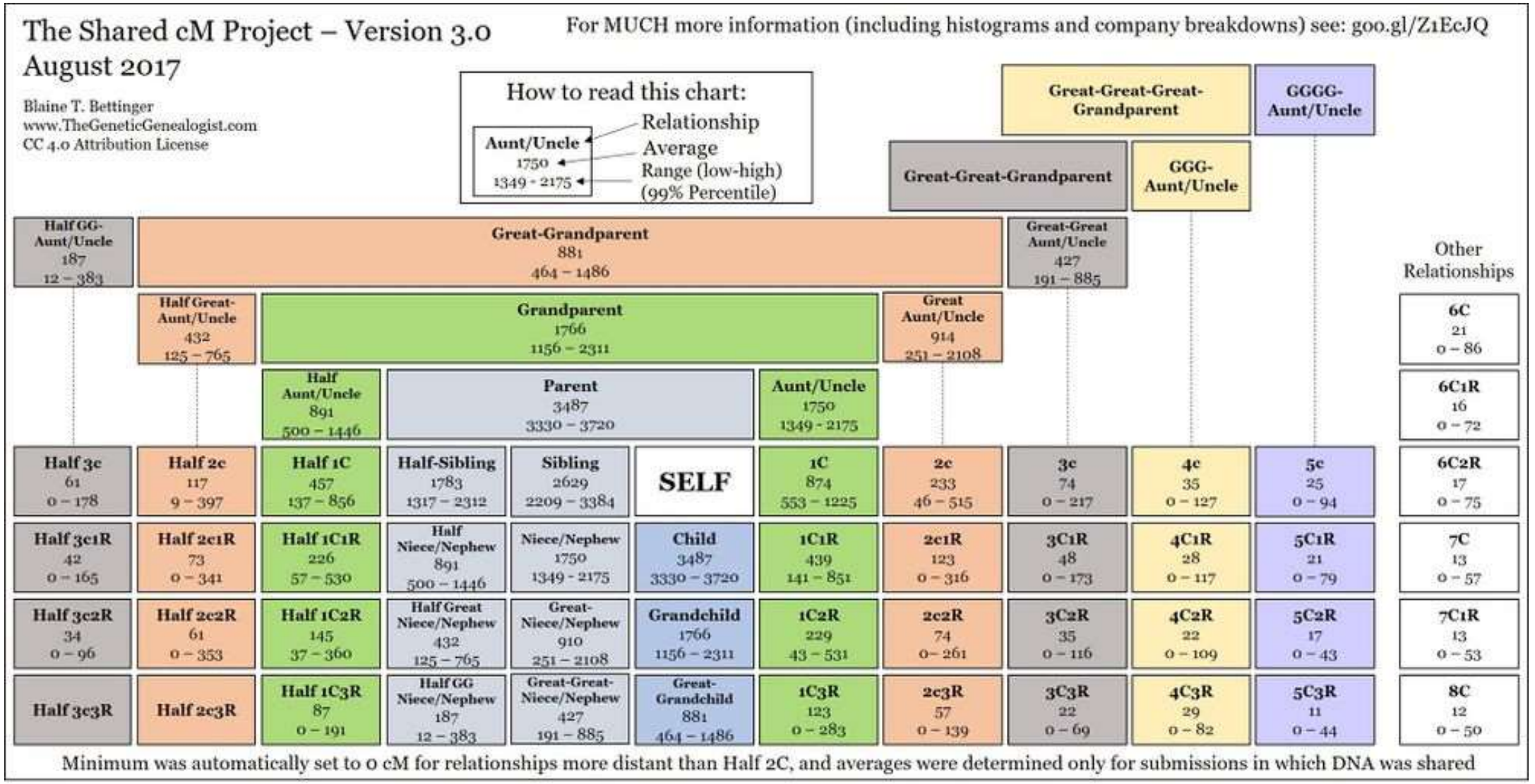
距離行列 類似度行列 固有値分解 グラフ

法数学勉強会
2019/01/26
京都大学(医)統計遺伝学分野
山田　亮

前回の話

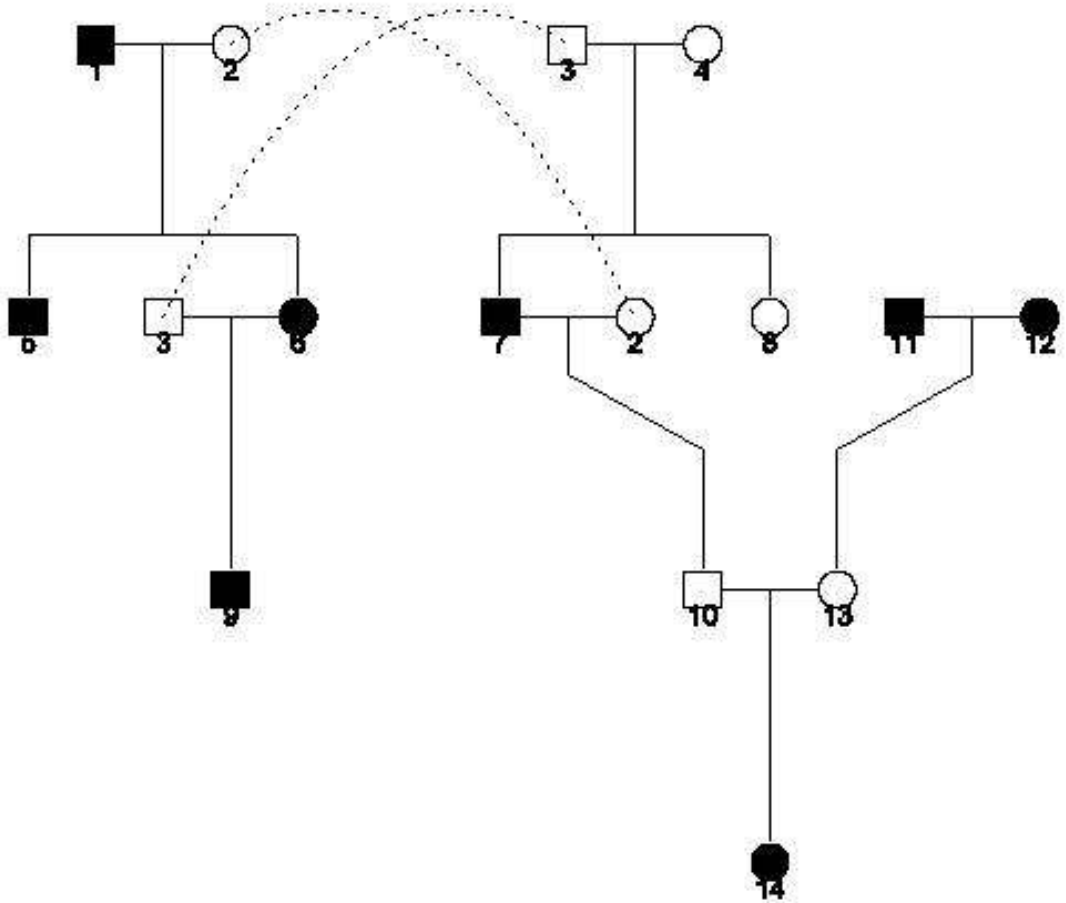
- ゴールデン・ステート・キラー
- DNA多型情報を用いて
- ２人の間の血縁関係を推定し
- (犯人を特定した)
- 「家系図を復元したか？」、「単に登録して2名の関係の推定の列挙をしただけか？」
  - →　多分、後者…

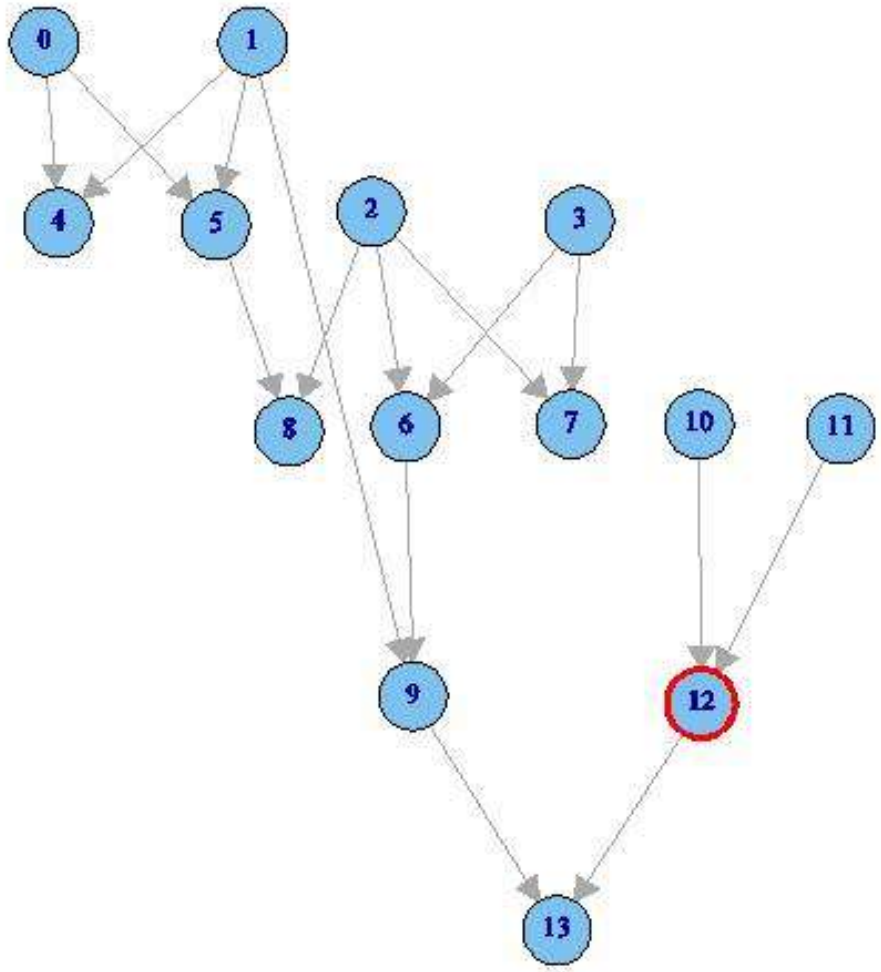
Autosomal DNA match



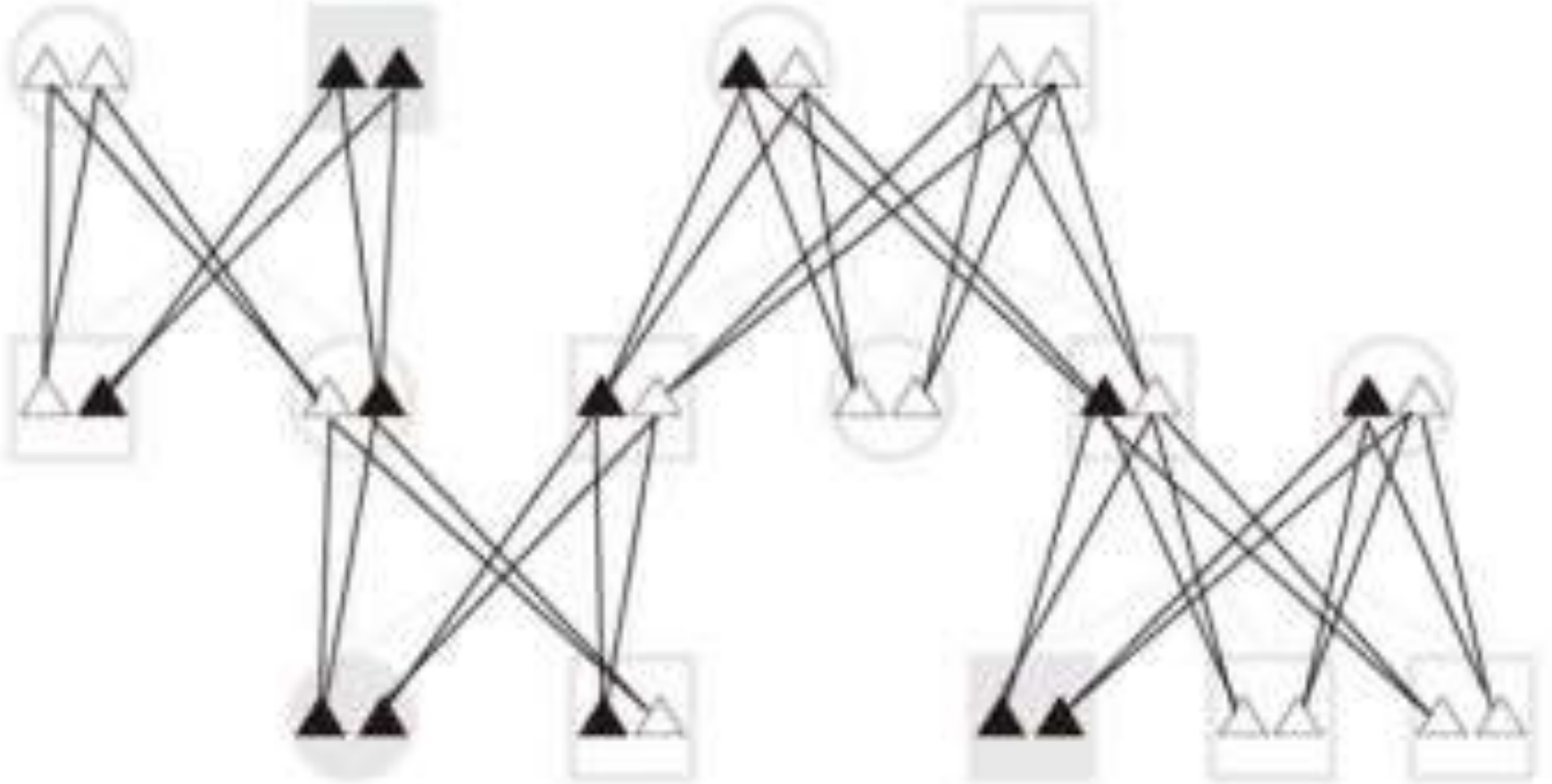
N人の関係を、2人ペアの関係の集まりとして見る話し

家系図





親子は1/2の関係
子によって「かすがい」された「夫婦」は、1/4の関係なのか？



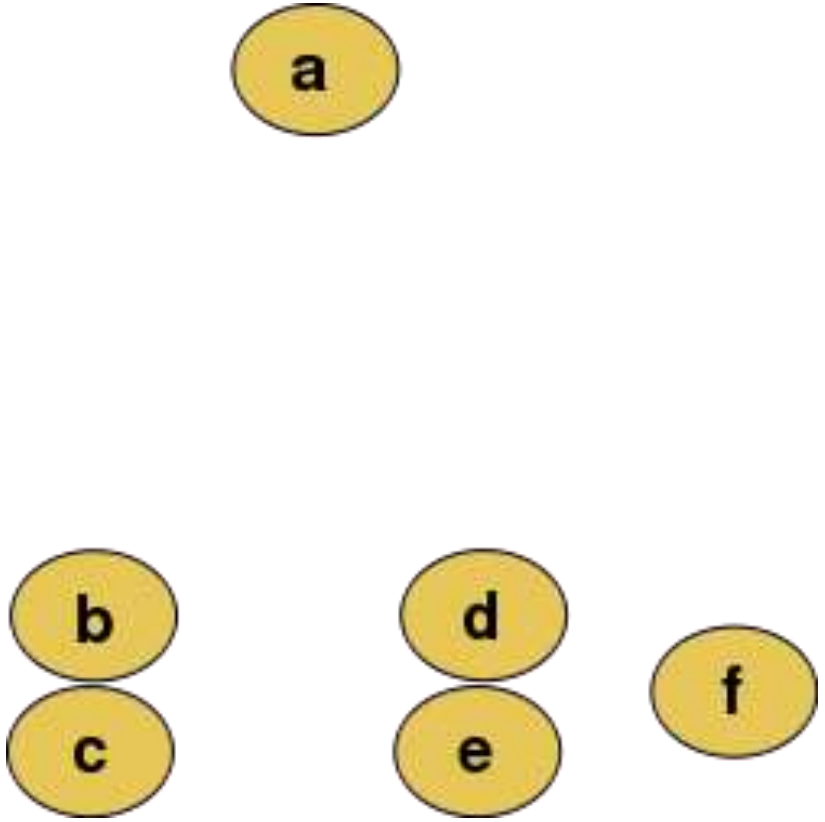
つながっているのは「染色体」
夫婦(の染色体)は子を介してつながっていない
孫(の染色体)を介してつながっている

たくさんの対象が作り上げる世界

- 二項定理
- 対象がN個あるとき
  - N個、それぞれについての情報：$N = {}_NC_1$
  - ペアに関する情報：${}_NC_2$
  - トリオに関する情報：${}_NC_3$
  - 4つ組に関する情報：${}_NC_4$
  - …
  - N個すべての組に特有の情報：${}_NC_N$
  - ${}_NC_1 + {}_NC_2 + \cdots + {}_NC_N = 2^N - 1$
  - $(1+x)^N = {}_NC_0\ x^0 + {}_NC_1\ x^1 + \cdots + {}_NC_N\ x^N$
    - $x = 1$のとき…

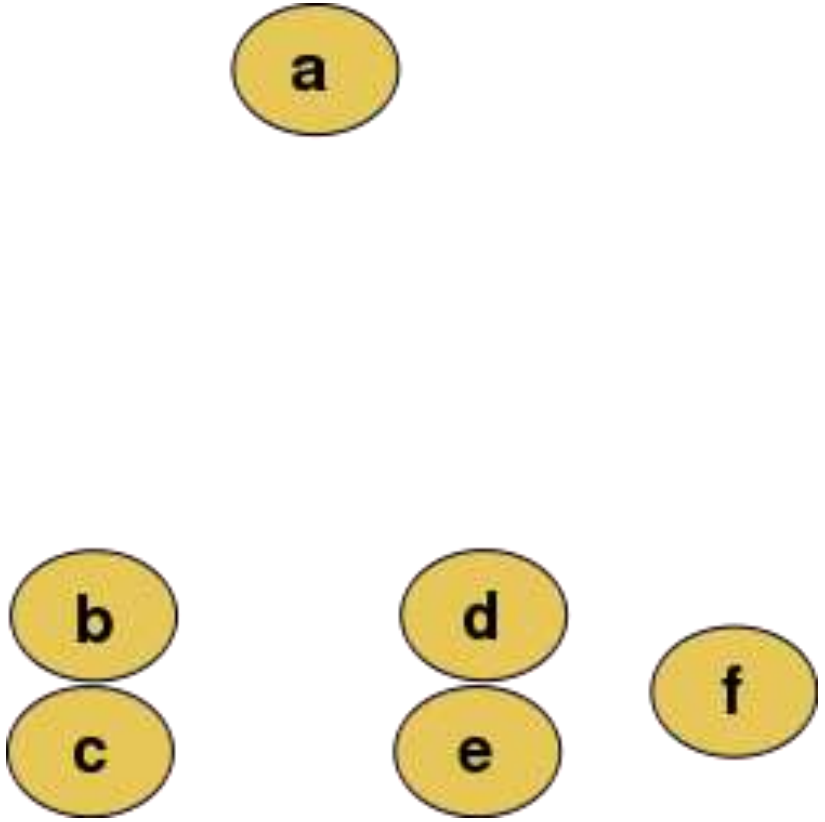
2^N は大変、 ${}_NC_2 = N(N-1)/2$で済ませる

距離行列



	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

対称行列、ペア数は三角部分



	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

距離行列から、「全体像」を復元する

- 多次元尺度構成法 Multi-Dimensional Scaling (MDS)
  - 多数の対象が空間に置かれている
  - 個々の対象の座標は調べられないが
  - ペアワイズの距離は調べ上げられる
  - 距離行列から、オリジナルの座標を復元する

MDS

- 第１点を原点に取る
- 第２点を、第１点との距離に応じて、X1軸上に取る
- 第３点を、第１点と第２点からの距離を満足するように、(X1,X2)平面上に取る
- …
- 第k点を、第1,2,…,k-1点からの距離を満足するように(X1,…,Xk-1)次元空間に取る
- N-1次元空間にきちんと配置できる

	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

MDS

- N個の対象はN-1次元空間座標に、うまく納められる
- 場合によってはもっと低い次元に納められる
- 低次元に納められれば
  - 次元縮約に成功した、ことになる

	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

MDS

- N個の対象はN-1次元空間座標に、うまく納められる
- 場合によってはもっと低い次元に納められる
- 低次元に納められれば
  - 次元縮約に成功した、ことになる
- 次元縮約を目的にした場合には、「固有値分解」して、できるだけ説明力の強い軸を取り出す工夫をしながら、座標配置を実現する

	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

MDS

- N個の対象はN-1次元空間座標に、うまく納められる
- 場合によってはもっと低い次元に納められる
- 低次元に納められれば
  - 次元縮約に成功した、ことになる
- 次元縮約を目的にした場合には、「固有値分解」して、できるだけ説明力の強い軸を取り出す工夫をしながら、座標配置を実現する

	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0

『もしも、元の空間がユークリッド空間だったなら』

MDS

- オリジナルの対象は、ユークリッド空間に置かれていたかどうかの保証はない
- ペアワイズの距離も、「ユークリッド距離」で測ったものとは限らない

- MDSをすると、固有値がすべて非負とは限らない
- 非負の固有値が出るということは、「空間をユークリッド的な平らな空間」に納めきれないということ
- 距離の測り具合が曲がっている（非ユークリッド的になっている）空間にあるとみなすこともできる

例

- 染色体の配列の違いを距離にする
  - 0010100
  - 0101100
  - 1101110 …
  - 多次元(0,1)立方体の頂点に配置されているとみなせる
  - わざわざユークリッド空間に配置しなおしたい？
  - →線形代数的な解析の対象に乗せることできる
- 配列の違いの距離の測り方に、「マンハッタン距離」と言われるものがある
- 「ユークリッド距離」ではなく「マンハッタン距離」なので、ユークリッド空間に納めきれない
- MDSの固有値に負のものが現れる

非ユークリッド空間(球面)に納める

- 染色体の配列の違いを『内積』的にとらえる
  - 似ているほど内積は大きい
- 類似度行列を作る
  - 以下の配列をベクトルとみて、対応する値が同じならカウントし異なればカウントしないことにする
  - まったく同一の配列の内積が１となるように補正する
    - 0010100
    - 0101100
    - 1101110 …
- 自配列との内積はすべて１→すべての配列は球面上の点となる
- 内積はcos(Θ)になるので、角度が配列の違いに対応する
  - 0000000 と 1111111とは、直角をなす

内積

$$V = \begin{pmatrix} v1 \\ v2 \\ \cdots \\ vn \end{pmatrix} \qquad V^T = (v1, v2, \ldots, vn)$$

- $V^T V = (v1, v2, \ldots, vn) \begin{pmatrix} v1 \\ v2 \\ \cdots \\ vn \end{pmatrix}$ = v1^2 + v2^2 + ⋯ + vn^2

内積行列

$$\begin{pmatrix} V1 \\ V2 \\ \cdots \\ Vn \end{pmatrix} (V1, V2, \ldots, Vn) = \begin{pmatrix} V1^T V1, V1^T V2, \ldots, V1^T Vn \\ V2^T V1, V2^T V2, \ldots, V2^T Vn \\ \ldots, \ldots, \ldots, \ldots, \\ Vn^T V1, Vn^T V2, \ldots., Vn^V n \end{pmatrix}$$

内積行列から座標を復元する

- 固有値分解
- H = V^{-1} Σ V
- Σは対角行列。対角成分は固有値
- 座標はＳＶ
  - ただし、Σ ＝ＳＳ、Sは固有値の平方根を対角成分とする対角行列

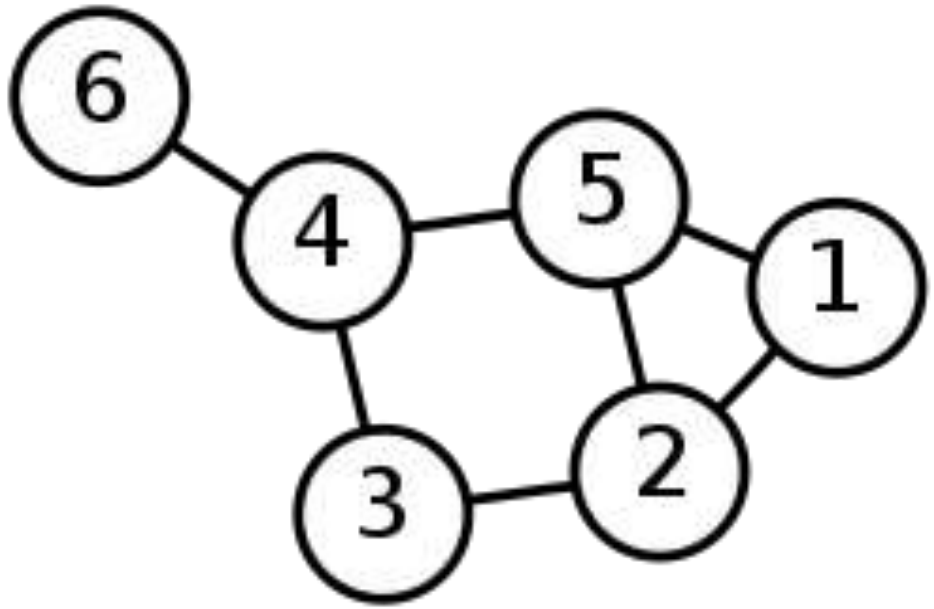
- 固有値がすべて非負なら、「ちゃんとした座標」に復元できる
- すべての対象はn次元球面上の点となる
- 配列の例の場合は、うまくこのようにできる
- 場合によっては負の固有値が出る→球面よりさらに曲がった空間を考えることになる

グラフ

- グラフは、ノードの集合とエッジの集合　からできている
- エッジはノードのペアに関して定まる

- グラフは「複数の対象＝ノード」と、その「ペア＝エッジ」を現したもの

- ${}_{N}C_{2}$ ＝ N(N-1)/2で済ませる仕組み



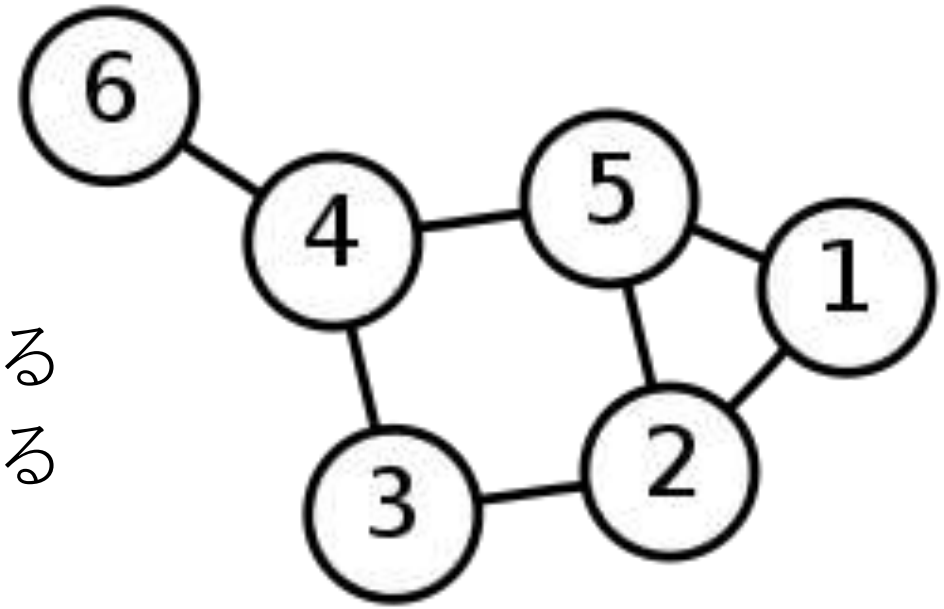
グラフ

- グラフは、ノードの集合とエッジの集合　からできている
- エッジはノードのペアに関して定まる

- グラフは「複数の対象＝ノード」と、その「ペア＝エッジ」を現したもの

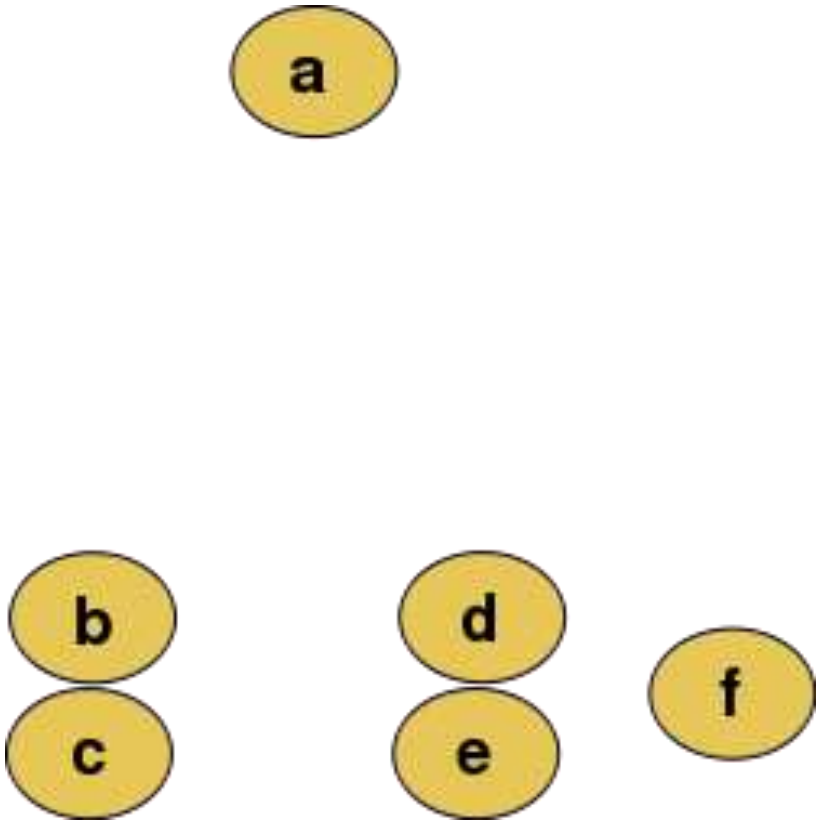
- ${}_{N}C_{2} = N(N-1)/2$で済ませる仕組み

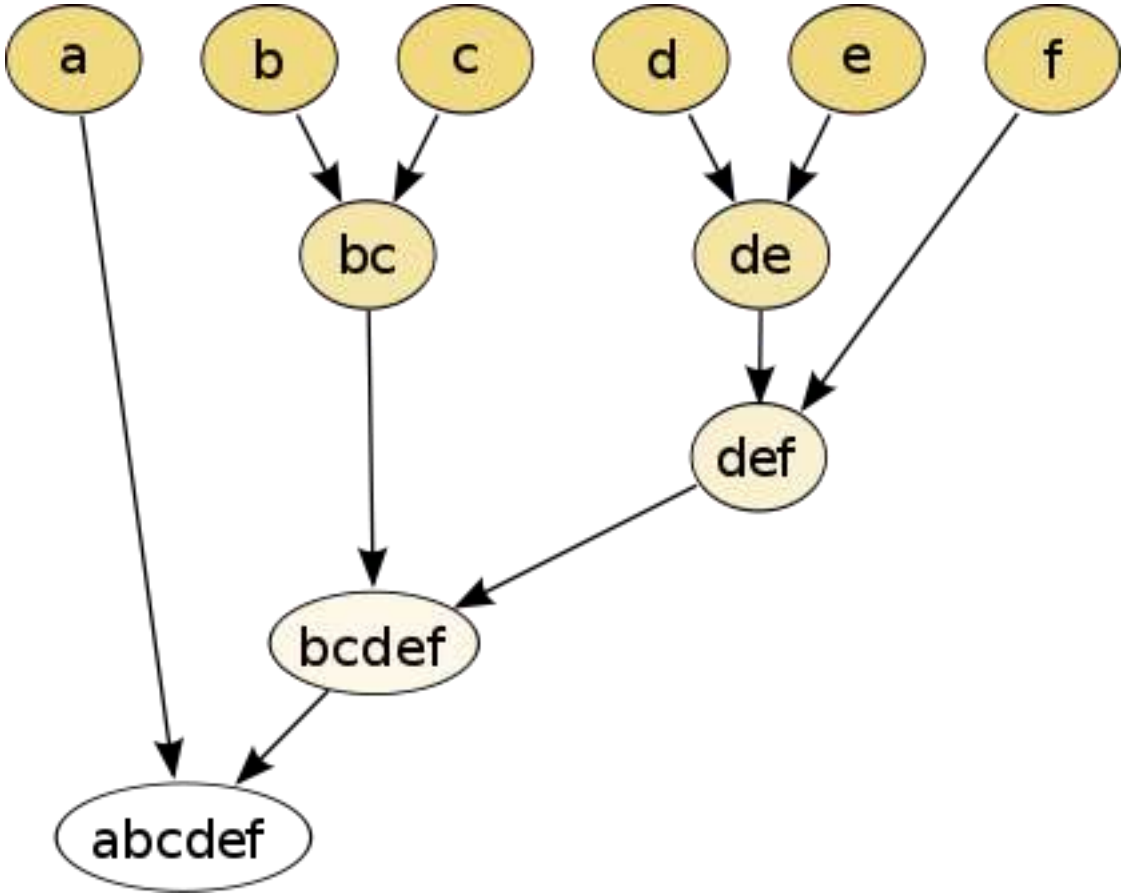
- エッジには長さを持たせることができる
- グラフ上の頂点間には距離が定義できる



距離行列からグラフを復元する

- 階層的クラスタリング



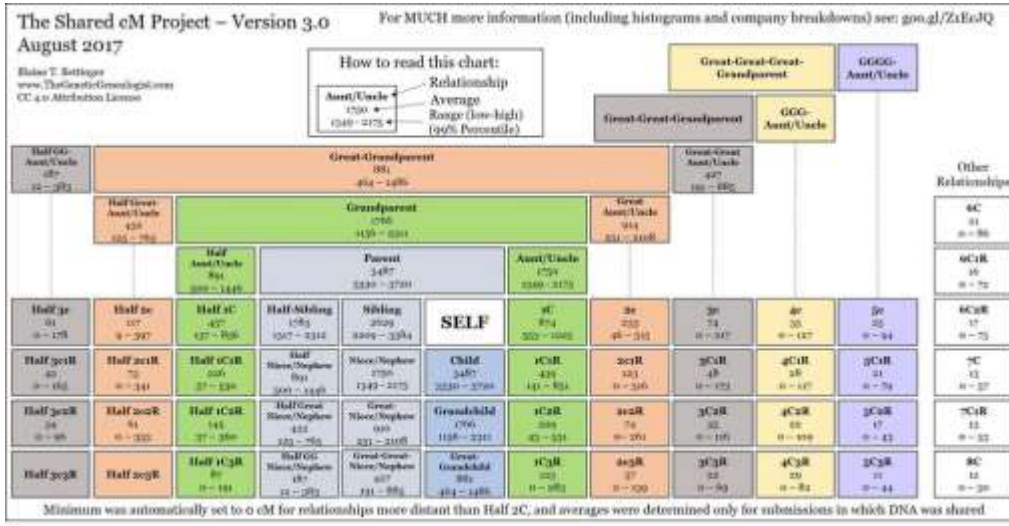


オリジナルが木構造で、その木(グラフ)上のグラフ距離が距離行列になっていれば、
完璧な木グラフが復元できる

オリジナルがユークリッド空間に配置されていればMDSで完璧なユークリッド座標が復元できた

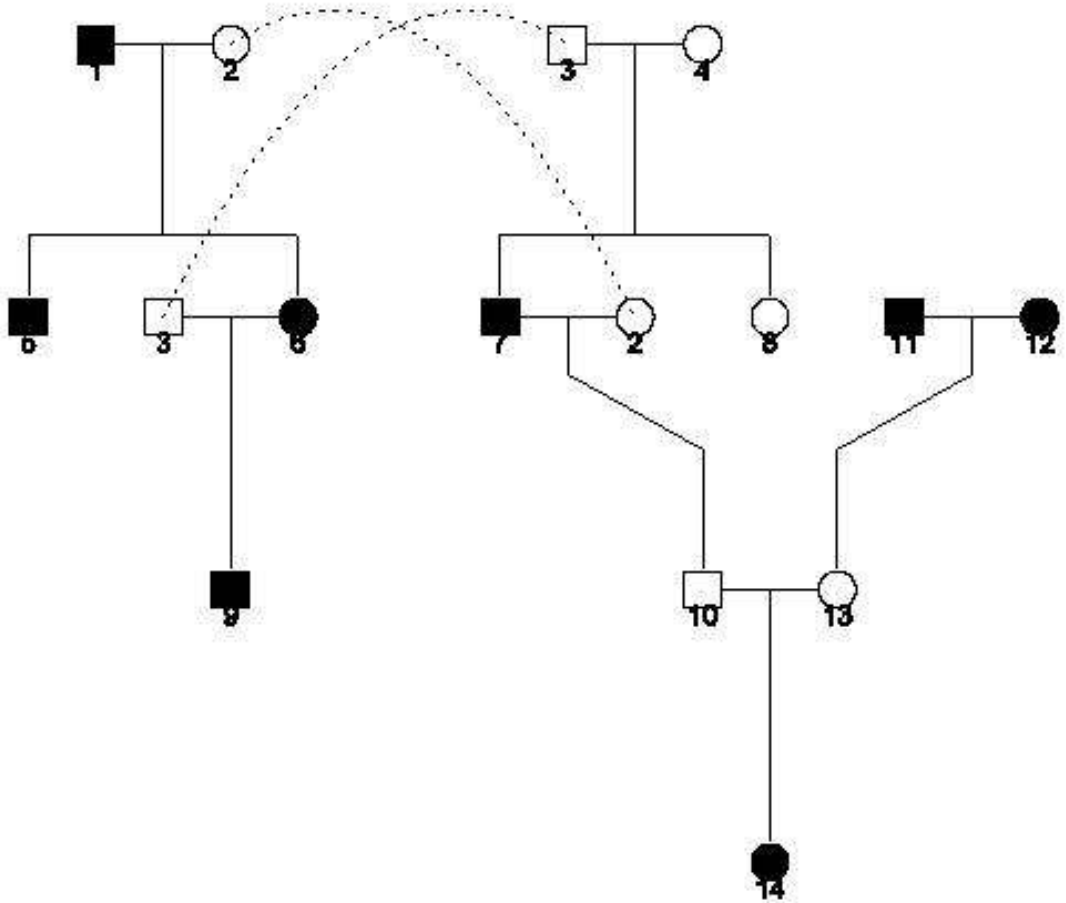
- グラフ距離が、木グラフ上の距離を満足していない場合には、ずれをどこかで吸収する必要がある
- 階層的クラスタリングのアルゴリズムがその不一致を吸収する

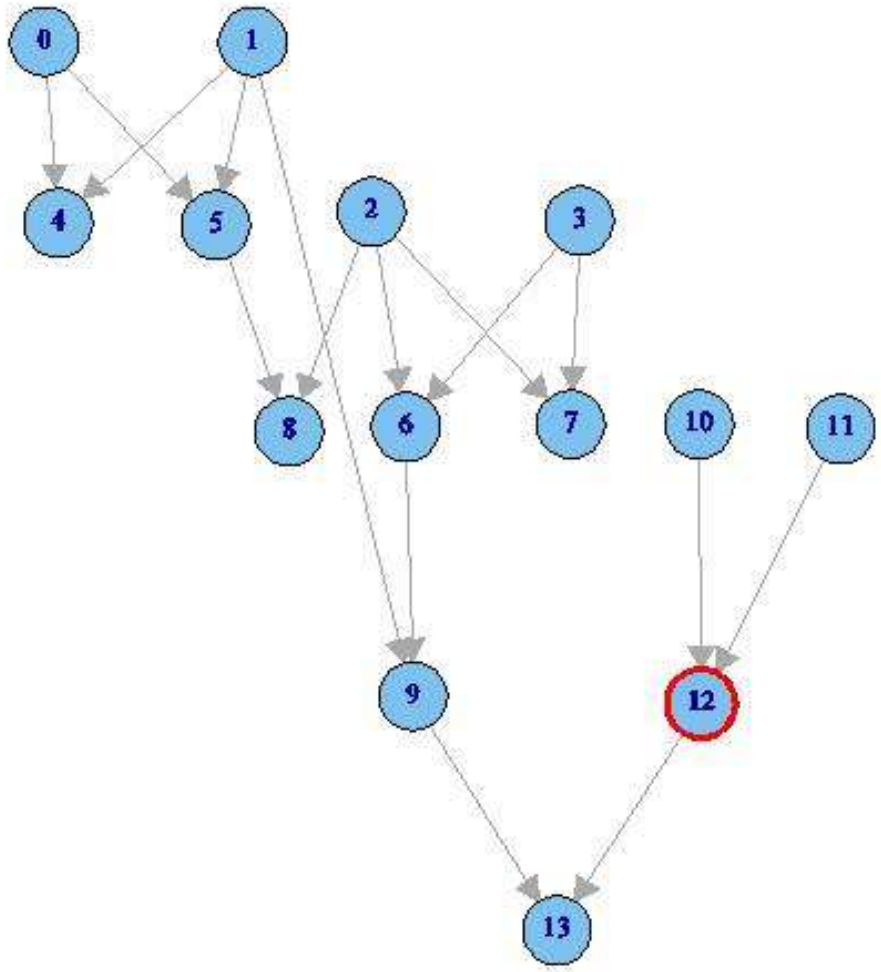
血縁度が作る類似度行列



- 自身は１
- 親子は1/2
- …
- 多型情報から得られる「血縁らしさ」の値を類似度に用いてもよい

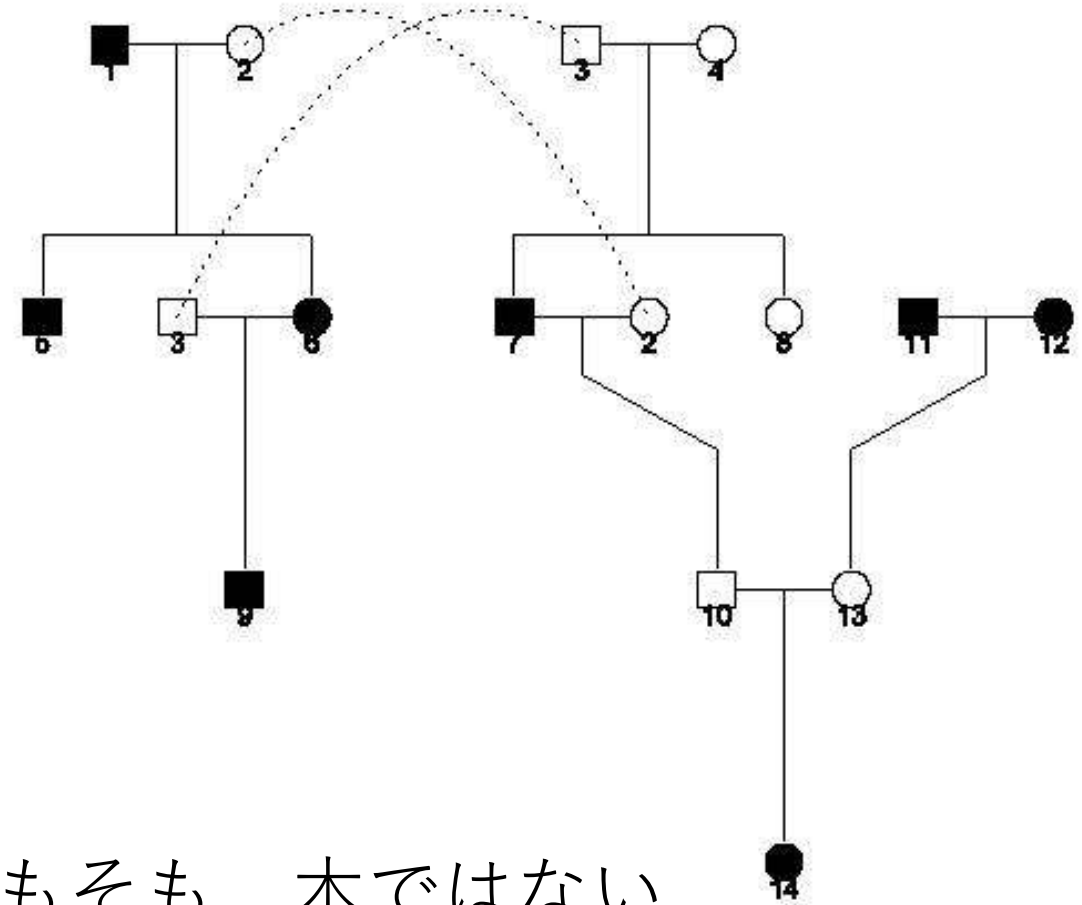
問題、再登場

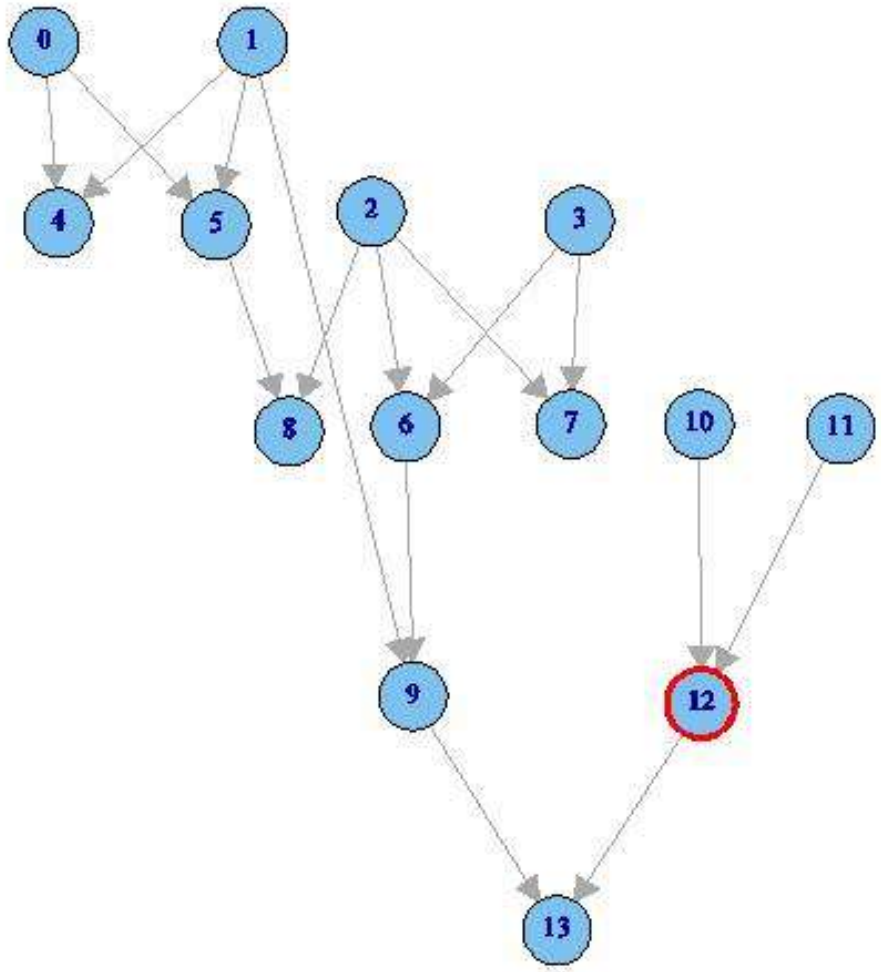




親子は1/2の関係
子によって「かすがい」された「夫婦」は、1/4の関係なのか？

家系図





そもそも、木ではない

木は

エッジの本数　＝　ノードの数－１

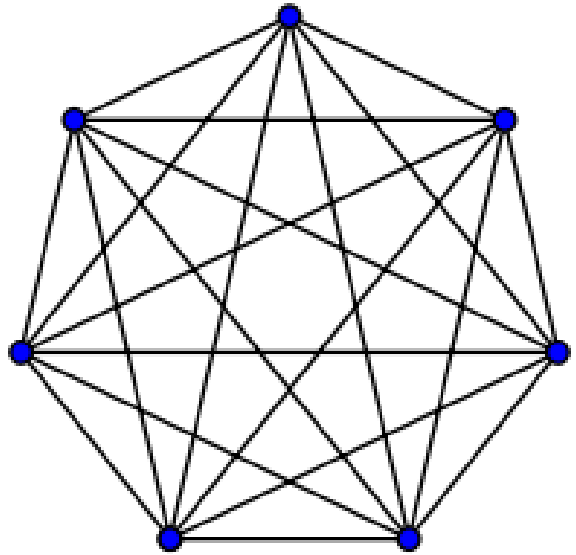
染色体・配列なら木にできる

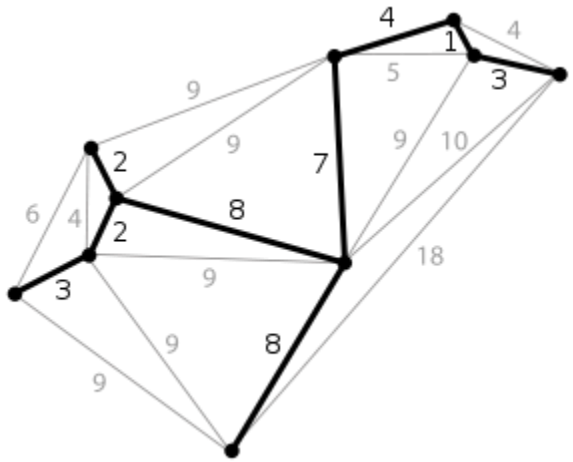
- 親の 2 本の染色体配列と、子の染色体配列とは、1/2一致の関係
- 祖父母の染色体配列と、孫の染色体配列とは、1/4一致の関係

- 配列一致程度、そのものを使うと、「グラフ上の距離」との関係がうまくいかない
- 対数を取ると…
  - 親子の染色体の違いは　log (1/2) = -1
  - 祖父母-孫の違いは、log(1/4) -2

配列の共有度の対数が作る距離行列(非類似度行列)から「木」を作る？

- 最小全域木
  - すべての配列ペアに「距離」がある
  - 配列をノードにして、すべての配列ペアにエッジを持たせる
    - すべてのノードペアにエッジがあるグラフを完全グラフと呼ぶ
  - すべてのノードを含む木であって、エッジの長さの総和が最小になるような木を最小全域木という
  - 最小全域木を見つけることは(計算機にとって)それほど大変ではない





配列の非類似度から最小全域木

- 祖父母・親子３代の配列があるとき、世代間を結ぶことは最小全域木的連結の実現になる(ことが多い)
- 親世代が抜けているときには、「世代を飛び越して」つないでやることもあり
- 非血縁者が結ばれることは、(ほぼ)ない

- すくなくとも、大まかな、配列伝達関係の骨格を、かなり単純に作り出せそう
- 血族婚により、染色体・配列の木構造が破綻しサイクルが生じる
  - 最小全域木を作った後で、木に採用されなかった、「強い類似度」のエッジを復活させることで血族婚に由来するサイクルの復活も可能か…

まとめ

- 多数の対象をペア間の類似度・非類似度の観察・定量することは、様々な場面で採用される方法
- オリジナルたちが、収まっていた構造(ユークリッド空間・木構造)を忠実に反映した距離行列が得られていれば、完璧な復元ができる
- そうでない場合は、歪みが生じる(負の固有値、曲がった空間、クラスタリングアルゴリズムによる残差処理)
- 血縁関係の場合は、２倍体による特殊性に注意
- 配列の場合は、対数を導入することで距離化できる
- 血族婚はサイクルを作る