Correctness of ancestral reconstruction

In the following section we will introduce a benchmark metric for ancestral sequence reconstruction, which we call "correctness of ancestral reconstruction" (COAR). The correctness of a reconstruction compared to the true evolutionary history can be measured by multiple similarity measures e.g. topological similarity, branch length similarity and sequence similarity between inferred and real ancestors. All these measures are inter-dependent e.g. the inferred sequences are affected by the branch lengths and the topology and the branch lengths are conditioned on a topology etc. And while inferring correct tree topology is important in its own right, the correctness of the inferred ancestral sequences are the foremost important objective of most BCR phylogenies when these sequences are used for applications involving DNA synthesis, protein expression and functional testing. For this reason, the sole purpose of the COAR metric is to capture the correctness of the inferred ancestral sequences. In particular, we would like to propose a loss function that does not penalize a phylogeny when minor parts of the tree topology is incorrect while ancestral sequence reconstruction is perfect.

The purpose of COAR is to compare two trees built with the same leaves; let us call these the true and inferred tree. When performing ancestral sequence reconstruction the desired result is often to reconstruct the internal nodes in the direct path going from a leaf to the root, as illustrated in Figure 28. This path is extracted by starting at a leaf node and traversing upwards, parent by parent, until the root is reached. In the following, this list of sequences will be referred to as the ancestral lineage. The correct ancestral lineage is the objective of COAR, and we construct the COAR value so it represents the expected per-site error in such a reconstruction. Following the example in Figure 28, often there will be small differences in tree topology between the true and inferred trees, and these will likely make the number of internal states in the ancestral lineages differ. This makes comparison difficult because two lists of different length cannot be element-wise compared. The lists could be made equal length by adding gaps, but then a systematic way of adding these would be necessary.

The basis of COAR is a list comparison progressing element-wise through the list i.e. element 1 in list 1 compared to element 1 in list 2, next, element 2 in list 1 compared to element 2 in list 2 etc. For lists of similar length the list comparison is easy, it will simply be the cumulated distance from list element comparisons, corresponding to the sum of Hamming distances between inferred and true ancestors in the lists. When lists are not equally long, one or more gaps must be introduced into one of the lists; we choose to do so in such a way that the list similarity is maximized. This is an alignment problem with matches/mismatches/gaps and it can be efficiently solved using the Needleman-Wunsch algorithm (13). We define it as a global alignment so that it has to start at the root and end at the leaf because both states are known for the true and inferred phylogenies. We further restrict the Needleman-Wunsch algorithm so that gaps are only allowed to be introduced into the shortest of the two lists being aligned, this forces the maximum number of node comparisons.

One interpretation of the COAR value is that it is the distance between the true and inferred mutation histories, as illustrated in in Figure 29. In this representation of an ancestral lineage the root and the leaf are two fixed states with a continuous mutation process running between them. The internal nodes

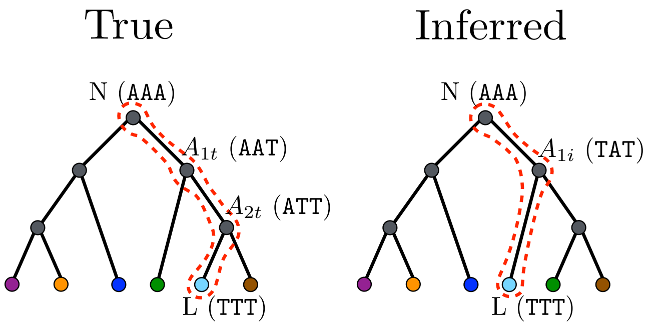

Figure 28: True vs. inferred tree with colored leaves and grey ancestral states. Reconstruction from the light blue leaf is marked by a dashed red line and annotated with genotypes in parenthesis. N is the naive sequence, L is the leaf sequence and the As are ancestors $1, 2, \ldots, n$ with either true or inferred marked by t or i, respectively, appended to the subscript. The inferred tree has misplaced the branch leading to the light blue node, resulting in a missing ancestral sequence.

in the ancestral lineage are discrete states in the continuous process, on the true tree these corresponds to actual cells but on the inferred tree they need not correspond to actual observed genotypes. Instead we can think about them as realizations along the continuous mutation process defined by the inferred tree. The COAR value is then a similarity measured between the true cell genotype and the inferred realizations, each sampled from the true and inferred mutation processes respectively, and in the case of a mismatch between the number of realizations and cells, a gap will be introduced in the alignment to compensate.

Using the aligned ancestral lineages it is now possible to derive a score, similar to a sequence alignment score. We use negative penalties for mismatches and zero points for matches, and furthermore normalize the alignment score to the smallest possible score (all mismatches) for that lineage, yielding the COAR value for a single lineage i:

$$\mathrm{COAR}_i = \frac{\mathrm{alignscore}(\mathrm{leaf}_i)}{\mathrm{alignscore}_{\mathrm{min}}(\mathrm{leaf}_i)}$$

Where alignscore is the score of the alignment between the true and inferred ancestral lineages and alignscore$_{min}$ is the smallest possible score given the number and length of the sequences in the ancestral lineages. The alignment score is defined in terms of penalties, so all values are less than or equal to zero. Since both numerator and denominator are negative the COAR value is positive.

COAR is defined in the range from 0 to 1, where 0 is a perfect ancestral sequence reconstruction and 1 is the worst. The COAR value is comparable across different trees, methods and datasets because of this normalization. Its value can be interpreted as the average per-site error across all the inferred ancestral lineage sequences. COAR for a single ancestral lineage can be expanded to the tree level by calculating the mean COAR value for the whole tree:

$$\mathrm{mean}(\mathrm{COAR}) = \sum_{i=1}^{N_L} \frac{\mathrm{alignscore}(\mathrm{leaf}_i)}{\mathrm{alignscore}_{\mathrm{min}}(\mathrm{leaf}_i)} \Bigg/ N_L$$

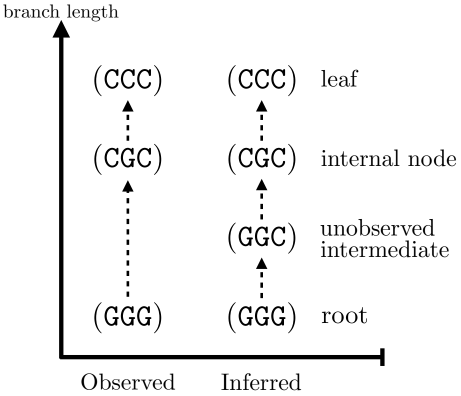

Figure 29: One interpretation of the COAR value is that it is the distance between the true and inferred mutation histories, here shown by the true and inferred ancestral lineage nodes of an example phylogeny. The true ancestral lineage (left side) represents actual observed cells where the genotype is a known constant. The inferred ancestral lineage (right side) represents the estimated genotypes at branching points along the inferred topology. In some cases there is a mis-correspondence between observed cells in the true phylogeny and the branching points in the inferred tree. These are treated as missing realizations and ignored in the alignment of the two mutation histories.

Where N_L is the number of leaves on the tree.

Calculating COAR values - example with a known root

As an example of how the COAR metric works we will present a small example, summarized in Figure 28 with the light blue leaf chosen for lineage reconstruction and the true and inferred ancestral lineages marked in each tree with red dashed lines. The root sequence is a known state called the naive sequence. Assume that the true phylogeny is known with corresponding ancestral sequences. Now take a leaf sequence on the tree and reconstruct its ancestral lineage by extracting the parent, the parent's parent, etc. until the root is reached, tabulated in Table 3. This ordered list of sequences constitute the reconstructed ancestral lineage for the chosen leaf and it always starts at the root and ends at the leaf, therefore we are imposing this as a restriction on the alignment. Furthermore, these two known states they do not count towards the COAR value.

	True	Inferred
Naive (N)	AAA	AAA
A_1	AAT	TAT
A_2	ATT	-
Leaf (L)	TTT	TTT

Table 3: Reconstructed ancestral lineage for true and inferred trees as shown and marked by red dashed line in Figure 28.

In the case of a wrongly inferred topology the true and inferred list of ancestral lineage sequences can have different length. It is therefore necessary to find a way of getting the best possible alignment between these two lists. We know the start and end of this alignment but the sequences in between are free to be shifted up or down to maximize the alignment fit. We adapt the Needleman and Wunsch dynamic program solution (13) to solve this as an alignment problem. A notable difference to the original algorithm is that it was intended to align two sequences of characters, like DNA or amino acids, while in this application a list of whole sequences are aligned.

The first step in the alignment algorithm is to calculate a score matrix of all pairwise sequence comparisons. For this example we use the negative Hamming distance as a score, however, the score function can be extended to reflect different situations, like imposing a larger penalty for non-synonymous rather than synonymous mutations. The score matrix is tabulated in Table 4.

	N	A_{1t}	A_{2t}	L
N	0	-1	-2	-3
A_{1i}	-2	-1	-2	-1
L	-3	-2	-1	0

Table 4: Score matrix based on all pairwise distances between the sequence in Figure 28.

Now we are ready to initializing the alignment grid used in the dynamic programming solution of the alignment problem. Initialization is started by inserting the scores of pure gap characters i.e. first row and first column (Table 5), and we enforce alignment of the two root sequences by setting these gap penalties to negative infinity. Similarly, we disallow introduction of gaps in the longest of the two lists, also by penalizing with negative infinity (Table 6). Setting certain gap penalties to negative infinity is a simple way of dealing with disallowed gaps and it also works well for implementations.

	-	N	A_{1t}	A_{2t}	L
-	0	-Inf	-Inf	-Inf	-Inf
N	-Inf	⇾			
A_{1i}	-Inf				
L	-Inf				

Table 5: The starting alignment grid, initialized with negative infinite gap penalties to disallow gap opening in the beginning of the alignment. The grid is filled up from left to right row by row, starting in the cell marked by ⇾.

Then the alignment grid is filled up, starting with the cell marked by ⇾ in Table 5, progressing to the rightmost cell and continuing in the same fashion on the next row. Cells are filled up using the following maximization:

$$C_{i,j} = \max\left\{(C_{i-1,j} + gp_{\text{down}}); (C_{i,j-1} + gp_{\text{right}}); (C_{i-1,j-1} + S_{i-1,j-1})\right\}$$

Where $C_{i,j}$ is the ith row and jth column cell in the grid, gp_{down} is the penalty of making a downwards gap, gp_{right} is the penalty of making a rightwards gap and $S_{i-1,j-1}$ is the score of aligning the ith, jth elements found by look-up in the score matrix (Table 4) In this example the longest list is that of the true ancestral lineage so in this list gaps are disallowed. In the inferred lineage gaps are allowed but not penalized: $gp_{\text{down}} = -\text{Inf}$ and $gp_{\text{right}} = 0$.

The grid is filled and the final alignment score is the number in the rightmost bottom cell (Table 6).

	-	N	A_{1t}	A_{2t}	L
-	0	-Inf	-Inf	-Inf	-Inf
N	-Inf	0	0	0	0
A_{1i}	-Inf	-Inf	-1	-1	-1
L	-Inf	-Inf	-Inf	-2	-1

Table 6: The filled alignment grid, ready for tracing back the best alignment. The rightmost bottom cell contains the score for the best alignment.

The last step is to traceback the best path through the alignment grid and return this as the list alignment. The traceback starts from the leaf sequence, in the right bottom corner, and ends with the naive sequence in the left top corner. A diagonal step is equivalent to a sequence match, a left move is introducing a gap character in the inferred list and a move up is introducing a gap in the true list. The best path is found by progressively moving upwards, choosing the move with:

$$\text{move}_{i,j} = \text{which}\left\{C_{i,j} = [(C_{i-1,j} + gp_{\text{down}}), (C_{i,j-1} + gp_{\text{right}}), (C_{i-1,j-1} + S_{i-1,j-1})]\right\}$$

Notice that this path has already been generated when the alignment grid was filled up and could be cached. The resulting alignment and the penalty for each position is tabulated in Table 7.

Lastly the alignment score is normalized by the smallest possible alignment score i.e. no similarity between sequences in the lists. This normalized number is the COAR value and it runs between 0 to 1. In the presented example we only calculated the COAR value for the reconstructed ancestral lineage from one leaf, but by doing the calculations on all leaves on the tree and taking the average, the mean COAR value for the whole tree would be computed.

True Inferred	N N	A_{1t} A_{1i}	A_{2t} -	L L
Penalty	0	-1	0	0
Max penalty	0	-3	0	0
COAR	-1/-3=0.333			

Table 7: The resulting alignment and the penalties for each position.

References

[1] Cui A, Di Niro R, Vander Heiden JA, Briggs AW, Adams K, Gilbert T, et al. A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. The Journal of Immunology. 2016;197(9):3566–3574.

[2] Harris TE. The theory of branching processes. Courier Corporation; 2002.

[3] Tas JM, Mesin L, Pasqual G, Targ S, Jacobsen JT, Mano YM, et al. Visualizing antibody affinity maturation in germinal centers. Science. 2016;351(6277):1048–1054.

[4] Kroese F, Timens W, Nieuwenhuis P. Germinal center reaction and B lymphocytes: morphology and function. In: Reaction Patterns of the lymph node. Springer; 1990. p. 103–148.

[5] Childs LM, Baskerville EB, Cobey S. Trade-offs in antibody repertoires to complex antigens. Phil Trans R Soc B. 2015;370(1676):20140245.

[6] Berek C, Milstein C. Mutation drift and repertoire shift in the maturation of the immune response. Immunological reviews. 1987;96(1):23–41.

[7] Kuraoka M, Schmidt AG, Nojima T, Feng F, Watanabe A, Kitamura D, et al. Complex antigens drive permissive clonal selection in germinal centers. Immunity. 2016;44(3):542–552.

[8] Phan TG, Paus D, Chan TD, Turner ML, Nutt SL, Basten A, et al. High affinity germinal center B cells are actively selected into the plasma cell compartment. Journal of Experimental Medicine. 2006;203(11):2419–2424.

[9] Ulrich HD, Mundorff E, Santarsiero BD, Driggers EM, Stevens RC, Schultz PG. The interplay between binding energy and catalysis in the evolution of a catalytic antibody. Nature. 1997;389(6648):271–275.

[10] Rieckmann JC, Geiger R, Hornburg D, Wolf T, Kveler K, Jarrossay D, et al. Social network architecture of human immune cells unveiled by quantitative proteomics. Nature immunology. 2017;18(5):583.

[11] Rieckmann JC, Geiger R, Hornburg D, Wolf T, Kveler K, Jarrossay D, et al. Social network architecture of human immune cells unveiled by quantitative proteomics. Nature Immunology. 2017;.

[12] Romppanen T. A morphometrical method for analyzing germinal centers in the chicken spleen. APMIS. 1981;89(1-6):263–268.

[13] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology. 1970;48(3):443–453.