CSCI8980: Applied Machine Learning in Computational Biology

Introduction to Bioinformatics

Rui Kuang

Department of Computer Science and Engineering
University of Minnesota
kuang@cs.umn.edu

History of Bioinformatics

Margaret Dayhoff

Thanks to Luce Skrabanek

History of Bioinformatics

1969 UNIX, ARPANET
1971 email
1973 ethernet
1974 internet
1975 Cray1 supercomputer
1977 Apple, Commodore and Tandy sell personal computers
1981 IBM PC
1984 Apple releases the Mac


Paul Berg

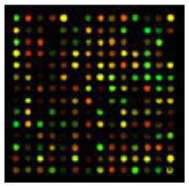
1972 first recombinant DNA molecule
1975 2D electrophoresis
1977 DNA sequencing
1980 multi-D NMR protein structure
1981 RNA secondary structure
1982 φ λ genome sequenced
1984 Epstein-Barr virus sequenced
1985 PCR
1987 Physical map of E. coli

1970 Needleman-Wunsch
1981 Smith-Waterman
1982 GenBank, GCG
1983 Miller-Lipman seq db searching alg
1985 FASTP
1986 SwissProt
1987 Human Genome Project
1988 FASTA, Clustal "bioinformatics" coined

NCBI

Thanks to Luce Skrabanek

History of Bioinformatics



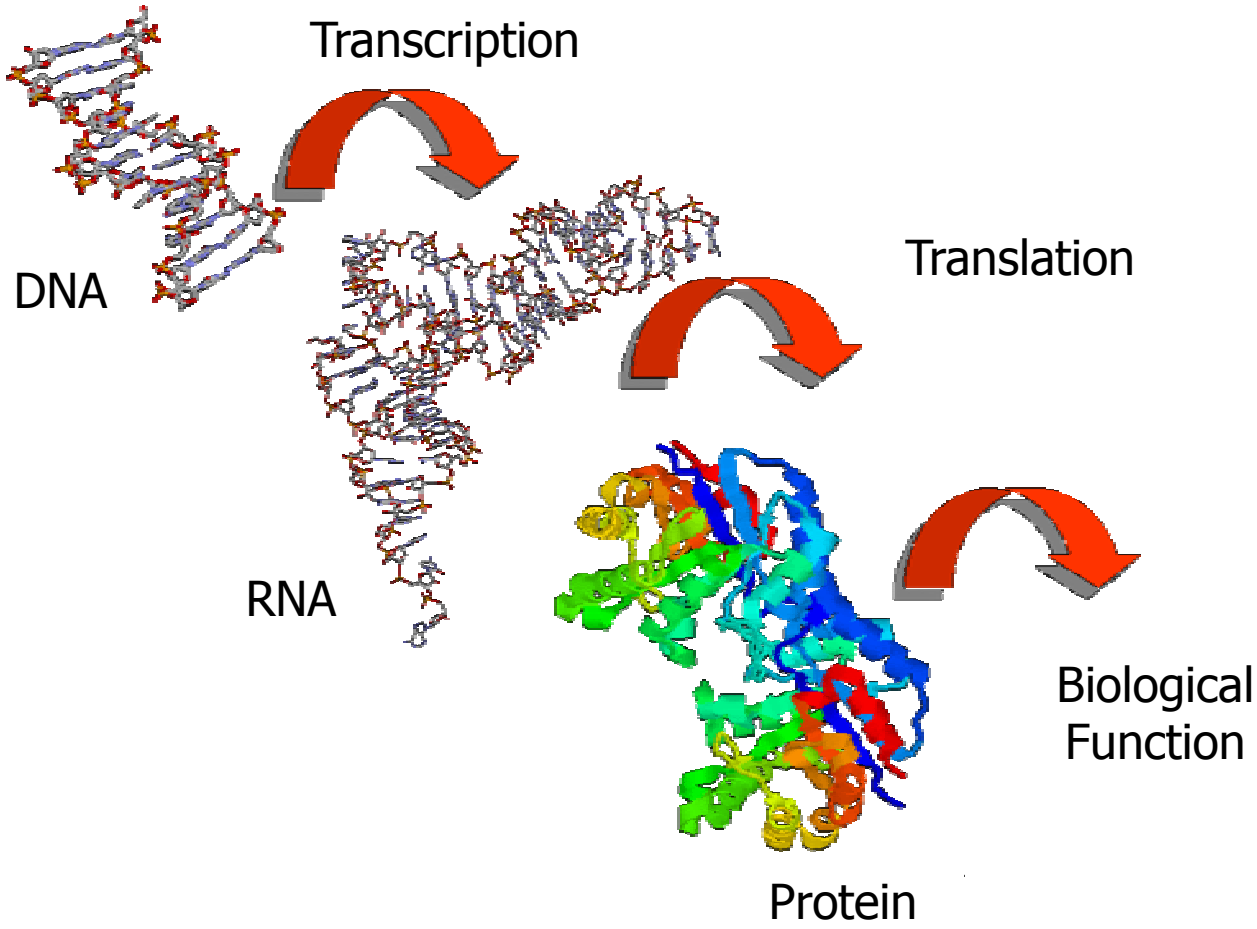
1991 WWW, Linux
1992 HP LaserJet 600x600 dpi printer
1994 Netscape
1995 Java
1997 Deep Blue beats Kasparov, DVDs
2003 XRAID

1991 ESTs
1995 Microarrays, First bacterial genomes
1996 Yeast
1997 E coli
1998 C. elegans
2000 Arabidopsis, Drosophila
2001 Human
2002 Mouse
2004 Rat

1990 BLAST
1991 First time "bioinformatics" appears in scientific literature
1993 ACeDB: first genome database
1995 Microarray analysis, SAGE
1997 PSI-BLAST, Pfam, GenScan
1998 Phred, Consed, GeneMark
1999 MFOLD
2000 GeneOntology, FASTA3
2002 Arachne

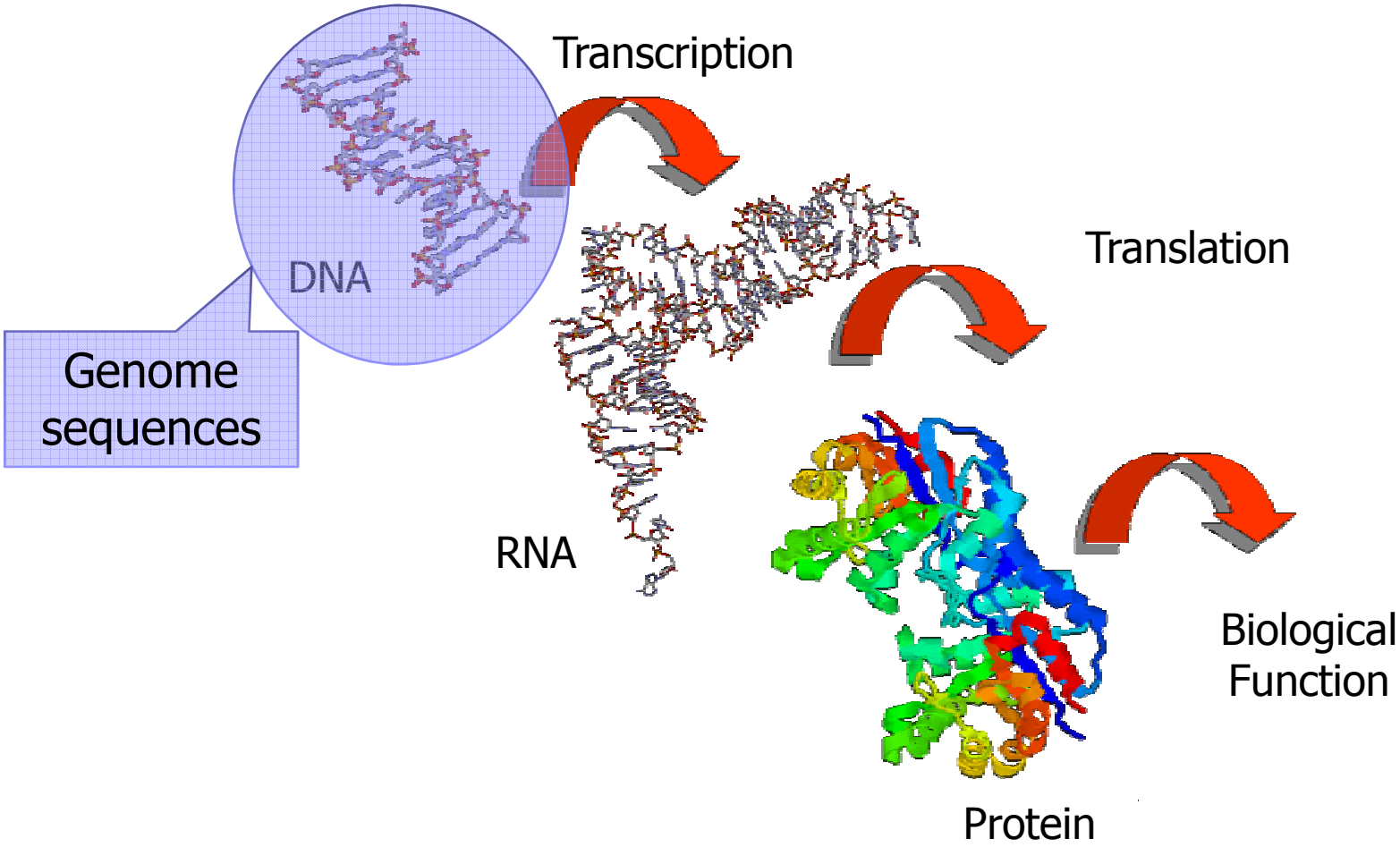
Thanks to Luce Skrabanek

Biological Data

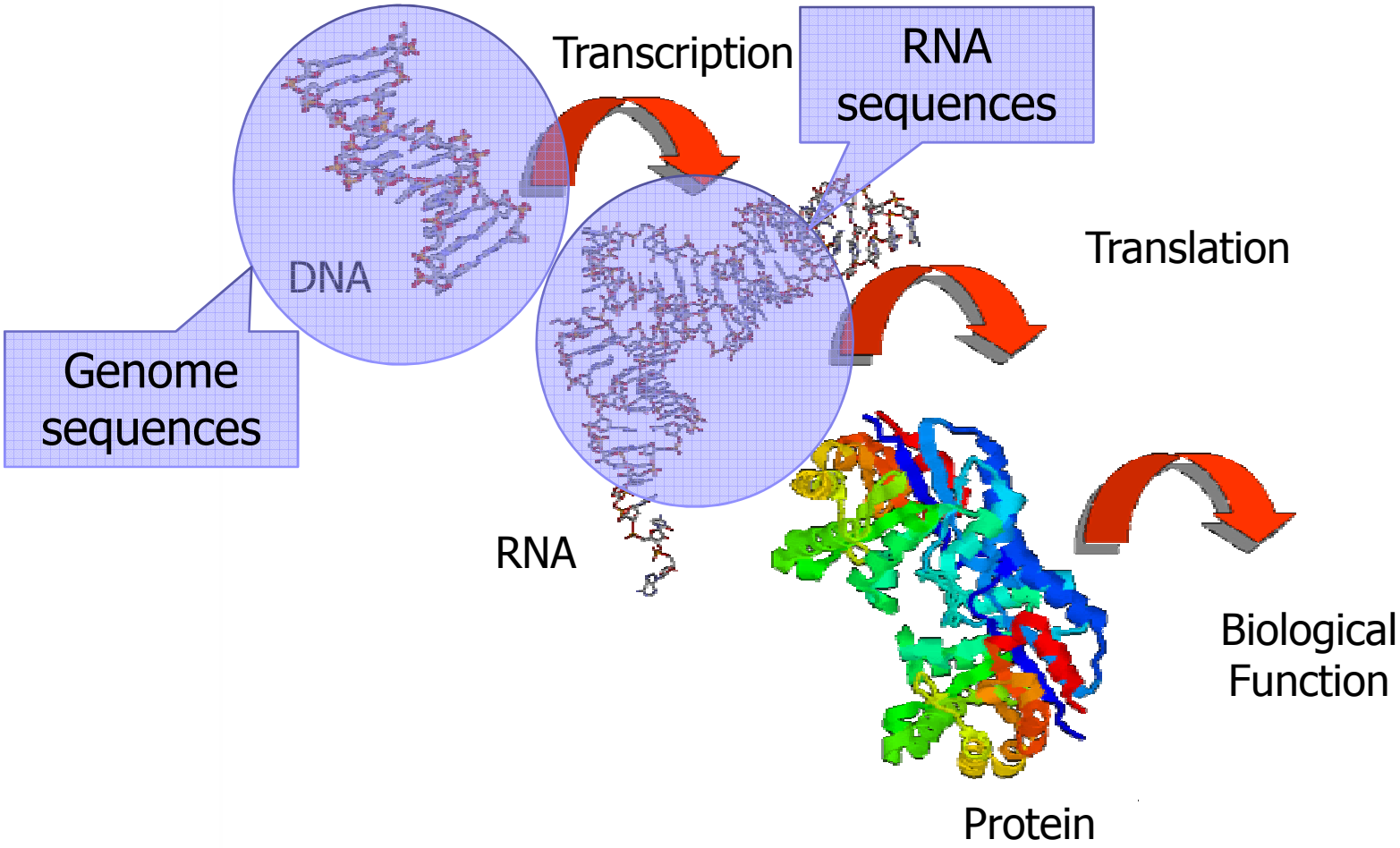


Jacques van Helden, David Gilbert and A.C. Tan, 2003

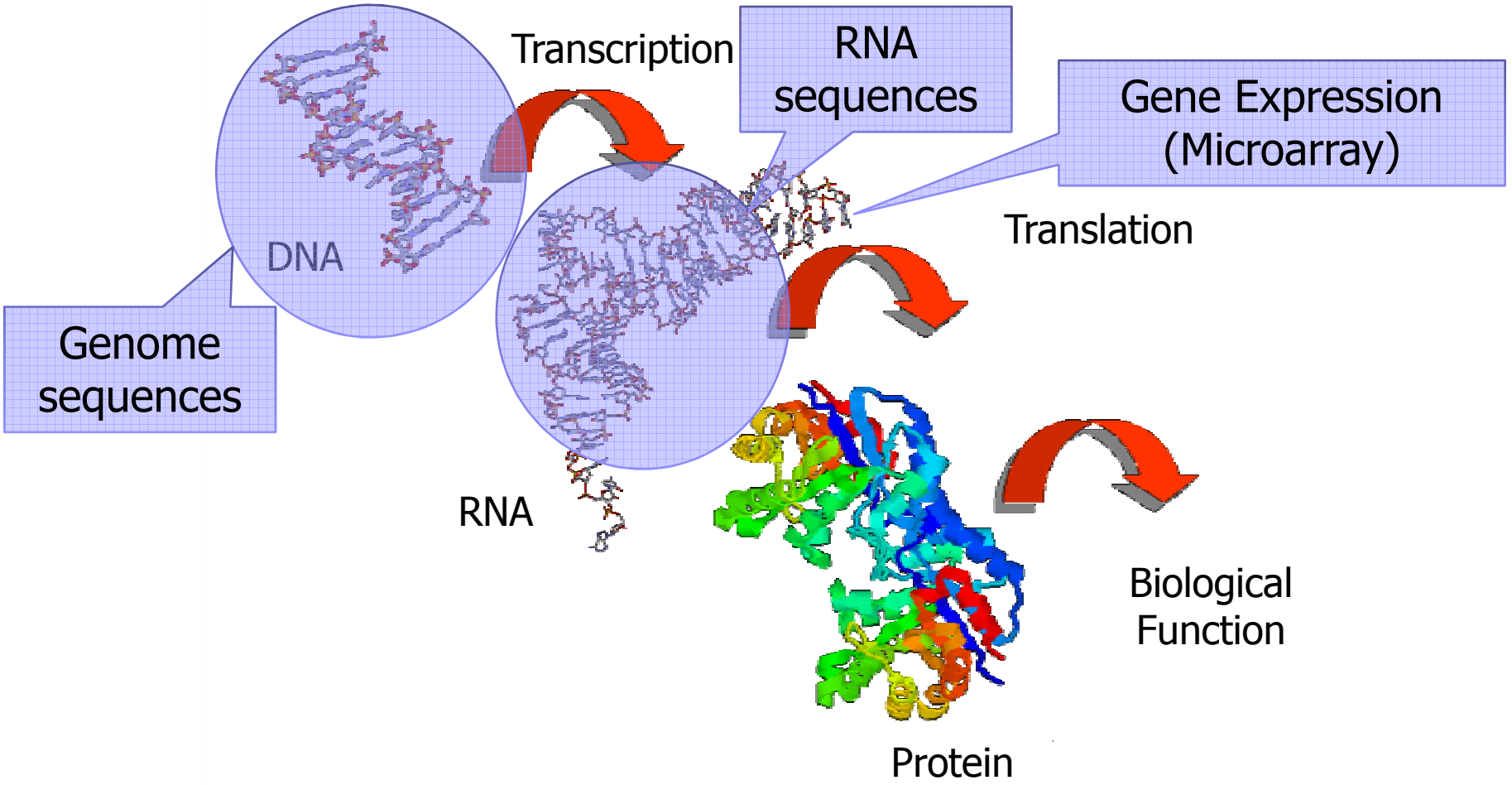
Biological Data



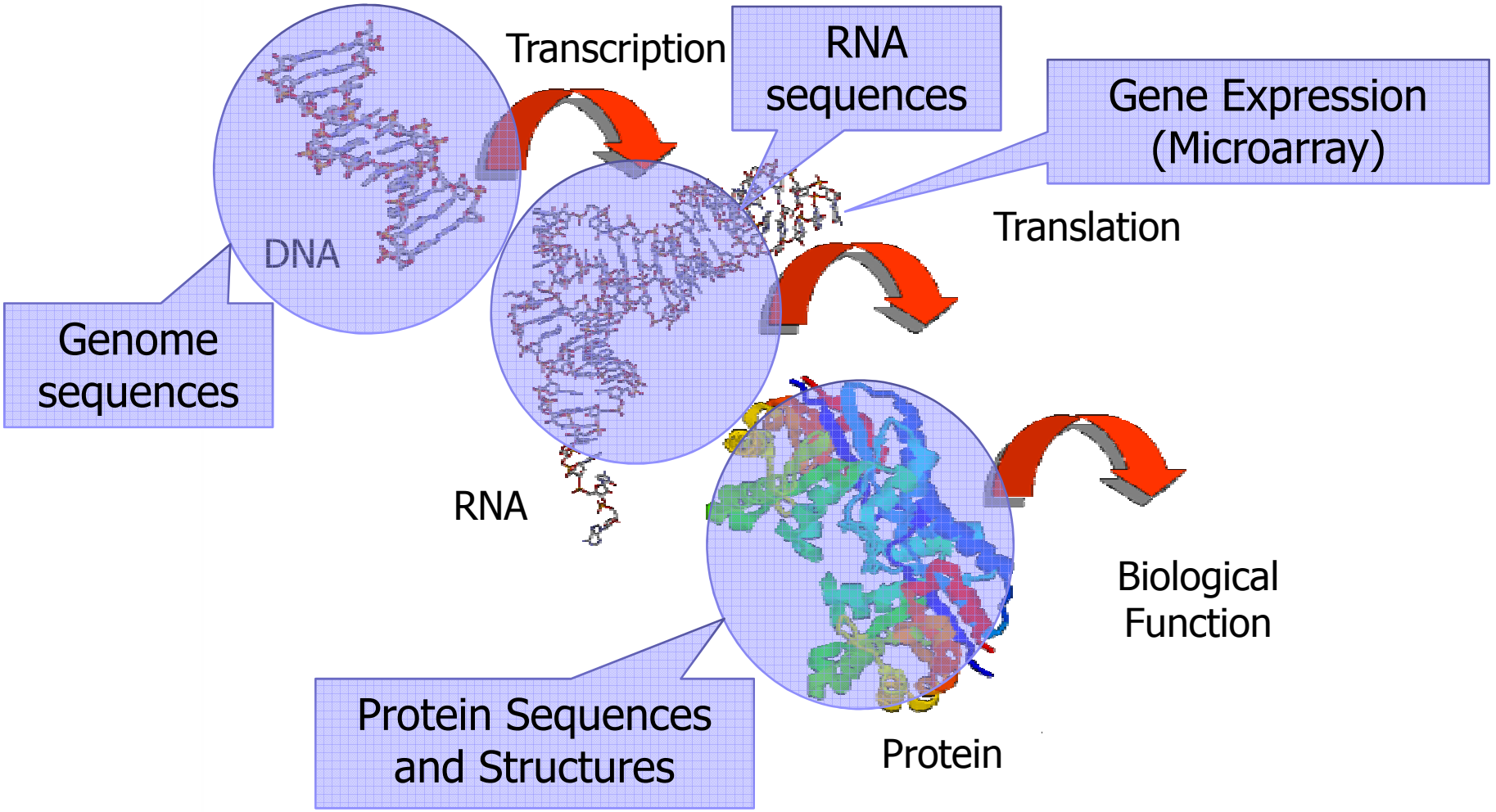
Biological Data



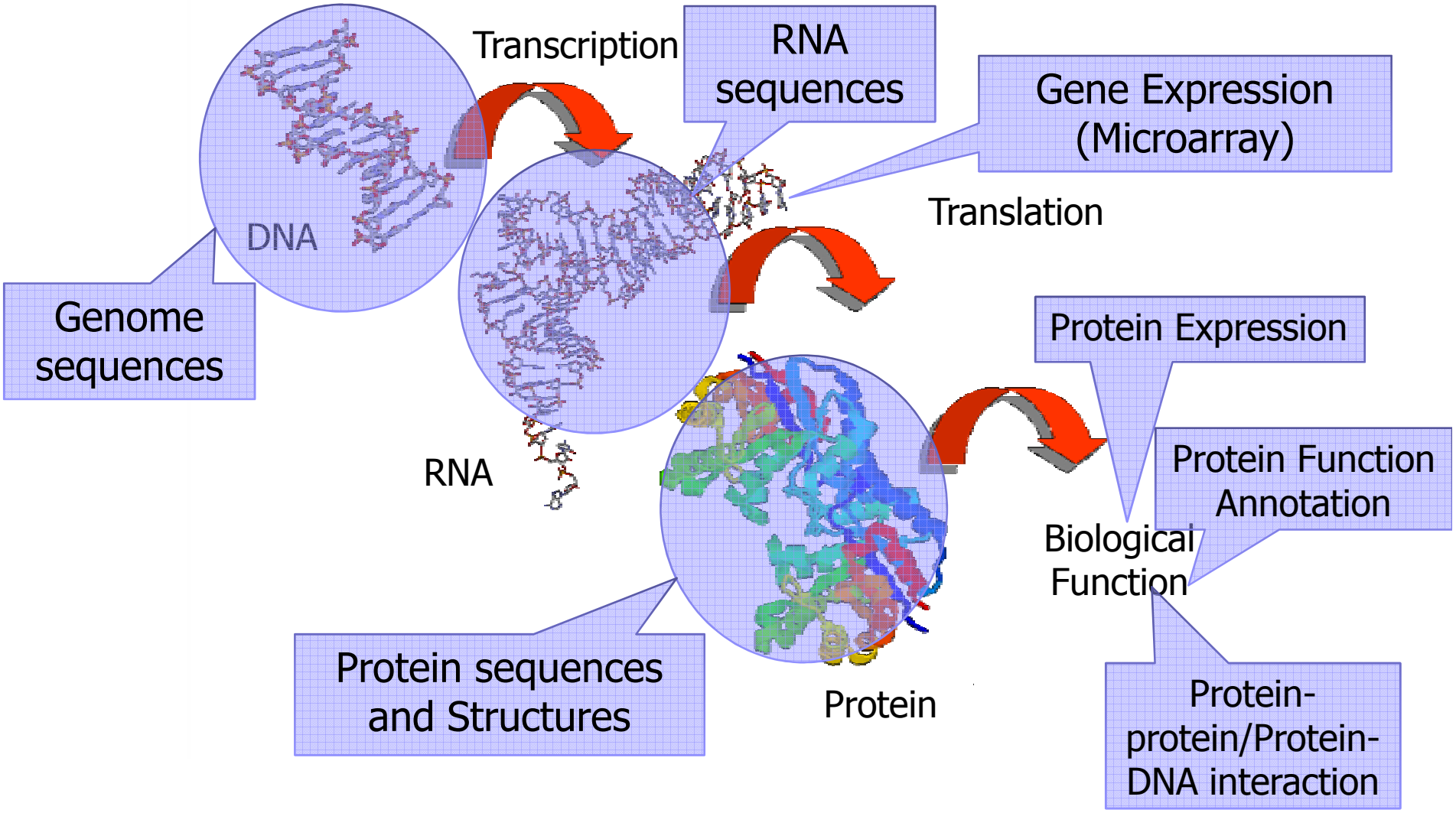
Biological Data



Biological Data



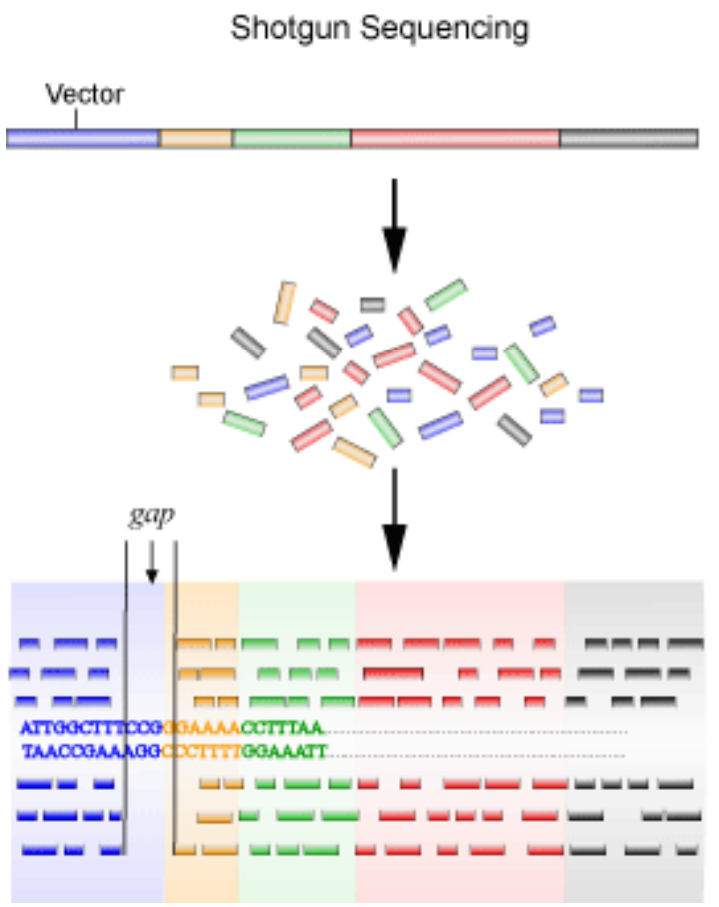
Biological Data



Other Data

- SNPs
- Organism-specific databases
- Genomes
- Molecular pathways
- Scientific literature
- Disease information
- ……

Combinatory Algorithms



- Get multiple copies of DNA segments.
- Alignment the segments to reconstruct the sequence.
- Closing the GAP with slow and expensive experiments.
- Combinatory algorithms for closing the gap with minimal number of pool tests.

CSCI8980: Applied Machine Learning in Computational Biology

Inferring Gene Regulatory Network with Bayesian Networks

Rui Kuang

Department of Computer Science and Engineering
University of Minnesota
kuang@cs.umn.edu

Cellular Networks

- Complex functions of cells are carried out by the coordinated activity of genes and their products
- C*ellular network* of interactions of 1000s of genes and their products
- New *high-throughput* genomic data, such as microarray data, enables *computational* study of cellular networks genome-widely.
- DNA (genes) ➔ mRNA ➔ proteins

Transcription

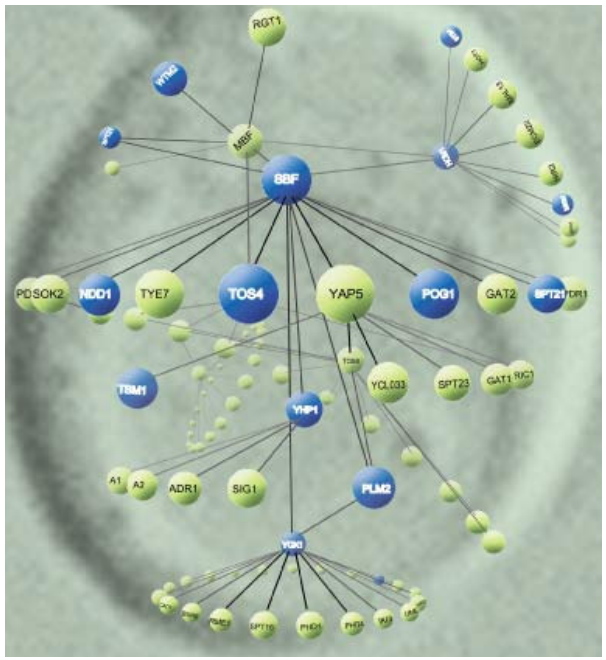


Figure: Snyder and Gerstein Labs

Gene Regulatory Networks

- *Gene regulatory networks*: switching on and off of genes by regulation of transcriptional machinery
- *Learning problem*: Model gene regulatory behavior using genome-wide data, extract hypotheses for wet lab testing
- *Descriptive models*, such as probabilistic graphical models, linear network models, clustering, are interpretable models to training data.
- Can check if local components of model reflect known biological mechanisms.

Gene Regulation

- *Regulatory proteins (transcription factors)* bind to non-coding *regulatory sequence (promoter)* of a gene to control rate of transcription

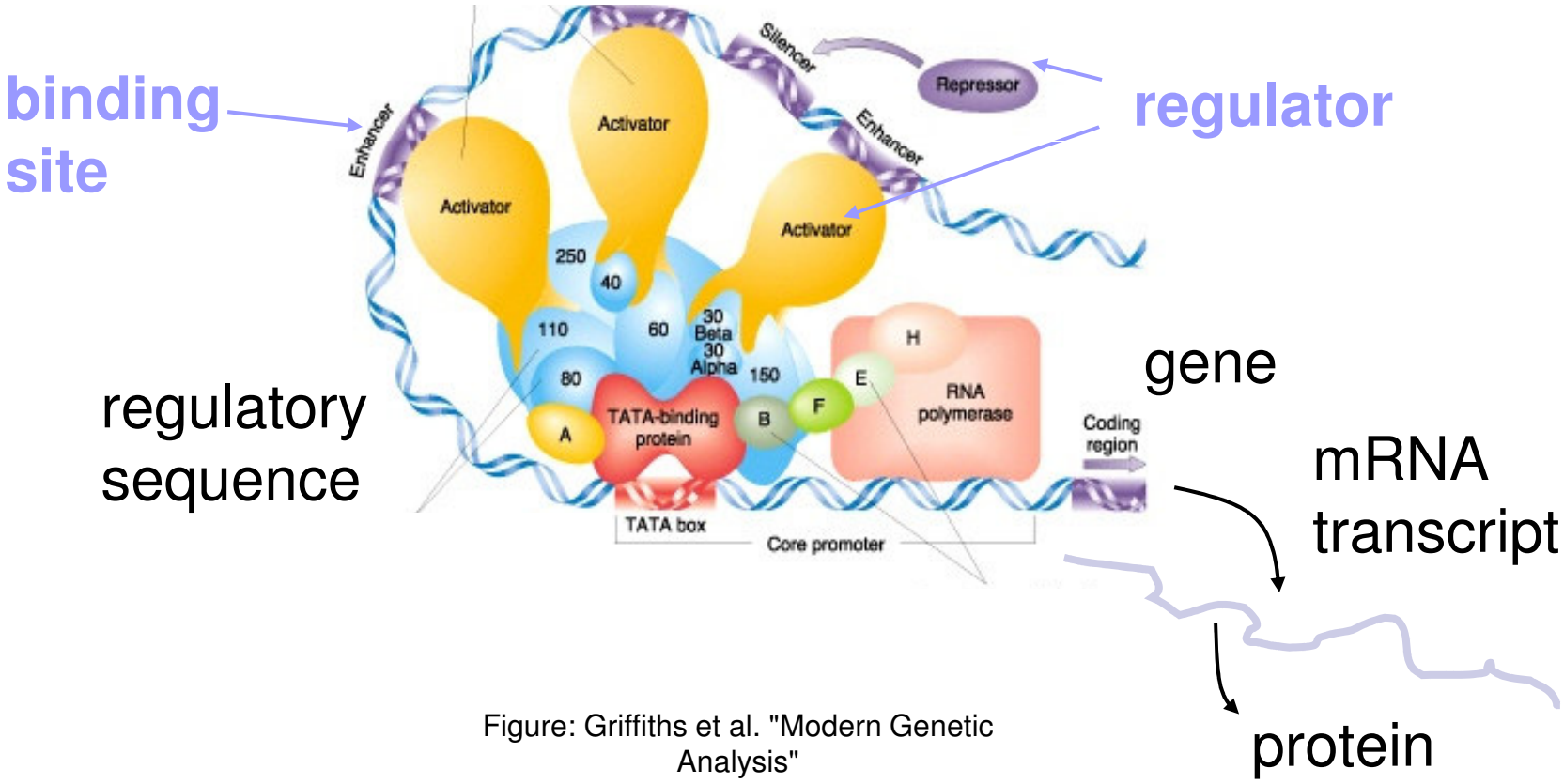


Figure: Griffiths et al. "Modern Genetic Analysis"

Gene Regulation

- *Regulatory proteins (transcription factors)* bind to non-coding *regulatory sequence (promoter)* of a gene to control rate of transcription

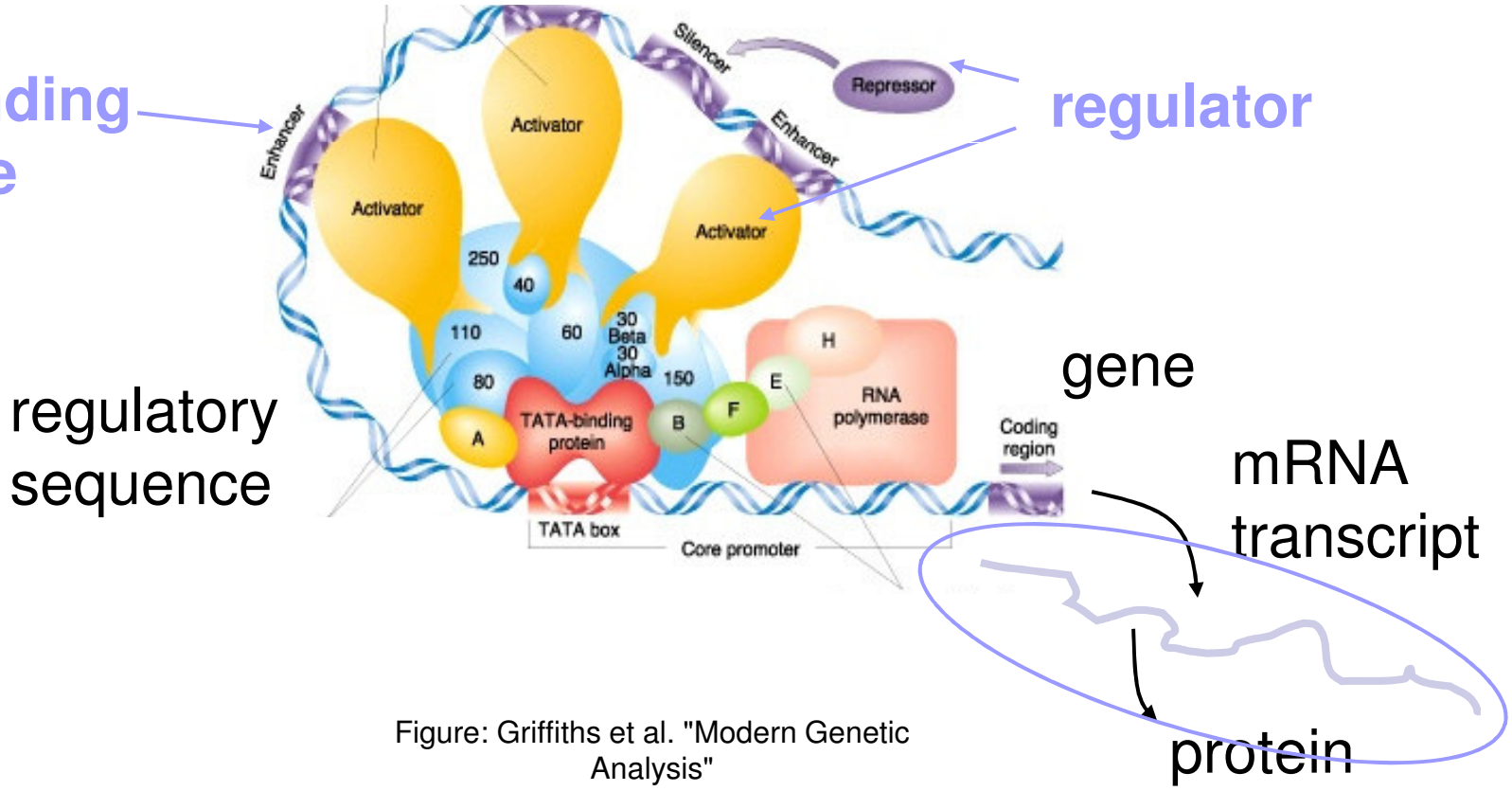
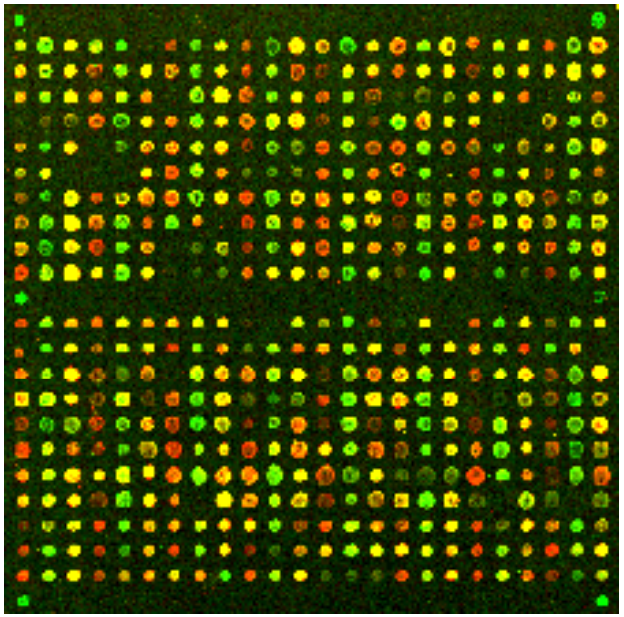


Figure: Griffiths et al. "Modern Genetic Analysis"

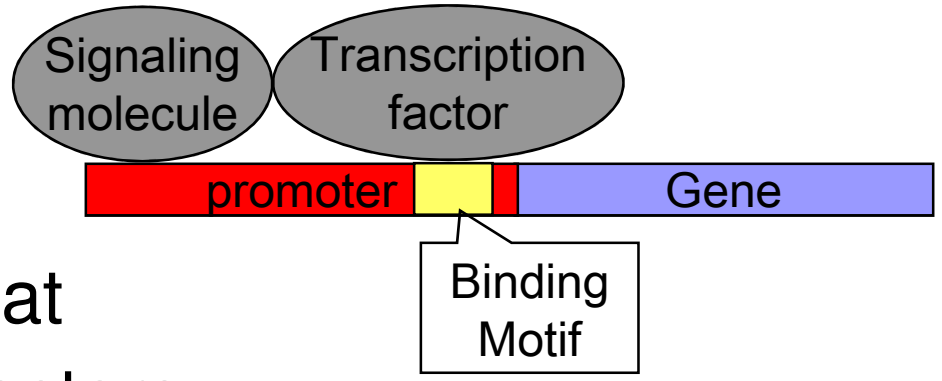
Genome-wide Expression Data



- *Microarray* (and other high-throughput) technologies measure *mRNA transcript expression levels* for 1000s of genes at once
- Noisy and sparse data
- Snapshot of the cellular system: *transcriptome*, i.e. protein expression not observed
- Difficult to infer regulatory relation between genes.

Regulatory Components in yeast

- For simple organisms like *yeast* (*S. cerevisiae*), previous studies and data sources the components needed in model:
  - Known and putative *transcription factors*
  - *Signaling molecules* that activate transcription factors
  - Known and putative *binding site "motifs"* in promoter regions
  - In yeast, regulatory sequence = 500 bp upstream region



Analyze Gene Expression Data

- Clustering
  - Groups genes with similar expression patterns
  - The gene clusters do not reveal the regulatory structure of the genes
- Boolean Networks
  - Deterministic models of the logical interactions between genes
  - Gene is in either on state or off state
  - Not feasible to learn from microarray data
- Bayesian Networks
  - Measure expression level of each gene
  - Gene as random variables affecting on others
  - Can possibly include other random variables, such as external stimuli, environment parameters, and biological factors

Model Validation of Genetic Regulatory Networks

- Using *Bayesian scoring* metric to choose the right network structure

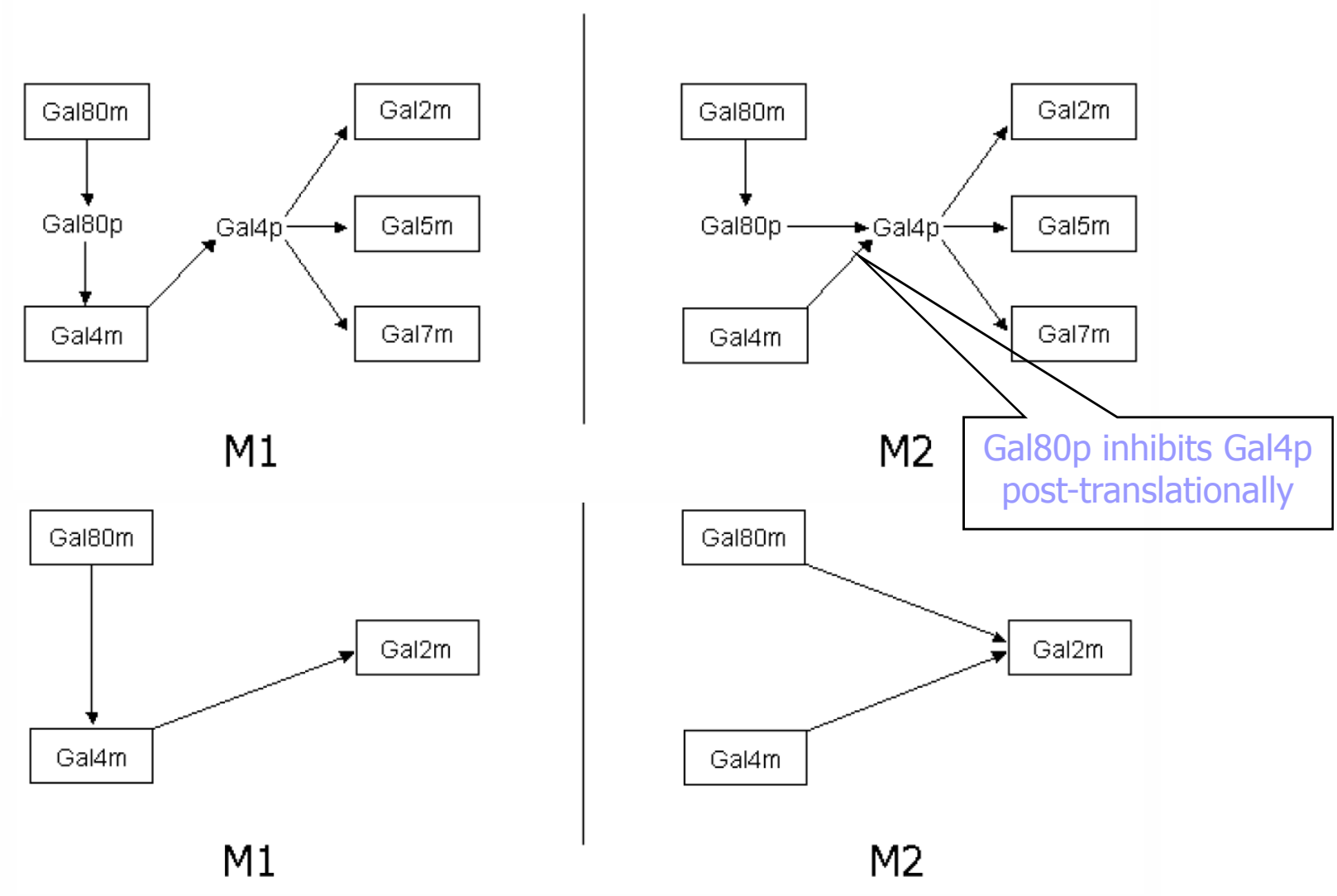
$$BayesianScore(S) = \log p(S \mid D)$$
$$= \log p(S) + \log p(D|S) + c,$$

where $p(D|S)$ is the likelihood function and $P(S)$ is a prior on the model S.

- Validated on the *galactose system* in S. cerevisiae
- Expression data: 52 genomes worth of Affymetrix GeneChip expression data

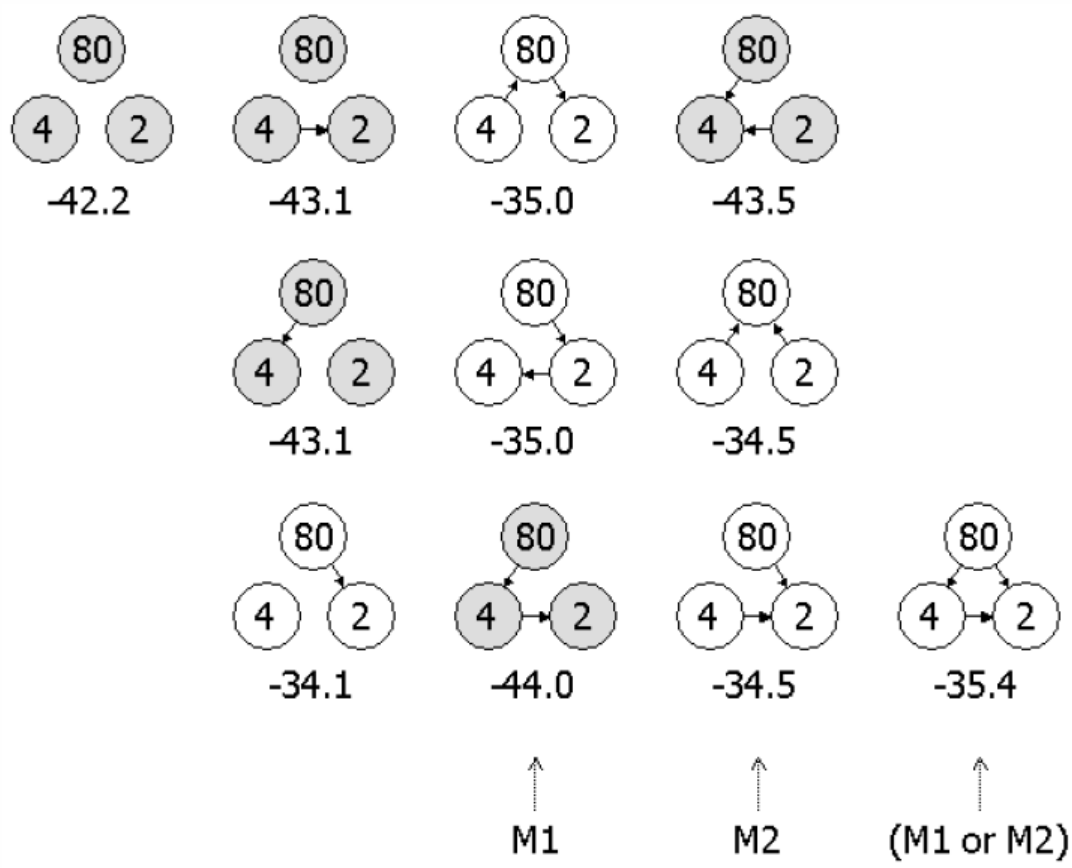
Hartemink et al. 2001

Hypothesis of Galactose System



Scoring Possible Structures

- Binary quantization of gene expression into up/down (3 binary random variables)



Scoring Possible Structures

- Binary quantization of gene expression into up/down (3 binary random variables)