# Bioinformatics

## Multiple sequence alignment

# Bioinformatics - lectures

- Introduction
- Information networks
- Protein information resources
- Genome information resources
- DNA sequence analysis
- Pairwise sequence alignment
- Multiple sequence alignment
- Secondary database searching
- Analysis packages
- Protein structure modelling

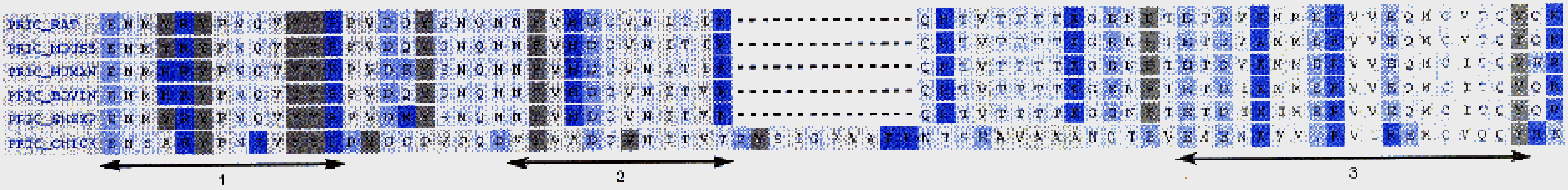
# Multiple sequence alignment

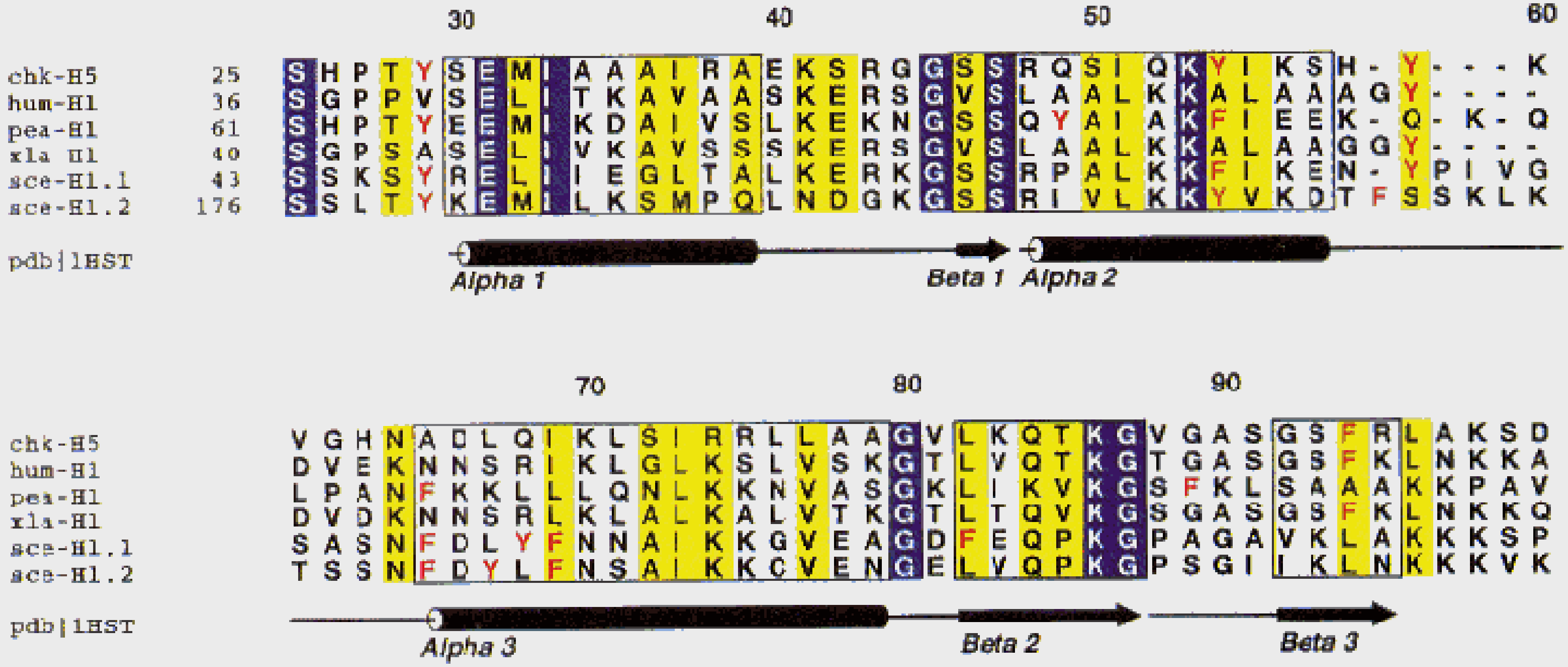
- multiple sequence alignment
- consensus sequence
- manual methods
- simultaneous and progressive methods
- databases of multiple sequence alignments
- hybrid approach for database searching

# Multiple sequence alignment

- Multiple sequence alignment is a 2D table in which the rows represent individual sequences and the columns the residue positions.
- Multiple sequence alignments are essential for analysis of sets of gene families.
- Sequence-based multiple sequence alignments - constructed according to similar strings of amino acid residues.
- Structure-based multiple sequence alignments - constructed according to structural evidence.

# Colour-coded multiple sequence alignments





```
chk-H5     25  S H P T Y S E M I A A A I R A E K S R G G S S R Q S I Q K Y I K S H - Y - - - K
hum-H1     36  S G P P V S E L I T K A V A A S K E R S G V S L A A L K K A L A A A G Y - - - -
pea-H1     61  S H P T Y E E M I K D A I V S L K E K N G S S Q Y A I A K F I E E K - Q - K - Q
xla H1     40  S G P S A S E L I V K A V S S S K E R S G V S L A A L K K A L A A G G Y - - - -
sce-H1.1   43  S S K S Y R E L I I E G L T A L K E R K G S S R P A L K K F I K E N - Y P I V G
sce-H1.2  176  S S L T Y K E M I L K S M P Q L N D G K G S S R I V L K K Y V K D T F S S K L K

chk-H5         V G H N A D L Q I K L S I R R L L A A G V L K Q T K G V G A S G S F R L A K S D
hum-H1         D V E K N N S R I K L G L K S L V S K G T L V Q T K G T G A S G S F K L N K K A
pea-H1         L P A N F K K L L L Q N L K K N V A S G K L I K V K G S F K L S A A A K K P A V
xla-H1         D V D K N N S R L K L A L K A L V T K G T L T Q V K G S G A S G S F K L N K K Q
sce-H1.1       S A S N F D L Y F N N A I K K G V E A G D F E Q P K G P A G A V K L A K K K S P
sce-H1.2       T S S N F D Y L F N S A I K K C V E N G E L V Q P K G P S G I I K L N K K K V K
```

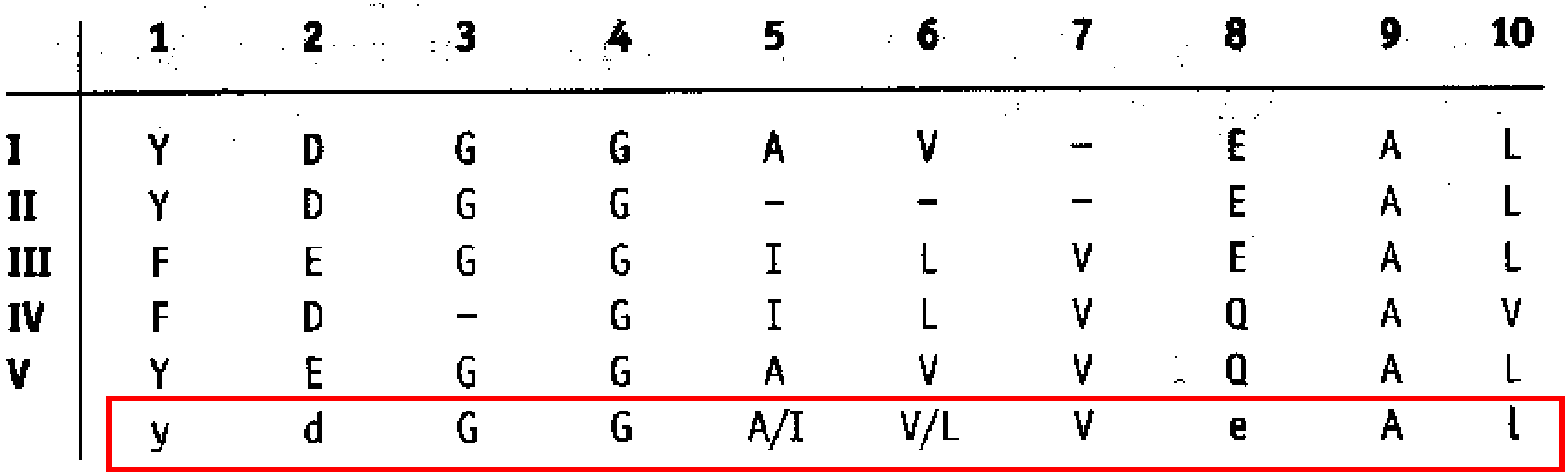
# Multiple sequence alignment

- Construction of a multiple sequence alignment:
  - positioning of residues within any sequence is preserved (absolute positions)
  - similar residues in all sequences are brought into vertical register (relative positions)
- All residues in any single column of an alignment will have the same relative position but different absolute position (unless the sequences are identical).

# Consensus sequence

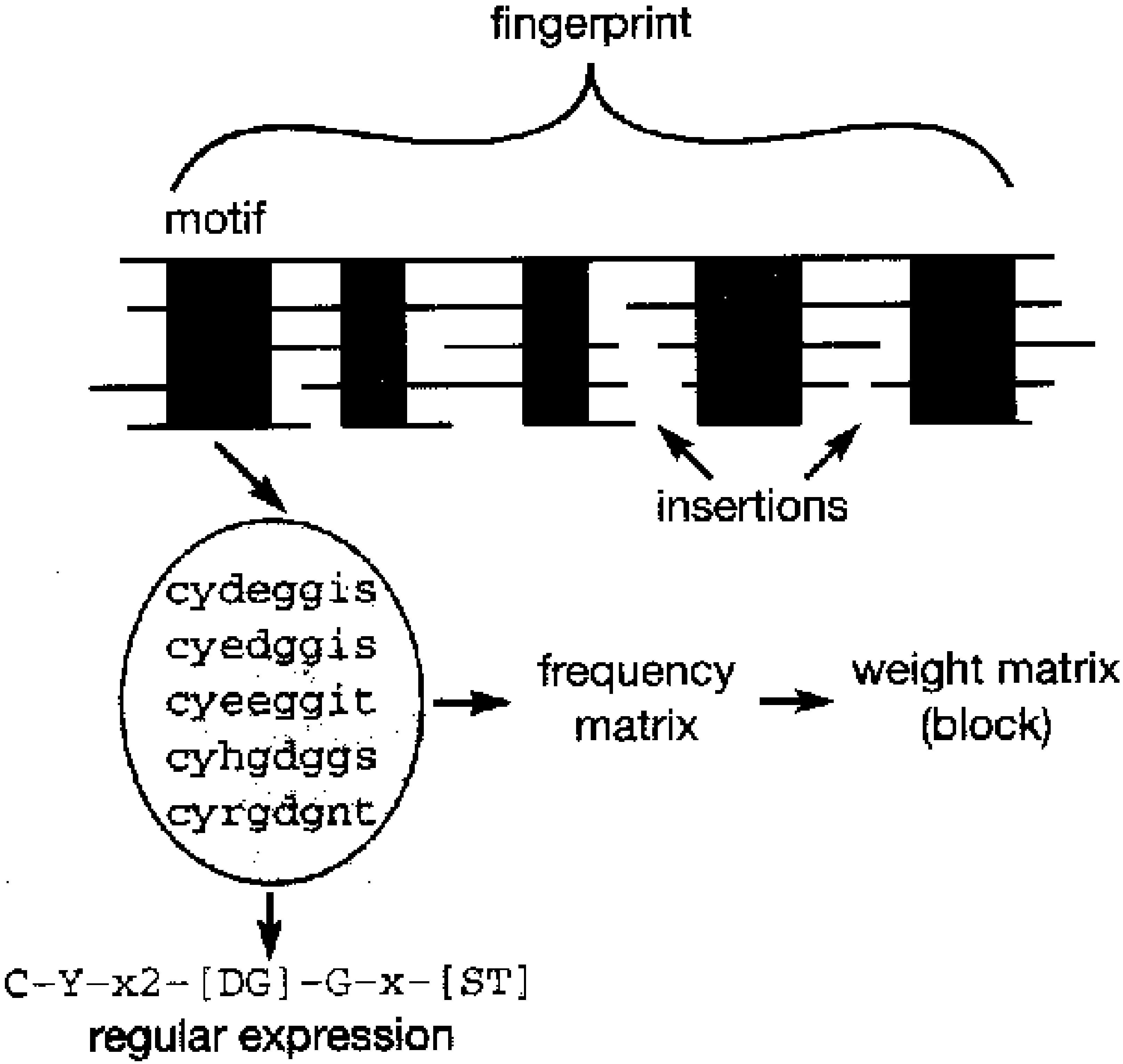
- The alignment table can be summarised by:
  - a single line: pseudo-sequence
  - unweighted matrix: fingerprint
  - ungapped block of residues (weighted): block
  - weighted matrix: profile

# Multiple alignment and the consensus sequence



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| I | Y | D | G | G | A | V | – | E | A | L |
| II | Y | D | G | G | – | – | – | E | A | L |
| III | F | E | G | G | I | L | V | E | A | L |
| IV | F | D | – | G | I | L | V | Q | A | V |
| V | Y | E | G | G | A | V | V | Q | A | L |
| | y | d | G | G | A/I | V/L | V | e | A | l |

# Multiple alignment and the profile, block and fingerprint



fingerprint

motif

insertions

```
cydeggis
cyedggis
cyeeggit
cyhgdggs
cyrgdgnt
```

frequency matrix

weight matrix (block)

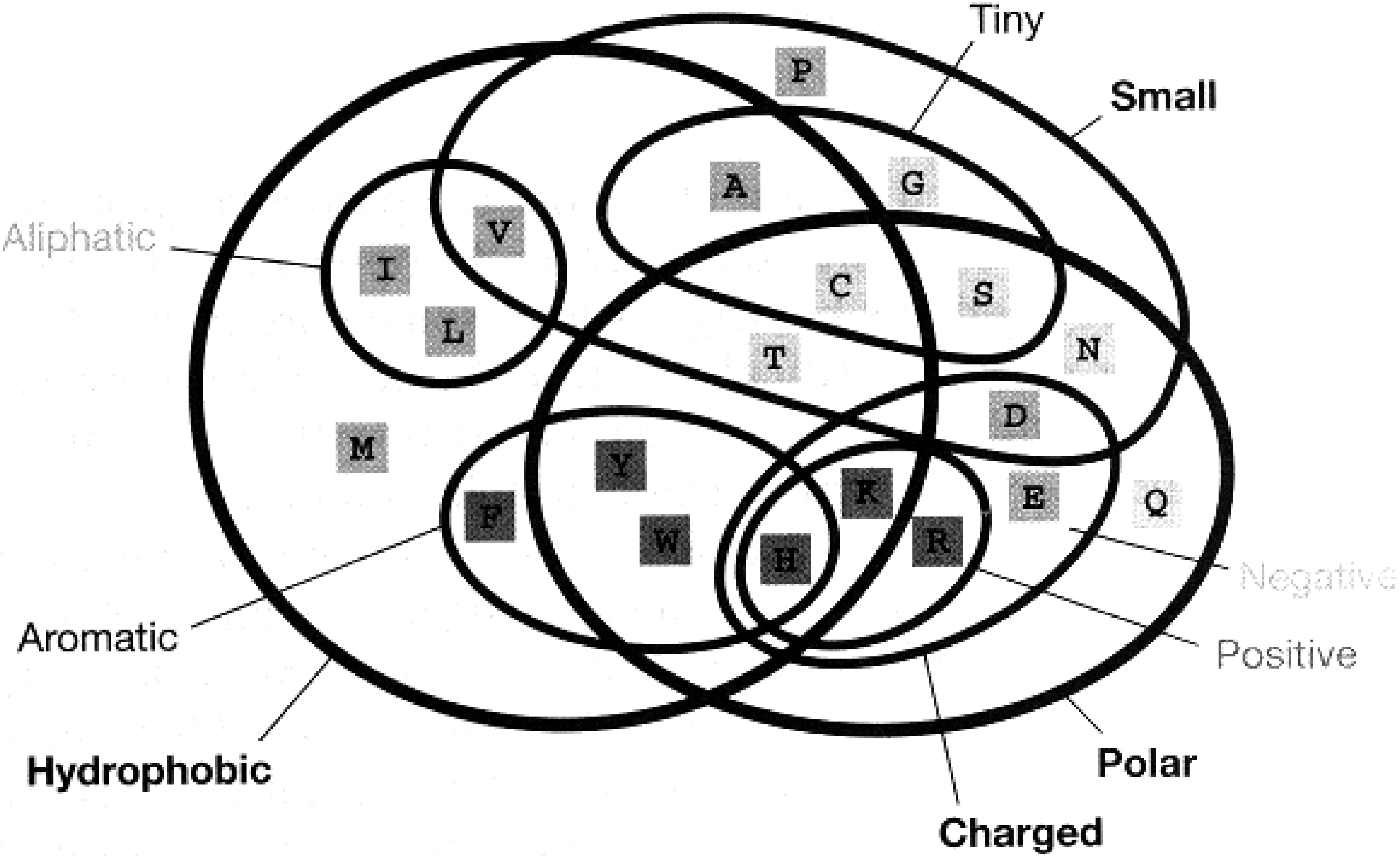
C-Y-x2-[DG]-G-x-[ST]

regular expression

# Manual methods

- Manual methods are subjective however they enable to incorporate  experimental evidences (e.g., mutagenesis data, structural knowledge) into the multiple alignment.
- Manual modification of the multiple alignments from automatic methods is the best approach.
- Intuitive colouring schemes assist the eye in spotting similarities.
- Quantitative evaluation of relatedness through calculation of residue identities/similarities.

| *Residue* | *Property* | *Colour* |
|---|---|---|
| Asp, Glu | Acidic | red |
| His, Arg, Lys | Basic | blue |
| Ser, Thr, Asn, Gln | Polar neutral | green |
| Ala, Val, Leu, Ile, Met | Hydrophobic aliphatic | white |
| Phe, Try, Trp | Hydrophobic aromatic | purple |
| Pro, Gly | Special structural properties | brown |
| Cys | Disulphide bond former | yellow |


Tiny
Small
Aliphatic
Aromatic
Hydrophobic
Charged
Polar
Positive
Negative
P
A
G
V
I
L
C
S
N
T
D
M
Y
K
E
Q
F
W
H
R

# Simultaneous methods

- Simultaneous methods align all sequences in a given set at once, rather than aligning pairs of sequences or building sequence clusters.
- Extension of 2D dynamic programming matrix to more dimensions.
- Number of dimensions = number of sequences.
- Suitable only for small sets of short sequences.

# Progressive methods

- Multi-dimensional programming matrix is not applicable to realistic problems - larger sets of longer sequences.
- CLUSTAL
  - 1. construction of evolutionary tree
  - 2. pairwise alignment of closely related sequences, addition of less related sequences
  - 3. final alignment, final evolutionary three
- CLUSTALW
  - positioning of gaps in closely related sequences according to their variability

# Databases of multiple alignments

- Multiple alignments bring together sequences from different species. This important evolutionary information can enhance sensitivity of database searches.
- Various abstractions (regular expressions, profiles, blocks, fingerprints or HMMs) can be searched against sequence databases. More information used in a query - higher sensitivity.
- Results of the searches using the multiple alignments are more difficult to interpret.

# Databases of multiple alignments

- Multiple alignments databases available via Web are produced automatically (e.g., PFAM) or manually (e.g., PRINTS).

- Iterative automatic methods may include false-positive sequences in the alignment which will corrupt it by insertion of many unrealistic gaps.

# zf-C2H2

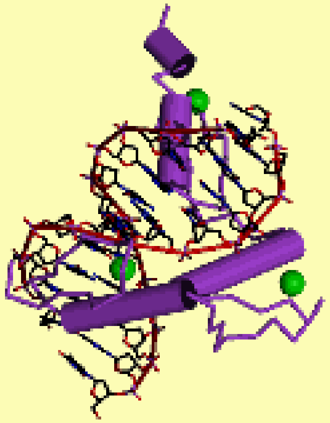


**Figure 1: 1a1h**
**Complex (zinc finger/dna)**
Qgsr (zif268 variant) zinc finger-dna complex (gcac site)

***Accession number:*** **PF00096**

**Zinc finger, C2H2 type**

The C2H2 zinc finger is the classical zinc finger domain. The two conserved cysteines and histidines co-ordinate a zinc ion. The following pattern describes the zinc finger. #-X-C-X(1-5)-C-X3-#-X5-#-X2-H-X(3-6)-[H/C] Where X can be any amino acid, and numbers in brackets indicate the number of residues. The positions marked # are those that are important for the stable fold of the zinc finger. The final position can be either his or cys. The C2H2 zinc finger is composed of two short beta strands followed by an alpha helix. The amino terminal part of the helix binds the major groove in DNA binding zinc fingers.

INTERPRO description (entry IPR000822)

Zinc finger domains [MEDLINE:88151019], PUB00005329 are nucleic acid-binding protein structures first identified in the Xenopus transcription factor TFIIIA. These domains have since been found in numerous nucleic acid-binding proteins. A zinc finger domain is composed of 25 to 30 amino-acid residues including 2 conserved Cys and 2 conserved His residues in a C-2-C-12-H-3-H type motif. The 12 residues separating the second Cys and the first His are mainly polar and basic, implicating this region in particular in nucleic acid binding. The zinc finger motif is an unusually small, self-folding domain in which Zn is a crucial component of its tertiary structure. All bind 1 atom of Zn in a tetrahedral array to yield a finger-like projection, which interacts with nucleotides in the major groove of the nucleic acid. The Zn binds to the conserved Cys and His residues. Fingers have been found to bind to about 5 base pairs of nucleic acid containing short runs of guanine residues. They have the ability to bind to both RNA and DNA, a versatility not demonstrated by the helix-turn-helix motif. The zinc finger may thus represent the original nucleic acid binding protein. It has also been suggested that a Zn-centred domain could be used in a protein interaction, e.g. in protein kinase C. Many classes of zinc fingers are characterized according to the number and positions of the histidine and cysteine residues involved in the zinc atom coordination. In the first class to be characterized, called C2H2, the first pair of zinc coordinating residues are cysteines, while the second pair are histidines.

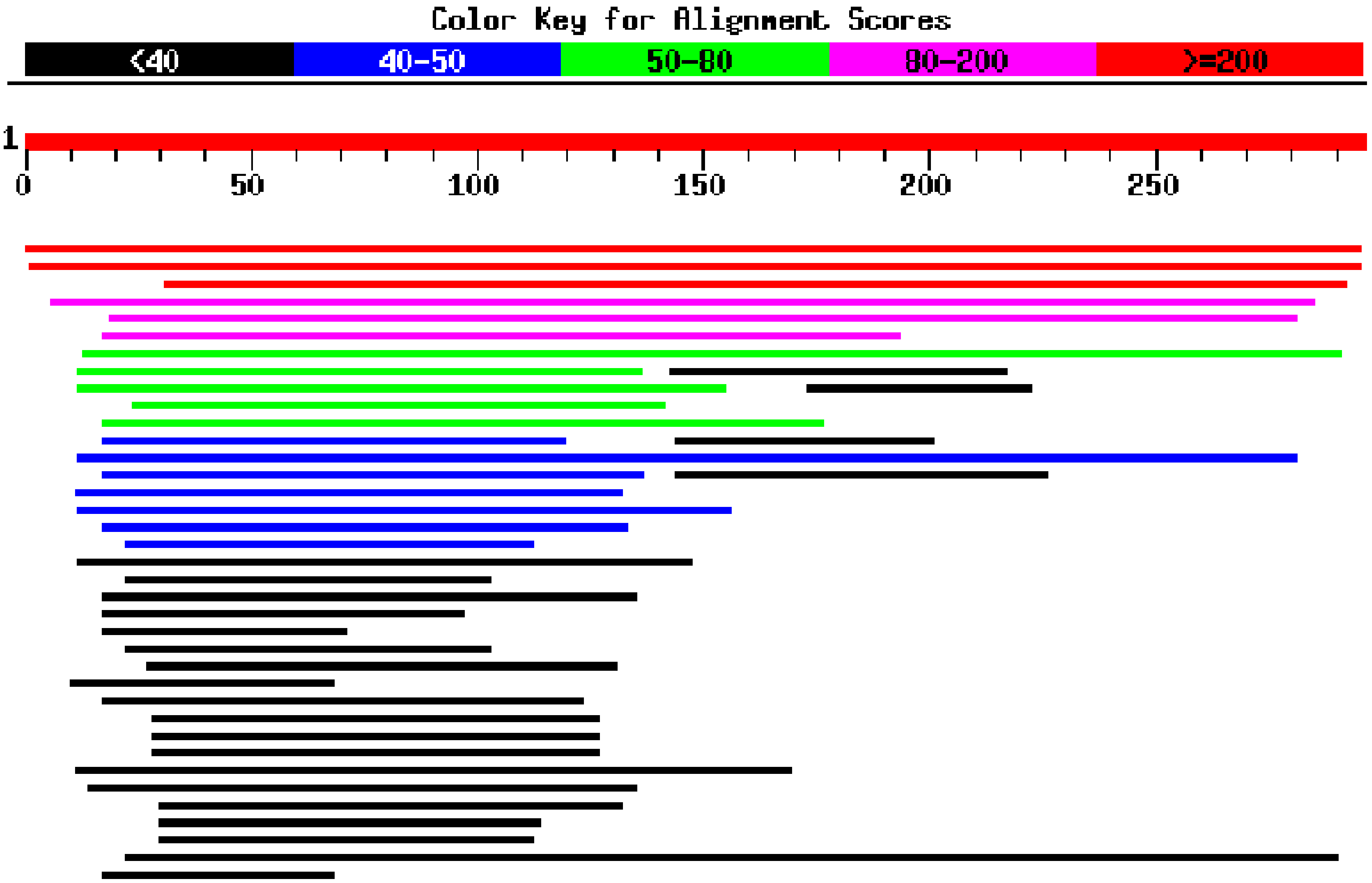
For additional annotation, see the PROSITE document PDOC00028 [ Expasy | SRS-UK | SRS-USA ]

| | | |
|---|---|---|
| TYY1_HUMAN/383-407 | YVCPF.DGCN...KKFAQSTNLKSHILT...H | P25490 |
| ZG52_XENLA/61-83 | YTCT...QCN...KQFSHSAQLRAHIST...H | P18727 |
| KRUP_DROME/306-328 | YTCE...ICD...GKFSDSNQLKSHMLV...H | P07247 |
| YKQ8_CAEEL/78-102 | YKCT...VCR...KDISSSESLRTHMFKQ.HH | P34303 |
| DEFI_CHICK/268-292 | YECP...NCK...KRFSHSGSYSSHISSK.KC | P36197 |
| ZFH1_DROME/389-413 | FGCD...NCG...KRFSHSGSFSSHMTSK.KC | P28166 |
| YL57_CAEEL/42-65 | YLCY...YCG...KTLSDRLEYQQHMLK..VH | P34437 |
| ZFA_MOUSE/542-564 | FKCD...ICL...LTFSDTKEVQQHALV...H | P23607 |
| BASO_HUMAN/719-742 | FQCD...ICK...KTFKNACSVKIHHKN..MH | Q01954 |
| HUNB_DROME/297-319 | FQCD...KCS...YTCVNKSMLNSHRKS...H | P05084 |
| SFP1_YEAST/598-623 | FKCPV.IGCE...KTYKNQNGLKYHRLH..GH | P32432 |
| ZG29_XENLA/62-84 | FVCT...VCG...KTYKYKHGLNTHLHS...H | P18717 |
| BASO_HUMAN/927-950 | ITCH...LCQ...KTYSNKGTFRAHYKT..VH | Q01954 |
| XFIN_XENLA/326-348 | YSCS...KCR...KTFKRWKSFLNHQQT...H | P08045 |
| XFIN_XENLA/503-525 | HKCS...KCD...LTFSHWSTFMKHSKL...H | P08045 |
| ZG44_XENLA/5-27 | FACT...KCK...RRFCSNKELFSHKRI...H | P18721 |
| FZF1_YEAST/72-94 | KACT...LCQ...KRFVTNQQLRRHLNS...H | P32805 |
| CF2_DROME/366-388 | HKCP...DCP...KTFKTPGTLAMHRKI...H | P20385 |
| SUHW_DROME/290-313 | INCP...DCP...KSFKTQTSYERHIFI..TH | P08970 |
| P43_XENBO/75-100 | HSCPT.AGCK...MTFSTKKSLSRHKLY..KH | P25066 |
| DISC_DROME/92-115 | VQCS...ICF...KTFCDKGALKIHFSA..VH | P23792 |
| RME1_YEAST/256-281 | LNCPF.PICQ...KTFRRKDAYKRHVAM..VH | P32338 |
| P43_XENBO/106-130 | LKCSV.PGCK...RSFRKKRALRIHVSE...H | P25066 |
| SRYD_DROME/307-329 | IICS...ICN...VSFKSRKTFNHHTLI...H | P07664 |
| SRYD_DROME/404-427 | GFCL...ICN...TTFENKKELEHHLQF..DH | P07664 |
| ACE2_YEAST/603-627 | FECLY.PNCN...KVFKRRYNIRSHIQT...H | P21192 |
| IKAR_MOUSE/488-512 | FECN...MCG...YHSQDRYEFSSHITRG.EH | Q03267 |
| SRYD_DROME/193-216 | QECT...TCG...KVYNSWYQLQKHISE..EH | P07664 |
| MFG2_MOUSE/336-358 | FECK...VCG...KSFKRESNLIQHGAV...H | P16373 |
| MFG3_MOUSE/344-366 | FECK...QCG...KIFSNGSYLLRHYDT...H | P16374 |
| MFG3_MOUSE/484-506 | FECK...ECG...KAFHFSSQLNNHKTS...H | P16374 |
| ZFX_HUMAN/488-510 | IECD...ECG...KHFSHAGALFTHKMV...H | P17010 |
| ZG52_XENLA/6-27 | FTCP...ECG...KRF.SQKSNCWHTED...H | P18727 |
| ZG8_XENLA/146-168 | FTCT...ECG...EHFANKVSLLGHLKM...H | P18737 |
| ZO61_XENLA/62-84 | FTCF...ECG...TCFVNYSWLMLHIRM...H | P18750 |
| HKR3_HUMAN/319-342 | FECP...KCG...KCYFRKENLLEHEAR..NC | P10074 |
| ZG28_XENLA/174-196 | FTCT...ECG...KCLTRQYQLTEHSYL...H | P18716 |
| ZG3_XENLA/6-28 | FMCT...KCG...KCLSTKQKLNLHHMT...H | P18718 |
| P43_XENBO/136-160 | SVCDV.PGCG...WKSTSAAKLAAHHRR...H | P25066 |
| ZKR1_CHICK/169-191 | HKCQ...HCG...KPFAGAAQLLAHSRG...H | P30373 |
| HUNB_DROME/733-757 | FKCN...MCG...EKCDGPVGLFVHMARN.AH | P05084 |
| P43_XENBO/45-69 | WKCGK.KDCG...KMFARKRQIQKHMKR...H | P25066 |
| ZG5A_XENLA/90-112 | FSCT...VCG...EMFTYRAQFSKHMLK...H | P18726 |
| TSH_DROME/466-490 | LKCM...RCG...ESFRSLGEMTKHMQET.QH | P22265 |

# Hybrid approach for database searching

- PSI-BLAST
  - Position-Specific Iterated - BLAST
  - algorithm by Altschul *et al.* (1997)
  - incorporates elements of both pairwise and multiple sequence alignment methods
  - procedure: initial search - creation of position specific profiles from the hits - new search ... in iterations
  - advantage: detects even very weak similarities
  - disadvantages: the profile can be diluted if low-complexity regions are not masked; inclusion of single false-positive sequence into the profile leads to bias towards unrelated sequences

# Graphic hit list from a database search using PSI-BLAST


Color Key for Alignment Scores
<40
>=200