BMEG3102 Bioinformatics

Lecture 6. Motifs and Domains

Qi Dou

Email: qidou@cuhk.edu.hk

Office: Room 1014, 10/F, SHB

BMEG3102 Bioinformatics

The Chinese University of Hong Kong

Lecture outline

1. General definitions

2. DNA motifs

3. Protein domains

Part 1

General Definitions

Protein binding sites

- Suppose a protein can bind certain locations in each of the following sequences:
  - AACCCGATACAGACGACCATTACGACC
  - GAGACGACATACATTACACCAA
  - CCGACTAAACCAGATACAGAGATTACAGCATAC
  - ACATCCATACAGACAAAAACATAGAGGGACGATT

- Where do you think the protein binds?
  - Assumption: The protein recognizes a certain pattern of its binding sites (due to shape and energy)

Protein binding sites

- Would be easier if you know the binding site in one of the sequences:
  - AACCCGATACAGACGACCATTACGACC
  - GAGACGACATACATTACACCAA
  - CCGACTAAACCAGATACAGAGATTACAGCATAC
  - ACATCCATACAGACAAAAACATAGAGGGACGATT

Protein binding sites

- From the example, you can use sequence alignment or other methods to find out the binding site in the other sequences:
  - AACCCGATACAGACGACCATTACGACC
  - GAGACGACATACATTACACCAA
  - CCGACTAAACCAGATACAGAGATTACAGCATAC
  - ACATCCATACAGACAAAAACATAGAGGGACGATT

- Notice that:
  - The different occurrences could be slightly different
  - There may be multiple binding sites in one sequence

Motifs and domains

- In general, we define motifs/domains as patterns that
  1. Appear frequently
     - May not be exactly the same in different occurrences, but highly similar
     - Are unlikely to occur “by chance”. In other words, they are “over-represented”
  2. Usually have known or predicted functional roles
  3. Are evolutionarily conserved

- There are many types of motifs and domains

Motifs and domains: Examples

DNA sequence motifs:

NAME	START	SITES	END	STRAND	MARGINAL SCORE
1	23	aaacgacgtaatgctacg	6	-	22.9
2	17	aaacgac.taat.ctacg	2	-	8.45
3	30	caacgaggta.tgcaacg	14	-	14.1
4	23	caaccacgtaatgcaacg	6	-	23.6
5	24	caaccacgtaatgcatag	7	-	17.5

Score: **51.1928**

logo with ssc

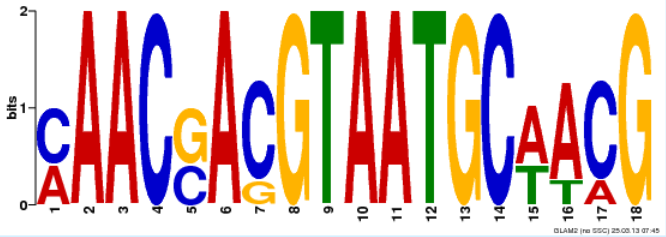


Image sources: http://rosalind.info/media/problems/meme/logo1.png, Wikipedia, Rfam

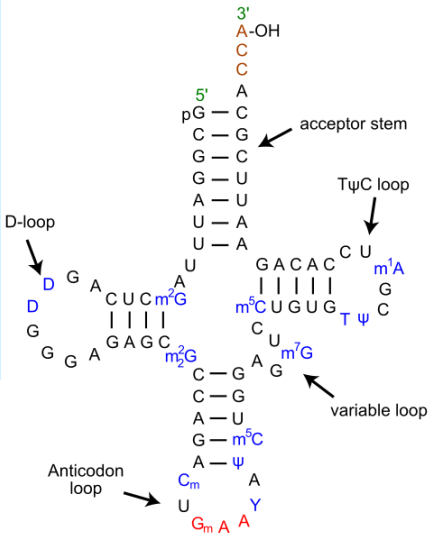
Motifs and domains: Examples

DNA sequence motifs:

NAME	START	SITES	END	STRAND	MARGINAL SCORE
1	23	aaacgacgtaatgctacg	6	-	22.9
2	17	aaacgac.taat.ctacg	2	-	8.45
3	30	caacgaggta.tgcaacg	14	-	14.1
4	23	caaccacgtaatgcaacg	6	-	23.6
5	24	caaccacgtaatgcatag	7	-	17.5

Score: **51.1928**

logo with ssc



RNA families:

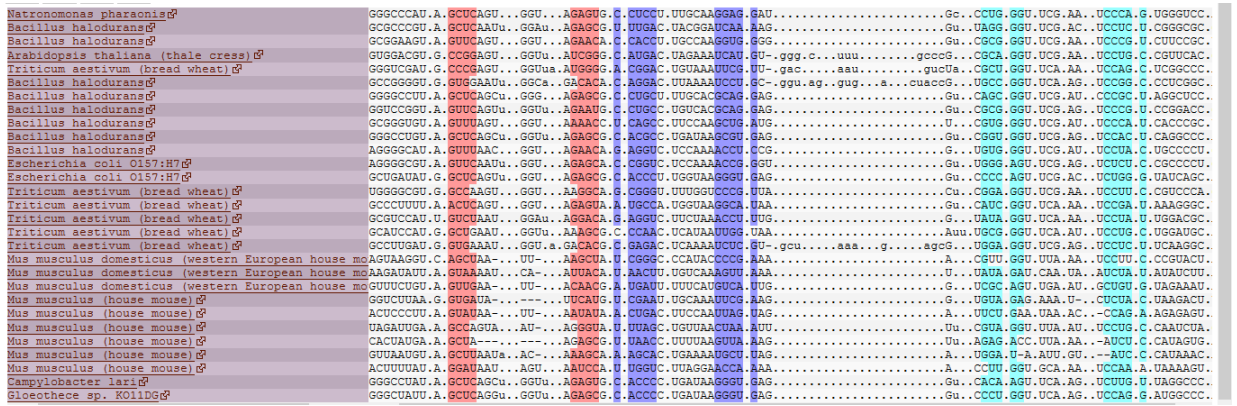
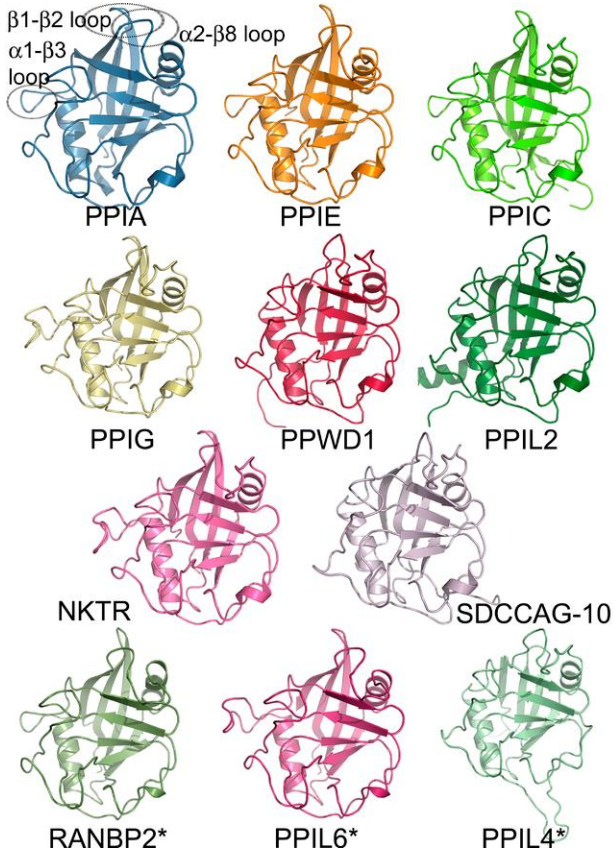


Image sources: http://rosalind.info/media/problems/meme/logo1.png, Wikipedia, Rfam

Motifs and domains: Examples

Protein domains:

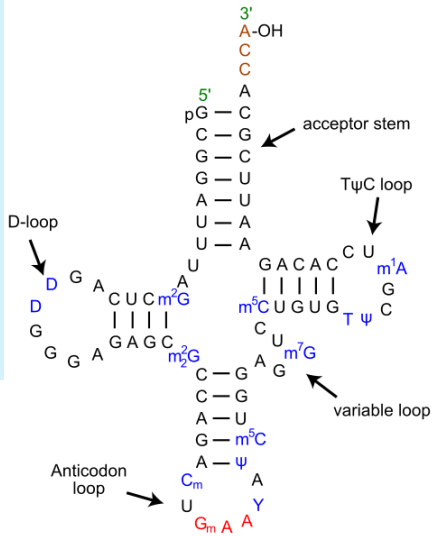


DNA sequence motifs:

NAME	START	SITES	END	STRAND	MARGINAL SCORE
1	23	aaacgacgtaatgctacg	6	-	22.9
2	17	aaacgac.taat.ctacg	2	-	8.45
3	30	caacgaggta.tgcaacg	14	-	14.1
4	23	caaccacgtaatgcaacg	6	-	23.6
5	24	caaccacgtaatgcatag	7	-	17.5

Score: **51.1928**

logo with ssc



RNA families:

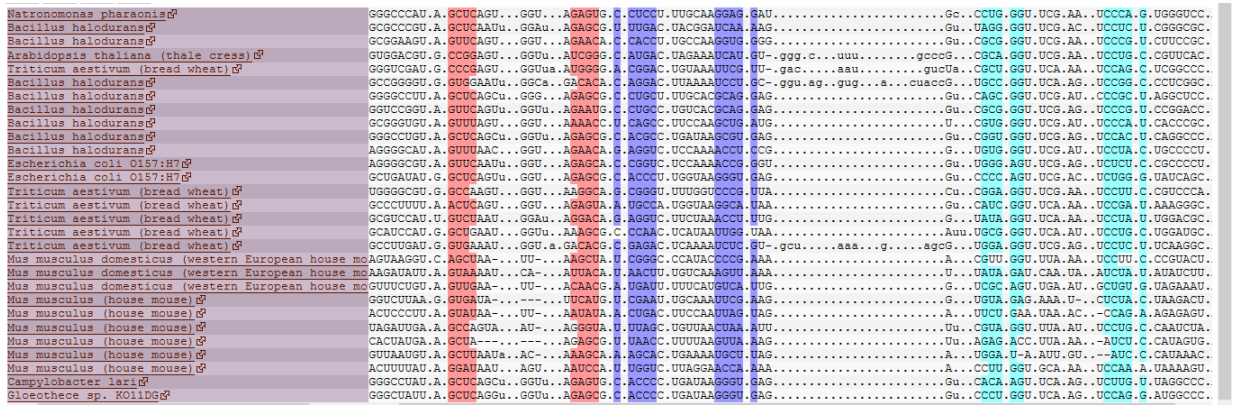


Image sources: http://rosalind.info/media/problems/meme/logo1.png, Wikipedia, Rfam

DNA motifs and protein domains

- Focuses of this lecture:
  - Transcription factor binding sites, which are short DNA regulatory sequences that frequently appear at specific genomic locations. Some of them are conserved across species.
  - Protein domains, which are similar sub-sequences on different proteins that serve particular functions. Again, some of them are evolutionarily conserved.

Motifs and domains: Differences

- In the literature, sometimes motifs and domains are distinguished by the followings:
  - A domain is assumed to possess certain functional or structural independence. A motif may not.
  - A domain is usually larger than a motif.

- In this lecture, we use these two terms more or less interchangeably.

Part 2

DNA Motifs

Regulatory regions

- DNA binding proteins (e.g., transcription factors) bind different types of DNA elements for different purposes. For example:
  - Promoters (around the transcription start site): To help the formation of the transcription machinery
  - Enhancers (usually further away from a gene): To enhance the expression of a gene
  - Silencers: To inhibit the expression of a gene
  - Insulators: To mark expression boundaries

Regulatory regions

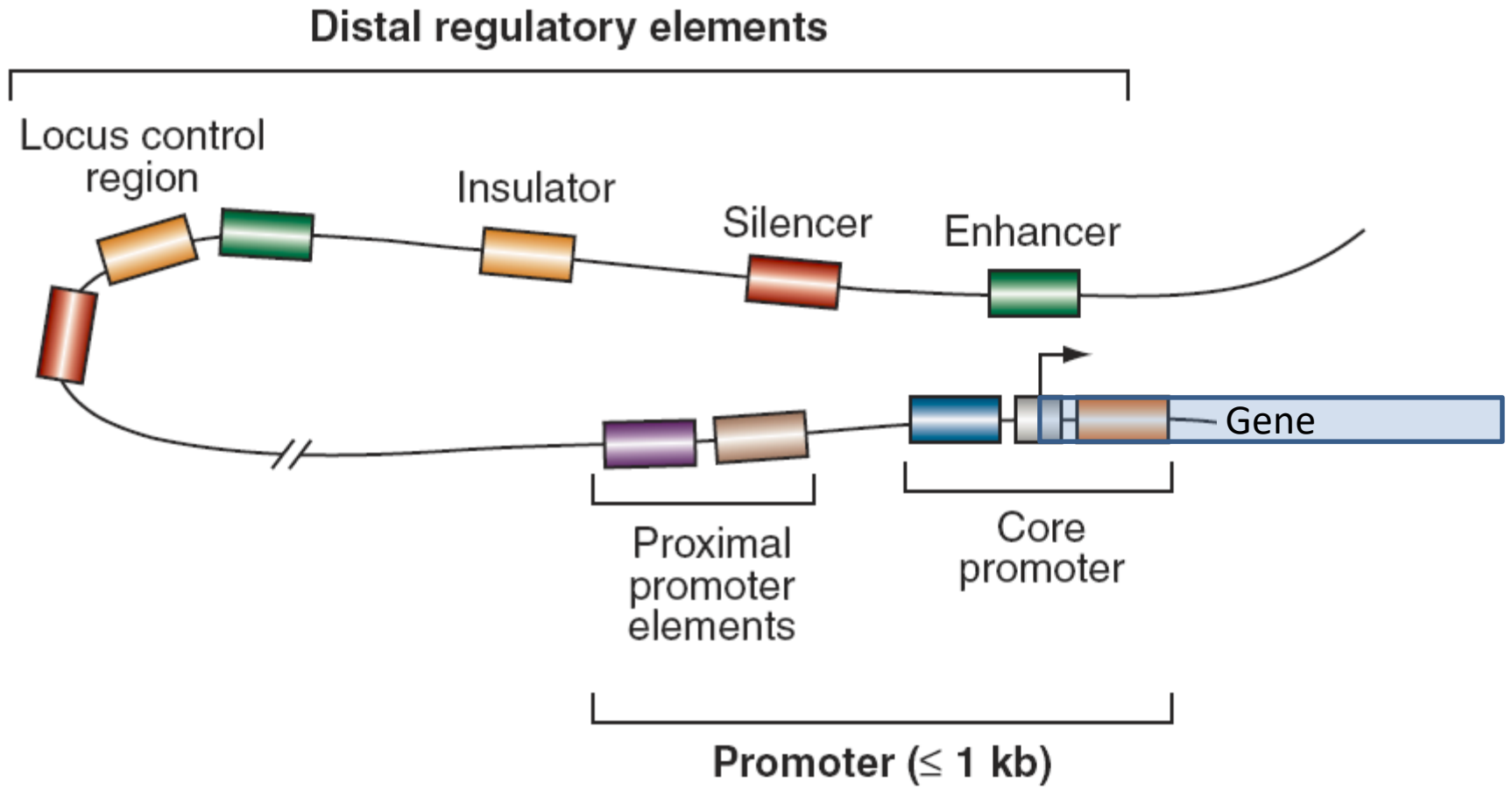


Image credit: Maston et al., *Annual Review of Genomics and Human Genetics* 7:29-59, (2006)

Transcription factor binding

- Where does a transcription factor bind?
  - Where DNA is accessible
  - Where there are special signals on the DNA (e.g., lack of methylation) and the surrounding proteins (e.g., histone modifications)
  - Where the DNA structure is suitable
  - Where the DNA sequence is suitable ← our focus
    - The DNA region bound by a transcription factor is called a transcription factor binding site (TFBS)
    - Usually quite short (e.g., 6-10bp)

Transcription factor binding

- Specific regions of transcription factors called the DNA binding domains recognize and bind the TFBS

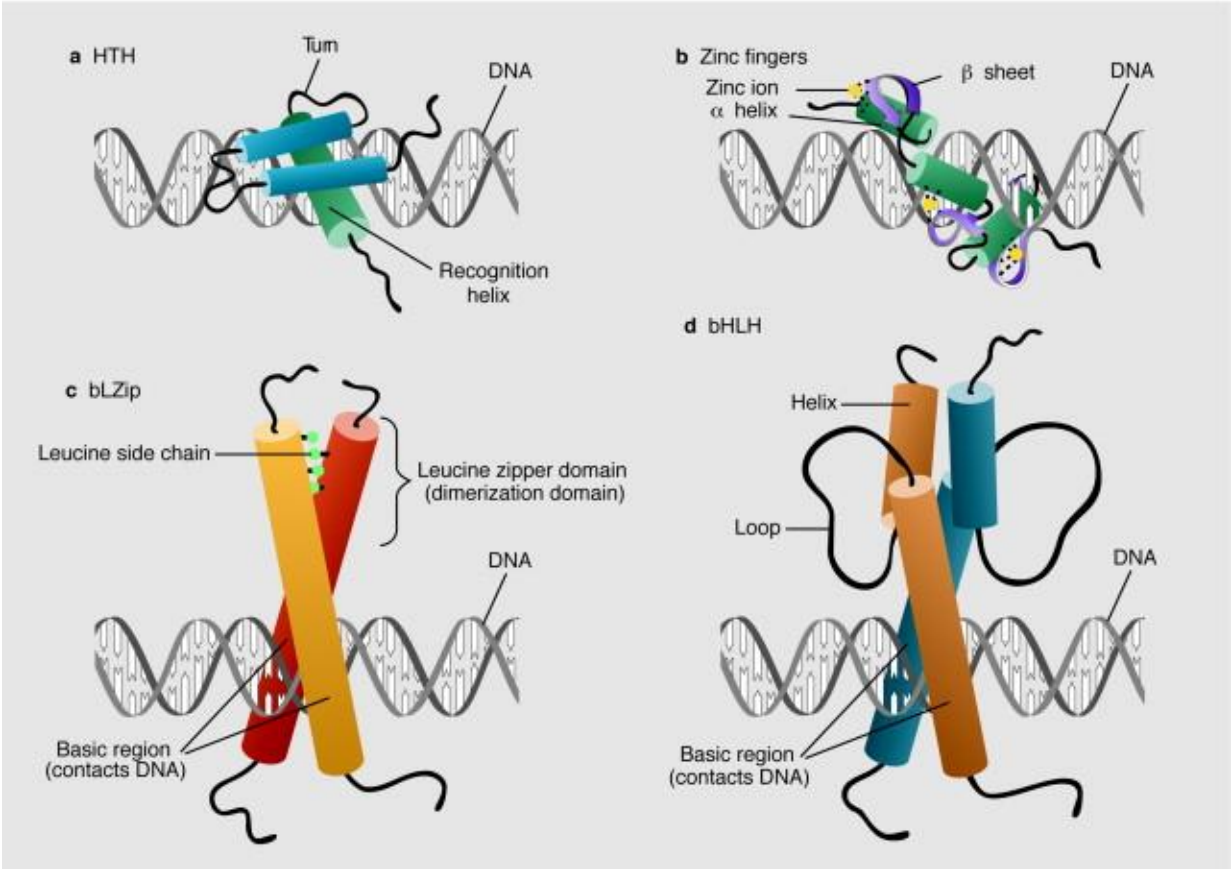


Image credit: Papavassiliou, *Molecular medicine Today* 4(8):358-366, (1998)

Representing motifs and domains

- How to represent a TFBS?
- If the pattern is very conserved, may use an exact representation
  - E.g., consensus sequence
- In most cases, need to capture the differences by statistical representations
  - E.g., position weight matrix
- For a representation that is more complex,
  - It can capture more information
  - It involves more parameters
    - Needs more data and time to estimate parameter values
    - Is more prone to over-fitting

Representations (patterns)

- Suppose we have the following TFBS sequences:
  - CACAAAC
  - CACAAAT
  - CGCAAAC
  - CACAAAC
- Consensus sequence:
  - CACAAAC
  - Problem: Information loss

Example source: http://conferences.computer.org/bioinformatics/CSB2003/NOTES/Liu_Color.pdf

Representations (patterns)

- Suppose we have the following TFBS sequences:
  - CACAAAC
  - CACAAAT
  - CGCAAAC
  - CACAAAC
- Consensus sequence:
  - CACAAAC
  - Problem: Information loss
- Degenerate sequence in IUPAC (International Union of Pure and Applied Chemistry) code (see http://www.bio-soft.net/sms/iupac.html):
  - CRCAAAY

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap (not used in motifs)

Example source: http://conferences.computer.org/bioinformatics/CSB2003/NOTES/Liu_Color.pdf

Representations (patterns)

- Suppose we have the following aligned TFBS sequences:
  - `CACAAAAC`
  - `CACAAA_T`
  - `CGCAAAAC`
  - `CACAAA_C`
- Regular expression (see http://en.wikipedia.org/wiki/Regular_expression for syntax)
  - E.g., `C[AG]CA{3,4}[CT]`

Representations (patterns)

- Position weight matrix

```
ATGGCATG
AGGGTGCG
ATCGCATG
TTGCCACG
ATGGTATT
ATTGCACG
AGGGCGTT
ATGACATG
ATGGCATG
ACTGGATG
```

	1	2	3	4	5	6	7	8
A	0.9	0.0	0.0	0.1	0.0	0.8	0.0	0.0
C	0.0	0.1	0.1	0.1	0.7	0.0	0.3	0.0
G	0.0	0.2	0.7	0.8	0.1	0.2	0.0	0.8
T	0.1	0.7	0.2	0.0	0.2	0.0	0.7	0.2

Example source: http://conferences.computer.org/bioinformatics/CSB2003/NOTES/Liu_Color.pdf

Representations (patterns)

- Position weight matrix

```
ATGGCATG
AGGGTGCG
ATCGCATG
TTGCCACG
ATGGTATT
ATTGCACG
AGGGCGTT
ATGACATG
ATGGCATG
ACTGGATG
```

	1	2	3	4	5	6	7	8
A	0.9	0.0	0.0	0.1	0.0	0.8	0.0	0.0
C	0.0	0.1	0.1	0.1	0.7	0.0	0.3	0.0
G	0.0	0.2	0.7	0.8	0.1	0.2	0.0	0.8
T	0.1	0.7	0.2	0.0	0.2	0.0	0.7	0.2

- Pseudo-counts: add a small number to each count, to alleviate problems due to small sample size

	1	2	3	4	5	6	7	8
A	10/14	1/14	1/14	2/14	1/14	9/14	1/14	1/14
C	1/14	2/14	2/14	2/14	8/14	1/14	4/14	1/14
G	1/14	3/14	8/14	9/14	2/14	3/14	1/14	9/14
T	2/14	8/14	3/14	1/14	3/14	1/14	8/14	3/14

Example source: http://conferences.computer.org/bioinformatics/CSB2003/NOTES/Liu_Color.pdf

Representations (patterns)

- Sequence logo
  - Nucleotide with the highest probability on top
  - Total height of the nucleotides at the *i*-th position,

$$h_i = 2 + \sum_{x \in \{A,C,G,T\}} p_{i,x} \log_2 p_{i,x} - \frac{4-1}{2n \ln 2}$$

    - $p_{i,x}$: probability of character *x* at position *i*
    - *n*: number of sequences
  - Height of nucleotide *x* = $p_{i,x} \cdot h_i$

	1	2	3	4	5	6	7	8
A	0.9	0.0	0.0	0.1	0.0	0.8	0.0	0.0
C	0.0	0.1	0.1	0.1	0.7	0.0	0.3	0.0
G	0.0	0.2	0.7	0.8	0.1	0.2	0.0	0.8
T	0.1	0.7	0.2	0.0	0.2	0.0	0.7	0.2

Identifying over-represented motifs (occurrence)

- The above representations are for known motifs: We know the exact DNA sequences of the TFBS
- In reality, how do we find out these sequences?
  - There are experiments that directly tell the rough binding locations of a protein
    - E.g., Chromatin immuno-precipitation followed by microarray (ChIP-chip) or sequencing (ChIP-seq/ChIP-exo)
  - If a TF is believed to regulate some genes by binding at their promoters, we can collect these promoter sequences
  - In both cases, the resolution is not high enough

Identifying over-represented motifs (occurrence)

- The corresponding motif discovery problem is as follows:
- Inputs: A set of sequences, each containing exactly one TFBS
  - There are other variations:
    - Each sequence contains one or more TBFS
    - Each sequence contains zero or one TFBS
    - Each sequence contains zero or more TFBS
- Goal: Find out the TFBS locations in the sequences
- Main idea: Identify common patterns in the sequences

Identifying over-represented motifs (occurrence)

- Different methods:
  1. Exhaustive search for all words of size up to *k*
     - Guaranteed to find best matches if the motif has a size ≤ *k*
     - The cost increases exponentially with respect to *k*
       - Indexing helps to a certain extent
     - More cost if inexact matches are allowed
  2. Multiple sequence alignment
     - Computationally hard
     - Many heuristics proposed
  3. Integrating auxiliary information for finding active binding site
     - Gene expression level (e.g., correlating number of binding sites with expression level)
     - Direct binding evidence (ChIP-chip or ChIP-seq)
     - Chromatin signals (e.g., DNA accessibility)

An illustrative example

- Suppose we know a protein binds some positions in the following sequences:
  - s_1= ACCGGCT
  - s_2= GTCAGCT
  - s_3= TCGGTAT
- 3-mer approach:
  - So, the binding site may be around CGG or GCT

3-mer	Number of sequences containing it
ACC	1
AGC	1
CAG	1
CCG	1
CGG	2
GCT	2
GGC	1
GGT	1
GTA	1
GTC	1
TAT	1
TCA	1
TCG	1

Statistical significance

- Now, consider the following situation:
  - A certain genome contains 80% `C`'s and `G`'s, and 20% `A`'s and `T`'s.
  - You have 100 sequences that contain the binding sites of a protein
    - 90 of them contain the pattern `GCGC`
    - 85 of them contain the pattern `ATAA`
  - Which one do you think is more likely to be the actual binding motif?

Statistical significance

- Statistically significant: something is unlikely to happen by chance
  - May suggest biological significance (why?)
- Steps to determine statistical significance:
  - Define a null hypothesis (background model)
  - Compute probability of occurrence given the null model
    - Direct computation
    - Simulation (more expensive, but usually more realistic)

Statistical significance

- Example: For a DNA sequence of length 4, assuming each base is independently and uniformly distributed, what is the chance of observing
  - 3 or more `A`'s on one strand?
    - Consider one strand: $(0.25)^4 + 4(0.75)(0.25)^3 = 0.0508$
    - Consider both strands: $0.0508 \times 2 = 0.1016$

Statistical significance

- What would be a good null model for a DNA sequence?
  - Independent, uniform?
    - Not quite true
  - Local dependence, uniform?
    - Better, but still missing global distribution
  - Local dependence, non-uniform?
    - Good, but more difficult to handle

Statistical significance

- What would be a good null model for a DNA sequence?
  - Independent, uniform?
    - Not quite true
  - Local dependence, uniform?
    - Better, but still missing global distribution
  - Local dependence, non-uniform?
    - Good, but more difficult to handle
- In general, good to get more realistic null distribution by
  - Sampling from permuted data
  - While preserving some key properties
    - For DNA, may want to preserve nucleotide frequencies, dinucleotide frequencies, etc.
  - Finally, get a distribution from the samples and see where the observed number is in the distribution

Part 3

Protein Domains

Protein domains

- Similar to DNA motifs, there are also enriched patterns on protein sequences that play special functions
  - With 20 amino acids (instead of 4 bases), even conserved patterns could contain a lot of variants
  - Need to consider similarity between amino acids
    - E.g., non-polar residues are in general more similar to each other than to polar or charged residues
- There are also patterns defined according to the structure
  - Sequence and structure are closely related

Protein domain: InterPro databases

- InterPro web site: “An integrated documentation resource for protein families, domains, regions and sites”
  - Gene3D: hidden Markov models
  - HAMAP: sequence matrices
  - PANTHER: hidden Markov models
  - Pfam: hidden Markov models

    Protein families http://pfam.xfam.org/
  - PIRSF: hidden Markov models
  - PRINTs: fingerprints (aligned position specific sequence matrices)
  - PROSITE: regular expressions, sequence matrices
  - PRODOM: sequence clusters
  - SMART: hidden Markov models
  - TIGRFAMs: hidden Markov models
  - SUPERFAMILY: hidden Markov models

Pfam

- For each family, the followings are provided:
  - An alignment of some representative *seed sequences* of the family
  - Profile HMM (a probabilistic model similar to PWM but which also considers relationships among positions)
    - Constructed from the seed sequences using HMMER3
    - Used to scan all protein sequences in UniProtKB to find occurrences
  - A *full alignment* that contains all sequences above a scoring threshold
- Additional information:
  - Domain architecture
  - Phylogenetic tree of sequences
  - Structural information, where available

Pfam

- For each family, the followings are provided:
  - An alignment of some representative *seed sequences* of the family
  - Profile HMM (a probabilistic model similar to PWM but which also considers relationships among positions)
    - Constructed from the seed sequences using HMMER3
    - Used to scan all protein sequences in UniProtKB to find occurrences
  - A *full alignment* that contains all sequences above a scoring threshold
- Additional information:
  - Domain architecture
  - Phylogenetic tree of sequences
  - Structural information, where available

Example: PF01036 Bacterial rhodopsins

- HMM logo (pattern)

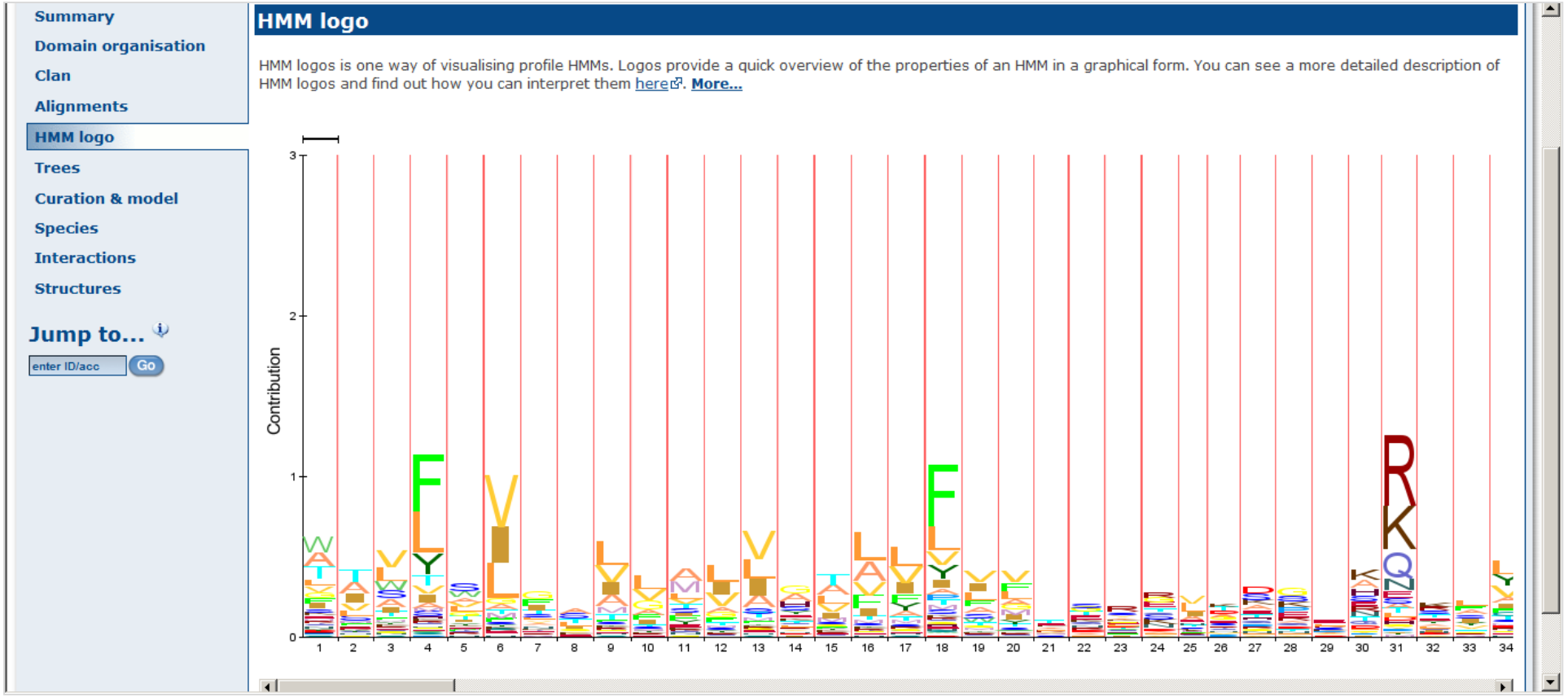


Image credit: Pfam

Example: PF01036 Bacterial rhodopsins

- Multiple sequence alignment (conservation)

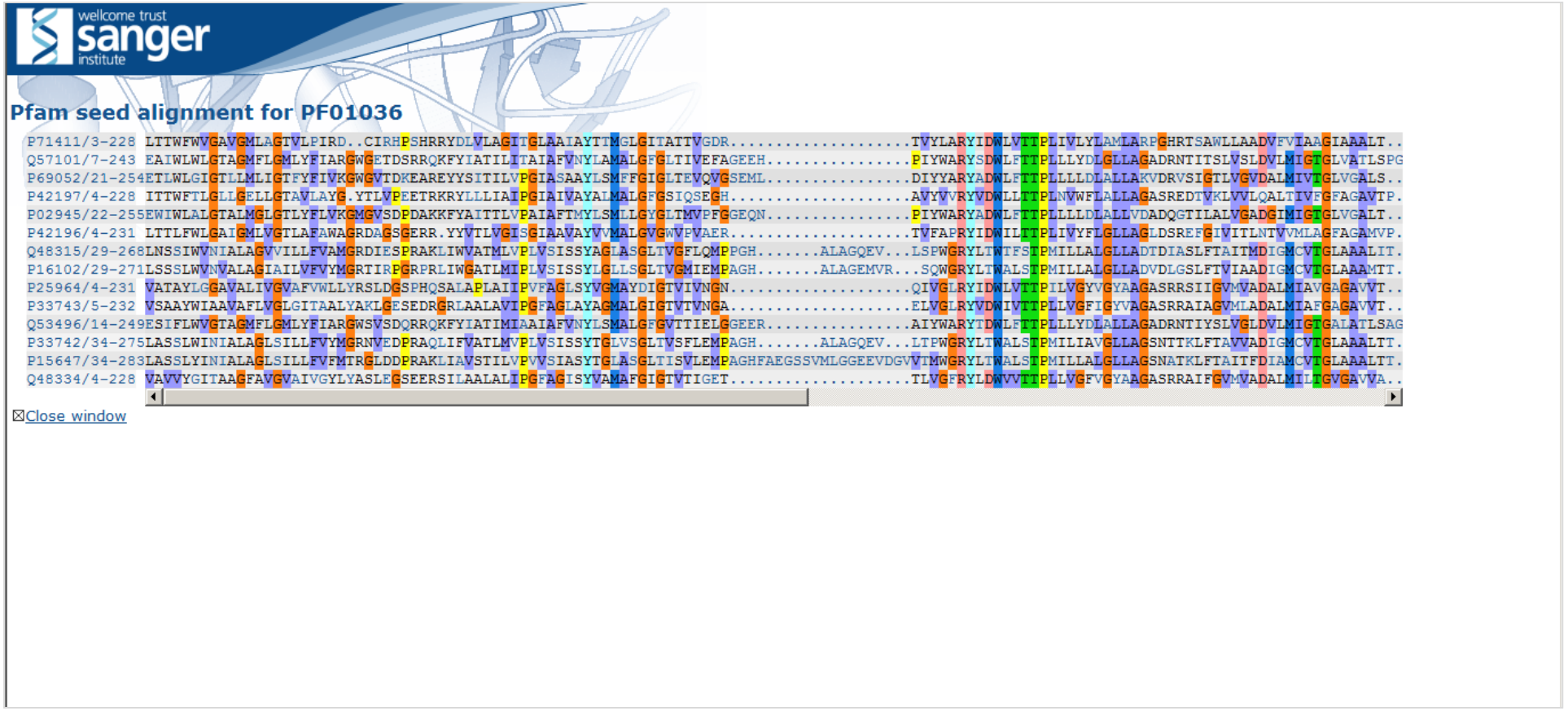


Image credit: Pfam

Example: PF01036 Bacterial rhodopsins

- Phylogenetic tree (conservation)

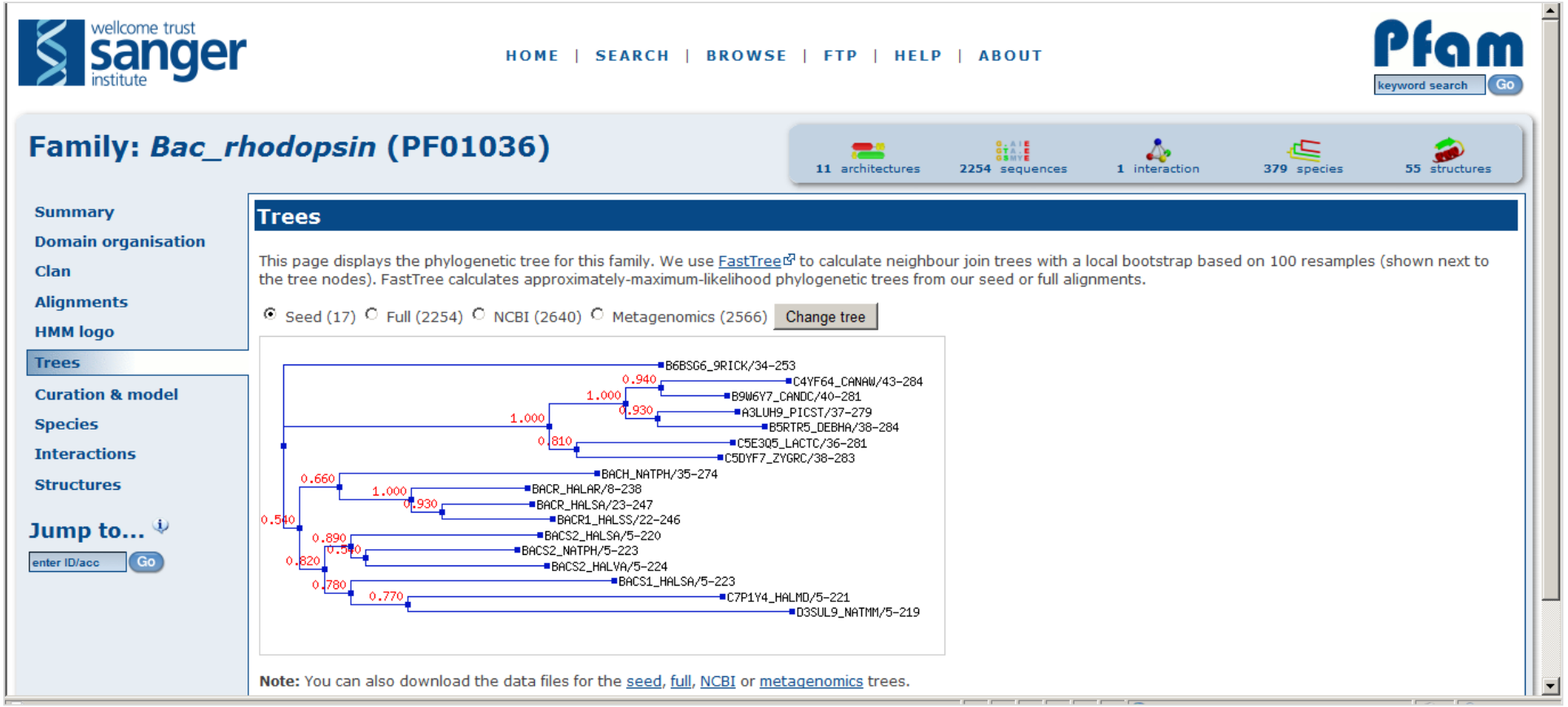


Image credit: Pfam

Example: PF01036 Bacterial rhodopsins

- Domain organization – one protein can have multiple domains

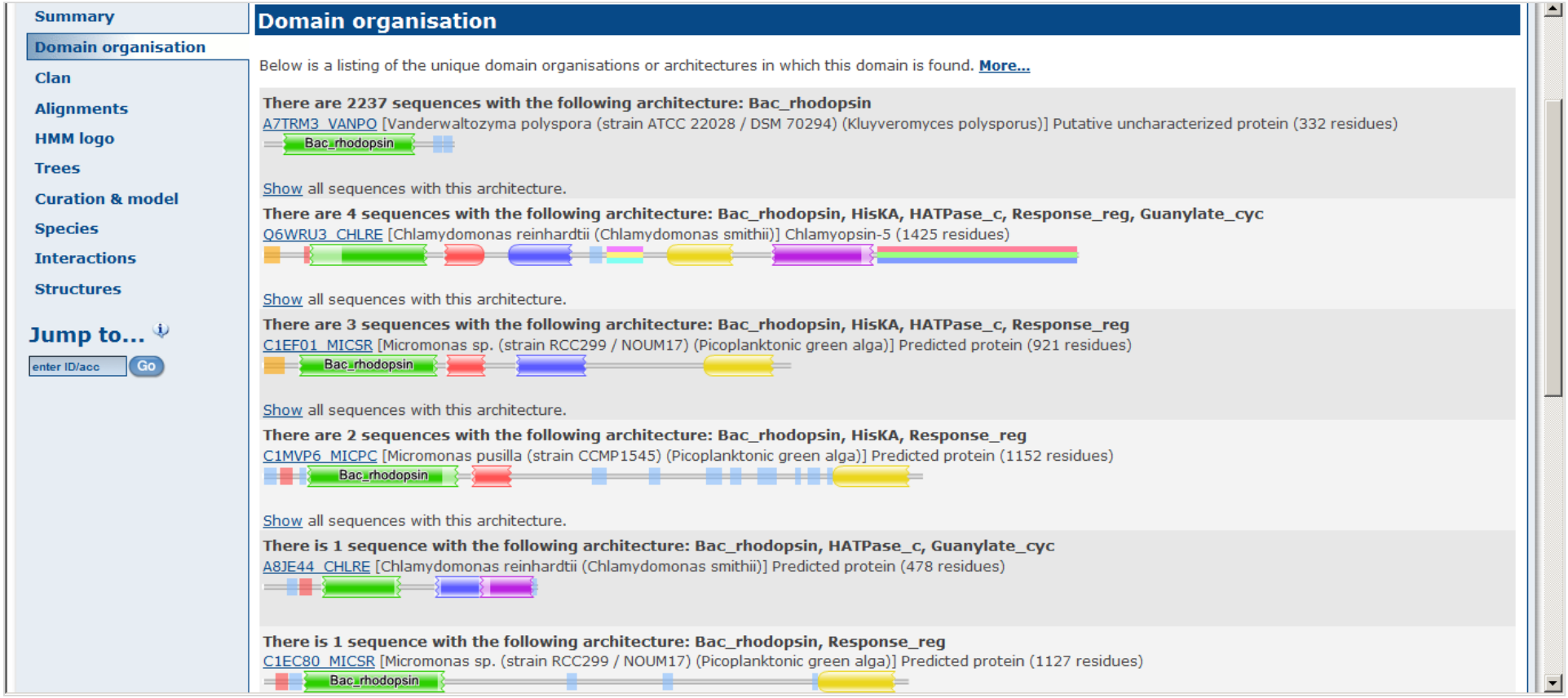


Image credit: Pfam

Example: PF01036 Bacterial rhodopsins

- Crystal structure (en route to function)

Image credit: Pfam

Pfam

- Four types of entries:
  - Family: collection of related proteins
  - Domain: structural unit found in multiple protein contexts
  - Repeat: stable only when multiple copies are present
    - Not to be confused with DNA repeats
  - Motif: short unit outside globular domains
- Entries are further grouped into clans based on similarity in either
  - Sequence
  - Structure
  - Profile HMM
- Two components
  - Pfam A: high quality, manually curated
  - Pfam B: lower quality, automatically curated
- Current: Pfam 29.0
  - 16,295 families
  - Over 73.5% of proteins in SWISSPROT and TrEMBL have at least one match to a Pfam-A family

Epilogue

Summary and Further Readings

Summary

- Motifs and domains
  - Patterns – Representations:
    - Exact
    - Probabilistic
  - Unusual number of occurrences
    - Statistical significance
  - Function
  - Conservation

Further readings

- Chapter 10 of *Algorithms in Bioinformatics: A Practical Introduction*
  - More biological background
  - Computational methods for finding motifs
  - Free slides available

Further readings

- D'haeseleer, How does DNA Sequence Motif Discovery Work? *Nature Biotechnology* 24(8):959-961, (2006)