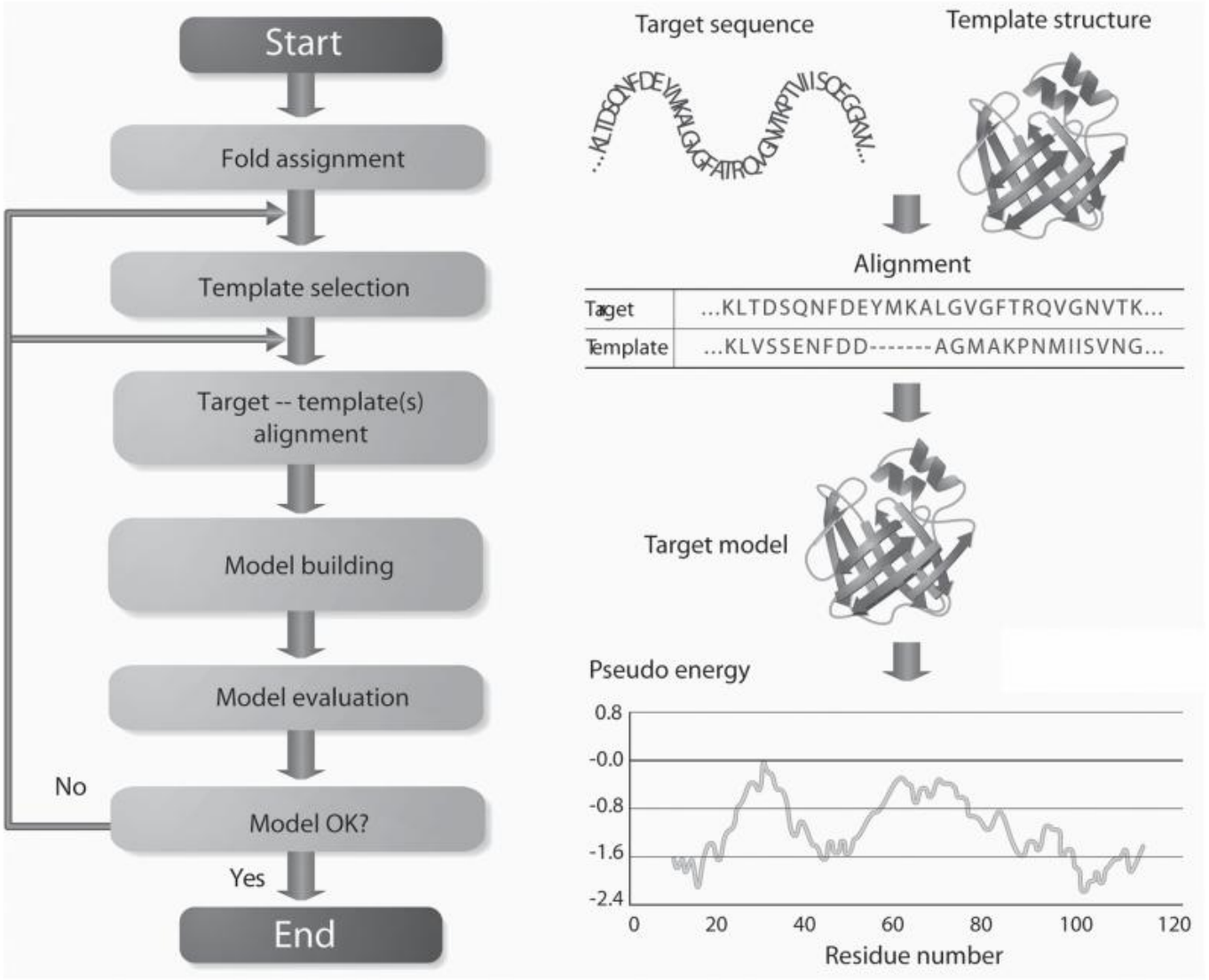
Week 10:
Homology Modelling (II) - HHpred

Course: Tools for Structural Biology

Fabian Glaser

BKU - Technion


Start
Fold assignment
Template selection
Target -- template(s) alignment
Model building
Model evaluation
Model OK?
No
Yes
End
Target sequence
Template structure
...KLTDSQNFDEYMKALGVGFTRQVGNVTK...
Alignment
Target ...KLTDSQNFDEYMKALGVGFTRQVGNVTK...
Template ...KLVSSENFDD-------AGMAKPNMIISVNG...
Target model
Pseudo energy
0.8
-0.0
-0.8
-1.6
-2.4
0
20
40
60
80
100
120
Residue number

Identify and align related structures by sequence methods is not an easy task

All comparative modeling programs strongly depend on a target-template alignment

The template search methods usually produce a non-optimal alignment, especially in the more difficult cases (< 30% ID)

Alignment methods depend on identity level

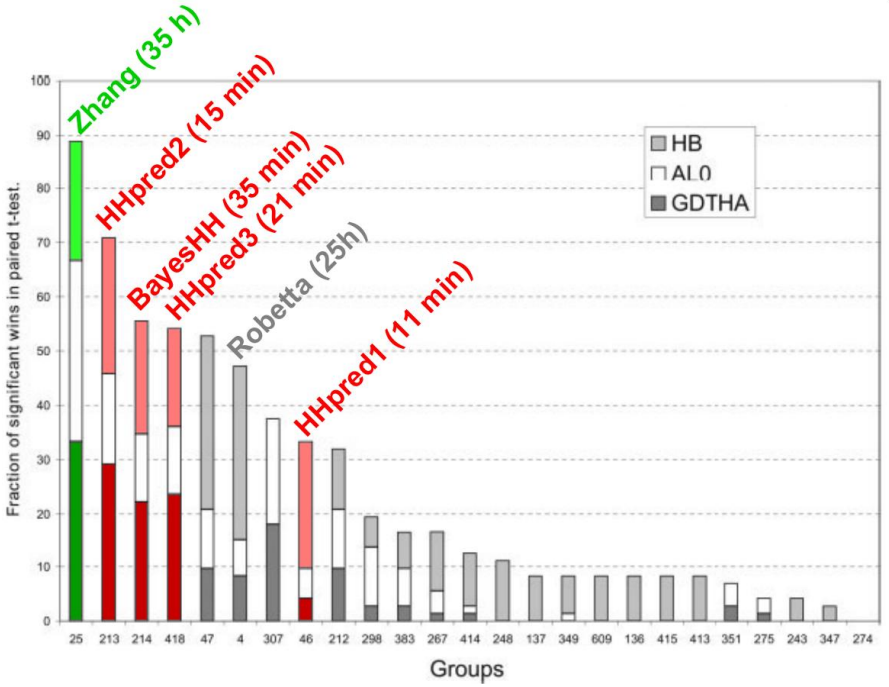
- **> 30% sequence identity**
  - Automatic methods for sequence-sequence alignment are usually accurate enough
- **< 30% sequence identity**
  - Manual alignment curation required
  - Use structural information (e.g. avoid gaps in secondary structure elements)
  - Misalignments are critical: each mistake in buried regions is estimated to cause a ~4Å deviation in the model!!
  - Therefore for this level of identity, more accurate methods are required

W244–W248 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki408

The HHpred interactive server for protein homology detection and structure prediction

Johannes Söding*, Andreas Biegert and Andrei N. Lupas

Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany



HHpred – Remote Homology detection & structure prediction

HHpred is a method for protein remote homology detection and 3D structure prediction based on the pairwise comparison of profile hidden Markov models (HMM-HMM alignment).

HHpred is as easy to use as BLAST or PSI-BLAST but at the same time is much more sensitive in finding remote homology.

HHpred accepts a single query sequence or a multiple alignment as input and it returns possible templates, E-value, etc.

HHpred can can also produce 3D-structural models calculated by the MODELLER software.

Profile based methods are more accurate

When searching for remote homologs, it is wise to make use of as much information about the query and database proteins as possible in order to better distinguish true from false templates.

For that purpose they use PROFILES

Profiles contain information about the importance of each position for finding other members of the protein family.

Profile methods are generally superior to sequence based methods.

What is a PSSM ?

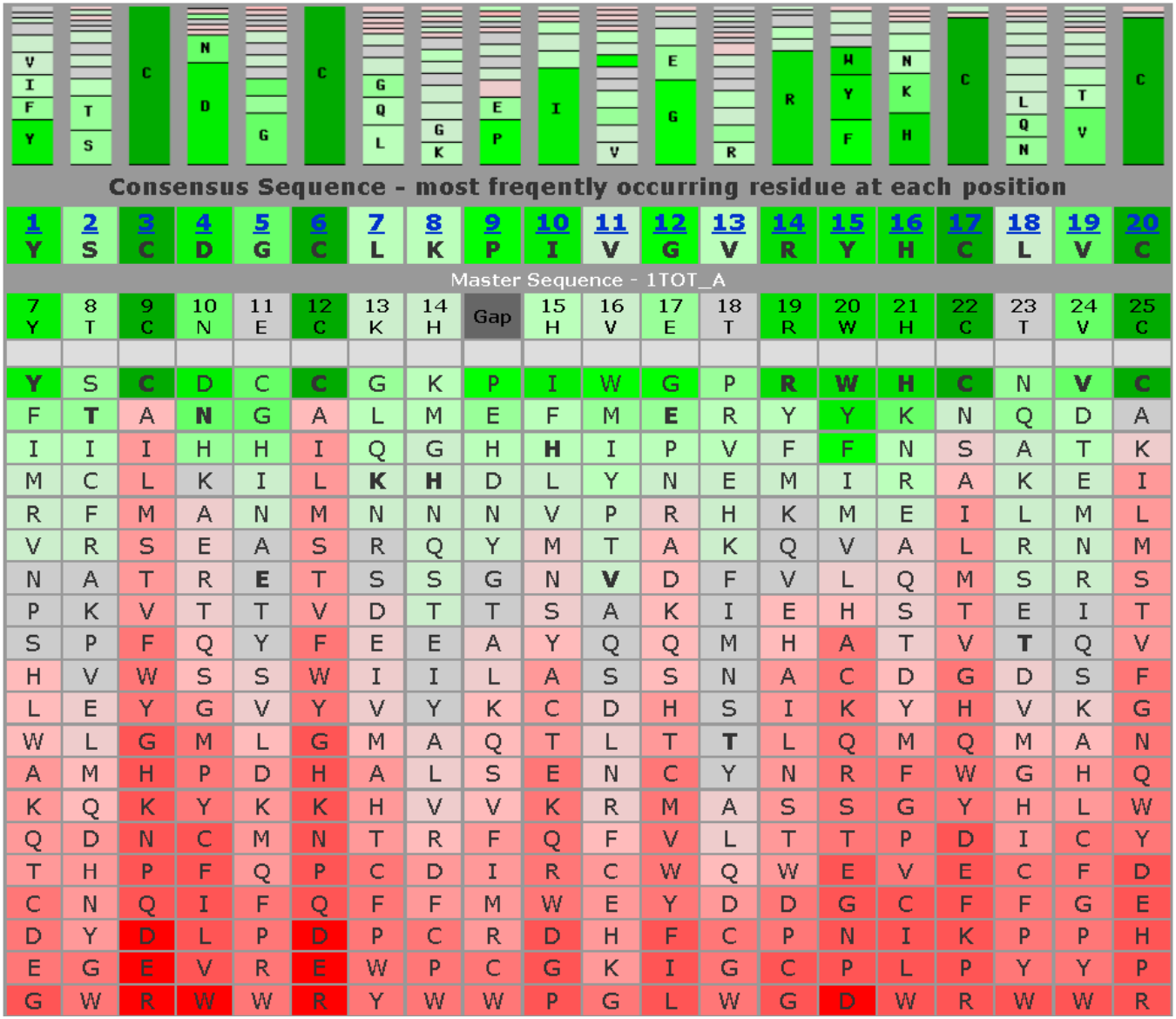
A PSSM, or Position-Specific Scoring Matrix, is a type of scoring matrix used in protein PSI-BLAST searches.

PSSM amino acid substitution scores are **position dependent** in a protein multiple sequence alignment.

Thus, a **Tyr-Arg** substitution may score different for different alignment positions.

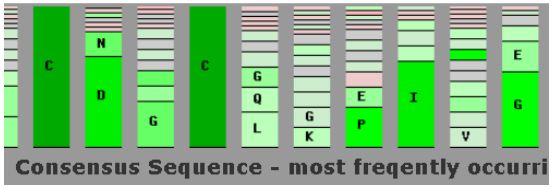
This is in contrast to position-independent matrices such as the PAM and BLOSUM matrices, in which the Tyr-Trp substitution receives the same score no matter at what position it occurs.

Position-Specific Scoring Matrix (PSSM) - NCBI Viewer

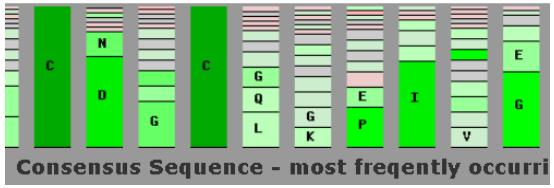


profile-profile comparison (alignment) method

1. Calculate template & target **profiles** by constructing alignments them with sequences from a NR database
2. Align the target and the template **profiles**



Align to



Profile 1

Profile 2

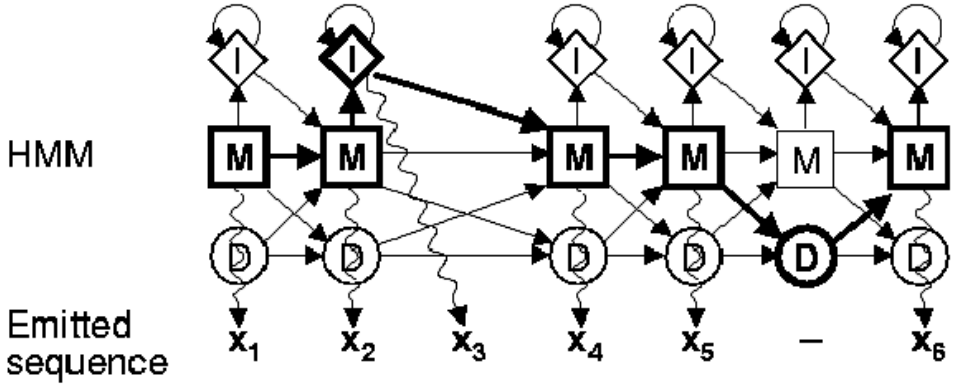
Profiles contain more evolutionary information about the family

Profile-profile alignment methods are able to provide better alignments with distant homologues

Alignment is done using HMM – Hidden Markov Model

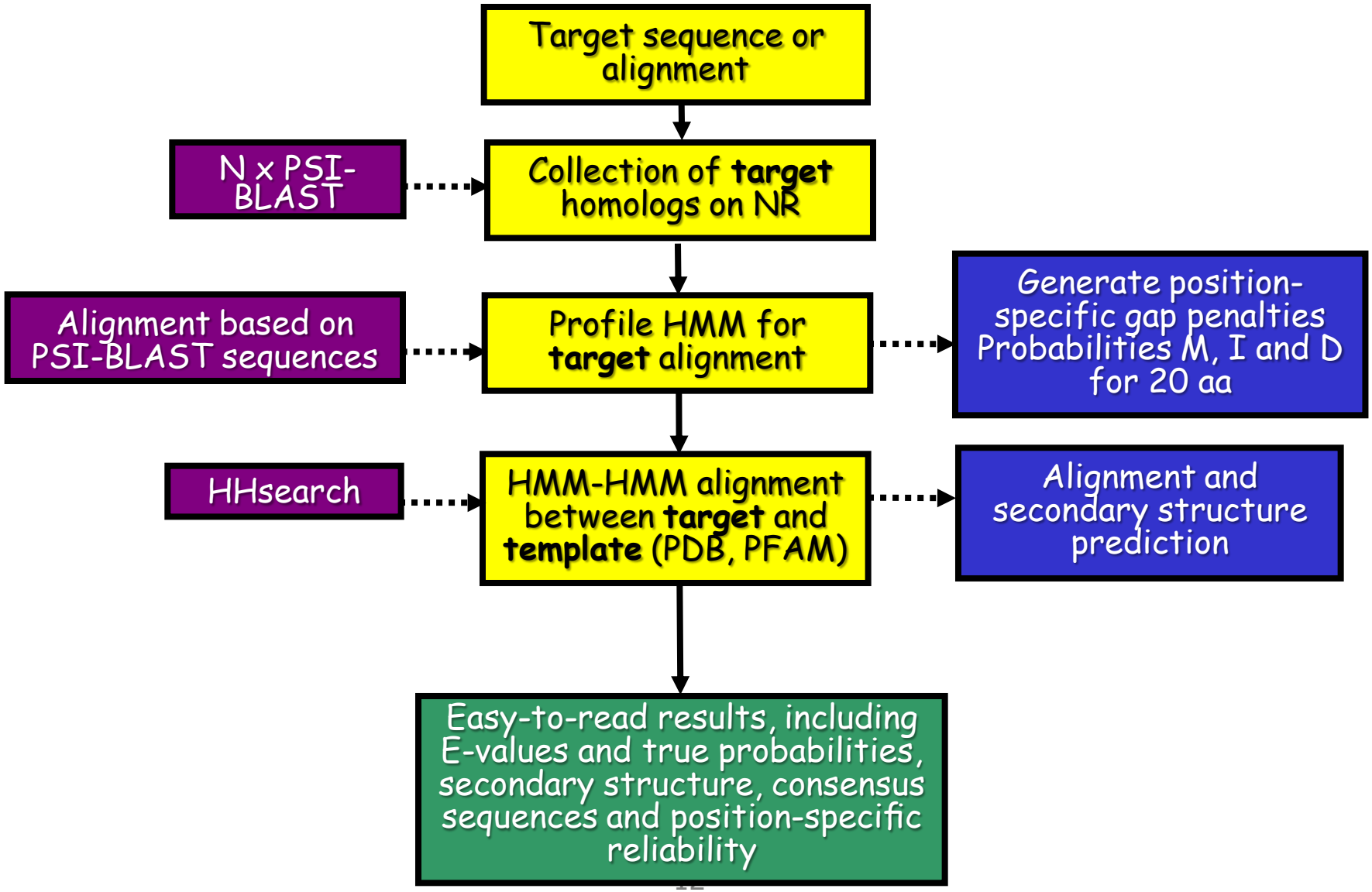
A Hidden Markov model (**HMM**) is a statistical method.

For a specific alignment, we can calculate an **HMM model** and extract information about a match (M), insertion (I) and a deletion states (D) for each column.



The extracted **model parameters** can then be used to align a new "query" sequence.

HHpred method



Criteria to choose within multiple PDB available templates

1. Higher sequence identity
2. Close subfamily
3. "Environment" similarity (solvent, pH, ligand, quaternary interactions)
4. The quality of the experimentally determined structure (RMSD)
5. Purpose of modeling (e.g. protein-ligand model)

How can I verify if a database distant match is really homologous?

1. Check probability and E-value
2. Check if homology is biologically suggestive or at least reasonable
3. Check secondary structure similarity
4. Check relationship among top hits
5. Check for possible conserved motifs (and their residues)
6. Check query and template alignments!
7. Use HHsenser to find more homologs for your query alignment
8. Verify predictions experimentally
9. Try out other structure prediction servers!

PIR Alignment used in Modeller

Alignment.ali (in PIR format)

Indicates that this is the template structure

Must match the PDB file name

Residues that take part in the alignment (pdb indexing!)

```
>P1;5fd1
structureX:5fd1:1     : :106   : :ferredoxin:Azotobacter vinelandii: 1.90: 0.19
AFVVTDNCIKCKYTDCVEVCPVDCFYEGPNFLVIHPDECIDCALCEPECPAQAIFSEDEVPEDMQEFIQLNAELA
EVWPNITEKKDPLPDAEDWDGVKGKLQHLER*
```

Indicates end of alignment

Indicates that this is the target sequence

Target name

```
>P1;1fdx
sequence:1fdx:1     : :54    : :ferredoxin:Peptococcus aerogenes: 2.00:-1.00
AYVINDSC--IACGACKPECPVNIIQGS--IYAIDADSCIDCGSCASVCPVGAPNPED-----------------
-------------------------------*
```

Indicates end of alignment

Residues that take part in the alignment