计算机科学
COMPUTER SCIENCE

蓝舌病毒基因组序列多元概率特征可视化分析

陈慧嫔，王琨，杨恒，郑智捷

引用本文

陈慧嫔，王琨，杨恒，郑智捷. 蓝舌病毒基因组序列多元概率特征可视化分析[J]. 计算机科学, 2022, 49(6A): 27-31.

CHEN Hui-pin, WANG Kun, YANG Heng, ZHENG Zhi-jie. Visual Analysis of Multiple Probability Features of Bluetongue Virus Genome Sequence[J]. Computer Science, 2022, 49(6A): 27-31.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300129

蓝舌病毒基因组序列多元概率特征可视化分析

陈慧嫔[1]　王　琨[1]　杨　恒[2]　郑智捷[1]
1 云南大学软件学院　昆明 650091
2 云南省亚热带动物病毒重点实验室　昆明 650224
(chuipin@foxmail.com)

摘　要　由于基因序列决定基因结构,基因结构反应生物功能性状,对病毒基因序列进行科学的数据可视化成为广泛应用的方法之一,对生物的基因序列进行可视化操作的需求日益增加。由此提出以最先进的生物信息分析学和分层结构化生物信息知识模型为基础,采用多元概率测度的方式对蓝舌病毒基因序列进行统计分析,结合计算机可视化的方法呈现蓝舌病毒在不同投影下的特征。与传统的病毒研究方法相比,此方法直观简洁,应用难度小。在不同的测量坐标下提供丰富的可视化呈现可反映蓝舌病毒的分类特征,将所提方法生成的结果与传统生物分析方法生成的系统发育树结果进行对照,可为同源性分析以及蓝舌病毒进化关系的研究提供参考,利于从多种角度深入研究蓝舌病毒。

关键词:蓝舌病毒;概率统计;数据可视化;基因序列
中图法分类号　TP391

Visual Analysis of Multiple Probability Features of Bluetongue Virus Genome Sequence

CHEN Hui-pin[1],WANG Kun[1],YANG Heng[2] and ZHENG Zhi-jie[1]
1 School of Software,Yunnan University,Kunming 650091,China
2 Yunnan Tropical and Subtropical Animal Virus Diease Laboratory,Yunnan Animal Science and Veterinary Institute,Kunming 650224,China

Abstract　The sequence of a gene determines its structure,and the structure of a gene reflects biological functional traits. Therefore,scientific data visualization of viral gene sequences has become one of the widely used methods. There is an increasing demand for visual manipulation of biological gene sequences. Therefore,based on the most advanced bioinformatics analysis and hierarchical structured bioinformatics knowledge model,the method of multiple probability measures is proposed to statistically analyze the bluetongue virus gene sequence,and combined with computer visualization methods,the characteristics of the bluetongue virus under different projections are presented. Compared with traditional virus research methods,this method is intuitive and concise,and it is easy to use. This method provides rich visualization under different measurement coordinates,reflecting the classification characteristics of bluetongue virus. Results generated by this method are compared with the phylogenetic tree generated by traditional biological analysis methods,which could provide references for homology analysis and the study of the evolutionary relationship of bluetongue virus. It is conducive to in-depth study of bluetongue virus from various angles.

Keywords　Bluetongue virus,Probability statistics,Data visualization,Gene sequence

1　引言

蓝舌病(Bluetongue,BT)是由蓝舌病病毒(Bluetongue Virus,BTV)感染引起的一种严重侵害反刍动物的烈性虫媒传染病,易发于山羊、绵羊、鹿、牛以及野生反刍动物,在我国被列为一类,在OIE被划定为A类的重要动物疾病。蓝舌病毒属于呼肠孤病毒科(Reoviridae)环状病毒属(Orbivirus)的双股RNA(dsRNA)病毒[1-2]。该病由于发病率和死亡率高,同时导致死产、流产以及产奶量和生育率降低,引发了严重的直接经济损失。同时,对反刍动物及其动物制品的贸易限制以及对动物进行疫苗接种、诊断和病媒控制的支出也引发了较大的间接经济损失[3-8]。为了保护我国畜牧业健康发展,采用更加科学有效的方法观察蓝舌病毒特征具有重要意义。

由于基因序列数据量大,因此利用可视化技术将序列全面展现在平面图象中,便可发挥人的视觉识别能力,找到其中可能存在的规律和现象。1983年,Hamori等提出了第一个使用图形的方法研究DNA序列的方案[9],随后基因序列可视化研究不断深入。Jeffery等[10]和Hao等[11]分别提出混沌游戏表示(CGR)及Hao序列分形表示这两种基于分形的DNA序列可视化表示方法。Yang等使用Z曲线可视化非编码RNA序列[12],Sun等提出3D图形的曲率序列和饶率序列来表示基因序列[13]。使用可视化方法研究生物基因序列的主要思想是将生物基因序列通过某种映射方法转换为图形关系,通过比较不同图形,可从序列编码的角度比较序列之间的异同。

由于蓝舌病毒基因组中第二个片段Seg-2及其翻译的

基金项目:国家自然科学基金(62041213)
This work was supported by the National Natural Science Foundation of China(62041213).
通信作者:郑智捷(conjugatelogic@yahoo.com)

蛋白 VP2 的序列变异决定了血清型，基因组中第六个片段 Seg-6 及其翻译的蛋白 VP5 也决定了血清型。因此本文主要聚焦于基因组中第二个片段 Seg-2，基于 A，C，G，T 这 4 种碱基之间的数量关系，采用多元概率（Multiple Probability）测度[14]，针对云南省热带亚热带动物病毒重点实验室提供的来自 48 个国家的 485 条蓝舌病毒第二段基因组序列数据，采用数据可视化的方法展现蓝舌病毒第二段基因组序列特征。

2 模型与方法

本文方法包括两个核心模块：多元概率测量模块和可视化呈现模块。

首先在多元概率测量模块中，将获取的基因序列按国家分类，并通过多元概率方法进行概率统计，保存统计结果。之后，在可视化呈现模块中，选择不同的测量指标对结果进行投影，以散点图的形式进行可视化呈现。

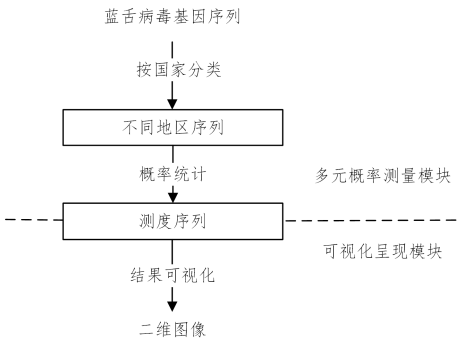


图 1 数据处理流程

Fig. 1 Data processing flow

2.1 多元概率测量模块

首先将数据集中的基因序列按国家进行分类，根据统计，数据集中共包含来自 48 个不同国家的蓝舌病毒基因序列，如表 1 所列。然后将分类后的数据输入多元概率测量模块，按照多元概率测度公式计算每个国家的多元概率值。

表 1 基因序列地区信息表

Table 1 Regional information of gene sequence

序号	国家	国家(英文)	序列数量
1	阿尔及利亚	Algeria	4
2	阿根廷	Argentina	5
3	澳大利亚	Australia	71
4	巴巴多斯	Barbados	1
5	波斯尼亚和黑塞哥维那	Bosnia and Herzegovina	1
6	保加利亚	Bulgaria	1
7	喀麦隆	Cameroon	4
8	中国	China	2
9	哥斯达黎加	Costa Rica	1
10	塞浦路斯	Cyprus	13
11	埃及	Egypt	1
12	法国	France	45
13	德国	Germany	2
14	直布罗陀	Gibraltar	1
15	希腊	Greece	24
16	危地马拉	Guatemala	1
17	洪都拉斯	Honduras	1
18	印度	India	51
19	印度尼西亚	Indonesia	10
20	以色列	Israel	16
21	意大利	Italy	47
22	牙买加	Jamaica	2
23	日本	Japan	14
24	科索沃	Kosovo	2
25	科威特	Kuwait	1
26	利比亚	Libya	3
27	马提尼克岛	Martinique	1
28	摩洛哥	Morocco	14
29	荷兰	Netherlands	3
30	尼日利亚	Nigeria	3
31	巴基斯坦	Pakistan	2
32	巴拿马	Panama	2
33	波兰	Poland	1
34	葡萄牙	Portugal	3
35	俄罗斯	Russia	1
36	塞尔维亚	Serbia	1
37	南非	South Africa	69
38	南韩	South Korea	1
39	西班牙	Spain	12
40	苏丹	Sudan	3
41	瑞士	Switzerland	1
42	台湾	Taiwan	2
43	突尼斯	Tunisia	4
44	土耳其	Turkey	5
45	英国	United Kingdom	1
46	美国	USA	27
47	津巴布韦	Zimbabwe	1
48	未知	unknown	4
总计			485

输入元素为完整的蓝舌病毒基因序列，序列的每个元素都由{A，C，G，T} 4 个符号之一组成。当基因序列包含 m 个元素时，可以统计出基因序列中 4 个符号的数量。m_s 为元素 s 的数量，其中 $s\in\{A,C,G,T\}$，P_s 为元素 s 出现的概率。因此，多元概率测度表示如下：

$$m=m_A+m_C+m_G+m_T$$

$$P_s=\frac{m_s}{m},s\in\{A,C,G,T\}$$

$$1=P_A+P_C+P_G+P_T$$

假设序列 X 中共有 100 个元素，其中 A 出现 37 次，C 出现 29 次，G 出现 11 次，T 出现 23 次，则 $P_A=0.37$，$P_C=0.29$，$P_G=0.11$，$P_T=0.23$，该值即为此序列的多元概率值。

2.2 可视化呈现模块

通过多元概率测量模块可得到各个国家的多元概率值，其中每个国家的多元概率值包含 4 个元素$\{P_A,P_C,P_G,P_T\}$。因此将多元概率值进行二维散点图投影时，投影坐标包含 6 种不同角度，即{A，C}，{A，G}，{A，T}，{C，G}，{C，T}，{G，T}。

可视化呈现模块中的处理过程如下：

1）选择国家；2）选择投影角度；3）将对应横纵坐标的值进行投影。假设选择的横纵坐标分别为{A，C}，即选取该国家的第一条序列，将多元概率测量模块得到的 P_A 值作为横坐标，P_C 值作为纵坐标，得到多元概率二维散点图的一个投影点，以此类推直到遍历完该国家的所有序列。

通过该方法，同一序列选择$\{P_A,P_C,P_G,P_T\}$中任意两种不同的值作为横纵坐标，变换坐标即可得到该序列不同角度的投影。对同一序列采用 6 种不同角度进行观察可发现更多基因序列细节，便于直观深入地了解蓝舌病毒。

3 可视化结果呈现

选择使用 Python 中的散点图方式对实验结果进行可视化呈现，并使用不同颜色对不同国家的蓝舌病毒序列加以区分，从{A，C}，{A，G}，{A，T}，{C，G}，{C，T}，{G，T}6 种维度可视化蓝舌病毒 segment2 数据集中的所有基因序列，可视化结果如图 2 所示。

(a)A,C 维度 (b)A,G 维度

(c)A,T 维度 (d)C,G 维度

(e)C,T 维度 (f)G,T 维度

图 2 蓝舌病毒 6 种维度可视化图像

Fig. 2 Bluetongue virus 6-dimensional visualization images

从图 2 中可观察到明显的区分特征,图中落在相同位置的点可推测其具有相同特征,产生集聚现象的部分可推测其具有相似特征。如 France,Tunisia,Morocco 这 3 个国家的数据分布集聚,部分数据重叠,其详细数据如表 2 所列。

表 2 France, Tunisia, Morocco 3 个国家可能具有相同特征的序列

Table 2 Sequences that may have the same characteristics in France, Tunisia and Morocco

序列	国家	多元概率 A	多元概率 C	多元概率 G	多元概率 T
JX861499	法国	0.31	0.174	0.252	0.263
KP821002	法国	0.31	0.174	0.252	0.263
KP821000	法国	0.31	0.174	0.252	0.263
KP821023	突尼斯	0.31	0.174	0.252	0.263
KP821003	法国	0.31	0.174	0.253	0.263
KP821015	摩洛哥	0.31	0.174	0.253	0.263
EU625362	摩洛哥	0.31	0.175	0.253	0.262
KP821014	摩洛哥	0.31	0.175	0.253	0.263

因此，从 6 种维度观察，蓝舌病毒的多元概率投影各有不同，其中的聚簇数量及规律可作为病毒的重要参考特征。

4 结果分析

经过前期大量测试及实验，为使结果更加简洁直观，选取蓝舌病毒 segment2 中前 100 条数据，分别使用进化分析软件 BEAST 生成的系统发育树（Phylogenetic Trees）以及多元概率方法生成多元概率测度数据，并将两种方法得到的结果进行比较。

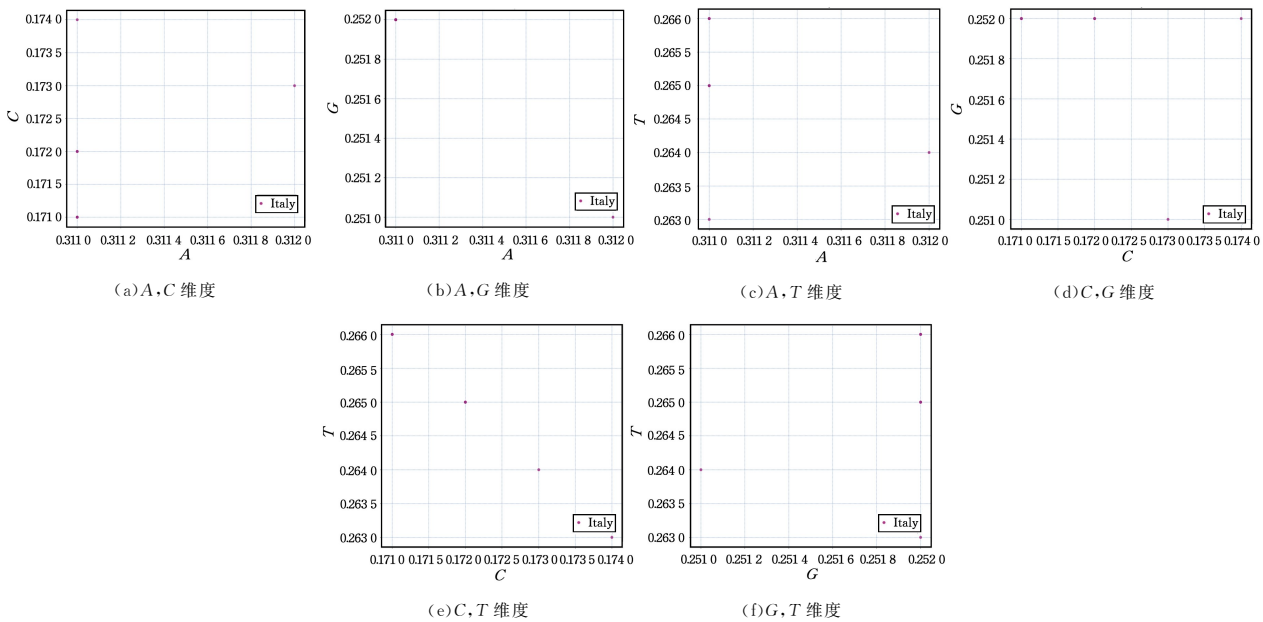
最终选取实验结果中呈现效果较好、具有明显代表性的意大利的蓝舌病毒数据为例进行结果分析。前 100 条数据中有 6 条来自意大利的蓝舌病毒，其信息如表 3 所列。

表 3 前 100 条数据中意大利蓝舌病毒信息表

Table 3 Information of Italian bluetongue virus in the first 100 data

序列	血清型	地区	年份
KJ661730	BTV-1	意大利	2013
KJ577095	BTV-1	意大利	2012
KJ577115	BTV-1	意大利	2010
KJ736002	BTV-1	意大利	2006
KJ019206	BTV-1	意大利	2013
KJ577125	BTV-1	意大利	2013

根据实验得到的多元概率数据生成{A,C},{A,G},{A,T},{C,G},{C,T},{G,T}6 种维度可视化结果。



(a)A,C 维度 (b)A,G 维度 (c)A,T 维度 (d)C,G 维度

(e)C,T 维度 (f)G,T 维度

图 3 前 100 条数据中意大利蓝舌病毒 6 种维度可视化图像

Fig. 3 6-dimensional visualization images of Italian bluetongue virus in the first 100 data

从该结果中可以看出，AG 维度中包含 2 个分类，除 AG 维度外，其余维度均包括 4 个分类，由此可推断这 6 条数据分为 4 个分类，其具体多元概率数值如表 4 所列。

表 4 前 100 条数据中意大利蓝舌病毒多元概率数值

Table 4 Multiple probability value of Italian bluetongue virus in the first 100 data

序号	序列	血清型	年份	多元概率 A	多元概率 C	多元概率 G	多元概率 T
1	KJ577095	BTV-1	2012	0.311	0.171	0.252	0.266
2	KJ019206	BTV-1	2013	0.311	0.171	0.252	0.266
3	KJ661730	BTV-1	2013	0.311	0.172	0.252	0.265
4	KJ577125	BTV-1	2013	0.311	0.172	0.252	0.265
5	KJ736002	BTV-1	2006	0.311	0.174	0.252	0.263
6	KJ577115	BTV-1	2010	0.312	0.173	0.251	0.264

根据表 4 中得到的结果，可将前 100 条数据中来自意大利的蓝舌病毒分为 4 类。

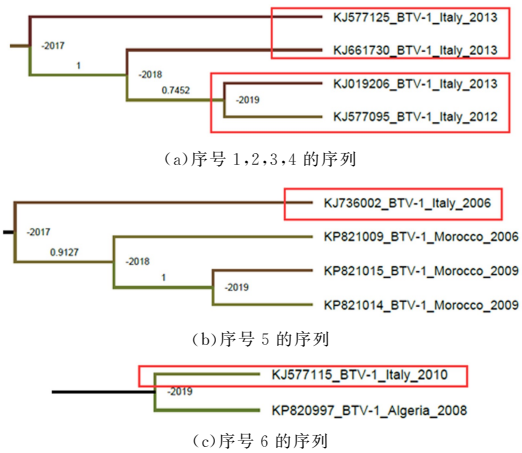
序号 1,2 的两条序列的多元概率 A、多元概率 C、多元概率 G，多元概率 T 的值相等，因此可归纳为类 I。

序号 3,4 的序列多元概率 A、多元概率 C、多元概率 G、多元概率 T 的值相等，但多元概率 C（与类 I 中多元概率 C 的值相差 0.001）以及多元概率 T（与类 I 中多元概率 T 的值相差 0.001）的值与类 I 中的值不相等，因此可归纳为类 II。

序号 5 的序列多元概率 C（与类 I 中多元概率 C 的值相差 0.003）和多元概率 T（与类 I 中多元概率 T 的值相差 0.003）的值与类 I 中的值不相等，因此可归纳为类 III。

序号 6 的序列多元概率 A（与类 I 中多元概率 A 的值相差 0.001）、多元概率 C（与类 I 中多元概率 C 的值相差 0.002）、多元概率 G（与类 I 中多元概率 G 的值相差 0.001）、多元概率 T（与类 I 中多元概率 T 的值相差 0.002）的值均与类 I 中的值不相等，因此可归纳为类 IV。

观察由蓝舌病毒 segment2 中前 100 条数据生成的系统发育树，其中 6 条来自意大利的蓝舌病毒如图 4 可知。



(a)序号 1,2,3,4 的序列

(b)序号 5 的序列

(c)序号 6 的序列

图 4 前 100 条数据中意大利蓝舌病毒发育树

Fig. 4 Italian bluetongue virus development tree in the first 100 data

由图 4 可知，多元概率测度得到的数据与系统发育树中的分支结构对应。

结束语 病毒基因序列分析问题一直是病毒学中受到主要关注的领域。Jeffrey 的研究表明，从元分析的层次结构及变体构造将各种病毒基因组序列作为独特的基因组索引集，并将相关信息映射到受限几何区域是可行的[15]。传统的复杂基因序列变异精细分析软件对生物基因序列进行分析并构建系统发育树的方法较为繁琐，参数设置复杂。无论是 Tu 等使用 BLAST(Basic Local Alignment Search Tools)软件对去毛鸟类进行的 DNA 分子鉴定[16]或是郭科使用 BEAST 软件对 MLB2 型星状病毒进行的分析[17]，其操作过程相比多元概率测度的方法都更为复杂。

本文建立了一种高效便捷的蓝舌病毒基因序列可视化模型，使用多元概率测度的统计学方法，并在此基础上对结果进行可视化操作可以简洁直观地呈现出蓝舌病毒的分类特征，其结果与进化分析软件 BEAST 生成的系统发育树结果相互吻合。

同时，此方法仍存在改进空间，相比传统的系统发育树结果，此方法无法呈现出具体的分支信息。如实验结果中的 KJ661730 及 KJ577125 两条序列，尽管多元概率测度方法可以使用简单的方式判断出二者的同源性，但由于其多元概率 A、多元概率 C、多元概率 G、多元概率 T 的值全部相同，无法像系统发育树结果中呈现的一样观察到二者其实位于不同的分支，因此多元概率测度方法对结果的进一步细分能力仍有待提升。

该模型易于理解和使用，初步分类结果符合实际，具有一定的科学性和实用性。后续可通过该模型对蓝舌病毒的其他基因组片段进行研究，并考虑提升分类精度的优化方法。

参 考 文 献

[1] PATEL A, ROY P. The molecular biology of Bluetongue virus replication[J]. Virus Research An International Journal of Molecular & Cellular Virology, 2014, 182: 5-20.

[2] PRASAD B V, SCHMID M F. Principles of virus structural organization[J]. Advances in Experimental Medicine and Biology, 2012, 726: 17-47.

[3] RUSHTON J, LYONS N. Economic impact of Bluetongue: a review of the effects on production[J]. Veterinaria Italiana, 2015, 51(4): 401-406.

[4] PINIOR B, BRUGGER K, KÖFER J, et al. Economic comparison of the monitoring programmes for bluetongue vectors in Austria and Switzerland[J]. The Veterinary Record, 2015, 176(18): 464.

[5] PINIOR B, LEBL K, FIRTH C, et al. Cost analysis of bluetongue virus serotype 8 surveillance and vaccination programmes in Austria from 2005 to 2013[J]. The Veterinary Journal, 2015, 206(2): 154-160.

[6] GREWAR J D. The economic impact of Bluetongue and other orbiviruses in sub-Saharan Africa, with special reference to Southern Africa[J]. Veterinaria Italiana, 2016, 52(3/4): 375-381.

[7] GETHMANN J, PROBST C, CONRATHS F J. Economic Impact of a Bluetongue Serotype 8 Epidemic in Germany[J]. Frontiers in Veterinary Science, 2020, 7: 65.

[8] MACLACHLAN N J, OSBURN B I. Impact of bluetongue virus infection on the international movement and trade of ruminants [J]. J Am Vet Med Assoc, 2006, 228(9): 1346-1349.

[9] HAMORI E, RUSKIN J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences[J]. Journal of Biological Chemistry, 1983, 258(2): 1318-1327.

[10] JEFFREY H J. Chaos game representation of gene structure [J]. Nucle-ic Acid Research, 1990, 18(8): 2163-2170.

[11] HAO B L. Fractals fromgenomes-exact solutions of a biology-inspired problem[J]. Physica A, 2000, 282: 225-246.

[12] YANG Y L, WANG J H. Research on Z-curve of non-coding RNA[C]// Proceedings of the Eleventh Chinese Conference on Biophysics. 2009: 183-184.

[13] SUN C F, CHENG Z. Visual processing method of gene sequence: China Patent, CN201610810914. 3 [P]. 2017-04-19.

[14] JEFFREY Z, ZHU M H. Input-Output Types of Fifteen Modules on Discrete and Real Measurements for COVID-19[M]. EC Neurology Special Issue, 2021(2): 71-85.

[15] JEFFREY Z, LIU J Z. A Visual Framework of Meta Genomic Analysis on Variations of Whole SARS-CoV-2 Sequences[M]. EC Neurology Special Issue, 2021(2): 49-70.

[16] TU F Y, HAN W J, WANG T, et al. A case of morphology and DNA molecular species of defeathered birds[J]. Chinese Journal of Forensic Sciences, 2020(5): 108-110.

[17] GUO K. Complete Genome amplification of Astrovirus MLB2 in Patients Following Hematopoietic Stem Cell Transplantation and the Evo; utionary Study of the Astrovirus[D]. Anhui: Anhui Medical University, 2020.

CHEN Hui-pin, born in 1999, bachelor. Her main research interests include computer biology science and cyber security.

ZHENG Zhi-jie, born in 1956, Ph. D, professor. His main research interest is variant construction.