



# Formal Ontology and Clinical bioinformatics

Anand Kumar  
IFOMIS, University of Leipzig

# Background: How the two worlds meet?

MedNet 2003, 4-7 December, Geneva



The 8th Annual World Congress on the Internet and Medicine "Internet in Health for All",

Clinical

Specific disease topics  
E-Health, Health support system  
Patient Management  
Patient health education



Molecular  
Biology

Biomedical Information on the Web  
Swiss-Prot, Swiss-Prot Variant pages  
Proteins, mutations, functions and structures

# Project overview: Ontology as a bridge?

## Diseases

- Pathological Processes
- Body site for diseases
- Diseases by staging
- Risk factors

## Anatomy

- Is-a, part-of
- Granular relationship

## Biological Processes

- Ontology
- Classification

## Swiss-Prot proteins

- Annotation: function, structure, mutation


SNOMED

FMA

GO



**Step 1: In search for a case study**  
**Disease ontology applied to**  
**Swiss-Prot proteins**



## **Problematic:** Choose a disease most represented in annotated Swiss-Prot human proteins with variants

- In Swiss-Prot, the disease information is stored in CC-disease, FT variant description, as well as in the ModSNP database.
- Although most of the diseases have a corresponding OMIM id, to know which proteins are associated with a ‘broader’ disease type (e.g. cardiovascular disease) is not trivial.

# Example

P08235 (MCR\_HUMAN): eg. VAR\_015626

- Early onset hypertension (MIM:605115)

hypertensive disorder

vascular disease

P02545 (LAMA\_HUMAN): eg. VAR\_009973

- CMD1A with quadriceps myopathy (MIM:607920)

- Dilated cardiomyopathy 1A (MIM:115200)

myocardial disease

heart disease

P37023 (KIR3\_HUMAN): eg. VAR\_006208

- Osler-Rendu-Weber syndrome 2 (MIM:600376)

vascular disease of the skin

vascular disease

Disease of  
Cardiovascular  
system

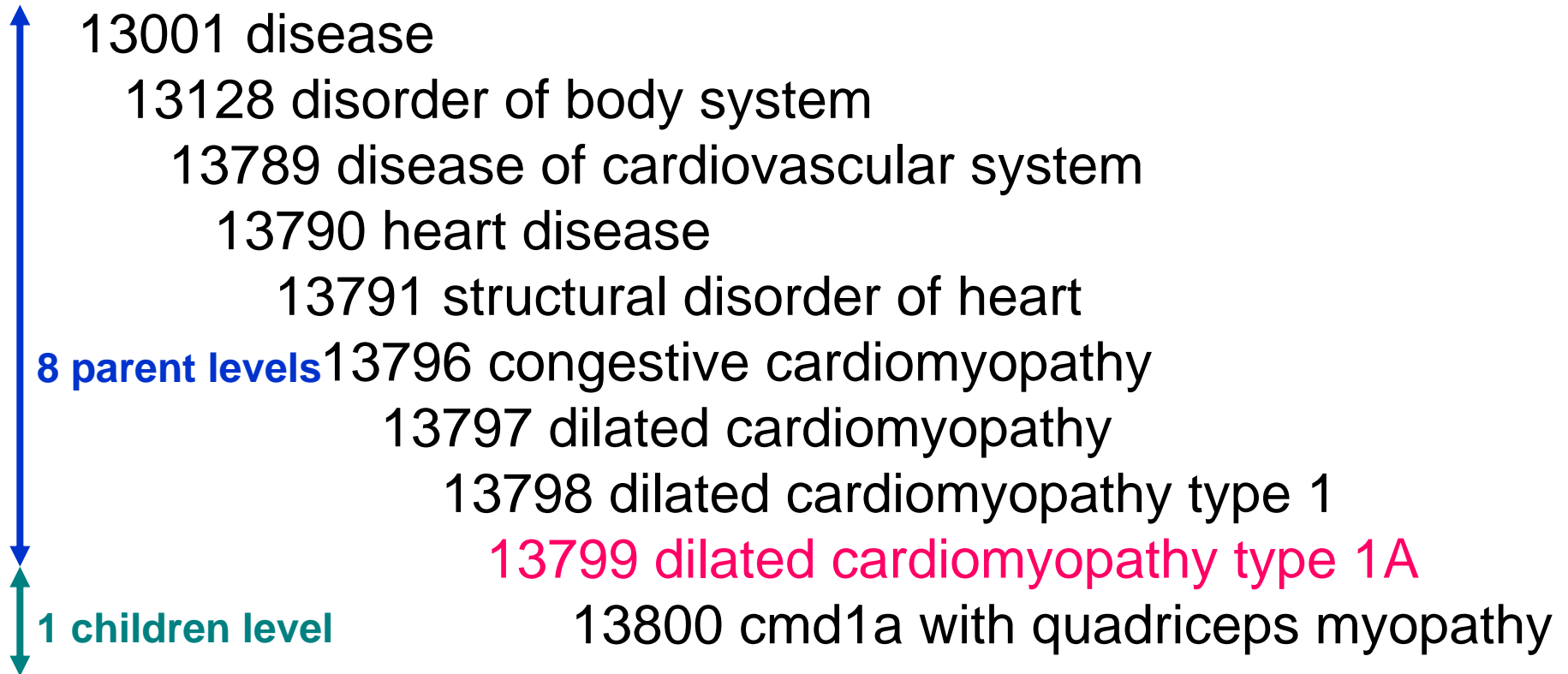
```
graph TD; A[vascular disease] --> B[Disease of Cardiovascular system]; C[heart disease] --> B; D[vascular disease of the skin] --> B;
```

The diagram illustrates the classification of three diseases into a common category. Three red arrows originate from the text labels: 'vascular disease' (from P08235), 'heart disease' (from P02545), and 'vascular disease of the skin' (from P37023). Each arrow points towards the central text 'Disease of Cardiovascular system', which is written in pink. The arrows from 'vascular disease' and 'vascular disease of the skin' are horizontal lines that turn downwards at the end. The arrow from 'heart disease' is a horizontal line that turns upwards at the end.

# SNOMED Clinical Terms®

- Clinical terminology is a structured list of terms for use in clinical practice by healthcare professionals. These terms cover areas such as diseases, operations, treatments, drugs, administrative items, and so on.
- SNOMED CT is the most comprehensive type of clinical terminology.
- SNOMED CT contains a detailed disease ontology at various levels of hierarchy.

# How a particular disease is classified under SNOMED CT?





## **Step 2: Develop ontological model for colon carcinoma**

# What is the need?

- Good data integration exists at the level of genes, RNAs and proteins
- Integration from pathological level of granularity to that of protein level is not well represented (OMIM, LocusLink)
- Need for a formal representation which respects various levels of granularities (Organ system, Organ, Tissue, Cellular, Molecular)
- Need for a integration of knowledge present within various life-sciences database
- Need for developing association rules
- Need for knowledge discovery for localization of processes and functions to the anatomical entities

# Disease Representation

- Disease classification based on Snomed CT
- Various aspects considered for classification (currently present in the form of multiple inheritance)
- Added from textbooks (deVita Principles of Oncology and Harrison Principles of Internal medicine)
  - Staging of diseases (TNM, Duke's, Modified Asler-Coller) :  
Based on Tumor size, Tumor type, Lymph node extent, Metastasis)
  - Screening (Patients screened based on their level of risk)
  - Risk factors (Pathological predisposing factors, Chemicals, Conditions running in families, Age)

# Disease classification

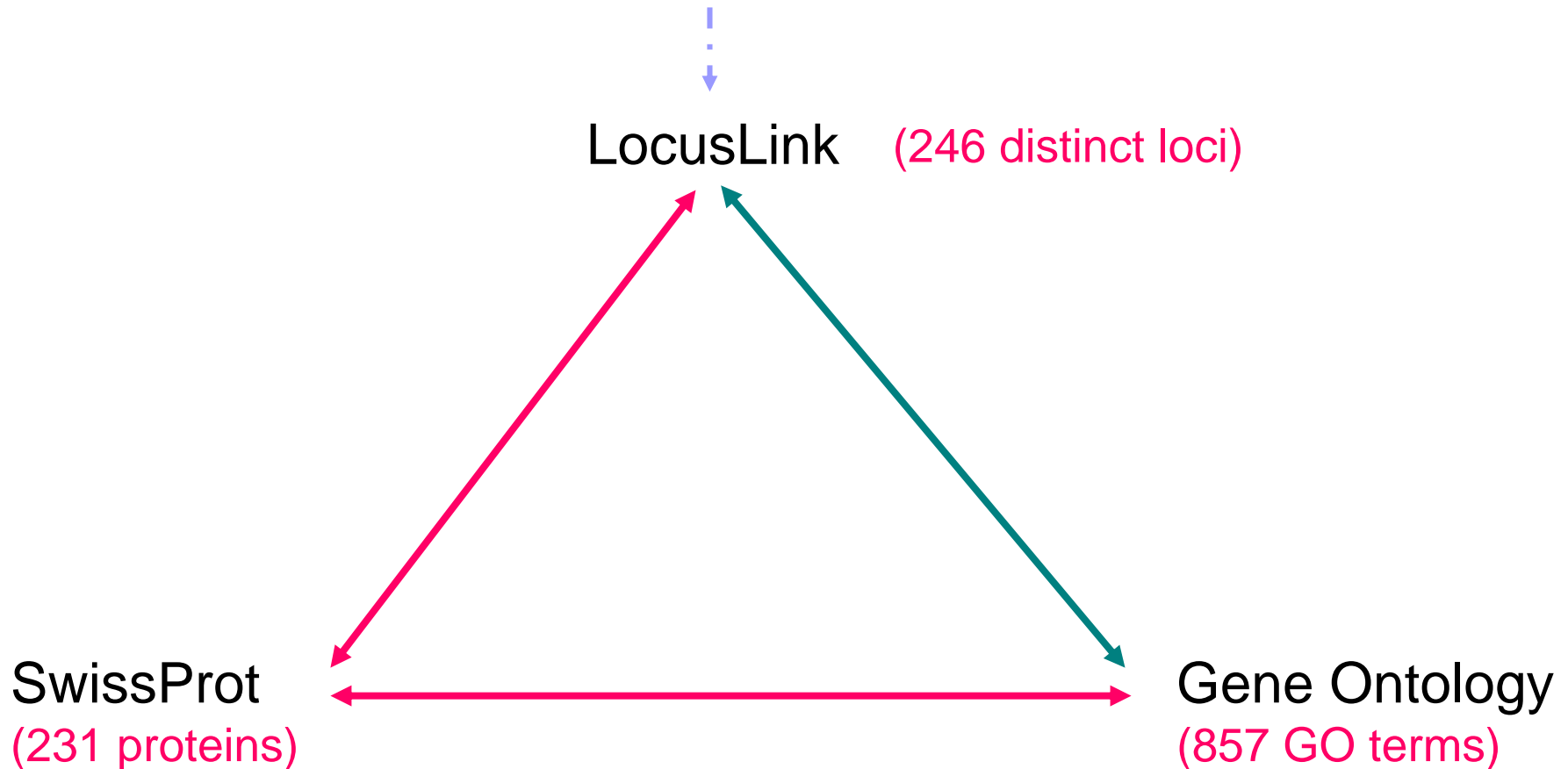
- Localization (Localization based on Organ system affected, Organ affected, Tissues involved, Pathologies, Causative)
- Pathology (Gross and Microscopic pathology, Size, Shape, Extent, Vessels involved, Nerves involved, Cell type, Nuclear characteristics, Staining etc.)
- To be added: Pharmacotherapeutics, Symptoms and Signs

# Anatomical and Histological Representation

- Anatomy of colon represented at Organ system, Organ, Tissue, Cell and Subcellular levels of granularity (Foundational Model of Anatomy)
- Gene Ontology's Cellular Component axis situated within the FMA axis
- Gross pathology mapped to the Carcinoma location
- Anatomical entities traced till Organ system level, in order to make the representation extendable and futurely compliant with different patho-physiological effects of the carcinoma
- Information regarding Clinical procedures, Carcinoma extent, Vascular invasion, Histological pathology being added
- Extensions being done to add relations like is-located-in, is-surrounded-by, etc. to make the anatomical representation deducible
- [Colon](#)

# Interlink between disease, LocusLink, Swiss-Prot, GO annotations

Colon cancer/Colon carcinoma



# Association rules based on GO annotations

- Association rules found considering the Gene Ontology annotations of SWISSPROT proteins
- Gene Ontology consists of three axes:
  - Cellular Component
  - Molecular Function
  - Biological Processes
- SWISSPROT contains over 20,000 annotations within GO originating from the human proteins; over 800 of which are from the proteins which are involved in the patho-physiology of Carcinoma of Colon
- Association between GO terms were established on the basis of these annotations
  - Statistical approach
  - Apriori-algorithm based approach
  - Dependency relations based on POS tagger

# Statistical approach

- GO terms are separated into three separate tables, depending on which of the three orthogonal axes they belong to
- The annotations which related a SWISSPROT protein to the terms within GO's cellular component axis were separated from those to terms within GO's molecular function axis and biological process axis
- All the GO terms which belong to two different axes but are annotated to the same protein are put together and three new tables are created: those annotations where – cc and mf terms, mf and bp terms, cc and bp terms, are annotated together.
- The distinct tuples present are grouped together and their count provides the weight as to how many times two GO terms from two distinct axes are annotated together

# Apriori-algorithm based approach

- Association rule induction originally developed for market basket analysis
- Aims at finding regularities in the shopping behaviour of customers
- With the induction of association rules one tries to find sets of products that are frequently bought together.
- This information is expressed in the form of rules like "If a customer buys bread and wine, he/she is likely to buy cheese, too."
- Algorithms for inducing association rules from a set of transactions (the "market baskets" or "shopping carts" bought by customers) usually work in two steps:
  - Frequent item sets are determined by searching the subset lattice of all items (Two approaches breadth-first and depth-first search)
  - association rules are constructed from the frequent item sets and filtered w.r.t. some quality criterion

# Apriori-algorithm based approach

- Support and Confidence are defined
- ribosome <- ribosome biogenesis; protein biosynthesis (0.2%, 93.2%)
- This rule says that there are 0.2% of the total annotations, put together ribosome biogenesis and protein biosynthesis, of which 93.2% (i.e. 82) are also annotated with the term ribosome.
- Formal ontological relations, already formalized in predicate logic, are applied between the entities, which would help to have deductions: has spatial projection, processual part of, facilitates, mediates, perpetrates
- Localization of Molecular functions to Cellular components
- Localization of Biological processes to Cellular components
- Composition of Biological Processes of Molecular Functions

# Association rules bases on GO annotations

- The association rule induction helps in:
  - Finding the links across GO axes
  - Finding missing links based on class-subclass relations
  - Pathway formation
  - Knowledge discovery in the form that the missing annotations can be found. For example, that between p53 and “transcription from Pol II promoter”; “induction of apoptosis by hormones” and “integral to plasma membrane
- Annotations lead to Proteins, Proteins lead to SWISSPROT and ....



# Application Examples

# Using the ontology model to answer some basic questions

- Clinical questions:

What are the risk factors involved in colon carcinoma?

Are there any molecular markers for diagnosis and/or prognosis?

# Using the ontology model to answer some basic questions

## ■ Clinical questions:

What are the risk factors involved in colon carcinoma?

Are there any molecular markers for diagnosis and/or prognosis?


## ■ Biological questions:

If p53 is a molecular marker, what's its role in the cell? What are the biological processes it takes part in?

Are there any colon-cancer related proteins that take part in these processes?

- [Case of p53](#)

- [Case of IGF-1A \[precursor\]](#)



# Using the ontology model to answer some basic questions

- Pioneering questions:

Is it possible to find new knowledge or derive new hypothesis from the ontology model?



## **Step 3: Formalisation**

## Formalization

- $A \text{ is-a } B \rightarrow \forall x(\text{inst}(x,A) \rightarrow \text{inst}(x,B))$
- $A \text{ infects } B \rightarrow \exists x\exists y (\text{inst}(x,A) \ \& \ \text{inst}(y,B) \ \& \ \text{infects}(x,y))$
- $A \text{ part-of } B =_{\text{def}} \forall x (\text{inst}(x,A) \rightarrow \exists y((\text{inst}(y,B) \ \& \ \text{part}(x,y))))$
- $\forall y((\text{inst}(y, B) \rightarrow \exists x((\text{inst}(x,A) \ \& \ \text{part}(x, y))))$
- $(\text{inst}(x, \text{colon}, t_1) \ \& \ \text{inst}(x, \text{colon}, t_2)) \rightarrow \forall t (t_1 \leq t \leq t_2) \rightarrow \text{inst}(x, \text{colon}, t)$
- $\forall x\forall y \text{ inst}(x, \text{mucosa of colon}, t) \rightarrow (\exists y (y, \text{colon}, t) \ \& \ \text{part}(y, x, t))$

# Formalization

- *organ system:* *digestive system, respiratory system, nervous system, ...*
- *organ:* *pharynx, esophagus, stomach, colon, ...*
- *organ part:* *mucosa of colon, submucosa of colon, ...*
- *tissue:* *epithelium of mucosa of colon, ...*
- *collection of cells:* *red blood cells within a test tube*
- *cell:* *colon epithelial cell, fibrocytes, ...*
- *subcellular:* *colon epithelial cell nucleus, colon epithelial cell membrane, ...*
- $\text{gr}(\text{digestive system}) = \text{organ system}$ ,  $\text{gr}(\text{colon}) = \text{organ}$ ,  $\text{gr}(\text{mucosa of colon}) = \text{tissue}$

# Formalization

## ■ TNM Classification

- T1: The tumor invades the submucosa, the second layer of the large intestine.
- T2: The tumor invades the muscularis propria.

■ **inst** (x, A) → **inst**(x, A) & gr(x) = level1; where gr stands for granularity

■ **inst** (x, B) → **inst**(x, B) & gr(x) = level2; where gr stands for granularity

■ **inst** (x, C) → **inst**(x, C) & gr(x) = level2; where gr stands for granularity

# Formalization

- $A \text{ part-of } B \rightarrow (\text{inst}(x, A) \ \& \ \text{inst}(y, B) \ \& \ \text{part}(x,y) \ \& \ \text{gr}(x) = \text{level1} \ \& \ \text{gr}(y) = \text{level2})$
- $(A \text{ part-of } B) \ \& \ (B \text{ part-of } C) =\text{def} (\text{inst}(x, A) \ \& \ \text{inst}(y, B) \ \& \ \text{inst}(z, C) \ \& \ \text{part}(x,y) \ \& \ \text{part}(y,z) \ \& \ \text{gr}(x) = \text{level1} \ \& \ \&\text{gr}(y) = \text{level2} \ \& \ \text{gr}(z) = \text{level3})$
- $(A \text{ part-of } B) \ \& \ (B \text{ part-of } C) \rightarrow (A \text{ part-of } C)$

# Formalization

- zooming action is an occurrent but  $\text{zoom}(\text{level1}, \text{level2})$  not a part of the ontology tree. It is an epistemology in order to describe reality.
- $\text{zoom}(\text{level1}, \text{level2}) \rightarrow (\text{inst}(x, A) \ \& \ \text{inst}(y, B) \ \& \ \text{part}(x,y) \ \& \ \text{gr}(x) = \text{level1} \ \& \ \text{gr}(y) = \text{level2} \ \& \ \text{description}(x, t1) \ \& \ \text{description}(y, t2) \ \& \ t1 > t2)$

## Formalization

- **inst** ( $x$ , T1-stage carcinoma of colon structure)  $=_{\text{def}}$   $\forall x \exists p \exists q \exists r \exists s$  (**inst** ( $p$ , carcinoma of mucosa of colon structure) & **inst** ( $q$ , carcinoma of submucosa of colon structure) & **inst** ( $r$ , carcinoma of muscularis layer of colon structure) & **inst** ( $s$ , carcinoma of serosa of colon structure) & has-anatomical-extent ( $x$ , ( $p, q$ )))
- **inst** ( $x$ , Grade 2 carcinomatous pathology of mucosa structure)  $=_{\text{def}}$   $\forall x \exists y$  (**inst** ( $x$ , carcinomatous pathology) & **inst** ( $y$ , non-polar nucleus of mucosal epithelium) & has-pathological-feature ( $x, y$ ))

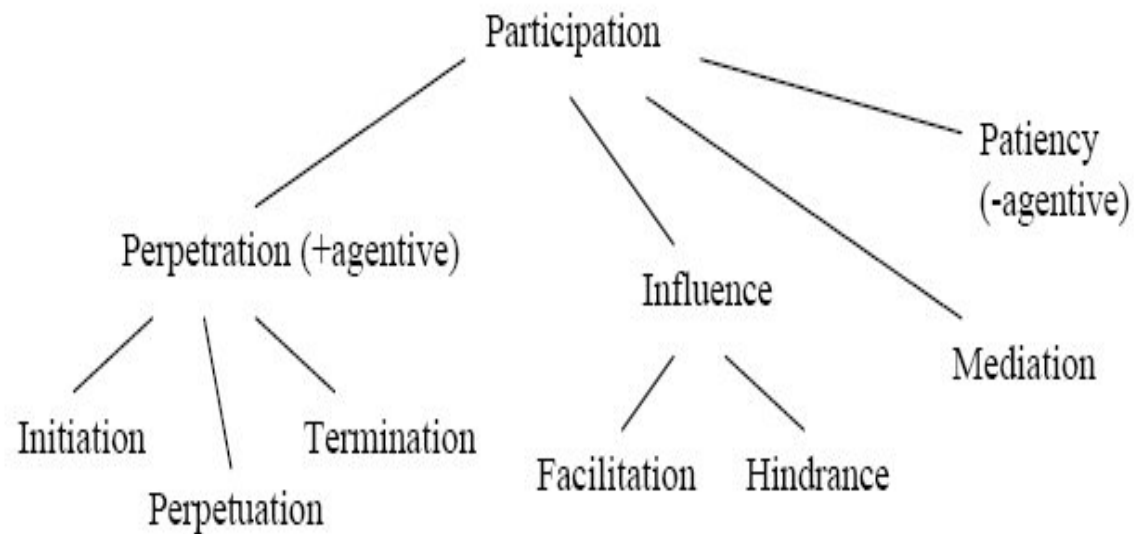
# Formalization

- **inst** ( $x$ , T1-stage carcinomatous process of carcinoma of colon) =<sub>def</sub>  
 $\forall x \exists p \exists q \exists r \exists s \exists t$  (<sub>gr=tissue</sub> **inst**<sub>gr=organ</sub> ( $x$ , carcinomatous process) & **inst**( $p$ ,  
mucosa of colon) & **inst** ( $q$ , submucosa of colon) & **inst** ( $r$ ,  
muscularis layer of colon) & **inst** ( $s$ , serosa of colon) & **inst** ( $t$ , colon)  
& has-anatomical-projection ( $x, (p, q)$ ) & part ( $p, t$ ) & part ( $q, t$ ) & part  
( $r, t$ ) & part ( $s, t$ ))

# Formalization

- A involved-in B =def  $\forall x \exists y (inst(x, A) \ \& \ inst(y, B) \ \& \ part(x,y))$
- B involving A =def  $\forall x \exists y (inst(x, C) \ \& \ inst(y, D) \ \& \ part(y,x))$
- (A involved-in B) part of (B involving A)
- No claim that B has only one part A
- A involved-in B is (involvement specification) of A
- B involving A is (involvement specification) of A
- A is A occurrent
- B is A occurrent
- Will need a granularity specification (probably to the anatomical axis)

# Formalization



# Formalization

- *electron transporter* initiates *electron transport* =<sub>def</sub>  $\forall x(\text{inst}(x, \text{electron transporter}) \rightarrow \exists y(\text{inst}(y, \text{electron transport})) \ \& \ \mathbf{initiates}(x,y,t))$
- $\mathbf{initiates}(x,y,t) \rightarrow \text{perpetrates}(x,y,t) \ \& \ \text{not}(\text{origin}(y,c) \ \& \ \text{before}(c,\text{start}(t)))$
- $\text{perpetrates}(x,y,z) \ \& \ \text{exists}(x,a) \ \& \ \text{occurs}(y,b) \rightarrow (\text{temporal-part}(z,a) \ \& \ \text{temporal-part}(z,b))$