

Classifying Molecules by their Roles

Satoko Yamamoto¹, Toshihisa Takagi², Ken-ichiro Fukuda³

(¹Institute for Bioinformatics Research and Development, Japan Science and Technology Agency, ²Graduate School of Frontier Sciences, University of Tokyo, ³Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology)

In the scientific literature, a molecule name may stand for concepts of various granularities, from concrete objects, such as Smad1 and ERK1, to abstract notions or categories, such as R-Smad and MAPK. Although biologists know that "Smad1 is one of the R-Smads, and ERK1 is one of the MAPKs", a computer has no such background knowledge. And this becomes a problem when one develops a literature knowledge based database, since the relation between Smad1 and R-Smads is lost. Typically, these relations between molecule names derive a hierarchical structure and a simple synonym dictionary would be insufficient to represent the relations between names. For example, what does "NF-kappaB" in the literature refer to? Mammals express five Rel (NF-kappaB) family proteins and they act as various homo- and heterodimers. Such information is not solvable completely with synonym. To resolve this problem, we are developing the MoleculeRole ontology, which connects the abstract molecule name to the concrete molecule name.

In order to perform a complicated query that needs to traverse the concept space of abstract and concrete molecule names, it is indispensable to manage the relation between the abstract name and the concrete substance unitarily, with consistency. Typically, however, this information on molecular function or families is stored in text form or hyperlink. For example, the "Ras" link in a clickable map takes the user to molecule information on "H-Ras."

To define a reusable and explicit classification system of molecules that enables a complicated search on textual knowledge database, the MoleculeRole ontology was built by the following methods. For the higher level of the ontology, the term referring to a molecule group was extracted from typical reviews and original literature in molecular biology, and all such terms were arranged in a hierarchical structure. This was designed with consideration of the structure's potential to become a general, acceptable classification for biologists. This classification is based on a conceptual classification of

molecular roles in protein interaction and signal transduction and is not identical to the families derived from sequence analysis. For example, it contains concepts such as "adapter protein", referring to the molecules that mediate molecular interactions, which is used as the branch point of a signal, and "signal regulator", which refers to molecules that control a signal to be positive or negative. Next, the entry of SwissProt/TrEMBL was manually attached to each leaf term of the ontology, and the higher level concept of the ontology was related with the corresponding term in Gene Ontology. The molecular complex was also added to the ontology, and a relation between a complex and its subunits was defined as a "part-of" relationship. The other class relation is "is-a". Moreover, the chemical compounds considered to be particularly important in the signal transduction field (e.g., second messenger, such as cAMP and calcium ion) were classified and added to the ontology. At present, the number of entries is 1124.

A resource like MoleculeRole ontology mitigates the burden of data curation sharply. It also becomes easy to carry out the kind of complicated search that is difficult to realize in a keyword-search based database. First, it becomes easy to fill in the gap between names. Although ERK1 cannot be found by a keyword-search even if MAPK is specified as a query term, it can be found by an ontology-based-search by carrying out a search that includes the child of MAPK on MoleculeRole ontology. Some databases realize this function by defining keywords and synonyms relative to data on each molecule (the keyword "MAPK" being attached to the "ERK1" molecule), but it is difficult to define a class relation, and it is inferior with regard to unitary management and the reusability of data. Second, it is possible to perform a query relaxation search that expands the concept relevant to the user-specified concept. By expanding the user-specified concept "MAPK" according to the ontology (protein serine/threonine kinase), information about all molecules with the same function can be acquired.

This work is part of our ongoing pathwaydatabase project (<http://www.inoh.org>). We are planning to curate more data and make the ontology open to the public.