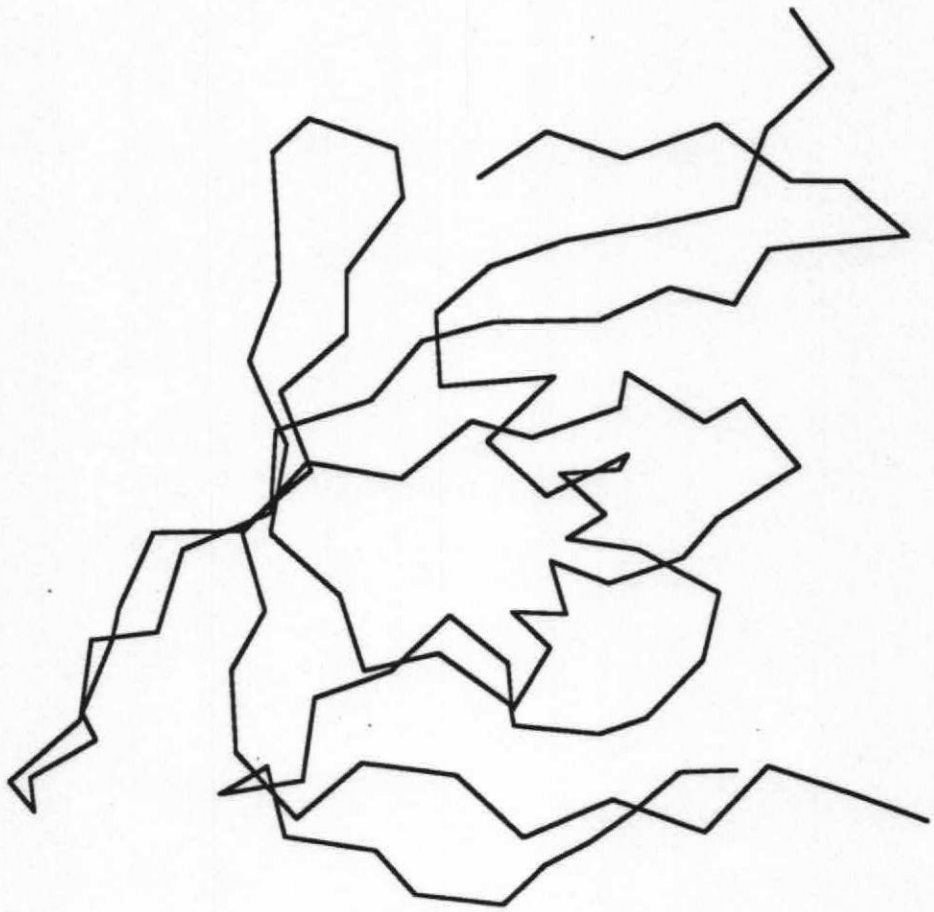
PROTEINS

SEQUENCE • STRUCTURE

FUNCTION



The biological activity of a protein can be understood in terms of its 3-dimensional structure

Structures

- Experimental determination is difficult
- 400 are known

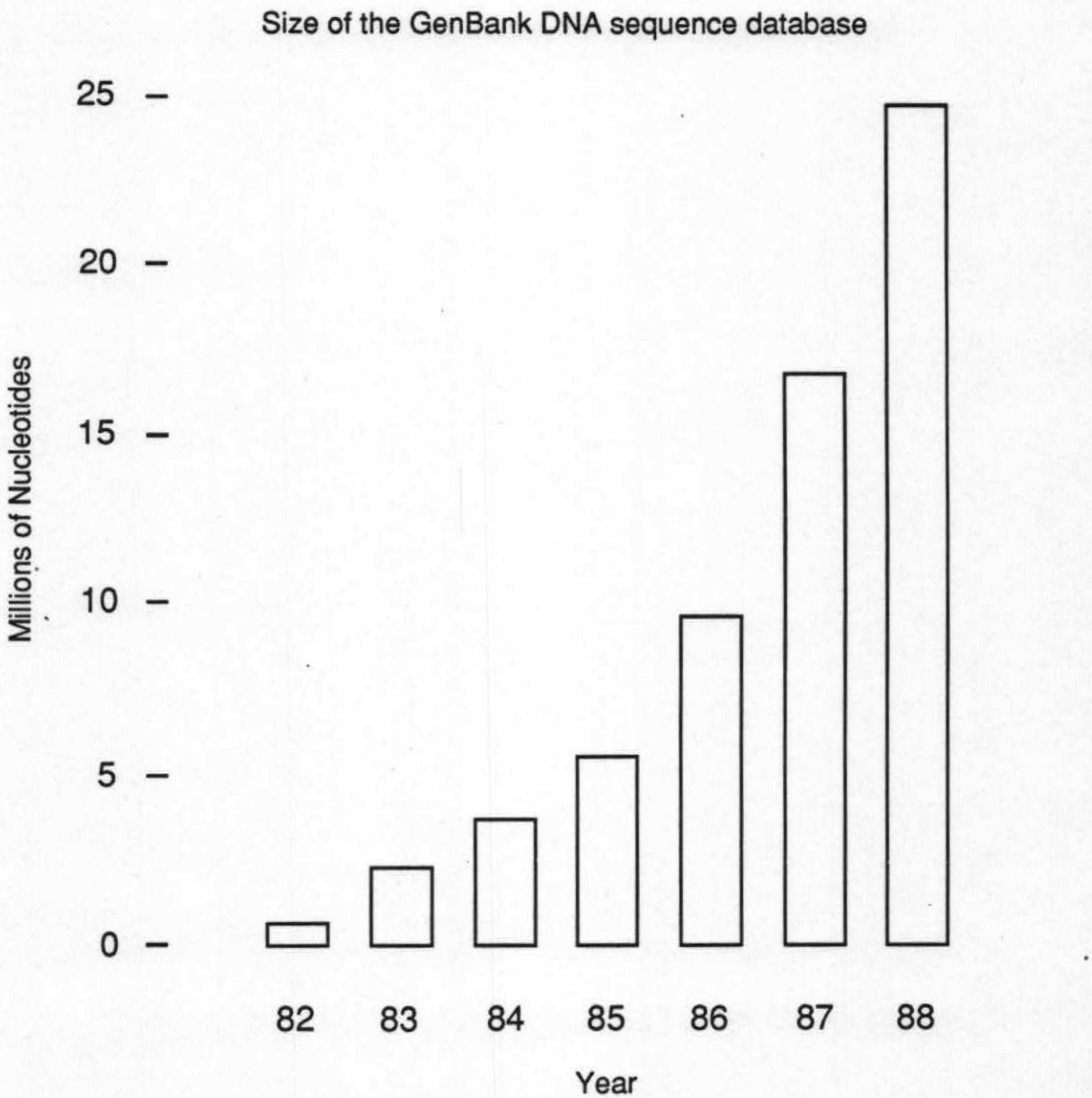
Sequences

- Determination straightforward
- 10,000 are known
- Contain all the information required to direct the folding of the protein

The Major Goal

Understanding the rules of protein folding would allow us to predict structure from sequence

The rate at which sequences are being determined is growing rapidly



Finding Similarities in Sequences

Motivation

- Conservation of a region of sequence between proteins implies functional importance

Method

- Dynamic programming is a sensitive method to identify conserved sequences
- With the CM we can apply the method to ALL pairs of proteins - over 50 million comparisons

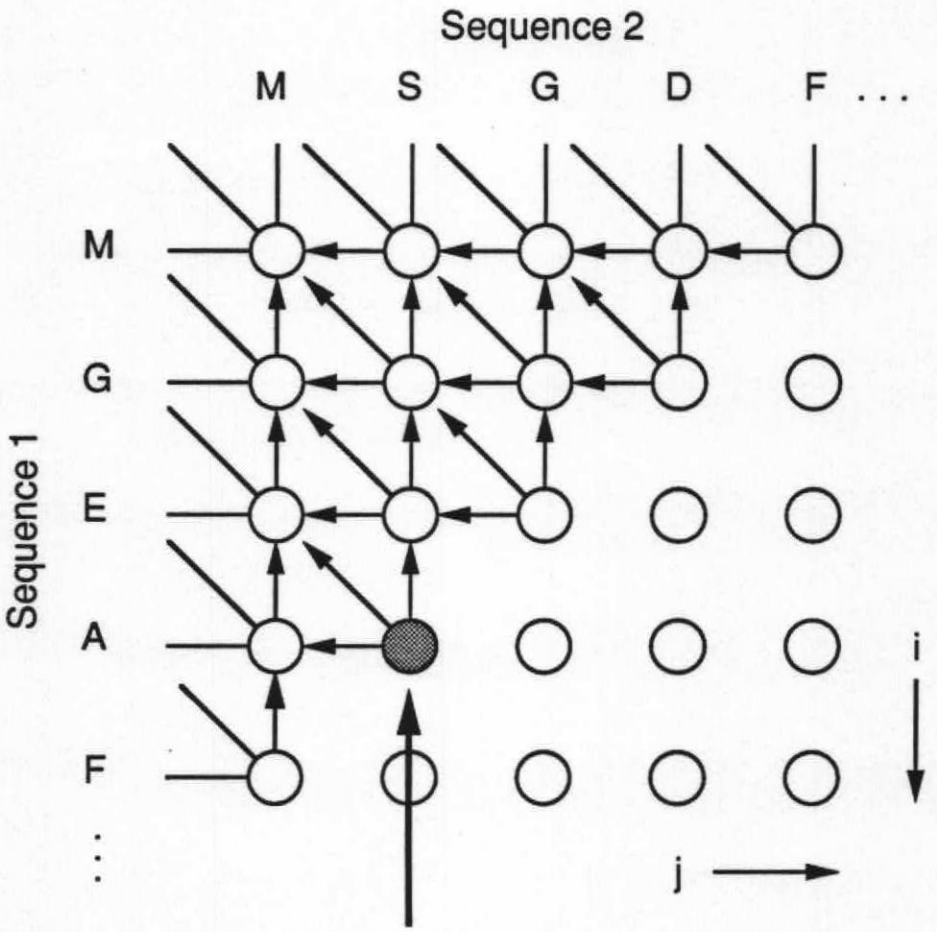
Application

- Families of related proteins can be identified
- Given the structure of one member of a family, we can reliably predict the structure of closely related proteins

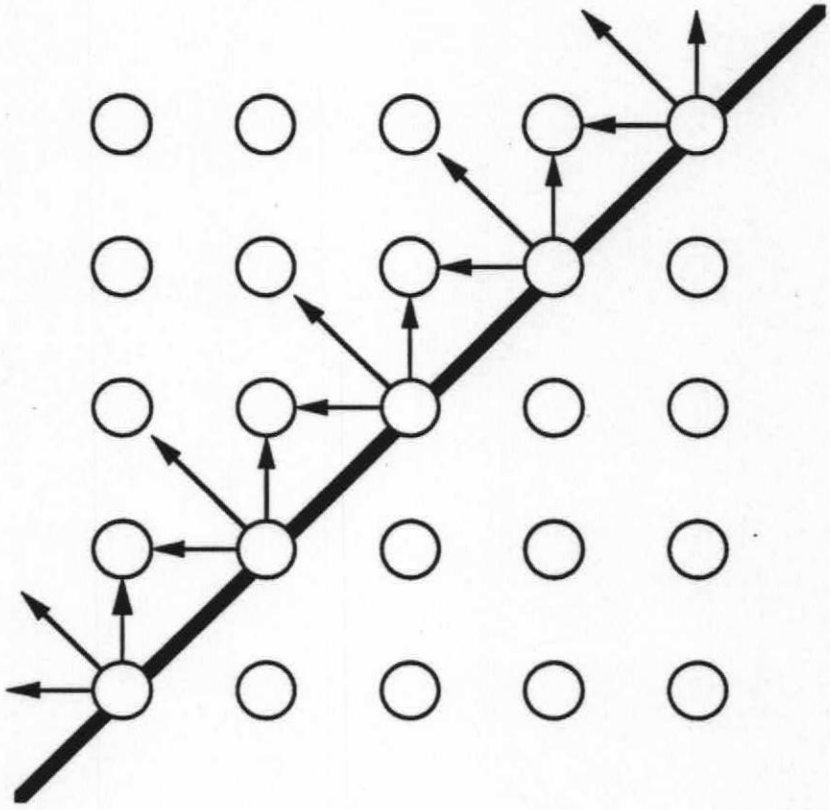
The recurrence relation

$$H_{i,j} = \max \begin{cases} 0 \\ H_{i-1,j-1} + s(a_i, b_j) \\ H_{i-1,j} + w \\ H_{i,j-1} + w \end{cases}$$

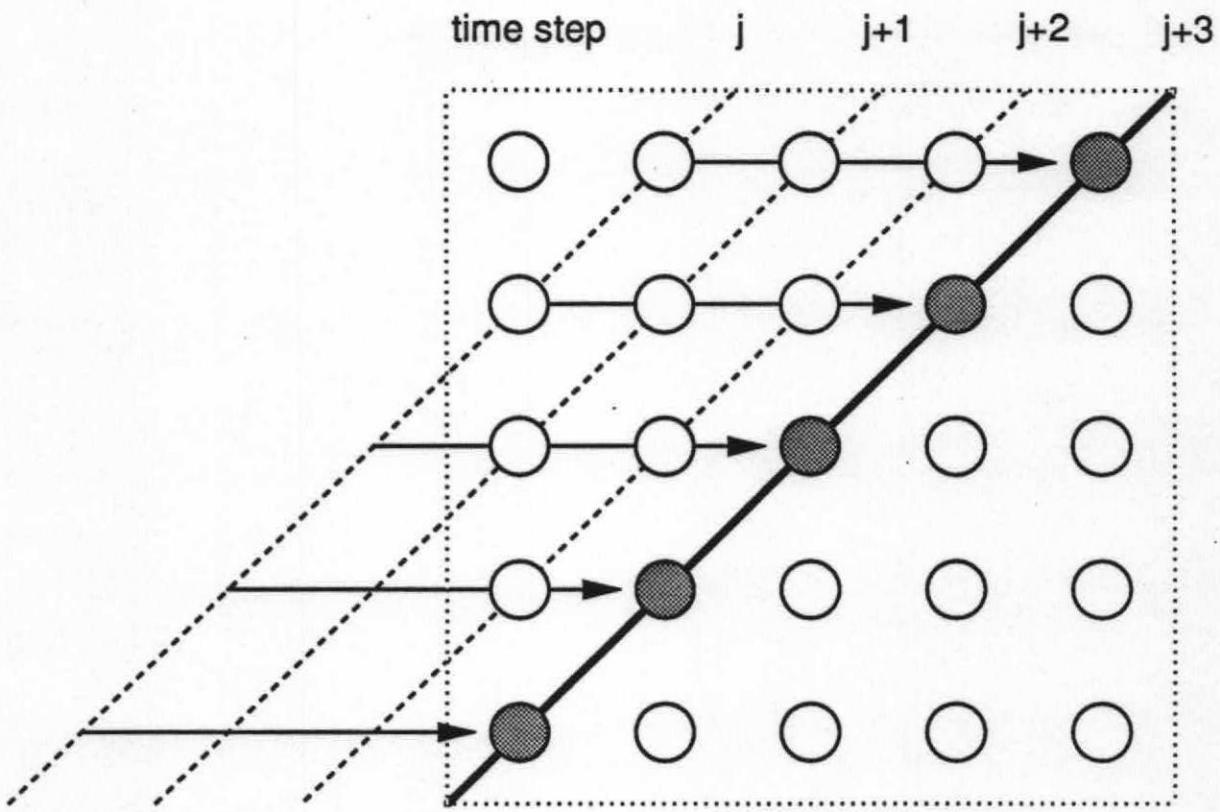
Calculation of the scoring matrix



Element (i,j) is dependent on the values of elements (i-1,j-1), (i,j-1) and (i-1,j)

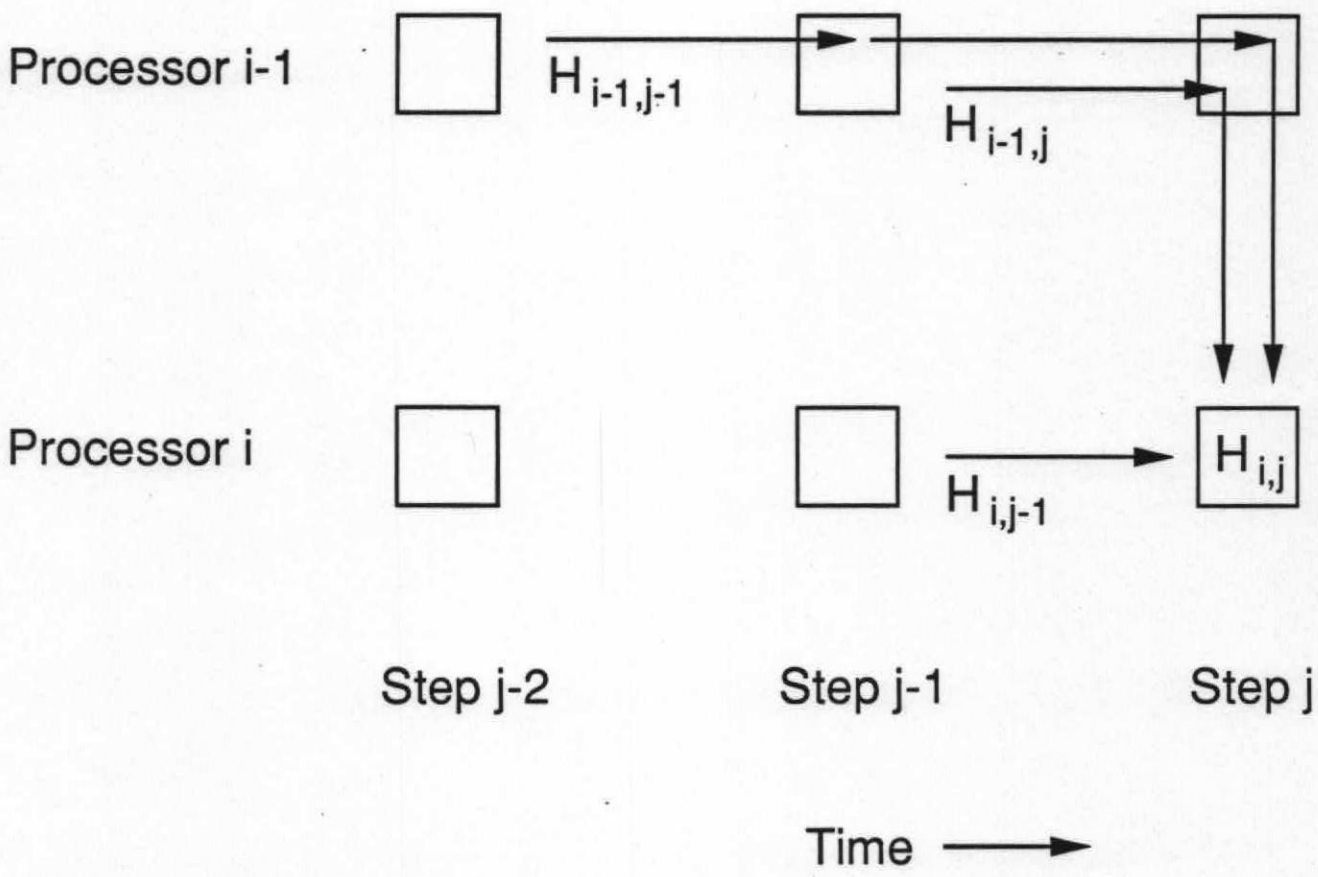


All elements on the same antidiagonal
can be computed in parallel



On each iteration the array of processors computes an entire antidiagonal; the matrix is traversed from left to right in successive iterations

The flow of information necessary to compute $H_{i,j}$



Performance

- Attained speed of 85 million matrix entries per second

- Peak of 100 million matrix entries per second

CM primitives used include :

- scans
- 1-d news transfers
- shared arrays
- indirect addressing

Sequence Patterns

Dispersed patterns of conserved amino acids often represent important functional sites

- A binding site for ATP

  G - G - - G K

- A Serine Protease active site

  DTSG

- A recognition site for DNA

  C - - - F - - - - - L - - H - - - - - C

These can be used to search the database for other examples

Multiple Sequence Alignment

- Instead of comparing pairs of sequences we can compare alignments

- In pairwise comparison all positions are equal. Conserved positions in a multiple alignment receive additional weight

- Pairwise comparison is not sensitive to find certain relationships that we know exist

- By comparing alignments against the database we hope to identify new structural and functional domains

AZURIN

```
1:  AECSVDIAGTDQMQFDKKAIEVSKSCKQFTVNLKHTGKLPRNVMGHNWVLTKTADMQAVEKDGIAAGLDNQYLKAGDTRVLAHTKVLGGGESDSVTFDVAKLAAGDDYTFFCSFPGHGALMKGTLKLVD
2:  AECSVDIAGNDQMQFDKKEITVSKSCKQFTVNLKHPGKLAKNVMGHNWVLTKQADMQGAVNDGMAAGLDNNYVKKDDARVIAHTKVIGGGETDSVTFDVSKLAAGEDYAYFCSFPGHFALMKGVLKLVD
3:  AQCEATIESNDAMQYNLKEMVVDKSCKQFTVHLKHVGKMAKAVMGHNWVLTKEADKEGVATDGMNAGLAQDYVKAGDTRVIAHTKVIGGGESDSVTFDVSKLTPGEAYAYFCSFPGHWAMMKGTLKLSN
4:  A-CDVSIEGNDSMQFNTKSIVVDKTCKEFTINLKHTGKLPKAAMGHNVVVSKKSDESAVATDGMKAGLNNDYVKAGDERVIAHTSVIGGGETDSVTFDVSKLKEGEDYAFFCSFPGHWSIMKGTIELGS
5:  AECKTTIDSTDQMSFNTKAIEIDKACKTFTVELTHSGSLPKNVMGHNLVISKQADMQPIATDGLSAGIDKNYLKEGDTRVIAHTKVIGAGEKDSLTIDVSKLNAAEKYGFFCSFPGHISMMKGTVTL-K
6:  AECKVTVDSTDQMSFDTKAIEIDKSCKTFTVDLKHSGNLPKNVMGHNWVLTTQADMQPVATDGMAAGIDKNYLKEGDTRIIAHTKIIGAGETDSVTFDVSKLKADGKYMFFCSFPGHIAMMKGTVTL-K
7:  AECKVDVDSTDQMSFNTKEITIDKSCKTFTVNLTHSGSLPKNVMGHNWVLSKSADMAGIATDGMAAGIDKDYLKPGDSRVIAHTKIIGSGEKDSVTFDVSKLTAGESYEFFCSFPGHNSMMKGAVVL-K
8:  AECSVDIQGNDQMQFNTNAITVDKSCKQFTVNLSHPGNLPKNVMGHNWVLSTAADMQGVVTDGMASGLDKDYLKPDDSRVIAHTKLIGSGEKDSVTFDVSKLKEGEQYMFFCTFPGHSALMKGTLTL-K
9:  AECSVDIQGNDQMQFSTNAITVDKACKTFTVNLSHPGSLPKNVMGHNWVLTTAADMQGVVTDGMAAGLDKNYVKDGDTRVIAHTKIIGSGEKDSVTFDVSKLKAGDAYAFFCSFPGHSAMMKGTLTL-K
```

```
1:  AEC-SVDIAGTD-QMQFDKKAIEVSKSCKQFTVNLKHTGKLPRNVMG--HNWVLTKTADMQAVEKDGIAAGLDNQYLKAGDTRVLAHTKVLGGGESDSVTF--DVAKLAAGDDYTFFCSFPGH-GALMKGTLKLVD
2:  AEC-SVDIAGND-QMQFDKKEITVSKSCKQFTVNLKHPGKLAKNVMG--HNWVLTKQADMQGAVNDGMAAGLDNNYVKKDDARVIAHTKVIGGGETDSVTF--DVSKLAAGEDYAYFCSFPGH-FALMKGVLKLVD
3:  AQC-EATIESND-AMQYNLKEMVVDKSCKQFTVHLKHVGKMAKAVMG--HNWVLTKEADKEGVATDGMNAGLAQDYVKAGDTRVIAHTKVIGGGESDSVTF--DVSKLTPGEAYAYFCSFPGH-WAMMKGTLKLSN
4:  A-C-DVSIEGND-SMQFNTKSIVVDKTCKEFTINLKHTGKLPKAAMG--HNVVVSKKSDESAVATDGMKAGLNNDYVKAGDERVIAHTSVIGGGETDSVTF--DVSKLKEGEDYAFFCSFPGH-WSIMKGTIELGS
5:  AEC-KTTIDSTD-QMSFNTKAIEIDKACKTFTVELTHSGSLPKNVMG--HNLVISKQADMQPIATDGLSAGIDKNYLKEGDTRVIAHTKVIGAGEKDSLTI--DVSKLNAAEKYGFFCSFPGH-ISMMKGTVTL-K
6:  AEC-KVTVDSTD-QMSFDTKAIEIDKSCKTFTVDLKHSGNLPKNVMG--HNWVLTTQADMQPVATDGMAAGIDKNYLKEGDTRIIAHTKIIGAGETDSVTF--DVSKLKADGKYMFFCSFPGH-IAMMKGTVTL-K
7:  AEC-KVDVDSTD-QMSFNTKEITIDKSCKTFTVNLTHSGSLPKNVMG--HNWVLSKSADMAGIATDGMAAGIDKDYLKPGDSRVIAHTKIIGSGEKDSVTF--DVSKLTAGESYEFFCSFPGH-NSMMKGAVVL-K
8:  AEC-SVDIQGND-QMQFNTNAITVDKSCKQFTVNLSHPGNLPKNVMG--HNWVLSTAADMQGVVTDGMASGLDKDYLKPDDSRVIAHTKLIGSGEKDSVTF--DVSKLKEGEQYMFFCTFPGH-SALMKGTLTL-K
9:  AEC-SVDIQGND-QMQFSTNAITVDKACKTFTVNLSHPGSLPKNVMG--HNWVLTTAADMQGVVTDGMAAGLDKNYVKDGDTRVIAHTKIIGSGEKDSVTF--DVSKLKAGDAYAFFCSFPGH-SAMMKGTLTL-K

10:  --L-DVLLGGDDGSLAFIPGNFSVS-AGEKITF---------KNNAGFPHNVVFDE--DEIPAGVDASKISMAEE-------DLLN-----AAGETYSVTL---SEK---G-TYTFYCA-P-HQGAGMVGKVTV-N
11:  --I-EVLLGSDDGGLAFVPGNFSIS-AGEKITF---------KNNAGFPHNVVFDE--DEIPAGVDASKISMPEE-------DLLN-----APGETYSVTL---SEK---G-TYSFYCS-P-HQGAGMVGKVTV-N
12:  --I-EIKLGGDDGALAFVPGSFTVA-AGEKIVF---------KNNAGFPHNIVFDE--DEVPAGVDASKISMSEE-------DLLN-----APGETYAVTL---SEK---G-TYSFYCS-P-HQGAGMVGKVTV-Q
13:  --I-DVLLGADDGSLAFVPSEFSIS-PGEKIVF---------KNNAGFPHNIVFDE--DSIPSGVDASKISMSZZ-------BLLN-----AKGETFEVAL---SNK---G-EYSFYCS-P-HQGAGMVGKVTV-N
14:  --I-EVLLGGDDGSLAFIPNDFSVA-AGEKIVF---------KNNAGFPHNVVFDE--DEIPSGVDAGKISMNEE-------DLLN-----APGEVYKVNL---TEK---G-SYSFYCS-P-HQGAGMVGKVTV-N
15:  --I-EILLGGDDGSLAFVPNNFTVA-SGEKITF---------KNNAGFPHNVVFDE--DEIPSGVDSGKISMNEE-------DLLB-----APGZVYZVZL---TZK---G-SYSFYCS-P-HQGAGMVGKVTV-N
16:  --V-EVLLGGGDGSLAFLPGDFSVA-SGEEIVF---------KNNAGFPHNVVFDE--DEIPSGVDAAKISMSEE-------DLLN-----APGETYKVTL---TEK---G-TYKFYCS-P-HQGAGMVGKVTV-N
17:  --I-EVLLGGGDGSLAFVPNDFSIA-KGEKIVF---------KNNAGFPHNVVFDE--DEIPSGVDASKISMDEN-------DLLN-----AAGETYEVAL---TEA---G-TYSFYCA-P-HQGAGMVGKVTV-N
18:  --L-DVLLGSDDGELAFVPNNFSVP-SGEKITF---------KNNAGFPHNVVFDE--DEIPSGVDASKISMDEA-------DLLN-----APGETYAVTL---TEK---G-SYSFYCS-P-HQGAGMVGKVTV-N
19:  --A-EVLLGSSDGGLVFEPSTFSVA-SGEKIVF---------KNNAGFPHNVVFDE--DEIPAGVDASKISMSEE-------DLLN-----APGETYAVTL---TEK---G-TYSFYCA-P-HQGAGMVGKVTV-N
20:  --V-EILLGGEDGSLAFIPSNFSVP-SGEKITF---------KNNAGFPHNVVFDE--DEVPSGVDSAKISMSED-------DLLN-----APGETYSVTL---TES---G-TYKFYCS-P-HQGAGMVGKVTV-N
21:  --L-EVLLGSGDGSLVFVPSEFSVP-SGEKIVF---------KNNAGFPHNVVFDE--DEIPAGVDAVKISMPEE-------ELLN-----APGETYVVTL---DTK---G-TYSFYCS-P-HQGAGMVGKVTV-Q
22:  --V-EVLLGASDGGLAFVPNSFEVS-AGDTIVF---------KNNAGFPHNVVFDE--DEIPSGVDAAKISMPEE-------DLLN-----APGETYSVKL---DAK---G-TYKFYCS-P-HQGAGMVGQVTV-N
23:  -DV-TVKLGADSGALVFEPSSVTIK-AGETVTW---------VNNAGFPHNIVFDE--DEVPSGANAEALS--HE-------DYLN-----APGESYSAKF---DTA---G-TYGYFCE-P-HQGAGMKGTITV-Q
24:  -ETYTVKLGSDKGLLVFEPAKLTIK-PGDTVEF---------LNNKVPPHNVVFDA--ALNPAKSADLAKSLSHK-------QLLM-----SPGQSTSTTFPADAPA---G-EYTFYCE-P-HRGAGMVGKITVAG
```

```
10:  -L-DVLLGGDDGSLAFIPGNFSVSAGEKITFKNNAGFPHNVVFDEDEIPAGVDASKISMAEEDLLNAAGETYSVTL---SEKGTYTFYCAPHQGAGMVGKVTV-N
11:  -I-EVLLGSDDGGLAFVPGNFSISAGEKITFKNNAGFPHNVVFDEDEIPAGVDASKISMPEEDLLNAPGETYSVTL---SEKGTYSFYCSPHQGAGMVGKVTV-N
12:  -I-EIKLGGDDGALAFVPGSFTVAAGEKIVFKNNAGFPHNIVFDEDEVPAGVDASKISMSEEDLLNAPGETYAVTL---SEKGTYSFYCSPHQGAGMVGKVTV-Q
13:  -I-DVLLGADDGSLAFVPSEFSISPGEKIVFKNNAGFPHNIVFDEDSIPSGVDASKISMSZZBLLNAKGETFEVAL---SNKGEYSFYCSPHQGAGMVGKVTV-N
14:  -I-EVLLGGDDGSLAFIPNDFSVAAGEKIVFKNNAGFPHNVVFDEDEIPSGVDAGKISMNEEDLLNAPGEVYKVNL---TEKGSYSFYCSPHQGAGMVGKVTV-N
15:  -I-EILLGGDDGSLAFVPNNFTVASGEKITFKNNAGFPHNVVFDEDEIPSGVDSGKISMNEEDLLBAPGZVYZVZL---TZKGSYSFYCSPHQGAGMVGKVTV-N
16:  -V-EVLLGGGDGSLAFLPGDFSVASGEEIVFKNNAGFPHNVVFDEDEIPSGVDAAKISMSEEDLLNAPGETYKVTL---TEKGTYKFYCSPHQGAGMVGKVTV-N
17:  -I-EVLLGGGDGSLAFVPNDFSIAKGEKIVFKNNAGFPHNVVFDEDEIPSGVDASKISMDENDLLNAAGETYEVAL---TEAGTYSFYCAPHQGAGMVGKVTV-N
18:  -L-DVLLGSDDGELAFVPNNFSVPSGEKITFKNNAGFPHNVVFDEDEIPSGVDASKISMDEADLLNAPGETYAVTL---TEKGSYSFYCSPHQGAGMVGKVTV-N
19:  -A-EVLLGSSDGGLVFEPSTFSVASGEKIVFKNNAGFPHNVVFDEDEIPAGVDASKISMSEEDLLNAPGETYAVTL---TEKGTYSFYCAPHQGAGMVGKVTV-N
20:  -V-EILLGGEDGSLAFIPSNFSVPSGEKITFKNNAGFPHNVVFDEDEVPSGVDSAKISMSEDDLLNAPGETYSVTL---TESGTYKFYCSPHQGAGMVGKVTV-N
21:  -L-EVLLGSGDGSLVFVPSEFSVPSGEKIVFKNNAGFPHNVVFDEDEIPAGVDAVKISMPEEELLNAPGETYVVTL---DTKGTYSFYCSPHQGAGMVGKVTV-N
22:  -V-EVLLGASDGGLAFVPNSFEVSAGDTIVFKNNAGFPHNVVFDEDEIPSGVDAAKISMPEEDLLNAPGETYSVKL---DAKGTYKFYCSPHQGAGMVGQVTV-N
23:  DV-TVKLGADSGALVFEPSSVTIKAGETVTWVNNAGFPHNIVFDEDEVPSGANAEALS--HEDYLNAPGESYSAKF---DTAGTYGYFCEPHQGAGMKGTITV-Q
24:  ETYTVKLGSDKGLLVFEPAKLTIKPGDTVEFLNNKVPPHNVVFDAALNPAKSADLAKSLSHKQLLMSPGQSTSTTFPADAPAGEYTFYCEPHRGAGMVGKITVAG
```

PLASTOCYANIN