Characterization of Regulatory Genomic Regions. Development of Databases and Sequence Analysis Tools

TRADAT

1995 - 1998

- Final Report -

Contents:

Scientific achievements - Overview

The concept of the TRADAT consortium included to provide an integrated platform for databases and software tools for the analysis of regulatory genomic regions. According to this concept, a set of databases was established and maintained, mainly EPD (Eukaryotic Promoter Database) and TRANSFAC (Transcription factors and their genomic binding sites). These data sources were successfully linked with each other and with a number of external databases. In addition to this integrated data resource, a series of software tools for the identification of individual regulatory elements and the characterization of their context were developed. These tools as well as the underlying patterns were subjected to systematic evaluation, optimization and experimental verification.

The concept of using weight matrices for the detection of transcription factor binding sites was extended and incorporated into higher order software tools. TRADAT contributed to the further development of MatInspector, ModelGenerator, FastM, and ModelInspector.

A number of new programmes have been developed for designing random oligonucleotides satisfying various user-defined constraints. These tools have been applied to generate the training and test oligonucleotide sets used in the experimental characterisation of NF1-binding sites. This work was performed in close collaboration between the laboratories of the TRADAT consortium.

In order to integrate the promoter with the gene structure information we have developed two integrative and WWW-accessible tools. The GeneBuilder system is able to predict the main functional signals (promoters, splice sites, TATA-box and Poly-A sites) and the coding regions (CDS). The program uses several parameters and modules for gene structure analysis, homology search and automatic annotations. The program package Theatre is designed for the comparative study of gene structures and analysis of genomic features, especially with respect to motifs likely to be involved in the regulation of gene expression. It exploits commonly used sequence analysis tools (ConsInspector, MatInspector, BLAST) and biological sequence databases (TRANSFAC and SPTR) to determine or predict the positions of coding regions, repetitive sequences and transcription factor binding sites in families of DNA sequences.

The tools developed by the TRADAT consortium as well as numerous additional routines were applied on the analysis of regulatory features of the Fugu genome and revealed several novel features of this extremely compact genome.

As a consequence of the TRADAT project public domain versions of EPD, TRANSFAC, MatInspector, FastM, ModelInspector, GeneBuilder and Theatre were installed freely on the WWW-servers of the TRADAT members and made available to the scientific community.

Scientific achievements - Main text

E. WINGENDER (COORDINATOR)
GBF - Gesellschaft für Biotechnologische Forschung mbH
Braunschweig, Germany

Work package 1, activity 1:

Within the context of the TRADAT project, the links of the TRANSFAC database, which is maintained at the GBF, Braunschweig, to external databases have been significantly extended. Thus, the links to the Eukaryotic Promoter Database (EPD) by P. Bucher, Epalinges (Switzerland) have been established in close cooperation with the Epalinges group. Also, cross-links with the Brookhaven database on protein structures (PDB) and with FlyBase (for Drosophila entries) have been established. In addition to these and other regular database cross-links, links to several additional external data sources such as the kidney development database (http://www.ana.ed.ac.uk/anatomy/database/kidbase/kidhome.html) or the tooth gene expression database (http://honeybee.helsinki.fi/toothexp/) have ben included in the TRANSFAC database, giving rise to a total of nearly 13,000 links. Also as part of this project, the GENE table has been established and maintained which serves as a link between TRANSFAC and TRRD (Transcription Regulatory Region Database; Institute of Cytology and Genetics, Novosibirsk, Russia). As a joint project with the ICG in Novosibirsk, the database COMPEL has been generated and maintained and appropriate crosslinks have been introduced into the TRANSFAC database as well. TRANSFAC has been successfully migrated from MS SQL Server to Oracle 7 to improve performance, version stability and backup security.

Work package 1, activity 2:

We succeeded in developing a Java-based input client for the TRANSFAC database. It connects directly to the relational database system(s) used for the maintenance of TRANSFAC (MS SQL Server; Oracle). This routine has been extensively tested during its use for internal updating. As a next step, it will be installed on the transfac server which is accessible by the „outside world“, storing the submitted data in a temporary database systems which is analogously structured to the TRANSFAC database and contains its full content to enable linking newly entered data sets to pre-existing ones.

Work package 1, activity 3:

A comprehensive and hierarchical transcription factor classification system has been developed and continuously updated and adapted. It is based on the features of the DNA-binding domains of the thus far cloned transcriptin factors. For this system, six levels of organization were defined and were called superclasses, classes, families, subfamilies (optional), genera and species. In total, four superclasses were identified, subdivided into 27 classes. The classification has been made publicly accessible on the transfac server under http://transfac.gbf.de/TRANSFAC/cl/cl.html .

Applying this classification scheme, it was possible to develop a semiautomatic routine which evaluates relationships among search patterns for TF binding sites (see below, WP 2, Activity 2). Another was to investigate correlations between the relationship of DNA-binding domains and the DNA motifs recognized by these domains.

Work package 2, activity 1:

The MATRIX table of the TRANSFAC database is the basis of a selected matrix library which is part of the MatInspector sequence analysis program (a joint development with the GSF group, see GSF project). To clear the output for redundancy, efforts have been made to develop and apply a general methodology for comparison of all available matrix search patterns. By correlating computed matrix distances with the relationship of the corresponding transcription factors according to the factor classification (see above), distance ranges could be identified where two matrices are definitely related and, thus, recognize largely overlapping if not identical sequence elements, or where they are necessarily unrelated. Nevertheless, it turned out that this process cannot be completely automatized and still will involve human expert knowledge.

To parametrize all search patterns individually and optimally, positive and negative sequence sets have been compiled for all matrices presently stored in the database. Negative training sets were derived from exon ≥2 sequences extracted from the EMBL data library. Positive training sets were obtained either from the TRANSFAC database (genomic binding sites) or from the original literature (e. g., sequences selected by CASTing approaches). The latter have been inserted into the database, for which a modification of its data structure was necessary. On the basis of the MatInspector algorithm, three threshold have been computed for each pattern: one minimizing the false positives, one minimizing the flase negative, and the third one representing the minimum of the sum of both error rates.

Moreover, the conventional matrix approach has been extended to generating and applying dinucleotide weight matrices. Since larger training sets are required for this approach than for the generation of single nucleotide matrices, it could be applied not for the whole set of transcription factors conventional matrices are available for. In total, we deduced 116 dinucleotide matrices. Up to know, 48 were systematicall compared with corresponding mononucleotide matrices, about half of them (25) exhibiting a significant improvement in terms of false positive and false negative matches. For certain transcription factors, no improvement was seen indicating that no nearest neighbour-influences are relevant for their binding sites.

Work package 3, activity 1:

Since the present methodology of, e. g., matrix search does not apply for composite elements, a method was developed to characterize and identify this kind of regulatory entity. This method considers the interdependent scoring quality of both constituent elements, but may also include their spacing and relative orientation. Applying this approach onto NFAT composite elements, we generated a search pattern that successfully identified nearly all known NF-AT sites with only a low number of false positives. It reveals a significantly high concentration of (potential) NF-AT sites specifically in T-cell specific promoters over muscle-specific promoters or random sequences indicating the specificity of the method. Scanning the whole EMBL database, a series of biologically highly interesting new potential composite elements of NFAT type was suggested, some of them are presently in the process of experimental validation.

To detect significant clusters of potential transcription factor binding sites (TSF) in genomic sequences analysed, e. g, with MatInspector or ConsInspector, a tool was developed which groups them according to their position and their scoring. We implemented a fuzzy clustering algorithm to enable individual elements to be assigned to different clusters at the same time. To clear for redundant matches which could pretend a high TFS density, a filter was incorporated which makes use of the results obtained from the overall matrix comparison described under WP2, activity 1.

L. MILANESI
CNR – Istituto Tecnologie Biomediche Avanzate
Milan, Italy

The following description of the GeneBuilder system collectively refers to
Work package 2, activity 4,
Work package 3, activity 3,
Work package 4, activity 2,

GeneBuilder system for sequence analysis.
In order to combine and visualize the results in predicting functional signals, coding regions and homology searches in databases we have developed the GeneBuilder system. The user may select various parameters to refine the gene structure prediction. For example, it is possible to build a gene model by selecting only the exons having high coding potential score without considering homology with a specific protein. It is also possible to refine the analysis by selecting different proteins from a list of the most homologous proteins automatically generated. Even in case of low homology GeneBuilder is able to predict the gene structure with high accuracy. The GeneBuilder system can be used for exploring alternative gene structure models by taking into account different properties of CDS and signal sequences. The interactive use of GeneBuilder for gene structure prediction is useful for finding alternative variants of transcription, translation initiation, polyadenylation and alternative splicing sites. The GeneBuilder system allows the user to explore alternative gene model variants by using homology with the proteins and to optimise the models by selecting different parameters. The graphical visualisation of the results is particularly important for the prediction of short exons and the analysis of several genes in a query sequence. GeneBuilder offers a selection of modules for a wide range of sequence analysis.

Module for searching and masking the repeated elements.
The presence of repeated elements can create problems in sequence analysis. For example some long repeated elements contain ORFs and can be recognized as potential genes. Also the output from the database search may be saturated by a number of highly scoring matches with repeated elements present in the sequence databases. The analysis of repeats can be very difficult due to high heterogeneity and short length of repeated elements. In this module the program searches for all potential repeats and masking them in the query sequence by comparing it with a collection of repeated elements (Jurka et al., 1992).

Module for homology database search.
Homology searches are very important for functional mapping, homology with a known functional region can suggest the function of a quary sequence. In particular, when the homologous protein sequence is already known and/or EST matches are detected, then the gene structure can be reconstructed with high accuracy. This module can be used to search the protein homology in SWISSPROT (Bairoch and Boeckmann, 1991) database and the nucleotide homology with EST sequence database.

Module for coding region prediction.
This module is used for finding the potential coding regions (CDS) by using global statistical properties of coding regions and the potential functional signals (splicing sites, start and stop codons) (Milanesi et al., 1993).

Module for revealing potential TATA-box and poly(A) signals.

These signals are usually very well conserved and they are located upstream and downstream of the gene structure. The prediction of these signals is based on the Hamming-Clustering method (Milanesi et al., 1996). The location of these signal together with the locations of the potential CDS, in some case, can help during the determination of the complete gene structure.

Module for CDS homology refinement.
The analysis of homology between potential peptides translated from predicted genes and proteins from databases can be very important particularly in the case of weak but significant similarities. This procedure can be repeated several times in order to refine the accuracy of the prediction.

Module for prediction of the potential binding sites of transcription factors.
The determination of potential binding sites can be very important in gene functional study. This module is based on MatInspector program (Quandt et al., 1995) and has been applied to the detection of potential binding sites of transcription factors. This program uses a library of weight matrices useful for finding binding site of Vertebrates, Fungi, Insects, Plants.

Module for CpG island prediction.
This module is used for prediction of the potential CpG islands. To locate the CpG islands we use the definition given by Gardiner-Garden and Frommer (1987): The CpG islands are regions greater than 200 bp in length which have more than 50% G+C ($p(G)+p(C) > 0.5$) and have a CpG content of at least 0.6 of that expected on the basis of the G+C content of the region ($p(CpG) > 0.6*p(C)*p(G)$). CpG iselands are frequently found at the 5' ends of genes. This procedure is used automatically by GeneBuilder, since it can be very useful in long sequences containing a number of genes to locate the gene rich regions.

Module for the user interface.
This module is able to accept the sequence to be analyzed in any format. All modules are offered and results presented interactively using Java applets and HTML or send to the user by electronic mail. The GeneBuilder program has been included in the WebGene launcher. Sequences can be input using the standard cut copy and paste commands or uploading from a user file. The cut and paste field can only transfer sequences of up to 20Kb due to the limitation of HTML browsers. The user can customize all relevant parameters before commencing the analysis.

Module for sequencing error correction.
As many sequencing strategies are error-prone, sequences may include apparent nucleotide substitutions, insertions and deletions. This fact should be taken into account in sequence analysis. Several algorithms have been developed for the analysis of frameshift errors (Xu et al., 1995; Fichant and Quentin, 1995) based on statistical properties of coding sequences. The accuracy of error correction is not very high, but can be significantly improved by using information from homologous proteins (States and Botstein, 1991; Posfai and Roberts, 1992). Error correction techniques should be used carefully, since there are many pseudogenes in eukaryotic genomes. In this module we have implemented two stages of error analysis: in the first we generate only a report on potential errors, the second we performs an automatic correction. In this procedure of revealing errors and then correcting them, we use dicodon potential and homology with a selected protein sequence. We also look for substitutions in stop codons in protein coding regions. This is a rarer type of error then the introduction of indels.

Availability: The GeneBuilder system is implemented as a part of the WebGene a the URL: http://www.itba.mi.cnr.it/webgene.

M. BISHOP
UK Human Genome Mapping Programme Resource Centre
Hinxton, United Kingdom

Work package 4, activity 1, and
Work package 5:

1. Objectives
Characterisation of regulatory genomic regions employing TRADAT databases and sequence analysis tools on a model minimalist vertebrate genome system of Fugu.

2. Main Achievements
Detailed information connected with the Fugu Landmark Mapping Project such as the project description, project protocols, cosmid sequences, clone sequences and the BLAST matches to the sequence databases is available via the UK HGMP-RC World Wide Web site at the following URL http://fugu.hgmp.mrc.ac.uk (Elgar et al., 1999, manuscript in progress). This genome survey scanning provides high throughput sequencing and the cloned sequences are not assembled to become contiguous fragments. The cosmids from the Fugu genomic library were selected randomly and sequenced using a shotgun scanning approach. Approximately 50 clones were sequenced per cosmid. Primer and vector sequences are identified and clipped. Databases comprising the positions of repetitive DNA sequences and coding sequences pertaining to the Fugu genome have been constructed and the results obtained were compared with equivalent studies carried out in other genomes.For example, we defined a new and comprehensive analysis in which 501 theoretically possible microsatellites with a repeat unit of 1-6 bases were used to query Fugu DNA (i.e. 11.338Mb). 6042 microsatellites were identified and categorised. In decreasing order, the twenty most frequently occurring microsatellites are AC, A, C, AGG, AG, AGC, AAT, AAAT, ACAG, ACGC, ATCC, AAC, ATC, AGGG, AAAG, AAG, AAAC, AT, CCG and TTAGGG. The twenty most frequently occurring microsatellites represent 81.79% of all microsatellites identified. Our results indicate that one microsatellite occurs every 1.876Kb of DNA in Fugu rubripes. 11.55% of the microsatellites are detected in open reading frames that are predicted protein coding regions. With respect to the proportion of microsatellites present in open reading frames and the total abundance (bp) of all microsatellites, the genome of Fugu rubripes is similar to the genome of other vertebrate species. Previous estimates performed indicate that approximately 1% of many vertebrate genomes are comprised of microsatellite sequences. However, many differences prevail in the abundance and frequency of the individual microsatellite classes. Many of the frequently occurring microsatellites in Fugu rubripes are known to code in other species for regions in proteins such as transcription factors, whilst others are associated with known functions, such as transcription factor binding sites and form part of promoter regions in DNA sequences of genes (for more information please see Edwards et al., 1998). Similar analyses have been carried out on chicken sequencing project (Clark et al., 1999) and an updated analysis on 30Megabases of Fugu DNA. The co-ordination of the UK HGMP Puffer Fish Website provides additional information regarding the morphology, anatomy, physiology and the behaviour of the puffer fish in relation to gene products and the regulation of gene expression (URLhttp://fugu.hgmp.mrc.ac.uk/fugu/pffp/pf.html).

Development of new tools

Theatre is designed for the comparative study of gene structures and analysis of genomic features, especially with respect to motifs likely to be involved in the regulation of gene expression.

Theatre exploits commonly used sequence analysis tools (ConsInspector, MatInspector, BLAST) and biological sequence databases (TRANSFAC and SPTR) to determine or predict the positions of coding regions, repetitive sequences and transcription factor binding sites in families of DNA sequences. The information is intuitively displayed and can reveal patterns that might not otherwise have been noticed. Theatre is of use in investigating function and evolution in divergently related DNA sequences. Theatre is a user-friendly, web-based environment for comprehensive feature formatting of DNA sequence alignments and can produce graphical displays and publication quality hard-copies of functional and structural features in aligned homologous DNA sequences.Availability: Theatre will be accessible for use by registered members of the Bioinformatics facilities of the UK Human Genome Mapping Project Centre Resource Centre (HGMP-RC). Registration is free to the academic community. Information regarding the registration process is available at the HGMP-RC's web site at the following URL http://www.hgmp.mrc.ac.uk.

T. WERNER
GSF – Zentrum für Umwelt und Gesundheit
Oberschleißheim, Germany

Work package 2, activity 1:

The matrix library is very important for the program MatInspector and it´s internal quality is crucial for the quality of the results. We identified three major areas crucial for the predictive value of MatInspector analyses.

First of all, the absolute number of matrices represented in the library is very important for general applicability of the tool. For this reason we extended the matrix library considerably in tight collaboration with the GBF group.

The second important point is the fact that matrices are not all completely independent of each other. This is best illustrated by several matrices which have been derived from binding sites of the same transcription factor determined under different experimental conditions. Although it is often impossible to combine these matrices into a single one, they are nevertheless tightly related. Therefore, the library produces redundant output and efforts have been made to develop and apply a general scheme for comparison of all available matrix descriptions. This was done in parallel in the GBF and the GSF group. In order to avoid redundancies without sacrificing valuable information present in overlapping matrices we introduced a new concept termed matrix families. A matrix family is defined as a group of matrices which represent variations of binding sites believed to be bound by the same or at least closely related factors. We developed a method allowing rapid and largely automatic grouping of the matrices in order to facilitate the matrix family definition.

We carried out complete family classification of all matrices in the MatInspector library. It is important to notice that our classification essentially agrees with the results of the much more elaborative studies of the GBF group which was based on matrix alignments (see GBF report for details). These independent approaches demonstrated that the method we used describes matrix similarities sufficiently well for our purpose and can be used routinely from now on.

The third important point is implementation of strict quality control during matrix generation in order to avoid production of matrices with poor recognition characteristics. Assessing this property after the often lengthy matrix definition process should be avoided in order to avoid wasting of resources and time on definition of unsuitable matrices. We were able to develop and successfully test a strategy which incorporates iterative matrix development including rigorous quality assessment at each step of the iteration. This way only matrices meeting clearly defined quality standards will be generated.

Another important question is the suitability of the scoring algorithm used for matrix searches. We carried out an extensive comparison of available weight matrix methods which has been published. As a main result we found that the quality of the matrices is more important than the scoring algorithm. Nevertheless, the MatInspector algorithm showed some advantages over other methods especially in quantitative predictions.

In conclusion work package 2 was successfully completed as far as the GSF project was concerned.

Work package 3, activities 1 and 2:

Our approach to describe functional contexts of promoter and other regulatory regions was further developed (joint effort between the TRADAT project and other resources of the group) and the

initial versions of three programs could be completed in the meantime. The program ModelGenerator is capable of deducing the organization of binding sites within a set of sequences representing similar promoters in line with our initial concept of a matrix of binding sites as described in the original program. One dimension of this matrix is now describing the linear extension of a promoter set (essentially the sequences) while other dimensions represent individual binding sites. ModelGenerator determines which binding sites are commonly found at specific positions or position ranges within the promoter. This would be equivalent to the elements of the promoter matrix which are occupied by individual elements as opposed to empty elements of this matrix which are forbidden for certain binding sites.

However, for practical reasons we came to term this structure a promoter model rather than a matrix since this is easier to imagine and to visualize than a multidimensional matrix which would be empty for most of its parts. Modeling with ModelGenerator is still a largely interactive process and requires expert knowledge in order to achieve optimal results. We showed that these results can describe a promoter class precisely.

Scanning of databases is made possible by another program called ModelInspector which uses models generated by ModelGenerator to identify matches in other sequences. This required the development of a unified scoring algorithm since ModelInspector has to cope with very distinct types of elements. For example, matrix and consensus elements score similarities in nucleotide sequences to a weight matrix while secondary structure elements are scored by the calculated binding energy of the structure. ModelInspector unifies these different scoring mechanisms and yields just two scores for each match, a score for element match quality and a score for the fit of relative element distances to the model.

ModelGenerator is demanding both predefined training sets of sequences as well as expert knowledge by the user. Therefore, it cannot be seen as a tool for the average user. Very often only few data about a single or very few sequences are known. Scientists are also most interested to check whether a peculiar setup of two binding sites found in one sequence is likely to be a mere statistical match or if such a combination is characteristic for a specific set of sequences.
We developed a new program called FastM because ModelGenerator cannot be used to answer these questions. FastM allows the user to define a model from scratch (requires no sequence data at all) and to use these models directly with ModelInspector in order to reveal the significance of the model. We developed a WWW-interface for FastM allowing direct use even by unexperienced users. However, this WWW-version was limited to models containing two binding sites in order to keep the load on our server manageable.

In cooperation with the Swiss groups (Mermod and Bucher) and the GBF group we were able to improve the binding site description of NF-1 binding sites. The MatInspector matrices for NF-1 are both of low quality and did not correlate in predictions with the experimentally verified binding affinities. P. Bucher showed that construction of a matrix with good correlation to the experimental results is possible. However, this required employment of the experimental affinity data which are not available for most matrices. Therefore, we attempted to build a FastM model of NF-1 binding sites composed of two half sites without using any of the experimental data. This model showed much better correlation with experimental data than the MatInspector matrices but could not match the accuracy of the matrix directly derived from the experimental data.
However, the approach showed that it is possible to improve binding site descriptions by modeling even when no experimental data are available.

P. BUCHER
ISREC
Epalinges, Switzerland

The main achievements by the ISREC team were:

1. The design, introduction and productive application of a new format of the Eukaryotic Promoter Database EPD allowing for representation of a great variety of new types of information.

2. The development of novel, user-friendly, WWW-based interfaces for accessing EPD, including the SEVIEW applet viewer which enables the user to navigate from EPD to EMBL to SWISS-PROT within the same graphic environment.

3. The development of bioinformatics support tools for experimental characterisation of protein binding sites.

Following is a concise description of the work carried out under each work package activity.

Work package 1, activity 4:

Maintenance, format extensions, and documentary improvements of the Eukaryotic Promoter Database EPD.

As mentioned above, the format of EPD was completely revised. Highlights of the new formats are: (i) an ID and accession number scheme conforming to the conventions of other major DNA and protein databases; (ii) "smart" cross-references defining the position of sequence objects in related databases (e.g. EMBL, TRANSFAC) relative to the transcription initiation site; (iii) support of secure sequence data retrieval using NI's rather than accession numbers as sequence identifiers; (iv) links between literature references and transcript mapping experiments; (v) unlimited space for documentary information. In a focused effort to populate the new data fields with useful information, every EPD entry was substantially revised and expanded. New software was developed for the maintenance of EPD in the new format. Along with the format revisions, about 100 new entries were added to EPD.

Work package 1, activity 5:

Enhanced cross-referencing of EPD with other databases including EMBL, SWISS-PROT, TRANSFAC, GDB, FlyBase, HOVERGEN, and bibliographic references.

A total of 4930 database cross-references, all manually checked, and 2143 automatically generated bibliographic links to MEDLINE have been added to EPD. The cross-referenced databases include those listed above, with the exception of HOVERGEN, plus the mouse genome database MGD. In addition, the organism (OS) line was redesigned such as to allow automatic hyperlinking to the NCBI taxonomy pages. A special effort has been made to interconnect TRANSFAC and EPD with the newly introduced "smart links" containing information that can be used by graphic visualisation tools for display of gene features along the DNA sequence. We further made an attempt to exhaustively cross-reference EPD with EMBL (each promoter in EPD is on average represented by two to three redundant EMBL sequences). This provides an effective mechanism to assemble more information about a specific promoter entry through indirect links, and to identify synonyms and redundancies in related databases.

Work package 1, activity 6:

Parallel efforts to enhance the interoperability of EPD with other databases and data access tools (ISREC).
Unique nucleotide sequence identifiers (NI's) were added to all EMBL cross-references to enhance the interoperability of the TRADAT databases (see also activity 1.7).

Five new interfaces for accessing EPD have been developed: (i) The ISREC-TRADAT document server providing access via HTTP hyperlinks to a variety of locally stored databases, including EPD, TRADAT, and SWISS-PROT. (ii) SEView, a java applet viewer for sequence objects. This tool exploits the smart links provided by the new EPD format in order to display the location of various sequence objects along a DNA sequence axis. (iii) The EPD query form allowing field-restricted searches for combinations of query-strings. (iv) A promoter sequence download page enabling the user to extract a Fasta- or EMBL-formatted database of promoter sequences corresponding to a subset of EPD of desired length. (v) Access via the SRS server of the SWISS EMBNet node.

Work package 1, activity 7:

Theoretical work on the problem of maintaining database compatibility within a federation of dynamic databases.

The TRADAT consortium has developed and maintained a network of asynchronously updated but interdependent databases. This could lead to data inconsistencies when information from different databases is combined over the network. For instance, a hyperlink may point to a recently deleted entry. More serious problems arise when the sequence of an EMBL entry is changed such that sequence positions indicated in EPD or TRANSFAC entries become incorrect. We have designed a system which ensures data coherence within a database federation without interfering with local maintenance procedures. The components of this systems are: (i) unique identifiers for different versions of database entries, (ii) network servers delivering archived older versions of database entries, (iii) control information (sequence strings, checksums) which allows on-the-fly verification of the compatibility of two documents, (iv) control information-based automatic correction mechanis restoring data compatibility between entries. All components of this system have been tested and productively incorporated into the automatic procedures used for adaptation of a current EPD release to a new EMBL or TRANSFAC release.

Work package 2, activity 2:

Development of weight matrix or profile descriptions of the binding specificities of major eukaryotic transcription factors.

A number of new programmes have been developed for designing random oligonucleotides satisfying various user-defined constraints such as falling within a certain length or G+C content range, matching a given consensus sequence in a unique fashion, or having a certain predicted affinity to a DNA binding protein. These tools have been applied to generate the training and test oligonucleotide sets used by the UNIL team in the experimental characterisation of NF1-binding sites.

N. MERMOD
Université de Lausanne
Lausanne, Switzerland

Efficient computational tools are being increasingly used for high throughput screening of newly determined DNA sequences in genome sequencing projects, so as to allow for instance the identification of candidate disease-responsible genes. In particular, effective DNA sequence analysis requires reliable information on the potential function and regulation of the identified genes. This is usually provided by algorithms that allow the prediction of the regulatory DNA sequences that act as binding sites for transcription factor proteins. At present, the efficacy of currently available algorithms and methods to predict eucaryotic regulatory sequence has in most cases not been evaluated experimentally.

Work package 2, activities 2, 3 and 5:

The UNIL contribution in the TRADAT project has mainly been two-fold. First, we evaluated the efficacy of various computer methods for the prediction of regulatory sequences. Second, we tested several experimental methods to measure protein-DNA interaction, and we chose one of them to analyze systematically a collection of binding sequences for a given transcription factor, the human CTF/NF1 protein. This work was performed in close collaboration with other laboratories of the TRADAT team, in particular ISREC, GSF and GBF. We also exchanged and provided scientific data with CNR and HGMP, for instance in the context of the modeling of the protein structure of CTF/NF1 proteins by the latter laboratory. In this sub-project, we decided to concentrate and analyze fully one such regulatory protein, rather than spread limited forces onto several types of regulatory sequences as originally proposed, because of the interest raised by initial assays with the CTF/NF1 protein, and also because of cuts in the funding of the original proposal. Results generated in this subproject provided partner TRADAT laboratories with a description of the biochemical parameters to consider in computer modeling of sequence specific DNA binding, and also with a collection of binding data as training and/or evaluation sets that was used for the development of new computer tools. These tools will and results will be made publicly available on the web sites of TRADAT, for instance through the TRANSFAC database.

In this study, we first measured the in vitro binding affinity of several DNA sequences for a human member of the CTF/NF1 family of transcription factors, and compared these experimental data with the output of currently available computer programs (Roulet et al., 1998). None of the tested prediction tools gave results that fitted well the experimental binding strength. Furthermore, the computer tools did not clearly allow an efficient qualitative prediction of binding sites, as attempts to define a cut-off line produced large proportions of false positive and/or false negative results. These results indicated that the reliability of the prediction tools is seriously limited by the set of training data available in the scientific literature, and also by the prediction methods and algorithms used to generate the computer tools. These results thus called for an assessment of the general quality of the data used to construct prediction tools and of the assumption and methods used to generate the prediction tools.

To address these issues in depth, we embarked in a systematic series of measurements for the in vitro binding affinities of potential CTF/NF1-binding DNA sequences, generated in collaboration with the ISREC and GBF laboratories. This study indicated that additional computational parameters, usually not taken into account by current weight matrix models, are required for accurate binding site prediction. For instance, current prediction methods make the assumption that the interactions of the regulatory protein with distinct base pairs are independent,

which implies that independent substitution should have additive effects on the binding strength. However, we found that this is not true for most of the combinations of substitutions that we evaluated. Second, this study indicated that the length of the binding sites is quite flexible, another parameter usually not taken into account by current tools. These findings were then used to generate a model for the binding of this transcription factor to regulatory sequences. This model formed the basis for the construction of new computer prediction tools by the ISREC and GSF laboratories. Our subsequent experimental evaluation of the novel algorithms and computer methods indicated that these accurately predict the binding affinity for natural and synthetic DNA sequences, thus validating the prediction tools. These findings thus not only highlighted some of the limitations of usual weight matrix tools, but, more importantly, they also formed the basis for the development of novel prediction methods. These new methods and concepts are likely to be of general significance for computer analysis, for instance to predict the many regulatory sites that consist of repeated or palindromic sequences.

Flow chart on objective fulfilment

No.	Milestone	Work Package	Status
1	Complete coverage of the current literature by EPD in its present format	1.4	largely done after reformatting of EPD[1]
2	Complete and accurate cross-referencing between EPD, TRANSFAC, and EMBL, and the delivery of data access software and network services supporting these links	1.1, 1.2, 1.3, 1.5, 1.6, 1.7	done
3	Library of weight matrices for transcription factor binding sites	2.1	done
4	Delivery of high precision weight matrices for TBP and NFI	2.2	done for NF-1/CTF[2]
5	Setting into operation of the standardized procedures to evaluate protein binding site prediction tools	2.3, 2.5	done
6	Final definition of filter criteria best correlated with potential of functionality	3.1	done
7	Methodology and format of context description	3.1, 3.2, 3.3	done
8	Integration of TRANSFAC and EPD into IGD	4.1, 4.2	integration in GeneBuilder and Theatre[3]

[1] To ensure long-term maintenance of EPD in a high quality, emphasis was given to reformatting and developing software for facilitated maintenance in the new format as well as to a complete revision of the preexisting data.
[2] The consortium decided to prefer a thorough in-depth analysis of the NFI DNA-binding properties and its functional implications rather than to include TBP as second independent system; nevertheless, work on a second transcription factor has been initiated.
[3] Since integration into IGD appeared not to provide an appropriate state-of-the-art tool for the purpose of the TRADAT project, the consortium decided to develop multilevel database and database-software integrations instead.

Deliverables:

After 1st year: TRANSFAC cross-references and EPD complete;
EPD format exension proposal available for discussion among partners;
software for generation of weight matrices and profile descriptions and trained neural networks available.

The databases will be publicly available, the software will be made available to the partners.

After 2nd year: Extended library of weight matrices available through TRANSFAC; EPD format extensions implemented; basic cross-referencing scheme implemented (TRANSFAC, EPD, EMBL, TRRD), access to promoter sequences via WWW servers. Some progress in the domain of software interoperability and IGD integration. Basic methodology of symbol description of regulatory regions.

Depending on the progress, first sequence analysis tools will be made publicly available.

All deliverables were made available to the partners and/or the public, respectively, until the end of the TRADAT project.

Co-operation links

During the five TRADAT workshops (see below) and between them, intensive mutual discussion caused a vivid exchange of ideas on all aspects of the project. This led to the exchange, installation and utilising programs, discussions and feedback on the application of sequence databases and sequence analyses tools employed in the study of minimalist model vertebrate genome and for the integration of individual tools in the TRADAT launcher and Theatre.

In particular, there were intensive contacts between the GBF and ISREC groups for establishing the EPD-TRANSFAC corss-links, and tight links between the GBF and GSF group were established in matrix-related work. Th latter cooperation included several work visits.

Both Swiss groups, GSF and GBF cooperated in the experimental verification of NF-1 binding sites and improvement of binding site descriptions by modeling rather than a single weight matrix. This includes Numerous visits between ISREC and UNIL contributors and a visit of N. Mermod (UNIL) to T. Werner (GSF).

Loose contacts have been established with the EU bioinformatics project coordinated by K. Wolfe, Dublin. The participation of the GBF group in the EU-funded CORBA project (coordinated by the EBI) provided contacts to the other participants of that project as well.

In 1998 the GSF group established links between the TRADAT project and Genomatix Software GmbH, a spin-off company from GSF. Genomatix provided the TRADAT project with advanced versions of the basic tools which augmented especially the development of the matrix generation strategy.

The following TRADAT Workshops were held:

Munich, 29.02. - 01.03.1996
Milan, 21.10. - 22.10.1996
Hinxton, 01.06. - 03.06.1997
Lausanne, 05.02. - 07.02.1998
Braunschweig, 19.11. - 20.11.1998

1st TRADAT Meeting
Munich, 29.02. - 01.03.1996

Participants: K. Frech, K. Quandt, R. Schneider, T. Werner (GSF), L. Milanesi (ITBA CNR), M. Bishop, Y. Edwards (HGMP RC), T. Heinemeyer, E. Wingender (GBF), P. Bucher, T. Junier (ISREC), N. Mermod (UNIL)

The 1st TRADAT meeting has been organised at the GSF in Munich by Dr. Thomas Werner.

This first workshop centered around the issues to be concerned by individual groups during the first year. Cooperations to be established during this first year focused on GBF and GSF who intended to collaborate in work package 2. Due to the reduced funding GSF had not employed personnel so far and proceeded with existing staff only. It was agreed to focus mainly on developments within the individual groups for the first year and to start cooperations during the second year.

2nd TRADAT Meeting
Milan, 21.10. - 22.10.1996

Participants: D. D'Angelo, L. Milanesi, I. Rogozin (ITBA CNR), M. Bishop, Y. Edwards (HGMP RC), T. Heinemeyer, E. Wingender (GBF), P. Bucher, R. Cavin-Périer, T. Junier (ISREC), N. Mermod, E. Roulet (UNIL)

The 2nd TRADAT meeting has been organised at the CNR ITBA in Milan by Dr. Luciano Milanesi.

During this meeting Wingender and Bucher introduced the database innovation in TRANSFAC, EPD and COMPEL. The new format of EPD was presented in detail and discussed. Bucher and Mermod suggested a technique for a systematic comparison of the results generated from the available tools with the DNA binding affinities realised by experimental approaches. Wingender reports on a new program developed by T. Werner, FastM, able to define models in the absence of a training set of sequences. Bishop reported the progress done at HGMP MRC in sequencing the genome of the pufferfish Fugu. Milanesi and Bishop presented the progress of more specialised tools for comparing and analysing the Fugu with Human genome. Milanesi introduced the progress done in developing the program for gene model prediction. The new method based on Hamming Clustering was shown for the TATA-box and Poly-A prediction. Wingender reported also the last development of the ModelGenerator/ ModelInspector package realised by the Werner group. In conclusion the TRADAT Web Server (http://www.itba.mi.cnr.it/tradat) was illustrated by Milanesi.

The 3rd TRADAT meeting.
Hinxton, 01.06. - 03.06.1997

Participants: M. Bishop, Y. Edwards (HGMP RC), T. Heinemeyer, E. Wingender (GBF), P. Bucher, R. Cavin-Périer, T. Junier (ISREC), N. Mermod, E. Roulet (UNIL), L. Milanesi (ITBA CNR), R. Schneider, T. Werner (GSF)

The 3rd TRADAT meeting has been organised at the HGMP RC in Hinxton by Drs. Yvonne Edwards and Martin Bishop.

The TRADAT collaboration comprises five work packages
1. Database issues
2. Methods for identificatipon of individual regulatory regions
3. Context analyses of regulatory elements.
4. Development of integrative software.
5. Application on a selected system.

The TRADAT project developments.

Edgar Wingender and Thomas Heinemeyer
1. TRANSFAC WWW server and TRANSFAC release 3.1
Cross referenced with EMBL, SWISSPROT, PIR, Flybase, PROSITE & EPD.
Integration with SRS 5.0 and DBGET.

2. Peer-reviewed electronic submission of new TFBS and TFs.
3. Status of transcription factor classification system
4. SAGA: Tool for the identification of structural characteristics such as minor
major grooves in DNA sequences using a genetic algorithm.
5. TFC: A tool for detecting potential TF-binding site clusters using fuzzy logic. TFC is platform independent and written in C.
6. Patternsearch (C-program works a bit like signalscan)

Thomas Werner
A new method to develop specific models for regulatory DNA regions.
Organisation analyses of promoter sequences using Coresearch, Matinspector, Consinspector, GenomeInspector and ModelGenerator. (ModelGenerator is not yet publicly available but can be used at GSF with T. Werner as it is still being developed).

Phillip Bucher, Thomas Junier, Rouayda Cavin
EPD Developments
1. EPD new format (release 50)
2. cross reference with EMBL, SWISSPROT, Flybase, OMIM, PIR.
3. flat file downloadable
4. netfetch server at ISREC - network-based access to EPD
5. EPD sequence download facility
6. graphical WWW interface with java applet

Nicolas Mermod, Emmanuelle Roulet, Phillip Bucher, Edgar Wingender
First comparisons of bioinformatics and experimental data for the NF1 binding sites. This study compared binding affinities of NF1 with the wild-type NF1 binding site and NF1 binding sites with base substitutions. Computed bioinformatics scores (such as the weighted matrix, Ci scores, core or matrix similarity values) are being assessed for their goodness of fit with experimentally determined NF1 binding affinities. Additionally, molecular recognition relationships between protein and DNA have been investigated.

Martin Bishop and Luciano Milanesi
GENEDB system providing a WWW-based interactive, intuitive GUI for obtaining and viewing representations of all genomic features from UTR, EPD, TRANSFAC, SWISSPROT, EMBL. GENEDB is under development and operates using various java applets and tools.

Luciano Milanesi
Software development of tools to detect features in DNA sequences
that modulate activation and termination of transcription.
1. Webgene: An integrated tool for gene structure prediction.
2. Poly-A prediction in 3' regions of eukaryotic genes.
3. TATA-box prediction in 5' regions of eukaryotic genes.
4. Transcription Initiation Site using Hamming-Clustering network, a new type of artificial neural network, specially designed to work with binary data. Each base is represented using 4 bit. Excellent GUIs.
5. Milan Server

Martin Bishop and Yvonne Edwards
1. Latest developments in the UK HGMP RC Fugu Landmark mapping project, and associated web sites developed including the Puffer Fish Web site.

2. Identification and functional comparative characterisation of repetitive DNA sequences in the genome of the Japanese Puffer-fish Fugu rubripes.
3. 3D protein Structure Prediction from sequence: How good are we?

4th TRADAT Meeting
Lausanne, 05.02. - 07.02.1998

Participants: P. Bucher, R. Cavin-Périer, T. Junier (ISREC), N. Mermod, E. Roulet (UNIL), M. Bishop, Y. Edwards (HGMP RC), L. Milanesi (ITBA CNR), T. Werner (GSF), E. Wingender (GBF)

The 4th TRADAT meeting has been organised at the ISREC in Epalinges by Dr. Philipp Bucher and Prof. Nicolas Mermod and their groups.

All participants present the latest results and progress in their TRADAT activities.
Special emphasis was given to the discussion on the comparative evaluation of matrix search tools and the correlation of these results with experimentally determined binding constants, as exemplified for transcription factor NFI/CTF.

T. Werner and E. Wingender report on the initial steps of establishing spin-off companies from GSF and GBF (Genomatix GmbH and BIOBASE GmbH, resp.).

5th TRADAT Meeting
Braunschweig, 19.11. - 20.11.1998

Participants: X. Chen, T. Heinemeyer, I. Liebich, T. Meinhardt, I. Reuter, F. Schacherer, E. Wingender (GBF), P. Bucher, R. Cavin-Périer (ISREC), N. Mermod, E. Roulet (UNIL), M. Bishop, Y. Edwards (HGMP RC), L. Milanesi (ITBA CNR), R. Schneider, T. Werner (GSF); as guest: H. Karas (CEO of BIOBASE GmbH, Braunschweig)

The 5th TRADAT meeting has been organised at the GBF in Braunschweig by Drs. Thomas Heinemeyer and Edgar Wingender.

During the 5th and last TRADAT workshop in Braunschweig, a final resume was made of what has been achieved by the participating groups. It became clear that all activities addressed in the Work Plan have been dealt with and were successfully finalized (see Scientific Report). The participants agree that the project as a whole has to be considered successful. Appointments have been made about which results still have to be jointly or separately published in near future, and upon future possibilities of cooperation, as far as the personnels situation permits to do so in the post-TRADAT era. Moreover, the participants agree that whenever possible, a new grant application shall be prepared and sent to the EC in the 5th Framework of Biotechnology. When the workshop was held, no detailed information was available about these possibilities.

The participants were also informed about the present status of the two companies founded by members of the consortium: Genomatix GmbH (T. Werner, Munich) and BIOBASE GmbH (E. Wingender, Braunschweig).

Illustrations

Fig. 1: Generalized gene structure comprising coding regions (green boxes) and regulatory regions (promoter, enhancers, silencers; red boxes). Promoter and enhancer regions consist of arrays of individual or composed transcription factor binding sites. In many cases, transcription factors autoregulate the expression of their own gene (white arrow).

Fig. 2a: Generation of a nucleotide distribution matrix. Sequence elements with a defined and experimentally proven function are aligned, frequencies of the four bases in all positions are counted and composed in a matrix. Matrix-based methods are the basic approach of the pattern recognition tools developed, used and optimized by the TRADAT consortium.

Fig. 2b: Use of a nucleotide distribution matrix for sequence interpretation. A nucleotide distribution matrix (represented by the pattern of white an dgreen squares indicating allowed and „forbidden“ nucleotides) has been generated as explained in the legend of Fig. 2a and is used to scan a genomic DNA sequence. Matching positions are then indicated likely to be binding sites of a certain transcription factor, in the example shown AP-1. Using the data of the TRANSFAC database, this suggestion is then translated into a functional information, here: element possibly mediating gene response to growth factors.

Fig. 3: Using the ModelGenerator program, a promoter model for mammalian actin genes has been derived. It comprises a number of ubiquitous transcription factor binding sites (USF, CAAT, SRF, and Sp1) as well as muscle-specific types of TATA-box (mTATA) and transcription initiator element (mINI). The individual elements can be detected by MatInspector using a TRANSFAC-derived matrix library, their presence in defined relative orientation and distance within genomic sequences is then detected by the program ModelInspector.

Fig. 4: Entry page of the GeneBuilder system for predicting gene structures.

Fig. 5: Using the GeneBuilder system, the user is guided through a number of subsequent steps to perform the in silico analysis of a sequence.

Fig. 6: Results output of the GeneBuilder system. The results of the individual GeneBuilder program modules are graphically displayed (top). The user may zoom into the map down to the nucleotide level (bottom).

Fig. 7: The TRANSFAC WWW server at the GBF, Braunschweig, provides access to the TRANSFAC database, to a number of additional external databases through the SRS system as well as to several new data sources developed in the same group (top left). TRANSFAC can be accessed through html forms (top right), some of the information provided is displayed in a graphical format (bottom left). Also included is the comprehensive transcription factor classification scheme developed in the context of the TRADAT project (bottom right).

Fig. 8: The WWW server of the GSF group (Munich; top left) offers access to a number of programs (bottom left). Among them, MatInspector is one of the most frequently used routines (top right). More advanced, ModelGenerator/ModelInspector is a program pair for generating specific promoter models and using them for the analysis of genomic sequences (bottom right).

Fig. 9: At the ISREC (Epalinges), the user can query the Eukaryotic Promoter Database (EPD).

The system enables the user also to download promoter sequences according to their biological species, number of upstream/downstream nucleotides etc.

Fig 10: Assembly of WWW server pages developed by the TRADAT consortium. The TRANSFAC server at GBF provides access to the TRANSFAC database and a number of additional database and software tools (top right), whereas the ISREC at Epalinges maintains a WWW server for the Eukaryotic Promoter Database (EPD, center). At the CNR ITBA, the WebGene launcher including the GeneBuilder package can be used (top left). The GSF group provides a bunch of software tools for the detection of transcription factor binding sites and their specific combinations such as FastM (bottom left). The tools developed in the TRADAT project have been applied on the analysis of the Fugu genome (bottom right).

Publications

The TRADAT project published 44 scientific papers (two of them in preparation or submitted for publication), 3 of them being joint publications of two or more partners:

GBF, Braunschweig
Project leader: E. Wingender

1. Kel, A. E., Kel-Margoulis, O., Babenko, V., Edgar Wingender (1999). Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. J. Mol. Biol. (submittted for publication).

2. Heinemeyer, T., Chen, X., Karas, H., Kel, A. E., Kel, O. V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., Wingender, E. (1999). Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms Nucleic Acids Res. 27, 318-322

3. Wingender, E., Chen, X., Heinemeyer, T., Kel, A. E., Liebich, I., Meinhardt, T., Schacherer, F., (1998). A hierarchy of databases for modeling gene regulatory mechanisms. Proceedings of the German Conference on Bioinformatics GCB '98, Cologne University; O. Zimmermann, D. Schomburg (eds.).

4. Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., Podkolodny N. L. and Kolchanov, N. A. (1998). Databases on Transcriptional Regulation: TRANSFAC, TRRD, and COMPEL Nucleic Acids Res. 26, 362-367.

5. Pickert, L., Klawonn, F. and Wingender, E. (1998). Trancription regulatory region analysis using signal detection and fuzzy clustering. Bioinformatics 14, 244-251.

6. Pickert, L., Reuter, I., Klawonn, F. and Wingender, E. (1997). Transcription regulatory regions revealed by signal detection and fuzzy clustering. Proceedings of the German Conference on Bioinformatics GCB '97, Kloster Irsee, Bavaria; H. W. Mewes, D. Frishman (eds.), 99-101.

7. Wingender, E., Kel, A. E., Kel, O. V., Karas, H., Heinemeyer, T., Dietze, P., Knüppel, R., Romaschenko, A. G., Kolchanov, N. A. (1997). TRANSFAC, TRRD and COMPEL: Towards a federated database system on transcriptional regulation. Nucleic Acids Res. 25, 265-268.

8. Wingender, E. (1997). Classification scheme of eukaryotic transcription factors. Mol. Biol. 31, 574-600; Mol. Biol. Engl. Tr. 31, 483-497 (1997).

9. Wingender, E., Karas, H., Knüppel, R. (1996). TRANSFAC Database as a Bridge between Sequence Data Libraries and Biological Function. Pacific Symposium on Biocomputing '97 (PSB'97), R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein (eds.). World Scientific, Singapore - New Jersey - London - Hong Kong 1996, pp. 477-485.

ITBA-CNR, Milan
Project leader: L. Milanesi

1. Rogozin, I.B. D'Angelo, D. and Milanesi L. (1999) Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences. Gene 8, 266, 129-137.

2. Kolchanov NA, Ponomarenko MP, Kel AE, Kondrakhin YuV, Frolov AS, Kolpakov FA, Goryachkovsky TN, Kel OV, Ananko EA, Ignatieva EV, Podkolodnaya OA, Babenko VN, Stepanenko IL, Romashchenko AG, Merkulova TI, Vorobiev DG, Lavryushev SV, Ponomarenko YuV, Kochetov AV, Kolesov GB, Solovyev VV, Milanesi L, Podkolodny NL, Wingender E, Heinemeyer T (1998) GeneExpress: a computer system for description, analysis, and recognition of regulatory sequences in eukaryotic genome. Ismb 1998;6:95-104

3. Milanesi, L. and Rogozin, I.B. (1998) Prediction of human gene structure. In: Guide to Human Genome Computing (2nd ed.) (Ed. M.J.Bishop), Academic Press, Cambridge, 215-259.

4. Rogozin, I.B. and Milanesi, L. (1997) Analysis of donor splice sites in different eukaryotic organisms. J. Mol. Evol., 45, 50-59.

5. Rogozin, I.B., Milanesi, L. and Kolchanov, N.A. (1996) Gene structure prediction using information on homologous protein sequence. Comput. Applic. Biosci., 12, 161-170.

6. Milanesi, L., Muselli, M. and Arrigo, P. (1996) Hamming-clustering method for signals prediction in 5' and 3' regions of eukaryotic genes. Comput. Applic. Biosci., 13, 399-404.

HGMP RC, Hinxton
Project leader: M. Bishop

1. Edwards Y. J. K., Frith M., Tivey A., Cheng, I., Elgar G. & Bishop M. J. Theatre: A Novel Tool for the Comparative Interpretation and Display of Functional and Structural Elements in DNA Sequences. Manuscript in preparation to be submitted to Bioinformatics by 28th February, 1999).

2. Elgar G., Clark M. S., Edwards Y. J. K., Meek S., Smith S., Umrania Y., Warner S., Williams G. & Bishop M. J. Characterization of the compact model genome of the Japanese puffer fish(Fugu rubripes) using a cosmid sequence scanning approach. Extended abstract published in the Proceedings for The German Conference in Bioinformatics (GCB 98), Cologne University, 7-10th October 1998.

3. Edwards Y. J. K., Frith M., Elgar G. & Bishop M. J. Theatre: a novel tool for the comparative investigation and display of evolutionary diversity of functional features in DNA sequences. Extended abstract published in the Proceedings for The German Conference in Bioinformatics (GCB 98), Cologne University, 7-10th October 1998.

4. Elgar G., Clark M. S., Edwards Y. J. K., Meek S., Smith S., Umrania Y., Warner S., Williams G. & Bishop M. J. Characterization of the compact model genome of the Japanese

puffer fish (Fugu rubripes) using a cosmid sequence scanning approach. Extended abstract published in the Proceedings for The International Conference on Bioinformatics of Genome Regulation and Structure. Institute of Cytology and Genetics, Novosibirsk, Russia, held on the 24th-31st August 1998. Volume 2, pp. 292-295.

5. Edwards Y. J. K., Frith M., Elgar G. & Bishop M. J. Theatre: a novel tool for the comparative investigation and display of evolutionary diversity of functional features in DNA sequences. Extended abstract published in the Proceedings for The International Conference on Bioinformatics of Genome Regulation and Structure. Novosibirsk, Russia, held on the 24th-31st August 1998. Volume 2, pp. 307-310.

6. Elgar G., Clark M. S., Edwards Y. J. K., Meek S. E., Smith S., Umrania Y., Warner S., Williams G. & Brenner S. (1998). Fugu Landmark Mapping Project: A cosmid sequence scanning approach to characterising the genome. Human Genome Meeting, Turin 28th-30th March 1998. Programme and Abstract Book, pp. 19, abstract number 81.

7. Elgar G., Clark M. S., Edwards Y. J. K., Meek S. E., Smith S., Umrania Y., Warner S. & Williams G. (1998). Preliminary data analysis on three human syntenic regions in Fugu. Human Genome Meeting, Turin, 28th-30th March 1998. Programme and Abstract Book, pp. 19, abstract number 82.

8. Clark M. S., Edwards Y. J. K., McQueen H., Meek S. E., Smith S., Umrania Y., Warner S., Williams G. & Elgar G. (1998). Sequence scanning chicken cosmids: An effective genome screen. Gene (in press).

9. Edwards Y. J. K., Elgar G., Clark M. S. & Bishop M. J. (1998). The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, Fugu rubripes: perspectives in functional and comparative genomic analyses. Journal of Molecular Biology, 278:843-854.

10. Frith M., Edwards Y. J. K., & Bishop M. J. (1997). Raising the curtain on Theatre: a new tool for the comparative investigation of functional features in DNA sequences. UK HGMP Resource Centre Genome News Winter 1997, pp. 15-17.

GSF, Oberschleißheim
Project leader: T-. Werner

1. Werner, T. (1998). Identification and characterization of promoters in eukaryotic DNA sequences. Mammalian Genome, in press.

2. Klingenhoff, A., Frech, K., Quandt, K., Werner, T. (1998). Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity, in press Bioinformatics.

3. Brack-Werner, R., Werner, T. (1998). Lentivirus Long Terminal Repeats (LTRs). In: Human immunodeficiency virus : Biology, molecular biology and Immunology, Ed. N. Saksena, Medical Systems, 57-83.

4. Lavorgna, G., Wagner, A., Bonicelli, E., Werner, T. (1998). Detection of potential target genes in silico?, Trends in Genetics 14, 375-376. *
(No grant citations allowed)

5. Morgenstern, B., Frech. K., Dress, A., Werner, T. (1998). DIALIGN: Finding local similarities by multiple sequence alignment, Bioinformatics 14, 290-294.

6. Frech, K., Quandt, K., Werner, T. (1998). Muscle actin genes: A first step towards computational classification of tissue specific promoters, In Silico Biol. 1, 0005.

7. Frech, K., Quandt, K., Werner, T. (1997). A new method to develop highly specific models for regulatory DNA regions. In Lecture Notes in Computer Sciences (LNCS) Bioinformatics, Springer Verlag, 79 - 87.

8. Frech, K., Quandt, K. Werner, T. (1997). Software for the analysis of DNA sequence elements of transcription. Comp. Appl. Biosci., 13, 89-97.

9. Blusch, J., H., Haltmeier, M., Frech, K., Sander, I., Leib-Mösch, C., Brack-Werner, R., Werner, T. (1997). Identification of endogenous retroviral sequences based on modular organization: proviral structure at the SSAV1 locus. Genomics 43, 52-61.

10. Frech, K., Danescu-Mayer, J., Werner, T. (1997). A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J. Mol. Biol. 270, 674-687.

11. Frech, K., Quandt, K. Werner, T. (1997). Finding protein-binding sites in DNA sequences: the next generation. Trends in Biochem. Sci., 22, 103-104.

12. Frech, K., Werner, T. (1996) Specific modeling of regulatory units in DNA-sequences, Pacific Symposium on Biocomputing 97 (Hrsg. R. Altman, A. K. Dunker, L. Hunter, T. E. Klein), 151-162.

13. Frech, K., Brack-Werner, R., Werner, T. (1996). Common modular structure of Lentivirus LTRs. Virology 224, 256-267.

ISREC, Epalinges
Project leader: P. Bucher

1. Cavin Périer R., Junier T., Bonnard, C. and Bucher P. (1999). The Eukaryotic Promoter Database EPD: recent developments. Nucleic Acids Res. 26, 355-359.

2. Roulet E., Fisch I., Junier T., Bucher P. and Mermod N. (1998) Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. In Silico Biology 1, 0004. <http://www.bioinfo.de/isb/1998/01/0004/>

3. Junier T. and Bucher P. (1998). SEView: a Java applet for browsing molecular sequence data. In Silico Biology 1, 0003. <http://www.bioinfo.de/isb/1998/01/0003/>

4. Cavin Périer R., Junier T. and Bucher P. (1998). The Eukaryotic Promoter Database EPD. Nucleic Acids Res. 26, 355-359.

5. Duret L. and Bucher P. (1997). Searching for regulatory elements in human non-coding sequences. Curr. Opin. Struct. Biol. 7, 399-406

University of Lausanne
Project leader: N. Mermod

1. Roulet, E., Bucher, P., Schneider, R., Wingender, E., Dusserre, Y, Werner, T. and Mermod N. In Silico predictions of DNA Binding Sequences for Mammalian Transcription Factor. Manuscript in preparation.

2. Roulet E, Fisch I, Junier T, Bucher P and Mermod N. (1998) Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. In Silico Biol. **1**, 1-8, <http://www.bioinfo.de/isb/1998/01/0004/>.

Press release

The regulation of gene expression is a key issue in modern molecular biological basic and applied research. The intepretation of genomic DNA sequences with regard to their regulatory potential therefore is an absolute requirement for rendering genome data into useful information. The TRADAT consortium established a number of database resources (mainly EPD and TRANSFAC), software tools (e. g., FastM, ModelGenerator, ModelInspector) and integrative WWW-based user interfaces (Theatre, TRADAT Launcher) that approach this goal. Application of these tools in connection with experimental validation achieved several scientific break-throughs in the analysis of regulatory elements and regions. The TRADAT launcher will be maintained by concerted efforts of the participating groups even after the TRADAT project has been finalized.
Two participants of the TRADAT consortium have founded start-up companies (Genomatix Software GmbH, Munich; BIOBASE Biological Databases GmbH, Braunschweig) which are prepared to exploit some of the TRADAT results.

Industrial communication

There is an increasing awareness in the pharmaceutical and biotechnical industry that genome-driven research and development will take over the leading role in the discovery and development of new products. In this context, gene expression studies become increasingly important as can be evidenced by the interest modern chip technology attracts. However, the overall „expression space“ defined by a multitude of spatial, temporal and conditional parameters is too large to be completely covered by any experimental approach. Therefore, in-depth analysis of genomic sequences will depend on a thorough computational („*in silico*“) investigation of their regulatory potential.

For this purpose, the six partners of the TRADAT consortium have developed and implemented a set of database and software tools, integrated under WWW-based user interfaces, which enable the user to perform gene identification analyses with special emphasis though not exclusive on the regulatory features.

It is a frequent observation that RD products from the public research, though being highly innovative and scientifically sound, lack the extent of user-friendliness that is required for granting a wide-spread use of the results. In fact, it is not even the task of publicly financed research groups to do this kind of work.

For this purpose, start-up companies have been founded by two TRADAT participants. Genomatix Software GmbH, Munich, and BIOBASE - Biological Databases GmbH, Braunschweig, started to develop appropriate products and offer them to the pharmaceutical and biotechnological industry.