Sequence Analysis 3A:

Introduction to sequence motifs

Katherine Dougan, PhD

Australian Centre for Ecogenomics
School of Chemistry & Molecular Biosciences
The University of Queensland

SCIE2100 | BINF6000 | Bioinformatics I - Introduction

Outline

- **Introduction to sequence motifs**
  - What are they?
  - What makes them difficult to identify?
- **Discrete representations**
  - Consensus sequences
  - Degenerate consensus sequences
  - Regular expressions
- **Examples of sequence motifs**

What is a *sequence motif*?

A sequence motif is a ***short, conserved*** nucleotide or amino acid sequence that is known, or predicted, to have a ***specific biological function***.

Sequence motif ≠ Structural motif

(At least not necessarily)

Sequence motifs can be difficult to find

```
 Seq 1: TCATTGGTCCTCAGGATCACGCGACAGGAAGTGTGGCGTAACCGGTTGACTGCCACATGCGCATTGGCTTCCAGGGCC
 Seq 2: TCATTGATGCGCATTGGCTTCCAGGGGTCCTCAGGATCACACCACAGCCCGCTGGAATAAGTGTGGCGTAAGCCACCC
 Seq 3: CACACCTTTAATTGTTGCAGGATGAATCAGAGGAGGTGTGGCAGTAAACAAGAATGAACCCCACAGCTTCACACTTCC
 Seq 4: TACTGGCGCCAGAGCCAATTTGCGTCATCTAACTAAAGATTTCAACAGCAGTGATATATCTTACTCAAGTGTGGCTAG
 Seq 5: CAAGGAGTGTGGATACAAAATTGCGCAACAGAGAGCCATCACTCACCACAGCCCGCTGGAATTCTCATGCTGGAGCAA
 Seq 6: CATTGTTGCAGGACCACAGCTTCGAGGTGTGGCAACACCTTTAGTAAACACTTCCTGAATCAGAGACAAGAATGAACC
 Seq 7: ACACATCCGTGTGGCGATTGGGCGGCGTAACCTCGCTTATTTGCATAGGCCGATTGCACAACCGGGCGGCGACCTCAG
 Seq 8: ACACGGCCGATTGCACAACCACCGACCTCAGTGGGAACCGGGCGGATCCGTGTGGCGATTGAGGCCGATTGCACCTC
 Seq 9: TTAAGAGAATGTCATTGCGGTGTGGCTGAGGGGAGGGAGAGGTGAGGGTGCAACTTGGGTAAAGGTTGTGGAGGCAT
Seq 10: TACCACTAGCTGCCCTAACTCTTACTAATTAGCGCCAGAGACAAAGTGTGGTTCAAAGATGCAGTGATATATTGCGTC
```

There’s a DNA-binding motif here…. can you find it?

Sequence motifs can be difficult to find

Seq 1: TCATTGGTCCTCAGGATCACGCGACAGGAA**ATGTGG**CGTAACCGGTTGACTGCCACATGCGCATTGGCTTCCAGGGCC

Seq 2: TCATTGATGCGCATTGGCTTCCAGGGGTCCTCAGGATCACACCACAGCCCGCTGGAATAA**GTGTGG**CGTAAGCCACCC

Seq 3: CACACCTTTAATTGTTGCAGGATGAATCAGAGGAG**GTGTGG**CAGTAAACAAGAATGAACCCCACAGCTTCACACTTCC

Seq 4: TACTGGCGCCAGAGCCAATTTGCGTCATCTAACTAAAGATTTCAACAGCAGTGATATATCTTACTCAA**GTGTGG**CTAG

Seq 5: CAAGGA**GTGTGG**ATACAAAATTGCGCAACAGAGAGCCATCACTCACCACAGCCCGCTGGAATTCTCATGCTGGAGCAA

Seq 6: CATTGTTGCAGGACCACAGCTTCGAG**GTGTGG**CAACACCTTTAGTAAACACTTCCTGAATCAGAGACAAGAATGAACC

Seq 7: ACACATCC**GTGTGG**CGATTGGGCGGCGTAACCTCGCTTATTTGCATAGGCCGATTGCACAACCGGGCGGCGACCTCAG

Seq 8: ACACGGCCGATTGCACAACCACCGACCTCAGTGGAACCGGGCGGGATCC**GTGTGG**CGATTGAGGCCGATTGCACCTC

Seq 9: TTAAGAGAATGTCATTGCGGT**GTGGCTG**AGGGGGAGGGAGAGGTGAGGGTGCAACTTGGGTAAAGGTTGTGGAGGCAT

Seq 10: TACCACTAGCTGCCCTAACTCTTACTAATTAGCGCCAGAGACAAA**GTGTGG**TTCAAAGATGCAGTGATATATTGCGTC

Their locations can *vary* and are *not obvious*…
This makes our jobs of finding them more difficult

Sequence motifs can be difficult to find

Seq 1: TCATTGGTCCTCAGGATCACGCGACAGGAA**ATGTGG**CGTAACCGGTTGACTGCCACATGCGCATTGGCTTCCAGGGCC

Seq 2: TCATTGATGCGCATTGGCTTCCAGGGGTCCTCAGGATCACACCACAGCCCGCTGGAATAA**GTGAGG**CGTAAGCCACCC

Seq 3: CACACCTTTAATTGTTGCAGGATGAATCAGAGGAG**GTCTGG**CAGTAAACAAGAATGAACCCCACAGCTTCACACTTCC

Seq 4: TACTGGCGCCAGAGCCAATTTGCGTCATCTAACTAAAGATTTCAACAGCAGTGATATATCTTACTCAA**GTGTCG**CTAG

Seq 5: CAAGGA**GTGTGG**ATACAAAATTGCGCAACAGAGAGCCATCACTCACCACAGCCCGCTGGAATTCTCATGCTGGAGCAA

Seq 6: CATTGTTGCAGGACCACAGCTTCGAG**GTGTAG**CAACACCTTTAGTAAACACTTCCTGAATCAGAGACAAGAATGAACC

Seq 7: ACACATCC**GTGTGA**CGATTGGGCGGCGTAACCTCGCTTATTTGCATAGGCCGATTGCACAACCGGGCGGCGACCTCAG

Seq 8: ACACGGCCGATTGCACAACCACCGACCTCAGTGGGAACCGGGCGGGATCC**GTGTGG**CGATTGAGGCCGATTGCACCTC

Seq 9: TTAAGAGAATGTCATTGCGGT**GTGGGTG**AGGGGGAGGGAGAGGTGAGGGTGCAACTTGGGTAAAGGTTGTGGAGGCAT

Seq 10: TACCACTAGCTGCCCTAACTCTTACTAATTAGCGCCAGAGACAAA**TTGTGG**TTCAAAGATGCAGTGATATATTGCGTC

The sequences of the motifs can also vary on less important bases.

This makes identifying them bioinformatically even more challenging.

Sequence motifs can be difficult to find

Seq 1: TCATTGGTCCTCAGGATCACGCGACAGGAA**ATGTGG**CGTAACCGGTTGACTGCCACATGCGCATTGGCTTCCAGGGCC

Seq 2: TCATTGATGCGCATTGGCTTCCAGGGGTCCTCAGGATCACACCACAGCCCGCTGGAATAA**GTGAGG**CGTAAGCCACCC

Seq 3: CACACCTTTAATTGTTGCAGGATGAATCAGAGGAG**GTCTGG**CAGTAAACAAGAATGAACCCCACAGCTTCACACTTCC

Seq 4: TACTGGCGCCAGAGCCAATTTGCGTCATCTAACTAAAGATTTCAACAGCAGTGATATATCTTACTCAA**GTGTCG**CTAG

Seq 5: CAAGGA**GTGTGG**ATACAAAATTGCGCAACAGAGAGCCATCACTCACCACAGCCCGCTGGAATTCTCATGCTGGAGCAA

Seq 6: CATTGTTGCAGGACCACAGCTTCGAG**GTGTAG**CAACACCTTTAGTAAACACTTCCTGAATCAGAGACAAGAATGAACC

Seq 7: ACACATCC**GTGTGA**CGATTGGGCGGCGTAACCTCGCTTATTTGCATAGGCCGATTGCACAACCGGGCGGCGACCTCAG

Seq 8: ACACGGCCGATTGCACAACCACCGACCTCAGTGGGAACCGGGCGGGATCC**GTGTGG**CGATTGAGGCCGATTGCACCTC

Seq 9: TTAAGAGAATGTCATTGCGGT**GTGGGTG**AGGGGGAGGGAGAGGTGAGGGTGCAACTTGGGTAAAGGTTGTGGAGGCAT

Seq 10: TACCACTAGCTGCCCTAACTCTTACTAATTAGCGCCAGAGACAAA**TTGTGG**TTCAAAGATGCAGTGATATATTGCGTC

We can describe a motif ***qualitatively***…

(Only *presence* versus *absence* of letters at this point…
no numerical information)

Discrete representations of motifs

- Consensus sequences
- Degenerate consensus sequences
- Rule / regular expressions

We can describe a motif ***qualitatively***…

(Only *presence* versus *absence* of letters at this point…
no numerical information)

Consensus sequences

A simple way of representing a motif, is by using a **consensus sequence**, or the most common letter at each position

However, this **does not allow for variability** in identifying positions where there are multiple options for the motif

Allowing for **degeneracy** in the consensus sequence does allow for some additional flexibility…

Alignment:

```
AATGCGGA
AATGTGGC
ACTGTGGC
CGTGTGGC
CGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGG
GGTGTGGG
```

Consensus: GGTGTGGC

Degenerate consensus sequences

This is still limiting, however, as there is no quantitative measure for variability at the different positions

R	A|G	puRine
Y	C|T	pYrimidine
S	G|C	Weak (weaker basepairs, fewer hydrogen bonds)
W	A|T	Strong (stronger basepairs, more hydrogen bonds)
K	G|T	Keto (both have a keto group)
M	A|C	aMine (both have an amine group)
B	C|G|T	not A (B comes after A)
D	A|G|T	not C (D comes after C)
H	A|C|T	not G (H comes after G)
V	A|C|G	not T or U (V comes after T and U)
N	A|C|G|T	aNy base

Alignment:

```
AATGCGGA
AATGTGGC
ACTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGA
GGTGTGGA
```

Consensus: GGTGTGGC

Degenerate Consensus: RVTGYGGM

Describing motifs using regular expressions

We can also describe a motif using a ***rule*** or ***regular expression…***

A – matches A
[AT] – matches A or T
{ AT } – matches neither A or T (i.e. G or C)
x – matches any symbol
x(3) – any 3 symbols

```
AATGCGGA
AATGTGGC
ACTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGC
GGTGTGGA
GGTGTGGA
```

[AG]{T}TG[CT]GG[AC]

Describing motifs using regular expressions

We can also describe a motif using a ***rule*** or ***regular expression…***

A – matches A
[AT] – matches A or T
{ AT } – matches neither A or T (i.e. G or C)
x – matches any symbol
x(3) – any 3 symbols

```
LVIEMLY
LVIECLY
LVIECLF
LVIEMLF
LVIEMLF
LVIEMLF
LVIEMLF
LVLEMLF
LVVEMLF
LVIEMLY
```

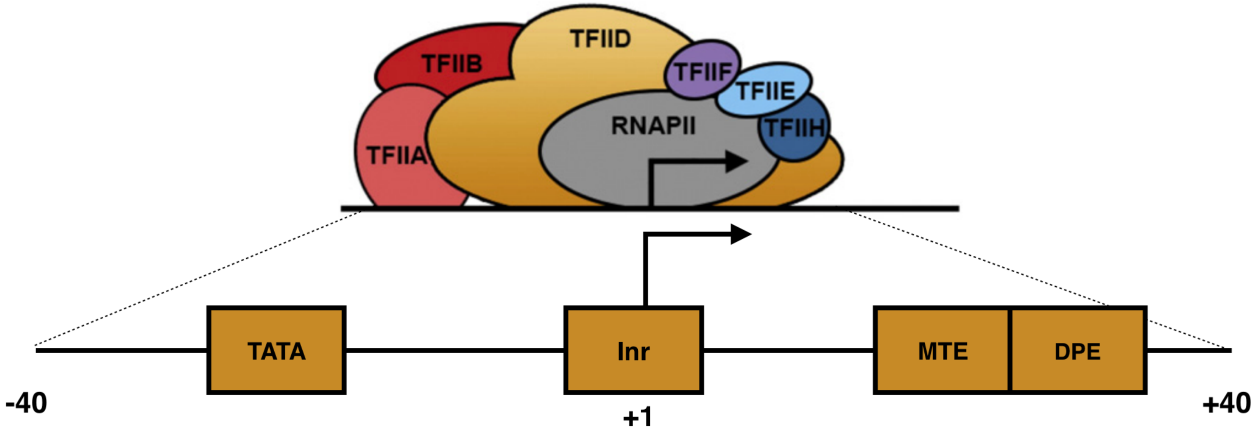
LV[ILV]E[MC]L[FY]

Like before, we are still limited in how much information this conveys. However, this can be informative in certain cases…

DNA: Transcription factor DNA-binding motifs

- Inr – Initiator Element
- MTE – Motif Ten Element
- DPE – Downstream Core Promoter Element

- MTE promotes transcription by RNA polymerase II
- MTE requires Inr, but independent of TATA and DPE
- MTE can compensate for mutations in TATA and DPE



Sequence Element	Approximate Position	Consensus Sequence
TATA Box	-30 to -23	TATAWAW
Inr	overlaps the transcription start site (+1)	BBCA$_{(+1)}$BW
MTE	+18 to +29	CSARCSSAACGS
DPE	+28 to +33	DSWYVY

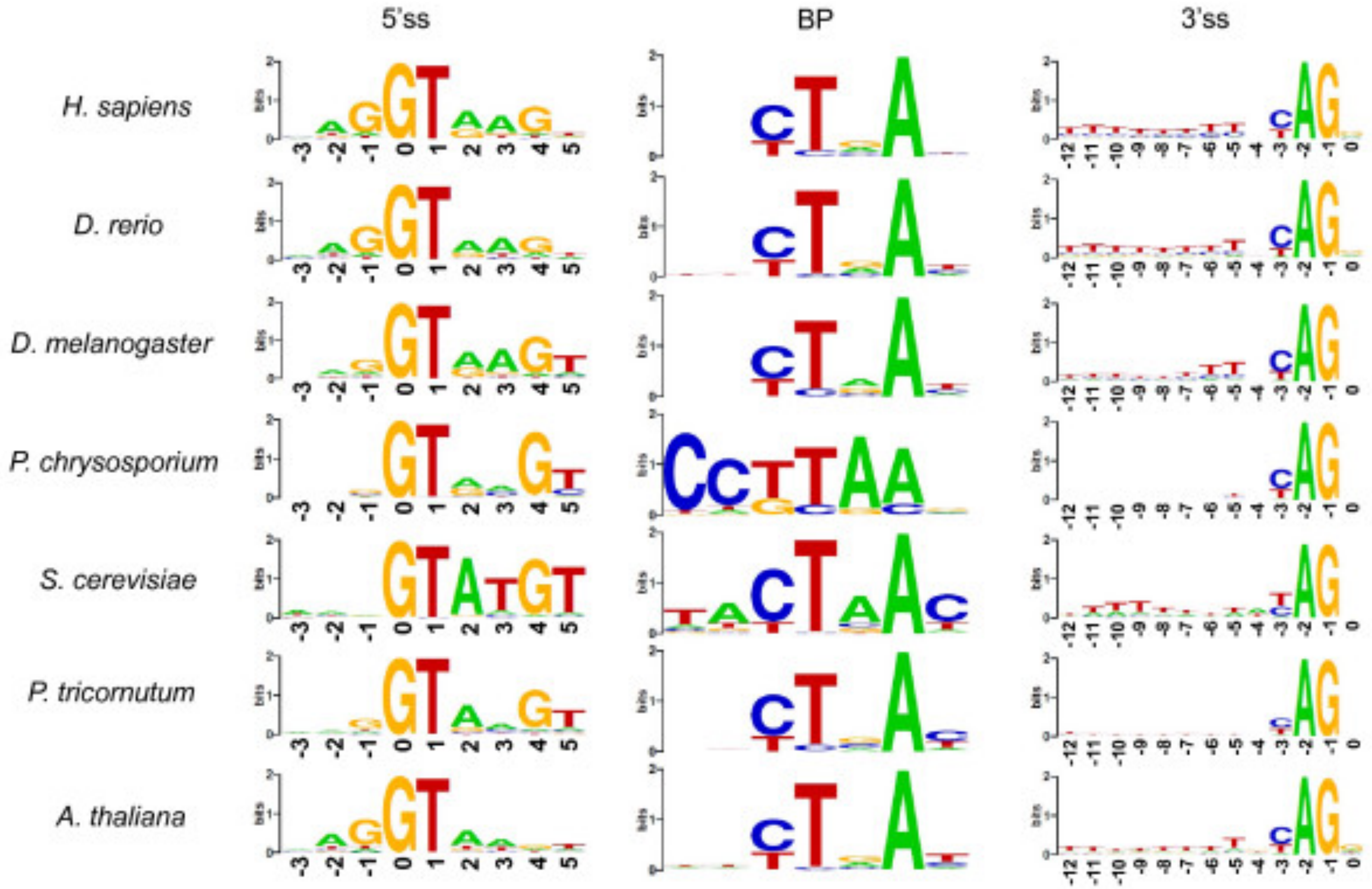
Nucleotide positions (column 2) are all relative to the A (+1) of the Inr.
IUPAC codes: D=A/G/T, S=G/C, R=A/G, V=A/C/G, W=A/T, Y=C/T, B= C/G/T

Iwata and Gotoh et al. 2011 in *BMC Genomics* 12:45

RNA: mRNA splice site motifs

Motifs for RNA splicing are generally very conserved across eukaryotes

If you're curious about exemptions to this.. read about dinoflagellate splice sites



Iwata and Gotoh et al. 2011 in *BMC Genomics* 12:45

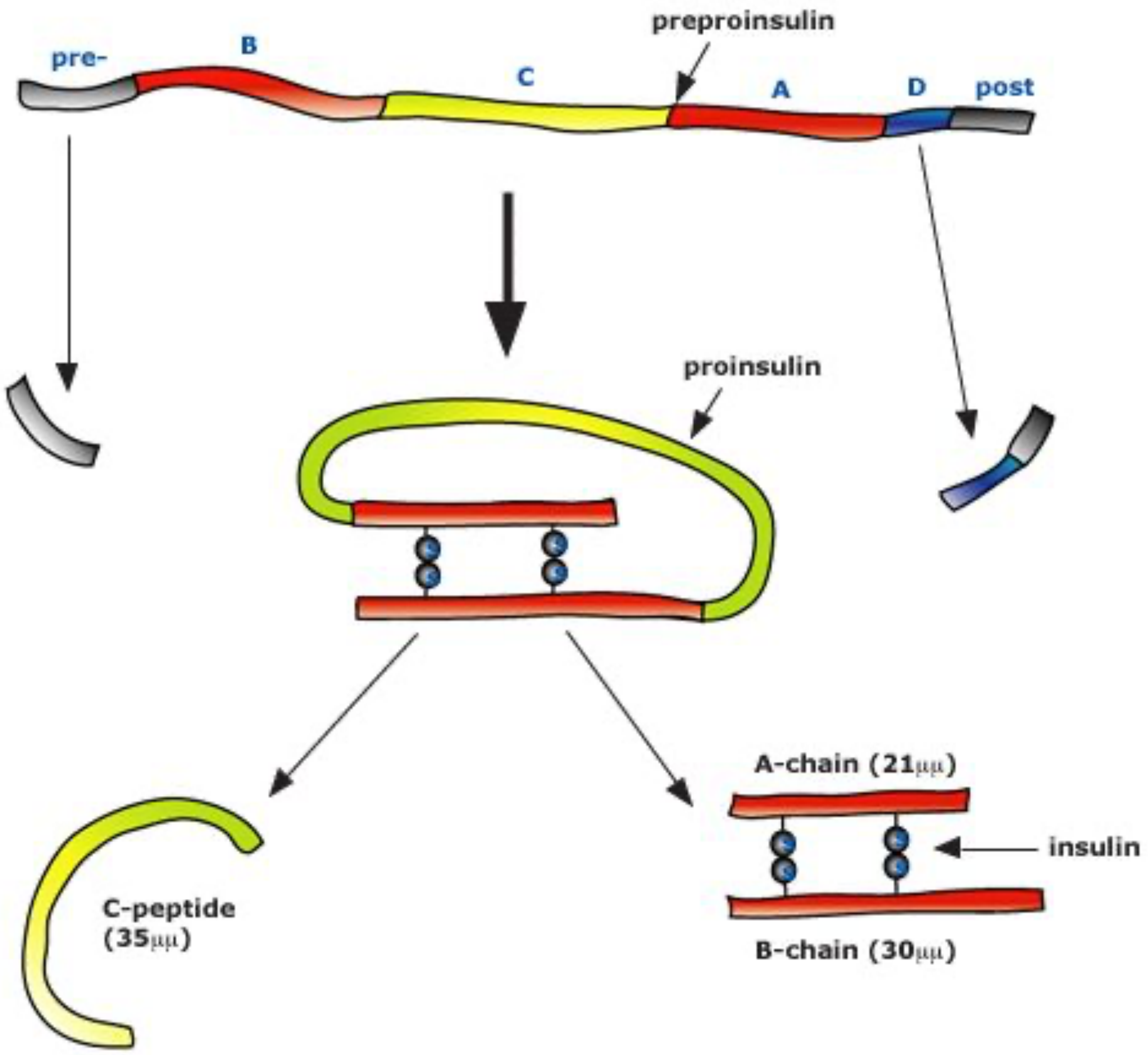
Protein: the insulin motif

Most of the insulin peptide gene sequence is not present in the final peptide product.

Even of the lingering sequence, only 8 cysteine residues are highly conserved as they are the location of the disulfide bonds that link the two chains

Distribution in cysteine residues is how they are classified in invertebrates

Chain A motif: x(5)-CC-x(3)-C-x(8)-C-x
Chain B motif: x(6)-C-x(11)-C-x(13)



Reflection

- *What are sequence motifs?*
- *What aspects of sequence motifs make them difficult to accurately locate and describe?*
- *What are the limitations of a discrete representation for a motif and why?*
- *How are exact consensus sequences, degenerate consensus sequences and regular expressions alike and dissimilar?*