3H1-01

クラスターBPによる二点連鎖解析

Two-point Linkage Analysis by Cluster BP

泉 祐介　　佐藤 泰介
Yusuke Izumi　Taisuke Sato

東京工業大学大学院／CREST
Graduate School, Tokyo Institute of Technology / CREST

In recent years, Bayesian networks have been applied to various kinds of probablistic tasks. For example, Superlink, a linkage analysis program, is based on a model represented by a Bayesian network and performs exact computation very quickly. In this paper, we perform approximate computation by cluster BP for two-point linkage analysis. We empirically show our method provides approximations good enough for practical use, as well as it might be useful for analysis with huge pedigree and/or many loci.

1. はじめに

不確実性のある知識の表現法として提案されたベイジアンネット [Jensen 01] が，近年ではさまざまな分野に応用されている．本論文で扱う連鎖解析もその 1 つであり，ベイジアンネットによるモデルに基づく連鎖解析ソフトウェアとして Superlink が開発されている [Fishelson 02]．この Superlink は従来のソフトウェアに比べて驚異的な高速化を実現し，これまで不可能であった大型の家系に対する解析も可能にした．

ところで，統計力学の観点から，ベイジアンネットを含むモデルクラスに対する周辺分布の近似計算法としてクラスター BP（belief propagation）[Sato 03] が提案されている．クラスター BP は，ある条件下で loopy BP と同様にループのあるベイジアンネットに対する単純で効率のよい近似計算法であるが，loopy BP が与える Bethe 近似よりもよい Kikuchi 近似を与える．

ここでは，疾患遺伝子の二点連鎖解析に対してクラスター BP を適用する．この適用によってクラスター BP の有用性を確かめるとともに，連鎖解析における近似計算の適用可能性を検証する．

2. 連鎖解析 [Strachan 97]

人間の遺伝物質は 23 対の**染色体**からなり*1，各個体は各対の一方を父親から，他方を母親から受け継ぐ．各染色体には**遺伝子座**または**座位**と呼ばれる部分があり，そこに遺伝子が存在する．遺伝子座ごとに存在する可能性のある遺伝子は何種類かあり，それらの各々を**対立遺伝子**という．ある遺伝子座に存在する 2 つの対立遺伝子の組を**遺伝子型**と呼ぶ．ただし我々が観測するのは病気の有無，血液型といった**表現型**である．表現型は遺伝子型により決まるが一意とは限らない．ある遺伝子型をもつ個体がある表現型をとる確率を**浸透率**と呼ぶ．

メンデルの**分離の法則**によれば，各個体は各座位ごとに 2 つある対立遺伝子の一方を等確率で選択して子に伝達する．子が複数いる場合，この伝達操作は各々の子に対して独立に行われる．また，**独立の法則**によれば，異なる座位にある対立遺伝子の分配は互いに独立に行われる．ただし，同じ染色体上にある座位間ではこの法則が成り立たない．すなわち，同じ染色体上にある 2 つの対立遺伝子は物理的に結合しているため，そのまま子に伝達される可能性が高い．これが**連鎖**である．一方，同じ染色体に存在する対立遺伝子が必ずしも結合したまま子に伝達されるとも限らない．減数分裂時に染色体が交差し，結合している対立遺伝子の組み合わせが変化する**組み換え**という現象があるためである．

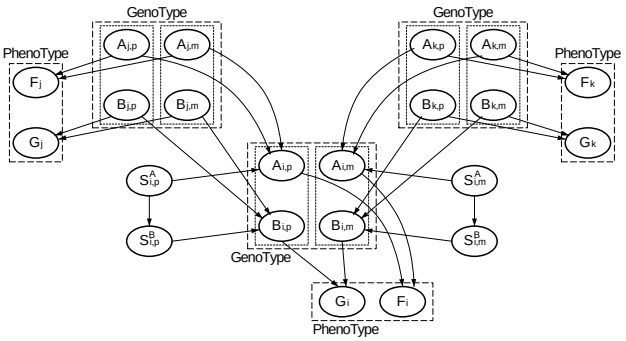

図 1: 親子 3 個体に対するベイジアンネット

一般に 2 つの座位が離れているほど，その座位間で組み換えが生じやすくなる．ある 2 座位間で組み換えが起こる確率をその座位間における**組み換え価**（θ で表す）といい，通常は $0 \leq \theta \leq 0.5$ の値をとる．$\theta = 0.5$ のときはその座位間では連鎖が生じない（自由組み換え）．組み換え価と距離との間には密接な関係があり，組み換え価を推定すればその座位間のおよその距離が得られる*2．

本論文では疾患遺伝子の二点連鎖解析を扱う．すなわち，すでに位置がわかっている座位（**マーカー座位**と呼ばれる）と**疾患座位**との組み換え価を推定する．得られた値から距離を計算すれば目的とする疾患座位の位置を推定できる．

2.1 ベイジアンネットによるモデル化

本論文では [Friedman 00] のモデルを用いる．図 1 は，親子 3 個体の二点連鎖解析をモデル化したものである．

[Friedman 00] では遺伝子座が $A, B, \ldots$，表現型が $F, G, \ldots$ で表されているので，それにならうこととした．図中の $A_{i,p}$，

連絡先: 泉 祐介，東京工業大学大学院 情報理工学研究科 計算工学専攻 佐藤研究室，〒152-8550 東京都目黒区大岡山 2-12-1，03-5734-2186，yuizumi@mi.cs.titech.ac.jp

*1 染色体には常染色体と性染色体があるが，本論文では前者のみを扱う．ただし多くの内容は後者の性染色体にもあてはまる．

*2 遺伝子座の正確な位置を得るには遺伝子を完全に解析するしかないが，これには膨大な人手と費用を要する．そのため，事前におよその位置を推定しておくことは重要である．

$A_{i,m}$ は個体 i の座位 A における父親由来，母親由来の対立遺伝子であり，F_i はこれに対応する表現型である．同様に $B_{i,p}$，$B_{i,m}$，G_i は座位 B における対立遺伝子と表現型を表す．$S^A_{i,p}$ のような変数は，父親（または母親）がもつ 2 つの対立遺伝子のうちどちらを継承するかを決める 2 値の確率変数で**セレクター**と呼ばれる．この変数が 0 ならば父親由来の，1 ならば母親由来の遺伝子を継承する．形式的な定義は次により与えられる（他のセレクターについても同様）．

$$A_{i,p} = \begin{cases} A_{j,p} & \text{if } S^A_{i,p} = 0 \\ A_{j,m} & \text{if } S^A_{i,p} = 1 \end{cases} \quad (j \text{ は } i \text{ の父親})$$

なお本論文では A，B を順にマーカー座位，疾患座位とする．マーカー座位 A に対する表現型 F は順序なし遺伝子型（父親由来と母親由来の遺伝子を区別しない遺伝子型）とする．疾患座位 B における対立遺伝子は疾患遺伝子 D と正常遺伝子 d を仮定する．この座位の表現型 G は病気の有無である．

3. クラスター BP

クラスター BP[Sato 03] は，n 個の離散的な確率変数の組 $X = (X_1, \ldots, X_n)$ の同時確率分布が

$$p(x) = \prod_{\alpha \in \mathbf{P}} \psi_\alpha(x_\alpha) \tag{1}$$

のようにポテンシャル関数 $\psi_\alpha(x_\alpha)$ の積で表されるモデルにおける周辺分布を，**tree 条件を満たすクラスターグラフ**上でのメッセージ交換により計算する手法である[*3]．この手法は統計力学における変分自由エネルギーの Kikuchi 近似を与える式 (1) において α は $(1, \ldots, n)$ の部分ベクトル，$\mathbf{P}$ はそのような部分ベクトルの集合である．x_α は $(x_1, \ldots, x_n)$ の α に対応する部分ベクトルを表す[*4]．便宜上ベクトルと集合は同一視し[*5]，$\{1, \ldots, n\}$ の部分集合を**クラスター**，$\mathbf{P}$ の元を**ポテンシャルクラスター**と呼ぶ．記法の簡便のためクラスター $\{2, 4\}$ を単に 24 のように記す．

いま $\mathbf{P}$ とは異なるクラスターの集合 $\mathbf{B}$ を考え，

$$\begin{cases} \forall \alpha \in \mathbf{P},\ \exists \alpha' \in \mathbf{B} \text{ s.t. } \alpha \subseteq \alpha' \\ \forall \alpha_1, \alpha_2 \in \mathbf{B},\ \alpha_1 \not\subset \alpha_2 \text{ and } \alpha_2 \not\subset \alpha_1 \end{cases}$$

を満たすものとする．このとき $\mathbf{B}$ の元を $\mathbf{P}$ **に適合する基本クラスター**と呼ぶ．さらに，この $\mathbf{B}$ に対し

$$\begin{cases} \mathbf{M} \equiv \{\beta \mid \beta = \alpha_1 \cap \cdots \cap \alpha_k (\neq \emptyset),\ k \geq 2,\ \alpha_i \in \mathbf{B}\} \\ \mathbf{U} \equiv \mathbf{M} \cup \mathbf{B} \end{cases}$$

とおく．$\mathbf{B}$ **に基づくクラスターグラフ** $\mathcal{G}_B$ は，$\mathbf{U}$ の元でラベル付けされた無向グラフで次の条件を満たすものをいう．

- $\mathbf{B}$ の各元はただ 1 回だけ頂点のラベルとして出現．
- 頂点 v_1，v_2 を結ぶ辺があれば，その辺は $\underline{v_1} \cap \underline{v_2}$ でラベル付けされている．

*3 ベイジアンネットでは条件付き確率 $p(x_i|\pi_i)$ がポテンシャル関数 $\psi_\alpha(x_\alpha)$ になる．

*4 $x = (x_1, x_2, x_3, x_4)$ に対し $\alpha = (2, 4)$ ならば $x_\alpha = (x_2, x_4)$.

*5 α の要素は順序づけられているので問題ない．

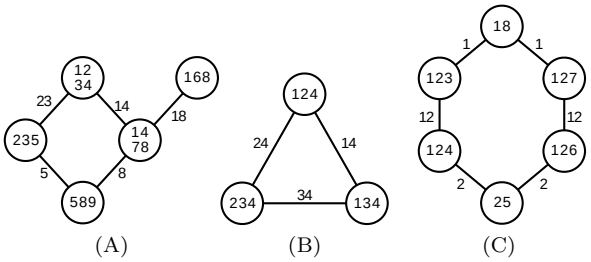

図 2: クラスターグラフの例

ここで $\underline{v}$ は v をラベル付けするクラスターを表す（以降も同様）．特に任意の $\alpha \in \mathbf{U}$ について，α を含むラベルでラベル付けされた頂点，辺からなる $\mathcal{G}_B$ の部分グラフが木をなすとき，かつそのときに限り $\mathcal{G}_B$ は **tree 条件を満たす**という．

図 2 はクラスターグラフの例である．(A) は tree 条件も満たす．一方，(B) は 4 を含む頂点と辺がループをなすため，また (C) は 12 を含む頂点と辺が非連結であるため，いずれも tree 条件は満たさない．

クラスター $\alpha \in \mathbf{U}$ のポテンシャルは

$$\phi_\alpha(x_\alpha) \equiv \prod_{\beta \in \mathbf{P},\, \beta \subseteq \alpha} \psi_\beta(x_\beta)$$

により与えられる．ただし計算に先立ち，辺に付随するポテンシャルを次の手順により頂点に付け換える．

1. 各辺 e について e につながる頂点の一方を e のポテンシャル $\phi_{\underline{e}}(x_{\underline{e}})$ の付け換え先 $\theta(e)$ に決める．
2. 頂点 v の**修正されたポテンシャル** $\phi'_v(x_{\underline{v}})$ を計算する．

$$\phi'_v(x_{\underline{v}}) = \frac{\phi_{\underline{v}}(x_{\underline{v}})}{\prod_{e:\theta(e)=v} \phi_{\underline{e}}(x_{\underline{e}})}$$

クラスター BP の計算式は次により与えられる．

$$m_{ij}(x_{\underline{i} \cap \underline{j}}) = \kappa \sum_{x_{\underline{i} \cap \underline{j}}} \phi'_i(x_{\underline{i}}) \prod_{k \in N(i) \setminus \{j\}} m_{ki}(x_{\underline{k} \cap \underline{i}}) \tag{2}$$

$$b_i(x_{\underline{i}}) = \kappa' \phi'_i(x_{\underline{i}}) \prod_{k \in N(i)} m_{ki}(x_{\underline{k} \cap \underline{i}}) \tag{3}$$

$m_{ij}(x_{\underline{i} \cap \underline{j}})$ が $\mathcal{G}_B$ の頂点 i から頂点 j に向かう多変数メッセージとなる．$N(i)$ は頂点 i に隣接する頂点集合を表す．実際の計算では式 (2) を満たす $\{m_{ij}(x_{\underline{i} \cap \underline{j}})\}$ を反復法により計算し，式 (3) から周辺分布 $p_{\underline{i}}(x_{\underline{i}})$ の近似 $b_i(x_{\underline{i}})$ を求める．

4. クラスター BP による二点連鎖解析

本節では Kamatani et al. による FJHN [*6]の家系図 [Kamatani 00] を対象として，クラスター BP による二点連鎖解析を行う．家系図中の個体数は 65 であり，そのうち 32 個体に対してはマーカー対立遺伝子の対が与えられている．なおマーカー座位は 9 座位あり，ここではその 9 座位の各々について疾患座位との組み換え価 θ を推定する．推定は θ を細かく刻み，それに対する尤度 $L(\theta)$ を計算することにより行う．実

*6 Familial Juvenile Hyperuricemic Nephropathy（家族性若年性高尿酸血症性腎症）の略．

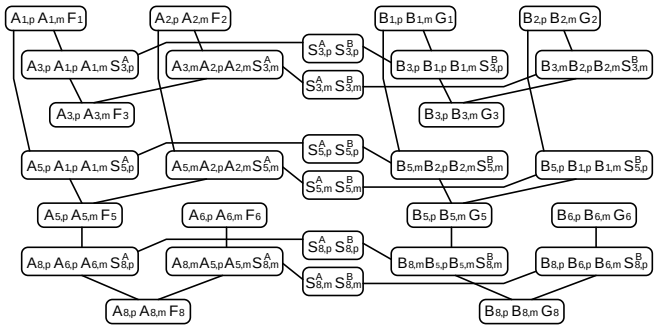

図 3: 方法 A におけるクラスターグラフの一部

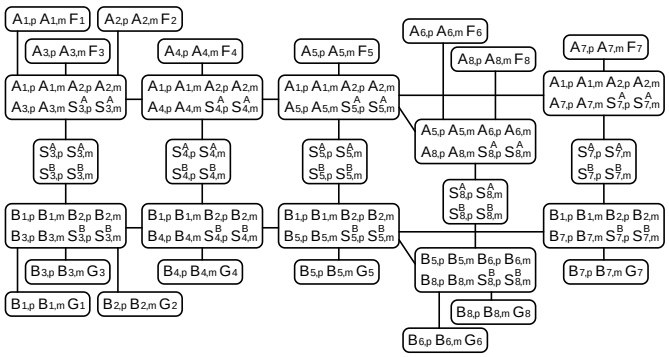

図 4: 方法 B におけるクラスターグラフの一部

際の推定にあたっては **LOD スコア**と呼ばれる自由組み換えとの尤度比の対数 $\log_{10} L(\theta)/L(0.5)$ を利用する．

計算は 2.1 節のモデルに基づく．クラスターグラフの定義方法による近似の精度，計算時間の差異を検討するため，2 通りのグラフを定義してそれぞれ尤度を計算し，それらの計算結果を Superlink による厳密計算の結果と比較する．

対立遺伝子の事前確率と浸透率については次を仮定する．ただし $\mathcal{F}$ は創始者（親のいない個体）を表す．また N はマーカー対立遺伝子の種類数である．

$$p(a_{i,p}) = p(a_{i,m}) = 1/N \quad (i \in \mathcal{F})$$

$$p(B_{i,p} = D) = p(B_{i,m} = D) = 0.01 \quad (i \in \mathcal{F})$$

$$p(G_i = \text{affected} \mid b_{i,p}, b_{i,m}) = \begin{cases} 0 & \text{if } b_{i,p} = b_{i,m} = d \\ 1 & \text{otherwise} \end{cases}$$

4.1 方法 A

親の存在する確率変数とその親変数からなるクラスターを基本クラスターとする[*7]．例えば，

$$\begin{aligned} \alpha^A_{i,p} &\equiv \{A_{i,p}, A_{j,p}, A_{j,m}, S^A_{i,p}\} \quad (i \notin \mathcal{F},\ j \text{ は } i \text{ の父親}) \\ \alpha^F_i &\equiv \{A_{i,p}, A_{i,m}, F_i\} \\ \alpha^S_{i,p} &\equiv \{S^A_{i,p}, S^B_{i,p}\} \quad (i \notin \mathcal{F}) \end{aligned}$$

$\alpha^A_{i,m}$，$\alpha^B_{i,p}$，$\alpha^B_{i,m}$，α^G_i，$\alpha^S_{i,m}$ も同様に定義できる．クラスターグラフの頂点は基本クラスターのみとし，α^F_i，α^F_j，$\alpha^S_{i,p}$ をそれぞれ $\alpha^A_{i,p}$ と，α^G_i，α^G_j，$\alpha^S_{i,p}$ をそれぞれ $\alpha^B_{i,p}$ と辺で結ぶ．α^F_i，α^F_k，$\alpha^S_{i,m}$ と $\alpha^A_{i,m}$，および α^G_i，α^G_k，$\alpha^S_{i,m}$ と $\alpha^B_{i,m}$ についても同様の辺を結ぶ．グラフの一部を図 3 に示す．

4.2 方法 B

$\alpha^A_i \equiv \alpha^A_{i,p} \cup \alpha^A_{i,m}$，$\alpha^B_i \equiv \alpha^B_{i,p} \cup \alpha^B_{i,m}$，$\alpha^S_i \equiv \alpha^S_{i,p} \cup \alpha^S_{i,m}$ と α^F_i，α^G_i を基本クラスターとする．方法 A に比べてクラスターが大きくなり，特に α^A_i と α^B_i は 8 変数からなる．クラスターグラフの頂点はやはり基本クラスターのみとする．辺については次の方針に従う．グラフの一部を図 4 に示す．

- α^A_i，α^B_i をそれぞれ α^S_i と結ぶ．
- 兄弟 $i_1, \ldots, i_n$ について $\alpha^A_{i_1}, \ldots, \alpha^A_{i_n}$，$\alpha^B_{i_1}, \ldots, \alpha^B_{i_n}$ をそれぞれ順に結ぶ．また共通の父親 j が非創始者のときは α^A_j と $\alpha^A_{i_1}$，α^B_j と $\alpha^B_{i_1}$ も結ぶ．母親についても同様．
- $i \notin \mathcal{F}$ のときは α^F_i と α^A_i を，$i \in \mathcal{F}$ のときは α^F_i と α^A_h（h は個体 i の第 1 子）を結ぶ．α^G_i についても同様．

*7 言い換えれば $\mathbf{B} = \mathbf{P}$．ただし，基本クラスターは互いに含まれないという条件があるので，親変数のない確率変数のみからなるクラスターは基本クラスターとしない．

座位	方法 A	方法 B
1	18 秒	41 分 02 秒
2	——	14 分 39 秒
3	6 秒	2 分 45 秒
4	7 秒	3 分 33 秒
5	49 秒	4 時間 21 分 08 秒
6	14 秒	14 分 58 秒
7	16 秒	14 分 57 秒
8	1 分 58 秒	4 時間 21 分 53 秒
9	39 秒	41 分 41 秒

表 1: 計算時間（Pentium4 3.4GHz にて測定）

4.3 計算結果

上で定義したクラスターグラフに基づいて計算した結果の一部を Superlink による厳密計算の結果とともに図 5 に示す．図中のグラフの横軸と縦軸には，それぞれ組み換え価 θ および得られた尤度から計算した LOD スコアをとっている．

方法 A では座位によって精度にいくらかのばらつきがあったが，どの座位においても $\theta = 0.00$ のときの尤度は厳密計算の結果と一致し，θ が大きくなるほど誤差も大きくなる傾向が見られた[*8]．一方，方法 B では厳密計算の結果とほぼ一致した．

計算時間については表 1 のとおりである．表からもわかるように方法 A は比較的短い時間で計算が終わるが，方法 B ではかなり膨大な計算時間を要した．ただし，表には載せていないが，Superlink の計算時間は方法 A と比べても格段に短い．

このほか方法 A では，一部の座位において特定の組み換え価に対する尤度を正常に計算できない現象が見られた．

5. 考察

計算精度については，ある程度の誤差は生じたが，もともと連鎖解析が「およその」位置を推定する手法であることを考えれば，実用的には十分な精度が得られたと言える．

ところで，ここでは二点連鎖解析を扱ったが，提案手法を多点連鎖解析に適用することも可能である．Friedman のモデルではセレクターに対して HMM を仮定しており，例えばセレクター S^C は S^B のみに依存する．したがって，方法 A を三点に拡張するとセレクターに関しては $\{S^B_{i,p}, S^C_{i,p}\}$，$\{S^B_{i,m}, S^C_{i,m}\}$ なる基本クラスターが増える．辺については，直感的には図 3 のグラフにおいて遺伝子座 A と B が $\{S^A_{i,p}, S^B_{i,p}\}$ や $\{S^A_{i,m}, S^B_{i,m}\}$ をはさんで接続されているように，遺伝子座 B と C が $\{S^B_{i,p}, S^C_{i,p}\}$ や $\{S^B_{i,m}, S^C_{i,m}\}$ をはさんで接続される．

*8 図中のグラフでは $\theta = 0.50$ のところで重なっているが，これは LOD スコアが $\theta = 0.50$ の尤度との比を表す値であるため．

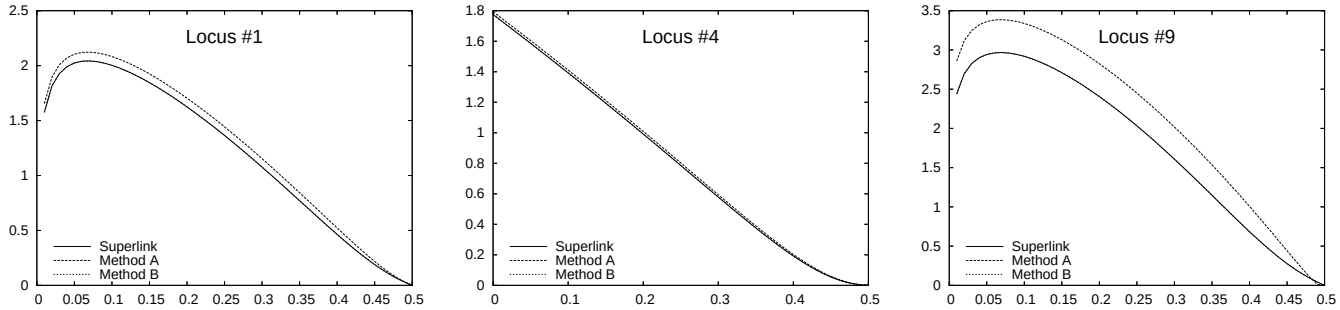

図 5: 計算結果（Superlink と方法 B のグラフは重なっている）

Friedman のモデルにおいて，確率変数の数は個体数，座位数に対して線形である．方法 A において，頂点の数は 4.1 節の定義にしたがえば確率変数の数を超えないので，頂点の数は個体数や座位数に対して線形オーダーになる．また，上で述べた方法にしたがって拡張した場合，5 個以上の変数を含む基本クラスターは現れることがないため，基本クラスターの大きさの最大値は一定である．各頂点のクラスターに含まれる変数の数がより多くなることはない．さらに提案手法では辺の数は頂点数に対して線形になる．したがって式 (2) の 1 回の反復における計算量は個体数や座位数に対して線形であると考えてよい．

ただし，個体数や座位数が増加すれば収束までの反復回数は増加すると思われる．よって全体の計算量が個体数や座位数に対して線形になるとは限らないが，実際の応用において計算精度があまり問題にならないときは一定回数で反復を終了させる場合もあるので，連鎖解析でもこれが可能であれば全体の計算量が線形になる．そのため，特に多くの座位を対象に解析する場合は，近似計算による解析が Superlink のような厳密計算による解析よりも有利になる可能性がある．

6. 今後の課題

現在はクラスターグラフを生成する部分とクラスター BP の計算を行う部分が別個のプログラムになっており，計算途中にパイプを通じた入出力が存在する．これが原因で計算時間が余計にかかっている面もあるため，まずはプログラムを一本化して計算時間を短縮すべきである．

前節で述べたように，提案手法の多点連鎖解析への適用は重要な課題である．また，方法 A では一部の値が正常に計算できない問題が生じたので，この問題に対する何らかの対処を施す必要があるだろう．

謝辞

本研究を遂行するにあたり，亀谷由隆氏には多くの場面で助言，支援をいただいた．ここに感謝の意を記す．

参考文献

[Fishelson 02] Fishelson, M. and Geiger, D.: Exact genetic linkage computations for general pedigrees, *Bioinfomatics*, Vol. 18, No. Suppl. 1, pp. S189–S198 (2002)

[Friedman 00] Friedman, N., Geiger, D., and Lotner, N.: Likelihood Computations Using Value Abstraction, in *Proceedings of Sixteenth Conference on Uncertainty in Artificial Intelligence* (2000)

[泉 03] 泉 祐介, 佐藤 泰介：クラスター BP による遺伝連鎖解析に向けて, 2003 年ベイジアンネットセミナー予稿集 (2003)

[Jensen 01] Jensen, F. V.: *Bayesian Networks and Decision Graphs*, Springer-Verlag (2001)

[Kamatani 00] Kamatani, N., et al.: Localization of a gene for familial juvenile hyperuricemic nephropathy causing underexcretion-type gout to 16p12 by genomi-wide linkage analysis of a large family., *Arthritis and Rheumatism*, Vol. 43, No. 4, pp. 925–929 (2000)

[Sato 03] Sato, T.: Cluster BP and cluster CCCP: Two simple methods for computing Kikuchi approximations, Technical Report TR03-0001, Dept. of Computer Science, Tokyo Institute of Technology (2003)

[Strachan 97] Strachan, T. and Read, A. P.: *Human Molecular Genetics*, BIOS Scientific Publishers (1997),（邦訳：村松正實，笹月健彦，木南凌，辻省次監訳，ヒトの分子遺伝学， メディカル・サイエンス・インターナショナル）