

Applications of Artificial Intelligence

Ralf Hofestädt

AG Bioinformatics / Medical Informatics

<http://cweb.uni-bielefeld.de/agbi/home/index.cw>

1. Bioinformatics
2. Motivation AI
3. Problem solver
4. Expert Systems (Example RAMEDIS)
5. Data Warehouse (Example PathAligner)

Bioinformatics

Computer Science and Biology

*-Modeling, Simulation, Visualization, Animation,
Artificial Intelligence, ...*

Biology and Computer Science

*-Finite Automata, Cellular Automata, L-System ...
-DNA-Computing, Genetic Algorithm, Neuronal
Networks, ...*

Hofestädt, R. (Hrsg.):

Bioinformatik 2000 - Forschungsführer Informatik in den Biowissenschaften.

BIOCOM, Berlin 1999

Bioinformatics - Germany

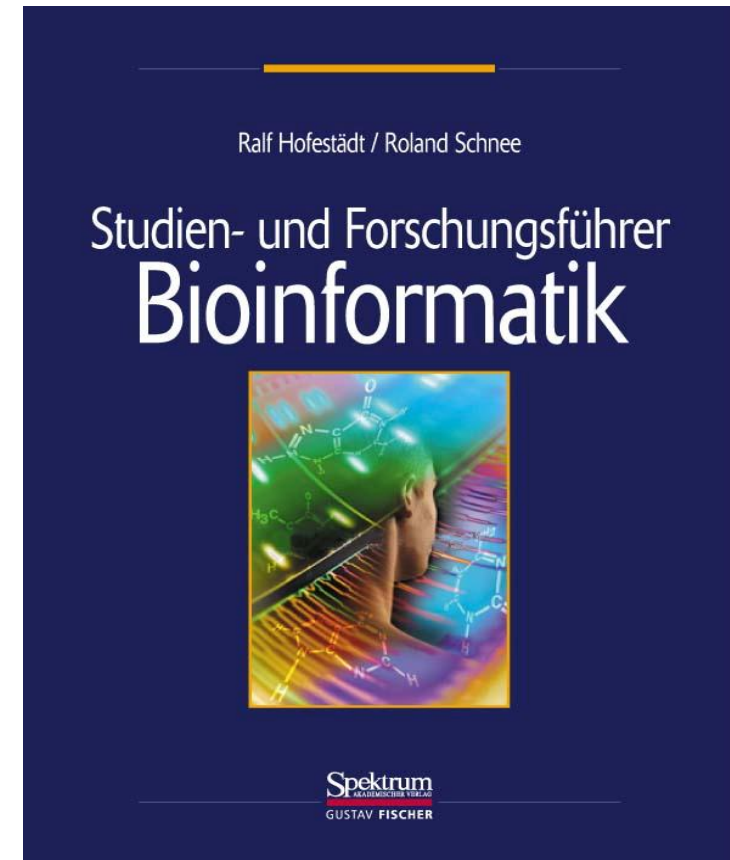
Curricula

Diploma

Bachelore

Master

Phd Courses



www.ff.imbio.de

Hofestädt, R. and Schnee R.(Hrsg.)

Studien- und Forschungsführer Bioinformatik.

Spektrum Akademischer Verlag, Heidelberg, 2002

AG Bio-/ Medizinische-Informatik

Bioinformatics - Germany

History - Highlights

- 1992 Workgroup Bioinformatics
German Society of Computer Science
- 1993 National Conference, Bonn
- 1996 German Conference on Bioinformatics
GCB, Leipzig
- 1999 DFG support for Bioinformatic Centers
- 2000 BMBF support for Bioinformatic Centers



www.ff.imbio.de

Hofestädt, R. and Schnee R.(Hrsg.)

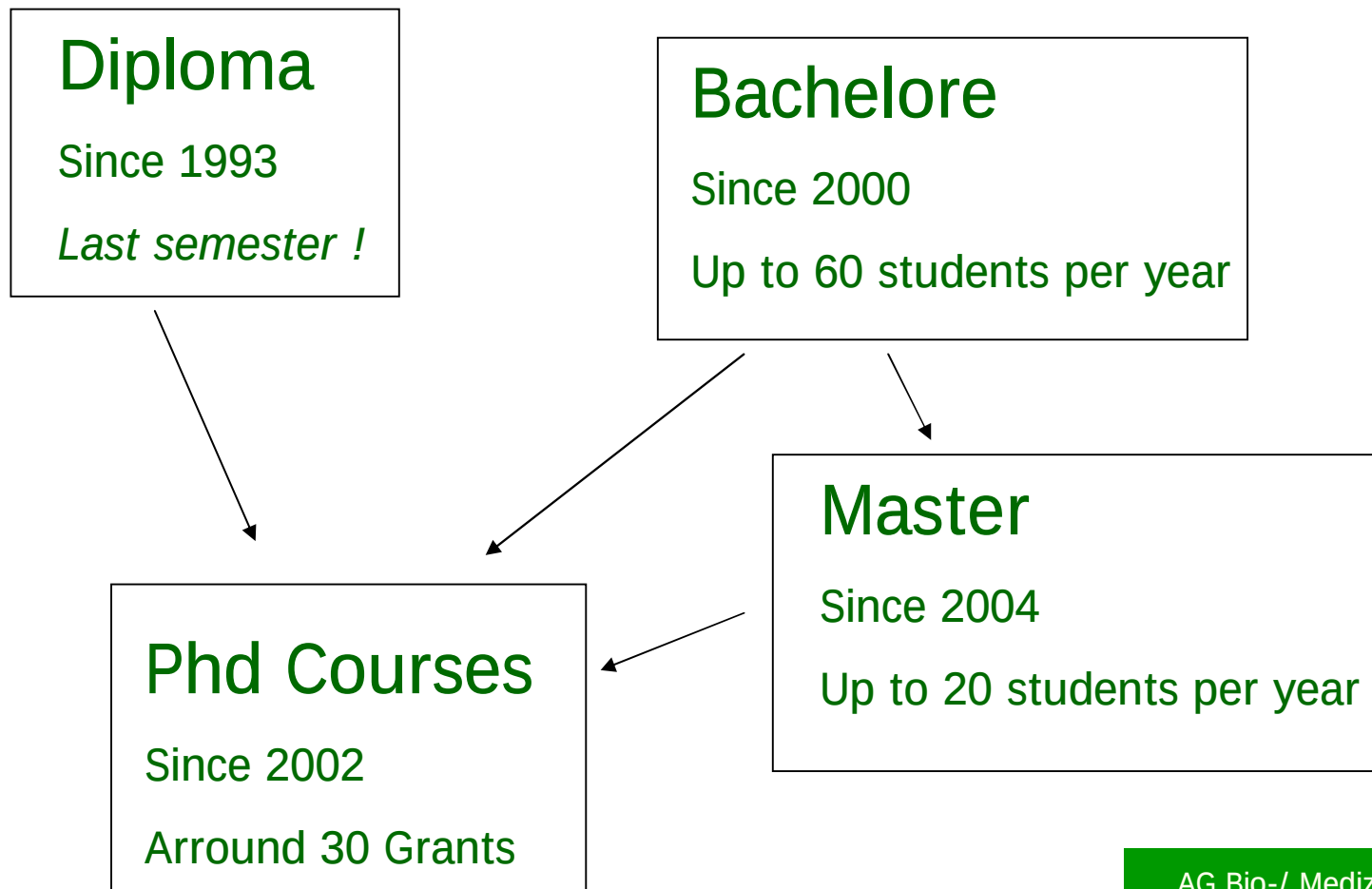
Studien- und Forschungsführer Bioinformatik.

Spektrum Akademischer Verlag, Heidelberg, 2002

AG Bio-/ Medizinische-Informatik

Bioinformatics - Bielefeld

Curricula



Bioinformatics - Bielefeld

Topics – Technical Faculty

Prof. Giegerich: Practical Computer Science and Sequence Analysis

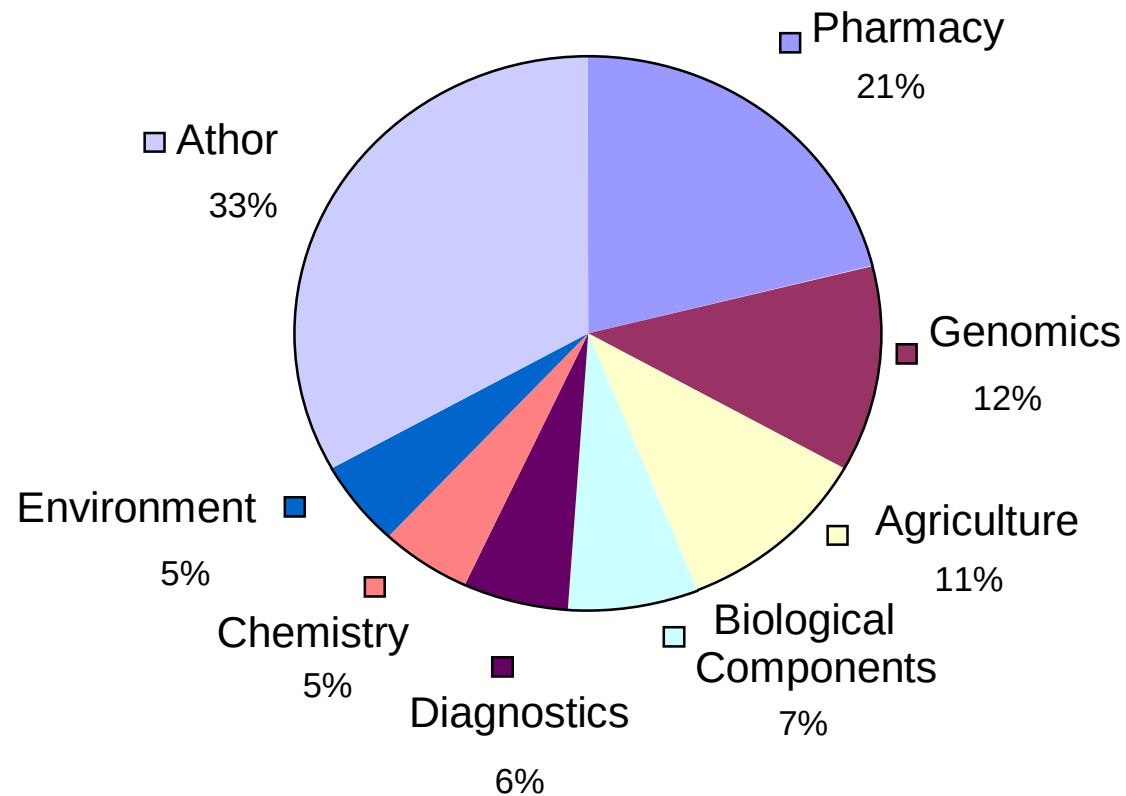
Prof. Hofestädt: Metabolic Networks and Medical Informatics

Prof. Stoye: Genominformatics

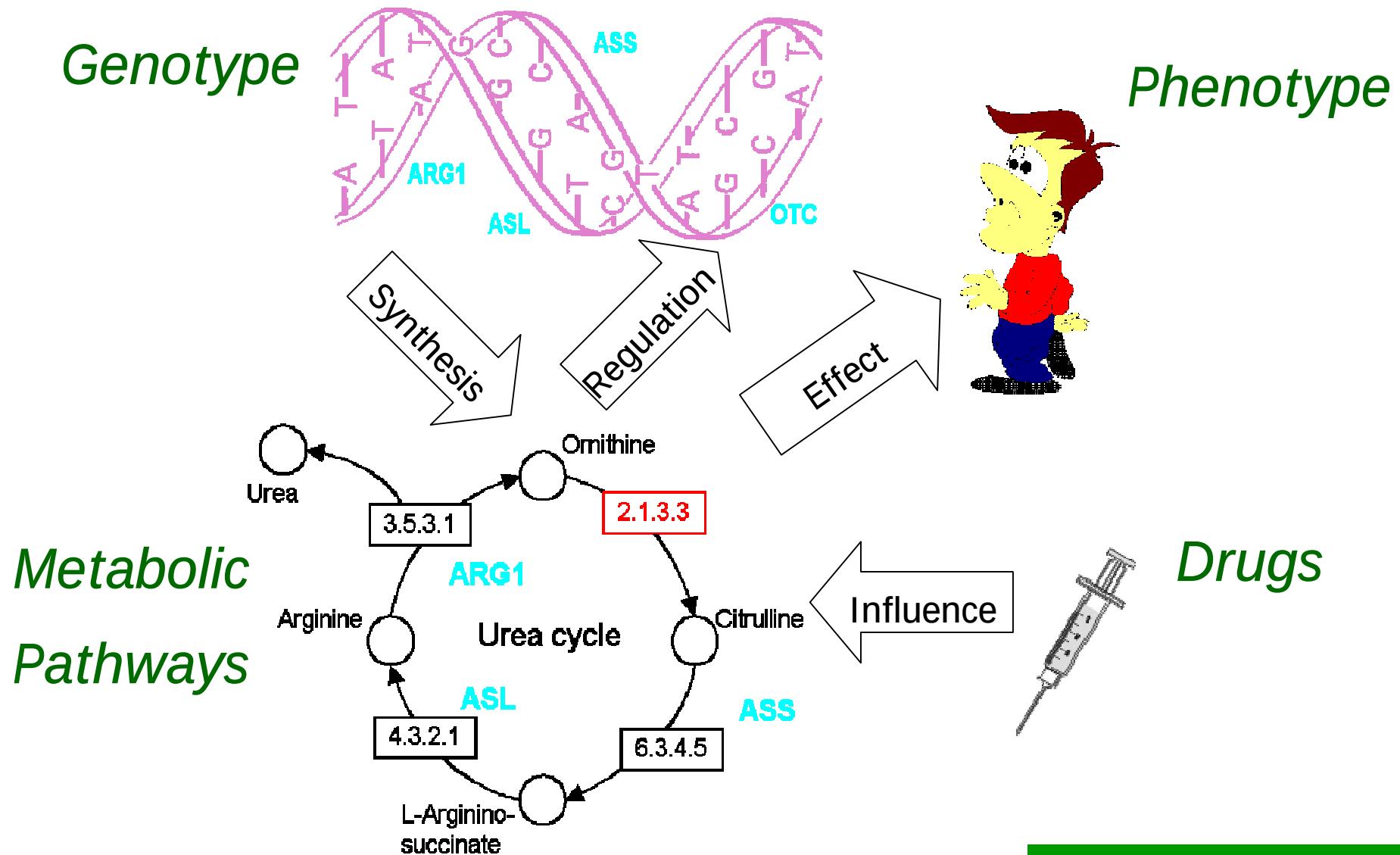
Prof. Naettkemper: Pattern Recognition

Application Area

Utilization of Bioinformatics

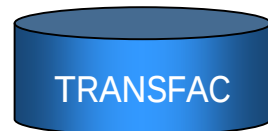
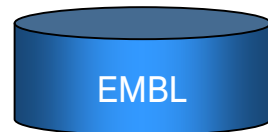


Genotype and Phenotype



Specific Database Systems

Genotype

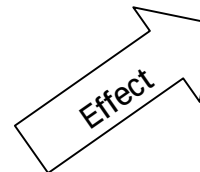
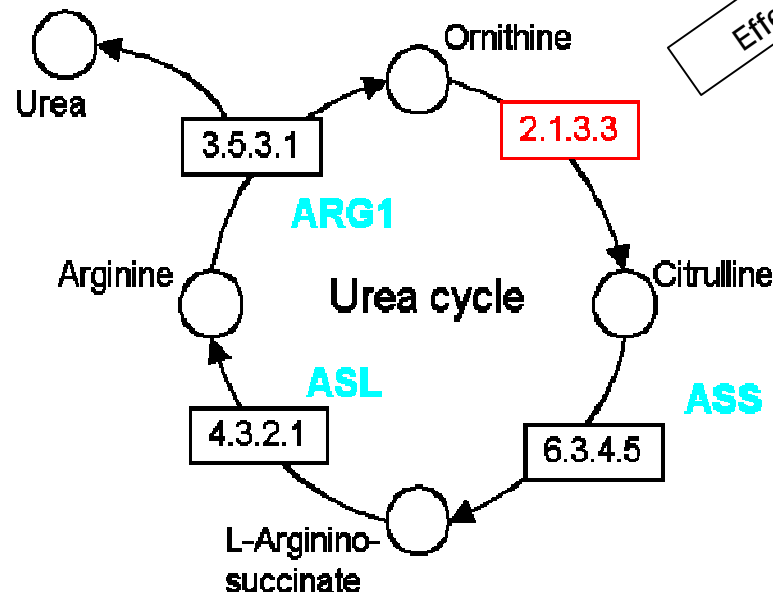
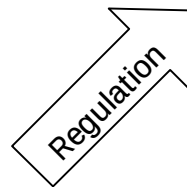
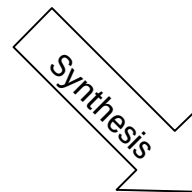
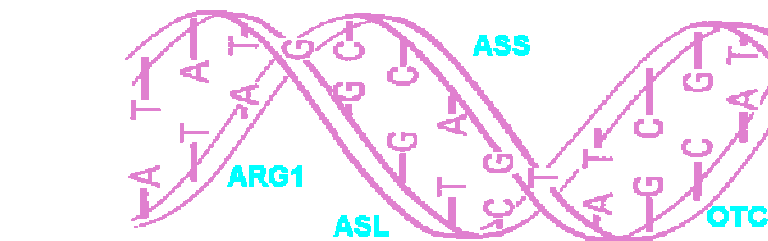


...

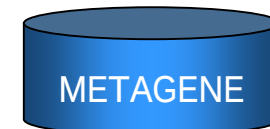
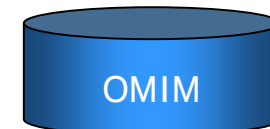
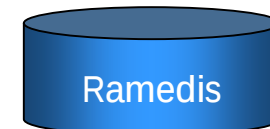
Metabolic Pathways



...

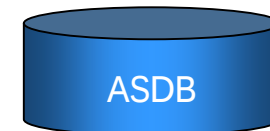


Phenotype



...

Drugs



...

Integrative Bioinformatics

Views to (complex) data

- Visualization*
- Statistic*
- Analysis Algorithms ...*

Data Mining / Information Fusion

- Integration of database systems and analysis tools*
- ARTIFICIAL INTELLIGENCE**

International Journal of Integrative Bioinformatics

<http://journal.imbio.de>

Algorithm

Algorithm ?

A system where each step is:

finite, definite, deterministic, traceable, input and output

Question:

Regarding a cooking recipe – is this a algorithm ?

Algorithm

Muhammed idn Musa abu Djafar al-Choresmi,

or

Al Khwarizmi or Al-Khowarizmi,

born ca. 780 south-east of the Aral-See, in Choresmien in
Usbekistan, died 850.

825 he wrote the book:

Exercises for trades and executor of a will

Arabic title: “**Kital al jabr w'al muqabata**”

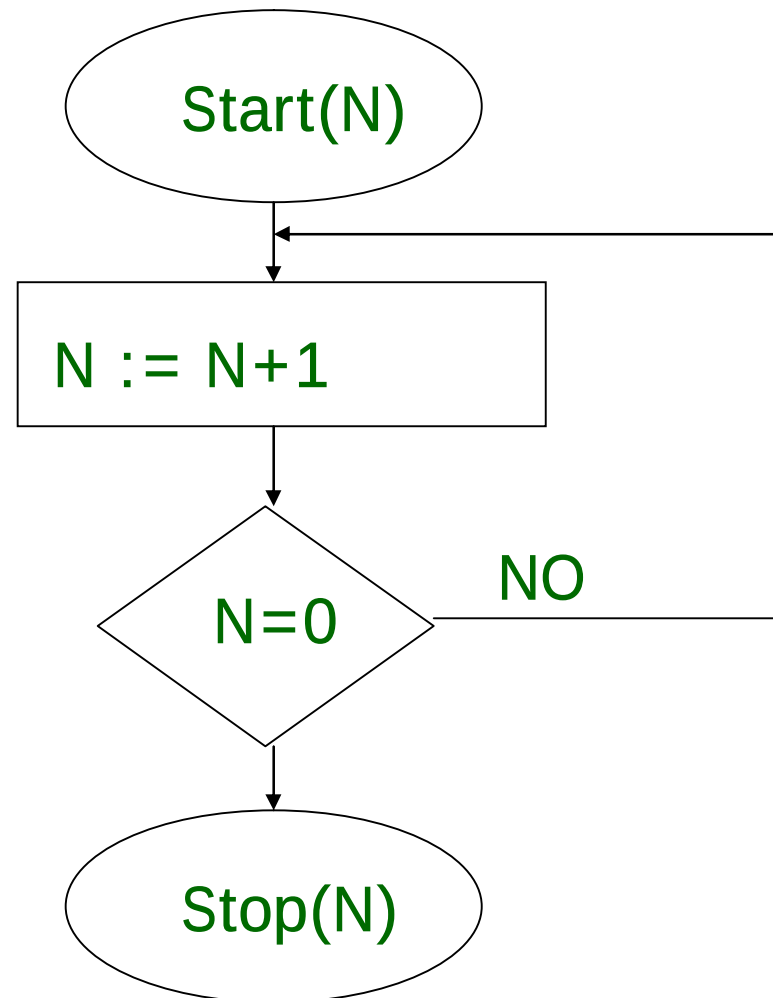
⇒ Algebra

Latin translation “**Liber Algorithmi**”.

Algorithm

Flow chart

Termination ?



$N \in \mathbb{N}$

Algorithm ?

Algorithm

Find largest element

The Algorithm has to find the least index i , so that L_i will be the largest element in L .

Specification of the problem:

Input: n from $\mathbb{N}$, L from $\mathbb{R}^n$

Function: $f: \mathbb{R}^n \rightarrow \{1, 2, \dots, n\}$

Output: $f(L) = i$ with $i = \min\{g \mid L_g = \max_{1 \leq k \leq n} L_k\}$

Algorithm

Algorithm FMAX: $i \leftarrow \text{FMAX}(L, n)$

Find maximum from Liste L with length $n > 1$. i is the first position of the list, where we can find the maximum for the first time.

F1: [choose favorit]

$i := 1$

F2: [search]

for all L_j ($j=i+1, \dots, n$)
execute step 3

F3: [compare]

if $L_j > L_i$, then $i := j$
now we have
 $L_i = \max_{1 < k < j} L_k$
(go back to step F2)

Algorithm

Definition: Computability

A problem is called computability iff we can construct a algorithm, which will solve the problem.

Question ?

Is each problem computability ?

Algorithm

Leibniz'scher Gedanke (Leibniz ideas) - 1694

For each problem we can find a calculator
which will solve the problem !

Gödel could show during the 20ties of the last century:

PK1 (predicate logic of the first level) is not computable !

We have problems which are not computable !

Artificial Intelligence

One more important problem !

COMPLEXITY (time and/or space)

Most of the problems, which are computable can not be solved with the algorithms and computers of today.

Therefore, we need other methods:

Artificial Intelligence

Artificial Intelligence

E. Rich: “Artificial Intelligence is the study of how to make computers do thinks at which, at the moment, people are better.”

The „Turing Test“ was introduced by Alan Turing (1912-1954)

[Computing Machinery and Intelligence \(Mind, Vol. 59, No. 236, pp. 433-460\)](#)

which he so boldly began by the following sentence:

I propose to consider the question "Can machines think ?“

Artificial Intelligence

When talking about the „Turing Test“ today:

We are connected to one person and one machine via a terminal, therefore we can't see our counterparts. Our task is to find out which of the two candidates is the machine, and which is the human only by asking them questions. If the machine can "fool" us, it is intelligent.

Artificial Intelligence

Research topics:

Pattern recognition, specific algorithms (genetic algorithm, neuronal networks ...), fuzzy systems, data mining, data warehouse, expert systems, case based reasoning, causal probabilistic networks, machine learning, knowledge representation, theorem proofer, robotics ...

Artificial Intelligence

Motivation 1:

Problem: time complexity = exponential and higher !

Most interesting questions (for example simulation of metabolic pathways, multiples alignment and so on).

What to do ?

Use Artificial Intelligence methods and try to solve the problem approximative.

Artificial Intelligence

Motivation 2:

Problem: uncertainty knowledge or rudimentary knowledge.

Most interesting questions (for example prediction of metabolic pathways, diagnostic processes).

What to do ?

Use Artificial Intelligence methods and try to solve the problem approximatively.

Artificial Intelligence

Motivation 3:

Problem: distributed and mass data (exponential growing of molecular data).

Most interesting questions (integration of molecular data and analysis).

What to do ?

Use “Data Warehouses”

Artificial Intelligence

Let us focus to the

- a) classical domain – automatic theorem solver
- b) most important tool – Expert Systems (ES)
- c) Application of ES for Bioinformatics RAMEDIS
- d) Data Warehouse (Pathaligner)

Artificial Intelligence

Theorem solver – idea

Definition of
axioms,
rules and
inference mechanism.

Problem: Exponential complexity of these methods !

Artificial Intelligence

Definition proof system

A proof system is a quadruple

	$S = (\Sigma, L, Ax, R)$ with
Σ	finite alphabet.
$L \subseteq \Sigma^*$	language of S .
$Ax \subseteq L$	axiom system of S .
$R = \{R_i : i \in I\}$	finite set of derivation rules.

Artificial Intelligence

Proof system for the logical calculus:

1) For a proof system $S = (Z, L, Ax, R)$

$$L := AUSD' \subseteq AUSD$$

is the set of logical expressions using only
the $\neg$ (negation) and $\rightarrow$ implication.

2) $Ax = \{Ax_1, Ax_2, Ax_3\}$ with

$$Ax_1: (p \rightarrow (q \rightarrow p))$$

$$Ax_2: ((p \rightarrow (q \rightarrow r)) \rightarrow ((p \rightarrow q) \rightarrow (p \rightarrow r)))$$

$$Ax_3: ((\neg p \rightarrow \neg q) \rightarrow ((\neg p \rightarrow q) \rightarrow p)) \quad (\text{Tautologien}).$$

3) $R = \{MP, SUB(v, b) : v \in V, b \in AUSD'\}$

With two rules:

MP Modus Ponens: $a, a \rightarrow b \rightarrow_{mp} b$

SUB (Substitution): $a \rightarrow SUB(a)$

Artificial Intelligence

Computer implementations need:

One simple rule !

Thats the so called **Resolution-rule** :

NOTATION

$\{a, c\}$

clause !

$\{b, \neg c\}$

a = literal

$\{a, b\}$

Resolution

IF one clause pair represents complementair literales

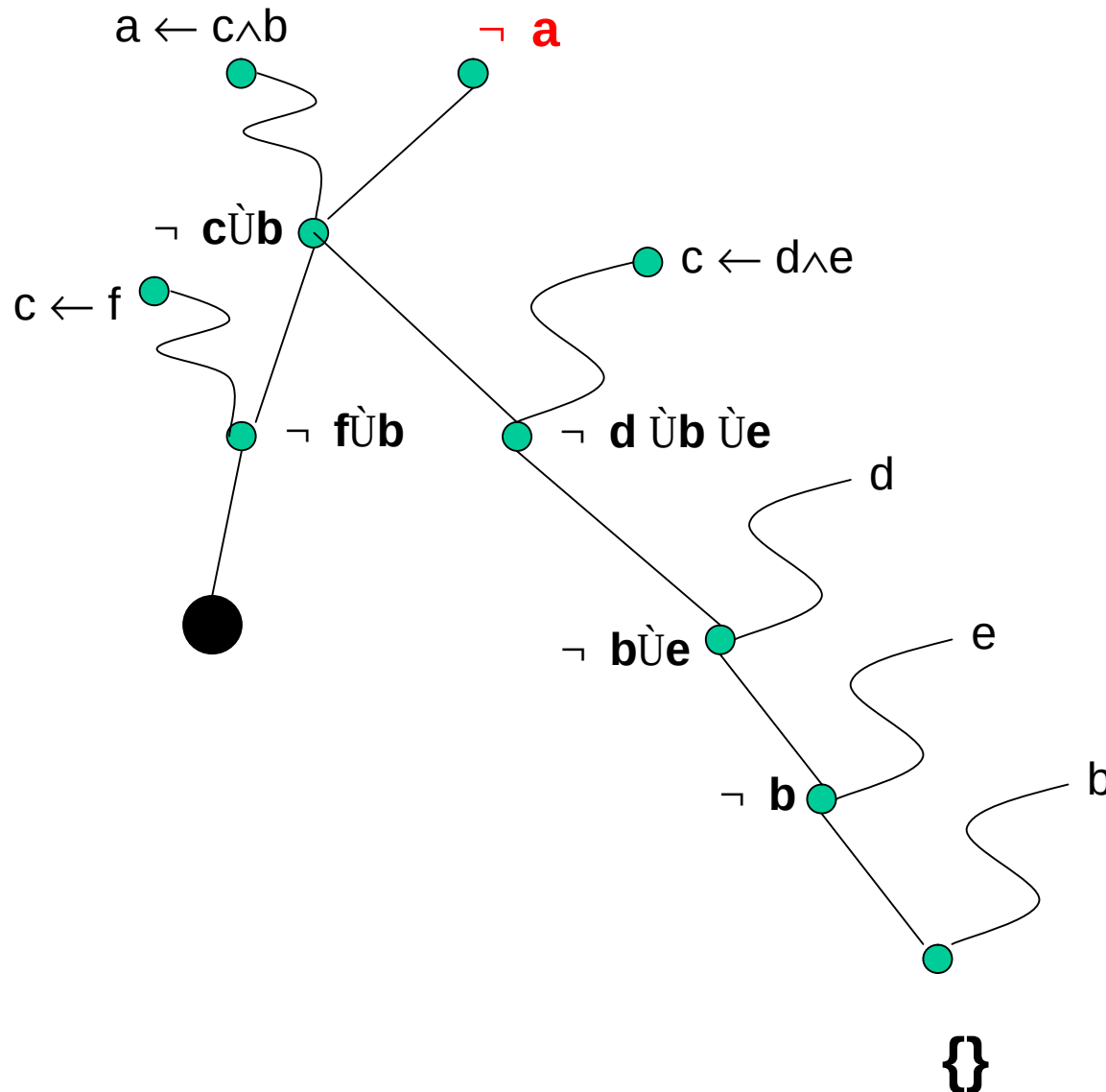
Delete the literal pair

Union of the clause sets

Artificial Intelligence

Example

Prolog Notation



- # 1. $a \neg c \cup b$

- 2. b**

- ### 3. c $\neg$ f

- ## 4. c $\neg$ dÙe

- ## 5. d

- ## 6. e

a ?

Artificial Intelligence

Depth search – PROLOG strategy

derivate (question)

sequential search a clause with the head „question“

IF clause head not found

IF backtracking possible

choose next backtracking point

OTHERWISE return NO

OTHERWISE

IF prämissen empty return YES

OTHERWISE

FOR ALL literals B_i of the selected clause
call from left to right derivate(B_i)

Artificial Intelligence

PROLOG is based on that strategy !

Problem ?

Cycles are possible !

Example:

1. b
2. $a \leftarrow b \wedge a$
3. a

We ask for a ?

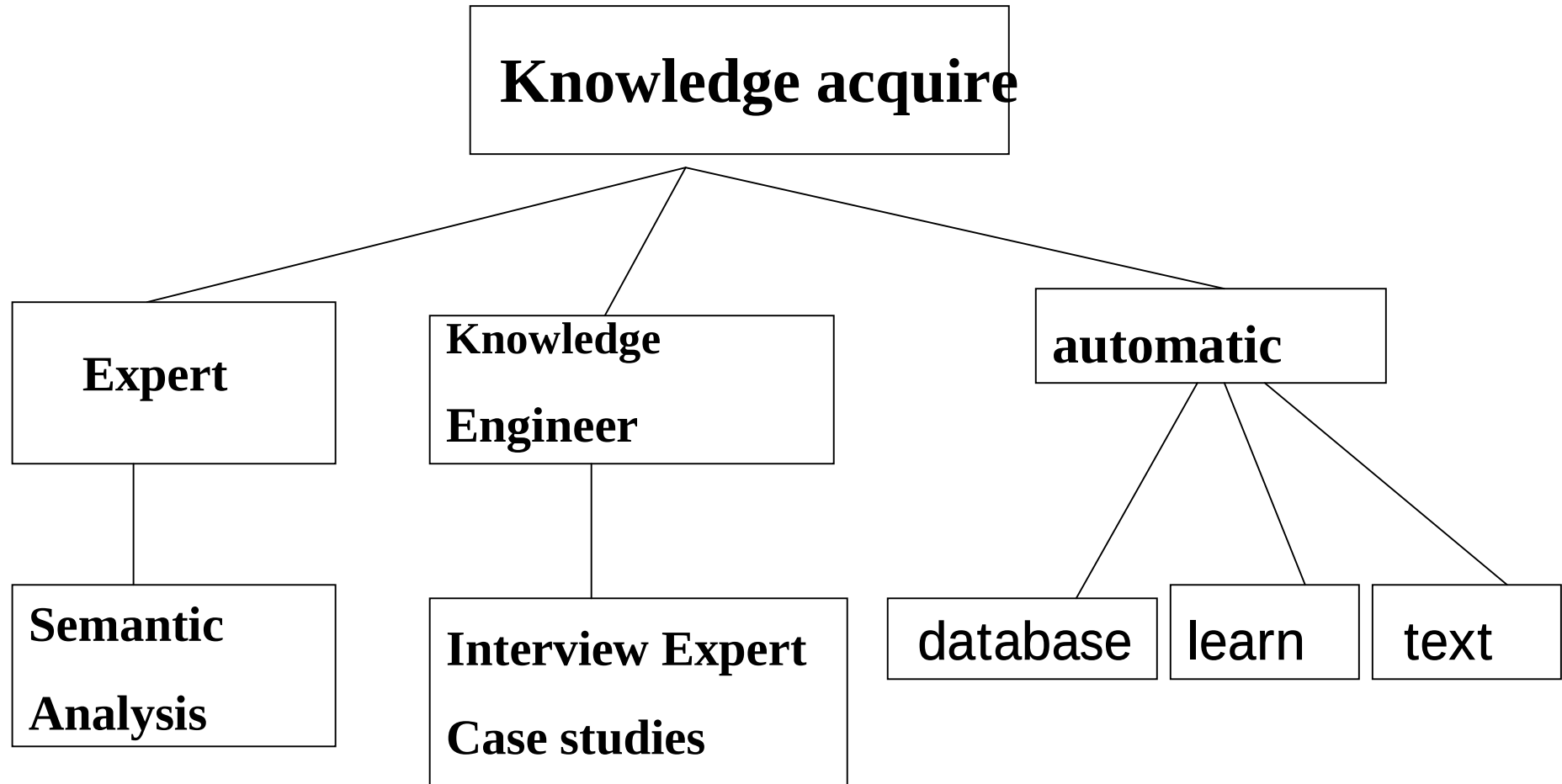
Infinite loop: see rule 2

Artificial Intelligence

Medical or molecular knowledge server

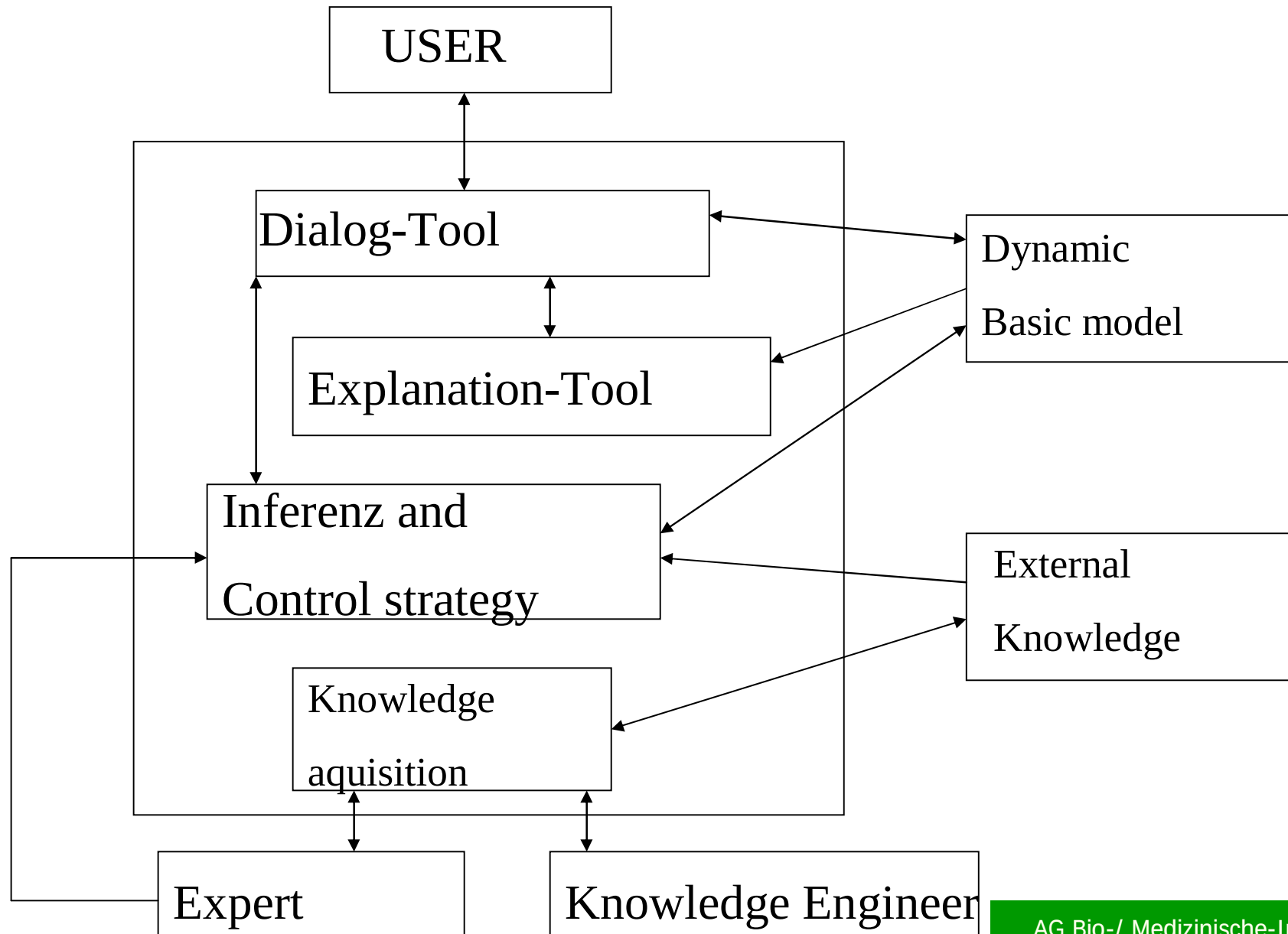
1. Electronical Networks
2. Literature-Database systems
Medline etc.
3. Systematic storing of data
METAGENE,
OMIM
4. Internet and information systems

Artificial Intelligence



Rule derivation

Expert System



Artificial Intelligence

Problem of expert systems ?

- a) Complexity of the inference algorithm
- b) Handle rudimentary knowledge

In the case of medical diagnosis:

The diagnosis process of a doctor „seems to be like art“!

Artificial Intelligence

New Methods to solve uncertainty problems:

Case based reasoning is one actual method !

Idea:

1. Define the case (parameters)
2. Select similar cases

Artificial Intelligence

CYRUS-System

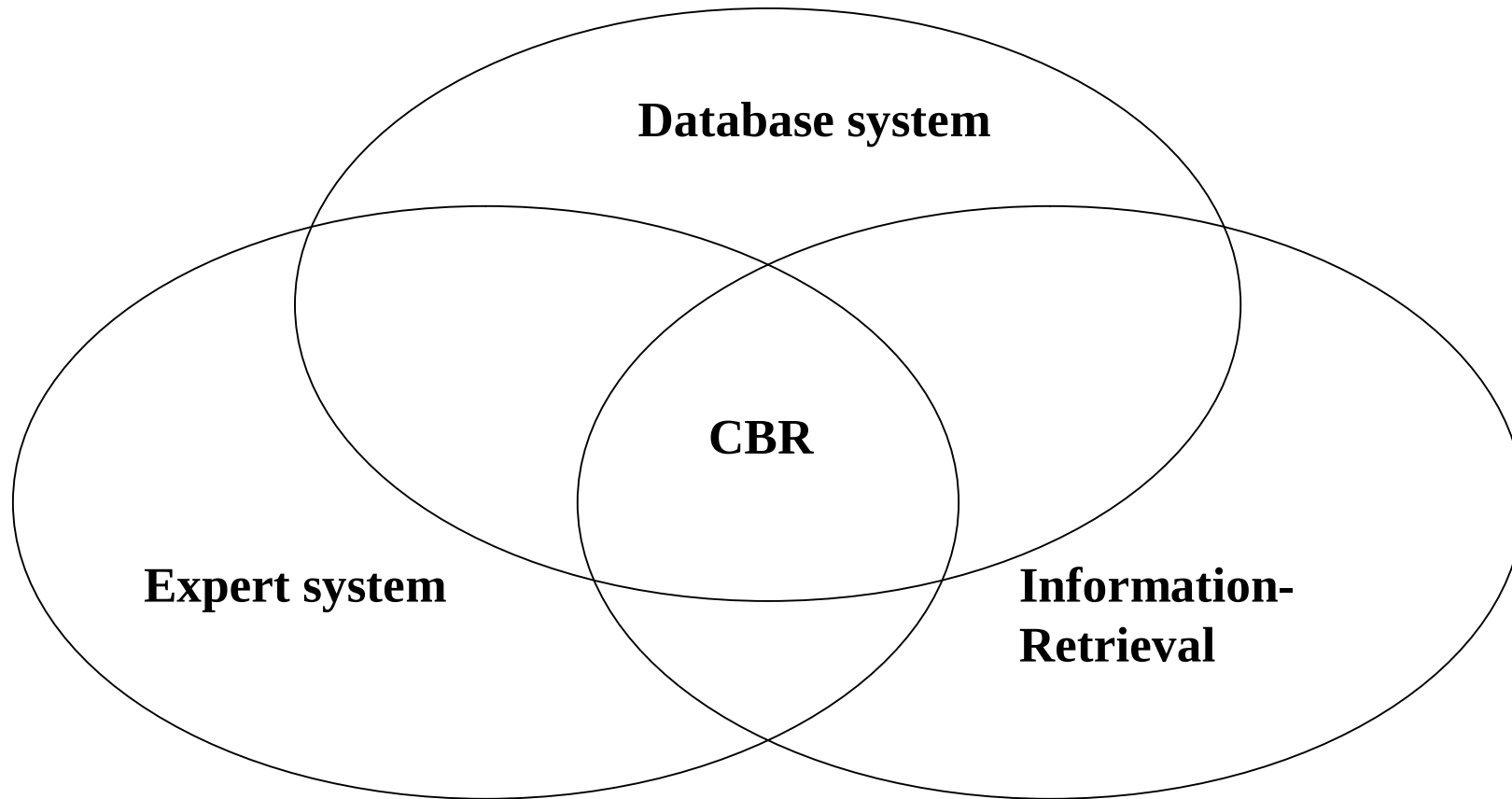
Janet Koldoner 1983

Yale University.

Question-Answer System:

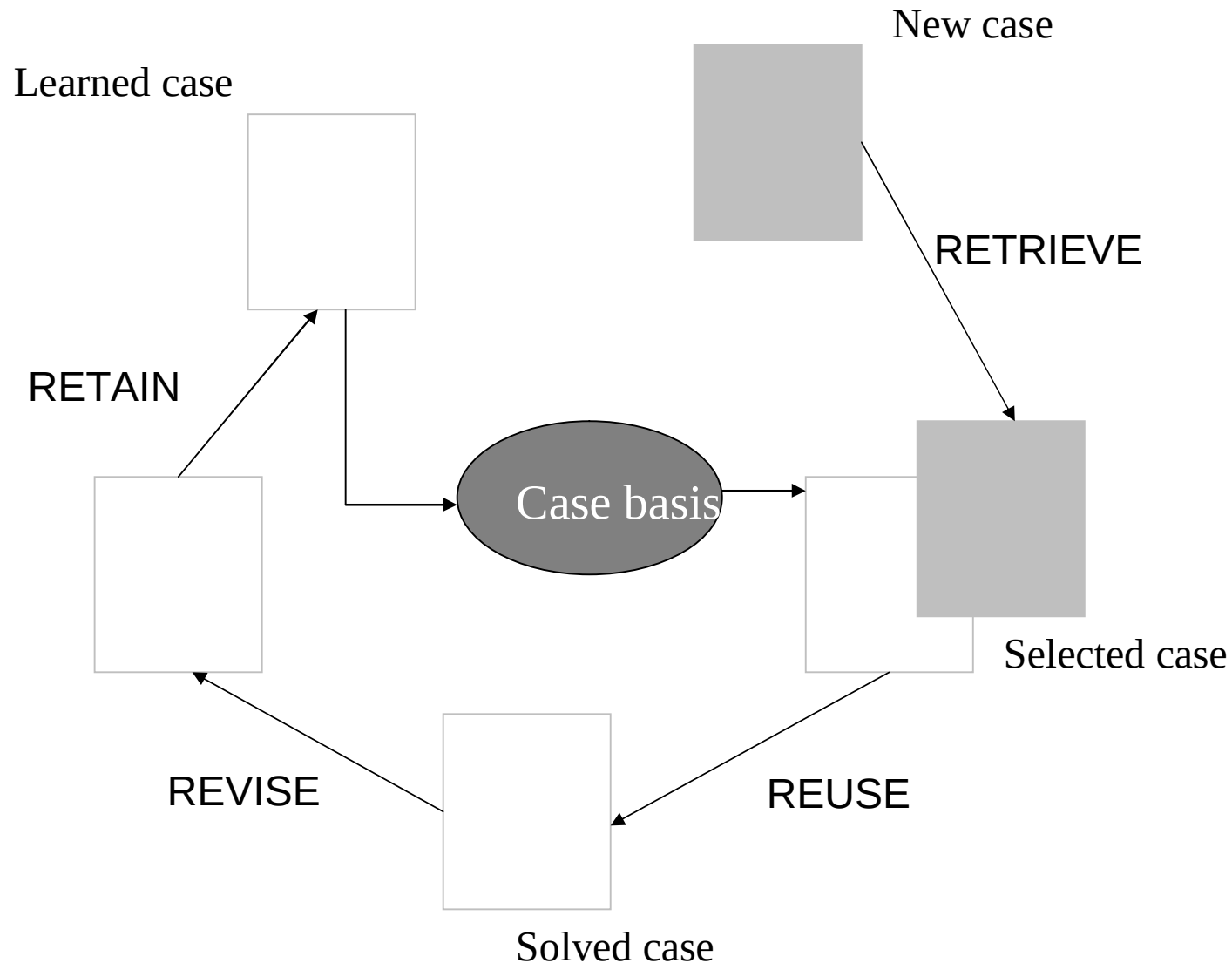
*Knowledge about different visits of the US mimister
Cyrus Vance.*

Artificial Intelligence



Context of Case Based Reasoning (CBR)

Artificial Intelligence



Ramedis

Ramedis : Rare Metabolic Diseases Database

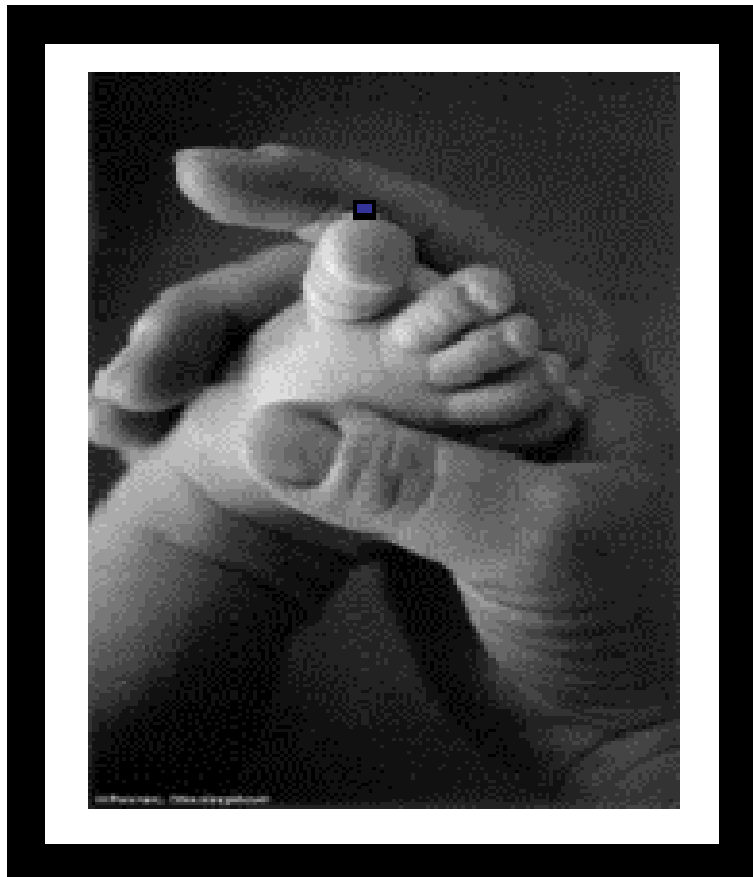
Web-based database and publication tool for inborn errors

- German Human Genome Project (DHGP)
- World-wide collection of seldom inborn errors
- Storage of patient cases
- <http://www.ramedis.de>

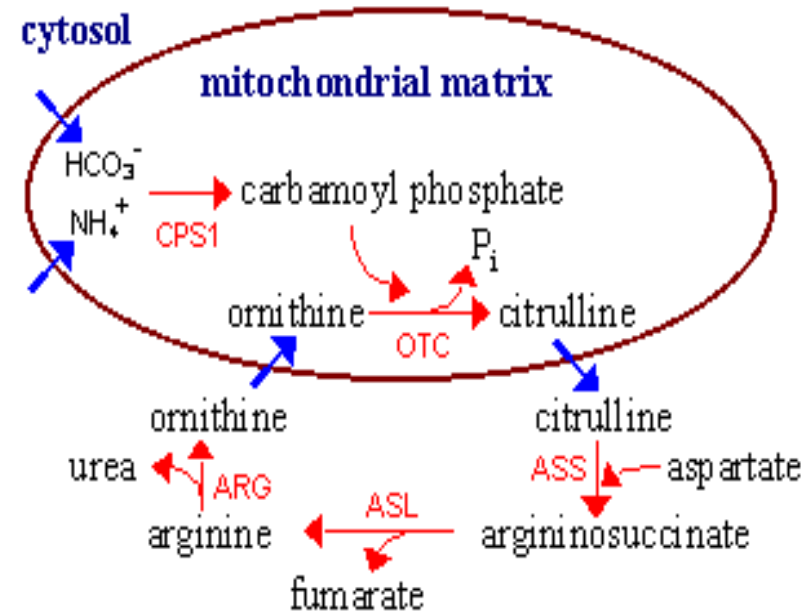
Cooperation with Prof. Dr. Trefz (Hospital Reutlingen)



An Biological Application: Urea Cycle Systems analysis



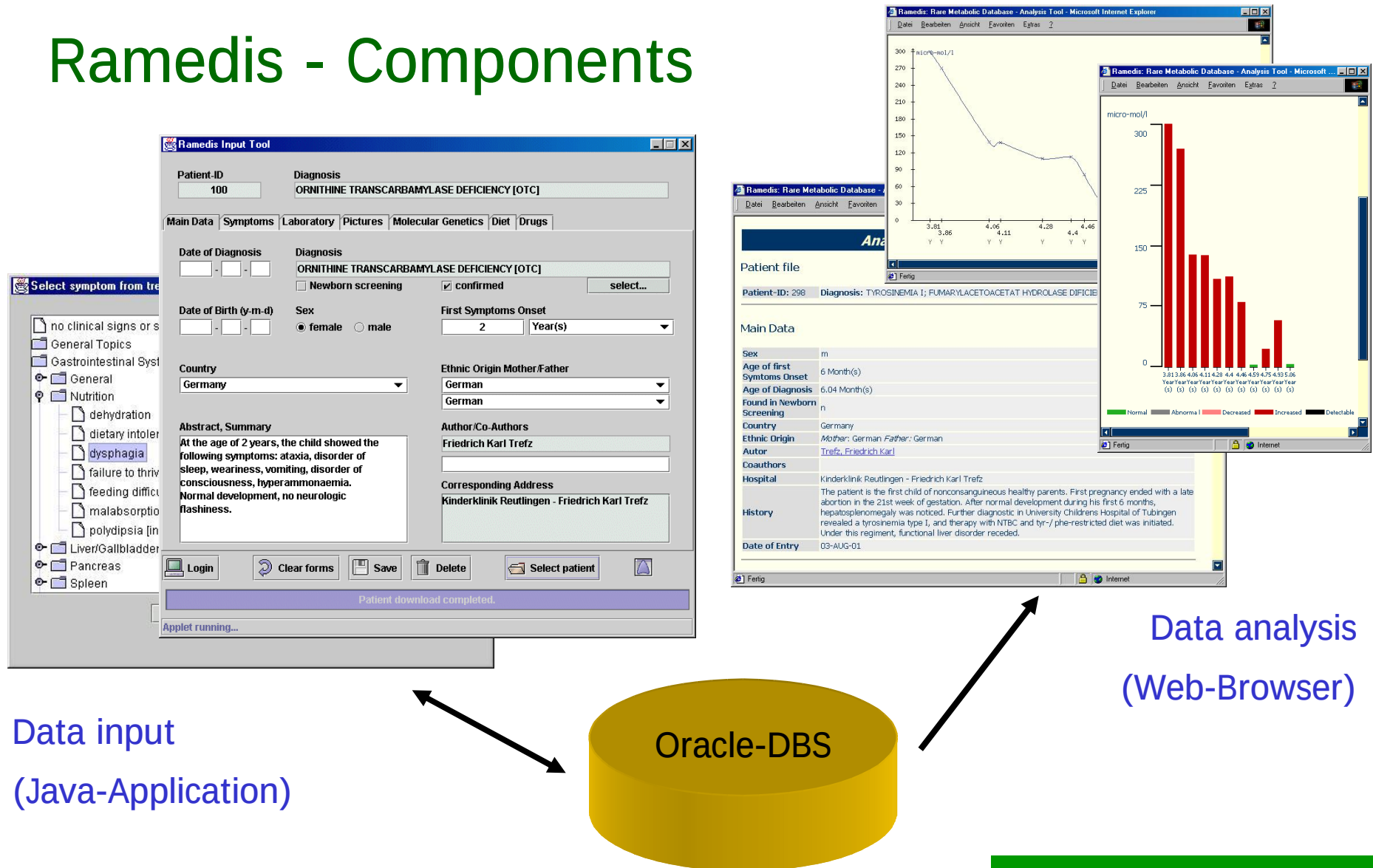
1 in 30,000 live birth !



Key enzymes in regulation of urea cycle in cells.
CPS1: Carbamyl phosphate synthetase, EC 6.3.4.16;
OTC: Ornithine transcarbamylase, EC 2.1.3.3;
ASS: Argininosuccinate synthetase, EC 6.3.4.5;
ASL: Argininosuccinate lyase, EC 4.3.2.1;
ARG: Arginase, EC 3.5.3.1

Ramedis

Ramedis - Components



Ramedis

CASE:

Problem description

Symptoms with occurrence and labor values

Problem solving

Diagnosis or differential diagnosis

Addition

Infos: Doctor and history of the disease

Ramedis

Case Based Reasoning

Query: *similar case*

Input (WEB mask):

sex, symptoms (5) and labor value (5) and ethnic birth

Case Retrieval (pre-selection)

Ramedis – pre-selection – partial equality (by Wess):

If one symptom or labor value is identical

=> element of the pre-selected case set

Ramedis

Similarity calculation

The calculation of the similarity:

$$\begin{aligned}\text{sim}(\text{case}) = & w1 * \text{sim}(\text{symptoms_case}) + \\ & w2 * \text{sim}(\text{labsubst_case}) + \\ & w3 * \text{sim}(\text{sex_case}) + \\ & w4 * \text{sim}(\text{ethnic_case})\end{aligned}$$

Variable w1 - w4 are the weight values.

Ramedis

CBR

New Case

Ramedis: Rare Metabolic Database - Analysis Tool - Microsoft Internet Explorer

Ramedis
Rare Metabolic Diseases Database

Analysis Tool

Search by combined values

Select symptoms	Select lab substances
Symptom 1	Lab Substance 1
Symptom 2	Lab Substance 2
Symptom 3	Lab Substance 3
Symptom 4	Lab Substance 4
Symptom 5	Lab Substance 5

Select ethnic origin

Mother	German
Father	German

Define weights

Symptoms		%
Lab Substances		%
Ethnic origin		%

Define number of rows

Number of shown cases	
-----------------------	----------------------

Start Search Reset

Similar Case

Ramedis: Rare Metabolic Database - Analysis Tool - Microsoft Internet Explorer

Ramedis
Rare Metabolic Diseases Database

Analysis Tool

Search result
for protein (natural)

PatID	Diagnosis	Conf.	Onset	Country	Ethnic Mother	Ethnic Father	Author
151	METHYLMALONYL-CoA MUTASE DEFICIENCY	y	13 Month (s)	Germany	German	German	Friedrich Karl Trefz
152	CITRULLINEMIA; ASS	y	4 Day(s)	Germany	Turkish	Turkish	Friedrich Karl Trefz
161	PROPIONIC ACIDEMIA; KETOTIC HYPERGLYCINEMIA [PA]	y	4 Day(s)	Germany	German	German	Friedrich Karl Trefz
173	METHYLMALONYL-CoA MUTASE DEFICIENCY	y	3 Week(s)	Germany	Turkish	Turkish	Friedrich Karl Trefz
204	METHYLMALONYL-CoA MUTASE DEFICIENCY	y	6 Day(s)	Germany	German	German	Friedrich Karl Trefz
209	METHYLMALONYL-CoA MUTASE DEFICIENCY	y	4 Day(s)	Egypt	Arab	Arab	Friedrich Karl Trefz
241	GLUTARIC ACIDURIA I	y	27 Month (s)	Germany	German	German	Published Cases
299	3-METHYL-CROTONYL-GLYCINURIA	y	7 Week(s)	Netherlands	Caucasian	Caucasian	Published Cases
301	CARBAMOYL PHOSPHATE SYNTHETASE DEFICIENCY [CPS]	y	2 Day(s)	Japan	Japanese	Japanese	Published Cases
302	CARBAMOYL PHOSPHATE SYNTHETASE DEFICIENCY [CPS]	y	3 Day(s)	Japan	Japanese	Japanese	Published Cases
306	PEARSON SYNDROM	y	3 Week(s)	Germany	German	German	Published Cases

Database Integration

SRS (Lion Bioscience)

PEDANT (BioMax

MOBY DICK

HUSAR

BioKleisli

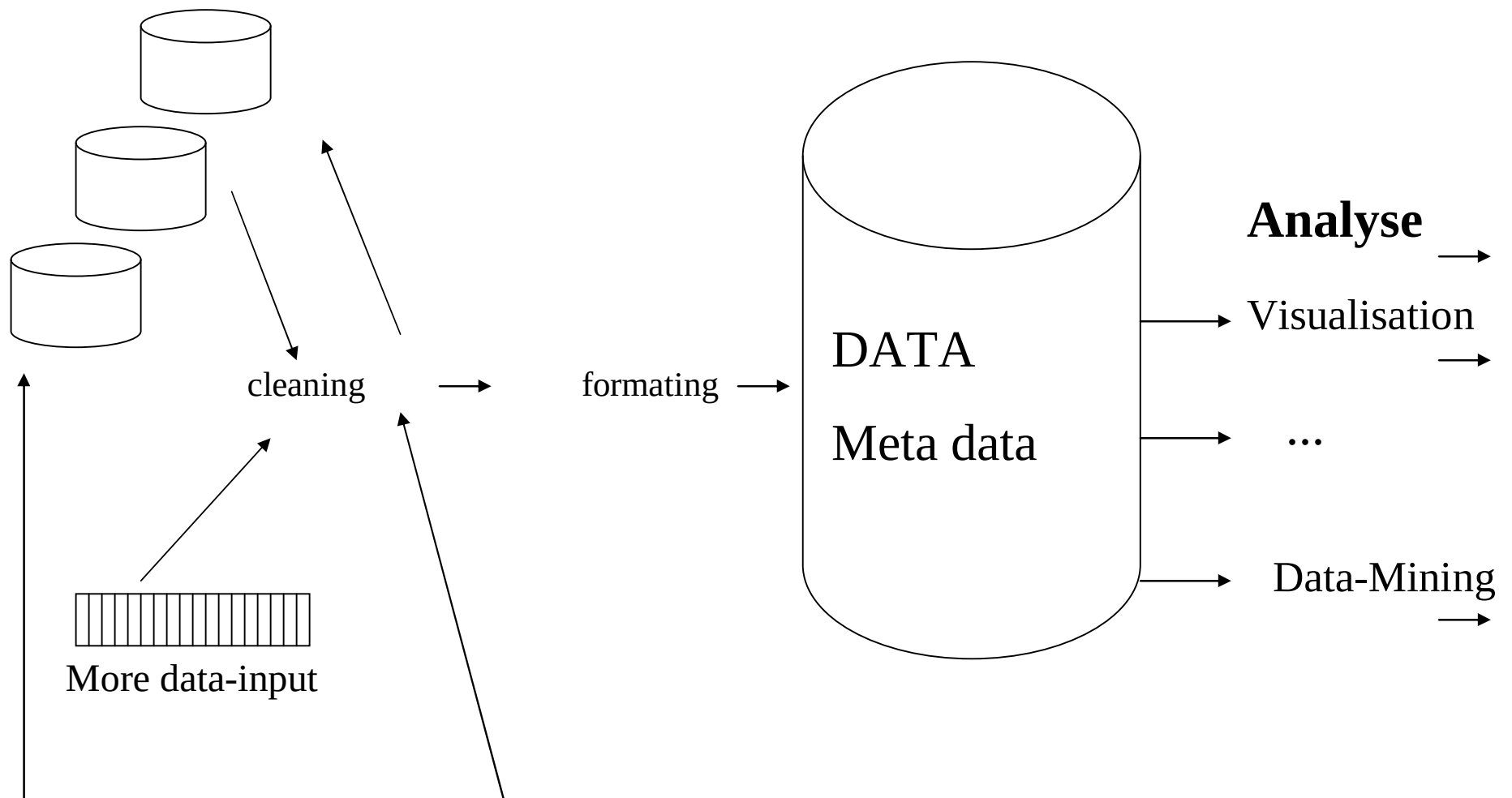
What Is There (WIT)

Biology Workbench

Integrated Genomic Database (IGD)

...

Date Warehouse



Data-Mining

Data-Mining

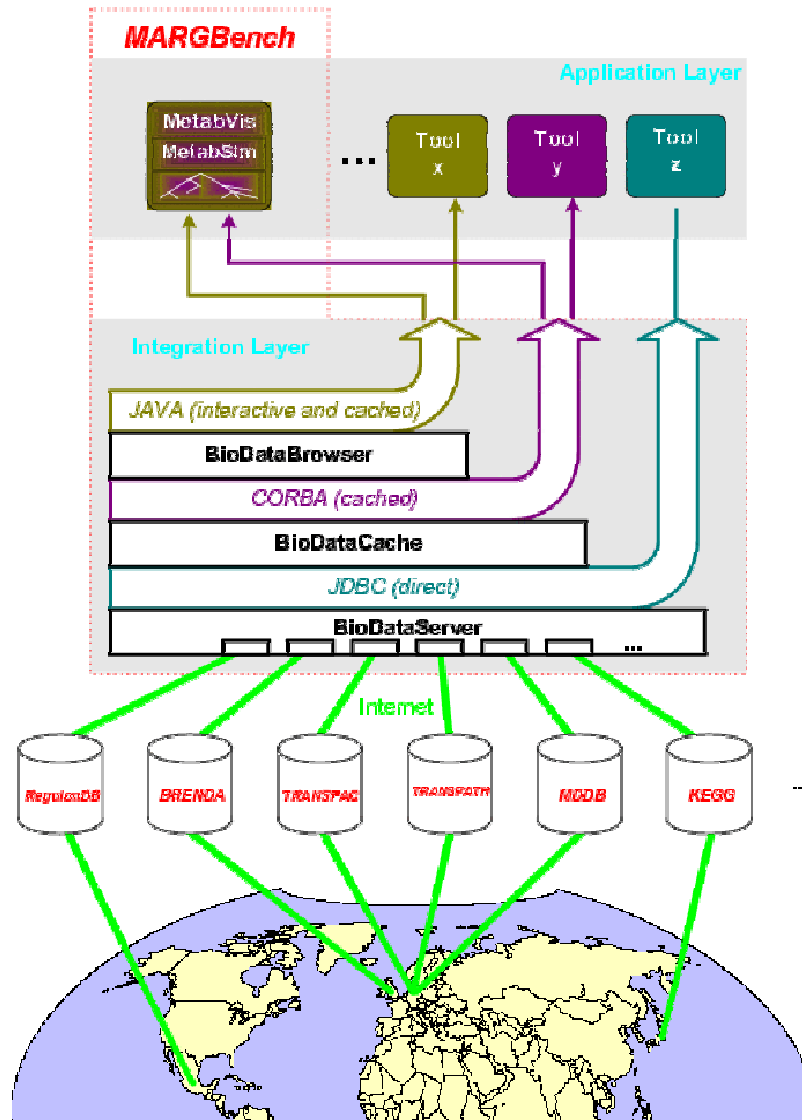
Detection of new information based on the analysis of patterns or rules inside the data (hugh).

Methods in use: Statistical analysis,
 optimization and
 computer based learning.

What else ?

- Neuronal Networks,
- Genetic Algorithms,
- Clustering and Segmentation ...

MARG – System Architecture



MARGBench

BioDataServer
BioDataCache
MetabSIM

*Integrative modeling,
analysis and
simulation
of metabolic processes.*

DFG Project

MARG

Network analysis: Problem and Algorithm

Application regarding object-oriented database systems



§ Path navigation

§ Transitive closure

§ Shortest pathway

§ ***Pathway Prediction***

§ Clique search

§ Minimum spanning tree*

§ Smallest tree (Steiner tree)*

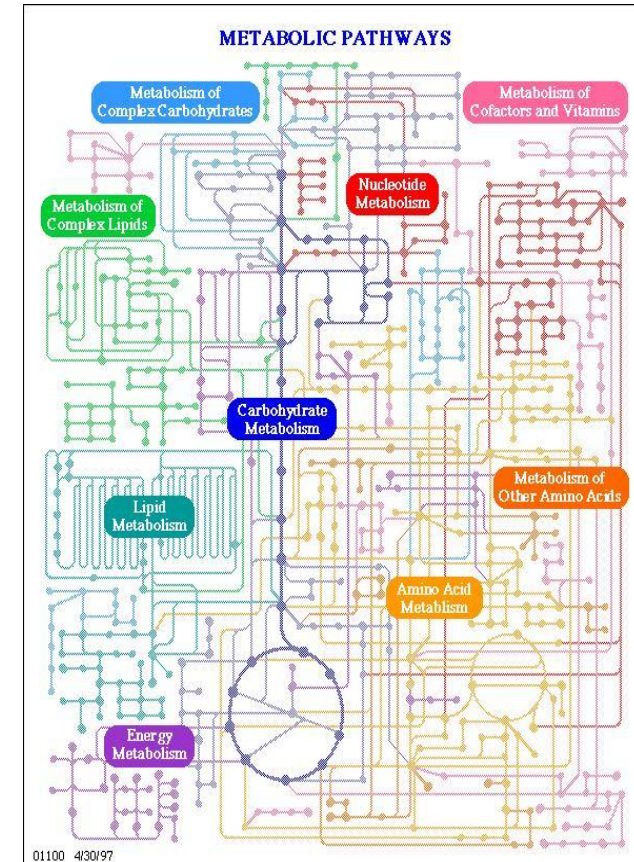
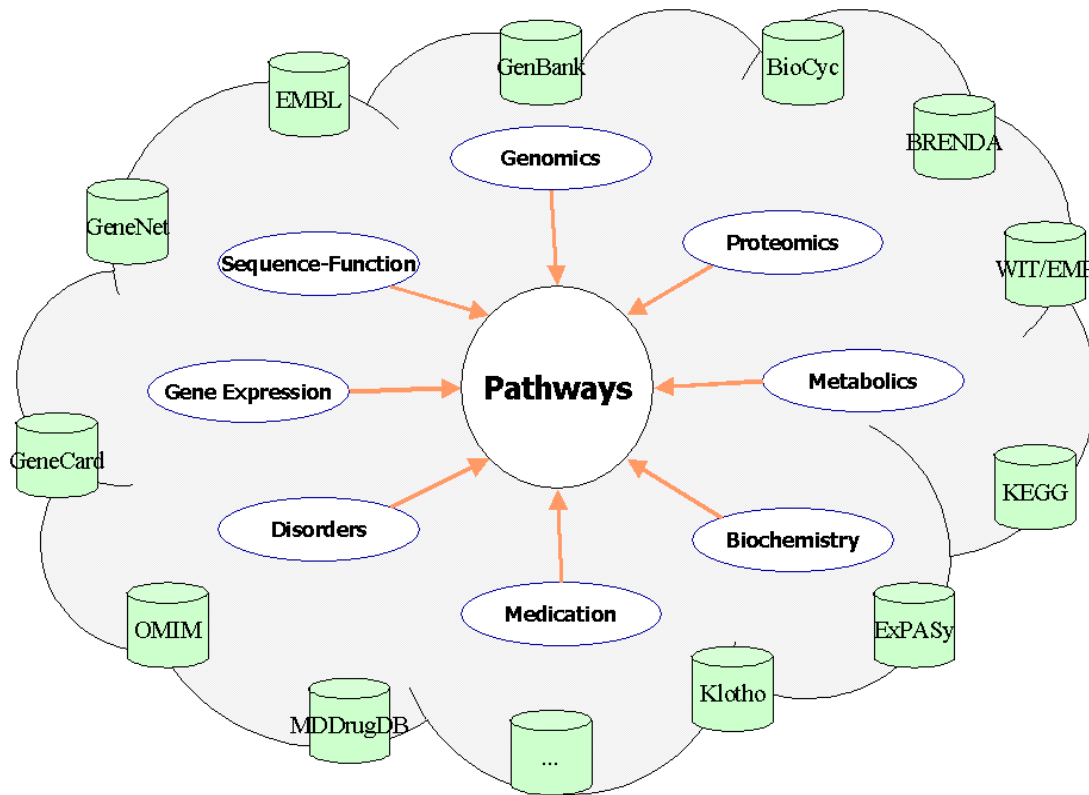
§ Boolean operations*

§ Interactive pathway synthesis

§ ...

**not implemented yet*

Pathway Prediction



- Rudimentary data and knowledge is available

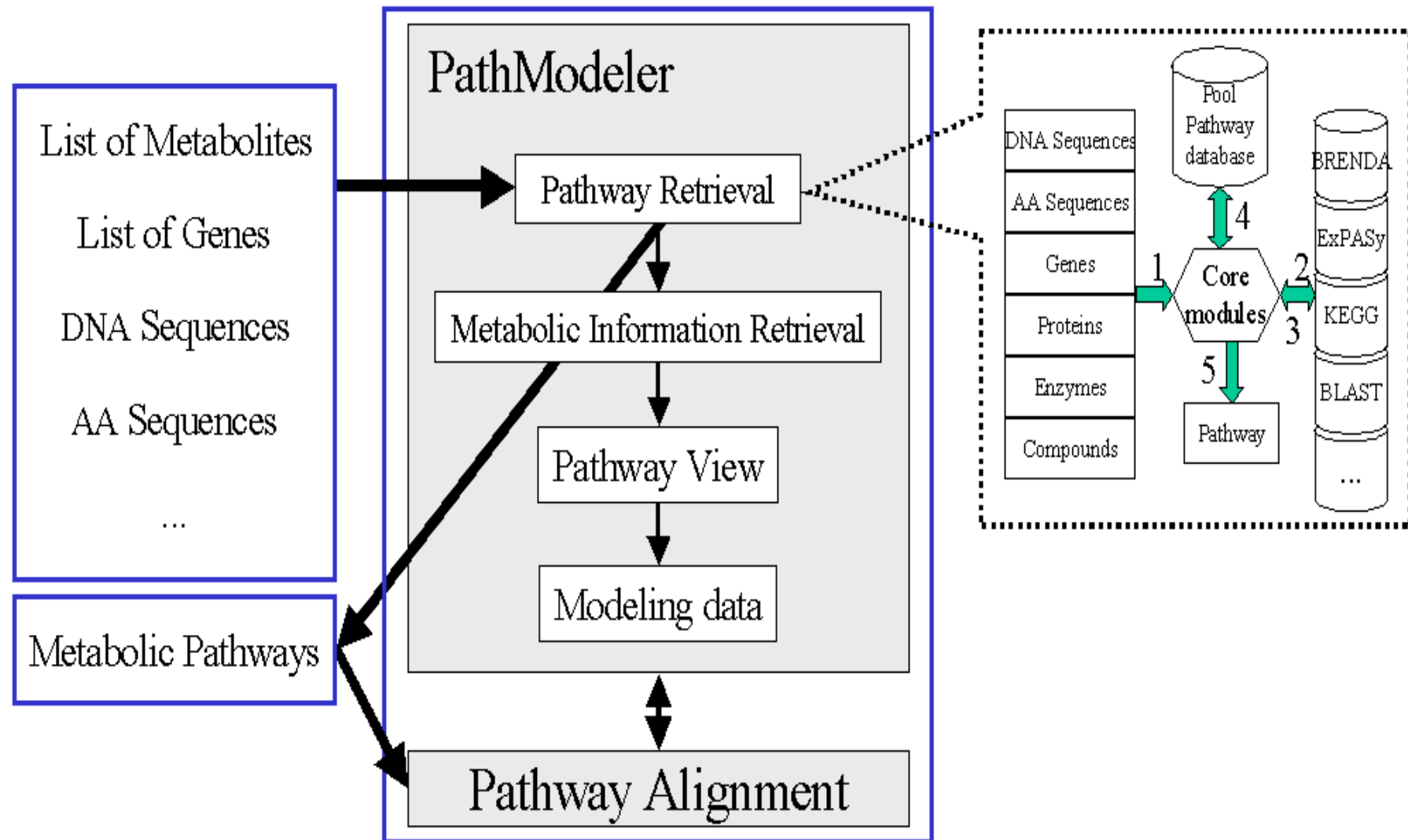
Motivation: Information system for retrieving and predicting of metabolic networks.

Problems

1. Given a rudimentary metabolic pathway (some genes, sequences, enzymes and metabolites), how can we predict the complete gene controlled metabolic pathway?
2. Given two gene controlled metabolic pathways, how can we calculate the similarity of both pathways?

We have designed the web based system **PathAligner**.

Concept



BiBiServ – PathAligner

Bielefelder
Bioinformatic
Server

The image displays two overlapping browser windows from the BiBiServ website. The background window is the 'PathAligner - Retrieval' page, which contains a form for submitting queries. The foreground window shows the 'PathAligner - Results' page, displaying the output of a query.

BiBiServ - Bielefeld University Bioinformatics Server

Tools Education Administration News Links

PathAligner - Retrieval

- Metabolic pathway reconstruction/retrieval
 - Reconstruction query

Please insert rudimental elements of the desired pathway, using "/" for multiple entries.

☒ Metabolite (Compound)
L-citrulline

☒ Enzyme and/or EC number
Arginase/6.3.4.16

☒ Gene
OTC

☒ Amino acid sequence
mifnlrillnnaafrnghnfmvrnfrcgqplqnkvqlkgrdiltlknftgeeikym
geylp1lqqkslgmifekrstrtr1stetglallgghpcflttqdihlgnestlc
rvyqgsdldtlakeasipiinglsdlyhpiqiladytlqehysslkgltlswigc
fgmhlqaatpkgyepdasvklaeqyakengtkliltndpleahgggnvliatdwis
qgyqvmtktakvaasdwtf1hclprkpeevddevfysprslvfpeaenrkwimav
pkf

or upload from local file

☐ Nucleotide sequence
CTGTGTTCACTGTTTCGAGTTACTTCTTCTGGTGTCTTCCCCAGGGATGGGTTCACT
TCCTGAGAGTAGCCCTGTTTTGTAGATTCTTCTGTGATGTAGAGACCTTCTCTGT
ACGACTGTTGTCAGTAGCTGCTTCCTAGTAGATAGCTGAGTGTCTTCTCCATCA
AGTCTGTCCACTTCAGTCATTGAAAAGTATTTAATGCCTAGAGTTTTCAAAATGT
CCACGTCTCTCAAGCCAAATATACACAATATCCTTGGCAGATATACAGGGAGTCACC
CACATCGGGAATCTTTCCTTTTAGTTCTTCAGGAGGAAAGATACAGGTTGTCCAT

or upload from local file

pathaligner - Results - Microsoft Internet Explorer

Address: http://bibiserv.techfak.uni-bielefeld.de/cgi-bin/pathaligner_0

PathAligner - Retrieval

- Metabolic pathway retrieval
 - Result

Your query contain compounds:
L-citrulline

Your query contain enzymes:
Arginase 6.3.4.16

Your query contain gene names:
OTC

Your query contain protein sequence:
mifnlrillnnaafrnghnfmvrnfrcgqplqnkvqlkgrdiltlknftgeeikymiwlsadlkfrikqkgeylp1lqq

It most likely belongs to the following pathways, if any! ;-)

☒ **Urea cycle pathway: 6.3.4.16->2.1.3.3->6.3.4.5->4.3.2.1->3.5.3.1**

☐ **Or your designed pathway::**

(C) Ming Chen

Done Local intranet

BiBiServ – PathAligner

Bielefelder
Bioinformatic
Server

PathAligner - Results - Microsoft Internet Explorer

Address http://bibiserv.techfak.uni-bielefeld.de/cgi-bin/pathaligner_EC

PathAligner - Retrieval

- Metabolic pathway retrieval
 - Metabolic information

RESULTS

The pathway EC entry is:
6.3.4.16\2.1.3.3\6.3.4.5\4.3.2.1\3.5.3.1

EC number	Km Reaction	Gene	Factor (Biobase password protected)	GeNetView (password)	Drug target	URL link to ExPASy
EC 6.3.4.16	Km Reaction	1373(CPS1)	unknown	NA	-	6.3.4.16
EC 2.1.3.3	Km Reaction	5009(OTC)	HNF-4alpha1 T00372 ; HNF-4alpha2 T02422 ; HNF-4alpha1 T02429 ; EBPbeta T00459 ; EBPalpha T00105 ;	OTC	-	2.1.3.3
EC 6.3.4.5	Km Reaction	445(ASS)	unknown	NA	-	6.3.4.5
EC 4.3.2.1	Km Reaction	435(ASL)	unknown	NA	-	4.3.2.1
EC 3.5.3.1	Km Reaction	383(ARG1) 384(ARG2)	unknown	NA	-	3.5.3.1

Network Data

KEGG Associated Pathway(s)

- [hsa00220 Urea cycle and metabolism of amino groups](#)
 - EC 2.1.3.3
 - EC 3.5.3.1
 - EC 4.3.2.1
 - EC 6.3.4.5
 - EC 6.3.4.16
- [hsa00251 Glutamate metabolism](#)
 - EC 6.3.4.16
- [hsa00252 Alanine and aspartate metabolism](#)
 - EC 4.3.2.1
 - EC 6.3.4.5
- [hsa00330 Arginine and proline metabolism](#)
 - EC 2.1.3.3
 - EC 3.5.3.1
 - EC 4.3.2.1
 - EC 6.3.4.5
 - EC 6.3.4.16
- [hsa00910 Nitrogen metabolism](#)
 - EC 6.3.4.16

(C) [Ming Chen](#)

Local intranet

BiBiServ – PathAligner

Bielefelder
Bioinformatic
Server

pathaligner - Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://bibitest.techfak.uni-bielefeld.de/cgi-bin/pathaligner_biopnml Go

 **PathAligner - Retrieval**

- Metabolic pathway retrieval
 - ▾ BioPNML code export

RESULTS

BiBiServ
Bielefeld University Bioinformatics Server
PathAligner

The pathway EC entry is:
6.3.4.16\\2.1.3.3\\6.3.4.5\\4.3.2.1\\3.5.3.1

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE net SYSTEM "BioPNML.dtd">
<BioPNML>
  <Petrinet id="pn1" type="Hybrid">
    <place id="6.3.4.16" type="continuous">
      <name id="0">
        <text>6.3.4.16</text>
        <value>m1</value>
      </name>
      <graphics>
        <size>10</size>
        <position x="-20" y="10"/>
        <color>red</color>
      </graphics>
      <initialMarking>
        <value>1</value>
      </initialMarking>
      <annotation>
        <name></name>
        <TypeRef/>
        <species>human</species>
        <location/>
        <concentration>1 (mM) </concentration>
        <comment/>
      </annotation>
    </place>
    <transition id="t6.3.4.16" type="continuous">
      <PathRef/>
      <reaction name="reaction_1" reversible="false">
        <enzyme>6.3.4.16</enzyme>
      </reaction>
    </transition>
  </Petrinet>
</BioPNML>
```

Save as

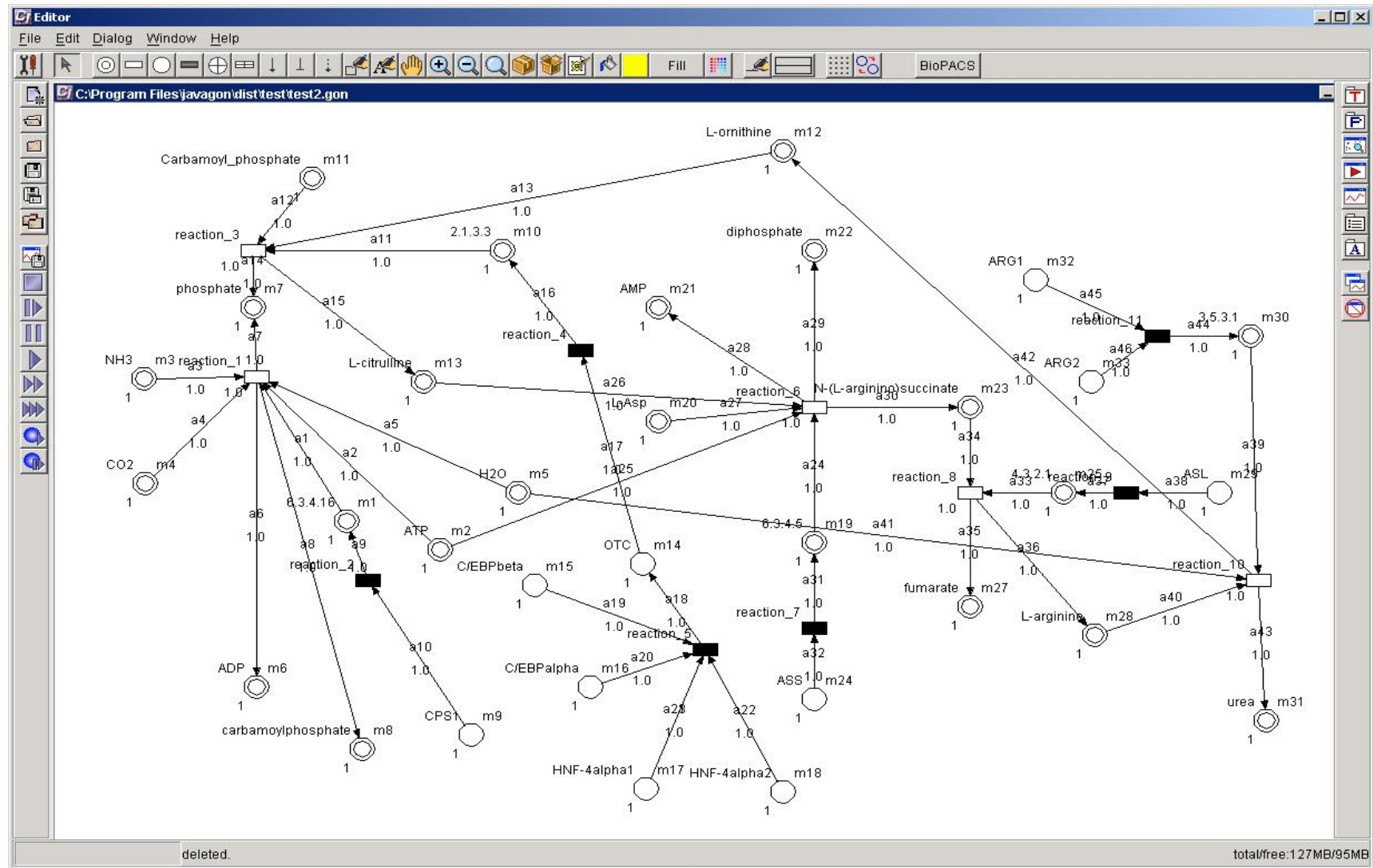
Translate into: SBML Go!

Note: Due to its immature, XML::XSLT might not support all functions. The XSLT file is available upon request.

(C) [Ming Chen](#)

Local intranet

BiBiServ – PathAligner



Alignment

Given a list of letters, METBOLIWAY and TABOLIXYZ, a good alignment might be:

MET - BOLIWAY
| | | |
TABOLI - XYZ

Many possible alignments (e.g. Dynamic programming algorithms) for two pathways, and different alignments may have different scores.

Our approach: *successive enzymatic reactions alignment*

M. Chen, S. Lin and R. Hofestädt

STCDB: Signal Transduction Classification Database.

Nucleic Acids Research, 2004, 32(1): D456-D458

Comparison of metabolic pathways

- Horowitz, 1945: **Retrograde evolution**, Sequential “disappearance” of key intermediary metabolites induces the recruitment of similar available substrates via new enzymes.
- Jensen, 1976: **Substrate ambiguity**, Enzyme recruitment from a pool of ancestral enzymes with basic functions and substrate ambiguity. Ordered regulated pathways evolved from these ancestral enzymes by gene duplication, followed by increased specialisation.
- Huynen, Dandekar and Bork (1999)
 - Citric acid cycle
 - Trends in Microbiology (1999)
- Dandekar, Schuster, Snel, Huynen and Bork (1999)
 - Glycolysis
 - Biochemical Journal (1999)
- Cordwell
 - Glycolysis, pentose phosphate pathway, tricarboxylic acid cycle
 - Archives in Microbiology (1999)
- Y. Tohsato et al. (2000) proposed a multiple (local) alignment algorithm utilizing information content that is extended to symbols having a hierarchical structure.
- Forst C.V. (2001) combined sequence information of involved genes with information of corresponding network.

Definitions

Let S be a finite set of EC numbers.

$E \in S^*$ is a *pathway* $(e_1 e_2 \dots e_n)$

The *length* of E is $|E|$.

The *empty string* (EC number) is e .

Edit Operation

An edit operation is an ordered pair

$$(a,b) \in (E \cup \{e\}) \times (E \cup \{e\}) \setminus \{(e,e)\}$$

An edit operation (a,b) is written as $a \rightarrow b$.

There are three kinds of edit operations:

$a \rightarrow e$ **deletion** of the enzyme a

$e \rightarrow b$ **insertion** of the enzyme b

$a \rightarrow b$ **replacement** of the enzyme a by the enzyme b

notice that $e \rightarrow e$ never happens.

Alignment

Let $P_1 = p_1 p_2 \dots p_m$ and $P_2 = p_1' p_2' \dots p_n'$ be two metabolic pathways.

An alignment of P_1 and P_2 is a sequence

$$(a_1 \rightarrow b_1, \dots, a_h \rightarrow b_h)$$

of edit operations such that

$$P_1' = a_1, \dots, a_h \text{ and } P_2' = b_1, \dots, b_h.$$

Notation

The alignment

$A = (2.4.2.3 \rightarrow 2.4.2.4, 3.5.4.5 \rightarrow e, 3.1.3.5 \rightarrow 3.1.3.5, e \rightarrow 2.7.4.9)$

of the pathways

$e_{2.4.2.3}e_{3.5.4.5}e_{3.1.3.5}$ and $e_{2.4.2.4}e_{3.1.3.5}e_{2.7.4.9}$

is written as follows:

$$\begin{pmatrix} 2.4.2.3 & 3.5.4.5 & 3.1.3.5 & e \\ 2.4.2.4 & e & 3.1.3.5 & 2.7.4.9 \end{pmatrix}$$

Similarity function s

s assigns to each edit operation a nonnegative real number.

$s(a,e)$ and $s(e,b)$ of the edit operation (a,e) and (e,b) is 0.

For all edit operations (a,b) $a \neq e$, $b \neq e$, say, $a = d_1.d_2.d_3.d_4$ and $b = d_1'.d_2'.d_3'.d_4'$, then the similarity function $s(a,b)$ is defined by:

$$s(a,b) = \begin{cases} 0, & \text{if } (d_1 \neq d_1'); \\ 0.25, & \text{if } (d_1 = d_1' \text{ and } d_2 \neq d_2'); \\ 0.5, & \text{if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 \neq d_3'); \\ 0.75, & \text{if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 = d_3' \text{ and } d_4 \neq d_4'); \\ 1, & \text{if } (d_1 = d_1' \text{ and } d_2 = d_2' \text{ and } d_3 = d_3' \text{ and } d_4 = d_4' \text{ i.e. } a=b). \end{cases}$$

Similarity

The similarity $s(A)$ of an alignment $A=(a_1 \rightarrow b_1, \dots, a_h \rightarrow b_h)$ is the sum of the costs of the edit operations:

$$s(A) = \sum_{i=1}^h s(a_i \rightarrow b_i)$$

2.4.2.3	3.5.4.5	3.1.3.5	2.7.4.14
2.4.2.4		3.1.3.5	2.7.4.9

$$s(E_1, E_2) = 1/4(0.75 + 0 + 1 + 0.75) = 0.625$$

PathAligner

The image displays three screenshots of the PathAligner web application interface, showing the alignment process and results.

Top Left Screenshot: PathAligner - Alignment

- Tools: Pairwise alignment, Enzyme-Enzyme pattern search
- Paste your metabolic pathway(s) as an EC number sequence, e.g. 1.1.1.1->2.2.2.2->3.3.3.3. EC numbers may be found by using [ECsearcher](#).
- Pathway 1: 6.3.5.5->2.1.3.3->6.3.4.5->4.3.2.1
- Pathway 2: 6.3.5.5->2.1.3.2->3.5.2.3->1.3.3.1->2.4.2.10->4.1.1.23

Top Right Screenshot: pathaligner - Results

- PathAligner - Alignment
 - E-E Pathway Alignment
 - EC number pattern based
- Your first pathway is:
6.3.5.5->2.1.3.3->6.3.4.5->4.3.2.1
- Scoring: 0.75
arginine biosynthesis I::2.1.3.3->6.3.4.5->4.3.2.1
- Scoring: 0.70
Urea cycle pathway::6.3.4.16->2.1.3.3->6.3.4.5->4.3.2.1->3.5.3.1
- Scoring: 0.60
arginine biosynthesis 4::1.4.1.3->2.6.1.13->2.1.3.3->6.3.4.5->4.3.2.1
- Scoring: 0.60
arginine biosynthesis III::6.3.5.5->2.1.3.3

Bottom Left Screenshot: PathAligner - Alignment (RESULTS)

- RESULTS
- Your first pathway is:
6.3.5.5->2.1.3.3->6.3.4.5->4.3.2.1
- The second pathway is:
6.3.5.5->2.1.3.2->3.5.2.3->1.3.3.1->2.4.2.10->4.1.1.23
- Scoring: 0.33
- (C) [Ming Chen](#)

Discussion

Goal of „Systems Biology“

Virtual Cell

Tasks of Bioinformatics

- Electronical infrastructure
- Information fusion
- **Handle the time complexity (AI methods)**
- Visualization