基于熵权法与 Topsis 的关键蛋白质识别模型

摘要

蛋白质是一类重要的生物大分子，但蛋白质的种类众多，仅在智人这一物种上就由 20 多万种。为更好研究蛋白质，借助蛋白质相互作用的特性来定量分析蛋白质的重要性。

首先，根据**网络**的相关知识，发现在有权网路中存在 5 个能够反映节点在网络中的重要性的指标，分别为**度中心性**、**中介中心性**、**接近度中心性**、**特征向量中心性**、**局部平均连接度**。分别运用这些指标对蛋白质网络中的蛋白质进行重要性排序，经过统计，发现 **P14618**、**P04637**、**Q96RG2** 三个蛋白在多个指标中均位列前五。认为这三个蛋白在网络中很关键。

为了得出一个定量分析指标，对蛋白质重要性进行定量排序，本文运用了**熵权法**和 **TOPSIS** 对指标进行加权，并运用 **Topsis 归一化**量化指标，最终得到所有蛋白质的重要性。对它们进行排序可得，前五位重要蛋白分别为，**P62753**、**P62269**、**P62277**、**P04637**、**P62701**。

最后，根据 PMID 数据分析各年份的研究热点。本文运用 pythn 爬虫在网络上抓取了 pubmed 上 PMID 所对应的论文发表的时间数据，截取 2000 年-2018 年的数据进行分析。运用问题二所得的评价方法，选取每年中重要性排名前 50 的蛋白质进行蛋白质功能分析。通过 **KEGG 数据库**查询蛋白质所代表的信号通路，从中归纳出蛋白质的功能标签，并按功能分类计算每年的研究热点方向。结论：例如**人类病毒性疾病**的研究热点变化趋势具有阶段性，契合当年的公共卫生需求；例如**遗传信息处理**和**细胞过程：运输与转录**在某些年份同时到达研究热点，认为它们之间具有领域的重叠等等。

关键字： 网络　度中心性　中介中心性　接近度中心性　特征向量中心性　局部平均连接度　熵权法　TOPSIS　KEGG 数据库

一、 问题重述

1.1 问题的提出

众所周知，蛋白质是最重要的生物大分子之一，在生物体内执行大量的功能，也是被研究最多的生物大分子（在 pubmed 中包含蛋白质的文献数大于包含核酸、脂类和糖类的论文总和）。研究蛋白质之间的相互作用（protein interaction）有助于揭示生命的秘密。

图 **1**　网络图示例：拟南芥

1.2 问题一

在 2013 年发表的一篇论文（Nat Methods. 2013 Aug;10(8):690-1.doi: 10.1038/nmeth.2561），将已经发表的文献中的蛋白质相互作用收集起来，每周加以更新，从 2013 年到现在的近 10 年间，已经有 70 多万对蛋白质相互作用被收集。该数据格式简单，每一行包含一对蛋白质之间的相互作用，用分号分隔了 8 个数据，即：Protein A, Gene A, Taxon A, Protein B, Gene B, Taxon B, Score, PMID. 根据该数据集或者借助其他公开发表的信息, 找出不超过 5 个重要的蛋白质，说明它们为什么重要。

1.3 问题二

在问题 1 的基础上，提出一种衡量蛋白质重要性的量化指标，试论证其合理性。

1.4 问题三

在该数据集的已有近 500 个版本（每周更新）中，寻找学术界研究兴趣的变化，结合问题 2 展开分析。

二、 问题分析

2.1 问题一分析

针对问题一，我们首先需要确定判断何为重要蛋白质的标准指标，再通过对数据集进行指标的处理及量化排序，获得在该指标下较为重要的蛋白质，并对各个指标的重要蛋白质进行对比统计分析，得出不超过五个重要蛋白质。

2.2 问题二分析

针对问题二，多种标准指标已经无法解决，题目需要一种量化指标衡量蛋白质重要性，所以我们需要将标准指标对重要性的影响力用加权的方式表示，再结合整个蛋白质网络对该蛋白质进行综合各个指标的总体评分，即可得到一个较为科学的用量化指标评判重要性的方法。

2.3 问题三分析

问题三要求在已更新的多个数据集中寻找学术界研究兴趣的变化。为此，首先下载这些数据。根据 PMID 找到他们的论文发表时间。并按照发表时间顺序，对所研究的蛋白进行分析。接着，搜索这些蛋白质在 KEGG 数据库中属于什么信号通路，并按功能对这些蛋白进行注释。画出其时序图，观察研究热点主要聚焦在什么功能上。

三、 模型的假设

在建立和分析模型时，以下假设至关重要：

- 数据集中所有蛋白质的功能和关联性已经被全面分析和记录。
- 使用的蛋白质网络指标（如度中心性、中介中心性等）能够充分反映蛋白质在网络中的重要性。
- 各指标的权重确定方法（如熵权法）是合理且能够反映蛋白质间的相对重要性。
- 模型分析的蛋白质网络在一定时间内是稳定的，即不受外部变化的显著影响。

四、 符号说明

符号	意义
$DC(u)$	无权网络中蛋白质节点 u 的度中心性
$DC^w(u)$	加权网络中蛋白质节点 u 的度中心性
$BC(u)$	无权网络中蛋白质节点 u 的中介中心性
$BC^w(u)$	加权网络中蛋白质节点 u 的中介中心性
$CC(u)$	无权网络中蛋白质节点 u 的接近度中心性
$CC^w(u)$	加权网络中蛋白质节点 u 的接近度中心性
$EC(u)$	蛋白质节点 u 的特征向量中心性
$EC^w(u)$	加权网络中蛋白质节点 u 的特征向量中心性
$LAC(u)$	无权网络中蛋白质节点 u 的局部平均连通性
$LAC^w(u)$	加权网络中蛋白质节点 u 的局部平均连通性

五、 模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 模型建立

根据查阅到的文献，我们获得了五个判断蛋白质重要性的标准指标，分别为度中心性、接近度中心性、中介中心性、特征向量中心性、局部平均连通性，其将每个蛋白质视为节点，将整个蛋白质数据集视为一个联通网络，用节点的中心性来表示其在蛋白质网络中的重要程度，指标的具体描述和计算公式分别为：

1 度中心性

它是刻画节点中心性最直接的指标，一个节点的度越大就意味其在网络中就越重要、越具有影响力。在无权网络中，一个节点的度是指其直接相连的邻居节点的个数，即对于蛋白质节点 u，其度中心性为：

$$DC(u) = \sum_{v} a_{uv} \tag{1}$$

而在加权网络中，一个节点的度是指其与邻居节点的边的权重之和，因此对于蛋白

质节点 u，其在加权网络中的度中心性为：

$$DC^w(u) = \sum_v w_{uv} \tag{2}$$

2 中介中心性（Betweenness Centrality，BC）

它以经过该节点的最短路径的个数来反映其中心性。在无权网络中，对于蛋白质节点 u，其介数中心性为：

$$BC(u) = \sum_s \sum_t \frac{\rho(s,u,t)}{\rho(s,t)}, s \neq u \neq t \tag{3}$$

其中，$\rho(s,t)$ 表示节点 s 和 t 之间最短路径的数量，$\rho(s,u,t)$ 表示 s 和 t 之间经过节点 u 的最短路径的数量。而在加权网络中，介数中心性的定义并没有发生本质变化，即对于蛋白质节点 u，其在加权网络中的介数中心性为：

$$BC^w(u) = \sum_s \sum_t \frac{\rho^w(s,u,t)}{\rho^w(s,t)}, s \neq u \neq t \tag{4}$$

$BC(u)$ 和 $BC^w(u)$ 最主要的区别是关于最短路径的计算方法。在无权网络中，两个节点之间的最短路径是指所含边数目最短的路径，而在加权网络中是指所含边权重之和最少的路径。

3 接近度中心性（Closeness Centrality，CC）

它反映了节点与其它节点之间的接近程度。在无权网络中，对于节点 u，其接近度中心性为：

$$CC(u) = \frac{N-1}{\sum_v d(u,v)} \tag{5}$$

其中 $d(u,v)$ 是节点 u 到 v 的最短路径。和加权网络中的中介度中心性 $BC^w(u)$ 一样，加权网络中的接近度中心性 $CC^w(u)$ 定义和 $CC(u)$ 概念相似，不同的是计算最短路径的方式。即对于蛋白质节点 u，其在加权网络中的接近度中心性为：

$$CC^w(u) = \frac{N-1}{\sum_v d^w(u,v)} \tag{6}$$

4 特征向量中心性（Eigenvector Centrality，EC）

它认为节点的重要性不仅取决于其邻居节点的数量，也和其邻居节点的重要性相关，对于节点 u，其特征向量中心性为：

$$EC(u) = \alpha_{\max}(u) \tag{7}$$

其中，$\alpha_{\max}$ 是对应于网络邻接矩阵 $A = a_{uv}$ 的最大特征值 $\lambda_{\max}$ 的特征向量，即 A 的

主特征向量，而 $\alpha_{\max}(u)$ 是 $G(V,E)$ 的第 u 个分量。同理在加权网络中，对于节点 u，其特征向量中心性为：

$$EC^w(u) = w_{\max}(u) \tag{8}$$

其中 $w_{\max}$ 是加权网络对应的邻接矩阵 $A^w = w_{uv}$ 的最大特征值的特征向量。

5 局部平均连通度（Local Average Connectivity-based，LAC）

它是指对于某一节点，将其邻居节点当作一个子图，统计每个邻居节点在该子图中的局部度数，然后再计算这些邻居节点的平均局部度数，反映该节点的邻居节点之间的紧密程度。在无权网络中，对于蛋白质节点 u，其局部平均连通度为：

$$LAC(u) = \frac{1}{|N_v|}\sum_{v\in N(v)} \deg_{C_u}(v) \tag{9}$$

其中，N_v 是蛋白质节点 u 的邻居节点的集合，v 是蛋白质节点 u 且 $v \in N(v)$,C_u 是由 N_v 构成的一个子图，$\deg_{C_u}(v)$ 表示邻居节点 v 在子图 C_u 中的局部度数。同理，在加权网络中，对于蛋白质节点 u，其局部平均连通度为：

$$LAC^w(u) = \frac{1}{|N_v|}\sum_{v\in N(v)} \deg^w_{C_u}(v) \tag{10}$$

5.1.2 模型求解

经计算得到各指标前五位蛋白质结果：

表 **1**　各指标前五位蛋白质

编号	度中心性	编号	中介中心性 ($\times 10^8$)	编号	接近度中心性 ($\times 10^{-4}$)	编号	特征向量中心性	编号	局部平均连通性
P11484	613.392	P62259	4.2474	**P14618**	0.2142	**P04637**	0.0029	P62701	61.4886
P10591	481.031	P02687	3.794	**Q96RG2**	0.2122	Q09472	0.002	P30050	56.5059
P05067	465.473	Q1EC66	2.9646	P22314	0.2111	Q00987	0.0019	P62906	56.4918
P10592	400.852	**Q96RG2**	2.7817	P08581	0.209	P03372	0.0019	P46777	56.1014
P04637	392.855	**P14618**	2.7578	P28482	0.2079	P0CG48	0.0019	P62424	55.661

经统计可以得到编号为 P14618、P04637、Q96RG2 在多个指标中均在前五位，说明这三个蛋白质是在蛋白质网络中占有重要地位的关键蛋白质。

5.1.3 结果分析

我们接着查阅了 P14618、P04637、Q96RG2 蛋白质的相关资料，并对结果作如下分析：

P14618 名称：Cyclin-dependent kinase inhibitor 1

功能：P14618 蛋白质参与调控细胞周期，通过抑制 CDK（Cyclin-dependent kinase）活性来控制细胞周期进程。除了 CDK 之外，它还能与 PCNA（Proliferating Cell Nuclear Antigen）等蛋白质相互作用，调控 DNA 复制和修复。

分析：细胞周期作为细胞生命阶段的代表概念，其中的 G1 期的分裂准备阶段涉及很多蛋白质的信号传递，而 P14618 在其中具有重要地位，与 CDK、PCNA 等蛋白质都有着重要的相互作用，而这也决定了其在蛋白质网络中的重要性。

P04637 名称：Cellular tumor antigen p53

功能：其在细胞周期调控中起到肿瘤抑制的作用，作为转录因子，其能够激活多种基因的转录，这些基因参与 DNA 修复、细胞周期抑制、凋亡和衰老等过程。其还能够诱导细胞凋亡、参与细胞衰老的调控。在细胞受到各种应激（如放射、化学药物、缺氧等）时，其会被激活并响应这些应激信号，调控细胞的生存和死亡。

分析：P04637 的功能繁多，在肿瘤抑制、转录、细胞衰老与凋亡、细胞应激反应中都起着很大作用，在发挥这些作用时，P04637 会与目标蛋白质相结合，而这种信息枢纽的特质也决定了其在蛋白质网络中不可或缺的作用。

Q96RG2 名称：Cell division cycle-associated 7-like protein

功能：Q96RG2 在细胞周期的 G1 期和 S 期之间的转换过程中发挥了重要作用，对细胞增殖起促进作用，另外，它也通过与 DNA 或其他转录因子相互作用，调节目标基因的转录活性。

分析：Q96RG2 可能通过与其他细胞周期蛋白和调控因子相互作用来调节细胞分裂和增殖，会涉及与其他蛋白质的结合，并且通过与其他转录因子相互作用来参与调控基因转录，说明其与很多蛋白质因子都具有结合特性，决定了其在蛋白质网络中的重要性。

5.2 问题二模型的建立与求解

5.2.1 熵权法指标加权

- **数据处理**

根据问题分析，我们先对数据进行初步处理获得各指标的基本数据特征如下：

表 **2**　蛋白质网络各指标数据特征

	度中心性	接近度中心性	中介度中心性	特征向量中心性	局部平均连通度
最大值	613.391999999958	0.000021422683	424739477.562511	0.002914006981	61.488584474886
最小值	0.033	0	0	0	0
方差	109.374696289837	0.000000000024	154508951116O1.2	0.000000002303	17.728439645236

可以发现如中介度中心性方差很大，数据变化很大，包含信息量较大，而特征向量

中心性方差较小，数据变化量小，包含信息量较小，各指标存在信息的不一致性，适合用熵权法加权。

- **模型建立**

熵权法利用信息熵来确定各评价指标的权重。信息熵反映了指标的信息量，信息量越大，权重越小。模型建立步骤如下：

1 标准化各指标数据

$$x_{ij} = \frac{a_{ij} - \min(a_{*j})}{\max(a_{*j}) - \min(a_{*j})} \tag{11}$$

其中，a_{*j} 为指标数据集。

2 计算第 j 个指标在第 i 个蛋白质上的概率分布

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \tag{12}$$

3 根据概率分布计算第 j 个指标的熵值：

$$e_j = -k \sum_{i=1}^{m} p_{ij} \ln(p_{ij}) \tag{13}$$

其中 $k = \frac{1}{\ln(m)}$。

4 计算差异系数

$$d_j = 1 - e_j \tag{14}$$

5 计算最终权重

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \tag{15}$$

- **模型求解**

利用熵权法求解得各指标的权重如下表所示：

表 **3**　各指标权重

	度中心性	接近度中心性	中介度中心性	特征向量中心性	局部平均连通度
权重	0.058104	0.684664	0.005047	0.035127	0.217058

5.2.2 Topsis 法归纳单一量化指标

- **模型简介**

TOPSIS 法（逼近理想解的排序方法）是一种多准则决策分析方法，通过计算各方案与理想解和负理想解的距离来进行排序。该方法的基本思想是最优方案应当距离理想解最

近，距离负理想解最远。其通过构建加权标准化决策矩阵、确定正负理想解、计算得出相对接近度的方法对方案进行排序，其中的相对接近度便可成为我们的单一量化指标。

- **模型建立**

结合熵权法获得的权重及 Topsis 法建立模型步骤如下：

1 使用各指标的权重 w_j 构建加权标准化决策矩阵 $V = [v_{ij}]$，公式为

$$v_{ij} = w_j x_{ij} \tag{16}$$

2 确定正理想解与负理想解

$$\begin{cases} A^+ = \{\max v_{ij} \mid j \in J\} \\ A^- = \{\min v_{ij} \mid j \in J\} \end{cases} \tag{17}$$

其中 A^+ 为正理想解，A^- 为负理想解。

3 计算各个蛋白质与理想解和负理想解的距离，公式为：

$$\begin{cases} S_i^+ = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^+)^2} \\ S_i^- = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^-)^2} \end{cases} \tag{18}$$

4 计算各个蛋白质的相对接近度

$$C_i = \frac{S_i^-}{S_i^+ + S_i^-} \tag{19}$$

- **模型求解**

经过 Topsis 法求解，得到了各个蛋白质的相对接近度，实现了单一指标的评判。

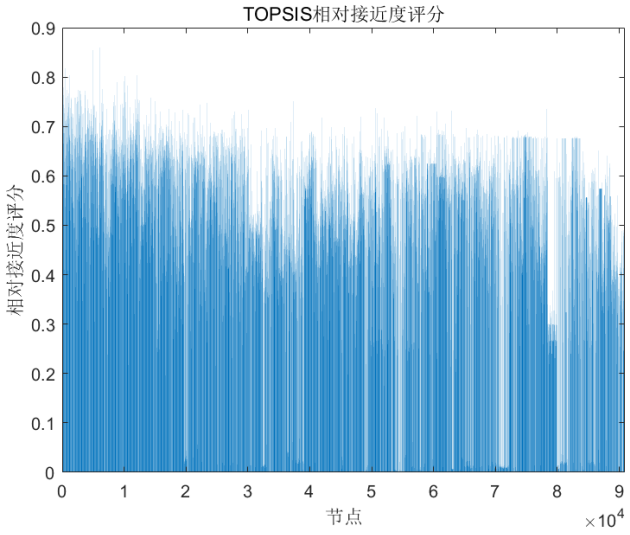

图 **2**　**Topsis** 相对接近度评分

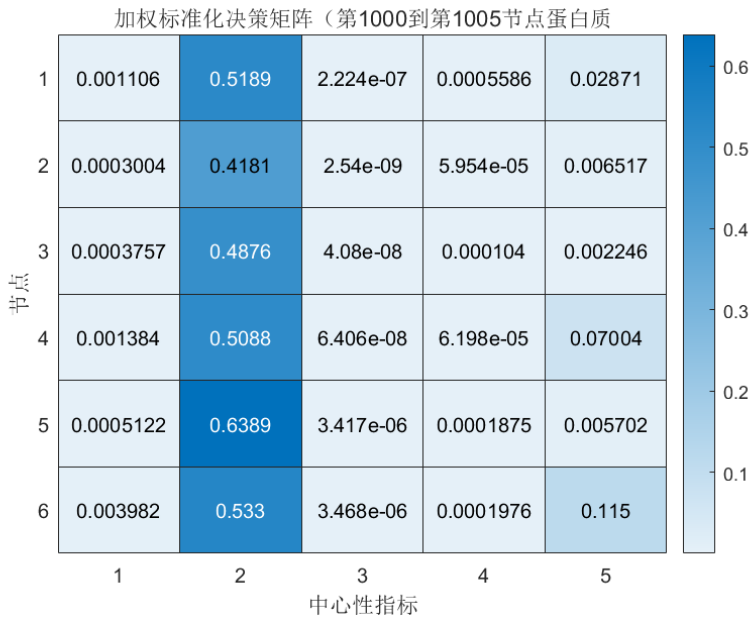

图 **3**　中心性指标

表 **4**　**Topsis** 法评分前五位蛋白质

蛋白质编号	评分
P62753	0.882457
P62269	0.88212
P62277	0.876674
P04637	0.873789
P62701	0.868244

5.3 问题三模型的建立与求解

5.3.1 时间信息查找

根据 PMID 的作用，可通过 PMID 精确查找一篇文献的发表时间。为了大量获取文献的发表时间，我们运用 python 爬虫进行批量数据抓取。

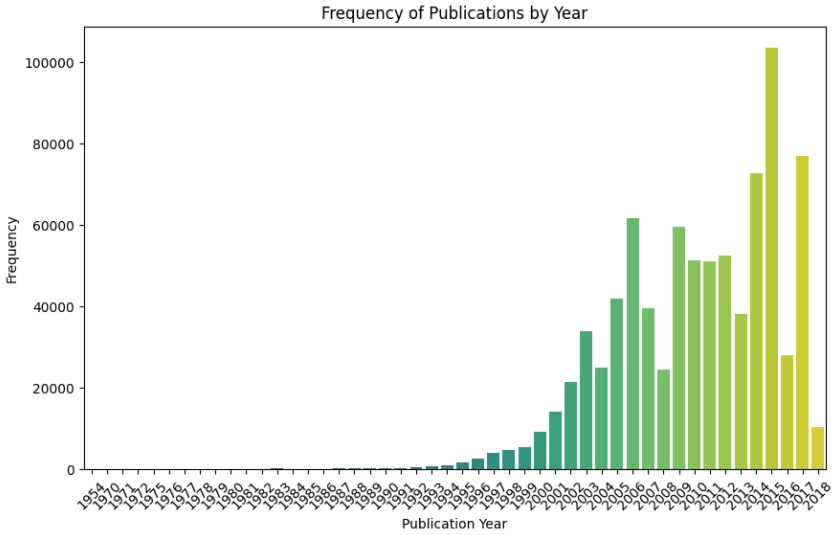

图 **4**　各年份发表论文数量柱状图

观察数据，发现相关论文从 2000 年开始才有稳定的上升，为了更好的反映研究热点，选取 2000 年至 2018 年的数据进行下一步分析。

5.3.2 蛋白质重要性分析

在以年划分的数据中，对蛋白质在按问题二给出的计算标准进行重要性进行排序，取每年数据的前 50 个蛋白质，认为他们是当年的热点研究蛋白。

5.3.3 蛋白功能查找

通过查阅文献可知，人类体中蛋白质数量超过 20 多万，并且，蛋白质不是孤立地发挥作用的，而是通过蛋白质互作用网络来影响生命体中的一些功能，直接分析蛋白而得出研究热点的做法是不可取的。研究热点应该是特定功能所代表的蛋白质集合集合。

为此，再次运用 python 爬虫抓取每年重要性前 50 的蛋白质所代表的信号通路及其功能，并对功能进行归类整理。由此得到当年的研究热点。

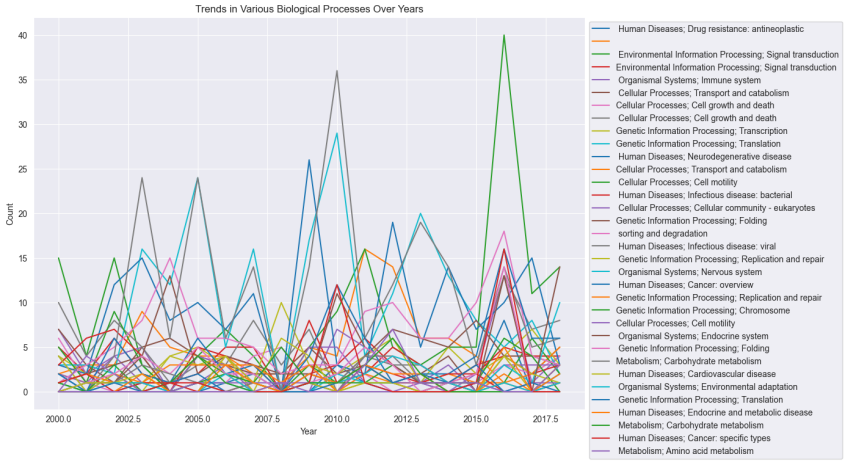

图 **5**　各年份研究热点功能折线图

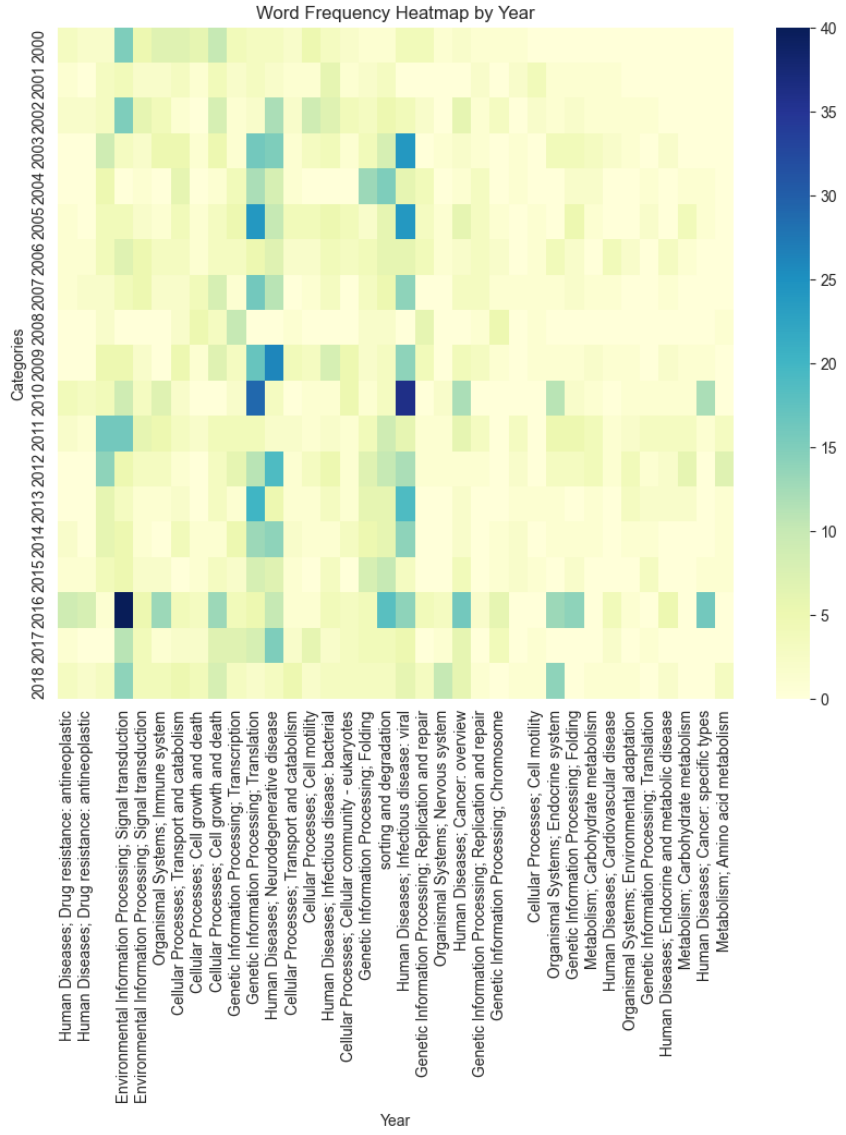

图 **6**　各年份研究热点功能热图

5.3.4 研究热点分析

由上文所得结果可知，研究热点的变化存在阶段性、突变性，即各年份的研究热点具有明显差异，而且研究热点的变化极大。因此我们得出以下一些比较显著的结论。

- 一、研究热点的生命周期
- 初期萌芽: 某些研究领域在图中最初的年份几乎没有活跃度，这可能意味着这些领域当时刚刚兴起，研究还未形成规模。例如，某些类别在 2000 年至 2005 年之间几乎没有词频，但在随后的几年中逐渐增长，说明这些领域可能在初期只有少数研究人员关注，随着时间推移，研究逐渐增加。
- 发展与巅峰: 图中的深蓝色区域表明某些研究领域在特定年份达到了其研究的巅峰状态。这可能是由于技术突破、重大科学发现或特定领域的政策支持所致。分析这些巅峰期可以揭示出关键的科学进展时间点。
- 衰退与成熟: 一些研究领域的词频在巅峰之后逐渐减弱，表明该领域可能逐渐成熟，或者已经解决了当时的主要科
- 二、研究领域的持久性与变迁
- 持久热门领域: 某些研究领域在整个时间范围内都保持较高的词频，如“Signal Transduction”（信号转导）和“Cellular Processes: Transport and catabolism”（细胞过程：运输和代谢）。这些领域可能涉及到基础性的问题，广泛应用于多个生物学和医学领域，因此能够长期保持研究的热度。
- 阶段性热门领域: 另一些领域如“Human Diseases: Infectious diseases: viral”（人类疾病：传染病：病毒性疾病）可能在特定的历史背景下，如疫情暴发或特定疾病的研究突破时，出现阶段性的研究高峰。这些研究的活跃期往往与社会需求或公共卫生事件密切相关。
- 三、多领域交叉和集成研究的趋势
- 跨学科的融合: 热图中可以观察到一些研究类别在多个时间点上都表现出一定的频次，且这些类别之间存在一定的交叉，这表明随着时间的推移，不同学科之间的交叉融合成为一种趋势。例如，“Genetic Information Processing: Transcription”（遗传信息处理：转录）与“Cellular Processes: Transport and catabolism”（细胞过程：运输和代谢）可能在某些年份中表现出同时高频，表明两者之间的关联性被越来越多的研究所关注。
- 综合研究的崛起: 随着科学研究的深入，研究者们越来越多地关注综合性研究。例如，基于系统生物学的方法可能整合了基因表达、代谢网络、信号通路等多个领域，推动了这些领域在研究中的频繁出现和相互影响。
- 四、外部驱动因素对研究热点的影响
- 技术进步: 例如，随着基因组学和蛋白质组学技术的发展，与“Genetic Information

Processing”（遗传信息处理）相关的研究可能在某些年份中频次增加，这反映了技术进步对科学研究的驱动作用。

- 公共卫生事件: 某些疾病领域（如“Human Diseases: Infectious diseases: viral”）的研究热度与当时的公共卫生事件有直接联系，如 SARS、H1N1 或 COVID-19 疫情，这些事件会显著推动相关领域的研究。
- 政策和资金导向: 政府或机构的研究资金分配和政策导向也会影响某些领域的研究热度，例如抗癌研究的长期持续性可能与各国政府对癌症研究的资金投入有关。。

参考文献

[1] 刘桂霞, 曹心恬, 赵贺. 基于特征图网络和多种生物信息预测关键蛋白质的深度学习框架 [J]. 吉林大学学报 (理学版), 2024, 62 (03): 593-605. DOI:10.13413/j.cnki.jdxblxb.2023227.

[2] 程启月. 评测指标权重确定的结构熵权法 [J]. 系统工程理论与实践,2010,30(07):1225-1228.

[3] 章穗, 张梅, 迟国泰. 基于熵权法的科学技术评价模型及其实证研究 [J]. 管理学报,2010,7(01):34-42.

[4] 虞晓芬, 傅玳. 多指标综合评价方法综述 [J]. 统计与决策,2004,(11):119-121.

[5] 杜挺, 谢贤健, 梁海艳, 等. 基于熵权 TOPSIS 和 GIS 的重庆市县域经济综合评价及空间分析 [J]. 经济地理,2014,34(06):40-47.DOI:10.15957/j.cnki.jjdl.2014.06.026.

[6] 杜挺, 谢贤健, 梁海艳, 等. 基于熵权 TOPSIS 和 GIS 的重庆市县域经济综合评价及空间分析 [J]. 经济地理,2014,34(06):40-47.DOI:10.15957/j.cnki.jjdl.2014.06.026.

[7] 蒋争凡, 翟中和. 细胞凋亡——当前生命科学研究的热点课题 [J]. 科学通报,1999,(18):1920-1928.

[8] 田德桥. 生命科学两用性研究关注热点的文献计量分析 [J]. 生物技术通讯,2016,27(05):684-687.

附录 A　问题一代码–matlab 源程序

```
% 计算度中心性
degreeCentrality = centrality(G, 'degree', 'Importance', G.Edges.Weight);

% 计算接近中心性
closenessCentrality = centrality(G, 'closeness', 'Cost', G.Edges.Weight);

% 计算中介中心性
betweennessCentrality = centrality(G, 'betweenness', 'Cost', G.Edges.Weight);

% 计算特征向量中心性
eigenvectorCentrality = centrality(G, 'eigenvector', 'Importance', G.Edges.Weight);



% 计算局部平均连通性
localAverageConnectivity = zeros(numnodes(G), 1);
for i = 1:numnodes(G)
nbrs = neighbors(G, i);
if numel(nbrs) > 1
subG = subgraph(G, nbrs);
localAverageConnectivity(i) = mean(degree(subG));
else
localAverageConnectivity(i) = 0;
end
end

% 获取每个指标前五的点的下标
numTopNodes = 5;

[~, degreeTopIdx] = maxk(degreeCentrality, numTopNodes);
[~, closenessTopIdx] = maxk(closenessCentrality, numTopNodes);
[~, betweennessTopIdx] = maxk(betweennessCentrality, numTopNodes);
[~, eigenvectorTopIdx] = maxk(eigenvectorCentrality, numTopNodes);
[~, localAvgConnTopIdx] = maxk(localAverageConnectivity, numTopNodes);

% 合并所有的下标
allTopIdx = [degreeTopIdx; closenessTopIdx; betweennessTopIdx; eigenvectorTopIdx;
    localAvgConnTopIdx];

% 记录重复的下标及其重复次数
[uniqueTopIdx, ~, idxCounts] = unique(allTopIdx);
repeatCounts = histcounts(idxCounts, 1:max(idxCounts)+1);

% 筛选出重复的下标及其重复次数
repeatedIdx = uniqueTopIdx(repeatCounts > 1);
```

```
repeatedCounts = repeatCounts(repeatCounts > 1);

% 按重复次数进行排序
[sortedCounts, sortIdx] = sort(repeatedCounts, 'descend');
sortedRepeatedIdx = repeatedIdx(sortIdx);

% 打印结果

disp('按重复次数排序的重复点的下标:');
disp(sortedRepeatedIdx');
disp('对应的重复次数:');
disp(sortedCounts');
```

附录 B　Topsis 代码–matlab 源程序

```
allCentrality=[degreeCentrality,closenessCentrality,betweennessCentrality,eigenvectorCentrality,localAverageC
% 标准化数据（最大-最小标准化）
data_min = min(allCentrality);
data_max = max(allCentrality);
data_norm = (allCentrality - data_min) ./ (data_max - data_min);

% 计算信息熵
[m, n] = size(allCentrality);
k = 1 / log(m);
E = -k * sum(data_norm .* log(data_norm + eps))';

% 计算冗余度
d = 1 - E;

% 计算权重
w = d / sum(d);

% TOPSIS法
% 构建加权标准化决策矩阵
data_weighted = data_norm .* repmat(w', m, 1);

% 确定正理想解和负理想解
V_pos = max(data_weighted);
V_neg = min(data_weighted);

% 计算各方案到正、负理想解的距离
D_pos = sqrt(sum((data_weighted - repmat(V_pos, m, 1)).^2, 2));
D_neg = sqrt(sum((data_weighted - repmat(V_neg, m, 1)).^2, 2));

% 计算相对接近度评分
```

```
C = D_neg ./ (D_pos + D_neg);
% 输出权重
[val,index]=maxk(C,10);
```

附录 C　获取论文时间数据——python 代码

```
import requests
import xml.etree.ElementTree as ET
from datetime import datetime
import pandas as pd
import os

def fetch_pub_date(pmid_list):
"""
获取一批PMID对应的发表日期。

参数:
pmid_list (list): PMID字符串的列表。

返回:
dict: PMID到发表日期的映射。
"""
pub_dates = {}
for pmid in pmid_list:
url =
    f"https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=pubmed&id={pmid}&retmode=xml"
try:
response = requests.get(url)
response.raise_for_status() # 如果状态码不是200，引发HTTPError
except requests.exceptions.RequestException as e:
print(f"Error fetching PMID {pmid}: {e}")
pub_dates[pmid] = None
continue

root = ET.fromstring(response.content)
pub_date = None
for item in root.findall('.//Item[@Name="PubDate"]'):
pub_date = item.text
pub_dates[pmid] = pub_date
return pub_dates

def convert_to_datetime(date_str):
"""
将字符串日期转换为datetime对象。
```

```
参数:
date_str (str): 日期字符串。

返回:
datetime: datetime对象，如果转换失败则返回None。
"""
if date_str:
try:
return datetime.strptime(date_str, "%Y %b %d")
except ValueError:
try:
return datetime.strptime(date_str, "%Y %b")
except ValueError:
try:
return datetime.strptime(date_str, "%Y")
except ValueError:
return None
return None

# 读取Excel文件
df = pd.read_excel('D:/zhuomian/实战模拟/2/pmid顺序.xlsx', sheet_name='Sheet2')
df = df.drop_duplicates()

# 分批处理，每批3000条数据
batch_size = 3000
pub_dates = {}

# 检查是否有已保存的中间结果文件
checkpoint_file = 'pub_dates_checkpoint.pkl'
if os.path.exists(checkpoint_file):
pub_dates = pd.read_pickle(checkpoint_file)
processed_pmids = set(pub_dates.keys())
else:
processed_pmids = set()

for start in range(0, len(df), batch_size):
end = min(start + batch_size, len(df))
batch_pmids = df.iloc[start:end]['PMID'].astype(str).tolist()

# 跳过已处理的PMID
batch_pmids_to_process = [pmid for pmid in batch_pmids if pmid not in processed_pmids]
if not batch_pmids_to_process:
continue

# 获取并保存发表日期
batch_pub_dates = fetch_pub_date(batch_pmids_to_process)
pub_dates.update(batch_pub_dates)
```

```
# 更新已处理的PMID集合
processed_pmids.update(batch_pub_dates.keys())

# 保存中间结果到文件
pd.to_pickle(pub_dates, checkpoint_file)

# 将结果添加到DataFrame中
df['Publication Date'] = df['PMID'].astype(str).map(pub_dates)
df['Publication Date'] = df['Publication Date'].apply(convert_to_datetime)

# 删除中间结果文件（可选）
#os.remove(checkpoint_file)

# 显示结果
print(df)
```