FASPAD

FAST SIGNALLING PATHWAY DETECTION
- MANUAL -

Falk Hüffner, Sebastian Wernicke, and Thomas Zichner

Institut für Informatik, Friedrich-Schiller-Universität Jena,
Ernst-Abbe-Platz 2, D-07743 Jena, Germany
{hueffner,wernicke,tzi}@minet.uni-jena.de

April 16, 2007

Contents

1 Introduction

FASPAD is a tool to detect candidates for signaling pathways in protein interaction networks[1]. In general, a signaling pathway is a cascade of successive protein interactions that the cell uses to react to various external and internal stimuli. The algorithmic approach is explained in the paper

> Falk Hüffner, Sebastian Wernicke, and Thomas Zichner:
> Algorithm engineering for color-coding to facilitate signaling pathway detection.
> In Proceedings of the 5th Asia-Pacific Bioinformatics Conference (APBC'07), Hong Kong. January 2007.
> Volume 5 in Advances in Bioinformatics and Computational Biology, pages 277–286, Imperial College Press.

Protein interaction networks indicate how well proteins of a organism can interact with each other. These networks are modeled by edge labeled graphs, where the proteins are represented by vertices and the interactions by edges. The label of an edge is the probability that the respective proteins can interact.

The tool searches for the best paths of a given length in a network, that is, the paths where the product of the edge probabilities is maximal. A list of the best pathway candidates is shown and the user has the possibility to display and examine further single paths.

The complete functionality of the tool is described in the next sections.

[1]The tool can also be used to find the minimum-weight path in a graph, that is, the path of a given length which minimizes the sum over its edge weights. You can find further information about this in Section 11.

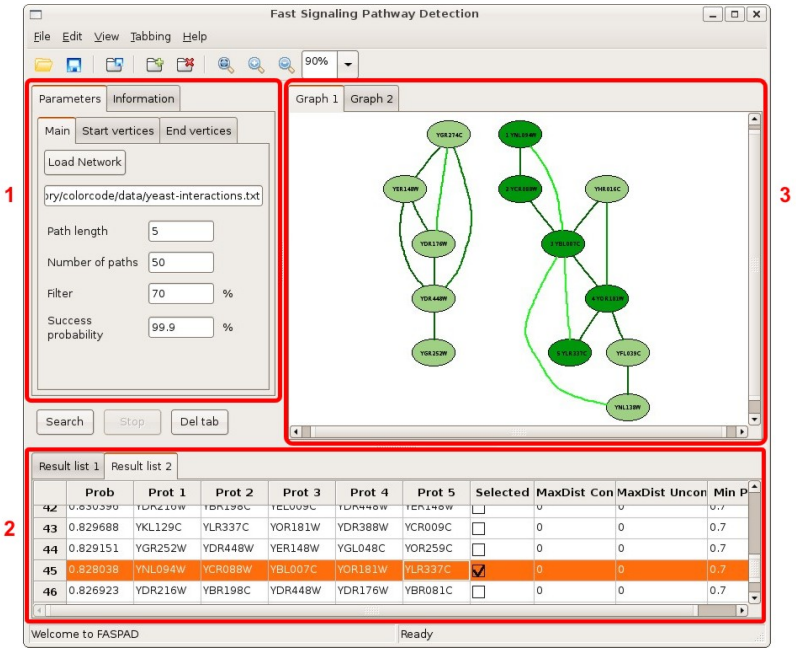

Figure 1: User interface of FASPAD

2 Overview

2.1 Graphical User Interface

Figure 1 shows the user interface of FASPAD, which is divided into three parts.

1 - Search Parameters. Here you enter all information concerning the search, that is, the network file as well as all other search parameter, for instance the pathway length. More information about this can be found in Sections 3 and 4.

2 - Result List. After performing the search (Section 5), the result paths are listed in this part of the user interface. For further information see Section 6.

3 - Graph Display. In this part, result pathways can be displayed graphically. It is also possible to interact with the path by clicking and Drag & Drop. This and other features concerning the graphical view are described in Section 7.

Figure 2: Tool bar of FASPAD

This tool contains also a menu and tool bar (Figure 2) to easily access the whole functionality as well as a status bar to display additional information.

2.2 Quick Start

If you do not want to wait to find your first signaling pathway candidates, you can find a short introducing example here. Our aim will be to find the 20 top-scoring signaling pathway candidates in the drosophila network that have the length of 9 proteins, start with *CG2072* or *CG8328*, and end with *CG9927*.
Here is a step-by-step instruction to fulfill the desired task.

- Obtain the protein interaction network of drosophila from the FASPAD web page (`http://theinf1.informatik.uni-jena.de/faspad/`).
- Run FASPAD.
- Load the interaction network (Sec. 3). For this, press the button `Load Network` in the tab `Parameters >> Main` on the left side and choose the file *drosophila-interactions.txt* in the standard dialog.
  Now you have to specify in which columns of the file the required information is stored. Select for *protein 1* column 1, for *protein 2* column 2 and for the interaction probabilities column 3. None of the first lines of the file have to be ignored and therefore this value should be set to 0.
- After the network is loaded, we can define the main parameters of the search (Sec. 4). According to our task, set `Path length` to 9 and `Number of paths` to 20. The parameters `Filter` and `Success probability` can stay unchanged.
- Now we want to define the start and end points of the search. We begin with the two start proteins. Go to the tab called `Start vertices`, write *CG2072* in the small text box and press `Add`. Then write *CG8328* and press `Add` again. If you want to remove a vertex, select it in the list box and press `Remove`. Now you can add the end point of the search (i.e. the protein *CG9927*) in the same way to the tab `End vertices`.
- After setting all the parameters, just press the `Search` button. You can follow the search process in the status bar (Sec. 5).
- When the search is finished, the result list is displayed on the bottom of the window. To display a single path, just click on the corresponding row in the table (Sec. 6).
- You can also display several paths together by using the check boxes in the select-column (Sec. 6.2). To display the path environment is a little bit more tricky and described in Section 6.3.
- By clicking on the vertices and edges you can get more information like the interaction probabilities (Sec. 7.3). This data is displayed in the `Information`-tab. You have also the possibility to rearrange the displayed graph in a typical drag&drop manner Sec. 7.2).
- If you want to save the result list or a displayed graph, use the menu entities `File->Save Result List as...` respectively `File->Save Result Graph as...` (Sec. 9).

Congratulation, you found your first signalling pathway candidates! But this was just a small example to give you an idea of FASPAD. Read the following sections to get to know the whole functionality of this tool.

3 Loading of a Network

Before you can start searching for signaling pathways, you have to load the protein interaction network you want to search in. For this, just click on the button called `Load Network` which is placed in the tab `Parameters >> Main` in the left part of the user interface. A standard file-open-dialog is shown where you can choose the network. The file has to contain all the interactions between the proteins together with their probabilities. In each line one interaction has to be stored and its properties have to be delimited by a tab or space character.

After selecting a file, an additional dialog opens where you can specify in which columns of the file the required data (identifier of protein 1, identifier of protein 2 and interaction probability) is stored.

The proteins can be denoted by a string, but they must not contain spaces. The interaction probabilities have to be floating-point numbers between 0 and 1.

You can also define whether the first lines of the file shall be ignored. This is, for instance, useful if the first line contains the column descriptors. Because of these adjustable parameters, FASPAD is able to handle almost all tab-delimited network files.

If the file format is correct, the file path and name of the network is shown in the text box below the button; otherwise, an error message box occurs.

4 Search Parameters

The search of signaling pathway candidates can be influenced by different parameters.

4.1 Main parameters

The main parameters can be found in the tab `Parameters >> Main`.

Path length. This specifies the number of proteins in the searched pathway. The longer the path is, the more time is needed to compute the results. FASPAD is able to search for paths up to the length of 31 proteins. But be careful with this parameter, because already a search for paths containing 15 to 20 proteins can take hours depending on the network.

Number of paths. This indicates the length of the result list. The n-best paths that fulfill all parameters are computed and shown. The less paths have to be computed the less time is needed.

Filter. To avoid that all found paths belong to only a few main pathways with small changes, you have the possibility to set a filter value. Paths with have more than this percent of proteins in common are filtered out. The higher the value, the faster the algorithm.

Success probability. The search algorithm used in this tool is a so-called "randomized algorithm". That is the results are correct only with a certain probability. This probability can be set here, a common value is 99.9 per cent. A higher value leads to a longer running time.

4.2 Start and End Vertices

There is the possibility to limit the proteins where a pathway is allowed to start and/or to end. This is for instance useful if want to search for pathways that start at a membrane protein and

end at a protein in the nucleus. For this you have the tabs `Start vertices` and `End vertices` under the tab `Parameters`. At the top of the tabs, you can choose whether you want all proteins or only a selection as possible start respectively end points. If you choose `Some`, only the proteins in the list below are considered. You can add a protein to the list by writing the name in the text box and clicking `Add` or pressing `Enter`. You can also load files which contain the names of proteins (one name per line). To do so just click on the button `File` and choose the file in the standard file-open-dialog. To delete proteins from the list just select them with the mouse and click on the button `Remove`.

5 The actual Search

5.1 Starting the Search

After loading a network and specifying all parameters, you can start the actual search process. To do so just click on the button `Search` which is situated below the `Parameters`-tab. If this button is disabled, no network is loaded.

5.2 During the Search

During the search you can find information about the progress in the status bar. Each search goes through the following three steps:

Bounds computation. This computation is needed to speed up the actual search. The running time of this computation is only dependent on the size of the network and not on path length or other parameters.

Pre-heating. Here the search is performed on a thinned out network. This results also in a faster search process. There are always 10 pre-heating steps.

Final search. This means that the actual search is running. The search is divided into several trials. How many trials need to be performed depends on the path length you are looking for and the desired success probability. Longer paths and a higher success probability leads to more trials. When the search is finished, "Search completed" is shown in the status bar. If no paths that fulfill all parameters could be found, "No paths found" is displayed.

5.3 Pausing or Aborting the Search

During the search, there are two important buttons. The former `Search` button can now be used to `Pause` the search. The second one is the `Stop` button which can be used to abort the search. In this case, all relevant paths which FASPAD already found up to this state of the search are shown in a result list. Of course the probability that these paths are the best in the network is less than the specified `success probability`.

6 Viewing the results

If the search was successful, a new tab with the results is displayed in the bottom part of the user interface. The results are arranged as a table. Each row represents one possible signaling pathway. The first column shows the probability of the path, that is, the product of the interaction

probabilities. The next n columns contain the protein names, where n is the length of the pathway. The four remaining columns are editable and used to control the display of the results. This will be described in more detail in the following subsections.

6.1 Single Paths

To view graphically all proteins of a single path as well as the interactions between them, click on the corresponding row in the table. The path will be displayed in the currently selected graph tab (you can find more information about tabbing in Section 8). The proteins are represented by ellipses which contain the name. Interactions are represented by splines. Their color indicates the probability; in the tab `Information` you can find a legend. Under texttttInformation, you can also get further data on a vertex or an edge by clicking on it.

6.2 Merging of Paths

There is the possibility to display two or more pathways simultaneously. This means that all proteins which are contained in at least one of the selected paths are displayed together with the interactions between them. Here each proteins occurs only once independent of the number of paths it is contained in.

To select several paths, tick the corresponding boxes in the "selected" column (you might have to click twice, once for selecting the cell, and once for selecting the checkbox). In addition to the ticked paths, the path in the currently selected row is displayed. The proteins of the latter are highlighted in a different color, and they are numbered according to the position in the path.

The selected paths do not have to belong to the same result list, as long the corresponding network is the same.

6.3 Path Environment

FASPAD can also display the adjacent proteins of a path. This is useful if you want to examine a signaling pathway closer, maybe you are looking for alternative or interchangeable proteins.

Theory. There are three parameters which control this functionality.

- **Maximal Distance Connected (maxDistCon)**
- **Maximal Distance Un-Connected (maxDistUnc)**
- **Minimal Probability (minProb)**

In addition to the proteins belonging to the selected path, all proteins of the network which fulfill at least one of the following two conditions are displayed.

- There are at least **two distinct paths** of length less or equal to *maxDistCon* which connect the protein with the pathway. Each edge of this two paths must have a probability higher than *minProb*.
- There is at least **one path** of length less or equal to *maxDistUnc* which connects the protein with the pathway. Each edge of this path must have a probability higher than *minProb*.

Application. Here are two application scenarios for which this functionality is helpful.

- *Task:* For all proteins of a path you want to display all interaction partners where the interaction probability is higher than *0.7*.
  Solution: Set *maxDistCon* to 0, *maxDistUnc* to 1, and *minProb* to 0.7.

- *Task:* You want to display all alternative protein interaction ways of length less or equal to three. That is, all paths that start and end in the main pathway and have a maximum length of three.
  Solution: Set *maxDistCon* to 3 and *maxDistUnc* to 0. The value for *minProb* depends on whether you want to display all paths or only those that use edges of a certain interaction probability.

Additional hints.

- Especially in networks of several thousand proteins, the path environment can be very large. Therefore you should choose the values *maxDistCon*, *maxDistUnc*, and *minProb* very carefully. Start with small distances (e.g. 1) and a high probability (e.g. 0.8) and change the values slowly if necessary.

- You can also "merge" path environments and not only single paths (see section 6.2 for merging of paths).

7 Working with Result Graphs

7.1 Zooming

To view each result graph in appropriate size, you can use the zoom functionality of FASPAD, either by the entries in the `View`-menu or by the buttons and the drop-down list in the task bar (figure 3: F-I).

Zoom in. The graph is displayed larger by setting the zoom to the next larger zoom factor.

Zoom out. The graph is displayed smaller by setting the zoom to the next smaller zoom factor.

Zoom fit. The zoom factor is adapted so that the whole graph fits into the current graph tab.

Drop-down list. You can also set a zoom factor directly by using this drop-down list.

7.2 Moving of Vertices

You have the possibility to rearrange the graph layout by moving single vertices. For this click with the left mouse button on a vertex and hold the mouse pressed while moving. If you reached the desired position, just release the mouse button. The edges connected to the vertex are moved accordingly.

7.3 Information about Proteins and Interactions

To get further information, click on a vertex or an edge with the left mouse button. The information is now shown in the status bar as well as the `Information`-tab.

Figure 3: Tool bar of FASPAD: A: Open result list; B: Save result list; C: Save result graph; D: Add new graph tab; E: Delete current graph tab; F: Zoom fit; G: Zoom in; H: Zoom out; I: Zoom drop-down list

Status Bar. If you click on a vertex the full name of it is shown. This is especially useful for long names which do not fit into the ellipses. If you click on an edge, the names of the corresponding vertices as well as the weight of the edge, that is, the interaction probability, are shown.

Information Tab. Additionally to the information which is also shown in the status bar, in the information tab you can find the vertex degree if you click on a vertex. That is the number of vertices which are connected by an edge to the selected one (in terms of signaling pathway detection, the number of proteins which can interact with the selected one).

8 Tabbing

To handle more than one result list or result graph FASPAD provides the functionality of tabbing (know from applications like *Mozilla Firefox*). Therefore, it is easy to work simultaneously with several result lists or graphs without the need of running different instantiations of the tool.

8.1 Result Graph Tabs

If a search is completed or you load an already existing result list from a file, a new result list tab is opened listing the detected paths. All previous tabs stay open, so you can still work with them. If you do not need a tab anymore, you can close it: first select it by clicking and second push the button `Del Tab` which is placed next to the buttons `Search` and `Stop`.

8.2 Result List Tabs

If a new result list is created by loading or a search, automatically a new graph tab is generated which displays the first (best) path. To display a desired path, just click on it and it will be shown in the currently selected graph tab (see Section 6). You can create and delete graph tabs at will by using the buttons in the tool bar (Figure 3: D-E).

9 Saving and Loading of Results

9.1 Storing of Result Lists

To save the current selected result list you have to go to the menu `File->Save Result List as...` or to click on the corresponding button in the tool bar (Figure 3: B). FASPAD offers two different file formats.

Character Separated Values (*.csv) Here a text file is created where each line contains one path. The first value in each line is the probability followed by the names of the proteins. All values are separated by a semicolon. This format is useful if you want to import your results into other applications like Microsoft Excel.

Internal format (*.rsl) This is an application-specific format in which the whole result list is stored, including all display values such as "selected" and "max dist". Also the file name and path of the network is stored. FASPAD is able to load these files (menu `File->Load Result List` or corresponding button in the tool bar (Figure 3: A)) so that you can continue your work later.

The standard file format for result lists is the internal format. To change this you have to click on `Browse for other folders` in the file save dialog and then you can choose the desired format in the lower right corner.

9.2 Storing of Result Graphs

To save the current selected result graph, you have to go to the menu `File->Save Result Graph as...` or click on the corresponding button in the tool bar (Figure 3: C). FASPAD offers three formats to store a displayed result graph.

Postscript (*.ps) Here the graph is stored as a vector graphic. This format is very useful if you want to include the picture into another document or if you want to print it. The output file is a postscript page. The size of the graph on this page depends on the current zoom value, but if necessary the graph is resized to fit on the page.

Portable Network Graphics (*.png) Here the graph is stored as a pixel graphic. Png is a common file format where the image is stored with lossless compression. It can be read by the most graphics programs.

The DOT Language (*.dot) Here the graph is stored in a descriptive language. That means it is stored which vertices and edges the graph contains as well as layout information like the colors. This format which is human readable is used mainly by the application *graphviz*.

The standard file format for result graphs is the PostScript format. To change this you have to click on `Browse for other folders` in the file save dialog and then you can choose the desired format in the lower right corner.

10 Option Panel

Some parameters of FASPAD can be changed to adapt the tool to the personal usage. You can open the option panel through the menu `Edit->Options...` (a screen shot is shown in Figure 4). All options are stored (under Unix/Linux in the file *.FASPAD* in the home directory and under Windows in the registry) so they remain after closing the program. The following parameters can be adjusted:

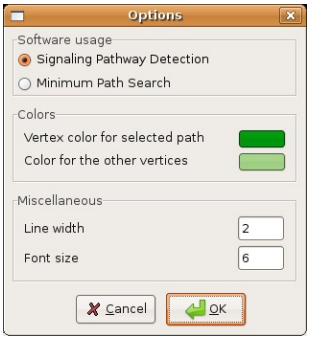

Figure 4: Option panel

Software usage. The tool is designed to detect signaling pathways in protein interaction networks. But it can be also used to find Minimum-Weight Paths in weighted graphs. This option is to select the field of application you want to use the tool for. The changes in the program which are related to it are described in section 11.

Vertex color for selected path. This color is used for all proteins of a selected pathway. You can change the color by clicking on it.

Color for the other vertices. This color is used for all proteins of the path environment (Section 6.3) as well as the not selected paths when displaying two or more.

Line width. The width of the lines which represent the interactions is constant and independent of the current zoom value. The width can be set here.

Font size. Here you can specify the font size which is used for the zoom value *100%*. The font size for other zoom values depends on it.

11 Using FASPAD to find Minimum-Weight Paths

You can use this tool not only to detect signaling pathways in protein interaction networks, but also to find minimum-weight paths, that is paths where **the sum** over all edges is minimized, in usual weighted graphs. For this you have to switch the program usage in the option panel `View->Options...` (for further information see Section 10).

The handling of FASPAD stays the same, but the following things are changed by choosing this option:

- The button to load a graph is called *Load Graph* instead of *Load Network*.
- The graph file has to contain edge weights instead of probabilities. So in a file, all edges have to be stored, one in each line. Like for network files (Sec. 3), you can specify in which columns the required data (identifier of vertex 1, identifier of vertex 2 and the edge weight) is stored.
  The vertices can be denoted by a string, but they must not contain spaces. The edge weights have to be floating-point numbers.

- In the first column of the result list the sum over all edge weights of the corresponding path is displayed instead of the product.
- In the last column of the result list you can enter the parameter max weight. When calculating the path environment (see Section 6.3), only edges with a weight less than this value are considered.
- When clicking on an edge the weight is displayed in the status bar as well as the information tab instead of a probability.

Acknowledgments

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).

The graphical user interface was built with the wxWidgets framework 2.8 written by Julian Smart and many others. We also used the GraphViz library version 2.8 by AT&T Research Labs.