半手足關係之重建在無父代資訊下

The Half Sibling Relationship Reconstruction without Parental Information

Yen Hung Chen[1] and Chih Hsuan Tzang[2]
Department of Computer Science,
University of Taipei, Taiwan
yhchen@utaipei.edu.tw[1], Wendytzang@Gmail.com[2]

摘要

本論文主要是探討一個組合最佳化問題：半手足關係重建問題(Half Sibling Relationship Reconstruction Problem)。給定 n 個物種(Species)，每個物種 i 有 ℓ 個locus，每個locus有兩個對偶基因以 $<a_{ij},b_{ij}>$ 表示，$1\le j\le \ell$，a_{ij} 及 b_{ij} 為物種 i 在第 j 個locus的2個對偶基因。在無法得知物種之間的父代資訊(Parental Information)下，我們想找出這 n 個物種是否有半手足關係(Half Sibling Relationship)，即分成 s 群，$1\le s\le n$，每個群內的物種必需是同父異母或同母異父的手足關係。本論文我們使用孟德爾遺傳4-allele定律來做為分群法則，即群內的每個物種的每個locus的所有對偶基因最多只能有4種。Half-Sibs Property定義為在群 S 內的每個物種 i，$i\in S$，其每個locus j，$1\le j\le \ell$，存在一個對偶基因 $P_j=<x_j,y_j>$（同一個父親或母親），使得 $a_{ij}\in P_j$ or $b_{ij}\in P_j$。半手足關係重建問題(Half Sibling Relationship Reconstruction Problem)定義為給定 n 個物種，每個物種 i 有 ℓ 個locus，每個locus有兩個對偶基因 $<a_{ij},b_{ij}>$，$1\le j\le \ell$，目的是要將 n 個物種進行分群使得分群數最少並滿足每個群必需服從Half-Sibs Property。半手足關係重建問題被證明為NP-Complete及APX-hard。本論文我們證明在 $\ell=1$ 時，此問題等價於節點涵蓋問題(Vertex Cover Problem)。因此在 $\ell=1$，我們有一個 $2-(2\ln\ln|V|/\ln|V|)$ 倍的近似演算法對於半手足關係重建問題。

1 緒論

手足關係(Sibling Relationships)的知識被廣泛地用在遺傳流行病學(genetic epidemiology)，保育生物學(conservation biology)及動物管理(animal management)。在有父(母)代的資訊下，要找物種間是否有手足關係是可以比較容易得知的[10]，然而在無父代的資訊下，如何重建手足關係的課題是生物學家這幾年所面臨的挑戰[1,3,5,6,7,13,14]。

基於孟德爾遺傳定律（Mendelian inheritance laws）: 4-allele rule 及 2-allele rule，Berger-Wolf 等人在 2005 年[3] (2007 年(期刊版)[7])定義了一個完全手足關係重建問題(Full Sibling Relationship Reconstruction Problem)在無父代的資訊下。給定 n 個物種(Species)，每個物種 i 有 ℓ 個 locus，每個 locus 有兩個對偶基因以 $<a_{ij},b_{ij}>$ 表示，$1\le j\le \ell$，a_{ij} 及 b_{ij} 為物種 i 在第 j 個 locus 的 2 個對偶基因，4-allele rule 為具有手足關係的物種群組 S 內每個物種其每個 locus 內的對偶基因最多只能有 4 個不一樣基因，即 $\forall$ $1\le j\le \ell$，$|\cup_{i\in S} a_{ij}\cup b_{ij}|\le 4$。而 2-allele rule 為在具有手足關係的物種群組 S 中每個物種其每個 locus 內的對偶基因內第一個基因（a_{ij}）最多只會有 2 個不一樣基因且第二個基因（b_{ij}）也最多只會有 2 個不一樣基因，即 $\forall$ $1\le j\le \ell$，$|\cup_{i\in S} a_{ij}|\le 2$ *and* $|\cup_{i\in S} b_{ij}|\le 2$。完全手足關係重建問題(Full Sibling Relationship Reconstruction Problem，FSRP) [3,4,7]為給定的 n 個物種，每個物種 i 有 ℓ 個 locus，每個 locus 有兩個對偶基因 $<a_{ij},b_{ij}>$，$1\le j\le \ell$，目的是要將這 n 個物種分群使其滿足 4-allele rule 或 2-allele rule 且群數要最小。Berger-Wolf 等人[3,7]證明 FSRP 為 NP-Complete 並轉換到集合涵蓋問題(Set Cover Problem)[8]後透過集合涵蓋的演算法解決 FSRP 在 4-allele rule[3,7]及 2-allele rule[4]並對一些物種進行電腦模擬。Ashley 等人[2]證明了 FSRP 不能設計出低於 1.0065 倍(當 locus 的數目為 $O(n^3)$)及 1.00014 倍(當 locus 的數目為 2）的近似演算法，除非 $RP=NP$。

因為 FSRP 的完全手足關係重建是基於同父同母的關係，所以在同一個群內的手足的每個 locus 會滿足其對偶基因最多 4 個不同(4-allele rule)或個別基因最多兩個不同(2-allele rule)，也就是基因是兩個來自父親兩個來自母親，因此最多 4 個不同或個別的基因是有兩個不同。但是當如果我們感興趣的是同父異母或同母異父的手足關係時，也就是半手足關係(Half Sibling Relationships)時，手足關係的重建就需要有不同的定義。因此 Sheikh 等人[11]定義了一個半手足關係重建問題(Half Sibling Relationship Reconstruction Problem)。給定 n 個物種，每個物種 i 有 ℓ 個 locus，每個 locus 有兩個對偶基因以 $<a_{ij},b_{ij}>$ 表示，$1\le j\le \ell$，a_{ij} 及 b_{ij} 為物種 i 在第 j 個 locus 的 2 個對偶基因，Half-Sibs Property 定義為在群 S 內的每個物種 i，$i\in S$，其每個 locus j，$1\le j\le \ell$，存在一個父代對偶基因 $P_j=<x_j,y_j>$(同

一個父親或母親)，使得 $a_{ij}\in P_j$ or $b_{ij}\in P_j$。半手足關係重建問題(Half Sibling Relationship Reconstruction Problem，HSRP)[11,12]定義為給定 n 個物種，每個物種 i 有 ℓ 個 locus，每個 locus 有兩個對偶基因$<a_{ij},b_{ij}>$，$1\le j\le \ell$，目的是要將 n 個物種進行分群使得分群數最少並滿足每個群必需服從 Half-Sibs Property。Sheikh 等人 [11] 證明 HSRP 為 NP-Complete 並轉換到集合涵蓋問題後透過集合涵蓋的演算法解決此問題並對一些物種進行電腦模擬。之後，Sheikh 等人[12]證明此HSRP為APX-hard並設計了一個Integer Linear Programming 解決此問題並對一些物種進行電腦模擬。因為目前針對 HSRP 較少談到近似演算法設計，本論文我們證明在 ℓ=1 時，HSRP 等價於節點涵蓋問題(Vertex Cover Problem)[8]。因此在 ℓ=1 時，透過節點涵蓋問題目前最好的近似演算法，我們有一個 *2-(2lnln|V|/ln|V|)*倍的近似演算法對於 HSRP。

底下我們舉例說明半手足關係重建問題。

給定 5 個已知基因的物種，每個物種內含有 2 個 locus，每個 locus 含有 2 個對偶基因如表 1 所示。表 2 為 HSRP 的最佳解(分成兩群)。每個群滿足 Half-Sibs Property：群 1 有 4 個手足，locus 1 存在一個父代對偶基因<35,7>，使得物種 1-4 內在 locus 1 會有一個基因屬於 35 或 7。locus 2 存在一個父代對偶基因<19,20>，使得物種 1-4 內在 locus 2 會有一個基因屬於 19 或 20。也就是<35,7>及<19,20>是在群 1 的手足中可能的父代的 locus1 及 locus2，因為是同父異母或同母異父，所以只有繼承雙親其中一個的對偶基因。但是物種 5 就無法加入，因為加入後物種 5 的 locus 1 會違反 Half-Sibs Property。Note: 加入物種 5 時 locus 2 不會違反 Half-Sibs Property。

表 1. Input instance for the HSRP

物種	Locus 1 <allelel1,allele2>	Locus 2 <allelel1,allele2>
1	<7,8>	<19,20>
2	<7,10>	<20,46>
3	<35,36>	<19,23>
4	<34,35>	<15,19>
5	<32,10>	<15,19>

表 2. The optimal solution for the HSRP

群 1: S_1	1,2,3,4
群 2: S_2	5

本論文第二章我們將描述我們的證明在 ℓ=1 時，HSRP 等價於節點涵蓋問題。之後，我們給一個結論跟未來方向。

2 半手足關係重建問題

本章中，我們證明 HSRP 等價於節點涵蓋問題(Vertex Cover Problem)當 ℓ=1。因此我們可以透過節點涵蓋問題最好的近似演算法[9]，我們可以得到一個 *2-(2lnln|V|/ln|V|)*倍的近似解在 HSRP 當 ℓ=1。

Half Sibling Relationship Reconstruction Problem (HSRP) [11,12]

Instance: n 個物種，每個物種 i 有 ℓ 個 locus，每個 locus 有兩個對偶基因以$<a_{ij},b_{ij}>$表示，$1\le j\le \ell$，a_{ij} 及 b_{ij} 為物種 i 在第 j 個 locus 的 2 個對偶基因。

Problem: n 個物種進行分群使得分群數最少並滿足每個群必需服從 Half-Sibs Property。Half-Sibs Property 即把物種分成 s 個群組 $S_1,S_2,\ldots,S_s$ 使得每個 S_k 內的每個物種 i，$i\in S_k$，其每個 locus j，$1\le j\le \ell$，都存在一個父代對偶基因 $P_j=<x_j,y_j>$，使得 $a_{ij}\in P_j$ or $b_{ij}\in P_j$，$1\le k\le s$。

Vertex Cover Problem (VCP) [8]

Instance:一個無向完全圖 $G=(V,E)$

Problem: 一個節點集合 C (Vertex cover)使得在 E 內的每個邊都可以在 C 內找到一個節點與之緊鄰(*incident*)且 C 的 cardinality 要最小。

定理 1: 半手足關係重建問題等價於節點涵蓋問題，當每個物種只有 1 個 locus 時。

證明: 給定一個 HSRP 的 Instance: n 個物種，每個物種 i 有 1 個對偶基因以$<a_i,b_i>$表示。我們現在轉換到 VCP 的 instance: $G=(V,E)$如下:

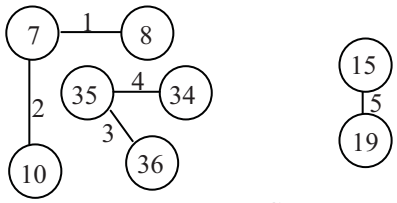
$V=\{v_{ai}\cup v_{bi}, 1\le i\le n\}$。即對每個對偶基因 $<a_i,b_i>$，我們造出對應的節點 v_{ai} 及 v_{bi}。

$E=\{(v_{ai},v_{bi})$, if $<a_i,b_i>$是物種 i 的對偶基因，$1\le i\le n\}$。即對每個物種 i，我們將其所對應的節點 v_{ai}，v_{bi} 透過一條邊連結。

轉換後，當在圖 G 中找到一個 vertex cover C 時，每次都在 C 內的任選兩個節點 v_x,v_y，並將任何緊鄰 v_x 或 v_y 的邊其對應的物種放到同一群內，並用 P=<x,y>代表此群的父代對偶基因，x 和 y 為節點 v_x 和 v_y 所對應的對偶基因，直到所有在 C 內的節點都被選過一次為止。因此對於 HSRP 我們會分成$\lceil |C|/2 \rceil$個群組。很顯然的，根據上述的做法我們找到的群組會滿足 Half-Sibs Property，因為每個群組內的物種 i 的對偶基因要不就是從 $a_i \in P$ 要不就是 $b_i \in P$ 且這些群組會是 HSRP 的最佳解，否則 C 就不會是 VCP 的最佳解。我們舉例說明轉換的過程。表 3 為 HSRP 的一個 instance。圖 1 為在表 3 的 instance 下所對應到 VCP 的一個 instance。{7,35,15}為圖 G 的最佳解針對 VCP。在 HSRP 的最佳解我們會把 S_1={1,2,3,4}分成一群，S_2={5}為另外一群 S_1 的父代對偶基因為<7,35>，S_2 的父代對偶基因為<15,any>，any 為任何的基因都滿足。

表 3. Input instance for the HSRP

物種	locus <allelel1,allele2>
1	<7,8>
2	<7,10>
3	<35,36>
4	<34,35>
5	<15,19>



G

圖 1. 轉換表 3 的 instance for the HSRP 到 instance $G=(V,E)$ for the VCP。

因此我們可以將 HSRP 的任何 instance 轉換到 VCP 的 instance，透過 VCP 的目前最好的近似演算法，我們會得到一個 HSRP 的近似演算法滿足相同倍率，目前 VCP 最好的近似演算法倍率是 $2-(2\ln\ln|V|/\ln|V|)$ [9]。反過來，我們可以將 VCP 的 instace $G=(V,E)$ 轉換到 HSRP 的 instance。我們將在 E 內的每個邊對應到每個物種。每個邊所連接的兩個節點對應到該物種的對偶基因，例如從圖 1 轉換到表 3。因此當我們找到 HSRP 的最佳解時，VCP 的最佳解也可以容易的得到。因此我們知道半手足關係重建問題等價於節點涵蓋問題，當每個物種只有 1 個 locus。 ❐

4. 未來研究方向

本研究探討半手足關係重建問題(HSRP)在無父代的資訊下，我們證明半手足關係重建問題等價於節點涵蓋問題當每個物種只有1個 locus時。然而此方法在locus個數超過1個時，會無法轉換成功。因此未來我們希望能設計一個更好的近似演算法來解決半手足關係重建問題在 locus的個數超過1個時候。

Acknowledgements

This work was supported in part by the Ministry of Science and Technology of the Republic of China under Contract MOST 103-2221-E-845-001 。 Corresponding author: Yen-Hung Chen。

Reference

[1] A., Almudevar and C. Field. Estimation of single generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 4, pp. 136–165, 1999.

[2] M.V. Ashley, T.Y. Berger-Wolf, P. Berman, W. Chaovalitwongse, B. DasGupta, and M.Y. Kao. On Approximating four covering and packing problems. *Journal of Computer and System Sciences*, Vol. 75, pp. 287–302, 2009.

[3] T.Y. Berger-Wolf, B. DasGupta, W. Chaovalitwongse, and M.V. Ashley. Combinatorial reconstruction of sibling relationships. *In Proceedings of the 6th International Symposium on Computational Biology and Genome Informatics*, pp. 1252–1255, 2005.

[4] T.Y. Berger-Wolf, S.I. Sheikh, B. DasGupta, M.V. Ashley, I.C. Caballero, W. Chaovalitwongse, and S.L. Putrevu. Reconstructing sibling relationships in wild populations. *Bioinformatics*, Vol. 23, pp. i49–i56, 2007.

[5] J. Beyer and B. May. A graph-theoretic approach to the partition of individuals into full-sib families. *Molecular Ecology*, Vol. 12, pp. 2243–2250, 2003.

[6] M.S. Blouin. DNA-based methods for pedigree reconstruction and kinship analysis

in natural populations. *TRENDS in Ecology and Evolution*, Vol 18, pp. 503–511, 2003.

[7] W. Chaovalitwongse, T.Y. Berger-Wolf, B. DasGupta, and M.V. Ashley. Set covering approach for reconstruction of sibling relationships. *Optimization Methods and Software*, Vol 22, pp. 11–24, 2007.

[8] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction to Algorithm* (3rd edition). Cambridge: MIT Press, 2001.

[9] E. Halperin. Improved approximation algorithms for the vertex cover problem in graphs and hypergraphs. *SIAM Journal on Computing*, Vol 31, pp. 1608–1623, 2000.

[10] A.G. Jones and W.R. Ardren. Methods of parentage analysis in natural populations. *Molecular Ecology*, Vol 12, pp. 2511–2523, 2003.

[11] S.I. Sheikh, T.Y. Berger-Wolf, A. Khokhar, I.C. Caballero, M.V. Ashley, W. Chaovalitwongse, C.A. Chou, and B. DasGupta. Combinatorial reconstruction of half-sibling groups. *In Proceedings of the 8th International Conference on Computational Systems Bioinformatics*, pp. 59–67, 2009

[12] S.I. Sheikh, T.Y. Berger-Wolf, A. Khokhar, I.C. Caballero, M.V. Ashley, W. Chaovalitwongse, and B. DasGupta. Combinatorial reconstruction of half-sibling groups from microsatellite data. *Journal of Bioinformatics and Computational Biology*, Vol 8, pp. 337–356, 2010.

[13] S.C. Thomas and W.G. Hill. Sibship reconstruction in hierarchical population structures using markova chain monte carlo techniques. *Genetical Research*, Vol 79, pp. 227–234, 2002.

[14] J. Wang. Sibship reconstruction from genetic data with typing errors. *Genetics*, Vol 166, pp. 1963–1979, 2004.