Part 1:
Data & Process Integration
Gene-EYe Integration-Platform

The Big Picture

**Genome Data Warehouse Layer (GDW Schema)**

Biological Entities -> Biological Concepts (e.g. Life Cycle)

**Genome DataBase Layer (GDB Schema)**

Relational Entities -> Biological Entities (e.g. Gene)

**Genome Data Store Layer (GDS Schema)**

Flat File Data -> Relational Entities (e.g. EMBL)

Data Integration – Information Flow

XML Data
dtd, xsd

Relational Data
schema

Flat file Data
manual

Scanner

Mapping

Analysis

Data Store

meta data Repository

DDL

scanner generator

data model

format

structure
### Meta Data Editor as an Eclipse Plugin

The editor is autogenerated from an UML-Model edited with Omondo UML in an Eclipse 2.1 IDE used in conjunction with the Eclipse Modelling Framework (EMF), a Model Driven Architecture plugin.

### Integration of Data Cleansing

**Problem**
- Errors at every activity during ETL:
  - extract
  - transform
  - loading
- Errors at any level of granularity:
  - Column, tuple, table, schema, database
- Genome data sources:
  - No final source of truth available
  - Must cope with ambiguity

**Proposed solution**
- Integrate constraints in meta data
  - OCL statements (expressive, but complicated evaluation necessary)
  - Fixed set of error classes (restricted expressiveness, but easy to implement)
- Formalize operational semantic of Data Cleansing functions
- Define a plugin API for ETL work
Conclusion Part I

• First Steps:
  – Well on our way: working prototyp
• BUT….
• more ahead
  – Better modelling (flexibility!) & cleansing
  – Improved cleansing pipeline

Part II:
Alternative Splicing
Commonly Accepted View = Predefined Exons/Introns

Alternative Splicing: ~ 60 % of Human Genes

→ Protein Synthesis Prescription 1

"Skipping" → Protein Synthesis Prescription 2

Multiple Exon Skipping
Facultative Promotors
Cryptic Exons
Mutually Exclusive Exons
Splice Site Attenuation
Intron Retention
Resplicing
......

Other Active ORFs

in any (?) Combination
Determinism in Exon/Intron Assignment to DNA

questionable

A Revised Understanding of Information
Recruitment for Protein Synthesis is
advisable, if not imperative

Chop DNA into $n$ Facultatively Exonic Fragments

Simulate Systematic Skipping

Theoretically $2^n$ Combinatorial Variants

$\rightarrow$ Known Ones Easy To Identify
$\rightarrow$ Dismiss Improper Concatenates
$\rightarrow$ Scan for Regulatory and Functional Signals
Theoretically $2^n$ Combinatorial Variants

- Known Spliceforms Easy To Identify
- Dismiss Improper Concatenates
- Scan for Regulatory & Functional Signals
- Evaluate Splice Junctions

Instead of "Generate & Test"

Re-Splicing

PfamXYZ

Linear Runtime Strategy

Dynamic Programming

Optimization Task

- Viterbi-Algorithm with Jumps
  & Translation Control
- Runtime $\sim nk$
Example

- Working Hypotheses concerning Molecular Etiology
- Evidencing
  - Directly (PCR)
  - EST
  - MS/MS-Data (non-public)

Conclusions Part II

→ Refrain from Static Exon/Intron Distribution
→ Pay Attention to Signals Occuring while Distant Information Fragments are being Concatenated
→ Look for Evidence in Independent Data Sets

Thank you
Questions??