

系统生物学中用于实验设计的交互式可视化工具和数据模型

摘要 - 在系统生物学的研究中实验设计十分重要，但往往得不到支持。为了改进实验设计，我们扩展了复杂鞘脂途径的可视化，以研究 SphinGOMAP 中的生物合成起源。我们使用 ganglio 系列鞘脂数据集作为测试平台，使用了 Java 通用网络/图形框架 (JUNG) 可视化工具包。我们的成果是一个交互式可视化工具和数据模型，用于脂质系统生物学的实验设计。通过允许 (i) 可选四种不同网络布局，(ii) 动态添加/删除屏幕上的分子和 (iii) 鼠标悬停以显示详细的分子数据，我们改进了当前 SphinGOMAP 在交互式可视化方面的性能。未来的工作将集中在系统地集成各种脂质相关数据，即 SphinGOMAP 生物合成数据、Lipid Bank 分子数据 (日本) 和 Lipid MAPS 代谢途径数据 (美国)。我们的目标是建立一个全面的交互式交流平台，以改良全球科学家在高通量脂质体系生物学研究中的实验设计。

关键词 - 实验设计，图形布局算法，交互式可视化，鞘脂途径

I. 引言

实验设计是系统生物学研究中一个重要的但常常得不到支持的方面。一方面，随着 DNA 微阵列和质谱等高通量技术不断推进，对高分辨率定量和时间数据的采集得以改进。另一方面，计算科学正在不断发展，产生了各种各样的生物信息学工具和算法，以便分析日渐增长的实验后数据量。在这个范例中，有限的资源有助于设计大规模的实验，这些实验涉及系统地采集持续时间内来自复杂生物系统的异质数据。

周密而明智的实验设计对于成功的系统生物学研究至关重要。由于技术和分析的进步，从研究相对独立的生物系统转向研究高通量系统生物学的势头正在不断增长。研究员能够在单变量实验的范例中清楚地人工阐述、设计实验，然而对于多变量实验，他们现在需要计算工具来进行可视化和处理。因此，一个实验设计的优劣，可以根据在分析实验后数据的过程中，它多大程度地减少计算负荷来评估。换句话说，健全完整的研究过程中，计算负载在整个实验设计、数据采集和实验后分析过程中保持平衡。

这种实验设计可以通过增加可用的先验知识的数量和质量来实现。随着社区标注和科学数据共享的力度不断加大，如果研究人员知道在哪里查找，那他们肯定不缺乏先验数据。但是，这种知识的质量并不能保证，即研究人员可能不知道如何看待。为了确保数量和质量，已经为各种生物学数据提出了多种标准。

为了提高先验知识的质量，我们提出了一个用于系统生物学实验设计的交互式可视化数据模型，使用 ganglio 系列鞘脂数据集的作为测试平台。通过实现 (i) 动态网络可视化，(ii) 可选四种不同网络布局 and (iii) 鼠标悬停以显示详细的分子数据，我们在交互式可视化方面扩展和改进了当前的 SphinGOMAP。这些功能提高了处理高容量、大规模系统生物学数据的用户体验。

II. 方法

基于来自 SphinGOMAP 的鞘脂分子数据和 JUNG 的程序化可视化库，我们设计和实现了交互式可视化工具。此外，我们扩展了我们的工具，来囊括与现有脂质资源（如 Lipid Bank 和 Lipid MAPS）有潜在相互作用的数据模型。

A. SphinGOMAP

SphinGOMAP 是用于组织和可视化鞘脂的途径图。SphinGOMAP 的明确目标是“促进关于神经鞘脂生物合成的'已知'和'未知'的对话并引导实验来完善这个模型”。 因此，SphinGOMAP 是随着新兴科学发现而发展的活跃文档。

当前的 SphinGOMAP 可以在交互和可视化方面进行改进。第一，SphinGOMAP 网站在 2007 年 10 月发布 2.0 版将静态文件中的大约 450 个化合物显示为微软 ppt 或 JPEG 图像。在目前的形式下，静态图像不允许用户与途径图自由交互。第二，各种鞘脂类家族（系列）必须分别在不同的文件中列出。这是因为鞘脂网络的密度和规模太大，无法在观察全景的同时又提供足够的细节。第三，SphinGOMAP 中的数据仅包含分子结构、类别、通用名称和 Lipid Bank ID。因此，路径图为了清晰度和一些细节牺牲了连通性。

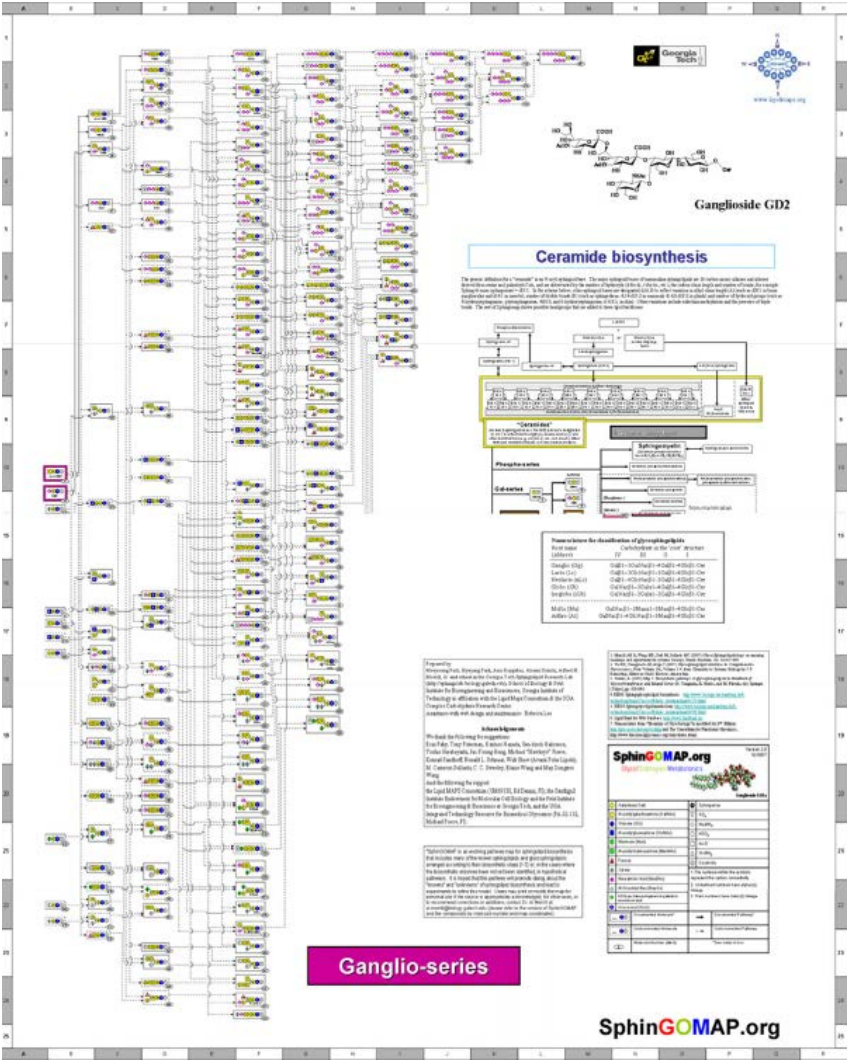


图 1. 目前 SphinGOMAP 中 ganglio 系列神经鞘脂的可视化结果。这可以作为微软 ppt 或静态 JPEG 图像文件提供，它交互性有限，并且为了清晰度和一些细节牺牲了连通性。

另一方面，在 SphinGOMAP 中部署了一个简单而直观的图例，用于表示鞘脂及其衍生物的生化官能团。该图例帮助用户一目了然识别不同的结构部分。这对于确定生物合成遗传中的递归模式特别有用。

使用数据模型和软件工具，我们通过交互、数据分离和有限的细节来解决这些问题。我们专注于 SphinGOMAP 的 ganglio 系列鞘脂作为我们的测试数据。该软件工具由开源可视化工具包实现。

B. JUNG - Java 通用网络/图形框架

用于网络/图形可视化的 JUNG API 提供通用的可扩展软件库，来对可以表示为图形或网络的数据进行分析和可视化。在这项工作中，我们整合了由 JUNG 软件库提供的 Kamada-Kawai、Fruchterman-Reingold、ISOMLayout 和 CircleLayout 算法，用 Java 图形 API 来渲染分子的结构。

分子的顶点是基于内部链结构生成的。这些链是在数据库过滤期间先验产生的。在数据库过滤中，我们根据公式提取分子结构，并根据链中分子的基数对它们进行重新排序。从较短的链开始，生成衍生物并将其映射到其父项。

分两步在内存中构造一棵树，来生成这种层次结构。第一步，我们选择单链分子来生成基本层次结构，第二步，将所有具有分支的分子分解成不同的链。例如，分子“Galb 1-3 GalNAcb 1-4(NeuAca2-3)Galb 1-4GlcCer”(b-beta) 分解成两条链，Galb 1-3 GalNAcb 1-4 Galb 14 GlcCer 和 NeuAca 2-3 Galb 1-4 GlcCer。我们在第一步里构建的基本层次结构中搜索这两个链，以提取父链。 请注意，每个分支分子有多个父亲。

一旦建立了这种关系，它将作为输入（哈希表结构）发送到可视化模块。可视化模块提取分子结构并作为图形在标签上渲染，该标签充当图结构的顶点。

C. Lipid Bank 和 Lipid MAPS

Lipid Bank 和 Lipid MAPS 是脂质数据的全球主要来源，分别来自日本和美国。一方面，在分子结构、科学通用的名称、光谱信息和文献方面，Lipid Bank 包含大量脂质（约 6000 个分子）的一手数据。另一方面，通过“表征脂质代谢物的全球变化（'脂质组学'）”，Lipid MAPS 侧重于更多的哺乳动物细胞内脂质相互作用方面的二手数据。Lipid MAPS 不仅包含生物活性脂质分子的结构数据和标注，还提供关于脂质代谢途径、实验方案、标准和时间过程结果的有关信息。Lipid MAPS 链接到相关分子如脂质相关蛋白的公共数据库。

III. 结果

图 2 给出了可视化的概述。在本节中，我们展示了一些值得注意的特性，例如 (a) 随着复杂度增加的动态网络可视化，(b) 可选不同网络布局 and (c) 鼠标悬停以显示详细分子数据。我们将 SphinGOMAP 中用于描述鞘脂分子上生物化学官能团的直观图例保留在可视化中。

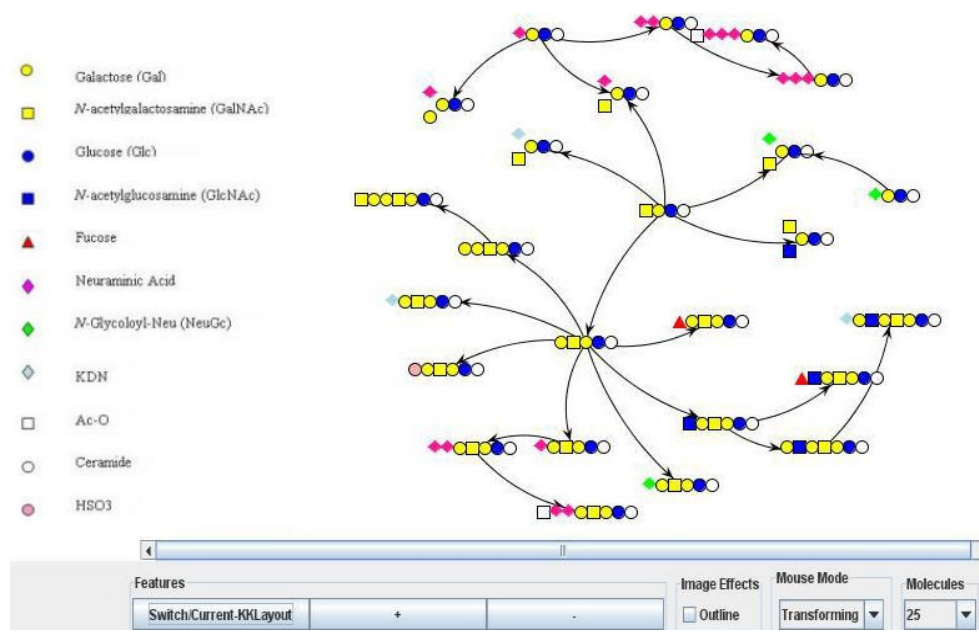


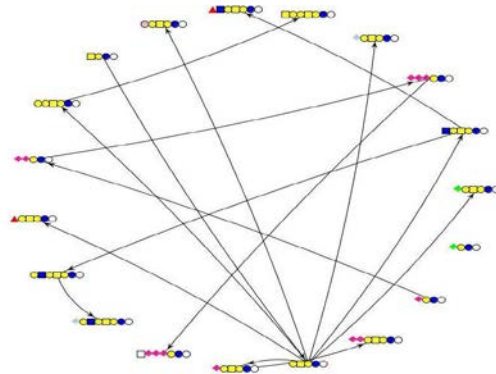
图 2. ganglio 系列鞘脂（选中 25 个分子）的交互式可视化的截屏。值得注意的功能包括随着网络复杂度的增加自发添加/删除分子、可选四种不同的网络布局，以及在鼠标悬停时列出详细的分子数据。

A. 随着复杂度增加的动态网络可视化

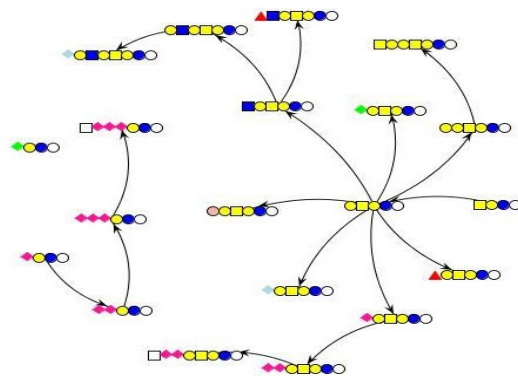
网络的复杂性是“实时”可视化的，因此用户可以自发地添加或删除可视化分子。为了做到这一点，所有的分子都会事先从数据库中过滤掉。我们推导出分子之间的父子关系，来作为内存中的层次结构，从最小的链开始，覆盖 SphinGOMAP 数据库中给定序列的所有分子。因此，分子可以以预定的等级顺序即有序序列自发地添加或删除。

B. 网络布局选择

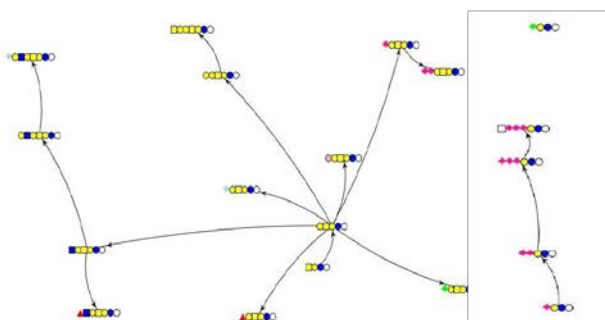
我们实现了四种不同网络布局的选择，如图 3 所示。在使用的四种布局算法中，Kamada-Kawai 算法是最易懂的，即使网络复杂度超过了 25 个分子。CircleLayout 适用于分子数量较少时，但在分子数量多时（例如 60 个分子）变得更复杂。



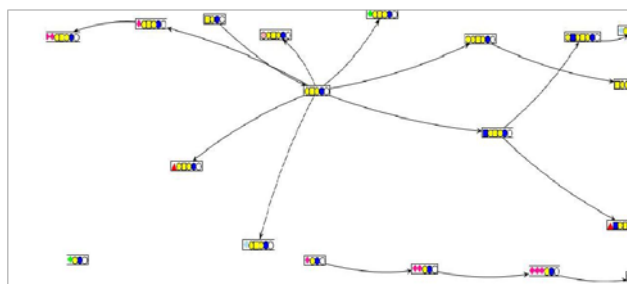
(a) Circle



(b) Kamad-Kawai



(c) ISOM



(d) Fruchterman-Reingold

图 3.25 个分子时的 Circle (a), Kamad-Kawai (b), ISOM (c) 和 Fruchterman-Reingold (d) 布局。不同的布局显示不同的网络特性，这些特性可能导致用户识别和进行适当的实验设计更改。

自组织图形布局算法 (ISOM) 使用尽可能多的空间来渲染图形，以便专门呈现每个关系。结果网络结构变得更清晰，但更加不可控。Fruchterman-Reingold (FR) 布局与 ISOM 布局类似，当屏幕上有 25 个分子，且最大边长小于 5 厘米时，效果不错。然而，在分子数量变多时，无法提供全景视图。因此，根据用户可能需要的网络分辨率级别，不同的布局可能会呈现不同的网络特性。

C. 鼠标悬停

图 4 所示的鼠标悬停功能允许用户查看详细的分子数据：通用名称、化学类别和 Lipid Bank ID。用这一功能将 Lipid Bank 和 Lipid MAPS 中更详细的鞘脂分子信息整合在一起，我们将 SphinGOMAP 的生物合成网络数据与一手数据和二手数据链接起来。

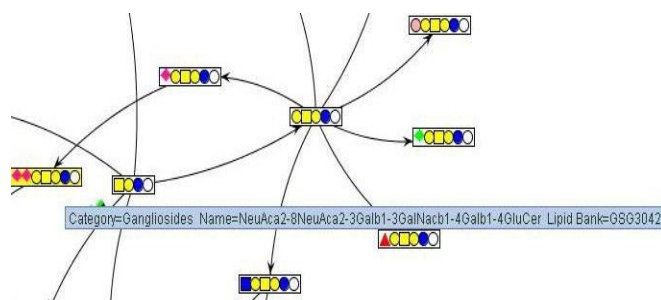


图 4. 鼠标悬停功能的放大图像。展示了详细的分子数据被;该功能将 Lipid Bank 和 Lipid MAPS 的数据合并链接, 以便与特定的鞘脂分子和途径进行更全面的交互。

IV. 讨论

采集大规模的生物网络数据是很重要的, 尤其是现已投入了难以估量的工作量和资源来进行实验和采集观测结果, 且还将继续投入。因此, 随着高通量技术带来的系统生物学越来越受关注, 我们必须合理地对数据组织给予同等重视。

为了便于系统生物学研究的实验设计, 我们在交互式可视化方面扩展并改进了当前的 SphinGOMAP。通过实现 (i) 动态网络可视化, (ii) 四种不同网络布局的选择以及 (iii) 鼠标悬停来显示详细的分子数据, 我们实现了这一点。这些特性提高了处理大批量、大规模系统生物学数据的用户体验。因此, 这项工作通过改善大规模复杂生物网络的可视化、交互性和可用性来促进系统生物学研究。

特别是对于实验设计, 典型用户这样可以使用我们的工具: (a) 基于动态网络可视化的特性, 追踪特定分子的生物合成起源, 或将其置于更广泛的下游衍生物或平行分子(化学遗传方面的“堂兄弟”)的环境中; (b) 用各种不同的网络布局来显示或确认先前未识别的生物合成关系, 以及 (c) 对有代谢数据知识的研究结果进行检查/并置, 这些结果通过与 Lipid MAPS 连接, 可以解释实验数据与先验网络模型之间的差异。

我们的实现体现了对以前的可视化在数据组织基于软件工具的呈现方面的具体改进。使用此工具, 我们旨在停止信息流的循环, 将先验知识和来自过去实验的社区反馈转化为更好的实验设计。未来的工作将集中在增加 SphinGOMAP、Lipid Bank 和 Lipid MAPS 之间的数据协作。通过连接生物合成起源数据、分子数据和代谢途径数据, 科学家可以期待与脂质分子信息的更多方面产生协同交互。

致谢

作者感谢 Christopher Haynes, Mitchell Parry Jeremy Allegood 和 Todd Stokes 提供技术援助和讨论; 感谢 Lipid MAPS 允许通过 SphinGOMAP 使用鞘脂生物合成数据。