Bio-REGNET

Retrieval of Patent Documents from Heterogeneous Sources using Ontologies and Similarity Analysis

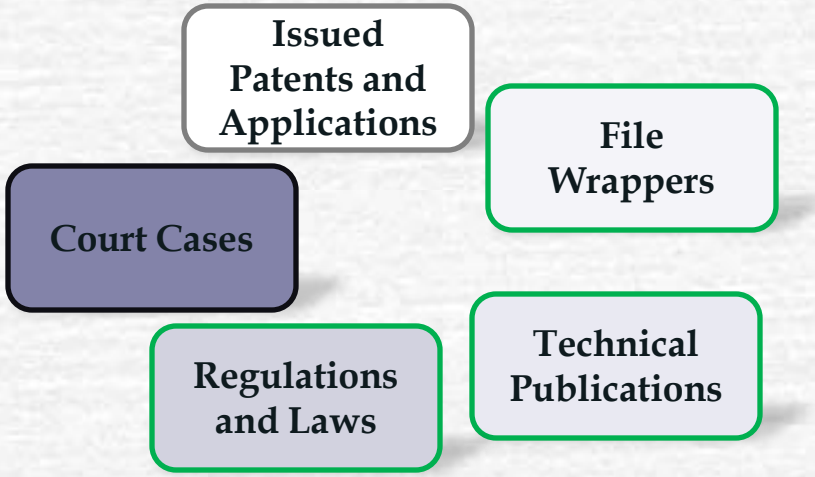
Siddharth Taduri, Gloria T. Lau, Kincho H. Law
Engineering Informatics Lab, Stanford University

Jay P. Kesan,
School of Law, University of Illinois Urbana-Champaign

09/21/2011

International Conference on Semantic Computing

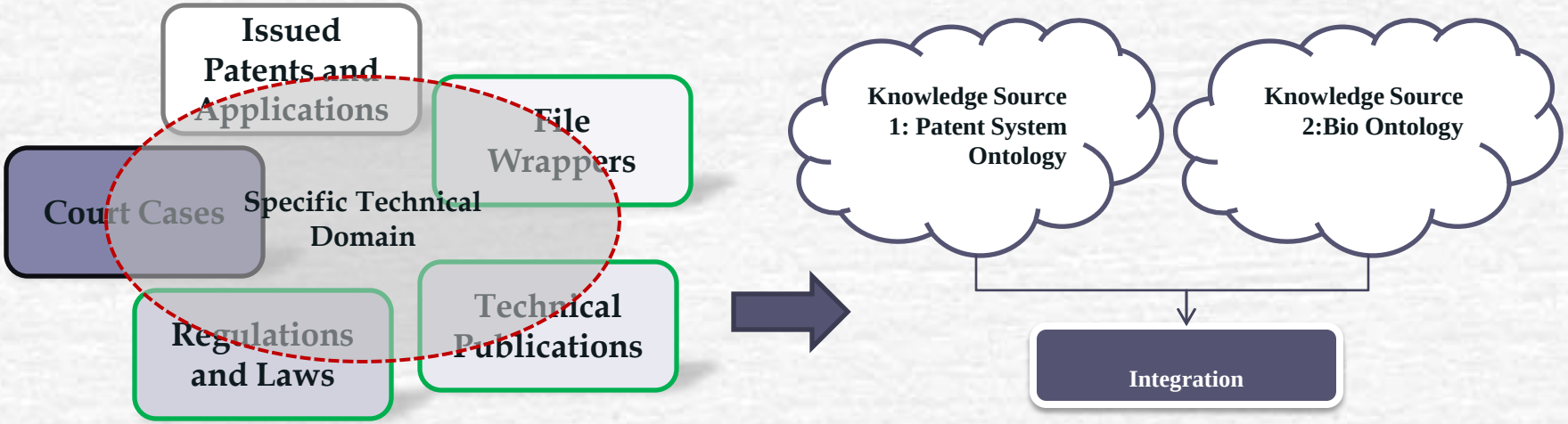
Problem Statement



- Patent Validity and Enforcement Questions involves analysis of documents in various domains – World-wide Patents, PTO File Wrappers, Scientific Publications and Court documents

- The information is siloed into several diverse information sources

Problem Statement



➢The sources are diverse in structure, formats, semantics and syntax

• How to develop and retrieve comprehensive knowledge of patents in a particular technological space?

Court Cases

927 F.2d 1200 (1991)

AMGEN, INC., Plaintiff/Cross-Appellant,
v.
CHUGAI PHARMACEUTICAL CO., LTD., and Genetics Institute, Inc., Defendants-Appellants.

Nos. 90-1273, 90-1275.
United States Court of Appeals, Federal Circuit.

March 5, 1991.

Suggestion for Rehearing Declined May 20, 1991.

…

…

Before MARKEY, LOURIE and CLEVENGER, Circuit Judges.

…

THE PATENTS

On June 30, 1987, the United States Patent and Trademark Office (PTO) issued to Dr. Rodney Hewick U.S. Patent 4,677,195, entitled "Method for the Purification of Erythropoietin and Erythropoietin Compositions" (the '195 patent). The patent claims both homogeneous EPO and compositions thereof and a method for purifying human EPO using reverse phase high performance liquid chromatography. The method claims are not before us. The relevant claims of the '195 patent are:

1. Homogeneous erythropoietin characterized by a molecular weight of about 34,000 daltons on SDS PAGE, movement as a single peak on reverse phase high performance liquid chromatography and a specific activity of at least 160,000 IU per absorbance unit at 280 nanometers.

* * * * * *

3. A pharmaceutical composition for the treatment of anemia comprising a therapeutically effective amount of the homogeneous erythropoietin of claim 1 in a pharmaceutically acceptable vehicle.
4. Homogeneous erythropoietin characterized by a molecular weight of about 34,000 daltons on SDS PAGE, movement as a single peak on reverse phase high performance liquid chromatography and a specific activity of at least about 160,000 IU per absorbance unit at 280 nanometers.

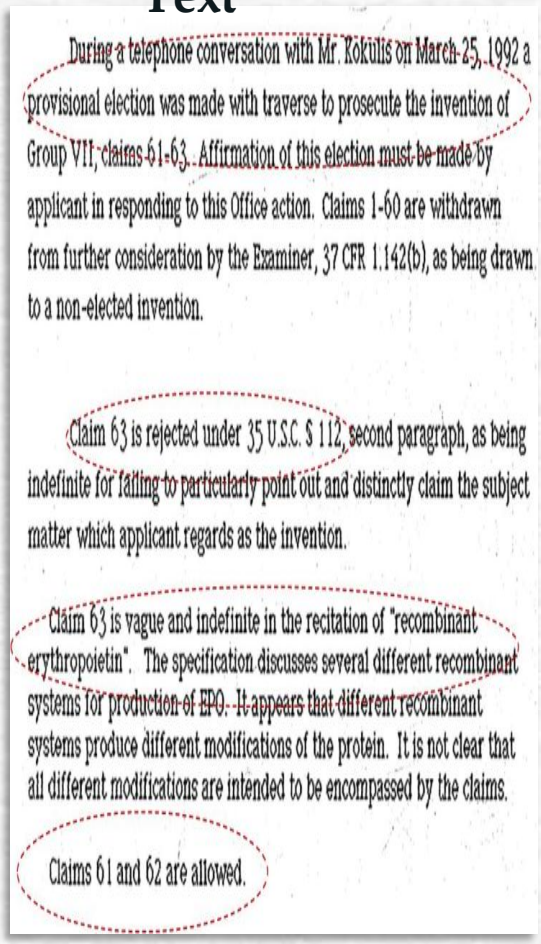
- Court Cases are not very well structured!
- Comparatively more difficult to parse information
- PACER – an electronic system to access databases for U.S. Courts – requires one to know party/assignee name, case number/type, etc. which may not be known

Patent File Wrappers

Events

- 1 - Application
  - Examples
  - Claims
  - Abstract
  - Declaration for Patent Application
  - Drawings
  - Application Transmittal
- 1.5 - Preliminary Amendment A
- 2 - Preliminary Amendment AA
  - Remarks
- Status Request
- 3 - Interview Summary
- 4 - Rejection
- 5 - Title Report
- 6 - Change of Address
- 7 - Extension of Time
- 8 - Interference Letter
- 9 - Interference Digest
- Miscellaneous Documents

Text



During a telephone conversation with Mr. Rokulis on March 25, 1992 a provisional election was made with traverse to prosecute the invention of Group VII, claims 61-63. Affirmation of this election must be made by applicant in responding to this Office action. Claims 1-60 are withdrawn from further consideration by the Examiner, 37 CFR 1.142(b), as being drawn to a non-elected invention.

Claim 63 is rejected under 35 U.S.C. § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

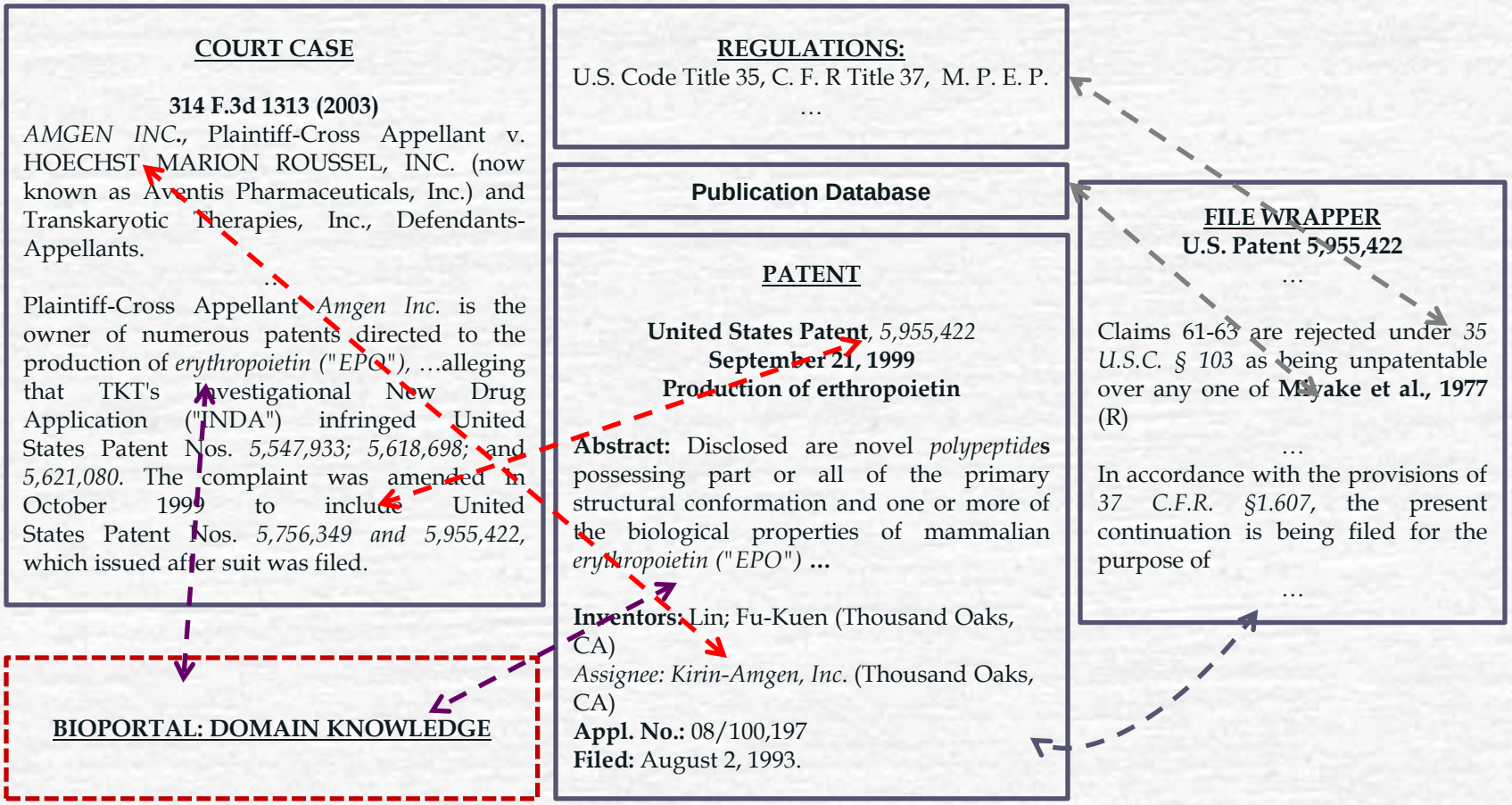
Claim 63 is vague and indefinite in the recitation of "recombinant erythropoietin". The specification discusses several different recombinant systems for production of EPO. It appears that different recombinant systems produce different modifications of the protein. It is not clear that all different modifications are intended to be encompassed by the claims.

Claims 61 and 62 are allowed.

- File Wrappers are folders which contain all documents exchanged between a patent applicant and the patent office
- Every File Wrapper is different! No standardized ordering of events
- The relevant information is embed within lots of irrelevant text
- File Wrappers are available as images requiring additional processing in order to extract text

Cross-Referencing

- There are many aspects of these documents which can be utilized; especially the cross-referencing between the documents



Patent System Ontology

- Established *semantics* allow us to reason over the classes, properties and instances to infer new facts

- *Documents* can be *connected* to form a *network* similar to citation networks. Only now we have not just citations, but other metadata such as co-inventorships, technological classification and other cross-domain relevancy metrics between documents (ex: patents occurring in court cases etc.)

- Can develop *rules* to perform additional inferences over the knowledge

Example Query

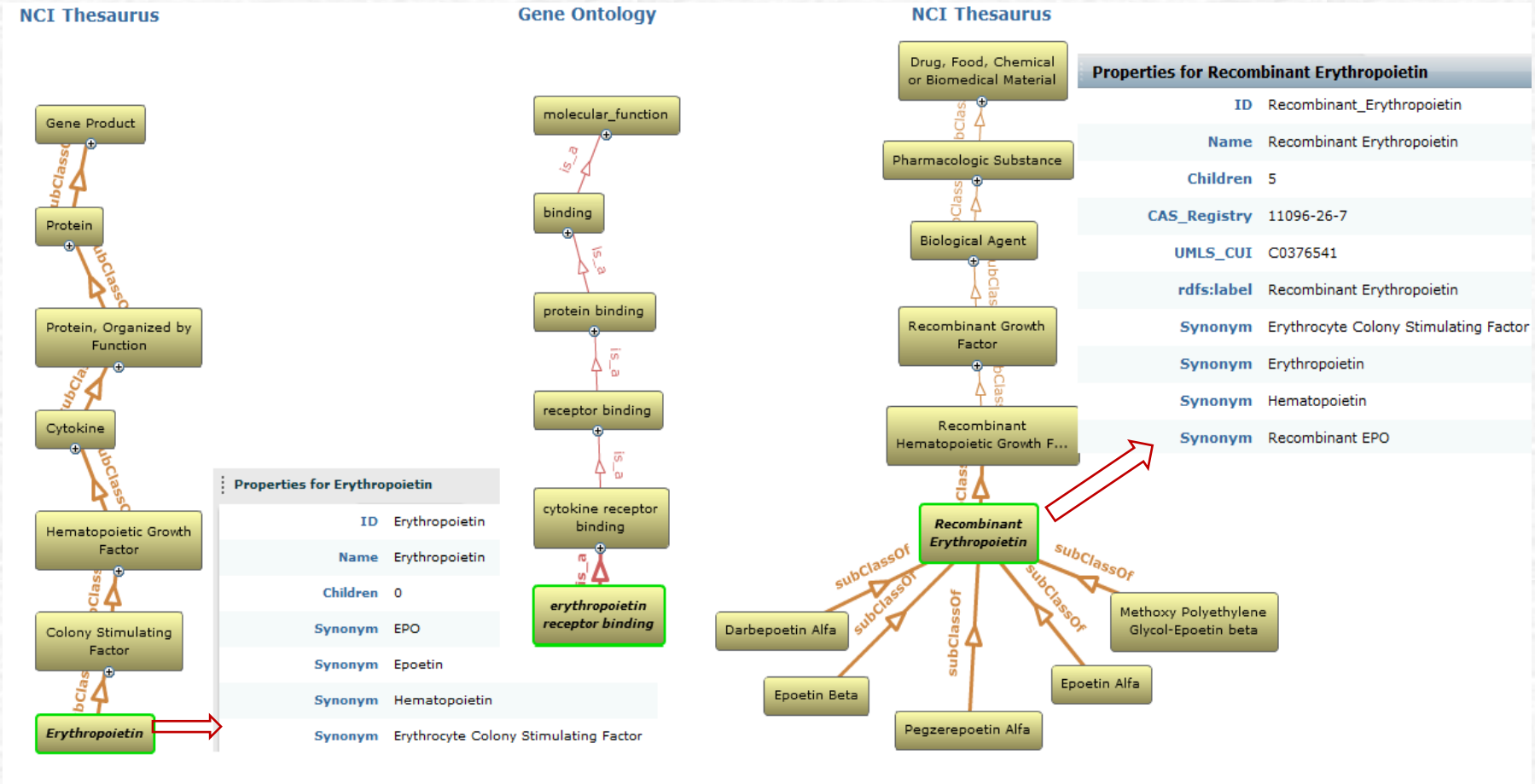
- Return all the patent documents which contain the keyword "erythropoietin" in the Claims and Assigned to "Amgen_Inc". What technology classes do these patent documents belong to?

- SPARQL Query:

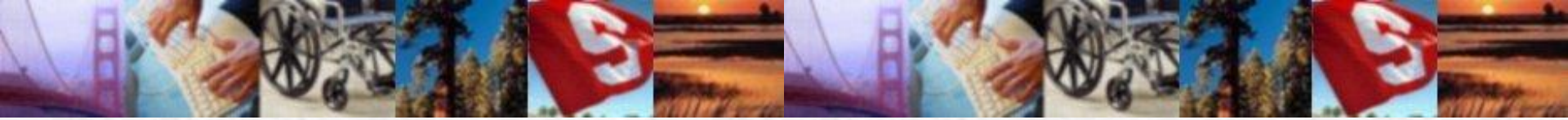
```
SELECT DISTINCT ?patent ?inventor
FROM <http://localhost:8890/PatentOntologyInferred>
WHERE{
        ?patent a ont:Patent .
        ?patent ont:hasAbstract ?abs .
        ?abs ont:resourceVal ?val .
        ?val bif:contains "erythropoietin" .

        ?patent ont:hasAssignee ont:Amgen_Inc .

        ?patent ont:hasInventor ?inventor
}
 Limit 10
```

Patent	Inventor
5856298	Strickland_Thomas_W
5885574	Elliott_Steven_G
7304150	Egrie_Joan_C
7304150	Elliott_Steven_G
7304150	Browne_Jeffrey_K
7304150	Sitney_Karen_C
7217689	Elliott_Steven_G
7217689	Byrne_Thomas_E
6319499	Elliott_Steven_G
5756349	Lin_Fu-Kuen

Domain (Bio) Ontologies



- Bio Ontologies serve as terminological standards in the domain



Expanded Query

Original Term: Erythropoietin

Synonyms: Erythropoietin, Recombinant Erythropoietin, erythropoietin receptor binding, Hematopoietin, Recombinant EPO, Erythrocyte Colony Stimulating Factor, Epoetin, EPO …

Children: Darbopoietin Alfa, Epoetin Alfa, Epoetin Beta …

Parents: Colony Stimulating Factors, cytokine receptor binding, recombinant hematopoietic growth factors…

Grand-Parents: hematopoietic growth factor, receptor binding, recombinant growth factor …

- An appropriate ranking function is to be applied to balance the more general terms. Heuristically, we assign a higher weight to synonyms, and a lower weight as we traverse away from the concept node

- Resulting Query: *"original term"* OR [*synonyms*]^weight OR [*children*]^weight OR ….

Use-Case: Erythropoietin

Current Corpus: experimental platform to test the overall effectiveness of the framework

- 5 Core patents – U.S. Patents 5,621,080, 5,756,349, 5,955,422, 5,547,933, 5,618,698
- 135 directly related patents (through citations) form our gold standard for computing formal measures such as Precision and Recall
- Total patent corpus of 1150 patents
- Identified over related 3000 publications through citations. These are available on PubMed and can be accessed through Entrez – A tool that provides a search interface to PubMed database
- Around 30 court cases, patent litigation involving major companies including Amgen, Hoechst Marion Roussel, Inc., Transkaryotic Therapies, Inc.
- BioPortal (http://bioportal.bioontology.org) is a comprehensive source of domain knowledge

Patent Ontology Stats

- 54 Classes, 40 Properties and over 15,000 individuals from 1150 patents, 30 court cases and one partially instantiated file wrapper

- Used Protégé-OWL to edit the ontology and Protégé-OWL/Jena API to programmatically instantiate physical documents

- Can query using any SPARQL endpoint such as Protégé or Virtuoso's Triple Store

- SWRL is used to declare rules. We use the Jess rule execution engine

Methodology

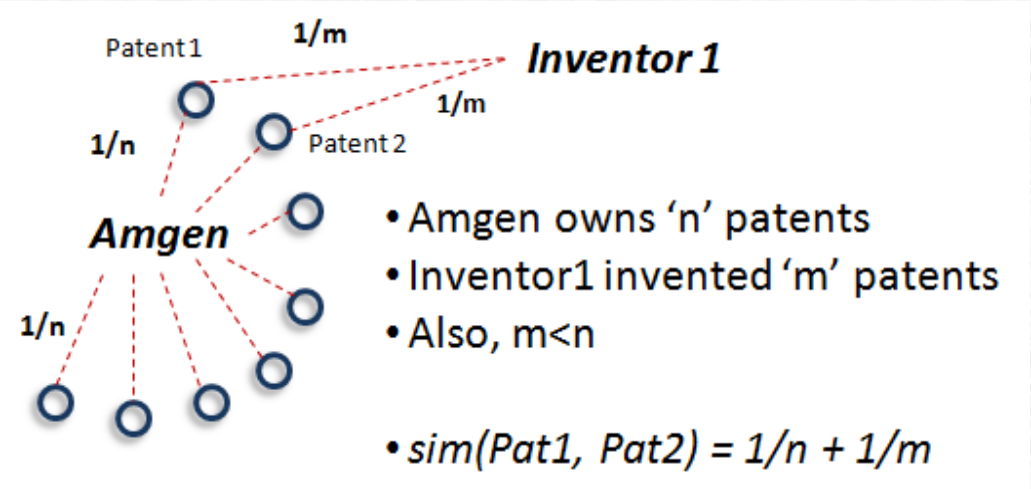
- The cross-references between document types and metadata of documents in the patent system are utilized through a rule-based system

- Structural dependencies between types of documents must be considered

- The application of bio-ontologies to each type of document is different due to the depth of technical terminology. This is controlled through the weighting vector

- Based upon an initial selection of documents by the user, we perform a similarity analysis between documents [User Relevancy Feedback]

Rules



- The declarative representation of the patent system ontology can facilitate reasoning through rules

- Different users may be interested in different aspects of the document (Users can use their own heuristics)

- The methodology allows users to select which rules apply during search

Rules

- Two patents share the same inventor:

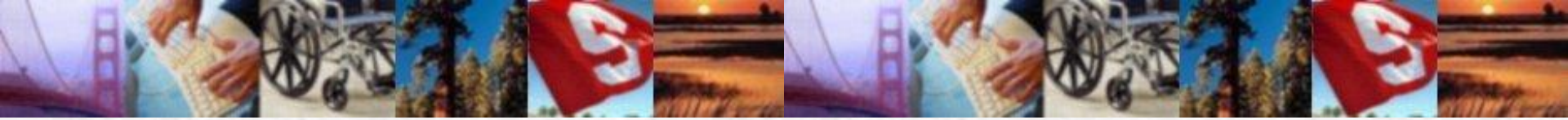
  IF hasInventor (?pat1, ?inv1) ^ hasInventor (?pat2, ?inv1) ^ owlDifferentFrom (?pat1, ?pat2) → hasSimilarDocument(?pat1, ?pat2)

- Same court case cites two different patents:

  IF patentsInvolved(?case, ?pat1) ^ patentsInvolved (?case, ?pat2) ^ owlDifferentFrom (?pat1, pat2) → hasSimilarDocument(?pat1, ?pat2)

- Rules are combined by using:

$$Sim(A,B) = \sum_{i=1}^{\#\ of\ Rules} Wi * inference(i)$$



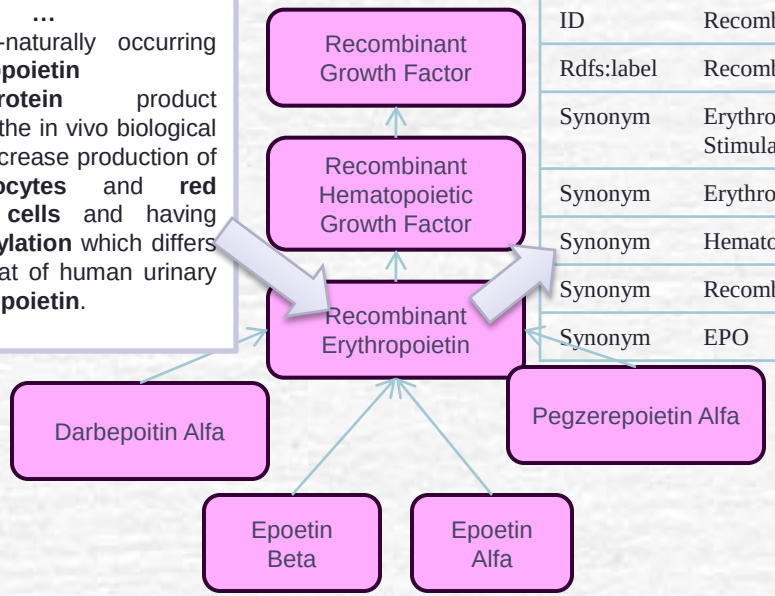
Text

Patent 5,547,933
…
Claims
…

1. A non-naturally occurring **erythropoietin glycoprotein** product having the in vivo biological … to increase production of **reticulocytes** and **red blood cells** and having **glycosylation** which differs from that of human urinary **erythropoietin**.

…

NCI Thesaurus



Properties for Recombinant Erythropoietin	
ID	Recombinant Erythropoietin
Rdfs:label	Recombinant Erythropoietin
Synonym	Erythrocyte Colony Stimulating Factor
Synonym	Erythropoietin
Synonym	Hematopoietin
Synonym	Recombinant EPO
Synonym	EPO

→ subClassOf property

W_{Pat} Weight Vector for Patents

W_{Case} Weight Vector for Cases

M Expanded Terms

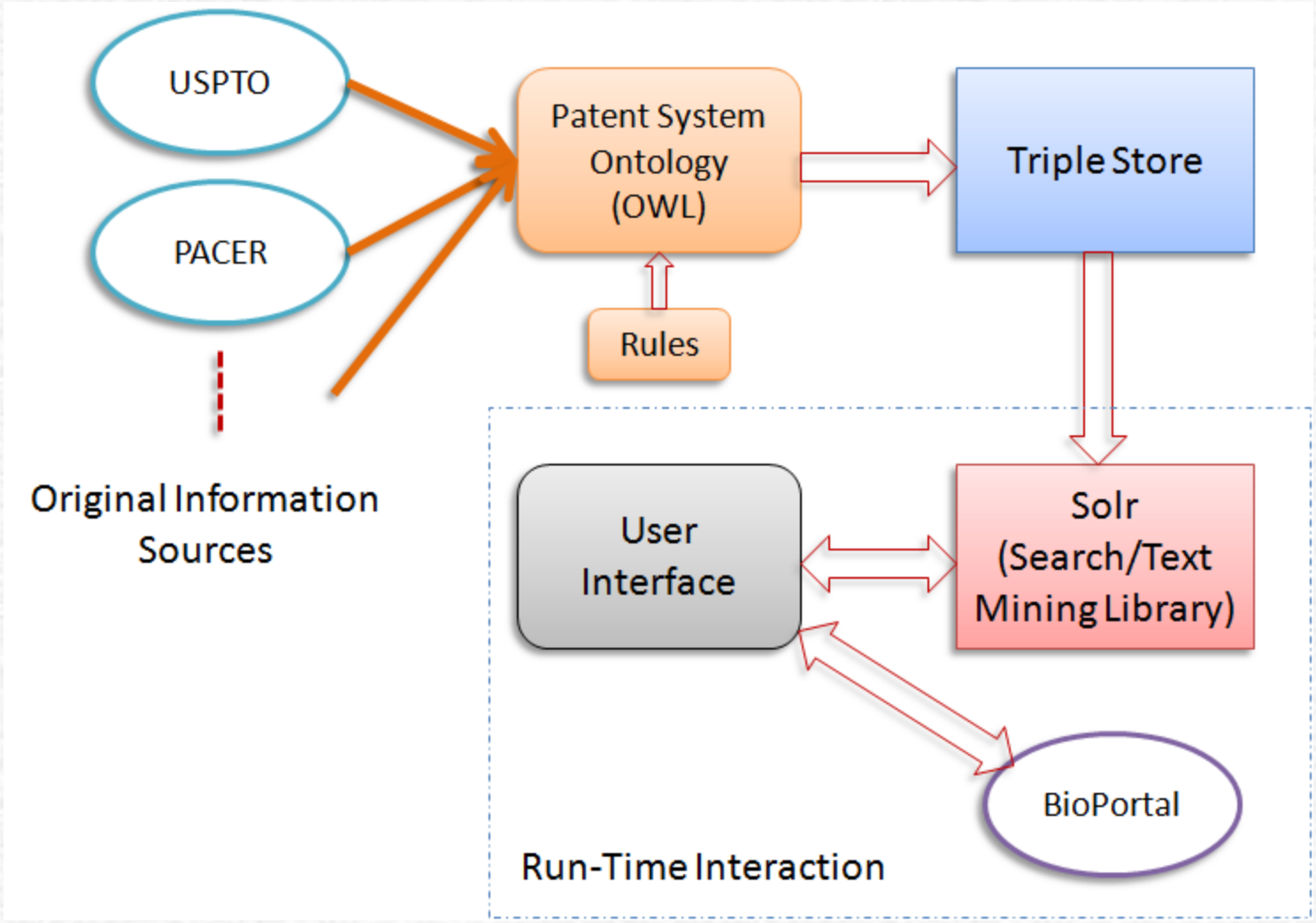
$$M = \begin{bmatrix} \{\text{non-naturally, erythropoietin, glycoprotein, biological, …, reticulocytes, glycosylation …}\} \\ \{\text{recombinant erythropoietin, epo, recombinant epo, hematopoietin, erythrocyte colony stimulating factor}\} \\ \{\text{recombinant hematopoietic growth factor}\} \\ \{\text{recombinant growth factor}\} \\ \{\text{darbapoietin alfa, epoetin beta, epoetin alfa…}\} \end{bmatrix}$$

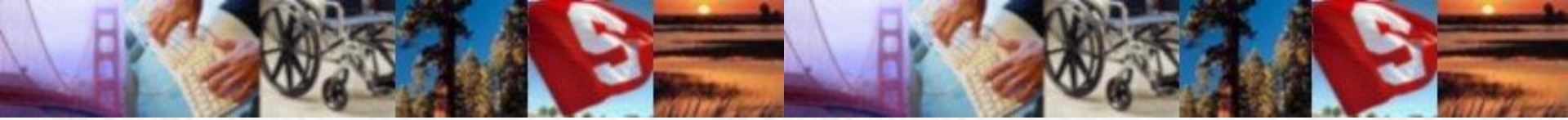
$$W_{Pat} = \begin{bmatrix} 1 \\ 0.5 \\ 0.2 \\ 0.2 \\ 0.1 \end{bmatrix} \quad W_{Case} = \begin{bmatrix} 1 \\ 0.5 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Generated Query:

$$Q_{Patent} = W_{Pat}^{T} * M$$
$$Q_{Case} = W_{Case}^{T} * M$$

Implementation





Result – Structural Dependency

Patent 5,955,422: Production of erythropoietin

Abstract

Disclosed are novel polypeptides possessing part or all of the primary structural conformation and one or more of the biological properties of mammalian erythropoietin ("EPO") which are characterized in preferred forms by being the product of procaryotic or eucaryotic host expression of an exogenous DNA sequence. Illustratively, genomic DNA, cDNA and manufactured DNA sequences coding for part or all of the sequence of amino acid residues of EPO or for analogs thereof are incorporated into autonomously replicating plasmid or viral vectors employed to transform or transfect suitable procaryotic or eucaryotic host cells such as bacteria, yeast or vertebrate cells in culture. Upon isolation from culture media or cellular lysates or fragments, products of expression of the DNA sequences display, e.g., the immunological properties and in vitro and in vivo biological activities of EPO of human or monkey species origins. Disclosed also are chemically synthesized polypeptides sharing the biochemical and immunological properties of EPO. Also disclosed are improved methods for the detection of specific single stranded polynucleotides in a heterologous cellular or viral sample prepared from, e.g., DNA present in a plasmid or viral-borne cDNA or genomic DNA "library".

Claims

What is claimed is:

1. A non-naturally occurring erythropoietin glycoprotein product having the in vivo biological activity of causing bone marrow cells to increase production of reticulocytes and red blood cells and having glycosylation which differs from that of human urinary erythropoietin.

2. The non-naturally occurring EPO glycoprotein product according to claim 1 wherein said product has a higher molecular weight than human urinary EPO as measured by SDS-PAGE.

9. A pharmaceutical composition comprising an effective amount a glycoprotein product effective for erythropoietin therapy according to claim 1, 2, 3, 4, 5 or 6 and a pharmaceutically acceptable diluent, adjuvant or carrier.

Patent 5,756,349: Production of erythropoietin

Abstract

Disclosed are novel polypeptides possessing part or all of the primary structural conformation and one or more of the biological properties of mammalian erythropoietin ("EPO") which are characterized in preferred forms by being the product of procaryotic or eucaryotic host expression of an exogenous DNA sequence. Illustratively, genomic DNA, cDNA and manufactured DNA sequences coding for part or all of the sequence of amino acid residues of EPO or for analogs thereof are incorporated into autonomously replicating plasmid or viral vectors employed to transform or transfect suitable procaryotic or eucaryotic host cells such as bacteria, yeast or vertebrate cells in culture. Upon isolation from culture media or cellular lysates or fragments, products of expression of the DNA sequences display, e.g., the immunological properties and in vitro and in vivo biological activities of EPO of human or monkey species origins. Disclosed also are chemically synthesized polypeptides sharing the biochemical and immunological properties of EPO. Also disclosed are improved methods for the detection of specific single stranded polynucleotides in a heterologous cellular or viral sample prepared from, e.g., DNA present in a plasmid or viralborne cDNA or genomic DNA "library".

COURT CASE

AMGEN INC. v. HOECHST MARION ROUSSEL, INC. and Transkaryotic Therapies, Inc.,

A. The '933 Patent

Amgen asserted the following three claims of the '933 patent against TKT:

1. A non-naturally occurring erythropoietin glycoprotein product having the in vivo biological activity of causing bone marrow cells to increase production of reticulocytes and red blood cells and having glycosylation which differs from that of human urinary erythropoietin.
2. The non-naturally occurring EPO glycoprotein product according to claim 1 wherein said product has a higher molecular weight than human urinary EPO as measured by SDS-PAGE.
9. A pharmaceutical composition comprising an effective amount of a glycoprotein product effective for erythropoietin therapy according to claim 1, 2, 3, 4, 5, or 6 and a pharmaceutically acceptable diluent, adjuvant or carrier.

Result – Combining Rules and Bio-Ontology

Score (Rules) "S1"	Score (BioAnnotation) "S2"	Total = 0.6*S1 + 0.4*S2
0.2	0.36	**0.264**

Patent 5,955,422: Production of erythropoietin

Abstract

Disclosed are novel polypeptides possessing part or all of the primary structural conformation and one or more of the biological properties of mammalian **erythropoietin** ("EPO") which are characterized in preferred forms by being the product of procaryotic or eucaryotic host expression of an exogenous DNA sequence. Illustratively, genomic DNA, cDNA and manufactured DNA sequences coding for part or all of the sequence of amino acid residues of EPO or for analogs thereof are incorporated into autonomously replicating plasmid or viral vectors employed to transform or transfect suitable procaryotic or eucaryotic host cells such as bacteria, yeast or vertebrate cells in culture. Upon isolation from culture media or cellular lysates or fragments, products of expression of the DNA sequences display, e.g., the immunological properties and in vitro and in vivo biological activities of EPO of human or monkey species origins. Disclosed also are chemically synthesized polypeptides sharing the biochemical and immunological properties of EPO. Also disclosed are improved methods for the detection of specific single stranded polynucleotides in a heterologous cellular or viral sample prepared from, e.g., DNA present in a plasmid or viral-borne cDNA or genomic DNA "library".

Patent 4,677,195: Method for the purification of erythropoietin and erythropoietin compositions

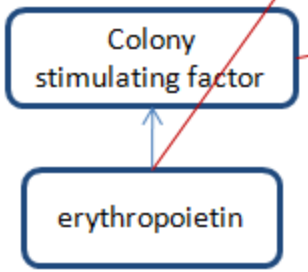
Abstract

A method for purifying erythropoietin is described. The method comprises treating partially purifying erythropoietin by reverse phase high performance liquid chromatography to obtain homogeneous erythropoietin having a molecular weight of about 34,000 daltons on SDS PAGE and moving a single peak on reverse phase HPLC. The homogeneous erythropoietin protein preferably has a specific activity of at least 120,000 IU, more preferably at least 160,000 IU per absorbance unit at 280 nm.

Patent 4,999,291: Production of human pluripotent granulocyte colony-stimulating factor

Abstract

Disclosed are novel polypeptides possessing part or all of the primary structural conformation and one or more of the biological properties of a mammalian (e.g., human) pluripotent granulocyte **colony-stimulating factor** ("hpG-CSF") which are characterized in preferred forms by being the product of procaryotic or eucaryotic host expression of an exogenous DNA sequence. Sequences coding for part or all of the sequence of amino acid residues of hpG-CSF or for analogs thereof may be incorporated into autonomously replicating plasmid or viral vectors employed to transform or transfect suitable procaryotic or eucaryotic host cells such as bacteria, yeast or vertebrate cells in culture. Products of expression of the DNA sequences display, e.g., the physical and immunological properties and in vitro biological activities of isolates of hpG-CSF derived from natural sources. Disclosed also are chemically synthesized polypeptides sharing the biochemical and immunological properties of hpG-CSF.

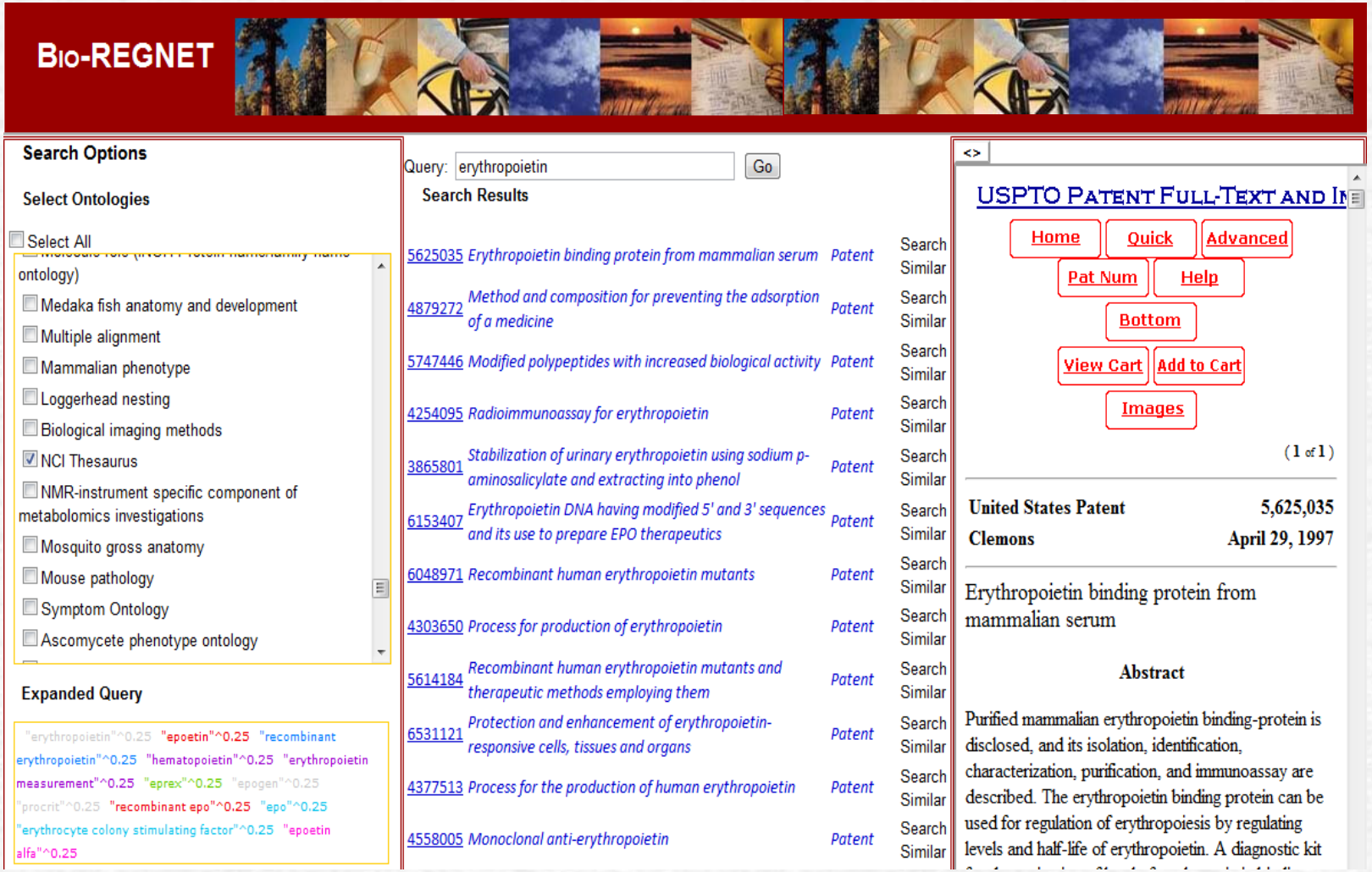


Score (Rules) "S1"	Score (BioAnnotation) "S2"	Total = 0.6*S1 + 0.4*S2
0.0	0.59	**0.236**

Future Work

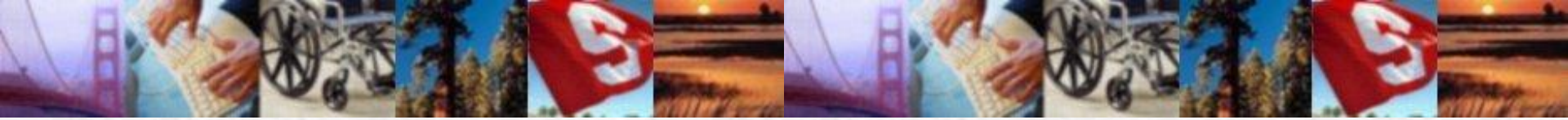
- Formal evaluation is hard due to the unavailability of well defined ground truths, but necessary
- Include other information sources – publications, regulations, laws
- Experiment with more use cases outside of the biomedical domain

Tool Snapshot



Acknowledgement

This research is partially supported by NSF Grant Number IIS-0811975 awarded to the University of Illinois at Urbana-Champaign and NSF Grant Number IIS-0811460 to Stanford University. Any opinions and findings are those of the authors, and do not necessarily reflect the views of the National Science Foundation.



Thank You: Questions?

Engineering Informatics Lab: http://eil.stanford.edu

Contact

Siddharth Taduri: staduri@stanford.edu
Gloria T. Lau: glau@stanford.edu
Kincho H. Law: law@stanford.edu
Jay P. Kesan: kesan@illinois.edu