Websites:
http://www.lifesciencesite.com
http://www.sciencepub.net

Emails:
editor@sciencepub.net
sciencepub@gmail.com

Life Science Journal

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

Linking of MLH1 Variants with Their Clinical Outcomes

Chunhong Zhang[1], Weisheng Ye[2]

[1]Department of Spinal Surgery III, [2]Department of Orthopedic Trauma, Tianjin Hospital, Tianjin, China
13132065700@163.com; yeweisheng@yahoo.com.cn

Abstract: In this study, we use the cross-impact analysis to establish a quantitative relationship between mutated primary structure of MLH1 protein and its clinical outcomes using the amino-acid distribution probability as a measure to determine the magnitude of changes in primary structure of MLH1 due to mutations. Then, we use the Bays' equation to calculate the probability of occurrence of long-QT syndrome under a new variant. Thereafter, we numerically compare altered protein functions with the help of amino-acid distribution probability. Finally, we use the amino-acid distribution probability to find the mutation trend in 155 variants. The results are not only meaningful for clinicians to have a concept on the possibility of occurring of long-QT syndrome when finding a new variant before any sophisticated and expensive tests, but also pave the ways for simulation of relationship between mutated primary structure of proteins and clinical outcome from molecular level.
[Zhang C, Ye W. **Linking of MLH1 Variants with Their Clinical Outcomes**. *Life Sci J* 2022;19(3):44-52]. ISSN 1097-8135 (print); ISSN 2372-613X (online). http://www.lifesciencesite.com. 7.doi:10.7537/marslsj190322.07.

Keywords: Bayes' law; cross-impact analysis; distribution probability; MLH1; variant

1. Introduction

Very likely, we should begin our studies on the genotype-phenotype relationship from determination of a relationship between a protein and a certain disease, and then we should strive the build a descriptively quantitative relationship between variant and its clinical outcome, because it is oftentimes that a mutation in a protein induces a disorder in clinical settings. With such quantitative models in hands, we may be able to predict a clinical outcome, which will be due to an undocumented mutation because biological evolution would generate new mutations along the time course. This would at least give a concept on what will result from a mutation.

At genetic level, human mismatch repair genes play an important role in DNA stability, and their inactivation leads to mutations that often loss the mismatch repair function [1]. Lynch syndrome, also known as hereditary non-polyposis colorectal cancer (HNPCC), is a dominant autosomal genetic disorder caused by germ line mutations in mismatch repair genes and the most common hereditary form of colorectal cancers, which is the second most leading cause of cancer related deaths in the western countries [2-5]. Therefore it is considered necessary to build a model that can quantitatively describe the relationship between mutations and clinical outcome.

For this reason, we need the methods, which can code a protein sequence as a numeric sequence. This can be achieved using a certain number to replace each amino acid in a protein, and generally the values representing physicochemical property of individual amino acid are the first option [6]. After such a replacement, a protein sequence becomes a numeric sequence useful for mathematical modeling. However, a limitation in such coding is that individual amino acid property does not reflect the characteristic of a whole protein and is unchangeable before and after mutation.

Over last decade, Doctors Wu and Yan have developed three approaches to quantify the characteristic of individual amino acid and characteristic of a whole protein [7-10], and their quantifications indeed differ before and after mutation, thus it is possible to use their approaches to build a quantitative relationship between mutated primary structure and changed function of protein.

2. Material and Methods

2.1 Data

The human MLH1 protein with its 155 variants (accession number P40692; update to July 11, 2012; sequence version 1, entry version 146) is obtained from UniProtKB/Swiss-Prot entry [11], of which 100 are missense variants, 2 are insertion and 13 are deletions.

2.2 Amino-Acid Distribution Probability

Amino-acid distribution probability is mainly related to the positions of amino acids along the protein, which is suitable for mutation analysis, and this approach has been used in some studies [7-10, 12-22].

For example, there are five tryptophans (W), which are the least abundant amino acids in MLH1. This protein is composed of 756 amino acids. Although it is very simple to determine the positions of these five tryptophans in MLH1 experimentally, it does not provide a quantitative measure that can be used for modeling. To estimate these five positions, the simplest way is to imagine dividing the MLH1 into five equal partitions, each contains about 151 amino acids (756/5 = 151.2). Then we may guess that each partition contains a W, which is one type of distribution for five tryptophans in MLH1. Naturally, we also may guess that one partition contains two tryptophans, three partitions contain one W, and one partition contains zero W, which is another type of distribution for five tryptophans in MLH1. Theoretically, there are totally seven different distributions for five tryptophans in five partitions in MLH1, and we can mathematically calculate the probability for each type of distribution of these five tryptophans (Table 1) according to the problem of subpopulations and partitions using the following equation [23], where r is the number of amino acids, n is the number of partitions, rn is the number of amino acids in the n-th partition, qn is the number of partitions with the same number of amino acids, and ! is the factorial. In the real world, there is only one possible distribution for these five tryptophans in MLH1, which is the actual distribution, whose probability is the actual distribution probability.

2.3 Amino-Acid Distribution Probability in MLH1 Variant

Theoretically the amino-acid distribution probability can be referred to the statistical mechanics, which classifies the distribution of elementary particles in energy states according to three assumptions of whether distinguishing each particle and energy state, i.e. Maxwell-Boltzmann, Fermi-Dirac and Bose-Einstein assumptions [23]. The approach used here is equivalent to the Maxwell-Boltzmann assumption.

Table 1. All possible distributions of five tryptophans (W) in 5 partitions of MLH1

1	2	3	4	5	Probability	Rank
W	W	W	W	W	0.0384	5
	W	W	W	WW	0.3840	1
		W	WW	WW	0.2880	2
		W	W	WWW	0.1920	3
			WW	WWW	0.0640	4*
			W	WWWW	0.0320	6
				WWWWW	0.0016	7

*, the actual distribution of five tryptophans in MLH1.

Generally, a variant is related to two types of amino acids, the original and mutated amino acids. In the cases of insertion or deletion, at least one type of amino acids would be affected. As each type of amino acids is associated with a certain distribution probability, a variant will change the distribution probability of relevant type of amino acids, that is, the amino-acid distribution probability differs before and after mutation so it is sensitive to mutation.

Table 2. Distribution pattern and probability of tryptophans and arginines in wild-type MLH1 and in its W666R variant

Partition	Wild-type		W666R variant	
	W	R	W	R
I	0	3	0	3
II	0	1	0	1
III	0	0	1	0
IV	2	2	3	2
V	3	1		1
VI		0		0
VII		1		1
VIII		2		2
IX		1		1
X		1		1
XI		2		2
XII		0		0
XIII		2		2
XIV		0		0
XV		0		0
XVI		1		1
XVII		1		1
XVIII		0		0
XIX		2		2
XX		0		0
XXI		2		2
XXII		1		1
XXIII		3		3
XXIV		4		4
XXV		1		1
XXVI		0		0
XXVII		0		0
XXVIII		1		1
XXIX		0		0
XXX		0		0
XXXI		0		0
XXXII		1		2
XXXIII		1		1
XXXIV		0		0
XXXV		1		1
XXXVI		1		1
XXXVII				0
Probability	0.0640	0.0301	0.1875	0.0297

W, Tryptophan; R, Arginine.

For example, a variant at position 666 changes tryptophan (W) to arginine (R) [24]. In above subsection, we have calculated the distribution probability of tryptophans (Table 1) before mutation, and now we show the calculation of their distribution probability after mutation. There are five tryptophans in wild-type MLH1, while there are 4 tryptophans in the variant (Table 2), for which we have $q_0 = 2$, $q_1 = 1$, $q_2 = 0$, $q_3 = 1$; and $r_1 = 0$, $r_2 = 0$, $r_3 = 1$, $r_4 = 3$, i.e. $\frac{4!}{2!\times 1!\times 0!\times 1!}\times\frac{4!}{0!\times 0!\times 1!\times 3!}\times 4^{-4} = 0.1875$. However, its distribution probability is 0.0640 before mutation, so this variant increases the distribution probability of tryptophans. On the other hand, there are 26 arginines in wild-type MLH1 and 37 arginines in the variant. Their distribution probabilities are 0.0301 and 0.0297 before and after mutation, so the variant decreases the distribution probability of arginines. The overall effect for this

variant on MLH1 is (0.1875 – 0.0640) + (0.0297 – 0.0301) = 0.1231, that is, the variant increases the distribution probability for MLH1.

Clearly, a variant changed the composition of amino acids, which can change the distribution pattern of corresponding amino acids in the protein, and consequently the distribution probability changed too. Table 3 lists the amino acid compositions and their distribution probability in wild-type human MLH1 and the variant that amino acids RVQ at positions 325-327 are missing. As can be seen, only 3 amino acids were missing in the variant, however, the distribution probability changed in 13 out of 20 kinds of amino acids. Hence, the distribution probability is sensitive to variants, and can serve as a measure for modeling.

This way, we have the quantitative measure for the mutated primary structure of MLH1 variants and we also have their documented clinical manifestations, therefore we can build a descriptively quantitative relationship between mutated primary structure and its clinical outcome.

2.4 Probabilistic Relationship

With the descriptively probabilistic method, we build the quantitative relationship between variant and clinical outcome. Our measure was amino-acid distribution probability and each individual variant related to its clinical outcome was presented as frequency, which we coupled by means of the cross-impact analysis, because the amino-acid distribution probability either increased or decreased after mutation, which was a 2-possibilty event, and the clinical outcome either occurred or did not occur after mutation, which was a yes-and-no event. Thereafter, we used the Bayesian equation to calculate the probability of occurrence of clinical outcome under mutation.

3. Results and Discussions

After above computation, we have the amino-acid distribution probability in wild-type human MLH1 and changed amino-acid distribution probabilities in its 155 natural variants, among which 138 variants are documented with colorectal cancer and 125 belong to the hereditary non-polyposis colorectal cancer (HNPCC), two variants relate to endometrial cancer, four variants relate to gastric cancer, while 13 variants are polymorphism. Thus, we can use the cross-impact analysis to build a quantitative relationship between the increase/decrease of distribution probability after mutations and corresponding clinical outcomes, because the cross-impact analysis is particularly suited for two relevant events coupled together [25-31].

Table 3. Amino acids, their compositions and distribution probability in wild-type human MLH1 and the variant that amino acids RVQ at positions 325-327 are missing

Amino acid	Number		Distribution probability	
	Wild-type	Variant	Wild-type	Variant
A	50	50	0.0026	0.0026
R	36	35	0.0301	0.0314
N	33	33	0.0506	0.0371
D	34	34	0.0416	0.0079
C	11	11	0.1010	0.1010
E	70	70	0.0100	0.0034
Q	30	29	0.0296	0.0147
G	40	40	0.0281	0.0137
H	19	19	0.0064	0.0064
I	48	48	0.0163	0.0114
L	76	76	0.0048	0.0069
K	47	47	0.0267	0.0225
M	14	14	0.0618	0.0618
F	26	26	0.0170	0.0363
P	34	34	0.0175	0.0175
S	69	69	0.0003	0.0000
T	45	45	0.0085	0.0060
W	5	5	0.0640	0.0640
Y	23	23	0.0712	0.0712
V	46	45	0.0174	0.0116

A, alanine; R, arginine; N, asparagine; D, aspartic acid; C, cysteine; E, glutamic acid; Q, glutamine; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; M, methionine; F, phenylalanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine.

Fig. 1 displays the cross-impact analysis on the relationship between changed amino acid distribution probabilities of MLH1 variants and their corresponding clinical outcomes. At the level of amino-acid distribution probability, *P(2)* and $P(\bar{2})$ are the decreased and increased probabilities induced by variants, and 102 and 53 variants result in the distribution probability decreased and increased, respectively. At the level of clinical outcome: (i) $P(1|\bar{2})$ is the impact probability (conditional probability) that cancer is diagnosed under the condition of increased distribution probability, and 46 variants have such an effect. (ii) $P(\bar{1}|\bar{2})$ is the impact probability that polymorphism is diagnosed under the condition of increased distribution probability, and 7 variants work in such a manner. (iii) *P(1|2)* is the impact probability that cancer is diagnosed under the condition of decreased distribution probability, and 96 variants play such a role. (iv) $P(\bar{1}|2)$ is the impact probability that polymorphism is diagnosed under the condition of decreased distribution probability, and 6 variant falls into this category. At the level of combined events, we can see the combined results of changed amino acid distribution probabilities and their corresponding clinical outcomes.

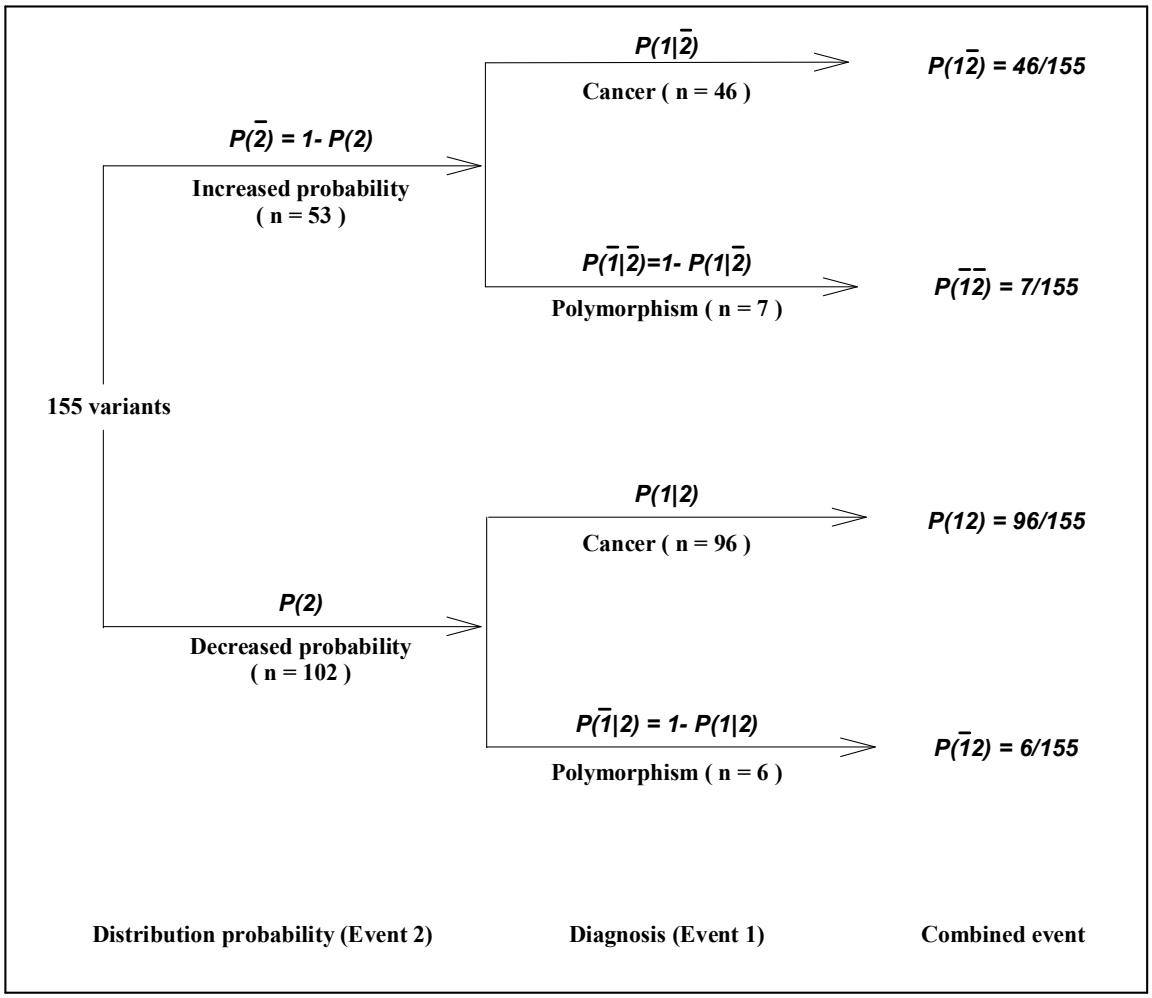

Figure 1. Cross-impact relationship among MLH1 variants, changed amino-acid distribution probability, and clinical outcomes.

Table 4. Computation on cross-impact analysis in Fig. 1

Event 2	$P(2) = 102/155 = 0.6581$
	$P(\bar{2}) = 1 - P(2) = 1 - 0.6581 = 0.3419 = 53/155$
Event 1	$P(1 \mid \bar{2}) = 46/53 = 0.8679$
	$P(\bar{1} \mid \bar{2}) = 1 - P(1 \mid \bar{2}) = 1 - 0.8679 = 0.1321 = 7/53$
	$P(1\mid 2) = 96/102 = 0.9412$
	$P(\bar{1} \mid 2) = 1 - P(1\mid 2) = 1 - 0.9412 = 0.0588 = 6/102$
Combined event	$P(1\bar{2}) = P(1 \mid \bar{2}) \times P(\bar{2}) = 46/53 \times 53/155 = 46/155 = 0.2968$
	$P(\bar{1}\bar{2}) = P(\bar{1} \mid \bar{2}) \times P(\bar{2}) = 7/53 \times 53/155 = 7/155 = 0.0452$
	$P(12) = P(1\mid 2) \times P(2) = 96/102 \times 102/155 = 96/155 = 0.6194$
	$P(\bar{1}2) = P(\bar{1} \mid 2) \times P(2) = 6/105 \times 102/155 = 6/155 = 0.0387$

Table 4 lists the computed probabilities according to the data in Fig. 1, from which two points can be found. (i) As $P(2)$ is larger than $P(\bar{2})$, a MLH1variant has a larger chance of decreasing the distribution probability. (ii) As the ratio is about 6 for $P(1 \mid \bar{2})$ versus $P(\bar{1} \mid \bar{2})$ and $P(1\mid 2)$ versus $P(\bar{1} \mid 2)$, a MLH1 variant has remarkably larger chance of causing cancer no matter its effect on the distribution probability.

The diagnosis of Lynch syndrome is complicated by the absence of a pre-morbid phenotype and germline mutation analysis is expensive and time consuming [32]. However, a variety of techniques have been used to screen molecular alterations at the DNA, RNA and protein level [33]. MLH1 methylation testing can identify the defect in MLH1 by denature gradient gel electrophoresis [34], immunohistochemistry [35]. Emerging technologies may substantially help the diagnosis in the future with high-throughput sequencing [36].

The Bayes' law [37], $P(1 \mid 2) = P(2 \mid 1)\frac{P(1)}{P(2)}$, indicates the probabilities of occurrences of two events. By now we can use it to determine the probability that the cancer occurs under a MLH1variant, $P(1)$, because $P(2)$ and $P(1 \mid 2)$ have already been defined in cross-impact analysis, while $P(2 \mid 1)$ is the probability that the distribution probability decreases under the condition of cancer.

As $P(1\mid 2) = 96/102 = 0.9412$ (Table 4), and $P(2 \mid 1) = 96/(46 + 96) = 0.6761$, $P(1) = \frac{P(1 \mid 2)}{P(2 \mid 1)} P(2) = \frac{0.9412 \times 0.6581}{0.6761} = 0.9161$, namely, the patient has a larger than 0.9 chance of being cancer when a new variant is found in MLH1. This is remarkable, because this estimate is very meaningful for clinical settings, because the clinicians would have a concept on the possibility of occurring cancer when finding a new variant before any sophisticated and expensive tests.

Indeed the analysis can furthermore go from clinical outcomes of mutations to functions of MLH1 variants. Of 155 MLH1 variants, only 44 are documented DNA mismatch repair function, which ranges from normal to complete lost. Fig. 2 displays the relationship between changed function for DNA mismatch repair and changed amino-acid distribution probability in MLH1 variants with statistical comparison. As can be seen, the decreased and lost functions in DNA mismatch repair are generally associated with the decrease in amino-acid distribution probability, while a normal function in DNA mismatch repair is generally associated with the increases in amino-acid distribution probability. Thus, the amino-acid distribution probability can be viewed as a measure that is directly linked to protein function. The higher the amino-acid distribution probability is, the higher the function protein has. Therefore the variants that decrease the amino-acid distribution probability are highly likely to reduce

the function. Actually, this amino-acid distribution probability also provides a way to statistically compare protein with its variants.

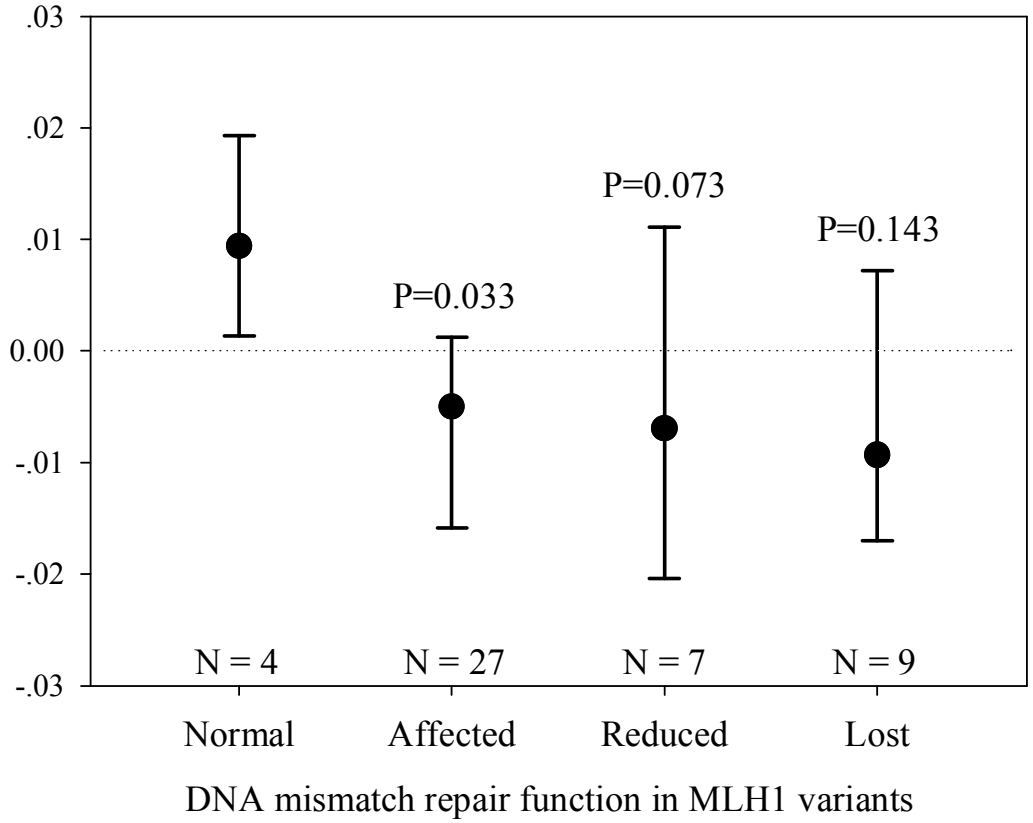

Figure 2. Comparison of changes between DNA mismatch repair function and amino-acid distribution probability in MLH1 variants. The data are presented as median with interquartile. The P values are obtained from Mann-Whitney rank sum test compared with normal group.

In this study, we use the distribution probability of amino acids as a measure to quantify 155 MLH1 variants, and then link their changes to clinical outcomes. The results shed lights on quantitative relationship between mutated primary structure and changed function of MLH1 protein, which paves the way for quantitative diagnosis of genetic disorder.

Corresponding Author:
Dr. Weisheng Ye
Department of Orthopedic Trauma
Tianjin Hospital
406 Jiefangnanlu Road, Hexiqu District, Tianjin, 300211, China.
E-mail: yeweisheng@yahoo.com.cn

References

[1]. Hitchins, M.P. Inheritance of epigenetic aberrations (constitutional epimutations) in cancer susceptibility. Adv. Genet., 2010, 70, 201-243.
[2]. Geiersbach, K.B.; Samowitz, W.S. Microsatellite instability and colorectal

cancer. Arch. Pathol. Lab. Med., 2011, 135, 1269-1277.
[3]. Rybak, C.; Hall, M.J. Interpretation of genetic testing for lynch syndrome in patients with putative familial colorectal cancer. J. Natl. Compr. Canc. Netw., 2011, 9, 1311-1320.
[4]. Hassen, S.; Ali, N.; Chowdhury, P. Molecular signaling mechanisms of apoptosis in hereditary non-polyposis colorectal cancer. World J. Gastrointest. Pathophysiol., 2012, 3, 71-79.
[5]. Parsons, M.T.; Buchanan, D.D.; Thompson, B.; Young, J.P.; Spurdle, A.B. Correlation of tumour BRAF mutations and MLH1 methylation with germline mismatch repair (MMR) gene mutation status: a literature review assessing utility of tumour features for MMR variant classification. J. Med. Genet., 2012, 49, 151-157.
[6]. Chou, K.C. Structure bioinformatics and its impact to biomedical science. Curr. Med. Chem., 2004, 11, 2105-2134.
[7]. Wu, G.; Yan, S. Lecture Notes on Computational Mutation. Nova Science Publishers, New York, 2008
[8]. Wu, G.;. Yan, S Randomness in the primary structure of protein: methods and implications. Mol. Biol. Today, 2002, 3, 55-69.
[9]. Wu, G.; Yan, S. Mutation trend of hemagglutinin of influenza A virus: a review from computational mutation viewpoint. Acta Pharmacol. Sin., 2006, 27, 513-526.
[10]. Yan, S.M.; Wu, G. Creation and application of computational mutation. J. Guangxi Acad. Sci., 2010, 26, 130-139.
[11]. The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res., 2010, 38, D142-D148.
[12]. Gao, N.; Yan, S.; Wu, G. Pattern of positions sensitive to mutations in human haemoglobin -chain. Protein Pept. Lett., 2006, 13, 101-107.
[13]. Yan, S.; Wu, G. Quantitative relationship between mutated amino-acid sequence of human copper-transporting ATPases and their related diseases. Mol. Divers., 2008, 12, 119-129.
[14]. Yan, S.; Wu, G. Connecting mutant phenylalanine hydroxylase with phenylketonuria. J. Clin. Monit. Comput. 2008, 22, 333-342.
[15]. Yan, S.; Wu, G. Quantitative relationship between mutated structure of human glucosylceramidase and Gaucher disease status. Int. J. Pept. Res. Ther. 2008, 14, 263-271.
[16]. Wu, G.; Yan, S. Building quantitative relationship between changed sequence and changed oxygen affinity in human hemoglobin -chain. Protein Pept. Lett., 2008, 15, 341-345.
[17]. Yan, S.; Wu, G. Descriptively probabilistic relationship between mutated primary structure of coagulation factor IX and clinical severity of hemophilia B. J. Appl. Res. 2009, 9, 100-106.
[18]. Yan, S.; Wu, G. Descriptively quantitative relationship between mutated N-acetylgalactosamine-6-sulfatase and mucopolysaccharidosis IVA. Biopolymers: Pept. Sci. 2009, 92, 399-404.
[19]. Yan, S.; Wu, G. Quantitative coupling of human antithrombin III mutations with their clinical outcomes. J. Guangxi Acad. Sci. 2009, 25, 183-186.
[20]. Yan, S.; Wu, G. Descriptively probabilistic relationship between mutated primary structure of von Hippel-Lindau protein and its clinical outcome. J. Biomed. Sci. Eng. 2009, 2, 117-122.
[21]. Yan, S.; Wu, G. Connecting KCNQ1 mutants with their clinical outcomes. Cli.n Invest. Med. 2009, 32, E28-E32.
[22]. Yan, S.; Wu, G. Linking mutated structure of adrenoleukodystrophy protein with X-linked adrenoleukodystrophy. Computer Methods Biomechan Biomed. Eng. 2010, 3, 403-411.
[23]. Feller, W. An Introduction to Probability Theory and Its Applications. 3rd ed, Vol, I. Wiley, New York, (1968) pp. 34-40.
[24]. Tournier, I.; Vezain, M.; Martins, A.; Charbonnier, F.; Baert-Desurmont, S.; Olschwang, S.; Wang, Q.; Buisine, M.P.; Soret, J.; Tazi, J.; Frebourg, T.; Tosi, M. A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum. Mutat., 2008, 29, 1412-1424.
[25]. Gordon, T.G. Cross-impact matrices – an illustration of their use for policy analysis. Futures, 1969, 2, 527-531.
[26]. Gordon, T.G.; Hayward, H. Initial experiments with the cross-impact matrix method of forecasting. Futures, 1968, 1, 100-116.
[27]. Enzer, S. Delphi and cross-impact techniques: an effective combination for

systematic futures analysis. Futures, 1970, 3, 48-61.
[28]. Enzer, S. Cross-impact techniques in technology assessment. Futures, 1970, 4, 30-51.
[29]. Sage, A.P. Methodology for Large-Scale Systems. McGraw-Hill, New York, (1977), pp. 165-203.
[30]. Wu, G. Application of cross-impact analysis to the relationship between aldehyde dehydrogenase 2 and flushing. Alcohol Alcohol., 2000, 35, 55-59.
[31]. Wu, G.; Yan, S. Prediction of mutation trend in hemagglutinins and neuraminidases from influenza A viruses by means of cross-impact analysis. Biochem. Biophys. Res. Commun., 2005, 326, 475-482.
[32]. van Lier, M.G.; Wagner, A.; van Leerdam, M.E.; Biermann, K.; Kuipers, E.J.; Steyerberg, E.W.; Dubbink, H.J.; Dinjens, W/N. A review on the molecular diagnostics of Lynch syndrome: a central role for the pathology laboratory. J. Cell. Mol. Med., 2010, 14, 181-197.
[33]. Pineda, M.; González, S.; Lázaro, C.; Blanco, I.; Capellá, G. Detection of genetic alterations in hereditary colorectal cancer screening. Mutat. Res., 2010, 693, 19-31.
[34]. Liu, T. Mutational screening of hMLH1 and hMSH2 that confer inherited colorectal cancer susceptibility using denature gradient gel electrophoresis (DGGE). Methods Mol. Biol., 2010, 653, 193-205.
[35]. Barrow, E.; Evans, D.G.; McMahon, R.; Hill, J.; Byers, R. A comparative study of quantitative immunohistochemistry and quantum dot immunohistochemistry for mutation carrier identification in Lynch syndrome. J. Clin. Pathol., 2011, 64, 208-214.
[36]. Geiersbach, K.B.; Samowitz, W.S. Microsatellite instability and colorectal cancer. Arch. Pathol. Lab. Med., 2011, 135, 1269-1277.
[37]. Wikipedia -- the free encyclopedia, Bayes' theorem. en.wikipedia.org/wiki/Bayes'_theorem. 2012.

1/3/2022