

### Methods for Receptors with Known Sequence

1. From Genome to Protein Sequences
2. From Sequences to Structure
  - Comparative Modeling
  - Knowledge-based Modeling
  - ab initio Modeling

### DNA Sequence Types in Databanks

- **mRNA** - may contain introns (non-coding sequences) and may not cover complete coding region
- **cDNA** - derived from mRNA by reverse transcription
- **genomic DNA** - from genome sequencing project; may contain introns, repeat regions, and other features; usually complete
- **GSS** (Genome Survey Sequence) - single-pass sequence with many errors
- **EST** (Expressed Sequence Tags) - short cDNA sequences prepared from mRNA from cell under particular conditions (e.g. disease); will not cover complete coding region.

### From Genome to Proteins

- the aim for drug design: DNA → RNA → protein → 3D-structure → binding
- sequences of nucleotides (A, C, G, T) available, the aim is to translate them to amino acid sequences
- coding regions for proteins (genes/cistrons) – no straightforward to find, annotation (sequence features) important
- amino acids coded by codons (triplets) but the sequence also contains start and stop codons...

### Genomic Information

- three most comprehensive data banks
  - **GenBank**  
www.ncbi.nih.gov/Genbank/
  - **EMBL-EBI Nucleotide Sequence Database**  
www.ebi.ac.uk/embl/
  - **DNA Data Bank of Japan**  
www.ddbj.nig.ac.jp
- daily exchange of data
- tools for processing the information

### Target Sequences for Drug Design

Target sequences are selected based upon various kinds of experimental information:

- **proteome mapping in disease**
  - the sequences that are over-expressed can be targeted
- **protein-protein interactions** – micro-arrays
  - interactions that are occurring in disease can be prevented by compounds that bind to one of the proteins in the binding area

### Translation Table(s)

- standard table
- other versions (15) available, e.g.
  - vertebrate
  - mitochondrial
  - bacterial
- little variation

1st	2nd				3rd
	T	C	A	G	
T	F Phe	S Ser	Y Tyr	C Cys	T
	F Phe	S Ser	Y Tyr	C Cys	C
	L Leu	S Ser	Ter	Ter	A
	L Leu	S Ser	Ter	W Trp	G
C	L Leu	P Pro	H His	R Arg	T
	L Leu	P Pro	H His	R Arg	C
	L Leu	P Pro	Q Gln	R Arg	A
	L Leu	P Pro	Q Gln	R Arg	G
A	I Ile	T Thr	N Asn	S Ser	T
	I Ile	T Thr	N Asn	S Ser	C
	I Ile	T Thr	K Lys	R Arg	A
	M Met	T Thr	K Lys	R Arg	G
G	V Val	A Ala	D Asp	G Gty	T
	V Val	A Ala	D Asp	G Gty	C
	V Val	A Ala	E Glu	G Gty	A
	V Val	A Ala	E Glu	G Gty	G

## Translation

---

- DNA contains two complementary antiparallel strands  
5'-AGCAGTCGATGCCGAATTCC-3'  
3'-TCGTCGACTACGGCTTAAGG-5'
- each strand can be read in two directions
- for each direction, there are three possible ways – six possible reading frames
- those that have stop codons early in the sequence can be discarded

## Protein Sequence Databases II

---

- **SWISS-PROT**
  - [www.expasy.ch/sprot/sprot-top.html](http://www.expasy.ch/sprot/sprot-top.html)
  - [www.ebi.ac.uk/swissprot/](http://www.ebi.ac.uk/swissprot/)
  - well annotated and cross-referenced
- **TrEMBL**
  - the same URL
  - a computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT

## Post-translational Modifications

---

- important for protein function/ligand binding
  - amidations  
-COO<sup>-</sup> → -CONH<sub>2</sub>
  - phosphorylations  
-OH → -OPO<sub>3</sub><sup>2-</sup>
  - ...
- can be identified by mass spectrometry
- annotated in the databases

## Protein Sequence Databases I

---

- Protein Information Resource (**PIR**)
  - [pir.georgetown.edu](http://pir.georgetown.edu)
  - annotated and classified sequences (PIR1), preliminary sequences (PIR2), unverified sequences (PIR3), conceptual sequences (PIR4)
- **PIR-NRL3D**
  - [pir.georgetown.edu/pirwww/dbinfo/nrl3d.html](http://pir.georgetown.edu/pirwww/dbinfo/nrl3d.html)
  - sequences of proteins in PDB
  - makes PDB available for similarity searches

## Protein Sequence Databases III

---

- Munich Information Center for Protein Sequences (**MIPS**)
  - [mips.gsf.de](http://mips.gsf.de)
- The Institute for Genomic Research (**TIGR**)
  - [www.tigr.org](http://www.tigr.org) (Rockville, MD)
  - microbial sequences
- many other databases available
- Lion Biosciences provide (free to academia) **Sequence Retrieval System** (SRS)
  - [downloads.lionbio.co.uk/publicsrs.html](http://downloads.lionbio.co.uk/publicsrs.html)
  - advanced search through 100's of libraries

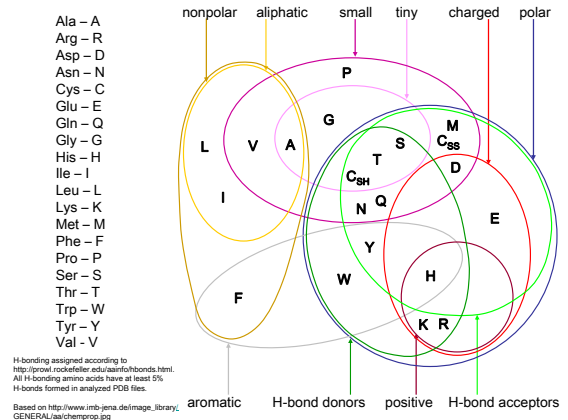
## Sequence Alignment I

---

- structure is more conserved than sequence
  - similarity in sequence implies similarity in structure
  - proteins with different sequences can have similar structures
- similarity in sequence also implies similarity in function
- comparison of the same protein in different organisms
  - evolutionary relationships (phylogenetic analysis)
- when doing alignments, one can look at
  - identical amino acids
  - amino acids with similar properties

## Sequence Alignment II

- sequence **identity**
  - the same amino acids at equivalent positions
- sequence **similarity**
  - amino acids with similar properties can be interchanged
- sequence **homology**
  - homology and similarity are often used as synonyms
  - in phylogenetic analysis, two sequences are homologous if they share a common ancestor

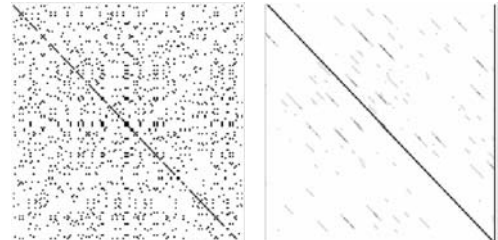


## Sequence Alignment: Procedure

- the compared sequences are
  - moved along each other
  - the gaps are introduced if needed, penalties
    - for opening a gap
    - for extending a gap (smaller)
- similarity matrix
  - rows and columns are individual positions
  - elements contain scores for comparison
    - highest scores for identity
    - lower scores for similarity
    - negative scores for dissimilar substitutions
- the best alignment has the highest score

## Sequence Alignment: Dotplots

- the plot of the relation between two sequences
  - the dots mark positions of identical AA
  - can be optimized by 'sliding window'



## Sequence Alignment: Two Sequences

- **lalign**
  - [www.ch.embnet.org/software/LALIGN\\_form.html](http://www.ch.embnet.org/software/LALIGN_form.html)
- **SIM**
  - [ca.expasy.org/tools/sim.html](http://ca.expasy.org/tools/sim.html)
- **align**
  - [www.ebi.ac.uk/emboss/align/](http://www.ebi.ac.uk/emboss/align/)
- the best way to align two sequences is to align the whole family
  - varying AAs less important
  - conserved AAs functionally important
  - gaps frequently in loops

## Sequence Alignment: Databases

- Basic Local Alignment Search Tool (**BLAST**)
  - [www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/)
  - uses local alignments
  - able to detect relationships among sequences which share only isolated regions of similarity
  - various enhancements and variants available
- **FASTA**
  - [www.ebi.ac.uk/fasta33/](http://www.ebi.ac.uk/fasta33/)
- **ClustalW**
  - [www.ch.embnet.org/software/ClustalW.html](http://www.ch.embnet.org/software/ClustalW.html)
- **T-COFFEE** (very good, comparatively slow)
  - [www.ch.embnet.org/software/TCoffee.html](http://www.ch.embnet.org/software/TCoffee.html)

## Sequence Alignment: Quality

- % identity
- Expect (E-) value
  - takes into account the size of the database
  - describes the number of hits one can "expect" to see just by chance when searching a database of a particular size
- E-value – rule of thumb
  - $E < 0.1$  good, random alignment
  - $E \sim 1$  means 50% chance alignment
  - $E > 10$  no relationship

## Alignment Plus Motifs

- Position-specific Iterated BLAST (**PSI-BLAST**)
  - [www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/)
- initial alignments generate a **profile**
  - frequencies of AAs in individual positions
- motifs can be incorporated easily
- the procedure is iterated until no statistically different sequences are found in the database
- very good alignment tool

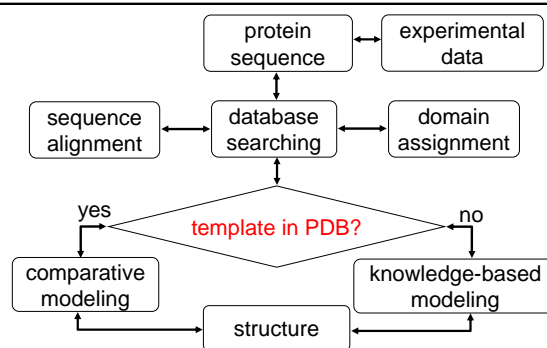
## What Is a Template?

identical AA	outcome
0 – 25 %	difficult to detect any homology
26 – 50 %	alignment can be a problem
51 – 75%	good comparative models, however, local errors can be large (a problem for drug design)
76 – 100 %	excellent comparative models with quality comparable to experimental structure (except a few side chains)

## Motifs

- specific consensus patterns discovered in multiple alignments
- discovered through profiles, Hidden Markov Models, position specific score matrices
- indicative of protein function
  - function can be identified even if overall similarity is low
- organized in motif databases
- many motif databases can be searched simultaneously by InterPro
  - [www.ebi.ac.uk/interpro/scan.html](http://www.ebi.ac.uk/interpro/scan.html)

## From Sequence to Structure



## Comparative Modeling

- identify and select related structures (templates)
- align target sequence with templates
  - for drug design, similarity >75%
  - gaps should be between secondary structure elements
- build a structural model for the target using information about template structures
  - in principle, copy template structure to target
  - gaps – loop search algorithms for structures from PDB-based libraries or optimization
- evaluate model (repeat if not satisfied)
  - procheck (bonds, dihedrals, noncovalent int's...)
  - molecular modeling (side chain optimization, MD)

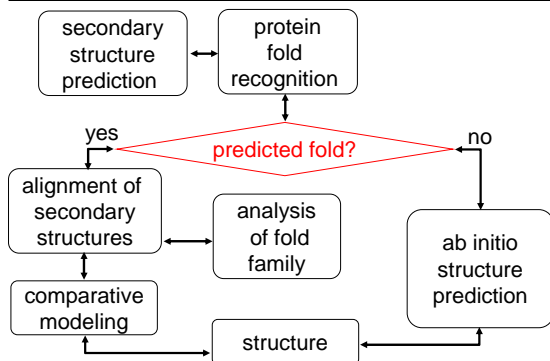
## Comparative Modeling: Software

- **Modeller** (Andrej Sali) – unix software
- **MODBASE** (<http://guitar.rockefeller.edu/modbase>) is a relational database of annotated comparative protein structure models for all available protein sequences matched to at least one known protein structure
- **SWISS-MODEL** – an easy to use web-server  
--- [www.expasy.ch/swissmod/SWISS-MODEL.html](http://www.expasy.ch/swissmod/SWISS-MODEL.html)

## Knowledge-based Modeling

- also called 'fold recognition', 'threading', 'recognition of remote homologies'
- sequences with low % identity can adopt the same 3D fold
- around 1000 folds are expected to occur in all proteins
- sequence is 'threaded' into each of the known folds in a database and resulting energy is evaluated by a score
- Methods: e.g. FUGUE, 3D-PSSM  
[www.sbg.bio.ic.ac.uk/~3dpssm/](http://www.sbg.bio.ic.ac.uk/~3dpssm/)

## Knowledge-based Modeling: Flow Chart



## Ab Initio Protein Folding

- **Rosetta method** – based on local structures observed in PDB; MC search of the possible combinations of likely local structures, minimizing a scoring function that accounts for nonlocal interactions such as compactness, hydrophobic burial, and pair interactions
- **Static approaches** – search for kinetically accessible minima in hypersurfaces generated by force-field-type description of electrostatic, H-bonding and dispersion interactions
- **Dynamic approaches** – simulations with different protein representations (coarse-grain Langevin simulations, torsional angular simulations, MD simulations)

## Protein Structure Databases

- **PDB**  
--- [www.rcsb.org/pdb](http://www.rcsb.org/pdb)

Derivative databases:

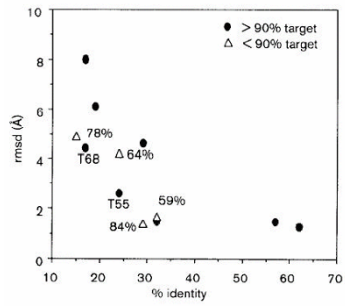
- **SumPDB** – annotated PDB  
--- [www.biochem.ucl.ac.uk/bsm/pdbsum/](http://www.biochem.ucl.ac.uk/bsm/pdbsum/)
- **Relibase+** - based on PDB, focusing on ligands and binding sites  
--- [relibase.rutgers.edu/](http://relibase.rutgers.edu/)
- **MSD** – Macromolecular Structure Database  
--- [www.ebi.ac.uk/msd/](http://www.ebi.ac.uk/msd/)

## What are the Best Methods?

- CASP – Community-wide experiment on the Critical Assessment of techniques for protein Structure Prediction - <http://predictioncenter.llnl.gov/>
- a competition organized every 2 years since 1994
- experimentalists provide unpublished structures
- the sequences are made available to all researchers
- researchers can predict the structure by  
--- homology modeling  
--- fold recognition (threading)  
--- ab initio modeling
- results are then compared against the real experimental structure.

## Quality of the Predictions I

- comparative modeling – CASP3



## Quality of the Predictions II

- threading, ab initio – CASP3

