Molecular Biology and Genetics Undergraduate Program
Department of Biological Sciences
Faculty of Arts and Sciences
Eastern Mediterranean University

This lecture is part of Spring 2020 – Covid-19 Pandemic – Online Education

Following rules are to be strictly followed:

- **Unauthorized photography, video or audio recordings of these sessions are NOT allowed.**
- **Lecture materials are not to be in-part or in-full distributed IN ANY FORM, ON ANY PLATFORM whatsoever.**

Introduction to protein 3D structure prediction and molecular modelling

Protein modelling

- The Protein-Structure gap
- Homology modelling: principles
- Sequence alignments
- Automated Homology Modeling
- SWISS-MODEL server

The Protein Structure Gap

Experimental protein structure solution (eg. by NMR or X-Ray crystallography) is labor intensive and expensive. For the majority of proteins in any given proteome, experimental structures are not available.

- Is it possible to **predict** 3-dimensional protein structures **computationally**?

- Which computational methods are **feasible** and applicable in a life science research context?

Protein Structure Prediction: Historic Concepts

The "threading / physics" concept: "The native conformation of a protein corresponds to a global free energy minimum of the protein / solvent system. To identify the correct fold, some form of energy calculation should be used to evaluate compatibility of the protein sequence with a structural conformation."

Ludwig Boltzmann
(20. February 1844 – 5. September 1906)

The "biology" perspective: "Homologous proteins have evolved by molecular evolution from a common ancestor over millions of years. If we can establish homology to a known protein, we can predict aspects of structure and function of a new protein by similarity."

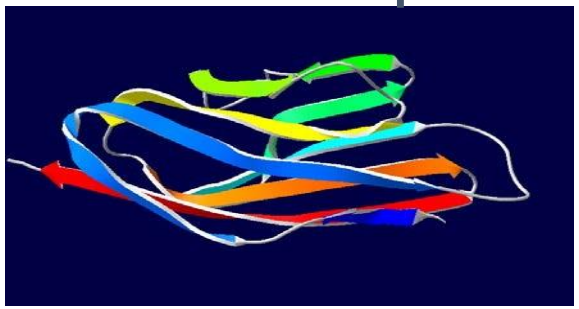
Charles Darwin
(12. February 1809 - 19. April 1882)

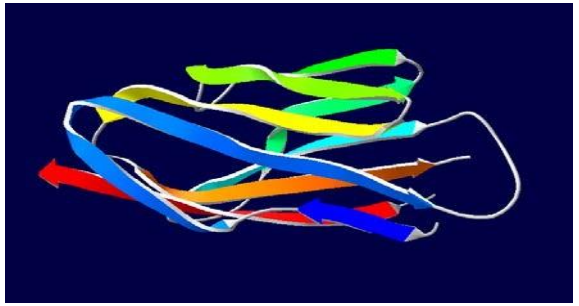
Homology (Comparative) Modeling

Evolution of Protein Structures

Protein structure is better conserved than sequence

Similar Sequence ➔ Similar Structure

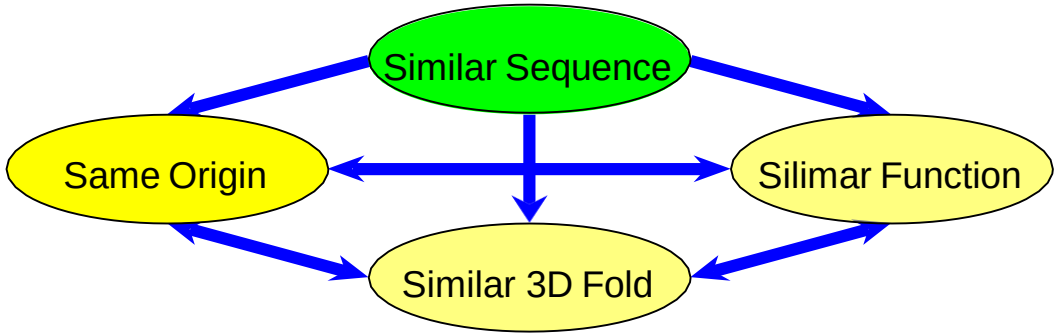




Homology modeling = Comparative protein modeling

Idea:	Using experimental 3D-structures of related family members (templates) to calculate a model for a new sequence (target).

Evolution of Protein Structures



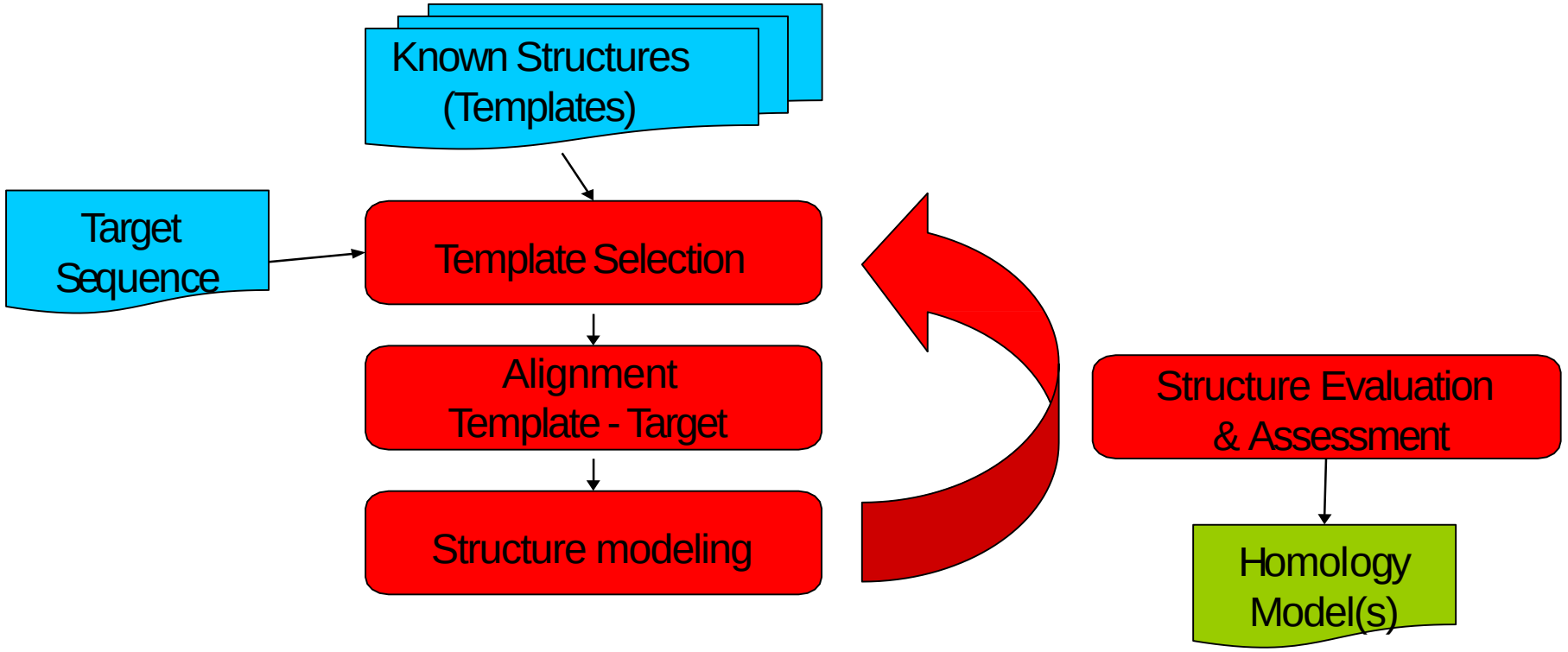
Protein structure is better conserved than sequence

Biological sequences evolve through mutation and selection

=> Selective pressure is different for each residue position in a protein (i.e. conservation of active site, structure, charge, etc.)

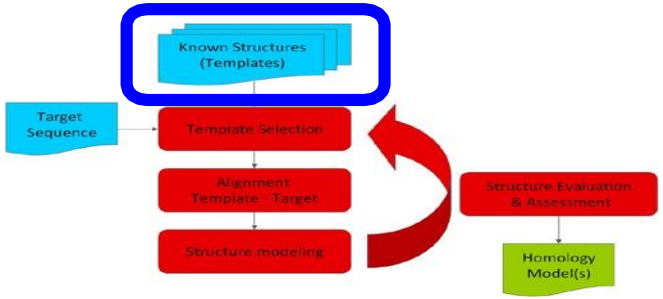
Alignments try to tell the evolutionary story of the proteins

Template-based Protein Modeling



General Workflow

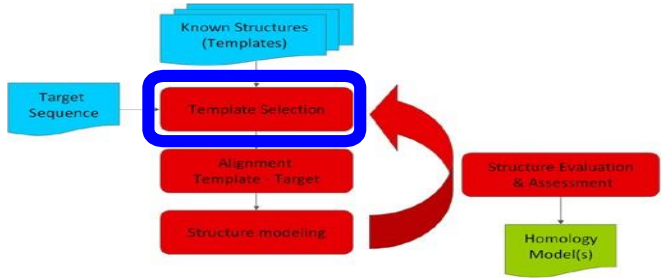
Comparative Protein Modeling



1. Template structure library

- Derived from the PDB http://www.pdb.org
- Remove incomplete and low resolution structures;
- Assign correct oligomeric states (e.g. PISA);
- Annotate bound ligands and non-protein molecules
- Cluster template sequences to remove redundancy
- Extract amino acid sequences and create searchable databases (BLAST) or profile libraries (HMMs);

Comparative Protein Modeling



2. Template selection / fold recognition

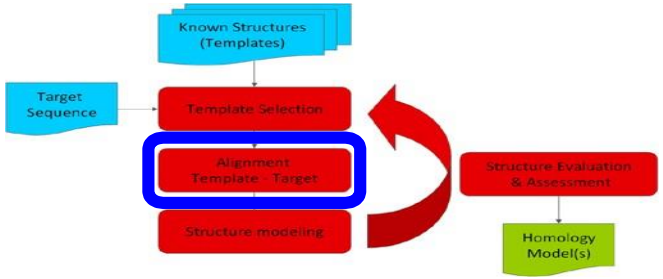
- search template library for homologous sequences using BLAST (close homologs) or HMM-HMM profile methods (remote homologs)
- ***Note:*** *Not all structures are of the same quality (resolution, experimental method).*
- ***Note:*** *Experimental conditions (e.g. bound ligands, cofactors, presence of DNA) can have a decisive role on the usefulness of a template - depending on the planned application.*

Hidden Markov Models (HMM Profile)

- Are a general statistical modeling technique for 'linear' problems like sequences or time series and have been widely used in speech recognition applications for twenty years.
- Within the HMM formalism, it is possible to apply formal, fully probabilistic methods to profiles and gapped sequence alignments, thus outperforming the well-known BLAST and FASTA algorithms in finding distantly related sequence homologues.
- The key idea is that an HMM is a finite model that describes a probability distribution over an infinite number of possible sequences.
- The HMM is composed of a number of states, which might correspond to positions in a 3D structure or columns of a multiple alignment.

(http://www.biopred.net/hmm.html)

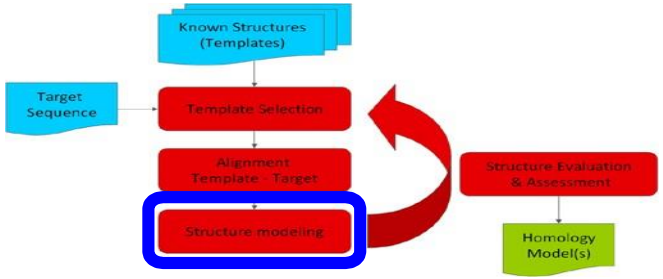
Comparative Protein Modeling



3. Target - Template alignment

- Pair wise or multiple sequence alignment works well for sequence identity>50%
- HMM-HMM profile methods for detecting remote homologs deliver high quality alignment which can be used directly.
- In case of ambiguous alignments, build several models for alternative alignment variants.
- ***Note:*** *Correct template alignments are crucial as none of the current modeling techniques can recover from incorrect input alignment*.

Comparative Protein Modeling

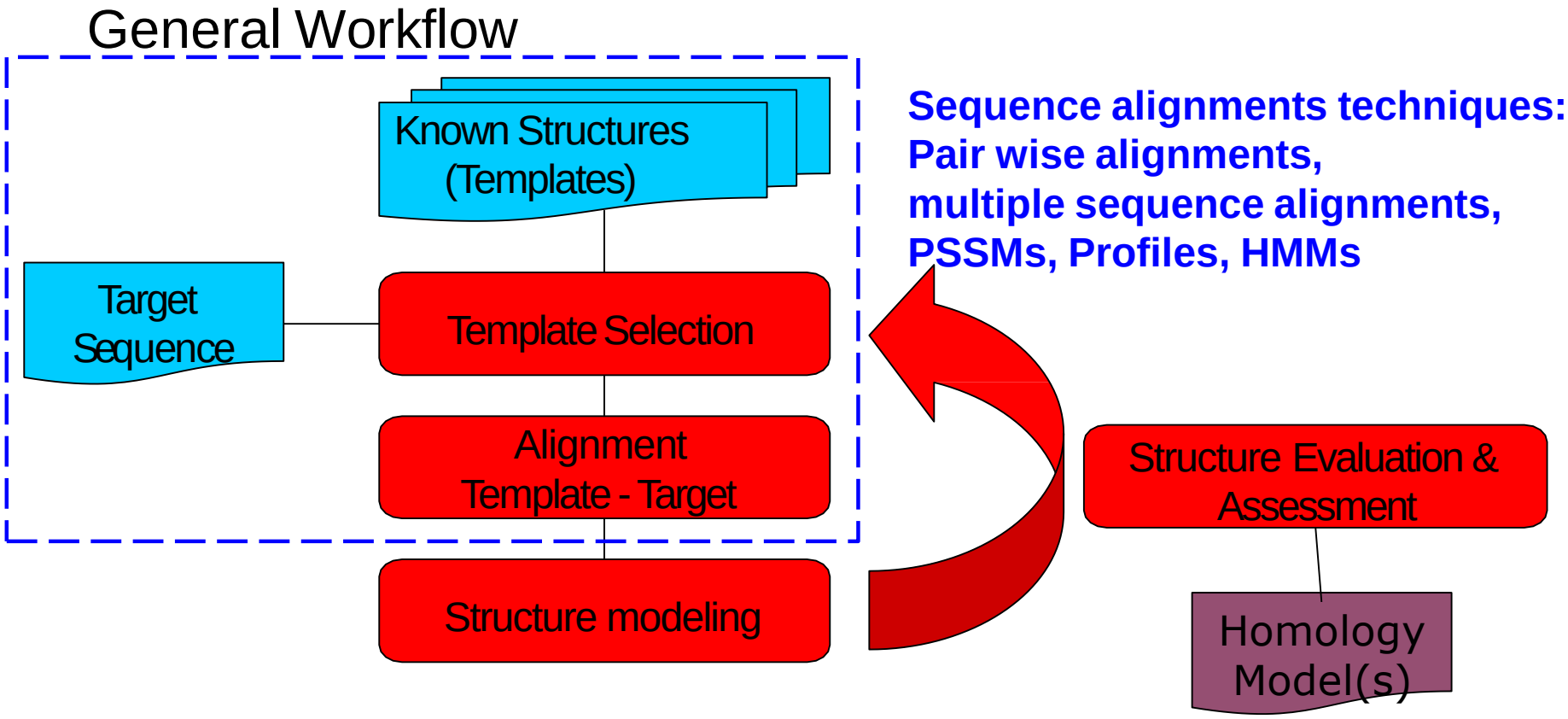


4. Building model coordinates

- Model coordinates are constructed based on the alignment to one or several template structures.

- Two main approaches for coordinate building:
  - Template based fragment assembly e.g. Composer (Sybyl, Tripos); SWISS-MODEL
  - Satisfaction of spatial restraints (Modeller)

Comparative Protein Structure Modeling



***PSSMs: Position-Specific Scoring Matrix

Homology Modelling Workflow