Computing Pathways in Bio-Models Derived from Bio-Science Text Sources

Troels Andreasen[1],Henrik Bulskov[1],
Jørgen Fischer Nilsson[2], Per Anker Jensen[3]

[1]Computer Science, Roskilde University,
[2]Mathematics and Computer Science , Technical University of Denmark,
[3]International Business Communication, Copenhagen Business School
{troels,bulskov}@ruc.dk, jfni@dtu.dk, paj.ibc@cbs.dk

Abstract. This paper outlines a system, OntoScape, serving to accomplish complex inference tasks on knowledge bases and bio-models derived from life-science text corpora. The system applies so-called natural logic, a form of logic which is readable for humans. This logic affords ontological representations of complex terms appearing in the text sources. Along with logical propositions, the system applies a semantic graph representation facilitating calculation of bio-pathways. More generally, the system affords means of query answering appealing to general and domain specific inference rules.

Keywords: Semantic text processing in bio-informatics; bio-models using natural logic and semantic graphs; pathway computation

1 Introduction

This paper addresses logic-based knowledge base bio-models derived from life science corpora. We discuss representation languages and reasoning principles for bio-models derived from actual life science sources. In particular, we describe and exemplify the intended query answering and pathway functionality, that is, the ability to compute conceptual pathways in the stored model.

This paper is a companion paper to our [1] and a follow-up on our [2]. The former paper describes the internal technical details of extracting information from life science corpora as well as how the representation and reasoning is performed computationally using the information.

2 Class Relationship Bio-models

Fundamentally, we model bio-systems formally as relationships between classes of concrete and abstract individuals. The classes comprise physical objects such as cells and organs and substances and amounts of substances such as blood. There are also classes comprising abstract entities such as properties and events, phenomena and processes.

The relationships in the model comprise static ones like taxonomic class inclusion and partonomic relationships, as well as dynamic interactions like causation. Further, there are assorted, concrete effect relationships such as transport of objects and substances. However, our approach at present does not make provision for narratives such as sequences of events: The knowledge base sentences are in principle unordered.

At first sight our approach resembles the well-known, rather simplistic entity-relationship models and RDF representations. However, our framework is unique in various respects, first of all in its *generativity*, that is, the ability of the models to accommodate arbitrarily complex concepts formed by composition of lexicalized classes and relationships as discussed in [3]. By way of examples, the virtually open-ended supply of concept terms such as *'cell in the liver that secretes hormone'*, *'arteria in pancreas'*, *'secrete from the exocrine pancreas'* are accomodated in the model by composing simple, given class terms into compound concept terms with an obvious resemblance to phrases in natural language, as these examples illustrate. All such encountered concepts are situated in the so-called 'generative ontology' in a manner so that they can be "de-constructed" and reasoned with computationally.

The generativity and the liaison to natural language specifications is achieved by adopting so-called 'natural logic' cf. [4–6] as the logical model language. In addition, the models come in the form of graphs with concepts as nodes and relations as edges. These two mutually supporting representation media for the bio-models are described briefly in the next sections. In section 5 we consider model fragments drawn from various medical text sources. In section 6 we exemplify and explain the pathway computation functionality of our system.

3 Models in Natural Logic

In the applied natural logic conception, a knowledge base or specification consists of a collection of descriptive sentences called 'propositions' in order to distinguish them from the natural language sentences from which they are derived. Propositions in the applied logic, dubbed NaturaLog, are of the following general form

$$Conterm_1\ Relterm\ Conterm_2$$

where

- The two *Conterm*s are atomic or compound concept terms.
- The *Relterm* is a relational term, in the simplest cases corresponding to a transitive verb, e.g. '*cause*', '*secrete*'. *Relterm*s are further used for qualification of concepts by prepositional phrases with prepositions like '*in*', '*via*' etc.

In the sample logical proposition betacell secrete insulin the two concept terms are atomic, and so is the intervening relational term. A bio-model or knowledge base comprising also the class inclusion ontology is made up of a finite, albeit possibly huge, collection of such NaturaLog propositions. As it appears, we use

sans serif font for these throughout. This model is then the basis for inferences and querying.

Propositions may contain complex structures: Compound *Conterms* consist of a class C with attached qualifications. In the more complex sample proposition

(cell that secrete insulin) is-located-in (pancreatic gland).

the first concept term consists of the atomic term cell adorned with a relative clause consisting of the relational term secrete followed by the concept term insulin. Relative clauses are indicated by the optional keyword 'that', merely to ease the reading. Relative clauses are assumed always to act restrictively. For instance, as a matter of principle cell that secrete insulin is recorded by the system as a sub-concept of cell in the concept inclusion structure in the ontology. Likewise, the second concept term pancreatic gland, is recognized as a sub-concept of the class gland in that all adjectives are also assumed to be interpreted restrictively. Parentheses are inserted for ease of reading and serve to ensure disambiguation. They may be omitted if there is no risk of ambiguity.

However, sub-class - and, more generally, sub-concept relationships may also be specified explicitly, namely by the relation term isa, corresponding to copula sentences. Example: betacell isa cell. By contrast, the propositions (cell that secretes insulin) isa cell and (pancreatic cell) isa cell are inferred by the system according to the principles mentioned. Still, (pancreatic cell) isa (cell located-in pancreas) (and *vice versa*) has to be provided.

As it appears, the natural logic propositions are perfectly readable, if somewhat stereotypical, by domain experts by virtue of their resemblance to natural language. The converse, challenging task of automating translation from manageable parts of natural language in scientific text sources into natural logic is approached in our [1].

3.1 Quantifiers and Recursion in Concept Terms

The above propositional form $Conterm_1\ Relterm\ Conterm_2$ is actually a special case of

$$Q_1\ Conterm_1\ Relterm\ Q_2\ Conterm_2$$

where the Qs are quantifiers, primarily 'all/every' or 'some'. Usually the quantifiers are absent with Q_1 then being interpreted as all and Q_2 as some by default. Accordingly, the example betacell secrete insulin is interpreted logically as the proposition all betacell secrete some insulin, where some insulin is meant to be some portion or amount of insulin. Generally speaking, the entities in a class of substance are taken to be all arbitrary, non-empty amounts of the substance.

This propositional form all $Conterm_1\ Relterm$ some $Conterm_2$ corresponds to the predicate logic formula $\forall x(Conterm_1[x] \rightarrow \exists y(Relterm[x,y] \wedge Conterm_2[y]))$, see further [1, 2], where we also discuss the relationship to description logic.

The introduced NaturaLog forms cover only those parts of binary predicate logic which are considered relevant for bio-modelling. Notable exclusions are logical negation and logical disjunctions. We plan to support a limited form of negation by way of the Closed World Assumption as in the concept term:

hormone that not secreted from endocrine gland, which might serve as a query to the knowledge base. The relational term secreted here is the reverse relation of secrete.

Recall that a concept term consists of a class C followed by one or more qualifications or restrictions, where restrictions consist of a relational term followed by a concept term: *Relterm Conterm*. In case of more than one restriction, these are to form a conjunction with and understood as logical conjunction proper. By contrast, two and-aligned concept terms within the same class are conceived of as a logical disjunction. Accordingly, concept terms have a finitely nested, recursive structure reflecting the syntax of natural language nominal phrases with possibly nested relative clauses and prepositional phrases. The handling of adjectives (ex. pancreatic gland) and compound nouns (ex. lung symptom), both assumed to be acting restrictively, as well as genitives is to be discussed elsewhere.

3.2 Ontologies

As mentioned above, a special case of the above propositions are class inclusion relationships corresponding to stylized copula sentences. For example, in the proposition pancreas isa (endocrine gland)isa is concept inclusion. The synonymy relation syn is construed as both way isa, cf. the declaration pancreas syn (pancreatic gland). Such propositions form the backbone of the ontology in our knowledge-based bio-models. Also partonomic propositions like betacell part-of (endocrine pancreas) are included in the ontology; cf. [7] for the various partonomic relations.

By contrast, a proposition like betacell secrete insulin is understood as an observational fact, an assertion, and therefore does not belong to the ontology proper. The concept of betacell would then be expected to be defined in some other way, which may or may not be part of the bio-model. However, the stated assertion might be replaced by the definitional proposition (cell that secrete insulin) syn betacell at the discretion of the domain expert. This proposition posits that all cells that secrete insulin (whatever their location), are to be called betacells.

4 Bio-models as Semantic Graphs

In our framework, the natural logic propositions constituting a bio-knowledge base are parallelled by an alternative representation in the form of directed graphs as commonly used in bio-models [8–10]. The graphs come about by decomposing compound and relational concept terms into their constituents in the form of triples [4]. These triples are re-conceived of as labeled directed edges between nodes. Every concept is associated with one node and *vice versa*.

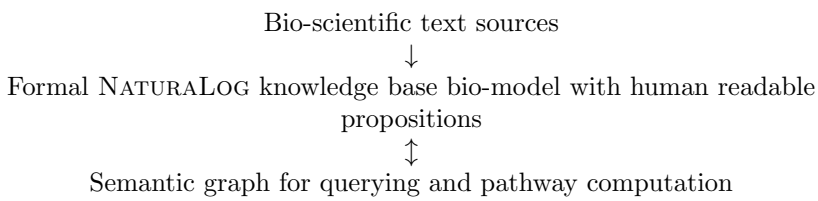
This semantic graph representation, which superficially resembles RDF, facilitates computation of relevant associations between concepts, namely by computation of connecting paths in the graph. For example, the subject concept in the proposition (cell that secrete insulin) located-in pancreas corresponding to

the natural language sentence *cell that secretes insulin is located in pancreas* is internally decomposed into the two triples

(cell-that-secrete-insulin) isa cell.
(cell-that-secrete-insulin) produce insulin.

where the added auxiliary concept (cell-that-secrete-insulin) is conceived of as an atomic name of a node defined by the two triples. An arc symbol as in '$\prec$' is inserted between the defining edges in the graph rendition to express that they form the definition of the concept, in casu (cell-that-secrete-insulin).

The given proposition, which is epistemically in observational mode, then becomes represented by the triple (cell-that-secrete-insulin) located-in pancreas. So in this way a distinction is made between definitional and assertive (observational) propositions. This ensures that the original propositions, whatever their complexity, can be regained from the semantic graph as indicated in the following figure:



5 Fragments of a Bio-model: A Case Study

Below follows logical knowledge base representation of excerpts from Wikipedia articles on endocrine glands and [11, 12]. The fragments concern the endocrine system and the gallbladder, and they are stated in an extended, relaxed form of NaturaLog, cf. [1].

The propositions set in parentheses below are present only implicitly in the considered corpora. Some propositions introduce sub-concepts by agglutination rather than by the use of separate words, calling for manual treatment. Conversely, some would-be compounds like *islet of Langerhans* and *Graves' disease* should not be decomposed, but should be kept as atomic class names.

endocrine gland secrete hormone.
gland isa organ.
(hormone isa protein.)
pituitary gland and pancreas and thyroid gland and adrenal glands
isa endocrine gland.
hypothalamus isa neuroendocrine gland.
(neuroendocrine gland isa endocrine gland.)
stomach produces grehlin, a hormone.[1]

[1] The apposition becomes a separate proposition, grehlin isa hormone.

target cell is affected by hormone.
liver and kidneys degrade [all] hormone.[2]

Thyroid gland

thyroid hormone derives from thyroglubin, a glycoprotein.
(glycoprotein isa protein.)
thyroxine and triiodothyronine isa thyroid hormone.
thyroxine syn T4.
triiodothyronine syn T3.
thyroxine and triiodothyronine increase cellular metabolism.
increased cellular metabolism increases oxygen use and heat production.
TSH causes secretion of thyroid hormone.
increased level of thyroid hormone inhibit
    (secretion from) pituitary gland and hypothalamus.
T3 and T4 promote gene (synthesis) and protein synthesis.
(hyperthyroidism syn increased secretion from thyroid gland.)
hyperthyroidism causes Graves' disease.
(hypothyroidism syn decreased secretion from thyroid gland.)
hypothyroidism causes cretinism in infants and myxedema in adults.
calcitonin reduces level of calcium in the blood by
    (inhibition of resorption in bone matrix) and
    (increase of deposit of calcium in bone).
increased level of calcium in the blood causes production of calcitonin.
calcitonin is-produced-by parafollicular cell in the thyroid gland.

Parathyroid glands

parathyroid glands secrete parathyroid hormone.
PTH syn parathyroid hormone.
parathyroid hormone causes increase of levels of calcium in blood.
decreased level of calcium in the blood causes release of PTH.
increased level of calcium in the blood inhibits release of PTH.

Pancreas

pancreas isa endocrine gland and exocrine gland.
(pancreas haspart endocrine gland and exocrine gland.)
endocrine pancreas has part islets-of-Langerhans.
Islets of langerhans release insulin and glucagon to the blood.
(alphacell and betacell are-located in (islet-of-Langerhans).)
alphacell (releases when low level of glucose in blood) glucagon.
glucagon promote (release of glucose from the liver).
betacell (release when increased level of glucose in blood) insulin.[3]

[2] The relation degrade having a negative connotation suggests use of the quantifier all instead of some in the linguistic object.
[3] The relation term is here compounded, forming a sub-relation of release.

insulin promotes (uptake of glucose) and metabolism in body cells.
reduced secretion of insulin causes diabetes mellitus.
polyuria and polydipsia and polyphagia isa symptom of diabetes mellitus.[4]

Pancreatitis

pancreatitis isa (inflammation of pancreas).
alcohol and gallstone may cause pancreatitis.
alcohol may cause chronic pancreatitis.[5]
gallstone may cause acute pancreatitis.

Backward references from propositions to the sentences from which they are derived are stored. One sentence may give rise to multiple propositions e.g. due to linguistic conjunctions, appositions, and parenthetical relative clauses. Furthermore, as seen, one proposition in general gives rise to multiple triples in the graph rendition by a decomposition introducing nodes for compound, auxiliary terms.

6 Computational Query Answering and Pathfinding

It is our tenet that bio-pathways can be computed in logical bio-models by mathematical composition of the relations corresponding to the edges in the semantic graph. This computation process is supported by logical inference rules, since inferred propositions may constitute shortcuts, as it were, in the graph view. For instance, the transitivity of inclusion, isa, conceptually shortens the distance from a concept to a superior concept in the ontology via intermediate concepts. Similarly for partonomic, causative and effect relations. In [2], the path finding is explained more abstractly as application of appropriate logical comprehension principles supporting the relation composition.

A miniature ontology, corresponding to a subset of the bio-model propositions listed in section 5, is visualised in the graph in figure 1. In addition, a candidate answer to the query

high calcium level in blood, gland?

is shown. The answer is provided as a pathway connecting the two concepts and the pathway can be seen as an explanation of how the two concepts are related. The reading of the answer can be

High calcium level in the blood causes production of calcitonin in the parafollicular cells in the thyroid gland, which is an endocrine gland, which is a gland.

[4] This proposition in the semantic graph is split into three simpler ones.
[5] The concept chronic pancreatitis is automatically bound to be recognized as a sub-concept of pancreatitis.

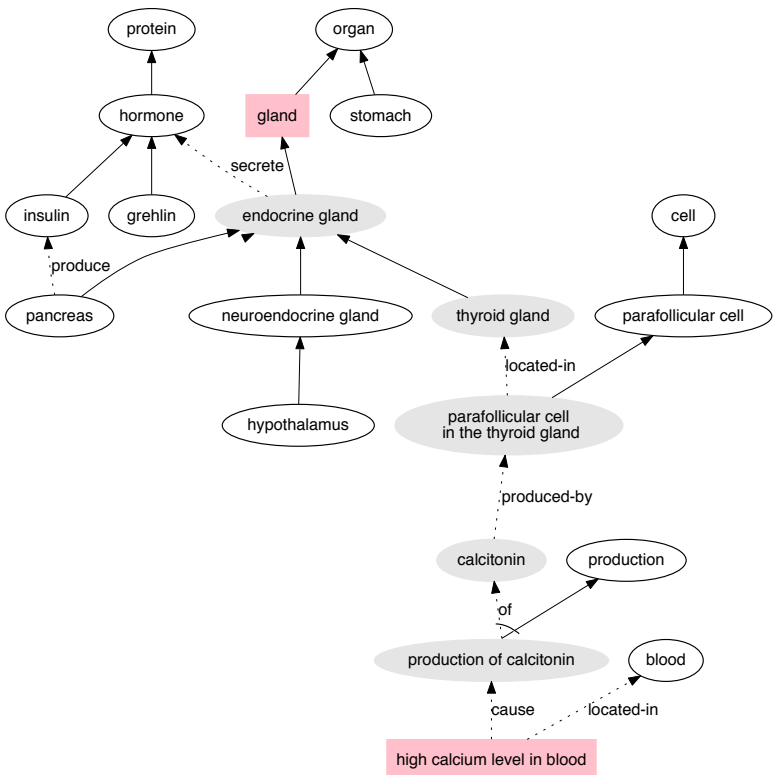


Fig. 1. A miniature ontology corresponding to a subset of the bio-model propositions listed in section 5. A pathway connecting "high calcium level in blood" and "gland" is shown. This pathway provides a candidate answer to a query specifying the two concepts (apparently the only one in this case).

In figure 2 an extended version of the ontology is shown. New knowledge has been added to the base such that the two concepts become connected by a new pathway corresponding to an alternative answer. The reading of this alternative answer can be

> *High calcium level in the blood is stimulated by parathyroid hormone, which is secreted by the parathyroid gland, which is an endocrine gland, which is a gland*".

A pathway computation, being more than a pure inferential process, in our system is also the composition of relations guided by appropriate path computation. In our framework this computation is reduced algorithmically to search for weighted paths between concept nodes in the graph representation, utilizing

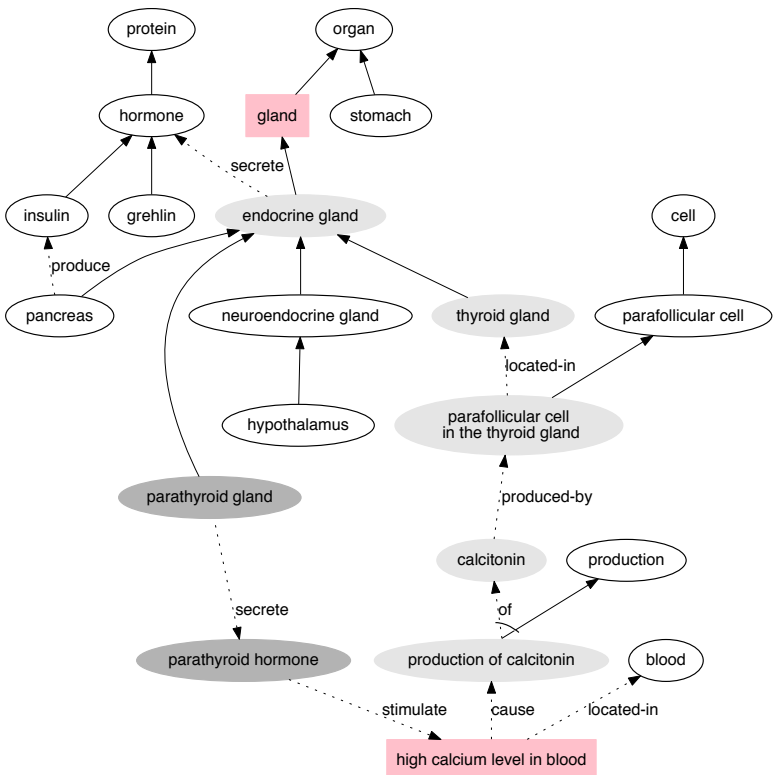


Fig. 2. A slightly extended version of the ontology shown in figure 1 now also presenting an alternative pathway that can contribute to the answer to the "high calcium level in blood", "gland" query.

standard heuristic algorithms in artificial intelligence. The intermediate propositional representations refer back to the source texts so that computed paths can be shown by highlighting excerpts in the texts.

A single concept may also be construed as a conceptual query asking for all the subsumed concepts. For example, the query

hormone produced-by endocrine gland?

would yield among others insulin, while the query

hormone not produced-by endocrine gland?

which uses an extended form of concept terms with negation (appealing to the Closed World Assumption), would yield the hormone grehlin.

7 Summary and Conclusion

We have described a system for querying and pathfinding in bio-models taking the form of logical knowledge bases derived from text sources. The applied logical language accommodates complex sentences, which can be queried by deductive means, and the supporting semantic graph form enables algorithmic pathfinding between concepts. A small scale prototype for demonstrating the functionality principles is under development. Computational translation of text sources into the logical form is a challenging problem which might be approached by adopting enriched forms of natural logic as a specification language for bio-systems.

References

1. T. Andreasen, H. Bulskov, J. Fischer Nilsson, P. Anker Jensen: Computing Conceptual Pathways in Bio-Medical Text Models In: Foundations of Intelligent Systems - 19th International Symposium, ISMIS 2014, Roskilde, Denmark, June 28-30, Proceedings (2014)
2. T. Andreasen, H. Bulskov, J. Fischer Nilsson, P. Anker Jensen, T. Lassen: Conceptual Pathway Querying of Natural Logic Knowledge Bases from Text Bases. In Proceedings of the 10th international conference on Flexible Query Answering Systems, H. Legind Larsen, M. J. Martin-Bautista, Amparo Vila, Andreasen, H. Christiansen (Eds.). Springer-Verlag, Berlin, Heidelberg (2013) 1-12
3. Andreasen, T., Bulskov, H., Jensen, P.A., Lassen, T.: Extracting Conceptual Feature Structures from Text. In: Foundations of Intelligent Systems - 19th International Symposium, ISMIS 2011, Warsaw, Poland, June 28-30, Proceedings (2011)
4. J. Fischer Nilsson, Diagrammatic Reasoning with Classes and Relationships, A. Moktefi & S.-J. Shin (eds.) *Visual Reasoning with Diagrams*, Studies in Universal Logic, Birkhäuser, Springer, 2013.
5. van Benthem, J.: Essays in Logical Semantics, Studies in Linguistics and Philosophy, Vol. 29, D. Reidel Publishing Company (1986)
6. van Benthem, J.: Natural Logic, Past And Future, Workshop on Natural Logic, Proof Theory, and Computational Semantics 2011, CSLI Stanford, `http://www.stanford.edu/~ icard/logic&language/index.html` (2011)
7. B. Smith & C. Rosse, The Role of Foundational Relations in the Aligment of Biomedical Ontologies, MEDINFO 2004, M. Fieschi *et al.*, 2004.
8. A. Vechina *et al.*, Representation of Semantic Networks of Biomedical Terms, IWBBIO 2013.
9. D. Miljkovic *et al.*, Incremental revision of biological networks from texts. IWBBIO 2013.
10. M. Quesada-Martinéz *et al.*, Analysis and Classification of Bio-ontologies by the Structure of their Labels, IWBBIO 2013.
11. Pancreatitis (2013), Retrieved March 1, 2013, from Wolters Kluwer Health's physician-authored clinical decision support resource *UpToDate* `http://www.uptodate.com/contents/acute-pancreatitis-beyond-the-basics` (2013)
12. Hypothalamus (2013), Retrieved March 1, 2013, from *emedicinehealth* `http://www.emedicinehealth.com/anatomy_of_the_endocrine_system/page2_em.htm#hypothalamus` (2013)