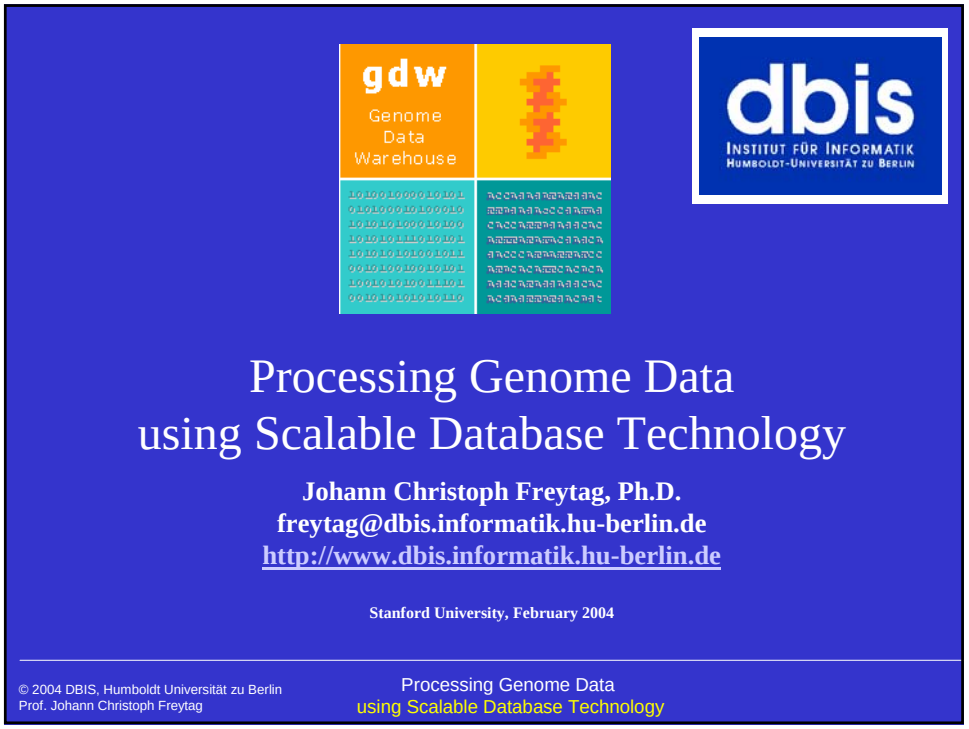
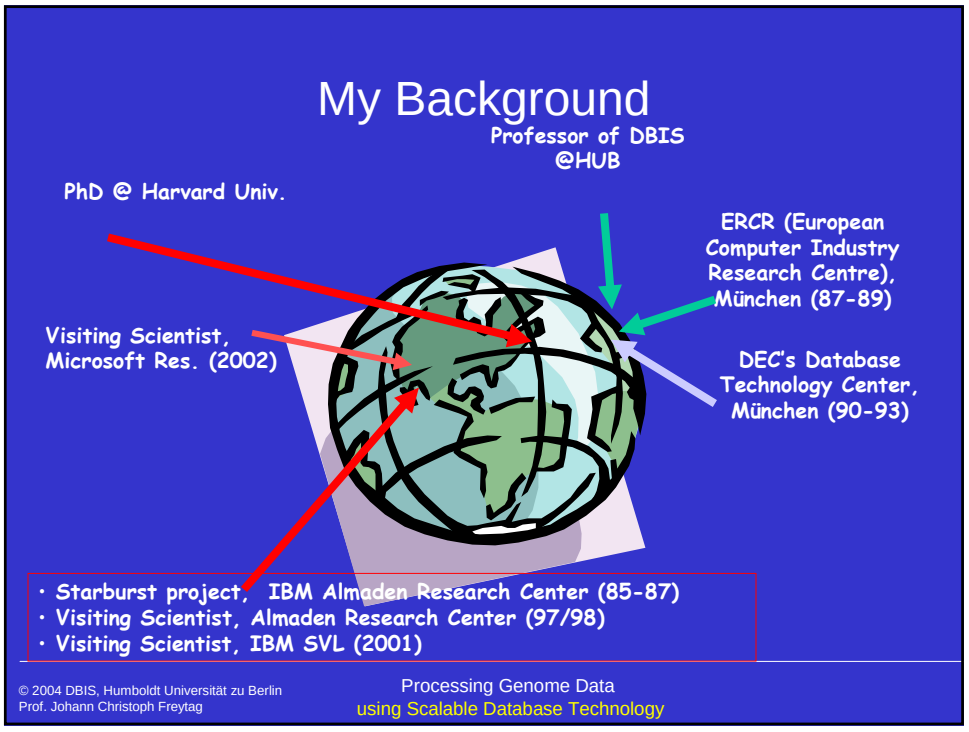
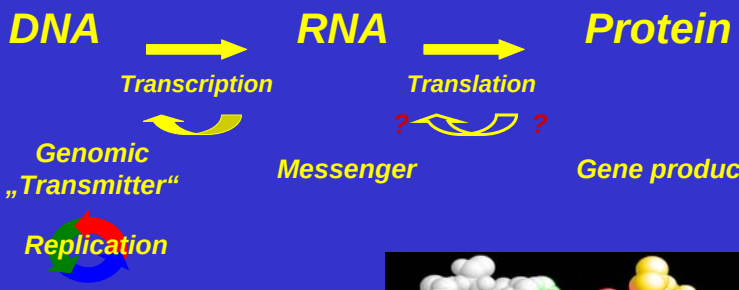
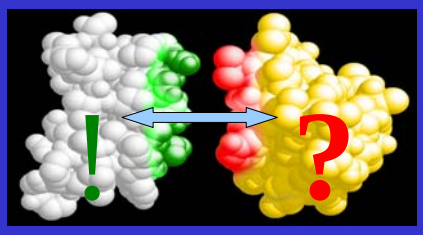
Prof. Johann-Christoph Freytag
DBIS, Humboldt-Universität zu Berlin





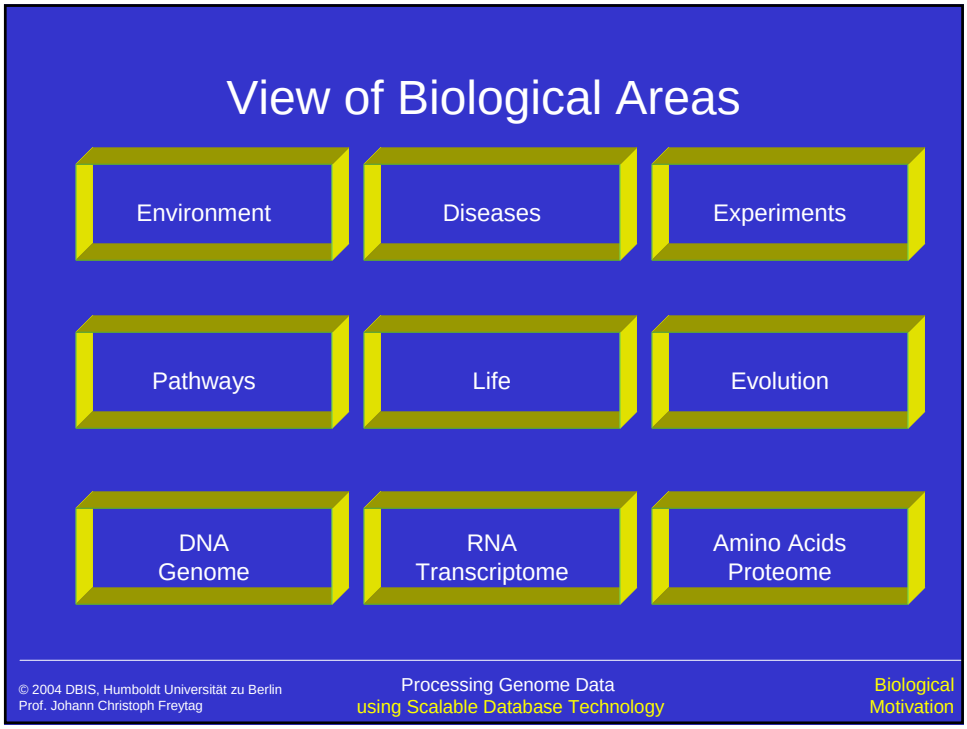
"What's the Meaning of Life"

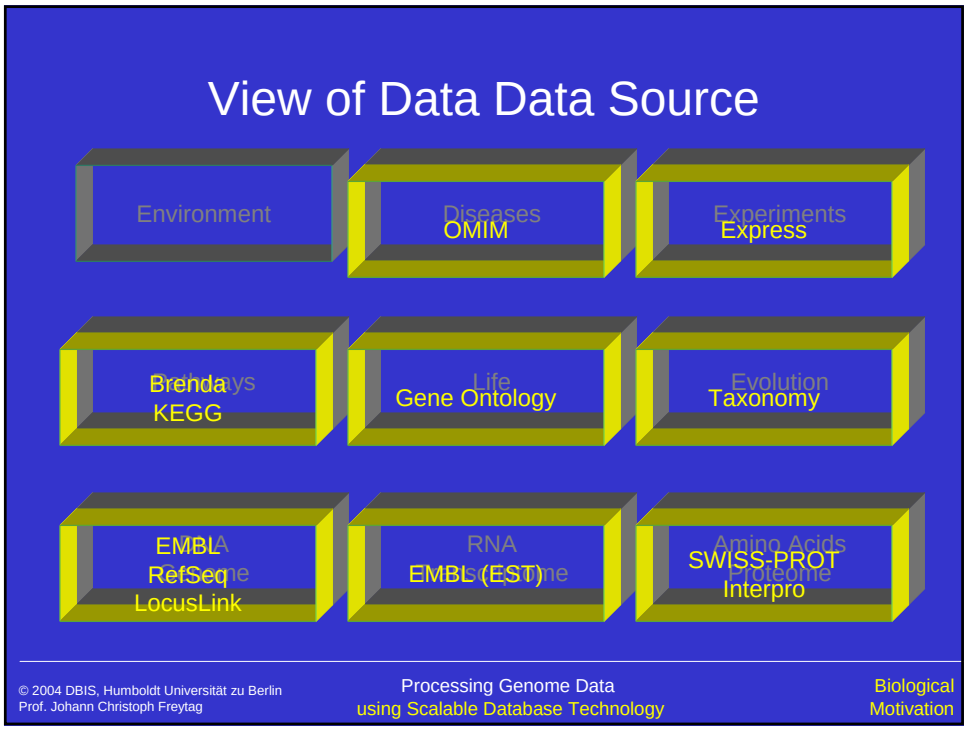




Overview

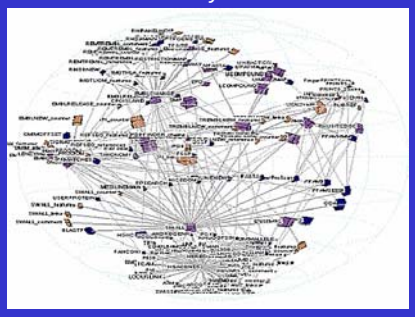
- (Biological) Motivation/Problems
- Using Database Technology
  - **Gene-EYe** Integration-Platform
  - Data Cleansing
  - BLAST-Integration into GDB
  - In-and-Out-the-Database: Using Workflow for "dry" Experiments
- Summary





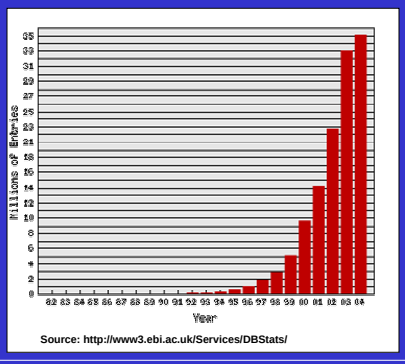
Complex Relationships

A graph depicting the relationships between 400+ biological data sources served by the EBI via SRS



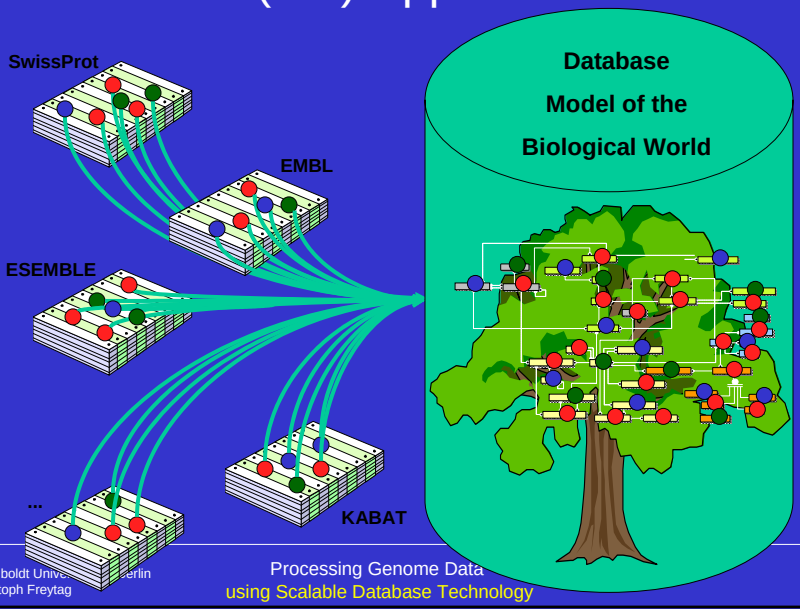
More than 400 Data Sources on the WEB

Database Growth of EMBL (# of records)



Processing Genome Data
using Scalable Database Technology

DBIS (our) Approach

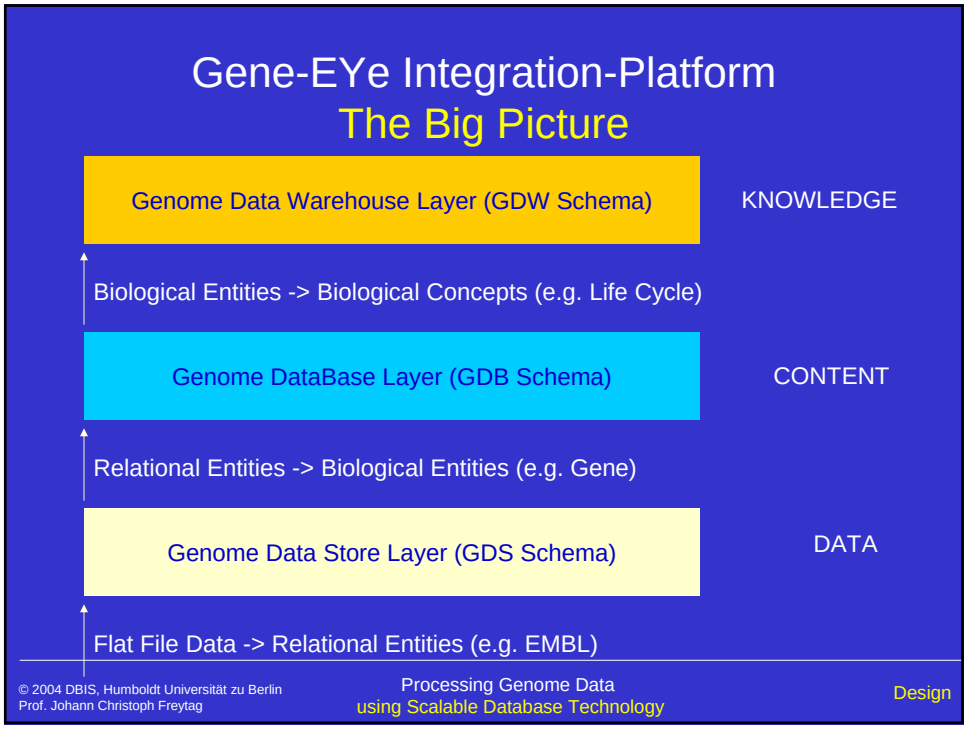


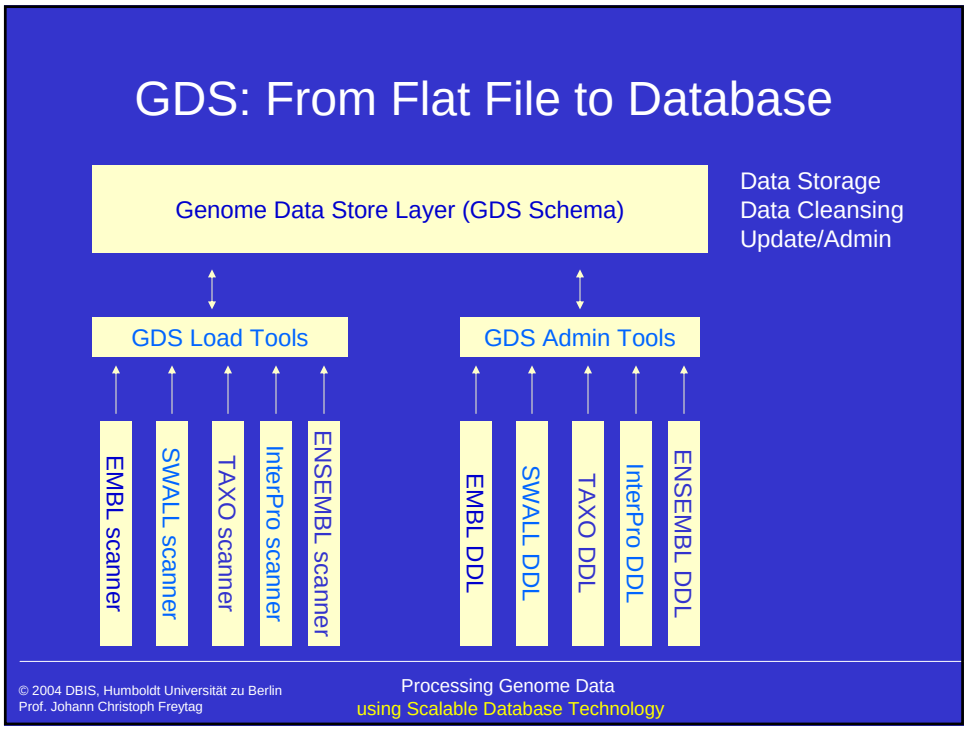
Processing Genome Data
using Scalable Database Technology

- (Biological) Motivation/Problems
- Using Scalable Database Technology
  - Gene-EYe Integration-Platform
  - Data Cleansing
  - BLAST-Integration into GDB
  - In-and-Out-the-Database: Using Workflow Concepts for “dry” Experiments
- Summary

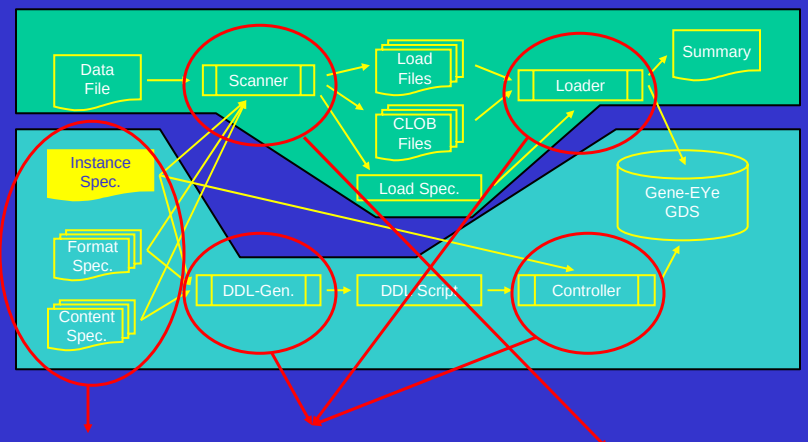
Gene-EYe Integration-Platform
Vision

- Provide mechanisms for
  - unified handling of different data sources
  - data source integration
  - change management
  - user defined data preparation
- Provide
  - relevant tools for sequence manipulation and retrieval
  - work flow support for operation and administration





The Data Import Pipeline - Revisited

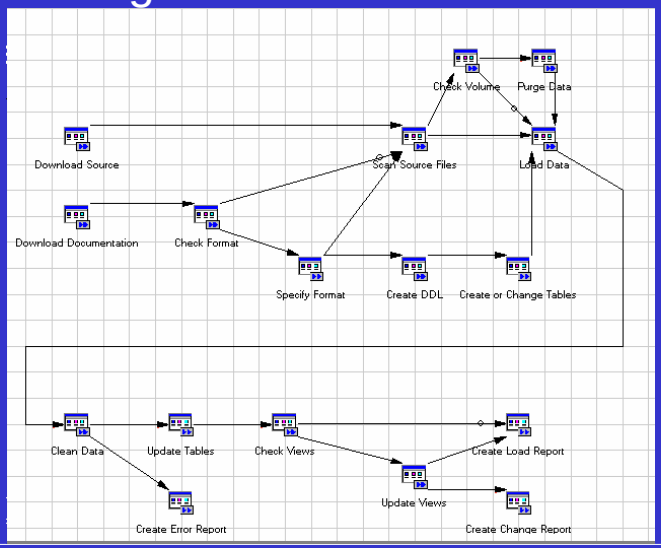


Phase 1: Property Files — Perl scripts — Hand crafted

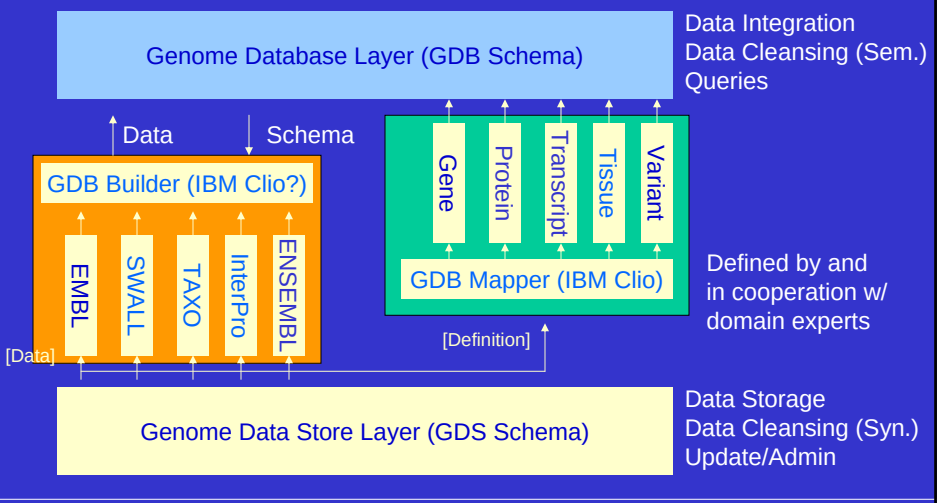
Phase 2: GEM[1] Repository — de.hui.dbis.geneeye.* (Java) — Autogenerate from Metadata

1: CWM compliant GeneEYe Metadata Repository

Modeling the Maintenance Process



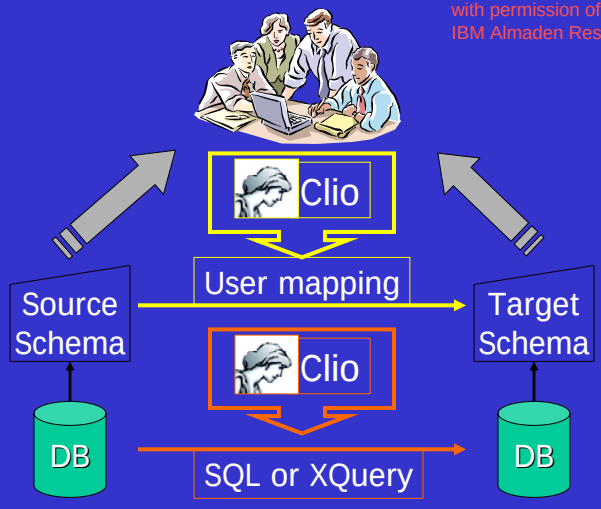
GDB-Layer: From Data to Biology



Processing Genome Data
using Scalable Database Technology

Schema Mapping with Clio

with permission of Dr. Felix Naumann
IBM Almaden Research Center



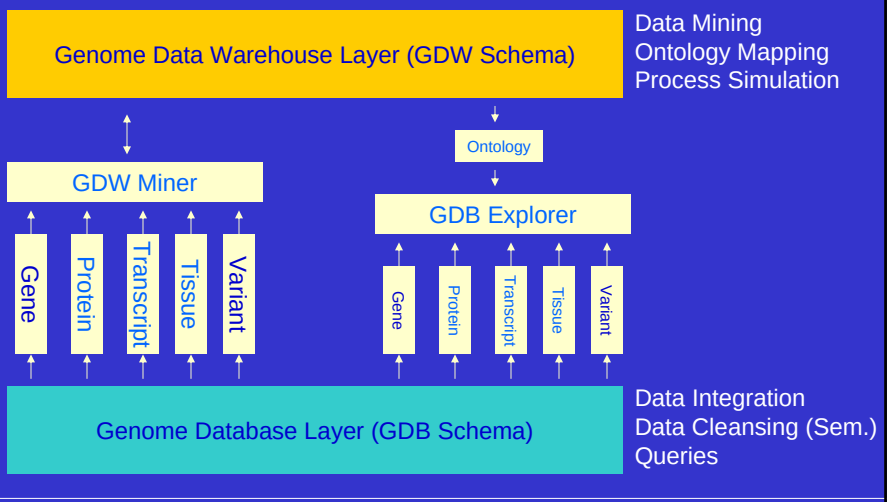
Processing Genome Data
using Scalable Database Technology

Clio Features

with permission of Dr. Felix Naumann
IBM Almaden Research Center

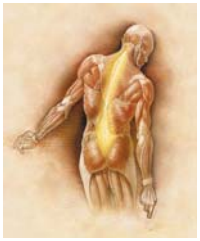
- Schema Viewer
  - Visual mapping between schema elements
- Attribute Matcher
  - Intelligent suggestions of likely mappings
- Data Viewer
  - Data examples for mapping queries
- Queries
  - SQL, XSLT, Xquery
  - Use and adhere to source and target schema constraints

GDW: Providing Facts for Research



- (Biological) Motivation/Problems
- Using Scalable Database Technology
  - Gene-EYe Integration-Platform
  - Data Cleansing
  - BLAST-Integration into GDB
  - In-and-Out-the-Database: Using Workflow Concepts for "dry" Experiments
- Summary

Errors in Genome Data



only **1.3 %** difference

The main difference are transcript copy numbers in the brain

A Sequence
agatcg
atagca
agcagc
0,23% – 2,58%

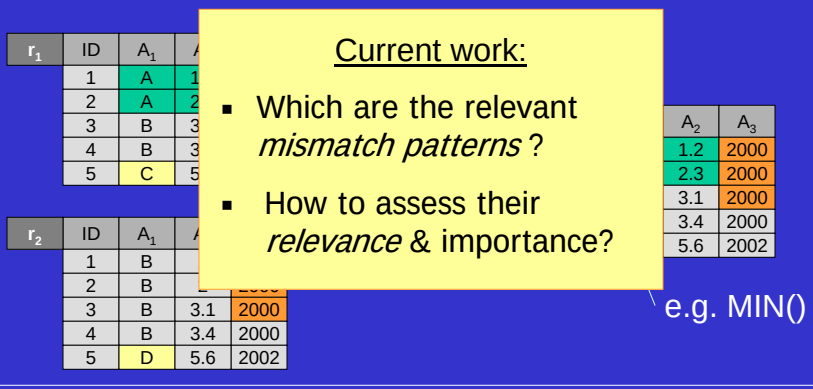
Annotation
5% - 30%

Sequence
QAERYDEMVESMKKVD
SVAYKNVIGARRASWY
EDKLKMIREYRQMVER

Annotation
CT AS IN
LING COM
DIRECTL
VED IN A
5% - 40%

Reliability-based Merging (cont.)

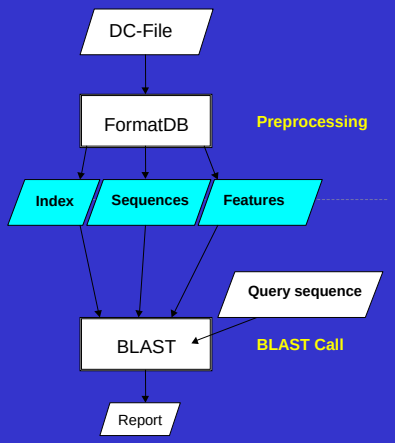
- Domain expert identifies reliable parts for merging
- Definition of a set of views for integration



- (Biological) Motivation/Problems
- Using Scalable Database Technology
  - Gene-EYe Integration-Platform
  - BLAST-Integration into GDB
  - In-and-Out-the-Database: Using Workflow Concepts for “dry” Experiments
- Summary

BLAST: General Introduction

- Algorithm/Package: Similarity Search
- Devloped by Altschul et al. (1990)
- Three Steps:
  1. Search for Word Pairs (Iseq, DSeq) of Length ***L*** on the Data Collection of Sequences above Threshhold T
  2. Expansion of each Word Pair until the Value V of their Alignment is Δ away from the local maximum
  3. Output of complete alignment (***H****igh-scoring* ***S****egment* ***P****air, HSP*), if Value(Alignment) > ***S***



Output: Powerset of Alignments

BLAST UDF Implementation

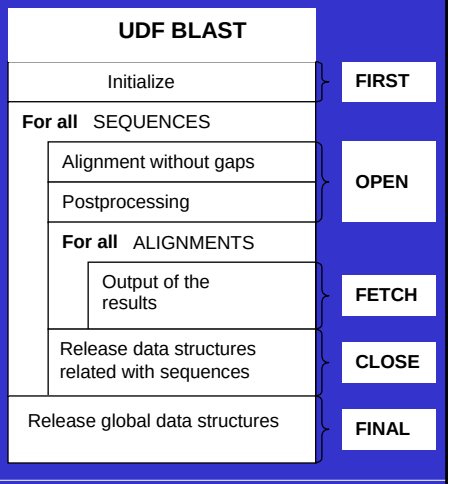
- Goal: Using BLAST in SQL-statements
- How?
  - BLAST-UDF implemented as Table Function
  - Use in SQL Query

```
SELECT *
FROM TABLE( BLAST(<Parameter>, <Query Sequence>,
                  <Comparison Sequence> ))
```

- Each call returns a set of alignments over Sequences in the Database

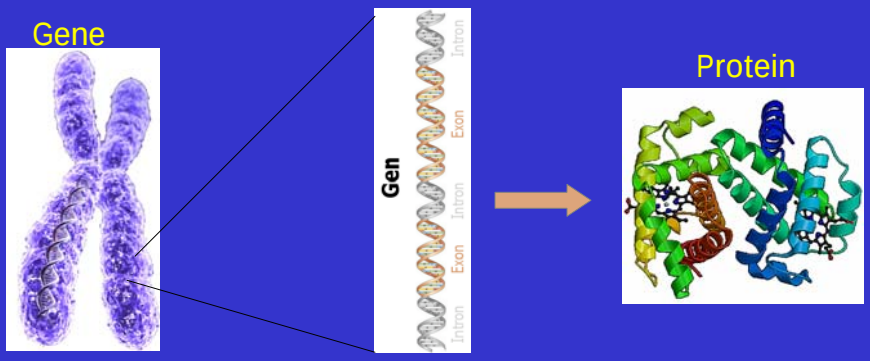
Structure of UDB Table Function

- Implementation:
  - Mapping of program into calling structure for table functions
  - Communiaction between the different calls via *scratchpad*
  - scratchpad: Storage area which remains intact and unchanged between UDF calls
    - Storage of data structures for different steps
    - especially for output from postprocessing: **`SeqAlign`**



- (Biological) Motivation/Problems
- Using Scalable Database Technology
  - Gene-EYe Integration-Platform
  - BLAST-Integration into GDB
  - In-and-Out-the-Database: Using Workflow Concepts for “dry” Experiments
- Summary

The Challenge: Exon Skipping

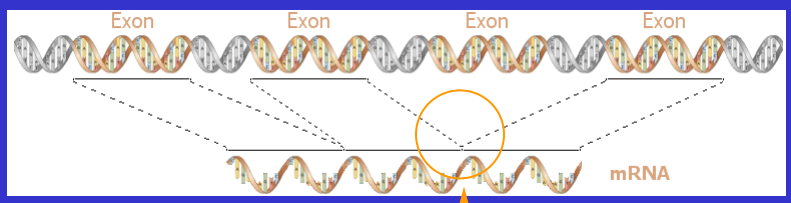


One Gen with 100 Exons ⇒ 2^{100} ~ 10^{30} Variations

n Exons within one Gene → linearly combined (splicing) → Used as Pattern for Protein Generation

Processing Genome Data
using Scalable Database Technology

Challenge: Exon Skipping

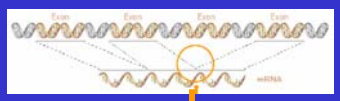


Do alternative fusion points new
funktional (i.e. biologigacl meaningful) „patterns“?

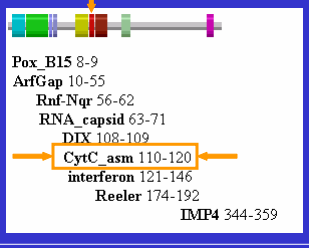
Processing Genome Data
using Scalable Database Technology

Functional Genomics: Gain of New Insight

First Horizon: Simple Exon Skipping

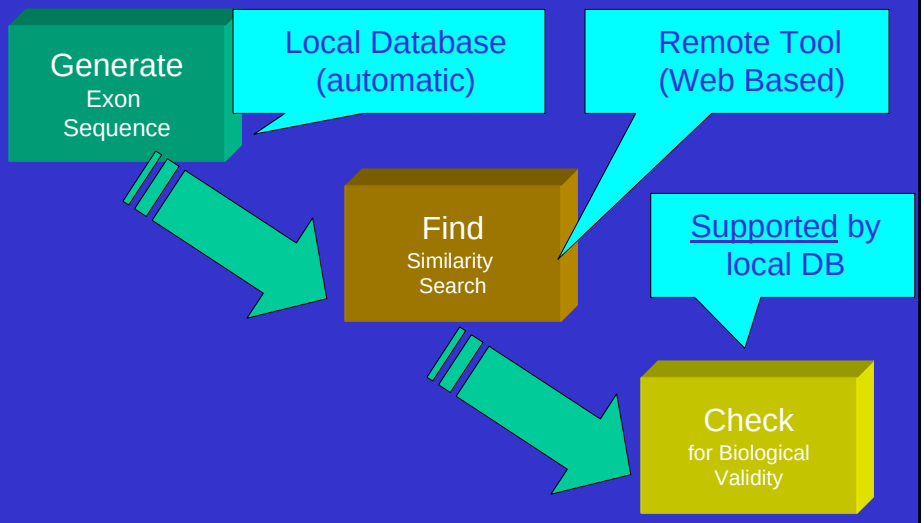


110110000111111

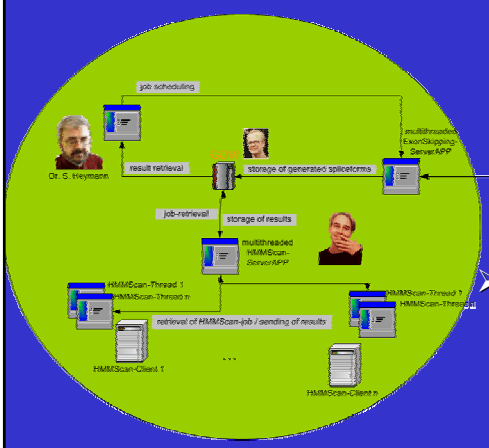


New Functionality!

Flow of Processing Steps

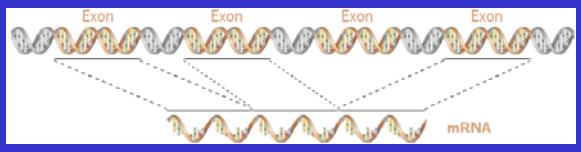


Implementation



- Some facts…
  - 60 days 100% load
    - One splice form per minute
    - So far: ca. 90.000 splice forms
  - First biolog. meaningful results

Cooperations

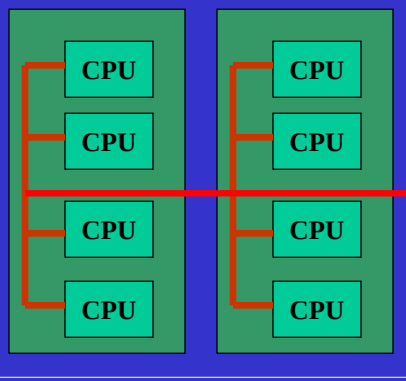


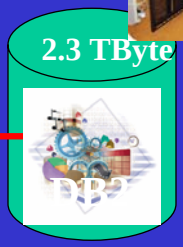
- Cooperation with
  - Univ. of Jena (Rolf Backofen)
  - Berlin Center of Bioinformatics (BCB)
    - Charite, FU, Max-Planck-Institut (M. Vingron)
  - Industry: IBM, small companies, …

Patrick Chappatte, Switzerland

Database Environment

IBM p-Server
(sponsored by IBM)





Summary

- Lesson learnt
  - Highly Dynamic Environment
    - Data: changes frequently
    - User: changes frequently
  - Provide a framework for
    - Date integration
    - Data processing
    - Data changes
    - Data dependencies…..
    - Meta data management
- Future Work
  - Query processing
    - Include domain knowledge
  - Data cleansing
  - Set of UDFs for biological data processing
  - Visualization of Data