Machine Learning Approaches to Biological Sequence Analysis

Byoung-Tak Zhang

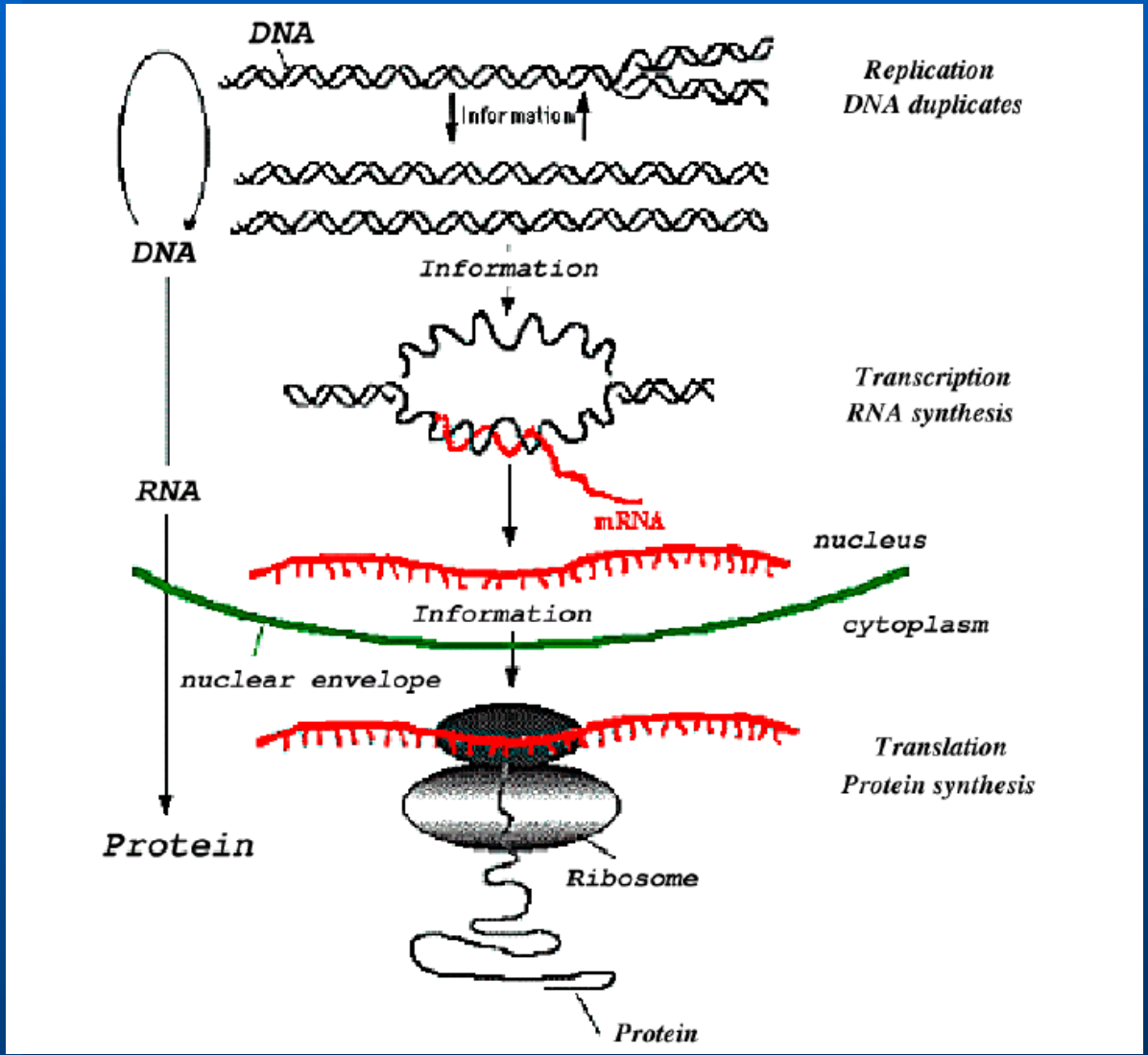
Center for Bioinformation Technology (CBIT) &
Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University

http://cbit.snu.ac.kr/
http://bi.snu.ac.kr/

Talk Outline

- Bioinformatics
- Machine Learning
- Gene Finding
- Promoter Prediction
- Protein Structure Prediction
- Summary

Molecular Biology: Central Dogma



DNA and Protein Sequences

DNA (Nucleotide) Sequence

SQ sequence 1344 BP; 291 A; C; 401 G; 278 T; 0 other

```
aacctgcgga aggatcatta  gcgggcccgc cgcttgtcgg cgcttgtcgg ccgccggggg
ccgagtgcgg gtcctttggg  ccgccggggg ggcgcctctg ccccccgggc ccgtgcccgc
cccaacctcc catccgtgtc  ccccccgggc ccgtgcccgc cggagacccc aacacgaaca
tattgtaccc tgttgcttcg  aacctgcgga aggatcatta ctgtctgaaa gcgtgcagtc
gcgggcccgc cgcttgtcgg  ccgagtgcgg gtcctttggg tgagttgatt gaatgcaatc
ccgccggggg ggcgcctctg  cccaacctcc catccgtgtc agttaaaact ttcaacaatg
ccccccgggc ccgtgcccgc  tattgtaccc tgttgcttcg gatctcttgg aacctgcgga
cggagacccc aacacgaaca  gcgggcccgc cgcttgtcgg ccgagtgcgg gtcctttggg
ctgtctgaaa gcgtgcagtc  agttaaaact ttcaacaatg cccaacctcc catccgtgtc
tgagttgatt gaatgcaatc  gatctcttgg ttccggctgc tattgtaccc tgttgcttcg
agttaaaact ttcaacaatg  tattgtaccc tgttgcttcg gcgggcccgc cgcttgtcgg
gatctcttgg ttccggctgc  gcgggcccgc cgcttgtcgg ccgccggggg ggcgcctctg
tattgtaccc tgttgcttcg  ccgccggggg ggcgcctctg agttaaaact ttcaacaatg
gcgggcccgc cgcttgtcgg  ccccccgggc ccgtgcccgc gatctcttgg ttccggctgc
ccgccggggg ggcgcctctg  cggagacccc tgttgcttcg tattgtaccc tgttgcttcg
ccccccgggc ccgtgcccgc  gcgggcccgc cgcttgtcgg gcgggcccgc cgcttgtcgg
cggagacccc tgttgcttcg  ccgccggggg cggagacccc ccgccggggg ggcgcctctg
gcgggcccgc cgcttgtcgg  gcgggcccgc cgcttgtcgg ccccccgggc ccgtgcccgc
ccgccggggg cggagacccc  ccgccggggg ggcgcctctg cggagacccc tgttgcttcg
```

Protein (Amino Acid) Sequence

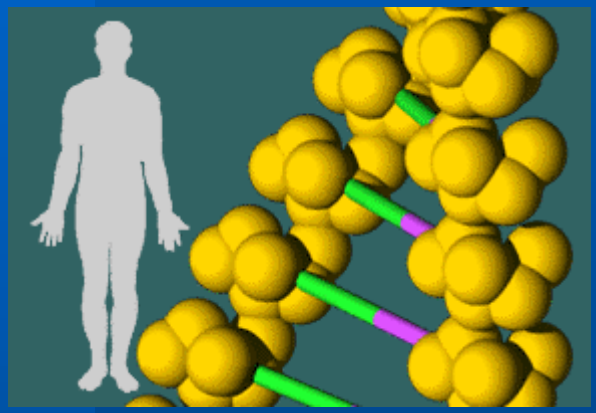
CG2B_MARGL Length: 388 April 2, 1997 14:55 Type: P Check: 9613 .. 1

```
MLNGENVDSR IMGKVATRAS SKGVKSTLGT RGALENISNV
ARNNLQAGAK KELVKAKRGM TKSKATSSLQ SVMGLNVEPM
EKAKPQSPEP MDMSEINSAL EAFSQNLLEG VEDIDKNDFD
NPQLCSEFVN DIYQYMRKLE REFKVRTDYM TIQEITERMR
SILIDWLVQV HLRFHLLQET LFLTIQILDR YLEVQPVSKN
KLQLVGVTSM LIAAKYEEMY PPEIGDFVYI TDNAYTKAQI
RSMECNILRR LDFSLGKPLC IHFLRRNSKA GGVDGQKHTM
AKYLMELTLP EYAFVPYDPS EIAAAALCLS SKILEPDMEW
GTTLVHYSAY SEDHLMPIVQ KMALVLKNAP TAKFQAVRKK
YSSAKFMNVS TISALTSSTV MDLADQMC
```

Some Facts

- 10^{14} cells in the human body.
- 3×10^{9} letters in the DNA code in every cell in your body.
- DNA differs between humans by 0.2% (1 in 500 bases).
- Human DNA is 98% identical to that of chimpanzees.
- 97% of DNA in the human genome has no known function.

Human Genome Project



Goals

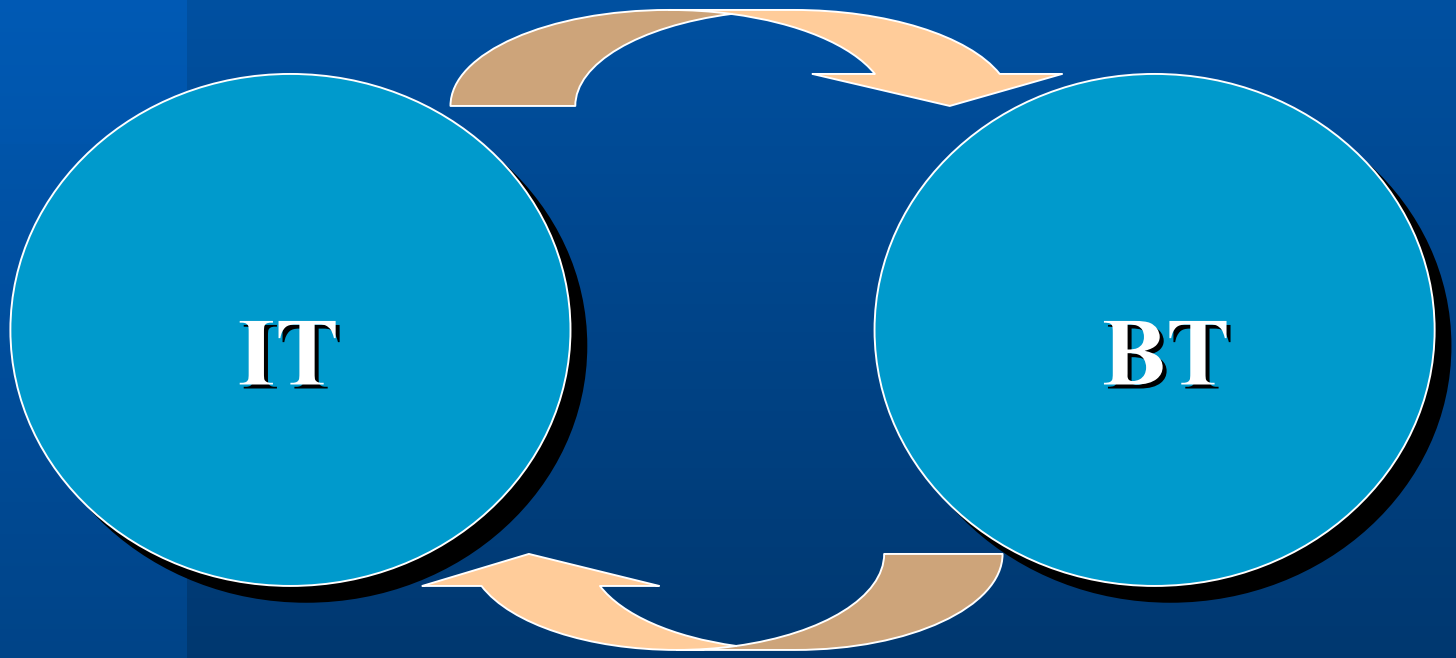
- Identify the approximate 40,000 genes in human DNA
- Determine the sequences of the 3 billion bases that make up human DNA
- Store this information in database
- Develop tools for data analysis
- Address the ethical, legal and social issues that arise from genome research

Genome Health Implications

- A New Disease Encyclopedia
- New Genetic Fingerprints
- New Diagnostics
- New Treatments

Bioinformation Technology (BIT)

Bioinformatics (*in silico* Biology)



Biocomputing (*in vivo* Computing)

Bioinformatics

What is Bioinformatics?

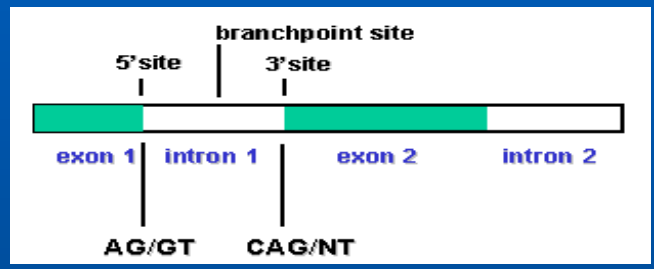
- ***Bio*** – molecular biology
- ***Informatics*** – computer science

- Bioinformatics – solving problems arising from biology using methodology from computer science.

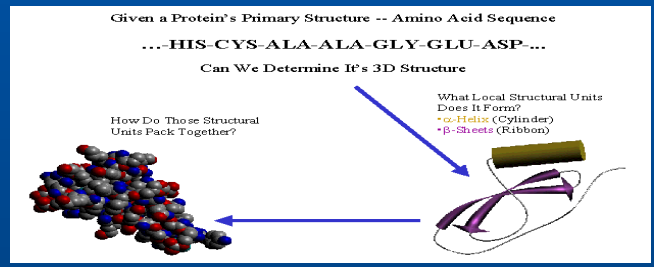
- Bioinformatics vs. Computational Biology
- Bioinformatik (in German): Biology-based computer science as well as bioinformatics (in English)

Topics in Bioinformatics

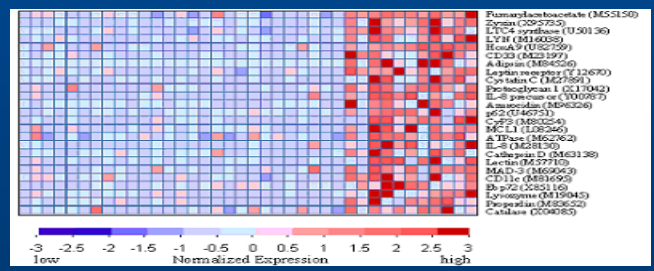
Sequence analysis
- Sequence alignment
- Structure and function prediction
- Gene finding



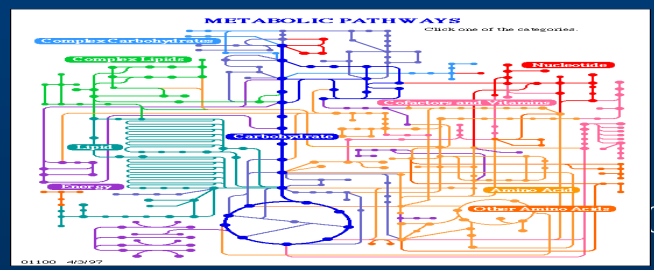
Structure analysis
- Protein structure comparison
- Protein structure prediction
- RNA structure modeling



Expression analysis
- Gene expression analysis
- Gene clustering

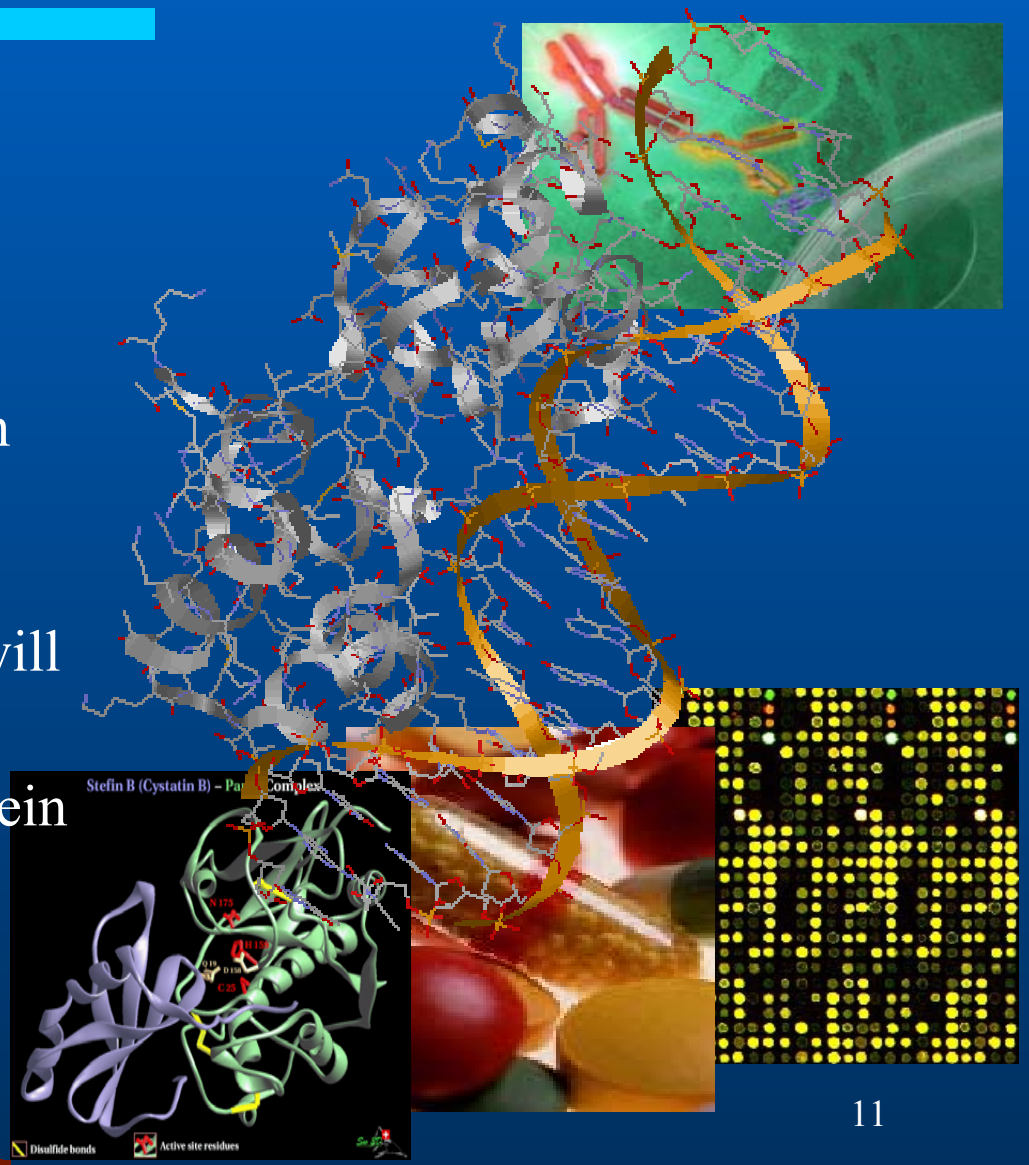


Pathway analysis
- Metabolic pathway
- Regulatory networks



Extension of Bioinformatics Concept

- Genomics
  - Functional genomics
  - Structural genomics
- Proteomics: large scale analysis of the proteins of an organism
- Pharmacogenomics: developing new drugs that will target a particular disease
- Microarray: DNA chip, protein chip

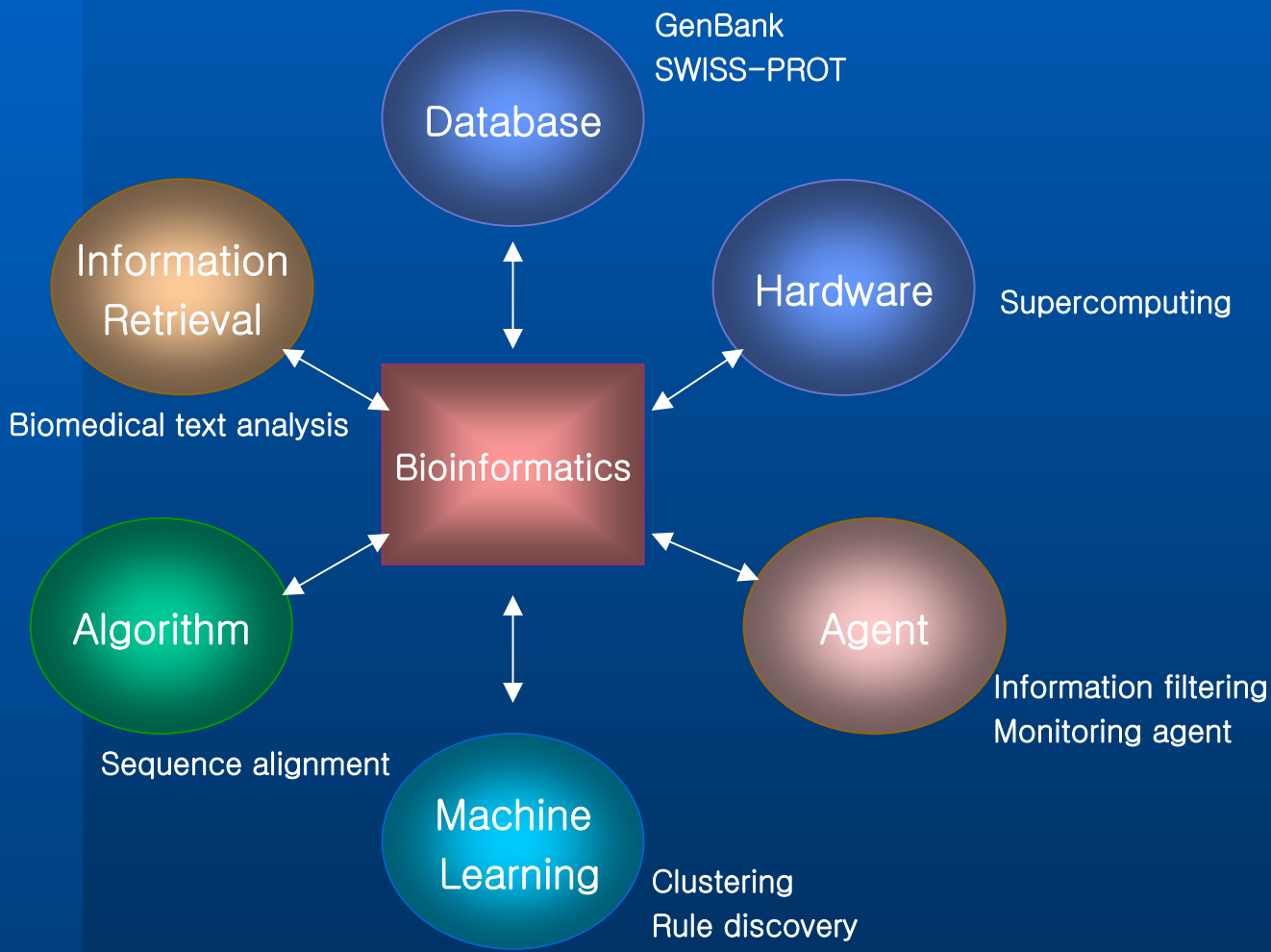


Applications of Bioinformatics

- Drug design
- Identification of genetic risk factors
- Gene therapy
- Genetic modification of food crops and animals
- Biological warfare, crime etc.

- Personalized Medicine
- E-Doctor

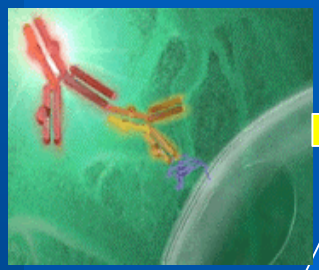
Bioinformatics as Information Technology



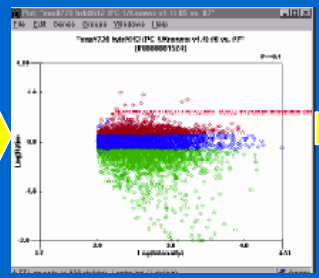
Background of Bioinformatics

- Biological information infra
  - Biological information management systems
  - Analysis software tools
  - Communication networks for biological research
- Massive biological databases
  - DNA/RNA sequences
  - Protein sequences
  - Genetic map linkage data
  - Biochemical reactions and pathways
- Need to integrate these resources to model biological reality and exploit the biological knowledge that is being gathered.

Areas and Workflow of Bioinformatics



AGCTAGTTCAGTACA
TGGATCCATAAGGTA
CTCAGTCATTACTGC
AGGTCACTTACGATA
TCAGTCGATCACTAG
CTGACTTACGAGAGT



Microarray (Biochip)

Structural Genomics	Functional Genomics	Proteomics	Pharmaco-genomics

Infrastructure of Bioinformatics

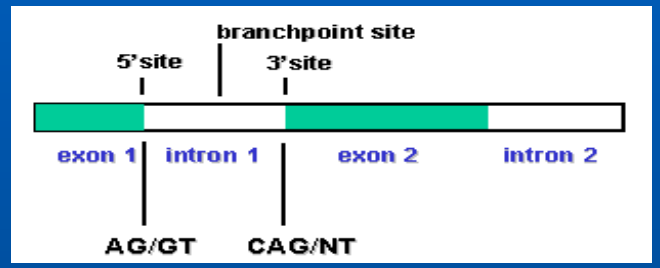
Topics in Bioinformatics

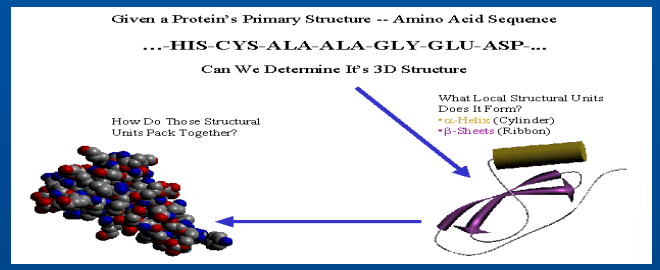
Sequence analysis
- Sequence alignment
- Structure and function prediction
- Gene finding

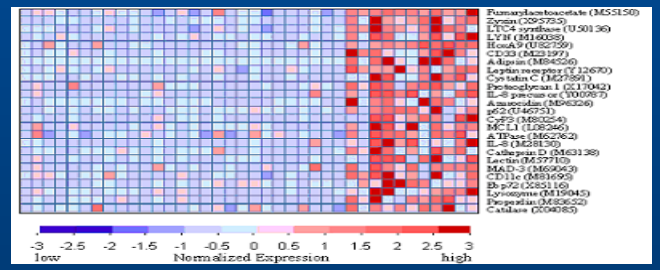
Structure analysis
- Protein structure comparison
- Protein structure prediction
- RNA structure modeling

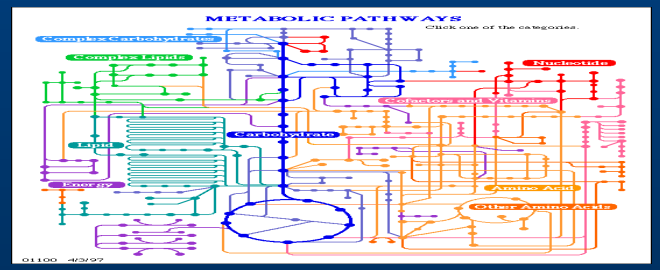
Expression analysis
- Gene expression analysis
- Gene clustering

Pathway analysis
- Metabolic pathway
- Regulatory networks









Machine Learning

Machine Learning

- Supervised Learning
  - Estimate an unknown mapping from known input- output pairs
  - Learn $f_{\mathbf{w}}$ from training set $D=\{(\mathbf{x},y)\}$ s.t. $f_{\mathbf{w}}(\mathbf{x}) = y = f(\mathbf{x})$
  - Classification: y is discrete
  - Regression: y is continuous
- Unsupervised Learning
  - Only input values are provided
  - Learn $f_{\mathbf{w}}$ from $D=\{(\mathbf{x})\}$ s.t. $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{x}$
  - Compression
  - Clustering
- Reinforcement Learning

Why Machine Learning?

- Recent progress in algorithms and theory
- Growing flood of online data
- Computational power is available
- Budding industry

Three niches for machine learning

- Data mining: using historical data to improve decisions
  - Medical records -> medical knowledge
- Software applications we can't program by hand
  - Autonomous driving
  - Speech recognition
- Self customizing programs
  - Newsreader that learns user interests

Methods in Machine Learning (1/2)

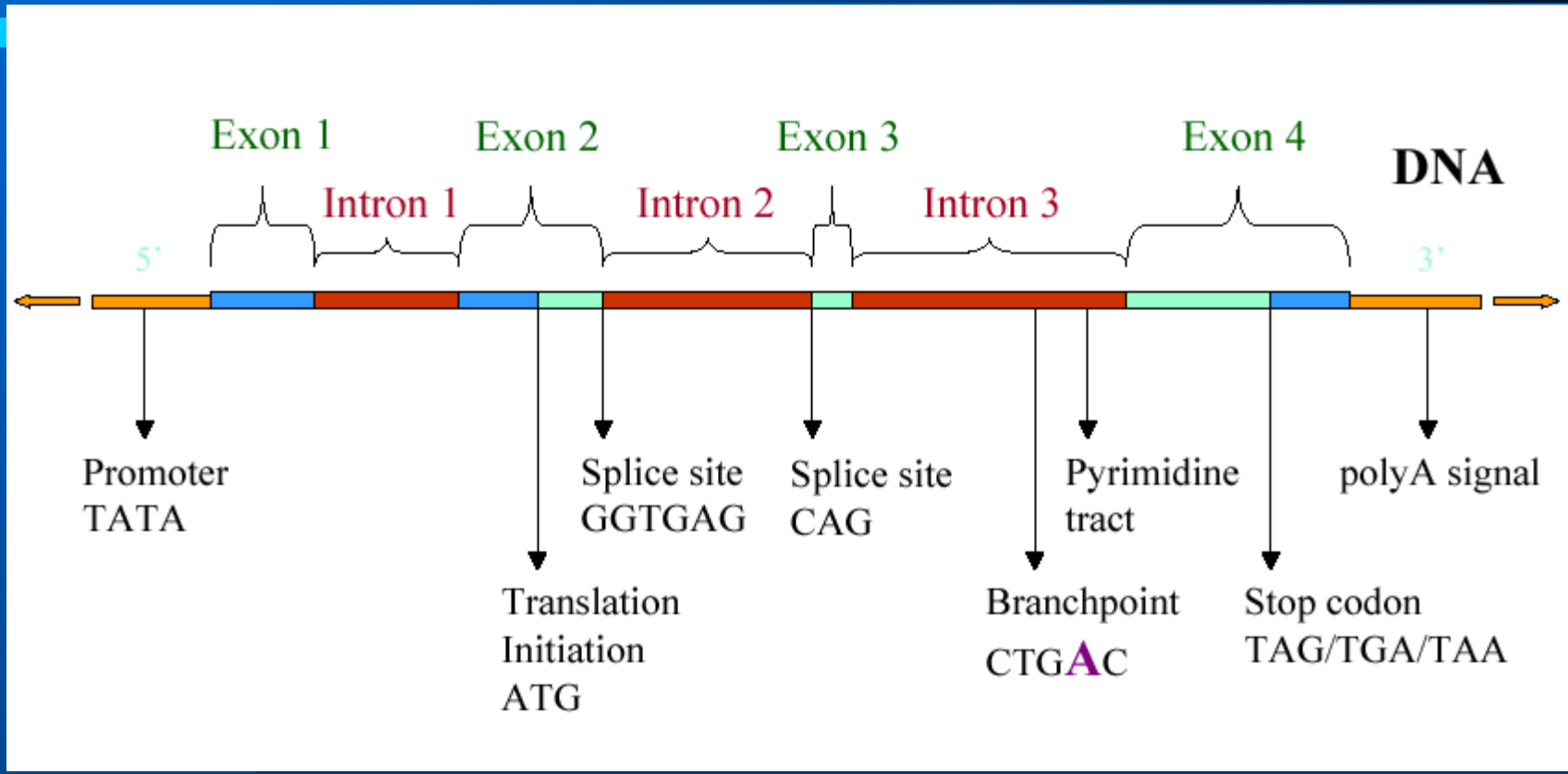
- Symbolic Learning
  - Version Space Learning
  - Case-Based Learning
- Neural Learning
  - Multilayer Perceptrons (MLPs)
  - Self-Organizing Maps (SOMs)
  - Support Vector Machines (SVMs)
- Evolutionary Learning
  - Evolution Strategies
  - Evolutionary Programming
  - Genetic Algorithms
  - Genetic Programming

Methods in Machine Learning (2/2)

- Probabilistic Learning
  - Bayesian Networks (BNs)
  - Helmholtz Machines (HMs)
  - Latent Variable Models (LVMs)
  - Generative Topographic Mapping (GTM)
- Other Machine Learning Methods
  - Decision Trees (DTs)
  - Reinforcement Learning (RL)
  - Boosting Algorithms
  - Mixture of Experts (ME)
  - Independent Component Analysis (ICA)

Gene Finding

DNA Structure



Gene-Finding Strategies

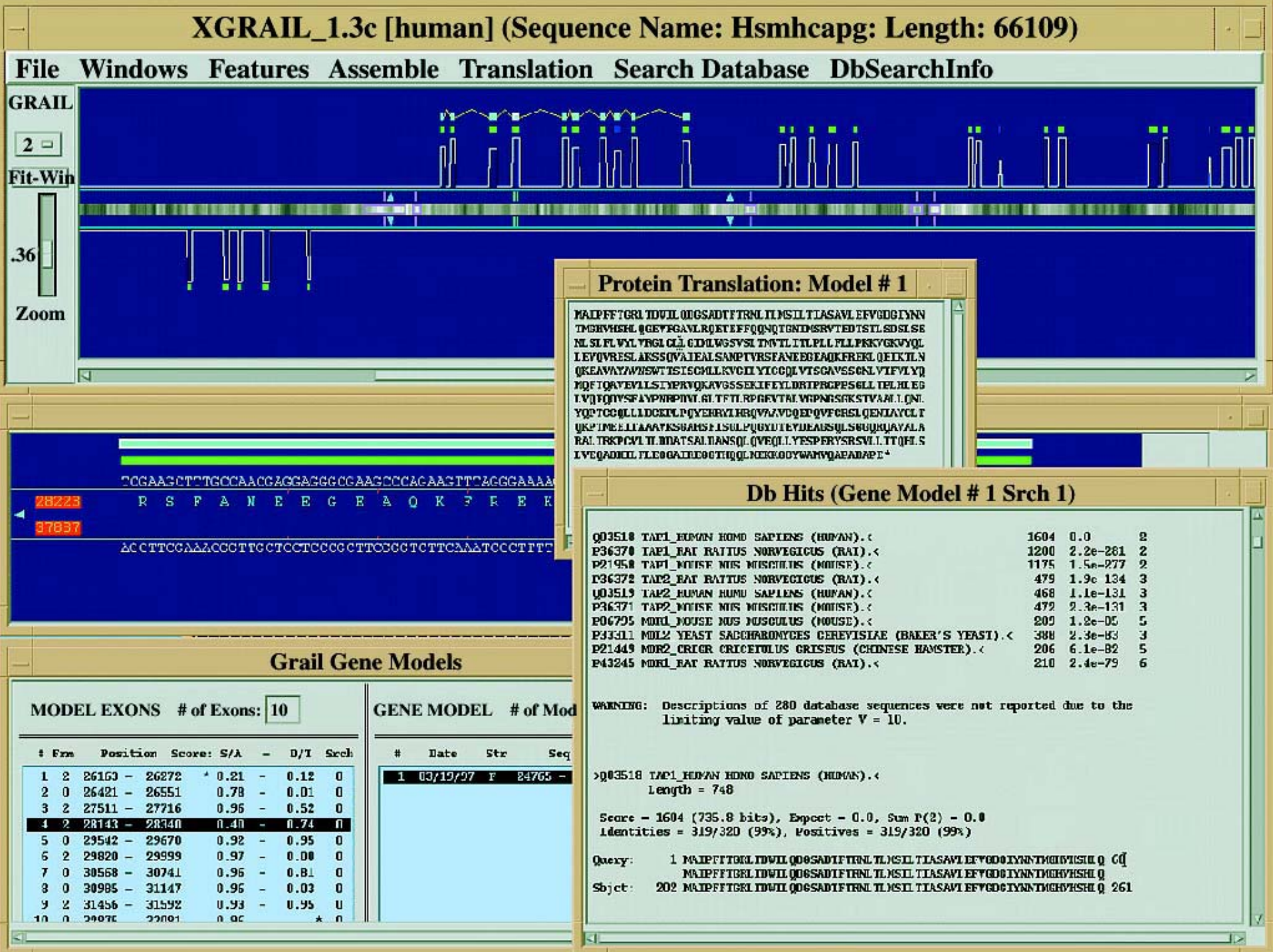
- **Gene Finding**
  - Goal: partitioning the genome into genes
  - Sequences represent coding or noncoding regions
  - Gene identification is a complex mathematical problem
- **Content-based Methods**
  - Rely on the overall, bulk properties of sequence
  - Particular codons, properties of repeats, compositional complexity of the sequences
- **Site-based Methods**
  - Presence or absence of a specific sequences, patterns, or consensus.
- **Comparative Methods**
  - Make determinations based on sequence homology

Gene Finding Programs

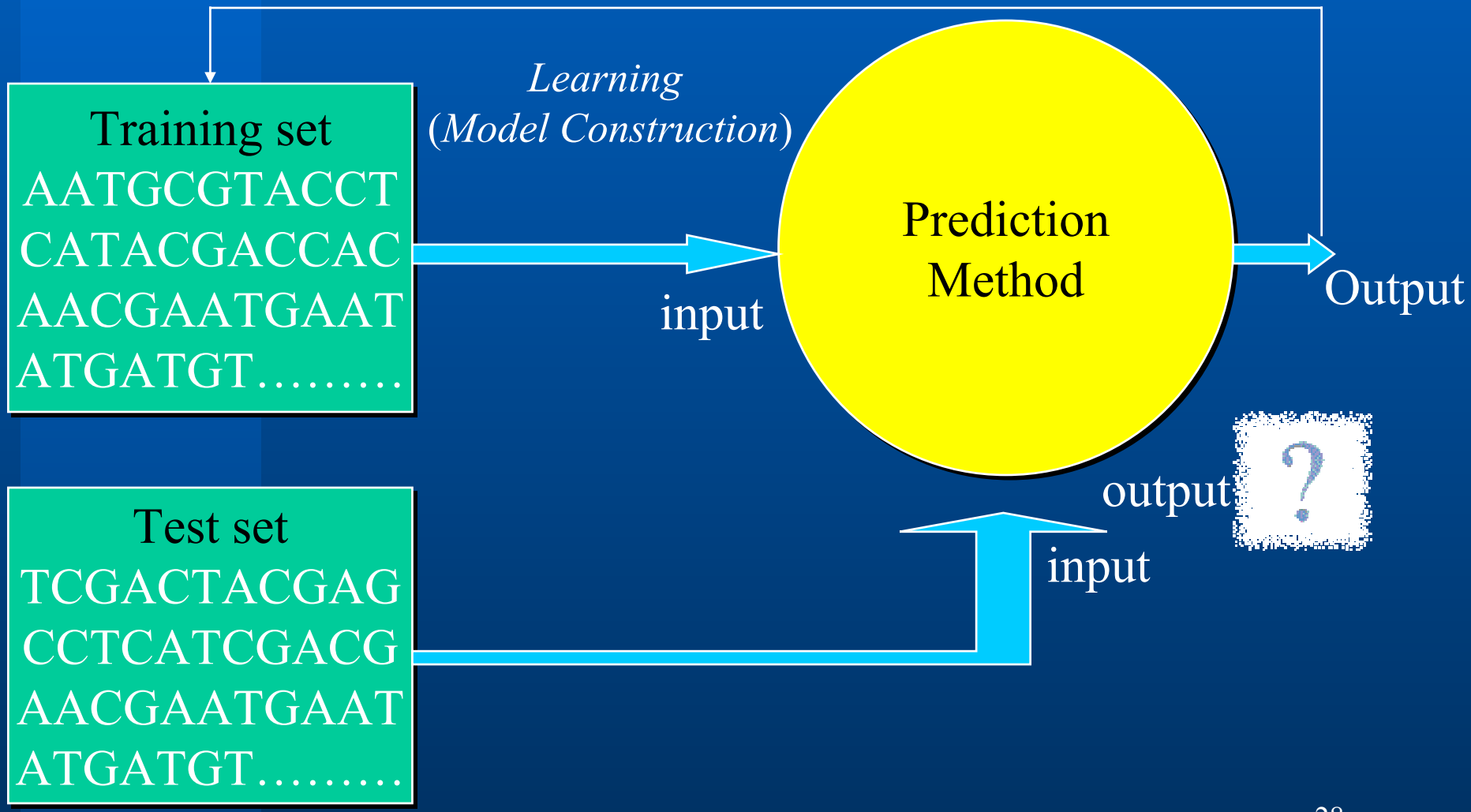
Name	Methods	Organism
ER	Discriminant Analysis	Human, Arabidopsis
GENSCAN (seems the most accurate)	Semi Markov Model	vertebrate, caenorhabditis, arabidopsis, maize
GRAIL	Neural Network	human, mouse, arabidopsis, drosophila, E.coli
GenLang	Definite Clause Grammer	Vertebrate, Drosophila, Dicot
GenView	Linear combination	Human, Mouse, Diptera
GeneFinder(FGENEH,etc.)	LDA	Human, E.coli, Drosophila, Plant, Nematode, Yeast
GeneID	Perceptron,rules	Vertebrate
GeneMark	5th-Markov	Almost all model organism
GeneParser	Neural networks	Human
Genie	GHMM	Human (vertebrate)
Glimmer	Interpolated Markov models (IMMs)	microbial
MORGAN	Decision Tree	vertebrate
MZEF	Quadratic Discriminant Analysis	Human, mouse, Arabidopsis, Pombe
NetPlantGene	Combined Neural Networks	A. thaliana
OC1	Decision tree	Human
PROCRUSTES	Spliced alignment	vertebrate
Sorfind	Rule base	Human
VEIL	HMM	vertebrate
Hogehoge	Wonderful method	extraterrestrial

GRAIL (Gene Recognition and Analysis Internet Link)

- Combine information from several exon-prediction algorithms
  - Each algorithm is designed to recognize a particular sequence property.
  - Use a neural network to provide more powerful exon recognition capabilities.
- Measure coding potentials
  - Determine the likelihood that a DNA segment is an exon
  - Frame-dependent 6 tuple preference model
    - Measure the strength of a potential splice junction or a translation start.
  - 5^{th} order non-homogeneous Markov chain model


XGRAIL_1.3c [human] (Sequence Name: Hsmhcapg: Length: 66109)
File Windows Features Assemble Translation Search Database DbSearchInfo
GRAIL
2
Fit-Win
.36
Zoom
Protein Translation: Model # 1
Db Hits (Gene Model # 1 Srch 1)
Q03518 TAP1_HUMAN HOMO SAPIENS (HUMAN).< 1604 0.0 2
P36370 TAP1_RAT RATTUS NORVEGICUS (RAT).< 1200 2.2e-261 2
P21958 TAP1_MOUSE MUS MUSCULUS (MOUSE).< 1175 1.5e-277 2
P36372 TAP2_RAT RATTUS NORVEGICUS (RAT).< 479 1.9e-134 3
Q03519 TAP2_HUMAN HOMO SAPIENS (HUMAN).< 468 1.1e-131 3
P36371 TAP2_MOUSE MUS MUSCULUS (MOUSE).< 472 2.3e-131 3
P06795 MDR1_MOUSE MUS MUSCULUS (MOUSE).< 209 1.2e-05 5
P33311 MDL2_YEAST SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).< 388 2.3e-83 3
P21449 MDR2_CRIGR CRICETULUS GRISEUS (CHINESE HAMSTER).< 206 6.1e-82 5
P43245 MDR1_RAT RATTUS NORVEGICUS (RAT).< 210 2.4e-79 6
WARNING: Descriptions of 280 database sequences were not reported due to the limiting value of parameter V = 10.
>Q03518 TAP1_HUMAN HOMO SAPIENS (HUMAN).<
Length = 748
Score = 1604 (735.8 bits), Expect = 0.0, Sum P(2) = 0.0
Identities = 319/320 (99%), Positives = 319/320 (99%)
Query: 1 MAIPFFTGRLIDWILQDGSADTFTRNLTLMSILTIASAVLEFVGDGIYNNTMGHVHSHLQ 60
MAIPFFTGRLIDWILQDGSADTFTRNLTLMSILTIASAVLEFVGDGIYNNTMGHVHSHLQ
Sbjct: 202 MAIPFFTGRLIDWILQDGSADTFTRNLTLMSILTIASAVLEFVGDGIYNNTMGHVHSHLQ 261
Grail Gene Models
MODEL EXONS # of Exons: 10
GENE MODEL # of Mod
Frm Position Score: S/A - D/I Srch
1 2 26163 - 26272 * 0.21 - 0.12 0
2 0 26421 - 26551 0.78 - 0.01 0
3 2 27511 - 27716 0.96 - 0.52 0
4 2 28143 - 28340 0.40 - 0.74 0
5 0 29542 - 29670 0.92 - 0.95 0
6 2 29820 - 29999 0.97 - 0.00 0
7 0 30568 - 30741 0.96 - 0.81 0
8 0 30985 - 31147 0.96 - 0.03 0
9 2 31456 - 31592 0.93 - 0.95 0
Date Str Seq
1 03/19/97 F 84765 -

Learning Concept



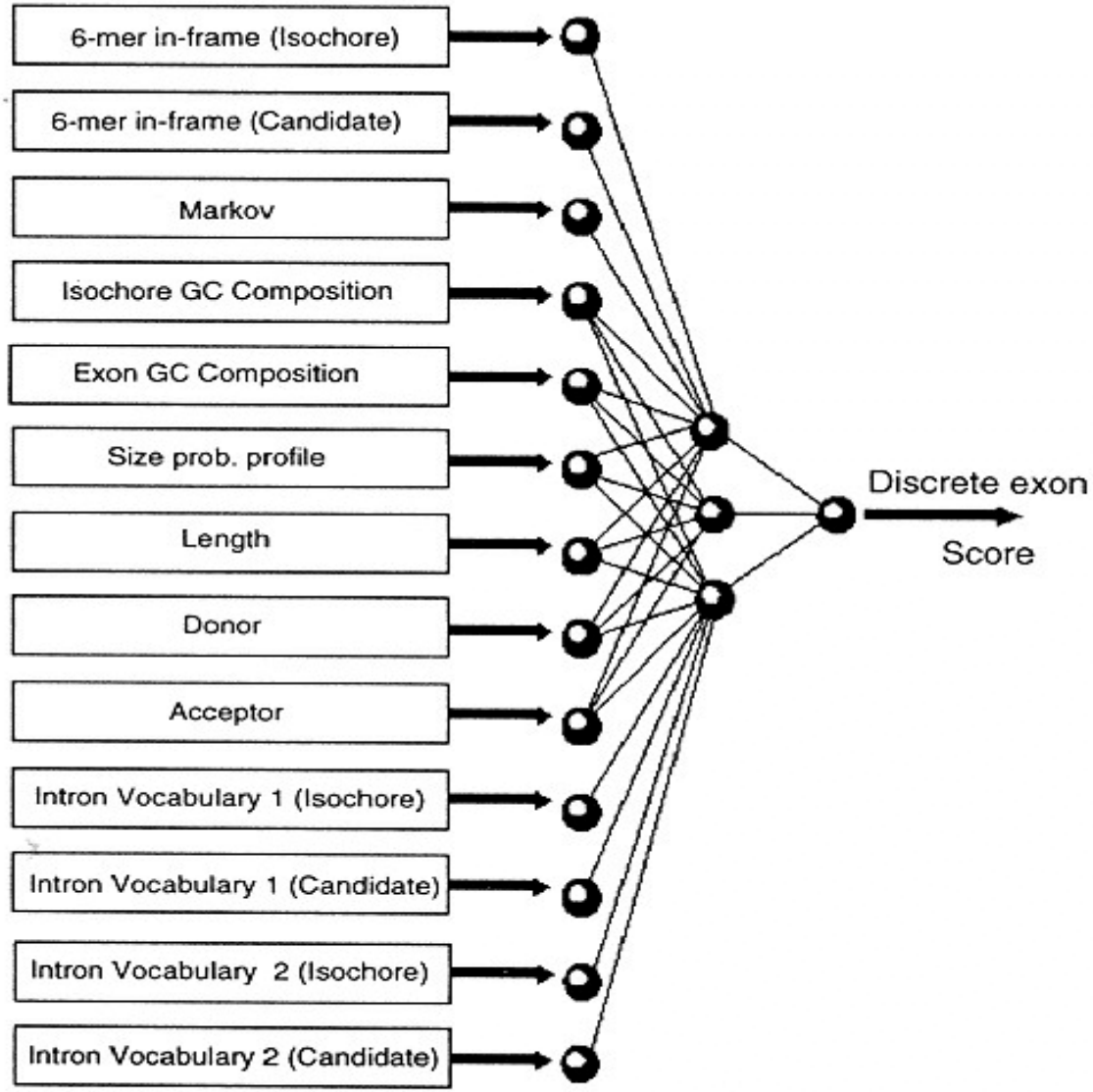


FIG. 1. Schematic of the neural network for evaluating internal protein coding exons in GRAIL.

Important Features

- 6-mer in-frame: higher frequencies of 6-mers in genomic DNA that are more commonly found in cording regions can be an indicator of the presence of an exon
- Harkov model: for gene recognition
- GC Composition: The recognition of coding regions using the 6-tuple method is known to have strong dependence on the G+C (bases G and C)
- Donor (end of exon/beginning of intron), Acceptor (end of intron/beginning of exon) => evaluate the region for potential splice sites (score)

Neural Networks in GRAIL

Known Sequence

CATATTCAAGAATTGAAGCGTGTAGT
CCTGACTTGAGAGCTGTAGATGACGT
GCTTATATGTTC……..

Preprocessing

	Coding potential value	GC Composition	Length	Donor	…	Intron vocabulary		Exon
x_1	0.7	0.8	0.1	0.3	…	0.9	0.2	t_1 1
x_2	0.4	0.2	0.6	0.1	…	0.4	0.5	t_2 0
x_3	0.2	0.9	0.3	0.1	…	0.8	0.3	t_3 0
⋮								
x_n	0.6	0.3	0.2	0.8	…	0.2	0.4	t_n 1

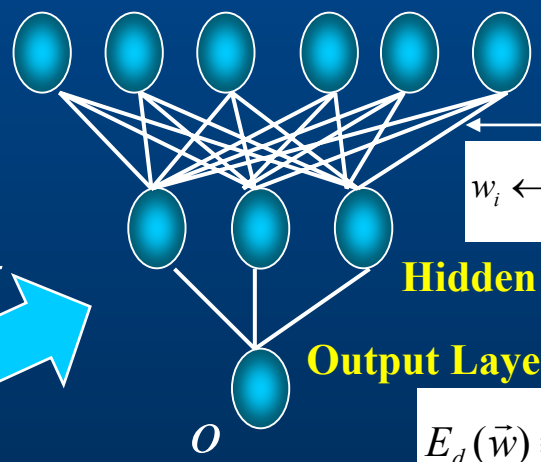
Training

Unknown Sequence

ATGACGTACGATCCCGTGACGGTGA
CGTGAGCTGACGTGCCGTCGTAGTA
ATTTAGCGTGA……..

x 0.6 0.3 0.2 0.8 … 0.2 0.4

Testing $f(x)$?



$$w_i \leftarrow w_i + \Delta w_i \quad , \quad \Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

$$E_d(\vec{w}) \equiv \frac{1}{2} \sum_{k \in outputs} (t_k - o_k)^2$$

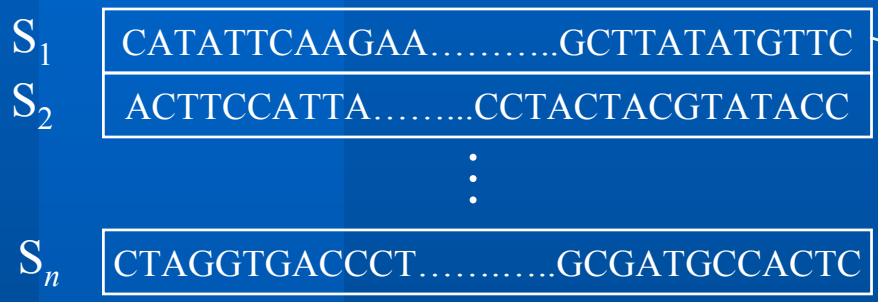
Web Site and Program

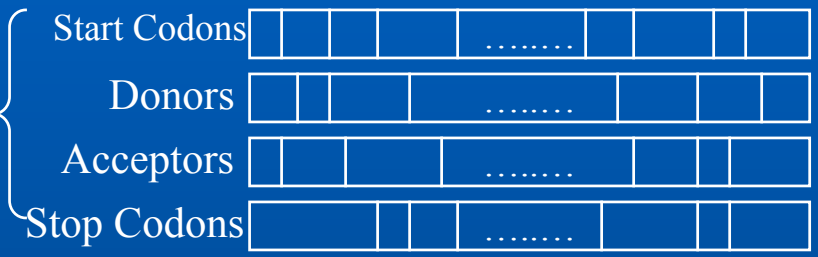
- Gene Recognition and Assembly Internet Link (Version 1.3)
  - http://compbio.ornl.gov/Grail-1.3/
- **GrailEXP** (Grail Experimental Gene Discovery Suite)
  - http://grail.lsd.ornl.gov/grailexp/
- XGRAIL (UNIX platform)
  - http://www.hgmp.mrc.ac.uk/Registered/Option/xgrail.html

MORGAN: A Decision Tree System for Gene Finding

- Integrated system for finding genes in DNA sequences
  - Parse a genomic DNA sequence into coding and non-coding regions.
  - Multi-frame Optimal Rule-based Gene Analyzer
  - Decision Trees (DTs)
  - Markov Chains (MCs)
  - Dynamic Programming (DP)

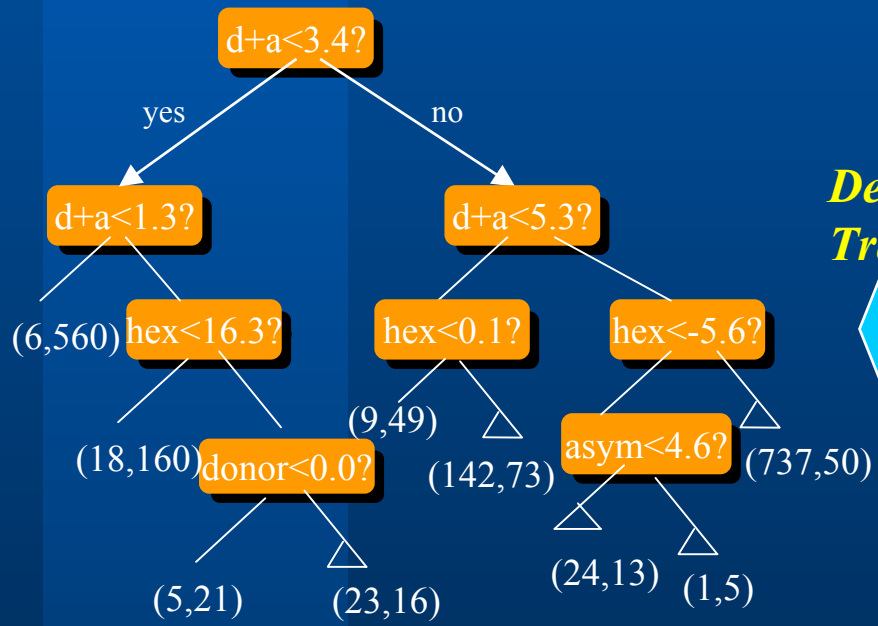
MORGAN: Training Procedure



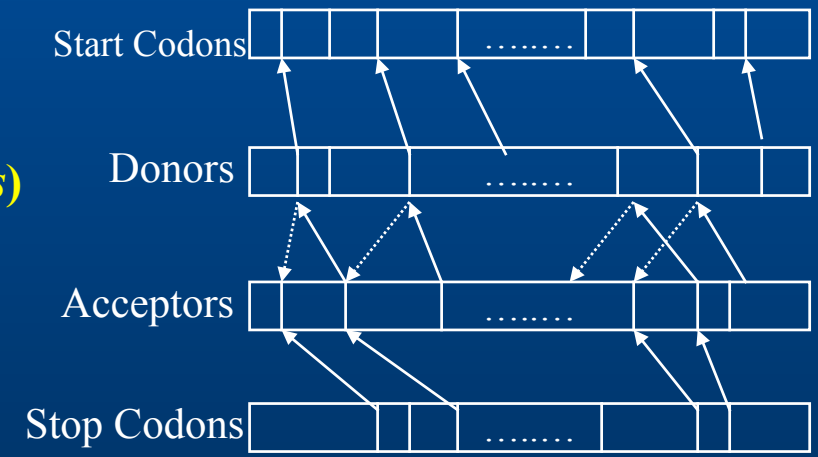


Markov Chains (MC)

Dynamic Programming (DP)



Decision Trees (DTs)



d+a Scoring

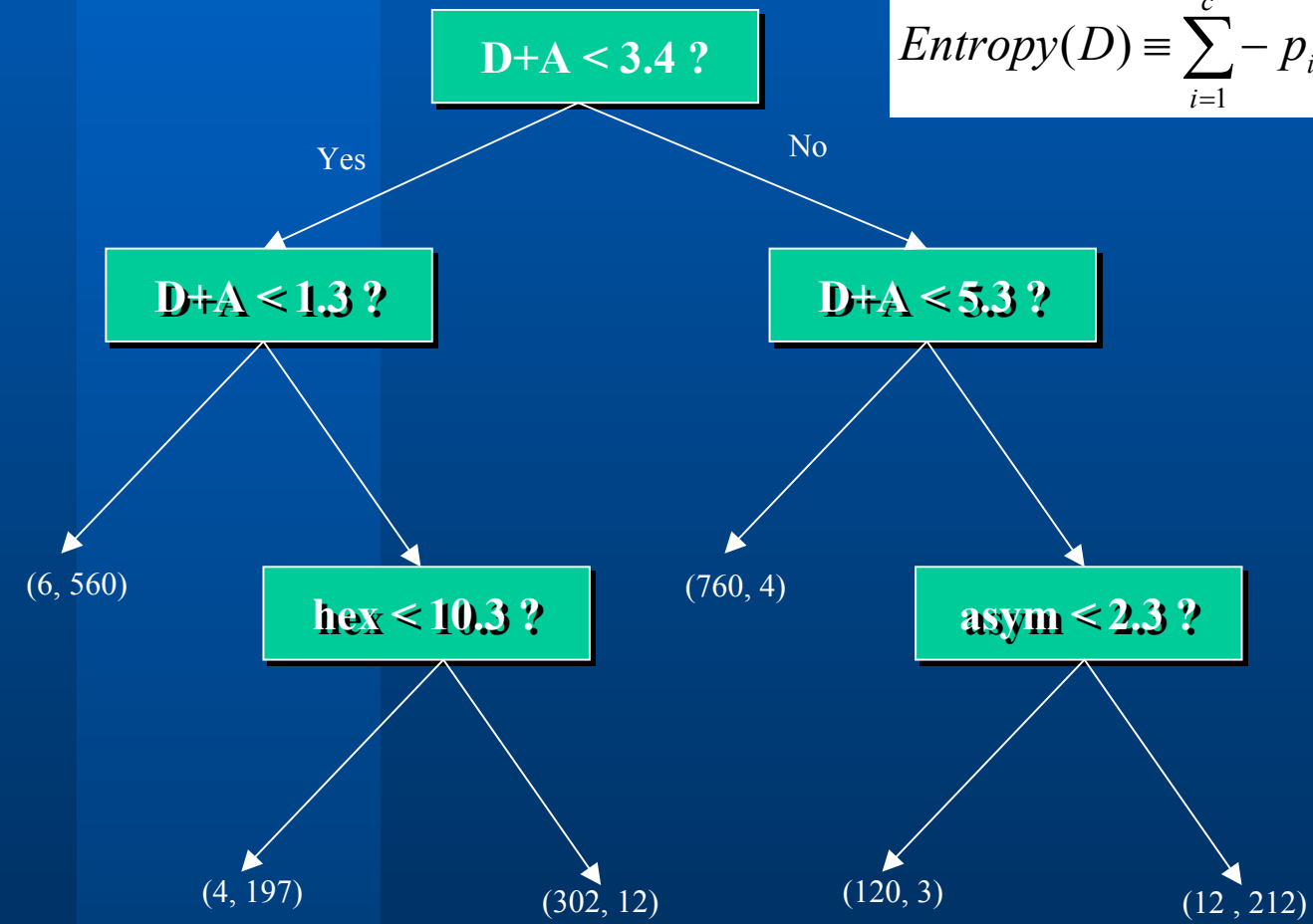
MORGAN: Training Set for Decision Trees

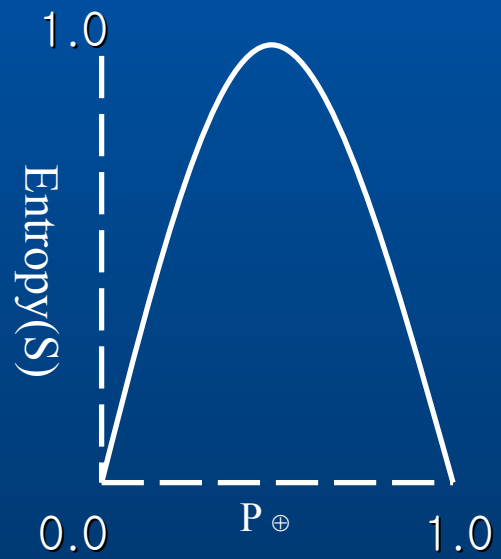
Position(n)	*D+A*	*Hex*	*Asymmetry*	*Exon*
P_1	1.5	0.3	1.8	**No**
P_2	4.3	- 2.5	4.1	**No**
P_3	6.9	10.7	6.3	**Yes**
P_4	3.2	2.5	5.5	**Yes**
P_5	2.5	- 2.5	4.3	**Yes**
P_6	0.7	0.3	1.4	**No**
P_7	3.1	12.5	2.9	**Yes**
P_8	1.8	1.4	3.3	**No**
P_9	2.4	0.1	2.1	**Yes**
P_{10}	3.2	8.3	6.2	**Yes**
P_{11}	1.5	0.1	3.1	**Yes**
⋮	⋮	⋮	⋮	**Yes**
P_n	1.0	1.8	4.5	**No**

Decision Tree Representation

$$Gain(D, A) \equiv Entropy(D) - \sum_{v \in Values(A)} \frac{|Dv|}{|D|} Entropy(Dv)$$

$$Entropy(D) \equiv \sum_{i=1}^{c} - p_i \log p_i$$





Markov Chains to Find Splice Sites

- MCs identify four signal types
  - Start signals, Donor sites, Acceptor sites, Stop codons

Ex) A simple Markov chain for a start codon

A 0.91	A 0.03	A 0.03
C 0.03	C 0.03	C 0.03
G 0.03	G 0.03	G 0.91
T 0.03	T 0.91	T 0.03

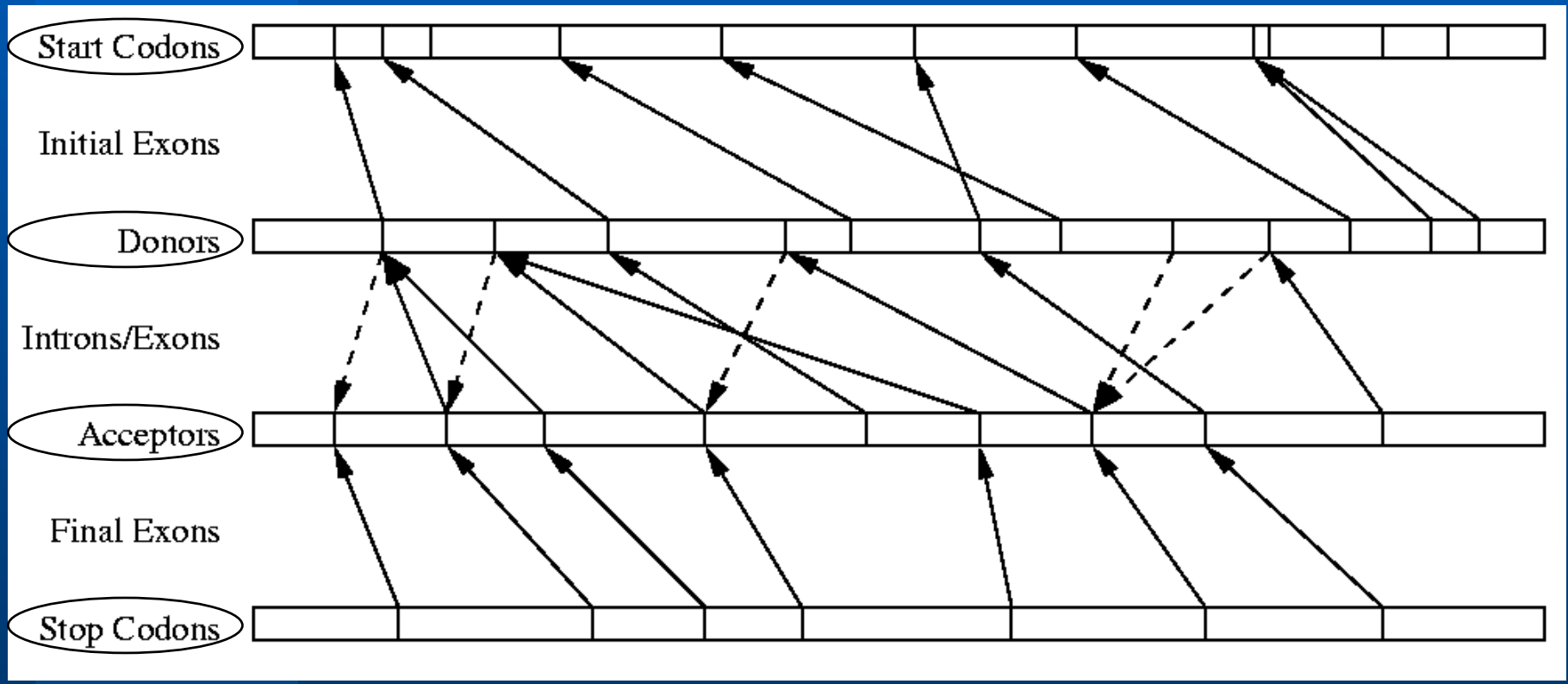
CTG

0.03 X 0.91 X 0.91 = 0.025

CTG: start codon?

$$P(M \mid CTG) = \frac{P(CTG \mid M)P(M)}{P(CTG)}$$

How the Dynamic Programming Algorithm Finds the Optimal Parse



Data and Experiments

- 570 vertebrate sequences

Data Set	Contents
Training Set	80%, 454 sequences, 2.3 million bases, 2146 exons
Test Set	114 sequences, 607924 bases, 499 exons
Second Test Set	80% identity to any sequence in the training set (97 sequences, 566962 bases)

Leading Gene-Finding Systems

Gene finder	Coding bases			Exact exons			ME
	Sn	Sp	AC	Sn	Sp	Avg	
MORGAN	0.81	0.83	0.79	0.59	0.59	0.59	0.17
GENSCAN	0.83	0.93	0.91	0.78	0.81	0.80	0.09
VEIL	0.83	0.72	0.73	0.53	0.49	0.51	0.19
Genie	0.78	0.84	0.77	0.61	0.64	0.62	0.15
FGENEH	0.77	0.85	0.78	0.61	0.61	0.61	0.15
GRAIL 2	0.72	0.87	0.75	0.36	0.43	0.40	0.25
GeneID	0.63	0.81	0.67	0.44	0.46	0.45	0.28
GeneParser2	0.66	0.79	0.67	0.35	0.40	0.37	0.29
GenLang	0.72	0.79	0.69	0.51	0.52	0.52	0.21
SorFind	0.71	0.85	0.73	0.42	0.47	0.45	0.24
Xpound	0.61	0.87	0.68	0.15	0.18	0.17	0.33

[a] AC is the approximate correlation proposed by Burset and Guigo (1996) [30] as a replacement for the correlation coefficient. Sensitivity (Sn) is the fraction of true coding bases that were correctly predicted as coding, and specificity (Sp) is the number of bases predicted to be in coding regions that actually were coding; their average is given in the Avg column. The "exact exon" columns show the corresponding results for prediction of whole exons. ME (missing exons) is the fraction of whole coding exons that are missed completely.

Promoter Region Prediction

What is Promoter Region ?

- A sequence that is used to ***initiate and regulate transcription*** of a gene. This is a ***crucial step in gene expression*** in general.
- (i) a gene region immediately ***upstream of a transcription initiation site***
  (ii) a *cis* - acting genetic element ***controlling the rate of transcription initiation*** of a gene
- Most genes in higher eukaryotes are transcribed from ***polymerase II dependent promoters.***

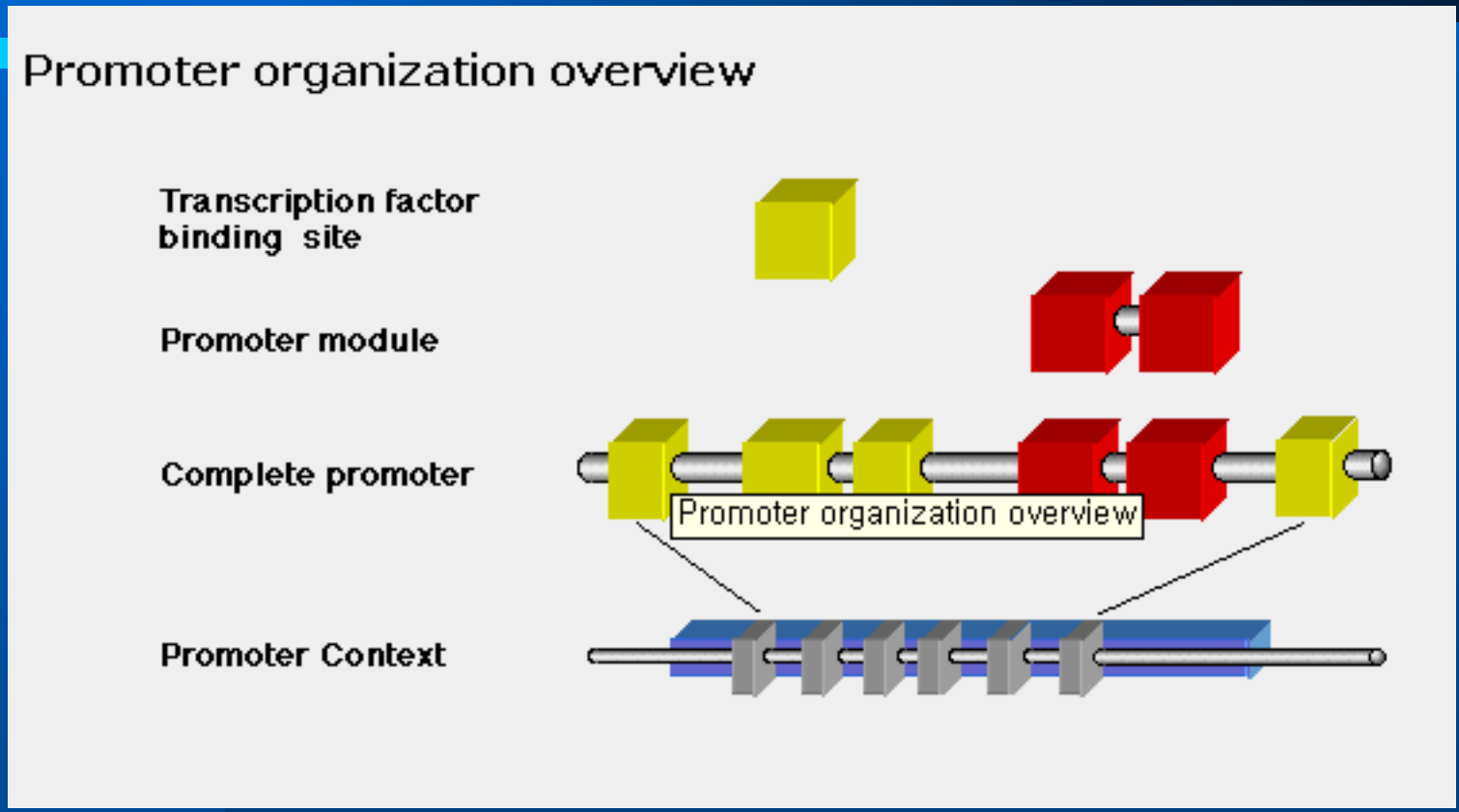
Why Promoter Region Prediction ?

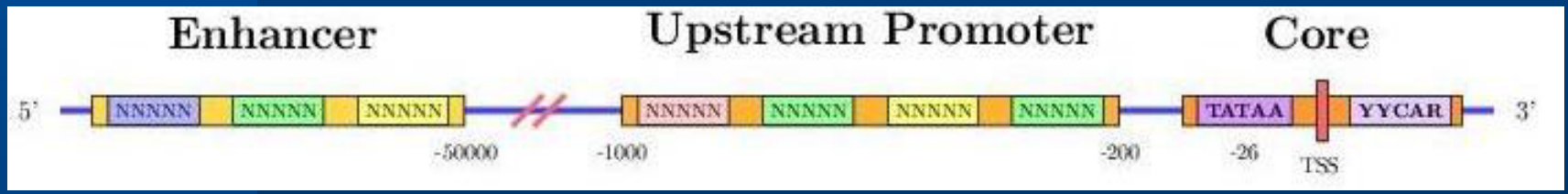
- ***Gene Finding***
- Determining the ***Correct Protein Translation***
- Determining the ***Expression Context***
  - DNA chip data analysis
- ***Genetic Network*** Analysis

Promoter Region Organization (1)

- **Promoters**
  - DNA regions which also contain ***transcription factor binding sites*** similar to enhancers but also include elements for specific ***initiation of transcription (core promoter).***
- **Enhancers**
  - DNA regions which are usually ***rich in transcription factor binding sites*** and/or repeats. They ***enhance transcription*** of the responsive promoter independent of orientation and position.

Promoter Region Organization (2)





Promoter Prediction: Method (1)

- **Pattern driven**

- ***Collecting*** a set of real ***transcription factor binding*** sites to build a characteristic representation or ***profile*** from them.
- ***Searching*** potential binding sites on the input sequences by using their characteristic ***profile***.
- ***Assembling*** found binding sites following some rules about these arrangements should be done to ***re-build*** the promoter region.

Promoter Prediction: Method (2)

- **Sequence driven**

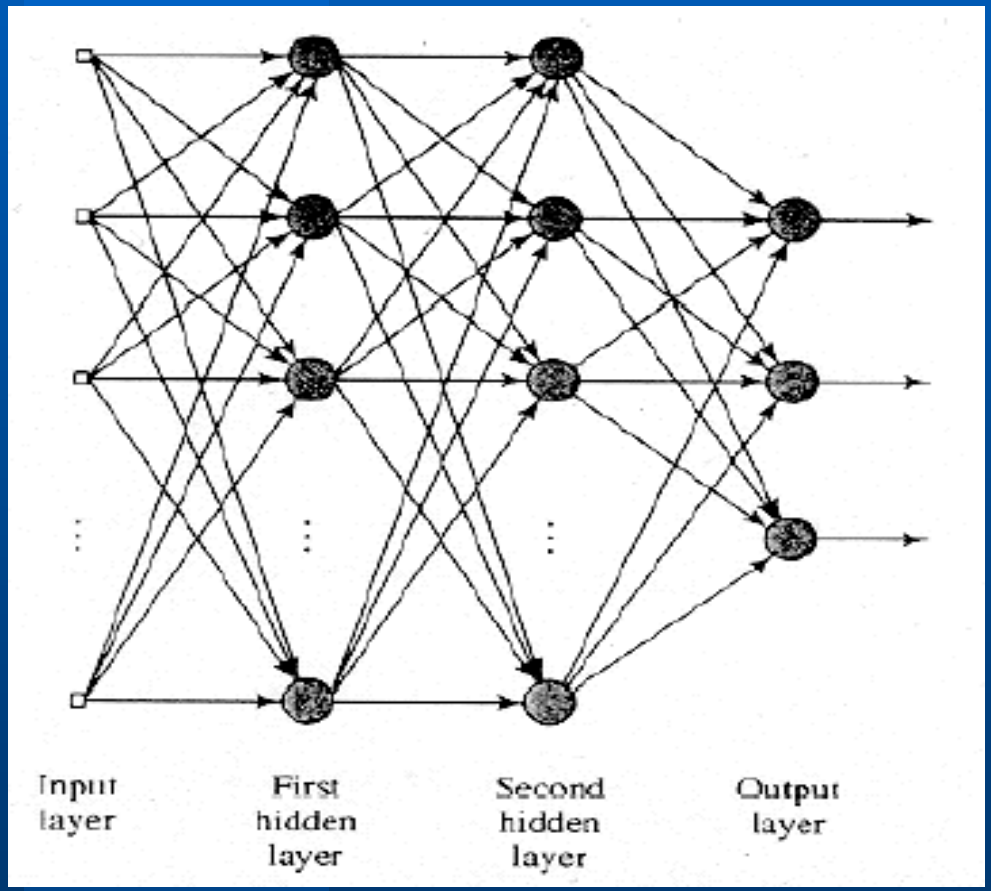
  Making different ***pairwise comparisons (alignments)*** between input sequences to form common patterns corresponding to ***well-conserved functional binding sites*** without using more information.

Promoter Prediction: Method (3)

- **Some recent approaches:**

- (Statistical) Discriminant analysis, regression analysis
- Consensus sequences (regular expressions)
- Position weight matrices
- ***Neural networks***
- ***Genetic Algorithm***
- Clustering of putative binding sites
- Oligonucleotide counts (word frequency), Markov models
- ***Hidden Markov models***
- Pairwise alignment, multiple alignment
- Iterative methods: Gibbs sampling

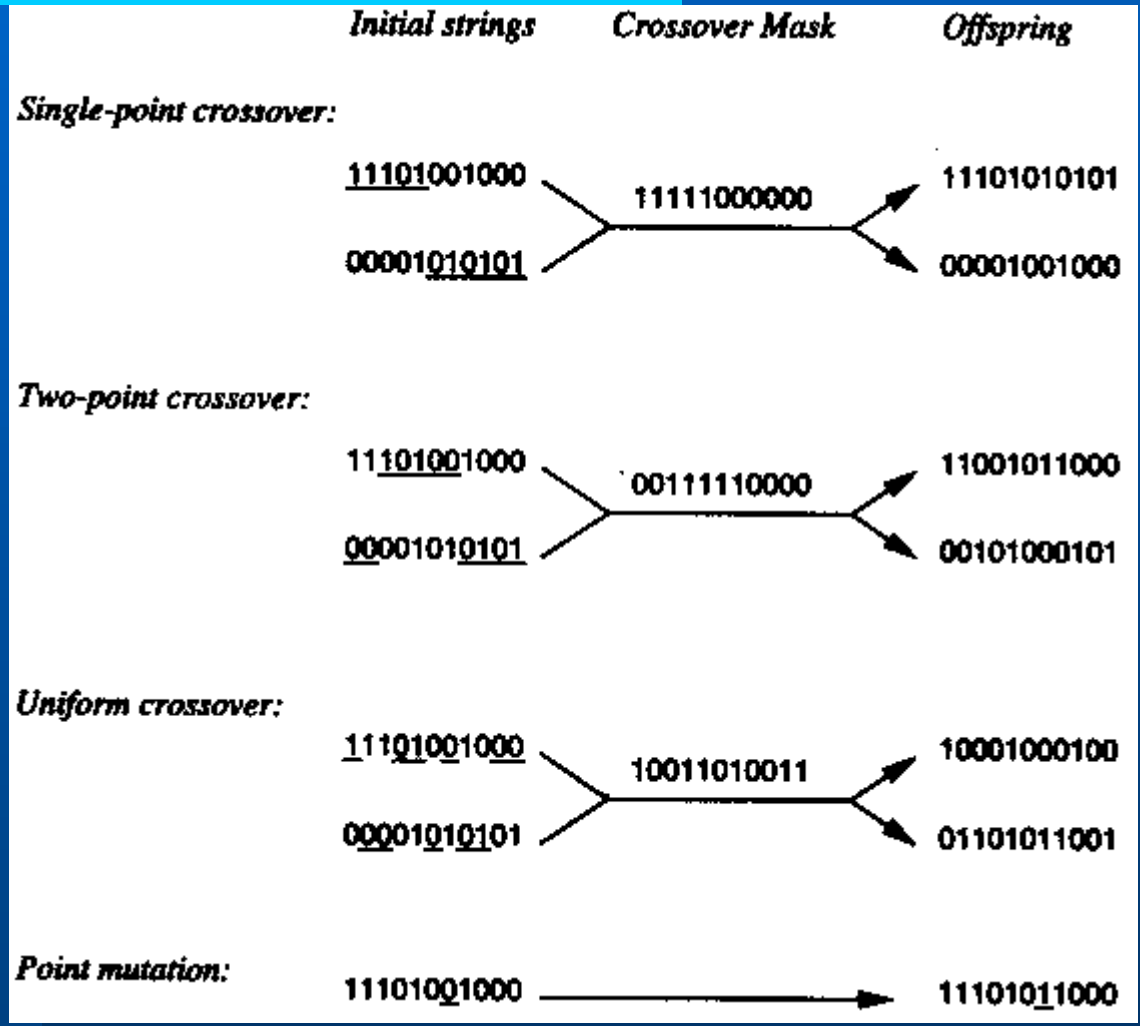
Neural Networks



Characteristic

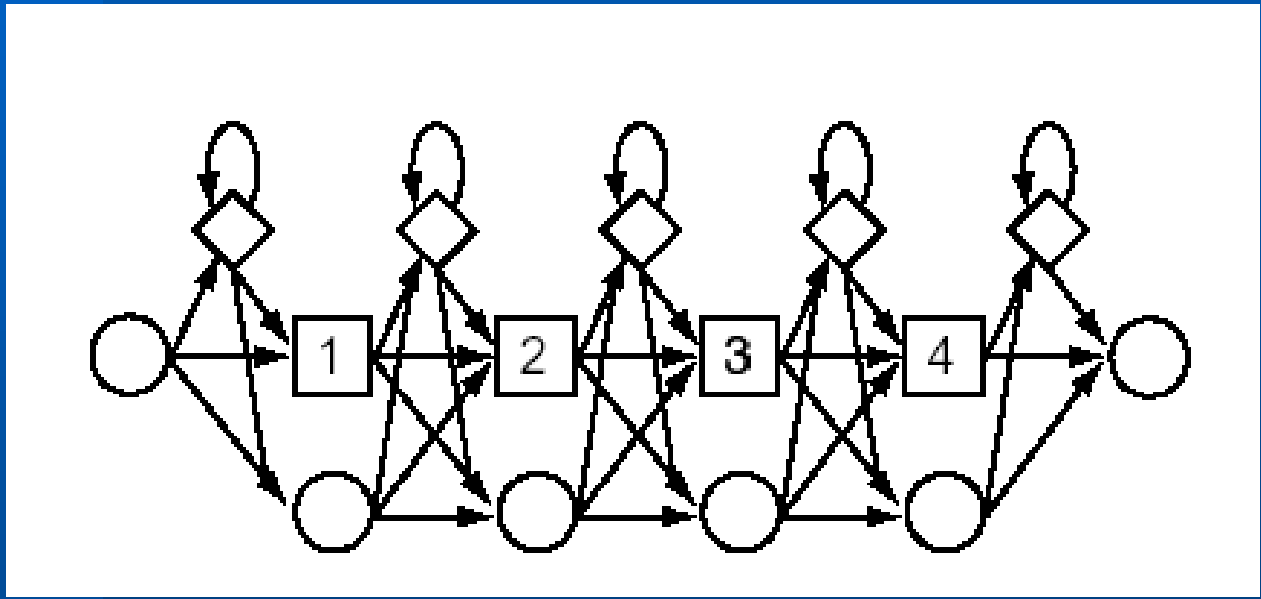
- Nonlinear I/O mapping
- Adaptivity
- Generalization ability
- Fault-tolerance (graceful degradation)
- Biological Analogy

Genetic Algorithm



Hidden Markov Model

< Profile HMM >



Deletion state Match state Insert state

Transition probability

Promoter Prediction: Reliability

- **Problems**

- Too much ***false positives*** (specifity) due to:

  1. Binding sites patterns are very short (5-15 bp)

  2. Usage of Transcription Factor databases: bias (new patterns)

  3. Relationships between Transcription factors are complex and degenerated

- **Improvement**

- Genome-wide expression data from microarrays

- Phylogenetic information from homologous genes

- New research about epigenetic information:

  CpG islands, DNA bendability, modules (cooperative sites)

Promoter Prediction Tools

- Audic/Claverie
  - Markov models of vertebrate promoter sequence.
- Autogene
  - Clustering algorithm based on the consensus site occurance)
- GeneID/Promoter2.0
  - Neural network and genetic algorithm
- NNPP
  - Time delay neural net architecture.
- TSSG/TSSW
  - Linear discriminant function

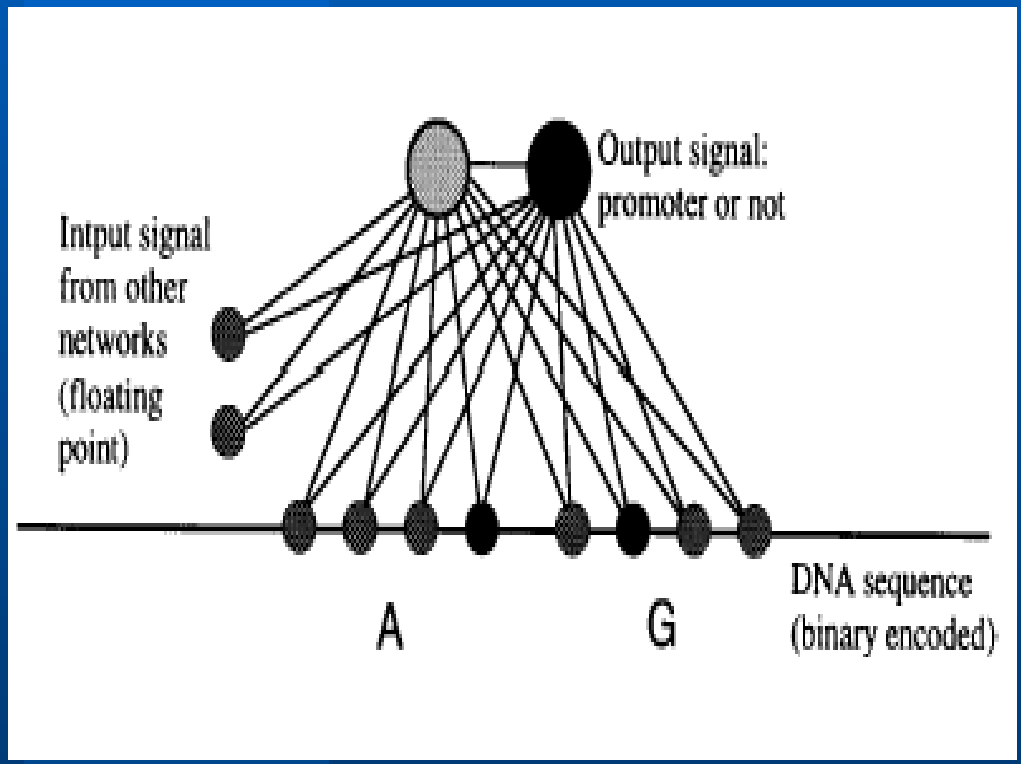
Table 1: Servers and software for promoter finding

Detection of pol-II promoters	
[1]Audic/Claverie	Send request to audic@newton.cnrs-mrs.fr
[2]CorePromoter	http://sciclio.cshl.org/genefinder/CPROMOTER/
[3]FunSiteP	http://transfac.gbf.de/dbsearch/funsitep/fsp.html
[4]ModelGenerator/ModelInspector	http://www.gsf.de/biodv/modelinspector.html
[5]PPNN	http://www-hgc.lbl.gov/projects/promoter.html
[6]PromFD 1.0	FTP to beagle.colorado.edu, directory: pub, file: promFD.tar
[7]PromFind	http://www.rabbithutch.com/
[8]Promoter 1.0	http://www.cbs.dtu.dk/services/promoter-1.0/
[9]Promoter Scan	http://biosci.umn.edu/software/proscan/promoterscan.htm
	http://bimas.dcrt.nih.gov/molbio/proscan/
[10]TSSG/TSSW	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
Detection of transcription factor binding sites	
[11]MatInd/MatInspector/FastM	http://www.gsf.de/biodv/matinspector.html
	http://www.gsf.de/biodv/fastm.html
[12]MATRIX SEARCH 1.0	Send request to chenq@boulder.colorado.edu
[13]PatSearch 1.1	http://transfac.gbf-braunschweig.de/cgi-bin/patSearch/patsearch.pl
[14]Signal Scan	http://bimas.dcrt.nih.gov/molbio/signal/
[15]TESS	http://www.cbil.upenn.edu/tess/
[16]TFSEARCH	http://pdap1.trc.rwcp.or.jp/research/db/TFSEARCH.html
General genefinders with pol-II detection and other feature detectors (MARs, CpG-islands)	
[17]GENSCAN	http://CCR-081.mit.edu/GENSCAN.html
[18]GRAIL	http://compbio.ornl.gov/Grail-1.3/
[19]MAR-Finder	http://www.ncgr.org/MarFinder/
[20]WebGene	http://itba.mi.cnr.it/webgene/

[1](Audic & Claverie, 1997), [2](Zhang, 1998b), [3](Kondrakhin *et al.*, 1995), [4](Frech *et al.*, 1997), [5](Reese *et al.*, 1996), [6](Chen *et al.*, 1997), [7](Hutchinson, 1996), [8](Knudsen, Submitted), [9](Prestridge, 1995), [10](Solovyev & Salamov, 1997), [11](Quandt *et al.*, 1995), [12](Chen & Stormo, 1995), [13](Wingender *et al.*, 1998), [14](Prestridge, 1997), [15](Schug & Overton, 1997), [16]Has not been published, [17](Burge & Karlin, 1997), [18](Matis *et al.*, 1996; Uberbacher *et al.*, 1996), [19](Singh, 1997), [20](Milanesi *et al.*, 1996; Milanesi & Rogozin, In press).

Promoter 2.0

http://www.cbs.dtu.dk/services/promoter/



- Neural Network
  - Input a small window of DNA sequence
  - Output of other neural networks.
- Genetic algorithm:
  - The ***weights*** in the neural networks are optimized to discriminate maximally between promoters and non-promoters.

Correlation coefficient: 0.63

Steen knudsen, Promoter 2.0:for the recognition of polll promoter sequences, Bioinformatics,Vol15, 356-361, 1999

McPromoter Finder (1)

- Integrate physical properties of DNA (DNA bendability, GC contents, CpG island) and DNA sequence.

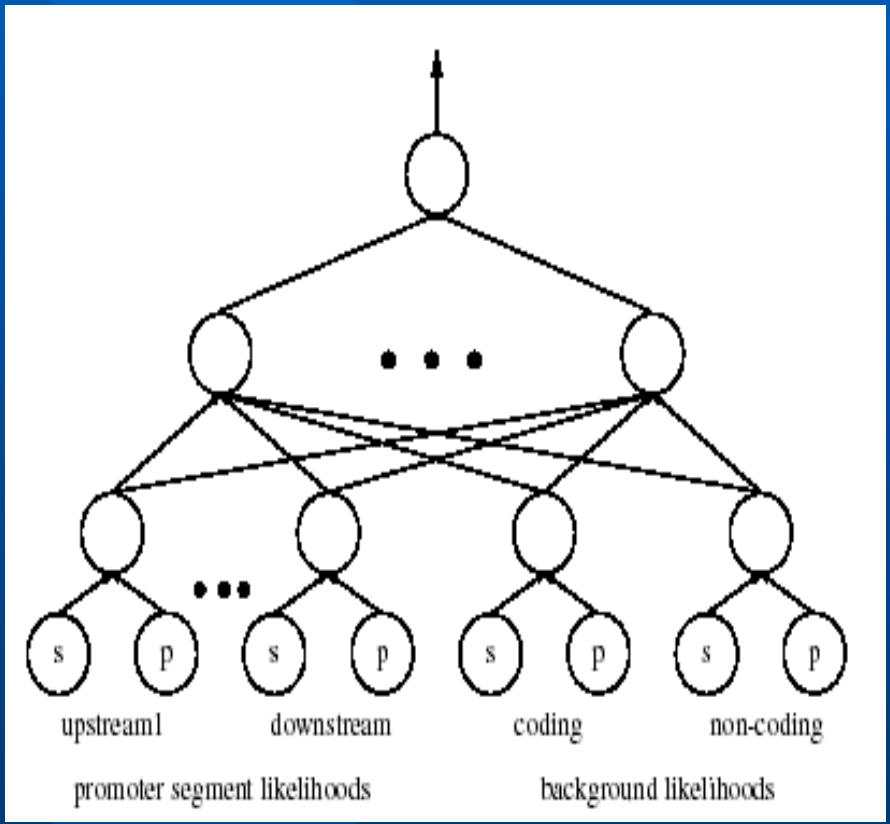
  1) ***Sequence*** likelihoods are modeled with ***interpolated Markov chains***

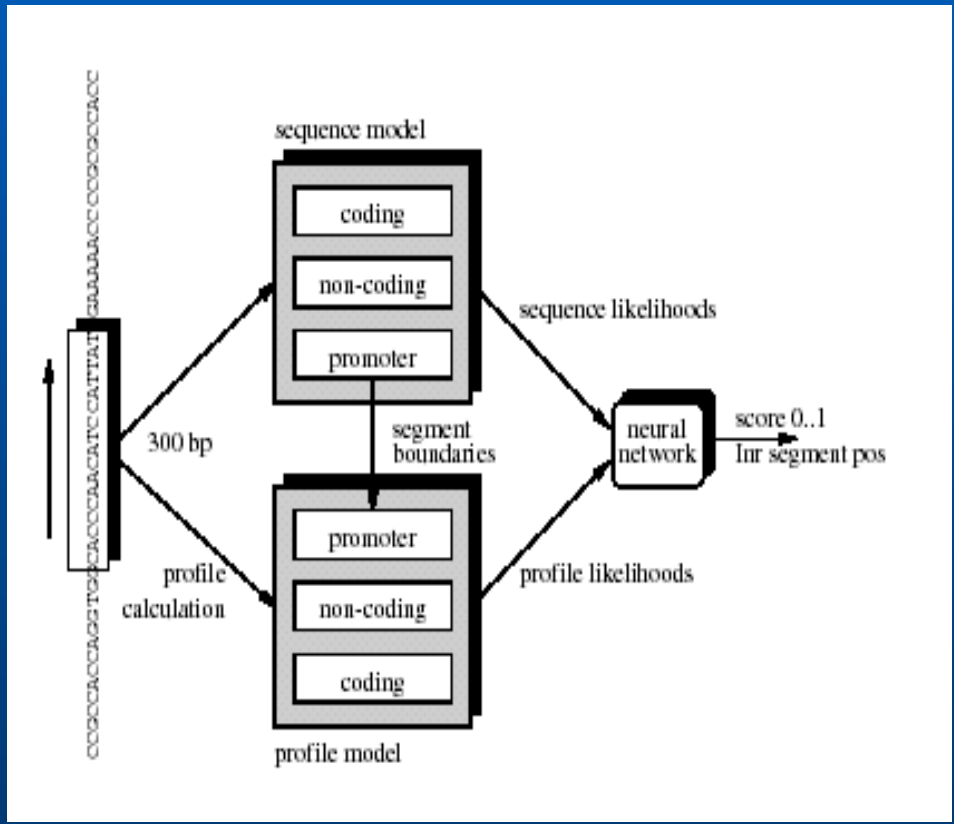
  2) ***Physical properties*** are modeled with ***Gaussian distribution***

- The models were trained on a representative set consisting of ***vertebrate promoters*** and ***human non-promoter*** sequences respectively on ***D. melanogaster promoters*** and ***non-promoters***

- The current classification performance on our human set: ***61% of the promoters recognized***, 1% of false positives (a ***correlation coefficient of 0.71***).

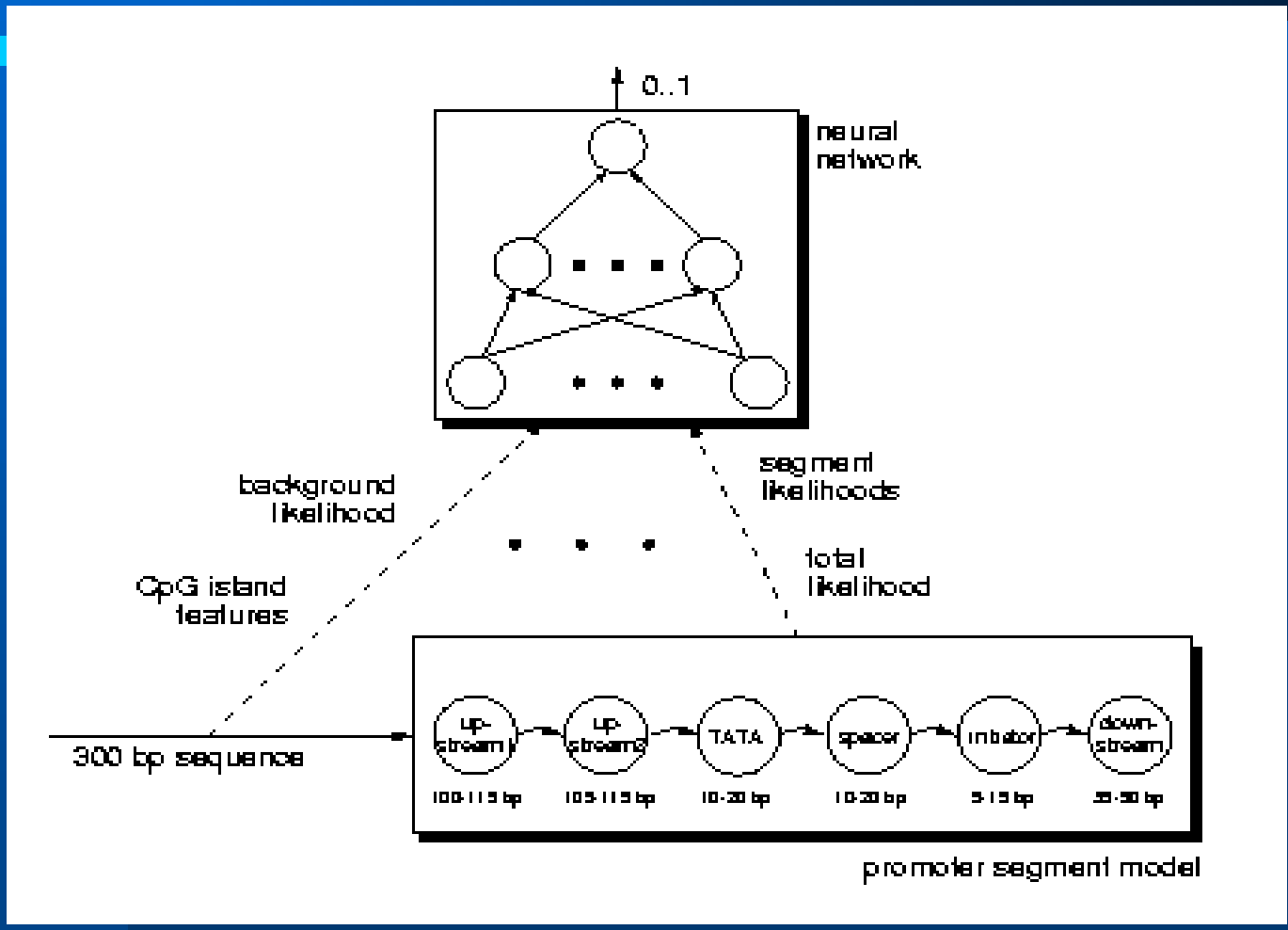
McPromoter Finder (2)





S: DNA sequence, **P**: Physical properties of DNA

McPromoter Finder (3)



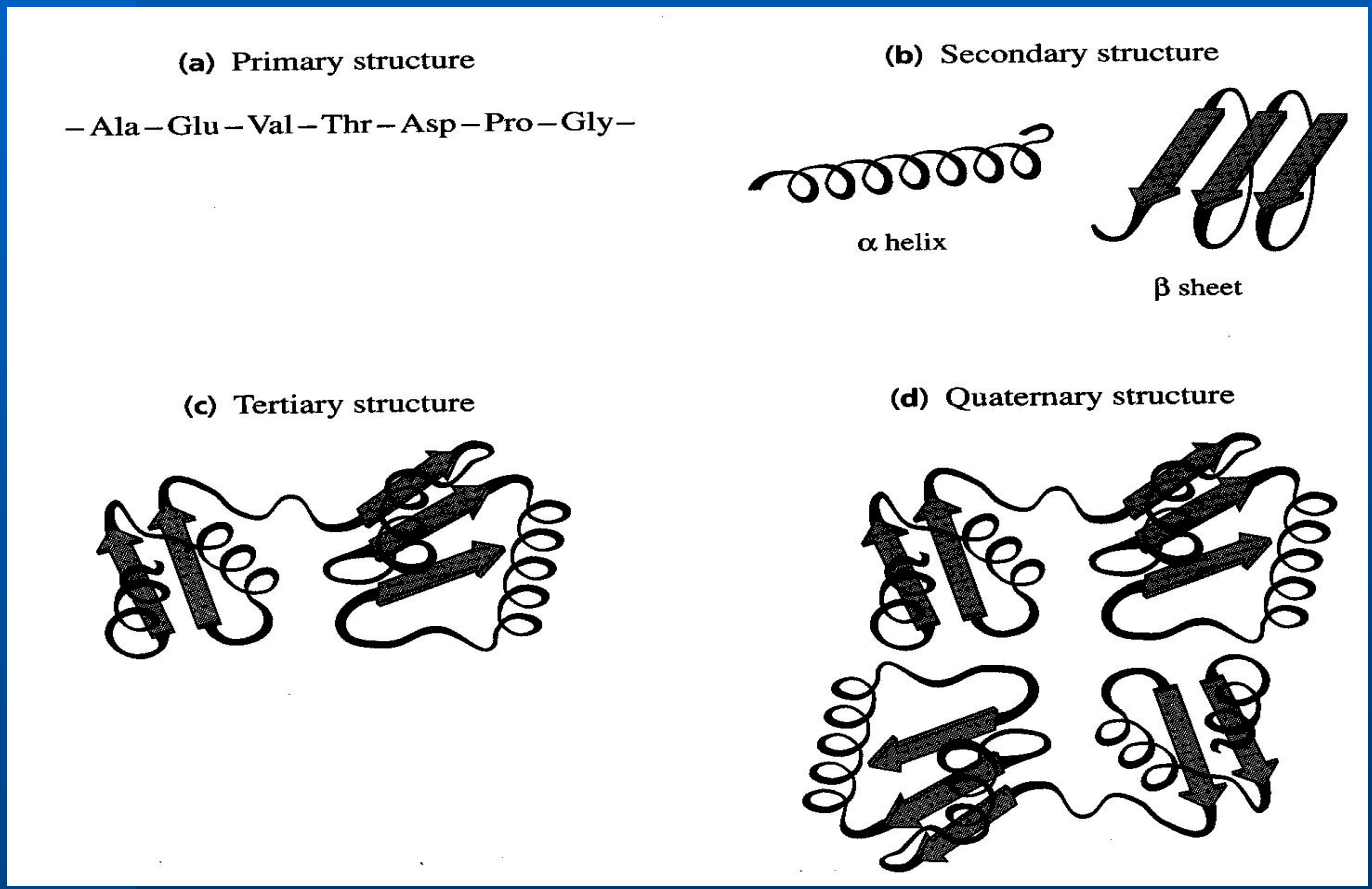
U. Ohler, H. Niemann, G. Liao and G. M. Rubin
Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition
Bioinformatics 17:S199-S206, 2001.

Protein Structure Prediction

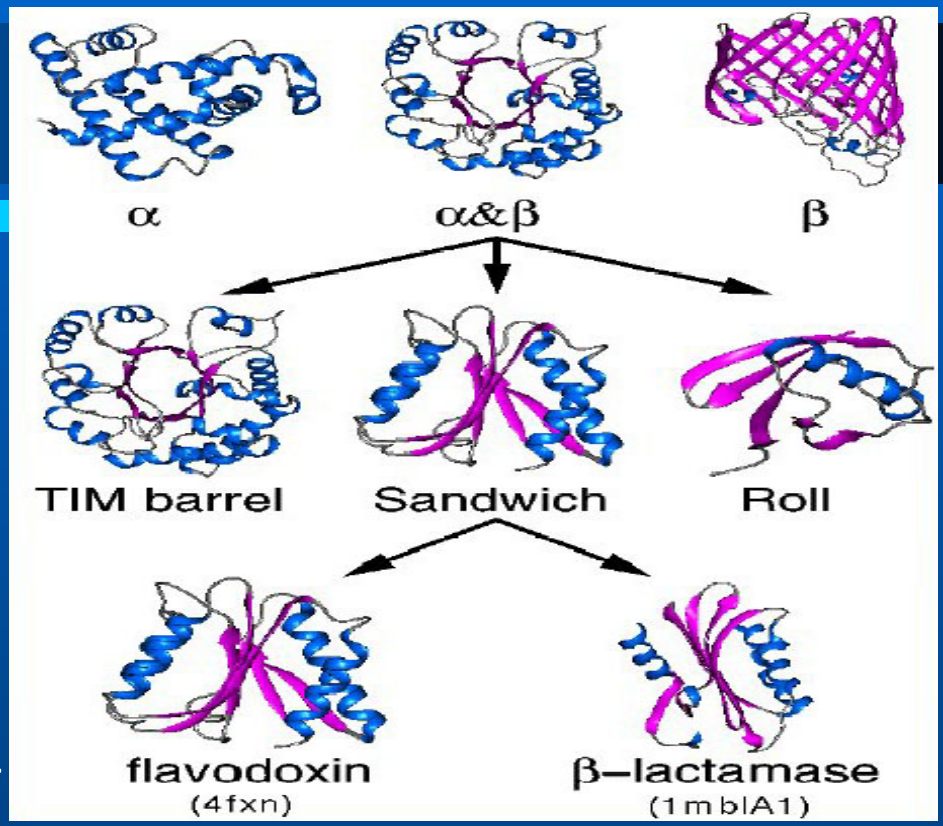
Outline

- Introduction to Protein Structure
- Introduction to HMM
- Computational Protein Structure Prediction
- SAM – HMMer & Pfam
- HMMstr
- Other Non-HMM Prediction Methods
  - SWISS-MODEL
  - VAST

Introduction to Protein Structure



Cont'd



Motif (protein sequence pattern): is recognizable combinations of α helices and β strands that appear in a number of proteins.

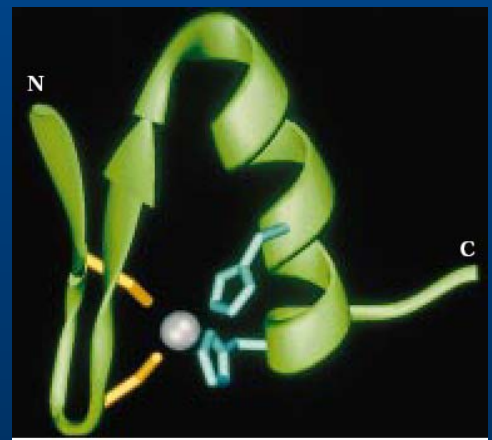
Domain consists of combinations of motifs, the size of domains varies from about 25 to 30 amino acid residues to about 300, with an average of about 100.

Protein family consist of members which has

1) Same function; and

2) Clear evolutionary relationship; and

3) Patterns of conservation, some positions are more conserved than the others, and some regions seem to tolerate insertions and deletions more than other regions, the similarity usually > 25% .



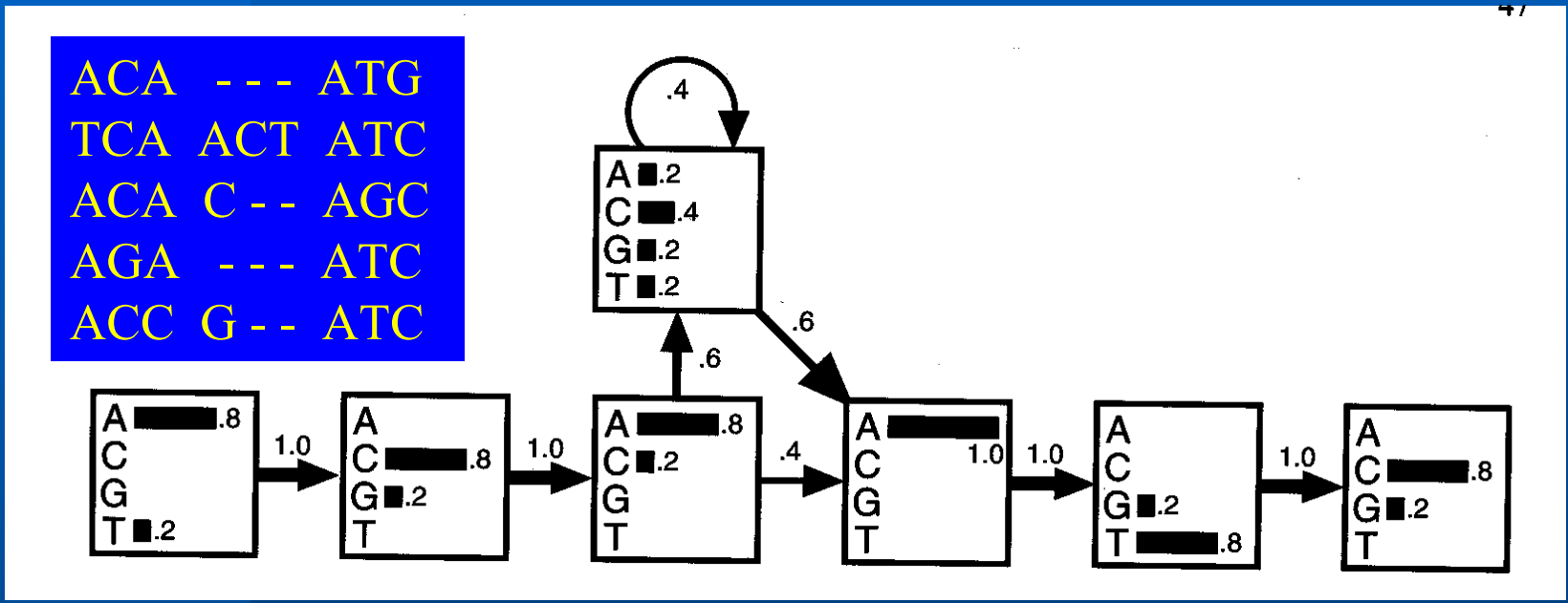
A zinc finger motif.

Introduction to Hidden Markov Models (HMM)

- A class of probabilistic models that describes a probability distribution over a potentially infinite number of sequences.
- Each state has a transition and an emission probability
  - Transition: from state to state transition (Transition probability)
  - Emission: each state emit output (Emission probability)
  - Only one output per state need not be required. Each output has emission probability. "Hidden" means this property
- HMM applications in computational biology

Application	Input	Output	Measure
Gene finding	Single sequences (O)	Likelihood for coding	$P(O)$
Secondary structure prediction	Sequence profile (O)	Secondary structure(D)	$P(D\|O)$
Structural context prediction	Sequence profile (O)	Context(C)	$P(C\|O)$
Dihedral angle region prediction	Sequence profile (O)	Dihedral angle region(R)	$P(R\|O)$
Protein design	Structure (D,C,R)	Sequence(O)	$P(O\|D,C,R)$
Sequence comparison	Sequence 1 (O_1) Sequence 2 (O_2)	Likelihood for alignment	$P(O_1 \sim O_2)$

Example of HMM Model (1) – DNA

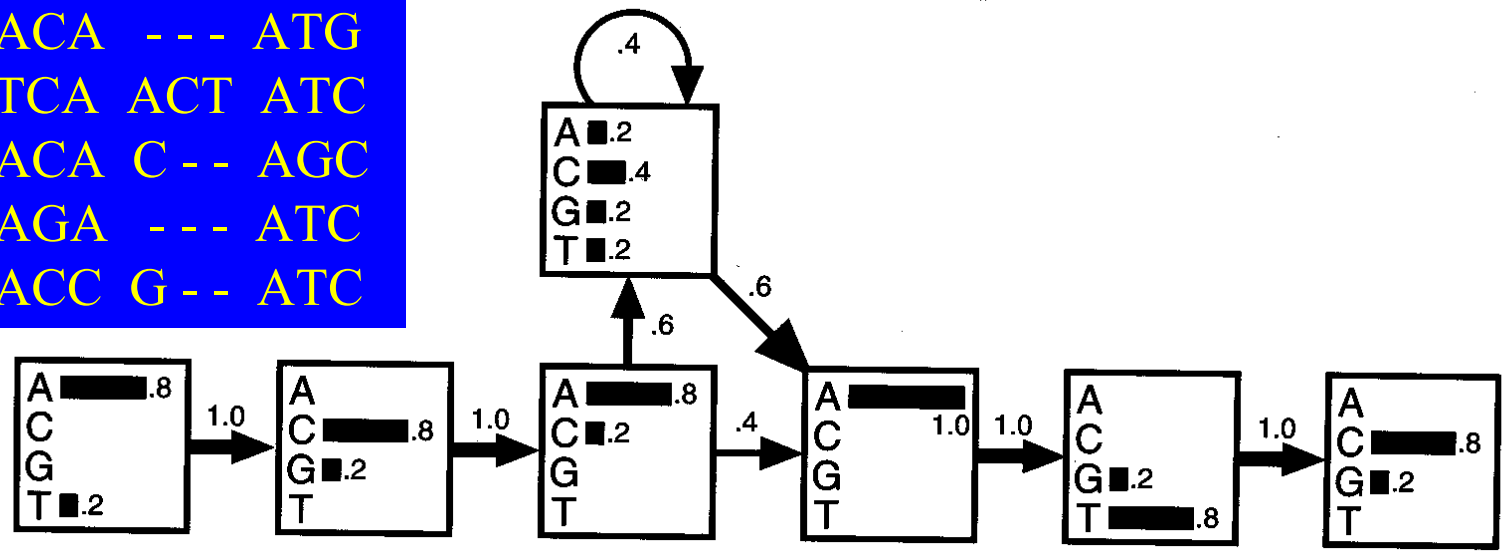


- A HMM model for a DNA motif alignments, The transitions are shown with arrows whose thickness indicate their probability. In each state, the histogram shows the probabilities of the four bases.

Example of HMM Model (2) – DNA

- Scoring

ACA - - - ATG
TCA ACT ATC
ACA C - - AGC
AGA - - - ATC
ACC G - - ATC

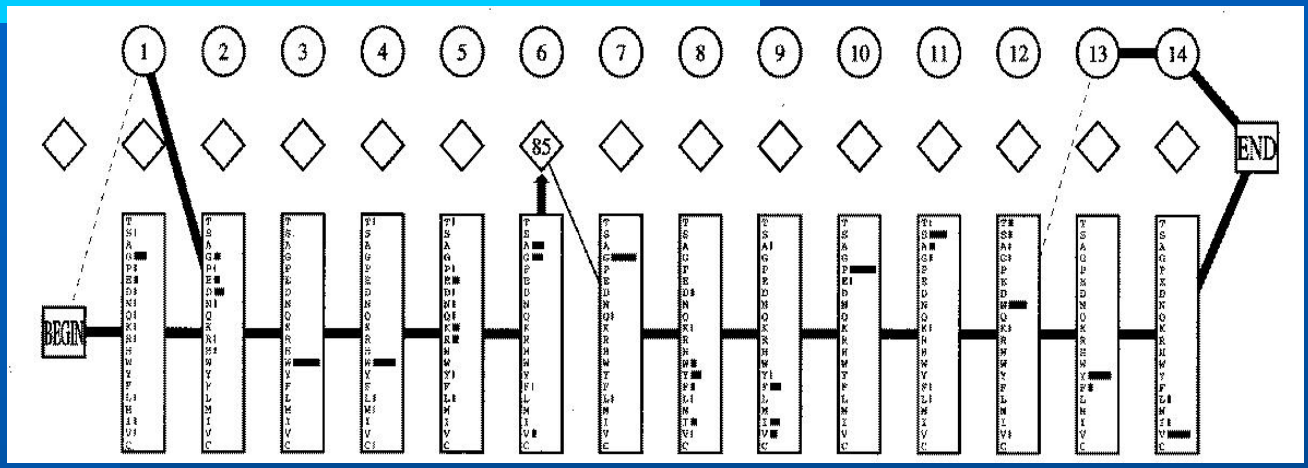


Highly implausible sequence: ACAC - - ATC

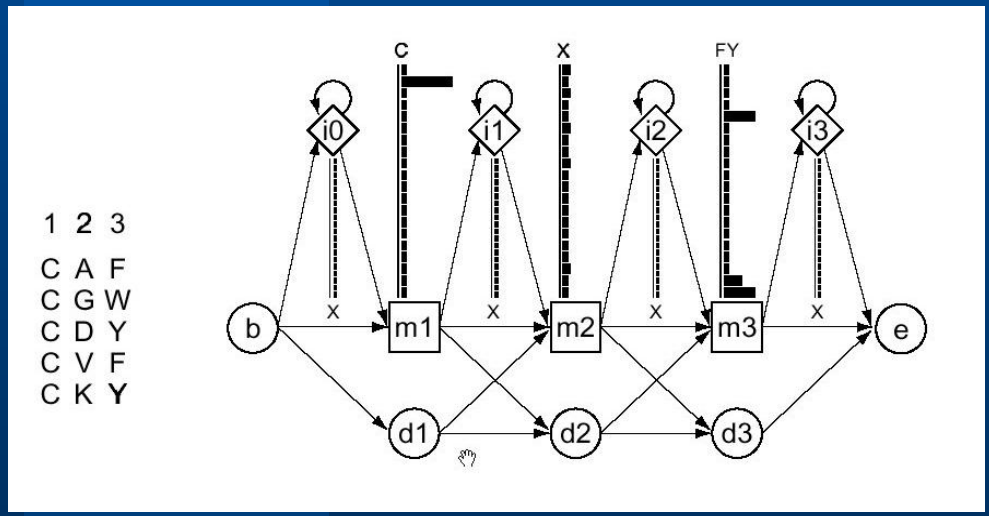
P (ACACATC) = 0.8x1 x 0.8x1 x 0.8x0.6 x 0.4x0.6 x 1x1 x 0.8x1 x 0.8
= 4.7×10^{-2}

Cf) log-odds score for sequence S = log [P(S)/$(0.25)^{L}$]
for this ACACATC sequense, log-odds score is 6.7

Example of HMM Model (Protein)

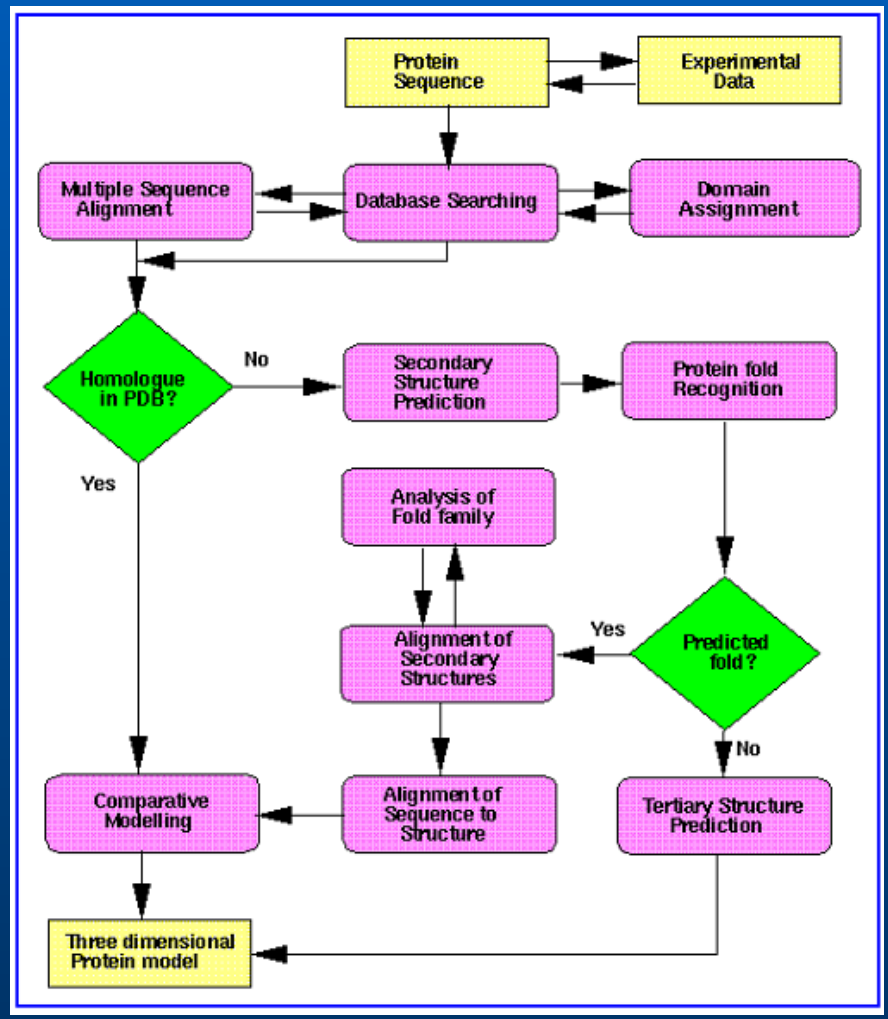


A small profile HMM (right) representing a short multiple alignment of five sequences (left) with three consensus columns.



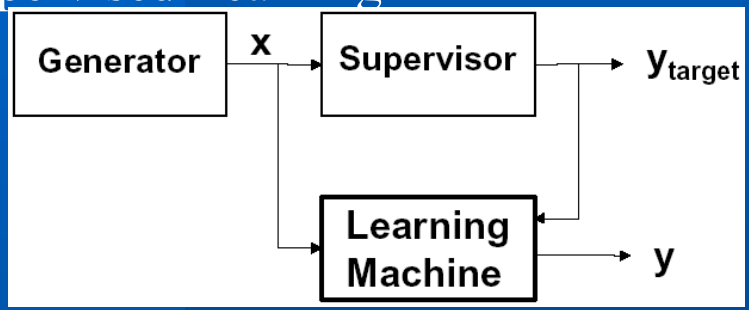
A linear hidden Markov model is a sequence of nodes, each corresponding to a column in a multiple alignment. In our HMMs, each node has a main state (square), insert state (diamond) and delete state (circle).

A Guide for Protein Structure Prediction



Computational Protein Structure Prediction

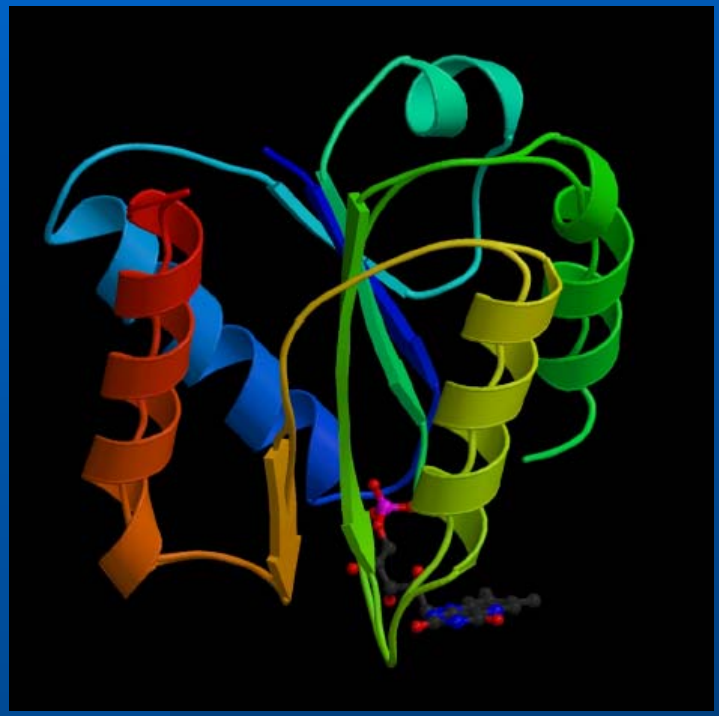
- Supervised Learning

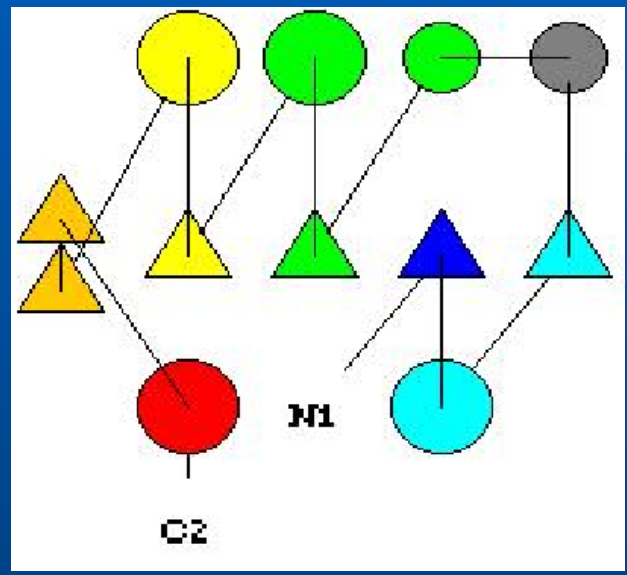


Training: Learn from (X, Ytarget)
Testing: Given X, output Y close to the supervisor's output Ytarget

  - For training, input data with correct output are required.
- Inductive Learning
  - The more complex mapping, the more data required
- Strategy
  - Mapping directly to tertiary structure is difficult
  - So, local aspects of structure that can be induced from the immediate sequence surrounding (Secondary structure prediction problem)
  - Folding problem

Secondary Structure Cartoons





- Helices & sheets & coil
- Secondary structures construct tertiary structures

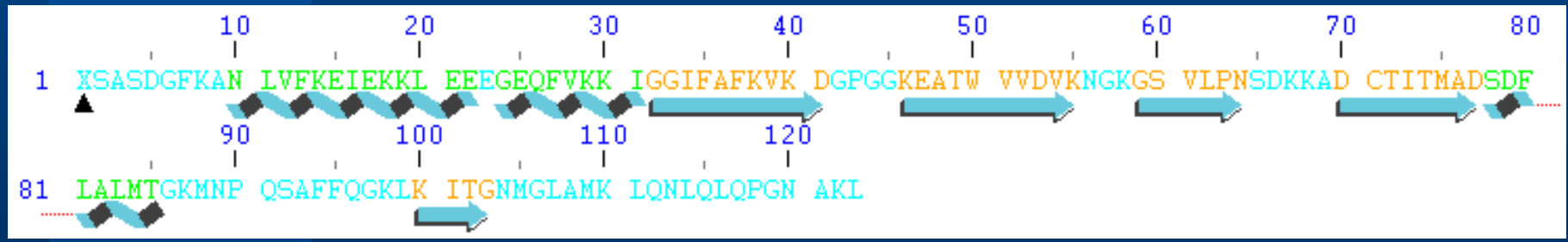
Example of NCBI

- This data is not obtained by prediction, but by experiments.
- But we verify that ‘secondary structures construct a tertiary structure’.
- Meaning of Sequence Details

  H, G, I (helix) E,

  B (beta strand) T (turn) S (bend)

```
 1 SSASDGFKAN LVFKEIEKKL EEEGEQFVKK IGGIFAFKVK DGPGGKEATW
         SGGG THHHHHHHHH HHHTHHHHHH H EEEEEEE  S SSS  EEE

51 VVDVKNGKGS VLPNSDKKAD CTITMADSDF LALMTGKMNP QSAFFQGKLK
   EEEESSTT E EETT     S EEEE  HHHH HHHHTTSS    SSTTTT
```



Why Care about Secondary Structure?

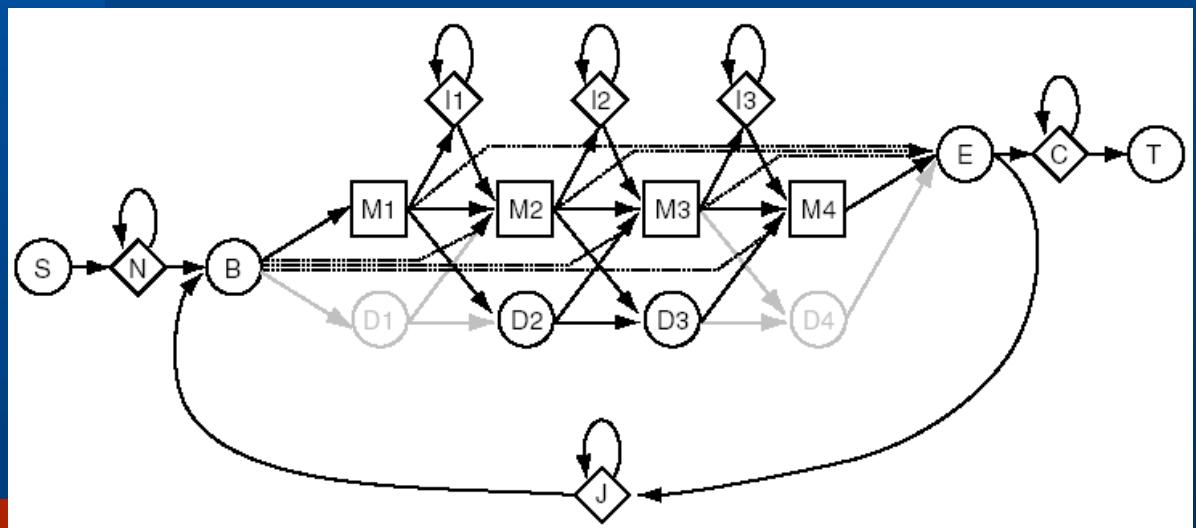
- Early stages of folding seem to involve nucleation around some secondary structures
- Possible to recognize fold class and many important structural features from SS alone
- Known secondary structure makes possible some tertiary structure prediction approaches
- Supports distinction between purely structural features and functional ones (e.g. active sites)
- Appears to be somewhat predictable from primary sequence.

Protein Structure Prediction with HMMs

- Most effective is homology modeling (Karplus)
  - Builds models of families with PDB structure
  - Uses "reverse null" model for log odds score, which reduces false positives from regions like amphiphthic helices which tend to match indiscriminantly
  - SAM & HMMer & Pfam
  - HMMstr
  - PSA
  - Signal Peptides: SignalP
  - Transmembrane Region: TMHMM , TMPRED

HMMer & PFAM – Related with SAM

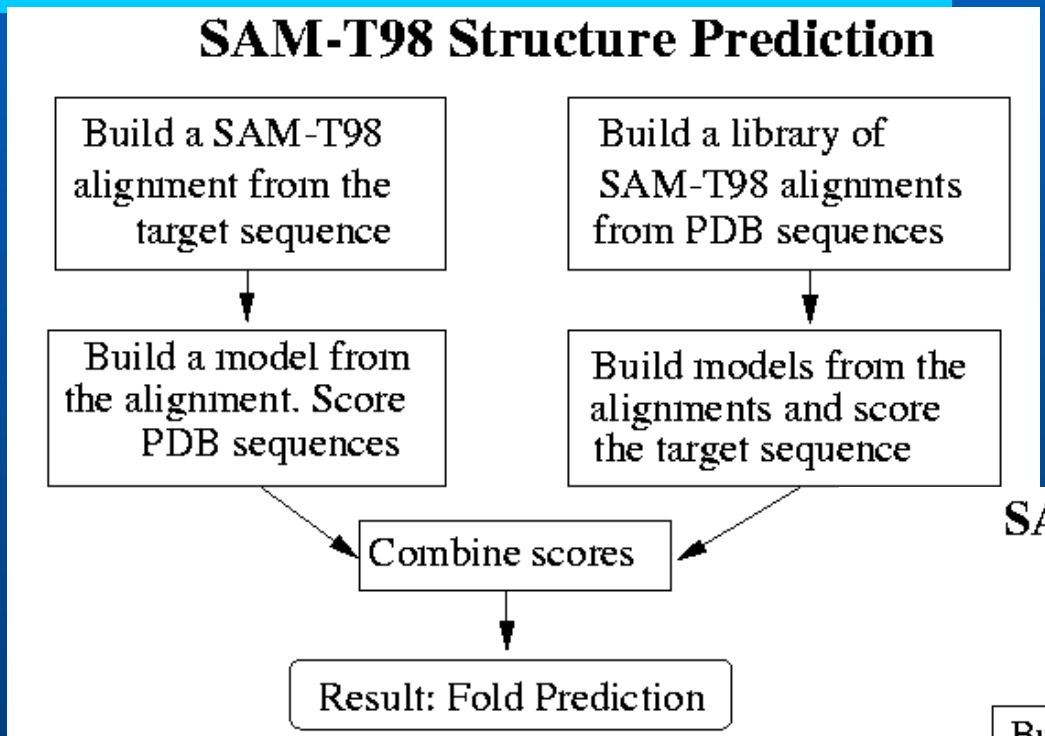
- HMMer is a tool for multiple alignment with HMM & to find motifs.
  - http://hmmer.wustl.edu/
- Pfam is a collection of protein families and domains. Pfam contains multiple protein alignments and profile-HMMs of these families. (focuses on "classical" domains with a high proportion of extracellular modules.) Pfam is constructed by HMMer
  - http://www.sanger.ac.uk/Software/Pfam/
- The HMMer 's database can be converted to SAM's.
- plan 7 model of HMMer

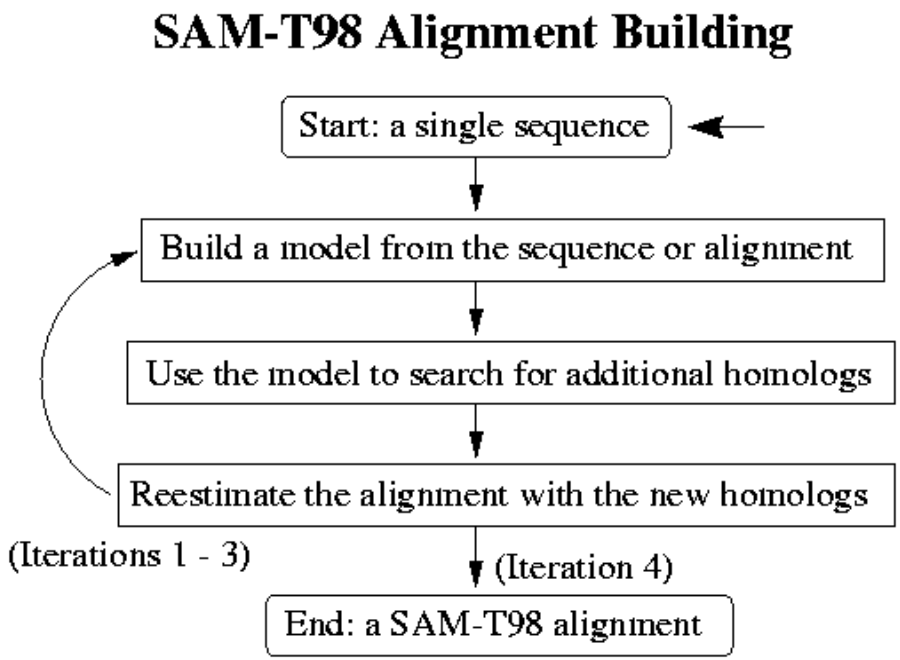


SAM

- http://www.cse.ucsc.edu/research/compbio/HMM-apps/
- The server has used UCSC's SAM-T98 method to create a library of HMMs, one per PDB structure (about 2500 HMMs total). You can search this database of HMMs with a protein sequence.
  - Compare Sequence Against Protein Model Library
  - Protein Query Against A Database
  - Tune Up a Multiple Alignments
  - Compare Two Alignments
  - Build SAM-T98 Alignment
  - Generate Weights for a Multiple Alignment
  - Build SAM-T98 HMM

Two Important Procedures of SAM



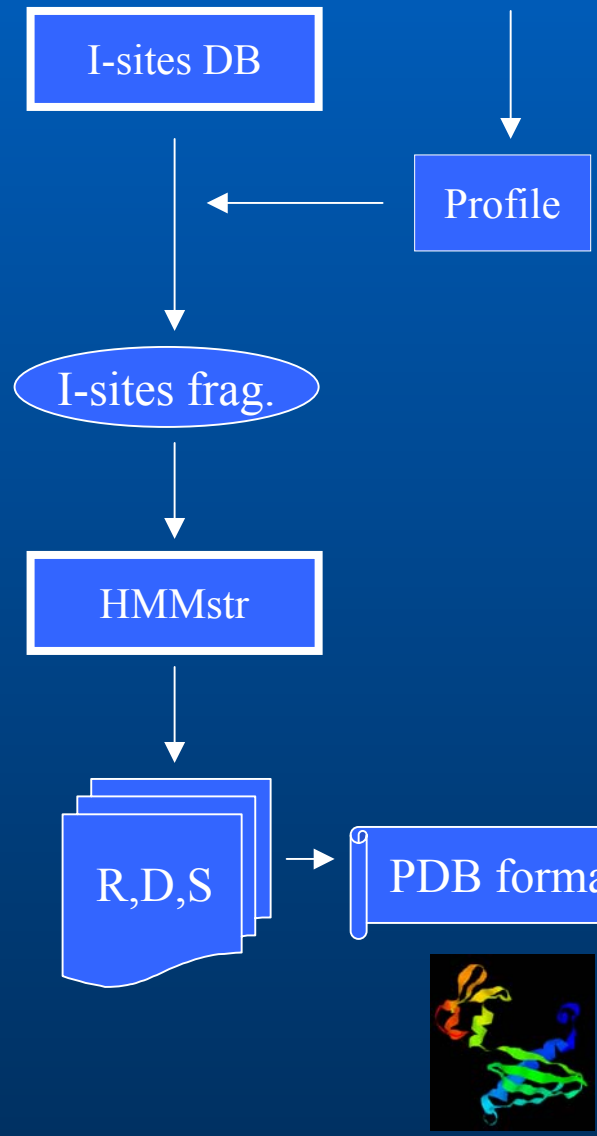


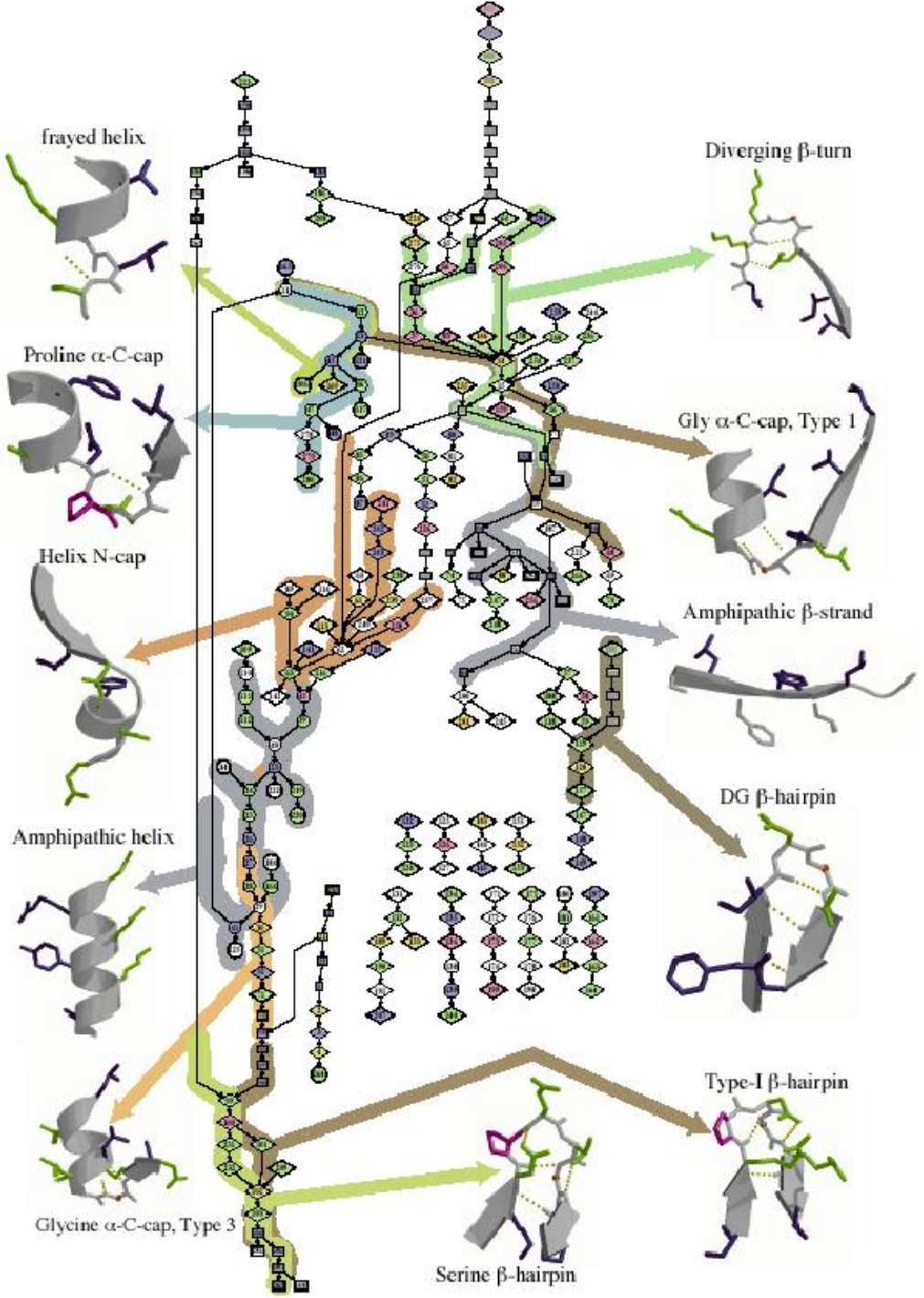
HMMstr

- An intereting approach to harder problems is the HMMstr "grammatical approach" to assembling sequences of local structural motifs
  - http://honduras.bio.rpi.edu/~isites/hmmstr/server.html
  - An interconnected sequence of HMM models for particular local structural motifs (such as hairpin turns or alpha helical n-terminal caps)
- Not world-beating predictive value, but an interesting approach, and generally competitive
- Shows the potentials for much complex structure in nested HMMs

HMMstr Procedure

- Running PSI-BLAST
  - Create sequence profile from alignment
  - Cannot make profile with only single sequence.
- Predicting I-sites
  - Find I-sites fragments
- Predicting backbone angles using I-sites fragments
- HMMstr prediction of sec. Struct and backbone angle
  - Create HMMstr-R, HMMstr-D, HMMstr-C
- Starting Rosetta
  - Create Tertiary Structure(PDB format)
  - PDB format can be shown with RasMol




frayed helix
Diverging β-turn
Proline α-C-cap
Gly α-C-cap, Type 1
Helix N-cap
Amphipathic β-strand
Amphipathic helix
DG β-hairpin
Type-I β-hairpin
Glycine α-C-cap, Type 3
Serine β-hairpin

Example of HMMstr

Model: HMMSTR-R, Trained on local structure (backbone angles)

Model: HMMSTR-D, Trained on secondary structure (three states)

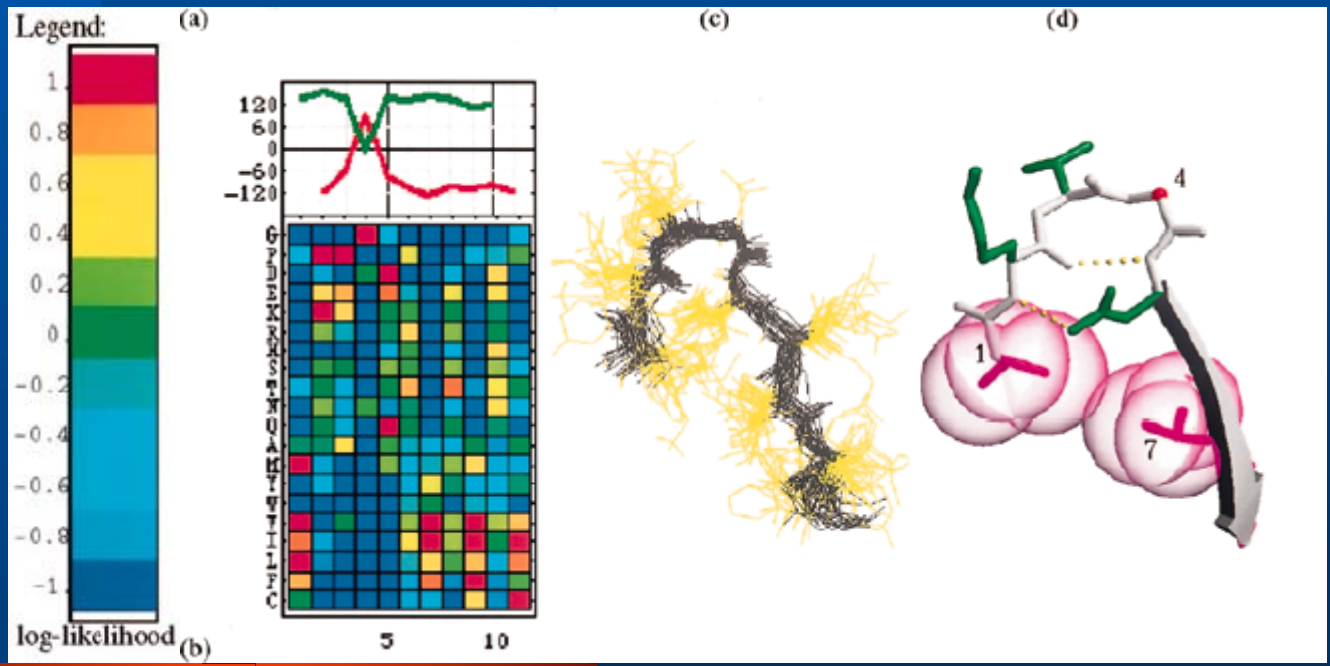
Model: HMMSTR-C,

```
   1          ....,....1....,....2....,....3....,....4....,....5
Seq           SSESDGFKSNLVFKEIEKKLEEEGEQFVKKIGGIFEFKVKDGPGGKESTW
Angles        BBHHBBEBHHHHHHHHHHHHHHHHHHHHHHHBBEEEEEEEBBBBBBEEEE
  confid      43333344556688899988887677666652467777765443355678
Sec struct    LLLLLLLLLLHHHHHHHHHHHHHHHHHHHLLEEEEEEELLLLLLLEEE
  confid      67778755444578888888775566655446657888765888865578
Context                                                nmmmmmnhhhhhhhnnm
  confid                                               44555554444444544

  51          ....,....6....,....7....,....8....,....9....,....0
Seq           TVDVKNGKGSVLPNSDKKADCTITMSDSDFLELMTGKMNPQSAFFQGKLK
Angles        EEEEBBHBEEEEEBBBBBBEEEEEEBHHHHHHHHHHHBBBBBEEEBEEEE
  confid      88776434467765444455677766666777776544433455643566
Sec struct    EEEEELLLLEEEELLLLLLEEEELLHHHHHHHHLLLLLLLLEELLEEE
  confid      98864887547656888765678757455777764556787544455455
Context       mmmmnhhhhnmnnddddddnmmn                        mmhhmmn
  confid      55555555544454444444555                        4466444

 101          ....,....1....,....2...
Seq           ITGNMGLAQKLQNLQKQPGNAKL
Angles        EEBBHHHHHHHHHHHHHHHHEE
  confid      553444567777777665444445
Sec struct    ELLLLHHHHHHHHHHLLLLLLE
  confid      44543345667766644565533
Context       m                     m
  confid      4                     4
```

Case of ‘1qnd’

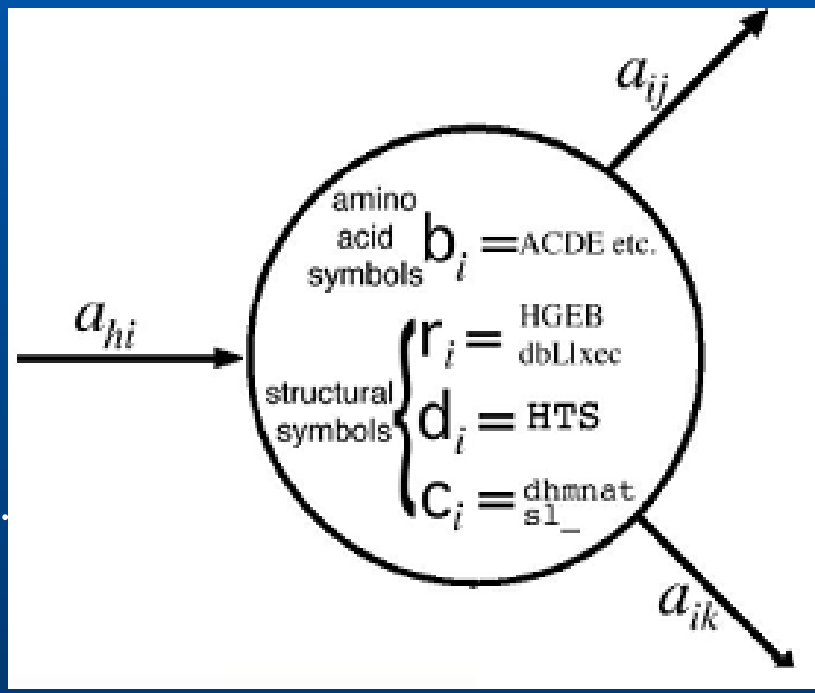
I-Sites

- I-sites: Invariant or Initiation sites
- I-sites library consists of an extensive set of short sequence motifs, length 3 to 19, obtained by exhaustive clustering of sequence segments from a non-redundant database of known structure



Markov State for HMMstr

- Each state emits an output symbol, representing sequence or structure
  - B: Corresponding to amino acids
  - R: Backbone angles
  - D: 3-state secondary structure
    - Helix, Strand , Turn
  - C:Strutural context
    - Hairpin, Diverging turn, Middle…



I-sites Clusters by Motif

I-sites clusters by motif (by number)

Choose a sequence/structure motif.
Click on the cluster identifier number to see the sequence profile and backbone angles.
Use your browser's FIND function to locate a cluster ID number.

Glycine-rich alpha-N-cap
3033 4018 7007 10011 10060 3029 8190 9001 9030 10026 10043 11074 4097

Diverging Type-II beta-turn, A
5116 10040 10094 6037 6488 7410 9024 9055 9128 11007 12018 12051 3004 6923 9019 9031

Glycine alpha C-cap, Type 2
8128 3257 9480 10523 11196 11648 12029 13022 15007 15079 15577 15610 3060 5050 5081 10912 11051 11506 12016 12908 12910 13907 15002 15025 15448 15905 5111

Serine alpha N-cap, type 2
5241 11105 13007

PG kink
5097 8300 11070

Polar alpha helix
3102

DG beta hairpin
3046 6040 7040 8069 9164 3010 7307 8188

Non-polar turn
4005

W loop
5171 9143

Diverging type-II turn, C
5073 6054

Miscellaneous motifs
10078 3111 3117 3133 3014 3016 3025 3030 3031 3051 3059 4103 4175 4179 5108 5123 8192

Glycine alpha-C-cap (Schellman) Type 1
10201 11151 11630 12013 12014 15006 15333 15585 15901 4227 7136 7812 11015 7135 11054 3069 4034 8020 8032 9013

Amphipathic alpha helix
11028 12046 13160 15106 15205 11176 10401 15548 13001 3023 5032 3043 4143 4026 4058 4064 6507 8055

Proline alpha C-cap
8165 9050 9172 10610 11024 11211 12021 13015 13027 13186 4008 11386 13028 15404 9075

DP alpha N-cap
6513 9022 11317 12019 12023 11047 5236 7256

Type-I beta hairpin
3009 3047 3148 9920 12031 13202 3027 6015 6111 7156 9102

Amphipathic beta strand
3116 3162 3020 4033 3260 4136 4190 5132 5036 6077 6355 8131 9174 8173

Serine beta hairpin
4137 7088 8093 8930 9931

Frayed alpha helix, type 2
4026 7038

Frayed non-polar helix
5252

Alpha loop, type 2
7015 8247 9060

Glycine alpha-C-cap (Schellman) Type 3
6283 7141 10034 11040 3106 8103 8162 3033

Serine alpha N-cap (N-capping box), type 1
13008 15353 15001 10554 12005 13010 15008 15015 15074 15286 7048 7124 8038 9021 9175 10095 11012 12001 12012 7248 11320 13005 13018 8150 9029

Alpha-alpha corner, type 1
4114 10063 11068 12010 13057 7111 10093 13066 3144

Alpha corner, type 2
7036 8157 11059

frayed alpha helix, type 1
13011 15004 15078 3077 5040 7037

Non-polar beta strand
3026 3035 3041 5057 5069 3034 3092 4021 6139 5107 6283 7428

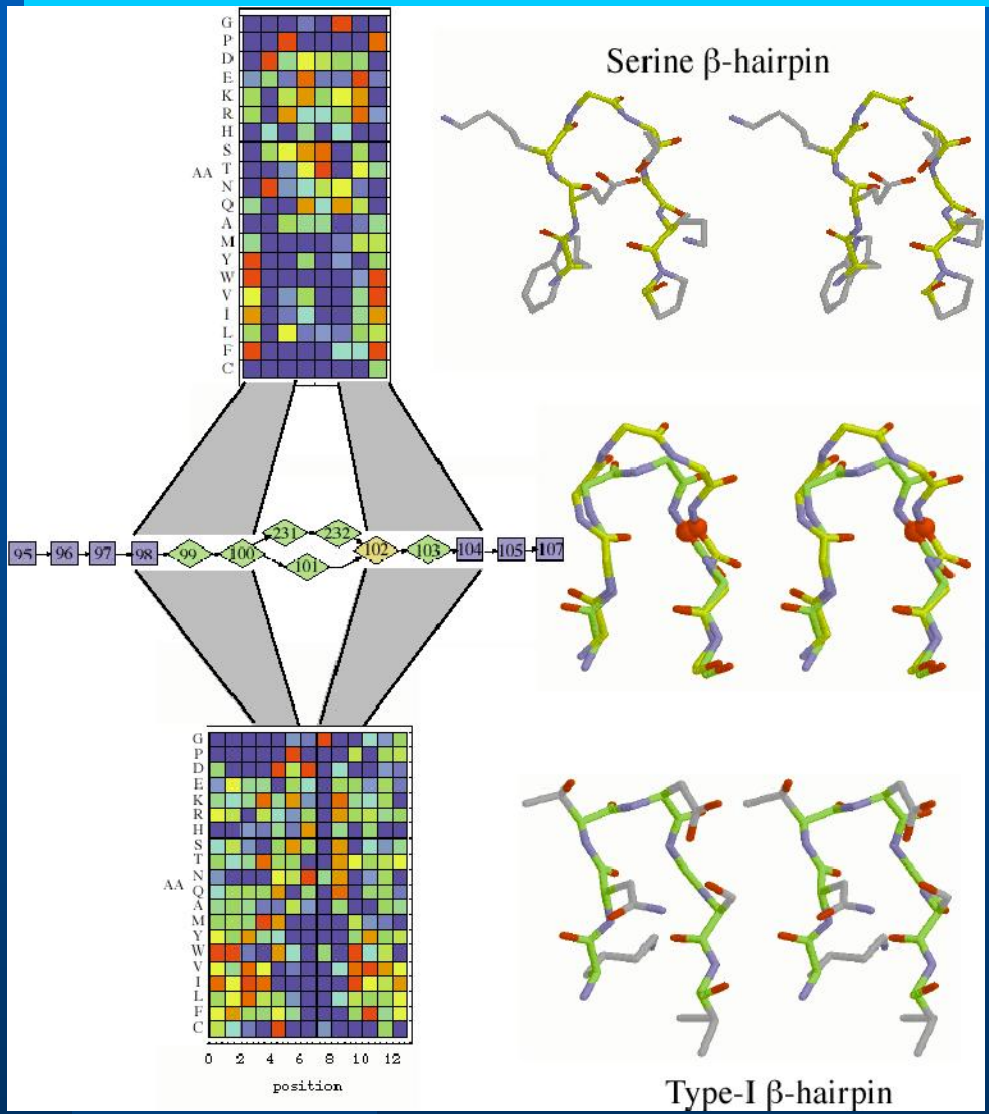
Non-polar alpha helix
4031 13025

Alpha loop, type 1
5015

Diverging type-II turn, B
5059 6220 8017 7010

Frayed alpha helix, type 3
8332

Merging of Two I-sites Motifs



- Shape of icon: Markov states
  - Rectangle: predominantly beta strand states
  - Diamond: predominantly turns
- Color of icon : a sequence preference
  - Blue: hydrophobic
  - Green: polar
  - Yellow: glycine
  - Etc

Example of HMMstr

Model: HMMSTR-R, Trained on local structure (backbone angles)

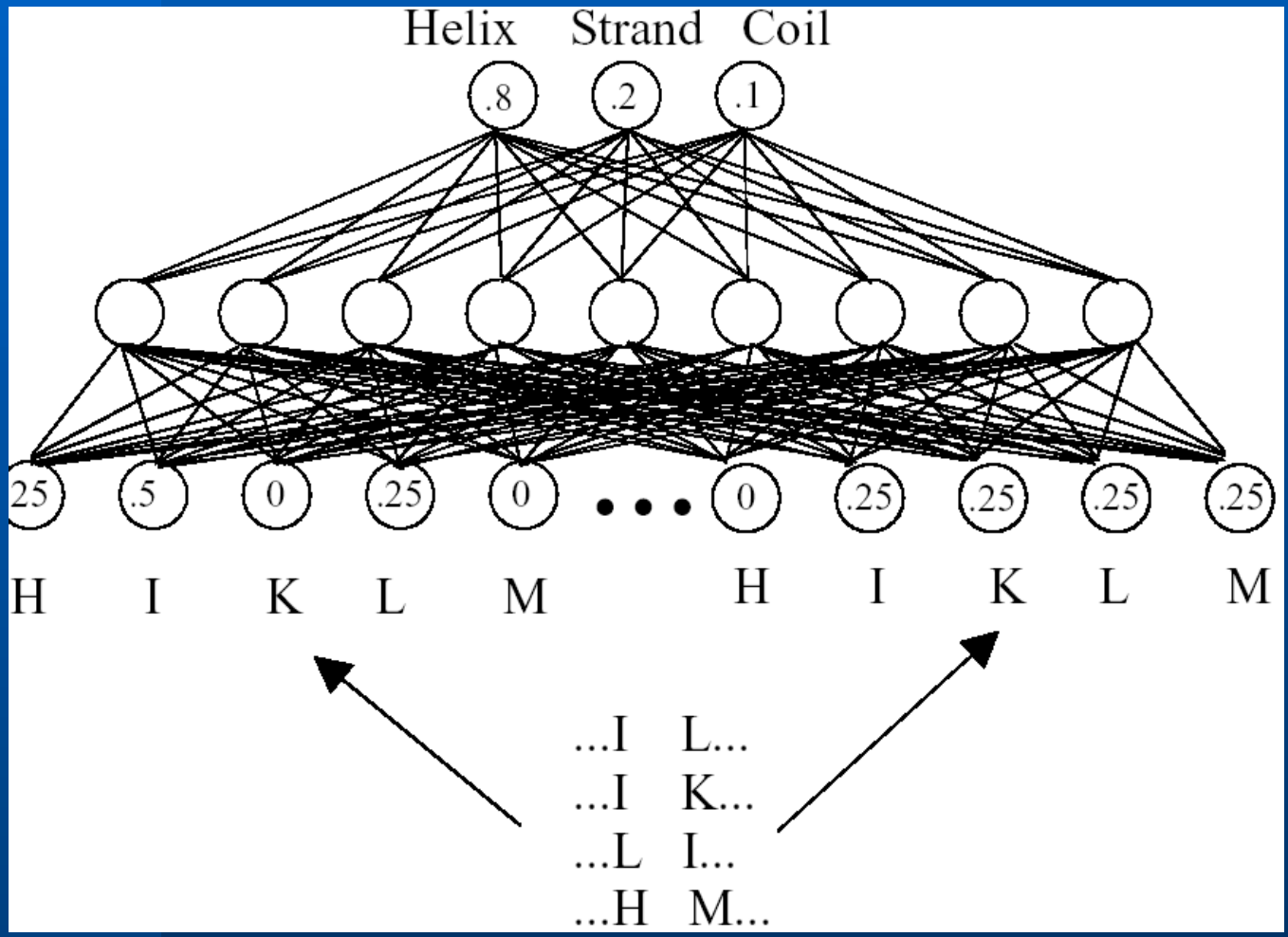
Model: HMMSTR-D, Trained on secondary structure (three states)

Model: HMMSTR-C,

```
   1          ....,....1....,....2....,....3....,....4....,....5
Seq           SSESDGFKSNLVFKEIEKKLEEEGEQFVKKIGGIFEFKVKDGPGGKESTW
Angles        BBHHBBEBHHHHHHHHHHHHHHHHHHHHHHHBBEEEEEEEBBBBBBEEEE
  confid      43333344556688899988887677666652467777765443355678
Sec struct    LLLLLLLLLLHHHHHHHHHHHHHHHHHHHLLEEEEEEELLLLLLLEEE
  confid      67778755444578888888775566655446657888765888865578
Context                                                nmmmmmnhhhhhhhnnm
  confid                                               44555554444444544

  51          ....,....6....,....7....,....8....,....9....,....0
Seq           TVDVKNGKGSVLPNSDKKADCTITMSDSDFLELMTGKMNPQSAFFQGKLK
Angles        EEEEBBHBEEEEEBBBBBBEEEEEEBHHHHHHHHHHHBBBBBEEEBEEEE
  confid      88776434467765444455677766666777776544433455643566
Sec struct    EEEEELLLLEEEELLLLLLEEEELLHHHHHHHHLLLLLLLLEELLEEE
  confid      98864887547656888765678757455777764556787544455455
Context       mmmmnhhhhnmnnddddddnmmn                        mmhhmmn
  confid      55555555544454444444555                        4466444

  101         ....,....1....,....2...
Seq           ITGNMGLAQKLQNLQKQPGNAKL
Angles        EEBBHHHHHHHHHHHHHHHHEE
  confid      553444567777777665444445
Sec struct    ELLLLHHHHHHHHHHLLLLLLE
  confid      44543345667766644565533
Context       m                     m
  confid      4                     4
```

Case of ‘1qnd’

Neural Network for Structure Prediction

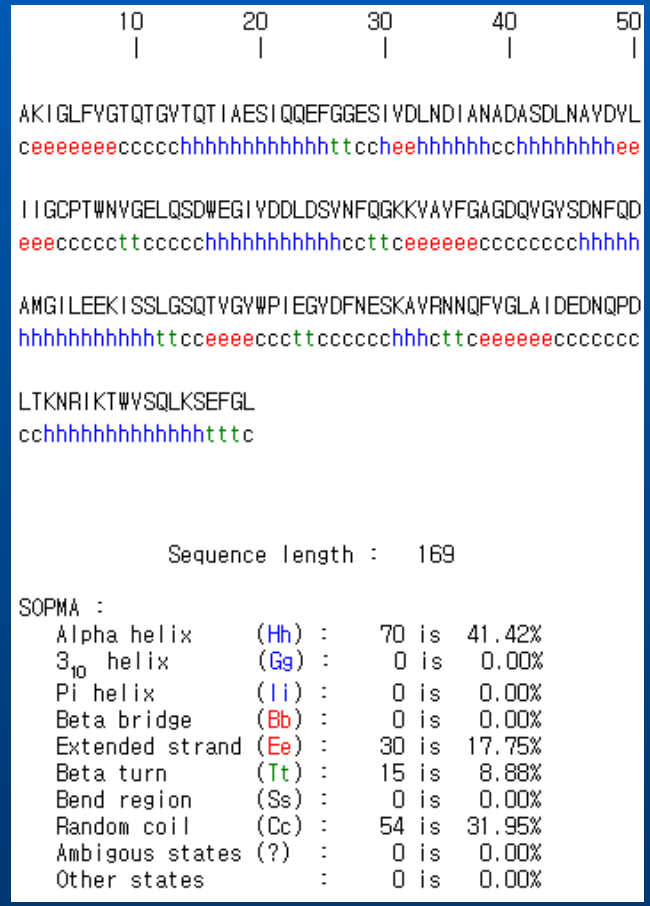


Other Non-HMM Secondary Structure Prediction Methods

- nnPredict: Using neural network
  - The nnPredict algorithm uses a two-layer, feed-forward neural network to assign the predicted type for each residue
  - Residues will be assigned as either being within a helix (H), strand (E) or neither (-). In case that no prediction can be made a "?" is returned to indicate that no confident assignment could be made.

```
Tertiary structure class: alpha/beta

Sequence:
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG
ELQSDWEGIYDDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYW
PIEGYDFNESKAVRNNQFVGLAIDEDNQPDLTKNRIKTWVSQLKSEFGL

Secondary structure prediction (H = helix, E = strand, - = no prediction)
----EEE------EEEHHHHHHH------EEEH--------------EEEE--------
---------------HHHH---EEEE-----------H--HHHHHHHH------E--E-
```

- Predator: using multiple sequence alignment
- PHD: using Neural Network and character of sequence
- PSIPRED: using Neural Network and Position-specific scoring Matrices(PSI-BLAST)
- JPred: using multiple sequence alignment and etc
- SOPMA
  - This self-optimized prediction method builds sub-databases of protein sequences with known secondary structures

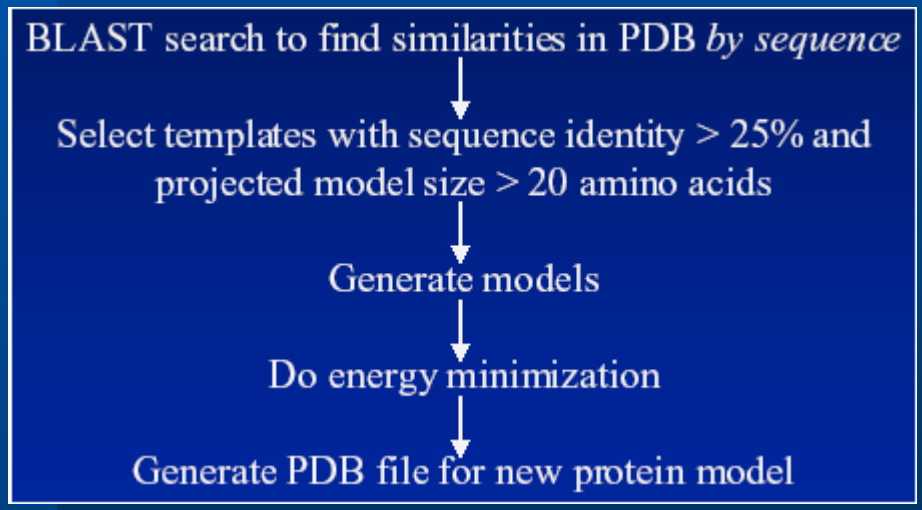


- PredictProtein: multi step predictive algorithm
  - Blast (for fast database search vs. SWISSPROT)
  - Maxhom (for multiple sequence alignment of similar sequences identified by BLAST)
  - ProSite (scanning for functional motifs) reported only if hit found
  - SEG (detection of composition-biased regions) reported only if more than 10 residues of low-complexity found
  - ProDom (scanning for the putative domain structure for your protein) reported only if hit found
  - Coils (prediction of coiled-coil regions) reported only if hit found
  - PHDsec (prediction of secondary structure)
  - PHDacc (prediction of solvent accessibility)
  - PHDhtm (prediction of transmembrane helices and their topology) reported only if hit found

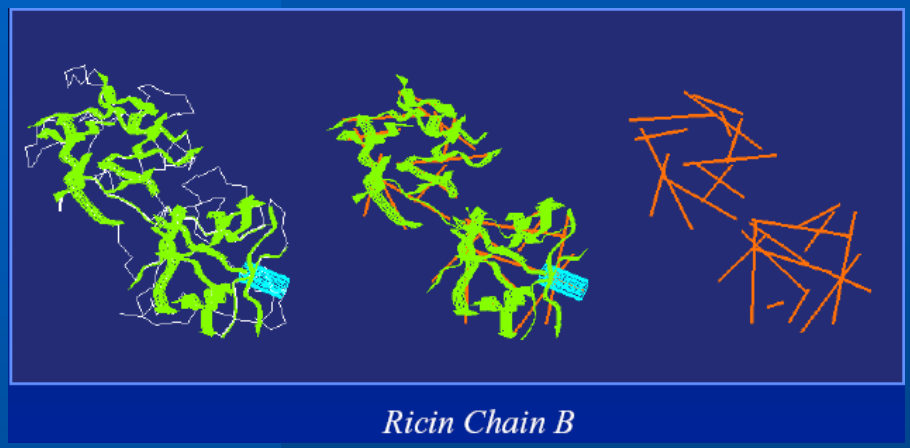
```
          ...,....1....,....2....,....3....,....4....,....5....,....6
AA        |AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG
PHD sec   | EEEEEEE    HHHHHHHHHHHHH     EEEEE HHH HHHH    EEEEE
Rel sec   |938999736982489999999997679824432132412786312419998615477 65
Detail:
prH sec   |000000000014689999999999821000011112565388764321000001111111
prE sec   |058998852000000000000000000036655421000000000148998741200 02
prL sec   |931000137985310000000000178985222344324511234554000114667776
```

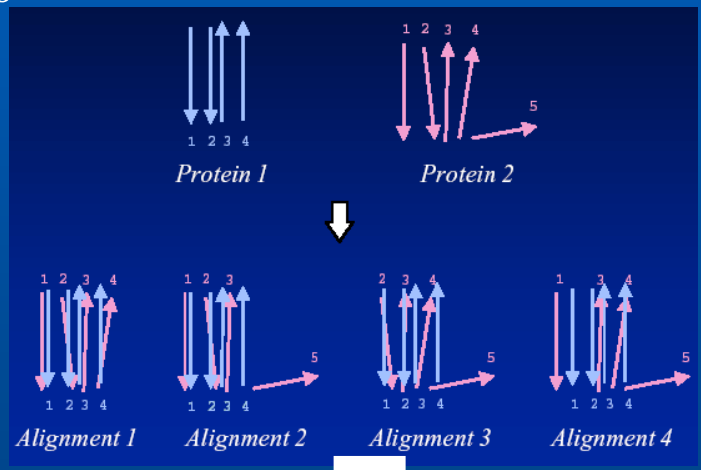
Other Non-HMM Tertiary Structure Prediction Methods

- Swiss-Model
  - Automated Protein Modeling Server
  - A free service that generates a PDB coordinate file of your protein sequence of interest
  - Methods of operation

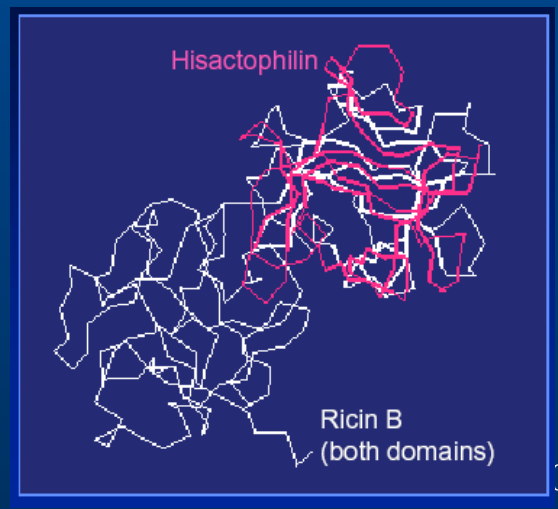


- VAST -Vector Alignment Search Tool finding "structure neighbors"
  - Step 1: Construct vectors for secondry structure elements
  - Step 2: optimally align structural vectors





  - Step 3: Refine residue-by-residue alignment using Monte Carlo



Summary

- AI and machine learning techniques have been successfully applied to bioinformatics
  - Neural networks
  - Decision trees
  - Hidden Markov models
- Typical examples include the following biological sequence and structure analysis problems
  - Gene finding
  - Promoter prediction
  - Protein structure prediction

Acknowledgements

Je-Gun Joung

Sung-Wook Chi

Cheol Han

- More information at

http://cbit.snu.ac.kr/

(SNU Center for Bioinformation Technology)

http://bi.snu.ac.kr

(SNU Biointelligence Laboratory)