Biophysics 生物物理学, 2016, 4(4), 49-55
Published Online November 2016 in Hans. http://www.hanspub.org/journal/biphy
http://dx.doi.org/10.12677/biphy.2016.44004

Hans汉斯

The Analysis of Minimum Spanning Tree of Proto-Oncogene Network

Fangping Wei[1*], Zhenxiong Lan[2]

[1]College of Physical Science and Technology, Guangxi University, Nanning Guangxi
[2]College of Computer and Information Engineering, Guangxi Teachers Education University, Nanning Guangxi
Email: *weifp@gxu.edu.cn

Received: Oct. 27th, 2016; accepted: Nov. 18th, 2016; published: Nov. 21st, 2016

Abstract

Human proto-oncogene is a gene which has a very close relationship with human cancer. It is possible to make the normal cells cancerous and cause cancer, when the structure of human proto-oncogene changed or it is over-expressed. Using the complex network theory to study the evolutionary and genetic relationship between proto-oncogene and cancer, the research shows that the conclusion is in good agreement with the actual situation. This paper studies the evolutionary and genetic relationship of the proto-oncogene sequence through constructing complex network and working out the method of minimum spanning tree network according to Prim algorithm, and analyzing three main parameters of network, average degree, average cluster coefficient, and average shortest path.

Keywords

Proto-Oncogene Sequence, CVTree Method, Complex Network, Minimum Spanning Tree

原癌基因网络的最小生成树分析

韦芳萍[1*]，蓝贞雄[2]

[1]广西大学物理科学与工程技术学院，广西 南宁
[2]广西师范学院计算机与信息工程学院，广西 南宁
Email: *weifp@gxu.edu.cn

*通讯作者。

文章引用: 韦芳萍, 蓝贞雄. 原癌基因网络的最小生成树分析[J]. 生物物理学, 2016, 4(4): 49-55.
http://dx.doi.org/10.12677/biphy.2016.44004

收稿日期：2016年10月27日；录用日期：2016年11月18日；发布日期：2016年11月21日

摘　要

人类原癌基因是与人类癌症关系非常密切的基因，当它的结构发生改变或过度表达时，就有可能使正常细胞发生癌变而导致癌症。本文通过构建复杂网络，然后根据Prim算法，做出最小生成树网络的方法，并分析网络的平均度、平均聚类系数和平均最短路径这三个主要参数来研究原癌基因序列之间的进化和亲缘关系。

关键词

原癌基因，CVTree方法，复杂网络，最小生成树

1. 引言

基因就是有遗传功能的 DNA 片段，不同的基因有着不同的遗传功能。而原癌基因是 DNA 中与细胞增殖有关的基因，当其结构发生改变导致表达激活时，细胞过度增殖，将会形成肿瘤，也就是癌症[1] [2] [3]。

目前人们一共发现了 500 多个人类原癌基因，共发现了 100 多种癌症。而这 100 多种癌症都是由这 500 多个原癌基因的表达激活而引起。例如，当基因 ARID1A 发生突变时，可能会引起肾癌、卵巢癌、乳腺癌、肝癌和前列腺癌等癌症；当基因 ITK 发生突变时，可能会引起淋巴瘤、肺癌和胰腺癌等癌症。最近在英国科学杂志《自然》中发表的一篇文章指出，通过对 30 种癌症的 7042 例癌症患者的突变基因的研究，发现了 22 种基因突变将会导致癌症，而在这 22 种基因中，几乎都是出现两种基因突变才引发癌症，子宫癌、胃癌和肝癌这三种癌症则需要 6 种基因突变才会引起[4]。本文通过构建原癌基因网络，利用网络参数构建癌症基因最小生成树，来分析癌症之间的进化关系。

CVTree 方法是近年来由郝柏林院士等人建立的同源比较方法里比较优越的方法，它通过计算两个基因组之间的组份矢量来构建进化距离，然后构建进化树(http://tlife.fudan.edu.cn/cvtree/) [5] [6] [7]。

复杂网络理论从建立至今已经有几十年的历史，目前正渐渐成熟完整。

一个复杂网络的最小生成树也是一个复杂网络，它包含了原网络中左右的节点，但节点之间相连的情况和原网络却不一样。一个无向且边带有权重的最小生成树，是把原网络中所有连接边的权重的和最小的树模型。现在最小生成树已经得到了非常广泛的应用，其涉及的领域有电力、航空还有生物科学领域。连接边的权重在不同领域分别代表了不同的意义[8] [9] [10]。

构建最小生成树有两种比较常用的方法，一种是 kruskal 算法，另一种是 prim 算法。这两种算法的主要区别是，kruskal 算法是顺序去边，而 prim 算法是顺序去端。本文采用 prim 算法构建最小生成树[11]。基于 prim 算法，便可找出无向且连接边带有权重的最小生成树。

2. 数据与方法

2.1. 数据来源

本文用到的原癌基因序列都是从美国国家生物技术信息中心(genbank) (http://www.ncbi.nlm.nih.gov/genbank) 下载。每个基因所含序列数大多都不止一条，下载总共得到了 4000 多条原癌基因序列。本文选择了与

15 种癌症对应的原癌基因，它们分别为膀胱癌(74)、大肠癌(448)、肺癌(623)、肝癌(136)、宫颈癌(319)、黑色素瘤(74)、甲状腺瘤(323)、淋巴瘤(570)、卵巢瘤(565)、脑肿瘤(59)、前列腺癌(587)、乳腺癌(668)、肾癌(300)、胃癌(138)和胰腺癌(513)，括号中的数量是癌症所对应的原癌基因序列总数。所有这 15 种癌症的原癌基因序列总数为 2229 条，因为同一个基因可能同时是属于几种癌症，所以各个癌症的序列之和并不为 2229。

2.2. 方法

2.2.1. 网络的构建

在用 cvtree 方法构建网络时，把每一条基因序列作为网络中的一个节点。本文将 15 种原癌基因分别作为一个单体构建一组 5 个不同 K 值的网络，K 分别取 6，9，12，15，18 这 5 个值。再将这 15 种癌症对应的 2229 条原癌基因序列作为一个总体单独构建一组 5 个不同 K 值的网络，总共构建的网络为 16 组。

用 CVTree 方法得到亲缘距离矩阵之后，我们得到了任意两条序列之间的亲缘距离 D，这时只要定一个标准值 Dis，并规定当任意两条序列的 D 值小于或等于标准值 Dis 时，这两条序列在网络中代表的节点相连，代表在标准值 Dis 下它们之间有亲缘关系；反之，两节点则不相连，代表在标准值 Dis 下它们之间没有亲缘关系。

2.2.2. 最小生成树的构建

本文利用 prim 算法构建网络最小生成树，先选定网络中任意一个节点作为子图，然后再找出这个子图到剩余子图中边长最短的边，将这条边对应的两个节点在邻接矩阵中的矩阵元置为 1，并将找到的节点加入子图，之后一直重复该步骤，直到子图包含了网络中所有节点时，便得到了最小生成数的邻接矩阵。

前面已经利用 CVTree 方法得到了 16 组不同 K 值的距离矩阵，通过 prim 算法计算这些矩阵边可得到最小生成树的邻接矩阵。两条序列的距离 D 就是它们之间边的权重。因为构建的网络为无向图，所以在距离矩阵中，对于任意两条序列有 $D_{ij} = D_{ji}$，且对角元全为 0，因此距离矩阵是关于对角对称的。

在原癌基因最小生成树网络中，边的权重直观体现了序列的亲缘关系，通过对最小生成树的分析，可以研究原癌基因的进化关系，进而研究癌症之间的转移机制等。

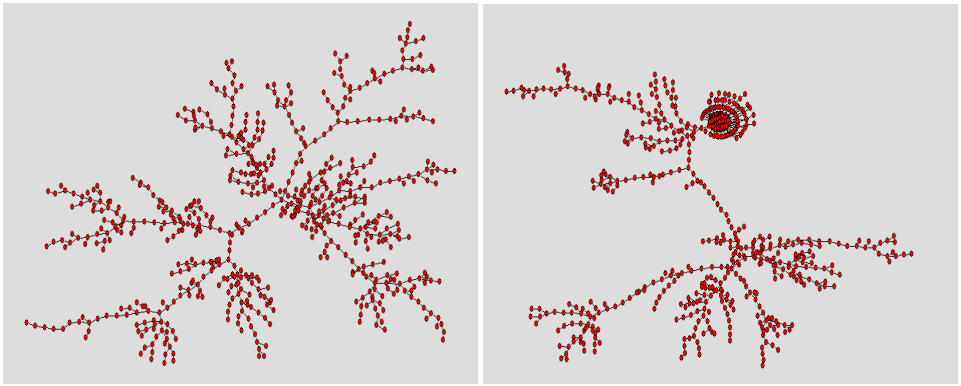
3. 结果与分析

3.1. 最小生成树的拓扑结构

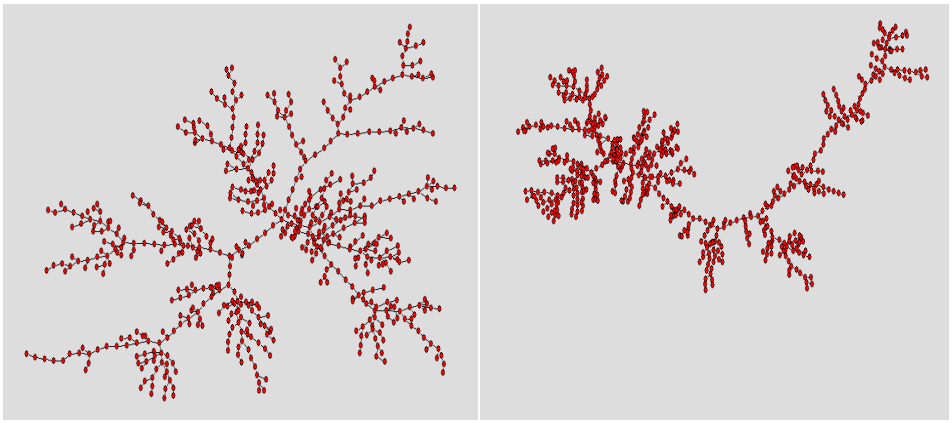
本文用基于 prim 算法而做出的 MST 最小生成树程序计算距离矩阵，得出了原癌基因序列的最小生成树邻接矩阵，并用 pajek 软件画出了 15 种原癌基因的 15 组最小生成树以及 15 组原癌基因总和的一组最小生成树。

图 1 给出了肺癌以及 15 种癌症对应的原癌基因序列在几个不同 K 值下的最小生成树的拓扑结构。分析图 1 可以发现，所有拓扑图中都有相同特点：所有节点中没有孤立点；只有极少数的节点连接边数大于 2，而其它绝大部分节点的连接边都只有一条或两条，不仅如此，在其它没列出来的癌症中，其最小生成树也都具有这个特点。由这一共同点我们可以看出原癌基因的进化特征将会比较相似；另外，这样的拓扑结构说明了序列之间的进化将会优先选择最短的路径进化。

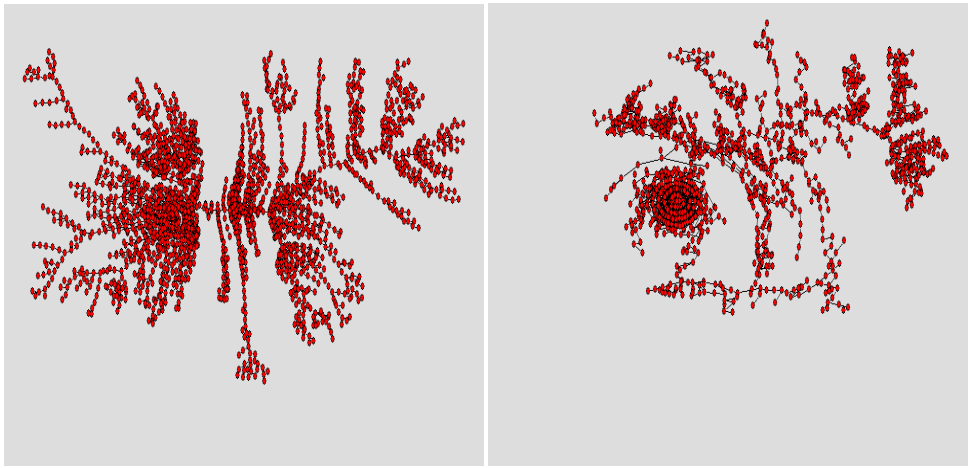
在同一种癌症不同 K 值的情况下进行横向比较，可以发现：随着 K 值的增大，将会出现一些具有较高的度的节点，且 K 值越大网络中最大的度也随着增大。节点的连接度越高，说明它在原癌基因序列进化中起到的作用也越高，在进化树中处于树干的作用。



(a) 原肺癌基因 K = 9　(b) 原肺癌基因 K = 15



(c) 原肺癌基因 K = 9　(d) 原肺癌基因 K = 15



(e) 所有序列 K = 6　(f) 所有序列 K = 15

Figure 1. Minimal spanning tree topology
图 1. 最小生成树的拓扑结构

3.2. 最小生成树的平均度

度是网络中三个重要的参数之一，分析原癌基因网络的度分布可以研究系统的亲缘进化关系特点。表 1 列出了 15 种癌症以及它们的总和的最小生成树在各个 K 值下的平均度。从表中可以看出：对于同一种癌症，K 的值并不影响最小生成树的平均度，不管 K 取何值，网络中连接边的总数都不变，这说明

Table 1. The average degree of the minimum spanning tree, the number behind the cancer name is the total number of cancers
表 1. 最小生成树的平均度，癌症名后面的数字是癌症所含的序列总数

K 值	6	9	12	15	18
原脑肿瘤基因 59	1.966102	1.966102	1.966102	1.966102	1.966102
原膀胱癌基因 74	1.972973	1.972973	1.972973	1.972973	1.972973
原黑色素瘤 74	1.972973	1.972973	1.972973	1.972973	1.972973
原肝癌基因 136	1.985294	1.985294	1.985294	1.985294	1.985294
原胃癌基因 138	1.985507	1.985507	1.985507	1.985507	1.985507
原肾癌基因 300	1.993333	1.993333	1.993333	1.993333	1.993333
原宫颈癌基因 319	1.993730	1.993730	1.993730	1.993730	1.993730
原甲状腺瘤基因 323	1.993808	1.993808	1.993808	1.993808	1.993808
原大肠癌基因 448	1.995536	1.995536	1.995536	1.995536	1.995536
原胰腺癌基因 513	1.996101	1.996101	1.996101	1.996101	1.996101
原卵巢癌基因 565	1.996460	1.996460	1.996460	1.996460	1.996460
原淋巴瘤基因 570	1.996491	1.996491	1.996491	1.996491	1.996491
原前列腺癌基因 587	1.996593	1.996593	1.996593	1.996593	1.996593
原肺癌基因 623	1.996790	1.996790	1.996790	1.996790	1.996790
原乳腺癌基因 668	1.997006	1.997006	1.997006	1.997006	1.997006
总原癌基因 2229	1.999103	1.999103	1.999103	1.999103	1.999103

网络平均紧密程度并没有变，而只是某些局部可能有变化，这与图 1 中最小生成树的情况相吻合；其次，随着序列数的增加最小生成树的平均度也跟着增大，序列条数越多，平均度就越大，但是随着序列数的增大，平均度增大的越缓慢，说明序列数越多的原癌基因，其最小生成树中节点连接得也越紧密。

再观察可以发现表 1 中的平均度几乎都约等于 2，由无向网络的度定义可知，每增加一个节点，网络中都只是多增加了一条边，即新加入的序列，和原网络中的一条序列亲缘关系较近。以上分析在一定程度上揭示了序列的自相似性和进化性。

3.3. 最小生成树的平均聚类系数

用 prim 算法得出最小生成树的邻接矩阵，在计算出的结果中发现，所有最小生成树的平均聚类系数都为 0。应为都是 0，所以就不再列表。聚类系数都为 0 说明了最小生成树网络中节点连接程度相对比较稀疏，也就是说序列之间的相互进化相对较难，同时也意味着不同原癌基因之间直接通过序列进化的可能性很低[12]。

3.4. 最小生成树的平均最短路径

15 种癌症以及它们的总和的最小生成树在各个 K 值下的平均最短路径在表 2 中给出。

最短路径是指从一个节点到另一个节点要经过的最少边数。最短路径越小，说明两条序列之间的亲缘关系就越近。综合分析表 2，可以看出大多数最小生成树网络的平均最短路径都服从 K 值越大平均最短路径就越小的规律，说明 K 值越大，整个网络中亲缘关系越密切；另外，在同一 K 之下，网络中序列数越多，平均最短路径在总体上也越大，说明序列越多，序列的平均亲缘关系将会越远。这可能是因为由祖先序列进化出来的子序列在漫长的进化过程中，也进化出了它的子序列，导致网络的拓扑结构越来

Table 2. The average shortest path to the minimum spanning tree, the number behind the cancer name is the total number of cancers
表 2. 最小生成树的平均最短路径，癌症名后面的数字是癌症所含的序列总数

K 值	6	9	12	15	18
原脑肿瘤基因 59	8.857978	6.132086	6.323787	4.980713	4.279369
原膀胱癌基因 74	8.825990	8.208812	8.445020	3.199926	3.199926
原黑色素瘤 74	8.626064	7.244354	10.363939	3.979267	3.928915
原肝癌基因 136	13.701634	10.297603	12.244553	6.094336	4.540741
原胃癌基因 138	12.754152	11.553052	10.973659	10.300645	6.789379
原肾癌基因 300	16.582430	16.081561	15.961405	14.048450	7.718484
原宫颈癌基因 319	17.757891	16.194949	23.276276	15.875633	20.791467
原甲状腺瘤基因 323	20.333519	21.101398	26.060535	14.084995	14.057920
原大肠癌基因 448	19.086739	19.625230	18.316085	22.743249	18.782788
原胰腺癌基因 513	20.066079	18.088435	20.369807	21.324394	13.651224
原卵巢癌基因 565	21.786042	21.513663	20.451553	25.345698	17.076797
原淋巴瘤基因 570	18.987069	21.136663	18.011797	12.775654	9.913533
原前列腺癌基因 587	24.107773	20.986912	21.630702	24.502980	16.243117
原肺癌基因 623	28.447673	22.420432	31.121469	23.778677	16.599051
原乳腺癌基因 668	22.128693	18.840294	19.090530	26.441278	16.683941
总原癌基因 2229	26.258700	23.655682	34.029590	36.193484	24.544119

越往外扩张，从而使得平均最短路径变大。细致观察可发现，序列多到一定程度时，平均最短路径的变化会趋于缓慢，说明可能在进化过程中祖先序列进化出来的子序列比其他序列更多，使得祖先序列在网络中处在中心位置，因此平均最短路径并不随序列数的增加而线性变化，而是渐渐趋于缓慢。这进一步说明了祖先序列在亲缘关系中的关键作用。另外，对于 DNA 序列的突变，一般认为是随机的，但是由于自然选择优胜劣汰的关系，只有一部分序列能继续生存。因此原癌基因的进化应该符合中性理论，且其进化可能是沿着最短最优的路径来进行。

4. 结论

本文收集了 15 种癌症的原癌基因序列，利用 CVTree 方法算出了 15 组序列和它们的总和序列共计 16 组距离矩阵，再通过距离矩阵，基于 prim 算法，分别计算出了它们的最小生成树邻接矩阵，并利用 pajek 软件画出最小生成树，最后计算了所有最小生成树的平均度、平均聚类系数和平均最短路径。对于同一种癌症的序列，在 5 个不同的 K 值下，最小生成树的平均度相等；所有的最小生成树的平均聚类系数在各个 K 值下均为 0；对于平均最短路径则呈现出随序列数的增加而总体增大的规律。通过对这三个网络参数的分析，揭示了原癌基因的进化可能是沿着最短的路径进行，并且其进化表现出了一定的自相似性；原癌基因网络中聚类系数为 0，说明网络连接较为稀疏，同时不同原癌基因序列之间亲缘关系并不密切。

基金项目

广西自然科学基金项目(No. 11262003)。

参考文献 (References)

[1] 吴一飞, 李灼日. 原癌基因 c-myc 与恶性肿瘤[J]. 医学临床研究, 2008, 25(9): 1698-1700.

[2] 兰晓瑜. C-myc 原癌基因启动区 G-四链体 DNA 序列对结肠癌细胞增殖的影响[D]: [硕士学位论文]. 太原: 山西医科大学, 2015: 5-30.

[3] Tabin, C.J., Bradley, S.M., Bargmann, C.I., *et al.* (1982) Mechanism of Activation of a Human Oncogene. *Nature*, **300**, 143-149. http://dx.doi.org/10.1038/300143a0

[4] Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., *et al.* (2013) Signatures of Mutational Processes in Human Cancer. *Nature*, **500**, 415-421. http://dx.doi.org/10.1038/nature12477

[5] Qi, J., Wang, B. and Hao, B.-L. (2004) Whole Proteome Prokaryote Phylogeny without Sequence Alignment: A *K*-String Composition Approach. *Journal of Molecular Evolution*, **58**, 1-11. http://dx.doi.org/10.1007/s00239-003-2493-7

[6] Qi, J., Luo, H. and Hao, B.-L. (2004) CVTree: A Phylogenetic Tree Reconstruction Tool Based on Whole Genomes. *Nucleic Acids Research*, **32**, W45-W47. http://dx.doi.org/10.1093/nar/gkh362

[7] Xu, Z. and Hao, B.L. (2009) CVTree Update: A Newly Designed Phylogenetic Study Platform Using Composition Vectors and Whole Genomes. *Nucleic Acids Research*, **37**, W174-W178. http://dx.doi.org/10.1093/nar/gkp278

[8] Albert, R. and Barabasi, A.L. (2002) Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, **74**, 47-97. http://dx.doi.org/10.1103/RevModPhys.74.47

[9] Barat, A. and Weigt, M. (2000) On the Properties of Small-World Network Models. *The European Physical Journal B-Condensed Matter and Complex Systems*, **13**, 547-560. http://dx.doi.org/10.1007/s100510050067

[10] Newman, M.E.J. and Watts, D.J. (1999) Renormalization Group Analysis of the Small-World Network Model. *Physics Letters A*, **263**, 341-346. http://dx.doi.org/10.1016/S0375-9601(99)00757-4

[11] Prim, R.C. (1957) Shortest Connection Networks and Some Generalizations., *The Bell System Technical Journal*, **36**, 1389-1401. http://dx.doi.org/10.1002/j.1538-7305.1957.tb01515.x

[12] 沈路明, 韦芳萍. 基于 cvtree 方法和复杂网络理论的癌症进化树分析[J]. 基因组学与应用生物学, 2014(2): 405-412.