# Supplementary File 4

# Joint-view trustworthiness (JVT) pseudocode and computational complexity analysis

*Pseudo-code of JVT*

$DS=\{ds\_{i}, ds\_{2}, ... , ds\_{n}\}$, – set of names of the datasets to load. Each $ds\_{i}$ is composed of 4 variables: the corresponding $x\_{i}$, $common\\_gene\\_names\_{i}$, $sample\\_lables\_{i}$, $AUC\\_label\_{i}$.

$x\_{i}$ – the actual dataset within each $ds\_{i}$, where features/genes are on rows and samples are on columns. Each $x\_{i}$ is organized having the same genes and in the same order on the rows

$common\\_gene\\_names\_{i}$ – an array of gene names common to each dataset

$sample\\_lables\_{i}$ – a 2D array where each row corresponds to $x\_{i}$ in each $ds\_{i}$ reporting for each sample the name of one of the two groups to which it belongs in $x\_{i}$

$AUC\\_label\_{i}$ – an array containing the name of one of the two groups of samples that should be used as positive class for AUC-ROC evaluation; each $AUC\\_label\_{i}$ corresponds to each $ds\_{i}$

$CoMa$ – an array containing the names of genes used for computing the combinatorial marker

$al$ – a numerical array of the same length as $CoMa$ containing one of the values in {$1,-1\}$ for each gene, indicating whether alignment is needed $(-1)$ or not $(1)$

$norm$ – type of gene normalization involved. Offered options: ‘*log10*’ and ‘*z-score*’

$T$ – an integer indicating the number of times the samples are drawn to build null model distributions

$Z$ – length of $common\\_gene\\_names$

$C$ – length of $CoMa$

$S$ – size of $DS$

$N\_{i}$ – number of samples of a $ds\_{i}$

$result$ – $2×(C+1)$ matrix, where rows 1 and 2 store minimum AUC score and p-value of minimum AUC score, respectively, as computed by JVT. Each column corresponds to each marker (single+combinatory).

|  |
| --- |
| **Alg1. Joint Multiview Trustworthiness** $JVT$$O(S(ZN+(T+C)(CZ+CN))+C(Z+T+S+C)+TS)$ |

1. $Input=$ datasets $DS=\{ds\_{i}, ds\_{2}, ... , ds\_{n}\}$, a Combinatorial Marker array $CoMa$, an array $al$, normalization type $norm$, sampling times $T$

2. find gene indices, $ind\\_CoMa$, in $CoMa$ via any $ds\_{i}$ using its $x\_{i}$ and $common\\_gene\\_names\_{i}$ // $O(ZC)$

 // *initialisation of variables begins*

3. generate a series $l=1 to Z$ // *used as indices to datasets* ; $O(Z)$

4.$te\leftarrow set\\_difference(l,ind\\_CoMa)$// $te$ *contains* $l$ *with* $ind\\_CoMa$ *removed* ;$O(ZC)$

5. $f\leftarrow length(te)$

6. append $ind\\_CoMa$ back to $l$ // $O(C)$

7. **if** $T < (Z-C)$ **OR** $T > (Z-C)$ **then** // *for randomisation; if* $T=Z-C$*, no need to randomise*

8.$rand\leftarrow $ uniformly generate $T$ random integers $\in \left[1,f\right]$ // $O(T)$

9.$l\leftarrow te(rand)$ // $l$ *is updated and now contains random indices from* $te$ *and has length* $= T$

10. append $ind\\_CoMa$ back to $l$ // $O(C)$

11. **end if**

12.$L=length(l)$ // $L=T+C$

13. initialise $L × S$ zero matrix, $ROC\\_area$, and $(T+1) × S$ zero matrix, $ROC\\_area\\_comb$

 // $ROC\\_area$ *and* $ROC\\_area\\_comb$ *are used to store single and combinatory gene AUC performance,* // *respectively, and have the same number of columns as the no. of datasets*

 // $O(S(T+C))$ *for* $ROC\\_area$ *matrix and* $O(S(T+1))$ *for* $ROC\\_area\\_comb$ *matrix*

// *initialisation of variables ends*

14. **for** $k= 1 to S$ **do** // *for each dataset* ; $O(S(ZN+(T+C)(CZ+CN))$

15. load dataset $ds=DS(k)$ // *this loads all the variables,* $x$*,* $common\\_gene\\_names$*,* //$sample\\_lables$ *and* $AUC\\_label$ *corresponding to dataset,* $ds$

16. perform *log* or *z-score* normalisation to $x$ corresponding to$ds$ based on $norm$
 // *if* $norm$ *does not contain offered options, use log normalisation as default;* $O(ZN)$

17. **for** $i=1 to L$ **do** // *for all sampling times* $T$ *and* $C$

 // $O((T+C)(ZC+CN+C+4N))$ **≈** $O((T+C)(CZ+CN)$)

18.pick the $i^{th}$ index in $l$ and select the corresponding gene, $rand\\_gene$ in $x$

19. $AUC=AUC\\_ROC\left(rand\\_gene,sample\\_lables,AUC\\_label\right)$

// *compute* $AUC\\_ROC$ *of* $rand\\_gene$ *in* $x$ *using the corresponding*
// $sample\\_lables $*and* $AUC\\_label $ ; $O(N)$

20. $ROC\\_area(i,k)=max(AUC , 1-AUC)$ // *for single gene* ;$i= $*row,* $k= $*column*

21. **if** $i\leq (T+1)$ **then** // *for combinatory gene*

22.  **if** $i\leq (T)$ **then**

23. $rand=$ uniformly generate $C$ random integers $\in \left[1,f\right]$ // $O(C)$

24. $R=te(rand)$

// $R$ *contains random indices from* $te$ *and has length* $= C$

25. **elseif** $i=(T+1)$ **then**

26.$R=ind\\_CoMa$ // *no randomisation; only marker indices used*

27. **end if**

28. $xt\leftarrow $ genes identified by $R$ indices in $x$ // $O(ZC)$

29. **for**$ sd=1$ **to** $C$ **do** // *for gene alignment ;* $O(C(N+N))$ ***≈*** $O(CN)$

30.  **if** $al(sd)==-1$ **then**

31. $m=mean(xt(sd))$ // $O(N)$

$ $// *take the mean of the gene identified by its* $sd$ *index*

32. $t(sd))=2×m - xt(sd))$ // *align and update gene* // $O(N)$

33. **end if**

34.  **end for**

35.$xt=column\\_mean(xt)$ //*take the column-wise mean of* $xt$ *genes* ;$O(CN)$

36.$AUC=AUC\\_ROC\left(xt,sample\\_lables,AUC\\_label\right)$

// *compute* $AUC$ *of* $xt$ *using the corresponding* $sample\\_lables $*and* // $AUC\\_label $ ; $O(N)$

37. $ROC\\_area\\_comb(i,k)= max(AUC , 1-AUC)$

38. **end if**

39. **end for**

40. **end for**

41.$ROC\\_area\\_min =$ compute the minimum of $ROC\\_area$ along each row // $O(S(T+C))$

 // $ROC\\_area\\_min$ is a vector containing the minimum AUC performance of single gene
 // across the datasets

42. $ROC\\_area\\_comb\\_min =$ compute minimum of $ROC\\_area\\_comb$ along each row // $O(S(T+1))$

 // $ROC\\_area\\_min$ is a vector containing the minimum AUC performance of combinatorial
 // gene across the datasets

43. initialise $2×(C+1)$ zero matrix, $result$ // *for result compilation* ; $O(2(C+1))$

44. **for** $i=1$ **to** $C$ **do** // *for single marker* ; $O(C(Z+T+C))$

45. $result(1,i)=$ $ROC\\_area\\_min$ value corresponding to the index $ind\\_CoMa(i)$ // $O(Z)$

46. $result(2,i)= $(number of times $ROC\\_area\\_min \geq result(1,i))/L$ // *p-value* ;$O(T+C)$

47. **end for**

48. $result(1,C+1)=$ $ROC\\_area\\_min\\_comb(T+1)$ // *for combinatory marker*

49. $result(2,C+1)= ($number of times $ROC\\_area\\_min \geq result(1,C+1))/(T+1)$

 // *p-value* ;$O(T+1)$

50.append *‘Comb Marker’* to $CoMa$

51. set row names of $result =$ {‘min AUC-ROC’, ‘JVT p-value’}

52. set column names of $result=CoMa$

53. $Output=result$

*Computational complexity of JVT*

We begin with the breakdown of computational complexity of JVT as presented in *Alg1* and then provide the overall associated complexity of JVT. From line 2 in *Alg1*, we can see the complexity of finding $C$ number of marker genes in a set of $Z$ number of common genes is $O(ZC)$. Similarly, lines 3, 4 and 6 have the complexity of $O(Z),$ $O(ZC)$ and $O(C)$, respectively. Line 8, where we generate $T$ random numbers and has the complexity of $O(T)$, where $T$ represents the sampling times, and line 9 runs at $O(C)$. In line no. 13, we create two zero matrices for storing the results, $ROC\\_area$ and $ROC\\_area\\_comb$, having the computational complexity of $O(S(T+C))$ and $O(S(T+1))$, respectively, where $S$ is the number of datasets involved. The loop in line 14 has a total complexity of $O(S(ZN+(T+C)(CZ+CN))$. This complexity can be arrived using lines 15 through 40, which compose the statements of the loop, as follows. Line no. 16, which normalizes each value in a given dataset has a complexity of $O(ZN)$. Line 17, which is also a loop, has a total complexity of $O((T+C)(CZ+CN))$. This can be arrived at by inspecting lines 19, 23, 28, 31, 32, 35 and 36. While, the complexities of these lines are easy to deduce and are given in the algorithm itself, lines 19, 29 and 36 are of particular interest here. Since line 19 computes the $AUC-ROC$ score of the gene with respect to the number of samples the gene carries, $N$, it has the computational complexity of $O(N)$. The same can be said of line 36. Line 19 is a loop and performs the gene alignment with a complexity of $O(CN)$. Putting all this together, we get the computational complexity of line 14 stated above. Furthermore, line 41, which computes the minimum $ROC\\_area$ of all the sampled genes, $T+C$, across all the datasets, $S$, has the complexity of $O(S(T+C))$. Similarly, for line 42, we have $O(S(T+1))$. Line 43 creates a $2×\left(C+1\right)$ zero matrix and intuitively runs at $O(2(C+1))$, and loop in line 44 at $O(C(Z+T+C))$. Finally, line 49 has the computational complexity of $O(T+1)$. Summing up the algorithm’s complexity and simplifying, we get,

$O(S(ZN+(T+C)(CZ+CN))+C(Z+T+S+C)+TS)$.

Clearly, if the values $S$, $N$ and $C$ are highly dominated by the $T$ and $Z$, which is the case in our study, the effect of $S$, $N$ and $C$ in the overall complexity of the JVT becomes negligible. Hence, the complexity boils down to $O(Z+T+ZT) ≈ O(ZT)$, i.e., the JVT’s complexity is non-linear in $Z$ and $T$.