Medizinische Universität Graz

The Gene Regulation Ontology (GRO): Quality, Requirements, Scope, Redesign Principles,

Stefan Schulz – Medical University of Graz (Austria)

with input from

Astrid Lægreid, NTNU (Norway)

Martin Kuiper, NTNU (Norway)

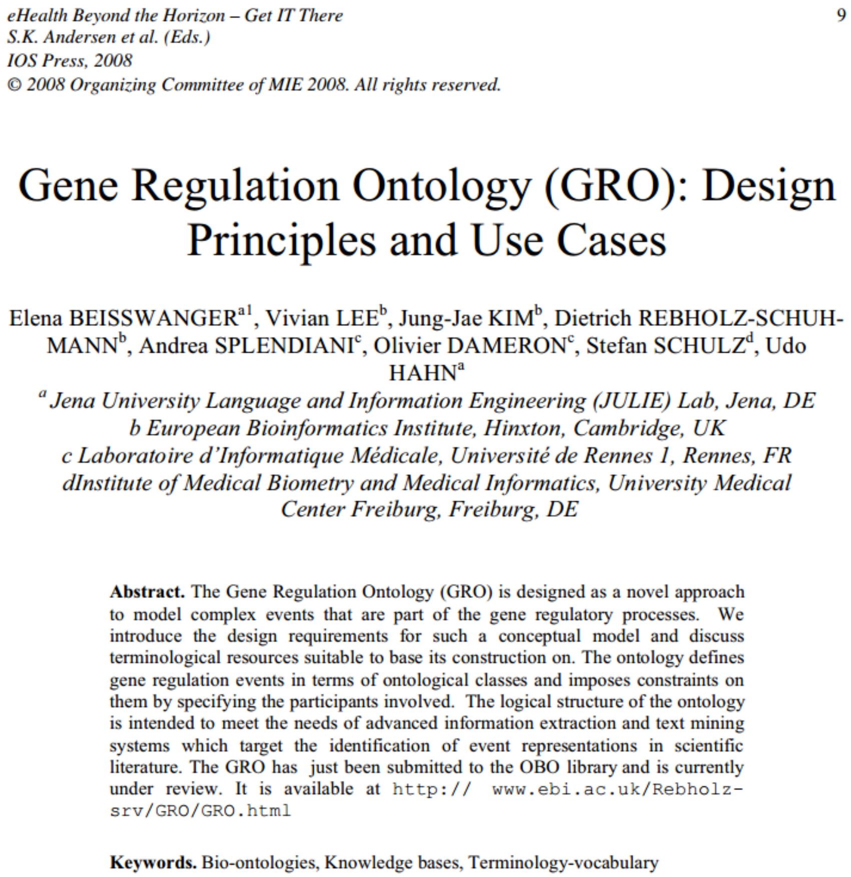
Jesualdo Tomás Fernández Breis, University of Murcia (Spain)

Chris Mungall, Bey Lab (U.S.)

Michel Dumontier, University of Maastricht (The Netherlands)

History of GRO

- FP6 BootSTREP (2006 – 09)
  - Focus: E. Coli gene regulation
- Design:
  - OBO Foundry principles
  - OWL-DL (Tbox only)
- Resources:
  - GO, SO, ChEBI, NCBI, TransFac, RO
  - UMLS and 150 MEDLINE abstracts
- Purpose:
  - Semantic annotation of documents
  - Basis for SWRL rules to derive biologica knowledge
- Use:
  - ISA software suite
  - KINO search engine
  - Annotations of GREC
  - BioNLP Shared Task 2013



eHealth Beyond the Horizon – Get IT There 9
S.K. Andersen et al. (Eds.)
IOS Press, 2008
© 2008 Organizing Committee of MIE 2008. All rights reserved.

Gene Regulation Ontology (GRO): Design Principles and Use Cases

Elena BEISSWANGER[a1], Vivian LEE[b], Jung-Jae KIM[b], Dietrich REBHOLZ-SCHUHMANN[b], Andrea SPLENDIANI[c], Olivier DAMERON[c], Stefan SCHULZ[d], Udo HAHN[a]

[a] Jena University Language and Information Engineering (JULIE) Lab, Jena, DE
b European Bioinformatics Institute, Hinxton, Cambridge, UK
c Laboratoire d'Informatique Médicale, Université de Rennes 1, Rennes, FR
dInstitute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Freiburg, DE

Abstract. The Gene Regulation Ontology (GRO) is designed as a novel approach to model complex events that are part of the gene regulatory processes. We introduce the design requirements for such a conceptual model and discuss terminological resources suitable to base its construction on. The ontology defines gene regulation events in terms of ontological classes and imposes constraints on them by specifying the participants involved. The logical structure of the ontology is intended to meet the needs of advanced information extraction and text mining systems which target the identification of event representations in scientific literature. The GRO has just been submitted to the OBO library and is currently under review. It is available at `http:// www.ebi.ac.uk/Rebholz-srv/GRO/GRO.html`

Keywords. Bio-ontologies, Knowledge bases, Terminology-vocabulary

Revisiting GRO in GREEKC

Creation of Google document on March 8, 2017:

Issues addressed:

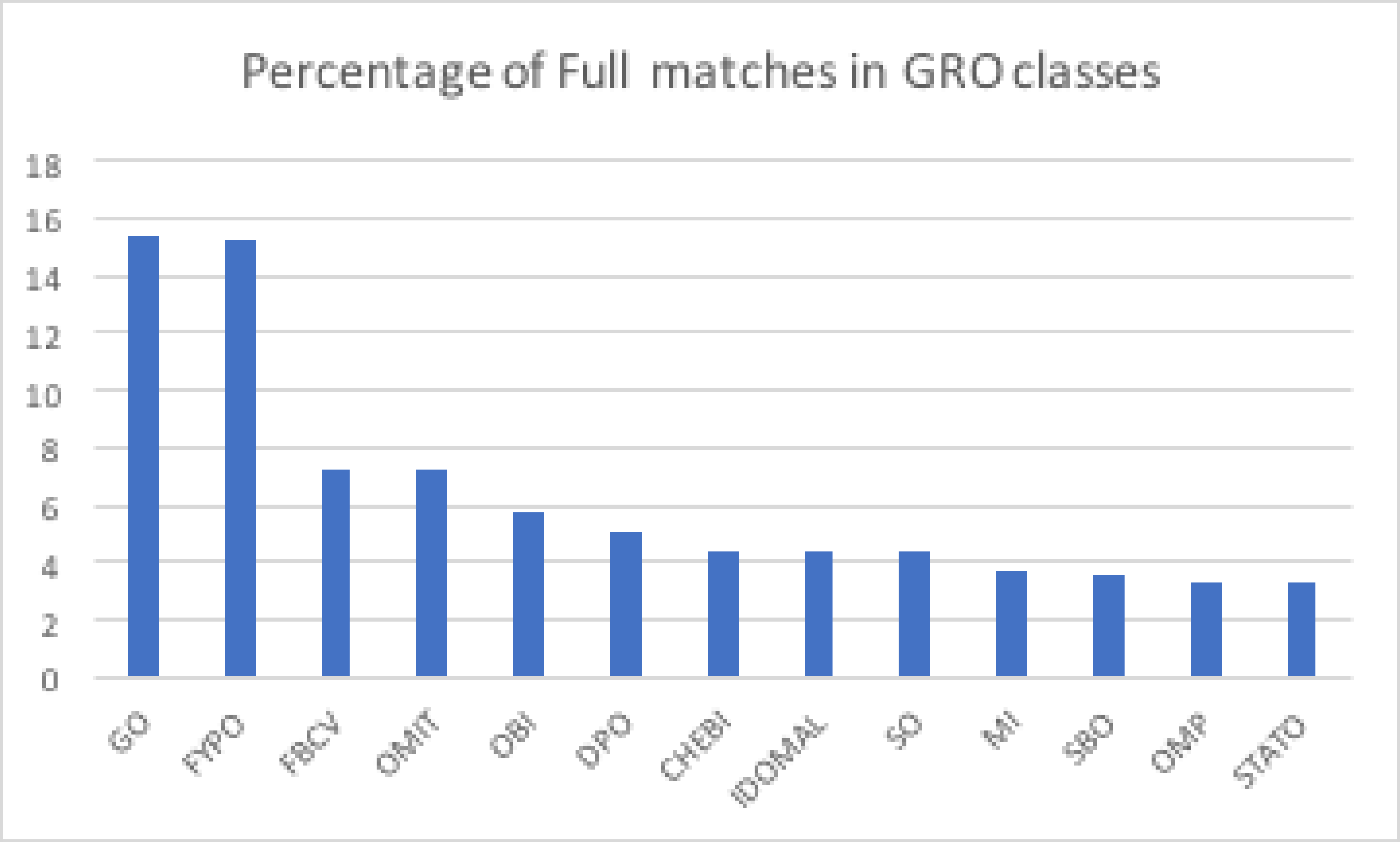
- Analysis of existing GRO 0.5 according to the OQUARE quality framework
  (Jesualdo Tomás Fernández Breis)
- Analysis of overlap of GRO 0.5 with other ontologies using OntoEnrich
  (Jesualdo Tomás Fernández Breis)
- Requirement analysis
  (Martin Kuiper, Astrid Lægreid)
- Start of redesign: GRO2, aligned with upper-level ontology BTL2
  (Stefan Schulz)
- Scoping / Limitations
  (Chris Mungall)

GRO 0.5 x OQuaRE

- Ontology metrics:
  - 507 classes,
  - 24 object properties
  - 9 datatype properties
  - **total of 2736 axioms**
  - 759 subclassOf axioms
  - 65 equivalentClasses
  - 91 disjointClasses
- OQuaRE metrics (detailed characteristics: 1 ☹ … 5☺):
  - Very good (> 4): controlled vocabulary; consistency, redundancy
  - Good (> 3 – 4): adaptability, analysability, testeability, changeability
  - Fair (> 2 – 3): results representation, formal relation support, reusability, changeability, replaceability
  - Poor (> 1 - 2): cohesion, availability
- OQuaRE assessment: GRO should be better related to other ontologies. Maintaineability should be improved

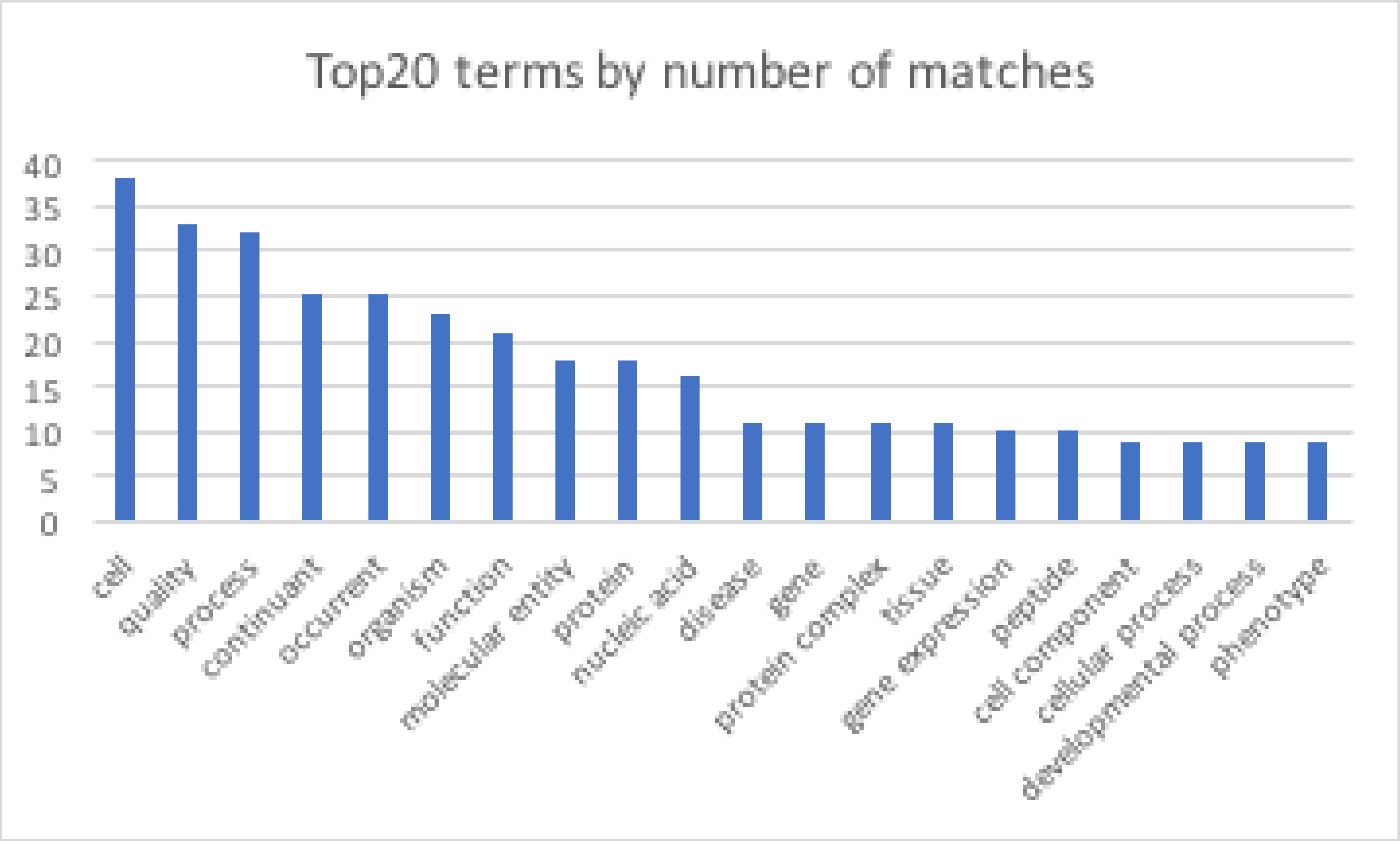
GRO 0.5 overlap (ontoEnrich)

- results of lexical alignment of GRO with OBO ontologies. Constraints: only full matches and a minimum coverage of 3%



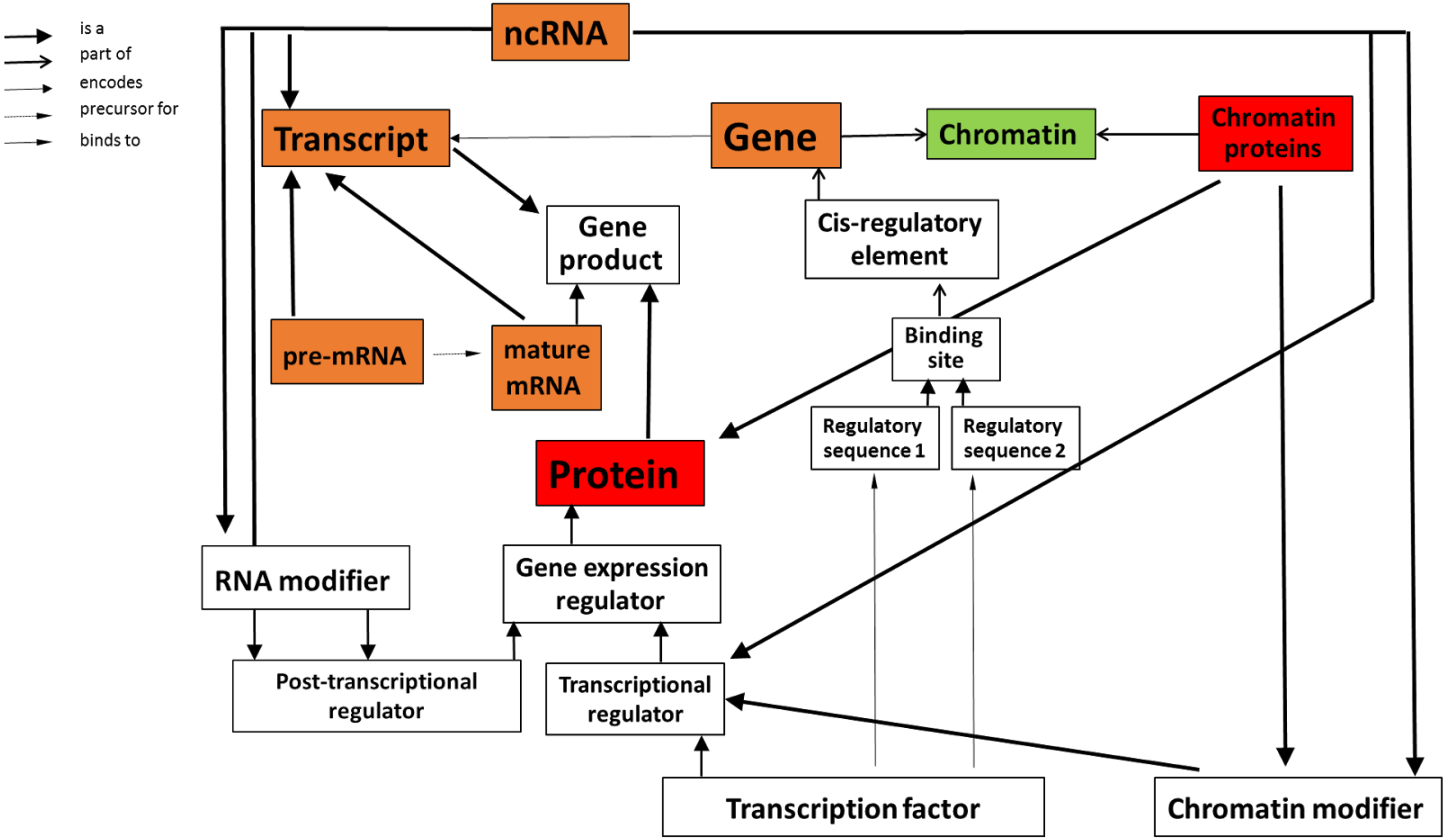
GRO 0.5 overlap (ontoEnrich)

- Top 20 GRO classes with largest amount of full matches



GRO Requirement analysis (I)

- What GRO needs to cover



Acknowledgements: Dietrich Rebholz-Schuhmann

GRO Content Requirements (I)

- Transcription factors (TFs)
  - DNA-binding or protein-binding
  - Complex (dimer, other)
  - TF/Complex activity (target gene (TG) specificity, activating/repressing, other; - all these change with nature of complex; each TF promiscuous regarding complex formations)
  - Posttranslational modifications (PTMs) (mostly phosphorylations), effect on activity
  - Ligand association (for TFs that are ligand-regulated nuclear receptors)
  - Localization (nuclear, cytoplasmic, other)
- Chromatin modifiers (proteins, non-coding RNAs, others)
  - Similar list to the above for TFs

GRO Content Requirements (II)

- Chromatin
  - 3D structure (nucleome) and dynamics (responses to internal and external cues), including interactions with nuclear lamina
  - Nucleome functional consequences for gene regulation (which TFs can bind where, can participate in which complexes, etc.)
- Cellular context
  - Cell type, origin (tissue, organ, developmental stage, etc.)
  - Cell status (primary, cancer cell line, other)
  - Cell state (proliferating, migrating, etc.)

GRO Functional requirements

- Structuring of content of knowledge bases
- Manual biocuration
- Text mining assisted biocuration
- Querying of knowledge bases
- Logical reasoning
- Building of molecular pathways linking intracellular signalling and gene regulation
- Taking cellular (tissue-/organ-/organismal-?) context into account
- Building and analysing directed networks (large, small) accounting for the integrated signalling and gene regulation molecular mechanisms in a cellular (tissue-/organ-/organismal-?) context

GRO Limitations

- Inherent limitations of ontologies
  - Express what is universally true
  - Borderline: including dispositions
    (some continuant has the ability to participate in processes of a certain kind)
- Representation of contingent / probabilistic / default knowledge outside the scope of ontologies
- Representation of quantitative information
- Representation of sequences – relation to SO?
  - sequences as physical segments of polymers vs.
  - sequences as abstract (information) entities
- Abox entities (individuals, instances)?
- Datatype properties?
- OWL full?

GRO2 methodology

- Goal: alignment with upper-level ontology BTL2 (BioTopLite v2)
- Reasons for upper-level alignment in general
  - avoids creating an specific upper level for every ontology
  - reduces the freedom of the modeller: improves interoperability
  - reduced the number of primitives: less ambiguities
- Reasons for BTL2 in particular:
  - good experiences with BTL2 in numerous projects
  - aligned with BFO, but with relations
  - highly axiomatised
  - more "pragmatic" than BFO
- Alignment done by ontology expert (Stefan Schulz)
- Time needed (including documentation): 5 hours
- Tools: Protégé 5; text editor

GRO →GRO2: steps 1-10

1. Import of BTL2
2. gro:roles under btl2:role
3. molecular entities under *btl2:monomolecular entity*
4. removal of disjoints that cause inconsistencies, e.g.

   complex molecular entity disjoint with *protein*
   complex molecular entity disjoint with *protein domain*
5. molecular function under *btl2:process*
6. Substitution of **hasFunction** (regarding BFO “activities”) by

   '**btl2:is bearer of**' some
   (*btl2:disposition* and (**'btl2:has realization'** only *'X activity'*)))
7. Function-activities related to their agents via ‘**btl2:has agent**’
8. substitution of other homonymous relations, e.g. **hasPart** by ‘**btl2:has part**’ (using text editor)
9. processes that give rise to other processes: ‘**btl2:has outcome**’, e.g.

   ‘**has patient**’ some ‘*gene expression*’ →
   'btl2:has outcome' some (**btl2:causes** some *'gene expression'*)
10. Some category shifts, e.g. *'experimental method'* subclass of *btl2:action*

GRO →GRO2: Steps 11-14

11. btl2 agents and patients haven’t processes in the range, therefore, e.g. **‘has agent’** some *transcription* → **‘btl2:is caused by’** some *transcription*
12. removal of datatype properties, e.g. **hasPolarity**; Values are modelled by DOLCE-style value regions, e.g.

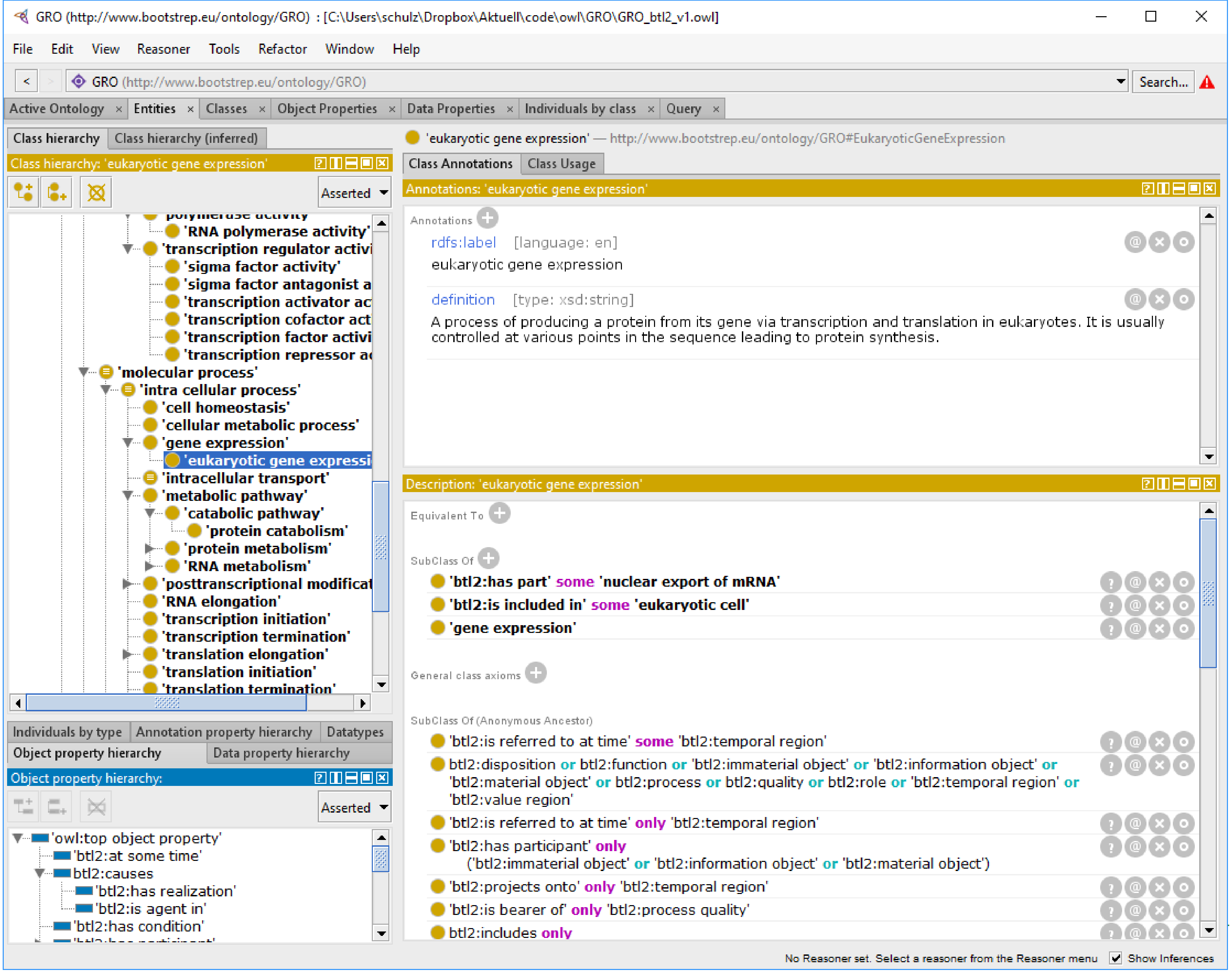
    Decrease subclassOf **'btl2:is bearer of'** some
    (*btl2:quality* and (**'btl2:projects onto'** some 'negative *polarity value region'*))
13. removal of redundant object properties:

    fromSpecies some *Eukaryote* →
    ‘btl2:at some time’ some (**‘btl2:is included in’** some *Eukaryote*)

    (there are alternative ways to express this, e.g. taxon qualities)
14. encoding: like realization of functions / dispositions: existential quantification too strong

    ORG subclassOf **encodes** some *Protein* →
    ORG subclassOf **'btl2:is bearer of'** some (*'genetic information'* and
    (**btl2:represents** only *protein*))

    reduction to bearer of protein-representing information sufficient?



GRO (http://www.bootstrep.eu/ontology/GRO) : [C:\Users\schulz\Dropbox\Aktuell\code\owl\GRO\GRO_btl2_v1.owl]

GRO (http://www.bootstrep.eu/ontology/GRO)

Class hierarchy: 'eukaryotic gene expression'

```
        polymerase activity
            'RNA polymerase activity'
        'transcription regulator activi
            'sigma factor activity'
            'sigma factor antagonist a
            'transcription activator ac
            'transcription cofactor act
            'transcription factor activi
            'transcription repressor ac
    'molecular process'
        'intra cellular process'
            'cell homeostasis'
            'cellular metabolic process'
            'gene expression'
                'eukaryotic gene expressi
            'intracellular transport'
            'metabolic pathway'
                'catabolic pathway'
                    'protein catabolism'
                'protein metabolism'
                'RNA metabolism'
            'posttranscriptional modificat
            'RNA elongation'
            'transcription initiation'
            'transcription termination'
            'translation elongation'
            'translation initiation'
            'translation termination'
```

'eukaryotic gene expression' — http://www.bootstrep.eu/ontology/GRO#EukaryoticGeneExpression

Annotations: 'eukaryotic gene expression'

Annotations

rdfs:label [language: en]

eukaryotic gene expression

definition [type: xsd:string]

A process of producing a protein from its gene via transcription and translation in eukaryotes. It is usually controlled at various points in the sequence leading to protein synthesis.

Description: 'eukaryotic gene expression'

Equivalent To

SubClass Of

- 'btl2:has part' some 'nuclear export of mRNA'
- 'btl2:is included in' some 'eukaryotic cell'
- 'gene expression'

General class axioms

SubClass Of (Anonymous Ancestor)

- 'btl2:is referred to at time' some 'btl2:temporal region'
- btl2:disposition or btl2:function or 'btl2:immaterial object' or 'btl2:information object' or 'btl2:material object' or btl2:process or btl2:quality or btl2:role or 'btl2:temporal region' or 'btl2:value region'
- 'btl2:is referred to at time' only 'btl2:temporal region'
- 'btl2:has participant' only ('btl2:immaterial object' or 'btl2:information object' or 'btl2:material object')
- 'btl2:projects onto' only 'btl2:temporal region'
- 'btl2:is bearer of' only 'btl2:process quality'
- btl2:includes only

Object property hierarchy:

```
'owl:top object property'
    'btl2:at some time'
    btl2:causes
        'btl2:has realization'
        'btl2:is agent in'
    'btl2:has condition'
    'btl2:has participant'
```

No Reasoner set. Select a reasoner from the Reasoner menu ☑ Show Inferences

Ideal world / blue sky

- Computable, well-performing ontology that covers all aspects of gene regulation
- Addressing a broad range of use cases
- Well connected with other ontologies
  - Upper level ontologies
  - Domain ontologies
- Following standards and good design principles
  - Self explaining labels
  - Formal and textual definitions / elucidations
  - Supporting synonyms
- Maintained by GR community / agreed quality criteria
  - Tools (content maintenance, content acquisition),
  - Benchmarks, competency questions
- Computable, ontology-based framework integrating bioontology and database content

Gaps

- Unclear scoping and specification for a Gene Regulation Ontology
- Better solved by combining existing ontologies
  - OBO principles (?)
- Overlap with existing resources, risk of creating / maintaining redundant content
- Different opinions regarding importance and choice of an upper level ontology and design principles
  - current upper level ontologies:
  - e.g. lexical polysemy: "sequence": information entity vs. material entity?
  - e.g. what are classes, what are relations, what are individuals?
- User friendly tooling
  - Content acquisition, ontology creation and curation
  - Quality assessment

GRO: Long-term perspective

- to discuss!