Overview

Research Summary: Protein Bioinformatics

As the 21st century unfolds, the biological sciences are being transformed by the advent of large-scale data. The sequencing of the human genome is a dramatic example of this. Simultaneous to this increase in biological data, computers and computation have had a transformative effect on the way information is handled, stored, and mined. These computational advances apply, of course, to many facets of life. The goal of my lab is to connect these two developments: harnessing computational advances for the analysis of large-scale biological data, principally by performing integrative surveys and systematic data mining.

More specifically, we are focused on protein bioinformatics: understanding the structure, function, and evolution of proteins through analyzing populations of them in databases and in whole-genome experiments. Overall we have four research foci, which follow a progression from surveying the overall genomic landscape to analyzing individual proteins and their interactions in more detail, to zooming in on the chemical structure of specific molecules.

1 Genomics: Mining and Annotating Intergenic Regions, especially in relation to Pseudogenes

We are involved in a number of large-scale collaborations (e.g. ENCODE) to probe the activity of intergenic regions with tiling array technology. We have developed tools to design, score and interpret these arrays and to highlight particular array artifacts. The overall conclusion from this work has been that much of the intergenic regions of the human genome appear to be active, both transcriptionally and in terms of protein binding. In connection with tiling array experiments, we have done an extensive amount of intergenic annotation, with a particular focus on mining intergenic regions for pseudogenes (protein fossils). We were, in fact, one of the first groups to perform comprehensive surveys of pseudogenes on a genome-wide scale in terms of protein families, which we did for human, worm, yeast and a number of other organisms. Collectively, our studies enable us to determine the common "pseudofolds" and "pseudofamilies" in various genomes and to address important evolutionary questions about the type of proteins that were present in the past history of an organism.

2 Proteomics: Using Networks to Mine Functional Genomic Data and Understand Protein Function

After the main elements of the human genome are identified, we need to characterize their function. We are trying to characterize gene function through molecular networks. We work on systematically integrating many weak functional genomic features with data mining techniques to predict protein networks (comprising protein interactions and other functional linkages). Some of the features integrated are obviously related to protein interactions (e.g. expression correlations), but many others such as gene essentiality are much less so. In addition, we have studied the structure of protein networks, both on a large scale in terms of global statistics (e.g. the diameter) and on a small scale in terms of local network motifs (e.g. hubs). In particular, we have correlated network hubs with gene essentiality. Most importantly, we extensively study the dynamics of networks. This has allowed us to show how a network dramatically changes in different conditions.

3 Structural Genomics: Analysis of Folds, Families and Functions on a Large Scale

Another area of research in our lab is structural genomics. Here, we conceptualize proteins not purely as character sequences or abstract network nodes, but more in terms of their molecular structure. We have examined the large-scale relationships between sequence, structure and function in order to understand the extent to which structural and functional annotation can reliably be transferred between similar sequences, particularly when similarity is expressed in modern probabilistic language. We have related the occurrence of protein folds and families to phylogeny and deep evolutionary history. Our studies enabled us to recognize that particular folds are more common in certain organisms than in others. Finally, as part of our work on structural genomics, we relate the properties of proteins with their eventual success at being purified and structurally characterized. This has been in the framework of a database and decision-tree mining framework that we have built for the NESG structural genomics consortium.

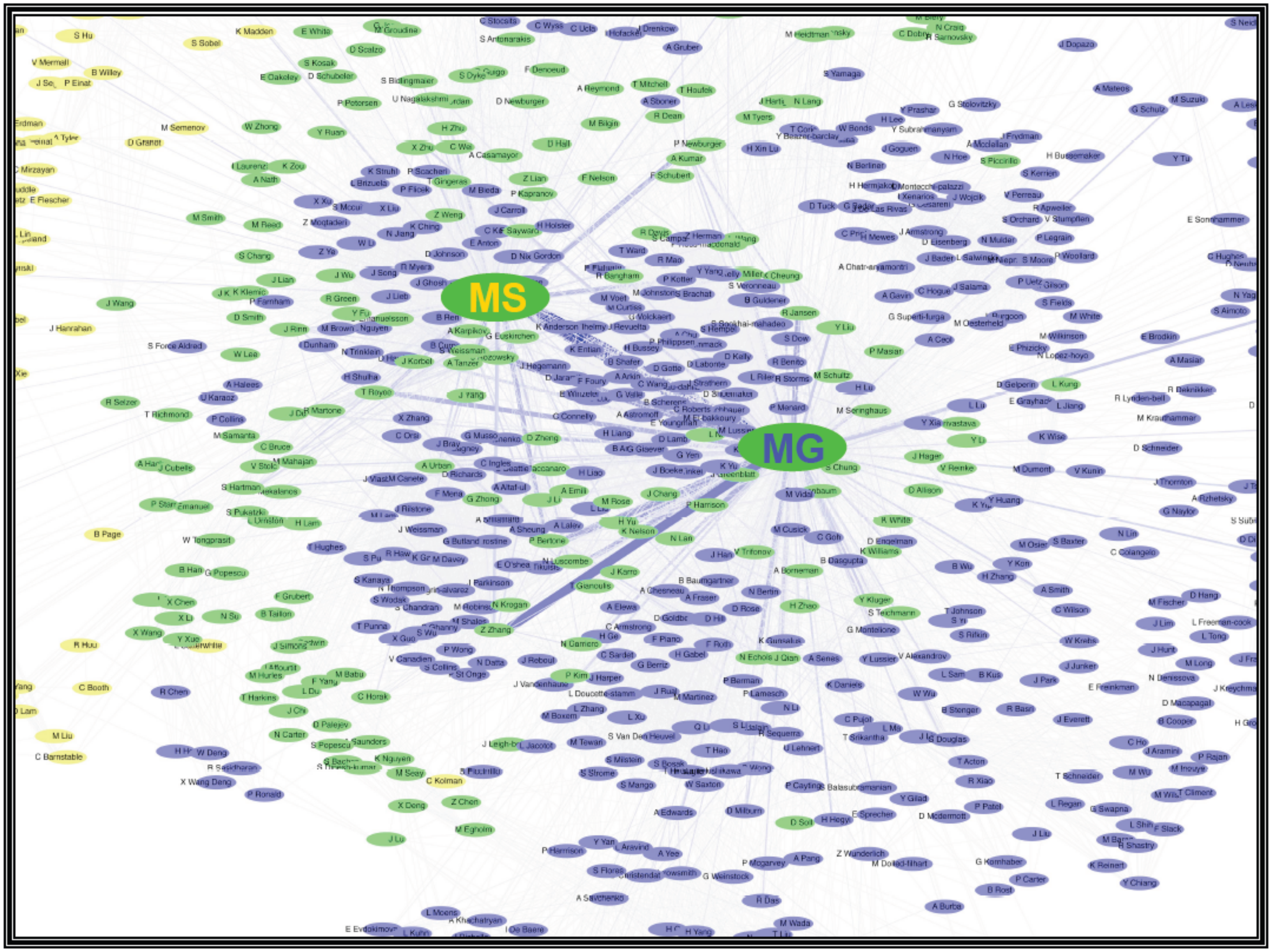
4 Computational Biophysics: Relating Macromolecular Motions and Packing

The final area of focus in the lab is analyzing small populations of structures in terms of their detailed 3D-geometry and physical properties. Here, we try to interpret macromolecular motions in terms of packing. We have set up a database of macromolecular motions and coupled it with simulation tools to interpolate between structural conformations; the database also has tools to predict likely motions based on simple models, such as normal modes and localized hinges connecting rigid domains. Part of this project involves devising a system for characterizing motions in a highly standardized fashion. Our motions classification scheme is motivated by the fact that protein interiors are packed exceedingly tightly, and the tight packing can greatly constrain a protein's mobility. We have developed tools for measuring and comparing the packing efficiency at different interfaces (e.g. inter-domain, protein surface, helix-helix, protein vs. RNA) using specialized geometric constructions (e.g. Voronoi polyhedra).

Summary & Broader Societal Issues

In summary, my lab acts a connector, bringing quantitative approaches from disciplines such as CS and applied math to bear on real questions and data in molecular biology. In particular, we have extensively applied classical computational approaches involving simulation, machine learning, and database design to biological problems. This often happens in the framework of practical, experimental collaborations, where we function as part of multi-disciplinary teams. Team participation is a key feature of the lab. Finally, as part of our mission to connect biology with computation, we have also extensively analyzed how a number of larger issues relating to computation in society impact biological research. In particular, we have examined how general aspects of e-publishing and digital libraries relate to biomedical databases and how various legal and security concerns significantly impact genomics database interoperation.

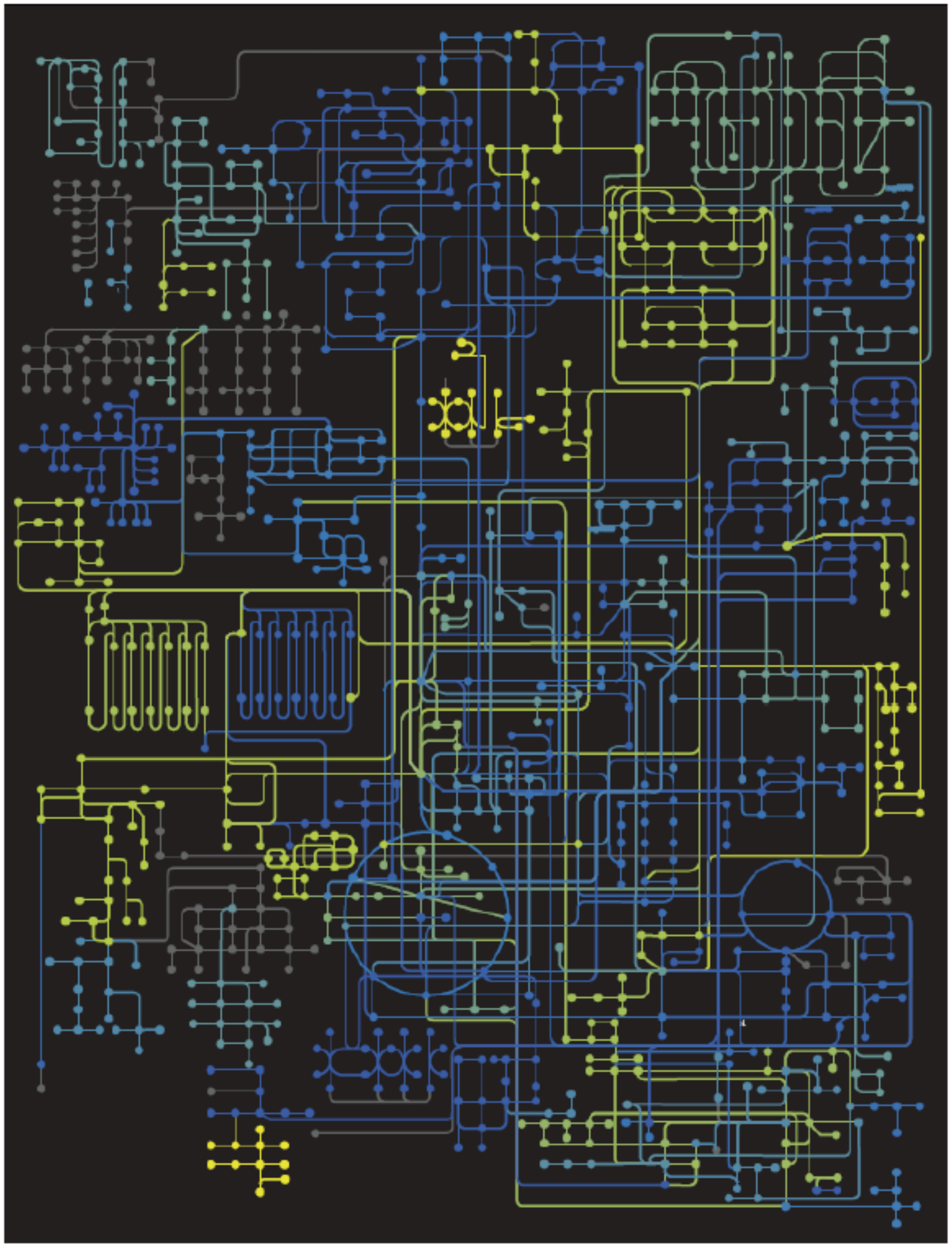
Networks.GersteinLab.org



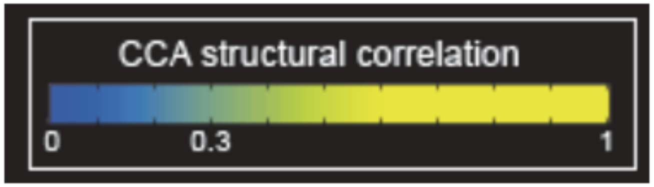
This is a research collaboration network centered on Dr. Mark Gerstein and Dr. Michael Snyder. Each eclipse stands for an individual researcher.

1

- Quantifying environmental adaptation of metabolic pathways in metagenomics.
- TA Gianoulis, J Raes, PV Patel, R Bjornson, JO Korbel, I Letunic, T Yamada, A Paccanaro, LJ Jensen, M Snyder, P Bork, MB Gerstein (2009) Proc Natl Acad Sci U S A 106: 1374-9.
- 2 figures

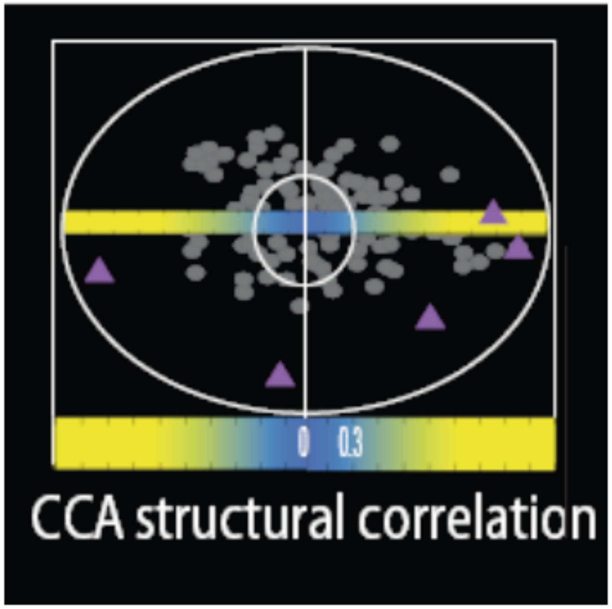


Strength of Pathway co-variation with environment

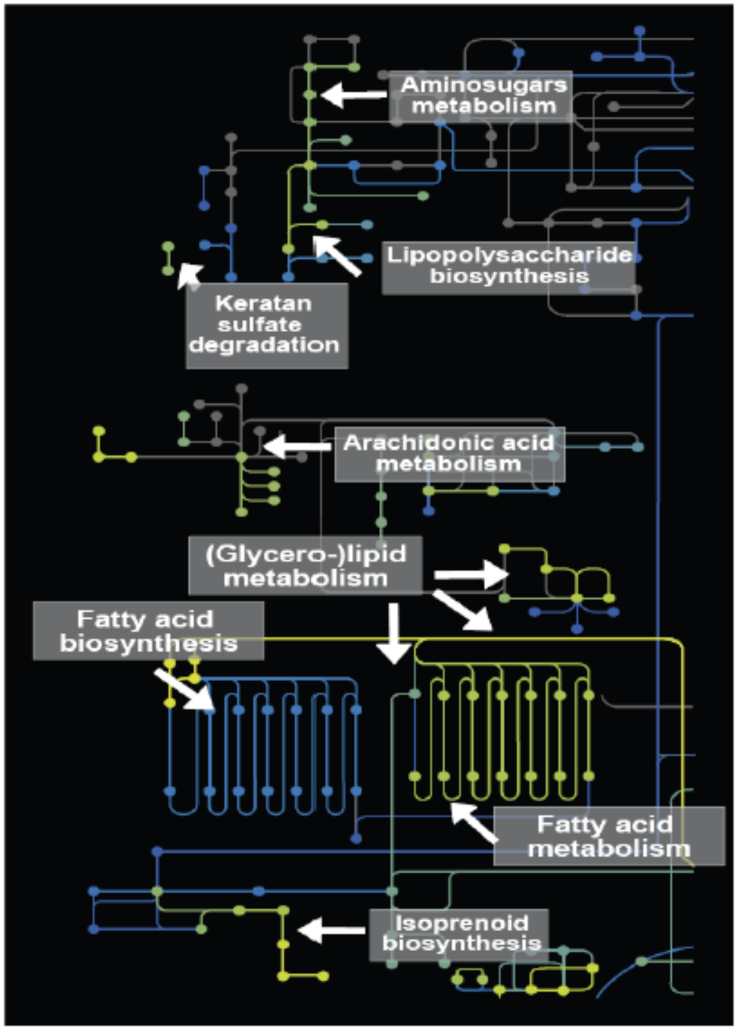


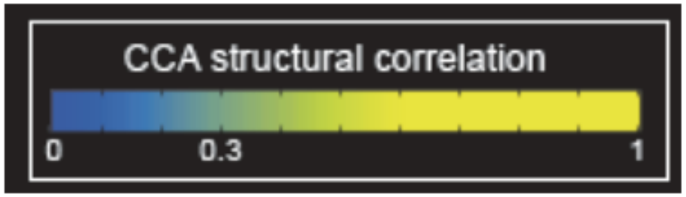
Environmentally invariant

Environmentally variant



[Gianoulis et al., PNAS (in press, 2009)]


Aminosugars metabolism
Lipopolysaccharide biosynthesis
Keratan sulfate degradation
Arachidonic acid metabolism
(Glycero-)lipid metabolism
Fatty acid biosynthesis
Fatty acid metabolism
Isoprenoid biosynthesis

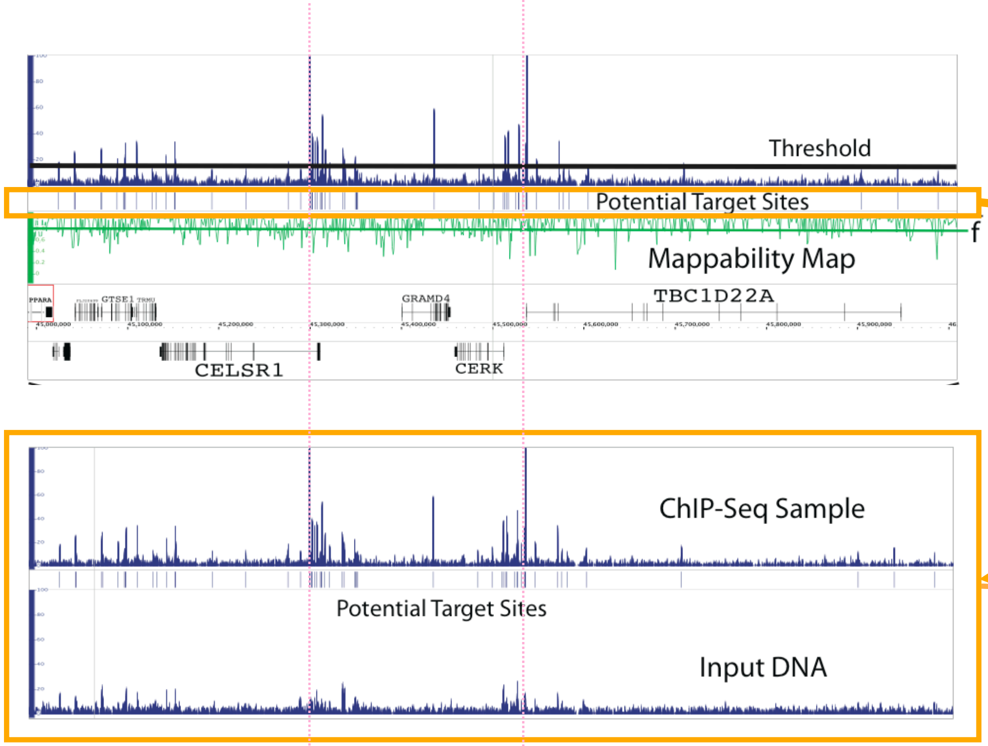


[Gianoulis et al., PNAS (in press, 2009)]

2

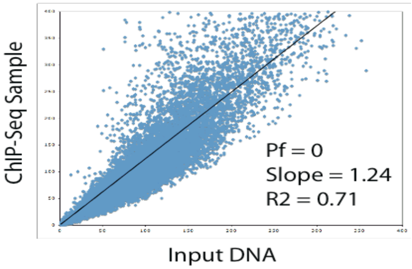
- PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.
- J Rozowsky, G Euskirchen, RK Auerbach, ZD Zhang, T Gibson, R Bjornson, N Carriero, M Snyder, MB Gerstein (2009) Nat Biotechnol 27: 66-75.
- 1 figure

PeakSeq: Scoring Relative to Controls

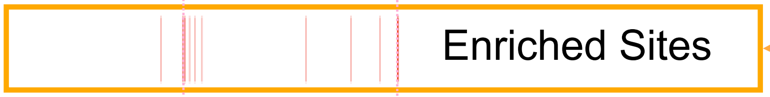


Filter for Potential Targets based on "Mappability" Simulation

Scale Input Relative to ChIP



Score Relative to Bionomial Expectation

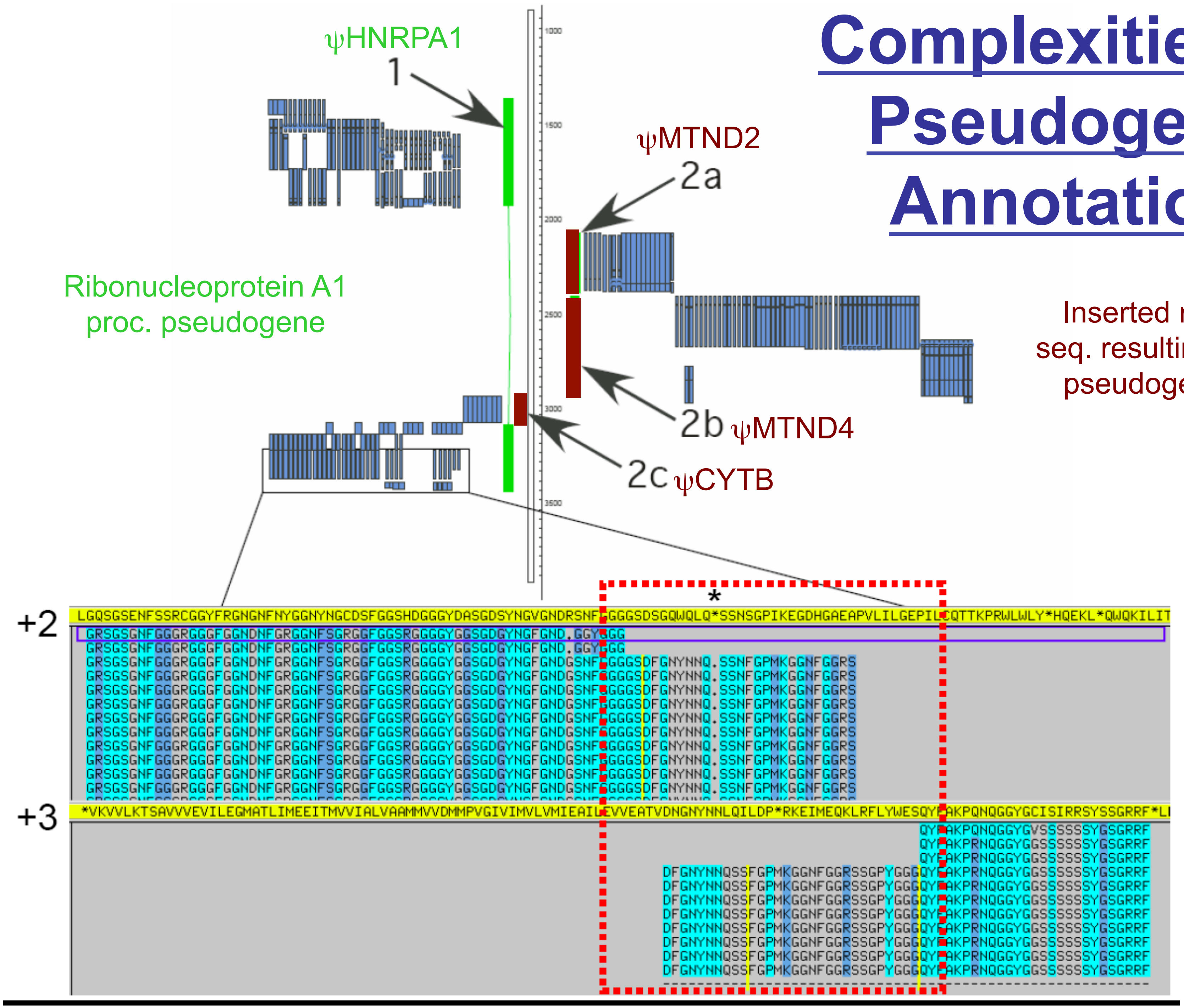


[Rozowsky et al. Nat. Biotech ('09)]

3

- Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution.
- D Zheng, A Frankish, R Baertsch, P Kapranov, A Reymond, SW Choo, Y Lu, F Denoeud, SE Antonarakis, M Snyder, Y Ruan, CL Wei, TR Gingeras, R Guigó, J Harrow, MB Gerstein (2007) Genome Res 17: 839-51.
- 2 figures

Complexities in Pseudogene Annotation



Ribonucleoprotein A1
proc. pseudogene

Inserted mito.
seq. resulting in 3
pseudogenes

+2

```
                                                                          *
LGQSGSENFSSRCGGYFRGNGNFNYGGNYNGCDSFGGSHDGGGYDASGDSYNGVGNDRSNFGGGGSDSGQWQLQ*SSNSGPIKEGDHGAEAPVLILGEPILCQTTKPRWLWLY*HQEKL*QWQKILIT
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGND.GGYGGG
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGND.GGYGGG
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
 GRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGSDFGNYNNQ.SSNFGPMKGGNFGGRS
```

+3

```
*VKVVLKTSAVVVEVILEGMATLIMEEITMVVIALVAAMMVVDMMPVGIVIMVLVMIEAILEVVEATVDNGNYNNLQILDP*RKEIMEQKLRFLYWESQYFAKPQNQGGYGCISIRRSYSSGRRF*L
                                                                                                  QYFAKPQNQGGYGVSSSSSSYGSGRRF
                                                                                                  QYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                                                  QYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    DFGNYNNQSSFGPMKGGNFGGRSSGPYGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    DFGNYNNQSSFGPMKGGNFGGRSSGPYGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    DFGNYNNQSSFGPMKGGNFGGRSSGPYGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    DFGNYNNQSSFGPMKGGNFGGRSSGPYGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    DFGNYNNQSSFGPMKGGNFGGRSSGPYGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    DFGNYNNQSSFGPMKGGNFGGRSSGPYGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    DFGNYNNQSSFGPMKGGNFGGRSSGPYGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    DFGNYNNQSSFGPMKGGNFGGRSSGPYGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF
                                                                    ---------------------------------------------------------
```

]

History of Pseudogene Preservation

representative pseudogenes drawn from 201 total

	A	B	C	D	E	F
human	☒	☒	☒	☒	☒	☒
chimp	☒	■	☒	☒	■	■
baboon	☒	☒	☒	☒	☒	■
macaque	☒	☒	☒	☒	☒	■
marmoset	☒	○	☒	■	■	■
galago	☒	○	☒	☒	■	■
rat	○	○	☒	■	☒	■
mouse	☒	○	☒	■	☒	■
rabbit	○	○	○	■	☒	■
cow	○	○	○	☒	○	■
dog	☒	○	○	■	☒	☒
rfbat	☒	○	○	☒	■	■
shrew	☒	○	○	☒	■	■
armadillo	☒	○	○	☒	○	■
elephant	☒	○	○	■	☒	■
tenrec	○	○	○	■	☒	☒
monodelphis	○	○	○	■	☒	■
platypus	○	○	○	■	☒	■
chicken	○	○	○	■	○	■
xenopus	○	○	○	○	○	☒
tetraodon	○	○	☒	■	○	☒
zebrafish	○	○	○	■	☒	■

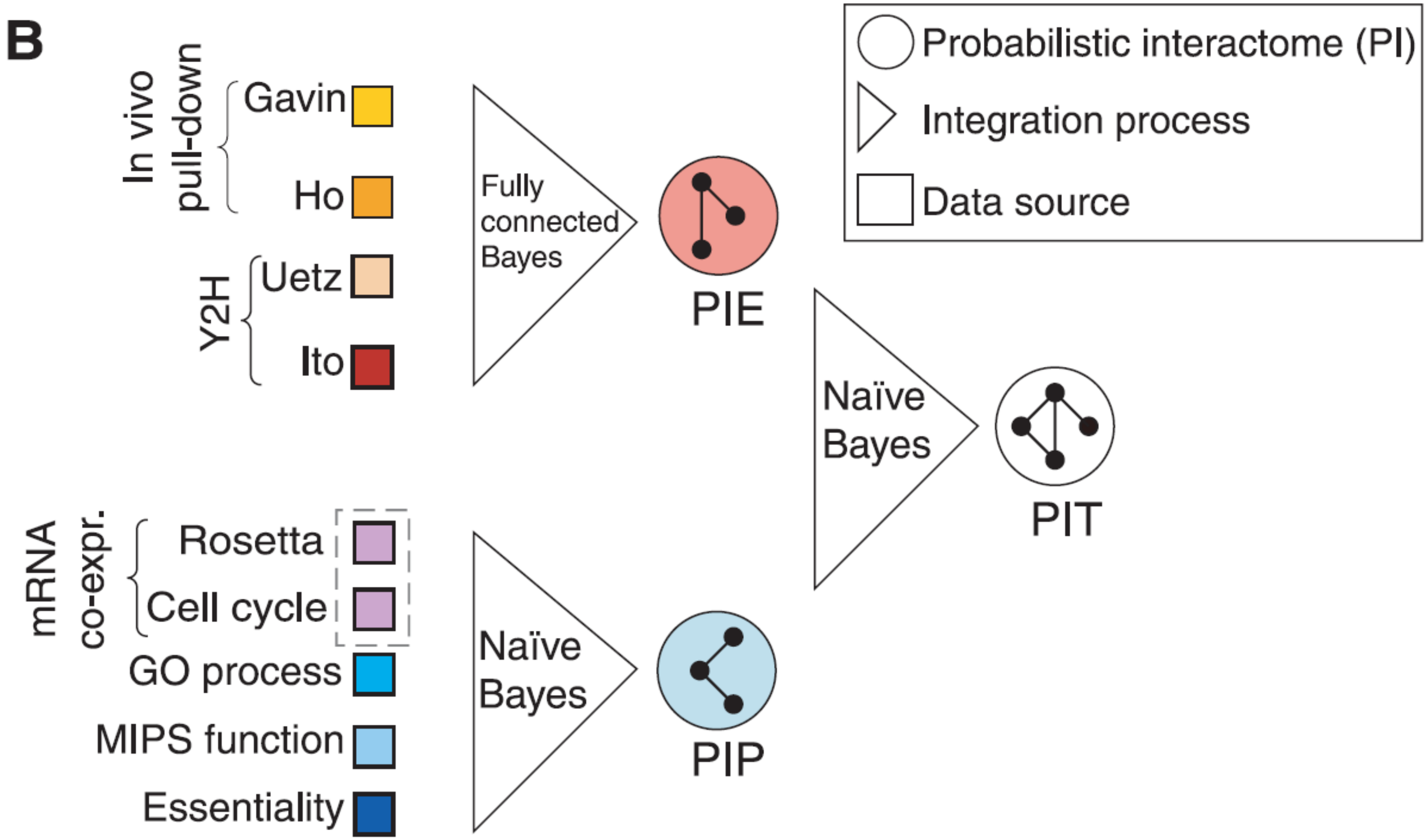
Based on alignment from ENCODE MSA group

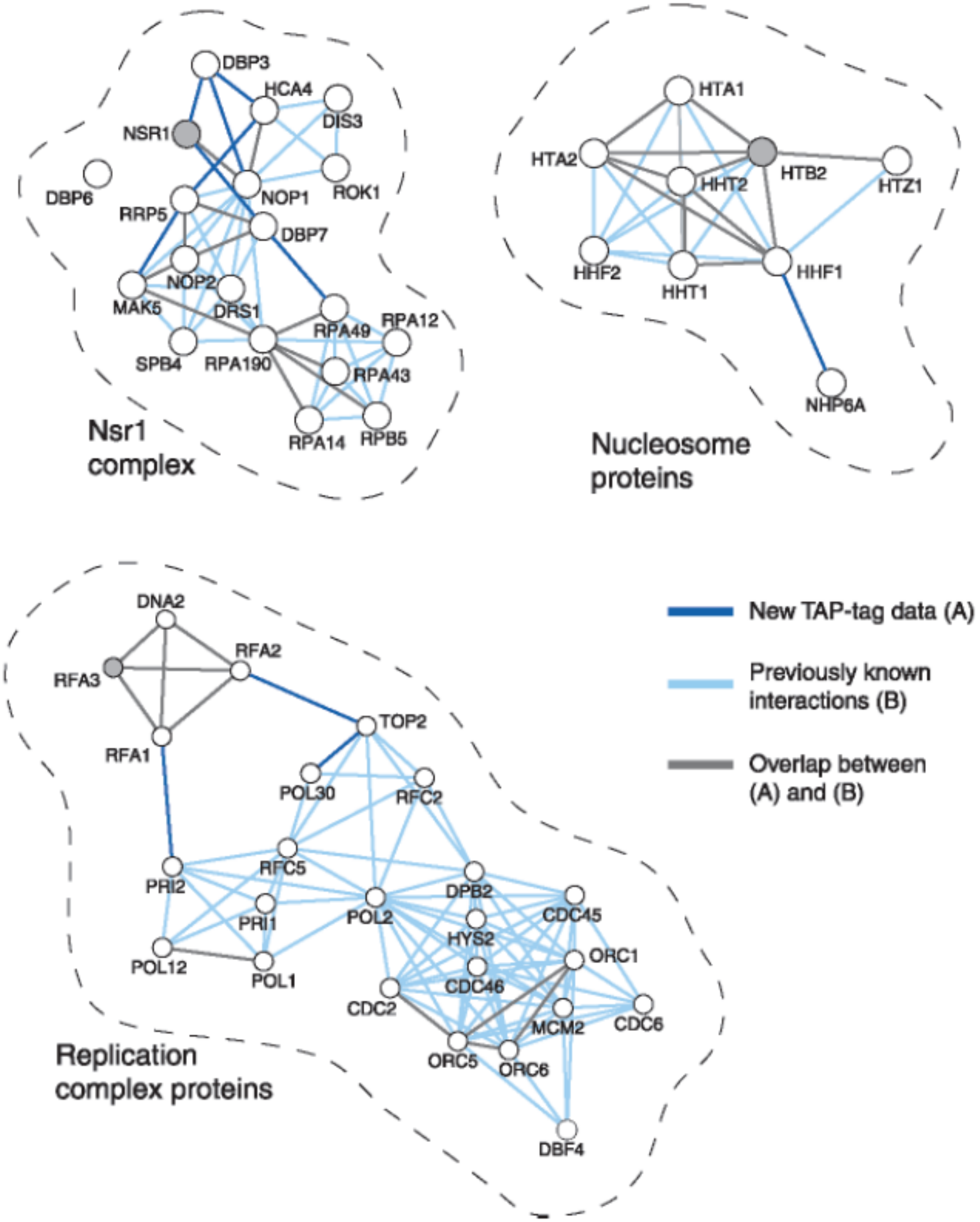
Zheng et al. (2007) Gen. Res.

Absent	○
Present with Disablement	☒
Present without Disablement	■

4

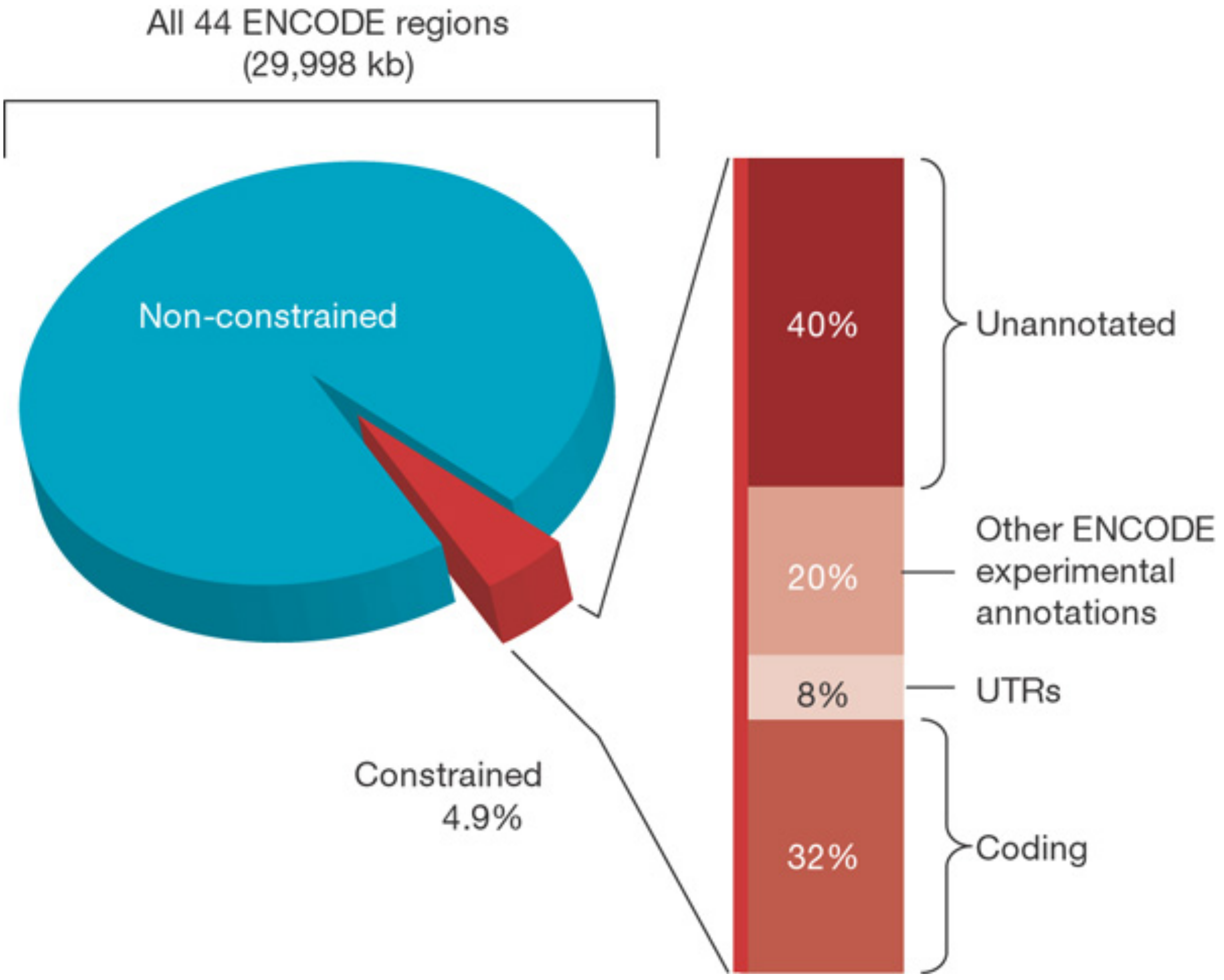
- A Bayesian networks approach for predicting protein-protein interactions from genomic data.
- R Jansen, H Yu, D Greenbaum, Y Kluger, NJ Krogan, S Chung, A Emili, M Snyder, JF Greenblatt, M Gerstein (2003) Science 302: 449-53.
- 2 figures

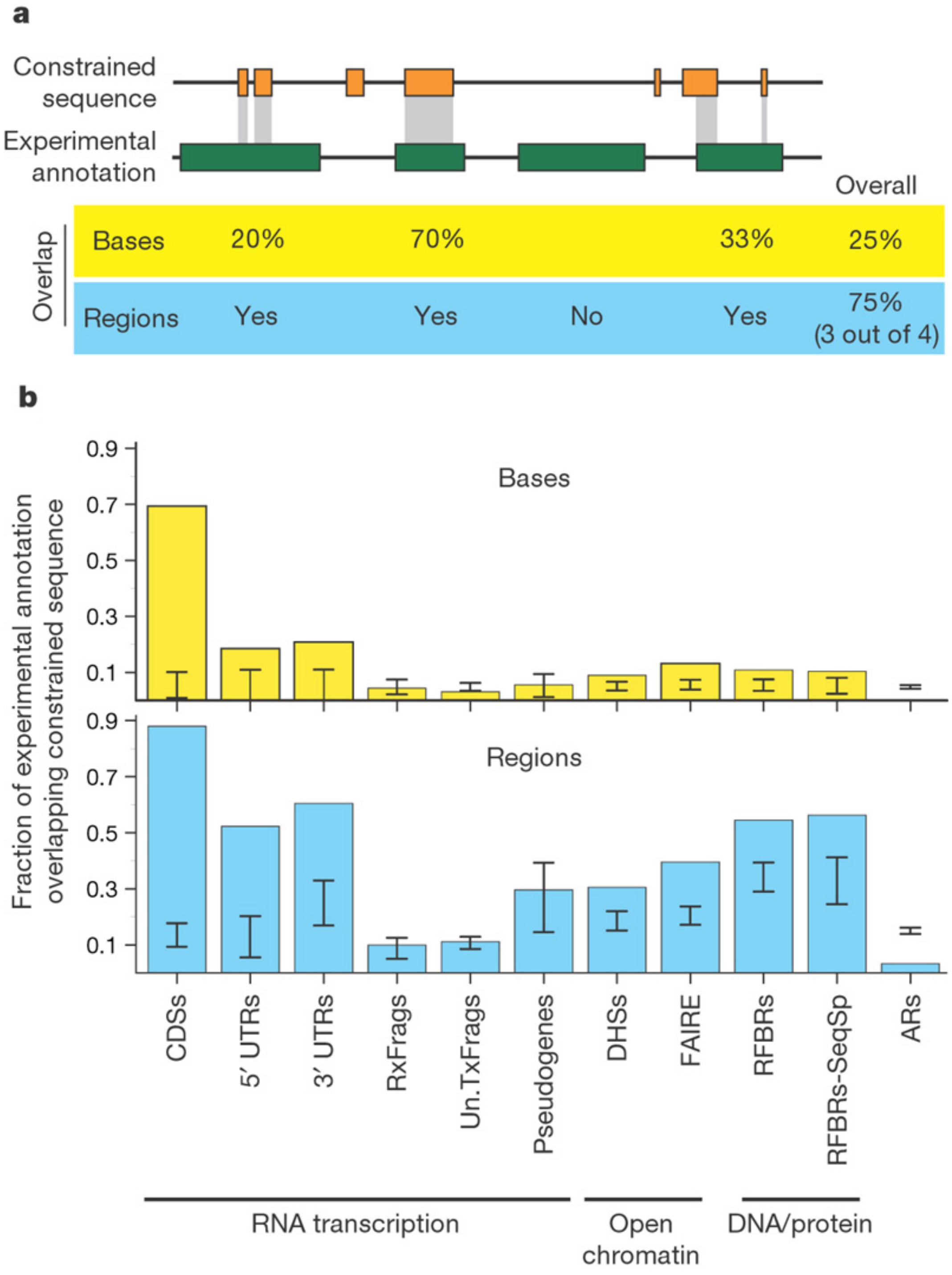

B
In vivo pull-down
Gavin
Ho
Y2H
Uetz
Ito
mRNA co-expr.
Rosetta
Cell cycle
GO process
MIPS function
Essentiality
Fully connected Bayes
PIE
Naïve Bayes
PIP
Naïve Bayes
PIT
Probabilistic interactome (PI)
Integration process
Data source


DBP3
HCA4
DIS3
NSR1
DBP6
NOP1
ROK1
RRP5
DBP7
NOP2
MAK5
DRS1
RPA49
RPA12
SPB4
RPA190
RPA43
RPA14
RPB5
Nsr1
complex
HTA1
HTA2
HTB2
HHT2
HTZ1
HHF2
HHT1
HHF1
NHP6A
Nucleosome
proteins
DNA2
RFA2
RFA3
TOP2
RFA1
POL30
RFC2
RFC5
PRI2
DPB2
PRI1
POL2
CDC45
HYS2
ORC1
POL12
POL1
CDC46
CDC2
MCM2
CDC6
ORC5
ORC6
DBF4
Replication
complex proteins
New TAP-tag data (A)
Previously known
interactions (B)
Overlap between
(A) and (B)

5

- Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.
- ENCODE Project Consortium (2007) *Nature* 447: 799-816.
- 2 figures


All 44 ENCODE regions
(29,998 kb)
Non-constrained
Constrained
4.9%
40%
Unannotated
20%
Other ENCODE experimental annotations
8%
UTRs
32%
Coding


a
Constrained sequence
Experimental annotation
Overall
Overlap
Bases
20%
70%
33%
25%
Regions
Yes
Yes
No
Yes
75% (3 out of 4)
b
Fraction of experimental annotation overlapping constrained sequence
Bases
Regions
0.9
0.7
0.5
0.3
0.1
CDSs
5′ UTRs
3′ UTRs
RxFrags
Un.TxFrags
Pseudogenes
DHSs
FAIRE
RFBRs
RFBRs-SeqSp
ARs
RNA transcription
Open chromatin
DNA/protein

6

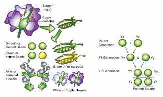
- What is a gene, post-ENCODE? History and updated definition.
- MB Gerstein, C Bruce, JS Rozowsky, D Zheng, J Du, JO Korbel, O Emanuelsson, ZD Zhang, S Weissman, M Snyder (2007) Genome Res 17:669-81.
- 1 figure

NimbleGen SYSTEMS, INC.

Gene: An Evolving Concept

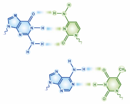
GENOME RESEARCH CSH

Gregor Mendel



The **laws of inheritance** were described.

1865

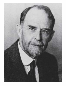


The **nucleic acids** were isolated and studied by Friedrich Miescher.

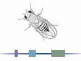
1869

The rediscovery of Mendel's work by Carl Correns, Erich von Tschermak-Seysenegg, and Hugo De Vries prompted the foundation of **genetics.**

1900

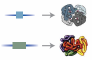


Thomas Morgan



Studies in *Drosophila melanogaster* suggest a **linear model of genes** on chromosomes, like 'beads on a string.'

1910



One gene, one enzyme; Then **one gene, one protein.**

Artificial **transmutation of the gene** by X-ray was reported by Hermann Müller.

1927

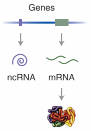
Francis Crick James Watson

The DNase experiment by Avery, MacLeod, and McCarty suggested **transformation is induced by DNA.**

1941 1944

The **DNA double helix** structure was solved.

1953



Protein

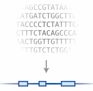
The '**Central Dogma**' of molecular biology was proposed by Francis Crick.

1958

The **first sequence** of a gene, *COAT_BPMS2*, was determined.

1972

The first large-scale **gene function** analysis using gene expression in yeast

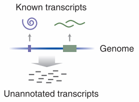


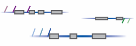
GENSCAN, a computer program for **gene structure prediction**, became available.

1994

The drafts of the **human genome sequence** were published.

1997 2001





Alternative promoters

The ENCODE Project highlighted the **complexity** of gene transcription and regulation.

2007

Gene as a discrete heredity unit — *Gene as a distinct locus* — *Gene as a physical molecule* — *Gene as a protein blueprint* — *Gene as transcribed code* — *Gene as ORF sequence pattern* — *Gene as annotated genomic entity* — *Gene as ...*

ZHENGDONG D. ZHANG MMVII

A term invented almost a century ago, 'gene,' with its beguilingly simple orthography, has become a central concept in biology. Given a specific meaning at its coinage, this word has evolved into something complex and elusive over the years, reflecting our ever-expanding knowledge in genetics and in life sciences at large. The stunning discoveries made in the ENCODE Project—like many before that significantly enriched the meaning of this term—are harbingers of another tide of change in our understanding of what a gene is.

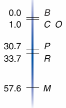
1909

The first appearance of the word '**gene**,' derived from the Greek *genesis* or *genos.*

Wilhelm Johannsen

1913

Alfred Sturtevant constructed the **first genetic map.**



1928

Griffith's experiment demonstrating type-switching in *pneumococcus* suggested a **transforming principle.**

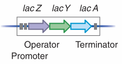
1952

Hershey and Chase determined that **DNA is the genetic material.**

Alfred Hershey Martha Chase

1961

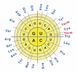
The **operon**, described by François Jacob and Jacques Monod, demonstrated **transcriptional control.**



lac Operon

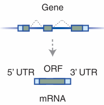
1965

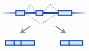
The **genetic code** was deciphered by Marshall Nirenberg, Har Gobind Khorana, and others.



1977

Introns and the mechanism of **RNA splicing** were discovered by Phillip Sharp and Richard Roberts demonstrating 'split gene structure.'





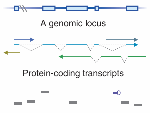
Alternative splicing

2003

The **ENCODE Project** was launched.

2007

The pilot phase of the ENCODE Project was finished. New gene models are proposed.



454 LIFE SCIENCES

POSTER DESIGN: ZHENGDONG D. ZHANG, YALE UNIVERSITY. IMAGE CREDITS: THE PORTRAITS OF GREGOR MENDEL AND WILHELM JOHANNSEN ARE USED BY COURTESY OF WIKIPEDIA. THE PORTRAIT OF THOMAS MORGAN IS LICENSED FROM THE CORBIS CORPORATION. THE PHOTOGRAPH OF FRANCIS CRICK AND JAMES WATSON IS COURTESY OF THE JAMES D. WATSON COLLECTION / COLD SPRING HARBOR LABORATORY ARCHIVES. THE PHOTOGRAPH OF ALFRED HERSHEY AND MARTHA CHASE IS COURTESY OF THE COLD SPRING HARBOR LABORATORY ARCHIVES. THE ILLUSTRATION OF MENDEL'S EXPERIMENTS WITH THE PEA PLANT IS GENEROUSLY PROVIDED BY JANE WANG, THE SCIENCE CREATIVE QUARTERLY. THE DRAWINGS OF PROTEIN MOLECULES ARE COURTESY OF IRVING GEIS. THE IMAGE OF THE BACTERIOPHAGE MS2 (PDB ID: 2MS2) IS KINDLY PROVIDED BY DR. JEAN-YVES SGRO, INSTITUTE FOR MOLECULAR VIROLOGY, UNIVERSITY OF WISCONSIN-MADISON. THE DRAWING OF THE DNA DOUBLE HELIX STRUCTURE IS PROVIDED BY THE NATIONAL LIBRARY OF MEDICINE. THE ENCODE PROJECT LOGO IS USED WITH PERMISSION FROM THE NATIONAL HUMAN GENOME RESEARCH INSTITUTE.

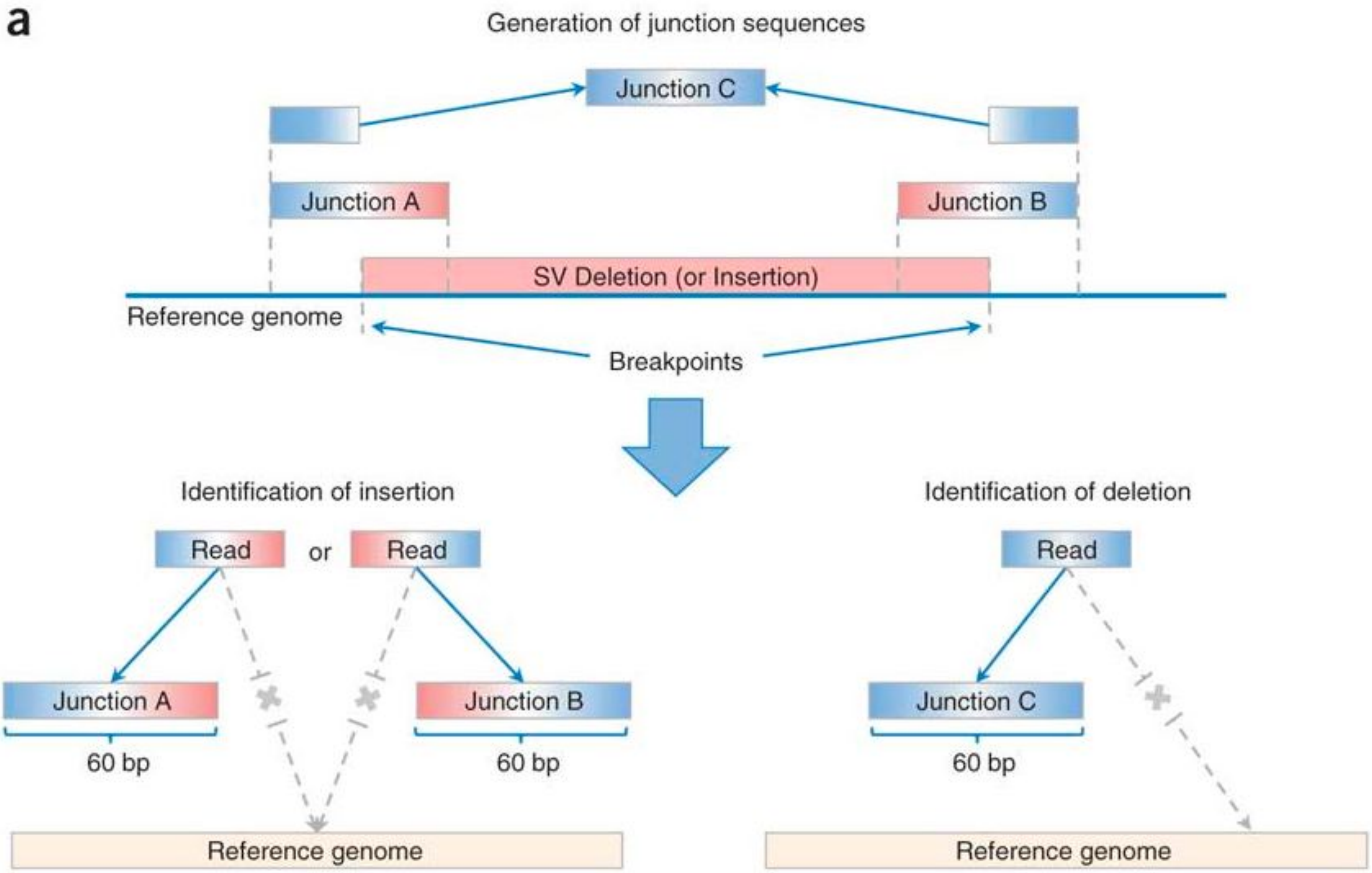
Roche Diagnostics

7

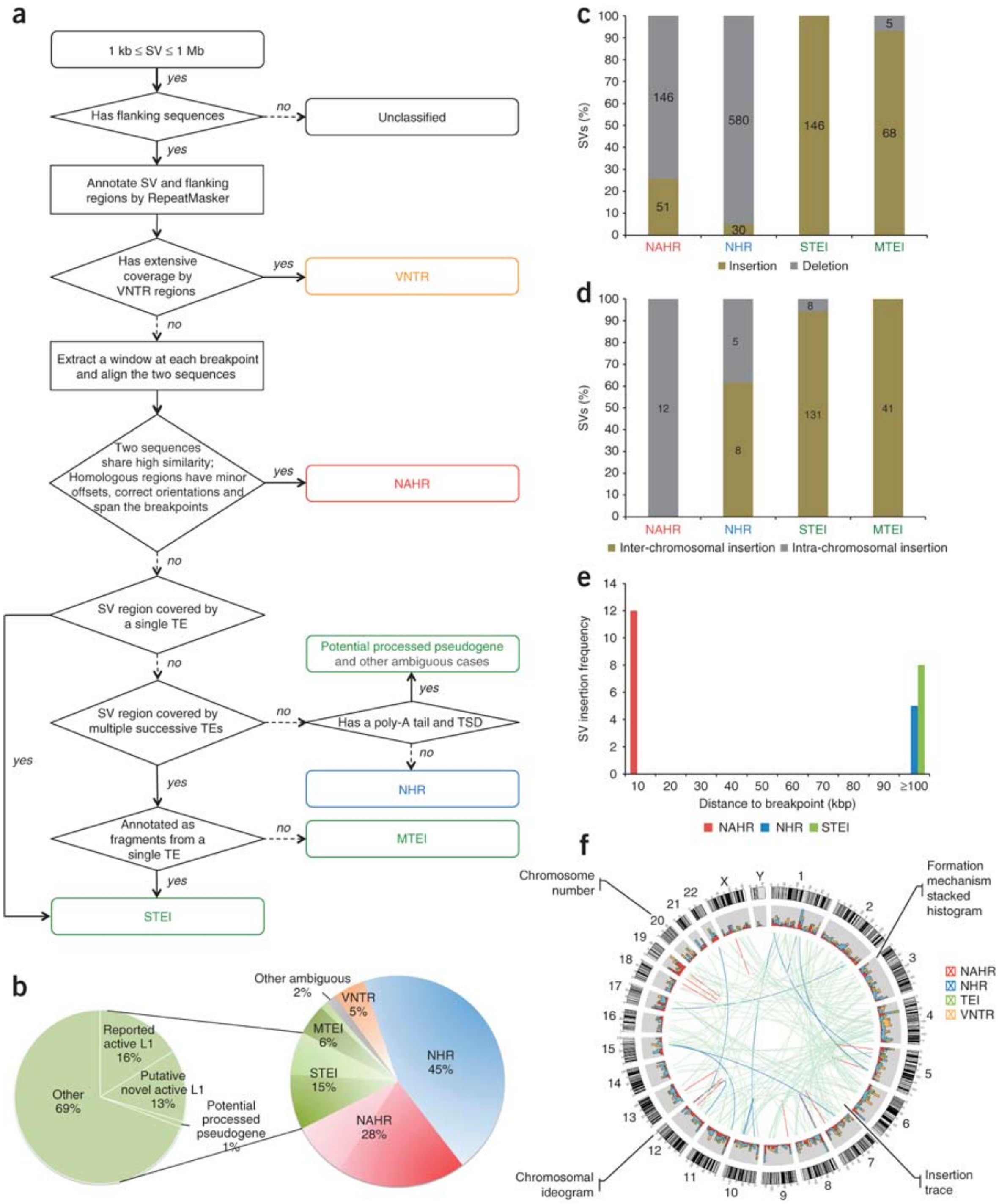
- "Personal genomics requires redefining privacy The human blueprint: dangerous secrets"
- D Greenbaum, M Gerstein. (2008) Insight, Nov. 2, Page 2 -- SF Chronicle
- No figures

8

- Nucleotide-resolution analysis of structural variants using BreakSeq
  and a breakpoint library.
- HY Lam, XJ Mu, AM Stütz, A Tanzer, PD Cayting, M Snyder, PM Kim, JO
  Korbel, MB Gerstein (2010) Nat Biotechnol 28: 47-55.
- 2 figures



Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match


a
1 kb ≤ SV ≤ 1 Mb
yes
Has flanking sequences
no
Unclassified
yes
Annotate SV and flanking regions by RepeatMasker
Has extensive coverage by VNTR regions
yes
VNTR
no
Extract a window at each breakpoint and align the two sequences
Two sequences share high similarity; Homologous regions have minor offsets, correct orientations and span the breakpoints
yes
NAHR
no
SV region covered by a single TE
no
Potential processed pseudogene and other ambiguous cases
yes
SV region covered by multiple successive TEs
no
Has a poly-A tail and TSD
no
yes
NHR
yes
Annotated as fragments from a single TE
no
MTEI
yes
STEI
b
Other 69%
Reported active L1 16%
Putative novel active L1 13%
Potential processed pseudogene 1%
Other ambiguous 2%
VNTR 5%
MTEI 6%
STEI 15%
NHR 45%
NAHR 28%
c
SVs (%)
100
90
80
70
60
50
40
30
20
10
0
146
51
580
30
146
5
68
NAHR
NHR
STEI
MTEI
Insertion
Deletion
d
SVs (%)
12
5
8
8
131
41
NAHR
NHR
STEI
MTEI
Inter-chromosomal insertion
Intra-chromosomal insertion
e
SV insertion frequency
14
12
10
8
6
4
2
0
10
20
30
40
50
60
70
80
90
≥100
Distance to breakpoint (kbp)
NAHR
NHR
STEI
f
Chromosome number
Formation mechanism stacked histogram
Chromosomal ideogram
Insertion trace
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
NAHR
NHR
TEI
VNTR

9

- Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context.
- PM Kim, JO Korbel, MB Gerstein (2007) Proc Natl Acad Sci U S A 104: 20274-9.
- 2 figures

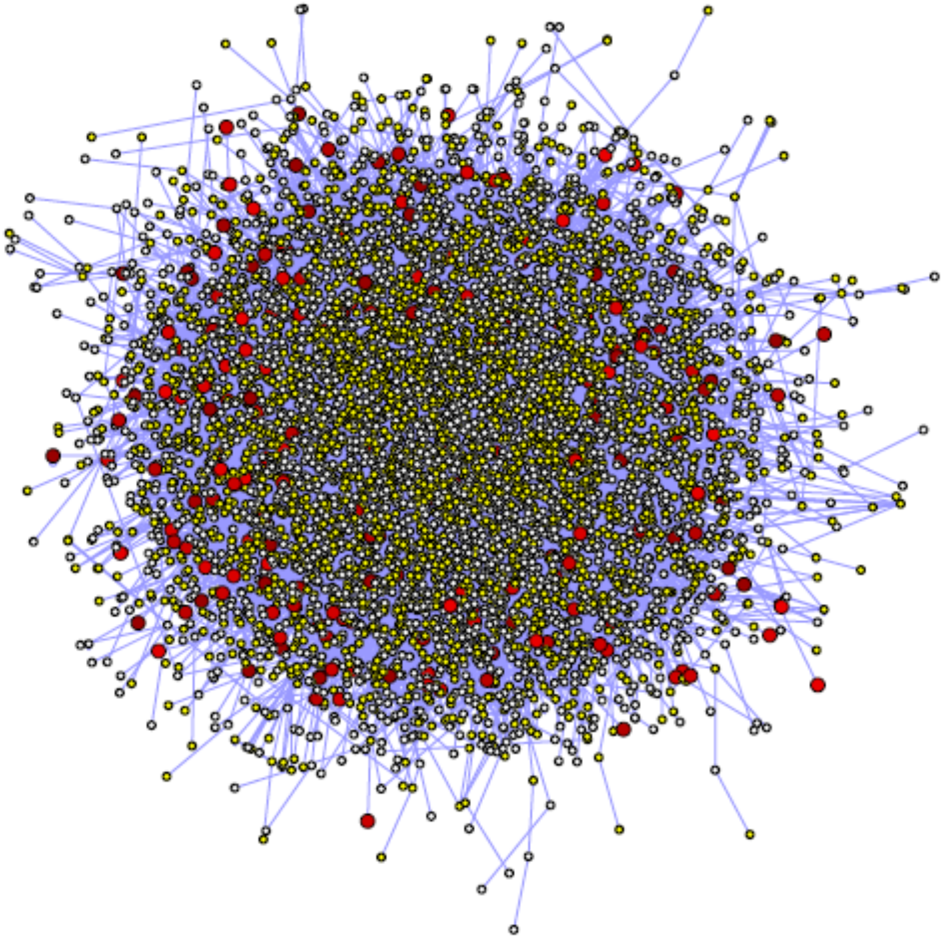


Fig. 1. The human protein interaction network and its connection to positive selection. Proteins likely to be under positive selection are colored in shades of red (light red, low likelihood of positive selection; dark red, high likelihood) (6). Proteins estimated not to be under positive selection are in yellow, and proteins for which the likelihood of positive selection was not estimated are in white (6).

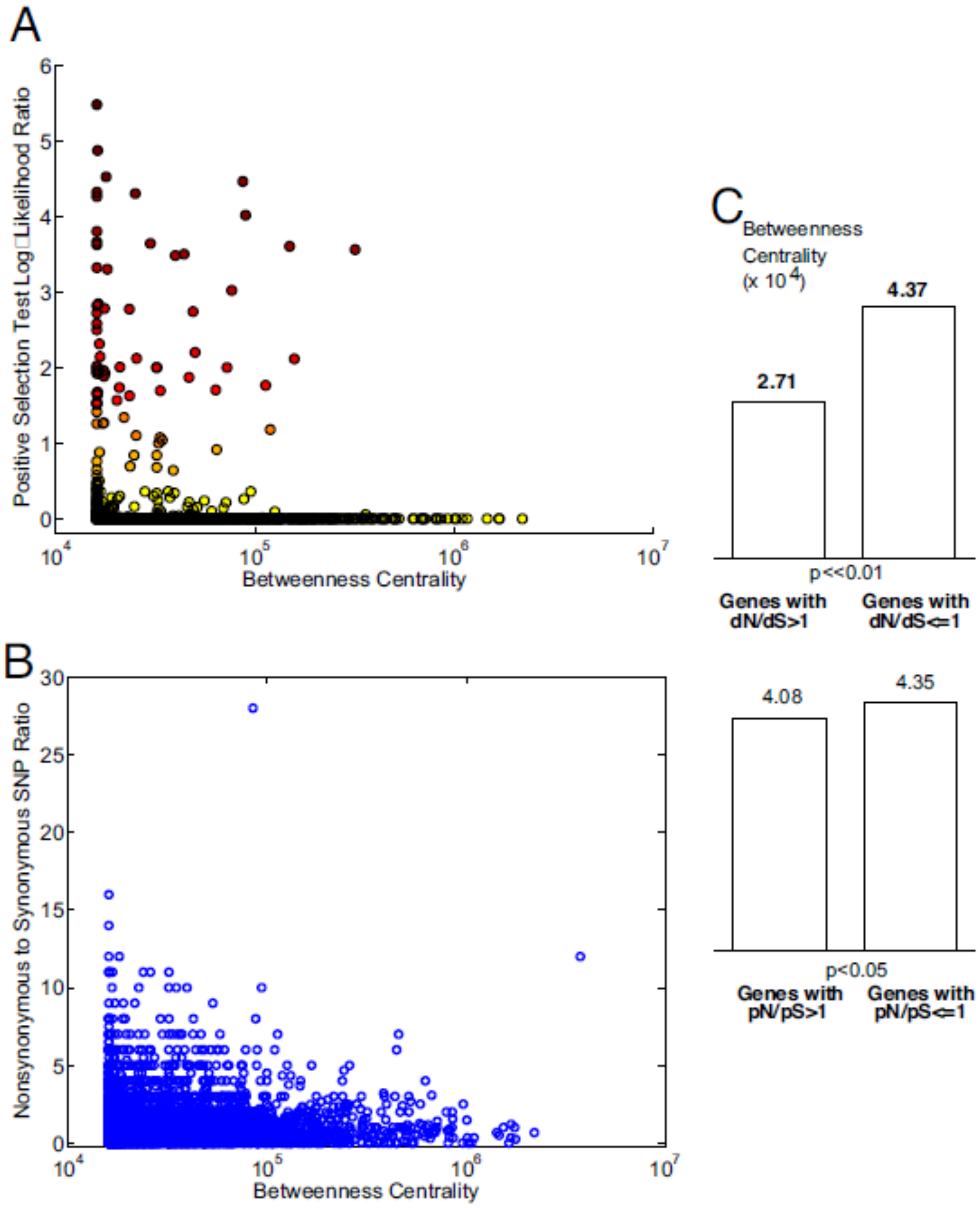


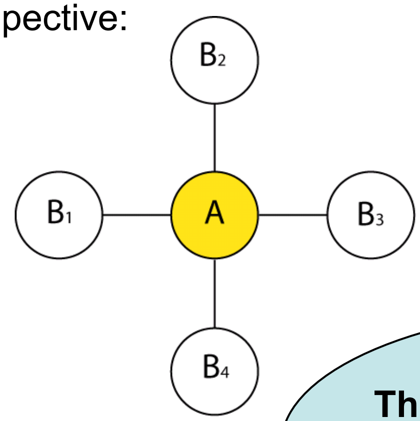
Fig. 2. Relationship of protein network centrality and single-nucleotide changes. (*A*) The periphery of the human interactome is strongly enriched for genes under positive selection. Shown is the correlation of the likelihood to be positively selected (6) and betweenness centrality (18). Dots are colored according to the same scheme as in Fig. 1. As expected for a highly significant Spearman rank correlation, almost all dots are near the *x* axis for high betweenness centralities, whereas high probabilities for positive selection are only observed at low betweenness centralities (Spearman $\rho = -0.06$, significant at *P* = 1.2e-06). (*B*) The periphery of the human interaction network is more variable on the protein sequence level. Shown is the ratio of nonsynonymous to synonymous SNPs vs. network centrality. A higher ratio (which corresponds to variability at the protein sequence level) tends to occur at the network periphery (Spearman $\rho = -0.1$, significant at *P* = 4.0e-04). (*C Upper*) Betweenness centrality of genes with some likelihood of being under positive selection (with a log-likelihood ratio >0) vs. all other genes. (*C Lower*) Betweenness centrality of genes with a high ratio of nonsynonymous to synonymous SNPs vs. genes with a low ratio of nonsynonymous to synonymous SNPs. The significance level of the differences is given as the Wilcoxon rank sum *P* value between the bars.

10

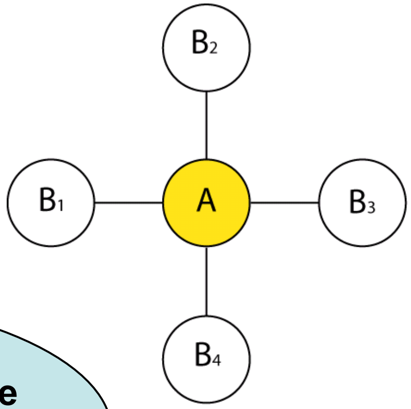
- Relating three-dimensional structures to protein networks provides evolutionary insights.
- PM Kim, LJ Lu, Y Xia, MB Gerstein (2006) Science 314: 1938-41
- 5 figures

MOTIVATION

Network perspective:

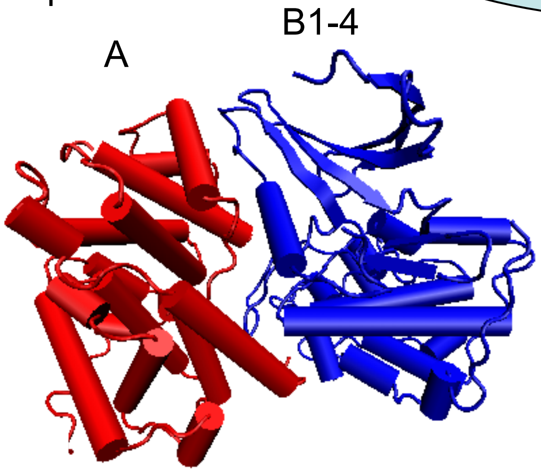


=



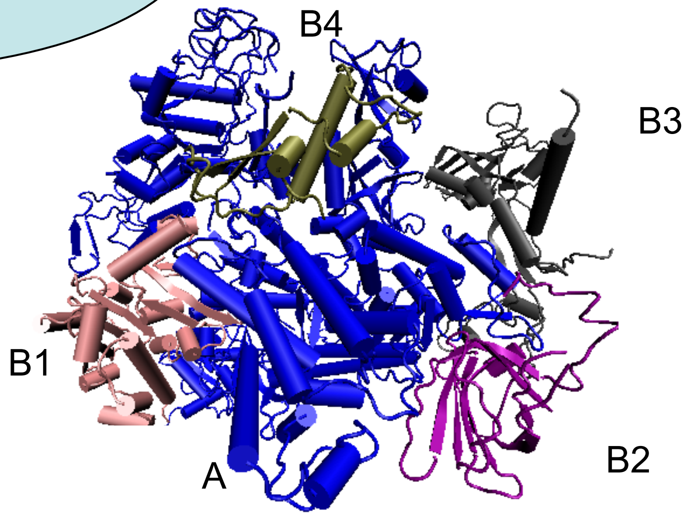
There remains a rich source of knowledge unmined by network theorists!

Structural biology perspective:



Cdk/cyclin complex

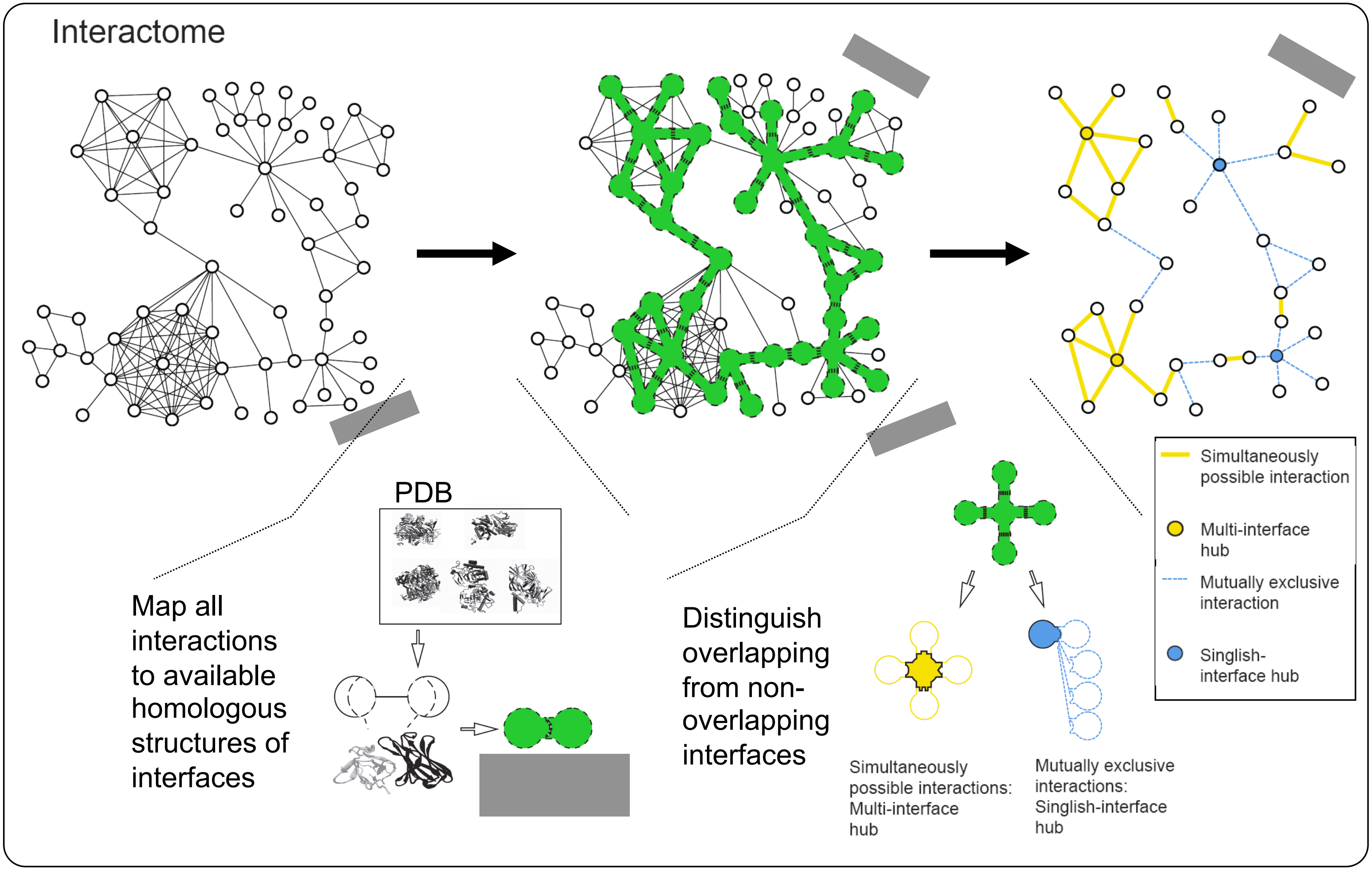
≠



Part of the RNA-pol complex

Source: Kim et al. *Science* (2006)

UTILIZING PROTEIN CRYSTAL STRUCTURES, WE CAN DISTINGUISH THE DIFFERENT BINDING INTERFACES


Interactome
PDB
Simultaneously possible interaction
Multi-interface hub
Mutually exclusive interaction
Singlish-interface hub

Map all interactions to available homologous structures of interfaces

Distinguish overlapping from non-overlapping interfaces

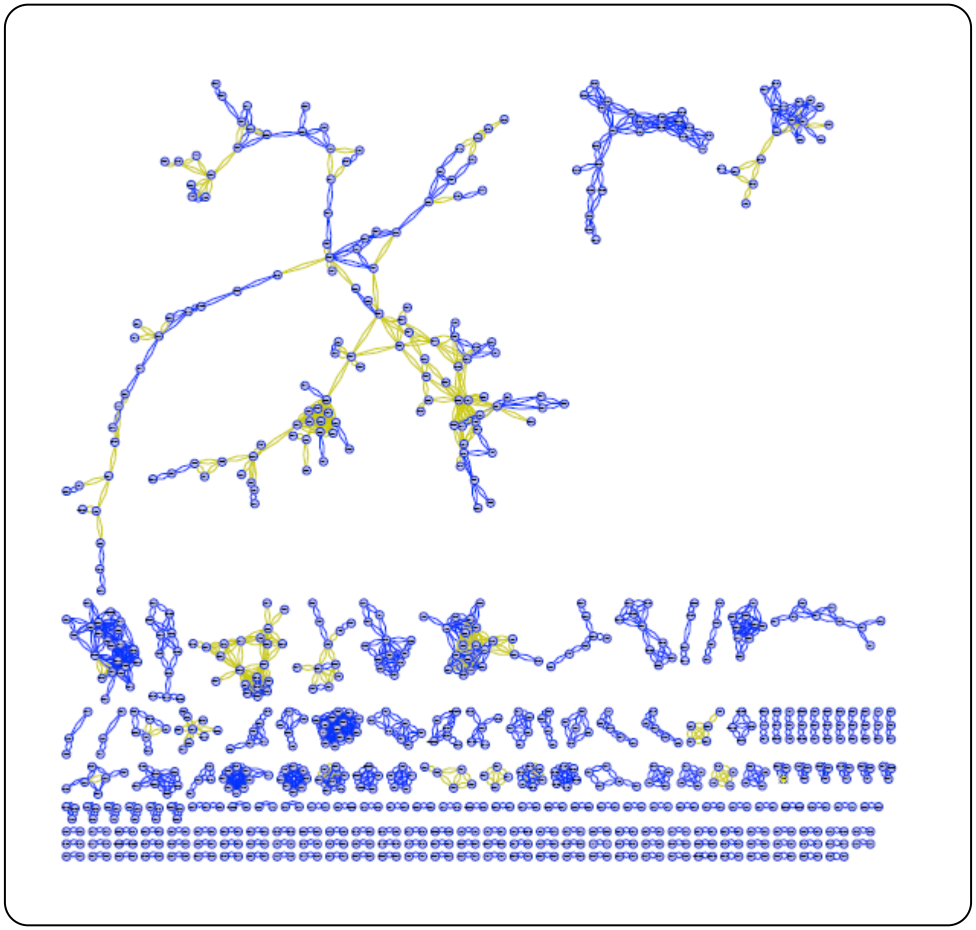
Simultaneously possible interactions: Multi-interface hub

Mutually exclusive interactions: Singlish-interface hub

Source: Kim et al. *Science* (2006)

THAT IS HOW THE RESULTING NETWORK LOOKS LIKE

- **The Structural Interaction Network (SIN)**



- Represents a “very high confidence” network
- Total of 873 nodes and 1269 interactions, each of which is structurally characterized
- 438 interactions are classified as mutually exclusive and 831 as simultaneously possible
- While much smaller than DIP, it is of similar size as other high-confidence datasets

Source: PDB, Pfam, iPfam and Kim et al. *Science* (2006)

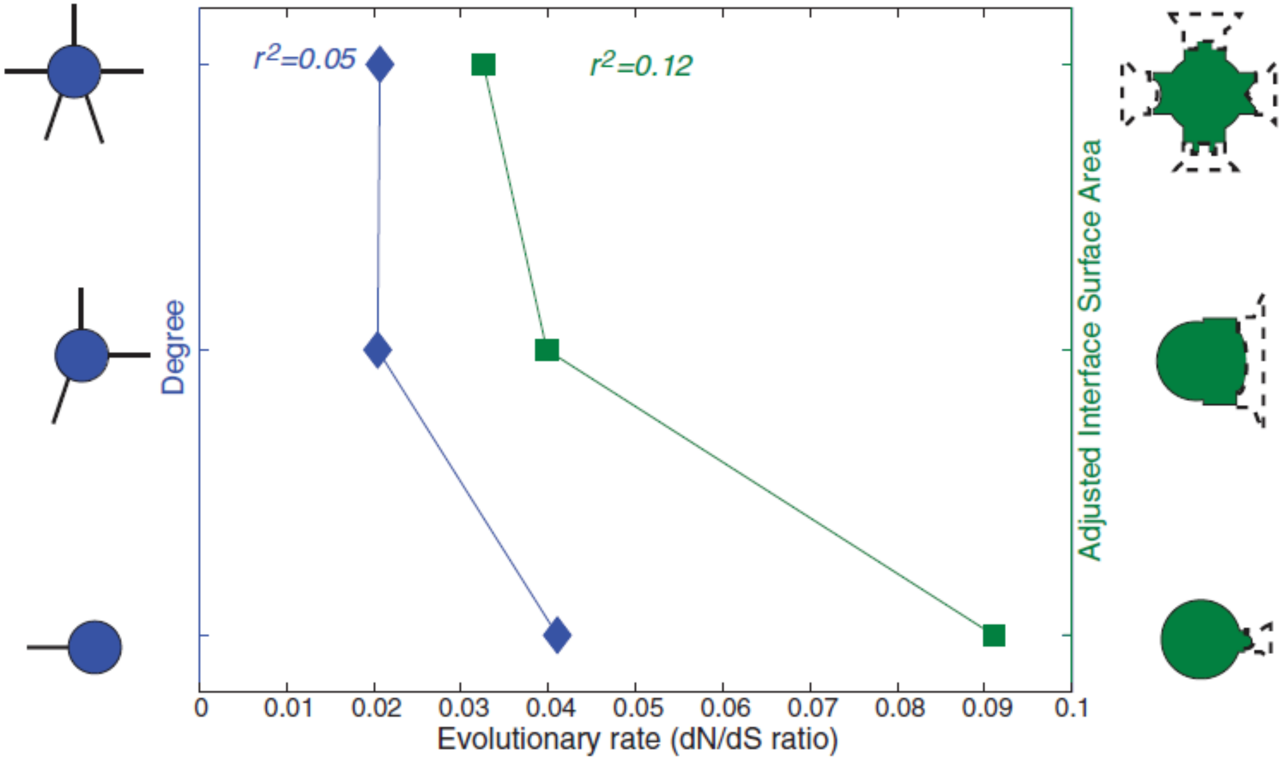


Fig. 2. Dependence of the average evolutionary rate (*dN/dS* ratio) of a protein with the degree and the interacting accessible surface area (adjusted by protein size, as estimated from molecular weight). For the degree correlation coefficient, we get $r^2 = 0.05$, and for the adjusted interface surface area, $r^2 = 0.12$, suggesting that more than twice as much of the variation in *dN/dS* is accounted for by adjusted interface surface area (12%) than by the degree (5%).

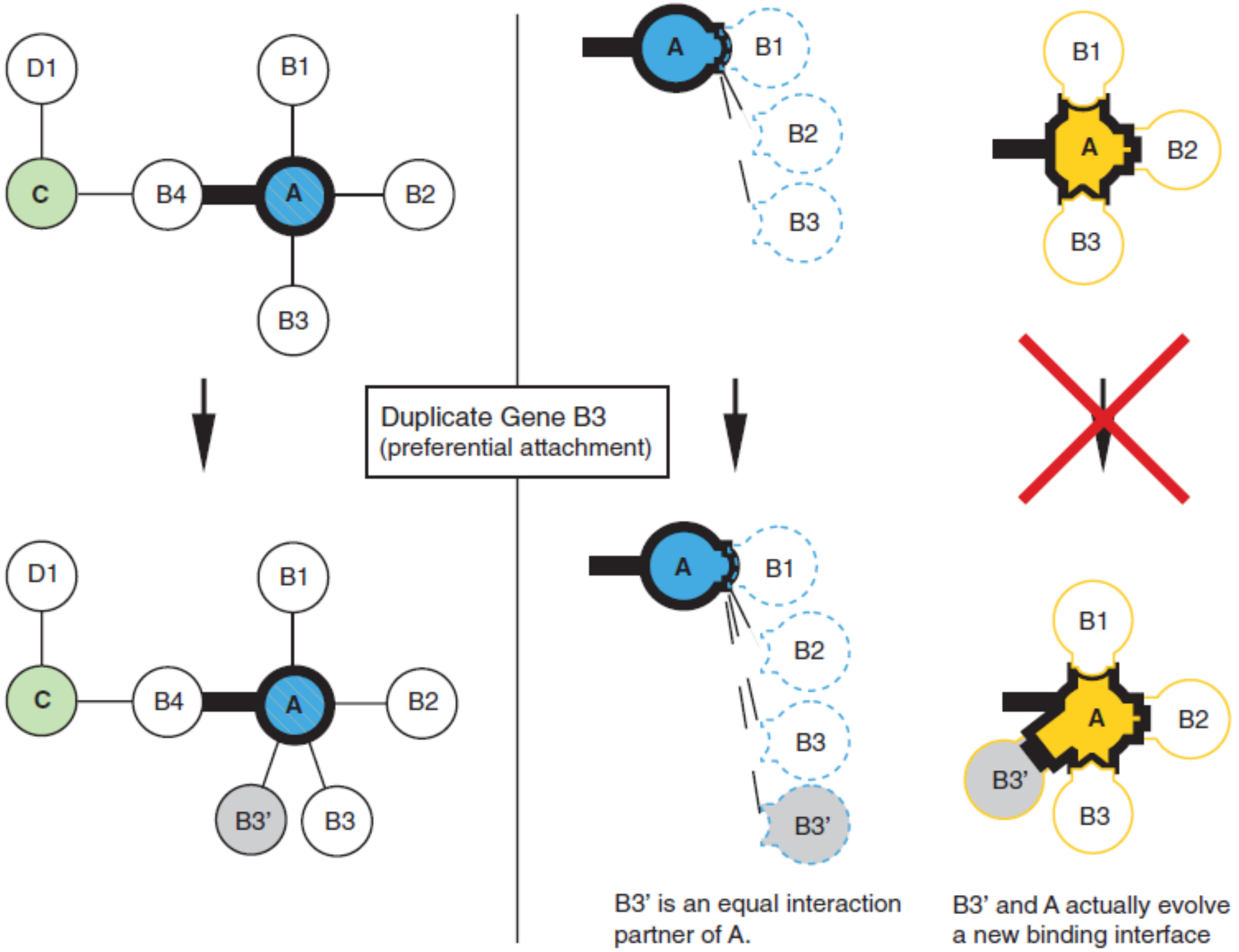
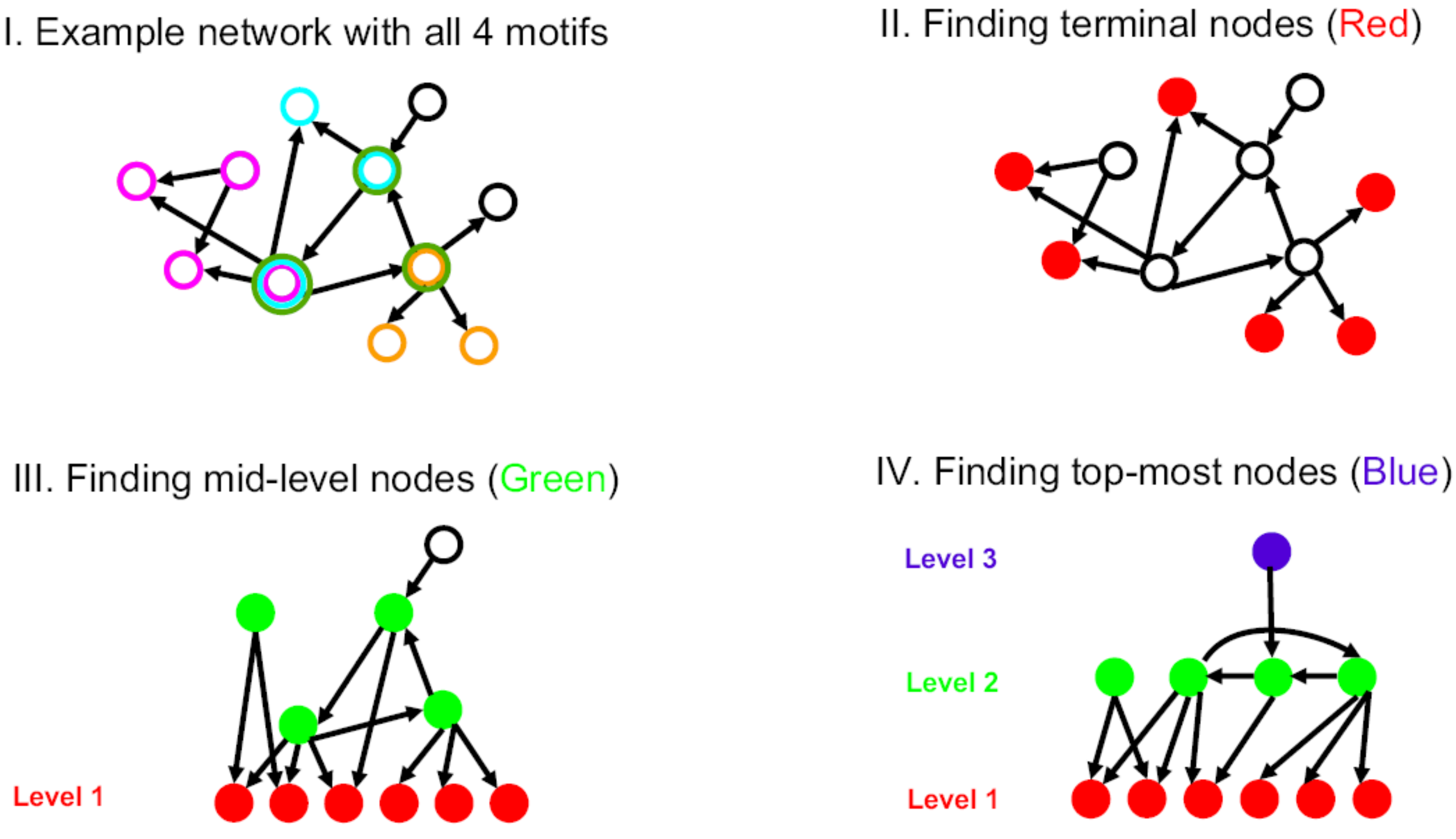

D1
B1
C
B4
A
B2
B3
Duplicate Gene B3
(preferential attachment)
D1
B1
C
B4
A
B2
B3'
B3
A
B1
B2
B3
A
B1
B2
B3
B3'
B3' is an equal interaction partner of A.
A
B1
B2
B3
A
B1
B2
B3'
B3
B3' and A actually evolve a new binding interface

Fig. 3. The concept of network evolution by gene duplication. A given protein may acquire a new interaction by duplication of an existing one. Given equal likelihood of any gene to be duplicated, a protein with many partners is more likely to get a new partner than one with few—hence, there is effective preferential attachment. For singlish-interface hubs, this mechanism is straightforward. However, for multi-interface hubs, it would then require coevolution of the hub and the duplicated gene to form a new interface.

11

- Genomic analysis of the hierarchical structure of regulatory networks.
- H Yu, M Gerstein (2006) Proc Natl Acad Sci U S A 103: 14724-31
- 5 figures

Determination of "Level" in Regulatory Network Hierarchy with Breadth-first Search



I. Example network with all 4 motifs

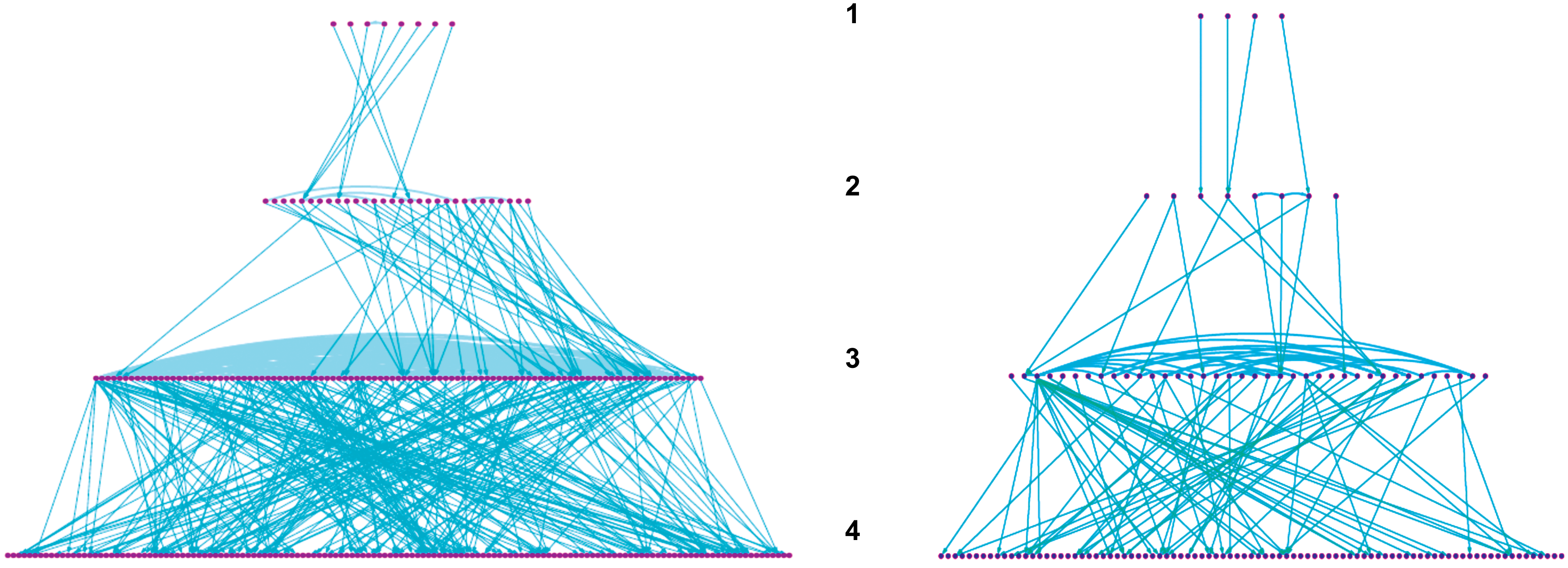
II. Finding terminal nodes (Red)

III. Finding mid-level nodes (Green)

IV. Finding top-most nodes (Blue)

[Yu et al., PNAS (2006)]

Regulatory Networks have similar hierarchical structures

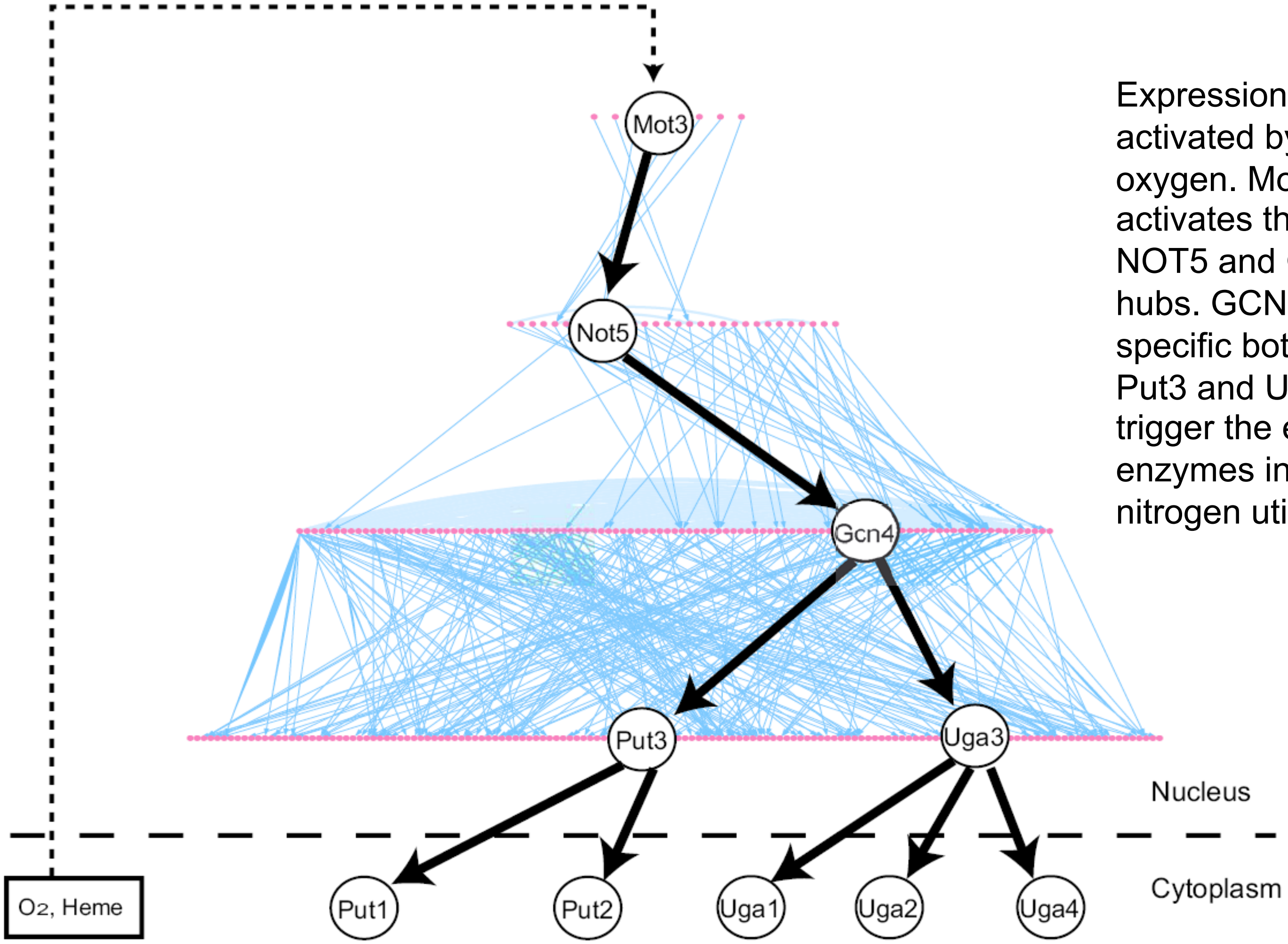


S. cerevisiae

E. coli

[Yu *et al.*, *Proc Natl Acad Sci U S A* (2006)]

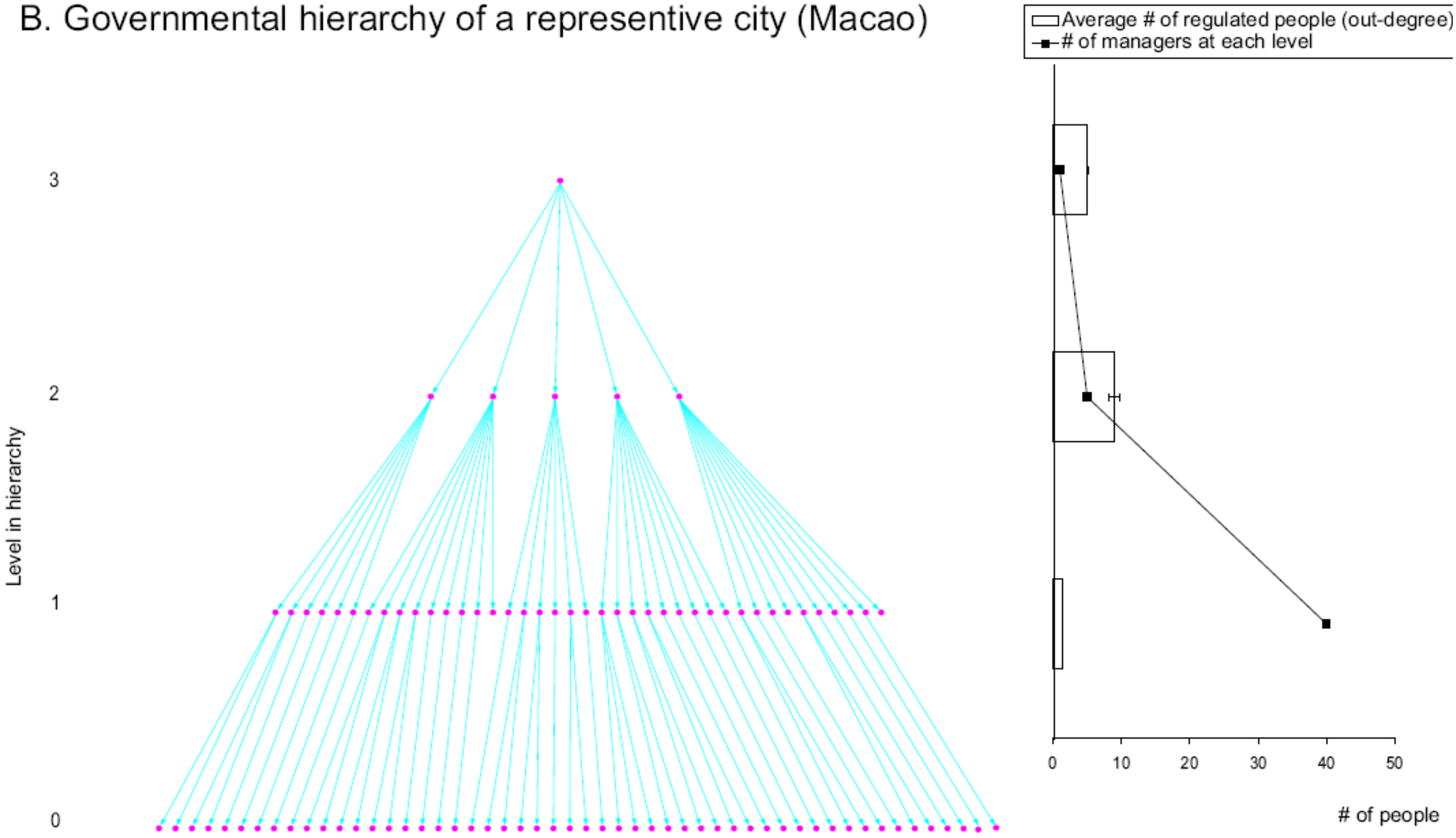
Example of Path Through Regulatory Network



Expression of MOT3 is activated by heme and oxygen. Mot3 in turn activates the expression of NOT5 and GCN4, mid-level hubs. GCN4 activates two specific bottom-level TFs, Put3 and Uga3, which trigger the expression of enzymes in proline and nitrogen utilization.

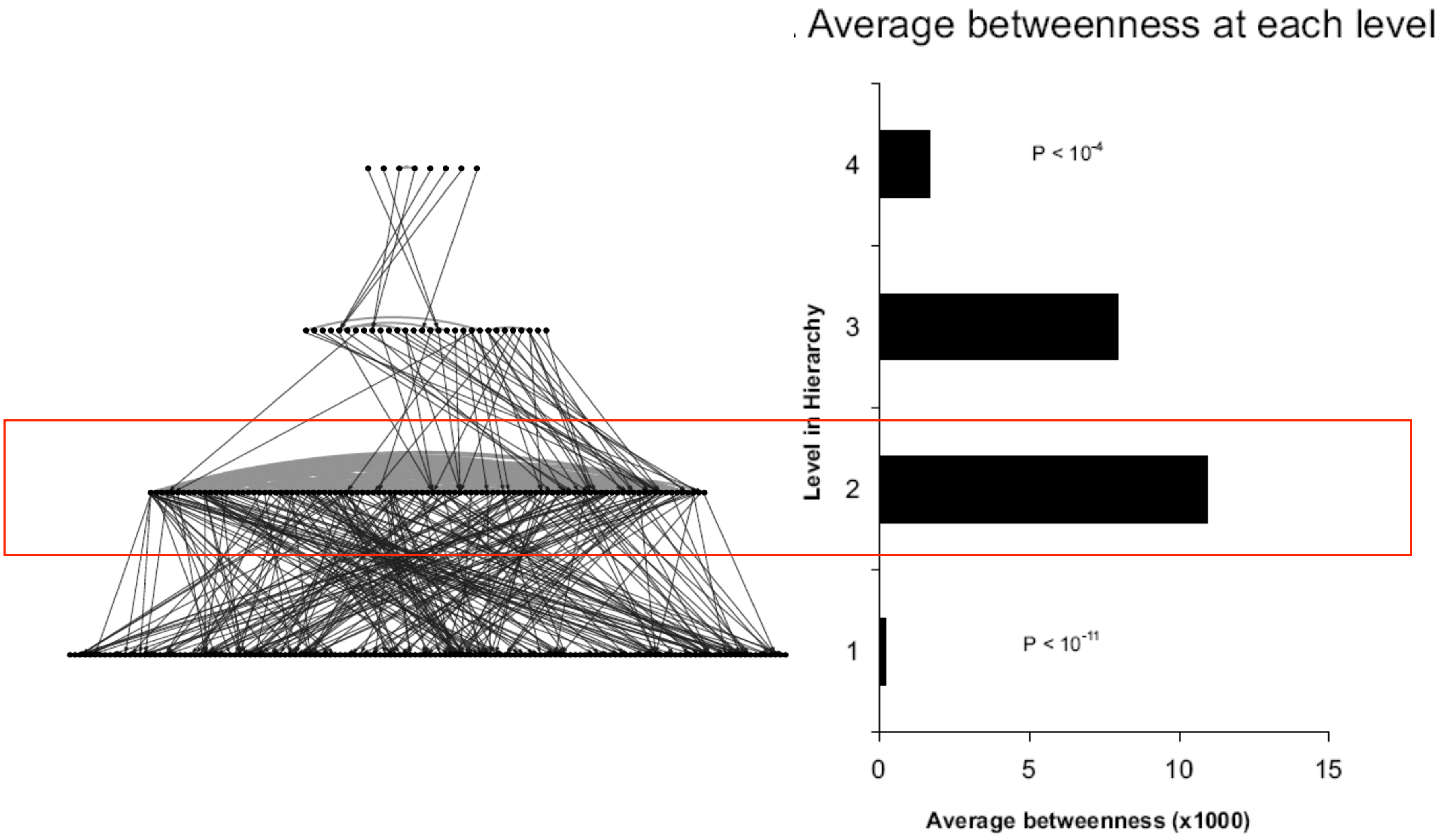
[Yu et al., PNAS (2006)]

Yeast Network Similar in Structure to Government Hierarchy with Respect to Middle-managers



B. Governmental hierarchy of a representive city (Macao)

Characteristics of Regulatory Hierarchy: Middle Managers are Information Flow Bottlenecks



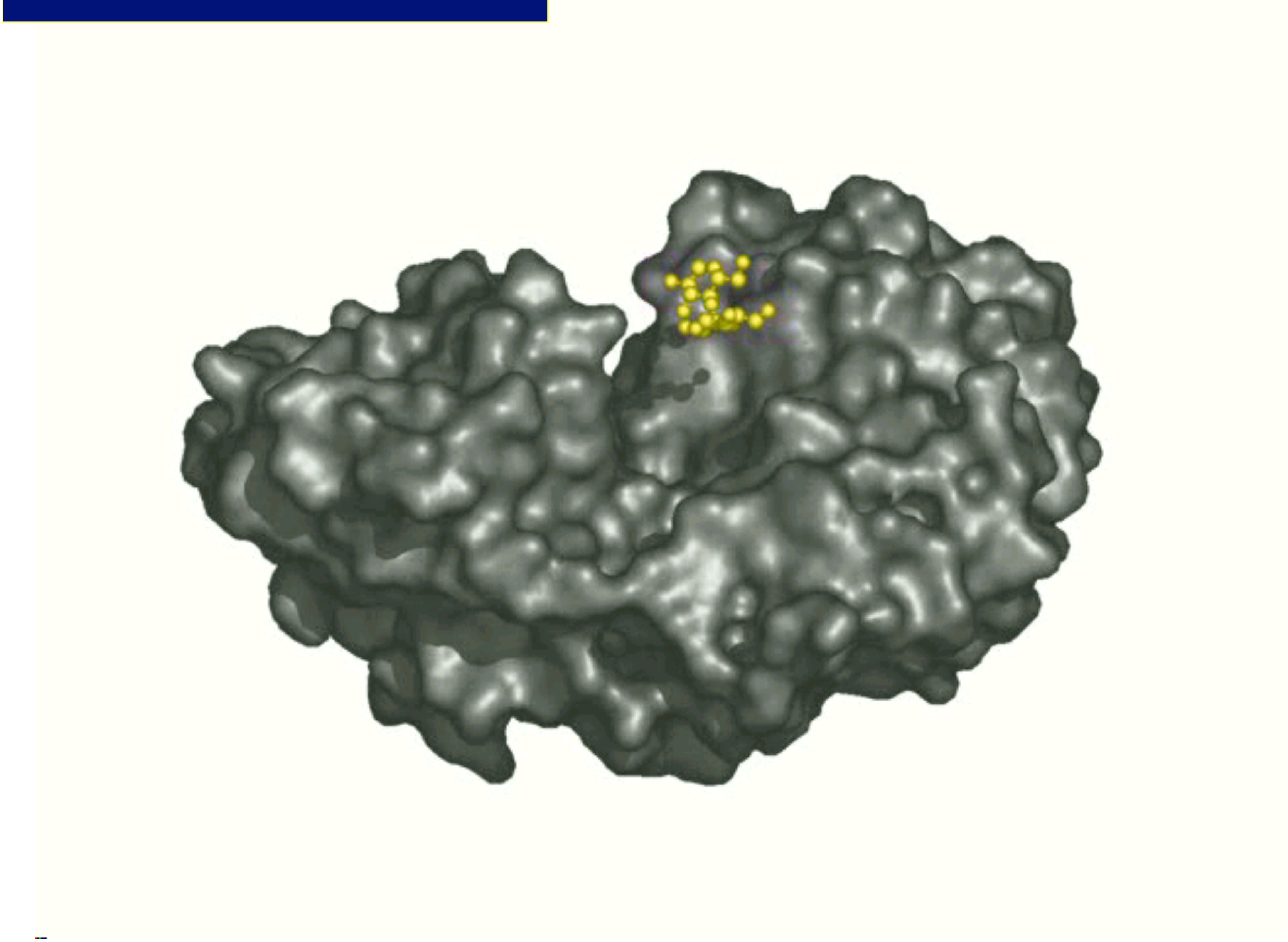
[Yu et al., PNAS (2006)]

12

- The Database of Macromolecular Motions: new features added at the decade mark.
- S Flores, N Echols, D Milburn, B Hespenheide, K Keating, J Lu, S Wells, EZ Yu, M Thorpe, M Gerstein (2006) Nucleic Acids Res 34: D296-301.
- 5 figures

Example "Morph": MBP

- 2 Known Crystal Structures
  (endpoints, not necessarily same seq.)
- Std. Geometric Stats. (from structure comparison)
- Pathway Interpolation

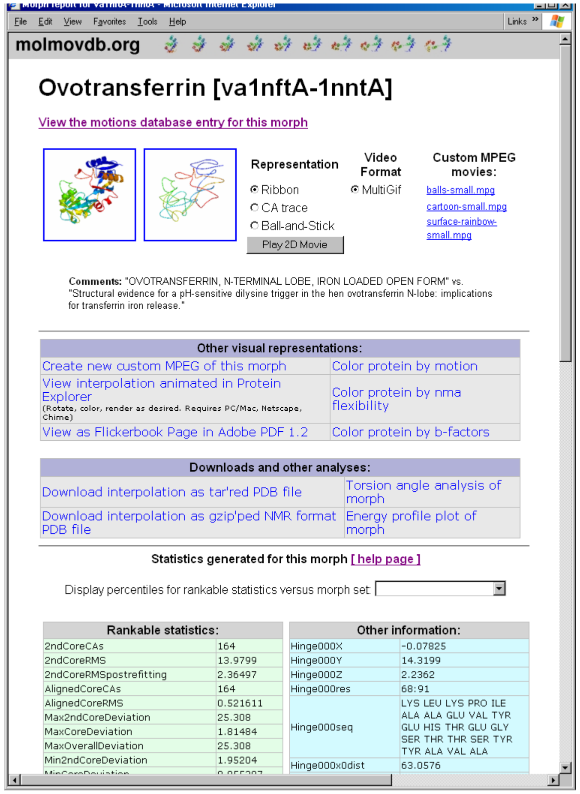


[Flores et al. (2006) *NAR* 34:D296.]

Motions collecting together and annotating Individual morphs into logical units

~19K morphs
(instances of conformational variability)
(384 canonical ones)

~200 classified motions



Morph report for va1nftA-1nntA - Microsoft Internet Explorer

molmovdb.org

Ovotransferrin [va1nftA-1nntA]

View the motions database entry for this morph

Representation	Video Format	Custom MPEG movies:
● Ribbon ○ CA trace ○ Ball-and-Stick Play 2D Movie	● MultiGif	balls-small.mpg cartoon-small.mpg surface-rainbow-small.mpg

Comments: "OVOTRANSFERRIN, N-TERMINAL LOBE, IRON LOADED OPEN FORM" vs. "Structural evidence for a pH-sensitive dilysine trigger in the hen ovotransferrin N-lobe: implications for transferrin iron release."

Other visual representations:	
Create new custom MPEG of this morph	Color protein by motion
View interpolation animated in Protein Explorer (Rotate, color, render as desired. Requires PC/Mac, Netscape, Chime)	Color protein by nma flexibility
View as Flickerbook Page in Adobe PDF 1.2	Color protein by b-factors

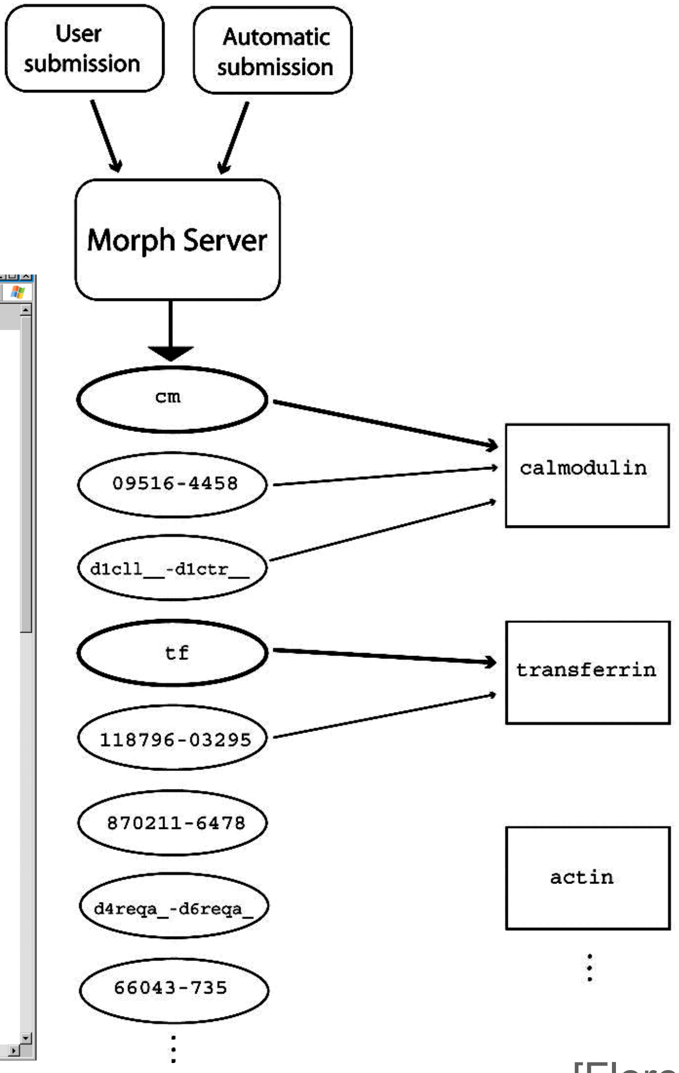
Downloads and other analyses:	
Download interpolation as tar'red PDB file	Torsion angle analysis of morph
Download interpolation as gzip'ped NMR format PDB file	Energy profile plot of morph

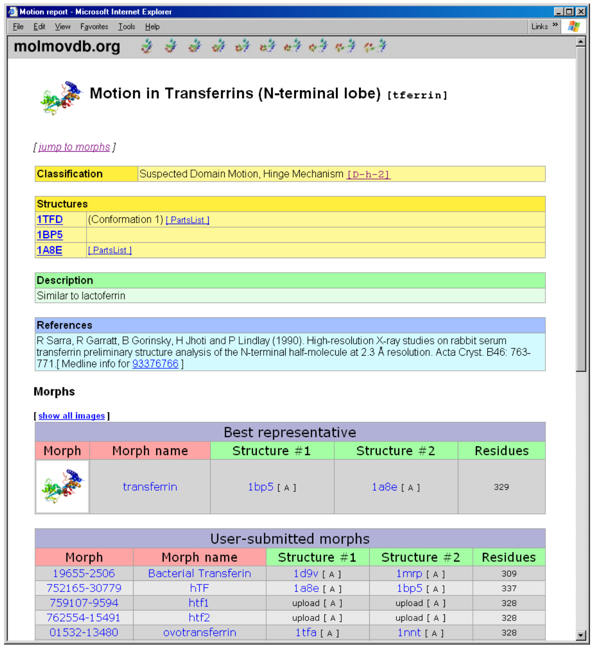
Statistics generated for this morph [help page]

Display percentiles for rankable statistics versus morph set:

Rankable statistics:	
2ndCoreCAs	164
2ndCoreRMS	13.9799
2ndCoreRMSpostrefitting	2.36497
AlignedCoreCAs	164
AlignedCoreRMS	0.521611
Max2ndCoreDeviation	25.308
MaxCoreDeviation	1.81484
MaxOverallDeviation	25.308
Min2ndCoreDeviation	1.95204

Other information:	
Hinge000X	-0.07825
Hinge000Y	14.3199
Hinge000Z	2.2362
Hinge000res	68:91
Hinge000seq	LYS LEU LYS PRO ILE ALA ALA GLU VAL TYR GLU HIS THR GLU GLY SER THR THR SER TYR TYR ALA VAL ALA
Hinge000x0dist	63.0576





Motion report - Microsoft Internet Explorer

molmovdb.org

Motion in Transferrins (N-terminal lobe) [tferrin]

[jump to morphs]

Classification	Suspected Domain Motion, Hinge Mechanism [D-h-2]

Structures	
1TFD	(Conformation 1) [PartsList]
1BP5	
1A8E	[PartsList]

Description

Similar to lactoferrin

References

R Sarra, R Garratt, B Gorinsky, H Jhoti and P Lindlay (1990). High-resolution X-ray studies on rabbit serum transferrin preliminary structure analysis of the N-terminal half-molecule at 2.3 Å resolution. Acta Cryst. B46: 763-771.[Medline info for 93376766]

Morphs

[show all images]

Best representative

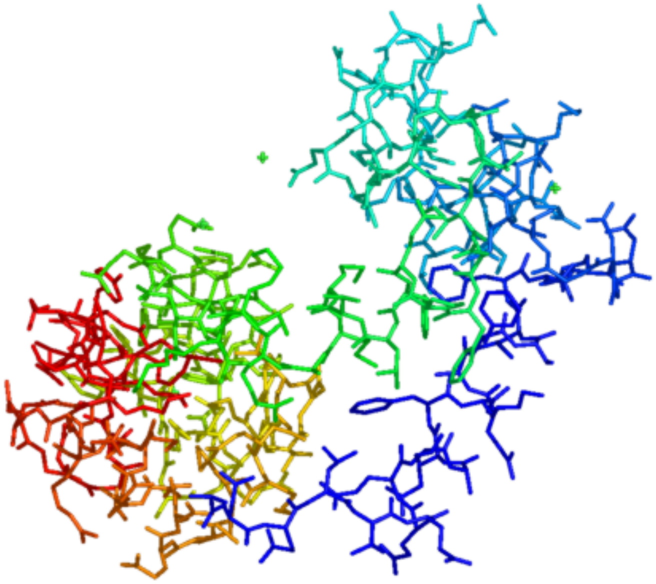
Morph	Morph name	Structure #1	Structure #2	Residues
	transferrin	1bp5 [A]	1a8e [A]	329

User-submitted morphs

Morph	Morph name	Structure #1	Structure #2	Residues
19655-2506	Bacterial Transferin	1d9v [A]	1mrp [A]	309
752165-30779	hTF	1a8e [A]	1bp5 [A]	337
759107-9594	htf1	upload [A]	upload [A]	328
762554-15491	htf2	upload [A]	upload [A]	328
01532-13480	ovotransferrin	1tfa [A]	1nnt [A]	328

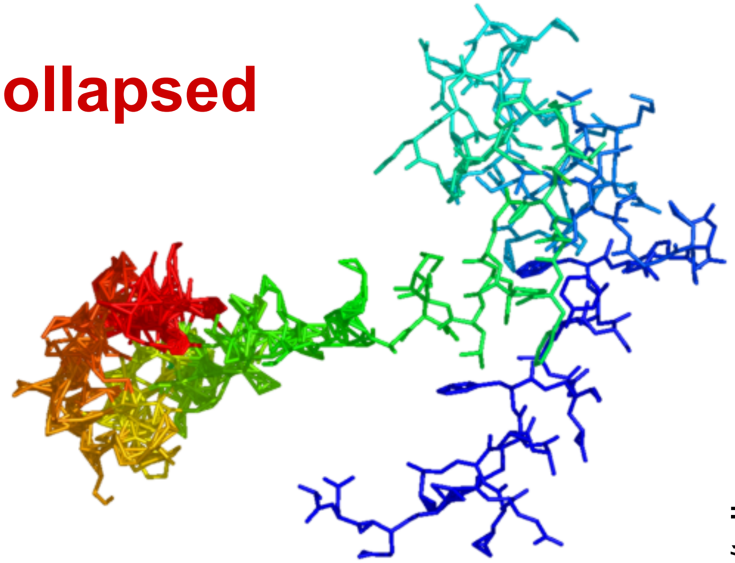
[Flores et al. (2006) *NAR* 34:D296.]

Adiabatic Mapping vs Linear Interpolation Strategies Compared with Calmodulin



Frame 4 (adiabatic)

Collapsed



Frame 4 (linear)

Transferrin hinge involves absence of steric constraints (continuously maintained interfaces), esp. at hinge