Machine Learning of Protein Interactions in Fungal Secretory Pathways

Jana Kludas, Mikko Arvas, Sandra Castillo, Tiina Pakula, Merja Oja, Celine Brouard, Jussi Jäntti, Merja Penttilä, Juho Rousu

PLOS ONE

RESEARCH ARTICLE

Machine Learning of Protein Interactions in Fungal Secretory Pathways

Jana Kludas[1], Mikko Arvas[2], Sandra Castillo[2], Tiina Pakula[2], Merja Oja[2], Céline Brouard[1], Jussi Jäntti[2], Merja Penttilä[2], Juho Rousu[1]*

1 Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland, **2** VTT Technical Research Centre of Finland, Espoo, Finland

* juho.rousu@aalto.fi

Citation: Kludas J, Arvas M, Castillo S, Pakula T, Oja M, Brouard C, et al. (2016) Machine Learning of Protein Interactions in Fungal Secretory Pathways. PLoS ONE 11(7): e0159302. doi:10.1371/journal.pone.0159302

Editor: Bin Liu, Harbin Institute of Technology Shenzhen Graduate School, CHINA

Received: February 15, 2016

Accepted: June 30, 2016

Published: July 21, 2016

Data Availability Statement: The sequence data for the two studied organisms were downloaded from SGD database (http://www.yeastgenome.org) for S. cerevisiae and from JGI Mycocosm database (http://genome.jgi.doe.gov/Trire2/Trire2.home.html) for T. reesei. The S.cerevisiae protein interaction data is taken from the paper: Feizi A, Österlund T, Petranovic D, Bordel S, Nielsen J (2013). Genome-scale modeling of the protein secretory machinery in yeast. PLoS computational biology 8: e63284. The gene expression correlation data used in T. reesei PPI network evaluation is given in the supplementary material.

Abstract

In this paper we apply machine learning methods for predicting protein interactions in fungal secretion pathways. We assume an inter-species transfer setting, where training data is obtained from a single species and the objective is to predict protein interactions in other, related species. In our methodology, we combine several state of the art machine learning approaches, namely, multiple kernel learning (MKL), pairwise kernels and kernelized structured output prediction in the supervised graph inference framework. For MKL, we apply recently proposed centered kernel alignment and *p*-norm path following approaches to integrate several feature sets describing the proteins, demonstrating improved performance. For graph inference, we apply input-output kernel regression (IOKR) in supervised and semi-supervised modes as well as output kernel trees (OK3). In our experiments simulating increasing genetic distance, Input-Output Kernel Regression proved to be the most robust prediction approach. We also show that the MKL approaches improve the predictions compared to uniform combination of the kernels. We evaluate the methods on the task of predicting protein-protein-interactions in the secretion pathways in fungi, *S.cerevisiae*, baker's yeast, being the source, *T. reesei* being the target of the inter-species transfer learning. We identify completely novel candidate secretion proteins conserved in filamentous fungi. These proteins could contribute to their unique secretion capabilities.

Introduction

Protein secretion is a fundamental cellular process that is required for transporting proteins into cellular compartments, the cell surface and the external space of the cell as well as for covalent modification i.e. disulphide bond formation and glycosylation of proteins. As can be expected from its central role, the protein secretion machinery is conserved in eukaryotes. Fundamental research to unravel its functioning has been carried out in the fungus *Saccharomyces cerevisiae* [1]. However, the baker's yeast *S. cerevisiae* of the subphylum Saccharomycotina does not naturally secrete large amounts of proteins unlike the filamentous fungi of the subphylum Pezizomycotina. For example the Pezizomycotina *Trichoderma reesei* (*Hypocrea jecorina*) is able to secrete its native cellulase proteins with yields of over 100 g/l in industrial

Funding: This work has been supported by the European Union FP7 Cooperation Work programme (grant 289126: BIOLEDGE) (http://cordis.europa.eu/fp7/home_en.html) and the Finnish Funding Agency for Innovation TEKES under the Living Factories strategic opening (dno:562/31/2014) (http://www.tekes.fi). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

cultivations [2]. With their protein secretion capabilities Pezizomycotina are central to industrial biotechnology. However, compared to *S. cerevisiae* their protein secretion machinery has not been studied in detail.

Protein-protein interaction data is very useful in defining cellular functions of proteins. Protein-protein interaction (PPI) is a term that covers various possible interactions between pairs of proteins from stable physical interactions to functional associations. While in many cases the molecular function of a new protein can be determined by sequence similarity searches, the molecular function tells little about the cellular function the protein might be carrying out in given conditions. However, the actual high throughput experimental measurement of protein-protein interactions has been limited to the most studied organism such as *E. coli*, *S. cerevisiae* and *Homo sapiens*.

The availability of verified annotations on protein function and interactions especially outside model organisms is currently a major bottleneck. In May 2015 the Genomes OnLine Database (https://gold.jgi-psf.org/) contained entries for almost 60 000 organisms. The number of sequenced species is growing exponentially and most importantly improvements in sequencing techniques allow the assembly of genome sequences of uncultivable micro-organisms from metagenomics samples [3]. In parallel the version 10.0 of protein-protein interaction database STRING [4] contained 2031 organisms. STRING combines a number of data sources, i.e. genomic neighbourhood, gene fusion, species co-occurrence, gene co-expression, experimental protein-protein interaction data and text mining results to predict protein-protein interactions. Furthermore, all taxonomic groups of organisms typically contain 10–20 percentage of lineage restricted genes i.e. genes that are not found in other taxonomic lineages [5]. Outside model organisms the function of these genes is typically unknown. With sequencing of uncultivable micro-organisms this percentage is likely to increase.

To bridge this gap computational PPI prediction has been intensely studied in the last decade. Early approaches focused on inferring functional PPIs from genomic context such as gene neighbors, gene clusters, Rosetta stone, and phylogenetic profiles as well as protein sequence co-evolution as reviewed in [6]. Lately, the research field has developed methods for predicting PPI networks of physical interactions [7, 8], pathway memberships [7] and more general biological networks such as gene regulatory networks and metabolic networks [8, 9]. Commonly experimental data is used as input features such as microarray/co-expression [8, 9], sometimes also more high level features such as domain knowledge, phylogenetic profiles and interologs [7]. In [8] a good review on approaches for *de novo* as well as for supervised biological network inference is given.

Transfer of protein interactions based on sequence homology is a widely used technique, but requires strict amino acid sequence identity cut-offs, for example above 80 percent, to be reliable [10, 11]. This limits its use to lineages where not much sequence diversification has occurred. For example homologous genes between species belonging to the fungal genus Aspergilli, of the subphylum Pezizomycotina, have only 68 percent average amino acid identity [12]. Furthermore, recently duplicated and hence sequence wise similar genes often change their function i.e. neofunctionalise. When comparing duplicated genes between species it has been found that orthologous gene pairs are more likely to retain functions than paralogous (for review [13]). However, using the orthology-paralogy relationships in function transfer would require that they would be first solved. Although numerous methods exist for this, on genome scale this is still not a trivial task. For example, the commonly used best-bi-directional hit technique can easily be misled in multigene families.

To overcome the sequence similarity requirement of annotation transfer, machine learning methods have been developed for PPI prediction over the last decade. In particular, the supervised network inference paradigm [14] takes the PPI prediction as a binary classification

problem, to predict, whether a pair of proteins interact or not. Thus, any general model for classification learning is applicable in this setting, including ensemble learners [15–17], Naive Bayes, and support vector machines (SVM). SVM models rely on so called pairwise kernels, where the similarities of protein pairs are compared to each other. Another class of PPI learning methods aim to predict interaction patterns by learning similarities between proteins in the protein interaction network. Output kernel trees [18] and input-output kernel regression [19] are recent examples of this kind of methods.

The above approaches have not been explicitly applied to cross-species transfer learning, perhaps due to the limited amount of verified PPIs in a majority of species. Beyond basic sequence comparisons, more advanced computational methods have been applied in the cross-species setting only sparingly. In [20], a cross-species cluster co-conservation method is proposed, that exploits phylogenetic profiles for predicting protein interaction networks. In [21], a link propagation approach was proposed relying on gene expression and sequence similarity, applied to cross-species metabolic network reconstruction.

There is a dire need for novel function and interaction prediction methods that would be locally available, able to cross large sequence similarity distances and not require the solving of orthology-paralogy relationships to cope with the rising amount of genomes. In this paper, we introduce a framework of machine learning methods that can be used for predicting physical or functional protein-protein interaction or more specific biological networks i.e. metabolic pathways depending on what type of training labels are used. Our method uses as features various sequence similarity and protein family analysis derived from the CoReCo pipeline [22]. Although our method relies partly on sequence similarity, it is, through a combination of methods, still able to predict for proteins that do not belong into any known protein family. Hence our method can give clues for PPIs of previously unknown proteins. Our method introduces recently proposed multiple kernel learning (MKL) methods [23] to supervised network inference, thus boosting the performance of the latter method family and making full use of the wide array of sequence-derived features.

We focus in predicting the secretion machinery in industrially relevant fungi, in particular, *T.reesei*. Our focus is in predicting functional protein-protein interactions (PPI) in the secretory pathway. As there are no verified protein interaction data available for these organisms, we assume the cross-species transfer learning setting, where the training data comes from *S. cerevisiae*, and prediction targets is *T.reesei*.

Materials and Methods

Data and preprocessing

Sequence data. In this paper the models are based on features that can be computational derived from protein sequence data. The sequence data for the two studied organisms were downloaded from SGD database (http://www.yeastgenome.org) for *S. cerevisiae* and from JGI Mycocosm database (http://genome.jgi.doe.gov/Trire2/Trire2.home.html) for *T. reesei*.

Protein-protein interaction data. The machine learning methods require a set of known PPIs to be used as ground truth for the model output, used for training and testing the model. We obtained our PPI data from the recently published genome-scale model of the yeast secretory machinery [24] that gathers knowledge of 50 years of research on secretion in *S.cerevisiae*. The authors identified 162 proteins to be involved in secretion that are assigned to 16 subsystems such as translocation, ER glycosylation, COP, Golgi processing etc. These protein complexes give 2200 undirected interactions between the 162 secretion proteins which are used as training labels.

Feature extraction. For our models we use several types of features to characterize the similarity of proteins as well as the similarity of protein pairs. For all protein sequences of the 2 organisms we computed the following features using the CoReCo pipeline [22]: sequence alignment with BLAST against the UniProt database as well as Global Trace Graph (GTG) [25], protein domains and functional sites gathered by InterProScan [26] from its member databases: Pfam [27], Panther [28], Gene3D [29], PRINTS [30], Prosite [31], PIRSF [32], SMART [33], and SUPERFAMILY [34] (See S1 Table for details on these data sources).

Artificial sequences. We used artificial data to test if the different biological network inference algorithm that have been developed for intra-species prediction also work for inter-species prediction with low sequence similarity. They are well below commonly used amino acid sequence identity cut-off values. For obtaining artificial sequences with varying levels of sequence similarity we altered the sequences of the 162 secretion proteins of *S. cerevisiae* based on Blosum matrices [35]. These matrices represent the substitution probabilities from an amino acid to an other amino acid in natural sequence data sets. Hence, they allow approximation of natural sequence evolution. We created four different data sets where we deleted and mutated 70%, 60%, 38% and 20% of the amino acids according to the Blosum30, Blosum40, Blosum62 and Blosum80 respectively. The Blosum matrices were downloaded from NCBI Blast site (ftp://ftp.ncbi.nih.gov/blast/matrices/). Each different Blosum matrix has been made by combining proteins that are no more similar than a given percentage (30%, 40%, 62% and 80%) to one single sequence and then comparing only those sequences [35]. In Fig 1, the percentage of amino acid sequence identity between the artificially mutated protein sequences of *S. cerevisiae* and *T. reesei* based on the Smith-Waterman alignment is shown. Based on visual comparison, the generated Blosum30 data set has a similar level of sequence similarity to *S. cerevisiae* as *T. reesei*. In the experiments, the artificially perturbed sequences were coupled with the labels of the corresponding labels of the original sequences.

Transcriptomic data analysis for biological network validation. The transcriptomic data for the validation of *T. reesei* PPI network was composed by eight publicly available data sets taken from Gene expression omnibus [36] plus eight in-house data sets. The public data sets contained 76 samples all together and the in-house data sets 499 samples. Once combined the

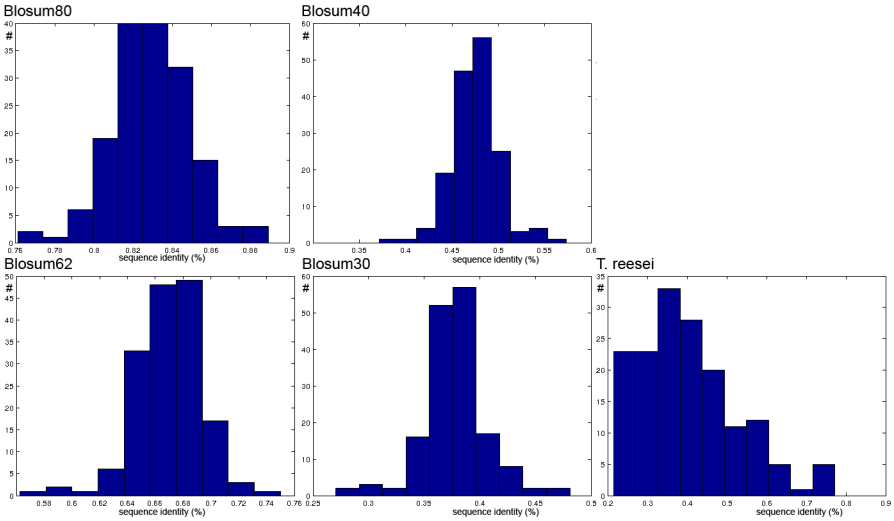

Fig 1. Frequency distribution of percentage of amino acid sequence identity between natural *S. cerevisiae* sequences and (1) sets of artificial sequences created from from *S. cerevisiae* with different Blosum matrices, (2) natural *T. reesei* sequences.

doi:10.1371/journal.pone.0159302.g001

final data set contained 575 samples and 9078 genes. Each data set was normalized separately using quantile normalization [37] and normalized again after they were combined using COMBAT normalization [38].

Problem formalization

Supervised graph inference has been introduced a decade ago in [14] and has been widely used for biological network reconstruction subsequently. Given a set of nodes $V = v_1, .. ,v_m$ a biological network can be defined as an undirected graph $G = (V, E)$ where $E \subset V \times V$ are the edges between the m vertices. The graph can be represented by a symmetric adjacency matrix $Y = (y_{ij})$ of size $m \times m$ where $y_{ij} = y_{ji} = 1$ if the nodes v^i and v^j are connected and $y_{ij} = y_{ji} = 0$ otherwise. We will also use the shorthand $y(v_i) = (y_{ij})_{j=1}^{m}$ to denote the connectivity pattern of protein v_i in the network. In addition, we assume that each node has assigned features $x(v_i) \in \chi$, for some input space χ.

The learning task is then defined as follows: given partial knowledge of the graph $G = (V, E)$ and the feature representation of the nodes, determine a function f: $V \times V \rightarrow \{0, 1\}$ that best approximates the unknown edges of the graph.

Note that the main difficulty for solving this problem is that the features are assigned to individual nodes and the labels to pairs of nodes [9]. To transform the task into a standard classification problem, we use a global approach that tries to find a feature representation for pairs of nodes. Another issue inherent to biological network inference is the substantial class imbalance since the number of positive interactions is small compared to the number of all possible interactions. Thus special care is needed for setting up the evaluation experiments, see e.g. [39]. First of all, the evaluation metrics should be chosen such that the class imbalance does not lead to incorrect conclusions (e.g AUPR metric explained below). Secondly, methods that predict for each protein an interaction profile (see OK3 and IOKR below), represented as a multilabel, a binary vector containing interaction labels for all other proteins, are able to mitigate the class imbalance, since in general the set of multilabels are diverse with no very frequent multilabel. In [9] it is recommended to perform cross validation on the nodes as cross validation on pairs tends to give too optimistic results. A schematic representation of the duality between the biological network and the adjacency matrix and the cross validation on nodes is given in Fig 2.

Finally, for performing inter-species biological network inference we use the protein sequences and their interactions from one species as training set and the protein sequences from the second species as testing set. Note that in this setting the training-testing interactions are not of interest and that the feature representation needs to be the same for training and testing proteins.

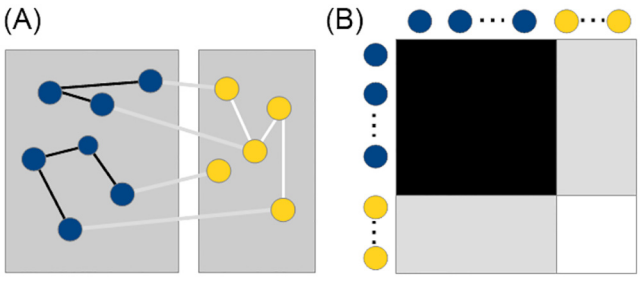

Fig 2. Schematic representation of the duality between (A) the PPI network and (B) the adjacency matrix for the proteins in the training set (blue) and testing set (yellow) and their interactions: training interactions (black), training-testing interactions (gray) and testing interactions (white).

doi:10.1371/journal.pone.0159302.g002

Inference Algorithms

In this section we present three different approaches for supervised network inference that we have applied to inter-species PPI network prediction. Additionally, we present different approaches for learning kernels that account for the relevance of a data source for the learning task.

Output kernel trees (OK3). have been proposed by [18] and are based on the kernel embedding of the graph where the kernel function is defined as k_Y: $V \times V \rightarrow \Re$ with $k_Y(v, v') = \langle \psi(v), \psi(v') \rangle$. The kernel $k_Y(v, v')$ is defined such that adjacent vertices have higher values of k_Y than non-adjacent ones. To achieve this, the diffusion kernel is commonly used $K_Y = \exp(-\beta L)$ where L is the Laplacian matrix of the graph $L = D - Y$ with D being the degree matrix and Y the adjacency matrix. Additionally, $\beta > 0$ is a user defined parameter that controls the diffusion degree.

The OK3 algorithm relies on the top-down induction algorithm widely used to learning decision trees (e.g. CART [40]). The methods start with a tree represented by a single leaf and then recursively partition (or split) the input data S until the data is homogeneous enough (in our case: the proteins in S have similar connectivity patterns). The data arriving to leaf L of the decision tree is split into two parts S_l and S_r, using a binary test $T_t(x) \in \{0, 1\}$ based on a value of a single input feature of x (e.g. does protein have a given motif or not). The two sets $S_l = \{x \in S | T(x) = 0\}$ and $S_r = \{x \in S | T(x) = 1\}$ will be recursively used to grow subtrees which then will be attached as the children of L.

For learning the decision trees on the input vectors $x_i = x(v_i)$, $i = 1..m$ the following score is maximized to select a test T to be inserted in the decision tree leaf given the set of inputs S routed to the current decision tree leaf:

$$Score(T, S) = var\{\psi(v) \mid S\} - \frac{N_l}{N} var\{\psi(v) \mid S_l\} - \frac{N_r}{N} var\{\psi(v) \mid S_r\} \quad (1)$$

where $\psi(v)$ is the output feature vector, N, N_l and N_r are the sizes of the training sample S and its left and right split, S_l and S_r, respectively. The variance of the output feature vectors in the set S can be easily computed using the kernel trick:

$$var\{\psi(v) \mid S\} = \frac{1}{N} \sum_{i=1}^{N} k_Y(v_i, v_i) - \frac{1}{N^2} \sum_{i,j=1}^{N} k_Y(v_i, v_j)$$

One main advantage of the OK3 approach is that the decision tree on the input features results in a ranking of relevant features for the learning task.

Then for prediction each leaf L is labeled with a prediction $\hat{\psi}_L = \frac{1}{N_l} \sum_{i=1}^{N_L} \psi(v_i)$ analog to standard regression trees where N_L are the number of samples that reach the leaf. Finally, the kernel value between two vertices v and v' where $x(v)$ reaches leaf L_1 and $x(v')$ leaf L_2 respectively can be approximated by thresholding

$$\hat{k}_Y(v, v') = \frac{1}{N_{L_1} N_{L_2}} \sum_{i=1}^{N_{L_1}} \sum_{j=1}^{N_{L_2}} k_Y(v_i^1, v_j^2)$$

where $v_i^k, i = 1, \ldots, N_{L_k}$ enumerate the vertices routed to leaf L_k. For improving the accuracy of the method an ensemble of decision trees also known as a random forest is used. In our experiments we used the C code provided by the authors [18].

Kernels on protein pairs. The main idea of the biological network reconstruction methods presented in [8] is to reformulate the task as a pattern recognition problem: given a training

set $\tau = \{(u_1, t_1),(u_2, t_2), .. ,(u_N, t_N)\}$ of patterns $u_i \in \Re^q$ with a binary label $t_i \in \{-1, 1\}$ infer a function f: $\Re^q \rightarrow \{-1, 1\}$ for any new pattern u. The main hindrance in doing so is that in network reconstruction the labels are defined on pairs of vertices and the input features or patterns on individual vertices. Thus in a first step a so called linear kernel on pairs of vertices induced by their input features is defined by their inner product $k_X(v, v') = x(v)^T x(v')$. These kernels k_X represent the similarity of any pair of protein sequences that are then used to compute kernels on pairs of protein pairs as follows

1. Direct product kernel: $k_{DRCT}((a, b), (c, d)) = k_X(a, c) * k_X(b, d)$
2. Tensor product pairwise kernel: $k_{TPPK}((a, b), (c, d)) = k_X(a, c) * k_X(b, d) + k_X(a, d) * k_X(b, c)$
3. Metric learning pairwise kernel: $k_{MLPK}((a, b), (c, d)) = (k_X(a, c) - k_X(a, d) - k_X(b, c) + k_X(b, d))^2$

Now a standard support vector machine (SVM) can be used to solve the binary classification task. Since PPI networks are undirected the tensor product kernel k_{TPPK} and the metric learning pairwise kernel k_{MLPK} are best suited for modelling the similarity between protein pairs.

Despite the method's good predictive performance it has a major drawback: the kernels between pairs of proteins can become quickly very large even for a reasonable amount of protein sequences. The space complexity for storing the kernel matrix turns out to be $O(m^4)$ where m is the number of proteins in the biological network which leads to serious scalability problems and usage of computational resources [21].

Input-Output Kernel Regression (IOKR). This method combines elements of the two previous algorithms that circumvent their respective disadvantages—on the input side it uses the simple kernels on protein pairs and on the output side it uses the diffusion kernel built from the adjacency matrix of the output graph. But the classification problem is addressed by solving a kernel learning problem using regularized regression [19, 41]. The method comes in two flavors: the supervised version learns only the kernel ridge regression model and the semi supervised one adds a smoothness constraint using the inputs of labeled data and auxiliary data, called unlabeled data.

As the OK3 method, IOKR proposes to solve the link prediction problem by learning an output kernel $k_Y : V \times V \rightarrow \mathbb{R}$, that encodes the similarities between the proteins in the interaction network. After learning this kernel, positive interactions can be predicted for the kernel values that are higher than some threshold θ:

$$f_\theta(v, v') = sgn(\hat{k}_Y(v, v') - \theta)$$

As k_Y is a kernel, its values can be written as: $k_Y(v, v') = \langle \psi(v), \psi(v') \rangle$, where ψ is called the output feature map. The IOKR method approximates the output feature map ψ with a function h and then build an approximation of the output kernel k_Y by taking the inner product between the values of this function:

$$\hat{k}_Y(v, v') = \langle h(v), h(v') \rangle .$$

Thus learning f_θ reduces to learn the single variable function h.

Then given models of the general form $h_M(v) = M\phi(v)$ and assuming a regularized square loss function the parameters of the supervised IOKR model can be estimated based on l training samples as follows:

$$\text{argmin}_M \sum_1^l \| h_M(v_i) - \psi(v_i) \|^2 + \lambda_1 \| M \|_F^2$$

where $\lambda_1 > 0$ is a regularization parameter that is tuned with cross validation for the experiments.

The method has also been extended to the semi-supervised setting where the input of unlabeled data is taken into account. The new cost function that has to be minimized is:

$$\text{argmin}_M \sum_1^l \| h_M(v_i) - \psi(v_i) \|^2 + \lambda_1 \| M \|_F^2 + \lambda_2 trace(h_M L_{X_n} h_M^T)$$

where $L_{X_n} = \exp(-\beta(D_n - K_{X_n}))$ denotes the diffusion kernel associated to input kernel matrix on labeled and unlabeled data. The last term constrains proteins that are similar to each other in input to be similar in the predicted interaction network. $\lambda_1 > 0$ and $\lambda_2 > 0$ are two regularization parameters that are tuned with cross validation for the experiments. Both minimisation problems lead to a closed form solution that can be found in Propositions 4 and 6 of [19].

Multiple Kernel Learning (MKL). The heterogeneous set of features that we extracted from the protein sequences is expected not to uniformly contribute information to the learned model which makes the uniform combination of the kernels over the different data sources suboptimal. Therefore we apply Multiple Kernel Learning (MKL) to take the feature's relevance into account. We focus on linear mixtures of kernels,

$$\mathbf{K}_{\boldsymbol{\mu}} = \sum_{q=1}^{r} \mu_q \mathbf{K}_q$$

where the weights μ_q are typically restricted to be non-negative to ensure the PSD property of the resulting mixture. Note that setting $\mu_q = 1$ for all kernels yields the uniform kernel combination. A major step forward in the MKL field was learning kernels based on centered kernel-target alignment [23]

$$\hat{\rho}(\mathbf{K}, \mathbf{K}_Y) = \frac{\langle \mathbf{K}_c, \mathbf{K}_Y \rangle_F}{\| \mathbf{K}_c \|_F \| \mathbf{K}_Y \|_F}$$

where $\langle . \rangle_F$ is the Frobenius product, $\| . \|_F$ the Frobenius norm, $\mathbf{K}_Y$ is a target kernel and $\mathbf{K}_c$ denotes a centered version of the input kernel K, achieved by the centering operation

$$\mathbf{K}_c = \left[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{m} \right] \mathbf{K} \left[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{m} \right]$$

where $\mathbf{1}$ denotes the vector of ones and $\mathbf{I}$ is the identity matrix.

This gives a simple improvement over the uniform combination of kernels be directly using the kernel-target alignment scores $\hat{\rho}(\mathbf{K}_q, \mathbf{K}_Y)$ as a mixture weights:

$$\mathbf{K}_{\boldsymbol{\mu}} \propto \sum_{k=1}^{p} \hat{\rho}(\mathbf{K}_k, \mathbf{K}_Y) \mathbf{K}_k$$

This MKL method is called ALIGN. In [23] it is claimed that the kernel centering is critical for the kernel alignment score to correlate well with performance.

The previously presented independent kernel alignment neglects the correlation between the base kernels which can be overcome by jointly maximization the alignment between the convex combination kernel with the target kernel and is also referred to as ALIGNF:

$$\max_{\boldsymbol{\mu}} \frac{\langle \mathbf{K}_{\boldsymbol{\mu}}, \mathbf{K}_Y \rangle_F}{\| \mathbf{K}_{\boldsymbol{\mu}} \|_F}$$

With the constraints that $\| \boldsymbol{\mu} \|_2 = 1$ and $\boldsymbol{\mu} \geq 0$ the alignment maximization problem can be

rewritten as:

$$\boldsymbol{\mu}^* = \mathbf{argmax}_{\mu} \frac{\boldsymbol{\mu}^T \mathbf{a}\mathbf{a}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \mathbf{M} \boldsymbol{\mu}}$$

where $\mathbf{a} = (\langle \mathbf{K}_1, \mathbf{K}_Y \rangle_F, \ldots, \langle \mathbf{K}_r, \mathbf{K}_Y \rangle_F)^T$ records the kernel-target alignments of the input kernels and $\mathbf{M} = (M_{ql})_{ql}$ with $M_{ql} = \langle \mathbf{K}_q, \mathbf{K}_l \rangle_F$ contains the pairwise kernel alignments between the input kernels. The problem can be solved by quadratic programming [23].

Another approach for optimizing the kernel target alignment has been proposed in [42]. The method aims at sparse combinations of kernels by regularizing the kernel weights by ℓ_p-norm, where $1 \geq p$ is simultaneously optimized. The proposed generalized ℓ_p-norm kernel target alignment formulation is as follows:

$$\min_{\boldsymbol{\mu} \geq 0} \lambda_1 \parallel \boldsymbol{\mu} - \boldsymbol{\mu}_0 \parallel_2^2 + \lambda_2 \sum_{i=1}^{r} \mu_i^p - \sum_{i=1}^{r} \mu_i \mathbf{a}$$

The squared Euclidean distance in the first term is an instantiation of Bregman divergence [42]

$$\bar{B}_F(\boldsymbol{\mu}) = F(\boldsymbol{\mu}) - (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \nabla F(\boldsymbol{\mu}_0)$$

for $F(\boldsymbol{\mu}) = \langle \boldsymbol{\mu}, \boldsymbol{\mu} \rangle$, and $\boldsymbol{\mu}_0$ is a fixed point in the domain of F (Following [42] we used $\boldsymbol{\mu}_0 = 0$ in our experiments.). Additionally, $\lambda_1 \leq 0$ and $\lambda_2 \leq 0$ are the regularization parameters. For implementing the sparsity inducing l_p regularizer p is systematically reduced towards unity till a sufficient level of sparsity is obtained. The solution of the path following is computed with a Predictor-Corrector algorithm [42].

Evaluation metrics for Binary Predictions

Binary classification problems are typically evaluated with the accuracy measure which is computed as the number of correctly predicted pairs divided by the total number of pairs. For highly imbalanced problems like network inference accuracy is not an appropriate measure because it favours the majority class and thus the non-interactions. In the following Receiver-Operator-Characteristic (ROC) and Precision-Recall (PR) curves are presented which are better suited for evaluating network inference predictions [43].

Both measures are based on a so called confusion matrix which is 2 x 2 for binary classification with the columns and rows representing the predicted and the actual classes respectively. Denoting interactions as positive and non-interactions as negative the confusion matrix is given in Table 1.

From this matrix several measures for model evaluation can be derived:

- True positive rate (TPR): also known as sensitivity or recall, is the number of true positives divided the number of the actual positives TP/P

Table 1. Confusion matrix indicating True positive (TP), False positive (FP), False negative (FN) and True negative predictions.

	Ground truth	
	P—Positive	**N—Negative**
Predicted Positive	TP—True Positive	FP—False Positive
Predicted negative	FN—false negative	TN—true negative

doi:10.1371/journal.pone.0159302.t001

- True negative rate (TNR): also known as specificity, is the number of true negatives divided by the number of actual negatives TN/N
- False positive rate (FPR): is the number of false positives divided by the number of actual negatives FP/N
- False negative rate (FNR): is the number of false negatives divided by the number of actual positives FN/P
- Precision: is the number of true positives divided by the number of predicted positives $TP/(TP + FP)$

All of these measures need to be combined in order to give a reliable performance measure of an algorithm e.g. specificity and sensitivity or precision and recall. Note as well that a threshold needs to be defined if predictions are confidence scores. For evaluating algorithms with varying confidence thresholds ROC and PR curves can be used.

ROC curves. plot the TPR over the FPR for varying confidence thresholds. More specifically, each threshold corresponds to a different confusion matrix and thus a different pair of values for TPR and FPR and a point on the ROC curve. The end points are always (0, 0) and (1, 1) and a perfect classifier would pass through the point (0, 1), while a random classifier would be a diagonal connecting (0, 0) and (1, 1). A common summary statistic of the ROC curve is the area under the ROC curve (AUROC). AUROC is one for a perfect classifier and 0.5 for a random one. For the highly imbalanced network prediction tasks even moderate FPR can lead to more FP predictions than TP predictions and hence a very low precision.

PR curves. plot the precision over the recall for varying confidence thresholds. The curve starts at a pseudo point (0, 1) and ends at $(1, P/(P + N))$ which corresponds to to predicting all pairs as positive. An optimal classifier would pass as well through (1, 1). The area under the PR curve (AUPR) is also a common summary statistic. As for AUROC one assumes that the higher the AUPR the better the performance of the method. One advantage of PR curves over ROC curves is that they allow to measure early precision where recall is low and thus gives a tool to evaluate the quality of the top ranks of the result list.

Results

We report here on three sets of experiments. First, we evaluate how the prediction methods perform under simulated sequence data, representing differing amount of genetic distance between the source and target species. Second, we check how well the methods separate the secretory pathway from the rest of the genome. Third, we evaluate the PPI prediction in the cross-species transfer learning from *S. cerevisiae* to *T. reesei*.

Network reconstruction for evolutionary distant sequences

Here we compare the performance of the network inference methods Output Kernel Trees (OK3), Tensor kernel SVM on protein pairs (PP), and supervised and semi-supervised Input-Output Kernel Regression (IOKR) for evolutionary distant species. As training data, we use the *S. cerevisiae* secretory pathway protein sequences as input and their functional interactions as labels. Then we try to predict these interactions in secretory pathway protein sequences that were perturbed using different BLOSUM matrices that correspond to different genetic distances.

Fig 3 depicts the Receiver operating characteristic curves (ROC) with associated area-under-curve (AUC) statistics for each inference method for the different evolutionary distances. As expected, all methods predict the better the smaller the distance with BLOSUM80

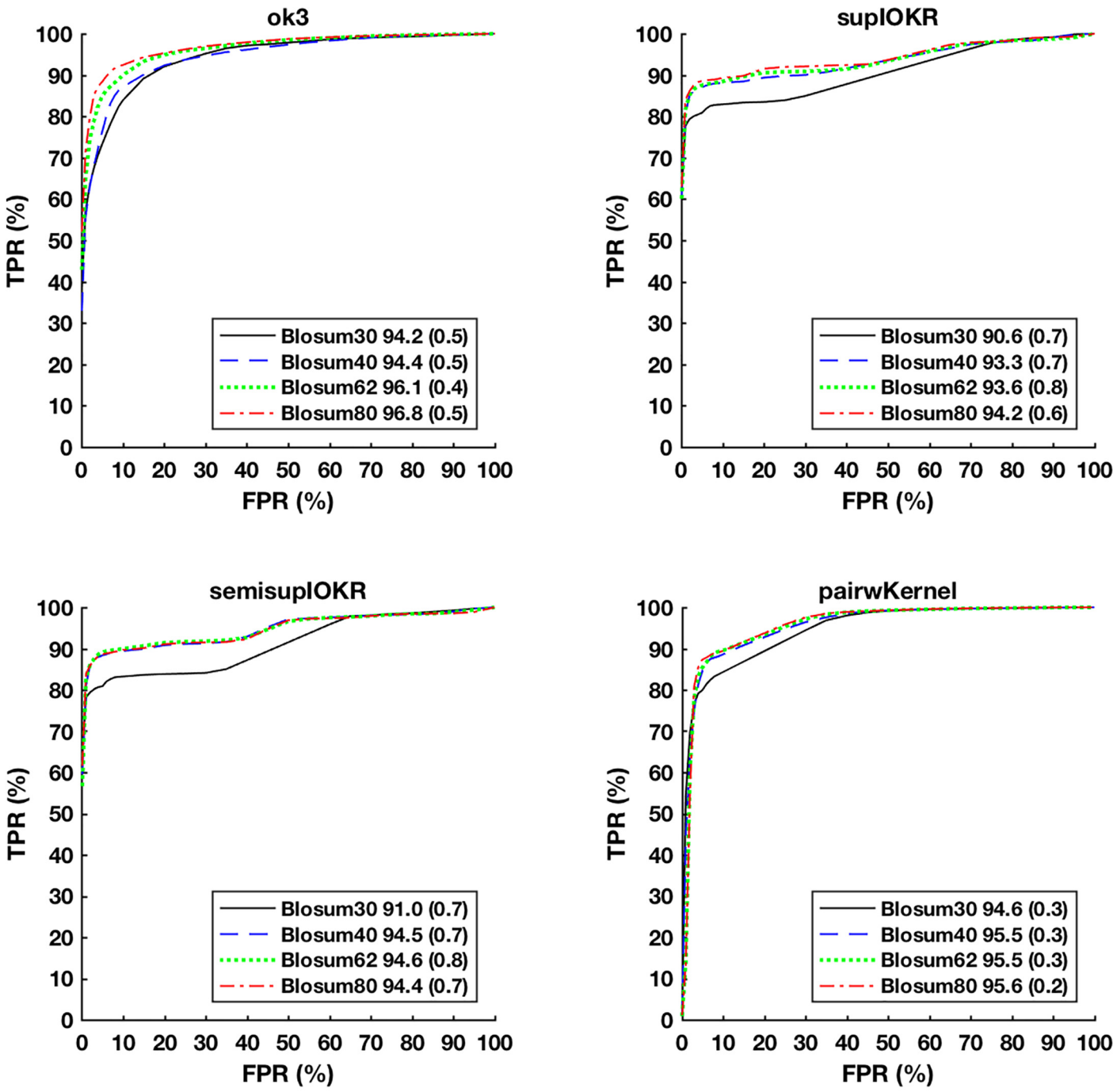

Fig 3. ROC curves for predicting PPIs in different artificial data sets with Output Kernel trees (OK3), Tensor kernels on protein pairs (Tensor Kernel on PP), and supervised and semi-supervised Input-Output Kernel Regression (IOKR). AUROC statistic of the associated curve is depicted in the figure legend (standard deviation in parenthesis).

doi:10.1371/journal.pone.0159302.g003

curves having the highest AUC and being closest to the top-left corner of the plots. The curves are averages of 20-fold cross-validation experiment.

In terms of AUC, OK3 obtains the best results, tensor kernel (PP) the second best and the IOKR methods being somewhat less accurate. However, closer examination of the method's

prediction performance for top-ranked interactions ($FPR < 0.1$) reveals that the IOKR methods in fact have the best early precision, thus would get the top-ranked interactions more accurately predicted than the competing methods.

The AUC statistics and the ROC curves of OK3 follow a smoothly worsening pattern with respect to the increasing evolutionary distance, while the other methods manifest a step change so that BLOSUM30 is markedly worse in AUC and lies clearly below the other curves.

Fig 4 depicts the Precision-Recall (PR) curves of the same experiment. Here, the IOKR methods clearly perform best, having close to perfect precision regardless of the evolutionary distance until recall level of 0.5 and then a sharp drop at recall levels of 0.7–0.9 depending on the evolutionary distance. In contrast, OK3 manifests a close to one precision only for BLOSUM80 and for recall levels up to 0.3. Pairwise kernels do not obtain a high precision and produces a pattern that is inverted with respect to the evolutionary distance, indicating a high number of false positives in the SVM classifier and possible overfitting when the evolutionary distance is small.

The ROC and PR curves together indicate IOKR as the best compromise, given that both a high overall accuracy and high initial precision are desirable for network reconstruction.

Identifying secretory pathway PPIs from full genome

Next, we check how well transfer learning of the secretory pathway works in the basic case of the source and target species being the same. In this experiment, the inference models were trained on the *S. cerevisiae* secretion proteins and their functional interactions, and the goal is to test the ability of the models to correctly identify the secretion pathway proteins among all *S. cerevisiae* proteins. In this setup, the ground truth is composed of PPIs between two secretory pathway proteins as the positive class and all other interactions as the negative class (true interactions between one or two non-secretory proteins as well as missing interactions between pairs of secretory pathway proteins).

Fig 5 depicts the results of a 5-fold cross-validation experiment for the different network inference methods. In the ROC space (left pane), Pairwise kernels and the two IOKR methods are close in performance, with the semi-supervised IOKR being marginally better than the two others. OK3, however, performs significantly worse than the other three methods. In the precision-recall space (right pane), the two IOKR methods are the most robust in the low-recall regime, with the semi-supervised variant maintaining 0.7 precision rate up to 0.6 recall rate. Pairwise kernels and OK3 demonstrate a different pattern: they suffer from a high false positive rate in the low-recall regime, but have a good precision in mid-recall regime, before tailing off.

Analyzing the ROC and PR results together, semi-supervised IOKR emerges as the best compromise, due to its good ROC behaviour and good precision in the low-recall regime. It appears that the semi-supervised aspect gives some protection for the method against false positives in the low to mid-recall levels.

Comparison of Multiple Kernel Learning methods

Next, we compare the different MKL methods ALIGN, ALINGF and p-norm path following on the reconstruction of the set of secretion proteins from the full genome of *S. cerevisiae* when using semi-supervised IOKR as predictor.

The results are shown in Fig 6. It can be seen that the MKL methods perform better than the simple sum of input kernels (UNIMKL) in terms of ROC curve as well as PR curves.

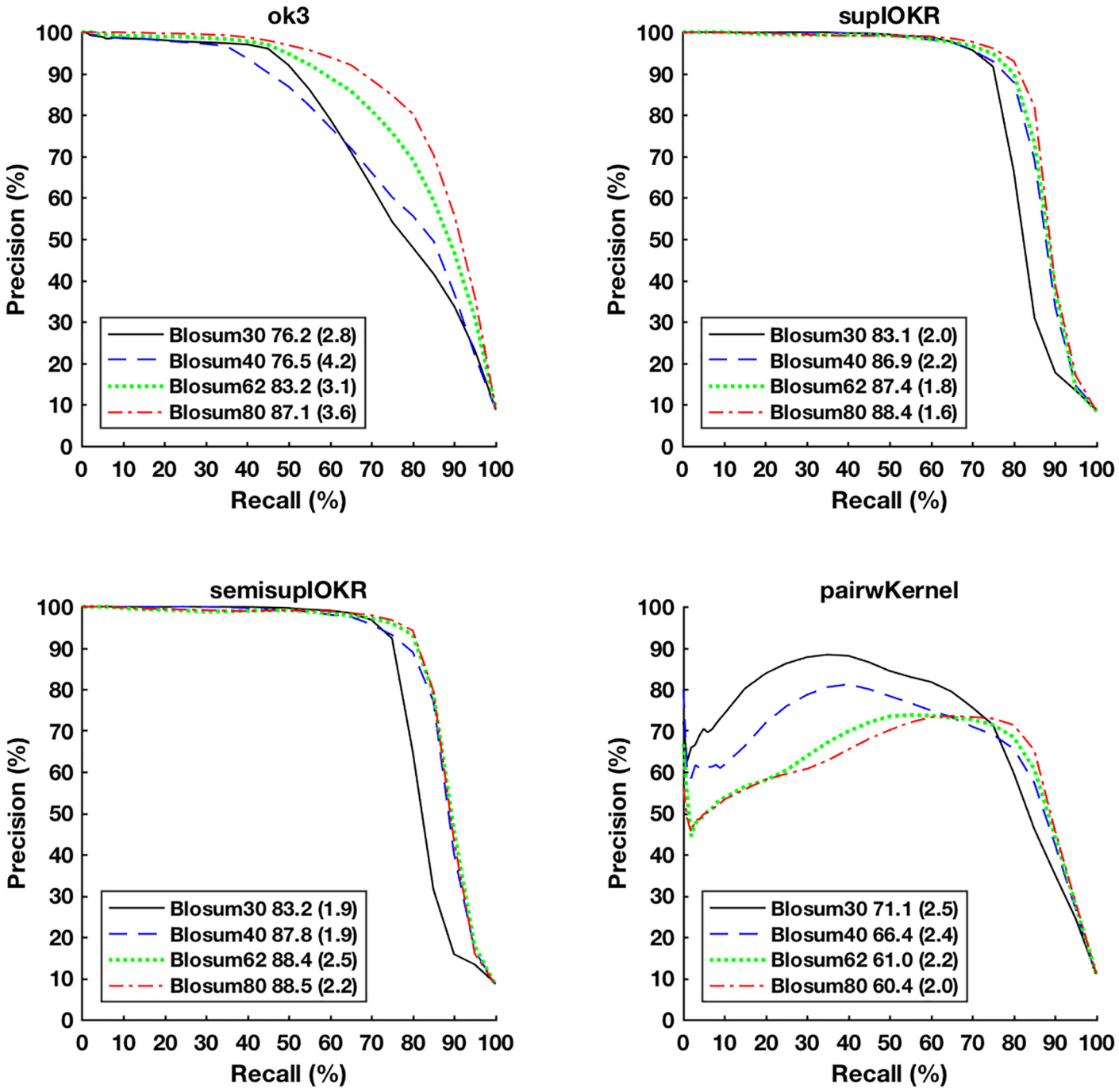

Fig 4. Precision-Recall (PR) curves for predicting PPIs in different artificial data sets with Output Kernel trees (OK3), Tensor kernels on protein pairs (Tensor Kernel on PP), and supervised and semisupervised Input-Output Kernel Regression (IOKR). AUPR statistic is shown in the legend for each curve (standard devation in parenthesis.

doi:10.1371/journal.pone.0159302.g004

Nonetheless, the gains of MKL are smaller than we expected them to be. Looking at the ROC curves, the p-norm path following MKL outperforms the other methods, whereas for the PR measure the simpler ALIGNF outperforms all other methods with p-norm path following being the second best.

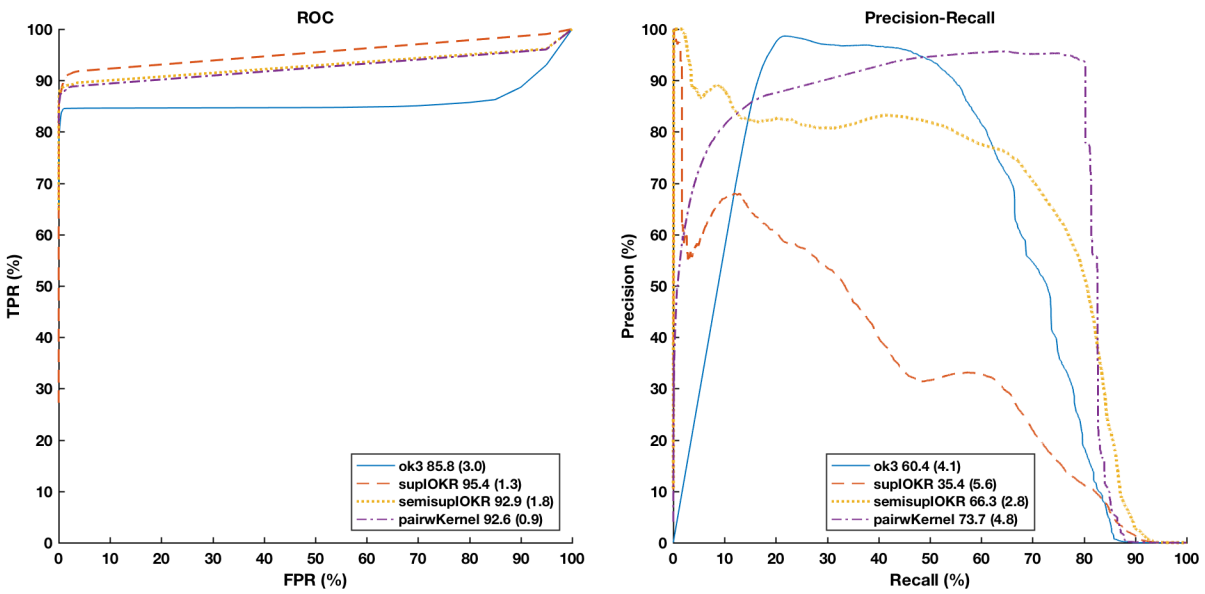

Fig 5. ROC curves and Precision-Recall (PR) curves for predicting secretory PPIs from the full *S. cerevisiae* genome with Output Kernel trees (OK3), Tensor kernels on protein pairs (Tensor Kernel on PP), and supervised and semi-supervised Input-Output Kernel Regression (IOKR). AUCROC and AUPR statistics are shown in the legend for each curve.

doi:10.1371/journal.pone.0159302.g005

Secretion network prediction for *Trichoderma reesei*

Finally, we evaluate the PPI prediction quality in an inter-species setup, where the training data comes from *S. cerevisiae* and the target species is *T. reesei*. However, no experimental protein interaction data exists for *T. reesei* that could be used as the ground truth. Thus we focus on qualitative analysis of the predicted *T. reesei* secretion network by expert knowledge. For

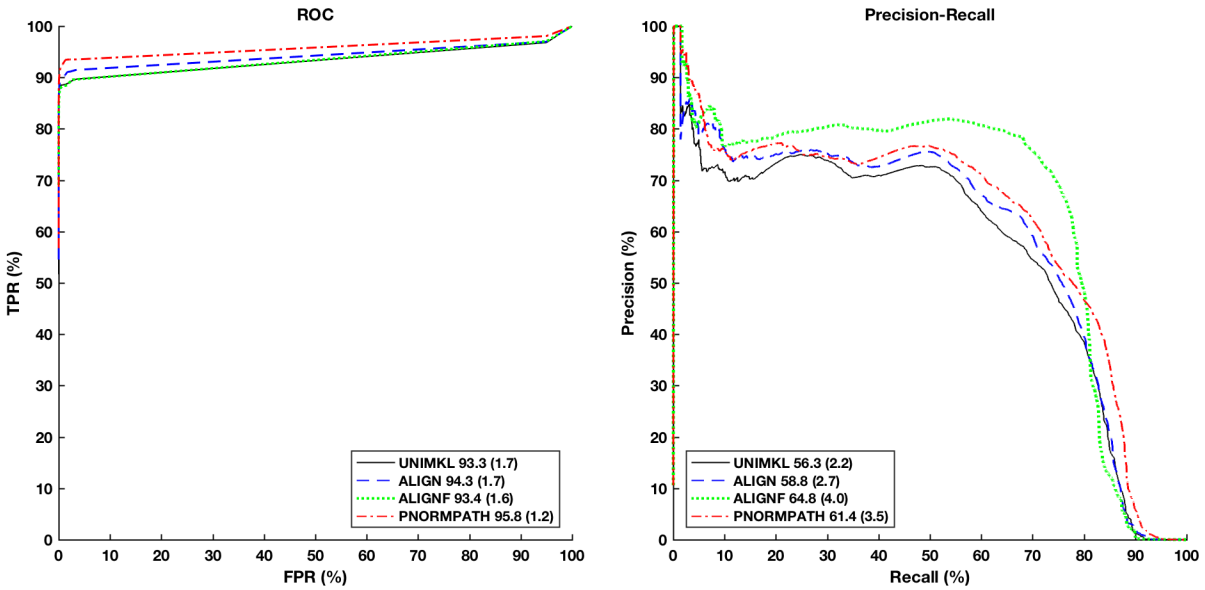

Fig 6. ROC curves and Precision-Recall (PR) curves for predicting secretory PPIs from the full *S. cerevisiae* genome with semi-supervised Input-Output Kernel Regression (IOKR) and different Multiple Kernel Learning (MKL) methods compared to no MKL (UNIMKL). AUCROC and AUPR statistics are shown in the legend for each curve.

doi:10.1371/journal.pone.0159302.g006

predicting the PPI in *T. reesei* we used semi-supervised IOKR and p-norm path following for learning the input kernel, since this method combination achieved the best performances in the previous experiments.

In order to validate the predicted *T. reesei* secretion network, its genes (*T. reesei* genome version 2.0 [44]) were annotated with a combination of sequence similarity based methods: best BLASTp [45] match to *S. cerevisiae* proteins, best BLASTp match to UniProtKB/Swiss-Prot [46], Interproscan domain predictions [26], PANNZER description line and GO-category predictions [47] and a manually curated set of *Aspergillus niger* protein secretion related genes [48].

The *T. reesei* secretion network contains in total 320 genes. According to the annotation described above 27 genes belong to the heterokaryon incompatibly family and are sequence wise very similar. This family contains a GTPase domain that could contain similar features as GTPases involved in secretion. 51 genes belong to other than secretion related categories of cellular function. 18 genes were annotated to be related to cell growth, cell wall synthesis and cell motility and six were found to be related to chromatin modification. In general these 24 proteins contain domains related to small molecule modifications of macromolecules such as glycosylation, phosphorylation, ubiquitinylation and methylation. Similar molecular functions are abundant in the known secretion pathway enzymes. 14 of the 51 were annotated as molecular and cellular function unknown (Column 'Class' in Table 2). Hence, manual annotation based

Table 2. Unknown genes and genes without any interactions in STRING in predicted *T. reesei* secretion network. Column 'Gene' contains the *T. reesei* gene ID. 'In STRING' tells if the gene has interactions in STRING. Columns 'Btw' and 'Deg' denote the betweenness and degree network statistics of the corresponding gene. Columns 'Class' and 'Putative secretion pathway component' are author assigned classifications. 'Taxon specificity' gives the largest taxonomic group the gene was found in.

Gene	STRING	Btw	Deg	Class	Protein family or Molecular function	Putative secretion pathway component	Taxon specificity
104448	NO	4075.6	66	Unknown		Protein folding	Pezizomycotina
111569	NO	217.4	43	Unknown		COPII	Pezizomycotina
112222	NO	1.0	2	Unknown	Tetratricopeptide-like helical	?	Pezizomycotina
46764	NO	0.4	7	Unknown		Protein glycosylation	Sordariomycetes
120000	YES	0.2	11	Unknown	Homeodoman superfamily	Protein folding	Trichoderma
105729	YES	0.0	11	Unknown		GPI-biosynthesis	Trichoderma
122293	NO	0.0	9	Unknown		Protein folding	Pezizomycotina
4791	NO	0.0	7	Unknown		COPII	Pezizomycotina
62041	NO	0.0	5	Unknown		Protein folding	Pezizomycotina
110342	YES	0.0	3	Unknown		Protein glycosylation	Pezizomycotina
104336	NO	0.0	2	Unknown	SnoaL-like domain	COPII	Trichoderma
110791	NO	0.0	2	Unknown		Protein glycosylation	Trichoderma
112490	NO	0.0	1	Unknown		?	Pezizomycotina
47424	YES	0.0	6	Unknown		GPI-biosynthesis	Sordariomycetes
65060	NO	0.0	17	Secretion	Golgi mannosyltransferase complex subunit	Golgi processing	Pezizomycotina
81211/ ANP1	NO	0.0	2	Secretion	Golgi mannosyltransferase complex subunit	Golgi processing	Fungi
78463	NO	0.0	1	Regulatory functions	TPR repeat protein		Pezizomycotina
5275	NO	0.0	1	Chromatin modification			Fungi
27219	NO	0.0	2	Cell growth, wall, motility	Alpha-galactosidase		Trichoderma
65137	NO	0.0	2	Cell growth, wall, motility	Endo-1,3-beta-glucanase		Pezizomycotina

doi:10.1371/journal.pone.0159302.t002

on sequence similarity suggests a minimum of 75% true positive rate and a maximum of 20% false positive rate. The predicted secretion network, excluding heterokaryon incompatibility family and other than secretion related genes in order to ease visual inspection of know genes, is shown in Fig 7. An alternative layout (S1 Fig) and a table format (S2 Table) of the network are also provided as supplementary material.

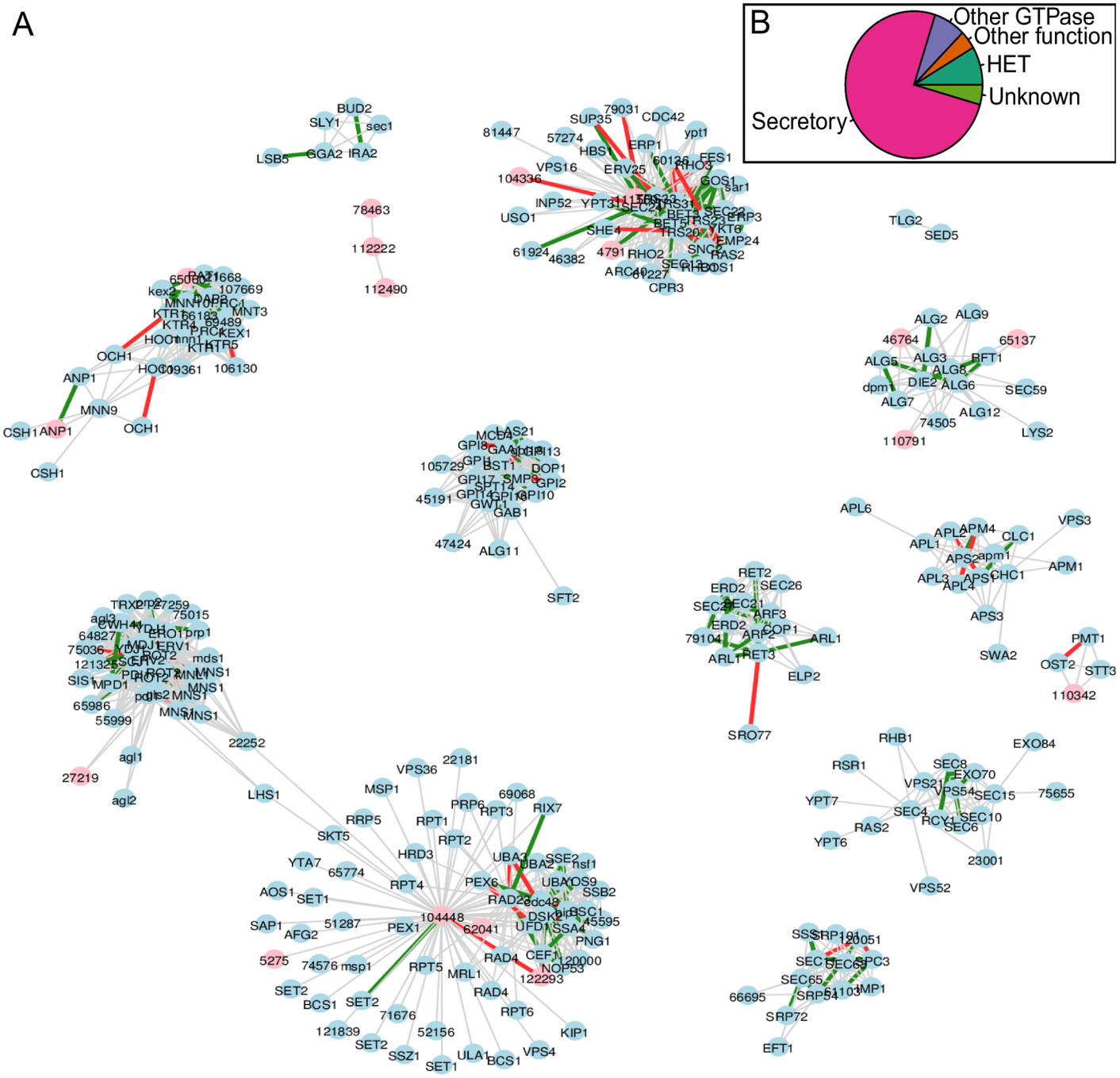

Fig 7. Predicted *T. reesei* secretion network. A) The proteins annotated as secretory (242) and unknown (14) are included. Proteins are nodes and they are labelled with best matching *S. cerevisiae* protein name or if no match was found with *T. reesei* gene ID number. Thick edges signify either negative (red) or positive (green) absolute Pearson correlation of > 0.3 in transcriptomic data. Pink nodes do not have any interactions in STRING. B) Pie chart of functional classes of the 320 proteins included in the *T. reesei* secretion network.

doi:10.1371/journal.pone.0159302.g007

16 of the 320 genes were found to have no interactions in the protein-protein interaction database STRING [4] (Column 'In STRING' in Table 2). For 2 of these we found strong similarity based evidence that they are part of the Golgi mannosyltransferase complex and could be considered as false negative predictions by STRING.

For each gene annotated as unknown, a putative role in the secretion pathway machinery was assigned based on their position in the predicted network (Column 'Putative secretion pathway component' in Table 2). To estimate the novelty of such secretion pathway components the taxonomic distribution of the unknown genes was estimated with multi-genome protein clustering [49] (Column 'Taxon Specificity' in Table 2). All unknown genes were found to be restricted to the subphylum Pezizomycotina or a smaller taxon with-in Pezizomycotina.

In order to further validate the *T. reesei* secretion network we used a combined transcriptomics data set of public and in-house data (see Methods). Pearson correlation of the expression values of all gene pairs that have a predicted PPI (an edge in the PPI network) was computed. The average of absolute values of these correlations was found to be 0.2 with an empirical p-value of $p < 0.05$. This p-value was calculated by rewiring the network 1000 times with the igraph function 'rewire' [50] and counting the average of absolute correlation each time. Absolute correlations above 0.3 are highlighted in Fig 7.

Discussion

Experimental measurement of protein-protein interactions is technically demanding and often different methods can give conflicting results [7, 51]. Also, even a reliably measured interaction might not have a detectable biological function. To circumvent such challenges we use an expert curated interaction network of functional associations derived from numerous experiments [24].

We tested several recent machine learning methods for the task of PPI prediction. Classification models tested included pairwise kernels, output kernel trees as well as supervised and semi-supervised input-output kernel regression. The methods differed in performance depending on whether ROC or PR was used as the evaluation metric. Semi-supervised IOKR proved to be the best compromise when both evaluation metrics were taken into account: it had the best PR performance and a reasonable ROC—this choice puts an emphasis on good performance in the positive class, required for reliable network reconstruction.

Multiple kernel learning methods tested included uniform kernel combination, methods based on centered kernel alignment as well as the newly proposed p-norm path following algorithm. In our tests, we found that generally p-norm path following performed best in the ROC metric while other methods were close to each other in performance. In the PR metric, ALIGNF outperformed the other methods and p-norm path following being second best. Altogether, p-norm path following seems to give the best performance, although the improvements of MKL over no MKL were smaller than expected.

To demonstrate our prediction approach we predict protein secretion network for *T. reesei*, an industrially important protein production organism, which has no experimentally verified PPIs to date. Novel understanding of their protein secretion network machinery could have significant impact in the generation of improved protein production strains through targeted engineering.

For *T. reesei* we find that the predicted network is well supported by sequence similarity based manual annotation and by transcriptomics data. Most importantly the predicted network includes 14 previously unknown genes that are taxonomically restricted to Pezizomycotina and hence could explain their exceptional protein secretion capabilities.

Finally we note that our set up does not need complex external database systems or specialized experimental data to be generated, but relies on data available through standard sequence searches, evaluated through fast machine learning models. Hence, our methods are amenable to local implementation as part of a genome annotation pipeline.

Supporting Information

S1 Table. Protein feature data sources. Protein feature data sources used in training the PPI prediction models.
(XLSX)

S2 Table. Trichoderma reesei secretion network. Predicted protein-protein interactions of Trichoderma reesei secretion pathway.
(XLSX)

S1 Fig. Alternative layout of predicted *T. reesei* secretion network. In this layout interactions of individual genes are easier to inspect with the cost of less clear overall structure.
(TIFF)

Acknowledgments

This work has been supported in part by the European Union FP7 Cooperation Work programme (grant 289126: BIOLEDGE) and the Finnish Funding Agency for Innovation TEKES under the Living Factories strategic opening (dno:562/31/2014).

Author Contributions

Conceived and designed the experiments: JK MA JR SC MO JJ MP. Performed the experiments: JK MA SC. Analyzed the data: JK MA JR SC. Contributed reagents/materials/analysis tools: TP MO CB. Wrote the paper: JK MA JR CB SC MO.

References

1. Schekman R (2010) Charting the secretory pathway in a simple eukaryote. Molecular biology of the cell 21: 3781–3784. doi: 10.1091/mbc.E10-05-0416 PMID: 21079008
2. Cherry JR, Fidantsef AL (2003) Directed evolution of industrial enzymes: an update. Current opinion in biotechnology 14: 438–443. doi: 10.1016/S0958-1669(03)00099-5 PMID: 12943855
3. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, et al. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature biotechnology 32: 822–828. doi: 10.1038/nbt.2939 PMID: 24997787
4. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, et al. (2014) String v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research: gku1003.
5. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC (2009) More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics 25: 404–413. doi: 10.1016/j.tig.2009.07.006 PMID: 19716618
6. Shoemaker BA, Panchenko AR (2007) Deciphering protein–protein interactions. part ii. computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3: e43. doi: 10.1371/journal.pcbi.0030043 PMID: 17465672
7. Browne F, Zheng H, Wang H, Azuaje F (2010) From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions. Advances in Artificial Intelligence 2010: 7. doi: 10.1155/2010/924529
8. Vert JP (2010) Reconstruction of biological networks by supervised machine learning approaches. Elements of Computational Systems Biology: 165–188.
9. Schrynemackers M, Küffner R, Geurts P (2013) On protocols and measures for the validation of supervised methods for the inference of biological networks. Frontiers in genetics 4: 262. doi: 10.3389/fgene.2013.00262 PMID: 24348517

10. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, et al. (2004) Annotation transfer between genomes: protein–protein interologs and protein–dna regulogs. Genome research 14: 1107–1118. doi: 10.1101/gr.1774904 PMID: 15173116
11. Lewis AC, Jones NS, Porter MA, Deane CM (2012) What evidence is there for the homology of protein-protein interactions? PLoS computational biology 8: e1002625. doi: 10.1371/journal.pcbi.1002645
12. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, et al. (2005) Sequencing of aspergillus nidulans and comparative analysis with a. fumigatus and a. oryzae. Nature 438: 1105–1115. doi: 10.1038/nature04341 PMID: 16372000
13. Gabaldón T, Koonin EV (2013) Functional and evolutionary implications of gene orthology. Nature Reviews Genetics 14: 360–366. doi: 10.1038/nrg3456 PMID: 23552219
14. Vert JP, Yamanishi Y (2004) Supervised graph inference. Advances in Neural Information Processing 17: 1433–1440.
15. Lin C, Chen W, Qiu C, Wu Y, Krishnan S, et al. (2014) Libd3c: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing 123: 424–435. doi: 10.1016/j.neucom.2013.08.004
16. Zou Q, Li X, Jiang Y, Zhao Y, Wang G (2013) Binmempredict: a web server and software for predicting membrane protein types. Current Proteomics 10: 2–9. doi: 10.2174/1570164611310010002
17. Zeng J, Li D, Wu Y, Zou Q, Liu X (2016) An empirical study of features fusion techniques for protein-protein interaction prediction. Current Bioinformatics 11: 4–12. doi: 10.2174/1574893611666151119221435
18. Geurts P, Touleimat N, Dutreix M, d'Alché-Buc F (2007) Inferring biological networks with output kernel trees. BMC Bioinformatics 8(Suppl 2). doi: 10.1186/1471-2105-8-S2-S4 PMID: 17493253
19. Brouard C, Szafranski M, d'Alché-Buc F (2011) Semi-supervised penalized output kernel regression for link prediction. In: Proceedings of the 28th International Conference on Machine Learning. pp. 593–600.
20. Karimpour-Fard A, Detweiler C, Erickson K, Hunter L, Gill R (2007) Cross-species cluster co-conservation: a new method for generating protein interaction networks. Genome Biology 8. doi: 10.1186/gb-2007-8-9-r185 PMID: 17803817
21. Kashima H, Yamanishi Y, Kato T, Sugiyama M, Tsuda K (2009) Simultaneous inference of biological networks of multiple species from genome-wide data and evolutionary information: a semi-supervised approach. Bioinformatics 25: 2962–2968. doi: 10.1093/bioinformatics/btp494 PMID: 19689962
22. Pitkänen E, Jouhten P, Hou J, Syed MF, Blomberg P, et al. (2014) Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. PLoS computational biology 10: e1003465. doi: 10.1371/journal.pcbi.1003465 PMID: 24516375
23. Cortes C, Mohri M, Rostamizadeh A (2012) Algorithms for learning kernels based on centered alignment. Journal of Machine Learning Research 13: 795–828.
24. Feizi A, Österlund T, Petranovic D, Bordel S, Nielsen J (2013) Genome-scale modeling of the protein secretory machinery in yeast. PLoS computational biology 8: e63284.
25. Heger A, Mallick S, Wilton C, Holm L (2007) The global trace graph, a novel paradigm for searching protein sequence databases. Bioinformatics 23: 2361–2367. doi: 10.1093/bioinformatics/btm358 PMID: 17823134
26. Jones P, Binns D, Chang HY, Fraser M, Li W, et al. (2014) Interproscan 5: genome-scale protein function classification. Bioinformatics 30: 1236–1240. doi: 10.1093/bioinformatics/btu031 PMID: 24451626
27. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The pfam protein families database. Nucleic Acids Research 40: D290–D301. doi: 10.1093/nar/gkr1065 PMID: 22127870
28. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD (2016) Panther version 10: expanded protein families and functions, and analysis tools. Nucleic acids research 44: D336–D342. doi: 10.1093/nar/gkv1194 PMID: 26578592
29. Yeats C, Lees J, Reid A, Kellam P, Martin N, et al. (2008) Gene3d: comprehensive structural and functional annotation of genomes. Nucleic acids research 36: D414–D418. doi: 10.1093/nar/gkm1019 PMID: 18032434
30. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, et al. (2012) The prints database: a fine-grained protein sequence annotation and analysis resource—its status in 2012. Database 2012. doi: 10.1093/database/bas019
31. Sigrist CJ, De Castro E, Cerutti L, Cuche BA, Hulo N, et al. (2012) New and continuing developments at prosite. Nucleic acids research: gks1067.
32. Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH (2006) Pirsf family classification system for protein functional and evolutionary analysis. Evolutionary Bioinformatics 2.

33. Letunic I, Doerks T, Bork P (2015) Smart: recent updates, new developments and status in 2015. Nucleic acids research 43: D257–D260. doi: 10.1093/nar/gku949 PMID: 25300481

34. de Lima Morais DA, Fang H, Rackham OJL, Wilson D, Pethica R, et al. (2011) Superfamily 1.75 including a domain-centric gene ontology method. Nucleic Acids Research 39: D427–D434. doi: 10.1093/nar/gkq1130 PMID: 21062816

35. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 89: 10915–10919. doi: 10.1073/pnas.89.22.10915

36. Edgar R, Domrachev M, Lash AE (2002) Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic acids research 30: 207–210. doi: 10.1093/nar/30.1.207 PMID: 11752295

37. Bolstad B (2001) Probe level quantile normalization of high density oligonucleotide array data. Unpublished manuscript.

38. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8: 118–127. doi: 10.1093/biostatistics/kxj037 PMID: 16632515

39. Song L, Li D, Zeng X, Wu Y, Guo L, et al. (2014) ndna-prot: identification of dna-binding proteins based on unbalanced classification. BMC bioinformatics 15: 1. doi: 10.1186/1471-2105-15-298

40. Beiman L, Friedman J, Olsen R, Stone C (1984) Classification and regression trees. Wadsworth International.

41. Brouard C, d'Alché-Buc F, Szafranski M (2015) Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels. Technical Report hal-01216708, The open archive HAL.

42. Jawanpuria P, Varma M, Nath S (2014) On p-norm path following in multiple kernel learning for non-linear feature selection. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). pp. 118–126.

43. Davis J, Goadrich M (2006) The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp. 233–240.

44. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, et al. (2008) Genome sequencing and analysis of the biomass-degrading fungus trichoderma reesei (syn. hypocrea jecorina). Nature biotechnology 26: 553–560. doi: 10.1038/nbt1008-1193a PMID: 18454138

45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215: 403–410. doi: 10.1016/S0022-2836(05)80360-2 PMID: 2231712

46. Consortium U, et al. (2014) Uniprot: a hub for protein information. Nucleic Acids Research: gku989.

47. Koskinen P, Törönen P, Nokso-Koivisto J, Holm L (2015) Pannzer-high-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics: btu851.

48. Wortman JR, Gilsenan JM, Joardar V, Deegan J, Clutterbuck J, et al. (2009) The 2008 update of the aspergillus nidulans genome annotation: a community effort. Fungal Genetics and Biology 46: S2–S13. doi: 10.1016/j.fgb.2008.12.003 PMID: 19146970

49. Gasparetti C, Faccio G, Arvas M, Buchert J, Saloheimo M, et al. (2010) Discovery of a new tyrosinase-like enzyme family lacking a c-terminally processed domain: production and characterization of an aspergillus oryzae catechol oxidase. Applied microbiology and biotechnology 86: 213–226. doi: 10.1007/s00253-009-2258-3 PMID: 19798497

50. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems: 1695.

51. Bonetta L (2010) Protein-protein interactions: Interactome under construction. Nature 468: 851–854. doi: 10.1038/468851a PMID: 21150998