# Marking Time in Developmental Biology: Annotating Developmental Events and their Links with Molecular Events

**Gail Sinclair**
School of Informatics
University of Edinburgh
Edinburgh EH8 9LW
c.g.sinclair@ed.ac.uk

**Bonnie Webber**
School of Informatics
University of Edinburgh
Edinburgh EH8 9LW
bonnie@inf.ed.ac.uk

**Duncan Davidson**
MRC Human Genetics Unit
Western General Hospital
Edinburgh EH4 2XU
Duncan.Davidson@hgu.mrc.ac.uk

## Abstract

Current research in developmental biology aims to link developmental genetic pathways with the processes going on at cellular and tissue level. Normal processes will only take place under specific sequential conditions at the level of the pathways. Disrupting or altering pathways may mean disrupted or altered development.

This paper is part of a larger work exploring methods of detecting and extracting information on developmental events from free text and on their relations in space and time.

## 1 Introduction

Most relation extraction work to date on biomedical articles has focused on genetic and protein interactions, e.g. the extraction of the fact that expression of Gene A has an effect on the expression of Gene B. However where genetic interactions are tissue- or stage-specific, the conditions that govern the types of interactions often depend on *where* in the body the interaction is happening (space) and *at what stage* of life/development (time).

For genetic pathways involved in development, it is critical to link what is happening at the molecular level to changes in the developing tissues, usually described in terms of processes such as *tubulogenesis* and *epithelialization* (both involved in the development of the kidney) and where they are happening.

The processes themselves are usually linked to *stages* rather than precise time points and spans like "6.15pm EST", "March 3", "last year". Within the developmental mouse community, there are at least two different ways of specifying the developmental stage of an embryo - Theiler stages (TS), and days post coitum/embryonic day (d.p.c./E). However, these cannot be simply mapped to one another as can days, weeks and years. Embryonic days are real time stages independent of the state of the embryo and dated from an assumption about when (approximately) the relevant coitus must have taken place, while Theiler stages are relative stages dependent on the processes an embryo is undergoing.

Developmental stages can be also be referred to implicitly, by the state of the embryo or the processes currently taking place within it. This is because during development, tissues form, change, merge or even disappear. So if the embryo is undergoing *tubulogenesis*, one can assume that its developmental stage is (loosely) somewhere between TS20 and birth. If the text refers to *induced mesenchyme* during a description of tubulogenesis, one can assume that this change in the mesenchyme is the (normal) consequence of the *Wolffian duct* invading the *metanephric mesenchyme*. The invasion is known to occur around 10.5 d.p.c so the *induced mesenchyme* must come into existence soon after this time.

Temporal links between developmental events may be indicated explicitly (e.g. *first, a tubule develops into a comma-shaped body, which then develops into an S-shaped body*), but they are more likely to be indicated implicitly by their ordering in the text and by associative (or "bridging") anaphora where the anaphor refers to the result of a previously mentioned process, e.g. the *induction of metanephric mesenchyme* as one event, and a subsequent mention of *induced mesenchyme* (an "associative" or "bridging" reference) within another event, suggesting the former event occurred before the latter.



*Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 46–53,
Sydney, July 2006. 

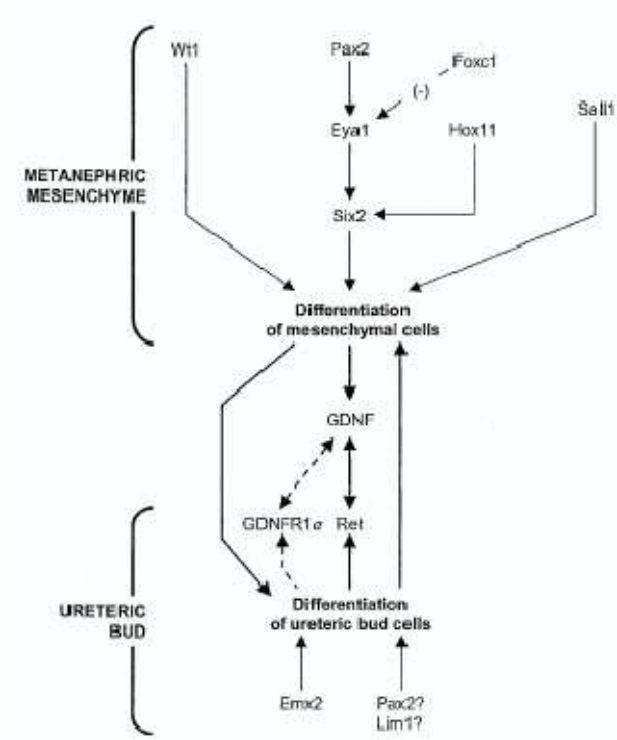


Figure 1: Partial genetic pathway for early kidney morphogenesis. The arrows show directed interactions between genes that are required for the specified processes. E.g Pax2 interacts with (*activates*) Six2 which, together with Sall1 and Wt1, is required for differentiation of the mesenchymal cells in the metanephric mesenchyme. Image taken from (Ribes et al., 2003).

This work on linking molecular and developmental events mentioned in text on development is also meant to deal with the problem that no one article ever fully describes a topic. The partial genetic interaction network in Figure 1 has been built from several different studies and not determined from just a single experiment. So not only does the information within one article need to be mined for useful information - the information across articles needs to be associated with each other with respect to temporal, spatial and experimental grounding. Eccles et al. (2002) states that Pax2 is required for differentiation of mesenchymal cells during kidney morphogenesis, while Sajithlal et al. (2005) states that Eya1 is required. However these two results by themselves do not help us determine whether these requirements are independent of one another or whether they are required at different stages or in different parts of metanephric mesenchyme or whether the two genes interact. The conditions involved in the experiments, most importantly the temporal conditions, can help to link the two events.

This work aims to develop methods for extracting information from text that will ground genetic pathways (molecular events) with regard to tissue location, developmental process and stage of embryonic development - that is, their **spatio-temporal context**. The task at hand is to recognise how biologists write about developmental events and then adapt existing or formulate new natural language processing techniques to extract these events and temporally relate them to each other. The resultant information can then be used both for database curation purposes and for visualisation, i.e. to enrich pathway diagrams such as Figure 1, with information such as when and where the interactions take place, what type of interactions are involved (physical, activation, inhibition), the origin of this information and other associated information.

## 2 Notions of Time

As previously mentioned, there are different ways of calibrating for developmental stages, and they cannot simply be mapped to one another. The two most common stage notations for mouse development are Theiler stages, TS, and Embryonic days, E (equivalent to days post coitum, d.p.c.). The latter are self explanatory in that they denote the 24 hour day and can be considered *real-time* staging.

The convention was originally that *E11* would represent the 24 hour period of the 11th day. It is, however, now common to find *E11.5* representing the same time period, but this is merely a change in convention due to standard practices of experimentation.

A Theiler stage on the other hand represents a non-fixed *relative* time period defined by the progress of development rather than directly in terms of the passage of time. Theiler Stages (Theiler, 1989) divide mouse development into 26 prenatal and 2 postnatal stages. In general, Theiler used external features that can be directly assessed by visual inspection of the live embryo as developmental landmarks to define stages. The Edinburgh Mouse Atlas Project (EMAP)[1] uses Theiler stages to organise anatomical terms in their Mouse Atlas Nomenclature (MAN). EMAP gives a brief description of each Theiler stage with TS25 as an example as follows:

> Skin wrinkled
> The skin has thickened and formed wrinkles and the subcutaneous veins are less visible. The fingers and toes have become parallel and the umbilical hernia has disappeared. The eyelids have fused. Whiskers are just visible.
> Absent: ear extending over auditory meatus, long whiskers.

An embryo is in TS25 at approximately 17 d.p.c. As can be seen in Figure 2, an embryo at E11 could be considered in Theiler stage 17, 18 or 19, i.e. Theiler stages can overlap one another with respect to Embryonic day. Indeed, here, TS17 can fully encompass TS18 in the dpc timeline.

The development of internal structures is approximately correlated with external developments, so except for fine temporal differences, the Theiler stages can be assumed to apply to the whole embryo. Theiler stages provide only gross temporal resolution of developmental events, and the development of internal structures often take place within the boundaries of one of these stages or overlapping stage boundaries. Thus, internal developmental processes can also have their own finer *relative* timeline or staging.

There is no ontology or reference book that comprehensively specifies this finer staging and the knowledge of the biologist as the reader of articles is relied upon. This work will contribute to making this deeper staging criteria explicit.

[1] Edinburgh Mouse Atlas Project - http://genex.hgu.mrc.ac.uk/

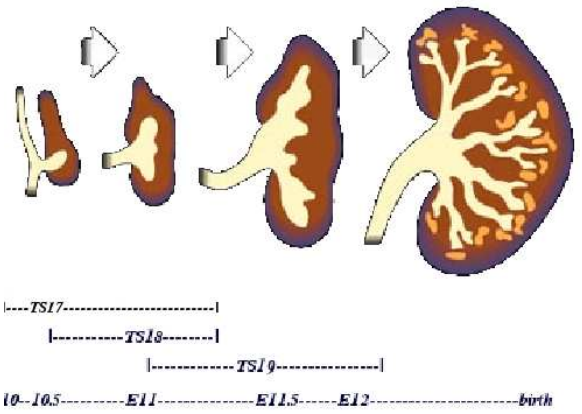


Figure 2: Graphic of kidney morphogenesis annotated with the two standard staging notations for mouse development. At E10.5 the Wolffian duct invades the metanephric mesenchyme forming the ureteric bud around E11. The bud then branches around E11.5 and continues to do so until birth, forming the ultimate functional units of the kidney - the nephrons. TS = Theiler Stage, E = Embryonic day/dpc. This image is adapted from http://www.sciencemuseum.org.uk/

## 3 Annotation

### 3.1 Event Classification

As a first step, a Gold Standard corpus of 988 sentences was developed with each sentence being classified as containing the description of a developmental and/or molecular event or not. 385 sentences were classified as positive, with 603 negative. Named entities within all these sentences were also annotated. Among these element types were *stage, process and tissue*. A Naive Bayes automatic classifier for sentence classification was developed using this Gold Standard resulting in a balanced F-score of 72.3% for event classification. (A manual rule-based approach resulted in an F-score of 86.6%, but this has yet to be fully investigated for automation. Guessing positive for all sentences would give a balanced F-score of 58.4%)

### 3.2 Event Specifications

Two event types are of interest in this work - *molecular* and *tissue* events. The former involve the action (and possible effect) of molecules dur-

ing development and the latter involves the development of the tissues themselves. A description of an event can be expected to contain the following elements:

- *molecular* or *tissue* event type (e.g. *expression, inhibition*)
- stage or temporal expression (e.g. *after* X, *subsequent to* X, *E11*)
- at least one of
  - molecule name, anatomical term, biological process term

The informational elements included within an event description can then be used to relate events to each other. Specifically, processes involve known tissues and are known to happen during certain stages, just as the relative order of processes, tissue formations and stages are known.

While an initial specification of an event may be associated with a single sentence, clause or phrase, not all the elements of relevance to this work may be specified there. In particular, an informational element of the event may be explicitly and fully stated in this initial event specification, or it may be underspecified or it may be missing. For those that are underspecified or missing, background knowledge about other elements and events may need to be taken into consideration in order for them to be fully resolved (see Section 4.2).

The following is a straightforward example where the given sentence specifies all the main elements required for a molecular event.

1. *At E11, the integrin α8 subunit was expressed throughout the mesenchyme of the nephrogenic cord.*
   - Molecular Event : *expression*
   - molecule name: *integrin α8*
   - anatomical term: *mesenchyme of the nephrogenic cord*
   - stage: *E11*

Example 2 shows that a single sentence may specify more than one event.

2. *Prior to formation of the ureteric bud, no α8 expression was evident within the mesenchyme that separates the urogenital ridge from the metanephric mesenchyme and within the metanephric mesenchyme itself.*
   - EVENT-0
     - Tissue Event : *formation of* anatomical term
     - anatomical term: *ureteric bud*
     - stage/temporal expression = missing
   - EVENT-1
     - Molecular Event: absence of expression
     - molecule name: *α8*
     - anatomical term: *mesenchyme that separates the urogenital ridge from the metanephric mesenchyme*
     - temporal expression:*Prior to* EVENT-0
   - EVENT-2
     - Molecular Event: absence of expression
     - molecule name: *α8*
     - anatomical term: *metanephric mesenchyme*
     - temporal expression: *Prior to* EVENT-0

EVENT-0 is not the focus of this sentence, but rather a *reference* event. Its attributes need to be recorded so that the stage of the other events can be determined.

TimeML (Pustejovsky et al., 2004) is a specification language designed for the annotation of temporal and event information. Although TimeML is not currently being used as a method of representation for this work, Example 1 above could be represented as follows:

**<SIGNAL sid="s1" type="temporal">**
***At***
**</SIGNAL>**
**<TIMEX tid="t1" type="STAGE" value="E11">**
***E11***
**</TIMEX>**
***the integrin α8***
**<EVENT eid="e1" class="molecular">**
***was expressed***
**</EVENT>**
***throughout the mesenchyme of the***
**<SIGNAL sid="s2" type="tissue">**
***nephrogenic cord***
**</SIGNAL>**

*nephrogenic cord* can be considered a signal of type "tissue" as it does not exist throughout the

whole of development and so can indicate or rule out time periods for this event description.

### 3.3 Event Time-Stamping

The relative timing of any biological processes mentioned in the event descriptions first needs to be determined before we can work out when the actual events described are taking place.

Schilder and Habel (2001) looked beyond the core temporal expressions and into prepositional phrases that contained temporal relations, i.e. *before, during*, etc and introduced the notion of noun phrases as event-denoting expressions. An event that is described as occurring "after the election" does not have an explicit time-stamp attached to it, but the knowledge about the timing of the election mentioned gives the reader a notion of when in absolute time the event occurred. This is similar to Example 2 above where Event-0 is the reference event, thus biological processes can be considered event-denoting expressions.

While Schilder and Habel rely on prepositional phrases to designate their event-denoting noun phrases, for this work propositional phrases are not necessarily required. The mention of a noun phrase by itelf may be enough. In developmental biology, tissues may only be extant for a limited period before they form into some other tissue and these can also be used as event-denoting expressions - for example, *comma-shaped bodies* are structures within the developing kidney that are only in existence for a relatively short time period - before the existence of the *S-shaped bodies* and after *epithelialization*. Therefore the mention of tissues as well as processes can help to pinpoint the timing of the event being described. While they may not ultimately bring us to the exact stage the event is occurring in, it can at least rule out some spans of time. We discuss this further in Section 4.2.

In order for events to be linked to one another, it is necessary to uniquely index each event and its elements. Mapping across indices will be utilised so that known relationships between elements can be represented. For example, E10 comes **before** E12, *tubulogenesis* occurs **during** *kidney morphogenesis*, and the *proximal tubule* is **part of** the *nephron*.

Of the elements types listed in Section 3.2, only the *molecule* element cannot be used to resolve developmental stage while *tissue, process, stage* and, of course, *temporal expression* can. Other elements are also of interest to the biologist and integral to development and molecular function, however they are not of use in the grounding of events in time.

## 4 Initial Investigations

This section demonstrates that one must look beyond the sentence in order to resolve the temporal aspects of events.

### 4.1 Evidence for Developmental Stage

Evidence sufficient to resolve developmental stage can come from many places. 314 positive sentences from the Gold Standard corpus and their context were examined, and the evidence required to resolve developmental stage for each of the events mentioned there was determined as shown in Table 1.

As can be seen from the table, only 48 out of the 314 event sentences (i.e. 15%) have the developmental stage in which the event is occurring *explicitly stated* in the given sentence, (e.g. Example 1 in Section 3). So other means need to be explored in order to ground events with respect to developmental stage. An event sentence may be a continuation of a topic, and so the specific developmental stage involved may well be stated in the immediately surrounding or related text.

Information in the immediately surrounding text (rows labelled *Following Sentence, Previous Sentence* and *Current Paragraph*) resolves the developmental stage of the event in 64 cases (i.e. 21%). This most commonly occurs by looking for the immediately previously mentioned stage, and in one case the next encountered stage.

Event sentences also often refer to figures, and so the stage being described in the caption (i.e. legend) of the referenced figure will often be the same as the one relevant to the sentence. (This was true of all sentences looked at that referenced a figure.) Figures, however, are generally only found in the *Results* sections and so this type of evidence is not often going to be of use for sentences found in other sections of an article.

Similarly, events can be described within the figure legends themselves. The concise and simple way in which legends are generally written mean that the explicit stage is commonly referred to, and so stage can be resolved using this referenced information(43 out of 47 cases, i.e. 91%).

| Source of Evidence | Abstract | Introduction | Results | Discussion | Methods | Totals |
|---|---|---|---|---|---|---|
| Time Irrelevant | 7 | 12 | 22 | 23 | 1 | 65 |
| Prior Knowledge | 17 | 33 | 31 | 45 | 0 | 126 |
| Following Sentence | 0 | 0 | 1 | 0 | 0 | 1 |
| Previous Sentence | 0 | 0 | 7 | 0 | 0 | 7 |
| Current Paragraph | 0 | 0 | 18 | 1 | 0 | 19 |
| Reference to Figure | 0 | 0 | 38 | 0 | 0 | 38 |
| Within Fig Legend | 0 | 0 | 43 | 0 | 0 | 43 |
| (time not resolved) | 0 | 3 | 1 | 0 | 0 | 4 |
| Explicitly Stated | 0 | 1 | 41 | 5 | 1 | 48 |
| (not relevant) | 0 | 0 | 1 | 0 | 0 | 1 |
| **Totals** | 24 | 49 | 165 | 74 | 2 | 314 |

Table 1: Location and type of evidence sufficient to resolve developmental stage in sentences. *Time Irrelevant* indicates that the event being described is not time critical, i.e. event is a constant over developmental timeline, or end result. *Prior knowledge* means temporal information other than that found in the current paragraph but associated with current event such as *tissue* and *process* is required for temporal resolution. This may be found in the current article or from previously curated information (assuming accurate terminology mapping.) Text from outside the current paragraph cannot be relied upon to be relevant to the current sentence without additional information. *time not resolved* means the stage could not be pinpointed using the figure legend. *not relevant* indicates that although an explicit stage was referred to within the sentence, this was not relevant to the event being described, e.g. event and stage in different clauses of the sentence.

Table 2 shows a similar table to Table 1, but deals only with those sentences found within figure legends. It shows where within the figure legend the required evidence for developmental stage can be found. As can be seen, in 80% of these cases the relevant developmental stage can be ascertained directly from the legend. It should be noted that figure legends in biological articles tend to be much lengthier than those from NLP articles.

In 21% of the event sentences, a specific developmental stage is not relevant to the fact being described (first row of Table 1), e.g. *the kidneys of the double mutants were located more caudal and medial than normal*. This sentence is describing an end result, i.e. an affected or normal kidney at birth (although this could, of course, be considered a developmental stage.) Alternatively, the time-irrelevant event being described could be a non-event, e.g. the fact that a gene is *never* expressed in a particular tissue. Similarly, this could be considered as the developmental stage range from conception to birth.

The significantly small proportion of event sentences located in Abstracts (24 of 314 total event sentences, less than 8%) demonstrates the need to use full text. Even where an event is described within an Abstract, it is rarely accompanied by associated processes or tissues specific enough to suggest the stage of development never mind an explicit timestamp, as it is, by necessity, only generally describing the whole article. The majority of BioNLP work is being done with the use of Abstracts only. This is because of their relative ease of access compared with full text, but methods developed using Abstracts only will not necessarily be as effective when applied to full text.

As can be seen, the majority of temporally-underspecified event sentences are situated in the *Results* section of the articles. Indeed, this is the section where most event sentences are to be found. This work is initially focussing on event descriptions found in Results sections of articles as these will focus on the work done by the authors and their findings and will not generally include modality in the event descriptions as Introduction and Discussion sections might. As shown above, the Methods section rarely contains event descriptions and when they do they are usually about what the experiment aims to show and so this should be repeated in the Results section.

### 4.2 Prior Knowledge

As mentioned earlier, if none of the above sources reveal the relevant stage of an event, then other elements within the sentence, such as *tissue* or *process*, need to be looked at so that prior knowledge

| Source of Evidence | Figure Legends |
|---|---|
| Time Irrelevant | 4 |
| Prior Knowledge | 4 |
| Following Sentence | 0 |
| Previous Sentence | 14 |
| Current Paragraph | 13 |
| Explicitly Stated | 11 |
| **Total** | 47 |

Table 2: Location and type of evidence sufficient to resolve developmental stage in sentences within figure legends. Rows as in Table 1, with *Current Paragraph* being equal to the whole of the legend.

about those elements can be exploited for developmental stage to be resolved. For example, given the sentence

> *Prior to formation of the ureteric bud, no α8 expression was evident within the mesenchyme that separates the urogenital ridge from the metanephric mesenchyme and within the metanephric mesenchyme itself.*

the developmental stage can be resolved if we know when the ureteric bud forms (TS17/E10.5). It could also be the case that the other tissues or processes mentioned have a specific lifetime within development and these could help to further pinpoint the timeline involved for the lack of *α8* expression. For example,

> *Pax2 was initiating in the metanephric mesenchyme undergoing induction.*

It is not so straightforward to assign a stage here, since the mesenchyme is constantly being induced from E11 (TS18) until birth (TS26), but we have at least discounted E1-E10 (TS1-TS17) as relevant stages.

Resources such as the Mouse Atlas Nomenclature (MAN) (Ringwald et al., 1994) will provide the initial prior knowledge in order to resolve developmental stage of events. This describes the different stages of development and the tissues in evidence at each stage, giving what is known as the *abstract mouse*. From this abstract mouse, we can ascertain the normal stage ranges where tissues exist and use this knowledge for temporal resolution, taking care not to assume that tissues do not necessarily exist within the same stage range in mutant mice than in wild-type. The prior knowledge databank can be recursively added to with facts from events already extracted from papers for use in further event extraction and their anchoring in time.

## 5 Future Work

### 5.1 Term Normalisation

There is no point extracting events descriptions if we cannot relate the events and their elements to each other. The event-denoting expressions identified need to be normalised so that it can be recognised when two terms are referring to the same element.

Inconsistent terminology in the biomedical field is a known problem (Sinclair et al., 2002). One gene can have several names (synonymy) just as the same name can be used for more than one gene (homonymy). Very often the synonyms bear no relation to one another since they were perhaps concurrently discovered in different laboratories and named. For example, the gene *insomnia* can also be known as *cheap date*, since experiments found that organisms without this gene have a tendency to fall asleep and are particularly susceptible to alcohol. The same anatomical part can also be referred to by different terms, e.g. the *Wolffian duct* is also known as the *nephric duct*, and the *metanephros* is another name for the *kidney*. There is also a lineage issue, where a tissue with one name (or perhaps more) develops into something with another name (e.g. the *intermediate mesoderm* gives rise to both the *Wolffian duct* and the *metanephric mesenchyme* which in turn both develop into the *metanephros*. The MAN includes this type of information.

Term normalisation is particularly important for the *process* and *tissue* elements. If these terms are not normalised, temporal knowledge about the terms may not be exploited and it may not be determined that events involving them are linked.

### 5.2 Event Elements

If the elements required to fully describe an event are explicitly stated within a simple sentence, then temporal grounding will be straightforward. However, this is unlikely to often be the case. More complex sentences will dictate the need for dependency relations to be determined so that each event's elements can be identified. Methods for dealing with missing or underspecified elements that are not resolved within the event description itself will be investigated.

A naive approach will first be investigated to fill these gaps: find the closest appropriate element in the previous context (varying the size of the window for how far back to look, such as current paragraph or last 3 sentences). An error analysis on this simple method will help to guide the amount of further work necessary to achieve equal success across all elements. For those elements that this method is ineffective, other methods will be developed incorporating features such as sensitivity to syntax, event type and location within article. Similarly, it will be established whether different techniques are required for missing information than for underspecified information. They will first be treated in the same manner with analysis determining whether they should be treated differently.

## 6 Conclusion

This ongoing work has shown the importance of relative time lines in order to link events to one another. The identification of event elements and their normalisation will then form a basis for reasoning over these elements with regards to first time-stamping of events and then temporally relating the events. The aim of many BioNLP studies is ultimately to reason over extracted events and, as such, the relative timing of these events is crucial. For example, if we know
1. tissue X is transformed into tissue Y at stage S and
2. molecule M is expressed in X at stage S-1,
then it can be reasoned that event 2 has an impact on event 1. This reasoning can be made more successful if we know as much about the events as possible, not just that tissue Y is formed and molecule M is expressed.

It has also been demonstrated that we not only need to look beyond the sentence level for temporal resolution but also beyond the article in order to replicate the reader's assumed level of background knowledge.

## References


J. F. Allen, Towards a general theory of action and time, *Artificial Intelligence*, vol 23, pp 123-154, 1984.

M. R. Eccles, S. He, M. Legge, R. Kumar, J. Fox, C. Zhou, M. French and R. W. Tsai, .PAX genes in development and disease: the role of PAX2 in urogenital tract development. *Int J Dev Biol*, vol 46, no 4, pp 535-44, 2002.

J. Pustejovsky, I. Mani, L. Belanger, B. Bogurev, B. Knippen, J. Littman, A. Rumshisky, A. See, S. Symonen, J. Van Guilder, L. Van Guilder and M. Verhagen, The Specification Langeuage TimeML. in The Language of Time: A Reader, Oxford University Press, 2004.

D. Ribes, E. Fischer, A. Calmont and J. Rossert, Transcriptional Control of Epithelial Differentiation during Kidney Development. *J Am Soc Nephrol*, vol 14, pp S9-S15, 2003.

M. Ringwald, R. A. Baldock, J. Bard, M. H. Kaufman, J. T. Eppig, J. E. Richardson, J. H. Nadeau and D. Davidson, A database for mouse development. *Science*, vol 265, pp 2033-2034, 1994.

G. Sajithlal, D. Zou, D. Silvius and P. X. Xu, Eya 1 acts as a critical regulator for specifying the metanephric mesenchyme. *Dev Biol*, vol 284, no 2, pp 323-36, 2005.

F. Schilder and C. Habel, From Temporal Expressions to Temporal Information: Semantic Tagging of News Message., in *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing, Toulouse, France*, pp 88-95.

G. Sinclair, B. Webber and D. Davidson, Enhanced Natural Language Access to Anatomically Indexed Data. in *Proceedings of the ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*, Philadelphia, pp 45-52.

K. Theiler, *The House Mouse. Atlas of Embryonic Development.* Springer Verlag New York, 1989.