

## 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 ...aggot ATAAATGATATAGTGGTTATA gtag...
EC ...acctt TTTTATTATTAAATCGTATTA gtcac...
VC ...ttata ACTAATAATTATAAAATATGT gtgtc...
YP ...gctga TGAAATGATATAATCGTTATA taaga...
  
```

...agegagcctgagcactcgaggcatctctgcacattcagc**atgggatgggcctcctctcctgtatgcgctgatga**...

5A2CCAACA	Rap1
5GTGGCAAA	Rpp1
5AAAGATCA	Gcn4
CTAGAA-TTC	HSE
55-TCC-C	Mig1/STRE
55CCAAT-A	Hap2,3,4
CACGTGA	Chf1
T-ACGGT	MCB
TTC-GAA	Lys14
CCGAT-GG	Leu3

Some known binding site motifs

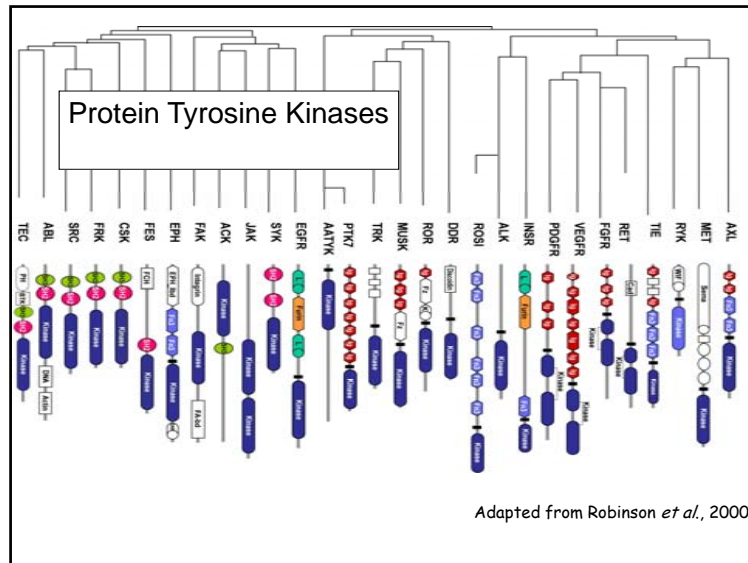
## Applications of Local MSA

### Conserved patterns in biological sequences

Example: Protein domains

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

Insulin receptor



## Protein domain databases

### Conserved Domain Database (CDD)

Representation: Position specific scoring matrices (PSSMs)

Structurally corrected local MSAs

CDART: Conserved Domain Architecture Retrieval Tool

### PFAM, SMART

Representation: Hidden Markov Models (HMM's)

Curated local MSA's

### More:

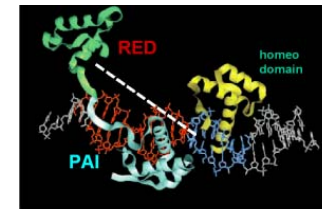
see Mount, Table 9.5

## Multi-domain protein example: PAX gene family

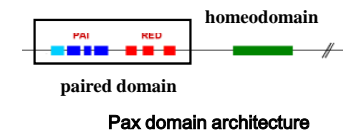
- Developmental regulatory genes that encode transcription factors
- Contain a DNA binding domain
- Early expressed during embryogenesis
- Role in morphological boundaries and early regionalization



<http://www.gene-regulation.com/info/pax.html>



Pax structure



<http://www.gene-regulation.com/info/pax.html>

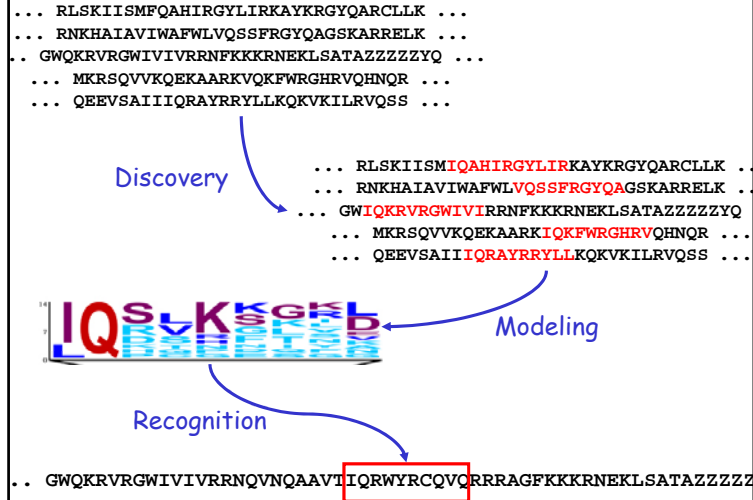
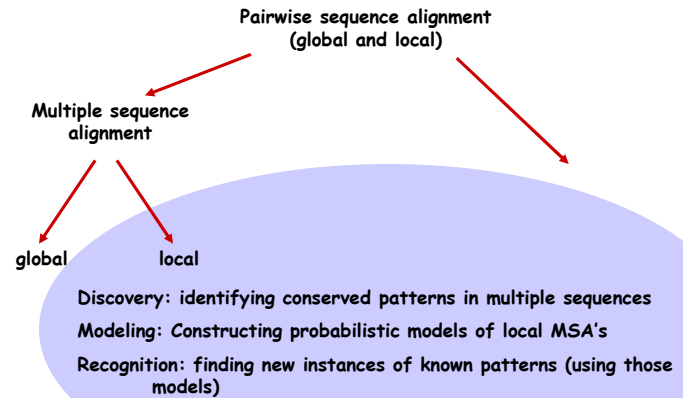
```

1  mphnsirsgh  gglnglqgaf  vngrplpevv  rqrivdlahq  gvrpcdisrq  lrvshgcvsk
61  ilgryyetgs  irpgviggsk  pkvatpkvve  kigykrqnp  tmfaweirdr  llaegvcdnd
121  tvpsvssinr  iirtkvqgpf  nlpmdscvat  kslspgthli  pssavtppes  pqdsdlssty
181  singllgiaq  pgndnrkrmd  dsdqdsclrs  idsqssssgp  rkhlrtdfst  qhhlealecp
241  ferqhypeay  aspshkgeq  glyplpllns  alddgkatlt  sstntplgrnl  sthqtypvva
301  dphspfaikq  etpelsssss  tpsslssaf  ldlqqvgsqg  pagasvppfn  afphaasvyg
361  qftggallsq  remvgptlpg  ypphiptsgq  gsyassaiaq  mvagseysgn  ayshtpyssy
421  seawrfpms  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

