

## Microarray Evaluation for Gene Regulation Analysis

Course Tutorial and Step by Step Example

[www.genomatix.de](http://www.genomatix.de)

Contact:

**Genomatix**

Landsberger Strasse 6

80339 München

Fon: +49-89-5997660

Fax: +49-89-59976655

e-mail: [info@genomatix.de](mailto:info@genomatix.de)

## Contents

1. Introduction.....	3
2. Theoretical Background.....	4
3. Practical Example: Evaluation of the Role of PDGF in Fibroblasts.....	5
3.1. Step 1: Statistical Analysis.....	6
3.2. Step 2: Literature Analysis.....	8
Step 2a: Functional Subcluster Analysis.....	10
Step 2b Transcription Factor Analysis.....	12
3.3 Step 3: Promoter Analysis .....	13
FrameWork Analysis.....	16
Promoter Database Scan .....	17
Intermediate Result Evaluation 1 .....	20
Intermediate Result Evaluation 2 .....	23
Intermediate Result Evaluation 3 .....	25
3.4. Step 4: Additional Statistical Analysis.....	26
3.5 Step 5: Merging of Results Into Biological Context.....	28
Pathway Mining Using BiblioSphere .....	29

This tutorial was compiled by Dr. Martin Seifert.

Martin Seifert has a Ph.D in Biology. After spending some time at Brunel University (London) he accepted the position of head of the microarray facility at the Department of Cell Biology of the Technical University of Munich. In spring 2004 he joined Genomatix adding his expertise in DNA-microarray analysis to Genomatix training team.

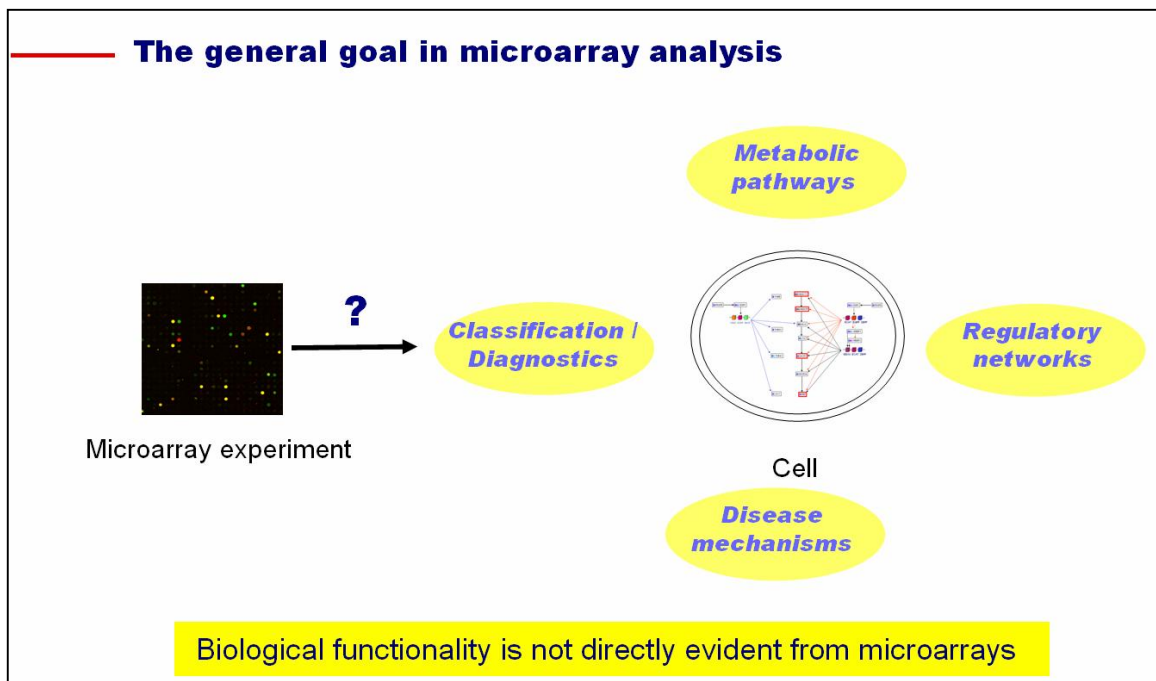
## 1. Introduction

Microarray mining is a challenging task because of the intrinsic superposition of several biological processes mirrored in the data. However, general aims of microarray analysis always comprise classification and diagnostics of samples, gaining insight into metabolic pathways and regulatory networks and finally learning more about disease mechanisms (Figure 1).

In this tutorial Genomatix presents a strategy for microarray mining based on the combination of statistical significance analysis of gene expression, literature-mining, and promoter analysis. The strategy is illustrated by a stepwise analysis of publicly available microarray data (a timeline experiment of PDGF-stimulated fibroblasts). The applied strategy reveals results beyond the detection of the major metabolic pathway known to be directly linked to the PDGF response: It is possible to identify cross-talking regulatory networks underlying the metabolic pathway without using a priori knowledge about the experiment. These regulatory networks involve the EGR1 factor, which is a known target of PDGF stimulation.

To reproduce the analysis all necessary data can be downloaded at:

<http://www.genomatix.de/download/tutorial/>



## 2. Theoretical Background

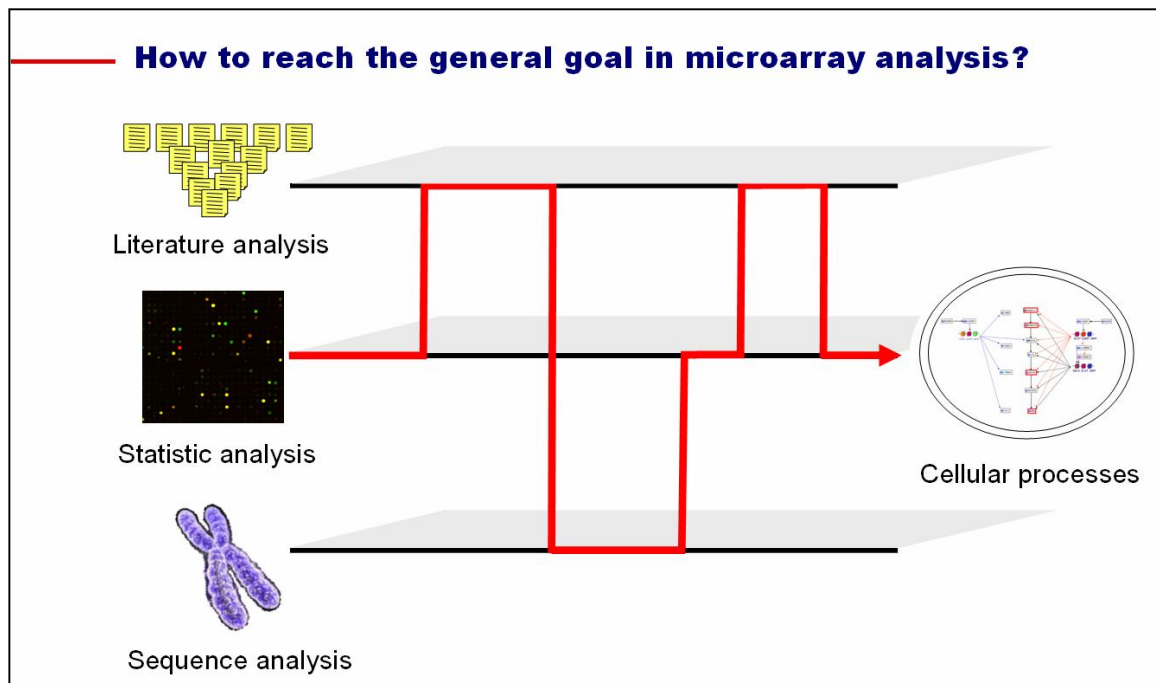
Microarray results reflect a multitude of simultaneous cellular processes although only subsets of expression changes are directly caused by the experimental conditions. Therefore, a major task for an in depth analysis is to identify genes whose expression changes due to the experimental setup and distinguish them from effects of biological diversity or general stress response of the cell.

In this tutorial we present a combinatorial strategy, which includes the analysis of gene promoters, for a biological evaluation of relationships between significantly regulated genes. The procedure is based on a combination of statistical-, literature- and promoter analysis and aims at establishing gene regulatory networks on molecular level.

No single method could solve the task for the following reasons:

Statistical analysis reveals mRNA with significantly changed expression levels but fails to assign these changes to biological events. Projecting microarray data onto pathway information from literature allows to associate genes with biological processes, but is restricted to current knowledge and fails to select those genes that are directly pertinent for the experimental conditions. Promoter analysis is capable of revealing targets of transcriptional coregulation but cannot discern molecular mechanisms of regulation directly from the initial microarray data.

We use the principle of biological consistency and comprehensiveness to complete the picture by combination of these methods, which also allows for the integration of genes missed by individual methods.



### 3. Practical Example: Evaluation of the Role of PDGF in Fibroblasts

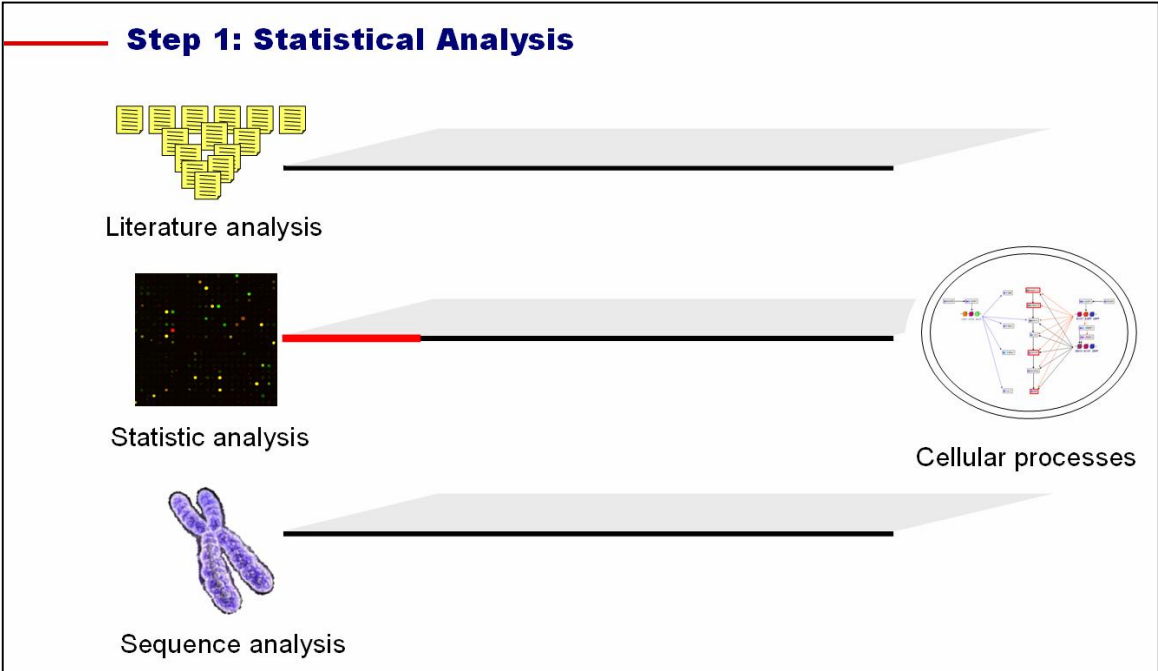
To illustrate the strategy the tutorial shows the analysis of data from a study of platelet-derived-growth-factor (PDGF) stimulation of human fibroblast cells, carried out in a time series study using cDNA microarrays [1]. The data set can be retrieved from the Gene Expression Omnibus (GEO) database (dataset GSE1484). <http://www.ncbi.nlm.nih.gov/geo/>

The analysis follows the steps below:

#### **Workflow of the project**

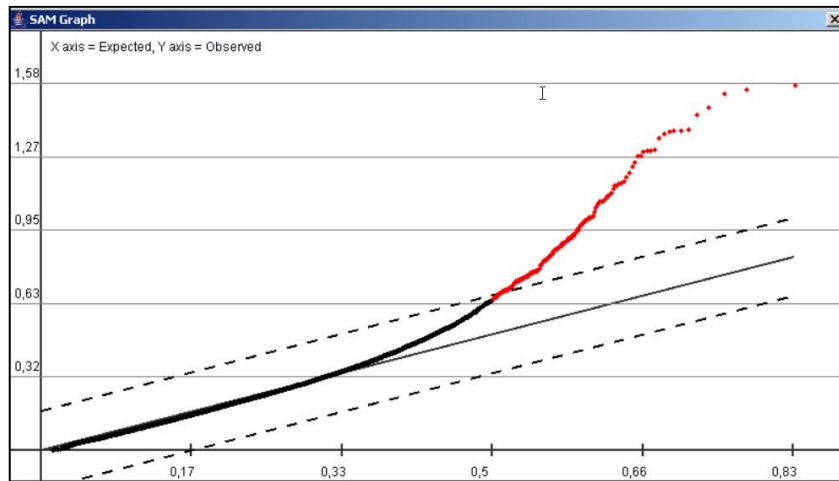
- 1 Find statistical clusters
- 2 Project statistical clusters onto biology and categorization of results by z-scoring (**BiblioSphere**)
- 3 Analyze functional groups for co-regulation (**EIDorado** & **GEMS**) and find additional potentially co-regulated genes (**ModelInspector**)
- 4 Carry out additional statistical analysis
- 5 Merge results into biological context

3.1. Step 1: Statistical Analysis



Microarray data were statistically analyzed using the "Multi Experiment Viewer - MeV included in the TM4 software package from TIGR ([www.tigr.org](http://www.tigr.org)): The Data was first normalized using total intensity normalization. Subsequently the SAM-algorithm [2] with a false discovery rate (FDR) below 5% was applied.

The resulting list contained 105 significant up regulated genes. This list of genes can be downloaded at [http://www.genomatix.de/download/tutorial/PDGF\\_tutorial.xls](http://www.genomatix.de/download/tutorial/PDGF_tutorial.xls) for subsequent analysis to follow the strategy.



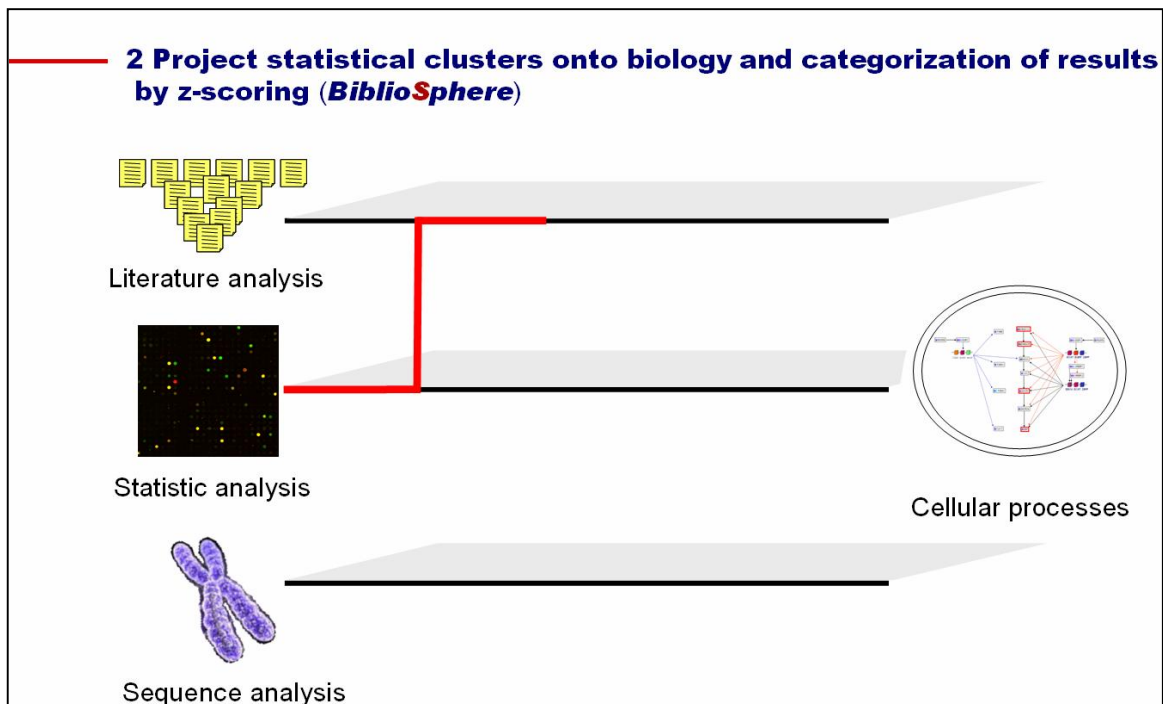
**Statistical analyzed microarray data data**

- Significance Analysis for Microarrays (SAM; FDR: 4,3%)
- 105 of 9928 gene spots are significantly up regulated (Chip: Hver1.2.1)

hours PDGF induction      1      4      10      24

The heatmap displays gene expression levels over time. The X-axis is labeled "hours PDGF induction" with values 1, 4, 10, and 24. The Y-axis represents individual gene spots. The color scale ranges from green (low expression) to red (high expression). A vertical red bar is present on the right side of the heatmap.

## 3.2. Step 2: Literature Analysis



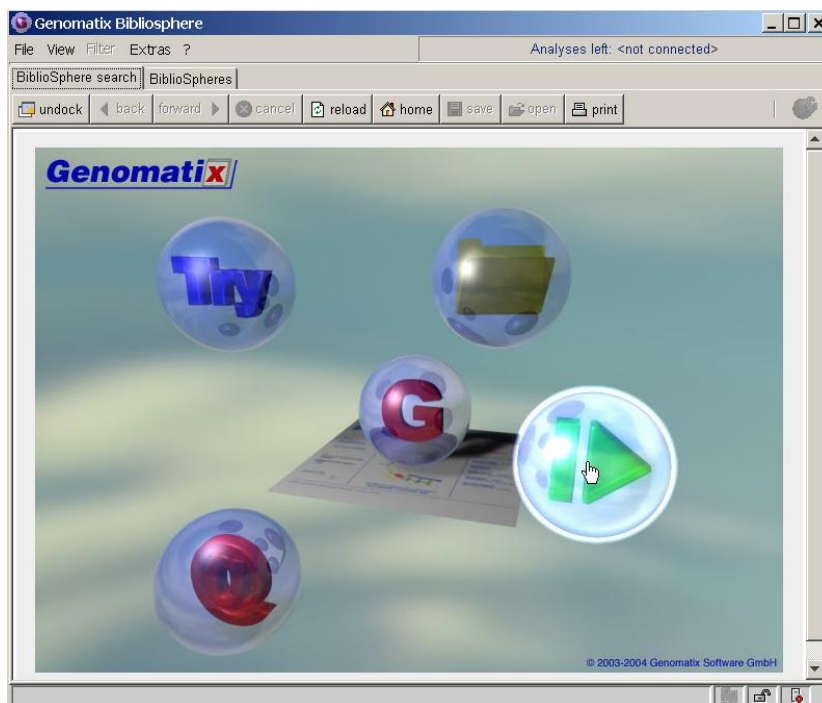
The 105 genes from step 1 are now used for a subsequent analysis using BiblioSphere.

BiblioSphere is a data-mining solution to extract and analyze gene relationships from literature databases and genome-wide promoter analysis. BiblioSphere contains literature data mining strategies using more than 350.000 quality checked gene names, synonyms and Genomatix proprietary semantic relation concepts. Based on PubMed, BiblioSphere currently searches over 12 million abstracts. The BiblioSphere program can be downloaded at:

<http://www.genomatix.de/products/BiblioSphere/bibliosphere6.html>

The aim of the analysis is to find whether there are functional sub-clusters within the 105 genes. And if there are transcription factors which are related with the genes.

BiblioSphere offers different entry possibilities: There is a single gene query to find the literature environment for single genes, a gene group query allows to find co-citations within a gene group plus the literature environment of the input genes. For this tutorial the third option “large cluster query” is applied. In the case of the tutorial we will first only have a look at the input genes and perform a functional filtering.



## Step 2a: Functional Subcluster Analysis

The list of 105 genes is submitted to BiblioSphere™ using the large cluster query option and the Gene Ontology filter “biological process” is applied to gain subgroups with the same GO annotation to find functionally correlated groups. This makes sense as e.g. most pathways belong to particular biological processes. Each resulting subgroup is scored by a z-score: The z-score for an item indicates how far and in what direction that item deviates from its distribution's mean. Z-scores are sometimes called "standard scores".

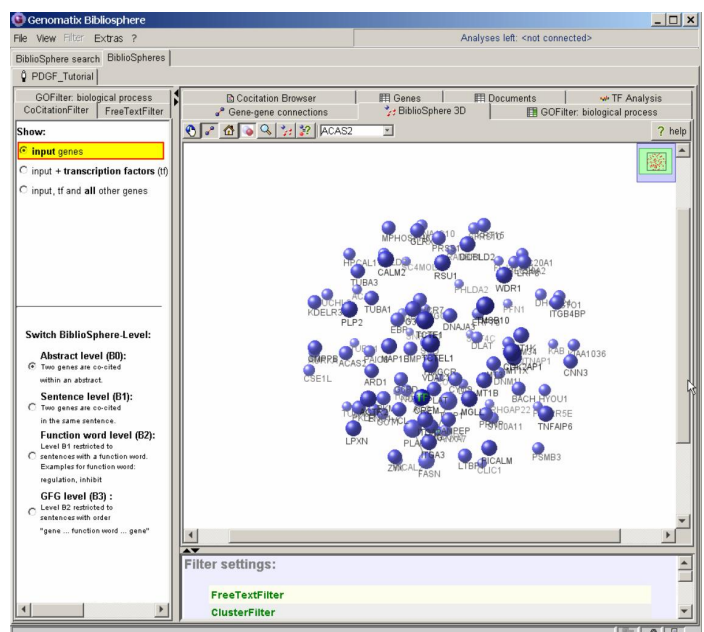
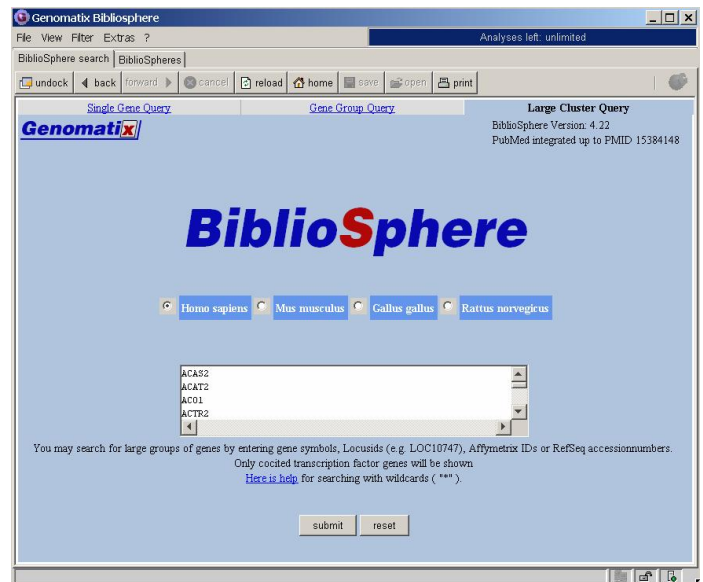
### Box 1

To follow the strategy please take the 105 significantly up regulated genes and put them into the large cluster query input screen of BiblioSphere.

Use human as species.

After the query you will receive a 3D BiblioSphere view.

Please go to the “CoCitation Filter” in the left window and click “Show input genes”



**Box 2**  
Apply the biological process filter from the filter pull down menu

Go to the GOFilter: biological process table and sort for the highest z-score by clicking on the column header

Click on the category with the highest z-score: sterol biosynthesis

Term	ID	O...	E...	Z...
sterol biosynthesis	GO:0016126	6	0.19	13.38
cholesterol biosynthesis	GO:0006695	5	0.17	11.9
germ cell migration	GO:0008354	1	0.01	11.18
removal of superoxide radicals	GO:0019430	1	0.01	11.18
negative regulation of neurogenesis	GO:0050768	1	0.01	11.18
regulation of membrane protein ectodomain	GO:0051043	1	0.01	11.18
negative regulation of dendrite morphogenesis	GO:0050774	1	0.01	11.18
glucose 6-phosphate utilization	GO:0006010	1	0.01	11.18
negative regulation of membrane protein	GO:0051045	1	0.01	11.18
sterol metabolism	GO:0016125	7	0.46	9.7
acetyl-CoA metabolism	GO:0006084	2	0.05	8.99
cholesterol metabolism	GO:0008203	6	0.43	8.56
alcohol metabolism	GO:0006066	12	1.55	8.5
steroid biosynthesis	GO:0006694	6	0.46	8.22
acetate metabolism	GO:0006083	1	0.02	7.85
regulation of dendrite morphogenesis	GO:0050773	1	0.02	7.85
acetate biosynthesis	GO:0019413	1	0.02	7.85
negative regulation of proteolysis and	GO:0045861	1	0.02	7.85

**Filter settings:**  
FreeTextFilter  
ClusterFilter  
There is no cluster center selected  
GOFilter: biological process  
The current filter value is "sterol biosynthesis".

**Box 3**  
Go to the "Genes" table. All genes not filtered out are flagged "no" all genes filtered out are flagged "yes".

Copy the genes flagged "no"

ge...	g...	re...	id...	u...	fit...	m...
1	SC4MOL	sterol-C4...	NONE	6307	NM_006745	no
2	DHCR7	7-dehydro...	NONE	1717	DHCR7	no
3	DHCR24	24-dehydr...	NONE	1718	NM_014762	no
4	EBP	emopamil...	NONE	10682	NM_006579	no
5	HMGCR	3-hydroxy...	NONE	3156	NM_000859	no
6	HMGCS1	3-hydroxy...	NONE	3157	NM_002130	no
7	HYOU1	hypoxia u...	NONE	10525	NM_006369	yes
8	PICALM	phosphati...	NONE	8301	NM_007166	yes
9	PPP2R5E	PPP2R5E	NONE	5529	NM_006246	yes
10	KRT15	keratin 15	NONE	3666	NM_002275	yes
11	NQO1	NAD(P)H...	NONE	1728	NM_000903	yes
12	MGLL	monoglyc...	NONE	11343	MGLL	yes
13	SNRPD2	small nucl...	NONE	6633	SNRPD1	yes
14	SNRPD1	small nucl...	NONE	6632	NM_006938	yes
15	GLRX	glutaredox...	NONE	2745	GLRX	yes
16	GOT1	glutamic-o...	NONE	2805	GOT1	yes
17	FASN	fatty acid...	NONE	2194	FASN	yes
18	MAP1B	microtubul...	NONE	4131	MAP1B	yes
19	BACH	brain acyl...	NONE	11332	NM_007274	yes

**Filter settings:**  
FreeTextFilter  
ClusterFilter  
There is no cluster center selected  
GOFilter: biological process  
The current filter value is "sterol biosynthesis".

**Result** Using BiblioSphere genes can be categorized in different biological ontologies provided by GO or MeSH. Subgroups of genes derived from experimental results or by biological criteria are differentially associated with these categories. The degree of association is expressed by the z-score.

The top-scoring group in our experiment has a z-score of 13 and contains 6 genes, scored in the category “sterol biosynthesis”. Thus, a purely data-driven approach correctly identifies a subgroup of genes known to be involved in PDGF response [1].

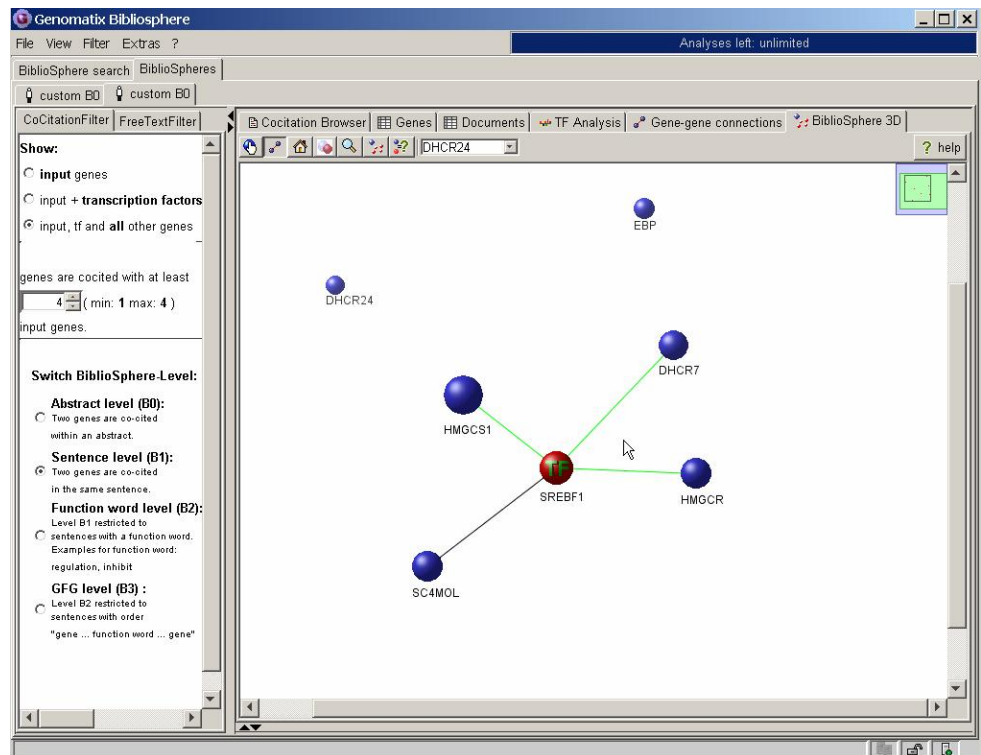
## Step 2b Transcription Factor Analysis

### Box 4

Paste the six genes from the category “sterol biosynthesis” into a new “gene group query” of BiblioSphere.

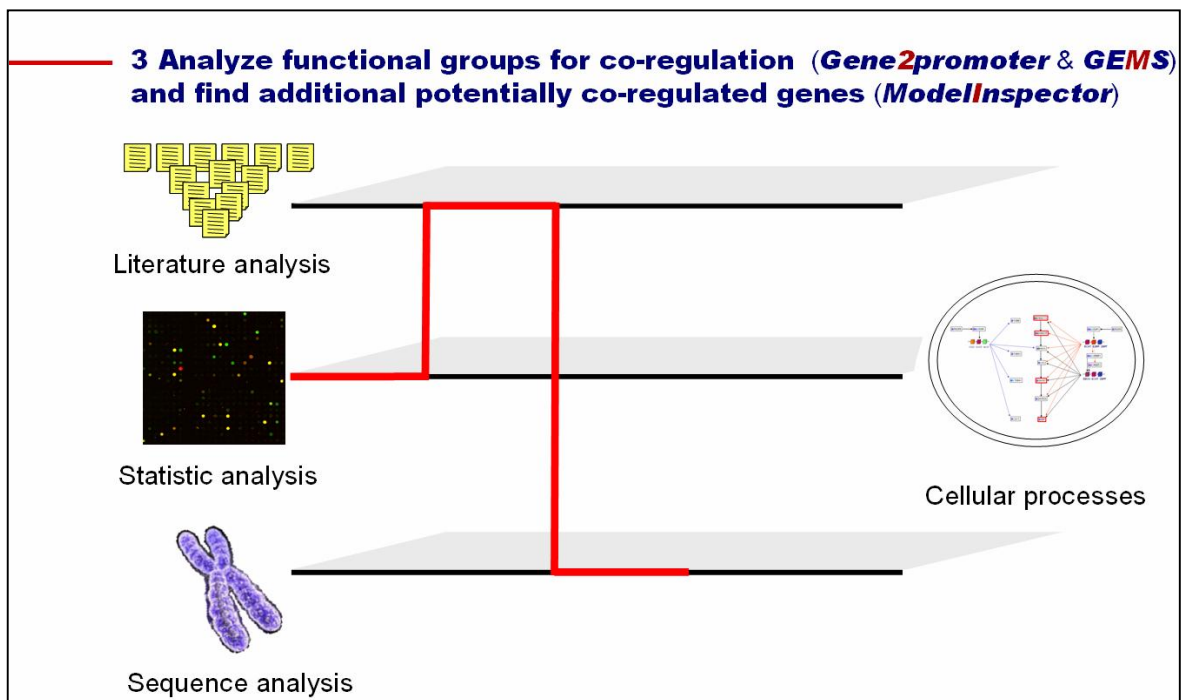
Have a look at “input genes and all other genes”

To make the search more stringent switch the BiblioSphere-CoCitation Level to “Sentence” and look for genes which are at least cocited with four of the six input genes.



The six genes from the “sterol biosynthesis” group (SC4MOL, HMGCS1, DHCR7, HMGCR, EBP, and DHCR24) are re-analyzed as a gene group and BiblioSphere is used for the identification of transcription factors which are cocited with as many of the genes of the “sterol biosynthesis” group as possible. This results in the identification of a single factor, SREBF1 (sterol regulatory element binding transcription factor 1), which is cocited with 4 input genes (SC4MOL, DHCR7, HMGCS1, HMGCR). Potential binding sites for SREBF1 (a member of the E-box family of transcription factors) can be verified on sequence level for three of the cocited genes (HMGCS1, HMGCR, and DHCR7; indicated by the green connecting lines). SREBF1 induction by PDGF had been already confirmed experimentally by Demoulin et al[1], supporting our independent finding from data analysis. In addition SREBF1 is known to be involved in the sterol-induction of HMGCS1 in hamster [3].

## 3.3 Step 3: Promoter Analysis



Coregulation of mammalian genes usually depends on sets of transcription factors rather than individual factors alone. Regulatory sequence elements are often organized into defined frameworks of two or more transcription factor binding sites and clusters of such motifs. The aim of the sequence analysis is to find such transcription factor motifs in regulated genes.

Therefore, the corresponding human promoter sequences for all six genes from step 2 are analyzed with the EIDorado™/Gene2Promoter system. EIDorado is a genome annotation database which includes promoter sequences at highest quality levels. EIDorado is based on a condensation of publicly available data plus Genomatix proprietary annotation, including promoters, transcription factor binding sites, promoter modules, scaffold/matrix attachment regions (S/MARs), and single nucleotides polymorphisms (SNPs) as well as comparative genomics. Gene2Promoter is a multiple identifier interface to query EIDorado.

## Box 5

To follow the strategy, paste the six genes from the category “sterol biosynthesis” into the Gene2Promoter interface at the GenomatixSuite main page (<http://www.genomatix.de/cgi-bin/eldorado/main.pl>). You can either enter gene symbols or Locus IDs. This will allow you to retrieve and select the promoters for subsequent promoter analysis.

Genomatix License Agreement Your Comments Personal Gene2Promoter Available Genomes

GenomatixSuite GEMS Launcher EIDorado Gene2Promoter BiblioSphere  
 FAQ Results Sequences Protocol Help

### Result for Loc3157, Loc3156... (6 loci) (6 seq.)

6 loci were read from input!		Legend	
6 loci were analysed		gold	experimentally verified 5' complete transcript
7 transcripts were found in the mapped sequences		silver	transcript with 5' end confirmed by PromoterInspector prediction
8 unique promoters were found		bronze	annotated transcript, no confirmation for 5' completeness

Results				
	Locus	Selection	Promoter	Transcript/TSS
Loc3157 found on:	Homo sapiens Chromosome 5 Contig NT_006576 (-)  3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble) HMGCS1 (Loc3157)  View this locus in EIDorado More Gene Info Comparative Genomics PubMed	<input checked="" type="checkbox"/>	P930396 43286167-43286771 (-) 605bp	AK095492 (12 exons) TSS = 501  NM_002130 (11 exons) TSS = 505
		<input checked="" type="checkbox"/>	P930397 43271733-43272333 (-) 601bp	CompGen promoter (no transcript assigned)
Loc3156 found on:	Homo sapiens Chromosome 5 Contig NT_006713 (+)  3-hydroxy-3-methylglutaryl-Coenzyme A reductase HMGCR (Loc3156)  View this locus in EIDorado More Gene Info Comparative Genomics PubMed	<input checked="" type="checkbox"/>	P861897 25226957-25227557 (+) 601bp	NM_000859 (20 exons) TSS = 501
		<input checked="" type="checkbox"/>	P861898 25232290-25232902 (+) 613bp	CompGen promoter (no transcript assigned)
Loc6307 found on:	Homo sapiens Chromosome 4 Contig NT_016354 (+)  sterol-C4-methyl oxidase-like SC4MOL (Loc6307)  View this locus in EIDorado More Gene Info Comparative Genomics PubMed	<input checked="" type="checkbox"/>	P935712 90743489-90744089 (+) 601bp	NM_006745 (6 exons) TSS = 501
Loc10682 found on:	Homo sapiens Chromosome X Contig NT_079573 (+)  emopamil binding protein (sterol isomerase) EBP (Loc10682)  View this locus in EIDorado More Gene Info Comparative Genomics PubMed	<input checked="" type="checkbox"/>	P935045 11229566-11230166 (+) 601bp	NM_006579 (5 exons) TSS = 501
Loc1718 found on:	Homo sapiens Chromosome 1 Contig NT_032977 (-)  24-dehydrocholesterol reductase DHCR24 (Loc1718)  View this locus in EIDorado More Gene Info Comparative Genomics PubMed	<input checked="" type="checkbox"/>	P951238 9171975-9172808 (-) 834bp	NM_014762 (9 exons) TSS = 525
Loc1717 found on:	Homo sapiens Chromosome 11 Contig NT_033927 (-)  7-dehydrocholesterol reductase DHCR7 (Loc1717)  View this locus in EIDorado More Gene Info Comparative Genomics PubMed	<input checked="" type="checkbox"/>	P964753 1382048-1382648 (-) 601bp	NM_001360 (9 exons) TSS = 501

Select genes by Gene Ontology annotation of

#### Available tasks for selected promoters above

<b>Select promoter length:</b> <input checked="" type="radio"/> Genomatix optimized length <input type="radio"/> User defined length: <input type="text" value="500"/> bp upstream of first TSS and <input type="text" value="100"/> bp downstream of last TSS (max.: 3000)	<b>Save selected promoters</b> <input checked="" type="checkbox"/> in GenBank format <input type="checkbox"/> in FASTA format <input type="checkbox"/> export to EXCEL format  <b>Submission to GEMS Analysis</b> <input type="checkbox"/> Search for common transcription factor binding sites <input checked="" type="checkbox"/> Search for a common framework of transcription factor binding sites <input type="checkbox"/> Search for promoter modules from the <input type="text" value="vertebrate"/> library <input type="checkbox"/> Multiple alignment of selected promoters <input type="checkbox"/> Multiple alignment plus common TF sites in aligned regions
--	---

For questions or bug reports, please contact [support@genomatix.de](mailto:support@genomatix.de)

© Genomatix Software GmbH 1998-2005 - All rights reserved. GenomatixSuite 3.3.0

**Box 6**  
 Leave all promoters in the analysis as “Selected” and proceed with “Search for a common framework of transcription factor binding sites”. This will open the FrameWorker program of GEMS-Launcher.

## Box 7

Use the “vertebrates” library and “all matrices”.

**Please enter the possible elements of the common framework:**

Library selection	
<a href="#">Matrix group</a> ( <a href="#">View transcription factor &lt;-&gt; matrix assignment</a> )	<input type="checkbox"/> Fungi <input type="checkbox"/> Other Functional Elements <input type="checkbox"/> Insects <input type="checkbox"/> Plants <input type="checkbox"/> Miscellaneous <input checked="" type="checkbox"/> Vertebrates
<a href="#">Matrix filters</a> (only available for vertebrates)	Select matrices associated with the following tissues ( <a href="#">show</a> all tissue associations): <div style="border: 1px solid black; padding: 2px;">             Adipose Tissue              Adrenal Glands              Antibody-Producing Cells              Antigen-Presenting Cells              Blastomeres           </div>
<input type="radio"/> - use <b>all</b> matrices from selected groups <input type="radio"/> - continue with <b>subset</b> definition from selected groups <input type="radio"/> - use <b>previously</b> defined matrix subsets	

## Box 8

The following screen gives you several options for FrameWorker parameters. Change only the “quorum constraints” to 3 of 8 (37%) and check the checkbox “Determine specificity of models”. This is the most stringent setting giving a framework of at least three elements and will bring the most specific model to the top of the following results. Leave the other parameters at their default setting (you can of course play with the settings and see what happens)

**FrameWorker Parameters**

<a href="#">Quorum constraint for framework</a>	Minimum percentage of input sequences to contain a framework: <input type="text" value="3 of 8 (37%)"/> %
<a href="#">Distance constraints for framework</a>	Minimum distance between two elements: <input type="text" value="5"/> (max: 300) Maximum distance between two elements: <input type="text" value="50"/> (max: 300)
<a href="#">Options</a>	<input type="checkbox"/> <b>Show intermediate models</b> (else only the longest model is shown) max. <input type="text" value="10"/> different models per model length (max. 100) <input type="checkbox"/> <b>Show detailed model matches</b> max. <input type="text" value="10"/> matches per model & sequence <input type="checkbox"/> include sequence of the matches <input checked="" type="checkbox"/> <b>Determine specificity of models</b> <input checked="" type="radio"/> Maximum number of elements in models: <input type="text" value="4"/> <input type="radio"/> Check ONLY for unordered pairs of elements
<a href="#">Your email address</a>	<input checked="" type="radio"/> Show result directly in browser window <i>(Note: Do not use this option for long-running jobs! If you should get a server-timeout message, please restart the job using the email option below.)</i> <input type="radio"/> Send the URL of the result to <input type="text" value="seifert@genomatix.de"/>
Result name	
Result name (optional)	<input type="text"/>
<input type="button" value="Start FrameWorker"/> <input type="button" value="Reset the Form"/>	

## Box 9

Scroll past the list of common elements to find the top-scoring framework. It consists of three elements: EBOX-ECAT-ZBPF. Please name the model and save it by clicking the "Save selected models" button at the bottom of the page.

**Model "model\_3el\_1":**

Save this model as

Element	Strand	Matrix sim.	Distance to next element	Common to	FW-Scores
1 V\$EBOX	+	Optimized (min. 0.94)	38 - 43 bp	3 sequences (37 %) 3 matches, 3 non-overlapping	1.00 / 1.00
2 V\$ECAT	-	Optimized (min. 0.93)	9 - 29 bp		
3 V\$ZBPF	-	Optimized (min. 0.87)	---		

**Graphical output:**

Align 100 bp

Save selected models

## Result Background and Explanation

For the analysis the program FrameWorker of the GEMS-Launcher package is used. GEMS Launcher is a software package for DNA analysis. GEMS includes software for transcription factor analysis, discovery of complex regulatory patterns and alignments.

FrameWorker is a software tool allowing the extraction of common motifs (frameworks) of transcription factor binding sites from a set of DNA sequences (in this case promoter sequences). The resulting top-scoring framework consists of three transcription factors: EBOX-ECAT-ZBPF.

The factor SREBF1, a member of the EBOX family was already identified from the BiblioSphere analysis. The framework was identified in the promoter of three of the six genes (HMGCS1, EBP, and DHCR7). Interestingly, while this yields further support for the binding sites already found by cocitations of SREBF1 with HMGCS1 and DHCR7, there had been no evidence for inclusion of EBP from the literature analysis.

## Promoter Database Scan

Subsequent to the definition of a framework, it is possible to scan DNA sequences for matches of such defined transcription factor motifs [4]. One comprehensive approach is to scan databases containing promoter sequences for matches. By this, it is possible to identify potential target genes of defined transcription factor motifs.

In order to find additional genes belonging to the emerging regulatory network governed by the identified framework, we can now scan all human promoters (Genomatix Promoter Database GPD, 50,109 promoters) with the task ModelInspector from GEMS Launcher.

ModelInspector uses a library of predefined models or models defined with FastM or FrameWorker to scan DNA sequences for matches to these models. A model consists of various individual elements (like transcription factor binding sites, repeats, hairpins), their strand orientation, their sequential order, and their distance ranges.

### Box 10

To follow the strategy, from the last step, saving the model the program will take you directly to the ModelInspector task. By clicking the “Start this task button” the program to search the promoter databases will be launched.

To search for matches to your new model you can use the GEMS task:  
**ModelInspector: Search for [user-defined models](#)**

**GEMS Launcher**

[or select one of these databases](#)

**GenBank Release 144 sections:**

Bacteria     Other Vertebrates     Rodents  
 Invertebrates     Plants     Viral  
 Other Mammalian     Primates

**Genomatix promoters:**

Anopheles Promoters     Drosophila Promoters     Plasmodium Promoters  
 Arabidopsis Promoters     Human Promoters (all)     Rat Promoters (all)  
 Chicken Promoters     **Human Promoters (known genes)**     Rat Promoters (known genes)  
 Chimpanzee Promoters     Mouse Promoters (all)     Rice Promoters  
 Dog Promoters     Mouse Promoters (known genes)

**Other databases:**

Anopheles Genome     Dog Genome     Rat Genome  
 Arabidopsis Genome     Drosophila Genome     RefSeq 8.0 (mRNA)  
 Buchers EPD (Rel. 80)     Human Genome     Rice Genome  
 Chicken Genome     Mouse Genome  
 Chimpanzee Genome     Plasmodium Genome

**You can restrict your search to sequences with**

the words \*  in keyword line  
 the words \*  in description line  
 the words \*  in all annotation lines  
\*) words can be separated by spaces or commas

Combine above restrictions with

### Box 11

In the database selection window please check “Human Promoters (known genes)” and proceed with “Load sequences”.

Parameters	
<a href="#">Model groups</a>	Model Group: <b>User-defined models</b> <input type="radio"/> - use all models from this library <input type="radio"/> - use <b>previously</b> defined model subsets <input checked="" type="radio"/> - continue with <b>subset</b> selection
<a href="#">Max. number of matches</a>	<input type="text" value="1000"/> (max: 20000)
<a href="#">Output filter</a>	<input checked="" type="radio"/> Show matches within all sequences <input type="radio"/> Show only a subset of matches based on sequence annotation from feature table: <input type="checkbox"/> sequences annotated as 3'UTR <input type="checkbox"/> sequences annotated as 5'UTR <input type="checkbox"/> sequences annotated as exons <input type="checkbox"/> sequences annotated as introns <input type="checkbox"/> sequences annotated as promoters <input type="checkbox"/> sequences annotated as repeats <input type="checkbox"/> all unannotated sequences
<a href="#">Your email address</a>	<input checked="" type="radio"/> Show result directly in browser window <small>(Note: Do not use this option for long-running jobs! If you should get a server-timeout message, please restart the job using the email option below.)</small> <input type="radio"/> Send the URL of the result to <input type="text" value="user@company.com"/>
Result name	
<a href="#">Result name (optional)</a>	<input type="text"/>
<input type="button" value="Continue"/> <input type="button" value="Reset the Form"/>	

### Box 12

In the following parameters screen leave the parameters default. Only change to “show results directly in browser window”

User-defined: [Tutorial\\_1](#)  
 User-defined: [UCI](#)

Please make sure you selected at least one checkbox!

### Box 13

Select the saved model on the next screen and “Start Task”

## Box 14

The ModelInspector search returns with a Match List giving detailed information about the position of matches in promoters from the searched promoter database (in this case all known human promoters). For the searched model 10 matches are retrieved.

Match List:						
Sequence	Model Name	Position	Strand	Model Score	Select	Match
P501173 (1 - 601) [DNA] sym=STX6 loc=Loc10228 taxid=9606 spec=Homo sapiens ctg=NT_0044487 str=(-) start=31400819 end=31401419 len=601 comm=syntaxin 6;(NM_005819/501/silver)	Tutorial_1	326 - 410	(+)	96.9 %	<input type="checkbox"/>	
P583520 (1 - 605) [DNA] sym=HMGCS1 loc=Loc3157 taxid=9606 spec=Homo sapiens ctg=NT_006576 str=(-) start=43286167 end=43286771 len=605 comm=3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble);(AK095492_1/501/gold;NM_002130/505/silver)	Tutorial_1	297 - 379	(+)	95.2 %	<input type="checkbox"/>	
P529035 (1 - 601) [DNA] sym=IDH2 loc=Loc3418 taxid=9606 spec=Homo sapiens ctg=NT_010274 str=(-) start=5611135 end=5611735 len=601 comm=isocitrate dehydrogenase 2 (NADP+), mitochondrial;(NM_002168/501/bronze)	Tutorial_1	487 - 408	(-)	90.2 %	<input type="checkbox"/>	
P534499 (1 - 601) [DNA] sym=NEED4L loc=Loc23327 taxid=9606 spec=Homo sapiens ctg=NT_025028 str=(+) start=3502144 end=3502744 len=601 comm=neural precursor cell expressed, developmentally down-regulated 4-like;(NM_015277/501/silver)	Tutorial_1	46 - 114	(+)	94.8 %	<input type="checkbox"/>	
P615672 (1 - 601) [DNA] sym=DHCR7 loc=Loc1717 taxid=9606 spec=Homo sapiens ctg=NT_039927 str=(-) start=1382048 end=1382648 len=601 comm=7-dehydrocholesterol reductase;(NM_001360/501/silver)	Tutorial_1	368 - 454	(+)	94.8 %	<input type="checkbox"/>	
P310311 (1 - 630) [DNA] sym=BOPI loc=Loc23246 taxid=9606 spec=Homo sapiens ctg=NT_037704 str=(-) start=82403 end=83032 len=630 comm=block of proliferation 1;(AK024840_1/530/gold;NM_015201/501/silver)	Tutorial_1	518 - 442	(-)	94.8 %	<input type="checkbox"/>	
P583533 (1 - 676) [DNA] sym=HSF1 loc=Loc3297 taxid=9606 spec=Homo sapiens ctg=NT_037704 str=(+) start=82130 end=82805 len=676 comm=heat shock transcription factor 1;(NM_005526/563/silver)	Tutorial_1	386 - 462	(+)	94.8 %	<input type="checkbox"/>	
P588169 (1 - 601) [DNA] sym=EBP loc=Loc10682 taxid=9606 spec=Homo sapiens ctg=NT_037704 str=(+) start=11229566 end=11230166 len=601 comm=enopamil binding protein (sterol isomerase);(NM_006579/501/bronze)	Tutorial_1	326 - 391	(+)	93.4 %	<input type="checkbox"/>	
P501174 (1 - 650) [DNA] sym=STX6 loc=Loc10228 taxid=9606 spec=Homo sapiens ctg=NT_0044487 str=(-) start=31400834 end=31401483 len=650 comm=syntaxin 6;(CompGen promoter, no transcript assigned)	Tutorial_1	390 - 474	(+)	96.9 %	<input type="checkbox"/>	
P613030 (1 - 601) [DNA] sym=LMNA loc=Loc4000 taxid=9606 spec=Homo sapiens ctg=NT_0044487 str=(+) start=6596006 end=6596606 len=601 comm=lamin A/C;(CompGen promoter, no transcript assigned)	Tutorial_1	97 - 21	(-)	91.3 %	<input type="checkbox"/>	

A total of 10 matches was found in 10 sequences.

Sequences searched: 36776 (23930905 bp).

## Evaluation of results

Model: Tutorial\_1

Number of input genes: 9  
Number of genes annotated in GO: 9  
Number of significant GO groups found: 9

GO group	z-score	# genes (observed)	# genes (expected)	list of genes	Locus IDs
steroid metabolism	9.28	3	0.10	HMGCS1, DHCR7, EBP	3157, 1717, 10682
lipid biosynthesis	8.24	3	0.12	HMGCS1, DHCR7, EBP	3157, 1717, 10682
sterol metabolism	13.55	3	0.05	HMGCS1, DHCR7, EBP	3157, 1717, 10682
sterol biosynthesis	21.22	3	0.02	HMGCS1, DHCR7, EBP	3157, 1717, 10682
cholesterol metabolism	14.05	3	0.04	HMGCS1, DHCR7, EBP	3157, 1717, 10682
lipid metabolism	4.45	3	0.37	HMGCS1, DHCR7, EBP	3157, 1717, 10682
alcohol metabolism	7.15	3	0.16	HMGCS1, DHCR7, EBP	3157, 1717, 10682
steroid biosynthesis	13.55	3	0.05	HMGCS1, DHCR7, EBP	3157, 1717, 10682
cholesterol biosynthesis	22.71	3	0.02	HMGCS1, DHCR7, EBP	3157, 1717, 10682

## Extract Sequences

Which matches?	<input type="radio"/> selected matches in above list <input checked="" type="radio"/> all matches
Extent of sequences	<input type="radio"/> complete sequence <input checked="" type="radio"/> match positions ± <input type="text" value="50"/> bp
Output file	<input type="text" value="suite_Seifert.ext"/>

Extract sequence of matches

Extract LocusIDs for BiblioSphere

Export matches to EXCEL format

## Further Evaluation of Matches

Search PubMed for	("gene name") AND ("promoter" OR "transcription factor") (where "gene name" is automatically extracted from the description lines) Note: This works only for annotated genomic DNA sequences from eukaryotes!
Extract gene names	<input type="text"/>

For questions or bug reports, please contact [support@genomatix.de](mailto:support@genomatix.de)

## Box 15

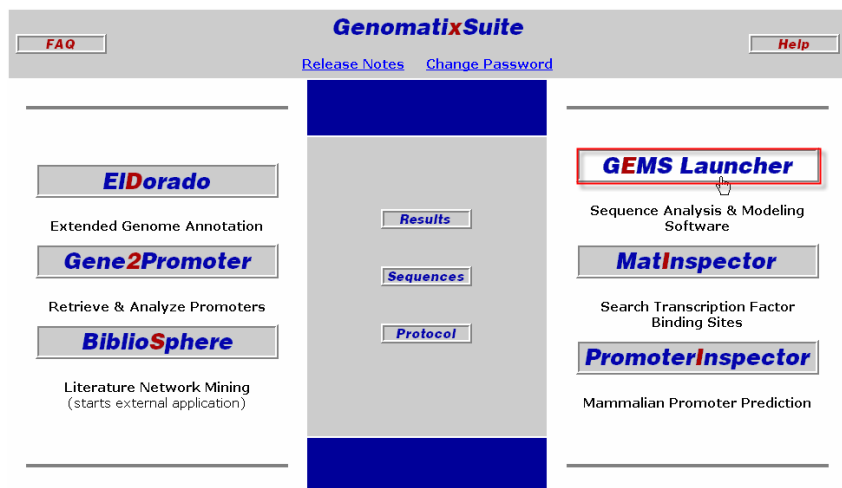
At the bottom of the results page an evaluation of results is carried out for the matched genes by z-scoring. I.e. a z-score is calculated by grouping the matched genes into GeneOntology groups applying the "biological process filter". Only highly significant groups are displayed (z-score > 4.0)

## Result Evaluation 1

The matchlist contains 10 matches for the framework (this is rather specific). Regarding the original functional focus, only the genes which were used in FrameWorker for the construction of the framework were retrieved. Since the distance variability between the single transcription factors influences the selectivity, the framework can be altered to be less selective by increasing the distance variability. This is carried out with the Gems-Launcher task “FastM “Modification/deletion of user-defined models (and subsets)””

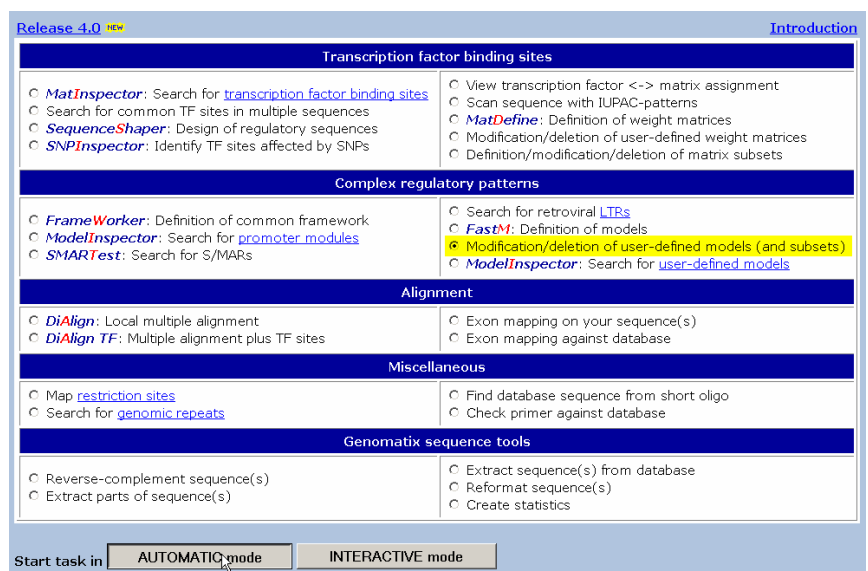
### Box 16

To follow the strategy, start GEMS Launcher from the GenomatixSuite main page



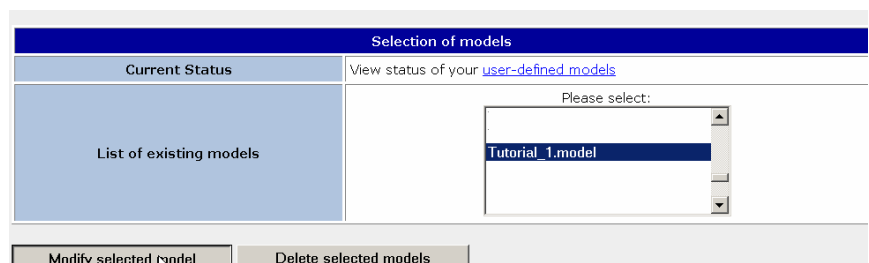
### Box 17

Check “Modification/deletion of user-defined models (and subsets)” and start the program in “AUTOMATIC mode”



### Box 18

Select the model you want to modify from the “List of existing models” and click “Modify selected model”



Box 19  
Click on "Change distance range 1".

Your model "Tutorial_1": Matrix V\$EBOX Matrix V\$ECAT Matrix V\$ZBPF			
	Name	Strand	Quality
Change element 1	Family: V\$EBOX	+	core sim. > 0.75 Optimized matrix sim.
Change distance range 1	38 to 43 bp		
Change element 2	Family: V\$ECAT	-	core sim. > 0.75 Optimized matrix sim.
Change distance range 2	9 to 29 bp		
Change element 3	Family: V\$ZBPF	-	core sim. > 0.75 Optimized matrix sim.

Box 20  
Change distance range to "minimum distance" 5 and "maximum distance" 100. Finish with "Show model"

Distance range definition	
Please select a distance range between element 1 and element 2:	<input type="radio"/> Close range (1 to 20 bp)
	<input type="radio"/> Medium range (30 to 50 bp)
	<input type="radio"/> Long range (100 to 200 bp)
	<input checked="" type="radio"/> Your own distance range: (maximum window size is 2000)
	minimum distance: <input type="text" value="5"/> (min: 0)
	maximum distance: <input type="text" value="100"/> (max: 3000)
<input type="button" value="Show Model"/>	

"Change distance range 2" in the same way (also use 5-100)

Your model "Tutorial_1": Matrix V\$EBOX Matrix V\$ECAT Matrix V\$ZBPF			
	Name	Strand	Quality
Change element 1	Family: V\$EBOX	+	core sim. > 0.75 Optimized matrix sim.
Change distance range 1	5 to 100 bp		
Change element 2	Family: V\$ECAT	-	core sim. > 0.75 Optimized matrix sim.
Change distance range 2	5 to 100 bp		
Change element 3	Family: V\$ZBPF	-	core sim. > 0.75 Optimized matrix sim.

If this is **not** the correct model, please select the corresponding "Change element"-button

You can also  or  your model!

Model parameters	
Element threshold	This value defines the <b>optimized threshold</b> for searching this model <input type="text" value="80"/> %
Model name	Change the name for your model? <input type="text" value="Tutorial_1modified"/>
	Comment to save with your model? <input type="text" value="generated by FrameWorker"/>

Box 21  
"Change the name of your model" and save

Box 22  
Perform the ModelInspector search with the modified model as described on page 13. (Be careful to select the modified model from the list)

To search for matches to your new model you can use the GEMS task:

**ModelInspector: Search for user-defined models**

GEMS Launcher

## Box 23 ModelInspector results. (truncated list)

**Output overview of ModelInspector matches (389 matches)**

go to: [ [Output overview](#) ] [ [Detailed output](#) ] [ [Statistics](#) ]

ModelInspector professional Release 5.1.1 October 2004

**Solution parameters:**

Sequence file: Human Promoters (known genes)  
 Models: User-defined/Tutorial\_1\_mod.model  
 Model score threshold: 80.0 % (abs. 2.40)  
 Output sorted by: match positions on the sequences  
 Maximum number of matches: 1000

**Match List:**

Sequence	Model Name	Position	Strand	Model Score	Select Match
P537935 (1 - 601) [DNA] sym=ARHGFE16 loc=Loc27237 taxid=9606 spec=Homo sapiens ctg=NT_004321 str=(+) start=686397 end=686997 len=601 comm=Rho guanine exchange factor (GEF) 16;(NM_014448/501/bronze)	Tutorial_1_mod	527 - 403	(-)	87.6 %	<input type="checkbox"/>
P589365 (1 - 1073) [DNA] sym=TDRKH loc=Loc11022 taxid=9606 spec=Homo sapiens ctg=NT_004487 str=(-) start=2253218 end=2254290 len=1073 comm=tudor and KH domain containing; (AK056402_1/972/gold;NM_006862/938/bronze)	Tutorial_1_mod	820 - 936	(+)	94.6 %	<input type="checkbox"/>

## Box 24 The Evaluation of results reveals several sterol metabolism associated categories. These categories include additional genes: LSS, MVK, SC5DL and SREBF2

GO group	z-score	# genes (observed)	# genes (expected)	list of genes	Locus IDs
tRNA aminoacylation for protein translation	4.68	4	0.55	DARS, RARSL, C20orf27, FARS1	1615, 57038, 54976, 10667
proline metabolism	5.86	2	0.11	PYCR2, PYCR1	29920, 5831
tricarboxylic acid cycle	4.43	3	0.36	FH, MDH2, IDH2	2271, 4191, 3418
tRNA aminoacylation	4.68	4	0.55	DARS, RARSL, C20orf27, FARS1	1615, 57038, 54976, 10667
cholesterol biosynthesis	5.37	4	0.45	HMGCS1, MVK, DHCR7, EBP	3157, 4598, 1717, 10682
porphyrin biosynthesis	4.43	3	0.36	ALAD, NFE2L1, FECH	210, 4779, 2235
amino acid activation	4.68	4	0.55	DARS, RARSL, C20orf27, FARS1	1615, 57038, 54976, 10667
sterol biosynthesis	6.35	5	0.51	HMGCS1, MVK, SC5DL, DHCR7, EBP	3157, 4598, 6309, 1717, 10682
response to metal ion	4.84	2	0.15	NEDD4L, MTF1	23327, 4520
response to inorganic substance	4.48	2	0.17	NEDD4L, MTF1	23327, 4520
heme biosynthesis	5.49	3	0.26	ALAD, NFE2L1, FECH	210, 4779, 2235
nucleobase, nucleoside, nucleotide and nucleic acid metabolism	4.52	73	45.86	MYOG, POU3F1, DARS, GBX2, SOX11, PTMA, ORC3L, RARSL, CREBL1, CPSF4, FLJ23311, ARNTL, ASCL2, FLJ21415, FANCA, NEUROD2, PPP1R9B, NFE2L1, ATP5G1, HOXB13, JAZ1, TGIF, SMAD2, RUVBL2, NPAS1, ERF, ZNF540, ZSCAN1, ZNF543, FLJ14981, MBD3, POLRMT, FLJ14009, C20orf27, SUHW1, SREBF2, ECGF1, SCML1, ARX, H2AFZ, EXOSC9, HNRPDL, TRIM33, DNMT3A, ELL2, DPYSL2, GDA, POLE, HMG31, CBX8, HEY2, ADCY4, HIF1A, YY1, ATXN3, CHX10, E2F1, HOXC13, ZNF295, JUN, POLD4, FARS1, BARHL1, BOP1, ZNF34, HSF1, COPEB, ELK1, MTF1, POLD2, LDHD, SP2, AMPD2	4656, 5453, 1615, 2637, 6664, 5757, 23595, 57038, 1388, 10898, 79733, 406, 430, 79794, 2175, 4761, 84687, 4779, 516, 10481, 23512, 7050, 4087, 10856, 4861, 2077, 163255, 284312, 125919, 84954, 53615, 5442, 79816, 54976, 129025, 6721, 1890, 6322, 170302, 3015, 5393, 9987, 51592, 1788, 22936, 1808, 9615, 5426, 3146, 57332, 23493, 196883, 3091, 7528, 4287, 338917, 1869, 3229, 49854, 3725, 57804, 10667, 56751, 23246, 80778, 3297, 1316, 2002, 4520, 5425, 197257, 6668, 271
growth hormone secretion	9.58	2	0.04	GAL, LTBP4	51083, 8425
tRNA modification	4.25	4	0.64	DARS, RARSL, C20orf27, FARS1	1615, 57038, 54976, 10667
heme metabolism	4.79	3	0.32	ALAD, NFE2L1, FECH	210, 4779, 2235
sterol metabolism	4.34	6	1.24	HMGCS1, MVK, SREBF2, SC5DL, DHCR7, EBP	3157, 4598, 6721, 6309, 1717, 10682
protein complex assembly, multichaperone pathway	5.29	2	0.13	RUVBL2, SURF1	10856, 6834
peptide hormone secretion	5.86	2	0.11	GAL, LTBP4	51083, 8425
response to unfolded protein	4.47	5	0.87	CREBL1, DNAJA1, HERPUD1, TRA1, HSF1	1388, 3301, 9709, 7184, 3297
proline biosynthesis	7.74	2	0.06	PYCR2, PYCR1	29920, 5831
glutamine biosynthesis	6.63	2	0.09	PYCR2, PYCR1	29920, 5831
steroid biosynthesis	4.34	6	1.24	HMGCS1, MVK, LSS, SC5DL, DHCR7, EBP	3157, 4598, 4047, 6309, 1717, 10682

## Result Evaluation 2

Within the total list of 389 hits in 375 promoters, four categories related to “steroid metabolism” were overrepresented (z-scores ranging from 4.43 - 6.35). In addition to the three genes used to define the framework, LSS, MVK, SC5DL and SREBF2 were identified. LSS is the lanosterol synthetase, which belongs to the same metabolic pathway (synthesis of cholesterol) as the three initial framework genes. LSS is present and up-regulated on the microarray. However, the gene failed to pass the statistical test with SAM and was selected only on molecular evidence. For SREBF2 and MVK the situation is the same as for LSS: they are on the chip and up-regulated but not statistically significant. SREBF2 is the “sterol regulatory element binding transcription factor 2”. Interestingly SREBF2 is an EBOX factor. MVK (mevalonate kinase) is an enzyme also belonging to the cholesterol synthesis chain. SC5DL is the sterol-C5-desaturase which also belongs to sterol metabolism (ergosterol). SC5DL is not on the chip therefore no information about expression change can be retrieved.

In total we have now identified seven genes belonging to the sterol metabolism to be co-regulated by the EBOX-ECAT-ZBPF framework!

### Reiteration of the Framework Analysis

To see whether matches for additional frameworks can be retrieved, the complete group of 10 genes (6 from the literature analysis, four additional from the database search, (categories related to “steroid metabolism”) are subjected to an additional FrameWorker analysis. The ten genes are:

HMGCS1, MVK, SC5DL, DHCR7, EBP, SREBF2, LSS, HMGCR, SC4MOL, DHCR24

#### Box 24

To follow the strategy enter the 10 genes identified by BiblioSphere and sequence analysis into Gene2promoter as described in Box 5 and Box 6. Proceed as described in Box 7 and Box 8.

Set the “Quorum constrains” to at least 8/19. Set the distance constraints to 5-100 bp.

Start the analysis.

#### Box 25

Frameworker will display this error message. This is due to an overlap in promoter sequences in the sequence pairs. Please go back to the Gene2promoter results window (hit your browser’s “Back” button three times) and uncheck one sequence each for the overlapping promoters, e.g. P920801 and P936407. Then restart the framework analysis. For “Quorum constraints” now use 8/17, for “Distance constraints” 5-100.

```
Sorry, but the GEMS Launcher task
"FrameWorker: Definition of common framework"
returned with the following error:
```

```
>
> Sorry, but the sequence pair(s)
>
> P920801 & P920799
> P936407 & P936405
>
> are identical in at least 50 percent of the shorter sequence.
> Please remove one sequence of each pair from your input and start FrameWorker again!
>
```

## Box 26

You will receive a FrameWorker result as displayed. Please save the model and proceed with a Modelinspector search as described in Box 10-13.

**Model "model\_3el\_1":**

Save this model as

Element	Strand	Matrix sim.	Distance to next element	Common to	FW-Scores
1 V\$ECAT	-	Optimized (min. 0.95)	10 - 92 bp	8 sequences (47 %) 41 matches, 8 non-overlapping	0.20 / 1.00 2.40701e-13
2 V\$EGRF	+	Optimized (min. 0.77)	6 - 98 bp		
3 V\$ZBPF	-	Optimized (min. 0.74)	---		

**Graphical output:**

**Sequences searched: 17 (10630 bp).**

Box 27  
Modelinspektor Results.  
total number of matches:  
961

## Output overview of ModelInspector matches (961 matches)

go to: [ [Output overview](#) ] [ [Detailed output](#) ] [ [Statistics](#) ]

ModelInspector professional Release 5.1.1 October 2004

### Solution parameters:

Sequence file: Human Promoters (known genes)  
Models: User-defined/Tutorial\_2.model  
Model score threshold: 80.0 % (abs. 2.40)  
Output sorted by: match positions on the sequences  
Maximum number of matches: 1000

### Match List:

Sequence	Model Name	Position	Strand	Model Score	Select Match
P537842 (1 - 637) [DNA] sym=HSPC150 loc=Loc29089 taxid=9606 spec=Homo sapiens ctg=NT_004487 str=(-) start=52719887 end=52720523 len=637 comm=HSPC150 protein similar to ubiquitin-conjugating enzyme; (AK000504_1/536/gold;NM_014176/501/bronze)	<a href="#">Tutorial_2</a>	<a href="#">335 - 485</a>	(+)	89.8 %	<input type="checkbox"/>
P622046 (1 - 894) [DNA] sym=C1orf22 loc=Loc80267 taxid=9606 spec=Homo sapiens ctg=NT_004487 str=(-) start=35132543 end=35133436 len=894 comm=chromosome 1 open reading frame 22; (AK023095_1/501/gold;NM_025191/794/bronze)	<a href="#">Tutorial_2</a>	<a href="#">407 - 511</a>	(+)	92.0 %	<input type="checkbox"/>

Box 28  
The Evaluation of results reveals significant sterol metabolism associated categories (truncated list).

Please export the whole match list with the “export to excel function” at the bottom of the page

### Evaluation of results

Model: Tutorial\_2

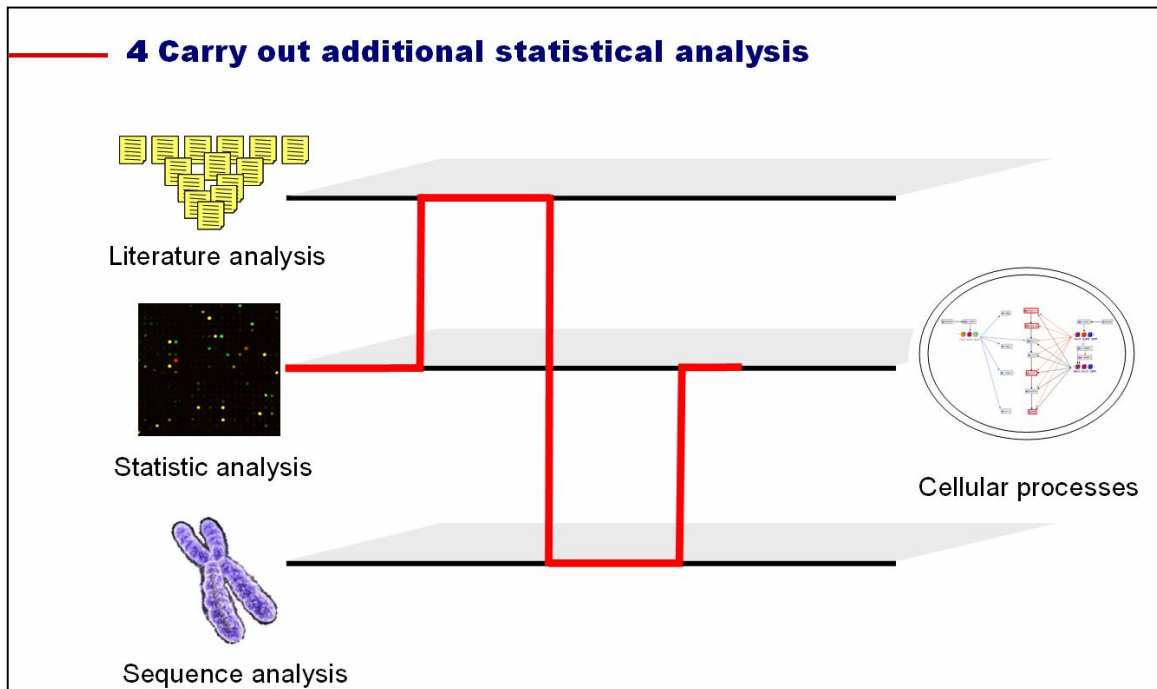
Number of input genes: 828  
Number of genes annotated in GO: 529  
Number of significant GO groups found: 26

GO group	z-score	# genes (observed)	# genes (expected)	list of genes	Locus IDs
steroid metabolism	4.36	16	5.81	HMGCS1, HMGCR, HSD17B8, OPRS1, MVK, STARD5, SREBF1, LSS, SREBF2, SC4MOL, CYP46A1, DHCR24, SC5DL, DHCR7, EBP, FDPS	3157, 3156, 7923, 10280, 4598, 80765, 6720, 4047, 6721, 6307, 10858, 1718, 6309, 1717, 10682, 2224
sterol metabolism	6.25	13	2.81	HMGCS1, HMGCR, OPRS1, MVK, SREBF1, SREBF2, SC4MOL, CYP46A1, DHCR24, SC5DL, DHCR7, EBP, FDPS	3157, 3156, 10280, 4598, 6720, 6721, 6307, 10858, 1718, 6309, 1717, 10682, 2224

## Result Evaluation 3

The second framework analysis results in a single framework consisting of three TFBSs (ECAT, EGRF, ZBPF), matching 8 of 10 genes (HMGCS1, DHCR7, HMGCR, EBP, LSS; MVK, SC5DL, SREBF2). The framework notably does no longer contain the SREBF1 binding site (EBOX). Database analysis (ModelInspector) with this framework yields a total of 961 matches in 828 genes and again the “sterol metabolism” associated categories were overrepresented (among others), this time 16 genes could be significantly related to sterol/steroid metabolism. The genes are falling into the categories associated to “sterol metabolism” are: **CYP46A1**, DHCR24, DHCR7, EBP, **FDPS**, HMGCR, HMGCS1, **HSD17B8**, LSS, MVK, **OPRS1**, SC4MOL, SC5DL, **SREBF1**, SREBF2 and **STARD5** (the 7 new identified genes are marked bold). SREBF1 was already identified as a likely regulator of the first framework and now as a target of the new framework!

## 3.4. Step 4: Additional Statistical Analysis

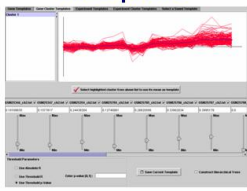


We already learned that some genes (MVK and LSS) which were found by the analysis strategy were not in the original cluster of 105 significantly up regulated genes. However after re-evaluation a sub-statistical up-regulation could be confirmed. Using the framework and its model matches as a complementary line of evidence, it is possible to revisit the statistical analysis with relaxed stringency. To relax the statistics PTM - template matching was used [5].

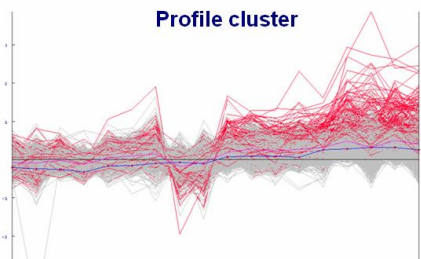
**Clustering by profile of the initially selected 105 genes**

- Expression cluster is extended by Pavlidid Template Matching (PTM)
- Cluster of 105 significant regulated genes is taken as template
- The threshold p-value is 0.1

**Initial profile**



**Profile cluster**



- Cluster is extended to 798 genes (including all 105 initial genes)

The underlying profile was derived from the 105 significantly regulated genes from step 1. 798 genes were retrieved with a threshold p-value of 0.1. This list of 798 genes can be downloaded at:

[http://www.genomatix.de/download/tutorial/PDGF\\_tutorial\\_2.xls](http://www.genomatix.de/download/tutorial/PDGF_tutorial_2.xls)

52 of these genes also were in the match list of the ECAT, EGFR, ZBPF framework (this can e.g. be confirmed by the “Query” function in MS Excel. Alternatively, download the match list at

[http://www.genomatix.de/download/tutorial/PDGF\\_tutorial\\_3.xls](http://www.genomatix.de/download/tutorial/PDGF_tutorial_3.xls)

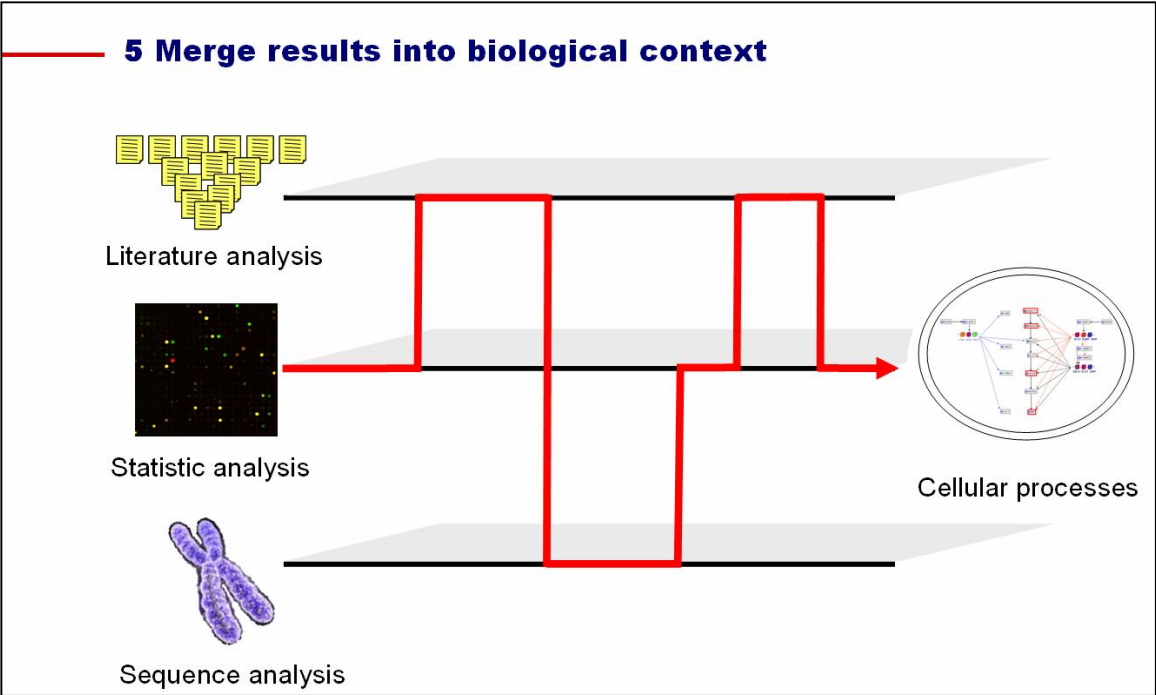
These 52 genes can now be submitted to BiblioSphere (cf. Box 1) and filtered with the GO-Filter “Biological Process”. Eight of the 52 genes belong to the GO-category “steroid metabolism”.

- 52 genes share a common framework and are co-expressed
- 8 genes belong to the GO-category "steroid biosynthesis":  
DHCR24, DHCR7, EBP, HMGCR, HMGCS1, LSS, MVK, SC4MOL

Eight genes are associated with steroid metabolism are supported by three lines of evidence:

1. Common up-regulation
2. Common framework
3. Common functional class (GO-annotation)

3.5 Step 5: Merging of Results Into Biological Context

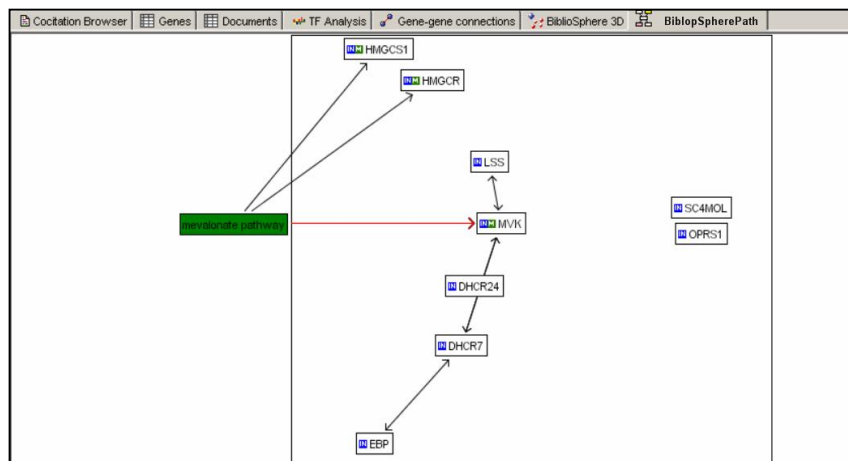


## Pathway Mining Using BiblioSphere Pathway Edition

The 54 genes supported by three lines of evidence were used as input for the BiblioSphere GO analysis. 9 of the 54 genes belong to the category “steroid biosynthesis” with a z-score of 17.51. The biological connection between the six top-scoring genes was analyzed with BiblioSphere Pathway Edition which will be released in March 2005. This reveals that seven genes are part of the mevalonate pathway which is closely linked to the sterol metabolism. SC4MOL and OPRS1 are not directly linked, they belong to the ergosterol synthesis.

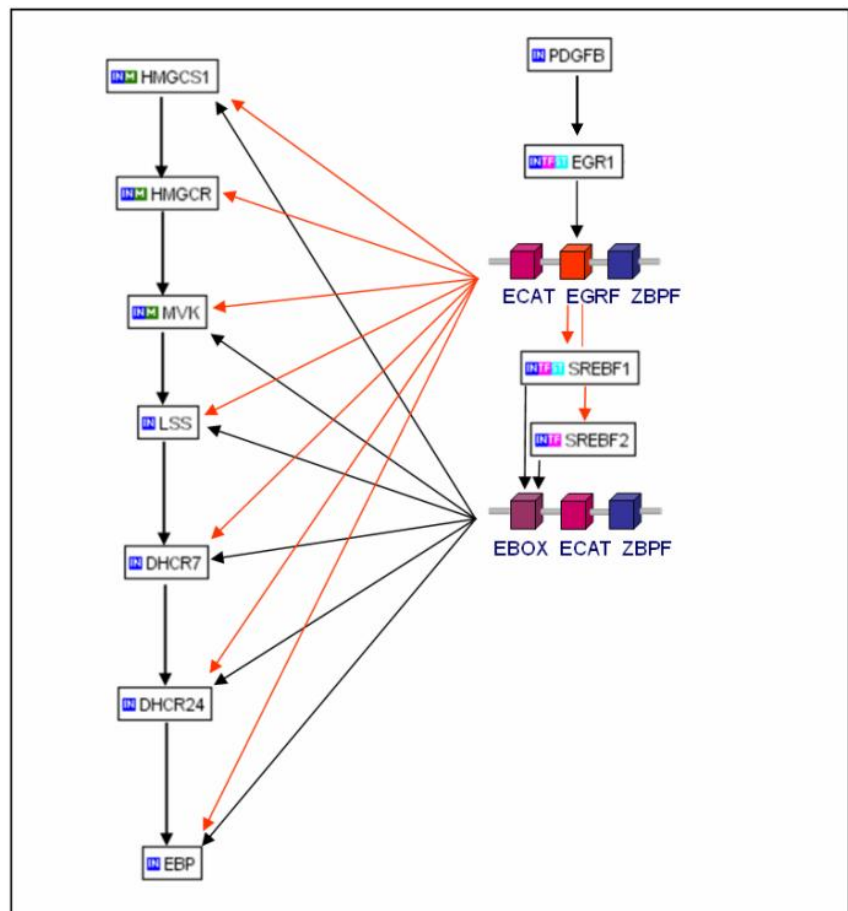
### Box 29

You can already use BiblioSphere for the GO-categorization. Please enter the list of 54 genes into BiblioSphere and proceed as described in Box 1-3. The BiblioSphere Pathway Edition will be released in March 2005.



### Box 30

Merging results into biological context. Displayed are the metabolic steps of the sterol biosynthesis and the referring signalling pathway. The results of the sequence analysis are integrated in this view.



The biological connection between the six top-scoring genes was analyzed with the BiblioSphere pathway mining tool. This reveals that seven genes are part of the mevalonate pathway which is closely linked to the sterol metabolism. SC4MOL and OPRS1 are not directly linked, they belong to the ergosterol synthesis.

## Additional Analysis Cycle: Tubulin Genes

From the literature analysis it is evident, that several more functional groups can be derived from the experimental data. Among others, there is a group of tubulin genes up-regulated under the experimental conditions. Please try the presented strategy and see what you can find out.

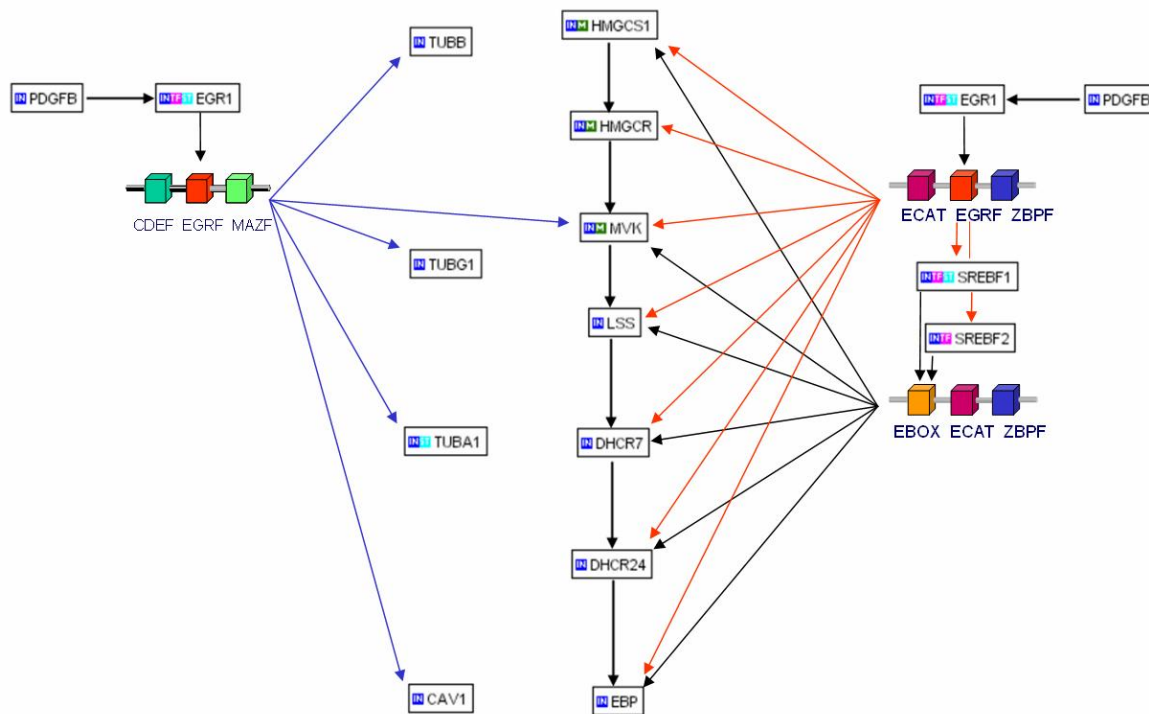
The genes for further analysis are: TUBA1, TUBA3, TUBAB, TUBB, TUBG1

### Box 31

To further follow the analysis go back to the to the BiblioSphere of the 105 original up-regulated genes. Go to the genes table, sort genes alphabetically by clicking on the genes column and select the 5 tubulin genes.

Carry out the subsequent analysis according to the steps from Box 5-15

id	gene	gene name	re...	id...	u...
86	SLC20A1	solute carrier family 20 (phosphate transporter), member 1	NONE	6574	SLC20A1
87	SLC20A2	solute carrier family 20 (phosphate transporter), member 2	NONE	6575	SLC20A2
88	SNRPD1	small nuclear ribonucleoprotein D1 polypeptide 16kDa	NONE	6632	SNRPD1
89	TCTE1	t-complex-associated-testis-expressed 1	NONE	6989	TCTE1
90	TCTE1	t-complex-associated-testis-expressed 1-like 1	NONE	6993	TCTE1
91	TIMP1	tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibi...	NONE	7076	TIMP1
92	TMSB10	thymosin, beta 10	NONE	9169	TMSB10
93	TNFAIP6	tumor necrosis factor, alpha-induced protein 6	NONE	7130	TNFAIP6
94	TOMM34	translocase of outer mitochondrial membrane 34	NONE	10953	TOMM34
95	TUBA1	tubulin, alpha 1 (testis specific)	NONE	7277	TUBA1
96	TUBA3	tubulin, alpha 3	NONE	7846	TUBA3
97	TUBA8	TUBA8	NONE	51807	TUBA8
98	TUBB	tubulin, beta polypeptide	NONE	7280	TUBB
99	TUBG1	tubulin, gamma 1	NONE	7283	TUBG1
100	UCHL3	ubiquitin carboxyl-terminal esterase L3 (ubiquitin thiolesterase)	NONE	7347	UCHL3
101	VCL	vinculin	NONE	7414	VCL
102	VDAC1	voltage-dependent anion channel 1	NONE	7416	VDAC1
103	WDR1	WD repeat domain 1	NONE	9949	WDR1
104	ZYX	zyxin	NONE	7791	ZYX



In the final picture, regulatory links between different functional gene groups can be discovered. Here, cell structure proteins TUBB, TUBA1, TUBG1, and CAV1 (which is involved in tubule formation) can be linked to genes of the sterol metabolism via a framework (CDEF EGRF MAZF) derived from the tubulin genes: There is a match in the promoter of MVK. Notably this framework also contains EGRF. Egr-1 is a transcription factor fitting to EGRF known to be a target of PDGF treatment in several cells [6, 7].

## Summary

The iterative combination of complementary lines of evidence, independent of prior knowledge, enables us to gain new insights about the underlying regulatory networks and cascades of PDGF signaling. Regulatory links between different functional groups, cell structure proteins (tubulins) and genes of the sterol metabolism are discovered.

It becomes clear that only relatively small sub-groups from all co-expressed genes can be related to co-regulation (typically less than 10). This is not a contradiction, since co-expression may be based on a variety of co-regulatory mechanisms and reflects the natural flexibility and complexity of living organisms.

## Literature

- 1 Demoulin, J.B. et al. (2004) Platelet-derived growth factor stimulates membrane lipid synthesis through activation of phosphatidylinositol 3-kinase and sterol regulatory element-binding proteins J Biol Chem 279, 35392-402
- 2 Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response Proc Natl Acad Sci U S A 98, 5116-21
- 3 Dooley, K.A., Millinder, S. and Osborne, T.F. (1998) Sterol regulation of 3-hydroxy-3-methylglutaryl-coenzyme A synthase gene through a direct interaction between sterol regulatory element binding protein and the trimeric CCAAT-binding factor/nuclear factor Y J Biol Chem 273, 1349-56
- 4 Frech, K., Danescu-Mayer, J. and Werner, T. (1997) A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter J Mol Biol 270, 674-87
- 5 Pavlidis, P. and Noble, W.S. (2001) Analysis of strain and regional variation in gene expression in mouse brain Genome Biol 2, RESEARCH0042
- 6 Kaufmann, K. and Thiel, G. (2001) Epidermal growth factor and platelet-derived growth factor induce expression of Egr-1, a zinc finger transcription factor, in human malignant glioma cells J Neurol Sci 189, 83-91
- 7 Hjoberg, J. et al. (2004) Induction of early growth-response factor 1 by platelet-derived growth factor in human airway smooth muscle Am J Physiol Lung Cell Mol Physiol 286, L817-25

Further reading and other tutorials: <http://www.genomatix.de/science/index.html>