Patterns, Profiles and Multiple sequence alignments

Arne Elofsson

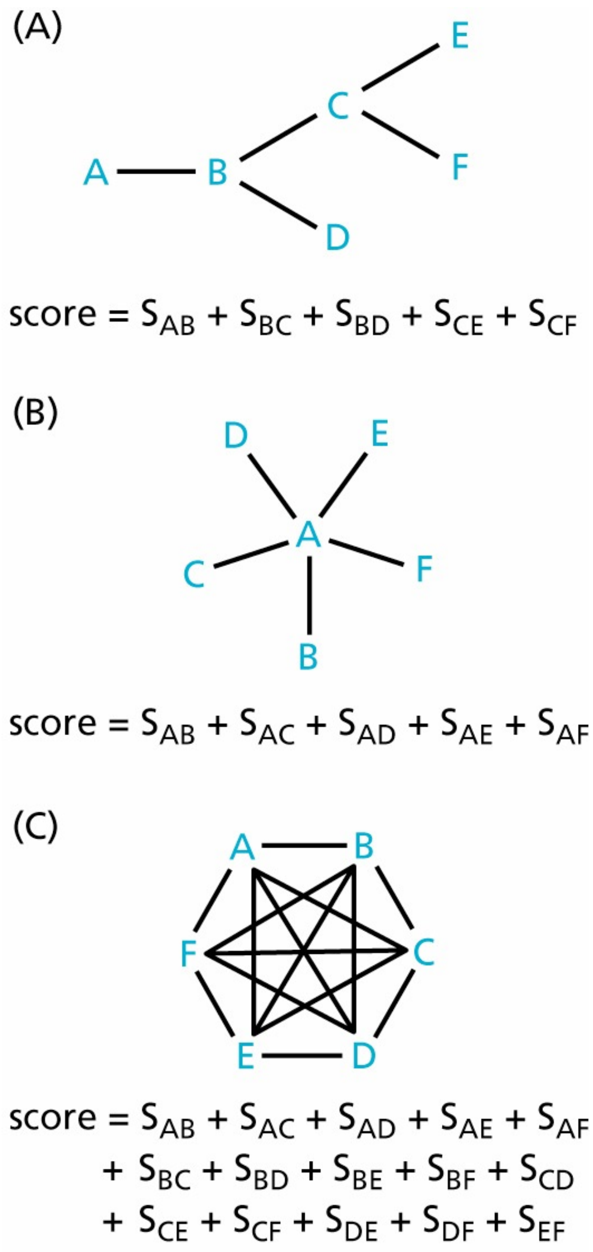
To read:
http://en.wikipedia.org/wiki/Multiple_sequence_alignment
http://en.wikipedia.org/wiki/Hidden_Markov_model
Extra
http://en.wikipedia.org/wiki/HMMER
http://en.wikipedia.org/wiki/HHpred_/_HHsearch
http://www.ploscollections.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.0030123
http://www.ncbi.nlm.nih.gov/turorials/BLAST

How to obtain MSAs

- Exact solution is impossible for a handful of sequences (2^N-1 alternatives)
- Popular methods include:
  - ClustalW
  - T-coffee
  - kalign
  - PSIBLAST

$$2^N - 1$$

Some scoring in MSAs

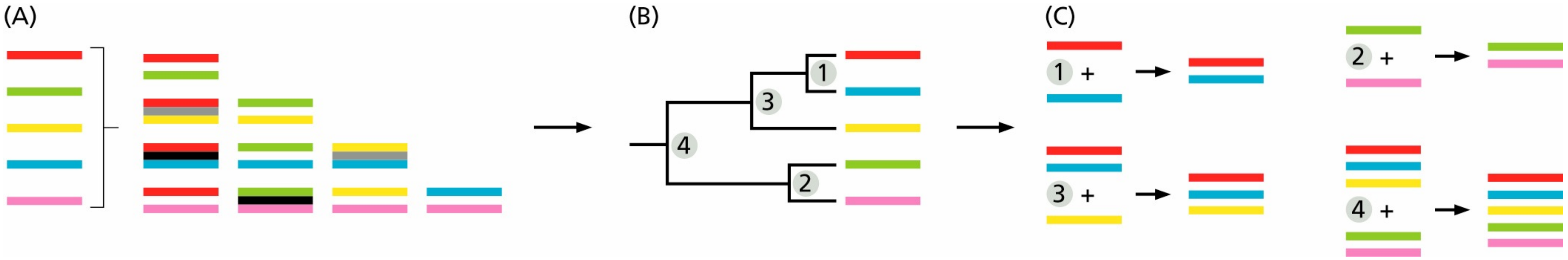


score = $S_{AB} + S_{BC} + S_{BD} + S_{CE} + S_{CF}$

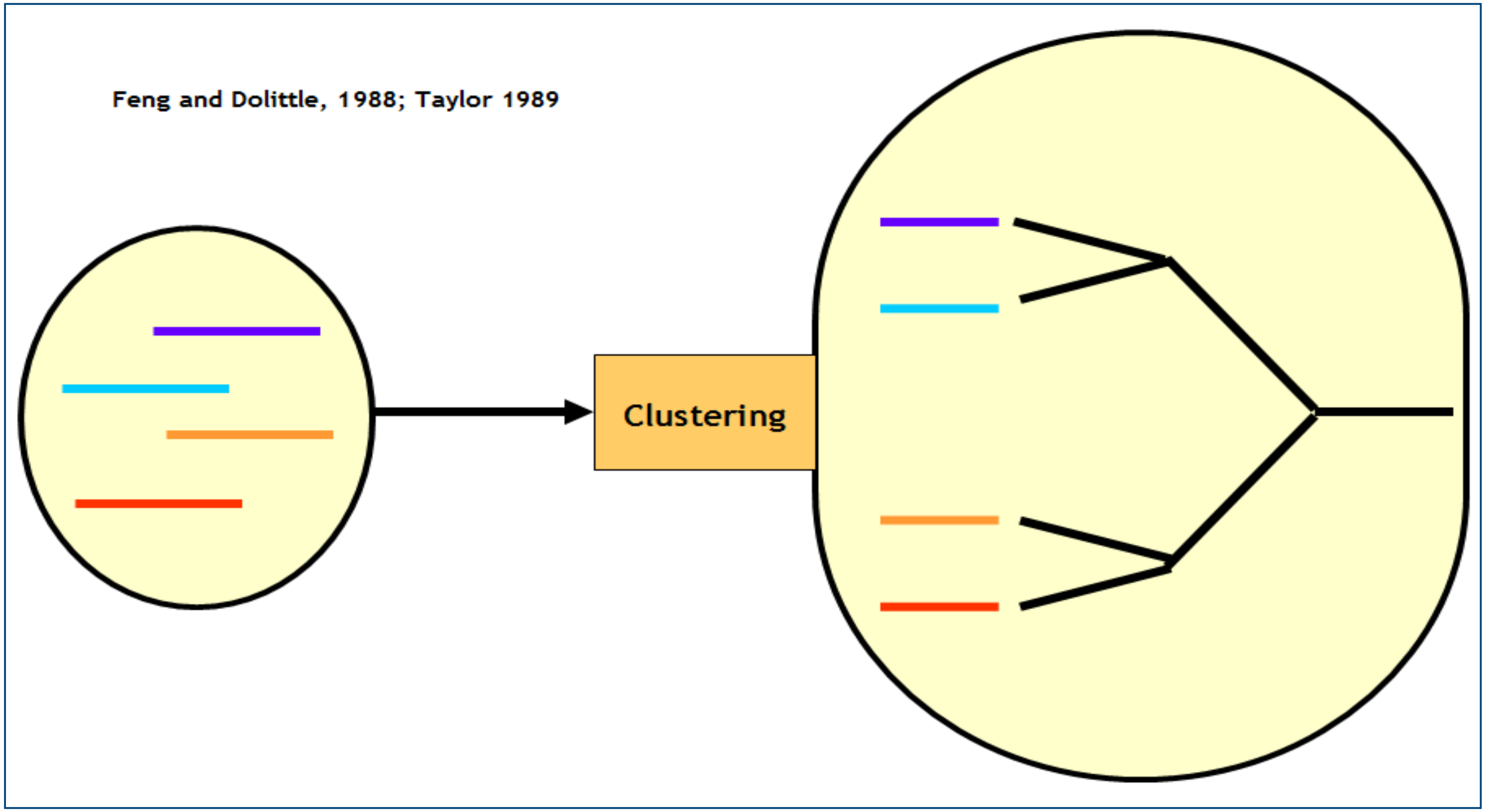
score = $S_{AB} + S_{AC} + S_{AD} + S_{AE} + S_{AF}$

score = $S_{AB} + S_{AC} + S_{AD} + S_{AE} + S_{AF}$
$+ S_{BC} + S_{BD} + S_{BE} + S_{BF} + S_{CD}$
$+ S_{CE} + S_{CF} + S_{DE} + S_{DF} + S_{EF}$

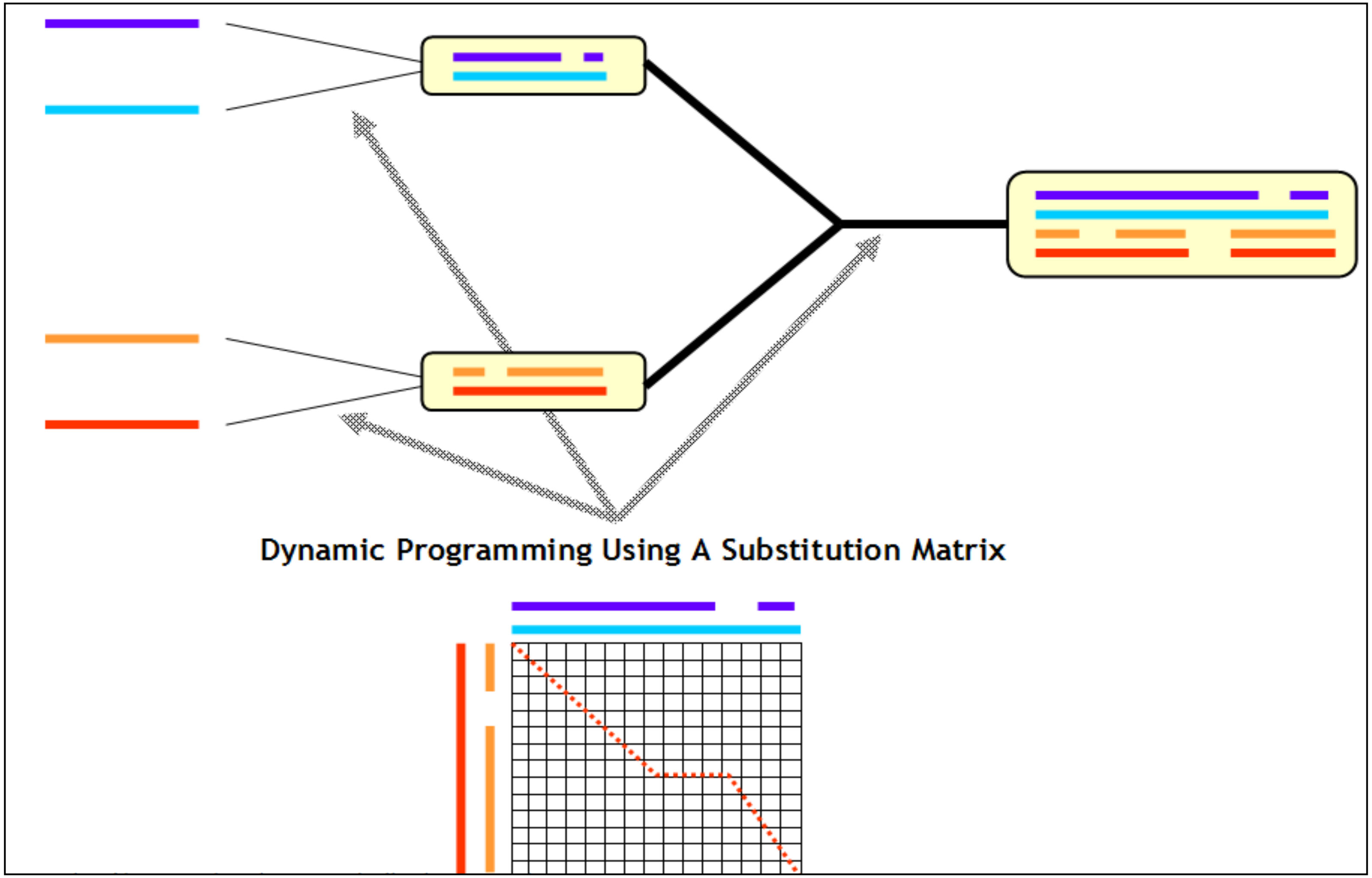
A progressive MSA



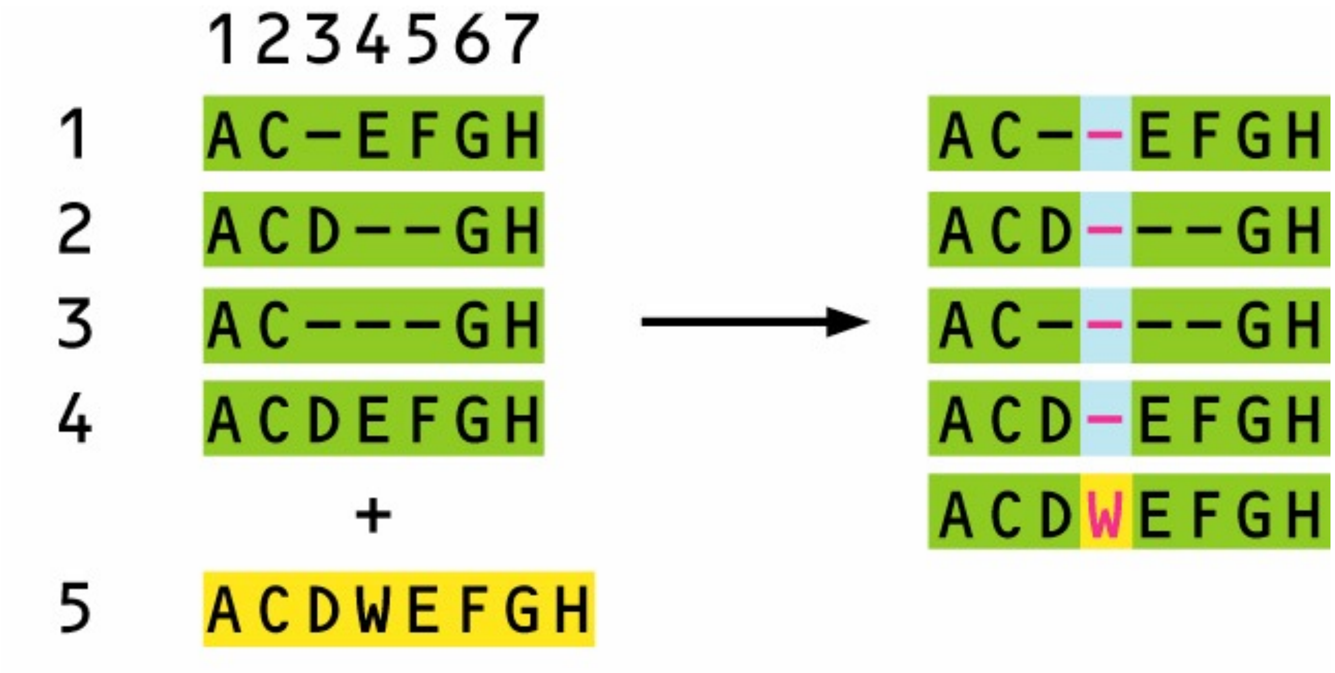
ClustalW



ClustalW

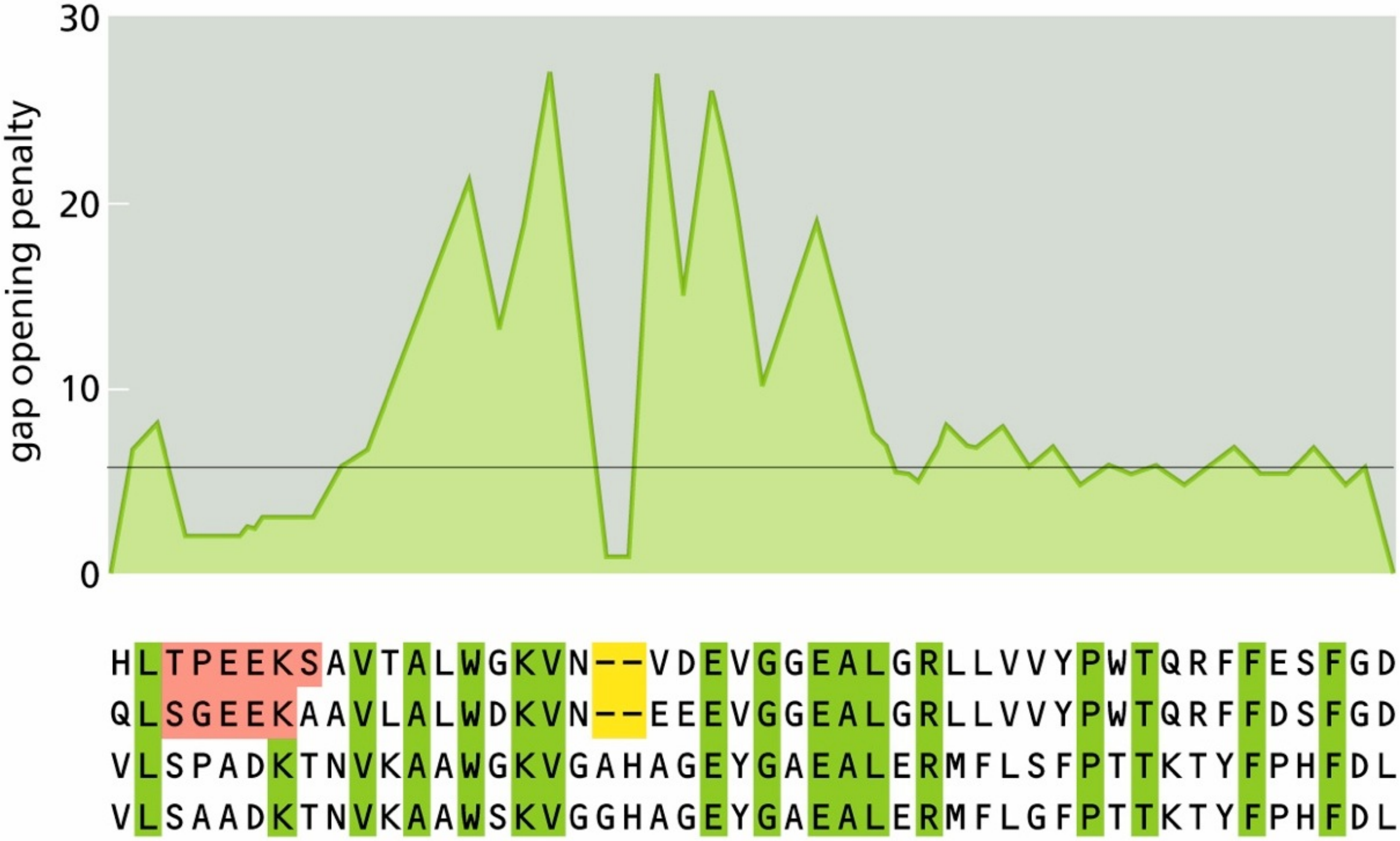


The gap scoring problem



```
  1234567
1 AC-EFGH          AC--EFGH
2 ACD--GH          ACD---GH
3 AC---GH   --->   AC----GH
4 ACDEFGH          ACD-EFGH
     +             ACDWEFGH
5 ACDWEFGH
```

ClustalW gap-opening penalty

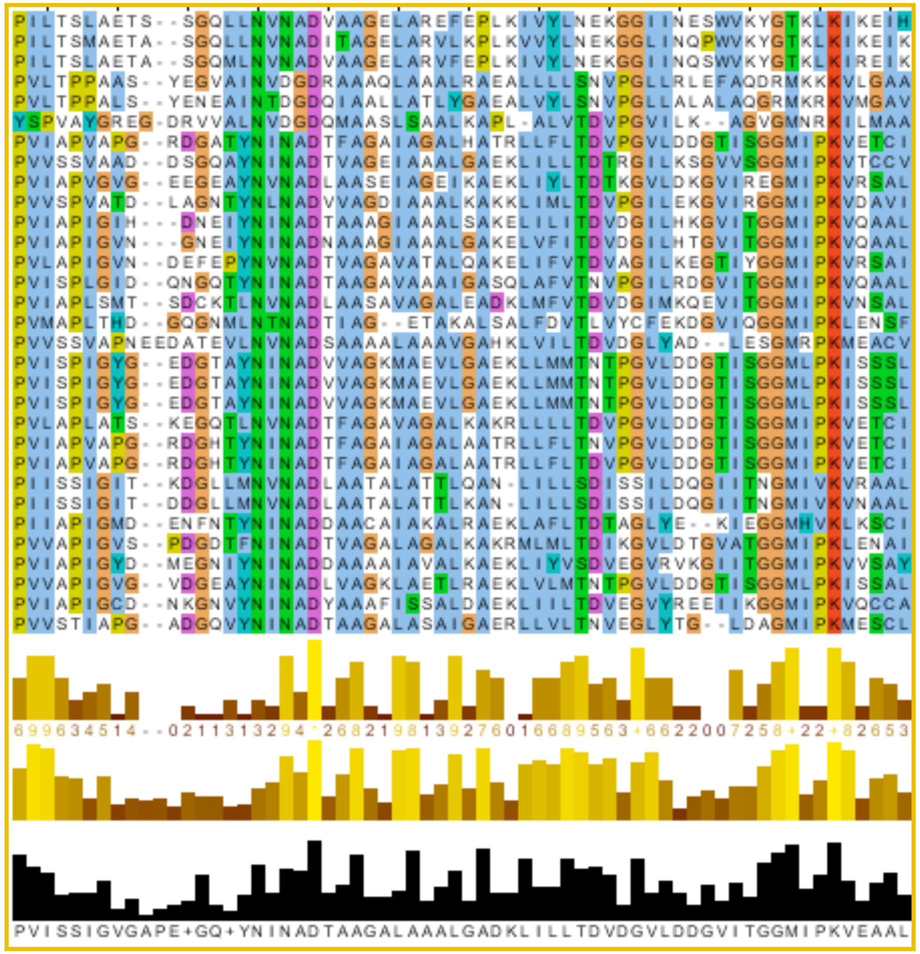


```
HLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGD
QLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGD
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDL
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDL
```

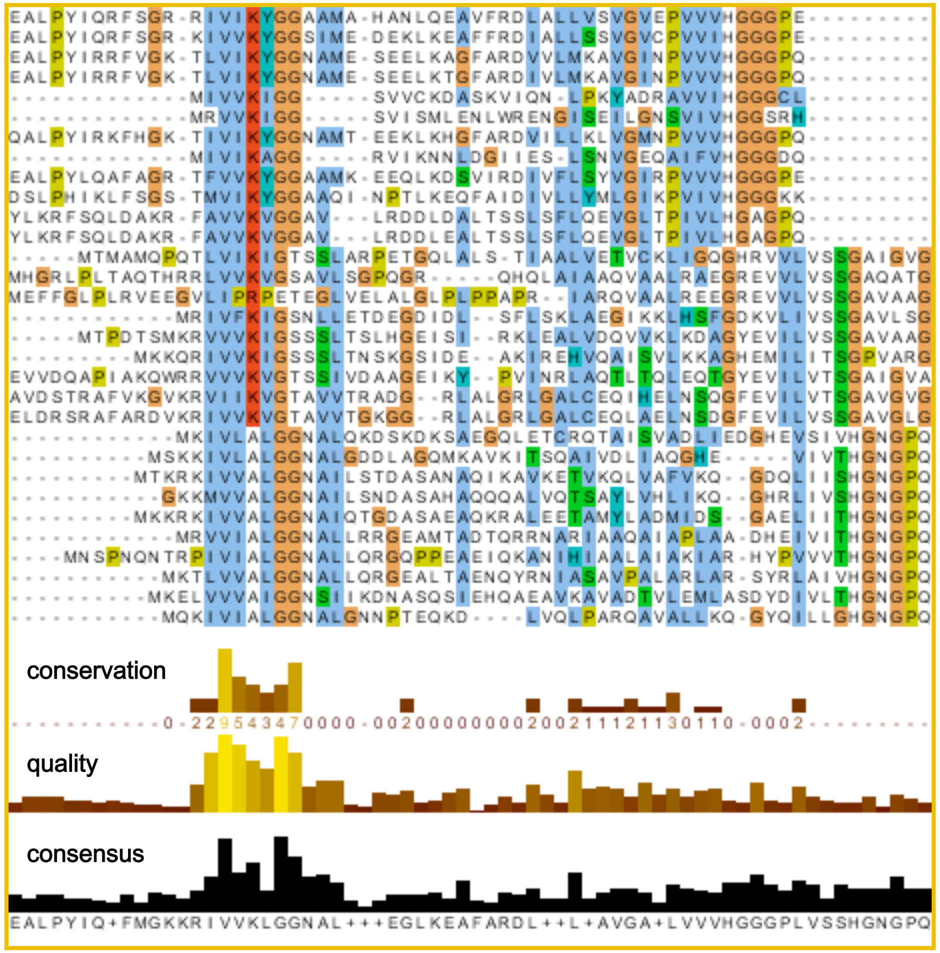
Multiple sequence alignments

- Some information that can be obtained from a multiple sequence alignment

Multiple sequence alignments
(good ones look pretty !!)


PVISSIGVGAPE+GQ+YNINADTAAGALAAALGADKLILLTDVDGVLDDGVITGGMIPKVEAAL

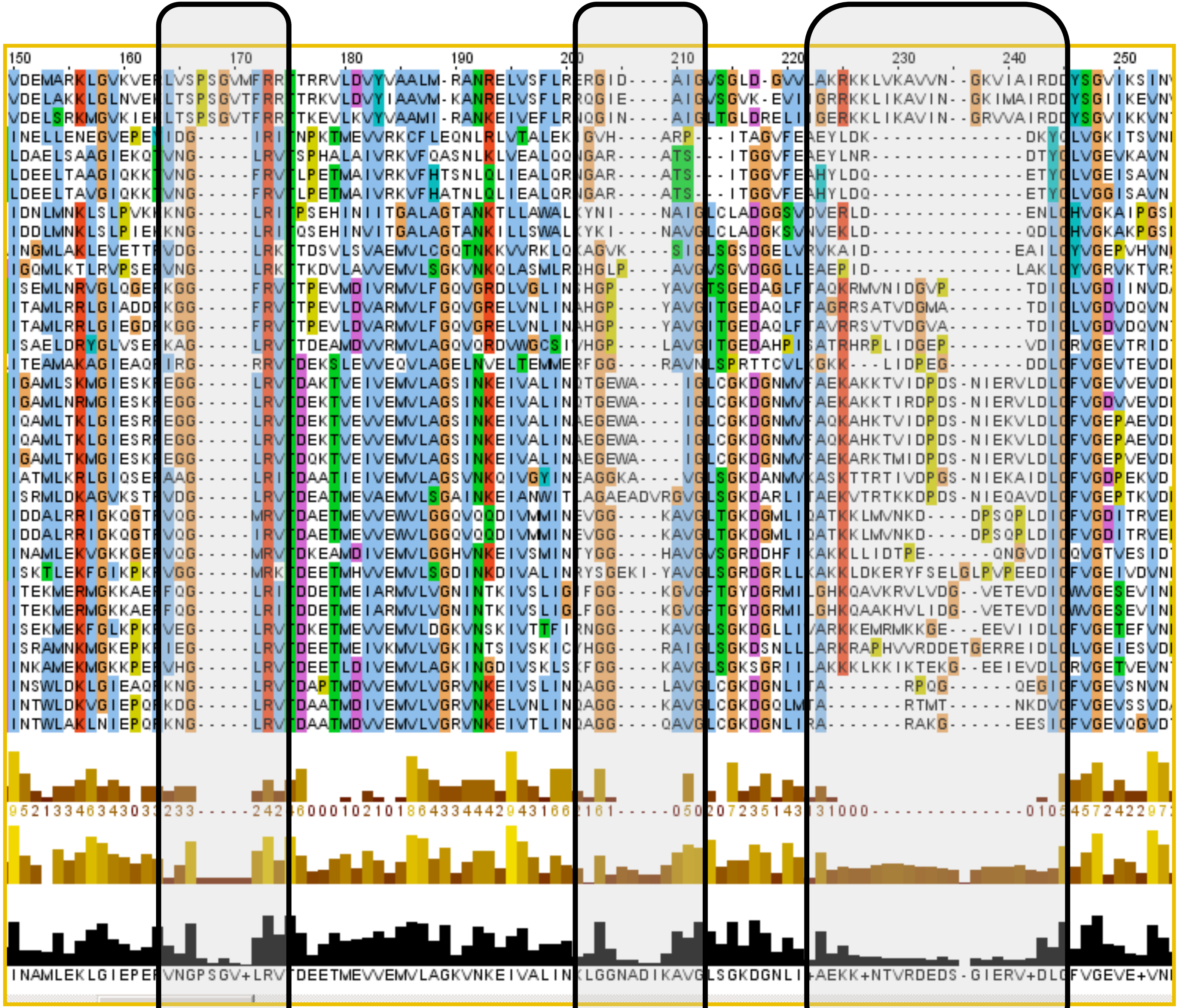
☺


conservation
quality
consensus
EALPYIQ+FMGKKRIVVKLGGNAL+++EGLKEAFARDL++L+AVGA+LVVVHGGGPLVSSHGNGPQ

?

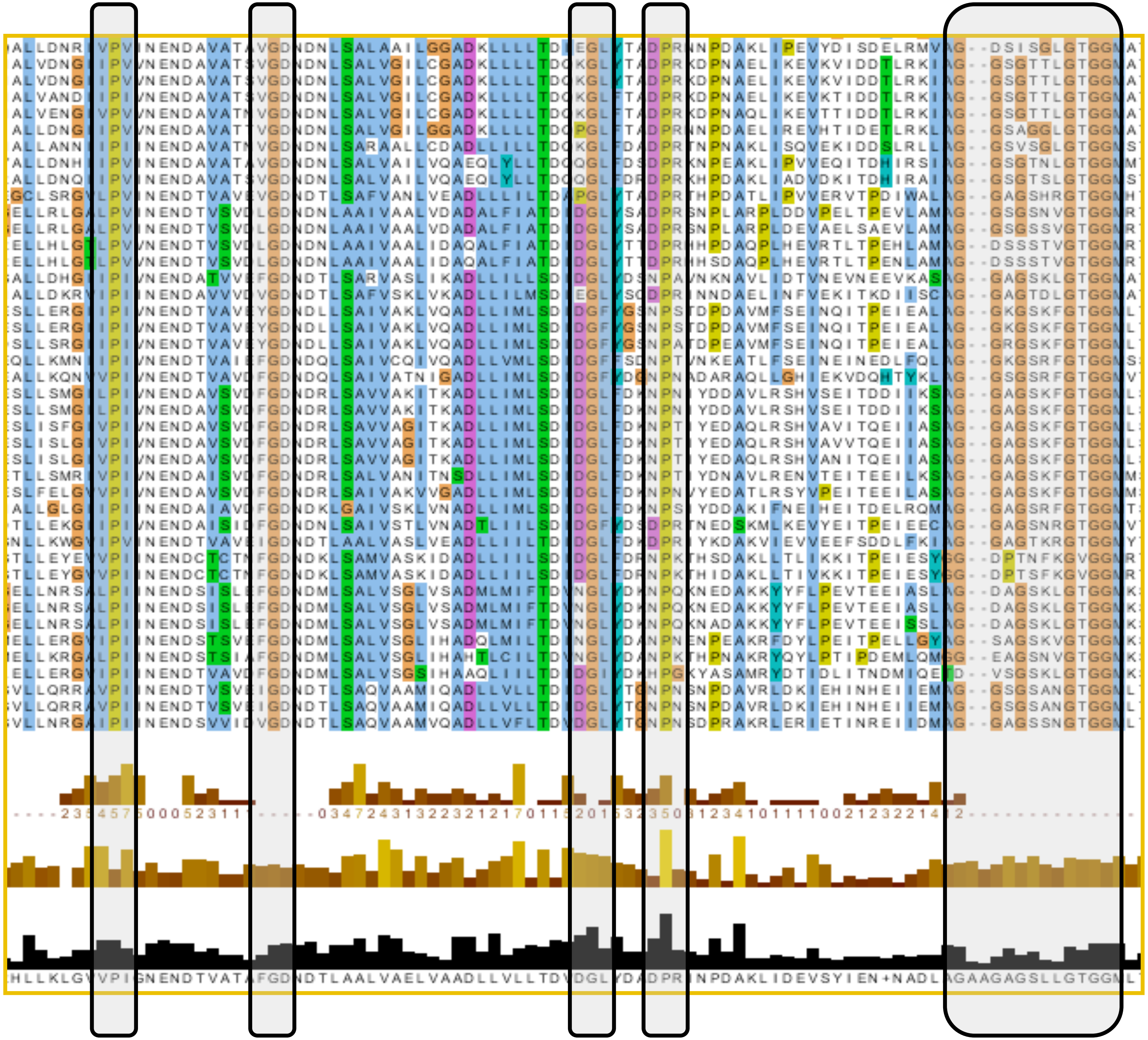
Features found in MSA:

The position of insertions and deletions suggests regions where surface loops exist…

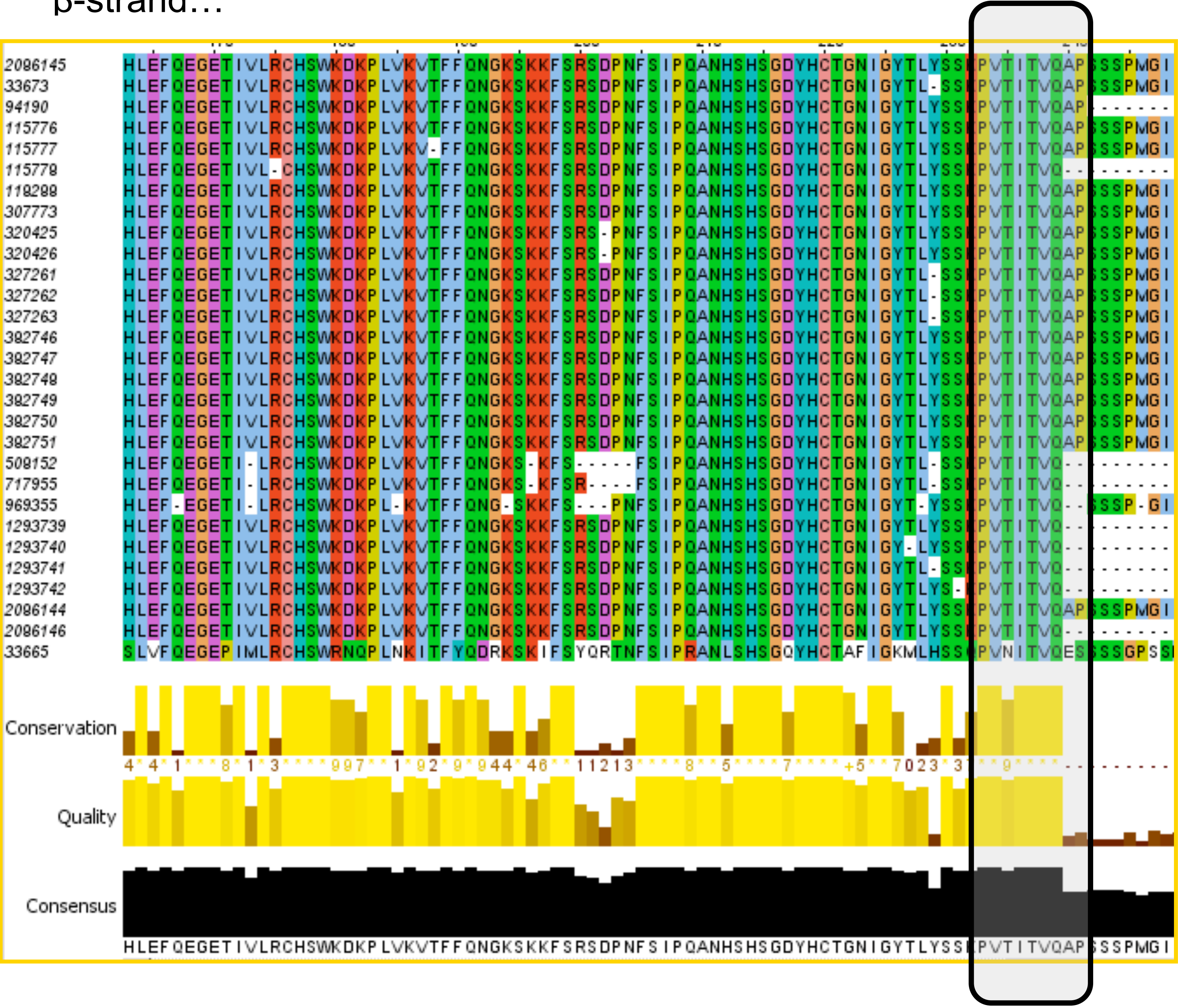


```
150       160       170       180       190       200       210       220       230       240       250
VDEMARKLGVKVERLVSPSGVMFRRTTRRVLDVYVAALM-RANRELVSFLRERGID----AIGVSGLD-GVVLAKRKKLVKAVVN--GKVIAIRDDYSGVIKSINY
VDELAKKLGLNVEKLTSPSGVTFRRTTRKVLDVYIAAVM-KANRELVSFLRRQGIE----AIGVSGVK-EVIIGRRKKLIKAVIN--GKIMAIRDDYSGIIKEVNY
VDELSRKMGVKIEKLTSPSGVTFRRTTKEVLKVYVAAMI-RANKEIVEFLRNQGIN----AIGLTGLDRELIIGERKKLIKAVIN--GRVVAIRDDYSGVIKKVNT
INELLENEGVEPEYIDG-----IRITNPKTMEVVRKCFLEQNLRLVTALEKIGVH----ARP---ITAGVFEAEYLDK--------------DKYGLVGKITSVNI
LDAELSAAGIEKQTVNG-----LRVTSPHALAIVRKVFQASNLKLVEALQQNGAR----ATS---ITGGVFEAEYLNR--------------DTYGLVGEVKAVN
LDEELTAAGIQKKTVNG-----FRVTLPETMAIVRKVFHTSNLQLIEALQRNGAR----ATS---ITGGVFEAHYLDQ--------------ETYGLVGEISAVN
LDEELTAVGIQKKTVNG-----FRVTLPETMAIVRKVFHATNLQLIEALQRNGAR----ATS---ITGGVFEAHYLDQ--------------ETYGLVGGISAVN
IDNLMNKLSLPVKKKNG-----LRITPSEHINIITGALAGTANKTLLAWALKYNI----NAIGLCLADGGSVDVERLD--------------ENLGHVGKAIPGSI
IDDLMNKLSLPIEKKNG-----LRITQSEHINVITGALAGTANKILLSWALKYKI----NAVGLCLADGKSVNVEKLD--------------QDLGHVGKAKPGSI
INGMLAKLEVETTRVDG-----LRKTTDSVLSVAEMVLCGQTNKKVVRKLQKAGVK----SIGLSGSDGELVRVKAID-------------EAILGYVGEPVHVN
IGQMLKTLRVPSERVNG-----LRKTTKDVLAVVEMVLSGKVNKQLASMLRQHGLP----AVGVSGVDGGLLEAEPID-------------LAKLGYVGRVKTVRS
ISEMLNRVGLQGERKGG-----FRVTTPEVMDIVRMVLFGQVGRDLVGLINSHGP----YAVGTSGEDAGLFTAQKRMVNIDGVP-------TDIGLVGDIINVDA
ITAMLRRLGIADDRKGG-----FRVTTPEVLDVARMVLFGQVGRELVNLINAHGP----YAVGITGEDAQLFTAGRRSATVDGMA-------TDIGLVGDVDQVN
ITAMLRRLGIEGDRKGG-----FRVTTPEVLDVARMVLFGQVGRELVNLINAHGP----YAVGITGEDAQLFTAVRRSVTVDGVA-------TDIGLVGDVDQVNT
ISAELDRYGLVSERKAG-----LRVTTDEAMDVVRMVLAGQVQRDVWGCSIVHGP----LAVGITGEDAHPISATRHRPLIDGEP-------VDIGRVGEVTRIDT
ITEAMAKAGIEAQRIRG-----RRVTDEKSLEVVEQVLAGELNVELTEMMERFGG----RAVNLSPRTTCVLKGKK---LIDPEG-------DDLGFVGEVTEVDI
IGAMLSKMGIESKREGG-----LRVTDAKTVEIVEMVLAGSINKEIVALINQTGEWA----IGLCGKDGNMVFAEKAKKTVIDPDS-NIERVLDLGFVGEVVEVDI
IGAMLNRMGIESKREGG-----LRVTDEKTVEIVEMVLAGSINKEIVALINQTGEWA----IGLCGKDGNMVFAEKAKKTIRDPDS-NIERVLDLGFVGDVVEVDI
IQAMLTKLGIESRREGG-----LRVTDEKTVEVVEMVLAGSINKEIVALINAEGEWA----IGLCGKDGNMVFAQKAHKTVIDPDS-NIEKVLDLGFVGEPAEVDI
IQAMLTKLGIESRREGG-----LRVTDEKTVEVVEMVLAGSINKEIVALINAEGEWA----IGLCGKDGNMVFAQKAHKTVIDPDS-NIEKVLDLGFVGEPAEVDI
IGAMLTKMGIESKREGG-----LRVTDQKTVEIVEMVLAGSINKEIVALINAEGEWA----IGLCGKDGNMVFAEKARKTMIDPDS-NIERVLDLGFVGEPVEVDI
IATMLKRLGIQSERAAG-----LRITDAATIEIVEMVLAGSVNKQIVGYINEAGGKA----VGLSGKDANMVKASKTTRTIVDPGS-NIEKAIDLGFVGDPEKVD
ISRMLDKAGVKSTRVDG-----LRVTDEATMEVAEMVLSGAINKEIANWITLAGAEADVRGVGLSGKDARLITAEKVTRTKKDPDS-NIEQAVDLGFVGEPTKVDI
IDDALRRIGKQGTRVQG-----MRVTDAETMEVVEWVLGGQVQQDIVMMINEVGG----KAVGLTGKDGMLIQATKKLMVNKD----DPSQPLDIGFVGDITRVEI
IDDALRRIGKQGTRVQG-----IRVTDAETMEVVEWVLGGQVQQDIVMMINEVGG----KAVGLTGKDGMLIQATKKLMVNKD----DPSQPLDIGFVGDITRVEI
INAMLEKVGKKGERVQG-----MRVTDKEAMDIVEMVLGGHVNKEIVSMINTYGG----HAVGVSGRDDHFIKAKKLLIDTPE------QNGVDIGQVGTVESIDT
ISKTLEKFGIKPKRVGG-----MRKTDEETMHVVEMVLSGDINKDIVALINRYSGEKI-YAVGLSGRDGRLLKAKKLDKERYFSELGLPVPEEDIGFVGEIVDVNI
ITEKMERMGKKAERFQG-----LRITDDETMEIARMVLVGNINTKIVSLIGIFGG----KGVGFTGYDGRMILGHKQAVKRVLVDG--VETEVDIGWVGESEVINI
ITEKMERMGKKAERFQG-----LRITDDETMEIARMVLVGNINTKIVSLIGLFGG----KGVGFTGYDGRMILGHKQAAKHVLIDG--VETEVDIGWVGESEVINI
ISEKMEKFGLKPKRVEG-----LRVTDKETMEVVEMVLDGKVNSKIVTTFIRNGG----KAVGLSGKDGLLIVARKKEMRMKKGE---EEVIIDLGFVGETEFVNI
ISRAMNKMGKEPKRIEG-----LRVTDEETMEIVKMVLVGKINTSIVSKICYHGG----RAIGLSGKDSNLLLARKRAPHVVRDDETGERREIDLGLVGEIESVDI
INKAMEKMGKKPERVHG-----LRVTDEETLDIVEMVLAGKINGDIVSKLSKFGG----KAVGLSGKSGRIILAKKKLKKIKTEKG--EEIEVDLGRVGETVEVNT
INSWLDKLGIEAQRKNG-----LRVTDAPTMDVVEMVLVGRVNKEIVSLINQAGG----LAVGLCGKDGNLITA-------RPQG------QEGIGFVGEVSNVN
INTWLDKVGIEPQRKDG-----LRVTDAATMDIVEMVLVGRVNKELVNLINQAGG----KAVGLCGKDGQLMTA-------RTMT------NKDVGFVGEVSSVDA
INTWLAKLNIEPQRKNG-----LRVTDAATMDVVEMVLVGRVNKEIVTLINQAGG----QAVGLCGKDGNLIRA-------RAKG------EESIGFVGEVQGVDT

95213346343033233-----2426000102101864334442943166 2161-----0502072351431310 00--------------0105457242297

INAMLEKLGIEPEPVNGPSGV+LRVTDEETMEVVEMVLAGKVNKEIVALINKLGGNADIKAVGLSGKDGNLI+AEKK+NTVRDEDS-GIERV+DLGFVGEVE+VNI
```

Conserved glycine or proline suggests a β-turn.

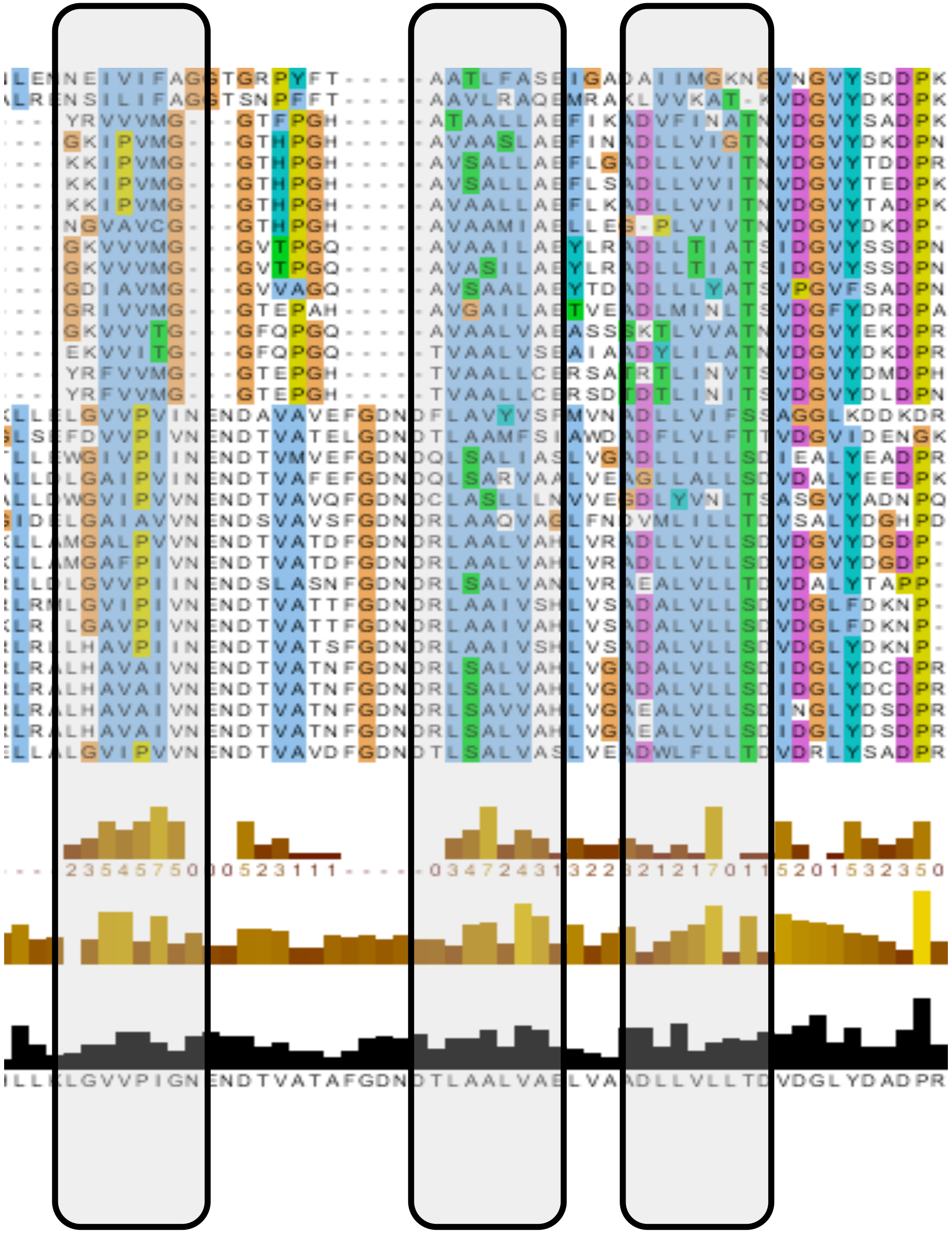


Residues with hydrophobic properties conserved at i, i+2, i+4 (etc) separated by unconserved or hydrophilic residues suggests a surface β-strand…

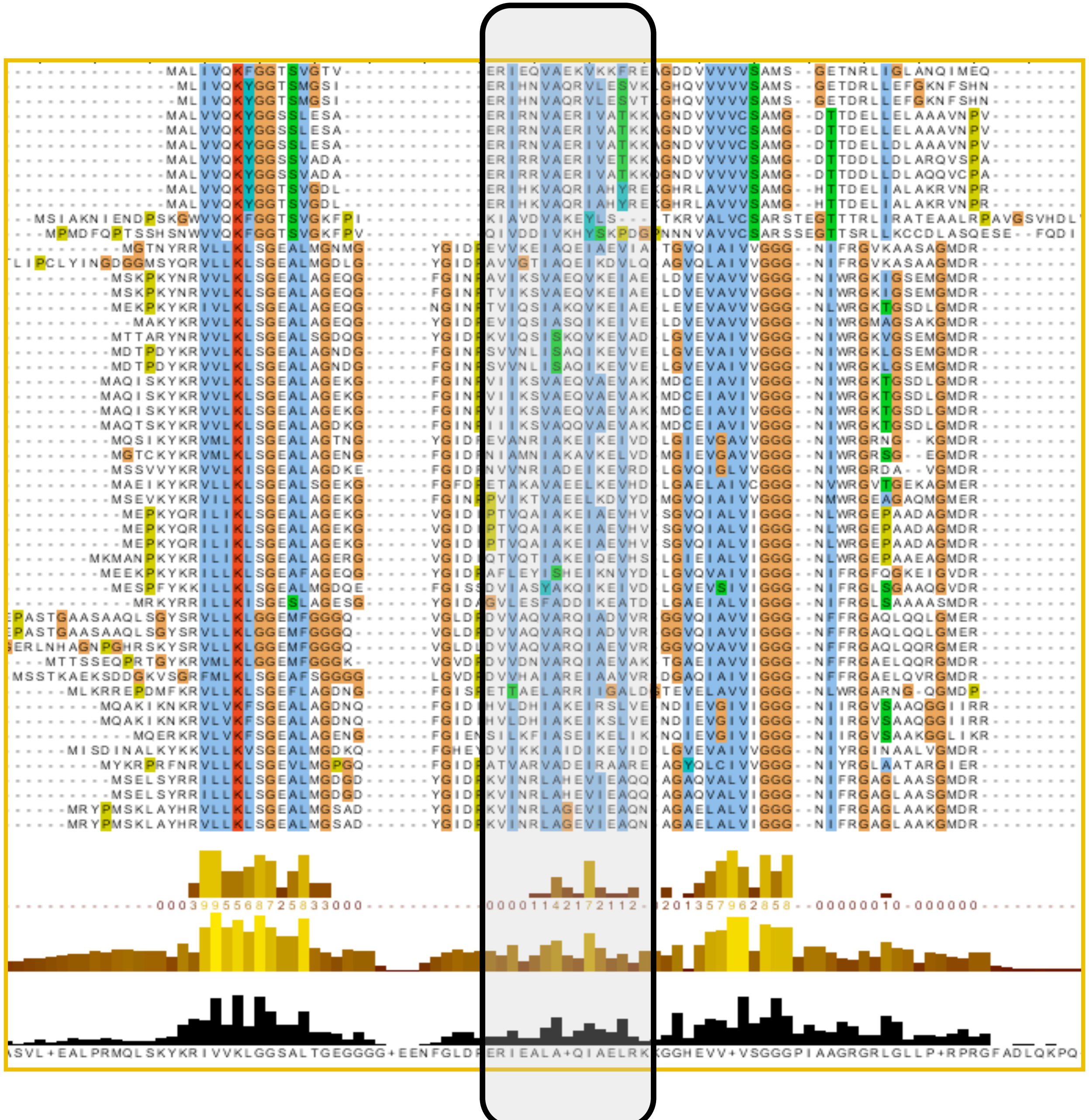


```
2086145   HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
33673     HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTL-SSKPVTITVQAPSSSPMGI
94190     HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAP-------
115776    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
115777    HLEFQEGETIVLRCHSWKDKPLVKV-FFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
115778    HLEFQEGETIVL-CHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQ---------
118288    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
307773    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
320425    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRS-PNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
320426    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRS-PNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
327261    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTL-SSKPVTITVQAPSSSPMGI
327262    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTL-SSKPVTITVQAPSSSPMGI
327263    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTL-SSKPVTITVQAPSSSPMGI
382746    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
382747    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
382748    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
382749    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
382750    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
382751    HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
508152    HLEFQEGETI-LRCHSWKDKPLVKVTFFQNGKS-KFS-----FSIPQANHSHSGDYHCTGNIGYTL-SSKPVTITVQ---------
717955    HLEFQEGETI-LRCHSWKDKPLVKVTFFQNGKS-KFSR----FSIPQANHSHSGDYHCTGNIGYTL-SSKPVTITVQ---------
969355    HLEF-EGETI-LRCHSWKDKPL-KVTFFQNG-SKKFS---PNFSIPQANHSHSGDYHCTGNIGYT-YSSKPVTITVQ--SSSP-GI
1293739   HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQ---------
1293740   HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGY-LYSSKPVTITVQ---------
1293741   HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTL-SSKPVTITVQ---------
1293742   HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYS-KPVTITVQ---------
2086144   HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
2086146   HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQ---------
33665     SLVFQEGEPIMLRCHSWRNQPLNKITFYQDRKSKIFSYQRTNFSIPRANLSHSGQYHCTAFIGKMLHSSQPVNITVQESSSSGPSS

Conservation
          4*4*1***8*1*3****997**1*92*9*944*46**11213****8**5****7****+5**7023*3***9****---------
Quality
Consensus
          HLEFQEGETIVLRCHSWKDKPLVKVTFFQNGKSKKFSRSDPNFSIPQANHSHSGDYHCTGNIGYTLYSSKPVTITVQAPSSSPMGI
```

A short run of hydrophobic amino acids (4 or 5 residues) suggests a buried β-strand…

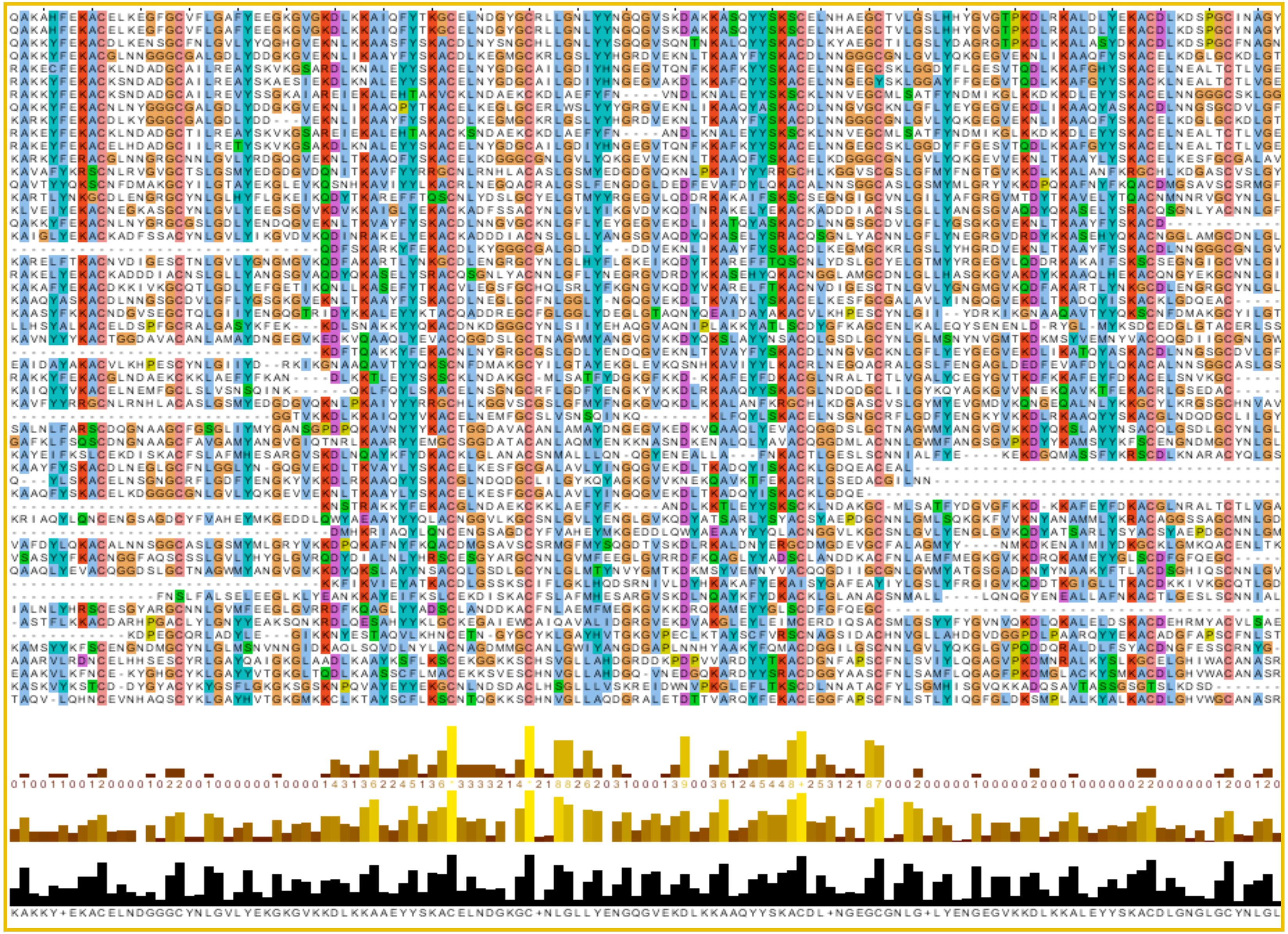


Pairs of conserved hydrophobic amino acids separated by pairs of unconserved or hydrophilic residues suggests an α-helix with one face packed in the protein core. Similarly, an i, i+3, i+4, i+7 pattern of conserved residues.”

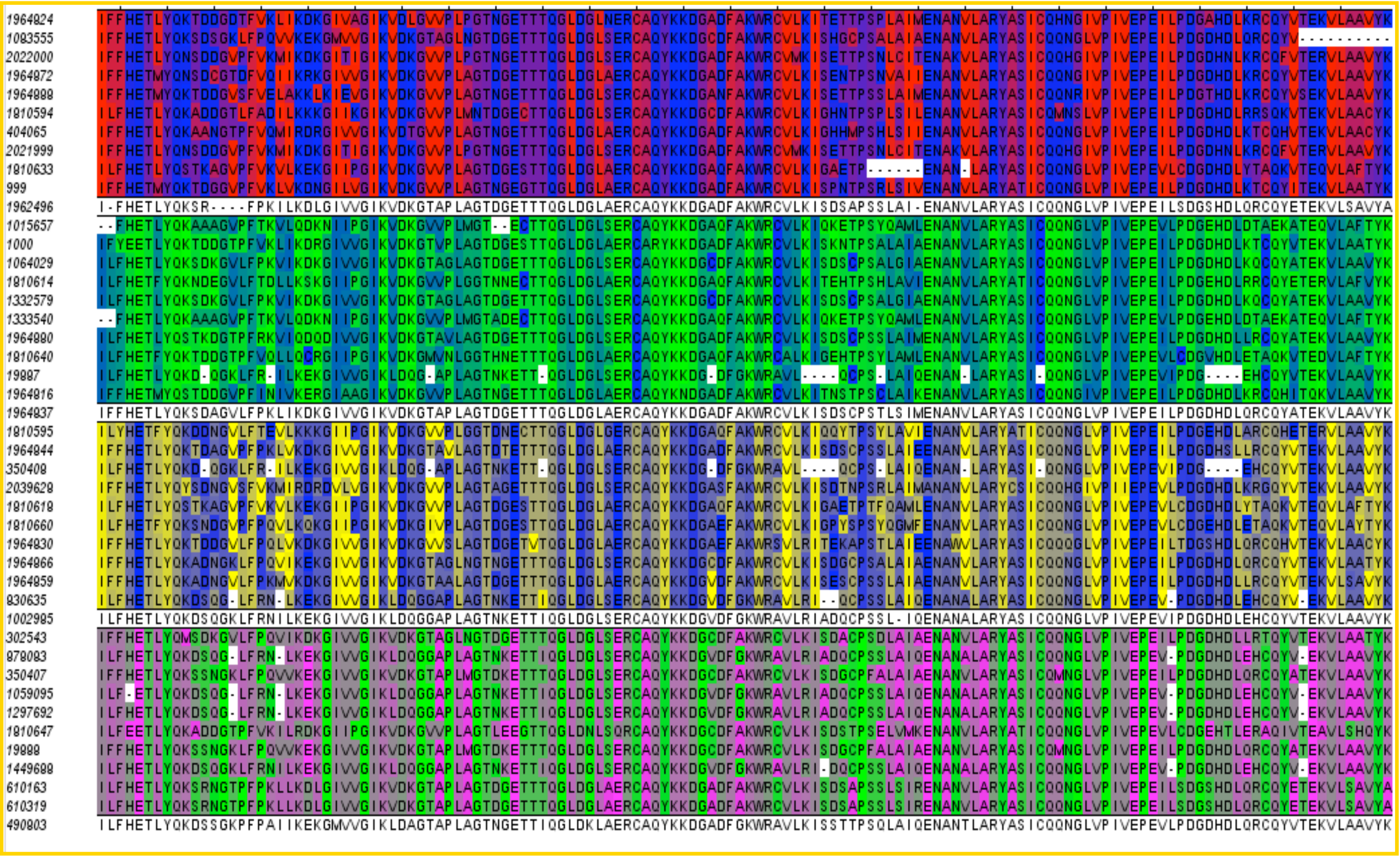


Cysteine is a rare amino acid, and is often used in disulphide bonds (pairs of conserved cysteines)

Charged residues (histidine, aspartate, glutamate, lysine, arginine) and other polar residues embedded in a conserved region indicate functional importance



Coloring your alignments



HYDROPHOBIC / POLAR

- 🟥 hydrophobic
- 🟦 polar

BURIED INDEX

- 🟦 buried
- 🟩 surface

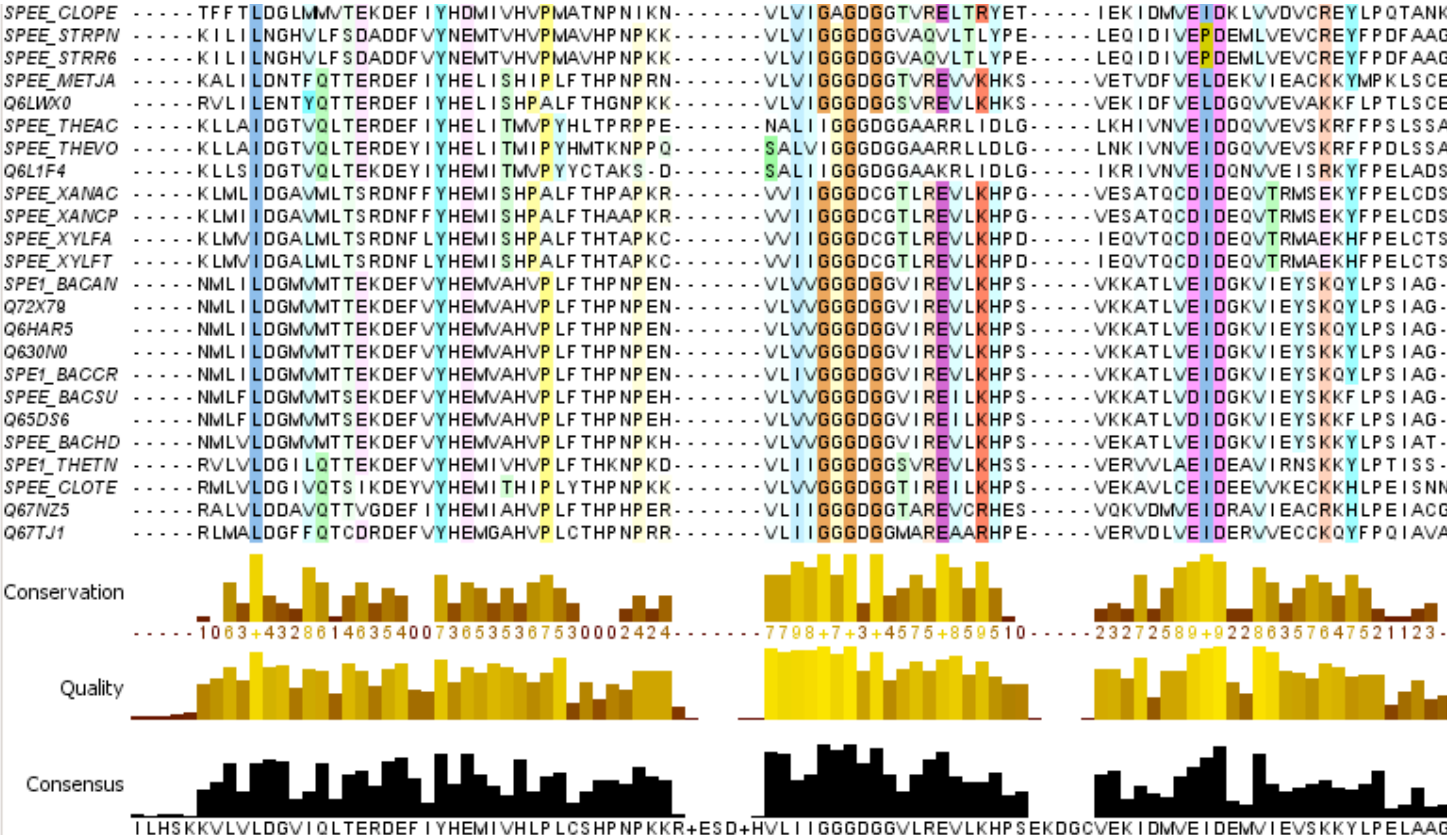
β-STRAND LIKELIHOOD

- 🟨 probable
- 🟦 unlikely

HELIX LIKELIHOOD

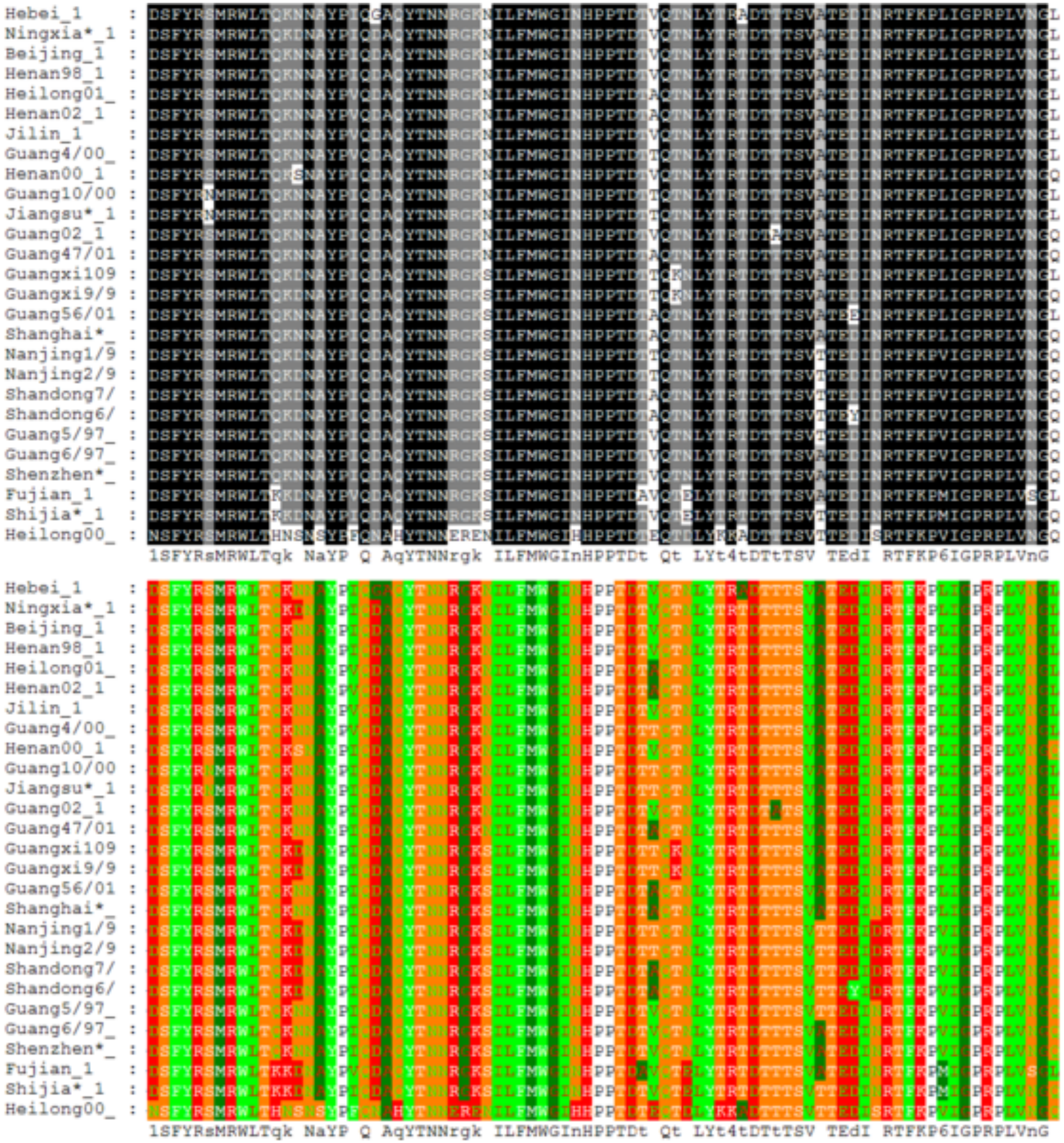
- 🟪 probable
- 🟩 unlikely

Coloring your alignment



```
SPEE_CLOPE   -----TFFTLDGLMMVTEKDEFIYHDMIVHVPMATNPNIKN-------VLVIGAGDGGTVRELTRYET-----IEKIDMVEIDKLVVDVCREYLPQTANK
SPEE_STRPN   -----KILILNGHVLFSDADDFVYNEMTVHVPMAVHPNPKK-------VLVIGGGDGGVAQVLTLYPE-----LEQIDIVEPDEMLVEVCREYFPDFAAG
SPEE_STRR6   -----KILILNGHVLFSDADDFVYNEMTVHVPMAVHPNPKK-------VLVIGGGDGGVAQVLTLYPE-----LEQIDIVEPDEMLVEVCREYFPDFAAG
SPEE_METJA   -----KALILDNTFQTTERDEFIYHELISHIPLFTHPNPRN-------VLVIGGGDGGTVREVVKHKS-----VETVDFVELDEKVIEACKKYMPKLSCE
Q6LWX0       -----RVLILENTYQTTERDEFIYHELISHPALFTHGNPKK-------VLVIGGGDGGSVREVLKHKS-----VEKIDFVELDGQVVEVAKKFLPTLSCE
SPEE_THEAC   -----KLLAIDGTVQLTERDEFIYHELITMVPYHLTPRPPE-------NALIIGGGDGGAARRLIDLG-----LKHIVNVEIDDQVVEVSKRFFPSLSSA
SPEE_THEVO   -----KLLAIDGTVQLTERDEYIYHELITMIPYHMTKNPPQ-------SALVIGGGDGGAARRLLDLG-----LNKIVNVEIDGQVVEVSKRFFPDLSSA
Q6L1F4       -----KLLSIDGTVQLTEKDEYIYHEMITMVPYYCTAKS-D-------SALIIGGGDGGAAKRLIDLG-----IKRIVNVEIDQNVVEISRKYFPELADS
SPEE_XANAC   -----KLMLIDGAVMLTSRDNFFYHEMISHPALFTHPAPKR-------VVIIGGGDCGTLREVLKHPG-----VESATQCDIDEQVTRMSEKYFPELCDS
SPEE_XANCP   -----KLMIIDGAVMLTSRDNFFYHEMISHPALFTHAAPKR-------VVIIGGGDCGTLREVLKHPG-----VESATQCDIDEQVTRMSEKYFPELCDS
SPEE_XYLFA   -----KLMVIDGALMLTSRDNFLYHEMISHPALFTHTAPKC-------VVIIGGGDCGTLREVLKHPD-----IEQVTQCDIDEQVTRMAEKHFPELCTS
SPEE_XYLFT   -----KLMVIDGALMLTSRDNFLYHEMISHPALFTHTAPKC-------VVIIGGGDCGTLREVLKHPD-----IEQVTQCDIDEQVTRMAEKHFPELCTS
SPE1_BACAN   -----NMLILDGMVMTTEKDEFVYHEMVAHVPLFTHPNPEN-------VLVVGGGDGGVIREVLKHPS-----VKKATLVEIDGKVIEYSKQYLPSIAG-
Q72X78       -----NMLILDGMVMTTEKDEFVYHEMVAHVPLFTHPNPEN-------VLVVGGGDGGVIREVLKHPS-----VKKATLVEIDGKVIEYSKQYLPSIAG-
Q6HAR5       -----NMLILDGMVMTTEKDEFVYHEMVAHVPLFTHPNPEN-------VLVVGGGDGGVIREVLKHPS-----VKKATLVEIDGKVIEYSKQYLPSIAG-
Q630N0       -----NMLILDGMVMTTEKDEFVYHEMVAHVPLFTHPNPEN-------VLVVGGGDGGVIREVLKHPS-----VKKATLVEIDGKVIEYSKKYLPSIAG-
SPE1_BACCR   -----NMLILDGMVMTTEKDEFVYHEMVAHVPLFTHPNPEN-------VLIVGGGDGGVIREVLKHPS-----VKKATLVEIDGKVIEYSKQYLPSIAG-
SPEE_BACSU   -----NMLFLDGMVMTSEKDEFVYHEMVAHVPLFTHPNPEH-------VLVVGGGDGGVIREILKHPS-----VKKATLVDIDGKVIEYSKKFLPSIAG-
Q65DS6       -----NMLFLDGMVMTSEKDEFVYHEMVAHVPLFTHPNPEH-------VLVVGGGDGGVIREILKHPS-----VKKATLVDIDGKVIEYSKKFLPSIAG-
SPEE_BACHD   -----NMLVLDGMVMTTEKDEFVYHEMVAHVPLFTHPNPKH-------VLVVGGGDGGVIREVLKHPS-----VEKATLVEIDGKVIEYSKKYLPSIAT-
SPE1_THETN   -----RVLVLDGILQTTEKDEFVYHEMIVHVPLFTHKNPKD-------VLIIGGGDGGSVREVLKHSS-----VERVVLAEIDEAVIRNSKKYLPTISS-
SPEE_CLOTE   -----RMLVLDGIVQTSIKDEYVYHEMITHIPLYTHPNPKK-------VLVVGGGDGGTIREILKHPS-----VEKAVLCEIDEEVVKECKKHLPEISNN
Q67NZ5       -----RALVLDDAVQTTVGDEFIYHEMIAHVPLFTHPHPER-------VLIIGGGDGGTAREVCRHES-----VQKVDMVEIDRAVIEACRKHLPEIACG
Q67TJ1       -----RLMALDGFFQTCDRDEFVYHEMGAHVPLCTHPNPRR-------VLIIGGGDGGMAREAARHPE-----VERVDLVEIDERVVECCKQYFPQIAVA

Conservation
             -----1063+43286146354007365353675300002424-------7798+7+3+4575+859510-----23272589+922863576475211 23-
Quality

Consensus
             ILHSKKVLVLDGVIQLTERDEFIYHEMIVHLPLCSHPNPKKR+ESD+HVLIIGGGDGGVLREVLKHPSEKDGCVEKIDMVEIDEMVIEVSKKYLPELAAG
```

Alignment of 27 avian influenza hemagglutinin protein sequences colored by residue conservation (top) and residue properties (bottom)



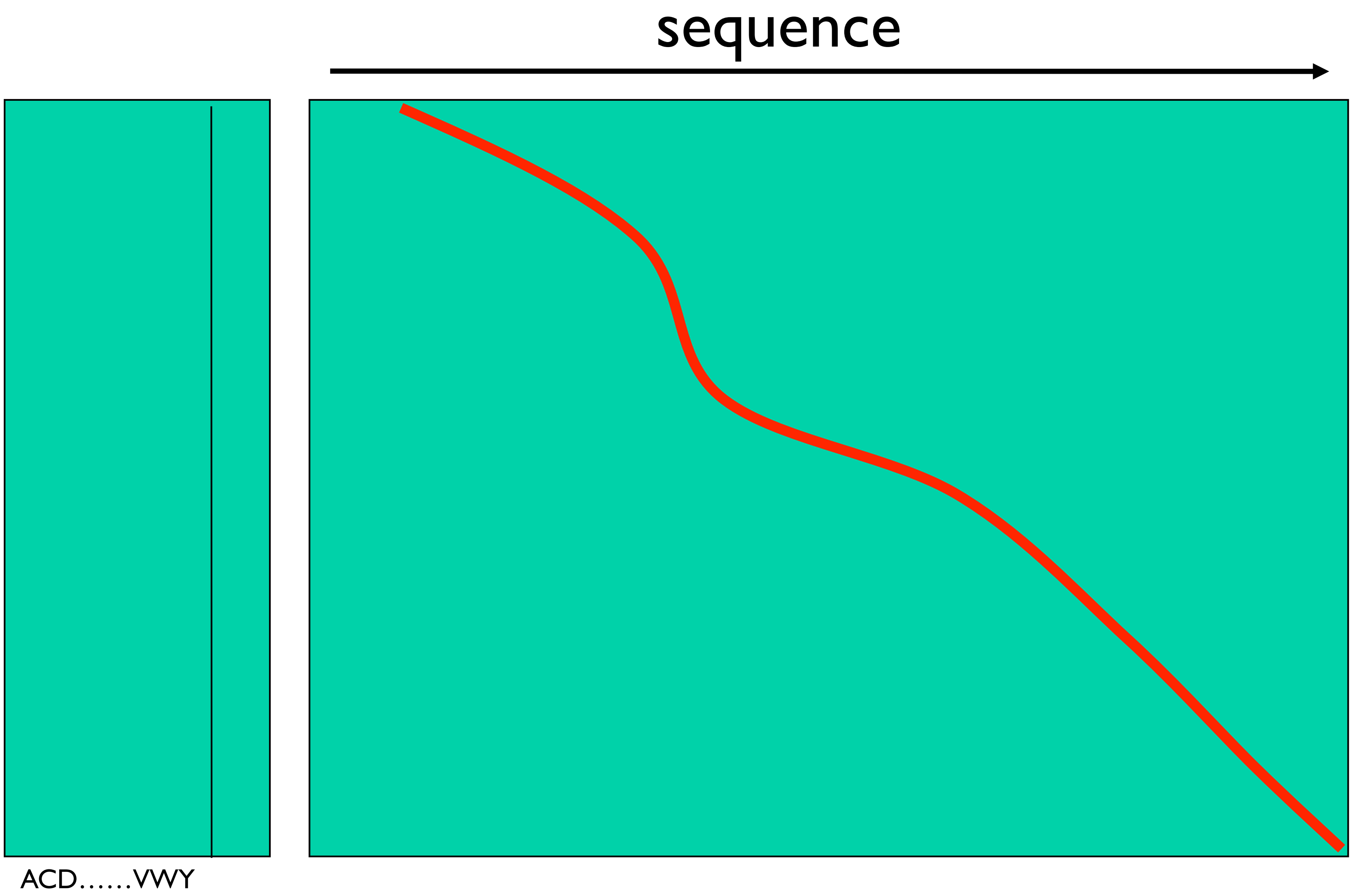
PSSMs

- How to use evolution to detect (more) homologs ?
  - Iterated sequence search
  - Patterns
  - Profiles (PSSMs, PSIBLAST)
  - HMMs

Searching with PSSMs

- PSSMs are profiles
  - without gap information
  - Without substitution table information
- Dynamic programming can be used
  - Identical algorithm as for Smith-Waterman
- Profile alignment vs. single sequence
  - Better alignments
  - Better detection

Profile-sequence alignment



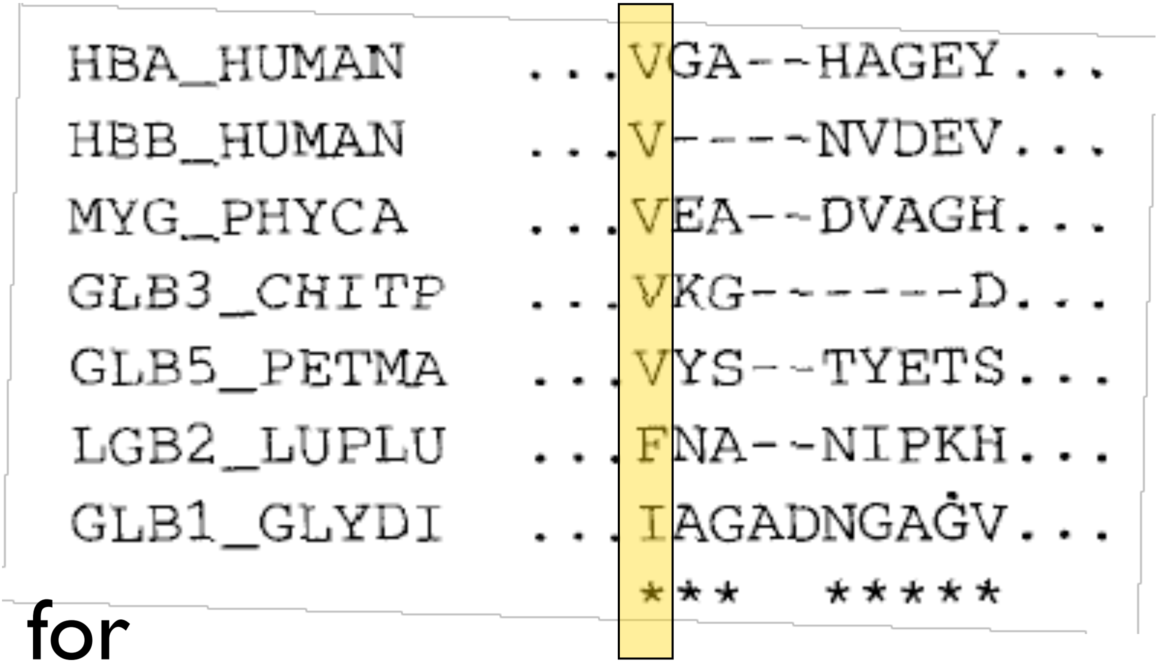
Average profiles

- Gribskov, McLachlan and Eisenberg 1987
  - No underlying probabilistic model, but rather assigned position specific scores for each match state and gap penalty
  - The score for each consensus position is set to the average of the standard substitution scores from all the residues in the corresponding multiple sequence alignment column
  - Gap costs

The “average” profile method

- Score for each residue is average score for that residue with all sequence in MSA
- Average score over all replacements:

Non-probabilistic or



```
HBA_HUMAN     ...VGA--HAGEY...
HBB_HUMAN     ...V----NVDEV...
MYG_PHYCA     ...VEA--DVAGH...
GLB3_CHITP    ...VKG------D...
GLB5_PETMA    ...VYS--TYETS...
LGB2_LUPLU    ...FNA--NIPKH...
GLB1_GLYDI    ...IAGADNGAGV...
                 ***  *****
```

The score for residue ‘a’ in column 1

$$\frac{5}{7}s(\mathrm{V},a)+\frac{1}{7}s(\mathrm{F},a)+\frac{1}{7}s(\mathrm{I},a)$$

s(a,b) : standard substitution matrix

Average profiles – example

- One position contains
  - 50% ILE ; 30% THR ; 20% VAL
- Calculate the score for ILE in this position
- Use the PAM250 Matrix
  - I-I=5
  - I-T=0
  - I-V=4
- Calculate
  - 0.5*5+0.3*0+0.2*4=3.3
- Integer

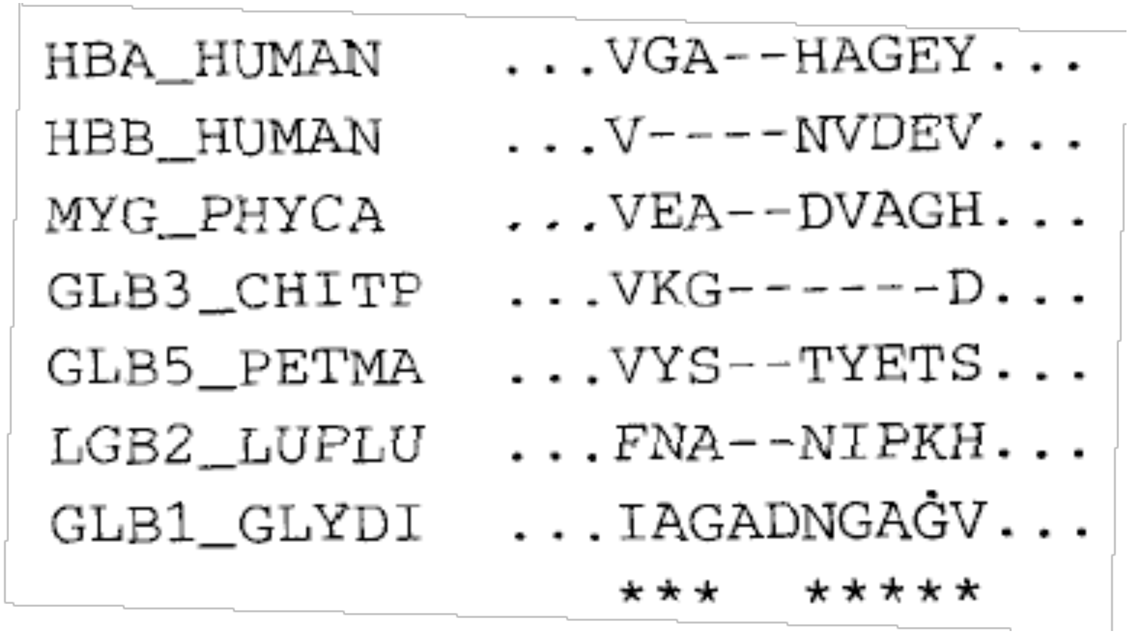
Average Profiles

- They also set gap penalties for each column using a heuristic equation that decrease the cost of a gap according to the length of the longest gap observed in the multiple alignment spanning the column

Problem With Average profiles

- If we had an alignment with 100 sequences, all with a cysteine (C), at some position, the probability distribution for that column for an “average” profile would be exactly the same as would be derived from a single sequence
- Doesn’t correspond to our expectation that the likelihood of a cysteine should go up as we see more confirming examples

Similar Problem With Gaps



```
HBA_HUMAN     ...VGA--HAGEY...
HBB_HUMAN     ...V----NVDEV...
MYG_PHYCA     ...VEA--DVAGH...
GLB3_CHITP    ...VKG------D...
GLB5_PETMA    ...VYS--TYETS...
LGB2_LUPLU    ...FNA--NIPKH...
GLB1_GLYDI    ...IAGADNGAĠV...
                 ***  *****
```

Scores for a deletion in columns 2 and 4 would be set to the same value

More reasonable to set the probability of a new gap opening to be higher in column 4

The amino acid frequency can be used for scoring

$$f_{u,b} = \frac{n_{u,b}}{N_{seq}} (EQ6.1)$$

$$m_{u,a} = \sum_{b} f_{u,b} s_{a,b} (EQ6.2)$$

Higher scoring for more conserved positions

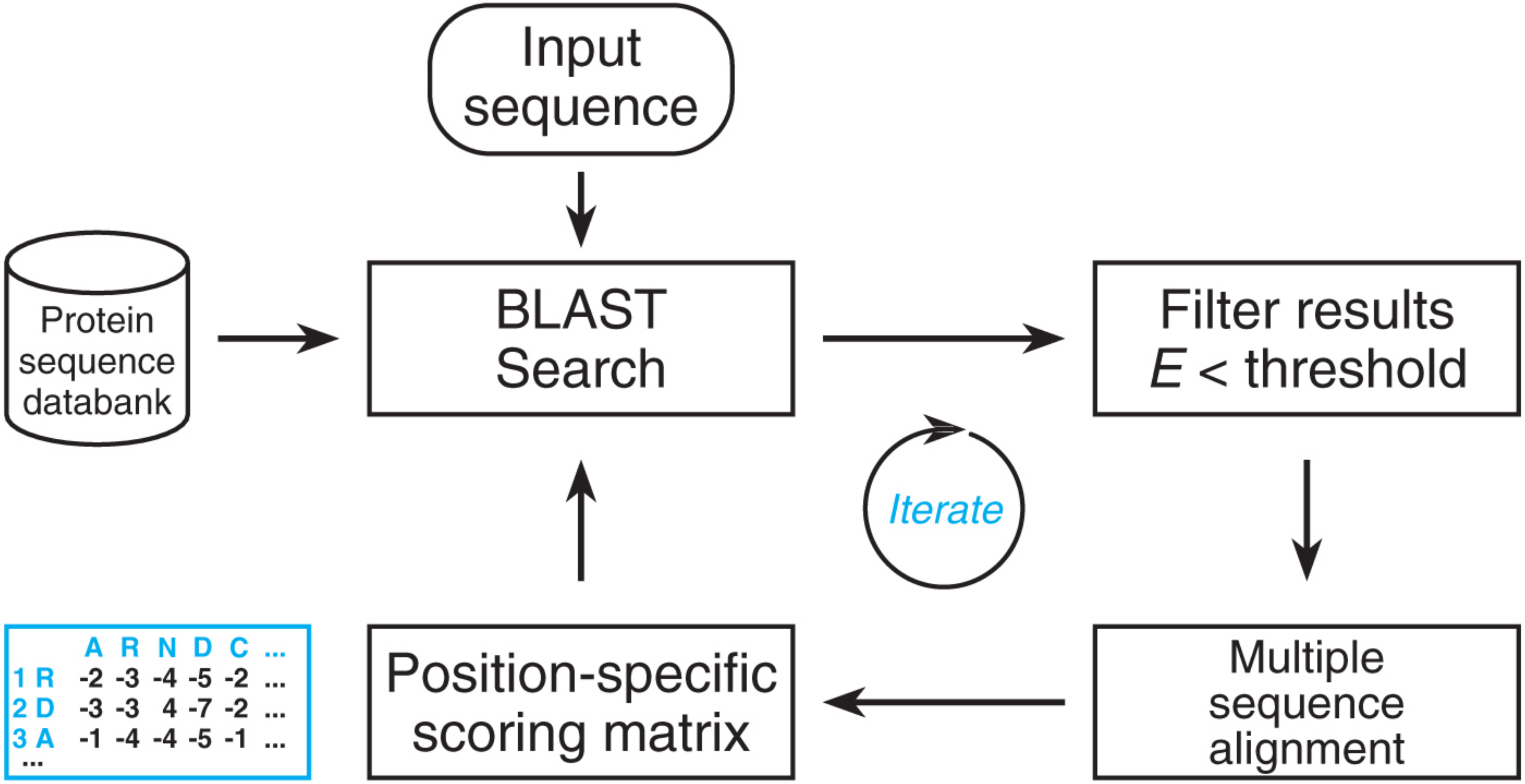
$$m_{u,a} = \sum_{b} \frac{\ln(1 - f'_{u,b})}{\ln(1/(N_{seq} + 1)} s_{a,b} (EQ6.3)$$

$$m_{u,a} = \log \frac{q_{u,a}}{p_a} (EQ6.4)$$

PSI-BLAST algorithm

- Input a single protein sequence and compares it to a protein database, using BLAST
- The program constructs a multiple alignment, and then a profile,
- The profile is compared to the protein database, again seeking local alignments.
- PSI-BLAST estimates the statistical significance of the local alignments found.
- Finally, PSI-BLAST iterates, by returning to step (2), an arbitrary number of times or until convergence.

Psiblast



Input sequence

Protein sequence databank

BLAST Search

Filter results *E* < threshold

Iterate

Multiple sequence alignment

Position-specific scoring matrix

	A	R	N	D	C	...
1 R	-2	-3	-4	-5	-2	...
2 D	-3	-3	4	-7	-2	...
3 A	-1	-4	-4	-5	-1	...
...						

PSI-BLAST

- Advantages
  - Fast (40 times faster than DP)
  - Significant better than DP
  - Good E-value estimates
- Disadvantages
  - Not optimal alignments

PSI-BLAST

- Important parameters
  - E-value cutoff
  - Number of iterations
  - Low complexity sequence filtering

PSI-BLAST in a nutshell

- With a protein sequence as query, use BLAST to search a protein sequence database.
- Collapse significant local alignments (those with *E*-value less than or equal to a set threshold *h*) into a multiple alignment, using the residue of the query sequence as alignment-column placeholders.
- Abstract a position-specific score matrix from the multiple alignment.
- Search the database with the score matrix as query.
- Iterate a fixed number of times, or until convergence.

PSI-BLAST live example

Sequence;

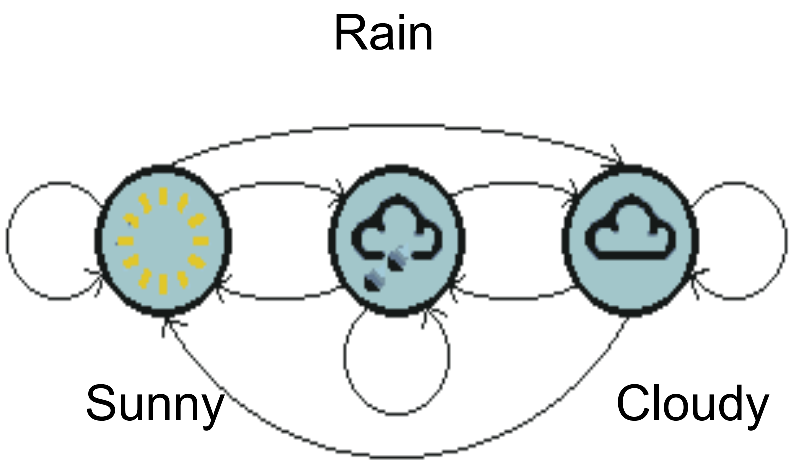
SISSRVKSVLLLGLQNAELAQKVGTTQQSIEQLENGKTRPRFLPELASAILGVSVDWLLNGT

Server:

http://www.ncbi.nlm.nih.gov/blast/

Run against Swissprot (faster)

Markov Chains



weather today

$$
\begin{array}{cc|ccc}
 & & \text{Sun} & \text{Cloud} & \text{Rain} \\
\hline
 & \text{Sun} & 0.5 & 0.25 & 0.25 \\
\text{weather yesterday} & \text{Cloud} & 0.375 & 0.125 & 0.375 \\
 & \text{Rain} & 0.125 & 0.625 & 0.375
\end{array}
$$

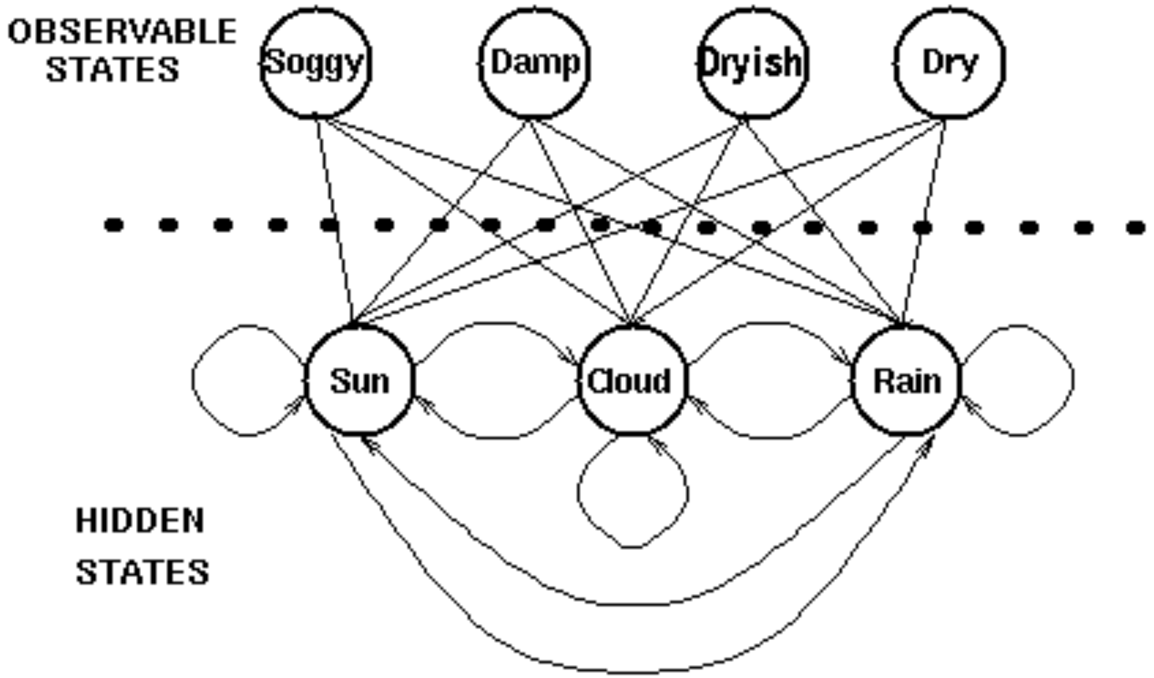
States : Three states - sunny, cloudy, rainy.

State transition matrix : The probability of the weather given the previous day's weather.

$$
\begin{array}{ccc}
\text{Sun} & \text{Cloud} & \text{Rain} \\
1.0 & 0.0 & 0.0
\end{array}
$$

Initial Distribution : Defining the probability of the system being in each of the states at time 0.

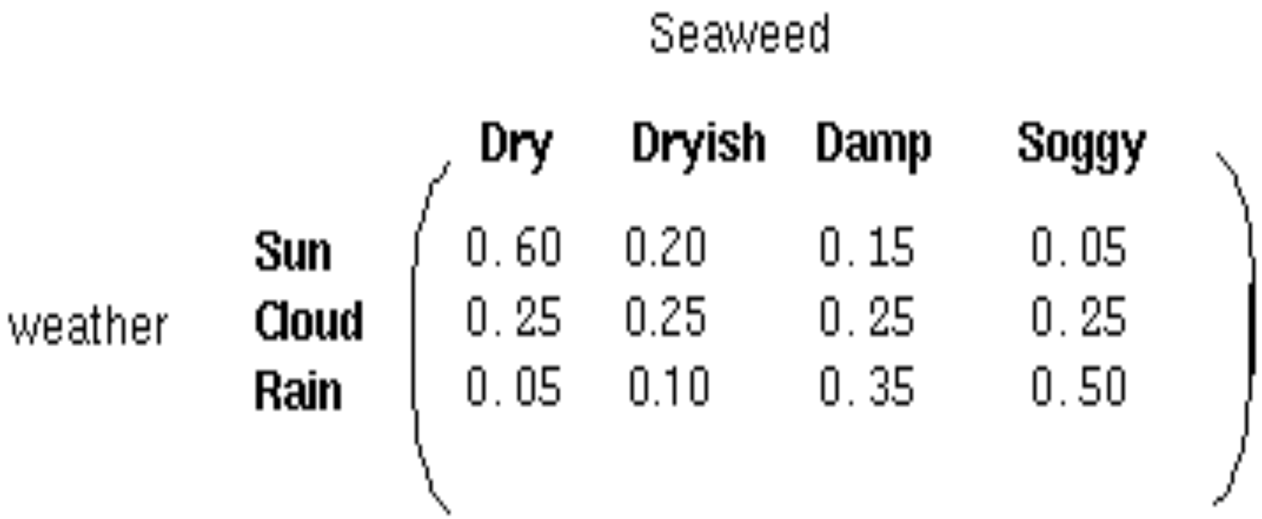
Hidden Markov Models



Hidden states : the (TRUE) states of a system that may be described by a Markov process (e.g., the weather).

Observable states : the states of the process that are `visible' (e.g., seaweed dampness).

Components of HMM



Seaweed

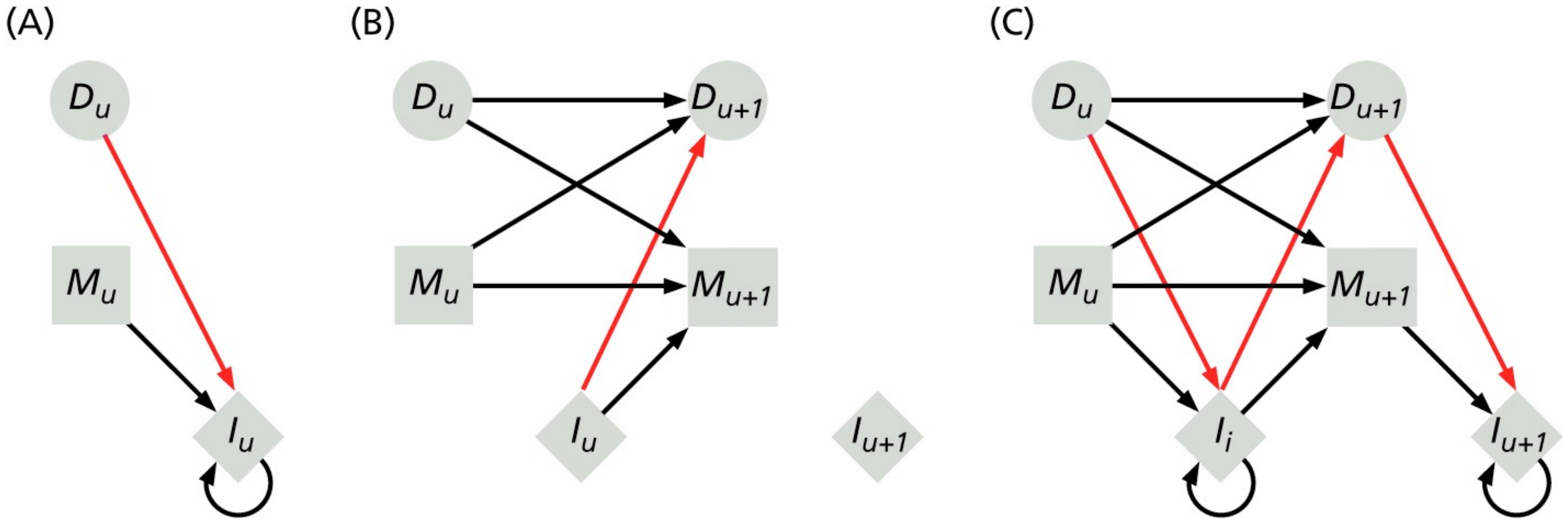
$$\text{weather}\quad\begin{array}{c|cccc} & \text{Dry} & \text{Dryish} & \text{Damp} & \text{Soggy} \\ \hline \text{Sun} & 0.60 & 0.20 & 0.15 & 0.05 \\ \text{Cloud} & 0.25 & 0.25 & 0.25 & 0.25 \\ \text{Rain} & 0.05 & 0.10 & 0.35 & 0.50 \end{array}$$

Output matrix : containing the probability of observing a particular observable state given that the hidden model is in a particular hidden state.

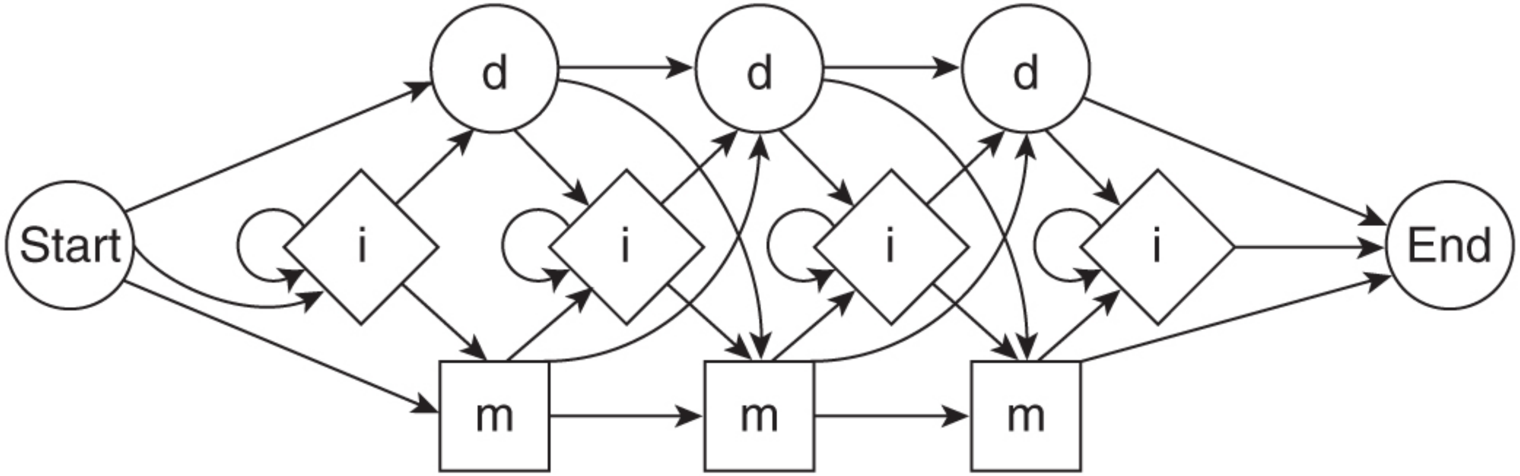
Initial Distribution : contains the probability of the (hidden) model being in a particular hidden state at time t = 1.

State transition matrix : holding the probability of a hidden state given the previous hidden state.

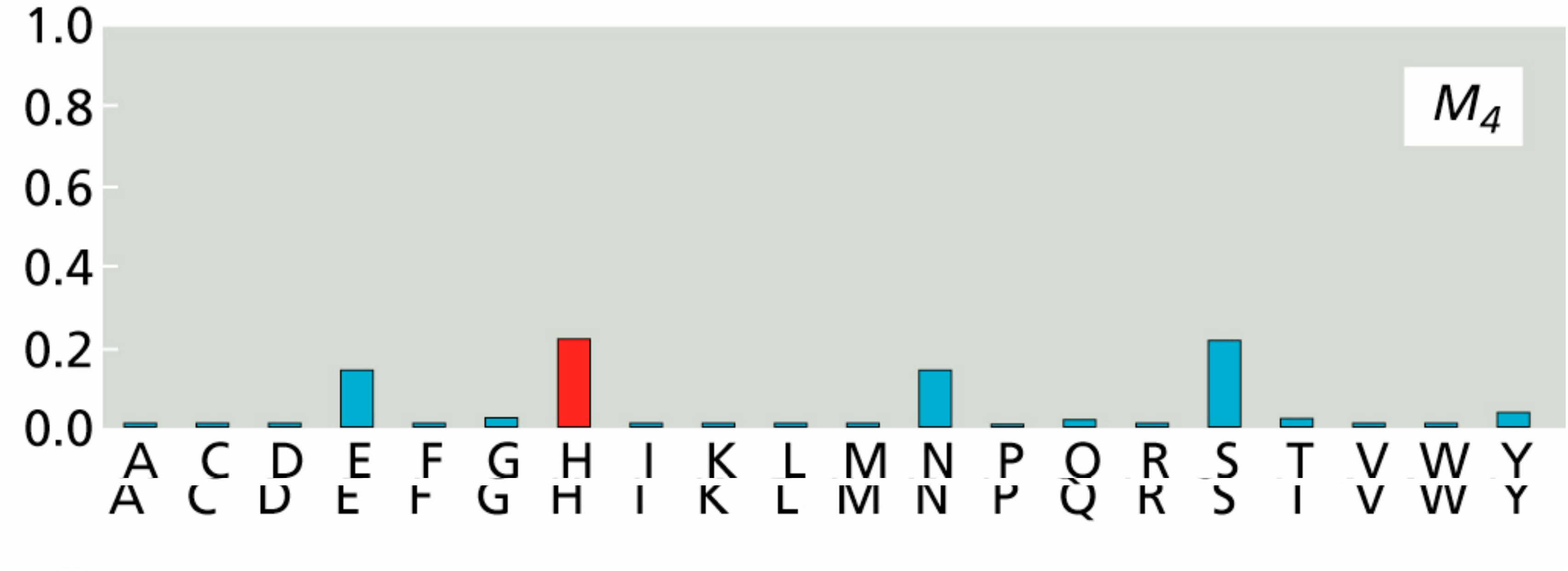
Profile HMMs



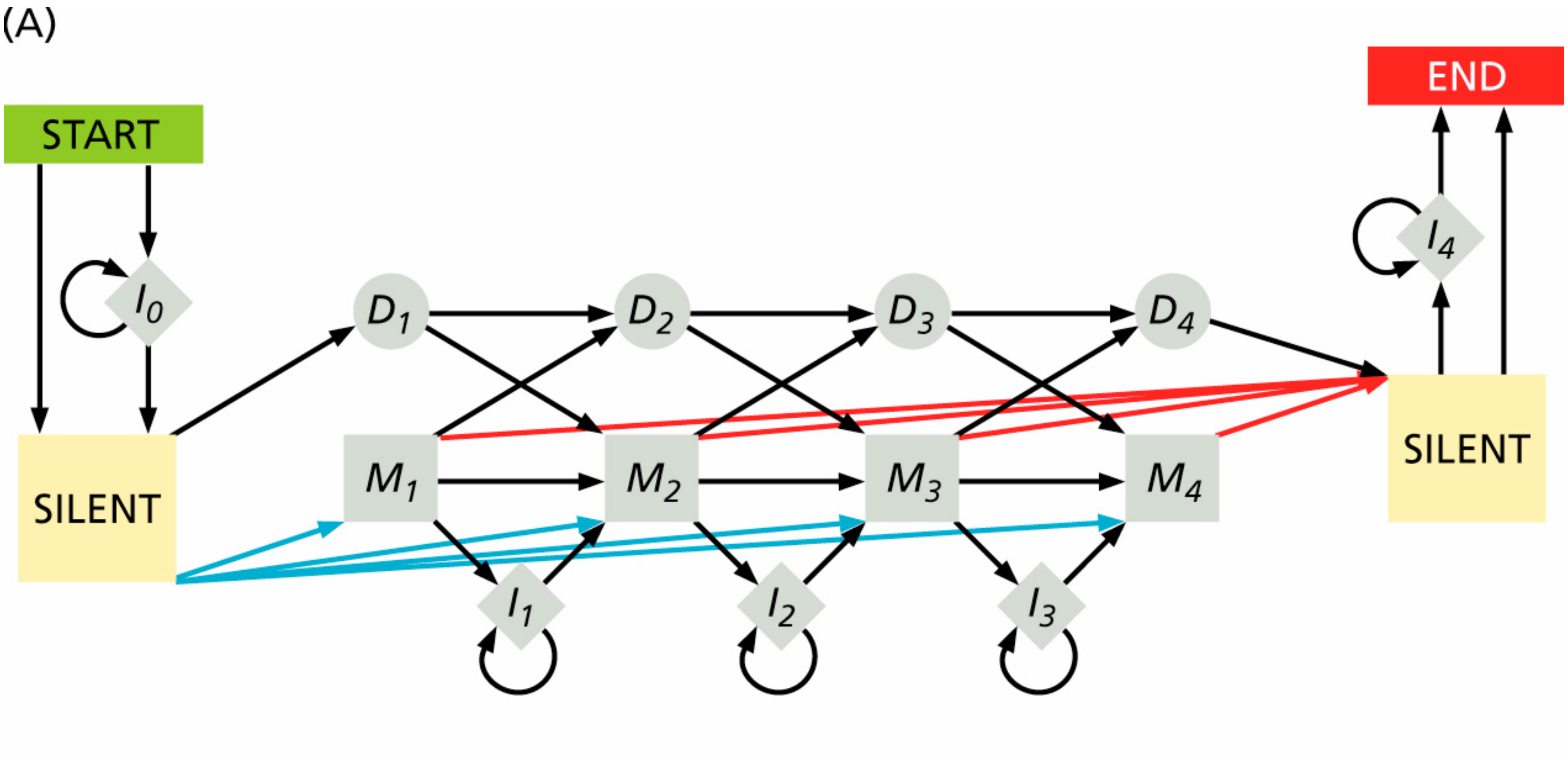
A profile HMM



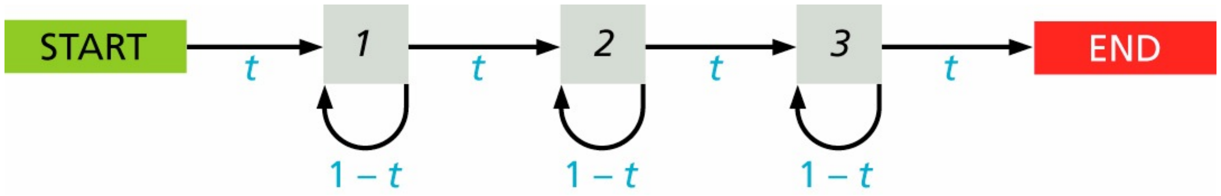
Emission probabilities

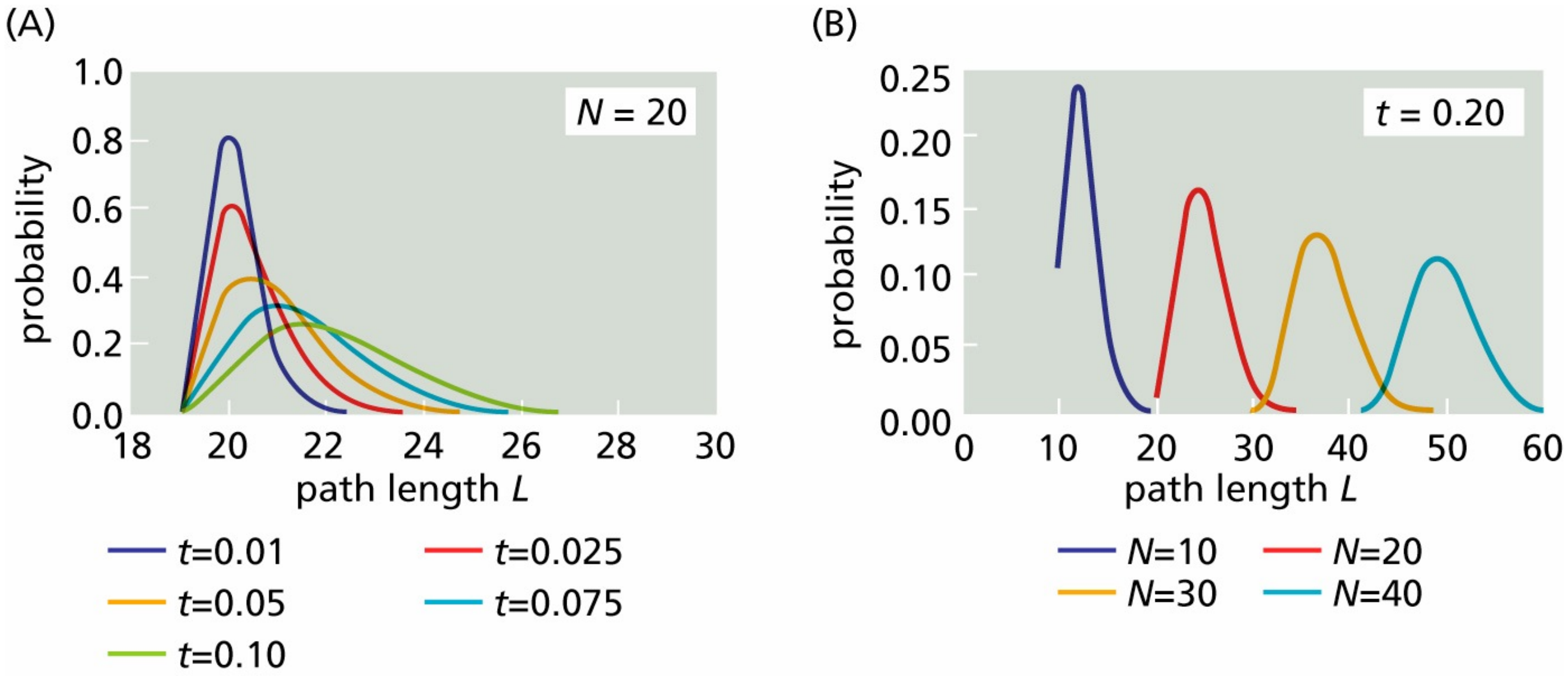


A local HMM



Length dependency of HMMs





Scoring an HMM

- The most probable path (Viterbi)
- Scoring for all paths (forward/backward)

Training an HMM

- Using unaligned sequences
- Baum-Welch expectations maximization
  - Estimating the number each emission and transition is used
- Starting with rough estimates

Profile HMMs: Effectiveness

- Advantages:
  - Expressive profiling method
  - Transparent method: You can view and interpret the model produced
  - Very effective at detecting remote homologs
- Disadvantages:
  - Slow – full search on a database of 400,000 sequences can take 15 hours
  - Have to avoid over-fitting and locally optimal models