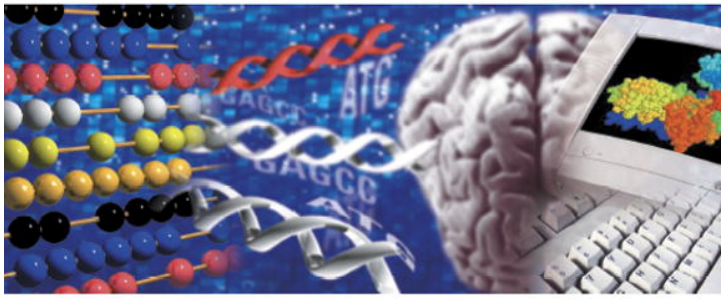
Volume 31, Issue 3, June 2007 ISSN 1476-9271

Computational Biology and Chemistry

Editors: James Crabbe, Andrzej Konopka



www.elsevier.com/locate/cbac

Available online at ScienceDirect www.sciencedirect.com

Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry 31 (2007) 173–177
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Optimal peeling order for pedigrees with incomplete genotypic information

Nadezhda M. Belonogova [a,b], Tatiana I. Axenovich [a,b,*]

[a] *Institute of Cytology & Genetics, Siberian Division of Russian Academy of Sciences (SD RAS), Lavrentyeva Ave. 10, Novosibirsk 630090, Russia*
[b] *Department of Cytology & Genetics, Novosibirsk State University, Novosibirsk, Russia*

Received 13 March 2007; accepted 17 March 2007

Abstract

The likelihood approach is common in linkage analysis of large extended pedigrees. Various peeling procedures, based on the conditional independence of separate parts of a pedigree, are typically used for likelihood calculations. A peeling order may significantly affect the complexity of such calculations, particularly for pedigrees with loops or when many pedigrees members have unknown genotypes. Several algorithms have been proposed to address this problem for pedigrees with loops. However, the problem has not been solved for pedigrees without loops until now.

In this paper, we suggest a new graph theoretic algorithm for optimal selection of peeling order in zero-loop pedigrees with incomplete genotypic information. It is especially useful when multiple likelihood calculation is needed, for example, when genetic parameters are estimated or linkage with multiple marker loci is tested. The algorithm can be easily introduced into the existing software packages for linkage analysis based on the Elston–Stewart algorithm for likelihood calculation.

The algorithm was implemented in a software package PedPeel, which is freely available at http://mga.bionet.nsc.ru/nlru/.

Keywords: Pedigree likelihood; Peeling; Algorithm; Graph theory; Software

1. Introduction

Large extended pedigrees comprise a valuable source of information for genetic mapping of complex traits. The likelihood approach is common in linkage analysis of such data. The likelihood of arbitrary pedigree can be written in a general form as follows

$$\mathrm{LH} = \sum_{\vec{G}} P(\vec{X}|\vec{G})P(\vec{G}),$$

where $\vec{X}$ is a matrix of observed phenotypes for all pedigree members and $\vec{G}$ is a matrix of their unobserved genotypes, and the summation is performed over all possible genotypic configurations (Elston and Stewart, 1971). The complexity of this formula calculation, or running time, is determined by the number of possible genotypic combinations and may be written as

$$\mathrm{CC} = \prod_{i=1}^{N} |g_i|,$$

where N is the pedigree size and $|g_i|$ is a number of possible genotypes for ith pedigree member. When the pedigree is large or the number of possible genotypes is large, this estimate is rather high for a practical use. Therefore, different "peeling" algorithms, which reduce the running time, are used for likelihood calculation (Elston and Stewart, 1971; Cannings et al., 1978; Lander and Green, 1987; Kruglyak et al., 1996). The idea underlying the peeling procedures is based on the fact that some portions of the likelihood function are conditionally independent and thus the likelihood of these parts may be evaluated sequentially. The memory space and running time necessary for likelihood calculation heavily depend on a proper selection of conditionally independent portions of the likelihood function and on the order of their peeling.

Among existing algorithms, Elston–Stewart peeling algorithm is optimal for pedigrees including hundreds of members

* Corresponding author at: Institute of Cytology & Genetics, Siberian Division of Russian Academy of Sciences (SD RAS), Lavrentyeva Ave. 10, Novosibirsk 630090, Russia. Tel.: +7 383 333 2840; fax: +7 383 333 1278.
E-mail address: aks@bionet.nsc.ru (T.I. Axenovich).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.03.004

(Elston and Stewart, 1971; Cannings et al., 1978). The algorithm is suited for a step-by-step reduction of the pedigree size through collapsing the genetic information about some pedigree members onto other pedigree members. Within the framework of Elston–Stewart algorithm, several algorithms for selection of optimal peeling sequences were proposed (Cannings et al., 1978; Harbron, 1995; Lange and Boehnke, 1983; Thomas, 1986; Fernandez and Fernando, 2002). However, all of these algorithms are applicable only to pedigrees with loops for calculation of their exact likelihood. Until recently, it was believed that the problem of determining the optimal peeling order is associated with looped pedigrees only. For pedigrees without loops the effective peeling algorithms use nuclear pedigrees (NPs) as conditionally independent portions of pedigree. In this case the memory space is no more $\sum_{i=1}^{N}|g_i|$ and running time is proportional to the number of NPs in the pedigree. Therefore, it was accepted that peeling order can be arbitrary when pedigree has no loops. This conclusion is correct as long as the pedigree does not include many members with unknown genotypes. The number of possible genotypes for such pedigree members may be several orders of magnitude greater than for genotyped pedigree members, and the optimal order of peelings plays a crucial role in decreasing the running time of the algorithm.

In this paper, we suggest a new algorithm for optimal selection of peeling order in zero-loop pedigrees with incomplete genotype information and test its efficiency using several large pedigrees.

2. Peeling procedure

Any pedigree can be represented by a set of NPs, with two NPs connected to each other by an individual belonging to both of them (Fig. 1A–B). These individuals are called connectors. An NP with a single connector is called a terminal NP. In pedigrees without loops every peeling operation condenses information about terminal NP on genotypes of corresponding connector. As a result, the number of NPs in the pedigree is decreased by one; for the NP adjacent to the peeled NP, the number of connectors is decreased by one. If the number of an NP's connectors gets down to one, the NP becomes a terminal one. To calculate the likelihood function for a large pedigree, all NPs must be sequentially peeled. For zero-loop pedigree this can be done using two NP peeling operations: peeling on a parent or peeling on one of the offspring. What kind of peeling operation will be used for given NP depends on the sequence of previous peelings. Since different peeling operations have different complexity of calculation, the optimal order of peelings plays a crucial role in decreasing the running time of the likelihood calculation.

To compare the running time for different peeling operations, lets consider a NP with a father (f), a mother (m) and a set of children numbered from 1 to n. Let x_i be phenotype and marker genotypes for the ith member of the pedigree, let g_i be the set of possible genotypes of this member. The set of possible genotypes consists of combinations of unobserved genotypes controlling given trait and a known marker genotypes. If the individual is not genotyped, a set of all possible marker genotypes has to be considered.

2.1. Peeling NP on a parent

Without loss of generality we consider the situation that the NP's connector is the father. As a result of this peeling the information about all members of the considered NP is condensed on the connector under each of its possible genotypes:

$$\Pr(X_f|g_f) = \Pr(x_f, x_m, x_1, \ldots, x_n|g_f) = \Pr(x_f|g_f) \times \sum_{g_m}\Pr(x_m, g_m)\prod_{i=1}^{n}\sum_{g_i}\Pr(g_i|g_f, g_m)\Pr(x_i|g_i). \quad (1)$$

The running time for computing this formula is proportional to

$$CC_{par} = |g_f||g_m|\sum_{i=1}^{n}|g_i|, \quad (2)$$

where $|g_j|$ is a number of possible genotypes for jth individual.

2.2. Peeling NP on one of the offspring

Consider peeling on the child with number j. As a result of this peeling the information about all NP members is condensed on this child for each of its possible genotypes:

$$\Pr(X_j, g_j) = \Pr(x_f, x_m, x_1, \ldots, x_n, g_j) = \Pr(x_j, g_j)\sum_{g_f}\Pr(x_f, g_f)\sum_{g_m}\Pr(x_m, g_m) \times \prod_{i\neq j}\sum_{g_i}\Pr(g_i|g_f, g_m)\Pr(x_i|g_i). \quad (3)$$

The running time for this formula is proportional to

$$CC_{off} = |g_f||g_m||g_j|\left(\sum_{i\neq j}|g_i| + 1\right). \quad (4)$$

The difference between running times for these two peeling operations is proportional to

$$CC_{par} - CC_{off} = |g_f||g_m|\{|g_j| - 1\}\sum_{i\neq j}|g_i|. \quad (5)$$

This difference is non-negative because the number of possible genotypes, $|g_i|$, is positive for each pedigree member. This difference is zero, only if the NP has a single child or if a child, who is NP's connector, has one possible genotype. The difference between running times for two peeling operations may be considerable. For example, when $n = 5$ and $|g| = 10$ for each member of the NP, the running time given by (4) is eight times greater than the one given by (2). In general, peeling on a parent is more preferable than peeling on one of the offspring. There are two special cases when optimal peeling operation may be always used.

Case #1: Optimal peeling on a parent is always possible for any terminal NP, having a parent as a connector.
Case #2: In any peeling order, the final NP can be always optimally peeled on a parent.

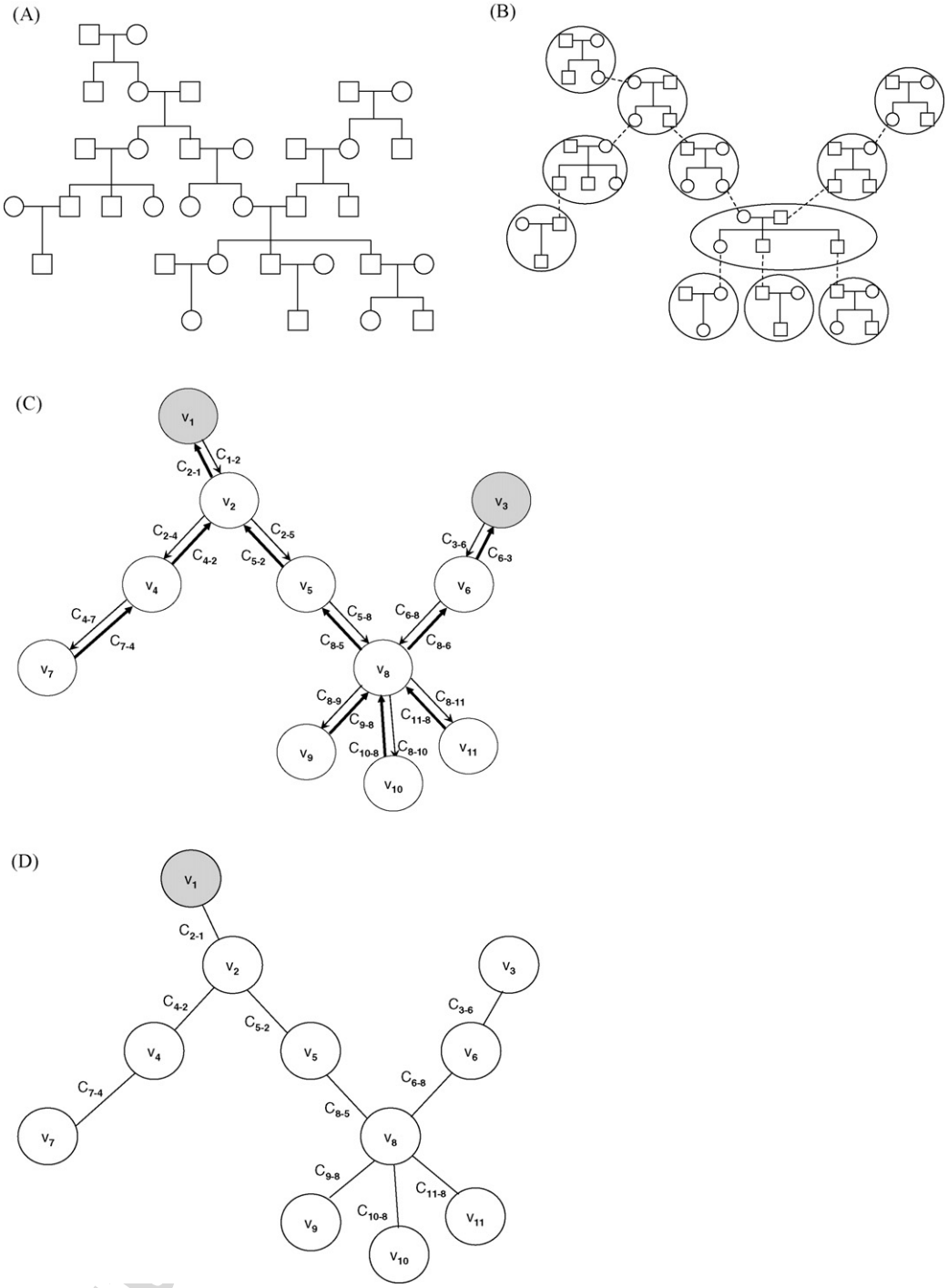


Fig. 1. Pedigree graph. (A) Initial pedigree; (B) the same pedigree represented as a set of nuclear pedigrees; (C) weighted pedigree graph (c_{i-j} is the weight of edge directed from V_i to V_j; bold arrow corresponds to preferable peeling on a parent; the potential final nodes are indicated by gray color); (D) rooted weighted tree with V_1 as a root.

When none of the pedigree's NPs have both parents with the ancestors, the likelihood may be optimally calculated using peelings on a parent only. If there is an NP with both parents having ancestors in the pedigree, the peeling on one of the offspring has to be used. In this case we must select what kind of NPs will be peeled on one of the offspring to ensure minimum running time of the pedigree likelihood calculation.

3. Optimal peeling order

The order of peelings determines what kind of peeling operations is used for each NP. The running time for pedigree likelihood calculation corresponding to a given peeling order is calculated as a sum of the running times for peeling operations specified for each NP. An optimal peeling order is the

one with minimum running time for pedigree likelihood calculation.

To select optimal peeling order, consider a pedigree without loops as a directed graph G with NPs as vertices and with weights assigned to directed edges. Every two adjacent vertices corresponding to two NPs, NP_i and NP_j with connector ij, are connected by two oppositely directed weighted edges (Fig. 1C). The weights of directed edges $NP_i \rightarrow NP_j$ and $NP_j \rightarrow NP_i$ are equal to the running time for peeling of NP_i or NP_j, respectively, on connector ij given by (2) or (4).

Select one vertex, NP_R, and create a simple weighted graph T with NPs as vertices and two NPs connected by an edge if and only if there is a directed edge between them in G. The weight of edge between NP_i and NP_j in graph T is equal to the weight of that edge between NP_i and NP_j in G, which is directed to NP_R. Observe that T is a rooted weighted tree with NP_R as a root (Fig. 1D).

Due to the structure of a graph G, each NP may be selected as a root and the corresponding rooted tree can be created. The peeling operation is equivalent to the cutting the leaf of T, i.e., a vertex adjacent to only one vertex. The peeling algorithm begins with the leaves and works up ending at the root. It is known that there is a unique path between any vertex and the root in T and the order of the leaves cutting is exclusively determined by specifying the root vertex (West, 2001). Therefore, we can calculate the running time, C, for the pedigree peeling as a sum of the weights of all edges of T plus the running time for peeling of root (final NP) given by (2). The rooted tree with minimum C value corresponds to the optimal peeling order.

In general, any vertex must be considered as a potential root and the values of C have to be calculated for all rooted tree graphs T corresponding to G. However taking into account the origin of the pedigree described by the graph G, several vertices may be excluded from the set of potential roots. Only a vertex corresponding to NP, where both parents have no ancestors, may be considered as a potential root. Indeed, any terminal NP with a parent having an ancestor may be peeled on this parent with the optimal running time (case #1) and selecting the adjacent NP as a root is more efficient. Moreover, last two NPs may be always peeled on their parents (case #2) and the NP with offspring as a connector must be final NP, or root. Thus, to select the optimal order of pedigree peelings, we must compare the values of C for those rooted tree graphs T, where the roots correspond to NPs in which both parents have no ancestors.

4. Algorithm

The procedure for selecting the optimal order of peelings may be summarized as follows. First, the pedigree is presented as a directed graph G with NPs defined as vertices. Second, the weight of each edge $NP_i \rightarrow NP_j$ of G is defined as a running time for peeling NP_i on connector ij given by (2) or (4). Third, the set of potential roots is defined as a set of NPs where both parents are ancestors of the pedigree. Finally, for each potential root the rooted weighted tree graph T corresponding to G is created and the total weight is calculated as a sum of weights of all edges plus the running time for the root peeling given by (2). The minimal total weight corresponds to the optimal rooted tree graph.

We implemented this algorithm in a software package PedPeel, which is freely available at http://mga.bionet.nsc.ru/nlru/.

5. Number of possible genotypes

The set of possible genotypes consists of combinations of all possible unobserved genotypes controlling trait and known marker genotype. In the simplest case of diallelic trait locus (alleles A and a), the number of possible genotype combinations equals to three for homozygous marker (AM_1/AM_1, AM_1/aM_1 and aM_1/aM_1) and four for heterozygous one (AM_1/AM_2, AM_1/aM_2 AM_2/aM_1 and aM_1/aM_2). If the individual is not genotyped, the set of possible marker genotypes includes all possible genotypes. The number of these genotypes is equal to $0.5\,h(h+1)$, where h is the number of possible haplotypes calculated as a product of the number of alleles controlling the trait and polymorphic alleles of the marker loci. Based on the pedigree data, some genotypes may be eliminated from the set of possible genotypes (Lange and Goradia, 1987; O'Connell and Weeks, 1999). However, when several generations of ancestors are not measured, we have to consider the largest possible number of genotypes for the majority of these ancestors. This number is 10 for diallelic marker, 55 for a marker with 5 alleles and 210 for a marker with 10 alleles.

6. Efficiency of the algorithm

To demonstrate the efficiency of our algorithm, we tested it on the likelihood calculation for all possible peeling orders in three large zero-loop pedigrees, which were produced from several pedigrees with multiple loops by breaking all loops. The origin of these pedigrees and selection of loop breakers were described in our previous paper (Axenovich et al., submitted for publication). The analyzed pedigrees differ by size and percentage of measured individuals (Table 1). We considered the marker locus with five alleles. The number of possible genotypes was assumed to be 4 for measured individuals and 55 for non-measured ones.

Running times were calculated for all possible orders of peelings, defined by selecting each NP as a root of the tree graph. The distributions of the total running times for the three pedigrees are shown in Table 1. In all cases, it was found that there is a unique peeling order with minimum running time. Any other order gave a value considerably larger than the optimal: 3.05×10^4 compared to 6.53×10^4 for the human and 4.13×10^6 compared to 1.24×10^7 for the silver fox pedigrees. The arctic fox pedigree provides the smallest difference between the optimal and the second best value for the running time of a peeling algorithm. However, the great part of distribution (333 of 359) was greater than 6.56×10^6 whereas the minimum value was 4.60×10^6. It means that the chance to select the peeling order with minimum or nearly minimum running time under arbitrary selection of peeling order is very small for all analyzed pedigrees. Table 1 shows that the average running time over all peeling orders is 1.80–3.18 times higher than the corresponding optimal value.

Table 1
Pedigree structures and running time for likelihood calculation

Pedigree	Number			Running time			
	Individuals	Measured	NPs	Maximum	Mean ± S.E.	Minimum (optimal peeling)	Ratio of mean to optimal
Human[a]	114	112	39	6.760×10^4	$6.557 \pm 0.093 \times 10^4$	3.054×10^4	2.15
Silver fox[b]	1845	1056	788	1.335×10^7	$1.311 \pm 0.001 \times 10^7$	4.126×10^6	3.18
Arctic fox[c]	952	617	359	2.664×10^7	$8.279 \pm 0.229 \times 10^6$	4.604×10^6	1.80

[a] Zero-loop fragment of pedigree from Dutch genetically isolated population (Pardo et al., 2005).
[b] Zero-loop fragment of the silver fox pedigree formed on the base of the breeding records of Experimental Farm of the Institute of Cytology and Genetics, Novosibirsk, Russia.
[c] Zero-loop fragment of the arctic fox pedigree formed on the base of the breeding records maintained at the Puskinsky fur farm, Moscow district, Russia (Axenovich et al., 2007).

Thus, the proposed algorithm may lead up to three-fold decrease of the running time.

Our algorithm can be easily introduced into those existing software packages for linkage analysis, which are based on the Elston–Stewart algorithm of likelihood calculation. Usually a peeling order is defined in preliminary analysis of pedigree structure and then this order is used for likelihood calculation. The algorithm is especially useful when multiple likelihood calculation is needed, for example, when the genetic parameters are estimated or the linkage with multiple marker loci is tested. Our algorithm was implemented in the software packages for segregation and linkage analysis, which are available from http://mga.bionet.nsc.ru/.

Acknowledgements

We would like to thank Professors Maria Axenovich and Pavel Borodin, for helpful comments. This research was supported by the joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.016.009), grant from the Russian Foundation for Basic Research (RFBR 07-04-00120), and the Program of Russian Academy of Sciences "Biodiversity and gene pools dynamics".

References

Axenovich, T.I., Zorkoltseva, I.V., Liu, F., Kirichenko, A.V., Aulchenko, Y.S. Breaking loops in large complex pedigrees. Hum. Hered., submitted for publication.

Axenovich, T.I., Zorkoltseva, I.V., Akberdin, I.R., Beketov, S.V., Kashtanov, S.N., Zakharov, I.A., Borodin, P.M., 2007. Inheritance of litter size at birth in farmed arctic foxes (*Alopex lagopus*, Canidae, Carnivora). Heredity 98 (2), 99–105.

Cannings, C., Thompson, E.A., Skolnick, E.H., 1978. Probability functions on complex pedigrees. Adv. Appl. Prob. 10, 26–61.

Elston, R.C., Stewart, J., 1971. A general model for the genetic analysis of pedigree data. Hum. Hered. 21 (6), 523–542.

Fernandez, S.A., Fernando, R.L., 2002. Technical note: determining peeling order using sparse matrix algorithms. J. Dairy Sci. 85, 1623–1629.

Harbron, C., 1995. A pedigree-based algorithm for finding efficient peeling sequences. IMA J. Math. Appl. Med. Biol. 12 (1), 13–27.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., Lander, E.S., 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. 58, 1347–1363.

Lander, E.S., Green, P., 1987. Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. U.S.A. 84, 2363–2367.

Lange, K., Boehnke, M., 1983. Extensions to pedigree analysis. V: Optimal calculation of Mendelian likelihoods. Hum. Hered. 33, 291–301.

Lange, K., Goradia, T.M., 1987. An algorithm for automatic genotype elimination. Am. J. Hum. Genet. 40, 250–256.

O'Connell, J.R., Weeks, D.E., 1999. An optimal algorithm for automatic genotype elimination. Am. J. Hum. Genet. 65, 1733–1740.

Pardo, L.M., MacKay, I., Oostra, B., van Duijn, C.M., Aulchenko, Y.S., 2005. The effect of genetic drift in a young genetically isolated population. Ann. Hum. Genet. 69, 288–295.

Thomas, A., 1986. Optimal computations of probability functions for pedigree analysis. IMA J. Math. Appl. Med. Biol. 3, 167–178.

West, D.B., 2001. Introduction to Graph Theory, second ed. Prentice Hall.