Protein-Protein Interaction Network

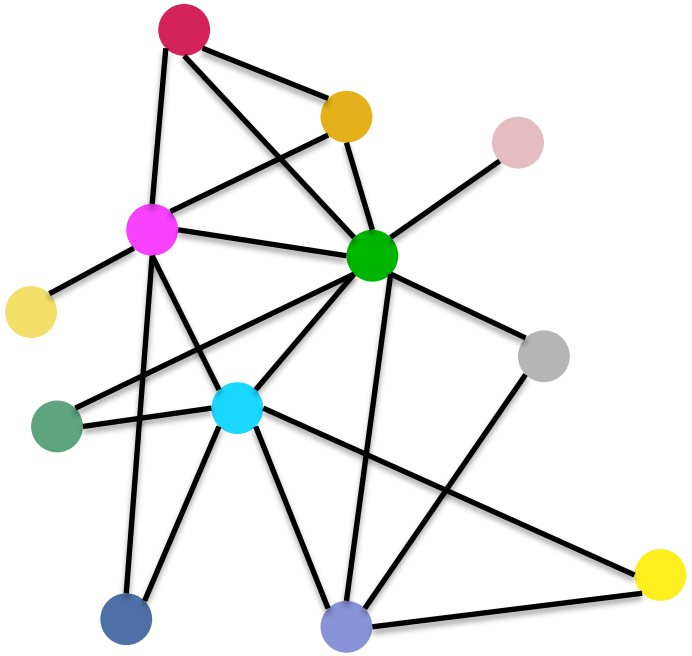
Lecture 2

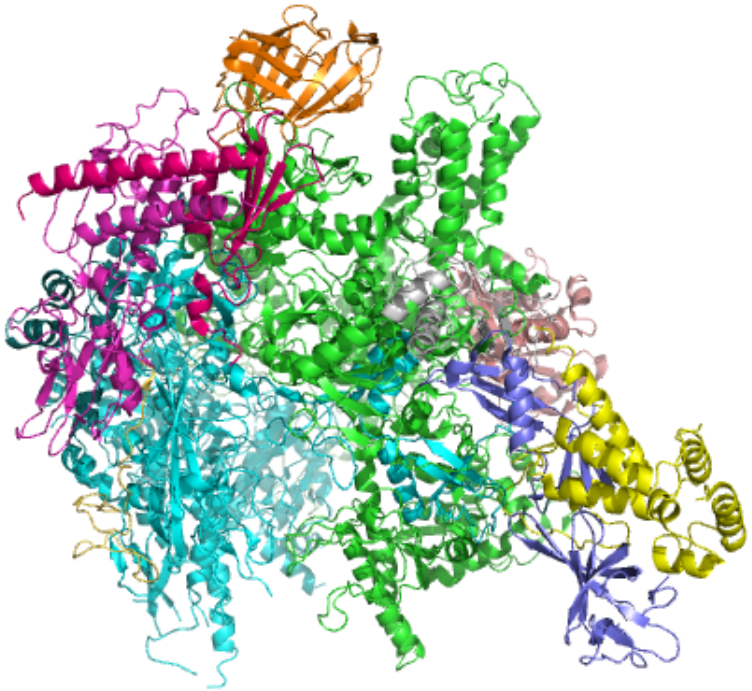
Recap of last class

- Protein-Protein Interaction Model
- How to get PPI
  - Experimental methods
  - Bioinformatic methods
- PPI databases
- network properties
- Analysis method
- Integration with other omic data

Protein Interactions

- 12-subunit RNA Polymerase II





$$Q = \frac{|E|}{V(V-1)/2} = \frac{20}{6 \times 11} = 0.3$$

PDB: 2B8K

Experimental methods

- **Co-immunoprecipitation** is considered to be the gold standard assay for protein–protein interactions, especially when it is performed with endogenous (not overexpressed and not tagged) proteins.
- **Pull-down assays** are a common variation of immunoprecipitation and are used identically, although this approach is more amenable to an initial screen for interacting proteins.
- **Chemical cross-linking** is often used to "fix" protein interactions in place before trying to isolate/identify interacting proteins.
- **Yeast two-hybrid assay**
- **Tandem Affinity purification**
- **Protein microarray**
- **Phage display**

Overlap of high-throughput interaction studies is LOW

	Ito Y2H	Uetz Y2H	Gavin TAP/ms	Ho FLAG/ms
Ito 2-hybrid	4363	186	54	63
Uetz 2-hybrid		1403	54	56
Gavin affinity			3222	198
Ho affinity				3596
Small scale	442	415	528	391

data from Salwinski & Eisenberg, Current Opinion in Structural Biology (2003) 13, 377-382

Outline

- Protein-Protein Interaction Model
- How to get PPI
  - Experiments: Y2H, MS, etc. (Assessing and filtering high throughput interaction data)
  - Bioinformatics
- PPI databases and network properties
- Analysis method
- Integration with other omic data

High throughput interaction data

- Not reliable
- Noisy

- Computational methods for improving the quality of interaction data
  - Assessment and validation

Assessing and filtering Criteria

- Promiscuity criteria
- Overlap criteria
- Topology criteria

Assessing and filtering Criteria

- Promiscuity criteria
  - In most high-throughput interaction studies, a few proteins are observed to interact promiscuously. Generally these are removed from the analysis.
  - Problem: some interactions may be real!
- Examples:
  - Using TAP/MS even without a bait, 17 proteins were found in pull-downs by Gavin et al. 49 other proteins found to have a similar frequency of interaction to these false positives were thrown out.
  - Using Yeast 2-hybrid, proteins were observed to make many interactions in many screens usually discarded as probably false positives.

Assessing and filtering Criteria

- Promiscuity criteria
- Overlap criteria
- Topology criteria

Assessing and filtering Criteria

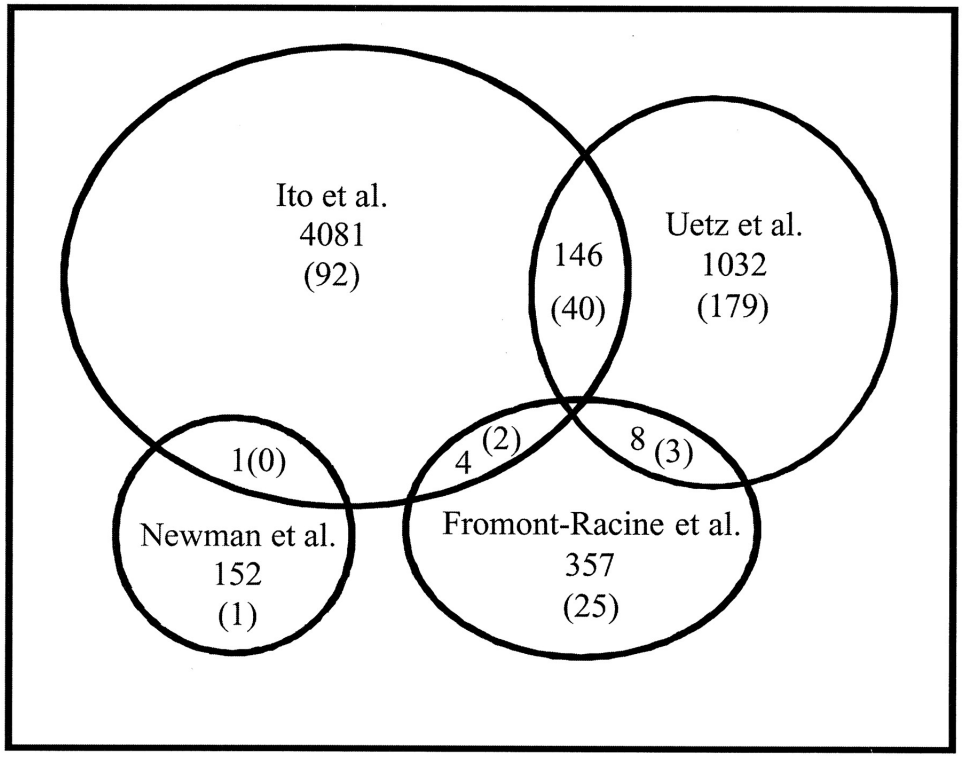
- Overlap criteria
  - An interaction has higher possibility to be real if two different types of methods discover it.
- Methods:
  - With interaction data.
  - With non-interaction data.

Assessing and filtering Criteria

With interaction data:

intersection is low!

E.g. compare Y2H and TAP/MS. Unfortunately, overlap is low.



Assessing and filtering Criteria

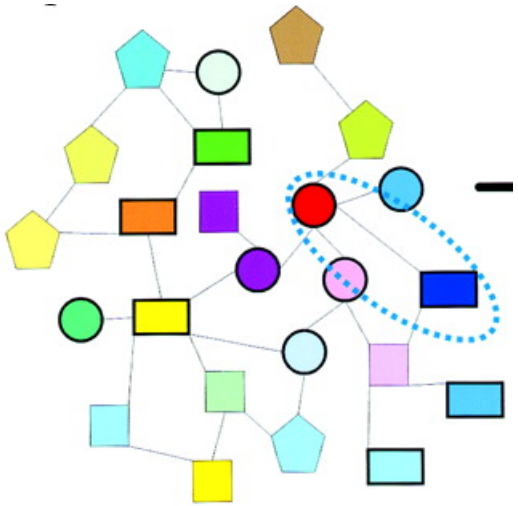
- Overlap criteria
- Methods:
  - With non-interaction data.
    - Expression Profile Reliability (EPR)
    - Homology methods -Paralogous Verification (PVM)
    - Domain Pair Verification (DPV)

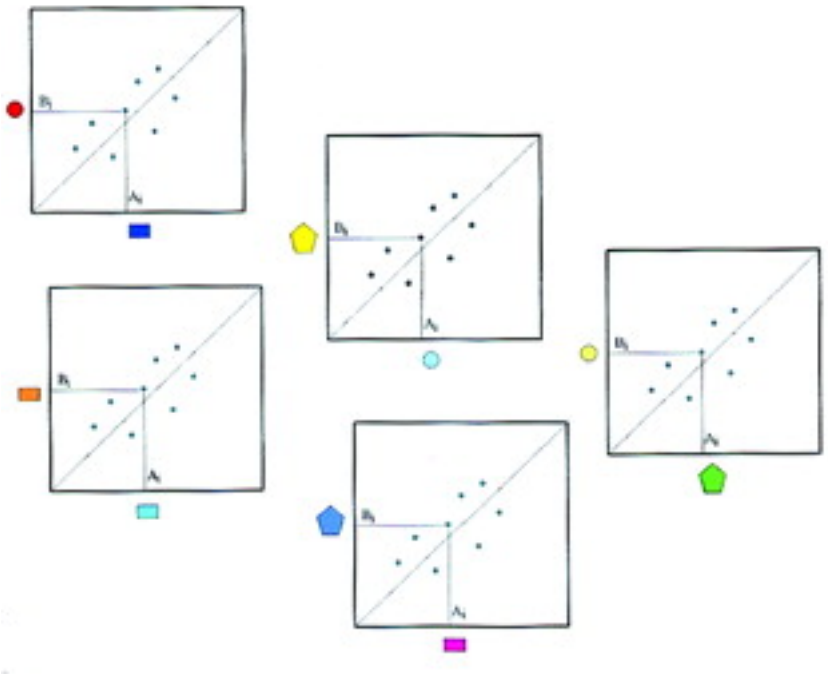
Deane et al. (2002) *Mol. Cell. Proteomics*

Expression Profile Reliability (EPR)

- Expression Profile Reliability Index (*EPR Index*) evaluates the quality of a large-scale protein-protein interaction data sets by comparing the expression profile.
- Two proteins have high possibility to interact with each other, if they co-express.

EPR

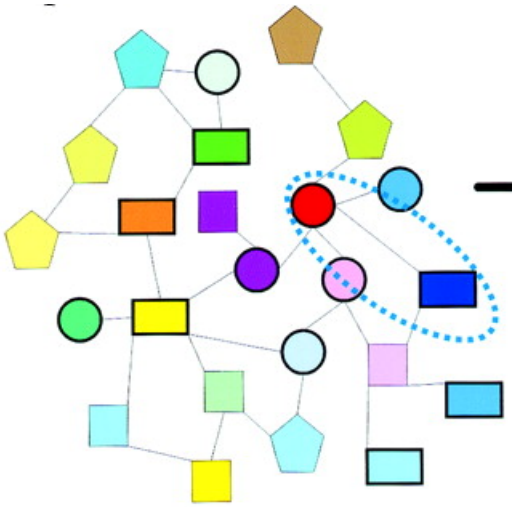


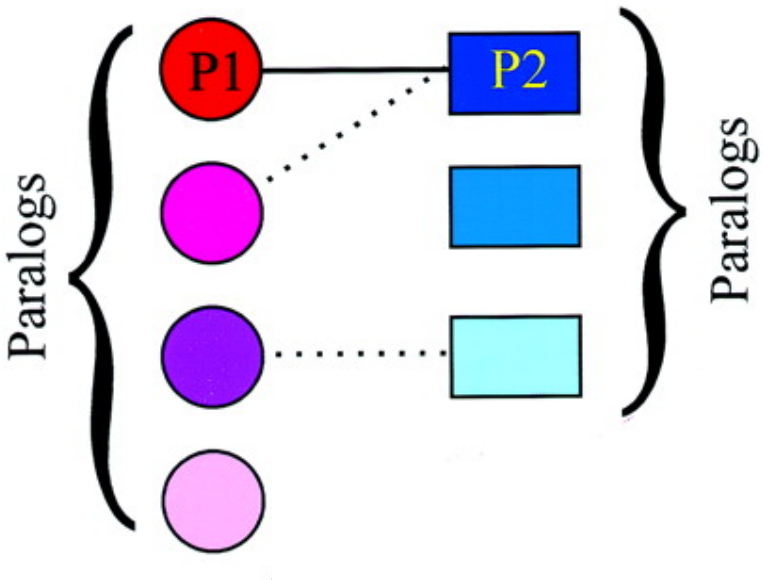


Collect the mRNA expression levels of the interaction pairs under several conditions, and calculate their expression correlations.

Deane et al. (2002) *Mol. Cell. Proteomics*

Paralogous Verification Method (PVM)





Count the number of paralogous interactions,
If the PVM score =2, they have a interaction.

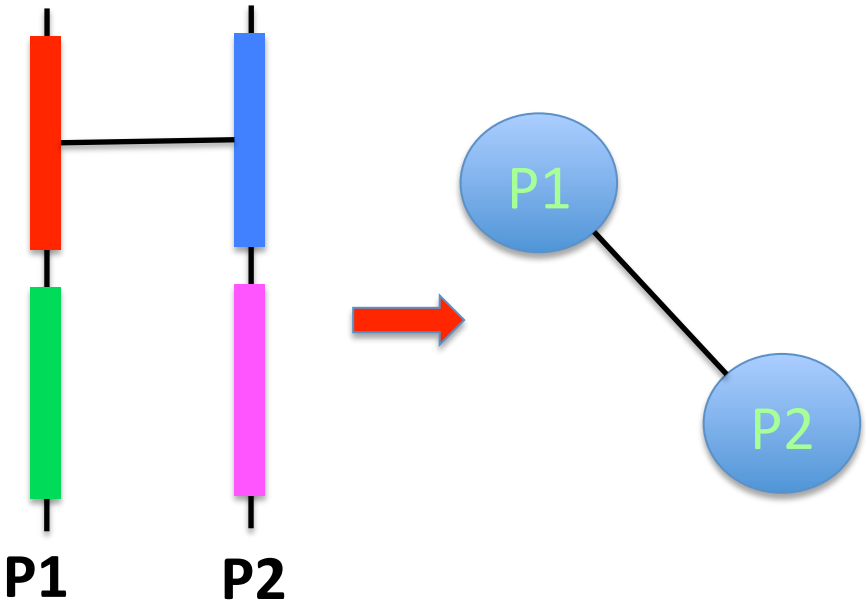
Homologous sequences are **paralogous** if they were separated by a gene duplication event: if a gene in an organism is duplicated to occupy two different positions in the same genome, then the two copies are paralogous.

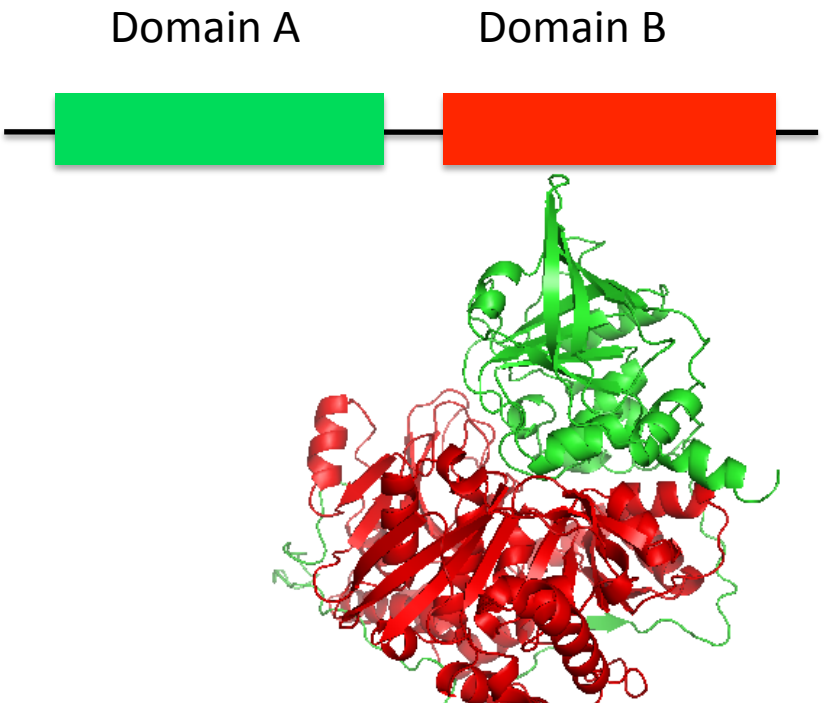
Paralogous Verification Method (PVM)

- PVM is very accurate; if a pair scores by PVM, it is almost certainly a true interaction.
- PVM does not have good coverage; it is not sensitive. PVM only confirms around 50% high-confidence samples. This is because many examples of paralogous complexes are sparse.

Domain Pair Verification (*DPV*)

- If two domains have an interaction, any two proteins that have those two domains also have interactions.
- Protein 3D structures can provide the atomic detains for protein interactions.
- The solved structures most are a single domain instead of a full length protein.



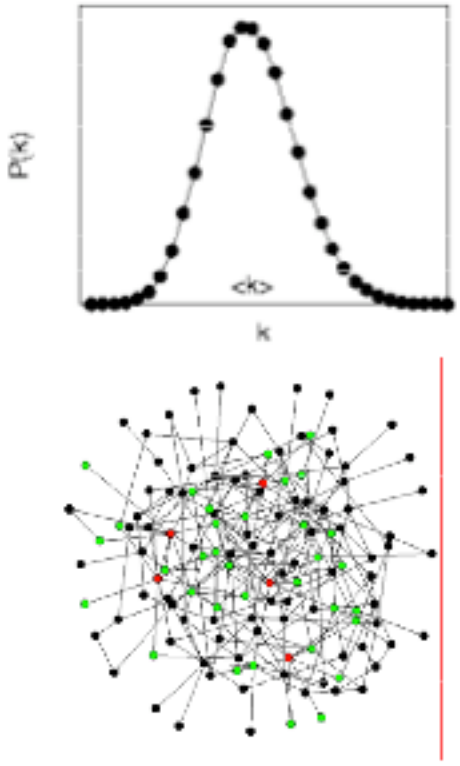


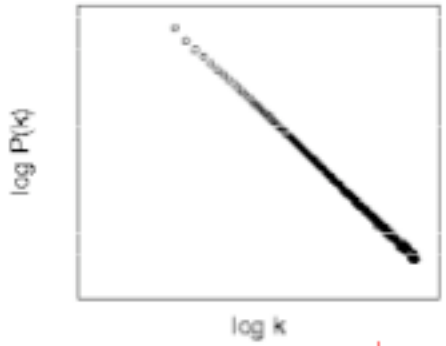
Assessing and filtering Criteria

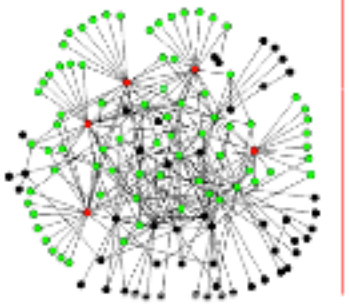
- Promiscuity criteria
- Overlap criteria
- Topology criteria

A scale free network

- Power-law degree distributions were found in diverse networks



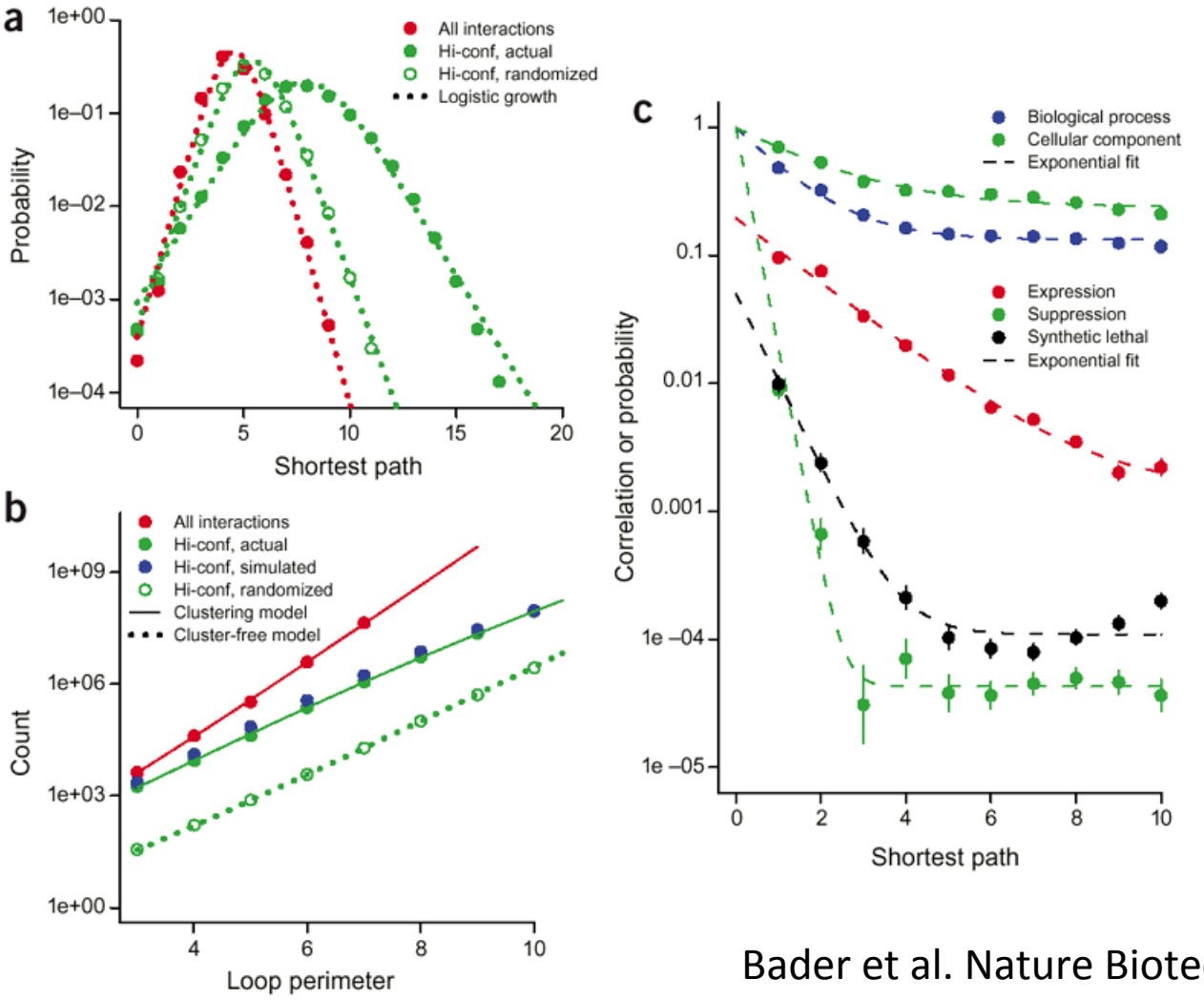




Large variability

Topology criteria

- Use information about the observed vs. expected interaction network.



Bader et al. Nature Biotechnology (2003) 22, 78-85

Outline

- Protein-Protein Interaction Model
- How to get PPI
  - Experiments: Y2H, MS, etc.
  - Bioinformatics
- PPI databases and network properties
- Analysis method
- Integration with other omic data

Why do we need bioinformatics way to generate PPI networks?

- Only model organisms have high throughput PPI data. For example, yeast and human. How about maize?
- High throughput method is expensive and time consuming.

Bioinformatics methods

- Homologous method to find Orthology
- Combination with other information, such as expression profile, GO annotations.
- Prediction
  - Sequence method
  - Structural based method
- Text mining

Orthologous proteins

- Homologous sequences are **orthologous** if they were separated by a speciation event: when a species diverges into two separate species, the divergent copies of a single gene in the resulting species are said to be orthologous.
- **Orthologs**, or orthologous genes (proteins), are genes in different species that are similar to each other because they originated from a common ancestor.

Orthology search

- Similarity search will be done using
  - BLASTP (Protein Basic Local Alignment Search Tool; Camacho 2009)
  - PSI-BLAST (Position-Specific Iterated Blast; Altschul *et al.* 1997).
  - Profile Hidden Markov Models will be generated from protein sequence databases and the search is done using HMMER3 (Eddy 1998; http://hmmer.org).

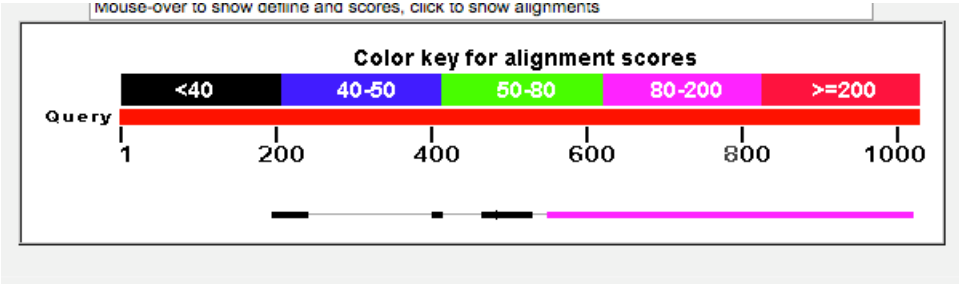
An Example

>At1g11720.1_ARATH

MAASGPKSSGPRGFGRRTTVGSAQKRTQKKNGEKDSNATSTATNEVSGISKLPAAKVDVQKQSSVVLNERNVLDRSDIEDGSDRLDKKTTDDDDLLEQKLKLERENLRRKEIETLA
AENLARGDRMFVYPVIVKPDEDIEVFLNRNLSTLNNEPDVLIMGAFNEWRWKSFTRRLEKTWIHEDWLSCLLHIPKEAYKMDFVFFNGQSVYDNNDSKDFCVEIKGGMDKVDFE
NFLLEEKLREQEKLAKEEAERERQKEEKRRIEAQKAAIEADRAQAKAETQKRRELLQPAIKKAVVSAENVWYIEPSDFKAEDTVKLYYNKRSGPLTNSKELWLHGGFNNWVDGLSIV
VKLVNAELKDVDPKSGNWWFAEVVVPGGALVIDWVFADGPPKGAFLYDNNGYQDFHALVPQKLPEELYWLEEENMIFRKLQEDRRLKEEVMRAKMEKTARLKAETKERTLKKF
LLSQKDVVYTEPLEIQAGNPVTVLYNPANTVLNGKPEVWFRGSFNRWTHRLGPLPPQKMEATDDESSHVKTTAKVPLDAYMMDFVFSEKEDGGIFDNKNGLDYHLPVVGGISK
EPPLHIVHIAVEMAPIAKVGGLGDVVTSLSRAVQELNHNVDIVFPKYDCIKHNFVKDLQFNRSYHWGGTEIKVWHGKVEGLSVYFLDPQNGLFQRGCVYGCADDAGRFGFFCHA
ALEFLLQGGFHPDILHCHDWSSAPVSWLFKDHYTQYGLIKTRIVFTIHNLEFGANAIGKAMTFADKATTVSPTYAKEAGNSVISAHLYKFHGIINGIDPDIWDPYNDNFIPVPYTSEN
VVEGKRAAKEELQNRLGLKSADFPVVGIITRLTHQKGIHLIKHAIWRTLERNGQVVLLGSAPDPRIQNDFVNLANQLHSSHGDRARLVLTYDEPLSHLIYAGADFILVPSIFEPCGLTQ
LIAMRYGAVPVVRKTGGLFDTVFDVDHDKERAQAQVLEPNGFSFDGADAPGVDYALNRAISAWYDGREWFNSLCKTVMEQDWSWNRPALEYLELY HSARK*

>GRMZM2G008263_P01_ZEAMA

MAATMGSISANGSYQTNRPSALKQAPHMQFQQCCNGGLRFLSKHSQSTRSKIQVAKRRATDNGIHPKTTGHRAPIVCSAGMTIVFVATEVHPWCKTGGLGDVVGGLPPALAA
MGHRVMTIAPRYDQYKDAWDTSVLVEVNIGDTVETVRFFHCYKRGVDRVFVDHPMFLEKVWGKTGAKLYGPTTGTDYRDNQLRFCLLCLAALEAPRVLNFNNSEYFSGPYGED
VVFVANDWHTAILPCYLKSMYKPNGIYKNAKVAFCIHNIAYQGRFARADFDLLNLPDSFLPSFDFIDGHVKPVLGRKLNWMKAGIIESDLVLTVSPHYVKELTSGPDKGVELDGVLR
TKPLEIGIVNGMDVYEWDPSTDKYISVKYDATTVTEARALNKESLQAEVGLPVDSSIPVIVFVGRLEEQKGSDILIAAIPEFVGENVQIIVLGTGKKKMEEELTQLEVKYPNNARGIAK
FNVPLAHMMFAGADFIIVPSRFEPCGLIQLQGMRYGVIPICSSTGGLVDTVEEGVTGFHMGSFNVECETVDPADVTAVASTVTRALKQYDTPAFHEMVQNCMAKDLSWKGPAK
KWEEVLLGLGVEGSRAGIDDAEEIAP LAKENVATP



Score = 165 bits (417), Expect = $1x10^{-44}$,

Identities = 158/547 (29%), Positives = 237/547 (44%),

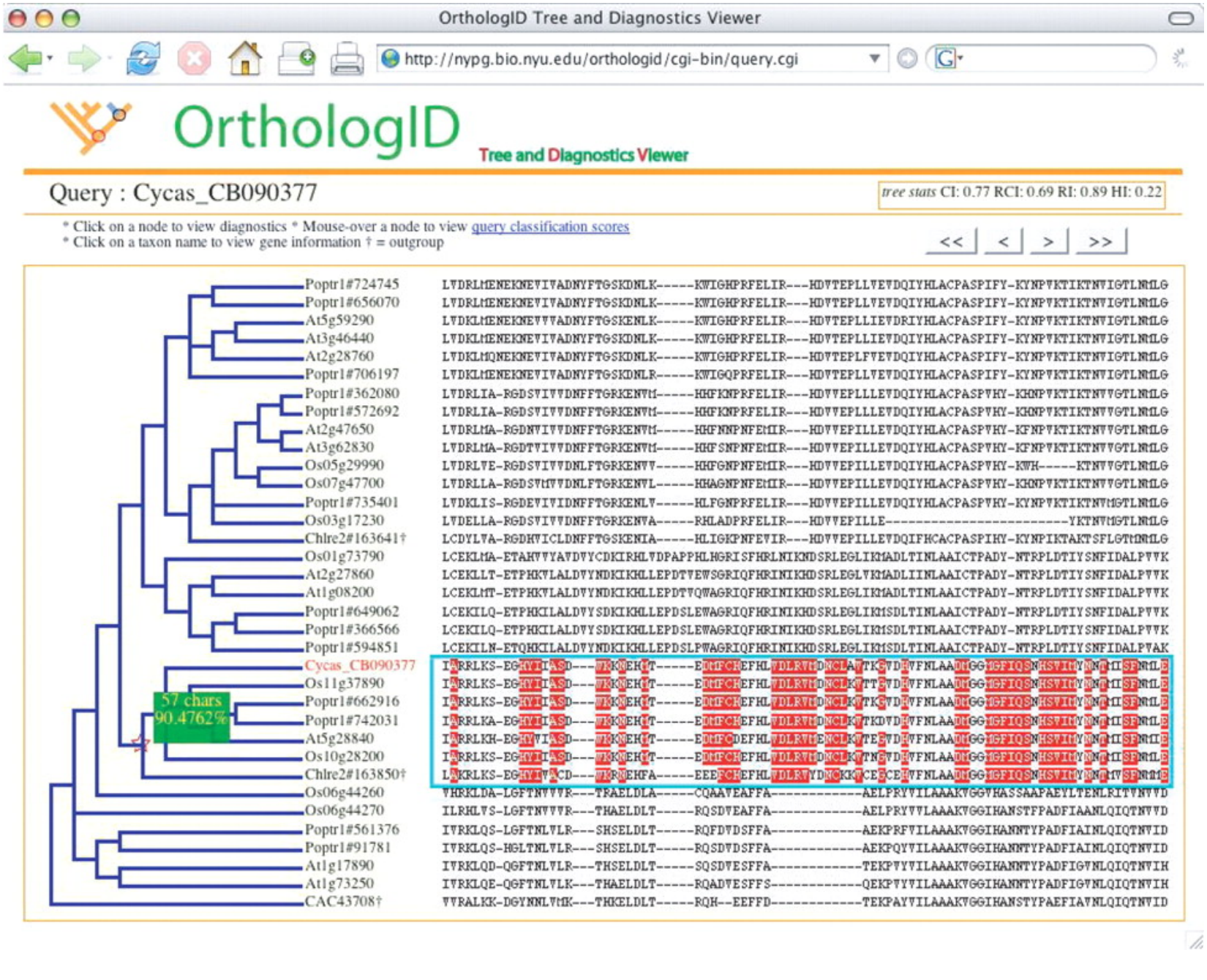
Gaps = 106/547 (19%)

Othology databases

- InParanoid (Berglund *et al.* 2008; http://inparanoid.sbc.su.se, 100 organisms: 1687023 sequences),
- OrthoMCL-DB (Chen *et al.* 2006; http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi, ortholog group predictions for 55 species)
- KEGG Orthology group (http://www.genome.jp/kegg/ko.html)

Othology databases

- OrthoMam (Ranwez *et al.* 2007; http://www.orthomam.univ-montp2.fr/orthomam/html/index.php, 36 organisms: 12777 sequences, Mammalian)
- OrthologID(Chiu *et al.* 2006; http://nypg.bio.nyu.edu/orthologid/, plants, 5 species, 137641 sequences)
- GreenPhylDB(Conte *et al.* 2007; http://greenphyl.cirad.fr, plants, 16 species,)

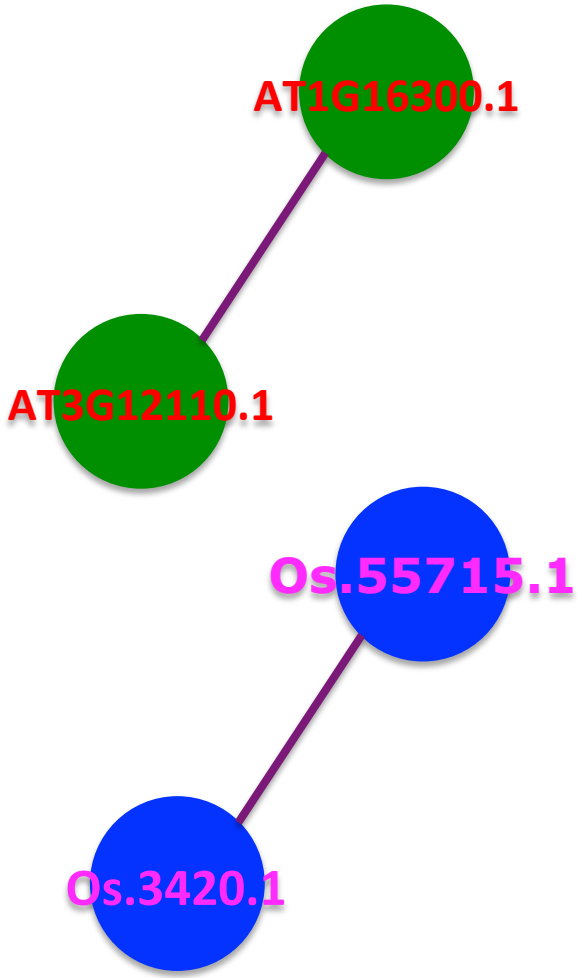

OrthologID Tree and Diagnostics Viewer
http://nypg.bio.nyu.edu/orthologid/cgi-bin/query.cgi
OrthologID
Tree and Diagnostics Viewer
Query : Cycas_CB090377
tree stats CI: 0.77 RCI: 0.69 RI: 0.89 HI: 0.22
* Click on a node to view diagnostics * Mouse-over a node to view query classification scores
* Click on a taxon name to view gene information † = outgroup
<< < > >>
57 chars
90.4762%
Poptr1#724745
Poptr1#656070
At5g59290
At3g46440
At2g28760
Poptr1#706197
Poptr1#362080
Poptr1#572692
At2g47650
At3g62830
Os05g29990
Os07g47700
Poptr1#735401
Os03g17230
Chlre2#163641†
Os01g73790
At2g27860
At1g08200
Poptr1#649062
Poptr1#366566
Poptr1#594851
Cycas_CB090377
Os11g37890
Poptr1#662916
Poptr1#742031
At5g28840
Os10g28200
Chlre2#163850†
Os06g44260
Os06g44270
Poptr1#561376
Poptr1#91781
At1g17890
At1g73250
CAC43708†

An example: Rice PPI

- http://www.harvest-web.org/

Rice	ATH
Os.3420.1	AT3G12110.1
…	…
Os.52771.1	AT5G60390.3
Os.55715.1	AT1G16300.1
Os.5492.1	AT3G56070.2
…	…

7000 15000



Bioinformatics methods

- Homologous method to find Orthology
- Prediction
  - Sequence method
  - Structural based method
- Text mining
- Infer from other networks, such as expression profile, GO annotations.

Predicting protein-protein interactions

- Sequence methods
- How can you predict that an interaction might occur between two proteins based purely on sequence data?

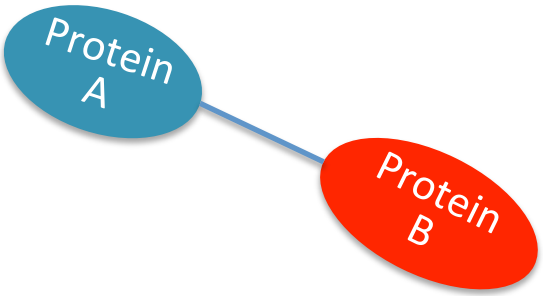
Valencia & Paz o s, (2002) Current Opin ion in Structural Biolog y 12, 368-373
Skrabanek et al. (2008) Mol Biotechnol. 38(1):1-17.

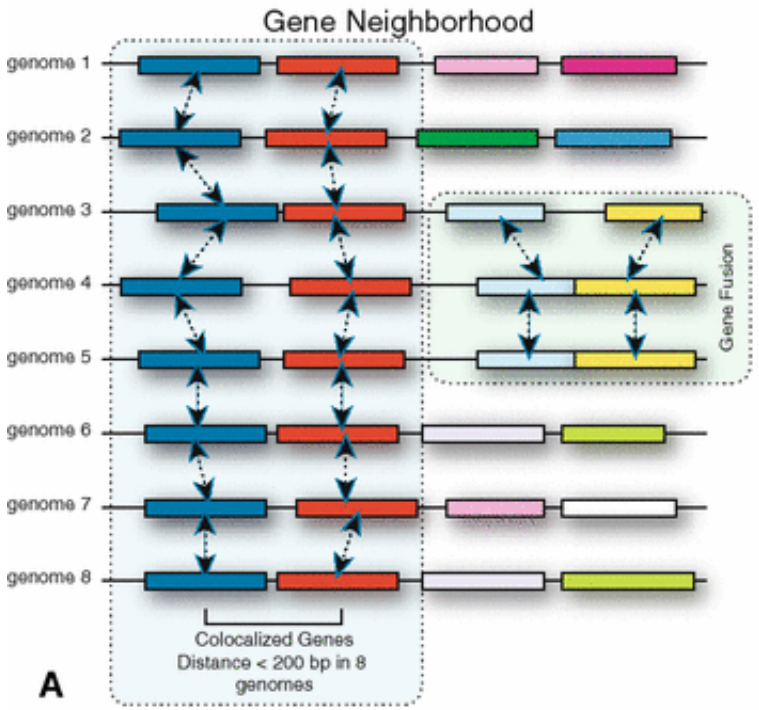
Prediction PPI with sequences

- Gene neighborhood
- Gene fusions
- Phylogenetic profiles
- Co-evolution
- Correlated Mutation

Prediction PPI with sequences

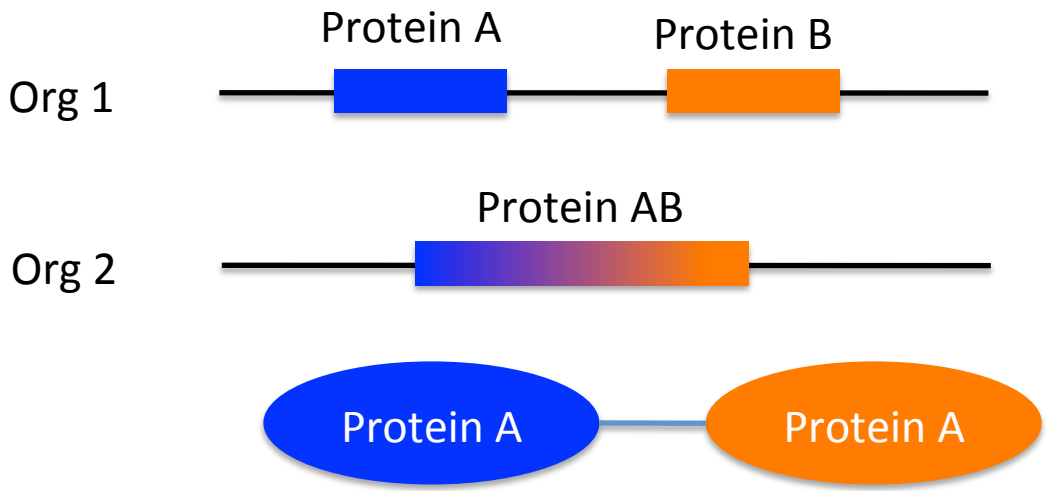
- Gene neighborhood
  - for bacteria, the arrangement of genes in operons means that interacting proteins are often encoded in adjacent sites in the genome





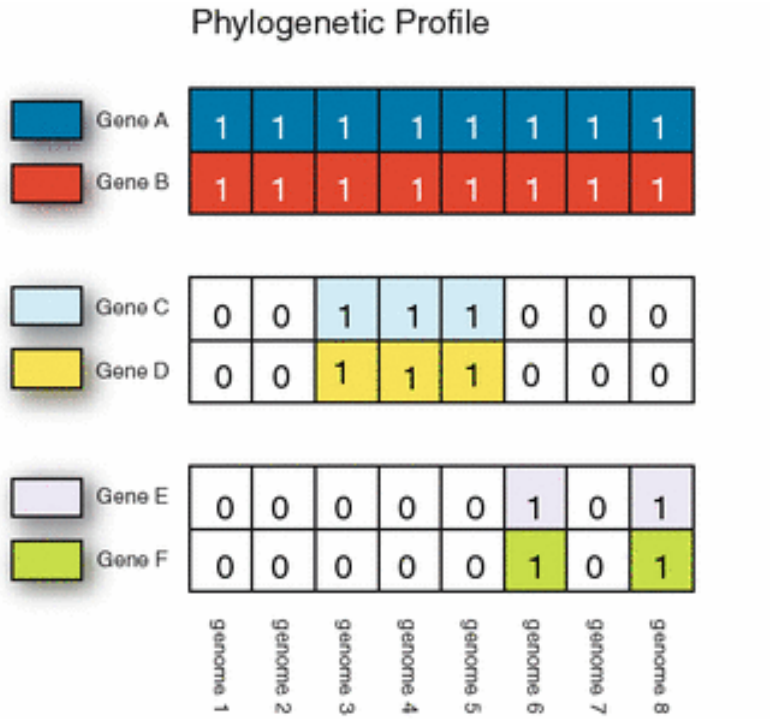
Prediction PPI with sequences

- Gene fusions
  - genes encoding interacting proteins in one organism are sometimes fused into a single gene in another. Look for these occurrences.



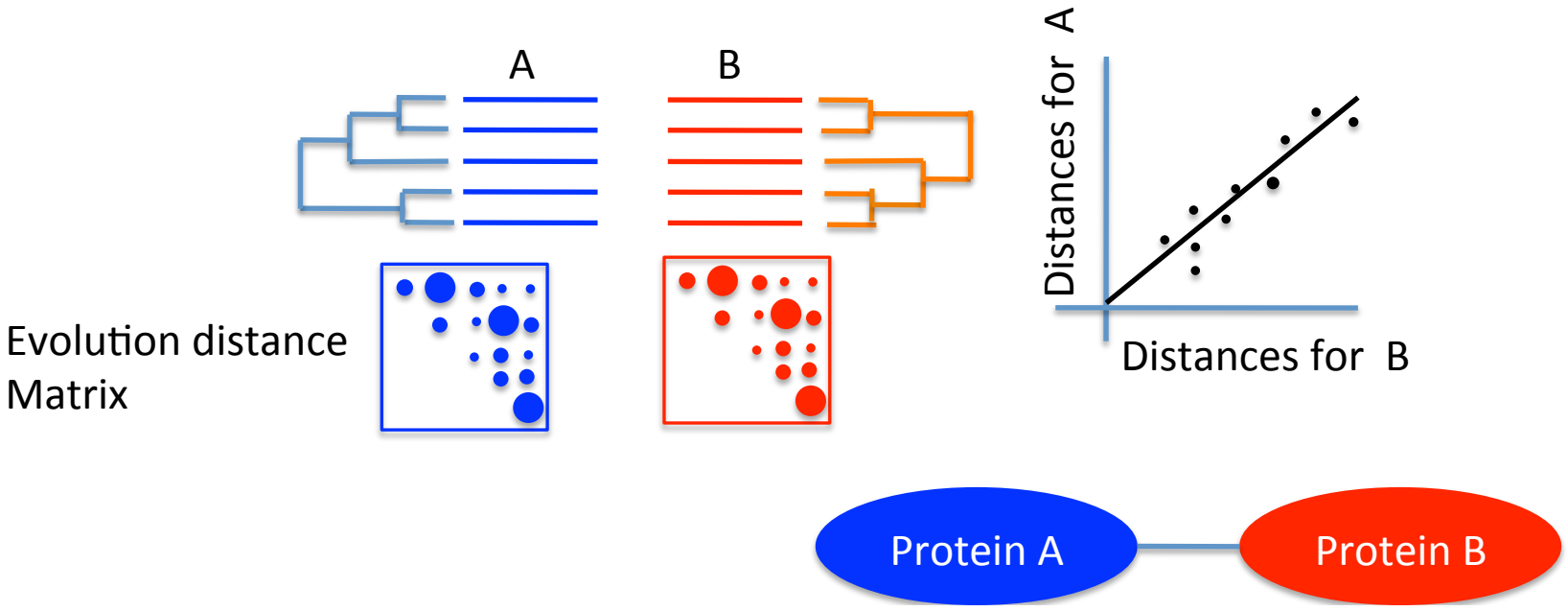
Prediction PPI with sequences

- Phylogenetic profiles
  - based on the joint presence/absence of a pair of proteins in a large number of genomes.



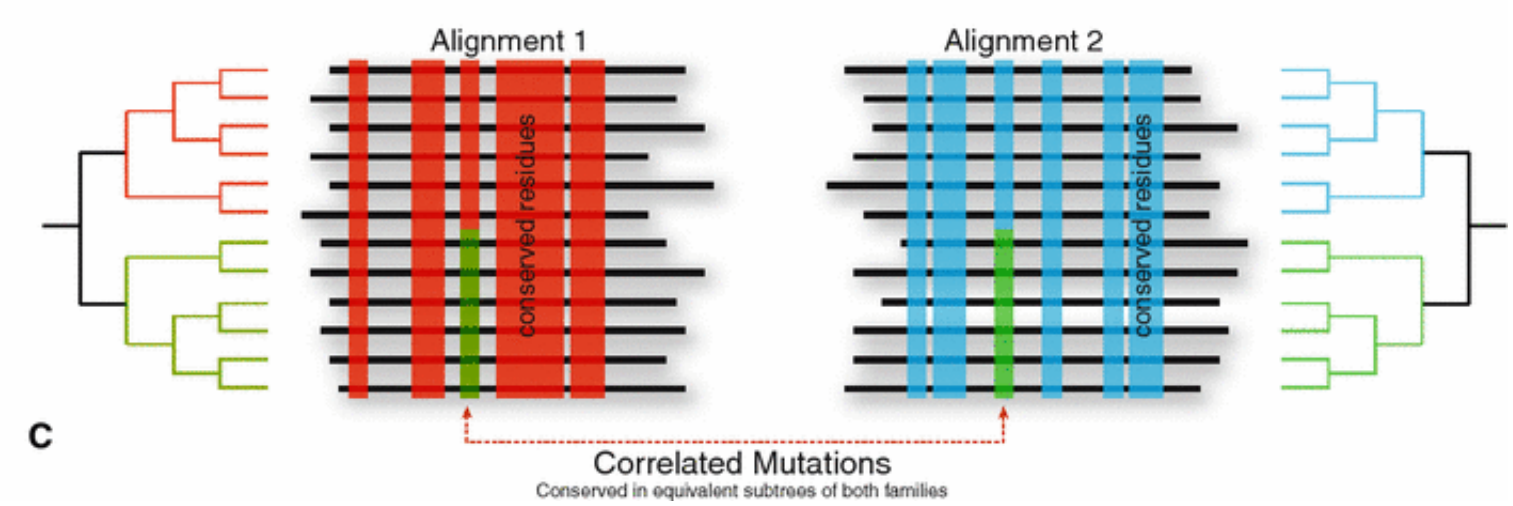
Prediction PPI with sequences

- Co-evolution
  - as assessed by similarity of phylogenetic trees. "mirrortree" method compares the distance matrices for generating trees;



Prediction PPI with sequences

- Correlated mutations
  - the idea is that interacting positions on different proteins should co- evolve so as to maintain the interface. Look for correlation between sequence changes at one position and those at another position in a multiple sequence alignment.



Süel et al. (2002) Nature Strut. Bio.
Pazos & Valencia (2002) Proteins

Prediction PPI with sequences

- Problems: they need lots of sequences, and the methods are very sensitive to the alignment method we used.

Web tools for PPI prediction with sequences

- AllFUSE (Enright *et al. 2001*, Gene fusions, http://www.ebi.ac.uk/research/cgg/allfuse/)
- STRING (Snel *et al.* 2000, Gene Co-Localization, gene-fusion, phylogenetic profiles, http://www.bork.embl-heidelberg.de/STRING/)
- WIT (Overbeek *et al.* 2000, Orthology/phylogenetic profiles/gene co-localization, http://wit.mcs.anl.gov/WIT2/)
- Predictome (Mellor *et al.* 2002, Gene Co-Localization, gene-fusion, phylogenetic profiles, http://predictome.bu.edu/)
- COGs (Tatusov *et al.* 1997, Orthology/phylogenetic profiles, http://www.ncbi.nlm.nih.gov/COG/)

Bioinformatics methods

- Homologous method to find Orthologs
- Prediction
  - Sequence method
  - Structural based method
- Text mining
- Infer from other networks, such as expression profile, GO annotations.

Structure-based methods

- Docking Method
- Threading Methods
- Structural Modeling Methods

Structures of protein interactions

- If two proteins have an interaction, they bind together as a certain conformation.
- For two give structures, if we can predict their docking conformation, we can predict their interaction.

Xylanase (Enzyme)

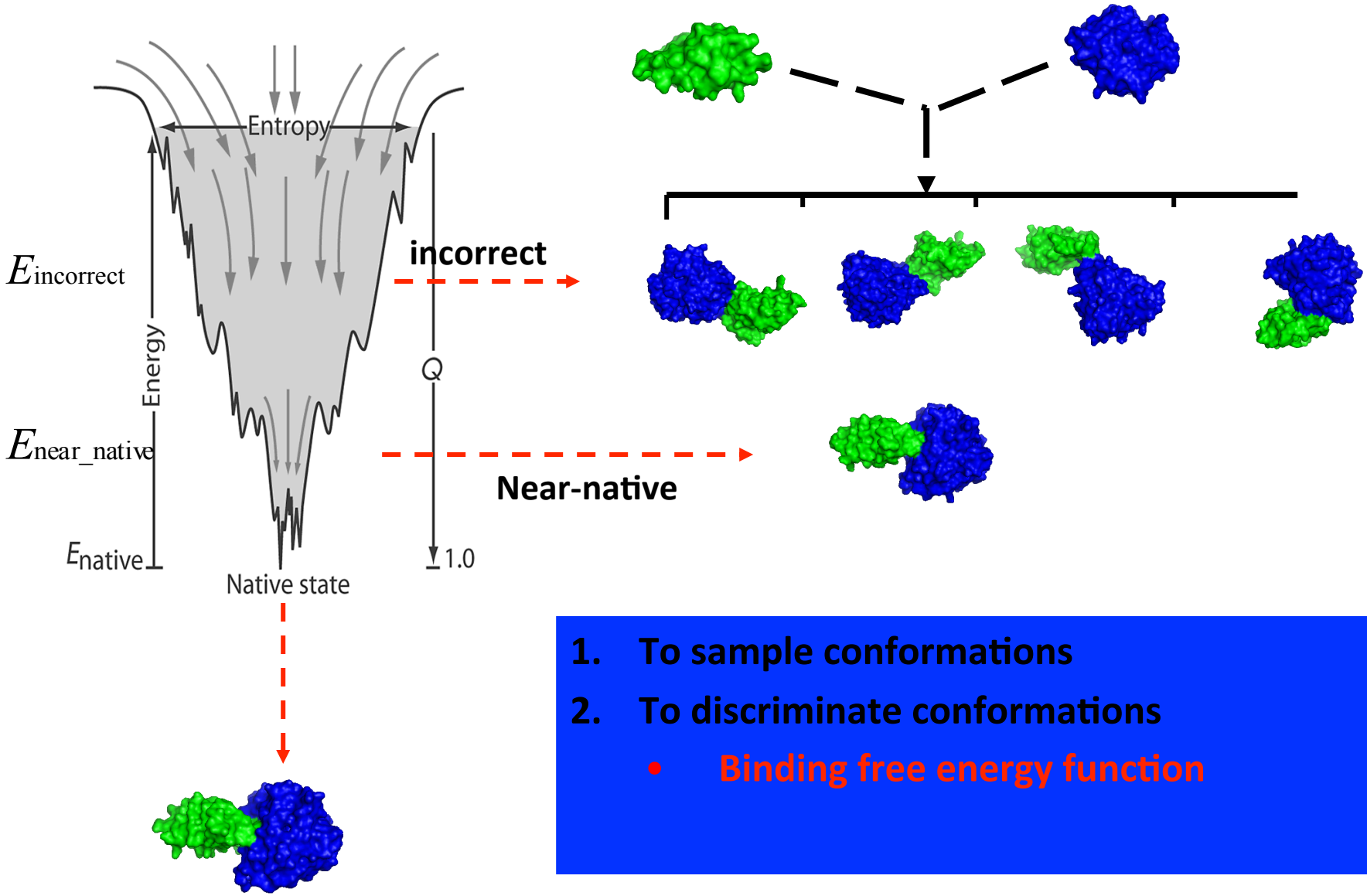
TAXI (Inhibitor)

PDB: 2B42

Docking methods

- Docking : how two known structures will interact
- Docking approaches require structures of both interacting components.

Docking Method



1. To sample conformations
2. To discriminate conformations
   - Binding free energy function

Docking servers

- Zdock: http://zdock.bu.edu/
- Hex: http://hex.loria.fr/
- RossetaDock: http://rosettadock.graylab.jhu.edu/
- GRAMM-X: http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/
- PATCH dock: http://bioinfo3d.cs.tau.ac.il/PatchDock/

Limitations of Docking methods

- No good energy scoring function to evaluate the docked structures.
- We don't have enough structures or good enough docking methods to make high-throughput prediction of protein-protein interactions practical at this point.
- Frequently, conformational changes accompany protein interactions. Docking methods generally require a structure of the bound conformation to predict interactions correctly. Modeling conformational flexibility is hard.

Structure-based methods

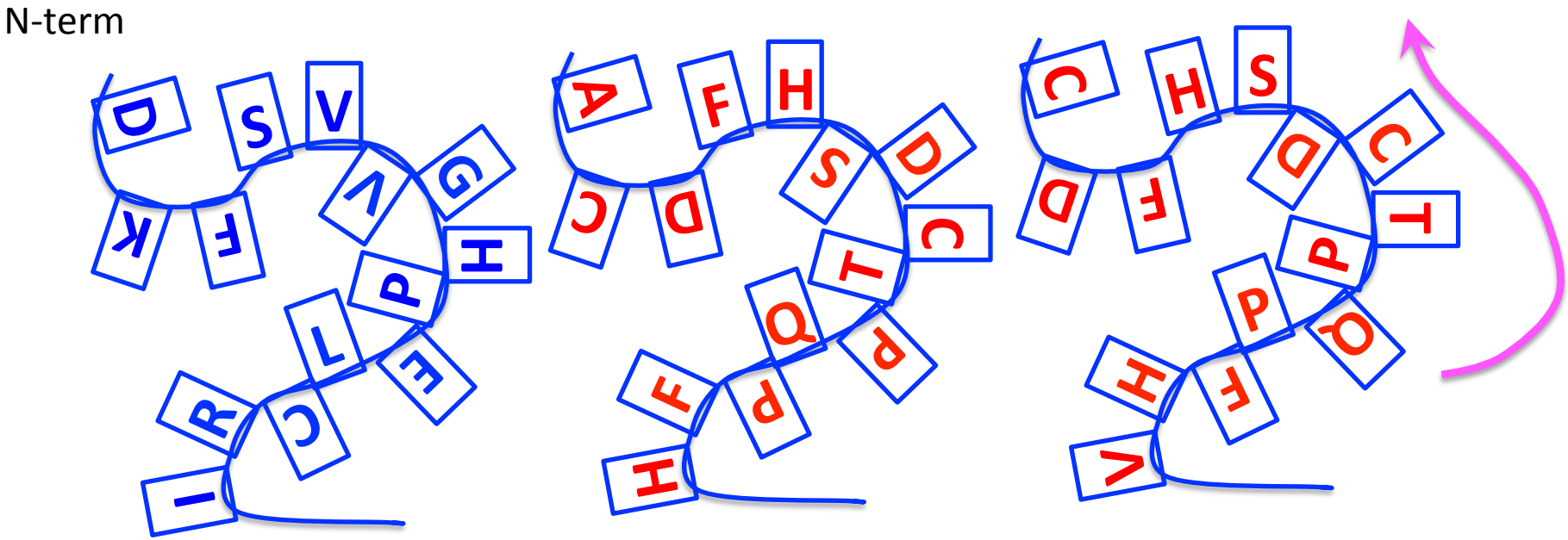
- Docking Method
- Threading Methods
- Structural Modeling Methods

Threading method

- For two proteins, we do not have structures for them.
- There are many protein binding complex structures in PDB.
- We may use the “threading method” to model the binding structures for two given proteins.

Threading method

- Protein-1: ACDFHSDCTPQPFHVISGAD……

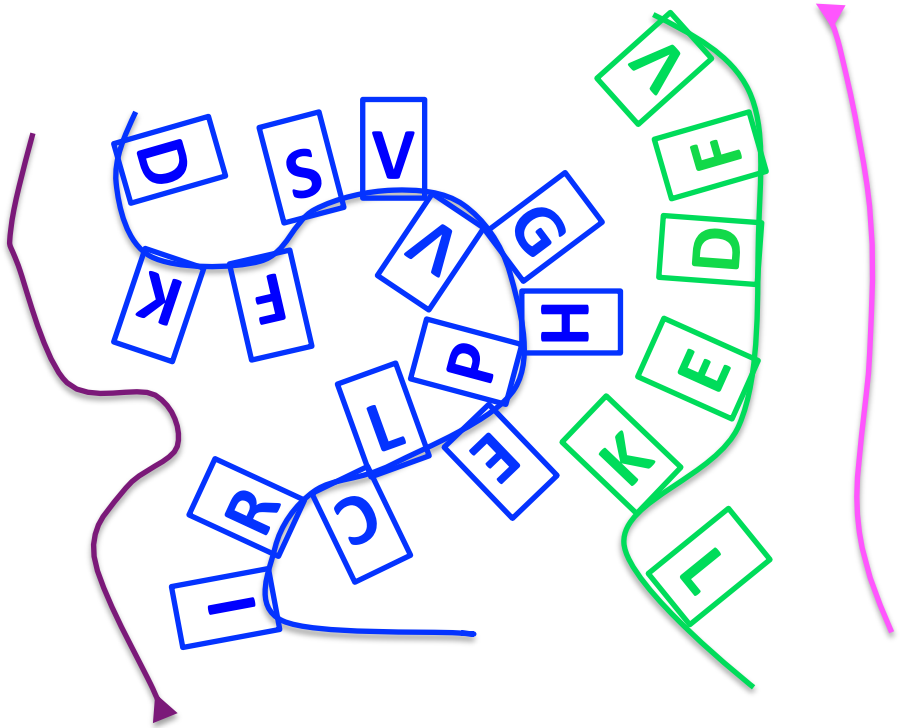


C-term

Structure template

Prediction Interactions by Threading method

- Protein-1: ACDFHSDCTPQPFHVISGAD……
- Protein-2: SKENYWAQLIHVGKSREYAI……



Complex Structure template

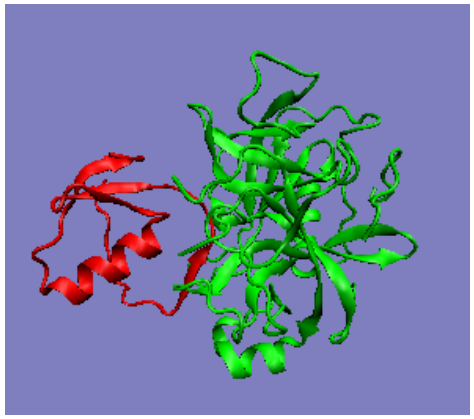
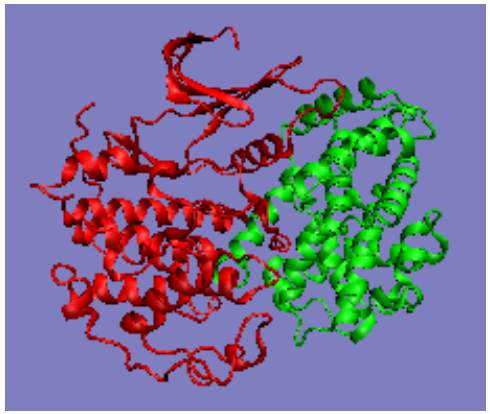
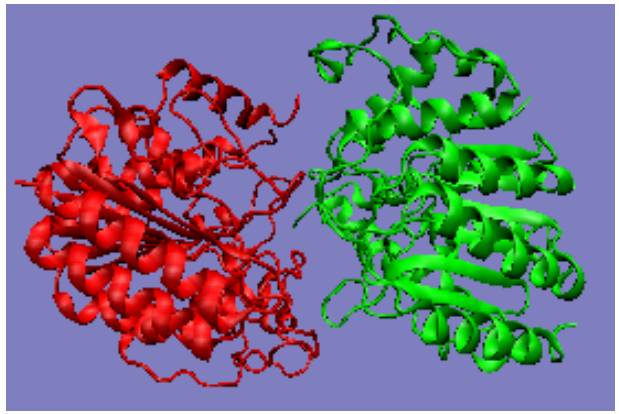
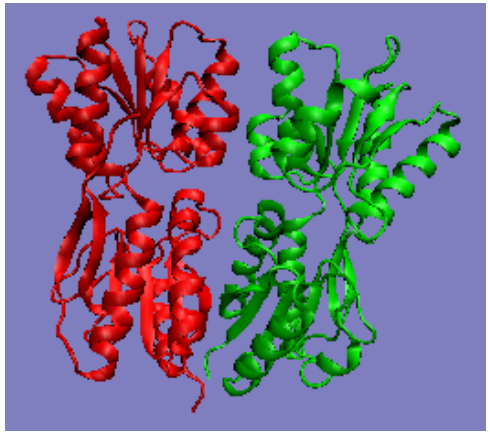
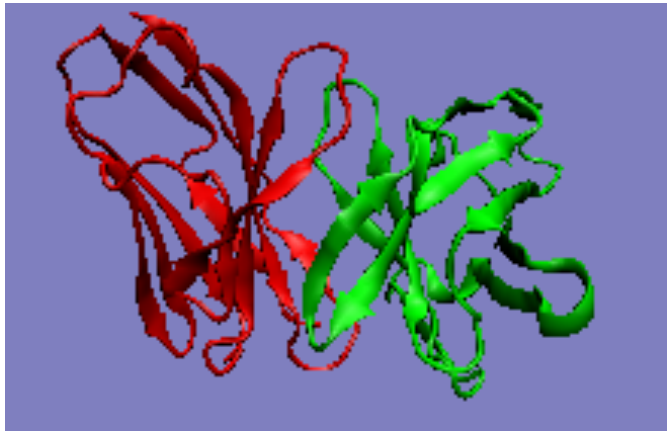
Threading methods

- Threading methods
  - Phase I: collect a complex structure library
  - Phase II: Thread each target sequence onto a library of folds
  - Phase III: Take pairs of fold assignments and thread the targets onto complexes of these folds (complexes of known structure) Evaluate an interfacial score to determine how complementary the fit is.

Lu et al., PROTEINS (2002) 49, 350-364,
Genome Research (2003) 13, 1146-1154

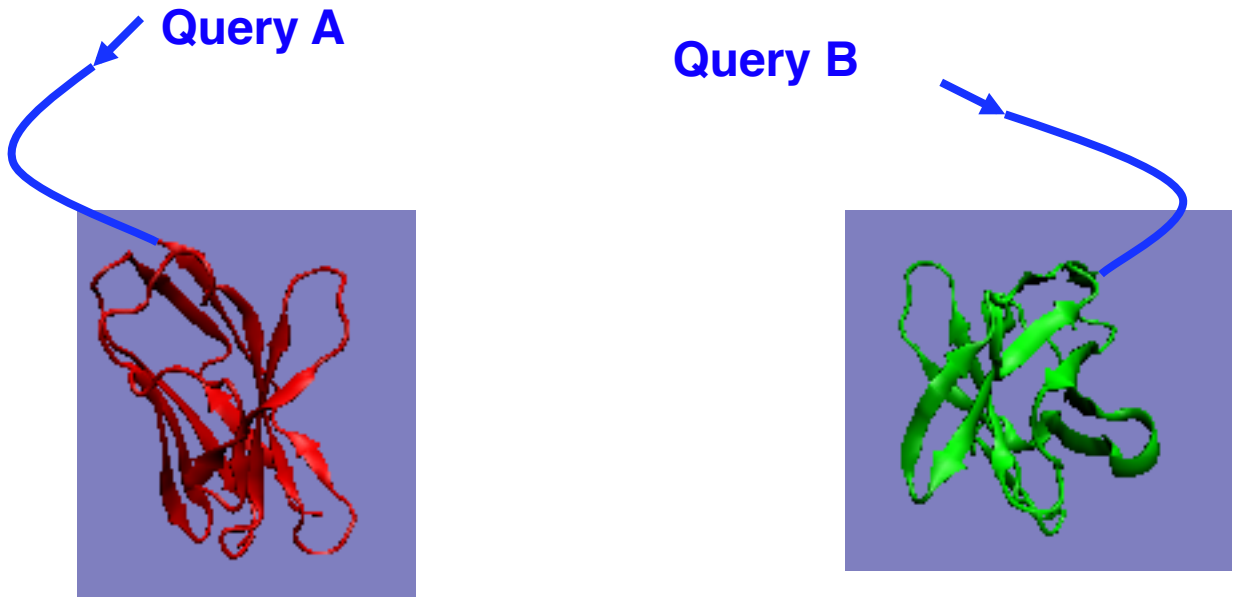
Method Description

1) Establish a library of dimeric templates ….

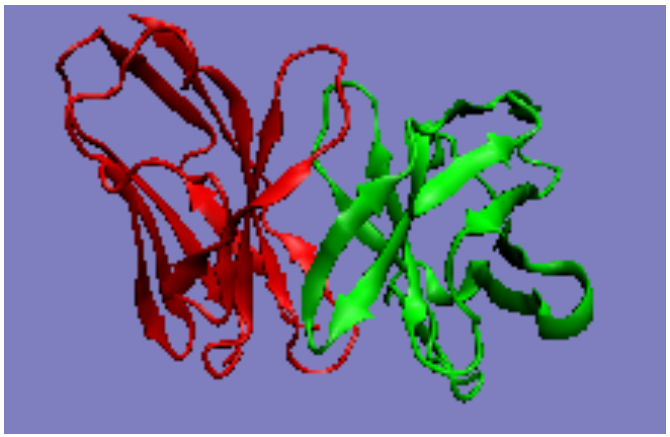
Method Description

2) Match query sequences to the structure of individual chains by threading method



Method Description

3) Verify binding by an energy scoring function



Is the binding affinity large enough? (binding threshold)

Threading methods

- Used library of 768 complexes, predicted 7,321 interactions for yeast proteins.
- Hard to assess performance. One way is to look at some property that you believe should correlate with interactions, e.g. co-localization or function.

Lu et al., PROTEINS (2002) 49, 350-364,
Genome Research (2003) 13, 1146-1154

Structure-based methods

- Docking Method
- Threading Methods
- Structural Modeling Methods

Predict protein interaction by Structural Modeling Methods

- For two proteins, we do not have structures for them.
- Similar to threading method, we may predict their structures.

Structure prediction methods

- ***Ab initio* methods**
  - based on physical principles rather than on previously solved structures
  - Not accurate
  - Time consuming
- **Template-based methods**
  - Predicting structure for a given sequence using a known structure template
  - The number of structure topologies is limited
  - Different sequences may encode similar structure

Structure prediction

Predicting structure for a given sequence using a known structure

1) Homology modeling: (Seq. ID>50%)

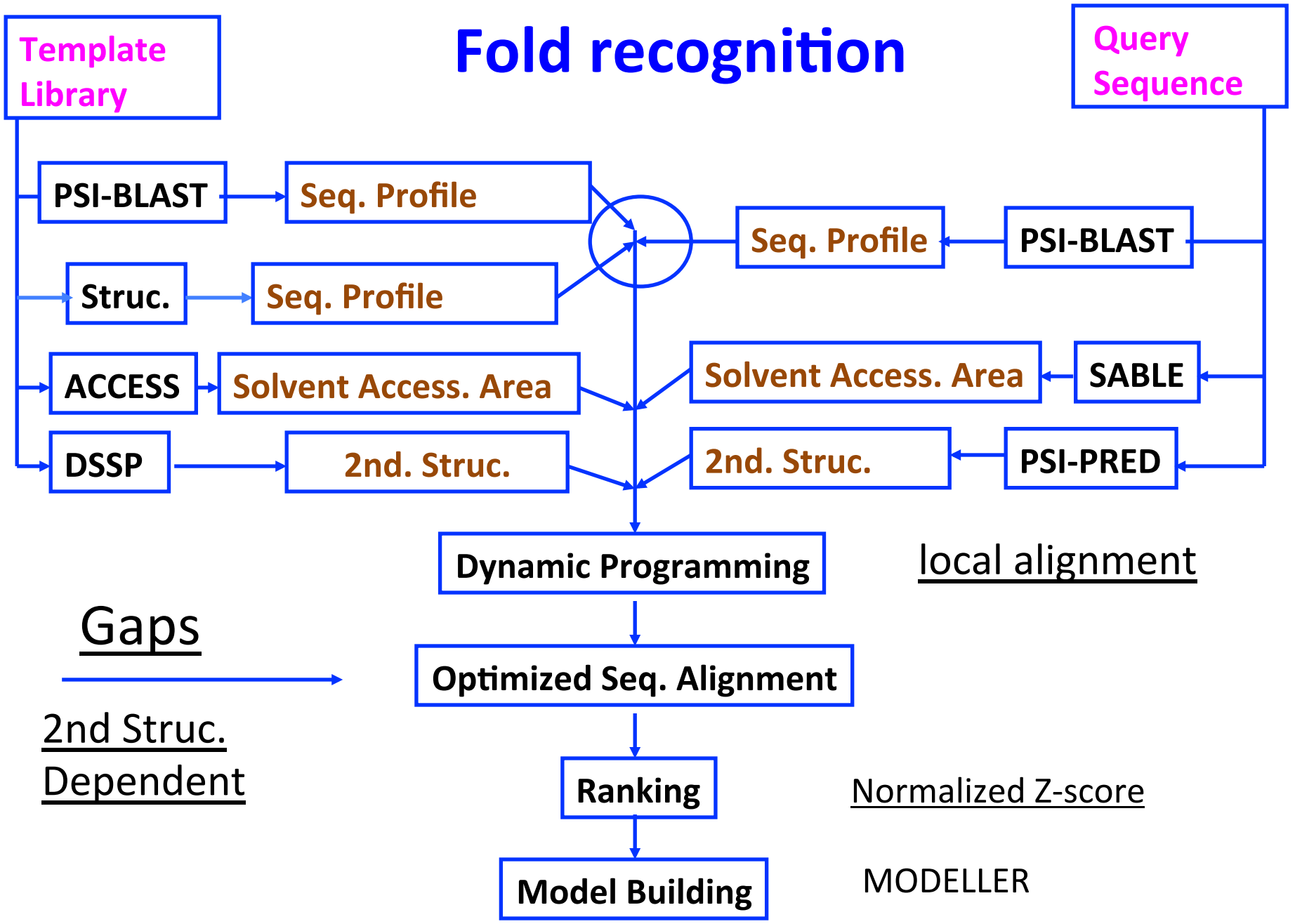
- High sequence identity with a sequence having known structures. (Any sequence level algorithm, such as BLAST, is enough)

2) Fold recognition (Remote/Structural homolog):

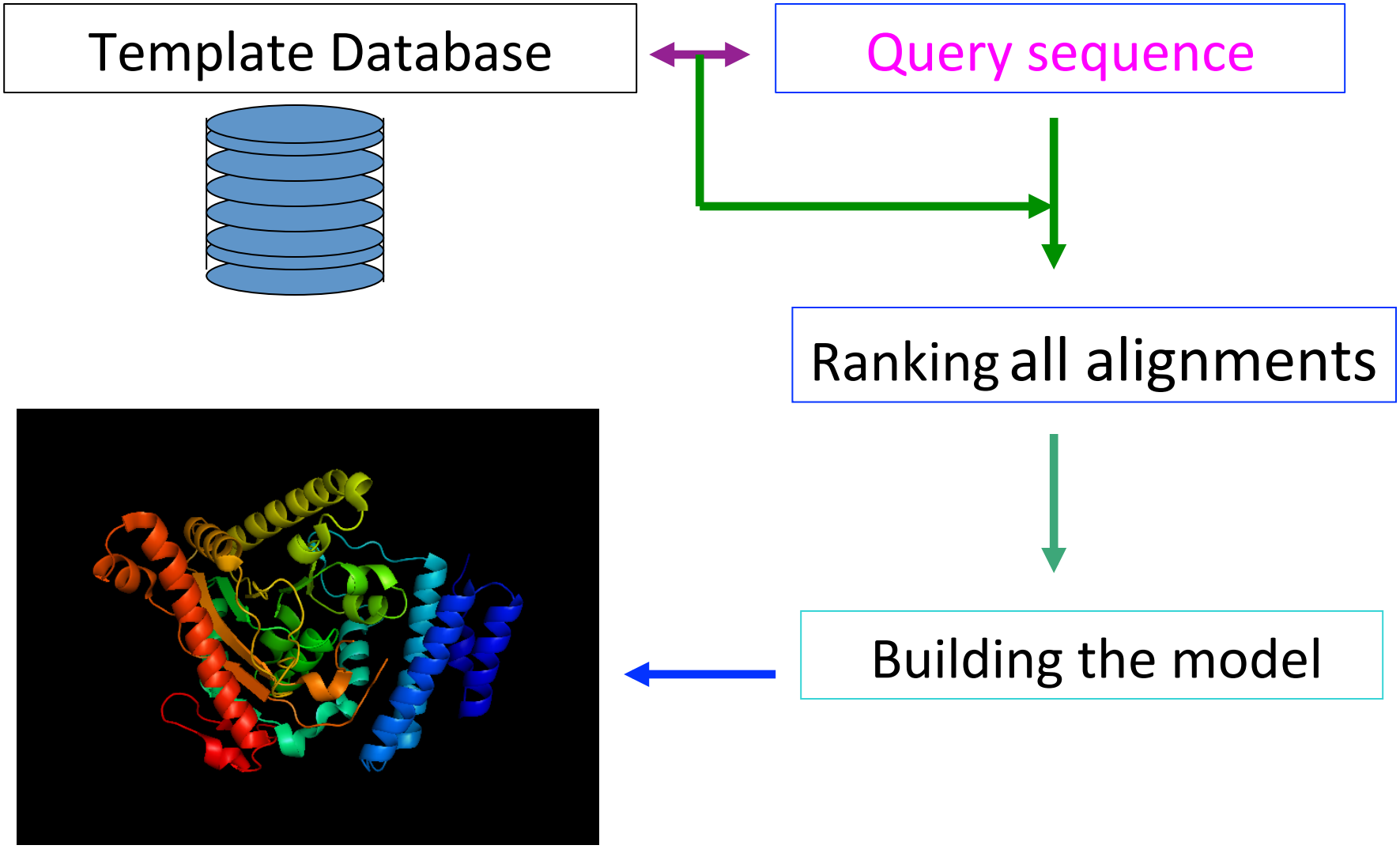
(Seq. ID <50%)

- Recognizing structurally homologous sequence without significant sequence identity with known structures

Fold recognition

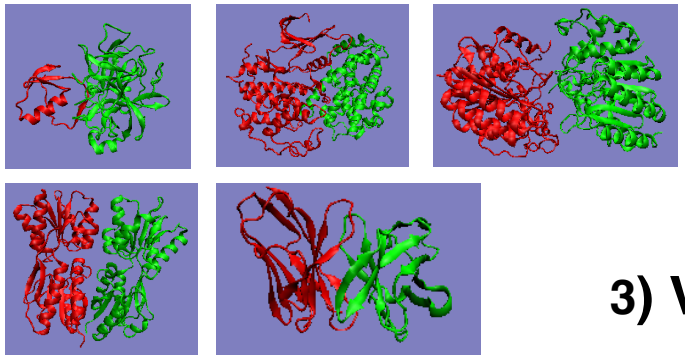


Structure Modeling



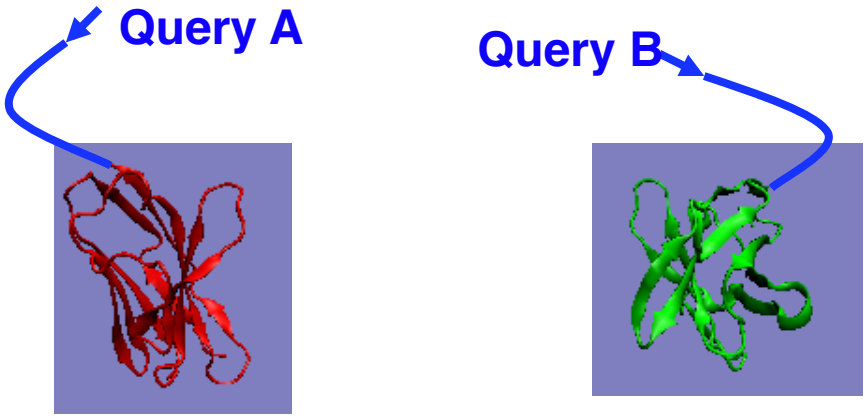
Method Description

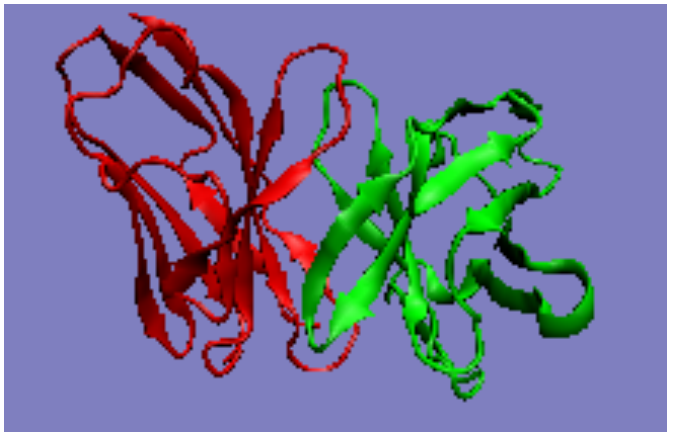
1) Establish a library of dimeric templates ….



3) Verify binding by an energy scoring function

2) Match query sequences to the structure of individual chains by modeling method





Is the binding affinity large enough? (binding threshold)

Structure Modeling Method

- ~65% accuracy when assessing whether different fibroblast growth factors bind to various receptors (4 structures available, 252 possible pairings evaluated).
- A library with 699 homodimers and 229 heterodimers, yeast 5887 proteins, predicted 2556 interactions. (Zhang)
- Not practical to apply at the genome level due to lack of homologous complexes with structures.

Bioinformatics methods

- Homologous method to find Ortholog
- Prediction
  - Sequence method
  - Structural based method
- Text mining
- Infer from other networks, such as expression profile, GO annotations.