Today in HST.480/6.092

Gil Alterovitz

Harvard-MIT
Division of Health
Science & Technology

Announcements

- Homework 2/3 due this Fri 5p
- Projects: In progress
- Today
  - Intro to Proteomics, Mass spec, scale-free networks
- Thurs
  - Intro to Proteomics Part II

Robotic Automation

Visit to new Novartis biomedical research center (built 2004)- near Random Hall. Email Gil for details.

Organization: Levels of Abstraction

- Part I: Sequence
- Part II: Expression
- Part III: Proteomics
- Part IV: Systems/Misc.

Proteomics: A Definition

- “The study of entire protein systems (proteomes): what are the component proteins, how they interact with each other, what kinds of metabolic networks or signaling networks they form”- Dr. Vihinen

Paradigm Shifts in Bioinformatics

- **Sequencing** (1980's to early 1990's)
  - DNA/RNA/Protein Sequence Analysis/sequence storage
- **3-D Protein Structure Prediction** (Mid-1980's-late 1990's)
  - Databases of Protein structures
- **DNA/RNA Microarray Expression Experiments** (Mid-1990's to 2000's)
  - Databases of expression data
- **Protein interaction experiments** (Early 2000's to Present)
  - Databases with pairwise interactions
- **Mass Spec proteomic pattern experiments** (Early 2000's to Present)
  - Databases with mass spec, protein identifications, proteomic patterns
- Integration of multiple modalities (Ongoing)

Networks in Bioinformatics/Proteomics

Image removed due to copyright considerations

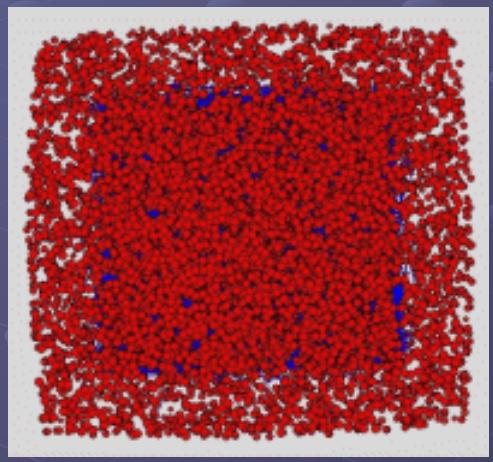


Image removed due to copyright considerations

Scale-free networks

Visualization

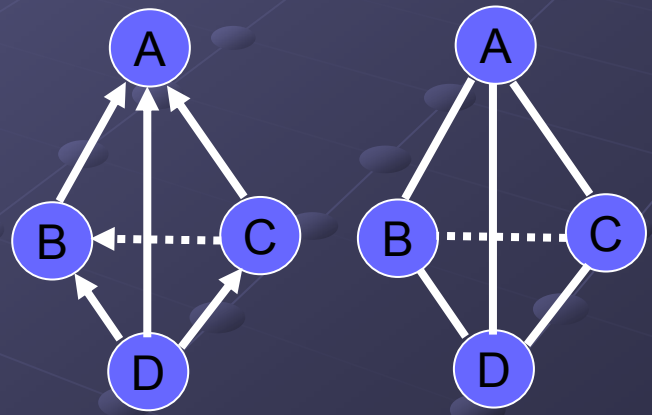
Network Analysis

Gil Alterovitz

HST.480/6.092

Representation

- Represented by a Graph G
  - G=(V,E)
  - V is a set of vertices and E is a set of edges between the vertices, namely:
    E = {(u,v) | u, v □ V}.
  - Node=Vertex
  - Arc=Edge
  - Directed vs. Undirected- no directionality (assume bidirectional)
  - Cyclic vs. Acyclic- no path exists from any vertex to itself
  - Direct Acyclic Graph = Bayesian Network



Networks

- Communication Networks
  - Nodes are routers/phones
  - Edges are phone lines

Image removed due to copyright considerations

Networks

Biological Networks

- Protein Interaction Networks
  - Nodes are yeast proteins
  - Edges are protein-protein interactions
- Gene regulation network
- Metabolism
  - Biochemical reactions

Image removed due to copyright considerations

Yeast Protein Interaction Network

Types

Type	Detail
Correlation graph (undirected graph)	The information about the positive / negative correlation between genes is described. Two related genes are connected with an undirected arc.
Cause-effect graph (direct graph)	Describing the relationship caused by a gene acting upon another gene. Causality is represented by a directed arc, whose direction shows the cause and effect.
Weighted graph (in the broad sense)	Some qualitative meaning is attached to a graph within its arcs. E.g., S-system or a Bayesian network.

Genome Informatics 14: 104–113 (2003)

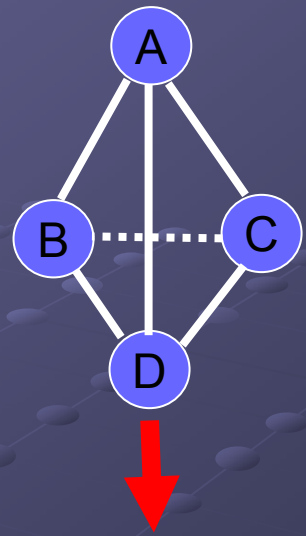
Adjacency Matrix

- Vertices: A,B,C,D
- Edges: A↔B, B↔C, C↔D, D↔A
- Represent as n x n matrix called:

  $\overline{\overline{A}}$ where n=Number of Vertices

- Place a 1 (or other weight for each edge) in matrix element:

  $\overline{\overline{A}}_{ij}$ where edge goes from i→j



$\overline{\overline{A}}$ To:

From:

	A	B	C	D
A	0	1	1	1
B	1	0	1	1
C	1	1	0	1
D	1	1	1	0

How many Edges?

- n^2 elements in matrix.
- Assume: no edges between self (i.e. no edge from A to A, etc.)
  - n^2-n elements
- However, since edges are bidirectional, we are double counting each edge.
  - Use only one of triangles:

Number of edges for k nodes =

$$\frac{n^2 - n}{2}$$

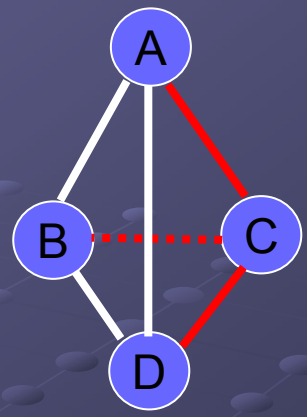
$\overline{\overline{A}}$ To:

From:

	A	B	C	D
A	0	1	1	1
B	1	0	1	1
C	1	1	0	1
D	1	1	1	0

Properties: Degree

- Neighbors
  - Vertices that have an edge between them.
- Degree
  - Number of edges linking a given vertex to its neighbors.
  - E.g. Degree is 3 for vertex C.



$\overline{\overline{A}}$

To:

From:

	A	B	C	D
A	0	1	1	1
B	1	0	1	1
C	1	1	0	1
D	1	1	1	0

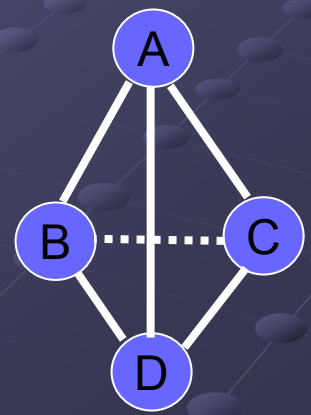
Properties: Clustering Coefficient

- Cluster- reflects tendency for neighbors of given vertex to be connected.
- Cluster Coefficient= Number of edges between neighbors of vertex i divided by total possible edges between k_i *neighbors* of vertex i.
  - If i=A, then k=3 and:

$$\frac{2*3}{3*(3-1)}=1$$

$$\frac{k_i(k_i-1)}{2}$$

$$C_i = \frac{2E_i}{k_i(k_i-1)}$$



- Average Cluster Coefficient: tendency of graph to form clusters = mean(Ci) for all vertices i =

$$C = \frac{\sum_{i=1}^{N} C_i}{N}$$

Erdös-Rényi Model (Random Network)

- Growth model
  - Edges to new nodes added from existing nodes with equal probability
- Degree distribution P(k), where k is the degree of node
- Average path length ~ ln N, where N is number of nodes

Poisson distribution

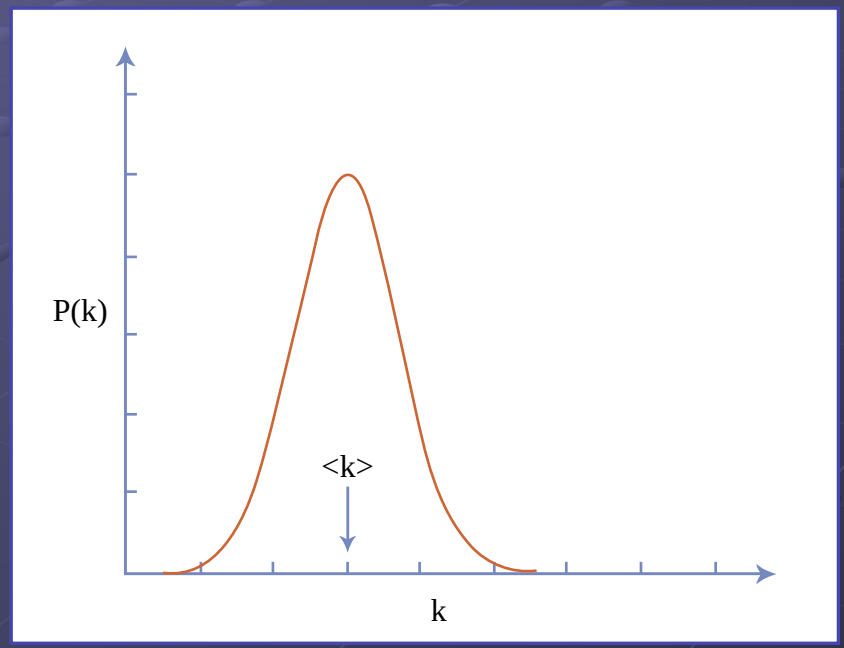


Figure by MIT OCW

R. Albert, A.-L. Barabasi, Statistical mechanics of complex networks, Rev. Mod. Phys., 74, 47, 2002

Scale-free Network

$P(k) \sim k^{-\gamma}$

$\gamma<3$ implies scale free.

$$\Pi(k_i) = \frac{k_i}{\Sigma_j k_j}$$

- Scale-free =
- Growth model.
  - Add a new node with m edges to existing network
  - Probability Π of adding edge from vertex i to a new vertex increases as to vertex i's degree (k_i) increases:

- Average path length ~ ln (ln N), where N is number of nodes. Therefore, more efficient signaling than random network.

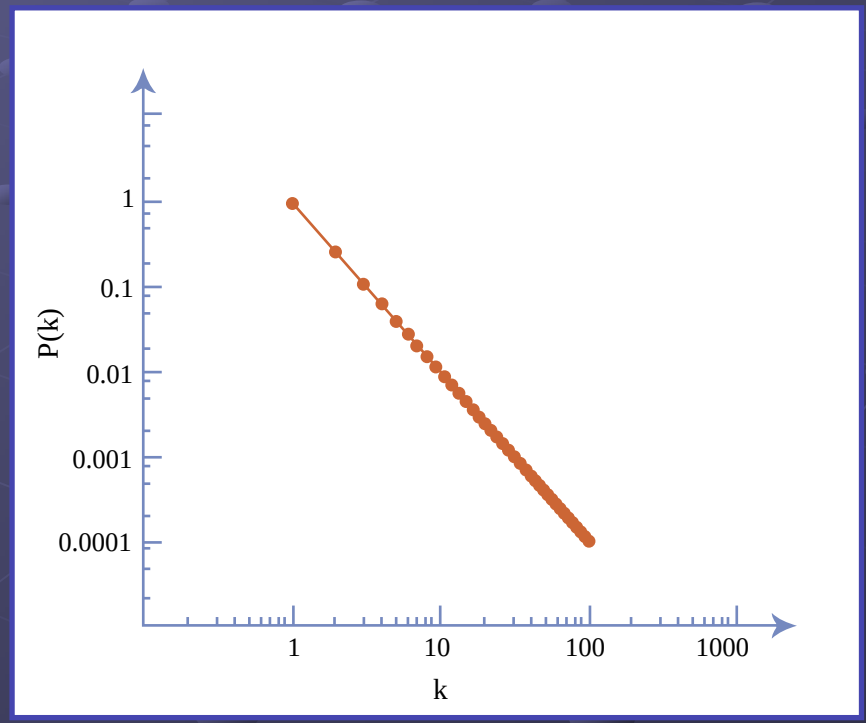


Figure by MIT OCW

Scale-free Network

Power-law distribution

Hubs

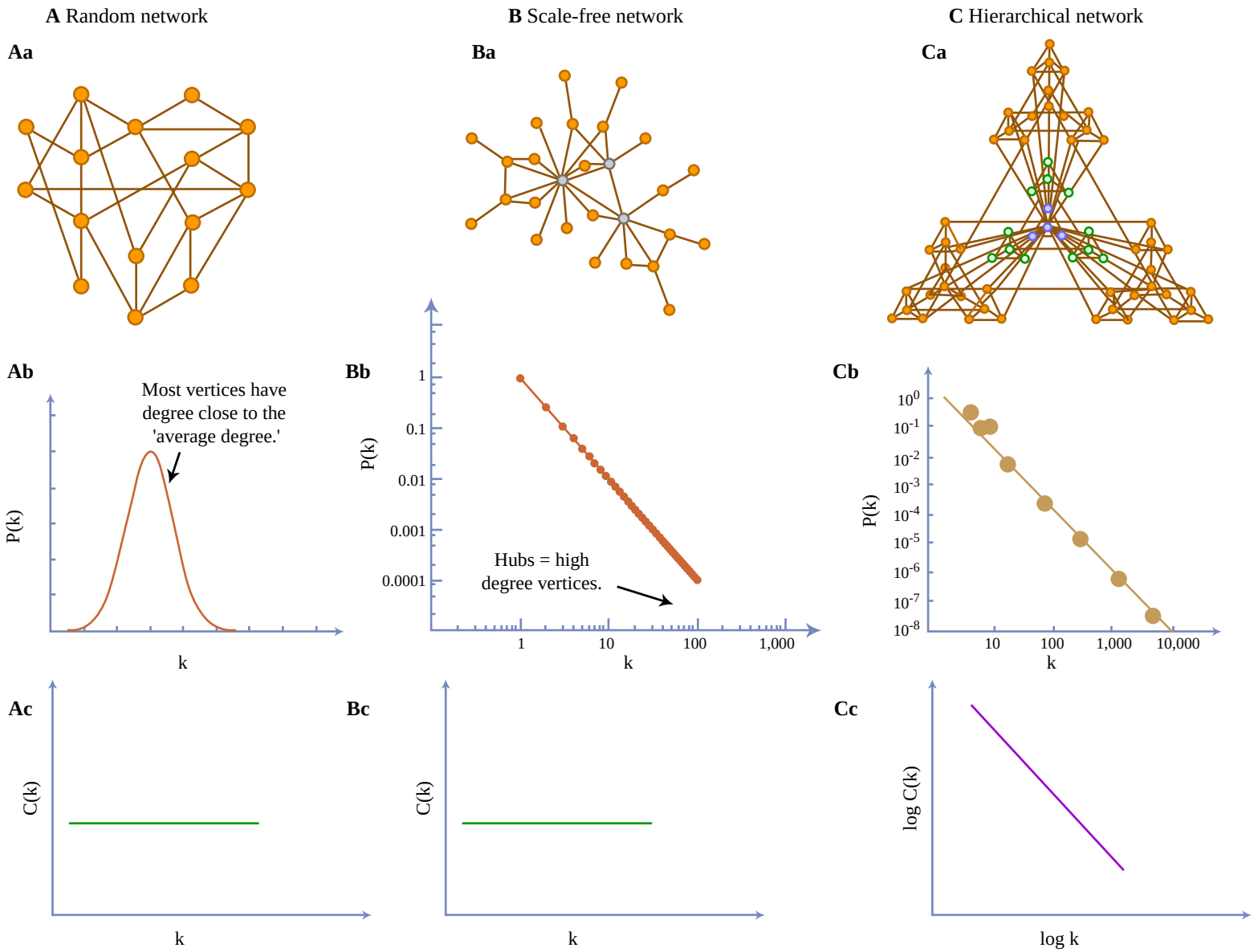
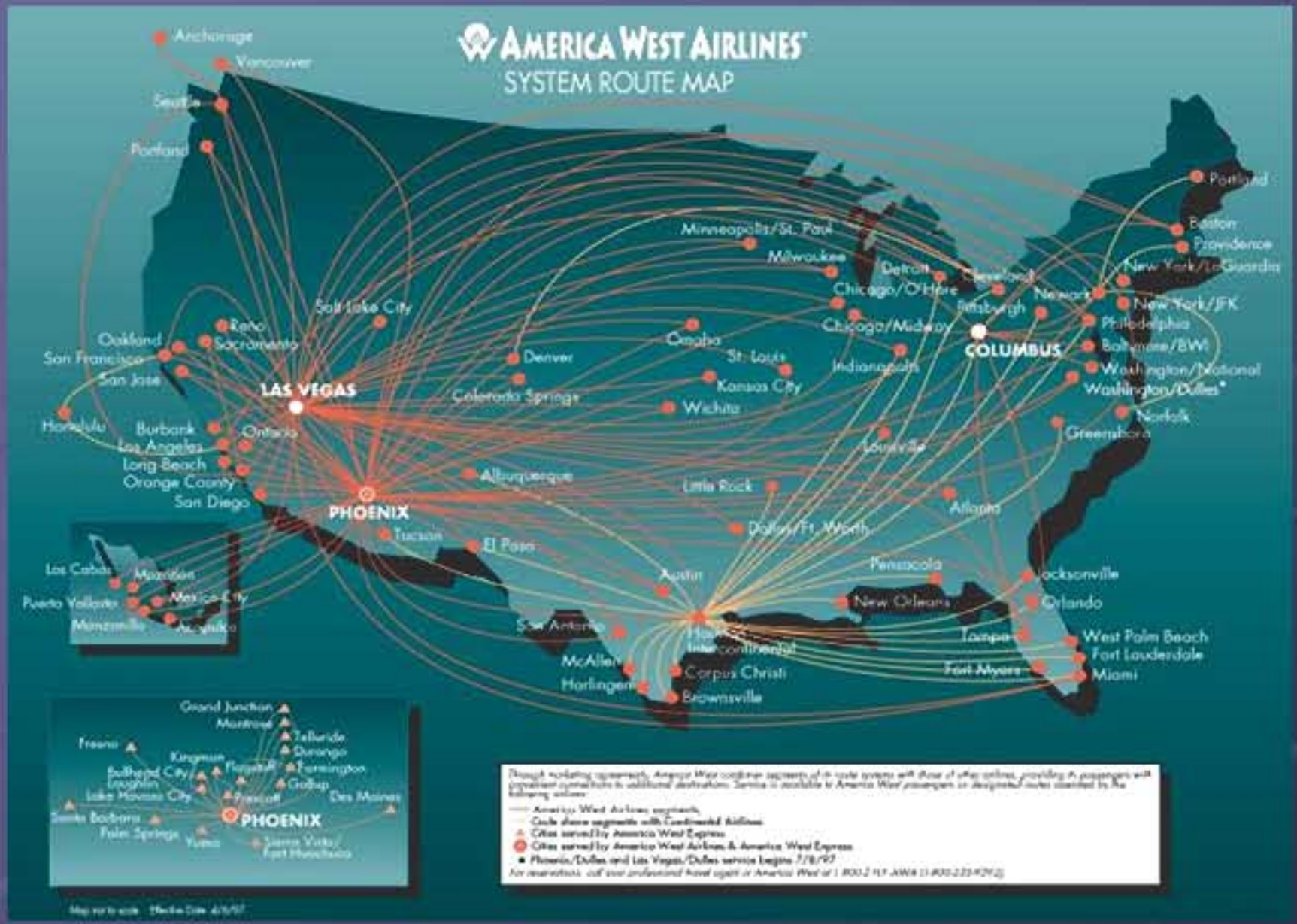


Figure by MIT OCW

Nature. 2000 Jul 27;406(6794):378-82.



Courtesy of America West Airlines. Used with permission.

Robustness Under Failure and Attack

- Measure of network operation: number of vertices in largest subgraph (a path exists any vertex to any other vertex). S is above number normalized by the original size of the graph.
- If failure is random hit:
  - Remove random node
  - Scale-free network is more likely to survive than random network

- If failure is targeted hit:
  - Remove node that causes maximum ‘damage’
  - Scale-free network is more vulnerable than random network

Image removed due to copyright considerations

Random

Image removed due to copyright considerations

Scale-free

Application: Protein-Protein Interactions

- Proteins (Vertices) with high degree (interact with many other proteins directly) are more essential than ones with a low degree.
- Knocking out high degree proteins more likely to result in catastrophic system failure.
  - Drug target applications

Image removed due to copyright considerations

Sample Protein Interaction Network (From Yeast)

Case Study: Lethality and Centrality for Yeast Proteins

- 1,870 proteins (vertices)
  2,240 interactions (edges)
- 93% of proteins are degree≤5
  - 21% are essential to yeast survival
- 0.7% of proteins are degree>15
  - 62% are essential
- Positively correlated: Correlation coefficient between lethality and connectivity is 0.75.

Image removed due to copyright considerations

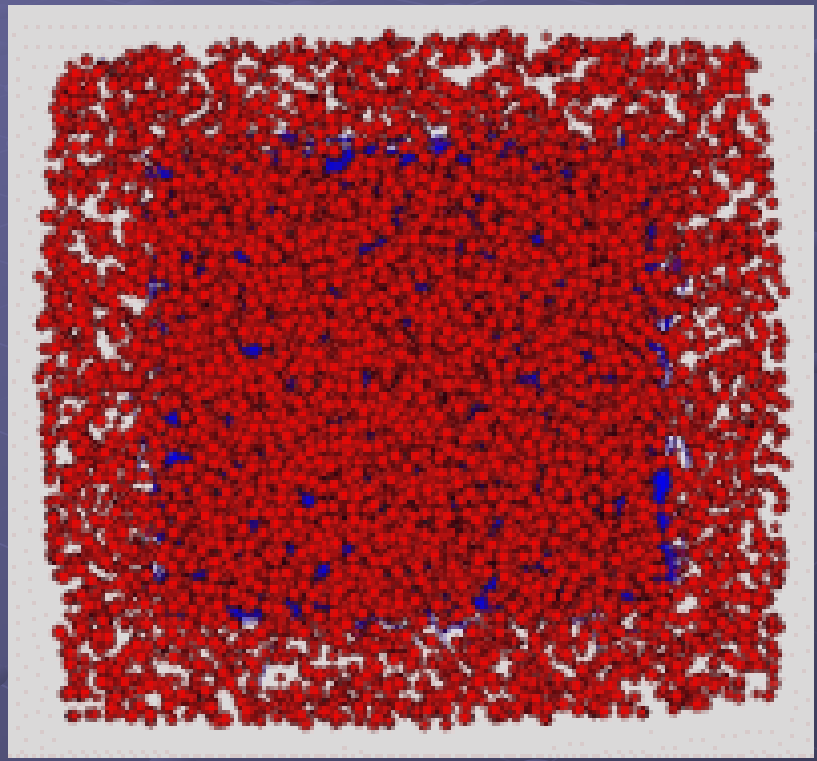
Complete Yeast Protein Interaction Network

Nature. 2001 May 3;411(6833):41-2.

Meta-Database Steps

- Parse XML/flat files of databases
- Convert Different protein identification numbers to NCBI Entrez Protein GI numbers (SeqHound Java API).
- Use SeqHound to find redundant GI's and select best annotated version protein from a group of database entries referring to the same protein sequence (redundant proteins).
- Merge databases (removing duplicates)
- Calculate molecular weight of different cleavage products based on NCBI Entrez annotated features
- Create hash/direct-lookup table for quick access via molecular weight

Visualization of Interactions



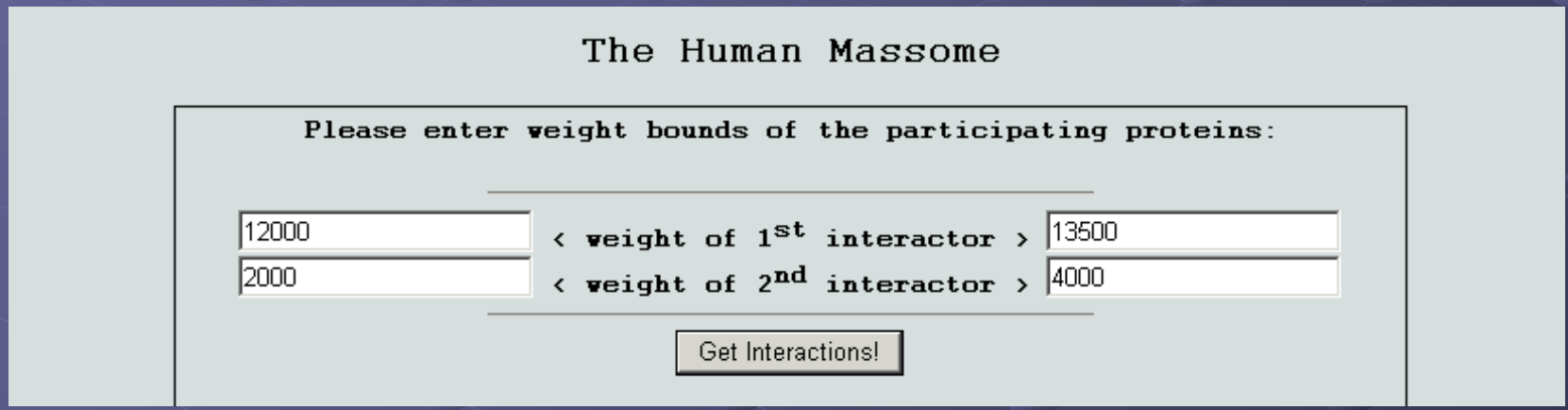
Blue = edges (interactions)
Red = vertices (proteins)

With Dima Patek

Harvard-MIT
Division of Health
Science & Technology

The Human Massome

Example: Found two proteins that bind. What are they?



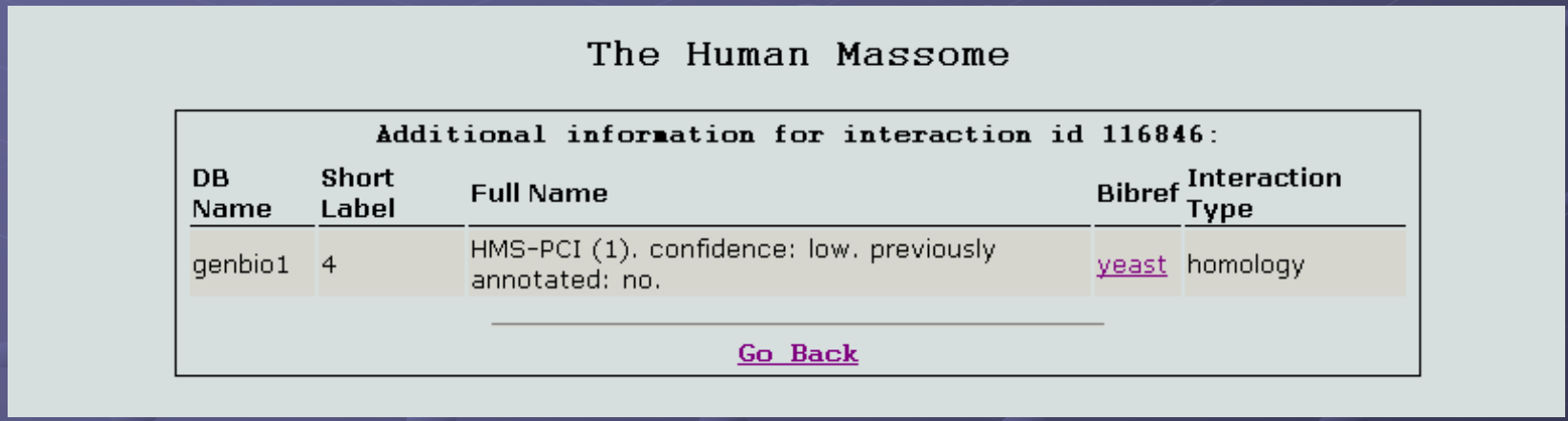
Example

The Human Massome

2 Interactions with participants weighing between (12000 , 13500) and (2000 , 4000):

ID	Name 1	GI 1	Weight 1	Name 2	GI 2	Weight 2
116846	acetyl-Coenzyme A carboxylase alpha isoform 5 [Homo sapiens].	38679980	12717.305	coatomer protein complex, subunit alpha [Homo sapiens].	4758030	2971.580

Example: Source of Interaction



The Human Massome

Additional information for interaction id 116846:

DB Name	Short Label	Full Name	Bibref	Interaction Type
genbio1	4	HMS-PCI (1). confidence: low. previously annotated: no.	yeast	homology

Go Back

Source: High-throughput mass spectrometric protein complex identification

Found: yeast proteins interacted. Found homologous proteins in human. Assume the human proteins interact.

From Interaction Networks to Signaling Pathways

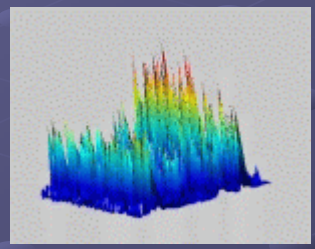
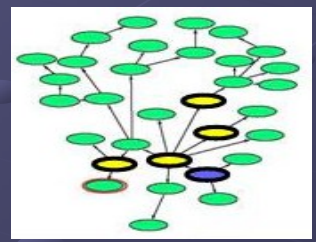
Assume just for this example: We don't know role of Fas-L

Following pathway, we can see "Fas-L involved in JNK Pathway" ->apoptosis

Image removed due to copyright considerations

Image removed due to copyright considerations

Proteomic Profiles Using Surface Enhanced Laser Desorption Ionization Time-of-Flight Mass Spectrometry (SELDI-TOF MS)

Gil Alterovitz

HST.480/6.092

Harvard-MIT
Division of Health
Science & Technology

The Promise of Proteomics…

PROTEOMICS

Searching for the real stuff of life

The discovery that humans have fewer genes than expected has thrust proteins into the research spotlight, says Victoria Griffith

Genetics and Medicine

Recruiting Genes, Proteins For a Revolution in Diagnostics

As companies create medicines for special conditions that require molecular testing. They are helping change the way common diseases are diagnosed

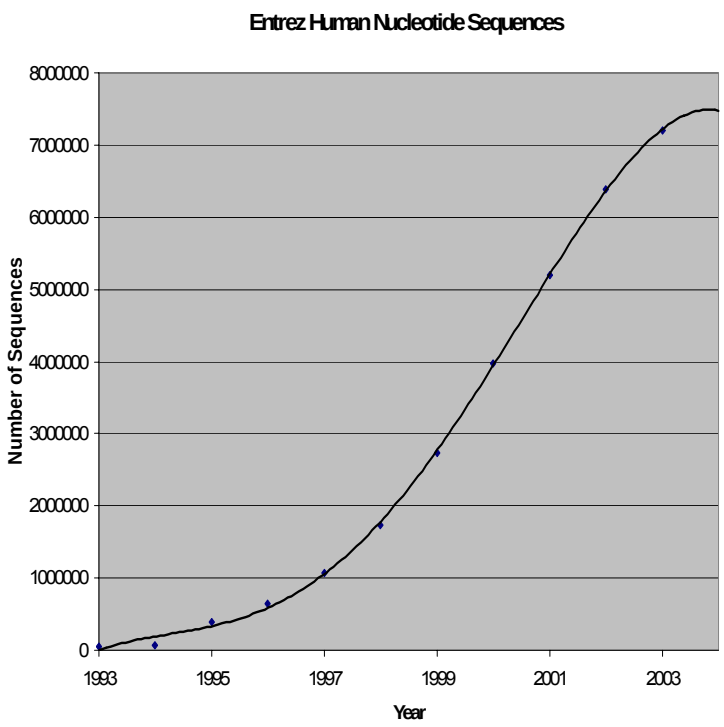
BIOTECH'S NEXT HOLY GRAIL

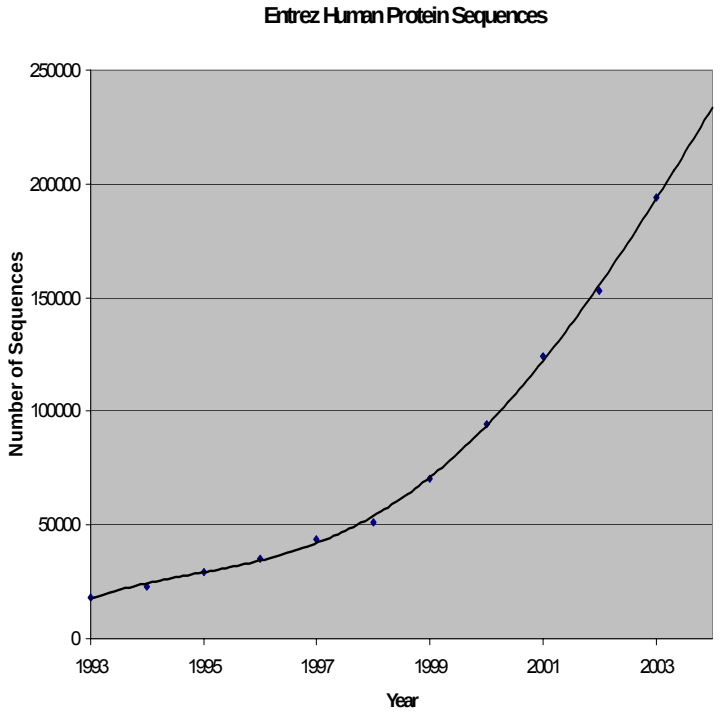
Now, companies are racing to Decipher the human protein set

Protein microarrays and proteomics

Gavin MacBeath

While the number of genetic sequences in Entrez is starting to saturate, the proteins being cataloged in Entrez is still growing exponentially each year*





* Alterovitz, G., Afkhami, E. & Ramoni, M. in *Focus on Robotics and Intelligent Systems Research*, ed. Columbus, F. Nova Science Publishers, Inc., New York, 2005 (In press).

1990's Genomics ➔ 2000's Proteomics

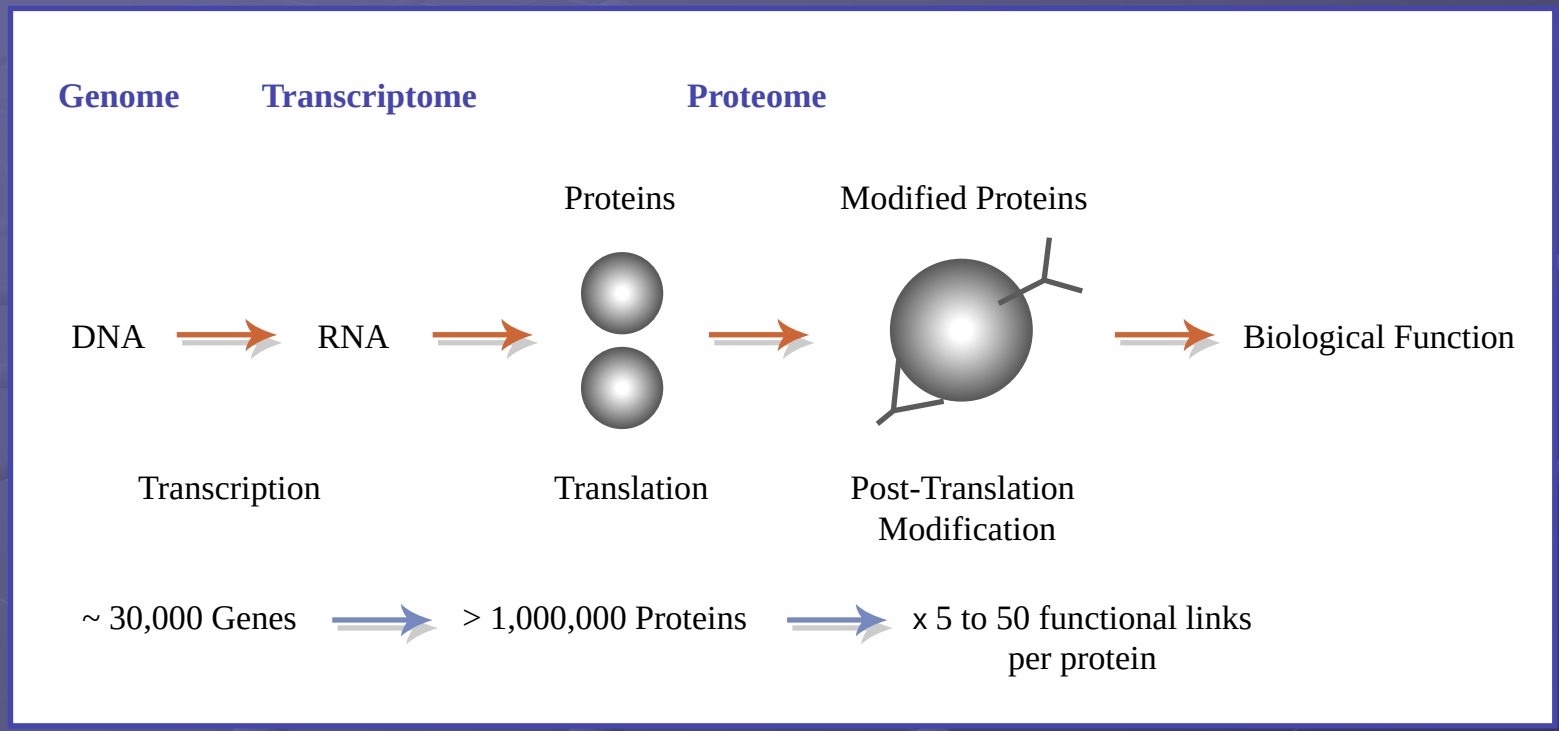


Figure by MIT OCW

Genes do not tell the whole story. We need to look at proteins.

Original Proteomic Cancer Profiling Paper

- Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. "Use of proteomic patterns in serum to identify ovarian cancer." *Lancet*. 2002. Feb 16;359(9306):572-7.

Image removed due to copyright considerations

Early Genomic Cancer Profiling Papers

- DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nat Genet*. 1996 Dec;14(4):457-60.

- Eric S. Lander , "The New Genomics: Global Views of Biology," *Science* 274, 536 (1996)

- Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP. "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nat Med*. 1998 Jul;4(7):844-7

The Promise: Old Proteomics ➔ New Proteomics, Surface Enhanced Laser Desorption and Ionization (SELDI)

- Parallelization
- Automation

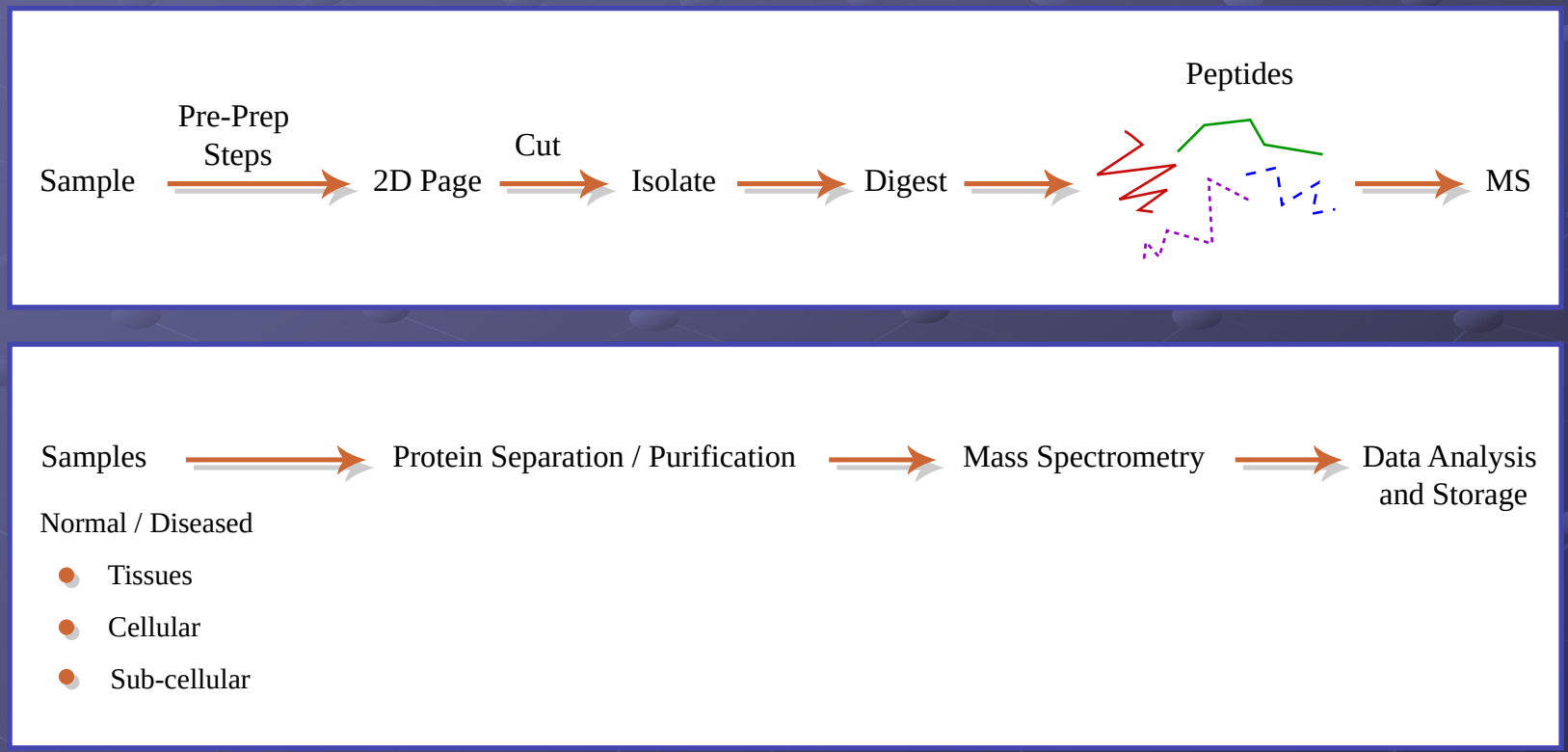
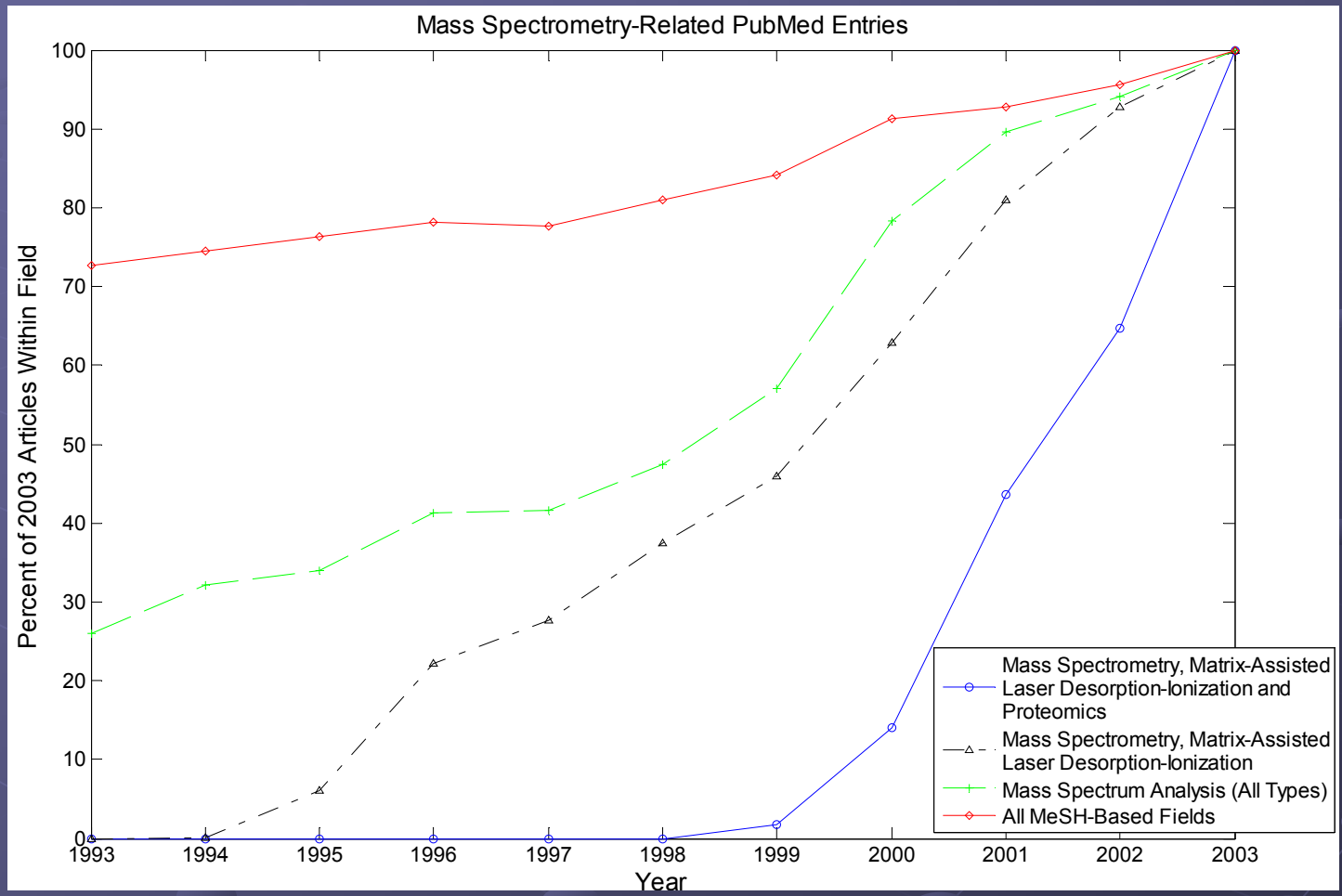


Figure by MIT OCW

Mass spectrometry is growing at a much faster rate in terms of papers compared to the general PubMed database



Alterovitz, G., Afkhami, E. & Ramoni, M. in *Focus on Robotics and Intelligent Systems Research*, ed. Columbus, F. Nova Science Publishers, Inc., New York, 2005 (In press).

New Flexibility with SELDI-TOF

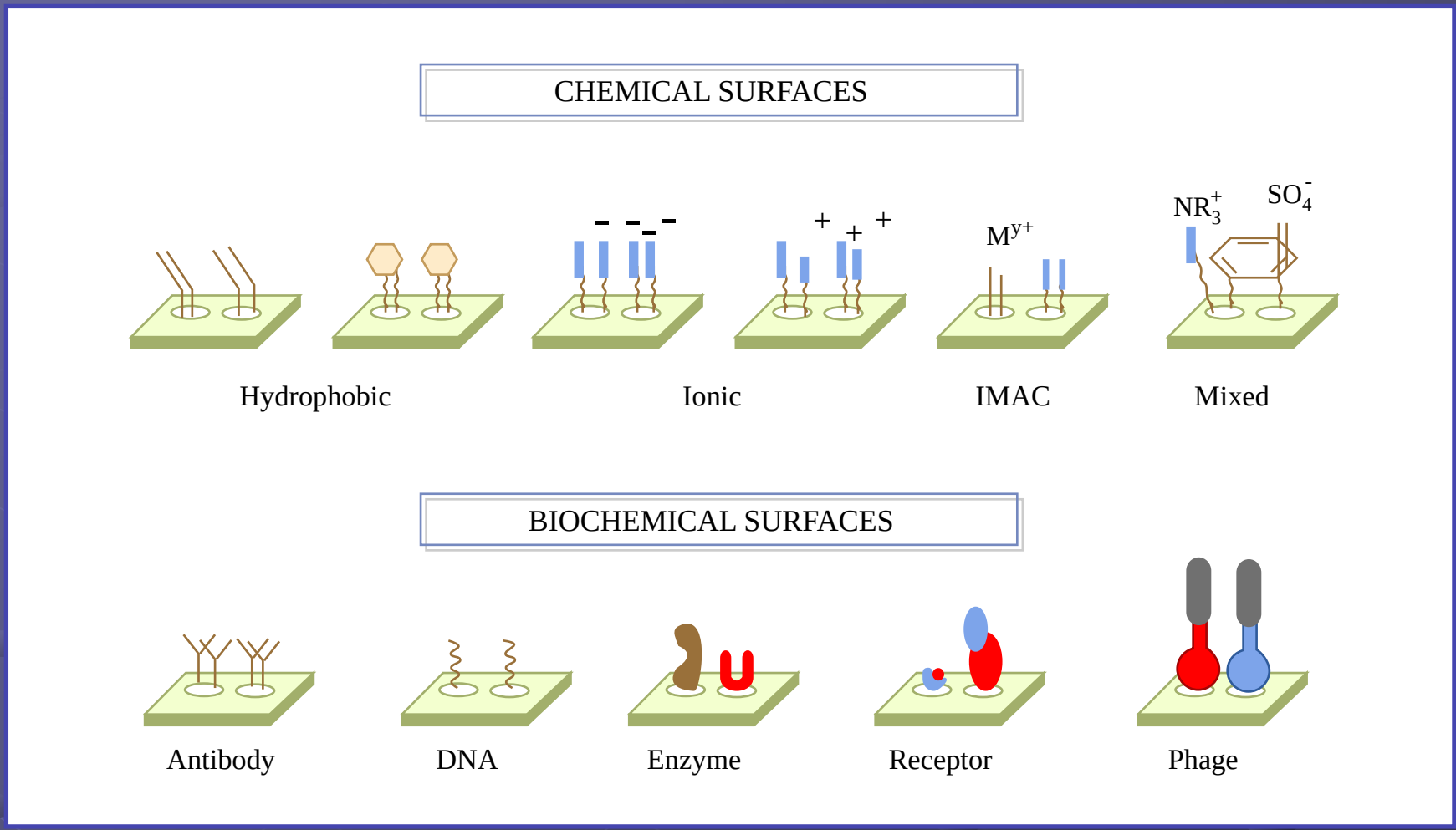


Figure by MIT OCW