

“Linear Sequence Analysis”

What can you learn from a (single) protein sequence?

- Calculate it's physical properties
 - Molecular weight (MW), isoelectric point (pI), amino acid content, hydrophathy (hydrophilic v. hydrophobic regions)
 - Does not take into account post-translational modifications of protein, so are usually not 100% accurate
- Identify sequence motifs and families
 - Signal sequences, transmembrane domains, coiled-coils, post-translational modification sites, secondary structure (non-homologous)
 - Domains, functional motifs (homologous)

“3-D Structure Analysis”

- Visualization
 - Domain structure, global fold, active sites, point mutations, SNPs, splice sites
- Evaluate structure “quality”
- Calculate physical properties
 - Surface areas, distances, side-chain conformations, contact maps
 - Structural alignment (ie similarity to other structures)
- Prediction
 - Physical properties: binding affinity, pKa's stability, specificity
 - 3D structure (homology modeling, fold recognition, de novo)
 - Advanced: protein design, “docking” of two proteins, active site modeling

Sequence Databases

- SwissProt (ExPASy)
 - Highly curated, updated less frequently
- TrEMBL (ExPASy)
 - Translated nucleotide sequences
 - Automatic translation, fast but less info
- UniProt (EBI)
 - Unified Protein Resource
 - Combines SwissProt, TrEMBL, PIR sequences

Sequence Analysis Sites

For protein sequences and tools to analyze them, the two major centers are:

- ExPASy : Expert Protein Analysis System
 - Many tools: <http://ca.expasy.org/tools/>
 - Databases: SwissProt, TrEMBL
- NCBI : Entrez Protein and Domains
- PIR: Protein Information Resource (folded into UniProt consortium; no longer major resource site)

More Sequence Databases...

- Non-redundant
 - NR (NCBI), UniRef (PIR/EBI)
- Reference
 - RefSeq (NCBI) – re-annotated by NCBI
- Domains/Families
 - Pfam – protein families (Sanger Center + 4 mirror sites)
 - SMART – Simple Modular Architecture Research Tool
 - CDD – Conserved protein Domain Database (NCBI), combines Pfam, SMART, and COGs databases
 - InterPro (based on UniProt, at EMBL-EBI)
 - Many others...

Structure Databases

- Experimental:
 - PDB: Protein Data Bank
 - Families:
 - SCOP, CATH, Dali database, Homstrad
- Models/Predictions
 - ModBase
 - SwissModel
- NOTE: All these databases are described in January Database issue of Nucleic Acids Research (plus other kinds of databases).
- Also, links to them

Protein Sequence Analysis Tools

- ExPASy Proteomics Tools
 - Calculate physical properties
 - Predict sequence motifs
 - what ExPASy calls 'Topology' : localization , TM domains
 - Signal sequences, postranslational modifications
 - Search pattern and profile collections
- PredictProtein and Meta-PP
 - A meta-server providing access to many servers with one submission form

Secondary Structure Prediction

- Three good methods:
 - Psipred
 - Sam-T02/T04/T06
 - PhD (PredictProtein)
- Compare a couple methods
- Use the three-state predictions

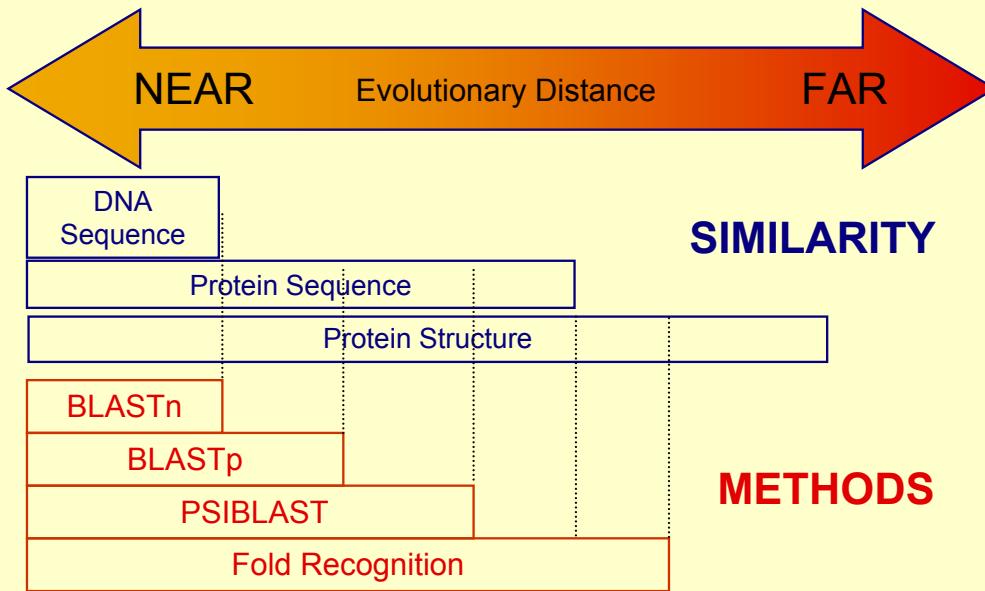
SEQUENCE <--> STRUCTURE <--> FUNCTION

- Evolutionary selection operates on **function**
- **Structure** is more closely linked to function than is sequence, so structure tends to be more conserved than sequence.
- Need to search farther in sequence space to find proteins with related structures and functions.

Detecting Remote Similarities

- Remote similarities can more easily be detected by comparing protein sequences
 - DNA sequences change faster than protein sequences (wobble position, redundant codons)
 - 4 letter DNA code vs. 20 letter amino acid code means that matches by chance are more likely in DNA; The protein code has more information in it!

Detecting Homology



Similar Sequences Share Similar Structures

- Compare all pairs of proteins in the same 'family' (pairs for which homology is **very** probable)
- Homologs do not necessarily share much sequence similarity.
- Proteins with >30% sequence identity almost always share the same fold

More structurally similar

