Linking Molecular Imaging Terminology to the Gene Ontology (GO)

P.K. Tulipano, W. Millar, J.J. Cimino

Pacific Symposium on Biocomputing 8:613-623(2003)

LINKING MOLECULAR IMAGING TERMINOLOGY TO THE GENE ONTOLOGY (GO)

P.K. TULIPANO,[1] W.S. MILLAR,[2] J.J. CIMINO[1]
Department of Medical Informatics[1]*, Department of Radiology*[2]
Columbia University, New York, NY 10032, USA
Email: tulipano@dmi.columbia.edu, wsm8@columbia.edu, cimino@dmi.columbia.edu

The rapidly developing domain of molecular imaging represents the merging of current advances in the fields of molecular biology and imaging research. Despite this merger, an information gap continues to exist between the scientists who discover new gene products and the imaging scientists who can exploit this information. The Gene Ontology (GO) Consortium seeks to provide a set of structured terminologies for the conceptual annotation of gene product function, process and location in databases. However, no such structured set of concept-oriented terminology exists for the molecular imaging domain. Since the purpose of GO is to capture the information about the role of gene products, we propose that the mapping of GO's established ontological concepts to a molecular imaging terminology will provide the necessary bridge to fill the information gap between the two fields. We have extracted terms and definitions from an already published molecular imaging glossary as well as molecular imaging research articles, and developed molecular imaging concepts. We then mapped our molecular imaging concepts to the existing gene ontology concepts as a method to comprehensively represent molecular imaging.

1 Introduction

Advances in medical imaging such as improved image resolution, new imaging agents, and experimental micro imaging devices, have stimulated interest in the *in vivo* assessment of molecular interactions and pathways.[1] These advancements coupled with the latest genomic discoveries have led to the rapid development of the molecular imaging domain. The integration of molecular sciences with imaging research has created a new cross-domain environment for molecular imaging scientists. Molecular imaging scientists are challenged with "learning the language" of the basic molecular sciences. In addition, they must keep up with the vast amount of information generated in the maturing fields of molecular imaging and molecular biology.

We have developed a controlled terminology of molecular imaging as the foundation for the integration of this cross-domain knowledge. The terminology will provide a comprehensive representation of the concepts in molecular imaging. We propose that linking our terminology to the concepts organized in the Gene Ontology will facilitate communications among the disciplines of molecular imaging and molecular biology.

2 Background

Molecular imaging is defined as “the *in vivo* characterization and measurement of biological processes at the cellular and molecular level through the use of imaging devices”.[2] While conventional imaging captures the phenotypic changes at the gross anatomic level that result from molecular processes, molecular imaging attempts to detect detailed information about the underlying molecular and cellular processes themselves. The key elements necessary for molecular imaging are: 1) highly specific imaging probes with high affinity for their targets and acceptable biological delivery, 2) identification of suitable targets, 3) appropriate amplification strategies, and 4) sensitive and fast imaging systems with high resolution.[3,4] The goals of molecular imaging research are the exploration of these key elements and the development of new agents, strategies and imaging techniques for *in vivo* imaging.[4] For example, EgadME is one of a new class of chemicals dubbed “smart contrast agents”,[5] so called because it is activated solely in the presence of the gene encoding beta-galactosidase. Researchers captured the MR signal produced by EgadME’s interaction with cells expressing the beta-galactosidase gene using conventional *in vivo* MR imaging.

Traditional biological techniques have already provided detailed molecular *in vitro* diagnostic information.[6] Molecular imaging research can draw upon these *in vitro* techniques to provide *in vivo* diagnostic information. For example, many of the detectable molecular imaging parameters, such as cell surface receptors and enzymatic activity, should be identical to those found *in vitro*. Information acquired about molecular function and pathways will aid in the determination of specific molecular targets as well as the development of novel imaging agents. The clinical success of a contrast agent or molecular probe, however, may be related to additional in vivo factors, such as biocompatibility and directional transport to the target molecule.[6]

The description, classification, and organization of biological objects has become increasingly important, particularly in bioinformatics.[7-9] The structuring of biological information can be accomplished through the use of ontological methods**.** The Gene Ontology (GO) Consortium was established in 1998 to develop shared, structured terminologies for molecular characteristics across three model organism databases: SGD, the *Saccharomyces* Genome database, FlyBase, the *Drosophila* genome database, and MGD/GXD, the Mouse Genome Informatics databases.[10] The GO project became involved in the development of a database resource that allows access to datasets that utilize a standardized terminology for genes and gene product.[10] The GO polyhierarchy consists of three ontologies: 1) molecular function, 2) biological process, and 3) cellular component. These three ontologies were chosen because they are common to all living organisms and are basic to annotations of information about genes and gene products.[11] The GO project has expanded considerably since its inception to include other databases such as WormBase and Rat Genome Database (RGD).[12]

Currently, an information gap exists between the molecular scientists who discover new gene products and the imaging scientists who can exploit the gene product functions into new noninvasive imaging methods.[4] A structured set of terminologies for the conceptual annotation of gene product function, process and location in databases would be useful in this regard; however, no such structured set of concept-oriented terminology exists for the molecular imaging domain.

The need for concept-oriented terminologies has been recognized and addressed by many medical informatics researchers.[13] The development of a controlled terminology is valuable in specifying and organizing concepts important in the domain. In addition, a controlled terminology provides the framework on which the development of informatics tools can evolve. Informatics tools such as automatic information retrieval systems, indexed image retrieval databases, and decision support systems have been developed and rely on existing terminologies such as the Unified Medical Language Systems (UMLS) and the Medical Entities Dictionary (MED).[14-17]

There are well established informatics standards for the development of a controlled terminology.[18,19] The Desiderata is a set of standards necessary for the development of a standard, reusable multipurpose terminology.[19] Some of the requirements of the Desiderata include domain content coverage, concept orientation, nonsemantic concept identifiers, polyhierarchy, multiple granularities, and multiple consistent views. A terminology must be able to provide appropriate coverage of the domain's concepts (*domain coverage*). The concept is the unit of representation and must have a single, coherent meaning within the terminology (*concept orientation*). The concepts must be represented by meaningless, unique identifiers that are free of hierarchical meaning (*nonsemantic identifiers*). Such identifiers allow for multiple classifications and rearrangement of concepts within a hierarchy. A terminology should have a hierarchical arrangement that allows assignment of concepts in one or more areas of the hierarchy (*polyhierarchy*). In addition, to provide for multiple user functionality, a terminology must provide different levels of granularity and must maintain a consistent view throughout its hierarchy (*multiple granularities and consistent views*).

A number of formal representations exist for the modeling of controlled terminologies. One particular model is frame-based that includes a directed acyclic graph (DAG) as its hierarchy structure.[20,21] In a DAG hierarchy, concept nodes are children of one or more parent nodes. The hierarchical relationships from child to parent are of 'is a' type. In a frame-based model, each concept node can also be viewed as a frame with named slots. Slots may have values associated with them. Each concept node can also be viewed as having nonhierarchical relationships to other nodes through named slots or semantic links. The MED, currently in place at Columbia University, is an example of a frame-based model that closely adheres to the guidelines enumerated in the Desiderata. The development of our controlled terminology is based on this representation.

We propose that the establishment of a controlled terminology of molecular imaging and its linkage to the GO's ontological concepts will provide the necessary bridge to fill the information gap between domains as well as facilitate communications and knowledge sharing between domains.

3 Methods

A literature review of the molecular imaging domain was performed in order to extract concepts and terms. Thirty molecular imaging papers were selected and manually reviewed; topics ranged from a broad molecular imaging overview[3] to more specific applications.[22,23] In addition, a molecular imaging glossary was retrieved and used as the starting point for collection of molecular imaging terms.[2]

The terms defined in the molecular imaging glossary were extracted and initially divided into four general classifications. Several iterations were required to classify, and then reclassify, the terms into appropriate classes. Terms from other articles were subsequently added. The final iteration involved the dissection of terms into general classifications, placement of terms into a hierarchy, and the assignment of concept node attributes.

A frame-based representation model was developed with a directed acyclic graph (DAG) as its hierarchical structure, similar to the representation found in the MED.[20] Each concept in the terminology was assigned a unique identifier and a unique name. Each of the concepts was assigned named attributes that may or may not have values. The top-level node concept and its four descendants are listed and defined in Table 1.

Table 1. Top Level Concepts Names and Definitions

Concept Name	Definition
Molecular Imaging Entity	A broad type for grouping physical and conceptual entities related to the domain of molecular imaging.
Imageable Probe	A broad type for any highly specific agent (such as a radiolabelled drug or conjugated antibody) used in imaging to report on an event.
Imageable Target	A broad type for any target (usually a protein or gene product) that interacts with an imageable molecular probe.
Amplification Technique	A method for increasing imageable signal.
Imaging Instrument	A device for determining the presence, measure, and time/spatial distribution of a quantity under observation

In addition, the GO's HTML browser, AmiGO!,[10] was used to search for terms that existed in our terminology. GO terms that mapped to our terminology were noted in the 'GO Code' slot for that particular concept.

4 Results

The Glossary of Molecular Imaging Terminology[2] contained a total of 197 terms and definitions. Forty-eight terms were related specifically to molecular imaging. Sixteen additional terms were obtained from the literature. In addition to terms, the molecular imaging glossary included six abbreviations that were specific to molecular imaging.

A frame-based knowledge model, based on the Medical Entities Dictionary, was created as the representation for the terminology. A directed acyclic graph (DAG) formed the hierarchical structure. Figure 1 represents a subset of the information we found relevant to the domain. The top-level parent is the concept *Molecular Imaging Entity*.

Direct descendants correspond to the four key elements of molecular imaging: *Imageable Targets*, *Imageable Probes*, *Amplification Techniques* and *Imaging Instruments.* Concepts were assigned named attributes that may contain corresponding values. Eight named attributes were assigned: 'Name', 'Synonyms', 'Abbreviations', 'Definition', 'GO Code', 'NonSemanticID', 'DescendantOf'. Hierachical relationships between concept nodes are 'is a' type. Figure 2 depicts a concept node frame with associated slot.

A search of all molecular imaging specific terms revealed 11 terms that were GO related. All GO codes were mapped through the named attribute of 'GO Code' in the concept frame. All of the GO terms found were mapped to our molecular imaging terminology through the general class of 'Imageable Targets'. The mapping of the gene ontology with the concepts of molecular imaging is listed in Table 2.

Table 2. Linking of GO concepts to our molecular imaging terminology

Concept Name	Molecular Imaging NonSemantic ID	GO Code
Enzyme	10	0003824
Cell Surface Receptor	37	0007166
Thymidine Kinase	32	0004797
Creatine Kinase	33	0004111
Tyrosinase	34	0009309
Somatostatin Receptor	35	0004994
Cytosine Deaminase	36	0004131
Beta-Galactosidase	38	0004565
Dopamine Receptor	39	0004952
NADPH-ferrihemoprotein reducatase	40	0003985
Gastrin Receptor	41	0015054

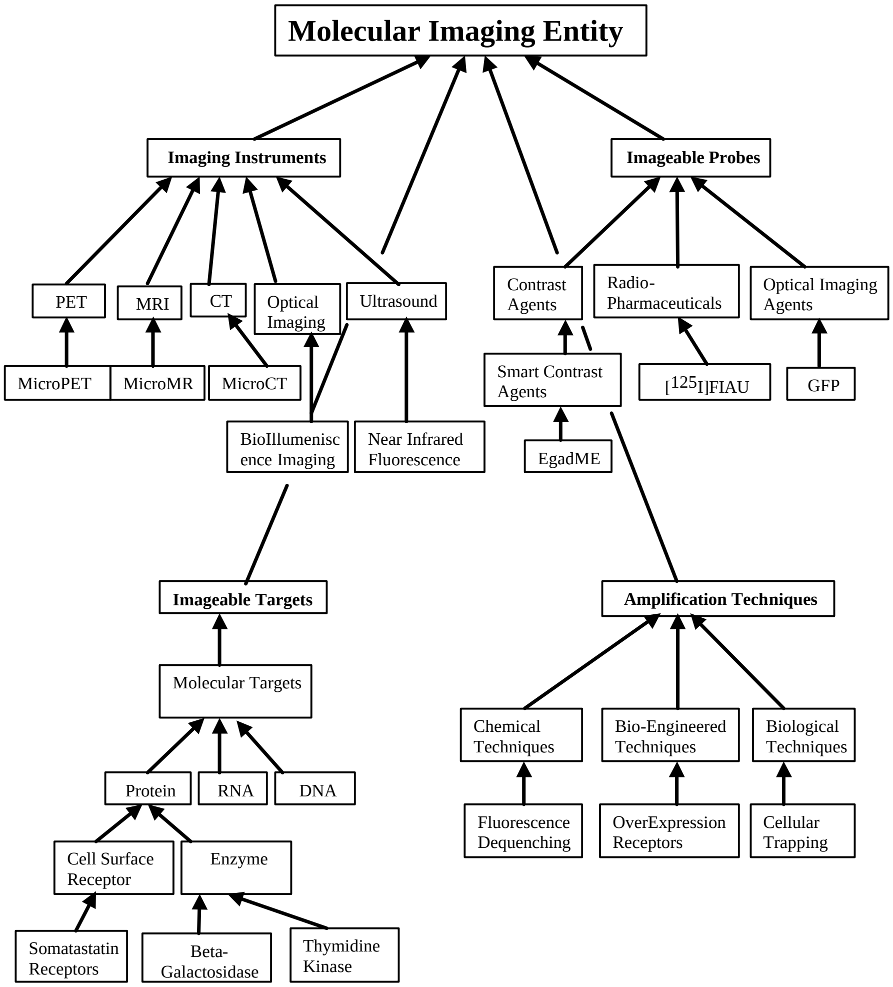

Figure 1. Molecular Imaging Representation. The direct descendants of the four key elements of Molecular Imaging (Imageable Probes, Imageable Targets, Imaging Instruments, and Amplification Techniques) and a subset of children nodes are depicted.

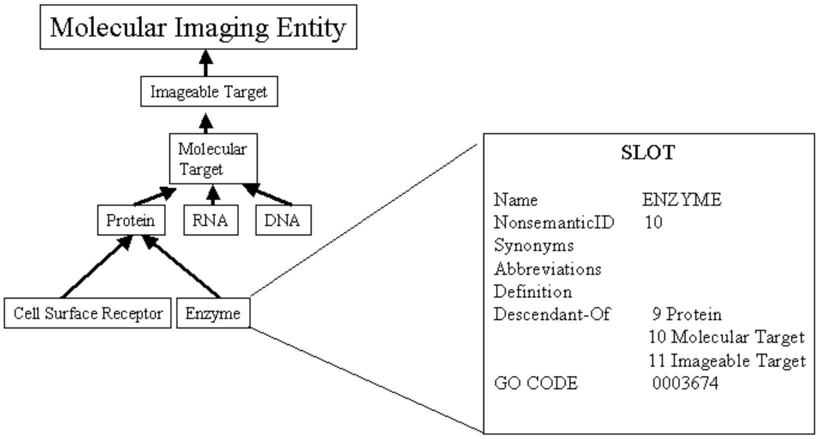

Figure 2. Concept Node Frame with Associated Slots and Slot Values

5 Discussion

The Glossary of Molecular Imaging Terminology developed by Wagenaar is a first attempt in defining terms for the molecular imaging domain.[2] Although not a comprehensive glossary, it does provide a foundation for the development of a molecular imaging specific terminology. The glossary is composed of terms relevant to the fields of molecular imaging and also the molecular sciences. One source of difficulty in extracting and classifying terms that exist in the realm of molecular imaging is that the 'language' used by basic biological scientists is also used in molecular imaging. The same terms may have several different meanings. For instance, the term, *amplification*, is defined in biology as 'an increase in the number of copies of a specific DNA fragment'. In imaging, however, amplification refers to an increase, not in the copies of DNA, but in the imageable signal.[3,24] Another interesting overlap of terms is the notion of a *reporter gene*. The *E. coli* beta-galactosidase gene has been extensively used in basic science research as a reporter gene. A reporter gene is defined as a gene that encodes an easily assayed and detectable protein. In molecular imaging, the beta-galactosidase gene is now used as an imaging reporter gene (also referred to as a *marker gene* in biology and a *imaging marker gene* in molecular imaging).[25] Its product is referred to as an *imaging reporter product* (also referred to as an *imaging molecular target*). These issues of concept ambiguity are resolved in our terminology through the use of nonsemantic identifiers as described in the Desiderata guidelines.[19] We uniquely identify concepts that correspond to a single, coherent meaning. These examples demonstrate that classifying terms specific to molecular imaging and do not cross into other related fields, is challenging since molecular imaging relies heavily on the information and resources previously developed in the basic scientific areas.

Despite the overlapping terms, concepts were created to represent molecular imaging terms as unambiguously as possible. There are four essential properties of molecular imaging. The direct descendants of the topmost concept node, 'Molecular Imaging Entity', are defined to reflect those key elements.[3,4] An 'Imageable Target' is essential since it is the target expressions and or pathway that *in vivo* imaging attempt to visualize. An 'Imageable probe' is required to probe for or locate the target of interest. Currently, 'Amplification Techniques' are required to increase the imageable signal because the resolutions of some molecular imaging instruments cannot detect the relatively small size of molecules involved in a gene expression or a molecular pathway. 'Imaging Instruments' are required to detect the 'Imageable Probe'. Other concept nodes can be added as the field matures; 'Smart Contrast Agents' and 'MicroPET' have been added to reflect the recent developments in molecular imaging.

Our controlled terminology uses a model similar to that found in the MED. In selecting this model, we expect that our terminology would follow the guidelines enumerated in the Desiderata. In terms of domain coverage, we have added concepts that we considered reflective of the current state of the domain at the time of its development. Molecular imaging is a dynamic and continuously evolving field. As such, the content of the molecular imaging terminology will also be expanding and evolving. The current structure supports the addition of new terms and the rearrangement of old terms. In addition, we provided each concept node with a slot 'NonsemanticID' that contains a meaningless, unique identifier. The directed acyclic graph structure supports multiple parents, which satisfies the Desiderata's notion of a polyhierarchy.

For example, in our terminology, 'green fluorescent protein' is considered both a child of 'optical imaging agent' and 'molecular target' because green fluorescent protein is an auto-fluorescent molecule;[26] that is, it requires no additional agent for its presence to be detected by imaging techniques. Therefore, concepts in our terminology can have multiple classifications. Unfortunately, the one desideratum not satisfied in our terminology is the notion of multiple levels of granularities. As the field progresses, new concepts and relationships will emerge that will contribute to the increase in finer granularity of the terminology.

The mappings in Table 2 demonstrate several salient points. Figure 3 is an example from Table 2 of the linking of the imageable target 'thymidine kinase' to the GO concept 'thymidine kinase'. The gene ontology mapped this item into two concepts. One is the concept of thymidine kinase as a subclass of kinases. In addition, 'thymidine kinase' mapped to the higher-level concept of 'transferase' in GO. The gene ontology links kinases and transferases to external gene databases. The gene databases contain the relevant gene information including the organism from which the genes were derived and the relevant literature associated with the gene. In one imaging strategy, the gene expression of thymidine kinase from the Herpes Simplex Virus 1 is imaged.[27] A possible inference that can be made from this mapping is that all transferases and kinases are imageable targets. In addition, it may be possible to image thymidine kinase expression from other organisms as well. Therefore, the possibility of such hypotheses being utilized and tested by imaging experts may prove useful. In addition to the generation of testable hypotheses, the linking of the two terminologies provides a connection between information found throughout research articles in both domains. The development of specialized computer applications that utilize this connection can facilitate automatic information retrieval from the enormous number of publications in molecular imaging and genomics. Automatic information retrieval applications are being developed for genomics. Our terminology may provide the foundation for such applications in molecular imaging.

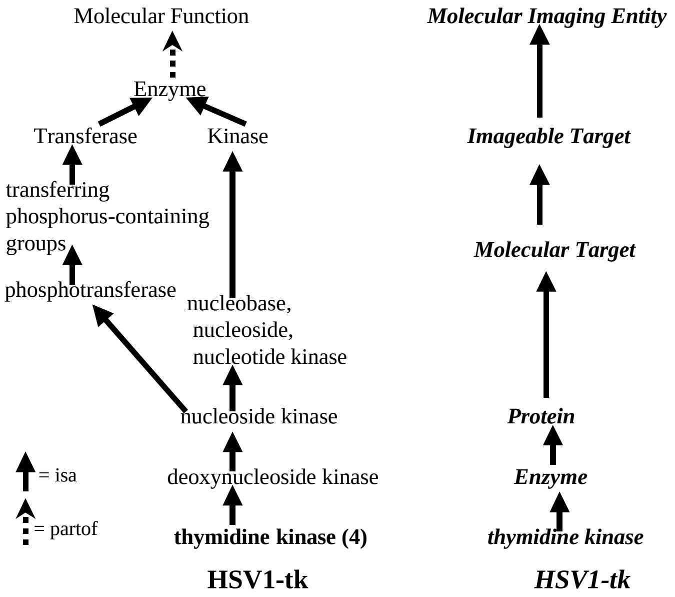

Figure 3. The association of the Herpes Virus Simplex 1 thymidine kinase gene product from GO to our molecular imaging terminology.

Two commonly used imaging targets were not originally described in GO. Green fluorescent protein (GFP), isolated from coelenterates such as the Pacific jellyfish, is an imageable target used in optical imaging.[28] There is no reference to this protein and it's function in the GO.[12] Another protein, luciferase (from the firefly Photinus pyralis), is a bioluminescence imageable target.[26] A query of the current version of the GO's gene products for luciferase did not retrieve it.[12]

6 Conclusion

We have developed a controlled terminology of molecular imaging to represent the concepts appropriately within the domain. We have linked our molecular imaging terminology to the concepts in the Gene Ontology. We proposed that this linking of concepts facilitates knowledge sharing among molecular imaging and molecular scientists. Our terminology may provide the foundation for use of automatic information retrieval applications in the emerging field of molecular imaging. Future work will include the expansion of the terminology and the determination of the complex relationships between concepts.

Acknowledgments

We wish to acknowledge our colleague, Neil Sarkar, for his helpful comments and insightful discussions. Travel support provided by Columbia University Corporate Affiliates Program.

References

1. R. Weissleder, "Molecular imaging: exploring the next frontier" Radiology 212, 609 (1999)
2. D.J. Wagenaar et al., "Glossary of molecular imaging terminology" Acad. Radiol. 8, 409 (2001)
3. M.G. Pomper, "Molecular imaging: an overview" Acad. Radiol. 8, 1141 (2001)
4. R. Weissleder and U. Mahmood, "Molecular imaging" Radiology 219, 316 (2001)
5. A.Y. Louie et al., "In vivo visualization of gene expression using magnetic resonance imaging" Nat. Biotechnol. 18, 321 (2000)
6. A. Hengerer and T. Mertelmeier, "Molecular Biology for Medical Imaging" electromedica 69, 44 (2001)
7. A. Rzhetsky et al., "A knowledge model for analysis and simulation of regulatory networks" Bioinformatics 16, 1120 (2000)
8. S. Schulze-Kremer, "Ontologies for molecular biology" Pac. Symp. Biocomput., 695 (1998)
9. R. Stevens et al., "Building a bioinformatics ontology using OIL" IEEE Trans. Inf. Technol. Biomed. 6, 135 (2002)
10. The Gene Ontology Consortium, "Creating the gene ontology resource: design and implementation" Genome Res. 11, 1425 (2001)

11. M. Ashburner et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium" Nat. Genet. 25, 25 (2000)
12. http://www.geneontology.org, 2002
13. C.G. Chute et al., "A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures" J. Am. Med. Inform. Assoc. 5, 503 (1998)
14. D.C. Berrios, "Automated indexing for full text information retrieval" Proc. AMIA Symp., 71 (2000)
15. J.J. Cimino et al., "Supporting infobuttons with terminological knowledge" Proc. AMIA Annu. Fall Symp., 528 (1997)
16. W.D. Bidgood, Jr. et al., "Image acquisition context: procedure description attributes for clinically relevant indexing and selective retrieval of biomedical images" J. Am. Med. Inform. Assoc. 6, 61 (1999)
17. D.A. Lindberg et al., "The Unified Medical Language System" Methods. Inf. Med. 32, 281 (1993)
18. P.L. Elkin et al., "Guideline for health informatics: controlled health vocabularies--vocabulary structure and high-level indicators" Medinfo. 10, 191 (2001)
19. J.J. Cimino, "Desiderata for controlled medical vocabularies in the twenty-first century" Methods Inf. Med. 37, 394 (1998)
20. J.J. Cimino et al., "Knowledge-based approaches to the maintenance of a large controlled medical terminology" J. Am. Med. Inform. Assoc. 1, 35 (1994)
21. J.J. Cimino et al., in Secondary "Designing an introspective, multipurpose, controlled medical vocabulary" (Washington, DC, 1989)
22. C.H. Tung et al., "Preparation of a cathepsin D sensitive near-infrared fluorescence probe for imaging" Bioconjug. Chem. 10, 892 (1999)
23. Y. Yu et al., "Quantification of target gene expression by imaging reporter gene expression in living animals" Nat. Med. 6, 933 (2000)
24. D. Hogemann and J.P. Basilion, ""Seeing inside the body": MR imaging of gene expression" Eur. J. Nucl. Med. Mol. Imaging 29, 400 (2002)
25. C. Bremer and R. Weissleder, "In vivo imaging of gene expression" Acad. Radiol. 8, 15 (2001)
26. M.J. Hickey et al., "Luciferase in vivo expression technology: use of recombinant mycobacterial reporter strains to evaluate antimycobacterial activity in mice" Antimicrob. Agents Chemother. 40, 400 (1996)
27. R.G Blasberg and Tjuvajev J.G , "Herpes simplex virus thymidine kinase as a marker/reporter gene for PET imaging of gene therapy", Q J Nucl Med 43(2):163-9 (1999)
28. M. Chalfie et al., "Green fluorescent protein as a marker for gene expression" Science 263, 802 (1994)