Supplemental Material: Variational Combinatorial Sequential Monte Carlo Methods for Bayesian Phylogenetic Inference

Antonio Khalil Moretti[1,*] **Liyi Zhang**[1,*] **Christian A. Naesseth**[1] **Hadiah Venner**[1] **David Blei**[1] **Itsik Pe'er**[1]

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Algorithm 1 Combinatorial Sequential Monte Carlo

Input: $\mathbf{Y} = \{Y_1, \cdots, Y_M\} \in \Omega^{NxM}$, $\theta = (\mathbf{Q}, \{\lambda_i\}_{i=1}^{|E|})$

1: Initialization. $\forall k$, $s_0^k \leftarrow \perp$, $w_0^k \leftarrow 1/K$.
2: **for** $r = 0$ **to** $R = N - 1$ **do**
3: **for** $k = 1$ **to** K **do**
4: RESAMPLE

$$\mathbb{P}(a_{r-1}^k = i) = \frac{w_{r-1}^i}{\sum_{l=1}^K w_{r-1}^l}$$

5: EXTEND PARTIAL STATE

$$s_r^k \sim q(\cdot | s_{r-1}^{a_{r-1}^k})$$

6: COMPUTE WEIGHTS

$$w_r^k = w(s_{r-1}^{a_{r-1}^k}, s_r^k) = \frac{\pi(s_r^k)}{\pi(s_{r-1}^{a_{r-1}^k})} \cdot \frac{\nu^-(s_{r-1}^{a_{r-1}^k})}{q(s_r^k | s_{r-1}^{a_{r-1}^k})}$$

7: **end for**
8: **end for**
9: **Output:** $s_R^{1:K}$, $w_{1:R}^{1:K}$

The proposal distribution for CSMC and approximate posterior for VCSMC can be written explicitly as follows:

$$Q_{\phi,\psi}\left(\mathcal{T}_{1:R}^{1:K}, \mathcal{B}_{1:R}^{1:K}, a_{1:R-1}^{1:K}\right) \coloneqq \left(\prod_{k=1}^K q_\phi(\mathcal{T}_1^k) \cdot q_\psi(\mathcal{B}_1^k)\right) \cdot \prod_{r=2}^R \prod_{k=1}^K \left[\frac{w_{r-1}^{a_{r-1}^k}}{\sum_{l=1}^K w_{r-1}^l} \cdot q_\phi\left(\mathcal{T}_r^k | \mathcal{T}_{r-1}^{a_{r-1}^k}\right) \cdot q_\psi\left(\mathcal{B}_r^k | \mathcal{B}_{r-1}^{a_{r-1}^k}, \mathcal{T}_{r-1}^{a_{r-1}^k}\right)\right]. \tag{1}$$

State $s_r^k = (\mathcal{T}_r^k, \mathcal{B}_r^k)$ is sampled by proposing forest $\mathcal{T}_r^k \sim q_\phi(\cdot | \mathcal{T}_{r-1}^{a_{r-1}^k})$ and branch lengths $\mathcal{B}_r^k \sim q_\psi(\cdot | \mathcal{B}_{r-1}^{a_{r-1}^k}, \mathcal{T}_{r-1}^{a_{r-1}^k})$ from UNIFORM and EXPONENTIAL distributions corresponding to Eq. 1 with ϕ and ψ denoting discrete and continuous terms.

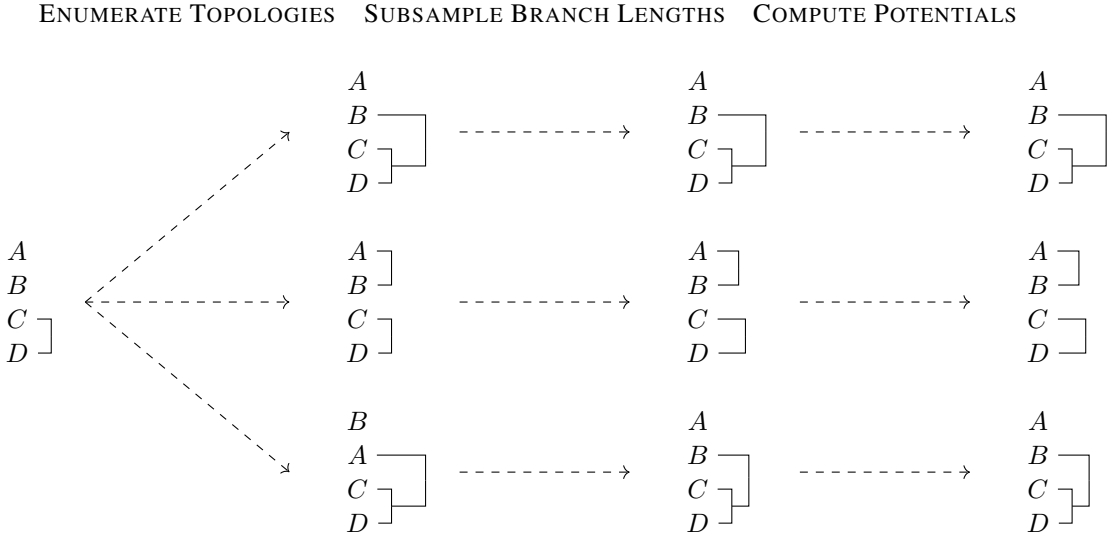

Figure 1: Overview of the NCSMC framework. The enumerated topologies for state $\{A, B, \{C, D\}\}$ are (*top*): $\{A, \{B, \{C, D\}\}\}$, (*center*): $\{\{A, B\}, \{C, D\}\}$ and (*bottom*): $\{B, \{A, \{C, D\}\}\}$. $M = 1$ *sub-branch* lengths are sampled for each edge. *Sub-weights* or *potentials* are computed (right). A single candidate is sampled to form the new partial state.

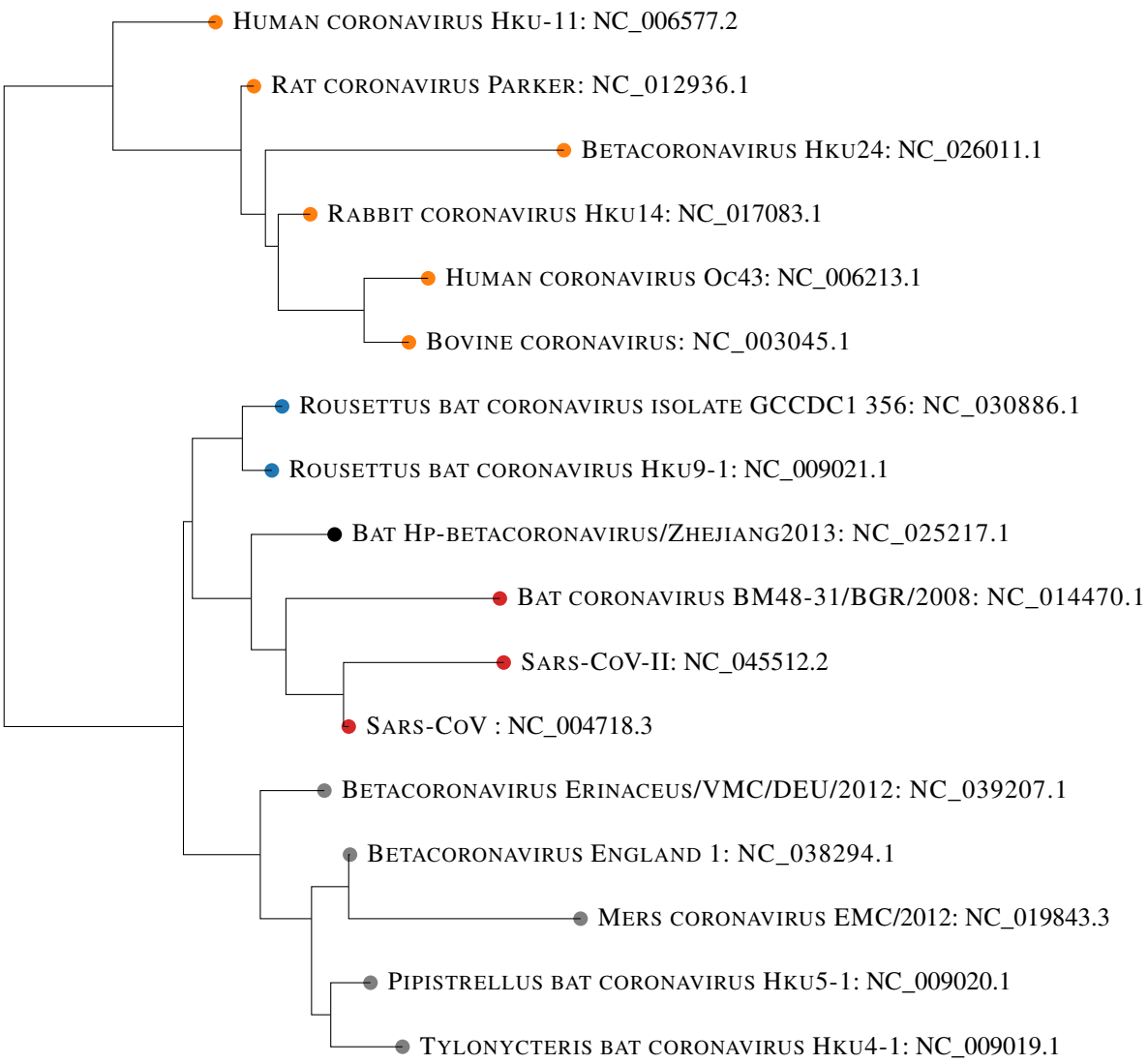

Figure 2: Overview of the betacoronavirus results. The data consists of 17 species of betacoronavirus across 36,889 sites. VNCSMC is run using $K, M = (256, 1)$. A single nonclock phylogeny is chosen based on maximum likelihood and displayed. Colors denote species from the four varying viral lineages: Embecovirus (orange *lineage A*); Nobecovirus (blue *lineage D*); Sarbecovirus (red *lineage B* including SARS-COV and SARS-COV-II); Merbecovirus (grey *lineage C*) and Hibecovirus (black *not classified into the four lineages*) are each partitioned in clades.

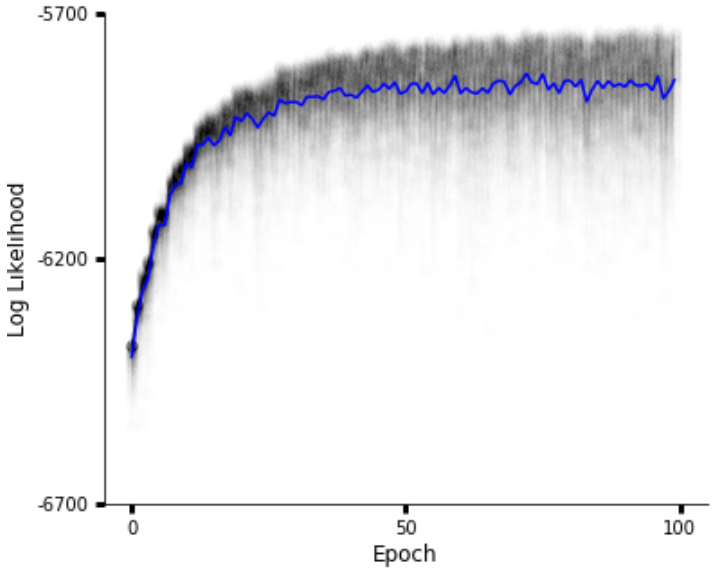

Figure 3: VNCSMC on the primates data with $K, M = (128, 1)$. The full distribution of log likelihood values for all particles across epochs is plotted in black. The average likelihood across samples is plotted in blue.

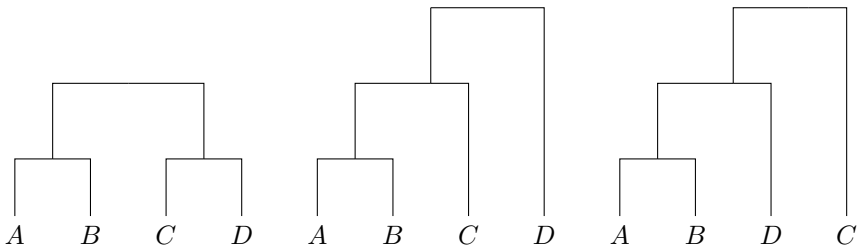

Figure 4: Overview of the dual representation of a partial state. The partial state $s_1^1 = \{P_{AB}, C, D\}$ for four taxa corresponding to Fig. 2 is illustrated using its dual representation $\mathcal{D}(s)$. The dual state $\mathcal{D}(s) \subseteq \mathcal{T}$ corresponds to the three complete tree topologies. (left): $\{\{A, B\}, \{C, D\}\}$ (center): $\{\{A, B\}, \{A, B, C\}\}$ and (right): $\{\{A, B\}, \{A, B, D\}\}$.

	VCSMC				VNCSMC			
K	*s/it*	*s/mit*	*time (minutes)*	ESS	*s/it*	*s/mit*	*time (minutes)*	ESS
4	5.17e-2	1.31e-2	0:22	3.98	4.01	1.17	6:32	3.99
8	5.58e-2	1.42e-2	0:28	7.96	4.27	1.24	7:09	7.88
16	3.11e-2	7.76e-2	0:30	15.79	4.83	1.53	8:15	15.62
32	5.78e-2	2.17e-1	0:49	31.72	5.98	1.59	10:17	31.00
64	9.80e-2	2.66e-1	1:23	62.92	8.33	2.09	14:33	62.59
128	1.35	3.48e-1	2:16	122.79	11.88	2.89	20:02	124.23
256	2.25	5.95e-1	3:52	252.02	21.77	4.98	36:51	252.43

Table 1: Empirical running times of VCSMC and VNCSMC. The Primates data consists of 12 taxa over 898 sites admitting 13,749,310,575 distinct tree topologies. Experiments were performed on a 2.4GHz 8-core intel i9 processor Macbook Pro with 64 GB memory and no GPU utilization. We profile using $K = \{4, 8, 16, 32, 64, 128, 256\}$ and $M = 1$. The left column provides seconds per iteration (*s/it*), the left center column provides seconds per minibatch (*s/mit*), the center right column provides total running time (minutes) across 100 epochs. The effective sample size is provided in the right columns.

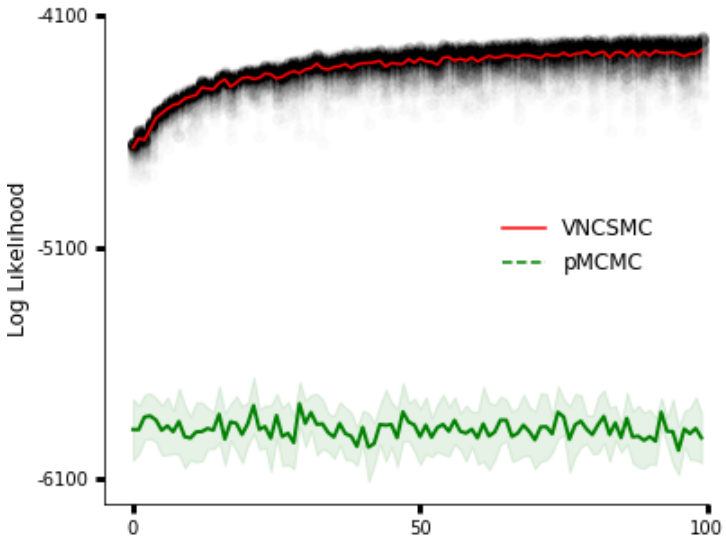

Figure 5: VNCSMC on the 9-taxa subset of primates data with $K, M = (128, 1)$. The full distribution of log likelihood values for all VNCSMC particles across epochs is plotted in black. The average likelihood across samples is plotted in red. Particle Gibbs [Wang and Wang, 2020] is run for 5000 iterations 10 times independently. The last 100 iterations for the 10 independent runs of Particle Gibbs are averaged and plotted in green. VNCSMC using 100 epochs outperforms Particle Gibbs using 5000 iterations.

References

Shijia Wang and Liangliang Wang. Particle Gibbs sampling for Bayesian phylogenetic inference, 2020.