



Functional Analysis by Association

FABA 1.6 User Manual

Copyright © 2004 by Chang Bioscience, Inc.
All rights reserved.

Table of Contents

Introduction	4
Compare FABAs with Yeast Two-hybrid	4
What Can FABAs Be Used For?	5
FABA 1.6 Installation	6
How Does FABAs Work?	8
A. Correlation score for individual dataset	8
B. Correlation score for multiple datasets	8
C. Gene expression data normalization	9
D. Missing data	9
E. Penalty for stable and low-signal genes	9
F. Noise levels	10
Power of FABAs	11
Realistic expectations	14
Data types	14
FABA Searches for Gene-Gene Correlations	15
How to Use FABAs Viewer	16
A. View Data	16
B. View history	16
C. Batch query	17
D. Reduce noises	18
E. Refine Your Query	19
F. Import Data	20
How to Use FABAs	21
A. Data Selection	21
B. Data Preparation	21
1. Preparation of info file	21
2. Preparation of data	21
(a) High-throughput data format	21
(b) Classification data format	21
(c) Correlation data format	21
C. Data upload	22
D. Project export	22
E. Project imports	22
F. Delete dataset and project	23
Think Again	23
Advanced Users	24
A. Data directory/file structure	24
B. Backup	24

C. Delete project and dataset	24
System Requirements.....	26
Public data resources	27
A. Gene expression data	27
B. Protein-protein interaction	27
C. Protein localization	27
D. Protein domains	27
E. Gene Ontology.....	27
Related Products	28
A. BioToolKit 300	28
B. The Electronic Protocol Book	28
Frequently asked questions	29
1. Does FABA have any limitations on the types of high-throughput data it can analyze?	29
2. Do I have to use Unigene ID as identification? Can I use my internal database ID?.....	29
3. Do you have data for any other species?.....	29
4. How does FABA help me? I don't have any HTP data.	29
5. How do you determine the noise level?.....	29
6. Why are some genes noisy?	29
7. Which noise level should I use?.....	30
8. FABA results seem to change for different versions of FABA. Why?	30
9. Since there are so many false correlations in each individual dataset, will FABA predictions be useful at all?	30
10. FABA ignores differences in array types and blindly compares microarray apples and oranges. How can such an approach produce any meaningful results?	30
11. Are you absolutely sure that array data from different platforms can be compared?.....	30
12. Why are you using only simple normalization options for microarray data?	30
License Requirements	32
Customer Support.....	32

Introduction

FABA stands for Functional Analysis By Association. FABA is designed to discover gene function by analyzing high-throughput data from different sources. Instead of analyzing individual data set, which frequently concentrates on one process and may be biased because of system specific experimental noises, FABA integrates data from different sources thus eliminating many false-positive results.

One of the most important sources of high-throughput data is the microarray gene expression data. Expression data from tens of thousands of microarray experiments are now publicly available to every researcher. And the amount of expression data is accumulating at a growing rate. FABA is also designed to analyze this humongous amount of expression data. Instead of mining data one set at a time, FABA empowers scientists with a tool to analyze tens sets of data and thousands of microarrays.

Compare FABA with Yeast Two-hybrid

	Yeast Two-Hybrid	FABA
LAB	Wet-Lab	Dry-Lab
Probe	Bait protein	Query gene
Interaction	Protein-Protein	Gene expression correlation + others
No, of baits	A few	Many
Positives	May have similar functions	May have similar functions
Limitations	Strong bait-target interactions False positives	Relevant data (e.g., gene expression) False positives
Cost	\$\$\$\$\$	\$\$\$ Can use Public data. Inexpensive
Time	Months	Days

In a sense, FABA is an *in silico* two-hybrid technology. The yeast two-hybrid technology uses a bait protein to fish for target proteins that physically interact with the bait. FABA on the other hand find genes that share similar properties (gene expression, protein localization, sequence similarities, etc.) of the query gene. Both find candidate genes that might share similar functions of the bait/query gene. Yeast two-hybrid can only detect proteins that have strong physical contacts. FABA can detect much broad gene-gene interactions, but relevant data must exist. For example, if we want to find out downstream genes of Her2 receptor in ovarian cancers, the FABA data must contain significant amount of data relevant to ovarian cancers or at least to other cancer types such as breast.

Both yeast two-hybrid and FABA will produce a large number of false positives. FABA however has one advantage: it can use a growing volume of public data (e.g., NCBI GEO database has archived >17,000 gene expression samples at the end of April, 2004). As a result FABA is a less expensive method for finding candidate genes with similar functions.

What Can FABA Be Used For?

- *Find genes that show similar expression patterns across thousands of microarray experiments.*
- *Find the significance of gene expression correlation by using gene expression in thousands of experiments as a reference.*
- *Find potential gene functions based on that genes with similar functions frequently show similar expression patterns.*
- *Find novel candidates of a signaling pathway.*
- *Analyze many types of high-throughput genome data.*
- *Incorporate gene ontology data into the analysis of high-throughput data.*
- *Verify results from other experiments such as yeast two-hybrid and genetic screening to find better leads.*

FABA 1.6 Installation

1. Check system requirements. Please make sure your computer has a minimum of 64 Mb of memory and Gbs of free hard drive space. FABA 1.6 demo can run on system with less memory and hard drive space, but if you intent to use it frequently then installing on the best computer in your lab will save you time in the future.

2. Download FABA 1.6 from www.ChangBioscience.com. Please download the install file appropriate for your computer platform. Installers are available for Windows/NT and Mac OS X. Please contact us if you have other systems such as Linus and Unix.

3. For Mac OS X, uncompress the installer file by double clicking on the install.zip icon. A new install icon will appear after uncompressing the file by Stuffit.

4. Start installation by double click the install icon. Mac OS X users, please make sure your system is not in the Classical Environment.

5. Follow the instructions of the install program to install the program.

6. If the installer can not start, please check whether you have completed the download. Check the installer's file size. You may need to download again or request a CD from Chang Bioscience.

7. Windows/NT users may experience an additional problem. The installer needs JAVA to start. Most computers already have JAVA JRE installed. Occasionally a few computers have not installed JAVA or need an upgrade. Please visit the following site to install JAVA JRE first:

<http://java.com/en/index.jsp>

8. After successful installation of FABA, users may start the program and view pre-run results.

9. Start FABA. Select a project and click on "Go". A new window will appear.

10. Select from the history pull down menu (right to the "Go" button) a previous result. The results will appear in a new window.

11. The demo version does not include data necessary for running your own query. Please request a free data CD from Chang Bioscience by emailing to info@changbioscience.com. The data CD contains expression data of more than 3,000 microarrays for human, mouse, and yeast.

12. After receiving the data CD, follow the instructions to import data (see also page 21).

13. You'll then be able to run your own query and add your own or new public data. You don't need the data CD if you would like to start a new project. Detailed instructions are listed elsewhere in this manual.

14. Please purchase a license before the demo license runs out so your work will not be interrupted.

15. We hope you will enjoy the FABA software, please don't hesitate to ask us if you have any questions. Although priorities will be giving to answer questions of licensed users, we'll attempt to answer all questions as soon as we can. Any suggestions and critics will be greatly appreciated.

16. Please join the FABA discussion forum to share your experiences with others. To join, please visit:

<http://www.changbioscience.com/forum/phpBB2/>.

How Does FABa Work?

To predict a gene's function, scientists analyze information from different sources. FABa is based on the same simple principle but doing so quantitatively. For each gene pair it calculates a correlation score in individual dataset. Such individual datasets could be microarray data (e.g., microarray data under a variety of stress conditions), protein localization, or sequence similarities. The correlation scores for these datasets will be similarity in transcription profiles, co-localization, and sequence similarities, respectively. Finally, a correlation score for each pair is calculated based on the scores for all the selected datasets. Genes with high correlation scores are more likely having similar functions because their shared characteristics.

A. Correlation score for individual dataset

(a) High-throughput data

The most frequent high-throughput data is the gene expression data. The data is a serial of measurements on different samples, and should be tab-delimited and in the following form:

Gene ID Data1 Data2 Data3 Data4 ...

The first column is unique gene identifications (e.g. Unigene number) for individual genes. Data in each column represent a separate measurement. The correlation between two genes is calculated as the Pearson correlation:

$$r_{ij} = \langle X_i, X_j \rangle.$$

Here X_i is the i -th row in the data matrix.

The correlation score is defined as

$$e = \begin{cases} \text{sign}(r) * (|r| - |r_0|) / (1 - |r_0|), & \text{if } |r| > |r_0| \\ 0, & \text{if } |r| = |r_0|. \end{cases}$$

here r_0 denotes the 95% confidence interval for the Pearson correlation r .

(b) Classification data

For classification data, the correlation score is defined as 1 for the same classes and 0 for different classes.

(c) Correlation data

The correlation coefficient r will be directly used as the correlation score e . Users are suggested to correct for the confidence intervals before uploading the data.

B. Correlation score for multiple datasets

The correlation score for multiple datasets are the sum of correlation scores for individual datasets:

$$E = \sum w_j e_j,$$

here w_j ($0 \leq w_j \leq 1$) is the weight parameter for dataset j .

C. Gene expression data normalization

Normalization is done for each dataset. Users may choose their preferred normalization method. For data uploaded by Chang Bioscience, all datasets are normalized such that the means for each column are identical within individual dataset.

D. Missing data

Missing data are ignored. Users may impute missing data before uploading. For data uploaded by Chang Bioscience, no missing data imputation is done.

E. Penalty for stable and low-signal genes

To minimize detecting gene-gene correlation for the large number of quasi-stable genes, the gene-gene correlation efficient is modified to penalize stable and low-signal genes.

$$r \rightarrow r * R / (1 + R),$$

here R is a parameter that is small if both genes are stable or have low expression values. R is much greater than 1 if either gene shows sufficient variation across samples. The R value will be 0 if both genes are constants.

F. Noise levels

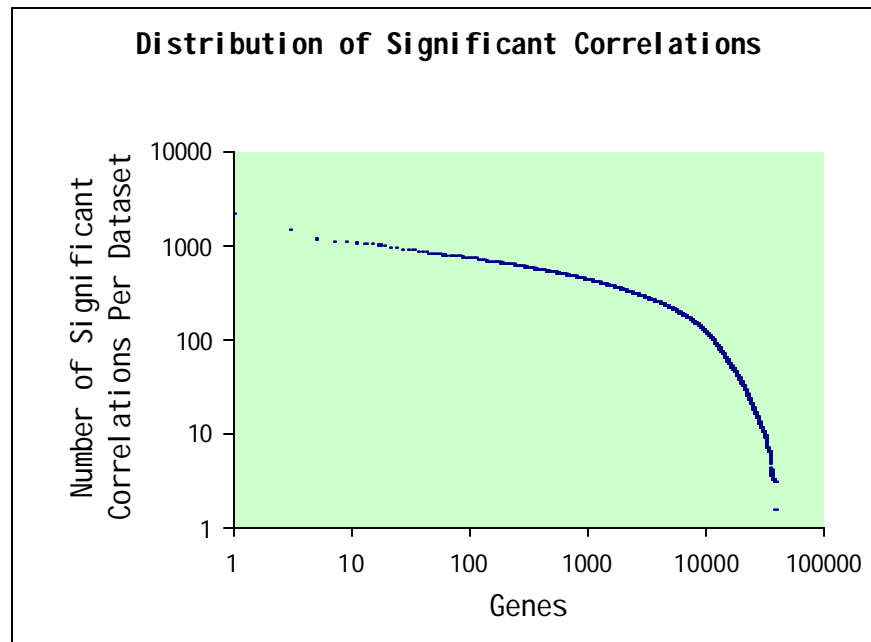


Figure 1. Statistics of significant correlations in gene expression data.

As shown in Figure 1, a small fraction of genes have a large number of significant correlations to other genes. These correlations are less informative because they may not represent biologically significant co-regulations. Reasons for such correlations include: (1) quasistable expressions, (2) array-system specific noises, (3) fluctuation with environmental conditions such as nutrients and stress, (4) hybridization noises, (5) improper normalization, and (6) probes contain over-represented sequence elements, just to name a few. The “noise” level for each gene is determined based on the following assumption: more significant correlations for a gene, less informative these correlations are, thus more “noisy” for the gene.

Based on the number of significant correlations a gene has, it is assigned one of twenty noise levels. Users can choose a noise level to remove all the genes with higher noise levels from final results.

Power of FABA

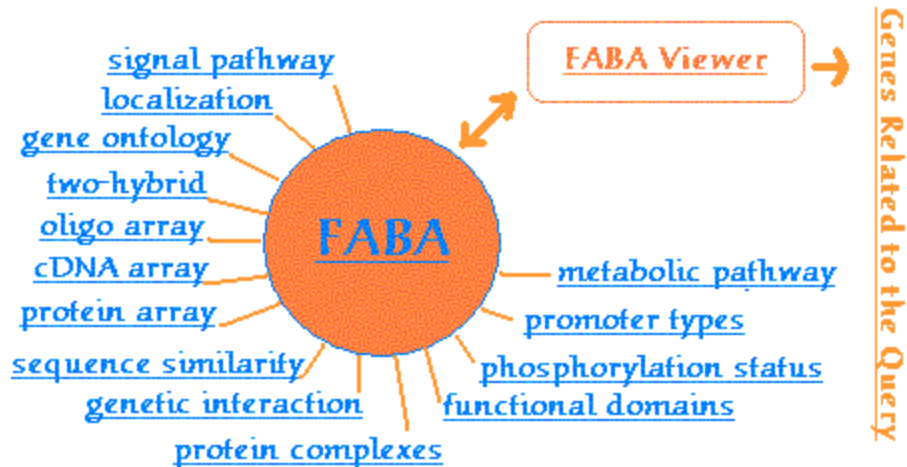


Figure 2. FABA can be used to analyze many types of high throughput data.

Shown in Figure 3 is a FABA example for *Saccharomyces Cerevisiae*. It includes SGD Gene Ontology data, yeast two-hybrid data, protein localization data, and 45 sets of microarray data (~900 arrays). The transcription profile of an uncharacterized ORF (YKL056C) is similar to those of ribosome proteins and DNA/RNA binding proteins, suggesting a role of this uncharacterized ORF in protein synthesis. Note the heatmap displays the correlation score between 2 genes in each set of experiments. The name of the query gene is shown on the top. The other gene member is shown on the right. Each column of the heatmap represents a set of experiments. Click on a column will show the data source. For example the dataset "Fermentation" is a collection of 12 microarray data published by Olesen et al. (FEMS Yeast Res. 2002 Dec; 2(4): 563-73). Red color represents a positive correlation and green color represents a negative correlation.

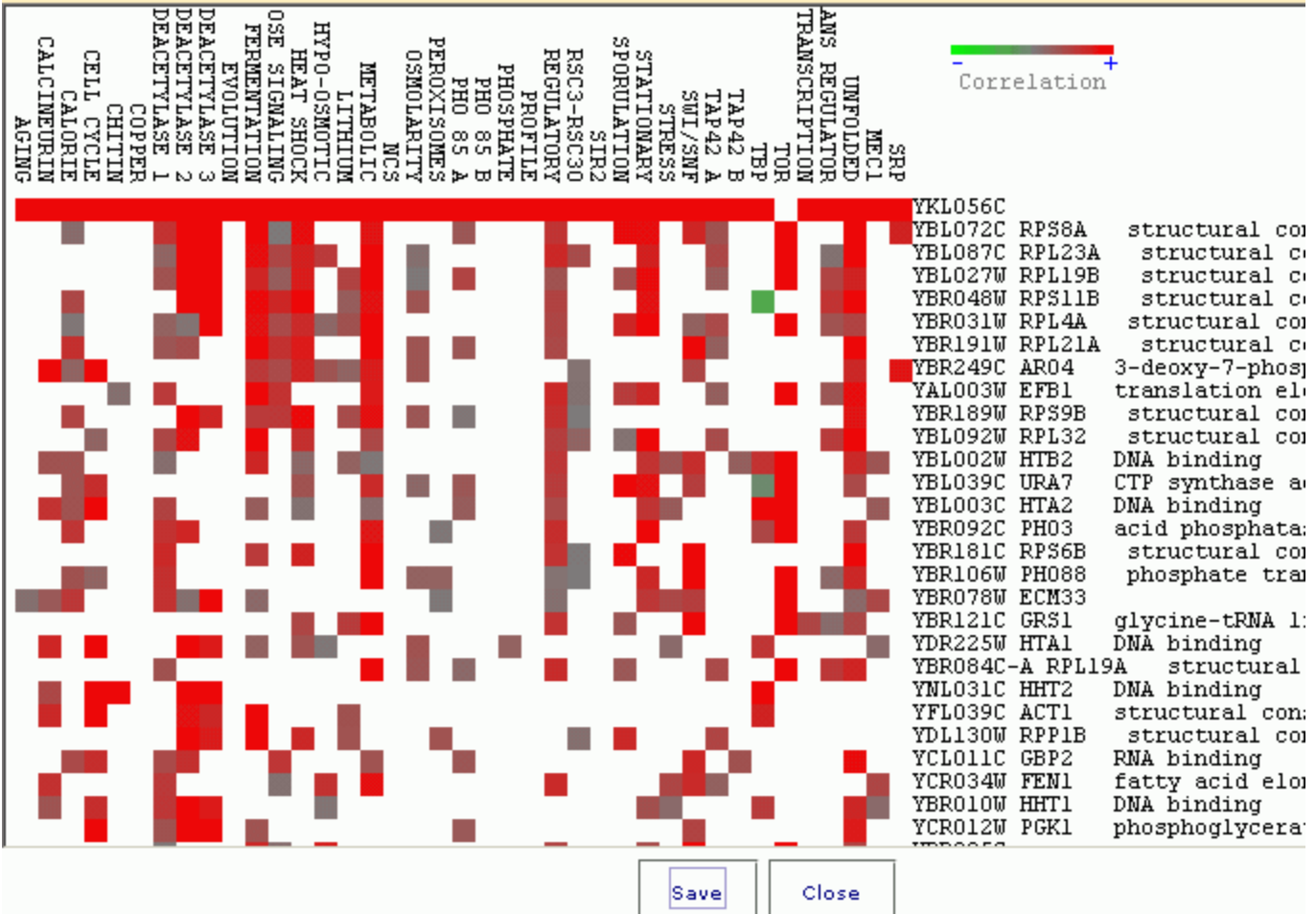


Figure 3. Genes correlated with YKL056C, a gene with unknown function. YKL056C shows strong correlation with ribosome subunits, suggesting its function in protein translation.

The color representation has different meanings in FABA compared to commonly used heatmap for microarrays: In FABA, a red square indicates that the two genes go up or down together in the specific set of data (if the data is array data); it does not suggest that the two genes are both up-regulated. Users will have to go to that dataset to find detailed information on how the two genes are correlated: they could be both up, or both down, and most likely both up under some conditions of the experiments and both down under other conditions. In the case of non-array data, red color simply represents same classification or positive correlation.



Powered By
Functional Analysis
By Association
© 2004, Chang Bioscience, Inc.



Figure 4. Tumor suppressor p53 associated genes. Known genes in the p53 pathway are marked by blue arrows and potential p53-related genes are marked by orange arrows.

Shown in Figure 4 is the result of querying a human gene expression datasets with the tumor suppressor gene p53. A total of 65 datasets and more than 1700 experiments are searched. Note many genes found to share similar gene expression profiles with p53 are known p53-related genes.

Realistic expectations

Although FABA is a powerful tool, and we have tried hard to make it easy to use, it is still a time consuming process to get the best result from FABA. Considerable amount of time is needed to carefully collect datasets from public or private sources, analyze using different queries and noise levels, and examine query results. But compared to other wet-lab techniques such as yeast two-hybrid, far Western, or genetic screening, FABA requires days of work instead of months for most wet-lab experiments. In addition, it is complementary to wet-lab approaches because its search is not limited to only one interaction.

As in wet-lab experiments, FABA will also produce a significant number of false positives. Users should have realistic expectations: as an evolving technology, FABA has advantages compared to other wet-lab technologies, but considerable software development is still needed to reduce the false prediction rate.

Data types

FABA can be used to integrate the following high-throughput data types:

1. *Microarray data independent of array platforms*
2. *Protein array data*
3. *Genome data such as sequence similarities, functional domains, promoter types, etc.*
4. *Genetic data such as genetic interactions and phenotype similarities*
5. *Protein localization and protein-protein interaction data*
6. *Metabolic and signaling pathway data*
7. *Gene Ontology data*
8. *Literature data*

Data that can be accepted by FABA are not limited to the afore-mentioned types. As a design principle, we want FABA to be flexible to analyze most if not all high-throughput data. We want FABA to address the central question in biology: gene function. We want FABA to be friendly software every biologist can use. Our efforts paid off. Users will find that FABA is a powerful yet easy to use tool for gene function studies.

FABA Searches for Gene-Gene Correlations

Behind the scene FABA searches for gene-gene correlations. There are many types of gene-gene correlations. We classified them into three categories for computational purposes: High-throughput, Classification, and Correlation. A few examples of each class are listed below:

1. High-throughput data

Examples: microarray, protein array, high-throughput mass spectrometry.

2. Classification

Examples: protein localization, conserved domain, phosphorylation status, signaling pathways.

3. Correlation

Examples: sequence similarities, genetic interactions, two-hybrid interactions, literature searches.

“High-throughput” data are serial measurements for each gene. For each gene there must be at least four measurements and there must be more than one gene in the dataset. The measurement data must be numerical. A typical “high-throughput” data is the microarray data, which measures the transcription levels of thousands of genes. Other examples include repetitive blood pressure measurements of hundreds of knockout mice under environmental conditions such as stress and diet. “High-throughput” data is not limited to experimental data. Literature data such as the frequency of gene names in thousands of publications may also be used as a rough estimate of existing knowledge of gene functions.

The common characteristics of data in the “high-throughput” category are: (1) the data are measurements for multiple genes; (2) there are multiple measurements (> 3); (3) each measurement contributes equally to the final analysis.

Classification data are classification based on one property of a gene. One example is a gene product’s cellular location. Each classification must be assigned an integer (e.g., 5 for nucleus, 23 for mitochondria membrane, etc.).

Correlation data are gene-gene correlation data. Examples include sequence similarities, genetic interactions, and protein-protein interactions. For each gene pair, the correlation must be scaled to the range of -1 to $+1$.

How to Use FABa Viewer

A. View Data

1. Start FABa. Select a project and click on "Go". A new window will appear.
2. In the query text field, type in a gene id, name, or keywords to search for your query. Hit "Enter" or click on "Go" to search. Search results will appear in the window below.
3. Select a query gene by clicking on it in the result window.
4. Click the "Associate" button to find genes with similar properties.

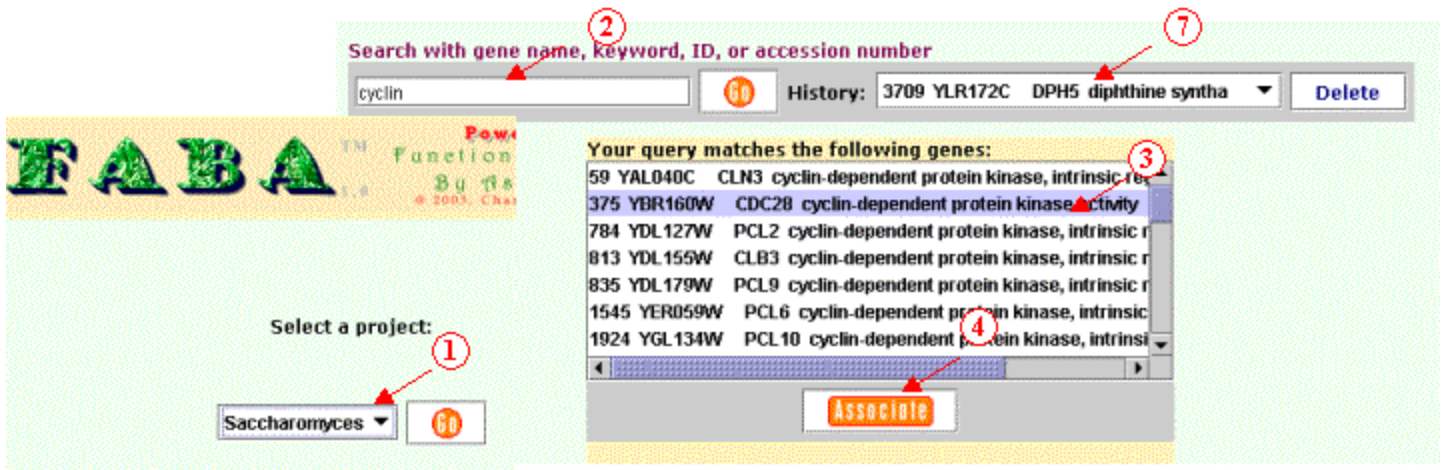


Figure 5. Steps for viewing FABa data.

B. View history

5. FABa will save previous searches in a list called "History."
6. Start FABa. Select a project and click on "Go". A new window will appear.
7. Select from the history pull down menu (right to the "Go" button) a previous result. The results will appear in a new window.
8. To delete saved results, click on the "Delete" button. A new window will appear. Check the results to be deleted and then click on "Delete" in the new window.

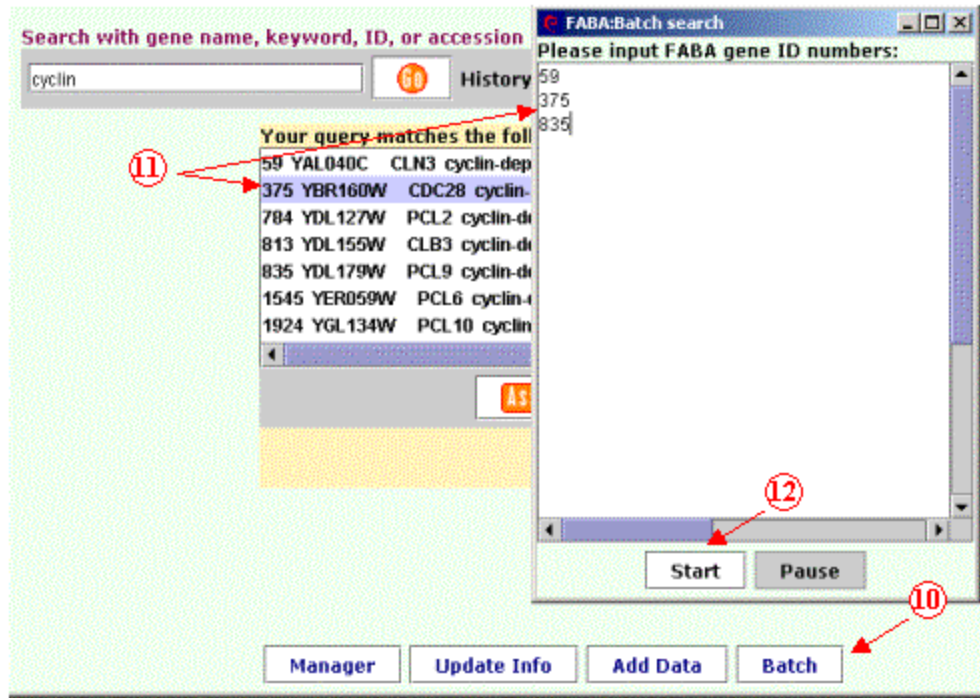


Figure 6. Steps for batch queries.

C. Batch query

9. To submit for multiple queries, start FABA, select a project and click on "Go".
10. In the new window, click on the "Batch" button (lower-right corner).
11. Input **FABA** gene ID numbers, separated by spaces (blank space or new line). Note the program can only recognize **FABA** gene ID numbers. To find FABA gene ID numbers, please use the search function in the project window. FABA gene IDs are the numbers in the first column of the search results.
12. Click on "Start" button to start the search.
13. To pause/resume search, click on the "Pause" button. Please be patient since the "Pause" command will not immediately take effect. It will wait until the query in progress is completed.
14. To view the progress of batch search, use the history pull-down menu.
15. To start a new batch search, pause the current batch search if any, close the batch search window, and then click on the "Batch" button again as in step 10.

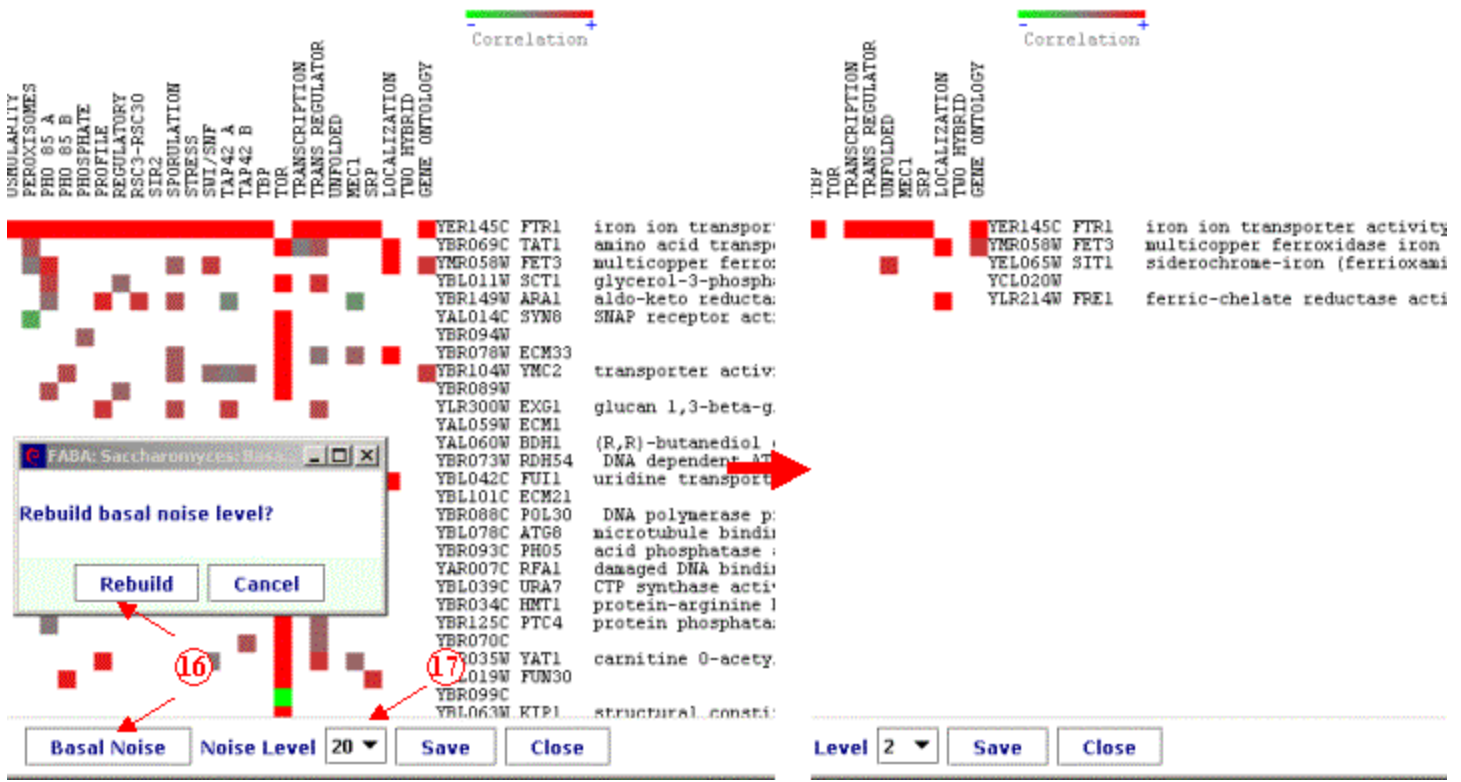


Figure 7. Using lower noise level to show most reliable correlations.

D. Reduce noises

16. If you have added/deleted data, you would need to rebuild the basal noise level before running reduce noise levels. Click on "Basal Noise" button and then "Rebuild" in the pop up window.

17. In the window that displays query results, change the noise level by using the pull down menu at the bottom of the result display. The noise level is scaled from 1 to 20, with 20 the highest and default.

18. Users are suggested to try several noise levels. A rigorous statistical test of noises in gene-gene correlation will be provided in the future versions of FABA.

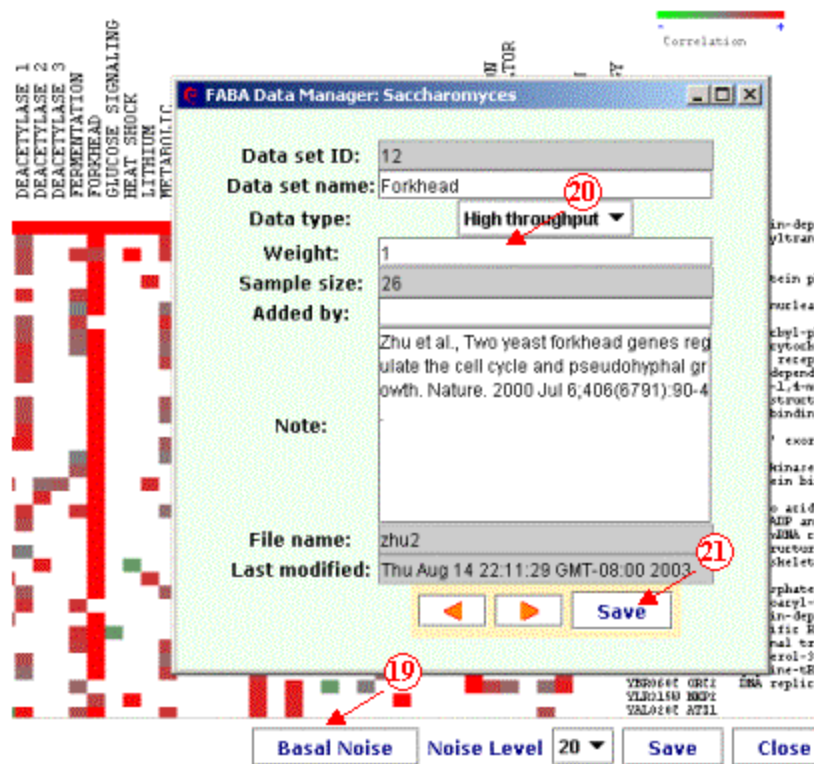


Figure 8. Steps for building basal noise level and changing dataset weight.

E. Refine Your Query

19. Carefully exam the heatmap. Click on a column to read the experiment information.
20. Reduce weight if the dataset contributes mostly noises. One symptom of noise is that a large number of genes show correlations with your query in a dataset, especially if these genes are "house-keeping" genes. (For microarray data, high-expression "house-keeping" genes may show the same noise pattern because of systematic experimental errors, e.g., improper normalization of data.) The other symptom is that the great majority of genes associated with your query in a dataset are known to have different functions. You may remove this dataset by decreasing weight or setting the weight to 0. Weight must be a non-negative number between 0 and 1.
21. Save your changes and run "Associate" again.
22. Repeat steps 16-21 if necessary.
23. Save the heatmap either by printing/capturing screens and paste into a drawing program (e.g., PowerPoint) or click on "Save" to save it as a JPEG file.
24. Start over with a different query. To find a target gene of a signaling pathway, one can use a signal gene or a known target gene as query. A known target may work better because of signaling cross talks and amplification.

F. Import Data

25. If you need to upload your own data, you must obtain a license for FABA. However, FABA Viewer allows users to import data already uploaded to FABA. Following the following steps to add a project created by Chang Bioscience Inc. or your colleague.
26. Uncompress FABA data sent to you if necessary.
27. Start FABA Viewer. Click on “New Project” button. A new window will appear.
28. The new window will show a project name “New”. Click on the “Edit” button next to “New.”
29. In the popup menu, replacing “New” with a unique project name. Click on “Save.”
30. Open the following folder: \$home\data. \$home is where you have installed FABA Viewer.
31. You will see a “project.txt” file (no need to open or save project.txt) and a number of folders in this folder. You should see a folder name that is the first four letters of your new project name, or first three letters followed by a number. If you are not sure which one is the one you just created, check for date of creation. Or open the folder; you should see only three empty files: datalist.txt, noise.txt, and masterGeneList.txt.
32. Copy all the files from the FABA data you received into this folder. When prompt, replace datalist.txt, noise.txt, and masterGeneList.txt. Files. Don’t drag in the top folder. All data files should be within \$home\data\yourproject, not in a subfolder of \$home\data\yourproject.
33. You are all set. Return to FABA Viewer and run your query.

How to Use FABa

A. Data Selection

Carefully select a number of datasets to be analyzed. Selections of dataset should base on relevance and quality. Good data that are not obviously relevant may actually be very helpful because they may have unexpected correlations and they may help define a baseline noise level. Bad data will contribute nothing but noises.

B. Data Preparation

1. Preparation of info file

Gene info data should be a tab-delimited text file in the following format:

Unique ID Accession Number Gene Name Description

The first column is required unique ID. This unique ID field should be identical for the same genes across all the datasets in one project. Suggested unique IDs are Systematic Name for *Saccharomyces cerevisiae* and Unigene ids for species such as human and mouse. The info file **REQUIRES** a header line to avoid mistakes in omitting the "Accession Number" column.

2. Preparation of data

(a) High-throughput data format

Data file should be tab-delimited text file. The first column should be the unique ID field. All other columns are data fields. All columns and rows must have the same length. Missing data can be represented by no space, empty space, or "NA." No header line is needed.

(b) Classification data format

Data file should be tab-delimited text file. The first column should be the unique ID field. The second column is the data field. Classification should be represented by an integer for each class. No missing data is allowed. No header line is needed.

(c) Correlation data format

Data file should be tab-delimited text file. The first and second column should be the unique ID fields. The third column is the correlation data field. Correlation values should be between -1 and $+1$. No missing data is allowed. No header line is needed.

C. Data upload

(a) Before uploading data, it is strongly suggested that users upload the info file first. This will speed up the upload of the data files. To upload info file, start FABA, go to your project, and click on "Update Info" button. A new window will appear. Follow the instructions to upload the info file.

(b) To upload data, click on the AddData button. A new window will appear. Select the correct data type and input the correct sample size (e.g., 10 if your dataset has 10 data columns excluding the id field) if high-throughput data. Click on "Upload" to upload.

(c) The data upload step is time consuming, since a lot of computations are done in the background. A batch utility is available for uploading multiple datasets. Users may run the batch uploading at night or during weekends. To batch upload several files, prepare a tab-delimited summary file for all the files to be uploaded. The first row of the summary file should be the column names:

<i>Name</i>	<i>Type</i>	<i>Scale</i>	<i>Weight</i>	<i>Sample Size</i>	<i>Added By</i>	<i>Note</i>	<i>File Name</i>
-------------	-------------	--------------	---------------	--------------------	-----------------	-------------	------------------

For the *Scale* field, the value must be exactly one of the following: "Unlog", "Log2", or "Log10". The *File Name* field should be the absolute pathway (location) of the data files (e.g. *c:\fabadata\yeast\his2.txt*).

A pause function is provided for batch upload. To prevent corruption of info file, the pause will stop uploading only if one dataset has been finished. Please be patient. The progress bar indicates the progress in the current file being uploaded. Click on "Resume" to continue uploading.

D. Project export

(a) To export a project for sharing with colleagues, open the project file (project.txt. Don't save any changes!) in \$home\data folder. \$home is where you have installed FABA. This file contains the directory name information for each project. The odd rows contain the project names and the even rows contain directory names.

(b) Find the corresponding directory in the \$home\data folder. Use file compress software (e.g. winzip, stuffit) to compress the entire folder into a single file. You may send the compressed file by email or burn it onto a CD.

E. Project imports

(a) Uncompress FABA data sent to you if necessary.

(b) Start FABA Viewer. Click on "New Project" button. A new window will appear. The new window will show a project name "New". Click on the "Edit" button next to "New."

(c) In the popup menu, replacing "New" with a unique project name. Click on "Save."

(d) Open the following folder: \$home\data. \$home is where you have installed FABA Viewer.

(e) You will see a "project.txt" file (no need to open or save project.txt) and a number of folders in this folder. You should see a folder name that is the first four letters of your new project name, or first three letters followed by a

number. If you are not sure which one is the one you just created, check for date of creation. Or open the folder; you should see only three empty files: datalist.txt, noise.txt, and masterGeneList.txt.

(f) Copy all the files from the FABAs data you received into this folder. When prompted, replace datalist.txt, noise.txt, and masterGeneList.txt. Files. Don't drag in the top folder. All data files should be within \$home\data\yourproject, not in a subfolder of \$home\data\yourproject.

(g) You are all set. Return to FABAs Viewer and run your query.

Warning: you cannot merge two projects into one! This function will be provided in the future upgrade.

F. Delete dataset and project

Delete function for dataset and project is not provided. Since uploading data is time consuming, delete function is intentionally omitted to avoid accidental loss of data. See instructions for Advanced Users below on how dataset and project can be deleted by manipulating FABAs files. To remove a dataset from a project, simply set the "Weight" to "0" (step 20).

Think Again

Although FABAs makes it much easier to find genes with similar functions by analyzing high-throughput datasets from different sources, it will be a mistake to think FABAs as a turnkey gene function finder. There are several reasons users must be very careful in reaching conclusions. First there are noises in the high-throughput data. For example as many as 20% of clones on certain microarrays may have mistaken identities. One of FABAs's strength is to reduce some of these noises by comparing data from different sources; however it is impossible to eliminate all the noises. Second FABAs searches mostly gene-gene correlations, which are indirect evidences that the two genes share similar functions. In the absence of any definitive evidence (e.g., genetic interactions, known function similarities), it is prudent to view FABAs results with a little suspicion. Third the datasets may not be ideal for answering your question. For example if you would like to find target genes of a signaling pathway in breast cancer, your datasets must contain significant number of breast cancer data. Fourth the query might not be optimal. If your datasets contain mostly microarray data, query with genes whose transcription levels change little will not likely be informative. Of course if the query gene is absent in most datasets, little correlation information will be available for the query.

To answer specific questions users are strongly suggested to collect their own data selection. Users are also suggested to run multiple queries and examine results carefully. FABAs results may still be noisy, but nevertheless it is a significant improvement from experiments such as single array dataset, differential display, or yeast two-hybrid.

Advanced Users

Only users experienced in editing and managing computer files should attempt to manipulate FABa data at the file level. Before any attempt in editing or deleting FABa data, please make a backup copy so the data can be restored if a mistake has been made. You may lose tens of hours of work if data files are improperly changed or deleted.

A. Data directory/file structure

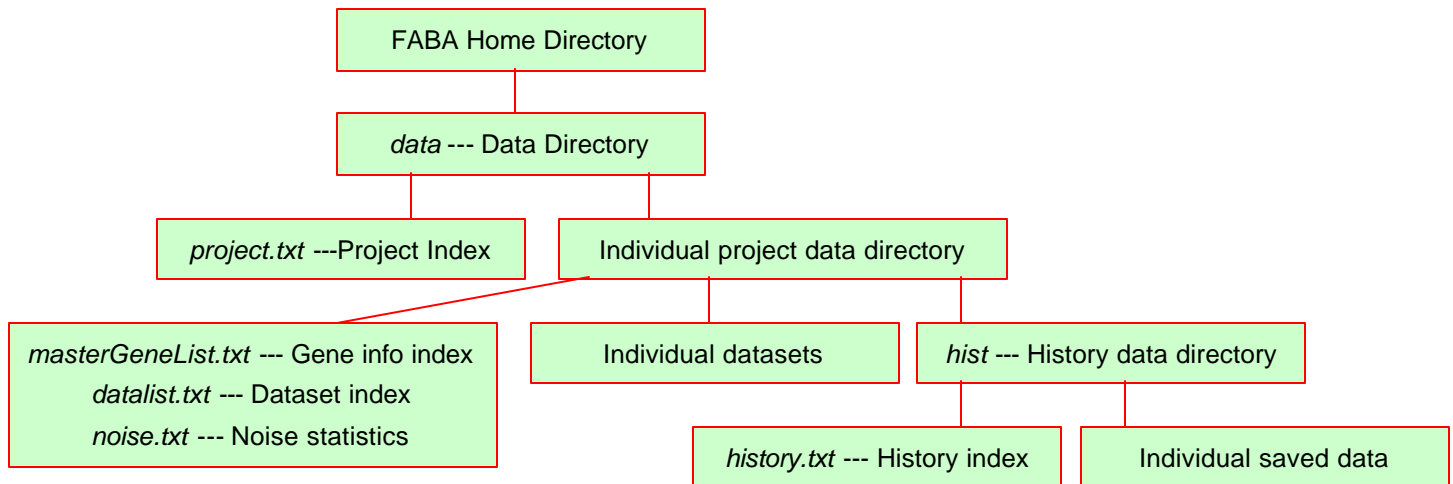


Figure 9. FABa file structures.

B. Backup

To backup all FABa data, find the directory named “data” within the home directory of FABa. Compress the *data* folder using data compression software such as WinZip or Stuffit. Move the compressed data file into a backup folder. To backup a project, first find out the project data directory (see C. Delete project and database) and compress the directory using WinZip or Stuffit.

To restore to the original data, uncompress the backup data and replace the files in the original folder. Please avoid uncompress into a subfolder. A frequent mistake will be uncompress the “data” folder within the original “data” folder (e.g., $\$home / data / data$). The correct structure should be $\$home / data$).

C. Delete project and dataset

(a) Delete a project

1. Open the project index file (*project.txt* in $\$home / data$ folder). The file should look like the following:

Mouse

mous
Saccharomyces
sacc
Yeast
yeas
Human
Huma

The odd rows are project names and the even rows below are the project folder name. For example, the folder for the project “*Saccharomyces*” is “*sacc*”, which is located at *\$home | data | sacc*.

2. To remove a project name from the FABA software project menu, delete the project name and project folder name, and save the project index file. For example after deleting the project “*Saccharomyces*”, the new project index file will look like the following:

Mouse
mous
Yeast
yeas
Human
Huma

To delete project data as well, remove the data folder for the project. For the project “*Saccharomyces*”, trash the entire “*sacc*” folder (*\$home | data | sacc*).

(b) Delete a dataset

Note: Users can remove a dataset from analysis by setting the weight for the dataset to zero (see step 20).

1. Find the data folder for the project as in (a) Delete a project.
2. Open the project folder (i.e., *\$home | data | sacc*).
3. In the project folder, open the dataset index file (i.e., *\$home | data | sacc | datalist.txt*). The file has the following columns:

<i>Data set ID</i>	<i>Data set name</i>	<i>Data type</i>	<i>Scale</i>	<i>Weight</i>	<i>Sample size</i>	<i>Added by</i>	<i>Note</i>	<i>File name</i>	<i>Last modified</i>
--------------------	----------------------	------------------	--------------	---------------	--------------------	-----------------	-------------	------------------	----------------------

Find your data set name, and under the File name column, find the corresponding data file name.

4. Delete the row for the dataset in *datalist.txt* file and save.
5. Find and delete the corresponding data file in the project folder (i.e., *\$home | data | sacc | swis* for the dataset *SWI/SNF*).

System Requirements

FABA is a JAVA application and runs on all platforms JAVA supports. Installers are tested on Windows/NT and Mac OSX. Installers on Linus and Unix are not tested (but should work). We'll be mining Gigabytes of data, a data size most biologists have never managed. It is strongly suggested that you install FABA on the best machine in your lab. Your computer system must have at least 64 Mb of memory and 2 Gb of free space. You may experience freezes when running FABA. Please avoid opening other applications to conserve resources when running FABA. If you plan to add your own high-throughput data (e.g., microarray data), you are advised to reserve 10 Gb of free space for each 1000 arrays to be uploaded. Upgrade your memory to at least 256 Mb if you expect 2000 arrays or more.

Please contact us if you would like us to custom upload your data.

Minimum System Requirements:

System:	Windows 98 and later Mac OSX and later Others: Linus, Unix, Sun (Inquire)
Memory:	>= 64 Mb, 256 Mb or greater suggested
Free storage space:	> 2 Gb, 10 Gb or more suggested

Public data resources

As examples we list below a few public data resources that can be used for FABA analysis. This list is far from complete.

A. *Gene expression data*

NCBI Gene Expression Omnibus (GEO): <http://www.ncbi.nlm.nih.gov/geo/>

Data depository of gene expression data, including microarray and SAGE data.

Stanford Microarray Database: <http://genome-www5.stanford.edu/>

Yale Microarray Database: <http://info.med.yale.edu/microarray/>

Yeast Microarray Global Viewer: <http://www.transcriptome.ens.fr/ymgv/>

B. *Protein-protein interaction*

UCLA database of interaction proteins (DIP): <http://dip.doe-mbi.ucla.edu/>

Biomolecular Interaction Network Database (BIND): <http://www.blueprint.org/bind/bind.php>

C. *Protein localization*

Yeast protein localization server: <http://bioinfo.mbb.yale.edu/genome/localize/>

D. *Protein domains*

Pfam: <http://www.sanger.ac.uk/Software/Pfam/>

E. *Gene Ontology*

Saccharomyces Genome Database (SGD): <http://www.yeastgenome.org/>

Disease genes (OMIM): <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>

Related Products

A. *BioToolKit 300*

BioToolKit 300 contains a number of software tools for biologists, including *GodlistManager* and *MicroHelper*. These tools will be handy for preparing FABA data.

GodlistManager has functions needed for batch GenBank search. One particularly useful option is to map accession numbers to Unigene IDs.

MicroHelper is useful for merging raw microarray data into a single file. It can also be used for filtration, normalization, and data transformation.

B. *The Electronic Protocol Book*

Perl language is frequently used by Bioinformatics scientists. A number of Perl scripts are included in *the Electronic Protocol Book*. They can be used to prepare data for FABA uploading.

Frequently asked questions

1. Does FABA have any limitations on the types of high-throughput data it can analyze?

The limitations are that you need to format your data into one of the following three types:

- 1. High-throughput data. Minimum sample size for this type is 4. The data should be in matrix form:
Gene ID Data1 Data2 Data3 Data4 ...
- 2. Classification. Data should be in two columns: the first column unique gene id, and the second column classification. The classification should be an integer.
- 3. Correlation: Data in three columns, the first two columns are gene ids for the pair and the third column correlation values.

2. Do I have to use Unigene ID as identification? Can I use my internal database ID?

Sure. As long as the identification is unique for each gene.

3. Do you have data for any other species?

We rely on public available high-throughput data. We'll appreciate if you could inform us a source of HTP data we missed. If there is little HTP for a particular species, there is little we can do.

4. How does FABA help me? I don't have any HTP data.

Everyone can benefit from the huge amount of public HTP data. Without the help of FABA it will take you days to weeks to check individually all the datasets. With FABA you can find correlated genes in minutes. The rigorous and consistent statistics eliminate many false positives. By comparing data from a large number of sources, we further reduce the number of false positives.

5. How do you determine the noise level?

The noise level is determined based on statistics of gene-gene correlations for all the experiments. Briefly a gene is considered noisy if it shows a large number of correlations to other genes. These genes have a large number of uninformative correlations, i.e., noisy in the overall gene-gene correlation that FABA measures.

6. Why are some genes noisy?

We speculate that most noisy genes (1) are quasistable expressions, (2) fluctuate with environmental conditions such as nutrients and stress, or (3) contain over-represented sequence elements.

7. Which noise level should I use?

There is no one level for all queries. Future version will provide better statistical guidelines for choosing an appropriate noise level.

8. FABA results seem to change for different versions of FABA. Why?

There are two reasons. First FABA search is not exhaustive because we need to limit the number of searches such that it can finish each query in a reasonable amount of time on personal computers. Newer version may have improvement in search speed, thus allowing more searches to be done. Second, functionally correlated genes may be stable under certain experimental conditions. Correlations of stable genes are not strong evidences of co-regulation. We are still experimenting to find an appropriate contribution of stable correlations to the final correlation score.

9. Since there are so many false correlations in each individual dataset, will FABA predictions be useful at all?

The false positives will be a serious problem. One goal of FABA is to identify some of the false positives and to reduce false positives by comparing correlations in different datasets. FABA is not perfect, but nevertheless it is an improvement over analyzing single dataset.

10. FABA ignores differences in array types and blindly compares microarray apples and oranges. How can such an approach produce any meaningful results?

FABA does not look for differences between microarray apples and oranges, but searches for reproducible gene-gene correlations in multiple datasets.

11. Are you absolutely sure that array data from different platforms can be compared?

Yes! Even more complicated meta-analysis can be achieved. In one study, by retaining reproducible data we analyzed together three NCI60 datasets (two different Affymetrix arrays and one cDNA array (signal channel only)). All three datasets measured gene expression of a panel of 60 cell lines used in the NCI Developmental Therapeutics Program (DTP). A hierarchical clustering showed that the great majority of samples clustered according to their histological origins. The clustering result is a significant improvement over a single dataset.

12. Why are you using only simple normalization options for microarray data?

Our priority is to build a rough but functional program. There are many other factors need to be examined and tested. Normalization is one of them but not necessarily the most significant one. For gene-gene correlation, we calculate the Pearson correlation. Poor normalization will introduce error, but that error may still be smaller than the uncertainty due to small sample size.

13. *What do you mean gene expressions fluctuate with environmental conditions such as nutrients and stress? Dose it mean I can increase my beta Actin if I eat more?*

We sometimes fail to appreciate how quickly our bodies respond to environmental changes. A sight of food makes our mouth wet, a single breath of cigarette relaxes us, and we die in five minutes without oxygen. Expression of a few genes will respond quickly to environmental changes. By the way, if one's body does not signal enough food intakes in a few minutes, one's total beta Actin will definitely increase in a short couple of days.

14. *Many of the genes found to be "noisy" by FABA seems to be "house-keeping genes." "House-keeping genes" have higher expression levels, thus should be less noisy.*

The definition of "noise" is the noise in the correlation score. Quasi-stable genes are "noisy" in FABA because they have many un-informative correlations with other quasi-stable genes.

15. *Are "house-keeping genes" stable genes?*

Although whether "house-keeping genes" are stable within a factor of 1 or 2 is not an interesting question in biology, it is a heated controversy in the microarray community because a number of scientists advocate data normalization based on "house-keeping genes" expressions. If you normalize array data using "house-keeping genes" as standards, then expressions for these genes are stable because that is the assumption. If a different normalization procedure is used, the "house-keeping genes" are quasi-stable most of the time. But one can always argue that the later normalization procedure is not optimal. Nevertheless the selection of "house-keeping genes" is quiet arbitrary and many change their expression levels during developments.

16. *I'm still confused about how can you ignore array types and data scales.*

FABA is based on the simple principle: if two genes go up and down together in only one set of data, the correlation is most likely random; if they go up and down in many sets of data, the correlation is most likely not random. The qualitative measure used to implement the simple principle in FABA 1.6 seems to work fine. But the measure itself is open to improvement.

17. *Are you worried about the quality of the public microarray data?*

These public data represent the current state of the art of microarray technology. Until we have a good and uniform criteria to evaluate each individual dataset, users need to be aware certain individual datasets may be of poor quality. On the bright side, a quick visual inspection of FABA results can in many cases reveal questionable datasets.

License Requirements

FABA license specifically requires users not to modify or extend FABA codes. The primary reason is that we would like to maintain a data standard such that data exchange between different groups is easy. This is also the reason we started FABA project in the first place. It will be counter-productive if different groups are not able to compare notes because different versions of software are used. Please contact us if you would like to develop similar software. We would be glad to collaborate or provide any help we can.

Customer Support

Thanks for using our software. If you have any questions or suggestions, please don't hesitate to contact us at info@changbioscience.com. Your feedback will help us to improve the software and will be greatly appreciated.

We'll answer your questions as soon as we can. But due to limitations of our resources, priority will be given to licensed users first. For non-licensed users, please be patient. Users can also post questions to the FABA user group at <http://www.changbioscience.com/forum/phpBB2/>.