LU TP 03-22
May 2003

# Bioinformatics-based Exploration of Regulatory Regions in the Human Genome

Srinivas Veerla

Masters Thesis in Bioinformatics

Thesis Advisors: Markus Ringnér and Jari Häkkinen

Complex Systems Division
Department of Theoretical Physics
Lund University


LUND
UNIVERSITY

**Abstract**

The microarray technology has made it possible to experimentally measure the expression levels of many individual genes simultaneously. Microarrays can, for example, be used to generate gene expression profiles of cells under different environmental conditions. These gene expression profiles, can then be used to cluster genes into co-expressed groups. It is reasonable to assume that co-expressed genes may also be co-regulated, and thus may share regulatory sequences in their non-coding regions. The aims of this project were two-fold; first to build a software package which can create a dynamic database ACID (array clone information database) consisting of information about the clones used in microarray experiments and second, by linking the clones to the genome, to investigate their transcriptional regulatory regions for presence of regulatory motifs and basal promoter regions.

# Contents



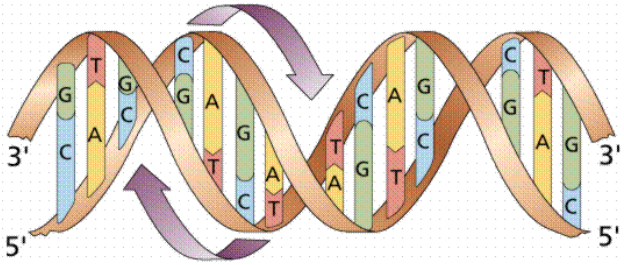
# List of Tables

# List of Figures

# 1 Introduction

All living cells, without any known exception, store their heriditary information in the form of double-stranded molecules of DNA (deoxyribonucleic acid). In the cell, DNA is stored in helical form with its complementary strand, and each strand is used as a template to make a complementary strand of itself (see figure 1), DNA is always formed of the same four types of nucleotides (bases) - A (adenine), T(thymine), G(guanine), C(cytosine). RNA (ribonucleic acid) is single stranded and it is a copy of one of the strands of DNA. RNA is slightly different from DNA, since it has U (uracil) instead of T (thymine).

Figure 1: Double-stranded DNA




Schematically a nucleotide has a 5' end and a 3' end. In a DNA strand the 3' end of each nucleotide is attached to the 5' end of the adjacent nucleotide. Therefore a DNA strand has a direction. By convention, double-stranded DNA is represented by two strands of sequences as follows; one is starting at the 5' end and continuing from left to right to the 3' end of the molecule, and is called the positive strand and the other is starting at the 3' end and continuing from left to right to the 5' end of the molecule, and is called a negative strand, see figure 2. RNA is represented by a single strand sequence and its direction is the same (5' to 3') as for the positive DNA strand

Figure 2: Postive and negative DNA strands

5'-------A T C G A C G A A T T A A C T G A G A C A ---- 3'

3'-------T A G C T G C T T A A T T G A C T C T G T ----5'

## 1.1 The central dogma

A segment of DNA sequence (information) corresponding to a protein is called a gene and it serves as the template for the synthesis of RNA, much as it does for its

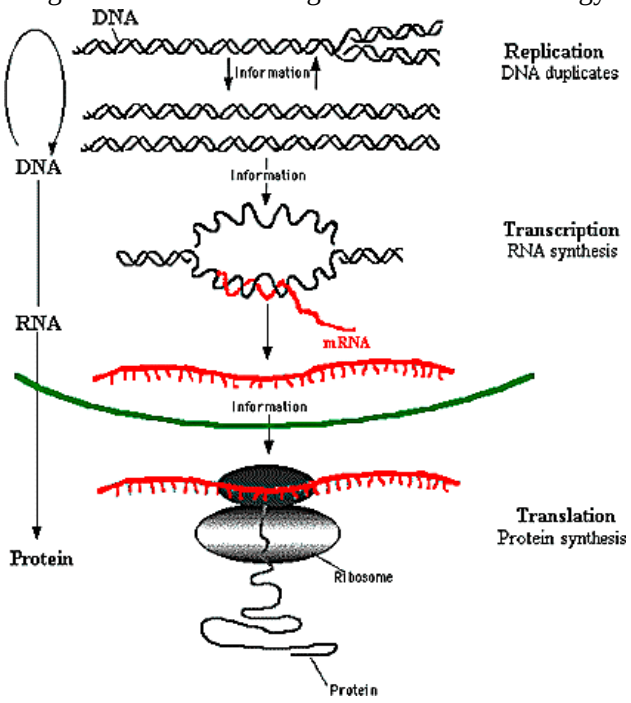
own replication. The majority of genes are expressed as the proteins they encode. The process occurs in two steps:

- Transcription = DNA to RNA
- Translation = RNA to protein

together they make up the "central dogma" of biology (see figure 3): DNA to RNA to protein. In general, proteins are the molecules that carry out the functions in cells.

DNA directs the synthesis of RNA by the transcription process, in which a gene is transcribed into a messenger RNA (mRNA). Messenger RNA carries coded information to ribosomes, which reads the mRNA sequence and use it to direct protein synthesis. This process is called translation. Figure 3, also illustrates the process of protein production. (Figure 3 is taken from National Health Museum Graphic gallery: http://www.accessexcellence.org/AB/GG/central.html).

Figure 3: The central dogma of molecular biology




In all cells, the expression of genes is regulated. Instead of manufacturing its full repertoire of possible proteins at full tilt all the time, cells adjust the rate of transcription of different genes independently, according to need. Stretches of

regulatory DNA are interspersed among the segments that code for protein, and special protein molecules, so-called transcription factors, bind to these noncoding regions to control the local rate of transcription. The quantity and organization of the regulatory and other noncoding DNA vary widely from one class of organisms to another. In this way, the total genetic information of a cell is embodied in its complete DNA sequence, the genome of the organism.

Protein-encoding genes in Eukaryotes (higher organisms) among other things, have exons, whose sequence actually encodes to protein; a transcription start site; a basal promoter or core promoter located within about 40 bp of the start site, and an "upstream" promoter, which may extend farther upstream.

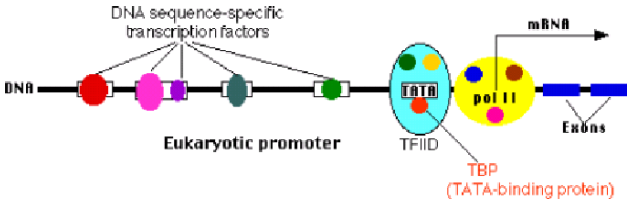



Figure 4: Transcription factors

The basal promoter is the transcription start site, where transcription of the gene into mRNA begins. The basal promoter typically contains a sequence of 6 bases (TATAAA) called the TATA box [11]. It is bound by Transcription Factor IID (TFIID) which is a complex of some 10 different proteins including TATA-binding protein (TBP) (see figure 4). A basal or core promoter is found in all protein-encoding genes.

Many different genes in many different types of cells share the same transcription factors - not only those that bind at the basal promoter but even some of those that bind upstream. What turns on a particular gene in a particular cell is probably its unique combination of promoter sites and the expression levels of the transcription factors that bind to them.

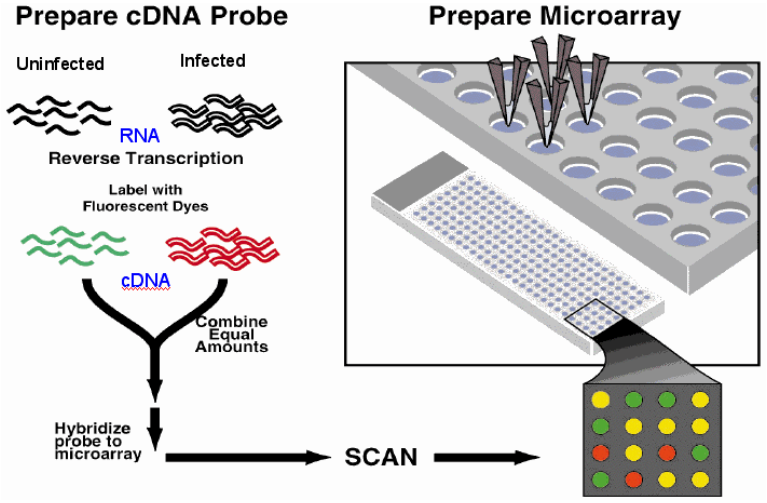
The latest estimates are that the DNA of a human cell, contains approximately 35,000 genes. They are expressed in the following ways: 1) Some of these genes are expressed in all cells all the time. These so-called housekeeping genes are responsible for the routine metabolic functions (e.g. respiration) common to all cells. 2) Some are expressed as a cell enters a particular pathway of differentiation. 3) Some are expressed all the time in only those cells that have differentiated in a particular way. 4) Some are expressed only as conditions around and in the cell

change. For example, the arrival of a hormone may turn on (or off) certain genes in that cell. It is therefore of great interest to investigate the expression profiles of different cells under varying conditions to characterize their genetic programs.

## 1.2 cDNA microarray technology

Microarrays are revolutionary measurement devices that allow biological exploration on a genomic scale (Schena et al. 1995). Microarrays allow for parallel analysis of the expression of thousands of genes, hence it is possible to investigate the global variation in transcriptional expression profiles. cDNA microarray assays use DNA base-pairing to indirectly measure sequence specific mRNA concentrations, which are believed to reflect the expression levels of genes. The central device of the method is a library of many DNA sequences called clones, which ideally are specific for one gene each, printed and immobilized onto a glass surface. RNA from a sample is converted to fluorescently tagged DNA, which is hybridised (base-paired) against the printed library, such that the amount of DNA that has hybridised (base-paired) to each clone can be detected via fluorescence microscopy. In this way, the RNA expression level in a sample can be measured for all clones printed on the microarray (see figure 5).

Figure 5: Microarray technology




Microarray experiments are not only about visualizing and quantifying the fluorescence signal from a microarry experiments, but also about analysing gene expression under experimental conditions versus reference conditions to determine

whether observed differences are significant or not. There are many sources of noise and variability in microarray data, including experimental sources as image scanning inconsistencies, issues involving computer interpretation and qualification of spots, hybridization variables such as temperature and time discrepancies between experiments, and experimental variation caused by differential probe labelling and efficacy of RNA extraction. It is assumed that co-expressed genes may also be co-regulated, and thus may share regulatory sequences in the non-coding regions surrounding them in the genome. Therefore, searching for the presence of common regulatory motifs for genes that are expressed in a similar fashion may further strengthen results from gene expression profiling studies.

## 2 Project goals

Microarray experiments often end up with a large number of genes which are of interest. However, to obtain information about the genes one by one may be exhausting, and therefore it is of use to build a database, which can assist microarray users to extract up-to-date information about all clones used in microarray experiments simultaneously. For this aim we developed **A**CID (**A**rray **C**lone **I**nformation **D**atabase) [5].

In microarray studies one often finds genes that are expressed in a similar fashion. It is assumed that genes which are expressed in a similar fashion may be regulated by common transcription factors. Therefore we decided to write an application tool which can identify common regulatory motifs in the upstream sequences of these genes. There are many tools available for investigating regulatory regions for simpler organisms like yeast (see for example [6, 7]), so we decided to write an application tool, which can investigate the regulatory regions for more complex mamalian species such as human and mouse. To make it simple for users, this application should only take microarray clone identifiers as input, so that microarray users need not to search for genome information (chromosome, start base, end base and strand) for each clone, to extract upstream regions and to find regulatory motifs. In addition to finding regulatory motifs, the application should give complete information about the clones from **A**CID.

mRNA RefSeqs (Reference Sequence) are considered to be a comprehensive, integrated, non-redundant set of full-length transcript (RNA) sequences, in other

words “well-defined genes” [1]. So we decided to use RefSeq information for our research questions. In particular, we use the mRNA RefSeq associated with a gene as a starting point and search for regulatory motifs in the upstream regions of RefSeqs positioned in the genome. A question is, whether a RefSeq start position reflects the transcriptional start site of the gene, which is crucial for our common regulatory motif search to make sense. If not, the upstream regions we extract from the genome sequence will not be the regulatory regions. To address this issue we investigated the presence of “core or basal promoters” in these upstream sequences.

# 3 Materials and methods

To get information that is helpful in solving our research questions, we decided to write packages (sets of related applications), which can retrieve information from the Santa Cruz Genome Browser [2] and Genome database [3], GenBank [15] and the UniGene database (http://www.ncbi.nih.gov/UniGene) and form a new dynamic database for cDNA microarray clones called **A**CID [5] (**A**rray **C**lone **I**nformation **D**atabase). The Santa Cruz browser & database contains genome sequences with annotation for many species including human and mouse. GenBank is a repository for measured sequences, including very many short sequences of mRNA. Many of these latter sequences correspond to the same full length transcript. Therefore, the mRNA sequences in GenBank are partitioned into non-redundant set of gene-oriented clusters called UniGene clusters, such that each cluster ideally corresponds to a unique gene. As more and more sequences become stored in GenBank, UniGene gets periodically updated and each version is assigned a build number. Packages were written in such way that we can retrieve information for any species present in UniGene. To increase the efficiency of retrieval of information from **A**CID, the packages create independent tables for each species. Depending on the UniGene build number of the species, the database can be updated independently for each species. Both UniGene and GenBank information were downloaded from ftp://ftp.ncbi.nih.gov/repository/UniGene.

Currently, **A**CID contains all *Homo sapiens* and *Mus musculus* cDNA clones present in UniGene. The information for each clone includes:

- GenBank accession numbers to sequence reads [15]
- direction (5’ or 3’ read) of the sequence reads

- which UniGene cluster it belongs to
- on which chromosome it is located
- cytoband information
- LocusLink information [1]
- OMIM (Online Medelian Inheritance in Man) information [14]
- its gene symbol and title
- its associated Gene Ontology terms [13]
- its associated mRNA Reference Sequence(RefSeq) [1].

**A**CID also contains information downloaded from the Genome Browser [3] about where RefSeqs [1] are positioned in genomes. Currently, this information is based on human genome assembly hg13 and mouse genome assembly mm2 [3]. The RefSeq information from the genome which we have in our database includes:

- which chromosome it is located on
- start and end base on the chromosome
- cytoband information

The above RefSeq information is required to retrieve the upstream regions from the genome sequence, to investigate for regulatory motifs as well as for TATA boxes.

For humans (using UniGene build number 160) ACID contains 2.5 million cDNA clones and 3.2 million EST sequences based on a approximately 111,000 UniGene clusters. 18,262 of the clusters are represented by at least one RefSeq mRNA sequence. For mouse (using UniGene build number 120), the corresponding numbers are 2.5 million clones, 3.1 million EST sequences, and approximately 90,000 clusters (12,860 with a RefSeq sequence). The database is implemented in MySQL and applications are written in Perl.

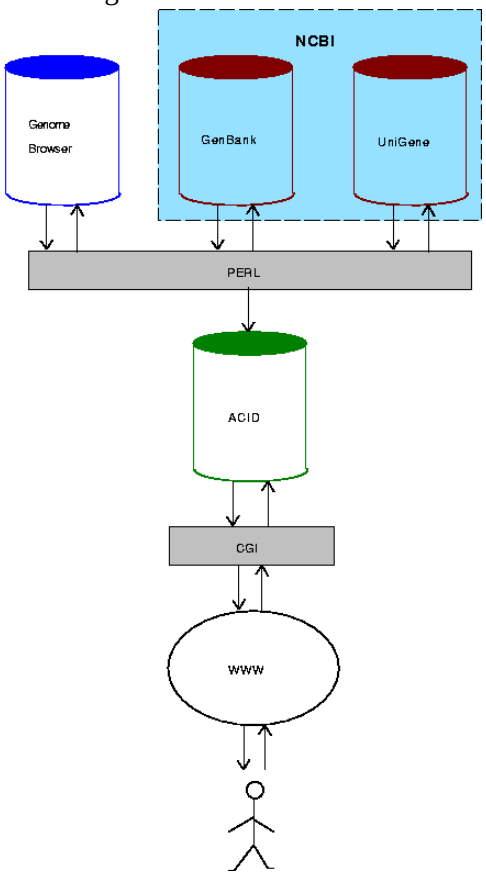
We have decided to write the applications in Perl because biological data is stored in enormous databases and text files. Sorting through and analyzing this data by hand (and it can be done) would take far too long, so the smart scientist writes computer tools to automate the process. Perl, with its highly developed capacity

to detect patterns in data, and especially strings of text, is the most obvious choice of programming language. The MySQL database is efficient and solace enough to implement bioinformatic projects and moreover it is free of cost.

## 3.1 ACID architecture

A huge amount of clone information is traditionally stored in a flat files by different organizations, for example in the Santa Cruz Genome Browser [2, 3], UniGene and GenBank [15]. To search information about clones we need to go to different web services. And to search information in a flat file is tedious and laborious job, so we decided to build **A**CID [5], which will hold publicly available information for clones used in microarray experiments. Our ambitious aim is to provide as much information about clones as possible. This database (**A**CID) will be updated automatically depending on the UniGene build number of the species, therefore it contains up-to-date information. Figure 6 shows the architecture of **A**CID. The first layer is the Perl layer, which downloads the given species information according to their build number. If the build number of the species in the UniGene database is greater than the build number of the species present in our database (**A**CID), then the information of that species is downloaded, which will be in the form of huge flat files. This Perl layer parses downloaded huge files to get the information we needed and stores it in **A**CID in an orderly fashion. For every new species, which are not present in **A**CID; the Perl layer dynamically creates new tables and stores the information of that species. In the next layer, the CGI layer, we use the CGI (Common Gateway Interface) scripting language to build a web page from where end users easily can search for the information using clone identifiers, RefSeqs or gene symbols as inputs. The web page is built in such a way that one can choose which information to be shown, for example, information related to the Gene Ontology, UniGene, the Genome Browser, and clone location (see figure 7). Maintainence of this database (**A**CID) is very simple, just to execute a small application of 5 to 6 lines and the complete database will be automatically updated with up-to-date information by comparing the build numbers present in **A**CID to UniGene build numbers of the species.

Figure 6: **A**CID architecture

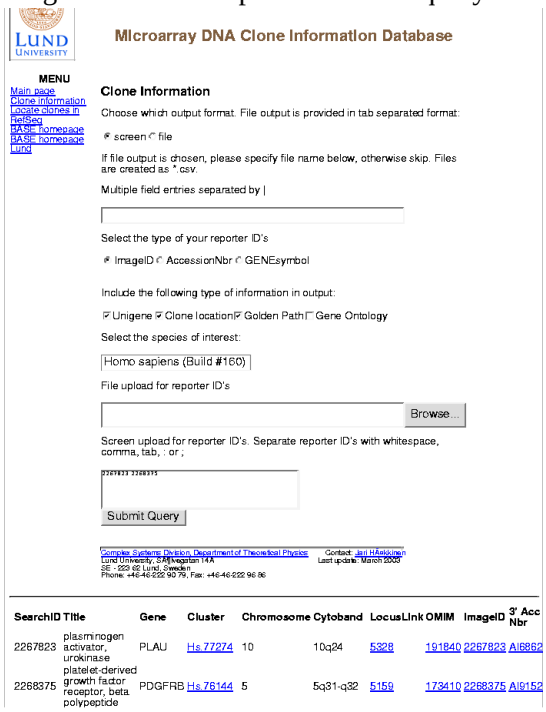



# 4 Motif search

## 4.1 Data for investigating regulatory motifs

We decided to test our application using data from a microarray experiment "Time-dependent transcriptional changes in a breast cancer cell line caused by hypoxia" [8]. Hypoxia, here by growing cells at a low oxygen rate, induces the activity of the HIF-1 protein [9], which is a transcription factor. HIF-1 is thought to play a major regulatory role in the cellular response to hypoxia [9]. The aim of that project was to investigate the time dependency of transcriptional changes due to in vitro hypoxia treatment of breast cancer cells. Microarrays containing 27,000 clones were used to analyze gene expression patterns at different time points of hypoxia treatment. The data set was filtered for quality and reduced to 15,000 genes. Of these 15,000 genes, 570 were found to be differentially expressed due to hypoxia. We wanted to investigate if the genes that were affected by the hypoxia treatment

Figure 7: An example of an ACID query.




contains a common regulatory motif. HIF-1 is known to bind to the hypoxia regulatory element (HRE) that has the consensus motif: 5'-G/C/T-ACGTGC-G/T-3' [10]. We use this motif sequence as input for our application tool to see if it is shared by the genes differentially expressed due to hypoxia. The motif sequence, which is one of the inputs to our application should be in matrix format; the values in the matrix correspond to each nucleotide at that position. We decided to assign the value 0.3 to G, C and T when they are in the first position and 0.5 to G and T when they are in the 8th position in the sequence. In this way, we turned the HRE consensus motif into the matrix in table 1.

Table 1: Sample motif matrix (G/C/T-ACGTGC-G/T)

| Pos | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| A | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 0.3 | 0 | 0 | 0 | 1 | 0 | 0 | 0.5 |
| G | 0.3 | 0 | 0 | 1 | 0 | 1 | 0 | 0.5 |
| C | 0.3 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |

Table 1 contains the weights of each nucleotide in the sequence at that po-

sition, which is used to calculate the score for a sequence. For example: if we find the sequence “GTGCTGCC” upstream of a gene, it will be given the score 0.3+0+0+0+1+1+1+0 = 3.3. Thus, each sequence found in an upstream region can be given a score for how similar it is to the motif of interest.

## 4.2 Application

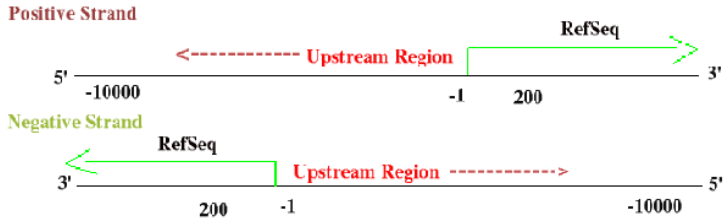
The purpose of this application is to search for shared regulatory motifs in the upstream regions of clustered genes. It is subdivided into two parts. The first involves automation of the process of extracting the sequences of the regions one wants, which could be of any length. The second phase involves searching for shared regulatory motifs in these extracted upstream regions using a given motif matrix.

### 4.2.1 First phase

The application retrieves the RefSeq information from **A**CID corresponding to a given set of microarray clones and uses the position information for each RefSeq to extract its upstream region, by scanning through the sequence of the relevant chromosome. If the RefSeq is located on a negative strand, then the application takes the downstream region of a given length in bases, starting from the end position of the RefSeq on the chromosome and reverse-complements it. This is done since the chromosomes which we have downloaded from Genome browser are the postive strand in the direction 5’ to 3’. If the RefSeq is located on a positive strand then the application takes the upstream region of a given length ending at the start position of the RefSeq on the chromosome.

The input used for retrieving the upstream regions are the start base and end base of the upstream region relative to the position of the RefSeq in the genome (eg., startbase -1 and endbase -10,000). If the startbase is a positive integer, for example, if the startbase given is 200 and the endbase given is -10,000, the application extracts 10,200 bases of upstream region, including the first 200 bases of the RefSeq (gene). This application is flexible to retrieve bases also from inside the RefSeq, since the RefSeq position on the chromosome may not exactly correspond to the transcription start site. So we decided to write an application which can even retrieve bases inside the gene, and this may be useful for other investigations.

Figure 8: Upstream region for a gene

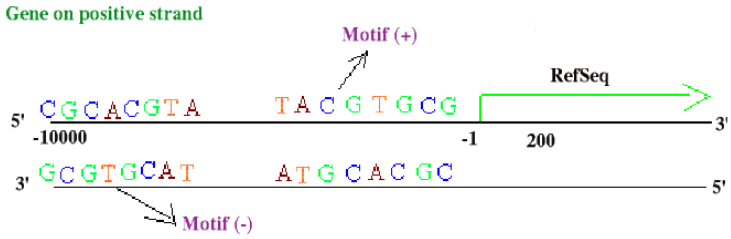



The green arrows in figure 8 shows the transcription of two RefSeqs (genes), one on each strand. On the positive strand, the RefSeq reads from 5' to 3', and the application takes the upstream region starting from the green line to a given length upstream in the direction to 5'. On the negative strand, again the RefSeq is also read from 5' to 3' and the application takes the upstream region from start (corresponds to the end base on the genome sequence) of the RefSeq to a given length upstream in the direction towards 5'. The range of the upstream region is denoted with starting base as -1 and ending ending base with the given length (eg., -10,000 bases).

### 4.2.2 Second phase

For each retrieved upstream region, the application scans through the sequence and searches for a given motif and its positions. We decided to find motifs and their positions on both strands (positive and negative), since proteins typically bind to double-stranded DNA. This application scans through the upstream sequence in windows with the same length as the motif we are looking for. Each base in a window is given a value according to its position in the motif matrix, the window is then given a score by summing up the values for each base.

Figure 9: Motif search for a gene on the positive strand




If the RefSeq is located on the positive strand the application searches for a

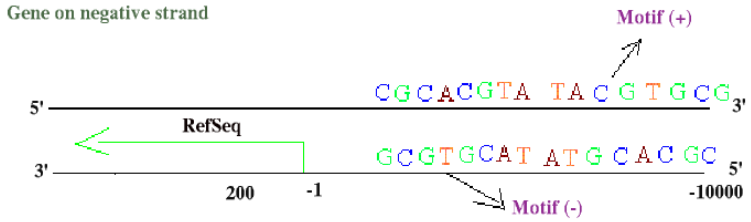
motif in the upstream region using a given motif matrix (example motif matrix shown in table 1). Figure 9 shows a “TACGTGCG” motif found on the positive strand and the score is 6.8 using the motif matrix in table 1.

Scoring:
T+A+C+G+T+G+C+G <=> 0.3+1+1+1+1+1+1+0.5 = 6.8

This application decides the presence of a motif depending on a given cutoff value for the score. If the cutoff value is greater than equal to 6 for the above motif matrix, then the motif sequence present on the upstream region is “T/A/G/C-ACGTGC-T/A/G/C”.

Since proteins typically bind to the double-stranded DNA, we also look for the motif on the negative strand even if the RefSeq is on the positive strand. Searching on the negative strand for the motif 5’-TACGTGCG-3’ corresponds to searching on the postive strand for its reverse complement 5’-CGCACGTA-3’ as can be seen in figure 9. Therefore the application looks for both the motif and its reverse-complement on the positive strand.

Figure 10: Motif search for a gene on the negative strand




If the RefSeq is located on a negative strand, the application follows the same procedure as it performed when the RefSeq was located on the positive strand. In this case the application complements and reverses the retrieved upstream region, since the chromosome sequence from the genome assembly is the positive strand in the direction 5’ to 3’, as shown in figure 10.

# 5 Basal promoter search

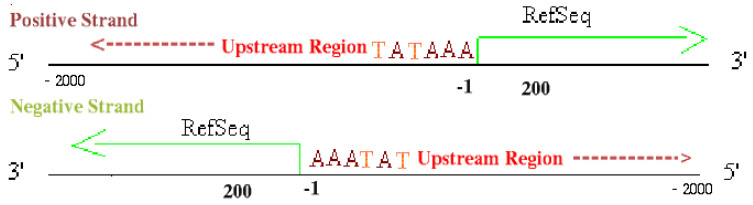
## 5.1 Data for basal promoter regions

We decided to investigate TATA boxes (basal promoter) [11] using the upstream regions for all RefSeqs present in **A**CID.

## 5.2 Application

The purpose of this application is to search for TATA boxes, that is searching for the sequence “TATAAA” [11] in the upstream regions for the RefSeqs retrieved from **A**CID. It is also subdivided into two parts. The first involves automation of the process of extracting out the region one wants (could be of any length). The second phase involves searching for TATA boxes in these upstream regions. This application function is similar to the previous application, but instead of searching for a given regulatory motif, it searches for “TATAAA” only on the same strand as the RefSeq (see figure 11)

Figure 11: TATA box search for genes




# 6 Results

## 6.1 Motif search

We executed the motif search application for both the hypoxia regulated set of clones and all the 15,000 genes that survived filtering. We used the motif matrix in table 1 and an upstream length of 10,000 (start base = -1 and end base = -10,000). The application stores results in a flat file in the following fashion:

Sample output: (Motif Search)

```
>136303|NM_021168|chr16|+|661572|700548|RAB40C|RAB40C, member RAS oncogene family

Motif   Strand  Seq     StartPos        EndPos  Score

GACGTGCT        -       AGCACGTC        -1174   -1181   6.8

TACGTGCT        -       AGCACGTA        -1448   -1455   6.8

>2547341|NM_006516|chr1|-|42395077|42428060|SLC2A1|solute carrier family 2, member 1

Motif   Strand  Seq     StartPos        EndPos  Score

NF      NF      NF      NF      NF      NF

>241705|NM_002863|chr14|-|45168037|45207149|PYGL|phosphorylase, glycogen; liver

Motif   Strand  Seq     StartPos        EndPos  Score

TACGTGCG        -       TACGTGCG        -8319   -8312   6.8

GACGTGCT        -       GACGTGCT        -364    -357    6.8

GACGTGCG        +       CGCACGTC        -4275   -4282   6.8

>824933|NM_013410|chr1|+|64533461|64612267|AK3|adenylate kinase 3

Motif   Strand  Seq     StartPos        EndPos  Score

GACGTGCG        +       GACGTGCG        -2400   -2393   6.8

CACGTGCG        +       CACGTGCG        -354    -347    6.8

CACGTGCT        -       AGCACGTG        -760    -767    6.8
```

The sample output contain clone information that begins with '>' and is separated by pipe symbols ("|"), it includes

- clone identifier
- RefSeq
- chromosome for RefSeq
- strand for RefSeq
- start position of RefSeq in Genome Browser
- end position of RefSeq in Genome Browser
- gene symbol
- gene name

It also contains motif search information that includes

- which motif the application found (Note: Motif direction is always from 5' to 3')

- on which strand
- the actual sequence in the upstream region
- start position of the motif in the upstream region
- end position of the motif in the upstream region
- score

Using **A**CID, we could associate 324 of the 570 clones found to be differentially expressed by hypoxia to a RefSeq with position in the human genome. The corresponding number for the other clones (not regulated by hypoxia) was 7438. We found that 85 clones of the hypoxia regulated group and 545 clones of the general group contain the given motif (see table 2).

Table 2: Results table for motif search in the hypoxia data set

| Clones | contain motif | donot contain motif | Total |
|---|---|---|---|
| Expression Regulated by Hypoxia | 85 | 239 | 324 |
| Expression not Regulated by Hypoxia | 545 | 6893 | 7438 |
| | | | 7762 |

The odds-ratio for the values in table 2 is (85/239)/(545/6893) = 4.5. This odds-ratio of 4.5 indicates that there is an association between the motif and the clones with expression affected by hypoxia. If there would be no association, the odds-ratio is expected to be equal to 1. To investigate the significance of the odds-ratio deviating from 1, we submitted the values of table 2 to Fisher's exact test [12] and got the p-value<=$2.2*10^{-16}$. This p-value roughly corresponds to the probability to get such an over-abundance of motifs for the hypoxia affected genes by random chance, given the overall number of genes in the data set with the motif.

## 6.2 Basal promoter search

A basal promoter search (TATA box search), searching for TATA boxes (TATAAA) [11] is important to confirm the results of our research question (motif search), since the start positions of RefSeqs should reflect the transcription start sites of the genes, for our common regulatory motif search to make sense. We decided to execute our application "Basal promoter search" for all RefSeqs present in our

database (**A**CID). The database **A**CID contains information for 17,295 RefSeqs. We executed the application for all 17,295 RefSeqs with upstream length 2000 bases and the sequence "TATAAA" as inputs, and came out with results as follows:

Table 3: TATA box search results

| contain sequence "TATAAA" | No. RefSeqs |
|---|---|
| Yes | 8,429 |
| No | 8,866 |
| Total | 17,295 |

Table 3 shows that 8,429 RefSeqs have TATA boxes (TATAAA) and 8,866 does not have TATA boxes (TATAAA) in their upstream regions of length 2000 bases.

Table 4: Number of TATA box hits for the RefSeqs

| Number of hits | Number of RefSeqs |
|---|---|
| 0 | 8,866 |
| 1 | 5020 |
| 2 | 2065 |
| 3 | 813 |
| 4 | 328 |
| 5 | 126 |
| >5 | 77 |
| Total | 17,295 |

Table 4 shows the number of TATA box hits for the RefSeqs. Table 5 shows for

Table 5: TATA box positions

| Upstream base-pairs range | No. of RefSeqs |
|---|---|
| 0 - 100 | 714 |
| 100 - 200 | 369 |
| 200 - 500 | 1507 |
| 500 - 1000 | 3269 |
| 1000 - 2000 | 7874 |

the upstream intervals 0 - 100, 100 - 200, 200 - 500, 500 - 1000 and 1000 - 2000 bases, the number of RefSeqs that have TATA boxes in the intervals. The output of the application is stored in a flat file in the following fashion:

Sample output: (TATA box search)

| | | | | | | |
|---|---|---|---|---|---|---|
| NM_020472 | 51 | chrX | - | 2 | PIGA | -1085,-340 |
| NM_002764 | 56 | chrX | + | 3 | PRPS1 | -888,-719,-459 |
| NM_002835 | 62 | chr7 | + | 2 | PTPN12 | -1720,-1202 |
| NM_003856 | 66 | chr2 | + | 2 | IL1RL1 | -1728,142 |
| NM_016232 | 66 | chr2 | + | 1 | IL1RL1 | 33 |
| NM_007327 | 105 | chr9 | + | 0 | GRIN1 | |
| NM_031846 | 167 | chr2 | + | 2 | MAP2 | -1292,-611 |
| NM_002600 | 188 | chr1 | + | 6 | PDE4B | -1896,-806,-798,-686,-592,-509 |

# 7 Discussion and outlook

We developed **A**CID, a database which contains information about microarray clones. To retrieve information about microarray clones from our database (**A**CID), we built a web page, which can be very useful for the microarray research community. This web page will become publicly available.

To illustrate the usefulness of **A**CID we built an application, which can find common regulatory motifs in the upstream sequences for genes by taking clone identifiers as input. We executed our application to find common regulatory motifs for genes that were similarly expressed by hypoxia treatment in a microarray experiment. We also built an application which can search for TATA boxes in the upstream regions of all RefSeqs (genes) and from the results of this application (TATA box search or basal promoter search), to address whether RefSeq start positions (which are used for retrieving upstream sequences) reflect the transcription start sites.

The results of the basal promoter search (TATA box search) shows that almost half of the upstream sequences of the RefSeqs contain TATA boxes (see table 3). Moreover, table 5 shows that these TATA boxes are found in many positions in the upstream sequences. Taken together, this tells us that RefSeq start positions may not perfectly reflect the transcription start sites. In the Genome Browser, there is an on-going effort to add a table that contains the transcription start site for each transcript, and in the future this information will also be available in **A**CID. It will be straight-forward to modify our application (motif search) to use this table instead of using RefSeq start positions for retrieving upstream sequences. It will be interesting to see how our results for the TATA boxes get modified. Nevertheless, looking at the specific application of investigating a common regulatory

motif for genes which are co-expressed under hypoxia, we find a significant association between the known motif (HRE) and the co-expressed genes. In the future, **A**CID is likely going to be extended with many more useful applications, which will benefit microarray experiment investigations. However, we note that current applications (motif search & basal promoter search) are not only applicable for the human genome but also the mouse genome, and that the motif search application is very useful already at this point.

## 8 Acknowledgments

Finally, credit where credit is due. I would like to thank my supervisor *Markus Ringnér* for being a wonderful person and for helping me in all stages of this project. I must also thank *Jari Häkkinen*, my second supervisor for providing necessary requirements to accomplish this project. Last not least I'd like to thank *Samuel Andersson* my co-partner in developing ACID and *Naomi Glarner* for providing microarray data to investigate common regulatory motifs.

## References

[1] *RefSeq and LocusLink: NCBI gene-centered resources*. Pruitt KD and Maglott DR (2001). Nucleic Acids Research 29:137-40.

[2] *The Human Genome Browser at UCSC.* Kent WJ, Charles W Sugnet, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D (2002). Genome Research 12:996-1006.

[3] *The UCSC Genome Browser Database.* Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, and Kent WJ (2003). Nucleic Acids Research 31:51-54.

[5] *ACID: a database for microarray clone information.* Ringnér M, Andersson S, Veerla S, Staaf J, and Häkkinen J (2003). LU TP 03-24, Lund University.

[6] *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, and Futcher B (1998). Molecular Biology of the Cell 9:3273-97.

[7] *Identifying regulatory networks by combinatorial analysis of promoter elements*. Pilpel Y, Sudarsanam P, and Church GM (2001). Nature Genetics 29:153-9.

[8] *Time-dependent transcriptional changes in a breast cancer cell line caused by hypoxia*. Glarner N (2003). Masters Thesis, LU TP 03-06, Lund University.

[9] *Signal transduction to hypoxia-inducible factor 1*. Semanza G (2002). Biochemical Pharmacology 64:993-998.

[10] *A nuclear factor induced by hypoxia via de novo protein synthesis binds to the human erythropoietin gene enhancer at a site required for transcriptional activation.* Semanza GL,Wang GL (1992). Molecular Cell Biology 12:5447-54.

[11] *Combinatorial regulation of transcription I: General aspects of transcription control*. Ernst P and Smale ST (1995). Immunity 2:311-319.

[12] *The use of multiple measurements in taxonomic problems.* Fisher RA (1936). Annals of Eugenics 7:179-188.

[13] *Creating the gene ontology resource: design and implementation.* The Gene Ontology Consortium (2001). Genome Research 11:1425-1433.

[14] *Online Mendelian Inheritance in Man OMIM: www.ncbi.nlm.nih.gov/entrez.* Parton MJ (2003). Journal of Neurology Neurosurgery and Psychiatry 74:703.

[15] *GenBank*. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003). Nucleic Acids Research 31:23-7.