Assignment of structure class to CD177

Søren Berg Padkjær
Nove Nordisk

Homology modeling plays a central role in determining protein structure in the structural genomics project. The importance of homology modeling has been steadily increasing because of the large gap that exists between the overwhelming number of available protein sequences and experimentally solved protein structures, and also, more importantly, because of the increasing reliability and accuracy of the method.

(From: Xiang: Advances in Protein Structure modelling, Curr Protein Pept. Sci. 2006 June ; 7(3): 217227.)

For sequences with less that 30% homology to a template, a number of alternative strategies have been developed.

These include

- template consensus sequences (sequence family) and
- profile analysis (conserved preference).

Profile methods have emerged as the primary approach in distant homology detection. E. g. PSI-BLAST [46] and hidden Markov models (HMMs) [47] have extended the boundaries of detectable sequence similarity.

PSI-BLAST,

- a pair-wise search of the database.
- a position specific score matrix (PSSM).
- matrix replaces the query sequence in the next round of database searching.
- iterated until no new significant alignments are found.

Although a major goal of the profile analysis has been remote homolog detection, an important side benefit has been significant improvement in alignment quality, even at levels of sequence identity for which pairwise alignment methods are known not to work.

CD177 is a protein with unknown 3D structure

```
> >UNIPROT_Q8N6Q3_CD177
MSAVLLLALLGFILPLPGVQALLCQFGTVQHVWKVSDLPRQWTPKNTSCDSGLGCQDTLM
LIESGPQVSLVLSKGCTEAKDQEPRVTEHRMGPGLSLISYTFVCRQEDFCNNLVNSLPLW
APQPPADPGSLRCPVCLSMEGCLEGTTEEICPKGTTHCYDGLLRLRGGGIFSNLRVQGCM
PQPGCNLLNGTQEIGPVGMTENCNRKDFLTCHRGTTIMTHGNLAQEPTDWTTSNTEMCEV
GQVCQETLLLLDVGLTSTLVGTKGCSTVGAQNSQKTTIHSAPPGVLVASYTHFCSSDLCN
SASSSSVLLNSLPPQAAPVPGDRQCPTCVQPLGTCSSGSPRMTCPRGATHCYDGYIHLSG
GGLSTKMSIQGCVAQPSSFLLNHTRQIGIFSAREKRDVQPPASQHEGGGAEGLESLTWGV
GLALAPALWWGVVCPSC
```

Only short homologues with low identity are detected in PDB with standard BLAST and PSI-Blast like:

```
> >PDB:2ING_X mol:protein length:213  Breast cancer type 1 susceptibility protein
          Length = 213

 Score = 29.6 bits (65), Expect = 3.9
 Identities = 20/75 (26%), Positives = 33/75 (44%), Gaps = 4/75 (5%)

Query: 186 NLLNGTQEIGPVGMTENCNRKDFLTCHRGTTIMTHGNLAQEPTDWTTSNTEMCEVGQVCQ 245
           +++NG    GP    E+ +RK F    RG  I  +G    +PTD      ++C    V +
Sbjct: 93  DVVNGRNHQGPKRARESQDRKIF----RGLEICCYGPFTNKPTDQLEWMVQLCGASVVKE 148

Query: 246 ETLLLLDVGLTSTLV 260
            +   L  G+   +V
Sbjct: 149 LSSFTLGTGVHPIVV 163
```

We need new ideas, new forms of stringent mathematical criteria for homology and new measures of information content and characterizations of structure for proteins.