Neurocomputing 145 (2014) 62–67

Contents lists available at ScienceDirect

Neurocomputing

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

Discovering functional patterns from pattern signatures of P53 protein sequence association structure

David K.Y. Chiu*, Ramya Manjunath

School of Computer Science, University of Guelph, Guelph, Ontario, Canada N1G2W1

ARTICLE INFO

Article history:
Received 16 December 2013
Received in revised form
10 June 2014
Accepted 14 June 2014
Available online 24 June 2014

Keywords:
Association network
Pattern signature
Statistical granular computing
Nanostructure analysis
p53 tumor suppressor protein

ABSTRACT

The relationship connecting the biomolecular sequence, the molecular structure, and the biological function is of extreme importance in nanostructure analysis of a protein. Previous studies involving multiple sequence alignment of biomolecules have demonstrated that associated sites are indicative of the structural and functional characteristics of biomolecules, comparable to methods such as consensus sequences analysis. In this paper, an association network structure is constructed from detected significant associated sites in aligned p53 sequence ensemble. From the structure, pattern signatures are measured. These signatures are then compared to selected functionality of the p53 proteins. The results indicate that the extracted site patterns are significantly associated with some known properties of p53, a tumor suppressor. Furthermore, when the sites are aligned with p63 and p73, the homologs of p53 without the same cancer suppressing property, using the common domains, the sites significantly discriminate between the human sequences of the p53 family. Therefore, the study confirms the importance of these detected sites that may indicate their differences in cancer suppressing property.

1. Introduction

Biological sequences when aligned can provide the common and discriminatory information about the individual residue of the biomolecule family. It can also provide information from which knowledge can be extracted that directs us towards the common functional sites of the molecule. Identifying the relationships between the sequences and their relationship to structure and biological functionality is an active area of research (for examples, see [6–8]). Identifying the sequence patterns that infer the functional characteristics of the biomolecule is vital in nanotechnology that represents the biomolecule as a nanostructure such as drug discovery [16].

Previous studies involving multiple sequence alignment of related species have shown that associated patterns of the sequences can reflect structural and functional characteristics of the biomolecule [3,7,9,10,12]. In an aligned sequence ensemble of proteins, associated sites represent sites with amino acid pairs significantly observed together. Two types of associations can be considered, the association between two sites (such as, say *X* and *Y* sites) and the association among multiple sites (such as W associated with *X*, *Y*, and *Z* sites). In this paper, a new method is proposed as a novel association structure generated from the aligned sequence ensemble of the biomolecule's family. The associations in the structure are evaluated using different levels, using insights from granular computing [23,24]. It involves the association testing of different sizes of a two-dimensional contingency table analysis such that the statistical associations between different outcome subsets can be evaluated (Fig. 2) [4,5,11]. Sequence sites with statistically significant association with other sites can then be used as pattern signatures [8,13].

In the proposed analysis, there are two phases. In the first phase, the molecular sites in the multiple sequence alignment are labeled into three different types depending on their site association characteristics: conserved sites (C-sites) with conservation patterns, interdependent sites (D-sites) with association patterns, and hypervariate sites (H-sites) that cannot be classified into the previous types. Next, the importance of these sites is evaluated by testing their significant association to pre-targetted functionality of the biomolecule such as known structural or functional patterns. In previous research, the patterns derived from associated sites were capable of inferring secondary and tertiary bonding structures [6], and have been used for the recognition of the ribosome binding sites in *E. coli* [15]. Similar sites can also have conformational, biochemical, and taxonomical significance [3,29]. In other studies, regions obtained from statistical patterns are shown to correspond to exon sub-regions [8] and the identification of the three-dimensional molecular core sites [7].

* Corresponding author.
E-mail address: dchiu@uoguelph.ca (D.K.Y. Chiu).

http://dx.doi.org/10.1016/j.neucom.2014.06.036
0925-2312/

2. Granular associations at different levels

One fundamental task of data analysis that is found to be extremely useful is the discovery, description and quantification of the associations embedded in a complex data type [26]. For complex biomolecular sequence data, it is analogous to an analysis of the nanostructure that represents the biomolecule. Typically, the associations of an event can be analyzed considering the observations from the complete outcome space. However, the associations from a given dataset can be a global or a local phenomenon (Fig. 2), that is, the associations can be local if only a subset of the complete outcome space is involved. The two phenomena between a local or a complete space can be quite different and their information may convey different characteristics. For example, when only a portion of the complete set of outcomes is relevant, then a local analysis on a subset may indicate a different magnitude that deviates from independence in the associations. Fig. 2 illustrates the probability distribution characteristics in terms of the local and global patterns may deviate differently from the expected pattern event. The diagram shows that at the global or a local level, the observation pattern event can deviate differently from the two different null hypotheses, denoted as H_0^1 and H_0^2 respectively. Information at one level then may not exist at the other [4,5]. Therefore using multiple levels of pattern analysis may provide a more complete basis for data abstraction and analysis, and can be very valuable for knowledge discovery in some datasets.

3. p53—Guardian of the genome and its homologs

Lane called the tumor suppressor protein p53 the “guardian of the genome” and it was referred to as the cellular gatekeeper, mainly because of its role related to human cancers [21]. Under stress conditions, such as DNA damage (from ionizing radiation, UV radiation, chemotherapeutic agents etc.), or heat shock, hypoxia, and oncogene over-expression, wild type p53 is activated and triggers diverse biological responses in cell cycle arrest, as well as DNA repair, apoptosis, and cellular senescence. Hence p53 prevents the replication of damaged DNA and maintains the integrity of the genome. On the other hand, the inactivation of p53 due to mutations, deletion, or interaction with cellular and viral proteins is a common event in the development of diverse types of cancer. Indeed, p53 is frequently inactivated in about 45–50% of all types of cancer [17,19,22]. Under normal conditions, the active p53 responds to the DNA damage in the cells and prevents the proliferation of damaged cells. However, when p53 is inactivated, it loses its biological function, permitting the proliferation of the cells that carry the damaged DNA, possibly leading to tumor formation. This molecule has then been actively studied world-wide ever since.

The human p53 protein [20] is 393 amino acids long and has three domains: an N-terminal transactivation domain (1–93), a sequence specific DNA binding domain (102–292) and a C-terminal oligomerization domain (323–393). In 1997 and 1998, the p73 and p63 respectively were identified as structural and functional homologs of p53 [25], together all three molecules as the p53 family. The overall domain structure of the family members is conserved, with similar transactivation domain (TAD), DNA binding domain (DBD), and oligomerization domain (OGD), even though their total lengths are quite different. However unlike p53, the genes encoding p63 and p73 are rarely mutated in human cancer, and knock-out mice studies [27,31] demonstrate developmental defects rather than a propensity for tumor formation. Hence one objective in this study is to investigate functional patterns related to the variability in the molecular sites that may indicate their differences.

4. Methodology

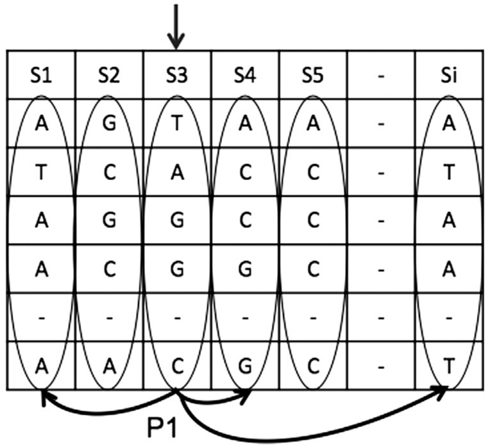
4.1. First phase: construction of a sequence association structure

Assuming that common and discriminative properties reflected from the p53 sequences from different species are important, the sequences are aligned to facilitate the discovery of these patterns. The first phase of our analysis then labels the aligned sites in the p53 sequences into different types based on the aligned variation and association characteristics. Three types of variability are identified [9,29]:

- *Associated sites* (*D-sites*): These sites indicate the observed amino acids in the aligned sites are significantly associated with the amino acids of multiple other sites, reflecting a complex interdependent relationship.
- *Invariant or conserved sites* (*C-sites*): These sites indicate the observed amino acids in the alignment are mostly the same, reflecting a constant or small variability.
- *Hypervariate sites* (*H-sites*): These sites are those sites that cannot be classified into one of the above.

Since D-sites are those with observed amino acids that can be associated with the amino acids from multiple other sites in the molecular alignment, the associated patterns can be considered as convergent association pattern. The association relationships are detected by using a suitable statistical test, using independence as the null hypothesis. Given the aligned sequences, each aligned site was tested for association with each of the other sites. In our case and depending on the sequence data, we accept a site to have convergent association when it is found to have statistically significant association with more than one site (Fig. 1).

A statistical test can be used to evaluate the significance of the association relationship between two distinct aligned sites. After all the statistical evaluations are applied, the interrelationships between the aligned sites in the molecule form a complex association structure, analogous to a nanostructure of the molecule. The analysis is then to identify meaningful network signatures, so that further association with identified functional properties can be evaluated. The goal is to relate whether a functional pattern can be linked to specific sites of the molecule from these pattern signatures. We hypothesize that in identifying sites from the association structure, the underlying functional relationships of the biomolecule may be revealed.



S1	S2	S3	S4	S5	-	Si
A	G	T	A	A	-	A
T	C	A	C	C	-	T
A	G	G	C	C	-	A
A	C	G	G	C	-	A
-	-	-	-	-	-	-
A	A	C	G	C	-	T

Fig. 1. Significant site-site pattern (P1) and site with convergent association patterns (S3).

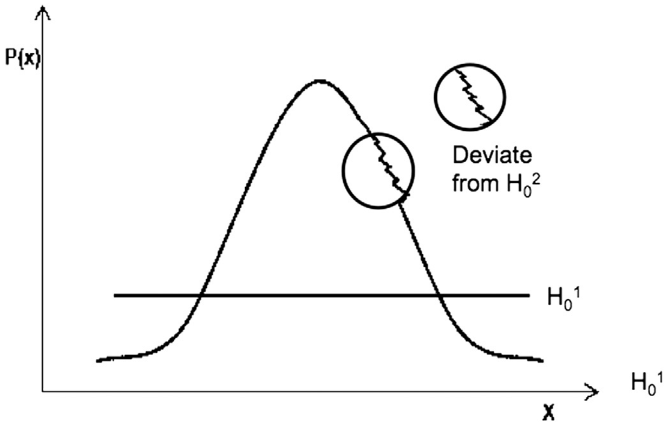


Fig. 2. Schematic probability curve showing deviation from hypothesis H_0^1 and H_0^2.

4.2. Selection of statistical test

In general, to evaluate associations between two aligned sites in sequences of a large sample size, chi-square test based on the construction of a two-dimensional contingency table can be applied [2]. To evaluate the significance of the association relationship between the sites, the null hypothesis is that the variables between two sites are independent. Based on a predefined significance level, the null hypothesis can be rejected such that the two sites are accepted as associated. This test is similar to the one used in [13], as they have the same null hypothesis and distribution. When the sample size is small, other tests such as Fisher's exact test can be used, resulting in sparse contingency tables. With multiple testings for association relationship in the alignment involving multiple site pairs, Bonferroni correction [1] can be applied to control the familywise error rate:

$$\alpha' = (\alpha)/n$$

where α is the significance level and n is the number of multiple tests.

4.3. Site pattern signatures using different sizes of two-dimensional contingency table analysis

To address the problem of sparse contingency table, as well as to detect different forms of granular associations discussed earlier, different outcome subsets from the two-dimensional contingency tables between two variables can be considered. Multiple levels of data association are constructed by using different sizes of outcome subspace in the two-dimensional contingency table. Three levels of analysis can be considered:

- Full contingency table analysis (the R_F Method)
- 2 × 2 contingency sub-table analysis (the $R_{2\times2}$ Method)
- Single cell contingency table analysis (the R_1 Method)

The standard full contingency table analysis (R_F Method) evaluates the association relationship between two distinct sites from an aligned sequence ensemble. After the contingency table relating two sites in the aligned sequences is generated, Fisher's exact test can be applied. The test detects the significance of the association between the two selected sites. If the test statistic is larger than the tabulated value at a pre-defined significance level, then the association is accepted as significant.

The 2 × 2 contingency sub-table analysis of a two-dimensional table ($R_{2\times2}$ Method) evaluates the association between the outcome subsets of, say *X* and *Y* of two distinct aligned sites, denoted as sub-*X*, and sub-*Y* that was selected using relevant criteria. Note that since we are evaluating the association of the amino acids between two sites in the human p53 sequence, the sub-table analysis is analogous to evaluating their frequencies based on a meaningful outcome subspace.

The criterion for selecting a 2 × 2 sub-table can be described as follows:

- Select the outcomes (other than that from the human sequence) with the highest marginal frequency.
- Create a sub-table involving the amino acid type observed from the human sequence, resulting in a 2 × 2 sub-table.

In a single cell contingency analysis method, the cell relating between the observed amino acid type in the human sequence of sites *X* and *Y* can be evaluated for significant association using method described in [18,30]. The hypothesis test is applied to identify significant deviation from independent associations. The test statistic *t* is computed based on the normal distribution on the difference between the observed and expected frequencies normalized. It is defined as

$$t = (obs_{xy} - exp_{xy})/(\sqrt{exp_{xy}})$$

where obs_{xy} and exp_{xy} are the observed and expected frequency between the two amino acid types in the sample. The statistic *t* can be adjusted when certain assumptions are not met using the adjusted test statistic calculated from the marginal probabilities in the contingency table [30]. To evaluate the statistical significance, if the test statistic is larger than the tabulated value at a pre-defined significance level, then the association is accepted as significant.

4.4. Second phase: evaluation of significant functional patterns to pattern signatures

In the second phase, the association between defined pattern signatures and a targeted functional characteristic of the p53 protein is evaluated.

Different types of statistical patterns generated previously are classified into seven pattern signatures:

- *Conserved sites pattern* (*CS*): It indicates sites with mostly constant value observation. This is the pattern of the C-sites.
- *$R_{2\times2}$ pattern* (*or $R_{2\times2}$*): It indicates sites identified as significantly associated using the described 2 × 2 contingency sub-table method.
- *R_1 pattern* (*or R_1*): It indicates sites identified as significantly associated using the single cell contingency table method.
- *$CS+R_{2\times2}$ pattern*: It indicates sites that are either conserved (CS) or associated using the 2 × 2 contingency sub-table method.
- *$CS+R_1$ pattern*: It indicates sites that are either conserved (CS) or associated using the single cell contingency table method.
- *$R_{2\times2}+R_1$ pattern*: It indicates sites that are associated either using the 2 × 2 sub-table or the single cell contingency table method.
- *$CS+R_{2\times2}+R_1$ pattern*: It indicates sites that are either conserved or associated either using the 2 × 2 sub-table or the single cell contingency table method.

It is assumed that knowledge about the functional characteristic of a site may not be completely known with many other factors affecting it. Therefore patterns of relationship may not be deterministic and could only be revealed probabilistically. The goal here is to analyze the significance of the association between the identified pattern signatures and a targeted functionality. If the association is significant, it then relates the sites with the pattern signature to the functional characteristic.

The statistical significance between the pattern signatures and the functional patterns is evaluated using a test of independence based on the sample distribution of the sites. Using a two-dimensional contingency table analysis method, the variables between a functional pattern (whether it is revealed on a site or not) and a network signature are evaluated. The chi-square statistical test is applied based on a pre-defined significance level. The null hypothesis assumes that the pattern signature and a functional pattern are independent and the alternate hypothesis otherwise. From the contingency table, the observed and expected frequencies are calculated and compared. The chi-square statistic is evaluated with one degree of freedom. The relationship is considered statistically significant if $\chi^2 > N_\alpha$, where N_α is the tabulated threshold value and α is the predefined significance level.

5. Experimental studies using the p53 protein alignment

The amino acid sequences used in the experiments are obtained from the UniProtKB database (http://www.uniprot.org). The database stored 34 different species of p53 sequences, three species of p63 sequences and 3 sequences of p73 sequences, among them including the human sequences.

In the first phase of the analysis, the multiple sequence alignment of 34 p53 sequences is obtained, using the ClustalW (Version 2.1) program [14]. The default settings are used, producing a total of 393 aligned sites. The alignment and the subsequent analysis indicate 115 sites as conserved sites (C-sites) with the CS pattern signature. The remaining 278 (393–115) aligned sites are identified as either the D-sites or the H-sites.

The three levels of association analysis, R_F, $R_{2\times2}$, and R_1, are then applied. Due to the small sample size and R_F generates largely sparse contingency tables, the method is excluded from further analysis. Using a 5% significance level and with the Bonferroni correction, the proposed $R_{2\times2}$ method identifies 107 D-sites and the R_1 method identifies 28 D-sites.

In the second phase of the analysis, functional characteristics are selected to evaluate whether they have significant relationship with the proposed pattern signatures. These evaluations are discussed below.

5.1. Analysis 1: comparing pattern signatures to P53/P63/P73 protein family

Since p53 has different tumor suppressing properties from its homologs in cancer patients, the differences can be indicated by the differences in the human sequence of the aligned sites between p53, p63 and p73, with respect to the observed amino acids. The human sequences of p53, p63, and p73 are aligned according to their common domains. The amino acid patterns are further characterized into 5 different types:

- Type I: The amino acids in the human sequence of p53, p63, and p73 are all the same.
- Type II: The amino acids in the human sequence of p53, p63, and p73 are all different.
- Type III: The amino acid in the human sequence of p53 is different from that of p63 and p73.
- Type IV: The amino acid in the human sequence of p63 is different from that of p53 and p73.
- Type V: The amino acid in the human sequence of p73 is different from that of p53 and p63.

Since Type III amino acid pattern differentiates p53 from the other two homologs of p63 and p73, this functional pattern is the most important in terms of how the protein functions are different with respect to tumor suppressing property. Fig. 3 shows that D-sites are mostly associated with Type III pattern (which discriminates between p53 and its homologs). In Table 1, it shows that the pattern signatures of CS, $R_{2\times2}$, CS+R_1, and $R_{2\times2}$+R_1 were stronger and statistically significant with 1% significance level. The $R_{2\times2}$+R_1 pattern signature is more significant than the individual effect of either $R_{2\times2}$ or R_1 (Table 1). The D-sites are positively associated with Type III patterns significantly, but not deterministically. C-sites with CS pattern signature are negatively associated with it (Table 2). Even though only half of the (R_1+$R_{2\times2}$) patterns observe Type III differences, it is a substantial increase as compared to the non-(R_1+$R_{2\times2}$) patterns and is highly significant. However, when the CS pattern signature is considered with the other patterns (CS+$R_{2\times2}$, and CS+$R_{2\times2}$+R_1), the chi-square value decreases drastically and is also weaker, indicating an interactive effect.

5.2. Analysis 2: comparing pattern signatures to the three-dimensional location of the P53 molecule

The 3-dimensional molecular structure for the DNA binding domain (DBD) is available in the PDB protein data bank (PDB ID: 1TUP) and can be displayed using the PyMOL software (www.pymol.org). The association testing between the pattern signatures and their site locations (whether in the exterior or interior) shows that none of the pattern signatures by itself is significantly associated, but when the C-sites and the D-sites are considered together, it shows an association with the molecular locations with 5% significant level. It shows that the combined sites with CS, $R_{2\times2}$, R_1 patterns all contributed to the association of the locations. The D-sites (with $R_{2\times2}$ and R_1 patterns) are mostly situated

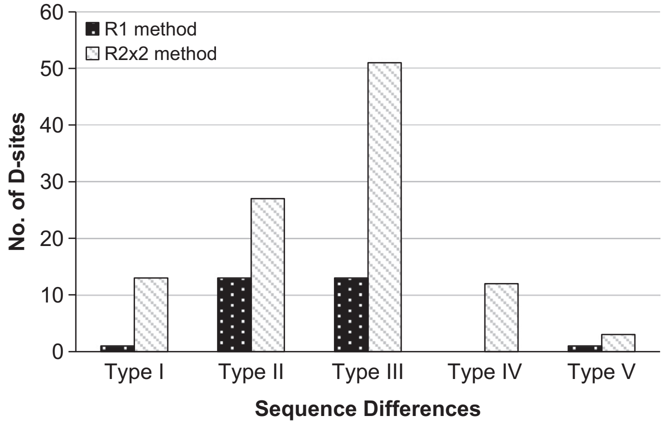


Fig. 3. Number of D-sites discriminating among the human sequence of p53, p63 and p73. Note that D-sites most distinguish p53 from its homologs (Type III sequence difference).

Table 1
Associations between pattern signatures and Type III sequence differences.

Pattern signatures	*P*-value	Association with Type III patterns	χ^2-Value
CS[a]	0.0001	Negatively associated	29.618
$R_{2\times2}$[a]	0.0001	Positively associated	19.088
R_1	0.0732	Not significant	3.210
CS+$R_{2\times2}$	0.2618	Not significant	1.259
CS+R_1[a]	0.0001	Negatively associated	17.713
$R_{2\times2}$+R_1 (D-sites)[a]	0.0001	Positively associated	24.147
CS+$R_{2\times2}$+R_1	0.7244	Not significant	0.124

[a] Indicates at least 99% confidence level.

Table 2
Observations between significant pattern signatures and Type III sequence differences.

Pattern signatures	Type III observations	Non-Type III observations	Total	*P*-values	Comments
CS pattern (C-sites)	14	109	123	< 0.0001	CS pattern is largely associated with non-Type III
Non-CS pattern	104	166	270		
Total	118	275	393		
($R_1+R_{2\times2}$) pattern (D-sites)	63	60	123	< 0.0001	Even though only half of the ($R_1+R_{2\times2}$) patterns observe Type III differences, it is a substantial increase and is highly significant
Non-($R_1+R_{2\times2}$) pattern(D-sites)	70	200	270		
Total	133	260	393		

Table 3
Site patterns and mutation frequency.

Sites	No. of sites	Total mutation frequency of all sites	Average mutation frequencies
D-sites ($R_1+R_{2\times2}$) pattern	44	2039	46.34
C-sites (**CS pattern**)	111	10,827	97.54
H-sites	30	1265	42.16

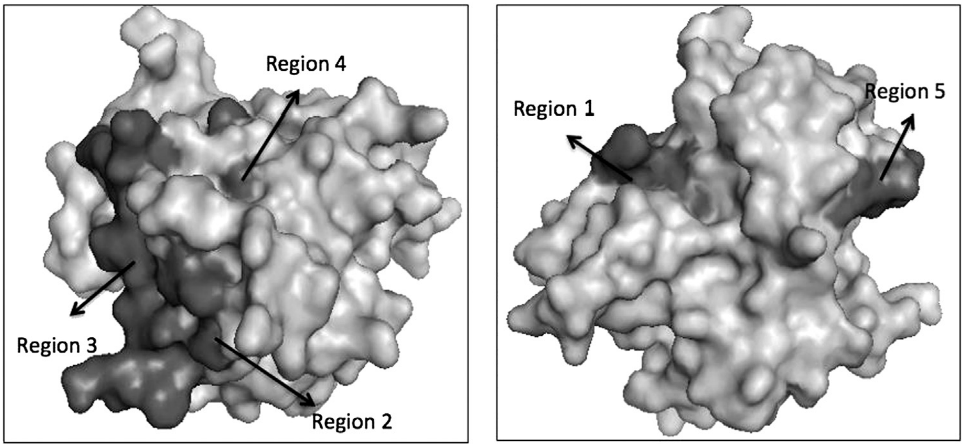


Fig. 4. Identified D-site regions in the 3D structure of the p53 protein core domain. They are indicated in black. All, except region 4, are in the exterior of the molecule.

in five regions. They are: region 1 (from 112:G to 115:H), region 2 (from 189:A to 192:Q), region 3 (from 200:N to 214:H), region 4 (from 233:H to 236:Y) and region 5 (from 261:S to 264:L) in the human p53 sequence. All these regions are in the exterior of the molecule, except region 4, which is in the interior of the molecule (Fig. 4). It could be explained that D-sites, which are statistically associated with multiple other sites, are likely more related to the exterior properties of the molecule.

5.3. *Analysis 3: comparing D-sites (R_1 and $R_{2\times2}$ patterns) to mutation frequency*

The mutation frequency for the amino acids in the human sequence p53 is available from the UMD TP53 mutation database [28]. It is noted that mutation rate is highest in the DNA binding domain (from site 102 to 292). Six sites have extremely high mutation frequency and are considered as mutation hotspots. They are excluded in the following analysis. The mutation frequencies calculated for the different site patterns are then depicted in Table 3. The table clearly shows that the mutation frequencies, both the total and the average frequencies for C-sites are much higher than that of the D-sites or the H-sites. It could be explained by that D-sites are more tightly linked to the other sites, and therefore less likely to be mutated. This explanation is consistent with the evaluation in Analysis 2.

6. Conclusions

Clearly, there are many analysis that could be done than those indicated here. The experimental studies on p53 protein confirm that the proposed evaluation is useful to identify association pattern signatures based on the characteristics of the network structure of the alignment. The proposed association analysis extracts statistically significant information based on the outcome subspaces using different sizes of the two-dimensional contingency table. The experiments on p53 show that the method identifies associated patterns as $R_{2\times2}$ and R_1 pattern signatures. The analysis further reveals that the defined pattern signatures can be compared to targeted structural and functional patterns of the molecule, allowing probabilistic uncertainty. In summary, the extracted association pattern signatures have proven to be useful in relating to some structural and functional characteristics. It is also useful in specifically identifying the sites that the associations occur.

Acknowledgments

The research is supported by Natural Sciences and Engineering Research Council of Canada, Canada, Discovery Grant #400297. We would like to thank Prof. Zeny Feng for checking the statistical tests.

References

[1] H. Abdi, Bonferroni and Šidák corrections for multiple comparisons, in: N.J. Salkind (Ed.), Encyclopedia of Measurement and Statistics, Sage, Thousand Oaks, CA, 2007.

[2] Y.M. M. Bishop, S.E. Fienberg, P.W. Holland, Discrete Multivariate Analysis: Theory and Practice, MIT Press, USA, 1975.

[3] D.K.Y. Chiu, X. Chen, A.K.C. Wong, Association between statistical and functional patterns in biomolecules, in: Proceedings of the Atlantic Symposium on Computational Biology and Genome Information Systems and Technology, Durham, USA, 2001, pp. 64–69 .

[4] D.K.Y. Chiu, B. Cheung, Hierarchical maximum entropy discretization. computing and information, in: R. Janicki, W. Koczkodaj (Eds.), Proceedings of the International Conference on Computing and Information (ICCI'89), Toronto, Canada, North-Holland, Amsterdam, 1989, pp. 237–242.

[5] D.K. Y. Chiu, B. Cheung, A.K. C. Wong, Information synthesis based on hierarchical maximum entropy discretization, J. Exp. Theor. Artif. Intell. 2 (1990) 117–129.

[6] D.K. Y. Chiu, T. Kolodziejczak, Inferring consensus structure from nucleic acid sequences, Comput. Appl. Biosci. 7 (1991) 347–352.

[7] D.K. Y. Chiu, T.W. H. Lui, NHOP: a nested associative pattern for analysis of consensus sequence ensembles, IEEE Trans. Knowl. Data Eng. 25 (10) (2013) 2314–2329.

[8] D.K. Y. Chiu, T.W. H. Lui, A multiple-pattern biosequence analysis method for diverse source association mining, Appl. Bioinform. 4 (2) (2005) 85–92.

[9] D.K. Y. Chiu, Y. Wang, Multipattern consensus regions in multiple aligned protein sequences and their segmentation, EURASIP J. Bioinform. Syst. Biol. 35809 (2006) 1–8.

[10] D.K. Y. Chiu, A.K. C. Wong, Multiple pattern associations for interpreting structural and functional characteristics of biomolecules, Inf. Sci. 167 (2004) 23–39.

[11] D.K. Y. Chiu, A.K. C. Wong, B. Cheung, Information discovery through hierarchical maximum entropy discretization and synthesis, in: Gregory Piatetsky-Shapiro, William J. Frawley (Eds.), Knowledge Discovery in Databases, MIT Press, Cambridge, MA, 1991, pp. 126–140.

[12] D.K.Y. Chiu, P.S.C. Xu, InfoBarcoding: selection of non-contiguous sites in molecular biomarker, in: Proceeding of the Computational Advances in Bio and Medical Sciences (ICCABS), 2011, pp. 69–74.

[13] K. Durston, D.K. Y. Chiu, A.K. C. Wong, G.C. L. Li, Statistical discovery of site inter-dependencies in sub-molecular hierarchical protein structuring, EURASIP J. Bioinform. Syst. Biol. (2012) 8.

[14] European Bioinformatics Institute tool for Multiple Sequence Alignment using clustalw2, ⟨http://www.ebi.ac.uk/Tools/msa/clustalw2.html/⟩.

[15] D. Frishman, A Mironov, M. Gelfand, Starts of bacterial genes: estimating the reliability of computer predictions, Gene 234 (1999) 257–265.

[16] A.J. Gonzalez, L. Liao, C.H. Wu, Predicting ligand-binding residues using multipositional correlations and kernel canonical correlation analysis, in: Proceedings of the 2010 IEEE International Conferernce on Bioinformatics and Biomedicine (BIBM), 2010, pp. 158–163.

[17] M.S. Greenblatt, W.P. Bennett, M. Hollstein, C.C. Harris, Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis, Cancer Res. 54 (1994) 4855–4878.

[18] S.J. Haberman, The analysis of residuals in cross-classified tables, Biometrics 29 (1973) 205–220.

[19] M. Hollstein, D. Sidransky, B. Vogelstein, C.C. Harris, p53 Mutations in human cancers, Science 253 (5015) (1991) 49–53.

[20] A.C. Joerger, A.R. Fersht, Structural biology of the tumor suppressor p53 and cancer-associated mutants, Adv. Cancer Res. 97 (2007) 1–23.

[21] D.P. Lane, Cancer and p53, guardian of the genome, Nature 358 (1992) 15–16.

[22] D.P. Lane, C.F. Cheok, S. Lain, p53-based cancer therapy, Cold Spring Harb. Perspect. Biol. (2010) 2a001222 (2010).

[23] T.Y. Lin, Granular computing, Lecture Notes in Computer Science, vol. 2639, Springer, Berlin (2003) 16–24.

[24] T.Y. Lin, From rough sets and neighborhood systems to information granulation and computing in words, in: Proceedings of the European Congress on Intelligent Techniques and Soft Computing, 1997, pp. 1602–1607.

[25] G. Melino, X. Lu, M. Gasco, T. Crook, R.A. Knight, Functional regulation of p73 and p63: development and cancer, Trends Biochem. Sci. 28 (2003) 663–670.

[26] W. Pedrycz, Granular Computing: An Emerging Paradigm, Physica-Verlag, Heidelberg, 2003.

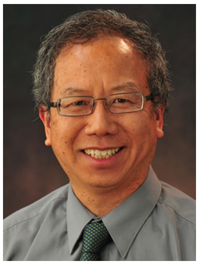
[27] T. Stiewe, The p53 family in differentiation and tumorigenesis, Nat. Rev. Cancer 7 (3) (2007) 165–168.

[28] The p53 website, ⟨http://p53.free.fr/⟩.

[29] A.K. C. Wong, T.S. Lui, C.C. Wang, Statistical analysis of residue variability in cytochrome C, J. Mol. Biol. 102 (2) (1976) 287–295.

[30] A.K.C. Wong, Y. Wang, High-order pattern discovery from discrete-valued data, IEEE Trans. Knowl. Syst. 9 (6) (1997) 877–893.

[31] A. Yang, M. Kaghad, D. Caput, F. McKeon, On the shoulders of giants: p63, p73 and the rise of p53, Trends Genet. (2002) 90–95.



David K.Y. Chiu is a professor at the School of Computer Science, University of Guelph, Canada. He is a former recipient of the Science and Technology Agency (STA) Fellowship of Japan and the Alley Heaps Chair of Computing Science at St. Francis Xavier University. He has published more than 100 technical papers in the areas of artificial intelligence, pattern recognition and bioinformatics. He has guest edited a special issue on bioinformatics in *Biomedical Engineering*. Previously, he has done research with the National Network Centers of Excellence of Canada and NCR Canada Ltd.

Ramya Manjunath did her research with the School of Computer Science, University of Guelph, Canada. She graduated with a Master of Science in Bioinformatics degree at the University of Guelph in 2012. She is currently working as a research assistant at Biodiversity Institute of Ontario, University of Guelph. Her research interests include NGS data analysis, data mining, data visualization, and pattern recognition.