Systems biology

PINALOG: a novel approach to align protein interaction networks—implications for complex detection and function prediction

Hang T. T. Phan and Michael J. E. Sternberg*
Division of Molecular Biosciences, Faculty of Natural Sciences, Imperial College, London SW7 2AZ, UK

ABSTRACT

Motivation: Analysis of protein–protein interaction networks (PPINs) at the system level has become increasingly important in understanding biological processes. Comparison of the interactomes of different species not only provides a better understanding of species evolution but also helps with detecting conserved functional components and in function prediction.

Method and Results: Here we report a PPIN alignment method, called PINALOG, which combines information from protein sequence, function and network topology. Alignment of human and yeast PPINs reveals several conserved subnetworks between them that participate in similar biological processes, notably the proteasome and transcription related processes. PINALOG has been tested for its power in protein complex prediction as well as function prediction. Comparison with PSI-BLAST in predicting proteasome and transcription related processes. PINALOG has been tested for its power in protein complex prediction as well as function prediction. Comparison with PSI-BLAST in predicting proteasome and transcription related processes. PINALOG has been tested for its power in protein complex prediction as well as function prediction. Comparison with PSI-BLAST in predicting proteasome and transcription related processes. PINALOG has been tested for its power in protein complex prediction as well as function prediction. Comparison with PSI-BLAST in predicting proteasome and transcription related processes. PINALOG has been tested for its power in protein complex prediction as well as function prediction. Comparison with PSI-BLAST in predicting proteasome and transcription related processes.

Availability and implementation: The PINALOG web-server is freely available from http://www.sbg.bio.ic.ac.uk/~pinalog. The PINALOG program and associated data are available from the Download section of the web-server.

Contact: m.sternberg@imperial.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on August 24, 2011; revised on February 11, 2012; accepted on March 6, 2012

1 INTRODUCTION

In recent years, there has been a substantial increase in the quantity of protein–protein interactions (PPIs) being detected by experimental methods. The analysis of PPI networks (PPINs) at the system level has become increasingly important in understanding biological processes. Comparison of the interactomes of different species not only provides a better understanding of species evolution but also helps with detecting conserved functional components and in function prediction. Several methods have been developed to align interactomes, both globally and locally, most of which use sequence similarity or network topology or both in establishing the equivalence.

PathBLAST (Kelley et al., 2003) marked the first local alignment method and was similar in principle to the BLAST search algorithm.

*To whom correspondence should be addressed.

Other local alignment methods include: NetworkBLAST (Sharan et al., 2005), which is the next generation of PathBLAST; MaWISH (Koyutürk et al., 2006), which adopts the evolutionary models of match, mismatch and deletion of the proteins; Graemlin 2.0 (Flannick et al., 2006), which infers an alignment from network modules; the Bayesian method (Berg and Lassig, 2006); the match and split algorithm (Narayanan and Karp, 2007); and Phunkee (Coote et al., 2007), which aligns proteins based on sequence and context in the networks. The global alignment of networks proves to be more challenging due to the complexity and scale of the problem. Graemlin 2.0 (Flannick et al., 2006) formulates a model for protein duplication, deletion and mutation and aligns the network progressively using a hill-climbing algorithm. The Markov-random field-based method (Bandypadhyay, 2006) and IsoRank (Singh et al., 2008) solve the problem by eigenvalue-based methods. GRAAL (Kuchaiev et al., 2010) uses only the network topological similarity to align the networks using a seed-and-extend method for graph alignment.

In many applications of PPIN alignment, the underlying objective is to equivalence proteins with related function and interacting partners. However, in most alignment methods only the sequence data or network topology are used to align the input networks despite the availability of other sources of information such as function annotation or gene expression. While sequences are informative in elucidating the orthologous relationships of proteins across species, they do not necessarily indicate functional similarity. The omission of function information could therefore result in many pairs of equivalenced proteins having little or no similarity of function. This in turn makes it less accurate in detecting conserved functional modules or predicting protein function. Recently, Ali and Deane (2009) have introduced functional similarity of proteins into a local alignment method. They used functional similarity in combination with sequence similarity in the local alignment match and split algorithm (Narayan and Karp, 2007) and identified similar functional subnetworks. The recently developed method MI-GRAAL (Kuchaiev and Przulj, 2011) presents a global alignment algorithm which in different information can be incorporated such as topological features, sequence similarity and functional similarity. However, there has been no assessment on how the inclusion of functional similarity might influence the resulting alignments and any subsequent applications. Here we describe PINALOG, a global network alignment algorithm. PINALOG forms the alignment between two PPINs based on the similarities of protein sequence and the protein function between the two networks. Functional similarity is formalized...
The design of the PINALOG alignment method is based on the observation of similarity in the extension mapping process. The sequence similarity of two functional similarity in the seed identification and with additional topological information. Node scoring scheme is performed by direct inheritance of protein complex data from a known species to an unknown one. PINALOG has been tested for complex prediction power using a human-yeast alignment. Protein function prediction is via direct transfer to an un-annotated protein from an annotated protein. The power of our function prediction is assessed in a cross-validated study.

2 METHODS

The design of the PINALOG alignment method is based on the observation of proteins that are not uniformly distributed throughout PPINs. Instead there are some proteins that form well connected subnetworks. Furthermore, it has been shown by Brohee and van Helden (2006) that protein complexes and functional modules form highly connected components in the PPINs. Therefore, it would be more reliable and efficient to align two PPINs by first finding highly similar protein pairs (seed protein pairs) from these highly connected protein subnetworks (referred to as communities from now on) in the networks and then extending the alignment to other proteins in the neighborhood of seed protein pairs.

Let A and B be the PPINs for species A and B. \( a_i \) and \( b_j \) are the \( i \)-th and \( j \)-th proteins in PPIN A and B, respectively. \( C^k_A \) is the \( k \)-th community in PPIN A, and \( C^k_B \) is the \( j \)-th community in B. PINALOG aligns A and B in three steps as summarized in Figure 1: community detection, community mapping and extension mapping.

**Step 1: Community detection of input networks using CFinder**

Communities in biological networks such as PPINs often indicate functional groupings of proteins in the network (Adamscek et al., 2006; Lewis et al., 2010; Song and Singh, 2009). There are several methods to detect communities within a network such as the minimum-cut method, hierarchical clustering or the Girvan-Newman algorithm. In PINALOG, we employ CFinder (Palla et al., 2005), a clique percolation method, which is capable of detecting overlapping communities. Denote a \( k \)-clique as a subgraph of the network composed of \( k \) proteins where all pairs of proteins interact. Then a \( k \)-clique community is defined as union of all \( k \)-cliques that are reachable from each other through adjacent \( k \)-cliques (two adjacent \( k \)-cliques have \( k-1 \) proteins in common). CFinder constructs communities by merging adjacent \( k \)-cliques (see Supplementary Material for more details).

**Node scoring scheme**

The similarity between two proteins is a combination of sequence and functional similarity in the seed identification and with additional topological similarity in the extension mapping process. The sequence similarity of two proteins \( a_i \) and \( b_j \) is calculated based on their BLAST bit score as

\[
S_{seq}(a_i, b_j) = \frac{S(b_j, a_i)}{S_{max}}
\]

where \( S(b_j, a_i) \) is the BLAST bit score value when aligning \( a_i \) and \( b_j \) and \( S_{max} \) is the highest score among all BLAST bit scores obtained. The functional similarity of two proteins annotated with GO terms is calculated based on the semantic similarity (Lord et al., 2003) of terms in \( g(a_i) \) and \( g(b_j) \). Semantic similarity depends on the rarity of the terms in the annotation database as well as their distance in the ontology. The semantic similarity between each term in \( g(a_i) \) and each term in \( g(b_j) \) is calculated, resulting in a \( k \times l \) table of semantic similarities. Let \( \text{rowScore} \) and \( \text{colScore} \) be the average of the row maxima and the average of column maxima in the table, respectively (see Supplementary Fig. S1-S3), the Schlicker’s similarity between \( a_i \) and \( b_j \) is then calculated as

\[
S_{func}(a_i, b_j) = \max(\text{rowScore}, \text{colScore})
\]

The similarity of two proteins in Step 2 is defined as a linear combination of functional similarity and sequence similarity

\[
S(a_i, b_j) = \theta S_{seq}(a_i, b_j) + (1 - \theta) S_{func}(a_i, b_j)
\]

In PINALOG \( \theta \) provides a relative weighting between sequence and functional similarity. The value of \( \theta \) is automatically decided based on the number of reciprocal best BLAST hits in the protein sequences between two input species, reflecting the closeness of two species and thus the contribution of sequence similarity in the overall score. Details of how \( \theta \) is calculated are given in Supplementary Material.

**Step 2: Community mapping to obtain seed protein pairs**

Community mapping is determined in two steps. The first step is to determine the highest score when equivalencing proteins in each community in species A with each community in species B. To this end, we define the similarity between two communities as the score \( F(C^k_A, C^m_B) \), the sum of similarities between protein pairs in the optimal equivalence (OptMap) of proteins in community \( C^k_A \) in species A with proteins in \( C^m_B \) in species B using the Hungarian method (Kuhn, 2005). The optimal equivalence is the mapping where the sum of protein pair similarities is largest. The Hungarian algorithm
is a combinatorial optimization algorithm solving the assignment problem in polynomial time (see Supplementary Material).

\[
F(c_i^a, c_i^b) = \sum_{a_i \in c_i^a, b_i \in c_i^b} n(a_i, b_i) \quad (4)
\]

The first step produces a community similarity matrix. This matrix is used in the second step to obtain the best equivalence of communities by maximizing the community scoring function:

\[
F(\text{core}) = \sum_{c_i^a \subseteq A, c_i^b \subseteq B} F(c_i^a, c_i^b) \quad (5)
\]

The maximization of \( F(\text{core}) \) is achieved by Hungarian method based on the community similarity matrix. For each pair of mapped communities, the Hungarian mapping of proteins between them are obtained and added to the list of seed protein pairs. A filtering step is used to retain only the top 15% pairs of mapped proteins as the seeds for extension, known as the core equivalences.

**Step 3. Extension mapping**

Extension mapping is performed by considering neighbours of proteins in the core equivalence as candidates for adding to the alignment. In addition to protein sequence and functional similarity, topological similarity in the PPINs is included in the form of neighbourhood similarity. Let \( N(a) \) and \( N(b) \) be the set of all first neighbours (proteins separated by one interaction) and second neighbours (proteins separated by two interactions) of \( a \) and \( b \) in \( E \). Let \( d(a_i, a_j) \) denote the distance between \( a_i \) and \( a_j \) in a network. The similarity between \( a \) and \( b \) in extension mapping is then defined as:

\[
\sum_{a_i \in N(a), b_j \in N(b)} \frac{1}{d(a_i, a_j) + 1 + d(b_j, b_i) + 1} \quad (6)
\]

By awarding the candidate pairs with a proportion of the score of neighbouring aligned pairs, PINALOG aims at mapping more neighbouring protein pairs. The Hungarian method is used to find the optimal equivalence candidates. These candidates are then added to the equivalences in the core. The process is repeated until no more pairs are added (see Supplementary Fig S1-1). As CFinder identifies communities that can be overlap, community mapping can result in many-to-many mappings in seed protein pairs, thus in the final alignment.

**3 RESULTS**

**3.1 Alignment of human PPIN with other species**

We have aligned different pairs of PPINs from human, yeast, fly, worm and mouse using: PINALOG; two state-of-the-art alignment methods IsoRank and Graemlin 2.0, the integrative method MIGRAAL and the naive BLAST approach by reciprocal best BLAST hit (denoted as BLAST). The PPINs of these species were obtained from IntAct (Aranda et al., 2010) and aligned by both PINALOG and IsoRank. Graemlin 2.0 requires a training set to learn the parameters for the alignment. Thus comparison with Graemlin can only be made using the data set in the package, which provides PPINs of fly and yeast (denoted as fly2 and yeast2). As IsoRank and MIGRAAL generate alignments with one-to-one mapping, for ease of comparison we reduce the final alignment of PINALOG to a one-to-one mapping (see Supplementary Material).

The aim of PINALOG is to obtain the best equivalence considering a balance of sequence homology, functional similarity and network topology. Thus one would expect differences in the alignment generated by a purely sequence-based approach and one from a network strategy such as PINALOG. However, there is no gold standard with which to compare the results, so we need to consider a range of metrics. Accordingly, in Table 1 and in Supplementary Material, we report NA, the number of aligned protein pairs; NC, the number of conserved interactions; NH, the number of protein pairs belonging to the same Homologene groups (Wheeler et al., 2005); NI, the number of interlogs (Walhout et al., 2000); and NF, the number of aligned protein pairs with functional similarity >0.5. Conserved interactions are interactions occurring in both species when two protein nodes forming an interaction in one species are equivalent to two protein nodes which also form an interaction in the other species. NH is a common measure of alignment quality, counting the number of protein pairs belonging to the same homologene groups identified by the Homologene algorithm. ‘Interlog’ refers to an orthologous pair of interacting proteins between different species. We use the definition of interlog by Yu, et al. (2004) to quantify NI. Both NH and NI describe the sequence similarity of protein pairs in the alignments. If the functional similarity between two proteins is >0.5, their functions are considered related (Schlicker et al., 2006), hence NF.

**Table 1. Alignment results of different pairs of species by PINALOG (PA) (obtained with the automatically generated \( \theta = 0.871 \), IsoRank (IR), MI-GRAAL (MG) and BLAST (BL) for human/yeast**

<table>
<thead>
<tr>
<th>Stat</th>
<th>PA</th>
<th>IR</th>
<th>MG</th>
<th>BL</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>5222</td>
<td>5674</td>
<td>5674</td>
<td>1818</td>
</tr>
<tr>
<td>NC</td>
<td>3319</td>
<td>717</td>
<td>4107</td>
<td>530</td>
</tr>
<tr>
<td>NH</td>
<td>3149</td>
<td>734</td>
<td>1347</td>
<td></td>
</tr>
<tr>
<td>NI</td>
<td>454</td>
<td>165</td>
<td>0</td>
<td>818</td>
</tr>
<tr>
<td>NF/NA</td>
<td>0.09</td>
<td>0.023</td>
<td>0</td>
<td>0.45</td>
</tr>
<tr>
<td>NI/NC</td>
<td>0.14</td>
<td>0.19</td>
<td>0</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The statistics (Stat) are NA, NC, NF, NH and NI which denote the number of aligned pairs of proteins, the number of conserved interactions, the number of protein pairs with functional similarity >0.5 and the number of Homologene pairs and interlogs. The ratios NI/NC and NF/NA are also given.
A common connected subgraph (CCS) from the alignment is defined of proteins in different species sequence similarity (blastp) of proteins in different species together with other information such as phylogeny and synteny. Since network-based methods equivalence pairs of protein together with their interaction edge, one would expect that a sequence method would identify more Homologenes than a network-based approach.

The PINALOG alignment has 3.5 times as many interlogs (NI) as IsoRank (460 versus 136) while MI-GRAAL fails to find any. This indicates that PINALOG finds many interlogs between species, which along with conserved interactions, might contribute to the functional similarity of PPINs across species. BLAST finds a few more interlogs than PINALOG (465 compared with 460). Considering one-to-one mapping of networks, we can take the NI of BLAST as the maximal number obtainable. Clearly PINALOG found most of these whereas IsoRank did not. However as PINALOG generated far more conserved interactions (NC) than IsoRank, it is helpful to consider the ratio of NI to NC. This ratio is 0.14 for PINALOG, 0.19 for IsoRank and 0.88 for BLAST. If the sole aim is to find interlogs, then BLAST is a superior approach to any of these three network alignment methods. However, the PINALOG alignment aims to establish equivalences based on function and interactions not just those from sequence.

We also assess whether these conserved edges produce large and dense connected subgraphs which is helpful in finding topologically similar areas of the aligned networks indicating functional similarity. A common connected subgraph (CCS) from the alignment is defined as a connected subgraph of the conserved network. The largest CCS from the conserved graph obtained from PINALOG is composed of 1858 proteins connected by 2774 interactions; while that of IsoRank has only 44 proteins connected by 87 interactions (Supplementary Material). Functional clustering of the human conserved network shows the advantage of the PINALOG alignment over MI-GRAAL followed by mapping onto the yeast clusters shows that there is little or no functional similarity between corresponding clusters. The average functional similarity between corresponding clusters of MI-GRAAL is 0.16 as compared with PINALOG 0.33. The correspondence in function of mapped clusters obtained from the conserved graph of the PINALOG alignment indicates that the resulting alignment shows the actual biological equivalence between the two input networks.

3.1.2 Alignment of the human PPIN with other species. We have also applied PINALOG to align the networks of human-fly, human-worm and human-mouse. The PPINs of fly, worm and mouse are relatively sparse compared with the yeast interactome. Therefore, the alignment between human and species other than yeast are expected to have far fewer conserved interactions. However, as these species are closer to human in terms of evolution, the resulting alignment should produce a smaller number of Homologenes. Taking into consideration these expected differences, the statistics for these comparisons (Supplementary Table SI-1) are broadly in keeping with the observations for human-yeast. For any comparison, PINALOG, IsoRank and MI-GRAAL identify a very similar number of aligned pairs (NA). In terms of conserved edges (NC), MA-GRAAL finds most, then PINALOG and finally IsoRank. In keeping with the use of functional similarity to establish the alignment, PINALOG finds more functional alignments (NF). PINALOG identifies more Homologene pairs than IsoRank and MI-GRAAL. In addition, PINALOG finds more interlogs that IsoRank and MI-GRAAL and for these comparisons the ratio N/NC is at least double for PINALOG compared with IsoRank.

3.1.3 Yeast2-fly2 alignment and Graemlin’s performance. In the alignment between yeast and fly interactomes, PINALOG and IsoRank perform comparably, apart from the PINALOG having more functional equivalences (NF). MI-GRAAL fails to find any Homologenes or interlogs. The values of NA and NC show that Graemlin’s alignment is far smaller than those from PINALOG, IsoRank and MI-GRAAL. Graemlin only finds 23 Homologenes compared with 241 from PINALOG and 2211 from IsoRank. PINALOG IsoRank and Graemlin find very similar number of interlogs (36, 34 and 36, respectively).

In summary, PINALOG generally performed better than IsoRank in most pairs of species in terms of conserved interactions, homologous pairs and interlogs. In the yeast-human alignment, although PINALOG finds more interlogs (NI), the fraction of conserved interactions which are interlogs is larger in IsoRank (see N/NC). Very similar results are obtained in the fly2-yeast2 equivalences for PINALOG and IsoRank. Comparison with Graemlin 2.0 shows that PINALOG and IsoRank generate far larger alignments with more Homologenes. Although MI-GRAAL uses the same source of data including sequence and functional similarity, it does not produce alignments with similar protein pairs in terms of either function or sequence. We explored the use of different parameters in MI-GRAAL to improve performance, but were unable to obtain improvements over the default values. Further analysis on the functional similarity of aligned protein pairs (see Supplementary Material and Supplementary Fig. SI-5) shows the advantage of utilizing the available functional information in the alignment,
Thus a high number of false positives (FPs) are expected. These complexes.

putative FPs in the PINALOG prediction may well be true protein single network-based methods. However, when inspecting the cell division chromatin modification, chromatin organization the 9 proteins of cluster 13 are enriched with terms related to M phase.

Another example is cluster 12 with 8 proteins that are related to M phase and another example is cluster 13 with 8 proteins that are related to M phase during cell division. The GOTermFinder analysis indicates that the putative FPs in the PINALOG prediction may well be true protein complexes.

BLAST performed as well as PINALOG and better than other single network-based methods. However, when inspecting the overlap in the predictions made by PINALOG and BLAST, we found that the overlap of the two predicted sets of clusters is not substantial, with only 50% of the predicted clusters matched. This suggests that PINALOG and BLAST could be used complementarily to predict protein complexes. This would provide a better coverage of the possible protein complexes. BLAST predicts clusters based on protein sequence, thus the prediction space is limited to sequence related proteins. In contrast, PINALOG predictions are made based on a combination of sequence, function and network topology, therefore the prediction space is extended to functionally related protein clusters that are reinforced by conserved interactions. We have combined the predictions made by PINALOG and BLAST into one predictor PINALOG+BLAST. This combination not only boosted the precision but also the recall rate, yielding a higher $F$-measure as compared with the individual methods (Fig. 2). We suggest that PINALOG and BLAST are used combination to help predicting protein complex more accurately.

5 BENCHMARKING OF PROTEIN FUNCTION PREDICTION

Another application of PINALOG is to predict the function of proteins by inheriting the annotation available of the aligned protein from the other species. PINALOG aims at providing function prediction (i) when there is no sequence homology; (ii) when there is a sequence homology between the un-annotated protein and the aligned annotated protein in the other species but the per cent sequence identity is low (typically <30%) to the closest annotated PSI-BLAST or to IsoRank.

A dataset of proteins to test the accuracy of function prediction was established with 415 GO annotated human proteins. These proteins had low per cent identity (<30%) to the closest annotated PSI-BLAST hit in the UniProtKB database. A 100-fold cross validation was performed. In each run of the cross validation, a set of proteins was selected from the test set and had their GO terms hidden. The test proteins which were aligned with an annotated protein in the other species
Table 2. Protein function prediction assessment by 100-fold cross validation, comparison with PSI-BLAST and IsoRank

<table>
<thead>
<tr>
<th></th>
<th>PINALOG-A</th>
<th>PSI-BLAST</th>
<th>IsoRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>Recall</td>
<td>0.14</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>0.28</td>
<td>0.43</td>
</tr>
<tr>
<td>MF</td>
<td>Recall</td>
<td>0.07</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>0.29</td>
<td>0.47</td>
</tr>
</tbody>
</table>

BP and MF stands for biological process and molecular function terms in GO.

were identified and the function transferred. Over the 100 runs, a total of 669 and 41 were transferred.
Functions of mapped proteins in yeast represent by GO terms in two categories BP (biological process) and MF (molecular function) were directly transferred to the human proteins in the test set.

The results were analyzed in terms of precision and recall, where precision is defined as tp/(tp+fp) and recall as tp/(tp+fn) being the number of true positives, false positives and false negatives of the predictions made. Our method is a binary predictor without a variable cut-off parameter and thus no precision/recall curve can be produced. Table 2 shows that PINALOG outperforms PSI-BLAST prediction with a superior recall rate at similar level of precision with BP category. The McNemar test (McNemar, 1947) for the statistical significance of the difference in the performance was used based on the number of misclassifications in each method. The test indicates that PINALOG predictions are significantly different from PSI-BLAST in BP terms at the level of $P=0.001$ significance level. PINALOG predicts more BP terms with a recall twice that of PSI-BLAST (14% versus 7%) at the same level of precision (~28%). For the MF category, PINALOG and PSI-BLAST share very similar levels of recall (~28%) while the precision of PSI-BLAST is slightly better (43% versus 47%). The McNemar test suggests no significant difference between them at the $P=0.001$ significance level. On the other hand, PINALOG outperforms IsoRank in both BP ($P=0.001$) and MF ($P=0.001$) categories. For example, in the BP category, PINALOG has almost twice the recall compared with IsoRank (14% versus 8%) at higher precision (28% versus 20%).

To summarize, in the challenging area where sequence which can be remote homologues not identifiable by PSI-BLAST. server Phyre2 (Kelley and Sternberg, 2009) which detects templates functions for six proteins, three of which have very general GO un-annotated proteins. Out of these nine proteins, ffPred provides et al.

terms. We compared our predictions to those from ffPred (Lobley et al., 2006) because the remaining five are mapped to proteins in yeast that have annotation only with very general GO terms. Although many methods have been developed to align PPINs, comparing their performance remains difficult due to the absence of an unambiguous correct alignment, the complexity of the networks and the differing aims of the alignment methods. For local alignment methods, the difficulty is less severe as some local alignments can be evaluated by their agreement with known protein complexes. However, this approach is less useful to evaluate global alignment methods. Often assessment is performed based on the correctness of mapping orthologous protein pairs. However, this assessment measure would always penalize equivalences based on network topology when this conflicts with equivalences between orthologous pairs. Therefore this assessment method is not ideal in evaluating the performance of network alignment algorithms. Accordingly, we have also used a series of other measures. PINALOG is flexible in allowing the use of sequence only alignment or sequence-function alignment depending on the requirement of the user. For the sequence-function alignment, parameters are automatically calculated from the input species and this will help non-experts. When aligning two species where one or both species are poorly annotated, alignment might be biased towards aligning well-annotated proteins whose functions are similar. Then it is advisable that sequence and network topology only are used to align the networks to avoid bias. We are developing a version of PINALOG to perform alignment of PPINs from multiple species. A web-server for PINALOG is available allowing users to upload necessary information and receive alignment results by email and a link to the results file (http://www.sbg.bio.ic.ac.uk/~pinalog/). The computing time of the alignment process depends on the size of the input networks. The typical computing time is ~24h with the longest run in our assessment being the human-yeast alignment which takes 24h on a computer with 2.8GHz processor with 8 GB memory (see Supplementary Material). Given the importance of the resulting alignment, this time is acceptable. The code for PINALOG is available from the website, Download section.

ACKNOWLEDGEMENT
We would like to thank Wellcome Trust for funding this project and Dr Mark Wass for valuable advice on function prediction.

Funding: Wellcome Trust

Conflict of Interest: MIIES is a director and shareholder in Equinox Pharma Ltd which markets software and services in the bioinformatics and chemoinformatics area.

REFERENCES


