BioFrameNet, a FrameNet Extension to the Domain of Molecular Biology

Recently there has been a dramatic increase in the frequency and number of publications of scientific texts on molecular biology, driven by advances in rapid, high throughput genomics technology, where analyses of all the genes of an entire genome can be carried out in a matter of hours. Results of such experiments are most often reported in journal articles. To take advantage of the resulting volume of literature, researchers have built tools for the automatic processing of molecular biology texts, with a common goal being to enable extraction of facts and relationships asserted and described in them. Tasks like extraction of complex assertions from texts require capabilities in higher level language processing, in particular the ability to link syntactic and semantic elements of lexical units, phrases, and clauses. FrameNet (Ruppenhofer, et al. 2006) is a well known lexical resource aimed at providing descriptions of these sorts of linkings for general English, and can thus, if adapted for this domain, be used for building tools that process molecular biology texts.

The present work introduces BioFrameNet, an extension of FrameNet to the molecular biology domain. First, we examine the syntactic and semantic combinatorial possibilities exhibited in the language of scientific writings for the domain to get a better understanding of its grammatical properties. To illustrate, the following sentences show how different Frame Element fillers may serve the same grammatical function:

1. IRS-3 expression blocked [$_{\text{TRANSPORTED\_ENTITY}}$ glucose/IGF-1 induced IRS-2] **TRANSLOCATION** [$_{\text{ORIGIN}}$ from the cytosol] [$_{\text{DESTINATION}}$ to the plasma membrane].

2. phosphorylation/dephosphorylation of the C-terminal region of PTEN serves as an electrostatic switch that controls the [$_{\text{DESTINATION}}$ membrane] **TRANSLOCATION** [$_{\text{TRANSPORTED\_ENTITY}}$ of the protein]

These examples use *translocation*, an event noun from the `Intracellular_Transport` frame. Nouns that combine with *translocation* in a compound can be realizations of different Frame Elements, here either the TRANSPORTED_ENTITY (1) or the TRANSPORT_DESTINATION (2). At the same time, a single Frame Element can have different grammatical realizations; here TRANSPORTED_ENTITY is either a bare noun modifier of the nominal predicate (1) or a post-predicate prepositional phrase (2). As is the case with the original FrameNet on which it is based, BioFrameNet facilitates the investigation of these sorts of combinatorial possibilities more precisely than other standard lexical resources (e.g. WordNet, Fellbaum 1998).

BioFrameNet follows FrameNet's grounding in Frame Semantics (Fillmore 1982), and in doing so offers a new perspective on the language of molecular biology. In providing a Frame Semantics description/analysis and cataloging of the grammatical structures used in the scientific language of molecular biology, this work shows that such an approach produces a semantically insightful analysis of this specialized language. The Frame Semantics analysis can then be used for automated language understanding work, such as content extraction and question answering, as in Fillmore et al. (2006) and Moschitti et al. (2003), both for the language in general.