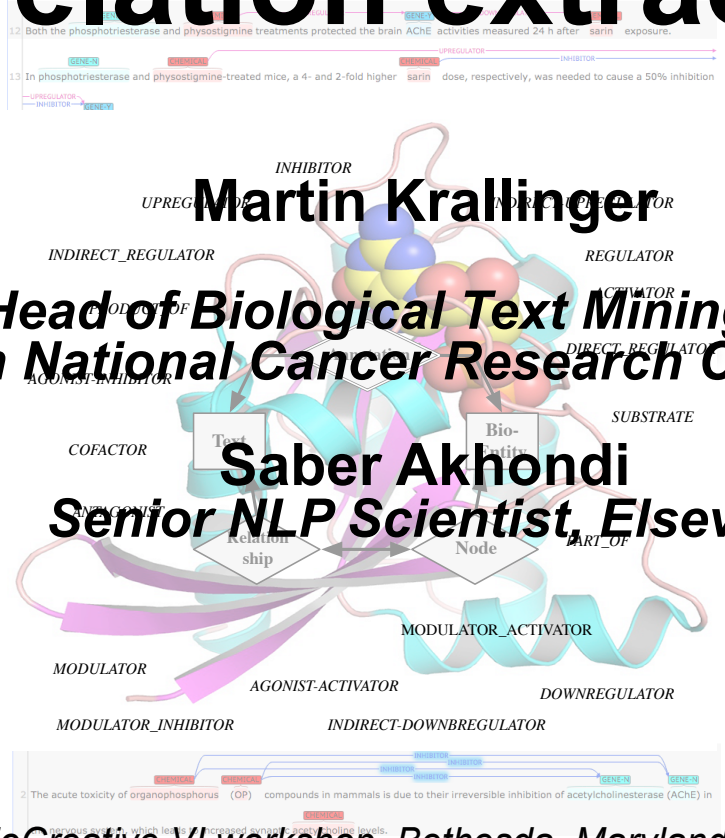
Overview of the Chemical-Protein relation extraction track



Martin Krallinger

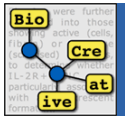
Head of Biological Text Mining Unit
Spanish National Cancer Research Centre (CNIO)

Saber Akhondi

Senior NLP Scientist, Elsevier

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

BioCreative VI workshop, Bethesda, Maryland (October 20th 2017)



BioCreative VI: Chemical-protein interaction (CHEMPROT) track

ChemProt track session at BioCreative VI

08:30 - 10:30	**TRACK 5 Text mining chemical-protein interactions** 8:30-9:00 Overview of the Chemical-Protein relation extraction track (Martin Krallinger / Saber A. Akhondi (CNIO / Elsevier) 9:00-9:15 Chemical-protein relation extraction with SVM, CNN, RNN and ensemble systems (Yifan Peng, NCBI, NLM, NIH) 9:15-9:30 Extracting Chemical-Protein Interactions using Long Short-Term Memory Networks (Sérgio Matos, University of Aveiro, Portugal) 9:30-9:45 Attention based Neural Networks for Chemical Protein Relation Extraction (Ravikumar Komandur Elayavilli, Mayo Clinic, USA) 9:45-10:00 Extracting protein-chemical compound interactions from literature (Pei-Yau Lung, Florida State University) 10:00-10:15 Knowledge-base-enriched relation extraction (Ignacio Tripodi, University of Colorado, Boulder) 10:15 - 10:30 CTCPI - Convolution Tree Kernel-based Chemical-Protein interaction detection (Po-Ting Lai, National Tsing-Hua University, Hsinchu, Taiwan)

Chemical-protein/gene interactions

- Increase in journal, patents describing chemical compounds gene/gene product interactions

- Compared to protein-protein and gene/chemical-disease relations, the detection of relations between chemical and proteins/genes is an under-explored research area

- Constant need for biological, pharmacological and clinical research

- Interested in integration, curation and storing of relationships between biological and chemical entities in databases.

Importance of chemical-protein/gene interactions

•Key relevance:

- **Biomedicine**
- **Molecular biology**
- **Precision medicine**
- **Systems biology**

• **Metabolic relations**

•Construction / curation of **metabolic pathways**

•Understand **drug metabolism** (CYPs)

•**Drug-drug interactions**

•**Adverse reactions**

• **Antagonist** and **agonist** interactions:

- **drug design**
- **drug discovery**
- **mechanism of action**

• **Inhibitor or activator** associations

•**Drug design**

•**Systems biology approaches** (Boolean models of network based approaches)

• **Drug-induced gene expression**

•**Models of drug response** (affect on up/down regulation of gene)

Databases with compounds / compounds interactions

DRUGBANK

STITCH 4.0

SuperTarget

Therapeutic Targets Database

PHARMGKB

REACTOME
A CURATED PATHWAY DATABASE

ChemProt

KEGG
Kyoto Encyclopedia of Genes and Genomes

HUMANCYC
A member of the BioCyc database collection

Small Molecule Pathway Database

ctd

Pathway Studio®

REAXYS®

SureChEMBL beta Open Patent Data

THOMSON REUTERS
INTEGRITY

GOSTAR
GVK BIO Online Structure Activity Relationship Database

ChEMBL

Eidogen·Sertanty

SciFinder®

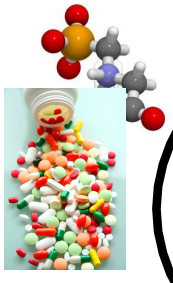
BioCreative VI: Chemical-protein interaction (CHEMPROT) track

Bio-entity relation types

BioCreative II, II.5, III

BioCreative IV, V, V.5

BioCreative I, II, V, V.5



Compounds, Drugs

Genes, gene products

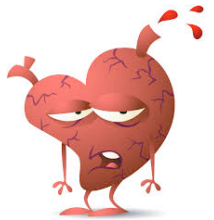
-Protein-Protein interactions
- Transcriptional gene regulation

- Drug-drug interactions

BioCreative IV, V

Synthesis, Reactions

Diseases, Therapeutic indications, AE



BioCreative IV, V

BioCreative VI: Chemical-protein interaction (CHEMPROT) track

CHEMPROT track entities

Compounds, Drugs

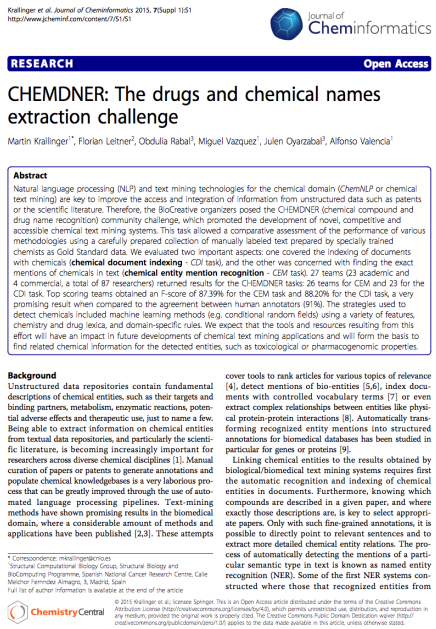
Chemical Entity Mentions (CEMs)

Genes, gene products, sequences,..

Gene and Protein Related Object (GPROs)

Based on previous work: CHEMDNER tracks BioCreative IV, V, V.5

- Krallinger, et al. (2015). CHEMDNER: The drugs and chemical names extraction challenge. *Journal of cheminformatics*, *7*(S1), S1.
- Krallinger, M. (2017). Evaluation of chemical and gene/protein entity recognition systems at BioCreative V. 5: the CEMP and GPRO patents tracks. *Proceedings of the BioCreative*, *5*, 3-11



Krallinger et al. Journal of Cheminformatics 2015, 7(Suppl 1):S1
http://www.jcheminf.com/content/7/S1/S1

Journal of Cheminformatics

RESEARCH Open Access

CHEMDNER: The drugs and chemical names extraction challenge

Martin Krallinger*, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, Alfonso Valencia

Abstract

Natural language processing (NLP) and text mining technologies for the chemical domain (*ChemNLP* or chemical text mining) are key to improve the access and integration of information from unstructured data such as patents or the scientific literature. Therefore, the BioCreative organizers posed the CHEMDNER (chemical compound and drug name recognition) community challenge, which promoted the development of novel, competitive and accessible chemical text mining systems. This task allowed a comparative assessment of the performance of various methodologies using a carefully prepared collection of manually labeled text prepared by specially trained chemists as Gold Standard data. We evaluated two important aspects: one covered the indexing of documents with chemicals (**chemical document indexing** - *CDI* task), and the other was concerned with finding the exact mentions of chemicals in text (**chemical entity mention recognition** - *CEM* task). 27 teams (23 academic and 4 commercial, a total of 87 researchers) returned results for the CHEMDNER tasks: 26 teams for CEM and 23 for the CDI task. Top scoring teams obtained an F-score of 87.39% for the CEM task and 88.20% for the CDI task, a very promising result when compared to the agreement between human annotators (91%). The strategies used to detect chemicals included machine learning methods (e.g. conditional random fields) using a variety of features, chemistry and drug lexica, and domain-specific rules. We expect that the tools and resources resulting from this effort will have an impact in future developments of chemical text mining applications and will form the basis to find related chemical information for the detected entities, such as toxicological or pharmacogenomic properties.

Background

Unstructured data repositories contain fundamental descriptions of chemical entities, such as their targets and binding partners, metabolism, enzymatic reactions, potential adverse effects and therapeutic use, just to name a few. Being able to extract information on chemical entities from textual data repositories, and particularly the scientific literature, is becoming increasingly important for researchers across diverse chemical disciplines [1]. Manual curation of papers or patents to generate annotations and populate chemical knowledgebases is a very laborious process that can be greatly improved through the use of automated language processing pipelines. Text-mining methods have shown promising results in the biomedical domain, where a considerable amount of methods and applications have been published [2,3]. These attempts cover tools to rank articles for various topics of relevance [4], detect mentions of bio-entities [5,6], index documents with controlled vocabulary terms [7] or even extract complex relationships between entities like physical protein-protein interactions [8]. Automatically transforming recognized entity mentions into structured annotations for biomedical databases has been studied in particular for genes or proteins [9].

Linking chemical entities to the results obtained by biological/biomedical text mining systems requires first the automatic recognition and indexing of chemical entities in documents. Furthermore, knowing which compounds are described in a given paper, and where exactly those descriptions are, is key to select appropriate papers. Only with such fine-grained annotations, it is possible to directly point to relevant sentences and to extract more detailed chemical entity relations. The process of automatically detecting the mentions of a particular semantic type in text is known as named entity recognition (NER). Some of the first NER systems constructed where those that recognized entities from

* Correspondence: mkrallinger@cnio.es
Structural Computational Biology Group, Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, Calle Melchor Fernández Almagro, 3, Madrid, Spain
Full list of author information is available at the end of the article

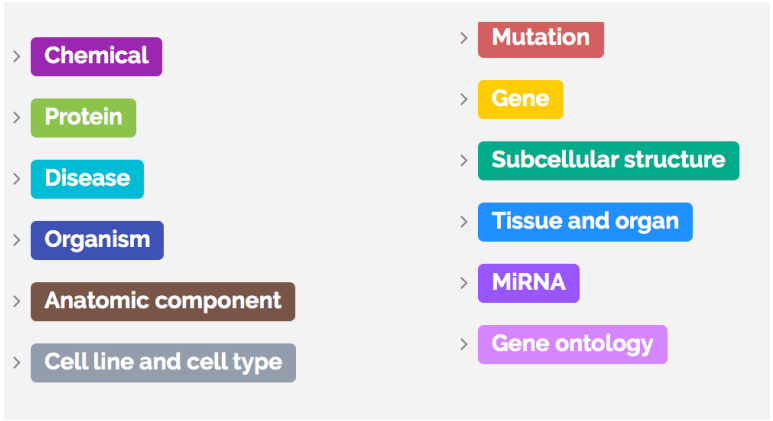
Chemistry Central

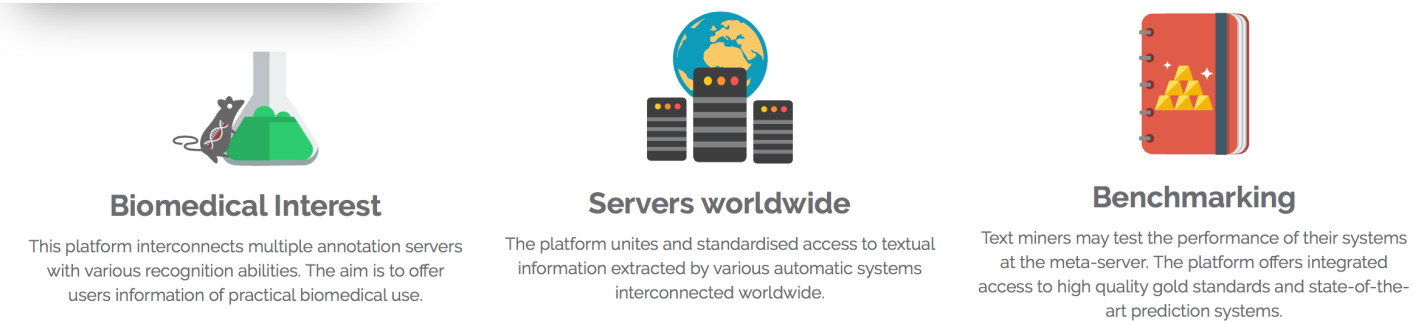
Bio-entity recognition systems

Text mining **evaluation of online systems:**

- TIPS (Technical interoperability and performance of annotation servers).
- CEMP (Chemical Entity Mention recognition)
- GPRO (Gene and Protein Related Object recognition)

BeCalm
Biomedical Annotation Metaserver





Pérez-Pérez, M., et al (2017). Benchmarking biomedical text mining web servers at BioCreative V. 5: the technical Interoperability and Performance of annotation Servers-TIPS track. Proceedings of the BioCreative, 5, 12-21.

BioCreative VI: Chemical-protein interaction (CHEMPROT) track

Bio-entity recognition systems

Server types

Chemical Disease Anatomical component Mutation Subcellular structure miRNA

Server	Chemical	Protein	Disease	Organism	Anatomic component	Cell line and cell type	Mutation	Gene	Subcellular structure	Tissue and organ	MiRNA	Gene ontology	Total
[redacted]			✓										1
[redacted]	✓	✓	✓			✓			✓	✓	✓		7
[redacted]	✓	✓	✓	✓	✓	✓	✓	✓					8
[redacted]	✓	✓	✓			✓	✓	✓					6
[redacted]	✓	✓	✓	✓	✓		✓	✓	✓	✓			9
[redacted]		✓		✓						✓			3
[redacted]	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		10
[redacted]	✓		✓	✓		✓			✓				5
[redacted]												✓	1
[redacted]	✓												1
[redacted]	✓												1
[redacted]			✓				✓				✓		3
[redacted]	✓		✓	✓				✓	✓	✓			6
Total	9	6	9	6	3	5	4	5	5	5	3	1	

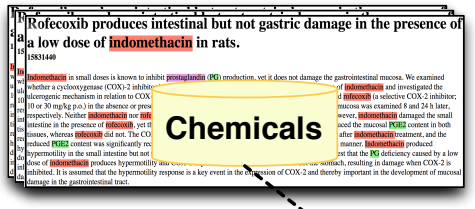
Showing 1 to 14 of 14 entries

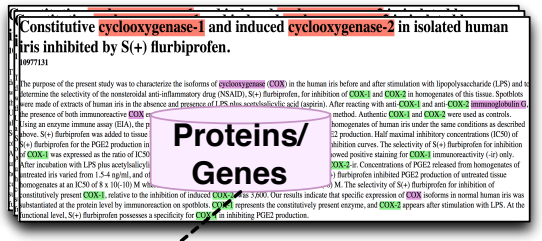
Protein Organism Cell line/type Gene Tissue/organ GO

Pérez-Pérez, M., et al (2017). Benchmarking biomedical text mining web servers at BioCreative V. 5: the technical Interoperability and Performance of annotation Servers-TIPS track. Proceedings of the BioCreative, 5, 12-21.

BioCreative VI: Chemical-protein interaction (CHEMPROT) track

Annotation process



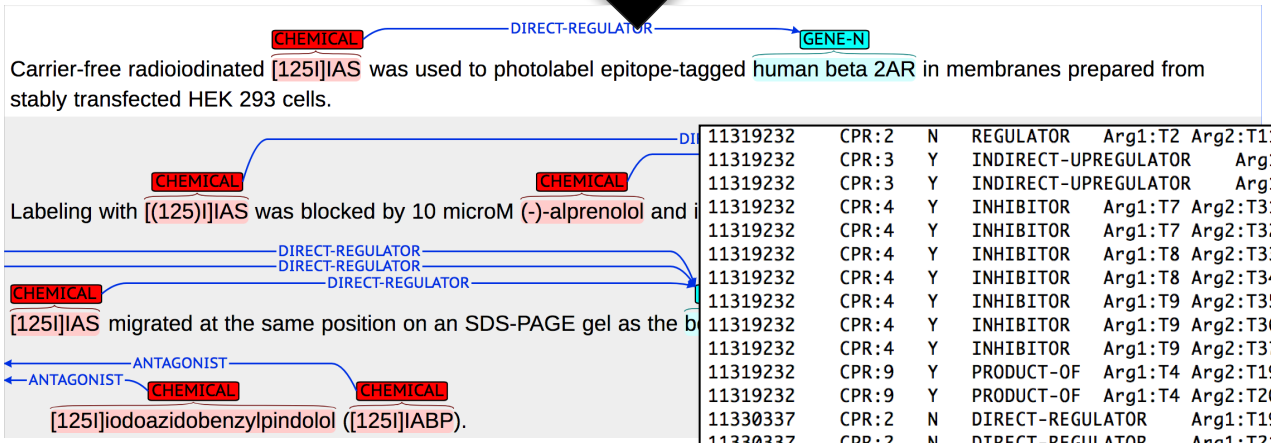



```
11319232    T1  CHEMICAL    242 251 acyl-CoAs
11319232    T2  CHEMICAL    1193    1201    triacsin
11319232    T3  CHEMICAL    1441    1448    sucrose
11319232    T4  CHEMICAL    1637    1652    triacylglycero
11319232    T5  CHEMICAL    1702    1711    acyl-CoAs
11319232    T6  CHEMICAL    176 184 acyl-CoA
11319232    T7  CHEMICAL    790 806 N-ethylmaleimide
11319232    T8  CHEMICAL    898 910 Troglitazone
11319232    T9  CHEMICAL    1012    1022    Triacsin C
11319232    T10 CHEMICAL    0   8   Acyl-CoA
```

Named Entity annotations

Relation annotations

```
11319232    T22 GENE-N  372 376 ACSs
11319232    T23 GENE-Y  509 513 ACS1
11319232    T24 GENE-Y  515 519 ACS4
11319232    T25 GENE-Y  525 529 ACS5
11319232    T26 GENE-Y  531 535 ACS1
11319232    T27 GENE-N  176 195 acyl-CoA synthetase
11319232    T28 GENE-Y  655 659 ACS4
11319232    T29 GENE-Y  705 709 ACS5
11319232    T30 GENE-N  197 200 ACS
11319232    T31 GENE-Y  824 828 ACS4
11319232    T32 GENE-N  840 843 ACS
11319232    T33 GENE-Y  921 925 ACS4
```

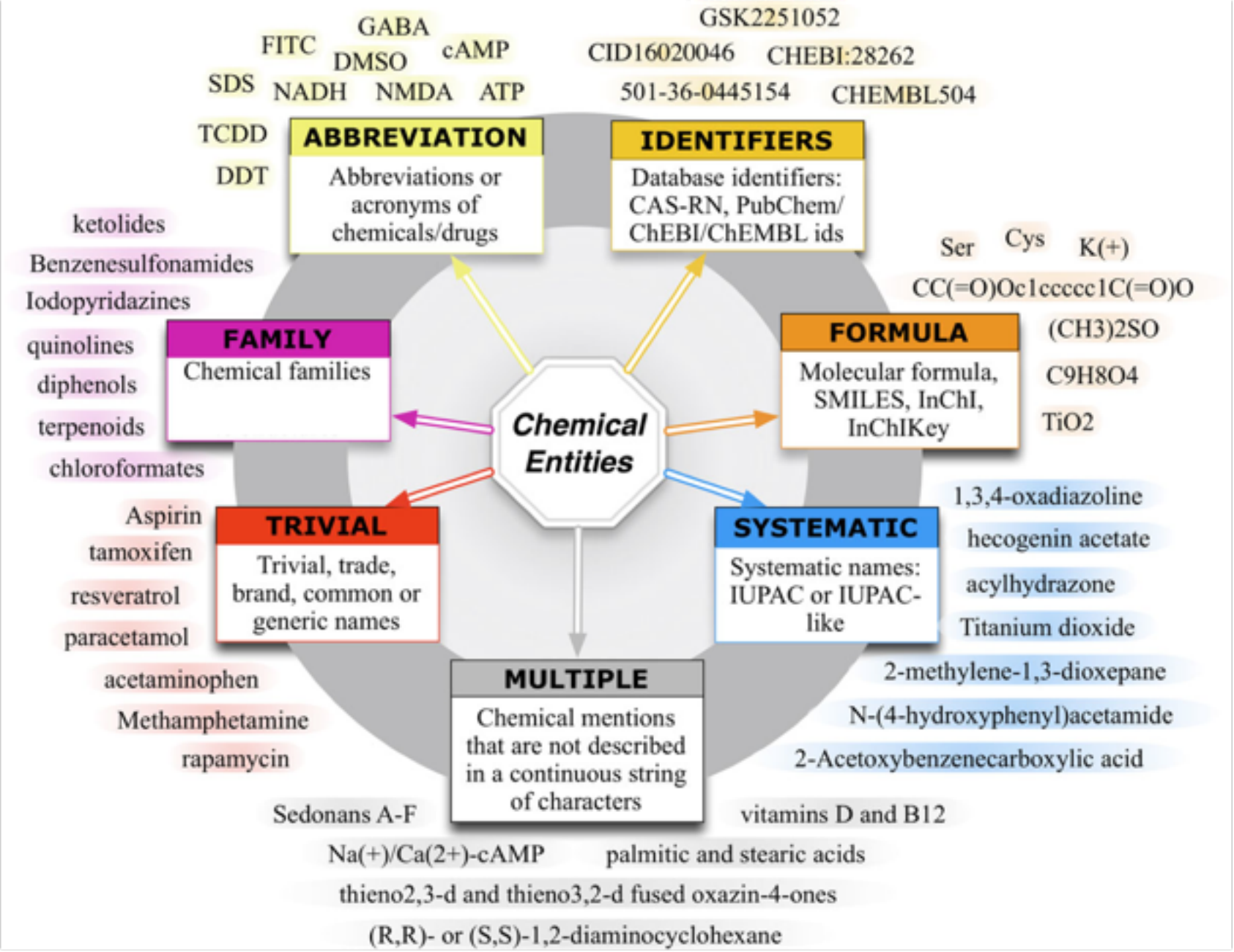



```
11319232    CPR:2   N   REGULATOR   Arg1:T2 Arg2:T11
11319232    CPR:3   Y   INDIRECT-UPREGULATOR    Arg1:T3 Arg2:T17
11319232    CPR:3   Y   INDIRECT-UPREGULATOR    Arg1:T3 Arg2:T18
11319232    CPR:4   Y   INHIBITOR   Arg1:T7 Arg2:T31
11319232    CPR:4   Y   INHIBITOR   Arg1:T7 Arg2:T32
11319232    CPR:4   Y   INHIBITOR   Arg1:T8 Arg2:T33
11319232    CPR:4   Y   INHIBITOR   Arg1:T8 Arg2:T34
11319232    CPR:4   Y   INHIBITOR   Arg1:T9 Arg2:T35
11319232    CPR:4   Y   INHIBITOR   Arg1:T9 Arg2:T36
11319232    CPR:4   Y   INHIBITOR   Arg1:T9 Arg2:T37
11319232    CPR:9   Y   PRODUCT-OF  Arg1:T4 Arg2:T19
11319232    CPR:9   Y   PRODUCT-OF  Arg1:T4 Arg2:T20
11330337    CPR:2   N   DIRECT-REGULATOR    Arg1:T19    Arg2:T47
11330337    CPR:2   N   DIRECT-REGULATOR    Arg1:T21    Arg2:T50
```

Chemical guidelines

- Defines what constitutes a chemical entity mention (CEM)
- Based on the previous CHEMDNER BioCreative IV, V and V.5 guidelines
- Only Chemical nouns (and specific adjectives, treated as nouns) are tagged (not reactions, prefixes or enzymes)
- Classification of CEMs into 7 classes
- GPNOM-rule annotation rules:
  - General rules
  - Positive rules
  - Negative rules
  - Orthography rules
  - Multiword rules

BioCreative VI: Chemical-protein interaction (CHEMPROT) track

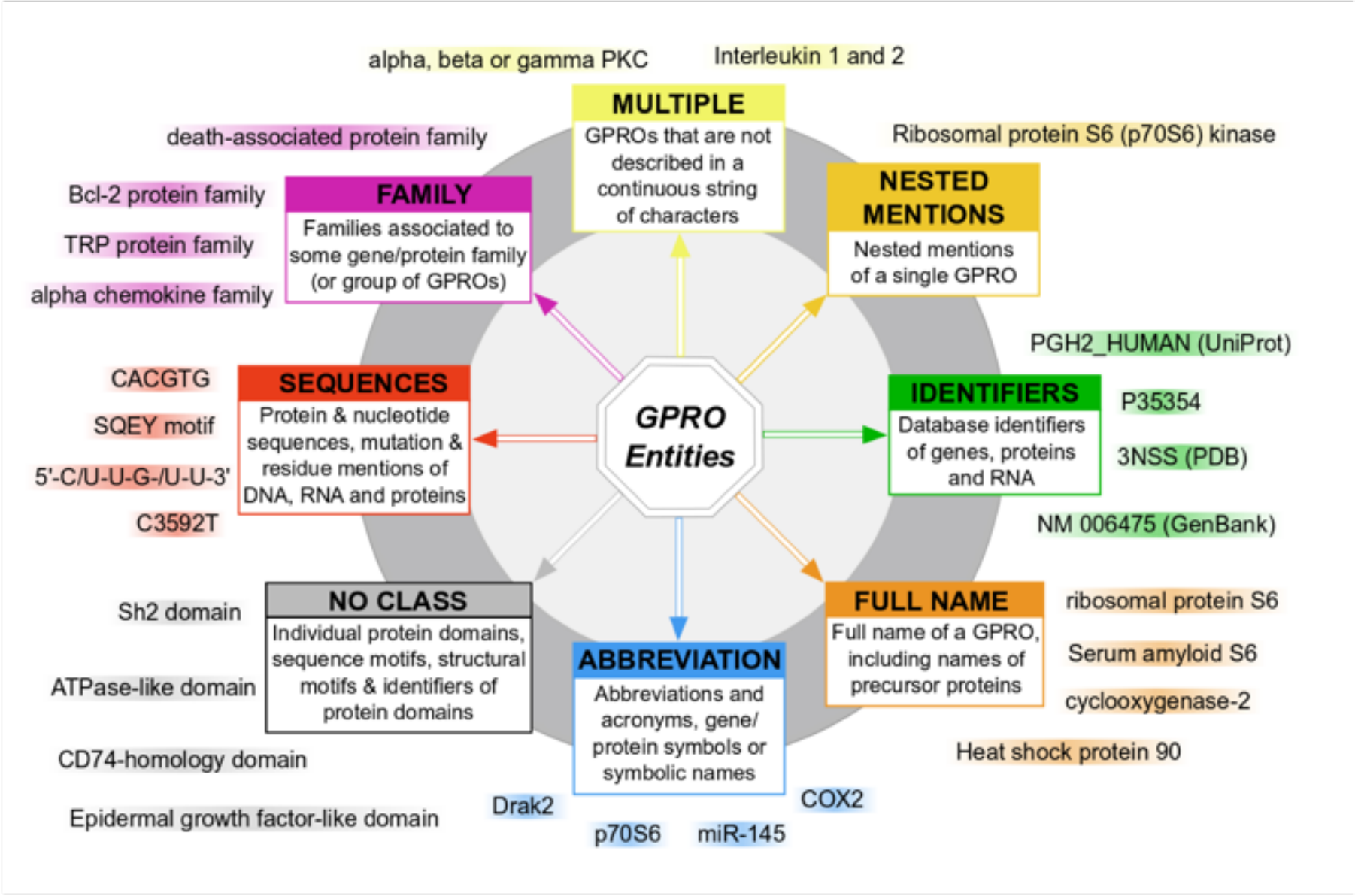


Gene/protein guidelines

- Defines what constitutes a Gene and Protein Related Object (GPRO)
- Genes, gene products (proteins, RNA), DNA/protein sequence elements and protein families, domains and complexes
- Used for CHEMDNER track of BioCreative V and V.5
- GPNOM-annotation rules: general, positive, negative, orthography, multiword
- Additional rules for assignment to 8 GPRO classes
- Includes GPRO grounding guidelines

  GPRO mentions that can be normalized to a bio-entity database record

GPRO mention classes



GPRO mention class types 1 and 2

NESTED MENTIONS	Nested mentions of a single GRPO	ribosomal protein S6 (p70S6) kinase
IDENTIFIER	Database identifiers of genes, proteins and RNA	P35354 PGH2_HUMAN (UniProt) 3NSS (PDB) NM 006475 (GenBank)
FULL NAME	Full name of a GPRO, including names of precursor proteins (cleaved proteins). Multi-word terms referring to specific gene/protein named entities.	ribosomal protein S6 Serum amyloid A3 cyclooxygenase-2 Heat shock protein 90
ABBREVIATION	Abbreviation and acronyms of GPROs, gene/protein symbols or symbolic names.	Drak2 p70S6 COX2 miR-145
NO CLASS	Individual protein domains, sequence motifs, structural motifs, identifiers of protein domains (PFAM).	Epidermal growth factor-like domain Sh2 domain CD74-homology domain ATPase-like motifs
SEQUENCE	Protein (amino acid) sequences, nucleotide sequences, mutation and residue mentions of DNA, RNA and proteins.	CACGTG SQEY motif 5′-C/U-U-G/U-U-3′ C3592T
FAMILY	GPRO families associated to some gene/ protein family (or group of GPROs).	death-associated protein family Bcl-2 protein family TRP protein family alpha chemokine family
MULTIPLE	GPROs that are not described in a continuous string of characters.	Interleukin 1 and 2 alpha, beta, or gamma PKC

GENE-Y tag

C1 type

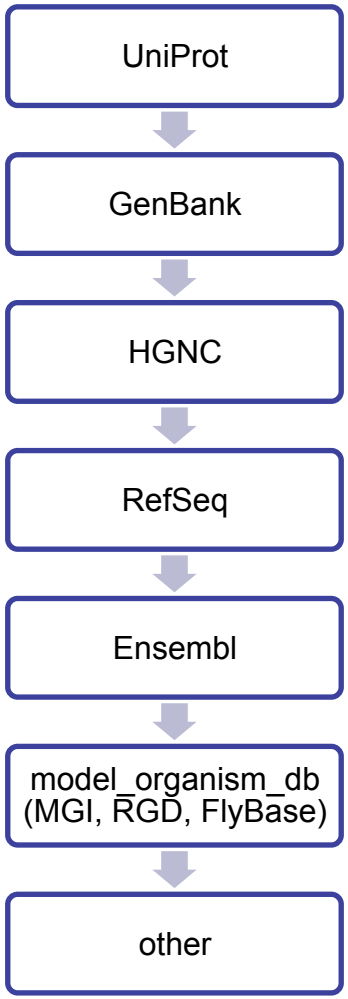
normalized to a bio-entity database record

GENE-N tag

C2 type

Can not be normalized to databases

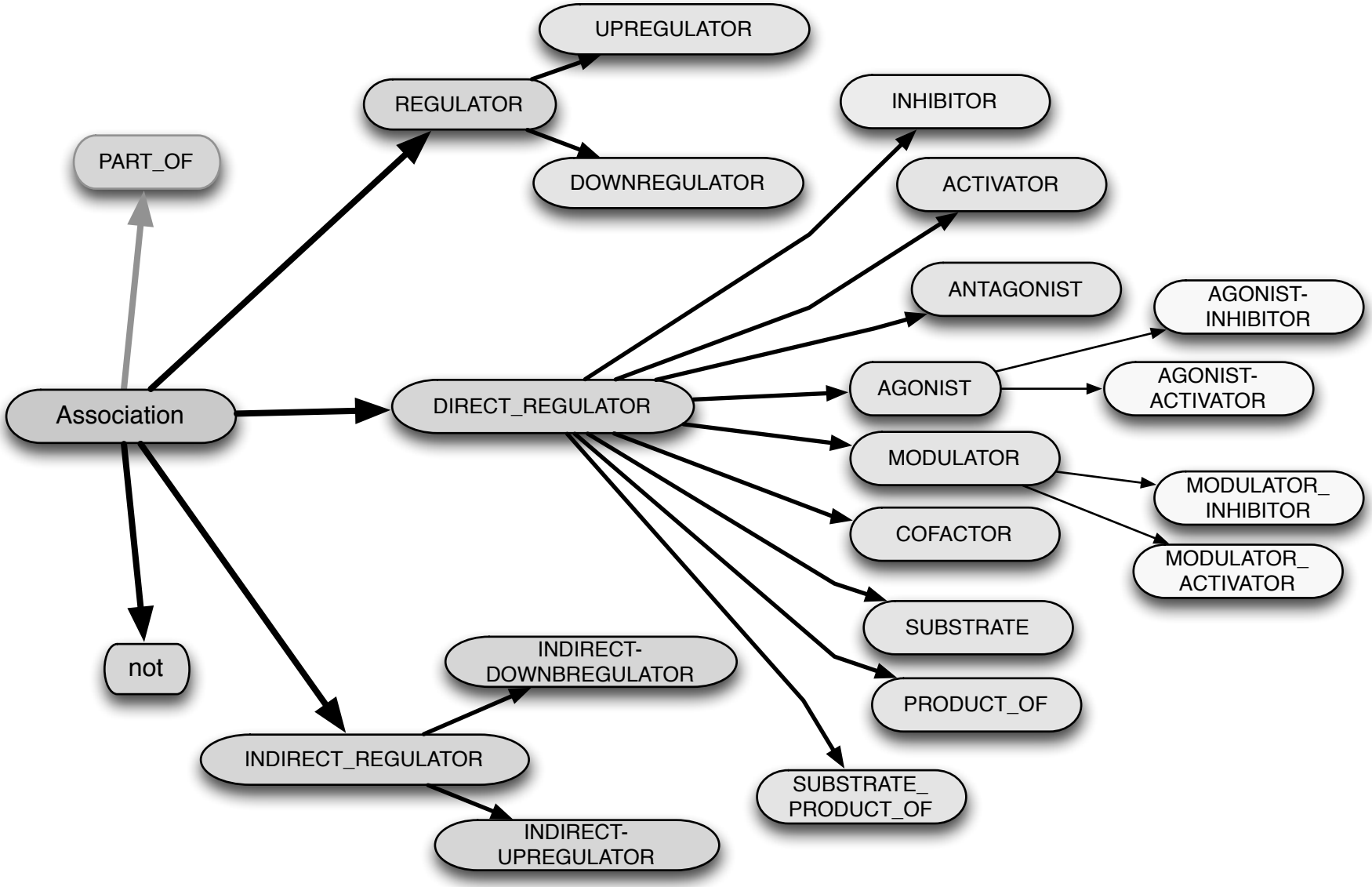
GPRO mention class types normalization



Chemical-Protein interaction guidelines

- Relation annotation carried was exhaustive for set of interactions types
- Other relationships between chemicals and genes (e.g. phenotype and biological responses) were not annotated
- The ChemProt relations were **directed**, i.e. only relations of “what a chemical does to a gene/protein" (chemical → gene/protein direction) were annotated and not vice versa
- Establish a homogeneous nomenclature and avoid redundant class definitions
- Based on revision of several resources: DrugBank, Therapeutic Targets Database, ChEMBL, assay normalization ontologies (BAO), Biological Expression Language (BEL) and SIGNOR
- These resources inspired subclasses definitions of the DIRECT REGULATOR class or the INDIRECT REGULATOR class

ChemProt relation types

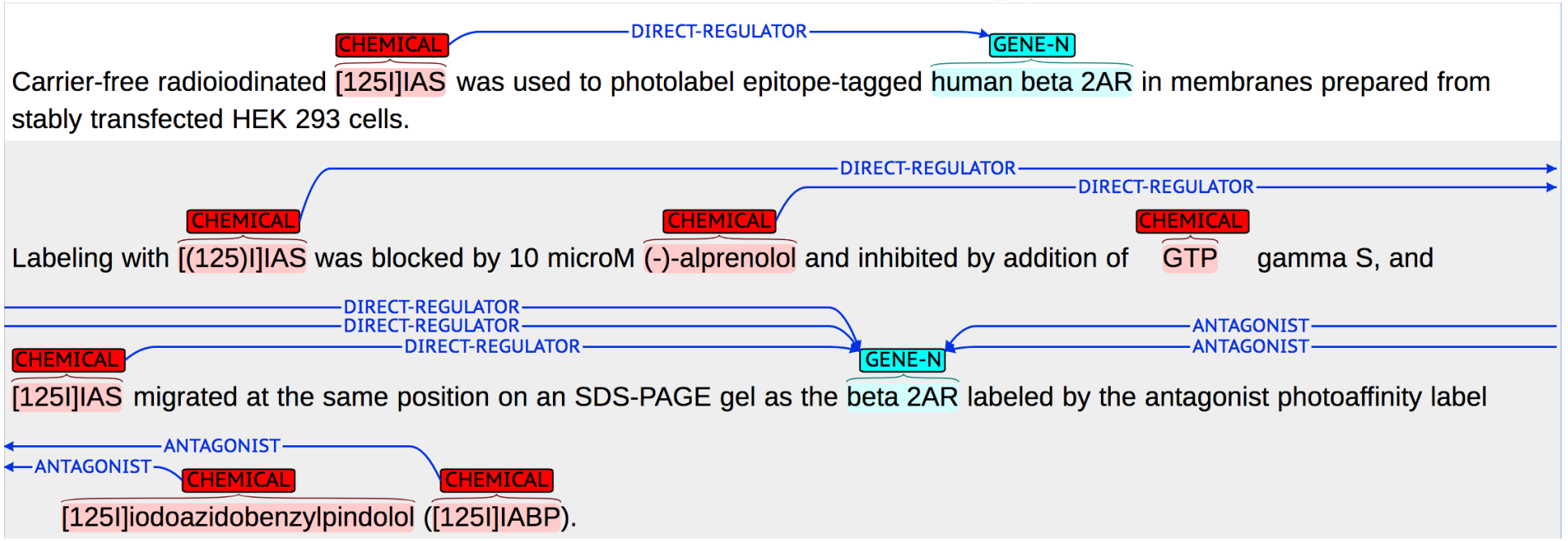


ChemProt relation types

Group	*Eval.*	*CHEMPROT relations belonging to this group*
CPR:1	N	PART_OF
CPR:2	N	REGULATOR|DIRECT_REGULATOR|INDIRECT_REGULATOR
CPR:3	Y	UPREGULATOR|ACTIVATOR|INDIRECT_UPREGULATOR
CPR:4	Y	DOWNREGULATOR|INHIBITOR|INDIRECT_DOWNREGULATOR
CPR:5	Y	AGONIST|AGONIST-ACTIVATOR|AGONIST-INHIBITOR
CPR:6	Y	ANTAGONIST
CPR:7	N	MODULATOR|MODULATOR-ACTIVATOR|MODULATOR-INHIBITOR
CPR:8	N	COFACTOR
CPR:9	Y	SUBSTRATE|PRODUCT_OF|SUBSTRATE_PRODUCT_OF
CPR:10	N	NOT

Grouped based on biological semantical classes

ChemProt relation types



1. Article identifier (PMID)
2. Chemical-Protein relation (CPR) group*
3. Evaluation type (Y: group evaluated, N: group not evaluated – extra annotation).
4. CHEMPROT relation (CPR)
5. Interactor argument 1 (chemical entity followed by the interactor term identifier)
6. Interactor argument 2 (gene/protein entity followed by the interactor term identifier)

ChemProt corpus overview

Dataset	Documents	Chemicals	Genes	All CPIs
Sample	50	683	606	339
Training	1.020	13.017	12.735	6.437
Development	612	8.004	7.563	3.558
Test	800	10.810	10.018	5.744
All	2482	32514	30922	16078

- Gene/protein mentions normalizable (GENE-Y): total of 20544 (63.20% of all gene/protein mentions)
- 2,599 additional abstracts were added to test set (total 3,399 for testing)

Annotations per ChemProt CPR group

CPR Class	Count Relations
CPR:4	5119
CPR:2	4258
CPR:3	2074
CPR:9	1873
CPR:6	741
CPR:10	698
CPR:1	677
CPR:5	500
CPR:7	73
CPR:8	62
CPR:0	3

Group	*Eval.*	*CHEMPROT relations belonging to this group*
CPR:1	N	PART_OF
CPR:2	N	REGULATOR|DIRECT_REGULATOR|INDIRECT_REGULATOR
CPR:3	Y	UPREGULATOR|ACTIVATOR|INDIRECT_UPREGULATOR
CPR:4	Y	DOWNREGULATOR|INHIBITOR|INDIRECT_DOWNREGULATOR
CPR:5	Y	AGONIST|AGONIST-ACTIVATOR|AGONIST-INHIBITOR
CPR:6	Y	ANTAGONIST
CPR:7	N	MODULATOR|MODULATOR-ACTIVATOR|MODULATOR-INHIBITOR
CPR:8	N	COFACTOR
CPR:9	Y	SUBSTRATE|PRODUCT_OF|SUBSTRATE_PRODUCT_OF
CPR:10	N	NOT

Annotations per ChemProt relation type

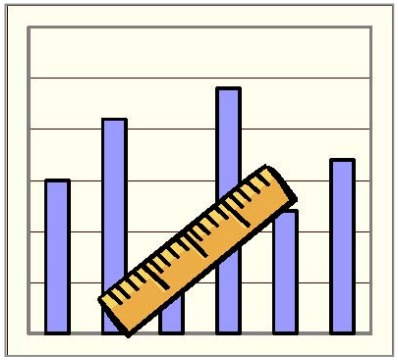
Relation type	Nr. Annotations
INHIBITOR	3758
REGULATOR	1876
DIRECT-REGULATOR	1652
SUBSTRATE	1284
INDIRECT-DOWNREGULATOR	1109
INDIRECT-UPREGULATOR	1042
ACTIVATOR	873
ANTAGONIST	741
INDIRECT-REGULATOR	730
NOT	698
PART-OF	677
PRODUCT-OF	570
AGONIST	465
DOWNREGULATOR	252
UPREGULATOR	159
COFACTOR	62
MODULATOR-ACTIVATOR	46
SUBSTRATE_PRODUCT-OF	19
AGONIST-ACTIVATOR	19
MODULATOR	16
AGONIST-INHIBITOR	16
MODULATOR-INHIBITOR	11
UNDEFINED	3

Evaluation metrics

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$



Main metric: Micro-averaged F-score

Automated predictions against manual annotations (Gold Standard)

Exact match evaluation

FN false negatives - incorrect negative classification results (type II errors)

FP false positives - incorrect positive classification results (type I errors)

TN true negatives - correct negative classification results (correct rejection)

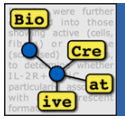
TP true positives - correct positive classification results (correct hit)

ChemProt participating teams

Team Id	Team Leader	Institution	Nr. runs
374	Sérgio Matos	Universidade de Aveiro	5
379	Sijia Liu	Mayo Clinic	4
394	Neha Warikoo	Academia Sinica	5
397	Atakan Yüksel	Boğaziçi University	1
403	Peter Corbett	Royal Society of Chemistry	1
404	Ignacio Tripodi	University of Colorado	5
417	Farrokh Mehryary	University of Turku	5
421	Cong Sun	DaLian University of Technology	1
424	Sangrak Lim	Korea University	2
427	Wei Wang	National University of Defense Technology	5
430	Yifan Peng	NCBI, NLM, NIH	5
432	Pat Verga	UMass Amherst	4
433	Pei-Yau Lung	Florida State University	2

ChemProt results across all interaction classes (I)

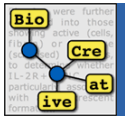
Team Id	Run	Precision	Recall	F-Score	# Predictions	TP	FN	FP
Co-mention	Abstract	0.0050	1.0000	0.0099	693635	3458	0	690177
Co-mention	Sentence	0.0437	0.9803	0.0837	77545	3390	68	74155
TEAM_430	RUN_5	0.7266	0.5735	**0.6410**	2729	1983	1475	746
TEAM_430	RUN_4	0.7311	0.5685	0.6397	2689	1966	1492	723
TEAM_430	RUN_1	**0.7437**	0.5529	0.6343	2571	1912	1546	659
TEAM_430	RUN_2	0.7283	0.5503	0.6269	2613	1903	1555	710
TEAM_430	RUN_3	0.7426	0.5382	0.6241	2506	1861	1597	645
TEAM_403	RUN_1	0.5610	**0.6784**	0.6141	4182	2346	1112	1836
TEAM_417	RUN_3	0.6608	0.5662	0.6099	2963	1958	1500	1005
TEAM_417	RUN_4	0.6105	0.6006	0.6055	3402	2077	1381	1325
TEAM_417	RUN_5	0.6088	0.5989	0.6038	3402	2071	1387	1331
TEAM_424	RUN_2	0.6704	0.5194	0.5853	2679	1796	1662	883
TEAM_424	RUN_1	0.6760	0.5159	0.5852	2639	1784	1674	855
TEAM_433	RUN_2	0.6352	0.5121	0.5671	2788	1771	1687	1017
TEAM_433	RUN_1	0.6276	0.4858	0.5477	2677	1680	1778	997
TEAM_417	RUN_1	0.6373	0.4462	0.5249	2421	1543	1915	878
TEAM_417	RUN_2	0.6337	0.4387	0.5185	2394	1517	1941	877
TEAM_374	RUN_5	0.5738	0.4722	0.5181	2846	1633	1825	1213
TEAM_379	RUN_5	0.5301	0.4639	0.4948	3026	1604	1854	1422
TEAM_374	RUN_2	0.5156	0.4670	0.4901	3132	1615	1843	1517
TEAM_379	RUN_2	0.4849	0.4913	0.4881	3504	1699	1759	1805
TEAM_379	RUN_4	0.5072	0.4306	0.4657	2936	1489	1969	1447
TEAM_432	RUN_4	0.4718	0.4453	0.4582	3264	1540	1918	1724
TEAM_379	RUN_1	0.4773	0.4375	0.4565	3170	1513	1945	1657



BioCreative VI: Chemical-protein interaction (CHEMPROT) track

ChemProt results across all interaction classes (II)

Team Id	Run	Precision	Recall	F-Score	# Predictions	TP	FN	FP
Co-mention	Abstract	0.0050	1.0000	0.0099	693635	3458	0	690177
Co-mention	Sentence	0.0437	0.9803	0.0837	77545	3390	68	74155
TEAM_432	RUN_3	0.4073	0.4783	0.4400	4061	1654	1804	2407
TEAM_374	RUN_4	0.4024	0.4193	0.4107	3603	1450	2008	2153
TEAM_427	RUN_5	0.2696	0.6663	0.3839	8545	2304	1154	6241
TEAM_427	RUN_4	0.2674	0.6602	0.3806	8538	2283	1175	6255
TEAM_427	RUN_3	0.2634	0.6622	0.3769	8695	2290	1168	6405
TEAM_404	RUN_2	0.3387	0.4078	0.3700	4163	1410	2048	2753
TEAM_374	RUN_1	0.6419	0.2577	0.3677	1388	891	2567	497
TEAM_404	RUN_1	0.3460	0.3913	0.3673	3910	1353	2105	2557
TEAM_427	RUN_2	0.2535	0.6478	0.3643	8838	2240	1218	6598
TEAM_427	RUN_1	0.2496	0.6417	0.3594	8889	2219	1239	6670
TEAM_404	RUN_4	0.3307	0.3641	0.3466	3807	1259	2199	2548
TEAM_374	RUN_3	0.5919	0.2403	0.3418	1404	831	2627	573
TEAM_404	RUN_5	0.3058	0.3603	0.3309	4074	1246	2212	2828
TEAM_394	RUN_3	0.2932	0.3271	0.3092	3857	1131	2327	2726
TEAM_394	RUN_5	0.2587	0.3456	0.2959	4619	1195	2263	3424
TEAM_432	RUN_2	0.5491	0.2021	0.2955	1273	699	2759	574
TEAM_394	RUN_2	0.2563	0.3456	0.2943	4662	1195	2263	3467
TEAM_394	RUN_1	0.2446	0.3407	0.2847	4816	1178	2280	3638
TEAM_404	RUN_3	0.3305	0.1666	0.2215	1743	576	2882	1167
TEAM_421	RUN_1	0.1618	0.3409	0.2195	7287	1179	2279	6108
TEAM_432	RUN_1	0.2211	0.2024	0.2114	3166	700	2758	2466
TEAM_397	RUN_1	0.6057	0.1102	0.1864	629	381	3077	248
TEAM_394	RUN_4	0.0729	0.0150	0.0249	713	52	3406	661



BioCreative VI: Chemical-protein interaction (CHEMPROT) track

ChemProt best f-score for each CPR class

Team Id	CPR class	Run	Precision	Recall	F-Score	Relations
430	CPR:6	RUN_3	0.8075	0.6587	**0.7256**	*ANTAGONIST*
430	CPR:4	RUN_1	0.7651	0.6707	0.7148	*DOWNREGULATOR|INHIBITOR| INDIRECT_DOWNREGULATOR*
430	CPR:5	RUN_5	0.7903	0.5026	0.6144	*AGONIST|AGONIST-ACTIVATOR| AGONIST-INHIBITOR*
430	CPR:3	RUN_1	0.6667	0.5143	0.5806	*UPREGULATOR|ACTIVATOR| INDIRECT_UPREGULATOR*
430	CPR:9	RUN_5	0.7008	0.4255	0.5295	*SUBSTRATE|PRODUCT_OF| SUBSTRATE_PRODUCT_OF*

Conclusions

- Engaged a considerable number of teams
- First time a chemical-protein/gene interaction relation track was posed
- Obtained result are valuable contributions for curation of chemical and biological data
- ChemProt corpus: a large collection of manually annotated mentions of chemical compounds and genes/proteins for improving/ evaluating bio-entity recognition systems
- ChemProt interactions types could foster sophisticated bio-entity relation extraction pipelines with very competitive results (esp. machine learning / artificial neural network based approaches)
- Certain relation classes were easier to extract, while difficult classes might need a more granular relation type definition

Future directions

• We plan to promote additional efforts focusing on these more granular relation types

• Will provide annotations of gene mention normalizations, i.e. biological database identifiers for the ChemProt corpus

• Plan to release detailed inter-annotator agreement results

• Ideal scenario for production of silver standard corpus using various systems and set up a community effort for curation of chemical interactions

•We are still annotating and will make a test set available for teams to calculate performance

Acknowledgements

cnio Centro Nacional de Investigaciones Oncológicas

- Martin Krallinger
- Jesús Santamaría
- Ander Intxaurrondo
- José Antonio López

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

cima
CENTER FOR APPLIED MEDICAL RESEARCH
UNIVERSITY OF NAVARRA
Molecular Therapeutics Program

- Obdulia Rabal
- Julen Oyarzabal*

GOBIERNO DE ESPAÑA · MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Inncorpora-Torres Quevedo PTQ-1-04781

To all BioCreative Teams!

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

- Alfonso Valencia

BIO CREATIVE

- BioCreative organizers
- Cecilia Arighi
- Lynette Hirschman

ELSEVIER

- Saber A. Akhondi
- Georgios Tsatsaronis
- Umesh Nandal
- Erin Van Buel
- Akileshwari Chandrasekhar
- Marleen Rodenburg
- Marius Doornenbal

Universidad de Vigo

- Analia Lourenço
- Martin Perez Perez
- Gael Perez Rodriguez
- Florentino Fernández Riverola

NTNU
Norwegian University of Science and Technology

- Astrid Laegreid

Congratulations

s.akhondi@elsevier.com

Elsevier price goes to team 430