

Motif alignment for DNA or protein sequences

BIOL 495S/ CS 490B/ MATH 490B/
STAT 490B Introduction to Bioinformatics

March 29 & April 1, 2002

Definitions of motif

- In the context of protein structures, a motif is a simple combination of a few secondary structural elements.
- For DNA sequences, an example of motif is the transcription factor binding sites in promoter regions (see textbook Box 3.3 p146-148).
- We describe motif as a conserved segment of residues in nucleotide or amino acid sequences.

An example of protein motif, helix-turn-helix motif

- The helix-turn-helix (HTH) structure was the first DNA-binding motif discovered.
- DNA-binding proteins are responsible for replicating the genome, for transcribing active genes, and for repairing damaged DNA, etc.
- From structural studies and sequence comparisons, many DNA-binding proteins can be grouped into classes that use related motifs for recognition.
- Some families, such as HTH, were first recognized because of structural similarities. Other families were first identified by sequence comparisons and later characterized by structural studies.

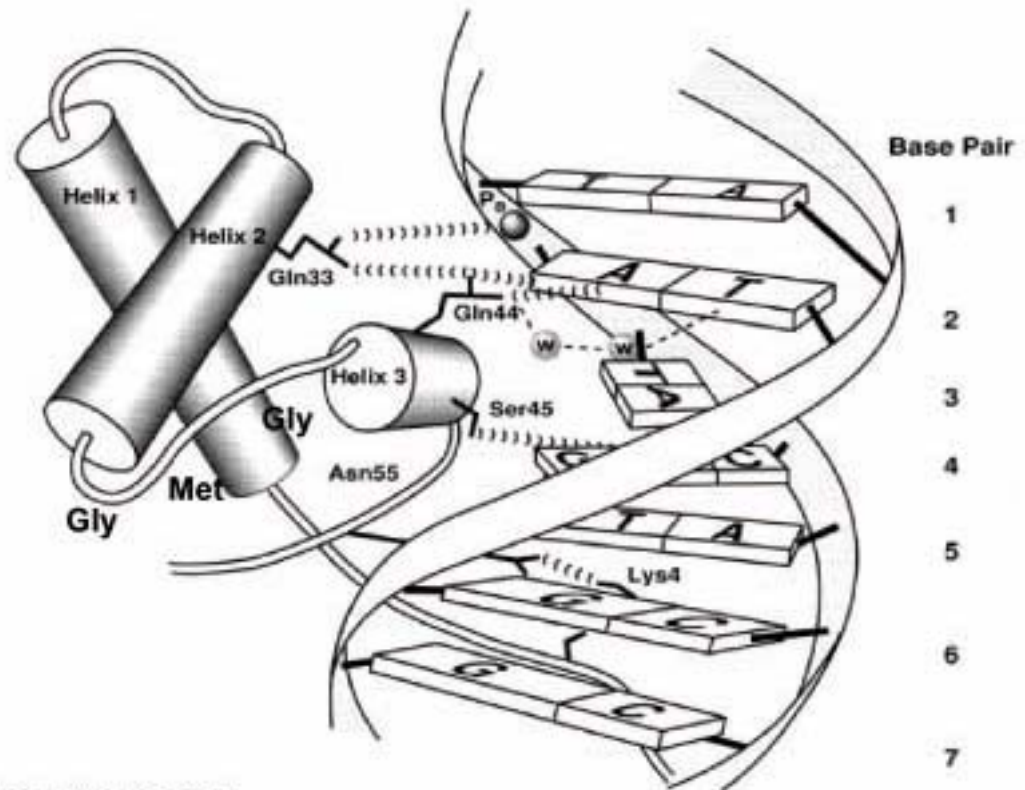
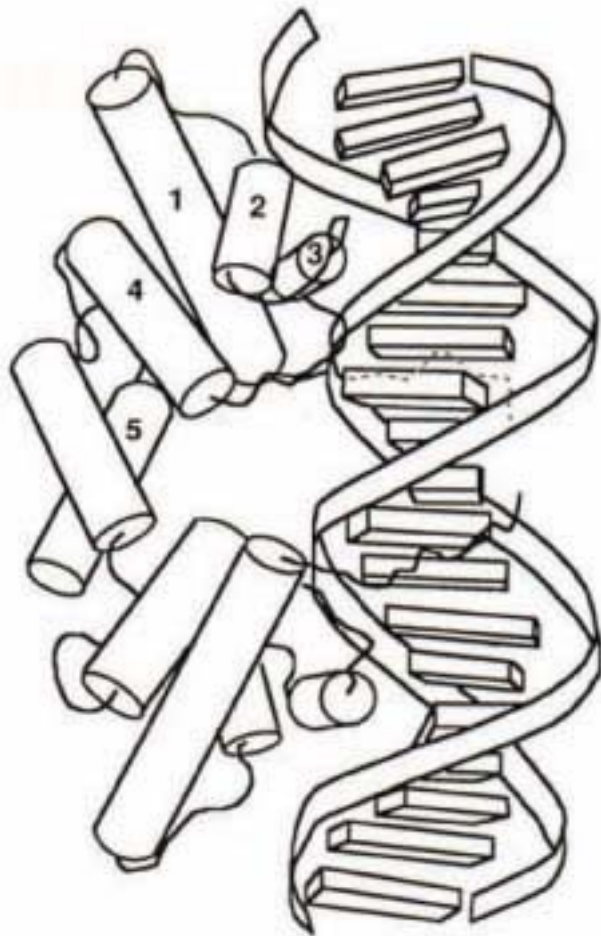
Helix-Turn-Helix (HTH) Motif

Basic structure: 2 (really 3) helices with hydrophobic core which stabilizes fold.

helix 1 - "stabilization helix"

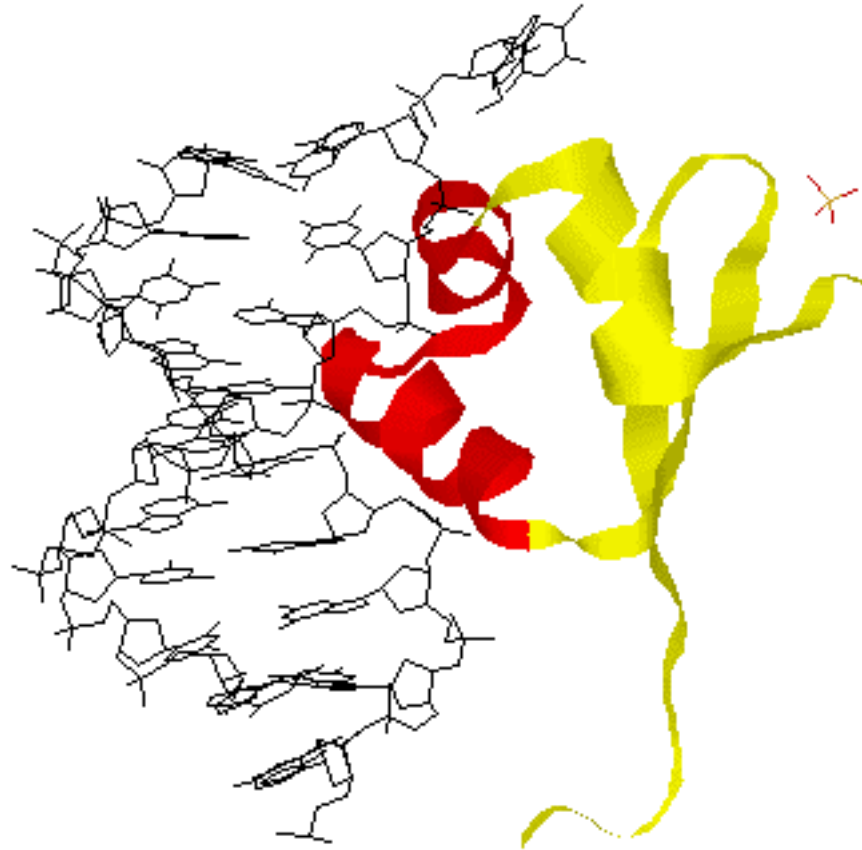
helix 2 - "positioning helix"

helix 3 - "recognition helix"



λ repressor

Cro view 2



Helix-turn-helix motif

- The first crystal structures of DNA-binding proteins were those of the λ Cro protein, the E. coli CAP protein, and the DNA-binding domain of λ repressor.
- Comparisons of these three proteins revealed a conserved recognition motif consisting of an α -helix, a turn, and a second α -helix.
- Sequence comparisons suggested HTH motif occurred in a large family of prokaryotic DNA-binding proteins. The overall sequence similarity can be very low, but the HTH motif regions have conserved residues.
- The most highly conserved residues in the HTH motif include a glycine in the turn and also several hydrophobic residues. The residues in the binding sites are critical to its structure and ability to bind to DNA.
- Recognition of these residue patterns provide information on the motif structures and functions.

Alignment of the helix-turn-helix sequences of phage and bacterial regulatory.

RQEIGQIVGCSRETVGRILK	CAP
QRELKNELGAGIATITRGSN	Trp
LYDVAEYAGVSYQTVSRVFN	Lac
IKDVARLAGVSVATVSRVIN	Gal
TRKLAQKLGVEQPTLYWHFK	Tet (Tn10)
QTRAALMMGINRGTLRKKLK	Fis protein
QAELAQKVGTTQQSIEQLEN	434 Rep
QTELATKAGVKQQSIQLIEA	434 Cro
QTKTAKDLGVYQSAINKAIH	lambda Cro
QESVADKMGMGQSGVGALFN	lambda Rep

$\alpha 2$	$\alpha 3$

Motif patterns can be described by amino acid frequencies

	Ser	Gly	Ala	Thr	Val
Some motif region	55.7%	82.2%	70.3%	54.0%	49.2%
Back-ground	6.8%	7.2%	7.8%	5.9%	6.6%

Random probability model of sequences

- Recall the random model of sequences. We assume the frequencies for amino acids are p_A, p_C, \dots, p_Y , and assume amino acids occur independently in sequences. Then the probability of observing a segment, for example VLAALMVALM, is

$$p_V^2 p_L^3 p_A^3 p_M^2$$

- Generally, the probability of observing n_1 alanine (A), n_2 cysteine (C), ..., n_{20} tyrosine (Y) in a sequence of length n , $n_1 + n_2 + \dots + n_{20} = n$ is from a multinomial distribution

with the form
$$\binom{n}{n_1, \dots, n_{20}} p_A^{n_1} \dots p_Y^{n_{20}}$$

Position-specific probability model of motifs

Amino acid distribution in the motif is assumed different from the random model. The frequencies of amino acids are position specific in the motif. Assume we have a motif with width = 3. Then the motif model can be represented in a table with amino acid frequencies for each position (column).

VGC		1	2	3
LGA	A	0.3	0	0.1
AGV	C	0	0	0.1
AGV	G	0	1	0
LGV	I	0	0	0.1
MGI	L	0.3	0	0
VGT	M	0.2	0	0.1
AGV	T	0	0	0.1
LGV	V	0.2	0	0.5
MGM	Others	0	0	0

Probability model of motifs

- Assume a motif has width w , then a $20 \times w$ matrix is used to describe the motif. Entry q_{ij} in the matrix is the frequency of observing amino acid type i in motif position j .
- When we have residue segments of a known motif, its probability model is obtained by counting residue frequencies in each column of the motif (that is maximum likelihood estimate).
- The probability model is valid under the assumptions that sequences with the common motif are independent, and positions inside the motif are independent.
- For simplicity, the motif model assumes no gap. In many cases, motifs are so conserved that gaps are not allowed for the structural or functional reason.

If we have a probability model for a motif, we can search new sequences for more motif segments. This can be done when we scan a new sequence and calculate probabilities for all possible segments with the same motif width. If a probability value for a segment is very high according to the position-specific motif model, this segment is a candidate of the motif.

While, a more challenge problem is to create a motif model from unaligned sequences. We need to align motif segments first, then build the motif model by maximum likelihood estimates described before. One common statistical strategy for identifying motifs is based on a method called Gibbs sampling. We will use DNA promoter sequences as an example this time.

Gibbs sampling strategy for motif alignment

Suppose the following sequences are promoter regions of a group of co-regulated genes (the sequences are shorter than what we would find in practice):

ACCGTGGTGT

TGGCACAAGC

GCCGATAGTC

AGTGGCGAAC

CCTGTGGTCA (sequence Z)

Initialize the iteration by selecting a random starting sequence (the last one in our example) and designating it as sequence Z.

The idea of each step in the iteration is to use the remaining sequences to find the location of the shared regulatory motif in Z.

Pick a random initial alignment for the remaining sequences and starting point for the motif of width $w=4$.

```
      * * * *  
      ACCGTGGTGT  
TGGCACAAGC  
      GCCGATAGTC  
      AGTGGCGAAC
```

For each position i , where $i=1,\dots,w$, in the alignment within the motif, tabulate the nucleotide frequencies q_{ij} :

	1	2	3	4
A	0.25	0.25	0.00	0.50
C	0.25	0.50	0.00	0.25
G	0.50	0.00	1.00	0.25
T	0.00	0.25	0.00	0.00

Compute the corresponding nucleotide frequencies p_j for the pool of sites outside the pattern:

$$\begin{aligned} p_A &= 5/24 = 0.21 & p_C &= 6/24 = 0.25 \\ p_G &= 7/24 = 0.29 & p_T &= 6/24 = 0.25 \end{aligned}$$

Select a starting point, x , for the motif in sequence Z :

* * * *

CCTGTGGTCA

Calculate the probability of the pattern using the values from the position-specific model (q's)

$$Q(x) = 0.5 \times 0.25 \times 1.0 \times 0.25 = 0.00625$$

and also the probability of the pattern using the “background” values from sequence regions outside the profile (p's)

$$P(x) = 7/24 \times 6/24 \times 7/24 \times 7/24 = 0.006203$$

The ratio of the two probabilities,

$$R(x) = Q(x) / P(x) = 1.01$$

is indicative of how likely it is that sequence Z has an example of the motif beginning at position x, so we select as the location of the motif the position with largest $R(x)$.

The current sequence Z is added to the alignment, with a new sequence designated as Z for the next cycle of the iteration.

Placement of the motif in sequence Z becomes more and more refined each iteration, and the complete sequence alignment (i.e., the placement of the motif's starting point in each sequence) eventually converges.

Why Gibbs sampling iterations work for detecting motifs

The idea driving Gibbs sampling methods is that the better the description of the motif probabilities (q 's), the more accurately the position of the motif in sequence Z can be identified. In its early stages, the locations of the motifs are chosen essentially at random. Frequencies in the p 's and q 's are similar, and consequently, values of $R(x)$ hover near 1. As the iterations continue, however, some of the motif locations are placed correctly by chance, leading to more pronounced differences between the profile frequencies in q and the background frequencies in p , higher ratio values, and better placement of the motif in that iteration's sequence Z .

For the example data, the alignment eventually converges to

```
      * * * *
      ACCGTGGTGT
      TGGCACAAGC
GCCGATAGTC
      AGTGGCGAAC
      CCTGTGGTCA
```

With q's

	1	2	3	4
A	0.00	0.20	0.00	0.00
C	0.00	0.00	0.00	0.40
G	0.00	0.80	1.00	0.00
T	1.00	0.00	0.00	0.60

The background frequencies in the final alignment are

$$\begin{aligned} p_A &= 9/30 = 0.30 & p_C &= 11/30 = 0.37 \\ p_G &= 8/30 = 0.27 & p_T &= 2/30 = 0.07 \end{aligned}$$

For consensus segment TGGT ratio of the probability from motif model and that of background is

$$\frac{q_{1T}q_{2G}q_{3G}q_{4T}}{p_T p_G p_G p_T} = \frac{1.0 \times 0.8 \times 1.0 \times 0.6}{0.07^2 \times 0.27^2} = 1343.7$$

And the ratio of a consensus segment TGGC is

$$\frac{q_{1T}q_{2G}q_{3G}q_{4C}}{p_T p_G p_G p_C} = \frac{1.0 \times 0.8 \times 1.0 \times 0.4}{0.07 \times 0.27^2 \times 0.37} = 169.5$$

The high ratios indicate the detected motif is a good one that is highly different from random null model.

More on Gibbs sampling motif alignment

Further details and extension of Gibbs sampling for motif alignment include assigning pseudocounts when computing frequencies, methods for selecting motif width, aligning multiple motifs, and adding gaps in the alignment.

Other motif alignment approaches

- Conserved domain database (CDD) in the NCBI website is a collection of known motifs. CDD currently contains domains derived from two popular collections, [Smart](#) and [Pfam](#), plus contributions from colleagues at NCBI. The source databases also provide descriptions and links to citations.
- Position-specific score matrices are prepared from the underlying known motifs. When we input a query sequence, it is compared to position-specific score matrices to detect a domain hit.
- In Smart and Pfam servers, HMM models are derived for known motifs and used to detect motifs of a new sequence.

CD-Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>

NCBI **CD-Search** Entrez ?

RPS-BLAST 2.2.2 [Dec-14-2001]

Query= [gi|306460|gb|AAA58376.1](#) achaete scute protein
(238 letters)

Database: oasis_sap.v1.54
3693 PSSMs; 718,011 total columns

Mouse-over boxes to display more information

NEW Show other proteins containing these domains

- .. This CD alignment includes 3D structure. To display structure, download [Cn3D v3.00!](#)

PSSMs producing significant alignments:

	Score	E
	(bits)	value
• gnl Smart smart00353 HLH, helix loop helix domain	55.8	3e-09
• gnl Pfam pfam00010 HLH, Helix-loop-helix DNA-binding domain	51.6	5e-08

- [gnl|Smart|smart00353](#), HLH, helix loop helix domain

Add query to multiple alignment, display sequences

CD-Length = 53 residues, 81.1% aligned
Score = 55.8 bits (133), Expect = 3e-09

```

Query: 136  VNLGFATLREHVPNGAANKKMSKVETLRSAVEYIRALQQLLDE 178
Sbjct: 11  INEAFDELRLSLPLPNKKLSKASILRLAIDYIKSLQEQLQK 53
  
```

Internet

Start | American S... | CatchWord... | Exceed | dlock | odds.stat | odds.stat | xbiff | motif.ppt | CD-Search... | 4:25 PM

Position-specific score matrix

- For each position in the derived motif, every amino acid is assigned a score.
- The scores are usually derived from standard substitution score matrices, BLOSUM or PAM.
- Highly conserved residue at a particular position is assigned a high positive score, and others are high negative scores.
- At weak conserved positions, all residues receive scores near zero.

Ten columns from the multiple alignment of seven globin protein sequences

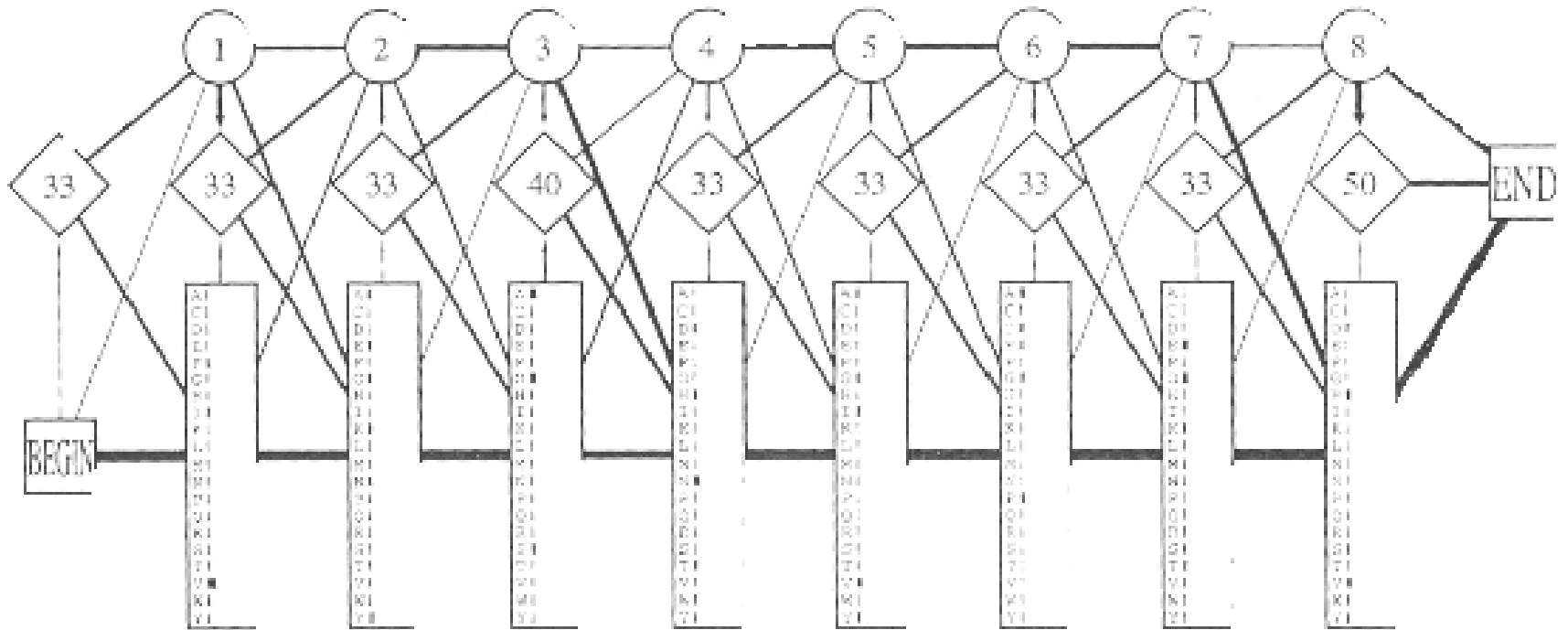
```
HBA_HUMAN      . . . VGA--HAGEY . . .
HBB_HUMAN      . . . V-----NVDEV . . .
MYG_PHYCA      . . . VEA--DVAGH . . .
GLB3_CHITP     . . . VKG-----D . . .
GLB5_PETMA     . . . VYS--TYETS . . .
LGB2_LUPLU     . . . FNA--NIPKH . . .
GLB1_GLYDI     . . . IAGADNGAGV . . .
```

The score for residue a in column 1 can be set to

$$\text{score}(1, a) = \frac{5}{7}s(V, a) + \frac{1}{7}s(F, a) + \frac{1}{7}s(I, a)$$

Analogous to the probability model of motif, for motif with width w , a $20 \times w$ matrix is derived as a position specific score matrix (PSSM). However, PSSM is a non-probability model of a motif, and it depends on the regular score matrix of BLOSUM or PAM.

We can use dynamic programming to align a sequence to the PSSM for searching motif in the query sequence.



A hidden Markov model derived from the alignment of seven globin protein sequences. Emission probabilities are shown as bars opposite the amino acids, and transition probabilities are indicated by the thickness of the lines.

Comparison of motif alignment to other alignment approaches

- Needleman-Wunsch algorithm, global alignment for a pair of sequences by dynamic programming.
- Smith-Waterman algorithm, local alignment for a pair of sequences by dynamic programming.
- BLAST, many pairwise alignments of one query sequence against a sequence database. It is local alignment however.
- ClustalW, multiple sequence alignment, from end-to-end. There is no significance measure of the multiple alignment result. The multiple alignment is obtained by progressively implementing pairwise alignments.
- Motif alignment, local multiple alignment. For diverse sequences, it is the best way to detect local conserved regions.

References

- Textbook Box3.3, p146-148.
- Lawrence et al. (1993), Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science Vol. 262, 208-214.