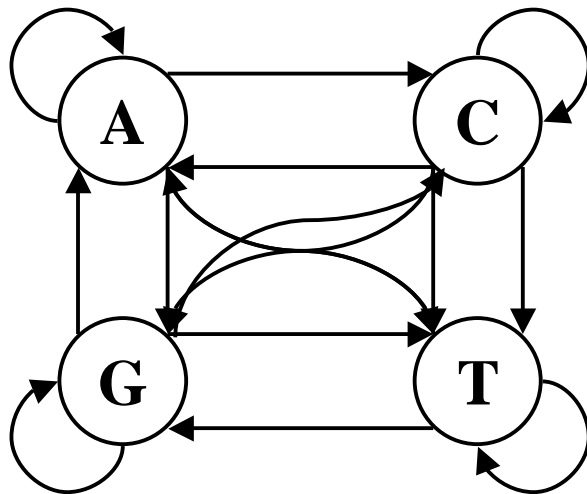


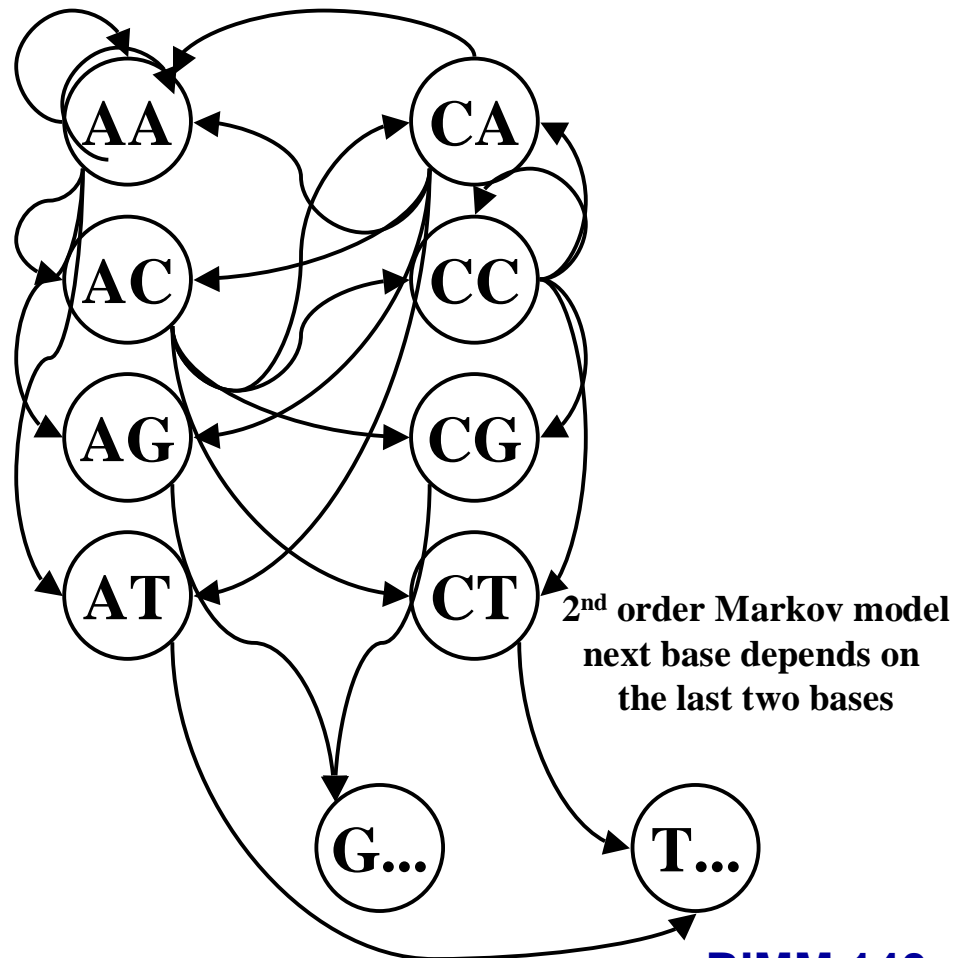
# Gene Modeling

## Search by Content

### Markov Models



1<sup>st</sup> order Markov model  
next base depends only  
on the current base



2<sup>nd</sup> order Markov model  
next base depends on  
the last two bases

# Gene Modeling

---

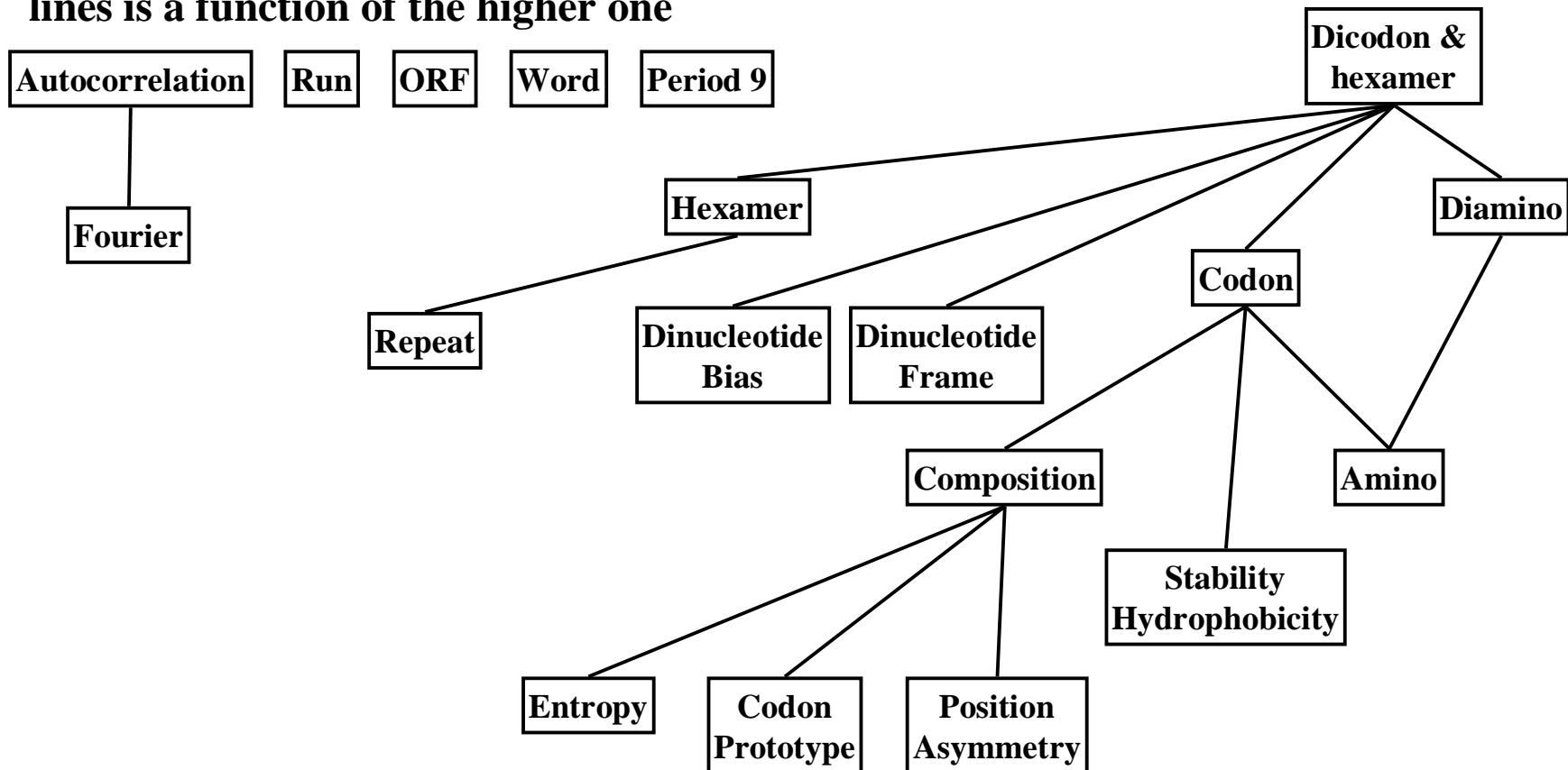
## *Search by Content*

- **Other methods (see Fickett and Tung, “Assessment of protein coding measures”, *Nucleic Acids Res.* 20, 6441-6450, 1992)**
  - Open reading frame (genes have longer ORF)
  - Runs ( R&Y stronger in noncoding, W&S in coding)
  - N-word counts, most commonly hexamer
  - Stability (tendency to mutate to same amino acid residue)
  - Other base asymmetry measures
  - Periodicity, such as period 9
  - Global patterns (GC content, CpG islands)

# Gene Modeling

## Search by Content

- Derivability of measures (from Fickett and Tung), lower function connected by lines is a function of the higher one



# *Gene Modeling*

---

## *Search by Content*

- **Methods that do not require a model**
  - Uneven base positional base frequencies
  - Testcode (Fickett)
- **Methods that require a model**
  - GC content
  - Codon usage/Codon preference
  - Hexamer statistics
  - Specific words

# *Gene Modeling*

---

## *Search by signal*

- **Splice donor/acceptor sequences**
- **Poly-A**
- **Cap signal**
- **ATG**
- **promoter elements**
  - -10,-35
  - TATA, CAAT, etc.
- **Usually recognized using weight matrix approach**

# Gene Modeling

---

## Search by site

- Eukaryotic transcription initiation site

	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
A	16	4	90	1	91	69	92	57	40	14	21	21	21	17	20
C	37	12	0	2	0	0	1	1	11	35	38	33	30	28	26
G	39	5	1	1	1	0	5	11	40	39	33	33	33	36	36
T	8	79	9	96	8	31	2	31	9	12	8	13	16	19	18

-----  
G T A T A A A A G G C G G G G

# Gene Modeling

---

## *Search by Site*

- **CCAAT-box**
  - Y Y Y R R C C A W W S R -212 .. -57
- **GC-box**
  - W R K R G G Y R K R K Y Y K -164 .. +1
- **cap-site**
  - K C W K Y Y Y Y +1 .. +5
- **Information about composite regulatory elements, transcription factors and eukaryotic promoters are collected in the following databases:**
  - TRANSFAC (Wingender et al., 1996).
  - TFD (Ghosh, 1993)
  - TRRD (Kel et al., 1995b)
  - COMPEL database (Kel et al., 1995a).
  - EPD (Bucher, 1988),

# Gene Modeling

---

## Search by Site

- Eukaryotic translation initiation site

	-6	-5	-4	-3	-2	-1	+1	+2	+3
A	18	19	24	68	23	15	100	0	0
C	21	40	58	2	55	53	0	0	0
G	47	23	12	30	16	23	0	0	100
T	13	18	6	0	7	9	0	100	0
	G	C	C	A	C	C	A	T	G

# Gene Modeling

---

## *Search by Site*

- Splice sites
- **The splicing of introns is part of a multi step process of RNA maturation which takes place in the nucleus to generate mature mRNA molecules for transport to the cytoplasm. This process involves several factors such as snRNP (small nuclear ribonucleoprotein particles) and hnRNPs (heterogeneous nuclear ribonucleoprotein particles). This complex assembly is called the spliceosome.**
- **It has been found that introns usually begin with GU (donor splice site) and end with AG dinucleotides (acceptor splice site).**
- **The branch point signal is located within the range of 10-50 bases upstream from the acceptor splice site (the lariat region).**



# Gene Modeling

---

## Search by Site

- Branch point signal

A	1	0	39	99	11		
C	76	8	15	1	45		
G	2	0	42	0	6		
T	21	91	4	0	38		
	C	T	G	A	C	A	T

# Gene Modeling

---

## *Search by Site*

- **Polyadenylation site**
- **Polyadenylation (cleavage of pre-mRNA 3' end and synthesis of poly-(A) tract) is a very important early step of pre-mRNA processing.**
- **Sites**
  - AATAAA, located 15-20 nucleotides upstream from the poly-(A)
  - ATTAAA, is nearly as active as the canonical sequence.
  - An additional signal with consensus YGTGTTY (diffusive GT-rich sequence) was revealed in region from 20 to 30 nucleotides downstream of poly-(A) site (site of cleavage) (McLauchlan et al., 1985).

# *Gene Modeling*

---

## *Search by Sites*

- **Identifying sites**
- **Log-odds matrix / window analysis**
- **Hidden Markov Model**
- **Neural network**

# *Gene Modeling*

---

- **GRAIL uses a combination of search by content and search by signal approaches to produce a complete gene model based on genomic DNA sequence**
- **GRAIL uses a neural net approach to combine information from a variety of "sensors" or indicators. Because of the neural net training, it doesn't matter too much if the sensors are highly correlated, they just get lower weights in the final prediction.**
- **System is trained on real genes from a specific organism**

# Gene Modeling

---

- **Hexamer in-frame, candidate region and 60 bases left and right**
  - Isochore, candidate region
  - Left and right regions are presumably introns, if they have high coding scores, it may indicate that the candidate region is too small.
- **Markov chain model (high AT and high GC models)**
- **Isochore GC content**
- **Exon GC content**
- **Coding region length profile**
  - Candidates with lengths corresponding to common lengths get higher scores
- **Candidate region length**

# *Gene Modeling*

---

## *Splice donor site*

- **Splice acceptor site**
- **Intron vocabulary (2 methods)**
  - Isochore, candidate region
  - Search for "words" that are common in introns but not exons

# Gene Modeling

---

- **Find candidate regions with specific edge signals, i.e. splice junctions**
- **Evaluate coding potential for all sensors**
- **Predict coding region using neural net**
- **Assemble gene model**
  - 1st coding region starts with ATG
  - last coding region ends with inframe stop codon
  - adjacent coding regions maintain translation frame
  - distance must be at least minimum intron size
  - Uses dynamic programming to optimize combination of coding regions

# *Gene Modeling*

---

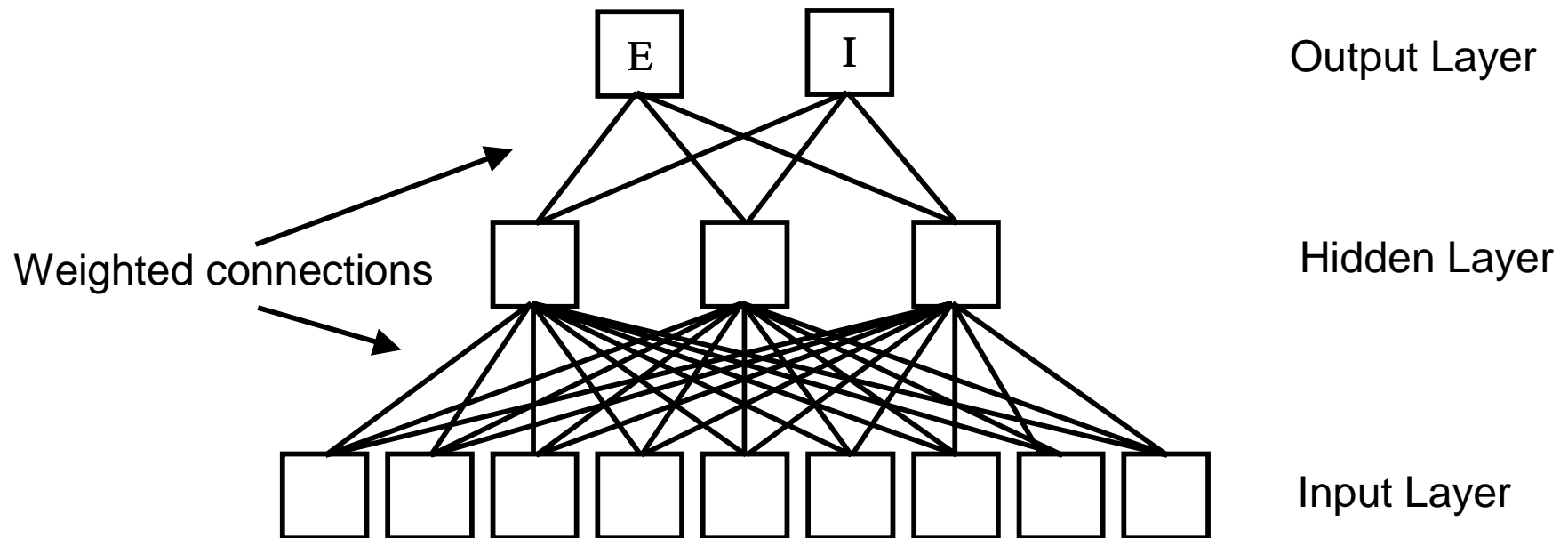
## *Neural Net*

- **Neural net methods provide essentially a black box method for making predictions based on a set of training data. Neural nets explicitly consider interactions between the various inputs - interactions that may be very complex.**
- **Each connection in the net has a weight associated with it.**
- **During training, weights are iteratively adjusted so that the prediction agrees with the training data. This process is typically known as back propagation.**
- **Weight matrices can be looked at as a neural net with no hidden layers (also called a perceptron).**

# Gene Modeling

## Search by Site

- **Artificial Neural Network (ANN)**
- **Must be trained with classified data**



# *Gene Modeling*

---

## *Other things GRAIL does*

- **Attempts to correct indels**
- **Detects CpG islands (frequently found at 5' end of genes)**
- **Promoters**
- **Poly-A sites**
- **Repetitive DNA**
- **Protein sequence searching**

# *Gene Modeling*

---

- **GRAIL II uses candidate region approach described above**
- **Many other systems**
  - Glimmer
  - Genemark (mostly for bacterial genomes)
  - GeneParser
  - GeneID
  - FGeneH
  - GenLang

# Gene Modeling

---

## Limitation

<i>Gene prediction method</i>	<i>Limitation</i>
<b>Ab initio (Hidden Markov Model (HMM)-based) methods</b>	<b>Poor sensitivity and specificity, leading to whole genes or exons being missed or wrongly predicted</b>
<b>Similarity to existing expression sequence tags (ESTs)</b>	<b>Contaminating ESTs derived from unspliced mRNA, genomic DNA and nongenic transcription</b>
<b>Similarity to existing gene/proteins</b>	<b>Unable to distinguish pseudogenes (non-protein coding) and novel genes undetected</b>
<b>Current approaches result in</b>	<b>Partial genes, fragmented genes, gene fusions and spurious predictions</b>