Biological Sequence Analysis

- How to model a sequence
- Blast BLAST
- Why Bayesian statistics?
- We need Markov's help
- Transforming human into mice

Further reading

- Biological sequence analysis---Probabilistic models of proteins and nucleic acids (R.Durbin, S.Eddy, A.Krogh and G.Mitchison, Cambridge University Press, 1998)
- Bioinformatics---Sequence and Genome Analysis (D.W. Mount, Cold Spring Harbor Lab Press, 2001; 2003 Chinese)
- Introduction to computational Biology---Maps, sequences and genomes (M.S.Waterman, Chapman & Hall, 1995)
- Bioinformatics---The machine learning approach (P.Baldi and S.Brunak, MTT, 2001; 2003 Chinese)
- Current topic in computational molecular biology (Edited by Jiang et al. Tsinghua Uni. Press and MIT, 2002)
- Genetic data analysis Ⅱ---Methods for discrete population genetic data (B.S.Weir, Sinauer Associates, Inc., 1996)
- 计算分子生物学导论（塞图宝等，科学出版社，2003；1th. ed. 1997 by Brooks/Cole)

How to model a sequence

- From genetic model to sequence model
- Some sequence models

Genetic model

- $Y=\mu+G+e$ (Johannsen, 1909)
- $Y=\mu+(A+D+I)+e$ (Fisher, 1918)
- $Y=\mu+(A+D+(AA+AD+DD))+e$ (Cockerham, 1954)
- $Y=\mu+\ldots\ldots$ (Zhu, 1996)

A same way for sequence…

- $L = X_{ij}$

  where *i* is A, T, G or C (nucleic acids)/ A, B, C, E, …(amino acids), and *j* is 1~10^5(for genes) / 0~10^{10}(for genomes)

- $L = X_{ij} + e$ *e: background*
- $L = X_{ij} + S_{ij} + e$ *S: species-specific*
- $L = X_{ij} + S_{ij} + F_{ij} + e$ *F: gene family-specific*

Probabilistic model

- When we talk about a model normally we mean a system that simulates the object under consideration;
- A probabilistic model is one that produces different outcomes with different probabilities;
- A probabilistic model can therefore simulate a whole class of objects, assigning each an associated probability.

R.S.A.G.

A simple sequence probabilistic model

- $L = q_A q_T q_G q_C$

  $(P = q_A q_T q_G q_C = \prod_{i=1}^{n} p_{x_i})$

- eiid and niid

HMM

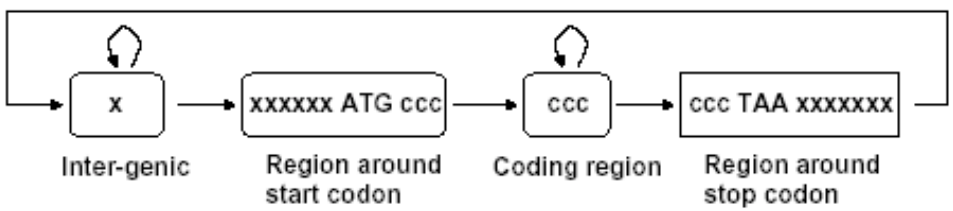
Markov Model (MM)

- Biological sequences can be modeled as the output of a stochastic process in which the probability for a given nucleotide to occur at position p depends on the k previous positions. This representation is called k-order Markov Model.

$$P(x_i|x_1, x_2, \ldots, x_{i-1}) = P(x_i|x_{i-k}, x_{i-(k-1)}, \ldots, x_{i-1})$$

Hidden Markov Model (HMM)

- In a HMM the biological sequences are modeled as the output of a stochastic process that progresses through a series of discrete states. Each state model correspond to a Markov Model.



(Krog, 1998)

HMM

Generalized Hidden Markov Model (GHMM)

- GHMMs are HMMs where states are arbitrary sub-models (e.g., neural networks, position weight matrices, etc.).
- The duration of a particular state depends on some probability distributions.

Principle of Markov Models

- Given a DNA string S, find the most probable path M in the model that generates S. This will be the most probable gene structure.

Markov derived models have many desirable properties

- Modeling: theoretically well-founded models.
- Efficient: $O(|M| \cdot |S|)$ where $|M|$ is the number of states in the model and $|S|$ is the length of the string.
- Scoring: theoretically well-founded scoring system.

Profile (weighted matrix)

- Domains can be defined by different methods:
  - **Pattern (regular expression): used for very conserved domains**
  - **Consensus seuquence**
  - **Profiles (weighted matrices): two-dimensional tables of position specific match-, gap-, and insertion-scores, derived from aligned sequence families; used for less conserved domains**
  - **Hidden Markov Model (HMM): probabilistic models; an other method to generate profiles.**
  - **Motif : A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains.**

Protein domain/family db

Interpro

PROSITE	Patterns / Profiles
ProDom	Aligned motifs (PSI-BLAST) (Pfam B)
PRINTS	Aligned motifs, OWL
Pfam	HMM (Hidden Markov Models)
SMART	HMM
TIGRfam	HMM

DOMO	Aligned motifs
BLOCKS	Aligned motifs (PSI-BLAST)
CDD(CDART)	PSI-BLAST(PSSM) of Pfam and SMART

General information about the entry

Entry name	**EPO_TPO**
Accession number	**PS00817**
Entry type	PATTERN
Date	OCT-1993 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
PROSITE documentation	PDOC00644

Name and characterization of the entry

Description	Erythropoietin / thrombopoeitin signature.
Pattern	P-x(4)-C-D-x-R-[LIVM](2)-x-[KR]-x(14)-C.

Numerical results

- SWISS-PROT release
- Total number of hits in
- Number of hits on prot
- Number of hits on prot
- Number of false hits (c
- Number of known mis
- Number of partial sequ because they are partia
- Precision (true hits / (tr
- Recall (true hits / (true

General information about the entry

Entry name	**INTEIN_C_TER**
Accession number	**PS50818**
Entry type	MATRIX
Date	MAY-2002 (CREATED); MAY-2002 (DATA UPDATE); MAY-2002 (INFO UPDATE).
PROSITE documentation	PDOC00687

Name and characterization of the entry

Description	Intein C-terminal splicing motif profile.

Matrix / Profile

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=22;
/DISJOINT: DEFINITION=PROTECT; N1=3; N2=20;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=0.8533; R2=0.02263959; TEXT='NScore';
/CUT_OFF: LEVEL=0; SCORE=290; N_SCORE=7.4; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=249; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: M0=-8; D=-20; I=-20; B0=-60; B1=-60; E0=-60; E1=-60; MI=-105; MD=-105; IM=-105; DM=-105;

                A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:           B0=0; B1=0; BI=-105; BD=-105;
/M: SY='Y'; M=-15,-11,-29, -9, -6,  3,-23, -5,-14, -1,-13,-10,-10, -2, -9,  0,-12, -8,-15, -5, 12, -9;
/M: SY='V'; M=  1,-25,-13,-27,-26, -1,-24,-27, 20,-18,  7,  6,-24,-26,-25,-17, -8,  1, 34,-27, -9,-26;
/M: SY='Y'; M=-19,-17,-28,-18,-17, 29,-28,  9, -2,-11, -1, -1,-16,-27,-11, -9,-18,-10, -9, 17, 55,-17;
/M: SY='D'; M=-13, 27, -1, 33,  5,-29,-12, -6,-30, -8,-25,-23, 15,-17, -6,-12,  4, -3,-22,-41,-21, -1;
/M: SY='L'; M= -6,-25,-20,-27,-22,  3,-30,-24, 23,-25, 26, 13,-24,-25,-21,-21,-18, -4, 20,-23, -3,-23;
/M: SY='T'; M= -1,  1,-18, -1,  4,-16,-10,-10,-14, -7,-16,-13,  3,-12, -3, -9, 11, 12, -9,-28,-11,  0;
/M: SY='V'; M=  3,-18, -3,-21,-22, -8,-21,-25, 10,-18,  1,  0,-17,-24,-22,-19, -4,  4, 23,-30,-13,-22;
/I:           I=-5; MD=-27;
/M: SY='E'; M=-10,  0,-25,  7, 15,-17,-19, -7,-21,  2,-19,-15, -5, 11,  0, -6, -6, -9,-21,-21, -9,  6; D=-5;
/I:           I=-5; MD=-27;
/M: SY='N'; M= -2, 10,-17,  7,  0,-18,  5, -1,-18, -2,-18,-12, 13,-11, -3, -3,  2, -3,-16,-22,-13, -2; D=-5;
/I:           I=-5; MI=-27; MD=-27; IM=-27; DM=-27;
/M: SY='H'; M=-12,  0,-26,  1,  5,-22, -6, 42,-25, -8,-18, -6,  5,-16,  7, -4, -4,-13,-24,-27,  0,  4; D=-5;
/I:           I=-5; DM=-27;
/M: SY='N'; M= -5, 14,-19,  2, -5,-10, -8, -2,-13, -4,-15,-11, 27,-20, -5, -2,  3,  0,-17,-30,-13, -5;
/M: SY='F'; M=-16,-25,-24,-30,-25, 40,-25, -8,  5,-22,  9,  3,-20,-29,-25,-17,-20,-10,  0,  8, 37,-25;
/I:           I=-6; MD=-32;
/M: SY='V'; M= -6,-27,-18,-30,-25,  7,-30,-20, 24,-24, 18, 12,-23,-26,-23,-20,-16, -4, 25,-21, -1,-25; D=-6;
/I:           I=-6; MI=-32; IM=-32; DM=-32;
/M: SY='A'; M= 15,-14,-14,-20,-15, -7,-14,-20,  1,-15, -2, -3,-12,-17,-13,-17,  4,  8,  8,-23,-11,-14;
/I:           I=-5; MD=-25;
```

More …

- NN(Neural network)
- Stochastic grammar: Chomsky hierachy
- ……

Conclusions

- Statistic methods for genetic models can also be used to sequence models. For example, ML, MCMC;
- One set of methods for biological sequence analysis is rooted in computer science, where there is an extensive literature on text string comparison methods;
- Complexity of model: number of parameters of model

- Sequence: *of* a gene, region, chromosome, or genome
- Three kinds of works: modeling, statistical analysis and algorithmics
- Sequence modeling: a long way to go

Exercise

- Try to construct a new model for biological sequence