Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2006, Article ID 35809, Pages 1–8
DOI 10.1155/BSB/2006/35809

Multipattern Consensus Regions in Multiple Aligned Protein Sequences and Their Segmentation

David K. Y. Chiu and Yan Wang

Department of Computing and Information Science, University of Guelph, Guelph, ON, Canada N1G 2W1

Received 23 November 2005; Revised 22 May 2006; Accepted 7 June 2006

Recommended for Publication by John Quackenbush

Decomposing a biological sequence into its functional regions is an important prerequisite to understand the molecule. Using the multiple alignments of the sequences, we evaluate a segmentation based on the type of statistical variation pattern from each of the aligned sites. To describe such a more general pattern, we introduce multipattern consensus regions as segmented regions based on conserved as well as interdependent patterns. Thus the proposed consensus region considers patterns that are statistically significant and extends a local neighborhood. To show its relevance in protein sequence analysis, a cancer suppressor gene called p53 is examined. The results show significant associations between the detected regions and tendency of mutations, location on the 3D structure, and cancer hereditable factors that can be inferred from human twin studies.

1. INTRODUCTION

Decomposing a sequence into regions can be extremely important in understanding the functional characteristics of the biomolecule. Performing this using multiple alignments of the sequence family can dramatically improve the reliability of the interpretation, as well as capturing the overall property beyond the original sequence. Thus consensus sequence, or frequency pattern along a segment across multiple aligned sequences, provides a convenient characteristic to indicate a commonly observed, and likely an intrinsic property of the sequences. A well-known example is the TATA binding protein, a DNA sequence (consensus TATAAA) upstream of the transcription start site in the promoter region of many eukaryotic genes. In addition, the notion of consensus structure (see Chiu and Kolodziejczak [1], Chiu and Harauz, [2]), proposed in the early 1990's, captures a different feature discovered from multiple aligned sequences. It confirms that a jointly inferred 2D, and even 3D structure, can be in some cases recovered from the aligned sequences, see Chiu and Harauz [2]. In these cases, the multiple aligned sequences can be treated as a sample observation of the sequence family. The detected pattern is analogous to an estimated overall feature of the biomolecules from the sequences. In this paper, we extend the notion further to propose multipattern consensus region that generalizes consensus sequence that has been found to be extremely useful in sequence analysis.

A multipattern consensus region is defined as a region segment given the multiple alignments of the sequences so that the segment is dominated by sites that are conserved or, in another instance, interdependent pattern characteristics. To define the patterns more rigorously, the patterns are detected based on statistical test of significance, rather than frequency count. Note that multipattern consensus region generalizes consensus sequence in that consensus sequence is a special case based on conservation patterns only. Because of the generalization, multipattern consensus regions can be more informative about the biomolecule, and allow analysis of these additional statistical properties as well. Previous studies have found various kinds of interdependent patterns in sequences to be very important in indicating the structural and functional characteristics of the molecule, see; Chiu and Harauz [2], Chiu and Liu [3]; Chiu and Wong [4]; Chiu and Lui [5]; and Greenblatt et al. [6].

There is another advantage in using statistical variation patterns in segmenting sequences into regions. One objective is to divide the aligned sequences into meaningful regions that have bearing on the functional characteristics of the biomolecule. However, which property is appropriate other than the original amino acid or nucleotide type may not be known. Identifying statistically significant patterns that consider both conserved and interdependent properties may provide a higher-level indicator of the unknown property, beyond the original amino acid or nucleotide type. Furthermore, statistical variation patterns are not exact, and can tolerant errors and inaccuracies.

Even though the notion of consensus region is in principle applicable to DNA or RNA sequences, these applications have not been explored using aligned sequences, using algorithms such as that by Boys and Henderson [7] and Li et al. [8]. One problem is the availability of meaningful multiple alignments for DNA and RNA sequences. Another problem is the difficulty in aligning these sequences due to problems such as segment rearrangement, see Chiu and Rao [9]. It is also possible that these sequences may behave differently since each unit in the sequence has only 4 possible types of nucleotides, compare to the usual 20 types of amino acids in proteins. Therefore this paper only focuses on evaluating consensus regions in multiple aligned protein sequences.

This paper presents an outline of the segmentation algorithm (see Yan [10]) for multipattern consensus regions in aligned protein sequences, similar to Zhang [11], but applied to statistical variation patterns rather than the original amino acids. The segmentation algorithm analyzes the sequences after identifying the initial label of the statistical variation patterns for each aligned site. The optimization of the segmentation algorithm can be computationally explosive, see Zhang [11]. We use a heuristic segmentation algorithm and adopt a split-and-merge strategy to divide the aligned sequences into multipattern consensus regions.

In the experiments, we apply the algorithm to analyze a biomolecule known as p53, a cancer suppressor. The detected multipattern consensus regions are compared to its 3D molecular model. We further analyze their relationship to known mutation properties and hereditable factors as observed in cancer occurrences between human twins in previous etiology studies, see Lichtenstein et al. [12], Magnusson et al. [13].

2. A RANDOM n-TUPLE REPRESENTATION

To model statistical variations involving sequences of discrete values, we represent the aligned sequences as outcomes of a random n-tuple, denoted as $X = (X_1, X_2, \ldots, X_n)$ (e.g., see Wong et al. [14]). Each variable in X is then a discrete-valued variable. For example, each unit in a sequence such as the amino acid residue of a protein sequence is an outcome of the corresponding random variable. The order of the variables in the random n-tuple is preserved, consistent with the alignment. Under this framework, each variable X_i $(1 \leq i \leq n)$ can be referred to as a feature variable of the sequences to be modeled. A realization of X is a sequence that can be denoted as $x = (x_1, x_2, \ldots, x_n)$, where x_i in x is referred to as a sequence attribute, and n is the length of the aligned sequences. Each x_i $(1 \leq i \leq n)$ can take up a sequence attribute value denoted as a_{ip}. A sequence attribute value a_{ip} is a value taken from the attribute value set, $\Gamma_i = \{a_{ip} \mid p = 1, 2, \ldots, L_i\}$. L_i is the size of the value set for variable X_i. If some sequences are shorter than the others, a null symbol representing a gap can be inserted. A multiple aligned ensemble of sequences can then be considered as the outcome observations of X. This general data model allows for different kinds of pattern detection to be analyzed.

3. TYPES OF STATISTICAL VARIATION PATTERNS

Using a scheme proposed by Wong et al. in [14], the statistical variation pattern of the outcome observations of a variable can be classified into four categories: (1) invariant, when all the outcomes are the same (labeled as I); (2) conserved, when most of the outcomes are dominated by a single type but not invariant (labeled as C); (3) interdependent, when values are strongly associated with other values (labeled as D); and (4) hypervariate when it cannot be classified into any of the above types (labeled as V).

The four proposed categories are intended to be inclusive and capture the variation characteristics from the aligned sequence ensemble. Conserved type and interdependent type may not be mutually exclusive. It is understood that an aligned site on a molecule can have both the effects of conservation and interdependency at different strengths.

3.1. Measure of conserved patterns

A conserved pattern at a point, say for a protein sequence, indicates that the observed amino acid residues in an alignment are not constant among the aligned sequences, even though they are observed to be mostly the same. However, because of its small variability, it may indicate intrinsic reason for its variability. The reason for its variability may not be known. There it is labeled differently from the invariant type.

Methods that evaluate variability of the outcomes of a variable X_i in X can be used to detect conserved pattern. We propose a measure referred to as the compositional redundancy (see Wong et al. [14]; Shannon [15]; and Gatlin [16]), which is defined as

$$R^{(1)}(X_i) = \frac{\log L_i - H(X_i)}{\log L_i}, \tag{1}$$

where $H(X_i)$ is the Shannon entropy function (see Shannon [15]) defined as

$$H(X_i) = -\sum_{p=1}^{L_i} P(X_i = a_{ip}) \log P(X_i = a_{ip}). \tag{2}$$

Note that $R^{(1)}(X_i) = 1$ when $H(X_i) = 0$, or that X_i is invariant. $R^{(1)}(X_i) = 0$ when $H(X_i)$ is maximized, with $H(X_i) = \log L_i$, or the occurrences of each type of the outcome of X_i are equiprobable. In other words, the higher the value of $R^{(1)}(X_i)$ is, the more conserved X_i is.

It is important though to distinguish a significant measure of $R^{(1)}(X_i)$ from those that are due to random perturbation. Assuming a binary decision determined from a statistical test of significance, we evaluate $R^{(1)}(X_i)$ empirically from the observed data. $R^{(1)}(X_i)$ has an asymptotic chi-square property, and a criterion for testing deviation from equiprobability of the feature composition can be used, see Gatlin [16]. However, when the sample size is small, a threshold identified from a clear "valley" in the histogram distribution in the observed sequences can be used. This heuristic method based on a threshold can still provide some meaningful interpretation of the pattern type Wong et al. [14].

3.2. Measure of interdependent pattern

Interdependent pattern indicates that values of the variable outcomes are strongly and significantly associated with values of other variables, see Chiu and Lui [3, 5]; Chiu and Wong [4]. Evaluation is based on the interdependency between values rather than the interdependency between their corresponding variables. It is used allowing those values of a variable that are statistically random to be disregarded and consider only the interdependent values of the variable in the calculation. The formula is indicated below in the statistical evaluation.

To consider only those that are statistically significant rather than due to random perturbations, we use the following method, based on the adjusted residual, see Wong and Wang [17]. After we identify all the statistically significant joint outcomes, the detected interdependencies as calculated from the function $I(\cdot)$ are summed, see Chiu and Lui [3, 5]; Chiu and Wong [4]. Note that the calculation is not based on the corresponding variables, but summing the individual values that are interdependent.

Consider the joint outcome of $X_i = a_{ip}$ and one of some other outcomes, say $X_j = a_{jq}$. The total interdependency for X_i at position i is calculated by a function $FD'(X_i)$. It is expressed as the summation of interdependency of all the values with $X_i = a_{ip}$. It is defined as

$$FD'(X_i) = \sum_{p=1}^{L_i} S(X_i = a_{ip}). \tag{3}$$

The function $S(\cdot)$ is defined as

$$S(X_i = a_{ip}) = \sum_{j=1,\, j\neq i} \sum_{q=1}^{L_j} I(X_i = a_{ip}, X_j = a_{jq}) \tag{4}$$

assuming that $(X_{i=}a_{ip}, X_j = a_{jq})$ is statistically significant.

$S(\cdot)$ is the calculated interdependency of a_{ip} (an outcome of the variable X_i as defined at position i on the aligned sequences) to the associated values in all other positions (as enumerated by the index j). It is formulated as the sum of the self-mutual information between the values, $(X_i = a_{ip}, X_j = a_{jq})$, provided that the interdependency calculated is statistically significant Chiu and Lui, see [3, 5]. Note that the summation represents the total significant interdependency of the sequences on the value a_{ip}, an outcome of X_i, and ignoring the other outcomes of X_i that are not interdependent. The objective is to give a measurement to account for the significant interdependency of the whole molecule at that point as defined by the value a_{ip}. It can be said that if the interdependency effect is known to occur at only some local neighborhood, then the enumeration of the index j can be restricted by a local window. However in general, the computation can be applied to the whole sequence.

The self-mutual information $I(X_i = a_{ip}, X_j = a_{jq})$ is defined in the usual way as

$$I(X_i = a_{ip}, X_j = a_{jq}) = \log\left(\frac{\text{prob}\,(X_i = a_{ip}, X_j = a_{jq})}{\text{prob}\,(X_i = a_{ip})\,\text{prob}\,(X_j = a_{jq})}\right). \tag{5}$$

Interdependence pattern calculated using $FD'(\cdot)$ is then based on summing the detected significant interdependency of $S(\cdot)$ of all the outcomes a_{ip} of the variable X_i. In other words, the calculation of $FD'(\cdot)$ represents the interdependencies at the position i on the aligned sequences. Since all the positions are calculated equally, the summation of the self-mutual information is calculated without weight.

Statistical significance of interdependency between joint values $(X_i = a_{ip}, X_j = a_{jq})$ can be evaluated in many ways. We use the following method.

Let $e = (X_i = a_{ip}, X_j = a_{jq})$ be the interdependence pattern between $X_i = a_{ip}$ and $X_j = a_{jq}$. The standardized residual $z(e)$ is defined as (see Haberman [18], Wong and Wang [17])

$$z(e) = \frac{\text{obs}(e) - \exp(e)}{\sqrt{(\nu \exp(e))}}, \tag{6}$$

where obs(e) is the observed frequency from the data ensemble and exp(e) is the expected frequency calculated from a prior model, usually based on the independence assumption. The statistics $z(e)$ has an asymptotic standard normal distribution and has a variance estimated by ν. The parameter ν can be estimated as

$$\nu = 1 - \text{prob}\,(X_i = a_{ip})\,\text{prob}\,(X_j = a_{jq}). \tag{7}$$

Thus $X_i = a_{ip}$ and $X_j = a_{jq}$ are significantly interdependent between them if $z(e) > \varepsilon(\alpha)$, where $\varepsilon(\alpha)$ is the tabulated value given a confidence level α. The expected frequency can be calculated from the marginal frequencies of $X_i = a_{ip}$ and $X_j = a_{jq}$. Note that the statistics $z(e)$ evaluates the statistical interdependency between the two values rather than their corresponding variables. It is based on a single entry in the contingency table rather than from the whole table. This is to disregard outcomes of the variable that may not be associated.

Assuming a high interdependency is distinguishable from those with a low one, we label X_i from the values of $FD'(X_i)$ using a threshold, taken as zero. For a small sample size, the threshold can be chosen to be higher, identified from the histogram distribution of the calculations from all the sites. For those points that have a calculated $FD'(\cdot)$ value higher than the threshold, then the position i of the aligned sequences is considered as expressing an interdependent pattern.

With these measures of conserved and interdependent patterns defined, the units of the aligned sequences can then be classified into one of the four statistical variation patterns as I-, C-, D-, or V-pattern type.

3.3. Sequence segmentation

Consider that a biosequence can be divided into regions based on the significant statistical variation pattern of each sequence unit from an aligned sequence ensemble. The segmentation has the following desirable properties.

(i) Each region is composed of contiguous neighboring sites, the majority of which have the same site pattern.
(ii) Adjacent regions may overlap with a common segment from the region boundaries.
(iii) Gaps between adjacent regions are allowed. That is, the start point of a region is not necessarily adjacent to the end point of the previous region. Similarly, the end point of a region may not be adjacent to the start point of the next region.
(iv) Some contiguous sites can be ignored if these sites do not form regions.
(v) Region length can vary and is not fixed. However, a minimum length can be imposed.

These properties are intended to be general, allowing flexibility in the segmentation process. Computationally, the optimal segmentation can be difficult to obtain. We use a heuristic algorithm similar to that by Zhang in [11] and described in more detail by Yan in [10] .

3.4. A segmentation algorithm

In order to identify multipattern consensus regions, we proposed the following segmentation algorithm. This algorithm takes the sequence and the detected statistical variation pattern of each site from the alignment as inputs. The algorithm outputs the sequence with the detected regions. The segmentation algorithm is composed of five phases.

In phase 1, regions are initiated based on the majority pattern type. A window of size w is moved along the sequence. For each window position, we count the number of sites for each type in that window, and find the pattern type with the maximum number of sites. The segment in the window is initiated as a region if the number of sites of the majority type is sufficiently large.

In phase 2, we merge adjacent regions detected if a statistical test of independence cannot distinguish between the regions based on their pattern types detected, see Kalbfleisch [19]; Haberman [18]. In this case, the distance between adjacent regions on the sequence needs to be sufficiently small. After phase 2, the boundaries of regions are tentatively determined.

Next, we identify the pattern type for the detected regions. In phase 3, we determine the type of each region based on the majority pattern type within that region. For each region, we count the number of sites for each pattern type, and find the type with the maximum count. Then the region is labeled according to that type.

In phases 4 and 5, we refine the boundaries and pattern types of regions. If the adjacent regions are of the same type and the gap between them is sufficiently small, we reapply a statistical test (see Wong and Wang [17]; Haberman [18]) on these two regions. The regions are merged if the statistical test fails to distinguish between them. In phase 5, the region boundaries are refined by removing sites adjacent to the boundaries whose type is different from the region type.

The segmentation algorithm is summarized as follows.

(1) Initiate regions based on high frequency count of a majority pattern in an observation window.
(2) Merge adjacent regions based on region length, statistical test of independence, and the size of gap between regions.
(3) Determine the region type according to the majority pattern type.
(4) Refine boundaries and pattern type of regions.

Applying the segmentation algorithm, sequences can be segmented based on the detected patterns. Even though not all the region types can be observed in a sequence, the four possible types are (1) mostly invariant; (2) mostly conserved; (3) mostly interdependent, and (4) mostly hypervariant.

4. EXPERIMENTAL EVALUATION

Our proposed method is tested on a dataset consisting of p53 protein sequences, known to be a tumor suppressor, taken from NCBI database and Protein Data Bank, EBI, see Berman et al. [20]. It is understood that p53 participates in the repairing of damaged DNA, and thus preventing the occurrence of cancers. Mutant p53 has lost these activities, leading to possible malignant transformation in cancers, see Hollstein et al. [21]; Levine et al. [22]; Levine [23]. It is found that p53 is frequently mutated in about 45%–50% in all types of cancers, see Hollstein et al. [21]; Greenblatt et al. [6]. In the experiments, p53 protein sequences from 31 species are retrieved from the SWISS-PROT database, see Boeckmann et al. [24, Figure 4]. These sequences are then aligned using ClustalW program version 1.8 [BCM Search Launcher: Multiple Sequence Alignments].

4.1. Identifying pattern type for each aligned site of the sequences

This experiment identifies the statistical variation patterns on each aligned position of the p53 sequences. First, we calculate the composition redundancy ($R^{(1)}$) and interdependency (FD') for each aligned position. From the histograms of the composition redundancy ($R^{(1)}$) and the interdependency (FD'), we identify the threshold as 0.57 and 600, respectively. Then, we label each site of the molecular sequence according to whether it is above or below the threshold.

Using this criterion, 86I-patterns, 55C-patterns, 188D-patterns, and 75V-patterns are identified. Since conservation and interdependence characteristics are not mutually exclusive, we found 11 patterns that can be classified into both types of C- and D-patterns.

4.2. Identify segmented regions

In this experiment, we segment the p53 sequence into regions based on the majority of the pattern types. The segmentation

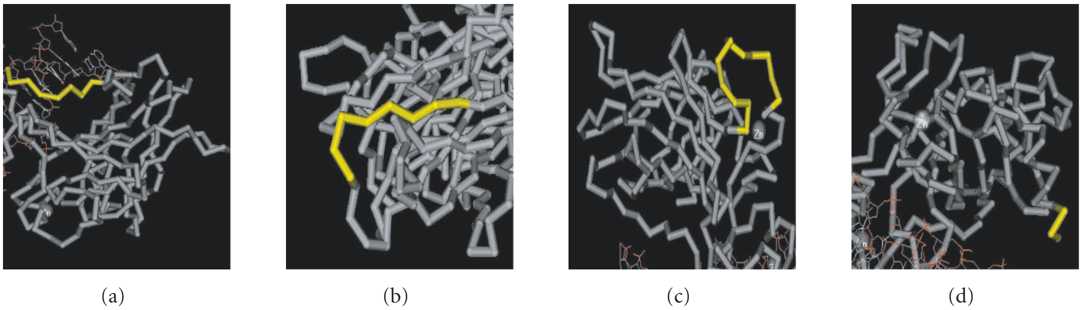


Figure 1: The four identified D-regions (sites 94–101, 143–150, 181–192, 287–289) in the core domains are shown in yellow and are at the exterior of the molecule.

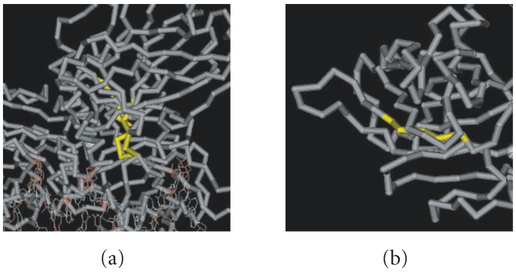


Figure 2: The two V-regions (sites 162–174, 232–236 shown in yellow) of the core domain are buried in the interior.

algorithm is then applied. Eighteen regions are identified (Figures 1, 2, and 3). Some adjacent regions have overlapping regions. Gap exists between some regions.

The result shows that the positions of the p53 sequences form clear regions. There are 7 D-regions, 5 I-regions, and 6 V-regions. The D-regions and the V-regions are mostly located at both terminals of the sequence. The 3 D-regions are located at the beginning of the sequence, and other 3 D-regions are located at the end of the sequence. The 3 V-regions are located at the beginning of the sequence, and 2 V-regions are located at the end of the sequence. The central domain of the sequence located between sites 170 and 280 is rich in I-regions. The C-patterns are isolated and do not form regions. The regions at the core domain are shown in Figures 1–3. The result shows that there are 4 D-regions (sites 94–101, 143–150, 181–192, 287–289), 5 I-regions (sites 172–179, 193–199, 215–223, 237–254, 265–282), and 2 V-regions (sites 162–174, 232–236) in the p53 core domain (sites 94 – –289). The sequences from the 4 D-regions are shown in Figure 4. The interdependency of the amino acids among the first 21 sequences, mostly among the higher animals, is clearly seen. The interdependency can go beyond the D-regions. Amino acids with low interdependency are screened out and do not contribute to the overall interdependency calculation in the equation.

4.3. *Multipattern consensus regions and molecular structure in P53*

We evaluate further our detected region patterns by comparing them to the three-dimensional structure of p53. The three-dimensional model is available from the National Center for Biotechnology Information (NCBI). In our experiment, we plot the identified regions in the core domain and analyze the relationship between these regions and the molecular structure. The three-dimensional-structure viewer software Cn3D is used in the plots.

All D-regions are located at the exterior and all I-regions and V-regions are buried inside the core domain (see Figures 1–3). This relationship is also observed in lysozymes (see Yan [10]) and cytochrome c (see Chiu and Wong [4]).

4.4. *Multipattern consensus regions and cancer patterns in P53*

It is known that the majority of the p53 mutations occur in the core domain, see Cho et al. [25]; Greenblatt et al. [6]; Hamroun et al. [26]. In this experiment, we evaluate the relationships between the mutations of the detected regions and different types of cancers at the core domain that contains sequence-specific DNA binding activity.

From the database of the International Agency for Research on cancer (IARC), we obtain records of cancer patients with observed p53 mutations. The version of collection we use contains 14050 records organized in 34 attributes, see Hamroun et al. [26]. The records include the location on the sequence where mutation occurs and the cancer type of the patients.

Comparing the locations when mutation occurs and the cancer type (Table 1), the mutated codons in I-regions are more likely to cause cancers in stomach, colon, rectum, liver and intrahepatic bile ducts, hematopoietic and reticuloendothelial systems, and nasopharynx. The mutated codons in D-regions are more likely to cause cancers in mouth, accessory sinuses, nasal cavity and middle ear, and head and neck. The mutated codons in V-regions are more likely to cause cancers in testis and breast.

Our results are compared to a study on hereditable factors causing cancers, see Magnusson et al. [13]; Lichtenstein et al. [12]. Our results (Table 1) show that the region patterns are significantly associated with cancers in stomach, colon, pancreas, lung, breast, cervix uteri, ovary, prostate gland, bladder, and hematopoietic and reticuloendothelial systems. The association between the region patterns and cancers in

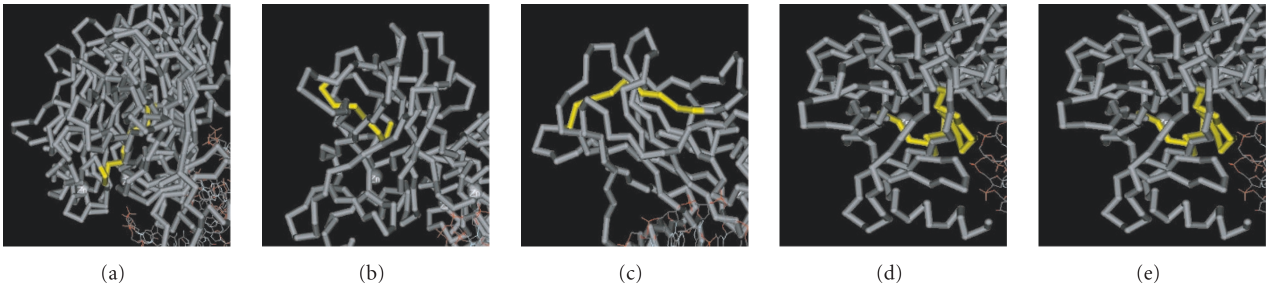


Figure 3: The 5 I-regions (sites 172–179, 193–199, 215–223, 237–254, 265–282 shown in yellow) of the core domain are buried in the interior.

Sequence code	D1	D2	D3	D4
p53_HUMAN	SSSVPSQK	VQLWVDST	RCSDSDGLAPPQ	ENL
p53_CERAE	SSSVPSQK	VQLWVDST	RCSDSDGLAPPQ	ENF
p53_MACFA	SSSVPSQK	VQLWVDST	RCSDSDGLAPPQ	ENF
p53_MACMU	SSSVPSQK	VQLWVDST	RCSDSDGLAPPQ	ENF
p53_CAVPO	SSSVPSHK	VQVWVESP	RCSDSDGLAPPQ	ENF
p53_CRIGR	SSSVPSYK	VQLWVNST	RSSEGDSLAPPQ	KNF
p53_MARMO	SSSVPSQN	VQLWVDST	RCSDSDGLAPPQ	ENF
p53_MESAU	SSSVPSYK	VQLWVSST	RSSEGDGLAPPQ	KNF
p53_MOUSE	SSFVPSQK	VQLWVSAT	RCSDGDGLAPPQ	ENF
p53_RAT	SSSVPSQK	VQLWVTST	RCSDGDGLAPPQ	ENF
p53_SPEBE	SSSVPSQN	VQLWVDST	RCSDSDGLAPPQ	ENF
p53_TUPGB	SSSVPSQK	VQLWVDSA	RCSDSDGLAPPQ	ENF
p53_CANFA	SSSVPSPK	VQLWVSSP	RCSDSDGLAPPQ	ENF
p53_CHICK	SPVVPSTE	VQVRVGVA	RCGGTDGLAPAQ	ENF
p53_FELCA	SSFVPSQK	VQLWVRSP	RCPDSDGLAPPQ	ENF
p53_RABIT	SSSVPSQK	VQLWVDST	RCSDSDGLAPPQ	ENF
p53_BOVIN	SSFVPSQK	VQLWVDSP	RSSDSDGLAPPQ	ENL
p53_EQUAS	—	VYLRISSP	RCSDSDGLAPPQ	ENF
p53_HORSE	SSFVPSQK	VQLLVSSP	RCSDSDGLAPPQ	ENF
p53_PIG	SSFVPSQK	VQLWVSSP	RSSDSDGLAPPQ	ENF
p53_SHEEP	SSFVPSQK	VQLWVDSP	RSSDSDGLAPPQ	ENF
p53_XENLA	SCAVPSTD	LLVRVESP	RSVEGEDAAPPS	DNY
p53_BARBU	TASVPVAT	VQMVVNVA	RTPD-DGLAPAA	SNF
p53_BRARE	TSTVPETS	VQMVVDVA	RTPD-DNLAPAG	SNF
p53_ICTPU	TSTVPVTS	VLMAVSSS	RSNDSDGPAPPG	SNF
p53_ORYLA	PTTVPVTT	IEVRVSKE	NEDS—VEHRS	ESR
p53_ONCMY	TSTVPTTS	VQIVVDHP	STSENEGPAPRG	INL
p53_PLAFE	SSTVPVVT	VEVLLSKE	TEDT—AEHRS	ESS
p53_TETMU	SPTVPVTT	VEVLLGKD	NEDS—AEHRS	TNS
p53_XIPMA	APTVPAIS	IGVLVKEE	SEDL—SDNKS	GNL
p53_XIPHE	APTVPAIS	IGVLVKEE	SEDL—SDNKS	GNL

Figure 4: The aligned sequences of the four D-regions: D1 (94–101), D2 (143–150), D3 (181–192), D4 (287–289). Note that some selected amino acids here are highly associated. Amino acids with low interdependency will be screened out. The association can go beyond the D-regions.

corpus uteri and cervix uteri is not significant. The comparison shows a strong correspondence among significant association between the region patterns and the cancers. This means that a significant association of the patterns with cancers also indicates a significant hereditable factors of cancers when human twins are followed. Because the current sequence's sample size is small, whether significant cancer association can be reflected by these detected patterns and the corresponding sites, should be evaluated further in the future.

5. DISCUSSIONS

The experiments show that multipattern consensus region generalizes previous notion of consensus sequence and is found to be useful in some sequence analysis problems. The

Table 1: Comparing results with hereditary studies of cancers in human twins.

Cancer type	I-regions		D-regions		V-regions		All regions	Hereditary factors
	Residual	α**	Residual	α	Residual	α		
Stomach	2.68	+ + +**	0.72		0.31		Significant	Significant
Colon	7.23	+ + +	−1.98	− −**	−3.34	− − −**	Significant	Significant
Pancreas	−0.8		0.01		−2.49	− −	Significant	significant
Lung	−3.78	− − −	−0.36		−0.04		Significant	Significant
Breast	−4.07	− − −	0.04		2.8	+ + +	Significant	Significant
Cervix uteri*	0.17		1.85		1.25		Not significant	Not Significant
Corpus uteri	0.61		0.37		0.19		Not Significant	Not significant
Ovary	2.29	++**	−2.31	− −	1.18		Significant	Significant
Prostate	−3.77	− − −	1.09		1.54		Significant	Significant
Bladder	−3.23	− − −	0.91		−1.71		Significant	Significant
Hematopoietic	3.61	+ + +	−3.07	− − −	0.23		Significant	Significant

* Cervix uteri was not found to be significant with hereditary factor according to Lichtenstein et al. [12] in human twins, but by Magnusson in et al. [13], a genetic link was found. We obtain a weak significant relationship ($\alpha > 90\%$) between the D-region and cervix uteri cancer. D-regions are all negatively associated with cancers when a significance relationship is found. Compared to a study we did earlier based on point relationships, the significance level is stronger, see Chiu et al. [27]. The result of D-regions is also consistent with that by Chiu and Lui in [5].

** α is the P-value indicating the significance level of association between the cancer type and the region type ("+" indicates a positive association and "−" a negative association. "+ + +" is above 99%; "++" is between 95% and 99%; "− − −" is below 1%; "−−" is between 1% and 5%).

experiments show that molecular sites in at least some protein biosequences can be classified meaningfully into region types.

In the experiments on region segmentation, comparisons between the detected region patterns and the three-dimensional structure of the molecule indicate a meaningful structural interpretation. I-regions are buried inside the interior of the biomolecule. This structural characteristic is possibly due to that these positions are invariant between species and are less affected. The D-regions are located at the exterior and affect the exterior shape of the molecule. These regions may play a more functional role in interactions between biomolecular processes as they relate between sites from one to another within the molecule.

Comparisons between the detected region patterns and the mutations in specific cancers also show significant correspondence that could be indicative of hereditable factors. Our method identifies the exact location in the molecule where the suggested correspondence may be traced.

6. CONCLUSION

In summary, it is possible that some sequences cannot be meaningfully segmented, that is, there is only one single segment in the whole sequence. In this paper, we have introduced the notion of multipattern consensus region in biosequence based on the statistical variation pattern of the aligned site in multiple sequences. It generalizes consensus sequence to incorporate interdependent characteristic, and thus provide a more flexible scheme to label statistical variations in multiple aligned sequences. The experimental results reveal that the multipattern consensus regions are well formed in p53. Comparing the region patterns and the structural characteristics, our detected consensus regions are associated with the molecular locations that are also related to mutations in different cancer types. Because ability to mutate can be related to genetic factors, their correspondence to hereditary study of cancers in human twins provides insights into a more specific indication of where in the molecule the hereditary effect might be reflected. Thus the experiments further support the notion that statistical variation patterns in sequence families can be indicative of their functionality at the very fine molecular level.

ACKNOWLEDGMENTS

This research is supported by the Discovery Grant of the NSERC of Canada and the Korea Research Foundation Grant (KRF-2004-042-C00020).

REFERENCES

[1] D. K. Y. Chiu and T. Kolodziejczak, "Inferring consensus structure from nucleic acid sequences," *Computer Applications in the Biosciences*, vol. 7, no. 3, pp. 347–352, 1991.

[2] D. K. Y. Chiu and G. Harauz, "A method for inferring probabilistic consensus structure with applications to molecular sequence data," *Pattern Recognition*, vol. 26, no. 4, pp. 643–654, 1993.

[3] D. K. Y. Chiu and T. W. H. Lui, "Integrated use of multiple interdependent patterns for biomolecular sequence analysis," *International Journal of Fuzzy Systems*, vol. 4, no. 3, pp. 766–775, 2002.

[4] D. K. Y. Chiu and A. K. C. Wong, "Multiple pattern associations for interpreting structural and functional characteristics of biomolecules," *Information Sciences*, vol. 167, no. 1–4, pp. 23–39, 2004.

[5] D. K. Y. Chiu and T. W. H. Lui, "A multiple-pattern biosequence analysis method for diverse source association mining," *Applied Bioinformatics*, vol. 4, no. 2, pp. 85–92, 2005.

[6] M. S. Greenblatt, W. P. Bennett, M. Hollstein, and C. C. Harris, "Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis," *Cancer Research*, vol. 54, no. 18, pp. 4855–4878, 1994.

[7] R. J. Boys and D. A. Henderson, "A Bayesian approach to DNA sequence segmentation," *Biometrics*, vol. 60, pp. 573–588, 2004.

[8] W. Li, P. Bernaola-Galván, F. Haghighi, and I. Grosse, "Applications of recursive segmentation to the analysis of DNA sequences," *Computers and Chemistry*, vol. 26, no. 5, pp. 491–510, 2002.

[9] D. K. Y. Chiu and G. Rao, "The 2-level pattern analysis of genome comparisons," *WSEAS Transactions on Biology and Biomedicine*, vol. 3, no. 3, pp. 167–174, 2006.

[10] W. Yan, "A segmentation algorithm for consensus regions in biosequences," M.S. thesis, Department of Computing and Information Science, University of Guelph, Guelph, Ontario, Canada, 2003.

[11] J. Zhang, "Analysis of information content for biological sequences," *Journal of Computational Biology*, vol. 9, no. 3, pp. 487–503, 2002.

[12] P. Lichtenstein, N. V. Holm, P. K. Verkasalo, et al., "Environmental and heritable factors in the causation of cancer: analyses of cohorts of twins from Sweden, Denmark, and Finland," *New England Journal of Medicine*, vol. 343, no. 2, pp. 78–85, 2000.

[13] P. K. E. Magnusson, P. Sparen, and U. B. Gyllensten, "Genetic link to cervical tumours," *Nature*, vol. 400, no. 6739, pp. 29–30, 1999.

[14] A. K. C. Wong, T. S. Liu, and C. C. Wang, "Statistical analysis of residue variability in cytochrome c," *Journal of Molecular Biology*, vol. 102, no. 2, pp. 287–295, 1976.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948, reprinted in C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Ill, USA, 1949.

[16] L. L. Gatlin, "The information content of DNA," *Journal of Theoretical Biology*, vol. 10, no. 2, pp. 281–300, 1966.

[17] A. K. C. Wong and Y. Wang, "High-order pattern discovery from discrete-valued data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 6, pp. 877–893, 1997.

[18] S. J. Haberman, "The analysis of residuals in cross-classified tables," *Biometrics*, vol. 29, pp. 205–220, 1973.

[19] J. G. Kalbfleisch, *Probability and Statistical Inference, Vol. 2: Statistical Inference*, Springer, New York, NY, USA, 2nd edition, 1985.

[20] H. M. Berman, J. Westbrook, Z. Feng, et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[21] M. Hollstein, D. Sidransky, B. Vogelstein, and C. C. Harris, "p53 mutations in human cancers," *Science*, vol. 253, no. 5015, pp. 49–53, 1991.

[22] A. J. Levine, J. Momand, and C. A. Finlay, "The p53 tumour suppressor gene," *Nature*, vol. 351, no. 6326, pp. 453–456, 1991.

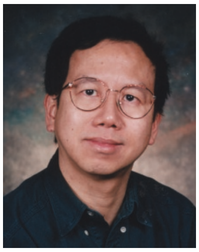
[23] A. J. Levine, "p53, the cellular gatekeeper for growth and division," *Cell*, vol. 88, no. 3, pp. 323–331, 1997.

[24] B. Boeckmann, A. Bairoch, R. Apweiler, et al., "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Research*, vol. 31, no. 1, pp. 365–370, 2003.

[25] Y. Cho, S. Gorina, P. D. Jeffrey, and N. P. Pavletich, "Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations," *Science*, vol. 265, no. 5170, pp. 346–355, 1994.

[26] D. Hamroun, S. Kato, C. Ishioka, M. Claustres, C. Beroud, and T. Soussi, "The UMD TP53 database and website: update and revisions," *Human Mutation*, vol. 27, no. 1, pp. 14–20, 2005.

[27] D. K. Y. Chiu, X. Chen, and 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 Technolgoy (CBGIST '01)*, pp. 64–69, Durham, NC, USA, March 2001.



David K. Y. Chiu is a Professor in the Department of Computing and Information Science and a graduate faculty in the Biophysics Interdepartmental Group at the University of Guelph, Ontario, Canada. He was a former recipient of the Science and Technology Agency (STA) Fellowship of Japan and a Visiting Researcher to Electrotechnical Laboratory (currently National Institute of Advanced Industrial Science and Technology) in Japan. He has been involved in the program committees of numeral conferences including AI, FLAIRS Uncertain Reasoning Track, International Conference on Computer Vision, Pattern Recognition and Image Processing, and he is the cochair of International Conference on Computational Biology and Genome Informatics in 2003 and 2005. He will be guest-editing a Special Issue on Bioinformatics in the journal Biomolecular Engineering. He is a Member of the International Advisory Board of Knowledge Engineering and Discovery Research Institute at the Auckland University of Technology.

Yan Wang received the M.S. degree in computing and information Science from the University of Guelph in Canada. During her study, she worked on developing computational methods to analyze biosequences. She received numerous scholarships, including the Ontario Graduate Scholarship. She was trained as an Ophthalmologist in China and was a Member of Chinese Medical Association. She has published in *Ophthalmology in China*. Currently, she is a Clinical Data Manager at MDS Pharma Services, MDS Inc.