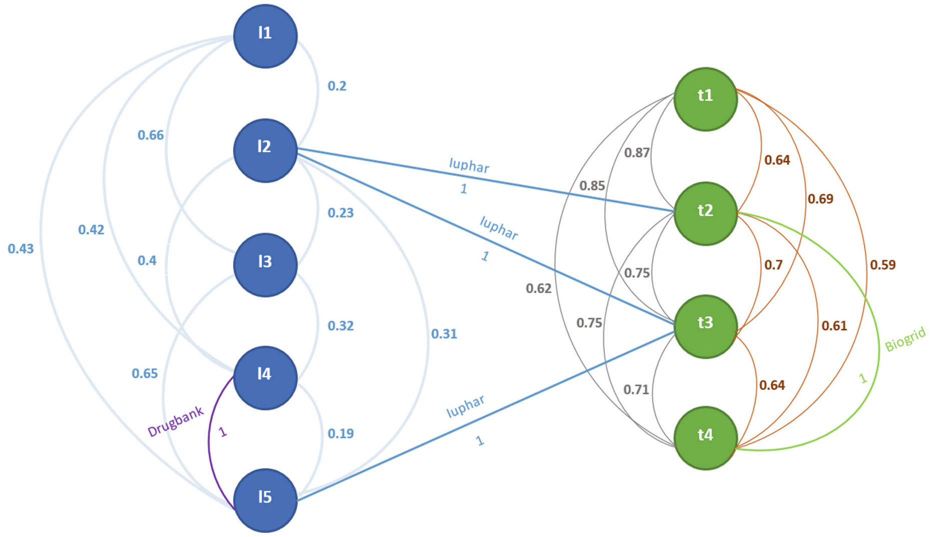

Fig. 2. An example of a multi-layer graph with 6 networks: ligand-protein network is in deep blue (IUPHAR), ligand-ligand networks are in light blue (ligand similarity network) and violet (DrugBank), protein-protein networks are in green (BioGrid), grey (protein similarity network based on substrings) and brown (protein similarity network based on motifs)

Definition 4 *(***Motif***). Protein motifs are patterns defined using biochemical background knowledge, often expressed in the form of regular expressions.*[1]

Definition 5 *(***Tanimoto coefficient***). The Tanimoto coefficient of two vectors* $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^d$ *is calculated as:* $coeff_{Tanimoto}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||^2 + ||\boldsymbol{y}||^2 - \boldsymbol{x} \cdot \boldsymbol{y}}$.

2.2 Problem Formulation

The problem setting we address in this paper is one of link prediction between ligands (drug candidates) and proteins (biological targets).

Definition 6 *(***Ligand-protein activity prediction***). For a given number of ligand-protein activity networks* $G^i_{lp} = \langle V_l \cup V_p, E_i, \lambda_v, \lambda_{e_i} \rangle$*, with* $u \in V_l$ *labeled with ligands identifiers,* $v \in V_p$ *labeled with protein identifiers,* $\forall (u,v) \in E, u \in V_l, v \in V_l$*, and* $\mathcal{A}_e = \{0,1\}$*, ligand-ligand networks* $G^i_l = \langle V_l, E^i_l, \lambda_v, \lambda_{e^i_l} \rangle$*, protein-protein networks* $G^i_p = \langle V_p, E^i_p, \lambda_v, \lambda_{e^i_p} \rangle$ *and a given* $(u,v) \notin E, u \in V_l, v \in V_p$ *predict, whether* $\lambda_e((u,v)) = 1$*.*

We limit ourselves to the relatively easier task of predicting whether there is activity or not, leaving the prediction of its *strength* as future work.

[1] An open-access database is available at http://prosite.expasy.org.

3 Related Work

The literature on compound-target activity prediction, even using networks, is too vast to discuss here. We therefore present a number of works illustrating the characteristics we discussed in the introduction. Ligand-protein activity, the use case we explore here, has been addressed in [13], which selects a ligand and target similarity measure each, and multiplies activity vectors of known ligands/targets with the similarity to new ligands/targets to derive predictions. In [14], the same group used ligand structural and pharmacological similarity, as well as genetic protein similarity, mapped ligands and targets into a shared feature space and predicted activity. The authors of [3] used three networks: ligand-ligand similarity, target-target similarity, ligand-target activity, evaluated four ligand similarity measures, settling on Tanimoto distance. The proposed method, NWNBI, exploits similarity weights and log-values of activity measurements to perform four-step network traversals. In [2], ligand similarity is calculated as weighted average of *two* similarity measures, and combined with a target similarity, and the interaction network into a three-layer network, which they refer to as "heterogeneous". They simulate random walk with restart by matrix multiplication, and show that only using a single similarity measure or ignoring the interaction network deteriorates results. Three networks are also used in [7], the authors discuss different options for similarity measures, and perform low-rank matrix factorization on the adjacency/similarity matrices. They address sparsity by giving non-existing links a small non-negative weight. Ligand-protein activity is also the subject of [1], which exploits the three-layer network to perform weighted nearest-neighbor classification. Gene-disease interactions have been considered in [12], using three layers, simulating random walk by matrix multiplication, using different numbers of steps for the two similarity networks. Using a similar bi-random walk idea, [9] consider microRNA-disease interactions, exploiting a three-layer network. The random walk with restart in [8] is symmetric (and functionally the same as in [2]), with the similarity networks constructed by averaging two similarity measures. They evaluate different parameter settings.

4 Exploring a Multi-layer Graph

As the preceding section shows, the standard setting employed consists of three networks, and to adhere to this setting, authors either choose a single similarity measure empirically, or combine similarity measures via user-specified weights. Instead, we propose to combine all available networks into a multi-graph having more than three layers. Once we have such a network, the question is how to exploit it, however, and here we hew close to the literature.

4.1 The Random Walk Model

A long-established method for exploring a network is the random walk [10], which proceeds roughly as follows: starting from a randomly selected node, it performs walks along edges of the graph at random. In every step, the edge to follow is chosen uniformly from all outgoing links (in the case of an unweighted graph) or proportional to link weights (in the case of a weighted graph). Node importance is based on how frequently the walker visits the node: a node with higher frequency is considered more important than a node with a low value (Fig. 3). This idea can be modified in a number of ways to improve network exploration: the walker can be constrained to perform at most *max_steps* steps, to not visit any of the last c vertices it encountered, or with small probability $1-\beta$ the process can be restarted at any time to avoid getting trapped by those vertices it mustn't visit. The product of the probabilities of edges the walker traversed gives the cumulative probability of a path between two nodes and can be used to *predict* a link between a starting node and an end node: if the path probability is greater than a given threshold, a new edge is predicted.

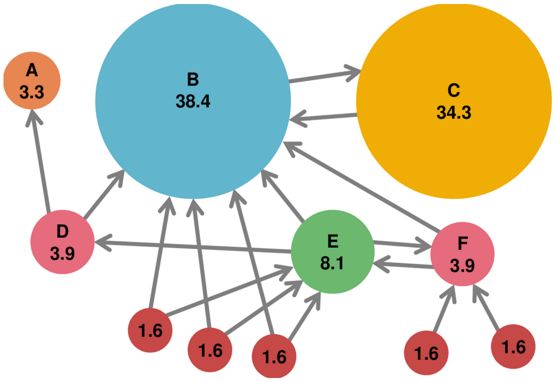

Fig. 3. Nodes importance example in a graph, taken from [6]

To extend this approach to multi-layer graphs, one needs to add how to choose the layer to walk in. We propose to select a network uniformly at random from the set of networks, and multiply the path strength by $\frac{1}{|\{G_l^i\}|+|\{G_p^j\}|+|\{G_{lp}^k\}|}$. Repeat the process until a user-defined target vertex is reached or the maximum number of steps have been performed. Due to the randomized nature, random walks are usually repeated several times to derive more robust estimates.

4.2 Network-Based Random Walk on Multi-layer Network

Instead of explicitly random walking as described above, random walks are often simulated via matrix multiplication of transition matrices. This is notably the approach proposed in [2], abbreviated as NRWRH. They define a transition matrix $M = \begin{bmatrix} M_{pp} & M_{pl} \\ M_{lp} & M_{ll} \end{bmatrix}$, in the manner described above, i.e. uniform probabilities for lp/pl-transitions, proportional probabilities for similarities, with an additional user-specified parameter $\lambda \in [0,1]$ affecting moves from ligands to proteins and vice versa. Given a ligand l_i, a starting vector $v_0 \in [0,1]^{|V_l|+|V_p|}$ is initialized with 1 at the position for $v \in V_l, \lambda_v(v) = l_i$, $\frac{1}{|\{(v,u) \in E_{lp}\}|}$ at the positions for the proteins linked to it, 0 otherwise. Protein entries in v_0 are multiplied with $1-\eta$, ligand entries with η, a user-defined parameter to bias the walk towards proteins ($\eta < 0.5$), or ligands ($\eta > 0.5$). The vector representing the probabilities that a walker starting with l_i finds itself in any of the nodes is calculated iteratively as $p_{t+1} = (1-\beta)M^T p_t + \beta p_0$ until $|p_{t+1} - p_t| < 10^{-10}$. This can be understood as the random walker walking "in all directions at the same time". The approach can be considered a simplified version of Personalized PageRank [5], simplified because edges are undirected and there is only a single starting vertex. Removing the starting vertices from the final state vector, and ranking entries gives predicted edges. We adapt this approach to a setting with $|\{G_l^i\}| + |\{G_p^j\}| + |\{G_{lp}^k\}| \geq 3$. While the algorithm stays essentially the same, we decompose the transition matrix into a matrix M for within-network/layer transitions, and a matrix N for between-network/layer transitions. We also do away with the user-dependent λ. Explicitly creating M in the manner shown above is easy for three layers but becomes much harder when different numbers can be involved. We hence construct $M = \begin{bmatrix} M_{G_p} & 0 & 0 \\ 0 & M_{G_l} & 0 \\ 0 & 0 & M_{G_{lp}} \end{bmatrix}$, with $M_{G_p} = \begin{bmatrix} M_{G_p}^1 & 0 & \dots & 0 \\ 0 & M_{G_p}^2 & \dots & 0 \\ 0 & 0 & \dots & M_{G_p}^{|\{G_p^i\}|} \end{bmatrix}$ derived from protein-protein similarity networks (M_{G_l}, $M_{G_{lp}}$ accordingly). The transition matrix $N = \begin{bmatrix} N_{G_p^1 \to G_p^1} & N_{G_p^2 \to G_p^1} & \dots & N_{G_{lp}^{G_{lp}^i} \to G_p^1} \\ \dots & \dots & \dots & \dots \\ N_{G_p^1 \to G_{lp}^{G_{lp}^i}} & N_{G_p^2 \to G_{lp}^{G_{lp}^i}} & \dots & N_{G_{lp}^{G_{lp}^i} \to G_{lp}^{G_{lp}^i}} \end{bmatrix}$ explicitly models possible layer transitions, with 1s on the main diagonal of a submatrix $N_{G_j \to G_i}$ for all nodes present in both layers, 0s otherwise. Note that this means that transition matrixes from ligand to protein layers (and vice versa) have zeros everywhere including the main diagonal. The initial state vector v_0 has dimensionality $(|V_p| \cdot |\{G_{p_i}\}| + |V_l| \cdot |\{G_{l_i}\}| + |V_l \cup V_p| \cdot |\{G_{lp_i}\}|)$ with entries for *all* vertices in *all* layers. It is initialized by setting the entry for the starting ligand and each linked protein to 1 in every network they are present. Matrices and state vectors are column-normalized – the entries of a column must sum to 1.

Our algorithm, NEtWork-basEd Random walk on MultI-layered NEtwork (NEWERMINE), is summarized in Algorithm 1. $(M_{norm}N)_{norm}$ can be precomputed, giving us a matrix that is functionally equivalent to M as defined in NRWRH, and used on every iteration of NEWERMINE to save computation time. At the end, v_{final} needs to be summarized by summing up for each vertex all corresponding entries, leading to a vector with dimensionality $|V_l \cup V_p|$ from which the edge ranking can be derived.

5 Experimental Evaluation

In order to allow reproducibility of our work, we evaluated our approach on publicly available data. In this part we provide a description of the data used

Algorithm 1: The NEWERMINE algorithm

```
Input  : adjacency matrix $M$, transition matrix $N$, $starting\_vertex$,
         $max\_steps$, $\eta$, $\beta$, $max\_diff$
Output: Probability scores $v_{final}$
$V_{0_l} \leftarrow$ initialize $starting\_vertex$
$V_{0_p} \leftarrow$ initialize targets for which an interaction with $starting\_vertex$ is known
$V_0 \leftarrow (1-\eta) \cdot V_{0_{l\,norm}} + \eta \cdot V_{0_{p\,norm}}$
$step \leftarrow 0$
repeat
    $step \leftarrow step + 1$
    $V_{step} \leftarrow \beta \cdot (M_{norm}N)_{norm} V_{step-1} + (1-\beta) \cdot V_0$
until $(|v_{step} - v_{step-1}| \leq max\_diff) \vee (step > max\_steps)$
return $v_{step}$
```

and the details of the experimental protocol. This is followed by the results and the discussion.

5.1 Experimental Settings

Datasets. In total we have used 4 datasets:

1. IUPHAR – an open-access database of ligands, biological targets and their interactions. We used version 2017.5 (released on 22/08/2017). The full dataset has 8978 ligands, 2987 proteins, and 17198 interactions (edges) between them[2]. In order to satisfy the designed setting conditions, we removed duplicate interactions (based on different affinity measures), leaving 12456 interactions in total. For existing interactions, we label an edge with 1 if the negative logarithm of the affinity measure is $\geq$5, non-interacting otherwise.[3] We treat all affinity measures available in the data (pKi, pIC50, pEC50, pKd, pA2, pKB) as equivalent.
2. DrugBank (DB) – an open-access database of drug-drug interactions. We used version 5.0.11 (released 20-12-2017). It has 658079 interactions of 3138 distinct drugs. 242922 of these interactions involve 1254 distinct ligands that are present in IUPHAR. The database was also used as a source of 2D representations of ligands to compute ligand similarities.
3. BioGrid (BG) – an open-access database of protein-protein interactions mined from a corpus of biomedical literature. We used version 3.4.154 (25/10/2017). It has 1482649 interactions of 67372 distinct proteins. Only 15410 of these interactions involve proteins present in IUPHAR (1925 distinct proteins).
4. NCBI Protein database – The National Center for Biotechnology Information proteins database[4] was used to obtain amino acids sequences to represent targets. The data was parsed from the website of NCBI and mapped to IUPHAR

[2] in ligands.csv, interactions.csv, and targets_and_families.csv, respectively.
[3] Cutoff proposed by researchers from CERMN (http://cermn.unicaen.fr).
[4] https://www.ncbi.nlm.nih.gov/protein/.

Table 1. Data set and network characteristics

Data set	Entities	Relations	Sparsity	Network			
				Vertices	Edges	Sparsity	CC
IUPHAR	11965	12456	0.00017	11965	12456	0.00017	443
DrugBank	3138	658079	0.1337	1254	122808	0.15631	1
BioGrid	67372	1482649	0.00065	1898	8658	0.0048	11
Ligand similarity	6821	23259610	1	6821	23259610	1	1
NCBI	1818	1651653	1	1818	1651653	1	1

using the RefSeq attribute (human protein sequence identifier) available in IUPHAR. The database was accessed 20/12/2017.

Ligands were mapped between networks by numerical identifiers provided by IUPHAR as well as by INN (International Non-proprietary Name) and Common name attributes. Proteins were mapped by IUPHAR identifiers as well as by Human Entrez Gene attribute.[5] In total we have built 6 networks:

1. a drug interaction network based on DrugBank,
2. a drug similarity network based on similarities calculated using the Tanimoto coefficient on binary vectors constructed by frequent subgraphs,
3. the drug-target interaction network based on IUPHAR,
4. a target interaction network based on BioGrid, and
5. two target similarity networks calculated using the Tanimoto coefficient on feature vectors constructed by *frequent substrings* and *Prosite motifs*.

Similarity networks' edges were labeled with labels $\in [0, 1]$, interaction networks with labels $\in \{0, 1\}$. Table 1 shows the characteristics of the data sets, and of the networks we derived from them. It is noticeable how sparse the data is, and also how this sparsity translates into disconnected parts of the network. Sparsity might result in a low performance of the traditional recommender systems approaches, while disconnected networks are challenging for random walker approaches.

Evaluation Protocol. To evaluate our approach, we used leave-one-out cross-validation: for each of the 12456 edges in the IUPHAR network, we remove it from the network, set the ligand as starting vertex, infer strengths for all possible ligand-target paths, remove ligand-target edges contained in the training data, and check whether the removed edge is found in the top-20 remaining paths[6] according to their strengths. If this is the case for an interacting edge, we consider it a *true positive*, otherwise a *false negative*. For negative examples, the relationship is inverse.

[5] Global Query Cross-Database Search System gene identifiers: https://www.ncbi.nlm.nih.gov/gene.

[6] Precision at 20.

Quality Measures. To evaluate our methods we use several performance measures:

- Accuracy: the ratio of true positives (TP) – drug-target links correctly classified as positives – and true negatives (TN) – drug-target links correctly classified as negatives – over all predictions: $Acc = \frac{TP+TN}{TP+FP+TN+FN}$.
- Area under receiver operating curve (AUC): evaluates whether true positives are usually ranked above or below false positives when sorting predictions by confidence.
- Precision: the ratio of TP over all drug-target links classified as positives: $Prec = \frac{TP}{TP+FP}$. Precision measures whether a model is specific enough to mainly classify links of the positive class as positive. This gives additional insight into the accuracy score.
- Recall: the ratio of TP over all positive links in the test data: $Rec = \frac{TP}{TP+FN}$. Recall measures whether a model is general enough to classify a large proportion of the positive class as positive.

In addition to this, we also report weighted versions of accuracy, precision, and recall that give us a more accurate assessment for unbalanced datasets. Due to the fact that the number of negative examples are smaller than that of the positives in our data, we assign a classification cost of 1 to positives and cost *neg_cost* to negatives, derived by: $neg\_cost = \frac{|D|}{2 \times |N|}$, where $|D|$ – number of examples, $|N|$ – number of negative examples. We then perform evaluation based on the costs defined: FN and TN receive score *neg_cost* for every negative example w.r.t. its real class, while FP and TP receives score 1 for positives.

Implementation. We implemented NEWERMINE in Python[7]. We used the networkx library to model the multi-layer network, the NumPy library to perform all matrix computations and the sklearn library for cost-based evaluation.

5.2 Experimental Results

Using Three-Layer Graphs. We first use NEWERMINE on a number of multigraphs aggregated from three networks each, the ligand-target network, one ligand-ligand network, and one target-target network. This is the setting used in the papers discussed in Sect. 3.

For the experiments we defined 6 possible combinations with IUPHAR, only ligand-target interaction network we have: (1) DrugBank + BioGrid, (2) DrugBank + Target similarity (TS) (substrings:str), (3) DrugBank + TS (motifs:mot), (4) Ligand similarity (LS) + BioGrid, (5) LS + TS (str), (6) LS + TS (mot). The basic properties of the combinations compared to the full graph are presented in Table 2. The results of the use of NEWERMINE with parameters $\eta = 0.2$, $\beta = 0.7$ (taken from [2]) are presented in Fig. 4. This is a rather conservative setting, equivalent to relatively few steps before the walker restarts.

[7] https://zimmermanna.users.greyc.fr/supplementary-material.html.

Table 2. Basic properties of different combinations of networks

Combination	Ligands	Targets	$\lvert V \rvert$	$\lvert E \rvert$	Sparsity	CC
DB + BG	7025	2307	9332	143922	0.003	87
DB + TS (str)	7025	2101	9126	1786917	0.042	103
DB + TS (mot)	7025	2101	9126	1786917	0.042	103
LS + BG	8056	2307	10363	23280724	0.434	21
LS + TS (str)	8056	2101	10157	24923719	0.4832	22
LS + TS (mot)	8056	2101	10157	24923719	0.4832	22
Six layers	8137	2502	10639	26706838	0.4719	1

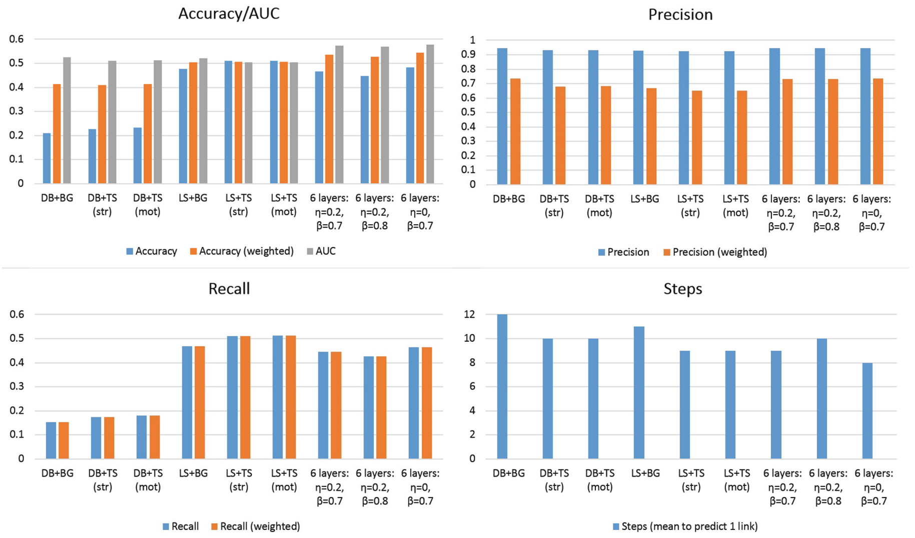

Fig. 4. Evaluation results of NEWERMINE for different combinations of three networks and the six-layer graph

The number of vertices in different networks depend on available IDs and structural information. In any case, the networks are sparse and they are not fully connected. Using similarity networks alleviates this situation somewhat and combining *all* networks leads to a single connected component (bottom row).

Figure 4 shows that using different three-layer graphs leads to rather different results. The arguably most notable result is that using ligand structural similarity instead of DrugBank network significantly improves accuracy and recall.

Using the Full, Six-Layer Graph. The results for NEWERMINE on the full multi-layer graph are also presented in the Fig. 4. We show additional values for η and β: $\eta = 0$ strongly biases the walk towards targets, we also consider $\beta = 0.8$

for $\eta = 0.2$. Using more layers decreases recall somewhat, but improves weighted accuracy (taking the lower proportion of negative examples into account), AUC score and precision. Different parameter values do not have a large effect on the results but change running times: increasing β also increases the number of steps necessary for convergence, and decreasing η decreases this number.

6 Conclusion and Perspectives

We have presented an approach for exploiting an arbitrary number of networks combined into a multi-layer network, proposing general matrix formulations to form intra- and inter-network transitions.

As we have demonstrated experimentally, combining different networks improves vertex reachability and therefore interaction prediction. So far, we have only exploited more than one protein similarity network, already achieving very good results. In future work, we intend to also integrate different ligand similarity semantics, and different databases indicating ligand-protein activity. Additionally, we intend to employ our approach for different target settings, e.g. for miRNG-disease links. Finally, we aim to move from the "active"/"inactive" setting to one where we predict the strength of the activity.

References

1. Buza, K., Peska, L.: ALADIN: a new approach for drug–target interaction prediction. In: Ceci, M., Hollmén, J., Todorovski, L., Vens, C., Džeroski, S. (eds.) ECML PKDD 2017. LNCS (LNAI), vol. 10535, pp. 322–337. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71246-8_20
2. Chen, X., Liu, M.X., Yan, G.Y.: Drug-target interaction prediction by random walk on the heterogeneous network. Mol. BioSyst. **8**(7), 1970–1978 (2012)
3. Cheng, F., Zhou, Y., Li, W., Liu, G., Tang, Y.: Prediction of chemical-protein interactions network with weighted network-based inference method. PloS One **7**(7), e41064 (2012)
4. Hattori, M., Okuno, Y., Goto, S., Kanehisa, M.: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. JACS **125**(39), 11853–11865 (2003)
5. Haveliwala, T.H.: Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. TKDE **15**(4), 784–796 (2003)
6. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, Cambridge (2014)
7. Lim, H., Gray, P., Xie, L., Poleksic, A.: Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. Sci. Rep. **6**, 38860 (2016)
8. Liu, Y., Zeng, X., He, Z., Zou, Q.: Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources. TCBB **14**(4), 905–915 (2017)
9. Luo, J., Xiao, Q.: A novel approach for predicting microrna-disease associations by unbalanced bi-random walk on heterogeneous network. J. Biomed. Inform. **66**, 194–203 (2017)

10. Pearson, K.: The problem of the random walk. Nature **72**(1867), 342 (1905)
11. Smith, T., Waterman, M.: Identification of common molecular subsequences. Mol. Biol. **147**, 195–197 (1981)
12. Xie, M., Hwang, T., Kuang, R.: Prioritizing disease genes by bi-random walk. In: Tan, P.-N., Chawla, S., Ho, C.K., Bailey, J. (eds.) PAKDD 2012. LNCS (LNAI), vol. 7302, pp. 292–303. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30220-6_25
13. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics **24**(13), i232–i240 (2008)
14. Yamanishi, Y., Kotera, M., Kanehisa, M., Goto, S.: Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics **26**(12), i246–i254 (2010)

A Hierarchical Ornstein-Uhlenbeck Model for Stochastic Time Series Analysis

Ville Laitinen(✉) and Leo Lahti

Department of Mathematics and Statistics, University of Turku, Turku, Finland
velait@utu.fi

Abstract. Longitudinal data is ubiquitous in research, and often complemented by broad collections of static background information. There is, however, a shortage of general-purpose statistical tools for studying the temporal dynamics of complex and stochastic dynamical systems especially when data is scarce, and the underlying mechanisms that generate the observation are poorly understood. Contemporary microbiome research provides a topical example, where vast cross-sectional and longitudinal collections of taxonomic profiling data from the human body and other environments are now being collected in various research laboratories world-wide. Many classical algorithms rely on long and densely sampled time series, whereas human microbiome studies typically have more limited sample sizes, short time spans, sparse sampling intervals, lack of replicates and high levels of unaccounted technical and biological variation. We demonstrate how non-parametric models can help to quantify key properties of a dynamical system when the actual data-generating mechanisms are largely unknown. Such properties include the locations of stable states, resilience of the system, and the levels of stochastic fluctuations. Moreover, we show how limited data availability can be compensated by pooling statistical evidence across multiple individuals or studies, and by incorporating prior information in the models. In particular, we derive and implement a hierarchical Bayesian variant of Ornstein-Uhlenbeck driven t-processes. This can be used to characterize universal dynamics in univariate, unimodal, and mean reversible systems based on multiple short time series. We validate the model with simulated data and investigate its applicability in characterizing temporal dynamics of human gut microbiome.

Keywords: Longitudinal analysis · Hierarchical models · Ornstein-Uhlenbeck process · Resilience · Stochastic processes

1 Introduction

Many natural and social systems are complex and cannot be studied in isolation. The underlying data-generating mechanisms are often largely unknown in such cases, and the observed dynamics can be characterized only indirectly [8]. Non-parametric models that focus on characterizing observed data properties, rather

W. Duivesteijn et al. (Eds.): IDA 2018, LNCS 11191, pp. 188–199, 2018.
https://doi.org/10.1007/978-3-030-01768-2_16

than modeling the underlying mechanisms, can provide valuable information on the system behavior. In the context of human microbiome dynamics, for instance, such non-parametric models have been used to describe and infer the presence of alternative ecosystem states [13], periodicity, stochasticity, and chaos [5,6]. In many real applications, the data is scarce, and new methods are needed in order to derive maximal information from limited observations.

Our study is motivated by the analysis of temporal dynamics of human gut microbiome. This refers to the totality of microbial communities living on skin, gastrointestinal tract and other body sites. Contemporary human microbiome research has largely focused on cross-sectional cohorts with limited follow-ups, providing information on the composition and inter-individual variation of the microbiome. The dynamics of these systems are yet, however, not well understood despite their clinical importance [1,10]. As understanding of these systems is accumulating, the research focus is beginning to shift from general descriptions towards actionable clinical applications and manipulation.

In this work, we show how key dynamical properties of poorly understood dynamical systems can be inferred from limited time series by pooling information can across multiple individuals. In the present work, we focus specifically on mean-reversible stochastic processes. Such dynamic behavior is frequently observed in the human gut microbiome. Many bacterial species in the human gut ecosystem have been reported to exhibit characteristic abundance levels around which they tend to fluctuate over time (see e.g. [13]). It has been reported that the average abundance levels of many gut bacteria remain relatively stable over long time periods but on a shorter (daily) time scale the abundances can exhibit considerable fluctuations [3]. Mean-reverting stochastic processes, in particular the Ornstein-Uhlenbeck (OU) process, provide well-established means to characterize key properties of such systems, including the location and resilience of the potential wells, speed of mean reversion, and volatility of abundance levels, even when the underlying mechanisms regulating those dynamics are unknown. Stochastic processes and generative probabilistic models provide a rigorous framework for the characterization of the observed dynamics in such cases, with wide applicability across different application domains [9,11,16,18].

We adapt and apply these techniques to model human gut microbiome dynamics. A key practical limitation of the existing methods in our application is that the available implementations of the OU process depend on the availability of long time series with dozens of time points or more. The currently available longitudinal data sets in typical human microbiome studies have more limited sample sizes and time series lengths, or sparse sampling intervals. Combined with high levels of variation and limited knowledge of the data-generating processes, these limitations form considerable challenges for the application of previously established stochastic models, such as the the OU process, in contemporary human microbiome research. In order to address these limitations, we derive, implement, and validate a hierarchical extension to the OU process. This can be used to recover key information of the system dynamics from limited data by aggregating information across short time series from multiple individ-