Protein Sequence Analysis

Characteristics of the sequence space and relationships with structure and function spaces

Florencio Pazos (CNB-CSIC)

Florencio Pazos Cabaleiro
Protein Design Group (CNB-CSIC)
pazos@cnb.uam.es
Protein sequences, structures, and functions

Molecular chaperonin

GroEL

(Dr. Jianpeng Ma, Harvard Univ.)
Sequence similarity relationships

Relationship protein sequence ↔ function due to the underlying duplication process (divergent evolution)

homologs /orthologs /paralogs
Sequence similarity relationships

Superfamily: Common origin but maybe not traceable by sequence homology
Family: Clear sequence homology. Function can be different.
Subfamily: Clear sequence homology and same function.  

Sometimes arbitrary
Clustering the whole sequence space

- [http://www.protonet.cs.huji.ac.il/](http://www.protonet.cs.huji.ac.il/)
Clustering the whole sequence space

• [http://www.protonet.cs.huji.ac.il/](http://www.protonet.cs.huji.ac.il/)
Protein domains
How many protein families?

Protein Structure

```
Protein Sequence:
MREYKLVVLGSGGVGKSALTQVFVGIFVDEYDPIEDSY
RKQVEVDQCQCMLEILDGTAGQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMLVGNKCDL
EDERVVGKEQGQNLRQCNCAFLESSAKSINVNEIFYDLVRQIN
```
Genome sequencing

Experimental determination of protein structures

By D. Devos
High-throughput determination of protein structures – Structural Genomics

Characteristics of the structure space

~50,000 structural domains

~2,000 unique folds


Characteristics of the structure space

Characteristics of the structure space

http://www.rcsb.org
Characteristics of the structure space


Characteristics of the structure space
Relationships with sequence space

Characteristics of the structure space
Relationships with sequence space

P-loop ATP hydrolases FOLD
Characteristics of the structure space

$\alpha/\beta$ structural CLASS

P-loop ATP hydrolases FOLD

http://scop.mrc-lmb.cam.ac.uk/scop/index.html

Relationship between structure and function spaces

Relationships between protein sequence structure and function spaces
The protein “Universe”

$10^{400}$ Possible sequences

$10^{10}$ Sequences in the biosphere

$10^5$ Families

$10^3$-$10^4$ Folds

$10^4$ Functions (GO)