Structural biomathematics: an overview of molecular simulations and protein structure prediction

Bernat Anton

Figure: Parc de Recerca Biomèdica de Barcelona (PRBB).

Contents

All the biological information of the human body is encoded in our DNA. **Human Genome Project**: Sequentiation of the whole human genome completed on 2001, by Francis Collins (*Public Project*) & Craig Venter (*Celera Genomics*).

- About 3 billion base pairs (A, C, T and G).
- Estimation of 30000 genes (around 3000bp per gene).
- Less than 2% of the genome codes for proteins.
- Unknown function for over the half of the discovered genes!

All the biological information of the human body is encoded in our DNA. **Human Genome Project**: Sequentiation of the whole human genome completed on 2001, by Francis Collins (*Public Project*) & Craig Venter (*Celera Genomics*).

- About 3 billion base pairs (A, C, T and G).
- Estimation of 30000 genes (around 3000bp per gene).
- Less than 2% of the genome codes for proteins.
- Unknown function for over the half of the discovered genes!

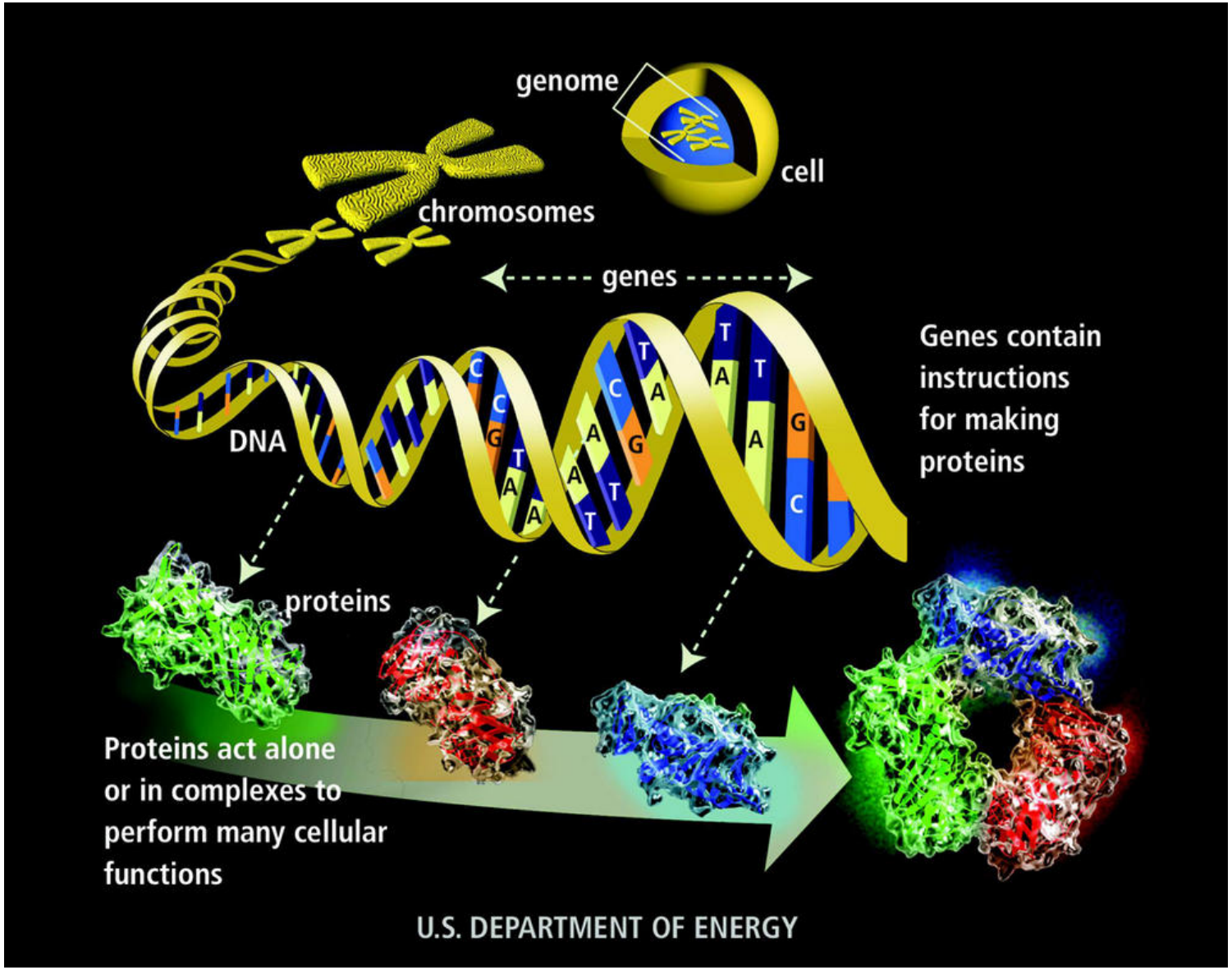
All the biological information of the human body is encoded in our DNA. **Human Genome Project**: Sequentiation of the whole human genome completed on 2001, by Francis Collins (*Public Project*) & Craig Venter (*Celera Genomics*).

- About 3 billion base pairs (A, C, T and G).
- Estimation of 30000 genes (around 3000bp per gene).
- Less than 2% of the genome codes for proteins.
- Unknown function for over the half of the discovered genes!

All the biological information of the human body is encoded in our DNA. **Human Genome Project**: Sequentiation of the whole human genome completed on 2001, by Francis Collins (*Public Project*) & Craig Venter (*Celera Genomics*).

- About 3 billion base pairs (A, C, T and G).
- Estimation of 30000 genes (around 3000bp per gene).
- Less than 2% of the genome codes for proteins.
- Unknown function for over the half of the discovered genes!

ATGAAAAAGACAGCTATCGCGATTGCAGTGGCACTGGCTGGTTTCGCTACCGTGGCC
CAGGCGGCCTCTGAGGGAAACAGTGACTGCTACTTTGGGAATGGGTCAGCCTACCG
TGGCACGCACAGCCTCACCGAGTCGGGTGCCTCCTGCCTCCCGTGGAATTCCATGAT
CCTGATAGGCAAGGTTTACACAGCACAGAACCCCAGTGCCCAGGCACTGGGCCTGG
GCAAACATAATTACTGCCGGAATCCTGATGGGGATGCCAAGCCCTGGTGCCACGTG
CTGAAGAACCGCAGGCTGACGTGGGAGTACTGTGATGTGCCCTCCTGCTCCACCTGC
GGCCTGAGACAGTACAGCCAGCCTCAGTTTCGCATCAAAGGAGGGCTCTTCGCCGA
CATCGCCTCCCACCCCTGGCAGGCTGCCATCTTTGCCAAGCACAGGAGGTCGCCCGG
AGAGCGGTTCCTGTGCGGGGGCATACTCATCAGCTCCTGCTGGATTCTCTCTGCCGC
CCACTGCTTCCAGGAGAGGTTTCCGCCCCACCACCTGACGGTGATCTTGGGCAGAAC
ATACCGGGTGGTCCCTGGCGAGGAGGAGCAGAAATTTGAAGTCGAAAAATACATTG
TCCATAAGGAATTCGATGATGACACTTACGACAATGACATTGCGCTGCTGCAGCTGA
AATCGGATTCGTCCCGCTGTGCCCAGGAGAGCAGCGTGGTCCGCACTGTGTGCCTTC
CCCCGGCGGACCTGCAGCTGCCGGACTGGACGGAGTGTGAGCTCTCCGGCTACGGC
AAGCATGAGGCCTTGTCTCCTTTCTATTCGGAGCGGCTGAAGGAGGCTCATGTCAGA
CTGTACCCATCCAGCCGCTGCACATCACAACATTTACTTAACAGAACAGTCACCGAC
AACATGCTGTGTGCTGGAGACACTCGGAGCGGCGGGCCCCAGGCAAACTTGCACGA
CGCCTGCCAGGGCGATTCGGGAGGCCCCCTGGTGTGTCTGAACGATGGCCGCATGA
CTTTGGTGGGCATCATCAGCTGGGGCCTGGGCTGTGGACAGAAGGATGTCCCGGGT
GTGTACACAAAGGTTACCAACTACCTAGACTGGATTCGTGACAACATGCGACCG
(SEQ ID NO:2)


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

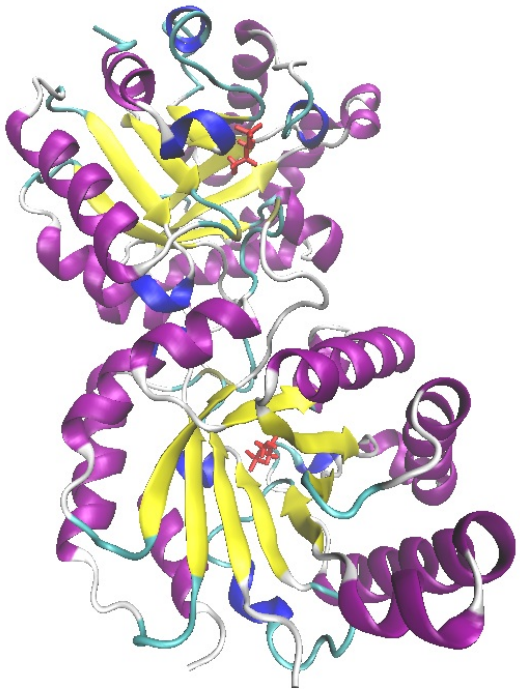
$$\text{DNA} \xrightarrow{\text{Transcription}} \text{RNA} \xrightarrow{\text{Translation}} \text{Protein}$$

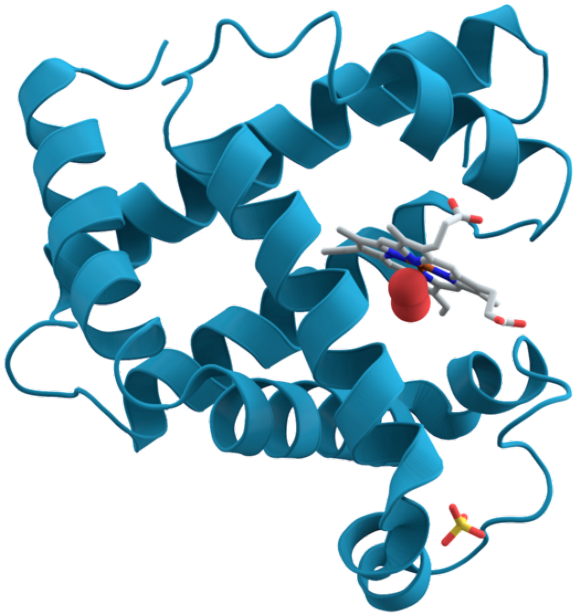
Standard genetic code

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA	Stop (*Ochre*)	UGA	Stop (*Opal*)	A
	UUG		UCG		UAG	Stop (*Amber*)	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

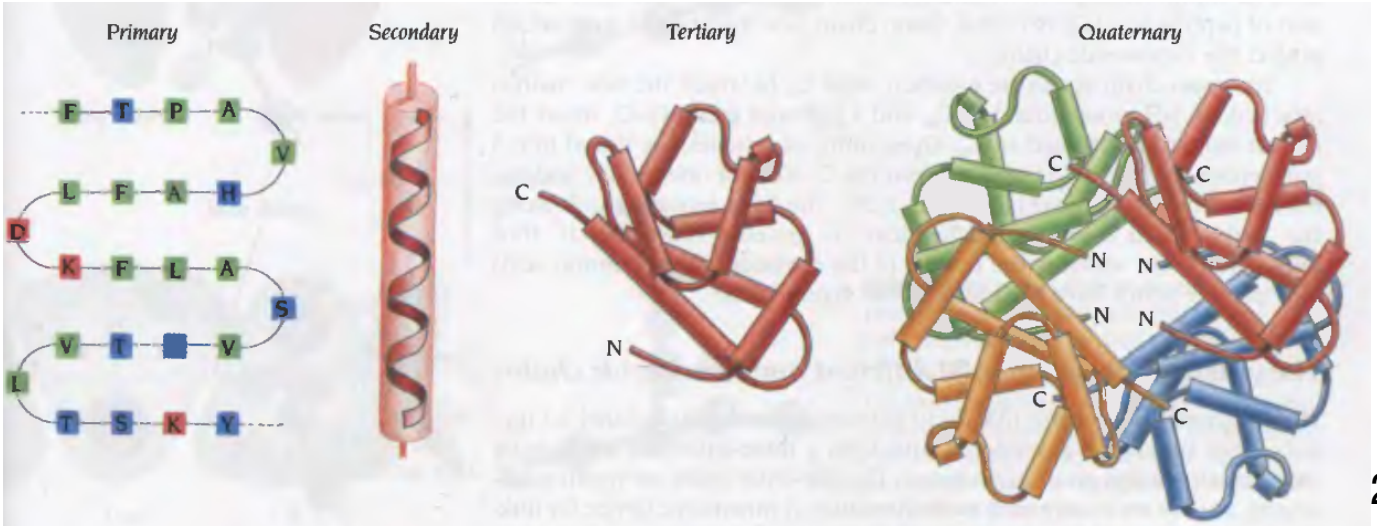
1

[1] Table taken from *Wikipedia*.





$$\text{Protein structure} \xrightarrow{???} \text{Protein function} \longrightarrow \text{Gene function}$$



[2]

- **Primary:** Amino acid linear sequence.
- **Secondary:** α-helices and β-strands.
- **Tertiary** / **Domains:** Functionally independent part of the sequence.
- **Quaternary:** Multi-subunit complex of domains or proteins.

[2] Figure taken from C.Branden & J.Tooze, *Introduction to Protein Structure*.

Main question:

How can we find the structure of a given protein?

- Crystallography.
- Nuclear magnetic resonance spectroscopy.
- Molecular simulation.
- Prediction of structure (structural biology).

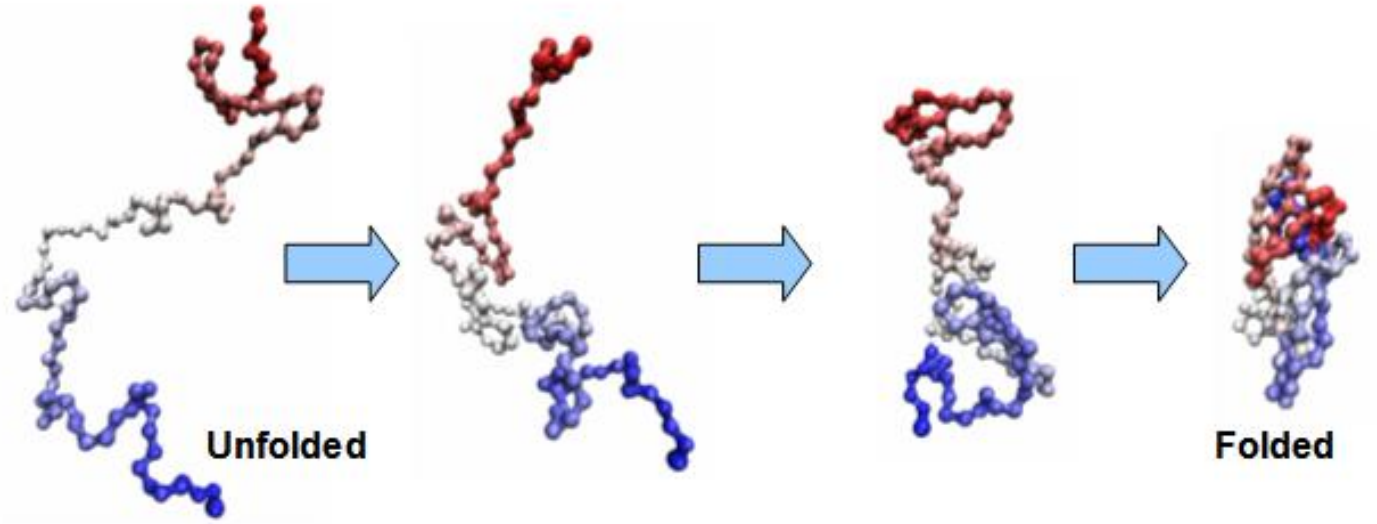
NOT AN EASY TASK!

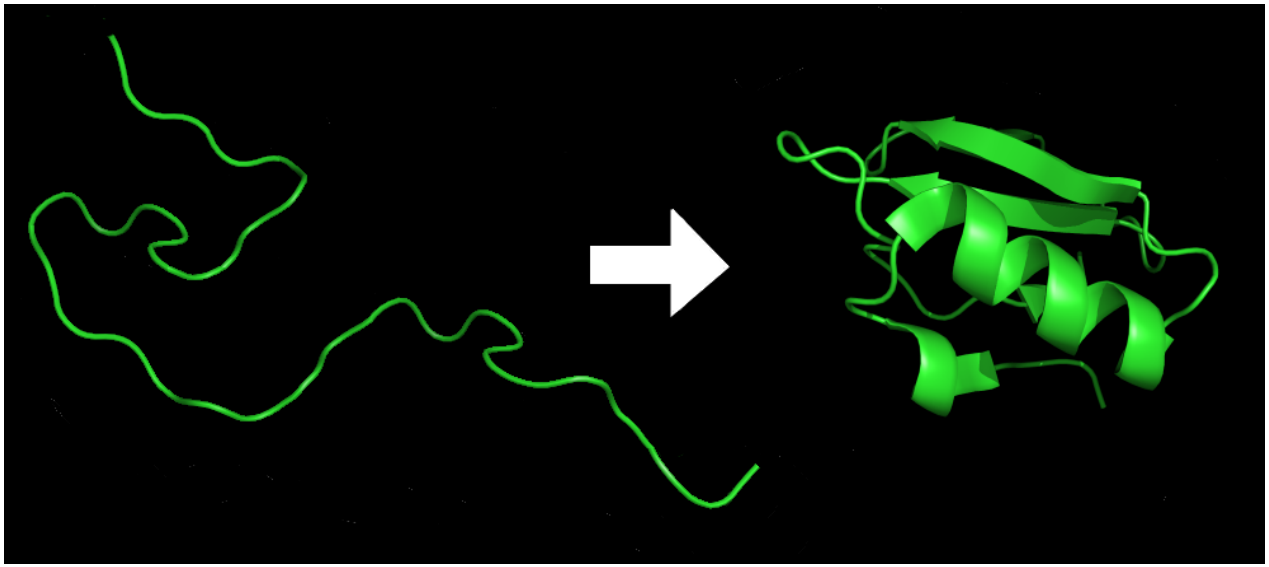
Main question:

How can we find the structure of a given protein?

- Crystallography.
- Nuclear magnetic resonance spectroscopy.
- Molecular simulation.
- Prediction of structure (structural biology).

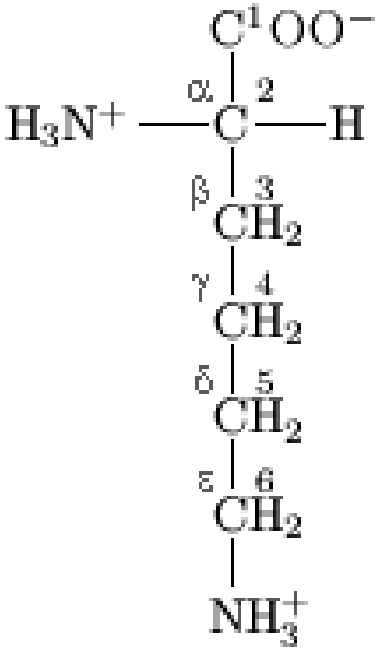
NOT AN EASY TASK!



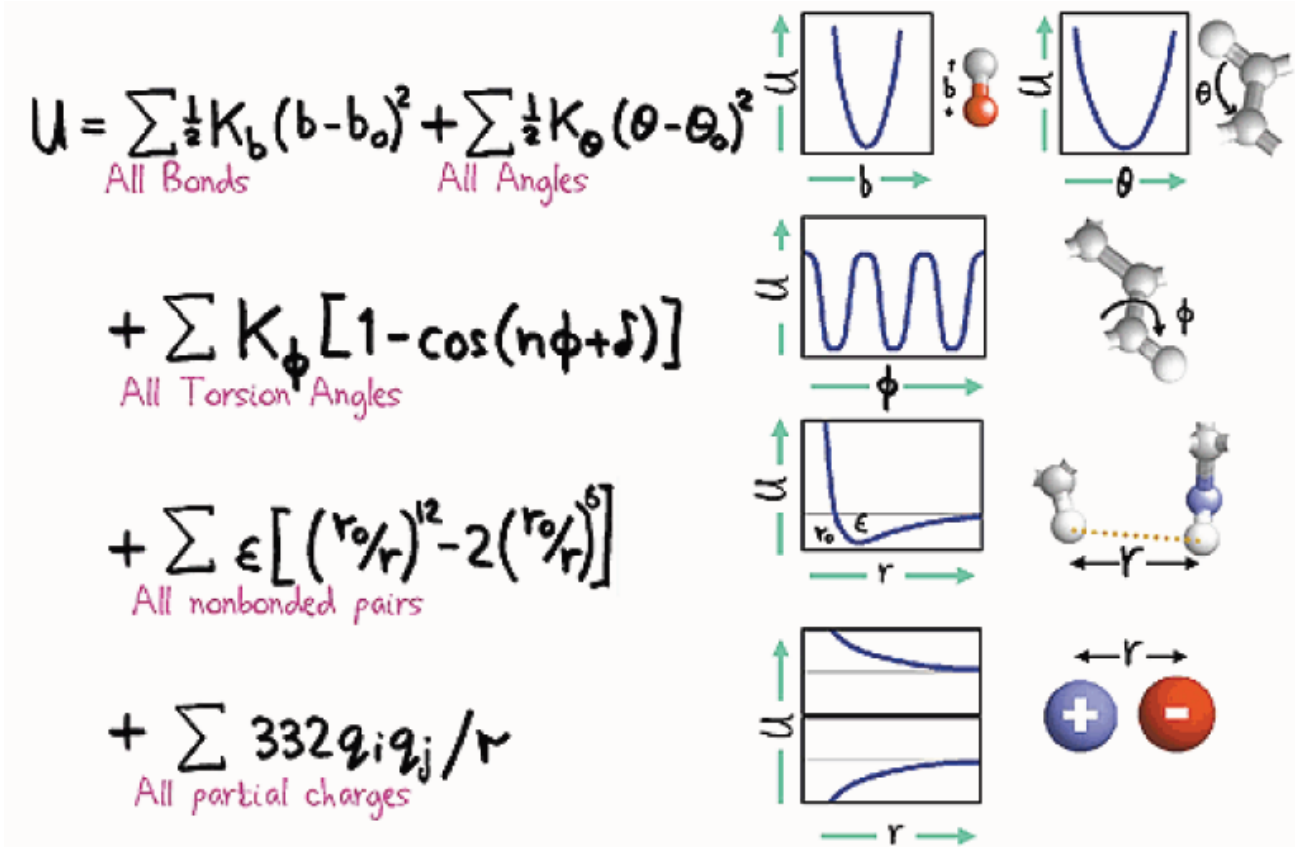


3

[3] Both images were obtained using VMD software



Lysine



Internal (mechanical) energy of the system

And these are not the only forces and energies implied in a molecular simulation!

PLC-β2 simulation

This simulation lasts around $20ns$, with timesteps of $4fs$[4], using the ACEMD software with the AMBER forcefield. The simulation has been visualized using VMD software.

The protein has 708 amino acids, for a total of around 150000 atoms in the simulation (counting water and lipid molecules).

In the simulation can be observed the folding of the X/Y linker in order to cover the hydrophobic active site of the protein.

[4] $1ns = 10^{-9}$seconds, $1fs = 10^{-15}$seconds

Afinsen's Dogma

The native structure of a protein is unique and is determined only by it's amino acid sequence. The folding to its native state is almost spontaneous.

Levinthal's Paradox

Due to the huge number of degrees of freedom in an unfolded protein, the number of possible conformations is astronomically large.

Then... how can proteins fold?

- Partially folded transition states.
- Funnel-like energy landscapes.
- ...?

Afinsen's Dogma

The native structure of a protein is unique and is determined only by it's amino acid sequence. The folding to its native state is almost spontaneous.

Levinthal's Paradox

Due to the huge number of degrees of freedom in an unfolded protein, the number of possible conformations is astronomically large.

Then... how can proteins fold?

- Partially folded transition states.
- Funnel-like energy landscapes.
- ...?

Let X, Y be two (discrete) random variables.

- The **(self-)information** of X is $I(X) = -log(P(X))$.
- The **entropy** of X is the measure of uncertainty associated with X: $S(X) = E(I(X))$.
- The **mutual information** of X and Y (also called **Kullback-Leibler divergence**) is

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) log \left(\frac{P(x,y)}{P(x)P(y)} \right)$$

Maximum Entropy Principle

Given a proposition that expresses testable information, the probability distribution that best represents the current state of knowledge is the one with largest entropy.

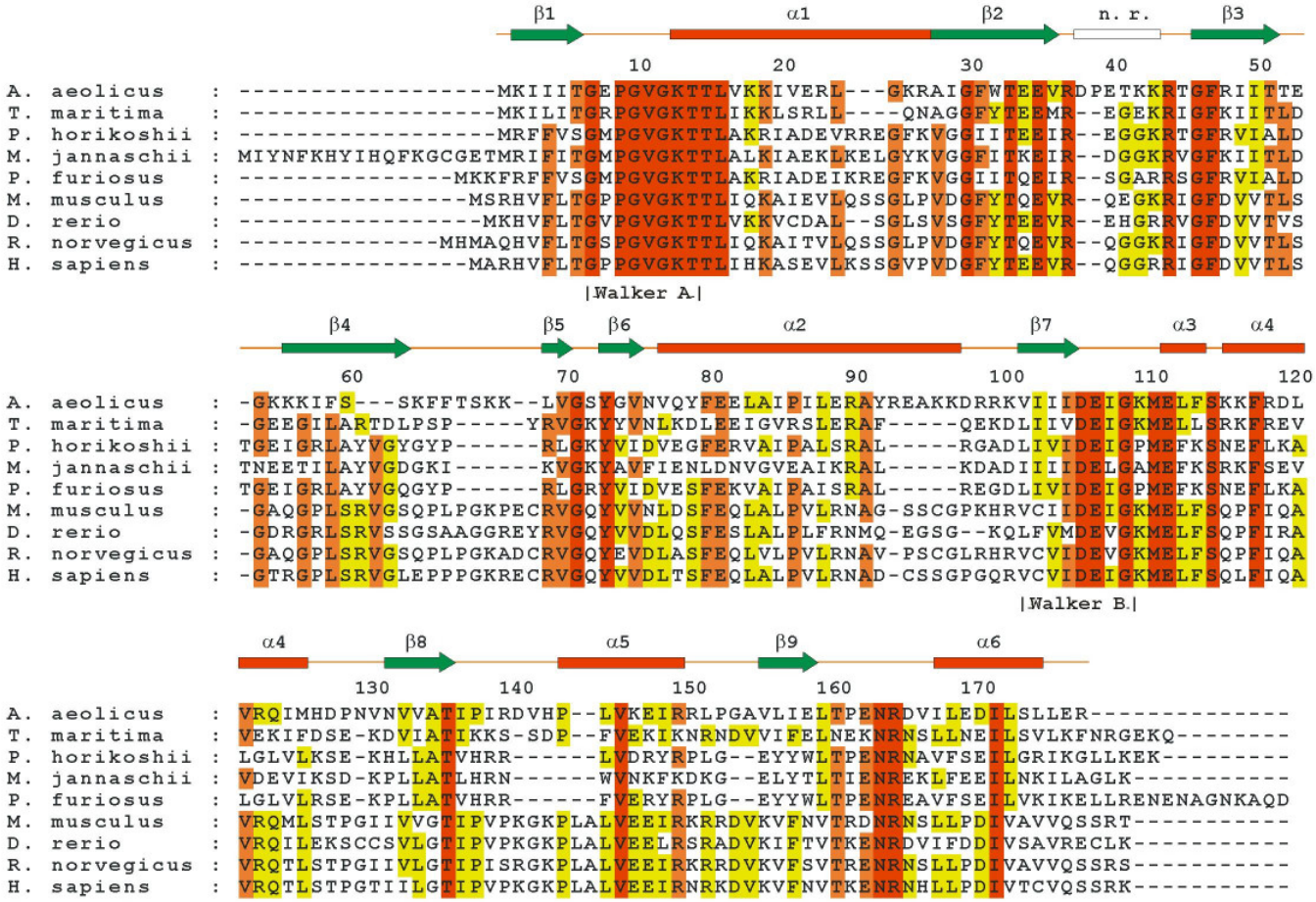


Figure: Multiple Sequence Alignment (MSA) for aaTHEP1.

From the previous MSA let's define:

$f_i(A)$ = 'frequency of apparitions of aa A in the position i of the MSA'

$f_{i,j}(A,B)$ = 'frequency of simultaneous apparitions of aa A and B in respective positions i and j of the MSA'

$$MI_{i,j} = \sum_{A,B} f_{i,j}(A,B) ln\left(\frac{f_{i,j}(A,B)}{f_i(A)f_j(B)}\right)$$

Be careful!

By definition, this mutual information of these frequencies is local in the amino acid chain, thus is noised by transitivity of correlations.

From the previous MSA let's define:

$$\begin{aligned} f_i(A) &= \text{'frequency of apparitions of aa A in the position i of the MSA'} \\ f_{i,j}(A,B) &= \text{'frequency of simultaneous apparitions of aa A and B in respective positions i and j of the MSA'} \end{aligned}$$

$$MI_{i,j} = \sum_{A,B} f_{i,j}(A,B) ln\left(\frac{f_{i,j}(A,B)}{f_i(A)f_j(B)}\right)$$

Be careful!

By definition, this mutual information of these frequencies is local in the amino acid chain, thus is noised by transitivity of correlations.

We want

$$P(A_1, \ldots, A_L)$$

a general model for the probability of a particular amino acid sequence $A_1 \ldots A_L$ to be member of the iso-structural family under consideration, and such that

$$\begin{aligned} P_i(A) &\approx f_i(A), \\ P_{i,j}(A,B) &\approx f_{i,j}(A,B), \end{aligned}$$

where

$$\begin{aligned} P_i(A) &= \sum_{A_k \neq A} P(A_1, \ldots, A_L), \\ P_{i,j}(A,B) &:= \sum_{A_k \neq A,B} P(A_1, \ldots, A_L). \end{aligned}$$

Many distributions satisfying this: Maximum Entropy Principle!!!!

We want

$$P(A_1, \ldots, A_L)$$

a general model for the probability of a particular amino acid sequence $A_1 \ldots A_L$ to be member of the iso-structural family under consideration, and such that

$$\begin{aligned} P_i(A) &\approx f_i(A), \\ P_{i,j}(A,B) &\approx f_{i,j}(A,B), \end{aligned}$$

where

$$\begin{aligned} P_i(A) &= \sum_{A_k \neq A} P(A_1, \ldots, A_L), \\ P_{i,j}(A,B) &:= \sum_{A_k \neq A,B} P(A_1, \ldots, A_L). \end{aligned}$$

Many distributions satisfying this: Maximum Entropy Principle!!!!

Optimization problem:

$$\begin{aligned} \text{maximize} \quad & S = -\sum_{A_i | i=1,\ldots,L} P(A_1,\ldots,A_L) ln P(A_1,\ldots,P_L) \\ \text{subject to} \quad & P_{i,j}(A,B) = f_{i,j}(A,B) \\ & P_i(A) = f_i(A) \end{aligned}$$

Solution: *disordered Q-state Potts model*

$$P(A_1,\ldots,A_L) = \frac{1}{\mathcal{Z}} exp\left\{ \sum_{1\leq i<j\leq L} e_{i,j}(A_i,A_j) + \sum_{1\leq i\leq L} h_i(A_i) \right\}$$

where:

- the parameters $e_{i,j}(A_i,A_j)$, $h_i(A_i)$ are the *Lagrange multipliers* of the system,
- $\mathcal{Z}$ is the normalization constant (*partition function*).

Optimization problem:

$$\begin{aligned} \text{maximize} \quad & S = -\sum_{A_i | i=1,\ldots,L} P(A_1,\ldots,A_L) ln P(A_1,\ldots,P_L) \\ \text{subject to} \quad & P_{i,j}(A,B) = f_{i,j}(A,B) \\ & P_i(A) = f_i(A) \end{aligned}$$

Solution: *disordered Q-state Potts model*

$$P(A_1,\ldots,A_L) = \frac{1}{\mathcal{Z}} exp\left\{ \sum_{1\leq i<j\leq L} e_{i,j}(A_i,A_j) + \sum_{1\leq i\leq L} h_i(A_i) \right\}$$

where:

- the parameters $e_{i,j}(A_i,A_j)$, $h_i(A_i)$ are the *Lagrange multipliers* of the system,
- $\mathcal{Z}$ is the normalization constant (*partition function*).

Geometrically, this probability distribution is given by the Boltzmann-Gibbs distribution:

$$P(A_1, \ldots, A_L) = \frac{1}{\mathcal{Z}} e^{-\mathcal{H}(A_1, \ldots, A_L)}$$

Formally, the marginals of this distribution are obtained from

$$\begin{aligned} \frac{\partial ln\mathcal{Z}}{\partial h_i(A)} &= -P_i(A) \\ \frac{\partial^2 ln\mathcal{Z}}{\partial h_i(A)\partial h_j(B)} &= -P_{i,j}(A,B) + P_i(A)P_j(B) \end{aligned}$$

but the direct computation is computationally prohibitive.

Geometrically, this probability distribution is given by the Boltzmann-Gibbs distribution:

$$P(A_1, \ldots, A_L) = \frac{1}{\mathcal{Z}} e^{-\mathcal{H}(A_1, \ldots, A_L)}$$

Formally, the marginals of this distribution are obtained from

$$\begin{aligned} \frac{\partial ln\mathcal{Z}}{\partial h_i(A)} &= -P_i(A) \\ \frac{\partial^2 ln\mathcal{Z}}{\partial h_i(A)\partial h_j(B)} &= -P_{i,j}(A,B) + P_i(A)P_j(B) \end{aligned}$$

but the direct computation is computationally prohibitive.

The Lagrange multipliers can be obtained using **Mean Field Aproximation** technique[5]:

- Introduce a new parameter α in the partition function (via the *disturbed Hamiltonian*):

$$\mathcal{H}(\alpha) = \sum_{i=1,\dots,L} exp\left\{\alpha \sum_{1\leq i<j\leq L} e_{i,j}(A_i, A_j) + \sum_{1\leq i\leq L} h_i(A_i)\right\}$$

- Consider the Legendre transform of the *Gibbs free energy* $\mathcal{F} = -ln\mathcal{Z}(\alpha)$ (*Gibbs potential*):

$$\mathcal{G}(\alpha) = ln\mathcal{Z}(\alpha) - \sum_{i=1,\dots,L} \sum_{A} h_i(A) P_i(A).$$

[5] Plefka, T., *Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model* (1982)

The Lagrange multipliers can be obtained using **Mean Field Aproximation** technique[5]:

- Introduce a new parameter α in the partition function (via the *disturbed Hamiltonian*):

$$\mathcal{H}(\alpha) = \sum_{i=1,\dots,L} exp\left\{\alpha \sum_{1\leq i<j\leq L} e_{i,j}(A_i, A_j) + \sum_{1\leq i\leq L} h_i(A_i)\right\}$$

- Consider the Legendre transform of the *Gibbs free energy* $\mathcal{F} = -ln\mathcal{Z}(\alpha)$ (*Gibbs potential*):

$$\mathcal{G}(\alpha) = ln\mathcal{Z}(\alpha) - \sum_{i=1,\dots,L} \sum_{A} h_i(A) P_i(A).$$

[5] Plefka, T., *Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model* (1982)

- Considering the *empirical connected correlation matrix*:

$$C_{i,j}(A,B) = f_{i,j}(A,B) - f_i(A)f_j(B).$$

As a consecuence of the functional form of the Legendre transform

$$\begin{aligned} h_i(A) &= \frac{\partial \mathcal{G}(\alpha)}{\partial P_i(A)} \\ (C^{-1})_{i,}(A,B) &= \frac{\partial h_i(A)}{\partial P_j(B)} = \frac{\partial^2 \mathcal{G}(\alpha)}{\partial P_i(A)\partial P_j(B)} \end{aligned}$$

- Expand the Gibbs potential up to first order Taylor expansion around $\alpha = 0$:

$$\mathcal{G}(\alpha) \approx \mathcal{G}(0) + \alpha \frac{\partial \mathcal{G}(\alpha)}{\partial \alpha}_{|\alpha=0}$$

A computation over the two terms of the Taylor expansion of $\mathcal{G}$ leads us to an expression which is easily derivable. First and second derivatives with respect the marginal distributions $P_i(A)$ provide self-consistent equations for the local fields, from which we obtain

$$(C^{-1})_{i,j}(A,B)_{|\alpha=0} = -e_{i,j}(A,B), \text{ for } i \neq j.$$

Finally, the parameters h_i can be estimated imposing empirical single-site frequency counts as marginal distributions and considering gauge conditions:

$$f_i(A) = \sum_B P_{i,j}(A,B).$$

A computation over the two terms of the Taylor expansion of $\mathcal{G}$ leads us to an expression which is easily derivable. First and second derivatives with respect the marginal distributions $P_i(A)$ provide self-consistent equations for the local fields, from which we obtain

$$(C^{-1})_{i,j}(A,B)_{|\alpha=0} = -e_{i,j}(A,B), \text{ for } i \neq j.$$

Finally, the parameters h_i can be estimated imposing empirical single-site frequency counts as marginal distributions and considering gauge conditions:

$$f_i(A) = \sum_B P_{i,j}(A,B).$$

This leads us to the effective pair probabilities

$$P_{i,j}^{Dir}(A,B) = \frac{1}{\mathcal{Z}_{i,j}} exp\left\{e_{i,j}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\right\}.$$

From which we can define its Kullback-Leibler divergence, that will be called **Direct Information**:

$$DI_{i,j} := \sum_{A,B} P_{i,j}^{Dir}(A,B) ln\left(\frac{P_{i,j}^{Dir}(A,B)}{f_i(A)f_j(B)}\right).$$

For each pair of positions in the sequence, we "know" if they are spatially close.

And now... what?

- Depending on the previous, not an unique folding for the protein is possible. We must remove knotted structures (*Alexander polynomial* or *Heegaard Floer homology*).
- A scoring method over the resulting foldings must be defined, in order to decide which one of the structures is better.
- A short simulation of the system may be run in order to optimize the energies of the folding.

For each pair of positions in the sequence, we "know" if they are spatially close.

And now... what?

- Depending on the previous, not an unique folding for the protein is possible. We must remove knotted structures (*Alexander polynomial* or *Heegaard Floer homology*).
- A scoring method over the resulting foldings must be defined, in order to decide which one of the structures is better.
- A short simulation of the system may be run in order to optimize the energies of the folding.

For each pair of positions in the sequence, we "know" if they are spatially close.

And now... what?

- Depending on the previous, not an unique folding for the protein is possible. We must remove knotted structures (*Alexander polynomial* or *Heegaard Floer homology*).
- A scoring method over the resulting foldings must be defined, in order to decide which one of the structures is better.
- A short simulation of the system may be run in order to optimize the energies of the folding.

For each pair of positions in the sequence, we "know" if they are spatially close.

And now... what?

- Depending on the previous, not an unique folding for the protein is possible. We must remove knotted structures (*Alexander polynomial* or *Heegaard Floer homology*).
- A scoring method over the resulting foldings must be defined, in order to decide which one of the structures is better.
- A short simulation of the system may be run in order to optimize the energies of the folding.

Figure: Chewbacca mounted on a squirrel wants to thank you for your assistance!