Evaluating and Analyzing Inconsistent RDF Data in a Semantic Dataset: EMAGE Dataset

Nwagwu Honour Chika

Cultural Communication and Computing Research Institute (C3RI)
Faculty of Arts, Computing, Engineering and Sciences
Sheffield Hallam University, United Kingdom

Honour.C.Nwagwu@student.shu.ac.uk

Abstract. This paper explains how to evaluate and analyse inconsistent Resource Description Framework (RDF) data by using EMAGE semantic (RDF) dataset as its use case. The author exploits the sub graph matching powers and mathematical functions of SPARQL query in evaluating inconsistent RDF data in a semantic dataset. He also proposes a mathematical method for calculating the amount of inconsistency in RDF data through a graph search approach. Finally, He analyzed the evaluated inconsistent RDF data.

Keywords: Triples, RDF data, Inconsistent data, Ontology, SPARQL queries

1 Introduction

EMAGE is a database of *in situ* gene expression data in the mouse embryo and an accompanying suite of tools to search and analyze the data (http://www.emouseatlas.org/emage/). EMAGE publishes *in situ* gene expression data for the developmental mouse. Its data is collected through a scrutinized process which involves assessing and tabulating of Biologist's experimental reports. These data include reports on gene expressions in mouse experiments which are reported elsewhere [11], the gene expression database (GXD), and laboratory reports among others. The Biologist's experimental report determines the strength of the expressed gene in a tissue of a mouse at a particular Theiler Stage. The Theiler stages correspond to a 28 days period associated with the developing mouse denoted by TS01 to TS28. More information about EMAGE datasets and mouse experiments can be found at the Edinburgh Mouse Atlas Project (EMAP) website [10, 12].

EMAGE's dataset can serve as a platform for Biologists to find solutions to the causes of abnormalities in organisms. Biologists can suggest answers to the cause of abnormalities in organisms through comparing the data indicating the strength of expressed gene in a healthy organism with that of unhealthy organism [9]. Neverthe-

less, data from some of the experiments which provide Biologists with this needed information can sometimes be inconsistent and these inconsistencies could be as a result of experimental error or simply a slight variation in experimental conditions [8]. Also, the accuracy of a dataset with inconsistent information can be increased through deleting the inconsistent data but at the cost of an increase in the incompleteness of the dataset. This cost can be avoided or minimized by properly evaluating and analyzing the degree of the inconsistency in the dataset. The author has explained how the inconsistency of RDF data can be identified, evaluated and analyzed. He has achieved this by explaining what RDF data model is in section 2.0, Identifying inconsistent RDF data in EMAGE dataset in section 3.0, Evaluating and analyzing inconsistent RDF data in section 4.0 and finally, the author presents his approach on how inconsistent RDF data can be evaluated and analyzed in section 5.0.

2 RDF data model

Information in semantic dataset is represented by RDF data in the form of triples and stored in a triple store. A triple consists of subject, predicate and an object. An illustration of a RDF triple is as shown in figure 1 below.

```
<http://www.cubist_project.eu/HWU#tissue_EMAP_42>
<http://www.w3.org/2001/01/rdf-schema#label> "embryo" .
```

Figure 1: A triple in EMAGE dataset

Each triple in RDF dataset represents a statement of a relationship between the entities denoted by the nodes that it links. RDF data can contain one or more triples. Each triple is composed of a subject, predict and an object. In RDF data, each subject of a triple is represented by a Universal Resource Identifier (URI) or blank node, each predicate is represented by a URI and each object node is represented by a URI, a blank node or a literal. For example in figure 1, the subject of the triple is a URI "http://www.cubist_project.eu/hwu#tissue_EMAP_42", the predicate is a URI "http://www.w3.org/2000/rdf-schema#label", and the object node is a literal "embryo". The author adopted turtle serialization format (http://www.w3.org/TR/turtle/) in this example. RDF data has other serialization formats for representing its data such as N-Triple, N3, RDF/XML and RDFa.

3 Identifying inconsistent RDF data in EMAGE dataset: SPARQL Query Language

Inconsistency exists in RDF data when the data does not conform to the rules governing their design. This is evident when there is a contradiction in the RDF data such that the RDF data contains both A and ¬A.

Inconsistency in EMAGE dataset is identified through identifying data which do not conform to EMAGE's textual annotation rules. These rules include the general "detected somewhere in" and "not detected everywhere in" rules which are used to propagate gene expression levels up and down the hierarchical structure of a particular EMAP anatomy. In addition, the expression level of a gene in a particular structure of a given Theiler stage in EMAGE dataset is reasoned through propagation approach. Through propagation approach, the associated level of gene expression in tissues that exhibit "is_part_of" relationship with other tissue(s) within a particular structure are propagated up or down the given structure in line with the chosen level of gene expression of that structure. As a consequence, gene expression levels could be inconsistent. This can be as a result of positive propagation (expressions propagated up the anatomy) that contradicts with an experimental result or negative propagation (expressions propagated down the anatomy) that contradicts with an experimental result. Also, gene expression can be completely contradictory (two experiments on the same tissue in which a gene is stated as detected in one experiment and not detected in the second experiment) or partly contradictory (two experiments on the same tissue in which the genes detected have different expression levels). Also, Inconsistency in EMAGE datasets has been categorized and defined [9] as either binary inconsistency: gene that is both expressed and not expressed in a given tissue of a Theiler stage and analogue inconsistency: involving varied strength levels of a particular gene in a given tissue of a Theiler stage.

In other to identify inconsistent RDF data, a subset of EMAGE RDF model dataset was stored in OWLIM-SE triple store (http://www.ontotext.com/owlim). The investigated dataset has 1,216,277 triples. The author applied appropriate SPARQL queries as to retrieve inconsistent data from the stored RDF dataset. He was able to detect binary inconsistency in the investigated dataset in some tissues which have "is_part_of" relationship with other tissues of the same hierarchical annotation structure. In these tissues, a gene is specified as "detected" and also specified as "not detected" in their related tissue. An example of EMAGE hierarchical annotation structure is shown in the figure 2 below. The SPARQL query in figure 3 identifies RDF data with binary inconsistency from Theiler stage 15 of the investigated HWU RDF model dataset. It can also be applied to any other Theiler stage by changing the Theiler stage number in the statement under label #3 of the query. Table 1 displays the result set. The author used the hash key (#) together with a unique number in the SPARQL query to identify comments that explain the SPARQL statement(s).

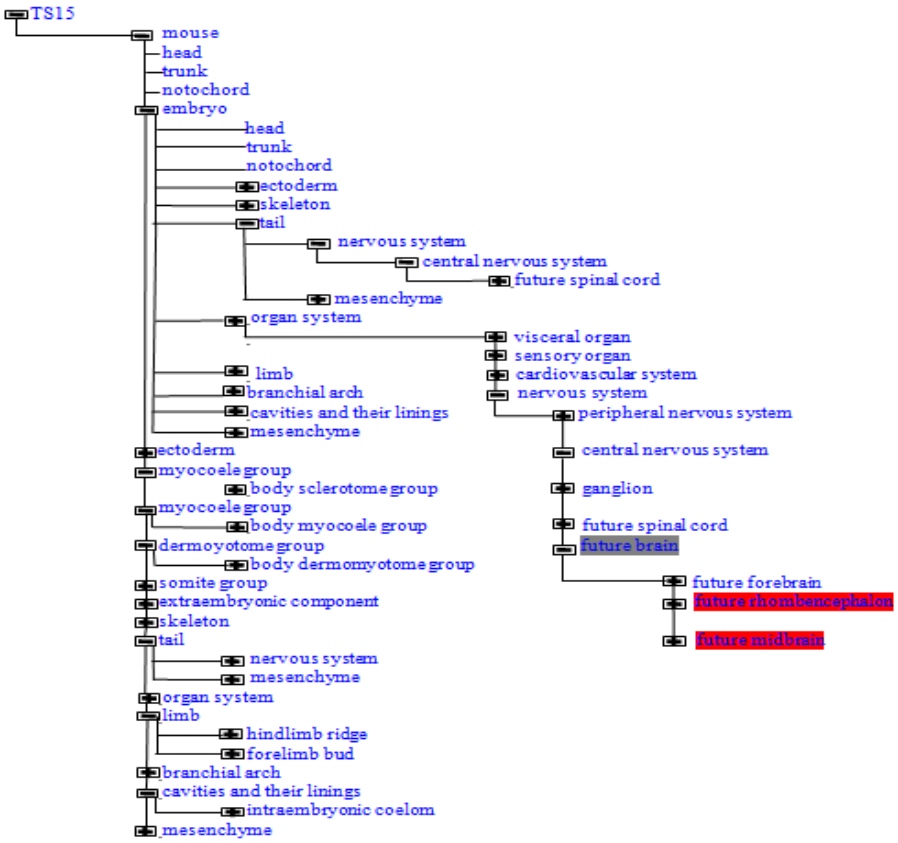

Figure 2: A subset of the Anatomy Ontology of Theiler stage 15 (drawn from http://www.emouseatlas.org/emap/ema/home.html)

To illustrate the different types of inconsistent data in the investigated dataset, the author used instances from Theiler stage 15. Figure 2 shows a subset of EMAP anatomy of Theiler stage 15.

Table 1: Binary inconsistent tissue experiments of Theiler stage 15

Gene_label	T_label	T_Experiment_label	Gene_strength	T2_label	T2_Experiment_label	Gene_strength2
Pax2	future midbrain	EMAGE:3530	hwu:level_detected	future brain	EMAGE:984	hwu:level_not_detected
Pax2	future midbrain	EMAGE:3879	hwu:level_detected	future brain	EMAGE:984	hwu:level_not_detected
Pax2	future rhombencephalon	EMAGE:3879	hwu:level_detected	future brain	EMAGE:984	hwu:level_not_detected

```
#1 Declare URI namespace
prefix hwu: <http://www.cubist_project.eu/HWU#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

#2 Select variables whose bindings are returned as solutions of the query
SELECT DISTINCT ?gene_label  ?t_label ?t_Experiment_label
?gene_strength    ?t2_label ?t2_Experiment_label ?gene_strength2
where { {

#3 Select a set of triple pattern that depicts the investigated RDF data: Set 'A'
?x rdf:type hwu:Textual_Annotation ; hwu:belongs_to_experiment
?y ;  hwu:in_tissue ?z ; hwu:has_involved_gene ?g ;
hwu:has_strength ?gene_strength .
?z hwu:has_theiler_stage hwu:theiler_stage_15  ; rdfs:label
?t_label .
?y rdfs:label ?t_Experiment_label .
?g rdfs:label ?gene_label
}
OPTIONAL #4 SPARQL key word which enables optional match
{
#5 Select optional variables contradicting set 'A' in another set: set 'B'
?b rdf:type hwu:Textual_Annotation ; hwu:belongs_to_experiment
?y2 ; hwu:in_tissue ?ztissue2 ; hwu:has_involved_gene ?g ;
hwu:has_strength ?gene_strength2  .
?ztissue2 rdfs:label ?t2_label .
?y2 rdfs:label ?t2_Experiment_label .

#6 Stipulate the relationship between set 'A' and set 'B'
?z hwu:is_part_of ?ztissue2 .

#7 Stipulate the necessary condition that can ascertain any
#7 possible contradictory values between set 'A' and set 'B'
Filter(?gene_strength  = hwu:level_detected && ?gene_strength2
= hwu:level_not_detected ) } }

#8 Aggregate values of variables to be returned
group by ?gene_label  ?t_label ?t_Experiment_label
?gene_strength    ?t2_label ?t2_Experiment_label ?gene_strength2

#9 Restrict expected results to allow only the output of contradictory values
having ((round((count(?t2_label))/(count(?t_label))*100)) > 0)

#10 Establish the order for the result set
order by ?gene_label
```

Figure 3: Query to identify binary contradictory RDF data in Theiler Stage 15

The result set in table 1 above, shows identified binary inconsistent RDF data in Theiler stage 15. As an example, some tissues (*future midbrain* and *future rhombencephalon* of experiments EMAGE:3530 and EMAGE:3879 respectively) with involved gene "Pax2" whose expression level are specified as "level_detected" were identified. *Future midbrain* and *Future rhombencephalon* have the same involved gene "Pax2" and a "is_part_of" relationship with the tissue "Future brain" whose expression level is specified as "level_not_detected" in EMAGE:984. These expression levels of *Pax2* as specified in these experiments contradict each other and do not abide with the semantics of the word "is_part_of" as utilized by EMAP. In addition, analogue inconsistency was detected in the investigated dataset in some tissues which have "is_part_of" relationship with other tissues. The identified analogue inconsistent data involve a gene with varied strength levels such as "strong" and "moderate" in tissues that have "is_part_of" relationship with other tissues. Analogue inconsistency in RDF data from Theiler stages 15 of the investigated dataset was identified by substituting the filter condition under label #7 of figure 3 with the below filter condition:

```
Filter(?gene_strength = hwu:level_strong && ?gene_strength2 = hwu:level_weak || ?gene_strength = hwu:level_moderate && ?gene_strength2 = hwu:level_weak || ?gene_strength = hwu:level_strong && ?gene_strength2 = hwu:level_moderate)
```

Table 2: Analogue inconsistent tissue experiments of Theiler stage 15

Gene_label	T_label	T_Experiment_label	Gene_strength	T2_label	T2_Experiment_label	Gene_strength2
Fkbp3	branchial arch	EMAGE:5349	hwu:level_strong	embryo	EMAGE:5349	hwu:level_weak
Fkbp3	limb	EMAGE:5349	hwu:level_strong	embryo	EMAGE:5349	hwu:level_weak
Msx1	2nd branchial arch mesenchyme	EMAGE:5411	hwu:level_moderate	2nd branchial arch	EMAGE:3839	hwu:level_weak
Nav2	epithelium	EMAGE:6026	hwu:level_strong	otocyst	EMAGE:6026	hwu:level_weak
Pax1	3rd branchial pouch endoderm	EMAGE:61	hwu:level_moderate	3rd branchial pouch	EMAGE:246	hwu:level_weak
Pax1	3rd branchial pouch endoderm	EMAGE:61	hwu:level_moderate	3rd branchial pouch	EMAGE:3938	hwu:level_weak
Smarcb1	branchial arch	EMAGE:5062	hwu:level_strong	embryo	EMAGE:5062	hwu:level_weak
Smarcb1	limb	EMAGE:5062	hwu:level_strong	embryo	EMAGE:5062	hwu:level_weak

The result set in table 2, shows the identified analogue inconsistent RDF data in Theiler stage 15. As an example from the table, some tissues (*Branchial arch* and *limb* of experiment EMAGE:5349) with involved gene "Fkbp3" whose expression levels are specified as "level_strong" have been identified from the investigated dataset. *Branchial arch* and *Limb* have "is_part_of" relationship with the tissue *Embryo*. Yet, *Fkbp3* has a level of expression "level_weak" in *Embryo* in the same experiment. These expression levels of "Fkbp3" as specified in the experiment contradict each other and do not abide with the semantics of the word "is_part_of" as utilized by EMAP. Examples from other EMAGE inconsistency types include the inconsistency from positive propagation: Gene "Pax2" was "detected" in *Future midbrain* in EMAGE:3879 and "not detected" in *Future brain* in EMAGE:984 (Table 1). *Future midbrain* is part of *future brain* and it is located at a lower part to *future brain* in the

anatomy structure of Theiler stage 15 (figure 2). *Future brain* should unavoidably have the same gene expression as *future midbrain* if gene expression is to be propagated up the anatomy. The strength level of *future brain* is therefore contradicted by not fully propagating *Future midbrain's* gene expression level up the anatomy and this result to 'an inconsistency of positive propagation'. On the other hand, *Future midbrain* should unavoidably have the same gene expression level as *future brain* if gene expression is to be propagated down the anatomy. The strength level of *future midbrain* was contradicted by not fully propagating the gene expression level in *future brain* down the anatomy and this result to 'an inconsistency of negative propagation'. Figure 2 shows the tree illustrating the hierarchical structure of *future midbrain* and *future brain* in Theiler stage 15.

4 Evaluating and analyzing inconsistent RDF data

There are two main methods of dealing with inconsistent data in a dataset: to diagnose and repair it, and reasoning with the inconsistency [3]. Also, various approaches such as [7, 8] have been proposed on reasoning with the inconsistent data. The act of addressing inconsistent data through identifying the inconsistency with the aim of repairing it through deleting the inconsistent data will inevitably increase the incompleteness of the dataset. More so, the use of various reasoning approaches on inconsistent dataset would produce varied result sets for a given approach on the dataset. These lapses can be addressed through measuring and detailing of the inconsistencies in the retrieved information from an inconsistent dataset.

Obviously, measuring inconsistency has been proven useful in analyzing diverse range of information types such as news reports [4]. However, there are a few approaches [1, 2] for measuring the inconsistencies of semantic datasets. There are other publications which verify and validate the RDF data held within a database [5, 6] but these works do not measure and analyze the amount of inconsistency in inconsistent information retrieved from the database. Consequently, the author assesses the amount of inconsistency in inconsistent information from a graph based approach. He achieves this through adopting the sub graph matching powers of SPARQL queries.

5 Approach

The amount of inconsistency in an investigated RDF data can be measured by evaluating the amount of contradiction in the RDF data against the likelihood of the contradiction to occur. This amount is assessed herein by calculating their ratio as a fraction of 100. The result educates us on how large/small the embedded contradiction in the RDF data is. As stated above, the amount of inconsistency in EMAGE's data from a graph based approach is herein assessed through adopting the mathematical and sub graph matching powers of SPARQL queries. This approach can be applied to all RDF dataset formats. It necessitates proper SPARQL query skills and adequate knowledge of the dataset by the dataset analyst. The amount of contradictions in the data under investigation against its total possibility to occur in the dataset is calculated as follows:

Xm = A RDF graph pattern in a RDF dataset
Xk = Contradictory sub graph of Xm

The interest is in calculating the amount of ***Xk*** in ***Xm*** such that

$\sum Xk$ = Total number of contradictions in ***Xk***
$\sum Xm$ = Total number of occurrence of ***Xm*** in the dataset

$$\text{Amount of Inconsistency in Xm} = \frac{\sum Xk}{\sum Xm} * \frac{100}{1}$$

In this investigation, the question "what amount of binary or analogue contradiction is present in the expression levels of the genes in each tissue experiment of Theiler stage 15" is answered. The amount of Binary or analogue inconsistency in RDF data from any of the Theiler stages of the investigated dataset is identified by adding the following SPARQL statement before label #2 of figure 3.

```
Select ?gene_label ?t_Experiment_label
round((count(?gene_strength2))/(count(?gene_strength))) * 100)
as ?amount_of_inconsistency)
{
```

And also substituting the aggregation statement under the label #8 of the query with the below statement:

```
Group by ?gene_label ?t_Experiment_label
```

The result set of the administered query on Theiler stage 15 is as displayed in table 3 and 4 below.

Table 3: Amount of binary inconsistency in tissue experiments of Theiler stage 15

Gene_label	T_Experiment_label	Amount_of_inconsistency (%)
Pax2	EMAGE:3530	50
Pax2	EMAGE:3879	33

Table 3 above, gives a more clarifying result set of each inconsistent experiment in Theiler stage 15 of the dataset than table 1. Rather than listing inconsistent experiments singly (like in table 1), the amount of its occurrence in the RDF data with the stipulated pattern is measured. These measures inform us of the amount of inconsistent assays in each tissue experiment of a particular Theiler stage in the dataset. As an example in EMAGE:3530, it can reliably be stated that half (50%) of the assays are binary inconsistent. While in EMAGE:3879, less than half (33%) of the assays results

are binary inconsistent. Consequently, decisions by Biologists to carry out further test or to remove existing experimental results from the dataset can be made.

Table 4: Amount of analogue inconsistency in tissue experiments of Theiler stage 15

Gene_label	T_Experiment_label	Amount_of_inconsistency (%)
Fkbp3	EMAGE:5349	20
Msx1	EMAGE:5411	8
Nav2	EMAGE:6026	2
Pax1	EMAGE:61	22
Smarcb1	EMAGE:5062	22

Figure 4 below, depicts a flowchart for measuring inconsistency of RDF dataset.

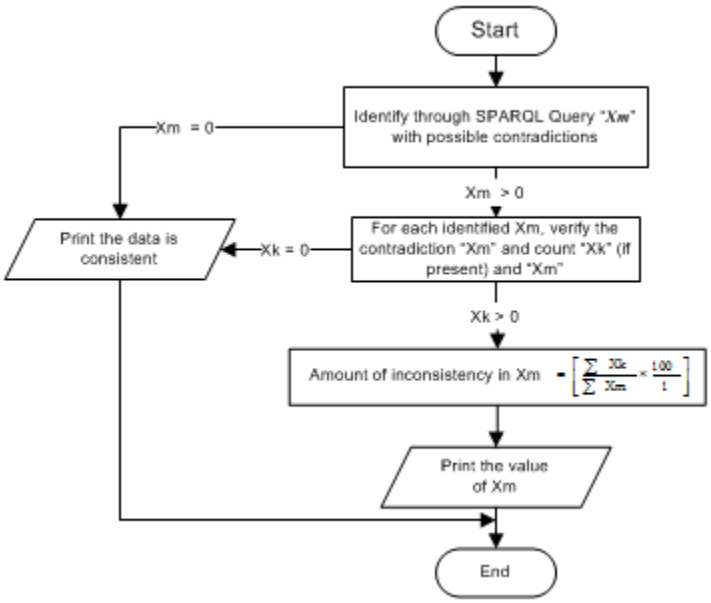

Figure 4: Flowchart for measuring inconsistency in RDF data

A tissue experiment can have several assays. The author's approach identifies the amount of these inconsistent assay(s) in their corresponding tissue experiment. For example, in Theiler stage 15 of the investigated dataset, there are 6 assays on EMAGE:3879, and 2 of them are binary inconsistent thus the amount of inconsistency in the experiment is calculated by dividing 2 with 6 and multiplied the result by 100. The importance of identifying the amount of inconsistency in a tissue experiment is to identify how valid the assay results of a particular experiment are.

6 Conclusion

Evaluating and analyzing inconsistent RDF data of a RDF model dataset is a field yet to be explored. Interestingly, it has been shown in this paper that the measure and analysis of inconsistent RDF data gives an insight to the soundness of the information under investigation. Nevertheless, the author hopes to improve on this research by automating these processes of identifying, evaluating and analyzing inconsistent RDF data.

The author acknowledges the partners of CUBIST project especially Heriot-Watt University and Sheffield Hallam University for their support and provision of his research datasets. He also acknowledges his two PhD supervisors “Simon Andrews” and “Simon Polovina" for their invaluable contributions and review of this work.

Reference

1. Grant, J., and Hunter, A. (2006). Measuring inconsistency in knowledgebases. Journal of Intelligent Information Systems, 27(2), 159-184.
2. Grant, J., and Hunter, A. (2008). Analysing inconsistent first-order knowledgebases. Artificial Intelligence, 172(8), 1064-1093.
3. Huang, Z., van Harmelen, F., and ten Teije, A. (2006). Reasoning with inconsistent ontologies: Framework, prototype, and experiment. Semantic Web Technologies: Trends and Research in Ontology-Based Systems, 71-93.
4. Hunter, A. (2002, July). Measuring inconsistency in knowledge via quasi-classical models. In PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (pp. 68-73). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
5. Jerven, B., Sebastien, G., and the UniProt Consortium. Catching inconsistencies with the semantic web: a biocuration case study
6. Jupp, S., Parkinson, H., and Malone, J. Semantic Web Atlas: Putting Gene Expression Data Into Biological Context.
7. Lembo, D., Lenzerini, M., Rosati, R., Ruzzi, M., and Savo, D. (2010). Inconsistency-tolerant semantics for description logics. Web Reasoning and Rule Systems, 103-117.
8. McLeod, K., and Burger, A. (2007). Using argumentation to tackle inconsistency and incompleteness in online distributed life science resources. In Proceedings of IADIS International Conference Applied Computing (pp. 489-492).
9. McLeod, K., and Burger, A. (2011). WP7 requirement document of CUBIST Consortium 2010-2013. Available at http://www.cubist-project.eu/fileadmin/CUBIST/user_upload/Deliverable/CUBIST_D7.1.1_HWU_v1.0.pdf
10. Richardson L, Venkataraman S, Stevenson P, Yang Y, Burton N, Rao J, Fisher M. Baldock RA, Davidson DR, Christiansen JH. EMAGE mouse embryo spatial gene expression database: 2010
11. Suda, Y., Hossain, Z. M., Kobayashi, C., Hatano, O., Yoshida, M., Matsuo, I., and Aizawa, S. (2001). Emx2 directs the development of diencephalon in cooperation with Otx2. Development, 128(13), 2433-2450.
12. Theiler, K. (1989). The house mouse: atlas of embryonic development (p. 168). New York: Springer-Verlag.