Text Mining at work: critical assessment of the completeness and correctness of knowledge-based, computable disease models for the pharmaceutical industry

Prof. Martin Hofmann-Apitius

Head of the Department of Bioinformatics

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

II-SDV conference, Nice, April 16, 2013

Where do I come from? → Fraunhofer Society

*Joseph von Fraunhofer (1787 – 1826)
Scientist, Inventor and Entrepreneur

- Founded 1949
- Europe´s largest applied research organisation
- 60 Research Institutes (7 Institutes in the US)
- 17.000 Employees
- Annual Budget about 1,5 Billion Euro
- Financial model: 2/4 industry collaborations
  1/4 public funding
  1/4 institutional funding

The Fraunhofer Institute Center Schloss Birlinghoven

- Largest research centre for informatics and applied mathematics in Germany
- Around 700 employees, thereof 500 scientists, approx. 200 students and trainees

What science are we doing?
SCAI Department of Bioinformatics: R&D in a nutshell

Fraunhofer SCAI Department of Bioinformatics R&D activities:

1. Information extraction in the life sciences:

   I. Text Mining - Recognition of named entities & relationships in text

   II. Image Mining - Reconstruction of chemical information from chemical structure depictions

2. Disease modelling (focus on neurodegenerative diseases

3. eScience, Grid-/Cloud- Computing and HPC (C

Making Scientific Content
available for Computing

My Dream: Direct Usage of Unstructured Information Sources for Disease Modelling

From Text Mining Technology
to
Modelling of Neurodegenerative Diseases

Why Modelling of Neurodegeneration?

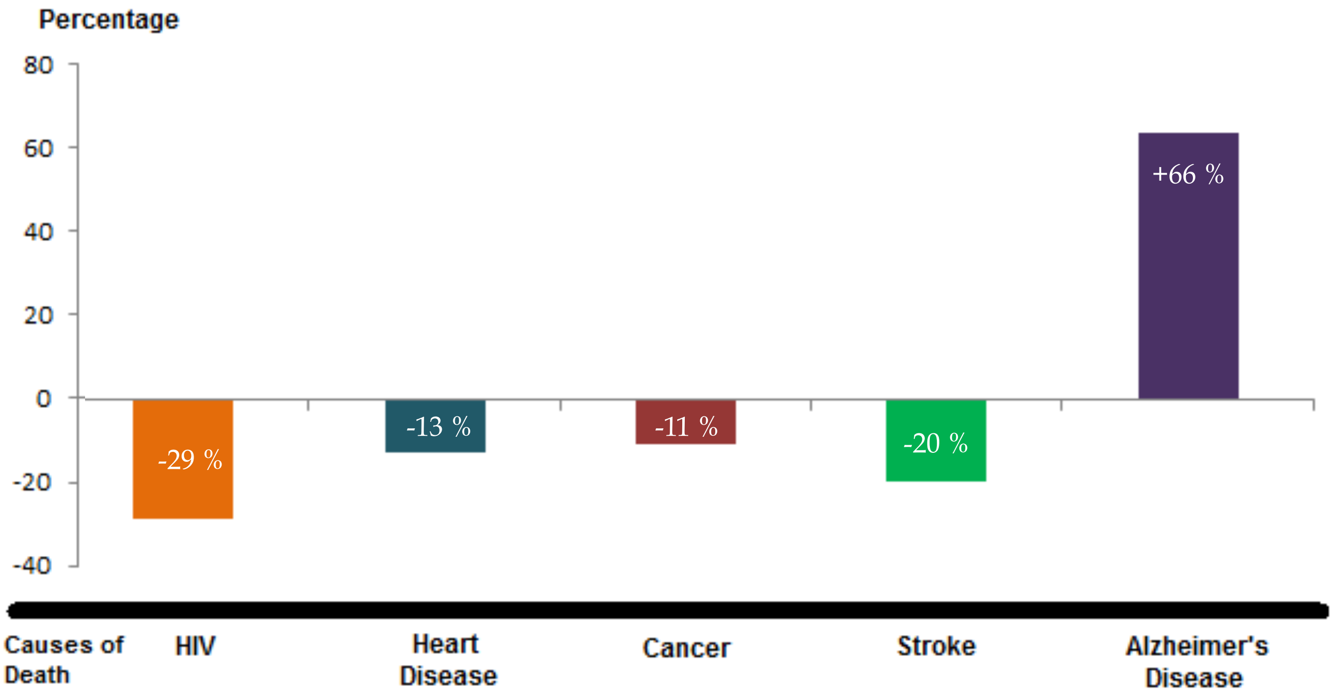
In 2009 the Federal Government of Germany decided to start a new research centre that focuses on translational research on neurodegenerative diseases. In fact, neurodegenerative diseases (Alzheimer, Parkinson, Multiple Sclerosis; Epilepsy; „rare“ NDDs) are a major societal challenge:

The total costs of Alzheimer is estimated to exceed 20 trillion US$ in the US in the years between 2020 - 2050. (source: Alzheimer.org). Current costs / year in the US (according to Alzheimer.org): **183 billion US$**

The incidence rate of Alzheimer and other dementias is almost 50% in the population older than 85 years. Next generation will regularly have a life span of >100 years.

Diseases specific mortality rate

Changes in selected causes of death in USA , 2000-2010[1]



[1] www.alz.org

Modelling Alzheimer´s Disease: Tools

- An ontology capturing relevant knowledge on Alzheimer´s Disease (**ADO**; Malhotra et al., “Alzheimer´s & Dementia”, in press)
- An ontology representing brain regions and cell types (**BRCO**)
- An excellent machinery for biomedical text mining (**ProMiner – UIMA enabled**) with top performing gene and protein name recognition
- A **biomarker terminology** that identifies biomarker candidates in the scientific literature (Younesi et al., BMC Med. Inf. Dec. Making (2012))
- A powerful formalism to capture knowledge and model it as a network of causal and correlative relationships: **BEL** (biological expression language)

Capturing Knowledge for Disease Models: BEL

Biological Expression Language (BEL)

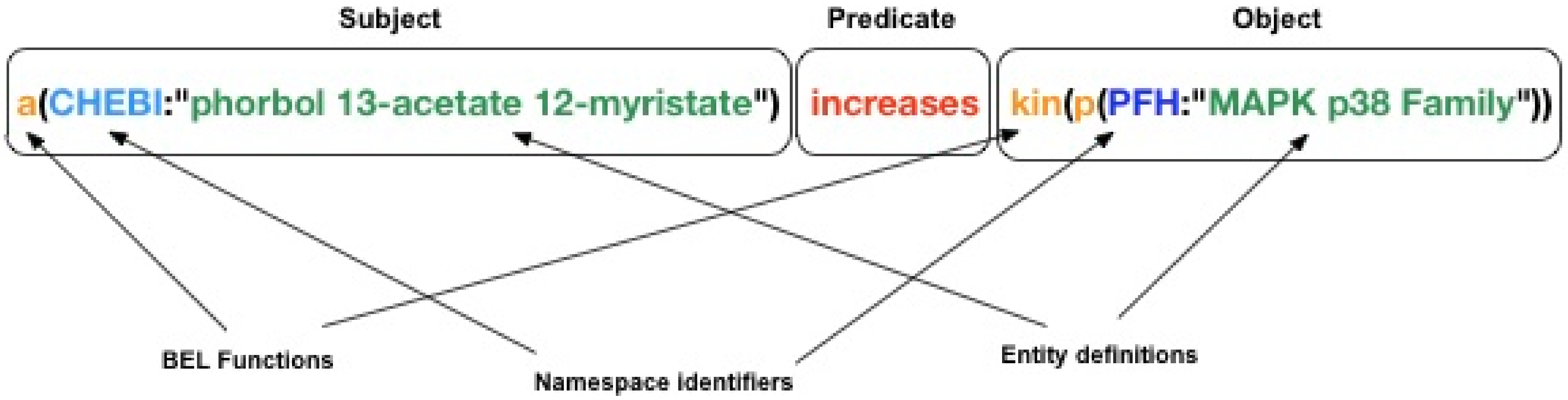
BEL is a language for representing scientific findings (e.g. what you would read in a journal article) in a computable form

- Captures qualitative causal and correlative relationships in context
  - Biological & experimental system in which the relationships were observed
  - Literature cited
  - Curation process

Advantages of BEL as a Language

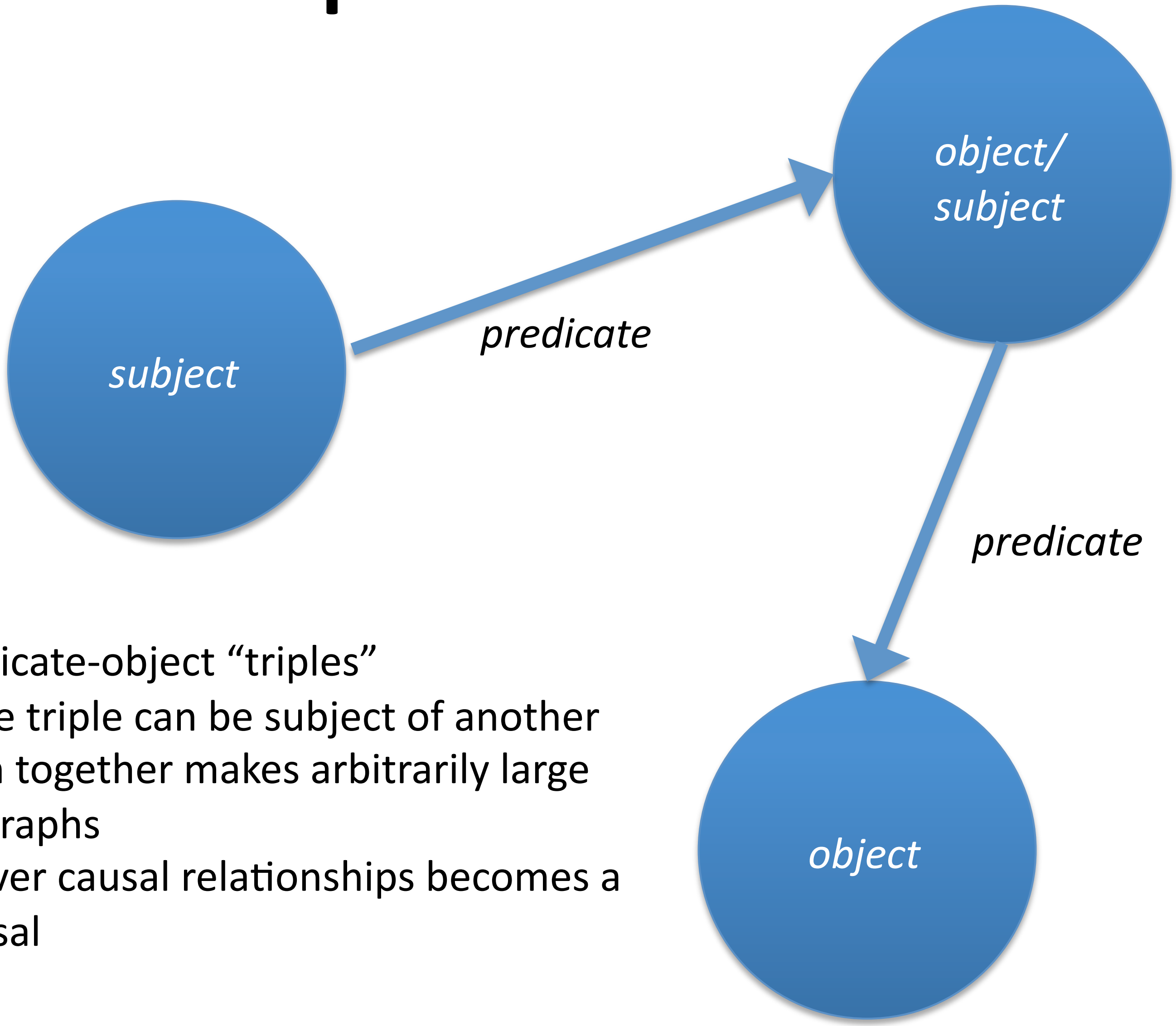
- Standard is small and easy to learn
- Easy to read and use
- Computable
  - Forms a graph-based knowledge base
- Supports both causal and correlative relationships as well as negative relationships
  - Suitable for recording a variety of experimental and clinical findings
- Can be used with almost any set of vocabularies and ontologies
  - Highly adaptable and easy to adopt
- Can be easily extended to annotate findings with use-specific contexts such as experimental and clinical parameters

BEL Language Structure



"PMA increases the kinase activity of p38 MAPKs"

BEL Forms Graphs



- Subject-predicate-object “triples”
- Object of one triple can be subject of another
- Putting them together makes arbitrarily large knowledge graphs
- Reasoning over causal relationships becomes a graph traversal

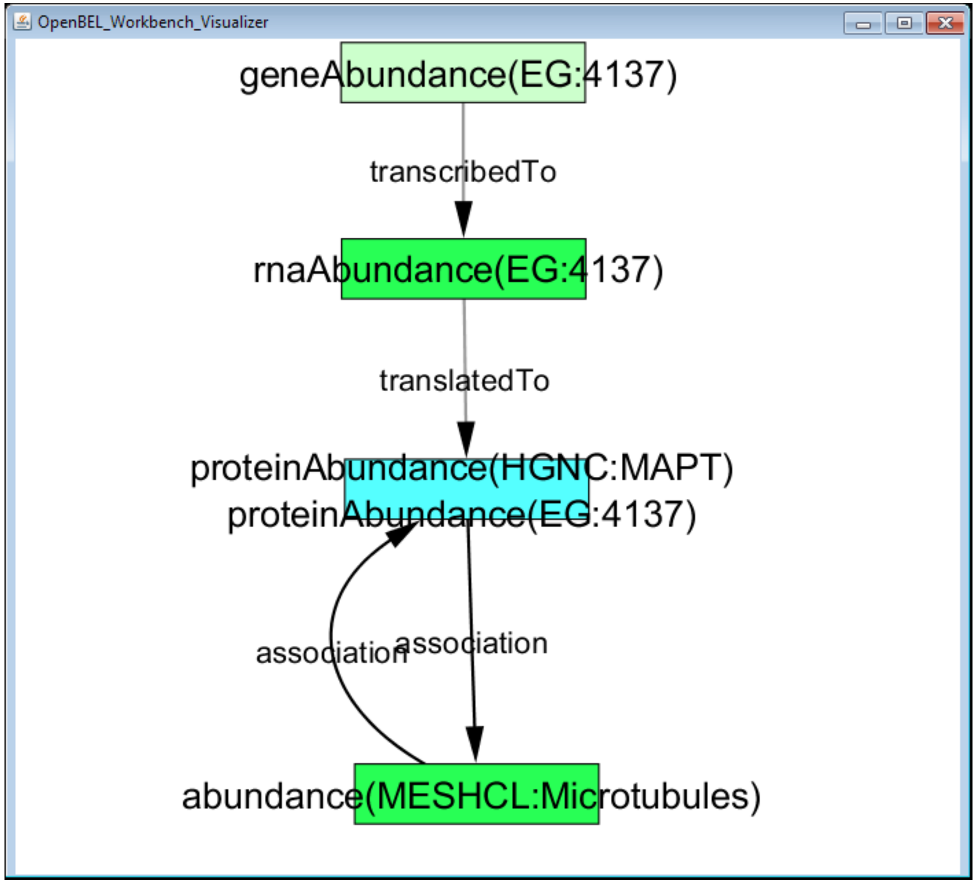
Example 3

```
SET Citation={"PubMed","Trends Neurosci. 2008 Sep;31(9):454-63. Epub 2008 Jul 31","18675468"}
SET Evidence = "In healthy neurons the axon contains relatively high amounts of microtubules which are stabilized by the protein tau."
SET Cell = "Neurons"
SET CellStructure = "Axons"
```

- Option 1 – Use association to connect Tau and microtubules
  - `p(HGNC:MAPT) -- a(MESHCL:Microtubules)`
- Option 2 – More detailed relationships, provide causal flow and binding information
  - `p(HGNC:MAPT) =| bp(GO:"microtubule depolymerization")`
  - `bp(GO:"microtubule depolymerization") =| a(MESHCL:Microtubules)`
  - `complex(p(HGNC:MAPT), a(MESHCL:Microtubules))`

Example 3

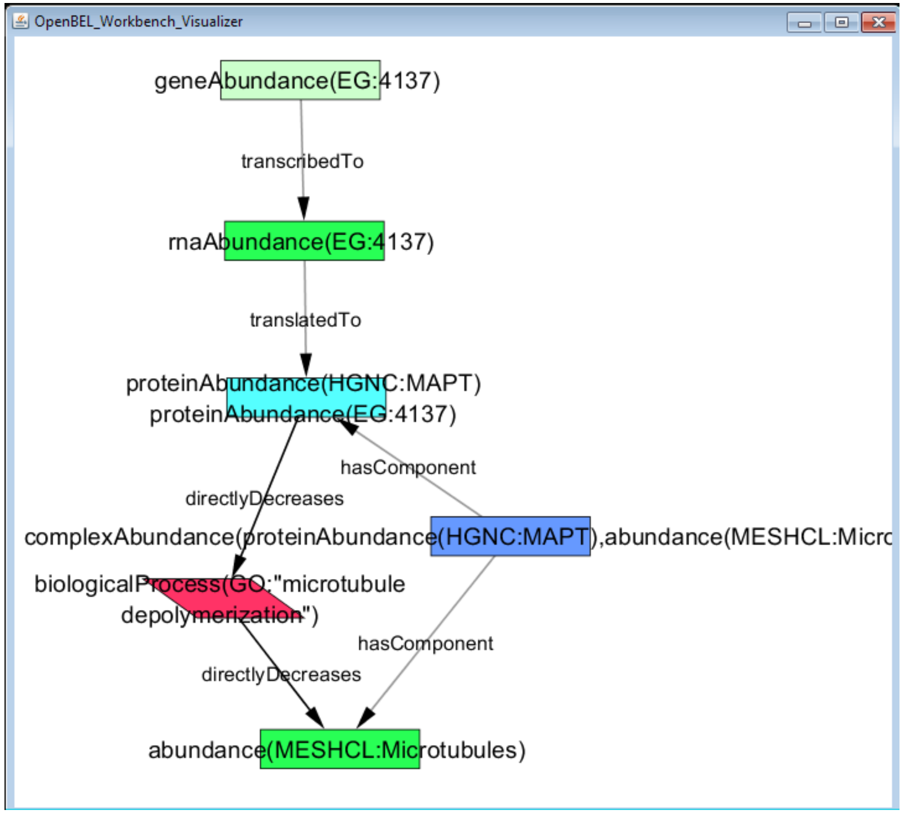
- Option 1 – Use association to connect Tau and microtubules
  - `p(HGNC:MAPT) -- a(MESHCL:Microtubules)`


OpenBEL_Workbench_Visualizer
geneAbundance(EG:4137)
transcribedTo
rnaAbundance(EG:4137)
translatedTo
proteinAbundance(HGNC:MAPT)
proteinAbundance(EG:4137)
association
association
abundance(MESHCL:Microtubules)

- Visualized in Cytoscape with Phase III
- Note expansion of association to include reciprocal relationship
  - `a(MESHCL:Microtubules) \`
    `-- p(HGNC:MAPT)`
- Note gene activation pathway
  - `g(EG:4137) transcribedTo \ r(EG:4137)`
  - `r(EG:4137) translatedTo \ p(EG:4137)`

Example 3

- Option 2 – More detailed relationships provide causal flow and binding information
  - `p(HGNC:MAPT) =| bp(GO:"microtubule depolymerization")`
  - `bp(GO:"microtubule depolymerization") =| a(MESHCL:Microtubules)`
  - `complex(p(HGNC:MAPT), a(MESHCL:Microtubules))`


OpenBEL_Workbench_Visualizer
geneAbundance(EG:4137)
transcribedTo
rnaAbundance(EG:4137)
translatedTo
proteinAbundance(HGNC:MAPT)
proteinAbundance(EG:4137)
hasComponent
directlyDecreases
complexAbundance(proteinAbundance(HGNC:MAPT),abundance(MESHCL:Micrc
biologicalProcess(GO:"microtubule
depolymerization")
hasComponent
directlyDecreases
abundance(MESHCL:Microtubules)

- Visualized in Cytoscape with Phase III
- Note expansion of complex with hasComponent relationships

  -
    ```
    complex(p(HGNC:MAPT), \
    a(MESHCL:Microtubules)) \
    hasComponent a(MESHCL:Microtubules)
    ```

  -
    ```
    complex(p(HGNC:MAPT), \
    a(MESHCL:Microtubules)) \
    hasComponent p(HGNC:MAPT)
    ```

An OpenBEL Model for APP Physiology

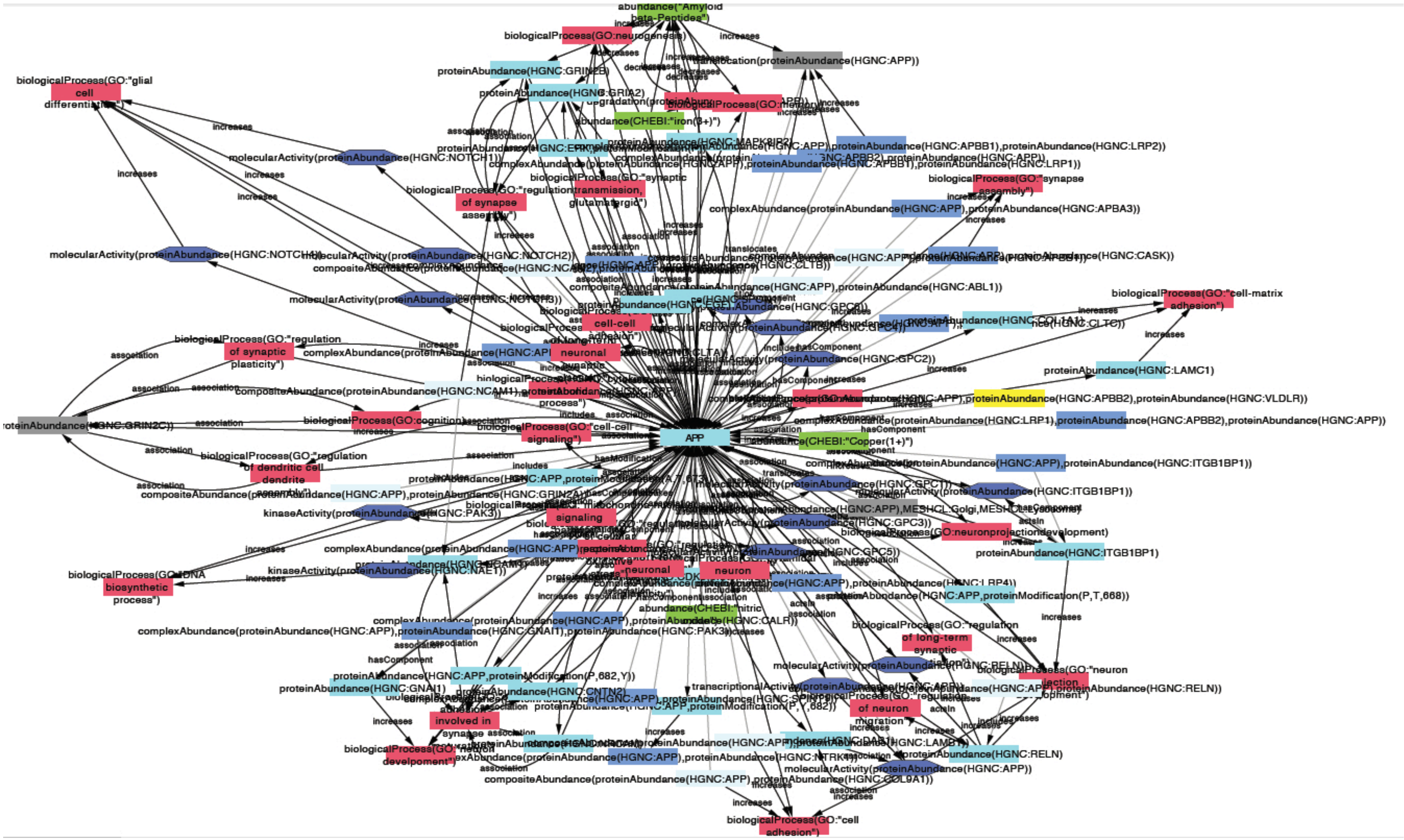
- Two models:
  - Human APP “normal physiology“
  - Human APP “Alzheimer Disease Condition”

- Modelling principles:
  - No mixing of rodent and human physiology
  - Only causal and correlative relationships
  - Issues: integration of genetics and proteolytic processing

APP Model Statistics:

- Human APP “normal physiology“
  - 8 modules representing defined pathway context
  - 965 BEL statements
  - 682 nodes and 1387 edges

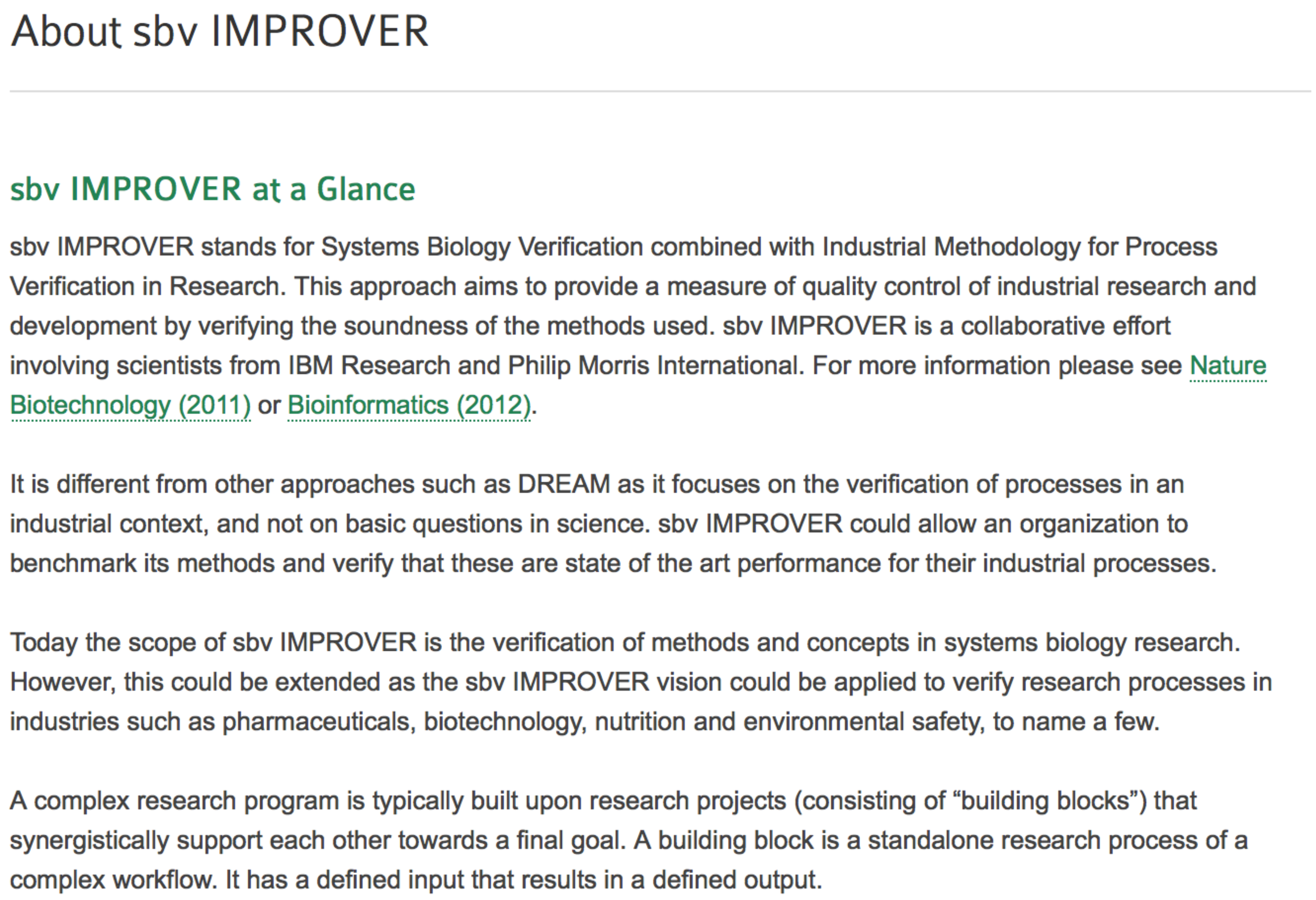
- Human APP “Alzheimer Disease Condition”
  - 8 modules representing defined pathway context
  - 1035 BEL statements
  - 1301 nodes and 3116 edges


APP
abundance("Amyloid beta-Peptides")
biologicalProcess(GO:neurogenesis)
translocation(proteinAbundance(HGNC:APP))
proteinAbundance(HGNC:GRIN2B)
proteinAbundance(HGNC:GRIA2)
biologicalProcess(GO:"glial cell differentiation")
biologicalProcess(GO:memory)
abundance(CHEBI:"iron(3+)")
molecularActivity(proteinAbundance(HGNC:NOTCH1))
proteinAbundance(HGNC:MAPK8IP2)
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:APBB1),proteinAbundance(HGNC:LRP2))
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:APBB1),proteinAbundance(HGNC:LRP1))
biologicalProcess(GO:"synaptic transmission, glutamatergic")
biologicalProcess(GO:"regulation of synapse assembly")
biologicalProcess(GO:"synapse assembly")
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:APBA3))
molecularActivity(proteinAbundance(HGNC:NOTCH4))
molecularActivity(proteinAbundance(HGNC:NOTCH2))
compositeAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:CASK))
compositeAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:ABL1))
molecularActivity(proteinAbundance(HGNC:NOTCH3))
biologicalProcess(GO:"cell-matrix adhesion")
proteinAbundance(HGNC:EGF)
proteinAbundance(HGNC:COL1A1)
biologicalProcess(GO:"regulation of synaptic plasticity")
complexAbundance(proteinAbundance(HGNC:APP))
molecularActivity(proteinAbundance(HGNC:GPC2))
proteinAbundance(HGNC:LAMC1)
compositeAbundance(proteinAbundance(HGNC:NCAM1),proteinAbundance(HGNC:APP))
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:APBB2),proteinAbundance(HGNC:VLDLR))
proteinAbundance(HGNC:GRIN2C)
biologicalProcess(GO:cognition)
biologicalProcess(GO:"cell-cell signaling")
compositeAbundance(proteinAbundance(HGNC:LRP1),proteinAbundance(HGNC:APBB2),proteinAbundance(HGNC:APP))
abundance(CHEBI:"Copper(1+)")
biologicalProcess(GO:"regulation of dendritic cell dendrite assembly")
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:ITGB1BP1))
molecularActivity(proteinAbundance(HGNC:GPC1))
molecularActivity(proteinAbundance(HGNC:ITGB1BP1))
compositeAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:GRIN2A))
kinaseActivity(proteinAbundance(HGNC:PAK3))
proteinAbundance(HGNC:APP),MESHCL:Golgi,MESHCL:Lysosomes
molecularActivity(proteinAbundance(HGNC:GPC3))
biologicalProcess(GO:neuronprojectiondevelopment)
proteinAbundance(HGNC:ITGB1BP1)
complexAbundance(proteinAbundance(HGNC:APP)
molecularActivity(proteinAbundance(HGNC:GPC5))
biologicalProcess(GO:DNA biosynthetic process")
kinaseActivity(proteinAbundance(HGNC:NAE1))
neuron
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:LRP4))
proteinAbundance(HGNC:APP,proteinModification(P,T,668))
abundance(CHEBI:"nitric oxide")
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:CALR))
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:GNAI1),proteinAbundance(HGNC:PAK3))
biologicalProcess(GO:"regulation of long-term synaptic")
molecularActivity(proteinAbundance(HGNC:RELN))
biologicalProcess(GO:"neuron projection development")
proteinAbundance(HGNC:APP,proteinModification(P,682,Y))
proteinAbundance(HGNC:GNAI1)
proteinAbundance(HGNC:CNTN2)
transcriptionalActivity(proteinAbundance(HGNC:APP))
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:RELN))
biologicalProcess(GO:"regulation of neuron migration")
biologicalProcess(GO:"neuron development")
involved in synapse
compositeAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:LAMB1))
proteinAbundance(HGNC:RELN)
complexAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:NTRK1))
molecularActivity(proteinAbundance(HGNC:APP))
compositeAbundance(proteinAbundance(HGNC:APP),proteinAbundance(HGNC:COL9A1))
biologicalProcess(GO:"cell adhesion")
increases
decreases
association
hasComponent
includes
translocates
actsIn
hasModification

Fraunhofer
SCAI

IMPROVER:

An Industrial Initiative for the Verification of Systems Biology Models



About sbv IMPROVER

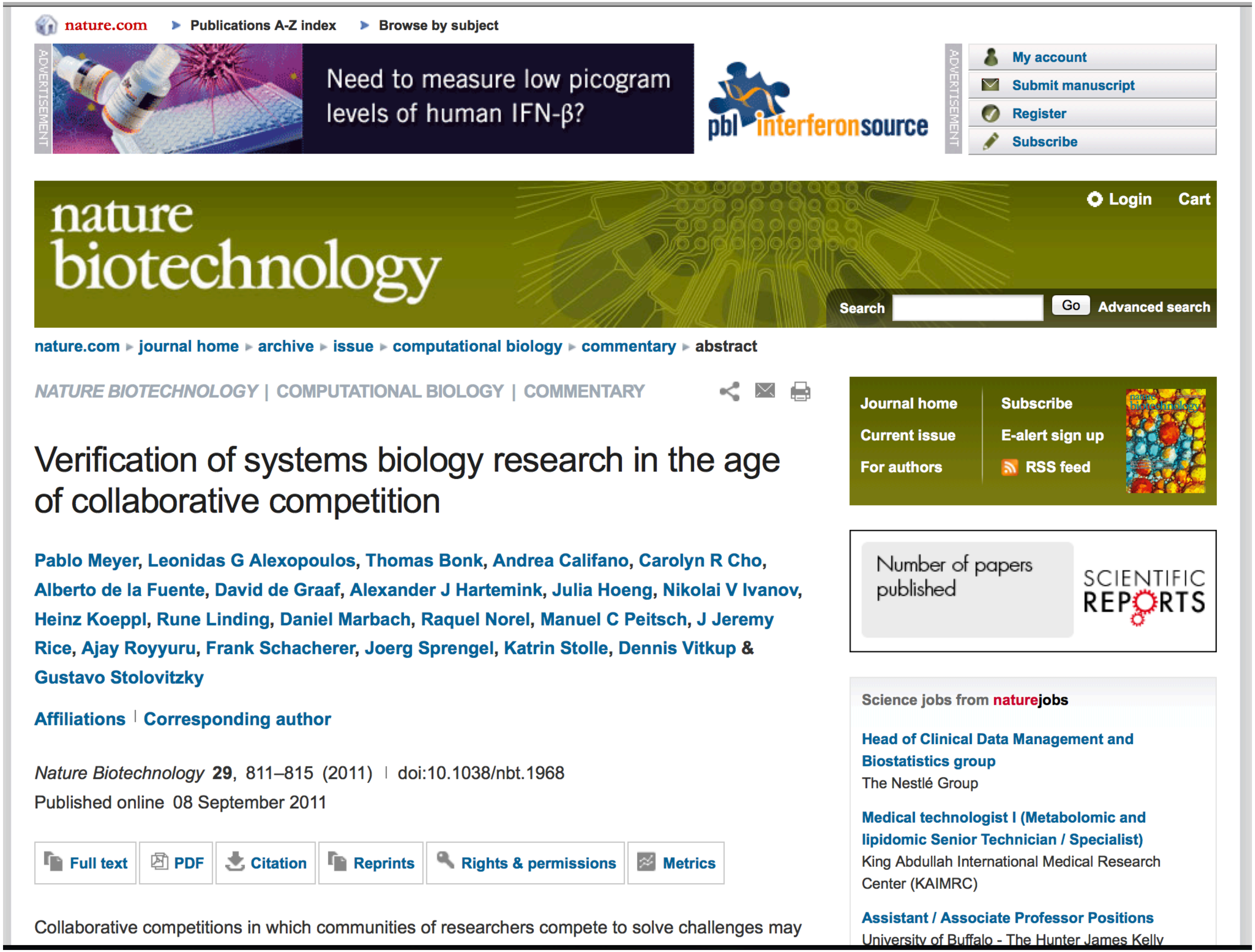
sbv IMPROVER at a Glance

sbv IMPROVER stands for Systems Biology Verification combined with Industrial Methodology for Process Verification in Research. This approach aims to provide a measure of quality control of industrial research and development by verifying the soundness of the methods used. sbv IMPROVER is a collaborative effort involving scientists from IBM Research and Philip Morris International. For more information please see Nature Biotechnology (2011) or Bioinformatics (2012).

It is different from other approaches such as DREAM as it focuses on the verification of processes in an industrial context, and not on basic questions in science. sbv IMPROVER could allow an organization to benchmark its methods and verify that these are state of the art performance for their industrial processes.

Today the scope of sbv IMPROVER is the verification of methods and concepts in systems biology research. However, this could be extended as the sbv IMPROVER vision could be applied to verify research processes in industries such as pharmaceuticals, biotechnology, nutrition and environmental safety, to name a few.

A complex research program is typically built upon research projects (consisting of “building blocks”) that synergistically support each other towards a final goal. A building block is a standalone research process of a complex workflow. It has a defined input that results in a defined output.



NATURE BIOTECHNOLOGY | COMPUTATIONAL BIOLOGY | COMMENTARY

Verification of systems biology research in the age of collaborative competition

Pablo Meyer, Leonidas G Alexopoulos, Thomas Bonk, Andrea Califano, Carolyn R Cho, Alberto de la Fuente, David de Graaf, Alexander J Hartemink, Julia Hoeng, Nikolai V Ivanov, Heinz Koeppl, Rune Linding, Daniel Marbach, Raquel Norel, Manuel C Peitsch, J Jeremy Rice, Ajay Royyuru, Frank Schacherer, Joerg Sprengel, Katrin Stolle, Dennis Vitkup & Gustavo Stolovitzky

Affiliations | Corresponding author

Nature Biotechnology **29**, 811–815 (2011) | doi:10.1038/nbt.1968
Published online 08 September 2011

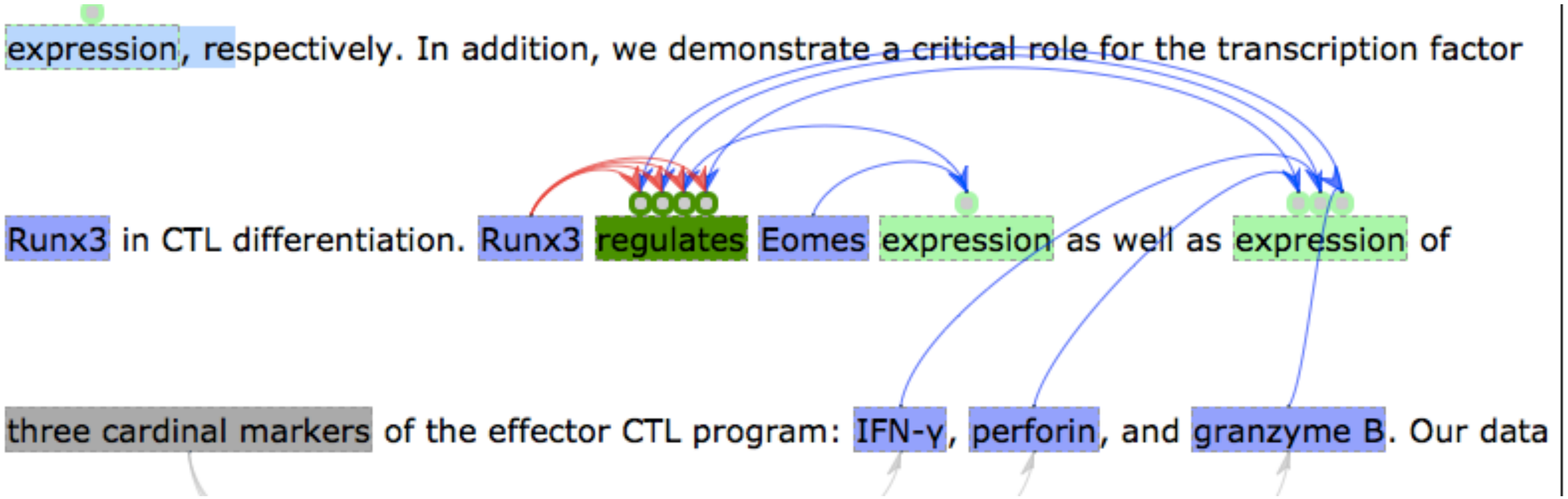
Full text | PDF | Citation | Reprints | Rights & permissions | Metrics

Collaborative competitions in which communities of researchers compete to solve challenges may

sbv IMPROVER
Symposium 2013

Species Translation Challenge,
Athens Greece, 29-31 October 2013

Automated Recognition and Extraction of Causal Statements



expression, respectively. In addition, we demonstrate a critical role for the transcription factor Runx3 in CTL differentiation. Runx3 regulates Eomes expression as well as expression of three cardinal markers of the effector CTL program: IFN-γ, perforin, and granzyme B. Our data

How do we Assess the Completeness and Correctness of Systems Biology Models?

Features for a Scoring Function

- xxxxxxxxxxxxxxxxxxxxx
- xxxxxxxxxxxxxxxxxxxxx
- xxxxxxxxxxxxxxxxxxxxx
- xxxxxxxxxxxxxxxxxxxxx

The Scoring Function at Work

Picture deleted ..... Sorry

Results too recent to share them with a wider community

We are willing and happy to share our results through a journal publication within the next year

Take Home Message

- Text Mining Technology allows for direct usage of extracted information for systems biology models / disease models
- The Biological Expression Language BEL is an easy-to-learn formalism ideally suited for capturing knowledge and knowledge – based modeling of complex diseases
- The industrial IMPROVER challenge addresses the issue of systems biology model verification, SCAI contributes a BEL extraction engine
- Fraunhofer SCAI has developed a complex scoring function to assess the completeness and correctness of disease models