

Fingerprint Kernels for Protein Structure Comparison (J. Mol. Inf., 2012)

Thomas Fober, Marco Mernberger, Gerhard Klebe, and Eyke Hüllermeier

Presenter: Jose Lugo-Martinez
Lab Meeting

October 8, 2012

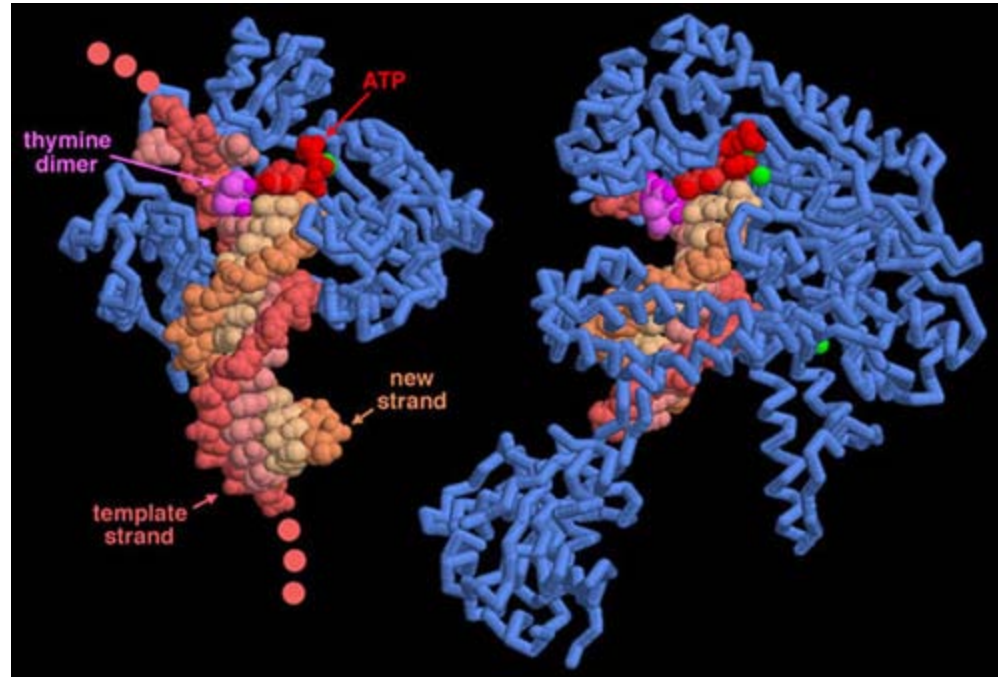


SCHOOL OF INFORMATICS
AND COMPUTING

INDIANA UNIVERSITY
Bloomington



Motivation



Goal: Find a meaningful similarity measure for protein structures comparison

TGFKBML



Graph Kernels

- Define kernels on pair of graphs

$$k(G, G') = \langle \Phi(G), \Phi(G') \rangle$$

$k(G, G')$ – measure of similarity between G and G'

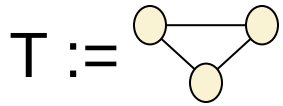
- Kernel Matrix K , where $K_{ij} = k(G_i, G_j)$ for $1 \leq i, j \leq n$

– Properties of K

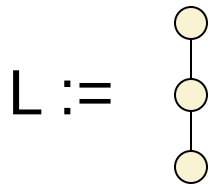
- I. Symmetric
- II. Positive semi-definite



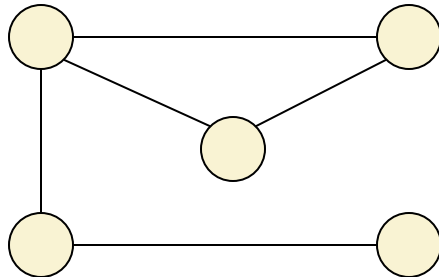
Graph Kernel Example



$$\Phi(G) = (\#(T), \#(L))$$

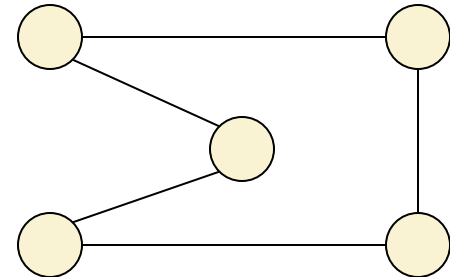


G_1



$$\Phi(G_1) = (1, 3)$$

G_2

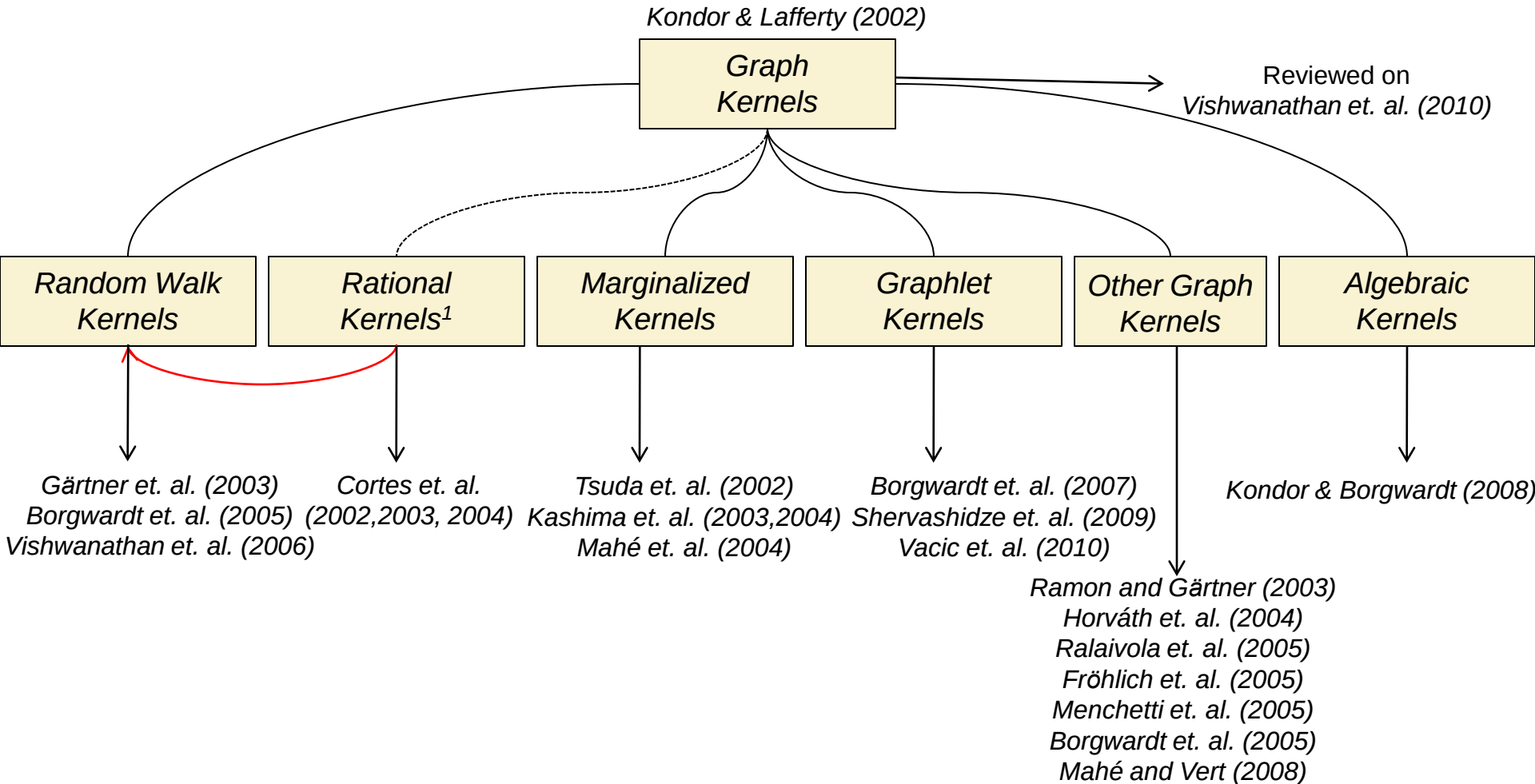


$$\Phi(G_2) = (0, 5)$$

$$k(G_1, G_2) = \langle \Phi(G_1), \Phi(G_2) \rangle = 15$$



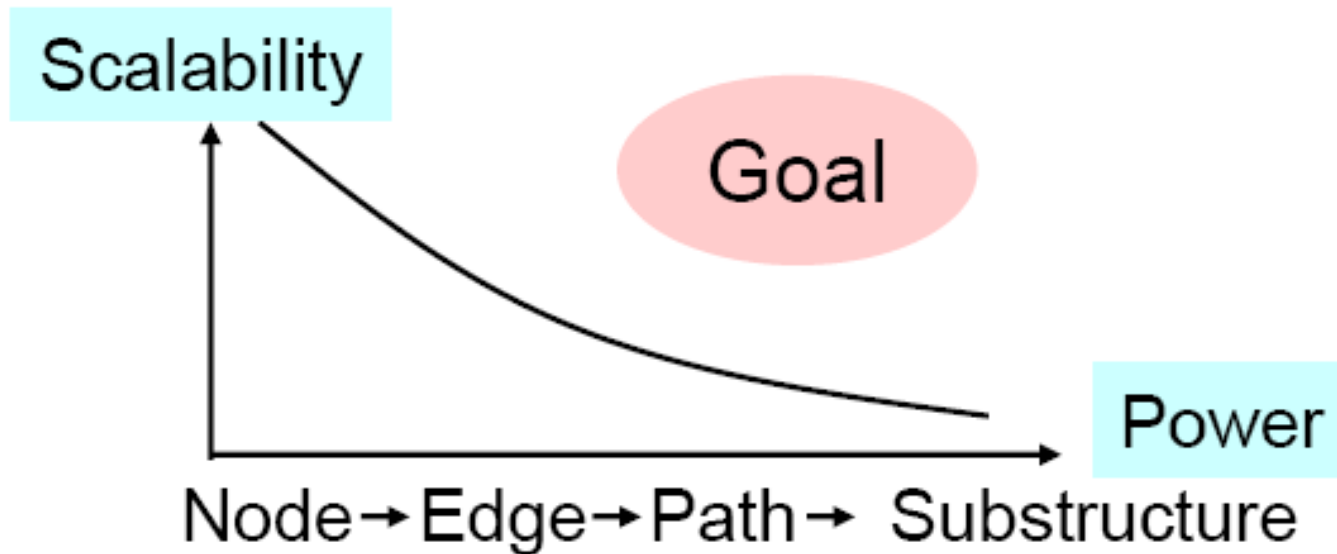
Graph Kernels Research Efforts



1. Certain Rational Kernels when specialized to graphs reduced to Random Walk Graph Kernel (*Vishwanathan et. al. 2010*)



Question: How to scale up graph kernels to vertex-labeled and edge-weighted graphs?



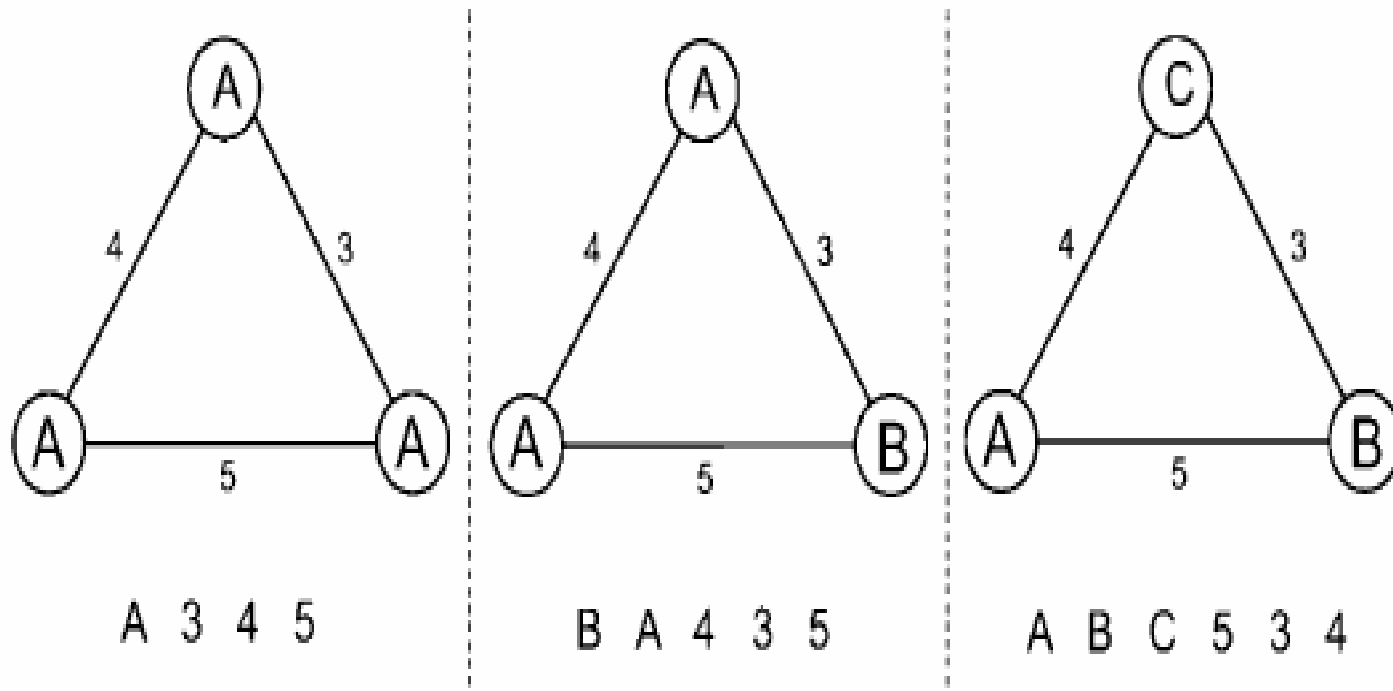


Modeling protein binding sites as graphs

- Geometry of a binding site is internally represented by a set of ***pseudocenters*** (7 types)
 - representing physicochemical properties of a surface patch within binding site
- pseudocenters are nodes, labeled using type
- distance between pseudocenters are edges weights



Example





Paper Contributions

- Extended *random walk* and *shortest path* kernels
- Novel kernel method, *fingerprint kernel*
- Use of *fuzzy-patterns*, thus, *fuzzy fingerprint kernel*



Fingerprint Kernel

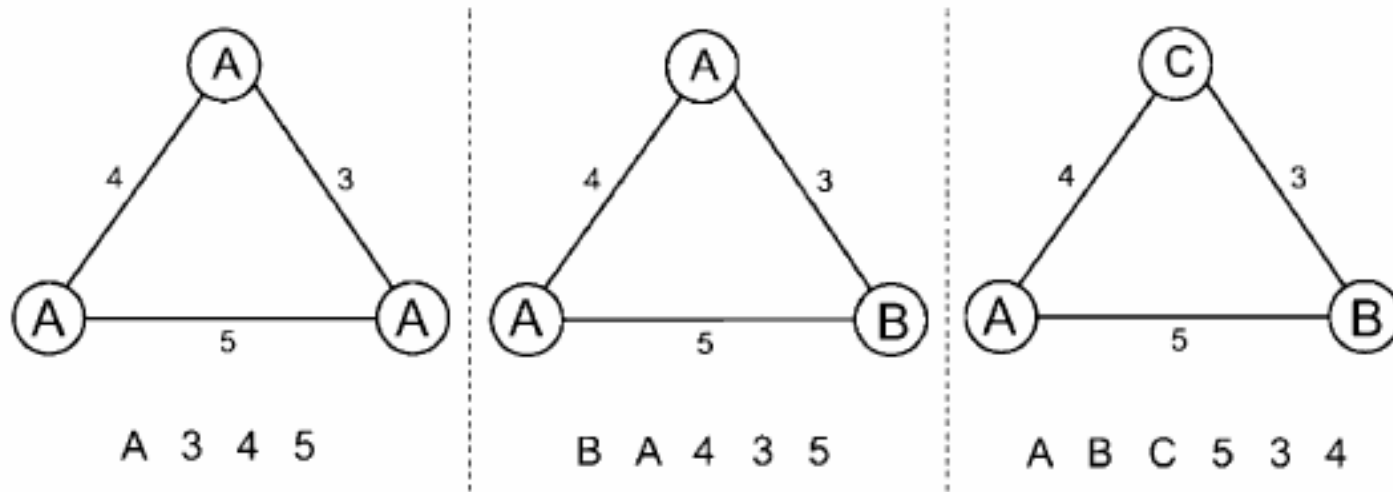
- **Idea:** Create a fingerprint vector of fixed length by counting all non-isomorphic graphs of size 3!
 - Sounds familiar?!?!?
- Count all non-isomorphic graphs of size 3 for n distinct node and k distinct edge labels

$$k_{FPH}(G, G') = \frac{1}{N(n, k)} \sum_{i=1}^{N(n, k)} \kappa_{\delta}([f_G]_i, [f_{G'}]_i)$$

$$k_{FPJ}(G, G') = \frac{\sum_{i=1}^{N(n, k)} \min([f_G]_i, [f_{G'}]_i)}{\sum_{i=1}^{N(n, k)} \max([f_G]_i, [f_{G'}]_i)}$$



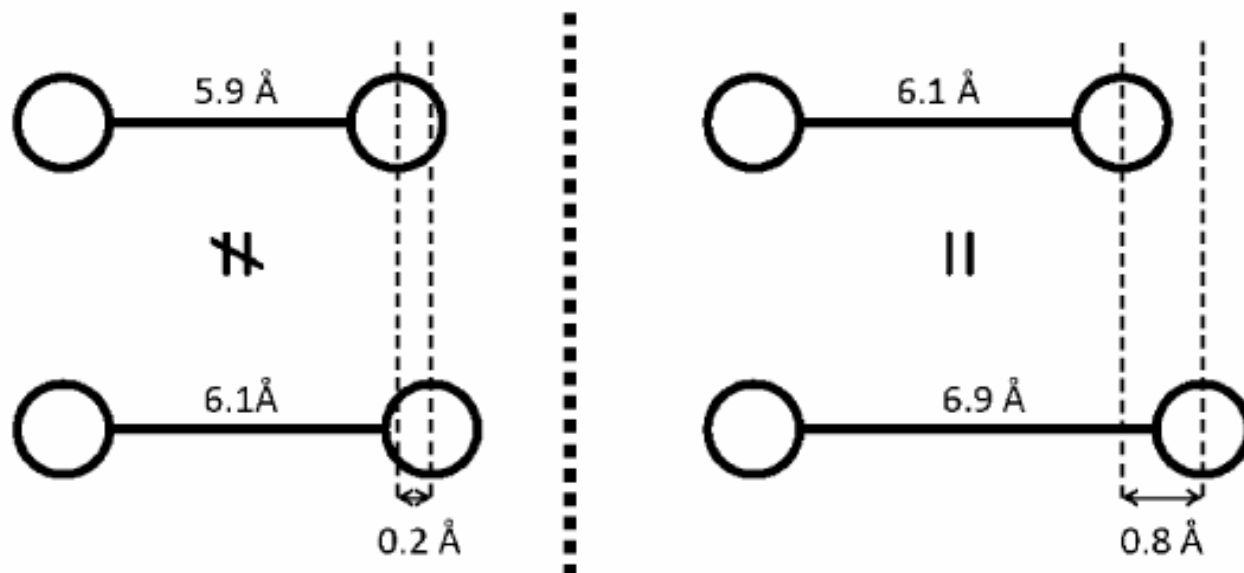
Let's do some combinatorics!!!



$$N(n, k) = \binom{n}{3} \cdot k^3 + n(n-1) \cdot k \cdot \binom{k+1}{2} + n \cdot \binom{k+2}{3}$$

Fuzzy Fingerprint Kernel

- Why?
 - problem of discontinuity on bin-boundaries





Solution

- Create a fuzzy partition of a domain X defined by a finite family of fuzzy subsets $F_i(x)$, x in X as:

$$F_i(x) = \max\{0, 1 - |x - i|\}$$

- $F_i(x)$ can be interpreted as the fuzzy subset of numbers “approximately equal to i ”



INDIANA UNIVERSITY

Thank You!

QUESTIONS?