

Information extraction using/finding terminology (from ontologies ??)

Summer School, Balatonfured

July 6th, 2004

Dietrich Rebholz-Schuhmann, MD, PhD
Group Leader Rebholz Group
EBI, WT Genome Campus
Hinxton, Cambridge, Uk



Goals:

- Identification and extraction of facts
- Normalization of facts
- Information extraction on-the-fly:
have a server solution ready which is fast and identifies different sets of normalized facts
- In principle process any type of text
- Support to curation teams
- Combine of information retrieval and information extraction



Type of facts we are looking for:

- Acronyms, e.g. HZF-1
- Descriptive names, e.g. thyroid hormone receptor, clones of humans
- Formalized facts, e.g. mutations (Cys/Val343)
- Associations and relationships between biological objects, e.g. cellular location of proteins

Issues of continuous threat:

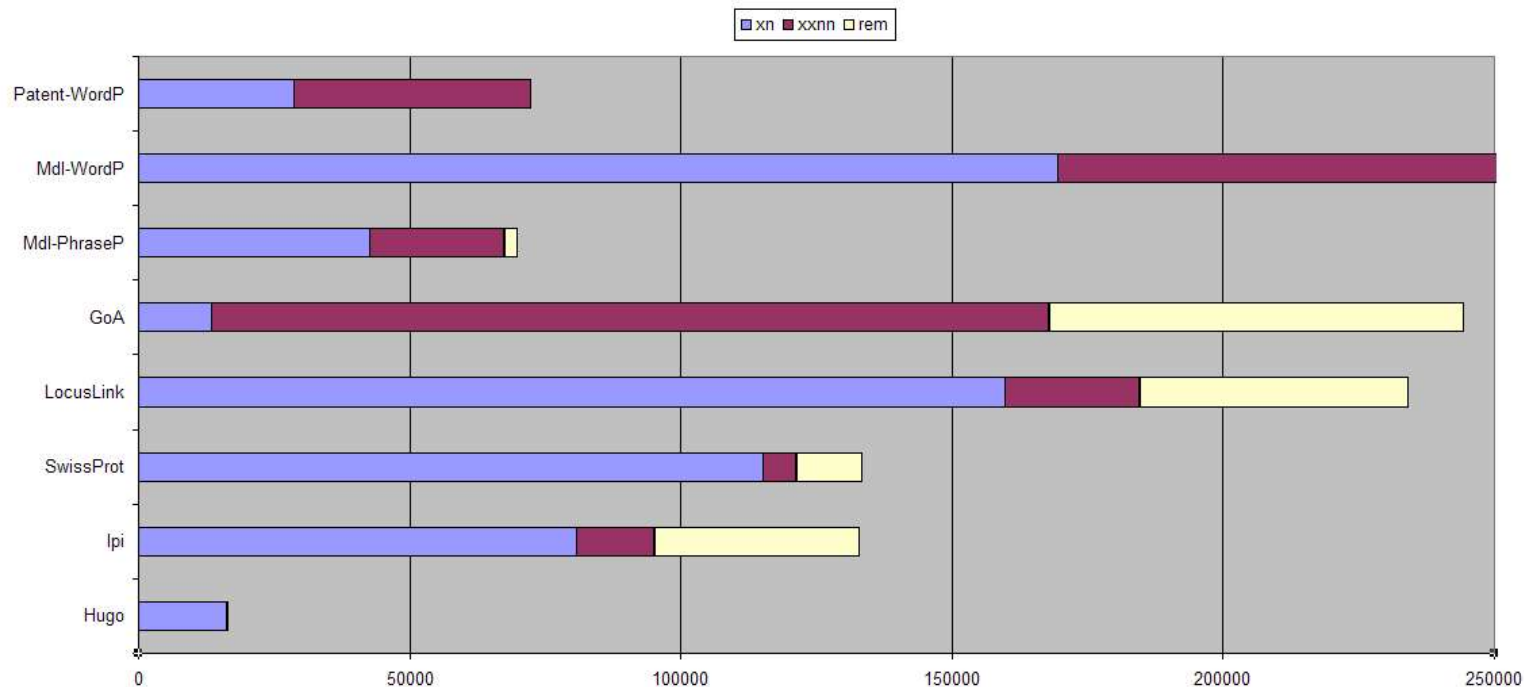
- Ambiguities are frequent, e.g. A2M
- Inconsistencies with biological databases



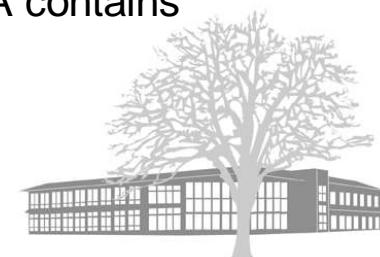
Collection of terms:

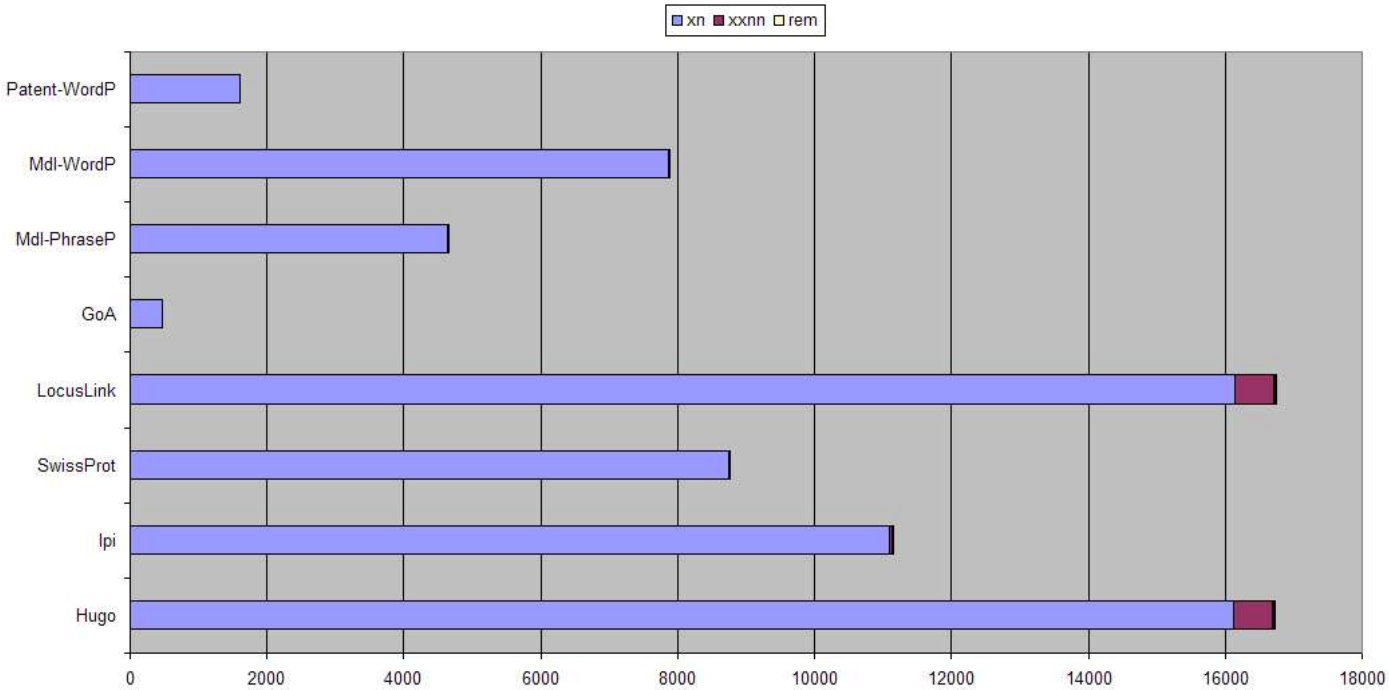
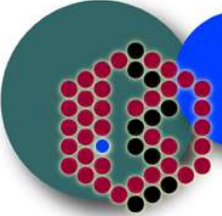
- Resources are Hugo, Ipi, Swiss-Prot, LocusLink, GoA
- Different categories:
 - single terms vs. multi word terms
 - XN = Single terms (characters, digits)
 - XXNN = Multi word terms (characters, digits)
 - REM = containing special characters
- 4 Patterns which explicitly identify PGNs in the text (PhraseP) in Medline and Patents
- Generalized word patterns out of single term PGNs





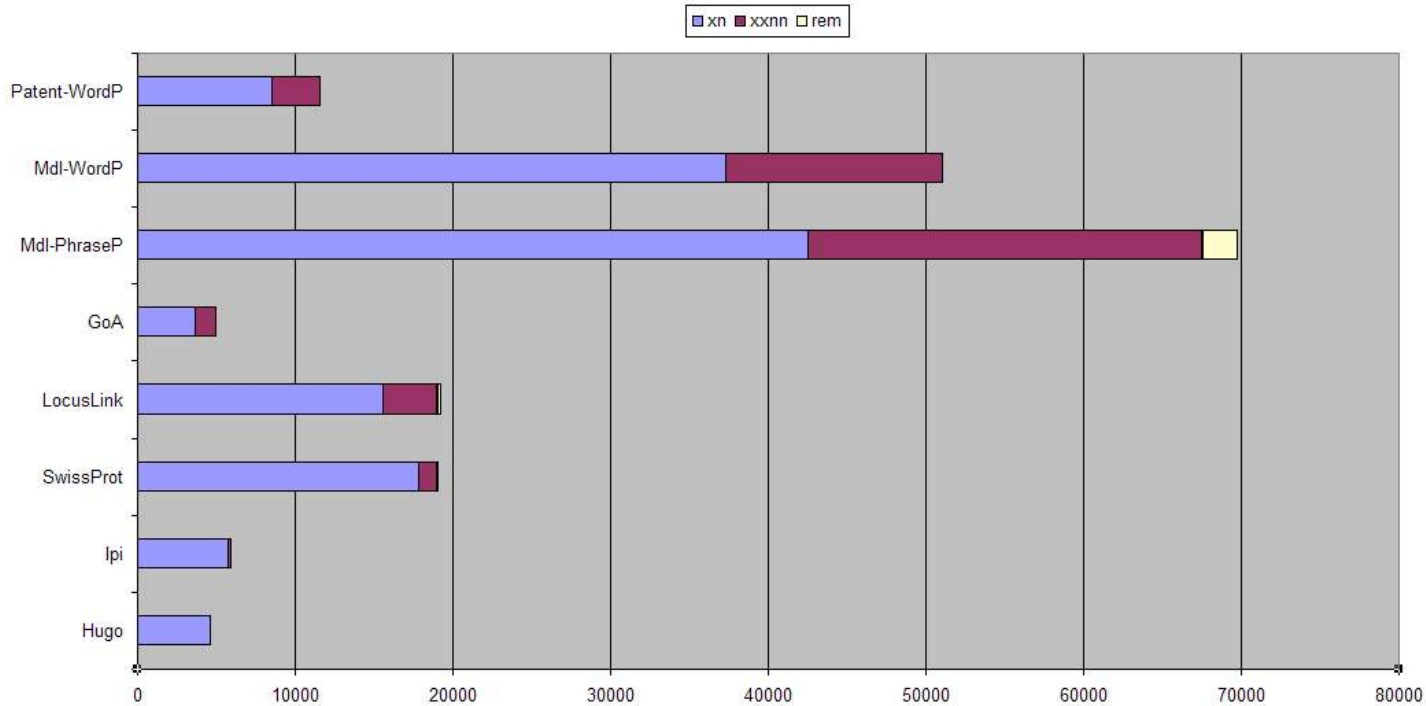
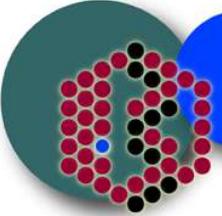
- The number of unique PGNs (in lowercase only) in Hugo, Ipi, Swiss-Prot, LocusLink, GoA, and in the sets of PGNs from Medline (phrase patterns: Mdl-PhraseP; word patterns: Mdl-WordP), and from the European patent abstracts (Patent-WordP) is shown. GoA is clearly different from the other databases, since GoA contains mainly descriptive names.





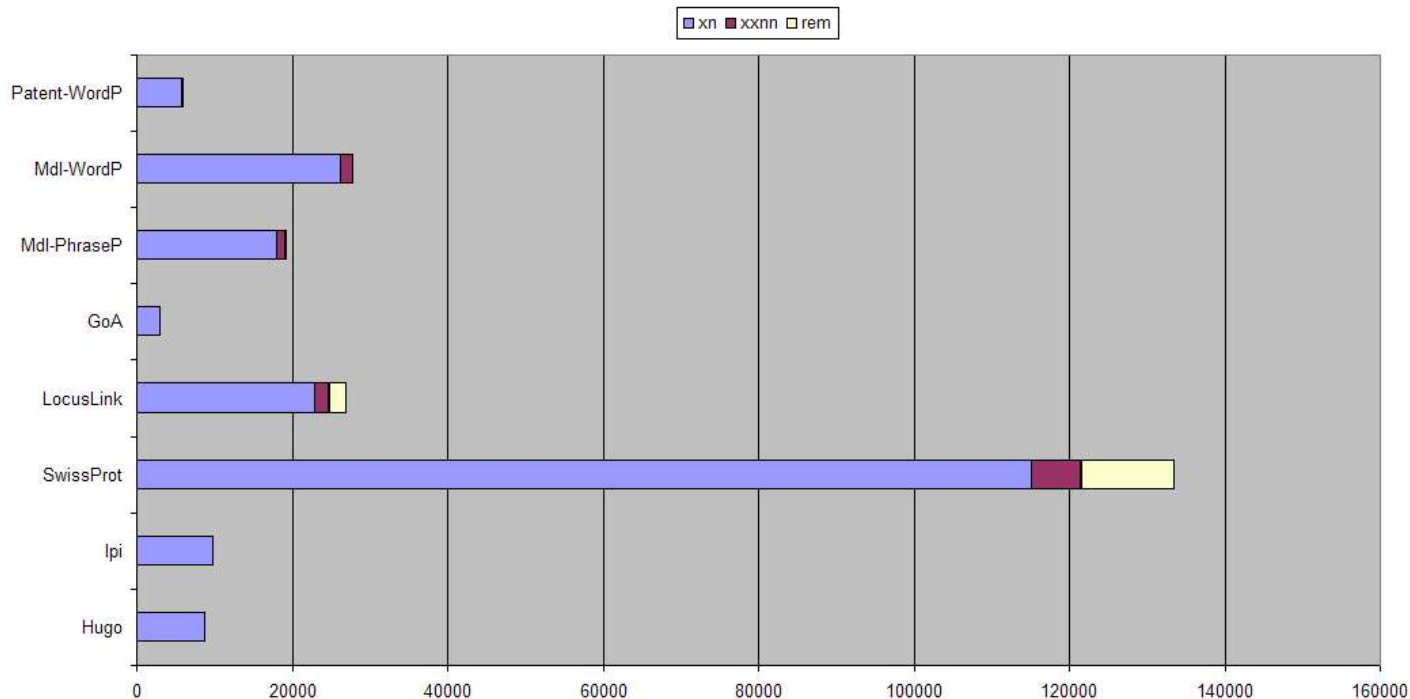
- **(Comparison of PGNs of Hugo to other data resources)** The figure shows the number of PGNs from Hugo, which have been identified in the other sources. All PGNs have been transformed into their lowercase representation. The entries of Hugo are contained in LocusLink, IPI and Swiss-Prot. At most half of the entries of Hugo can be found in Medline with the help of word patterns (Mdl-WordP).



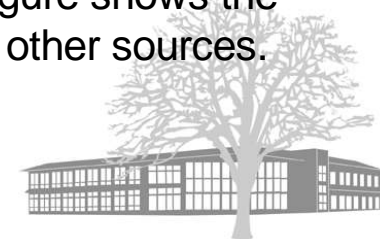


- **(Comparison of PGNs from Medline to other sources)** The figure shows the number of PGNs from Medline (phrase patterns), which have been identified in the other sources. The coverage of the PGNs from the word patterns in comparison to LocusLink and Swiss-Prot is even higher than the coverage of the Hugo PGNs (ref. to Fig. 3). The coverage for the European patent abstracts is as well high.





- **(Comparison of PGNs from Swiss-Prot to other sources)** The figure shows the number of PGNs from Swiss-Prot, which have been identified in the other sources. A large number of PGNs do not reappear in the literature.



- Different term representations
 - NF kappa B, NF kB, nuclear factor kappa B
 - TIF2, TIF-2, transcription intermediary factor-2, transcriptional intermediate factor 2
 - Parathyroid Cell calcium-sensing receptor, Calcium-sensing receptor, calcium sensing receptor, Ca⁺ receptor, calcium receptor
 - 5-hydroxytryptamine 1A receptor, 5-HT-1A, Serotonin receptor, 5-HT1A, HTR1A, 5HT1A
- Orthographic variation
 - Hyphens, slashes: amino acid and amino-acid
 - Lower / upper case: NF-KB and NF-kb
 - Spelling variations: tumour and tumor
 - Latin/Greek transcriptions: oestrogen and estrogen
 - Acronyms: RAR and retinoic acid recptor



- Higher complexity variation:
 - Different reductions:
thyroid hormone receptor and thyroid receptor
SB2 gene and SB2
 - Embedded variants forms:
CBR is CREB-binding protein, and
CREB is cAMP-response element-binding protein
 - Overabstraction: dystrophin (muscular dystrophy,
Duchenne and Becker types) holds for DXS143, DXS164,
DXS206, DDXS 230, DXS239, DXS268, DXS269,
DXS270, DXS272
- Term ambiguity
 - Genes such as: an, by, can, for, with, we



Identification of GO terms

- 3 main categories:
 - molecular function (of an entity/concept),
 - biological process (a concept)
 - cellular component (an entity)
- Node: terms / terminology
- Edges: semantic relations, e.g. `is_kind_of`, `is_part_of`
- Computation based on the graph:
 - nearness of terms
 - similarity of terms
 - Identify term dependence, e.g. substring relationship



McCray/Browne/Bodenreicher: “The Lexical Properties of the Gene Ontology (GO)”, NLM (AMIA 2002)

- Molecular function (5626 terms), biological process (4677 terms), cellular component (1077 terms)
- The longer a term ~ the less likely to be found in text
- The more non-alphanumeric characters ~ the less likely to be found in text

	Molecular Function	Biological Process	Cellular Comp.	Total Terms
GO strings	5626	4677	1077	11380
In UMLS	2436	256	370	3062 (27%)
Passed NLP filter	4338	3730	907	8975 (79%)
In corpus	2125	1318	570	4013 (35%)
Full term in lexicon	636	166	204	1006 (9%)
	Molecular Function	Biological Process	Cellular Comp.	
MeSH	2269	164	331	
SNOMED	1119	86	161	



- Task 2a: “**Recover**” **text** that provides evidence for the associated GO annotation of the associated protein (both known from the databases)
 - RegExp matching
 - Matching of phrase patterns from GO to the text
 - Scoring of single terms over a sliding window
 - Calculate Levenstein distance between GO term and text
 - Index GO terms and use document as Query (Vector space model)
 - Trained Naïve Bayes classifiers
 - Results are overall similar: 10% of terms identified at 30% precision, remaining 90% at less than 10% precision
- Task 2b: **Identify GO terms from the text** (+ provide this evidence) in conjunction with a contained PGN (GO annotation in the text)
- Task 2c: **Select papers** which appear to be suitable to give a



5275	N N N	143	ADJ ADJ N N
4507	N N	140	N N (N)N
1287	N	140	N ADJ N
1209	ADJ N N	137	ADJ N N N N
752	N N N N	135	N PREP ADJ N
734	ADJ N	135	ADJ N PREP N
406	ADJ N N N	110	ADJ ADJ N
400	ADJ N PREP N N	106	N N N N N
312	N PREP N N	100	ADJ N PREP ADJ N N
180	N ADJ N N	99	ADJ N PREP ADJ N
167	N PREP N	92	N PREP ADJ N N
167	N N PREP N	92	ADJ N PREP N N N
153	EN N		

Analysis of GO terms

- N combinations the most frequent
- Adj N combinations at second place
- Prepositional phrases optional semantic relations
- Special characters at low frequency



1413822	N N	42050	ADJ ADJ N N N
726689	N N N	36011	DET N N N
675261	ADJ N N	34368	DET ADJ N N
302340	ADJ N N N	32888	N N N N N
223185	ADJ N	25564	N N ADJ N
168443	N N N N	20817	DET ADJ N
92701	ADJ ADJ N N	18962	N ADJ N N N
61791	ADJ ADJ N	18946	ADJ N ADJ N
58378	ADJ N N N N	17182	N N ADJ
58293	DET N N	17020	EN N N N
51961	N ADJ N	16084	DET ADJ N N N
51779	EN N N	15818	ADV ADJ N N
47886	N ADJ N N	14771	ADJ N ADJ N N

Analysis NPs of combinations of terms from Medline containing at least one GO noun term

- N combinations the most frequent
- Adj N combinations at second place
- BUT: difference in length and in pattern
- Terms tend to be longer



* map kinase * activity

GO terms concerned with MAP kinase

MAP kinase phosphatase activity
 MAP kinase activity
 MAP kinase 1 activity
 MAP kinase 2 activity
 MAP kinase kinase kinase activity
 MAP/ERK kinase kinase activity
 MAP kinase kinase kinase kinase activity

MAP kinase kinase activity
 MAP-kinase scaffold activity
 MAP-kinase anchoring activity
 activation of MAP/ERK kinase kinase
 MAP-kinase scaffold protein activity
 MAP-kinase anchor protein activity
 cytoplasmic translocation of MAP kinase

Exact matches in Medline can be found as follows:

map kinase activity
 map kinase kinase activity
 map kinase kinase kinase activity

map kinase phosphatase activity
 map kinases activity

A selection of noun phrases which contain a match to MAPK:

adhesion dependent map kinase activity
 alf4 induced map kinase activity
 attenuated oxytocin induced map kinase activity
 endothelin 1 stimulated p38 map kinase activity
 enhanced insulin induced map kinase activity
 epidermal growth factor stimulated map kinase activity
 map kinase phosphotransferase activity



* aspartate * activity

The following GO terms have been identified via an exact match.

3 hydroxyaspartate aldolase activity
aspartate aminotransferase activity
aspartate ammonia lyase activity
~~aspartate carbamoyltransferase activity~~

aspartate kinase activity
aspartate oxidase activity
aspartate racemase activity
aspartate transaminase activity

The following concepts from Medline are not known to GO and refer to **enzymes**. The identified named entities can be classified according to the ending ' -ase' as enzymes.

aspartate 4 decarboxylase activity
aspartate aminopeptidase activity
aspartate beta semialdehyde dehydrogenase activity
aspartate dehydrogenase activity
aspartate phosphotransferase activity
aspartate transcarbamoylase activity
n acetylaspartate aminohydrolase activity
l aspartate:quinine oxidoreductase activity



The following list entries can be found, which refer to named entities, which are obviously no enzyme activities, but refer to the **activity of (a complex of) protein(s)**. Does GO provide additional information to classify the encountered terminology automatically?

```
aspartate antagonist activity
aspartate channel activity
aspartate induced channel activity
adp induced malate aspartate shuttle activity
malate aspartate reduced nicotinamide adenine dinucleotide shuttle activity
malate aspartate shuttle activity
```

Receptor refers obviously to a protein. Again, does GO provide additional information to classify the encountered terminology automatically?

```
aspartate receptor activity
aspartate receptor antagonist activity
aspartate receptor antagonist ap5 increased activity
aspartate receptor controlled motor activity
aspartate receptor mediated spontaneous activity
aspartate receptor mediated synaptic activity
prolonged aspartate receptor mediated activity
prolonged aspartate receptor mediated postsynaptic activity
```



Physiological parameters (phenotypes)

Generalized patterns from from a list of

- "abnormal * B cell morphology * development"
 - "abnormal * T cell physiology"
 - "abnormal * gland morphology"
 - "increased * levels of * hormone"
 - "increased number of * cells"
 - "decreased * * * level"
 - "abnormal * * system physiology"
 - "abnormal * * of * morphology"
 - (continued)
-
- RNase protection assays demonstrated that Epo or dimethyl sulfoxide induction [increased steady-state mRNA levels](#) 10- to 20-fold after 24 to 48 hours
 - However, although superoxide dismutase [decreased superoxide anion levels](#) in the presence of L-NAME or in endothelium-denuded rings, it no longer inhibited the tone
 - Insulin-like growth factors, their receptors, binding proteins, and binding protein proteases are important in normal and [abnormal ovarian follicle development](#).

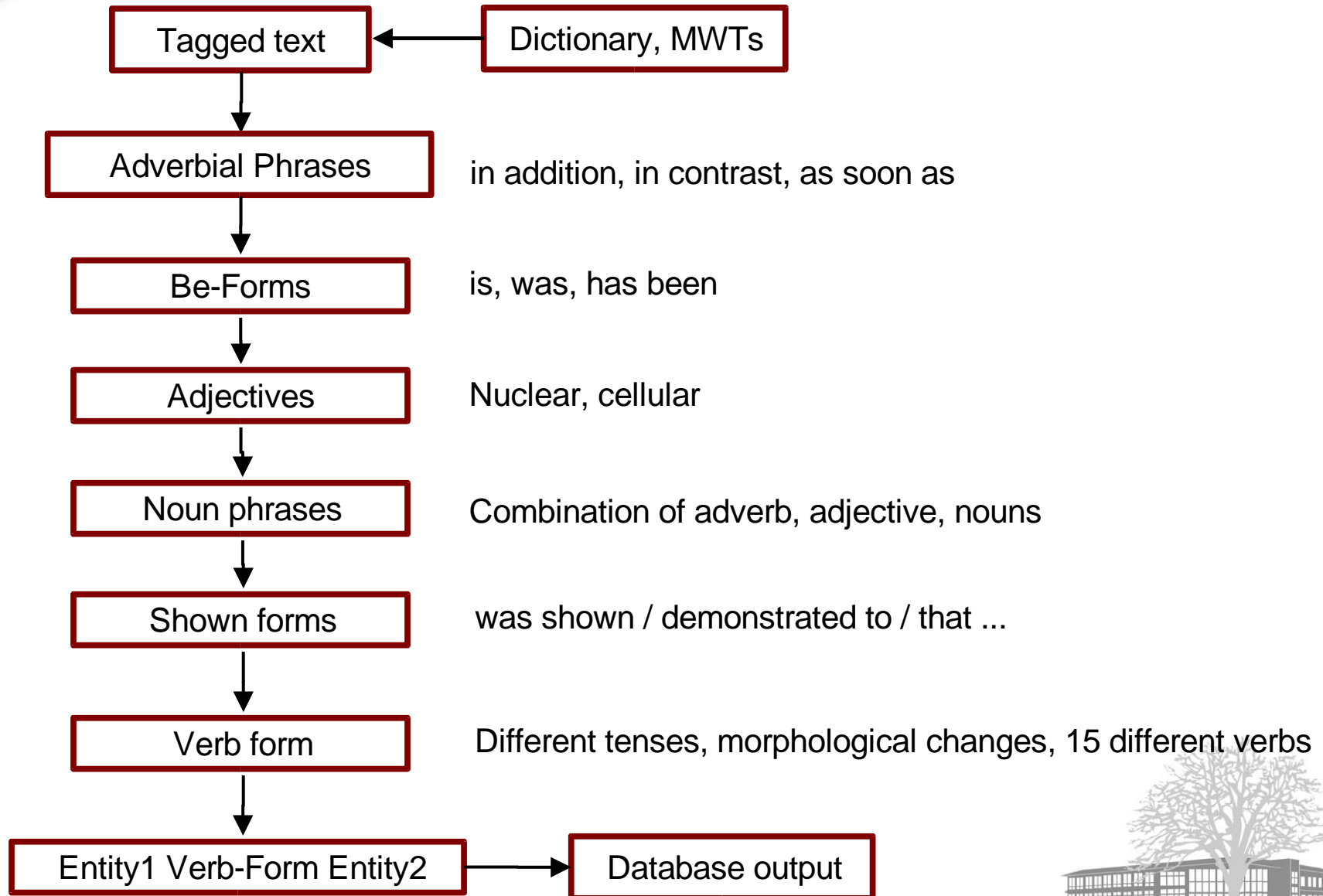


- Arg506 to Gln
- valine 804-->leucine
- Ile15 to Thr15
- Pro12Ala
- arginine(3500)---->glutamine
- C282Y
- A1166-->C
- 677C-->T
- 1166A/C,
- 359 (Ile/Leu)
- nucleotide 383T-->C
- codon 113 and His-->Arg
- Cys/Val343
- Val-->Ala at codon 113
- IVS1-2A-->G
- codon 241 and codon 247, where the single base changes from C to T
- methionine to threonine substitution at residue 235
- methionine for valine at position 30
- Ser-->Leu change at amino acid 217
- Heterozygosity for the IVS-I-5 (G-->C) mutation
- A fourth mutation, 433 --2(A-->G) transition, was identified at the splice-acceptor site in intron 2



- Dictionaries of terms
- Complex term identification
 - Automatic term recognition
 - Matching of complex terms
- Tagging tools
- Cascaded Finite State Transducer as core parsing technology
- Disambiguation tools
- Postprocessing of information
- Viewing tools to improve the workflow





- List of verbs used

assemble, attach, bind, complex, contact, couple, dimerize, dissociate, inhibit, interact, link, precipitate, regulate

BUT: *associate* too unspecific (inactivated in server)

- Negative results can be extracted/indicated, but this leads to all kind of conflicts:

- How to deal with double negation ?

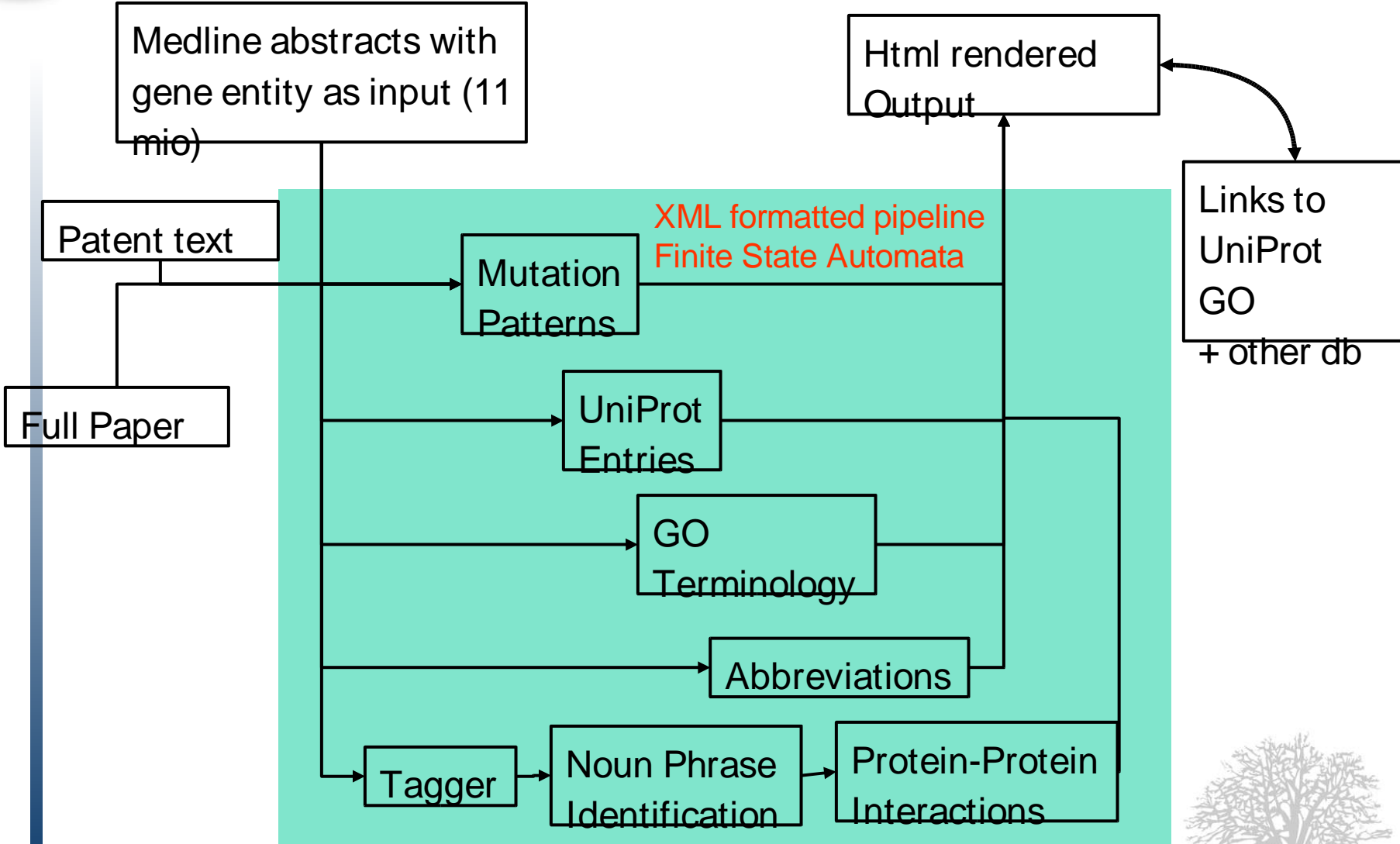
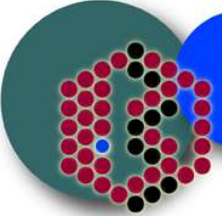
- How to deal with ‘ does not inhibit’ ?

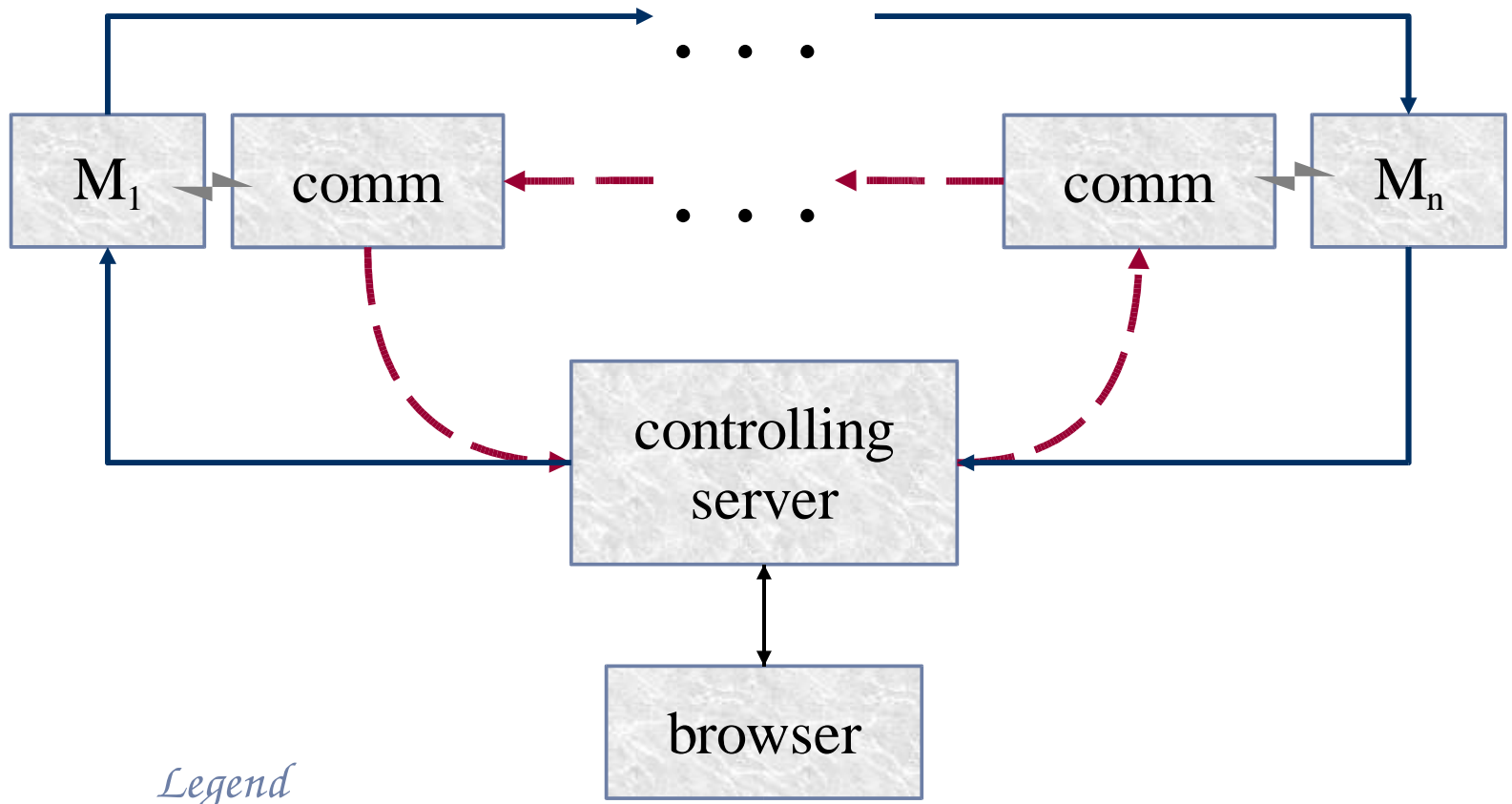
- Others claim to use 94 verbs:

activate, abolish, accelerate, alter, amplify, assemble, associate, attach, augment, bind, block, catalyze, complex, conjugate, down-regulate, enhance, express, form, induce, inhibit, interact, ligate, link, mediate, modify, promote, regulate, stimulate, suppress, synthesize, target, trigger, up-regulate, ubiquitinate



But also: infect, localize, prevent, modulate



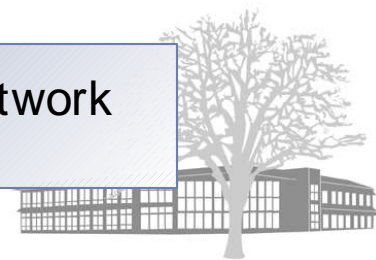




Legend

- M_i module
-  request
-  data

- data travels **n+1** times over the network
- modules **do** work in parallel



-WHATIZIT-

Whatizit can tell you the meaning of words found in your text, depending on the kind of information you want to see highlighted.

What kind of information do you want to be highlighted in the text? Select one of these:

Whatiz the text you want to highlight?

Nuclear receptors (NRs) complexed with agonist ligands activate transcription by recruiting coactivator protein complexes. In principle, one should be able to inhibit the transcriptional activity of the NRs by blocking this transcriptionally critical receptor-coactivator interaction directly, using an appropriately designed coactivator binding inhibitor (CBI). To guide our design of various classes of CBIs, we have used the crystal structure of an agonist-bound

Submit

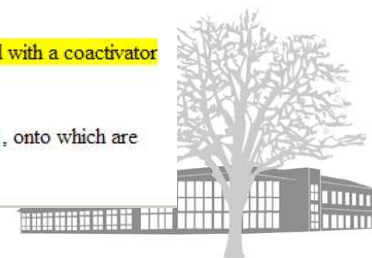
Result:

Nuclear receptors complexed with agonist ligands activate transcription by recruiting coactivator protein complexes

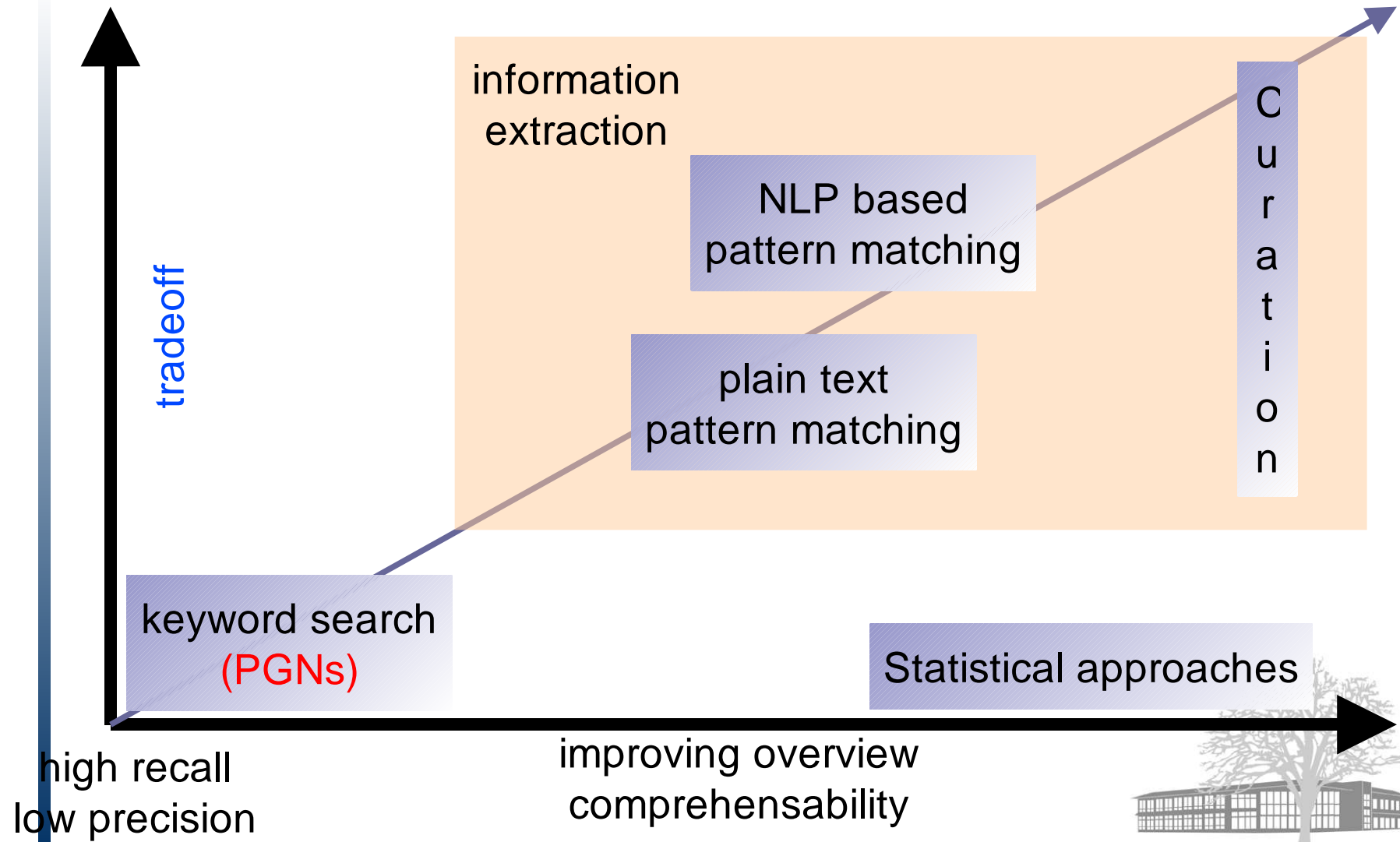
In principle , one should be able to inhibit the transcriptional activity of the NRs by blocking this transcriptionally critical receptor-coactivator interaction directly , using an appropriately designed coactivator binding inhibitor

To guide our design of various classes of **CBIs** , we have used the crystal structure of an agonist-bound estrogen receptor ligand binding domain complexed with a coactivator peptide _{p₂} containing the LXXLL signature motif bound to a hydrophobic groove on the surface of the LBD

One set of CBIs , based on an outside-in design approach , has various heterocyclic cores that mimic the tether sites of the three leucines on the peptide helix , onto which are appended leucine residue-like substituents



low recall
high precision



	Recall	Precision
cited SNPs in 1 letter code	151 / 191 (79,1 %)	151 / 152 (99.3 %)
cited SNPs in 3 letter code + fullname	52 / 78 (66,7 %)	52 / 54 (96,3 %)
cited SNPs	204 / 273 (74,7 %)	204 / 207 (98,6 %)
contained SNPs	143 / 190 (75,3 %)	143 / 146 (97,9 %)
contained SNP-gene pair	57 / 162 (35,2 %)	57 / 61 (93,4 %)

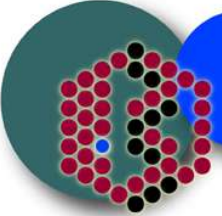
- Cited: Several citations in the same abstract are counted separately.
- Contained: Several citations in the same abstract are counted only once.
- No phrase patterns available for the association between SNPs and genes.



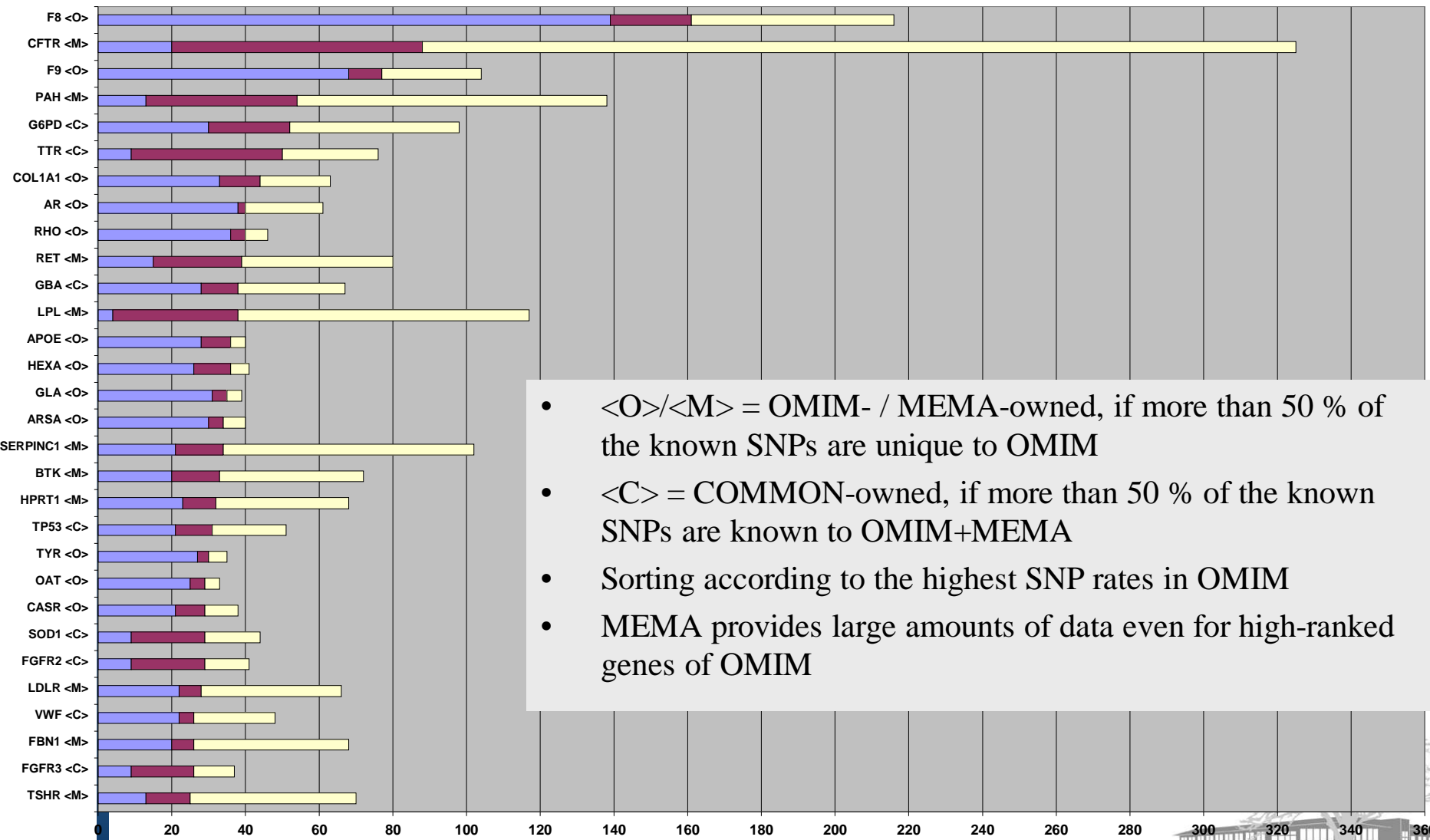
	Nr of Genes	Polym.	SNP (substitution)			SNP (subst.)
			nsSNP	nucleotidic	ambigue	Total
OMIM	1215	10083	6699	207	0	6906
extracted by MEMA	2115	24351	20503	2376	1117	23996
common to OMIM + MEMA	782	1887	1826	38	0	1864
unique to OMIM	433	8196	4873	169	0	5042
unique to MEMA	1333	22464	18677	2338	1117	22132

- MEMA = Mutation Extraction from Medline Abstracts
- nsSNP = non-synonymous SNPs, represented by AA substitutions
- 1887 SNP-gene pairs belonging to 782 genes, are common to OMIM + MEMA
- OMIM contains a large portion of deletion and insertion SNPs
- MEMA is optimized to SNPs of the substitution type

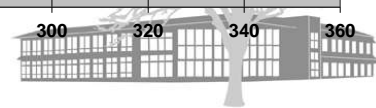


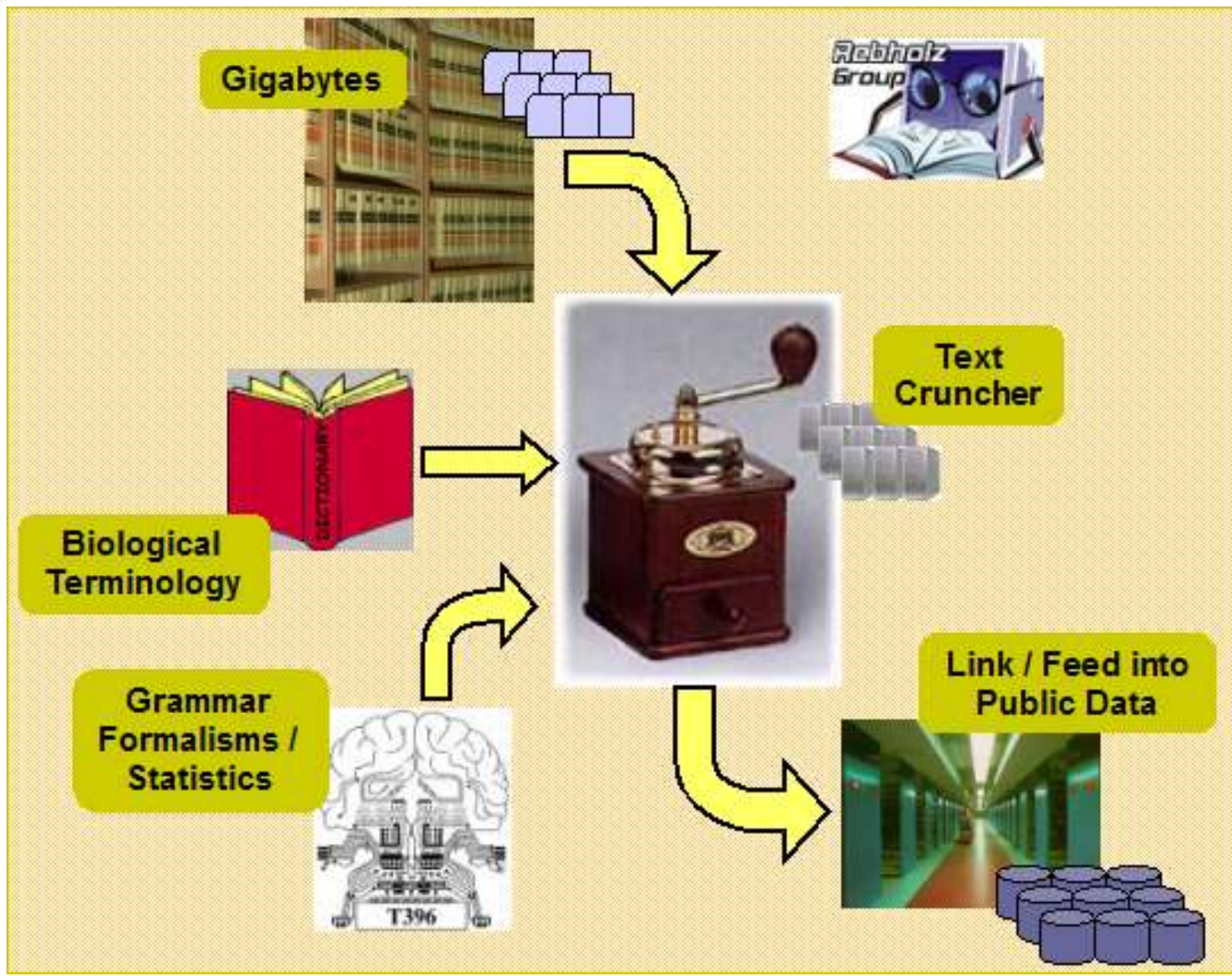
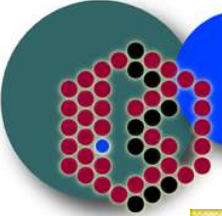


Omim Common Mema



- <O>/<M> = OMIM- / MEMA-owned, if more than 50 % of the known SNPs are unique to OMIM
- <C> = COMMON-owned, if more than 50 % of the known SNPs are known to OMIM+MEMA
- Sorting according to the highest SNP rates in OMIM
- MEMA provides large amounts of data even for high-ranked genes of OMIM





Conclusions:

- The use of good terminological resources is essential.
- How far we profit from ontological relationships, is unclear at present.
- There is no clear distinction between a descriptive name and a describing noun phrase, which leads into a mapping problem.
- RegExp approaches and ML approaches can be integrated into a given pipeline of extraction components on a distributed computing system (all ready to use).



Outlook:

- Improve indexing: special indexes and/or semantic tags (Lucene), interactive IE through indexing (Linguamatics)
- No distinction between Medline + Full paper (resources@EBI), which needs Pdf processing
- Identification / extraction / reduction of facts.
- Individualized IE engines through Whatizit, whoever needs it / wants to assess it
- User feedback on the given modules



Special thanx to the collaborations:

- Automatic Term Recognition: Sophia Ananiadou (Salford Uni), Goran Nenadic (Umist)
- GO team: exact matches for GO and UniProt (Michael Ashburner, Evelyn Camon)
- COSMIC team (Sanger): mutations + phenotypes (Simon Forbes)
- Interact: Protein-protein interactions + PGNs (Henning Hermjakob)
- UniProt curation team: (Rolf Apweiler)
- Peter Stoehr + Rodrigo Lopez: TM services IE on documents available from EBI

