Penalized likelihood for sparse contingency tables

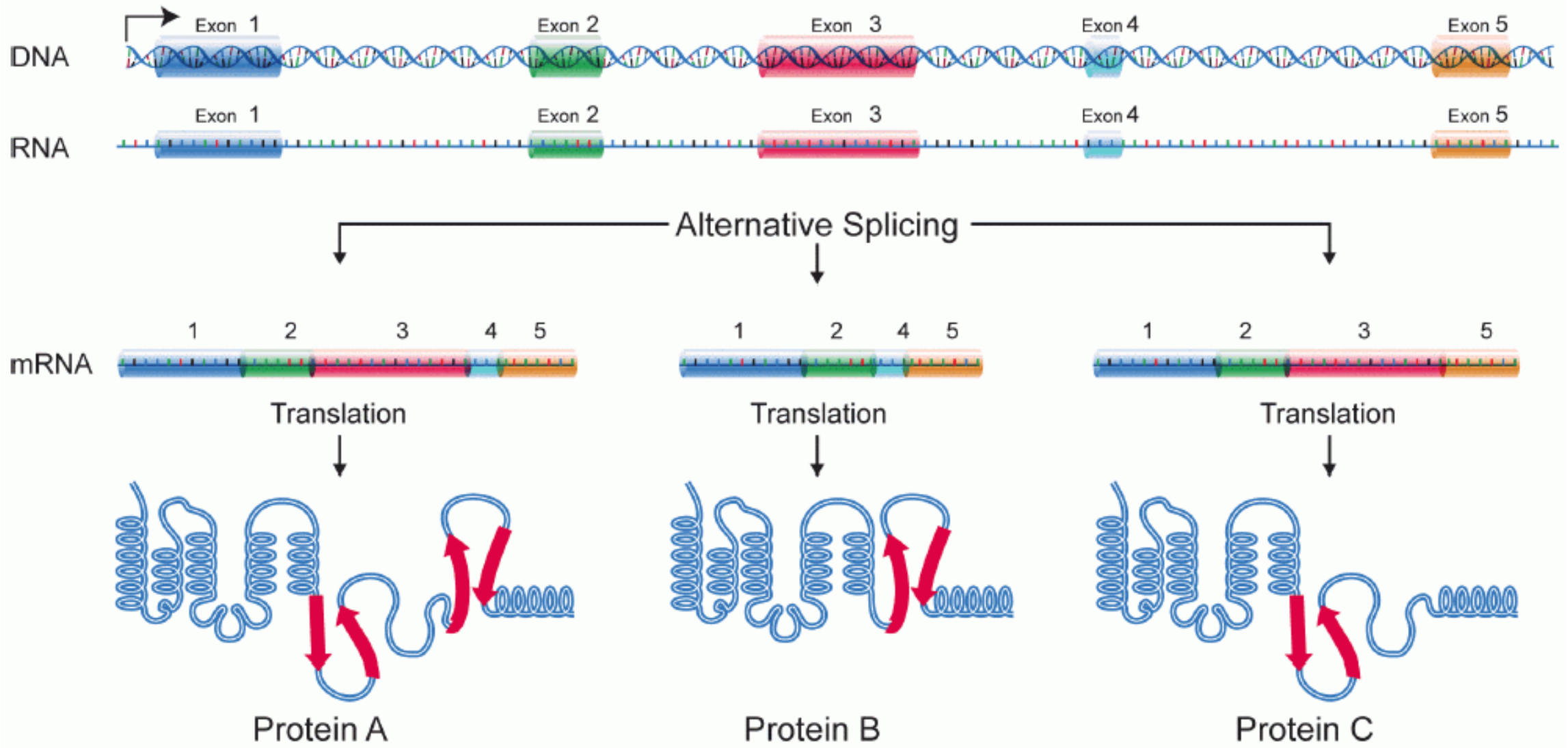
——with an application to full-length cDNA libraries

属性数据分析 第10组

朱珉珲 蒋翌坤 李渭栋 蒋慧仪 胡文博

* 生物小课堂

- 人类基因数量实际上并不多，大致有20000-25000，和其他更简单的生物基因数量差不多。
- **alternative RNA splicing**，指在基因转录时，intron（内含子）被删除，只有**被选择到**的exon（外显子）保留，组成mRNA。这个过程通过从基因上选择不同的exon，使一个基因可以生成很多不同蛋白质。
- **Full-length cDNA library**提供了详细的exon组合信息。但是由于可能的exon组合数量十分巨大，使得得到的列联表是**稀疏(sparse)**的，其中维度代表exon，单元格代表观测到的变体。
- 通过判断在alternative RNA splicing中exon之间是否有交互效应，可以简化寻找在功能测定中的肽内相互作用这一难题。


DNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
RNA
Exon 1
Exon 2
Exon 3
Exon 4
Exon 5
Alternative Splicing
mRNA
1 2 3 4 5
1 2 4 5
1 2 3 5
Translation
Translation
Translation
Protein A
Protein B
Protein C

背景知识

- 假设有q个分类变量$C = (C_1, \dots, C_q)$，每个分类变量有g_q个可能的取值，于是一共有$m = \prod_{i=1}^{q} g_j$个不同的分类组合。
- 为了书写简单，我们将每一个分类组合映射到一个自然数上。

$$f: (c_1, \dots, c_q) \leftrightarrow i \in \{1, \dots, m\}$$

- 当有n个观测值时，q维列联表有m个单元格，单元格观测到的数量服从多项分布。

背景知识

- Log-linear Model

完整模型:

$$\log p_i = \beta_\emptyset + \sum_{l\in\{1,\dots,q\}} \beta_l c_l + \sum_{\substack{j,k \\ j<k\in\{1,\dots,q\}}} \beta_{jk} c_j c_k + \cdots + \beta_{12\dots q} c_1 c_2 \dots c_q$$

即$\log p = X\beta$，X为设计矩阵

- 在**alternative RNA splicing**中，假设exon有q个，$c_j \in \{-1, 1\}$，表示exon j是否被选择到。
- 当假设一些变量的交互效应不存在时，模型变为

$$\log p = X_a \beta_a$$

其中X_a为X中一些列向量被移除。

背景知识

- Graphical Model
- 通过图的方式可以很好的表示变量间的条件相关性。顶点表示不同的分类变量。
- 如何得到边：如果$\forall a \subset b, \beta_b = 0$，有$\beta_a \neq 0$，则在所有对应$a$的顶点间画边。
- 从图中可以轻松看出边际独立和条件独立：如果a和b被c隔开，则a和b条件独立于c。

模型选择方法

- 模型选择 层次模型 vs 非层次模型
- 层次模型：如果$\beta_{\mathrm{ij}} = 0$,那么就有对于$\forall k, \beta_{ijk} = 0$.
- 层次模型是非层次模型的一个子集，其优势在于系数之间存在一定的约束关系使得模型得到了简化，便于计算，系数稳定性较强。
- 层次模型不是生物学中的一个明显的特征。生物特性中基因表达并没有分层的性质。其因子之间的相互作用是非常复杂的。
- 与非层次模型相比，层次模型的系数等级次序的稳定性较强，不会随着设计矩阵X设定的不同而变化。

模型选择方法

- $\boldsymbol{L_1}$ - *regularization* 模型选择
- LASSO线性回归模型:

$$\hat{\beta}^{\lambda} = \arg\min_{\beta}\left[\sum_{i}(Y - X\beta)_i^2 + \lambda\sum_{j}\left|\beta_j\right|\right]$$

- 可能存在的问题：研究的对象是sparse contingency table!

模型选择方法

- $\boldsymbol{L_1}$ - *regularization* 模型选择

$$\hat{\beta}^{\lambda} = \arg\min_{\beta}\left[-l(\beta) + \lambda\sum_{j}|\beta_j|\right], \text{式中} l(\beta) = \log P_{\beta}[n] \propto \sum_{i=1}^{m}\frac{n_n}{n}(X\beta)_n$$

- 增加约束条件概率和为1：$s.t.\sum_{i=1}^{m}\exp\{(X\beta)_i\} = 1$
- 可能存在的问题：如果设计矩阵X发生变化，那么可能导致模型中的系数发生不稳定的变化。

模型选择方法

- $\boldsymbol{L_1}$ - *regularization* 模型选择
- 引入**Group- $\boldsymbol{L_1}$ - Penalty** 介于$\boldsymbol{L_1}$和$\boldsymbol{L_2}$之间的惩罚项

$$\hat{\beta}^{\lambda} = \arg\min_{\beta}\left[-l(\beta) + \lambda \sum_{\substack{a\subseteq C \\ a\neq\varnothing}} \left\|\beta_a\right\|_{l_2}\right] \qquad \sum_{\substack{a\subseteq C \\ a\neq\varnothing}} \left\|\beta_a\right\|_{l_2}, \left\|\beta_a\right\|_{l_2}^2 = \sum_{j} (\beta_a)_j^2$$

- λ的选择：交叉验证。
- 训练集训练计算$p(\hat{\beta}^{\lambda})$，测试集计算负对数似然值：$\frac{-\sum_{i=1}^{m} n_{test,i} \cdot \log\left(p_i(\hat{\beta}^{\lambda})\right)}{\sum_{i=1}^{m} n_{test,i}}$

模型选择方法

- *Level-L_1-regularization*模型选择

- 完全饱和模型 ⟶ 低阶交互作用的子模型：只用主效应拟合模型
- 得到与|C|层次相对应的 |C| 分数，其中分数最低的层次被选中
- 避免出现只包含单个高阶相互作用，而忽略大部分低阶交互作用的情况
- 倾向于选择较稀疏、能够更好地分层的模型

模型选择算法

- $\boldsymbol{L_1}$-*regularization* for factors with two levels的算法

$$\begin{cases} \sum_{a \subseteq C} \|\beta_a\|_{\ell_2}, \text{ where } \|\beta_a\|_{\ell_2}^2 = \sum_j (\beta_a)_j^2 \\ \sum_{i=1}^{m} \exp\{(\mathbf{X}\boldsymbol{\beta})_i\} = 1. \end{cases} \quad \text{等价于} \quad g(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \sum_{i=1}^{m} \exp(\mu_i) + \lambda \sum_{\substack{a \subseteq C \\ a \neq \varnothing}} \|\beta_a\|_{\ell_2}$$

- 记$\mathcal{A}$为主动交互项集合，则对$\forall a \in \mathcal{A}$，都有$\beta a \neq 0$

$$\nabla g_{\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}}, \lambda) = -\mathbf{X}_{\mathcal{A}}^t \{\frac{\mathbf{n}}{n} - \cdot \exp(\mathbf{X}_{\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}})\} + \lambda(0, sign(\boldsymbol{\beta}_{\mathcal{A}}))^t$$

$$\nabla^2 g_{\mathcal{A}}(\boldsymbol{\beta}_{\mathcal{A}}, \lambda) = \mathbf{X}_{\mathcal{A}}^t diag\{\exp\{\mathbf{X}\boldsymbol{\beta})\}\mathbf{X}_{\mathcal{A}}.$$

模型选择算法

- $\boldsymbol{L_1}$-*regularization* for factors with two levels的算法

(1) Start with $\hat{\boldsymbol{\beta}} = (-\log(m), 0, \ldots, 0)$

(2) Set: $\lambda_0 = 1$, $\mathcal{A} = \{\varnothing\}$ and $t = 0$.

(3) While $(\lambda_t > \lambda_{min})$

(3.1) $\lambda_{t+1} = \lambda_t - \varepsilon$

(3.2) $\mathcal{A} = \mathcal{A} \cup \{j \notin \mathcal{A} : |[\mathbf{X}^t \cdot \frac{\mathbf{n}}{n} - \exp(\mathbf{X}\hat{\boldsymbol{\beta}})]_j| > \lambda_{t+1}\}$

(3.3) $\hat{\boldsymbol{\beta}}$ is updated as $\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t - \nabla^2 g_{\mathcal{A}}(\hat{\boldsymbol{\beta}}_{t'}, \lambda_{t+1})^{-1} \cdot \nabla g_{\mathcal{A}}(\hat{\boldsymbol{\beta}}_{t'}, \lambda_{t+1})$.

(3.4) $\mathcal{A} = \mathcal{A} \setminus \{j \in \mathcal{A} : \{\hat{\boldsymbol{\beta}}_{t+1,j}| < \delta\}$

(3.5) $t = t + 1$

- 该算法得到的数据对$(\beta \mathrm{t}, \lambda t)$是一系列惩罚参数$\lambda t$的估计值$(\mathrm{t}=\varepsilon, 2\varepsilon, \cdots)$
- 其中，步长ε的选择代表了计算复杂度与精度之间的权衡

TESTING-DATA

Independently simulated using N(0,1)

True underlying interaction vector β consisting of **5 factors** (1 to 5) of **2 levels**

345 + 235 + 234 + 135 + 123 + 14 generators

➞ all third and fourth order interactions are absent

➞ five of ten second order interactions and all first order interactions are present

TESTING-CRITERIA

标准	公式	说明
MSS	$MSS = 1 - \frac{1}{m}\sum_{i=1}^{m}\left\lvert I_{\{\beta_i \neq 0\}} - I_{\{\hat{\beta}_i \neq 0\}}\right\rvert$	得以正确分配的模型项占比
RMSE	$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}_i - \beta_i)^2}$	/
SPREAD	$SPREAD = \frac{1}{m}\sum_{i=1}^{m}\hat{\sigma}_i$	$\hat{\sigma}_i$表示$\hat{\beta}_i$标准差估计值；用以表示不同数据集下β的变化情况
NLS	$NLS(\hat{\beta}) = -\sum_{i=1}^{m} p_i \cdot \log(p_i(\hat{\beta}))$	比较不同的估计流程

TESTING-RESULTS

	MSS	NLS	RMSE	SPREAD
		Penalty-based regularization methods:		
ℓ_1-regularization	69.7%	2.20	0.228	0.144
Level-ℓ_1-regularization	89.7%	2.22	0.237	0.179
Relaxed ℓ_1-regularization	82.2%	2.22	0.233	0.154
ℓ_2-regularization	-	2.20	0.238	0.130
		MCMC without model selection:		
$\sigma^2 = 2$	-	2.32	0.747	0.401
$\sigma^2 = 1$	-	2.27	0.467	0.287
$\sigma^2 = 1/2$	-	2.24	0.294	0.201
		MCMC with model selection:		
$\sigma^2 \sim \Gamma^{-1}(2,3)$	81.5%	2.23	0.294	0.231
$\sigma^2 = 2$	76.6%	2.25	0.431	0.342
$\sigma^2 = 1$	78.4%	2.24	0.331	0.265
$\sigma^2 = 1/2$	76.6%	2.23	0.281	0.225
		MCMC with hierarchical model selection:		
$\sigma^2 \sim \Gamma^{-1}(2,3)$	84.1%	2.22	0.255	0.180
$\sigma^2 = 2$	80.6%	2.29	0.415	0.284
$\sigma^2 = 1$	83.4%	2.26	0.308	0.221
$\sigma^2 = 1/2$	83.4%	2.24	0.247	0.178
$\sigma^2_1 = 1/10$	86.3%	2.20	0.236	0.097
$\sigma^2 = 1/100$	69.7%	2.28	0.420	0.033

TESTING-RESULTS

	MSS	NLS	RMSE	SPREAD
		Penalty-based regularization methods:		
ℓ_1-regularization	69.7%	2.20	0.228	0.144
Level-ℓ_1-regularization	89.7%	2.22	0.237	0.179
Relaxed ℓ_1-regularization	82.2%	2.22	0.233	0.154
ℓ_2-regularization	-	2.20	0.238	0.130

使用l_2惩罚项代替l_1惩罚项

使用两个惩罚参数而非单一的参数λ
在一定条件下，这一方法是优于Lasso方法的

TESTING-RESULTS

在NLS、RMSE与SPREAD的角度上，文献中提出的**penalty-based regularization** 方法并不逊色于Bayesian方法
* **Level - l_1-regularization**与**Relaxed l_1-regularization**的结果更优于MCMC方法。

需要注意的是，Bayesian 方法的结果与初始值的关系有待进一步探索
* 对于MCMC方法而言，σ^2的不同初始值/不同先验分布的设定会导致迥异的结果，这也为实际应用带来了一定的困难。

Model selection的模型 优于 无model selection的模型
Hierarchical模型 优于 non-hierarchical模型

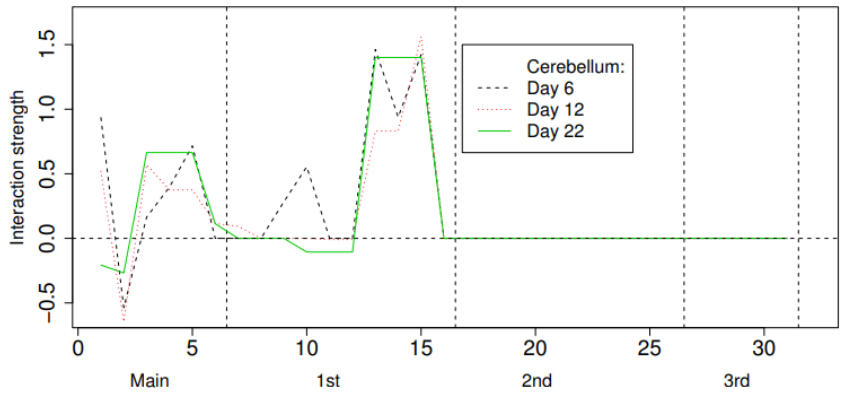
- Level - l_1-regularization在模型选择方面表现最好（MSS↑，NLS与RMSE↓），且参数λ可以通过交叉验证获得;
- MCMC方法则为模型控键的不确定性提供了更多信息。

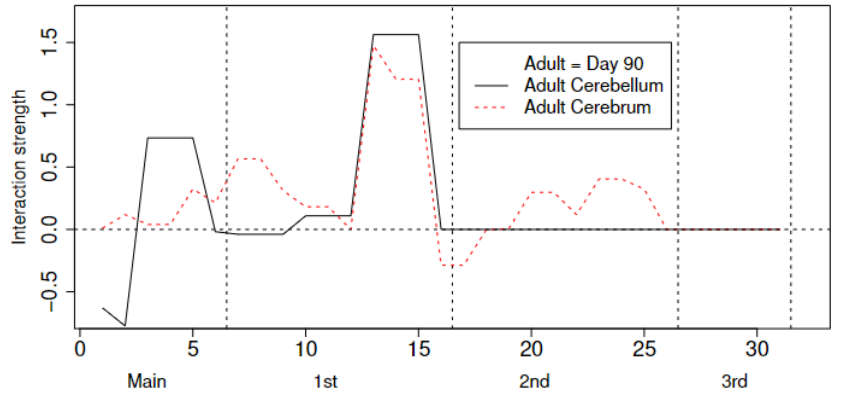
应用

- 使用$\boldsymbol{L_1}$-penalization，论文针对哺乳动物itpr1基因研究了它在剪切时的交互效应
- 该基因编码第二信使*三磷酸肌醇（InsP3）*的受体之一。该基因经历alternative RNA splicing，有七个转录的变异位点，其中$q = 5$在单基因库中被完全评估。研究构建五个单基因库，一个来自成年大鼠大脑，另外四个来自出生后不同阶段的小脑，分别为第6天、第12天、第22天和第90天（成年）。每个单基因库包含转录后的RNA数量在179至277之间
- 该基因在cDNA水平上与人类受体基因相似度为89%，在氨基酸水平上为95%。

应用

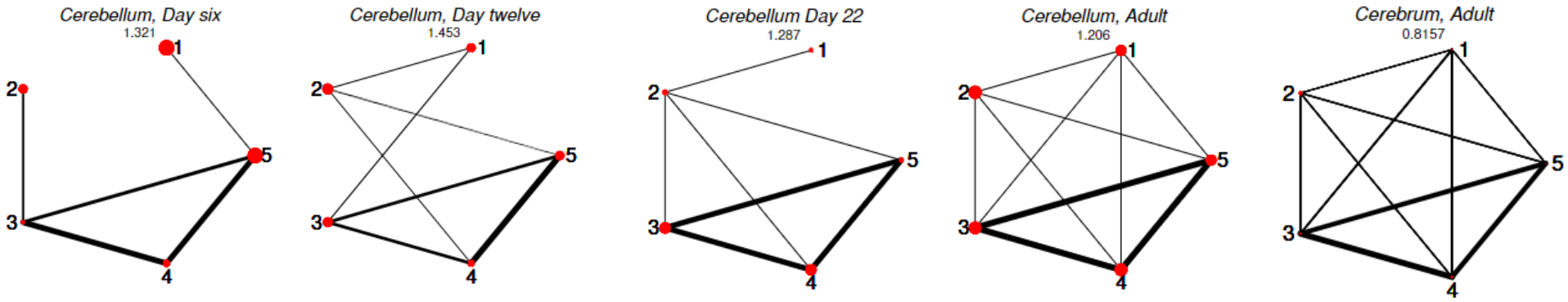
- **相互作用向量图**。左图显示了大鼠在出生后第6天、第12天和第22天小脑组织对应的$\hat{\beta}$。右图显示了大鼠在90天时小脑和大脑组织对应的$\hat{\beta}$。在图中，系数的顺序从左到右依次排列，例如对于二阶相互作用，排列顺序是123，124，125，...，345





应用

- **条件独立图**。对于每个图，预测概率分数（负对数似然值）作为衡量Goodness of fit的指标。注意到外显子三、四和五之间的强烈相互作用。



应用

从结果来看:

- 该基因的外显子主要以成对方式相互作用，没有高阶的相互作用。
- 主效应在不同发育阶段基本上相同。
- 外显子三、四和五之间的强烈相互作用。
- 在大鼠小脑发育过程中，相互作用模式的最大变化发生在第6天到第12天
- 大鼠成年时小脑和大脑相互作用模式的变化，比在不同发育阶段的小脑中的模式更为复杂，涉及多个二阶相互作用。

结论

- 论文提供了一个高效的算法level- $\boldsymbol{L_1}$ -regularization来识别分类变量间的交互效应。通过模拟实验，该算法相较于MCMC与$\boldsymbol{L_1}$ -regularization有着很好的拟合结果。
- 该算法可以通过交叉验证来优化参数λ，计算成本低。
- 论文提供的算法可以用于处理高维、稀疏的列联表数据；尤其是在生物计算、统计领域，变量间的交互效应是关键特征之一，这能够识别在复杂系统中的组成部分是如何相互协同，对生物性结果产生影响。