Content-rich biological network constructed by mining PubMed abstracts

by Hao Chen and Burt M. Sharp
BMC Bioinformatics 5(147) October 2004
DOI: 10.1186/1471-2105-5-147

presented by Christopher Maier for
INLS 279: Bioinformatics Research Review
2005-02-08

Motivating Problem

"Despite the existence of text-mining methods that identify biological relationships based on the textual co-occurrence of gene/protein terms or similarities in abstract texts, knowledge of the underlying molecular connections on a large scale, which is prerequisite to understanding novel biological processes, lags far behind the accumulation of data."

Motivating Problems

- Tools based on term co-occurrence are plentiful and computationally efficient, but do not give information regarding the nature of interactions
- Tools using NLP techniques have been made, but for testing purposes and are often unavailable to the scientific community

Behold: CHILIBOT!

- CHIp LIterature RoBOT
- Leverages NLP techniques to create interaction networks among biological concepts, genes, proteins, and drugs
- Characterizes the nature of the interaction
- Creates networks that reflect the scale-free nature of biological networks

CHILIBOT Architecture

- Web-based Perl application running on *NIX platform
- Synonym database (like ARGH)
- NCBI E-Fetch interface to PubMed
- POS tagging of sentences, followed by shallow parsing
- Visualization by AiSee

Example Network

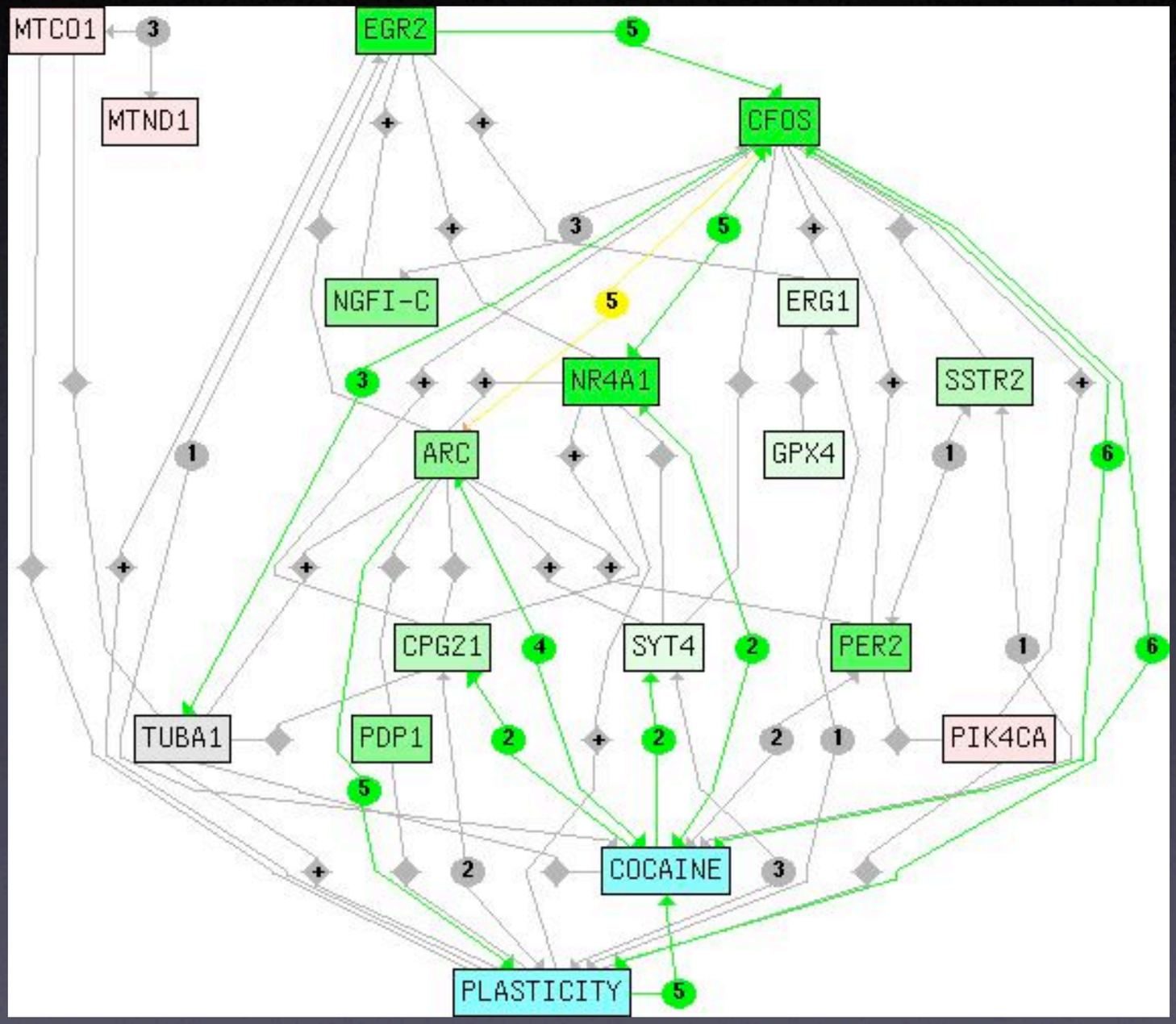


Fig. I

Evaluation and Verification

- DIP: Database of Interacting Proteins
  - Manually curated database of protein interaction relationships
- Try to recall known relationships and classify them correctly

Effect of Increasing Numbers of Abstracts

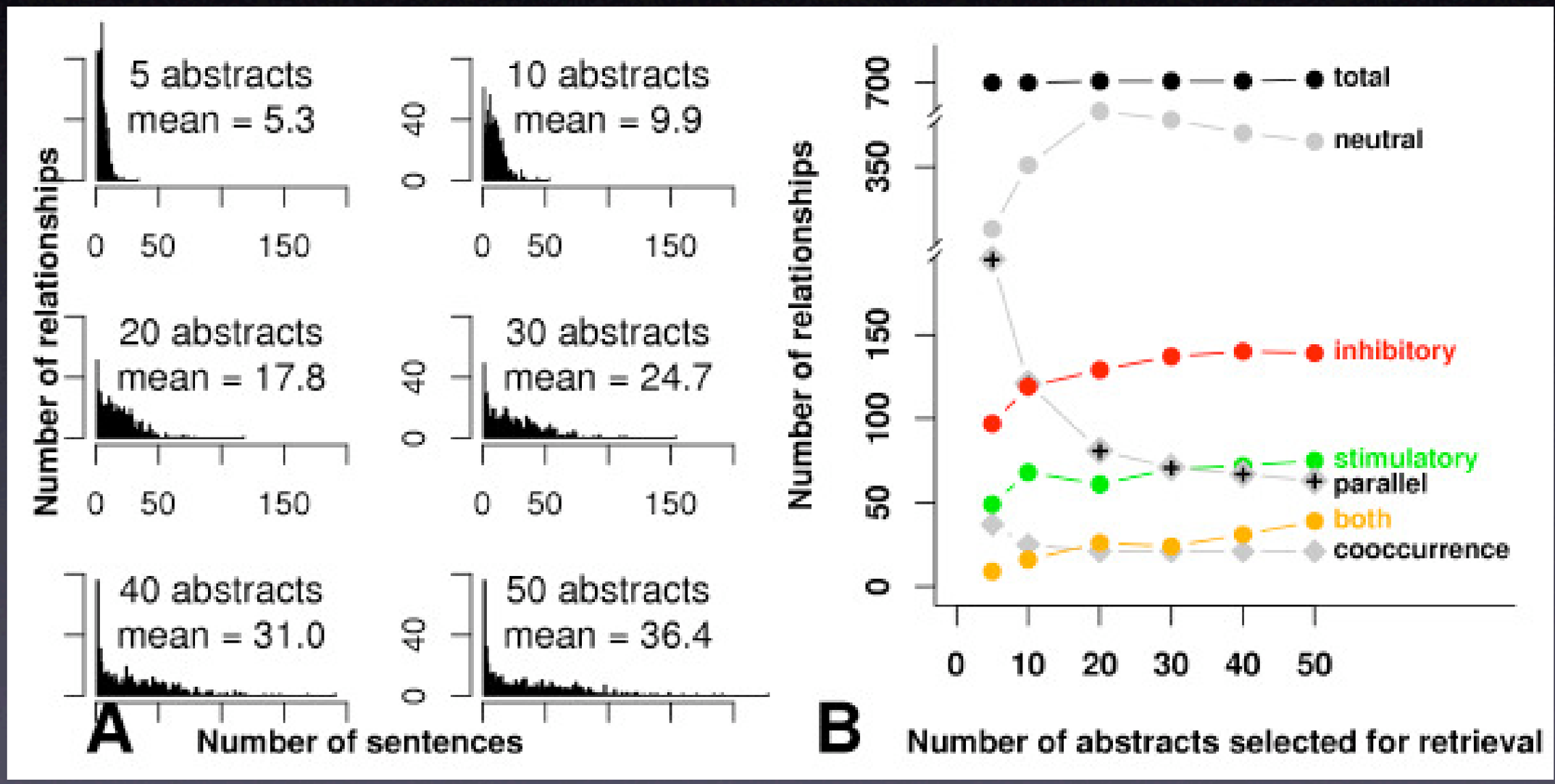


Fig. 3

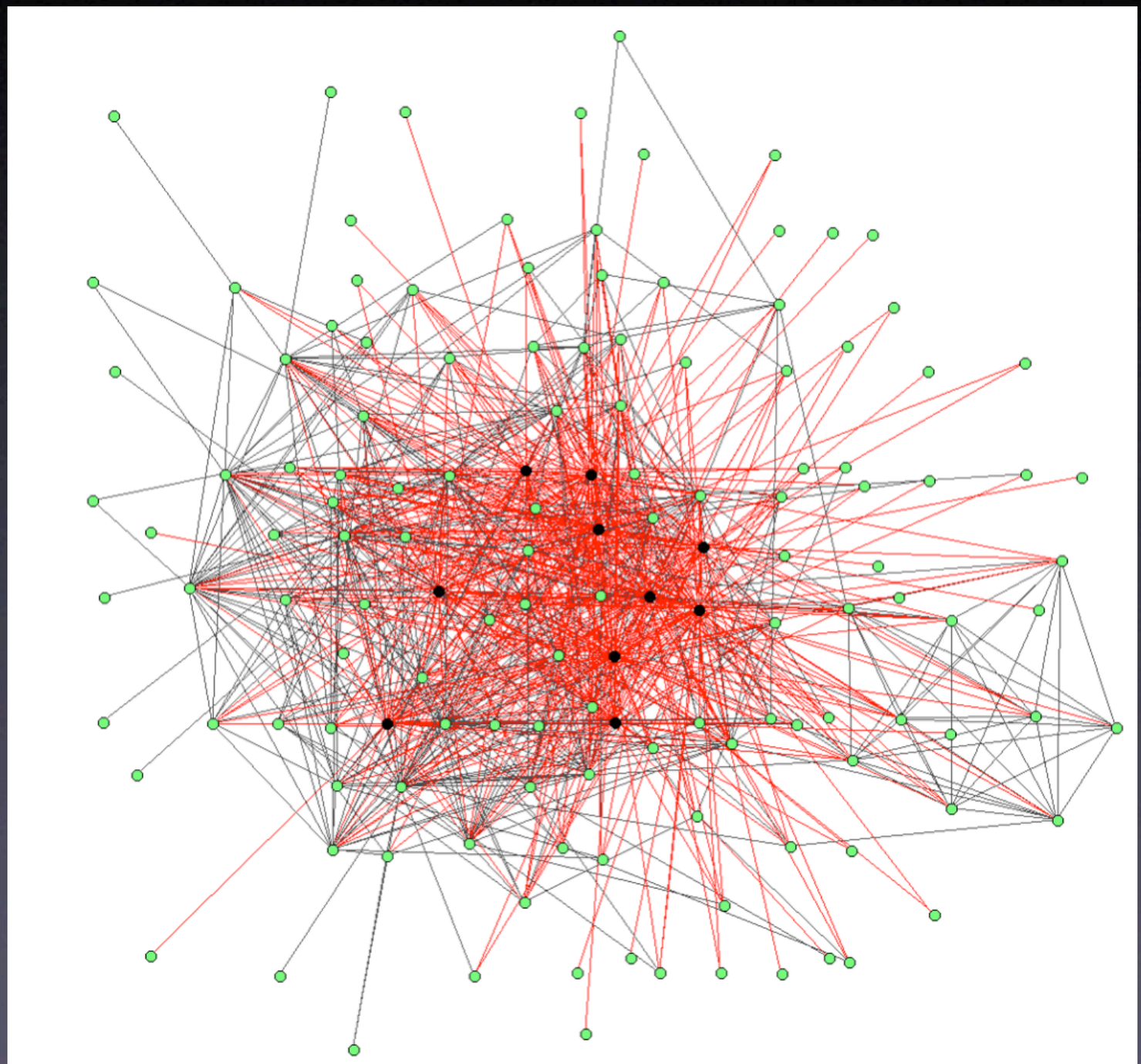
Failure to detect known DIP relationships

Reason	Undetected relationships (%)
member of protein family (name generalization)	30.8
incomplete synonym list	26.5
no reference at abstract level	22.1
other	20.6

Hypothesis Suggestion

- Swanson's A-B-C model of "undiscovered public knowledge"
- Scan network for pairs of concepts that share no direct link, but have tertiary nodes in common
- Retrospective and prospective studies

Bibliometric Networks Appear to be Scale-Free



Thoughts

- How did they determine the fold increase / decrease of a protein?
- Scale-free network bits seemed randomly tacked-on
- Would like to see specific examples of sentences and rules applied to them