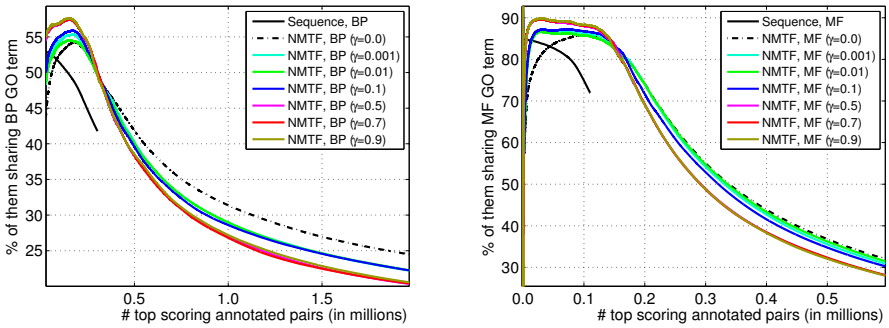
Supplementary Material
Fuse: Multiple Network Alignment via Data Fusion

Vladimir Gligorijević*, Noël Malod-Dognin*, and Nataša Pržulj**

Department of Computing, Imperial College London, United Kingdom.
n.przulj@imperial.ac.uk

1 Study of different values of parameter γ

To determine the influence of the regularization parameter γ, we perform NMTF runs for $\gamma \in \{0.0, 0.001, 0.01, 0.1, 0.5, 0.7, 0.9\}$. For each run, we evaluate the functional consistency of the protein associations that are predicted by NMTF (Fig. 1), as well as the functional consistency of the obtained clusters after the alignment step (Fig. 2). In both cases, we follow the same methodology as in the main document.

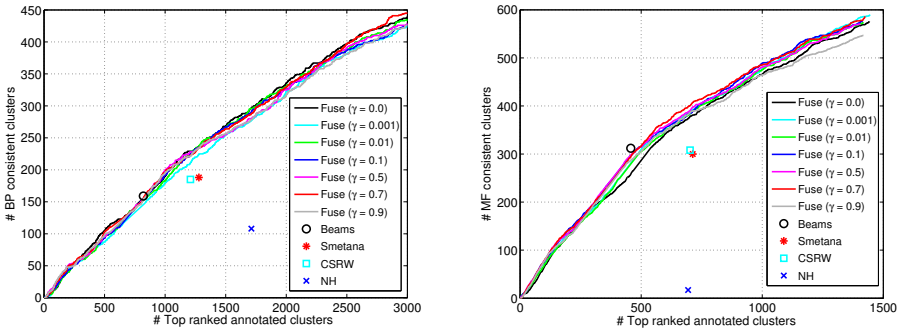


Supplementary Figure 1. Functional consistency of NMTF and sequence-based associations for different values of γ. For all of sequence similarities and NMTF based similarities with different values of γ, we plot the cumulative number of annotated protein pairs (x-axis) against the percentages of them sharing GO terms (y-axis). Biological process (BP) and molecular function (MF) annotations are presented on the left and on the right panels, respectively.

As presented in supplementary Figure 1, NMTF achieves similar functional consistencies for $\gamma \in \{0.5, 0.7, 0.9\}$, with $\gamma = 0.7$ produces the highest number of consistent protein pairs with the top association scores. Moreover, the corresponding curves are above the ones corresponding to $\gamma \in \{0.0, 0.001, 0.01, 0.01\}$; this indicates that a larger influence of PPI networks (measured by parameter γ) results in higher functional consistency of the predicted associations between proteins. Finally, after the alignment step,

* Both authors contributed equally
** Corresponding author.

the highest functional consistency of the produced clusters is achieved for $\gamma = 0.7$ (see supplementary Figure 2). Therefore, in the main document and in the the rest of this supplementary material, we consider the results of Fuse with $\gamma = 0.7$.



Supplementary Figure 2. Functional consistency of Fuse's clusters for different values of γ. For Fuse's protein clusters with different values of γ, we plot the cumulative number of annotated clusters containing proteins from all 5 species (x-axis) against the number of them that are functionally consistent (y-axis). Clusters are ranked according to the scores computed as the sum of association scores between their proteins. We also report the total number of annotated and consistent clusters for the competing aligners (points in the plots). Biological process (BP) and molecular function (MF) annotations are considered separately in the left-hand-side and right-hand-side panels, respectively.

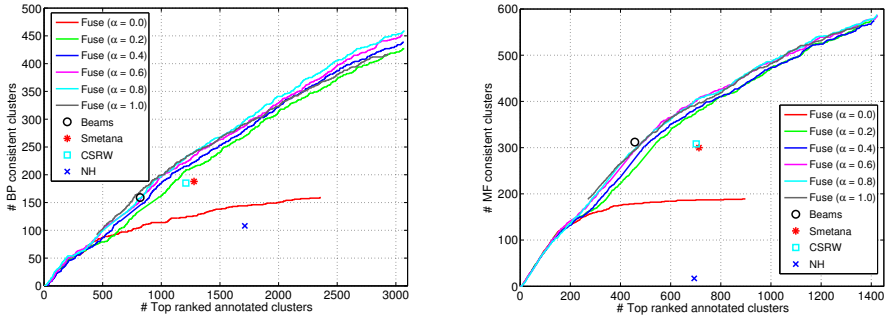
2 Study of different values of parameter α

To estimate the influence of parameter α onto the functional consistency of Fuse's clusters, we evaluate the functional consistency of the obtained clusters after the alignment step, when using $\gamma = 0.7$ and when α varies in $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

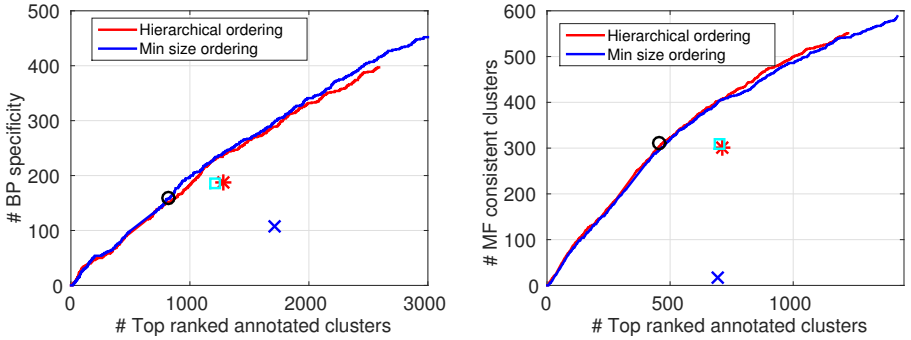
As presented in supplementary Figure 3, Fuse achieves the highest consistency of its clusters for $\alpha = 0.8$. Interestingly, with only the sequence taken into account ($\alpha = 1.0$) our alignment method already outperforms the other aligners.

3 Study of different alignment step orderings

In the alignment step of Fuse, different ordering strategies can be used to merge the networks. We consider two such strategies: 1) "min-size", which merges networks from the smaller towards the larger one; and 2) "Hierarchical", which merges networks according to the phylogenetic tree constructed from the pairwise alignments of the networks.



Supplementary Figure 3. Functional consistency of Fuse's clusters for different values of α. For Fuse's protein clusters with different values of α, we plot the cumulative number of annotated clusters containing proteins from all 5 species (x-axis) against the number of them that are functionally consistent (y-axis). Clusters are ranked according to the scores computed as the sum of association scores between their proteins. We also report the total number of annotated and consistent clusters for the competing aligners (points in the plots). Biological process (BP) and molecular function (MF) annotations are considered separately in the left-hand-side and right-hand-side panels, respectively.



Supplementary Figure 4. Functional consistency of Fuse's clusters when using different alignment step orderings. For Fuse's protein clusters with different ordering strategies, we plot the cumulative number of annotated clusters containing proteins from all 5 species (x-axis) against the number of them that are functionally consistent (y-axis). The clusters are ranked according to the scores computed as the sum of association scores between their proteins. We also report the total number of annotated and consistent clusters for the competing aligners (points in the plots). Biological process (BP) and molecular function (MF) annotations are considered separately in the left-hand-side and right-hand-side panels, respectively.

To evaluate the influence of the ordering strategy, we compare the functional consis-

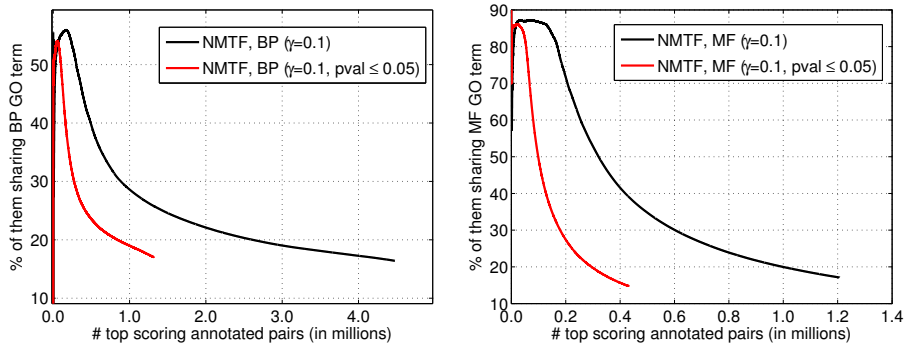
tency of the obtained clusters after the alignment step, when Fuse is set to use either min-size, or hierarchical ordering (γ and α are set to 0.7 and 0.8, respectively).

As presented in supplementary Figure 4, Fuse achieves the highest functional consistency when merging from the smallest network towards the largest one.

4 Permutation test for determining p-values of predicted associations

To determine the statistical significance of NMTF predicted associations, we use the following permutation test. We perform 200 NMTF runs, each time with a different random permutation of relation matrices. In each run, for each pair of species, i and j, we permute the entries in the corresponding relation matrix $\mathbf{R}_{ij}$ while keeping the same distribution of degrees of proteins in species i and j. After factorizing the permuted relation matrices, we infer predictions from each reconstructed matrix $\hat{\mathbf{R}}_{ij}$. We define the p-value of a predicted association as a fraction of NMTF runs in which this association was observed.

As presented in supplementary Figure 5, filtering associations according to p-value results in filtered associations having lower functional consistencies in comparison with the non-filtered ones.



Supplementary Figure 5. Cluster BP (left panel) and MF (right panel) consistency of all NMTF-predicted protein associations (black line) and filtered NMTF-predicted associations with p-value ≤ 0.05 (red line).