

基于复杂网络理论的肺癌特异性基因识别研究

于 晓

(上海理工大学管理学院, 上海 200093)

✉yu_xiao127@163.com



摘 要: 基于复杂网络理论分别构建以肺癌组织和健康肺组织基因为节点, 基因间调控关系为边的基因调控网络, 并从拓扑结构、分布特征、节点中心性三方面分析网络特性, 挖掘网络核心节点生物功能差异性以识别出肺癌特异性基因。结果发现, 肺癌组织和健康肺组织基因调控网络拓扑参数极其相似且两者都为无标度网络, 两网络核心节点集高度重叠, 但非重叠部分核心节点的生物功能十分特殊, 并据此识别出肺癌特异性基因。该方法识别出的肺癌特异性基因, 能够成为潜在肺癌生物标记物, 为肺癌的早期诊断提供帮助, 同时该方法能够适用于其他疾病特异性基因的识别。

关键词: 复杂网络; 肺癌; 中心性; HUB节点

中图分类号: TP319 **文献标识码:** A

Research on Specific Gene Identification of Lung Cancer Based on Complex Network Theory

YU Xiao

(School of Management, University of Shanghai for Science and Technology, Shanghai 200093, China)

✉yu_xiao127@163.com

Abstract: Based on complex network theory, the study constructs lung cancer tissue and healthy lung tissue gene regulatory networks with genes as nodes and inter-gene regulatory relationships as edges. The paper analyzes network characteristics from three aspects: topological structure, distribution characteristics, and node centrality. By comparing the biological function of the key nodes of networks to identify lung cancer-specific genes, it is found that the topological parameters of the two gene regulatory networks are extremely similar and they are both scale-free networks. The key nodes set of the two networks are highly overlapping, but the non-overlapping key nodes' biological functions are very special, and then the lung cancer-specific genes are identified base on this. The lung cancer-specific genes identified by this method may be potential biomarkers for lung cancer, which promotes the early diagnosis of lung cancer. The method can also be applied in the identification of other disease-specific genes.

Keywords: complex network; lung cancer; centrality; HUB node

1 引言(Introduction)

肺癌是最常见和严重的癌症之一, 其发病率和死亡率在世界大多数国家都处于首位, 且增长率惊人^[1]。在中国, 肺癌已成为致死率最高的癌症^[2], 根据中国卫生部于2006年进行的第三次全国性死因抽样调查, 肺癌造成的死亡率自1990年以来增加了75.77%, 并且在剔除年龄结构变化因素后再次增加了33.25%。尽管吸烟这一肺癌的最相关危险因素已被广泛认识, 但是对肺癌发病机制和诊断治疗等一系列问题的研究依旧进展缓慢。其次, 由于肺癌通常在早期不会引发任何症状, 许多患者的相关体征和症状仅在疾病发展时发生,

这让肺癌的确诊存在相当大的难度, 若在肺癌早期通过检测特异性基因使其确诊并得到充分的治疗, 其死亡率将会在很大程度上降低。此外, 大多数肺癌患者仍在接受化学疗法, 但这些患者中的大多数经过化学疗法后并未好转, 这是因为化学疗法并没有解决患者潜在遗传缺陷这一导致肺癌发生的根本病因。因此, 现在面临的巨大挑战是如何准确地识别这些潜在遗传缺陷, 以便使治疗与患者相匹配, 获得更高的治愈率。这一系列的问题让识别有诊断用途的肺癌生物标记成了攻克肺癌这一高致死率疾病的关键。本文以复杂网络理论为工具, 提出了系统性识别肺癌特异性基因的方法: 根

据肺癌组织中基因关系构建肺癌组织基因调控网络,分析网络拓扑结构和分布规律,结合中心性算法挖掘网络中的关键节点,根据其生物功能推测肺癌特异性基因。该方法为医学工作者们揭示肺癌发生机制和针对性治疗肺等方面提供了依据。

2 文献综述(Literature review)

网络生物学是生物医学研究中一个快速发展的领域,它提供了一个总体的观点,即遗传疾病的发生不是单个基因突变的结果,而是由基因网络环境的扰动引起。因此,认识这些基因相互作用网络的拓扑结构,以及确定在其结构和调控中起关键作用的基因是理解复杂遗传性疾病的关键。

现代生物网络的研究着重分析网络的拓扑结构^[3],基于对大规模生物网络的研究通常比临床实验室更有优势^[4]。此外研究表明生物网络中关键基因的发现对解决生物问题有重要意义,因为疾病生物网络中的关键基因往往与致病基因有着高度的相关性^[5]。在对蛋白质相互作用网络的研究中验证了这一观点,外国研究者们发现癌症致病基因或蛋白质在网络中存在非常高的中心性^[6],同时癌症相关致病基因或蛋白质之间更倾向于密切联系使其成为一个社区^[7],在社区之内它们存在更广泛的相互作用关系。

在众多识别生物网络关键节点的方法中,节点中心性得分是一种快速且可重现的方法,其根本思想是根据节点在网络中所处位置的关键性对其进行排名。这些网络节点中心性度量通常基于路径长度和随机游走,例如接近度中心、介数中心、聚类系数。同时,针对不同类型生物网络不同中心性算法的表现也不尽相同,例如, Kim等人^[8]发现,基于最短路径的介数中心性度量可以作为检测关键性基因的一种可靠标准。此外, He等人^[9]表明,在蛋白质相互作用网络中接近度中心性并不能准确预测出关键蛋白质。此外,子图中心性在高密度网络中的效果拔群,但在酵母菌蛋白质相互作用网络这种较稀疏网络中预测关键蛋白质的能力尚未超过度中心性。

国内针对肺癌特异性基因的研究工作主要集中在单个基因的突变检测上^[10,11],很少有从系统或者网络角度出发的相关研究,对此我们利用复杂网络理论进行了系统性对比实验,通过对比肺癌与健康肺组织基因调控网络的拓扑参数以分析两个网络的差异性。其次针对上述文献中单一拓扑中心性算法难以适配的问题,我们提出全局与局部中心性相结合观点,利用四种局部和七种全局中心性算法综合确定关键基因。

3 研究设计(Research design)

3.1 基因调控网络构建

我们所用到的数据来自文献[12],研究者们提供来自32个人体不同细胞、组织、器官的基因调控数据,以边表的形式存储。我们选定其中肺癌组织和健康肺组织的基因调控数据作为我们的实验数据,然后对其进行预处理,根据权重分布规律剔除了一部分噪声,将剩下的核心数据构建成了肺癌和健康肺组织基因调控网络。其中健康肺组织基因调控网络如图1(a)所示,肺癌基因调控网络如图1(b)所示。图中的节点分别代表参与肺癌组织和健康肺组织细胞调控的基因,网络中的边代表这些基因之间的相互作用关系。

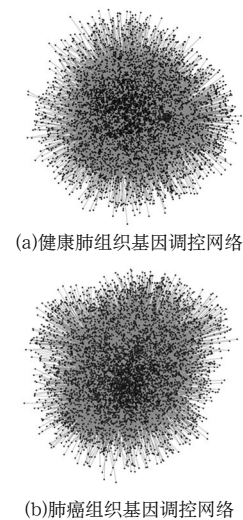


图1 健康肺组织基因调控网络和肺癌组织基因调控网络
Fig.1 Gene regulatory networks in healthy lung tissue and lung cancer tissue

3.2 网络拓扑参数与分布特征

本文计算了肺癌组织基因调控网络和健康肺组织基因调控网络的拓扑参数,通过对比发现肺癌组织和健康肺组织基因调控网络拓扑参数极其相似,如表1所示,这就解释了肺癌细胞和健康肺细胞在基因调控关系结构上的同源性。其中,网络直径被定义为两个节点对间的最大距离,而节点间距离是指连接两个节点所需要的最小边数;网络半径与网络直径相似,被定义为两个节点对间的最小距离;网络集中性用来描述网络拓扑结构与星形网结构的相似程度,相似性越高该数值越接近于1;网络密度被定义为网络中实际存在边的数量与网络最大可容纳边数量的比值;最短路径长度表示从原始节点到目标节点所需经过的最小边数;特征路径长度被定义为存在于网络中所有最短路径的均值。

表1 健康肺组织与肺癌组织基因调控网络拓扑参数

Tab.1 Topological parameters of gene regulatory network in healthy lung tissue and lung cancer tissue

参数	健康肺组织	肺癌组织
节点数	5112	5802
网络直径	6	6
网络半径	4	4
网络集中性	0.248	0.262
网络密度	0.002	0.002
最短路径长	26127432	33657402
特征路径长	3.234	3.229
平均邻居节	12.516	12.937

分别统计肺癌组织和健康肺组织基因调控网络的度分布,其中健康肺组织基因调控网络的度分布如图2(a)所示,肺癌组织基因调控网络的度分布如图2(b)所示。图中横坐标代表节点度,纵坐标代表网络中对应横坐标节点度下的节点数,由图可知网络度分布均存在明显下降趋势,说明网络属于无标度网络,即网络中的大多数节点只存在很少的连边与此同

时网络中存在为数不多连边数巨大的HUB节点，低连通节点与高连通节点之间的边在网络中占优势^[13]。无标度网络的特征是它们具有许多外围节点和几个高度连接的节点，这些网络具有通用的组织原则：“小世界”属性(即，它们通过节点之间相对较短的路径高度连接)。拥有这种属性网络具有鲁棒性，可以淘汰外围节点并保留核心网络功能。

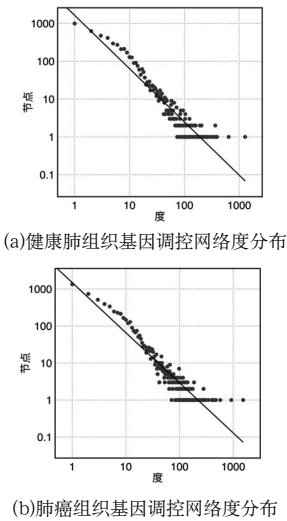


图2 健康肺组织和肺癌组织基因调控网络度分布
Fig.2 Degree distribution of gene regulatory networks in healthy lung tissue and lung cancer tissue

3.3 网络节点中心性分析

网络中心性算法是度量节点在网络中重要性的度量方法，这里分别用到四种局部性中心性和七种全局性中心性算法综合挖掘网络中的HUB节点。我们先通过计算得到两个网络中所有基因节点的11种中心性得分，再对11种中心性得分累加归一化作为最终的中心性得分。最终中心性得分排在前面的基因在对应组织基因调控网络中有强中性心，据此挑选得分排名前10的节点作为对应组织基因调控网络的HUB做后续分析，见表2。

表2 肺癌组织和健康肺组织基因调控网络中心性得分
top10基因

Tab.2 Centrality score of top10 genes in gene regulatory networks of lung cancer tissue and healthy lung tissue

肺癌	中心性得分	健康肺	中心性得分
SP1	9.0302	SP1	9.0301
SP4	6.5496	SP4	5.5513
KLF4	4.7453	POU2F1	5.0073
POU3F2	4.4085	RFX2	4.9569
KLF7	4.4049	ERG	4.4258
RFX2	4.3652	KLF4	4.3898
CTCF	4.2648	ETS1	4.2843
KLF16	4.2171	RFX3	4.2101
SHOX2	4.0746	CTCF	4.1884
RFX3	4.0107	KLF16	4.1614

其中，四种局部性算法的度量标准为：节点度(Degree)，最大邻居连通分量(Maximum Neighborhood Component)^[14]，最大邻居连通分量密度(Density of Maximum Neighborhood Component)^[14]，最大集团中心性(Maximal Clique Centrality)^[15]。其中，节点度被定义为与该节点直接相连其他网络中节点的数量；最大邻居连通分量定义为节点邻居网络中最大连通分量包含的节点数，这里的邻居网络是节点与它的邻居所构成的网络；最大邻居连通分量密度则为节点邻居网络实际边数与所能容纳最大边数的比值；最大集团中心性则与节点所在的最大集团中连通分量个数相关。

七种全局性算法的度量标准为：接近度(Closeness)^[16]、偏心率(Eccentricity)^[17]、辐射率(Radiality)^[18]、障碍率(Bottleneck)^[19]、压力值(Stress)^[20]、介数(Betweenness)^[21]、边扩散度(Edge Percolated Component)^[22]。其中，接近度定义为节点到其他节点距离的倒数和，该值越高节点越容易到达其他节点；偏心率定义为节点到距离它最远的节点距离的倒数；辐射率以网络直径为标准，比较节点到其他节点的距离远近；障碍率则是在网络中构造以节点为根的最小生成树，通过统计遍历该最小生成树过程中所经过网络中的其他节点来定量根节点的中心性；压力值定义为通过该节点最短路径的数量，压力值越高，该节点更倾向于位于其他节点对的最短路径上；介数被定义为通过该节点最短路径数与网络中最短路径总数的比值，介数值越大说明该节点越倾向于处于网络中枢轴位置；边扩散度则按照相等概率连续除去节点所在网络中的边，执行多次，并记录每次操作后当前网络中该节点的邻居节点数，算法的结果取决于这些记录。

3.4 特异性基因的确定

为了识别肺癌组织中的特异性基因，我们重点关注仅存在于肺癌组织却不存在于健康肺组织中的HUB基因，这些HUB基因如图3所示，由图可知肺癌组织和健康肺组织HUB基因大部分重合，仅存在三个肺癌独有的HUB基因。通过在Genecards^[23]中匹配这些肺癌独有HUB基因的功能，能够进一步得到它们的生物功能信息。这里Genecards是人类基因的综合数据库，提供以基因数据为中心的基因组、转录组、蛋白质组、遗传和临床相关的生物功能信息。

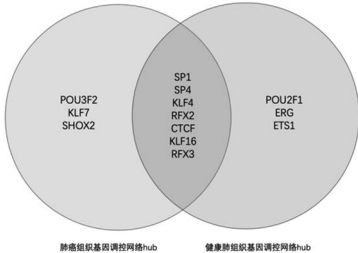


图3 健康肺组织和肺癌组织HUB基因韦恩图
Fig.3 Venn diagram of Hub genes in healthy lung tissue and lung cancer tissue

根据Genecards的匹配结果发现肺癌组织独有的三个HUB基因(POU3F2、KLF7、SHOX2)确实在生物功能上与重合基因存在差异，我们因此将其作为肺癌特异性基因。具体来说，多于一半的HUB基因同时在于肺癌组织网络和

健康肺组织网络中重复出现,它们为SP1、SP4、KLF4、RFX2、CTCF、KLF16、RFX3,且它们大多为不同类型的转录因子,主要参与调控那些不论癌细胞还是正常细胞都要经历的细胞过程,包括细胞分化、细胞生长、细胞凋亡。而我们重点关注的肺癌组织基因调控网络独有HUB基因为:POU3F2、KLF7、SHOX2,其中POU3F2编码的蛋白质参与神经元分化,并增强促肾上腺皮质激素释放激素调节基因的激活,它的过表达与黑色素瘤细胞增殖有关。KLF7编码的蛋白通过抑制胰岛 β 细胞中胰岛素的表达和分泌,以及通过调节脂肪细胞中脂细胞因子的分泌来促进2型糖尿病的发展。SHOX2也是一种蛋白质编码基因,与它相关的疾病包括Turner综合征和DeLange综合征,其患者先天身形矮小。从简单功能上已经能看出这三个肺癌独有HUB基因与其他基因的差异。

接下来,为了验证所挖掘出肺癌特异性基因(POU3F2、KLF7、SHOX2)的有效性,我们查阅相关肺癌临床实验文献试图从生物实验角度使其特异性再次得到验证。结果是,对于SHOX2基因,多个研究已经证明SHOX2基因DNA甲基化是检测肺癌的强有力的生物标记物^[24,25]。对于KLF7基因也有研究证明与邻近的正常组织相比,患者肺腺癌(Lung Adenocarcinoma, LAC)组织中的KLF7表达升高,且KLF7的高蛋白水平与肿瘤大小相关,此外,高的KLF7表达水平与肺腺癌患者较差的临床结局显著相关,表明了KLF7作为新型预后生物标志物和治疗靶标的潜在作用^[26]。以上结果均表明本文利用复杂网络节点中心性方法所识别出的肺癌特异性基因与临床生物实验所得出的结论是一致的,这证明了本文所提出方法的有效性。

4 结论(Conclusion)

随着基因组学、蛋白质组学、转录组学的发展,生物大数据得以涌现,针对癌症这一复杂人类疾病的相关研究也逐渐从临床实验转移到数据分析上来。本文正是在这一背景下,借助复杂网络理论这一系统性工具对肺癌基因数据进行建模分析,借助多种拓扑中心性算法,以及人类基因综合数据库提出在生物大数据中挖掘癌症特异性基因的系统性方法:(1)构建癌症与对照组基因调控网络。(2)分析网络拓扑结构和分布规律。(3)计算网络节点中心性得分。(4)比对并匹配HUB节点生物功能。

通过分析发现基因调控网络存在广泛的无标度特性,即人类器官组织对应的基因调控系统由少数重要基因来主导控制,它们调控着整个系统的行为,识别这些基因对认识人类细胞的生物进程有重要意义。此外,健康肺组织细胞和肺癌组织细胞对应的基因调控网络从拓扑结构到分布特征上都极其相似,这也证明了同处病变与健康组织的同源性。最后,

利用本文所提出系统性方法识别出的肺癌特异性基因,其肺癌相关性也在本文中得到了证实,该基因将作为一个重要的肺癌生物标记物利用在临床检测上。总的来说,本文方法对从大规模癌症基因数据中识别特异性基因提供了可靠思路,而癌症特异性基因的存在将使癌症在早期更容易被识别,进而提升治愈率。

参考文献(References)

- [1] Alzahouri K, Martinet Y, Briancon S, et al. Staging practices of primary non-small-cell lung cancer: a literature review[J]. *European Journal of Cancer Care*, 2006, 15(4): 348-354.
- [2] Chen W, Zhang S, Zou X. Estimation and Projection of Lung Cancer Incidence and Mortality in China[J]. *Chinese journal of lung cancer*, 2010, 13(5): 488-493.
- [3] Carlsson G. Topology and data[J]. *Bulletin of the American Mathematical Society*, 2009, 46(2): 255-308.
- [4] Furney SJ, Alba MM, Lopezbigas N. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations[J]. *BMC Genomics*, 2006, 7(1): 165-175.
- [5] Park D, Park J, Park SG, et al. Analysis of human disease genes in the context of gene essentiality[J]. *Genomics*, 2008, 92(6): 414-418.
- [6] Sun J, Zhao Z. A comparative study of cancer proteins in the human protein-protein interaction network[J]. *BMC Genomics*, 2010, 11(S3): 1471-1481.
- [7] Gandhi TKB, Zhong J, Mathivanan S, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets[J]. *Nature Genetics*, 2006, 38(3): 285-293.
- [8] Kim J, Kim I, Han SK, et al. Network rewiring is an important mechanism of gene essentiality change[J]. *Scientific Reports*, 2012(2): 900-907.
- [9] He X, Zhang J. Why do hubs tend to be essential in protein networks?[J]. *PLoS genetics*, 2006, 2(6): 826-834.
- [10] 周建平, 梁立轩, 李志芳. 非小细胞肺癌细胞系中关键致癌相关基因突变检测[J]. *攀枝花学院学报*, 2016, 33(05): 71-74.
- [11] 俞训彬, 陈小岩, 陈灵锋. 采用Illumina测序技术检测非小细胞肺癌驱动基因关键位点突变[J]. *临床与实验病理学杂志*, 2019, 35(7): 861-862.
- [12] Marbach D, Lamarter D, Quon G, et al. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases[J]. *Nature Methods*, 2016, 13(4): 366-370.
- [13] Maslov S, Snead K. Specificity and Stability in Topology of Protein Networks[J]. *Science*, 2002, 296(5569): 910-913.
- [14] Lin CY, Chin CH, Wu HH, et al. Hubba: hub objects analyzer—

- a framework of interactome hubs identification for network biology[J].Nucleic Acids Research,2008,36(S2):438-443.
- [15] Chin CH,Chen SH,Wu HH,et al.cytoHubba: identifying hub objects and sub-networks from complex interactome[J].BMC Systems Biology,2014,8(S4):1752-1761.
- [16] Sabidussi G.The centrality index of a graph[J].Psychometrika,1966,31(4):581-603.
- [17] Dankelmann P,Goddard W,Swart C S.The Average Eccentricity of a Graph and its Subgraphs[J].Utilitas Mathematica,2004,65(2):41-51.
- [18] Thomas WV,Robert KF.Integration and radiality:Measuring the extent of an individual's connectedness and reachability in a network[J].Social Networks,1998,20(1):89-105.
- [19] Chin CS,Samanta MP.Global snapshot of a protein interaction network—a percolation based approach[J].Bioinformatics,2003,19(18):2413-2419.
- [20] Alfonso S.Structural parameters of communication networks[J].Bulletin of Mathematical Biophysics,1953,15(4):501-507.
- [21] Barthélemy M.Betweenness centrality in large complex networks[J].European Physical Journal B,2004,38(2):163-168.
- [22] Przuji N,Wigle DA,Jurisica I.Functional topology in a network of protein interactions[J].Bioinformatics,2004,20(3):340-348.
- [23] Rebhan M,Chalifa-Caspi V,Prilusky J,et al.GeneCards:a novel functional genomics compendium with automated data mining and query reformulation support[J].Bioinformatics,1998,14(8):656-664.
- [24] Ilse P,Biesterfeld S,Pomjanski N,et al.Analysis of SHOX2 Methylation as an Aid to Cytology in Lung Cancer Diagnosis[J].Cancer genomics & proteomics,2014,11(5):251-258.
- [25] Ilse P,Biesterfeld S,Pomjanski N,et al.SHOX2 DNA Methylation Is a Tumour Marker in Pleural Effusions[J].Cancer genomic & proteomics,2013,10(5):217-223.
- [26] Cai XD,Zhou YB,Huang LX,et al.Reduced expression of Krüppel-like factor 17 is related to tumor growth and poor prognosis in lung adenocarcinoma[J].Biochemical and Biophysical Research Communications,2012,418(1):67-73.

作者简介:

于 晓(1991-),男,硕士生.研究领域:复杂网络.

(上接第8页)

的算法应用性更强。整体上来讲准确率较所对比方法也有明显提升。

5 结论(Conclusion)

为解决学习者正视屏幕时出现的分心状态无法被检测到的问题,本文以眼动行为分析为突破点,提出了一种基于RNN的眼动分析算法,实现了在线学习者视线焦点在屏幕上时,对学习者的阅读、搜索、分心三种学习状态的检测。与其他同类眼动分析算法相比,本文算法能有效地将当前时刻的隐藏状态与之前时刻隐藏状态和当前时刻输入特征相关联,提高模型对特征的表达能力,分类准确率有明显提升。该算法具有设备要求较低,实施简单等优点。整体效果良好,具有较高的实用价值。对于眼动分析中视线在屏幕下方运动时出现了一定误差的问题,有待进一步的作出优化。

参考文献(References)

- [1] Lee H, Kim Y, Park C. Classification of human attention to multimedia lecture[C]. 2018 International Conference on Information Networking(ICOIN). IEEE, 2018: 914-916.
- [2] Gunawan F E, Wanandi K, Soewito B, et al. Detecting the early drop of attention using EEG signal[C]. 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics(EECSI). IEEE, 2017: 1-6.
- [3] Jie W, Hai-yan L, Biao C, et al. Application of educational data mining on analysis of students' online learning behavior[C]. 2017 2nd International Conference on Image, Vision and Computing(ICIVC). IEEE, 2017: 1011-1015.
- [4] Xu R, Chen J, Han J, et al. Towards emotion-sensitive learning cognitive state analysis of big data in education: deep learning-based facial expression analysis using ordinal information[J]. Computing, 2019: 1-16.
- [5] 王昌海, 申红雪, 张王卫, 等. 一种基于人脸识别的课堂教学监控系统[J]. 软件工程, 2020, 23(01): 48-50.
- [6] 吴沧海, 熊焕亮, 王映龙. 远程学习中学习状态判断的情感计算研究[J]. 软件导刊(教育技术), 2013, 12(07): 24-27.
- [7] 王凤娇, 田媚, 黄雅平, 等. 基于眼动数据的分类视觉注意模型[J]. 计算机科学, 2016, 43(01): 85-88; 115.
- [8] 江波, 王小霞, 刘迎春, 等. 在线测评中的学习者眼动行为分析——以浙江工业大学的眼动实验为例[J]. 现代教育技术, 2018, 28(05): 19-25.
- [9] 王晓明, 赵歆波. 基于深度神经网络的个体阅读眼动预测[J]. 清华大学学报(自然科学版), 2019, 59(06): 468-475.
- [10] Yi J, Sheng B, Shen R, et al. Real time learning evaluation based on gaze tracking[C]. 2015 14th International Conference on Computer-Aided Design and Computer Graphics(CAD/Graphics). IEEE, 2015: 157-164.

作者简介:

郑茜元(1993-),男,硕士生.研究领域:智能计算.

郑虹(1974-),女,博士,副教授.研究领域:智能计算.本文通讯作者.

侯秀萍(1964-),女,硕士,教授.研究领域:智能计算.