DBMS-based EST Clustering and Profiling for Gene Expression Analysis

Hong Hai Do, Prof. Dr. Erhard Rahm
Institut für Informatik, Universität Leipzig

Dr. Knut Krohn, Prof. Dr. Ralf Paschke
IZKF, Universität Leipzig
Content

- Motivation
- Previous work
- DBMS-based expression analysis
- Data model for sequence data
- Database population
- Analysis possibilities
- Summary
Motivation

- **Expressed Sequence Tags (EST)**
  - single pass reads from randomly selected cDNA clones
  - 300-500 base pairs
  - sequences for protein coding regions: fragments of genes

- **Availability of huge amounts of ESTs in public databases**
  - GenBank
  - dbEST
  - UniGene, Stack, TIGR

- **Steady growth due to numerous sequencing projects**
  - double in size every 22 months (GenBank)
Gene Expression Analysis Using ESTs

- Previous work:
  - Vasmatzis et al. (1997): Discovery of three new genes specifically expressed in human prostate
  - Rosenthal et al. (1999): Investigation of genes specifically expressed in normal and tumor tissues
  - other tissues/organs: thyroid (Univ. Leipzig), retina, ...

- Main approach: *In silico* differential display - expression analysis using clustered ESTs
  - EST clusters represent candidate genes
  - existence and number (count) of ESTs of a gene: qualitative and quantitative expression level of the gene in the corresponding organ or tissue
  - sequence annotation available in databases: organism, organ, tissue, cell type, disease, clone, library, ...
In silico Differential Display

1. Extraction of relevant ESTs
2. Sequence comparison
3. Clustering
4. Profile determination
5. Expression profile

EST database (dbEST, UniGene)

Sub-bank

Vasmatzis: # different tissues in cluster

# non-specific ESTs
# specific ESTs
Thyroid Investigation (Univ. Leipzig)

- Identification of 2200 thyroid-related sequences from dbEST
  - 1/3 are known genes (excluded)
  - 1/3 with repetitive elements (excluded)

- 38 EST subbank-specific clusters, i.e. containing only thyroid-related sequences
  - 10 clusters with more than 3 ESTs

- PCR experiments
  - expression of 6 clusters confirmed
  - 5 clusters with higher expression in thyroid than in other tissues (brain, heart, liver, lung, muscle, testis, kidney)
  - 1 cluster with expression only in thyroid
Limitations of Previous Implementations

- Largely manual process
  - manual execution of analysis steps
  - manual evaluation of results

- Small data sets in flat file-based local data management
  - small number of ESTs taken into account
  - concentration on a particular tissue or organ in the investigation
  - small spectrum of other tissues or organ for comparison

- Generation of simple, often imprecise expression profiles
  - expression profiles of EST clusters with respect to tissues and organs

- Limited analysis possibilities
  - fixed query forms
  - access to “canned“ results from investigations previously conducted under a particular focus
  - online analysis not possible
DBMS-based Gene Expression Analysis

- Use of DBMS technologies for data management (IBM DB2, Oracle, Microsoft SQL Server, etc.)
  - huge data sets
  - flexible und declarative query languages (SQL)
  - security, performance, availability, ...

- Data integration and preparation
  - construction of a uniform data model for sequence data and analysis
  - extraction, transformation and cleaning of sequence data from different sources (public EST/cluster databases)
  - pre-computation of sequence similarities, clusters, cluster profiles to support interactive analysis

- Data analysis:
  - web-based query forms for simple queries
  - complex queries with user-friendly interfaces to SQL
  - integration of commercial tools, e.g. for data mining
Modeling Sequence Data

<table>
<thead>
<tr>
<th>Consensus</th>
<th>UniqueRegions</th>
<th>OverlapRegions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><code>Base 0</code></td>
<td><code>Base n</code></td>
</tr>
<tr>
<td></td>
<td>0 1 2 3 4 5 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequence 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequence 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequence 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequence 4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster / CandidateGene</th>
<th>SequenceOverlap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><code>AAGTTAGTAA</code></td>
</tr>
<tr>
<td></td>
<td><code>AAGTCGGTAA</code></td>
</tr>
<tr>
<td></td>
<td><code>AAGTCAGTAA</code></td>
</tr>
</tbody>
</table>

Candidate SNPs
Data Model for Sequence Data

- Publication
- Method
  - Sequence
  - Comparison
  - ClusterProfile
  - Cluster
  - Region
  - ConsensusSequence
- Annotation
- UniqueRegion
- OverlapRegion
- SNP
Sequence and Sequence Similarity

### Annotation hierarchy

- **Organism**
  - **Tissue**
    - **Organ**
      - **Celltype**
      - **Disease**

### Comparison
- Sequence1ID
- Sequence2ID
- MethodID
- ClusterID
- StartPosition1
- StartPosition2
- OverlapLength
- Score
- Probability

### Method
- MethodID
- MethodName
- Parameter

### Sequence
- SequenceID
- dbESTID
- AnnotationID
- PublicationID
- ESTName
- GenBankGI
- CloneID
- Source
- DNAType
- SequenceString
- Quality
- Description

### Annotation
- AnnotationID
- Organism
- Organ
- Tissue
- Disease
- Celltype

### Publication
- PublicationID
- MedLineID
- Title
- Authors
- Journal
- [Year]
Cluster and Expression Profile

- **Sequence**
  - **Annotation**
  - **Comparison**
  - **Cluster**
    - ClusterID
    - ClusterName
    - GeneSymbol
    - Chromosome
  - **ClusterProfile**
    - ClusterID
    - AnnotationID
    - ExpressionLevel
  - **ConsensusSequence**
    - ClusterID
    - SequenceString
Sequence Alignment and SNP
Data Integration and Cleaning

- Selection of relevant data sources
  - ESTs from dbEST
  - clusters from UniGene, STACK oder TIGR

- Implementation of source-specific programs (wrappers) for data import
  - identification of useful sequences
  - transformation of data from text files (e.g. FASTA) according to the data model
  - loading: persistent storage in the relational database

- Data cleaning and preparation
  - filter sequences with >5% Ns
  - trim leading vector sequences (sequences from automated sequencers), polyA tail
  - mask low complexity and repetitive regions

- Update
  - e.g. periodical import of new sequences from dbEST
Sequence Comparison and Clustering

- Use of DBMS-proprietary functions for text comparison and text search, e.g. *like*, *score* in Oracle
  - low cost, good performance
  - inflexible: Standard-Scoring-Function for natural language and not optimized for sequences, alignment not possible

- Use of existing programs (BLAST, FASTA, ...)
  - flexible parameterization, optimized for sequences
  - operate on flat files (input and output)

- Ideal: Extension of the DBMS with an API to BLAST, etc.
  - sequence comparison and search within DBMS

- Assessing cluster quality:
  - storing only similarity scores greater than a threshold (e.g. 95% in a defined number of nucleotides) in the database
  - minimal number of ESTs per cluster
  - minimal number of ESTs per tissue/organ
Analysis Possibilities (1)

- Support large amounts of sequence data: Broad spectrum of
  - ESTs
  - tissues, organs, etc.
  - genes

- Flexible generation of expression profiles
  - qualitative as well as quantitative expression profiles
  - more accurate expression profiles

- Integrated platform for sequence analysis
  - gene expression analysis
  - consensus sequences, sequence alignment, SNP, etc.
Interactive online analysis possible due to pre-computed sequence similarities and clusters

Support common analysis tasks:
- discovery of new candidate genes, which are specifically expressed in a particular tissue or organ

Support comparative analysis tasks:
- large-scale (at genome level) correlation analysis: identification of correlated expression patterns of
  - tissues with respect to genes
  - genes with respect to tissues
- discovery of candidate genes with similar functions, and tissues and cells with similar expression profiles

Support automatic approaches
Summary

- **In silico** differential display for expression analysis:
  - simple and effective approach for predicting (new) specifically expressed candidate genes
  - computation-intensive, time-consuming process due to large amounts of sequence data
  - currently performed largely manually

- Application of DBMS technologies
  - local DBMS-based management of all kinds of data, in particular sequence and annotation, sequence similarity and cluster
  - pre-computation of sequence similarities, clusters: performance for online analysis
  - improvement and automation of the previous *in silico* differential display approach
  - problems: integration of existing tools for sequence comparison, alignment and clustering

- Combination with other methods for expression analysis
  - Microarrays, RT-PCR, SAGE