Mapping between sequence and response spaces in the olfactory system in Drosophila

Alberto Guggiola [1], Matteo Marsili [2]

[1]Laboratoire de Physique Théorique de l'ENS, 24, Rue Lhomond, 75005, Paris, France
guggiola@lpt.ens.fr

[2]The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, 34014, Trieste, Italy

1 Abstract

The objective of this work is to map the differences among the sequences of olfactory receptor genes to the differences in how the associated neurons respond to chemical stimuli; i.e., having defined a metric in the two spaces, one would look for a relation of the kind $\mathbf{d_{ij}^{sequences}} \sim \mathbf{f}\left(\mathbf{d_{ij}^{responses}}\right)$.

Moreover, one would like to infer the most important parts of the sequences for the recognition of stimuli, i.e. subsequences able by themselves to justify the differences among the response patterns.

2 Are the spaces correlated?

The response space

The data

The DoOR Dataset is used, containing the responses of 51 different Olfactory Receptors when exposed to 204 pure chemicals.

- No information about natural odorants
- No time-dependent variation of the responses
- Data originally coming from heterogeneous experimental setups

The metric

For each couple of ORs the similarity between their response patterns is measured using the Kendall τ:

$$\tau_{ij} = \left\langle \mathrm{sign}\left[(x_i^\alpha - x_i^\beta)\cdot(x_j^\alpha - x_j^\beta)\right]\right\rangle_{\alpha,\beta}$$

The null model

For each neuron, the responses to stimuli are randomly permuted, so to obtain a τ_{rand} to compare the true results with.

The sequence space

The sequences

The sequences of the Odorant Receptor genes of Drosophila are dowloaded and aligned. They all share a common domain structure (alternance of seven transmembrane helices with extracellular domains and cytoplasmic regions).

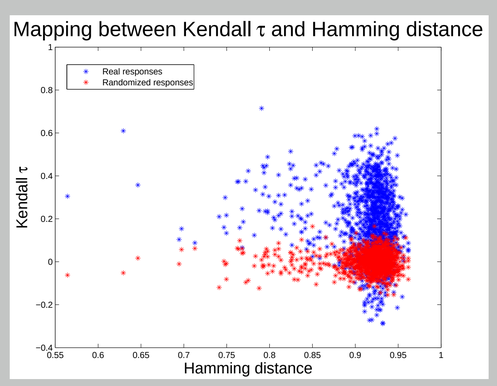
The metric

The Hamming distance between sequences (or subsequences) is considered. Some attention has to be paid about considering as *aligned* two gaps in the same position.

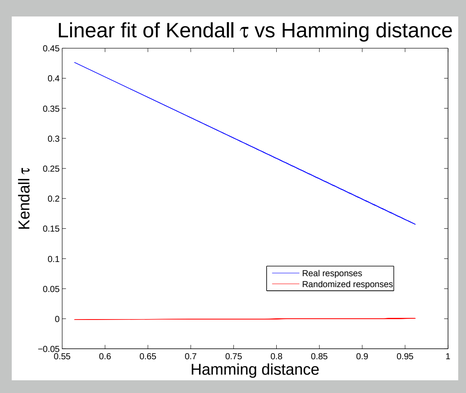
Correlations between the two spaces

Can the complete sequences explain the response pattern?

Even if the result is quite noisy, it is found that the τ_{real} are significantly anticorrelated with the Hamming distance, whereas the τ_{rand} are not.



Correlation is not clearly visible from the scatter plot of the data



τ_{real} is correlated with d_{ij}, while τ_{rand} is not

	Correlation between τ_{ij} and d_{ij}	P-value
Real responses	-0.15	$\mathbf{10^{-8}}$
Randomized responses	0.004	0.9

3 How to select subsequences?

Most significant positions

General idea

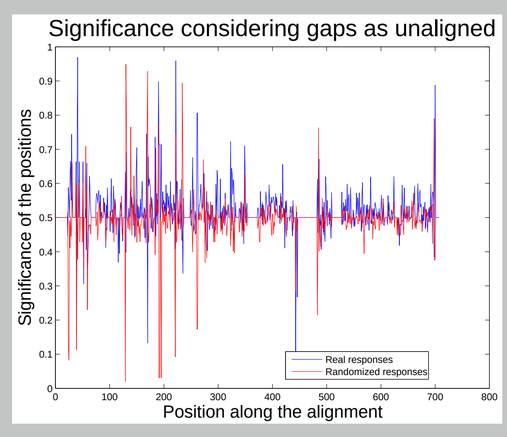
If three sequences i, j, k are considered, such that in the position p of the alignment i and j show the same amino acid, whereas in k there is a different one, then the more the position p is able to explain the response pattern, the more is likely that $\tau_{ij} > \tau_{ik}$ and $\tau_{ij} > \tau_{jk}$

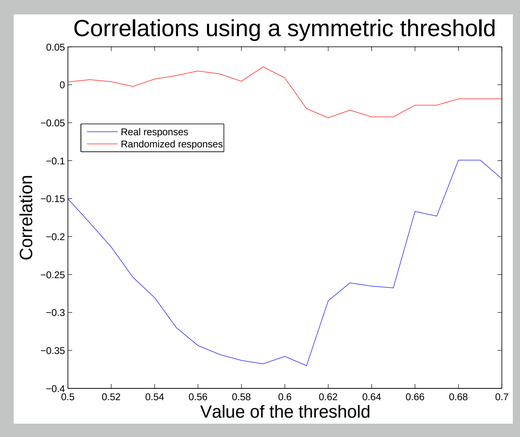
Formal definition of significance

For each position in the alignment, all the triples of the kind discussed are checked, and a mean value is calculated:

$$\sigma_p \sim \sum_{i\le j\le k} \mathbb{I}_{p_j=p_i}\mathbb{I}_{p_k\neq p_i}\left[\theta(\tau_{ij}-\tau_{ik}) + \theta(\tau_{ij}-\tau_{jk})\right]$$

Are the subsequences constituted by the most significant positions correlated with the responses?



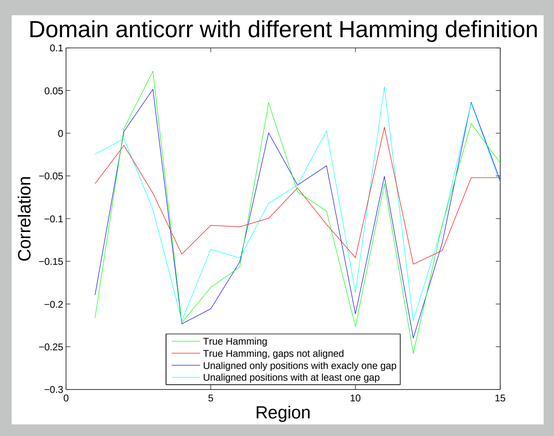


- Maximum anticorrelation for intermediate thresholds ($\sim$ **90** a.a. subsequences)
- **BUT** Unclear biological significance of the selected positions

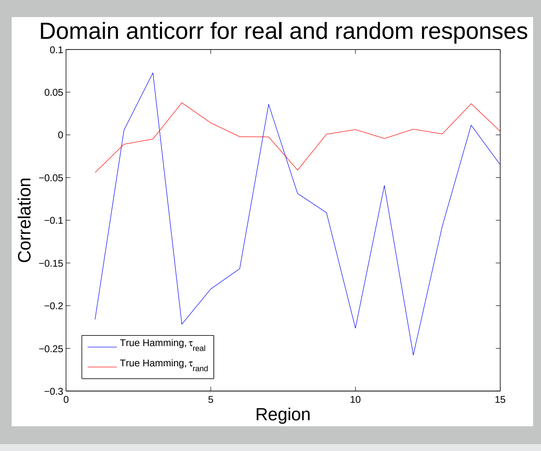
Most significant structural domains

- The structural domains are separately aligned
- The correlations between the Hamming distances between subsequences and the similarity in the responses are studied

Are different regions differently correlated with the responses?



- Significant correlations in some regions (II, V and VI helices)
- Robust w.r.t. different definitions of distance



- Randomized responses $\rightsquigarrow$ correlations compatible with zero
- Real responses $\rightsquigarrow$ correlations compatible with zero in some regions, different from zero in others

4 Perspectives

- Verify the biological role of the structural regions found to be the most correlated with the responses
- Computationally evaluate the capability of single positions in the considered regions to explain the response pattern
- Eventually, verify experimentally if those positions are actually important in the odorant recognition processes

5 References

- Munch D., Galizia C.G., *DoOR: The Database of Odorant Responses*. Chemosense: 13 (2011), 4, pagg. 1-6
- The UniProt Consortium, *Update on activities at the Universal Protein Resource (UniProt) in 2013* Nucleic Acids Res. 41: D43-D47 (2013).
- Sievers F., Higgins D. et al. *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega*, doi:10.1038/msb.2011.75
- Laissue P.P., Vosshall L.B., *The Olfactory Sensory Map in Drosophila*, Brain Development in Drosophila Melanogaster, 2008, Landes Bioscience and Springer Science, Chapter 7