DNA methylation changes with age as a complex system: a parenclitic network approach to a family-based cohort of patients with Down Syndrome

M. Krivonosov[1], T. Nazarenko[2,*], M.G. Bacalini[3], C. Franceschi[1,3], A. Zaikin[1,2,4], and M. Ivanchenko[1]

[1]Department of Applied Mathematics and Laboratory of Systems Biology of Ageing, Lobachevsky University, Nizhny Novgorod, 603950, Russia
[2]Department of Mathematics and Institute for Women's Health, University College London, London, WC1H 0AY, UK
[3]IRCCS Istituto delle Scienze Neurologiche di Bologna (ISNB), Italy
[4]Department of Paediatrics and Paediatric Infectious Diseases, Institute of Child Health, Sechenov First Moscow State Medical University, 119146 Moscow, Russia
[*]t.nazarenko@ucl.ac.uk

ABSTRACT

Network models are a powerful tool to represent, analyze and unfold the complexity of a large-dimensional data system at the fundamental level. The main advantage of network analysis is the opportunity to identify network disease signatures which we use in this paper for patients with Down Syndrome. One of the new methods based on the reconstruction of relations between system features is a Parenclitic Networks approach enabling setting links even for functionally unlinked features. In our work, we develop and generalize the Parenclitic Networks approach using Down Syndrome as a Case study. We present our open-source implementation to make the method more accessible to all researchers. The software includes a complete workflow to construct Parenclitic Networks and demonstrate as a generalization that any machine learning algorithm can be chosen as a kernel to build edges in the network using geometric (SVM) and probabilistic (PDF) approaches as examples. We also present a new approach (PDF-adaptive) that allows to automatically to solve one of the main problems of building a network - choosing a "cut-off threshold". We apply our implementation to the problem of detecting the network signature of Down Syndrome in DNA methylation data from the family-based cohort. We demonstrate the first insights into how Parenclitic Networks can be used not only as constructs to reduce data dimension and solve the classification problem using network characteristics, but also to find a network signature of transitional changes. We also show that this approach turns out to be more informative in terms of identifying biological essences (gene ontology analysis) than analytical approaches that consider individual CpG sites.

Introduction

Epigenetic modifications are chemical and/or physical changes in chromatin that can be preserved across cell divisions. Among epigenetic modification, DNA methylation is the most studied[1]. This is in part due to the fact that this molecular layer tends to be stable and can be assessed by cost-effective approaches, like the Infinium microarrays, that allow to measure DNA methylation ten of thousands of CpG loci with single base resolution and at the genome-wide level[2]. In addition, a bulk of literature had demonstrated that DNA methylation changes occurs both in physiological (for example, development, sex differentiation or aging) and in pathological (for example, cancer, age-related diseases or genetic syndromes) conditions[1,3,4].

So far, the studies that have investigated DNA methylation changes associated to a certain condition have used analytical approaches that consider single CpG sites (differentially methylated positions) or groups of chromosomally adjacent CpG sites (differential methylated regions)[5,6]. However, the epigenetic regulation of the genome is a complex and highly integrated system, in which DNA methylation modules are not isolated[7]. Analytical approaches that consider the cross-talk between DNA methylation alterations in different regions of the genome could therefore provide new insights on the epigenetic remodelling occurring in various physiological and pathological conditions, as recently demonstrated[8–11]. In this framework, parenclitic network analysis has the potential to provide deeper insights into the regulation of DNA methylation profiles, intended as a complex system. In contrast to many other network analysis algorithms, a parenclitic approach enables to represent methylation data in the form of a network even if links between different methylation probes or regions are unknown.

This approach is based on consideration of pairwise interactions of system features and determination of a link between these features as a deviation in each such pair of the Case group from the Control group. As a measure of deviation Zanin et al. proposed using the length of the perpendicular from each analysed point to a linear regression that was build on a Control group on the plane of two coordinates forming this pair[12]. Parenclitic networks have been successfully applied to problems of the detection of key genes and metabolites in different diseases[12–15] see[16] for a review and[17] for a discussion of applications for brain research. In[18] we have applied this methodology to implement a machine learning classification of human DNA methylation data carrying signatures of cancer development. Later[19], based on the understanding that the interactions of two features (at least in biological systems of biomarkers) often cannot be described by linear model, it was proposed to use 2-dimensional kernel density estimation (2DKDE) to model the Control distribution. The deviation is then defined as a $1-p$-value for a point lying inside the 2DKDE-grid, where p is the total probability of all points with larger probability as the point under consideration (thus the smaller the $1-p$, the farther the point is from the center of distribution and the greater the distance for it) and as normalized distance to the grid for the points outside the 2DKDE-grid. The main difficulty in using such approaches is the laboriousness of producing a complete analysis package from the beginning to the end (consisting of many technically-complex nested steps) and presence of difficult-to-solve issues (like making "cut-off thresholds" decisions for network edges design).

The main aim of our work is generalization of parenclitic approach, exploration it capabilities, making it more accessible to all researchers and demonstration of using it on real data. First, we present the implementation of a complete workflow to construct Parenclitic Networks; demonstrated that any machine learning algorithm can be chosen as a kernel to build edges in the network using geometric (SVM) and probabilistic (based on the probability density function - PDF) approaches as examples. We also present a new approach (PDF-adaptive) that allows to automatically to solve one of the main problems of building a network - choosing a “cut-off threshold”. We give all the necessary computational estimates for the kernel and indicate their features in connection with their work with specific data. Secondly, we apply our implementation to the problem of detecting the age dependent network signature of Down Syndrome on methylation data. In this study we demonstrate the first insights into how Parenclitic networks can be used not only as constructs to reduce data dimension and solve the classification problem using network characteristics (like "Case-Control" that was considered in previous works), but also as a network signature of transitional changes.

Results

Implementation of workflow to construct Parenclitic Network

The implementation of a complete workflow to construct Parenclitic Networks is available at https://github.com/mike-live/parenclitic (see Materials and Methods for details).

Kernels of Parenclitic Network

Our implementation demonstrates that any machine learning classification algorithm can be chosen as a kernel to build edges in the network using both geometric ideas (for example, SVM - see details in Methods, Parenclitic algorithm, Kernel: SVM) and probabilistic ones (for example, PDF - see details in Methods , Parenclitic algorithm, Kernels: PDF reconstruction). We also present a new PDF-adaptive approach (see details in Methods, Parenclitic algorithm, Kernel: PDF-adaptive) that allows to automatically to solve one of the main problems of building a network - choosing a “cut-off threshold”.

By default, the PDF-adaptive core is proposed (for reasons of high adaptability to data and a reasonably high calculation speed). However, the kernel can be selected based on the following considerations:

- **Asymptotics of method**: it affects the computational speed, generally it is proportional to $n^2 \cdot \text{Kernel}$, where Kernel can be: SVM $= m^2 ... m^3$; PDF $= m \cdot L + m^2, (L \sim 10^5)$; PDF-adaptive $= m^2$ (m is a number of samples in dataset, n is a number of features);

- **Features of the data:** SVM is applicable only for pairs separated by a hyperplane (possibly with bends); PDF is applicable for complex spatial configurations of classes with the ability to Control a common threshold in probability for all pairs of parameters; PDF-adaptive is applicable for complex spatial configurations of classes with the ability to automatically select for each pair of parameters of its adaptive threshold;

- **Features of the Kernels:** SVM is more robust and cannot find multiple areas of the same class; PDF simple takes into account the division of the plane into several groups belonging to the same class and able to find extra groups if there are strange outliers; PDF-adaptive is similar to the previous one, but loses a feature of the probabilistic probability cutoff approach and becomes less robust than the previous one.

Applying to detection of signature of Down Syndrome on methylation data

Background

As an application and demonstration of our implementation, we considered a publicly available dataset (GSE52588) in which whole blood DNA methylation was assessed by the Infinium HumanMethylation450 BeadChip in a cohort including persons affected by Down Syndrome (DS), their unaffected siblings (DSS) and their mothers (DSM)[20] (29 families in total).

Previous studies have demonstrated that Down Syndrome (DS), which is caused by a full or partial trisomy of chromosome 21, is characterized by a profound remodelling of DNA methylation patterns, which involves not only chromosome 21 but is widespread across the genome[20–23]. Among the different cohorts investigated so far, the GSE52588 dataset is a family-based cohort, which allows to evaluate not only the epigenetic remodelling in DS, but also the potential contribution of genetic background and environment, which are at least in part shared by the members of the same family[20]. DMPs and DMRs analysis on this dataset have highlighted epigenetic alterations in genes involved in developmental functions (including haematological and neuronal development), metabolic functions and regulation of chromatin structure, in line with the results on independent cohorts[20–23].

DS is considered a segmental progeroid syndrome, as a phenotype of premature/accelerated aging occurs in a subset of organs and systems, including the immune and the nervous systems[24,25]. Among the progeroid phenotypes of DS, the high rate and early occurrence (by the age of 40 years) of Alzheimer's disease is one of the most prominent[26]. In this framework, Horvath et al re-analysed the GSE52588 dataset, together with other methylation datasets on DS, in order to investigate the epigenetic age of DS patients using a epigenetic clock based on the weighted combination of 353 CpG sites[27]. Depending on the tissue and/or the age range of the cohort, a premature or accelerated epigenetic aging was observed, and more specifically in GSE52588 dataset DS persons were 4.6 years older than age-matched Controls.

Building different types of Parenclitic Networks

Age distribution of subjects from the 3 groups are reported in Fig.1 (B). As depicted in Fig.1 (A), this family-based model (which minimizes potential confounding effects, as members of the same family tend to share the same habits/environment) allows 3 different comparisons, in which 2 phenotypes (DS syndrome and aging) are differently combined: the comparison DS vs DSS is informative of epigenetic remodeling associated to the syndrome, independently from the age of the subjects; the comparison DSS vs DSM is informative of epigenetic remodeling occurring during ageing in euploid subjects; finally, in the comparison between DS and DSM the subjects differ for both the presence of the syndrome and for the different age. Importantly, as the DSM group in sex-biased, the probes showing a sex-dependent DNA methylation profile were removed, as described in *Preprocessing Data* paragraph of *Materials and Methods* section.

We fixed Control and Case groups (see the example in Fig.1 (C-D), where DSM is a Control group and DS is a Case group). The connection (an edge in the network) between two sites (vertices in the network) will be determined as follows. In each plane of two CpG sites, we can determine the area (in this Case, using KDE-2D) covering a Control group. In this study, we use the PDF-adaptive kernel, which automatically selects the best border which dividing Control and Case groups. After fixing such a boundary, all samples inside the Control-area don't get an edge, and all samples outside this region get an edge (Fig.1 (D)). Only those edges (and vertices connected by them) in which the separation of the Control groups from the Case group through the constructed boundary occurred with an accuracy of 90% will fall into the final network configuration.

Important note: *Our goal is not only to determine the signature (select CpG sites) using the network approach, but also to determine the information content of this approach in comparison with the methods of selecting features through a review of each site separately. Therefore, we exclude CpGs loci which show good performance of Case/Control separation in one-dimensional distributions (see for details Preprocessing Data paragraph of Materials and Methods section).*

We create four different Parenclitic networks, setting different Control-Case combinations (by manipulating the Y mark in our algorithm, see paragraph *Input data* in *Material and Methods* section):

1. **M-Control Network** — DSM group is *Control* group, DS group is *Case* group, DSS group is *Test* group;
2. **S-Control Network** — DSS group is *Control* group, DS group is *Case* group, DSM group is *Test* group;
3. **AGE-Control Network** — DSS group is *Control* group, DSM group is *Case* group, DS group is *Test* group;
4. **DS-Control Network** — DS group is *Control* group, DSS and DSM groups sre *Case* group.

A visual representation of the process of building networks in each design is presented on Fig. 2 (A (M/S/AGE/DS), B (M/S/AGE/DS)), and on Fig. 2 (C (M/S/AGE/DS)) - networks for members of Family 1 are presented. We can say that in each construction the Parenclitic Network Algorithm established a multidimensional boundary in the selected set (own for each construction) of attribute pairs through which, the set of the Control group is very clearly separated from the Case group. Samples of the Control group are located on one side of such a multidimensional border (that is, they have almost no edges in their networks), and the group of Cases turned out to be on the other side of this multidimensional border, which is indexed by

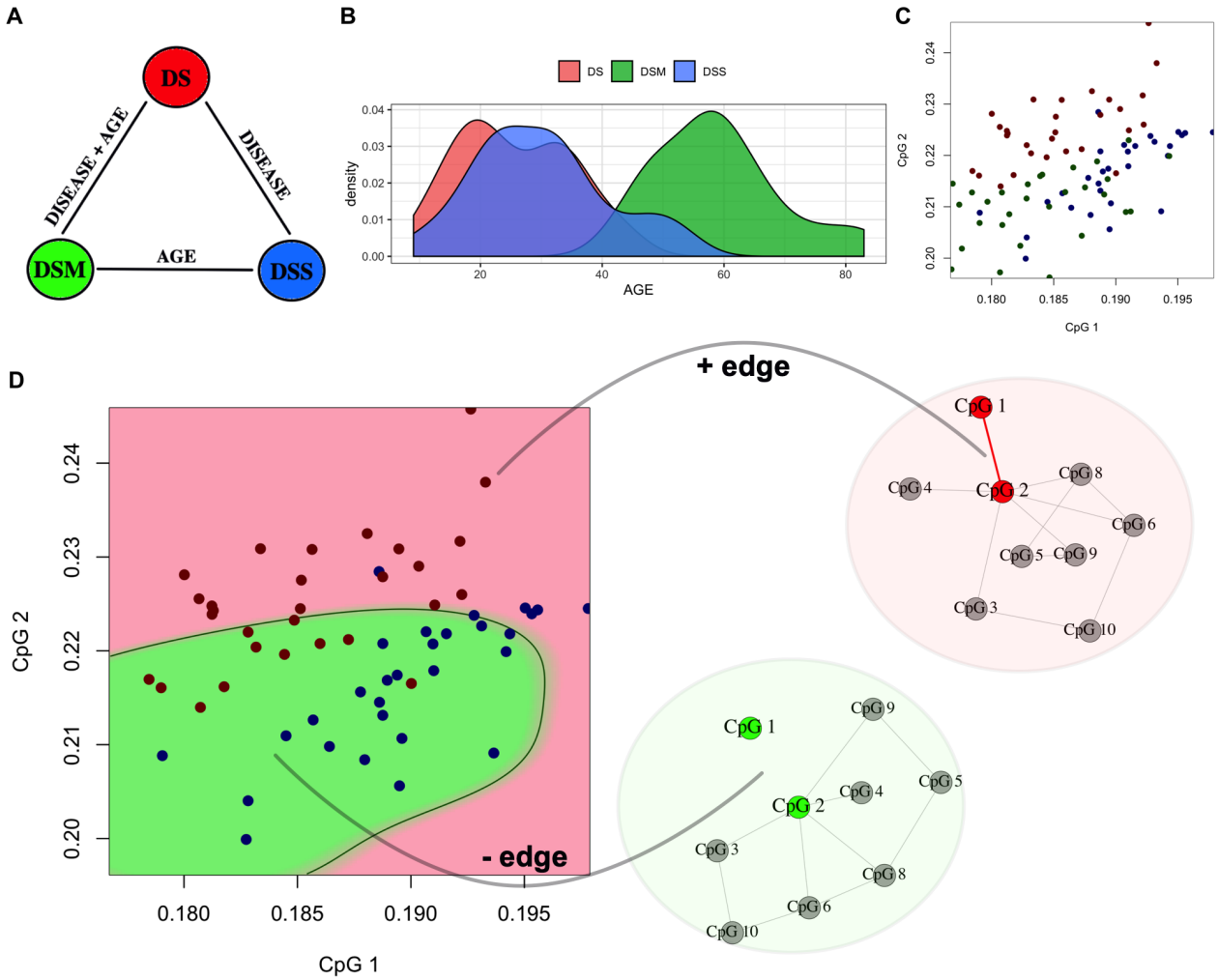

Figure 1. **(A)** Differences between the three groups under consideration by different types of properties; **(B)** Age-distribution for M/S/DS groups; **(C)** Example of plane of CpGs loci pair and original plot of points; **(D)** Building PDF area on Controls (DSM in this example) and constructing of edge between two CpGs loci if the point is “outside" the Controls-area and the absence of edge if the point is “inside" the Controls-area).

an edge in each pair of selected signs, as an indication that they lie on the other side of Control group. Thus, the illustrations of the networks for the participants in the Control (see Mother network on Fig. 2 (C (M)), Sibling network on Fig. 2 (C (S)), Sibling network on Fig. 2 (C (AGE)), DS network on Fig. 2 (C (DS))) and Case groups (see DS network on Fig. 2 (C (M)), DS network on Fig. 2 (C (S)), Mother network on Fig. 2 (C (AGE)), Mother or Sibling network on Fig. 2 (C (DS))) clearly demonstrate the global topological differences on the selected attribute systems. It is interesting to consider the participants in the Test group in the first three constructions (see Sibling network on Fig. 2 (C (M)), Mother network on Fig. 2 (C (S)), DS network on Fig. 2 (C (AGE))). This networks are characterized by an intermediate state (visually, they are not as dense as the Cases, but also not as discharged as the Controls). In order to investigate such intermediate states and establish their relationship with age, we turn to the analysis of the topological characteristics of the received networks for each participant in the dataset in each network construction.

Fig. 3 (A) shows the Venn Diagram for the vertices (CpG sites) of networks, which demonstrates that each network is predominantly unique in its attributes. The resulting networks are also shown on Fig. 3 (B,C,D,E) (enlarged images can be found at the links given in the caption of this figure).

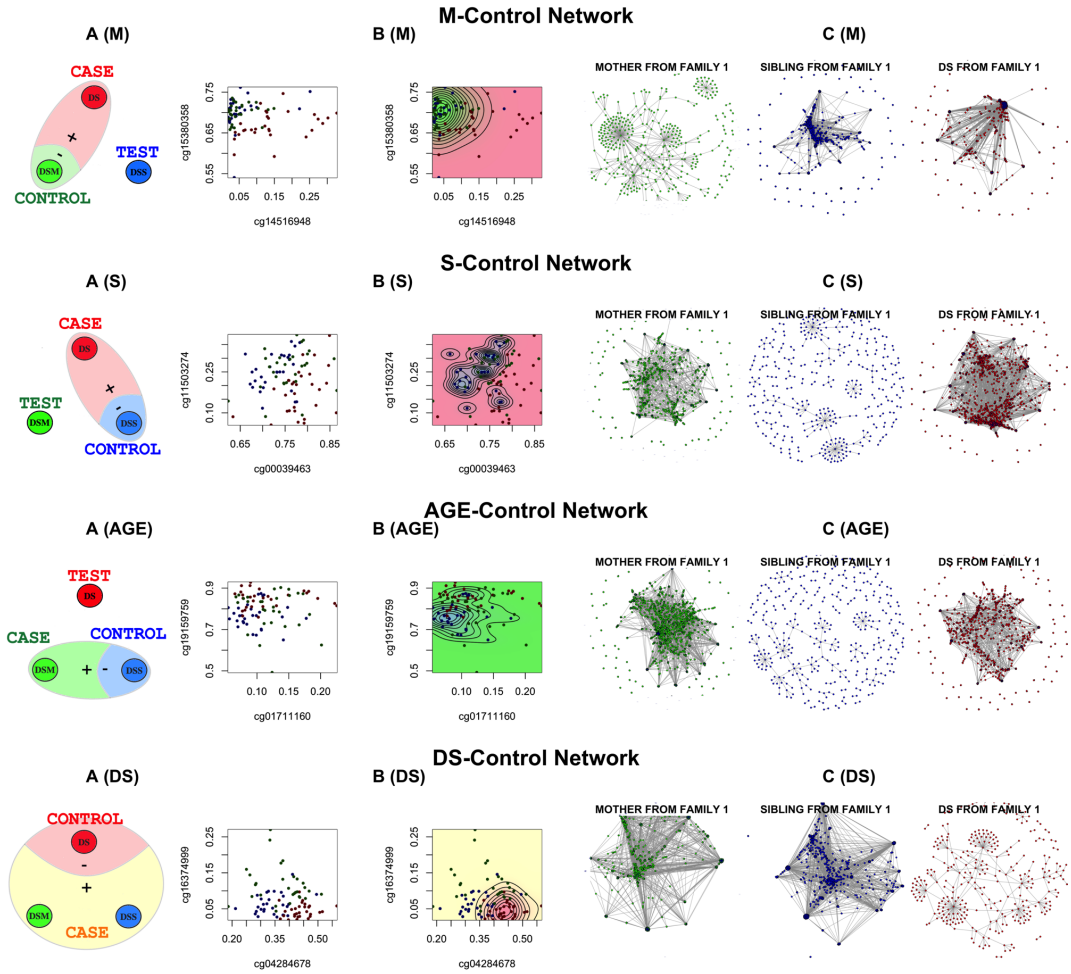

Figure 2. **(A-(M/S/AGE/DS))** Definition of Control and Case groups; **(B-(M/S/AGE/DS))** Illustration of the construction of the Control-area by Control group in each design and the automatic determination of the best border (accuracy of separation of Control and Case groups); **(C-(M/S/AGE/DS))** Illustration networks for members of one family: DSM (green network), DSS (blue network), DS (red network. Participants of the Control group is characterized by a discharged network structure; participants of the Case group is characterized by a dense and connected network structure, and participants of the Test group (for the first two networks) is characterized by an intermediate state (between discharged and density).

Networks analysis and identification of network-age-dependent effects

We previously demonstrated that persons with DS tend to show a higher biological age compared to age-matched Controls, as predicted on the basis of DNA methylation patterns according to Horvath's algorithm. As not all the DS subjects show the same epigenetic age acceleration, we considered the network characteristics respect not only in to age (AGE), but also to epigenetic age acceleration (RESIDUALS), calculated as described in the Materials and Methods section.

Since the main idea of the Parenclitic approach is to obtain topologically different objects (graphs) for Cases (mainly with edges on the set of selected features) and Controls (mainly without edges), the analysis of the characteristics of these two groups can show only a strong difference between the two groups (for example, the "number of edges" in the Cases graphs will be significantly larger than in the Controls graphs) and not show differences "inside" each group (for example, the "number of edges" in the Cases graphs will be approximately the same, since during the construction of the network we included only those signs in which the Cases/Controls separation between occurred with high accuracy (0.9)). Of particular interest is the study of *Test* group (a group not involved in building a network and different from the other two groups in a certain sense).

We analyzed all characteristics of the obtained networks and found that the characteristics for the *Test* groups (in the S-Control, M-Control and AGE-Control configurations) not only lie between the characteristics of two other groups, but also demonstrate an age-dependent transition from one group to another. In Fig. 3) (F), as an example, we give only one characteristic - number of zero degrees nodes (number of nodes without connections) and it relationship with AGE (first column

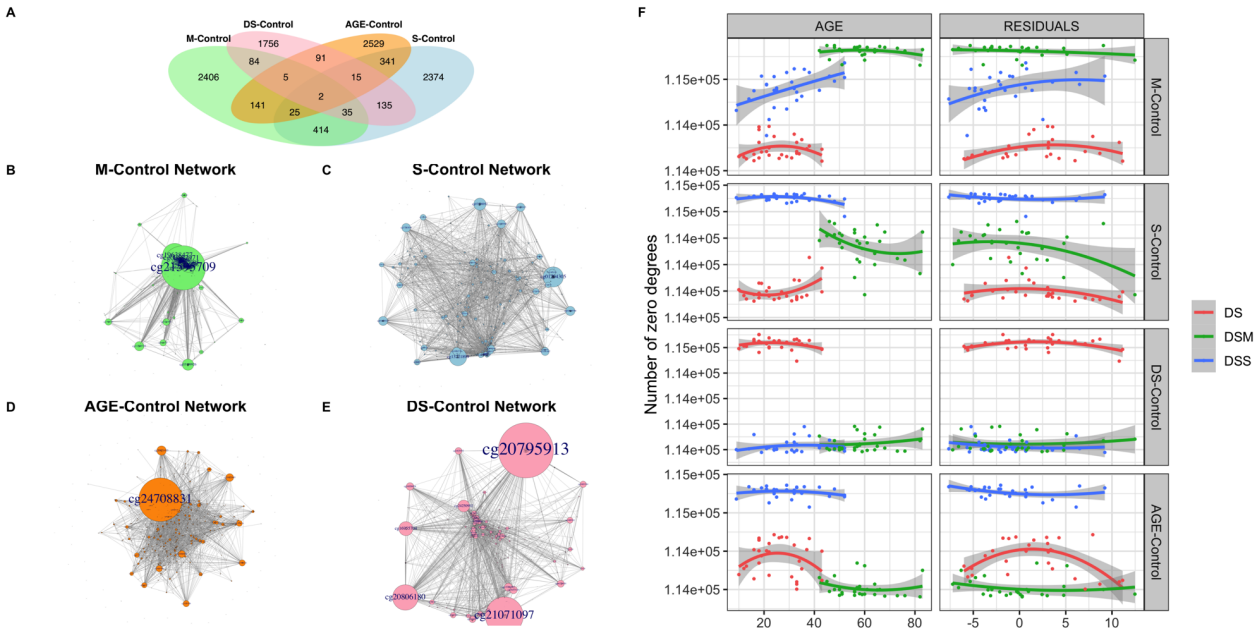

Figure 3. **(A)** Venn Diagram for the vertices (CpG sites) of networks; size of nodes in each image **(B)** (M-Control Network), **(C)** (S-Control Network), **(D)** (AGE-Control Network) and **(E)** (DS-Control Network)) are associated with their degree (the greater the degree of the vertex, the more connections it has). Enlarged presentations of networks and their constructions (in *csv* files) can be found on the links on https://tatiananazarenko.github.io/PN–DS/; **(F)** Dependence of characteristics (here, for example, only Number of zero degrees nodes) of individual networks in M/S/DS/AGE-network design versus AGE and RESIDUALS.Plots for all characteristics can be found in Supplementary Materials.

of Fig. 3 (F)) and RESIDUALS (second column of Fig. 3 (F)). Plots for all characteristics can be found in Supplementary Materials.

Considering the characteristics of networks relative to age allows us to make the following interpretations for each network configuration:

- **M-Control Network:** This is the state of the CpGs loci system in which the DSM and DS groups are separable (have different methylation) with high accuracy (this means that in each plane of two CpGs loci the DS group lies outside the PDF-area constructed on DSM group and this state is indexed for the DS by the edge in the network structure). As can be seen on Fig. 3 (F: AGE/M-Control) the older the individual from DSS group, the closer his characteristic to the DSM group. This means that the older the individual, the more often (on the set of CpGs pairs selected in the network) he falls into in the DSM area (the more the number of its zero degrees). At the same time, the younger he is, the closer he is to the DS group (the more edges and the smaller the number of zero degrees). From this we can conclude that methylation on the system of CpGs loci selected in the M-network changes in a healthy population with age, but methylation state of DS is characteristic of a very young healthy population.

- **S-Control Network:** This is the state of the CpGs loci system in which the DSS and DS groups are separable (have different methylation) with high accuracy (this means that in each plane of two CpGs loci the DS group lies outside the PDF-area constructed on DSS group and this state is indexed for the DS by the edge in the network structure). As can be seen on Fig. 3 (F: AGE/S-Control the older the individual from DSM group, the closer his characteristic to the DS group. This means that the older the individual, the more often (on the set of CpGs pairs selected in the network) he falls into in the DS area (the less the number of its zero degrees). At the same time, the younger he is, the closer he is to the DSS group (the less edges and the higher the number of zero degrees). From this we can conclude that methylation of the system of CpGs loci selected in the S-network changes in a healthy population with age but methylation state of DS is characteristic of a very old healthy population.

- **DS-Control Network:** This is the state of the CpGs loci system in which the DSS and healthy population (DSS and DSM together) are separable (have different methylation) with high accuracy (this means that in each plane of two CpGs loci the healthy population lies outside the Control-area constructed on DS group and this state is indexed for healthy

population by the edge in the network structure). In this Case, we rigidly separate a healthy population (regardless of age) and we can conclude that system of CpGs loci pairs selected in this network characterize the pure state of DS disease regardless of age.

- **AGE-Control Network:** This network included simply age-dependent CpG loci. By analyzing the characteristics of networks, we can argue that the characteristics of DS on average lie closer to the characteristics of DSM (this means that DS, the more often (on the set of CpGs pairs selected in the network) fall into in the DSM area (the less the number of its zero degrees) than to theirs healthy peers), which indicates that DS shows age-related acceleration.

Gene Ontology Results

To prepare starting list for performing of parenlitic analysis we use list of CpGs described in **Preprocessing data** section. Developed approaches are very sensitive to features which are one-dimensional good separators. Therefore, we exclude those CpGs for the particular network configuration. Next, based on obtained background lists for each network configuration we perform parenclitic analysis.

Finally, the list of CpG probes retrieved by parenclitic analysis in each comparison was assessed for the enrichment in Gene Ontology (GO) terms. While the DS-Nodes comparison did not return any significantly enriched GO, 23, 37 and 77 GO terms were enriched in lists of CpG probes resulting from the S-Nodes comparison, M-Nodes comparison and AGE-Nodes comparison respectively. A large overlap was observed between these 3 comparisons, with 19 terms in common. To gain more insights on the ontology shared between the comparisons or specific for them, we applied REVIGO to the lists of GO terms to remove redundant terms and we focused on the "Biological Processes" description (Table. 1). We found 5 GO terms (pattern specification process, regionalization, cell fate commitment, skeletal system development, appendage development) common to all the comparisons, and other 5 GO terms (embryonic organ morphogenesis, epithelial cell differentiation, forebrain development, camera-type eye development, neuron migration) were shared between the S-Nodes comparison and the M-Nodes comparison. Interestingly, the same GO analysis on the results of 1D comparison ((Table. 1), Supplementary Materials) returned a lower number of enriched GO in the S-Nodes comparison and in the M-Nodes comparison , and no overlap between these two comparisons and the AGE-Nodes comparison.

Term ID	Description	Frequency	S-Nodes comparison (log10 p)	M-Nodes comparison (log10 p)	DS-Nodes comparison (log10 p)	AGE-Nodes comparison (log10 p)
GO:0007389	*pattern specification process*	0.15%	-5.684	-5.8894		-7.9355
GO:0003002	*regionalization*	0.12%	-5.1203	-5.382		-8.644
GO:0045165	*cell fate commitment*	0.09%	-4.9626	-4.8861		-5.0339
GO:0048562	*embryonic organ morphogenesis*	0.08%	-4.7282	-5.3354		
GO:0021953	*central nervous system neuron differentiation*	0.04%	-3.804			
GO:0001501	*skeletal system development*	0.11%	-2.9274	-3.1411		-2.5157
GO:0035107	*appendage morphogenesis*	0.06%	-2.859			
GO:0048736	*appendage development*	0.06%	-2.8043	-3.8576		-3.3526
GO:0007517	*muscle organ development*	0.08%	-2.3705			
GO:0072073	*kidney epithelium development*	0.03%	-2.1971			
GO:0030855	*epithelial cell differentiation*	0.14%		-3.5217		
GO:0030900	*forebrain development*	0.07%		-3.4096		
GO:0043010	*camera-type eye development*	0.07%		-3.3543		
GO:0001764	*neuron migration*	0.03%		-2.2874		
GO:0097485	*neuron projection guidance*	0.07%				-4.2434
GO:0001655	*urogenital system development*	0.07%				-4.0381
GO:0007156	*homophilic cell adhesion via plasma membrane adhesion molecules*	0.10%				-3.8404
GO:0030509	*BMP signaling pathway*	0.04%				-3.8404

GO:0072001	*renal system development*	0.07%				-3.8404
GO:0007423	*sensory organ development*	0.16%				-3.8404
GO:0061138	*morphogenesis of a branching epithelium*	0.04%				-3.7958
GO:0001763	*morphogenesis of a branching structure*	0.05%				-3.6196
GO:0048608	*reproductive structure development*	0.16%				-3.4037
GO:0048705	*skeletal system morphogenesis*	0.05%				-3.3695
GO:0060173	*limb development*	0.04%				-3.3526
GO:0061458	*reproductive system development*	0.16%				-3.2725
GO:0030534	*adult behavior*	0.04%				-3.0895
GO:0043583	*ear development*	0.05%				-2.8072
GO:0019933	*cAMP-mediated signaling*	0.01%				-2.6638
GO:0007187	*G-protein coupled receptor signaling pathway, couple to cyclic nucleotide second messenger*	0.05%				-2.4709
GO:0060541	*respiratory system development*	0.05%				-2.3026
GO:0007416	*synapse assembly*	0.04%				-2.1716
GO:0035270	*endocrine system development*	0.03%				-2.1445
GO:0007188	*adenylate cyclase-modulating G-protein coupled receptor signaling pathway*	0.04%				-2.1072
GO:0060485	*mesenchyme development*	0.06%				-2.1072
GO:0007600	*sensory perception*	0.43%				-2.0474

Table 1. Gene Ontology results for Nodes of S/M/DS/AGE-Control Networks. It can be seen here that several processes that significant for the S-Control and M-Control Networks at the same time (e.g. *pattern specification process, regionalization, cell fate commitment,skeletal system development,appendage development*) turn out to be significant also for the AGE-Control (which, on a par with the study of network characteristics, confirms the fact that the selected processes in these networks turn out to be age-dependent in a healthy population). It is also interesting that the *embryonic organ morphogenesis* process turned out to be significant for S-Control and M-Control Networks, but not for the AGE-Control. This indicates that this process is not characteristic of development in a healthy population at an age, but is different for DS from a healthy population of any age. At the same time, it can be seen that both the S-Control and M-Control Networks contain unique processes (S-Control: *central nervous system neuron differentiation, appendage morphogenesis, muscle organ development, kidney epithelium development*; S-Control: *epithelial cell differentiation, forebrain development, camera-type eye development, neuron migration*), which indicates that the networks (in addition to the general factor - AGE-dependence) contain unique disease properties characteristic.

Materials and Methods

Parenclitic implementation

Parenclitic is an open source python library. It can be distributed through PyPI repository. Current version 0.1.6 provides functionality described in this paper.

Our package provides 3 main features:

- Build, save and load parenclitic network.
- Choose or create kernel to identify edges.
- Compute network metrics based on python-igraph package.

Performance tests are done on machine 256Gb RAM, x2 CPU Intel Xeon Gold 6238 CPU @ 2.10GHz, 22 cores per socket, 2 threads per core, 88 threads in sum. Computation time of CpG dataset with 87 samples and 114674 features is 147562 seconds $\sim$ 41 hours.

Parenclitic algorithm

In this section we describe a framework of the generalized parenclitic algorithm. The main goal of the algorithm is to identify deviations of a sample by pair of features. Final representation of those relations is a network where features corresponds to nodes and feature connections are links between nodes.

Input data

A structure of input data for Network construction is as follows: (X_i, Y_i), where X_i - feature vector of i-th subject, $Y_i \in \pm 1, \pm 2$ - label of i-th subject: number (1 or 2) defines set of subjects involved in construction a network (1 - involved, 2 - not involved) and sign defines set of Controls and Cases ($"-"$: Control group for construction KDE area, $"+"$: Cases - group for study of deviation from Control).

Network construction

Network construction consists of 3 steps:

1. First of all, we consider subjects on the two-dimensional plane of couple of features (i, j). There is generally a Control, Case and Test subject groups, or only a Control and Case groups. For certainty, consider an example where DSM is a Control group and others are a Test group (Fig. 1 (C,D)).
2. At this step, a special classification method is used to determine the distance from subject to Control group. This is kernel method which determines subject deviations from a normal state. Based on subject features it should predict if subject deviated from Control group or not and how much it deviates (Fig. 1 (D)).
3. Finally, we reconstruct individual network for subject such that nodes are a features and links between 2 nodes designate a subject deviation by corresponding features. More precisely there is link if subject deviate from Control group. ((Fig. 1 (D)).

A classification method on the second step is a kernel of parenclitic algorithm. As such a method, we suggest 3 different approaches considered in the next 3 subsections. Samples with 2 selected features and labels represents an input for those algorithms. Algorithms perform classification subjects on 2 classes: deviated from Control group and not deviated.

Kernels

- **Kernel: PDF reconstruction.** As one of these classifiers, the distance threshold obtained from a PDF estimation can be used. The idea is to reconstruct a probability density function (PDF) of Control group subjects on two dimensional plane of two selected features. Let it be, for example, the KDE (kernel density estimation) method with Gaussian kernel. There are 2 groups: Control group and test group. So, algorithm can be described by the following sequence of steps:

    1. Compute parameters of KDE for Control group samples.
    2. Calculate distance from subject to PDF. Let us define such distance on the plane of two features:

$$d(x,y) = \int\limits_{\mathrm{PDF}(\eta,\nu)>\mathrm{PDF}(x,y)} \mathrm{PDF}(\eta,\nu)\, d\eta\, d\nu, \tag{1}$$

where (x,y) - values of two selected features for subject. The meaning of this distance formula is a probability for subject to be more probable to Control group PDF. So, the closer subject in value of PDF(x,y) to the maximum value PDF(x^*,y^*), the smaller the distance $d(x,y)$.

3. Choose threshold of significance level. As $d(x,y)$ is a probability, it has value in range $[0,1]$. Therefore, one can choose a constant threshold value for the distance. Initially, there is no rule for threshold defining, so it can be varied in the allowable range with some step. For certainty, let us define set of thresholds $\mathbf{thr} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

4. Identify subject deviation from a Control group by a comparison between the distance for this subject and the threshold value. If the subject is far from Control group ($d(x,y) > \mathbf{thr}$), this deviation is significant.

- **Kernel: SVM classification.** Due to the long time of computation for the PDF kernel approach, we suggested faster distance classifier method. The methylation data analysis of Down Syndrome blood shows that almost all interesting feature pairs represent simple point cloud configurations. The Support Vector Machine classifier is well applicable to such data. In the general Case, one can use different kernels inside the SVM and different parameters for the best results. To finalise all steps for the identification of deviations from the Control group one should:

  1. Construct separating hyperplane by SVM algorithm on the samples with labels C_i from both groups: Control and test.

  2. Check if SVM provides us with a good separation or not. Division on 2 sets can be assessed by setting a score, and if the score is too small no need to consider those couple of features. So, if $score < \mathbf{thr_s}$ then there is no way to determine correct deviations. Also this score cut off help to reduce number of links.

  3. Calculate signed distance from sample to hyperplane $d(x,y)$.

  4. Choose threshold value for distance (by default is 0). Here one can apply the same approach as for choosing a threshold in the PDF approach.

  5. Identify subject deviation from the Control group by a comparing a distance for this subject and the threshold value. If the subject is far from Control group ($d(x,y) > \mathbf{thr}$) deviation is significant (the same rule as in PDF approach)

- **Kernel: PDF-adaptive (the best threshold method).** The calculation of a distance in previous approaches helped us to reduce a two-dimensional problem to the one-dimensional one. Based on those distance function a decision rule is defined. The only essential complication in previous kernel algorithms is a choice of the threshold value. To overcome it we propose the following: let's choose a threshold value in an adaptive manner. Choosing a threshold defines the 2D plane division into two complex sets of points: deviated or not. Subjects can be in both sets, but it can be in expected set or not. But on the one-dimensional distances it just a two ranges: $[0;\mathbf{thr}]$ and $(\mathbf{thr};1]$. Here, however, the problem of the best set splits arises. To the best of our knowledge it has already been solved in decision trees algorithm, using a measure of an information gain for all possible splits and choosing the one that maximizes it. So, only the third step in the PDF reconstruction method is practically changed.

  There are finite number of threshold separators where subsets of samples are changed. For all of them, one can calculate information gain metric and take the threshold giving the maximum value.

Network analysis

Revealing internals of investigated complex system can be done by applying network analysis. Insight to the structure of networks give us a possibility to measure integral system state observed for different subjects. Also, centrality measures locates the most important or interesting nodes (features for initial system) in network.

In the previous section, the algorithms produced network for each subject, where nodes represent features, and links — deviation by couple of features from it's expected state. In this study the following metrics is considered:

- Metrics that associate a number with each node (feature):Betweenness, Pagerank, Closeness, Eigenvector centrality, Degrees;

- Metrics that associate a number with each links: Distance;

- Metrics that characterize graph as whole: Efficiency, Robustness, Max component size, Number of nodes, Number of links;

- Other metrics: Component sizes;
- Statistics for vector metrics: Minimal value, Maximal value, Mean, Standard deviation, Number of zeros;

Full metric description includes:

- Degree centrality is defined as the number of links incident upon a node. The degree can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network (such as a virus, or some information)
- Betweenness centrality is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex. *The betweenness centrality type measures how often each graph node appears on a shortest path between two nodes in the graph. Since there can be several shortest paths between two graph nodes s and t, the centrality of node u is:

  $$c_u = \sum_{s,t \neq u} \frac{n_{st}(u)}{N_{st}}$$

  $n_{st}(u)$ is the number of shortest paths from s to t that pass through node u, and N_{st} is the total number of shortest paths from s to t. If the graph is undirected, then the paths from s to t and from t to s count only as one path (divide the formula by two).
- PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank of undirected graph statistically close to its degree distribution. *The pagerank centrality type results from a random walk of the network. At each node in the graph, the next node is chosen uniformly from the set of successors of the current node (neighbors for the undirected Case). The centrality score is the average time spent at each node during the random walk.
- Closeness centrality of a node is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes. The closeness centrality types use the inverse sum of the distance from a node to all other nodes in the graph. If not all nodes are reachable, then the centrality of node i is:

  $$c_i = \left(\frac{A_i}{n-1}\right)^2 \frac{1}{C_i},$$

  A_i is the number of reachable nodes from node i (not counting i), n is the number of nodes in G, and C_i is the sum of distances from node i to all reachable nodes. If no nodes are reachable from node i, then c_i is zero.
- Eigenvector centrality is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. The eigenvector centrality type uses the eigenvector corresponding to the largest eigenvalue of the graph adjacency matrix. The scores are normalized such that the sum of all centrality scores is 1.
- Distance to dividing line from SVM; or a PDF distance defined in PDF kernel description.
- Component is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the initial graph. There are several components. Size of component is a number of vertices in this subgraph.
- Efficiency of a network is a measure of how effective is the information exchange in the system. Average efficiency is defined as:

  $$\frac{1}{n(n-1)} \sum_{s \neq t} \frac{1}{d_{st}}$$

  where n denotes number of nodes and d_{st} denotes the length of the shortest path between a node s and another node t.
- Robustness of the ability to withstand failures and perturbations, is a critical attribute of many complex systems including complex networks. Robustness calculated as number of steps in process of removing nodes with high degrees until links presented in graph.

Complexity analysis

Evaluation of network analysis complexity affecting the computation time provided in this paper can be done using the algorithmic complexity. The main pipeline includes analysis of pairs of features and performing some special kernel estimation for them. So, the complexity of a framework algorithm is $O(n^2 \cdot T_{kernel}(m))$, where m denote the number of samples and n — the number of features. The computational time of the kernel depends on implementation details. Therefore, the following is a detailed description of the PDF kernel calculation.

PDF kernel detailed implementation

The algorithm of this kernel supposes that input data consists of $d = 2$ features and m samples.

- First step is a calculation of two-dimensional KDE function. It is represented by a sum of Gaussian functions and can be evaluated as $O(d^3 + m \cdot d^2)$. Hence, in the Case $d = 2$, complexity is $O(m)$.

- Next, distance from subjects to distribution is calculated. There are several techniques to compute an integral. Here we suggest the Monte Carlo integration method. It does not have the highest accuracy, but it can help calculate all of the integrals at once in the fastest way.

  Let us consider the Monte Carlo method to integrate PDF over the whole plane. To start we should do sampling from reconstructed distribution and sum up them. Let us define random variables $(X_j, Y_j), j = \overline{1, L}$ equally distributed same as the reconstructed two dimensional PDF. Next, we generate pair of numbers (x_j, y_j) from those distributions.

$$\frac{1}{L}\sum_{j=1}^{L} \mathrm{PDF}(x_j, y_j) \xrightarrow[L\to\infty]{} \int\limits_{\mathbb{R}^2} \mathrm{PDF}(\eta, \nu) d\eta d\nu = 1 \tag{2}$$

  This phase can be done in $O(L \cdot d^2 \cdot m)$ time, where L denote the number of integration points. But we are interested in computing of integral that select only those PDF values that are greater than some constant value $PDF(x, y)$. In this way we obtain:

$$\frac{1}{L}\sum_{j=1}^{L} \mathrm{PDF}(x_j, y_j) \cdot [\mathrm{PDF}(x_j, y_j) > \mathrm{PDF}(x, y)] \xrightarrow[L\to\infty]{} \int\limits_{\mathrm{PDF}(\eta,\nu)>\mathrm{PDF}(x,y)} \mathrm{PDF}(\eta, \nu) d\eta d\nu \tag{3}$$

  It costs $O(L \cdot d^2 \cdot m^2)$ for computing m distances. First optimization of this formula is a calculating PDF values only once. To achieve this we order (x_j, y_j) by value of $p_j = PDF(x_j, y_j)$ in descending order and compute cumulative sum $s_j = \sum_{k=1}^{j} p_j$. This takes $O(L + L \cdot \log L + L \cdot d^2 \cdot m)$ time. Values of m distances for subjects can be obtained by binary search over p_j from s_j. Overall this step takes $O(L + L \cdot \log L + L \cdot d^2 \cdot m + m \cdot \log m)$. By experiments, we suggest $L = 10^4$ or $L = 10^5$ to achieve a balance between computational time and accuracy.

- After the distance values are known, a construction of links using threshold takes linear time in m.

It takes a lot of time to compute distance values. But if only the deviation presence matters it can be performed without heavy integral computations. To achieve this, notice that distance monotonically decreased with PDF value increased and for one constant $p = PDF(x, y)$ value there is only one distance value $d = d(x, y)$. This means that we can solve the problem of threshold defining using d or p values in a manner. In the last section, we define the adaptive threshold algorithm which gives the same result on both p and d sequences. So, for adaptive threshold technique: computation of p costs $O(m^2)$ and best split identifying based on information gain costs $O(m \cdot \log m + m)$. Full complexity estimation is $O(n^2 \cdot m^2)$.

The last kernel based on the SVM method doesn't use any additional computations, consequently, the complexity of the kernel is the same as SVM on 2 features with m samples (also it depends on internal parameters and implementation). Full complexity estimation is $O(n^2 \cdot m^2)$.

Application

Preprocessing Data

Raw *.idat* files extracted and pre-processed using minfi Bioconductor package[28]. Data were normalized using the preprocessFunnorm function implemented in[29], and probes with a detection p-value higher that 0.05 in more than 1% of samples

were removed. Furthermore, cross-reactive and polymorphic probes, as reviewed by Zhou et al[30], and probes on the X and Y chromosomes were removed.

Finally, 412993 probes from the original array were used. Due to the large number of features it is necessary to reduce dimensionality.

In our study, we are concentrating only on Islands and Shores regions (for the preservation of the biological meaning): CpGs set - a subset HM450 probes from Islands and Shores regions (114674 CpGs). We exclude gender-specific CpGs to decrease bias evolved by gender variety. We use list of gender-specific CpGs[31] with 5665 items. 113630 remains in the list after removing those CpGs.

Our goal is not only to determine the signature (select CpG sites) using the network approach, but also to determine the information content of this approach in comparison with the methods of selecting features through a review of each site separately. Therefore, we preliminarily check whether the Control and Case groups (for each network design) are separated in the one-dimensional Case (when we consider each site separately and use one-dimensional distributions). If this happens with an accuracy of more than 75%, we do not include such a vertex in the network. From such sites we form a separate list of 1D. Split border for 1D classification is constructed based on maximizing of the information gain. Such technique usually used in decision trees.

DNAmAGE and RESIDUALS calculation

DNAmAGE (DNA methylation age) was calculated using the online tool available at https://dnamage.genetics.ucla.edu/. RESIDUALS (epigenetic age acceleration) values were calculated as the residuals of the linear regression between Horvath's epigenetic age and chronological age.

Gene Ontology analysys

Gene Ontology enrichment was performed using the methylgometh function implemented in the methylGSA Bioconductor package[32]. As the parenclitic approach does not return a p-value, the input file for methylgometh function was created by assigning an arbitrary p-value of 0.001 to all the CpG sites selected by the parenclitic analysis, and an arbitrary p-value of 0.9 to all the remaining CpG sites on which the parenclitic analysis was run. A significance cut-off o 0.01 was then used. REVIGO (http://revigo.irb.hr/) was used to remove redundant GO terms.

Discussion

In this study, we presented our open-source implementation of Generalized Parencltic Network analysis to make it more accessible to all researchers. One of the main methodological advance is an introducing new kernels and discussing the possibility of their selection based on the goals of the problem. We believe that the simple integration of new kernels into the overall implementation will allow researchers to use not only the proposed methods, but also connect and test their own ideas.

As a Case study we applied our implementation to the construction of DNA methylation Parencltic Networks for a family-based cohort patients with Down Syndrome and revealed a network-age-dependence effects. We gave interpretations of the constructed networks as a group of functions *decelerated* and *accelerated* in the age scale for DS patients. Our study allowed for the first time to decompose the epigenetic signature accompanying DS patients into such categories based on the hidden links between the covariates.

We believe that the networks built in this work are not the end of the study, but only the beginning, because new information obtained in the course of network analysis can be now widely analysed by biologists and clinicians in order to identify molecular mechanisms resulting in and accompanying Down Syndrome. Additionally, this network analysis may help to identify new molecular targets for treatment of patients with Down syndrome to prevent their accelerated ageing. The networks built in this study require more detailed analysis and can help researchers involved in Down Syndrome research to discover new interpretations based on the interactions detected. We particularly highlight here that the main idea of Parenclitic Network approach, consisting in the separation states of group Case-Control, can be also apply for analyzing the transition in time (by age or other continuous scale) from one such state to another. The *test* set (which is located between two critical states within the framework of the process under study) can become an additional improvement in the study of complex systems.

In our work, we used different constructions of Parenclitic Networks by determining critical states (Control and Case) in different ways. For example, in the S-Control Network, where peers of DS patients were selected as the Control group, and DS patients as the Case group. Firstly, we identified attributes in which these two groups were well separated. Then, all the selected attributes of the system were understood as a multidimensional region (consisting of selected attribute pairs), in which a multidimensional boundary separating the Case and Control groups (obtained by PDF-adaptive algorithm) was fixed (as the union of all boundaries in attribute pairs). Mothers set (as *test* set) was passed through such a fixed construction and it was shown what their place is between the two boundary states. Since it was important in our task to associate the differences between Control-Case groups with age-related changes, we have investigated the connection between *test*-networks (namely,

their topological characteristics) with age and showed that the transition between two critical conditions is highly associated with age. Namely, that transitions across Network borders are associated with a change in age in a healthy population and remain almost unchanged in patients with DS.
This design can be applied to a sample of patients of any disease (for example, cancer), in which, in addition to critical states (healthy - diagnosed patient), there are data on intermediate conditions (for example, analyzes of diagnosed patients at earlier time, when they could be considered as healthy people). We believe that the results of such an analysis based on Generalized Parenclitic Networks will not only help in the early diagnosis of the disease (by identifying critical transition marks) and risk assessment, but also shed light on the process itself, through attributes involved in it.

References

1. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of dna methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607, DOI: 10.1038/s41580-019-0159-6 (2019).
2. Bibikova, M. *et al.* High density dna methylation array with single cpg site resolution. *Genomics* **98**, 288 – 295, DOI: https://doi.org/10.1016/j.ygeno.2011.07.007 (2011). New Genomic Technologies and Applications.
3. Cerrato, F. *et al.* Dna methylation in the diagnosis of monogenic diseases. *Genes* **11**, 355, DOI: 10.3390/genes11040355 (2020).
4. Robertson, K. D. Dna methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610, DOI: 10.1038/nrg1655 (2005).
5. Li, D., Xie, Z., Le Pape, M. & Dye, T. An evaluation of statistical methods for dna methylation microarray data analysis. *BMC Bioinforma.* **16**, DOI: 10.1186/s12859-015-0641-x (2015).
6. Maksimovic, J., Phipson, B. & Oshlack, A. A cross-package bioconductor workflow for analysing methylation array data. *F1000Res* **5**, 1281, DOI: 10.12688/f1000research.8839.3 (2016).
7. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52, DOI: 10.1038/35011540 (1999).
8. Cui, Z.-J., Zhou, X.-H. & Zhang, H.-Y. Dna methylation module network-based prognosis and molecular typing of cancer. *Genes (Basel)* **10**, DOI: 10.3390/genes10080571 (2019).
9. Horvath, S. *et al.* Aging effects on dna methylation modules in human brain and blood tissue. *Genome Biol* **13**, R97, DOI: 10.1186/gb-2012-13-10-r97 (2012).
10. Lund, J. B. *et al.* Weighted gene co-regulation network analysis (wgcna) of promoter dna methylation on all-cause mortality in old-aged birth cohorts finds modules of high-risk associated biomarkers. *J Gerontol A Biol Sci Med Sci* DOI: 10.1093/gerona/glaa066 (2020).
11. Yuan, L. & Huang, D.-S. A network-guided association mapping approach from dna methylation to disease. *Sci Rep* **9**, 5601, DOI: 10.1038/s41598-019-42010-6 (2019).
12. Zanin, M. & Boccaletti, S. Complex networks analysis of obstructive nephropathy data. *Chaos: An Interdiscip. J. Nonlinear Sci.* **21**, 033103 (2011).
13. Zanin, M., Menasalvas, E., Sousa, P. A. & Boccaletti, S. Preprocessing and analyzing genetic data with complex networks: An application to obstructive nephropathy. *Networks & Heterog. Media* **7**, 473 (2012).
14. Zanin, M., Menasalvas, E., Boccaletti, S. & Sousa, P. Feature selection in the reconstruction of complex network representations of spectral data. *PloS one* **8** (2013).
15. Zanin, M. *et al.* Knowledge discovery in spectral data by means of complex networks. *Metabolites* **3**, 155–167 (2013).
16. Zanin, M. *et al.* Combining complex networks and data mining: why and how. *Phys. Reports* **635**, 1–44 (2016).
17. Papo, D., Buldú, J. M., Boccaletti, S. & Bullmore, E. T. Introduction: Complex network theory and the brain. *Philos. Transactions: Biol. Sci.* **369**, 1–7 (2014).
18. Karsakov, A. *et al.* Parenclitic network analysis of methylation data for cancer identification. *PloS one* **12**, e0169661 (2017).
19. Whitwell, H. J., Blyuss, O., Menon, U., Timms, J. F. & Zaikin, A. Parenclitic networks for predicting ovarian cancer. *Oncotarget* **9**, 22717–22726, DOI: https://doi.org/10.18632/oncotarget.25216 (2018).
20. Bacalini, M. G. *et al.* Identification of a dna methylation signature in blood cells from persons with down syndrome. *Aging* **7**, 82–96, DOI: 10.18632/aging.100715 (2015).

21. Do, C., Xing, Z., Yu, Y. E. & Tycko, B. Trans-acting epigenetic effects of chromosomal aneuploidies: lessons from down syndrome and mouse models. *Epigenomics* **9**, 189–207, DOI: 10.2217/epi-2016-0138 (2017).

22. Henneman, P. *et al.* Widespread domain-like perturbations of dna methylation in whole blood of down syndrome neonates. *PLoS One* **13**, e0194938, DOI: 10.1371/journal.pone.0194938 (2018).

23. Laufer, B. I., Hwang, H., Vogel Ciernia, A., Mordaunt, C. E. & LaSalle, J. M. Whole genome bisulfite sequencing of down syndrome brain reveals regional dna hypermethylation and novel disorder insights. *Epigenetics* **14**, 672–684, DOI: 10.1080/15592294.2019.1609867 (2019).

24. Franceschi, C. *et al.* Accelerated bio-cognitive aging in down syndrome: State of the art and possible deceleration strategies. *Aging Cell* **18**, e12903, DOI: 10.1111/acel.12903 (2019).

25. Martin, G. M. Genetic syndromes in man with potential relevance to the pathobiology of aging. *Birth Defects Orig Artic Ser* **14**, 5–39 (1978).

26. Lott, I. T. & Head, E. Dementia in down syndrome: unique insights for alzheimer disease research. *Nat Rev Neurol* **15**, 135–147, DOI: 10.1038/s41582-018-0132-6 (2019).

27. Horvath, S. *et al.* Accelerated epigenetic aging in down syndrome. *Aging Cell* **14**, 491–5, DOI: 10.1111/acel.12325 (2015).

28. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics* **30**, 1363–9, DOI: 10.1093/bioinformatics/btu049 (2014).

29. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369, DOI: 10.1093/bioinformatics/btu049 (2014). https://academic.oup.com/bioinformatics/article-pdf/30/10/1363/17344721/btu049.pdf.

30. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22–e22, DOI: 10.1093/nar/gkw967 (2016). https://academic.oup.com/nar/article-pdf/45/4/e22/25364844/gkw967.pdf.

31. Singmann, P. *et al.* Characterization of whole-genome autosomal differences of dna methylation between men and women. *Epigenetics & Chromatin* **8**, DOI: 10.1186/s13072-015-0035-3 (2015).

32. Ren, X. & Kuan, P. F. methylgsa: a bioconductor package and shiny app for dna methylation data length bias adjustment in gene set testing. *Bioinformatics* **35**, 1958–1959, DOI: 10.1093/bioinformatics/bty892 (2019).

Acknowledgements (not compulsory)

Authors acknowledge support by the grant of the Ministry of Education and Science of the Russian Federation Agreement No. 074-02- 2018-330. AZ, TN acknowledge support by the MRC grant MR/R02524X/1.

Author contributions statement

AZ, CF, MI designed the study, MK developed software, TN developed methodology of age-related analysis, MK, TN, MGB conducted the graph and data analysis. All authors have written and reviewed the manuscript.

Additional information

Authors declare no competing interests.