Big Data for Government Symposium

http://www.ttcus.com

@TECHTrain

TTC™
Technology Training Corporation

Linkedin/Groups:
Technology Training Corporation

Effectively Exploiting Big Data with Semantics: A Pilot Project

Thomas C. Rindflesch, National Library of Medicine

Fredrik Salvesen, YarcData

Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.

Background

- Basic biomedical research is crucial to medicine
  - Complexity of molecular pathophysiology
  - Challenges development of new therapies
- Big data can facilitate progress
  - MEDLINE: Biomedical research literature
- Need effective, automatic access to information in this data source

Background

- Basic biomedical research is crucial to medicine
  - Complexity of molecular pathophysiology
  - Challenges development of new therapies
- Big data can facilitate progress
  - MEDLINE: Biomedical research literature
- Need effective, automatic access to information in this data source
- Semantic processing

Challenges

- All data must be searched simultaneously to discover hidden relationships
- Response time needs to be improved over that available with commodity hardware
- Need a solution that can support 50 billion triples
  - Electronic medical record
  - Structured biomedical data
  - Web content

Solution

- Convert data to RDF graph database
- Use purpose build hardware
  - YarcData's Urika graph appliance
  - Designed for graph database discovery
- Support large expansion, real time response, and limited performance degradation
- Pilot project for proof of concept

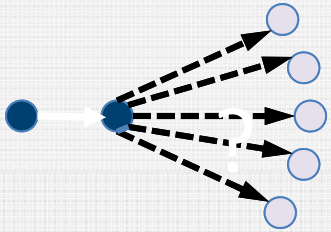
Pilot project components

- Data and metadata in RDF
  - Nearly 24 million MEDLINE citations
  - Nearly 70 million semantic predications
  - Both converted to 2.2 billion RDF triples
- Effective computing infrastructure
  - YarcData Urika graph appliance
- Application to manipulate semantic content of text
  - Semantic MEDLINE

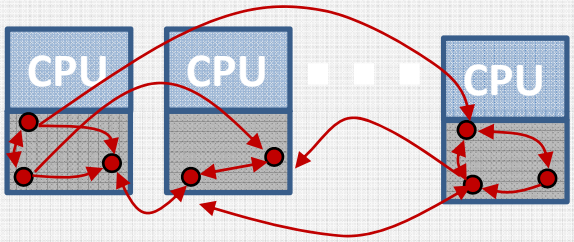
The YarcData approach

Research challenge:

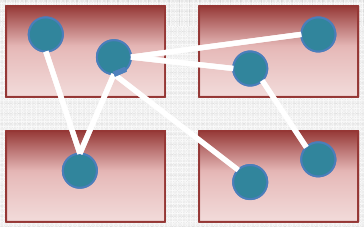
uRiKA



Large Shared Memory Architecture
Up to 512 TB



XMT2 Massively Multi-Threaded Processors
128 Threads



Scalable IO
Up to 350TB per Hour

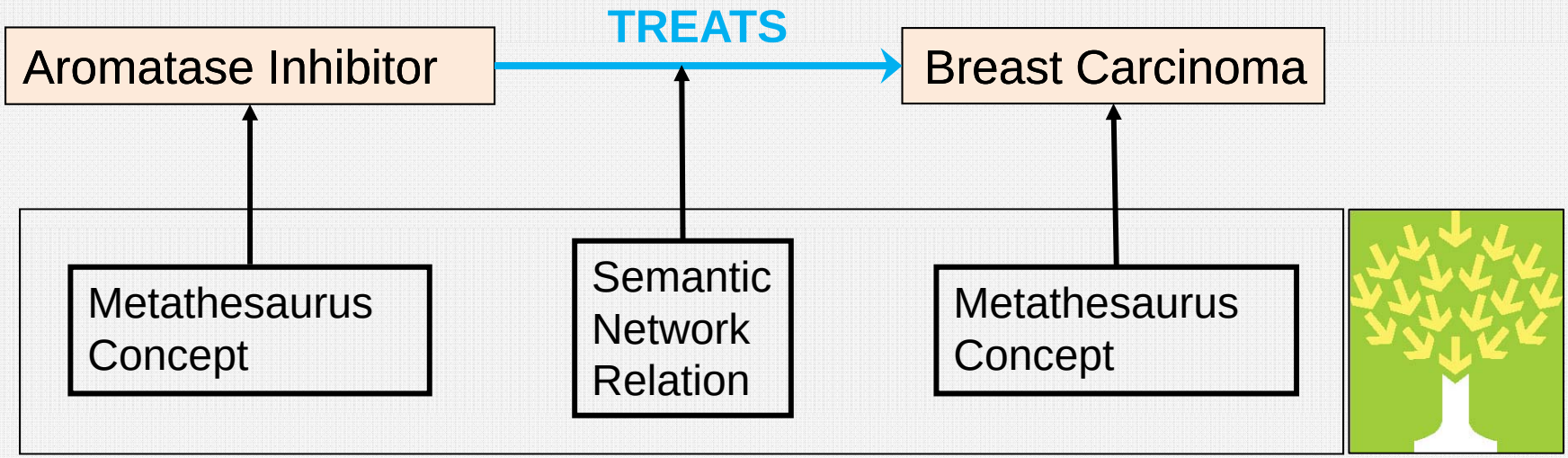
Real-time, Interactive Analytics
on Large Graph Problems

Semantic processing: SemRep

- Developed at National Library of Medicine
- Depends on domain knowledge
  - Unified Medical Language System (UMLS)
- Computable representation of meaning
  - Semantic predications

SemRep: semantic predication

Exemestane after non-steroidal aromatase inhibitors for post-menopausal women with advanced breast cancer



Unified Medical Language System

Web application: Semantic MEDLINE

- Uses nearly 70 million semantic predications
  - From all of MEDLINE
- To guide the user through content
- Exploits existing IR system
  - PubMed
- Displays results as an interactive graph

Semantic MEDLINE overview

Document retrieval

MEDLINE citations

SemRep: semantic processing

Semantic predications

Automatic summarization

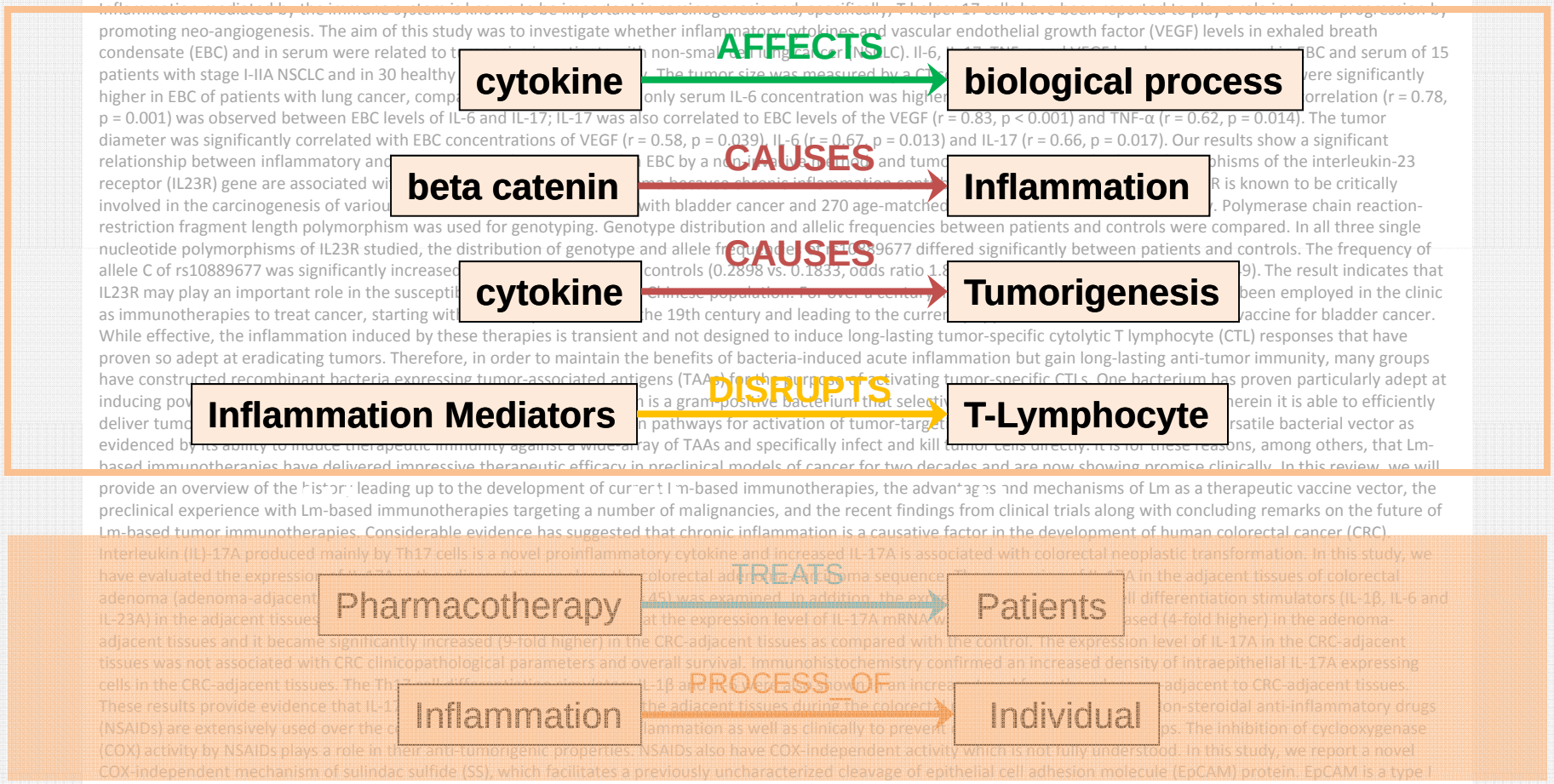
Graphical summary

Biomedical information management

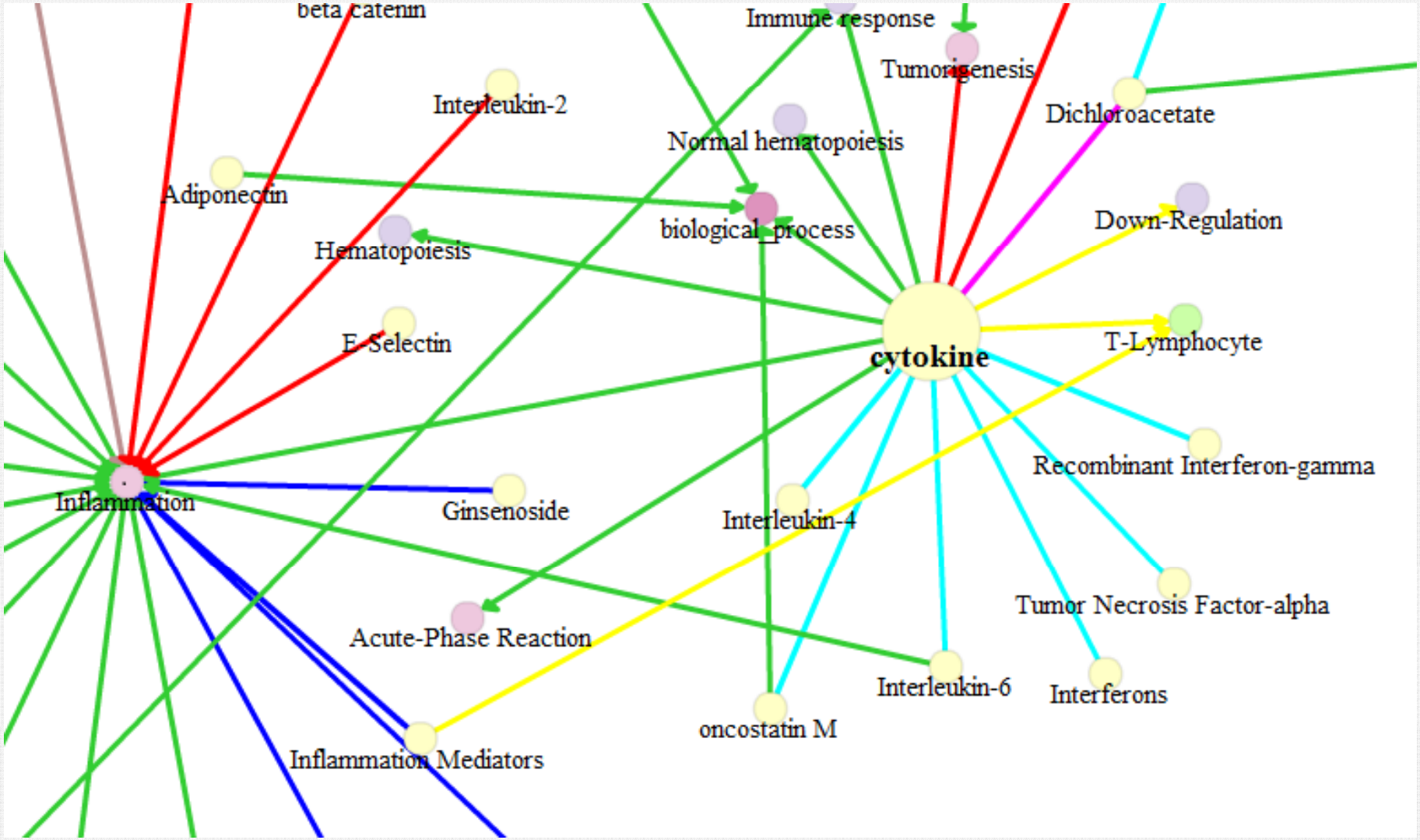
Use case: Inflammation and cancer

- With some exceptions, cancer therapy is not effective
- Scientific basis
  - Traditionally: kill cancer cells
  - More recently: manipulate non-cancer cells (immune system)
- Goal: look for trends in cancer immunotherapy

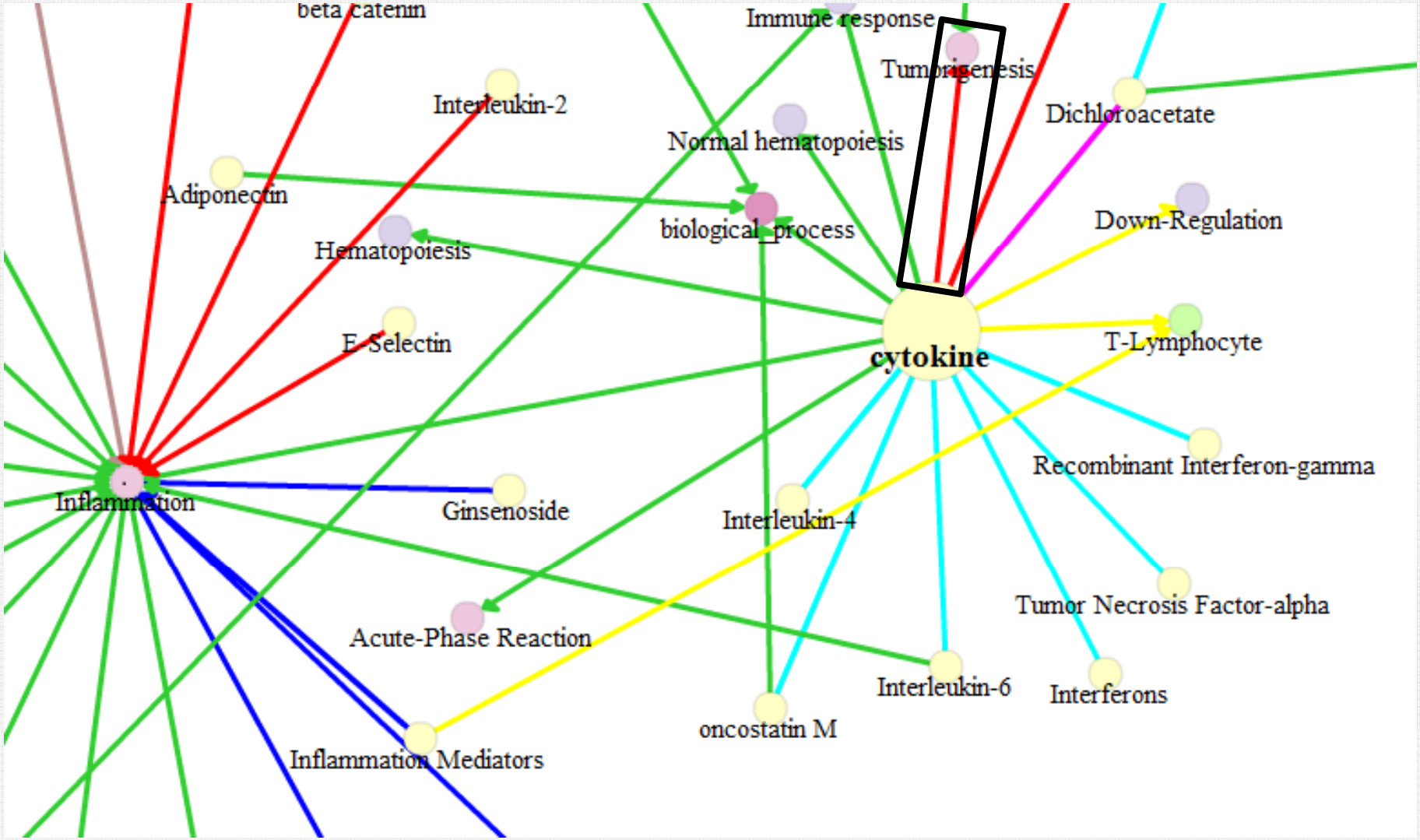
SemMed: semantic predications



Semantic predications as a graph



Cytokine CAUSES Tumorigenesis



MEDLINE citation

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Display Settings: Abstract

Send to:

Cancer Invest. 2014 Jun;32(5):197-205. doi: 10.3109/07357907.2014.898156. Epub 2014 Mar 26.

Interleukin-8 and interleukin-17 for cancer.

Zarogoulidis P[1], Katsikogianni F, Tsiouda T, Sakkas A, Katsikogiannis N, Zarogoulidis K.

Author information

Abstract

Pro-inflammatory cytokines have been associated with chronic inflammation and inflammatory diseases. Increased levels of interleukins (ILs) have been associated with inflammatory disease exacerbation. ILs levels have been observed to be associated with advance stage cancer for several types of cancer and a poor prognostic maker for malignant disease. Moreover; increased levels of cytokines induce tumorigenesis. There are several paradigms such as the hepatocellular

Exploiting semantic processing

- Research
  - Literature-based discovery (LBD)
    - Hypothesis generation
  - Discovery browsing
    - Investigate salient aspects of a topic
- Trends
  - Discern trends: Where is research headed?
  - Guide trends: Where should it be headed?

Acknowledgments

- Michael J. Cairelli, D.O.
- Marcelo Fiszman, M.D., Ph.D.
- Halil Kilicoglu, Ph.D.
- Graciela Rosemblat, Ph.D.
- Dongwook Shin, Ph.D.
- YarcData team
  - Aaron Bossert
  - Ted Slater
  - Tim White
  - Fredrik Salvesen

Additional information

- Slides, including 7 minute live demo: http://www.youtube.com/watch?v=ShfI4SNzNO4
- Slides, including 20 minute live demo: http://www.youtube.com/watch?v=6frNAmPD0mo
- Thomas C. Rindflesch (tcr@nlm.nih.gov)
- Fredrik Salvesen (fredrik@salvesen.me)