

## FASPAD: fast signaling pathway detection

Falk Hüffner\*, Sebastian Wernicke and Thomas Zichner

Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

Received on March 8, 2007; revised on April 16, 2007; accepted on April 19, 2007

Advance Access publication April 26, 2007

Associate Editor: Anna Tramontano

### ABSTRACT

**Summary:** FASPAD is a user-friendly tool that detects candidates for linear signaling pathways in protein interaction networks based on an approach by Scott *et al.* (Journal of Computational Biology, 2006). Using recent algorithmic insights, it can solve the underlying NP-hard problem quite fast: for protein networks of typical size (several thousand nodes), pathway candidates of length up to 13 proteins can be found within seconds and with a 99.9% probability of optimality. FASPAD graphically displays all candidates that are found; for evaluation and comparison purposes, an overlay of several candidates and the surrounding network context can also be shown.

**Availability:** FASPAD is available as free software under the GPL license at <http://theinf1.informatik.uni-jena.de/faspad/> and runs under Linux and Windows.

**Contact:** hueffner@minet.uni-jena.de

### 1 INTRODUCTION

Protein interaction networks model the interaction of proteins in an organism. More precisely, they represent proteins as nodes and their mutual interactions as edges. Each edge can be labeled by a so-called ‘interaction reliability’ (Scott *et al.*, 2006), which can be interpreted as the probability that the respective proteins interact with each other in one or more signaling pathways, where a signaling pathway is a cascade of successive protein interactions that the cell uses to react to various external and internal stimuli.

Our tool focuses on the detection of *linear* pathways, i.e. paths where no node occurs more than once. These are easy to understand and analyze and can serve as a seed structure for experimental investigation of more complex mechanisms (Ideker *et al.*, 2001). There have been some efforts to design algorithms that automate the discovery of linear pathways in protein interaction networks (Scott *et al.*, 2006; Steffen *et al.*, 2002). The basic idea of these approaches is that a ‘good’ candidate path is a path that maximizes the product of its edge probabilities. While both Steffen *et al.* (2002) and Scott *et al.* (2006) demonstrated that this approach indeed yields biologically meaningful results, there are two problems that have remained so far:

- Detecting paths that maximize the product of edge probabilities is algorithmically very hard, i.e. an NP-hard problem. Thus, existing approaches require several hours for relevant path lengths of  $\sim 5$ –12 (Scott *et al.*, 2006).

- There is no user-friendly tool which allows for a quick and comfortable evaluation of pathway candidates that are generated by this approach.

The FASPAD tool addresses both of these issues: first, it relies on recent algorithmic insights by Hüffner *et al.* (2007) which drastically improve the running time over previous approaches so as to allow for an interactive exploration of pathway candidates. Second, FASPAD features a graphical display of candidates with various useful features, such as the ability to produce an overlay of several candidates or to show the surrounding network context of a candidate.

### 2 IMPLEMENTATION AND FEATURES

FASPAD is written in C++. The graphical user interface is based on the wxWidgets library, which makes it portable to several operating systems.

The workflow with FASPAD is divided into two main phases: first, a set of parameters must be specified that describes the paths the user is interested in. Second, following the actual calculation of pathway candidates, the user can graphically display them and examine their interactions with each other. The results can be saved both graphically and in text format. Both workflow phases are detailed in the following subsections, the interface is exemplified in Figure 1.

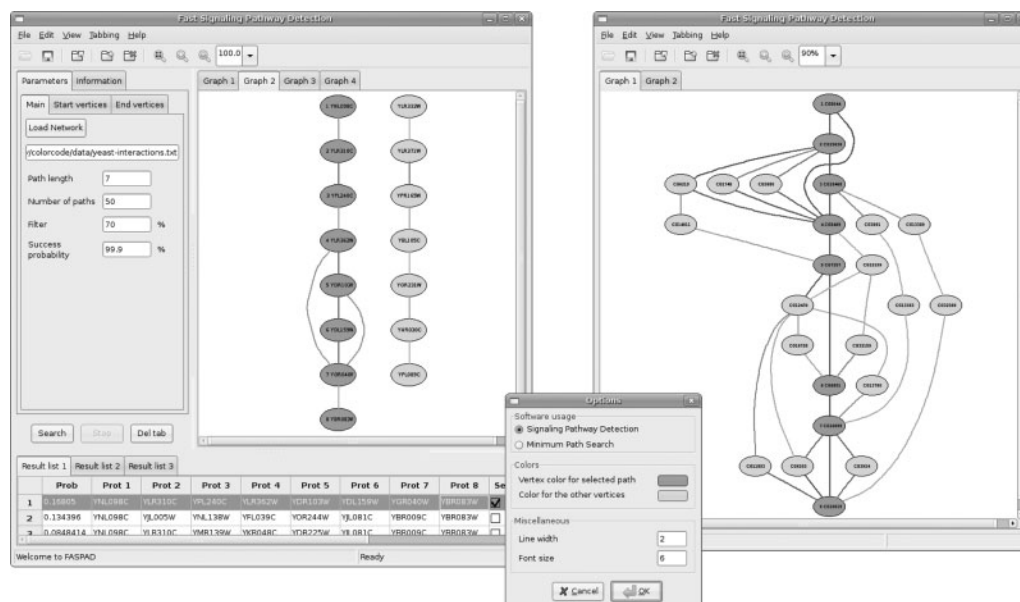
The use of FASPAD is not limited to protein interaction networks; optionally, it can search paths that minimize the sum of edge weights in arbitrary networks, at long as the path length requested is not too large.

#### 2.1 Search parameters

The user first specifies a file that contains the actual protein interaction network. The file format that is used by FASPAD is a simple text format, which facilitates the conversion of various data sources to it. Various parameters can be set for a candidate search:

- *Path length.* FASPAD supports path lengths up to 31 proteins. (Considering that signaling pathways quite rarely consist of more than 15 proteins, this should not pose any restrictions for practical use.)
- *Number of paths.* Typically, one is not only interested in a single ‘best’ path, but rather wants a set of ‘good’ paths to further examine them. The size of this set can be specified with a parameter.

\*To whom correspondence should be addressed.



**Fig. 1.** Two of the pathway candidates found by FASPAD in the yeast network by Scott *et al.* (2006) are displayed on the left: they correspond to the cell wall integrity pathway and the filamentous growth pathway known from literature. The right window shows a path in the *Drosophila* network (dark nodes) with its context.

- **Filter.** Looking at a number of paths with the best scores, one will typically find many small variations of the top scoring path. These are not interesting for manual examination. Therefore, it is possible to request that paths must not have more than a certain percentage of their proteins in common.
- **Success probability.** The search algorithm which FASPAD uses is randomized, i.e. the results are optimal only with a certain probability. This probability can be specified by the user in order to balance accuracy with running time. (Increasing the probability slows down the algorithm by a logarithmic factor.)
- **Start and end nodes.** Often, it is interesting to restrict the search to candidates that start and end with certain groups of proteins, e.g. candidates that start with a membrane protein and end with a transcription factor. Any such restriction can be specified in FASPAD.

With most parameter choices that are relevant in practice, the actual calculation of pathway candidates takes only a few seconds. The candidates that are found are displayed in table form.

## 2.2 Graphical display

In the most simple form, a single pathway candidate is displayed graphically. The user can examine the nodes and the edges and their probability. The nodes are arranged automatically; this arrangement can be tweaked in a drag and drop manner.

Often, one would like to examine whether there are small variations of a path that ‘make more sense’, even if their score is not as good, for example due to measuring errors. For this, it is possible to display also some surrounding context of the path. For example, one can display all proteins that interact with the displayed proteins with a probability higher than 70%, or display possible shortcuts within the path.

In order to compare several pathway candidates found by FASPAD, it is possible to select more than one path for display at the same time. This can be combined with the context display option.

To make it possible to generate different views on the same data set, or to compare results for different parameter settings or even different data sets, several ‘tabs’ can be used, each displaying a combination of paths with a certain layout. The result list can be saved and restored later.

Finally, the generated paths can be exported in a simple text format. The path display can be exported as PostScript vector graphics, as .png graphics file, or in the ‘dot’ format, which is a network layout description format used by graphviz.

## ACKNOWLEDGEMENTS

We thank Nadja Betzler (Jena) for testing the program. This work was supported by the Deutsche Telekom Stiftung and the Deutsche Forschungsgemeinschaft (DFG), projects PEAL (Parameterized Complexity and Exact Algorithms, NI 369/1), OPAL (Optimal Solutions for Hard Problems in Computational Biology, NI 369/2) and PIAF (Fixed-Parameter Algorithms, NI 369/4).

*Conflict of Interest:* none declared.

## REFERENCES

- Hüffner, F. *et al.* (2007) Algorithm engineering for color-coding to facilitate signaling pathway detection. In *Proceedings of the 5th Asia-Pacific Bioinformatics Conference (APBC 07) Advances in Bioinformatics and Computational Biology*. London: Imperial College Press. Vol. 5 pp. 277–286.
- Ideker, T. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.
- Scott, J. *et al.* (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J. Comput. Biol.*, **13**, 133–144.
- Steffen, M. *et al.* (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics*, **3**, 34.