Universität für Bodenkultur

Department für Biotechnologie

Institut für Angewandte Mikrobiologie

Betreuer:

Univ.Prof. Dipl.-Ing. Dr.nat.techn. Karl Bayer

Dipl.-Ing. Dr.nat.techn. Wolfgang Ernst

# Further Development of an automated Inter-species Analysis and Annotation Tool for Agilent Microarrays – iMAT

Diplomarbeit

zur Erlangung des Akademischen Grades

Diplomingenieur

vorgelegt von

**Nora Katharina Nicole Neumann**

März 2010

# KURZBESCHREIBUNG

CHO (Chinese Hamster Ovary) Zellen sind eine der interessantesten und populärsten Säugetier Zelllinien für die Produktion von therapeutisch wichtigen rekombinanten Proteinen.

Der Grund dafür liegt in ihrer Fähigkeit die produzierten rekombinanten Proteine so zu glykosilieren, dass sie eine ähnliche Struktur zu humanen Glykoproteinen aufweisen. Deswegen können sie ohne weitere Bearbeitungsschritte bei Menschen zur Anwendung gebracht werden. Um die Leistungsfähigkeit dieses Expressionssystems zu verbessern, werden Transkriptomanalysen mit Microarrays durchgeführt. Derzeit sind nur begrenzte Sequenzinformationen zu CHO Zellen verfügbar, da das Hamster Genom bis jetzt noch nicht sequenziert wurde und demzufolge auch noch kein öffentlich verfügbares Microarray entwickelt werden konnte.

Es konnte bereits nachgewiesen werden, dass speziesübergreifende Transkriptomanalysen mit bestehenden Microarray Plattformen gut annotierter, genetisch nah verwandter Organismen, wie zum Beispiel Maus (*mus musculus*), wichtige Expressionsdaten erzeugen. In vorangegangen Arbeiten wurde das Anwendungspotential von Maus Microarrays für die Analyse von Hamster Zellen bereits positiv evaluiert. In diesem Projekt wurden zusätzliche Hamster Sequenzinformation, die erst seit kurzem zur Verfügung stehen, verwendet, um die Daten aus vorangegangenen Experimenten zu bestätigen. Nukleotidsequenzen diverser Säugetieren wurden mit der entwickelten Software analysiert, um ein Probenset zu identifizieren, welches als konserviert in den diversen Spezies angesehen werden kann und somit die Möglichkeit einer generischen Microarray Plattform untersucht. Diese herausgefilterte Teilmenge identifizierter Proben wurde

mit den Signalintensitäten der Expressionsanalysen aus Hitzeschock-Experimenten mit Hamsterzellen verglichen und die Korrelations-Koeffizienten der experimentellen Daten ermittelt, um so die Software zu evaluieren.

Die gewonnenen Informationen wurden für die Erstellung einer Skala, auf drei Parametern basierend (iMAT Score, % Sequenzidentität und der Anzahl von aufeinanderfolgenden übereinstimmenden Basenpaaren), genutzt, welche es dem Benutzer ermöglichen soll, die Ergebnisse der Software leichter zu interpretieren. Diese Skala sowie die Software, die in diesem Projekt entstand, können als wertvolles Tool im Bereich der speziesübergreifenden Microarray-Analyse gesehen werden, welches ermöglicht, derzeit erhältliche Microarray Plattformen auf die Anwendbarkeit für derartige Experimente zu untersuchen.

# ABSTRACT

Chinese Hamster Ovary cells are one of the most interesting and popular mammalian hosts for the production of therapeutically important proteins. This is because they produce recombinant glycoproteins that have a similar structure to normal human glycoproteins, therefore they are already biologically active in humans. To improve the overall performance of Chinese hamster cell systems for producing medical proteins, transcriptome analysis is important. Since only limited sequence information about the Chinese hamster is available, a species-specific microarray for public access has not yet been developed. Alternatively, existing microarray platforms for closely genetically related, well-annotated organisms, such as mouse, have been proven to yield valuable expression data in cross-species transcriptome analysis. In previous works, the application of mouse microarrays for analysing Chinese hamster cells has already been evaluated. In this project additional hamster sequence data that became available more recently, were used to add more confidence to the data derived from previous studies. Sequences from other mammals were also used to obtain a set of probes conserved amongst several species, to support the approach of developing a generic microarray chip in the future. Furthermore, the usage of sequence alignment programs, a custom global alignment algorithm (iMAT) and automated annotation in this project resulted in a distinct subset of probes derived from iMAT analysis results. These probes were investigated using signal intensity values from expression analysis using heat shock studies on CHO cells. Finally, the differences in the correlation coefficient values, which were calculated from the experimental signal intensities obtained from these heat shock experiments, were used to create a novel reliability scale based on three parameters (iMAT score, % sequence homology and consecutive

number of matching base pairs). This scale, along with software, deliverable in this project, are invaluable new additions to the field of inter-species microarray analysis, and will help to investigate microarray chips for their feasibility in inter-species experiments.

# ACKNOWLEDGEMENTS

I would like to express my gratitude to Univ.Prof. Dipl.-Ing. Dr.nat.techn. Karl Bayer for the time and support of my application at Cranfield University and the supervision of my thesis at the University of Natural Resources and Applied Life Sciences in Vienna.

I would also like to thank Dr. Wolfgang Ernst and Dr. Friedemann Hesse for accepting me for this project.

Especially, I would like to thank Dr. Wolfgang Ernst for the support and guidance during this project and the time he spent to discuss the project with me either in person or over the telephone.

Similarly I would like to express my appreciation to Dr. Conrad Bessant for his supervision of the project at Cranfield University.

Many thanks to Jennifer Mead for her time to help in the course of the project.

Last but not least I would like to thank Laurie Tonon and Sofia Barreira for their support during the last phase of my project.

# INDEX

KURZBESCHREIBUNG ..... III
ABSTRACT ..... V
ACKNOWLEDGEMENTS ..... VII
INDEX ..... IX
TABLE OF TABLES ..... X
TABLE OF FIGURES ..... XI
1 INTRODUCTION AND LITERATURE REVIEW ..... 13
1.1 BACKGROUND ..... 14
1.2 CHO CELLS ..... 16
*1.2.1 Recombinant protein production ..... 16*
*1.2.2 CHO cell line development and selection ..... 17*
1.3 MICROARRAYS ..... 18
*1.3.1 Gene expression profiling and usage of microarrays ..... 18*
*1.3.2 Hybridisation ..... 20*
*1.3.3 Selection of microarray platform for inter-species experiment ..... 20*
*1.3.4 Agilent 60-mer oligonucleotide microarrays ..... 21*
1.4 BIOINFORMATICS TOOLS ..... 22
*1.4.1 Sequence comparison tools ..... 22*
*1.4.2 Pair wise alignment tools ..... 23*
*1.4.3 Multiple alignment tools ..... 25*
1.5 AUTOMATED SEQUENCE ALIGNMENT, FILE MANIPULATION AND ANNOTATION ..... 25
1.6 STARTING POINT FOR THIS RESEARCH PROJECT ..... 26
1.7 PROJECT OBJECTIVES ..... 28
*1.7.1 Reliability index (Scoring Scale) ..... 29*
*1.7.2 Cross-species conservation ..... 29*
*1.7.3 Automation and User Interface ..... 29*
2 MATERIALS AND METHODS ..... 30
2.1 DATA ..... 30
*2.1.1 Sequence Data ..... 30*
*2.1.2 Experimental Data ..... 30*
2.2 PLATFORM ..... 31
2.3 PROGRAM WORKFLOW ..... 31
2.4 PART 1 – SEQUENCE COMPARISON ..... 32
*2.4.1 BLAST ..... 32*
- *Stand-alone BLAST ..... 32*
- *Workflow of sequence comparison ..... 33*

*2.4.2 Global alignment ..... 35*
2.5 PART 2 – ANNOTATION ..... 37
*2.5.1 Agilent ID annotation ..... 37*
*2.5.2 BLAST/iMAT hits annotation ..... 40*
*2.5.3 Inter-species conserved probe sets ..... 43*
*2.5.4 Global alignment analysis ..... 43*
*2.5.5 iMAT report ..... 46*
2.6 PART 3 – ISC – PROBE SET COMPARISON ..... 47
2.7 GRAPHICAL USER INTERFACE (GUI) ..... 47
2.8 CORRELATION COEFFICIENT ..... 48
3 RESULTS AND DISCUSSION ..... 51

3.1 VALIDATING IMAT WITH MOUSE PROBES AGAINST MOUSE DATABASE ..... 52
3.2 INTER-SPECIES CONSERVED PROBES – CHO – RELIABILITY INDEX ..... 53
*3.2.1 Hamster sequence data:* ..... *53*
*3.2.2 CHO sequence alignment against mouse probes* ..... *54*
*3.2.3 ISC subset selection by annotation matches* ..... *56*
*3.2.4 ISC subsets and experimental data* ..... *58*
*3.2.5 ISC subset annotation and consecutive base pairs > 15* ..... *63*
*3.2.6 ISC subset annotation and homology ≥ 90 %* ..... *64*
*3.2.7 ISC subset based on annotation, iMAT score, % sequence homology and presence of consecutive base pairs > 15* ..... *65*
*3.2.8 ISC subset based on iMAT score, homology, consecutive base pairs* ..... *66*
3.3 CROSS – SPECIES SEQUENCE ALIGNMENT ANALYSIS ..... 69
*3.3.1 Validating iMAT with mammalian and rodent databases* ..... *69*
*3.3.2 Mouse probes versus rat, human and hamster* ..... *74*
*3.3.3 ISC subsets based on annotation matches and cross-species sequence analysis 76*
3.4 IMAT – GRAPHICAL USER INTERFACE ..... 79
*3.4.1 Sequence comparison tab* ..... *80*
*3.4.2 Annotation tab* ..... *83*
*3.4.3 Inter-species conserved sets tab* ..... *84*
4 CONCLUSION AND FUTURE WORK ..... 87
4.1 SUMMARY OF THE PROJECT AIMS MET AND DISCUSSION OF THE RESULTS ..... 87
4.2 CHALLENGES ..... 89
4.3 FUTURE WORK ..... 90
5 REFERENCES: ..... 91
A. APPENDIX FOR ISC SUBSET SELECTION ..... 95

# TABLE OF TABLES

TABLE 1-1 OVERVIEW OF BLAST VARIANTS AVAILABLE ..... 24
TABLE 3-1 MOUSE VS. MOUSE SEQUENCE ALIGNMENT AND THE FOUND HITS IN THE UNIGENE DATABASE ..... 52
TABLE 3-2 HAMSTER SEQUENCE ALIGNMENT FOUND HITS RESULTS AT AN E-VALUE OF 1 AND 100 000 ..... 55
TABLE 3-3 HAMSTER-ANNOTATION MATCHES SUBSET OBTAINED WITH GENE NAME COMPARISON ..... 56
TABLE 3-4 IMAT SCORE GROUPS ..... 57
TABLE 3-5 COMPARISON OF IMATSCORE, % HOMOLOGY WITH AND WITHOUT CONSECUTIVE BASE PAIRS ..... 68
TABLE 3-6 MAMMALIAN DATABASES THAT WERE SELECTED TO BE ALIGNED AGAINST MOUSE PROBES ..... 70
TABLE 3-7 SEQUENCE ALIGNMENT HUMAN, RAT, HAMSTER ..... 70
TABLE 3-8 SEQUENCE ALIGNMENT RODENTS ..... 71
TABLE 3-9 SEQUENCE ALIGNMENT PRIMATES ..... 71
TABLE 3-10 SEQUENCE ALIGNMENTS OF RAT, HUMAN AND HAMSTER AGAINST MOUSE PROBES AT AN E-VALUE OF 100,000 ..... 74
TABLE 3-11 FOUND PROBES IN ALL THREE SPECIES ..... 74
TABLE 3-12 ISC SUBSETS RESULTS AND CONSERVED PROBES RAT, HUMAN, HAMSTER ..... 76
TABLE 3-13 PARAMETER ANALYSIS OF CROSS-SPECIES CONSERVATION BASED ON ANNOTATION MATCHES ..... 77

# TABLE OF FIGURES


FIGURE 2-1 FORMATDB TERMINAL ENTRY: EXAMPLE COMMAND FOR EXECUTING FORMATDB ..... 33
FIGURE 2-2 OLIGONUCLEOTIDE SEQUENCE FILE: EXAMPLE OF NECESSARY FORMATTING ..... 33
FIGURE 2-3 STANDALONE BLAST WORKFLOW IN IMAT, WITH OLIGONUCLEOTIDE FILE AND ORGANISM DATABASE AS THE INPUT DATA DELIVERING THE FIRST IMAT RESULTS FILE ..... 34
FIGURE 2-4 GLOBAL ALIGNMENT WORKFLOW ..... 36
FIGURE 2-5 WORKFLOW OF THE AGILENT ID ANNOTATION TAKING THE AGILENT PROBE IDS FROM THE IMAT RESULTS FILE AND PRODUCING A AGILENT ID ANNOTATION FILE ..... 39
FIGURE 2-6 REPORT FILE OF THE ANNOTATED HITS FROM THE IMAT RESULTS FILE ..... 41
FIGURE 2-7 WORKFLOW OF UNIGENE ID/ ENSEMBL ID ANNOTATION TAKING THE IMAT HITS FROM THE RESULTS FILE AND PRODUCING A REPORT FILE AS SHOWN IN FIGURE 2-6 AND AN ADDITIONAL ISC PROBE SET FILE ..... 42
FIGURE 2-8 GLOBAL ALIGNMENT ANALYSIS USING THE IMPROVED IMAT ALGORITHM ..... 45
FIGURE 2-9 EXAMPLE GRAPHICAL OUTPUT FOR ALIGNMENT ANALYSIS. THIS OUTPUT IS INTENDED TO GIVE THE USER AN OVERVIEW OF THE RESULTS OBTAINED IN THE ALIGNMENT ANALYSIS STEP ..... 46
FIGURE 3-1 MOUSE VS. MOUSE HIT DISTRIBUTION, GRAPHICAL OVERVIEW OF THE NUMBER OF PROBES THAT FALL INTO THE IMAT SCORE CATEGORIES. ..... 52
FIGURE 3-2 ISC SUBSET BASED ON ANNOTATION MATCHES AT AN E-VALUE OF 1 AND ITS PROBE DISTRIBUTION BASED ON IMAT SCORE GROUPS ..... 57
FIGURE 3-3 ISC SUBSET BASED ON ANNOTATION MATCHES AT AN E-VALUE OF 100,000 AND ITS PROBE DISTRIBUTION BASED ON IMAT SCORE GROUPS ..... 57
FIGURE 3-4 SCATTER PLOT OF SIGNAL INTENSITIES OF ISC SUBSET BASED ON ANNOTATION MATCHES AND AN IMAT SCORE 180 AT AN E-VALUE 1 ..... 59
FIGURE 3-5 SCATTER PLOT OF SIGNAL INTENSITIES OF ISC SUBSET BASED ON ANNOTATION MATCHES AND AN IMAT SCORE 180 AT AND E-VALUE 100,000 ..... 59
FIGURE 3-6 LOG2 PLOT OF SIGNAL INTENSITIES OF GENES WITH IMAT SCORE 180-150 E-VLAUE 1 ..... 60
FIGURE 3-7 LOG2 PLOT OF SIGNAL INTENSITIES OF GENES WITH IMAT SCORE 180-150 E-VALUE 10000 ..... 60
FIGURE 3-8 RELIABILITY SCALE BASED ON ESTABLISHED PARAMETERS OF IMAT SCORE, % SEQUENCE HOMOLOGY AND PRESENCE OF CONSECUTIVE BASE PAIRS ..... 68
FIGURE 3-9 INTERSECTION OF IMAT HITS OF ALL SPECIES EXCEPT MOUSE ..... 72
FIGURE 3-10 SEQUENCE COMPARISON TAB: PROJECT NAME, E-VALUE, BLAST HITS ENTRY FIELDS ..... 80
FIGURE 3-11 SEQUENCE COMPARISON: OLIGONUCLEOTIDE FILE SELECTION, ORGANISM DATABASE ..... 81
FIGURE 3-12 EXAMPLE OF THE META_DATA.TXT FILE WHICH CONTAINS ALL NECESSARY PATHS TO THE ORGANISM DATABASES ..... 82
FIGURE 3-13 ADD NEW DATABASE CONFIRMATION ..... 82
FIGURE 3-14 ADD NEW DATABASE CONFIRMATION MESSAGE ..... 83
FIGURE 3-15 ANNOTATION TAB ..... 83
FIGURE 3-16 ISC - COMPARISON ..... 84
FIGURE 3-17 EXAMPLE OF INTERACTIVE HELP WINDOW IN IMAT ..... 86
FIGURE A-1 SCATTER PLOT OF ISC SUBSET BASED ON MATCHING ANNOTATION AND CONSECUTIVE BASE PAIRS LENGTH > 15 ..... 95
FIGURE A-2 SCATTER PLOT OF ISC SUBSET BASED ON MATCHING ANNOTATION AND SEQUENCE HOMOLOGY ≥ 90% AT AND E-VALUE OF 100000 ..... 95
FIGURE A-3 ISC SUBSET BASED ON ANNOTATION MATCHES AND IMAT SCORE ≥150, ≥ 90% HOMOLOGY, > 15 BASE PAIRS ..... 96
FIGURE A-4 ISC SUBSET BASED ON ANNOTATION MATCHES, IMAT SCORE ≥ 70, SEQUENCE HOMOLOGY ≥ 60 % AND A CONSECUTIVE BASE PAIR STRETCH > 15 NUCLEOTIDES ..... 96
FIGURE A-5 ISC SUBSET SELECTION BASED ON IMAT ≥150, %HOMOLOGY ≥ 90, CONSECUTIVE BASE PAIR STRETCH > 15 ..... 97
FIGURE A-6 ISC SUBSET SELECTION BASED ON IMAT ≥ 70, % SEQUENCE HOMOLOGY ≥ 60 %, CONSECUTIVE BASE PAIR STRETCH > 15 ..... 97

# 1 Introduction and Literature Review

Regulation of gene expression plays an important role in controlling biological processes in living cells. Cell development and the associated biochemical processes are determined by the cellular proteomes and the proteome of a cell is regulated by gene expression. The transcriptome can be regarded as an indirect "readout" of the proteome, and offers information on the biochemical status of a cell (Chen *et al.*, 2006).

Modern gene expression analyses are commonly carried out with microarray experiments allowing monitoring the whole transcriptome of a given organism at a certain time point, under certain conditions.

Today, the availability of microarray chips is still limited to model organisms, mostly human and rodents. Development of high density DNA microarrays is not only complex, expensive and time consuming but also requires considerable knowledge of the genomic sequence for the species of interest (Ernst *et al.*, 2006).

Due to these restrictions, one alternative is to use commercially available microarrays of closely related species for global gene expression studies to yield highly valuable data without the species-specific arrays.

Indeed, previous studies have already revealed high sequence conservation within mammals and especially within rodents (Makalowski and Boguski, 1998).

One aim of this project is to create a fully automated analysis workflow, including a user interface, sequence alignment, annotation and alignment analysis iMAT[1] results, to provide an easy to use tool that enables the researcher to find suitable microarray platforms for inter-species experiments, when no commercially available microarrays are to hand.

[1] iMAT: Inter-species microarray analysis tool

Another aim of this project is to validate the results of the previously developed tool iMAT with the newly available CHO[2] sequences as well as with sequence information of other mammalian databases.

## 1.1 Background

Recombinant protein therapeutics provide innovative and effective therapies and are used today to treat different human diseases. The production of recombinant protein therapeutics relies on the fact that they must be synthesised in their biologically active from. This requires post-translational modifications, such as glycosylations. The necessary glycoproteins are synthesised only in mammalian cells lines. Other popular microbial hosts lack the cellular machinery to achieve this.

Since the establishment of tissue culturing for CHO cells, they have been used in several biomedical studies, ranging from cell cycle to toxicology studies, for example (Jayapal *et al.*, 2007).

In 1957, Dr. Theodore T. Puck first isolated an ovary from a female Chinese hamster and established the first cell line on culture plates. During that time the low chromosome number of Chinese hamsters (2n=22) was considered especially useful for tissue culture studies. Although there have been major advances in cell line development, cell selection still remains empirical due to the large variation between experiments and lack of understanding of mammalian cell culture processes, such as underlying cytogenetic events. One cannot predict how clones that were selected and characterized on bench top bioreactors will behave in large-scale bioreactors.

These difficulties can be explained because of our limited knowledge of the biology and physiology of this mammalian cell line. Transcriptome analysis offers one

[2] CHO: Chinese Hamster Ovary

approach to understand underlying regulatory mechanisms as well as improving the overall performance of the CHO expression system.

Although CHO cells are widely used as host cells for protein expression, the whole genome of Chinese hamsters has not yet been sequenced, and therefore no whole genome microarray has been developed (Jayapal *et al.*, 2007).

As an alternative for species-specific microarrays to study gene expression profiles, commercially available arrays for mouse or rat have been proven to be feasible for CHO hybridisation experiments. The high sequence conservation among mammals and especially within rodents has already been proven and result in sufficient probe signal intensities (Ernst *et al.*, 2006).

Since 2006, over 27,000 unique non-overlapping transcript sequences of the CHO Genome were identified. Although these sequences were used for the creation of a proprietary CHO DNA microarray by the CHO Consortium (Jayapal *et al.*, 2007), the need for a way to analyse data from the cross-species microarray approach still is strong, as more sequences became available since 2007 and the hamster genome has not yet been sequenced.

This project is therefore aiming to provide an easy-to-use bioinformatics tool to select a suitable microarray platform for cross-species analysis, to meet this need.

## 1.2 CHO Cells

This section is intended to highlight the favourable features of CHO cells, as hosts for recombinant protein production, as well as the importance of CHO cells in medical applications. Furthermore light is shed on how CHO cell lines are selected.

### *1.2.1 Recombinant protein production*

The selection of host cells for recombinant protein production has an important influence on the desired product. The ability of the host cells to fold proteins correctly and express proteins with post-translational modifications means that the protein products are suitable in terms of their solubility, stability, biological activity and safety for humans (Jayapal *et al.*, 2007).

For the production of recombinant proteins, as well as any other products, economy and quality of production procedures and efficacy of the platform, play an important role. Pressure to find the most suitable expression systems is becoming more and more important, as systematic genomics research increases the number of possible gene targets (Gellissen *et al.*, 2005). Especially microarray analysis are regarded as a central analysis step, making this project even more interesting in terms of finding the best microarray platform for species for which microarrays are not commercially available.

Although mammalian cells are more demanding in terms of cultivation than bacterial expression systems like *E.coli*, mammalian cells are the preferred platform for therapeutically active proteins for administration in humans.

Also, compared to other eukaryotic platforms such as yeast, which are also capable of modifying recombinant proteins, only mammalian cells have the ability to glycosylate the proteins in an authentic structure (Sandig *et al.*, 2005).

### *1.2.2 CHO cell line development and selection*

During the early cell biology studies of CHO, particular mutants deficient in the enzyme dihydrofolate reductase (DHFR) with auxotroph nutritional requirements were identified (Urlaub and Chasin, 1980).

To obtain a high yielding CHO cell line from a variety of parental lines, mostly the DHFR[3] selection system is used. It allows selection of stable clones as well as acceptable gene amplification. Gene amplification is provided when CHO cells are cultivated in presence of methotrexate (MTX), a folic acid analogue. It blocks the DHFR activity; therefore the cells react with an increased expression of DHFR for survival. Ensuring the cell had a DHFR containing gene construct, co-amplification of this transfected gene is provided (Kaufman *et al.*, 1983).

After amplification, the clone with highest productivity and growth rates, as well as best product quality, needs to be isolated and evaluated in lab size reactors.

Selected clones should fulfil various requirements such as

- Low nutritional requirements but high growth rate
- High product yield
- Safe and stable products
- Ability to grow in suspensions/bioreactors
- Low cell mortality
- Desired glycosylation

[3] DHFR is a monomeric enzyme, that mediates the transformation of folic acid to tetrahydrofolate(THF)

Extensive screening, medium optimization and process monitoring and control to raise the productivity levels aid the cell line development.

Previous studies focused on

- the optimisation of glycosylation patterns, which are similar to those of humans (Jenkins *et al*., 1996) and therefore fully biological functional.
- Apoptosis engineering, cell cycle engineering and metabolomic pathway engineering to enhance productivity of mammalian cells (Kuystermans *et al*., 2007)

## 1.3 Microarrays

This chapter will describe the use of microarrays in the context of gene expression profiling and the inter-species approach applied in this project.

### *1.3.1 Gene expression profiling and usage of microarrays*

Gene expression analysis aims to measure the expression of thousands of genes simultaneously for one population at a particular point in time. Microarray[4] technology reveals underlying genetic mechanisms, such as up and down regulation (Southern, 2001, Brown and Botstein, 1999).

There are different variations of this technique but all use the attachment of a large number of probes[5] (spots) to a solid surface, which represent either a whole genome, or a specific subset of genes.

[4] a collection of microscopic samples arranged in an orderly manner attached to a solid surface.
[5] Immobilised nucleic acid known sequence on the chip

Microarrays are either used to provide qualitative (detection of sequences) or quantitative (measure expression levels of genes) information. There are two main technologies available:

- **Spotted microarrays:**

  500-1000 base pairs (cDNA[6])/25-100-mers (oligo) immobilised to a surface using robot spotting up till 80,000 spots per slide

  "dual-channel" or two –colour microarrays

  Two samples, differently labelled are hybridised to the same slide and the relative expression levels are detected. Probes can be either oligonucleotides, cDNA or fragments of PCR[7] products

- **Oligonucleotide microarrays:**

  18-80-mers are immobilised *in situ* (on-chip) or with other methods such as photolithography

  In case of one-colour hybridisation, only one sample is hybridised onto the microarray and absolute expression levels are measured. Probes are complementary mRNA[8] sequences.

  Agilent employs this technique to produce chips with 50-60 base pairs in length by in situ synthesis. Each spot then represents one gene or gene region.

  Affymetrix, the second most popular manufacturer of oligonucleotide microarrays, produces arrays with 25 base pair length. One gene is split up into11-20 25-mer probes, therefore one spot only represents one small part of the gene or the gene region.

---

[6] cDNA=complementary DNA
[7] Polymerase chain reaction
[8] mRNA: generated from the transcription of a cDNA template. Mature RNA means, that introns were spliced out and it serves as the template for protein translation.

Each probe has a target that it should bind specifically during hybridisation[9]. Targets are labelled with either a detectable molecule or a form of dye, mostly fluorophores. The signal emitted gives a value of expression of the gene, if it contains the target sequence. Regardless of which array platform is used, both serve the purpose of binding a specific sequence (Jaluria *et al.*, 2007).

For this project Agilent 60-mer oligonucleotide microarrays are relevant.

### *1.3.2 Hybridisation*

Hybridisation results are critical for the outcome of the gene expression analysis with microarrays.

Hybridisation is dependent on various parameters, such as length of nucleic acid sequence (the longer the sequence the less likely a cross-hybridisation will take place), percentage of homology (in case of cross-species hybridisations) and the type of nucleotides that are involved in forming the hydrogen bonds (G-C bonds are more stable then A-T bonds). But also preparation of the samples, as well as preparation of the microarray itself, influences the hybridisation process aside from the thermo dynamic challenge.

### *1.3.3 Selection of microarray platform for inter-species experiment*

"Inter-species" and "cross-species" is an approach where microarray platforms of closely related species are used for gene expression profiling experiments. In this project microarrays of mouse, which is closely related to Chinese hamster were employed. The inter-species approach relies on the assumption that closely related species have conserved transcripts. Therefore, microarrays for mouse should be able to detect their orthologs in Chinese hamster (Wang *et al.*, 2004).

[9] Hydrogen bond between two single stranded nucleic acid sequences

Whole genome chips consist of the most variable regions of a particular gene to make the oligonucleotide array as species-specific as possible, it follows that closely related species that share high sequence similarity will also differ in exactly those variable regions. However, cross-species hybridisation results must be analysed with caution keeping the possibility of false-positive results in mind.

Recent studies have shown that microarray platforms require a specific length of perfect matches between targets and probe to yield a specific hybridisation signal (Ernst *et al.*, 2006, Yee *et al.*, 2008). When comparing both platforms it seems that the longer the oligonucleotide sequence on the chip the more likely a specific hybridisation signal will occur.

Other studies investigated the response of different mouse and hamster cell lines to the same stimuli, highlighting similarities and differences in the response (De Leon Gatti *et al*., 2007). It showed that mouse hybridoma cells and CHO cells can be regarded as responding in a similar way to the treatments.

### *1.3.4 Agilent 60-mer oligonucleotide microarrays*

A comparison of gene coverage of different microarray platforms in 2006 revealed that differences in coverage were highly conserved across the chromosomes (Verdugo and Medrano, 2006).

For this project the sequence information of Agilent's 60-mer oligonucleotide arrays were used, as well as CHO sequence information from the CHO consortium.

The 60-mer oligonucleotides are synthesised with Agilent's SurePrint technology, a non-contact inkjet printing process. This platform is suitable for various applications such as gene expression analysis, where either one or two colours are used. Another application can be comparative genomic hybridisation analysis. On the more recent chip that is available more than 41,000 mouse genes and transcripts are immobilised

including public domain annotations. Additional content is available for example from UCSC, RefSeq, Ensembl and UniGene databases, to name a few.

Probe selection is verified using NCBIs Genome build 32 and probes are validated with Agilent's laboratory validation process (Agilent, 2009).

## 1.4 Bioinformatics tools

This section gives more information on the bioinformatics approaches for inter-species microarray data analysis, focussing on the sequence information.

Data derived from microarray experiments must be further analysed to view the biological meaning of those results, when using the cross-species approach.

The inter-species microarray experiment (as for any other microrarray experiment) produces two main types of data:

- Nucleotide sequences of the oligonucleotide probes and the gene sequence of the applied samples.
- Hybridisation signal intensities measured in the form of fluorescence signal of the labelled targets in the microarray experiment. These values relate to the level of hybridisation of probe and target.

Bioinformatics techniques can help to identify inter-species homology and the similarity between targets and probe sequences. Furthermore, statistical methods can be used for filtering microarray data or for calculating the probability of hybridisation, as well as for extracting relevant information from the data.

### *1.4.1 Sequence comparison tools*

For this project sequence comparison will be used to find similarities between sequences in closely related species. Available algorithms look for similarities rather than exact matches, therefore it is possible to find orthologous gene sequences. Gene

sequences are called orthologs, when descending from a common ancestor. Scoring matrices are employed to calculate the similarity between two or more aligned sequences.

For sequence comparison, pair wise alignment tools and multiple alignment tools such as BLAST[10], GAP[11], ClustalW, and ClustalX (Altschul *et al.*, 1990, Needleman and Wunsch, 1970, Larkin *et al*., 2007, Thompson *et al*., 1994) are available.

### *1.4.2 Pair wise alignment tools*

Pair wise alignment tools search for the highest possible score and work either on a local (BLAST) or on a global sequence level.

Local alignment tools are used to calculate the optimal similarity between sub regions of the sequences (Frazer *et al*., 2003). They are based on the Smith and Waterman Algorithm, a further development of the Needleman and Wunsch algorithm, which was created for global alignments. Global alignments calculate the optimal score of two compared sequences over their entire length (Frazer *et al*., 2003).

Both of the mentioned algorithms are based on dynamic programming approaches that tend to be very time-consuming and computationally extensive.

Sequence alignment tools used more frequently today are based on a heuristic approach like BLAST and BLAT, which break sequences into short words, compare the sequences and extend them to high score alignments until the substitution matrix score decreases again. These two approaches differ in the way of scanning through the sequences, in their way of extending the high score alignments of the 3 letter words and how they handle their alignments (Kent, 2002).

---

[10] BLAST: Basic local alignment tool

[11] GAP: Global Alignment Program

Nevertheless, BLAST is still the most popular tool used today for sequence comparisons and, more importantly, for this project for homology detection between oligonucleotide probes and transcripts (Adjaye *et al.*, 2004, Wang *et al.*, 2004).

**Table 1-1 Overview of BLAST variants available**

| | |
|---|---|
| **blastp** | Compares protein sequence against a given amino acid sequence database |
| **blastn** | Compares a DNA sequence against a given nucleotide sequence database |
| **PSI-blast** | Position-Specific Iterative Blast: for finding distant protein relatives |
| **blastx** | Compares a in all reading frames translated nucleotide sequence against a protein database |
| **tblastn** | Compares a protein sequence against a database of in all 6 reading frames translated nucleotide sequences |
| **tblastx** | Compares a in all reading frames translated nucleotide sequence against a dtabase of in all 6 reading frames translated nucleotide sequences to find very distant relationships |
| **megaBLAST** | Used for comparing large number of query sequences, faster than BLAST |

For this project the blastn variant was used to compare oligonucleotide and transcript sequences. Table 1-1 gives an overview of all available BLAST variants.

Apart from these variants, many specialised variants are available - further reference is available at the NCBI website (NCBI-Blast,May 2009).

Regardless of which BLAST variant was used, a list of hits (if found), together with the chosen hits score and E-value, is shown as the result.

The score is calculated as the sum of gap "penalty" scores according to the substitution matrix. The "E-value" or "Expectation value" reports the number of hits that occur just by chance when searching against the database. It is a statistically significant threshold. Where the lower the E-value, the more significant the alignment. The E-value not only takes the length of the sequence into account but also the size of the database (it gets higher the larger the database, gets lower the longer the query sequence).

### *1.4.3 Multiple alignment tools*

Multiple alignment tools like Clustal W (Thompson *et al*., 2002) and T-Coffee (Notredame *et al.*, 2000) are able to compare more than just two sequences, especially if they are unknown sequences and highlight their similarities. Multiple alignment tools can also help in the decision as to which microarray platform could be used in a cross-species experiment.

They can find information about phylogenetic relationships between different species. Multiple sequence alignment works in a similar way as pair wise comparison, but instead of dynamic programming mostly heuristic approaches to compare all sequences in combination with hierarchical cluster analysis are employed.

The described tools for pair wise and multiple alignments are available as a web based version or downloadable stand-alone version to run batch comparisons on a local machine.

## 1.5 Automated sequence alignment, file manipulation and annotation

Perl[12] was first introduced in the 1980s. It is a stable cross-platform programming language and licensed under the GNU General Public License (GPL). It was developed by Larry Wall and intended mainly for text manipulations and parsing[13], hence it is useful for sequence analysis programs, like iMAT in this project.

It is a multi-purpose interpreted language for various tasks such as systems administration, web development and graphical user interface (GUI) development. It is extensible due to various third party modules available over CPAN (Comprehensive Perl Archive Network) and supports procedural and object-oriented Programming (The Perl Directory, May 2009).

---

12 Practical Extraction and Report language

13 Parsing: extracting meaningful information from a text

Bioperl is a collection of more than 500 Perl modules that cover vast areas of bioinformatics. Bioperl modules are written object oriented and an open source project maintained by international volunteers. (Bioperl, 2009)

BioMart offers an easy to use solution to query biological databases to gather additional biological information on for example probe IDs on microarray chips. It can also be easily accessed via a Perl API[14] (Smedley *et al*., 2009).

## 1.6 Starting point for this research project

The work of this project is based on previous students work in 2005 and 2006 (Mead, 2005, Güzlek, 2006).

During these studies, Mead and Güzlek investigated the initial assumption of the feasibility of using Agilent's 60-mer oligonucleotide microarrays for cross-species transcriptome analysis. Other studies also reported, that a minimum sequence identity of ≥ 16 base pairs is required to yield a successful hybridisation signal (Kane *et al*., 2000).

In the course of their projects and with additional work by Andreas Schlattl (an internship student at the ACBT Working group at the University of Applied Life Sciences and Natural Resources in Vienna), part of iMAT[15] was developed using standalone BLAST and an additional global alignment, where the 60-mer sequences were aligned across their whole length against obtained BLAST hits.

This resulted in an output file that reported the iMAT score, which was assigned to each globally aligned probe against BLAST hits and could reach maximum 180. Also UnigeneIDs, GenbankIDs, Blast Score, E-value and the number of totally found perfect matches are reported.

[14] API: Application programming Interface, set of routines and data structures provided.

[15] iMAT: Interspecies microarray analysis tool

This previous iMAT software has been further improved in this MSc project. Currently the iMAT score serves as the most significant parameter for the inter-species data, and has been given priority in this project as is explained later.

## 1.7 Project objectives

Based on talks with the supervisors the following specifications for the program were established, which led to the project objectives as described below (1.7.1, 1.7.2, 1.7.3):

- Provide additional information for the user such as % homology between two globally aligned sequences and information on consecutive base pairs to indicate a possible outcome of a cross-species microarray experiment. Also to give additional information on the iMAT score.

- Automate gene annotation for Agilent IDs and Unigene IDs or Ensembl Transcript IDs. Find gene names and additional unique IDs for the given IDs in sequence comparison file.

- Create a GUI for the program, not only for input, but it should also keep the user informed throughout the process. It should also be platform independent

- The developed tool should allow future addition of different Agilent oligonucleotide microarrays sequence files and addition of downloaded organism databases, either from NCBI or Ensembl, where necessary.

- One result file should be created containing annotated Agilent IDs and the best matched hit from the transcript database including its annotation, iMAT score, consecutive base pair information, and % homology between Agilent oligonucleotide sequence and the highest scoring hit. Additional files with all available information on the hits should still be provided for the user as additional and detailed information.

### 1.7.1 Reliability index (Scoring Scale)

The first aim of the project is to create a reliability index that integrates additional data, such as: gene annotation (finding gene names); percent of homology (identities) between two sequences; the degree of redundant transcripts that are found for every probe and the indication of consecutive base pairs in the two matching sequences.

### 1.7.2 Cross-species conservation

In addition, as since 2006 as a result of the work by the CHO Consortium, the number of available CHO transcripts through the CHO consortium nearly tripled. iMAT will therefore be used to determine the inter-species homology with these new CHO transcripts by aligning CHO transcripts and mouse probes from Agilent.

At first iMAT is going to be validated by aligning the available mouse oligonucleotide probe sequences against the UniGene Mouse database, and then against other mammal databases.

### 1.7.3 Automation and User Interface

One aim of the project is to automate the whole workflow to enable a reasonable time frame for obtaining reliable probes for any given species for which no commercially available microarrays exist and to create a more user-friendly interface for iMAT. This component of the project is a software deliverable for the bench scientist to use.

Although this project focuses on the usability of mouse oligonucleotide microarray for hybridising hamster probes, iMAT also has the potential to help in determining the feasibility of human oligonucleotide microarrays for related primate species, for example.

# 2 Materials and Methods

In this chapter the employed methods and algorithms will be mentioned. Selected platforms will be described as well as the obtained result data.

## 2.1 Data

### *2.1.1 Sequence Data*

Agilent oligonucleotide sequence data from Agilent G4121B mouse chip was used to query different rodent and mammalian databases. Sequence databases were downloaded mainly from NCBI UniGene (ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene) and unspliced transcript databases from Ensembl with BioMart (http://www.ensembl.org/biomart/martview/045d51431cefad1fa2ceed16179bc333).

UniGene offers a 'clustered view' of an organism's transcriptome. Each UniGene entry represents a set of transcript sequences that seem to come from the same transcription locus, either gene or expressed pseudogene. In addition, information on protein similarities, gene expression and genomic location is provided (NCBI-UniGeneAugust 2009).

Ensembl BioMart offers access to unspliced transcript sequences for the available species. Additional information, apart from the sequence, can be selected individually. In this project, Ensembl Gene ID and Ensembl Transcript ID were used.

For the hamster sequence information, a new sequence database in FASTA format was created. A detailed description will be mentioned later.

### *2.1.2 Experimental Data*

Experimental microarray signal data was used from experimental setup as described by (Güzlek, 2006).

## 2.2 Platform

The whole project was carried out on a MacBook Pro with an Intel Core 2 Duo 2.8 GHz Processor and 4 GB RAM. The program was written in Perl (version 5.8.8 preinstalled). For automated sequence comparison, annotation, file parsing and creating the GUI different Perl modules, such as Perl Tk and BioPerl were used.

## 2.3 Program Workflow

A general program setup was developed consisting of three major parts, 'Sequence Comparison', 'Annotation' and 'ISC[16] probe sets'. The first part is based on previous scripts developed by Jennifer Mead (2005), Hacer Gülzek (2006) that were assembled by Andreas Schlattl (2006). Parts of this script have been used and modified to fit the project objectives of this MSc project.

In the following subchapters (2.4, 2.5, 2.6) the software setup of iMAT will be explained in more detail, including a detailed description of each individual step of the program (Sequence Comparison, Annotation and ISC probe sets). Furthermore the Workflow of the program steps as well as the outcome files will be presented as well as their utilisation within the program.

[16] ISC = inter-species conserved

## 2.4 Part 1 – Sequence Comparison

The first part of the program calculates the sequence alignments using conventional sequence alignment algorithms (BLAST) in combination with a custom global alignment (2.4.2).

### *2.4.1 BLAST*

As already mentioned in the introduction section, BLAST is one of the most popular heuristic approaches for a local alignment of two sequences. It is reckoned to be faster than dynamic programming approaches. The power and effectiveness of BLAST lies in its capability to break the DNA sequence into small substrings and therefore it deals with small sections of a sequence at a time.

BLASTN, which is used in this project for aligning the microarray sequences against hamster transcripts and other sequence databases, is designed to compare nucleotide sequences.

- ***Stand-alone BLAST***

For this project the stand-alone version (version 2.2.20 – May 2009) from NCBI BLAST was downloaded via ftp (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST).

This file package contains several executable files, blastall and formatdb (both Unix Executable Files) being the most important of them for this project.

Formatdb (Figure 2-1) is used to format protein or nucleotide databases into a BLAST-friendly format. If no preformatted databases are available those databases can be either in ANS.1 or FASTA format but may not exceed a specific size of more than 4 billion letters (Research Computing Center,June 2009, NCBI-formatdb,August 2009).

```
mitch-199-205:Hs_seq master$ formatdb -i Hs.seq.all -p F -v 2000
```

**Figure 2-1 formatdb terminal entry: example command for executing formatdb**

Different options and more detailed description can be accessed online. (http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/formatdb_fastacmd.html)

- ***Workflow of sequence comparison***

The oligonucleotide sequence data has first to be transformed into tab delimited text (.txt) files (Figure 2-2), including the probe identifier from the microarray and its sequence. It has been agreed with the client to leave this step manual and not include it in the program workflow, because the format of the original sequence file released by Agilent varies in every release version.

| ID | Sequence |
|---|---|
| A_51_P401471 | GGGCTTAATTATTACCAAATTCTTAGAAGCTGTGTCTCCCAGACTGTAACCATTGAAGAA |
| A_51_P315841 | GCTCCCTGTCTAAGTGGTAAGGTGGGGATTGTCTCCATCTTTGTCATAATAAAGCTGAGA |
| A_51_P437938 | TTATAGAAGATCCATGGGACTAAACAACTATGGGCTAAGAATGTGTCTGGGGAATTACCT |
| A_51_P322989 | CATCCTGAAGCATTGTGGGTTCCTTTCAATGTTGTTATACTCTTCCCCTCTAGTTATGGT |
| A_51_P110068 | GAATAAGGCATTTCTCTATTGTTTTGAGGGGGCCTATGGTAAATCAAATTAACCTACCCC |
| A_51_P326529 | CAAAACCATAGAGTCTGTTTTCCAGTAGTCTTGATTCGTATAAAATAATGACTTCCTTCC |
| A_51_P515158 | TCCAGCTCTTTCCAAAAGACGATGTCACACACAGCCTCTCTAGCAGTCTTACTGAAGATT |
| A_51_P517662 | GTGTGACTATCACCGTTAGAGCTGTTATTTTTATGACTCCTTTGAGTTGGATGTTGAGTC |
| A_51_P396875 | AAAGCAAGGAACTCCTTCTACCCTTAGAATTTCCAACATTCCTTCTTATTCATCAGCTGC |

**Figure 2-2 Oligonucleotide sequence file: example of necessary formatting**

The sequence database is downloaded in FASTA format from NCBI or Ensembl and prepared for BLAST by applying formatdb and setting the files in the correct location on the computer ie. correct folder location. The formatdb step is now included in the program workflow. The user is than able to specify different program settings through a GUI, which is discussed in more detail later.

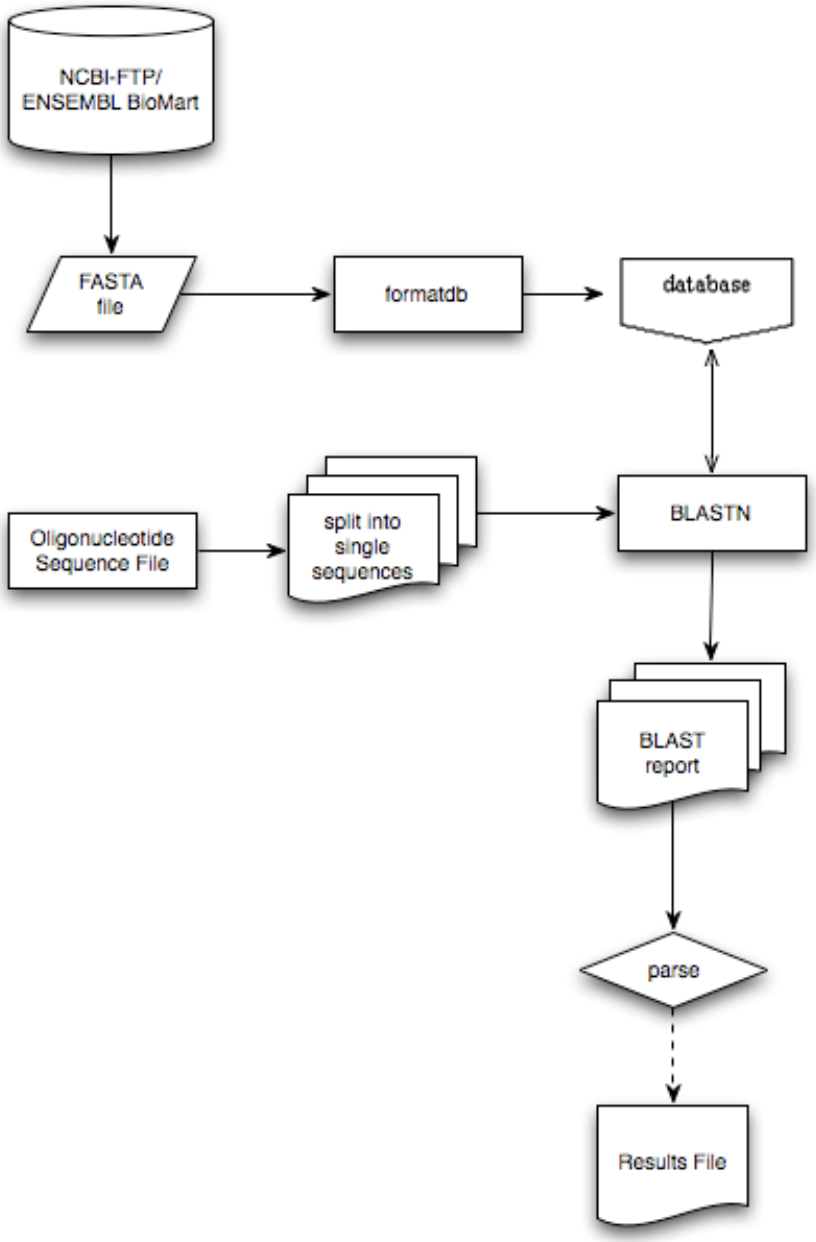


**Figure 2-3 Standalone BLAST workflow in iMAT, with oligonucleotide file and organism database as the input data delivering the first iMAT results file**

As shown in Figure 2-3 single sequences from the oligonucleotide file are aligned against the sequences from the downloaded, formatted database to find the best match between as many nucleotides as possible. It is matched on a local level of the sequence, not over its whole length. A BLAST report for each individual query sequence is created by BLAST with the probe ID as the identifier, and then it is parsed for further use in the program. The purpose of the BLAST step in this program

is to indentify as many sequence matches between the query Agilent oligonucleotide sequence and the transcript sequences of a downloaded organism database as possible. The quality of the local alignment of the two sequences depends on the user-specified criterion (E-value).

### *2.4.2 Global alignment*

In this step the best BLAST hits (meaning the highest scoring below a set E-value) are taken and a global alignment over all 60 nucleotides of each query sequence (Agilent oligonucleotide sequence) and it's found BLAST hit sequences (subject sequences) is performed. Andreas Schlattl developed this algorithm during his internship. During this step, three different score matrices are generated: one for both sequences and one for each sequence individually.

In the first step, perfect matches between the nucleotides are rewarded with a score of +3, whereas gap initiations, gap elongations and mismatches are penalised with -1, resulting in an additional matrix of matches.

Through calculation of the maximum value from the comparison of two given scores from two initial matrices at a time (and taking gap penalties into account), three new matrices with scores for the matches are created.

Each maximum value from those three matrices (given three values at a time) is taken for each possible nucleotide alignment. These are compared to each other resulting in a final highest iMAT score for each possible alignment.

This iMAT score assumes a maximum possible value of 180, being the equivalent to 60 perfect matches between the two sequences. The final global alignment between the two sequences is parsed into separate alignment files, one for each oligonucleotide probe on the microarray, and is named after the Agilent identifier on the microarray.

This alignment file is later used in the alignment analysis step. Figure 2-4 shows the workflow of the global alignment step.

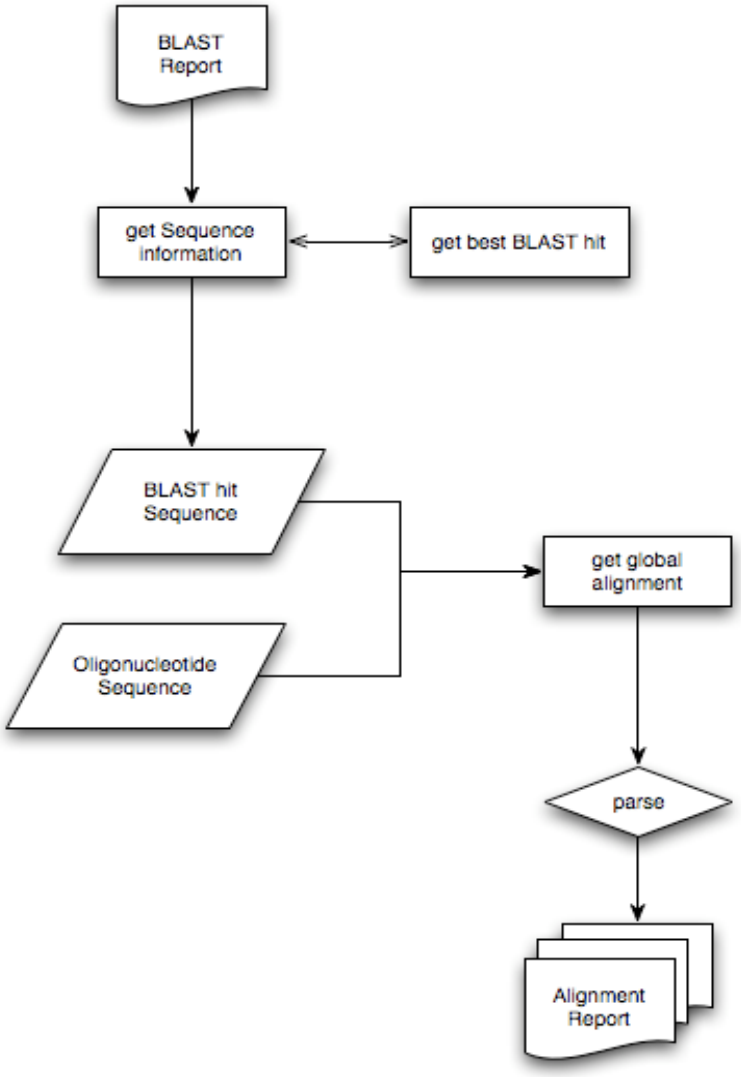


**Figure 2-4 Global Alignment Workflow**

## 2.5 Part 2 – Annotation

This describes the second part of iMAT, which analyses the sequence alignments and automatically annotates the obtained results from the sequence comparison step.

### *2.5.1 Agilent ID annotation*

The annotation file for the probe IDs provided was already outdated and incomplete.

Results from BLAST and the global alignment are parsed into a first result file in a comma separated values (.CSV) format. For each probe ID the user can specify (via the GUI) how many high scoring BLAST hits shall be aligned in the global alignment step. Together with additional information such as UniGene ID or Ensembl Gene/Transcript ID as well as the calculated iMAT score, a report file is created.

To enrich the information resulting from BLAST and the global alignment, an automated annotation (with gene name and additional identifier) of the global alignments (and the Agilent probes) was specified as one aim of this project.

This annotated information is stored in a .CSV file for easy manipulation by the user.

The latest official annotation file release from Agilent for the G4121B mouse chip in (2007) only contained 19311 annotated Agilent IDs.

Given that Agilent IDs are not the most common identifier in most biological databases, Ensmbl BioMart's Perl API was used to harvest additional information based on the Agilent ID as the query ID. This allowed annotation of even more Agilent IDs in this first annotation step.

The decision to use the Ensembl BioMart was not only based on BioMart being the only Perl API accessible online source that accepted Agilent IDs as a query ID. But rather, it also allows a very rapid access, even when a lot of identifiers need to be annotated, and is particularly useful for data-mining like searches as used in iMAT

(searches for gene names and additional identifier). Furthermore, the Perl API can be easily modified if further information is needed in the future, such as gene description, KEGG pathway IDs and many more. Search results are automatically parsed into a .CSV file, which can be used for additional information for the user, as well as in the further steps in the annotation process.

If the first BioMart annotation approach did not yield additional annotated Agilent IDs, the second identifier, such as GenBank ID or Ensembl Transcript ID, given in Agilent's official annotation file was used. In this case a separate, additional BioMart query was performed with GenBank /Ensmebl Transcript IDs as the query IDs.

This more complex approach yielded a total of 19,635 annotated Agilent IDs, which was important for identifying inter-species conserved probes later on in this project. **Figure 2-5** presents a basic overview of the annotation step for Agilent IDs.

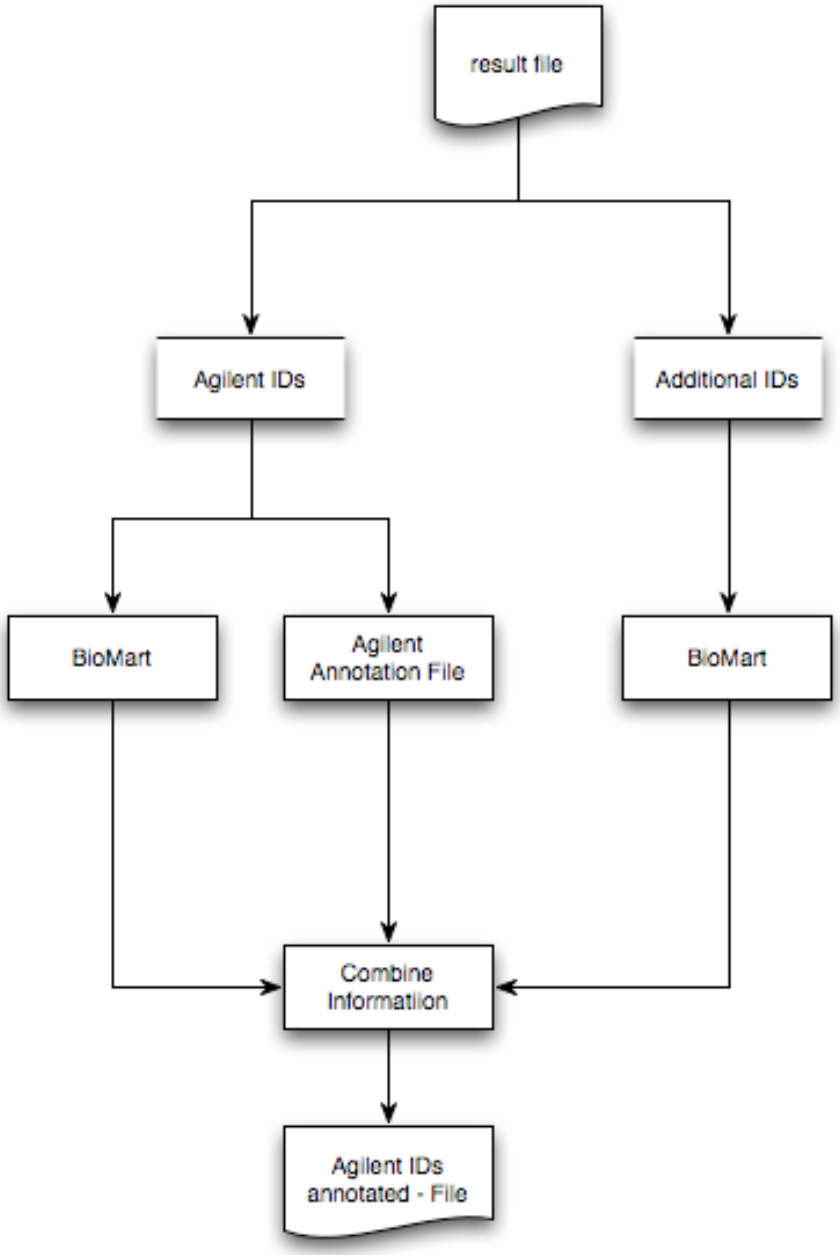


**Figure 2-5 Workflow of the Agilent ID annotation taking the Agilent probe IDs from the iMAT results file and producing a Agilent ID annotation file**

### 2.5.2 BLAST/iMAT hits annotation

For each Agilent ID in the resulting .CSV file the best hit IDs are taken to further annotate them in a similar way as the Agilent IDs. The first step takes the primary identifier for each hit as the query ID for the Ensembl BioMart query.

If this first search doesn't yield additional information, the IDs with missing information are stored for further annotation. Depending on from where the species database was downloaded, either Ensembl or UniGene, a second annotation step is performed as shown in Figure 2-7.

For UniGene databases NCBI's Entrez Programming Utilities (eUtils) are used to access additional available information.

eUtils is a NCBIs own tool to gain access to information stored in Entrez databases outside of the regular web query interface (NCBI-eUtils,August 2009).

There are seven eUtils provided for gaining access, ESearch and EFetch are used in this project. With ESearch the search for information according to the primary query ID is performed in Entrez's gene database to get the designated information for the UniGene ID representing a BLAST hit.

EFetch is the method used for actually retrieving the available information regarding the query ID. The retrieval mode thus the file format is important in this part. In this project the extensible mark-up language format (.XML) is used to parse only specific information from the retrieval file (gene name and Ensembl ID).

For Ensembl databases, the Ensembl Gene ID or Ensembl Transcript ID is used as the query ID for BioMart to retrieve additional information for a certain identifier.

All gained information, annotation of Agilent IDs, annotation of hit IDs and their iMAT score is parsed into a separate report file (Figure 2-6) providing the user with

all information for each annotated Agilent ID and its globally aligned hits from the sequence alignment step.

| Agilent ID | Gene Name | add. ID | UnigeneID | Gene Name | Ensembl Gene | Blast Score | UnigeneID | Gene Name | Ensembl Gene | Blast Score | UnigeneID | Gene Name | Ensembl Gene | Blast Score | UnigeneID | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A_51_P22450 | Bag1 | NM_009736 | Hs.377484 | BAG1 | ENSG0000010 | 168 | Hs.377484 | BAG1 | ENSG0000010 | 168 | Hs.377484 | BAG1 | ENSG0000010 | 168 | Hs.377484 | |
| A_51_P44276 | 2310061N02R | AK010010 | Hs.158688 | EIF5B | ENSG0000015 | 84 | Hs.158688 | EIF5B | ENSG0000015 | 84 | Hs.507805 | COG6 | ENSG0000013 | 82 | Hs.507805 | |
| A_51_P40860 | Slc27a4 | NM_011989 | Hs.719285 | HEPACAM | ENSG0000016 | 94 | Hs.719285 | HEPACAM | ENSG0000016 | 94 | Hs.719285 | HEPACAM | ENSG0000016 | 94 | Hs.719285 | |
| A_51_P41378 | Commd3 | NM_147778 | Hs.514227 | GFAP | ENSG0000013 | 96 | Hs.719196 | TMEM87A | ENSG0000010 | 89 | Hs.715605 | PPP6C | ENSG0000011 | 71 | | |
| A_51_P28687 | Ttll11 | AK016577 | Hs.715522 | EXOC5 | ENSG0000007 | 109 | Hs.99195 | GPR125 | ENSG0000015 | 88 | Hs.99195 | GPR125 | ENSG0000015 | 88 | Hs.369920 | |
| A_51_P36262 | Tnnt1 | NM_011618 | Hs.707356 | TNNT1 | ENSG0000010 | 125 | Hs.707356 | TNNT1 | ENSG0000010 | 125 | Hs.21417 | ZCCHC12 | ENSG0000017 | 83 | Hs.707356 | |
| A_51_P39760 | Stam2 | NM_019667 | Hs.518475 | EIF4A2 | ENSG0000015 | 101 | Hs.451802 | RP11-490G2.1 | ENSG0000023 | 98 | Hs.558364 | ORC4L | ENSG0000011 | 51 | | |
| A_51_P20394 | Olfr652 | NM_147048 | Hs.715569 | LRP3 | ENSG0000013 | 82 | Hs.715569 | LRP3 | ENSG0000013 | 81 | Hs.715569 | LRP3 | ENSG0000013 | 81 | Hs.715569 | |
| A_51_P16344 | Carhsp1 | NM_025821 | Hs.158932 | APC | ENSG0000013 | 93 | Hs.158932 | APC | ENSG0000013 | 93 | Hs.172550 | PTBP1 | ENSG0000001 | 86 | Hs.172550 | |
| A_51_P27060 | Dusp15 | NM_145744 | Hs.657632 | CPSF2 | ENSG0000016 | 84 | Hs.657632 | CPSF2 | ENSG0000016 | 84 | Hs.657632 | CPSF2 | ENSG0000016 | 84 | Hs.657632 | |
| A_51_P49452 | Alg11 | BC061469 | Hs.496710 | STAG2 | ENSG0000010 | 98 | Hs.716396 | HSPA5 | ENSG0000004 | 93 | Hs.68257 | GTF2F1 | ENSG0000012 | 82 | Hs.68257 | |
| A_51_P23936 | Krt82 | NM_053249 | Hs.719174 | UGCGL1 | ENSG0000013 | 77 | Hs.719174 | UGCGL1 | ENSG0000013 | 77 | Hs.719174 | UGCGL1 | ENSG0000013 | 77 | Hs.719174 | |
| A_51_P49301 | Cmya5 | AJ511265 | Hs.482625 | CMYA5 | ENSG0000016 | 119 | Hs.482625 | CMYA5 | ENSG0000016 | 119 | Hs.482625 | CMYA5 | ENSG0000016 | 119 | Hs.482625 | |
| A_51_P37757 | Pdx1 | NM_008814 | Hs.516253 | ASPRV1 | ENSG0000017 | 95 | Hs.599470 | GOT2 | ENSG0000012 | 90 | Hs.599470 | GOT2 | ENSG0000012 | 90 | Hs.567297 | |
| A_51_P44275 | Arhgef19 | NM_172520 | Hs.369056 | SP100 | ENSG0000006 | 87 | Hs.369056 | SP100 | ENSG0000006 | 82 | Hs.682638 | RP11-544M22 | ENSG0000022 | 79 | Hs.682638 | |
| A_51_P30818 | Sfrs1 | NM_00107816 | Hs.715839 | CTNNA3 | ENSG0000018 | 91 | Hs.715839 | CTNNA3 | ENSG0000018 | 91 | Hs.715839 | CTNNA3 | ENSG0000018 | 88 | Hs.715839 | |
| A_51_P37039 | Carm1 | NM_021531 | Hs.371416 | CARM1 | ENSG0000014 | 147 | Hs.371416 | CARM1 | ENSG0000014 | 141 | Hs.371416 | CARM1 | ENSG0000014 | 141 | Hs.371416 | |
| A_51_P22037 | 1700060H10R | NM_133751 | Hs.533468 | FAM29A | ENSG0000014 | 98 | Hs.533468 | FAM29A | ENSG0000014 | 98 | Hs.533468 | FAM29A | ENSG0000014 | 98 | Hs.533468 | |
| A_51_P36707 | Il9 | NM_008373 | Hs.654973 | AC008865.3-2 | ENSG0000021 | 85 | Hs.271791 | ATR | ENSG0000017 | 81 | Hs.271791 | ATR | ENSG0000017 | 81 | Hs.478150 | |

**Figure 2-6 Report file of the annotated hits from the iMAT results file**

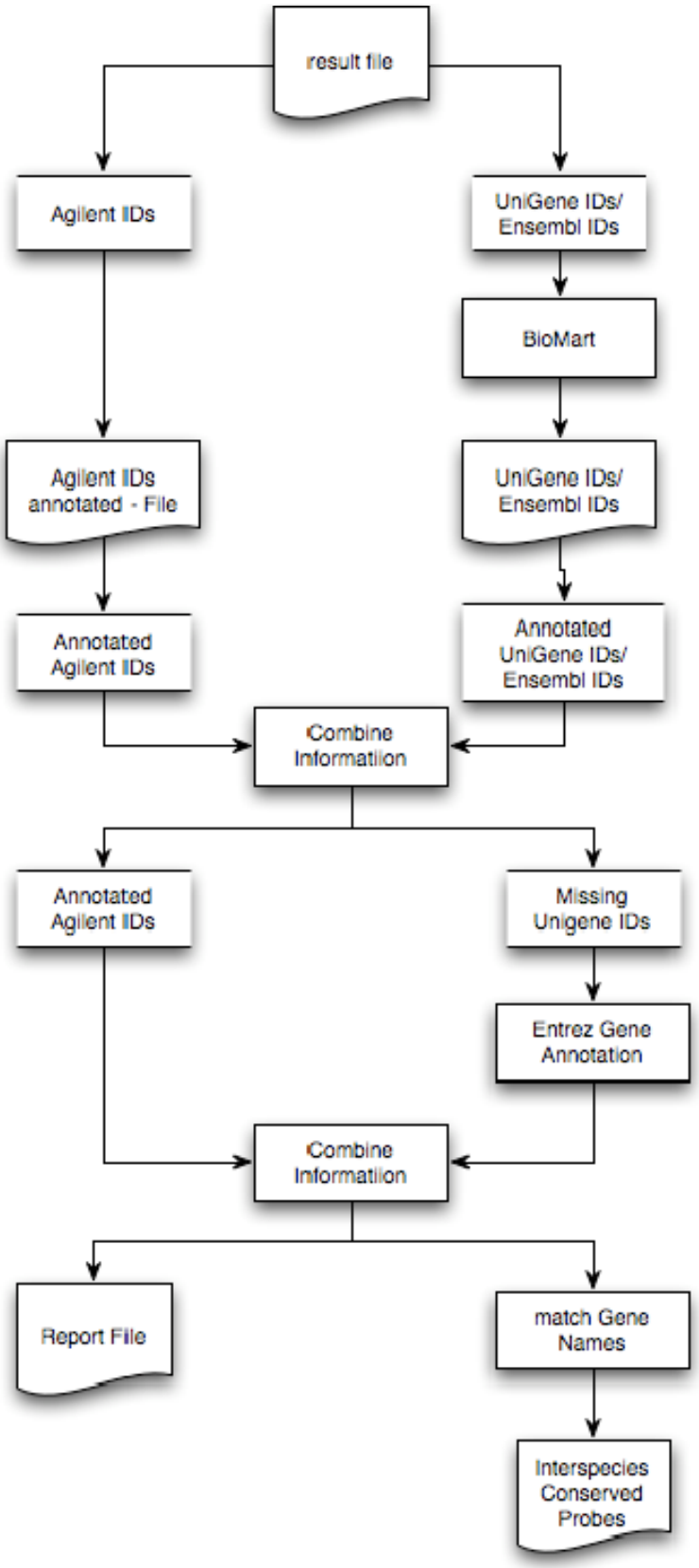


**Figure 2-7 Workflow of UniGene ID/ Ensembl ID annotation taking the iMAT hits from the results file and producing a report file as shown in Figure 2-6 and an additional ISC probe set file**

### 2.5.3 Inter-species conserved probe sets

An analysis for matching annotation terms (gene symbols) between the Agilent IDs and the hit IDs is performed in addition to the Biomart and Entrez gene search. Those results are stored in a separate file.

This allows the user more detailed analysis and indicates, as specified in "Aims and Objectives", conserved genes between different species.

The inter-species conserved (ISC) probe set is regarded as a subset of the total probe set that demonstrates conservation. The assumption is that this set of oligonucleotide probes successfully binds to its targets in the sample applied on the chip under certain experimental conditions, thus it is more likely to produce a better signal intensity. For defining ISC probe sets, a certain threshold should be set. In the previous projects the E-value was regarded as the most useful criterion, as it describes the number of BLAST hits that are expected to occur within a given organism database. In this case the lower the E-value of a BLAST hit the more significant this hit is statistically (Güzlek, 2006).

In this project, the focus is shifted to the matches in sequence annotation (gene names) between Agilent probes and their hits in the database, because it is the aim of this project to add certainty to this specific subset of the results. The combined application of three parameters: (1) the iMAT score, (2) the % sequence homology and (3) the length of consecutive base pair stretch were used to investigate the ISC subset selection further.

### 2.5.4 Global alignment analysis

As one aim of this project is the generation of a reliability scale based on various parameters to provide the user with more information on the possible outcome of the cross-species microarray experiment.

For this purpose an additional analysis step of the global alignment was added to provide the user with additional information such as the % of sequence homology between two aligned sequences, and the number of consecutive base pair matches.

In this step the alignment file for each Agilent probe is re-examined and comparisons of the two globally aligned sequences (Agilent oligonucleotide sequence and corresponding BLAST hit) are performed. More specifically the perfect matches between the sequences are also counted as well as the number of consecutive matching base pairs.

Previous studies have shown that not only high compliance between these two rather short (60 base pairs) sequences is necessary, but also a certain amount of consecutive matching base pairs (>14, ≥ 16) in order to generate hybridisation signals above the background in the microarray experiment (Yee *et al.*, 2008, Ernst *et al.*, 2006). The outcome of these calculations (namely % sequence homology and consecutive base pair matches) and the iMAT scores are stored in two separate files. One file containing Agilent probe IDs and information on each BLAST hit that has been globally aligned, and the other only containing the Agilent probe information and its maximum hit information (Figure 2-8). In addition a result file is created containing annotated Agilent probe IDs and highest scoring annotation matches.

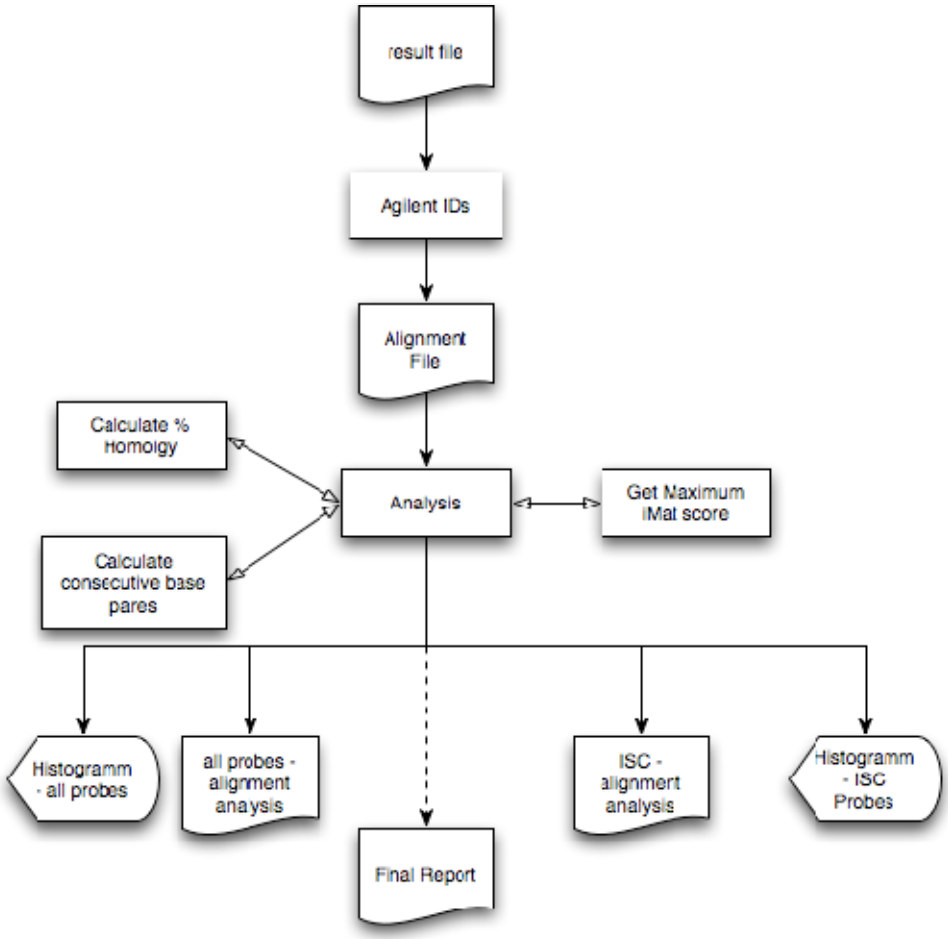


**Figure 2-8 Global Alignment Analysis using the improved iMAT algorithm**

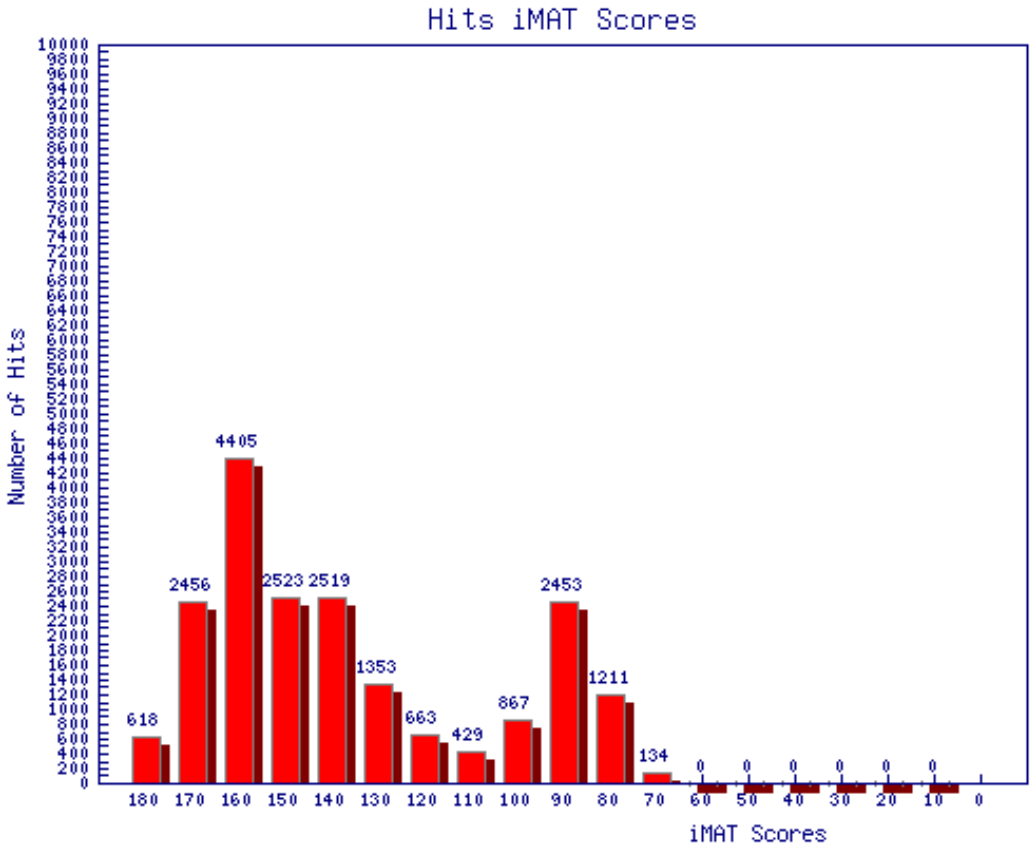


**Figure 2-9 Example graphical output for alignment analysis. This output is intended to give the user an overview of the results obtained in the alignment analysis step**

During the alignment analysis step graphical results are generated; an example is shown in Figure 2-9, which informs the user of the number of hits of the results and how the hits fall into iMAT score categories (180, 179-170, 169-160, … ) at a glance. This graphical report is created for the automated ISC subset selection data based on annotation matches and for the iMAT report data.

Again, the chosen parameter settings from the user input are used to allow the user to specify which probe hits are included in the final iMAT report.

### *2.5.5 iMAT report*

To provide the user with a structured "overview" of all analyses, a final results file is created. This file is a list of annotated Agilent IDs with the highest matching iMAT score hit. It includes gene names of Agilent IDs and their highest scoring hit information. The highest scoring hit information includes gene name as well as the

corresponding iMAT score, the consecutive base pairs and the % homology between the two sequences.

In addition, result and report files from previous steps are retained to allow the user to further investigate annotations or iMAT scores for one Agilent ID, if they wish.

The file format .CSV was chosen again to allow easier file parsing for future extensions of the program, and to provide the user with a file format that can be easily analysed with common programs such as MS Office Excel. Moreover it is possible for the user to sort the results in a desired way to assess interesting information easily.

## 2.6 Part 3 – ISC – probe set comparison

To investigate the possibility of cross-species analysis further this inter-species probe set comparison part was developed.

This part is solely for analysis of different ISC analysis results or iMAT report files. Here the user can select up to five different organisms result sets and compare them to each other to see which probes are conserved among all species. A results file in a specified folder is created. In addition the user can select the iMAT score, the number of consecutive base pairs and % homology again as filters to specify the selection of the ISC subset. A separate report and a graphical analysis file are created to inform the user of how many probes were apparent in all organisms for their given settings.

## 2.7 Graphical User Interface (GUI)

Every step of the program described before can be influenced by user entry through a GUI to ease the usage of iMAT, as the previous scripts have been command line based.

The GUI was developed with Perl Tk, a GUI toolkit initially developed as a Tcl[17] extension and available to download over CPAN. The reason for usage of Perl Tk was that Java interfaces are capable of calling and initialising Perl scripts. However, call-backs and progress information are harder to establish between two different platforms like Perl and Java compared to staying in one language alone. Furthermore there was no exact specification given by the user on how to create a GUI, only that the application should not run on a server, thus precluding the use of Perl CGI for the iMAT program. Thus, the usage of Perl Tk met the platform independence and had less disadvantages than other possible solutions for the required program set up.

## 2.8 Correlation Coefficient

The linear correlation coefficient (CC) is used to investigate the relationship between the signal intensities of two given samples. For example in this project signal intensities from mouse and CHO microarray experiments were compared to each other. The CC can assume values between zero and one, one meaning two identical datasets have been compared or that all data pairs had the same relationship to each other across the whole dataset. The closer the value gets to zero, the less similarities have been found between two datasets. Therefore, the closer the CC is to one, the better the reproducibility of cross-species experiments.

Jennifer Mead established this method as a method to evaluate cross-species data in 2005. Details on how to calculate the CC, therefore, can be taken from "Development of Methods to evaluate inter-species gene expression data" (Mead, 2005).

The experimental data used to calculate the CC in this project has been used previously (Hacer Gülzek, 2006). As there are now many more hamster sequences available, this approach was reapplied on the new dataset, aligning 20,868

[17] Tcl: Tool Command Language (www.tcl.tk)

oligonucleotide probes against 43,178 hamster sequences using the BLASTN and the global alignment algorithm.

# 3 Results and Discussion

This chapter outlines the features of the created program. In addition the testing of iMAT with test data is illustrated and results of several compared species are shown.

First mouse probes were aligned against the latest UniGene mouse database to investigate the accuracy of the software algorithms and validated the software.

By aligning mouse probes against the custom created hamster sequence database in this project the potential of the mouse microarray serving as a feasible platform for inter-species experiments with CHO was investigated in more detail (chapter 3.2).

A main project objective was to create a reliability index (based on annotation matches, % sequence homology, consecutive base pairs and the iMAT score) to add further confirmation to the iMAT results.

The influence of the different parameters on the iMAT results were investigated by comparing the results calculated with iMAT (e.g. the identified probes) to the correlation coefficient (chapter 2.8) of the signal intensities of those probes derived from heat shock experiments of mouse (3T3) and CHO dhfr-cells.

iMAT already preselects inter-species conserved subsets based on matches of the gene name. Therefore the influence of each individual parameter (annotation match, % sequence homology, consecutive base pairs and the iMAT score) was thoroughly tested separately (chapters 3.2.3, 3.2.5, 3.2.6) and in combination (chapter 3.2.7, 3.2.8) to identify the most stringent criterion, to gather further information on each parameter and based on the results develop a reliability scale.

In addition analysing cross-species conserved probes between Hamster, Rat and Human tested the suitability of the mouse microarray as a "generic" platform (chapter 3.3), as those species were the most interesting for this project.

## 3.1 Validating iMAT with mouse probes against mouse database

To validate the E-value settings, the global alignment step and annotation, mouse probes were aligned against the mouse database (Mus musculus) downloaded from NCBI UniGene[18] (results shown in Table 3-1).

**Table 3-1 Mouse vs. mouse sequence alignment and the found hits in the UniGene database**

| Organism | All Queries | Found | No Hit | Found(%) | Av. Score |
|---|---|---|---|---|---|
| Mouse | 20868 | 20800 | 68 | 99.67% | 176.95 |

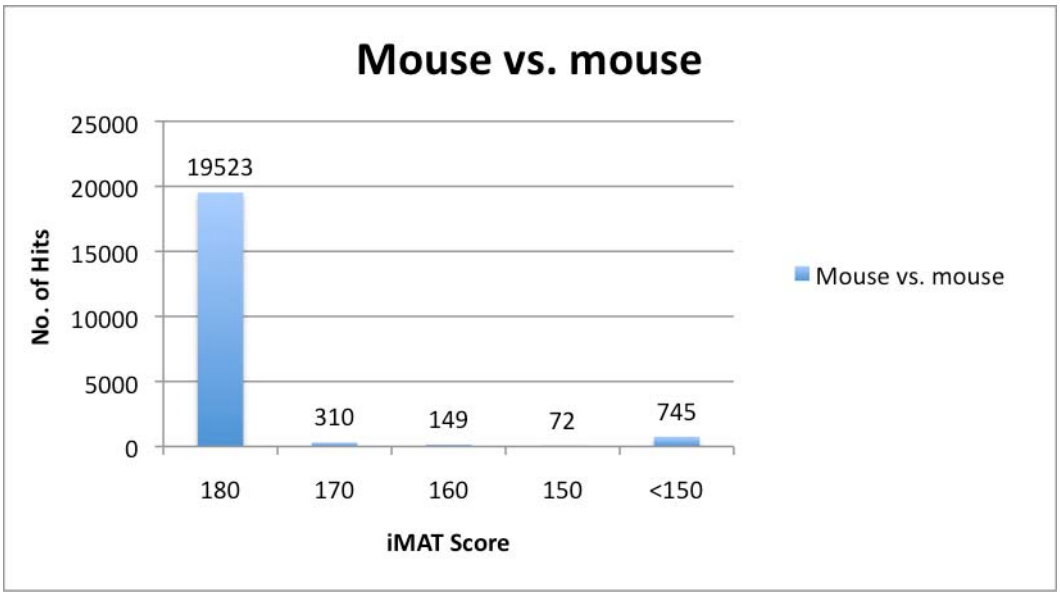


**Figure 3-1 Mouse vs. mouse hit distribution, graphical overview of the number of probes that fall into the iMAT score categories.**

Out of 20,868 probes sequences only 20,798 could be successfully aligned against the UniGene mouse sequence database at an E-value of 1. From these 20,2798 sequences 19,523 hits had the highest iMAT score of 180 (Figure 3-1). The distribution of the remaining sequences was spread amongst lower iMAT scores ranging down till a lowest score of 41.

Out of these alignment results, 16,935 sequence hits could be matched according to their gene names to the ones of the Agilent probes (19,636 total annotated). The

[18] The latest database version (May 2009) comprised 4.215.085 total sequences in clusters and 78,825 UniGene clusters.

missing matching IDs and low iMAT scores can be explained by the nature of the UniGene database (clustered transcripts) and the relatively short sequence of the oligonucleotides on the microarray (60 mer). Furthermore, the usage of IDs from unavailable databases as well as uncurated databases might give an explanation.

## 3.2 Inter-species conserved probes – CHO – Reliability Index

One of the main aims of the project was to investigate sequence similarities between hamster sequences and the mouse probe sequences from Agilent's microarray in order to confirm and predict which sequences are more likely to generate a successful hybridisation signal in a cross-species microarray experiment and also to detect sequence similarities between rodent species as well as other mammalian species.

The intention of this project is also to show that sequence conservation exists between several mammalian organisms. Bearing this in mind, the idea was to show that a set of Agilent probes on a commercially available microarray, such as mouse is very likely to hybridise to targets of several other mammalian samples. In theory, depending on the evolutionary and phylogenetic relation of two organisms, one organism specific microarray could be used in experiments with other organisms, whose transcripts are either not very well known or for which no commercially available microarrays exist. This approach might lead to generic probe sets for a generic microarray chip in the future. To investigate this possibility further hamster sequence data was aligned against mouse probes from Agilent's mouse microarray.

### *3.2.1 Hamster sequence data:*

At the moment no hamster genome sequence database and only sparse transcript data is publicly available, as only limited sequencing efforts have been undertaken. With

new and fast sequencing methods recent developments indicate that sequencing of the Chinese hamster genome will be achieved soon.

Projects on cross-species microarray experiments from previous students (Güzlek, 2006, Mead, 2005) dealt with a different dataset of hamster sequences downloaded from the NCBI taxonomy database.

In this year's project the latest sequence release from the CHO Consortium (http://hugroup.cems.umn.edu/CHO/cho_index.html) was used, obtained by the ACBT working group and as being part of the Institute for Applied Microbiology at the University of Natural Resources and Applied Life Sciences (BOKU) Vienna, Austria. Additional sequences from NCBI taxonomy database were downloaded. All those sequences (contigs[19] and EST[20]s from the Consortium and Entrez Nucleotide sequences) were combined into a single FASTA file as an individual Chinese hamster database, resulting in 43,178 sequences. To take care of a certain amount of redundancy only single read ESTs and contig sequences were used from the consortium sequence release.

### *3.2.2 CHO sequence alignment against mouse probes*

To investigate the influence of the E-value on sequence comparison and annotation results hamster sequences were aligned against the mouse probes at the default E-value of iMAT with a value of 100,000. This setting was intended to find as many as possible hits in the databases and to determine whether inaccuracy of the BLAST hits was reflected in lower iMAT scores or less accurate annotation.

[19] Contig: from shotgun DNA sequencing, a contiguous overlapping set of genes is derived. It is used to deduce the original sequence from the DNA Source

[20] EST: Expressed sequence tags, a short sub set of a transcribed DNA, it may be used to identify gene transcripts and is derived from a one-shot sequencing from cloned mRNA resulting in a relatively low quality nucleotide fragment.

The mouse probes were aligned against the custom created hamster sequence database. Depending on the E-value selection, a different amount of sequence matches was found.

The first run was performed using an E-value of 1 and a setting of 10 for the number of global alignments of the best ranked BLAST hits.

A total number of 14,495 out of 20,868 Agilent probes (69.46 %) showed hits in the database at an E-value of one (Table 3-2).

**Table 3-2 Hamster sequence alignment found hits results at an E-value of 1 and 100 000**

| Organism | All Queries | Found | No Hit | Found (%) | Average Score | |
|---|---|---|---|---|---|---|
| Hamster | 20868 | 14495 | 6373 | 69.46 | 109 | E-value 1 |
| Hamster | 20868 | 20864 | 4 | 99.98 % | 106.63 | E-value 100000 |

In Table 3-2 “all queries” is the number of oligonucleotide probes that were queried against the hamster database. The number of hits is the number of hits found in the database below a certain E-value. The average score refers to the global iMAT scoring (max. 180).

As already mentioned, during the project the question was raised of how the E-value selected by the user influences the outcome of all the analysis. For this purpose, a second sequence alignment at an E-value of 100 000 (results Table 3-2) was performed to find as many sequence hits as possible and to investigate the influence of the global alignment on further analysis methods such as annotation and alignment analysis.

Compared to the sequence alignment results at an E-value of 1 it becomes apparent that a very high E-value influences the outcome, as many more BLAST hits were found than at a low E-value. Still, as a second global alignment step is used, the average iMAT score now is lower than at an E-value of 1. Reasons for this are, that much more global alignment sequence pairs (20854 vs. 14495), which include also

hits of lower confidence, due to lower stringency, account for the global score. With a setting of 100 000 iMAT now is able to align more hits globally than at the lower E-value of 1.

### 3.2.3 ISC subset selection by annotation matches

Contrary to previous projects, the inter-species conserved probe sets (ISC) were automatically selected due to a match of gene names in iMAT. These preselected sets were further investigated according to the iMAT score. The iMAT score is an indicator for the homology between two globally aligned sequences. As mentioned before out of 20,868 Agilent probe IDs, 19,636 probes could be annotated and were stored in a separate file. The information of this Agilent annotation file was used to compare the gene symbols of the probe IDs with the gene names of the obtained hits. If the gene names match, the probe ID information and the according hits information are stored in a separate 'iscxxx.csv' file. For the further analysis these files were used.

**Table 3-3 Hamster-annotation matches subset obtained with gene name comparison**

| Organism | Annotated Probe IDs | Found Matches | E-value | Average Score | CC of matches |
|---|---|---|---|---|---|
| Hamster | 19636 | 4114 | 1 | 109 | 0.735 |
| Hamster | 19636 | 4250 | 100000 | 106.63 | 0.724 |

As shown in Table 3-3 the overall CC values for both data sets were calculated. As the values were lower than in years above, when only comparing specific subsets according to low E-values (<1) and non-redundant sequence entries, the current results were further investigated in different iMAT score groups (Table 3-4) to find out how many of the annotated hits fell into the high scoring groups and lower scoring groups.

**Table 3-4 iMAT score groups**

| | |
|---|---|
| Group 1 | 180 |
| Group 2 | 179 -170 |
| Group 3 | 169-160 |
| Group 4 | 159-150 |
| Group 5 | 149-140 |
| Group 6 | 139-130 |
| Group 7 | 129-120 |
| Group 8 | 119-110 |
| Group 9 | 109-100 |

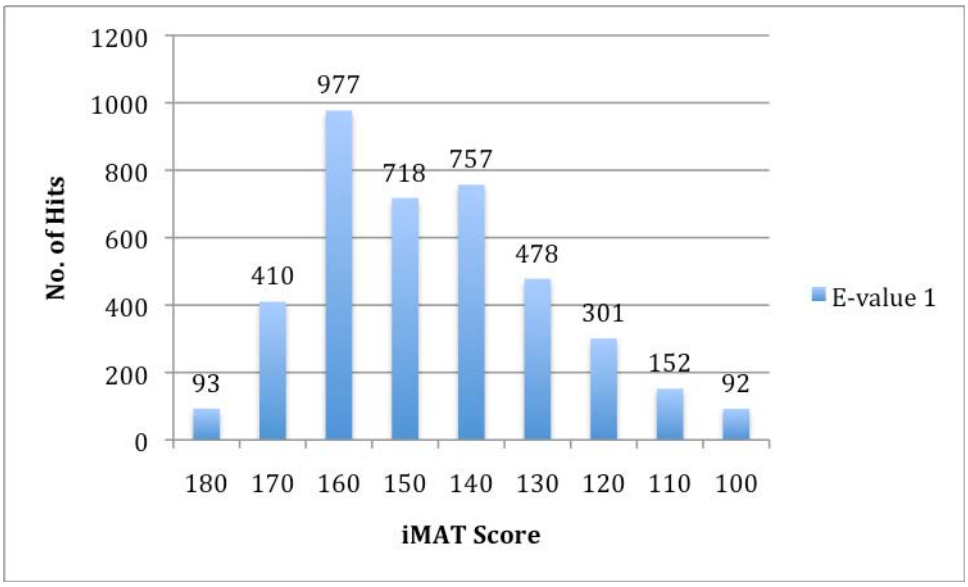


**Figure 3-2 ISC subset based on annotation matches at an E-value of 1 and its probe distribution based on iMAT score groups**

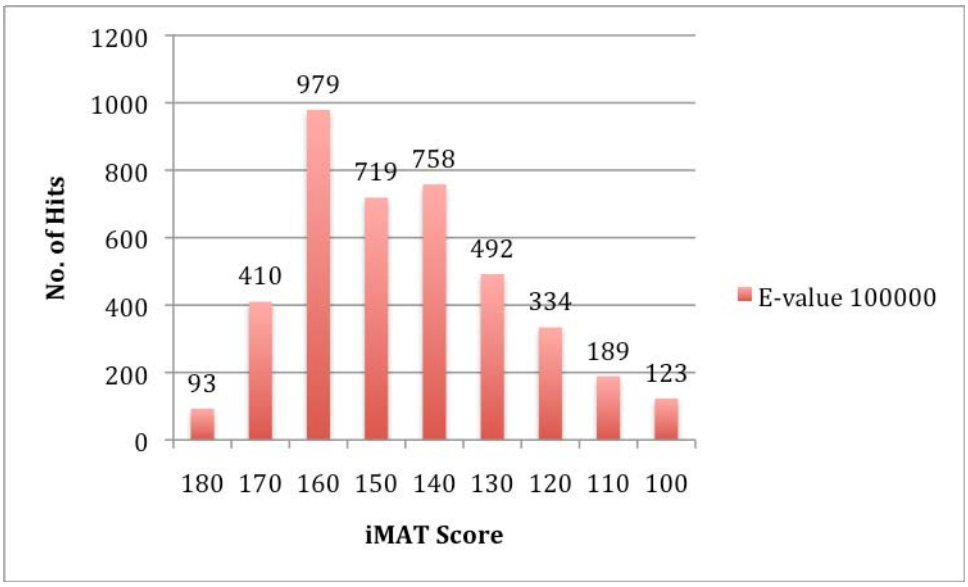


**Figure 3-3 ISC subset based on annotation matches at an E-value of 100,000 and its probe distribution based on iMAT score groups**

Comparing the two charts (Figure 3-2, Figure 3-3) for the distribution of the iMAT scores in the two ISC subsets at different E-values, it is clear that especially in the top iMAT score region (180-150) results of sequence comparisons only have very small differences. This means that although in the second sequence comparison a very high E-value of 100,000 was set, still the same genes were found and even more genes could be identified when matching the gene names. In summary, the E-value setting does not influence the quality of the iMAT results.

### *3.2.4 ISC subsets and experimental data*

The previous part (3.2.3) highlights how iMAT identifies probes for a possible inter-species microarray experiment and shows the distribution of the identified probes based on the iMAT score. Based on their iMAT score the probes were selected and further investigated by using the signal intensities from experimental data of both, mouse and CHO microarray experiments for the corresponding probes. Their correlation coefficients (CC) were calculated to test the accuracy of iMAT.

The correlation was calculated for the signal intensities of all "found probes" derived from mouse (3T3) and CHO dhfr-cells, in the course of a heat shock experiment at an E-value of 1 and 100,000. Determined CC were 0.763 (76.3%) for E-value = 1 and 0.757 (75.7 %) E-value = 100,000, respectively. The results of the former dataset are slightly better, but also contain substantially fewer values; therefore this minor difference in CC results most probably from the different dataset sizes, which were 14,495 hits found compared to 20,864 hits found (Table 3-2).

In addition, further analysis was performed on different groups of probes with different iMAT scores to show a relationship between sequence homology of the aligned probe pairs and their calculated CC.

In Figure 3-4 and 3-3 the group with the highest iMAT scores (180), containing 93 perfect matching and correctly annotated genes are shown.

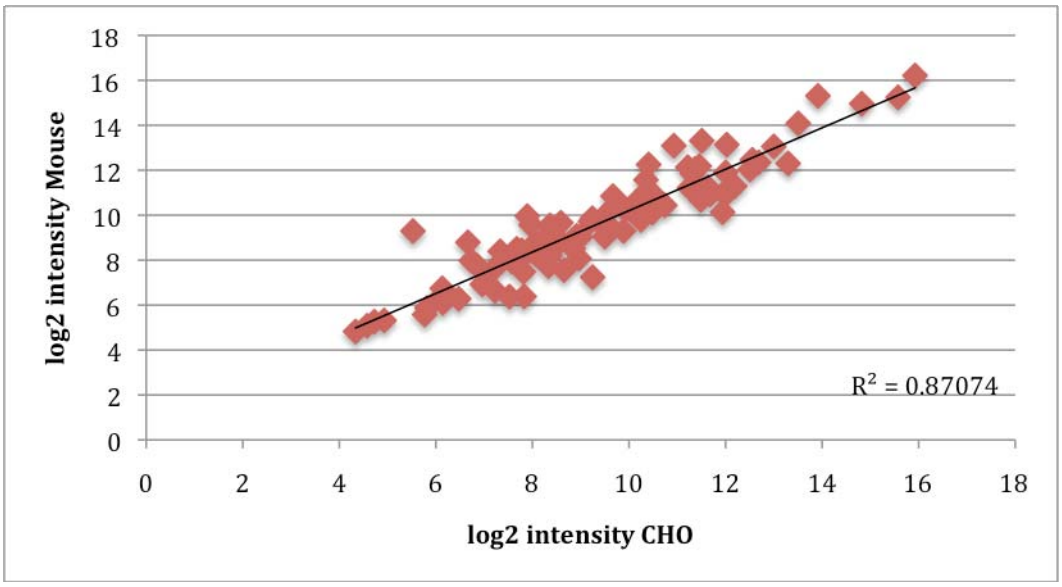


**Figure 3-4 Scatter plot of signal intensities of ISC subset based on annotation matches and an iMAT score 180 at an E-value 1**

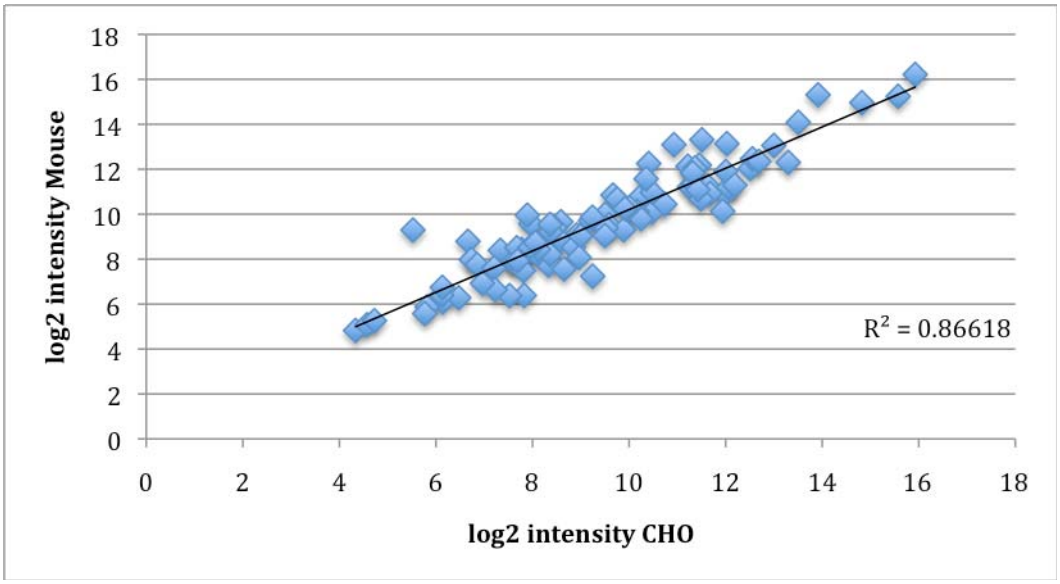


**Figure 3-5 Scatter plot of signal intensities of ISC subset based on annotation matches and an iMAT score 180 at and E-value 100,000**

Both of the graphs (Figure 3-5, Figure 3-4) show the mean log2 signal intensities of that subset of genes that was derived from hybridisation experiments of five mouse RNA samples versus five CHO RNA samples detected on Agilent mouse microarrays as described in the previous students project (Güzlek, 2006). The mouse and hamster

signal intensities of these 93 top iMAT scores gave a correlation coefficient value of 0.93 (93.07% E-value 1 and 93.31 % E-value 100 000) for both E-value settings.

To investigate the correlation of the signal intensities for different iMAT score groups, first the CC of signal intensities of values for an iMAT score of ≥150 were investigated. This range was also used in the previous project (Güzlek, 2006, Mead, 2005). This step was intended to help in the development of a reliability scale based on different parameters such as iMAT score, % sequence homology and the presence of consecutive base pair matches > 15.

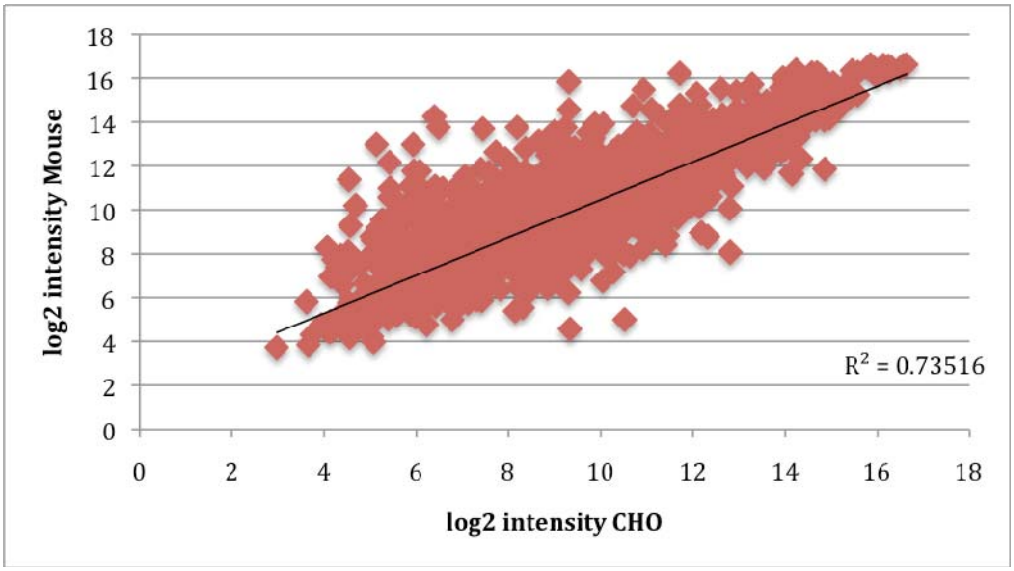


**Figure 3-6 log2 plot of signal intensities of genes with iMAT score 180-150 E-vlaue 1**

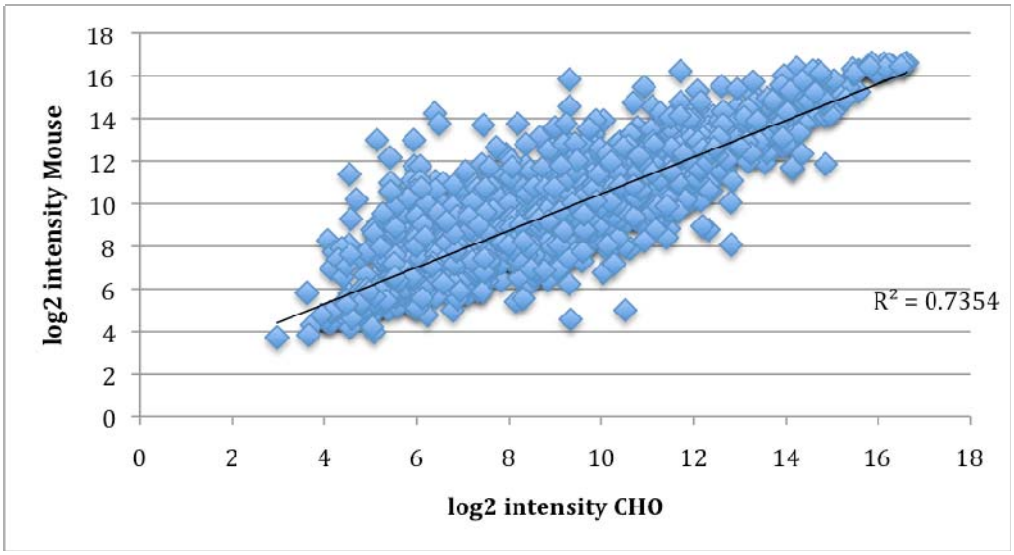


**Figure 3-7 log2 plot of signal intensities of genes with iMAT score 180-150 E-value 10000**

This iMAT score range includes 2,198 signal intensity values at an E-value of 1 (Figure 3-6) and 2,201 signal intensity values at an E-value of 100,000 (Figure 3-7). The correlation coefficients were 0.85 (85.74 %) for E-value 1 and 0.876 (85.76 %) for E-value 100,000. Compared to previous work, where the selection of the ISC subsets was based on the iMAT score, subsets based on the matching annotation and iMAT score yielded slightly better results than in the years before, although a greater number of probes fell into the category.

Therefore, ISC subset selection based on annotation matches can be one way to select probes that are conserved amongst different species.

The amount of selected genes in the different categories remained approximately the same compared with previous projects. An additional analysis of only the two best iMAT score groups (180-170) resulting in 503 signals still was approximately as accurate as the results in 2006 (CC of 0.89, 89.65 % with 315 genes) (Güzlek, 2006) compared with 0.89 (89.56 %,) but still worse than the manually edited list from 2005 with a CC of 0.91 (91.4 %) for 123 genes (Mead, 2005, Güzlek, 2006). An explanation for that might be the smaller number of genes that were considered for calculating the CC, as more genes are included also a higher number of less reliable genes are in the result lists.

Compared to E-value selection of previous years, the results for selection based on annotation matches and iMAT scores proved to be equally accurate and sometimes slightly better when taken into account that for each analysis more probes were available. Normally the CC value gets higher for a decreased number of probes (Güzlek, 2006).

Although a greater number of probes was investigated based on annotation matches and iMAT score the similar CC values compared to precious projects could be calculated proofing that ISC subset selection based on annotation matches is accurate. As a proof of concept, genes only based on the iMAT scores only were selected and analysed again. This resulted in the same CC values as already mentioned above.

These iMAT score groups fall in acceptable BLAST E-value ranges starting with $10^{-27}$. In previous works, that compared EST and matches of annotated genes, E-values of $10^{-15}$ were regarded as high quality results (Adjaye *et al*., 2004) and E-values of below $10^{-10}$ could still be regarded as significant (Wlaschin *et al*., 2005). Therefore, the results can be very well regarded as trustworthy and ISC subset selection is possible based on annotation matches and iMAT score because the E-values of the individual alignments are in acceptable ranges.

It also became apparent that the E-value setting barely influences the sequence alignment since iMAT scoring adds an additional level of confidence. Therefore, further analyses were carried out with the results of the higher E-value setting as ISC subset of annotation matches contained more probes.

In order to investigate how the other parameters such as a consecutive matching base pair length and % sequence homology between cross-species gene pairs influence the CC values and influence inter-species conserved probe subset selection additional analyses were performed. The results of this analysis should help in the creation of the reliability scale as mentioned in this projects aims and objectives (Chapter 1.7.1). At first only the values above an iMAT score of ≥ 150, the influence of consecutive base pairs and a certain threshold of % sequence homology was investigated, than a combination of all of the parameters on the ISC subset with matching annotation and

on all obtained hits was tested to verify the influence of annotation matches on the CC values.

### *3.2.5 ISC subset annotation and consecutive base pairs > 15*

As reported by previous studies (Yee *et al.*, 2008, Ernst *et al.*, 2006, Kane *et al.*, 2000) a stretch of consecutive matching nucleotides is needed to provide a hybridisation signal above the background signal in a microarray experiment.

The actual length needed depends on the microarray platform that is being used, either Affymetrix or Agilent, and the nature of the probes represented on the microarray. On Affymetrix oligonucleotide microarrays one gene is represented as a collection of up to 26 probes each 25-mers long, which represents the whole gene length (Affymetrix, 2009). In contrast Agilent oligonucleotide microarrays represent one gene mostly as one 60-mer oligonucleotide sequence (Agilent, 2009). A longer probe on the microarray is reckoned to be more targets specific whereas shorter probes can be spotted in higher density on the chip. For Agilent oligonucleotide microarrays a length of at least 16 consecutive base pairs yields a hybridisation signal above the background.

This setting was used to investigate how many probes that matched in their annotation contain the necessary consecutive base pairs length for emitting a hybridisation signal.

By selecting the E-value of 100,000 and a consecutive base pair match of > 15 out of the probes, which matched according to their gene names, 3,265 signal intensity values were obtained. The CC of these values was 0.755 (75.5 %).

This is actually worse than with other selection methods. But this selection does not necessarily take the overall sequence homology between 2 sequences into account, as one stretch of 16 continuous base pair matches in the beginning or the end of the

sequence can also yield a successful hybridisation signal above the background although the overall sequence homology might only be 50 %. A high iMAT score on the other hand does not automatically mean that a continuous stretch of more than 15 base pairs occurs in this sequence alignment. Also hits above an iMAT score of 150 can fall into the group of missing consecutive base pair matches.

More over the % sequence homology does not correlate with the consecutive base pair matches, as also hits with a sequence homology of above 90 % appear to be missing the necessary complementary stretch, therefore sometimes not hybridising to the microarray.

Further analysis were performed to find out which parameters, either iMAT score, % sequence homology or presence of consecutive base pair stretches > 15, influence the correlation of the signal intensities the most i.e. giving the best correlation values. These analysis were done to find different areas among the probe sets to develop a novel and easy to use reliability scale, allowing the user to see at a glance which probes are highly likely to yield reliable hybridisation signals in an inter-species microarray experiment. It was also intended to give the user a kind of "grey area" based on sequence alignments and annotation, to indicate probes that are likely to yield hybridisation signals in a cross-species microarray experiment, but can't be seen as certain as well as a "black list" of probes that should not be considered in further analysis.

A scatter plot of the results can be seen in appendix A, Figure A-1.

### *3.2.6 ISC subset annotation and homology ≥ 90 %*

As part of the alignment analysis the percent of sequence homology between two globally aligned sequences was calculated to confirm a certain similarity and to help

in establishing a reliability scale based on several parameters. As a threshold a sequence homology of ≥ 90 % was selected.

By application of this threshold 1,787 signal intensities out of 4,250 possible intensities were obtained. The CC value of this subset was 0.869 (86.9%). This is slightly better than the CC of values selected with an iMAT score ≥ 150 (0.85 for 2,201 signal intensities).

The relation between the % homology and the iMAT score can explain this result. The iMAT score translates into the % homology between two sequences. An iMAT score of 150 equals a homology of mostly 87.1 %. Differences can still occur, as the iMAT score is calculated as a result of a global alignment rewarding matches and mismatches and gaps with different penalties, whereas the homology is calculated by rewarding matches between the nucleotides only.

A scatter plot of the signal intensity values can be seen in appendix A, Figure A-2.

### *3.2.7 ISC subset based on annotation, iMAT score, % sequence homology and presence of consecutive base pairs > 15*

The application of all three thresholds iMAT score ≥ 150, consecutive base pairs > 15, % sequence homology ≥ 90 % resulted in a subset of 1,739 probes out of 4,250 probes from the matched annotation subset. This selection intended to get the best hits in a specific range only with the selection of 90% homology being the most stringent parameter and only high scoring hits with the presence of the necessary stretch of consecutive base pairs. This information was used working towards a reliability scale.

The CC value was 0.869 (86.9 %), the same as for selecting based on homology only.

In order to be able to create a reliability scale a second threshold was set between the iMAT score of ≥70 - <150, a sequence homology between <90 and ≥ 60 % and consecutive base pairs > 15.

Out of all annotation matches 1,233 probes fell into this group, the CC value was 0.656 (65.6%). This means that in this group probes are still likely to bind to targets but with a smaller certainty than for the first group investigated.

Selection of consecutive base pairs was crucial at this point. Analysis within the same range of threshold values only reached a CC value of 0.496 (49.6%) when consecutive base pair matches were disregarded. Considering these results the selection of consecutive base pair matches seems to be an important threshold below a certain sequence homology, but again not the only parameter that needs to be considered.

Scatter plots of the results above can be seen at appendix A, Figure A-3 and Figure A-4.

### 3.2.8 ISC subset based on iMAT score, homology, consecutive base pairs

To investigate if there were differences in ISC subset selection based on annotation matches compared to all obtained results, additional analysis on all the probes with their highest scoring matches was performed. This was done in order to proof that not only the probes with annotation matches are influenced by the three parameters and to show that the established parameters and thresholds for a reliability scale are applicable on all obtained results.

The first group was selected from the results for an E-value of 100,000. The threshold settings were selected as follows: iMAT score $\geq$ 150, sequence homology $\geq$ 90 %, consecutive base pair stretch > 15. Out of all found hits (20,864) 2,009 (versus 1,739 probes out of 4,250 annotation matching probes) fell into this selection. The calculated CC value was 0.874 (87.4%) which was consistent with the results from the subset obtained based on annotation matches and all three thresholds together.

To establish a medium range for the reliability scale the next subset had probes with hits in the iMAT score range of ≤150 and ≥ 70, between < 90 and ≥ 60 % homology between the sequences and a consecutive base pair stretch of at least 16. The CC was 0.722 (72.2%). 3,513 values fell into this category.

This higher value can be easily explained, as the selection based on annotation matches does not consider the best hit for every probe, but only the best hit within the set with a matching gene name. So the hit with the best iMAT score and sequence homology can be the one with the same gene name, but does not necessarily have to.

Through the different selection of the subsets it becomes apparent that relying in annotation matches should not be the only parameter on which a subset selection can be based.

Indeed, annotation matches are a good indicator if good annotation information is available for a specific organism, but if dealing with a poorly annotated organism, which is likely to be the case, when considering cross-species microarray experiments when considering cross-species microarray experiments considering annotation alone is not a reasonable option. It can surely be a helpful criterion, but should not be regarded as the one and only parameter for finding inter-species conserved probes.

Nevertheless, both subset selections follow a similar trend in the number of probes and their CC values, enabling the creation of a schematic scale, based on the three different parameters: the iMAT score, % homology between sequences and the presence of a consecutive base pair stretch of more than 15 nucleotides (Figure 3-8).

This scale is intuitive and novel it is an easy point of reference for the user to gauge the suitability of a microarray platform for a selected organism, for which no commercially available microarray exist.

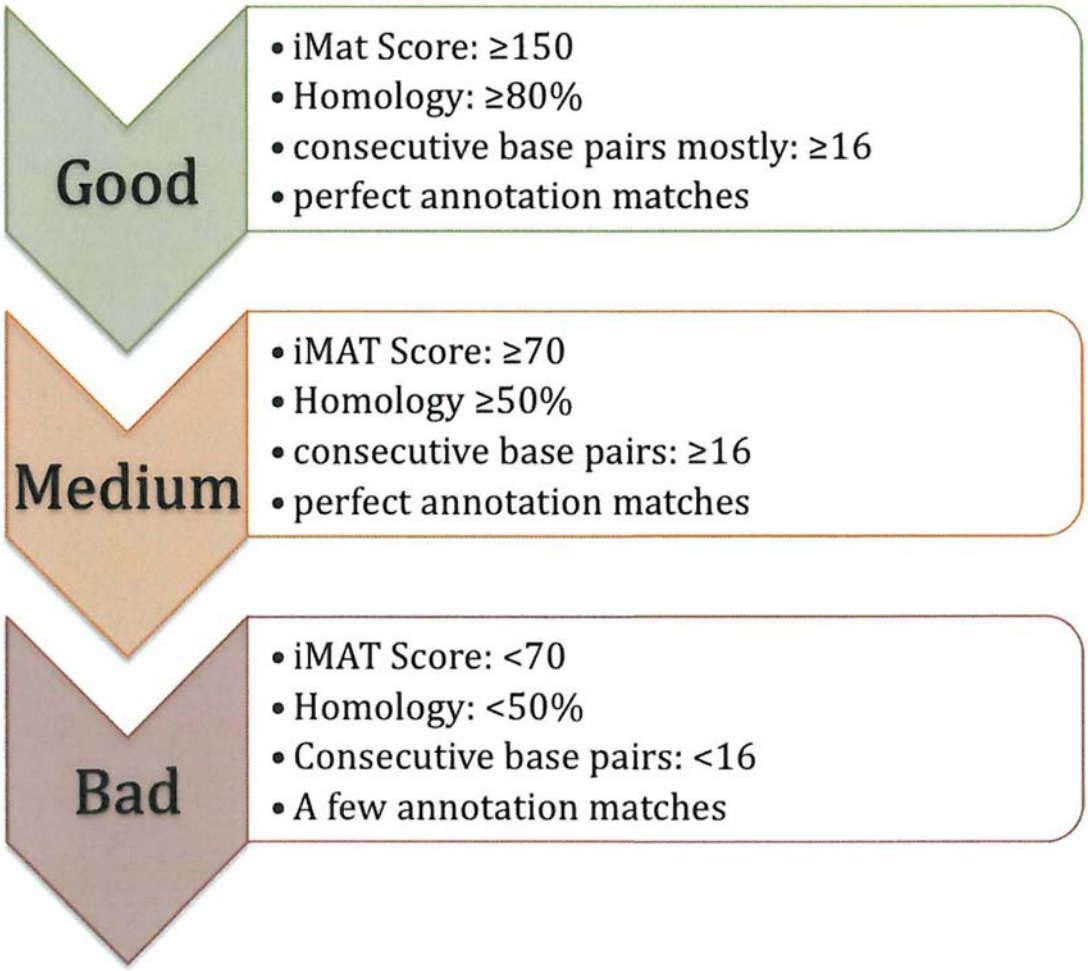


**Figure 3-8 Reliability scale based on established parameters of iMAT score, % sequence homology and presence of consecutive base pairs**

For comparison the same settings were used to determine the correlation of the subsets without continuous base pair stretches but within the same iMAT score ranges and homology ranges. All those probes were again selected from the dataset at an E-value of 100 000.

**Table 3-5 Comparison of iMATscore, % homology with and without consecutive base pairs**

| iMAT score | % Homology | Cons. Bp | Hits | CC | Annotated |
|---|---|---|---|---|---|
| 180-150 | 90% | Yes | 112 | 0.842(84.2%) | Yes |
| <150-70 | <90%-60% | Yes | 758 | 0.496(49.6%) | Yes |
| 180-150 | 90 % | No | 177 | 0.839(83.9%) | No |
| <150-70 | <90-60% | No | 13293 | 0.684(68,4%) | No |

Comparing the results in Table 3-5 justified the selection of the reliability scale above. It also revealed, that a consecutive base pair length of at least 16 nucleotides is

an important parameter, but the CC values might also be misleading, as they do not indicate how high the signal intensities are. They reveal how similar the signal intensities of two compared microarray experiments were. For example having low signal intensity for one probe in both microarray experiments (mouse on mouse and hamster on mouse) will give a high correlation, only if they differ (mouse on mouse higher than hamster on mouse) they CC value will get lower.

It also becomes apparent, that values above an iMAT score of ≥ 150 still have a good CC value, although they do not have the required 16 consecutive base pairs. For this group, the overall homology for the two aligned sequences is still very high indicated by the percentage of sequence homology and of course the iMAT score.

## 3.3 Cross – species sequence alignment analysis

### *3.3.1 Validating iMAT with mammalian and rodent databases*

As the hamster genome is due to be sequenced this year, the investigation of inter-species conserved probes amongst several other mammals was of interest for this project for example investigating the possibility of mouse microarrays as a generic microarray platform for other species. Once the hamster sequencing is complete iMAT can still be used for the prediction of any other cross-species microarray analysis.

The tables below show the results of the sequence alignments of mouse probes against several mammalian and rodent databases. All those sequence alignments were performed with iMAT at an E-value specified as one. As described in Table 3-6 this year's project focussed more on the homology between rodent species and mammals.

**Table 3-6 Mammalian Databases that were selected to be aligned against mouse probes**

| Organism | Latin Name | File Name | Number of Sequences |
|---|---|---|---|
| Rat | *Rattus norvegicus* | Rn.seq.all | 881,300 |
| Human | *Homo sapiens* | Hs.seq.all | 9,906,206 |
| Macaque | *Macaca mulatta* | Mmu.seq.all | 68,573 |
| Hamster | *Cricetulus griseus* | Cho.seq.all | 43,178 |
| Chimpanzee | *Pan troglodytes* | Ptr.seqt.ens | 40,215 |
| Rabbit | *Oryctolagus cuniculus* | Ocu.seqt.ens | 547 |
| Squirrel | *Spermophilus tridecemlineatus* | Str.seqt.ens | 18,359 |
| Guinea Pig | *Cavia porcellus* | Cp.seqt.ens | 13,476 |
| Tree Shrew | *Tupaia belangeri* | Tbe.seqt.ens | 4117 |
| Kangaroo Rat | *Dipodomys ordii* | Dor.seqt.ens | 19,126 |

Based on the results of the last student's project (Güzlek, 2006) it became apparent that it was necessary to adapt the previous scripts to be able to work with Ensembl sequence databases as Ensembl provides a wider range of species than NCBI UniGene. Also UniGene databases start building after a certain amount of sequence entries are available thus restricting the databases species with a superior level of sequence information.

Table 3-7, Table 3-8 and Table 3-9 show all the analysis results performed during this project.

**Table 3-7 Sequence alignment human, rat, hamster**

| Organism | Rat | Human | Hamster |
|---|---|---|---|
| All Queries | 20868 | 20868 | 20868 |
| Found | 20664 | 12110 | 14495 |
| No Hit | 204 | 8758 | 6373 |
| Found(%) | 99.02 % | 58 .03% | 69.46 % |
| Av. Score | 139 | 126.52 | 109 |

Among rat, human and hamster 9,035 probes had hits in all databases.

**Table 3-8 Sequence alignment rodents**

| Organism | Rabbit | Squirrel | Guinea Pig | Kangaroo Rat |
|---|---|---|---|---|
| All Queries | 20868 | 20868 | 20868 | 20868 |
| Found | 17279 | 15575 | 16755 | 15668 |
| No Hit | 3589 | 5293 | 4113 | 5200 |
| Found(%) | 82.80 % | 74.64 % | 80.29 % | 75.08 % |
| Av. Score | 87.97 | 102.23 | 102.64 | 102.92 |

Among the rodent species 9,681 probes had hits in all species databases.

**Table 3-9 Sequence alignment primates**

| Organism | Macaque | Chimpanzee |
|---|---|---|
| All Queries | 20868 | 20868 |
| Found | 17805 | 14087 |
| No Hit | 3063 | 6781 |
| Found(%) | 85.32 | 67.51 % |
| Av. Score | 104.61 | 114.51 |

Among the primates only 12,682 probes had hits in both species.

All the queries show the number of Agilent oligonucleotide mouse probes that were queried for cross-species conserved probe sets against the different databases. The number of hits is the number of hits found in the database at an E-value below one (between the probe and the gene).

A comparison of all nine different species resulted in 4,622 probes that had hits across all species.

These results also include the low-homology values from alignments between mouse probes and transcripts of another species. The homology of two sequences is indicated by the iMAT score (details see Chapter 2.4.2.) and the % sequence homology.

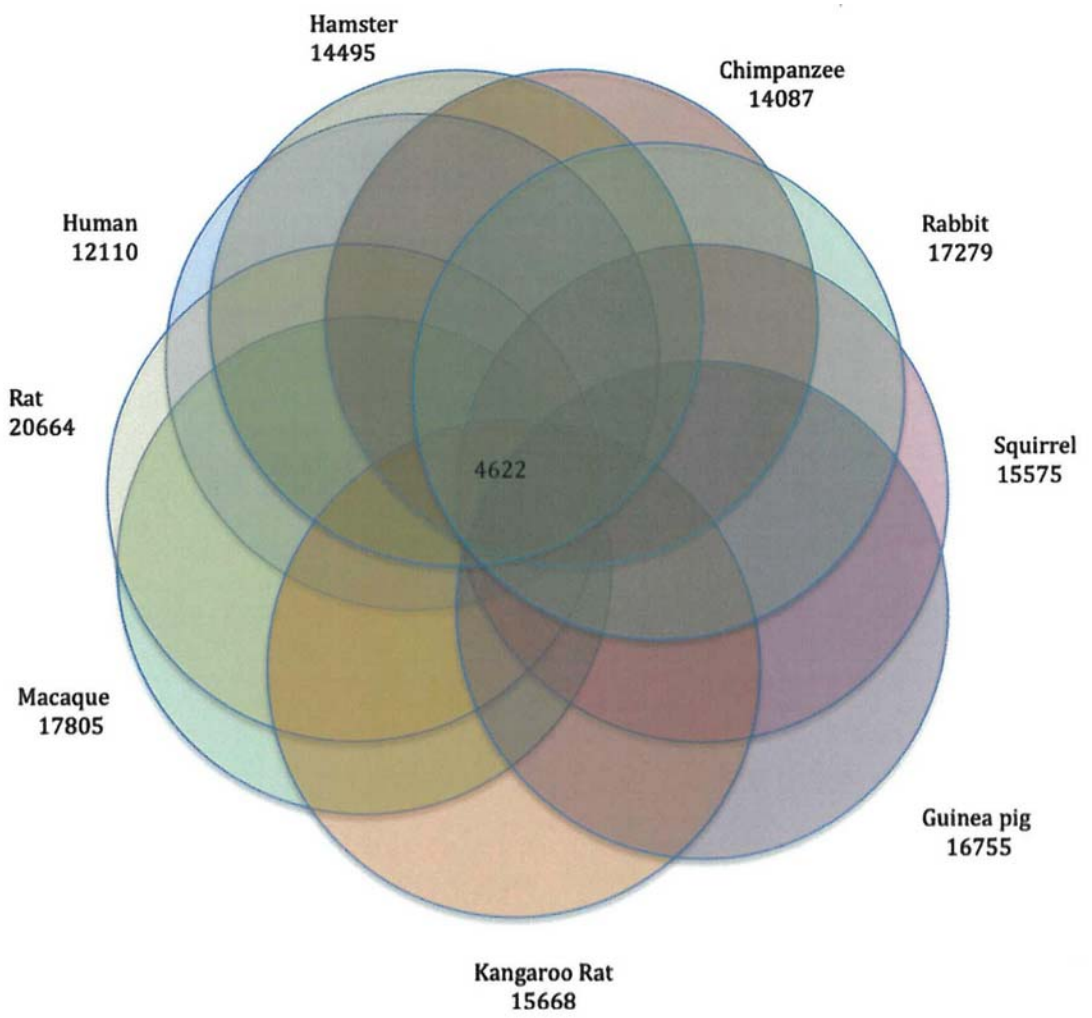


**Figure 3-9 Intersection of iMAT hits of all species except mouse**

Figure 3-9 shows all found hits across all investigated species. All found iMAT alignments were considered regardless of their score, only the E-value of below 1 was taken into account. The idea behind it was, that probes present in all species are likely to give strong signal intensities in a cross-species experiment. As already shown above Table 3.7-3.9 between closely related organisms a higher number of probes can be regarded as conserved. This analysis was done to show, that a generic microarray

chip with generic sequences is possible for a certain group of species, for example for rodents or primates.

The average iMAT score of probes present in all species was 117, which is below the more stringent threshold of 180-150. These thresholds were set empirically[21] to assess hybridisation of the target.

Still the more stringent threshold does not provide any guarantee that a probe will bind to the correct target and that successful hybridisation will occur during the experiment. The more reliable higher threshold ≥ 150 was used to identify alignments that are highly likely to fulfil the requirements for a successful hybridisation in a microarray experiment.

For further analysis, only human, rat and hamster were selected, to investigate the homology across these species compared to mouse probes.

[21] The criterion for choosing these scores as a threshold was that alignments having this score also have a certain level of homology provide hybridisaton signals

### 3.3.2 Mouse probes versus rat, human and hamster

Sequence alignments of those three species against mouse were performed at two different E-values one and 100,000 (Table 3-10).

**Table 3-10 Sequence alignments of rat, human and hamster against mouse probes at an E-value of 100,000**

| Organism | Rat | Human | Hamster |
|---|---|---|---|
| All Queries | 20868 | 20868 | 20868 |
| Found | 20864 | 20864 | 20864 |
| No Hit | 4 | 4 | 4 |
| Found(%) | 99.98 | 99.98 | 99.98 |
| Av. Score | 139.29 | 114.75 | 106.63 |

The results of the sequence alignment at an E-value of one can be seen in Table 3-7 for the three different organisms. Because of the higher E-value almost all probes had a hit in each database due to this lower stringency.

First, all of the best iMAT score hits at the two different E-values were compared to each other to find the conserved mouse probes across those three species.

**Table 3-11 Found probes in all three species**

| E-Value | iMAT 180 | iMAT ≥ 150 | Total hits across three species at iMAT 180-150 |
|---|---|---|---|
| 1 | 16 | 1284 | 10,644 |
| 100000 | 24 | 1284 | 10,665 |

Table 3-11 above shows the conserved probes of both sequence alignment analysis at an E-value of one and 100000 for perfectly matching alignments (score = 180) and the probe set with iMAT scores 180 < >150.

At first, it was calculated how many probes of the mouse microarray had hits across the three species at the respective iMAT scores.

For the E-value of one and an iMAT score of 180, 942 different Agilent probe IDs had hits across the three species. Out of those only 16 probes were conserved in all 3 species for an iMAT score of 180. Although less (434 different) Agilent probe IDs

had hits across the three different species at an E-value of 100 000 and an iMAT score of 180, more probes (24) were conserved in all of the three species. Even though more iMAT hits were found at the lower more accurate E-value (one), the iMAT hits at the high E-value of 100000 seem to yield sequence alignments with better sequence conservation across the three studied species.

As in chapter 3.2.4 the iMAT score was set to a value of ≥ 150 to investigate the confidence area further.

Comparing both results (Table 3-11) it seems beneficial to set a more generous E-value for cross-species analysis, especially when looking for highly conserved genes across different species, which should fall into the highest iMAT score range. In addition the number of probes with hits across the 3 species across the selected iMAT score range (180-150) becomes almost equivalent (10,644 at an E-value of one, 10,665 at an E-value 100 000).

Setting a very high E-value in this case (of above a conventional selected value), for example to 100 000 instead of one or below one, does not seem to influence the detection of cross-species conserved probes.

To investigate the feasibility of using matches of gene names as a parameter to identify cross-species conserved probes, the ISC subsets of all three species (based on matches of the gene name between the probe and an identified hit), were compared to each other and further studied.

### 3.3.3 ISC subsets based on annotation matches and cross-species sequence analysis

To test the suitability of the mouse microarray as a "generic" chip for different species and the sequence conservation across different species the most interesting species (Rat, Human and Hamster) for this project were chosen. Probe subsets of annotated genes only were considered.

**Table 3-12 ISC subsets results and conserved probes Rat, Human, Hamster**

| Organism | Hamster | Rat | Human | Total hits across three species | Hits in all three |
|---|---|---|---|---|---|
| Found | 4114 | 7385 | 5046 | 9775 | 1308 |
| E-value | 1 | 1 | 1 | 1 | 1 |
| Organism | Hamster | Rat | Human | Total hits across three species | Hits in all three |
| Found | 4249 | 9592 | 5224 | 11034 | 1749 |
| E-value | 100 000 | 100 000 | 100 000 | 100 000 | 100 000 |

Table 3-12 shows how many probes based on annotation matches were found for each of the three species at both E-value settings. As already mentioned in Chapter 3.2.3, ISC subsets were automatically selected based on annotation matches. Again the probes having hits across three species were measured, which are shown in Table 3-12 (across three). Out of the 9775 probes, which showed hits across the three species at an E-value of one, 1308 probes were "conserved". Interestingly, using an E-value of 100 000 a total number of 11034 probes were identified across the three species out of which 1749 probes were "conserved".

Because of the higher number of conserved probes at an E-value of 100 000 this dataset was used for further analysis such as % sequence homology and consecutive base pairs.

The same criteria as in Chapter 3.2 were used to investigate the cross-species homology for human, rat and hamster ISC subsets based on annotation matches.

As shown in Table 3-13, 7,473 probes were found across all three species by applying the iMAT score threshold of ≥ 150. Out of these 7,473 probes 740 were present in all three species.

**Table 3-13 Parameter analysis of cross-species conservation based on annotation matches**

| iMAT 180-150 | Homology 100-90% | Base pairs >15 | Combination of all 3 parameters | |
|---|---|---|---|---|
| 7473 | 6476 | 9943 | 6,373 | Total hits across three species |
| 740 | 541 | 1236 | 509 | Hits in all three |

By using the % sequence homology as a threshold 6476 probes were found in total across all three species but only 541 probes were identified as "conserved" (Table 3-13). Based on a consecutive base pairs length of greater than 15, 9943 probes were discovered to have matches across all three species. 1236 were identified in all three of them.

Viewing these results the percentage of homology is the most stringent threshold but as already mentioned in chapter 3.2.6 it doesn't necessarily mean a successful hybridisation signal in a microarray experiment. On the other hand just because a sequence has 16 or more consecutive complementary nucleotides it doesn't necessarily mean the bound target is specific to the particular probe. The iMAT score gives an initial indication of a homology between sequences but is not enough information to predict a successful hybridisation results in a microarray experiment.

However, a combination of all 3 thresholds (Table 3-13 "Combination of all three" iMAT ≥ 150, consecutive base pairs >15, homology ≥ 90 %) revealed that out of all the annotation matches 6,373 probes totally across all three species fell into these categories. 509 matches were present in all three species.

In comparison this selection seems quite similar to the results of % homology but it also reveals that if the homology is above 90% it doesn't necessarily mean that these sequences also have a consecutive base pair match of >15.

Analysing matches of gene annotation is one way to select ISC sets but cannot serve as the only criterion as already mentioned in chapter 3.2.8. But it helps to indicate over which iMAT score ranges, ranges in homology and number of consecutive base pairs matching genes can still be found.

Still the results indicate, although small, a number of 509 probes can be found across all three species as "conserved". These probes all matched in the annotation and fell into the stringent selection of parameters.

The analysis of the inter-species conserved probes not only revealed how the different parameters are dependent on each other but also indicated the information content of each criterion. Only a combined application of those parameters can serve as a reliability scale to indicate if a microarray platform might be suitable for specific inter-species experiments.

The cross-species analysis revealed that iMAT can identify cross-species conserved probes and showed that the closer species are related to each other and the stronger the homology towards the microarray platform is, the better is the indication for a "generic" microarray.

## 3.4 iMAT – Graphical User Interface

When starting the iMAT program with the command 'perl iMAT.pl' in the terminal or command window (inside the program folder) the user has three different options of how to proceed, 'Sequence Comparison', 'Annotation' and 'Inter-species conserved set', which are represented as three tabbed windows.

The first option 'Sequence Comparison' includes BLASTN and the global alignment. The second option allows the annotation of previously processed sequences via the first step, and the third option allows the user to create custom inter-species probe conservation analysis. The different graphical implementations of iMAT are explained in the following sections.

### 3.4.1 Sequence comparison tab

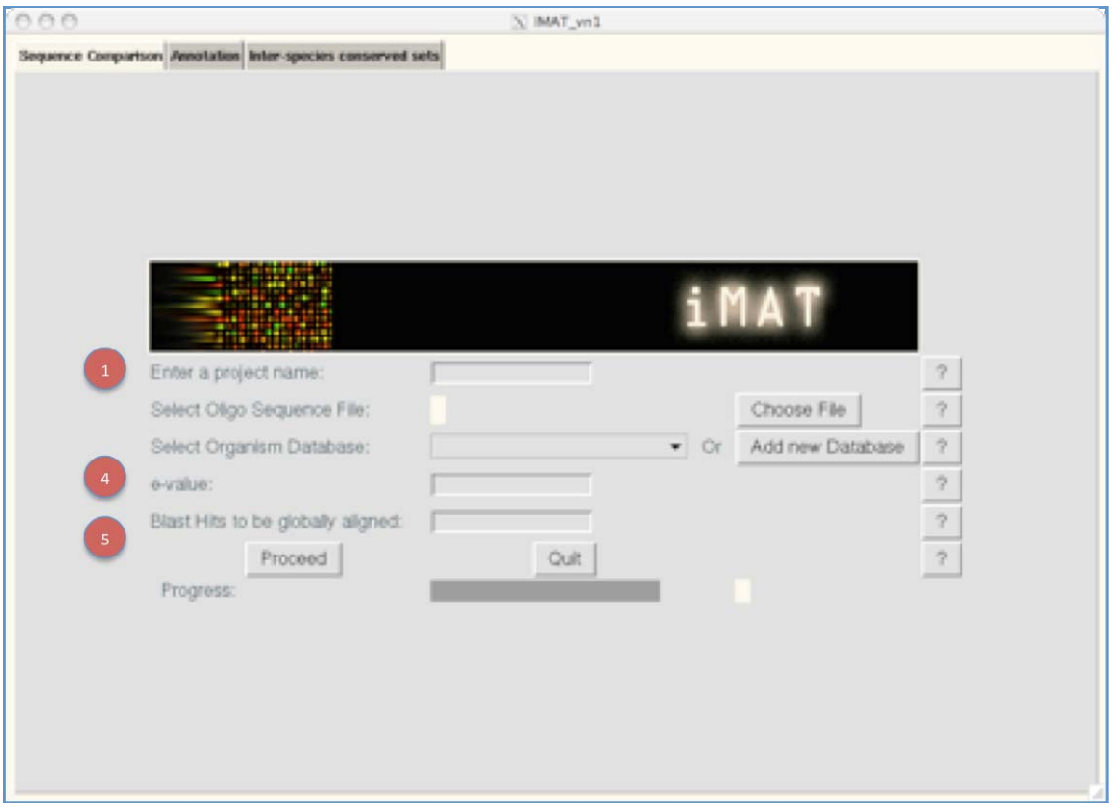


**Figure 3-10 Sequence comparison tab: project name, E-value, BLAST hits entry fields**

1 (Figure 3-10) This part creates the project folder for the analysis in the program directory. It should be a unique name, as it serves as a file name for the different result and report files created with iMAT.

4 (Figure 3-10) The E-value influences the outcome of the sequence alignment. A too low E-value might yield fewer results. Too high an E-value lowers the accuracy of the BLAST hits, but this is partly put into perspective by the later global alignment as it only takes the number, as specified, of best BLAST hits to be globally aligned. If left blank the default E-value of 100,000 will be set.

5 (Figure 3-10) In this step the user can select how many high scoring BLAST hits shall be globally aligned with the global alignment part within iMAT. The global

alignment has the purpose to align probe sequences over their full length against the sequence hits that were found by BLAST.

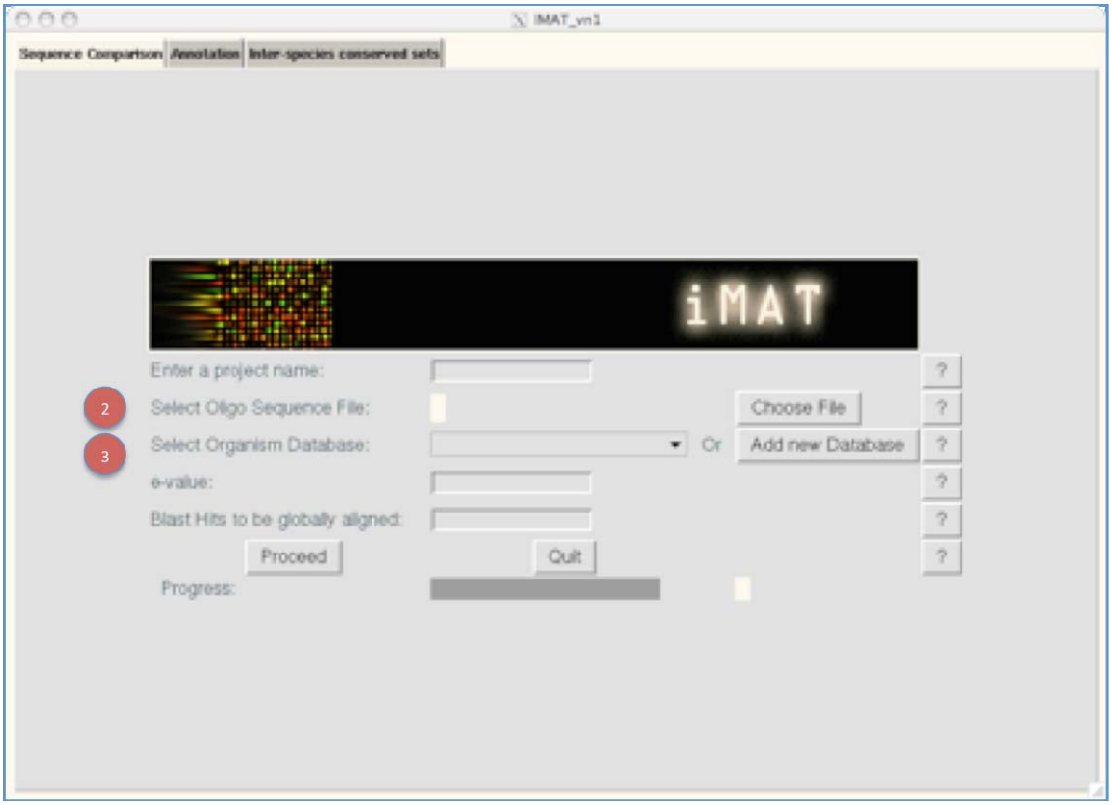


**Figure 3-11 Sequence comparison: oligonucleotide file selection, organism database**

2 (Figure 3-11) Here the user enters the path to the Agilent oligonucleotide sequence file, which is already in the tab-delimited format as shown in Figure 2-2, including Agilent identifier and oligonucleotide sequences have to be specified. This enables iMAT to perform the BLASTN alignment and the further global alignment.

3 (Figure 3-11) To start the sequence alignment a pre-formatted database must be selected. It is also possible in this step to add a new organism database, if a new organism database must be aligned against microarray probe sequences. The databases must be pre-formatted that the standalone BLAST version is able to read the sequences and compare them to each other.

When adding a new organism database some steps still remain manually as agreed upon with the user. The organism name and the path to the database have to be entered into the meta_data.txt file (Figure 3-12) located inside the program directory.

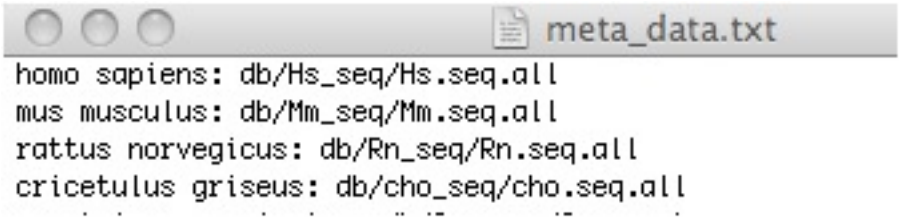


**Figure 3-12 Example of the meta_data.txt file which contains all necessary paths to the organism databases**

The database file can be selected with the button "add new database" which prompts a file selection window (Figure 3-13) to appear so the user can navigate to the database file. Adding the database will prompt the user to confirm that the database will be formatted by the program. As mentioned in chapter 2.4.1 a specific formatting of an organism database is needed to perform BLAST sequence alignment.

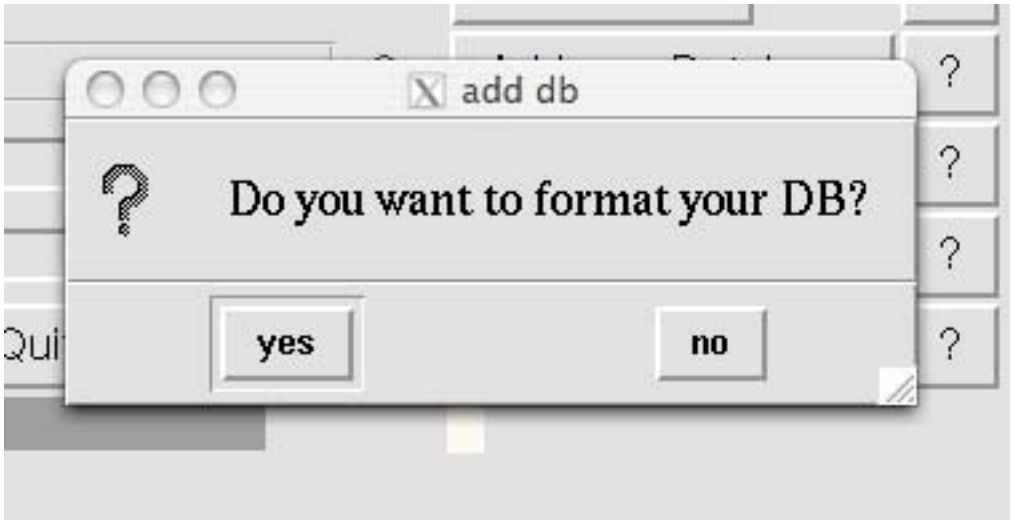


**Figure 3-13 Add new database confirmation**

After the formatting the database a message will be displayed to inform the user if it has been successfully formatted (Figure 3-14).

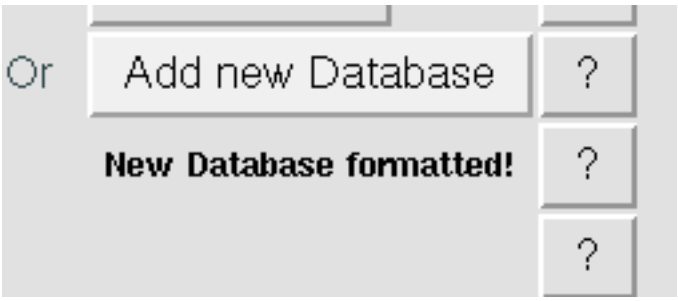


**Figure 3-14 Add new database confirmation message**

Now the newly formatted database can be selected from the dropdown list where all organism databases are listed. Including this step in the program ensures that the former manual step of formatting the databases is now part of the program and does not have to be performed outside iMAT.

### *3.4.2 Annotation tab*

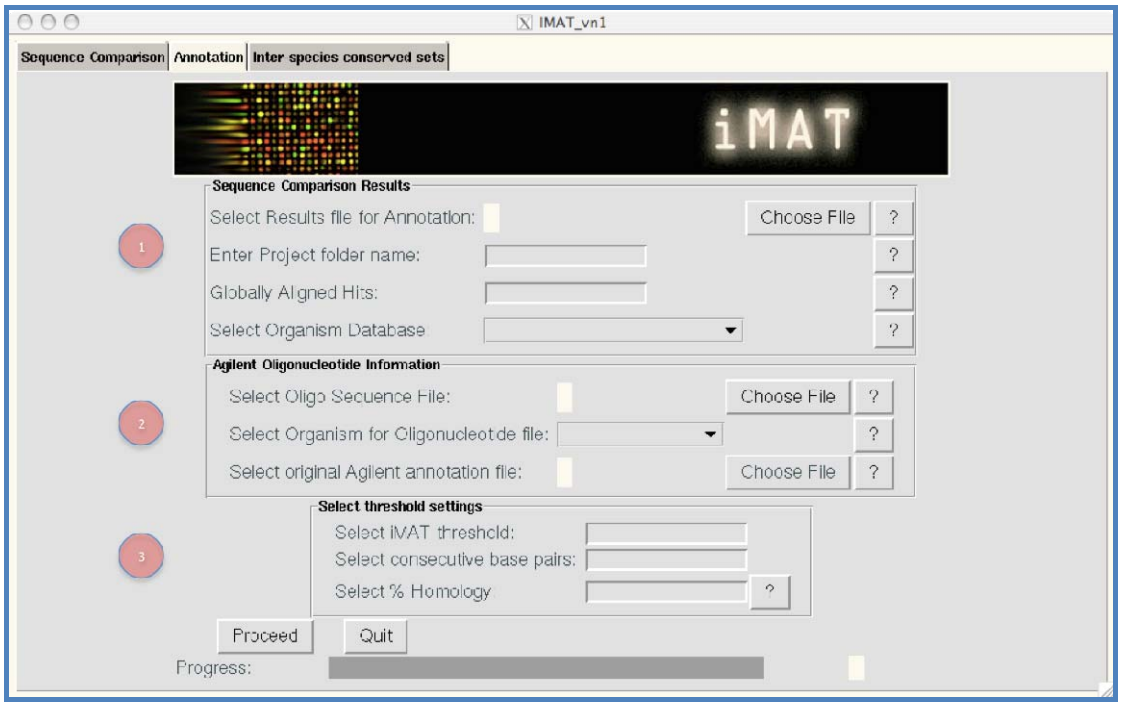


**Figure 3-15 Annotation tab**

1 (Figure 3-15) The first part uses results files from the sequence comparison step.

2 (Figure 3-15) The second part uses the information contained in user prepared files such as oligonucleotide sequences and available annotation data from Agilent.

3 (Figure 3-15) The third part is for selecting the thresholds for the alignment analysis. To make it easier for the user to identify the different sections they were framed. The cut-off values relate to the alignment analysis, allowing the user to set a minimum iMAT score, % homology and if the hits should have a minimum of 16 matching consecutive base pairs.

### 3.4.3 Inter-species conserved sets tab

As one of the aims of this project was to perform cross-species analysis, this part was created to perform these analyses within iMAT.

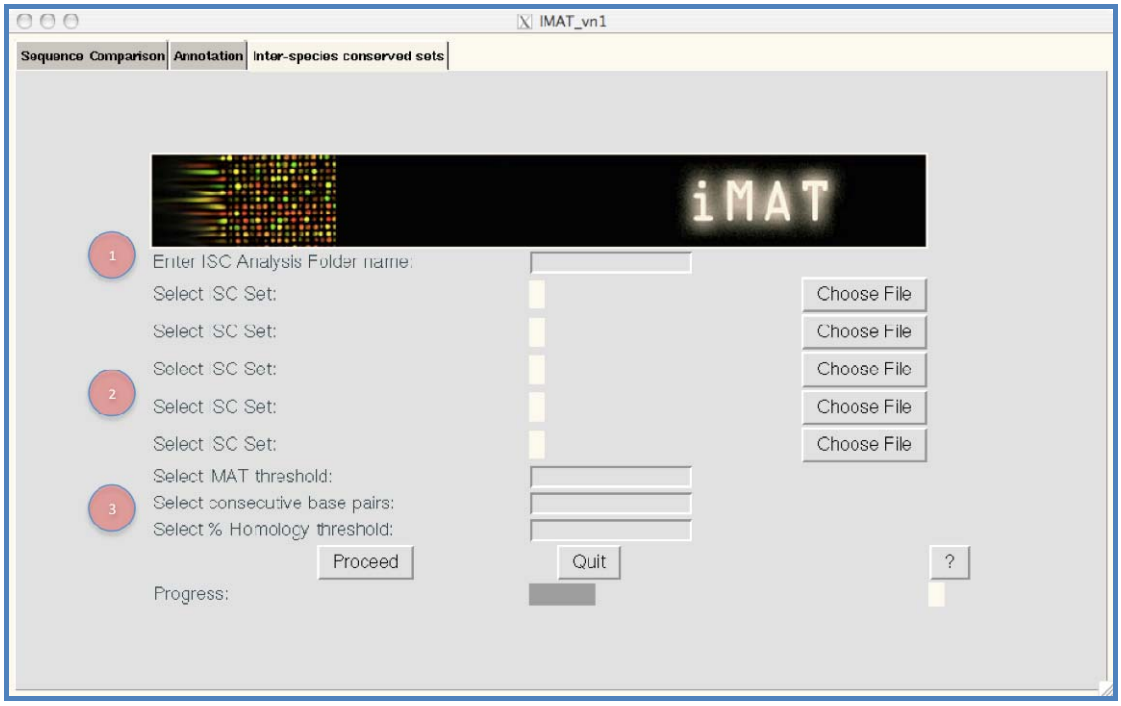


**Figure 3-16 ISC - comparison**

This section is intended for the user to investigate inter-species conserved probe sets of microarray probes aligned against different organisms. In the future it could also help to investigate which microarray would fit best for a selected species.

1 (Figure 3-16) A unique analysis folder name must be entered for storing of the result files. As more than one result file from the annotation step will be compared with each other this analysis is not associated with a specific project any more.

2 (Figure 3-16) Two to five analysis files, created by the annotation and alignment analysis step, can be selected and compared with each other. A results file with the microarray probe identifier and their annotation will be created as well as a chart indicating how many probes of the microarray are similar to all selected organisms.

3 (Figure 3-16) This part is intended to allow the user to set specific thresholds for this analysis. For example at which iMAT score the different result files shall be compared with each other. This means that only results with a specified iMAT score will be taken, and checked in how many of the selected files they exist.

iMAT checks of all entered data. It checks if the entered project folder does not exist yet and removes illegal characters (tab separator or whitespace). It checks if selected files are valid, if the entered numbers for BLAST hits to be globally aligned and the E-value are in a correct format (positive number) and if the threshold settings are within acceptable range.

Prior to the analysis files have to provided in a distinct format consistent with the original Agilent annotation files from the Agilent website. These can change the format with every new release; therefore it was not possible to parse the files into the necessary format within the automated program process.

In case the sequence comparison has already been performed the user is able to annotate previously obtained results. This is useful because the sequence comparison can take a very long time, depending on the E-value, BLAST settings and used computer (up to 4 days). So this was implemented as a way to store the initial result file but allow the user to perform additional analysis on the sequence comparison results.

Sequence comparison and annotation in this manner can only be performed separately and not at the same time, as iMAT does not support parallel processing.

In addition, help for the user is provided throughout the program interface as indicated by the question mark buttons. This interactive help (Figure 3-17) is useful to provide the user with a guideline through the whole program and not having to read user manuals before being able to start analysis.

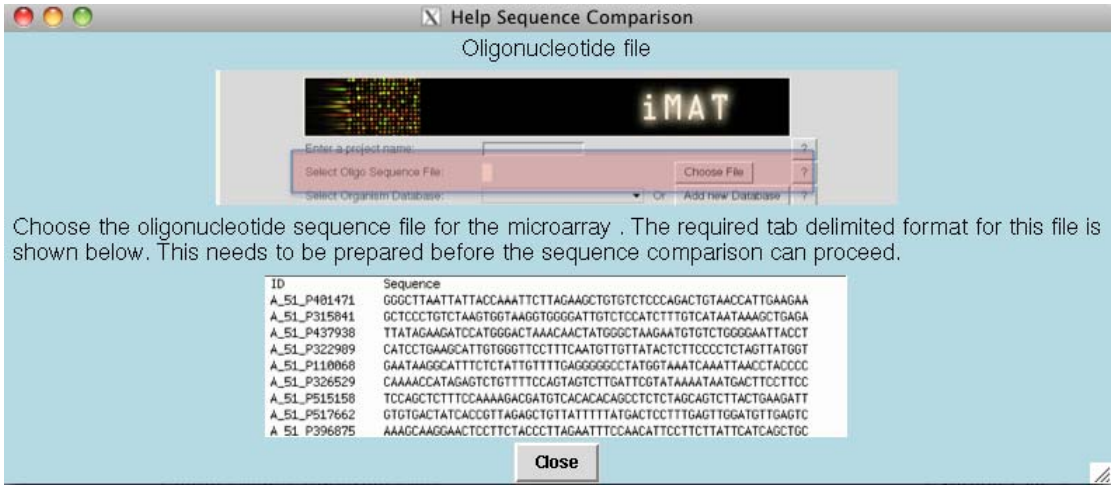


**Figure 3-17 Example of interactive help window in iMAT**

Those windows pop up in real time when the according help button for an entry file is pressed to explain the user which files are supposed to be used in this step to ensure a smooth program workflow.

# 4 Conclusion and future work

This chapter summarises how the aims of the project were met, as well as an interpretation of the results obtained. Furthermore, the challenges met and future work is explained.

## 4.1 Summary of the project aims met and discussion of the results

The main aim of the project was to develop a platform that helps to find cross-species conserved sequences in order to facilitate in the knowledge of a possible outcome for cross-species microarray experiments. Secondly it was important to create a novel informative scale for the user to decide how suitable a microarray might be for a selected species. Thirdly it was intended to show that cross-species conservation exists among mammalian organisms. To meet these goals several analysis steps were implemented to gather additional gene information and analyse the global sequence alignment of the BLAST hits.

After an initial validation process by aligning all the Agilent mouse oligonucleotide sequences of the microarray that was used for same- and cross-species expression studies against the UniGene mouse sequence database, different annotation steps were developed to enrich the sequence information with gene names and additional identifiers. Several approaches were combined to ensure a complete as possible annotation of found probes through the sequence comparison. More specifically BioMart's Perl API was used in combination with NCBI's eUtils to automate the process of gene annotation, this being the first aim of the project.

A graphical user interface was developed in Perl Tk to ease the use of this automated sequence comparison and annotation tool as specified. As many steps as possible were automated for the user to ensure a workflow within a reasonable time frame.

Several result files are created throughout the process. The most informative results are parsed into one single results file containing all information.

In addition graphical outputs were developed to give the user an additional summary about the results data.

Since 2006 the number of available sequences for hamster has tripled. To investigate the reliability of the results calculated with iMAT the probes identified by iMAT were compared to the correlation coefficient (chapter 2.8) of the signal intensities of those probes derived from heat shock experiments of mouse (3T3) and CHO dhfr-cells. The linear correlation coefficient (CC) is used to investigate the relationship between the signal intensities of two given samples. The overall correlation for the whole dataset of aligning mouse probes against the hamster database was calculated with 0.756 (75.6 %).

As iMAT is using a nucleotide BLAST alignment in combination with a custom global alignment the E-value settings were investigated (chapters 3.2.2, 3.2.3 and 3.2.4) to assess if differences in the E-value actually influence the final output of the sequence comparison. Results of the subsets selected showed mostly not a lot difference in correlation of signal intensities derived from the cross-species experiments. Most significant was the difference in total hits found when probes were aligned against the different organism databases. An explanation is, that the organisms investigated are all studied to a different extent, meaning that databases of well-known and well-studied organisms contain much more information and of course sequence entries.

At an E-value of 100,000 almost all mouse probes (4 IDs were missing in all alignments) could be aligned against the rat, human and hamster sequence database. As the setting for BLAST hits to be globally aligned was maintained the same (ten)

throughout all experiments every time only the best ten BLAST hits were aligned globally. Concluding that even if a very unreliable E-value is selected, the global alignment of only the best hits takes care of inaccurate BLAST results. Setting a very high E-value for the BLAST step results in more hits that can be globally aligned hence this global alignment can be further investigated for calculation of % homology and consecutive base pair stretches and more hits can be annotated, if annotation is available. By applying the different parameters, annotation matches, % homology, consecutive base pair stretches and iMAT score, it became apparent that selecting only one parameter for the identification of inter-species conserved probes and the homology between different organisms alone was not informative and accurate enough.

This suggested to base the reliability scale on parameters such as iMAT score, % homology between the sequences, consecutive base pair stretches and matches in gene annotation between probe and found hit.

Based on the various subset analysis (chapters 3.2.3, 3.2.5, 3.2.6, 3.2.7 and 3.2.8) a reliability scale such as shown in Figure 3-8 could be an informative scale indicating how trustworthy the obtained analysis results with iMAT are.

The results showed that the iMAT algorithm is very well capable of performing local sequence alignments, additional global alignments and annotate the hits in order to identify cross-species conserved probes and give insights on cross-species homology in general as an indication for a “generic” microarray.

## 4.2 Challenges

During the course of this projects sometimes challenges were encountered, that slowed down the work. These challenges are common to the particular field of work of Bioinformatics and Transcriptomics. When using different database sources for

information enrichment challenges with gene annotation and different database identifiers are easily encountered for finding truly unique identifier in transcript databases or simply the conversion of database identifiers as well as the connection between different biological databases relating specific information.

Standards might already have been developed but sometimes they are not well adopted by the community.

## 4.3 Future Work

At the end of this project several approaches for annotation of the found hits as well as global alignment analysis were combined. Together with a graphical user interface the user is now provided with an easy to use application. This prototype was implemented and tested, but there is still a lot of work to be done to enrich information gathering and add reliability.

The first improvement that should be made is to add the possibility to analyse signal intensity values or correlation coefficients of signal intensities in context with the sequence comparison and annotation to add confidence to predicted results.

It could also be useful to add more selection parameters to the interface so the user can select manually, which annotation information to retrieve for example gene description, gene ontology terms, etc.

Further sequence alignment methods such as ClustalW could be investigated in combination with iMATs own global alignment. This could help in comparison to the current usage of BLASTN and the global alignment and give more confidence to identified cross-species conserved probes.

As Agilent is not the only microarray platform producer it might prove to be valuable for the community to extend the iMAT platform to the Affymetrix format.

# 5 REFERENCES:


NCBI - formatdb 14.08.2009http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/formatdb_fastacmd.html

NCBI -eUtils 14.08.2009http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=coursework&part=eutils

NCBI-Blast 28.05.2009http://blast.ncbi.nlm.nih.gov/Blast.cgi

NCBI-UniGene 14.08.2009http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene

Research Computing Center 10.06.2009http://rcc.uga.edu/applications/bioinformatics/ncbiblast2210/formatdb.html#C

The Perl Directory: About Perl 28.05.2009http://www.perl.org/about.html

ADJAYE, J., HERWIG, R., HERRMANN, D., WRUCK, W., BENKAHLA, A., BRINK, T. C., NOWAK, M., CARNWATH, J. W., HULTSCHIG, C., NIEMANN, H. & LEHRACH, H. (2004) Cross-species hybridisation of human and bovine orthologous genes on high density cDNA microarrays. *BMC Genomics,* 5.

Affymetrix website 20.08.2009http://www.affymetrix.com/products_services/research_solutions/index.affx

Agilent Technologies 26.05.2009http://www.chem.agilent.com/en-us/products/instruments/dnamicroarrays/wholemousegenomeoligomicroarraykit/pages/default.aspx

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

Bioperl 20.08.2009http://www.bioperl.org/wiki/Main_Page

BROWN, P. O. & BOTSTEIN, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet,* 21**,** 33-7.

CHEN, Z., WANG, W., LING, X., LIU, J. & CHEN, L. (2006) GO-Diff: Mining functional differentiation between EST-based transcriptomes. *BMC Bioinformatics,* 7**,** 72.

DE LEON GATTI, M., WLASCHIN, K. F., NISSOM, P. M., YAP, M. & HU, W.-S. (2007) Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells undergoing butyrate treatment. *Journal of Bioscience and Bioengineering,* 103**,** 82-91.

ERNST, W., TRUMMER, E., MEAD, J., BESSANT, C., STRELEC, H., KATINGER, H. & HESSE, F. (2006) Evaluation of a genomics platform for cross-species transcriptome analysis of recombinant CHO cells. *Biotechnology Journal,* 1**,** 639-650.

FRAZER, K. A., ELNITSKI, L., CHURCH, D. M., DUBCHAK, I. & HARDISON, R. C. (2003) Cross-Species Sequence Comparisons: A Review of Methods and Available Resources. *Genome Research,* 13**,** 1-12.

GELLISSEN, G. P. D., STRASSER, A. W. M. & SUCKOW, M. (2005) Key and Criteria to the Selection of An Expression Platform. IN PROF. DR. GERD, G. (Ed.) *Production of Recombinant Proteins.*

GÜZLEK, H. (2006) Development of Bioinformatics Systems for Cross-species Transcriptome Analysis. *Cranfield Health.* Silsoe, Cranfield University.
JALURIA, P., KONSTANTOPOULOS, K., BETENBAUGH, M. & SHILOACH, J. (2007) A perspective on microarrays: current applications, pitfalls, and potential uses. *Microb Cell Fact,* 6**,** 4.
JAYAPAL, K. P., WLASCHIN, K. F., HU, W. S. & YAP, M. G. S. (2007) Recombinant protein therapeutics from CHO Cells - 20 years and counting. *Chemical Engineering Progress,* 103**,** 40-47.
JENKINS, N., PAREKH, R. B. & JAMES, D. C. (1996) Getting the glycosylation right: implications for the biotechnology industry. *Nat Biotechnol,* 14**,** 975-81.
KANE, M. D., JATKOE, T. A., STUMPF, C. R., LU, J., THOMAS, J. D. & MADORE, S. J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res,* 28**,** 4552-7.
KAUFMAN, R. J., SHARP, P. A. & LATT, S. A. (1983) Evolution of chromosomal regions containing transfected and amplified dihydrofolate reductase sequences. *Mol Cell Biol,* 3**,** 699-711.
KENT, W. J. (2002) BLAT,ÄîThe BLAST-Like Alignment Tool. *Genome Research,* 12**,** 656-664.
KUYSTERMANS, D., KRAMPE, B., SWIDEREK, H. & AL-RUBEAI, M. (2007) Using cell engineering and omic tools for the improvement of cell culture processes. *Cytotechnology,* 53**,** 3-22.
LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-2948.
MAKALOWSKI, W. & BOGUSKI, M. S. (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc Natl Acad Sci U S A,* 95**,** 9407-12.
MEAD, J. (2005) Development of Methods to Evaluate Inter-species Gene Expression Data. *Department of Analytical Science and Informatics.* Silsoe, Cranfield University.
NEEDLEMAN, S. B. & WUNSCH, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology,* 48**,** 443-453.
NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. (2000) T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology,* 302**,** 205-217.
SANDIG, V., ROSE, T., WINKLER, K. & BRECHT, R. (2005) Mammalian Cells. IN PROF. DR. GERD, G. (Ed.) *Production of Recombinant Proteins.*
SMEDLEY, D., HAIDER, S., BALLESTER, B., HOLLAND, R., LONDON, D., THORISSON, G. & KASPRZYK, A. (2009) BioMart--biological queries made easy. *BMC Genomics,* 10**,** 22.
SOUTHERN, E. M. (2001) DNA microarrays. History and overview. *Methods Mol Biol,* 170**,** 1-15.
THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics,* Chapter 2**,** Unit 2.3.
THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through

sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22**,** 4673-80.

URLAUB, G. & CHASIN, L. A. (1980) Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. *Proceedings of the National Academy of Sciences of the United States of America,* 77**,** 4216-4220.

VERDUGO, R. A. & MEDRANO, J. F. (2006) Comparison of gene coverage of mouse oligonucleotide microarray platforms. *BMC Genomics,* 7**,** 58.

WANG, Z., LEWIS, M., NAU, M., ARNOLD, A. & VAHEY, M. (2004) Identification and utilization of inter-species conserved (ISC) probesets on Affymetrix human GeneChip(R) platforms for the optimization of the assessment of expression patterns in non human primate (NHP) samples. *BMC Bioinformatics,* 5**,** 165.

WLASCHIN, K. F., NISSOM, P. M., GATTI MDE, L., ONG, P. F., ARLEEN, S., TAN, K. S., RINK, A., CHAM, B., WONG, K., YAP, M. & HU, W. S. (2005) EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotechnol Bioeng,* 91**,** 592-606.

YEE, J. C., WLASCHIN, K. F., CHUAH, S. H., NISSOM, P. M. & HU, W.-S. (2008) Quality assessment of cross-species hybridization of CHO transcriptome on a mouse DNA oligo microarray. *Biotechnology and Bioengineering,* 101**,** 1359-1365.

# A. APPENDIX for ISC subset selection

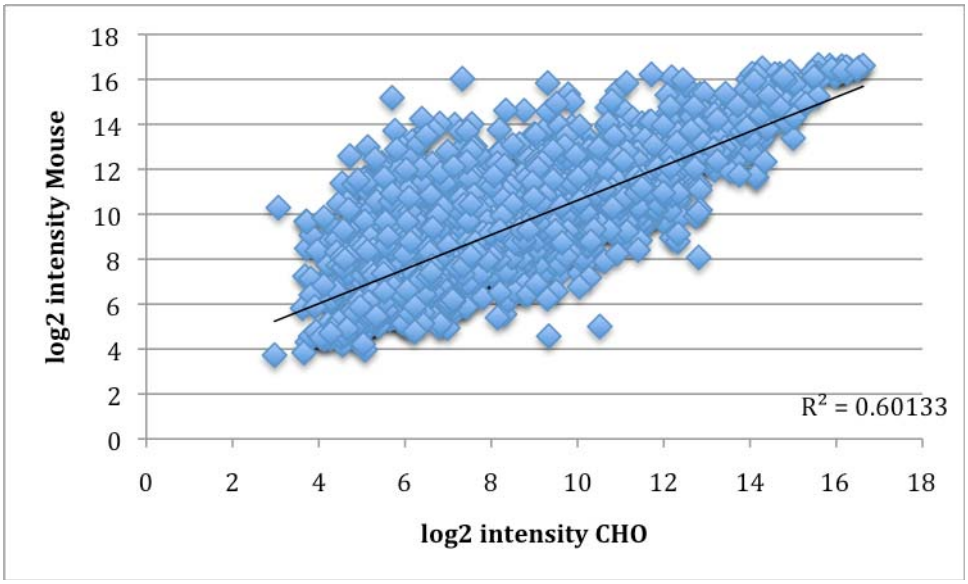


**Figure A-1 Scatter plot of ISC subset based on matching annotation and consecutive base pairs length > 15**

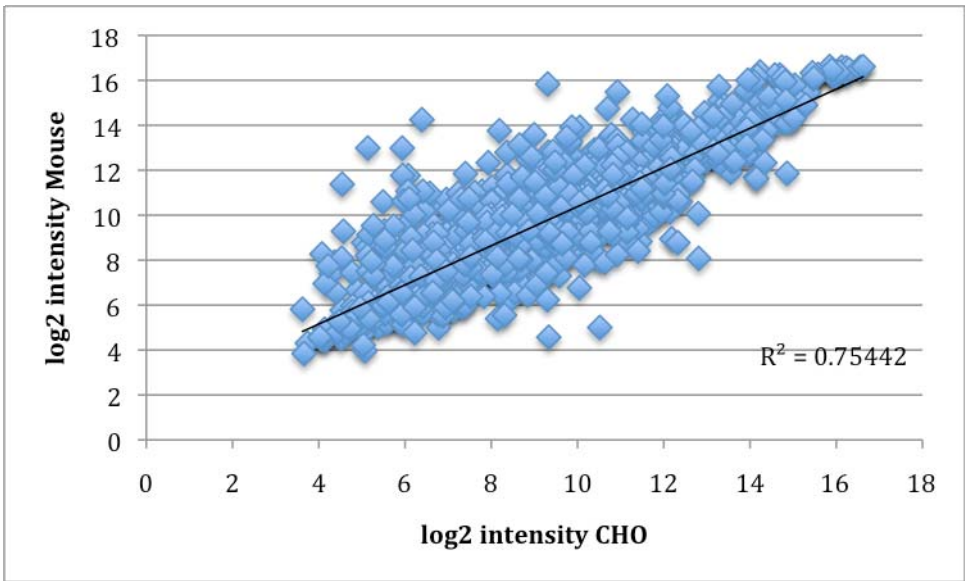


**Figure A-2 Scatter plot of ISC subset based on matching annotation and sequence homology ≥ 90% at and E-value of 100000**

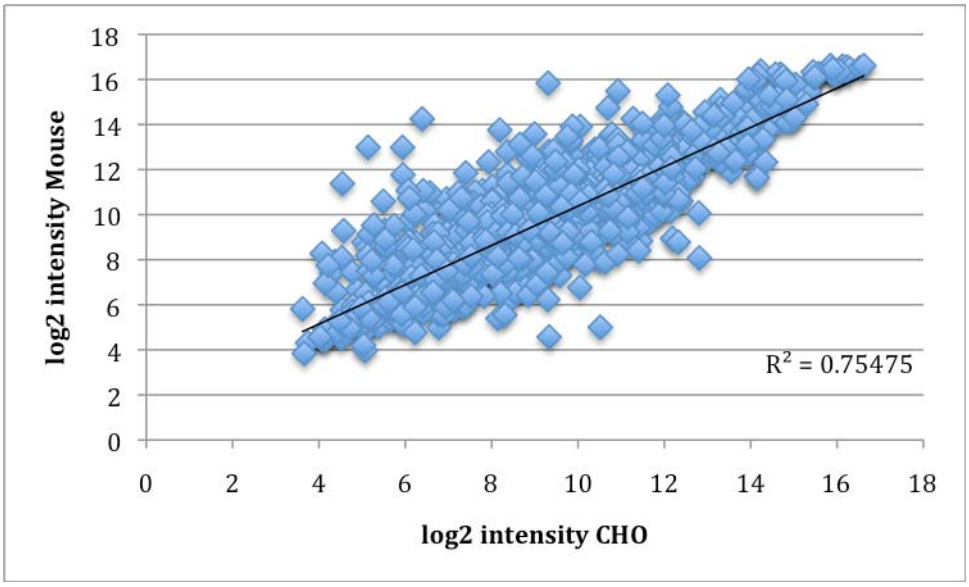


**Figure A-3 ISC subset based on annotation matches and iMAT score ≥150, ≥ 90% homology, > 15 base pairs**

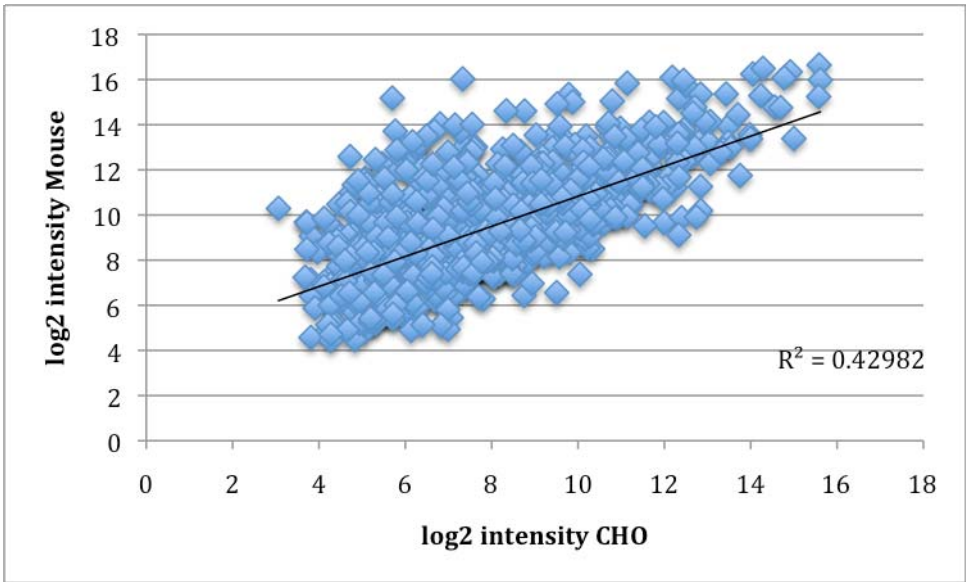


**Figure A-4 ISC subset based on annotation matches, iMAT score ≥ 70, sequence homology ≥ 60 % and a consecutive base pair stretch > 15 nucleotides**

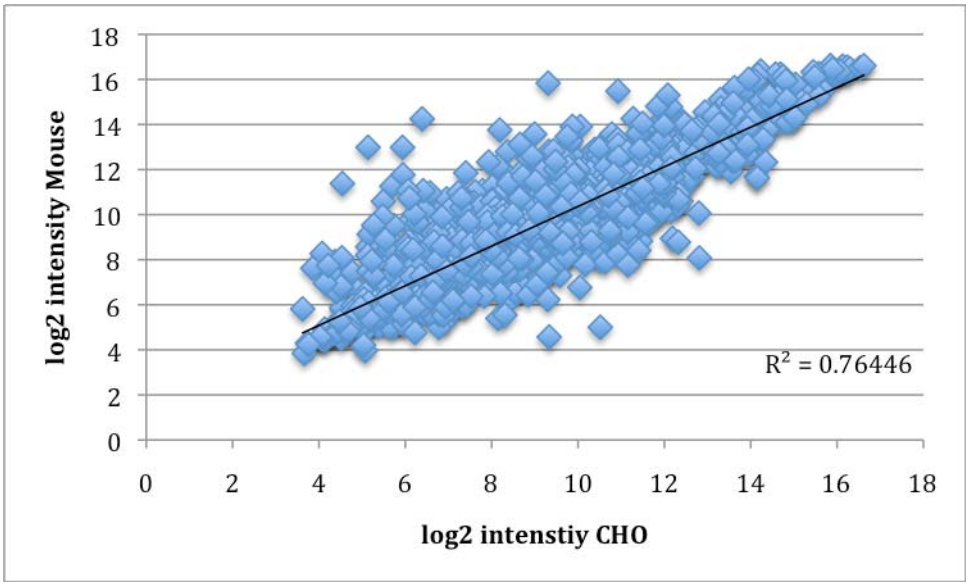


**Figure A-5 ISC subset selection based on iMAT ≥150, %homology ≥ 90, consecutive base pair stretch > 15**

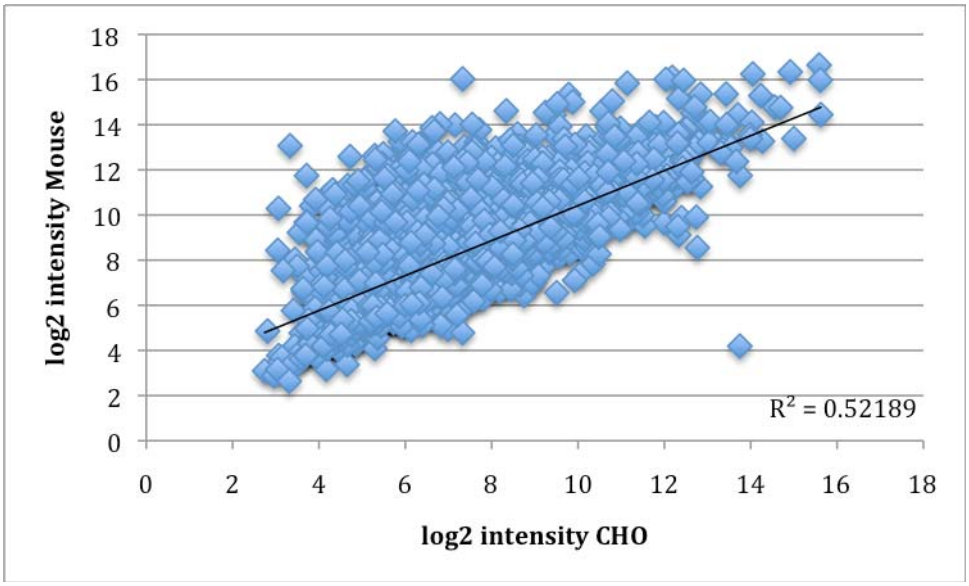


**Figure A-6 ISC subset selection based on iMAT ≥ 70, % sequence homology ≥ 60 %, consecutive base pair stretch > 15**