NatalieQ: A web server for protein-protein interaction network querying

Mohammed El-Kebir1,2,3†*, Bernd W Brandt4†, Jaap Heringa2,3,5 and Gunnar W Klau1,2,3,5

Abstract

Background: Molecular interactions need to be taken into account to adequately model the complex behavior of biological systems. These interactions are captured by various types of biological networks, such as metabolic, gene-regulatory, signal transduction and protein-protein interaction networks. We recently developed NATALIE, which computes high-quality network alignments via advanced methods from combinatorial optimization.

Results: Here, we present NATALIEQ, a web server for topology-based alignment of a specified query protein-protein interaction network to a selected target network using the NATALIE algorithm. By incorporating similarity at both the sequence and the network level, we compute alignments that allow for the transfer of functional annotation as well as for the prediction of missing interactions. We illustrate the capabilities of NATALIEQ with a biological case study involving the Wnt signaling pathway.

Conclusions: We show that topology-based network alignment can produce results complementary to those obtained by using sequence similarity alone. We also demonstrate that NATALIEQ is able to predict putative interactions. The server is available at: http://www.ibi.vu.nl/programs/natalieq/.

Keywords: Network alignment, Protein-protein interaction, Sequence similarity, Topology, Wnt signaling pathway

Background

To adequately model complex behavior of biological systems one needs to take molecular interactions into account. These interactions are captured by various types of biological networks such as metabolic, gene-regulatory, signal transduction and protein-protein interaction (PPI) networks. Recent advances in technological developments and computational methods have resulted in large amounts of network data. For instance, STRING [1], a database of experimentally verified and computationally predicted protein interactions, grew from 261,033 proteins in 89 organisms in 2003 to 5,214,234 proteins in 1,133 organisms in January 2014. However, the development of solid methods for analyzing network data is lagging behind, particularly in the field of comparative network analysis. Here, one wants to detect commonalities between biological networks from different strains or species, or derived from different conditions. In contrast to traditional comparison at sequence level, topology-based comparison methods explicitly take interactions into account and are thus more suitable to compare networks. Subnetworks with shared interactions across species allow for improved transfer of functional annotations from one species to the other by using more information than sequence alone [2].

We have developed NATALIEQ, a web server for accurate topology-based protein-protein interaction network queries. It provides an interface to the general network alignment method NATALIE [3,4], which is fast and supports various scoring schemes taking both node-to-node correspondences and network topologies into account. Briefly, NATALIE views the network alignment problem as a generalization of the well-studied quadratic assignment problem and solves it using techniques from integer linear programming.

Currently, only few web servers for comparative network analysis exist. The PathBLAST web server [5] reports exact and approximate hits in a target PPI network.
for a user-defined simple query, expressed as a linear path of up to five proteins. The NetworkBLAST web server [6] finds locally-conserved protein complexes between species-specific PPI networks. NetAligner [7], a recent web server, allows the comparison of user-defined networks or whole interactomes within a set of fixed species using a heuristic network alignment with no guarantees on the optimality of the identified solutions.

Our contribution is twofold. First, NATALIEQ employs a new scoring function to produce high-quality pairwise alignments between a user-specified query network of arbitrary topology and interactomes of several model species and human. The score of an alignment is primarily based on the number of conserved interactions, while sequence similarity is used as a secondary, subordinate optimization goal. In addition, the alignments computed by the underlying NATALIE algorithm come with a quality guarantee that often proves their optimality. Second, through an interactive visualization of the alignment, the user can quickly get an overview of conserved and non-conserved interactions and can use the protein descriptions of the nodes to assess the alignment. We illustrate a usage scenario of the web server on the Wnt signaling pathway and demonstrate that NATALIEQ is able to predict putative interactions that are not detected by other methods.

Implementation

Network alignment algorithm

NATALIE, the alignment method of NATALIEQ, is applicable to any type of network and supports any additive score function taking both node-to-node correspondences and topology into account. Here, we take as input a pair of PPI networks whose nodes and edges correspond to proteins and their interactions. Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two PPI networks whose edges have a confidence value above a user-defined threshold $c_{\text{min}}$. We denote by $E(v_1, v_2)$ the $E$-value of proteins $v_1 \in V_1$ and $v_2 \in V_2$ obtained by an all-against-all sequence alignment. Typically, $G_1$ is a smaller query network such as a specific pathway of interest, and $G_2$ is a large species-specific PPI network.

A network alignment is a partial injective function $a : V_1 \rightarrow V_2$ with the additional requirement that if $v_1 \in V_1$ is aligned then $a(v_1) \in \{v_2 \in V_2 \mid E(v_1, v_2) \leq E_{\text{max}}\}$. That is, every node $v_1 \in V_1$ is related to at most one node $v_2 \in V_2$ with $E$-value $E(v_1, v_2)$ below a pre-specified cutoff $E_{\text{max}}$ and vice versa. We score the topology component of an alignment $a$ as follows

$$t(a) = \frac{1}{\min(|E_1|, |E_2|)} \sum_{v \in E_2} w(u, a(u), v, a(v))$$

with

$$w(u, a(u), v, a(v)) = \begin{cases} 1 & \text{if } (a(u), a(v)) \in E_2, \\ 0 & \text{otherwise.} \end{cases}$$

This score is also known as edge correctness and denotes the fraction of edges from the smaller query network that have been aligned. The problem of global pairwise network alignment is to find the highest-scoring alignment. Should there be several alignments with the same maximum edge correctness, we would prefer the alignment with the highest overall bit score as obtained by an all-against-all sequence alignment. We achieve this in the following way. Let $b(v_2, v_2) \in [0, 1]$ be the normalized bit score of aligning protein $v_1 \in V_1$ with protein $v_2 \in V_2$. The total score of an alignment $a$ is then

$$s(a) = t(a) + \frac{1}{1 + \min(|E_1|, |E_2|) \cdot \min(|V_1|, |V_2|)} \cdot \sum_{u \in V_1} b(u, a(u)).$$

That is, the score component is ensured to be strictly smaller than the score contribution of one conserved edge. Therefore ties among alignments with the same edge correctness are broken in favor of those with the highest overall bit score.

We use NATALIE to compute alignments with maximum total score. A specific feature of NATALIE is that any identified solution comes with an upper bound on the optimal score value. In the NATALIEQ setting with small query networks, the upper bound equals the score of the alignment found, thereby proving its optimality. The identified alignment is not necessarily optimal if there is a gap between the score and the upper bound. In that case the relative size of the gap provides a bound on the error due to suboptimality. In a recent study [4] on aligning PPI networks of six different species, NATALIE was compared to state-of-the-art network alignment methods, evaluating the number of conserved edges as well as functional coherence of the modules in terms of Gene Ontology annotation. The study established NATALIE as a top network alignment method with respect to both alignment quality and running time.

Databases

We currently provide eight model species from STRING [1] and IntAct [8] as target databases. We added textual descriptions to the protein IDs. For the STRING networks, these descriptions are available as a separate publicly available download. We retrieved the protein descriptions for the IntAct networks by cross-referencing the IntAct UniProt identifiers with the Swiss-Prot and
TrEMBL databases [9]. To allow NATALIEQ to take protein sequence information into account, we stored the amino acid sequences of the proteins in separate FASTA files per network. We retrieved these sequences from the STRING and IntAct databases. The target databases will be updated upon new releases of STRING and IntAct.

Processing
NATALIEQ computes a network alignment in a two-step fashion implemented in a Perl wrapper script. First, the wrapper invokes BLAST [10,11] to create pairwise protein alignments between the sequences corresponding to the nodes of the query and target network. Next, the wrapper invokes NATALIE [3,4] for different $E$-value cut-offs $E_{\text{max}} \in \{0, 10^{-100}, 10^{-50}, 10^{-10}, 1, 10, 100\}$. Each cut-off $E_{\text{max}}$ imposes restrictions on the allowed pairings, that is, only pairs $(u, a(u))$ with $u \in V_1$ whose $E$-value is at most $E_{\text{max}}$ are allowed. During these computations, which take a few minutes for a typical network query, the user is updated about the progress and may bookmark the unique web page for this run or leave an e-mail address to be notified upon completion.

Results and discussion
Web server
The input of NATALIEQ consists of a query network that can be in several formats: a simple edge list format, Cytoscape’s SIF format, IntAct’s MITAB format or STRING’s text-based format. The input file format is automatically detected. Optionally, the user can provide a FASTA file containing the protein sequences corresponding to the network nodes. In case no FASTA file is supplied and the node labels correspond to UniProt, RefSeq or GI identifiers, the corresponding sequences are retrieved automatically from the NCBI Protein database [12]. The user can select one of two well-known protein interaction databases (IntAct or STRING) and one of currently eight model species as target network. Options are the score function and the confidence threshold $c_{\text{min}}$. We support two score functions: topology, which is the scoring function as defined previously, as the default option, and sequence only, which results in the best network alignment in terms of sequence similarity, disregarding topological information.

The output page first gives an overview of the results for the different $E$-value cut-offs (Figure 1). The user can select a result for detailed inspection. Interesting results to inspect are, for example, the one with best sequence similarity among the top-scoring topological similarities or the one with best topological score at lowest $E$-value cut-off. The detailed view starts with summary statistics about the input networks and the computational process (Figure 2). It then displays an interactive network alignment visualization using the Javascript D3 library (http://mbostock.github.com/d3/), which is a data-driven framework for information visualization. The visualization (Figure 3) shows the aligned part of the two networks, overlaying nodes and links using red color for the query and grey for the target network. Thus, a matched query-target node or link pair will be colored in both red and grey. This interactive network visualization shows the user which parts of the query and target networks are matched. Hovering over nodes and links displays tool-tips with protein names and descriptions and link confidence, respectively, and allows for a quick overview of the alignment. If the user clicks on a node, information about that node is shown in a separate table, which in addition to the protein names and descriptions includes the bit score and $E$-value of the BLAST pairwise alignment and a hyperlink to the original database for more information about the target protein. The interface allows for a more detailed analysis by toggling the visibility of node labels, background target nodes and edges, unmatched query nodes and edges, and unmatched target edges.

In addition, the detailed view shows tables containing aligned query-target nodes, edges conserved in both query and target network, edges in the query network that remain unaligned, and unaligned edges in the target network whose incident nodes are aligned (Figure 4). The interactive visualization can be exported to a static SVG file and the user can download the alignment and the interaction tables for further off-line analysis. We support Cytoscape [13] by providing Cytoscape-compatible files containing the entire alignment and query network as well as matched parts of the target network.

Case study: Wnt signaling pathway
To illustrate the capabilities of NATALIEQ, we consider a biological case study involving the Wnt signaling pathway whose abnormal signaling has been associated with cancer. This pathway is initiated by binding of secreted Wnt
signaling proteins to the cell surface receptors Frizzled and LRP. This causes the activation of the signaling protein Dishevelled, which in turn inhibits the assembly of the degradation complex GSK-3β/axin/APC/β-catenin. As a result, the degradation of β-catenin is prevented causing it to accumulate in the nucleus. There, β-catenin forms a complex with LEF-1/TCF thereby displacing Groucho. The newly formed complex induces the transcription of various Wnt target genes, including c-myc which is a proto-oncogene encoding for a protein involved in cell growth and proliferation [14].

We manually constructed a PPI network of the pathway by using a subset of the proteins involved, namely WNT1, A2MR (LRP1), FZD1 (Frizzled-1), DVL1 (Dishevelled),
Interactions in query network but not in target network

| Query node | Target node | E-value | Target confidence |
|------------|-------------|---------|-------------------|
| WNT1       | A2MR        | 3.6e-92 | 0.887 TLE1 MYC    |
| WNT1       | FZD1        | 2.7e-36 | 0.854 DVL1 MYC    |
| FZD1       | CTNNB1      | 6.17e-01 | 0.639 CTNNB1 MYC |
| DVL1       | AXIN1       | 5.8e-01 | 0.346 FZD1 MYC    |
| AXIN1      | TCF7        | 2.3e-01 | 0.346 FZD1 MYC    |
| GSK3B      | APC         | 4.6e-01 | 0.928 CTNNB1 TLE1|
| APC        | CTNNB1      | 6.0e-01 | 0.928 CTNNB1 TLE1|
| CTNNB1     | TCF7        | 3.6e-01 | 0.928 CTNNB1 TLE1|

Interactions in target network but not in query network

| Target node | Query node | Target confidence |
|-------------|------------|-------------------|
| WNT1        | A2MR       | 0.887 TLE1 MYC    |
| WNT1        | FZD1       | 0.854 DVL1 MYC    |
| FZD1        | CTNNB1     | 0.639 CTNNB1 MYC |
| DVL1        | AXIN1      | 0.346 FZD1 MYC    |
| AXIN1       | TCF7       | 0.346 FZD1 MYC    |
| GSK3B       | APC        | 0.928 CTNNB1 TLE1|
| APC         | CTNNB1     | 0.928 CTNNB1 TLE1|
| CTNNB1      | TCF7       | 0.928 CTNNB1 TLE1|
| TCF7        | MYC        | 0.928 CTNNB1 TLE1|
| MYC         | TCF7       | 0.928 CTNNB1 TLE1|
| TCF7        | TLE1       | 0.928 CTNNB1 TLE1|
| TLE1        | MYC        | 0.928 CTNNB1 TLE1|
| MYC         | TCF7       | 0.928 CTNNB1 TLE1|

Figure 4 NATALIEQ alignment tables for the alignment of the Wnt query network against the target PPI network (STRING) of *D. melanogaster* using the sequence only score function at E-value cut-off 1. Blue entries are links to the STRING database.
AXIN1, GSK3B, CTNNB1 (β-catenin), APC, TCF7, TLE1 (Groucho), and MYC. For each of these proteins, we obtained their respective sequences from the STRING database. The edges we used correspond to the interactions described above. The query network consists of 11 nodes and 17 edges and is available as the example network file on the main page of NATALIEQ.

As a first sanity check, we queried against the human PPI network from STRING with link confidence threshold $c_{\text{min}} = 0.1$. For all $E$-value cut-offs, NATALIEQ found the optimal alignment where indeed all interactions are present and all query proteins are aligned with their identical counterparts in the human network as we could verify from the descriptions and interaction tables in the output.

For our next experiment, we used the PPI network of *D. melanogaster* as target. See also Figures 1–4 for an illustration. To study whether topological information improves comparative analysis, we compare the results of NATALIEQ using both the topology and sequence only score functions. We see that in the resulting sequence only alignments for $E$-value cut-offs larger than $10^{-10}$ one interaction of the query network is not mapped. This is the interaction between A2MR and FZD1. The counterpart of FZD1 in the sequence only alignment is FBpp0075485 with a bit score of 519 ($E$-value: $5 \cdot 10^{-177}$). The web server also provides the BLAST output, which shows that FZD1 is indeed sequence-wise most similar to FBpp0077788. NATALIEQ with the topology score function at $E$-value cut-offs larger than $10^{-10}$ is able to match all (17) query interactions and pairs FZD1 and FBpp0077788 with a bit score of only 150 ($E$-value: $6 \cdot 10^{-38}$). Although the bit score is less than the one obtained in the sequence-only alignment, the interaction A2MR–FZD1 is now present in the target network and has a normalized confidence of 0.172. So using NATALIEQ, we find that FZD1 may functionally be more related to FBpp0077788 than its sequence-wise most similar counterpart FBpp0075485. This hypothesis is corroborated by UniProtKB/SwissProt annotation indicating that the protein FBpp0077788 contains a Frizzled domain. Running the same example using the NetAligner web server [7] results in only 5 conserved interactions using default settings.

This example illustrates how NATALIEQ can facilitate the transfer of functional annotation across species. For instance, we could transfer functional annotation concerning the Wnt pathway between the human and fly networks by using the alignments we obtained.

**Conclusions**

We developed NATALIEQ, a web server for global pairwise network alignment of a pre-specified query PPI network to a selected target network. The underlying alignment method computes alignments with a worst-case bound on their quality. For the biological query networks we considered, the optimality gap was closed and provably optimal alignments with respect to the used score function were thus found. The user can quickly get an overview of the alignment through the interactive visualization, where conserved and non-conserved interactions are easily visible.

Currently, we support eight different target species from both STRING and IntAct. NATALIEQ is extendible, and we will add more target networks in the future. In addition, we plan to exploit the general applicability of the underlying NATALIE method by facilitating the identification of network motifs through more sophisticated query networks where nodes are labeled by GO terms and edges are labeled by different interaction types, such as inhibition and activation.

**Availability and requirements**

- **Project name:** NATALIEQ
- **Project home page:** http://www.ibi.vu.nl/programs/natalieq/
- **Operating system(s):** Platform independent
- **Programming language:** PHP and Perl
- **Other requirements:** modern web browser (Internet Explorer 9 or higher, Firefox, Chrome or Safari)
- **Any restrictions to use by non-academics:** no license required

**Competing interests**

The authors declare they have no competing interests.

**Authors’ contributions**

NATALIEQ was conceived by all authors. MEK, BWB and GWK designed and implemented the web interface and processed the data. All authors contributed to the writing of the manuscript and approved the final manuscript.

**Authors’ information**

Mohammed El-Kebir and Bernd W Brandt are joint first authors and both authors contributed equally.

**Acknowledgements**

We thank Sonja Boas for providing crucial insights on the Wnt signaling pathway case study. BWB was supported by the University of Amsterdam under the research priority area “Oral Infections and Inflammation”.

**Author details**

1. Life Sciences, Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, the Netherlands. 2. Centre for Integrative Bioinformatics VU (IBIVU), VU University Amsterdam, De Boelelaan 1081A, 1081 HV Amsterdam, the Netherlands. 3. Amsterdam Institute for Molecules, Medicines and Systems (AIMMS), Amsterdam, the Netherlands. 4. Department of Preventive Dentistry, Academic Centre for Dentistry Amsterdam (ACTA), University of Amsterdam and VU University Amsterdam, Amsterdam, the Netherlands. 5. Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, the Netherlands.

Received: 11 April 2013 Accepted: 20 March 2014

Published: 1 April 2014
References

1. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 2013, 41(D1):D888–D892.

2. Atlas N, Sharan R. Comparative analysis of protein networks: hard problems, practical solutions. Commun ACM 2012, 55(5):98–97.

3. Klau GW. A new graph-based method for pairwise global network alignment. BMC Bioinformatics 2009, 10:S9.

4. El-Kebir M, Heringa J, Klau GW. Lagrangian relaxation applied to sparse global network alignment. In Pattern Recognition in Bioinformatics (Proc. PRIB 2011), Volume 7036 of Lecture Notes in Computer Science: Edited by Loog M, et al. Berlin: Springer; 2011:225–236.

5. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res 2004, 32(W83–88).

6. Kalaev M, Smoot M, Ideker T. NetworkBLAST: comparative analysis of protein networks. Bioinformatics 2008, 24(4):594–596.

7. Pacheco RA, Ceol A, Aloy P. NetAligner—a network alignment server to compare complexes, pathways and whole interactomes. Nucleic Acids Res 2012, 40(W157–W161).

8. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedrazzi M, Pfeiffenberger E, Ponnas R, Raghunathan A, Rohebert R, Orchard S, Hermjakob H. The IntAct molecular interaction database in 2012. Nucleic Acids Res 2012, 40:D841–D846.

9. The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 2012, 40:D71–D75.

10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990, 215(3):403–410.

11. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389–3402.

12. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the national center for biotechnology information. Nucleic Acids Res 2007, 35(Database):D5-D12.

13. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011, 27(3):431–432.

14. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell. 4th edition. New York: Garland Science Taylor & Francis Group; 2002.

doi:10.1186/1752-0509-8-40
Cite this article as: El-Kebir et al. NatalieQ: A web server for protein-protein interaction network querying. BMC Systems Biology 2014 8:40.