

Please consult the instruction sheets accompanying this form before filling it out

1. principal investigator institute address telephone/fax email	Dr A.H.C. van Kampen Bioinformatica Laboratorium Academisch Medisch Centrum Universiteit van Amsterdam Meibergdreef 9 Postbus 22660 1100 DD Amsterdam tel.: 020-5667096 fax: 020-5664440 a.h.vankampen@amc.uva.nl	Dr R. Versteeg Afd. Anthropogenetica Academisch Medisch Centrum Universiteit van Amsterdam Meibergdreef 9 Postbus 22660 1100 DD Amsterdam tel.: 020-5665243 fax: 020-6918626 r.versteeg@amc.uva.nl	Dr H.J. Bussemaker Swammerdam Institute for Life Sciences Universiteit van Amsterdam Kruislaan 318 1098 SM Amsterdam tel.: 020-5257795 fax: 020-5257934 bussemaker@science.uva.nl	Dr H.N. Caron Afd. Kinderoncologie en Anthropogenetica EKZ, Academisch Medisch Centrum Universiteit van Amsterdam Meibergdreef 9 Postbus 22660 1100 DD Amsterdam tel.: 020-5666592 fax: 020-6918626 h.n.caron@amc.uva.nl
2. title	A sequence-based Human Transcriptome Map: Identification and computational analysis of landmarks in the human genome			
3. BMI research themes	A / C / D			
4. summary of a. overall aim and key objectives b. approach c. elements of innovation d. relevance for BMI program <i>please indicate a,b,c and d in your text</i>	<p>Overall aim and key objectives: The project aims at the establishment of an integrated bioinformatics, computational and genomic research infrastructure. We will apply bioinformatics tools to integrate the human genome sequence with high throughput mRNA analysis data to generate a complete expression profile of the human genome. This will be used to identify long-range transcriptional domains. We will analyze the DNA sequences that define the transcriptional domains and control long-range transcriptional activity. A similar transcriptome map for the mouse will be constructed to complement these studies. Establishment of new approaches for the implementation of database applications and the development of computational analysis methods for large DNA data sets are integral part of these aims.</p> <p>At the Academic Medical Center (Dept. of Human Genetics and the Bioinformatics Laboratory), we have recently established the Human Transcriptome Map (HTM), which integrates a radiation hybrid map of the human genome with the expression data of 4.4 million mRNAs identified by the SAGE technology (Caron et al., <i>Science</i> 291, 2001, 1289-92). The HTM (http://bioinfo.amc.uva.nl/HTM) generates gene expression profiles for any chromosomal region in twelve normal and pathologic tissue types. The map revealed an unexpected clustering of highly expressed genes to specific chromosomal regions (Figure 1). It provides a new tool to search for genes over-expressed or silenced in cancer. At the Swammerdam Institute for Life Sciences we developed algorithms to identify regulatory DNA sequences in complete genomes with the use of genome wide expression profiles (Bussemaker et al, <i>Nature Genet.</i> 27, 2001, 167-71). In this BMI program, we will integrate the research lines of the three participating groups to reach the following key objectives:</p> <p>I Integration of the Human Transcriptome Map with the full genomic sequence. The present Human Transcriptome Map is based on the human radiation hybrid map (GeneMap'99). We will construct an improved and more comprehensive version of this map for which we will use the draft genomic sequence as backbone. We will integrate this map with other biological databases, e.g., genomic annotations and micro-array data. To overcome technical limitations to the swift integration of new data sources, we will use a new approach for implementing database applications. As this map is an important tool for identification of cancer-related genes in a wide range of human tumor types, it will be made available on internet.</p> <p>II) Identification of transcriptional domains in the Human Transcriptome Map. The HTM identifies several types of transcriptional domains. Most prominent are Ridges, which are clusters of highly expressed genes. Several subtypes of Ridges can be recognized, e.g. gene-dense, gene-poor and telomeric Ridges. The HTM furthermore identifies different kinds of weakly expressed regions. The new sequence-based HTM will likely reveal a more detailed expression landscape. Recognition and delineation of landmarks requires sophisticated clustering analysis and statistical analyses, which have to</p>			

be developed. The transcriptional domains will be analyzed for relationships with other genomic features included in the database, e.g. CG content, repeats, functional gene categories, known regulatory sequences and protein binding domains.

III) Identification of regulatory DNA sequences defining expression domains.

We will apply computational analysis to identify DNA elements characteristic for the transcriptional domains. We have previously developed algorithms to identify cis-regulatory elements in non-coding DNA by correlating genome-wide expression data with genome sequence information for *Saccharomyces cerevisiae* (Bussemaker et al., 2001). These methods will be applied and further developed to identify control elements in the different types of human transcriptional domains obtained with the HTM.

IV) Generation of a Mouse Transcriptome Map.

We will apply the algorithms used for the Human Transcriptome Map to establish a mouse transcriptome map. Analysis of mouse transcriptional domains will facilitate identification of the human regulatory DNA sequences, as regulatory elements are probably conserved in evolution. Furthermore, this map will open the road to future experimental analysis of such regulatory sequences by generation of transgenic ES cells and/or mice.

b. Approach.

To approach the key objectives we will integrate the disciplines of bioinformatics, genomics and computational sequence analysis, while using technologies and methods that are standards in the field of bioinformatics. We have experience with the use and maintenance of local copies of public biological databases. Our insight in genomic databases (structure, weak and strong points) enables an optimal development of bioinformatics tools. We also have experience with multivariate data analysis, development of computational algorithms and global optimization techniques. Data analysis software will be developed in Perl, C and the statistical package Splus. This software will be interfaced to the database directly via ODBC and through Perl/DBI scripts. We will use a relational DBMS to implement the central data repository. A web-interface for the database will make it accessible over the internet.

c. Elements of innovation.

This work integrates two major technological spearheads of molecular genetics: the human genome sequence and high throughput mRNA expression analysis. The development of innovative bioinformatical approaches forms the basis for this integration. A new type of database application architecture is necessary to enable rapid updating of the relational database with continuously renewing biological databases, as well as to implement a flexible query interface to the database. Identification and computational analysis of the transcriptional domains requires upgrading of recent algorithms developed for yeast genomics. The project therefore integrates advanced molecular genetics, bioinformatics and computational analysis.

d. Relevance for Biomolecular Informatics.

Our collaboration in this project further adds to the bioinformatics research infrastructure at the Academic Medical Center and the Swammerdam Institute of Life Sciences at the University of Amsterdam. Moreover, since the HTM is the first genome wide transcriptome map, the development and accessibility on internet of HTM and sub-databases like human and mouse tag-to-gene databases, will contribute to cancer research and SAGE-analyses of biomedical and biological research groups. Bioinformatics research will advance from innovative database application architecture and computational algorithms. Expertise, methodologies, software and databases will be disseminated and will find applications in related research areas. Our collaborations at the national and international level will add to the impact of our results.

5. key words	<ul style="list-style-type: none"> - Human genome project - Serial analysis of gene expression - Genome-wide transcription profiling - Genome architecture - Data integration 	<ul style="list-style-type: none"> - Human Transcriptome Map - DNA microarrays - Multivariate data analysis - Regulatory elements - Comparative genomics 														
6. participants	<table border="1"> <thead> <tr> <th>Name</th> </tr> </thead> <tbody> <tr> <td>Dr. A.H.C. van Kampen</td> </tr> <tr> <td>Ing B.D.C. van Schaik</td> </tr> </tbody> </table>	Name	Dr. A.H.C. van Kampen	Ing B.D.C. van Schaik	<table border="1"> <thead> <tr> <th>discipline</th> </tr> </thead> <tbody> <tr> <td>Bioinformaticist</td> </tr> <tr> <td>Bioinformaticist</td> </tr> </tbody> </table>	discipline	Bioinformaticist	Bioinformaticist	<table border="1"> <thead> <tr> <th>Paid by</th> </tr> </thead> <tbody> <tr> <td>AMC</td> </tr> <tr> <td>AMC</td> </tr> </tbody> </table>	Paid by	AMC	AMC	<table border="1"> <thead> <tr> <th>hrs/week</th> </tr> </thead> <tbody> <tr> <td>5</td> </tr> <tr> <td>20</td> </tr> </tbody> </table>	hrs/week	5	20
Name																
Dr. A.H.C. van Kampen																
Ing B.D.C. van Schaik																
discipline																
Bioinformaticist																
Bioinformaticist																
Paid by																
AMC																
AMC																
hrs/week																
5																
20																

Dr. P.J. de Groot	Chemometrician	AMC	2
A.C.M. Luyf	Bioinformaticist	AMC	1
Ing R. Waaijer	Technician	AMC	1
Dr. R. Versteeg	Molec. Biologist	AMC	5
Dr. H.N. Caron	Paediatrician, Molec. Biol.	KNAW/AMC	5
Dr. D. Geerts	Cell biologist	AMC	3
Dr. H.J. Bussemaker	Physicist	UvA	5
Dr. P. van Bommel	Informaticist	KUN	adv.
Prof. Dr. R. van Driel	Molec. Biologist	UvA	adv.
Dr. B. van Steensel	Molec. Biologist	KNAW/UvA	adv.

7. requested support	personnel			requested over the years					
	level	months	fte	2001	2002	2003	2004	2005	2006
	postdoc	48	1.0	2	12	12	12	10	
	postdoc	48	1.0	2	12	12	12	10	
	postdoc	48	1.0	2	12	12	12	10	
	technician	24	1.0	2	12	10			
	consumables (kf)			7.5	27.5	27.5	27.5	20	
	equipment (kf)			20	10	10	10		

8. **a. justification of requested personnel:**
 This project aims to integrate several of the major developments in biomedical research. They are: the establishment of the human genomic sequence, high throughput mRNA analyses, innovative bioinformatics to integrate complex heterogeneous databases, statistical and clustering analysis approaches and computational analysis of very large datasets. As all these fields go through a period of fast developments, their integration in an optimal configuration requires a prolonged period of research. Based on our recent experience with comparable projects with a more limited aim, we foresee that the here proposed software and database development, the application of analytical tools and the integration with rapidly developing public genome databases will reach a stage of consolidation only in the fourth grant year. As the three requested post-docs represent different disciplines that should interact from the start of the project on, we feel it necessary and essential that they can continue their work for four years. A three year appointment of the post-docs would prematurely restrict the development of the project. The requested technician for programming is necessary for two years.

b. justification of consumables
 To store local copies of the public biological databases that we use during the project we need 75Gbyte of hard disk space to extend our Unix server. The price for this space is FL 25.750 per year (FL 210,- per year/Gbyte and FL10.000 for maintenance each year).

c. justification of equipment
 We ask for four computers (PCs) for the requested personnel (4xFL5.000,-) and support of minor equipment.

d. requested support from own institute or other sources
 The AMC will make the investments to set up a central Oracle platform, which will be used during this project. Furthermore, personnel support is given by the dept. of Human Genetics and the Bioinformatics Laboratory.

9. overall aim, key objectives
9. overall aim
 The project aims at the establishment of an integrated bioinformatics, computational and genomic research infrastructure. Successful realization of this goal requires an integral approach within a well-defined bio-medical and genomics research line. We aim to apply bioinformatics and computational data analysis to identify key elements in the long-range organization of transcription domains in the human genome. We will integrate the human genome sequence with data from high throughput mRNA analyses of human tissues. This will result in chromosomal expression profiles that will be used to define long-range

10. international position and collaboration

11.

a. Selected publications
by the applicant(s)

b. International literature
(key references marked with *)

transcriptional domains in the genome. Integral part of this effort is the development of an advanced database application design that facilitates integration of databases related to gene function and chromosome topography and that can easily be queried. The different types of expression domains to be identified with this database will be analyzed by computational sequence analyses for hallmarks representing regulatory sequences. Establishment of a mouse transcriptome map should complement these efforts.

1. Establishment of a second-generation Human Transcriptome Map database.

The present version of the Human Transcriptome Map (HTM) is based on a radiation hybrid map with the rough position of 24.000 human genes (Caron et al., 2001). It gives gene expression profiles for any chromosome or chromosomal region in a large set of normal and malignant human tissues. We will develop a next generation HTM that is based on the full genomic sequence, as provided by the databases of the (public) Human Genome Project. For the high throughput mRNA analyses we use the SAGE (Serial Analysis of Gene Expression) technology (Velculescu et al., 1995). SAGE is based on the extraction of a 10 base pair 'tag' from each transcript in a tissue and the sequencing of thousand of these tags. SAGE gives the absolute expression level of each gene per e.g. 100.000 transcripts in a tissue. The quantitative aspect of SAGE implies that SAGE libraries obtained all over the world can directly be compared. About one hundred public SAGE libraries are presently available (Lal et al., 1999). They total to about 4.4 million SAGE tags, each representing a transcript expressed in a human tissue. We will integrate all public SAGE libraries with the human genome sequence. This will result in an application that gives gene expression profiles for any chromosome or chromosomal region in a large series of normal and malignant tissues. SAGE libraries are being made from an increasing number of tissues at varying differentiation stages. The application will therefore eventually show the gene expression profiles of most human tissues and their differentiation stages. The availability of this application on internet will facilitate identification of cancer-related genes in the wide variety of human cancers for which SAGE libraries are included in the database.

9./10./11. continued

2. Data integration and the development of a new architecture for database applications.

During the development of the HTM application, we experienced that implementation, modification and extension of the underlying relational database and the corresponding user interfaces is time consuming. This limits rapid scientific advances. Therefore, we will develop a new architecture for the HTM, which should result in an application of which the database is dynamic and can be modified without having to re-program the user interfaces. This database system integrates gene expression data with data obtained from public biological databases (e.g., genomic sequence), and provides the basis for the analysis of long-range transcriptional domains. To obtain a high quality HTM database we will develop algorithms that mark errors or inconsistencies in the public databases before integration.

3. Characterization of transcriptional domain types in the human genome

The current version of the Human Transcriptome Map reveals a higher order structure of transcriptional domains in the genome. The most prominent domains consist of clusters of several tens to hundreds of genes with a much higher gene expression than the genomic average. We identified about 30 such clusters, called Ridges. There appear to exist subtypes, as some Ridges are extremely gene dense, while others have a normal gene density. Furthermore, the HTM gives a first view of several less defined transcriptional domains, e.g. stretches of hardly expressed genes. The current HTM is based on the radiation hybrid map of the genome, which has a limited resolution and counts many errors. The sequence-based HTM will be more reliable, include more genes and will provide more precise information on the genomic position of the coding regions. Furthermore, the database will present additional information, e.g. regarding repeat sequences and other genomic features. The genomic landscape will therefore show up in more detail, with probably additional types of transcriptional domains. Two approaches will be pursued to identify and define these transcriptional landmarks. First, we will implement user-interfaces that allow the visual exploration of large quantities of heterogeneous data on a genome-wide scale. Secondly, we will use multivariate data analysis methods (e.g., hierarchical clustering, principal component analysis, multi-dimensional scaling) to identify and characterize long-range transcriptional domains. These methods were originally not designed to cope with the quantity and heterogeneity of data presently encountered in genomics and will therefore be adapted and improved.

4. Identification of regulatory DNA sequences defining expression domains

To understand genome-wide transcription regulation, it is essential to understand the interplay between elements controlling long range transcriptional domains and the more commonly studied transcription factors controlling individual genes. We will develop computational algorithms that identify regulatory elements in DNA sequences and a mathematical model that quantitatively describes the variability in transcriptional activity across the human genome. The quantitative model will provide information about the activity of the regulatory proteins that interact with the regulatory elements. To identify and model the different elements that regulate individual genes and long-range transcriptional domains we will proceed in two steps.

First, we will search for sequence elements that characterize Ridges. There are about 30 of these domains of high gene expression. Availability of the genomic sequence of these regions will permit a search for motifs consistently more prevalent in the promoters of the genes within these domains, as well as for elements dispersed throughout the Ridges. In addition, we will search for elements characterizing boundaries of Ridges. Similar searches will be performed for other types of transcriptional domains to be identified in this project.

Secondly, we will develop models to find and quantify transcription factor binding sites in the promoters and enhancers of all human genes, and investigate to what extent the resulting model explains the variability in transcriptional activity across the genome, including the long-range transcriptional domains. For this we will extend our modeling approach using REDUCE, which we recently applied to yeast (Bussemaker et al, 2001).

Thirdly, we will relate the HTM to boundary elements that isolate euchromatin from heterochromatin. We will initially focus on boundary elements of the polycomb group, which have a predictable influence on chromatin domain structure (Van der Vlag et al, 2000). The identification of these DNA elements and their position in the genome has only just started and it is expected that various classes of such elements exist. We will investigate whether they relate to the long-range transcriptional domains. Once all these elements have been identified, a quantitative model will be constructed that can be fit to the genome-wide expression data, and whose parameters provide information about the activity of the regulatory proteins that correspond to the DNA elements. We will analyze how much of the variability of gene expression across the genome is predicted by the model.

5. Construction of a mouse transcriptome map

The algorithms used for the Human Transcriptome Map will be used to establish a mouse transcriptome map. The first question to be asked is whether the murine genome harbors similar transcriptional domains as the human genome. If this is indeed the case, comparison of syntenic mouse and human transcriptional domains will facilitate identification of regulatory DNA sequences, as they are likely to be conserved. Already identified human regulatory sequences can be validated by a search for similar sequences in the corresponding mouse domains. An even more interesting question that can be answered by the murine transcriptome map is whether individual long-range transcriptional domains (e.g. ridges) are integrally conserved during evolution. It is well established that the mouse and human genomes are chopped up and remixed versions of an ancestral genome. The question is whether fragmentation and rearrangement of this ancestral genome has been random or instead followed the boundaries set by the transcriptional domains. Comparison of the mouse and human transcriptome in conjunction with a detailed map of mouse-human synteny can give the answer. The mouse transcriptome map will finally provide a basis for future experimental testing of long-range regulatory sequences in transgenic mouse or ES cells.

6. Use of micro-array data and integration with SAGE

Integration of micro-array data in the HTM is attractive, but fundamental differences between SAGE and micro-array technologies exist. Construction of the Human Transcriptome Map is based on two features unique to the SAGE technology. Firstly, SAGE is quantitative, i.e. it establishes the transcript number of each gene per e.g. hundred thousand mRNAs in a tissue. Secondly, SAGE data are universal, implying that all worldwide established SAGE libraries can directly be compared. These features are not met by most currently used micro-array technologies. Micro-array analyses usually compare two mRNA sources, and identify the differences. Such analyses would not identify Ridges, as these regions are highly expressed in all tissues. Furthermore, these data are not universal,

but depend on the reference mRNA used. The micro-array systems that measure hybridization intensities of one mRNA source (e.g. Affymetrix chips) are possibly less linear than SAGE, due to differences in hybridization dynamics for different probes. However, micro-arrays represent a quick and easy method, making integration in HTM desirable. We will therefore organize the relational database in a way that the publicly available micro-array datasets can be related to the HTM data. This should reveal whether array data can provide information on long range expression profiles. It could be envisioned that specific tumor samples have disturbed expression of entire transcriptional domains due to translocations. In addition, in a collaborative research line with the Max Planck Institute for Molecular Genetics in Berlin (Prof. Dr. H.-H. Ropers), we will perform micro-array analyses of cell lines also analyzed by SAGE. These data will assess the use of array data for the HTM.

10. International position and collaboration

Genome-wide expression profiles have only been generated for yeast (Velculescu et al., 1997) and for man (Caron et al., 2001). Yeast showed an evenly distribution of highly and weakly expressed genes over the genome. Therefore, virtually nothing is known about the long range transcriptional domains identified in the HTM. It is possible that they relate to known nuclear substructures, which will be analyzed in a parallel research line (collaboration Prof. R. van Driel). Development of analysis software for SAGE libraries is rapidly proceeding. The NCBI SAGEmap databases that relate SAGE tags to the corresponding mRNAs are highly useful (Lal et al., 1999). Where this database aims at completeness, our AMCtagmap database (Caron et al., 2001) aims at restricting to reliable tags. Both approaches are therefore complementary. Furthermore, several methods for statistical analysis of SAGE data have been established (e.g., Zhang et al., 1997; Kal et al., 1999).

Many bioinformatics efforts relate to databases and data integration. These approaches include multi-database query systems (Stevens, 2000), data warehouses that do or do not homogenize data (Etzold et al., 1996; Leser et al., 1998), and standards for data representation and exchange (Life Sciences Research Task Force, 1997). Advances in the XML technology enable the potential useful for data management (Achard et al., 2001). Peculiarities of bioinformatics databases were already outlined (e.g., Davidson et al., 1995 and Karp 1995). Many approaches were followed for the creation of (graphical) interfaces for biological databases (e.g., Stein et al., 2001; Searls, 1995; Etzold et al., 1996; Fischer et al., 1999). The approach that we propose in this project emphasizes independency between user interface and database.

The field of multivariate data analysis (e.g. Kruskal, 1964) and global optimization (e.g., Kirkpatrick et al., 1983; Lucasius and Kateman, 1995) is well established. Potential pitfalls have been documented (e.g., Byron, 1995). They are widely applied, but the quantity and heterogeneity of biological data are still challenging (e.g., Sherlock, 2000).

The most common way for identifying regulatory elements in DNA sequences is to group genes into disjoint clusters (Eisen et al., 1998), and subsequently analyze the upstream regions within each cluster (Lawrence et al., 1993; Neuwald et al., 1995; Helden et al., 1998). The approach of this project is based on the identification of these elements from gene expression ratios for a large number of genes and the upstream regulatory DNA sequence for these genes.

Collaborations for this project:

We collaborate with Prof. E.D. Siggia at The Rockefeller University for sequence and gene expression analysis algorithms, Dr. G. Riggins (Dept. of Pathology, Duke Medical Center and NCBI-SAGE library repository) and Dr. A. Lash (NCBI, NIH, Bethesda) for the NCBI tag-extraction approach and the use and updating of SAGE libraries and with Prof. H.-H. Ropers (Max Planck Institute for Molecular Genetics, Berlin) for micro-array analysis. We collaborate with Prof. R. van Driel and Dr. B. van Steensel (SILS, University of Amsterdam) on the relationship of transcriptional domains and nuclear architecture. Databases and daily updates are i.a. obtained from the CMBI (KUN, Nijmegen). We collaborate with Dr. P. van Bommel (Computing Science Institute, Information Retrieval and Information Systems, KUN, Nijmegen) for the development of database applications. At the University of Amsterdam, we collaborate with Prof. L.O. Hertzberger (Informatics Institute) in the Virtual Laboratory project, the Dept. of Clinical Epidemiology and Biostatistics (AMC) for advanced statistical analyses, the Dept. of Information and Communication Technology (AMC) for software engineering and the Dept. of Clinical Informatics (AMC) for development of computational methods.

11a Selected publications by the applicants

Applicant,	First author,	Last author,	co-author,	presentations,	abstracts
van Kampen:	7,	0,	3,	10,	10
Versteeg:	7,	17,	17,	35,	20
Bussemaker:	15,	0,	20,	32,	9
Caron:	13,	0,	13,	19,	11

9./10./11.
 Continued

- *1. Boon, K, Caron HN, van Asperen R, Hermus M-C, van Sluis P, Roobeek I, Weis I, Voute PA, Schwab M, Versteeg R, N-myc enhances the expression of a large set of genes functioning in protein synthesis. **EMBO J** 20 (2001):1383-1393. [14]
- *2. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. **Nature Genet** 27 (2001):167-71. [30.7]
- *3. Caron, H., van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus M.-C, van Asperen R, Boon K, Voûte PA, Heisterkamp S, van Kampen A, Versteeg R, The Human Transcriptome Map reveals a clustering of highly expressed genes in chromosomal domains. **Science** 291 (2001):1289-1292. [24.6]
4. Pauws E, van Kampen AHC, van de Graaf SAR, de Vijlder JJM, Ris-Stalpers C., Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. **Nucl Acids Res**, 2001, in press [5.7].
5. Spieker, N, Beitsma M, van Sluis P, Chan A, Caron H, Versteeg R, Three chromosomal rearrangements in neuroblastoma cluster within a 300 kb region on 1p36.1. **Genes, Chromosomes and Cancer** (2001), in press [4.9]
- *6. Spieker, N, van Sluis P, Beitsma M, Boon K, van Schaik BDC, van Kampen AHC, Caron H, Versteeg R, The MEIS1 oncogene is highly expressed in neuroblastoma and amplified in cell line IMR32. **Genomics** 71 (2001):214-221 [3.4]
- *7. Bussemaker HJ, Li H, Siggia ED, From the cover: building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. **Proc Natl Acad Sci USA** 97 (2000):10096-100. [10.3]
- *8. Spieker, N., Beitsma M., van Sluis P, Roobeek I, den Dunnen JT, Speleman F, Caron H, Versteeg R, An integrated 5 Mb physical, genetic and radiation hybrid map of an 1p36.1 region implicated in neuroblastoma pathogenesis. **Genes, Chromosomes and Cancer** 27 (2000):143-152. [4.9]
9. Van Kampen A, van Schaik B, Pauws E, Michiels E, Ruijter J, Caron H, Versteeg R, Heisterkamp S, Leunissen JA, Baas F, van der Mee, M, USAGE: a web-based approach towards the analysis of SAGE data. **Bioinformatics** 16 (2000):899-905. [2.3].
10. van Limpt V, Chan A, Caron H, Sluis PV, Boon K, Hermus MC, Versteeg R. SAGE analysis of neuroblastoma reveals a high expression of the human homologue of the drosophila delta gene. **Med Pediatr Oncol.** 35 (2000):554-8. [1.5]
11. van der Drift P, Chan A, Westerveld A, and Versteeg R, Multiple MSP pseudogenes in a local repeat cluster on 1p36.2: an expanding genomic graveyard? **Genomics** 62 (1999), 74-81. [3.4]
12. Van Kampen AHC, Buydens LMC, The ineffectiveness of recombination in a genetic algorithm for the structure elucidation of a heptapeptide in torsion angle space. A comparison to simulated annealing. **Chemom Intell Lab Syst** 36 (1997):141-152. [1.7]
13. Caron, H., P. van Sluis, J. de Kraker, J. Bökkerink, M. Egeler, G. Laureys, R. Slater, A. Westerveld, P.A. Voûte, R. Versteeg. Allelic loss of chromosome 1p identifies neuroblastoma patients with a high risk of an unfavourable outcome. **New England J. Medicine** 334 (1996):225-230. [28.8]
14. Cheng NC, Chan A, Beitsma M, Op den Camp I, Speleman F, Westerveld A, Versteeg R, A human modifier of methylation for class I HLA genes (MEMO-1) maps to chromosomal bands 1p35-36.1. **Human Molec. Genetics** 5 (1996) 309-317. [9.4]
15. Caron H, van Sluis P, van Hoeve M, de Kraker J, Bras H, Slater R, Mannens M, Voûte PA, Westerveld A, Versteeg R, Allelic loss of chromosome 1p36 in Neuroblastoma is of preferential maternal origin and correlates with N-myc amplification. **Nature Genetics** 4 (1993), 187-190. [30.7]

Book chapters

- Versteeg, R. Genetics of solid childhood tumours; in: Genetics for the clinician (R. Hennekam ed.) Baillière's Clinical Paediatrics series, Baillière Tindall London, 1998, pp.277-292.
- White, P. and R. Versteeg. Allelic loss and neuroblastoma suppressor genes. in: Neuroblastoma (G.M. Brodeur, T. Sawada, Y. Tsuchida, P.A. Voûte, eds.) Elsevier Science Amsterdam, 2000, pp. 57-68.
- Caron, H and Hogarthy, M. Imprinting of MYCN, 1p and other loci. in: Neuroblastoma (G. Brodeur, T. Sawada, Y. Tsuchida, P.A. Voûte, eds.) Elsevier Science Amsterdam, 2000, pp. 101-11.

11b International literature

- Achard F, Vaysseix G, Barillot E, XML, bioinformatics and data integration. **Bioinformatics** 17 (2001):115-125.
- *Byron JTM, Non-uniqueness and inversions in cluster analysis. **Appl Statist** 44 (1995) 117-134.
- Cremer T, Kreth G, Koester H, Fink RH, Heintzmann R, Cremer M, Solovei I, Zink D, Cremer C, Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture. **Crit Rev Eukaryot Gene Expr** 10 (2000):179-212.
- *Davidson SB, Overton C, Buneman P, Challenges in integrating biological data sources. **J Comput Biol** 2 (1995):557-572.
- Eisen MB, Spellman PT, Brown PO, Botstein D, Cluster analysis and display of genome-wide expression patterns. **Proc Natl Acad Sci U.S.A.** 95 (1998):14863-14868.
- Etzold T, Ulyanov A, Argos P, SRS: information retrieval system for molecular biology data banks. **Meth Enzymol** 266 (1996):114.
- *Flint J, Tufarelli C, Peden J, Clark K, Daniels RJ, Hardison R, Miller W, Philipsen S, Tan-Un KC, McMorrow T, Frampton J, Alter BP, Frischauf AM, Higgs DR. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. **Hum Mol Genet** 10 (2001):371-382.
- Fischer S, Crabtree J, Brunk B, Gibson M, Overton G, BioWidgets: data interaction components for genomics. **Bioinformatics** 15 (1999):837-846.
- Gregory, PD. Transcription and chromatin converge: lessons from yeast genetics. **Curr Opin Genet Dev** 11 (2001):142-147.
- Kal AJ, van Zonneveld AJ, Benes V, van den Berg M, Koerkamp MG, Albermann K, Strack N, Ruijter JM, Richter A, Dujon B, Ansorge W, Tabak HF, Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. **Mol Biol Cell** 10 (1999):1859-1872
- *Karp PD, A strategy for database interoperability. **J Comput Biol** 2 (1995):573-586.
- *Kirkpatrick S, Gelatt CD, Vecchi MP, Optimization by simulated annealing. **Science** 220 (1983): 671-680.
- *Kruskal JB, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. **Psychometrika** 29 (1964):1-26.
- *Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K, Papadopoulos N, Vogelstein B, Kinzler KW, Strausberg RL, Riggins GJ, A public database for gene expression in human cancers. **Cancer Res** 59 (1999):5403-5407.
- *Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. **Science** 262 (1993): 208-214.
- Leser U, Lehrach H, Crollius HR, Issues in developing integrated genomic databases and application to the human X chromosome. **Bioinformatics** 14 (1998):583-590.
- *Lucasius CB, Kateman G, Understanding and using genetic algorithms. **Chemom Intell Lab Syst** 25 (1994):99-146.
- Lyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. **Nature** 409 (2001):533-538.
- Life Sciences Research Task Force of the Object Management Group, 1997, <http://www.omg.org/homepages/lsr>
- McCue LA, Thompson W, Carmack CS, Ryan MP, Liu JS, Derbyshire V, Lawrence CE Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. **Nucl. Acids Res** 29 (2001) 774-782.
- *Neuwald AF, Liu JS, Lawrence CE, Gibbs motif sampling: detection of bacterial outer membrane protein repeats. **Protein Sci.** 4 (1995):1618-1632.
- Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. **Science** 290 (2000):2306-2309.
- Searls, D, bioTK: componentry for genome informatics graphical user interfaces. **Gene** 163 (1995):GC1-16.
- *Sherlock G, Analysis of large-scale gene expression data. **Curr Opin Immunol** 12 (2000): 201-205.
- Siepel A, Farmer A, Tolopko A, Zhuang M, Mendes P, Beavis W, Sobral B, ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. **Bioinformatics** 17 (2001):83-94.
- Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J, WormBase: network access to the genome and biology of *Caenorhabditis elegans*. **Nucl Acids Res** 29 (2001):82-86.
- Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton N, Goble C, Brass A, TAMBIS: Transparent access to multiple bioinformatics sources. **Bioinformatics** 16 (2000):184-185.
- Van der Vlag J, den Blaauwen JL, Sewalt RG, van Driel R, Otte AP, Transcriptional repression mediated by polycomb group proteins and chromatin-associated repressors is selectively blocked by insulators. **J Biol Chem** 275, (2000):697-704.

*van Helden J, Andre B, Collado-Vides J, Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. **J Mol Biol** **281** (1998): 827-842.
 Van Steensel, B, Delrow, J., and Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. **Nat Genet** **27** (2001):304-308.
 *Velculescu VE, Zhang, L, Vogelstein, B, Kinzler KW, Serial analysis of gene expression. **Science** **270** (1995):484-487.
 Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW, Gene expression profiles in normal and cancer cells. **Science** **276** (1997):1268-1272.

12.
 a. Previous research related to this proposal
 b. Preliminary results

12a. Previous research related to this proposal

The Bioinformatics Laboratory (PI: A. van Kampen) supports biomedical research in the Academic Medical Center (AMC) through participation in research projects. We harbor expertise in the field of (statistical) data analysis, applied mathematics, database technology, information technology and molecular biology/chemistry. As part of collaborations with biomedical research groups new data analysis methodologies, tools and databases are developed. We developed the USAGE database application for the (statistical) analysis of SAGE data (van Kampen et al., 2001). Much experience was gained from this application with the implementation of research databases and web-enabled user interfaces. We further develop a 'metabolic' database to support research on metabolism of *S. cerevisiae*. This database integrates DNA micro-array data with information such as sub-cellular localization of enzymes, functional gene categories and activity of transcription factors. Graphical user-interfaces and multivariate data analysis methods support the interpretation of these data. Other projects include the development of a DNA micro-array database based on the EBI ArrayExpress model, multivariate analysis of (biological) data of patients with atherosclerotic vascular disease, and a study of variations in polyadenylation cleavage sites (Pauws et al., 2001). Major efforts were devoted to the establishment of a local large-scale 'bioinformatics infrastructure', which comprises hardware, bioinformatics software, local copies of up-to-date public databases (e.g. GenBank, Unigene, SwissProt) and other tools. Due to its complexity, one person of our group is dedicated to the maintenance of this infrastructure. The infrastructure includes two servers. We use a SUN E5500 server (UltraSparc, three 400 MHz processors, 8MB cache, 2 Gbyte RAM memory) running the Solaris 2.6 operating system. 200Gbyte of disk space is presently available for running projects. Additionally, we use a NT server for running Microsoft SQL server, which is used for our database applications. This infrastructure forms the basis of most of our bioinformatics projects and allows us to carry out the project described in this proposal.

The Neuroblastoma research group (PI: R. Versteeg) at the Dept. of Human Genetics (AMC, UvA) aims to identify the molecular genetic defects underlying the human childhood tumor neuroblastoma. Neuroblastomas have a variable clinical course, ranging from spontaneous regression to swift progression. Our model system consists of 250 tumor and control samples of neuroblastoma patients, as well as 25 neuroblastoma cell lines. The tumors are analysed for gene defects and by high throughput mRNA analysis, as provided by the SAGE technology. Presently 175.000 SAGE tags have been identified, each representing a mRNA expressed in neuroblastoma. We focus on development of bioinformatical tools to identify the genes belonging to these tags and to identify transcripts with a role in neuroblastoma pathogenesis. One example is the Human Transcriptome Map that is applied to identify candidate genes in chromosomal regions disrupted in neuroblastoma. A series of genes and pathways with a role in neuroblastoma pathogenesis has thus been identified (Spieker et al., 2001, van Limpt et al., 2000, and manuscripts in preparation). Furthermore, we focus on the identification of the N-myc downstream pathway in neuroblastoma. The N-myc oncogene is frequently amplified in aggressive neuroblastoma. SAGE libraries of neuroblastoma cell lines with and without ectopic N-myc expression identified over 350 genes that are either up- or down-regulated (Boon et al., 2001). The hierarchy of this pathway is studied. The ultimate goal of our research program is to understand the different pathways and their interactions that cause neuroblastoma, in order to identify candidate drug targets for innovative therapies.

A major theme of the research of Harmen Bussemaker at the Swammerdam Institute of Life Sciences (SILS) at the UvA is the study of regulation based on quantitative modeling of gene expression data using complete genome sequences. In close collaboration with the

12.continued

group of Siggia at Rockefeller University in New York City, two methods for discovering regulatory elements based on statistical analysis on a genome-wide scale were developed. MobyDick is a method for discovering over-represented motifs in DNA sequences in an intrinsic way that does not require reference frequencies and thus is well-suited to whole-genome analysis. It was successfully applied to the combined upstream regions of yeast to find putative new regulatory elements, and validated by comparison with a database of known transcription factor binding sites. REDUCE is a different method that can be used to discover and then quantify the effect of regulatory elements, based on an integrated analysis of genome sequence and genome-wide expression data. It was also tested on yeast and found to be highly successful. REDUCE has been applied in collaboration with several groups at SILS (Grivell, Klis, Van Steensel), AMC (Tabak), and UU (Holstege). Finally, GOQ (Gene Ontology Quantification) was recently developed as a tool for scoring gene categories (function, location, pathways, etc) based on expression data. In the context of the Amsterdam ICES-KIS project, there exists a close collaboration with both the Microarray Division at SILS and the Informatics Institute of the UvA aimed at the development of a database and tooling infrastructure for micro-array studies

12b. Preliminary results

The Human Transcriptome Map

We recently developed the Human Transcriptome Map (HTM), which generates gene expression profiles for any chromosomal region in normal and pathological tissue types (Caron et al., 2001). The HTM is based on computational analyses and database integration. The HTM integrates GeneMap'99, which gives the chromosomal position of 45,049 ESTs and genes, with over 4.4 million SAGE transcript tags. A major problem of SAGE is to identify the genes that belong to the experimentally obtained 10 bp tags in SAGE libraries. Previous software was designed to identify all possible gene assignments, and therefore accepted a high percentage of erroneous assignments as well. We therefore could not use these programs for the HTM. We designed new algorithms to identify the reliable SAGE tags for all human genes. We used the Est and mRNA sequences in the Unigene database. As these sequences are full of sequence errors and Unigene is not error-proof as well, we designed elaborate algorithms to correct for such errors. The application selected 422,088 reliable 3' transcripts from 807,165 transcript sequences, extracted 80,416 tags and rejected 27,790 of them as resulting from sequence errors in the Est database. The resulting AMCtagmap thus identified 52,626 tags for all human Unigene clusters. The final tag-to-gene assignment was checked by hand for 287 tags and showed an error rate of 6%. This was sufficiently low to apply the program to the Human Transcriptome Map. The next step in construction of the HTM was the integration of a large number of databases (e.g., Unigene, RHdb, GeneMap'99, AMCtagmap, CGAP SAGE libraries etc) into a relational database. Finally a web-enabled user interface was built (<http://bioinfo.amc.uva.nl>).

The resulting Human Transcriptome Map reveals a higher order organization of the genome, as there is a strong clustering of highly expressed genes. These domains, called Ridges, are found in all normal and pathologic tissues analyzed and thus represent a fundamental higher order organization of the human genome. Preliminary analysis of these domains for physical characteristics suggests that 50% of them have a high gene density. Besides this fundamental insight in genome organization and long-range transcriptional control, practical applications for cancer research exist. When global positional information is available for chromosomal defects in cancer, inspection of the HTM quickly identifies candidate genes over- or under-expressed in the chromosomal region. Several amplified oncogenes in neuroblastoma were thus identified (Spieker et al., 2001 and unpublished data).



Fig. 1. Expression profile of human chromosome 11 (p arm to the left, q arm to the right). Expression levels of 1208 genes are shown as a moving median with a window size of 39 genes. Two central Ridges are seen, as well as a telomeric Ridge at the p arm (Caron et al., 2001).

REDUCE

We recently developed a method, named REDUCE (Regulatory Element Detection Using Correlation with Expression), designed to provide a mathematical model to relate expression patterns to the occurrence of regulatory DNA sequences (Bussemaker et al., 2001). The parameters of this model correspond to the activity of each of a relatively small set of transcription factors. Effectively, the model exploits the information about the cell-wide regulatory network that is hidden in the non-coding part of the genome sequence. If a set of genes is co-regulated because the same transcription factor binding site occurs in the promoter region of each of the corresponding DNA sequences, then the expression levels for these genes can be described in the model by this sequence motif and a single parameter. The sequence motifs on which the model is based are found by an automated procedure that selects motifs until no significant correlation with expression remains, as was shown in detail for *S. cerevisiae*. Surprisingly, about 35% of the variation in expression is already captured by the simplest class of models that use consensus sequences to characterize protein binding sites, and assumes independent contributions when multiple elements occur in the promoter region of a gene.

13.
 a. Main lines of experimental approach

a. Main lines of experimental approach.

The project will develop along the steps in the following scheme. The period of the different steps and contribution of the requested post-docs is indicated. The different steps of this scheme are described in detail in 13b, following the same numbering.

b. Practical investigation scheme: more detailed outline for the whole period

Step	Postdoc	Year
1. Development of a sequence-based HTM database.	A, B	1
2. Identification of transcriptional domains by visual inspection and univariate statistics	A, B, C	1
3. Identification of regulatory elements for genes in these domains	C	1, 2
4. Extension of the HTM with additional genomic databases, array databases etc.	A, B	2, 3
5. Development of new architecture for HTM database application	B	2, 3, 4
6. Identification and characterization of transcriptional domains by multivariate techniques.	B, C	2, 3
7. Development of a Mouse Transcriptome Map	A, B	3, 4
8. Computational analysis of the murine and human domains for regulatory DNA sequences	C	3, 4
9. Comparative analysis of the human and murine transcriptome maps	A, B, C	4
10. Formulation of molecular genetic experiments to falsify the results of this bioinformatics project.	A, B, C	4

Postdoc A will be based at the dept. of Human Genetics (AMC, UvA). His/her background is in principle molecular genetics, with an interest in computational genomics. It is essential for this project that this post-doc has a throughout insight in genomic databases and especially in the molecular genomic technology used to obtain the data. Genomic databases are in full swing and therefore contain errors, ambiguous data and imprecise data. Furthermore, many entries suggest a precision that is only based on imprecise techniques (e.g. nucleotide numbers in the full human genomic sequence). The project therefore requires the continuous input of an experienced molecular biologist.

Postdoc B will be based at the Bioinformatics Laboratory (AMC, UvA). His/her background is in informatics with broad experience with the design and implementation of relational databases and extensive expertise with SQL. This postdoc should have a strong interest in software engineering in general and be interested in molecular biology.

Postdoc C will be based at the Swammerdam Institute of Life Sciences (UvA). His/her background will be in physics or applied mathematics. This postdoc should have general

knowledge of computational algorithms and experience with the development of new algorithms. Furthermore, she/he should be interested in the analysis of biological data. The technician will be based at the Bioinformatics Laboratory (AMC, UvA) and will carry out part the programming required for the e.g. development of database parser, filters and (graphical) user interfaces.

For all the work described in this proposal we will use standard technologies such as Perl and C for software development, relational databases (Microsoft SQL server and Oracle) for the database development, Splus package for statistical analysis, and HTML, Java(script), Perl and CGI scripting for the development of user-interfaces. All software and databases will be running on central computer servers and will be accessible via the intranet/internet by using a web-browser.

13b. Practical investigation scheme (numbers refer to the scheme in 13a)

1. Development of a sequence based HTM database.

Several database present partially annotated versions of the public draft sequence of the human genome (EMBL Ensemble, NCBI Genome View, UCSC Genome Assembly). We currently evaluate these databases in order to select the most appropriate database as basis for the new transcriptome map. We will develop the algorithms to assign SAGE tags to the annotated expressed sequences in this database. We will probably use of the human Unigene database that we previously used for the HTM. This will permit a smooth integration of the sequence data with the AMCtagmap database (Caron et al., 2001). To integrate these data we will implement a relational database. We will also implement a (graphical) user interface that allows querying the database. This will enable queries like 'fold induction of genes between any tissue'; 'average or mean expression level of genes in any region' etc. and the graphical presentation of these results. This interface will be directly connected to the database via predefined SQL queries, SQL templates. Following this trajectory, we expect to have a functional database within the first year of the project that can produce the data required for following parts of the project. However, since we plan to develop a new architecture in the second part of the project we design this first version of the HTM database in such a way that it can easily be translated to the new design.

2. Identification of transcriptional domains by visual inspection and univariate statistics

Upon visual inspection of the 'whole chromosome views' (Figure 1) of the first generation HTM we observed a clustering of highly expressed genes in domain-like structures. It required elaborate statistical testing to prove that these domains did not arose by chance. The first generation HTM suggested several other interesting domains that have not yet been analyzed and validated for statistical significance. The sequence-based second generation HTM will be much more detailed. We will start with a visual inspection of the new HTM for putative transcriptional domains. Firstly, we will identify the regions of very high or a very low expression. These regions will be defined in relation to gene density and other relatively straightforward features like CpG content and genomic repeats. We expect to define in this way a series of putative domain types, which will be tested for significance. This implies the calculation (e.g. via monte carlo) of the probability to this number of domains across the genome if the order of the genes is randomly permuted (Caron et al., 2001).

3. Identification of regulatory elements for genes in these domains

We will extend the application of the REDUCE method from yeast to human, where the required framework for associating DNA sequence with expression data is provided by the HTM. We expect that the proximal promoters in human can be analyzed without major modifications, although the regulatory elements are distributed over far larger regions than in yeast. However, we will implement suffix tree algorithms for the counting of substring patterns to obtain a significant speed-up of the algorithm, which make the analysis of the larger human genome feasible. In addition, this allows us to search for a wider range of sequence motifs. To improve the performance of the REDUCE algorithms we will extend the method to use position weight matrices rather than the currently used consensus motifs to describe DNA elements.

4. Extension of the HTM with additional genomic databases, array databases, etc.
For the further characterization of the long-range transcriptional domains we will extend the HTM database such that it allows the integration of other biological databases. We will integrate data about repeats, gene density, CG and CpG content across the genome and functional gene categories. We will also include the regulatory elements that were identified in step 3. At this stage we will also include DNA micro-array data, protein expression data and protein-DNA interaction data. We expect that the latter two types of data become available in large quantities during the project. We will also extend the user-interface to allow visualization of large quantities of heterogeneous data on a genome-wide scale.
5. Development of new architecture for HTM database application
Based on experience with other projects we expect that the rate at which new information becomes available or new research (i.e., database) questions arise can hardly be met by the speed of software and database development. We will therefore develop a three layer architecture for our database application, consisting of the user interface, the 'query module', and a dynamic relational database. This architecture will be easier to update than traditional relational database approaches, without extensive user interface programming or re-modelling of the database. The first layer will be the user interface to the HTM database. This interface should be sufficiently expressive to allow most queries that are also possible by a direct use of SQL. A second requirement is that the user interface can be automatically or semi-automatically adapted if the database model changes. In terms of SQL, an 'expressive' user interface implies capabilities that go beyond the 'query by example' strategy. Following data warehousing and data mining principles, the user interface must allow the compilation of data cubes (e.g., merging datasets), ordering of data, or to summarize particular subsets of data. The second layer comprises the 'query module', which is software component that serves as the middleware between the user interface and the relational database. Its machinery decides how to compose a SQL query from user input. The complexity of this module depends on the expressiveness allowed by the user interface and complexity of the data model. The module will, however, not be restricted to particular database implementation. Consequently, if the database is modified this information should automatically become available to the module. The third layer comprises the relational database. The requirements imposed on the data model have a large impact on the complexity of the user interface and the query module. In the context of our new architecture we will investigate what data model is optimal. This includes the requirement that the database must be dynamic, i.e., permit the automatic addition of new tables or attributes to the table and to relate these with existing tables. This may imply that the data model of the first version of the database is adapted, although we try to anticipate on this.
6. Identification and characterization of transcriptional domains by multivariate techniques
To identify and characterize the long-range transcriptional domains in the HTM we will use multivariate data analysis techniques (e.g., hierarchical clustering, k-means, principal component analysis, multi-dimensional scaling). These analysis are not straightforward for several reasons. First, the data set combines interval-scaled variables (e.g., expression levels, gene density) and discrete variables (e.g., functional gene classification). Scaling issues and missing values (e.g., gene function) increase the complexity of the analysis. Secondly, most multivariate data analysis techniques can be parameterized in many ways. Consequently, one therefore can 'tune' these methods until results are satisfactory (e.g., clusters of domains with similar characteristics). However, since the interpretation of these results is not always obvious, they may results from mathematical artifact due to incorrect parameterization. Alternatively, different (parameterizations of these) methods may give different clusters of domains with other biological interpretations. For these reasons, we will perform a comparison of these methods with an emphasis on the biological and mathematical interpretation of the results. In addition we will apply statistical methods (e.g., bootstrapping) to establish the significance of these results. To overcome possible limitations of these techniques, we will develop new methods based on global optimization techniques (e.g., simulated annealing (Kirkpatrick et al., 1983) and genetic algorithms (Lucasius et al., 1994)), which are more robust for analyzing complex data sets.
7. Development of Mouse Transcriptome Map

We will construct a Mouse Transcriptome Map using similar algorithms and database architecture as for the Human Transcriptome Map. The map will be based on the murine genomic sequence, which is currently established in by a public-private consortium and will be available in draft in the first half of the grant period. We will identify the SAGE tags for all mouse genes with the algorithms used for the AMCtagmap. In short, we will use mouse Est and mRNA databases (with almost 2 million mouse Ests), select the 3'-end sequences and extract the 10-bp SAGE tags from them. These tags will be subjected to computational analysis and sequence alignment routines using the mouse Unigene database (NCBI), to identify and reject tags resulting from sequencing errors in the individual Est and mRNA sequences in the databases. The resulting Mouse-tagmap lists all tags for all known mouse genes. This database will be valuable for researchers constructing mouse SAGE libraries and will be accessible on the internet. Construction of the Mouse Transcriptome Map finally requires mouse SAGE libraries. A growing number of mouse libraries is currently established and the first large library is published by the NCBI SAGE library repository. In addition, mouse SAGE libraries are currently sequenced at the Neurozintuigen Laboratory (Dr. F. Baas) at the Academic Medical Center in Amsterdam. Details of the construction of the Mouse Transcriptome Map will depend on the structure of the annotated mouse genome sequences. As these annotations built on the experience with human annotation projects, our experience with human maps will facilitate the construction of the mouse map.

The mouse map will be analyzed for Ridges and other transcriptional domains by the same approach as used for the Human Transcriptome Map. Visual inspection and statistical validation will identify transcriptional domains. Subsequently, more sophisticated cluster analysis methods will identify additional landmarks in the mouse genome.

8. Computational analysis of the murine and human domains for regulatory DNA sequences

The analysis of enhancer regions in human sequences is more complex than the analysis of proximal promotor regions, since the enhancer elements may be very distant (10-100kb) from the proximal promotor region. This situation therefore comes close to identification of regulatory sequences dispersed throughout large transcriptional domains. To make these analyses feasible, we will make use of the human-mouse cross-species comparison of non-coding DNA to separate candidate regulatory regions from 'junk' DNA on basis of sequence conservation. Similar approaches and algorithms were recently developed (McCue et al, 2001) and are used at the group of Siggia at Rockefeller. To exploit this cross-species comparison we will collaborate with the latter group to integrate phylogenetic comparison methods with REDUCE. Such an approach can be very beneficial as reported by Flint et al. (2001). We will also extend our computational methods to include sequence signals regulating long-range transcription domains. For instance, boundary elements separating regions of high and low expression may be characterized by looking for sequence motifs that correlate with these regions. We will build a mathematical model that allows us to identify the activities of regulatory proteins via their binding sites for different tissues or developmental stages. This will contribute to a further understanding of the differences between various tissues in terms of regulatory pathways.

9. Comparative analysis of the human and murine transcriptome maps

We will finally integrate the human and mouse transcriptome maps into a Comparative Human-Mouse Transcriptome Map (CHMTM). To this end, we will use maps of human-mouse synteny and align the transcriptomes of both species. The present versions of the human-mouse synteny maps have a limited resolution, but detailed sequence-based maps are being developed (e.g. for human chromosome 19), and will become available within the next two years. The aligned mouse and human transcriptome maps will answer the question whether mouse and man have the same long-range chromosomal architecture. Insight in this fundamental question is also of major importance to translate these findings into molecular biological experiments, as the mouse genome can relatively easy be manipulated. In addition, the CHMTM will reveal whether the boundaries of the large blocks of mouse-man synteny coincide with the boundaries of transcriptional domains in mouse and man. This should reveal whether the mammalian genomes are variants of only a few hundred functional building blocks, or in contrast can result from unlimited rearrangements that are not restricted by a domain architecture of the genome.

chromosomes but follow a structured plan. Highly active genes are found close together in large clusters, suggesting that they form a kind of factories that boost gene activity. In this project, we aim to improve this database application significantly. We will use state-of-the-art information technology to relate the complete human DNA sequence to gene activity levels in a wide range of tissues. Analytical tools will then identify the different domains that are formed by the genes on the human chromosomes. We will analyze the properties of these domains and identify the segments of the human genome that determine the overall structure that controls the activity of all human genes. The vast amount of data that have to be integrated requires the establishment of a solid informatics infrastructure.

18.
signature*

date: 27 march 2001

the applicant:
AHC van Kampen

- :with your signature you declare also to have taken notice of the accompanying instruction sheets