

Applications of Local MSA

Conserved patterns in biological sequences

Example: Transcription factor binding sites

```

SP ...gcttt AATTTTCACTATATACTATAA cgatt...
ST ...cagat ATAAATGATATAGTGGTTATA gttaa...
ST ...atctt TTTTATTATTAAATCGTATTA gcagc...
EC ...aggct ATAAATGATATAGTGGTTATA gttag...
EC ...acctt TTTTATTATTAAATCGTATTA gtcac...
VC ...ttata ACTAATAATTATAAAATATGT gtgtc...
YP ...gctga TGAAATGATATAATCGTTATA taaga...

```

...agcgagcctgagcactcgaggcatctctgcacattcagc**atgggatgggctctctgctctgatgcgctgatga...**

Sequence	Transcription Factor
5A2CCa2ACA2	Rap1
oGTGGCAAA2	Rpo4
aa22GA2TCA	Gcn4
CT2GAA2TTC	HSE
222T2C2CC2C2	Mgl1/STRE
222CCAAT22A2	Rap2,3,4
CACGTG22222	Chf1
22ACGCGT222	MCB
2TTC222Gaa22	Lys14
oCC22T222G2	Leu3

Some known binding site motifs

Applications of Local MSA

Conserved patterns in biological sequences

Example: Protein domains

- Fold independently
- Carry out specific functions
- Found in diverse contexts
- Conserved in evolution

Insulin receptor

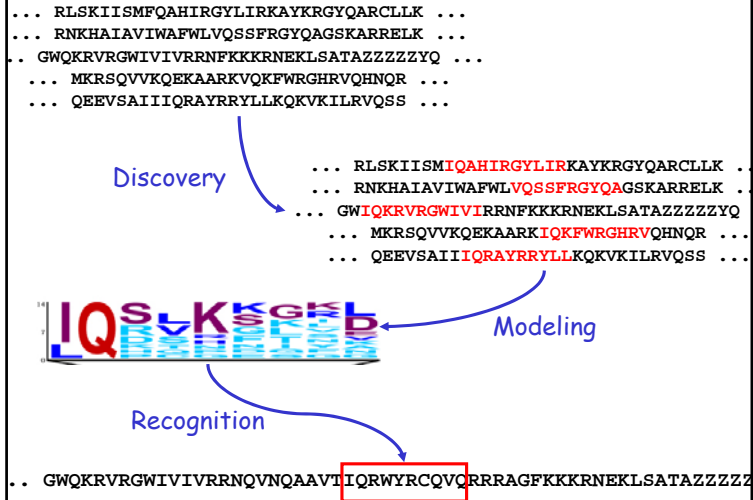
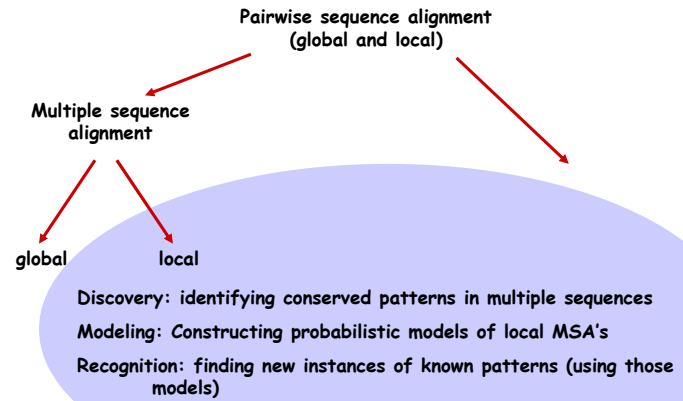

```

1  mphnsirsgh  gglnglqgaf  vngrplpevv  rqrivdlahq  gvrpcdisrq  lrvshgcvsk
61  ilgryyetgs  irpgviggsk  pkvatpkvve  kigdykrqnp  tmfaweirdr  llaegvcdnd
121  tvpevssinr  iirtkvqgpf  nlpmdscvat  kslspgthli  pssavtppes  pqdsdlssty
181  singllgiaq  pgndnrkmd  dsdqdsclrs  idsqssssgp  rkhlrtdtfs  qhhlealecp
241  ferqhypeay  aspshkgeq  glyplplns  alddgkatlt  sstntplgrn  sthqtypvva
301  dphspfaikq  etpelsssss  tpsslssaf  ldlqqvgsq  pagasvppfn  afphaasvyg
361  qftgqallsq  remvgptlpg  ypphiptsgq  gsyassaiaq  mvagseysgn  ayshtpyssy
421  seawrfpns  llspyyys  tsrpsappts  atafdh1

```

paired box gene 8 [Mus musculus]
gi|6754990|ref|NP_035170.1|[6754990]

[CDART: Conserved Domain Architecture Retrieval Tool](#)



Local Multiple Sequence Alignment Probabilistic Framework

- Discovery
 - Given multiple sequences, often unaligned, find a conserved pattern (e.g., the Pax domain)
- Representation
 - Given a local MSA for the Pax domain, construct probabilistic model
- Recognition (using model)
 - Given a new sequence, does it contain the Pax domain?
 - Find all sequences with Pax domains in the data base.

Local MSA Methods

- Discovery:
 - Hidden Markov Models (HMMs)
 - Gibb's sampler
 - PSI BLAST
- Modeling:
 - Position Specific Scoring Matrices (PSSMs)
 - HMMs
- Recognition:
 - Depends on model

Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates...

*Assume pattern has already
been discovered.*

Input: local MSA, $k \times n$ matrix
 $A[i,j]$: j th symbol in i th sequence

Output: Scoring matrix, $|\Sigma| \times n$
 $S[i,j]$: score of symbol i at position j

Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates...

Example:

WEIRD
WEIRD
WEIRE
WEIQH

[See spreadsheets...](#)

Position Specific Scoring Matrices

PSSM's, profiles, weight matrices, templates...

Given $A[l,k]$ (k sequences, l positions),

– Frequency of aa i at position j

$$F[i, j] = \frac{n_i}{k}$$

– Propensity of aa i at position j

$$P[i, j] = \frac{F[i, j]}{b_i}$$

– Log odds scoring matrix

$$S[i, j] = \log_2 P[i, j]$$

Amino acid background frequencies

D	0.052
E	0.062
H	0.023
I	0.053
Q	0.041
R	0.051
W	0.014

Pseudocounts

Example:

```

AAAAA
CCCCC
DDDDD
. . .
YYYYY
WEIRD
WEIRD
WEIRE
WEIQH
    
```

$$F[i, j] = \frac{n_i + b}{k + |\Sigma| b}$$

The pseudocount, **b**, avoids the problem of zero entries in the frequency matrix (and negative infinity in the log odds scoring matrix.)

Frequently, **b = 1**, is chosen.

Also, see Durbin, 5.6

