# IC251 – Basics of Bioinformatics (4 Credits)

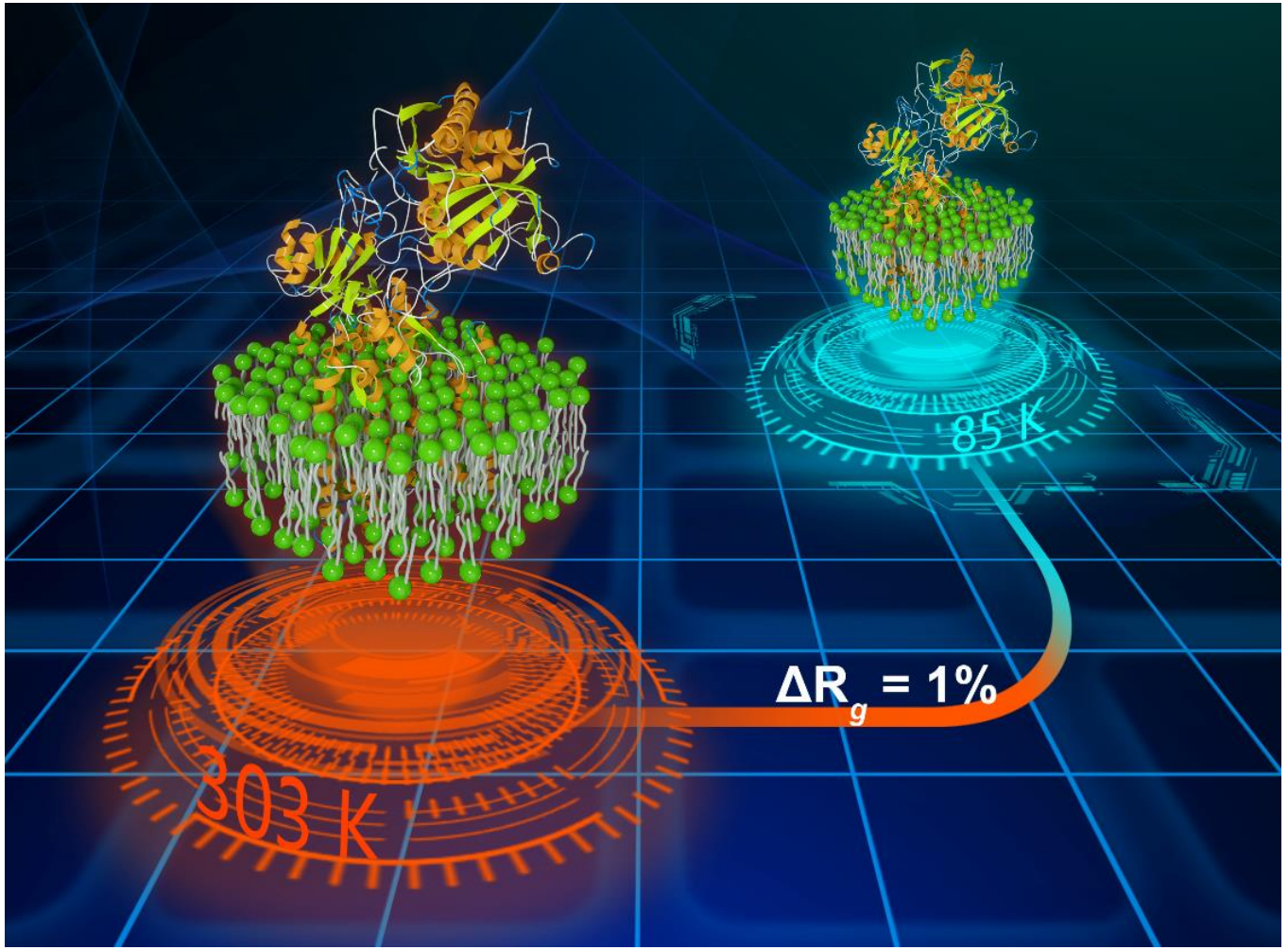
# Lecture – Molecular visualization

# Representation and visualization

## Molecular visualization

### What is it?

Example of protein model in membrane




**Source:** Cover Graphics, Rukmankesh Mehra et al., 2020, *Physical Chemistry Chemical Physics*, 22, 5427-5438 https://doi.org/10.1039/C9CP06723J

# Representation and visualization

## Molecular graphics

**What is it?**

- Formally, molecular graphics refers to a visualization of molecular objects.
- But this term is also used as a synonym of molecular modeling.
- Molecular visualization is an interdisciplinary problem between chemistry and computer sciences.
- Currently, virtual chemistry on screen is a routine and highly interactive user-friendly systems are expected to be a part of each molecular platform.
- Molecular representations here can range from the atoms to the surfaces.
- Molecular representation examples include stick, ball and stick, CPK (Corey, Pauling and Koltun) or space filling, surface, ribbon, cartoon.

J. Polanski, in Comprehensive Chemometrics, 2009 (Volume 4)

# Representation and visualization

## Molecular graphics

### Why do we require it?

- The development of molecular graphics has had a profound effect on our ability to view, interrogate, and model molecular structure.
- The most important advantages are the ability to visualize and manipulate the three-dimensional structure of molecules and to provide rapid and detailed analyses of molecular properties, especially when closely coupled to molecular calculations.

RODERICK E. HUBBARD, in Guidebook on Molecular Modeling in Drug Design, 1996

### How do we do it?

- Numerous molecular graphics packages are available that allow the user to view and manipulate molecular structures, providing insight into how the structure of the molecule might be related to its chemical or biological behavior.

A.J. ABRUNHOSA et al., in Quantitative Functional Brain Imaging with Positron Emission Tomography, 1998

**Major molecular graphics tools** ⇨

- ✓ Maestro
- ✓ VMD
- ✓ PyMol
- ✓ Swiss-PDB Viewer

# Lecture - Protein 3D structure prediction

# Protein 3D structure prediction

- ✓ The process by which the 3-dimensional structure of a protein is identified by using either its homology to the known protein structures (from experiments) or by optimizing it's energy.
- ✓ Predict atomic model from the amino acid sequence.

Terms used:

Target protein – to be predicted

Template protein – existing structure

## Methods

- Homology modeling
- Threading / fold recognition
- *Ab initio*

→ Knowledge-based methods

→ Use existing protein structural information to predict new (i.e. use experimental structure)

# Protein 3D structure prediction

## Methods

### Homology modeling

Also known as comparative modeling.

Close sequence homology with existing structure.

≥30% sequence identity

### Threading / fold recognition

Structural similarity with existing structure.

May or may not be similar at sequence level.

Approx. < 30% sequence identity

### *Ab initio*

Simulation based approach.

Predicts based on physicochemical principles governing protein folding.

Do not use structural templates.

No sequence identity

**Accuracy:** Homology modeling > Threading > *Ab initio*

# Methods

## Homology modeling

### Principle:

**If two proteins share a high enough sequence similarity, they are likely to have very similar three-dimensional structures.**

**If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence.**

**Homology modeling produces an all-atom model based on alignment with template proteins.**

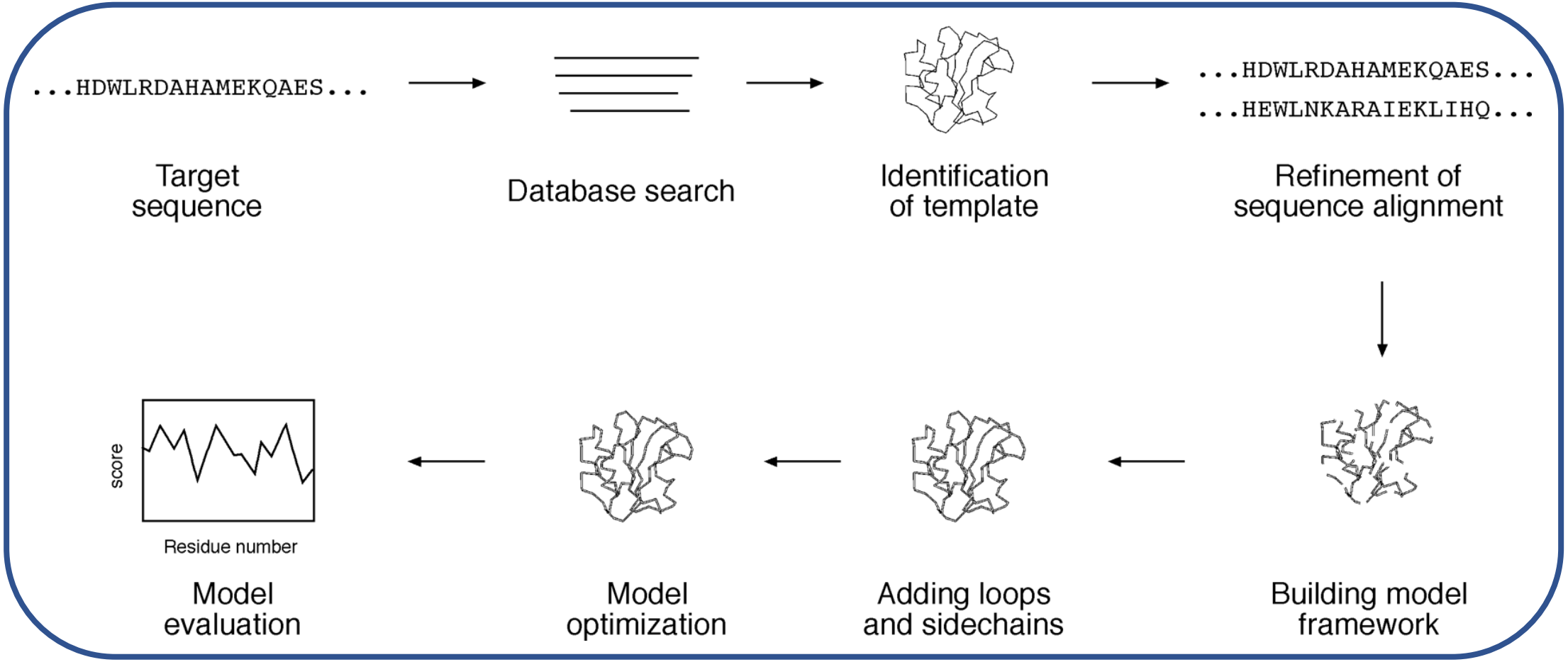
Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

### Homology modeling

Steps




Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

## Homology modeling

### Template Selection

Search Protein Data Bank (PDB) for existing experimental structures sharing high sequence identity.

PDB search is usually performed using BLAST.

Can use one or multiple templates.

Often structure(s) with the highest percentage identity and highest resolution is selected as a template.

As a rule of thumb, a database protein should have at least 30% sequence identity with the query sequence to be selected as template

### Sequence Alignment

The full-length sequences of the template and target proteins need to be aligned.

Most critical step in homology modeling, which directly affects the quality of the final model.

Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

## Homology modeling

Contd…

### Sequence Alignment

Incorrect alignment at this stage leads to incorrect designation of homologous residues and therefore to incorrect structural models.

Should be visually inspected to ensure that conserved key residues are correctly aligned.

If necessary, manual refinement of the alignment should be carried out to improve alignment quality.

### Backbone Model Building

Residues in the aligned regions of the target protein can assume a similar structure as the template proteins.

i.e. the coordinates of the corresponding residues of the template proteins can be simply copied onto the target protein.

If the two aligned residues are identical, coordinates of the side chain atoms are copied along with the main chain atoms.

Source: Essential Bioinformatics by Jin Xiong

# Methods

## Homology modeling

Contd…

### Backbone Model Building

If the two residues differ, only the backbone atoms can be copied. The side chain atoms are rebuilt in a subsequent procedure.

### Loop Modeling

In the sequence alignment for modeling, there are often regions caused by insertions and deletions producing gaps in sequence alignment.

The gaps cannot be directly modeled, creating “holes” in the model.

Closing the gaps requires loop modeling, which is a very difficult problem in homology modeling and is also a major source of error.

Source: Essential Bioinformatics by Jin Xiong

# Methods

## Homology modeling

Contd…

### Side Chain Refinement

Once main chain atoms are built, the positions of side chains that are not modeled are determined.

Modeling side chain geometry is very important in evaluating protein–ligand interactions at active sites and protein–protein interactions at the contact interface.

Most current side chain prediction programs use the concept of rotamers, which are favored side chain torsion angles extracted from known protein crystal structures.

A collection of preferred side chain conformations is a rotamer library in which the rotamers are ranked by their frequency of occurrence.

In prediction of side chain conformation, only the possible rotamers with the lowest interaction energy with nearby atoms are selected.

# Protein 3D structure prediction

## Methods

## Homology modeling

Contd…

### Model Refinement

Refined structural irregularities such as unfavorable bond angles, bond lengths, or close atomic contacts.

Corrected by applying the energy minimization procedure on the entire model, which moves the atoms in such a way that the overall conformation has the lowest energy potential.

The goal of energy minimization is to relieve steric collisions and strains without significantly altering the overall structure.

Another often used structure refinement procedure is molecular dynamic simulation.

### Model Evaluation

The final homology model is evaluated to make sure that the structural features of the model are consistent with the physicochemical rules.

This involves checking anomalies in ϕ–ψ angles (in Ramachandran plots), bond lengths, close contacts, and so on.

Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

### Homology modeling

### Popular tools for homology modeling

- Modeller
- SWISS-MODEL
- I-TASSER
- PRIME

Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

## Threading and fold recognition

### Principle:

**There are only small number of protein folds available, compared to millions of protein sequences.**

**This means that protein structures tend to be more conserved than protein sequences.**

**Consequently, many proteins can share a similar fold even in the absence of sequence similarities.**

**This forms the principle to predict protein structures beyond sequence similarities.**

Source: Essential Bioinformatics by Jin Xiong

# Methods

## Threading and fold recognition

### Definition:

**Threading or structural fold recognition predicts the structural fold of an unknown protein sequence by fitting the sequence into a structural database and selecting the best-fitting fold.**

**The comparison emphasizes matching of secondary structures, which are most evolutionarily conserved.**

**Therefore, this approach can identify structurally similar proteins even without detectable sequence similarity.**

Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

## Threading and fold recognition

**Pairwise energy based** → **Threading**

**Profile based** → **Fold Recognition**

Both terms are often used interchangeably

A protein sequence is searched for in a structural fold database to find the best matching structural fold using energy-based criteria.

Steps in Threading

1. Align the query sequence with each structural fold (at the sequence profile level).
2. Adjust the local alignment to get lower energy and thus better fitting.
3. Build a crude model for the target sequence by replacing aligned residues in the template structure with the corresponding residues in the query.
4. Calculate the energy terms of the raw model.
5. Rank models based on the energy terms to find the lowest energy fold.

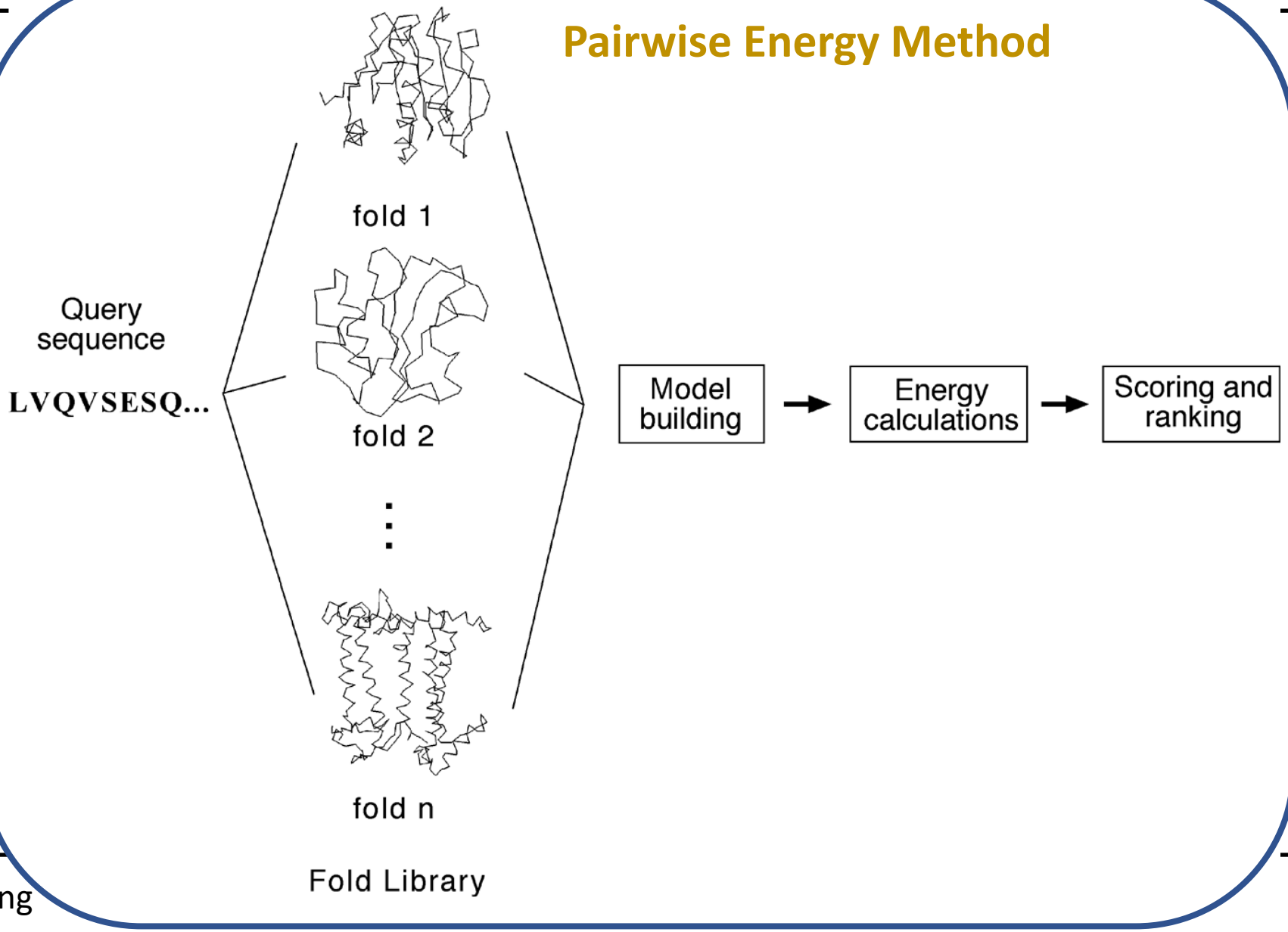
Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

**Steps in Threading**

### Pairwise Energy Method




Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

## Threading and fold recognition

**Pairwise energy based** → **Threading**

**Profile based** → **Fold Recognition**

Both terms are often used interchangeably

A structural profile is constructed for a group of related protein structures.

### Steps in Fold Recognition

- The structural profile is generated by superimposition of the structures to expose corresponding residues.
- The profile contains scores that describe the propensity of each of the twenty amino acid residues to be at each profile position.
- Scores contain information for secondary structural types, the degree of solvent exposure, polarity, and hydrophobicity of the amino acids.
- For the query sequence, secondary structure, solvent exposure and polarity are predicted.
- Compared with propensity profiles of known folds to find the fold that best represents the predicted profile.

Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

***Ab initio***

## Principle:

**If no suitable experimental structure (template) exists in the database, then homology modelling and threading will not work.**

**However, proteins in nature fold on their own without checking what the structures of their homologs are in databases.**

**There is some information in the sequences that provides instruction for the proteins to "find" their native structures.**

**Most proteins fold spontaneously into a stable structure that has near minimum energy known as native state.**

**This folding process appears to be nonrandom; however, its mechanism is poorly understood.**

Source: Essential Bioinformatics by Jin Xiong

# Protein 3D structure prediction

## Methods

### *Ab initio*

### Steps:

The *ab initio* prediction method attempts to produce all-atom protein models based on sequence information alone without the aid of known protein structures.

The perceived advantage of this method is that predictions are not restricted by known folds and that novel protein folds can be identified.

However, because the physicochemical laws governing protein folding are not yet well understood, the energy functions used in the *ab initio* prediction are at present rather inaccurate.

The folding problem remains one of the greatest challenges in Bioinformatics today.

Because the native state of a protein structure is near energy minimum, the prediction programs are thus designed using the energy minimization principle.

Searching for all possible structural conformations is not yet computationally feasible.

Source: Essential Bioinformatics by Jin Xiong

# Thank you

## Wish you all the best!

**Source:** https://www.cleanpng.com/png-smiley-clip-art-emoticon-happiness-megadeth-dave-t-6671171/download-png.html