Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

Local Multiple Sequence Alignment

Sequence File Formats

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Localized Alignments

- Just like with pairwise alignments, we may not be interested in the global alignment of multiple sequences, but rather only specific regions that are conserved.
- Local Alignment of MSAs are important:
  - Given regions of genomic DNA occurring upstream or before a certain gene, there might be sequences where transcription factors bind to the DNA so that the gene can be transcribed.
  - Thus, if we are interested in determining if there is any signal in the regions upstream of a certain family of genes across several different organisms, it would be important to only find the conserved region, and not try to align all of the genomic DNA.
  - Localized alignments of protein sequences can yield information about conserved domains found in otherwise unrelated proteins.

Approaches to Local Alignment

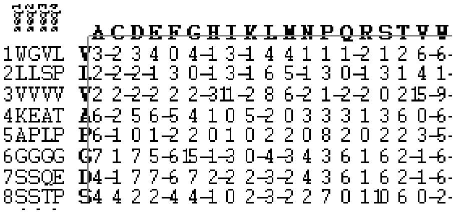
- Profile Analysis
- Block Analysis
- Pattern-searching or statistical methods

Profile Analysis…

- Profiles are found by first multiply aligning the sequences, determining which regions are the most highly conserved,
- and then creating a scoring matrix for the alignment of the highly conserved region.
- Profile is composed of:
  - Columns:
    - one for each residue;
      - columns for insertions and deletions as well
  - Rows:
    - one for each position in the conserved region or motif

…Profile Analysis

- Profiles describe a MSA by a scoring matrix:



		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W
1WGVL	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6
2LLSP	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1
3VVVV	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9
4KEAT	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6
5APLP	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5
6GGGG	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	6	1	6	2	-1	-6
7SSQE	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6
8SSTP	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2

Profile Searches

- Once a profile is created, it can be used to search a target sequence or database for possible matches to the profile using the profiles scores to evaluate the likelihood at each position.
- Profile scores evaluate likelihood of a match at each position

Copyright 2000 N. AYDIN. All rights reserved.

Drawback to Profiles

- Profiles only as representative as the variation in the training sets.
- Thus, there is a bias in the profile towards the training data.
- Training sets can be erroneous if not carefully constructed

Calculating Profiles

- Each cell is the log-odds score
  - The value of an individual cell is calculated as the log odds score of finding a particular residue in a particular location in an alignment divided by the probability of aligning the two amino acids by random chance using a particular scoring scheme (such as PAM250, BLOSUM80, …).
    - PAM (Percent Accepted Mutation)
    - BLOSUM (Blocks Substitution Matrix)
  - Additional penalties must be calculated for gap opening and gap extension in the profile as well.
- Some methods take in sequence weights as well
  - One method (average method) weighs the proportion of the amino acids found in a particular column, and weights the score of matching the consensus residue at a given position to that particular residue.

Shannon Entropy

- One method to calculate the observed column variation given the expected variation in the evolutionary model is to use an information measure known as entropy.
  - Entropy is the amount of information of the observed column variation if expected variation in the evolutionary model is known
- The smaller the entropy, the more conserved a column is.

Entropy…

- The entropy (***H***) for a single column is calculated by the following formula:

$$H = -\sum_{residues(a)} f_a \log(p_a)$$

- ***a***: is a residue (amino acid),
- ***f_a***: frequency of residue ***a*** in a column,
- ***p_a*** : probability (expected frequency) of residue ***a*** in that column

…Entropy…

- *H* is calculated for each 20 ancestor amino acids and for a large number of evolutionary distances (PAM1, PAM2, PAM4, ...).
- The distance that gives the minimum value for *H* for each column-possible ancestor combination is the best estimate of the distance that generates the column diversity from that ancestor.
- This analysis provides 20 possible models (M_a for a = 1,2,3....20) as to how the amino acid frequencies in a column (*F*) may have originated.

…Entropy…

- The next step in the evolutionary profile construction determines the extent to which each M_a predicts *F* by the Bayes conditional probability analysis.

$$P(M_a|F) = P(M_a) \times P(F|M_a) / \sum_{\text{all}\,a\text{'s}} P(M_a) \times P(F|M_a)$$

  - where the prior distribution $P(M_a)$ is given by the background amino acid frequencies and

$$P(F|M_a) = P_{aa1}^{faa1} \times P_{aa2}^{faa2} \times P_{aa3}^{faa3} \ldots\ldots P_{aa20}^{faa20}$$

  - i.e., the product of the expected amino acid frequencies in M_a raised to the power of the fraction observed for each amino acid in the msa column.

Copyright 2000 N. AYDIN. All rights reserved.

…Entropy

- From $P(M_a|F)$, the weights for each of the 20 possible distributions that give rise to the msa column diversity are calculated as follows:

$$W_a = P(M_a|F) - P(M_{random}|F)$$

  - where W_a is the weight given to M_a and $P(M_{random}|F)$ is calculated as above using amino acid distribution.

Log-odds score…

- Another measure of creating a profile is by using log-odds score.
- In this method,
  - the $\log_2$ of the ratio of observed/background frequencies is calculated for each position.
  - What results is the amount of information available in an alignment given in bits.
- A new sequence can then be searched to see if it possibly contains the motif.

- Profiles can also indicate log-odds score:
  - $\text{Log}_2(\text{observed} \div \text{expected})$
- Result is a bit score

…Log-odds score

- The log odds scores for the profile (Profile_{ij}) are given by

$$\text{Profile}_{ij} = \log\left[\sum_{\text{all } a\text{'s}} (W_{ai} \times P_{aij})/P_{\text{random}j}\right]$$

where

- W_{ai} is the weight of an ancestral amino acid a at row i in the profile,
- P_{aij} is the frequency of amino acid j in the PAM amino acid distribution that best matches at row i,
- $P_{\text{random}j}$ is the background frequency of amino acid j.

BLOCK Analysis…

- Blocks are similar to profiles in the sense that
  - they represent locally conserved regions within a MSA.
- However, the difference is that ...
  - blocks lack insert and delete (indels) positions in the sequences.
  - Instead, every column includes only matches and mismatches
- Blocks can be determined either
  - by performing a multiple sequence alignment, or
  - by searching a database for similar sequences of the same length.

…BLOCK Analysis…

- Generally determined by performing multiple alignment first

- Ungapped regions are then separated into blocks

- Algorithms have been developed for searching for blocks

…BLOCK Analysis

- Statistical approaches to finding the most alike sequences have been proposed, such as
  - the Expectation-Maximization algorithms and
  - the Gibbs sampler.

- In any case, once a set of blocks has been determined, the information contained within the block alignment can be displayed as a sequence profile.

Copyright 2000 N. AYDIN. All rights reserved.

BLOCKS Programs

- A global sequence alignment will usually contain ungapped regions that are aligned between multiple sequences.
- These regions can be extracted to produce blocks.
- Two widely used programs:
  - BLOCKS
  - eMOTIF

http://www.blocks.fhcrc.org/blocks/process_blocks.html

http://dna.stanford.edu/emotif/

Example…

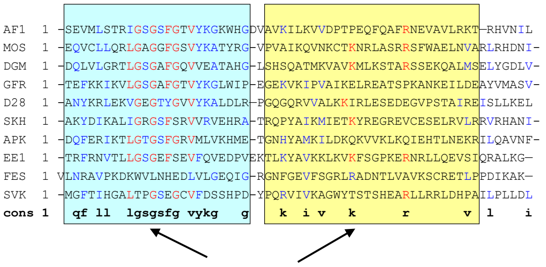
- 10 Truncated Kinase proteins
  - Approximately 75 residues in length
    - A protein kinase is a kinase enzyme that modifies other proteins by chemically adding phosphate groups to them (phosphorylation).
    - The human genome contains about 500 protein kinase genes and they constitute about 2% of all human genes.
    - Protein kinases are also found in bacteria and plants.
    - Up to 30% of all human proteins may be modified by kinase activity, and kinases are known to regulate the majority of cellular pathways, especially those involved in signal transduction.

…Example…

```
>D28     CD28  S. CEREVISIAE CELL CYCLE CONTROL PROTEIN KINASE
ANYKRLEKVGEGTYGVVYKALDLRPGQGQRVVALKKIRLESEDEGVPSTAIREISLLKEL
>SKH   SKH  HELA MYSTERY PUTATIVE PROTEIN KINASE
AKYDIKALIGRGSFSRVVRVEHRATRQPYAIKMIETKYREGREVCESELRVLRRVRHANI
>APK   CAPK  BOVINE CARDIAC MUSCLE CYCLIC AMP-DEPENDENT (ALPHA)
DQFERIKTLGTGSFGRVMLVKHMETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNF
>EE1   WEE1  S. POMBE MITOTIC INHIBITOR
TRFRNVTLLGSGEFSEVFQVEDPVEKTLKYAVKKLKVKFSGPKERNRLLQEVSIQRALKG
>GFR   EGFR  HUMAN EPIDERMAL GROWTH FACTOR RECEPTOR
TEFKKIKVLGSGAFGTVYKGLWIPEGEKVKIPVAIKELREATSPKANKEILDEAYVMASV
>DGM  PDGF RECEPTOR, MOUSE KINASE REGION
DQLVLGRTLGSGAFGQVVEATAHGLSHSQATMKVAVKMLKSTARSSEKQALMSELYGDLV
>FES  THIS IS VFES TYROSINE KINASE
VLNRAVPKDKWVLNHEDLVLGEQIGRGNFGEVFSGRLRADNTLVAVKSCRETLPPDIKAK
>AF1   RAF1  HUMAN C-RAF-1 ONCOGENE
SEVMLSTRIGSGSFGTVYKGKWHGDVAVKI LKVVDPTPEQFQAFRNEVAVLRKTRHVNIL
>MOS   CMOS  HUMAN C-MOS ONCOGENE
EQVCLLQRLGAGGFGSVYKATYRGVPVAIKQVNKCTKNRLASRRSFWAELNVARLRHDNI
>SVK   HSVK  HERPES SIMPLEX VIRUS PUTATIVE PROTEIN KINASE
MGFTIHGALTPGSEGCVFDSSHPDYPQRVIVKAGWYTSTSHEARLLRRLDHPAILPLLDL
```

…Example…

Multiple Alignment created using ClustalW; Colors Added using BoxShade



```
AF1   1 -SEVMLSTRIGSGSFGTVYKGKWHGDVAVKILKVVDPTPEQFQAFRNEVAVLRKT-RHVNIL
MOS   1 -EQVCLLQRLGAGGFGSVYKATYRG-VPVAIKQVNKCTKNRLASRRSFWAELNVARLRHDNI-
DGM   1 -DQLVLGRTLGSGAFGQVVEATAHG-LSHSQATMKVAVKMLKSTARSSEKQALMSELYGDLV-
GFR   1 -TEFKKIKVLGSGAFGTVYKGLWIP-EGEKVKIPVAIKELREATSPKANKEILDEAYVMASV-
D28   1 -ANYKRLEKVGEGTYGVVYKALDLR-PGQGQRVVALKKIRLESEDEGVPSTAIREISLLKEL
SKH   1 -AKYDIKALIGRGSFSRVVRVEHRA-TRQPYAIKMIETKYREGREVCESELRVLRRVRHANI-
APK   1 -DQFERIKTLGTGSFGRVMLVKHME-TGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNF-
EE1   1 -TRFRNVTLLGSGEFSEVFQVEDPVEKTLKYAVKKLKVKFSGPKERNRLLQEVSIQRALKG-
FES   1 VLNRAVPKDKWVLNHEDLVLGEQIG-RGNFGEVFSGRLRADNTLVAVKSCRETLPPDIKAK-
SVK   1 -MGFTIHGALTPGSEGCVFDSSHPD-YPQRVIVKAGWYTSTSHEARLLRRLDHPAILPLLDL
cons 1   qf ll  lgsgsfg vykg   g    k  i v   k      r       v  l    i
```

BLOCKS Server located blocks

…Example…

```
• Taking this alignment, blocks can be generated using
  the BLOCKS server:

ID   x6676xbli; BLOCK
AC   x6676xbliA; distance from previous blocks=(1,1)
DE   ../tmp/6676.blin
BL   UNK motif;  width=24; seqs=10; 99.5%=0; strength=0
AF1              (    1) SEVMLSTRIGSGSFGTVYKGKWHG  41
MOS              (    1) EQVCLLQRLGAGGFGSVYKATYRG  48
DGM              (    1) DQLVLGRTLGSGAFGQVVEATAHG  49
GFR              (    1) TEFKKIKVLGSGAFGTVYKGLWIP  41
D28              (    1) ANYKRLEKVGEGTYGVVYKALDLR  61
SKH              (    1) AKYDIKALIGRGSFSRVVRVEHRA  54
APK              (    1) DQFERIKTLGTGSFGRVMLVKHME  46
EE1              (    1) TRFRNVTLLGSGEFSEVFQVEDPV  55
FES              (    1) LNRAVPKDKWVLNHEDLVLGEQIG 100
SVK              (    1) MGFTIHGALTPGSEGCVFDSSHPD  73
//
```

…Example

```
ID   x6676xbli; BLOCK
AC   x6676xbliB; distance from previous blocks=(2,2)
DE   ../tmp/6676.blin
BL   UNK motif;  width=28; seqs=10; 99.5%=0; strength=0
AF1             (  27) AVKILKVVDPTPEQFQAFRNEVAVLRKT  87
MOS             (  27) PVAIKQVNKCTKNRLASRRSFWAELNVA  75
DGM             (  27) SHSQATMKVAVKMLKSTARSSEKQALMS  92
GFR             (  27) GEKVKIPVAIKELREATSPKANKEILDE  83
D28             (  27) PGQGQRVVALKKIRLESEDEGVPSTAIR  83
SKH             (  27) RQPYAIKMIETKYREGREVCESELRVLR  74
APK             (  27) GNHYAMKILDKQKVVKLKQIEHTLNEKR  85
EE1             (  27) TLKYAVKKLKVKFSGPKERNRLLQEVSI  77
FES             (  27) GNFGEVFSGRLRADNTLVAVKSCRETLP 100
SVK             (  27) PQRVIVKAGWYTSTSHEARLLRRLDHPA  92
//
```

Copyright 2000 N. AYDIN. All rights reserved.

Statistical Methods for Aiding Alignments

- Commonly used methods for locating motifs:
  - Expectation-Maximization (EM)
  - Gibbs Sampling

Expectation-Maximization…

- EM algorithm has been used to identify both conserved domains in unaligned proteins and protein-binding sites in unaligned DNA sequences, including sites that may include gaps
- In the EM algorithms,
  - the starting point is a set of sequences expected to have a common sequence pattern that may not be easily detectible.
  - An initial guess is made as to the location and size of the site of interest in each of the sequences.
  - These initial sites are then aligned.
  - Approximate length of signal must be given
- Randomly assign locations of this motif in each sequence

…Expectation-Maximization…

- The EM algorithm consists of two steps, which are repeated consecutively:
  - Expectation Step
    - In the expectation step, background residue frequencies are calculated based on those residues that are not in the initially aligned sites.
    - Column specific residues are calculated for each position in the initial motif alignment.
    - Using this information, the probability of finding the site at any position in the sequences can then be calculated.
    - Residues not in a motif are background
  - Frequencies used to determine probability of finding site at any position in a sequence to fit motif model

…Expectation-Maximization

- Maximization Step
  - In the maximization step, the counts of residues for each position in the site as found in the expectation step are used to calculate the location within each sequence that maximally aligns to the motif pattern calculated in the expectation step.
  - This is done for each of the sequences.
- Once a new motif location has been calculated, the expectation step is repeated.
- This cycle continues until the solution converges.

Example of EM - initial alignment…

TCAGAACCAGTTATAA**ATTTAT**CATTTCCTTCTCCACTCCT
CCCACGCA**GCCGCC**CTCCTCCCCGGTCACTGACTGGTCCTG
TCGACCCTCTGAACCTATCAGGGACCA**CAGTCA**GCCAGGCAAG
AAAACACTTGAG**GGAGCA**GATAACTGGGCCAACCATGACTC
GGGTGAATGGTACTGCT**GATTAC**AACCTCTGGTGCTGC
AGCCTAGAGT**GATGAC**TCCTATCTGGGTCCCCAGCAGGA
GCCTCAGGATCCAGCACACAT**TATCAC**AAACTTAGTGTCCA
CATTATCAC**AAACTT**AGTGTCCATCCATCACTGCTGACCCT
TCGGAACAAGGCAAA**GGCTAT**AAAAAAAATTAAGCAGC
GCCCCTTCCCCA**CACTAT**CTCAATGCAAATATCTGTCTGAAACGGTTCC
CATGCCCTCAAGTGTGCAGATTGGT**CACAGC**ATTTCAAGG
GATTGGTCACAGCAT**TTCAAG**GGAGAGACCTCATTGTAAG
TCCCCAACTCCCAACTGACCTTAT**CTGTGG**GGGAGGCTTTTGA
CCTTATCTGT**GGGGGA**GGCTTTTGAAAAGTAATTAGGTTTAGC
ATTATTTTCCTTATCAGAAGC**AGAGAG**ACAAGCCATTTCTCTTTCCTCCCGGT
AGG**CTATAA**AAAAAATTAAGCAGCAGTATCCTCTTGGGGGCCCCTTC
CCAGCACACACACTTATC**CAGTGG**TAAATACACATCAT
TCAAATAGGTACGGATAAG**TAGATA**TTGAAGTAAGGAT
ACTTGGGGTTCCAGTTTGATAAGAAAAGACTT**CCTGTG**GA
TGGCCGC**AGGAAG**GTGGGCCTGGAAGATAACAGCTAGTAGGCTAAGGCCAG
CAACCACAACCTCTGTATCCGGTAGTGGCAGATGGAAA
CTGTATCCGGTAG**TGGCAG**ATGGAAAGAGAAACGGTTAGAA
GAAAAAAAATAAATGAAGTCTGCC**TATCTC**CGGGCCAGAGCCCCT
TGCCTTGTCTGTTGTAGATAATGAATCTATCCTCCA**GTGACT**
GGCCAGGCTGAT**GGGCCT**TATCTCTTTACCCACCTGGCTGT
CAACAGCAGGTCCTACTATCGCCTCCCTCT**AGTCTC**TG
CCAACCG**TTAATG**CTAGAGTTATCACTTTCTGTTATCAAGTGGCTTCAGCTATGCA
GGGAGGGTGGGGCCCCTATCTCTCCTA**GACTCT**GTG
CTTTGTC**ACTGGA**TCTGATAAGAAACACCACCCCTGC

begin with an initial, random alignment:

…Example of EM - Residue Counts…

- From this alignment, the frequency of each base occurring is calculated.
- In this case, the motif we are searching for is six bases wide.
  - Therefore, we need to calculate seven different sets of frequencies:
    - One for the background,
    - one for each of the columns in the motif.
- Calculating the total counts, we get:

Nucleotide	Motif Position (0 = Backgorund)							
	0	1	2	3	4	5	6	Total
A	279	6	12	6	6	11	7	48
C	280	8	3	5	7	7	7	37
G	225	9	8	10	7	5	8	47
T	262	6	6	8	9	6	7	42
Total	1046	29	29	29	29	29	29	174

Copyright 2000 N. AYDIN. All rights reserved.

…Example of EM - Residue Frequencies…

- After calculating the observed counts for each of the positions, we can convert these to observed frequencies:

Nucleotide	Motif Position (0 = Backgorund)						
	0	1	2	3	4	5	6
A	0.267	0.209	0.414	0.209	0.209	0.379	0.241
C	0.267	0.276	0.103	0.172	0.241	0.241	0.241
G	0.216	0.310	0.276	0.345	0.241	0.172	0.276
T	0.250	0.209	0.209	0.276	0.310	0.209	0.241

  - Frequency of nucletide *a* for the background (Col0):
    - # of nucletide *a* in Col0 in Row*a* / # of all nucletides in Col0
  - Frequency of *a* in Col*c*:
    - # of nucletide *a* in Col*c* / # of all nucletides in Col*c*

…Example of EM - Residue Frequencies…

- However, in order to alleviate the issue of zero counts and overtraining of the data, pseudocounts are introduced to the observed counts:
  - In this case, frequency of nucleotide *a* in Col*c*:

$$P_{ca} = (n_{ca} + b_{ca}) / (N_c + B_c)$$

P_{ca}: Probability of residue *a* in column c ; n_{ca}: count of *a*'s in column *c* ; b_{ca}: pseudocount of *a*'s in column *c* ;
N_c: total count in column *c* ; B_c: total pseudocount in column *c*

  - Chosing a pseudocaunt is arbitrary
  - For example, assuming that 4 nucleotides have equal probabilities, if total pseudocount (B_c) is chosen as 1, pseudocount of each nucletide will be $b_{ca}= B_c/4$.
  - Note that a different pseudocount scheme is used in the following table

Nucleotide	Motif Position (0 = Backgorund)						
	0	1	2	3	4	5	6
A	0.267	0.256	0.296	0.256	0.256	0.289	0.263
C	0.267	0.263	0.230	0.243	0.256	0.256	0.256
G	0.216	0.240	0.233	0.246	0.226	0.213	0.233
T	0.250	0.241	0.241	0.254	0.261	0.241	0.248

…Example of EM - Maximization Step…

- In the expectation step, the residue frequencies for the motif are used to estimate the composition of the motif site.
- The expectation step attempts to maximally discriminate between sequence within and not within the site.
- For each sequence, each possible motif location is considered in order to find the most probable location given the current motif.
- Consider the first sequence:
  - `TCAGAACCAGTTATAA`**ATTTAT**`CATTTCCTTCTCCACTCCT`
  - There are 41 residues; 41- 6 + 1 = 36 sites to consider

- Starting from the first site (`TCAGAA`), 36 scores for the first sequence are calculated

…Example of EM - Residue Frequencies…

- Let us consider the eigth site CAGTTA.
  - In order to calculate site score, observed frequency table is used:

Nucleotide	Motif Position (0 = Backgorund)						
	0	1	2	3	4	5	6
A	0.267	0.256	**0.296**	0.256	0.256	0.289	**0.263**
C	0.267	**0.263**	0.230	0.243	0.256	0.256	0.256
G	0.216	0.240	0.233	**0.246**	0.226	0.213	0.233
T	0.250	0.241	0.241	0.254	**0.261**	**0.241**	0.248

Position:	1	2	3	4	5	6
	C	A	G	T	T	A

S_{CAGTTA} = 0.263×0.296×0.246×0.261×0.241×0.263

S_{CAGTTA} = 0.000317

…Example of EM - Maximization Step…

	1	2	3	4	5	6	1*2*3*4*5*6	RANDOM	ODDS
TCAGAA	.241	.230	.256	.226	.289	.263	0.000244	0.000274	0.89
CAGAAC	.263	.296	.246	.256	.289	.256	0.000363	0.000362	1.00
AGAACC	.256	.233	.256	.256	.256	.256	0.000256	0.000362	0.71
GAACCA	.240	.296	.256	.256	.256	.263	0.000313	0.000362	0.87
AACCAG	.256	.296	.243	.256	.289	.233	0.000317	0.000362	0.88
ACCAGT	.256	.230	.243	.256	.213	.248	0.000193	0.000274	0.71
CCAGTT	.263	.230	.256	.226	.241	.248	0.000209	0.000257	0.81
CAGTTA	**.263**	**.296**	**.246**	**.261**	**.241**	**.263**	**0.000317**	**0.000257**	**1.23**
AGTTAT	.256	.233	.254	.261	.289	.248	0.000283	0.000241	1.18
GTTATA	.240	.241	.254	.256	.241	.263	0.000238	0.000241	0.99
TTATAA	.241	.241	.256	.261	.289	.263	0.000295	0.000297	0.99
TATAAA	.241	.296	.254	.256	.289	.263	0.000353	0.000297	1.19
ATAAAT	.256	.241	.256	.256	.289	.248	0.000290	0.000318	0.91
TAAATT	.241	.296	.256	.256	.241	.248	0.000279	0.000297	0.94
AAATTT	.256	.296	.256	.261	.241	.248	0.000303	0.000297	1.02
AATTTA	.256	.296	.254	.261	.241	.263	0.000318	0.000297	1.07
ATTTAT	.256	.241	.254	.261	.289	.248	0.000293	0.000278	1.05
TTTATC	.241	.241	.254	.256	.241	.256	0.000233	0.000278	0.84

…Example of EM - Maximization Step…

TTATCA	.241	.241	.256	.261	.256	.263	0.000261	0.000297	0.88
TATCAT	.241	.296	.254	.256	.289	.248	0.000332	0.000297	1.12
ATCATT	.256	.241	.243	.256	.241	.248	0.000229	0.000297	0.77
TCATTT	.241	.230	.256	.261	.241	.248	0.000221	0.000278	0.80
CATTTC	.263	.296	.254	.261	.241	.256	0.000318	0.000297	1.07
ATTTCC	.256	.241	.254	.261	.256	.256	0.000268	0.000297	0.90
TTTCCT	.241	.241	.254	.256	.256	.248	0.000240	0.000278	0.86
TTCCTT	.241	.241	.243	.256	.241	.248	0.000216	0.000278	0.78
TCCTTC	.241	.230	.243	.261	.241	.256	0.000217	0.000297	0.73
CCTTCT	.263	.230	.254	.261	.256	.248	0.000255	0.000297	0.86
CTTCTC	.263	.241	.254	.256	.241	.256	0.000254	0.000297	0.86
TTCTCC	.241	.241	.243	.261	.256	.256	0.000241	0.000297	0.81
TCTCCA	.241	.230	.254	.256	.256	.263	0.000243	0.000318	0.76
CTCCAC	.263	.241	.243	.256	.289	.256	0.000292	0.000339	0.86
TCCACT	.241	.230	.243	.256	.256	.248	0.000219	0.000318	0.69
CCACTC	.263	.230	.256	.256	.241	.256	0.000245	0.000339	0.72
CACTCC	.263	.296	.243	.261	.256	.256	0.000324	0.000339	0.95
ACTCCT	.256	.230	.254	.256	.256	.248	0.000243	0.000318	0.76

Copyright 2000 N. AYDIN. All rights reserved.

…Example of EM - Maximization Step…

- The six base site CAGTTA beginning at base 8 is calculated to have the highest odds probability.
- Therefore, it is chosen as the new site in sequence 1.
- This is repeated for each of the sequences.
- In the maximization step, the newly chosen sites for each of the sequences are used to recalculate the frequency table.
- The expectation/maximization cycle is then repeated, until the results converge on a set of motifs.

…Example of EM - Maximization Step

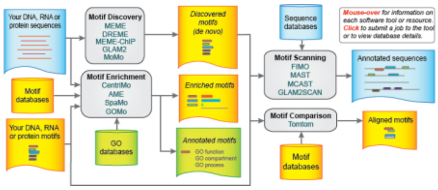
- Before:
  - Random Alignment
- TCAGAACCAGTTATAAA**ATTTAT**CATTTCCTTCTCCACTCCT
- After:
  - Maximal location (given random motif alignment) (first round)
- TCAGAAC**CAGTTA**TAAATTTATCATTTCCTTCTCCACTCCT

Available E-M Programs

- MEME – Uses E-M algorithms as explained
  - **Multiple EM for Motif Elicitation (MEME)** is a program developed that uses the expectation-maximization methods as described previously.
    - ParaMEME searches for blocks using the EM algorithm,
    - MetaMEME searches for profiles using Hidden Markov Models (HMMs).
- MEME locates one or more ungapped patterns in a single DNA or protein sequence, or in a series of sequences.
- A search is conducted on a variety of motif widths in order to determine the most likely width for the profile.
  - This likelihood is based on the log likelihood score calculated after the EM algorithm.

The MEME Suite

- Motif-based sequence analysis tools



- http://meme-suite.org/index.html

MEME Software

- One of three types of motif models can be chosen:
  - OOPS (One expected Occurrence Per Sequence)
    - simplest model type since it assumes that there is exactly one occurrence per sequence of the motif in the dataset.
  - ZOOPS (Zero or One expected Occurrence Per Sequence)
    - generalization of OOPS
    - assumes zero or one motif occurrences per dataset sequence
  - TCM (Two-Component Mixture)
    - assumes that there are zero or more non-overlapping occurrences of the motif in each sequence in the dataset

  - Bailey, Timothy L. and Charles Elkan. "The Value of Prior Knowledge in Discovering Motifs with MEME." Proceedings. International Conference on Intelligent Systems for Molecular Biology 3 (1995): 21-9 .
  - https://tlbailey.bitbucket.io/papers/cs95_143.pdf

MEME Software

- Various prior knowledge can be added to MEME, including
  - the expected number of motifs,
  - the expected length of the motif,
  - whether or not the motif is palindromic
    - only applicable for DNA sequences

Copyright 2000 N. AYDIN. All rights reserved.

Gibbs Sampling…

- Similar in nature to the EM algorithms.
  - – Combines both EM and simulated annealing techniques in order to determine a maximal local alignment of multiple sequences.
  - – Goal is to find most probable pattern by sampling from motif probabilities to maximize model÷background probabilities
    - The idea behind Gibbs sampling is to determine the most probable pattern common to all of the sequences by sliding them back and forth until the ratio of the motif probability to the background probability is a maximum.

…Gibbs Sampling…

- Predictive Update Step
  - – Random motif start position chosen for all sequences except one
  - – Initial alignment used to calculate residue frequencies for motif and background
  - – Similar to the Expectation Step of EM
- Sampling Step
  - – Model probability÷background probability normalized and weighted
  - – Motif start position chosen based on a random sampling with the given weights
  - – Different than EM algorithm

…Gibbs Sampling

- Process repeated until residue frequencies in each column do not change
- The sampling step is then repeated for a different initial random alignment
  - – Sampling allows escape from local maxima
  - – Employs a shifting routine that will take a current multiple motif alignment, and shift it a few bases to the left or the right, in order to see if only part of the motif is being found
  - – A range of motif sizes can be explored in Gibbs sampling as well
- Gibbs sampling can be extended
  - – to search for multiple motifs in the same set of sequences,
  - – to find a pattern in only a fraction of the sequences.
- In addition, certain model-specific parameters can be enforced, such as palindromic sequences

Hidden Markov Models…

- Hidden Markov Models (HMMs)
  - – probabilistic models for studying sequences of symbols.
- HMMs can model matches, mismatches, insertions and deletions of symbols.
- HMMs have been deeply rooted in speech recognition problems.
- In speech recognition, the problem is the phonemes (or words) that have been spoken in a particular time frame.

…Hidden Markov Models…

- Consider the difficulty.
  - – Everyone you meet has a different voice.
  - – Everyone speaks with a slight variation
    - this might be caused by an accent, the person having a cold, or differences in physiological development.
- However, humans are able to distinguish what the speaker is saying.
  - – The idea behind speech recognition is to take in a spoken word and to try to fit it to a specific model of possible words.
    - This may in fact be close to what the brain does

…Hidden Markov Models

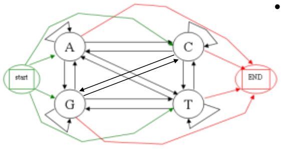
- Problems in sequence analysis are similar.
- For instance,
- given an amino acid sequence, we may want to determine the protein family to which it belongs.
- The amino acid sequence can be treated similarly to the speech signal in a given frame, and the amino acids can be treated as the phonemes.

Copyright 2000 N. AYDIN. All rights reserved.

Markov Chain

- A probabilistic model that generates a sequence where the probability of a symbol depends upon the previous symbol.
  - A traffic light is an example of a Markov chain.
- A Markov Chain can be used to model a random DNA sequence, where there are four states:
  - A, C, G, T
    - one for each letter in the alphabet.
- When we are given a certain state, there is a transition from that state to another state with an associated probability
  - called a transition probability.
- An example Markov Chain can be drawn as follows:

Markov Chain



- The key property of a Markov chain is that
  - the probability of a symbol S at position $p(S_p)$ depends only upon the previous symbol S at position $p_{-1}(S_{p-1})$, and not on the entire previous sequence.
- Since the probability of a symbol is dependent upon the previous symbol, a prime example for the use of Markov chains is in the detection of CpG islands, which are rich in the dinucleotide CG.
  - **CpG** (CG) **islands** is a short stretch of DNA in which the frequency of the CG sequence is higher than other regions.
    - "p" simply indicates that "C" and "G" are connected by a phosphodiester bond.

Markov Chain

- The process of methylation in biological systems will typically convert the nucleotide C to a T with a high probability when a CG nucleotide is encountered.
  - As a result, there will be an overabundance of the dinucleotide TG, and an underabundance of the dinucleotide CG.
- If we ignore the start and end states for now, we can see that there are sixteen different transitions.
  - A study of regions of genomic DNA has determined normal genomic transition probabilities to be the following,
    - where the FROM node is labeled along the rows to the left, and the TO node is labeled along the columns above:

Markov Chain

	A	**C**	**G**	**T**
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

- The model shown above can then assign these weights to the edges of the graph

Markov Chain

- In some regions of the genome, such as the promoter region of genes, methylation is suppressed.
  - In these regions, the dinucleotide CG is found in greater quantities.
- In fact, the nucleotides C and G are found to a greater degree than elsewhere in the genome.
  - A study of regions of genomic DNA where CpG islands exist has determined the transition probabilities to be the following:

Markov Chain

	A	**C**	**G**	**T**
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

- A new model just like the one above can have its transition properties assigned according to the new table.
- Now we have two different models:
  - the first where CpG islands are absent,
  - the second where CpG islands are present.

Copyright 2000 N. AYDIN. All rights reserved.

Markov Chain

- Let’s call the first model the non-CpG model and the second model the CpG model.
- Given a new sequence, how would we determine whether it belongs to the non-CpG model or the CpG model?
- Remember, the key property of a Markov chain
  - the probability of a symbol S at position $p(S_p)$ depends only upon the previous symbol S at position $p_{-1}(S_{p-1})$,
    - not on the entire previous sequence.

Markov Chain

- Therefore, to find the probability that a sequence fits a model,
  - you would multiply all of the conditional probabilities:

    $P(x) = P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})\ldots P(x_2|x_1)P(x_1)$
- which can be rewritten as:

$$P(x) = P(x_1)\prod_{i=2}^{L} a_{x_{i-1}x_i}$$

- where $a_{x_{i-1}x_i}$ is the probability from residue at posistion i-1 to the residue at position i

Markov Chain

- Let’s consider for now that in the non-CpG model, P(A) = P(T) = 0.3; P(C) = P(G) = 0.2,
  - so that A and T are more probable.
- In the CpG model, consider P(A) = P(C) = P(G) = P(T) = 0.25.
- Now consider the sequence: GGCGACG
- The probability for this sequence:

  P(G)P(G|G)P(C|G)P(G|C)P(A|G)P(C|A)P(G|C)

Markov Chain

- For the non-CpG model can be calculated as:

  (0.20)(0.298)(0.246)(0.078)(0.248)(0.205)(0.078) = 0.000000453499
- For the CpG model can be calculated as:

  (0.25)(0.375)(0.339)(0.274)(0.161)(0.274)(0.274)(0.125) = 0.0010526
- Given this information, it is more likely that this sequence fits the CpG model.
- One thing to note is how quickly the probability gets to zero.
  - This shows the importance of using log statistics.

Using Markov models for discrimination

- How different the non-CpG and CpG models are in relation to each other?
  - If they are not different enough, then there is not enough information to determine from which model a particular sequence is derived.
- In order to test whether we are able to discriminate between the two models, a log ratio is taken for each of the scores in the two previous tables to create a third table, where each entry, x, in the new table is equal to:

  $\log_2(P(x|\text{CpG model}) / P(x|\text{ non-CpG model}))$

Using Markov models for discrimination

- The resulting table is as follows:

	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

- Using this log-odds ratio table as the scores, we can then see that
  - a sequence with a negative score will belong to the non-CpG model,
  - a sequence with a positive score will belong to the CpG model.

Copyright 2000 N. AYDIN. All rights reserved.

Position Specific Scoring Matrix (PSSM)

- Position Specific Scoring Matrices incorporate information theory in order to gain a measure of how much information is contained within each column of a multiple alignment.
  - The information contained within a PSSM is a logarithmic transformation of the frequency of each residue in the motif.
- One problem with creating a model of a sequence alignment that is then used to search databases is that there is a bias towards the training data
  - Some residues may be underrepresented
  - Other columns may be too conserved

Pseudocounts…

- Solution:
  - Introduce Pseudocounts to get a better indication
- The goal of adding pseudocounts is to obtain an improved estimate of the probability p_{ca} that amino acid a is in column c in all occurrences of the blocks, and not just the ones in the present sample.
- The current estimate of p_{ca} is f_{ca}, the frequency of counts in the data.
- A simplified Bayesian prediction improves the estimate of p_{ca} by adding prior information in the form of pseudocounts

…Pseudocounts

- Now the estimated probability is changed from a frequency of counts in the data to the following form:

$$P_{ca} = \frac{n_{ca} + b_{ca}}{N_c + B_c}$$

  - P_{ca}: Probability of residue a in column c
  - n_{ca}: count of a's in column c
  - b_{ca}: pseudocount of a's in column c
  - N_c: total count in column c
  - B_c: total pseudocount in column c
- These probabilities are then converted into a log-odds form (usually $\log_2$ so the information can be reported in bits) and placed in the PSSM .

Searching PSSMs

- In order to search a sequence against a PSSM, the value for the first residue in the sequence occurring in the first column is calculated by searching the PSSM.
- Similarly, the value for the residue occurring in each column is calculated.
- These values are added (since they are logarithms) to produce a summed log odds score, S.
- This score can be converted to an odds score using the formula 2^S.
- The odds scores for the motif beginning at each position can be summed together and normalized to produce a probability of the motif occurring at each location.

Information in PSSMs

- Information theory can give an appreciation for the amount of information contained within each sequence.
- When there is no information contained within a column, the amount of uncertainty can be measured as
  - $\log_2 20 = 4.32$ for amino acids (20 amino acids)
  - $\log_2 4 = 2$ for nucleic acid sequences (4 nucletides)
- If only one amino acid is found in a particular column, then the uncertainty is 0 (there is only one choice).
- If there are two amino acids occurring with equal probability, then there is an uncertainty to deciding which residue it is.

Measure of Uncertainty

- The amount of uncertainty for a particular column is measured as the entropy, as introduced previously

$$H_C = -\sum_{residues\ (a)} f_{ac} \log(p_{ac})$$

- The uncertainty for the whole PSSM can be calculated as a sum over all columns:

$$H_c = \sum_{allcolumns} H_c$$

Copyright 2000 N. AYDIN. All rights reserved.

Relative Entropy

- In addition to the entropy measure given before, a relative entropy measure could be calculated as well.
  - Relative entropy takes into account not only the data in the columns of the motif, but also the overall composition of the organism being studied.
- Relative entropy can be measured as:

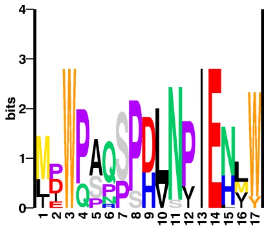
$$R_C = -\sum_{residues(a)} f_{ac} \log_2 (p_{ac} / b_a)$$

- $\boldsymbol{b_a}$ is background frequency of residue $\boldsymbol{a}$ in the organism

Sequence Logos…

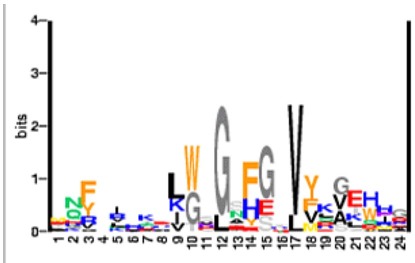
- One way to look at a particular PSSM is to view it visually.
  - Sequence logos are one way to do so, by illustrating the information in each column of a motif.
- Such a graph can indicate which residues and which columns are the most important as far as sequence conservation is concerned.
  - The height of the logo is calculated as the amount by which uncertainty has been decreased
  - If the frequency in the column is less than the frequency in the background, then a negative relative entropy can be computed, which can be shown by an inverted character in the logo.

…Sequence Logos…



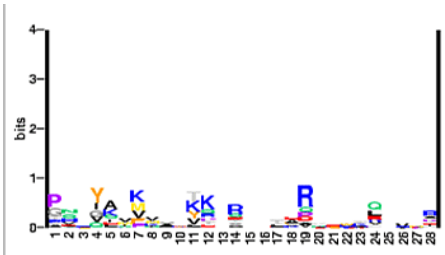
Logo of Gibbs Block D (Tc1) 9 sequences

…Sequence Logos…



PSSM of x6676xbIIA (x6676xbII;) 10 sequences.

…Sequence Logos



PSSM of x6676xbIIB (x6676xbII;) 10 sequences.

Sequence Editors

- Allow manual editing of alignments
- Add color to alignments
- Prepare images for publication
- Some sequence editors:

– BoxShade	http://www.ch.embnet.org/software/BOX_form.html
– Serial Cloner	http://serialbasics.free.fr/Serial_Cloner.html
– GenBeans	http://www.genbeans.org/
– GeneStudio	http://genestudio.com/
– Seqtools	http://www.seqtools.dk/
– GENtle	http://gentle.magnusmanske.de/
– pDRAW32	http://www.acaclone.com/
– DAMBE	http://dambe.bio.uottawa.ca/DAMBE/dambe.aspx

Copyright 2000 N. AYDIN. All rights reserved.

Sequence File Formats

- We have been using DNA and amino acid sequences already
- What is the typical format for these?
  - ANSWER: Many different options
- In order to standardize sequence data, The Nomenclature Committee of the International Union of Biochemistry and the International Union of Pure and Applied Chemistry (IUPAC) has established a standard code to represent bases that are uncertain or ambiguous.

Standard Codes (IUPAC)

- IUPAC nucleotide codes and corresponding bases:

  A = adenine
  C = cytosine
  G = guanine
  T = thymine
  U = uracil
  R = G or A (purine)
  Y = T or C (pyrimidine)
  K = G or T (keto)
  M = A or C (amino)

  S = G or C
  W = A or T
  B = G or T or C
  D = G or A or T
  H = A or C or T
  V = G or C or A
  N = A or G or C or T (any)
  . or - = gap
- Any other character represents an error that will not be tolerated by nearly all sequence analysis programs.

Standard IUPAC Codes

- IUPAC standard single letter and three letter amino acid codes:

A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamine or Glutamic acid
X	Xaa or Xxx	Any amino acid

Fasta File Format

- Fasta sequence format is one of the most basic and widespread sequence formats.
- A sequence in fasta format has as its first line a descriptor beginning with a '>' character.
- The proceeding lines contain the sequence (either nucleotide or amino acid) using standard one-letter symbols.
- This format is extremely useful for sequence analysis programs, since it is devoid of numerical and non-sequence characters (with the exception of the newline character).

Fasta File Format

- Example Fasta Sequence:

```
>gi|27819608|ref|NP_776342.1| hemoglobin, beta [beta globin] [Bos taurus]
MLTAEEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPKVKAHGKKVLDSF
SNGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARNFGKEFTPVLQADFQKVVAGVANAL
AHRYH
```

- first line begins with '>', followed by gi,
  - next field surrounded by '|' is GenBank identifier
- the keyword 'ref'
  - field will be the reference for the version of this sequence.
- final field is the description

Fasta File Format

- Example Fasta Sequence:

```
>gi|27819608|ref|NP_776342.1| hemoglobin, beta [beta globin] [Bos taurus]
MLTAEEKAAVTAFWGKVKVDEVGGEALGRLLVVYPWTQRFFESFGDLSTADAVMNNPKVKAHGKKVLDSF
SNGMKHLDDLKGTFAALSELHCDKLHVDPENFKLLGNVLVVVLARNFGKEFTPVLQADFQKVVAGVANAL
AHRYH
```

- nearly all sequence based programs treat anything following the '>' as a comment
- a few sequence analysis programs expect sequences to be in a strict fasta format

Copyright 2000 N. AYDIN. All rights reserved.

GenBank

- GenBank is the National Center for Biotechnology Information’s nucleic acid and protein sequence database.
- It is the most widely used source of biological sequence data.
- GenBank file format contains information about the sequence, including literature references, functions of the sequence, locations of various features, etc.
- information organized into fields, each with an identifier, justified to the farthest left column.
- Some identifiers have additional subfields.
- sequence data lies between the identifier ORIGIN and the ‘//’ which signals the end of a GenBank record.

GenBank Record

```
LOCUS       HBB                 145 aa           linear   MAM 22-JAN-2003
DEFINITION             hemoglobin, beta [beta globin] [Bos taurus].
ACCESSION               NP_776342
VERSION                NP_776342.1  GI:27819608
DBSOURCE               REFSEQ: accession NM_173917.1
KEYWORDS               .
SOURCE                  Bos taurus (cow)
ORGANISM               Bos taurus          Eukaryota; Metazoa; Chordata; Craniata;
            Vertebrata; Euteleostomi; Mammalia; Eutheria;     Cetartiodactyla;
            Ruminantia; Pecora; Bovoidea;          Bovidae;                 Bovinae;                Bos.
REFERENCE              1  (residues 1 to 145)
AUTHORS                Duncan,C.H.
JOURNAL                Unpublished (1991)
COMMENT                PROVISIONAL REFSEQ: This record has not yet been subject to          final
                       NCBI review.  The reference sequence was derived from M63453.1.
FEATURES               Location/Qualifiers     source        1..145
```

ASN.1

- Abstract Syntax Notation (ASN.1):
  - formal description language developed to encode various data to be easily connected across computer systems
- ASN.1 is highly structured and detailed
- ASN.1 format contains all of the other information found in other formats

Copyright 2000 N. AYDIN. All rights reserved.