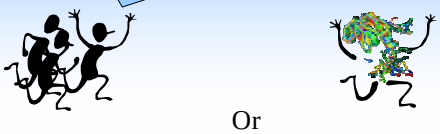
GET THAT PROTEIN!



Or
Implementation of and access to biological databanks

Patrick Lambrix
Department of Computer and Information Science
Linköpings universitet

1

Outline

- Biological databanks
- Storing and accessing textual information
- Accessing multiple biological databanks (link-driven federations, view integration)

2

Beyond the scope

- Non-textual data
- Alignments (e.g. BLAST)
- Visualization
- Data mining

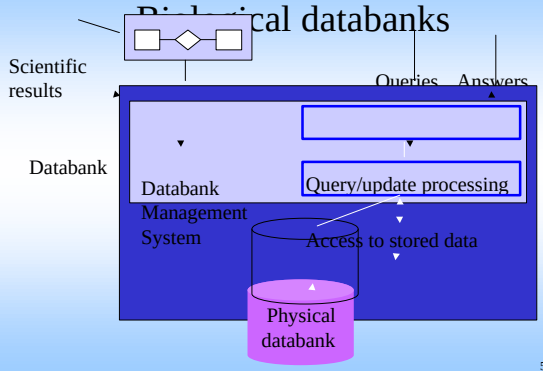
3

Biological databanks

- Biological data in electronic form
- well-known sources:
  e.g. SWISS-PROT, EMBL, DDBJ, PDB, GENBANK, KEGG, ACEDB
- Used in every day research

4

Biological databanks



5

Outline

- Biological databanks
- Storing and accessing textual information
- Accessing multiple biological databanks
  - link-driven federations
  - view integration

6

Storing and accessing textual information

- What information is stored?
- How is the information stored?
  - high level
- How is the information retrieved?

7

What information is stored?

- Model the information
  - Entity-Relationship model (ER)
  - Unified Modeling Language (UML)

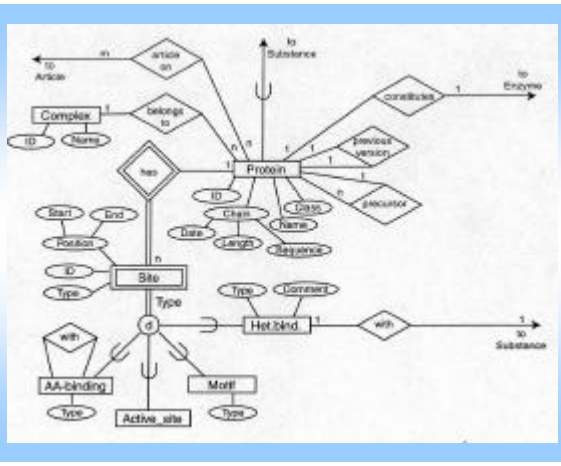
8

What information is stored? - ER

- entities and attributes
- entity types
- key attributes
- relationships
- cardinality constraints

- EER: sub-types

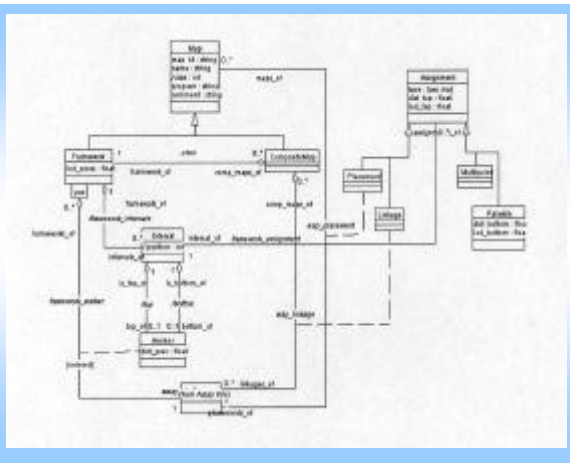
9



What information is stored? - UML

- classes, objects
- attributes
- operations
- associations, links

11

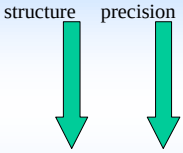


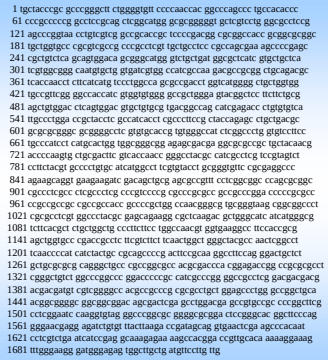
Storing and accessing textual information

- What information is stored?
- How is the information stored?
  - high level
- How is the information retrieved?

Storing textual information

- Text (IR)
- Semi-structured data
- Data models (DB)
- Rules + Facts (KB)



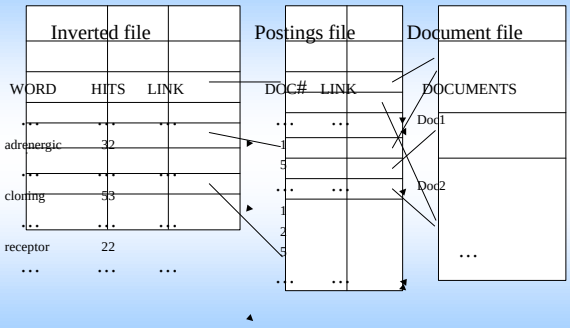


DEFINITION	Homo sapiens adrenergic, beta-1-, receptor
ACCESSION	NM_000684
SOURCE ORGANISM	human
REFERENCE	1
AUTHORS	Frielle, Collins, Daniel, Caron, Lefkowitz, Kobilka
TITLE	Cloning of the cDNA for the human beta 1-adrenergic receptor
REFERENCE	2
AUTHORS	Frielle, Kobilka, Lefkowitz, Caron
TITLE	Human beta 1- and beta 2-adrenergic receptors: structurally and functionally related receptors derived from distinct genes

Storing textual information - Text - Information Retrieval

- search using words
- conceptual models:
  boolean, vector, probabilistic, …
- file model:
  flat file, inverted file, ...

IR - File model: inverted files



IR - formal characterization

Information retrieval model: (D,Q,F,R)

- D is a set of document representations
- Q is a set of queries
- F is a framework for modeling document representations, queries and their relationships
- R associates a real number to document-query-pairs (ranking)

19

IR - conceptual models

Classic information retrieval

- Boolean model
- Vector model
- Probabilistic model

20

Boolean model

	adrenergic	cloning	receptor		
Doc1	yes	yes	no	-->	(1 1 0)
Doc2	no	yes	no	-->	(0 1 0)

Q1: cloning and (adrenergic or receptor)
--> (1 1 0) or (1 1 1) or (0 1 1) Result: Doc1

Q2: cloning and not adrenergic
--> (0 1 0) or (0 1 1) Result: Doc2

21

Boolean model

Advantages

- based on intuitive and simple formal model (set theory and boolean algebra)

Disadvantages

- binary decisions
  - words are relevant or not
  - document is relevant or not, no notion of partial match

22

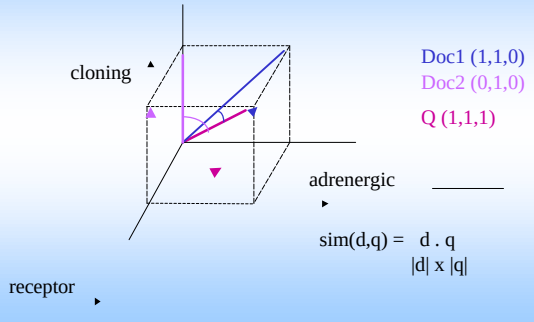
Boolean model

	adrenergic	cloning	receptor		
Doc1	yes	yes	no	-->	(1 1 0)
Doc2	no	yes	no	-->	(0 1 0)

Q3: adrenergic and receptor
--> (1 0 1) or (1 1 1) Result: empty

23

Vector model (simplified)



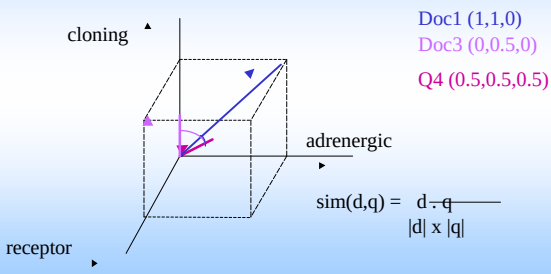
24

Vector model

- Introduce weights in document vectors
  (e.g. Doc3 (0, 0.5, 0))
- Weights represent importance of the term for describing the document contents
- Weights are positive real numbers
- Term does not occur -> weight = 0

25

Vector model



26

Vector model

- How to define weights? tf-idf

$d_j (w_{1,j}, \ldots, w_{t,j})$

$w_{i,j}$ = weight for term k_i in document d_j
= $f_{i,j}$ x idf_i

27

Vector model

- How to define weights? tf-idf

term frequency $freq_{i,j}$: how often does term k_i occur in document d_j?

normalized term frequency:

$f_{i,j} = freq_{i,j} / max_l\ freq_{l,j}$

28

Vector model

- How to define weights? tf-idf

document frequency : in how many documents does term k_i occur?

N = total number of documents

n_i = number of documents in which k_i occurs

inverse document frequency idf_i: log (N / n_i)

29

Vector model

- How to define weights for query?

recommendation:

$q = (w_{1,q}, \ldots, w_{t,j})$

$w_{i,q}$ = weight for term k_i in q
= $(0.5 + 0.5\ f_{i,q})$ x idf_i

30

Vector model

- Advantages
- term weighting improves retrieval performance
- partial matching
- ranking according to similarity

Disadvantage

- assumption of mutually independent terms?

31

Probabilistic model

weights are binary ($w_{i,j} = 0$ or $w_{i,j} = 1$)

R: the set of relevant documents for query q

Rc: the set of non-relevant documents for q

$P(R|d_j)$: probability that dj is relevant to q

$P(Rc|d_j)$: probability that dj is not relevant to q

$sim(d_j,q) = P(R|d_j) / P(Rc|d_j)$

32

Probabilistic model

$sim(d_j,q) = P(R|d_j) / P(Rc|d_j)$

(Bayes' rule, independence of index terms, take logarithms, $P(k_i|R) + P(\text{not } k_i|R) = 1$)

--> SIM(dj,q) ==

$$\sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times (\log(P(k_i|R) / (1- P(k_i|R))) + \log(P(k_i|Rc) / (1- P(k_i|Rc))))$$

33

Probabilistic model

- How to compute $P(k_i|R)$ and $P(k_i|Rc)$?
  - initially: $P(k_i|R) = 0.5$ and $P(k_i|Rc) = n_i/N$
  - Repeat: retrieve documents and rank them

  V: subset of documents (e.g. r best ranked)

  V_i: subset of V, elements contain k_i

  $P(k_i|R) = |V_i| / |V|$

  and $P(k_i|Rc) = (n_i - |V_i|) / (N - |V|)$

34

Probabilistic model

- Advantages:

- ranking of documents with respect to probability of being relevant

- Disadvantages:

- initial guess about relevance
- all weights are binary
- independence assumption?

35

IR - measures

$$\text{Precision} = \frac{\text{number of found relevant documents}}{\text{total number of found documents}}$$

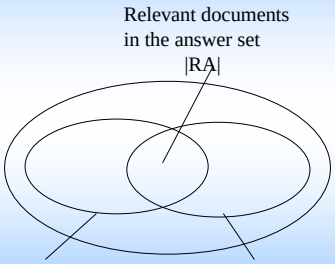
$$\text{Recall} = \frac{\text{number of found relevant documents}}{\text{total number of relevant documents}}$$

36

IR - measures

Precision =
|RA| / |A|

Recall =
|RA| / |R|



37

PERL

Practical Extraction and Reporting Language

```
while ($line=<INPUTFILE>) {
  if ($line=~ /^DEFINITION/ ) {
     ($head, $rest) = split("DEFINITION",$line);
     print OUTPUTFILE $rest;
     print OUTPUTFILE "\n";
  }
}
```

38

Storing textual information - Databases

- Relational databases:
  - model: tables + relational algebra
  - query language (SQL)
- Object-oriented databases
  - model: persistent objects, messages, encapsulation, inheritance
  - query language (e.g. OQL)
- Systems: GDB (R), ACEDB (OO)

39

Relational databases

PROTEIN

PROTEIN-ID	ACCESSION	DEFINITION	SOURCE
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human

DESCRIBED-IN

PROTEIN-ID	REFERENCE-ID
1	1
1	2

ARTICLE-AUTHORS

REFERENCE-ID	AUTHOR
1	Frielle
1	Collins
1	Daniel
1	Caron
1	Lefkowitz
1	Kobilka
2	Frielle
2	Kobilka
2	Lefkowitz
2	Caron

ARTICLE-TITLE

REFERENCE-ID	TITLE
1	Cloning of the cDNA for the human beta 1-adrenergic receptor
2	Human beta 1- and beta 2-adrenergic receptors: structurally and functionally related receptors derived from distinct genes

40

SQL queries

```
select source
from protein
where accession = NM_000684;
```

PROTEIN

PROTEIN-ID	ACCESSION	DEFINITION	SOURCE
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human

41

SQL queries

```
select title
from protein, article-title, described-in
where protein.accession = NM_000684
  and protein.protein-id
                = described-in.protein-id
  and described-in.reference-id
                = article-title.reference-id;
```

DESCRIBED-IN

PROTEIN-ID	REFERENCE-ID
1	1
1	2

PROTEIN

PROTEIN-ID	ACCESSION	DEFINITION	SOURCE
1	NM_000684	Homo sapiens adrenergic, beta-1-, receptor	human

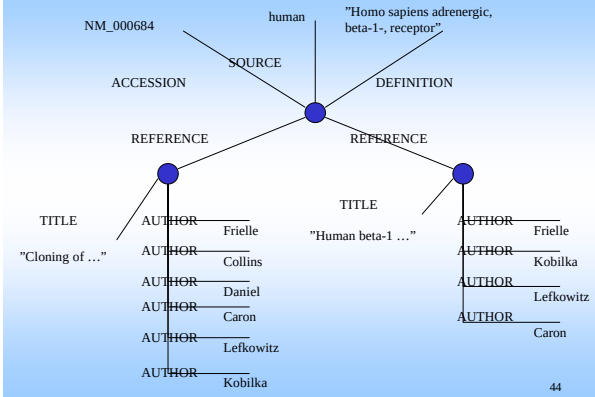
ARTICLE-TITLE

REFERENCE-ID	TITLE
1	Cloning of the …
2	Human beta 1- …

42

Storing textual information - Semi-structured data

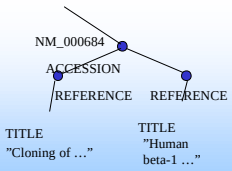
- Connected network of nodes
- query: path search
- systems: Genbank, OMIM



Queries

select source
from protein
where fills accession NM_000684;

select reference.title
from protein
where fills accession NM_000684;



Storing textual information - Knowledge bases

- Often based on a logic
- Query answering based on inference mechanism
- Knowledge bases often fit in main memory
- Useful for ontologies

Storing textual information - Knowledge bases

(F) source(NM_000684, Human)
(R) source(P?,Human) => source(P?,Mammal)
(R) source(P?,Mammal) => source(P?,Vertebrate)

Q: ?- source(NM_000684, Vertebrate)
A: yes

Q: ?- source(x?, Mammal)
A: x? = NM_000684

Outline

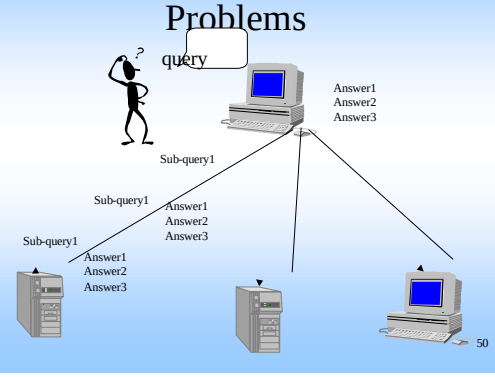
- Biological databanks
- Storing and accessing textual information
- Accessing multiple biological databanks
  - link-driven federations
  - view integration

Access to multiple databanks - Problems

- Users of biological databanks need to have good knowledge about where to find information and how to retrieve the information from the sources.
- Representations in different databanks may not conform or not be consistent with each other.

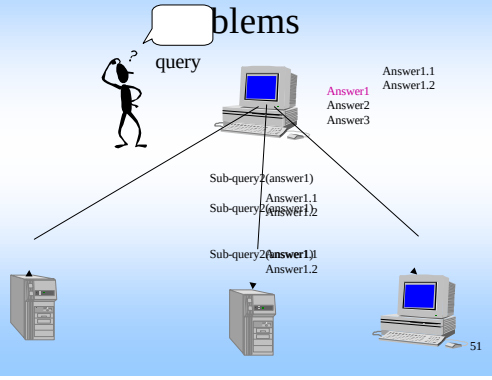
49

Access to multiple databanks - Problems



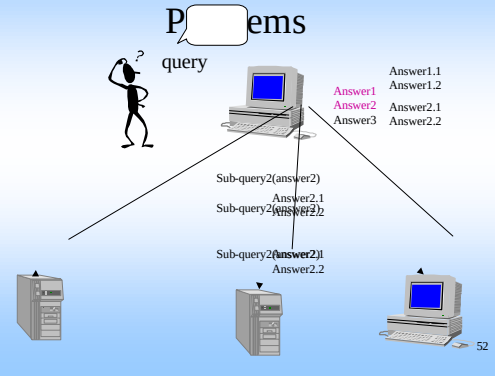
50

Access to multiple databanks - Problems



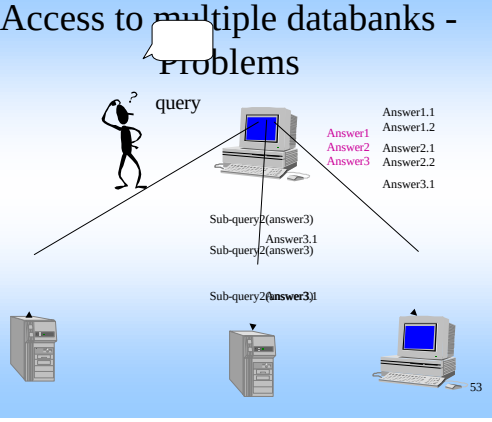
51

Access to multiple databanks - Problems



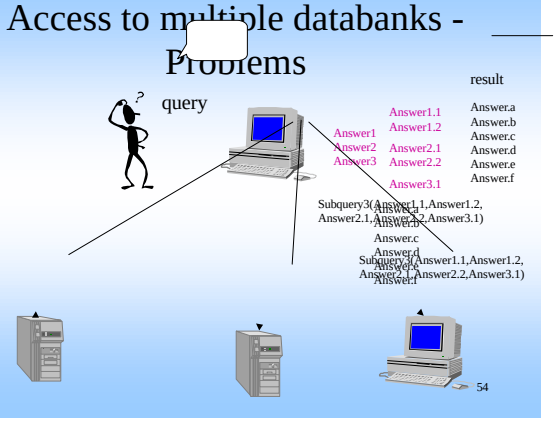
52

Access to multiple databanks - Problems



53

Access to multiple databanks - Problems



54

Current approaches

- Link-driven federations
- View integration

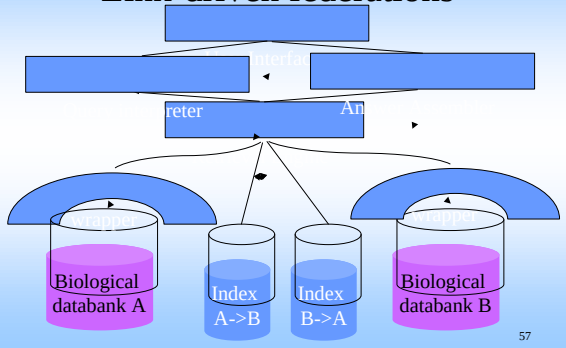
55

Current approaches

- Link-driven federations
  - create explicit links between databanks
  - query: retrieve interesting results and use web links to go to other related data sources
  - SRS, Entrez

56

Link-driven federations



57

Link-driven federations

- Advantages
  - complex queries
  - fast
- Disadvantages
  - good knowledge required
  - syntax-based
  - discrepancies in terminology not solved

58

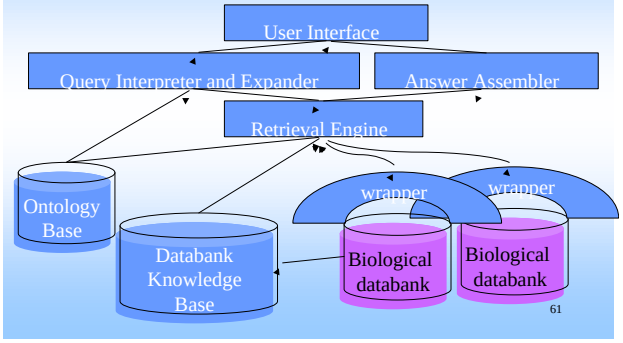
59

Current approaches

- View integration
  - define global schema over the databanks
  - query: using a high-level query language
  - BioKleisli, K2, TAMBIS, BioTRIFU

60

View integration



61

View integration

- Advantages
  - complex queries
  - good knowledge not required
  - solution for discrepancy issue
  - semantics-based

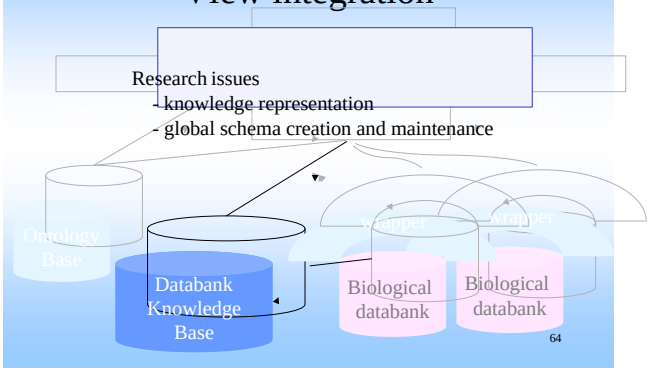
62

View integration

- Disadvantages
  - more computation
  - view maintenance

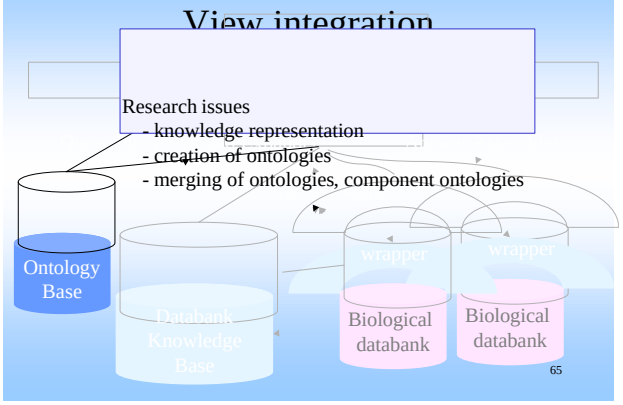
63

View integration



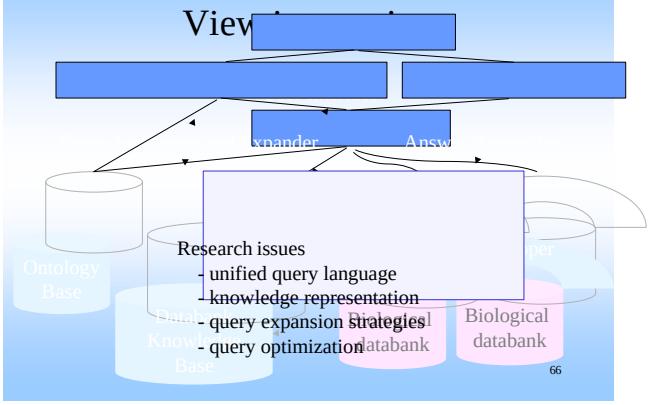
64

View integration



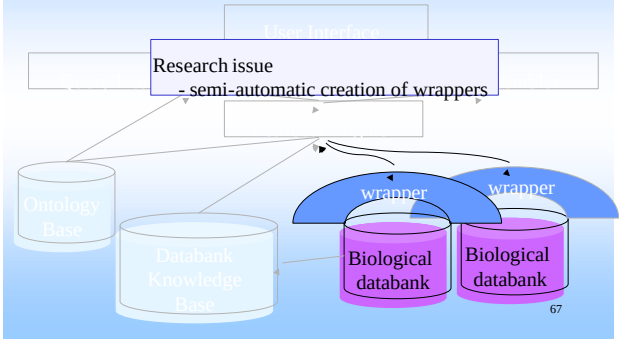
65

View integration



66

View integration



67

Related work at IISLAB Laboratory for Intelligent Information Systems

- Transparent access to multiple biological databanks
- Live help systems (web, ML)
- Extraction of information from text (IE,ML)
- TRIFU (KBIR)
- Description logics (KR)

68

69

Literature

- Information Retrieval

Baeza-Yates, R., Ribeiro-Neto, B., *Modern Information Retrieval*, Addison-Wesley, 1999.

Schwartz, R., Wall, L., *Learning Perl*, O'Reilly & Associates, 1994.

- Databases

Elmasri, R., Navathe, S., *Fundamentals of database systems*, 3rd edition, Addison Wesley, 2000.

70

Literature

- Semi-structured data

Abiteboul, S., Quass, D., McHugh, J., Widom, J., Wiener, J., 'The Lorel query language for semi-structured data', *International Journal on Digital Libraries*, 1(1):68-88, 1997.

Lambrix, P., Shahmehri, N., 'Querying documents using content, structure and properties', *Journal of Intelligent Information Systems*, 15(3):287-307, 2000.

71

Literature

- Multiple Databanks

SRS Users manual version 6.07

Davidson, S., Overton, C., Buneman, P., 'Challenges in integrating biological data sources', *Journal of Computational Biology*, 2:557-572, 1995.

72

Literature

- Multiple Databanks

Davidson, S., Crabtee, J., Brunk, B., Schug, J., Tannen, V., Overton, C., Stoeckert, C., 'K2/Kleisli and GUS: Experiments in integrated access to genomic data sources', *IBM Systems journal*, 40(2):512-531, 2001.

Goble, C., Stevens, R., Ng, G., Bechhofer, S., Paton, N., Baker, P., Peim, M., Brass, A., 'Transparent access to multiple bioinformatics information sources', *IBM Systems Journal*, 40(2):532-551, 2001.

Assignment

Choose one of:

- Challenges in integrating biological databanks
- Compare integration systems
- own topic - literature
- own topic - implementation