Co-expression networks of addiction-related genes in the mouse and human brain

Upgrades to the website:
`http://addiction.brainarchitecture.org/`

Contents

1 Definitions and notations

- `'big12'`: the non-hierarchical partition of the left hemisphere into 12 regions (and `'Basic cell groups and regions'` that are left out of the computation of localization scores because they are too patchy)
- `'fine'`: the non-hierarchical partition of the left hemisphere into 94 regions. This annotation. It is compatible with `'big12'` in the sense that each region in `'fine'` intersects one and only one region in `'big12'`.
- `'cortex'`: the non-hierarchical partition of the left hemisphere into layers (NB: this does not cover the whole of the cerebral cortex as annotated in `'big12'` and `'fine'`.
- $E(v, g)$: the expression energy at voxel v of gene g. With data at a spatial resolution of 200 microns, the voxel x gene matrix E hase size $(49742, 3041)$.
- Localization scores of gene g in region ω_r: the fraction of the (squared) L^2-norm of the

expression energy of gene g that is comprised in region ω_r:

$$\lambda_r(g) = \frac{\int_{\omega_r} E(v,g)^2 dv}{\int_{\omega} E(v,g)^2 dv},$$

where ω is the total volume of the region of the brain that is charted in the annotation from which ω_r is taken (see [1] for definitions).
- Reference scores for region ω_r:
uniform reference: fraction of the atlas taken up by region ω_r:

$$\lambda_{\omega_r}^{\text{unif}} = \frac{\int_{\omega_r} dv}{\int_{\omega} dv} = \frac{\text{Vol}(\omega_r)}{\text{Vol}(\omega)}.$$

average reference:

$$\lambda_{\omega_r}^{\text{average}} = \frac{\int_{\omega_r} (\sum_g E(v,g))^2 dv}{\int_{\omega} (\sum_g E(v,g))^2 dv}.$$

2 Marker genes: presentation of scores and anatomical search

2.1 Rendering of localization scores

• **Visualization of existing results as histograms.** The localization scores for `'big12'`, `'fine'`, and `'cortex'` are now available only in matrix form, which makes them difficult to visualize. We postponed the visualization last year due to time constraints. At least for `'big12'` we should have a histogram on the model of what is available on the Allen website for expression level and expression density. It would be shown upon clicking the button `View localization scores`.
• **References: two additional histograms shown at the end of the page.** The uniform and average references should be shown on the same screen.
• **Fine annotation.** The same visualization technique could be applied to the `'fine'` annotation, but given the large number of regions (94), it will be difficult less illuminating. A possibility would be to make each region of `'big12'` clickable. Upon clicking a region the user will see the break-up of the localization scores into the regions of `'fine'` included in that region, which will keep the number of fine regions of order 10.

2.2 Data files

`localizationScoresBig12.mat`
`regionNamesBig12.mat`
`geneNames.mat`

uniformRefLocalizationScoresBig12.mat
averageRefLocalizationScoresBig12.mat
localizationScoresBig12 is a region x gene matrix with the lines arranged in the same order as the names of regions in regionNamesBig12 and columns arranged in the same order as in regionNamesBig12.mat.
Same goes with the strings 'Fine' and 'CortexLayers' substituted to 'Big12'. The order of genes does not change.

localizationScoresFine.mat
regionNamesFine.mat
uniformRefLocalizationScoresFine.mat
averageRefLocalizationScoresFine.mat

localizationScoresCortexLayers.mat
regionNamesCortexLayers.mat
uniformRefLocalizationScoresCortexLayers.mat
averageRefLocalizationScoresCortexLayers.mat

Matlab code snippet: localization_rendering.m
Example. The user wants to study the localizations properties of 'Gabra6', 'Pak7', 'Prox1', 'Dpp6' in the 'Big12' atlas
>> options = struct('identifier', 'big12');
>> genesStudied = 'Gabra6', 'Pak7', 'Prox1', 'Dpp6' ;
>> localizationRendering = localization_rendering(geneNames, options);

2.3 Anatomical search of marker genes

Have an additional frame below Co-expression data on the front page http://addiction.brainarchi called Anatomical search, with three possible actions modelled on Search By Gene Category:
- Marker genes: global scores
- Marker genes: local scores
- Co-markers of pairs of regions

For each of these actions, the user will be able to choose one of the three annotations 'big12', 'fine' and 'cortex'. Let us assume that the user has clicked one of these three annotations. He has the option to restrict to addiction-relatad genes (one button), and he can enter the number of genes he wants to see (the field Number of markers will be filled with a default value of 10, that the user will be able to modify by typing another value).

2.3.1 Marker genes: global scores

The user sees a list of the regions in the annotation and clicks one, call it ω_r.
If the field `Number of markers` has been filled with the value N_m, a table containing the N_m genes with highest localization score λ_{ω_r}, sorted by descending value of ω_r, is displayed. The table has four columns (clickable `Image Series`, `Gene Name`, `Gene Info`, `Entrez Id`, `localization score`). For each of the regions and each of the annotations, the sorted lists of genes in the full dataset and in the set of addiction-related genes can be precomputed and stored in the form of mat files ($2 \times (12 + 94 + 7)$ files for the three annotations, the factor of two coming from the two lists of genes, `'All genes'` and `'Addiction-related genes'`).

2.3.2 Marker genes: local markers

The user sees a list of the regions in the annotation and clicks one, call it ω_r. There are genes that have high expression in a given region, lower expression in its neighborhood, and some regions of higher expression away from the region ω_r: they *separate* the region from the rest of the brain, and are termed *local markers*. They can be detected by computed the norm of the difference between the gene-expression profiles masked around the region of interest, and the characteristic function of the region (see [1]).

The scores are not very easy to interpret, but there is a ranking of genes (that generically has large conflicts with the ranking according to localization scores). The best N_m local markers can be displayed in the same way as described above, with the column `Localization score` replaced by `Rank as global markers`. See the appendix (page 25) of [1] for results for $\omega_r =$ Striatum and $\omega_r =$ Striatum.

Again the sorted lists of genes in the full dataset and in the set of addiction-related genes can be precomputed and stored in the form of mat files ($2 \times (12 + 94 + 7)$ files for the three annotations, the factor of two coming from the two lists of genes, `'All genes'` and `'Addiction-related genes'`).

2.3.3 Co-markers of pairs of regions

The user chooses two regions ω_r and ω_s from the list of regions in the chosen annotation. The genes can be ranked by localization scores for the region made of the reunion $\omega_r \cup \omega_s$, which is not in the atlas but can be easily constructed from the annotation.

Again there is a ranking of genes. See the last appendix (page 26) of [1] for results for $\omega_r =$ Striatum and $\omega_s =$ Cerebellum.

The website displays a table of the N_m best co-markers of ω_r and ω_s.

The sorted lists of genes in the full dataset and in the set of addiction-related genes can be precomputed and stored in the form of mat files ($2 \times (12*11/2+94*93/2+7*6/2)$ files for the three annotations, the factor of two coming from the two lists of genes, `'All genes'` and `'Addiction-related genes'`).

3 Sets of genes: statistical analysis

In the present section, we assume the user has selected a list of genes. This set consists of N_g genes, corresponding top the columns of indices $(c_1, \ldots, c_{N_g})$ in the matrix E. We describe additional buttons that could be added to the frame `Your selections` below `View Localization Scores`. They could be called `Neuroanatomical properties` and `Study Co-expression networks`.

The co-expression matrix for the full set of genes is pre-computed (call it C, with $C_{gg'}$ the cosine of the angle between the gene-expresssion vectors of gene g and gene g' in voxel space). We want to make some statistics on the thresholded co-expression networks that are currently dispayed on the website (by studying the average and maximum size of connected components of the graph for selected genes, and comparing those quantities to the ones obtained for random sets of genes of the same size).

3.1 Connected components of a thresholded co-expression graph

Some of the following computations are already done in real time for the display of the co-expression networks. We would like to make an analysis of the results available to the user.

The coefficients of the co-expression networks are numbers between 0 and 1 by construction. We define the following thresholding procedure on co-expression networks: given a threshold ρ between 0 and 1, and a co-expression network C, put to zero all the coefficients that are lower than this coefficient. This defines a thresholded co-expression network C_ρ such that:

$$C_\rho(i,j) = C_\rho(i,j) \times \mathbf{1}(C(i,j) \geq \rho).$$

By construction we have $C_0 = C$. The graph corresponding to the matrix C_ρ has connected components, and each connected component has a certain number of genes in it. If the initial co-expression network has G genes, for every integer k between 1 and G we can count the number $N_\rho(k)$ of connected components that have exactly k genes in them.

We can study the average size of connected components of thresholded co-expression networks

$$\mathcal{G}(\rho) = \frac{\sum_{k=1}^{G} k N_\rho(k)}{\sum_{k=1}^{G} N_\rho(k)}$$

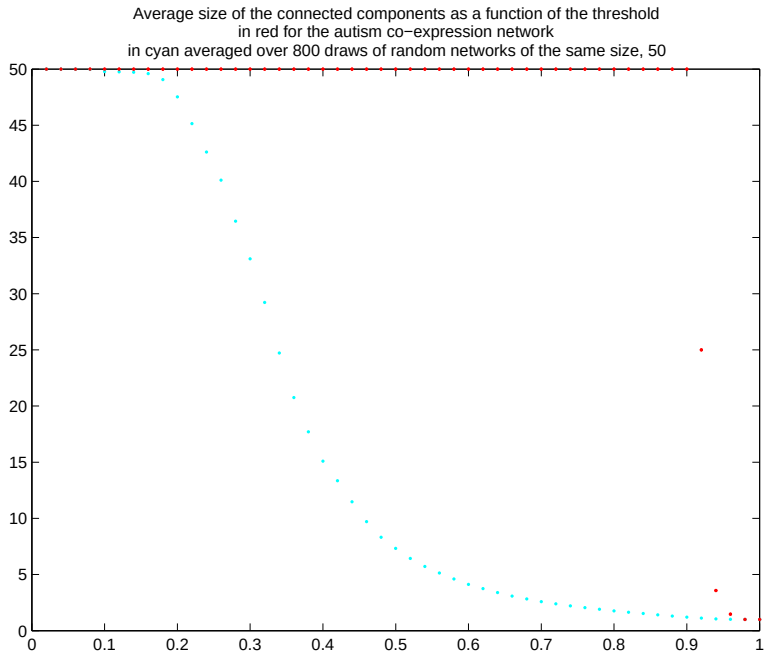

Figure 1: Average size of the connected components of thresholded co-expression graphs as a functio of the threshold, for a set of 50 autism-related genes in red, and averaged over 800 draws of random sets of 50 genes from the Allen Gene Expression Atlas (in turquoise). At a high threshold of 0.9, the 50 genes of interest are all connected.

and the maximum size of connected components

$$\max\{k > 0, N_\rho(k) > 0\}.$$

as a function of the threshold. We see that $\mathcal{G}(0)$ is the size of the set of genes, as the whole set is connected. At large thresholds every single singe is disconnected from the other genes, as having co-expression equal to one is equivalent to having exactly the same expression across the whole brain. So at threshold 1 all the connected components have size one, and $\mathcal{G}(1) = 1$.

3.1.1 Statistical significance

We can repeatedly draw random sets of N_g genes from the full dataset, compute the same quantities at each draw, and average the results across the draws (the number of draws can be fixed, and the results pre-computed, stored and accessed by values of N_g). These averages and the values obtained for the N_g genes of interest can be drwan on the same figure, as in **Figure 1** (ignore the reference to autism in the title of the figure, it was drawn for other purposes).

3.1.2 'Discovery genes'

Another button 'Discover co-expressed genes' could return a list of genes from the set of 'all genes' that are more co-expressed than a chosen threshold. This can be done using the co-expression matrix C and the ordered list of gene names. The list would be presented as a column, and each entry would have an annotation saying whether it is in out list of addiction-related genes or not.

3.2 Neuroanatomical properties

3.2.1 Quantities to evaluate in real time

The sum of gene-expression energies in the slected list of genes:

$$S(v) = \sum_{k=1}^{N_g} E(v, c_k).$$

$$S_{\text{norm}}(v) = \frac{S(v)}{\sqrt{\int_\omega S(v)^2 dv}}.$$

This function has a certain profile across regions of the brain. Is it exceptional in some region(s) of the brain?

The sum of all the genes in the dataset (or in the set of addiction-related genes), is precomputed:

$$S^{\text{all}}(v) = \sum_{g=1}^{G=3041} E(v, g).$$

$$S^{\text{all}}_{\text{norm}}(v) = \frac{S^{\text{all}}(v)}{\sqrt{\int_\omega S^{\text{all}}(v)^2 dv}}.$$

For a given annotation, say `'Big12'`, one wants to know how much the normalized sum of selected genes $S_{\text{norm}}(v)$ deviates from the sum of all genes. One computes the following logarithm of ratios for each region in the annotation:

$$L_r = \log\left(\frac{\sum_{v\in\omega_r} S_{\text{norm}}(v)}{\sum_{v\in\omega} S^{\text{all}}_{\text{norm}}(v)}\right).$$

This quantity is positive (resp. negative) at index r if the selected genes are over-expressed (resp. under-expressed) in the region ω_r.

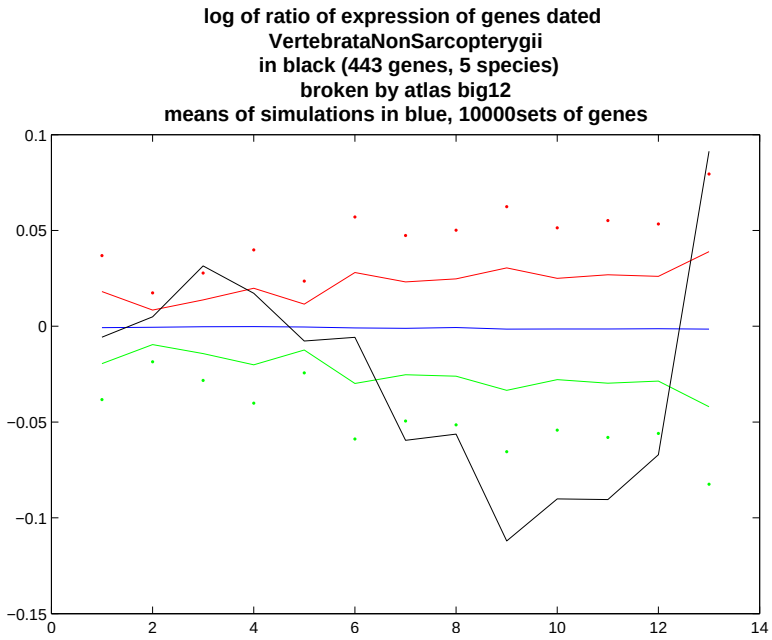

Figure 2: L_r shown as a function of label r across `Big12` for a set of 443 genes (ignore the species consideration in the title), with the regions arranged in the following order: 'Basic cell groups and regions', 'Cerebral cortex', 'Olfactory areas', 'Hippocampal region', 'Retrohippocampal region', 'Striatum', 'Pallidum', 'Thalamus', 'Hypothalamus', 'Midbrain', 'Pons', 'Medulla', 'Cerebellum'. The average of the log ratios across 10,000 random sets of genes of the same size are plotted in blue. The red (green) lines (dots) show the average values plus (minus) one (two) standard deviations from the average. We have over-expression in cerebellum and under-expression in hypothalamus.

3.2.2 Statistical significance

How exceptional are these effects? One can precompute the log-ratios defined above for sets of genes of size N_g (with a large number of sets), and compute the average values and standard deviations of L_r at fixed values of N_g).

This will give rise to figures of the type of **Figure 2** (for a set of 443 genes whose log-ratio in cerebellum is more than two stds higher than average across 10000 draws of the same size (the labels of the regions would ghave to be put in by `xTickLabel`):

References

[1] P. Grange and P.P. Mitra, *Determination of optimal sets of genes as markers of anatomical regions in the mouse brain*, `[arXiv:1105.1217 [q-bio.QM]]`.