# ASSOCIATING GENES AND PROTEIN COMPLEXES WITH DISEASE VIA NETWORK PROPAGATION

ARCHIT SINGH

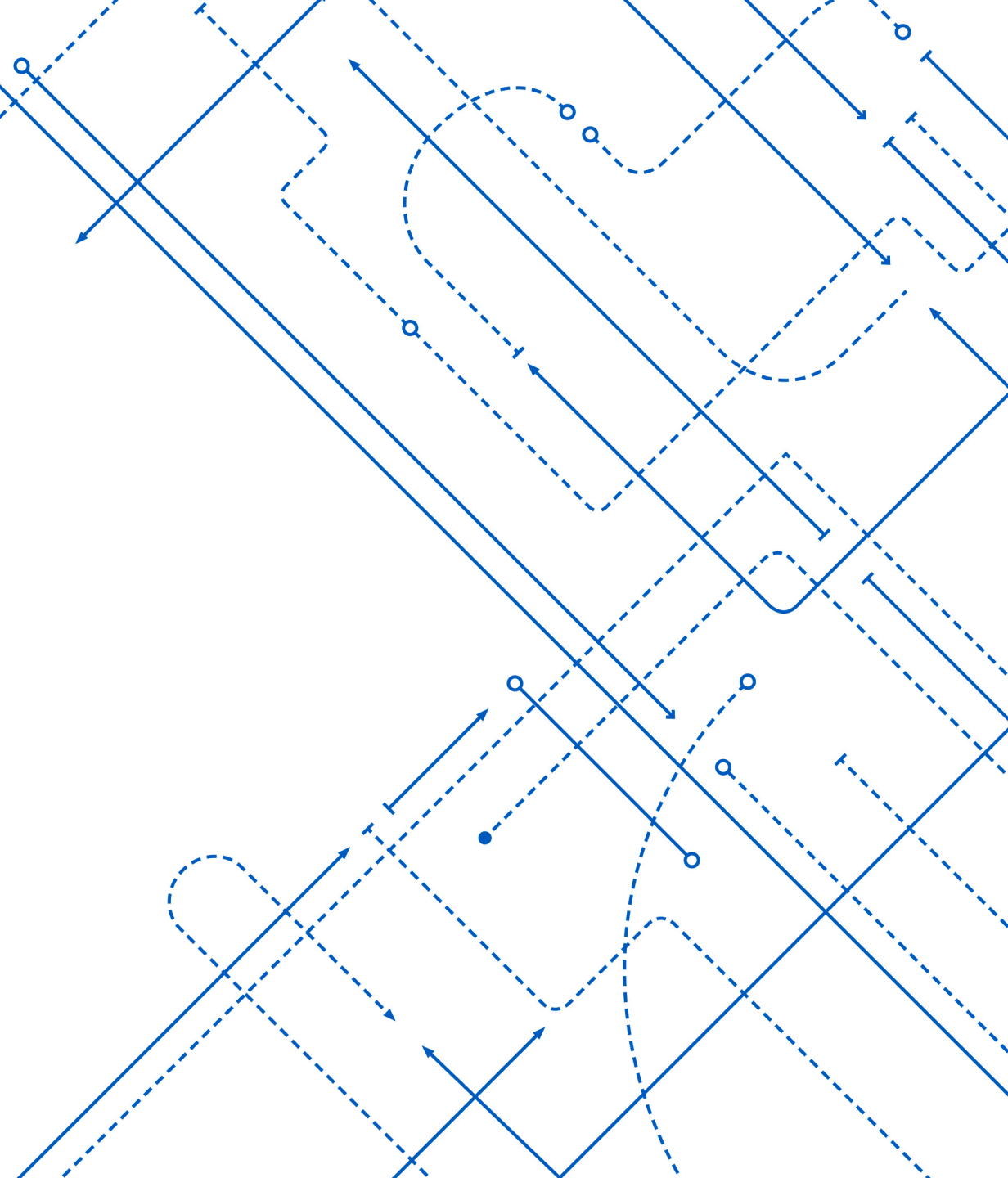
SHUBHAM SHARMA

University at Buffalo
School of Engineering and Applied Sciences

University at Buffalo
School of Engineering and Applied Sciences

# CONTENTS





# Introduction

- Genes - A gene is a basic unit of heredity in a living organism. Genes are coded instructions that decide what the organism is like, how it behaves in its environment.
- Proteins - Proteins are large, complex molecules that play many critical roles in the body. They are necessary for building the structural components of the human body, such as muscles and organs.
- PPI - Protein–protein interactions (PPIs) are the physical contacts of high specificity established between two or more protein molecules
- Genes and proteins - Most genes contain the information require to make proteins. The journey from gene to protein is one that is complex and controlled within each cell and it consists of two major steps – transcription and translation

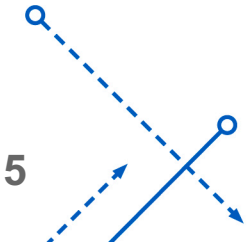
# Challenges

- Identification of disease-causing genes is a fundamental challenge in human health
- Recently, several studies have tackled this challenge via a network-based approach, motivated by the observation that genes causing the same or similar diseases tend to lie close to one another in a network of protein-protein or functional interactions.
- Most of these approaches use only local network information in the inference process and are restricted to inferring single gene associations.
- This research provides a global, network-based method for prioritizing disease genes and inferring protein complex associations, which we call PRINCE

# Existing Solutions and their Shortcomings

- Lage K. et al. – Devised a method of scoring a protein wrt a disease based on the involvement of its direct neighbors in a similar disease
- Kohler et al. – Performed this by grouping diseases into families. For a given disease they employed a random walk from the known genes to prioritize the candidate genes
- Wu et al. – Devised a method to score a candidate gene based on correlation between vectors of similarity between the disease and diseases with known causal genes and the vector of closeness between the gene g and the known disease genes.

- These methods either reveal a local network of genes or suggest just the prioritization and not the association between genes



# PRINCE to the rescue - (PRIoritizatioN and Complex Elucidation)

- Iterative network propagation method
- Gene Prioritization Function

$$F(v) = \alpha \left[ \sum_{u \in N(v)} F(u) w'(v,u) \right] + (1-\alpha) Y(v)$$

$$F = \alpha W' F + (1-\alpha) Y \Leftrightarrow F = (I - \alpha W')^{-1} (1-\alpha) Y$$

$$W' = D^{-1/2} W D^{-1/2}$$

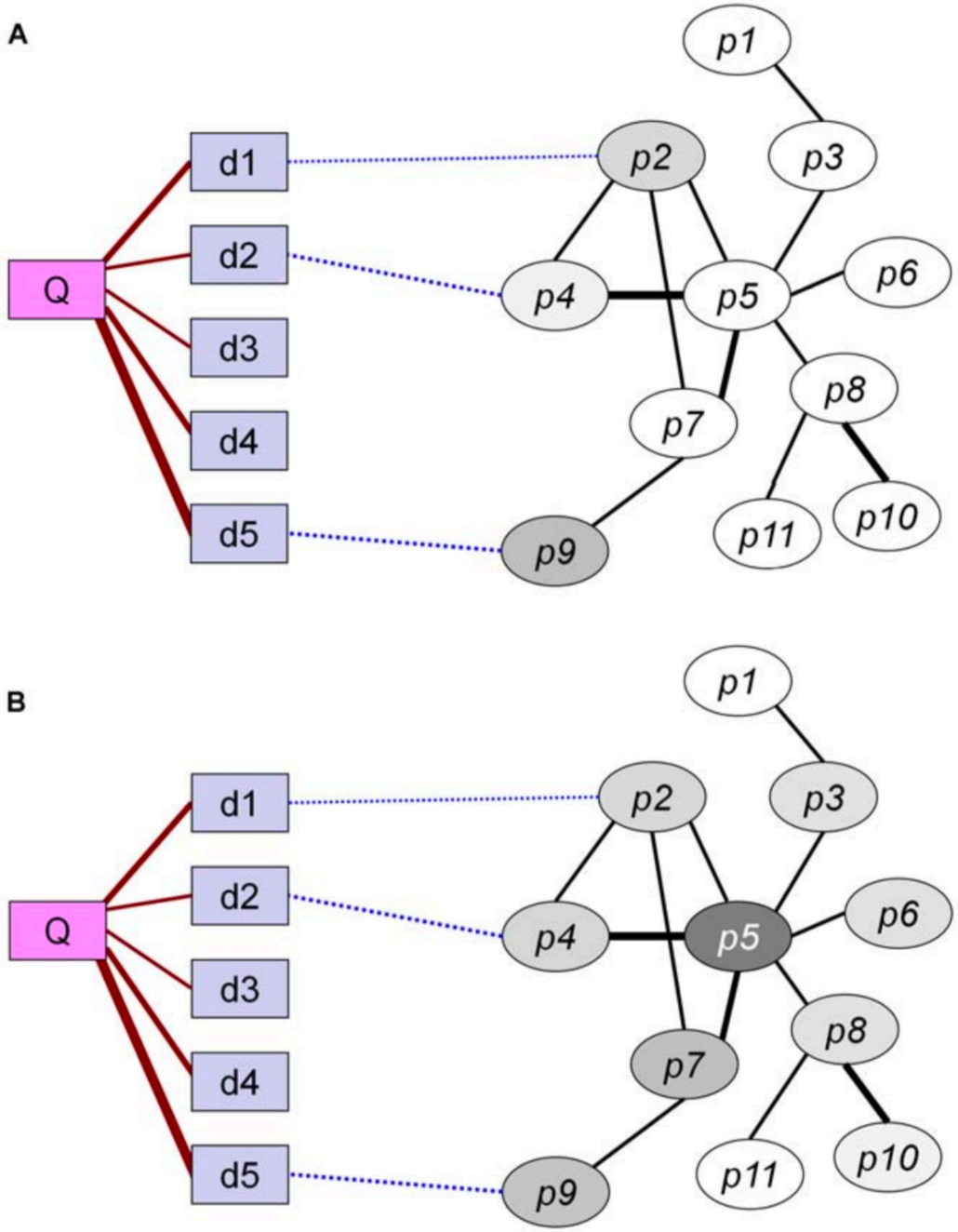
- Computing prior information vector Y using similarity between diseases

$$L(x) = \frac{1}{1 + e^{(cx+d)}} \qquad Y(v) = L(S(q,p))$$

- Iteration

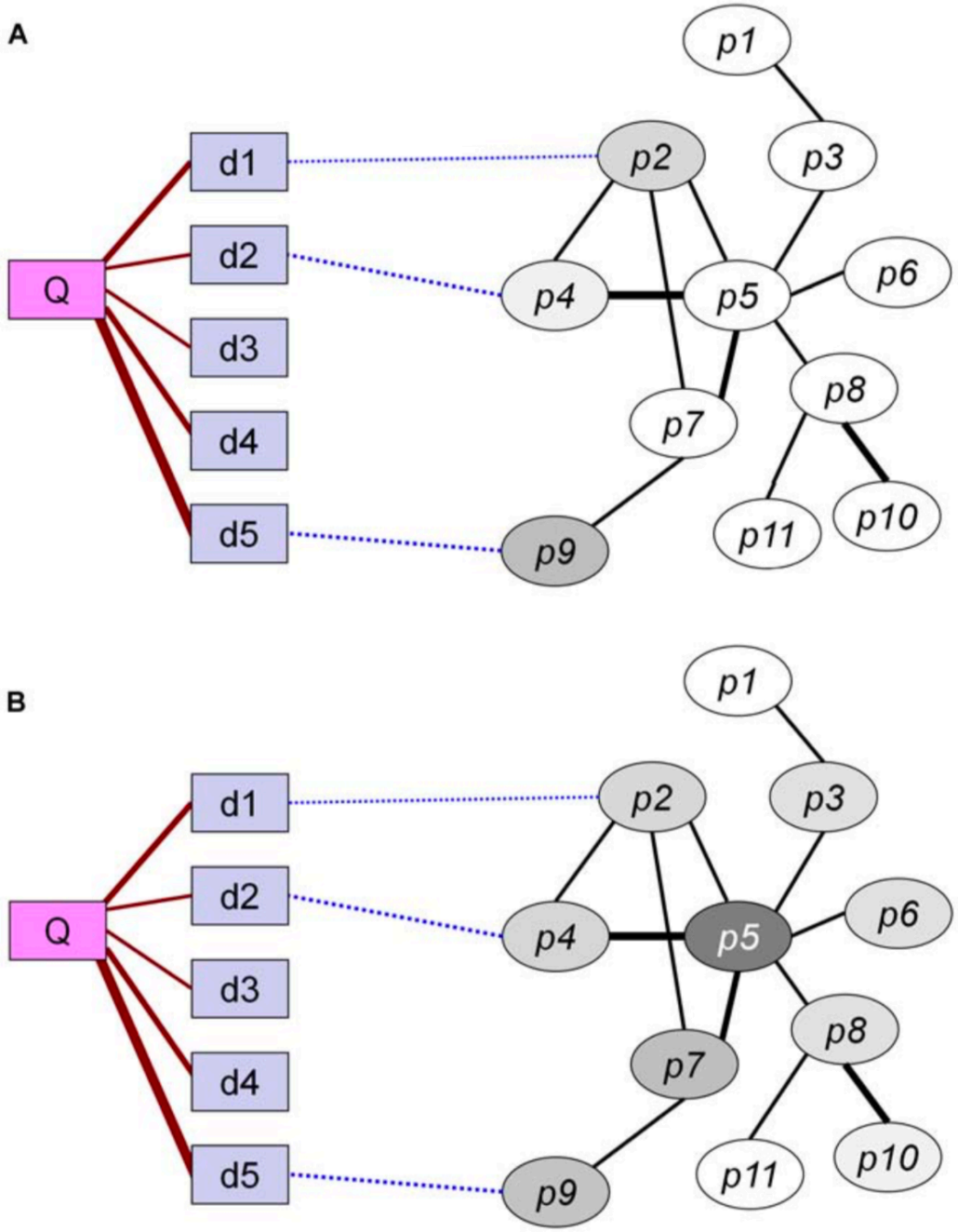
$$F^t := \alpha W' F^{t-1} + (1-\alpha) Y$$

Convergence of flow after several iterations



# PRINCE (PRIoritizatioN and Complex Elucidation)

- Parameter Tuning – tuned using cross validation
  - c – Parameter controlling the logistic regression transformation
  - α - Relative importance of prior information in the association assignement
  - Number of Propagation iterations
- Powerful method with demonstrated ability
- Global Network Approach
- Normalization of PPI weights as well as disease-disease similarities



# Comparison with other methods

- Leave-one-out cross validation
- Precision versus Recall graph when varying the rank threshold 1<k<100
- As can be seen from the graph Precision takes a dip and Recall sees an increase with the increasing value of k

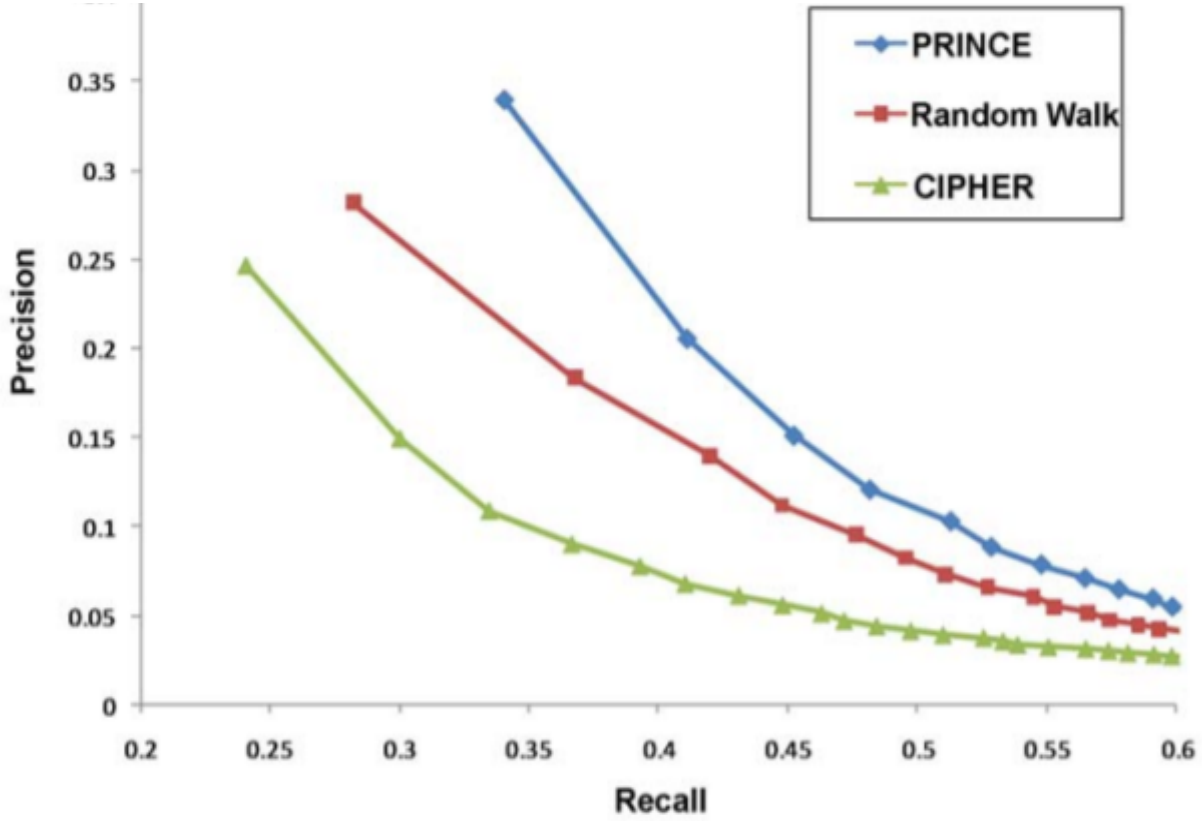



**Figure 2. A comparison of prioritization algorithms.** Performance comparison for PRINCE, Random Walk and CIPHER in a leave-one-out cross-validation test over 1,369 diseases with a known causal gene. The figure shows recall versus precision when considering the top $k\%$ proteins for various values of $k$.

# Limitations

- Needs prior phenotypic information
- Fails to utilize other relevant data like differentially expressed genes
- Depends on accurate and comprehensive Protein-Protein Interaction data


ThankYou
Merci
Salamat
Kop
Hvala
Tack
kun
Xie
Gra
Dhanyavad
Khopjai
Gomapsupnida
Terima
Danke
dank
Enkosi
Euxaristo
Kun
Shukriya
Shokriya
Mamnoon
Kasih
Todah
Shokrun
Spaas
Mul
Ači
Ngiyabonga
Gamsahapnida
Dank
Takk
Cam
Dankie
Kruthagnathalu
Arigatou
Dhonnobaad
Asante
Dhan
daa
Hain
ederim
Or
Go
Grazie
Falemínderit
Dhanyavaad
Dhanyavaadaalu
Waad
Daw
Dakujem
krap
Grazzi
raibh
Gracias
Nandree
Blagodariya
Fyrir
Dekuju/Dekujeme
Te ekkür
Dziękuję
Diolch
Kiitos
Sheun
umesc
Shnorhakalutiun