Protein complex similarity based on Weisfeiler-Lehman labeling

Bianca K. Stöcker[1,2], Till Schäfer[4], Petra Mutzel[4], Johannes Köster[1,2,3], Nils Kriege[4], and Sven Rahmann[1,4]

[1]**Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, 45147 Essen, Germany**
[2]**Algorithms for Reproducible Bioinformatics, Genome Informatics, Institute of Human Genetics, University Hospital Essen, University of Duisburg-Essen, 45147 Essen, Germany**
[3]**Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston MA 02215, USA**
[4]**Department of Computer Science, TU Dortmund University, 44221 Dortmund, Germany**

Corresponding author:
Sven Rahmann[1]

Email address: Sven.Rahmann@uni-due.de

ABSTRACT

Being able to quantify the similarity between two protein complexes is essential for numerous applications. Prominent examples are database searches for known complexes with a given query complex, comparison of the output of different protein complex prediction algorithms, or summarizing and clustering protein complexes, e.g., for visualization. While the corresponding problems have received much attention on single proteins and protein families, the question about how to model and compute similarity between protein complexes has not yet been systematically studied. Because protein complexes can be naturally modeled as graphs, in principle general graph similarity measures may be used, but these are often computationally hard to obtain and do not take typical properties of protein complexes into account. Here we propose a parametric family of similarity measures based on Weisfeiler-Lehman labeling. We evaluate it on simulated complexes of the extended human integrin adhesome network. Because the connectivity (graph topology) of real complexes is often unknown and hard to obtain experimentally, we use both known protein-protein interaction networks and known interdependencies (constraints) between interactions to simulate more realistic complexes than from interaction networks alone. We empirically show that the defined family of similarity measures is in good agreement with edit similarity, a similarity measure derived from graph edit distance, but can be much more efficiently computed. It can therefore be used in large-scale studies and simulations and serve as a basis for further refinements of modeling protein complex similarity.

INTRODUCTION

Proteins fulfill manifold tasks in living cells, but they rarely act alone. Indeed, most cellular functions are enabled only when proteins physically interact with other proteins, forming protein complexes. DNA transcription is a typical example, where RNA polymerase II, general transcription factors, cell type specific transcription regulators and mediator proteins interact.

Understanding protein complex formation and function is one of the big challenges of cell biology, approached by both experimental techniques and computational modeling. While the constituent protein sequences can be obtained from the genome (even that can be challenging), the computational prediction of real protein complexes from protein interaction networks appears to be much more difficult as evidenced by the recent literature on the topic; see Bhowmick and Seah (2016) for a survey, or Srihari et al. (2017) for a textbook introduction. Fortunately, new experimental technologies are about to enhance our understanding of complexes significantly in the near future, e.g. high-resolution protein-protein docking (Park et al., 2015; Vakser, 2014; Kozakov et al., 2017; Wass et al., 2011). Large scale generation of libraries of cell

lines having two or more endogenously tagged fluorescent proteins (Boutros et al., 2015) and recent high-throughput and multiplexed implementations of fluorescence correlation spectroscopy allow us to systematically measure endogenous concentrations, binding constants and high-order complexes in such libraries of cell lines (Hwang et al., 2006; Wobma et al., 2012; Grecco et al., 2016; Wachsmuth et al., 2015).

When studying biological entities such as protein sequences or protein complexes, a fundamental task is to define a measure of similarity between two such entities. For protein sequences, there is a well-established theory based on scoring matrices and alignment scores (Pearson, 2013). For protein complexes, it appears that no systematic effort to quantify similarity has been made yet. The purpose of the present article is therefore to discuss the different options to define a similarity measure on protein complexes and to propose a reasonable and computationally tractable definition of protein complex similarity.

Establishing a similarity measure is not only important fundamentally, but there are many immediate applications, of which we mention the following three examples.

Database search: In the *database search problem* we are given a query complex and a large collection (database) of complexes, and the task is to find the complexes in the database that are most similar to the query. Obviously, a meaningfully defined similarity measure is essential for this task.

Comparing predictions: Several complex prediction methods predict putative complexes by locating dense regions in a protein interaction network (Drew et al., 2017; Hernandez et al., 2017; Ma and Gao, 2012; Pellegrini et al., 2016), and for comparing complexes predicted by different algorithms, it is of interest to compute a maximum-weight matching between the output of two algorithms, where the weighting is given by a similarity function.

Summarizing and clustering: When stochastically simulating complex formation based on available knowledge such as possible interactions and interaction constraints, it is helpful to aggregate the simulation output to focus on frequently seen or typical complexes, ignoring small differences. Aggregation or clustering by similarity thereby reduces data size and complexity. Such a task first and foremost requires a way to quantify the similarity between two protein complexes.

When there are many (say, tens of thousands of) different complexes subject to pairwise comparison, a similarity measure must be efficiently computable to be of practical interest.

Models for protein complexes. We first need to discuss models for protein complexes on different levels of detail, namely the *set*, *multiset*, and *graph* models.

While intuition suggests that protein complexes can be naturally described as graphs with proteins as vertices and physical interactons as edges, there are in fact different ways to formally describe a protein complex. In the following, we briefly mention the most prominent ones with their advantages and disadvantages. We start with a given set P of all proteins of an organism, the building blocks of the complexes.

Set: In its most simple form, a protein complex can be defined as a set (in the mathematical sense, i.e., without multiplicities) of proteins, i.e., as a subset of P. We use the standard notation $\{p_1, p_2, \ldots, p_n\}$ for sets. Sets neither capture the multiplicities nor the nature of the physical interactions between the constituent proteins of a complex. However, some experimental techniques (e.g. immunoprecipitation with mass spectrometry) only give such set-type information, and several existing databases only provide this type of information, e.g. the CORUM database provided by the Munich Information center for Protein Sequences MIPS (Ruepp et al., 2010).

Multiset: Formally, a multiset is a function $C : P \to \mathbb{N}_0$ that assigns a multiplicity to each protein $p \in P$ with $C(p) = 0$ for proteins p that are not part of the complex. We also use the multiset notation $C = \{\!\{p_1, p_1, p_2\}\!\}$ to express that $C(p_1) = 2$, $C(p_2) = 1$ and $C(p) = 0$ for all other $p \in P$. Defining a protein complex as a multiset of proteins gives a more accurate representation of the complex, but still does not consider the interaction topology.

Graph: To add more information, we can define a protein complex as an undirected graph $C = (V, E, \ell)$ with labeled vertices V, such that each vertex $v \in V$ represents a protein and hence has a label

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

$\ell(v) \in P$, each edge $e \in E \subseteq V \times V$ represents a physical interaction between the corresponding proteins, such that E is symmetric and C is connected. The graph description provides the interaction topology. We call this representation a *protein complex graph* and define its *size* as $|C| := |V| + |E|$. (This representation could be further refined by considering the different domains of each protein and specifying precisely which domains interact.)

For the set and multiset models, a similarity measure is readily given by the *Jaccard similarity* (see Methods). For graphs, the *graph edit distance* has been proposed for pattern recognition tasks more than 30 years ago (Sanfeliu and Fu, 1983). A graph edit distance between graphs C and C' measures the total costs of the edit operations required to transform C into C'. Defining similarity via graph edit operations appears intuitive, but has computational disadvantages, as edit distance computation on graphs is hard in general. More specifically, the graph edit distance generalizes the classical maximum common subgraph problem (Bunke, 1997), which is NP-complete (Garey and Johnson, 1979) and hard to approximate with given guarantees (Kann, 1992). Recently, a binary linear programming formulation for computing the graph edit distance has been proposed (Lerouge et al., 2017), which allows to compare graphs of moderate size using state-of-the art general purpose solvers. However, when we want to compare many complexes, evaluating the edit distance between all pairs becomes infeasible in practice.

In this article, we therefore propose an efficient alternative: We define a family of similarity measures on graphs by resorting to the (efficiently computable) Jaccard similarity, while still taking the graph structure into account. This is achieved by so-called Weisfeiler-Lehman labeling of the vertices (Weisfeiler and Lehman, 1968), propagating vertex labels between neighbors. This approach is different from recent work that approximates and bounds the graph edit distance (Riesen et al., 2015) and has the advantage of scaling better to large-scale studies.

The remainder of the article is structured as follows. In the Methods section, we define a parametric family of similarity measures based on Weisfeiler-Lehman labeling and the precise definition of graph edit similarity we compare against. In the Results section, we describe how we obtain pairs of protein complexes, for which we compare Weisfeiler-Lehman similarity and edit similarity. The simulated protein complexes take known protein interaction networks and additionally constraints between interactions into account, and therefore should represent more realistic complexes than arbitrary connected subgraphs of protein interaction networks. Finally, we discuss limitations and possible extensions of this work.

METHODS

Our goal is to define a similarity measure between protein complexes that captures not only the (multisets of the) constituent proteins, but also the interaction topology (graph structure). Similarities derived from graph edit distance offer this property, but as mentioned above, they are hard to compute. Therefore, we introduce a parameterized family of similarity measures on protein complexes, which are based on multiset comparisons of vertex labels in the complex graph and take the local neighborhood of each protein into account by using Weisfeiler-Lehman labels.

Jaccard similarity of sets and multisets

To compare sets or multisets, Jaccard similarity coefficients are an established measure.

Let $M \subseteq U$ and $M' \subseteq U$ be two subsets of a common universe U. Then the *Jaccard similarity* between M and M' is defined as

$$J_{\text{set}}(M, M') := \frac{|M \cap M'|}{|M \cup M'|} \in [0,1]\,. \tag{1}$$

This definition is extended to multisets as follows. Recall that multisets M and M' are functions $U \to \mathbb{N}_0$, assigning multiplicities $M(o)$ and $M'(o)$ to each object $o \in U$. (The set definition can be seen as the special case where the value set is only $\{0,1\}$ instead of $\mathbb{N}_0$.) Then the *Jaccard similarity* between M and M' is defined as

$$J_{\text{multiset}}(M, M') := \frac{\sum_{o \in U} \min\{M(o), M'(o)\}}{\sum_{o \in U} \max\{M(o), M'(o)\}} \in [0,1]\,. \tag{2}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

A parametric family of protein complex similarity measures

Instead of comparing the protein complexes directly by their graph topology and labeling, we extract and compare multisets of features of the protein complexes. Weisfeiler and Lehman (1968) developed an iterative label refinement procedure to derive a canonical graph representation for graph isomorphism testing. The same procedure is often used to define graph similarities or graph kernels (Shervashidze et al., 2011). For the latter purpose, the vertex labels of each Weisfeiler-Lehman iteration are used as features of the graphs. Initially, the feature multiset of a graph consists of the union of all vertex labels. After the initialization, the vertex labels are iteratively augmented by the labels of the neighboring vertices from the previous iteration, thereby encoding the (local) graph structure in the vertex labels. Let us now formally define the process.

Definition 1 (Weisfeiler-Lehman labeling of iteration i for a protein complex graph)**.** Let $C = (V, E, \ell_0)$ be a protein complex graph with label function $\ell_0 : V \to L_0 := P$. Furthermore, let $N(v) := \{u \mid \{v,u\} \in E\}$ denote the neighbors of vertex $v \in V$. Then, the Weisfeiler-Lehman labeling of iteration i is defined as a re-labeling of the protein complex graph: It replaces the labeling function $\ell_0 : V \to L_0$ with a labeling function $\ell_i : V \to L_i$. The value of ℓ_i for a vertex $v \in V$ is recursively defined as

$$\ell_i(v) := (\ell_{i-1}(v), \{\!\{\ell_{i-1}(u) \mid u \in N(v)\}\!\}). \tag{3}$$

Note that the second component of the new label is a multiset.

To avoid that the length of labels increases in each iteration, label compression is performed after each step in practice. This is achieved by a one-to-one mapping of the labels $\{\ell_i(v) \mid v \in V\}$ to integer labels.

Given the Weisfeiler-Lehman labeling function of a protein complex graph for some iteration i, we can now define the multiset of Weisfeiler-Lehman features for iteration i.

Definition 2 (Weisfeiler-Lehman feature set of iteration i for a protein complex graph)**.** Let $C = (V, E, \ell_0)$ be a protein complex graph with label function $\ell_0 : V \to L_0 = P$. Then, the Weisfeiler-Lehman features of iteration i are defined as multiset $\mathrm{WL}_i(C) = \{\!\{l_i(v) \mid v \in V\}\!\}$. Note that $\mathrm{WL}_0(C)$ corresponds to the initial multiset of protein names.

To compare two complexes C and C', we compare the iteration sequences of Weisfeiler-Lehman features $(\mathrm{WL}_i(C))_{i\geq 0}$ and $(\mathrm{WL}_i(C'))_{i\geq 0}$, by computing a convex combination of the Jaccard similarities for each iteration. Let $w = (w_i)_{i\geq 0}$ be a weight sequence with $w_i \geq 0$ for all $i \geq 0$ and $\sum_{i\geq 0} w_i = 1$. For w as just defined, let

$$S_w(C, C') := \sum_{i\geq 0} w_i \cdot J_{\mathrm{multiset}}(\mathrm{WL}_i(C), \mathrm{WL}_i(C')), \tag{4}$$

where J_{multiset} is given by Eq. (2). This defines a family of similarity measures between complexes with values in $[0,1]$, parameterized by a convex combination $w = (w_0, w_1, \dots)$.

It is easy to see that, as long as $w_0 > 0$, we have $S_w(C, C') = 0$ if and only if the protein sets of C and C' are disjoint. If $S_w(C, C') < 1$, the protein complex graphs are not isomorphic. However, $S_w(C, C') = 1$ does not necessarily imply that C and C' are isomorphic even if $w_i > 0$ for all i: There exist examples of non-isomorphic graphs G, G' with $\mathrm{WL}_i(G) = \mathrm{WL}_i(G')$ for all $i \geq 0$. (As a simple example, take G to be a cycle of six vertices, and G' to be two cycles of three vertices, all with the same label.) On the other hand, there exist classes of graphs, such as the so-called CR-graphs, for which the implication "$S_w(C, C') = 1 \Rightarrow C, C'$ are isomorphic" is true if $w_i > 0$ for all i (Arvind et al., 2015).

In practice, we may assume that most protein complexes are non-adversarial graphs with sufficiently simple structure such that their Weisfeiler-Lehman features are appropriate to characterize their similarity. In fact, we put forward the hypothesis that using a single iteration is frequently sufficient for practical purposes, and we set $w_i := 0$ for $i \geq 2$ in our computational experiments (see Results) and only have a single free parameter $w_0 \in [0,1]$ that defines $w_1 := 1 - w_0$. In the following, we write ω for w_0. In this case, S_ω is efficiently computable: A proof of the following lemma can be found in the work of Shervashidze et al. (2011).

Lemma 3. *For $\omega \in [0,1]$, each of the one-parameter similarity measures*

$$S_\omega(C, C') := \omega \cdot J_{multiset}(WL_0(C), WL_0(C')) + (1 - \omega) \cdot J_{multiset}(WL_1(C), WL_1(C'))$$

can be computed in $O(|C| + |C'|)$ time, where $|C| = |V| + |E|$.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

A similarity measure based on the graph edit distance

To compare the family of Weisfeiler-Lehman multiset-based similarity measures defined above with graph edit distance, we start with a formal definition of the edit-based similarity. We allow the following elementary operations to edit a graph: vertex deletion, vertex insertion, vertex relabeling, edge deletion, and edge insertion. A sequence $(o_1,\dots,o_k)$ of such edit operations that transforms a graph G into another graph H is called an *edit path* from G to H. Each operation o is assigned a cost $c(o)$, which is zero for substituting vertices and edges with the same label. We use a cost of 1 for all operations except vertex relabeling which has a cost of 2, corresponding to one deletion and one insertion (leaving the edges in place). Note that deleting or inserting a vertex of degree k otherwise has cost $k+1$ for deleting k edges and the vertex itself. We denote the set of all possible edit paths from G to H by $\Upsilon(G,H)$.

Definition 4. Let G and H be labeled graphs. The *graph edit distance* from G to H is defined by

$$d(G,H) = \min\left\{ \sum_{i=1}^{k} c(o_i) \;\middle|\; (o_1,\dots,o_k) \in \Upsilon(G,H) \right\}. \tag{5}$$

Intuitively, the graph edit distance preserves a subgraph G' of G that is also contained in H using zero-cost substitutions, deletes the vertices and edges in G that are not in G' and then inserts vertices and edges to obtain an isomorphic copy of H. Therefore all non-zero costs can be attributed to the elements which are in one of the graphs, but not in their common subgraph. In this sense the graph edit distance is similar to the symmetric difference of two sets. This observation motivates the following normalized similarity measure derived from the graph edit distance. We define the *graph edit similarity* as

$$J_{\text{graph}}(G,H) := \frac{|G|+|H|-d(G,H)}{|G|+|H|+d(G,H)} \in [0,1], \tag{6}$$

where $|G| := |V(G)|+|E(G)|$. Note that the graph edit distance between G and H is at most $|G|+|H|$, which is achieved by deleting all vertices and edges of G and inserting all vertices and edges of H. In this case the graph edit similarity is zero. Similarly, $J_{\text{graph}}(G,H) = 1$ if and only if $d(G,H) = 0$. In this respect the similarity measure resembles the Jaccard similarity. In fact, we can show a deeper relation to the multiset Jaccard similarity.

Lemma 5. *For two protein complexes, let C, D denote their protein multisets and G, H their protein complex graphs. For the edge-free graphs $G' = (V(G),\emptyset)$ and $H' = (V(H),\emptyset)$ it holds that $J_{graph}(G',H') = J_{multiset}(C,D)$.*

Proof. An optimal graph edit path is obtained as follows: We substitute the vertices with common labels free of cost, which are $Z = \sum_{p\in P} \min\{C(p),D(p)\}$ in total. We delete the remaining $|G'|-Z$ vertices in G' and insert $|H'|-Z$ vertices to obtain an isomorphic copy of H' at a total cost of $|G'|+|H'|-2Z = d(G,H)$. Instead we may also substitute up to $||G'|-|H'||$ vertices, each at cost two, which results in the same total cost. Using the fact that $|G'| = \sum_{p\in P} C(p)$ and $|H'| = \sum_{p\in P} D(p)$, we obtain the result by calculating

$$\begin{aligned} J_{\text{graph}}(G',H') &= \frac{|G'|+|H'|-d(G',H')}{|G'|+|H'|+d(G',H')} = \frac{Z}{|G'|+|H'|-Z} = \frac{Z}{\sum_{p\in P} C(p)+\sum_{p\in P} D(p)-Z} \\ &= \frac{\sum_{p\in P}\min\{C(p),D(p)\}}{\sum_{p\in P} C(p)+D(p)-\min\{C(p),D(p)\}} = \frac{\sum_{p\in P}\min\{C(p),D(p)\}}{\sum_{p\in P}\max\{C(p),D(p)\}} = J_{\text{multiset}}(C,D). \end{aligned}$$

□

Lemma 5 shows that the graph edit similarity can indeed be seen as a natural extension of the multiset Jaccard similarity to graph structured data.

For our computations, we used a recent binary linear programming formulation to compute the graph edit distance exactly (Lerouge et al., 2017). The approach was implemented in Java, and all instances were solved using an academic license of Gurobi 7.5.2 on Linux x86-64.

RESULTS

Hypothesis

We hypothesize that the Weisfeiler-Lehman based family of similarity measures S_ω defined in Eq. (4) approximates well the edit distance based similarity defined in Eq. (6) for typical protein complexes. The similarity measures S_ω have the advantage that they can be efficiently computed (see Lemma 3).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

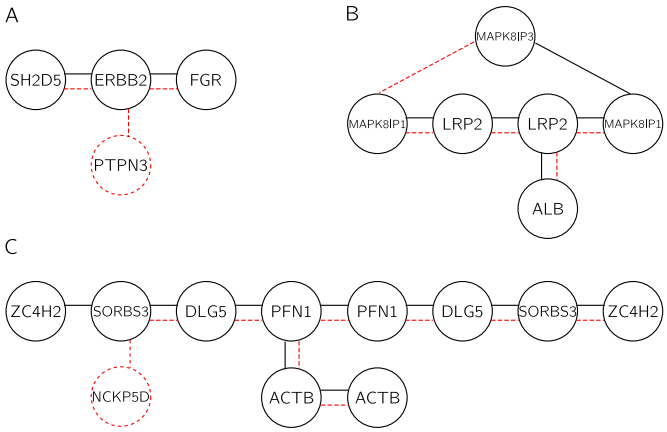


Figure 1. Three exemplary pairs of protein complexes: Each labeled node is a protein instance, each edge represents a protein interaction, and solid black vs. dashed red edges distinguish between the two complexes. **A:** Edit similarity 0.714; WL similarity in $[0.4, 0.75]$ depending on weight ω. **B:** Edit similarity 0.838; WL similarity 1.0 (independent of ω). **C:** Edit similarity 0.9; WL similarity in $[0.667, 0.818]$ depending on ω.

Data generation

As mentioned in the Introduction, obtaining real protein complex graphs is difficult at the moment, because experimental techniques that resolve the (graph) topology of the complexes are only being developed. Therefore we resort to the simulation of complexes, based on two types of knowledge: possible protein-protein interactions, formalized by a *protein interaction network*, and *constraints between protein interactions*.

Formally, a protein interaction network is an undirected graph $N = (P, I)$, where P is the set of protein types of a cell (or an organism), and $I \subset P \times P$ indicates the pairs of protein types that may potentially physically interact. Since N describes the entirety of possible interactions, any protein complex can be seen as a connected subgraph of N.

It is important to realize that protein interactions are not independent of each other, but interdependent. Those interaction dependencies are generated by two major mechanisms. On the one hand there is allosteric regulation, in which the capability of a protein to bind other proteins is affected by a conformational change upon one interaction (Laskowski et al., 2009). The other key mechanism is steric hindrance that prevents proteins from binding simultaneously to too close or identical protein domains leading to mutual exclusiveness of interactions (Sánchez Claros and Tramontano, 2012). The dependencies between interactions constrain the set of possible protein complexes and their assembly paths. Therefore, for understanding the design and function of intracellular protein networks it is important to consider the dependencies between protein interactions. One possible model for this are *constrained protein interaction networks*, where the protein interaction network is enhanced by the interaction dependencies (*constraints*) modeled as propositional logic formulas (Stöcker et al., 2017).

With constrained protein interaction networks, we can stochastically simulate complex formation based on the available knowledge and obtain a detailed interaction topology (which proteins physically interact) for each complex.

To evaluate the Weisfeiler-Lehman based similarity ("WL similarity") against the edit distance based similarity ("edit similarity"), we computed both similarity measures on selected pairs of 100 000 simulated protein complexes from the extended human adhesome network as presented by Stöcker et al. (2017). Since edit similarity computations are computationally costly, we only computed the edit similarity on 500 000 candidate pairs from these simulated complexes. These candidate pairs were generated for all pairs of complexes that have at most 20 proteins (larger complexes are so rare that high similarities are unlikely), that have a size difference of protein multisets of at most 10, and that share at least one protein. The candidate pairs were sorted descendingly after the number of shared proteins and then the edit distance based similarity was computed on the first 500 000 candidate pairs. The resulting edit similarity values were classified into bins of width 0.1. Because most pairs of complexes share a small

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

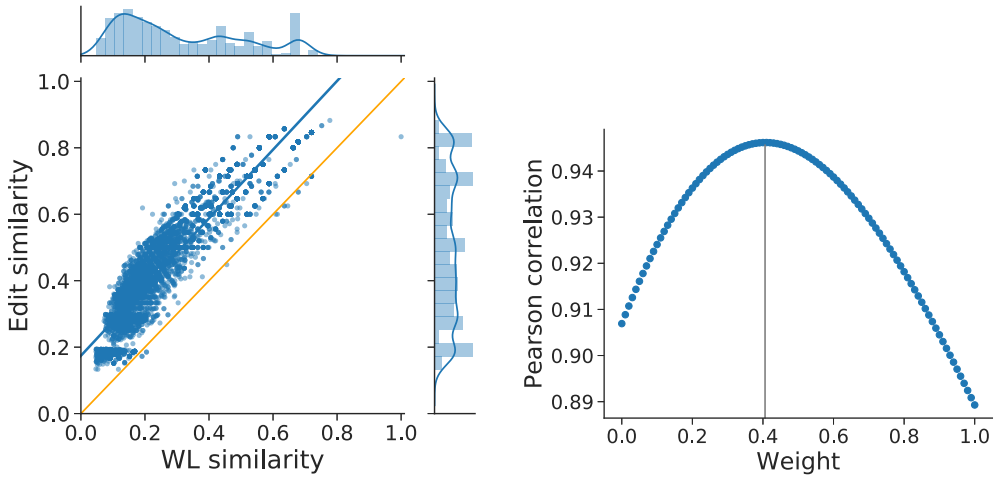


Figure 2. Left: Scatterplot comparison of edit similarity and WL similarity for weight $\omega = 0.41$, including marginal distributions and least-squares regression line. **Right:** Pearson correlation between edit similarity and WL similarity as a function of weight ω. The maximum correlation (0.946) occurs for $\omega = 0.41$, shown on the left side.

number of proteins, we find many pairs with small edit similiarity (but none in the range $[0.0, 0.1[$ because we required one common protein) and comparatively few pairs with edit similarity above 0.5. To achieve a uniform distriution among bins for the comparison, we randomly selected 1000 pairs from each bin, excluding the bin $[0.9, 1.0[$ which contained a single pair. This yielded 8000 pairs of complexes from 8 bins.

Similarity comparison

We first consider three exemplary pairs (Figure 1 A–C) with edit similarities of approximately 0.7, 0.8 and 0.9, respectively, the latter being the most similar observed pair.

In example A, an additional protein (PTPN3) is added to an existing complex, a linear chain of 3 proteins. The edit similariy is $10/14 = 0.714$, the WL similarity is between 0.75 for $\omega = 1$ and and 0.4 for $\omega = 0$. Because the edit similarity is between the extreme WL similarities, there exists a unique weight $\omega^* \approx 0.898$, for which WL and edit similarities agree for this particular complex pair. Example B is an noteworthy case, because the WL similarity is 1.0, independent of ω, because the vertex labels are identical even after the first Weisfeiler-Lehman iteration. (Further iterations would show a difference.) The edit similarity is $20/24 = 0.83$, which is obtained by attaching ALB to the other LRP2 protein. In example *C*, one protein is replaced by another one in a fairly large complex. The edit similarity (0.905) is relatively high and outside the WL similarity range between 0.667 for $\omega = 0$ and 0.818 for $\omega = 1$.

Because most protein complexes are small and do not exhibit properties of examples B or C, the overall agreement between WL similarity and edit similairty is high: For each of the selected complex pairs, we computed the exact edit similarity and the WL similarity for each weight $\omega \in W := \{0.0, 0.01, 0.02, \ldots, 1.0\}$. Let e be the vector of edit similarity values and $s(\omega)$ the corresponding vector of WL similarity values using weight ω. To compare the similarity measures, we calculated both the Pearson correlation coefficient and the cosine similarity of e and $s(\omega)$ for all $\omega \in W$. As can be seen from Figure 2, the highest values occur for ω between 0.38 and 0.44 and the maximum Pearson correlation coefficient is obtained for $\omega = 0.41$. For the cosine similarity, the maximum value is reached for weight $\omega = 0.69$, but the function is less peaked, and values above 0.4 lead to high agreement (Figure 3).

Overall, we find good agreement between edit similarity and WL similarity for sufficiently large values of ω, i.e., if the Jaccard similairity of the constituent protein multiset has sufficiently high weight.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

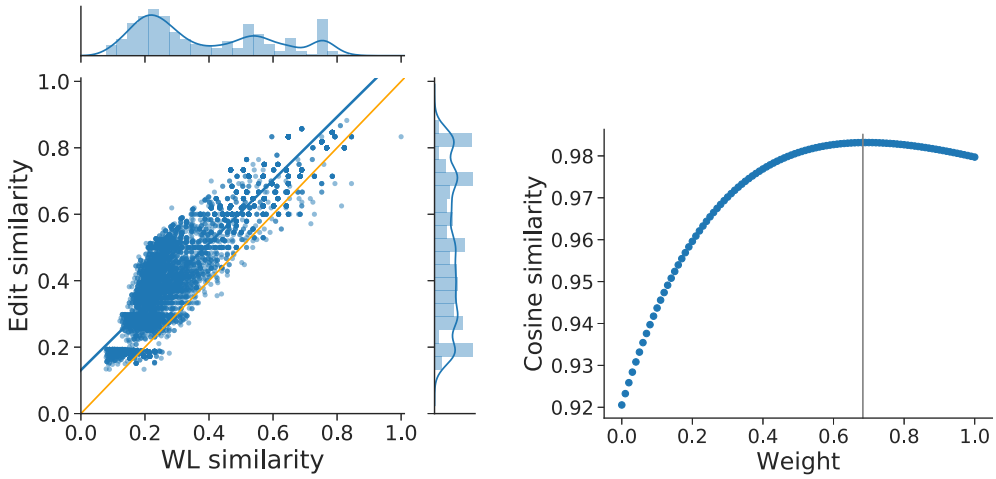


Figure 3. Left: Scatterplot comparison of edit similarity and WL similarity for weight $\omega = 0.69$, including marginal distributions and least-squares regression line. **Right:** Cosine similarity between edit similarity and WL similarity as a function of weight ω. The maximum cosine similarity (0.983) occurs for $\omega = 0.69$ shown on the left side.

Reproducibility

The performed data analysis is available as a reproducible Snakemake (Köster and Rahmann, 2012) workflow[1].

DISCUSSION

Our motivation to consider protein complex similarity was to reduce the complexity of the simulation output of our constrained protein interaction network simulator (Stöcker et al., 2017), and we were surprised to see that apparently, no similarity measures have been proposed in the literature. Depending on the underlying representation (set, multiset or graph), different alternatives suggest themselves. However, most graph-based measures are both theoretically and practically hard to compute for larger complexes or for large amounts of complexes. While different tractable graph similarity measures have been proposed, e.g. by Conte et al. (2004) or by Vishwanathan et al. (2010), or approximate graph edit distance (Riesen et al., 2015), none of these appear to be specifically tailored to the properties of protein complexes (often less than ten vertices; sparse). Our proposal to define the similarity as a convex combination of two Jaccard coefficients (protein label multiset and Weisfeiler-Lehman label multiset after one iteration) has two additonal advantages. First, using Jaccard coefficients allows to efficiently pre-filter for high similarity using locality sensitive hashing. Second, for weight $\omega = 1$ of the 0-th WL iteration, our measure reduces to the natural similarity measure of the multiset representation. Our framework hence allows for a smooth transition between multiset and graph representation. The comparison to an edit-based similarity seems to indicate that the protein label multiset plays an important role if one wants to approximate the edit similarity.

From a biological point of view, a high similarity between two complexes should indicate a high probability that they share the same function and can substitute each other in a cellular process. If such information were available, we could evaluate each similarity measure with regard to how it relates to common function. At present, when not even the interaction topology of most complexes has been determined, the corresponding data is out of reach, and such an evaluation is not feasible. In this situation, we suggest that edit similarity is a measure that corresponds to intuition about similarity and that any reasonable similarity measure should be close to edit similarity. The measure we propose has this property (for any weight $\omega \in [0, 1]$) but offers the advantage that it can be quickly computed and scales to millions of complex pairs.

Both WL similarity and edit similarity, as previously defined, have limitations from a biological point

[1] https://doi.org/10.5281/zenodo.1178084

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

of view in the sense that they do not consider similarities between proteins: Two proteins are either equal or distinct. However, if two proteins are closely related, should they be treated as equal or distinct? In the former case, we lose resolution. In the latter case, we would benefit from a fine-grained similarity function between proteins (e.g. a modification of p is very similar to p, a protein with some common domains is somewhat similar to p, but a completely disjoint protein in terms of domains has similarity zero). In this sense, the question of how to best measure protein complex similarity is far from settled.

FUNDING

This work was supported by Deutsche Forschungsgemeinschaft (DFG) Collaborative Research Center (SFB) 876, projects A6 and C1, and by Mercator Research Center Ruhr (MERCUR), project Pe-2013-0012 (UA Ruhr professorship).

REFERENCES

Arvind, V., Köbler, J., Rattan, G., and Verbitsky, O. (2015). On the power of color refinement. In Kosowski, A. and Walukiewicz, I., editors, *Fundamentals of Computation Theory - 20th International Symposium, FCT 2015, Gdańsk, Poland, August 17–19, 2015, Proceedings*, volume 9210 of *Lecture Notes in Computer Science*, pages 339–350. Springer.

Bhowmick, S. S. and Seah, B. (2016). Clustering and summarizing protein-protein interaction networks: A survey. *IEEE Trans. Knowl. Data Eng.*, 28(3):638–658.

Boutros, M., Heigwer, F., and Laufer, C. (2015). Microscopy-based high-content screening. *Cell*, 163(6):1314–25.

Bunke, H. (1997). On a relation between graph edit distance and maximum common subgraph. *Pattern Recognition Letters*.

Conte, D., Foggia, P., Sansone, C., and Vento, M. (2004). Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(3):265–298.

Drew, K., Lee, C., Huizar, R. L., Tu, F., Borgeson, B., McWhite, C. D., Ma, Y., Wallingford, J. B., and Marcotte, E. M. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol Syst Biol*, 13(6):932.

Garey, M. and Johnson, D. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Grecco, H. E., Imtiaz, S., and Zamir, E. (2016). Multiplexed imaging of intracellular protein networks. *Cytometry A*, 89(8):761–75.

Hernandez, C., Mella, C., Navarro, G., Olivera-Nappa, A., and Araya, J. (2017). Protein complex prediction via dense subgraphs and false positive analysis. *PLoS One*, 12(9):e0183460.

Hwang, L. C., Gösch, M., Lasser, T., and Wohland, T. (2006). Simultaneous multicolor fluorescence cross-correlation spectroscopy to detect higher order molecular interactions using single wavelength laser excitation. *Biophys J*, 91(2):715–27.

Kann, V. (1992). On the approximability of the maximum common subgraph problem. In *Proceedings of the 9th Annual Symposium on Theoretical Aspects of Computer Science (STACS'92)*, volume 577 of *LNCS*, pages 377–388, London, UK, UK. Springer-Verlag.

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., and Vajda, S. (2017). The cluspro web server for protein-protein docking. *Nat Protoc*, 12(2):255–278.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

Laskowski, R. A., Gerick, F., and Thornton, J. M. (2009). The structural basis of allosteric regulation in proteins. *FEBS Letters*, 583(11):1692–1698.

Lerouge, J., Abu-Aisheh, Z., Raveaux, R., Héroux, P., and Adam, S. (2017). New binary linear programming formulation to compute the graph edit distance. *Pattern Recognition*.

Ma, X. and Gao, L. (2012). Discovering protein complexes in protein interaction networks via exploring the weak ties effect. *BMC Syst Biol*, 6 Suppl 1:S6.

Park, H., Lee, H., and Seok, C. (2015). High-resolution protein-protein docking by global optimization: recent advances and future challenges. *Curr Opin Struct Biol*, 35:24–31.

Pearson, W. R. (2013). Selecting the Right Similarity-Scoring Matrix. *Curr Protoc Bioinformatics*, 43:1–9.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

Pellegrini, M., Baglioni, M., and Geraci, F. (2016). Protein complex prediction for large protein protein interaction networks with the core&peel method. *BMC Bioinformatics*, 17(Suppl 12):372.

Riesen, K., Ferrer, M., and Bunke, H. (2015). Approximate Graph Edit Distance in Quadratic Time. *IEEE/ACM Trans Comput Biol Bioinform*. Epub ahead of print.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes – 2009. *Nucleic acids research*, 38(suppl 1):D497–D501.

Sánchez Claros, C. and Tramontano, A. (2012). Detecting mutually exclusive interactions in protein-protein interaction maps. *PLoS One*, 7(6):e38765.

Sanfeliu, A. and Fu, K. S. (1983). A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*.

Shervashidze, N., Schweitzer, P., van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. (2011). Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561.

Srihari, S., Yong, C. H., and Wong, L. (2017). *Computational Prediction of Protein Complexes from Protein Interaction Networks*. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA.

Stöcker, B. K., Köster, J., Zamir, E., and Rahmann, S. (2017). Modeling and simulating networks of interdependent protein interactions. *bioRxiv*.

Vakser, I. A. (2014). Protein-protein docking: from interaction to interactome. *Biophys J*, 107(8):1785–93.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242.

Wachsmuth, M., Conrad, C., Bulkescher, J., Koch, B., Mahen, R., Isokane, M., Pepperkok, R., and Ellenberg, J. (2015). High-throughput fluorescence correlation spectroscopy enables analysis of proteome dynamics in living cells. *Nat Biotechnol*, 33(4):384–9.

Wass, M. N., Fuentes, G., Pons, C., Pazos, F., and Valencia, A. (2011). Towards the prediction of protein interaction partners using physical docking. *Molecular systems biology*, 7:469.

Weisfeiler, B. and Lehman, A. A. (1968). A reduction of a graph to a canonical form and an algebra arising during this reduction. *Nauchno-Technicheskaya Informatsiya*, 2(9):12–16. In Russian.

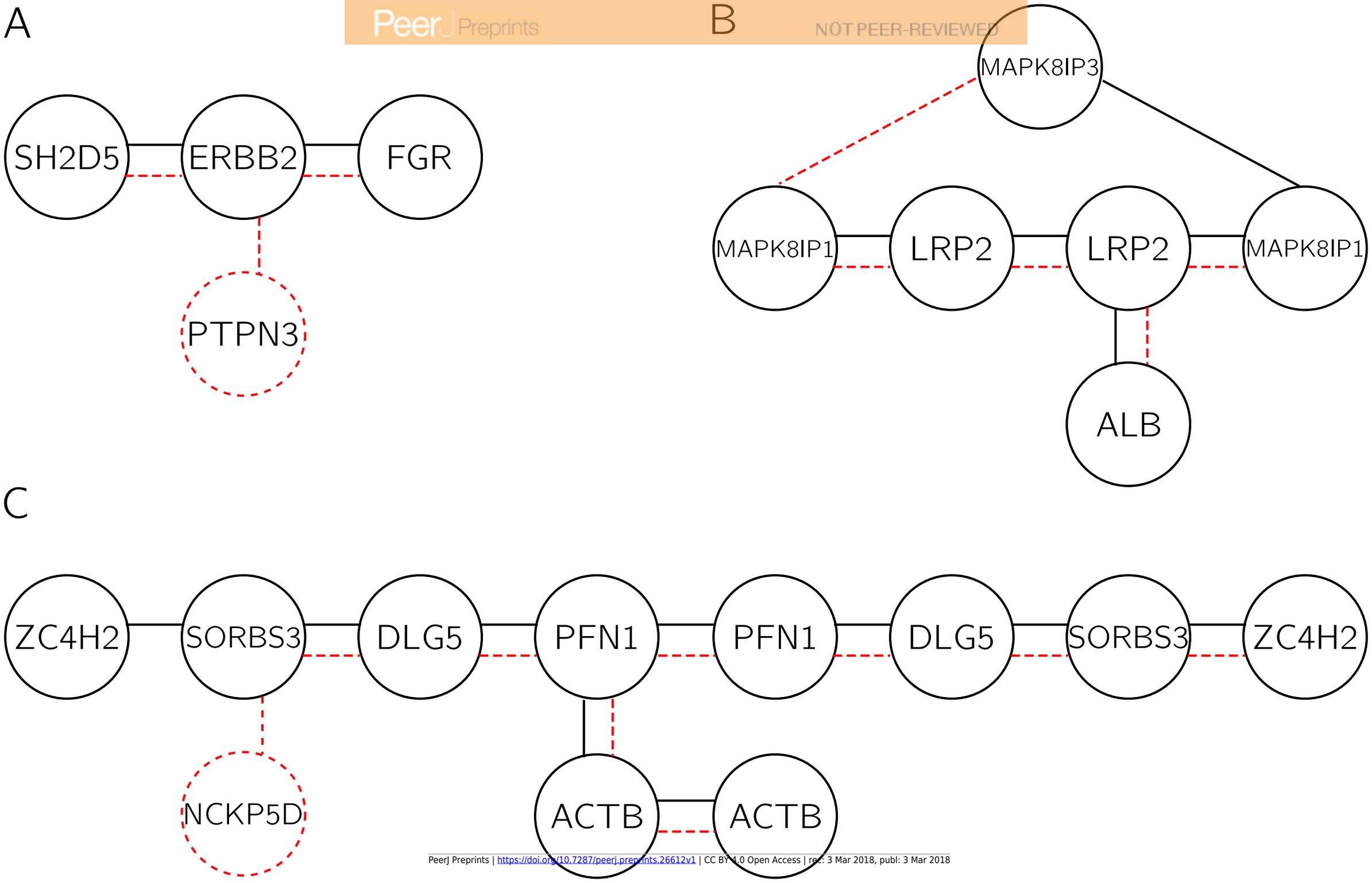
Wobma, H. M., Blades, M. L., Grekova, E., McGuire, D. L., Chen, K., Chan, W. C. W., and Cramb, D. T. (2012). The development of direct multicolour fluorescence cross-correlation spectroscopy: towards a new tool for tracking complex biomolecular events in real-time. *Phys Chem Chem Phys*, 14(10):3290–4.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

Figure 1(on next page)

Three exemplary pairs of protein complexes.

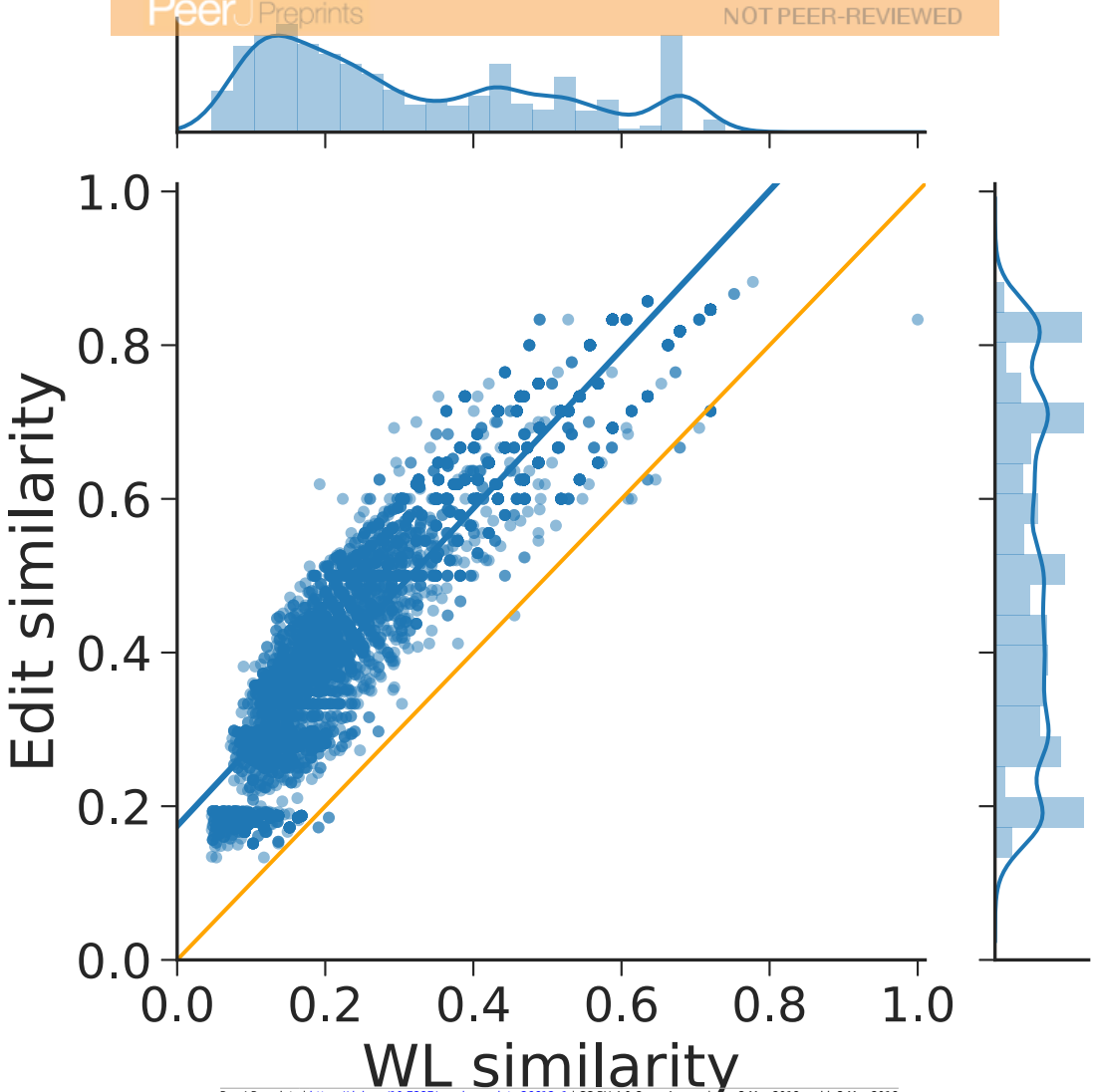
Each labeled node is a protein instance, each edge represents a protein interaction, and solid black vs. dashed red edges distinguish between the two complexes.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

Figure 2(on next page)

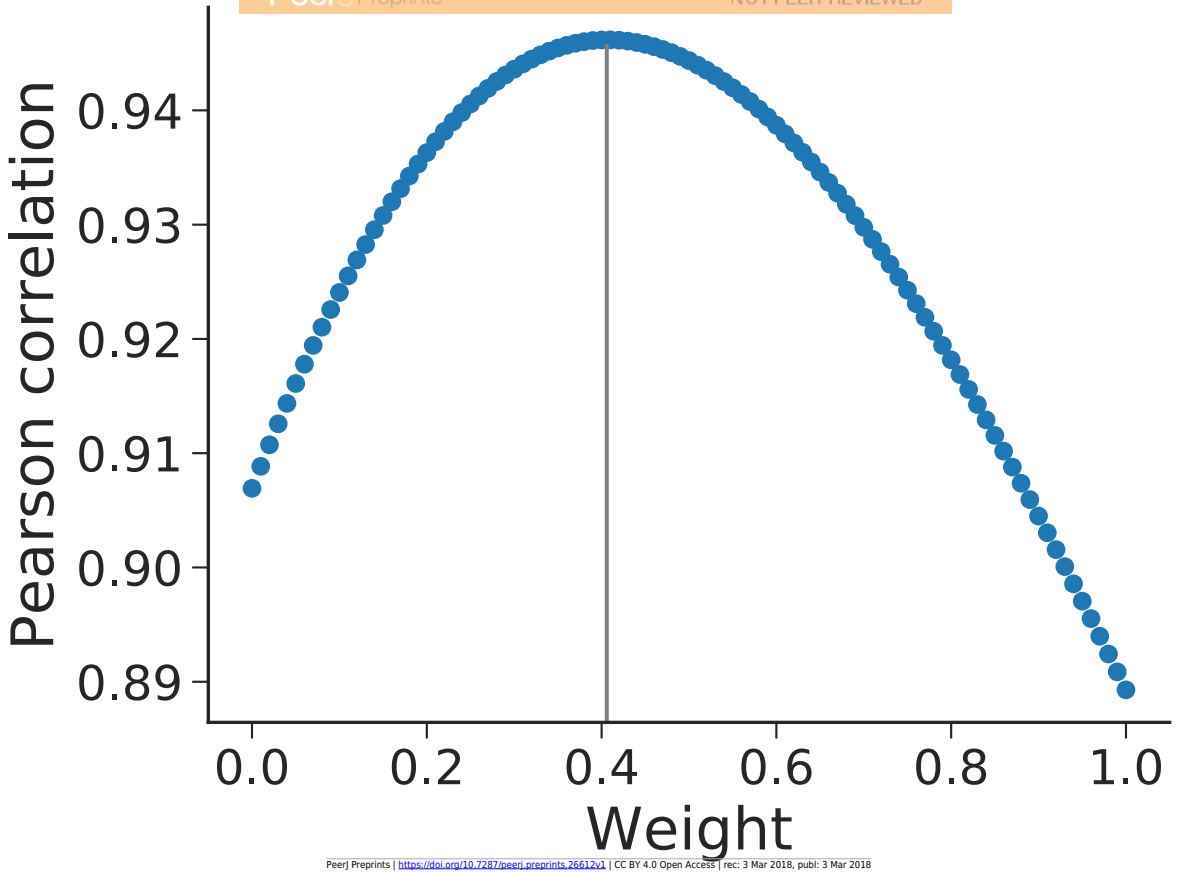
Scatterplot comparison of edit similarity and WL similarity for weight $\omega=0.41$, including marginal distributions and least-squares regression line.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

Figure 3(on next page)

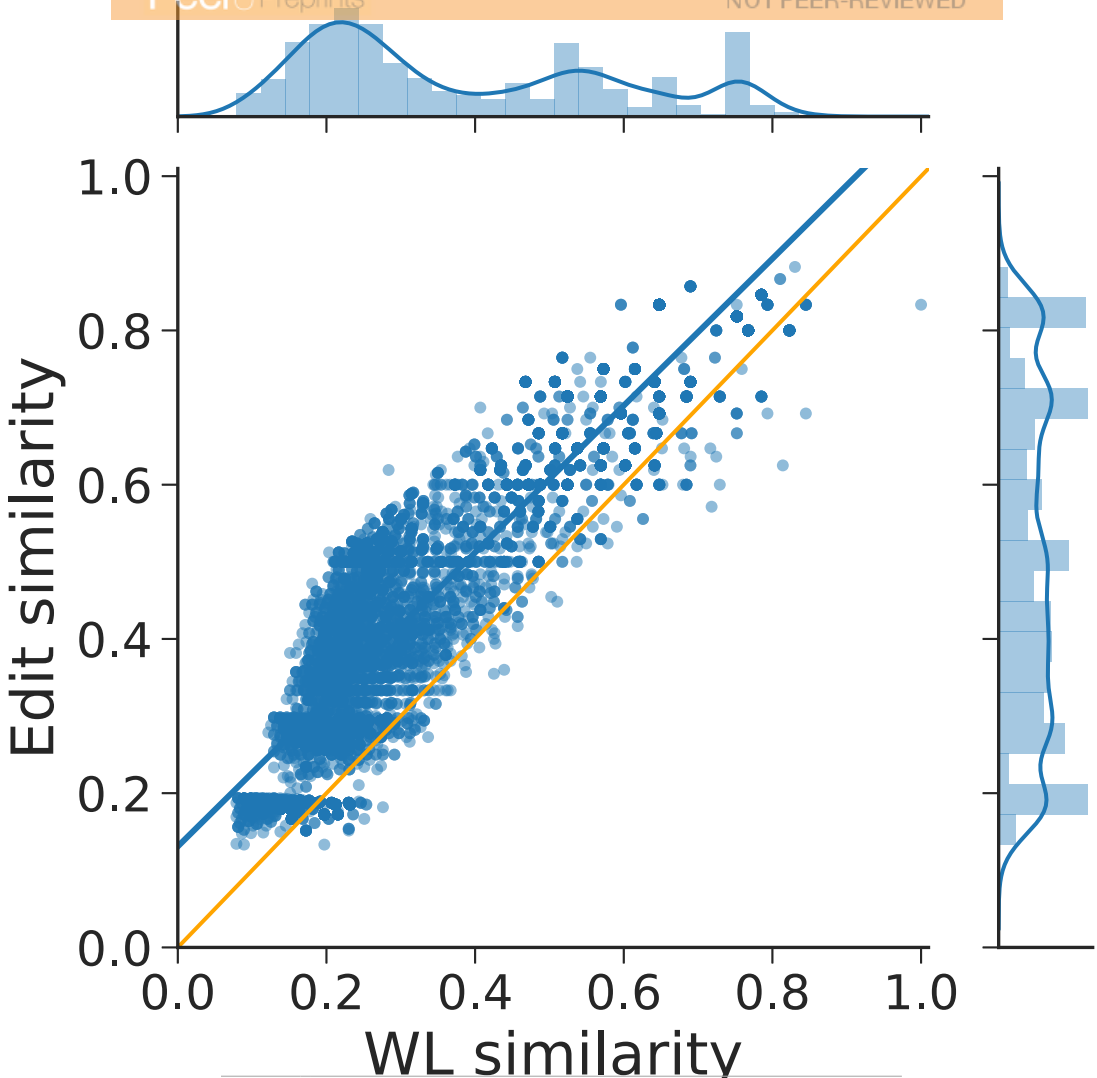
Pearson correlation between edit similarity and WL similarity as a function of weight~ω. The maximum correlation (0.946) occurs for $\omega=0.41$, shown on the left side.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

Figure 4(on next page)

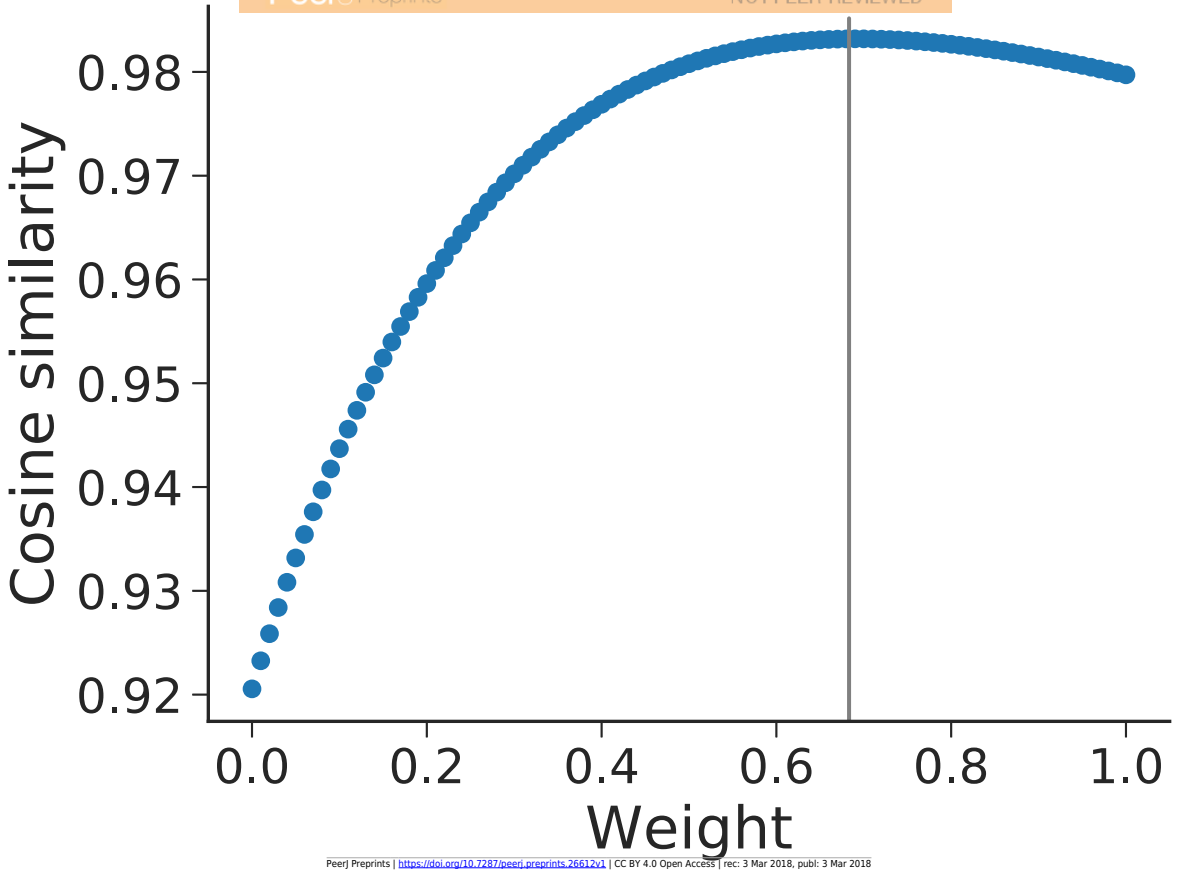
Scatterplot comparison of edit similarity and WL similarity for weight $\omega=0.69, including marginal distributions and least-squares regression line.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018

Figure 5(on next page)

Cosine similarity between edit similarity and WL similarity as a function of weight~ω. The maximum cosine similarity (0.983) occurs for $\omega=0.69$ shown on the left side.



Peerj Preprints | https://doi.org/10.7287/peerj.preprints.26612v1 | CC BY 4.0 Open Access | rec: 3 Mar 2018, publ: 3 Mar 2018