A Protein Ontology from Large-scale Textmining?

Fraunhofer Institut
Algorithmen und Wissen-
schaftliches Rechnen

Protege-Workshop Manchester, 07-07-2003

Kai Kumpf, Juliane Fluck and Martin Hofmann

Instructive mistakes: a narrative

•Aim: Protein ontology that supports reasoning about and curation of protein interaction data created from text mining

•What do we have? Gene ontology, protein repositories (database schemas and class diagrams) and protein databases covering various aspects of proteins

•What does a typical protein repository look like? And why that does not suffice.

•What data do we have at our disposal

•Protein classification

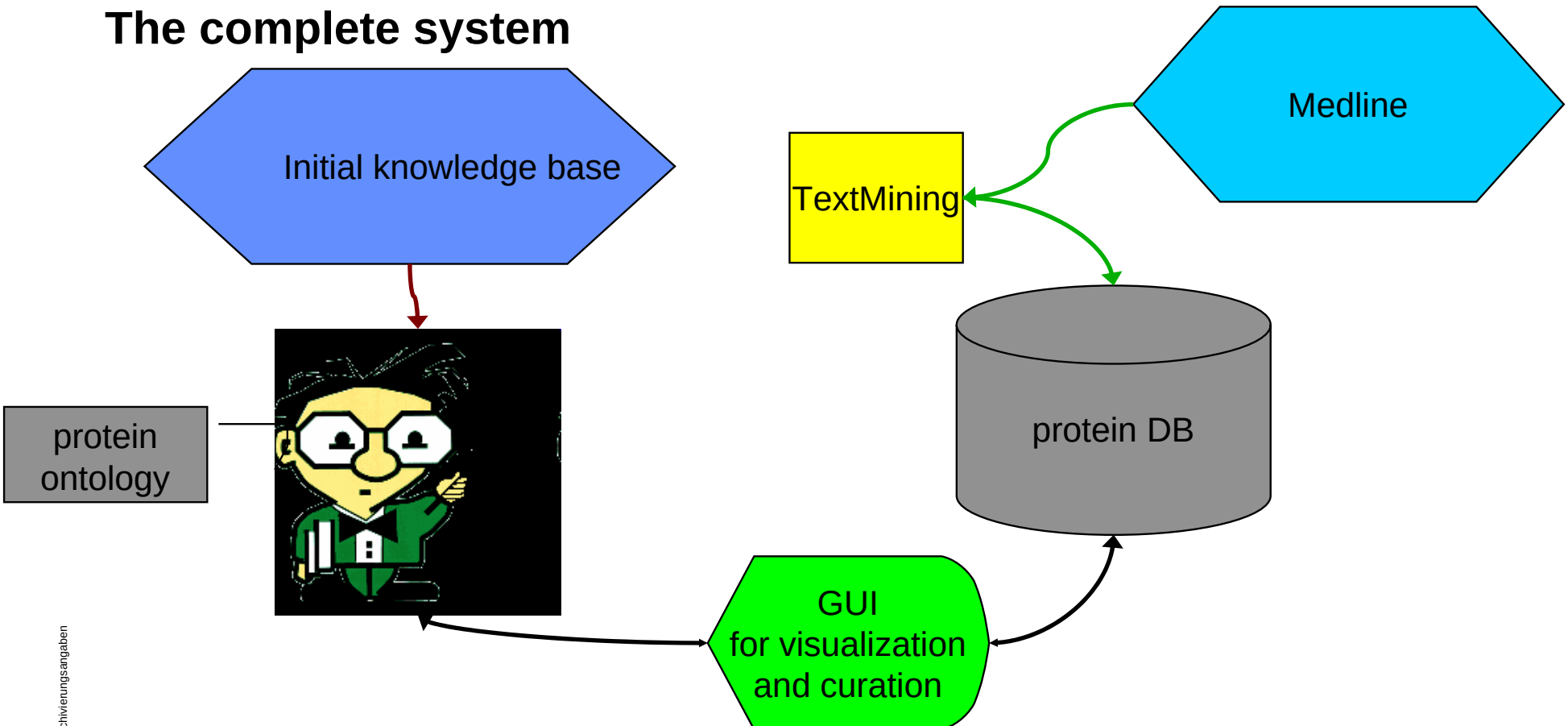
•… and how this can lead into trouble

•Protein families as a (last?) resort

•A temporary conclusion

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

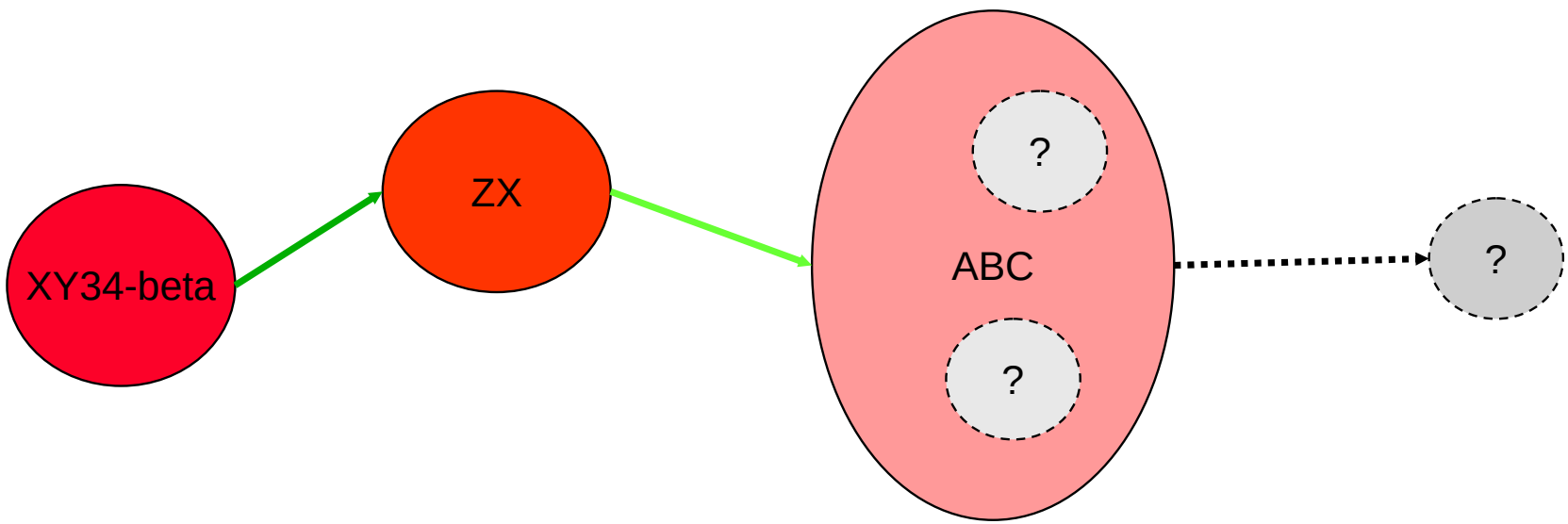
The complete system



Fraunhofer SCAI Institut
Algorithmen und Wissen-
schaftliches Rechnen

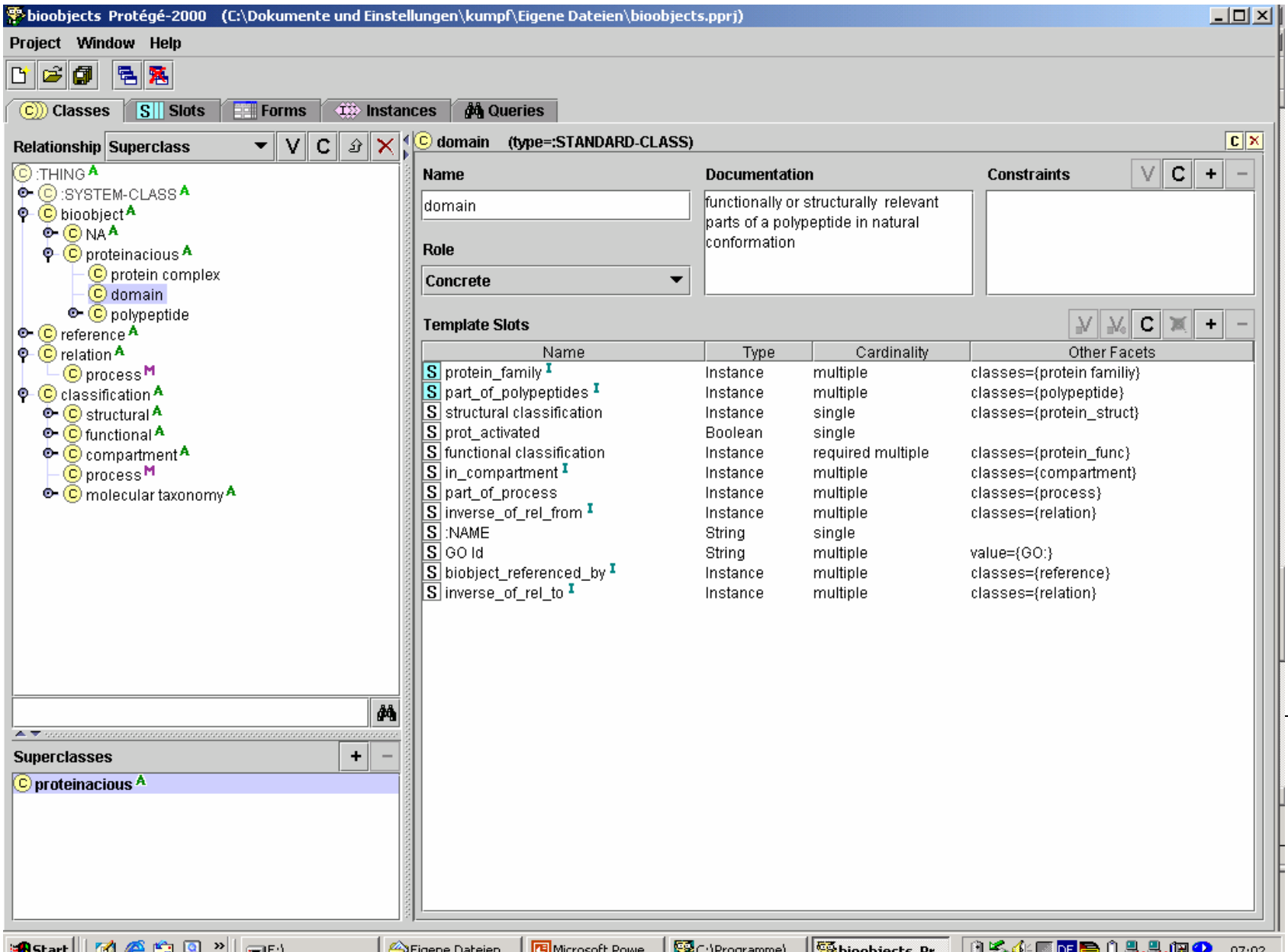
Text mining results

We found that XY34-beta phosphorylates ZX, thus triggering signal transduction via the ABC pathway.



Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

The ontology framework which isn't



What‘s wrong with this picture?

This is an ontology modeling the structure of a protein database instead of a protein ontology.

The modelers stopped short of starting a real, *biological* knowledge base.

Different functional types of proteins have to be included as classes, instead of instances.

!

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Protein ontologies

•What do we have at our disposal? No protein ontology, to start with.

- GO, gene ontology
  - distinguishes function, process compartment
- Biological pathway DBs: (aMAZE), KEGG, BioCyc, BIND, DIP, WIT, biopax
- Protein family DBs: InterPro,SMART, PFAM, PROSITE, BLOCKS, PRINTS, CATH
- General protein DBs: SWISS-PROT...
- SwissProt Keywords - ~ 880
- MEDLINE abstracts and MESH headings (protein relevant) - ~ >1000.000

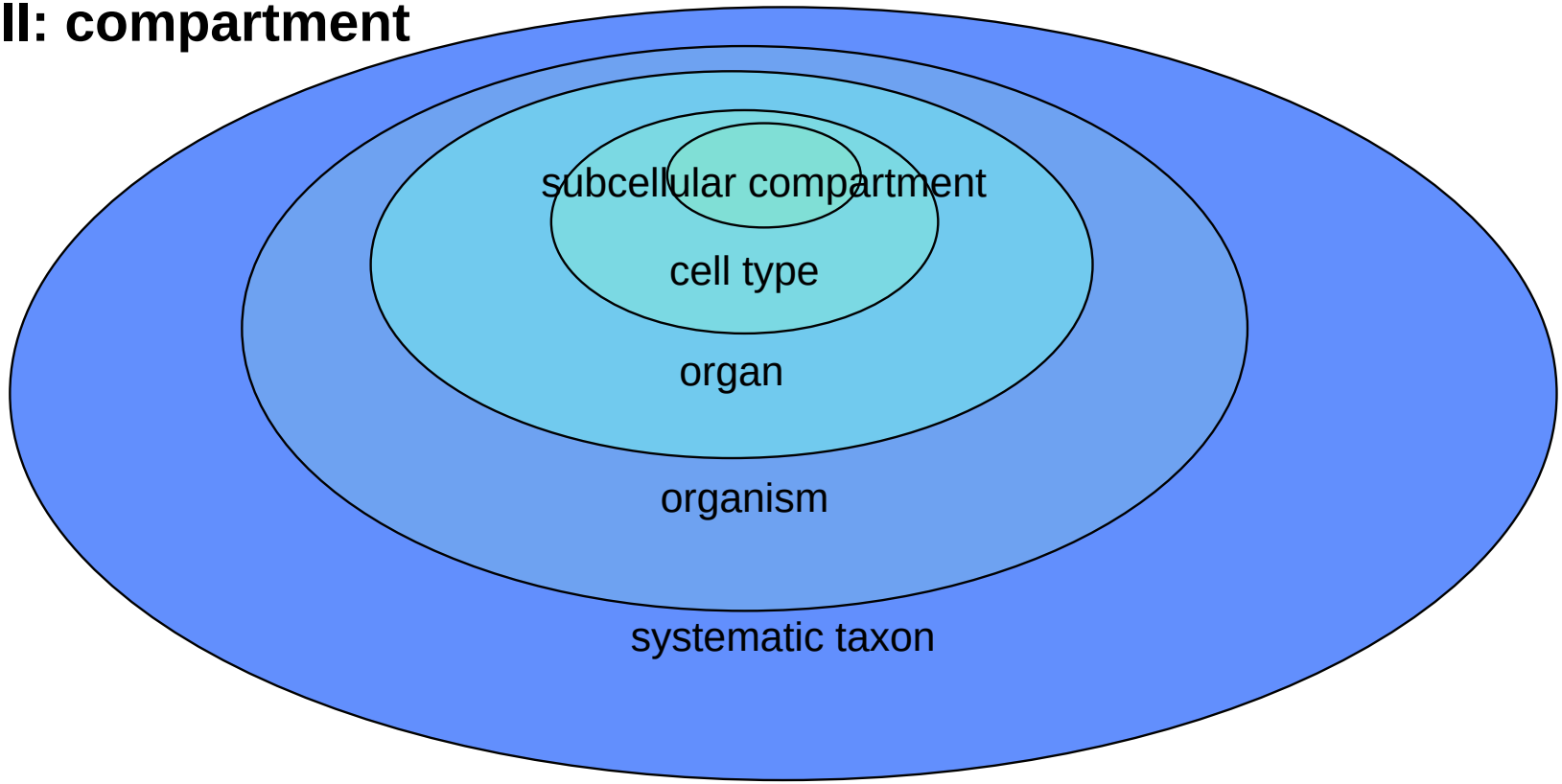
Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

How can proteins be characterized? I: structure / function / process

- **Structure** (SCOP): alpha/beta/mixed, ...
- **Function**: enzymes, non-enzymes / structural proteins
    - enzymes distinguishable by EC classification: Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, Ligases
    - But: this is a chemists view of proteins. Distinction is too coarse by far.
- **Processes** == Metabolic Pathways
- Compartments: Where do the proteins act, where are they modified, do they act in different compartments simultaneously, serially? See next page.

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

How can proteins be characterized?
II: compartment



Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

The great protein confusion

- **Biological** side
  - Even one domain does not always fulfill the same functions: activated/inactivated forms,
  - complexes can behave differently from their constituent parts
- **Database** side
  - long list of synonyms and different IDs / accession no.s
    - Proteins can be named differently in different stages of their life cycle
    - Some modifications are not even listed with names
  - relations, interactions, functions as free text descriptions
    - text mining?!
  - GO 2 InterPro, SwissProt keywords, etc. looks promising, but GO mixes up hierarchical levels (e.g. psychological vs. molecular processes)
    - solution: GONG and GOAT?

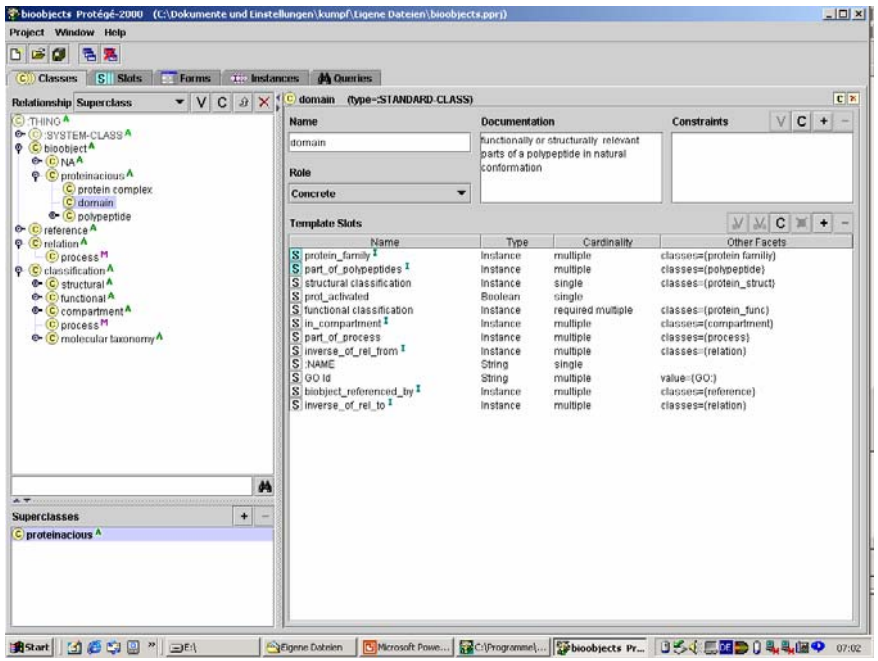
Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Protein families jump to rescue

- Families are built from and capture *more than just on part of the* essential classification features.
- Protein functions can often be deducted from their major domains.
- Domains/Proteins don't come in a lot of flavors.
  - ❑ Families appear to constitute a "natural" taxonomy, more homologous proteins with similar sequence than analogous proteins
- Families can be constructed from real evolutionary relationships (PIR Superfamily DB)
- Families can be modified so as to incorporate more classification features
  - ❑ New clusters or other cluster combinations might result from extending the feature space

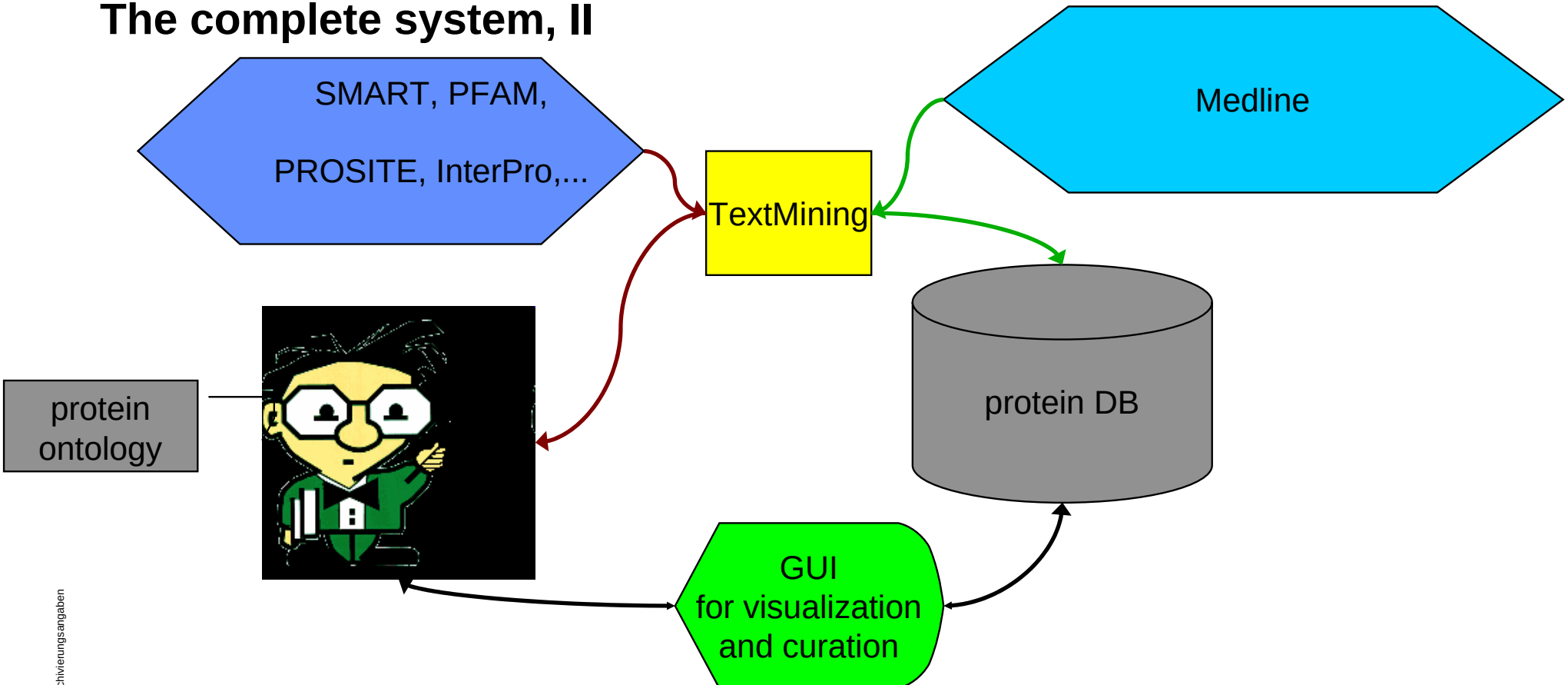
The ontology framework which could be what needs to be done?

1. Choose the family classifications that appear relevant as the domain knowledge basis
   1. Individual function
   2. Pathways
2. Build a proper ontology 8-)
   1. Bottom level classes are still families
   2. The instance level is individual, concrete protein names
3. Let text mining engine churn out relationships
4. Add to ontology, assigning probabilities / plausibility measures on the way, based on what is already there, iteratively
   1. Either we know the family of the protein beforehand and check for plausibility of interactions or
   2. The family is unknown, then we need to reason about plausible kinship



Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

The complete system, II



Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Is that all there is to it?

Obviously: **NO**.

What then *is* the normative protein family classification?

A definitive answer is still pending.

Thank you for your kind attention.

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen