. The hydropathy index of those sequences was analyzed and validated with respect to entropic score.

To provide a comprehensive understanding of mentioned changes, the PDB structures of those particular sequences were also observed. Depending on the PDB, structure networks were shown along with the community cluster network and betweenness centrality plot. In Figs. 3-8, the structure network analysis, Community cluster network and betweenness centrality plots were shown for four diverse cancer types. Though we had performed the analysis of multiple proteins from multiple families, most diverse samples were shown in this article. From structure network analysis, the dependencies on residues at different secondary structural orchestrations could be observed. From the experimental outcomes, ordered structures had a diverse set of community clusters based on conservation of residue-residue interaction than disordered structure and also betweenness centrality graph was well distributed than disordered structures. Each

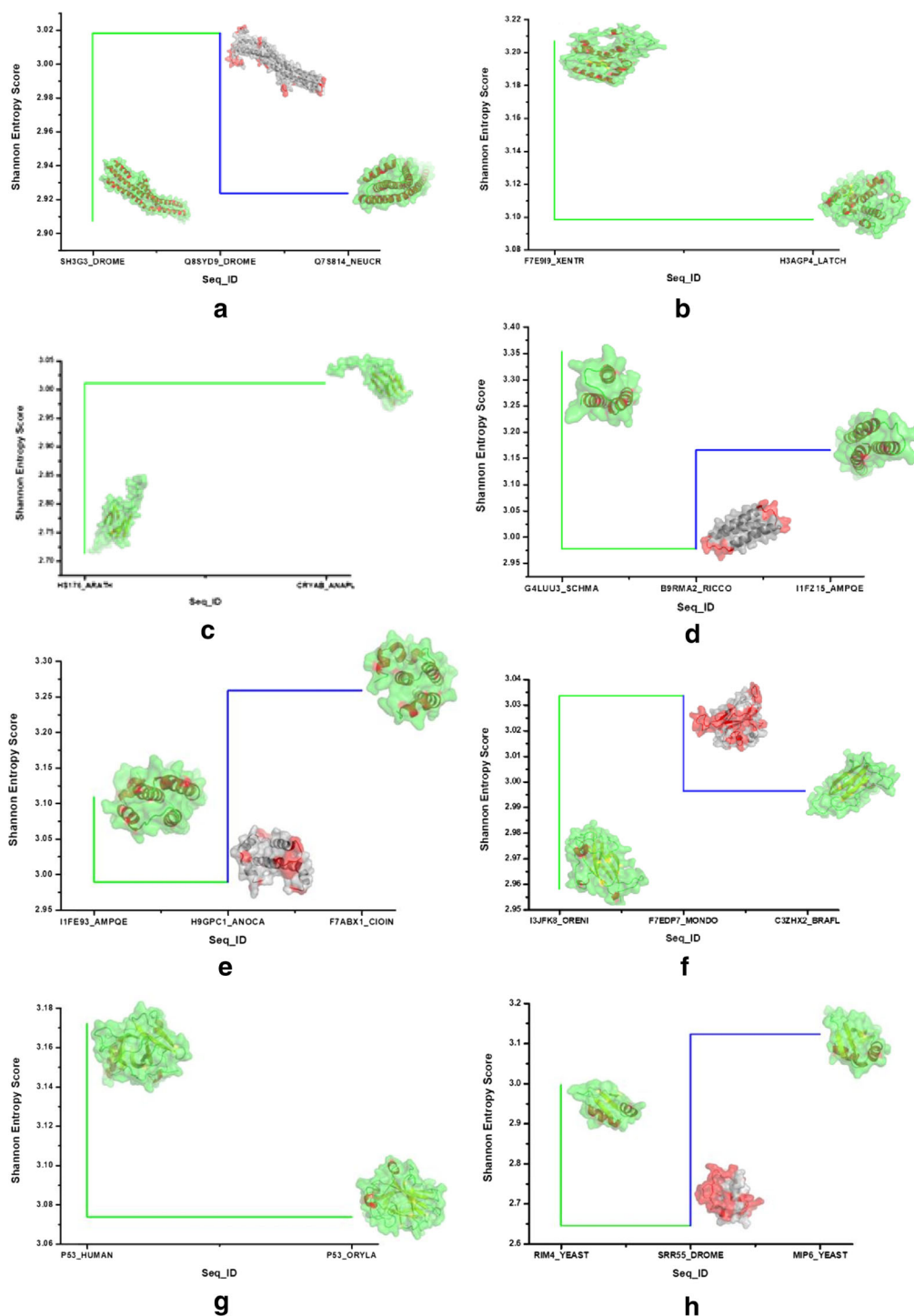


Fig. 2 Representing the change in evolutionary trend in **a.** Endophilin-B1, **b.** Breast cancer type 2 susceptibility protein, **c.** Heat shock protein beta-1, **d.** BAG family molecular chaperone regulator 1, **e.** Apoptosis regulator Bcl-2, **f.** Calpain-type cysteine protease DEK1, **g.** Cellular tumor antigen p53 and **h.** RNA-binding protein 38 proteins responsible for four cancer types such as breast cancer, prostate cancer, acute lymphoblastic leukemia and lymphoma respectively

directly coupled pair, obtained from structure space analysis, were found to be situated either in common cluster or in two densely connected clusters. The betweenness

centrality was calculated to unveil the influence of a particular node on the internal dynamics of different structure.

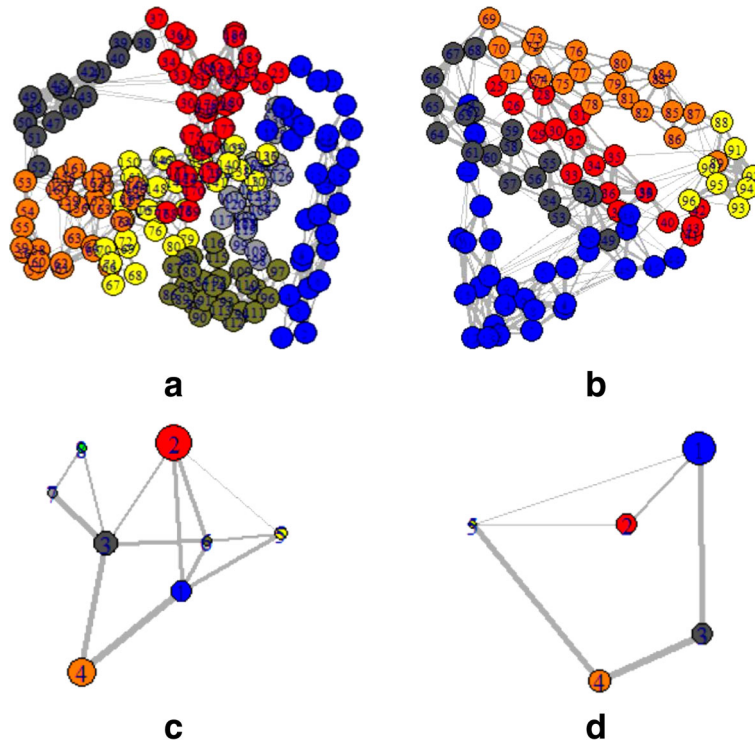
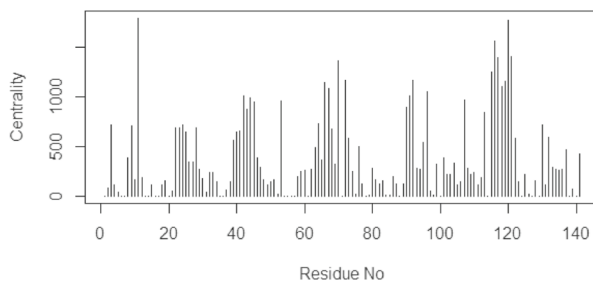
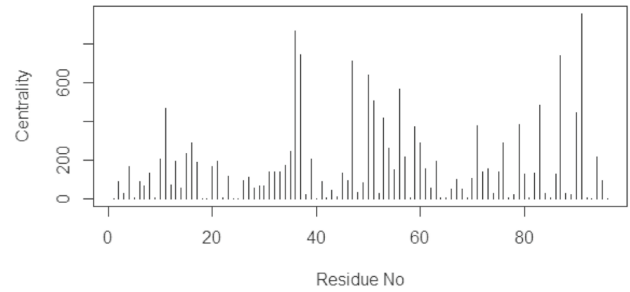


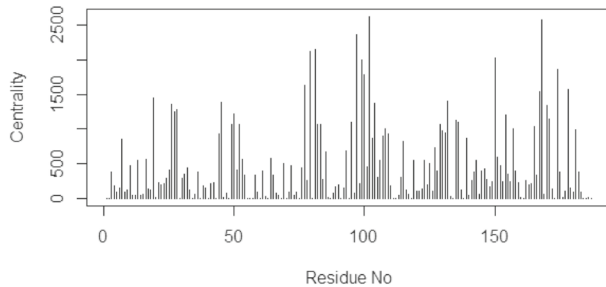
Fig. 3 Structure network of Acute Lymphoblastic Leukemia for two spices such as **a.** *Monodelphis domestica* and **b.** *Anolis carolinensis* which represent high and low entropic score respectively. Similarly, the Community cluster network of **c.** *Monodelphis domestica* and **d.** *Anolis carolinensis* are shown



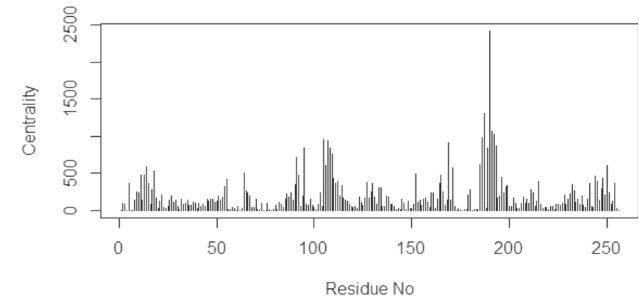
a



b



c



d

Fig. 4 Betweenness centrality plot of Acute Lymphoblastic Leukemia for **a.** *Monodelphis domestica* and **b.** *Anolis carolinensis* which represent high and low entropic score respectively. Similarly, the plot is shown for two spices affected by breast cancer such as **c.** *Xenopus tropicalis* and **d.** *Drosophila melanogaster*

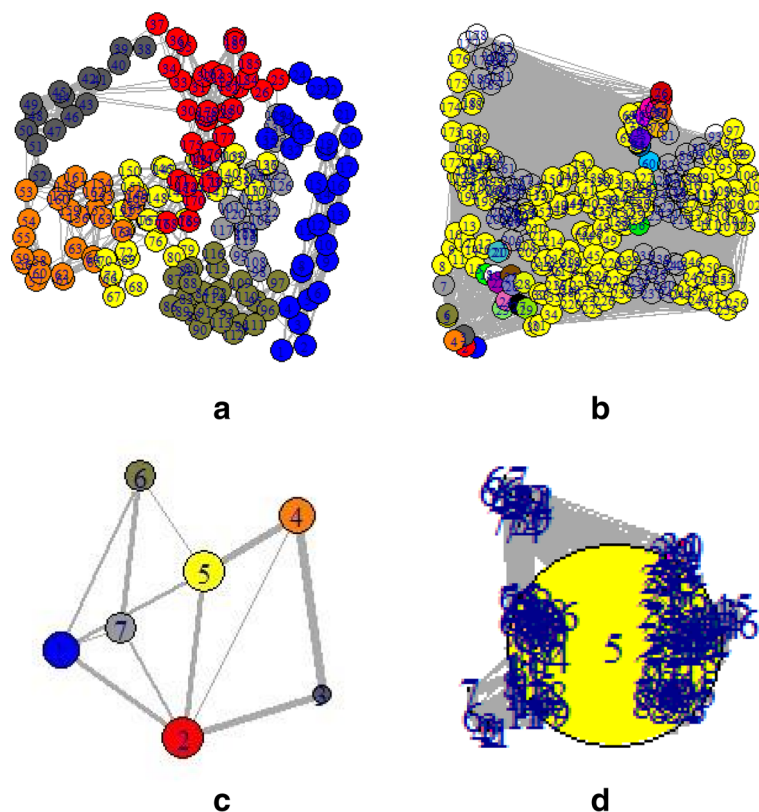


Fig. 5 Structure network of Breast Cancer for two species such as **a.** *Xenopus tropicalis* and **b.** *Drosophila melanogaster* which represent high and low entropic score respectively. Similarly, the Community cluster network of **c.** *Xenopus tropicalis* and **d.** *Drosophila melanogaster* are shown

Discussion

In case of sequence biology, scoring of hydropathic index described the complexity of protein primary structure. It helped to understand the propensity of protein in terms of structural order or disorder. As mentioned before, each of the structurally affected human proteins and their family were chosen from four diverse cancer types. Observing the trend of Shannon entropic score of the consensus sequence of each family, structural transformation of the proteins could be considered as the reason behind diseased conditions. In Table 1, the Shannon entropic scores of all selected families were reported. Among them, only breast cancer and leukemia have one of each ordered family propensity. This fact justified the number of conserved structural motif of the family had ordered propensity. Hence most of the proteins of the family were ordered. Subsequently, the proteins with disordered propensities in consensus sequence scores were showing the disordered trend. Statistically, the entropic scores for consensus were significant. In Table 2, entropic scores of the human proteins were compared with their average hydrophobic index. Mostly proteins with disordered entropic scores were showing compatibility with

average hydrophobic index. Hydrophobic index was justifying the spontaneous folding capacity of the protein. So lower hydrophobic index was indicating towards higher disordered propensity. Hence the compatibility between two different scoring systems could be clearly observed from Table 2.

Thereafter, the sequence specific information was compared with the three-dimensional structures of proteins. In Fig. 3, two proteins of the leukemia were shown in terms of Gaussian network model based structure networks and their highly conserved community clusters. Similarly, in Figs. 5, 6 and 8, the structure network and community clusters were given for breast cancer, lymphoma and prostate cancer respectively. Depending on the number of shortest path on each residue, betweenness centrality plots were given in Figs. 4 and 7. From Fig. 4, the distribution of residual dependencies in terms of betweenness centrality for leukemia and breast cancer were given. The residual for the random leukemia protein sample in Fig. 4a, has justified the entropic score of its family. In Fig. 4a, the residual distribution was highly conserved at certain residual points whereas, in Fig. 4b, dependencies in terms of centrality scores were well

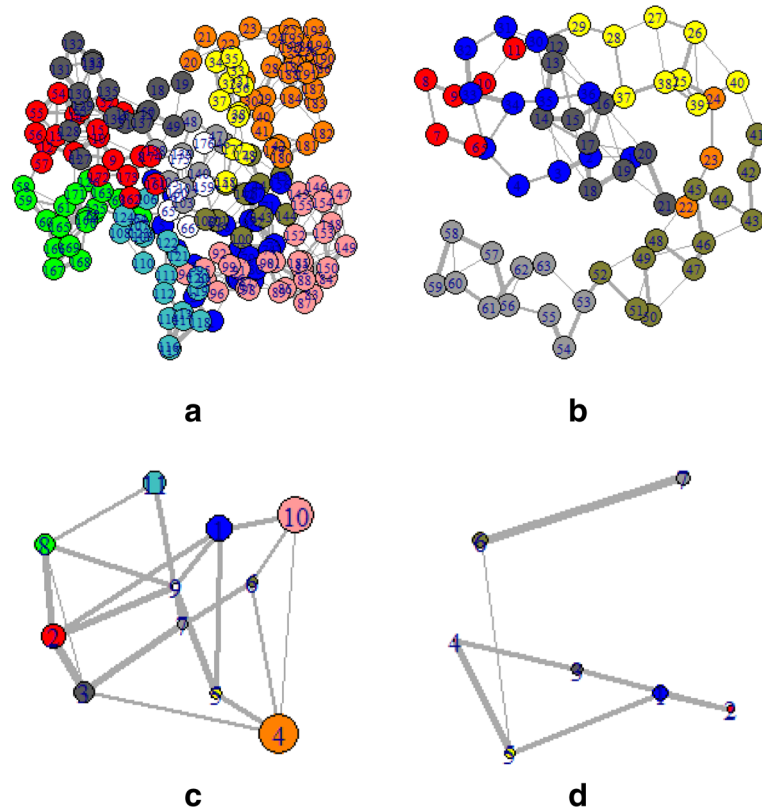


Fig. 6 Structure network of Lymphoma for two spices such as Lymphoma for **a.** Homo sapiens and **b.** Drosophila melanogaste which represent high and low entropic score respectively. Similarly, the Community cluster network of **c.** Homo sapiens and **d.** Drosophila melanogaste are shown

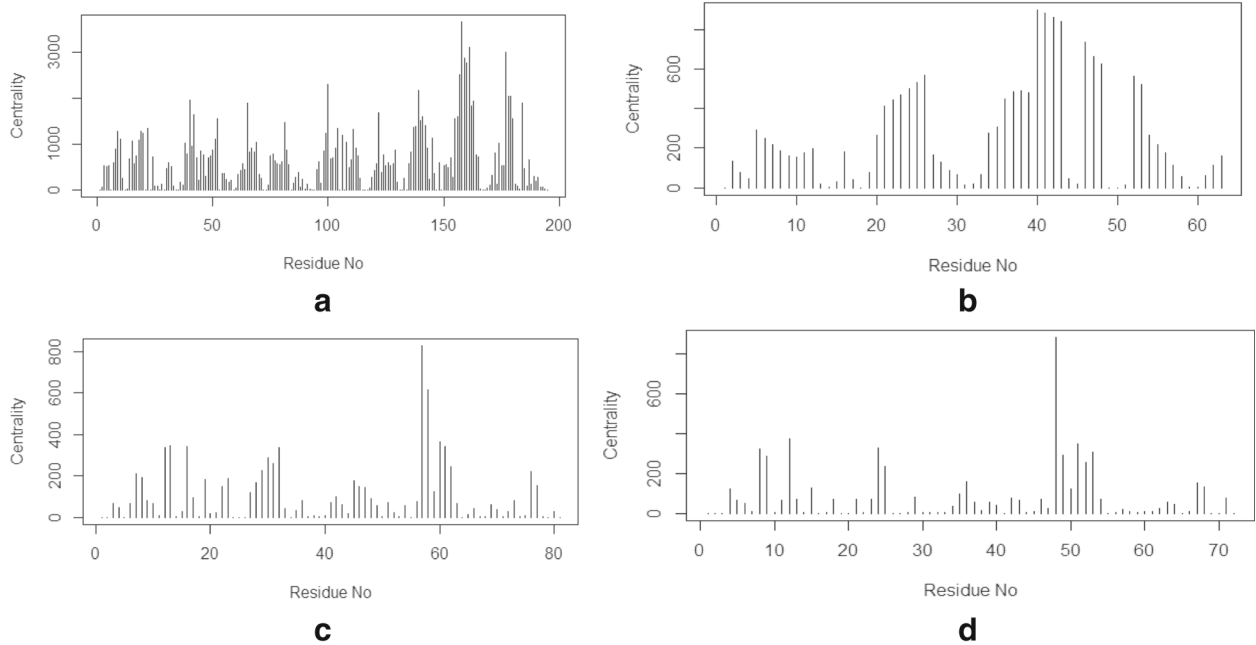


Fig. 7 Structure network of Lymphoma for two spices such as Lymphoma for **a.** Homo sapiens and **b.** Drosophila melanogaste which represent high and low entropic score respectively. Similarly, the Community cluster network of **c.** Homo sapiens and **d.** Drosophila melanogaste are shown

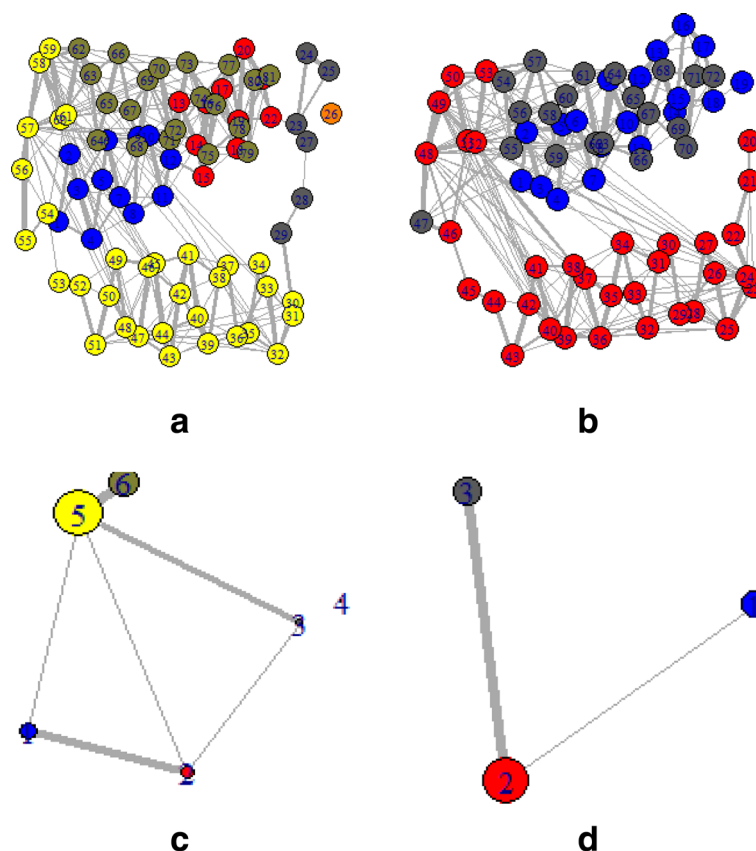


Fig. 8 Structure network of prostate cancer for **a.** *Aspergillus niger* and **b.** *Ricinus communis* which represent high and low entropic score respectively. Similarly, the Community cluster network of **c.** *Aspergillus niger* and **d.** *Ricinus communis* are shown

Table 1 The Shannon entropy and t-test value of four selected diverse cancer types

Disease	Protein family	Shannon entropy	Significant score	p-value
Prostate cancer	Heat shock protein beta-1	3.36	True	7.3e-36
	BAG family molecular	3.39	True	3.4e-40
Breast cancer	Breast cancer type 2	4.01	True	1.0e-30
	Endophilin-B1	2.87	True	5.8e-38
Acute lymphoblastic leukemia	Apoptosis regulator	3.34	True	1.9e-27
	Bcl-2	2.79	True	3.8e-25
Lymphoma	Calpain-type cysteine	4.01	True	1.0e-07
	Cellular tumor antigen p53 RNA-binding protein 38	3.38	True	3.8e-96

distributed throughout the sequence. Though the family of Fig. 4b was maintaining ordered trend this particular protein structure was showing a disordered trend in terms of individual entropic score. That is why few higher peaks in the plot have been seen. Similarly, in Fig. 4c and d, the centrality plotting for breast cancer samples were given which were following the similar trend like leukemia samples. In Fig. 7, centrality distribution for lymphoma and prostate cancer were shown. Comparing with sequence information, the family with higher disordered propensity was showing conservation at certain

Table 2 The hydrophobic index value of the selected proteins for four cancer types

Disease	Uniprot Id	Entropic score	Average score of hydrophobic index
Prostate cancer	P04792	2.96	0.60
Breast cancer	P38398	3.16	0.41
Acute lymphoblastic leukemia	Q8RVL1	2.94	0.58
Lymphoma	P04637	3.17	0.37
	Q9H0Z9	3.03	0.55

residual points even in the ordered samples of the family. Hence the path of evolutionary transformations of the proteins from the family could be described from these observations. In human samples, the sudden structural changes were following the common mentioned path of transformations by disrupting the amount of average hydrophobic amino acids. Again the spontaneous folding capacity of the structure could be affected.

Conclusion

In this article, we have proposed a method based on sequence complexity calculation of each protein families using Shannon entropic scoring for different malignancies. For four different cancer types viz., prostate cancer, lymphoma, acute lymphoblastic leukemia and breast cancer, eight different protein families were selected which structurally involved with the diseases. The objective was to observe the structural transformation of proteins in an evolutionary timespan. It was successfully shown that the entropic scoring based on amino acid distributions in the sequence helped to understand structured or unstructured propensity of proteins and their families. The results, obtained from entropic studies were complemented by hydrophobic indexing of the sequences. To map the sequence on structure, a structure space analysis was also performed. For each structure, the changes in residual dependencies were observed based on variation in betweenness centrality. Distribution of centrality for the structures were showing a compatible pattern with sequence dependent information. More precisely, structural orchestrations of proteins were varying with entropic scores accordingly. Finally, the experimental outcomes and comparative analyses suggested the evolutionary path of transformation in protein structures which could be comprehended by theoretical entropic scoring based on the conserved residual distribution in protein sequences.

Additional files

Additional file 1: The HMM logo or signature of the selected protein families. (PDF 479 kb)

Additional file 2: Shannon entropy score of proteins under Endophilin-B1 family responsible for Breast cancer. (XLSX 10 kb)

Additional file 3: Shannon entropy score of proteins under Breast cancer type 2 susceptibility protein family responsible for Breast cancer. (XLSX 10 kb)

Additional file 4: Shannon entropy score of proteins under Heat shock protein beta-1 protein family responsible for Prostate cancer. (XLSX 10 kb)

Additional file 5: Shannon entropy score of proteins under BAG family molecular chaperone regulator 1 protein family responsible for Prostate cancer. (XLSX 16 kb)

Additional file 6: Shannon entropy score of proteins under Apoptosis regulator Bcl-2 protein family responsible for acute lymphoblastic leukemia. (XLSX 15 kb)

Additional file 7: Shannon entropy score of proteins under Calpain-type cysteine protease DEK1 protein family responsible for acute lymphoblastic leukemia. (XLSX 15 kb)

Additional file 8: Shannon entropy score of proteins under Cellular tumor antigen p53 protein family responsible for lymphoma. (XLSX 9 kb)

Additional file 9: Shannon entropy score of proteins under RNA-binding protein 38 protein family responsible for lymphoma. (XLSX 11 kb)

Abbreviations

DNA: Deoxyribonucleic acid; IDPs: Intrinsically disordered proteins; MSA: Multiple sequence alignment; NMA: Normal mode analysis; PDB: Protein data bank; SE: Shannon entropy

Acknowledgments

The work of SS is supported by DST-INSPIRE fellowship. The work of AD is supported by the of UGC-UPE-II.SC thanks UGC for the senior research fellowship for the work performed at CSIR-Indian Institute of Chemical Biology. Most importantly we thank the reviewers for their valuable comments and suggestions which help us to improve the paper.

Funding

Not applicable.

Availability of data and materials

All data generated or analyzed during this study are either taken from publicly available databases or included in this article.

Declarations

Not applicable.

About this supplement

This article has been published as part of *BMC Bioinformatics, Volume 19 Supplement 13, 2018: 17th International Conference on Bioinformatics (InCoB 2018): bioinformatics*. The full contents of the supplement are available at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-13>.

Authors' contributions

SS and SC have conceptualized the paper. AD performs the experiments. SS, SC and AD have scripted the manuscript. UM and KC have corrected and edited the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

We used publicly available data for human and animal samples and cell line studies. No human and animals are directly involved.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Computer Science and Engineering, Jadavpur University, 700032 Kolkata, India. ³CSIR-Indian Institute of Chemical Biology, Raja S.C. Mullick Road, 700032 Kolkata, India. ²Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, 02138 Massachusetts, USA.

Received: 29 October 2018 Accepted: 30 November 2018

Published: 4 February 2019

References

1. Sugase DHJK, Wright PE. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*. 2002;447:1021–5.

2. Dunker AK, et al. What's in a name? why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins*. 2013;1(1):24157.
3. Babu MM. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem Soc Trans*. 2016;44(5):1185–200.
4. Luhesli LM, Dobson CM. Bridging the gap: From protein misfolding to protein misfolding diseases. *FEBS Lett*. 2003;583(16):2581–6.
5. Shmygelska A, Hoos HH. An adaptive bin framework search method for a beta-sheet protein homopolymer model. *BMC Bioinforma*. 2007;8(1):136.
6. Mann M, Jensen ON. Proteomic analysis of post-translational modifications. *Nat Biotechnol*. 2003;21:255–61.
7. Das S, Pal U, Das S, Bagga K, Roy A, A M, C MN. Sequence complexity of amyloidogenic regions in intrinsically disordered human proteins. *PLoS ONE*. 2014;9(3):89781.
8. Brown CJ, Johnson AK, Dunker AK, W DG. Evolution and disorder. *Curr Opin Struct Biol*. 2011;21(3):2441–6.
9. Forman-Kay JD, Mittag T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure*. 2013;21(9):1492–9.
10. Morlando M, Pelullo CP, Di Giuseppe G. Prostate cancer screening: Knowledge, attitudes and practices in a sample of men in Italy. a survey. *PLoS ONE*. 2017;12(10):0186332.
11. Molla MS, Katti DR, Katti KS. In vitro design of mesenchymal to epithelial transition of prostate cancer metastasis using 3d nanoclay bone-mimetic scaffolds. *J Tissue Eng Regen Med*. 2018;12(3):727–37.
12. Berg KD, Thomsen FB, Mikkelsen MK, Ingimarsdottir IJ, Hansen RB, Kejs AM, Brasso K. Improved survival for patients with de novo metastatic prostate cancer in the last 20 years. *Eur J Cancer*. 2017;72:20–7.
13. Kyriakopoulou K, Kefali E, Piperigkou Z, Bassiony H, Karamanos NK. Advances in targeting epidermal growth factor receptor signaling pathway in mammary cancer. *Cell Signal*. 2018;51:99–109.
14. Kurozumi S, Yamaguchi Y, Kurozumi M, Ohira M, Matsumoto H, Horiguchi J. Recent trends in microRNA research into breast cancer with particular focus on the associations between microRNAs and intrinsic subtypes. *J Hum Genet*. 2017;62:15–24.
15. McGuire A, Brown JAL, Kerin MJ. Metastatic breast cancer: the potential of miRNA for diagnosis and treatment monitoring. *Cancer Metastasis Rev*. 2015;34:145–55.
16. Yue ZX, Gao RQ, Gao C, Liu SG, Zhao XX, Xing TY, Niu J, Li ZG, Zheng HY, Ding W. The prognostic potential of coilin in association with p27 expression in pediatric acute lymphoblastic leukemia for disease relapse. *Cancer Cell Int*. 2018;18(105).
17. Siegel SE, Advani A, Seibel N, Muffly L, Stock W, Luger S, Freyer DR, Douer D, Johnson RH, DeAngelo DJ, Hayes, Coccia PF, Bleyer A. The prognostic potential of coilin in association with p27 expression in pediatric acute lymphoblastic leukemia for disease relapse. *Cancer Cell Int*. 2018;18(105).
18. Shwang HS, Yoon DH, Hong JY, Park CS, Lee YS, Ko YH, Kim SJ, Kim WS, Suh C, Huh J. The cell-of-origin classification of diffuse large B cell lymphoma in a Korean population by the lymph2cx assay and its correlation with immunohistochemical algorithms. *Ann Hematol*. 2018;97(12):2363–72.
19. UniProt C. Uniprot: A hub for protein information. *Nucleic Acids Res*. 2015;43:204–12.
20. Sickmeier M, Hamilton JA, LeGall T, Vacic V, S CM, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. Disprot: the database of disordered proteins. *Nucleic Acids Res*. 2007;35:786–93.
21. Tang D, et al. Expression of heat shock proteins and heat shock protein messenger ribonucleic acid in human prostate carcinoma in vitro and in tumors in vivo. *Cell Stress Chaperones*. 2005;10(1):46–58.
22. Calderwood SK, Gong J. Heat shock proteins promote cancer: It's a protection racket. *Trends Biochem Sci*. 2016;41(4):311–23.
23. Bruchmann A, Roller C, Walther TV, Schäfer G, Lehmusvaara S, Visakorpi T, Klocker H, Cato AC, Maddalo D. Bcl-2 associated athanogene 5 (bag5) is overexpressed in prostate cancer and inhibits er-stress induced apoptosis. *BMC Cancer*. 2013;13:96.
24. So A, Hadaschik B, Sowery R, Gleave M. The role of stress proteins in prostate cancer. *Curr Genom*. 2007;8(4):252–61.
25. Islam MN, Paquet N, Fox D, Dray E, Zheng XF, Klein H, Sung P, Wang W. A variant of the breast cancer type 2 susceptibility protein (brca) repeat is essential for the recq1 helicase to interact with rad51 recombinase for genome stabilization. *J Biol Chem*. 2012;287(27):23808–18.
26. Karbowski M, Jeong SY, Youle RJ. Endophilin b1 is required for the maintenance of mitochondrial morphology. *J Cell Biol*. 2007;166(7):1027–39.
27. Li J, Barylko B, Eichorst JP, Mueller JD, Albanesi JP, Chen Y. Association of endophilin b1 with cytoplasmic vesicles. *Biophys J*. 2016;111(3):565–76.
28. Findley HW, Gu L, Yeager AM, Zhou M. Expression and regulation of bcl-2, bcl-xl, and bax correlate with p53 status and sensitivity to apoptosis in childhood acute lymphoblastic leukemia. *Blood*. 1997;89(8):2986–93.
29. Kang MH, Reynolds CP. Bcl-2 inhibitors: targeting mitochondrial apoptotic pathways in cancer therapy. *Clin Cancer Res*. 2009;15(4):1126–32.
30. Zhu DM, Uckun FM. Calpain inhibitor ii induces caspase-dependent apoptosis in human acute lymphoblastic leukemia and non-hodgkin's lymphoma cells as well as some solid tumor cells. *Clin Cancer Res*. 2000;6(6):2456–63.
31. Mikosik A, et al. Increased μ – calpain activity in blasts of common b-precursor childhood acute lymphoblastic leukemia correlates with their lower susceptibility to apoptosis. *PLoS ONE*. 2015;10(8):0136615.
32. Gaidano G, Ballerini P, Gong JZ, Inghirami G, Neri A, Newcomb EW, Magrath IT, Knowles DM, Dalla-Favera R. p53 mutations in human lymphoid malignancies: association with burkitt lymphoma and chronic lymphocytic leukemia. *Proc Natl Acad Sci*. 1991;88(12):5413–7.
33. Xu-Monette ZY, et al. Dysfunction of the tp53 tumor suppressor gene in lymphoid malignancies. *Blood*. 2012;119(16):3668–83.
34. Bracken AP, Helin K. Polycomb group proteins: navigators of lineage pathways led astray in cancer. *Nat Rev Cancer*. 2009;9:773–84.
35. Pereira B, Billaud M, Almeida R. Rna-binding proteins in cancer: Old players and new actors. *Trends Cancer*. 2017;3(7):506–28.
36. Redfern OC, Dessailly B, Orengo CA. Exploring the structure and function paradigm. *Curr Opin Struct Biol*. 2008;18(3):394–402.
37. Schneider TD. Consensus sequence zen. *Appl Bioinforma*. 2002;1(3):111–9.
38. Crooks GE, Hon G, Chandonia JM, Brenner SE. Weblogo: A sequence logo generator. *Genome Res*. 2004;14:1188–90.
39. Romero P, Obradovic Z, Li X, Garner E, Brown C, Dunker A. Sequence complexity of disordered protein. *Proteins*. 2000;42(1):38–48.
40. Kallberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J. Template-based protein structure modeling using the raptorx web server. *Nat Protoc*. 2012;7:1511–22.
41. Chakrabarty B, Parekh N. Naps: Network analysis of protein structures. *Nucleic Acids Res*. 2016;44(W1):375–82.
42. Brinda KV, Vishveshwara S. A network representation of protein structures: implications for protein stability. *Biophys J*. 2005;89(6):4159–70.
43. Kannan N, Vishveshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. *J Molecular Biol*. 1999;292:441–64.
44. Girvan M, Newman MEJ. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*. 2002;99:7821–6.
45. O'Mahony M. *Sensory Evaluation of Food: Statistical Methods and Procedures*; 1986.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

