Chemical Space Networks using Tanimoto Coefficient

Subarna Palit

M.Sc. Life Science Informatics

13.12.2016

Chemical Space

- Chemical space is of paramount importance in the field of Chemo-informatics and drug design
- To assess a set of target compounds for their molecular properties and associate chemical relationships with biological activities
- Distribution of all synthetically accessible molecules across a hypothetical multi-dimensional space
- Two major approaches

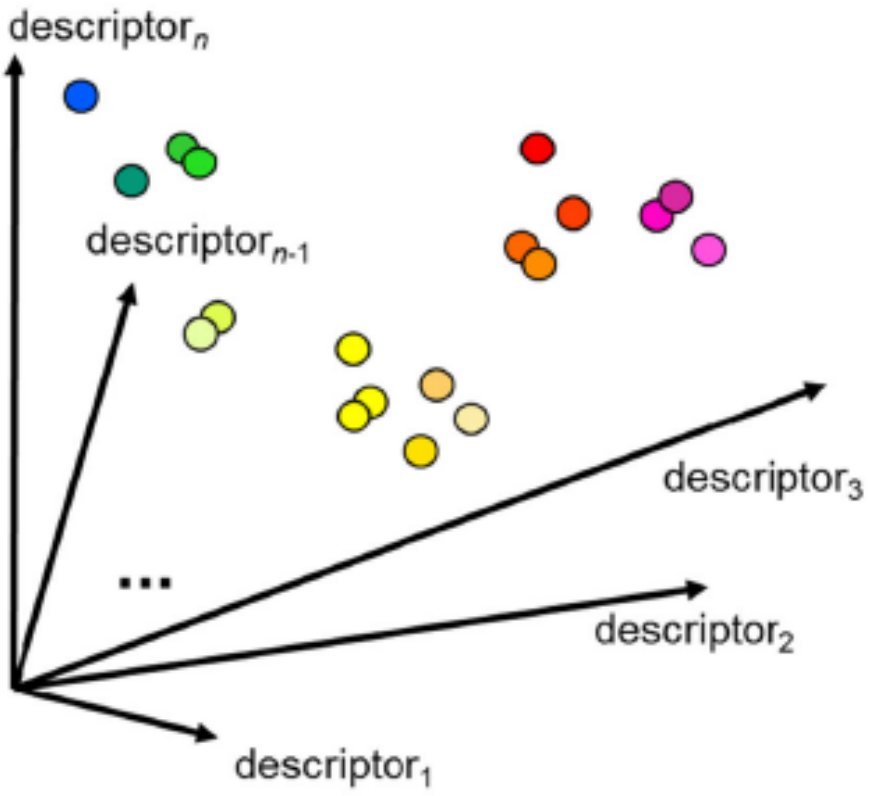
Coordinate-based	**Coordinate-free**

Vogt M, Stumpfe D, Maggiora GM, Bajorath J. Lessons learned from the design of chemical space networks and opportunities for new applications. J Comput Aided Mol Des. 2016 Mar;30(3):191-208 Accessed on 29.11.2016

Coordinate-based Chemical Space Representation

+ Traditionally molecular descriptors are used as coordinates
+ Increasing distance between molecules correlates with dissimilarity

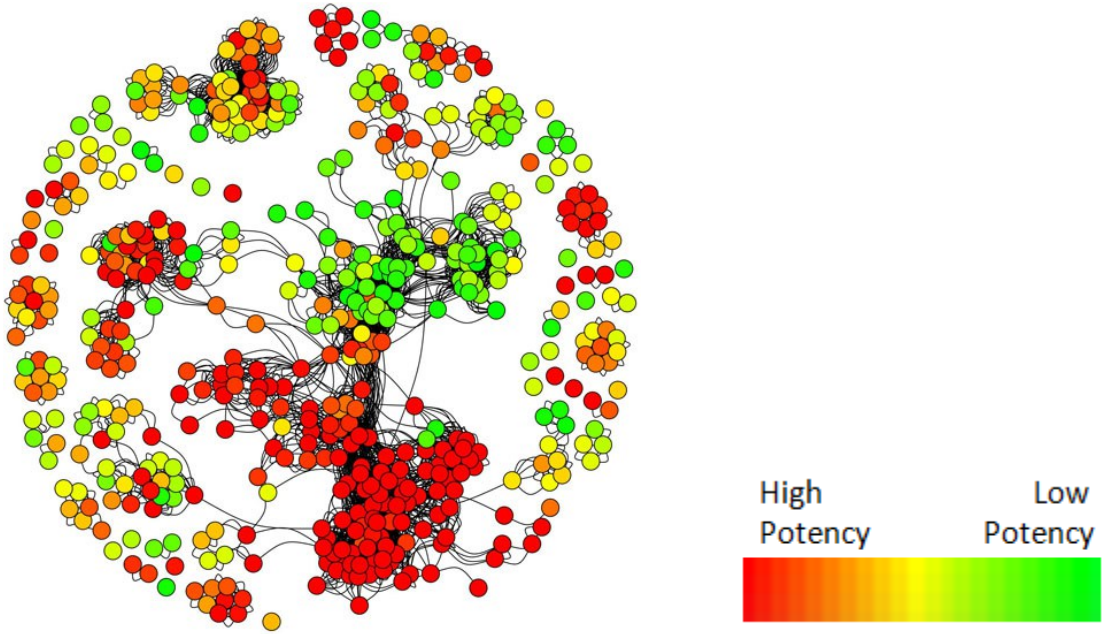
– High-dimensional making it difficult to visualize; dimensionality reduction leading to loss of vital chemical information



Vogt M, Stumpfe D, Maggiora GM, Bajorath J. Lessons learned from the design of chemical space networks and opportunities for new applications. J Comput Aided Mol Des. 2016 Mar;30(3):191-208 Accessed on 29.11.2016

Coordinate-free Chemical Space Network

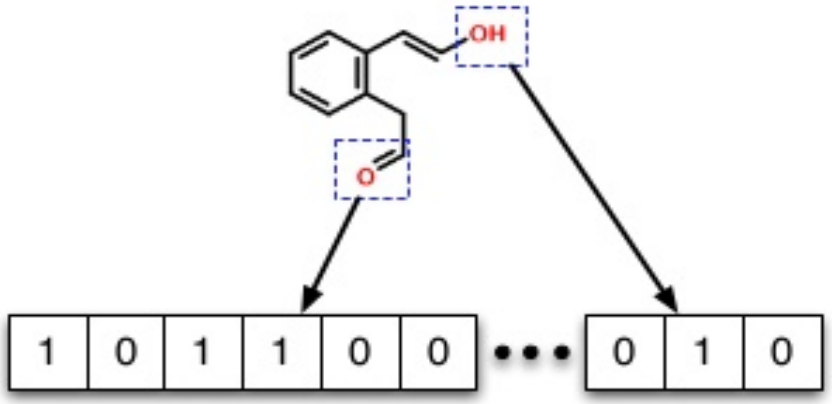
- Network-based representation of biologically relevant chemical space
  - nodes represent molecules, edges represent pair-wise molecular similarity relationships
- CSN for a set of active compounds



Vogt M, Stumpfe D, Maggiora GM, Bajorath J. Lessons learned from the design of chemical space networks and opportunities for new applications. J Comput Aided Mol Des. 2016 Mar;30(3):191-208 Accessed on 29.11.2016

Fingerprint-based similarity

- Fingerprints are a way of encoding the structure of a molecule
- Most common being a series of bits representing the presence (1) or absence (0) of particular substructures in the molecule
- Tanimoto coefficient is most widely used for binary fingerprints



Tanimoto Coefficient

A	1	0	1	1	0	1
B	1	1	0	1	0	0
C	1	0	0	1	0	0

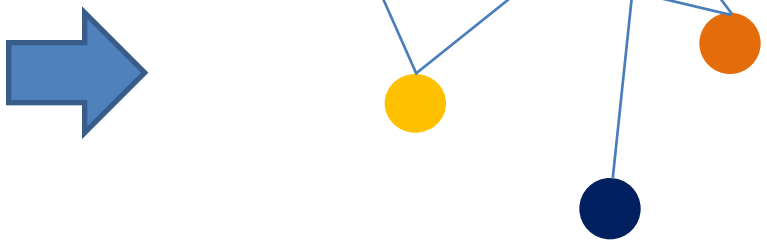
$$Tc = \frac{C}{A+B-C}$$

A = bits set to 1 in structure A, B = bits set to 1 in structure B, C = number of 1 bits common to both
Range is 0 to 1

Vogt M, Stumpfe D, Maggiora GM, Bajorath J. Lessons learned from the design of chemical space networks and opportunities for new applications. J Comput Aided Mol Des. 2016 Mar;30(3):191-208 Accessed on 29.11.2016

Tanimoto similarity threshold values (THR-CSN)

	A	B	C	D	E
A	1	T_{AB}	T_{AC}	T_{AD}	T_{AE}
B		1	T_{BC}	T_{BD}	T_{BE}
C			1	T_{CD}	T_{CE}
D				1	T_{DE}
E					1



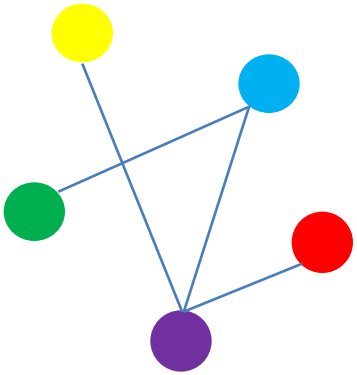
As threshold values are varied each value yields a different so-called threshold CSN (THR-CSN)

Vogt M, Stumpfe D, Maggiora GM, Bajorath J. Lessons learned from the design of chemical space networks and opportunities for new applications. J Comput Aided Mol Des. 2016 Mar;30(3):191-208 Accessed on 29.11.2016

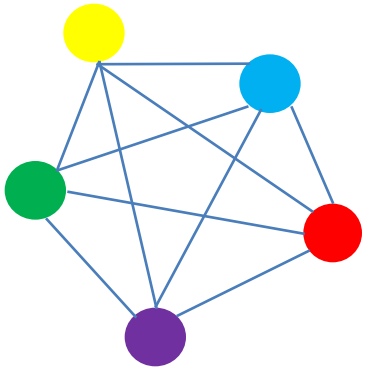
Edge Density

- Network density is defined as the ratio of actual edges in a network over all potential edges
- While a low density network with many singletons is not very informative, highly dense network get more difficult to interpret

LOW



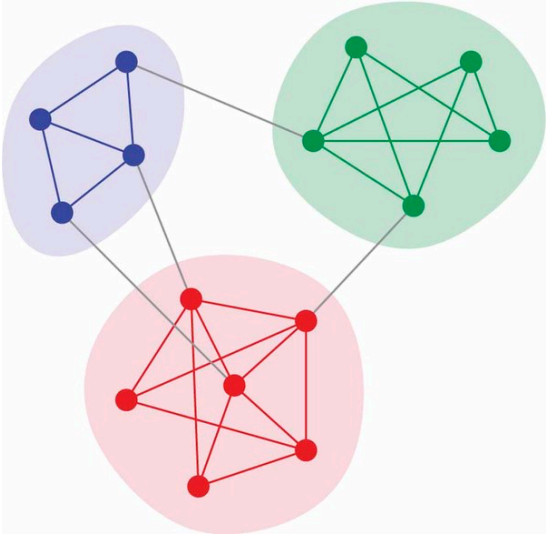
HIGH



Vogt M, Stumpfe D, Maggiora GM, Bajorath J. Lessons learned from the design of chemical space networks and opportunities for new applications. J Comput Aided Mol Des. 2016 Mar;30(3):191-208 Accessed on 29.11.2016

Community Structure and Modularity

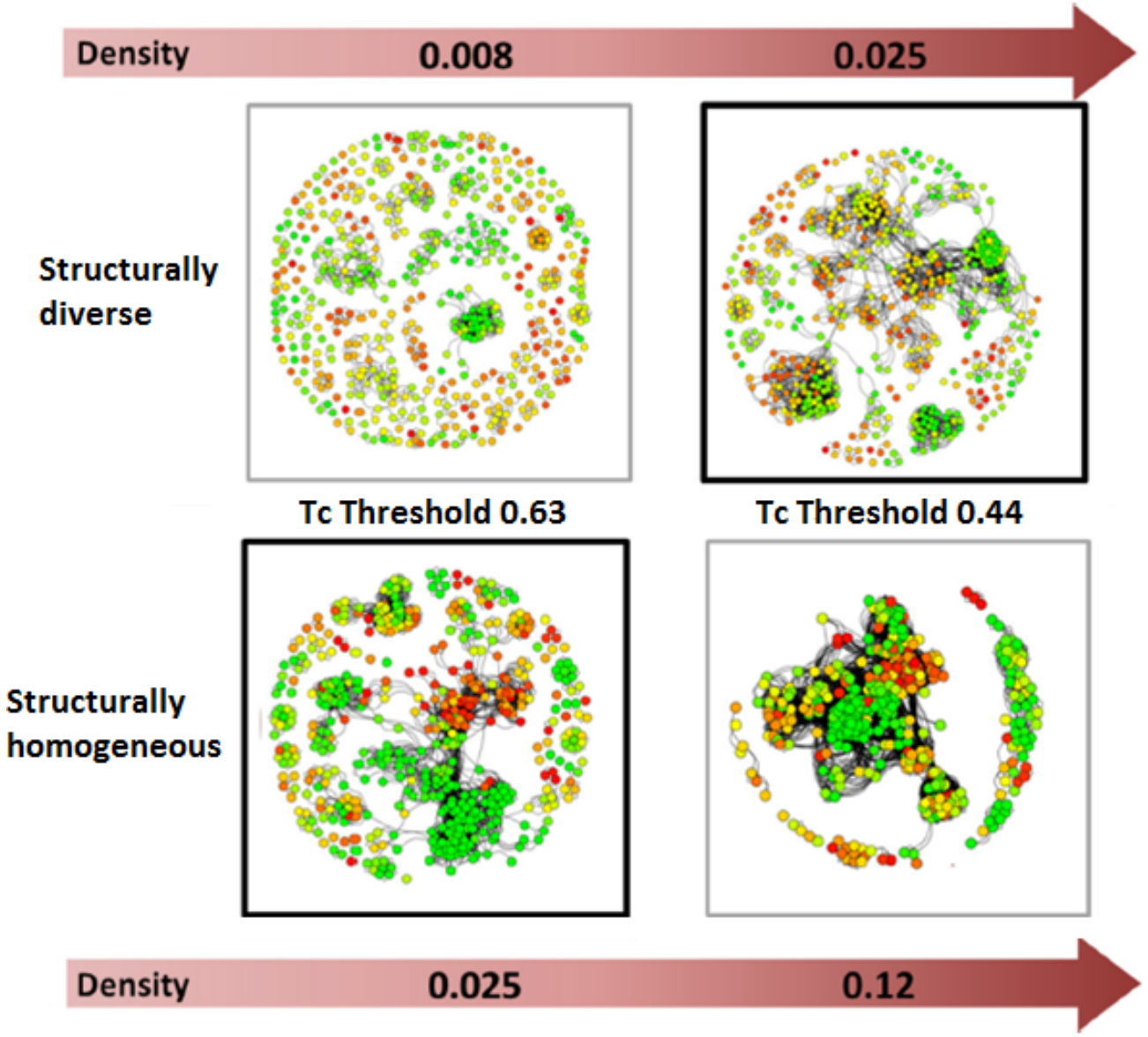
- Communities are groups of nodes in a network
  - more links between nodes within the same community
  - lesser links between nodes in different communities
- Network modularity generally correlates with community structures accounting for the nodes separated into well-defined modules



https://data.graphstream-project.org/talks/CSSS2012/gs-communities.html#/home Accessed on 29.11.2016

THR-CSNs

At low edge densities, CSNs have been found to display well-resolved topologies which can be compared



Zwierzyna M, Vogt M, Maggiora GM, Bajorath J (2015) Design and characterization of chemical space networks for different compound data sets. J Comput Aided Mol Des 29:113–125 Accessed on 29.11.2016

Discussion

- Useful in SAR exploration – community structures leading to differences between similarity relationships
- Display and analysis of only moderately sized and biologically relevant compound sets

Zwierzyna M, Vogt M, Maggiora GM, Bajorath J (2015) Design and characterization of chemical space networks for different compound data sets. J Comput Aided Mol Des 29:113–125 Accessed on 29.11.2016

THANK YOU !