EM Algorithm (Expectation-Maximization Algorithm): Numerical optimization routine that is often helpful for making maximum likelihood inferences

Best for cases where hard to analytically find m.l.e. for observed data but would be easy if could observe some hidden data

e.g., if could observe paths and emissions for HMM

e.g., if could observe times at which sequences change and actual changes that sequences experience at those times

e.g., if could observe genotype rather than phenotype for estimating allele frequencies (ABO blood groups)

Genotype	Hardy-Weinberg Probabilities	Phenotype
AA	p_A^2	A
AB	$2p_A p_B$	AB
AO	$2p_A p_O$	A
BB	p_B^2	B
BO	$2p_B p_O$	B
OO	p_O^2	O

Phenotype Counts are:

N_{A-}

N_{AB}

N_{B-}

N_{OO}

When only phenotypes are observed, likelihood is ...

$$(p_A^2+2p_A p_O)^{N_{A-}} (2p_A p_B)^{N_{AB}} (p_B^2+2p_B p_O)^{N_{B-}} (p_O^2)^{N_{OO}}$$

Difficult to analytically find max. likelihood estimates of allele frequencies with messy formula above ...

Genotype	Hardy-Weinberg Probabilities	Number Observed
AA	p_A^2	N_{AA}
AB	$2p_A p_B$	N_{AB}
AO	$2p_A p_O$	N_{AO}
BB	p_B^2	N_{BB}
BO	$2p_B p_O$	N_{BO}
OO	p_O^2	N_{OO}

$$\text{Likelihood} = (p_A^2)^{N_{AA}} (2p_A p_B)^{N_{AB}} (2p_A p_O)^{N_{AO}} (p_B^2)^{N_{BB}} (2p_B p_O)^{N_{BO}} (p_O^2)^{N_{OO}}$$

above is proportional to ...

$$p_A^{(2N_{AA}+N_{AB}+N_{AO})} p_B^{(2N_{BB}+N_{AB}+N_{BO})} p_O^{(2N_{OO}+N_{AO}+N_{BO})}$$

(maximum likelihood estimates of allele frequencies are proportions of alleles in samples)

Basic idea of EM algorithm applied to ABO blood groups:

Observed data are numbers each blood type: $N_{A-}, N_{AB}, N_{B-}, N_{OO}$

Full data would be: $N_{AA}, N_{AB}, N_{AO}, N_{BB}, N_{BO}, N_{OO}$

Full data log-likelihood proportional to:

$$(2N_{AA}+N_{AB}+N_{AO}) \log p_A + (2N_{BB}+N_{AB}+N_{BO}) \log p_B + (2N_{OO}+N_{AO}+N_{BO}) \log p_O$$

Sufficient statistics:

$$N_A = 2N_{AA}+N_{AB}+N_{AO}, N_B = 2N_{BB}+N_{AB}+N_{BO}, N_O = 2N_{OO}+N_{AO}+N_{BO}$$

Note: full-data log-likelihood is linear in sufficient statistic values

Full-data parameter max. likelihood estimates would be:

$$\widehat{p_A} = \frac{N_A}{2N}$$

$$\widehat{p_B} = \frac{N_B}{2N}$$

$$\widehat{p_O} = \frac{N_O}{2N}$$

$$N = N_{AA} + N_{AB} + N_{AO} + N_{BB} + N_{BO} + N_{OO}$$

$$2N = N_A + N_B + N_O$$

Step 1: Start with initial guesses to parameter values ${p_A}^{(0)}, {p_B}^{(0)}, {p_O}^{(0)}$. Set $h = 0$.

Step 2: Calculate expected sufficient statistic values N_A, N_B, N_O conditional upon ${p_A}^{(h)}, {p_B}^{(h)}, {p_O}^{(h)}$ and conditional upon observed data.

In other words, set:

$$ {N_A}^{(h)} = E[N_A | {p_A}^{(h)}, {p_B}^{(h)}, {p_O}^{(h)}, N_{A-}, N_{AB}, N_{B-}, N_{OO}] $$

$$ {N_B}^{(h)} = E[N_B | {p_A}^{(h)}, {p_B}^{(h)}, {p_O}^{(h)}, N_{A-}, N_{AB}, N_{B-}, N_{OO}] $$

$$ {N_O}^{(h)} = E[N_O | {p_A}^{(h)}, {p_B}^{(h)}, {p_O}^{(h)}, N_{A-}, N_{AB}, N_{B-}, N_{OO}] $$

Because each genotype is diploid,

$$E[N_A|p_A, p_B, p_O, N_{A-}, N_{AB}, N_{B-}, N_{OO}] = \frac{2{p_A}^2 + 2p_A p_O}{{p_A}^2 + 2p_A p_O} N_{A-} + N_{AB}$$

$$E[N_B|p_A, p_B, p_O, N_{A-}, N_{AB}, N_{B-}, N_{OO}] = \frac{2{p_B}^2 + 2p_B p_O}{{p_B}^2 + 2p_B p_O} N_{B-} + N_{AB}$$

$$E[N_O|p_A, p_B, p_O, N_{A-}, N_{AB}, N_{B-}, N_{OO}] =$$

$$\frac{2p_A p_O}{{p_A}^2 + 2p_A p_O} N_{A-} + \frac{2p_B p_O}{{p_B}^2 + 2p_B p_O} N_{B-} + 2N_{OO}$$

Step 3: Find max. likelihood estimates if expected values actually observed

$$p_A{}^{(h+1)} = \frac{N_A{}^{(h)}}{2N}$$

$$p_B{}^{(h+1)} = \frac{N_B{}^{(h)}}{2N}$$

$$p_O{}^{(h+1)} = \frac{N_O{}^{(h)}}{2N}$$

Step 4: Decide whether to terminate EM. If not, set h=h+1 and go to Step 2.

Probabilistic models of nucleotide substitution (when sites evolve identically and independently)

Let q_{ij} be the instantaneous rate of change at a site from nucleotide type i to type j

Q will refer to the matrix of instantaneous rates (Q will have 4 rows and 4 columns because i and j can each by any of 4 nucleotide types)

For a nucleotide that starts as type i at time 0, the probability that nucleotide is type j at time t is denoted $p_{ij}(t)$.

$p_{ij}(t)$ is referred to as a *transition probability*.

Rate Matrix for General Time Reversible Model

FROM \ To	A	C	G	T
A	$-\mu(a\pi_C + b\pi_G + c\pi_T)$	$\mu a\pi_C$	$\mu b\pi_G$	$\mu c\pi_T$
C	$\mu a\pi_A$	$-\mu(a\pi_A + d\pi_G + e\pi_T)$	$\mu d\pi_G$	$\mu e\pi_T$
G	$\mu b\pi_A$	$\mu d\pi_C$	$-\mu(b\pi_A + d\pi_C + f\pi_T)$	$\mu f\pi_T$
T	$\mu c\pi_A$	$\mu e\pi_C$	$\mu f\pi_G$	$-\mu(c\pi_A + e\pi_C + f\pi_G)$

Consider a **very very** small amount of evolutionary time Δt.

When $i \neq j$,

$$p_{ij}(\Delta t) \doteq q_{ij}\Delta t$$

Also,

$$p_{ii}(\Delta t) \doteq 1 - \sum_{j,j\neq i} q_{ij}\Delta t$$

Equivalently,

$$p_{ii}(\Delta t) \doteq 1 + q_{ii}\Delta t$$

where

$$q_{ii} = \sum_{j,j\neq i} -q_{ij}$$

(in preceding equations, $\doteq$ can be replaced by $=$ when the limit as Δt approaches 0 is taken)

With any rate matrix (including above), the matrix of transition probabilities $P(t)$ can be determined from the rate matrix Q and the amount of evolution t via

$$P(t) = e^{Qt} = I + \frac{(Qt)}{1!} + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \ldots,$$

where I is the identity matrix.

Felsenstein 1981 model assumes sequence positions evolve independently and identically.

Let π_j be the probability that a residue is type j. π_j is often called the equilibrium probability of residue type j.

$$p_{ij}(\infty) = \pi_j$$

For Jukes-Cantor model, $\pi_j = 1/4$ for all 4 nucleotide types j.

Computing $p_{ij}(t)$ for the Felsenstein 1981 model

The Felsenstein 1981 model assumes that this is how nucleotide substitution occurs:

1. For each site in the sequence, an "event" will occur with probability s per unit evolutionary time.

2. If no event occurs, the residue at the site does not change.

3. If an event occurs, the probability that a residue is type i after the event is π_i.

What is the probability that no event occurs in t units of evolutionary time?

$$(1-s) \times (1-s) \times (1-s) \ldots (1-s) = (1-s)^t.$$

When s is close to 0,

$$1 - s \doteq e^{-s}.$$

So,

$$\Pr(\text{no event}) = (1 - s)^t \doteq e^{-st}.$$

When s is redefined as an instantaneous rate per unit evolutionary time, the approximation becomes an equality:

$$\Pr(\text{no event}) = e^{-st}.$$

Therefore,

$$\Pr(\text{at least one event}) = 1 - \Pr(\text{no event}) = 1 - e^{-st}.$$

If there have been no "events", then the residue cannot possibly have changed after an amount of evolution t.

If there has been at least one event, then the residue is type j with probability π_j.

Poisson process: "s" is rate of events, "t" is time duration

Pr(no events in time t) = e^{-st}

Prob density first event at time t = se^{-st} (the exponential distribution)
mean of exponential distribution with parameter s is 1/s

Prob(K events in time t)= $\frac{e^{-st}(st)^K}{K!}$ (the Poisson distribution)

expected # events in time t = st
variance # events in time t = st

Jargon: If stochastic process has variance of # events in time t exceeding its mean, it is called **overdispersed**. If variance is less than mean, it is **underdispersed**.

Therefore,

$$\begin{aligned} p_{ii}(t) &= \Pr(\text{no events}) + \Pr(\text{at least one event})\pi_i \\ &= e^{-st} + (1 - e^{-st})\pi_i. \end{aligned}$$

For $i \neq j$,

$$\begin{aligned} p_{ij}(t) &= \Pr(\text{at least one event})\pi_j \\ &= (1 - e^{-st})\pi_j. \end{aligned}$$

Notice that s and t appear only as a product. s and t cannot be separately estimated. Only their product can be estimated.

The Felsenstein 1981 model and other models of sequence evolution have a property known as **Time Reversibility**.

Time reversibility means that $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$ for all i, j, and t.

Equivalently, $\pi_i q_{ij} = \pi_j q_{ji}$ for all i and j.

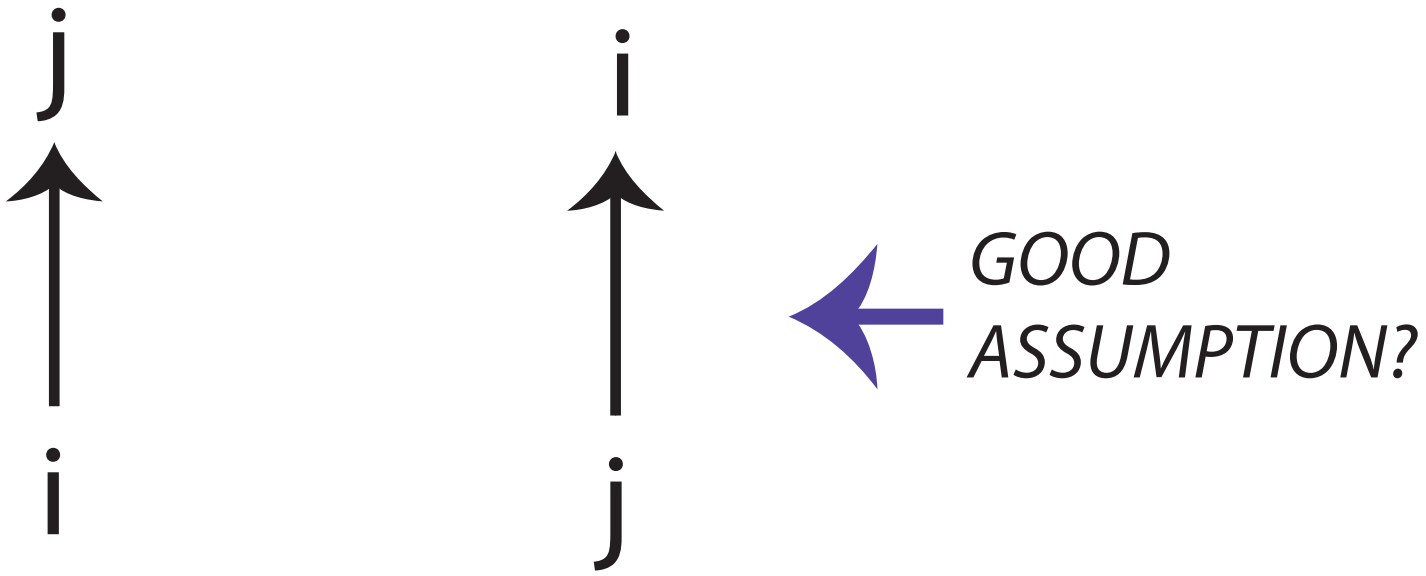
For phylogeny reconstruction, time reversibility means that we cannot (on the basis of sequence data alone) hope to distinguish which of two sequence is ancestral and which is the descendant.

The practical implication of time reversibility for phylogeny reconstruction is that maximum likelihood cannot infer the position of the root of the tree unless additional information information exists (e.g., which taxa are the outgroups) or additional assumptions are made (e.g., a molecular clock).

Widely used models of sequence change assume "time reversibility"

What is "time reversibility" ? (see also "Detailed Balance")

Intuition: "time reversibility" = "evolution has no direction"



Time Reversibility means that stationary probability of ancestor multiplied by transition probability are identical for 2 above scenarios, no matter what nucleotides represented by "i" and "j"

Consider a data set consisting of 2 aligned sequences.

Let n_{ij} be the number of sites where the first sequence has nucleotide type i and where the second sequence has nucleotide type j.

Let n represent the entire data set (i.e., n_{ij} for all i and j in $\{A, C, G, T\}$).

The Felsenstein 1981 parameters are $\pi = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ and st.

The likelihood is $\Pr(n \mid \pi, st)$ and maximum likelihood estimates of π and st are obtained by finding the combination of π and st that maximizes $\Pr(n \mid \pi, st)$.

The likelihood is

$$\Pr(n \mid \pi, st) = \prod_i \prod_j (\pi_i p_{ij}(t))^{n_{ij}}$$

where i and j are both in $\{A, C, G, T\}$.

Likelihoods can be a bit complicated to maximize...

Consider a site that has nucleotide type i in one sequence and nucleotide type j at the same position in another sequence.

For the Felsenstein 1981 and other time-reversible substitution models, we can consider i to be the ancestor and j to be the descendant.

Let t be the amount of time separating the ancestral and descendant state.

The time of the ancestral state i will be 0.

Because transition probabilities for Felsenstein 1981 model have s and t confounded (i.e., only product of s and t can be estimated), we will consider t to be 1 and will focus on estimating s.

The time of the descendant state j will then be 1.

Our "observed" data are i at the beginning of the timespan and j at the end of the time span.

The "unobserved" data are the character state of our site at all times greater than 0 and less than 1. These unobserved data are also sometimes referred to as "latent" data or missing data.

Now, imagine that we actually do know the state of our site at all times between 0 and 1.

This means that we know exactly how many substitutions occurred and we know exactly which substitutions occurred and we know exactly when each substitution occurred.

We will refer to this complete history as a site path and we will denote this site path by ρ.

Notation:

Assume ρ specifies exactly K nucleotide substitutions.

We will have $t(z)$ be the time of the z^{th} substitution $(1 \leq z \leq K)$.

For convenience, set $t(0) = 0$ and $t(K+1) = 1$.

We will have $i(z)$ be the nucleotide type of the site immediately after the z^{th} substitution $(1 \leq z \leq K)$.

For convenience, set $i(0) = i$ and $i(K+1) = i(K) = j$.

Earlier, we defined q_{ij} as the rate of change from i to j. Here, we maintain that definition with the exception of the meaning of q_{ii}. To account for hidden events in the Felsenstein 1981 model, we will have q_{ii} be $s\pi_i$.

This means $q_{ij} = s\pi_j$ for all i and j.

Define

$$q_{i\bullet} = \sum_j q_{ij}$$

t(k+1)=1 | i(k+1)=j

If we had complete information (i.e., information about ρ as well as about i at time 0 and j at time 1), our likelihood would be

t(k) | i(k)

$$\Pr(i, j, \rho \mid s, \pi) = \Pr(i \mid s, \pi)\Pr(j, \rho \mid i, s, \pi) = \pi_i \Pr(\rho \mid i, s, \pi)$$

$$= \pi_i \Big(\prod_{z=1}^{K} e^{-q_{i(z-1),\bullet}(t(z)-t(z-1))} \frac{q_{i(z-1),i(z)}}{q_{i(z-1),\bullet}} q_{i(z-1),\bullet} \Big)$$

$$\times e^{-q_{i(K),\bullet}(t(K+1)-t(K))}$$

t(2) | i(2)

$$= \pi_i \Big(\prod_{z=1}^{K} e^{-q_{i(z-1),\bullet}(t(z)-t(z-1))} q_{i(z-1),i(z)} \Big) e^{-q_{i(K),\bullet}(t(K+1)-t(K))}$$

Note: If we did not allow hidden events and if $q_{i\bullet}$ was defined by summing q_{ij} over all j that were not equal to i, then the above equation would apply to any Markov model of sequence change (even if i and j represent sequences rather than sites).

t(1) | i(1)

t(0)=0 | i(0)=i

For $w \in \{A, C, G, T\}$, let $c(w)$ be the count of times when $i(z) = w$ for $0 \leq z \leq K$.

With Felsenstein 1981 model, we have the likelihood

$$\begin{aligned}
\Pr(i, j, \rho \mid s, \pi) &= \pi_i \Big(\prod_{z=1}^{K} s\pi_{i(z)} e^{-s(t(z)-t(z-1))} \Big) e^{-s(t(K+1)-t(K))} \\
&= \pi_i \Big(\prod_{z=1}^{K} \pi_{i(z)} \Big) s^K e^{-s} \\
&= \pi_A^{c(A)} \pi_C^{c(C)} \pi_G^{c(G)} \pi_T^{c(T)} s^K e^{-s}
\end{aligned}$$

The log-likelihood is

$$\log \Pr(i, j, \rho \mid s, \pi) = K \log s - s + \sum_{w \in \{A,C,G,T\}} c(w) \log \pi_w$$

Maximum likelihood estimators are

$$\widehat{\pi}_w = \frac{c(w)}{c(A) + c(C) + c(G) + c(T)}$$

and

$$\hat{s} = K$$

.

If we knew s and π and had only the observed residues i and j, what would be our estimates of K, $c(A)$, $c(C)$, $c(G)$ and $c(T)$? First, note that

$$\begin{aligned}
E[K|K \geq 1, s] &= \sum_{z=0}^{\infty} z\Pr\left(K = z \mid K \geq 1, s\right) \\
&= \sum_{z=0}^{\infty} z\frac{\Pr\left(K = z, K \geq 1 \mid s\right)}{\Pr\left(K \geq 1 \mid s\right)} \\
&= \sum_{z=1}^{\infty} z\frac{\Pr\left(K = z \mid s\right)}{\Pr\left(K \geq 1 \mid s\right)} \\
&= \frac{1}{\Pr\left(K \geq 1 \mid s\right)} \sum_{z=1}^{\infty} z\Pr\left(K = z \mid s\right) \\
&= \frac{1}{1 - e^{-s}} \sum_{z=0}^{\infty} z\Pr\left(K = z \mid s\right) = \frac{E[K|s]}{1 - e^{-s}} = \frac{s}{1 - e^{-s}}
\end{aligned}$$

Note that

$$\begin{aligned} E[K|i,j,\pi,s] &= E[K|K=0,i,j,\pi,s]\Pr\left(K=0 \mid i,j,\pi,s\right) + \\ &\quad E[K|K\geq 1,i,j,\pi,s]\Pr\left(K\geq 1 \mid i,j,\pi,s\right) \\ &= \Pr\left(K\geq 1 \mid i,j,\pi,s\right)E[K|K\geq 1,i,j,\pi,s] \end{aligned}$$

If $i = j$,

$$\begin{aligned} E[K|i,j,\pi,s] &= \Pr(K \geq 1 \mid i,j,\pi,s) E[K|K \geq 1,i,j,\pi,s] \\ &= \frac{\Pr(K \geq 1, i, j \mid \pi, s)}{\Pr(i,j \mid \pi,s)} E[K|K \geq 1,s] \\ &= \frac{(1-e^{-s})\pi_j}{e^{-s}+(1-e^{-s})\pi_j} \times \frac{s}{1-e^{-s}} \\ &= \frac{s\pi_j}{e^{-s}+(1-e^{-s})\pi_j} \end{aligned}$$

If $i \neq j$,

$$\begin{aligned} E[K|i,j,\pi,s] &= \Pr(K \geq 1 \mid i,j,\pi,s) E[K|K \geq 1,i,j,\pi,s] \\ &= 1 \times \frac{s}{1-e^{-s}} \end{aligned}$$

Let $\delta(a, b)$ be 1 if $a = b$ and 0 otherwise.

For $w \in \{A, C, G, T\}$,

$$
\begin{aligned}
E[c(w)|i,j,\pi,s] &= E[c(w)|i,j,\pi,s,K=0]\Pr(K=0 \mid i,j,\pi,s) + \\
&\quad E[c(w)|i,j,\pi,s,K\geq 1]\Pr(K\geq 1 \mid i,j,\pi,s) \\
&= \delta(i,j)\delta(i,w)\frac{e^{-s}}{e^{-s}+(1-e^{-s})\pi_j} + \\
&\quad \delta(i,j)(2\delta(i,w)) + (\frac{s}{1-e^{-s}}-1)\pi_w)\frac{(1-e^{-s})\pi_j}{e^{-s}+(1-e^{-s})\pi_j} \\
&\quad +(1-\delta(i,j))(\delta(i,w)(1-\delta(j,w))+(1-\delta(i,w))\delta(j,w) \\
&\quad +(\frac{s}{1-e^{-s}}-1)\pi_w) \times 1
\end{aligned}
$$

K=0

P(K>=1) when i=j

i=j=w

hidden sub of type w

i not equal to j

hidden sub of type w

P(K>=1) when i not equal to j

Earlier, we showed that the log-likelihood is

$$\log \Pr\left(i, j, \rho \mid s, \pi\right) = K \log s - s + \sum_{w \in \{A,C,G,T\}} c(w) \log \pi_w$$

The values of K, $c(A)$, $c(C)$, $c(G)$, and $c(T)$ are termed **sufficient** statistics because they capture all of the information from i, j, and ρ upon which the likelihood depends.

In the usual situation, i and j constitute our "observed data" and the remainder of ρ represents missing information.

The idea of the EM-algorithm applied to Felsenstein 1981 model is to:

1. Start with initial guesses as to the parameters: $s^{(0)}$ and $\pi^{(0)}$ and set $h = 0$

2. Calculate expected values of the sufficient statistics conditional upon $s^{(h)}$ and $\pi^{(h)}$. This means calculate

$$K^{(h)} = E[K|i, j, \pi^{(h)}, s^{(h)}]$$

and

$$c(w)^{(h)} = E[c(w)|i, j, \pi^{(h)}, s^{(h)}].$$

3. Pretend the expected values of the sufficient statistics were actually observed and then apply the maximum likelihood formulae for the observed plus missing data:

$$\pi_w^{(h+1)} = \frac{c(w)^{(h)}}{c(A)^{(h)} + c(C)^{(h)} + c(G)^{(h)} + c(T)^{(h)}}$$

and

$$s^{(h+1)} = K^{(h)}$$

4. Decide whether to terminate EM algorithm. If not, set h=h+1 and go to Step 2.

Note: Above algorithm formulated for silly case where data set consists of a single site. Real algorithm is straightforward to extend to case where data set consists of many sites.

Note: Each EM cycle is guaranteed not to lower likelihood.

Application of the EM algorithm for 10,000 sites from two sequences, simulated according to Felsenstein 1981 model.

	0	1	2	3	4	5	...	10	True
s	0.100	0.305	0.397	0.442	0.465	0.477	...	0.491	0.500
π_A	0.200	0.387	0.397	0.400	0.401	0.402	...	0.402	0.400
π_C	0.250	0.299	0.299	0.299	0.299	0.299	...	0.299	0.300
π_G	0.400	0.212	0.205	0.204	0.203	0.203	...	0.202	0.200
π_T	0.150	0.102	0.099	0.098	0.097	0.097	...	0.097	0.100

EM algorithm in general (expanded and slightly modified from pages 324-325 of Durbin et al. text) ...

We have observed data x and missing information y and model parameters θ

Likelihood for observed data x

$$\Pr(x \mid \theta) = \sum_y \Pr(x, y \mid \theta)$$

$$\Pr(y \mid x, \theta) = \Pr(x, y \mid \theta) / \Pr(x \mid \theta)$$

So,

$$\log \Pr(y \mid x, \theta) = \log \Pr(x, y \mid \theta) - \log \Pr(x \mid \theta)$$

So,

$$\log \Pr(x \mid \theta) = \log \Pr(x, y \mid \theta) - \log \Pr(y \mid x, \theta)$$

Let θ^t represent one set of parameter values (with the hope that the sequence $\theta^1, \theta^2, \ldots, \theta^t, \theta^{t+1}$ represents a series of parameter sets that produce increasingly higher likelihoods $\Pr(x \mid \theta)$).

We multiply both sides of previous equation by $\Pr(y \mid x, \theta^t)$ and then sum over y to get

$$\sum_y \Pr(y \mid x, \theta^t) \log \Pr(x \mid \theta) =$$
$$\sum_y \Pr(y \mid x, \theta^t) \log \Pr(x, y \mid \theta) - \sum_y \Pr(y \mid x, \theta^t) \log \Pr(y \mid x, \theta)$$

$$\log \Pr(x \mid \theta) = \sum_y \Pr(y \mid x, \theta^t) \log \Pr(x, y \mid \theta) - \sum_y \Pr(y \mid x, \theta^t) \log \Pr(y \mid x, \theta)$$

We rename the first expression on the right

$$Q(\theta|\theta^t) = \sum_y \Pr(y \mid x, \theta^t) \log \Pr(x, y \mid \theta) = E[\log \Pr(x, y \mid \theta)|x, \theta^t]$$

Determining $Q(\theta|\theta^t)$ is the "E" step of the EM-algorithm. For our Felsenstein 1981 example, we had

$$E[\log \Pr(i, j, \rho \mid s, \pi)|i, j, s^t, \pi^t] =$$

$$E[K|i, j, s^t, \pi^t] \log s - s + \sum_{w \in \{A,C,G,T\}} E[c(w)|i, j, s^t, \pi^t] \log \pi_w$$

(See * on p.123)

Because

$$\begin{aligned}\log \Pr(x \mid \theta^t) &= \sum_y \Pr(y \mid x, \theta^t) \log \Pr(x, y \mid \theta^t) - \sum_y \Pr(y \mid x, \theta^t) \log \Pr(y \mid x, \theta^t) \\ &= Q(\theta^t|\theta^t) - \sum_y \Pr(y \mid x, \theta^t) \log \Pr(y \mid x, \theta^t),\end{aligned}$$

we have

$$\log \Pr(x \mid \theta) - \log \Pr(x \mid \theta^t) = Q(\theta|\theta^t) - Q(\theta^t|\theta^t) + \sum_y \Pr(y \mid x, \theta^t) \log \frac{\Pr(y \mid x, \theta^t)}{\Pr(y \mid x, \theta)}.$$

A fancy name for the part on the right side after $Q(\theta|\theta^t) - Q(\theta^t|\theta^t)$ is the **Kullback-Leibler divergence** of $\Pr(y \mid x, \theta)$ from $\Pr(y \mid x, \theta^t)$. This divergence must be ≥ 0 (proof not shown). This means

$$\log \Pr(x \mid \theta) - \log \Pr(x \mid \theta^t) \geq Q(\theta|\theta^t) - Q(\theta^t|\theta^t)$$

Therefore, the left side of above equation is not negative if the right side is not negative. We can be sure the right side is not negative by choosing the value θ^{t+1} that is the maximum over all θ of $Q(\theta|\theta^t)$. This is the "M" step of the EM-algorithm.