An Intelligent Agents Architecture for DNA-microarray Data Integration

Mauro Angeletti, Rosario Culmone and Emanuela Merelli

Università di Camerino

NETTAB
Genova, 17 May 2001
Outline

• Motivations
• Main choices in the system - IMAB
• System’s architecture
• Mobile agent model
• Preliminary results
• On-going work
Motivations

Electronic diagnostic tool = diagnosis
- Significant amount of data (ORFs, Experiments, ...) disseminated and duplicated in a myriad of different dbs and repositories
- Different access modes
- Several actors (researchers, doctors, ...)
- ...

Our-goal:
- define a general platform IMAB (Intelligent Mobile Agent platform for Biological data) to support “genetic data analysis”
- define a declarative language IMABL to specify agents
What is an agent?

An agent is a program capable of acting autonomously in order to accomplished tasks on behalf of its user.

- There are several dimensions to classifying existing software agents:
  - mobility, ability, reactivity ...

- In our context, an agent is a program that can move between nodes by preserving the status and managing information (i.e. knowledge) useful to perform its mission.
Main choices
Choice 1: all XML

- It’s a W3C standard for data transfer
- Many tools are or will be available
- It’s easy to use
- ...

- Bio data DTDs are under definition
Choice 2: all declarative

- Fast deployment
- Easy administration
  - e.g., dynamic change of rules, experiment categories ...
- Automatic verification
  - e.g., correctness of protocols
  - e.g., fairness of an extraction system
- etc.
Choice 3: all autonomous

- Each node is an autonomous elaborative unit
- Each node can have its operating system
- Each node can have a DBMS and/or XML repository
Choice 4: all JVM

- Each node must have a JVM
How it works ...

Single agent

The People

The agent platform

The Bio Data
How it works ...
System’s architecture
System’s architecture

Agent ... Agent
Agent platform
Operating system
Hardware
Agent model
Agent model

• Basic features
  λ communication,
  λ memory,
  λ cloning,
  λ moving

  λ knowledge mng

  • Application features
    λ Mission
    λ Working Tools
    λ Knowledge

  ability to exchange information
  with human customer or communicate
  and synchronize with others agents
  ability to instance a copy of itself
  ability to move the code from one site to another
  ability to believe, decide and enrich
Computational Analysis of Biological Data
Agent’s Mission

- Bio data search
- Bio data integration
- Bio data clustering
- Bio data extraction
- Bio pattern recognition
- Bio knowledge discovering
- Bio data prediction
- ...

Agent’s Working tools

- **Optimization**
  - Combinatorial algorithms
  - Heuristics algorithms
  - CLP: Constraints Logic Programming
  - ...

- **Classification**
  - Neural nets
  - Kohonen self-organizing map
  - ...

- **Knowledge Discovering**
  - Data mining
  - Multidimensional analysis
  - ...
Agent’s Knowledge

- Basic knowledge (well-established)
  \[ \lambda \text{dtd} \]

- Extended knowledge
  \[ \lambda \text{dtd} \rightarrow \text{dtd}', \text{new knowledge} \]

- Local knowledge
  \[ \lambda \text{any, not in XML format} \]
Application: analysis of gene expression
Data: Biological data related to an “organism”

- Experiment = set of inter-related hybridisations
- Hybridisation = collection of experimental data (spots).
  One spot for each ORF (Open Reading Frame)
- ORF = minimum bio data unit

Mission: analysis of hybridisation data

- Clustering by experiments
- Clustering by ORF

Tool:

- Kohonen self-organizing map

Basic Knowledge: MAML DTD

Extended Knowledge: Kohonen map
Microarray experiments

“...normally an experiment should include a set of hybridisations which are inter-related and performed in a limited period of time.”

MIAME (Minimal Information About Microarray Experiments) document by MGED
(Microarray Gene Expression Database group, UK)
Hybridisation

Each hybridisation is constituted by a collection of experimental data (spots) usually one spot for each ORF (Open Reading Frame).

- The intensity of each spot quantifies the expression of the related ORF under the chosen experimental conditions.
MicroArray Markup Languages

- MAML proposed by the European Molecular Biology Lab (EMBL) and the European Bioinformatic Institute (EBI) and recently submitted to OMG (Object Management Group)
- GEML (proposed by a public-private community, the GEML community [21]).
- BSML (Bioinformatic Sequence Markup Language) proposed by Visualgenomic, Inc. USA [20].
MAML DTD structure

- 1. Experimental design: the set of the hybridization experiments as a whole;
- 2. Array design: each array used and each element (spot) on the array;
- 3. Samples: samples used, the extract preparation and labeling;
- 4. Hybridizations: procedures and parameters;
- 5. Measurements: images, quantitation, specifications;
- 6. Controls: types, values, specifications.
Major goals pursued during the analysis of hybridization data

- data mining
- model-based/model-free
- functional classification
- clustering
Clustering

• Clustering using experiments euc. distance
  λ (ORF “guilty-by-association”)
  λ Present use: discover gene function

• Clustering using ORFs euc. distance
  λ (transcriptional fingerprint)
  λ Future use: diagnostic tool
Kohonen Algorithm

1. We define with $w_{ij}(t)$ the weights between the input neuron $ith$ and the $jth$ neuron in the map at time $t$. The weights initial values are randomly assigned in the $[0,1]$ range.

2. Given an input $x_0(t), x_1(t), ..x_n(t)$, where $x_i(t)$ is the $ith$ input

3. Calculate the distance $d_i$ between input $i$ and each output neuron $j$

   $d_j^2 = \sum (x_i(t)-w_{ij}(t))^2$

4. Select neuron with minimum distance, $j^*$

5. Modify the weights of the input neuron $i$ and $j^*$ and its neighbours $N_i(j^*)$

   $w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t)-w_{ij}(t))$ for each $j$ in $N_i(j^*)$ and $0 \leq i \leq n$

   $\eta(t)$ is a gain function $0 \leq \eta(t) \leq 1$

6. Cycle from step 2
Preliminary results
Database source

- 147 distinct hybridisation experiments
- 6053 ORFs

Agent’s Platform

- *Macondo* [Ciancarini97]
- *MJada* for agents coordination and synchronization
Kohonen algorithm

- Completely implemented in Java
- 9 clusters
- The algorithm converges in 10000 cycles
The output
On-going work
We are working on ...

- Knowledge representation and mng
  - Model
  - Manipulation language

- IMABL
  - Declarative language for agent definition