ASR paper review

April 18, 2020

1 Joint Reconstruction

Yang, 1996: ASRV (stochastic but independent)

- Subtitution rate variation among all sites in gene and protein sequences (except pseudogenes)
- uniform substitution rate → # of substitutions per site $\sim Poisson(x)$
  gamma-distributed substitution rate → # of substitutions per site $\sim NegBinomial(k, p)$
- Assume d expected substitutions per site, the rate variation term r, the average number of replacement each position: $d \times r$, $f(r; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{\beta r} r^{\alpha - 1}$
- $\alpha = \beta$, $\mathbb{E}(Gamma(\alpha, \beta)) = 1$, larger α → more "constant" rate.
- **discrete-gamma model**: k categories to approximate the continuous gamma distribution

Pupko, 2000

- PAML (1995) → inefficient implementation of joint reconstruction (exponential)
- Propose linear (# of sequences) DP algorithm for joint ML ancestral reconstruction:

$$\begin{aligned} max\ P(v \mid data) &= max\ \frac{P(data \mid v)P(v)}{P(data)} \propto max\ P(data \mid v)P(v) \\ &= max\ P(root)\prod_{p,c} P_{p,c}(t) \end{aligned}$$

$P_{p,c}(t)$: predecessor → child replacement probability with branch length t

Traverse **bottom-up** from each leave node (taxa), compute quantity $L_i(x)$ and character state $C_i(x)$ at node x (assume diffdrent sites along the sequence at x evolve independently):

$L_i(x)$:likelihood of the best reconstruction subtree given **predecessor** of x is character i
$C_i(x)$: most likely character at node x given **predecessor** of x is character i

  - Assume leave node (OTU) y with observed character j: $C_i(y) = j$ for all possible predecessor character i, $L_i(y) = P_{i,j}(t_y)$

- Assume unvisited internal node z, whereas both children nodes x, y has been visited (known $L_j(x), C_j(x), L_j(y), C_j(y)$). For each possible character i at predecessor node of z, compute:
  $L_i(z) = max_j \ P_{i,j}(t_z) \cdot L_j(x) \cdot L_j(y)$
  $C_i(z) = \underset{j}{argmax} \ L_z(i)$

Traverse **top-down** from root to leaves, assign most likely ancestral character at each node x (assume its predecessor is node z:

- recover the assigned character i at predecessor node z
- assign $C_y(i)$ for current node

Pupko, 2002: Branch & Bound

- Pupko (2000) DP algorithm couldn't integrate with ASRV (gamma distribution): different choices of among-site rate may result in different reconstruction in internal nodes

- Branch & bound (NP) guarantees global ML, applicable for large # of leaf taxa:
  Early pruning regions of search space, where upper bound of the partial reconstruction (σ) is lower than the best earlier construction (σ^*).
  $C(\sigma)$: set of all extensions of partial construction σ
  o : vector of leaf (OTU) observation sequences
  Estimate upper bound of σ:
  $$B(\sigma) \geq \underset{\sigma' \in C(\sigma)}{max} \ P(\sigma \mid o)$$
  If $\exists \ \sigma^*$ s.t. $B(\sigma^*) > B(\sigma)$, prune σ (ignoring all possible remaining assignments of the partial reconstruction)

  Two types of $B(\sigma)$:

  $$\underset{\sigma' \in C(\sigma)}{max} \ P(\sigma \mid o) \leq \sum_{\sigma' \in C(\sigma)} P(\sigma' \mid o) = P(\sigma \mid o) \tag{1}$$

  $$\underset{\sigma' \in C(\sigma)}{max} \ P(\sigma \mid o) \leq \underset{\sigma' \in C(\sigma)}{max} \sum_r P(\sigma' \mid r, o) P(r) \leq \sum_r \underset{\sigma' in C(\sigma)}{max} \ P(\sigma' \mid r, o) P(r) \tag{2}$$

  With fixed evolution rate r, $max_\sigma P(\sigma \mid r, o)$ calculated in Pupko (2000) DP algorithm. Calculate both bounds, $B(\sigma) = min((1), (2))$

```
function Reconstruct():
    $\sigma^* \leftarrow \{\}$
    $BestScore \leftarrow -\infty$
    DFS({})
    return $\sigma^*$
funcition DFS($\sigma$):
    if $\sigma$ is full reconstruction then
        if $P(\sigma \mid o) > BestScore$ then
            $\sigma^* \leftarrow \sigma$
            $BestScore = P(\sigma \mid o)$
        end
    else
        if $B(\sigma) \leq BestScore$ then
            return
        else
            $H \leftarrow$ HTU $\notin \sigma$
            for $a \in \Sigma$ do
                $\sigma' \leftarrow sigma \cup \{H = a\}$    (extend $\sigma$)
                DFS($\sigma'$)
            end
        end
    end
```

2 Marginal Reconstruction

Koshi, 1996

- Major differences from Felsentein (1981): extend the marginal reconstruction to protein sequences from only nucleotide sequences, with updated empirical matrices of multiple structural associations (e.g. α-helix, folding vs. unfolding). Some matrices are outdated as of today (PAM), while others were tested with tree topology of 3 layers & 4 sequences (not sure how they generated the matrices, but must be an empirical method).

- The post-order traversal is exactly the same as Felsentein's algorithm. The major difference with Pupko (2019): Koshi only exemplified the reconstruction of the root node with probabilistic assignment. Assume root node v with assignment i: $M_i(v) = \pi_i \cdot U_i(v)$ (see Pupko 2019 below). This is a special condition of marginal reconstruction of ancestral nodes. They didn't present the formula of other internal nodes ("This method is easily generalizable to recreate the amino acids at other locations in the phylogenetic tree besides the root" , page 2)

Pupko, 2019: Protein 3D structure to infer ASRV

- AA substitution propensities vary among structural parts of the protein: solvent accessibility

- Le and Gascel (2010): array of replacement matrices → evolutionary coonstraints of different regions of proteins.

- Focus on marginal reconstruction (prob. of each ancestral character assignment in each node → averaged over all possible assignments in other internal nodes of the tree

- (1). Single substitution model: for each amino acid a, assume stationary probability π_a, for each pair of amino acid a, b, $P(a \to b \mid t)$ based on Markov process. Assume E is all the observed leaf node characters, want to calculate posterior probability $P(a \mid E)$ for possible character a at all nodes. DP algorithm to find ancestors in all nodes in $O(n)$:

  - post-order (up) traversal $U(v)$: **node probability given the subtree of node** v:
    v is leaf node: $U_i(v) = \delta_{ij}$ (whether $i = j$ and j is observed character)
    v with child nodes (s_1, s_2): $U_i(v) = \sum\limits_j P(i \to j \mid t_{v,s_1}) U_j(s_1) \cdot \sum\limits_j P(i \to j \mid t_{v,S_2}) U_j(s_2)$
  - pre-order (down) traversal $D(v)$: **node probability given remaining nodes of the tree excluding edge** $predecessor f \to v$:[1]
    v is root node: $D_i(v) = 1 \; \forall i$
    v is internal node ($\nexists g$): $D_i(v) = \sum\limits_k P(i \to k \mid t_{f,b}) U_b(k)$
    Otherwise: $D_i(v) = \sum\limits_j P(i \to j \mid t_{g,f}) D_j(f) \cdot \sum\limits_k P(i \to k \mid t_{f,b}) U_k(b)$
  - "tree traversal" ($M(v)$): **marginal distribution for node** v **given the entire data, combining** U_i **&** D_i:
    v is root node: $M_i(v) = \pi_i \cdot U_i(v)$
    Otherwise: $M_i(v) = \pi_i \cdot U_i(v) \cdot \sum\limits_j P(i \to j \mid t_{f,v}) D_j(v)$

  In general:
  $M_i(v) \iff P(E \mid v = i) P(v = i) \; \forall \; i$ in node V;
  $P(v = i \mid E) = \dfrac{P(E \mid v = i) P(v = i)}{P(E)} = \dfrac{M_i(v)}{\sum_i M_i(v)}$

- (2). Integrate with ASRV
  Extra input: weight of each rate per site
  Assume caregories of n discrete Gamma distribution with shape parameter α, each category with weight $\dfrac{1}{n}$. Let $R = \{r_1, ..., r_n\}$ represents set of all possible rates. Rewrite $P(a \mid E)$ for character a at an arbitrary ancestral node:

  $$P(a \mid E) = \frac{P(E \mid a) P(a)}{P(E)} = \frac{\sum_r P(E \mid a, r) P(a) P(r)}{P(E)} = \frac{\sum_r P(E \mid a, r) \pi_a P(r)}{P(E)}$$

  Calculate the total marginal probabilities of at node v:

  $$P(E \mid a) \pi_a = \sum_r M_{a,r}(v) P(r)$$

  To extend $M_i(v)$ to $M_{i,r}(v)$, update each DP parameter by multiplying branch length with rate: $U_i(v), D_i(v) \to U_{i,r}(v), D_{i,r}(v)$ $(r \in R, i \in \Sigma_{AA})$

- (3). Integrate with "structural-aware" mixture of substitution models
  Extra input: weight of each matrix per site
  Classify each site into either "E" state (exposed) or "B" state (buried) probabilistically with

[1] g: grandfather node; f: father node; b: brother node; $g \to f \to v$ & b

associated weights (e.g. site s_i has solvent accessibility 25% →, assign $\frac{1}{3}$ of B matrix value and $\frac{2}{3}$ of E matrix value). Let $M = \{m_1, ..., m_n\}$ with weights $W = \{w_1, ..., w_n\}$.

For any arbitrary node v, for each matrix $m_i \in M$, calculate $U(v), D(v), M(v)$ for all rates w.r.t. m_i. Assume we calculate the probability of node v having character a:

$$P(E \mid a)\pi_a = \sum_m \sum_r M_{a,r,m}(i) \cdot P(r) \cdot w_m$$

Similar to extension in (2), $M_{i,r}(v), U_{i,r}(v), D_{i,r}(v) \rightarrow M_{i,r,m}(v), U_{i,r,m}(v), D_{i,r,m}(v)$. The calculation step in each node increased from c to $c|R||M|$

Assume $c = 20$, $R = 8$, $M = 2$, N homologous sequences, DP time complexity $O(20N) \rightarrow O(320N) \sim O(N)$ as N increases (still quite efficient compared to the exponential complexity).