CPS260/BGT204.1 Algorithms in Computational Biology **October 28, 2003**

Lecture 17: Gene Prediction

Lecturer: Pankaj K. Agarwal *Scribe: Haige Shen*

The problem of finding genes coding for a protein sequence presents different types of difficulties between eukaryotes and prokaryotes. For prokaryotes, there are no introns and intergenic regions are small, but genes may often overlap each other and the translation starts are difficult to predict correctly. In this lecture, all the discussions on gene prediction are for eukaryotes.

Gene/Transcribed region • A eukaryotic gene is defined as being composed of a transcribed region and of regions that *cis*-regulate the gene expression. The currently existing gene prediction software look only for the transcribed regions of genes, which is then called "the genes".

Exon • A coding area in a gene. Some exons may be non-coding.

Intron • A non-coding area in a gene. Introns are removed during the splicing machanism that leads to the mature mRNA.

UTRs • Untranslated Terminal Regions. The non-coding transcribed regions, which are located upstream of the translation initiation (5'-UTR) and downstream of the translation stop (3'-UTR).

Intergenic region • The region between two transcribed regions.

Promoter • A region sitting in the intergenic region, immediately upstream of the gene and not overlapping with it.

Currently, two essencial types of information are used to predict genes in a genome sequence: content sensors and signal sensors.

17.1 Content Sensors

Content sensors are measures that try to classify a DNA region into types, for example, coding regions versus non-coding regions. They can be further classified as extrinsic content sensors and intrinsic content sensors.

Extrinsic content sensors Extrinsic content sensors are exploited from the training based on unbiased sets of known coding regions. Sufficient similarities between a genomic sequence region and protein or DNA sequences in a database are detected by using local alignment methods, such as Smith-Waterman algorithm, FASTA and BLAST, to determine whether it is a transcribed and coding region.

Similarities with three types of sequences may provide information about exon/inton locations.

- Protein sequences. First and most widely used.

- cDNA sequences.
- Genomic DNA. This is based on the assumption that coding sequences are more conserved than non-coding ones.

Poor quality of databases and lack of guarantee to find a similar sequence weaken the extrinsic approaches. Moreover, insufficient accuracy and missing small exons are also the problems of such approaches.

Intrinsic content sensors Intrinsic content sensors are used to discriminate coding from non-coding regions in eukaryotes based on statistical differences between the sequence composition of these two types of regions. Many measures of coding potential have been investigated. Some measures are widely used and proven to be effective, for instance, the nucleotide composition (G/C rich and A/T rich), codon composition, hexamer frequency, and base occurrence periodicity. Additionally, the measure of CpG islands, which are regions that often occur in promotors where the frequency of the dinucleotide CG is not as low as it typically is in the rest of the genome, is informative for gene detection. Combinations of several such measures can be effective.

Markov models in intrinsic content sensoring Markov Models are widely used to find coding areas in nucleotide sequences. Figure 17.1 shows nucleotide sequence Π.

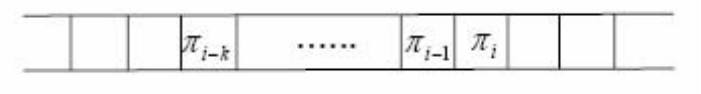

Figure 17.1: A nucleotide sequence

0th-order Markov model: The probability of appearance of a given base (A, T, G or C) at a given position π_i is independent of the any previous nucleotides, i.e.,

$$P(\pi_i|\Pi) = P(\pi_i).$$

1st-order Markov model: The probability of π_i is only dependent on the previous base, i.e.,

$$P(\pi_i|\Pi) = P(\pi_i|\pi_{i-1}).$$

k*th-order Markov model:* The probability of π_i is dependent on the previous *k* bases, i.e.,

$$P(\pi_i|\Pi) = P(\pi_i|\pi_{i-1}, \pi_{i-1}, ..., \pi_{i-k}).$$

Complex three-periodic Markov models (Figure 17.2) have been introduced to characterize coding sequence in most gene prediction programs, such as GeneMark and GeneScan. In these models, coding regions are defined by three Markov models, one for each position inside a codon.

The 5th order Markov model is applied to exploiting hexamer composition. In this model, two probability tables, one for coding regions and one for non-coding regions, are prepared in advance. Each table will be of

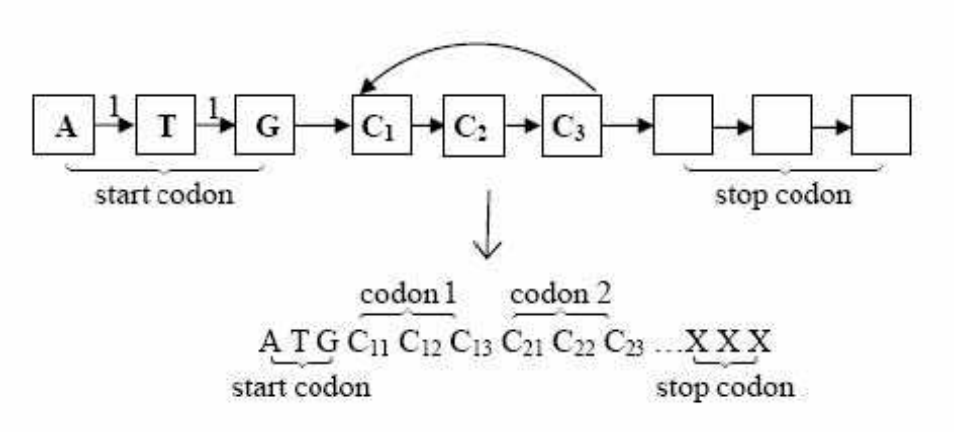

Figure 17.2: Three-periodic Markov Model

size 4^6. A window of 6-base width slides along a nucleotide sequence, so that the probability of observing the sixth base given the 5 preceding bases is registered. By refering to the two probability tables, the likelihood of this sequence being a coding region is estimated. This model is hard to detect donor and acceptor sites when exons are too short.

Usually the larger the order of Markov model, the finer it can characterize the coding regions. However, it requires a larger number of known coding sequences for reliable estimation.

17.2 Signal Sensors

Signals are specific functional sites which are inside or at the boundaries of the various genomic regions and involved in the various levels of protein encoding gene expression. For example, transcription factor binding sites, TATA boxes, donor sites, acceptor sites, branch points, poly-A site, initiation site (generally ATG with exceptions), stop codons. Figure 17.3 shows the structure of a gene with some of the signals marked.

The existing approaches to finding a signal include multiple alignment-based consensus sequence matching method, 0th-order Markov model-based positional weight matrix (PWM) method, and higher order Markov model-based weight array model (WAM), etc.. Table 1 (Figure 17.4) lists currently available splice site detection programs and the methods they used.

Here we describe GenScan model, which was developed by Burge and Karlin [Burge1997] using the signal sensors. Figure 17.5 illustrates the GenScan Model of genome sequence structure. Each circle or diamond represents a functional unit of a gene or genomic region. E_{sngl} is a single-exon (intronless) gene (translation start codon to stop codon). E_{init} is the initial exon (translation start codon to donor site). $E_k(0 \leq k \leq 2)$ is a phase k internal exon (acceptor site to donor site). $I_k(0 \leq k \leq 2)$ is a phase k intron, i.e., a intron starting after the jth base of codon. Internal exons are similarly divided according to the phase of the previous intron, which determines the codon position of the first base-pair of the exon, hence the reading frame. Donor and acceptor sites, translation and termination signals are considered as part of the forward exon. This GenScan

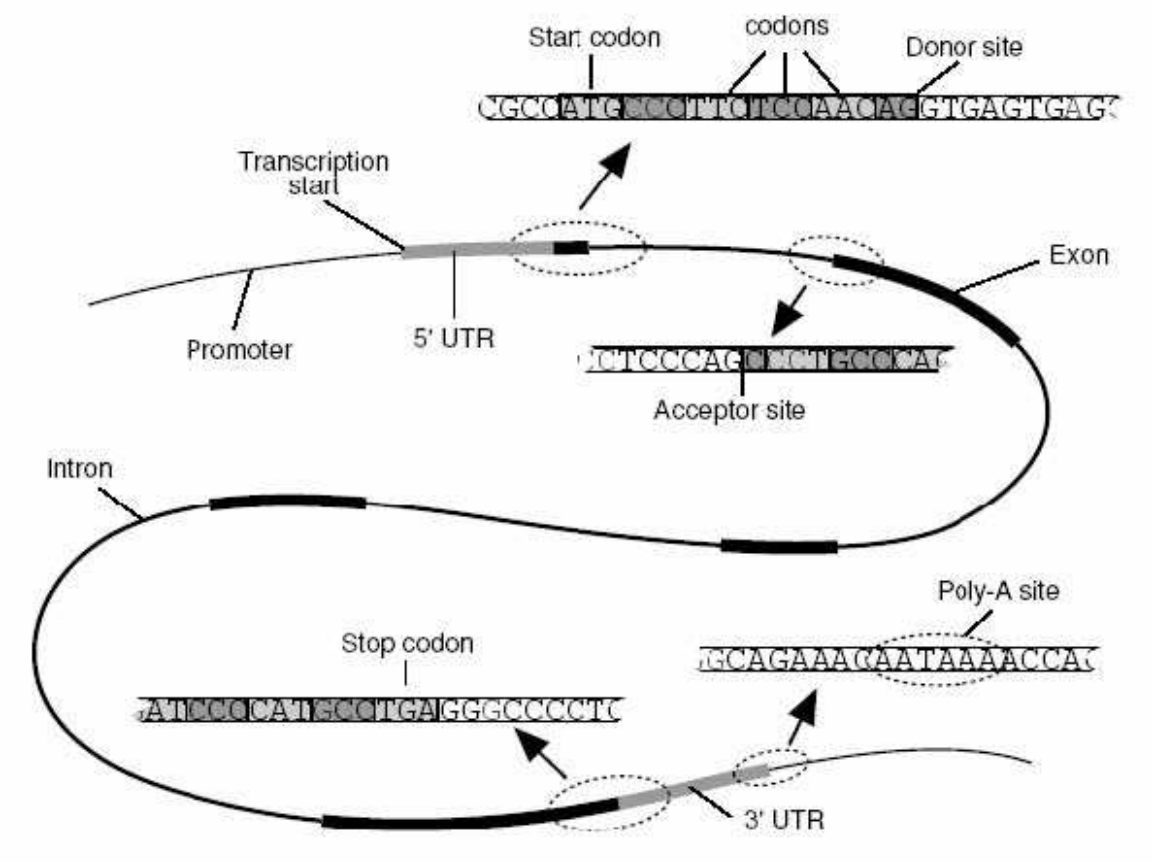

Figure 17.3: The Structure of a gene with some of the important signals shown [Krogh1998]

Table 1. Splice site prediction programs

Program	Organism	Method
GeneSplicer (152)	*Arabidopsis*, human	HMM + MDD
NETPLANTGENE (42) (http://www.cbs.dtu.dk/services/NetPGene/)	*Arabidopsis*	NN
NETGENE2 (43) (http://www.cbs.dtu.dk/services/NetGene2/)	Human, *C.elegans*, *Arabidopsis*	NN + HMM
SPLICEVIEW (39) (http://l25.itba.mi.cnr.it/~webgene/wwwspliceview.html)	Eukaryotes	Score with consensus
NNSPLICE0.9 (44) (http://www.fruitfly.org/seq_tools/splice.html)	*Drosophila*, human or other	NN
SPLICEPREDICTOR (40,153) (http://bioinformatics.iastate.edu/cgi-bin/sp.cgi)	*Arabidopsis*, maize	Logitlinear models: (i) score with consensus; (ii) local composition
BCM-SPL (http://www.softberry.com/berry.phtml; http://genomic.sanger.ac.uk/gf/gf.html)	Human, *Drosophila*, *C.elegans*, yeast, plant	Linear discriminant analysis

HMM, hidden MM; MDD, maximal dependence decomposition; NN, neural networks.

Figure 17.4: Table 1 [Mathe2002]

model is generally formulated as an non-homogeneous Hidden Markov Model.

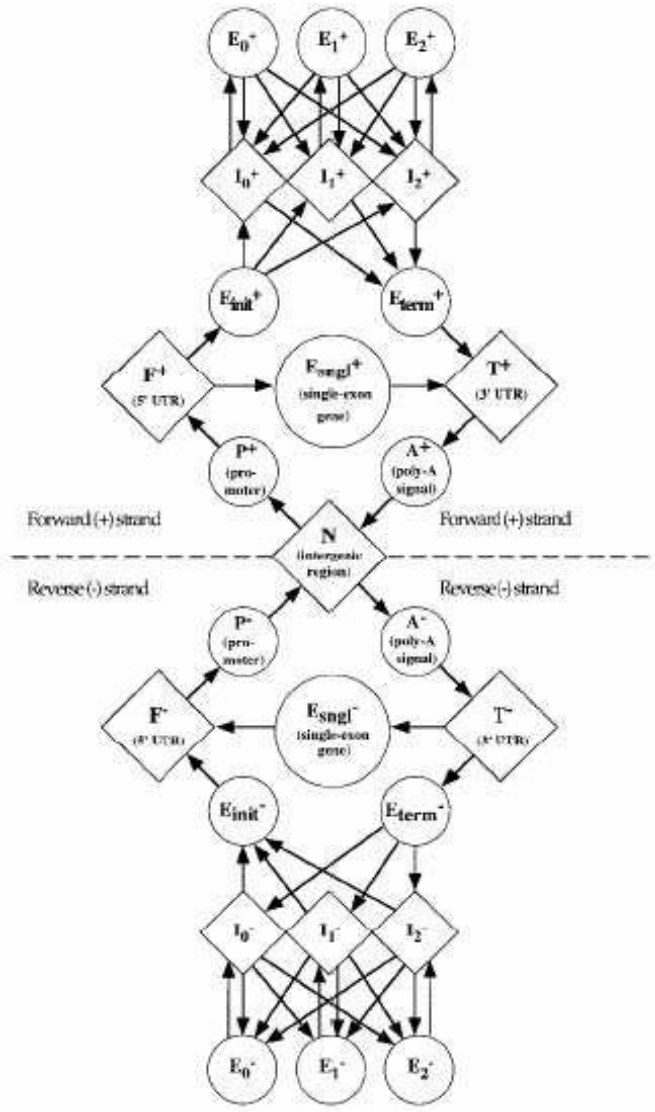

Figure 17.5: GenScan model

Non-homogeneous HMM First, since gene density and gene length are strongly related to CpG content of a genomic sequence as shown in Table 2 (Figure 17.6), separate initial and transition probability distributions are estimated for sequences in each categorie of the CpG content ($< 43\%, 43\% - 51\%, 51\% - 57\%, > 57\%$).

Second, the states of the model correspond to sequence segments of highly variable length. Exon length is

Table 2. Gene density and structure as a function of C + G composition: derivation of initial and transition probabilities

Group	I	II	III	IV
C + G% range	<43	43-51	51-57	>57
Number of genes	65	115	99	101
Est. proportion single-exon genes	0.16	0.19	0.23	0.16
Codelen: single-exon genes (bp)	1130	1251	1304	1137
Codelen: multi-exon genes (bp)	902	908	1118	1165
Introns per multi-exon gene	5.1	4.9	5.5	5.6
Mean intron length (bp)	2069	1086	801	518
Est. mean transcript length (bp)	10866	6504	5781	4833
Isochore	L1 + L2	H1 + H2	H3	H3
DNA amount in genome (Mb)	2074	1054	102	68
Estimated gene number	22100	24700	9100	9100
Est. mean intergenic length	83000	36000	5400	2600
Initial probabilities:				
Intergenic (N)	0.892	0.867	0.540	0.418
Intron (I_0^+, I_1^+, I_2^+, I_0^-, I_1^-, I_2^-)	0.095	0.103	0.338	0.388
5′ Untranslated region (F^+, F^-)	0.008	0.018	0.077	0.122
3′ Untranslated region (T^+, T^-)	0.005	0.011	0.045	0.072

The top portion of the Table shows data from the learning set of 380 genes, partitioned into four groups according to the C + G% content of the GenBank sequence; the middle portion shows estimates of gene density from Duret *et al.* (1995) for isochore compartments corresponding to the four groups above; the bottom portion shows the initial probabilities used by GENSCAN for sequences of each C + G% compositional group, which are estimated using data from the top and middle portions of the Table. All of the values in the top portion are observed values, except the proportion of single-exon genes. Since single-exon genes are typically much shorter than multi-exon genes at the genomic level (due to the absence of introns) and hence easier to sequence completely, they are probably substantially over-represented in the learning set relative to their true genomic frequency; accordingly, the proportion of single-exon genes in each group was estimated (somewhat arbitrarily) to be one half of the observed fraction. Codelen refers to the total number of coding base-pairs per gene. Data for subsets III and IV are estimated from the Duret *et al.* (1995) data for isochore H3 assuming that one-half of the genes and 60% of the amount of DNA sequence in isochore H3 falls into the 51 to 57% C + G range. Mean transcript lengths were estimated assuming an average of 769 bp of 5′UTR and 457 bp of 3′UTR per gene (these values derived from comparison of the "prim_transcript" and "CDS" features of the GenBank annotation in the genes of the learning set). To simplify the model, the initial probabilities of the exon, polyadenylation signal and promoter states are set to zero. All other initial probabilities are estimated from the data shown above, assuming that all features are equally likely to occur on either DNA strand. The initial probability for all intron states was partitioned among the three intron phases according to the observed fraction of each phase in the learning set. Transition probabilities were estimated analogously.

Figure 17.6: Table 2 [Burge1997]

critical to splicing. Very small (< 50bp) are skipped by the splicing matches, while very long ($>\sim 300$bp) exons could be not handled by the splicing machinery. With this consideration, seperate length distribution functions are derived empirically for initial, internal, and terminal exons and for single-exon genes. Intron and intergenic lengths are modeled as geonetric distributions with q estimated for each CpG group separately. Figure 17.7 shows the length distributions for the exons and introns of the learning dataset.

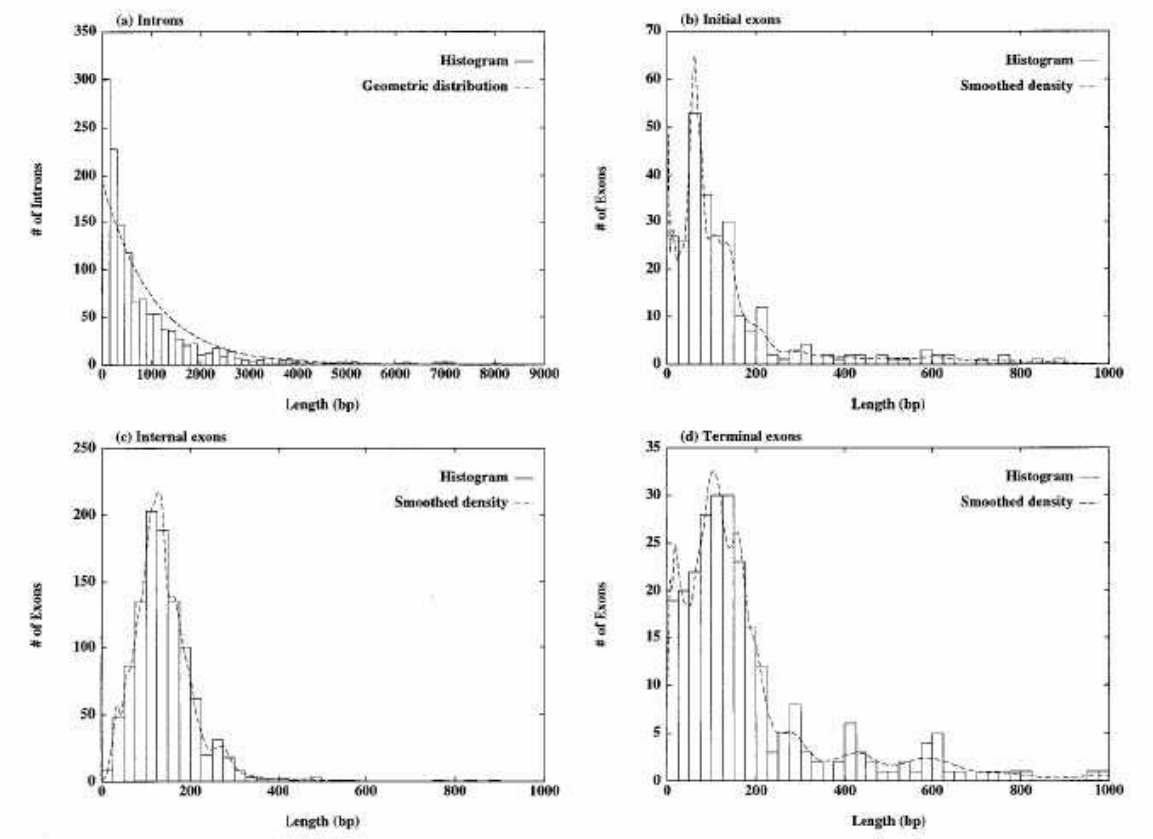

Figure 17.7: Length distribution for (a) 1254 introns; (b) 238 initial exons; (c) 1151 internal exons; (d) 238 terminal exons from the 238 multi-exon genes of the leaning data set. [Burge1997]

Signal Models Many models of biological signal sequences such as splicing sites and transcriptional and translational signals (poly-A, promoters, etc.) have been constructed. In GenScan model, one of the earliest and most influenctial approaches, weight matrix method (WMM), is used. In this method, the frequency $p_j^{(i)}$ of each nucleotide j at each position i of a signal of length is derived from a collection of aligned signal sequences (See Figure 17.8). Then the likelihood of generating a nucleotide sequence $X = x_1, x_2, ..., x_n$ is $P(X) = \prod_{i=1}^{n} p_{x_i}^i$. WMM has been improved by taking account of the dependencies between adjacent/consecutive nucleotides. Such developed models are known as weight array model (WAM).

Recently Korf et al developed a new model for gene prediction called TWINSCAN [Korf2001], which is a

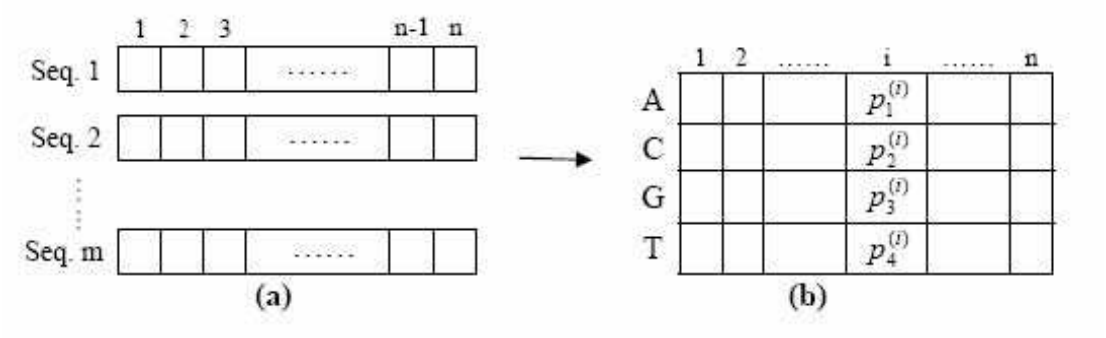

Figure 17.8: Weight metrix method for signal modeling. (a) Multi-sequence alignment for a particular signal; (b) Weight Matrix.

comparative genomic based gene-structure prediction system. It integrates cross-species similarity of high-throughput genome sequences into extended GenScan probability model (GenScan++) to exploit the homology between two related genomes. In similarity searching procedure, the local alignments reported by BLAST are used to construct a conservation sequence, which pairs one of the three cases, match, mismatch and gap, with each nucleotide of the target sequence.

17.3 Gene Expression Analysis

In any living cell, different subsets of its genes are expressed in different stages of certain biological processes. The particular genes expressed at a given stage and their relative abundance are the characteristics of a cell and crucial to its proper function. Measuring gene expression levels in different stages, different body tissues, and different organisms, therefore, provides the information in understanding biological processes.

One of approaches to measuring gene expression profiles is cDNA and oligonucleotide microarray technology, which makes it possible to view the expression of many thousands of genes simultaneously. There are also some other mRNA level methods such as RT-PCR and SAGE generating gene expression data.

Here we will give introduction to some machine learning methods in the context of gene expression data analysis.

Machine learning basically has two categories of methods:

- Supervised learning;
- Unsupervised learning.

Supervised learning solves the problem of classification. The definition of classes is pre-defined, and a labelled example dataset is available for machine training, so that new data can be put into these classes correctly. An example of such problems is determining tumor types (pre-defined) based on sample gene expression patterns. Many algorithms have been developed for supervised learning, such as neural networks, support vector machines and decision trees.

Unsupervised learning is addressed to the problem of discovering classes from unlabelled data or aggregating data into some clusters according to certain patterns discovered from this dataset itself. This type of problem is called clustering. Clustering techniques are applied to gene expression data for identifying subsets of genes that behave similarly under the set of tested conditions (gene clustering) or for identifying subsets of samples that can be characterized by different gene expression patterns (sample clustering). Its the first step in extracting information from mass of gene expression data set, which is crucial for solving the problems such as gene function exploration and gene regulations.

Clustering

$S = \{s_1, s_2, ..., s_N\}$ – a set of objects;
$d(s_i, s_j), 1 \leq i, j \leq N$ – distance between s_i and s_j;
$C = \{c_1, c_2, ..., c_K\}$ – a set of clusters; $c_i \subset S, 1 \leq i \leq K$, and $\bigcup_{i=1}^{K} c_i = S$. There is no consensus of definition of the cluster-set.
The goal of clustering is to partition S into a set of clusters, which satisfy the rule that $d(s_i^{(k)}, s_j^{(k)}) < d(s_i^{(k)}, s_h^{(l)})$, where $s_x^{(y)}$ is arbitary object in cluster y.

If we represented the objects and distances by a graphic model $G = (V, E)$, where $V \rightarrow S$,
and weights of edges $W(E) \rightarrow \{d(s_i, s_j) | s_i, s_j \in S\}$ (No edge corresponds to that d is ∞). Then the clustering problem turns out to be a max cut problem, i.e. finding the cut of maximum weight (Figure 17.9).

In the next class, we will focus on some important clustering algorithms.

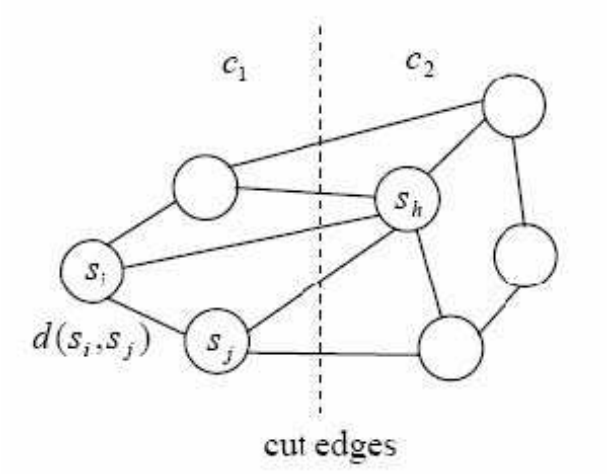

Figure 17.9: A graphic model of clustering

References

[Burge1997] C. Burge and Samuel Karlin. *Prediction of complete gene structures in human genomic DNA*, J. Mol. Biol., 268(1997), 78–94.

[Korf2001] I. Korf, P. Flicek, D. Duan, and M. R. Brent. *Integrating genomic homology into gene structure prediction*, Bioinformatics, 17(2001), Suppl. 1, S140–S148.

[Krogh1998] A. Krogh. *Computational methods in molecular biology: an introduction to hidden markov models for biological sequences*, Elsevier, 1998, 45–63.

[Mathe2002] C. Mathe, M. Sagot, T. Schiex, and P. Rouze. *Current methods of gene prediction, their strengths and weaknesses*, Nucleic Acid Research, 30(2002), 4103–4117.