BMC Bioinformatics

Poster presentation

Extracting Genetic Pathways From Text and Grounding at the Spatio-Temporal Level

Gail Sinclair*, Bonnie Webber and Duncan Davidson

Address: School of Informatics, University of Edinburgh, Edinburgh EH9 3JZ, UK.

Email: Gail Sinclair* - csincla1@inf.ed.ac.uk

* Corresponding author

from BioSysBio: Bioinformatics and Systems Biology Conference
Edinburgh, UK, 14–15 July 2005

Published: 21 September 2005

BMC Bioinformatics 2005, **6**(Suppl 3):P26

The molecular biology literature is believed to contain a wealth of information that has not yet made it into any structured database. Of particular current interest is information about genetic pathways, and there are many ongoing studies into automatically extracting such pathways from the molecular biology literature. In the area of development, however, pathways must be linked to changes in the developing tissues, usually described in terms of cellular processes and where they are happening. Thus our aim is to extract complementary information regarding the tissue location in which pathways are active, along with the biological process they are active in and the stage of embryonic development in which the process occurs.

Temporal information in developmental biology text has a rather different character than newswire (e.g. 5 pm, last year). With respect to murine developmental staging, there are at least two separate ways of explicitly specifying the developmental stage of the embryo – Theiler stages (TS), and days post coitum/embryonic day (d.p.c./E). These cannot be simply mapped to one another as can days, weeks and years. Stages can also be referred to implicitly, by the state of the embryo or the processes currently taking place within it, e.g. tubulogenesis = circa TS20 to birth.

To start, we have collected a corpus of articles about one aspect of kidney development, annotating instances of information about specific gene expression in tissues and about the processes involved. (Inter-annotator agreement on this gold standard is at 94%.) From deeper linguistic analysis of similar (automatically retrieved) sentences, the task at hand is in recognising how biologists write about sequential events and then adapting existing/formulating new natural language processing techniques to extract these events and relate them to each other. These techniques can then be evaluated by way of literature describing a different part of kidney development.