LSM2241
Sequence Patterns and Profiles

Greg Tucker-Kellogg
dbsgtk@nus.edu.sg

9 September 2015

Outline

Where we left off, and where we are going

Motifs and consensus

Patterns and profiles

Detecting motifs using patterns and profiles

Roundup and next time

Topic

Where we left off, and where we are going

Motifs and consensus

Patterns and profiles

Detecting motifs using patterns and profiles

Roundup and next time

What is sequence alignment good for?

- Allows us to ask fundamental questions about relations between sequences
- Describes relationships between sequences in terms of *similarity*
- Provides the tools to assess *homology*
- **Sometimes**, (e.g., via PAM matrices) describes relationships using an explicit theoretical model of a biological process (divergence of sequences through evolution)

Searching sequence databases using BLAST

- The **BLAST** family of programs allows us to search large sequence databases for sequences similar to a query sequence
- **BLAST** uses a heuristic strategy to search sequences databases for High Scoring Pairs
- **BLAST** searches do not provide guaranteed optimal alignments

Example: Comparative genomics of *Acitenobacter* genomes

OPEN ACCESS Freely available online

PLOS | ONE

Evolution of a Pathogen: A Comparative Genomics Analysis Identifies a Genetic Pathway to Pathogenesis in *Acinetobacter*

Jason W. Sahl[1]*, John D. Gillece[1], James M. Schupp[1], Victor G. Waddell[2], Elizabeth M. Driebe[1], David M. Engelthaler[1], Paul Keim[1,3]

1 Department of Pathogen Genomics, Translational Genomics Research Institute, Flagstaff, Arizona, United States of America, **2** Arizona Department of Health Services, Bureau of State Laboratory Services, Phoenix, Arizona, United States of America, **3** Center for Microbial Genetics and Genomics, Northern Arizona University, Flagstaff, Arizona, United States of America

Abstract

Acinetobacter baumannii is an emergent and global nosocomial pathogen. In addition to *A. baumannii*, other *Acinetobacter* species, especially those in the *Acinetobacter calcoaceticus-baumannii* (*Acb*) complex, have also been associated with serious human infection. Although mechanisms of attachment, persistence on abiotic surfaces, and pathogenesis in *A. baumannii* have been identified, the genetic mechanisms that explain the emergence of *A. baumannii* as the most widespread and virulent *Acinetobacter* species are not fully understood. Recent whole genome sequencing has provided insight into the phylogenetic structure of the genus *Acinetobacter*. However, a global comparison of genomic features between *Acinetobacter* spp. has not been described in the literature. In this study, 136 *Acinetobacter* genomes, including 67 sequenced in this study, were compared to identify the acquisition and loss of genes in the expansion of the *Acinetobacter* genus. A whole genome phylogeny confirmed that *A. baumannii* is a monophyletic clade and that the larger *Acb* complex is also a well-supported monophyletic group. The whole genome phylogeny provided the framework for a global genomic comparison based on a blast score ratio (BSR) analysis. The BSR analysis demonstrated that specific genes have been both lost and acquired in the evolution of *A. baumannii*. In addition, several genes associated with *A. baumannii* pathogenesis

Sahl et al. 2013

The limits of substitution matrices

- Substitution matrices (**BLOSUM**, etc.) represents each change independent of position
- For example, a `Ser` ⟶ `Ala` substitution is given the same penalty no matter where it occurs in the sequence
- We know this is not always a good idea!
  - sometimes a `Ser` ⟶ `Ala` substitution may destroy the function of a protein
  - In that case, should it have the same penalty?

Where we are heading

- We want to use a group of sequences to characterize a sequence family
  - We *start* with a multiple sequence alignment
  - We *end* with a profile or pattern representation
- We want to be able to use this representation
  - *As a query* to search a sequence database to ask “What sequences in the database are members of this family?”
  - *As a database* to ask “in what family or families does my sequence belong?”

Some key terms

Sequence Motif A biological sequence pattern that is widespread and has, or is assumed to have, biological significance

Pattern A qualitative description of a sequence motif, usually in the form of a regular expression

Profile A quantitative description of a pattern or motif using a position dependent scoring system

Topic

Where we left off, and where we are going

Motifs and consensus

Patterns and profiles

Detecting motifs using patterns and profiles

Roundup and next time

Sequence Motifs occur in a wide range of roles

Functional motifs

- enzyme active sites
- functional domains

Regulatory motifs

- post-translational modification sites
  - phosphorylation sites
  - glycosylation sites
  - myristylation sites
- signal sequences guiding cellular localization

DNA motifs

- restriction enzyme recognition sites
- promoter sites
- transcription factor binding sites

RNA motifs

- RNA splicing sites
- miRNA recognition sites

Two simple DNA sequence motifs

EcoRI DNA binding sequence `GAATTC`

- Exactly six nucleotides (note that the double stranded sequence is palindromic)
- Average frequency of occurrence is $4^6 = 4096$ bp in random DNA

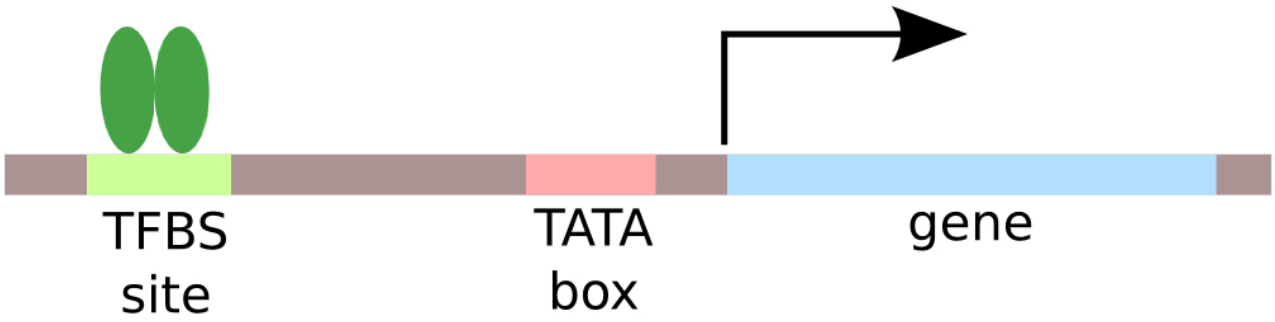
HindII DNA binding sequence `GTYRAC`

- Six nucleotides, but two are degenerate (`Y` means `C` or `T`, `R` means `A` or `G`)
- Average frequency of occurrence is $2^2 \times 4^4 = 1024$ bp in random DNA

Adapted from D'Haeseleer 2006 Patrik D'haeseleer. "What are DNA sequence motifs?" *Nature Biotechnology*, **24** (4):423–5, April 2006

DNA regulatory motifs I

We often consider promoters and other elements in DNA sequences that serve to regulate gene expression



The transcription factor binding site sequence influences binding affinity and regulatory strength. The cartoon above fits the **GAL4** transcriptional activator and many others.

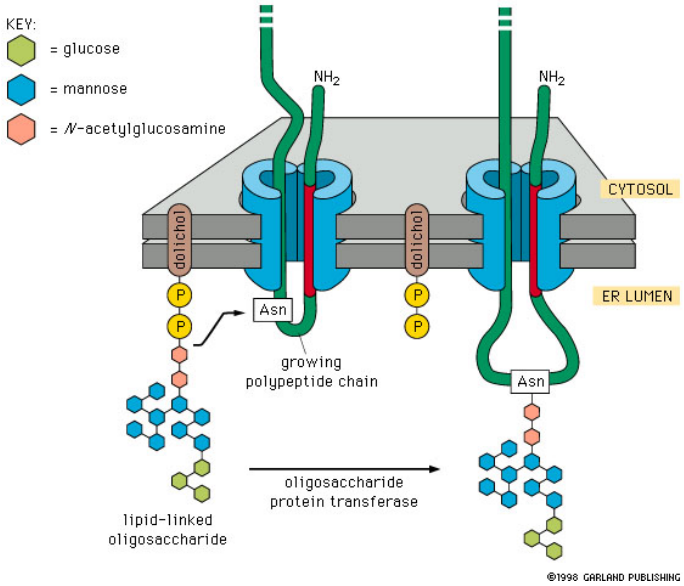
DNA regulatory motifs II

factor	Structural type	recognition sequence	Binds as
SP1	Zinc finger	5'-GGGCGG-3'	monomer
AP-1	Basic zipper	5'-TGA(G/C)TCA-3'	dimer
C/EBP	Basic zipper	5-ATTGCGCAAT-3'	dimer
HS Factor	Basic zipper	5'-XGAAX-3'	trimer
ATF/CREB	Basic zipper	5'-TGACGTCA-3'	dimer
c-Myc	Basic-helix-loop helix	5'-CACGTG-3'	dimer
Oct-1	Helix-turn-helix	5'-ATGCAAAT-3'	monomer
NF-1	Novel	5'-TTGGCXXXXXGCCAA-3'	dimer

Examples of transcription factor binding motifs, binding transcription factor proteins (both activators and repressors)

Perhaps not surprisingly, these short binding sequences are very hard to discriminate correctly in real DNA. There are many false positives.

Example motif: protein N linked glycosylation



The rules

- The modification occurs at an `Asn` residue (N)
- The next amino acid can be anything except `Pro`
- Either a `Ser` or a `Thr` must occur after that
- The next amino acid can be anything except a `Pro`

The PROSITE pattern

PROSITE pattern: `N-{P}-[ST]-{P}`

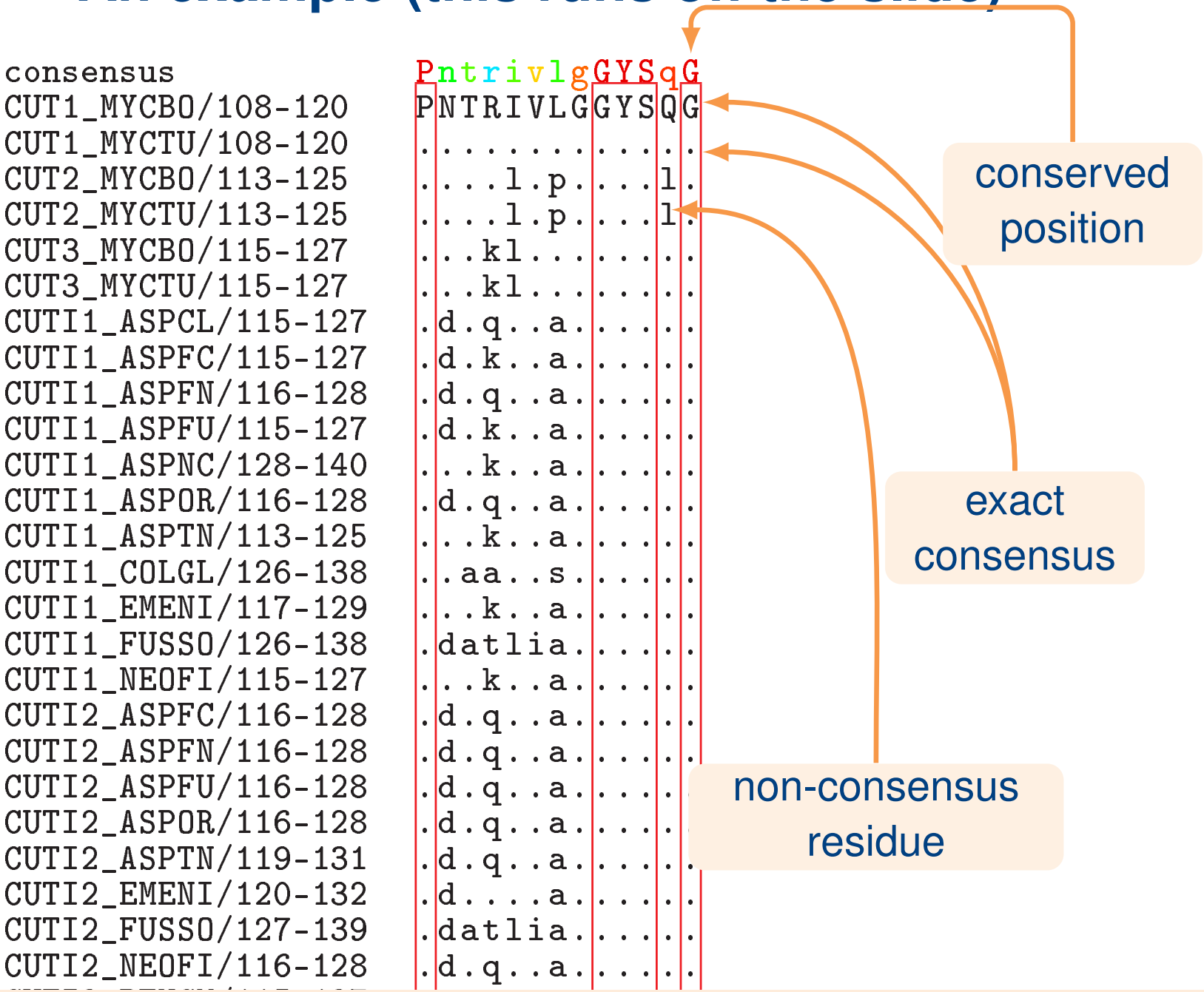
How do we work with sequence motifs?

- How should we describe them?
- How can we know whether or not a given sequence fits a given motif?
- How can we discover new motifs?
- How can we detect our new motif in sequence databases?
- How can we test a new sequence against known motifs?

Consensus: the simplest description

- Line up sequences in an MSA
- Each position is decided by vote
- Plurality wins

An example (this runs off the slide)



Limitations of consensus

- What if the consensus is weak?
- The vote is winner-take-all, so consensus ignores minority members of a sequence alignment, even large minorities
- What if the consensus is nonexistent?
  - Suppose any amino acid can be at a particular position? This cannot be represented by ordinary consensus
    - A degenerate consensus sequence can handle this in many cases, such as the HindII example above.
- How can a consensus handle variable-length gaps?

Consensus sequences are most useful for highly conserved sequence patterns

Topic

Where we left off, and where we are going

Motifs and consensus

Patterns and profiles

Detecting motifs using patterns and profiles

Roundup and next time

Some databases of patterns and profiles

- Protein databases
  - PROSITE database of protein domains, families and functional sites
  - CDD, Conserved Domain Database at NCBI
  - HMMER, profile Hidden Markov Models of biological sequences
- Nucleic acid databses
  - JASPAR database of transcription factor profiles
  - TRANSFAC (public) database of transcription factors. The commercial version is larger

PROSITE

- PROSITE[1] is an online resource describing protein *domains*, *families* and *functional sites* as well as **patterns** and **profiles**
- PROSITE **patterns** are regular expressions to describe a set of sequences.
  - A “regular expression” is a formalism for finding matches to a string of text.
  - The allowed expressions tell us what strings can be matched

Pattern notation for PROSITE

Notation	meaning
`A`	The amino acid `A`
`[ABC]`	any one of `A` or `B` or `C`
`X`	any amino acid at all
`{AB}`	any amino acid **except** `A` or `B`
`A(2)`	`AA` (an `A` repeated exactly 2 times)
`x(2,5)`	`xx` or `xxx` or `xxxx` or `xxxx`

Additional comments

- The one-letter abbreviation for amino acids is used in all PROSITE patterns
- Adjacent amino acids in PROSITE notation are separated by a hyphen ('-'). This is how gaps are represented in an MSA. How would you represent gaps in an MSA using PROSITE patterns?

Sequence Profile

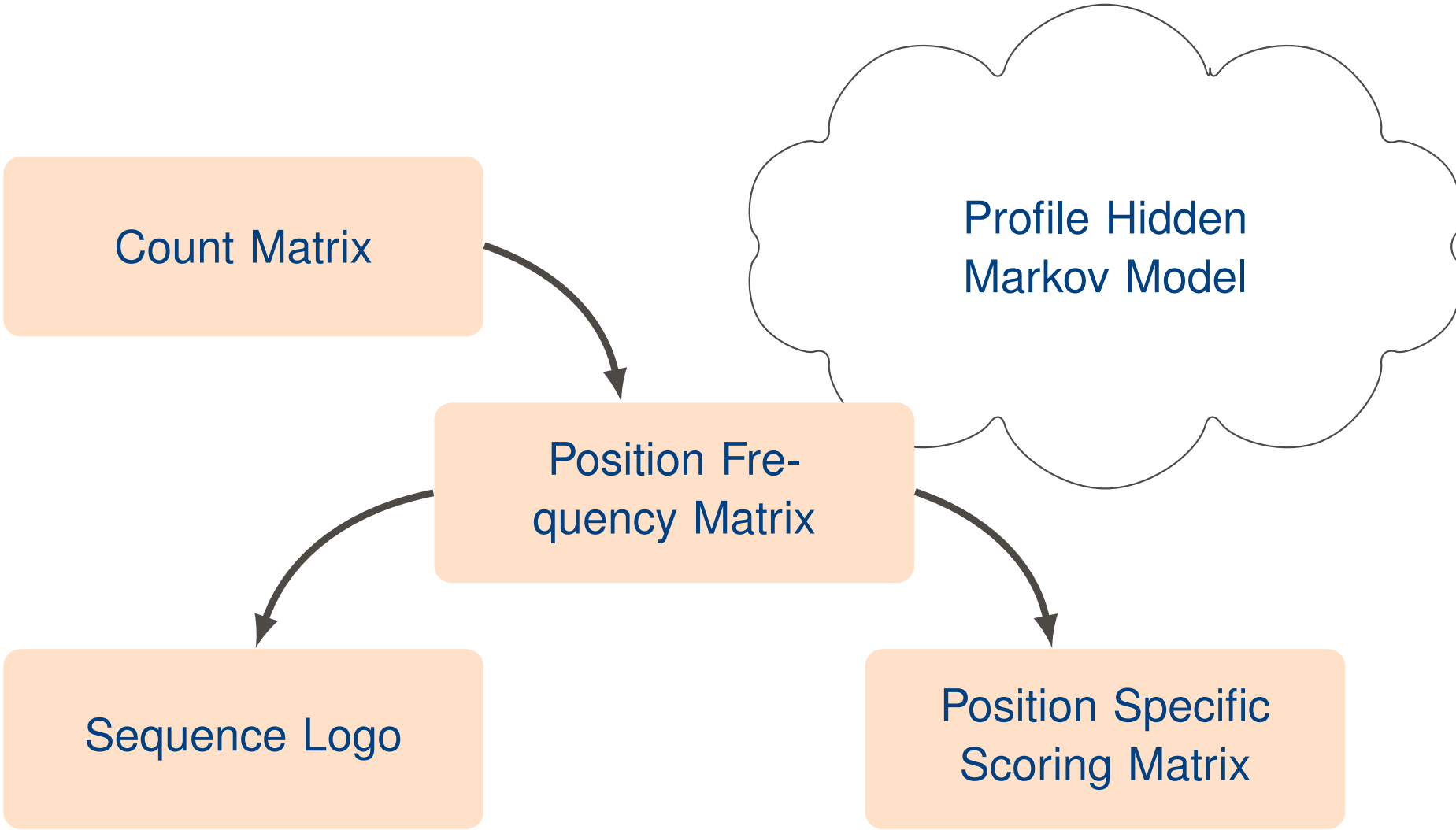
Definition

A quantitative description of a motif that includes information on frequencies or probabilities at each position

There are a number of approaches to profiles

- Position frequency matrix (PFM)
- Position specific scoring matrix (PSSM)
- Graphical sequence logos
- Profile Hidden Markov Model (HMM)

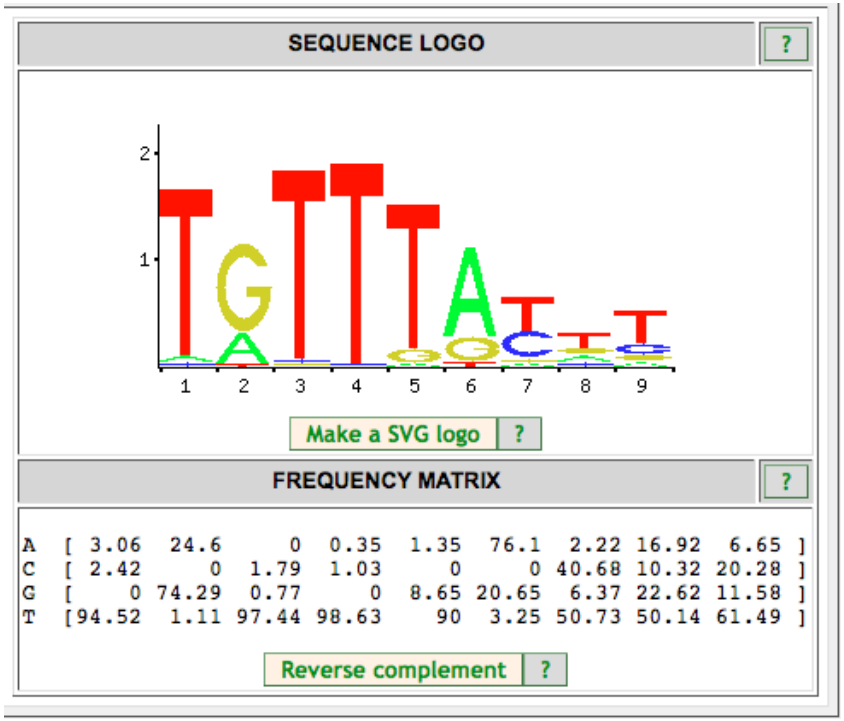
Deriving profiles from aligned sequences (see Excel workbook)



Position Frequency Matrix (PFM)

- A **Count Matrix** simply counts the occurences of each amino acid or nucleotide at each position
- A **Position Frequency Matrix** converts the counts to frequency of amino acid or nucleotide at each position.
- For nucleotide i at position j, we assign the matrix value of p_{ij} to be the corresponding frequency of occurrence

A PFM and “Sequence Logo” from JASPAR



JASPAR entry for the Forkhead class

- The graphical representation is called a **sequence logo**
- Hey, I thought these were frequencies? Why does the *y* axis go to 2?

The information in profiles can be represented in *bits*

For four possible nucleotides, the total possible information is **two bits**, because $\log_2(4) = 2$.

The actual information at any position i is given by

$$R_i = 2 - (H_i + e_n)$$

where H_i is calculated from the frequency matrix values $f_{a,i}$ as the uncertainty (or Shannon entropy) at position i

$$H_i = -\sum f_{a,i} \times \log_2 f_{a,i}$$

The other term e_n is a correction for small samples, which we will use in our spreadsheet

Position Specific Scoring Matrices

In the real world of protein sequences

- we might not want to use a strict consensus
- we might not want to use a yes/no criterion like patterns
- we might not have enough observations to describe all possibilities well by a PFM

What do we do?

- Rather than using frequencies, we would like to use *likelihoods* of occurrence at each position.

PSSM II, transforming to log odds

The simplest alteration, common for DNA motifs, is

$$M_{kj} = \log_2 \left(\frac{p_{kj}}{p_j} \right)$$

where p_j is the probability of any particular nucleotide or amino acid i

To account for smaller numbers of aligned sequences, p_{kj} is usually replaced by p'_{kj}

$$p'_{kj} = \frac{C_{kj} + p_j}{Z + 1}$$

Where C_{kj} is corresponding count matrix value.

The result is a Position Specific Scoring Matrix (PSSM) which serves as a scoring model of a sequence for a motif

$$M_{kj} = \log_2 \left(\frac{\frac{C_{kj}+p_j}{Z+1}}{p_j} \right)$$

Different motif representations

a

```
HEM13   CCCATTGTTCTC
HEM13   TTTCTGGTTCTC
HEM13   TCAATTGTTTAG
ANB1    CTCATTGTTGTC
ANB1    TCCATTGTTCTC
ANB1    CCTATTGTTCTC
ANB1    TCCATTGTTCGT
ROX1    CCAATTGTTTTG
```

b

```
YCHATTGTTCTC
```

c

```
A 002700000010
C 464100000505
G 000001800112
T 422087088261
```

d

e

f

Different representations of a motif: **a** MSA, **b** consensus, **c** count matrix, **d** count matrix as a logo **e** SeqLogo in bits, **f** SeqLogo with correction for GC content

Patrik D'Haeseleer. "What are DNA sequence motifs?" *Nature Biotechnology*, **24**:423–5, April 2006

Topic

Where we left off, and where we are going

Motifs and consensus

Patterns and profiles

Detecting motifs using patterns and profiles

Roundup and next time

Using Profile and Pattern databases

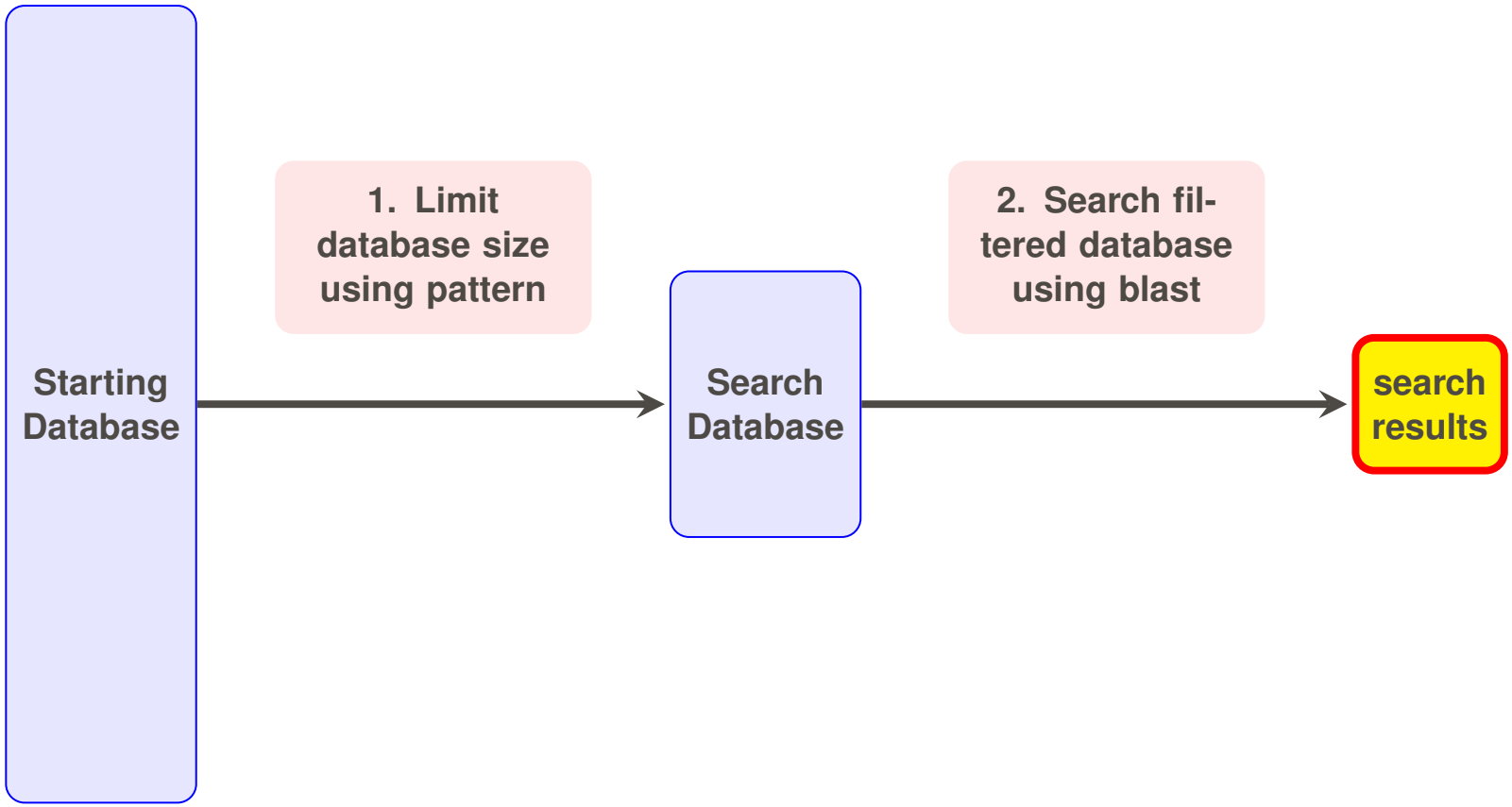
- Databases such as Prosite and JASPAR catalog motifs using patterns and profiles.
- Many other databases and tools are available. (See the MEME suite, for example.) Some, such as TRANSFAC, are commercial products with limited academic offerings
- You can construct a new pattern or profile for search or testing, often starting with a multiple sequence alignment of sequences of interest.

Pattern Hit Initiated BLAST: Φ-BLAST

- Suppose we have a query sequence which is a member of a protein family
- We are interested to find distant related family members
- If we have a pattern for the family, we can restrict the size of the search database using **PHI-BLAST**

INPUT	a query sequence and a pattern (usually a loose pattern)
OUTPUT	sequences matching the pattern, in decreasing order of **BLAST** *E* value
METHOD	Build a database of sequences matching the pattern, search within them

Graphical view of PHI-BLAST

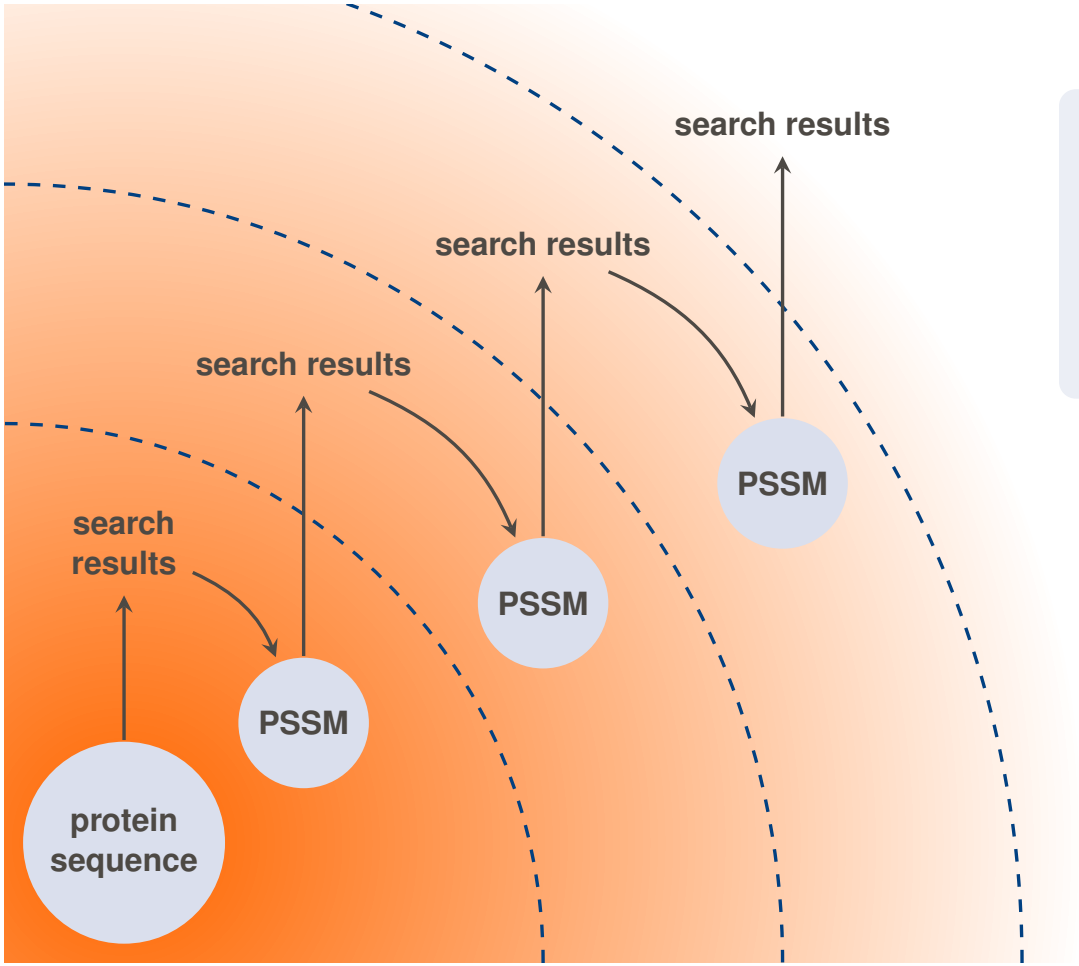


Using profiles with BLAST: Ψ-BLAST

Suppose we don't have a pattern. Position Specific Iterated BLAST (**PSI-BLAST**) handles this by the following approach

1. perform initial search using `blastp`
2. Automatically construct MSA from top hits
3. Create profile (PSSM) from MSA
4. Query the database using the profile
5. Go back to step 2
6. After a few iterations, stop.

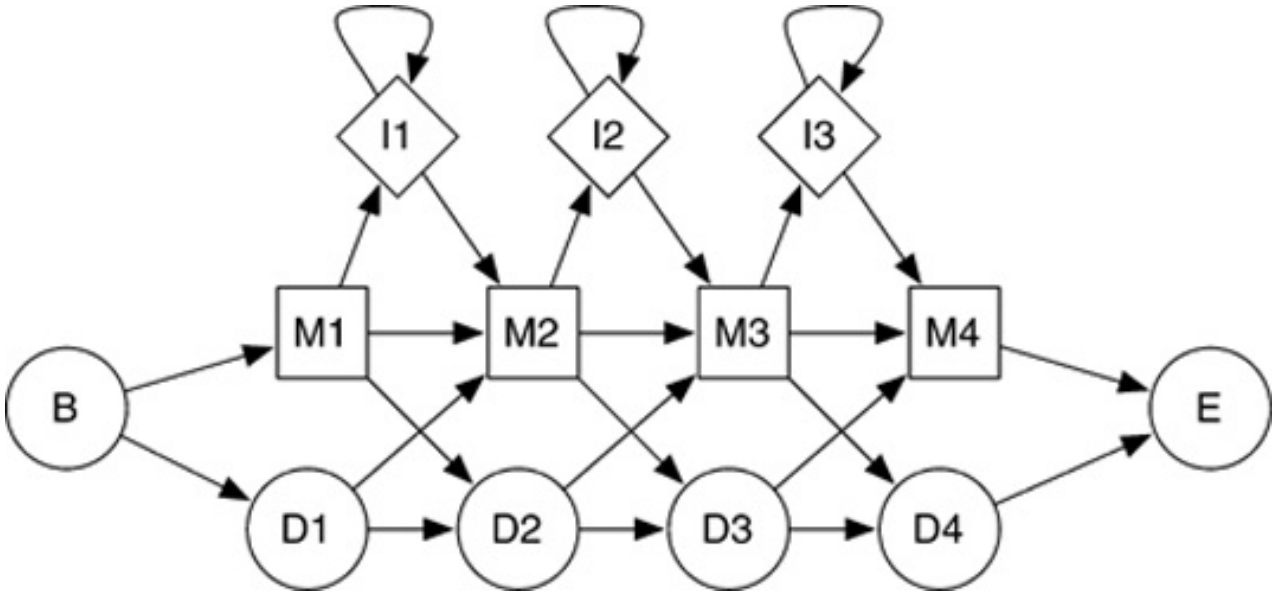
Graphical view of PSI-BLAST



The first round of PSI-BLAST is an ordinary `blastp`. Later rounds search deeper into distant relationships using Position Specific Scoring Matrices constructed from the results of earlier rounds.

Advanced: Profile Hidden Markov Models

Hidden Markov Models can represent profiles by a model that assigns probabilities to sequences.



A profile HMM representation of a sequence of length 4

Each `M` represents a match, each `I` represents an insertion, each `D` represents a deletion. `B` and `E` are the beginning and end of the sequence.

Why Profile HMMs?

- A full probabilistic representation of profiles
- Position-specific information is captured explicitly
- Position-specific insertion and deletion penalties are captured in the model
- PFAM is a database of Profile HMMs
- The tool **HMMER** can be used to generate and search HMMs

Topic

Where we left off, and where we are going

Motifs and consensus

Patterns and profiles

Detecting motifs using patterns and profiles

Roundup and next time

What we've learned

- Interesting biology is encoded in sequence motifs
- Position matters within sequence motifs
- Because position matters, pairwise alignment has a hard time detecting motifs
- Patterns can be used to define motifs qualitatively
- Profiles can be used to score motifs quantitatively
- Patterns and profiles can be used to search sequence databases
- Sequences can be used to search pattern and profile databases

Bibliography


D'Haeseleer, Patrik (2006). "What are DNA sequence motifs?" In: *Nature Biotechnology* 24.4, pp. 423–5. DOI: 10.1038/nbt0406-423 (cit. on p. 12).


Sahl, Jason W et al. (2013). "Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in acinetobacter." In: *PloS ONE* 8.1, e54287. DOI: 10.1371/journal.pone.0054287 (cit. on p. 6).