On the relevance of theoretical physics approaches for quantitative life sciences

Giorgio Parisi

In this talk I will shortly discuss:

- Why theoretical physics.
- The advantages of theoretical physics.
- A few examples.

Quantitative Biology

We have an overflow of data from all fields of biology. Just a few randomly chosen examples:

- Inside the cell: Genome, proteonome, metabolism.
- Many cells behaviour: We can record single cell movements during tissue developments.
- Collective movements of animals: we can measure the simultaneous movement of thousands of flocking birds with high accuracy (e.g. 10 cm.).

Why theoretical physics

We have to analyse huge set of data. We have to obtain conclusions extracting all the possible information from the data. Sometimes we are near the thermodynamic limit and concepts from statistica mechanics are relevant.

New tools coming from statistical mechanics are very useful in these problems.

A few examples:

- Powerful heuristic methods to the study of random optimisation problem (e.g. survey propagation method).
- Reverse statistical mechanics
  - Statistical Mechanics: You know the Hamiltonian and you have to compute the properties of the configurations.
  - Reverse statistical mechanics: You know some instances of the configurations and you have to compute the Hamiltonian.

The advantages of theoretical physics.

We have to make sense of the data. We want to get insight on their meaning. We need to compare the data with a model or better find a model that describe the data in a reasonable way.

The model must the simplest one, but not too simple.

Model building is an art we have to capture in the model the essence of the phenomena we study disregarding the inessential.

Universality

The collective (emergent) qualitative behaviour (and in same cases of the quantitative behaviour) of a large number of agents does not depend on the details of the interaction among agents.

There are wide Universality classes.

Universality

The collective (emergent) qualitative behaviour (and in same cases of the quantitative behaviour) of a large number of agents does not depend on the details of the interaction among agents.

There are larges Universality classes.

Knowing this principle helps to produce nice models without loosing too much time in introducing inessential complications, although in a few case it may lead to the construction of models based on spherical cows.

An example of a simple quantitative explanations (Ugo Bastolla).

If you compare different islands in the same archipelago (e.g. Galapagos) you find empirically that the number of different species belongs to the same group (e.g. birds) is roughly given by

$$NumberOfSpecies \propto Surface^{1/4}$$

Models are unable to explain these data if we consider islands isolated and speciation is the only source of diversity.

In presence of a sustained immigration the same model predicts

$$NumberOfSpecies \propto Immigration^{1/2} \qquad Immigration \propto Boundary \propto Surface^{1/2}$$

Hence

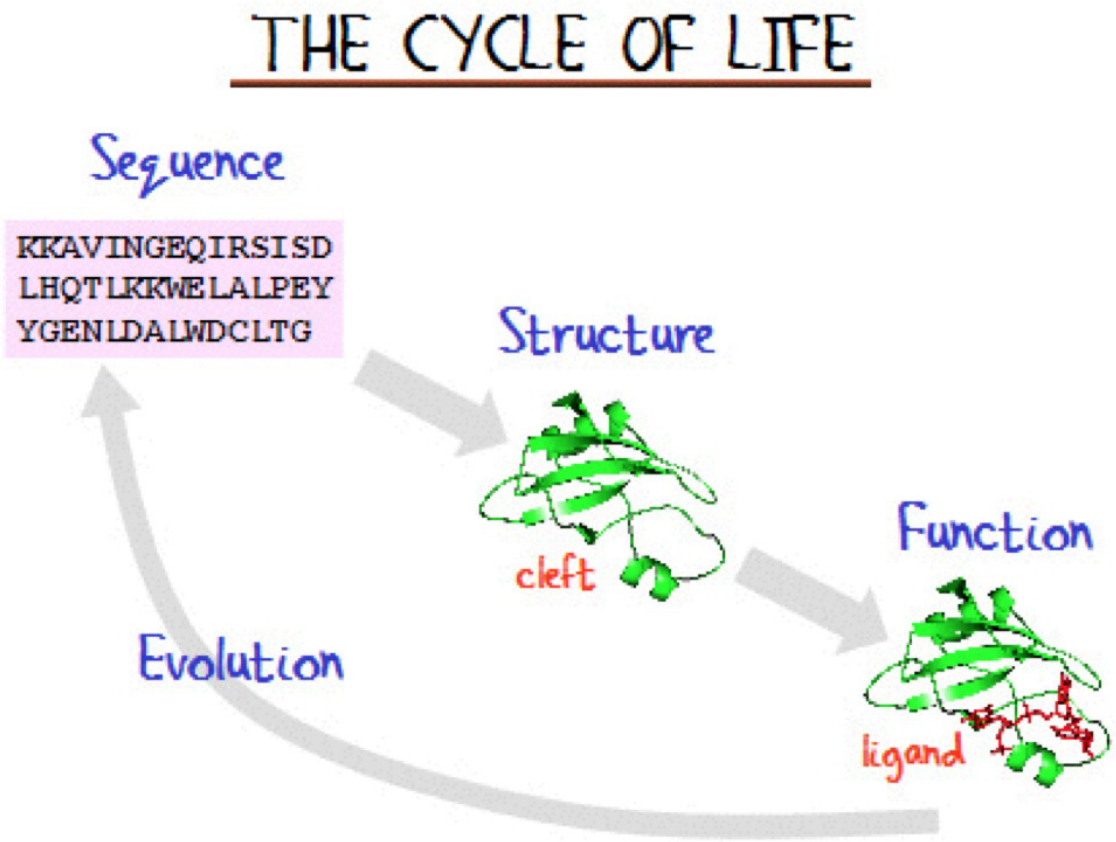
$$NumberOfSpecies \propto Surface^{1/4}$$

How to reconstruct the tertiary structure of a protein from its sequence, without doing long computer simulation of the folding and using a detailed information on the interactions among amino acids?

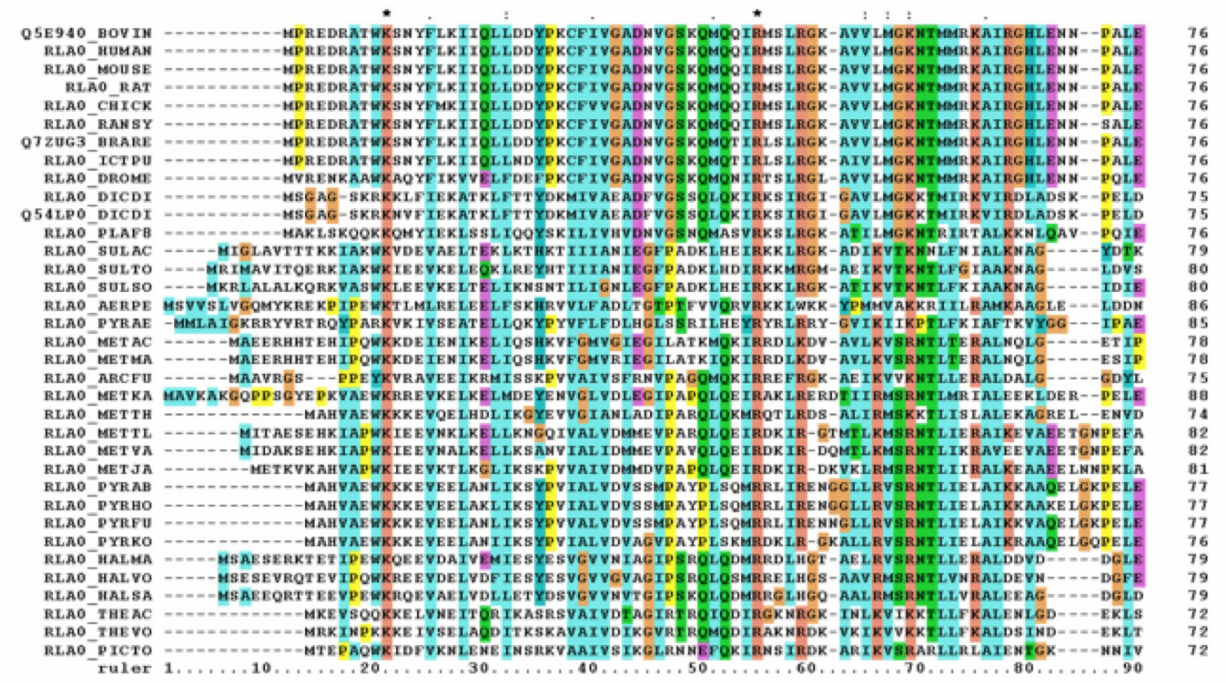
I will describe the work of Baldassi, Zamparo, Feinauer, Procaccini, Zecchina, Weigt, Pagnani

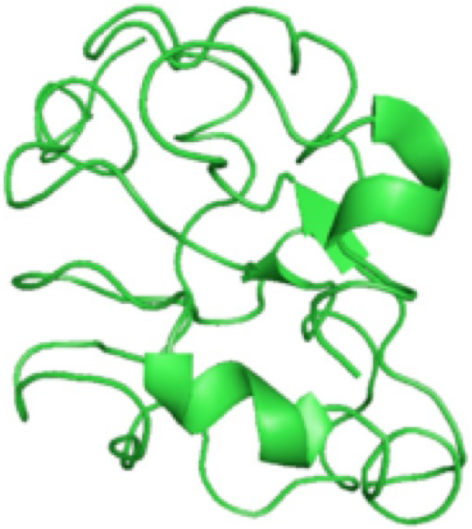
We can use the information coming from considering or the order of 10^2 different sequences to determinecontacts (i.e. aminoacids at the distance less that 7 Amstrongs).

Proteins are not just long molecules …



Inference of protein contacts from co-evolution



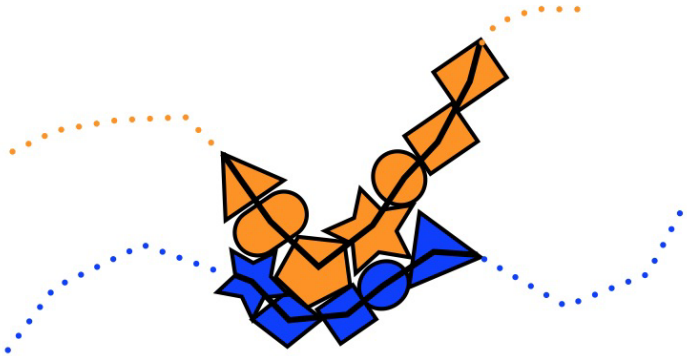


Ribosomal protein L10: from protopedia.org

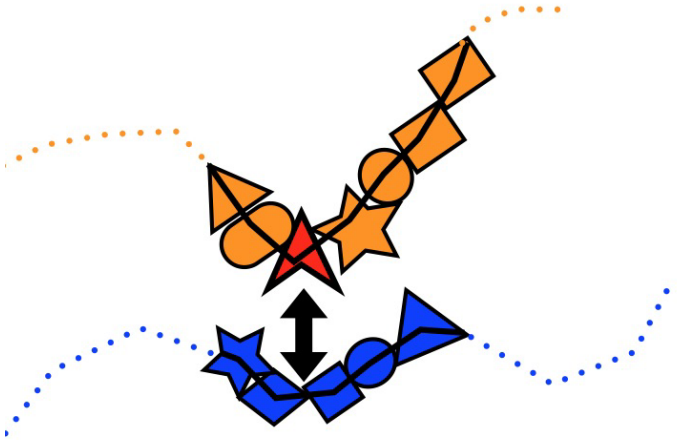
Sequence of homologous proteins (family) vary among species
Tertiary and quaternary structures are *virtually* the same

Random mutation are normally disruptive.
Mutations are highly correlated in order to preserve function

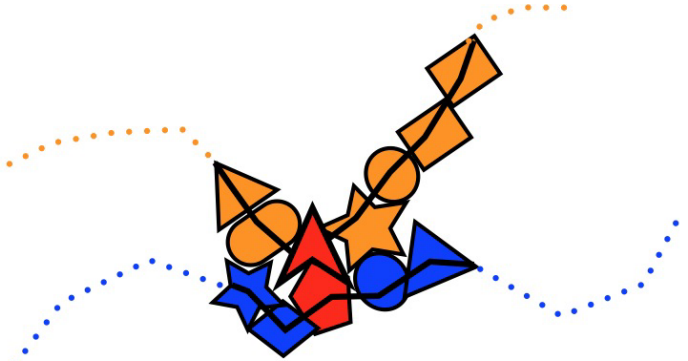
Residues proximity induces correlations



wild type



disruptive mutation



compensatory mutation

Up to what degree correlations unveil structure ?

One can measure the correlation. But correlation does not implies direct interaction.

A interacts with B, B interacts with C: A and C may not interact, but they are correlated because of B.

Method: we write the probability in the space of sequence.

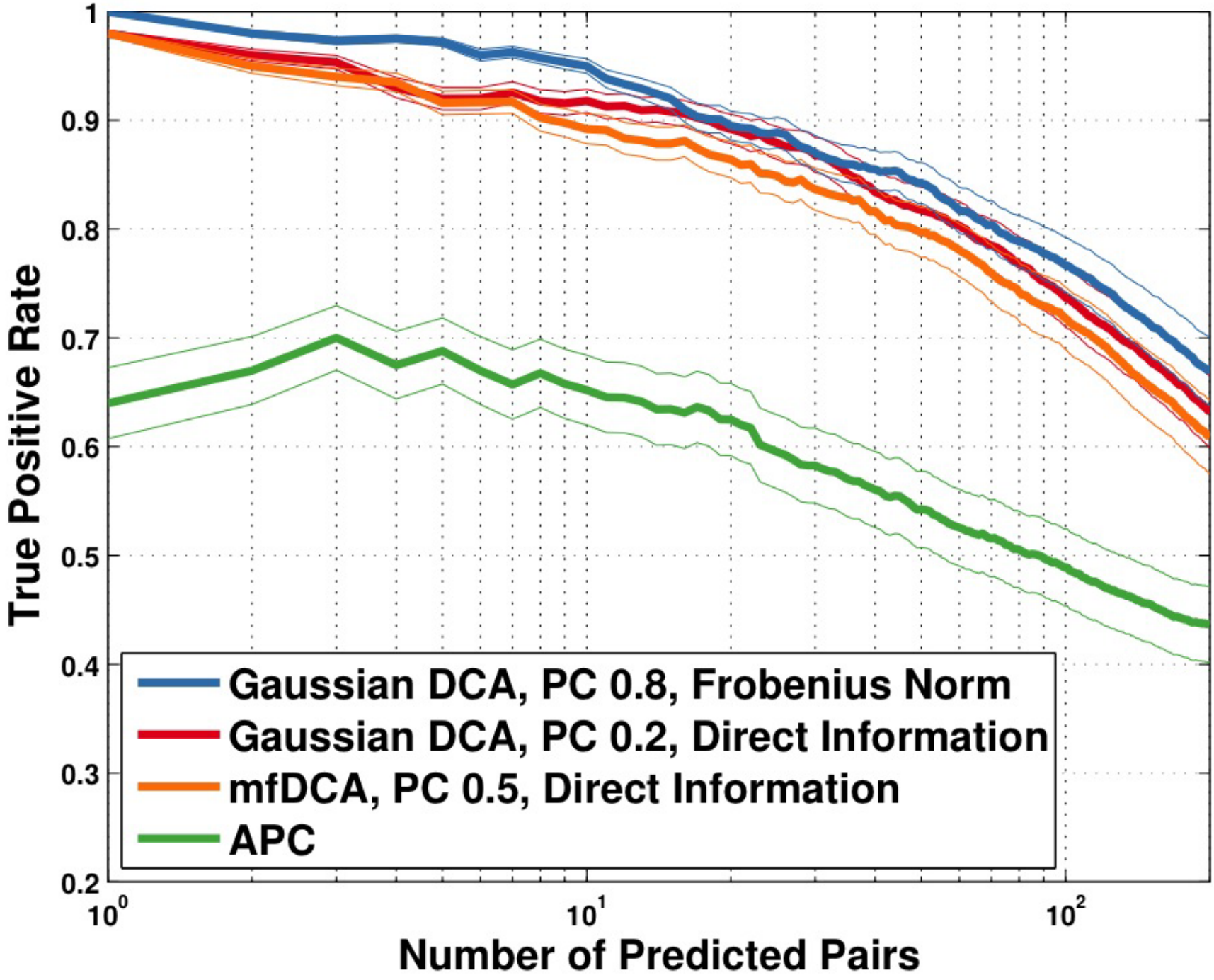
$$P(sequence) \propto \exp(-H(Sequence))$$

H reflect the evolutionary pressure and it is assumed to be of the form

$$H = \sum_{i,k} F_{i,k}(A_i, A_k)$$

F can be reconstructed from the sequence using techniques of *reverse* statistical mechanics.

The pairs with the largest F are mostly likely to be in contact.


True Positive Rate
Number of Predicted Pairs
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
10^0
10^1
10^2
Gaussian DCA, PC 0.8, Frobenius Norm
Gaussian DCA, PC 0.2, Direct Information
mfDCA, PC 0.5, Direct Information
APC