Metaheuristic Optimizations in Bioinformatics and Machine Learning

Vladimir Filipović
Faculty of Mathematics, University of Belgrade

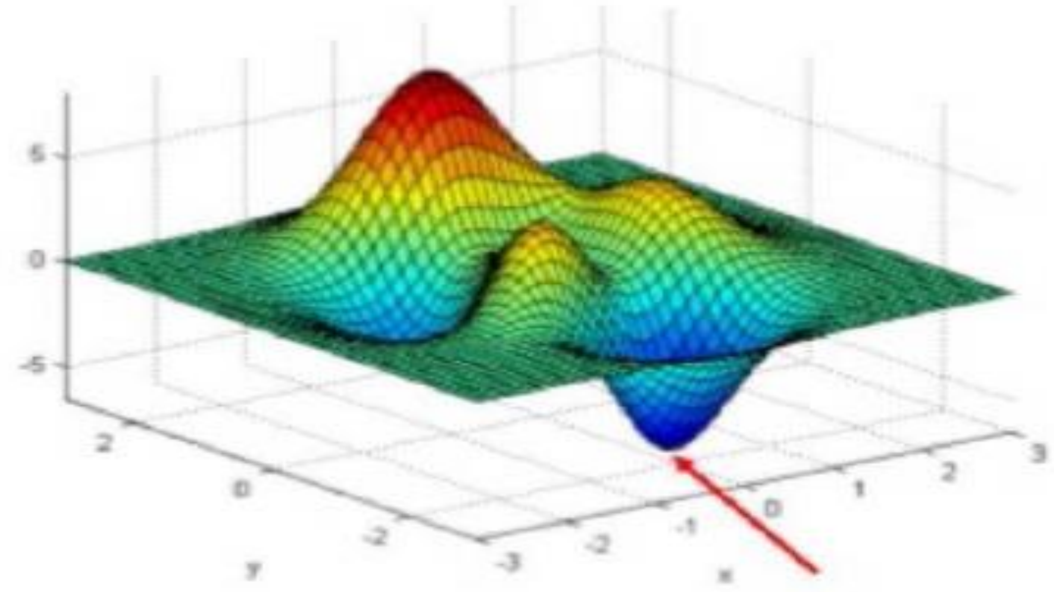
Contents

INTRODUCTION

- Tremendous growth of interest in applications of optimization in bioinformatics and machine learning
- Optimization is frequently used for designing and modeling complex systems
- To address the usage of metaheuristic we focus on three problems:
  - Dimensionality Reduction Problem
  - Maximum Betweeness Problem
  - Maximum Edge k-plex Partitioning Problem

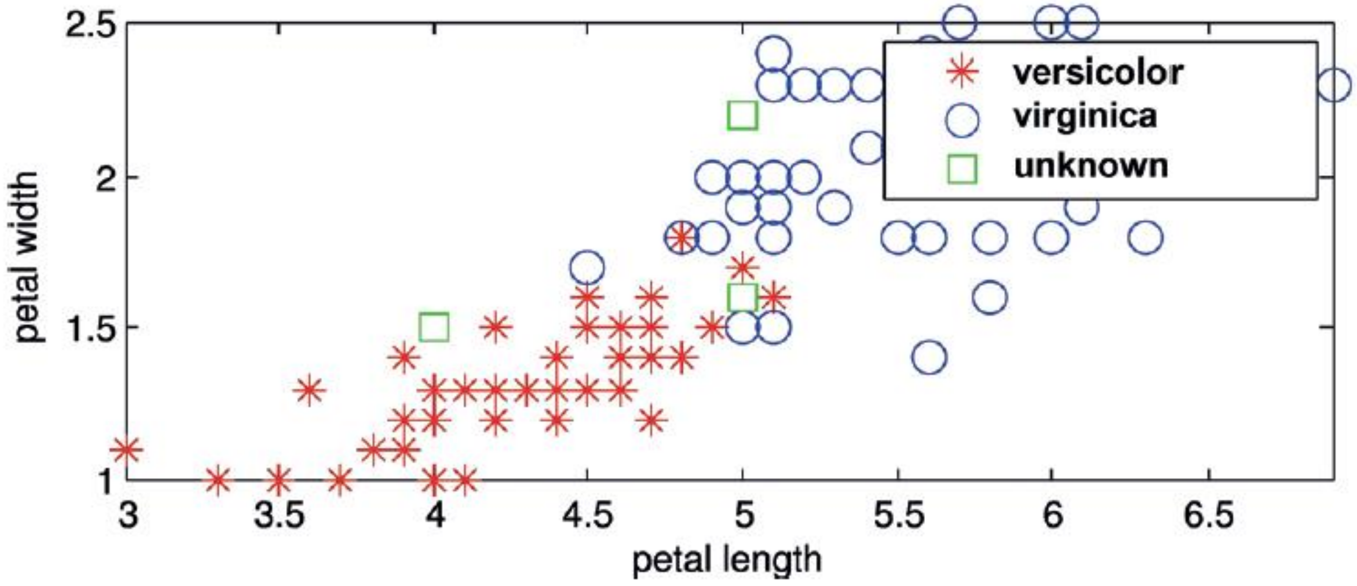
  and on two metaheuristic optimization methods:
  - Electromagnetism-like Metaheuristics
  - Variable Neighborhood Search

OPTIMIZATION PROBLEMS



- Following elements are known:
  - search space S
  - solution space X, $X \subseteq S$
  - objective function f, $f\colon S \longrightarrow R$
- In minimization optimization problems, the goal is to calculate $x^* \in X$, such that $f(x^*) = min\{f(x)|x \in X\}$.

CLASSIFICATION



- Data mining is one of the most popular and exciting discipline of applied informatics
- Data mining includes classification, which predicts a certain outcome based on a given input

DIMENSIONALITY REDUCTION PROBLEM

- Two benefits for applying dimensionality reduction by feature selection for the classification process:
  - By eliminating unnecessary features, it is possible to eliminate dataset noise that degrades the quality of the classification model
  - The problem dimension is decreased and the efficiency is increased
- There are three general types of dimensionality reduction algorithms:
  1. wrapper methods that use a classification algorithm as a black box for the evaluation of feature subsets
  2. filter methods that form feature subsets in the preprocessing phase, and do not depend on the employed classification algorithm
  3. embedded methods that form a feature subset in the training process and are specific to a given classification algorithm
- Wrapper method is considered, where the 1-NN and SVM are used as classification mechanisms

MAXIMUM BETWEENESS PROBLEM

- This problem has important applications in bioinformatics (for solving some physical mapping problems in molecular biology):
  - During radiation hybrid experiments, X-rays are used to fragment chromosomes
  - If the markers are more distant, the probability that the given dose of an X-ray will break a chromosome is greater
  - By estimating the frequency of the breaking points, and thus the distances between markers, it is possible to determine their order within a chromosome in a manner analogous to meiotic mapping
  - Improvement of the radiation experiment can be achieved by finding the total ordering of the markers that maximizes the number of satisfied constraints

MAXIMUM BETWEENESS PROBLEM

- Problem belongs to a class of discrete optimization problems

- Formal definition:

For a given finite set S of n objects $S = \{x_1, x_2, \ldots, x_n\}$

and a given set C of triples $(x_i, x_j, x_k) \in S \times S \times S$

the betweeness problem is a problem of determination of the total ordering of the elements from S, such that triples from C satisfy the "betweeness constraint" (element x_j is between the elements x_i and x_k).

Maximum Betweeness Problem (MBP), deals with finding the total ordering that maximizes the number of satisfied constraints

K-PLEX PARTITION PROBLEM

- Partitioning networks into high density subnetworks, especially cliques, has already been proven as a useful technique for obtaining new information in understanding complicated relations between biological elements
- On the other hand, a number of biological networks classes contain only sparse networks, where partitioning into cliques can be too restrictive method, so many potentially useful information about the interference of biological objects can be neglected
- Here, partitioning is followed by the principle that the objects in each partition are still highly connected in a particular way, but not so restrictively to form a clique
- By relaxing cliques to sparse graphs, biological objects become connected in semantically or functionally logical groups - k-plex

K-PLEX PARTITION PROBLEM

- We deal with the partitioning of the edge-weighted networks into k-plex components, where a subset of some n vertices in a network is a k-plex if the degree of each vertex in the subnetwork induced by this subset is at least $n - k$
- The aim of the Maximum Edge-weight k-plex Partitioning Problem is to find the k-plex partitioning with the maximal total weight of edges
- In the case where k=1, the k-plex is a clique and the Maximum edge-weight k-plex partitioning problem is brought down to the Maximum Edge-weight Clique Partitioning Problem

K-PLEX PARTITION PROBLEM

- Formal definition:

Let a network be denoted as $G = (V, E)$, where $V = \{1,2, \dots, n\}$ is the set of nodes and $E \subseteq V \times V$ is the set of edges.

With uv we simply denote the edge $\{u, v\} \in E$.

With real numbers $w_{uv} > 0$ we denote the weight of the edge connecting nodes u and v. We call u and v the end-vertices of the edge uv.

Let $k \geq 1$be an integer. A set of nodes S is called k-plex if the degree of each node in the subnetwork induced by S is at least $n - k$.

The weight of a k-plex is the sum of all its edge weights.

The weight of the whole partition is the sum of the weights of all its k-plex components.

The Maximum Edge-weight k-plex Partitioning Problem is then defined as finding such a partition of V which is of the maximum total weight and each component is a k-plex

METAHEURISTIC OPTIMIZATION METHODS

- Metaheuristic methods are generalized computational intelligence methods that can be successfully adopted to various problem domains
- They are trying to obtain the optimal solution, or the solution that is close to optimal one
- Metaheuristic algorithms are characterized with approximation and non-determinism
- Basic metaheuristics concepts are abstractly represented - they should be adapted to problem domain, otherwise they should won't obtain enough good solution
- Metaheuristic methods can be population-based (Evolutionary algorithms, Particle Swarm Optimization, Electromagnetism-based Metaheuristics, etc.) or single-solution (Taboo Search, Simulated Annealing, Variable Neighborhood Search etc.)

ELECTROMAGNETISM-LIKE METAHEURISTICS

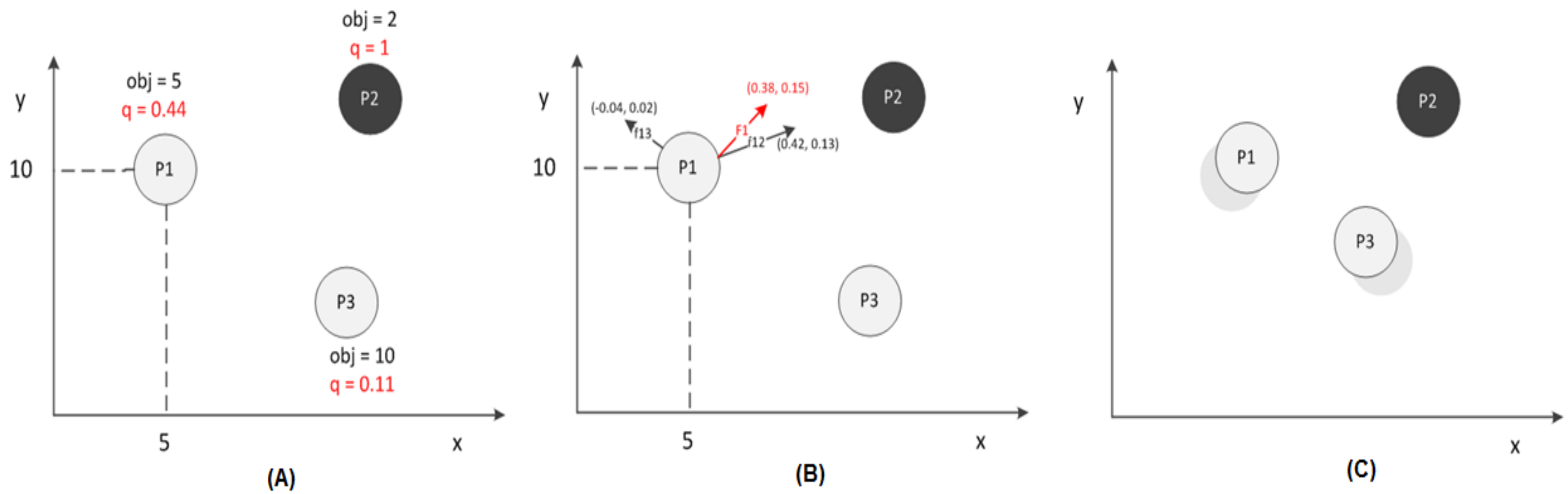
- Electromagnetism-like Metaheuristic (EM) represents a population-based optimization technique inspired by mechanisms of interaction among electrically charged particles (EM points)
- Method, introduced by Birbill and Fang, employs a proficient search process governed by EM points, where each of them represents single candidate solution of the underlying problem
- EM points that represent better solutions are awarded with higher charge. This is crucial for leading the search process towards promising solution regions, because EM points with higher charge attract other points more strongly. The exact attraction-repulsion relationship is given in formula analogues to Coulomb's Law.

EM ALGORITHM

input: N_{it}, M

$\mathbf{p}$ = createInitialPoints(M);
for $iter \leftarrow 1$ **to** N_{it} **do**
    for $i \leftarrow 1$ **to** M **do**
        objFunction($\mathbf{p_i}$);
    end
    charges($\mathbf{p}$);
    forces($\mathbf{p}$);
    relocate($\mathbf{p}$);
end
printSolution();

EM ILLUSTRATION



$$q_i = e^{-N\frac{obj(\boldsymbol{p_{best}})-obj(\boldsymbol{p_i})}{\sum_{k=1}^{M} obj(\boldsymbol{p_{best}})-obj(\boldsymbol{p_k})}}$$

$$\boldsymbol{F_i} = \begin{cases} \sum_{j=1,j\neq i}^{M} (\boldsymbol{p_j} - \boldsymbol{p_i}) \dfrac{q_j \times q_i}{\|\boldsymbol{p_j} - \boldsymbol{p_i}\|^2}, & p_j^{obj} < p_i^{obj} \\ \sum_{j=1,j\neq i}^{M} (\boldsymbol{p_i} - \boldsymbol{p_j}) \dfrac{q_j \times q_i}{\|\boldsymbol{p_j} - \boldsymbol{p_i}\|^2}, & p_j^{obj} \geq p_i^{obj} \end{cases}$$

VARIABLE NEIGHBORHOOD SEARCH

- Variable Neighborhood Search (VNS) method is a robust single-solution metaheuristic, introduced by Mladenović and Hansen
- The main searching principle of a VNS is based on the empirical evidences:
  - multiple local optima are correlated in some sense (usually close to each other)
  - a local optimum found in one neighborhood structure is not necessarily a local optimum for some other neighborhood structure

VNS ALGORITHM

input: n_{min}, n_{max}, it_{max}, $itrep_{max}$, t_{max}, $prob$, k
output: $\mathbf{x}$

$\mathbf{x} \leftarrow$ initializeSolution();
$n \leftarrow n_{min}$;
$it \leftarrow 1$;
while $it < it_{max} \wedge (it - it_{lastimpr}) < itrep_{max} \wedge t_{run} < t_{max}$ **do**
    $\mathbf{x}' \leftarrow$ shaking($\mathbf{x}$,n);
    $\mathbf{x}'' \leftarrow$ localSearch($\mathbf{x}'$, k);
    $move \leftarrow$ shouldMove($\mathbf{x}$, $\mathbf{x}''$, $prob$);
    if $move$ **then**
        $\mathbf{x} \leftarrow \mathbf{x}''$;
    else if $n < n_{max}$ **then**
        $n \leftarrow n + 1$;
    else
        $n \leftarrow n_{min}$;
    $it \leftarrow it + 1$;
end
return $\mathbf{x}$;

VNS ALGORITHM

- The main loop VNS algorithm usually imposes three main procedures: shaking, local search (LS) and neighborhood change.
  - Shaking - in order to escape local suboptimal solutions, a new solution within a parametrized neighborhood of the current best solution is generated.
  - Local search - starting from the new solution obtained in the previous step, other possible solutions within local neighborhood are systematically examined with the aim of finding the local optimum.
  - Neighborhood change - depending on the success of the previous two procedures, the current neighborhood size is adjusted. More precisely, when the current best solution is changed, neighborhood size is reduced to minimal, otherwise it is cyclically increased by 1 (cycle ends at maximal neighborhood size)
- Procedures are iteratively called, until no further improvements of the best solution can be made inside the current neighborhood. When that appears, the algorithm steps into the next neighborhood

Metaheuristic Optimizations in Bioinformatics and Machine Learning

RESULTS

- EM is proposed for Dimensionality Reduction Problem

Detailed elaboration in papers: A. Kartelj "An Improved Electromagnetism-like Method for Feature Selection," Journal of Multiple-Valued Logic and Soft Computing, vol. 25, no. 2, pp. 169-187, 2015. and V. Filipović, "Optimization, classification and dimensionality reduction in biomedicine and bioinformatics," Biologia Serbica, vol. 39, no. 1, pp. 83-98, 2017.

Results and data are publicly available in the GitHub repository https://github.com/vladofilipovic/documents-science-public/tree/main/conferences/canu-2022/suplemental

RESULTS

- EM is porposed for Maximum Betweeness Problem

Detailed elaboration in paper: V. Filipović, A. Kartelj and D. Matić, "An electromagnetism metaheuristic for solving the Maximum Betweenness Problem," Applied Soft Computing, vol. 13, no. 2, pp. 1303-1313, 2013.

Results and data are publicly available in the GitHub repository https://github.com/vladofilipovic/documents-science-public/tree/main/conferences/canu-2022/suplemental

Metaheuristic Optimizations in Bioinformatics and Machine Learning

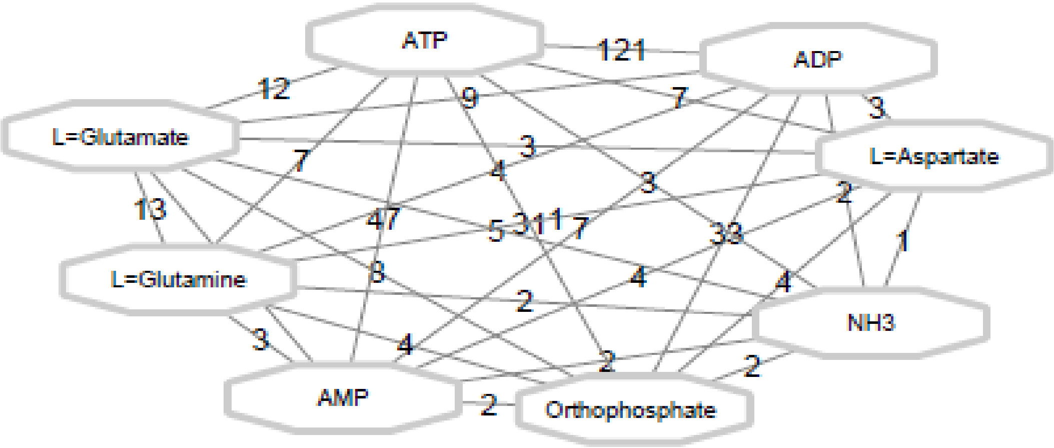
RESULTS

- VNS is proposed for Maximum Edge-weight k-plex Partition Problem

Detailed elaboration in paper: M. Grbić, A. Kartelj, S. Janković, D. Matić and V. Filipović, " Variable Neighborhood Search for Partitioning Sparse Biological Networks into the Maximum Edge-Weighted k-Plexes," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 5, pp. 1822-1831, 2019.

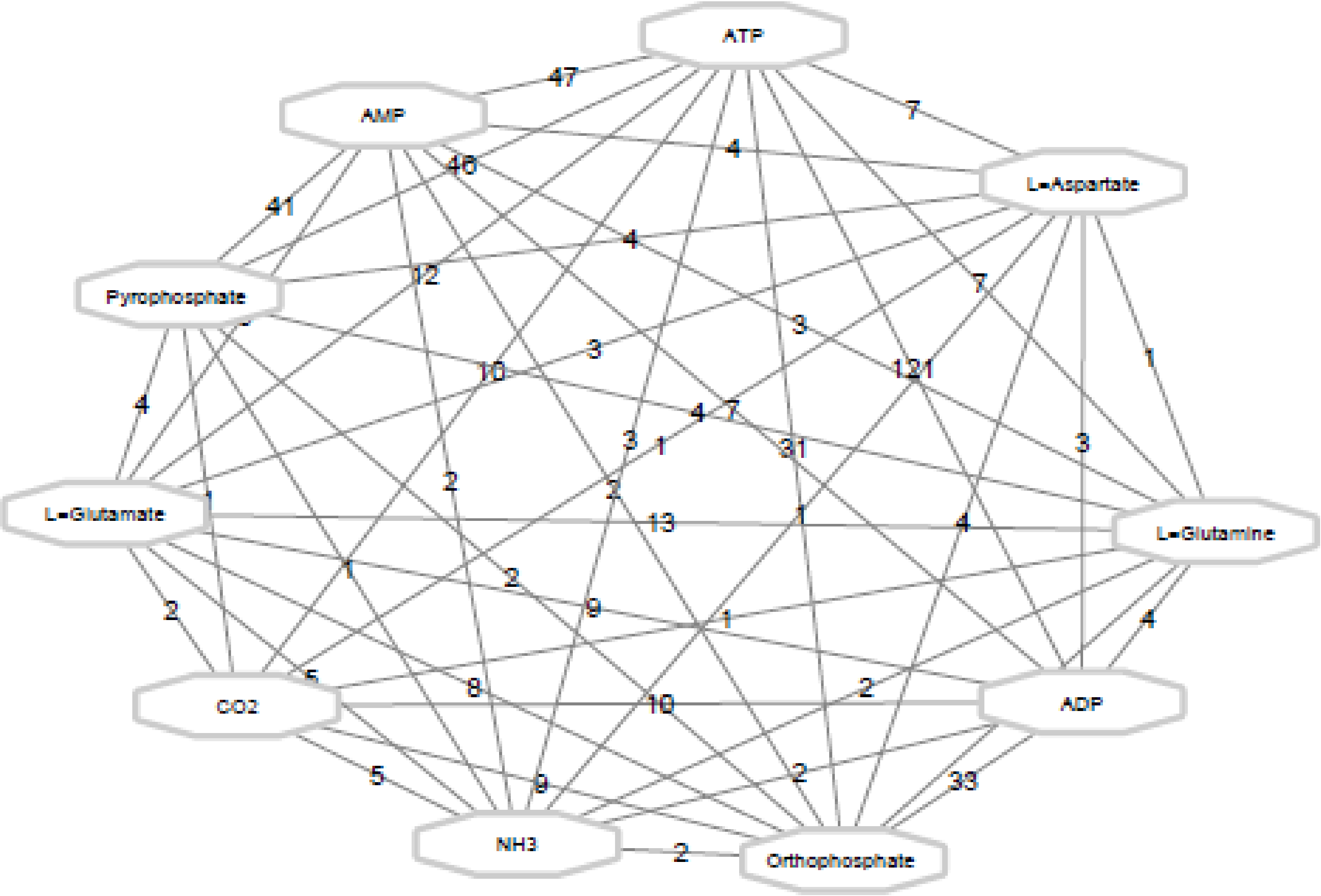
Results and data are publicly available in the GitHub repository https://github.com/vladofilipovic/documents-science-public/tree/main/conferences/canu-2022/suplemental

RESULTS



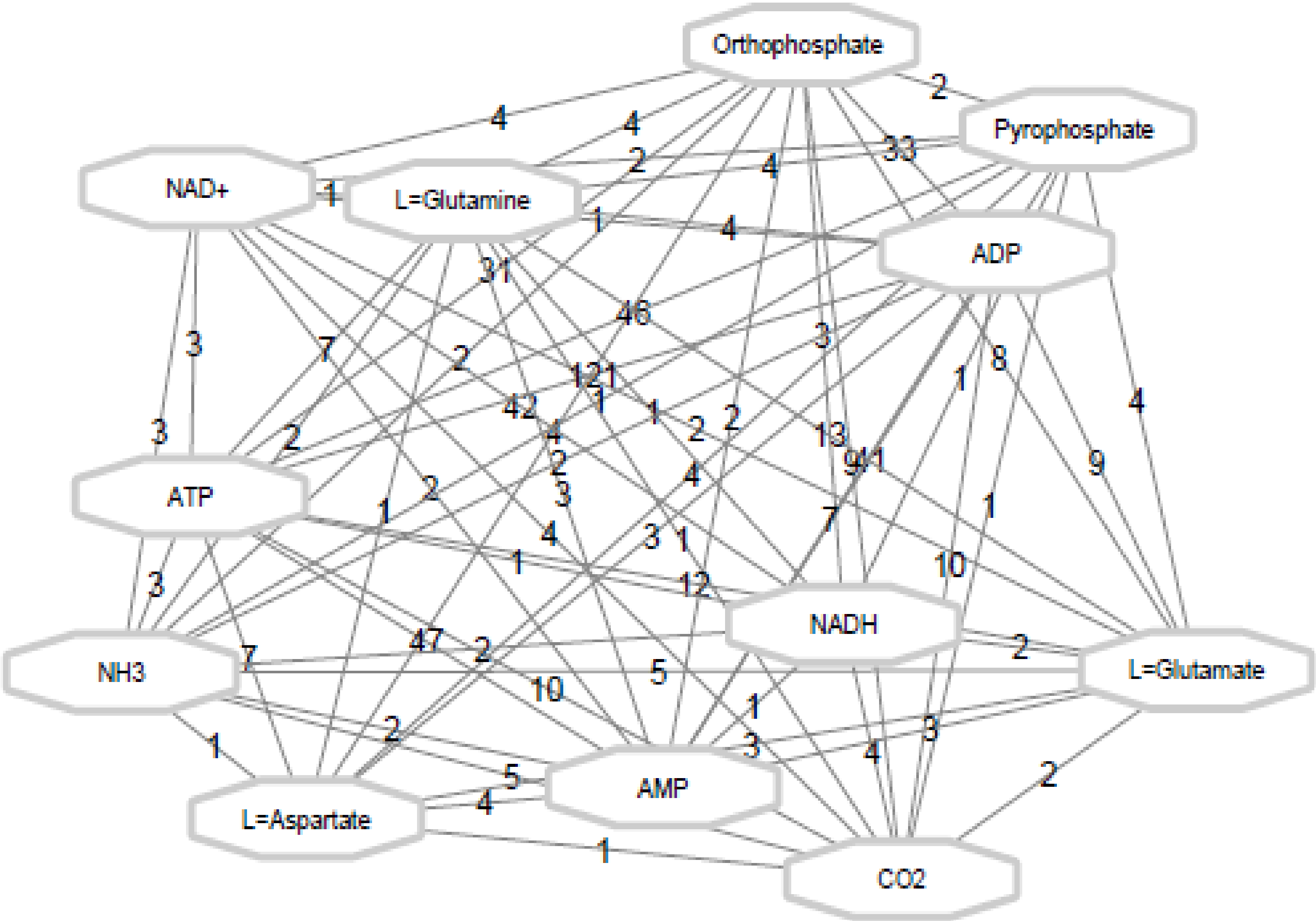
ATP
ADP
L=Glutamate
L=Aspartate
L=Glutamine
NH3
AMP
Orthophosphate

RESULTS


ATP
AMP
L-Aspartate
Pyrophosphate
L-Glutamate
L-Glutamine
CO2
ADP
NH3
Orthophosphate

Metaheuristic Optimizations in Bioinformatics and Machine Learning

RESULTS



TOPOLOGY SENSITIVE METAHEURISTICS

- The main motivation for integrating topology and metaheuristics comes from the notion that metaheuristics might use the topological regularities inside the solution space to better maneuver through it
- This can become especially useful when the solution space becomes extremely large.
  - In such situation, classical metaheuristics might use too much resources in order to search the solution space.
  - Although this sounds like it could lead to premature convergence to local optima, we stress that our conceptual design essentially generalizes and encompasses the classical metaheuristic algorithms.
- Proposed metaheuristics, during its execution, will gradually converge to its classical variants
- Fitness landscape analysis, which includes analysis of local optima positions, is very important for design of such metaheuristics

TOPOLOGY SENSITIVE METAHEURISTICS

- If some topological regularity in fitness landscape is detected, that regularity can be exploited and used for designing metaheuristic that will perform better than the alternatives
- Topology-based models and techniques already achieved good results in revealing hidden structures and detecting new regularities, so it can be expected that it will be helpful in this domain
- The most important topology (more precisely, algebraic topology) concepts in this domain are simplicial complexes, homology groups and persistent homologies.
- In-depth discussion of the concepts described in this section are given in the paper: A. Kartelj, V. Filipović, S. Vrećica and R. Živaljević, "Topologically sensitive metaheuristics," arXiv, 2020.

TOPOLOGY SENSITIVE EM

- TEM is designed as a generalization of EM that builds on m-simplex data, where for special case m=0 that algorithm becomes a classical EM
- The main difference in TEM, in comparison to classical EM, is in the movement step
  - For each solution point, within TEM, we try to find new solution position inside the solution space that will form a m-simplex with other m solution points from the current population
  - In EM, the movement is controlled partially by forces and partially by the randomness. In TEM, the movement is also controlled by forces, but now the randomness is restricted with respect to parameter m
  - This means that for $m>0$, the set of possible positions from which the new position is randomly chosen now becomes smaller
  - If, for a given solution point and current simplex size m, the movement is not possible, the simplex size is reduced by 1.

TOPOLOGY SENSITIVE EM

- The overall effect that we expect TEM will have on the search process in comparison to EM is increased preservation of the same or similar topological regularity through time (if this regularity exists)
- We also think that the expansion of already existing simplices, especially large ones, is well motivated - because the existence of regular formation of local optima itself is an indicator that more new local optima may be found around that formation
- Another important observation is that since TEM falls back to classical EM, we can expect that TEM will be generally applicable, i.e. if the topological regularity is low and cannot be exploited, TEM should work at least as good as classical EM (though performance might get deteriorated)

TOPOLOGY SENSITIVE VNS

- TVNS is essentially conceived as a generalization of VNS that builds on m-simplex data (with special case m=0 being a classical VNS)
- We will also sometimes refer to (m+1)-simplex neighborhood which corresponds to collection of all valid simplices that can be formed by adding 0-simplex to observed m-simplex
- Therefore, 1-simplex neighborhood correspond to classical VNS, while m-simplex neighborhoods where m>0 refer to its topological generalizations

TOPOLOGY SENSITIVE VNS

- The main loop of TVNS should be made in such a way that the sequence of neighborhood structures, that are now parametrized by m and k, starts with the most restrictive neighborhood and after that proceeds with the sequence of more relaxed ones
- Therefore, the neighborhoods will start with smallest neighborhood size k=k_min and the most restrictive simplex structure m=m_max, and further proceed with reduction of m by 1
- When m reaches 0, it basically means that classical VNS algorithm is to be performed
- After that, the k is increased by 1 and m is reset to m_max
- The full cycle through neighborhoods is done when k reaches k_max and m reaches 0. If, at some moment, current solution is improved, both k and m are reset to its initial values.

CONCLUSIONS

1. Two metaheuristic optimization methods aimed at solving specific problems in bioinformatics and machine learning are described and the obtained results are analyzed.
2. Topological enhancement for those methods are proposed.
3. Further research will be focused on theoretical characteristics of the proposed enchantments, on design and execution of computational experiments.

BIBLIOGRAPHY

V. Filipović, A. Kartelj and D. Matić, "An electromagnetism metaheuristic for solving the Maximum Betweenness Problem," Applied Soft Computing, vol. 13, no. 2, pp. 1303-1313, 2013.

M. Grbić, A. Kartelj, S. Janković, D. Matić and V. Filipović, " Variable Neighborhood Search for Partitioning Sparse Biological Networks into the Maximum Edge-Weighted k-Plexes," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 5, pp. 1822-1831, 2019

V. Filipović, "Optimization, classification and dimensionality reduction in biomedicine and bioinformatics," Biologia Serbica, vol. 39, no. 1, pp. 83-98, 2017.

A. Kartelj, "An Improved Electromagnetism-like Method for Feature Selection," Journal of Multiple-Valued Logic and Soft Computing, vol. 25, no. 2, pp. 169-187, 2015.

A. Kartelj, N. Mitić, V. Filipović and D. Tošić, "Electromagnetism-like Algorithm for Support Vector Machine Parameter Tuning," Soft Computing, vol. 18, no. 10, pp. 1985-1998, 2013.

A. Kartelj, V. Filipović, S. Vrećica and R. Živaljević, "Topologically sensitive metaheuristics," arXiv, 2020.

Thanks for your attention ☺

vladaf@matf.bg.ac.rs

vladofilipovic@hotmail.com