

The GenomeMatrix
Data mining from biological databases and data
sources for building integrated functional
genomics information

Diplomarbeit

Berlin, Januar 2005

Vorgelegt von

Reha Yildirimman

Matr.Nr. 172338

Durchgeführt am

Max Planck Institut für molekulare Genetik

Berlin

Eingereicht an

der Technischen Universität Berlin

an der Fakultät III für Prozesswissenschaften

am Institut für Biotechnologie

bei Prof. Dr. H. Görisch

Die selbständige und eigenständige Anfertigung versichert an Eides statt.

Berlin, den

Unterschrift

Danksagung

Mein besonderer Dank geht an Herrn Prof. Dr. H. Görisch der Fakultät III der Technischen Universität Berlin für die wissenschaftliche Betreuung und Ermöglichung der externen Diplomarbeit, welche mir neue Sichtweisen und Perspektiven ermöglichte.

Ich bedanke mich bei Herrn Dr. Schleusener für die sehr hilfreiche Unterstützung und Betreuung der Diplomarbeit. Viele Ideen und Anregungen haben mir sehr geholfen, diese Arbeit umzusetzen.

Großer Dank gebührt Herrn Prof. Dr. Lehrach und Frau Dr. M.L. Yaspo für die Bereitstellung des Themas und für das Ermöglichen dieser Diplomarbeit am Max Planck Institut für Molekulare Genetik. Die Arbeit gewährte mir sehr lehrreiche Einblicke und Erfahrungen im Bereich der Bioinformatik.

Für die sehr freundliche Integration in die Arbeitsgruppe und für die Hilfestellung bei Arbeiten möchte ich mich bei allen Mitarbeitern der Arbeitsgruppe recht herzlich bedanken.

Ich bedanke mich sehr bei Alia Benkahla, Hans-Jörg Warnatz und Marc Sultan für Ihre intensive Unterstützung und umfassende Hilfe.

Im Besonderen bedanke ich mich bei meinen Freunden und v.a. Anne Schiller, Götz Grandpierre und Holger Röbel, die mich stets moralisch unterstützt und bei vielen Sorgen und Problemen geholfen haben.

Dedicated to my parents.

1	INTRODUCTION.....	8
1.1	OVERVIEW: FUNCTIONAL GENOMICS AND COMPUTATIONAL NEEDS	8
1.2	DATA MINING.....	10
1.3	THE GENOMEMATRIX.....	11
1.4	OBJECTIVES OF THE PROJECT.....	12
2	SPECIFICATIONS.....	14
2.1	BASICS OF THE GENOMEMATRIX	14
2.1.1	<i>The Graphical User Interface (GUI)</i>	14
2.1.2	<i>Organism specific matrices</i>	15
2.1.3	<i>Orthology relationships</i>	18
2.1.4	<i>Usage of the GM</i>	19
2.1.4.1	Step 1: Selection of data classes	20
2.1.4.2	Step 2: Arrangement of data classes	21
2.1.4.3	Step 3: Query	21
2.2	METHODOLOGY	22
2.2.1	<i>Basic Local Alignment Search Tool (BLAST)</i>	22
2.2.2	<i>BlastSummary</i>	23
2.2.3	<i>ENSEMBL transcript to gene correlation table</i>	24
2.2.4	<i>ENSEMBL peptide to gene correlation table</i>	24
2.2.5	<i>External IDs correlation table</i>	25
3	DATA MINING FOR THE GENOMEMATRIX.....	27
3.1	OVERVIEW	27
3.1.1	<i>Retrieving data</i>	28
3.1.2	<i>Correlating data</i>	28
3.1.3	<i>Extracting data</i>	28
3.1.4	<i>Creating information files</i>	28
3.2	MAINTENANCE OF DATA CLASSES.....	28
3.2.1	<i>Biomolecular Interaction Network Database (BIND)</i>	28
3.2.1.1	Data source	29
3.2.1.2	Correlation	31
3.2.1.3	Data extraction and creation of all data class files	31
3.2.2	<i>Database of Interacting Proteins (DIP)</i>	32
3.2.2.1	Data source	33
3.2.2.2	Correlation	34
3.2.2.3	Data extraction and creation of all data class files	35
3.2.3	<i>GeneNest</i>	36
3.2.3.1	Data source	37
3.2.3.2	Correlation	37

Index

3.2.3.3	Data extraction and creation of all data class files	38
3.2.4	<i>Induced Mutant Resource (IMR)</i>	39
3.2.4.1	Data source	39
3.2.4.2	Correlation	40
3.2.4.3	Data extraction and creation of all data class files	41
3.2.5	<i>BioMedNet (BMN)</i>	41
3.2.5.1	Data source	42
3.2.5.2	Correlation	43
3.2.5.3	Data extraction and creation of all data class files	44
3.2.6	<i>Bay Genomics GeneTrap</i>	44
3.2.6.1	Data source	45
3.2.6.2	Correlation	46
3.2.6.3	Data extraction and creation of all data class files	46
4	APPLICATION OF THE GM INTO THE C21 GM	48
4.1	THE CHROMOSOME 21 GENOMEMATRIX	48
4.1.1	<i>Concept of the c21 GM</i>	48
4.1.2	<i>The HSA21 database</i>	49
4.2	ALTERATION PROCEDURES	49
4.2.1	<i>Correlation system between GM and c21 GM</i>	49
4.2.2	<i>Backbone tables</i>	50
4.2.3	<i>Orthology relationships</i>	53
4.2.4	<i>Importing GM data classes into the c21 GM</i>	56
4.3	DATA MINING FOR THE C21 GM.....	57
4.3.1	<i>In silico expression</i>	57
4.3.1.1	Data source	57
4.3.1.2	Creation of all data class files	57
4.3.2	<i>Whole mount in situ hybridization</i>	58
4.3.2.1	Data source	58
4.3.2.2	Creation of all data class files	58
4.3.3	<i>Brain sections in situ hybridization</i>	59
4.3.3.1	Data source	59
4.3.3.2	Creation of all data class files	59
5	DISCUSSION	61
5.1	DATA MINING FOR THE GENOMEMATRIX	61
5.2	THE C21 GM - A BIOINFORMATIC TOOL.....	63
5.3	FUTURE PROPOSAL.....	64
6	ABBREVIATIONS	65
7	LIST OF REFERENCES	67
8	APPENDIX	70

Index

8.1	DVD.....	70
8.2	LIST OF DATA SOURCES	70
9	ZUSAMMENFASSUNG	73

1 Introduction

1.1 Overview: functional genomics and computational needs

The complete sequencing of the human genome¹ rather marked the beginning than the end of the largest endeavor in genomic research [1,2,3]. Further exploration of functionally relevant regions of the genome and of all the protein-coding² genes together with the analysis of their function form the next logical steps, creating a new concept called “functional genomics” [4,5,6,7]. The fundamental strategy in a functional genomics approach is to expand the scope of biological investigation from studying single genes to studying all genes at once in a systematic fashion. Functional genomics information emerges from the data generated by high throughput (large-scale) experimental methodologies (e.g. gene expression profiles [9], protein-protein interactions [10,11], analysis of mutants and their phenotypes [12]) applied to many genes at a time. Approaches from different areas of science are required to solve the arising problems of how to extract, classify, assemble, correlate, evaluate, integrate and manage data from these results[13,14]. The application of computer technology creates new ways and possibilities to solve these problems, forming the field of bioinformatics. Bioinformatics will perform a critical and expanding role in functional genomics by providing the tools that assist in bridging the gap between sequence and function [15,16,17,18].

For obvious reasons, many experiments related to gene function cannot be performed on human subjects and are instead carried out on so-called “model organisms” (e.g. fly, worm, yeast) [19,20,21]. During animal evolution, species have diverged from common ancestors through changes in their DNA but still share a significant set of genes that remained very similar in their structure and key functionality. Although there are some exceptions and although a certain level of caution in data interpretation is necessary, it is still possible to apply the concept of “data transferability” from one organism to the other, i.e. the function of a gene discovered in mouse could be similar in human. The

¹ All of the genetic information carried by an organism

² A protein-coding gene consists of a promoter followed by the coding sequence for the protein and then a terminator.

systematic generation of experimental data for genes originating from one organism allows us to increase our knowledge of gene function for similar genes encoded by other organisms and related through orthology³ [22]. Gene expression profiles, protein-protein interactions, or the effects of gene mutants such as knockouts⁴ are just a few examples of experimental data which collectively provide a mass of information useful for the functional analysis of genes, referred to as functional genomics information. A tremendous international effort has already generated a wealth of data aimed at systematic analysis of gene function. This data is provided by various online databases, most of them containing information with the main focus on a specific organism (e.g. mouse) and/or data type (e.g. protein-protein interaction).

Their total number as well as the spectrum of focuses is large and still growing constantly. Despite this unmanageable amount several databases can be regarded as representative for their main focus. A few examples are given below:

The ENSEMBL database [24,25,26] produces and maintains automatic annotation on metazoan⁵ genomes. The system automatically assembles available sequences of a genome into DNA stretches, analyses these to predict genes and provides cDNA and peptide sequences for all predictions. ENSEMBL proposes gene catalogs for human, mouse and other species, on which many other experiments are based.

WormBase [27] and FlyBase [28] are characteristic examples for databases focused on a specific organism. They offer a high amount of different types of data solely limited to one organism.

FlyBase is a model organism database providing genetic and molecular data for *Drosophila*. FlyBase includes data on all species from the family *Drosophilidae*, with a primary focus on *Drosophila melanogaster*. It provides among other things information on genes, mutant alleles, expressions and properties of transcripts and proteins.

³ Orthology: orthologous genes are homologous genes in different species that arose from a single gene in the last common ancestor of these species

⁴ An alteration of a gene that results in loss of function.

⁵ Any animal of the subkingdom *Metazoa*, all animals except protozoans and sponges

WormBase is a model organism database providing a comprehensive data resource for *Caenorhabditis* biology and genomics. Some types of available data are sequences, protein-protein interactions and gene expressions.

BIND and DIP are examples of databases with a main focus on interaction information, whereas BIND provides information on interactions between diverse types of objects and DIP provides information on interactions solely between proteins.

The Biomolecular Interaction Network Database (BIND) is a collection of records providing molecular interactions information. A BIND record represents an interaction between two or more biological objects that is believed to occur in a living organism. A biological object can be a protein, DNA, RNA, ligand⁶, molecular complex, gene, photon or an unclassified biological entity. [29]

The Database of Interacting Proteins (DIP) catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The provided data is curated manually and automatically using computational approaches. [30]

The value of a database for functional genomics research is influenced by the type, amount and quality of integrated data, by the accessibility provided to the user and finally by the system functionality. Problems of access, overview and usability arise whenever the functional analysis of one or more genes requires information distributed across multiple such sources. A basic approach to solve this problem is to extract data from the sources of interest for further processing. [31] The access to extracted data in contrast to the access via a database allows further processing techniques such as data mining.

1.2 Data mining

⁶ A small molecule that binds to a receptor/protein.

The term data mining [32,33] is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. In the context of functional genomics such an extraction means the correlation of data to a gene in order to enrich the knowledge about gene function. Correlation is done by identifying relationships or common patterns between these sets of data, either directly or via additional data. The applied method depends on the type of data used. All methods applied are listed and described in more detail in chapter 2.2.

The process of data mining was used throughout this thesis to correlate data extracted from databases or data sources to either human or mouse genes. These correlations were integrated into the GenomeMatrix database.

1.3 The GenomeMatrix

The GenomeMatrix (GM) is a newly developed online database system, providing access to biological data on genes, their products and their functions from a wide range of data sources (see chapter 8.2) via a special graphical user interface. This mere collection of various, freely accessible data distributed widely over multiple locations simplifies investigations. The GM focuses primarily on completely sequenced model organisms and uses the most up to date gene catalog available for an organism (e.g. ENSEMBL in human). All data integrated into the GM is associated to genes of these organisms. Further relationships between data from different organisms through orthology provide additional useful information and functionality. This system layout provides the possibility of gene specific queries on data independent from the original data source build-up, classification, functionality or usability.

The graphical user interface allows the display of diverse levels of data and their relationships in a simple, efficient way. Data is displayed via a matrix of colored boxes, wherein columns represent different genes and rows represent the different types of information related to the genes. The information is, as far as possible, encoded by the color of the individual box. The color indicates the mere existence of an information, but can also encode the information itself (e.g. molecular function of a gene product, expression level of a gene), if the information is not too complex to be expressed via a color scale (e.g. a picture of a mouse carrying a mutation in this gene, a protein structure). Every information indicated by color in the matrix is accessible by clicking on the colored box. The interface is capable of displaying data on a large number of

genes and types of information in parallel. This selection can be adjusted in number, type and arrangement to fit any desired needs.

This pool of information combined with a powerful visualisation technique provides a useful tool for any type of research and is the frame of reference for this thesis. The GM is produced by the joint venture of the MPI-MG⁷ and RZPD⁸. It is publicly accessible under <http://www.genomematrix.org/>.

1.4 Objectives of the project

The project objectives were focused on building integrated functional genomics information in the context of the GM.

The first objective aimed at data mining from specific data sources. Many different sets of data had to be integrated into the GM and maintained over the project's duration in terms of updates. Data from these sources had to be extracted, analyzed and processed to comply with structural and textual specifications of the GenomeMatrix (chapter **2.1**). This task required the design of data mining procedures for each data source as well as the development of programs to execute necessary assignments.

The purpose of this data integration was to enable access to data from these sources via the GM system. All aspects of this objective are described in more detail in chapter 3.

The second objective was the adaptation of the GM to create a new system focused specifically on human chromosome 21, further referred to as c21 GM. A manually curated gene catalog was used for this chromosome. Additional to this modification, private data sets containing systematic functional genomics data had to be integrated into the c21 GM. This task required the development of several strategies and methods. Core elements of the GM had to be modified to render the system functional under the new conditions, and procedures for data mining had to be altered to suit the new specifications. GM data had to be integrated into the c21 GM according to these

⁷ MPI-MG: Max Planck Institute for molecular genetics

⁸ RZPD: Deutsches Ressourcenzentrum für Genomforschung GmbH

specifications. Furthermore, programs were generated to assist the execution of necessary assignments.

The purpose of these modifications is to enhance the quality of gene definitions inside the c21 GM. The new gene catalog defines genes with a higher accuracy compared to the gene catalog used in the GM (chapter 3). Given that all data integrated is related to genes, an increase in the quality of the gene prediction raises the quality of data relationships and thus the overall content quality of the system. The resulting c21 GM is equivalent to the GM in terms of appearance, functionality and usability, but differs in content quantity and quality. All aspects of this objective are described in more detail in chapter 4.

2 Specifications

2.1 Basics of the GenomeMatrix

In the following, the key elements of the GM systematics are described to illustrate the context of performed assignments in this thesis.

2.1.1 The Graphical User Interface (GUI)

The GUI [34] of the GM displays genes and corresponding information as a matrix of boxes (see Fig.1). Columns represent different genes, rows represent different types of information - further called data classes - related to the genes. The list of data classes can be selected and ordered without restrictions, whereas the list of genes can be adjusted in a few ways. A single gene or multiple genes of interest can be chosen for display. Multiple genes are either a number of specific genes in any arbitrary order or one gene in its chromosomal context, in which the gene of interest is arranged with an adjustable number of surrounding genes upstream and downstream of the chromosome. Additional to any type of selection and arrangement, the first row is reserved to represent general information on the selected list of genes.

An empty box indicates unavailable information, a colored box indicates available information. Color is an essential carrier of information in the GM. Depending on the data class a colored box indicates either the mere existence of information via one specific color, or the existence and the quality of information via a specific range of colors. Furthermore a colored box provides direct access to the information it indicates whenever a user accesses it through a mouse click. Details on the type of content and display are listed in chapter 3.

On the top of each column is a label stating the gene name or alternatively the gene identifier associated to the column. Besides each row is a label providing the abbreviation for the organism (e.g. Hs for *Homo sapiens*) and the data class name associated to the row. Each row label serves as a link and by clicking on the label a new window opens, displaying information about the data class.

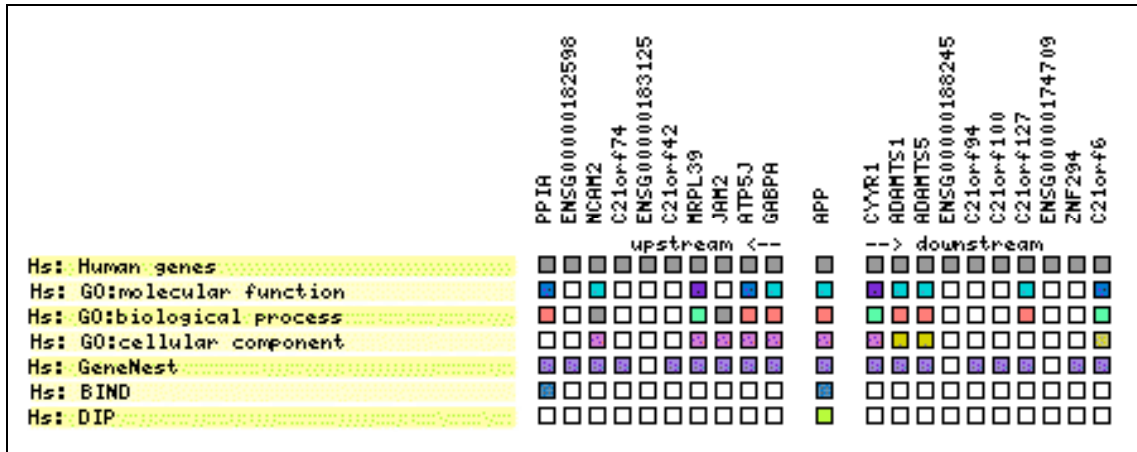


Fig. 1: A GUI display for a range of genes and data classes. In the row for the data class “Hs: BIND“ two boxes colored in blue indicate the mere existence of BIND information for the genes PPIA and APP. In the row of the data class “GO: molecular function”, different colored boxes can be found, each indicating the existence of molecular function information for the genes PPIA, NCAM2, MRPL39, ATP5J, GABPA, APP, CYR1, ADAMTS1, ADAMTS5, C21orf127, and C21orf6. Each color is unique and defines a specific molecular function. The associations between color and molecular function are stated in the information about the data class, accessible by clicking on the row label “GO: molecular function.”

2.1.2 Organism specific matrices

Each organism, and all data related to it, is represented by one organism specific matrix. The core of each matrix is a backbone table (see Tab. 1). It is constructed from the genomic sequence of the associated organism and resembles its gene catalog. Each row of a backbone table represents one gene and states the following eight values:

1. Gene identifier
2. Chromosome number
3. Chromosomal start position
4. Chromosomal end position
5. Value for red of RGB⁹
6. Value for green of RGB
7. Value for blue of RGB
8. Location of a HTML¹⁰ file.

ENSG00000085465,1,111255673,111269126,153,153,153,human/18_34/gene_summary/ENSG00000085465.card.HTML.gz

⁹ RGB: Red-Green-Blue color system based on values ranging from 0 to 255. The combined values define one color.

¹⁰ HTML: HyperText Markup Language used for creating hypertext documents on the World Wide Web.

Specifications

ENSG00000085491,1,108032123,108097413,153,153,153,human/18_34/gene_summary/ENSG00000085491.card.HTML.gz
 ENSG00000085511,6,161322251,161447697,153,153,153,human/18_34/gene_summary/ENSG00000085511.card.HTML.gz

Tab. 1: A section of the human backbone table showing 3 rows. Each row represents one human gene and provides information on gene identifier, chromosome number, start, and end position as well as values defining the RGB color of the associated GUI box for this gene. The last value states the location and name of the data file associated to this gene.

The gene identifier is a value associated to the source of the used genomic sequence and labels the gene. Values 2 to 4 describe the position of the gene on a chromosome. Values 5 to 7 define the RGB color of the box in the first row of GUI that represents general information on this gene. Value 8 defines the location of a HTML file containing the general information on the gene. This file is loaded when the RGB colored box is clicked.

GM relies on completely sequenced model organisms. At the beginning of the project matrices for five organisms were available (see Tab. 2).

Organism	Source of genomic sequence	Matrix name
<i>Homo sapiens</i>	ENSEMBL	Hs
<i>Mus musculus</i>	ENSEMBL	Mm
<i>Caenorhabditis elegans</i>	WormBase	Ce
<i>Drosophila melanogaster</i>	FlyBase	Dm
<i>Saccharomyces cerevisiae</i>	SGD [35]	Sc

Tab. 2 Overview of genomic sequence sources for available organisms in the GM as well as the corresponding matrix name.

The actual data content of a matrix is stored in data classes. A data class comprises all data from one source (e.g. a database) that has been related to the genes of an organism. Every data class consists of three types of files:

1. A reference file, further called identifier file (see Tab. 3). It contains the following five values separated by a comma:

1. Gene identifier
2. Value for red in the RGB
3. Value for green in the RGB
4. Value for blue in the RGB
5. Location of a data file.

```
ENSG00000188170,70,130,180,human/21_34/BIND/ENSG00000188170.BIND.txt.gz,
ENSG00000117601,70,130,180,human/21_34/BIND/ENSG00000117601.BIND.txt.gz,
ENSG00000180210,70,130,180,human/21_34/BIND/ENSG00000180210.BIND.txt.gz,
```

Tab. 3: A section of the BIND data class identifier file, showing 3 rows. Each row states a gene identifier, 3 values defining the RGB color of the associated GUI box and the location and name of the data file containing all BIND information associated to the gene identifier.

Each row in the identifier file represents the association of all data class information to one gene and is represented by a colored box in the GUI. The gene identifier is a value referring to one gene in the backbone table. Values 2 to 4 define the RGB color of the box in the GUI. Value 5 defines the location of a data file containing details about the data class information for the gene. This file is loaded when the colored box is clicked.

2. The data files (see Tab. 4), each containing all data class information related to one gene, a short description of the data class as well as the data class name. It includes HTML tags in order to enable a proper visual display of the content. The content of a data class file is inserted into a blank HTML file whenever the colored box of the associated data class and gene is clicked. A data file is labeled according to the associated gene and data class, i.e. the data file ENSG00000188170.BIND.txt contains all information from the data class BIND associated to the gene ENSG00000188170.

```
class: Link to the BIND Database
descr: The human BIND non-redundant protein sequence set (2255 entries), were linked to the ENSEMBL ids using BLASTP.
<TABLE BORDER="1" ALIGN="CENTER">
<TR>
<TH>ENSEMBL gene entry</TH>
<TH>BIND Protein Interactions</TH>
</TR>
<TR>
<TD>ENSG00000188170</TD>
```

```
<TD><A href="http://www.bind.ca/cgi-bin/bind/dataget?get=search&rectype=4&type=int&id=121062">121062</A></TD>
</TR>
<TR>
<TD>ENSG00000188170</TD>
<TD><A href="http://www.bind.ca/cgi-bin/bind/dataget?get=search&rectype=4&type=int&id=121035">121035</A></TD>
</TR>
```

Tab. 4 A section of a BIND data file showing BIND information linked to the human gene ENSG00000188170. HTML tags are included beside the BIND information for proper further display.

3. An information file, further called info file. It provides general information about the data class:

1. Short description of the data class
2. Source of the downloaded files
3. Downloading date
4. General statistic about the data class

The info file of a data class is read and inserted into a blank HTML file, when the label of a data class in the GUI is clicked. It includes HTML tags in order to enable a proper visual display of the values 1 to 4.

Example: The GM is searched for all available information regarding the human gene ENSG00000188170. The identifier file of each data class is searched for a match between a gene identifier and ENSG00000188170. If a match is found, the row of the identifier file is read. The box for ENSG00000188170 in the GUI is colored according to the RGB values defined in the row. If the colored box is clicked, the data file for ENSG00000188170 is loaded from the location defined in the row and displayed.

2.1.3 Orthology relationships

Orthology in the GM was computed using the InParanoid software [36] with standard parameters. It is based on pairwise similarity BLAST¹¹ [37] scores (see chapter 2.2.1). InParanoid first detects the mutual best hits between sequences from two different

¹¹ BLAST: Basic Local Alignment and Search Tool

species. These so-called main orthologs form the core of an orthologous group. Furthermore, other sequences are added to this group if they are closely related to one of these main orthologs. These members of the orthologous group are called in-paralogs. A confidence value (from 100% down to 5%) is provided for each in-paralog that shows how closely related it is to the main ortholog.

The computed orthology relationships between genes of two organisms are stored in one table (see Tab. 5), each row stating a gene identifier, the identifier of the corresponding ortholog and the confidence value defined by the InParanoid program.

ENSG00000105221,ENSMUSG00000004056,100
ENSG00000105223,ENSMUSG00000003363,100
ENSG00000105227,ENSMUSG00000001951,100

Tab. 5 A section of the orthology relationship table for human and mouse showing 3 rows. Each row states an orthology relationship between a human gene identifier to a mouse gene identifier of ENSEMBL, and the confidence value of this relationship.

Using these relationships between genes from different organisms, data in a matrix can be associated to data in the other matrices.

2.1.4 Usage of the GM

After selecting an organism of interest, different parameters can be selected in three steps to define and access the GUI.

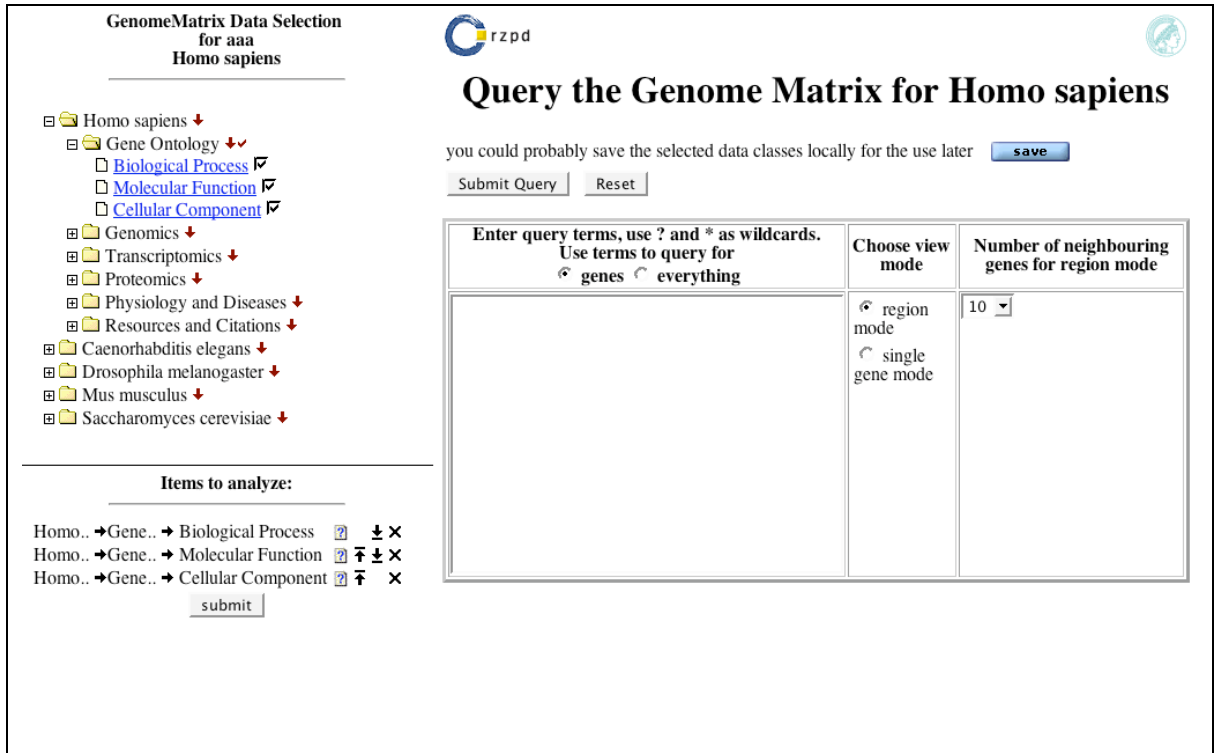


Fig. 2: A screenshot of the GM website displaying all areas of the online GM selection interface. In the top left part data classes can be selected for display in the GUI, in the bottom left part selected data classes can be arranged according to the desired arrangement in the GUI. The right part show the query interface.

2.1.4.1 Step 1: Selection of data classes

The selection of data classes is possible via a tree-like folder system in the upper left part of the website (see Fig. 2). Folders are used to group related content and to simplify overview. The top level consists of folders, each representing an organism specific matrix. Only matrices with existing orthology relationships to the organism of interest are available. Below each matrix folder is the second level of folders, each representing a group of content related data classes available in the matrix. The following types of groups are displayed:

1. Gene ontology
2. Genomics
3. Transcriptomics
4. Proteomics
5. Physiology and diseases
6. Resources and citations.

Below each group folder is a list of data classes available in this group.

Fig. 3 A GUI display resulting from the following parameters. Homo sapiens was the organism of interest. From the human matrix the data classes BIND, DIP, GeneNest [38,39] and the data class group GeneOntology[40] were selected. From the mouse matrix the data class group GeneOntology and the data classes Bay Genomics GeneTraps[41], Induces Mutants, Mouse Knockout Mutants, Bind and DIP were selected. The GM was queried for the human gene DSCAM in its chromosomal context, showing 20 genes upstream and downstream the chromosome.

2.2 Methodology

The following methods are frequently used for many assignments throughout this thesis and range from programs to specific correlation tables.

2.2.1 Basic Local Alignment Search Tool (BLAST)

BLAST [37] is a set of similarity search programs used to identify sequences that share similarity to a query sequence. It has been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm that seeks local as opposed to global alignments and is therefore able to detect relationships among sequences that share only isolated regions of similarity.

The BLASTP 2.2.1 program was used to compare protein sequences. It was set to keep low complexity¹² results and filter out low quality matches while using default parameters for remaining options. The thresholds used to filter the BLAST output are:

1. a score larger than 100
2. an identity percentage larger than 98%

The BLASTN 2.2.1 program was used to compare nucleotide sequences. It was set to filter out low quality matches while using default parameters for remaining options.

1. a score larger then 100
2. an identity percentage larger than 96%

¹² A region of protein sequence enriched for a single amino acid.

The result of a comparison made with any BLAST program is further referred to as BLAST result. The use of either BLAST program to compare two sequence files with each other is further referred to as blasting or to blast. Independently from the type of BLAST program used the smaller file is always blasted against the larger file. Every BLAST was applied using a cluster of computers to accelerate processing.

2.2.2 BlastSummary

BLAST result files can be very large and contain a high amount of information useless for further processing in the GM. BlastSummary is a PERL¹³ script used to extract specific information out of a BLAST result file and store these in a file – further referred to as RES file (see Tab. 6). The description of each BLAST match corresponds to one line in the RES file. In the case of blasting database A against database B, the RES file contains the following information:

1. Identifier a from database A
2. Identifier a length
3. Match start position in a
4. Match end position in a
5. Identifier b from database B
6. Identifier b length
7. Match start position in b
8. Match end position in b
9. Score of the match between a and b
10. E-value of the match between a and b
11. Identity percentage of the match between a and b.

ENST00000326632 1407 38 1407 Hs425208.1 1968 83 1776 1237 0.0 98%
ENST00000326632 1407 38 1407 Hs457620.2 1994 11 1946 1227 0.0 98%
ENST00000326632 1407 1 1407 Hs349333.1 2359 333 2260 1225 0.0 97%

Tab. 6 A section of a RES file showing 3 rows of summarized BLAST matches between ENSEMBL transcript sequences and GeneNest sequences.

¹³ PERL: Practical Extraction and Report Language. A programming language.

The resulting RES file is much smaller in file size than the BLAST result file and enables easier and faster further processing of BLAST matches. The BlastSummary script is applied on every BLAST result in every part of this thesis and will not be additionally mentioned.

2.2.3 ENSEMBL transcript to gene correlation table

The following correlation table (see Tab. 7) provides the correlations between an ENSEMBL transcript identifier and the corresponding ENSEMBL gene identifier.

ENST00000000233	ENSG00000004059
ENST00000000412	ENSG00000003056
ENST00000000442	ENSG00000173153

Tab. 7 A section of the ENSEMBL transcript to gene correlation table showing 3 rows. Each row states the relationship between an ENSEMBL transcript and an ENSEMBL gene identifier.

A FASTA¹⁴ file containing all ENSEMBL transcript sequences is available for human and mouse at the ENSEMBL ftp server and provides this correlation via the sequence headers (see Tab. 8).

>Transcript: ENST00000000233 Database:core Gene: ENSG00000004059 Clone:AC000357 Contig:AC000357.1.1.45309 Chr:7 Basepair:126782948 Status:known
--

Tab. 8 The header of the ENSEMBL transcript sequence for ENST00000000233.

All transcript identifiers and the corresponding gene identifiers were extracted from each header of the human and mouse FASTA file and stored in the ENST-ENSG.corr file for human and in ENSMUST-ENSMUSG.corr for mouse. Every further reference to the human and mouse correlation tables is done via these file names. The downloaded FASTA files are based on the same ENSEMBL release as the ENSEMBL genomic sequence used for the human and the mouse backbone table, respectively.

2.2.4 ENSEMBL peptide to gene correlation table

¹⁴ A sequence in FASTA format begins with a single-line description, followed by lines of sequence data.

The following correlation table (see Tab. 9) provides the correlation between an ENSEMBL peptide identifier and the corresponding ENSEMBL gene identifier.

ENSP0000000233	ENSG0000004059
ENSP0000000412	ENSG0000003056
ENSP0000000442	ENSG00000173153

Tab. 9 Section of the ENSEMBL peptide to gene correlation table showing 3 rows. Each row states the relationship between an ENSEMBL protein and an ENSEMBL gene identifier.

A FASTA file containing all ENSEMBL protein sequences is available for human and mouse at the ENSEMBL ftp server and provides this correlation via the sequence headers (see Tab.10).

>Translation:ENSP0000000233	Database:core	Gene:ENSG0000004059	Clone:AC000357
Contig:AC000357.1.1.45309	Chr:7	Basepair:126782948	Status:known

Tab. 10 The header of the ENSEMBL peptide sequence for ENSP0000000233.

All peptide identifiers and the corresponding gene identifiers were extracted from each header of the human and mouse FASTA file and stored in the ENSP-ENSG.corr file for human and ENSMUSP-ENSMUSG.corr file for mouse. Every further reference to the human and mouse correlation tables is done via these file names. The downloaded FASTA files are based on the same ENSEMBL release as the ENSEMBL genomic sequence used for the human and the mouse backbone table, respectively.

2.2.5 External IDs correlation table

This table (see Tab. 11) provides the correlation between ENSEMBL gene identifiers and several identifiers from other resources. Each row states the following values, separated by comma:

1. ENSEMBL gene identifier
2. Type of associated external identifier
3. External identifier

ENSMUSG00000025921,SPTREMBL,Q8VCH7,

Specifications

ENSMUSG00000025921,EMBL,BC019796, ENSMUSG00000025921,protein_id,AAH19796.1,
--

Tab. 11 Section of the external IDs correlation table showing 3 rows. Each row states the relationship between an ENSEMBL gene identifier and an external identifier.

The table is provided by the RZPD for every ENSEMBL release applied in the GM.

3 Data mining for the GenomeMatrix

3.1 Overview

The objective in this part of the project was the integration of data from specific sources as data classes into the GM for the organisms *Homo sapiens* and *Mus musculus*.

Data sources for *Homo sapiens*:

1. BIND [29]
2. DIP [30]
3. GeneNest [38,39]

Data sources for *Mus musculus*:

1. BIND
2. DIP
3. GeneNest
4. IMR [42]
5. BioMedNet [43]
6. Bay Genomics GeneTraps [41]

Integration of data into the GM implies several steps:

1. Retrieving data
2. Correlating data
3. Extracting data
4. Creating information files

Over the project's period of time, new versions of these data sources were released on an irregular basis, requiring repetition of data integration for these data sources. A new release of the human or mouse genomic sequence by ENSEMBL required a reintegration of all data sources into the organism's matrix.

The accompanying DVD contains all generated PERL scripts, source files and the resulting data classes for every performed data integration.

3.1.1 Retrieving data

Retrieving data from a data source can result in a complex task itself. Some data sources provide their content as downloadable files, in other cases data has to be manually extracted. Missing documentation, insufficient naming convention or the alteration of data access or structure are just a few examples of possible problems.

3.1.2 Correlating data

Implementing data into the GM means classifying which data is related to which gene. The most important factor here is the type of data provided, which affects the appropriate correlation method.

Some data sources provide sequences as main information. Correlation is done by blasting the sequences against the transcript, protein or the genomic sequences. A BLAST result describes the correlation of a sequence and thus all data associated to this sequence with a gene. Other data sources provide relationships between their data and external identifiers. In this case correlation is done mainly by using the external IDs correlation table to relate data from these sources to genes.

3.1.3 Extracting data

After correlating data to a gene the corresponding description is extracted from the data source and stored in a data file. Extracting data is done via PERL scripts developed for each data class.

3.1.4 Creating information files

The identifier file and the info file are needed beside the data files to form a data class. The info file is created manually and contains general information about the data class. The identifier file is created automatically by the same PERL script used for extracting data and creating all data files.

3.2 Maintenance of data classes

3.2.1 Biomolecular Interaction Network Database (BIND)

The BIND is a collection of records providing molecular interactions information [29]. A BIND record represents an interaction between two or more biological objects that is believed to occur in a living organism. A biological object can be a protein, DNA, RNA, ligand, molecular complex, gene, photon or an unclassified biological entity. The contents of BIND include high-throughput data submissions and hand-curated information gathered from the scientific literature.

In the GM, the BIND data class is available for the organisms *Homo sapiens* (see Tab. 12) and *Mus musculus* (see Tab. 13). The procedure described below for creating the BIND data class is identical for both organisms. The differences between the human and the mouse BIND data class arise from the use of organism specific source files. Mouse source files are correlated to genes in the *Mus musculus* matrix, human source files are correlated to genes in the *Homo sapiens* matrix.

ENSEMBL version	Number of BIND entries	Number of matched genes
9.30a.1	115	90
11.31.1	115	85
18.34.1	407	355
21.34	2255	1340

Tab. 12 Overview of the results for each performed integration of human BIND data. The number of BIND entries, the version of the ENSEMBL release used for the human gene catalog and the number of human ENSEMBL genes matched to BIND entries are stated.

ENSEMBL version	Number of BIND entries	Number of matched genes
9.30a.1	295	125
11.31.1	141	125
17.30.1	170	152
21.32	800	513

Tab. 13 Overview of the results for each performed integration of mouse BIND data. The number of BIND entries, the version of the ENSEMBL release used for the mouse gene catalog and the number of mouse ENSEMBL genes matched to BIND entries are stated.

3.2.1.1 Data source

Two files were downloaded from the BIND ftp server per organism:

1. An organism specific sequence file containing FASTA formatted protein sequences pertaining to proteins involved in protein-protein interactions (see Tab. 14)
2. An organism specific ASN.1¹⁵ ASCII¹⁶ file containing the description of the interactor(s) of the proteins that are in file 1 and the description of the interaction (see Tab. 15)

Interactions between human/human and mouse/mouse proteins, respectively, are to be integrated into the BIND data class of the corresponding organism. The protein sequences in file 1 resemble the information necessary to correlate the interactions information in file 2 to genes in the GM.

```
>gi4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVVNDEVGGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHG
KVLGAFSDGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQK
VVAGVANALAHKYH
```

Tab. 14 FASTA formatted BIND protein sequence for gi4504349. The gene identifier is marked in red.

Each header of a protein sequence contains a gene identifier (**gi**). All bio-molecular interaction information in file 2 is connected to such a gene identifier and thus to a protein sequence.

```
a {
  short-label "HBB" ,
  id
  protein {
    gi 4504349 ,
    di 0 ,
    other {
      general {
```

¹⁵ Abstract Syntax Notation

¹⁶ American Standard Code for Information Interchange

```
db "LocusLink" ,
tag
id 3043 } ,
```

Tab. 15 Section of the ANSI.1 ASCII BIND file. The gene identifier is highlighted in red.

3.2.1.2 Correlation

The organism specific protein sequences in file 1 were blasted against the appropriate organism protein sequences from ENSEMBL using BLASTP. The best match of a BIND protein to an ENSEMBL protein was considered as the resulting correlation (see Tab. 16). ENSP-ENSG.corr reveals the ENSEMBL gene associated to an ENSEMBL peptide, its correlated BIND protein, and thus to all information about bio-molecular interactions involving the BIND protein.

```
gi|4504349|ref|NP_000509.1| 147 1 147 ENSP00000333994 147 1 147 305 1e-83 100%
gi|4502261|ref|NP_000479.1| 464 1 464 ENSP00000236260 464 1 464 926 0.0 100%
gi|4502261|ref|NP_000479.1| 464 1 464 ENSP00000307953 260 1 260 395 e-110 99%
```

Tab. 16 Section of RES file for BIND showing 3 rows of BLAST matches summarized via the BlastSummary script.

Example: The identifier gi4502261 correlates to the ENSEMBL peptide ENSP00000236260 instead of ENSP00000307953 because gi4502261 and ENSP00000236260 share a score higher than gi4502261 and ENSP00000307953. All information about bio-molecular interactions involving gi4502261 in BIND is implicitly related to ENSP00000236260.

3.2.1.3 Data extraction and creation of all data class files

A PERL script was generated in order to create all BIND data files and the BIND identifier file. Four parameters are necessary to run this script:

1. Location of the file 2 (containing the interactions description)
2. Location of the RES file
3. Location of the ENSP-ENSG.corr file (human) respectively ENSMUSP-ENSMUSG.corr (mouse)
4. The organism. Depending on the organism specified (*Homo sapiens* or *Mus musculus*), the appropriate bio-molecular interaction information is used.

Using these parameters the following operations are performed:

1. The matches with the highest scores are read out of the RES file
2. The ENSEMBL protein IDs in the extracted matches are correlated to the ENSEMBL gene IDs via ENSP-ENSG.corr
3. All bio-molecular interaction information correlated to the specified organism and found in a filtered match are stored in a BIND data file
4. All ENSEMBL gene IDs and the correlated BIND data file location are stored in the BIND identifier file.

The BIND info file was created, containing the following information:

1. The source of all files downloaded
2. The date of download
3. Short description of the data class and the total number of ENSEMBL genes connected to the BIND data class.

Example: The protein sequence gi4504349 is matched to the ENSEMBL protein ENSP00000333994. This ENSEMBL protein is associated to the ENSEMBL gene ENSG00000188170 via the ENSP-ENSG.corr file. Thus, all bio-molecular interaction connected to gi4504349 is inferred to the gene ENSG00000188170. This information is stored in the file ENSG00000188170.BIND.txt. Additionally, the following row was added to the BIND.identifier file (see Tab. 17):

ENSG00000188170,70,130,180,human/21_34/BIND/ENSG00000188170.BIND.txt.gz,
--

Tab. 17 The row for the gene ENSG00000188170 added in the BIND.identifier file

3.2.2 Database of Interacting Proteins (DIP)

The DIP catalogs interactions between proteins, which were experimentally determined [30]. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The provided data is curated manually and automatically using computational approaches.

In the GM the DIP data class is available for the organisms *Homo sapiens* (see Tab. 18) and *Mus musculus* (see Tab. 19). The procedure described below for creating the DIP data class is identical for both organisms. The differences between the human and the mouse DIP data class arise from the use of organism specific source files. Mouse source

files are correlated to genes in the *Mus musculus* matrix, human source files are correlated to genes in the *Homo sapiens* matrix.

ENSEMBL version	Number of DIP entries	Number of matched genes
9.30a.1	685	566
11.31.1	685	569
18.34.1	705	588
21.34	894	768

Tab. 18 Overview of the results for each performed integration of human DIP data. The number of DIP entries, the version of the ENSEMBL release used for the human gene catalog and the number of human ENSEMBL genes matched to DIP entries are stated.

ENSEMBL version	Number of DIP entries	Number of matched genes
9.30a.1	177	103
11.31.1	177	104
17.30.1	178	107
21.32	201	132

Tab. 19 Overview of the results for each performed integration of mouse DIP data. The number of DIP entries, the version of the ENSEMBL release used for the mouse gene catalog and the number of mouse ENSEMBL genes matched to DIP entries are stated.

3.2.2.1 Data source

Two files were downloaded from the DIP ftp server:

1. A XML¹⁷ file containing the description of the interactor(s) of the proteins that are in file 2 and the description of the interaction themselves
2. A FASTA file containing protein sequences of all the proteins involved in protein-protein interactions in DIP.

File 1 contains protein interaction information to be integrated into the DIP data class of the corresponding organism (see Tab. 21). The protein sequences in file 2 (see Tab. 20) resemble the information necessary to correlate data in file 1 to genes in GM.

¹⁷ eXtended Mark-up Language

```
>DIP:1N|sw:P19527|pir:A21762|gi:112046
KARMSLARAELEKRIDSLMDEIAFLKKVHEEEIAELQAQIQYAQISVEMDVSSKPDLSAALKDIRA
QYEKLAAKNMQNAEEWFKSRFTVLTESAAKNTDAVRAAKDEVSESRLLKAKTLEIEACRGMNE
ALEKQLQELEDKQNADISAMQDTINKLENELRSTKSEMARYLKEYQDLLNVKMALDIEIAAYRK
LLEGEETKLSFTSVGSITSGYSQSSQVFGRSAYSGLQSSSYLMSARAFPAYYTSHVQEEQSEVEETI
EATKAEEAKDEPPSEGESEEEEEKKDE
```

Tab. 20 FASTA formatted sequence for the DIP protein DIP:1N

```
<node uid="DIP:1N" id="G:952" name="NFL_RAT" class="protein">
  <feature name="SWP:P19527" class="cref">
    <src>SwissProt</src>
  </feature>
  <feature name="PIR:A21762" class="cref">
    <src>PIR</src>
  </feature>
  <feature name="GI:112046" class="cref">
    <src>NCBI</src>
  </feature>
  <att name="descr">
    <val>neurofilament triplet L protein</val>
  </att>
  <att name="taxon">
    <val>10116</val>
  </att>
  <att name="organism">
    <val>Rattus norvegicus</val>
  </att>
</node>
```

Tab. 21 Section of the DIP data file which contains all protein-protein interaction information available from the DIP database.

The header of all sequences in file 2 contains a DIP identifier. Every protein interaction information in file 1 is connected to a DIP identifier and thus a protein sequence.

3.2.2.2 Correlation

The protein sequences in file 2 were blasted against the organism specific peptide sequences from ENSEMBL using BLASTP (see Tab. 22). The best match between a DIP protein in ENSEMBL was considered as the resulting correlation. ENSP-ENSG.corr reveals the ENSEMBL gene associated to an ENSEMBL peptide, its correlated DIP protein and thus to all protein interaction information involving the DIP protein.

DIP:6N lsw:P08700 pir:A24427 gi:418834 152 1 152 ENSP00000296870 152 1 152 306 5e-84 100%
DIP:8Nlsw:P04626 pir:A24571 gi:86984 1255 1 1255 ENSP00000269571 1255 1 1255 2629 0.0 100%
DIP:12Nlsw: pir:A25828 gi:89457 209 2 209 ENSP00000229695 215 6 213 446 e-126 100%

Tab. 22 Section of a RES file showing 3 rows of summarized BLAST matches between DIP proteins and ENSEMBL proteins.

Example: All information about protein-protein interactions involving DIP:6N is correlated to ENSP00000296870.

3.2.2.3 Data extraction and creation of all data class files

A PERL script was generated in order to create all DIP data files and the DIP identifier file. Four parameters are necessary to run the script:

1. Location of the file 1 (containing the interactions description)
2. Location of the RES file
3. Location of the ENSP-ENSG.corr file (human) respectively ENSMUSP-ENSMUSG.corr (mouse)
4. The organism. Depending on the organism specified (*Homo sapiens* or *Mus muslucus*) the appropriate protein interaction information is taken.

In order to extract interactions concerning human or mouse proteins only, all sequences from other species were filtered out. Using these parameters, the following operations are preformed:

1. The matches with the highest scores are taken out of the RES file
2. The ENSEMBL protein IDs in the extracted matches are correlated to ENSEMBL gene IDs via the ENSP-ENSG.corr file

3. All protein interaction information correlated to the specified organism and a filtered match, is stored in a DIP data file
4. All ENSEMBL gene IDs and the correlated DIP data file location are stored in the DIP identifier file.

The DIP info file containing the following information was created:

1. The source of all files downloaded
2. The date of download
3. Short description of the data class and the total number of ENSEMBL genes connected to the DIP data class.

3.2.3 GeneNest

GeneNest provides data that represents each gene by a single cluster of ESTs and/or mRNAs. Further subdivision of a cluster into contigs may be caused by alternative splicing, genomic sequences, or artefacts like chimeric sequences. Consensus sequence derived from GeneNest contigs are a basis for mapping genes onto the genome, and for analysis of splice isoforms. [38,39]

The GeneNest data class is available for the *Homo sapiens* (see Tab. 23) and *Mus musculus* (see Tab. 24). The procedure described below for creating the GeneNest data class is identical for both organisms. The difference between generating the human and the mouse GeneNest data class arises from the use of organism specific source files. The mouse source file is correlated to genes in the *Mus musculus* matrix, and human source file is correlated to genes in the *Homo sapiens* matrix.

ENSEMBL version	Number of GeneNest entries	Number of matched genes
9.30a.1	184307	20546
11.31.1	185446	21891
18.34.1	207382	20675
21.34	219069	20891

Tab. 23 Overview of the results for each performed integration of human GeneNest data. The number of GeneNest entries, the version of the ENSEMBL release used for the human gene catalog and the number of human ENSEMBL genes matched to GeneNest entries are stated.

ENSEMBL version	Number of GeneNest entries	Number of matched genes
9.30a.1	43768	18008
11.31.1	73971	19339
17.30.1	105786	19862
21.32	111428	22335

Tab. 24 Overview of the results for each performed integration of mouse GeneNest data. The number of GeneNest entries, the version of the ENSEMBL release used for the mouse gene catalog and the number of mouse ENSEMBL genes matched to GeneNest entries are stated.

3.2.3.1 Data source

One organism specific file containing FASTA formatted nucleotide sequences (see Tab. 25) was retrieved from the GeneNest source for human and mouse. In the case of the GeneNest data class, the information to be integrated into the GM is the correlation between ENSEMBL genes and GeneNest contigs. Therefore, only one data source file was necessary.

```
>Hs2.1 /db=Hs7 /gene="NAT2" /contig=112 /reads=25 /EST=21 /mRNA=4 /gb=NM_000015
/clone=IMAGE:1870937 /len=1344 /product="Homo sapiens N-acetyltransferase 2 (arylamine N-
acetyltransferase)";
GACTTCCCTTGCAGACTTTGGAAGGGAGAGCACTTTATTACAGACCTTGAAGCAAGAGG
ATTGCATTCAGCCTAGTTCCTGGTTGCTGGCCAAAGGGATCATGGACATTGAAGCATATT
TTGAAAGAATTGGCTATAAGAACTCTAGGAACAAATTGGACTTGGAAACATTAAGTACA
```

Tab. 25 The human GeneNest nucleotide sequence for contig 1 in the cluster Hs2. The GeneNest identifier indicating the cluster and contig number is colored red. The source of this sequence is stored in GeneNestHs.GeneNest.fasta, located in HUMAN21.34/resources/.

The header of each sequence provides an identifier (e.g. **Hs2.1**) defining the cluster and contig ID (e.g. for Hs2.1 **Hs2** is the cluster, and **1** is the contig) related to this sequence.

3.2.3.2 Correlation

The ENSEMBL transcripts were blasted against GeneNest consensus sequences using BLASTN (see Tab. 26). The best match of a GeneNest consensus sequence to an ENSEMBL transcript was considered as the resulting correlation. ENSEMBL transcripts and their correlated GeneNest consensus sequences were associated to an ENSEMBL gene via ENST-ENSG.corr.

ENST00000339213 570 1 570 Hs446901.1 1692 436 843 553 e-156 98%
ENST00000326183 918 1 529 Hs446901.1 1692 928 1456 1009 0.0 99%
ENST00000342730 1357 1 668 Hs247828.1 882 219 882 1126 0.0 97%

Tab. 26 Section of a RES showing 3 rows of BLAST matches between ENSEMBL transcripts and GeneNest contigs, summarized via the Blast Summary script.

Example: Contig1 of the GeneNest cluster Hs446901 is correlated to the ENSEMBL transcript ENST00000326183 instead of ENST00000339213 due to a higher score.

3.2.3.3 Data extraction and creation of all data class files

The description of the matches between genes and GeneNest contigs are to be integrated into GM. The formatted BLAST result – the RES file – contains this data. The descriptions of the selected correlations are to be stored into the appropriate data file. A PERL script was generated to create the GeneNest data files and the GeneNest identifier. Two parameters were necessary to run the script:

1. Location of the ENST-ENSG.corr file (human) respectively ENSMUST-ENSMUSG.corr (mouse)
2. Location of the RES file.

According to these parameters specified at script execution time the following operations were performed:

1. The matches with the highest scores were extracted from the RES file
2. The ENSEMBL transcript IDs in the extracted matches were correlated to ENSEMBL genes via the ENST-ENSG correlation table
3. For each selected BLAST match the values for ENSEMBL transcript identifier, ENSEMBL transcript length, Match coordinates into the ENSEMBL transcript, Human GeneNest identifier, Human GeneNest length, Match coordinates into the human GeneNest, Match BLAST-score, Match E-value, and Match identity percentage were read out of the RES file and stored in the GeneNest data files

4. The GeneNest sequences were correlated to the ENSEMBL genes by combining (1) and (2). The locations of all data files were stored in the GeneNest identifier file.

A GeneNest info file containing the following information was created:

1. The source of all files downloaded
2. The date of download
3. Short description of the GeneNest data class
4. The number of matches between ENSEMBL gene IDs and GeneNest entries.

3.2.4 Induced Mutant Resource (IMR)

The IMR database contains information about genetically engineered strains of mice [42]. These mice were altered by gene transfer (transgenics), homologous recombination (gene targeting), and chemical mutagenesis. The information about these strains includes a description of the mutant phenotype, husbandry requirements, and genetic typing protocols.

The IMR data class is available for *Mus musculus* only (see Tab. 27).

ENSEMBL version	Number of IMR strains	Number of matched genes
9.30a.1	854	385
11.31.1	854	385
17.30.1	916	411
21.32	941	460

Tab. 27: Overview of the results for each performed integration of mouse IMR data. The number of IMR strains, the version of the ENSEMBL release used for the mouse gene catalog and the number of mouse ENSEMBL genes matched to IMR strains are stated.

3.2.4.1 Data source

One HTML file containing the IMR Index of Strains was downloaded from the IMR website. It contains mutation strain names associated to gene symbols, locus names and web links to the corresponding IMR data. The file was formatted (see Tab. 28), and data

(e.g. HTML tags) useless for further processing was removed. Mutation strains information is to be integrated into the IMR data class.

```
rptgene&gid=28">Apc
adenomatosis polyposis coli
datasht&sid=3">C57BL/6J-ApcMin
****
```

Tab. 28: A section of the formatted HTML file source.txt. The values highlighted in red show the gene symbol Apc and the associated strain name C57BL/6J-ApcMin

3.2.4.2 Correlation

Mutation strains were correlated to ENSEMBL genes using gene symbols provided with each mutation strains information. The external IDs correlation table (see Tab. 29) was used to correlate an ENSEMBL gene identifier to the gene symbols associated with a mutation strains information.

```
ENSMUSG00000005871,RefSeq,NM_007462,
ENSMUSG00000005871,LocusLink,11789,
ENSMUSG00000005871,SWISSPROT,APC_MOUSE,
ENSMUSG00000005871,EMBL,M88127,
ENSMUSG00000005871,EMBL,U02937,
ENSMUSG00000005871,protein_id,AAB59632.1,
ENSMUSG00000005871,protein_id,AAA03443.1,
ENSMUSG00000005871,MarkerSymbol,Apc,
ENSMUSG00000005871,AFFY_MG_U74,MG-U74A:101447_at,
ENSMUSG00000005871,AFFY_MG_U74v2,MG-U74Av2:101447_at,
ENSMUSG00000005871,AFFY_MG_U74v2,MG-U74Bv2:163408_at,
ENSMUSG00000005871,AFFY_MOE430,MOE430A:1420956_at,
ENSMUSG00000005871,AFFY_MOE430,MOE430A:1420957_at,
ENSMUSG00000005871,AFFY_MOE430,MOE430A:1450056_at,
ENSMUSG00000005871,AFFY_Mu11Ksub,Mu11KsubB:Msa.444.0_at,
ENSMUSG00000005871,Sanger_Mver1_1_1,H3125A10_1,
ENSMUSG00000005871,Sanger_Mver1_1_1,H3080G10_1,
ENSMUSG00000005871,Sanger_Mver1_1_1,H3083H09_1,
```

Tab. 29: Section of external Ids correlation table showing all identifiers associated to the ENSEMBL mouse gene ENSMUSG00000005871. The row with values highlighted in red show the association between the marker symbol Apc and the gene ENSMUSG00000005871.

Example: The strain C57BL/6J-ApcMin associated to the gene symbol Apc (see Tab.) was correlated to the ENSEMBL gene ENSMUSG0000005871 using the external IDs table (see Tab.). The correlation was done here via the listed MarkerSymbol.

3.2.4.3 Data extraction and creation of all data class files

A PERL script was generated to create all IMR data files and the IMR identifier file.

Two parameters are necessary to run this script:

1. Location of the formatted file 1
2. Location of the external IDs correlation table.

Using these parameters the following operations are performed:

1. The gene symbols associated to all strains are correlated to the ENSEMBL gene id via the external IDs table
2. The correlated strain name, the gene symbol, the ENSEMBL gene identifier and a web-link to the IMR data sheet are stored in the data files
3. All ENSEMBL gene IDs and the location of the data files are stored in the IMR identifier file.

The IMR info file containing the following information was created:

1. The source of the file downloaded,
2. the date of download,
3. a short description of the data class and the total number of matched ENSEMBL genes.

3.2.5 BioMedNet (BMN)

The Mouse Knockout & Mutation Database (MKMD) [43] is BioMedNet's comprehensive database of phenotypic and genotypic information on mouse knockouts and classical mutations. It contains data concerning over 3,000 unique genes. Unfortunately the MKMD in BioMedNet did close on 31 December 2004.

The BioMedNet data class is available for *Mus musculus* only (see Tab. 30).

ENSEMBL version	Number of BMN entries	Number of matched genes
9.30a.1	3432	1579
11.31.1	3432	1753
17.30.1	3853	1872
21.32	4191	2078

Tab. 30: Overview of the results for each performed integration of mouse BMN data. The number of BMN entries, the version of the ENSEMBL release used for the mouse gene catalog and the number of mouse ENSEMBL genes matched to BMN entries are stated.

3.2.5.1 Data source

Two types of files were downloaded, 27 from the BMN website and one from the MGI ftp server:

1. 27 HTML files, each containing a part of the BioMedNet Mouse Knockout & Mutation Database
2. One Tab-delimited text file containing MGI Marker associations to GenBank [44] sequence information (see Tab. 32).

The BioMedNet Mouse Knockout & Mutation Database contained gene names, alphabetically stored in 27 HTML files (type 1). All files were formatted, and all information useless for further processing (e.g. HTML tags) was removed. All formatted files were concatenated into a single file, further called source.txt (see Tab. 31). The gene names in source.txt are to be integrated into the BMN data class.

Transcription factor E2a
Transcription factor E2A
Transcriptional intermediary factor 1, beta
Transducer of ErbB-2.1
Transferrin
Transferrin receptor

Tab. 31: Section of the source.txt showing 6 sample rows, each stating one gene name.

File 2 (see Tab. 32) contains the following tab-delimited data per row that enables this correlation:

1. MGI Marker Accession ID
2. Marker Symbol
3. Status
4. Marker Type
5. Marker Name
6. cM Position
7. Chromosome
8. GenBank Accession IDs
9. Unigene ID
10. RefSeq ID

MGI:98822	Tfrc	O	Gene	transferrin receptor	21.2	16	AI195355	AI426448	AJ426432
AK011596	AK048228	AK088961	AU015758	AW910858	BC054522	BC067			
403	M29618	X57349	28683	NM_011638					

Tab. 32 Section of data file containing identifier associations to GenBank. The gene name and RefSeq ID are highlighted in red.

3.2.5.2 Correlation

The genes names were matched to ENSEMBL genes using file 2 and the external IDs correlation table (see Tab. 33). Complete gene names are missing in the external IDs table. File 2 contains associations between these gene names and other types of identifiers. The external IDs table contains associations between some of these identifiers and the ENSEMBL gene identifiers. A match of a gene name from BMN with any external identifier for an ENSEMBL gene was considered as the resulting correlation.

ENSMUSG00000022887,LocusLink,17174,
ENSMUSG00000022797,RefSeq,NM_011638,
ENSMUSG00000022797,LocusLink,22042,

Tab. 33 Section of the external Ids file with highlighted RefSeq ID and ENSEMBL mouse gene identifier associated to it.

Example: The gene name “Transferrin receptor” can be associated to the RefSeq identifier NM_011638 via the file 2 (MRK_Sequence.rpt). This RefSeq identifier is further associated to the ENSEMBL gene

ENSMUSG00000022797 via the external IDs correlation table. Thus all informations available in BioMedNet for “Transferrin receptor“ is related to ENSMUSG00000022797.

3.2.5.3 Data extraction and creation of all data class files

A PERL script was generated in order to create all BMN data files and the BMN identifier file. Three parameters are necessary to run this script:

1. Location of the external IDs correlation table
2. Location of source.txt
3. Location of the file2.

Using these parameters, the following operations are performed:

1. Gene names are matched to RefSeq-[45], GenBank-, Marker Symbol- or MGI identifiers via file 2
2. Gene names are correlated to an ENSEMBL gene id via the matched identifiers, using the external IDs table
3. All correlated gene names are stored in a BMN data file
4. All ENSEMBL gene IDs and BMN data file location are stored in the BMN identifier file.

The BMN info file containing the following information was created:

1. The source of all files downloaded.
2. The date of download.
3. Short description of the data class and the total number of connected ENSEMBL genes.

3.2.6 Bay Genomics GeneTrap

BayGenomics is a consortium of research groups. Its major goal is to identify genes relevant to cardiovascular and pulmonary diseases. Gene-trap vectors are used to inactivate thousands of genes in mouse embryonic stem (ES) cells for the purpose of generating knockout mice, incorporated into the BayGenomics GeneTrap database. [41]

The BayGenomics data class is available for *Mus musculus* only (see Tab. 34).

ENSEMBL version	Number of BayGenomics entries	Number of matched genes
9.30a.1	270	171
11.31.1	270	171
17.30.1	270	172
21.32	270	169

Tab. 34: Overview of the results for each performed integration of mouse BayGenomics data. The number of BayGenomics entries, the version of the ENSEMBL release used for the mouse gene catalog and the number of mouse ENSEMBL genes matched to BayGenomics entries are stated.

3.2.6.1 Data source

Two files were downloaded from the BayGenomics website:

1. HTML file containing a list of mutated known genes with mouse ES cell clones
2. HTML file containing phenotype analysis of productive secretory trap insertions.

Information in file 1 is related to a gene name, gene symbol and one external identifier. Information in file 2 is related to a gene name and gene symbol. All files were formatted, and data (e.h. HTML tags) useless for further processing was removed. File 1, further called `known_genes.txt` (see Tab. 35), contains the following tab-delimited data per row:

1. Gene name
2. Mouse symbol
3. External ID
4. Human symbol
5. Number of Hits

a disintegrin and metalloprotease domain 19 (meltrin beta)	Adam19	NM_009616	ADAM19	1
a disintegrin and metalloprotease domain 23	Adam23	NM_011780	ADAM23	3
ADP-ribosylation factor 1 (5'UTR)	Arf1	NM_007476	ARF1	1

Tab. 35: Section of `known_genes.txt` show 3 rows, each holding 5 values: gene name, mouse symbol, external ID, human symbol and number of hits.

File 2, further called mouse.txt (see Tab. 36), contains the following tab-delimited data per row:

1. ES cell line identifier
2. Gene name
3. Mouse Symbol
4. Chromosome number
5. Status
6. Bgal expression
7. Phenotype
8. Comment

Ex005	"mannosidase 2, alpha 1"	Man2a1	Un	T	viable
Ex054	baculovirus IAP repeat-containing	Birc6	Un	M	"lethal, neonatal"
Ex057	amyloid beta (A4) precursor like protein 2	Aplp2	9	T	viable

Tab. 36: Section of mouse.txt showing 3 rows.

The gene names listed in both files are to be integrated into the BayGenomics data class.

3.2.6.2 Correlation

BayGenomics GeneTrap data was correlated to ENSEMBL genes using the stated gene symbol or external ID. The external IDs correlation table was used to correlate an ENSEMBL gene identifier to either gene symbol or external ID associated with BayGenomics GeneTrap information.

3.2.6.3 Data extraction and creation of all data class files

A PERL script was generated to create all BayGenomics data files and the BayGenomics identifier file. Three parameters are necessary to run this script:

1. Location of the external Ids correlation table
2. Location of the known_genes.txt file
3. Location of the mouse.txt file.

Using these parameter the following operations are performed:

1. The gene symbols associated to BayGenomics GeneTrap genes were correlated to the ENSEMBL gene ID via the external Ids correlation table
2. The correlated gene symbols, the ENSEMBL gene identifier and a web-link to the BayGenomics website are stored in the data files
3. All ENSEMBL gene Ids and the location of the data files are stored in the BayGenomics identifier file.

The BayGenomics info file containing the following information was created:

1. The source of the files downloaded
2. The date of download
3. Short description of the data class and the total number of matched ENSEMBL genes.

4 Application of the GM into the c21 GM

4.1 The chromosome 21 GenomeMatrix

4.1.1 Concept of the c21 GM

The c21 GM was created in cooperation with the work group of ML Yaspo at the MPI-MG, which was involved in the sequencing of the chromosome 21 [46] as well as the establishment of its gene catalog. Furthermore, the group is involved in generating systematic functional genomics data for chromosome 21. The c21 GM is intended to be a specialised bioinformatics tool focused on this chromosome. It was created to contain diverse data sets generated by the community of researchers working on this topic as well as the publicly accessible data of the GM.

The GM structure is modular enough to enable an alteration of the system to adapt to new specifications and conditions. Data is related to a gene of an organism, whereas all genes of an organism are defined by its gene catalog. This modular system makes up the matrix of an organism. In addition to the systematic advantages of usability and functionality, the most important aspect of a bioinformatics tool is the quality of data it manages. Every relationship between data and genes depend on the quality of the data source, on the accuracy of the used gene catalog taken as a reference, and on the quality of the correlation method used to link genes and data.

The GM system was adapted for the human chromosome 21 and its orthologous mouse genes. The HSA21 database [47] was selected as a reference for the human chromosome 21 and mouse orthologs gene catalogs, which were manually curated in contrast to the human and mouse gene catalogs in GM, which were automatically generated by the ENSEMBL database. Associated to the genes, the data content of the c21 GM is composed of manually curated data as well as of all applicable GM data which was integrated into the new system to broaden content quantity and quality. The specific datasets are described in detail in chapter .

The overall approach was to retain the features and functionality of the GM system, as well as to increase the quantity and improve the quality of data integrated, keeping changes to a minimum. All new data as well as the GM data had to be correlated to the HSA21 gene catalog, implicating several modifications to existing procedures (chapter 4.2). The resulting c21 GM enables access to an extended pool of information.

4.1.2 The HSA21 database

The HSA21 database contains gene expression data related to human chromosome 21 mouse gene orthologs. It was build upon the gene catalog of chromosome 21 curated by the working group of ML Yaspo using the latest built of the DNA sequence of the human chromosome 21 and of the corresponding mouse gene catalogue (MMU21 genes) calculated by orthology relationships. The expression patterns associated to the MMU21 genes are composed of a large-scale mRNA *in situ* hybridization screen performed at a critical stages of embryonic (E 9,5) and brain (E 14.5 and neonatal development), and an *in silico* EST mining (Expressed Sequence Tags mining). This chromosome-scale expression annotation associates many of the genes tested with a potential biological role and suggests candidates for the pathogenesis of Down's syndrome. [47]

4.2 Alteration procedures

Several alterations were necessary to adapt the GM to the concept of the c21 GM. A correlation system between GM and c21 GM (chapter 4.2.1) was created to enable direct importation of data from the GM (chapter 4.2.4). Backbone tables had to be modified or created (chapter 4.2.2) in order to suit the new gene identifier nomenclature of the HSA21 catalog. The manually curated orthology data in the HSA21 database (chapter 4.2.3) was preferred over the data automatically generated for the GM by the program InParanoid (chapter 2.1.3).

4.2.1 Correlation system between GM and c21 GM

A frame of reference was created by correlating genes in the GM to the genes that are in the c21 GM, thus correlating the ENSEMBL gene catalog to the HSA21 gene catalog. In order to associate all known transcripts to each gene, all human and mouse gene references from the HSA21 database were blasted against the human and mouse

ENSEMBL transcripts, respectively, using BLASTN (see Tab. 37 and Tab. 38). The best matches were considered as the resulting correlation between an HSA21 gene identifier and an ENSEMBL transcript identifier.

hsa21- ABCC13 _AF418600 3364 1 2814 Transcript: ENST00000285661 2501 1 2231 1411 0.0 100%
hsa21-ABCC13_AF418600 3364 294 712 Transcript:ENST00000318742 510 1 419 831 0.0 100%
hsa21-ABCC13_AF418600 3364 597 2812 Transcript:ENST00000318761 1903 1 1657 585 e-165 100%

Tab. 37 Section of a RES file showing 3 rows of summarized BLAST matches between human ENSEMBL transcript sequences and human HSA21 sequences.

mmu21- TPTE _AJ3111313 2675 4 2156 Transcript: ENSMUST00000068928 2153 1 2153 4252 0.0 99%
mmu21-TPTE_AJ3111313 2675 4 2675 Transcript:ENSMUST00000033870 2549 1 2549 4159 0.0 99%
mmu21-LIPI_BB663289 659 4 659 Transcript:ENSMUST00000046480 2084 1 656 1281 0.0 99%

Tab. 38 Section of a RES file showing 3 rows of summarized BLAST matches between mouse ENSEMBL transcript sequences and mouse MMU21 sequences.

The ENST-ENSG and ENSMUST-ENSMUSG correlation tables reveal the ENSEMBL genes correlated to the human and mouse ENSEMBL transcripts, respectively. The resulting correlations between c21 GM genes and GM genes are further referred to as hsa21.ENSG for human (see Tab. 39) and mmu21.ENSMUSG (see Tab. 40) for mouse. The version number of the referred to ENSEMBL database is appended.

ABCC13 ENSG00000155288
ABCG1 ENSG00000160179
ADAMTS1 ENSG00000154734

Tab. 39 Section of hsa21.ENSG18 showing 3 sample rows.

mmu21-OLIG1_AB038696	ENSMUSG00000046160
mmu21-IFNAR2_NM-010509	ENSMUSG00000022971
mmu21-IL10RB_NM-008349	ENSMUSG00000022969

Tab. 40 Section of mmu21.ENSMUSG17 showing 3 sample rows.

4.2.2 Backbone tables

In order to increase the amount of available information inside the c21 GM the yeast, worm, and fly matrices were imported from the GM. The following identifier nomenclatures were selected in the c21 GM and GM matrices (see Tab. 41):

Matrix	GM identifier nomenclature	c21 GM identifier nomenclature
<i>Homo sapiens</i>	ENSEMBL	HSA21
<i>Mus musculus</i>	ENSEMBL	ENSEMBL & HSA21
<i>D. melanogaster</i>	FlyBase	FlyBase
<i>S. cerevisiae</i>	SGD	SGD
<i>C. elegans</i>	Wormbase	Wormbase

Tab. 41 Overview of all identifier nomenclature sources used for the c21 GM and GM matrices of *Homo sapiens*, *Mus musculus*, *D. melanogaster*, *S.cerevisiae* and *C.elegans*.

The list of human genes, their chromosomal positions and gene identifier nomenclature were extracted from the HSA21 database. The c21 GM backbone table for *Homo sapiens* (see Tab. 42) was directly build on this data, applying the GM backbone table layout (chapter 2.1.2):

TPTE ,21,7669047,7753226,255,0,0,human/data/hsa21/TPTEmain.html.gz,
PRED1 ,21,7910043,7910115,0,0,255,human/data/hsa21/PRED1main.html.gz,
C21orf99 ,21,11076976,11086203,255,0,0,human/data/hsa21/C21orf99main.html.gz,
FGF7L ,21,11383451,11385008,0,0,255,human/data/hsa21/FGF7Lmain.html.gz,

Tab. 42 Section of the human backbone table of the c21 GM showing 4 rows.

A new reference HTML file was created for every gene, each opening the HSA21 database website and scrolling to the position of the referred to gene (see Fig. 4).

Application of the GM into the c21 GM




Human gene [Accession No.]	Description and Gene Ontology	CDY Ov%.	Mutants	Human in silico expr.	Mouse Ortholog [Accession No.]	mouse in silico expr.	RT-PCR E1E.5/P2	whole mount annotation	ISH image whole mount	brain section annotation	ISH image brain section
ZNF294 [AB018257]	Description: human mRNA for KIAA0714 protein Gene Ontology: enzyme / protein folding and degradation RING finger. Domains: putative E3 ubiquitin-protein ligase (EC 6.3.2.19).	CDY		EST	Znf294 [NM_128324] clone#36 [IMAGE:RZPD]	EST		wh_emb:ws		OB, CTX, vFB	 download
C21orf8 [AK011713]	Description: Unknown function Gene Ontology: unknown / unknown	CDY		EST	no ortho available						
C21orf5 [AK000706] [U19208]	Description: chromosome 21 open reading frame 5 Gene Ontology: unknown / unknown Domains: RWD; Domains in RING finger and WD repeat containing proteins and CEXDC-like helicases subfamily related to the UBCs Domains.			EST	ORF5 [NM_016324] clone#247 [IMAGE:RZPD] clone#251 [IMAGE:RZPD]	EST	- / -	clone#247 wh_emb: not detected clone#251 wh_emb: not detected		Ubiquitous	 download
USP16 [AF126736] [AK021104] [AF111323] [AK021247]	Description: Ubiquitin processing protease, EC 3.1.2.15. Gene Ontology: enzyme / protein folding and degradation. Domains: Ubiquitin_Carboxyl-terminal Hydrolase-like zinc finger.	CDY		EST	Usp16 [NM_024258] clone#16 [IMAGE:RZPD]	EST	+ / +	wh_emb: not detected		Not detected	
CCT8 [U13827]	Description: T-complex protein 1, theta subunit Gene Ontology: chaperone / protein folding and degradation	CDY		EST	Cct8 [NM_020867] clone#47 [IMAGE:RZPD]	EST		wh_emb:ws		OB, CTX, HP, vFB	 download
C21orf12 alien: FRG.071	Description: alien prediction only Gene Ontology:			EST	no confirmed ortholog						

Fig. 4 Screenshot of the HSA21 database website. The entry row of the human gene ZNF294 is highlighted.

The list of mouse genes, their chromosomal positions and gene identifier nomenclature were extracted from the HSA21 database. This data was used to modify the existing GM backbone table for *Mus musculus* in order to function as the new c21 GM backbone table for *Mus musculus*.

All ENSEMBL mouse gene entries in the GM backbone table which are absent in the mmu21.ENSMUSG17 file were removed because they were not in the HSA21 database. 21 mouse gene entries absent from the GM backbone table and present in the HSA21 mouse gene catalogue were added (see Tab. 43).

mmu21-C21orf119
mmu21-C21orf33
mmu21-C21orf34
mmu21-C21orf37
mmu21-C21orf41
mmu21-C21orf5
mmu21-C21orf51
mmu21-C21orf98
mmu21-FLJ37539
mmu21-H2BFS
mmu21-HUNK

mmu21-KRTAP6-1
 mmu21-KRTAP6-3
 mmu21-KRTAP10-10
 mmu21-KRTAP11-1
 mmu21-PRED15
 mmu21-PRED37
 mmu21-PRED38
 mmu21-PRED62
 mmu21-SLC5A3
 mmu21-SYNJ1

Tab. 43 The list of mouse genes added to the GM backbone table of mouse, rendering it functional as a backbone table in the c21 GM.

The resulting backbone table includes the ENSEMBL and HSA21 gene identifier nomenclatures (see Tab 44.).

ENSMUSG00000051730,2,70818531,70832865,153,153,153,mouse/17_30/gene_summary/ENSMUSG00000051730.html.gz,
 ENSMUSG00000051731,3,130717065,130717693,153,153,153,mouse/17_30/gene_summary/ENSMUSG00000051731.html.gz,
 mmu21-C21orf119,1,1,2,153,153,153,human/data/hsa21/C21orf119.html.gz,
 mmu21-C21orf33,1,1,2,153,153,153,human/data/hsa21/C21orf33.html.gz,

Tab. 44 Section of the mouse backbone table. Each row represents one gene (see chapter 2). The first 2 rows are entries from the GM mouse backbone table, the last 2 rows show added entries. The complete backbone tables is mm.backbone, stored in GM/chr21NEW/.

For each added mouse gene a new reference HTML file was created, opening the HSA21 database website for the referred to gene.

The GM backbone tables for the organisms *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Caenorhabditis elegans* were directly integrated into the c21 GM. No changes were necessary due to unchanged identifier nomenclatures.

4.2.3 Orthology relationships

The manually annotated orthology data from the HSA21 database (see Tab. 46) was preferred to the result of the program InParanoid (chapter 2.1.3) and was applied to the c21 GM.

Orthology relationships in the HSA21 database	Table name
<i>Homo sapiens</i> to <i>Mus musculus</i>	human_2_mouse.lst
<i>Homo sapiens</i> to <i>Drosophila melanogaster</i>	human_2_fly.lst
<i>Homo sapiens</i> to <i>Saccharomyces cerevisiae</i>	human_2_yeast.lst
<i>Homo sapiens</i> to <i>Caenorhabditis elegans</i>	human_2_worm.lst

Tab. 45 Overview of existing c21 GM orthology tables and their table names.

Orthology data had to be adapted in order to keep compatibility to GM and c21 GM data as well as new identifier nomenclatures. The GM orthology table layout (chapter 2.1.3) was applied for each orthology data set.

ABCG1 worm:CAA93461.1 fly:AAF52835.1 yeast:NP_014796.1
ADAMTS1 worm:CAA93287.1 fly:AAF46065.1
ADAMTS5 worm:CAA93287.1 fly:AAF46065.1

Tab. 46 Section of manually annotated orthology relationship data showing 3 sample rows.

HSA21 orthology data for human and mouse uses HSA21 gene identifiers for both organisms. These had to be adapted to match the gene identifiers present in the human and mouse c21 GM backbone tables. A combination of ENSEMBL and HSA21 gene identifiers were used in the mouse c21 GM backbone table (chapter 4.2.2). All HSA21 mouse gene identifiers in the orthology data, which are now referred to via ENSEMBL gene identifiers, have to be exchanged with the corresponding ENSEMBL identifiers (see Tab. 47). The modification was done using a generated PERL script.

Example: The human gene ZNF295 points to the mouse gene MMU21-ZNF295 in the HSA21 database orthology table. MMU21-ZNF295 is ENSMUSG00000047364 in the c21 GM mouse backbone table. Replacing the mouse identifier MMU21- ZNF295 by ENSMUSG00000047364 in the HSA21 database orthology table renders this relationship functional inside the c21 GM.

ZNF294,ENSMUSG00000041098,100,
ZNF295,ENSMUSG00000047364,100,
C21orf119,mmu21-C21orf119,100,

Tab. 47: Section of human to mouse orthology relationship table in the c21 GM showing 3 rows.

HSA21 Orthology data points human HSA21 gene identifiers to *Drosophila melanogaster* NCBI protein identifiers. FlyBase gene identifiers are used in the backbone of the c21 GM fly matrix. All NCBI [44] protein identifiers in the orthology data have to be exchanged with the corresponding FlyBase identifiers (see Tab. 48). The modification was done using a generated PERL script and a correlation table.

ADARB1,FBgn0026086,100, PFKL,FBgn0003071,43.83, SOD1,FBgn0033631,100,

Tab. 48: Section of the human to fly orthology relationship table in the c21 GM showing 3 sample rows.

HSA21 Orthology data points human HSA21 gene identifiers to *Saccharomyces cerevisiae* NCBI protein identifiers. SGD gene identifiers are used in the backbone of the c21 GM yeast matrix. All NCBI protein identifiers in the orthology data have to be exchanged with the corresponding SGD identifiers (see Tab. 49). The modification was done using a generated PERL script and a correlation table.

USP25,S0004920,100, DNMTA1,S0002547,100, CCT8,S0003545,100,

Tab. 49: Section of the human to yeast orthology relationship table in the c21 GM showing 3 sample rows.

HSA21 Orthology data points human HSA21 gene identifiers to *Caenorhabditis elegans* NCBI protein identifiers. Wormbase gene identifiers are used in the backbone of the c21 GM worm matrix. All NCBI protein identifiers in the orthology data have to be exchanged with the corresponding Wormbase identifiers (see Tab. 50). The modification was done using a generated PERL script and a correlation table.

C21orf19,C37C3.8a,90.22, ADARB1,T20H4.4,100, SLC19A1,C06H2.4,11.2,
--

Tab. 50: Section of the human to worm orthology relationship table in the c21 GM showing 3 sample rows.

4.2.4 Importing GM data classes into the c21 GM

GM data classes were inserted into the c21 GM to increase data content. Modifications on the gene catalogs and thus the backbone tables (chapter 4.2.2) defined the importing procedures. An unchanged backbone table enabled direct implementation of all related GM data classes. A modified backbone table implicated the process of several alterations before importing the related GM data classes.

A data class identifier file stands for all connections between a gene identifier and a data class. Changing the gene identifiers in a data class identifier file (chapter 4.2.1) means converting the whole data class to the new nomenclature. This data structure (chapter 2.1.2) limits all necessary alterations to one file. All data files containing the actual data class information are left unchanged.

The use of HSA21 gene identifiers in the human c21 GM backbone table rendered all human GM data classes inoperative in the c21 GM. ENSEMBL gene identifiers used in all human GM data classes had to be replaced by HSA21 gene identifiers. A generated PERL script and the hsa21.ENSG18 were applied to modify all human GM data class identifier files. All ENSEMBL entries absent in hsa21.ENSG18 were removed. The resulting identifier files render the connected data classes functional inside the c21 GM.

Example: The human ENSEMBL gene ENSG00000142166 is IFNAR1 in the c21 GM. The row for ENSG00000142166 in the human GM DIP data class identifier file is imported into the c21 GM after changing ENSG00000142166 to IFNAR1.

ENSG00000142166 ,173,255,47,human/18_34/DIP/ENSG00000142166.DIP.txt.gz

Tab. 51 Section of a GM human DIP identifier file

IFNAR1 ,173,255,47,human/18_34/DIP/ENSG00000142166.DIP.txt.gz
--

Tab. 52 Section of the resulting c21 GM DIP identifier file.

The c21 GM mouse backbone table was modified with added gene entries. Old gene entries were still based upon the GM mouse identifier nomenclature (ENSEMBL), rendering mouse GM data classes functional for these gene entries. A PERL script was generated to remove entries from all mouse GM data classes, which were absent in the mouse c21 GM backbone table.

All GM data classes from the Dm, Sc and Ce matrices were directly imported into the c21 GM due to an unchanged identifier nomenclature.

4.3 Data mining for the c21 GM

The presences of human as well as mouse *in silico* expressions, mouse whole mount ISH data and mouse brain sections ISH data inside the HSA21 database were integrated as individual data classes into the c21 GM.

4.3.1 *In silico* expression

In silico EST mining results represent an electronic, qualitative expression profile obtained for each chromosome 21 gene. They are classified by tissue type and available for *Homo sapiens* and *Mus musculus* inside the HSA21 database. The procedure of creating a data class from this source is identical for both organisms and differs only in the use of an organism specific source file.

4.3.1.1 Data source

One file per organism was retrieved containing the list of genes with associations to data on *in silico* EST mining results. The reference to the existence of such information inside the HSA21 database was implemented as *in silico* expression information into the c21 GM. This source file includes all information necessary to create all data class files.

4.3.1.2 Creation of all data class files

A PERL script was generated to create all *in silico* expression data files and the *in silico* expression identifier file. The following operations are performed:

1. A HTML file functioning as a data file was generated for each listed gene, including a link to the entry of this gene at the HSA21 database website

2. The identifier file was directly generated out of the source file by associating RGB color values as well as the location of a generated data file to each listed gene.

The *in silico* expression info file containing the following information was created:

1. The source of the downloaded file
2. The date of download
3. Short description of the data class and the total number of matched HSA21 genes

4.3.2 Whole mount *in situ* hybridization

Whole mount *in situ* hybridization (whole mount ISH) information from the HSA21 database shows gene expressions images for whole mount embryos, at the stage of development E14.5, associated to data annotations. The whole mount ISH data class is available for *Mus musculus* only.

4.3.2.1 Data source

One file containing the list of mouse genes with associations to data on whole mount ISH was retrieved from the HSA21 database. The reference to the existence of such information inside the HSA21 database was implemented as whole mount ISH information into the c21 GM. This source file includes all information necessary to create all data class files.

4.3.2.2 Creation of all data class files

A PERL script was generated to create all whole mount ISH data files and the whole mount ISH identifier file. The following operations are performed:

1. A HTML file functioning as a data file was generated for each listed gene, including a link to the entry of this gene at the HSA21 database website
2. The identifier file was directly generated out of the source file by associating RGB color values as well as the location of a generated data file to each listed gene.

The whole mount ISH info file containing the following information was created:

1. The source of the downloaded file
2. The date of download
3. Short description of the data class and the total number of matched HSA21 genes

4.3.3 Brain sections *in situ* hybridization

Brain sections *in situ* hybridization (whole mount ISH) information from the HSA21 database shows gene expressions images for brain sections embryos, at the stage of development P2, associated to data annotations. The brain sections ISH data class is available for *Mus musculus* only.

4.3.3.1 Data source

One file containing the list of mouse genes with associations to data on brain sections ISH was retrieved from the HSA21 database. The reference to the existence of such information inside the HSA21 database was implemented as brain sections ISH information into the c21 GM. This source file includes all information necessary to create all data class files.

4.3.3.2 Creation of all data class files

A PERL script was generated to create all brain sections ISH data files and the brain sections ISH identifier file. The following operations are performed:

1. A HTML file functioning as a data file was generated for each listed gene, including a link to the entry of this gene at the HSA21 database website
2. The identifier file was directly generated out of the source file by associating RGB color values as well as the location of a generated data file to each listed gene.

The brain sections ISH info file containing the following information was created:

1. The source of the downloaded file
2. The date of download
3. Short description of the data class and the total number of matched HSA21 genes

5 Discussion

This work demonstrates the significance and advantage of data mining as well as the need for an adjustable bioinformatics tool in the context of the GM database system. These two aspects are discussed in detail in the following, followed by proposals for the future.

5.1 Data mining for the GenomeMatrix

The first objective was the integration of data on human and mouse from specific data sources into the GM. The result is one data class for each integrated source. A data class is the outcome of an applied data mining strategy and represents all information from one source that could be related to the genes of a gene catalog present in the GM (chapter 3).

Example: Human protein-protein interactions provided by the DIP database are to be integrated into the GM. Data from the DIP database is used to relate human DIP interactions to genes of the human ENSEMBL gene catalog present in the GM. All interactions that can be related to a human gene are integrated into the GM and all integrated interactions form the human DIP data class.

The resulting advantages are the possibilities of data access and data visualization provided in the GM. In reference to a functional genomics approach, data access is to a large extent enhanced by the correlation of information from a data source to genes regardless of type. Arraying information according to associated genes in contrast to types of information simplifies functional genomics analysis on genes by offering a pool of different data for each gene.

Example: The combined associations between the mouse gene ENSMUSG00000004056 and information about protein-protein interactions inside the DIP data class, information about existing GeneNest associations inside the GeneNest data class, and information about gene knockout data inside the IMR data class form a pool of information on this gene. This pool represents a previous evaluation of different information in terms of gene association. A functional analysis on this gene with regard to these sources is simplified to a large extent by the mere existence and availability of this pool inside the GenomeMatrix.

Choosing the data classes of interest defines the content of such a pool of data. The GUI of the GM is capable of providing multiple pools of data on a gene through the display of a number of different data sources related to a number of different genes via a matrix of colored boxes (see chapter 2.1.1). Columns represent different genes and rows represent the different types of information related to the genes. The color indicates the mere existence of information, and can also encode the information itself. This layout reduces data comparison of different data sources or investigation on a single data source to a large extent. Data from different sources can now be compared and associated to each other directly in one unified, graphical display next to each other without the need to access each data source separately. Further, the possibility of displaying data classes from different organisms related through orthology provides additional information useful in functional genomics and increases the significance of such direct comparison and association.

Inconsistencies encountered in naming conventions of identifiers as well as data structure had a great impact on data integration in terms of effectiveness, possible automation and thus human involvement. The dimension of data mining used in this thesis is only made possible by the application of computer technology due to the amount and level of data necessary for analysis. This in turn calls for the development and deployment of programs to produce human-readable information from complex data. The functionality of a program is consequently defined by its capability to identify this data for further processing. Consistent data structure, as well as identifier naming convention, are essential to enable the recognition of valuable parts in data, defined by rules stated in the code of the program. Any type and degree of change to this data structure or identifier naming convention renders the data unreadable, which makes automated procession impossible. A program would have to be manually modified in order to be rendered functional again.

The manual control of results as well as sources becomes mandatory due to the possibility and occurrence of such inconsistencies, furthermore reducing the reasonable extend of combined data mining applications and thus the possible level of automation. The following examples show the possible impact of such an inconsistency.

In the course of data mining from the DIP database, the procedures and programs generated had to be modified due to changes of data structure in the DIP source file containing protein-protein interaction information. The complete alteration of the way how interaction data was stored demanded an analysis of the new structure and, concluding from this, the modification of existing data mining procedures.

In the case of data mining from the BioMedNet database, the sudden absence of necessary additional information rendered the developed data mining procedure for this data source nonfunctional. The phenotypic and genotypic information on mouse knockouts and classical mutations was provided by the BioMedNet database solely via gene names. Without any additional information like sequences or known identifiers, one possibility to correlate these gene names to data like ENSEMBL gene identifier is to seek a new source that gives a direct relationship between BioMedNet gene names to ENSEMBL gene identifier. Another possible solution is the access to information that gives an indirect relationship between BioMedNet gene names to an identifier with already known relationship to ENSEMBL, like e.g. GenBank accession numbers.

The elemental questions raised and answered during the development of data mining strategies can help to solve inconsistency problems. A possible solution is to apply these questions each time data mining is applied:

Is the desired information clearly defined by its source, or is additional information necessary?

How can this information be extracted from its source?

What is necessary to associate extracted data to the system of interest?

5.2 The c21 GM - a bioinformatic tool

An existing pool of detailed internal information on the human chromosome 21 formed the demand for superior data management and visualization. The modular system design as well as the integrated type of data visualization were the crucial factors for choosing the GM system as a bioinformatics tool.

The internal design of the GM system had to be adapted to comply with the provided definition of the human chromosome 21 gene catalog as well as to further offer GM data which have a relationship to the new catalog. The most significant alteration of the GM in terms of content quality and system structure was the integration of the new gene catalog, which established a new gene identifier nomenclature. This had a great impact on multiple sites given that every type of data relationship within the GM is expressed through gene identifiers. As a result, all relationships based on exchanged human gene identifiers were rendered non-functional. A variety of programs were developed to integrate the gene catalog as well as to reestablish all non-functional data relationships for data still possessing a relation to the new gene catalog. In addition, several new internal data sets were integrated to broaden the content of the system.

The c21 GM can be regarded as a bioinformatic tool with some content limitations that in return provide its advantages over the regular GM. The focus on a single chromosome of one organism in contrast to the spectrum of complete genomes offered in the GM affects its field of application and usefulness to the scientific community. The c21 GM is intended for scientists interested in a large pool of different types of data related to a high-quality definition of the human chromosome 21.

5.3 Future proposal

An open-source GM system that is capable of automatically implementing any given gene catalog as well as adapting all available data from the regular GM to suit this gene catalog would increase the value of the GM for scientists of all fields. The creation of the c21 GM could represent the first step towards this proposal, followed by further enhancement of the already developed procedures, methods and programs.

6 Abbreviations

ASN.1	Abstract Syntax Notation
ASCII Interchange	American Standard Code for Information
BIND	Biomolecular Interaction Network Database
BMN	BioMedNet
BLAST	Basic Local Alignment and Search Tool
BLASTP	BLAST program for protein
BLASTN	BLAST program for nucleotide
c21 GM	chromosome 21 GenomeMatrix
cDNA	copy DeoxyriboNucleic Acid
Ce	<i>Caenorhabditis elegans</i>
DIP	Database of Interacting Proteins
Dm	<i>Drosophila melanogaster</i>
DNA	DeoxyriboNucleic Acid
e.g.	exempli gratia
ENSP-ENSG.corr table	human ENSEMBL protein to gene correlation table
ENSMUSP-ENSMUSG.corr table	mouse ENSEMBL protein to gene correlation table
ENSMUST-ENSMUSG.corr table	mouse ENSEMBL transcript to gene correlation table
ENST-ENSG.corr table	human ENSEMBL transcript to gene correlation table
ES	Embryonic Stem
EST	Expressed Sequence Tags
FASTA	A sequence in FASTA format begins with a single-
	line description, followed by lines of sequence data
ftp	file transfer protocoll
GM	GenomeMatrix

Abbreviations

GUI	Graphical User Interface
Hs	<i>Homo sapiens</i>
HSA21	
hsa21.ENSNG	human HSA21 to ENSEMBL correlation table
HTML	HyperText Markup Language
i.e.	id est
ID	identifier
IMR	Induced Mutant Resource
ISH	<i>in situ</i> hybridization
<i>in situ</i>	in place
<i>in silico</i>	computational
MKMD	Mouse Knockout & Mutation Database
Mm	<i>Mus musculus</i>
MMU21	
MPI-MG	Max Planck Institute for Molecular Genetics
mRNA	messenger RiboNucleic Acid
mmu21.ENSMUSG	human MMU21 to ENSEMBL correlation table
NCBI	National Center for Biotechnology Information
PERL	Practical Extraction and Report Language
RES	result of the appliance of BlastSummary
RGB	Red-Green-Blue color system
RNA	RiboNuclei Acid
RZPD	Deutsches Ressourcenzentrum für
Genomforschung	GmbH
Sc	<i>Saccharomyces cerevisiae</i>
SGD	<i>Saccharomyces</i> Genome Database
XML	eXtended Mark-up Language

7 List of references

1. Pennisi E. Finally, the book of life and instructions for navigating it. *Science* 2000; 288: 2304–2307.
2. Ewing B, Green P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* 2000; 25: 232–234.
3. Liang F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* 2000 ; 25: 239–240.
4. Pandey A, Mann M. Proteomics to study genes and genomes. *Nature* 2000; 405: 837–846.
5. Kemmeren P. *et al.* Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* 2002; 9: 1133–1143.
6. Snyder M, Fields S. Protein analysis on a proteomic scale. *Nature* 2003; 422: 208-215.
7. Kao, C.M. Functional genomics technologies: creating new paradigms for fundamental and applied biology. *Biotechnol. Prog.* 1999; 15, 3, 304-311.
8. McKusick, V.A. Genomics: structural and functional studies of genomes. *Genomics* 1997; **45, 2**, 244-249.
9. Julio E. Celis *et al.* Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.* 2000; 480(1): 2-16. Review.
10. Legrain P. *et al.* Protein–protein interaction maps: a lead towards cellular functions. *Trends in Genetics* 2001; 17: 346-352.
11. Jones S, Thornton JM, Analysis and classification of protein-protein interactions from a structural perspective, *Protein-Protein Recognition*, Oxford University Press, 2000.
12. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature.* 2003; 421(6920): 231-237.
13. Brutlag, D.L. Genomics and computational molecular biology. *Curr. Opin. Microbiol.* 1998; 1, 3, 340-345.
14. Ashburner, M. and Goodman, N. Informatics: genome and genetic databases. *Curr. Opin. Genet. Dev.* 1997; 7, 6, 750-756.
15. Kanehisa, M. Grand challenges in bioinformatics. *Bioinformatics* 1998; 14, 4, 309.
16. Boguski, M.S. Bioinformatics. *Curr. Opin. Genet. Dev.* 1994; 4, 3, 383-388.

17. Hunter, L. Progress in computational molecular biology. *Sigbio News*. 1999; 19, 3, 9-12.
18. Waterman, M.S. *Introduction to Computational Biology*. Chapman and Hall, New York, 1995.
19. Adams MD. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* 2000; 287: 2185–2195.
20. The *C. elegans* sequencing consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 1998; 282: 2012–2018.
21. Goffeau A. *et al.* The yeast genome directory. *Nature* 1997; 387: 100–105.
22. Chervitz SA. *et al.* Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 1998; 282: 2022–2027.
23. Walhout AJM, Boulton SJ, Vidal M. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* 2000; 17:88-94.
24. James A. Cuff *et al.* The Ensembl Computing Architecture. *Genome Res.* 2004; 14: 971-975.
25. Val Curwen *et al.* The Ensembl Automatic Gene Annotation System. *Genome Res.* 2004; 14: 942-950.
26. Ewan Birney *et al.* An Overview of Ensembl. *Genome Res.* 2004; 14: 925-928.
27. Lincoln D. Stein *et al.* WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* 2001;29(1): 82-86.
28. Rachel A. Drysdale *et al.* FlyBase: genes and gene models. *Nucleic Acids Res.* 2005; 33: D390-5.
29. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*2003; 31(1):248-50.
30. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: The Database of Interacting Proteins. *NAR* 2000; 28:289-91
31. Rainsford, C.P. and Roddick, J.F. Database issues in knowledge discovery and data mining. *Aust. J. Inf. Syst.* 1999; **6, 2**, 101-128.
32. Ming-Syan, C., Jiawei, H. and Yu, P.S. Data mining: an overview from a database perspective. *IEEE Trans. Knowl. Data Eng.* 1996; **8, 6**, 866-883.
33. W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. *AI Magazine*, Fall 1992, pgs 213-228.

34. Frank Brown et al. Chemoinformatics: What is it and How does it Impact Drug Discovery. *Ann. Reports Med. Chem.* 1998; 33: 375-384.
35. J. Michael Cherry et al. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* 1998; 26(1):73-79.
36. Remm M, Storm CEV and Sonnhammer ELL. Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. *JMB* 2001; 314:1041-1052.
37. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* 1990; 215:403-410.
38. Haas,S., Kirby,S., Peters,M., Toussaint,B., Lehrach,H., Poustka,A., Vingron,M., Korn,B. Consensus Sequence Database of all human Genes and ESTs; Minimal Set of Human EST clones available. *DHGP Xpress* 1998.
39. Haas,S.A., Beissbarth,T., Rivals,E., Krause,A., Vingron,M. GeneNest: automated generation and visualization of gene indices. *Trends Genet.* 2000; 16 (11), 521-52.
40. Michael Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25(1): 25-29.
41. Doug Stryke et al. BayGenomics: a resource of insertional mutations in mouse embryonic stem cells. *Nucleic Acids Res.* 2003; 31(1): 278-281.
42. Mouse Mutant Resource Web Site, The Jackson Laboratory, Bar Harbor, Maine. World Wide Web (<http://www.jax.org/mmr/>)
43. Jim Giles. Elsevier waves goodbye to BioMedNet web portal. *Nature.* 2003; 426(6968): 744.
44. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. Genbank. *Nucleic Acids Res.* 2000; 28, 1, 15-18.
45. Kim D. Pruitt et al. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2005; 33: D501-504.
46. The chromosome 21 mapping and sequencing consortium The DNA sequence of human chromosome 21. *Nature* 2000; 405, 311-319.
47. A gene expression map of human chromosome 21 orthologues in the mouse *Nature* (2002) 420 : 586 – 590.

8 Appendix

8.1 DVD

Appended to this thesis is a DVD containing supplementary information. It contains all generated PERL scripts, necessary source files as well as the resulting data sets produced over the course of this work.

8.2 List of data sources

The GenomeMatrix made use of the following data sources during the time of this thesis:

01. ENSEMBL project- a joint project between EMBL - EBI and the Sanger Institute which produces and maintains automatic annotation on eukaryotic genomes
02. FlyBase- a Drosophila genome database
03. SWISS-PROT Protein Knowledgebase- an annotated protein sequence database
04. Gene Ontology Consortium- to produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing
05. The GeneNest database
06. The RZPD- Human and Mouse Mammalian Gene Collection
07. Mouse GeneTrap clones provided by the German GeneTrap Consortium (GGTC)
08. Mouse in-situ clones provided by B.Herrmann, MPI f. Immunbiologie
09. BIND- The Biomolecular Interaction Network Database
10. DIP- The Database of Interacting Proteins
11. PSF- Protein Structure Factory
12. PDB Protein Data Bank
13. Full length human cDNA clones from GFP(Green fluorescent protein)-fusion-protein experiments. Data provided by Stefan Wiemann, DKFZ, Heidelberg, Germany
14. Human protein expression clones generated by K.Buessow at the MPI for Molecular Genetics

15. KEGG database (Kyoto Encyclopedia of Genes and Genomes)- Mapping of genes to enzyme information and pathway map(s).
16. The Human Tetraodon Ecores
17. Zebrafish in-situ clones provided by Pia Aanstad and Alberto Musa from the MPI for Molecular Genetics
18. Mouse Genome Database (MGD) & Gene Expression Database (GXD) - databases for the mouse *Mus musculus*
19. WormBase - database for the nematode *Caenorhabditis elegans*
20. Bay Genomics GeneTraps database (Lab of Dr. William Skarnes, U.C. Berkeley).
The list of known genes and the list of mouse mutants
21. "Induced Mutant Resource Index of Strains" was downloaded from the The Jackson Laboratory database
22. The BioMedNet- Mouse Knockout & Mutation Database
23. Germline - Collection of *Caenorhabditis elegans* gene expression data

9 Zusammenfassung

Die rasche Technologieentwicklung der letzten Jahrzehnte hat den Naturwissenschaften neue Perspektiven eröffnet, aber auch neue Problemstellungen hervorgebracht. Weltweit konnte eine enorme Masse an Daten erzeugt werden, die aber zu einem großen Teil noch entziffert, eingegliedert, zusammengestellt, zueinander in Beziehung gesetzt und verwaltet werden müssen. Die Anwendung von Computertechnologie eröffnete neue Möglichkeiten, um die mit dieser Datenverwaltung zusammenhängenden Probleme zu lösen. Eine neue Disziplin war geboren, die Bioinformatik.

GenomeMatrix ([http://www.genomematrix.org/.](http://www.genomematrix.org/)) ist eine neuentwickelte Online-Datenbank, die den Zugang zu biologischen Daten aus einer Vielzahl von verschiedenen Ressourcen ermöglicht, wie zum Beispiel Angaben über Gene, deren Produkte und Funktionen. Im Rahmen dieser Arbeit wurde mittels Prozeduren des sogenannten „Data mining“ der Inhalt diverser Datenbanken (BIND, DIP, GeneNest, IMR, BayGenomics GeneTrap sowie BioMedNet) mit einem Referenz-Katalog für Gene (ENSEMBL) in Beziehung gesetzt. Diese neu entstandenen Datensätze wurden in die GenomeMatrix eingebunden. Zudem wurde die GenomeMatrix benutzt, um ein System mit Schwerpunkt auf dem menschlichen Chromosome 21 zu entwickeln. Die Chromosom 21-spezifische GenomeMatrix bietet sich als bioinformatisches Werkzeug an, das mit neuen Daten aus internen und externen Quellen beliebig erweitert werden kann.

Die GenomeMatrix koppelt den generierten Pool an Informationen mit einer benutzerfreundlichen Oberfläche. Sie ergänzt damit viele der herkömmlichen Datenbanken und stellt Forschern aus Naturwissenschaft und Medizin ein wertvolles Instrument für die funktionelle Genomanalyse zur Verfügung.