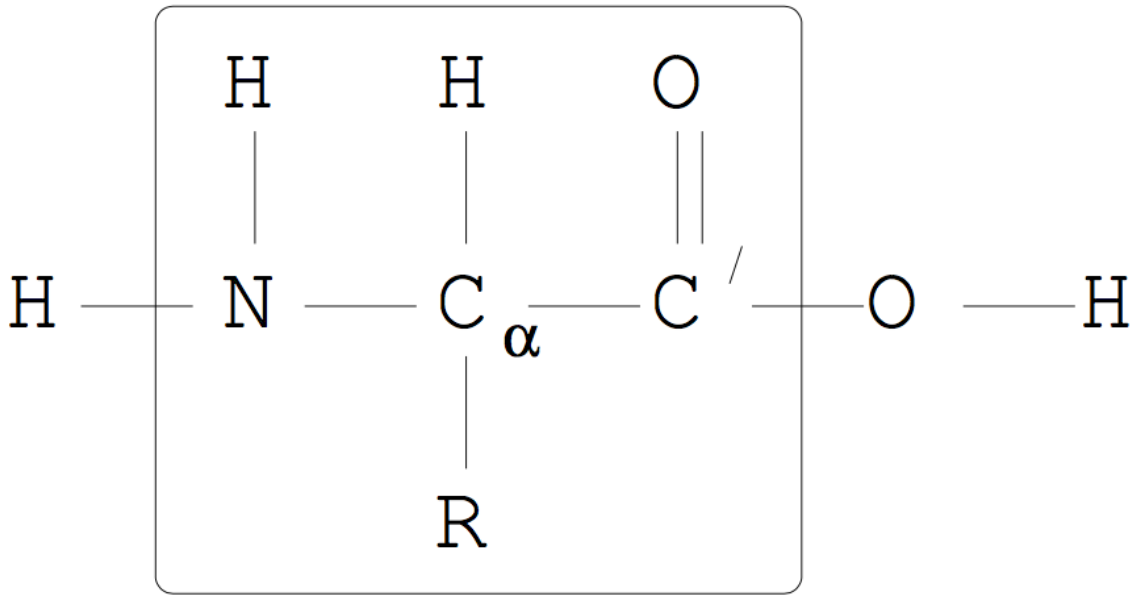
Protein Docking

Peter Mühlbacher

June 23, 2015

What Are Amino Acids?

Figure: Illustration of an amino acid. [Neumaier, 2006]



- amino acids polymerize in a specific sequence to a chain
- the repeating N, C_α, C-pattern is called the protein's backbone

Structures

Primary Structure

amino acid sequence

Secondary Structure

spatial arrangement of amino acid residues that are nearby in the sequence

Tertiary Structure

its three-dimensional structure, as defined by the atomic coordinates

Quaternary Structure

spatial arrangement of multiple folded proteins and the nature of their interactions [Berg, Tymoczko, Stryer, 2002]

What Is Protein Docking And Why Is It Important?

The Problem

- given tertiary structures, find the most likely quaternary structure
- evaluate its affinity

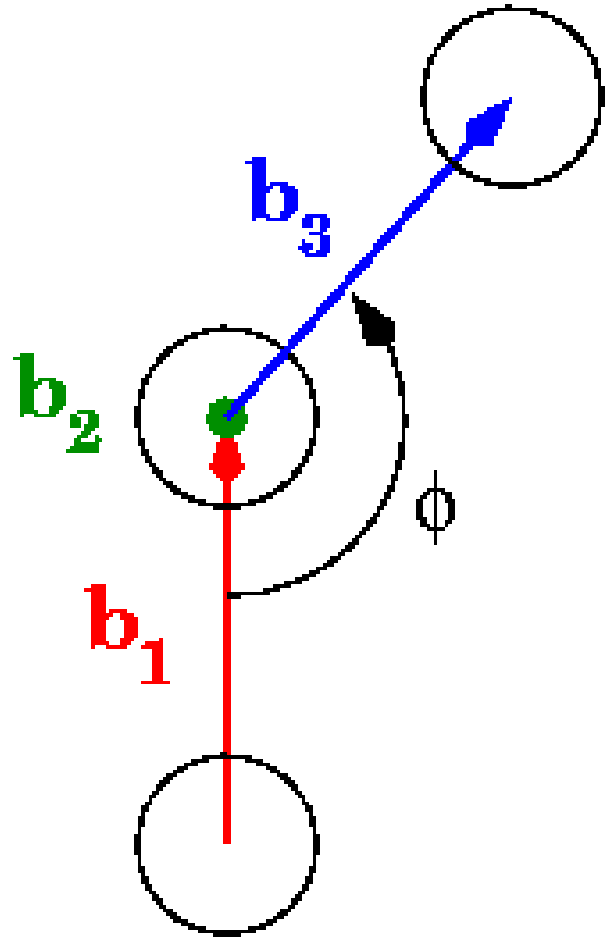
Examples For Quaternary Structures

- hemoglobin
- DNA polymerase
- ion channels

Understanding how proteins interact enables drug design.

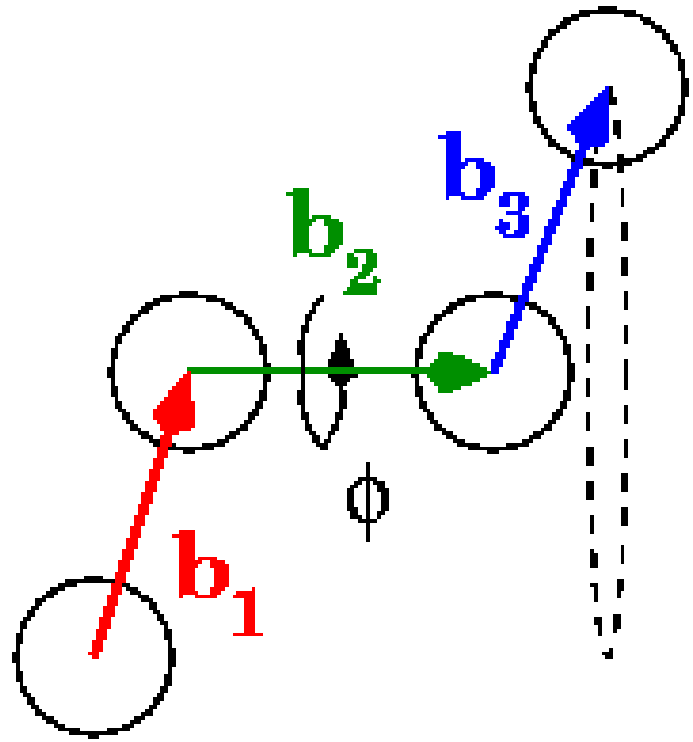
Vocabulary

- bond angle



Vocabulary

- bond angle
- dihedral angle



Vocabulary

- bond angle
- dihedral angle
- internal coordinates

set of bond lenghts, bond angles and dihedral angles; usually only some dihedral angles are allowed to vary though; protein can be seen as a tree-like graph structure

useful to alter a molecule in a chemically meaningful way

Vocabulary

- bond angle
- dihedral angle
- internal coordinates
- absolute coordinates

set of cartesian coordinates of every single atom in a molecule

useful to calculate potentials

Different Approaches

The problem can be casted as a minimization problem: $\operatorname{argmin}_X U(X)$. One can differentiate between several approaches.

Classes of Potential Functions

- **Force fields:** Typically, the potential U is a sum over bond, angle, dihedral angle, electrostatic and van der Waals energies.
- **Knowledge-based/empirical methods:** Compare segments with experimentally determined data.

Choosing a Set of Parameters

- **Rigid docking:** The parameters X consist of translation and rotation of the smaller protein in $\mathbb{R}^3$.
- **Flexible docking:** In addition to the 6 parameters of rigid docking, internal parameters (mostly dihedral angles, as bond angles and lengths are relatively stable) are used.

The Potential Function In Use

- let $X = x, y, z, \alpha, \beta, \gamma, \theta_1, \ldots, \theta_n$ free parameters,
- $M^1(X) = M^1$ be the flexible and M^2 the fixed molecule,
- LJ_{M_i, M_j} the Lennard-Jones potential, modelling pairwise van der Waals forces, depending on the types of the atoms M_i, M_j and
- $\|M_i - M_j\|$ the euclidian distance of their cartesian coordinates.

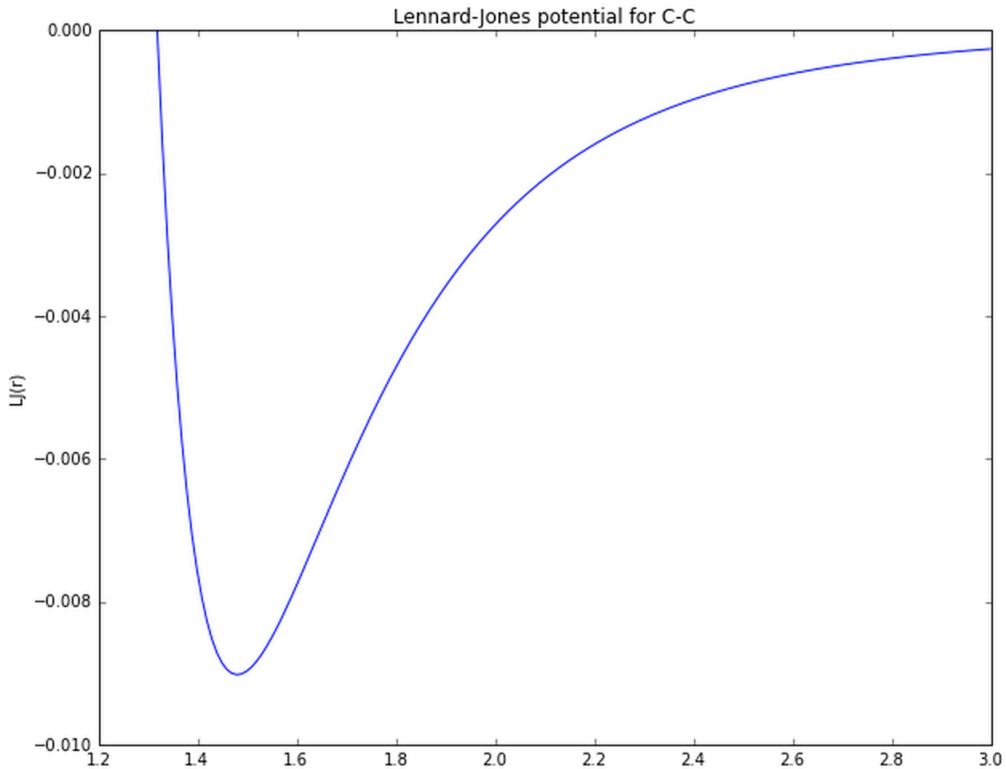
$$U(X) = \underbrace{\sum_{i=1}^{|M^1|} \sum_{j>i}^{|M^1|} LJ_{M_i^1, M_j^1}(\|M_i^1 - M_j^1\|)}_{\text{internal energy of } M^1} + \underbrace{\sum_{i=1}^{|M^1|} \sum_{j=1}^{|M^2|} LJ_{M_i^1, M_j^2}(\|M_i^1 - M_j^2\|)}_{\text{interaction of } M^1 \text{ and } M^2}$$

Physical Interpretation of LJ

Setting

$$LJ_{M_i,M_j}(r) = \frac{A_{M_i,M_j}}{r^{12}} - \frac{B_{M_i,M_j}}{r^6}$$

for $A_{M_i,M_j}, B_{M_i,M_j} \in \mathbb{R}^+$ yields the following with a strong divergence at $r = 0$.



Physical Interpretation of LJ

Setting

$$LJ_{M_i,M_j}(r) = \frac{A_{M_i,M_j}}{r^{12}} - \frac{B_{M_i,M_j}}{r^6}$$

for $A_{M_i,M_j}, B_{M_i,M_j} \in \mathbb{R}^+$ leads to the following interpretation:

- A corresponds to the strength of the Pauli-repulsion.
- B corresponds to the attractive long-range term.
- $\varepsilon := \min_r LJ(r) = \frac{B^2}{4A}$ is the depth of the potential well.
- $r_m := \operatorname{argmin}_r LJ(r) = \sqrt[6]{2\frac{A}{B}}$ determines the equilibrium distance of the two elements M_i, M_j.

Meaningfulness of the Potential Function

weaknesses

- Various other physical properties are neglected, e.g. dipole-dipole interactions leading to secondary structures like the α-helix.

strengths

Meaningfulness of the Potential Function

weaknesses

- Various other physical properties are neglected, e.g. dipole-dipole interactions leading to secondary structures like the α-helix.

strengths

- Speed.

Meaningfulness of the Potential Function

weaknesses

- Various other physical properties are neglected, e.g. dipole-dipole interactions leading to secondary structures like the α-helix.

"strengths"

- Speed.
- All models are simply fitted mathematical functions $\rightarrow$ treating $C, C_\alpha, O, N, \ldots$ as different elements might account for other physical properties to some extent.

Meaningfulness of the Potential Function

weaknesses

- Various other physical properties are neglected, e.g. dipole-dipole interactions leading to secondary structures like the α-helix.

"strengths"

- Speed.
- All models are simply fitted mathematical functions $\rightarrow$ treating $C, C_\alpha, O, N, \ldots$ as different elements might account for other physical properties to some extent.
- It works to a certain extent, i.e. $\nabla_X U(X) \approx 0$ if X describes the initial configuration which should be an equilibrium state.

Finding Optimal Parameters

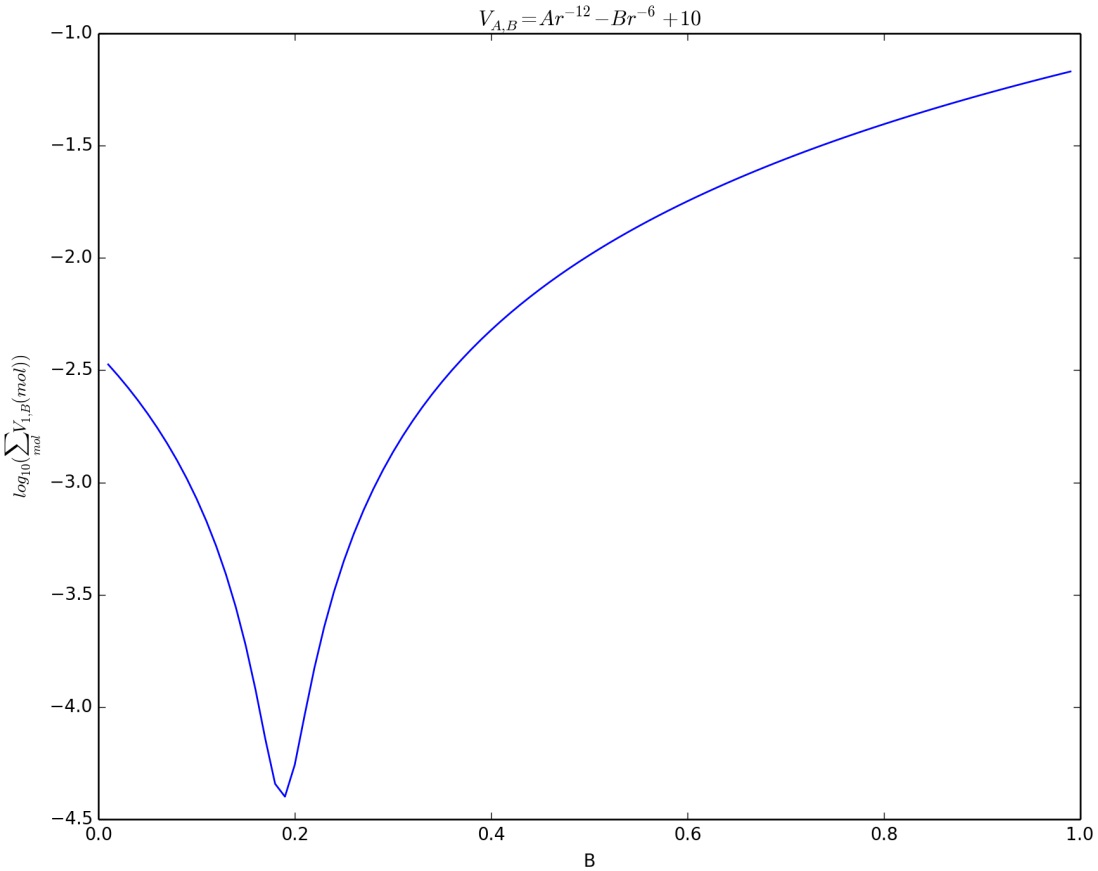
Observe that "M is in equilibrium" $\Leftrightarrow \nabla_X U_{A,B}(X_M, M) = 0$.
Thus, it is natural to choose $A, B \in (R^+)^{n \times n}$, where n is the number of different substances one wants to differentiate between, as

$$\underset{A,B}{\operatorname{argmin}} \sum_{M^i \in \text{training set}} \|\nabla_X U_{A,B}(X_{M^i}, M^i)\|.$$

However, this is an overdetermined system $\rightarrow$ fix $A_{11} = const$ and express all other parameters in terms of A_{11}, i.e.: $\tilde{A}_{ij} = A_{11}A_{ij}$, $\tilde{B}_{ij} = A_{11}B_{ij}$.

Finding Optimal Parameters

Figure: Projection of the optimization problem in $n(n+1)-1$ dimensions onto a 1-dimensional subspace (B_{11} is the only non-constant parameter, $A_{ij} = 1, B_{ij} = B_{11}$)



Numerical Challenges

Conversion Between Internal And Absolute Coordinates

While we want to write the potential as a function of the internal coordinates, it can only be calculated in terms of absolute coordinates $\rightarrow$ conversion is done often and should thus be implemented efficiently, i.e. minimize rounding errors and duration of computation

Computation of the Gradient

The potential is not simply a sum over rational functions as the distance between two atoms depends on the internal coordinates $\rightarrow$ the above conversion has to be included as well; as a result, computing the gradient explicitly is not only a tedious, but also an unnecessary task because of accumulating rounding errors $\rightarrow$ automatic backwards differentiation

The End

References

Arnold Neumaier (2006)

Molecular Modeling of Proteins

Berg JM, Tymoczko JL, Stryer L. (2002)

Biochemistry. 5th edition.