TR-NLP-UT-2006-2*

GENIA Ontology

Jin-Dong Kim[1,3,4], Tomoko Ohta[1], Yuka Teteisi[2], Jun'ichi Tsujii[1,3,4]

[1]University of Tokyo,
[2]Kogakuin University,
[3]University of Manchester,
[4] National Centre for Text Mining

Date of Release: 15 November 2006

* All technical reports from Tsujii Laboratory can be found at http://www-tsujii.is.s.u-tokyo.ac.jp/TR/.

1. Introduction

The GENIA ontology is designed to support semantic annotation to the GENIA corpus. This document describes the version 2.0 of the ontology which now is divided into two ontologies.

The GENIA term ontology is to support the GENIA term annotation. Biologically meaningful terms found in literature are to be marked and classified based on the term ontology through the term annotation. The term ontology defines biological entities, organisms and anatomical entities and most of the concepts in the ontology are mapped to categories in MeSH. Section 2 describes the GENIA term ontology.

The GENIA event ontology is to support the GENIA event annotation. Biological processes and molecular functions mentioned in literature are to be identified through the event annotation based on the event ontology. The concepts in the event ontology are mapped with categories in the Gene Ontology. Section 3 describes the GENIA event ontology.

The entire GENIA ontology is encoded in the Web Ontology Language (OWL) recommended by the World Wide Web Consortium (W3C). Currently, only hierarchy of concepts are defined using the rdfs:subclassOf relation.

2. GENIA term ontology

The GENIA term ontology 2.0 is comprised of three sub-ontologies covering chemicals, organisms, and anatomical parts of organisms. It has been evolved from the former GENIA ontology which has been distributed together with the GENIA term corpus version 3.01.

In the following sections the each sub-ontology is described in terms of the hierarchy and definition of concepts in the ontology. The concepts defined in the GENIA term ontology 2.0 have been mapped to categories in MeSH, and as such the definitions from MeSH will be frequently excerpted.

2.1. GENIA Chemicals ontology

The GENIA chemicals ontology is intended to cover any chemical substance. It roughly corresponds to the *Chemicals and Drugs Category* of MeSH excluding the *Pharmaceutical Preparations* and *Chemical Actions and Uses* sub-categories of it as the ontology is intended not to cover any verbal concepts

Compound	Corresponds to most of the sub-categories in the *Chemicals and Drugs Category* of MeSH excluding the *Pharmaceutical Preparations* and *Chemical Actions and Uses* sub-categories. The *Elements* category under *Inorganic Chemicals* category is also separated out into an individual concept, *Element*.
Organic_compound	Corresponds to the *Nucleotides* category of MeSH. "*The monomeric units from which DNA or RNA polymers are constructed. They consist of a purine or pyrimidine base, a pentose sugar, and a phosphate group. (From King & Stansfield, A Dictionary of Genetics, 4th ed)*"
Amino_acid_compound	Corresponds to the *Amino Acids, Peptides, and Proteins* category of MeSH. "*Amino acids and chains of amino acids connected by peptide linkages.*"
Amino_acid_monomer	Corresponds to the *Amino Acids* category of MeSH. "*Organic compounds that generally contain an amino (-NH2) and a carboxyl (-COOH) group. Twenty alpha-amino acids are the subunits which are polymerized to form proteins.*"
Peptide	Corresponds to the *Peptides* category of MeSH. "*Members of the class of compounds composed of AMINO ACIDS joined together by peptide bonds between adjacent amino acids into linear, branched or cyclical structures. OLIGOPEPTIDES are composed of approximately 2-12 amino acids. Polypeptides are composed of approximately 13 or more amino acids. PROTEINS are linear polypeptides that are normally synthesized on RIBOSOMES.*"
Protein	Corresponds to the *Proteins* category of MeSH. "*Linear POLYPEPTIDES that are synthesized on RIBOSOMES and may be further modified, crosslinked, cleaved, or assembled into complex proteins with several subunits. The specific sequence of AMINO ACIDS*

	determines the shape the polypeptide will take, during PROTEIN FOLDING, and the function of the protein."
Protein_molecule	A single protein molecule.
Protein_family_or_group	A group of evolutionarily or functionally related proteins
Protein_domain_or_region	Covers both protein domains and protein regions. A protein domain is an independently folded unit within a protein, often joined by a flexible segment of the polypeptide chain. A protein region is any region in a protein sequence.
Protein_complex	A group of two or more associated proteins
Protein_substructure	A secondary structure or combination of secondary structures of a protein
Nucleic_acid_compound	Corresponds to *Nucleic Acids, Nucleotides, and Nucleosides* category of MeSH. "*Complex compounds of high molecular weight occurring in living cells. These are basically of two types, ribonucleic (RNA) and deoxyribonucleic (DNA) acids, both of which consist of nucleotides (nucleoside phosphates linked together by phosphate bridges).*"
Nucleotide	Corresponds to *Nucleotides* category of MeSH excluding the *Polynucleotides* category. "*The monomeric units from which DNA or RNA polymers are constructed. They consist of a purine or pyrimidine base, a pentose sugar, and a phosphate group. (From King & Stansfield, A Dictionary of Genetics, 4th ed)*"
Polynucleotide	Corresponds to *Ploynucleotides* category of MeSH. "*This list includes those paired at least once with this heading in MEDLINE and may not reflect current rules for allowable combinations.*"
DNA	Corresponds to *DNA* category of MeSH. "*A deoxyribonucleotide polymer that is the primary genetic material of all cells. Eukaryotic and prokaryotic organisms normally contain DNA in a double-stranded state, yet several important biological processes transiently involve single-stranded regions. DNA, which consists of a polysugar-phosphate backbone possessing projections of purines (adenine and guanine) and pyrimidines (thymine and cytosine), forms a double helix that is held together by hydrogen bonds between these purines and pyrimidines (adenine to thymine and guanine to cytosine).*"
DNA_molecule	A single DNA molecule
DNA_family_or_group	A group of evolutionarily or functionally related DNA
DNA_domain_or_region	
DNA_substructure	A secondary structure or combination of secondary structures of a protein
RNA	Corresponds to *RNA* category of MeSH. "*A polynucleotide consisting essentially of chains with a repeating backbone of phosphate and ribose units to which nitrogenous bases are attached. RNA is unique among biological macromolecules in that it can encode genetic information, serve as an abundant structural component of cells, and also possesses catalytic activity. (Rieger et al., Glossary of Genetics: Classical and Molecular, 5th ed)*"
RNA_molecule	A single DNA molecule
RNA_family_or_group	A group of evolutionarily or functionally related proteins
RNA_domain_or_region	
RNA_substructure	A secondary structure or combination of secondary structures of a RNA
Lipid	Ccorresponds to *Lipid* category of MeSH.

	"A generic term for fats and lipoids, the alcohol-ether-soluble constituents of protoplasm, which are insoluble in water. They comprise the fats, fatty oils, essential oils, waxes, phospholipids, glycolipids, sulfolipids, aminolipids, chromolipids (lipochromes), and fatty acids. (Grant & Hackh's Chemical Dictionary, 5th ed)"
Carbohydrate	Corresponds to *Carbohydrate* category of MeSH. *"The largest class of organic compounds, including STARCH; GLYCOGEN; CELLULOSE; POLYSACCHARIDES; and simple MONOSACCHARIDES. Carbohydrates are composed of carbon, hydrogen, and oxygen in a ratio of Cn(H2O)n."*
Organic_compound_other	Corresponds to *Organic Chemicals* category of MeSH. *"A broad class of substances containing carbon and its derivatives. Many of these chemicals will frequently contain hydrogen with or without oxygen, nitrogen, sulfur, phosphorus, and other elements. They exist in either carbon chain or carbon ring form."*
Inorganic_compound	Corresponds to *Inorganic Chemicals* category of MeSH. *"A broad class of substances encompassing all those that do not include carbon and its derivatives as their principal elements. However, carbides, carbonates, cyanides, cyanates, and carbon disulfide are included in this class."*
Element	Corresponds to the *Elements* category of MeSH. It covers substances that cannot be decomposed or transformed into other chemical substances by ordinary chemical processes.

2.2. GENIA Anatomy Ontology

The GENIA anatory ontology corresponds to the anatomy category of MeSH. On the top level, there are four classes: Body_region, Tissue, Cell and Cell_component. The Tissue and Cell classes are further classified to natural one and cultured one.

Body_region	Corresponds to the *Body Regions* category of MeSH. *"Anatomical areas of the body."*
Tissue	Corresponds to the *Tissues* category of MeSH. *"Collections of differentiated CELLS, such as EPITHELIUM; CONNECTIVE TISSUE; MUSCLES; and NERVE TISSUE. Tissues are cooperatively arranged to form organs with specialized functions such as RESPIRATION; DIGESTION; REPRODUCTION; MOVEMENT; and others."*
Cell	Corresponds to the *Cells* category of MeSH excluding the *Cell Structure* sub-category which is separated out into *Cell_component.* *"The fundamental, structural, and functional units or subunits of living organisms. They are composed of CYTOPLASM containing various ORGANELLES and a CELL MEMBRANE boundary."*
Cell_cultured	It corresponds to the *Cells, Cultured* category of MeSH. *"The fundamental, structural, and functional units or subunits of living organisms. They are composed of CYTOPLASM containing various ORGANELLES and a CELL MEMBRANE boundary."* The name of this class has been changed from *Cell_line*, as it covers not only cell lines but also other cultured cell types including clone or hybrid cells.
Cell_natural	Naturally existing cells. It is an
Cell_component	Corresponds to the *Cell Structures* category of MeSH. *"Components of a cell."*

2.3. GENIA Organisms Ontology

The GENIA organisms ontology corresponds to the organisms category of MeSH. It classifies a organism to:a multicellular or unicellular organism or a virus. The multicellular organism is further classified to natural or genetically-modified one.

Multicellular_organism	Corresponds to most of the *Organisms* category of MeSH excluding unicellular organisms.
Unicellular_organism	Covers all types of unicellular organisms including yeast, bacteria and so on.
Virus	Corresponds to the *Viruses* category of MeSH. "*Minute infectious agents whose genomes are composed of DNA or RNA, but not both. They are characterized by a lack of independent metabolism and the inability to replicate outside living host cells.*"

3. GENIA event ontology

The GENIA event ontology is designed to support the GENIA event annotation of which the main purpose is to find natural language expressions mentioning biological processes and molecular functions.

The GENIA event ontology is intended to have connections with the Gene Ontology to raise the utility of the ontology and the annotation. In the following two subsections the biological process and molecular function ontologies of GENIA is described respectively. When necessary, the definition in GO will be excerpted in double quotation marks.

3.1. Biological process

The biological process concept corresponds to the same name of category in GO. Although they share the same definition, it should be noted that the biological process defined in GENIA ontology covers much smaller concepts than in GO. In the GENIA ontology, only cellular processes, processes related to viral life cycle and regulation process of them are covered.

"A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cellular physiological process or signal transduction. Examples of more specific terms are pyrimidine metabolism or alpha-glucoside transport. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps."

Cellular_process	Corresponds to the GO term "*cellular process*" except two sub-terms: "*regulation of cellular process*" and "*cellular metabolism*". The former is classified as the general "*Regulation*" class and the latter is as the general "*Metabolism*" class here. *"Processes that are carried out at the cellular level, but are not necessarily restricted to a single cell. For example, cell communication occurs among more than one cell, but occurs at the cellular level."*
Cell_adhesion	Corresponds to the GO term "*cell adhesion*". *"The attachment of a cell, either to another cell or to an underlying substrate such as the extracellular matrix, via cell adhesion molecules."*
Cell_communication	Corresponds to the GO term "*cell communication*". *"Any process that mediates interactions between a cell and its surroundings. Encompasses interactions such as signaling or attachment between one cell and another cell, between a cell and an extracellular matrix, or between a cell and any other aspect of its environment."*
Cell_differentiation	Corresponds to the GO term "*cell differentiation*". *"The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history. Differentiation includes the processes involved in commitment of a cell to a specific fate."*
Cell_recognition	Corresponds to the GO term "*cell recognition*". *"The process by which a cell in a multicellular organism interprets its surroundings."*

Cellular_physiological _process	Corresponds to the GO term "*cell physiological process*" except the sub-term "*cellular metabolism*" which is classified as the general "*Metabolism*" class here. *"The processes pertinent to the integrated function of a cell."*
Physiological_process	Corresponds to the GO term "*physiological process*" except the sub-term "*cellular physiological process*" which is classified as a subclass of the *Cellular_process* class here.
Localization	Corresponds to the GE term "*localization*". *"The processes by which a cell, a substance, or a cellular entity, such as a protein complex or organelle, is transported to, and/or maintained in, a specific location.*"
Metabolism	Corresponds to the GO term "*metabolism*". *"Processes that cause many of the chemical changes in living organisms, including anabolism and catabolism. Metabolic processes typically transform small molecules, but also include macromolecular processes such as DNA repair and replication, and protein synthesis and degradation."*
Protein_metabolism	Corresponds to the GO term "*protein metabolism*". *"The chemical reactions and pathways involving a specific protein, rather than of proteins in general. Includes protein modification."*
Protein_biosynthesis	Corresponds to the GO term "*Metabolism*". *"The chemical reactions and pathways resulting in the formation of a protein, rather than of proteins in general."*
Translation	Corresponds to the GO term "*translation*". *"A ribosome-mediated process in which the information in messenger RNA (mRNA) is used to specify the sequence of amino acids in a polypeptide chain."*
Protein_catabolism	Corresponds to the GO term "*protein catabolism*". *"The chemical reactions and pathways resulting in the breakdown of a protein by the destruction of the native, active configuration, with or without the hydrolysis of peptide bonds."*
Protein_modification	Corresponds to the GO term "*protein modification*". *"The covalent alteration of one or more amino acid residues within a protein, resulting in a change in the properties of that protein."*
Protein_amino_acid _phosphorylation	Corresponds to the GO term *"protein_amino_acid_phosphorylation"*. *"The process of introducing a phosphoric group on to a protein."*
Protein_amino_acid _dephosphorylation	Corresponds to the GO term "*protein_amino_acid_dephosphorylation*". *"The process of removing one or more phosphoric residues from a protein."*
Protein _ubiquitination	Corresponds to the GO term "*protein_ubiquitination*". *"The process by which one or more ubiquitin moieties are added to a protein."*
Protein _deubiquitination	Corresponds to the GO term "*protein_deubiquitination*". *"The removal of one or more ubiquitin moieties from a protein."*
DNA_metabolism	Corresponds to the GO term "*DNA metabolism*". *"The chemical reactions and pathways involving DNA, deoxyribonucleic acid, one of the two main types of nucleic acid, consisting of a long, unbranched macromolecule formed from one, or more commonly, two, strands of linked deoxyribonucleotides."*
DNA_modification	Corresponds to the GO term "*DNA modification*". *"The covalent alteration of one or more nucleotide sites in DNA, resulting in a change in its properties."*
DNA_recombination	Corresponds to the GO term "*DNA recombination*". *"The processes by which a new genotype is formed by reassortment of genes resulting in gene combinations different from those that were present in the parents. In eukaryotes genetic recombination can occur by chromosome assortment, intrachromosomal recombination, or nonreciprocal interchromosomal recombination. Intrachromosomal*

	recombination occurs by crossing over. In bacteria it may occur by genetic transformation, conjugation, transduction, or F-duction."
DNA_repair	Corresponds to the GO term "*DNA repair*". *"The process of restoring DNA after damage. Genomes are subject to damage by chemical and physical agents in the environment (e.g. UV and ionizing radiations, chemical mutagens, fungal and bacterial toxins, etc.) and by free radicals or alkylating agents endogenously generated in metabolism. DNA is also damaged because of errors during its replication. A variety of different DNA repair pathways have been reported that include direct reversal, base excision repair, nucleotide excision repair, photoreactivation, bypass, double-strand break repair pathway, and mismatch repair pathway."*
RNA_metabolism	Corresponds to the GO term "*RNA metabolism*". *"The chemical reactions and pathways involving RNA, ribonucleic acid, one of the two main type of nucleic acid, consisting of a long, unbranched macromolecule formed from ribonucleotides joined in 3',5'-phosphodiester linkage."*
RNA_modification	Corresponds to the GO term "*RNA modification*". *"The covalent alteration of one or more nucleotides within RNA, resulting in a change in the properties of the RNA."*
Transcription	Corresponds to the GO term "*RNA metabolism"* *"The synthesis of either RNA on a template of DNA or DNA on a template of RNA."*
Gene_expression	The phenotypic manifestation of a gene or genes by the processes of genetic transcription and genetic translation.
Regulation	Corresponds to the GO term "*regulation of biological process*". "*Any process that modulates the frequency, rate or extent of a biological process. Biological processes are regulated by many means; examples include the control of gene expression, protein modification or interaction with a protein or substrate molecule.*"
Positive_regulation	Corresponds to the GO term "*positive regulation of biological process*". "*Any process that activates or increases the rate, frequency or extent of a biological process. Biological processes are regulated by many means; examples include the control of gene expression, protein modification or interaction with a protein or substrate molecule.*"
Negative_regulation	Corresponds to the GO term "*negative regulation of biological process*". "*Any process that stops, prevents or reduces the rate, frequency or extent of a biological process. Biological processes are regulated by many means; examples include the control of gene expression, protein modification or interaction with a protein or substrate molecule.*"
Viral_life_cycle	Corresponds to the GO term "*viral life cycle*". "*A set of processes by which a virus reproduces. Usually, this is by infection of a host cell, replication of the viral genome, and assembly of progeny virus particles. In some cases the viral genetic material may integrate into the host genome and only subsequently, under particular circumstances, 'complete' its life cycle.*"
Initiation_of_viral_infection	Corresponds to the GO term "*initiation of viral infection*". "*Processes involved in the start of virus infection of cells.*"
Viral_genome_expression	Corresponds to the GO term "*viral genome expression*". "*The achievement of highly specific, quantitative, temporal and spatial control of virus gene expression within the limited genetic resources of the viral genome.*"

3.2. Molecular function

The GENIA event ontology defines some molecular functions. In the ontology, molecular functions are regarded as potential events.

"*Molecular function describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.*"

Molecular_function	Corresponds to the GO category "molecular function"
Binding	Corresponds to the GO term "binding". "The selective, often stoichiometric interaction of a molecule with one or more specific sites on another molecule."
Catalysis	Corresponds to the GO term "catalytic activity". "*Catalysis of a biochemical reaction at physiological temperatures. In biologically catalyzed reactions, the reactants are known as substrates, and the catalysts are naturally occurring macromolecular substances known as enzymes. Enzymes possess specific binding sites for substrates, and are usually composed wholly or largely of protein, but RNA that has catalytic activity (ribozyme) is often also regarded as enzymatic.*"

3.3. ETC.

The GENIA event ontology also defines two concepts prepared to support annotation to text.

Experimental_process	Specially designed experimental process. Use of this concept for annotation is limited to
Correlation	correlation between events or entities mentioned in text without referring to specific type of correlation.