3-5-2-2 参考 EM-algorithm による haplotype 頻度の推定

2 SNP の観測データからは 9 genotype の観測値が得られる。そのような観測値を最も得やすい 4 haplotype 頻度の分布を求めるのに用いる標準的アルゴリズムが Expectation-Maximization algorithm (EM-algorithm)と呼ばれるものである。

2 SNP の 9 genotype 観測値から 4 haplotype 本数を推定するにあたり、genotype は 2 種類に分けられる。片方の種類に分類される genotype は特定の haplotype の本数を一意的に決定する(確定的 haplotype 数)。もう片方に分類される genotype は特定の haplotype の本数を確率的に決定する(非確定的 haplotype 数)。

EM-algorithm では非確定的 haplotype 数を 4 haplotype に均等に分配し、確定的 haplotype 数と合算し、その値を基に haplotype 頻度の初期推定値を算出する。ついで、その haplotype 頻度推定値を基に、非確定的 haplotype 数を分配し、新たな haplotype 頻度の推定値を得る。ついで、新たな haplotype 頻度推定値を基に非確定的 haplotype 数の再分配を行う。この作業を繰り返すと haplotype 頻度は収束することが知られ、しかもその収束推定値は現実のデータとの整合性が高いことが知られている。エクセルファイル“2 SNP Haplotype 頻度推定”でも EM-algorithm を使用している。

参考のため、その具体的な数式を以下に掲載する。

X/x、Y/y はそれぞれ１つの SNP の 2 つのアレルを表している。
a1…a9 は X/x、Y/y の作る 9 genotype の観測値を表している。
a1 は XXYY、a2 は XXYy、a3 は XXyy、… a8 は xxYy、a9 は xxyy に相当する。
また、p_{XYn}は haplotype XY の世代 n における推定頻度を表す。ただし、下付き数字 n は 0 の場合、haplotype 推定初期値に対応し、1 以上の場合は上記で説明した非確定的 haplotype 数の分配作業の回数に相当する。
S＝a1+a2+…+a9 である。

$$p_{XY0}=(a1+1/2\times a2+1/2\times a4+1/4\times a5)/S$$
$$p_{Xy0}=(a3+1/2\times a2+1/2\times a6+1/4\times a5)/S$$
$$p_{xY0}=(a7+1/2\times a4+1/2\times a8+1/4\times a5)/S$$
$$p_{xy0}=(a9+1/2\times a6+1/2\times a8+1/4\times a5)/S$$

$$p_{XYn}=(a1+1/2\times a2+1/2\times a4+(p_{XYn-1}\times p_{xyn-1}/(p_{XYn-1}\times p_{xyn-1}+p_{Xyn-1}\times p_{xYn-1}))\times a5)/S$$
$$p_{Xyn}=(a3+1/2\times a2+1/2\times a6+(p_{Xyn-1}\times p_{xYn-1}/(p_{XYn-1}\times p_{xyn-1}+p_{Xyn-1}\times p_{xYn-1}))\times a5)/S$$
$$p_{xYn}=(a7+1/2\times a4+1/2\times a8+(p_{Xyn-1}\times p_{xYn-1}/(p_{XYn-1}\times p_{xyn-1}+p_{Xyn-1}\times p_{xYn-1}))\times a5)/S$$
$$p_{xyn}=(a9+1/2\times a6+1/2\times a8+(p_{XYn-1}\times p_{xyn-1}/(p_{XYn-1}\times p_{xyn-1}+p_{Xyn-1}\times p_{xYn-1}))\times a5)/S$$