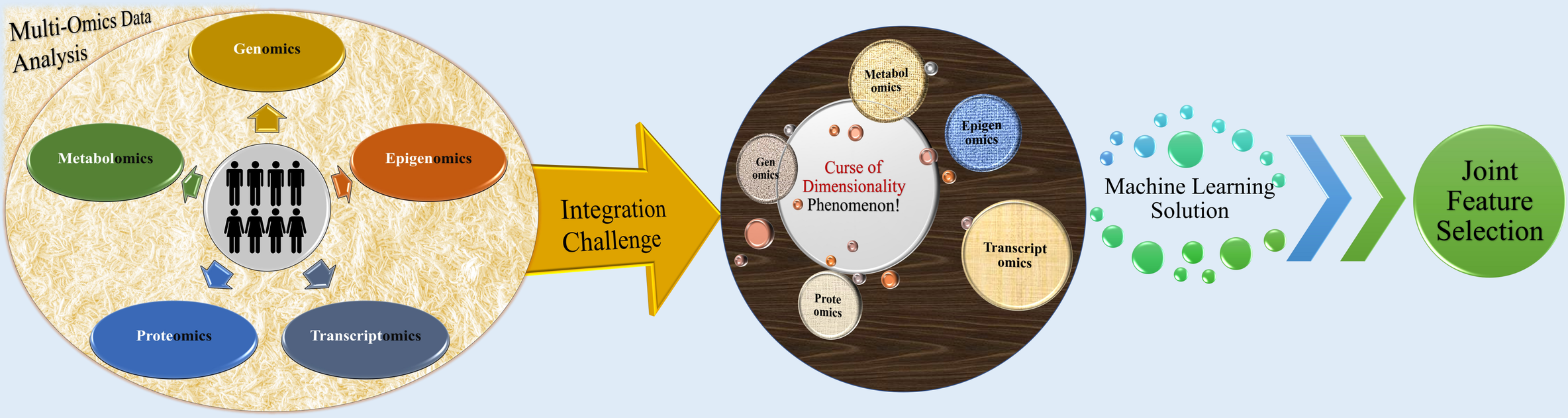
Multi-agent Feature Selection for Integrative Multi-omics Analysis

Author: Sina Tabakhi, Ph.D. Student
Supervisor: Haiping Lu, Senior Lecturer

Machine Learning Research Group

Introduction

* **Motivation:** Diagnose, treat, and cure cancers through the availability of massive biological omics data presented to biologists and data scientists.
* **Aim:** Obtain a deep understanding of complex molecular mechanisms that lead to diseases via multi-omics integration.
* **Challenge:** Mitigate the curse of dimensionality phenomenon which is the consequence of the multi-omics integration task.
* **Solution:** Utilize a feature selection technique to simplify the integration process possessed by high dimensionality datasets.
* **Previous efforts:** Apply feature selection independently to each omics dataset as a preprocessing step which neglects inter-omics interactions.
* **Hypothesis:** Can a joint feature selection for multi-omics data help improve the classification accuracy?



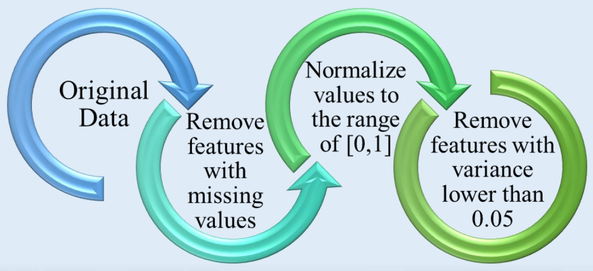
Materials & Methods

Multi-omics Data:

* Public multi-omics datasets such as The Cancer Genome Atlas (**TCGA**) have collected comprehensive profiles of several cancer types for multiple molecular layers. The **ovarian cancer** data from the TCGA are selected to conduct the experiments.

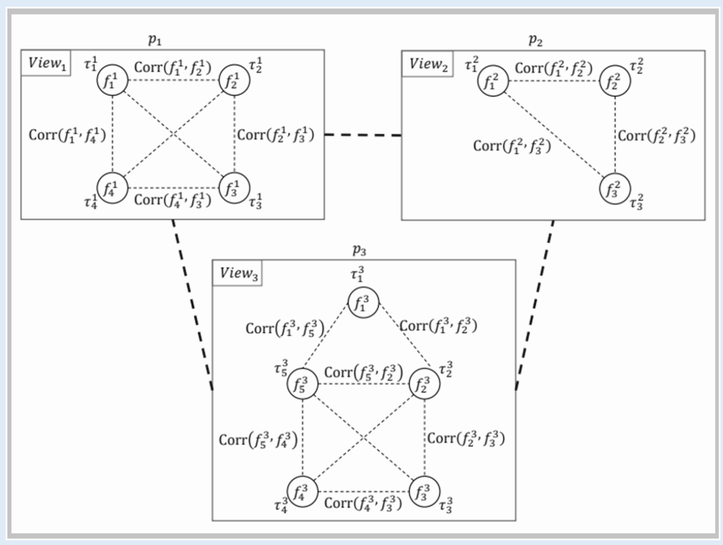
Omics Type	#Features	#Samples
DNA methylation	27,578	616
Gene-level copy number variation	24,776	579
Gene expression RNA-seq	20,530	308

* To ensure the robustness of computation, data have been preprocessed as follows:



Multi-Agent Feature Selection Architecture:

* This study aims to design a multi-agent architecture for **multi-view** (i.e. multi-omics) feature selection to consider different omics data together.
* The search space should be modeled as a suitable **graph** for a multi-agent algorithm before starting the feature selection procedure, illustrated in the following figure.



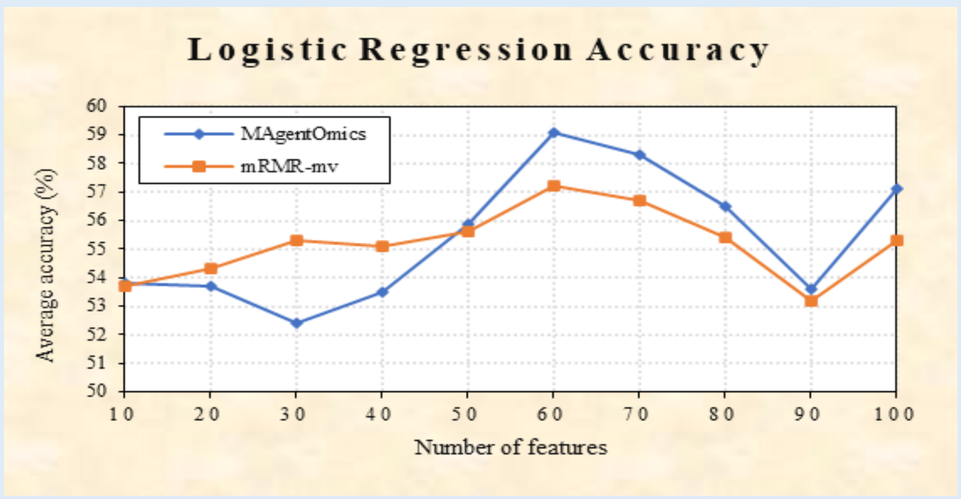
* Below is the proposed multi-agent feature selection **algorithm**.

Input
$\mathbb{D} =< (X^1, X^2, ..., X^v), y >$: multi-view dataset.
N_I: maximum number of iterations.
N_A: number of agents placed in each view.
Output
$\mathbb{D}' =< X', y >$: final single dataset $X', d' \times n$.

Calculate $\text{corr}(f_i^k, f_j^k), \forall k = 1, 2, ..., v$.
Calculate $\text{rel}(f_i^k), \forall k = 1, 2, ..., v$.
$\tau_i^k(0) \leftarrow c, \forall k = 1, 2, ..., v$. ▷ Initialize pheromone
$p_k \leftarrow \frac{1}{v}, \forall k = 1, 2, ..., v$. ▷ Initialize probability
for $t = 1$ to N_I **do**
  for $k = 1$ to v **do**
    Put N_A agents on a randomly chosen node.
  end for
  for $k = 1$ to v **do**
    for $a = 1$ to N_A **do**
      Form new feature subset
      Evaluate the generated subset
    end for
  end for
  Select the current-best solution at t-th iteration.
  Update the pheromone values
  Update probability distribution
end for
Choose the global-best solution found.
Construct $\mathbb{D}'$ based on the global-best solution.

Results

* The performance of the proposed method, **MAgentOmics**, as an unsupervised feature selection method is evaluated in comparison to the **mRMR-mv** [1], which is a supervised multi-view feature selection method.



Conclusion

* Tackled the high-dimensionality challenge of integrative multi-omics analysis via a multi-agent system.
* Assessed the relative importance of each view in the feature selection process.
* Demonstrated the MAgentOmics method outperforms the mRMR-mv supervised feature selection method.

References

[1] Y. El-Manzalawy, T.-Y. Hsieh, M. Shivakumar, D. Kim, and V. Honavar, "Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data," BMC Medical Genomics, vol. 11, no. 3, p. 71, 2018.

Contact

SCAN ME