Cluster based prediction of SH2 domain-peptide interactions using Graph Kernel

Vasumathi Jayakumar - 3210535

Department of Bioinformatics, University of Freiburg
Supervised by : Kousik Kundu

June 28, 2013

Overview

- Introduction
- Motivation
- Researches and Results
- Our research
- Result
- Conclusion

Introduction

Introduction

- Protein-protein interactions

Introduction

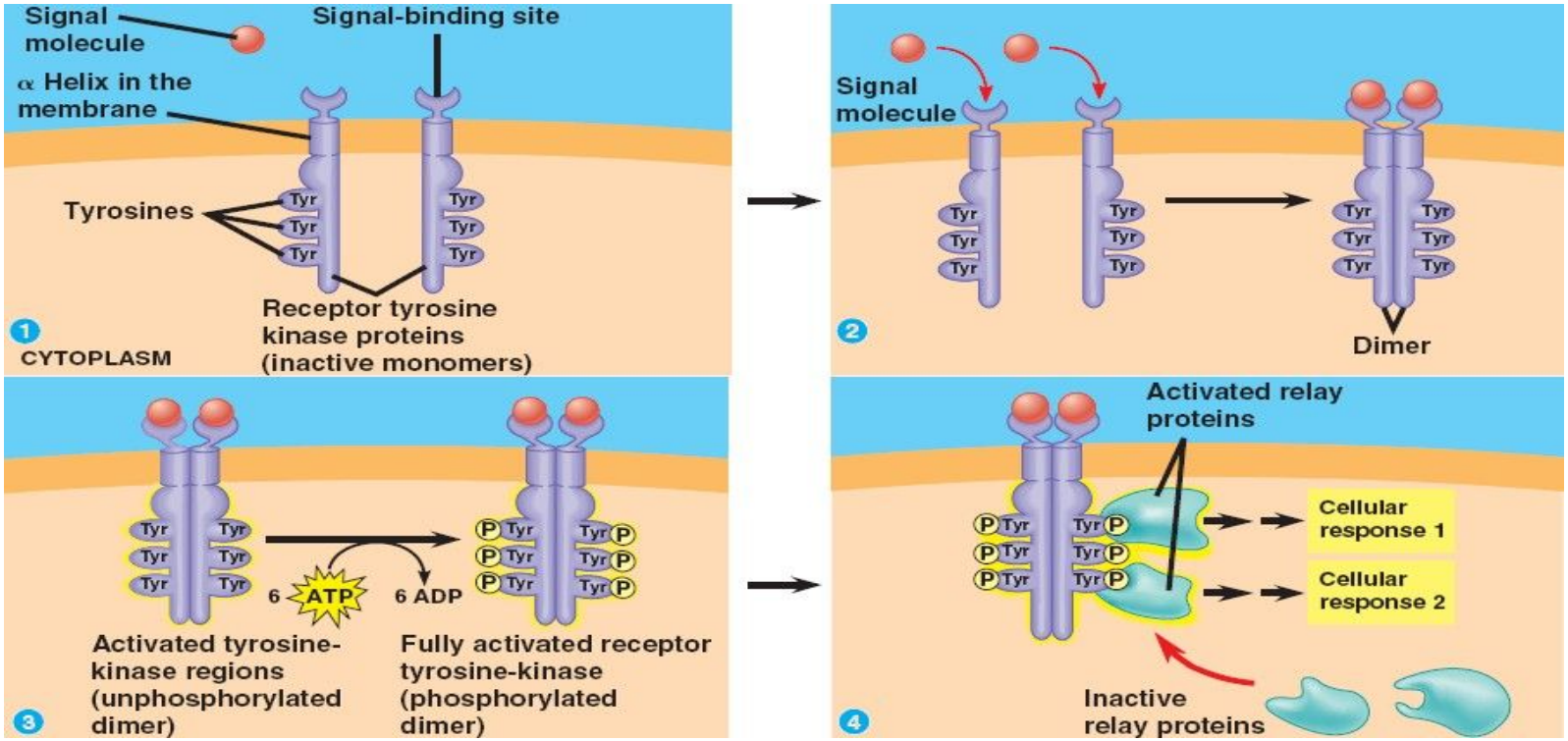
- Protein-protein interactions
- **cellular processes** - signalling, Cell communication, etc.

Introduction

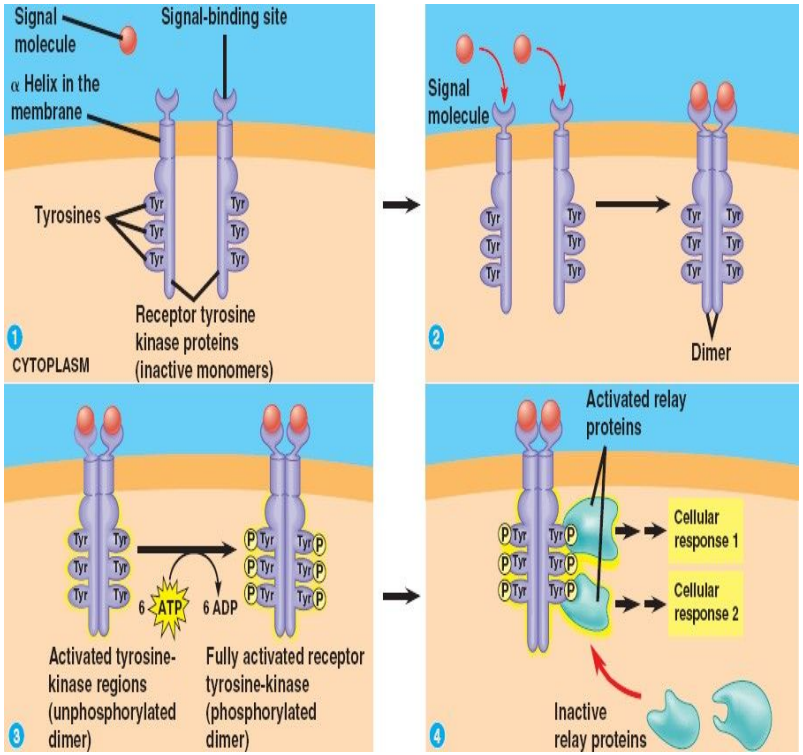
Introduction

- **PRMs**
  -Peptide-recognintion modules

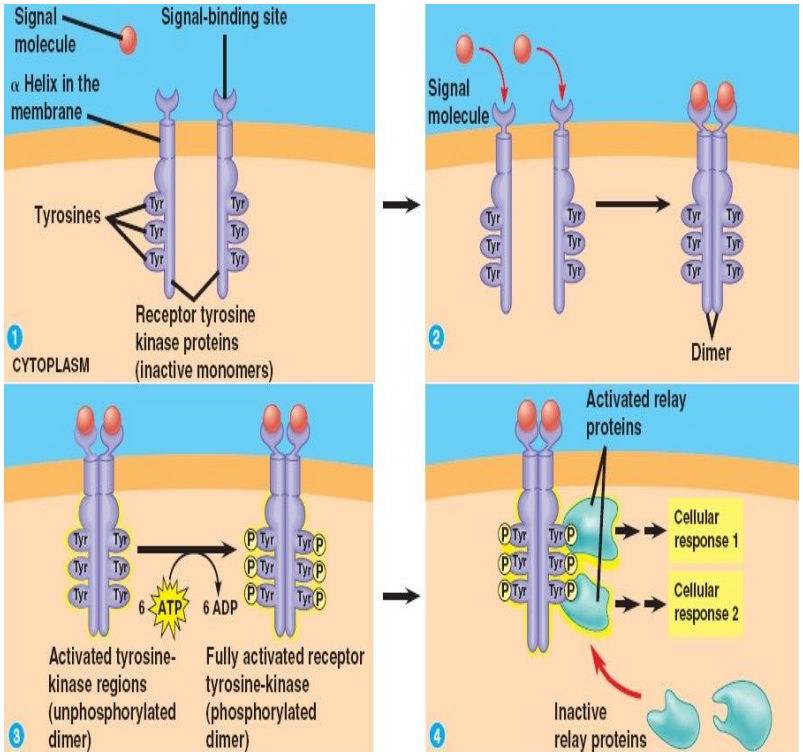
Receptor tyrosine kinases(RTKs)



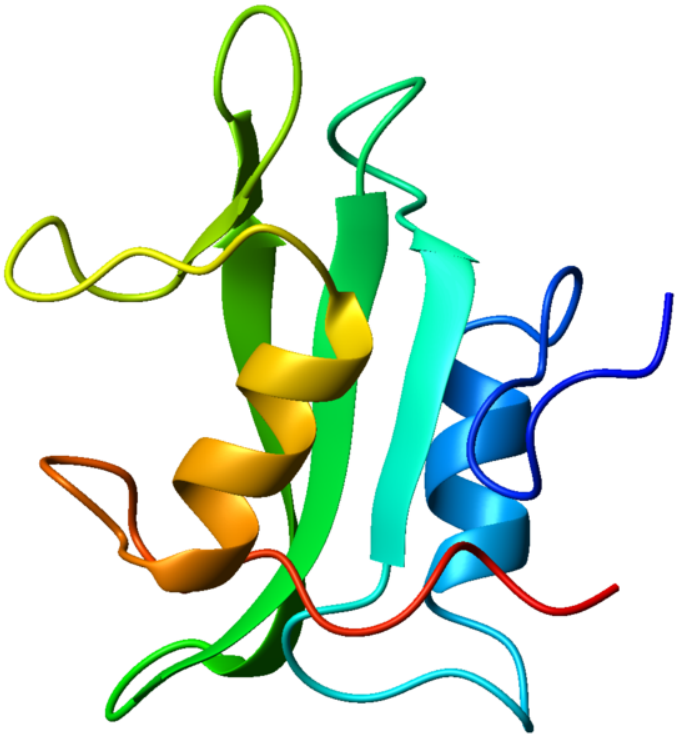
Introduction

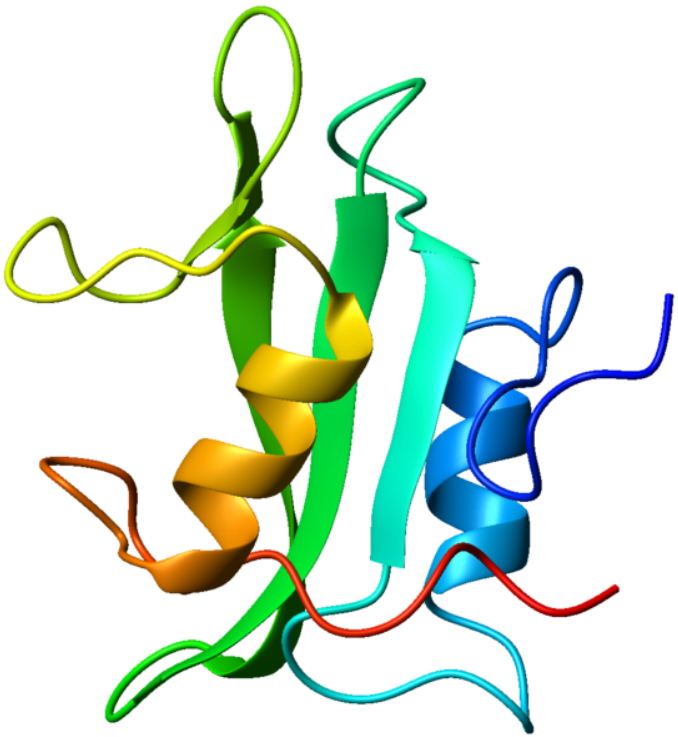


Introduction



- Receptor tyrosine kinases(RTKs)
  - Src homology 2 (SH2)
  - Peptide tyrosine binding (PTB)





- main for cellular communication
- found in intracellular signal transducing proteins
- Large beta sheet flanked by two alpha-helices
- 120 SH2 domain in 110 human proteins
- Binds with distinct phosphopeptides.
- Domain mutation causes many human disease

Motivation

Previous Researches

- Scansite
- SMALI
- Dompep

Previous Researches

Scansite

- Most popular tool, *Yaffe et. al.* in 2003
- Based on position specific scoring matrices(PSSMs)
- Derived from chemically synthesized peptide array libraries

SMALI

- SMALI - Scoring matrix-assisted ligand
- Recent approach, *Li et. al.* in 2008
- Based on (PSSMs)
- Derived from OPAL (oriented peptide array libraries)

Dompep

- More recent approach, *Li et. al.* in 2011
- Based on linear SVM (support vector machine)
-

Results of researches

- Position take important role in binding
- Used inear models - Complex dependencies between amino acids cannot be reflected
- Uses only positive interactions

Other approaches

- Uses structural information of SH2-peptide complex and
- Energy models derived from the structure
- A few approach - CoMFA, FoldX algorithm
- Computationally very expensive
- Depends of solved structures - available for few SH2-peptide complexes

Our approach

- Non-linear models
- Graph kernel approach
- Considered negative interactions

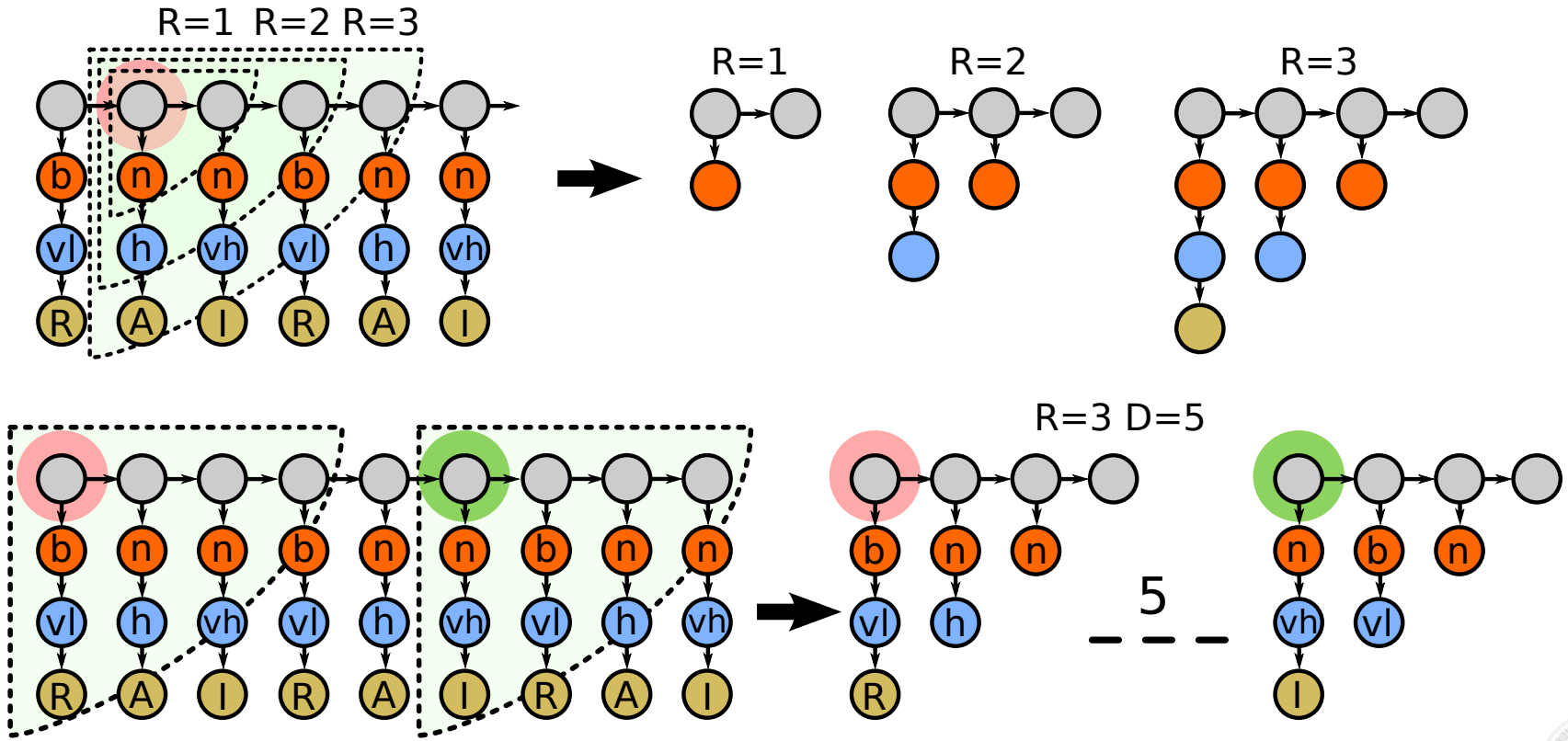
Graph Kernel Approach

- Computation of similarity measure between graphs in terms of a dot prodcut function – Graph kernel
- *Costa and Grave, 2010* - Neighborhood Subgraph Pairwise Distance Kernel(NSPDK)

NSPDK

- An instance of *decomposition kernel*
- A composite kernel operates over all possible "parts"
- Parts - *"neighborhood subgraphs"*
- Increasing radii $r < r_{max}$
- Distance not greater than d_{max}

NSPDK



Data Collection

- Microarray Dataset I (positive and negative)
- Microarray Dataset II (positive and negative)
- Netphores Dataset (positive)
- Positive interactions - 1098
- Needleman Wunsch alignment - SH2 domains
- MCL clustering of alignment - isolatoin value
- Identity $>= 60\%$
- Mafft alignment - SH2 domains
- Interactive Tree Of Life (ITOL)
- Finalized the clusters

Machine Learning

- Divide data set
- 75% - training set
- 25% - test set
- Used tool **EDeN**
- Find *Optimal parameter* - 5 fold *Cross Validation*
- Model 75% training set with optimal parameter
- Test 25% test set over the models
- Calculate performance using *Perf*
- Result
- Interactive Tree Of Life (ITOL)
- Finalized the clusters

Result

- Calculate performance using **Perf**
- Sensitivity, Specificity, Precision, AUC Precision, AUC ROC