

**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, proteome, 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 statistical 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 some 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 some cases of the quantitative behaviour) of a large number of agents does not depend on the details of the interaction among agents.

There are large Universality classes.

Knowing this principle helps to produce nice models **without losing too much time in introducing inessential complications**, although in a few cases 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

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

Hence

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

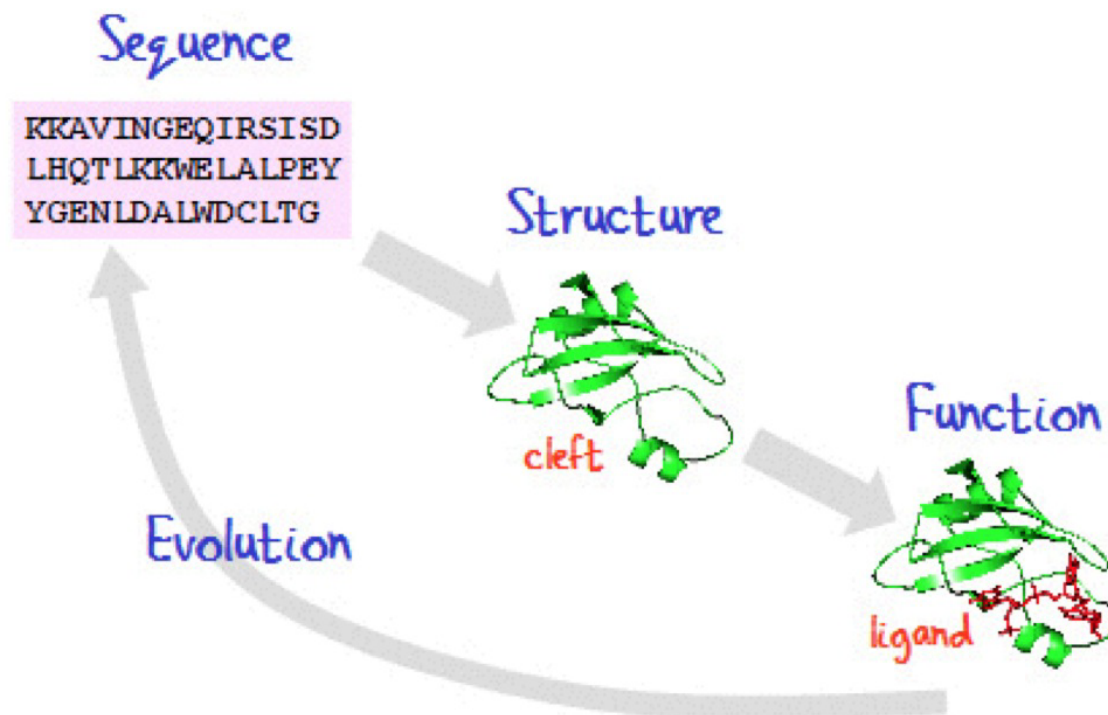
How to reconstruct the tertiary structure of a protein from its sequence, without doing long computer simulation of the folding and using a detailed information on the interactions among amino acids?

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 determine contacts (i.e. aminoacids at the distance less than 7 Amstrongs).

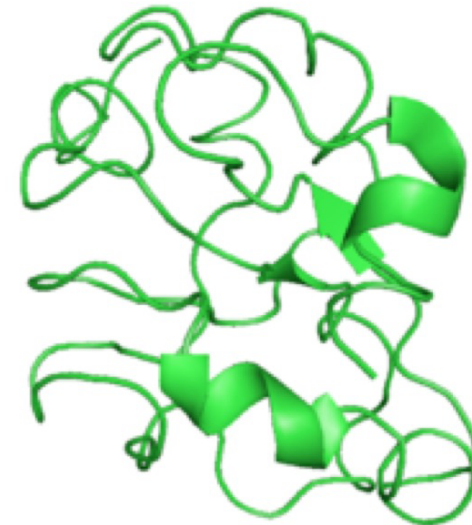
Proteins are not just long molecules ...

THE CYCLE OF LIFE



Inference of protein contacts from co-evolution

Q5E940_BOVIN	-----MPREDRATWKS	NYFLKIIQLLDDY	PKCFIVGADNVG	SKOMQIRMSLRGK	-AVVLMGKNTMMRKAIRG	HLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKS	NYFLKIIQLLDDY	PKCFIVGADNVG	SKOMQIRMSLRGK	-AVVLMGKNTMMRKAIRG	HLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKS	NYFLKIIQLLDDY	PKCFIVGADNVG	SKOMQIRMSLRGK	-AVVLMGKNTMMRKAIRG	HLENN--PALE	76
RLA0_RAT	-----MPREDRATWKS	NYFLKIIQLLDDY	PKCFIVGADNVG	SKOMQIRMSLRGK	-AVVLMGKNTMMRKAIRG	HLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKS	NYFLKIIQLLDDY	PKCFIVGADNVG	SKOMQIRMSLRGK	-AVVLMGKNTMMRKAIRG	HLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKS	NYFLKIIQLLDDY	PKCFIVGADNVG	SKOMQIRMSLRGK	-AVVLMGKNTMMRKAIRG	HLENN--SALE	76
Q7ZUG3_BRARE	-----MPREDRATWKS	NYFLKIIQLLDDY	PKCFIVGADNVG	SKOMQIRMSLRGK	-AVVLMGKNTMMRKAIRG	HLENN--PALE	76
RLA0 ICTPU	-----MPREDRATWKS	NYFLKIIQLLDDY	PKCFIVGADNVG	SKOMQIRMSLRGK	-AVVLMGKNTMMRKAIRG	HLENN--PALE	76
RLA0_DROME	-----MVRENKAANKAQY	FIKVVLEFDFPKCF	IVGADNVGSKOMQIR	MSLRGL-AVVLMGKNT	MMRKAIRG	HLENN--POLE	76
RLA0_DICDI	-----MSGAG-SKRK	KLFIEKATKLF	FTYDKMIVAEAD	FVGS-SOLQKIRKS	IRGI-GAVLMGKNT	MIRKVI	75
Q54LP0_DICDI	-----MSGAG-SKRK	NVFIKATKLF	FTYDKMIVAEAD	FVGS-SOLQKIRKS	IRGI-GAVLMGKNT	MIRKVI	75
RLA0_PLAFB	-----MAKLSKQ	KKQMYIEKLS	LIQQYSKILIV	HYDVGNSMASV	HKSLRGK-AEILMGKNT	RI	76
RLA0_SULAC	----MIGLAV	TTTKKIAK	KVDEVAELTE	KLKTHKTI	IIIANIEGFPAD	KLHEIRKKLRGK-ADIKV	79
RLA0_SULTO	----MRIMAV	ITQERK	IAKWKIEEYK	ELKREYHT	IIIANIEGFPAD	KLHEIRKKLRGK-AEIKV	80
RLA0_SULSO	----MKRLAL	LALKQK	VASWKLE	EYKELTEL	IKNSNTILIGN	LEGFPADKLHEIRKKLRGK-AEIKV	80
RLA0_AERPE	MSVVS	LVGQMYKREK	PIPEW	KTLMRELE	ELFSKIR	RVVLFADLTGPTFYVVRV	86
RLA0_PYRAE	MMLA	IGKRRYVRT	QYPA	RKVKIV	SEATELLQ	KVYVFLFDLHGLSRI	85
RLA0_METAC	-----MAEER	HTEHIPQ	WKKDEIEN	IKELIQSHK	VFGMVL	EGILATKMKQIRRD	78
RLA0_METMA	-----MAEER	HTEHIPQ	WKKDEIEN	IKELIQSHK	VFGMVL	EGILATKMKQIRRD	78
RLA0_ARCFU	-----MAAVR	GS--PDEY	KVRAVEE	IKRMIS	SKYVVAI	YSFRNVPA	75
RLA0_METKA	MAVK	KGGDPP	SGYE	PKVAE	WKRREYKEL	ELMDEYENVGLYDLE	88
RLA0_METTH	-----MAHVA	EWKKEVE	QLHDLIK	GYEVV	GIANLAD	IPAROLQKMRQT	74
RLA0_METTL	-----MITAE	SEHKIAP	WKIEE	VNKLKELL	KNGQIVAL	VDMMEVPA	82
RLA0_METJA	-----MIDAK	SEHKIAP	WKIEE	VNKLKELL	KNSAN	VIALIDMMEVPA	82
RLA0_METVA	-----METK	VKAHVA	WKIEE	VKT	LKGLIK	SKYVVAIYDMDVPA	81
RLA0_PYRAB	-----MAHVA	EWKKEVE	EELANL	IKSYV	VIALYD	VSSMPAYPLSQM	77
RLA0_PYRHO	-----MAHVA	EWKKEVE	EELANL	IKSYV	VIALYD	VSSMPAYPLSQM	77
RLA0_PYRFU	-----MAHVA	EWKKEVE	EELANL	IKSYV	VIALYD	VSSMPAYPLSQM	77
RLA0_PYRKO	-----MAHVA	EWKKEVE	EELANL	IKSYV	VIALYD	VAGVAPYPLSK	76
RLA0_HALMA	MSAE	SERKTET	IPWK	QEEYDAIV	MIESY	SVGVVNIAGIPSR	79
RLA0_HALVO	MSAE	SEVRQTE	IPQ	WREVEY	DELVD	FIESY	79
RLA0_HALSA	MSAE	EQRTTE	IPWK	QREYAE	LVDLLET	YDSGVVNVYTGIP	79
RLA0_THEAC	-----MKE	VSQK	KELYNE	ITRIK	ASRSVA	YDLAGIR	72
RLA0_THEVO	-----MRKIN	PKKKE	YSELAD	ITK	SKAVAI	YDKGVRE	72
RLA0_PICTO	-----MTE	PAQ	WKID	FYK	LENEI	NSRKYAAIYSIKGL	72
ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90						



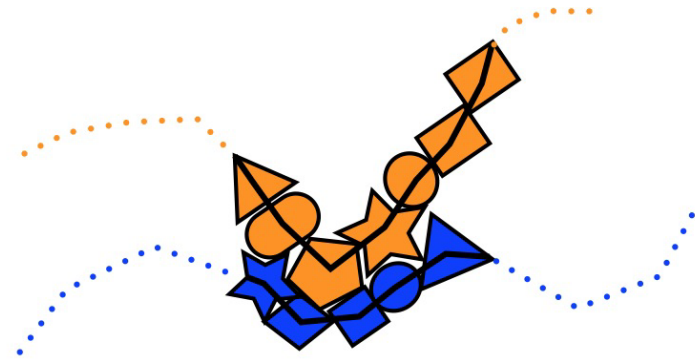
Ribosomal protein L10: from protopedia.org

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

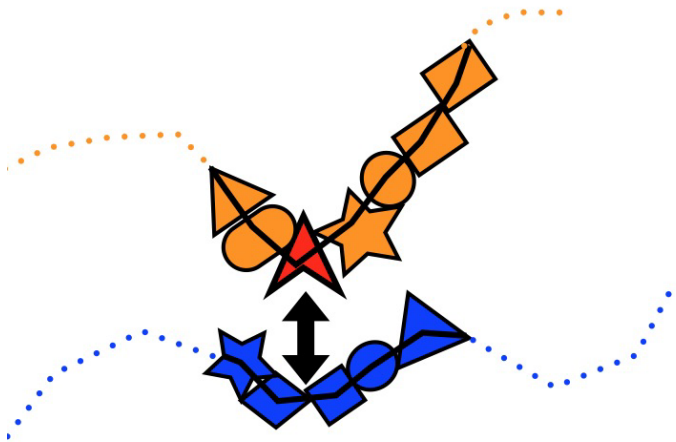
Random mutation are normally disruptive.

Mutations are highly correlated in order to preserve function

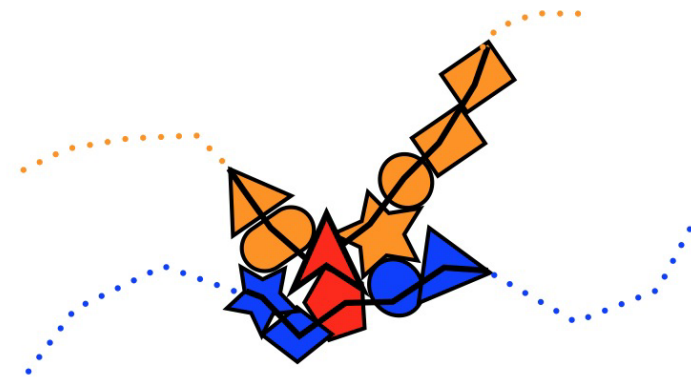
Residues proximity induces correlations



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.

$$P(\text{sequence}) \propto \exp(-H(\text{Sequence}))$$

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.

