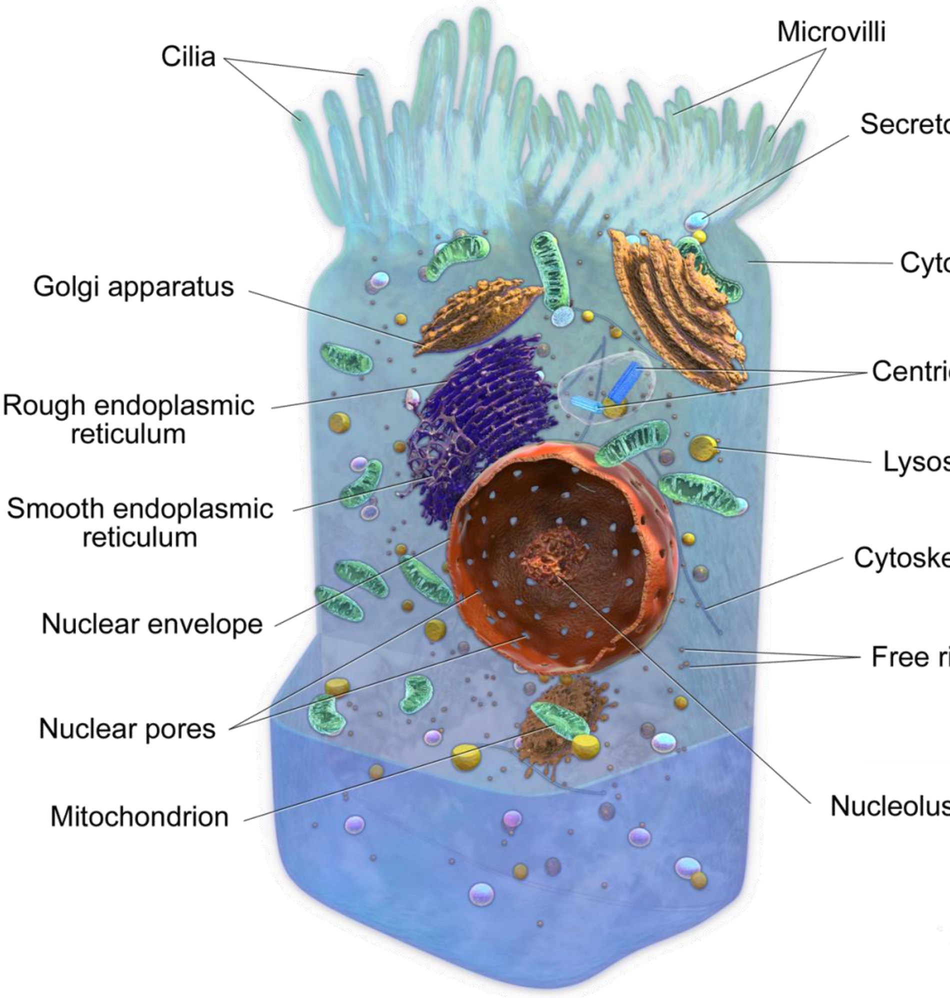
Applications of Machine Learning in Computational Biology

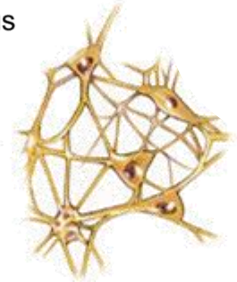
Narges Razavian

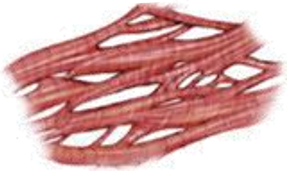
New York University

Slides thanks to James Galagan@Board Institute Su-In Lee@Univ of Washington
Rainer Breitling@ Univ of Glasgow Christopher M. Bishop@ ECCV 2004



Anatomy of a Cell


Neural cells


Cardiac muscle

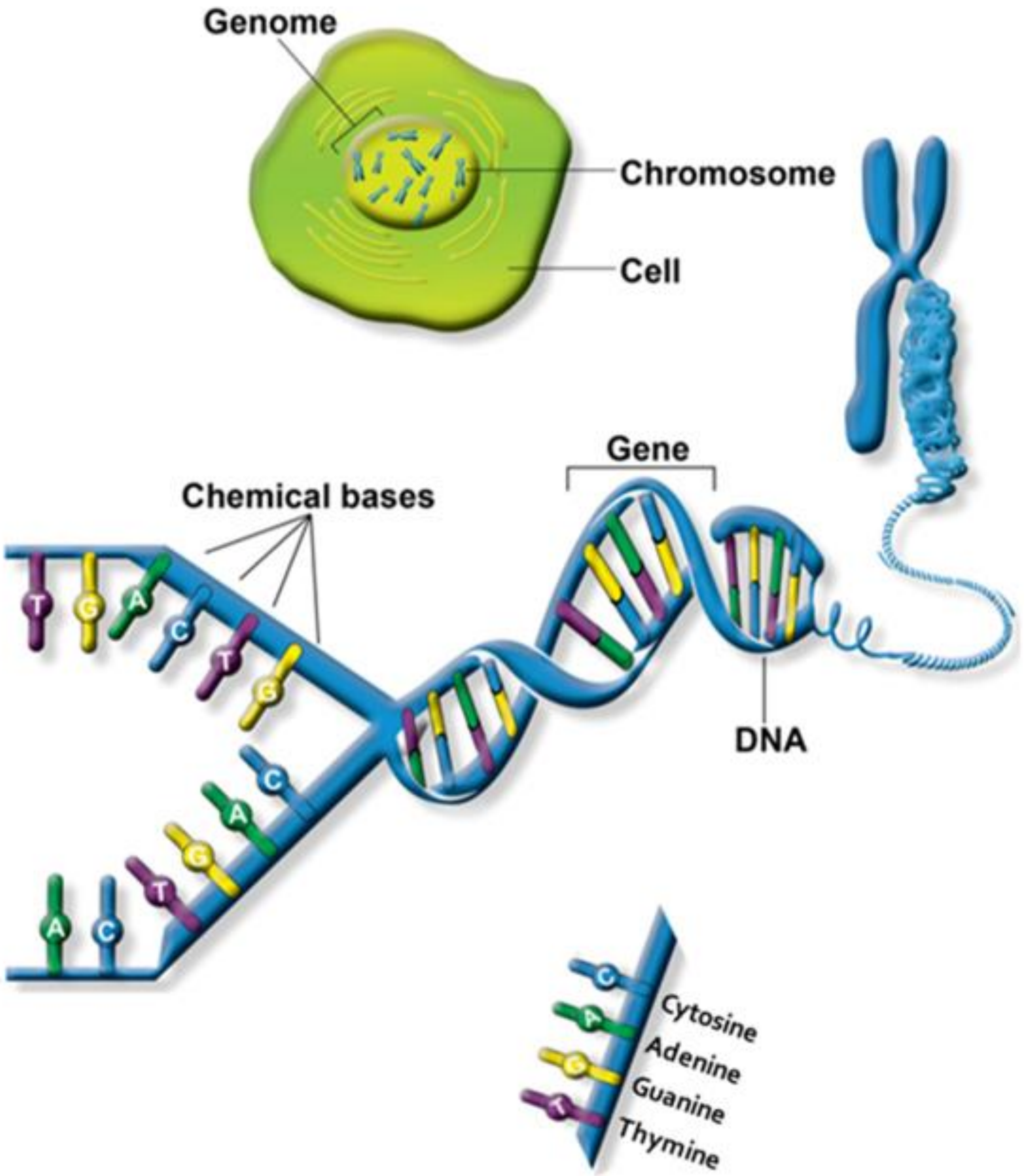
Blood cells

SNP SNP

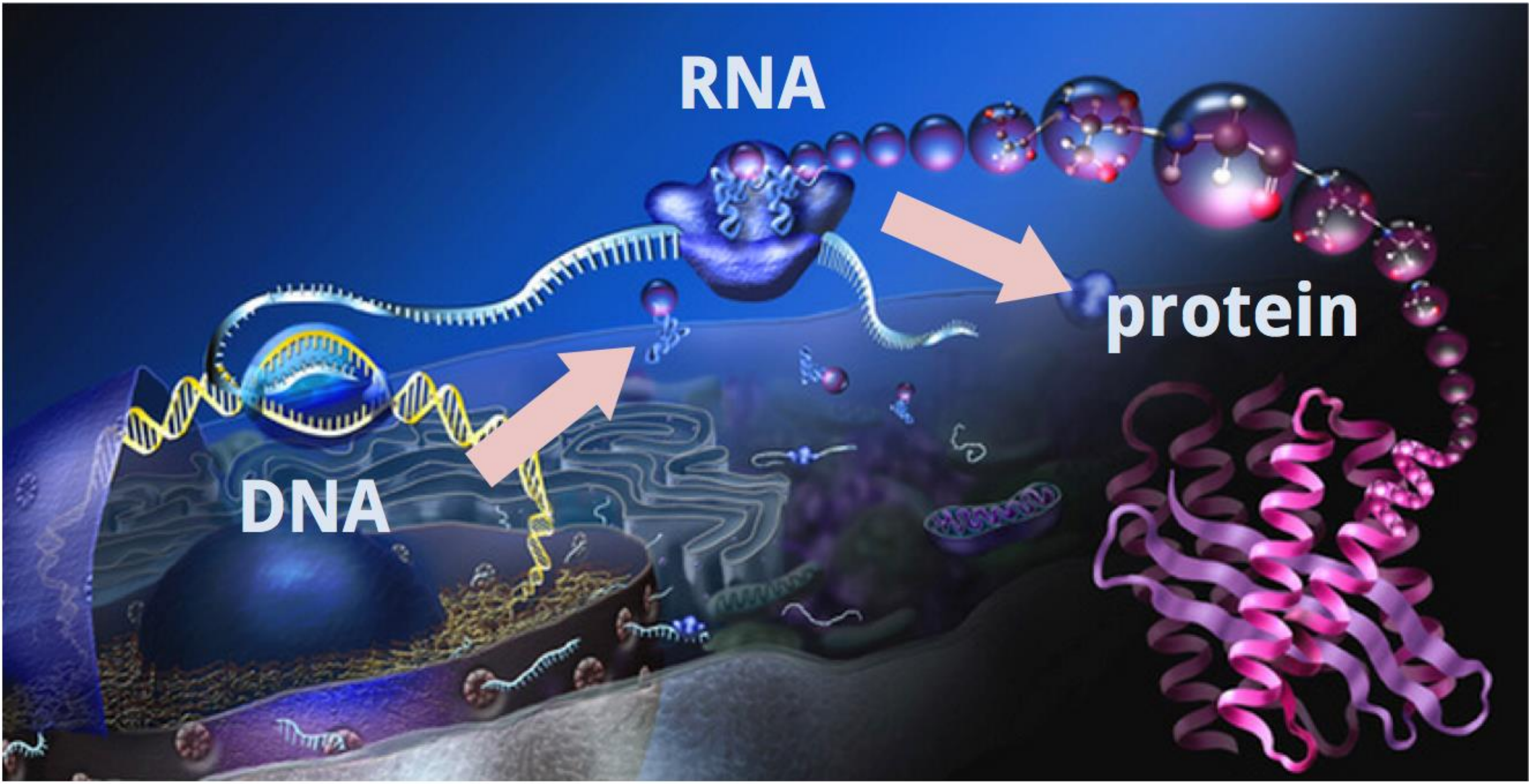
CTACT?CTGCCCTTAAG?TCCC

SNP SNP

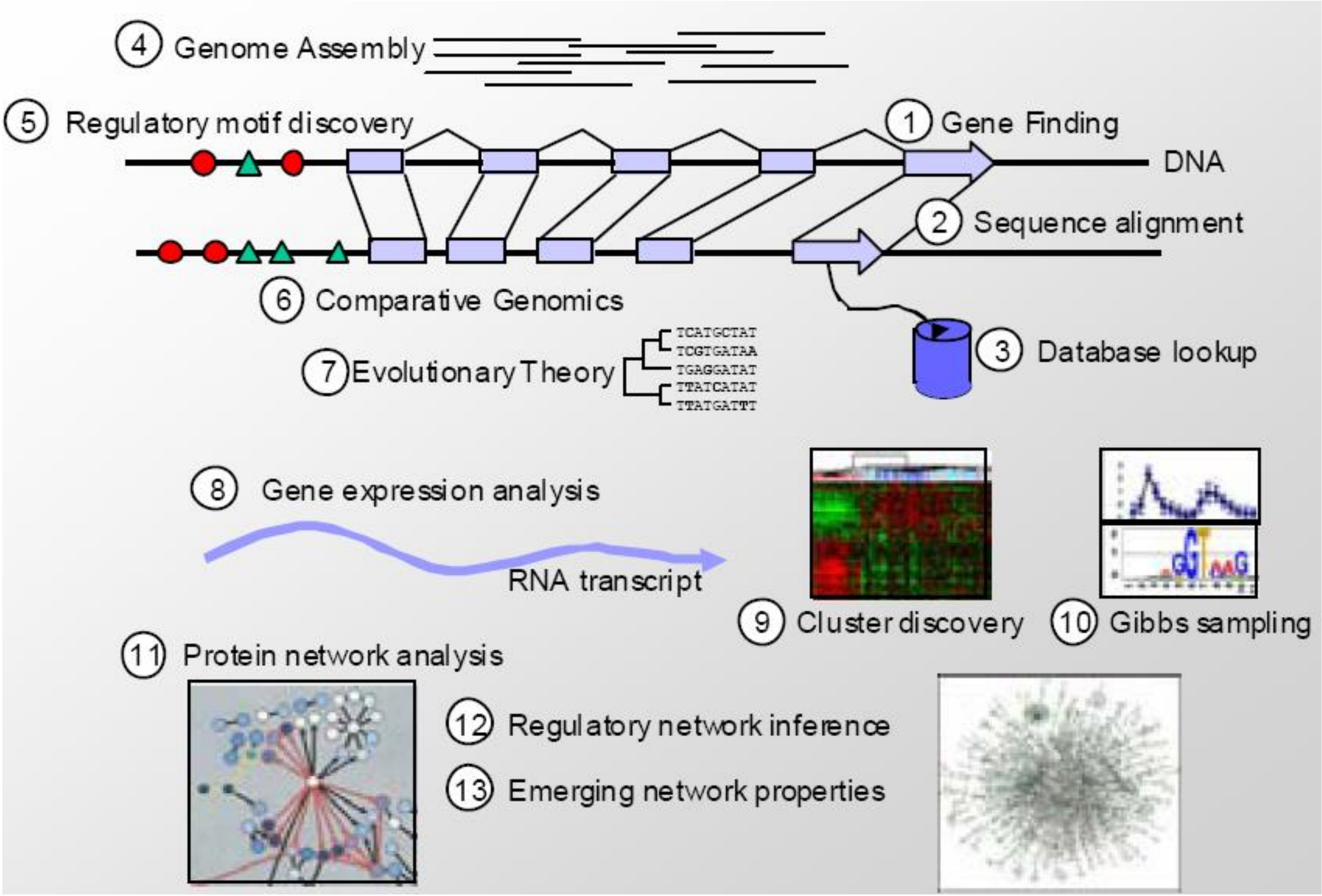
GATGA?GACGGGAATTCC?GGG



Central Dogma of Biology



Examples of Challenges involved



Slide Credit: Manolis Kellis

Application : Decoding Sequences and Motif Discovery

Motif Discovery

```
GCGTCTGACGGCGCACCGTTCGCGCTGCCGGCACCCCGGGCTCCATAATGAAAATCATGT
TCAGTAAGCTACACTCTGCATATCGGGCTACCAACGAAATGGAGTATCGGTCATGATCTT
GCCAGCCGTGCCTAAAAGCTTGGCCGCAGGGCCGAGTATAATTGGTCGCGGTCGCCTCGA
AGTTAGCTTATGCAATGCAGGAGGTGGGGCAAAGTTCAGGCGGATCGGCCGATGGCGGGC
GTAGGTGAAGGAGACAGCGGAGGCGTGGAGCGTGATGACATTGGCATGGTGGCCGCTTCC
CCCGTCGCGTCTCGGGTAAATGGCAAGGTAGACGCTGACGTCGTCGGTCGATTTGCCACC
TGCTGCCGTGCCCTGGGCATCGCGGTTTACCAGCGTAAACGTCCGCCGGACCTGGCTGCC
GCCCGGTCTGGTTTCGCCGCGCTGACCCGCGTCGCCCATGACCAGTGCGACGCCTGGACC
GGGCTGGCCGCTGCCGGCGACCAGTCCATCGGGGTGCTGGAAGCCGCCTCGCGCACGGCG
ACCACGGCTGGTGTGTTGCAGCGGCAGGTGGAACTGGCCGATAACGCCTTGGGCTTCCTG
TACGACACCGGGCTGTACCTGCGTTTTCGTGCCACCGGACCTGACGATTTCCACCTCGCG
TATGCCGCTGCGTTGGCTTCGACGGGCGGGCCGGAGGAGTTTGCCAAGGCCAATCACGTG
GTGTCCGGTATCACCGAGCGCCGCGCCGGCTGGCGTGCCGCCCGTTGGCTCGCCGTGGTC
ATCAACTACCGCGCCGAGCGCTGGTCGGATGTCGTGAAGCTGCTCACTCCGATGGTTAAT
GATCCCGACCTCGACGAGGCCTTTTCGCACGCGGCCAAGATCACCCTGGGCACCGCACTG
GCCCGACTGGGCATGTTTGCCCCGGCGCTGTCTTATCTGGAGGAACCCGACGGTCCTGTC
GCGGTCGCTGCTGTCGACGGTGCACTGGCCAAAGCGCTGGTGCTGCGCGCGCATGTGGAT
ATGGAGTCGGCCAGCGAAGTGCTGCAGGACTTGTATGCGGCTCACCCCGAAAACGAACAG
GTCGAGCAGGCGCTGTCGGATACCAGCTTCGGGATCGTCACCACCACAGCCGGGCGGATC
GAGGCCCGCACCGATCCGTGGGATCCGGCGACCGAGCCCGGCGCGGAGGATTTCGTCGAT
CCCGCGGCCCACGAACGCAAGGCCGCGCTGCTGCACGAGGCCGAACTCCAACTCGCCGAG
```

Sequence Annotation

```
GCGTCTGACGGCGCACCGTTCGCGCTGCCGGCACCCCGGGCTCCATAATGAAAATCATGT
TCAGTAAGCTACACTCTGCATATCGGGCTACCAACGAAATGGAGTATCGGTCATGATCTT
GCCAGCCGTGCCTAAAAGCTTGGCCGCAGGGCCGAGTATAATTGGTCGCGGTCGCCTCGA
AGTTAGCTTATGCAATGCAGGAGGTGGGGCAAAGTTCAGGCGGATCGGCCGATGGCGGGC
GTAGGTGAAGGAGACAGCGGAGGCGTGGAGCGTGATGACATTGGCATGGTGGCCGCTTCC
CCCGTCGCGTCTCGGGTAAATGGCAAGGTAGACGCTGACGTCGTCGGTCGATTTGCCACC
TGCTGCCGTGCCCTGGGCATCGCGGTTTACCAGCGTAAACGTCCGCCGGACCTGGCTGCC
GCCCGGTCTGGTTTCGCCGCGCTGACCCGCGTCGCCCATGACCACTGCCACGCCTGGACC
GGGCTGGCCGCTGCCGGCGACCAGTCCATCGGGTGC[illegible]CGCGCACGGCG
ACCACGGCTGGTGTGTTGCAGCGGCAGGTGGAACTGC[illegible]TGGGCTTCCTG
TACGACACCGGGCTGTACCTGCGTTTTCGTGCCACCG[illegible]TGACGATTTCCACCTCGCG
TATGCCGCTGCGTTGGCTTCGACGGGCGGGCCGGAGGAGTTTGCCAAGGCCAATCACGTG
GTGTCCGGTATCACCGAGCGCCGCGCCGGCTGGCGTGCCGCCCGTTGGCTCGCCGTGGTC
ATCAACTACCGCGCCGAGCGCTGGTCGGATGTCGTGAAGCTGCTCACTCCGATGGTTAAT
GATCCCGACCTCGACGAGGCCTTTTCGCACGCGGCCAAGATCACCCTGGGCACCGCACTG
GCCCGACTGGGCATGTTTGCCCCGGCGCTGTCTTATCTGGAGGAACCCGACGGTCCTGTC
GCGGTCGCTGCTGTCGACGGTGCACTGGCCAAAGCGCTGGTGCTGCGCGCGCATGTGGAT
ATGGAGTCGGCCAGCGAAGTGCTGCAGGACTTGTATGCGGCTCACCCCGAAAACGAACAG
GTCGAGCAGGCGCTGTCGGATACCAGCTTCGGGATCGTCACCACCACAGCCGGGCGGATC
GAGGCCCGCACCGATCCGTGGGATCCGGCGACCGAGCCCGGCGCGGAGGATTTCGTCGAT
CCCGCGGCCCACGAACGCAAGGCCGCGCTGCTGCACGAGGCCGAACTCCAACTCGCCGAG
```

Gene

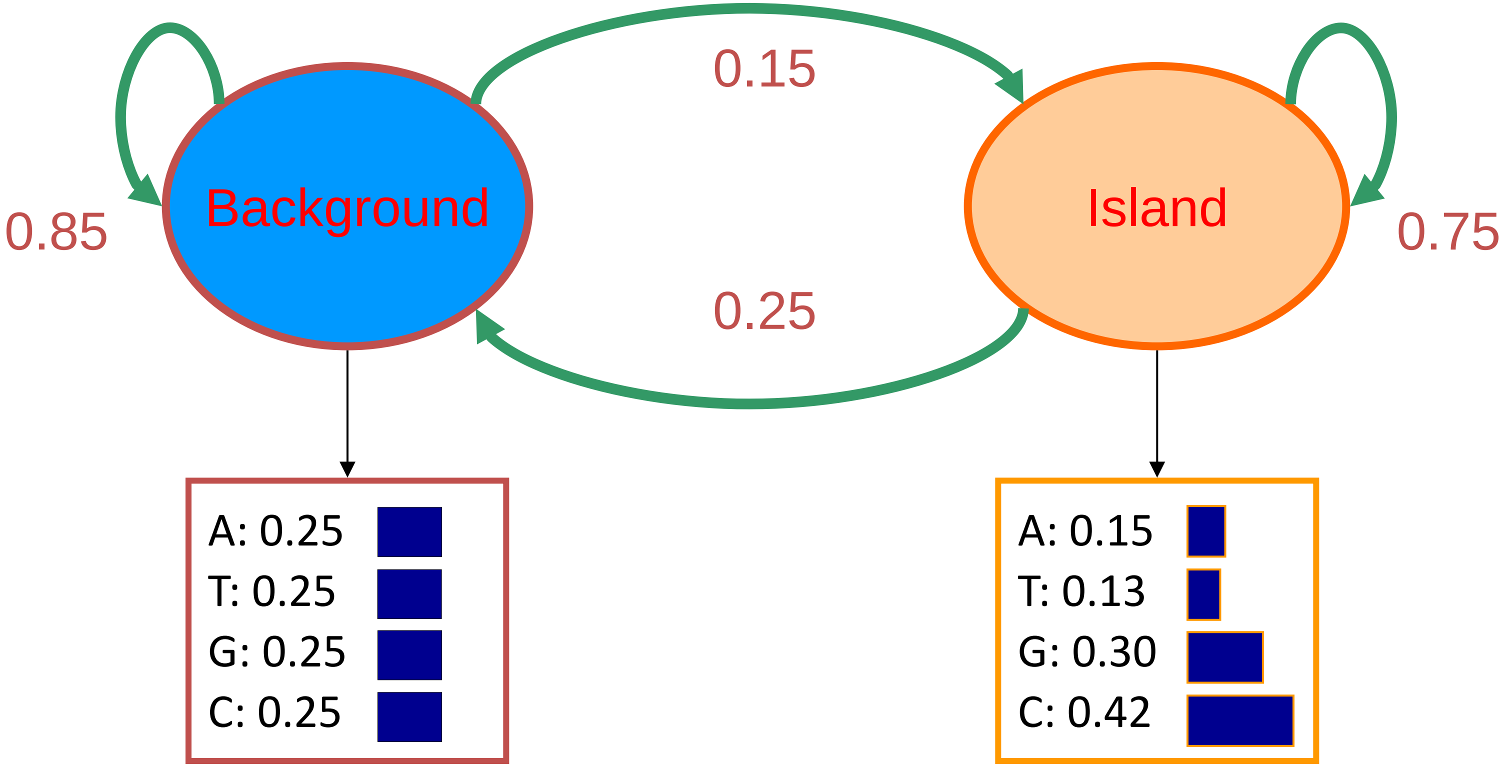
Sequence Annotation

```
GCGTCTGACGGCGCACCGTTCGCGCTGCCGGC ATAATGAAAATCATGT
TCAGTAAGCTACACTCTGCATATCGGGCTACC TATCGGTCATGATCTT
GCCAGCCGTGCCTAAAAGCTTGGCC GTCGCGGTCGCCTCGA
AGTTAGCTTATGCAATGCAGGAGGTGGGGCAAAGTTCAGGCGGATCGGCCGATGGCGGGC
GTAGGTGAAGGAGACAGCGGAGGCGTGGAGCGTGATGACATTGGCATGGTGGCCGCTTCC
CCCGTCGCGTCTCGGGTAAATGGCAAGGTAGACGCTGACGTCGTCGGTCGATTTGCCACC
TGCTGCCGTGCCCTGGGCATCGCGGTTTACCAGCGTAAACGTCCGCCGGACCTGGCTGCC
GCCCGGTCTGGTTTCGCCGCGCTGACCCGCGTCGCCC ACGCCTGGACC
GGGCTGGCCGCTGCCGGCGACCAGTCCATCGGGTGC CGCGCACGGCG
ACCACGGCTGGTGTGTTGCAGCGGCAGGTGGAACTGC TGGGCTTCCTG
TACGACACCGGGCTGTACCTGCGTTTTCGTGCCACCG TTCCACCTCGCG
TATGCCGCTGCGTTGGCTTCGACGGGCGGGCCGGAGGAGTTTGCCAAGGCCAATCACGTG
GTGTCCGGTATCACCGAGCGCCGCGCCGGCTGGCGTGCCGCCCGTTGGCTCGCCGTGGTC
ATCAACTACCGCGCCGAGCGCTGGTCGGATGTCGTGAAGCTGCTCACTCCGATGGTTAAT
GATCCCGACCTCGACGAGGCCTTTTCGCACGCGGCCAAGATCACCCTGGGCACCGCACTG
GCCCGACTGGGCATGTTTGCCCCGGCGCTGTCTTATCTGGAGGAACCCGACGGTCCTGTC
GCGGTCGCTGCTGTCGACGGTGCACTGGCCAAAGCGCTGGTGCTGCGCGCGCATGTGGAT
ATGGAGTCGGCCAGCGAAGTGCTGCAGGACTTGTATGCGGCTCACCCCGAAAACGAACAG
GTCGAGCAGGCGCTGTCGGATACCAGCTTCGGGATCGTCACCACCACAGCCGGGCGGATC
GAGGCCCGCACCGATCCGTGGGATCCGGCGACCGAGCCCGGCGCGGAGGATTTCGTCGAT
CCCGCGGCCCACGAACGCAAGGCCGCGCTGCTGCACGAGGCCGAACTCCAACTCGCCGAG
```

Promoter Motif

Gene

A *Generative* Model



0.15

0.85

Background

Island

0.75

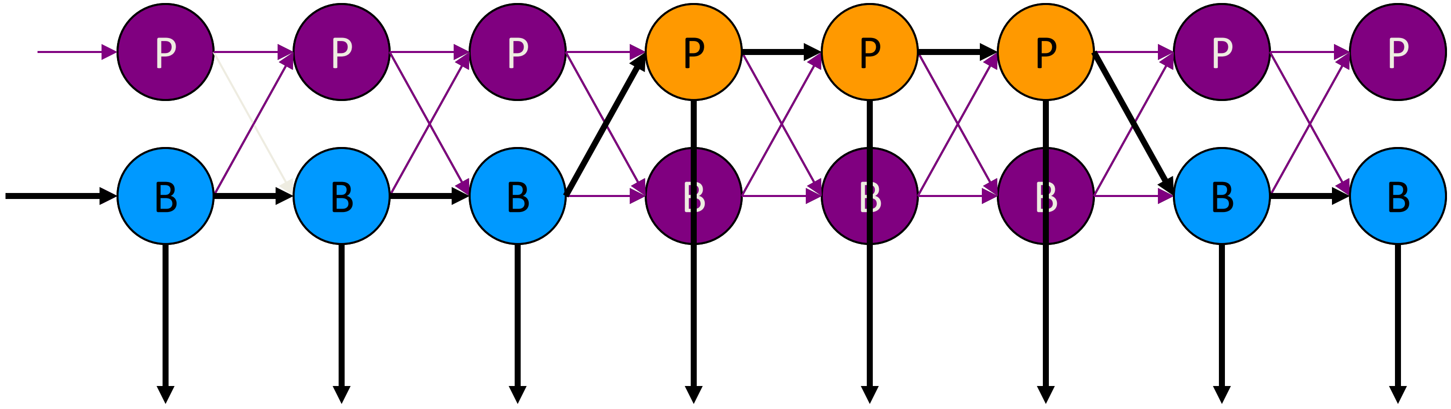
0.25

A: 0.25
T: 0.25
G: 0.25
C: 0.25

A: 0.15
T: 0.13
G: 0.30
C: 0.42

TAAGAATTGTGTCACACACATAAAAACCCTAAGTTAGAGGATTGAGATTGGCA
GACGATTGTTCGTGATAATAAACAAGGGGGGCATAGATCAGGCTCATATTGGC

A *Generative* Model(cont.)



S: G C A A A T G C

$P(L_{i+1}|L_i)$

	B_{i+1}	P_{i+1}
B_i	**0.85**	**0.15**
P_i	**0.25**	**0.75**

P(S|B)

A: 0.25	■
T: 0.25	■
G: 0.25	■
C: 0.25	■

P(S|P)

A: 0.42	■
T: 0.30	■
G: 0.13	■
C: 0.15	■

Fundamental HMM Operations

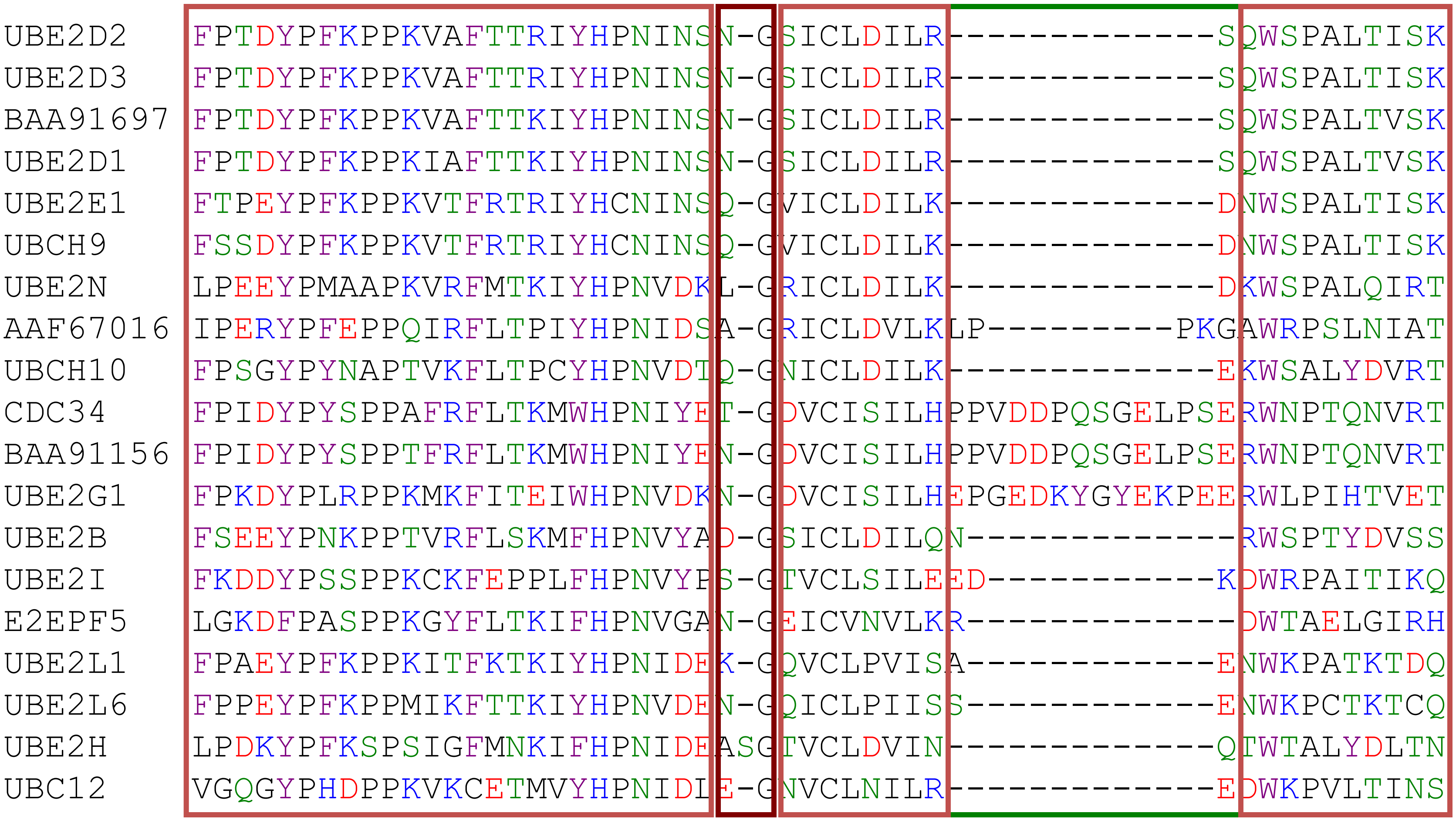
Computation	Biology
Decoding • *Given* an HMM and sequence S • *Find* a corresponding sequence of labels, L	Annotate pathogenicity islands on a new sequence
Evaluation • *Given* an HMM and sequence S • *Find* P(S|HMM)	Score a particular sequence (not as useful for this model – will come back to this later)
Training • *Given* an HMM w/o parameters and set of sequences S • *Find* transition and emission probabilities the maximize P(S | params, HMM)	Learn a model for sequence composed of background DNA and pathogenicity islands

Application: Modeling Protein Families

Modeling Protein Families

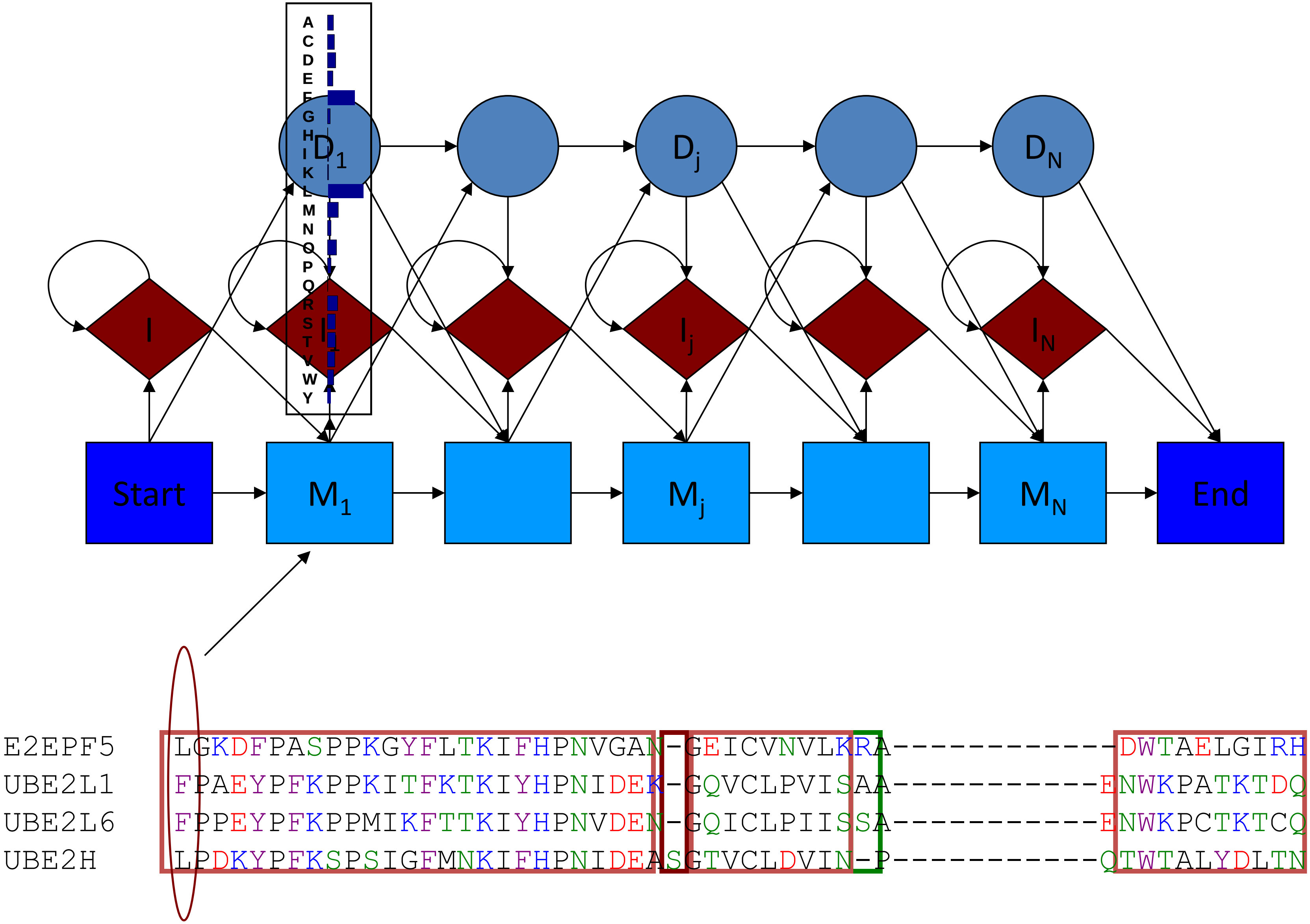
- Given amino acid sequences from a protein family, how can we find other members?
  - Can search databases with each known member – not sensitive
  - More information is contained in full set
- The HMM Profile Approach
  - Learn the statistical features of protein family
  - Model these features with an HMM
  - Search for new members by scoring with HMM

Human Ubiquitin Conjugating Enzymes



```
UBE2D2    FPTDYPFKPPKVAFTTRIYHPNINSN-GSICLDILR-------------SQWSPALTISK
UBE2D3    FPTDYPFKPPKVAFTTRIYHPNINSN-GSICLDILR-------------SQWSPALTISK
BAA91697  FPTDYPFKPPKVAFTTKIYHPNINSN-GSICLDILR-------------SQWSPALTVSK
UBE2D1    FPTDYPFKPPKIAFTTKIYHPNINSN-GSICLDILR-------------SQWSPALTVSK
UBE2E1    FTPEYPFKPPKVTFRTRIYHCNINSQ-GVICLDILK-------------DNWSPALTISK
UBCH9     FSSDYPFKPPKVTFRTRIYHCNINSQ-GVICLDILK-------------DNWSPALTISK
UBE2N     LPEEYPMAAPKVRFMTKIYHPNVDKL-GRICLDILK-------------DKWSPALQIRT
AAF67016  IPERYPFEPPQIRFLTPIYHPNIDSA-GRICLDVLKLP---------PKGAWRPSLNIAT
UBCH10    FPSGYPYNAPTVKFLTPCYHPNVDTQ-GNICLDILK-------------EKWSALYDVRT
CDC34     FPIDYPYSPPAFRFLTKMWHPNIYET-GDVCISILHPPVDDPQSGELPSERWNPTQNVRT
BAA91156  FPIDYPYSPPTFRFLTKMWHPNIYEN-GDVCISILHPPVDDPQSGELPSERWNPTQNVRT
UBE2G1    FPKDYPLRPPKMKFITEIWHPNVDKN-GDVCISILHEPGEDKYGYEKPEERWLPIHTVET
UBE2B     FSEEYPNKPPTVRFLSKMFHPNVYAD-GSICLDILQN-------------RWSPTYDVSS
UBE2I     FKDDYPSSPPKCKFEPPLFHPNVYPS-GTVCLSILEED-----------KDWRPAITIKQ
E2EPF5    LGKDFPASPPKGYFLTKIFHPNVGAN-GEICVNVLKR-------------DWTAELGIRH
UBE2L1    FPAEYPFKPPKITFKTKIYHPNIDEK-GQVCLPVISA------------ENWKPATKTDQ
UBE2L6    FPPEYPFKPPMIKFTTKIYHPNVDEN-GQICLPIISS------------ENWKPCTKTCQ
UBE2H     LPDKYPFKSPSIGFMNKIFHPNIDEASGTVCLDVIN-------------QTWTALYDLTN
UBC12     VGQGYPHDPPKVKCETMVYHPNIDLE-GNVCLNILR-------------EDWKPVLTINS
```

Profile HMM



```
E2EPF5    LGKDFPASPPKGYFLTKIFHPNVGAN-GEICVNVLKRA------------DWTAELGIRH
UBE2L1    FPAEYPFKPPKITFKTKIYHPNIDEK-GQVCLPVISAA-----------ENWKPATKTDQ
UBE2L6    FPPEYPFKPPMIKFTTKIYHPNVDEN-GQICLPIISSA-----------ENWKPCTKTCQ
UBE2H     LPDKYPFKSPSIGFMNKIFHPNIDEASGTVCLDVIN-P-----------QTWTALYDLTN
```

Using Profile HMMs

Computation

Biology

Decoding

Find sequence of labels, L, that maximizes P(L|S, HMM)

Align a new sequence to a protein family

Evaluation

- *Find* P(S|HMM)

Score a sequence for membership in family

Training

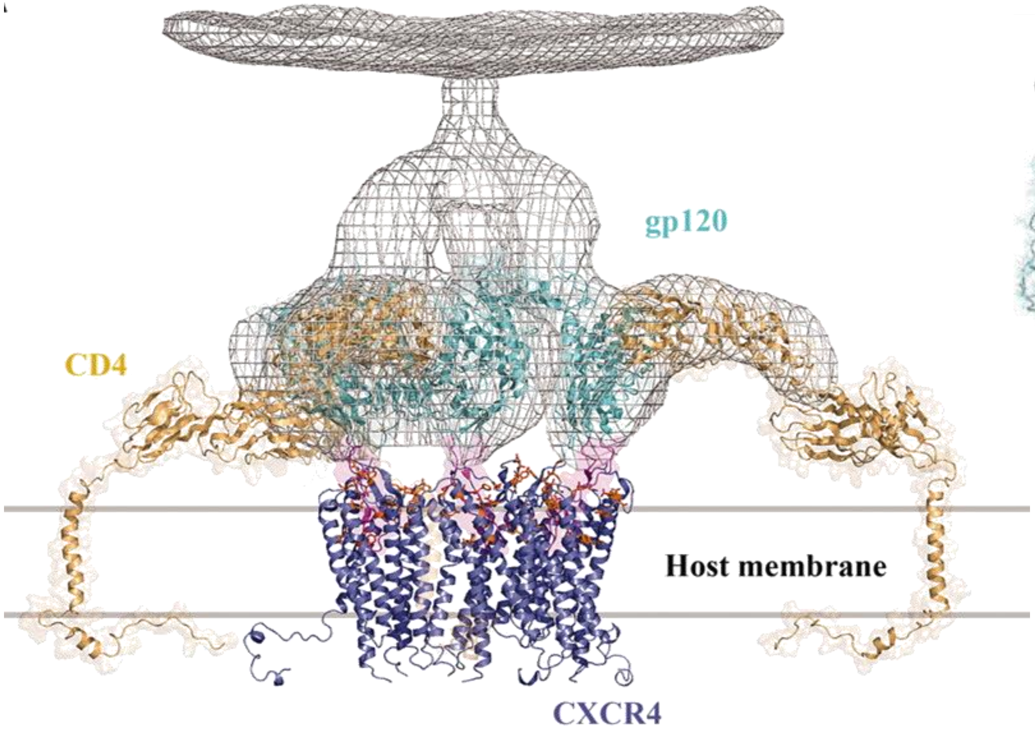
- *Find* transition and emission probabilities the maximize P(S | params, HMM)

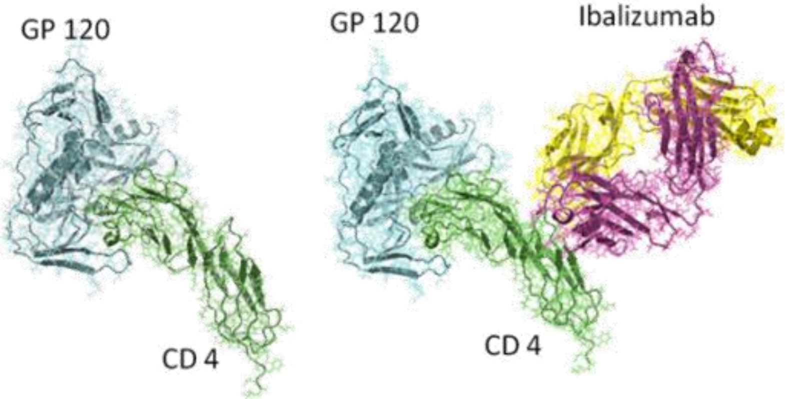
Discover and model family structure

Application: Modeling Protein Dynamics

Background

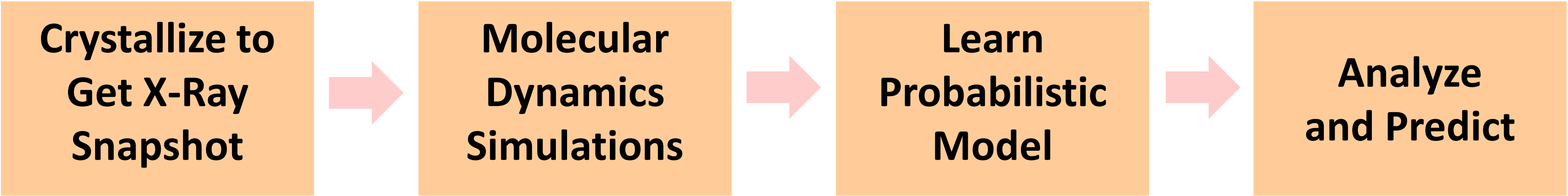
- **Proteins:** Molecular machines, composed of a sequences of Amino Acid sub-units





Background:

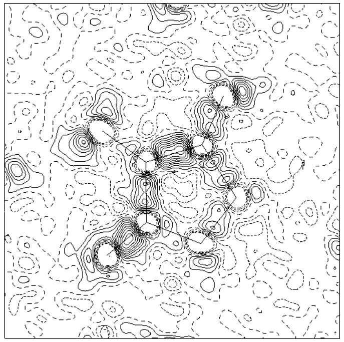
- Protein functional analysis pipeline

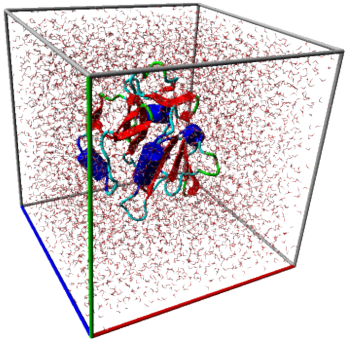

Crystallize to Get X-Ray Snapshot

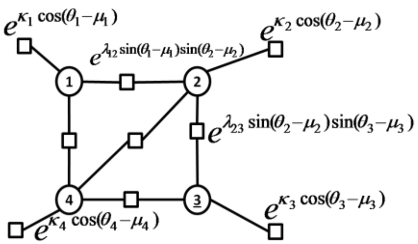
Molecular Dynamics Simulations

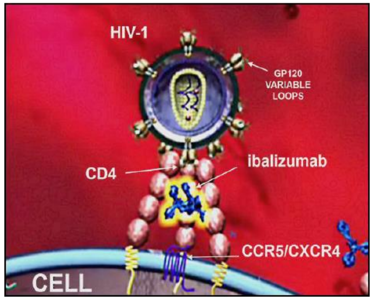
Learn Probabilistic Model

Analyze and Predict






$e^{\kappa_1 \cos(\theta_1-\mu_1)}$
$e^{\lambda_{12} \sin(\theta_1-\mu_1)\sin(\theta_2-\mu_2)}$
$e^{\kappa_2 \cos(\theta_2-\mu_2)}$
1
2
$e^{\lambda_{23} \sin(\theta_2-\mu_2)\sin(\theta_3-\mu_3)}$
4
3
$e^{\kappa_3 \cos(\theta_3-\mu_3)}$
$e^{\kappa_4 \cos(\theta_4-\mu_4)}$



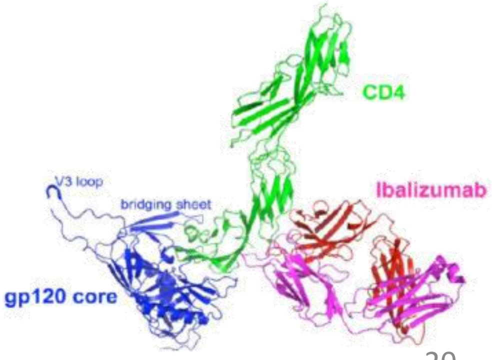
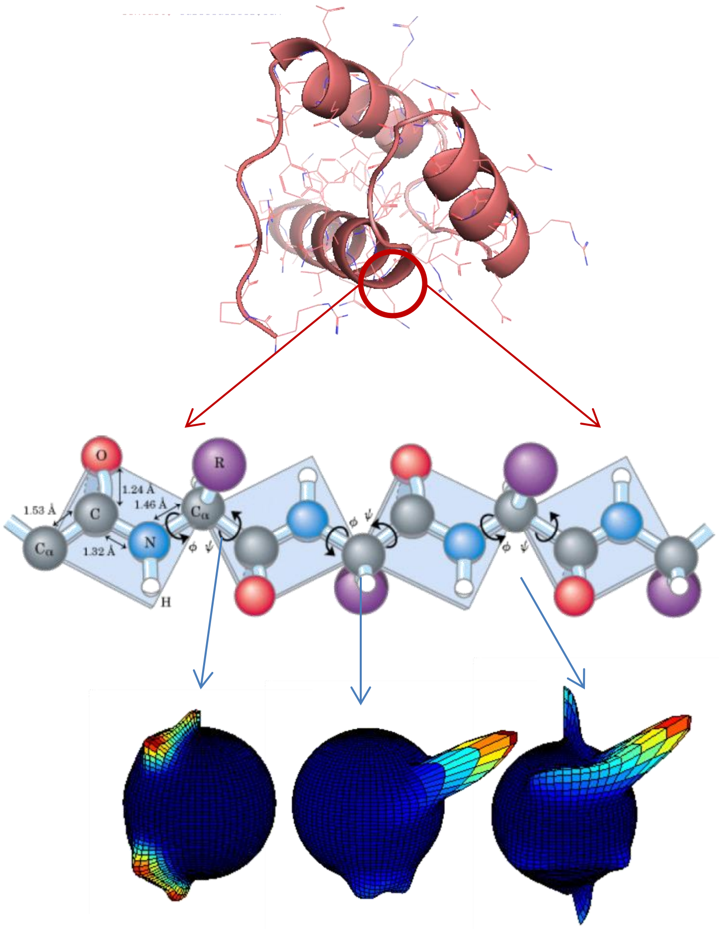


Image: H khanlou, et.al. "Durable Efficacy and Continued Safety of Ibalizumab in Treatment-Experienced Patients", Infectious Diseases Society of America (IDSA) October 2011

Modeling Protein Tertiary Structure



10 second Reminder! Probability Theory

- Sum rule

$$\boxed{p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y})}$$

- Product rule

$$\boxed{p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}$$

- From these we have Bayes' theorem

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}$$

  – with normalization

$$p(\mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$$

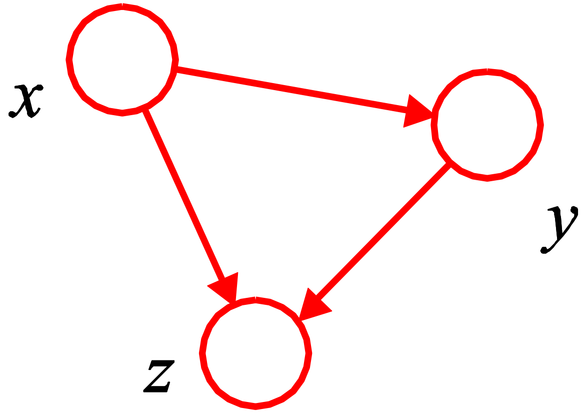
10 second Reminder(cont.)! Decomposition

- Consider an arbitrary joint distribution

$$p(x, y, z)$$

- By successive application of the product rule

$$\begin{aligned} p(x, y, z) &= p(x)p(y, z|x) \\ &= p(x)p(y|x)p(z|x, y) \end{aligned}$$

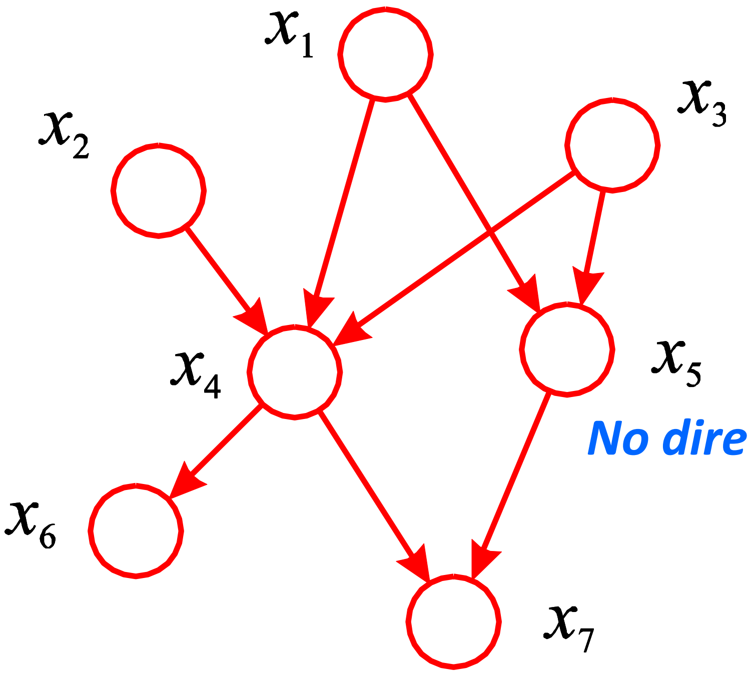


Directed Acyclic Graphs

- Joint distribution

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i|\mathrm{pa}_i)$$

where pa_i denotes the parents of i



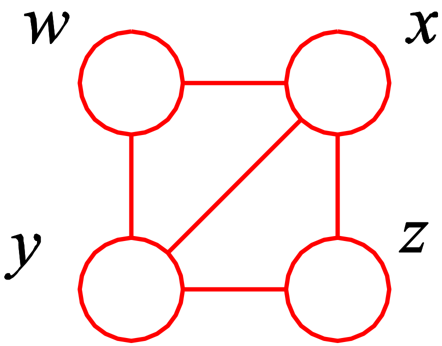
No directed cycles

Undirected Graphs

- Provided $p(\mathbf{x}) > 0$ then joint distribution is product of non-negative functions over the *cliques* of the graph

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C)$$

where $\psi_C(\mathbf{x}_C)$ are the *clique potentials,* and Z is a normalization constant



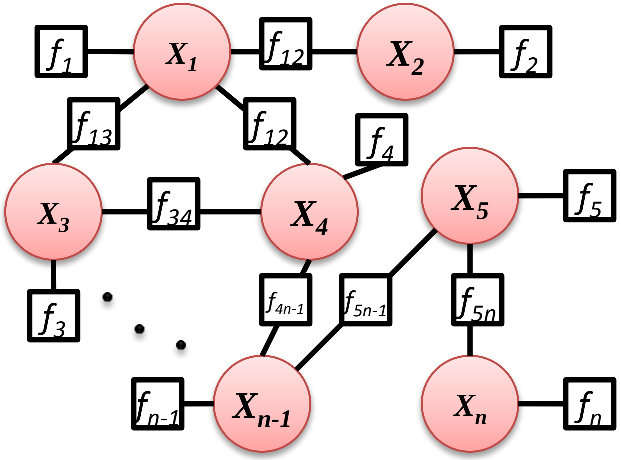
$$p(w, x, y, z) = \frac{1}{Z} \psi_A(w, x, y) \psi_B(x, y, z)$$

Undirected Graphical Models

- Pairwise Undirected graphical models (single and bivariate potentials only)

$$\text{Markov Random Field as A Factor Graph}$$

$$P(X) = \frac{\prod_{i=1}^{n} f_i(X_i) \prod_{\substack{eij=1 \\ i \neq j}} f_{ij}(X_i, X_j)}{\int \prod_{i=1}^{n} f_i(X_i) \prod_{\substack{eij=1 \\ i \neq j}} f_{ij}(X_i, X_j)\, dX_1..dX_n}$$

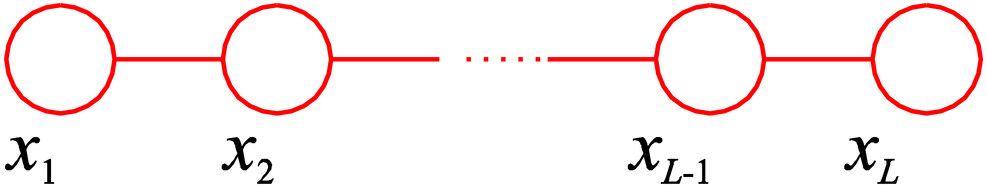


Question:

- Each potential has some parameters. How to estimate them from training data?
  - – Could do gradient descent on the likelihood of the data, (if we knew z)
  - – Often iterative process
- How to compute z?
  - – Belief propagation (next slides)

Message Passing

- Example



- Find marginal for a particular node

$$p(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_L} p(x_1, \ldots, x_L)$$

  - for M-state nodes, cost is $O(M^L)$
  - exponential in length of chain
  - but, we can exploit the graphical structure (conditional independences)

Message Passing

- Joint distribution

$$p(x_1, \ldots, x_L) = \frac{1}{Z}\psi(x_1, x_2) \ldots \psi(x_{L-1}, x_L)$$

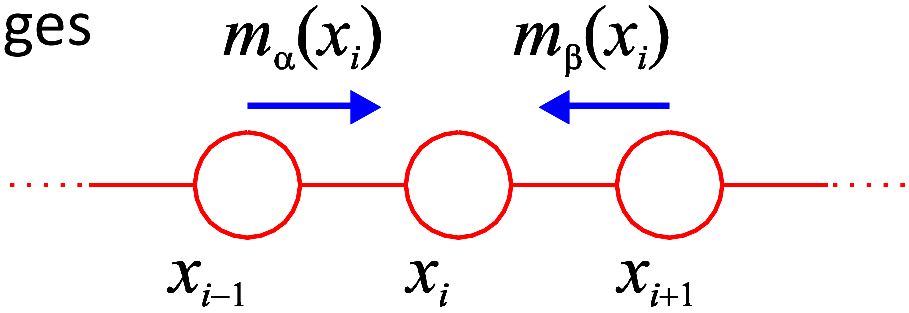
- Exchange sums and products

$$\begin{aligned} p(x_i) \quad = \quad & \frac{1}{Z} \overbrace{\cdots \sum_{x_2} \psi(x_2, x_3) \left[\sum_{x_1} \psi(x_1, x_2) \right]}^{m_\alpha(x_i)} \\ & \underbrace{\cdots \sum_{x_{L-1}} \psi(x_{L-2}, x_{L-1}) \left[\sum_{x_L} \psi(x_{L-1}, x_L) \right]}_{m_\beta(x_i)} \end{aligned}$$

Message Passing

- Express as product of messages

$$p(x_i) = \frac{1}{Z} m_\alpha(x_i) m_\beta(x_i)$$



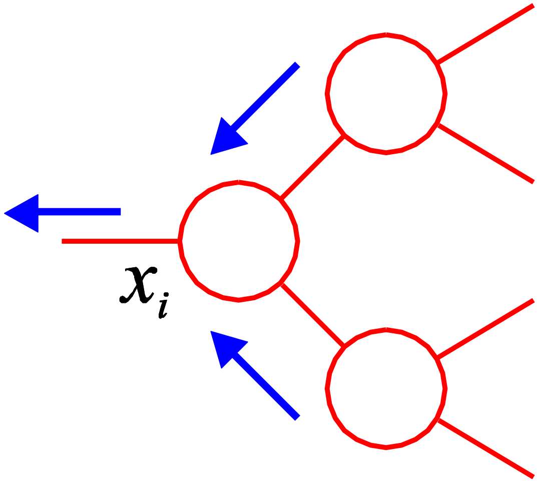
- Recursive evaluation of messages

$$\begin{aligned} m_\alpha(x_i) &= \sum_{x_{i-1}} \psi(x_{i-1}, x_i) m_\alpha(x_{i-1}) \\ m_\beta(x_i) &= \sum_{x_{i+1}} \psi(x_i, x_{i+1}) m_\beta(x_{i+1}) \end{aligned}$$

- Find Z by normalizing $p(x_i)$

Belief Propagation

- Extension to general tree-structured graphs
- At each node:
  - form product of *incoming* messages and local evidence
  - marginalize to give *outgoing* message
  - one message in each direction across every link



- No convergence guaranteed if there are loops!

Inference and Learning

- Data set

$$D = \{\mathbf{x}_n\}, \qquad n = 1, \ldots, N$$

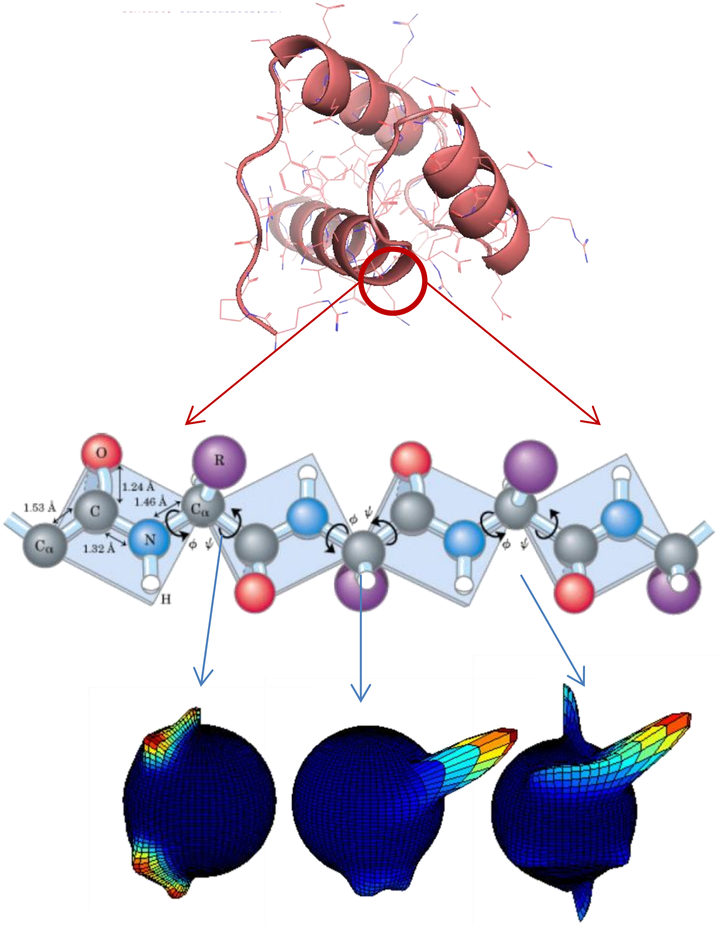
- Likelihood function (independent observations)

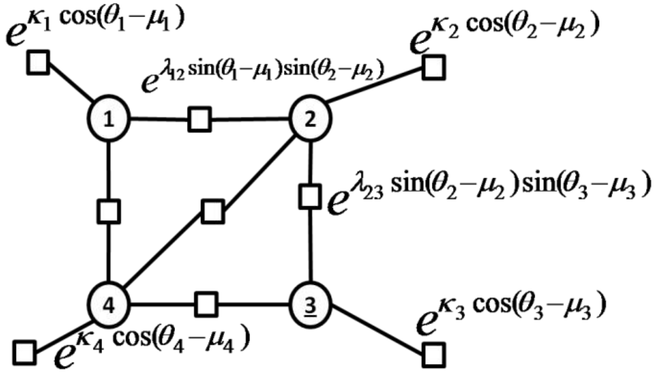
$$L(\boldsymbol{\theta}) = p(D|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta})$$

- Maximize (log) likelihood

$$\boldsymbol{\theta}_{\mathsf{ML}} = \arg\max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta})$$

Modeling Protein Tertiary Structure

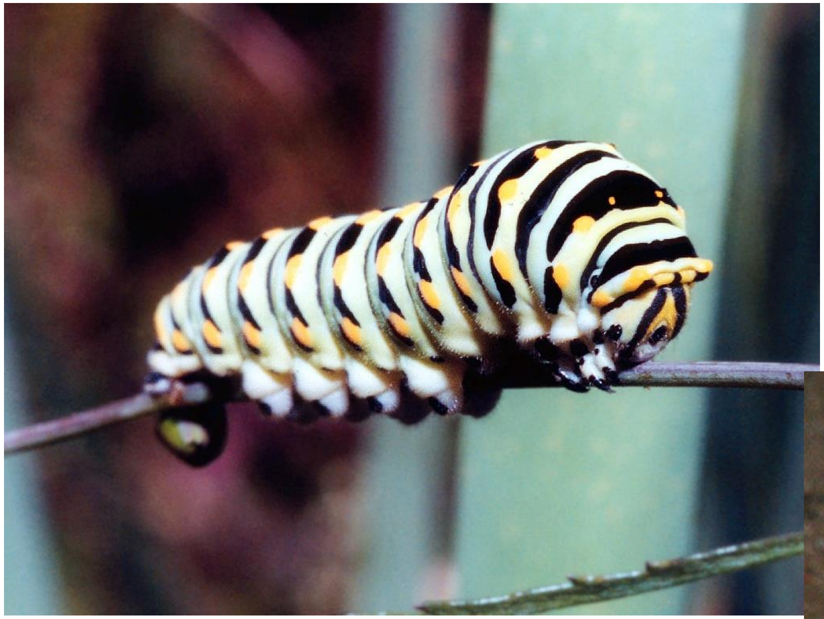



$e^{\kappa_1 \cos(\theta_1 - \mu_1)}$
$e^{\kappa_2 \cos(\theta_2 - \mu_2)}$
$e^{\lambda_{12} \sin(\theta_1 - \mu_1)\sin(\theta_2 - \mu_2)}$
1
2
$e^{\lambda_{23} \sin(\theta_2 - \mu_2)\sin(\theta_3 - \mu_3)}$
4
3
$e^{\kappa_3 \cos(\theta_3 - \mu_3)}$
$e^{\kappa_4 \cos(\theta_4 - \mu_4)}$

- Optimize Pseudo-likelihood of training data, to estimate parameters

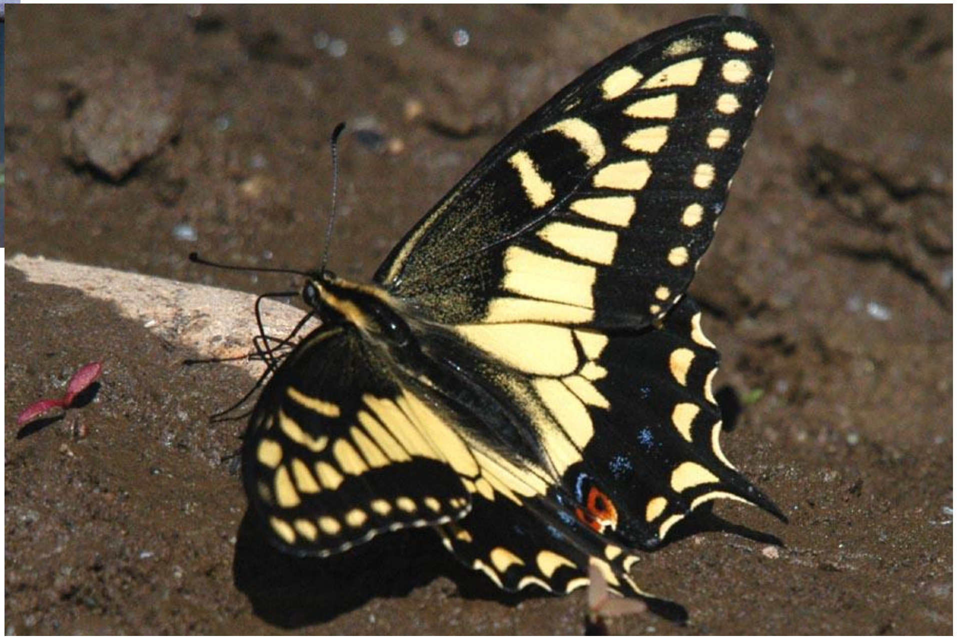
Application: Microarray Gene Expression Analysis

The dramatic consequences of gene regulation in biology

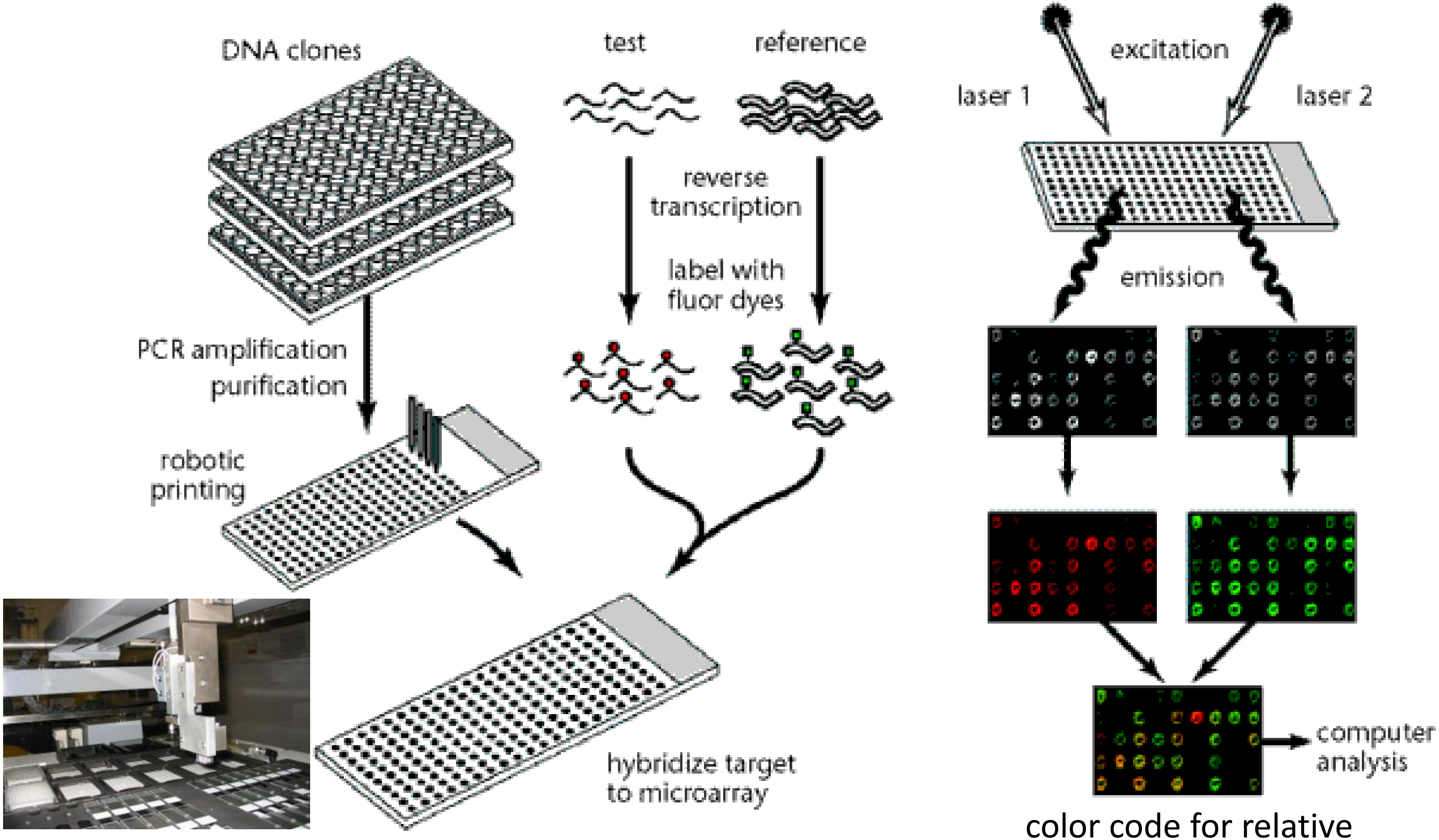


Anise swallowtail, *Papilio zelicaon*

Same genome →
Different tissues
- Different physiology
- Different proteome
- Different expression pattern



cDNA microarray schema

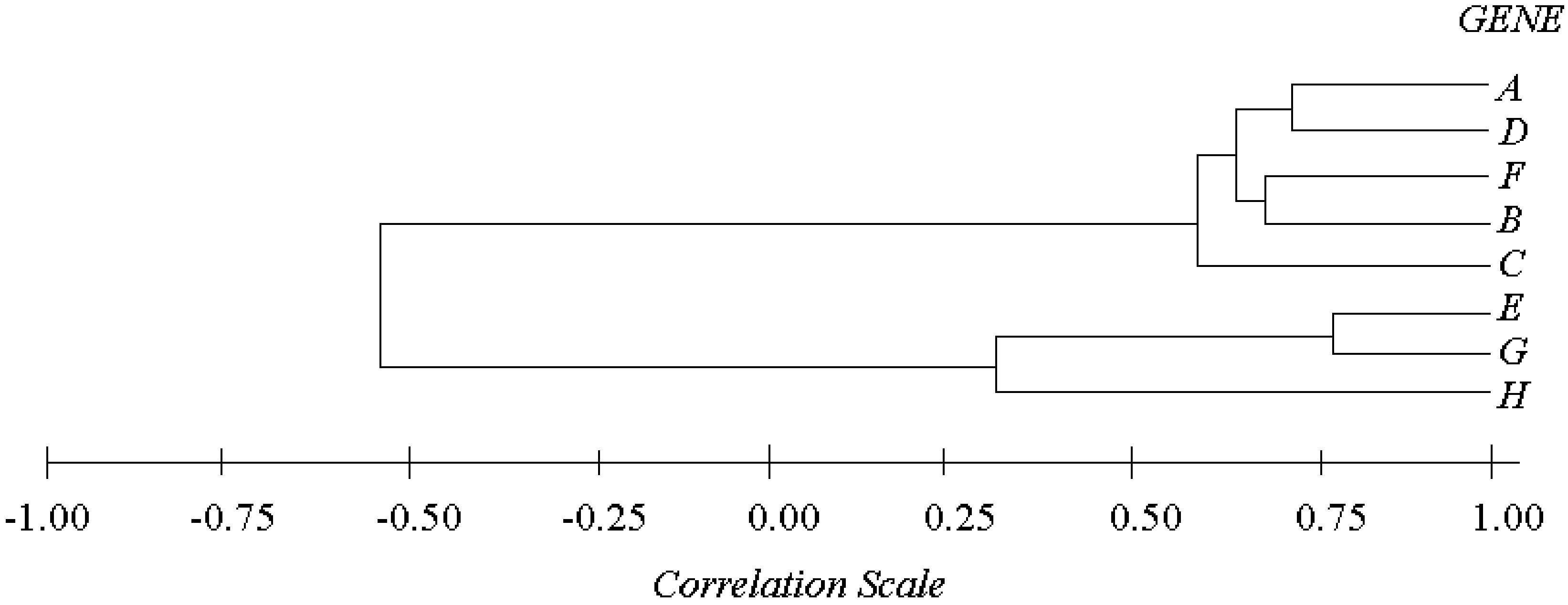


From Duggan *et al. Nature Genetics* **21**, 10 – 14 (1999)

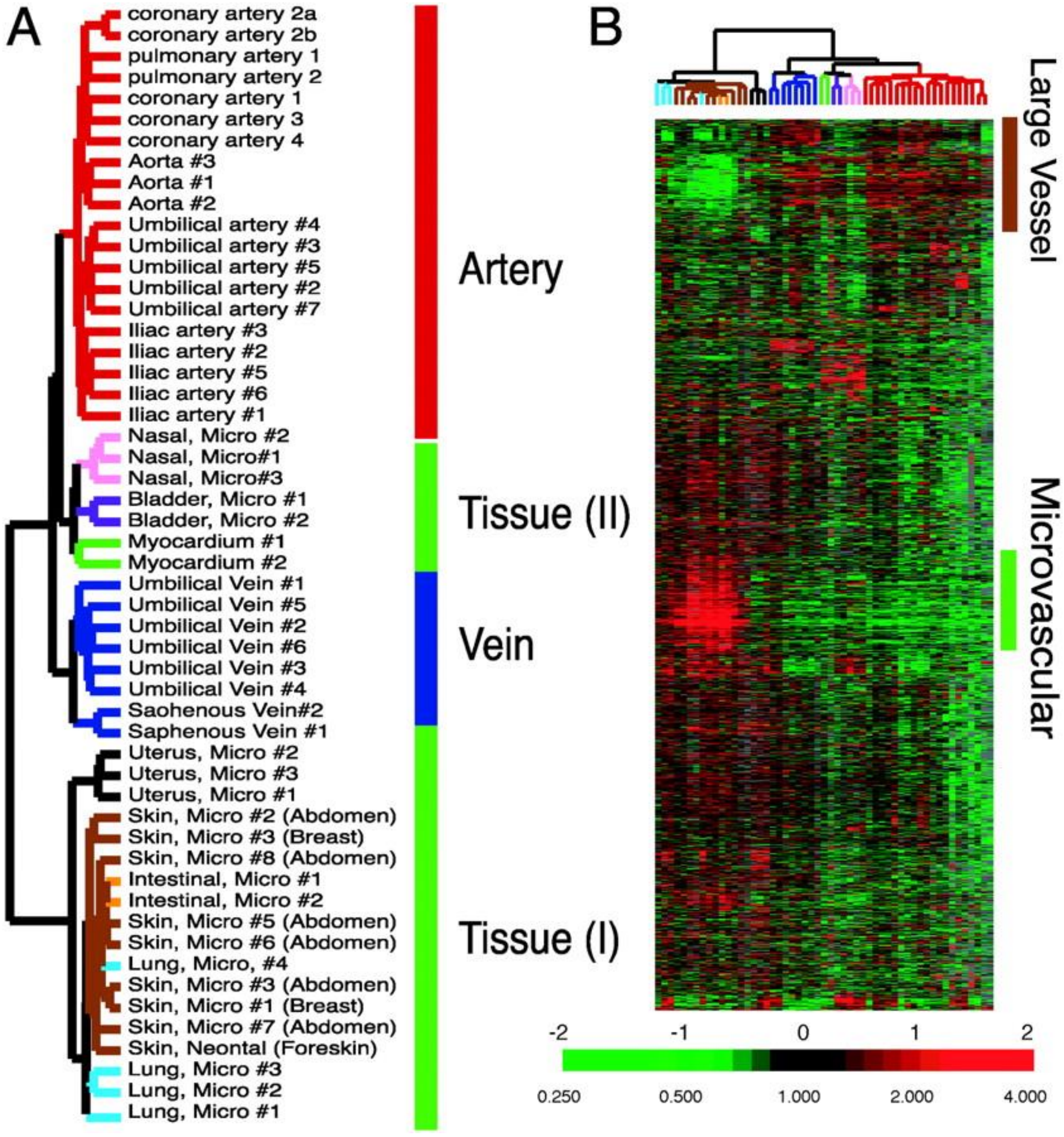
color code for relative expression

Hierarchical clustering

- Combine most similar genes into agglomerative clusters, build tree of genes
- Do the same procedure along the second dimension to cluster samples
- Display as a heatmap



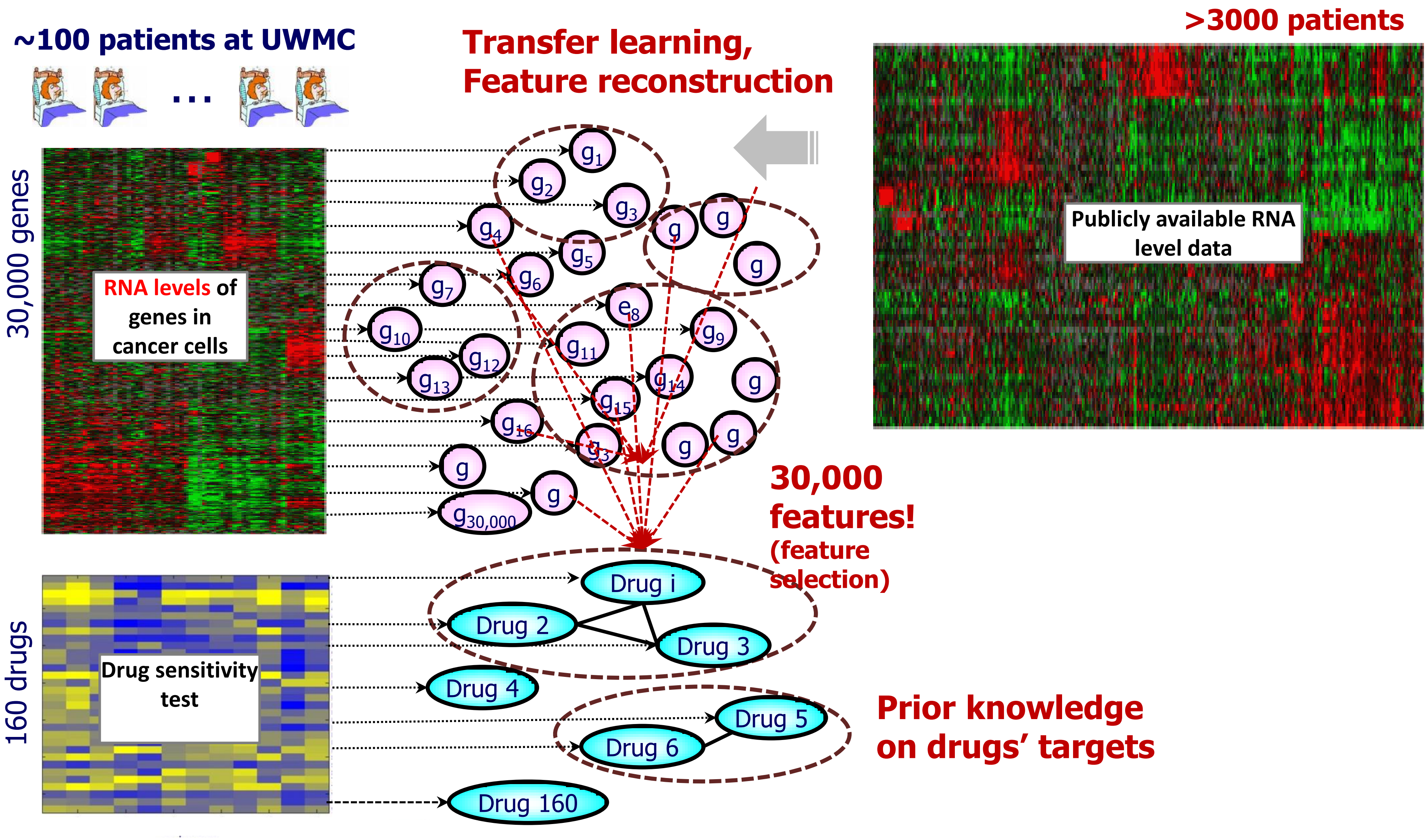
Hierarchical clustering results



Chi et al., PNAS | **September 16, 2003** | vol. 100 | no. 19 | **10623-10628**

"Endothelial cell diversity revealed by global expression profiling"

Personalized cancer treatment



~100 patients at UWMC

Transfer learning, Feature reconstruction

>3000 patients

Publicly available RNA level data

RNA levels of genes in cancer cells

30,000 features! (feature selection)

Drug sensitivity test

Prior knowledge on drugs' targets

Other applications

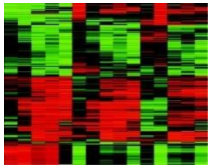
- Predicting phenotype (symptoms) given:

A few histologic features

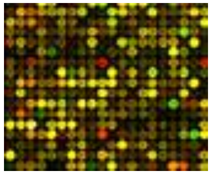
Epigenetics (Methylation)

...ACGTAGCTAGCT
AGCTAGCTGATGC
TAGCTACGTGCT...

DNA sequence



RNA levels of genes



Protein levels of genes

  - Predictive Models Can be:
    - Generative (i.e. Bayesian Network)
    - Discriminative (i.e. Regression, SVM, KNN)

Many more exciting research to come! ☺