Chemoinformatics and biological network

A focus on a ligand-target interaction network

Author: Vaishali CHAUDHARI

Supervisors: Prof. Bruno CREMILLEUX (GREYC), Prof. Thierry LECROQ (LITIS), Dr. Bertrand CUISSART (GREYC), Dr. Lina F. SOUALMIA (LITIS)

Université de Caen Normandie

4 July 2018

Introduction

Context

- Large amount of existing biological data related literature (PubMed) .
- Interaction between biological entities represented as a network.
- Several databases existing for biological data.
- Difficult in study because data is scale-free, complex, sparse (a vast majority of the interactions are missing).

Introduction

Context

- Large amount of existing biological data related literature (PubMed) .
- Interaction between biological entities represented as a network.
- Several databases existing for biological data.
- Difficult in study because data is scale-free, complex, sparse (a vast majority of the interactions are missing).

Our work:

- Part 1
  - Biological network: resources, types, pathways
  - Tools: CytoScape, Bioconductor (packages such as rBiopaxparser, PaxtoolsR)
  - Databases

Introduction

Context

- Large amount of existing biological data related literature (PubMed) .
- Interaction between biological entities represented as a network.
- Several databases existing for biological data.
- Difficult in study because data is scale-free, complex, sparse (a vast majority of the interactions are missing).

Our work:

- Part 1
  - Biological network: resources, types, pathways
  - Tools: CytoScape, Bioconductor (packages such as rBiopaxparser, PaxtoolsR)
  - Databases
- Part 2
  - Ligand-target interaction network
  - Simple random walk on network

Outline

Part 1
Biological network

Biological network

What is a network?

A linked list of interconnected nodes.

Biological network

What is a network?

A linked list of interconnected nodes.

What is a biological network?

- **Biological network**
  - **Series of interconnected pathways.**
  - **Nodes** - protein, peptide, gene, DNA, RNA, biomolecules etc.
  - **Edges** - pathways

Biological network

What is a network?

A linked list of interconnected nodes.

What is a biological network?

- **Biological network**
  - **Series of interconnected pathways.**
  - **Nodes** - protein, peptide, gene, DNA, RNA, biomolecules etc.
  - **Edges** - pathways
- **Pathway**

  Series of biological relationships such as interactions, regulations, reaction, activation, inhibitions, transformations.

Biological network

What is a network?

A linked list of interconnected nodes.

What is a biological network?

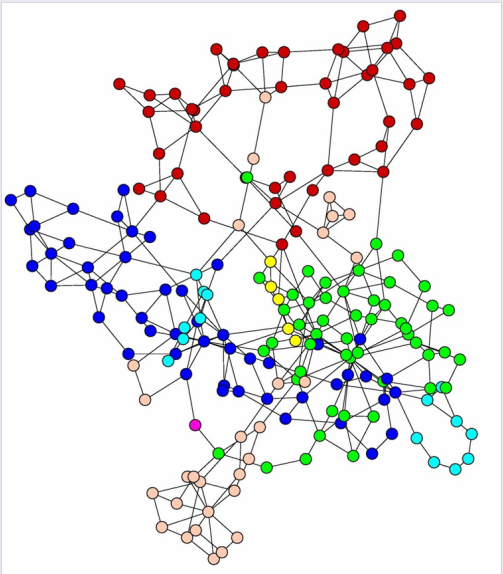
- **Biological network**
  - **Series of interconnected pathways.**
  - **Nodes** - protein, peptide, gene, DNA, RNA, biomolecules etc.
  - **Edges** - pathways
- **Pathway**

  Series of biological relationships such as interactions, regulations, reaction, activation, inhibitions, transformations.
- **XML based pathway representation format**

  BioPAX, PSIMI, SBML, KGML

Types of biological network (I)

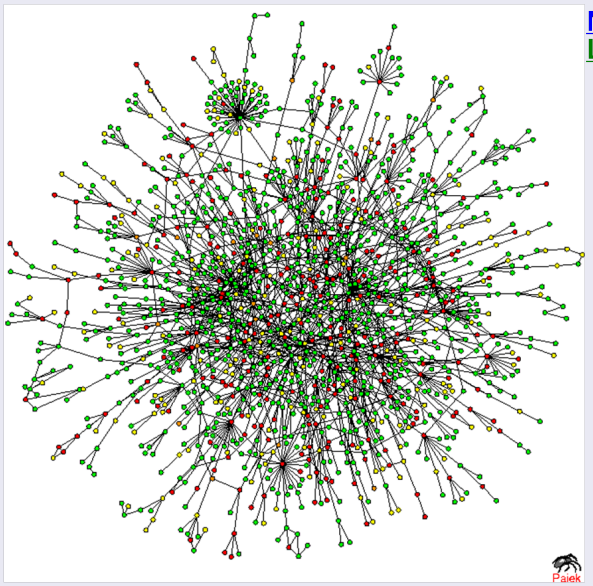
- **Metabolic network**
  - Represent chain of chemical reactions of small molecules, called metabolites or substrates.
  - Metabolic reactions are reversible i.e. they occur in both directions.
  - Undirected network represented by complete metabolic pathways.
  - Databases: KEGG, Reactome, SMPDB, WikiPathways, MANET



Metabolic network example [RSM+02]

Types of biological network (II)

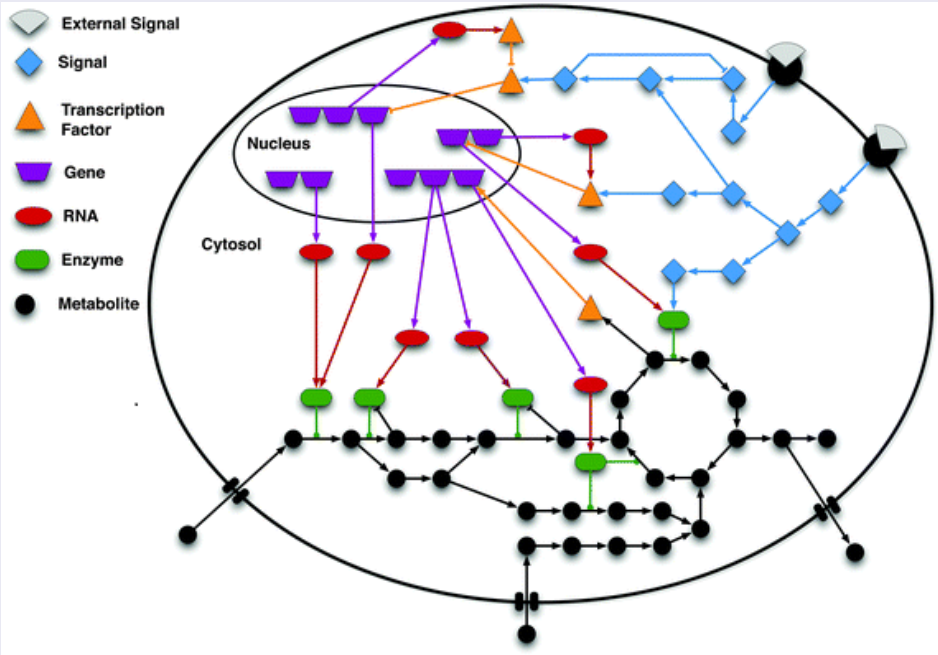
- **Protein-protein interaction network**
  - A protein can interact with another protein, e.g., to build a protein complex or to activate it.
  - Protein-protein interactions form large undirected network.
  - Databases: BIND, BioGRID, CCSB Interactome, NetPro, BioLip-protein-ligand binding database



Protein interaction network example[JMBO01]

Types of biological network (III)

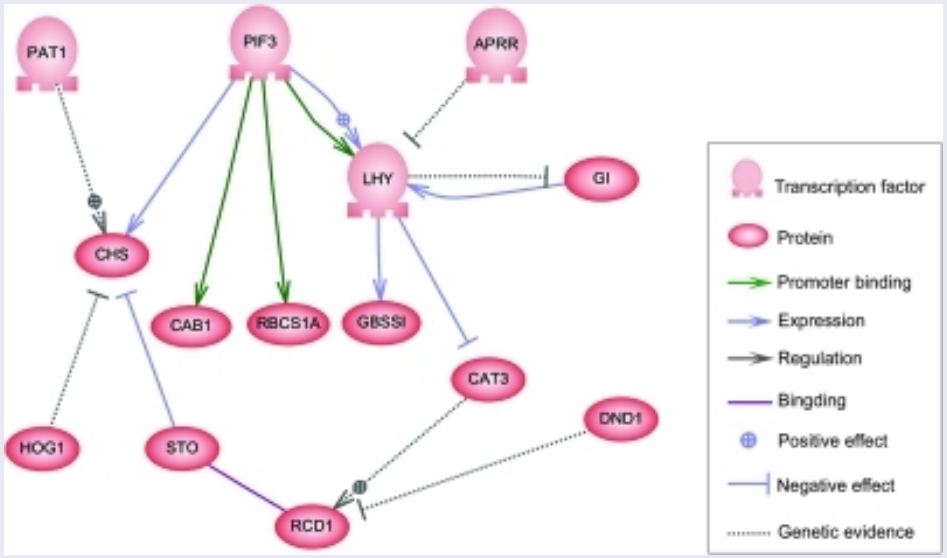
- **Signal transduction network example**
  - The complete network of all directed signal transduction pathways
  - Represent communication between cells. First molecule in a pathway receives a signal, it activates another molecule.
  - Databases: KEGG, MiST, NCI / Nature Pathway Interaction



Signal transduction network example [GBR+13]

Types of biological network (IV)

- **Gene regulatory network**
  - Directed network of gene regulatory pathways
  - It models the indirect regulation and interaction of genes.
  - Databases: KEGG, BioGPS, GeneMania, BioCarta, Reactome



Gene regulatory network [Wik17c]

Pathway visualization

- **Input**
  - BioPAX (.OWL) or KGML or SBML file

Pathway visualization

- **Input**
  - BioPAX (.OWL) or KGML or SBML file
- **Where to find input file?**
  - `https://www.ncbi.nlm.nih.gov/biosystems`
  - `http://reactome.org/PathwayBrowser/`
  - `http://www.genome.jp/kegg/pathway.html`

Pathway visualization

- **Input**
  - BioPAX (.OWL) or KGML or SBML file
- **Where to find input file?**
  - `https://www.ncbi.nlm.nih.gov/biosystems`
  - `http://reactome.org/PathwayBrowser/`
  - `http://www.genome.jp/kegg/pathway.html`
- **Tools**
  - Cytoscape or Bioconductor package (rBiopaxparser or PaxtoolsR) + R programming

Pathway visualization

- **Input**
  - BioPAX (.OWL) or KGML or SBML file
- **Where to find input file?**
  - `https://www.ncbi.nlm.nih.gov/biosystems`
  - `http://reactome.org/PathwayBrowser/`
  - `http://www.genome.jp/kegg/pathway.html`
- **Tools**
  - Cytoscape or Bioconductor package (rBiopaxparser or PaxtoolsR) + R programming
- **Output**
  - Network with pathways

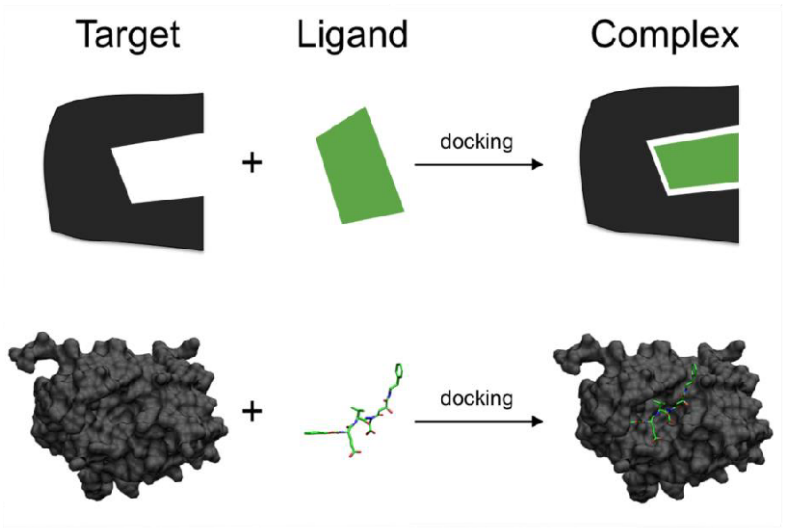
Part 2

Ligand-target interaction network

Ligand-target interaction

Ligand

- Low molecular weight chemical compound (weight < 900 daltons) .
- Ligands are drugable molecules [Wik18]
- **Ligand classes:**
  Synthetic organic, Metabolite, Natural product, Endogenous peptide, Inorganic, Approved drug



Stable complex using ligand and target [Wik17b]

Ligand-target interaction

Target or receptor

- Partner molecule to a ligand in complex forming process and most known targets are **proteins**.
- Proteins are found in a human body.
- **Target classes:**
  Enzyme, ion channels (VGIC, LGIC) , GPCR, nuclear receptor

Ligand-target interaction

Target or receptor

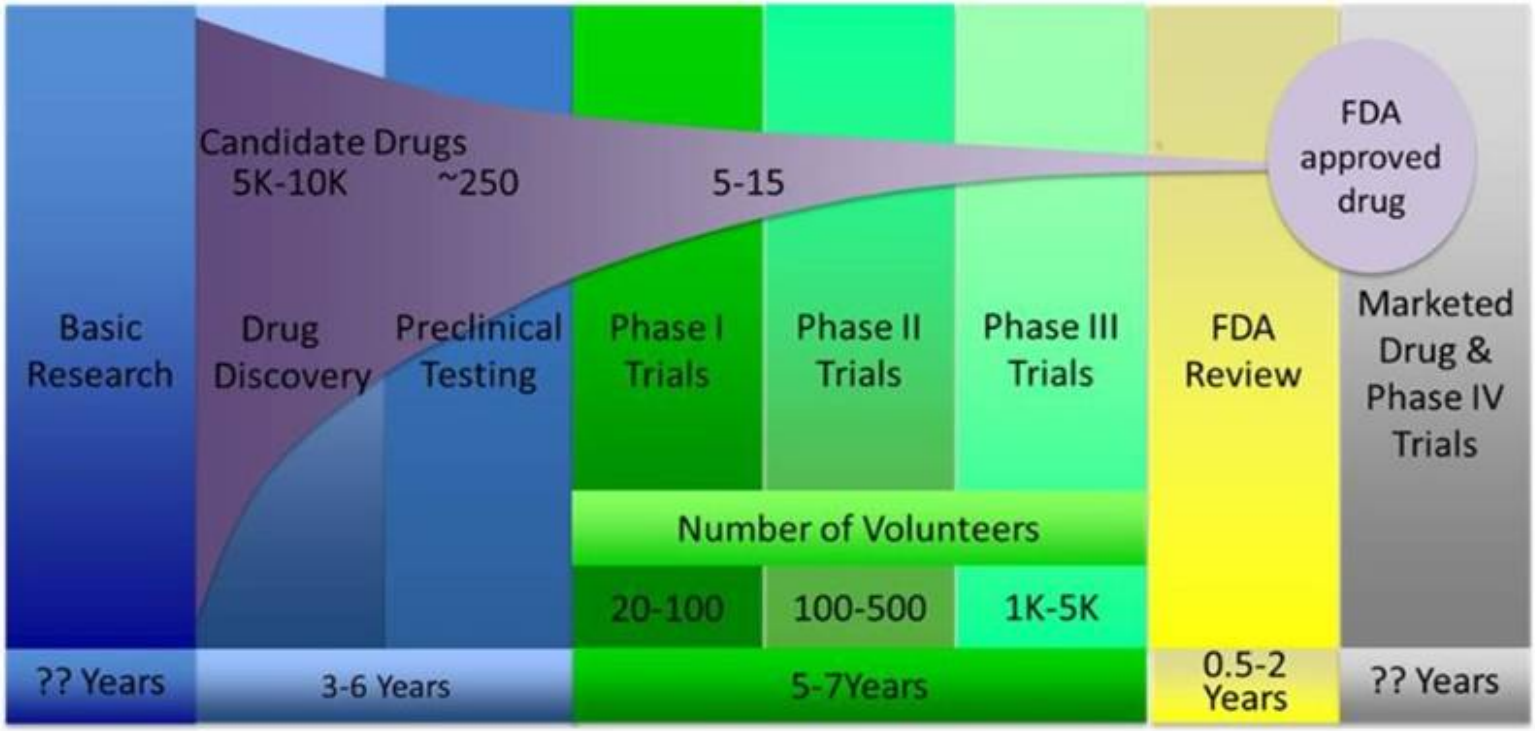
- Partner molecule to a ligand in complex forming process and most known targets are **proteins**.
- Proteins are found in a human body.
- **Target classes:**
  Enzyme, ion channels (VGIC, LGIC) , GPCR, nuclear receptor

Ligand-target interaction network

- Network is formed by using nodes (ligands, targets) and edges (interaction)
- Undirected network
- Predict new ligand-target interaction using topology of network
- Biggest database for interactions: IUPHAR / BPS database

Ligand-target interaction network

Main application: Drug discovery



https://www.linkedin.com/pulse/modern-drug-discovery-mike-jaqua

Dataset and software

Dataset

- Dataset from IUPHAR/BPS Guide to pharmacology (GtoPdb) database in PostgreSQL format
- Size of database = 18.6 MB
- 8872 ligands, 2900 target proteins, 16979 ligand-target interactions
- Number of literature data = 31,239 publications

Software

- **Programming language:** Python
- **Database:** PostgreSQL

Simple Random Walk

Simple Random Walk on a Graph

- Start from a random node at time t = 0
- Choose a neighbor randomly (including previous node) and move there
- Repeat until $t = \infty$ where ∞ : sufficiently large.

- Random walk on a graph $G = (V, E)$:
  - a sequence of state vectors: $(p^0, p^1, p^2, \ldots, p^n)$

- p^t = random walker state vector at time t

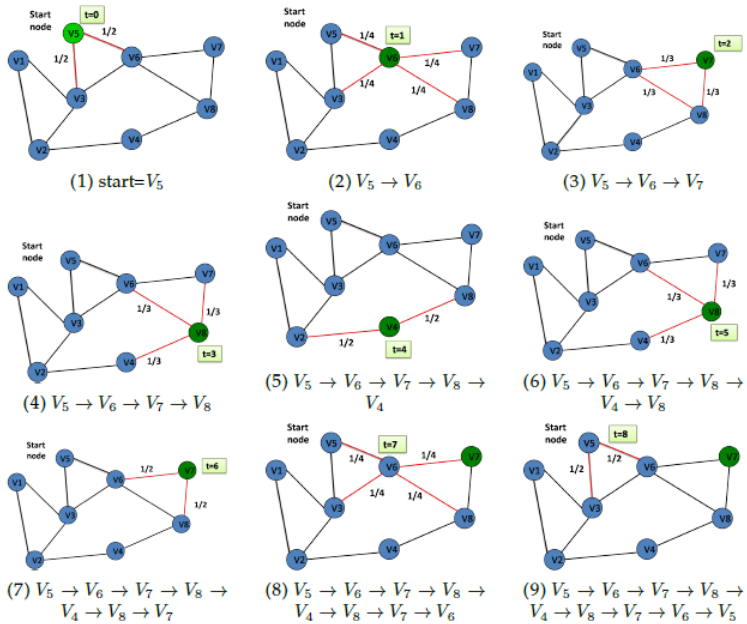
$$p^t = P * p^{t-1} \tag{1}$$

where P: transition probability matrix

- Markov chain with memory 1

Simple Random Walk: illustration

- $t = 1$, $p^1 = P * p^0$, paths of length 1 from p_0,
- $t = 2$, $p^2 = P * p^1 = P^2 * p^0$, paths of length 2,
- $t = 3$, $p^3 = P * p^2 = P^3 * p^0$, paths of length 3,
- $t = \infty$, $p^\infty = P * p^{t-1} = P^\infty * p^0$, long paths.



Random walk example

Different behaviours of a walker

Simple random walk

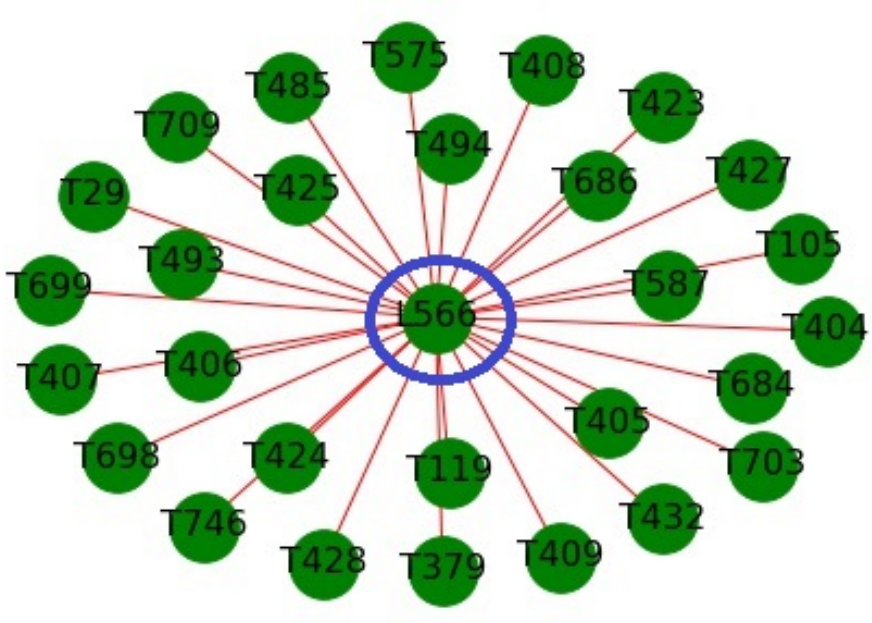
- $p = (p_1, \ldots, p_n))$, $q = (q_1, \ldots, q_n)$ two state vectors,
- Euclidean distance:

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \tag{2}$$

- Cyclic (as a Markov Chain with memory 1)
- Convergent (the distance between two successive states is decreasing)
- Stochastic (none of the other behviours)

Ligand-target interaction network

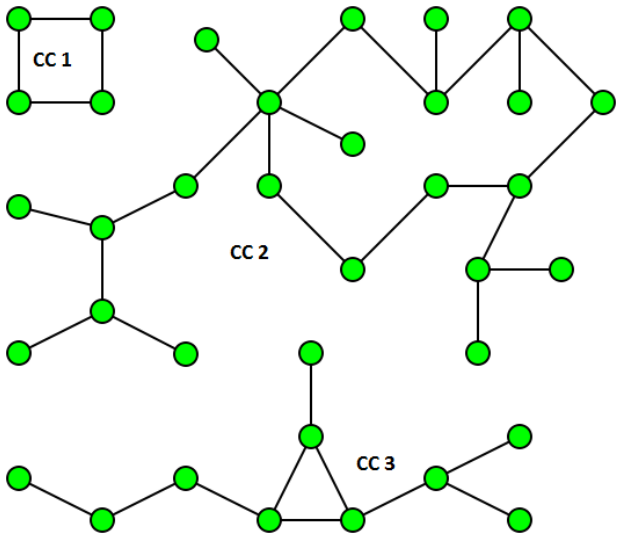
- Interaction network of ligand-target



Ligand-target interaction network

Connected Component (CC)

- Connected component of an undirected graph G is a connected subgraph of G of maximum size.



Connected component example [Wik17a]

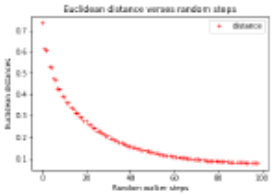
- **Total 490 CC in ligand-target interaction network of IUPHAR database**

10 largest Connected Component (CC) with ligand distribution in IUPHAR

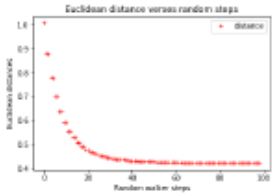
		Ligand Class or Family						
Connected Component (CC)	Total number of nodes in CC	Metabolite	Natural Product	Inorganic	Peptide	Antibody	Synthetic organic	Total
1	5588	313	154	27	580	27	3597	**4698**
2	248	3	1	0	0	0	235	**239**
3	154	0	1	0	91	0	53	**145**
4	113	0	0	0	95	0	12	**107**
5	103	0	0	0	67	0	32	**99**
6	101	0	1	0	16	70	8	**95**
7	96	0	0	0	43	0	50	**93**
8	84	0	0	0	72	0	9	**81**
9	65	0	0	0	47	0	11	**58**
10	58	0	1	0	37	0	17	**55**

Experimental results

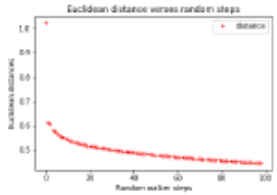
10 largest CC with Euclidean distance after 100 steps



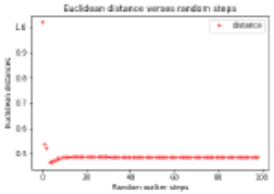
(1) CC1 (5588 nodes)



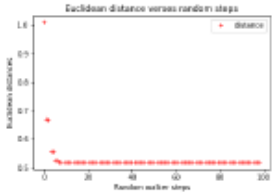
(2) CC2 (248 nodes)



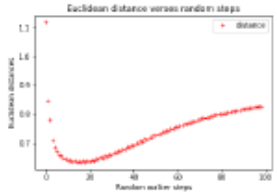
(3) CC3 (154 nodes)



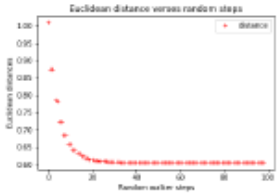
(4) CC4 (113 nodes)



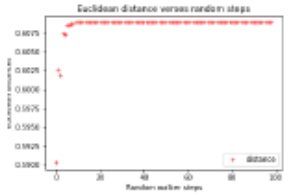
(5) CC5 (103 nodes)



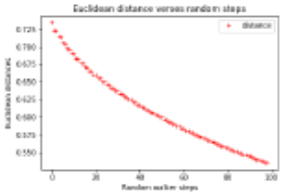
(6) CC6 (101 nodes)



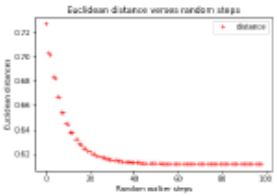
(7) CC7 (96 nodes)



(8) CC8 (84 nodes)



(9) CC9 (65 nodes)



(10) CC10 (58 nodes)

Experimental results

10 largest Connected Component (CC) Convergence status after 100 steps

Connected Component (CC)	Converged	Cyclic	Stochastic
1	Yes		
2	yes		
3	Yes		
4		Yes	
5		Yes	
6			Yes
7		Yes	
8			Yes
9			Yes
10	Yes		

Experimental results

Cyclic or non-cyclic state vectors detection after 100, 1000, 10000 steps

Connected Component (CC) with number of nodes	Iteration					
	100		1000		10000	
	Number of a cyclic state	Number of a non-cyclic state	Number of a cyclic state	Number of a non-cyclic state	Number of a cyclic state	Number of a non-cyclic state
CC 1 (5588 nodes)	0	5588	0	5588	5588	0
CC 2 (248 nodes)	0	248	248	0	248	0
CC 3 (154 nodes)	0	154	154	0	154	0
CC 4 (113 nodes)	0	113	113	0	113	0
CC 5 (103 nodes)	103	0	103	0	103	0
CC 6 (101 nodes)	0	101	101	0	101	0
CC 7 (96 nodes)	96	0	96	0	96	0
CC 8 (84 nodes)	84	0	84	0	84	0
CC 9 (65 nodes)	0	65	0	65	0	65
CC 10 (58 nodes)	4	54	58	0	58	0

Conclusion

- Convergence is not guaranteed with simple random walk.
- When number of iterations are increased most of the state vectors are in cyclic behavior.
- We can predict interactions using last state vector.
- In future, extension of simple random walk i.e Random Walk with Restart probability may give good results.
- Interaction or link prediction is better with RWR.

References

- Emanuel Gonçalves, Joachim Bucher, Anke Ryll, Jens Niklas, Klaus Mauch, Steffen Klamt, Miguel Rocha, and Julio Saez-Rodriguez, *Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models*, Molecular BioSystems **9** (2013), no. 7, 1576–1583.
- Hawoong Jeong, Sean P Mason, Albert-Laszlo Barabasi, and Zoltan N Oltvai, *Lethality and centrality in protein networks*, arXiv preprint cond-mat/0105306 (2001).
- Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási, *Hierarchical organization of modularity in metabolic networks*, science **297** (2002), no. 5586, 1551–1555.
- Wikipedia, *Connected component (graph theory) — wikipedia, the free encyclopedia*, 2017, [Online; accessed 14-September-2017].
- ———, *Docking (molecular) — wikipedia, the free encyclopedia*, 2017, [Online; accessed 13-September-2017].

ANY
QUESTIONS
...