# 系统发生网络构建算法综述

王 娟，郭茂祖

（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

**摘 要：** 物种的进化史通常被描述成一棵有根系统树，但是当物种进化过程中发生网状进化事件（如，杂交、重组和水平基因转移）时，物种的进化史不再适合被描述成系统树。系统发生网络是系统树的一般化，也是被用来描述物种的进化史，并可以描述物种的网状进化事件。而且系统发生网络也可以可视化冲突数据集，如由不同的基因得到的物种树。因此，系统发生网络的研究是生物信息的一个重要领域。介绍了系统发生网络的概念、发展、研究现状，总结了现有的系统发生网络构建算法。

**关键词：** 系统发生网络；网状进化事件；隐式网络；显式网络

**中图分类号：** TP301 **文献标识码：** A **文章编号：** 2095－2163(2014)01－0032－04

## A Survey of Phylogenetic Network Construction Algorithms

**WANG Juan, GUO Maozu**

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

**Abstract:** The evolutionary history of species is traditionally represented with a rooted phylogenetic tree, but when the evolution of species involves significant amounts of reticulate events (e.g., hybridization, recombination and horizontal gene transfer), the evolutionary history of the species has been described as the system is no longer suitable for tree. Phylogenetic networks are a generalization of phylogenetic trees, which is used to describe the evolutionary history of the species, and can describe the reticulate evolutionary events species. And the system network can also be visualized conflict data sets, such as obtained by different genes of the species tree. Therefore, research on phylogenetic networks is an important field of biological information. This paper introduces the system network concept, development, research present situation, summarizes the existing system network construction algorithm.

**Key words:** Phylogenetic Network; Reticulate Event; Abstract Network; Explicit Network

## 0 引 言

通常用系统树来表示一组分类单元的进化关系，这一模式有利于假设的讨论和检验。然而当描述更复杂的进化关系时，系统树的功能则略显不足。随着研究的逐渐深入，科学家们发现有些物种在进化过程中发生了网状进化事件，如反转（reversal）、移位（translocation）和转位（transposition）、重组（recombination）、水平基因转移（horizontal gene transfer，HGT）、杂交（hybridization）、基因转移或者基因重复和丢失[1-6]等，则此时生物的父代即不止一个，系统树不能描述各代之间的进化关系，因此促动了系统发生网络（phylogenetic network）的出现。系统发生网络构建方法及理论分析的研究是计算生物学的一个重要方向。系统发生网络是系统树的一般形式，又可译作系统演化网络、系统进化网络、进化网络。该种网络更适合那些发生了网状进化事件的数据，而且，对于树式进化模式（碱基的替代、插入、删除等）进化而来的数据，系统发生网络也可以实现数据中冲突信息的清晰表达，如由于不完全谱系分类机制或者是由于进化模型假设的不足引起的冲突信息[7]。系统发生网络是一个无环图，图中有些节点的父节点个数 ≥ 2（这种节点也被称为网络节点），如果图中没有网络节点，那么这时的系统发生网络就是一棵树。

系统发生网络根据拓扑结构分为无根（unrooted）网络和有根（rooted）网络；根据功能分为隐式（implicit）和显式（explicit）网络[8]。隐式网络（例如分割网络和准中位数网络）则可用来表示冲突信息，这些冲突信息可能来自各种原因，如模型误设（model misspecifi cation）；而显式网络则是尽力捕获生物进化过程中的网络进化事件，如杂交（hybridization）[9-10]、重组（recombination）[11-15]及水平基因转移（horizontal gene transfer，简称 HGT）[7,16-18]。显式网络中的内部节点表示祖先物种，且其中的网络节点对应所考虑的生物进化过程[14-16]，而隐式网络中网络节点没有任何生物解释。显式网络通常是有根的，因为生物进化过程本质上是有向的。然而有根系统发生网络可能是隐式网络，这取决于对相应网络进行构建和解释的具体方式[8]。

## 1 无根系统发生网络构建算法

无根系统发生网络是无根树的一般化。无根系统发生网络都是隐式网络，主要包括两类：分割网络（Split network）和准中位数网络（Quasi－median network）。在无根系统发生

收稿日期：2013－10－29
基金项目：国家自然科学基金（60932008，61172098）；高等学校博士学科点专项科研基金（20112302110040）；中央高校基本科研业务费专项资金（HIT. ICRST. 2010 022）。
作者简介：王 娟（1983－），女，内蒙古集宁人，博士研究生，主要研究方向：生物信息、分子进化、生物数学；
郭茂祖（1966－），男，山东夏津人，博士，教授，博士生导师，主要研究方向：机器学习、计算生物学、生物信息等。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

网络方面,分割( Split) 的概念起了重要作用。下面将详细给出分割的定义。

**定义 1**　设 $X$ 是一物种集合,$A$ 和 $B$ 是 $X$ 的非空子集,且 $A\cap B=\emptyset$ 和 $A\cup B=X$,则 $S=A\mid B$ 称为 $X$ 的一个分割。

有时将分割 $A\mid B$ 记为 $\frac{A}{B}$ 或者 $\frac{B}{A}$。分割 S 的大小记为 size( S) = min{ $\mid A\mid$, $\mid B\mid$} 。大小为 1 的分割称为是平凡的( trivial) 分割,否则称为非平凡的( non - trivial) 分割。设 $T$ 是 $X$ 上的一棵无根系统树,那么 $T$ 上的每一边定义了 $X$ 的一个分割。

分割网络可以从很多不同的数据集( 如距离矩阵、无根系统树集、序列及四分体) 构建得到。从这些数据构建分割网络时,大部分算法都是首先计算出一个加权分割集( 这里的权重可能表示的是距离或者特征变化量等) ,然后再由此加权分割集得到分割网络。由加权的分割集构建分割网络主要有两种方法: 凸包算法( convex hull) [19] 和圆形网络算法( circular network) [20] 。对于任何一分割集,凸包算法都能为 S 构建一个无根系统发生网络,且最坏情况是此网络包含指数级的节点数和边数。而圆形网络算法构建的网络仅包含平方级的节点数和边数。

从距离矩阵得到加权分割集的方法主要有 Neighbor - Net 方法[21] 和分割分解方法[22] 。从无根系统树构建加权分割集的主要方法有一致分割网络( consensus split network) 方法[23-24] 和 Z - 闭包( Z - closure) 算法[25-26] 。软件 Spitl-Tree4[27] 是一个用来推导无根系统发生网络的非常方便的工具,此软件可以从序列、距离、树或者是分割来推导得出无根系统发生网络,软件中收集了很多方法,如 Neighbor - net 方法以及 Z - 闭包算法。

## 2　有根系统发生网络构建算法

有根系统发生网络分为显式网络和隐式网络。显式网络理论上能很好地反映分类单元间的网状进化事件,由于进化是有向的,所以显式网络是有根的。Maddison 基于 rSPR ( rooted Subtree Prune and Regraft) 距离构建了系统发生网络[28] 。Nakhleh 等[29] 对 Maddison 的算法作了改进,提出了构建含有一个网络节点的系统发生网络的多项式算法,且此算法通过对基因树压缩的方式考虑了基因树中所带有的误差,使得此算法更具有实际应用价值。Wang 等[30] 及 Gusfield 等[31] 提出了从序列特征构建重组系统发生网络的算法。

Hein[32] 首次对构建系统树的最大简约法延伸到构建系统发生网络上。此后,Nakhleh 等[33] 旨在促进系统发生网络的构建和评估,而为每个网络定义了最简标准。文献 [33] 中提出的算法 Net2Trees 可用来计算网络的最简值,Net2Trees 算法的时间复杂度是指数级的。之后 ,Jin 等[34] 改进了这一 Net2Trees 算法,并提出了解决此问题的线性时间算法[35] 。以上介绍的最大简约法都是用相同的方式定义网络的最简值,都是将网络包含的所有树的最简值的最小值作为此网络的最简值。Kannan 等[36] 提出了另一种网络最简值的定义,即可定义为网络所有边的替换代价之和,并将计算系统树最优简约值( optimum parsimony score) 的 Sankoff 等[37-38] 方法延伸到系统网络上。

Jin 等[39] 提出了构建系统发生网络的最大似然法,首先,基于树的似然值给出此网络的似然值计算公式,且设计了启发式算法来计算此值,然后利用分支定界启发式算法及 EM 算法搜索最优网络,并且对真菌和质体中 的 15 种生物及古细菌中的 14 种生物分别构建了水平基因转移网络。Snir 等[40-41] 为构建和分析系统发生网络提出了一个新的概率模型 NET - HMM。模型中结合了最大似然法及马尔科夫模型,且假设 DNA 序列或者核苷酸序列上的相邻位点的进化是相互依赖的,这一假设与生物实际过程更为相符。在此模型中,隐状态是系统发生网络所包含的树。

隐式网络方面,Huson 等提出的 cluster network 方法是利用网络弹出算法( network - popping algorithm) 来构建有根隐式网络方法[42] 。此方法首先构建哈塞图( Hasse diagram) ,然后在此基础上以添加边的方式构建网络节点。其后 Huson 等[43] 提出了 galled network 方法,这是首先利用种子增长算法( seed - growing algorithm) 找出输入树集合的 RMCS 问题的解,即,去掉一些物种后的树集是不冲突的,这时可以为不冲突的树集构建一棵系统树 T,最后再将去掉的物种添加到 T 上,从而得到系统发生网络。Van Iersel 等又提出了 CASS 方法[7] ,此方法所构建的网络与实际生物网络更加相符,但是当所构建的网络很大时,该方法速度较慢,运行时间也长,不利于使用者在较短时间内得到结果网络。

程序 Dendroscope[44] 主要可用来计算有根系统发生网络,其中包含一些构建隐式网络的方法,如 CASS 方法、galled network 方法及 cluster network 方法; 程序中还包括一些构建显式网络的方法,如杂交网络方法。

## 3　结论与展望

本文对现有的系统发生网络构建方法进行概述。系统发生网络主要用于两种方式: 描述发生了网状进化事件的物种进化史、表示冲突的进化信息。随着数据量的增加,提出快速有效的构建系统发生网络的方法则已成为刻不容缓的研究任务。将系统发生网络应用到实际生物研究必将成为下一步的发展趋势。

## 参考文献:


[1] DELWICHE C F, PALMER J D. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids [J] . Molecular Biology and Evolution, 1996, 13(6):873 - 882.

[2] DOOLITTLE WF. Phylogenetic classification and the universal tree [J] . Science, 1999, 284(5423):2124 - 2128.

[3] GRIFFITHS R C, MARJORAM P. Ancestral inference from samples of DNA sequences with recombination [J] . Journal of Computational Biology, 1996, 3(4):479 - 502.

[4] RIESEBERG L H. Hybrid origins of plant species [J] . Annual review of Ecology and Systematics, 1997:359 - 389.

[5] SNEATH P. Cladistic representation of reticulate evolution [J] . Systematic Zoology, 1975, 24(3):360 - 368.

[6] SYVANEN M. Cross - species gene transfer; implications for a new theory of evolution [J] . Journal of theoretical Biology, 1985, 112


?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

(2):333 - 343.

[7] VAN IERSEL L, KELK S, RUPP R, et al. Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters [J]. Bioinformatics, 2010, 26(12):i124 - i131.

[8] HUSON D H, SCORNAVACCA C. A survey of combinatorial methods for phylogenetic networks [J]. Genome Biology and Evolution, 2011, 3:23.

[9] MADDISON W P. Gene trees in species trees. [J]. Systematic Biology, 1997, 46(3):523 - 536.

[10] LINDER C R, RIESEBERG L H. Reconstructing patterns of reticulate evolution in plants [J]. American Journal of Botany, 2004, 91(10):1700 - 1708.

[11] HEIN J. A heuristic method to reconstruct the history of sequences subject to recombination [J]. Journal of Molecular Evolution, 1993, 36(4):396 - 405.

[12] SONG YS, HEIN J. Constructing minimal ancestral recombination graphs [J]. Journal of Computational Biology, 2005, 12(2):147 - 169.

[13] HUSON D H, KLOEPPER T H. Computing recombination networks from binary sequences [J]. Bioinformatics, 2005, 21(suppl 2): 159 - 165.

[14] GUSFIELD D. Optimal, effi cient reconstruction of root - unknown phylogenetic networks with constrained and structured recombination [J]. Journal of Computer and System Sciences, 2005, 70(3):381 - 398.

[15] GUSFIELD D, BANSAL V. A fundamental decomposition theory for phylogenetic networks and incompatible characters [C]//Research in Computational Molecular Biology. Berlin Heidelberg: Springer, 2005:217 - 232.

[16] GUSFIELD D, HICKERSON D, EDDHU S. An efficiently computed lower bound on the number of recombinations in phylogenetic networks: Theory and empirical study [J]. Discrete Applied Mathematics, 2007, 155(6):806 - 830.

[17] NAKHLEH L. Evolutionary phylogenetic networks: models and issues [M]//Problem Solving Handbook in Computational Biology and Bioinformatics. Berlin Heidelberg: Springer, 2011:125 - 158.

[18] SEMPLE C. Hybridization networks [M]. Canterbury: Department of Mathematics and Statistics, University of Canterbury, 2006: 1 - 38.

[19] BANDELT H J, FORSTER P, SYKES B C, et al. Mitochondrial portraits of human populations using median networks [J]. Genetics, 1995, 141(2):743.

[20] DRESS AW, HUSON D H. Constructing splits graphs [J]. Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2004, 1(3):109 - 115.

[21] BRYANT D, MOULTON V. Neighbor - net: an agglomerative method for the construction of phylogenetic networks [J]. Molecular biology and evolution, 2004, 21(2):255 - 265.

[22] BANDELT H J, DRESS A W. A canonical decomposition theory for metrics on a finite set [J]. Advances in mathematics, 1992, 92(1):47 - 105.

[23] HOLLAND B, MOULTON V. Consensus networks: a method for visualising incompatibilities in collections of trees [M]//Algorithms in bioinformatics. Berlin Heidelberg: Springer, 2003:165 - 176.

[24] HOLLAND B R, HUBER K T, MOULTON V, et al. Using consensus networks to visualize contradictory evidence for species phylogeny [J]. Molecular Biology and Evolution, 2004, 21(7):1459 - 1461.

[25] WHITFIELD J B, CAMERON S A, HUSON D H, et al. Filtered Z - closure supernetworks for extracting and visualizing recurrent signal from incongruent gene trees [J]. Systematic biology, 2008, 57(6):939 - 947.

[26] HUSON D H, DEZULIAN T, KLOPPER T, et al. Phylogenetic super - networks from partial trees [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2004, 1(4): 151 - 158.

[27] HUSON D H, BRYANT D. Application of phylogenetic networks in evolutionary studies [J]. Molecular biology and evolution, 2006, 23(2):254 - 267.

[28] MADDISON W P. Gene trees in species trees [J]. Systematic biology, 1997,46(3):523 - 536.

[29] NAKHLEH L, WARNOW T, LINDER C R. Reconstructing reticulate evolution in species: theory and practice [C]//Proceedings of the eighth annual international conference on Research in computational molecular biology. New York: ACM Press, 2004: 337 - 346.

[30] WANG L, ZHANG K, ZHANG L. Perfect phylogenetic networks with recombination [J]. Journal of Computational Biology, 2001, 8(1):69 - 78.

[31] GUSFIELD D. Optimal, efficient reconstruction of root - unknown phylogenetic networks with constrained and structured recombination [J]. Journal of Computer and System Sciences, 2005, 70(3):381 - 398.

[32] HEIN J. Reconstructing evolution of sequences subject to recombination using parsimony [J]. Mathematical biosciences, 1990, 98(2):185 - 200.

[33] NAKHLEH L, JIN G, ZHAO F, et al. Reconstructing phylogenetic networks using maximum parsimony [C]//Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE. California: IEEE Computer Society Press, 2005:93 - 102.

[34] JIN G, NAKHLEH L, SNIR S, et al. Efficient parsimony - based methods for phylogenetic network reconstruction [J]. Bioinformatics, 2007, 23(2):e123 - e128.

[35] JIN G, NAKHLEH L, SNIR S, et al. A new linear - time heuristic algorithm for computing the parsimony score of phylogenetic networks: Theoretical bounds and empirical performance [M]//Bioinformatics Research and Applications. Berlin Heidelberg: Springer, 2007:61 - 72.

[36] KANNAN L, WHEELER W C, et al. Maximum Parsimony on Phylogenetic networks [J]. Algorithms for Molecular Biology, 2012, 7: 9.

[37] SANKOFF D. Minimal mutation trees of sequences [J]. SIAM Journal on Applied Mathematics, 1975, 28(1):35 - 42.

[38] SANKOFF D, ROUSSEAU P. Locating the vertices of a Steiner tree in an arbitrary metric space [J]. Mathematical Programming, 1975, 9(1):240 - 246.

(下转第37页)

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

控制,可对所有器件引脚、SFR 总线和 I/O 口弱上拉功能实现监测和控制。

在设计中,文中采用了 C8051F340 型号,其时钟频率为 48MHz,64K 字节的 flash 闪存和 4 352 字节的 RAM,内部振荡器精度达 0.25%,完全可以省去外部晶振而不会影响通讯波特率,在通讯接口方面除了 UART 串口外,还增加了 SPI 和 SMBUS 等,由此降低了硬件设计的难度[6]。其应用电路图如图 4 所示。

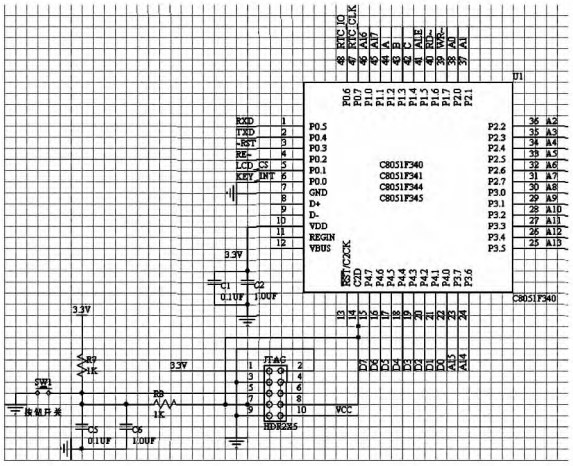


图 4 C8051F340 应用电路图

**Fig. 4 The circuit diagram of C8051F340**

## 3 软件设计

系统整体的软件设计主要分为两部分: 门禁机底层代码的设计和主机界面信息管理的代码设计,两者间的沟通联系就是借助通讯协议。因而,对于系统分析师来说,设计一个性能优良的通讯协议尤为重要。为达此目的,就必须对系统整体非常了解,对客户的需求和系统实现的功能也必须高度清楚,其中可以借助一些系统分析工具或语言来实现系统的分析规划,同时对复杂的系统进行建模[7]。下面分别对这两部进行分析和阐述。

### 3.1 底层设计

底层代码设计只需画出流程图,按照模块化设计思路展开设计,由于 C8051 系列可以通过 JTAG 在线仿真调试,借助基于 Keil C 平台上开发的 C 语言,其编译的效率和汇编相差不到 10%。整个底层代码主要分为: 射频模块、通讯模块、I/O 输入输出模块、存储模块。

### 3.2 管理界面设计

管理界面一般分为前台界面信息管理和后台数据库,从软件模块化设计和其后的维护升级来看,多是采用三层结构: 前台界面层、中间层、后台数据层。其中,前台界面层可以采用高级语言如: Visual C + +、Visual Basic、C + + Builder、Delphi,基于 Web 则可采用. net 构架的 C#或者 Java 语言,可随编者喜好任意选择,但控制标准则为设计界面简洁、且操作方便直观。中间层主要为 DLL 库或者 Activex 控件,其中包含一些对底层门禁机操作的代码,这将随硬件改动而同时升级,一般无需改动[7]。后台数据层就是基于 ODBC 或者 JDBC 的数据库连接,一般采用 Microsoft 公司 SQL Server,数据库架构的灵活设计,可以方便今后数据库的迁移升级,如改为 ORACLE、DB2 等,也利于日后整合至企业的 ERP 系统里。

## 4 结束语

本文通过结合作者在实际工程应用中的体会介绍了门禁系统的组成。近几年来,由于识别技术高速发展和嵌入式系统性能提高及其成本的相应降低,门禁系统融合了生物识别、RFID 身份识别以及图像识别等技术,开展了有关研究,且其中各有优缺点。如何实现快速有效地识别,以及在不同的场合下提高识别的效果则是目前正在研究的重点和热点问题。

## 参考文献:

[1] 瞿小玲. 基于 RFID 的低功耗智能门禁系统的设计与研究[D]. 成都: 成都理工大学,2012.

[2] 游战清,李苏剑. 无线射频识别技术(RFID)理论与应用[M]. 北京: 电子工业出版社,2004: 285 - 287.

[3] 王汝琳. 智能门禁控制系统[M]. 北京: 电子工业出版社,2004,9: 159 - 163.

[4] 陆永宁. IC 卡应用系统[M]. 南京: 东南大学出版社,2000: 215 - 220.

[5] 何立民. 从 Cygnal C8051F 看 8 位单片机发展之路[J]. 单片机与嵌入式系统应用,2002: 56 - 62.

[6] Michael Jackson. 软件开发问题框架: 现实世界问题的结构化分析[M]. 北京: 机械工业出版社,2005: 3 - 6.

[7] Steve Hoberman. 数据建模: 分析与设计的工具和技术[M]. 北京: 机械工业出版社,2004: 377 - 382.

[8] Dean Leffingwell, Don Widrig. 软件需求管理: 用例方法[M]. 北京: 中国电力出版社,2004: 234 - 240.

(上接第 34 页)

[39] JIN G, NAKHLEH L, SNIR S, et al. Maximum likelihood of phylogenetic networks[J]. Bioinformatics, 2006, 22(21): 2604 - 2611.

[40] SNIR S, TULLER T. Novel phylogenetic network inference by combining maximum likelihood and hidden Markov models[M]//Algorithms in Bioinformatics. Berlin Heidelberg: Springer, 2008: 354 - 368.

[41] SNIR S, TULLER T. The NET - HMM approach: Phylogenetic network inference by combining maximum likelihood and hidden markov models[J]. Journal of bioinformatics and computational biology, 2009, 7(4): 625 - 644.

[42] HUSON D H, RUPP R. Summarizing multiple gene trees using cluster networks[M]//Algorithms in Bioinformatics. Berlin Heidelberg: Springer, 2008: 296 - 305.

[43] HUSON D H, RUPP R, BERRY V, et al. Computing galled networks from real data[J]. Bioinformatics, 2009, 25(12): 85 - 93.

[44] HUSON D H, SCORNAVACCA C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks[J]. Systematic biology, 2012, 61(6): 1061 - 1067.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net