

中国秘史---类比暴露组学和基因组学联系的研究 从历史智能到人工智能

王德奎 (Wang Dekui)

y-tx@163.com

摘要: 暴露组学研究是点对多点做相关---健康研究的真相,是多对多互相影响,控制实验当然是必要的,但如果数据是来自观测研究,那这问题就几乎无解。受研究共同体的视野限制,如果只关心那些强信号,可能忽略了那些弱信号。但这里的强弱,是仪器决定的,不是生物学意义决定的。或许很多人的研究,可以讲一个故事,但很难回答一个真实的问题。这只是现存问题的很小一部分,每一点的进展都可能对上下游研究产生颠覆式影响。对研究方法论的标准化、可重复化及与对基础研究进展的快速整合,是必要的。或许十年后回看今天的暴露组学,很多人可能惊叹:为什么大量的资源被浪费在了毫无意义的研究上?不过这就是科研的现状---无法预知今天的愚蠢,但更重要的则是要意识到当前的问题---暴露组学处在新研究的黎明期,即幸运也不幸。幸运的是大家起跑点都差不多;不幸的是只要跑,摔跟头几乎是必然的。

[王德奎 (Wang Dekui). 中国秘史---类比暴露组学和基因组学联系的研究,从历史智能到人工智能. *Academia Arena* 2019;11(5):50-52]. ISSN 1553-992X (print); ISSN 2158-771X (online). <http://www.sciencepub.net/academia>. 9. doi: [10.7537/marsaaj110519.09](https://doi.org/10.7537/marsaaj110519.09).

关键词: 暴露组学; 研究;健康; 研究; 真相

理论与基础

暴露组学与口传智慧论

读加拿大滑铁卢大学于淼博士 2018 年 11 月 12 日在科学网的博文《暴露组学的黎明》,其中类似的“主流物质数据库的覆盖情况”图,有 60 个从 1 到 708206 的自然数中选的大小不同的数字,填写在被六种颜色组成类似多角花瓣曲线交叉分割的区域内。

此图联系邱嘉文研究员 2018 年 2 月 14 日在科学网的博文《数学,微积分,概念关系》中,类似的“数学,微积分,概念关系境界”图---这不是五颜六色类似花瓣曲线图,而类似多线条交叉串联起的被五种颜色区分的椭圆形的 28 个气球。对这两者内涵的相似,类比口传智慧和文本智慧很有感触。于淼是 2011 年以来哈工大化工与化学学院的教授、博士生导师,他 2007 年在英国 Warwick 大学获理学博士学位,又先后在丹麦奥胡斯大学、美国哈佛大学、麻省理工学院从事博士后研究工作。而邱嘉文研究员是珠海诚开智能科技有限公司的副经理。他们两人从事的工作不同,但他们的科研分析的“心灵的境界”,却有相似的地方---《数学,微积分,概念关系》的文字很少,不妨全摘录如下:

“去年 9 月女儿考上了心仪的大学和心仪的专业,想到当她考上高中的时候,我曾送了她‘心灵的境界’说。这回,她感到高数学起来有些困难,于是我送了她这幅图。她表示:‘可以。’现在分享给其他家长的孩子”。把邱嘉文研究员归类分列画的

“数学,微积分,概念关系境界”图,联系于淼教授表达的“主流物质数据库的覆盖情况”图,是为能更好地阐明“暴露组学”比“基因组学”科研的多头性,以及其中复杂关系的联系,由此来理解口传智慧比文本智慧的深邃。在《暴露组学的黎明》一文中,于淼教授说的“暴露组学”很新鲜,由于很多人是第一次才听说,不妨摘录一些如下:

“暴露组学从名词提出,到现在大概 10 年有余。从 2017 年开始,进入高速发展阶段;今天的暴露组学,大致处于 20 年前基因组学的发展阶段。美国最早开始提暴露组学的是 NIH---加州伯克利、埃默里大学都是暴露组学起步比较早的地方。西奈山医学院 2017 年成立的美国第一家暴露组学研究所,是借助美国医学院间的网络,来推动暴露组学研究。2018 年 11 月在美国西奈山医学院召开的第二届暴露组学会议,就相当于 20 年前基因组学的发展阶段。也许接下来的 10 年内,它的研究经费与成果可能出现“井喷”。

暴露组学研究什么呢?它的基本问题跟基因组学差不多---一个人健康与否,基因组学认为更多依赖基因。伴随测序技术的进步,针对个人的测序,已经是可负担的了。但暴露组学认为,人的健康状况除了基因外,还要考虑表观遗传、蛋白组、代谢组与日常暴露,甚至还要考虑诸如地理位置、社会经济地位、肠道微生物组等的作用。

健康是目标,这里预测变量却非常多,很明显不是一个单因素模型。所有暴露组学属于面向问题

的高度综合性学科，基础包括不限于统计学、生命科学、数据科学、社会科学、环境科学、分析化学、毒理学、公共卫生、医学、遥感、传感、自动化、信息科学等诸多学科；目前并不知道哪个学科更重要。但很明显，任何一个学科都可能成为回答终极问题的短板，而且几乎每一个学科，都有短板且学科间交流壁垒不是一般的高。例如，从环境分析化学与数据科学这两个学科来说，当前如果要评价暴露水平，首先得知道有什么？

也就是目的性分析---但就暴露组学而言，并无法事先知道样品里有什么？所以更多研究是借鉴代谢组学的方法，利用高分辨质谱，来对未知物进行信息采集。这里信息采集的终点是色谱质谱峰，然而高分辨质谱全扫描的结果，往往混杂大量源内反应形成的加合物、碎片或物质本身的同位素峰；这导致虽然可以同时收集上万峰，但形成这些峰的化合物可能只有峰数的十分之一，且这些峰会共相关。

如果想讨论物质间的相关性，而使用了峰数据，那么估计会有偏。同时，峰识别的算法，也通常对全扫数据很不友好，会看到大量不应该被当作峰的数据被选成了峰，积分效果也是一塌糊涂。这一点，从分析化学角度是不可接受的。另一个问题是对未知峰的标注，现在流行的方法，是先跑全扫筛出差异峰，然后把那些峰去打二级质谱，有的则直接对差异峰去标注。这里使用一级质谱定性是风险很高的，下游的通路分析会因此不靠谱。而且就算找到一级质谱的匹配，也无法确认是否是同分异构体。而同分异构体的生物活性千差万别，更不用说当前主流数据库各搞各的，覆盖范围有局限性，唯一的标注也并不意味定性。二级质谱定性当前有很多软件可以做，但基本都是欠拟合状态，训练用的数据基本依赖可获取标准或社区用户共享，想做未知物十分困难---当前主流物质数据库的覆盖情况，如“主流物质数据库的覆盖情况”图中最大的三个物质库还没列，因为数据搞不到，或搞得到但处理起来太费劲。

目前能汇总整理这些信息的地方并不多，而且处理有些库的数据时，发现数据整理问题很大，格式不标准。如果不是专业人士，光是数据提取就得懵圈。另外，分析通量也是一个容易被忽略的问题。即使分析上的问题都解决了，下面的问题就是统计分析---用什么模型？为什么用这种模型？眼下都没法检验，也说不上哪个好哪个坏，其实都不怎么样。统计模型的复杂性可高可低，一般说高了，过拟合，而低了，欠拟合。不是说不能一次性尝试几百种统计模型或机器学习模型，关键如何解释？线性模型与层级模型是两种最有解释力的模型，但预测性谁能用谁知道？直接上神经网络不是不行，就是不好解释。精巧的统计模型面对错综复杂的数据，

难怪临床上喜欢多元线性回归。另一个相关问题是代谢物或暴露物有差异，环境研究可能没有分组，或者说分组后并无法进行效应预测。虽可以用效应诱导分析来做，但效应终点还是相对固定的。

此时预测多个毒性终点，不过如何把荷质比转成结构，也可说是一团乱麻。多个毒性终点也意味着不同的健康模型，有没有基于多个健康模型的宏模型呢？回答这个问题，只能依赖合作研究。跟健康相关研究还有个问题，就是无穷混杂因素。例如，有的知道年龄、性别、种族等；有的在建模时是忽略的，甚至根本意识不到可能是混杂因素。

暴露组学研究是点对多点做相关---健康研究的真相，是多对多互相影响，控制实验当然是必要的，但如果数据是来自观测研究，那这问题就几乎无解。受研究共同体的视野限制，如果只关心那些强信号，可能忽略了那些弱信号。但这里的强弱，是仪器决定的，不是生物学意义决定的。或许很多人的研究，可以讲一个故事，但很难回答一个真实的问题。这只是现存问题的很小一部分，每一点的进展都可能对上下游研究产生颠覆式影响。

对研究方法论的标准化、可重复化及与对基础研究进展的快速整合，是必要的。或许十年后回看今天的暴露组学，很多人可能惊叹：为什么大量的资源被浪费在了毫无意义的研究上？不过这就是科研的现状---无法预知今天的愚蠢，但更重要的则是要意识到当前的问题---暴露组学处在新研究的黎明期，即幸运也不幸。幸运的是大家起跑点都差不多；不幸的是只要跑，摔跟头几乎是必然的。

以上摘录《暴露组学的黎明》的文字很多，关键是把“暴露组学”对应“口传地方史”考证，把“基因组学”对应“书报公开史”考证，想说明“口传地方史”考证，比“书报公开史”考证，复杂和困难的问题很多。为啥要作“口传地方史”考证？正如暴露组学研究的基本问题跟基因组学差不多---是关系一个人健康与否？基因组学认为更多依赖基因，但暴露组学认为，人的健康状态除了基因外，还要考虑表观遗传、蛋白组、代谢组与日常暴露，甚至还要考虑诸如地理位置、社会经济地位、肠道微生物组等的作用。

Author:

王德奎 (Wang Dekui)
y-tx@163.com

参考文献 (References)

1. Baidu. <http://www.baidu.com>. 2019.
2. Google. <http://www.google.com>. 2019.
3. Journal of American Science. <http://www.jofamericanscience.org>. 2019.

4. Life Science Journal. <http://www.lifesciencesite.com>. 2019.
5. Marsland Press. <http://www.sciencepub.net>. 2019; <http://www.sciencepub.org>. 2019.
6. National Center for Biotechnology Information, U.S. National Library of Medicine. <http://www.ncbi.nlm.nih.gov/pubmed>. 2019.
7. Nature and Science. <http://www.sciencepub.net/nature>. 2019.
8. Stem Cell. <http://www.sciencepub.net/stem>. 2019.
9. Wikipedia. The free encyclopedia. <http://en.wikipedia.org>. 2019.

5/25/2019