Identification of Functional Modules in Protein Complexes via Hyperclique Pattern Discovery

Hui Xiong

Department of Computer Science & Engineering
University of Minnesota - Twin Cities

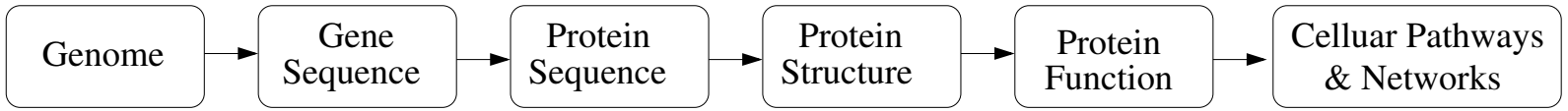
Overview

⇒ Introduction

- Protein Complex Data
- Functional Modules in Protein Complexes
- Hyperclique Pattern Discovery
- Experimental Results
- Summary

Introduction

- The pathway of bioinformatics.



- The Conventional Reductionist Approach.
  - One-gene-one-protein-at-a-time basis.
- Systems Biology
  - Global or integrative approaches.
  - A system-level understanding of behaviors and interactions among all the individual components of the cell.
  - Computational science can play an important role in systems biology.

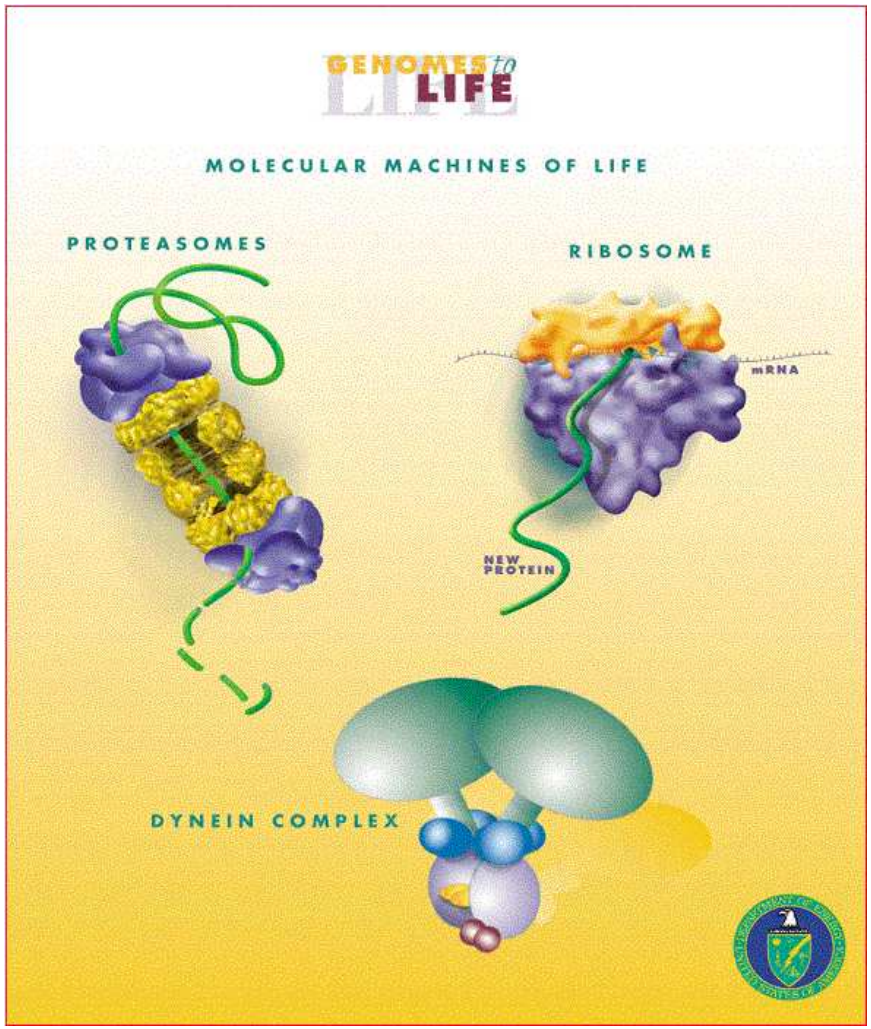
Overview

- Introduction

⇒ Protein Complex Data

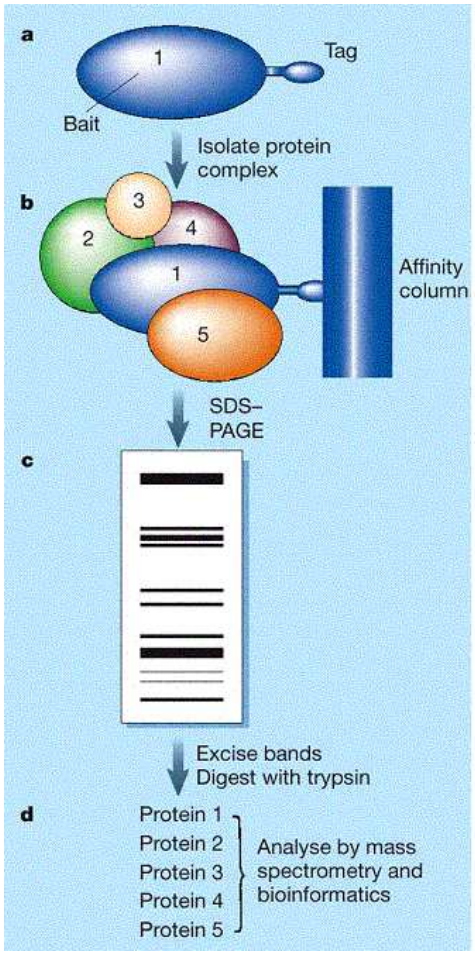
- Functional Modules in Protein Complexes
- Hyperclique Pattern Discovery
- Experimental Results
- Summary

Protein Complexes



- Multi-protein complexes contain functionally related proteins.
- Cellular process carried out by multi-protein complexes.
- Perform higher order functions.

Protein Complex Data



- Protein complexes from Large-scale experimental studies.
- The TAP-MS dataset by Gavin et al. 2002: Tandem affinity purification (TAP) - mass spectrometry (MS).
- Use bait proteins.
- Capture 232 multi-protein complexes.
- The TAP-MS dataset is the most reliable one (Deng, $etal.$)

Protein Complex	Proteins
c1	p_1, p_2
c2	p_1, p_3, p_4, p_5
c3	p_2, p_3, p_4, p_6

Overview

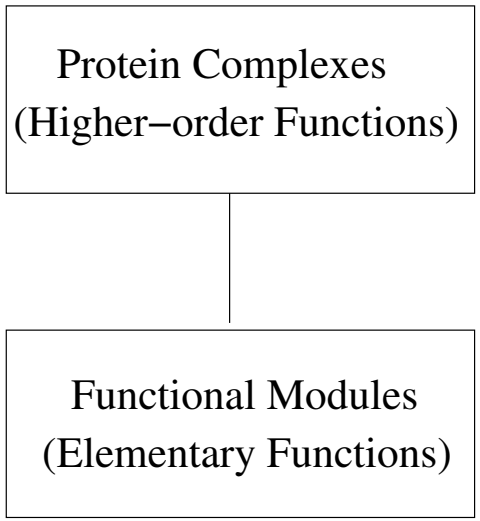
- Introduction
- Protein Complex Data

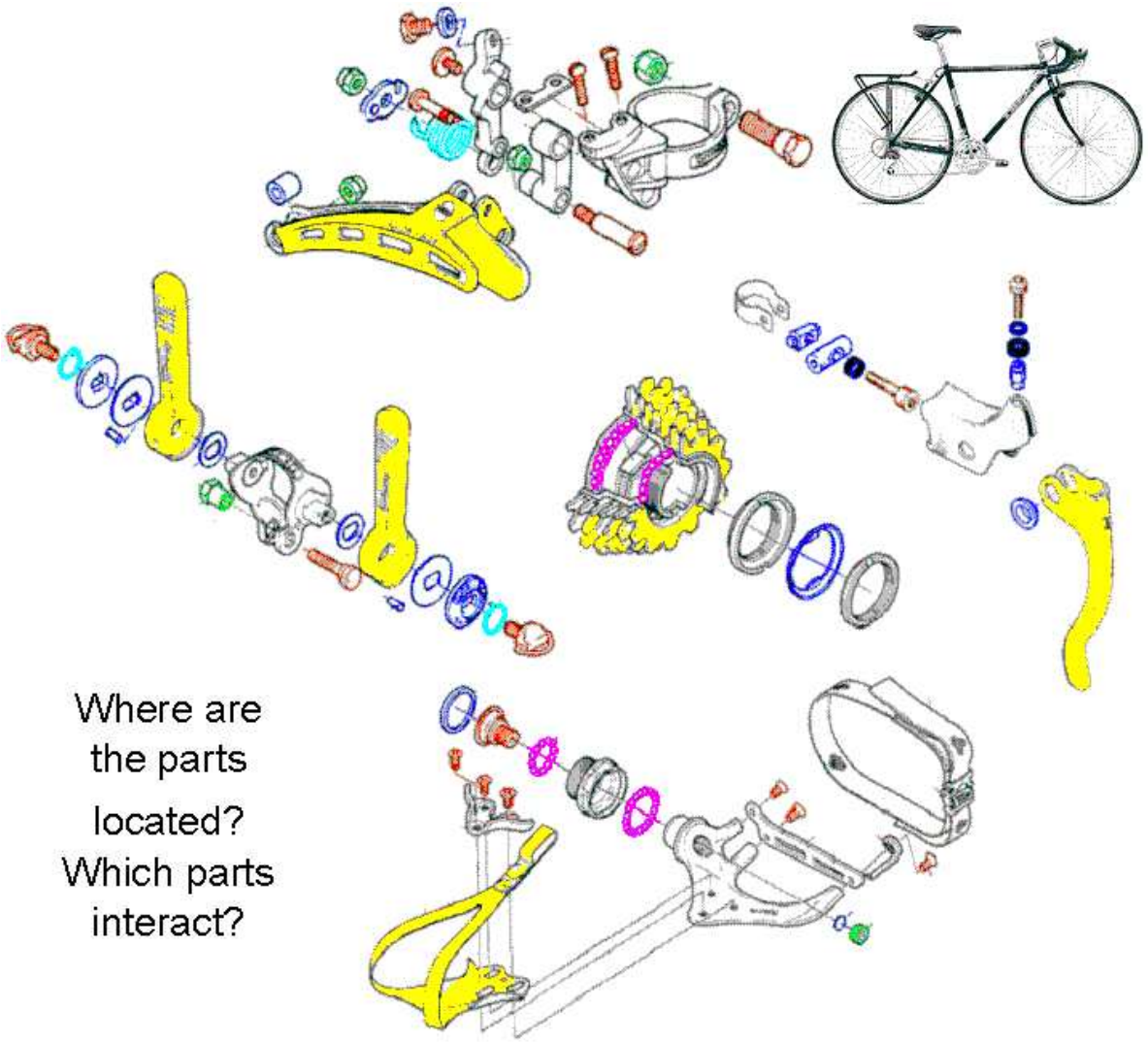
$\Rightarrow$ Functional Modules in Protein Complexes

- Hyperclique Pattern Discovery
- Experimental Results
- Summary

Functional Modules in Protein Complexes

- Functional Modules
  Functional modules are groups of proteins involved in common elementary biological function.





What are the shared parts (bolt, nut, washer, spring, bearing), unique parts (cogs, levers)? What are the common parts -- types of parts (nuts & washers)?

How many roles can these play? How flexible and adaptable are they mechanically?

Where are the parts located? Which parts interact?

(C) Mark Gerstein, Yale

Overview

- Introduction
- Protein Complex Data
- Functional Modules in Protein Complexes

⇒ Hyperclique Pattern Discovery

- Experimental Results
- Summary

Hyperclique Pattern Concepts

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Egg
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Protein Complex	Proteins
c1	p_1, p_2
c2	p_1, p_3, p_4, p_5
c3	p_2, p_3, p_4, p_6

- Pattern
  - A collection of one or more items
    E.g. {Milk}, {Beer, Diaper}
- Support Count (σ)
  - Frequency of occurrence of a pattern.
    E.g. σ({Bread, Milk, Diaper}) = 2
- Support (Agrawal et al. 1993)
  - Fraction of transactions that contain a pattern.
  - E.g. supp({Bread, Milk, Diaper})= 2/5 =40%

Hyperclique Pattern Concepts

- **Definition 1** *The* $\mathbf{h\text{-}confidence}$ *of a pattern* $P = \{i_1, i_2, \cdots, i_m\}$*, denoted as* $hconf(P)$*, is* $\frac{supp(\{i_1,i_2,\ldots,i_m\})}{max_{1\leq k\leq m}\{supp(\{i_k\})\}}$.

- For a pattern $P = \{A, B, C\}$, assume that:

  - $supp(\{A\}) = 0.1$, $supp(\{\mathsf{B}\}) = 0.1$, $supp(\{C\}) = 0.06$, $supp(\{A, B, C\}) = 0.06$.

- Hence, the h-confidence of the pattern P is:

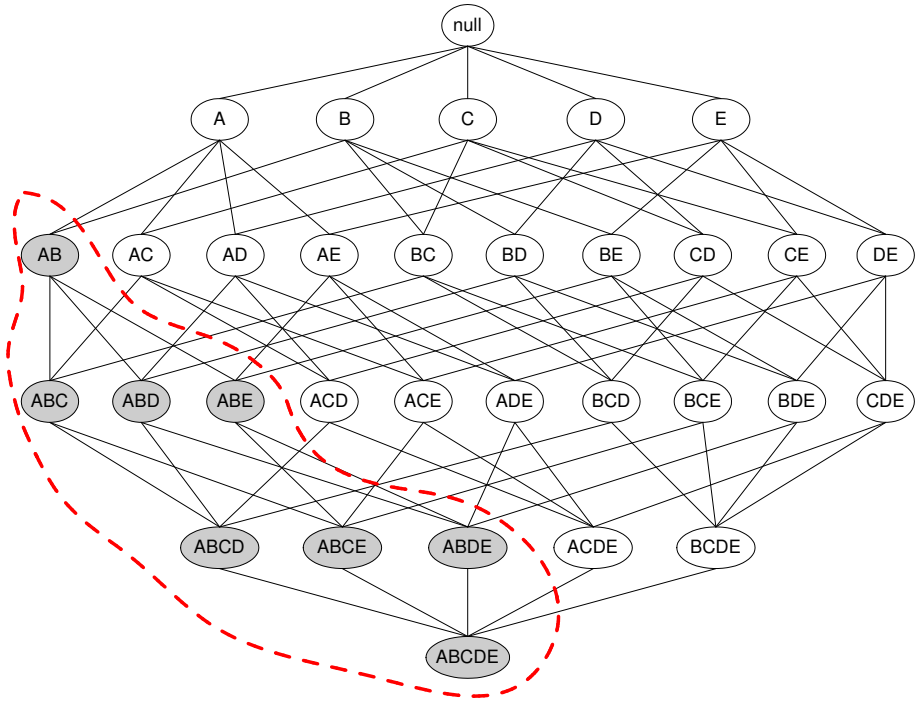
  - $hconf(P) = \frac{supp(\{A,B,C\})}{max\{supp(\{A\}),supp(\{B\}),supp(\{C\})\}} = 0.06/0.1 = 0.6$.

- **Definition 2** *A Pattern* P *is a* $\mathbf{Hyperclique\ Pattern}$ *if* $hconf(P) \geq h_c$*, where* h_c *is a user-specified minimum h-confidence threshold.*

- e.g. for a h-confidence threshold 50%, if the h-confidence of a pattern $P = \{A, B, C\}$ is 60%, then the pattern P is a hyperclique pattern.

The Anti-monotone Property of the H-confidence Measure

- **Lemma 1** *The h-confidence measure has the anti-monotone property. In other words, if* $P' \subseteq P$*, then* $hconf(P') \geq hconf(P)$*.*
- For instance, if the h-confidence of pattern {A, B} is not satisfied with the h-confidence threshold h_c, then the whole subtree of {A, B} can be pruned.



H-confidence as a measure of association

- **Theorem 1** *Given a hyperclique pattern* $P = \{i_1, i_2, \ldots, i_m\}$ *at the h-confidence threshold* h_c*, for two items* i_l *and* i_k *such that* $\{i_l, i_k\} \subset P$*, we have* $cosinesim(i_l, i_k) \geq h_c$*, where* $cosinesim(i_l, i_k) = \frac{supp(\{i_l, i_k\})}{\sqrt{supp(\{i_l\})supp(\{i_k\})}}$*, which is the cosine similarity between* i_l *and* i_k*.*
- The cosine similarity is uncentered correlation coefficient.
  - ◇ Pearson's Correlation Coefficient:

$$S(x_1, x_2) = \frac{\Sigma_{k=1}^{n}(x_{1k} - \overline{x_1})(x_{2k} - \overline{x_2})}{\sqrt{\Sigma_{k=1}^{n}(x_{1k} - \overline{x_1})^2 \Sigma_{k=1}^{n}(x_{2k} - \overline{x_2})^2}}$$

  - ◇ Uncentered Correlation Coefficient:

$$S(x_1, x_2) = \frac{\Sigma_{k=1}^{n} x_{1k}x_{2k}}{\sqrt{\Sigma_{k=1}^{n} x_{1k}^2 \Sigma_{k=1}^{n} x_{2k}^2}}$$

- Clique View: A hyperclique pattern can be viewed as a clique with items as vertices (there is an edge if the cosine similarity between two items is above the h-confidence threshold, h_c).

Overview

- Introduction
- Protein Complex Data
- Functional Modules in Protein Complexes
- Hyperclique Pattern Discovery

⇒ Experimental Results

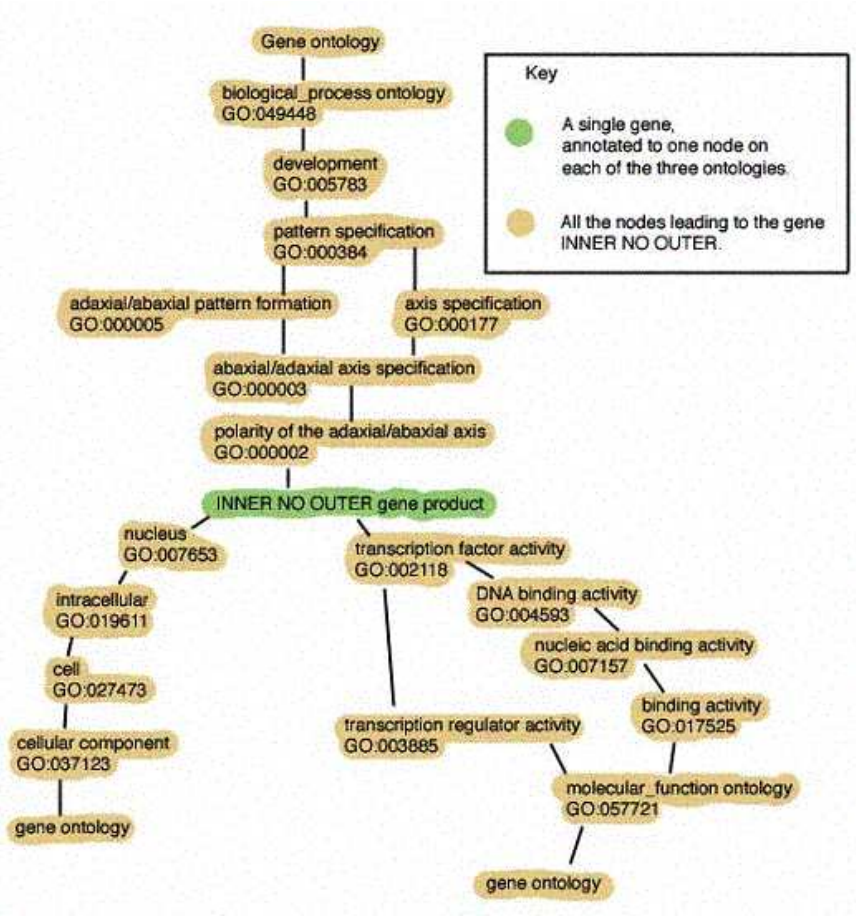
- Summary

Examples of Identified Hyperclique Patterns

- Examples of hyperclique patterns identified at a support threshold 0 and an h-confidence threshold 0.6. Detailed results are available at our project web site.
- http://www.cs.umn.edu/∼huix/pfm/pfm.html

Yeast Protein Complex Data		
Hyperclique patterns	Supp	Hconf
{Cus1, Msl1, Prp3, Prp9, Sme1, Smx2, Smx2, Smx3, Yhc1}	1.25%	100%
{Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Scl1}	1.7%	66.7%
{Cwc2, Ecm2, Hsh155, Prp19, Prp21, Snt309}	1.7%	100%
{Emg1, Imp3, Imp4, Kre31, Mpp10, Nop14, Sof1, Utp15, Noc4}	1.25%	100%

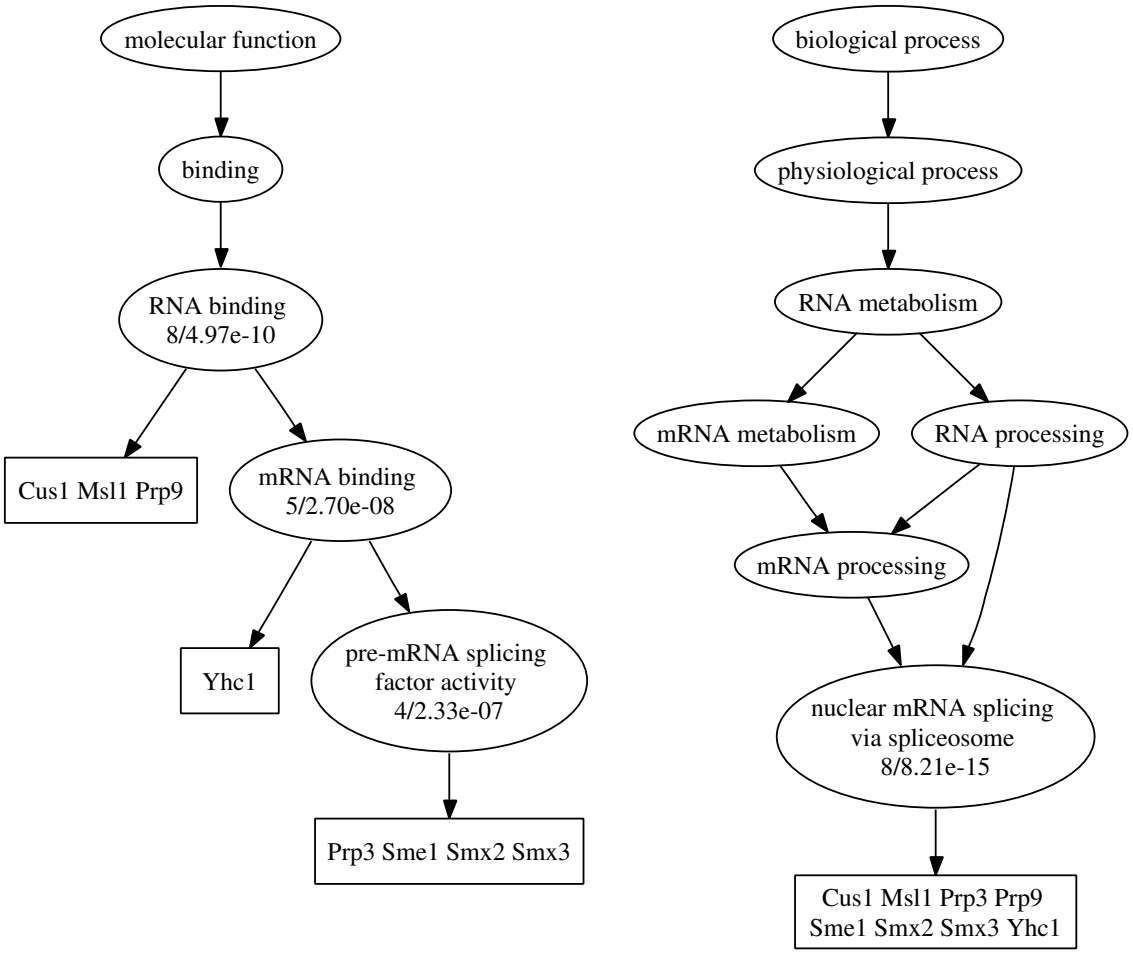
Gene Ontology (GO)



- Three separate ontologies: Biological Process, Molecular Function, Cellular Component.
- Organized as a DAG describing gene products (proteins and functional RNA).
- Collaborative effort between major genome databases.

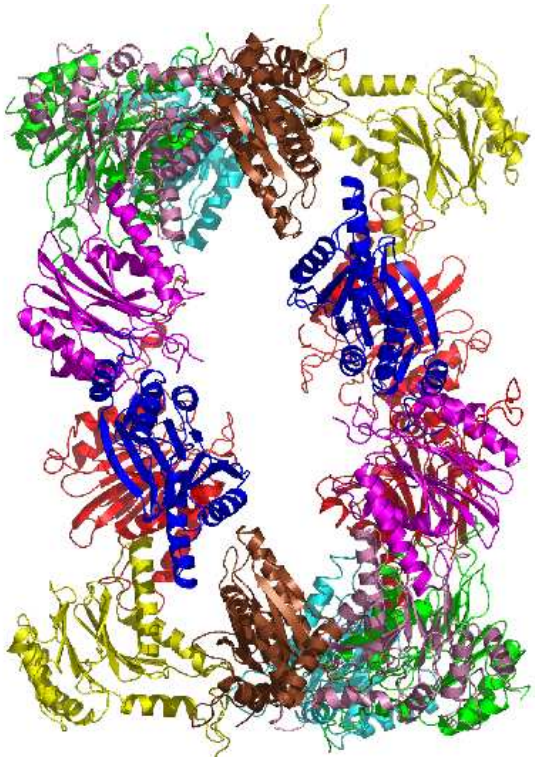
http://www.geneontology.org

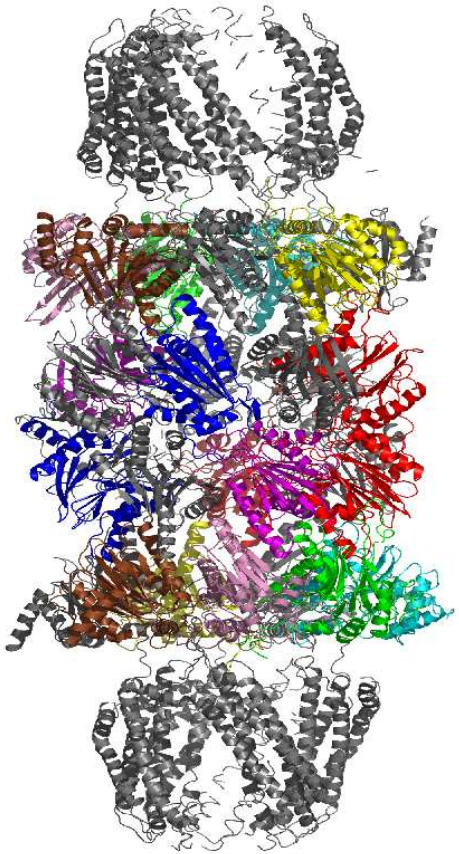
Analysis of Hyperclique Patterns Using Gene Ontology (GO)



- Pattern {Cus1, Msl1, Prp3, Prp9, Sme1, Smx2, Smx3, Yhc1}
- Function annotation: *RNA binding* with p-value 4.97e-10
- Biological process annotation: *nuclear mRNA splicing via spliceosome* with p-value 8.21e-15.

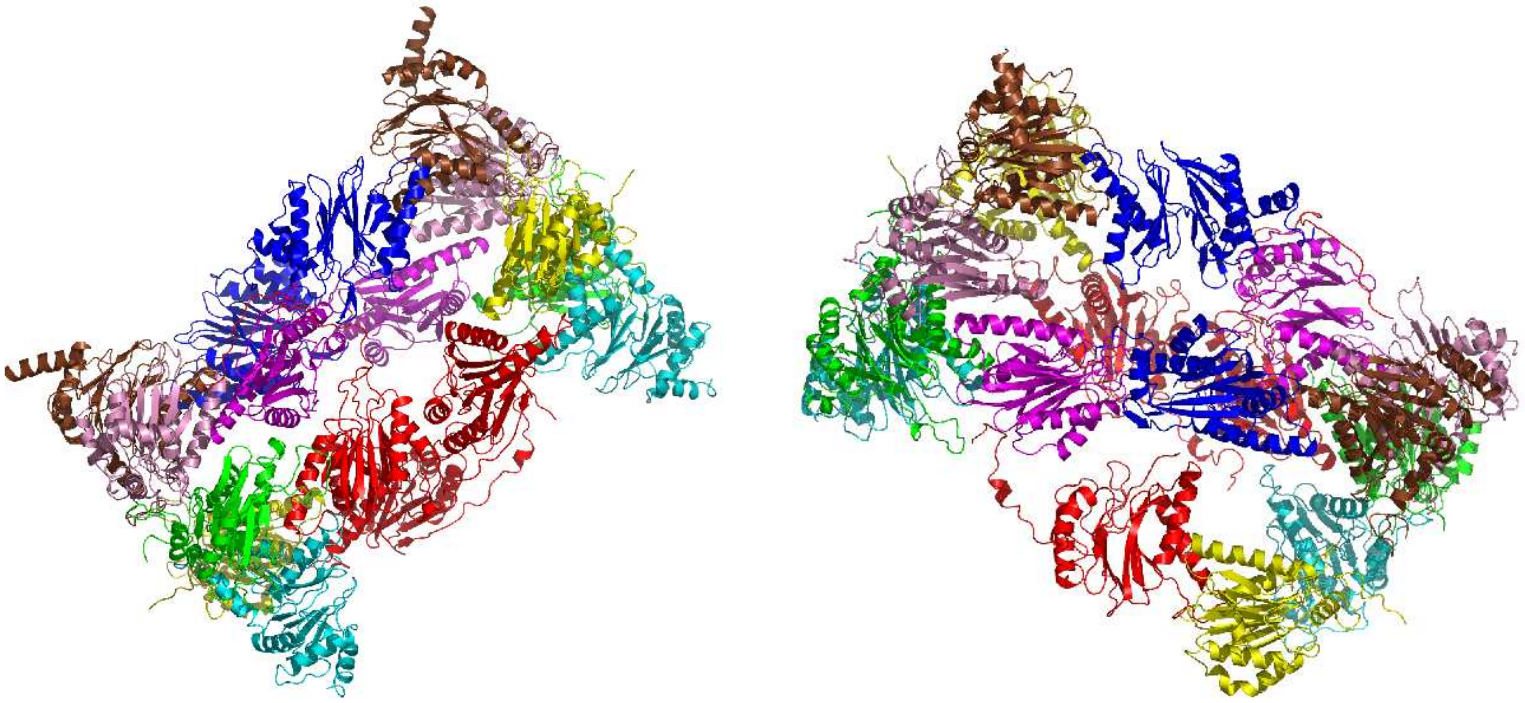
Hyperclique Patterns as Functional Modules





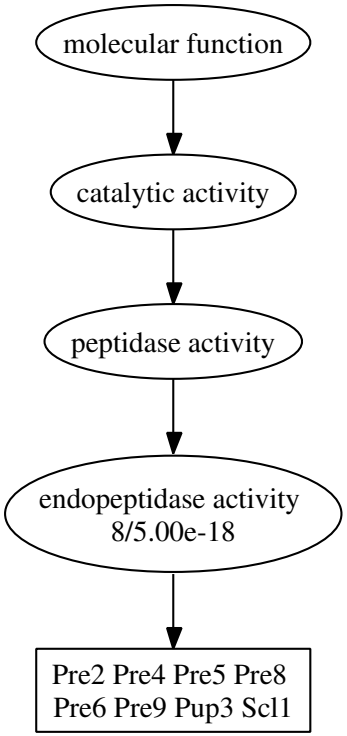
- Pattern {Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Scl1}
- 3-D structure data of proteasome is from the Protein Data Bank (http://www.rcsb.org/pdb)
- PyMol (pymol.sourceforge.net) visualizing the 3-D structure of proteins.

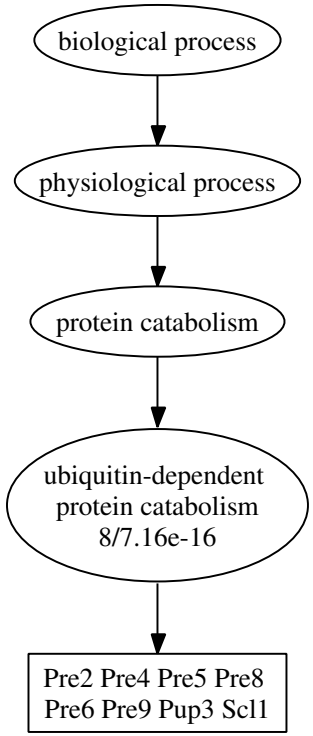
Hyperclique Patterns as Functional Modules



- Physical evidence implying that proteins in the same pattern tend to physically interact together to form a compact structure and perform a common function.

Gene Ontology Annotation





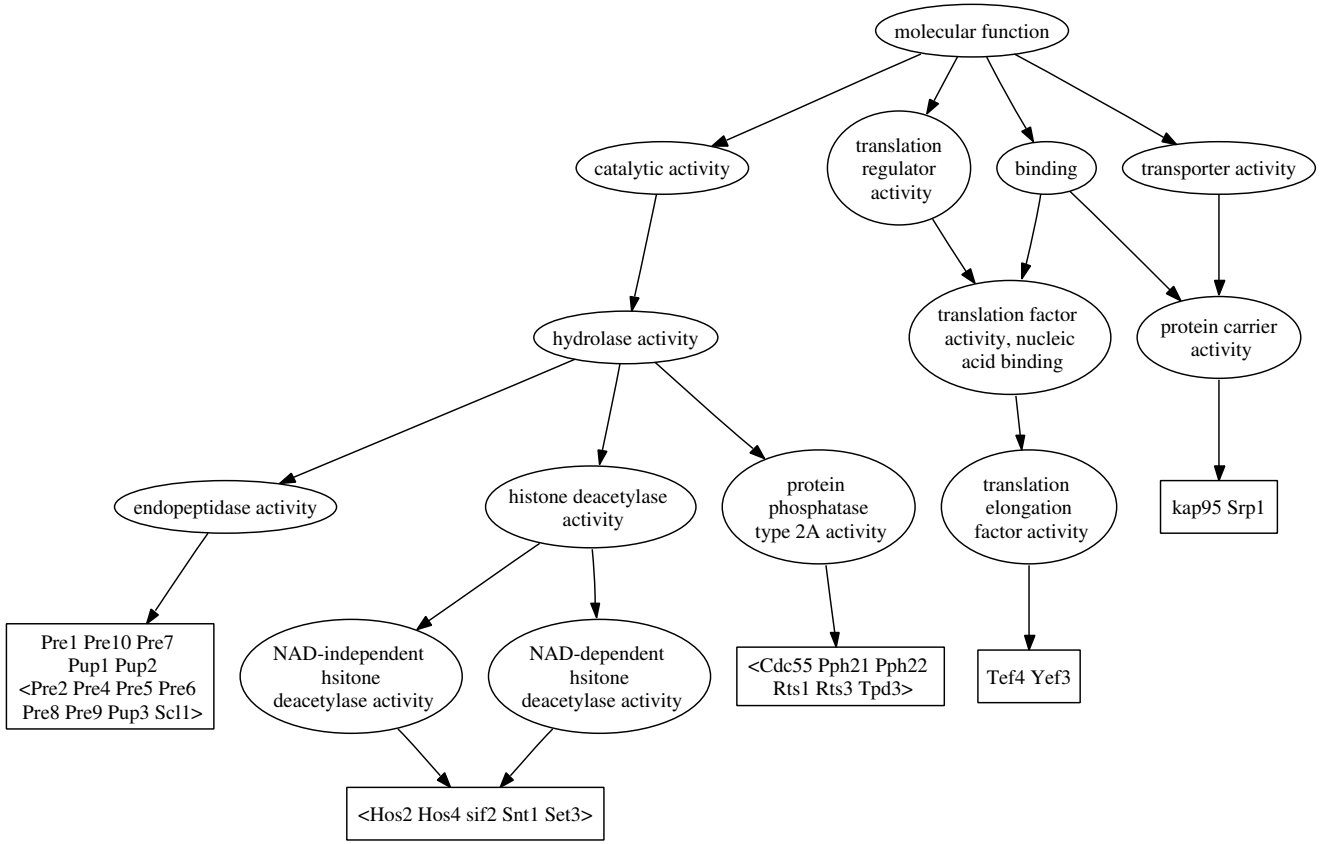
- All proteins in the pattern {Pre2, Pre4, Pre5, Pre6, Pre8, Pre9, Pup3, Scl1} perform the same biological function, *endopeptidase activity* and involve in the same biological process, *ubiquitin-dependent protein catabolism*.

Hyperclique Patterns as Functional Modules

CID	Protein Complexes	Function Category
106	Blm3 Dam1 Dbp9 Ecm29 Est3 Gfa1 Ino4 Kap95 Lys12 Mds3 Nud1 Pda1 Pdb1 Pre10 **Pre2** Pre3 **Pre4 Pre5 Pre6 Pre8 Pre9** Pse1 **Pup3** Rgr1 Rpt3 Rpt5 **Scl1** Spa2 Srp1 Ulp1 YFL006W YGR081C YMR310C YPL012W Yra1	Protein Synthesis and Turnover
148	Cdc6 Ecm29 Gfa1 Mlh2 Nas6 Pgk1 Pre1 **Pre2** Pre3 **Pre4 Pre5 Pre6** Pre7 **Pre8 Pre9 Pup3** Rpn10 Rpn11 Rpn12 Rpn13 Rpn3 Rpn5 Rpn6 Rpn7 Rpn8 Rpn9 Rpt1 Rpt2 Rpt3 Rpt4 Rpt5 Rpt6 **Scl1** Ubp6	Protein Synthesis and Turnover
157	Blm3 Cdc6 Ecm29 Mlh2 Pgk1 Pre1 Pre10 **Pre2** Pre3 **Pre4 Pre5 Pre6** Pre7 **Pre8 Pre9 Pup3** Rgr1 Rpn10 Rpn11 Rpn12 Rpn13 Rpn3 Rpn5 Rpn6 Rpn7 Rpn8 Rpn9 Rpt1 Rpt2 Rpt3 Rpt4 Rpt5 Rpt6 **Scl1** Ubp6 YFL006W	Protein Synthesis and Turnover
151	Blm3 Cdc55 Cin1 Erg13 Hhf2 Hos2 Iml1 Kap95 Kel1 Lte1 Myo5 Pfk1 Pph21 Pph22 Pre1 Pre10 **Pre2 Pre4 Pre5 Pre6** Pre7 **Pre8 Pre9** Pup1 Pup2 **Pup3** Rrd2 Rts1 **Scl1** Sif2 Srp1 Tdh2 Tdh3 Tef4 Tpd3 YBL104C YCR033W(Snt1) YGL245W YGR161C YIL112W(Hos4) YKR029C(Set3) Yef3 Yor1 Yra1 Zds1 Zds2	Signalling

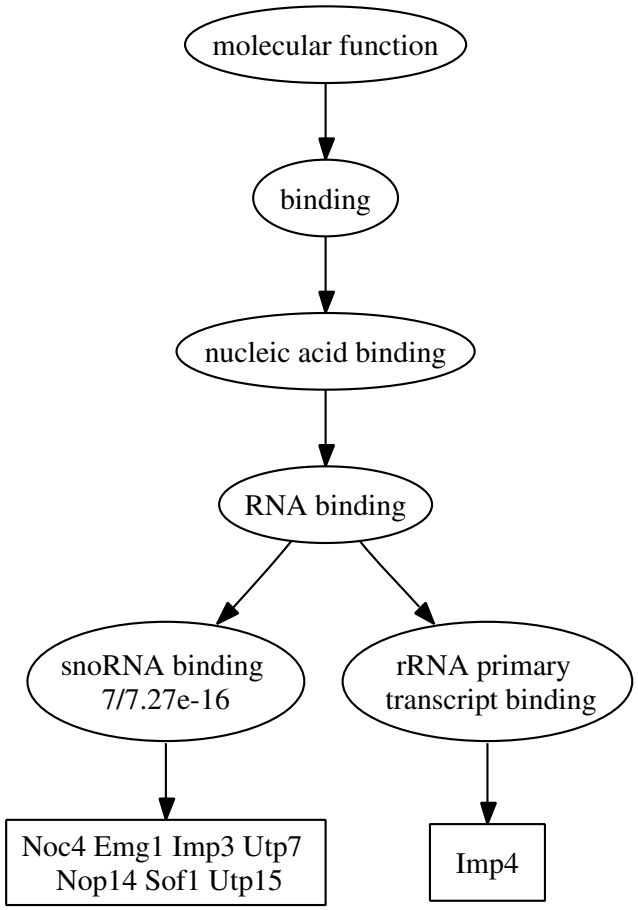
- The Hyperclique pattern {Pre2, Pre4, Pre5, Pre8, Pup3, Pre6, Pre9, Scl1} contained in four experimentally determined protein complexes.

Subgraph of GO function corresponding to the complex 151



- Proteins within a pair of $< >$ form a hyperclique pattern.
- Hyperclique patterns as functional modules to participate in a common complex.

Functional Annotation of Uncharacterized Proteins



- The hyperclique pattern:{Noc4, Emg1, Imp3, Imp4, Utp7, Mpp10, Nop14, Sof1, Utp15}.
- The protein `Mpp10` has no functional annotation.
- Physical interactions also identified among `Imp3`, `Imp4`, and `Mpp10` by two-hybrid genetic screen.
- Conclusion: Infer that `Mpp10` has the function “RNA binding”.

Overview

- Introduction
- Protein Complex Data
- Hyperclique Pattern Discovery
- Functional Modules in Protein Complexes
- Experimental Results

$\Rightarrow$ Summary

Summary

- Hyperclique pattern discovery for protein functional module extraction.
  - Go annotations validates hyperclique patterns as functional modules.
  - Physically evidence implying that proteins in the same pattern tend to physically interact together.
  - Functional Annotation of Uncharacterized Proteins

Acknowledgement

- Lawrence Berkeley National Laboratory
  - Dr. Chris Ding and Dr. Xiaofeng He
  - Dr. Stephen R. Holbrook and group members in the Holbrook Lab
- University of Minnesota - Twin cities
  - Professor Vipin Kumar

Questions?

- Personal Homepage - http://www.cs.umn.edu/∼huix

Thank You !