第 2 组的数据挖掘作业 4

设计数据挖掘研究方案

小组成员：高鹏昺 蒋世豪 李进雄 周亮 苏金涛 刘昊轩

1. 研究目的

探究生物节律基因和睡眠时间与年龄、寿命的关联性。

2. 理论背景介绍

通过对相关论文的研读，可以初步看出，人类的年龄与 DNA 甲基化和基因表达之间存在一定的关系，虽然其中的关系不是非常明显，但是通过前人的探索，可以知道存在两种不同的差异甲基化——即正关联和负关联的差异甲基化，关联着调控模式。既然这种差异基因能够调控年龄，那么与寿命也存在一定的关联性。

之前的论文单单只是探究了基因的调控，也就是人类的内部因素，那么来自于外界的刺激，比如睡眠时间或者其他的因素对于调控的影响尚不可知。同时，根据我们的常识，一个人的作息其实对于寿命的影响很大，那么生物节律基因又对年龄表现和寿命有什么影响呢？这就需要我们分析相关性，并通过相关数据挖掘方案进行挖掘。

其次，不同个体之间必然存在着差距，差异甲基化调控着人类的年龄表现，但是我们无法从一个人的外表体征明确给出一个人的年龄。所以我们需要研究个体的实际年龄与该年龄的正常（平均）表现。比如我们采集的样本年龄为 30-40，但是该年龄区间的个体都经常处于加班、睡眠不足等状况，那么最为直观的生理年龄表现就要更大（就是通常说的“显老”）。所以我们需要通过探究生物节律基因和睡眠时间对于差异甲基化的正常表达的影响，来规范我们的样本。

3. 数据挖掘方法

3.1 基因关联性分析

我们获取了不同年龄段的人的生物节律基因数据、平均睡眠时长数据、还有差异甲基化的相关数据。

首先，我们对于两者对于差异甲基化的影响也只是存在猜想阶段，所以有必要进行初步的验证尝试，即做关联性分析。

（1）确定反映系统行为特征的参考数列和影响系统行为的比较数列

反映系统行为特征的数据序列，称为参考数列。影响系统行为的因素组成的数据序列，称比较数列。

（2）对参考数列和比较数列进行无量纲化处理

由于系统中各因素的物理意义不同，导致数据的量纲也不一定相同，不便于比较，或在比较时难以得到正确的结论。因此在进行灰色关联度分析时，一般都要进行无量纲化的数据处理。

（3）求参考数列与比较数列的灰色关联系数ξ（Xi）

所谓关联程度，实质上是曲线间几何形状的差别程度。因此曲线间差值大小，可作为关联程度的衡量尺度。对于一个参考数列 X0 有若干个比较数列 X1, X2,…, Xn，各比较数列与参考数列在各个时刻（即曲线中的各点）的关联系数ξ（Xi）可由下列公式算出：其中 ρ 为分辨系数，一般在 0~1 之间，通常取 0.5。

是第二级最小差，记为 Δmin。 是两级最大差，记为 Δmax。

为各比较数列 Xi 曲线上的每一个点与参考数列 X0 曲线上的每一个点的绝对差值，记为 Δoi(k)。

所以关联系数ξ（Xi）也可简化如下列公式：

$$\xi_{0i} = \frac{\Delta(\min) + \rho\Delta(\max)}{\Delta_{0i}(k) + \rho\Delta(\max)}$$

（4）求关联度r_i

因为关联系数是比较数列与参考数列在各个时刻（即曲线中的各点）的关联程度值，所以它的数不止一个，而信息过于分散不便于进行整体性比较。因此有必要将各个时刻（即曲线中的各点）的关联系数集中为一个值，即求其平均值，作为比较数列与参考数列间关联程度的数量表示，关联度r_i

公式如下：

$$r_i = \frac{1}{N}\sum_{k=1}^{N}\xi_i(k)$$

ri——比较数列 xi 对参考数列 x0 的灰关联度，或称为序列关联度、平均关联度、线关联度。

ri 值越接近 1，说明关联性越好。

（5）关联度排序

因素间的关联程度，主要是用关联度的大小次序描述，而不仅是关联度的大小。将 m 个子序列对同一母序列的关联度按大小顺序排列起来，便组成了关联序，记为{x}，它反映了对于母序列来说各子序列的“优劣”关系。若 r0i>r0j，则称{xi}对于同一母序列{x0}优于{xj}，记为{xi}>{xj} ；r0i 表示第 i 个子序列对母数列特征值。

灰色关联度分析法是将研究对象及影响因素的因子值视为一条线上的点,与待识别对象及影响因素的因子值所绘制的曲线进行比较,比较它们之间的贴近度,并分别量化,计算出研究对象与待识别对象各影响因素之间的贴近程度的关联度,通过比较各关联度的大小来判断待识别对象对研究对象的影响程度。

3.2 基因序列比对分析

生物学序列通常是指核苷酸(DNA、RNA)或氨基酸序列，生物学序列分析比较、比对、索引和分析生物学序列。比对(alignment)是对序列排序以便获取最大程度的一致性，它也表示序列之间的相似程度。通过序列比对得到的相似性在确定两个序列同源的可能性时很有用。

生物序列比对的问题可以描述如下：对于给定的两个或多个输入生物序列，识别具有长保守自序列的相似序列。如果比对序列个数恰为 2，则称该问题为双序列比对；否则为多序列比对。待比较和比对的序列可以是核苷酸或氨基酸。对于核苷酸来说，如果两个符号相同则它们对齐；对于氨基酸来说，如果两个符号相同，或者一个可以通过可能自然出现的替换从另一个得到，则它们对齐。比对有局部比对和全局比对两种方式，前者表示仅有部分序列进行比对，后者需要在序列的整个长度上进行比对。

3.3 TF-IDF 基因序列挖掘

这里的应用，我们主要考虑的是创新性的（在我们的认知范围）将文本数据挖掘的方法 TF-IDF 作为一种辅助特征引入，像文本挖掘中的 TF-IDF 的作用一样，让程序自主提取出一些关键性的基因，我们的想法比较简单，就是首先将基因序列映射，然后提取中固定数量的 TF-IDF 值较高的一些基因作为研究的辅助特征，下面是我们的应用方式。

把 TF-IDF 的核心思想应用到本课题，如果一个词基因在一类人中的出现频率比较高，然后在我们的数据集的所有人中出现的次数比较少，这个基因就是一个比较好的分类特征。TF 是一个基因在某类人的出现频率，IDF 是逆向出现频率。

TF 是一个基因在某类人的出现频率，对一个基因 i，一类人 j，$tf_{i,j}$的计算公式是：

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{k,j}}$$

$n_{i,j}$是基因 i 在人群 j 中的次数，分母是人群中人的个数。

IDF 是用如下公式计算的：

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

|D|:是总的人数

$|\{j : t_i \in d_j\}|$有基因 i 的人的个数

最终一个基因的 TF-IDF 值计算方法如下：

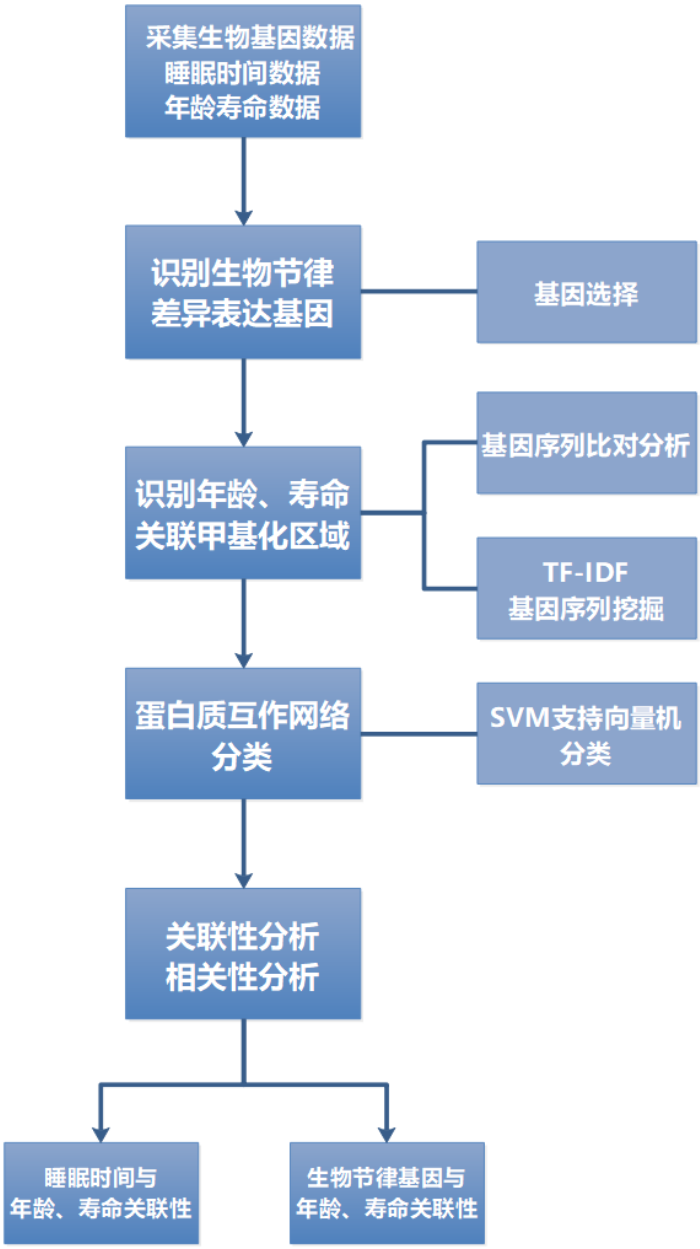
$$tfidf_{ij} = tf_{i,j} * idf_i$$

我们可以提取 TF-IDF 值较高的一些特征作为一些分类特征。

4. 方案流程

① 采集生物基因数据、睡眠时间和年龄寿命数据
② 识别生物节律差异表达基因
③ 识别年龄、寿命关联甲基化区域
④ 蛋白质互作用网络分类
⑤ 基因关联性、相关性分析

流程图:



5. 输入数据

(1) 生物基因及其注释数据
(2) 睡眠时间和年龄寿命数据
(3) 基因表达数据
(4) DNA 甲基化数据
(5) 蛋白质互作网络数据

6. 输出数据

1. 生物节律差异表达基因数据
2. 年龄、寿命关联甲基化区域数据
3. 蛋白质互作网络分类拓扑数据
4. 基因关联性、相关性数据

7. 实验结果

(1) 睡眠时间与年龄、寿命的关联性
(2) 生物节律基因与年龄、寿命的关联性