Inferring the Association Network from p53 Sequence Alignment Using Granular Evaluations

David K.Y. Chiu and Ramya Manjunath

University of Guelph, Guelph, Ontario, Canada
{dchiu@uoguelph.ca}

Abstract. The relationship connecting the biomolecular sequence, the molecular structure, and the biological function is of extreme importance in nanostructure analysis such as drug discovery. 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, a new method to detect associated sites in aligned sequence ensembles is proposed. It involves the use of multiple sub-tables (or levels) of two-dimensional contingency table analysis. The idea is to incorporate analysis by using a concept known as granular computing, which represents information at different levels of granularity. The analysis involves two phases. The first phase includes labeling of the molecular sites in the p53 protein multiple sequence alignment according to the detected associated patterns. The sites are consequently labeled into three different types based on their site characteristics: 1) conserved sites, 2) associated sites and 3) hypervariate sites. In the second phase, the significance of the extracted site patterns is evaluated with respect to targeted structural and functional characteristics of the p53 protein. The results indicate that the extracted site patterns are significantly associated with some of the known functionalities of p53, a cancer suppressor. Furthermore, when these sites are aligned with p63 and p73, the homologs of p53 without the same cancer suppressing property, based on 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 could indicate their differences in cancer suppressing property.

Keywords: Data-mining, association network, protein sequence alignment, granular computing, bioinformatics.

1 Introduction

Biological sequences when aligned can provide the common or discriminatory information about the individual residue of the biomolecule family. It can also provide the information from which knowledge can be extracted that directs us towards the 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 Chiu & Kolodziejczak, 1991, Chiu & Lui, 2005, Chiu & Liu,

A. Ngom et al. (Eds.): PRIB 2013, LNBI 7986, pp. 220–230, 2013.

2012). Identifying the sequence patterns that represent the functional characteristics of the biomolecule is vital in nanostructure analysis such as drug discovery (González, Liao, & Wu, 2010).

Previous studies, involving the multiple sequence alignment of related species have indicated that various kinds of interdependent or associated patterns can be indicative of the structural and functional characteristics of the biomolecule (Chiu, Chen, & Wong, 2001; Chiu & Lui, 2005; Chiu & Wong, 2004; Chiu & Lui, 2009; Chiu & Wang, 2006; Chiu & Xu, 2011). In this paper, a new method in inferring the association network in aligned sequence ensembles is proposed. It is derived from the concept of granular computing, where information is extracted at different levels of granularity or resolution (Lin et al. 1997, 2003). It involves the use of different sizes of two-dimensional contingency table analysis by focusing on the statistical associations between different outcome subsets (Chiu & Cheung 1989, Chiu et al. 1990, 1991). Furthermore, molecular sites with association patterns having multiple relationships with other sites demonstrate convergent information (Durston et al. 2012).

In the proposed analysis, there are two consecutive phases. First, the molecular sites in the multiple sequence alignment are labeled into three different types based on their site association characteristics: conserved sites (C-sites), interdependent sites (D-sites), and hypervariate or other sites (H-sites). Next, the importance of these sites is evaluated by testing their association to the functionality of the biomolecule such as known structural or functional characteristics.

In an aligned sequence ensemble, associated sites refer to sites that have statistical significance relationship with another site. In proteins, they represent sites with amino acid pairs observed together. Two types of associations can be considered, the association between two sites (such as X and Y sites) and the association among multiple sites (such as W, X, Y, and Z sites). Previous studies using multiple sequence alignment have observed that associated sites can predict the functional sites in biomolecules. For example, the patterns derived from associated sites were capable of inferring secondary and tertiary bonding structures (Chiu & Kolodziejczak, 1991), and have been used for the recognition of the ribosome binding sites in E. coli (Frishman, 1999). Similar sites can also have conformational, biochemical, and taxonomical significance (Wong, Liu, & Wang, 1996; Chiu et al., 2001). In other studies, regions obtained from statistical patterns are shown to correspond to exon sub-regions (Chiu & Lui, 2005) and the identification of the three-dimensional molecular core sites (Chiu & Lui, 2012).

2 Associations at Different Levels

One of the fundamental tasks of data mining is the discovery, description and quantification of the associations within the data (Pedrycz, 2001). Typically, the information from the associations in an event is detected considering the complete outcome space. However, the associations in the given dataset can be a global or a local phenomenon (Fig. 1). The two phenomena can be quite different and their information hence may convey different characteristics.

Figure 1 depicts a probability distribution curve with two different phenomena, local and global deviations from the expected pattern event. At the global and local levels, the observation pattern event deviates from two different null hypotheses, H^1_0 and H^2_0, respectively. At the global level, the observed data have defined deviation, whereas at the local level, the data can further deviate from the locally expected distribution.

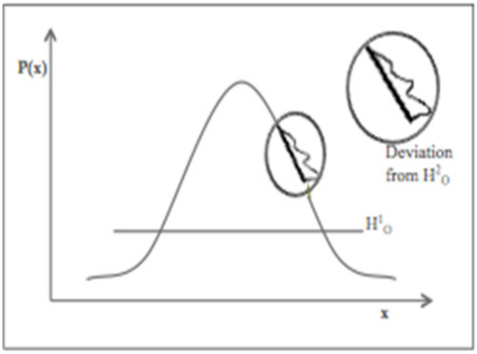


Fig. 1. Probability curve showing deviation from H^1_o and H^2_o

Hence, the information at one level of resolution may not exist at another, and this information may be significant (Chiu et al., 1991). Therefore focusing on multiple levels of resolution provides a more complete basis for data abstraction and knowledge discovery and can be extremely valuable for some datasets.

3 p53- Guardian of the Genome and Its Homologs

Lane (1992) first called the tumor suppressor protein p53 the "guardian of the genome" and Levine (1997) called it the "cellular gatekeeper". This molecule has been actively studied world-wide ever since. Under stress conditions, such as DNA damage (such as from ionizing radiation, UV radiation, and chemotherapeutic agents), heat shock, hypoxia, and oncogene over-expression, wild type p53 is activated and triggers diverse biological responses in cell cycle arrest, DNA repair, apoptosis, and cellular senescence. Hence p53 prevents the replication of damaged DNA and maintains the integrity of the genome.

The human p53 protein (Joerger & Fersht, 2007) 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).

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 (Greenblatt et al. 1994; Lane et al., 2010, Hollstein et al., 1991). Under normal conditions, the active p53 responds to the DNA damage in the cells and prevents the proliferation of damaged cells. When p53 is inactivated, it loses its biological function, permitting the proliferation of cells that carry damaged DNA, eventually leading to tumor formation.

In 1997 and 1998, p73 and p63, respectively, were identified as structural and functional homologs of p53 (Melino et al., 2003). The overall domain structure of the p53 family members is conserved and consists of a transactivation domain (TAD), DNA binding domain (DBD), and oligomerization domain (OGD). Unlike p53, the genes encoding p63 and p73 are rarely mutated in human cancer, and knock-out mice studies demonstrate developmental defects rather than a propensity for tumor formation.

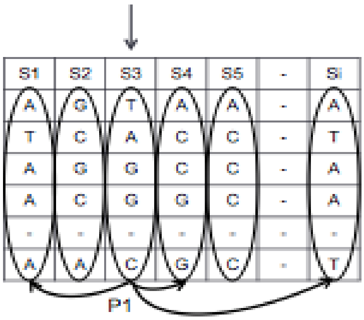
4 Methodology

4.1 The First Phase of Analysis

In the first phase of our proposed analysis, the aligned sites in the p53 protein multiple sequence alignment were labeled into different types based on aligned site characteristics. The three different types of sites were also discussed in (Wong et al. 1976; Chiu & Wang, 2006):

- Associated sites (D-sites): The D-sites indicated the sites with observed amino acid values multiply associated with the values of other sites, reflecting a complex interdependent relationship.
- Invariant or conserved sites (C-sites): The C-sites indicated the sites mostly with the same amino acid value, reflecting constant value observation.
- Hypervariate sites (H-sites): The H-sites indicated the sites that could not be classified into either the D-site or the C-site types.

In multiple sequence alignment of a biomolecule, convergent association pattern (such as D-sites) represented the sites that have association relationship with other sites converging on them. The association relationship between sites was detected by using a suitable statistical hypothesis test. In an aligned ensemble, each aligned site was statistically tested for association with all other sites. In our case, when a site was found to be significantly associated with more than one site, it was considered to have a convergent association pattern that reflected a multiple interdependence relationship. For example, in Figure 2, site S3 was tested for association with all the other sites and the sites associated with S3 were indicated by the P1 (site-site) pattern.



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. 2. Site-site pattern (P1) (modified from Chiu & Xu, 2011)

A statistical hypothesis test was used to evaluate the association relationship between two distinct sites in the aligned sequences. The goal was to evaluate whether or not a site was significantly associated with other sites in the aligned ensemble. It was hypothesized that in identifying the network of association patterns, the underlying functional structure of the biomolecules may be revealed.

4.2 Selection of Statistical Test

In general with large sample size, chi-square test can be applied to evaluate the significance of the association relationship between the site variables. Here, the sample size was small resulting in sparse contingency tables. Thus Fisher`s exact test could be applied.

4.3 Correction for Multiple Testing

In this phase, each aligned site in the alignment was tested for association relationship with all other sites. With multiple hypotheses tested, Bonferroni correction was applied to control the familywise error rate:

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

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

4.4 Detection of D-sites Using Different Sizes of Contingency Tables

The use of the proposed method based on granular association facilitated the identification of D-sites in the aligned sequence ensembles for different outcome subsets between two variables. Multiple levels of data abstraction were constructed by using different sizes of the two-dimensional contingency tables. Based on three different sizes of the contingency table, three levels of analysis could be employed:

- Full contingency table analysis (R_F)
- 2x2 contingency sub-table analysis (R_{2x2})
- Single cell contingency table analysis (R_1)

4.5 Full Contingency Table Analysis (R_F Method)

The standard full contingency table analysis evaluates the association relationship between two distinct sites from an aligned sequence ensemble. After the contingency table relating two sites in aligned sequences is generated, Fisher's exact test can be applied to each relationship. The test detects the significance of the association between the two selected distinct sites. The null hypothesis is that the site variables, say X and Y, are independent and the alternate hypothesis otherwise. If the test statistic is larger than the tabulated value at a pre-defined significance level, then the association is accepted as significant.

4.6 2x2 Contingency Sub-table Analysis (R_{2x2} Method)

The 2x2 contingency sub-table analysis of a two-dimensional table evaluates the association between the outcome subsets, denoted as sub-X, and sub-Y that was selected by using relevant criteria from the full contingency table. There were two criteria for selecting a sub-table, analogous to the use of two different but similar estimators.

The first selection criterion for selecting the 2x2 sub-table can be described as follows:

- Select the first two outcomes from a full contingency table with the highest marginal frequency.
- Create a sub-table involving the human amino acid in the two sites.

The second similar selection criterion can be used:

- Select the human amino acid in the two X and Y sites.
- Select the non-human amino acid in the X and Y sites with the highest marginal frequency.

After the 2x2 sub-table is constructed, the test of independence was applied to the two sites.

4.7 Single Cell Contingency Table Analysis (R_1 Method)

With a full contingency table constructed relating between, say sites X and Y, the cell with the observed amino acid in the human sequence of site X and site Y was selected. The hypothesis test is then applied to identify significant associations. The test statistic is computed based on the normal distribution on the difference between the observed and expected frequencies (Haberman, 1973, Wong & Wang, 1997). If the test statistic is larger than the tabulated value at a pre-defined significance level, then the association is accepted as significant. In another words, the single cell contingency table analysis is applied to evaluate the association between two different sites of the human sequence based on the distribution obtained from the aligned sequence ensemble.

4.8 The Second Phase of Analysis

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

As described before, the different types of statistical patterns can be classified into seven different categories:

- Conserved sites pattern (CS): It indicates sites with mostly a constant value observation.
- R_{2x2} pattern: It indicated sites identified as significantly associated using the 2x2 contingency sub-table method.
- R_1 pattern: It indicated sites identified as significantly associated sites by the single cell contingency table method.

- CS + R_{2x2} pattern: It indicated sites that are either conserved or identified as associated sites by the 2x2 contingency sub-table method.
- CS + R_1 pattern: It indicated sites that are either conserved or identified as significantly associated sites by the single cell contingency table method.
- R_{2x2} + R_1 pattern: It indicated sites identified as significantly associated sites by either the 2x2 contingency sub-table or the single cell contingency table method.
- CS + R_{2x2} + R_1 pattern: It indicated sites that are either conserved or identified as significantly associated sites by the 2x2 contingency sub-table or the single cell contingency table method.

The goal here is to analyze the association between the identified patterns and targeted functionalities to determine if they are significantly associated. This analysis would be useful in identifying significant functional association, possibly leading to the discovery of specific functional sites with the desirable properties. In the experiments, we had considered six different p53 functionalities, including structural characteristics and amino acid differences between p53 and its homologs of p63 and p73. There are five different types of discrimination between p53, p63 and p73, as:

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

4.9 Test of Independence in the Second Phase of Analysis

The statistical significance between the generated site patterns and the functional characteristics is evaluated using a test of independence from the construction of a new 2x2 contingency table, indicating whether the pattern and the functionality are significantly associated or not. The variable on the rows in the table indicated a targeted functionality (e.g. polarity) and the variable on the columns indicated the generated site pattern (e.g. CS pattern). The chi-square statistical test is then applied.

The null hypothesis assumes that the pattern (P) and the functionality (F) are independent and the alternate hypothesis otherwise. From the observed frequency table, the observed and expected frequencies are then calculated. The chi-square statistic is computed with one degree of freedom based on the deviations between the observed frequencies from the expected frequencies. The association relationship between the variables P and F is considered to be statistically significant if $\chi2 > N_\alpha$, where N_α was the tabulated threshold value with one degree of freedom and α is the confidence level.

5 Experimental Studies Using the p53 Protein Alignment

The amino acid sequences used in the experiments were 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.

In the first phase of analysis, the multiple sequence alignment of 34 p53 sequences was obtained, using the alignment from the ClustalW (Version 2.1) program. The following ClustalW default settings were used. (The pairwise alignment parameters were: protein weight matrix = Gonnet, gap open penalty = 10, and gap extension penalty = 0.1; the multiple alignment parameters were: protein weight matrix = Gonnet, gap open penalty = 10, gap extension penalty = 0.2, gap separation distances = 5, end gaps = off, and clustering method = neighbor joining.) The alignment indicated 115 sites as conserved sites and these sites were labeled as C-sites. The remaining 278 (393-115) aligned sites were employed in the experiments, to identify the D-sites and the H-sites.

The three levels of data abstraction methods, R_F, R_{2x2}, and R_1, were applied generating the labeled sites (as D-sites). Due to the small sample size of the data and R_F generates largely sparse contingency tables, hence the method were excluded from further analysis.

In the R_{2x2} method, two selection criteria (as two estimators) were used to select a sub-table from a full contingency table. In the p53-aligned data, it was found that both criteria selected similar D-sites as expected.

The proposed R_{2x2} method identified 107 D-sites with a 5% significance level after using the Bonferroni correction. In the transactivation domain (TAD), DNA binding domain (DBD), and oligomerization domain (OD), there were 20, 52, and 34 sites identified respectively.

The R_1 method identified 28 D-sites with a 5% significance level after using the Bonferroni correction. In the transactivation domain (TAD), DNA binding domain (DBD), and oligomerization domain (OD), there were 20, 4, and 4 sites identified respectively.

In the second phase of analysis, the human sequences of p53, p63, and p73 were aligned according to their common domains. This alignment was used to identify the discriminating types between the p53 family members.

The five different types used to discriminate among the human sequences of p53, p63, and p73 molecules were described. The association relationship between the defined patterns and the discriminating types were analyzed. The number of D-sites selected in type III was high in both the R_{2x2} and R_1 methods. Since type III differentiated p53 from the other two family members of p63 and p73. This relationship between the defined patterns was the most important.

The observed chi-square values and p-values for association testing between each pattern and discriminate type III were noted. Figure 3 shows that D-sites are mostly associated with type III (which discriminate between p53 and its homologs). The frequencies clearly demonstrated that the patterns CS, R_{2x2}, CS + R_1, and R_{2x2} + R_1 were stronger and statistically significant with type III discrimination with 0.01% significance level. The R_{2x2} + R_1 pattern was more significant than the individual effect of either R_{2x2} or R_1. However, when the CS pattern was considered with the other patterns (CS + R_{2x2}, and CS + R_{2x2} + R_1), the chi-square value decreased drastically and was also weaker. The results can be interpreted as follows:

- The patterns, CS, R_{2x2}, CS + R_1, and R_{2x2} + R_1, had different effect in discriminating between p53 and p63/p73.
- When the patterns, CS, R_{2x2}, and R_1, were considered together, the effects cancelled each other or that the CS pattern had an interactive effect with the D-sites effect.

6 Discussions and Conclusions

The experimental studies on p53 protein multiple sequence alignment confirm that the proposed granular association evaluation method is useful to identify and label associated site patterns. The method extracts information based on an outcome subspace in the data by using different resolutions (or sizes) of the two-dimensional contingency table. The experiments on p53 showed that the method identifies associated patterns in the R_{2x2} and R_1 analyses. Also, the R_{2x2} method identifies a higher number of sites than the R1 method, and the number of sites associated with each site may differ. The second phase of analysis revealed that the defined patterns can be associated with some targeted structural and functional properties of the p53 protein. In summary, the extracted association patterns have proven to be useful in discovering sites with some structural and functional properties of a protein molecule.

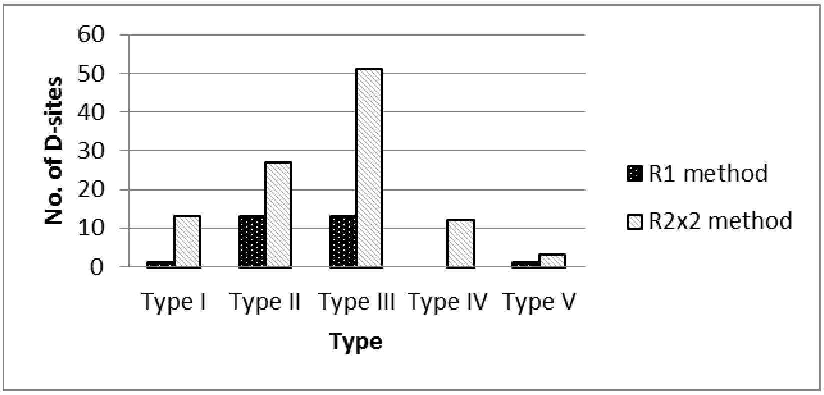


Fig. 3. Number of D-sites in discriminating among the human sequence of p53, p63 and p73. Note D-sites most distinguish p53 from its homologs (as in Type III).

Acknowledgements. The research is supported by Natural Sciences and Engineering Research Council of Canada, Discovery Grant #400297.

References

1. Chiu, D.K.Y., Chen, X., Wong, A.K.C.: 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, pp. 64–69 (2001)

2. Chiu, D.K.Y., Cheung, B.: Hierarchical Maximum Entropy Discretization. Computing and Information. In: Proceedings of the International Conference on Computing and Information (ICCI 1989), pp. 237–242. North-Holland, Toronto (1989)
3. Chiu, D.K.Y., Cheung, B., Wong, A.K.C.: Information Synthesis based on Hierarchical Maximum Entropy Discretization. Journal of Experimental and Theoretical Artificial Intelligence 2, 117–129 (1990)
4. Chiu, D.K.Y., Kolodziejczak, T.: Inferring Consensus Structure from Nucleic Acid Sequences. Computational Applications in Biosciences 7, 347–352 (1991)
5. Chiu, D.K.Y., Lui, T.W.H.: NHOP: A Nested Associative Pattern for Analysis of Consensus Sequence Ensembles. IEEE Trans. on Knowledge and Data Engineering (2012) (in press)
6. Chiu, D.K.Y., Lui, T.W.H.: A Multiple-pattern Biosequence Analysis Method for Diverse Source Association Mining. Applied Bioinformatics 4(2), 85–92 (2005)
7. Chiu, D.K.Y., Wang, Y.: Multipattern Consensus Regions in Multiple Aligned Protein Sequences and their Segmentation. EURASIP J. Bioinformatics Syst. Biol. 35809, 1–8 (2006)
8. Chiu, D.K.Y., Wong, A.K.C.: Multiple Pattern Associations for Interpreting Structural and Functional Characteristics of Biomolecules. Information Science 167, 23–39 (2004)
9. Chiu, D.K.Y., Wong, A.K.C., Cheung, B.: Information Discovery through Hierarchical Maximum Entropy Discretization and Synthesis. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 126–140. MIT Press, Cambridge (1991)
10. Chiu, D.K.Y., Xu, P.S.C.: InfoBarcoding: Selection of Non-Contiguous Sites in Molecular Biomarker. In: Proceeding, Computational Advances in Bio. and Medical Sciences (ICCABS), pp. 69–74 (2011)
11. Durston, K., Chiu, D.K.Y., Wong, A.K.C., Li, G.C.L.: Statistical Discovery of Site Inter-Dependencies in Sub-Molecular Hierarchical Protein Structuring. EURASIP J. on Bioinformatics and Systems Biology 2012, 8 (2012)
12. European Bioinformatics Institute tool for Multiple Sequence Alignment using clustalw2, `http://www.ebi.ac.uk/Tools/msa/clustalw2.html/`
13. Frishman, D., Mironov, A., Gelfand, M.: Starts of Bacterial Genes: Estimating the Reliability of Computer Predictions. Gene 234, 257–265 (1999)
14. Gonzalez, A.J., Liao, L., Wu, C.H.: Predicting Ligand-Binding Residues using Multi-Positional Correlations and Kernel Canonical Correlation Analysis. In: Proc. 2010 IEEE Intern. Conf. of Bioinformatics and Biomedicine (BIBM), pp. 158–163 (2010)
15. Greenblatt, M.S., Bennett, W.P., Hollstein, M., Harris, C.C.: Mutations in the p53 Tumor Suppressor Gene: Clues to Cancer Etiology and Molecular Pathogenesis. Cancer Research 54, 4855–4878 (1994)
16. Haberman, S.J.: The Analysis of Residuals in Cross-Classified Tables. Biometrics 29, 205–220 (1973)
17. Hollstein, M., Sidransky, D., Vogelstein, B., Harris, C.C.: p53 Mutations in Human Cancers. Science 253(5015), 49–53 (1991)
18. Joerger, A.C., Fersht, A.R.: Structural Biology of the Tumor Suppressor p53 and Cancer-Associated Mutants. Advanced Cancer Research 97, 1–23 (2007)
19. Lane, D.P.: Cancer and p53, Guardian of the Genome. Nature 358, 15–16 (1992)
20. Lane, D.P., Cheok, C.F., Lain, S.: p53-based Cancer Therapy. Cold Spring Harb. Perspect. Biol., 2, a001222 (2010)
21. Lin, T.Y.: Granular computing. In: Wang, G., Liu, Q., Yao, Y., Skowron, A. (eds.) RSFDGrC 2003. LNCS (LNAI), vol. 2639, pp. 16–24. Springer, Heidelberg (2003)

22. Lin, T.Y.: From Rough Sets and Neighborhood Systems to Information Granulation and Computing in Words. In: European Congress on Intelligent Techniques and Soft Computing, pp. 1602–1607 (1997)
23. Melino, G., Lu, X., Gasco, M., Crook, T., Knight, R.A.: Functional Regulation of p73 and p63: Development and Cancer. Trends Biochem. Sci. 28, 663–670 (2003)
24. Pedrycz, W.: Granular Computing: An Emerging Paradigm. Physica-Verlag, Heidelberg (2003)
25. The p53 website, `http://p53.free.fr/`
26. The UniProtKB database, `http://www.uniprot.org`
27. Wong, A.K.C., Lui, T.S., Wang, C.C.: Statistical Analysis of Residue Variability in Cytochrome C. J. Molecular Biology 102(2), 287–295 (1976)
28. Wong, A.K.C., Wang, Y.: High-Order Pattern Discovery from Discrete-Valued Data. IEEE Trans. on Knowledge Systems 9(6), 877–893 (1997)