Parallel Exchange of Randomized SubGraphs for Optimization of Network Alignment: PERSONA

Erhun Giray Tuncay, Rıza Cenk Erdur, and Tim Conrad

Abstract—The aim of Network Alignment in Protein-Protein Interaction Networks is discovering functionally similar regions between compared organisms. One major compromise for solving a network alignment problem is the trade-off among multiple similarity objectives while applying an alignment strategy. An alignment may lose its biological relevance while favoring certain objectives upon others due to the actual relevance of unfavored objectives. One possible solution for solving this issue may be blending the stronger aspects of various alignment strategies until achieving mature solutions. This study proposes a parallel approach called PERSONA that allows aligners to share their partial solutions continuously while they progress. All these aligners pursue their particular heuristics as part of a particle swarm that searches for multi-objective solutions of the same alignment problem in a reactive actor environment. The actors use the stronger portion of a solution as a subgraph that they receive from leading or other actors and send their own stronger subgraphs back upon evaluation of those partial solutions. Moreover, the individual heuristics of each actor takes randomized parameter values at each cycle of parallel execution so that the problem search space can thoroughly be investigated. The results achieved with PERSONA are remarkably optimized and balanced for both topological and node similarity objectives.

Index Terms—Global network alignment, protein-protein interaction networks, actor systems, particle swarm optimization

1 Introduction

Network Alignment generates node mappings between networks of organisms in question in order to compare them functionally. Alignment results can be used in various areas such as predicting functions of unannotated proteins, revealing mechanisms of certain diseases and reproducing a rooted phylogenetic tree based on the discovered evolutionarily conserved pathways or protein complexes and detected functional orthologs across species [1]. Most Global Network Alignment algorithms rely upon the assumption that the functions of smaller networks map one-to-one to the functions of bigger networks homologously unlike most Local Network Alignment algorithms that focus on overlapping highly conserved subnetworks by allowing many-to-many node mappings [2], [3], [4].

One of the most significant drawbacks of the existing one-to-one Global Network Alignment algorithms is the lack of homogeneity across the mappings of an alignment in terms of quality. This problem arises since the main heuristic of an alignment algorithm is applicable only for a certain proportion of mappings. This means it may be beneficial to alter the alignment heuristic during the progression of an alignment process according to its current performance and it is worth evaluating the contributions of every new set of mappings to an alignment individually. Such an interactive approach can be achieved by means of a querying mechanism capable of classifying the contributions of a particular set of mappings in terms of various topological similarity and node similarity metrics as well as various alignment heuristics that prioritize different metrics. The interactivity can further be extended with a collaborative infrastructure that orchestrates a population of aligners and enables exchanging significant mappings among population members that are strong in different metrics.

This study proposes a hybrid approach that combines several fundamental alignment heuristics and meta-heuristic search tools to design a custom multi-objective optimization workflow and a population of custom designed aligners that interact with each other collaboratively for solving the Global Network Alignment problem against multiple objectives. The most significant traits of PERSONA are adaptation to new data sets, adjustment among objectives with high precision and providing mature alignments that are balanced with respect to all objectives. This paper is structured as follows: First, globally recognized performance objectives of the Global Network Alignment problem are introduced. Later on, the multi-objective optimization approach of PERSONA is introduced by explaining the role of employing various heuristics in the process, the execution steps of the algorithm and the architecture of the

• *Erhun Giray Tuncay is with Freie Universität Berlin, 14195 Berlin, Germany, and also with the TIB - Leibniz Information Centre for Science & Technology, 30167 Hannover, Germany. E-mail: giray.tuncay@tib.eu.*
• *Rıza Cenk Erdur is with Ege University, 35040 İzmir, Turkey. E-mail: cenk.erdur@ege.edu.tr.*
• *Tim Conrad is with Zuse Institute Berlin, 14195 Berlin, Germany, and also with Freie Universität Berlin, 14195 Berlin, Germany. E-mail: conrad@zib.de.*

Manuscript received 1 January 2021; revised 10 December 2022; accepted 12 December 2022. Date of publication 22 December 2022; date of current version 5 June 2023.
(Corresponding author: Erhun Giray Tuncay.)
This article has supplementary downloadable material available at https://doi.org/10.1109/TCBB.2022.3231489, provided by the authors.
Digital Object Identifier no. 10.1109/TCBB.2022.3231489

framework. Following the methodology, performances of state-of-the-art multi-objective aligners are compared with respect to various data sets. Finally, the results are discussed and the achieved highlights were summarized as a conclusion.

2 Method

The PERSONA methodology is designed to succeed in multiple objectives that make most sense for a global Protein-Protein Interaction Network comparison. However, it is not currently possible to make a prioritization among these objectives or identify robust aggregate objectives out of the existing ones. For this reason, we represent the Global Network Alignment problem as a multi-objective optimization problem in order to achieve optimal performance in each of these objectives. In this context, we propose the collaborative methodology that depends on exchanging essential subgraph information throughout this chapter. Besides, we also introduce a heuristics suite enabling to design a custom singular alignment to generate multiple aligners with different characteristics. Subsequently, we explain the graph specific persistency infrastructure that stores the essential network characteristics and alignments in progression as well as the concurrent computation architecture that handles the collaboration tasks among the entities of the whole system for implementing the methodology.

2.1 Performance Objectives

One major compromise for solving a network alignment problem is the trade-off among multiple alignment objectives that evaluate alignment performance. Alignment performance can be computed with respect to topological similarity and node similarity metrics. Topological similarity objectives aim to address the functional similarities of the organisms in terms of their network structure and interaction patterns and evaluates node pairs that contribute to similar interaction patterns from both of the organisms to be aligned. On the other hand, node similarity objectives aim to address the functional similarity of node pairs from both of the compared organisms individually without considering their network structures. In this study, alignment quality is evaluated with several well known topological and node similarity objectives used in other Global Network Alignment studies [5], [6], [7], [8], [9] as summarized below:

Edge Consistency (EC) is an early topological similarity objective computed as the ratio of the edges in the smaller network mapped to the edges in the bigger network to all the edges in the smaller network. Mathematically $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ represents two graphs or networks while $g : V_1 \rightarrow V_2$ represents the alignment between two networks. Therefore, the mathematical representation of EC [10] becomes:

$$\frac{|\{(u,v) \in E_1 : (g(u), g(v)) \in E_2\}|}{|E_1|} \times 100. \tag{1}$$

Symmetric Sub Structure (S^3) is another topological similarity objective derived from EC that penalizes unaligned edges in both the larger and the smaller network. In S^3 G[V] is represented as an induced subnetwork of G with the node set V, while E(G) is represented as the edge set of G. If we denote $f(E_1) = \{(g(u), g(v)) \in E_2 : (u,v) \in E_1\}$ and $f(V_1) = \{g(v) \in V_2 : v \in V_1\}$ then S^3 [11] becomes:

$$\frac{|f(E_1)|}{|E_1| + |E(G_2[f(V_1)])| - |f(E_1)|}. \tag{2}$$

Largest Common Connected SubGraph (LCCS) is the number of aligned node pairs in the largest connected subgraphs of an alignment that exist in both networks with an exact copy. The necessity for this measure is due to the significance of aligning large and contiguous subgraphs rather than a number of small disconnected network regions that would not give an equivalent insight into common topology of two networks [12]. For computing the aligned LCCS, first LCCS is calculated for the aligned nodes of both networks seperately and then their intersection is computed by considering the member node pairs. The ratio of these node pairs to the size of the whole alignment gives the corresponding normalized score for this objective.

Gene Ontology Consistency (GOC) is a node similarity objective in which every aligned protein pair contributes to the overall GOC score by adding it the ratio of the size of the intersection of their common GO [13] terms to the size of the union of their GO terms. The overall GOC score is computed by summing up these ratios coming from all aligned protein pairs. The Jaccard-Index based formula below represents the contribution of a single aligned node pair to the overall GOC score [14]:

$$GOC(u,v) = \frac{GO(u) \cap GO(v)}{GO(u) \cup GO(v)}. \tag{3}$$

Gene Ontology Enrichment (GOE) is the percentage or number of aligned protein pairs that share at least one GO term annotation. This metric is a node similarity objective that is more adapted to the Global Network Alignment problem [15].

Resnik Similarity is one of the most popular semantic similarity measures [16]. The method computes the similarity between two terms as the Information Content of the Most Informative Common Ancestor in a particular Annotation Corpus [17]. The Information Content of a concept can be considered as its likelihood particularly in the Gene Ontology Annotation Database based on the annotation frequency [17], [18]. Resnik method is a pairwise term semantic similarity measure that is not directly applicable to genes and proteins and it can be adapted to proteins with a mixing strategy based on the average or maximum of all term pairwise similarities as well as the average of similarity between best matching terms [16].

Biological Sequence Similarity (BS) is a node similarity measure calculated by summing up the BLAST [19] bit scores or the e-values of the aligned protein pairs [8], [20]. The biological sequence similarity for all the possible node pairs from both of the aligned organisms can be measured with BLASTP [21], [22] application using the corresponding FASTA [23] texts of the proteins with user defined values for parameters such as word-size and e-value.

The exact biological relevance of all these objectives are yet to be discovered. For this reason, they can not precisely be prioritized among each other and it becomes mandatory to review all objective scores as a whole to get a better meaning. Nevertheless, it is possible to establish a particular aggregate objective function that evaluates multiple objectives intuitively in order to yield decisive results. We used such intuitive objective functions in Section 3.4 for efficient evaluation of our experiments.

2.2 Collaborative Method

Global Network Alignment is inherently a multi-objective optimization problem since the prioritization of the objectives among each other is a grey area despite certain definitions that highlight the prior significance of node similarity objectives over topological ones due to the verifiable evolutionary conservation information [24]. Since there is no obvious prioritization among objectives, it makes sense to preserve a set or population of alignments that are stronger in different objectives and able to learn from each other. PERSONA is inspired by the concept of Particle Swarm Optimization [25] as a meta-heuristics approach that intends collaboration among a population of aligners to optimize multiple objectives.

The original definition of Particle Swarm Optimization problem needs to be revised by concepts of multi-objective optimization as well as domain-specific properties for defining Global Network Alignment as a multi-objective problem. The most essential multi-objective optimization definition required for this revision is that a sample solution of a problem is considered pareto optimal [26] or non-dominated if improving the score of one of the objectives would require worsening other objective scores in the current solution space. Additionally, the concept of pareto dominance is also important for comparing two solutions. According to this definition, solution x pareto dominates another solution y, if it is strictly better than y with respect to at least one objective and is at least equivalent to y with respect to the remaining objectives [8]. Based on these definitions, a possible domain-specific revision is defining multiple swarm leaders out of the non-dominated solutions to map the fair number of problem objectives. Yet, it is possible that excessive number of non-dominated solutions exist and some of them are repetitive since they reside within the same neighborhood or in close distance. As a result, a leader selection methodology becomes an essential part of the process [27]. The selected leaders among the non-dominated solutions heavily effect the convergence rate and diversity of solutions. On the other hand, some algorithms restrict the number of stored non-dominated solutions by filtering a relatively useless set of them in order to improve performance. This is not an easy task since an exact quantification is not possible among multiple objectives and it may lead to a compromise in the diversity of the population [28].

There is a need to establish a robust trade-off mechanism in Particle Swarm Optimization in order to achieve a balanced pareto optimality among objectives while preserving the diversity in the population for preventing premature convergence. PERSONA uses a straightforward strategy of assigning each objective a dynamic leader with maximum objective performance to limit the number of global leaders in the swarm. Furthermore, PERSONA stores all the historical leader solutions in an archive in order to be capable of recalling their distinct characteristics for diversity. Additionally, PERSONA performs a parallel execution where each aligner of its population is scheduled to periodically execute its individual heuristics within its particular alignment strategy while the collaboration tasks are also scheduled to interactively perform amongst all aligners. The aligner population is diversified by focusing each member to primarily different objectives so that distinct regions of the search space can be scanned by the population. The meta-heuristic search tools described below are the flexible building blocks of the collaborative methodology:

Broadcasting and Following Leading Aligners of Each Objective: The global best aligner of each objective broadcasts its corresponding partial solution as a subgraph to the remaining follower aligners periodically.

Exchanging Partial SubGraphs with respect to Objective Score comparison: Every aligner sends their scores in each objective to all other aligners for pairwise multi-objective comparison and check whether the receiver or the sender pareto dominates the other by being superior in all the objective scores. If this is the case, then the pareto dominated party receives the whole alignment of the other one as a subgraph. Otherwise, the original sender sends one of its significant subgraphs by random objective selection.

Removing Low Scoring Mappings for Unprogressive Objective Scores: Low scoring mappings are removed by random objective selection when their respective alignment does not improve for a long period. This operation is useful for avoiding local maximizations.

Random Search: A number of random mappings are occasionally added to each alignment so that divergent solutions can be searched in the search space.

Recalling Historically Significant Partial Solutions: A random instance from The Global Best Scores History is broadcasted periodically to serve as the social memory or in other words experience of the aligner population.

Periodically Calling Various Heuristics with Random Parameters and Certain Probabilities: Every aligner of the population periodically performs its particular alignment approach based on certain heuristics with random parameter values in certain boundaries and probabilities of occurrence as part of its individual behavior. Further explanation can be found in Section 2.3 about possible heuristics that each aligner may perform individually.

The crucial collaboration tasks that these tools perform are based on extracting and exchanging the most significant subgraphs as a solution subset for each of the above mentioned objectives. The exchanged subgraph entities are extracted via special queries for each objective. These queries are explained in the Supplementary Material in detail, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2022.3231489. Apart from that, actors that may be defined as objects that encapsulate a state or behavior [29] are used as individual aligners in PERSONA due to their capability of exchanging messages with others and storing individual state information. In this context, a Tail-Chopping algorithm based scheduling approach [29] is used for choosing the next available aligner while performing interactive tasks such as exchanging results between aligners or broadcasting leading alignments that are shown in Algorithms 1 and 2 :

Algorithm 1. Send and Receive Policy for Individual Comparison and Exchange ***SG=SubGraph, LOBO= List of Better Objectives, RSO = Randomly Selected Objective*

(a) Send Policy

procedure SEND ($router$)
  for all $a \in aligners$ **do**
    $receiver \leftarrow scheduleNextReceiver(router))$
    $sendObjectiveScores(a, receiver, SG(message))$
  end for
end procedure

(b) Receive Policy

procedure ONRECEIVE ($message$)
  $senderScores \leftarrow receiveScores(message)$
  if $ParetoDominate(senderScores, receiverScores)$ **then**
    $addAllPossibleMappings(sender, receiver)$
    else if $ParetoDominate(receiverScores, senderScores)$ **then**
  $sendAllPossibleMappings(SG(receiver), sender)$
else
  $LOBO \leftarrow listBetterObjectives(sender, receiver)$
  $RSO \leftarrow randomlySelectOneObjective(L)$
  $sendBackSubGraph(SG(R), sender)$
end if
end procedure

Algorithm 2. Send and Receive Policy for BroadCast and Following Leading Aligners

(a) Send Policy for Broadcast

procedure BROADCAST
  for all $o \in objectives$ **do**
    $alignerID \leftarrow findAlignerWithBestScores(o)$
    $key \leftarrow markSubGraph(o, alignerID)$
    $broadcastBestAlignersSubGraph(key, noSender)$
  end for
end procedure

(b) Receive Policy for Following Leading Aligners

procedure ONRECEIVE ($message$)
  $key \leftarrow receiveSubGraphMark(message)$
  $addSubGraphToAlignment(key)$
end procedure

2.3 Aligner Heuristics Suite

PERSONA proposes various heuristics that can compose the characteristic behaviors of an individual aligner. Each aligner may have multiple heuristics as part of their behavior and each heuristic may have a user defined probability of occurrence at each execution cycle. These heuristics have been implemented with the Cypher Query Language [30] of the Neo4J Graph Database [31] infrastructure in order to make the search operations based on explicit reasoning. The Alignment Heuristics Suite developed as part of the study can be described in three main groups and it is further explained in the Supplementary Material, available online:

Seed and Extend Approaches: In this group of heuristics, the significant seeds are identified with respect to topological centrality scores of seeding pairs in addition to their node similarity thresholds. The alignment is intended to propagate to neighboring edges after identifying the central node pairs with a chosen centrality approach from this group. The topological centrality approaches present in this group are described by several studies [32], [33], [34], [35] as follows:

Page Rank: is the iteratively accumulated transitive influence or connectivity of each node distributed over its neighbors. The influence is computed by counting the frequency of hitting each node during a random traversal.

Betweenness Centrality: is the frequency that a node acts as an intermediary node between other nodes on a shortest path. This metric indicates the global importance of a node in terms of providing access and connectivity to other nodes.

Closeness Centrality: is the average distance of a node to all other nodes in a network based on the shortest path.

Harmonic Centrality: is a variant of closeness centrality that is based on the inverse of the distances of all other nodes rather than their distances.

Connectivity Degrees: is the number of local duples, triples or quadruples that a node is a part of.

Cluster Mapping Approach: The assumption behind this approach is that the reciprocal clusters of the compared organisms have similar interaction patterns that might indicate common functionalities as proposed by previous studies [36]. Therefore aligning interaction clusters having significant node similarity would also yield topologically strong mappings. The interaction clusters are identified either by Louvain Modularity or Label Propagation algorithms [32] as part of each heuristic. Focusing on mappings within clusters also contribute to improve the LCCS objective performances of the alignments.

Prioritization of node pairs with significant node similarity: The heuristics designed by this approach are simply based on focusing on node pairs with high BS, GOC and GOE but starting with edges or favoring edges wherever possible.

It is technically possible for a domain expert to generate a complete and ideal alignment of an organism pair by using a combination of the above mentioned heuristics with proper parameter values intuitively. Besides, a domain expert may easily improve an existing aligner by simply removing the ineffective mappings and adding effective mappings with these heuristics. Alternatively, the interactivity required for collaboration has been achieved with actors that exchange partial solutions in order to make the process more generic and expertise independent.

2.4 Alignment Process

The multi-objective optimization process of PERSONA is constructed upon a scheduling mechanism that calls the above mentioned meta-heuristic tools in a particular workflow. The user is provided some flexibility to build a custom workflow with a different order of steps and different execution periods of the meta-heuristic tools. The exact workflow designed and tested as part of this study is presented in Fig. 1. Additionally, the workflow of the whole alignment process can be summarized as four main steps that are explained below in further detail:

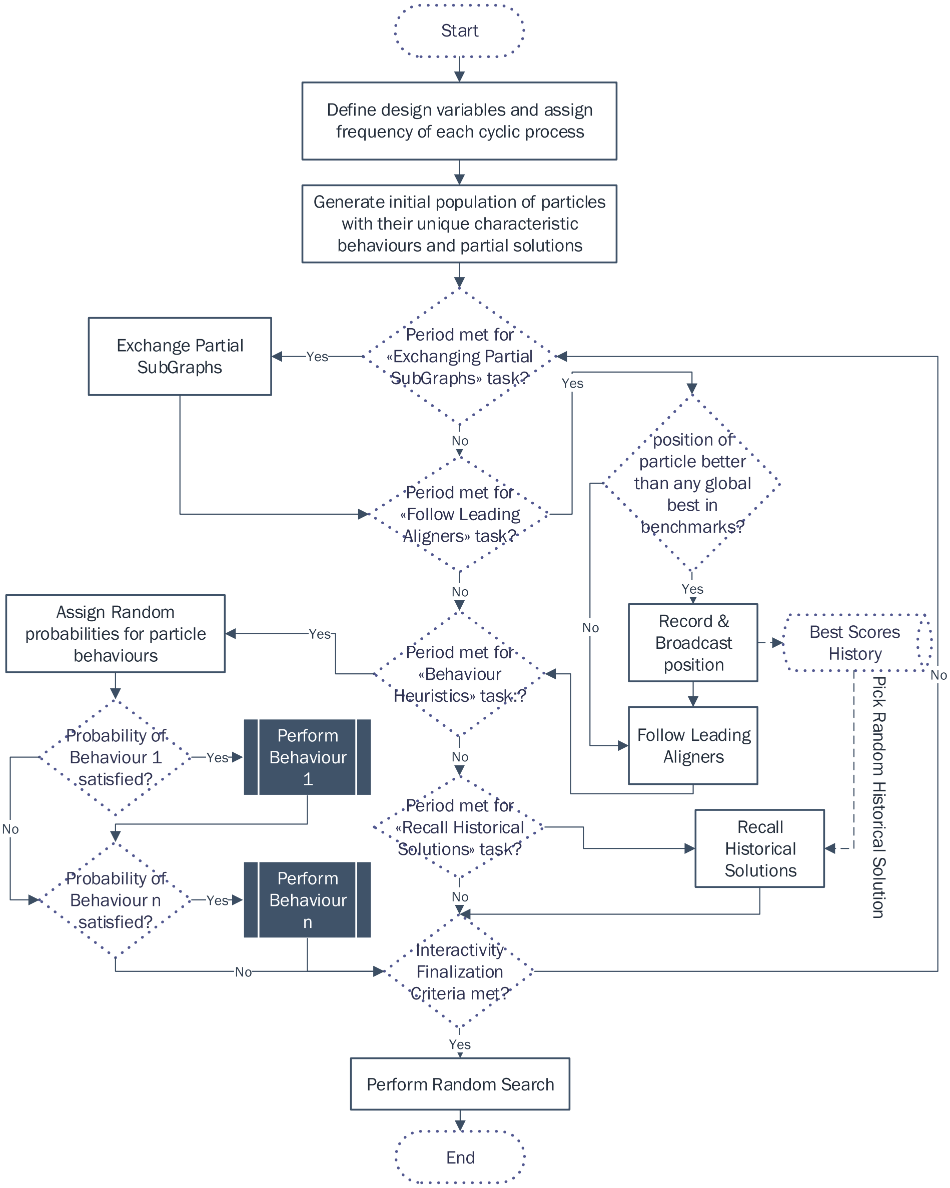


Fig. 1. Workflow of a Particle from the Swarm. The main steps are *Definition*, *Initialization*, *Collaboration* and *Post-Processing*. The schedule of *Collaboration* and *Post-Processing* steps in the workflow depend on their initially designated execution period. In this sense, these steps are executed until a stopping criterion is satisfied. The stopping criterion may either be a repetitive failure of alignment progress or a specific number of scheduled cycles depending on the design of the respective experiment.

1) Definition: Off-line jobs such as the cluster, connectivity and centrality computations are performed initially due to their computationally demanding nature. Subsequently, certain network characteristics are discovered and stored for the alignment strategy to be network independent. These characteristics include summary statistics and percentile based meta-data of centrality scores, biological sequence similarities and shared gene ontology terms between node pairs. Next, the number of aligners and the periodical schedule of the specific alignment jobs of aligners and exchange jobs are set for execution. In this scope, certain heuristics are scheduled as part of each alignment behavior while global and pairwise exchange policies are also defined and scheduled.
2) Initialization: Later on, some characteristic heuristics are used to initialize each aligner with proper similarity thresholds in order to propagate the alignment up to a mature state. Any available heuristic can be used for this purpose but the seed-and-extend heuristics are the most powerful candidates for the maturation intended in this step since they can identify topologically central mappings that also have the highest possible node similarity values initially. These heuristics can further be repeated by gradually lowering their node similarity thresholds. Alternatively, incomplete external alignments may be used to initialize each aligner. As a result, aligners achieve their initial partial solutions before starting the consecutive interactive phase of the application.
3) Collaboration: This step is the collaborative step where all aligners in the system act with a constant frequency of self improvement and a constant frequency of exchange that is defined and scheduled in

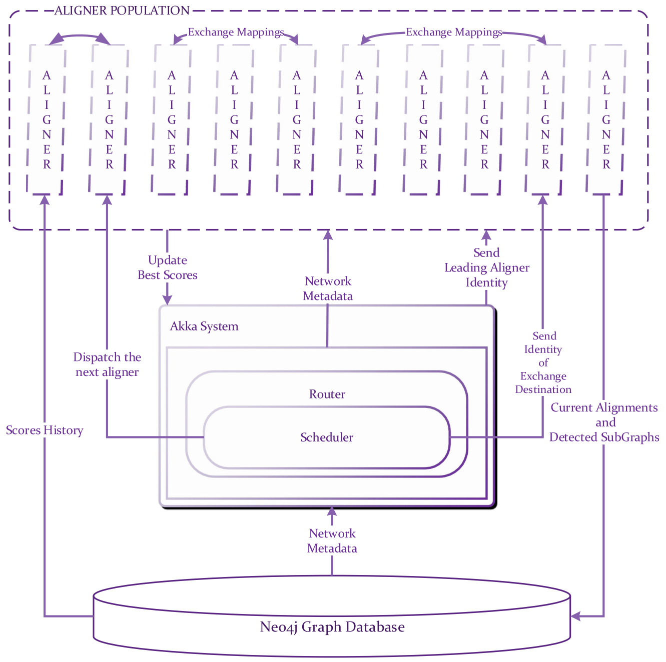


Fig. 2. Cooperative Architecture of the Swarm. This architecture demonstrates the information exchange infrastructure as a basis of cooperation among the particles in the swarm. Each particle in the swarm is implemented as a *Typed Actor* with standard send and receive policies for group and individual behavior formalized in Algorithm 1 and 2. The architecture mainly relies on a *Neo4J Graph Database* that stores individual aligner states and exchanged subgraphs along with a *Akka Concurrency System* that schedules execution order of sender and receiver aligners by means of a *Scheduler* and *Router*. The scheduler and router identifies and assigns upcoming senders and receivers based on their identity that is used in retrieving their latest state and significant subgraphs. Best Scores in each objective is updated after each mapping task and notified to the whole swarm.

the *Definition* step. Each aligner uses its own set of heuristics with randomized occurrence probabilities and parameter values as part of its behavior at every cycle. The parameter values are randomly assigned to each heuristic regarding their value interval during this process. Moreover, most significant mappings of leading aligners of each objective are periodically sent to other aligners and are also marked for future use during an alignment process. Finally, mappings with no contribution are removed as well as the ones that violate the one-to-one mapping restriction in each cycle. Meanwhile, a counter counts the number of cycles that an aligner has been unprogressive for and when the counter of an aligner reaches the threshold value, then the system randomly removes a limited number of minimally contributing mappings.

4) Post-Processing: This step is used for fine-tuning the alignments achieved through the PERSONA step. The standard procedure of this step is completing the alignment with random search for finding random mappings with positive effect. The standard procedure is applied by adding a limited number of random mappings and removing the ones with no positive effect in a loop until the alignment is complete. The pareto dominated alignments are filtered as a final task.

Generally, all optimization techniques require adaptations for preventing premature convergence and improving access to different regions of the search space. In PERSONA, every aligner specializes in particular set of heuristics and improves its alignment mostly with them. On the other hand, aligners can not reach certain regions of the search space if they are restricted with their own heuristics and that's why they need to collaborate to optimize their solutions. The collaboration prevents their convergence to a premature solution.

2.5 Architectural Design

PERSONA is implemented with the Typed Actor paradigm of Akka Concurrency Framework [29] utilizing tools such as Future Messages for concurrent completion of tasks, Typed Actors for implementing distinct population members with standard message receiving patterns, Routers and Schedulers for determining the subsequent message recipients as well as a Neo4J Graph Database infrastructure for persistence of aligner states and exchanged subgraphs. Fig. 2 shows the flow of information among all essential entities of the method (see Supplementary Material, available online, for more information about the whole design).

3 Comparison of Results

Since PERSONA aims to achieve a balanced blend of results, we chose its competitors among aligners that possess a potential to achieve balance in multiple objectives. We performed the experiments with four distinctly extracted real world data sets for detailed interpretation. This chapter explains the process of these tasks in detail.

3.1 Implementation Characteristics of PERSONA Population

We have implemented PERSONA with a special experimental setup that consists of a population of maximum ten customly designed aligners with various different and complementary characteristics. We designed each member of the population by using a combination of items from the previously mentioned Aligner Heuristics Suite as part of its behavior. Thus, each member may be regarded as an individual alignment algorithm implemented for this study. The distinguishing feature for each aligner was its primary heuristic. Some aligners relied mainly on a unique selection of a centrality detection algorithm while the remaining ones either had a main behavior of node pair prioritization based on node similarity or a cluster mapping approach based on a unique clustering technique from our Aligner Heuristics Suite. We intended to create diversity in the population by dedicating a single clustering or centrality detection heuristic to a particular aligner of the population. The respective aligner names in the population based on their primary heuristic is listed below:

- Page Rank Seeding Aligner,
- Betweenness Centrality Seeding Aligner,
- Closeness Centrality Seeding Aligner,
- Harmonic Centrality Seeding Aligner,
- Connectivity Degrees Seeding Aligner,
- Cluster Mapping Aligner with Label Propagation,

- Cluster Mapping Aligner with Louvain Modularity,
- Sequence Similarity Prioritizing Aligner,
- Sequence Similarity Seeding Aligner,
- Hybrid Centrality and Sequence Similarity Aligner

All the above mentioned aligners in our experiments performed their primary heuristics with high probabilities as part of their main behavior in each interaction cycle. Besides, all aligners performed secondary or complementary heuristics from the heuristics suite in lower probabilities than their primary heuristics. The main reason for executing heuristics with non-standard probabilities was to increase randomness and flexibility to search for optimized solutions. Additionally, we also randomized the output of each heuristic by random value assignment for the parameters that it requires. The secondary heuristic that we used most frequently was "Heuristic for Forming Edge Pair Mappings from Existing Node Pair Mappings" since it enables to propagate with edges of particular node similarity from seed mappings. Besides, we also used "Heuristic for Removing Inductive Mappings" and "Heuristic for Removing Low Scoring Mappings" frequently for especially opening up search space productively for every aligner. Since each aligner was composed of a combination of heuristics, its probability of achieving a balance among the favored objectives is improved remarkably. Finally, we employed the Post-Processing step of PERSONA with an individual alignment approach that starts with Biological Similarity Seeding. The propagation after seeding was maintained by forming edge pairs with an incremental constraint relaxation strategy in each cycle.

3.2 Competitors & Simulation Environment

NETAL [6], SPINAL [14], PISwap [37] and HubAlign [5] may be defined as a particular class of competitors for PERSONA since they are all deterministic aligners that evaluate topological and biological inputs. NETAL uses a greedy method by evaluating an alignment scoring matrix. HubAlign is based on preliminarily evaluating and scoring the topological and biological importance of proteins to identify hub nodes to align and then assigning alignment scores to protein pairs by considering sequence similarity and the importance score. PISwap is a method that iteratively refines the initial alignments of custom heuristics with topological information while compromising sequence information achieved by the well-known Hungarian algorithm [38]. On the other hand, SPINAL performs a fine-grained conflict resolution and following a coarse-grained construction of estimate scores. We executed these aligners iteratively in their most powerful range of each application parameter for producing significant sets of results.

SANA [9], PROPER [39], MAGNA++ [40] and PINALOG [41] form another class of competitors for PERSONA since they employ non-deterministic optimization by evaluating topological and biological inputs in order to generate a single alignment. SANA follows a simulated annealing based optimization approach. PROPER generates a seed of high sequence similarity protein pairs based on percolation matching and then progresses only with structural mapping. MAGNA++ is an improvement version over the original MAGNA [11] method that combines existing 'parent' alignments into superior 'children' alignments and then evolves this process over multiple generations. It enables maximization of a node conservation measure simultaneously with the chosen edge conservation measure and provides automatical utilization of all available resources by means of parallelization. Finally, PINALOG method combines information from protein sequence, function and network topology information and it consists of 3 fundamental steps starting with preliminary detection of communities with CFinder [42], followed by community mapping with respect to similarities and finalized with extension mapping of proteins in the neighbourhood of the core protein pairs.

We finally evaluated other unique alignment algorithms such as GEDEVO [43], [44] and OptNetAlign [8] for the purpose of comparing a diverse set of approaches. GEDEVO is a graph comparison tool that generate a single alignment based on the so-called Graph Edit Distance (GED) model where one graph is to be transferred into another one with a minimal number of edge insertions and deletions. The optimization methodology of this tool relies mainly on topological information but it can also be extended to utilize biological similarity. Conversely, OptNetAlign performs multi-objective optimization with respect to functional, biological and topological inputs based on a genetic algorithm that employs Uniform Partially Matched Crossover and hill climbing on a population of pareto optimal alignment results. PERSONA generates a population of alignments similar to OptNetAlign but it rather performs the pareto optimality check as a final step. Nevertheless, it manages to generate alignments that are stronger in various objectives due to the different nature of aligner behaviors in its population.

We principally chose most of these competitors due to their two-sided nature that compromise between node similarity and topological similarity measures. Another reason for choosing them was their applicability due to their existing documentation and source codes. We intended to compare our method with aligners such as IBNAL [45] and SSAlign [46] that assume annotational, biological and topological inputs simultaneously along with OptNetAlign and PINALOG due to their capability of mapping experimentally verified annotations but their source code was unavailable. We executed each competitor algorithm with the most possible balanced objective function based on the provided application parameters, performance related instructions of the authors and recommended modes for an effective experimental setup. In this regard, we identified the most balanced parameter combination of for each algorithm after several experiments and provided the resulting commands along with more detailed information in the Supplementary Material, available online. The source code, application and execution instructions of PERSONA is also available on the github repository https://github.com/giraygi/ppi-alignment.

3.3 Data Sets

We evaluated competitor algorithms along with PERSONA by comparing C. Elegans (CE), S. Cerevisiae (SC), M. Musculus (MM) and H. Sapiens (HS) with D. Melanogaster (DM) based on their protein, network, pairwise biological sequence similarity and annotation data. We initially used the earliest benchmark data set Isobase [47] for HS-DM comparison since it was tested by several algorithms in the

TABLE 1
Network Sizes of Data Sets

	Nodes Left	Nodes Right	Edges Left	Edges Right	Similarity Links	Average Common Annotations
DM-CE Intact	8532	4950	26289	11550	5669	9.07
DM-SC BioGRID	7937	5831	34753	77149	132007	1.64
HS-DM Isobase	9633	7518	36386	25830	97172	1.26
DM-MM Mentha	10827	9674	45706	31577	178415	5.93

The first compared organism in the first column is denoted as "Left", whereas the second compared organism is denoted as "Right" in the other columns. The "Similarity Links" column denotes the number of available BS links throughout all possible node pairs between the first and second compared organisms. "Average Common Annotations" column denotes the average number of common annotations throughout all possible node pairs between the first and second compared organisms.

literature. Later on, some distinctive data sets extracted by recent aligners such as PROPER and SANA were also used in evaluating DM-CE and DM-SC comparisons respectively. The network data of PROPER was retrieved from Intact 2016 [48] and integrated with sequence similarities from UniProt [49] as well as experimentally verified terms denoted by "EXP", "IDA", "IMP", "IGI", "IEP" and "IPI" codes in Gene Ontology Annotation (UniProt-GOA) [50]. The data set of SANA was retrieved from BioGRID 2017 [51] database with a complete list of protein-protein interactions, experimentally verified GO terms and sequence similarities. Finally, we evaluated the competing alignment algorithms in DM-MM comparison with the 12.07.2021 dated release of Mentha [52] data set that integrates experimentally curated PPI data of several molecular interaction databases by providing automated access with the Proteomics Standard Initiative Common Query Interface (PSICQUIC) [53], [54] in compliance with International Molecular Exchange (IMEx) [55] policies. We further integrated the sequence similarities computed by BLASTP from FASTA texts in UniProt and annotations from UniProt-GOA into Mentha networks in a similar fashion with the Intact data set.

As a notable remark, PROPER has used Intact 2016 data set with a considerably high threshold for BS that ignores lower similarity values and complicating its competitors in achieving high BS scores. For this reason, the respective data set will be referred as "Trimmed Intact 2016" in the following text. On the other hand, we generated the respective Annotation Corpus of each data set and used the FastSemSim python library [56] for computing Resnik Similarity of GO terms of protein pairs with the same maximum best match mixing strategy chosen by SANA based on the fact that Resnik similarity of the most specific common ancestor has been shown to provide more reliable results than aggregating similarities [57]. Additionally, we removed the repeating gene ontology terms in BioGRID 2017 data set to extract another gene ontology instance for using the data set conveniently with the more general objective GOC. We tested PERSONA with the exact form of these data sets used by the aforementioned aligners for keeping the comparison conditions uninfluenced. Table 1 represents the number of nodes, edges and sequence similarity links along with the average number of common annotations among possible node pairs for each organism pair evaluated in Section 3.4

Most of the data sets used by the competitors were not compatible due to the different input varieties they required. For this reason, we carried out complex transformation procedures for being able to test all aligners with data sets mentioned in this study along with additional ones for future experiments. We stored resulting data sets in a github repository https://github.com/giraygi/ppi-alignment-data with some of the transformation procedures in https://github.com/giraygi/ppi-alignment-converters/. We tested the competitor algorithms with these data sets.

3.4 Results

The alignments generated with the leading "balanced" aligners are mostly not able to pareto dominate each other since they are not completely superior than one another in all desired objectives. Therefore, it becomes necessary to make an assumption about the significance of each objective. For this reason, it makes sense to group resembling objectives and to make an assumption about the significance of each objective and each group of objectives. As part of a meaningful classification of objectives, the edge similarity based topological measures can be grouped with each other while the functional measures based on shared annotations can be proposed as another group. We can then include the undoubtedly meaningful primitive measure of edge similarity as well as the more developed S^3 that is able to penalize the dense-to-sparse mappings into the edge similarity based topological measures group. On the other hand, it is also meaningful to group GOC and GOE as part of another single objective since they interpret the functional coverage by directly using the same annotational data. Thus, it becomes easier to assign weights of importance to groups rather than individual objectives.

Consequently, we can consider five groups of objectives being Edge Similarity (EC and S^3), Global Topological Coverage (LCCS), Annotational Coverage (GOC and GOE), Sequence Similarity (BS) and Semantic Similarity (Resnik). By assigning a coefficient of 1 to each group and simply dividing the coefficients within the groups to group sizes, EC, S^3, GOC and GOE get coefficients of 0.5 whereas LCCS, Resnik and BS get coefficients of 1. Therefore, it is possible to conclude a simplistic multi-criteria decision making model for either a Weighted Sum Model (WSM) or a Weighted Product Model (WPM) [58] as an abstract aggregated objective function in order to be able to choose between different pareto optimal solutions without giving priority to any objective group or individual objective other than their designated coefficient. The WSM approach would also require the results to be comparable so that they can be used as part of the same equation. We normalized all objectives defined in Section 2.1 with Total Sum Scaling and Z-Score Normalization techniques in order to obtain

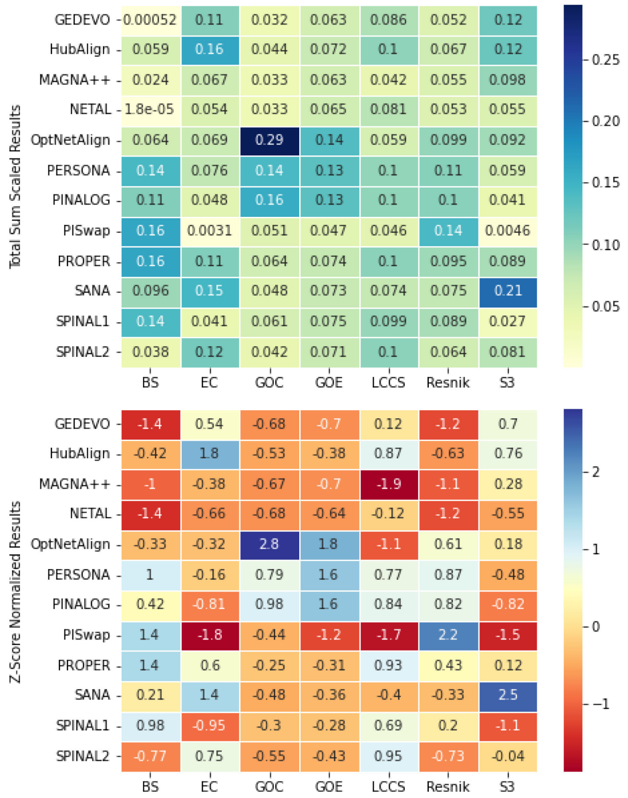


Fig. 3. Objective scores on DM-CE intact 2016 data set.

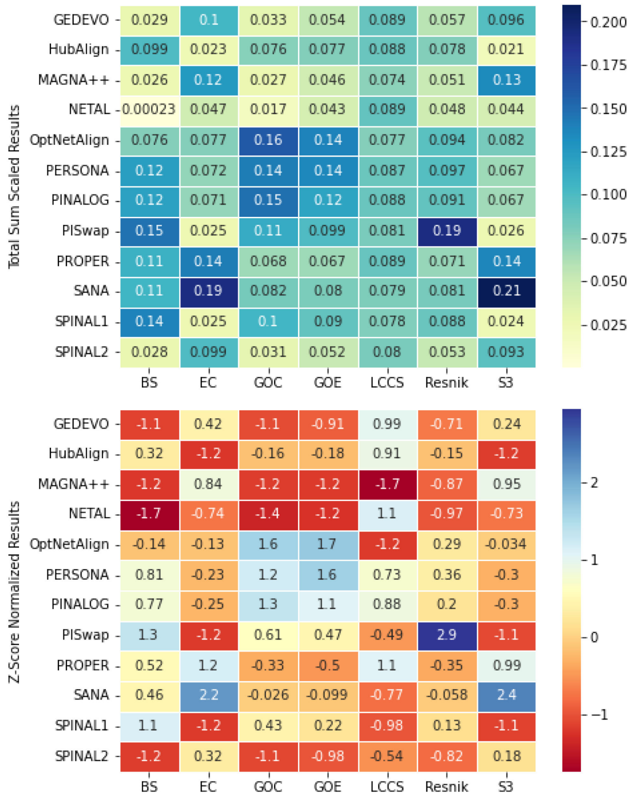


Fig. 4. Objective scores on DM-SC BioGRID 2017 data set.

comparable results for this purpose. We employed the formulas (4) and (5) in the interpretation of results for comparing the alignments that are not able to pareto dominate each other.

$$WSM = \frac{EC}{2} + \frac{S^3}{2} + LCCS + \frac{GOC}{2} + \frac{GOE}{2} + BS + Resnik \tag{4}$$

$$WPM = \sqrt{EC \times S^3 \times GOC \times GOE} \times LCCS \times BS \times Resnik \tag{5}$$

Subsequently, the formula (6) represents the relative performance of Algorithm K to Algorithm L since all the objectives are benefit criteria and the higher values of them represent better performance accordingly. The normalized performance of the Algorithm K on the j_{th} objective is represented as a_{K_j} in the formula. Besides, W is the vector of the same objective coefficients used in the formulas (4) and (5).

$$P(A_{K/}A_L) = \prod_{j=1}^{n} (a_{K_j}/a_{L_j})^{W_j}, for K, L, = 1, 2, 3, ..m. \tag{6}$$

Average performances of algorithms in the full set of objectives are represented in Figs. 3, 4, 5 and 6. Z-Score Normalized Results of each column sum up to 0 and Total Sum Scaled Results of each column sum up to 1 for each objective in these figures. For aggregating average performances into a single alignment score, both WSM and WPM were applied for Total Sum Scaled data whereas only WSM was applied for Z-Score Normalized data since it includes negative values. Higher aligner scores represent higher performance in the respective $WSM_{TotalSum}$, $WSM_{Z-Score}$ and $WPM_{TotalSum}(Other)/(PERSONA)$ calculations of Fig. 7.

We carried out all experiments with an Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz processor and 16 GB RAM on a 64 bit Linux platform. One exception was PINALOG since we alternatively carried out its execution with the supercomputer at Freie Universität Berlin called CURTA [59] after it endangered the safety of the personal computer due to its 24 hours execution period without termination and severe heating problems. The CPU time required by each algorithm on DM-CE Trimmed Intact 2016 Data Set was shown in Table 2. We executed time consuming algorithms for 3 hours except for PINALOG and MAGNA++ that both lack such a time parameter. For this reason, we configured MAGNA++ to run for 2000 cycles with all data sets and this period corresponded to 4.5 hours with DM-CE while more for other data sets. On the other hand, PINALOG ran entirely for 24 hours with DM-CE data set. We distributed the time consumption proportionally to the number of produced alignments where applicable.

4 Discussion

Our approach for comparing multi-objective performance of PERSONA with the competitor GNA algorithms in this study is based on an abstract aggregated objective function as it is explained in Section 3.4. For this purpose, we have assigned a coefficient of significance to each alignment objective and we have used aggregate WSM and WPM scores as the primary source of comparison since it is very hard to interpret the relative performance of each competitor algorithm in each individual objective. The aggregate

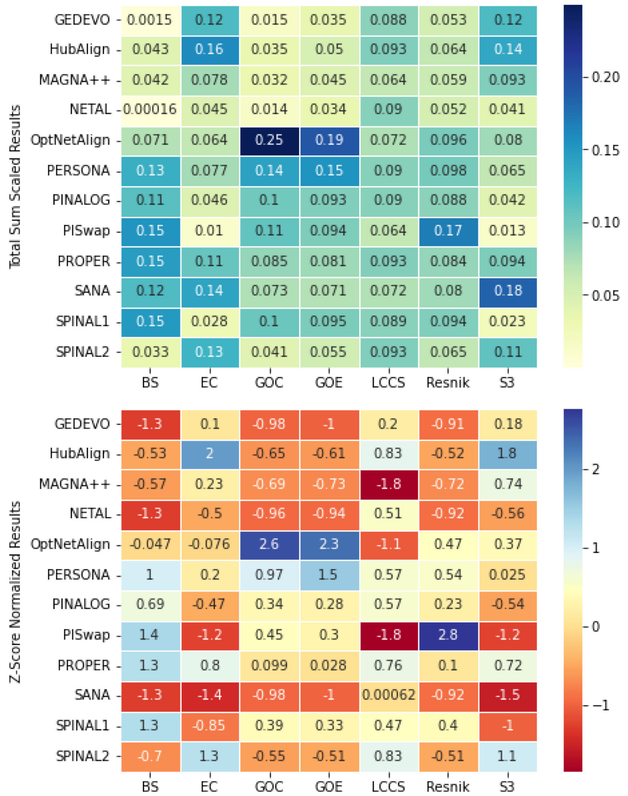


Fig. 5. Objective scores on HS-DM isobase data set.

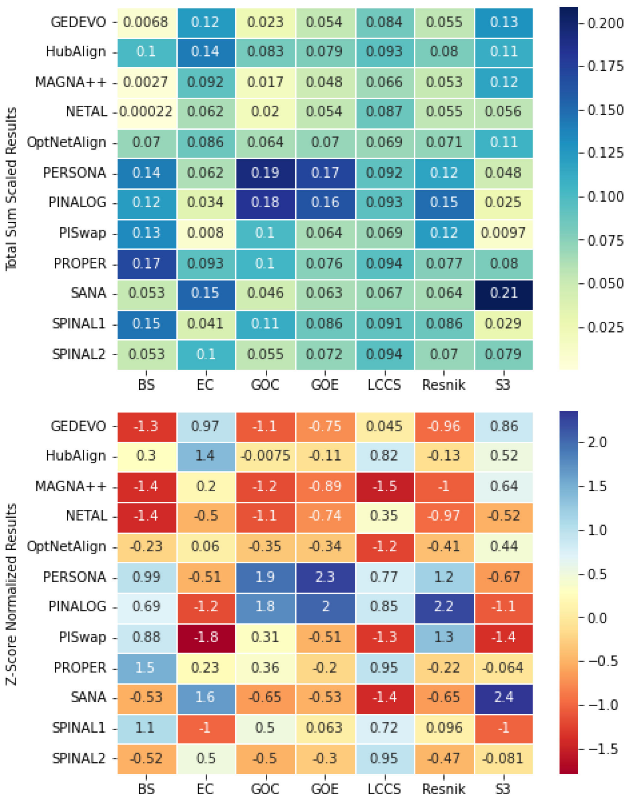
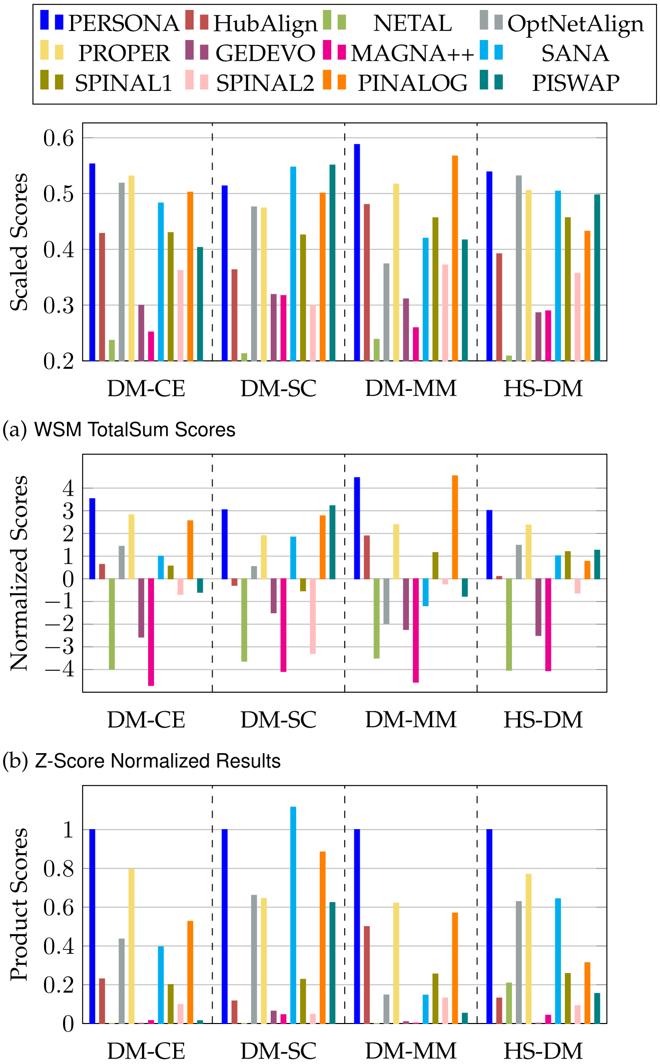


Fig. 6. Objective scores on DM-MM Mentha 2021 data set.



(a) WSM TotalSum Scores

(b) Z-Score Normalized Results

(c) WPM Scores

Fig. 7. Aggregate scores with all datasets.

TABLE 2
Performances on DM-CE Trimmed Intact 2016

	CPU Seconds
PROPER	12.1
PISWAP	13.40
NETAL	57.84
PERSONA	240
HubAlign	247.06
Spinal	579.14
OptNetAlign	2457.42
SANA	10252.56
MAGNA++	22154.33
GEDEVO	42114.59
PINALOG	89082.05

scores in Fig. 7 show that PERSONA is able to achieve a consistently high performance whereas other aligners occasionally achieved better results in certain data sets. PERSONA ranked the first in all three types of aggregation for the DM-CE, DM-MM and HS-DM data sets while ranking the second on DM-SC data set. On top of the aggregate scores, individual objective scores in Figs. 3, 4, 5 and 6 also gave some remarkable insights about the alignment methodologies. Furthermore, dual interpretations of objectives also revealed the trade-offs or compromises inherently carried within each competitor algorithm. In this context, PERSONA achieved a considerably high performance on GOC, GOE, Resnik and BS. Additionally, it achieved a relatively good performance of LCCS. However, the collaboration strategy of PERSONA increased the probability of leaving unaligned edges due to trying to keep the balance for other objectives. Consequently, it achieved relatively poor results in S^3 as the unaligned edges are penalized.

There were some key findings about the effects of data sets over alignment performance. As a first example, it was observed that PROPER focuses only on top ranking sequence similarities by default and consequently it gains an advantage in DM-CE comparison with Trimmed Intact 2016 and DM-MM comparison with Mentha data sets in terms of BS quality due to the high number of top ranking pairwise similarities they provide. It is also worth mentioning that the Trimmed Intact 2016 Data Set which was generated for the PROPER paper ignores pairwise BS scores less than 150 bits by trimming an essential part of the respective data resulting in a performance compromise for competitor algorithms that are able to consider low similarity node pairs in their alignment approach. As another example, HubAlign did not perform well in the pairwise comparison of DM-SC on primarily the topological similarity objectives despite its relatively better performance on the other pairwise comparisons. This shows that the algorithm is not capable of adapting certain topological network characteristics of the respective organism pair. As the third example, OptNetAlign achieved distinctive performance on GOC objective with DM-CE and HS-DM data sets but did not manage to show the same performance on other data sets. As a final example, SANA was overall more successful in the dataset prepared for its application. This maybe be due to problems of replicating their complex input formats in other data sets or it may as well show that the algorithm has a highly specific approach of handling data.

There were other key findings about the performances of competitors. For instance, PISwap consistently achieved very high Resnik Similarity Scores in all data sets despite its moderate scores in GOC and GOE. It also achieved the best aggregate Scaled and Normalized scores on DM-SC BioGRID 2017 data set whereas SANA achieved the best aggregate Product Score for the same data set. On the other hand, PINALOG performed better with annotational inputs and consequently its results without them were discarded. It also performed comparably better in functional and biological objectives despite its lower performance in topological objectives. It is also worth noting that the first mode of SPINAL called SPINAL1 generated better aggregate results than its second mode SPINAL2 despite its relatively lower performance on topological similarity objectives. This result was mainly due to the fact that topological improvement of SPINAL2 has resulted in a more significant compromise in the node similarity objectives and SPINAL1 has managed to generate more balanced alignments. On the other hand, NETAL did not offer any input methodology for node similarity data although stated otherwise in its documentation. For this reason, it did not generate any positive results in BS and it could generate barely positive results in GOC and GOE. Similarly, GEDEVO was not able to produce moderate results in the functional and biological objectives which may be due to its requirement of BLAST distance matrix input rather than a BLAST similarity input used in other algorithms. Finally, MAGNA++ was able to achieve moderately positive results on BS and yet it was not able to achieve positive results in the functional objectives of GOC, GOE and Resnik.

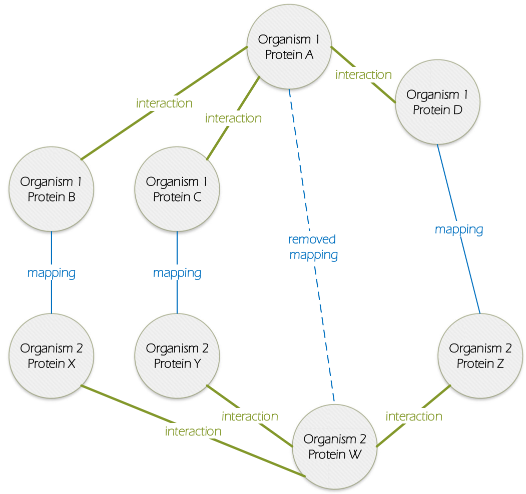


Fig. 8. Removed mapping effecting topological similarity. The removed mapping demonstrated with a dashed line has a critical effect upon topological similarity since it forms 3 aligned edges contributing to topological similarity. However, the remaining mappings cannot form any aligned edges without it.

The aim of this study is developing a generic and balanced approach that would perform well on every kind of network generated through various data sources. On the meta-heuristic level, the application harnesses the differences between concurrently executing aligners by comparing their current results in multiple objectives and sending the superior subgraphs to low scoring aligners so that the population achieves a balanced performance. It is also possible that some exchanged mappings may not contribute to the receiving aligner since some of its node pairs are already occupied with other mappings. In such cases, only the possible mappings of the superior subgraph are used in alignment but then it becomes possible that the remaining mappings have mostly lost their superiority characteristics. Fig. 8 illustrates such a topological similarity loss upon removal of the central mapping between proteins A and W from a subgraph. PERSONA employs a removing policy in each execution cycle in order to filter such remaining mappings that has no contribution. However, it may also be possible to predict the compromises or trade-offs beforehand and prevent sending useless subgraphs with a more precise conflict resolution policy. Yet, such a policy may conversely create some overhead due to alignment score computation of intersecting subgraphs beforehand.

The current version of PERSONA does not filter pareto dominated solutions until the final step. Filtering them during the Collaboration step may also make sense in terms of dealing with an elite and useful set of alignments. On the contrary, pareto dominated solutions may still include significant subgraphs to be evaluated by a superior alignment with collaboration as a remark of partial performance. The significant partial alignments generated throughout the alignment process may also be used as building blocks of a superior global alignment or shed light to a local alignment solution. In this scope, strong partial alignments achieved by PERSONA or other aligners can easily be merged with each other by PERSONA infrastructure. One key feature introduced by PERSONA for this task is the accurate measurement of partial performances in each objective.

PERSONA would also demonstrate a remarkable performance in a hypothetical objective that unifies biological, topological and annotational similarity objectives since it enables querying individually and topologically significant node pairs simultaneously as part of a single heuristic. Individually significant node pairs are individual mappings that possess significant biological and annotational similarity with each other regardless of their interactions with other nodes. On the other hand, topological significance refers to multiple node pairs that form a remarkable number of aligned edges when they are mapped with each other. Individually and topologically significant node mappings identified by this feature may compose the core of an alignment since they would require minimal or no compromise in any objective. Subgraphs composed of such mappings may also be stored as a regional benchmark in the collective memory to be evaluated by each population member for merging with its individual alignment. Apart from that, these significant subgraphs may further be merged with themselves in different combinations in order to build superior alignments with minimal compromise in most objectives.

Another future improvement in the data model could be modeling the Dynamic Network Alignment problem [60] mentioned in DynaWAVE [61] on top of the current data model of PERSONA. The current data model may easily be adapted to the Dynamic Network Alignment problem that addresses the temporal component of protein-protein interactions since it only requires node properties for activation and deactivation times of the interactions. Last but not the least, a many-to-many local or global alignment solution may also be incorporated into this application by removing the one-to-one mapping restriction from its built-in search tools and heuristics libraries. Such a solution might either be used as part of a complete alignment procedure or an aggregation procedure of external one-to-one alignments in the same fashion with Ulign [62].

5 Conclusion

PERSONA has been tested with a number of data sets published in different years to demonstrate the evolution of available data in time with new findings. Consequently, it enables an in-depth analysis of alignment performance with a temporal component. In this context, the results of this study prove that PERSONA is generating high performance alignments with all data sets thanks to its collaborative approach. Thus, it becomes obvious that it follows a network independent and robust methodology compared to other aligners that demonstrate an unstable performance on various different data sets. Therefore, it may be concluded that the particle swarm inspired meta-heuristic optimization approach of PERSONA is able to adapt the characteristics of evolving networks due to its collaboration strategies as a swarm, individual alignment heuristics as behaviors, interpretation of network meta-data, randomized discovery in the search space and so on.

PERSONA is able to handle local maximums by removing some of the minimally contributing mappings of an alignment for opening search space in each execution cycle. In order not to repeat weak mappings, certain tasks of the algorithm such as assigning parameter values in alignment heuristics, removal of minimally contributing mappings, search in the post-processing phase, selecting historically significant partial solutions or selection of partial alignments to be exchanged in each cycle are implemented in a randomized fashion. Such randomized tasks enabled access to different regions in the search space so that stronger mappings could be discovered without converging prematurely. It is also essential to decide how to initialize the particles of the swarm since the initial mappings heavily effect the final alignments. The reason is that, initial mappings narrow down the solution space and they may also be exchanged among the concurrently progressing aligners even if some of them are removed from an alignment in the further cycles of execution.

This study showed that PERSONA is able to achieve remarkable results among multiple objectives of the Global Network Alignment problem. The performance of PERSONA and the competitor aligners were evaluated with multiple criteria decision making tools of WPM and WSM to reach a conclusion in this sense. On the contrary, balance among multiple objectives may still be compromised for achieving superiority on some of the objectives based on user priorities. User priorities can always be targeted by altering aligner behaviours. However, it should be noted that most aligners make more significant compromises in the objectives that they ignore compared to the superior results that they achieve in the objectives that they focus. Therefore, the necessity of storing and further utilization of partial solutions with significant topological similarity and node similarity proves to be true for preventing huge compromises in contradicting objectives. Proper PERSONA methods should be employed for further utilization of such partial solutions.

Acknowledgments

The authors would like to thank Professor Dr. Tolga CAN from the Computer Science Department of Colorado School of Mines for his encouraging and remarkable ideas.

References

[1] Y. Zhu, Y. Li, J. Liu, L. Qin, and J. X. Yu, "Discovering large conserved functional components in global network alignment by graph matching," *BMC Genomic.*, vol. 19, no. S7, Sep. 2018, Art. no. 670. [Online]. Available: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-5027-9

[2] L. Meng, A. Striegel, and T. Milenković, "Local versus global biological network alignment," *Bioinformatics*, vol. 32, no. 20, pp. 3155–3164, Oct. 2016. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw348

[3] F. E. Faisal, L. Meng, J. Crawford, and T. Milenković, "The post-genomic era of biological network alignment," *EURASIP J. Bioinf. Syst. Biol.*, vol. 2015, no. 1, Dec. 2015, Art. no. 3. [Online]. Available: https://bsb-eurasipjournals.springeropen.com/articles/10.1186/s13637-015-0022-9

[4] S. Maskey and Y.-R. Cho, "Survey of biological network alignment: Cross-species analysis of conserved systems," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 2090–2096. [Online]. Available: https://ieeexplore.ieee.org/document/8983132/

[5] S. Hashemifar and J. Xu, "HubAlign: An accurate and efficient method for global alignment of protein–protein interaction networks," *Bioinformatics*, vol. 30, no. 17, pp. i438–i444, Sep. 2014. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu450

[6] B. Neyshabur, A. Khadem, S. Hashemifar, and S. S. Arab, "NETAL: A new graph-based method for global alignment of protein–protein interaction networks," *Bioinformatics*, vol. 29, no. 13, pp. 1654–1662, Jul. 2013. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt202

[7] J. Crawford, Y. Sun, and T. Milenković, "Fair evaluation of global network aligners," *Algorithms Mol. Biol.*, vol. 10, no. 1, Dec. 2015, Art. no. 19. [Online]. Available: http://www.almob.org/content/10/1/19

[8] C. Clark and J. Kalita, "A multiobjective memetic algorithm for PPI network alignment," *Bioinformatics*, vol. 31, no. 12, pp. 1988–1998, Jun. 2015.

[9] N. Mamano and W. B. Hayes, "SANA: Simulated annealing far outperforms many other search algorithms for biological network alignment," *Bioinformatics*, vol. 33, no. 14, pp. 2156–2164, Jul. 2017.

[10] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj, "Topological network alignment uncovers biological function and phylogeny," *J. Roy. Soc. Interface*, vol. 7, no. 50, pp. 1341–1354, Sep. 2010.

[11] V. Saraph and T. Milenković, "MAGNA: Maximizing accuracy in global network alignment," *Bioinformatics*, vol. 30, no. 20, pp. 2931–2940, Oct. 2014.

[12] O. Kuchaiev and N. Pržulj, "Integrative network alignment reveals large regions of global network similarity in yeast and human," *Bioinformatics*, vol. 27, no. 10, pp. 1390–1396, May 2011.

[13] M. Ashburner et al., "Gene Ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, no. 1, pp. 25–29, May 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/

[14] A. E. Aladağ and C. Erten, "SPINAL: Scalable protein interaction network alignment," *Bioinformatics*, vol. 29, no. 7, pp. 917–924, Apr. 2013.

[15] Y. Sun, J. Crawford, J. Tang, and T. Milenković, "Simultaneous optimization of both node and edge conservation in network alignment via WAVE," in *Proc. Int. Workshop Algorithms Bioinf.*, 2015, pp. 16–39.

[16] P. H. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: Assessment with biological features and issues," *Brief. Bioinf.*, vol. 13, no. 5, pp. 569–585, Sep. 2012. [Online]. Available: https://academic.oup.com/bib/article/13/5/569/411449

[17] C. Pesquita, "Semantic similarity in the gene ontology," in *The Gene Ontology Handbook*, C. Dessimoz and N. Škunca Eds., New York, NY, USA: Springer, 2017, pp. 161–173.

[18] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artif. Intell.*, 1995, pp. 448–453.

[19] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[20] C. Clark and J. Kalita, "A comparison of algorithms for the pairwise alignment of biological networks," *Bioinformatics*, vol. 30, no. 16, pp. 2351–2359, Aug. 2014.

[21] S. F. Altschul et al., "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[22] C. Camacho et al., "BLAST+: Architecture and applications," *BMC Bioinf.*, vol. 10, no. 1, Dec. 2009, Art. no. 421.

[23] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proc. Nat. Acad. Sci. USA*, vol. 85, no. 8, pp. 2444–2448, Apr. 1988. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC280013/

[24] M. Cannataro and P. H. Guzzi, *Data Management of Protein Interaction Networks*. Oxford, U.K.: Wiley, 2011.

[25] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. Int. Conf. Neural Netw.*, 1995, pp. 1942–1948.

[26] J. S. Arora, "Multiobjective optimum design concepts and methods," in *Introduction to Optimum Design*. Amsterdam, The Netherlands: Elsevier, 2004, pp. 543–563. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780120641550500173

[27] M. Mahmoodabadi, A. A. Safaie, A. Bagheri, and N. Nariman-zadeh, "A novel combination of particle swarm optimization and genetic algorithm for pareto optimal design of a five-degree of freedom vehicle vibration model," *Appl. Soft Comput.*, vol. 13, no. 5, pp. 2577–2591, May 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494612005121

[28] Z. Fan, T. Wang, Z. Cheng, G. Li, and F. Gu, "An improved multiobjective particle swarm optimization algorithm using minimum distance of point to line," *Shock Vib.*, vol. 2017, pp. 1–16, 2017. [Online]. Available: https://www.hindawi.com/journals/sv/2017/8204867/

[29] M. K. Gupta, *Akka Essentials: A Practical, Step-by-Step Guide to Learn and Build Akka's Actor-Based, Distributed, Concurrent, and Scalable Java Applications*. Birmingham, U.K.: Packt, 2012.

[30] N. Francis et al., "Cypher: An evolving query language for property graphs," in *Proc. Int. Conf. Manage. Data*, 2018, pp. 1433–1445.

[31] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases*, 2nd ed., Beijing, China: O'Reilly, 2015.

[32] M. Needham and A. E. Hodler, *Graph Algorithms: Practical Examples in Apache Spark and Neo4j*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2019.

[33] A. de Bernardi Schneider et al., "StrainHub: A phylogenetic tool to construct pathogen transmission networks," *Bioinformatics*, vol. 36, no. 3, pp. 945–947, Feb. 2020. [Online]. Available: https://academic.oup.com/bioinformatics/article/36/3/945/5550625

[34] L. Igual and S. Seguí, *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*, 1st ed., Cham, Switzerland: Springer, 2017.

[35] E. G. Tuncay, "Graph based methods to retrieve and predict epidemiological statistics," Robert Koch Institute, Berlin, Res. Rep. TR2016/DG/04/A1-01/0844, Sep. 2020.

[36] A. Alcalá, R. Alberich, M. Llabrés, F. Rosselló, and G. Valiente, "AligNet: Alignment of protein-protein interaction networks," *BMC Bioinf.*, vol. 21, no. 6, Nov. 2020, Art. no. 265.

[37] L. Chindelevitch, C.-Y. Ma, C.-S. Liao, and B. Berger, "Optimizing a global alignment of protein interaction networks," *Bioinformatics*, vol. 29, no. 21, pp. 2765–2773, Nov. 2013.

[38] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, no. 1/2, pp. 83–97, Mar. 1955. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/nav.3800020109

[39] E. Kazemi, H. Hassani, M. Grossglauser, and H. P. Modarres, "PROPER: Global protein interaction network alignment through percolation matching," *BMC Bioinf.*, vol. 17, no. 1, Dec. 2016, Art. no. 527. [Online]. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1395-9

[40] V. Vijayan, V. Saraph, and T. Milenković, "MAGNA++: Maximizing accuracy in global network alignment via both node and edge conservation," *Bioinformatics*, vol. 31, no. 14, pp. 2409–2411, Jul. 2015.

[41] H. T. T. Phan and M. J. E. Sternberg, "PINALOG: A novel approach to align protein interaction networks—implications for complex detection and function prediction," *Bioinformatics*, vol. 28, no. 9, pp. 1239–1245, May 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3338015/

[42] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, Jun. 2005. [Online]. Available: https://www.nature.com/articles/nature03607

[43] R. Ibragimov, M. Malek, J. Guo, and J. Baumbach, "GEDEVO: An evolutionary graph edit distance algorithm for biological network alignment," in *Proc. German Conf. Bioinf.*, 2013, Art. no. 12. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2013/4229/

[44] M. Malek, R. Ibragimov, M. Albrecht, and J. Baumbach, "CytoGEDEVO—Global alignment of biological networks with Cytoscape," *Bioinformatics*, vol. 32, no. 8, pp. 1259–1261, Apr. 2016.

[45] A. Elmsallati, A. Msalati, and J. Kalita, "Index-based network aligner of protein-protein interaction networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 1, pp. 330–336, Jan./Feb. 2018.

[46] A. Elmsallati, S. Roy, and J. K. Kalita, "Exploring symmetric substructures in protein interaction networks for pairwise alignment," in *Proc. Int. Conf. Bioinf. Biomed. Eng.*, 2017, pp. 173–184.

[47] D. Park, R. Singh, M. Baym, C.-S. Liao, and B. Berger, "IsoBase: A database of functionally related proteins across PPI networks," *Nucleic Acids Res.*, vol. 39, pp. D295–D300, Jan. 2011.

[48] S. Orchard et al., "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D358–D363, Jan. 2014.
[49] The UniProtConsortium, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, Jan. 2017.
[50] R. P. Huntley et al., "The GOA database: Gene ontology annotation updates for 2015," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1057–D1063, Jan. 2015. [Online]. Available: http://academic.oup.com/nar/article/43/D1/D1057/2437623/The-GOA-database-Gene-Ontology-annotation-updates
[51] C. Stark, "BioGRID: A general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D535–D539, Jan. 2006.
[52] A. Calderone, L. Castagnoli, and G. Cesareni, "mentha: A resource for browsing integrated protein-interaction networks," *Nature Methods*, vol. 10, no. 8, pp. 690–691, Aug. 2013. [Online]. Available: http://www.nature.com/articles/nmeth.2561
[53] B. Aranda et al., "PSICQUIC and PSISCORE: Accessing and scoring molecular interactions," *Nature Methods*, vol. 8, no. 7, pp. 528–529, Jul. 2011. [Online]. Available: https://www.nature.com/articles/nmeth.1637
[54] N. del Toro et al., "A new reference implementation of the PSICQUIC web service," *Nucleic Acids Res.*, vol. 41, pp. W601–W606, Jul. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3977660/
[55] S. Orchard et al., "Protein interaction data curation: The International molecular exchange (IMEx) consortium," *Nature Methods*, vol. 9, no. 4, pp. 345–350, Apr. 2012.
[56] M. Mina and T. Sanavia, "FastSemSim: Fast and easy evaluation of semantic similarity measures on biomedical ontologies," in *Proc. Bioinf. Italian Soc. Annu. Meet.*, 2014, pp. 108–109.
[57] J. Pandey, M. Koyutürk, S. Subramaniam, and A. Grama, "Functional coherence in domain interaction networks," *Bioinformatics*, vol. 24, no. 16, pp. i28–i34, Aug. 2008.
[58] E. Triantaphyllou, *Multi-Criteria Decision Making Methods: A Comparative Study*, P. M. Pardalos and D. Hearn Eds., Boston, MA, USA: Springer, 2000. [Online]. Available: http://link.springer.com/10.1007/978–1-4757-3157-6
[59] L. Bennett, B. Melchers, and B. Proppe, "CURTA: A general-purpose high-performance computer at ZEDAT, Freie Universität Berlin," 2020, Art. no. 5S. [Online]. Available: https://refubium.fu-berlin.de/handle/fub188/26993
[60] V. Vijayan, D. Critchlow, and T. Milenković, "Alignment of dynamic networks," *Bioinformatics*, vol. 33, no. 14, pp. i180–i189, Jul. 2017.
[61] V. Vijayan and T. Milenković, "Aligning dynamic networks with DynaWAVE," *Bioinformatics*, vol. 34, no. 10, pp. 1795–1798, May 2018. [Online]. Available: https://academic.oup.com/bioinformatics/article/34/10/1795/4781097
[62] N. Malod-Dognin, K. Ban, and N. Pržulj, "Unified alignment of protein-protein interaction networks," *Sci. Rep.*, vol. 7, no. 1, Apr. 2017, Art. no. 953. [Online]. Available: https://www.nature.com/articles/s41598–017-01085-9

Erhun Giray Tuncay received the BS degree in mechanical engineering from the Izmir Institute of Technology, in 2003, and the MS degree in computer engineering from Dokuz Eylül University, in 2007. He is currently working toward the PhD degree with the Bioinformatics and Computer Science Programmes, Freie Universität Berlin and he is working as a data scientist & ontology engineer for the TIB - Leibniz Information Centre for Science & Technology. He has worked and studied as a visiting researcher with the Electrical Engineering and Computer Science Department, Kumamoto University in 2010 and with Robert Koch Institute in 2019. His main research interests include artificial intelligence, bioinformatics, and multi-agent systems.

Rıza Cenk Erdur received the BS, MS, and PhD degrees in computer engineering from Ege University. He has worked as a research assistant ('94-'01), assistant professor ('01-'12) and associate professor ('12-ongoing) with the Ege University. His research interests include multi-agent systems, Semantic Web, knowledge engineering, and software engineering.

Tim Conrad received the degree in bioinformatics and computer science from Freie Universität Berlin and Monash University Melbourne, and the PhD degree from the Mathematics Institute of Freie Universität Berlin, in 2008, where he is currently an assistant professor. He is also serving as the Department head for the Mathematics of Complex Systems, Visual and Data-Centric Computing as well as Bioinformatics in Medicine Departments, Zuse Institute Berlin. His research interests include medical informatics, machine learning, network analysis, and proteomics.