生命的化学, 2023, 43(5): 648-656 www.life.ac.cn doi: 10.13488/j.smhx.20230130

综述

脊索动物嗅觉受体基因命名法的发展

韩 伟[1]， 赵素文[2,3*]， 黄行许[1*]， 林 峰[4*]

([1]之江实验室智能计算平台研究中心，杭州 311121; [2]上海科技大学iHuman研究所，上海 201210;
[3]上海科技大学生命科学与技术学院，上海 201210; [4]之江实验室人工智能研究院，杭州 311121)

摘要：嗅觉受体属于G蛋白偶联受体家族，在脊索动物的整个生命周期中都扮演着至关重要的角色。与其他多数基因家族不同，嗅觉受体家族是一个成员数量庞大的超基因家族，为它们合乎逻辑的命名可以更好地对该家族进行描述、分析和讨论，也可以为机器学习程序从庞大的嗅觉受体数据库自动构建相应的蛋白结构和功能知识库提供语义信息。由于脊索动物嗅觉受体演化速度很快、基因数量庞大、假基因比率高、在物种及染色体上分布差异巨大等多方面的原因，给嗅觉受体基因合理的命名较为困难。三十多年来，伴随着嗅觉受体研究领域的发展，嗅觉受体基因命名法也经历了多次迭代，在每个阶段都发挥着积极的作用。随着测序技术和生物信息学算法工具的发展，随之而来的是新注释的海量的嗅觉受体基因，这使已有的嗅觉受体基因命名法变得越来越难以适应大数据挖掘和知识工程的系统开发，因此迫切需要一个能满足当下需求的嗅觉受体基因命名法。

关键词：脊索动物；嗅觉受体；基因；命名法

Advances in nomenclature for chordate olfactory receptor genes

HAN Wei[1], ZHAO Suwen[2,3*], HUANG Xingxu[1*], LIN Feng[4*]

([1]Research Center for Intelligent Computing Platforms, Zhejiang Lab, Hangzhou 311121, China; [2]iHuman Institute, ShanghaiTech University, Shanghai 201210, China; [3]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China; [4]Research Institute of Artificial Intelligence, Zhejiang Lab, Hangzhou 311121, China)

Abstract: Olfactory receptors are members of the G protein-coupled receptor family, playing a crucial role throughout the entire lifespan of chordates. Distinguished from most gene families, the olfactory receptor family is a vast superfamily consisting of a large number of members. Providing logical names for these receptors enables better description, analysis, and discussion of the family. Additionally, it offers semantic information that assists machine learning programs in automatically constructing the corresponding protein structure and functional knowledge bases from extensive olfactory receptor databases. However, naming olfactory receptor genes appropriately poses significant challenges, given the rapid evolution, the large number of genes, a high incidence of pseudogenes, and substantial variations in their distribution across species and chromosomes in chordates. Over the past three decades, the field of olfactory receptor research has experienced significant development, resulting in several iterations of olfactory receptor gene nomenclature, all of which have played a positive role at its respective stage. The advancement of sequencing technologies and

收稿日期：2023-02-21
基金项目：国家自然科学基金项目(32122024)； 上海市生物大分子与精准医药前沿科学研究基地
第一作者：E-mail: hanwei2@zhejianglab.edu.cn
***通信作者：**林峰，E-mail: asflin@zhejianglab.com；黄行许，E-mail: huangxx@shanghaitech.edu.cn；赵素文，E-mail: zhaosw@shanghaitech.edu.cn

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

bioinformatics algorithms tools has brought about a massive influx of newly annotated olfactory receptor genes. This overwhelming number has rendered the existing olfactory receptor gene nomenclature increasingly inadequate to accommodate the needs of big data mining and knowledge engineering system development. Consequently, there is an urgent need for a new olfactory receptor gene nomenclature that can satisfy current demands.

Key Words: chordates; olfactory receptors; gene; nomenclature

嗅觉是脊索动物最重要的感觉之一，在寻找食物、交配、躲避危险、识别个体、标记领域等方面起着关键作用[1-3]。嗅觉的实现主要依赖嗅觉受体基因(olfactory receptor gene，*Olfr*)的表达。对于脊索动物来说，大部分的嗅觉受体基因在嗅觉组织或者器官内表达出嗅觉受体(olfactory receptors，ORs)，用于识别环境中的气味分子；也有部分嗅觉受体基因在多种非嗅觉组织或器官中表达，即异位表达，并被证明与多种疾病密切相关[4-6]。这些证据意味着嗅觉受体不仅可以作为嗅觉功能的执行者，还可以作为潜在的药物靶点，具有重要的研究价值。

嗅觉受体属于G蛋白偶联受体家族，具有七次跨膜螺旋结构域[7]，作为脊索动物中的超基因家族它的数量及分布极具特点。第一个特点是基因数量众多。大多数的脊索动物具有1 000多个嗅觉受体基因(包含功能基因与假基因)，占据了G蛋白偶联受体家族中相当一部分比例。例如，人类的功能性嗅觉受体基因的数量占G蛋白偶联受体家族成员的约50%[8]。第二个特点是嗅觉受体基因在不同物种基因组中的数量差异十分明显。在脊索动物中，嗅觉受体基因数量从数个到数千个不等，反映了物种对不同生存环境的适应[9-13]。第三个特点是嗅觉受体基因在染色体上广泛且不均匀的分布，数量众多的嗅觉受体基因成簇地分布在物种的大多数染色体上。以人的嗅觉受体基因在染色体上的分布为例，近千条嗅觉受体基因分布在除了20号和Y染色体外的其他22条染色体上，并且他们在每条染色体上分布的数量并不均匀，其中11号染色体上分布了大多数的嗅觉受体基因，而10、12和X号染色体上都仅存在一条嗅觉受体基因[14-17]。第四个特点是脊索动物嗅觉受体假基因比例很高。就目前Han等[9]注释的近1 700个脊索动物的约80万个嗅觉受体基因来看，功能性嗅觉受体与嗅觉受体假基因的比例大约是10∶9。综上，脊索动物嗅觉受体基因具有数量庞大、在物种及染色体上分布差异大、假基因比例高等特点，而这些恰恰导致了嗅觉受体基因的命名等基础问题难以解决。

此前，该领域的研究者曾提出过多种嗅觉受体基因的命名方案[18-26]，但是，目前这些方案或多或少存在一些问题。一方面，这些命名法存在基因标识符混乱，甚至是嗅觉受体亚家族和家族错误分配等缺陷，这将给生物实验科学家带来嗅觉受体基因功能认识上的困惑。在目前最流行的嗅觉受体基因命名法的规则下，部分嗅觉受体基因被明确分成了OR5、OR8和OR9三个基因家族。但是，有研究表明，这三个嗅觉受体基因家族的功能是一致的[27-30]。Han等[9]对这三个家族的序列进行分析，发现来自这三个家族的序列之间的相似性完全满足同一个嗅觉受体家族的标准，即他们应该被划分到同一个嗅觉受体家族。更多类似的例子还有OR1/3/7、OR2/13、OR41/42、OR51/52和OR61/62等[9,27-30]。另一方面，在已经完成全基因组测序的脊索动物中，有超过80%物种的全部嗅觉受体基因没有被注释和命名，这极大限制了后续功能和内源性配体鉴定等科学研究的开展[9]。随着生物信息学算法工具的发展，嗅觉受体基因的发现越来越依赖一个合乎计算生物学逻辑的命名规范，以使机器学习程序能更有效地从已有的嗅觉受体基因数据库中挖掘海量的潜在信息，自动构建相应的蛋白结构和功能知识库，并有助于实验生物学家获得更多嗅觉受体功能分类上的信息。

1　嗅觉受体基因命名法的发展

嗅觉受体无论是作为气味分子的探测器还是作为潜在的药物靶点都具有重要的研究价值，但是它在相当长的时间内都没能真正意义上步入主

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

流科学的殿堂。随着1991年Linda Buck和Richard Axel发现嗅觉受体，嗅觉受体几乎在一夜之间被推向主流神经科学研究领域[7]，他们也凭此杰出的成就获得了2004年的诺贝尔生理学或医学奖。自此，人们对嗅觉受体的研究热情逐渐高涨。但是，嗅觉受体基因命名在相当长的一段时间内都处于相对混乱的状态。

1.1 早期的嗅觉受体基因命名法

截至20世纪末，研究者们已经从多个物种中注释出了数百个嗅觉受体基因。此时的研究者对于嗅觉受体基因的命名法尚未达成共识，分别尝试基于克隆名称(如*HGMP07E*[18])、克隆方法或环境(如*HPFH1OR*[19])、染色体位置(如*OR17-23*[20])、全基因组范围的顺序编号和随机编号(如*OLFR1*[18]、*ORL3001*[21]和*ZF2A*[22])、随意指定(如gen5 [23])等方式为少数的嗅觉受体基因命名。对于嗅觉受体基因命名来说，这个阶段是百家争鸣的时期，尚未出现被研究者广泛认可的命名法。简单来说，这些命名法只是单纯地为嗅觉受体基因分配标识符，并没有明确的命名逻辑。在嗅觉受体基因数量较少的研究初期，这些命名法尚可满足需求。但是，随着测序技术的飞速发展，获得物种嗅觉受体基因数据越来越容易，此时混乱的命名法为大量嗅觉受体基因命名变得十分困难[31]。然而，大量嗅觉受体基因只有被有意义的、合乎逻辑的命名才能被更好地描述和讨论。此时，研究者已经认识到人们迫切需要一个新的嗅觉受体基因命名法[32]。

1.2 具有明确逻辑的嗅觉受体基因命名法

2000年，研究者着手构建具有内在逻辑的嗅觉受体基因命名法，其中Ziegler等[24]和Glusman等[25]提供的方案最具代表性。

Ziegler等[24]提供的命名法整合了嗅觉受体基因的物种、染色体编号、受体类型、等位基因、功能基因及假基因等相关信息。以hs6M1-6为例，其中前缀字母“hs”代表该基因所属的物种信息，在本例中是物种*Homo sapiens*的缩写；前缀字母后的数字“6”代表该基因所在染色体的编号，在本例中是6号染色体；之后的字母和数字组合“M1”代表了嗅觉受体基因的类型和亚型，在本例中是主嗅觉上皮嗅觉受体基因的亚型1；连接符后的数字“6”代表该基因一个任意但唯一的识别号。此外，该命名法还提供拓展信息。如果已知该嗅觉受体基因的等位基因，则用星号和额外的数字描述相应的等位基因，如hs6M1-6*01中的“01”代表了一个独特的等位基因识别号。如果已知一个嗅觉受体基因是假基因，那么会在该基因标识符后添加一个后缀“P”，如hs6M1-4P。相较于之前的嗅觉受体基因命名法，Ziegler等[24]提出的命名法通过特定的内在逻辑整合了嗅觉受体基因多方面的信息，有利于研究者们分享知识。但是，这种简单的信息整合方式忽略了嗅觉受体序列内部的关系，对于研究嗅觉受体演化、家族关系以及区分受体间关系并没有实质性的贡献。该命名法中每个物种嗅觉受体基因都需要研究者手动地为其命名，不适合为不断涌现的新的嗅觉受体基因命名[33-37]。此外，在新型研究范式下，该命名法所能够提供的符合计算生物学逻辑的语义信息十分有限。

Glusman等[25]提出了一种基于趋异演化模型(divergence evolutionary model)的嗅觉受体超家族命名法。他们全面地搜索并整理了当时数据库中的嗅觉受体基因序列，获得了25个物种的780条嗅觉受体基因序列，然后基于这些序列构建系统发生树，并依据被广泛接受的平均距离AD(N)阈值(公式1)，从系统发生树上划分出嗅觉受体的家族与亚家族[25,38,39]。具体来说，在系统发生树中平均距离AD(N)0.4的节点代表的嗅觉受体被划分为同一个家族，平均距离AD(N)0.6的节点代表的嗅觉受体被划分为同一个亚家族。基于嗅觉受体家族和亚家族信息并借鉴前人的研究，Glusman等[25]提出了新的嗅觉受体基因命名法[40-43]。以OR5AN1为例，其中前缀字母“OR”是Olfactory receptor的缩写，作为该基因标识符的根符号；前缀字母后的数字“5”代表该基因所属的嗅觉受体家族编号；之后的字母“AN”代表该基因所属的嗅觉受体亚家族；最后的数字“1”代表该基因一个任意但唯一的识别号。与Ziegler等[24]提出的命名法相似，如果已知一个嗅觉受体基因是假基因，那么会在该基因标识符后添加一个后缀“P”，如OR3A5P。Glusman等[25]提出的命名法是首次在嗅觉受体研究领域内使用亚家族和家族命名逻辑，通过基因标

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

识符可以清晰地分析嗅觉受体家族和亚家族关系。这种基于亚家族和家族的命名逻辑也被广泛地用于其他基因家族的命名法，因此，这种命名法更容易被研究者接受[40,41,43-45]。经过多年的发展，Glusman等[25]提出的命名法为人类嗅觉受体基因分配的标识符已经被广泛地使用，并被人类基因命名委员会(HUGO Gene Nomenclature Committee)采纳[46,47]。该命名法还以人类的嗅觉受体基因标识符为基础，拓展到其他几个物种，但是在此之后的多年间，除了人类的嗅觉受体基因标识符被广泛接受外，其他哺乳动物仍然在使用多种不同的命名法[48]。

$$\mathrm{AD(N)}=\sum_{i,j}\frac{100-\mathrm{PID}\left(a_i,b_j\right)}{100};\forall a_i\in A,\forall b_j\in B$$

公式1

其中A和B分别是系统发生树中节点N连接的两个子树；a_i是子树A中的任意一个节点，代表一条受体序列；b_j是子树B中的任意一个节点，代表另一条受体序列；$\mathrm{PID}\left(a_i,b_j\right)$是$a_i$和$b_j$之间的序列同一性(identity)。

Glusman等[25]提出了基于亚家族和家族的嗅觉受体基因命名法之后，这种命名逻辑迅速地被其他研究者接受并发展。Zozulya等[14]在Glusman等[25]命名法的基础上增加了物种和染色体编号等信息，提出了新的嗅觉受体基因命名法，如hOR22.01.01 (hOR11H1)。这种命名法似乎是Ziegler等[24]和Glusman等[25]两种命名法的嵌合体，也许是该命名法的规则太过于复杂，之后并没有被研究者广泛使用。Zhang等[48]借鉴了Glusman等[25]的命名逻辑，并将其应用于小鼠嗅觉受体基因的命名，如MOR1-1。

1.3 趋于统一的嗅觉受体基因命名法

2020年，Olender等[26]在他们之前命名法的基础上提出了脊椎动物嗅觉受体基因的统一命名法。该方法继承并发展了之前的命名逻辑，以人的嗅觉受体基因序列为核心，通过检测其他物种与人嗅觉受体基因序列之间的相似性来分配基因标识符，提出了相互最大相似度(mutual maximum similarity，MMS)算法用于嗅觉受体基因的命名(图1)。该算法的具体流程如下：首先，对于一组给定的嗅觉受体蛋白序列，使用BLASTP程序搜索人类所有的嗅觉受体蛋白序列所构建的数据库，从而获得了all-versus-all的同一性矩阵；然后，依次执行图中左侧黄色区域中编号为1-5的五个步骤；(1)识别同一性矩阵中值大于或等于80%的最佳匹配，并分配与人类相同的基因符号；(2)识别同一性矩阵中值大于或等于80%的次优候选序列，分配人类最佳匹配的基因符号并添加字母B、C等后缀，但是字母P除外，因为它是为假基因保留的特殊符号；(3)在几个非人类的模式动物中，顺序地执行1-3步骤；(4)对于未分配基因符号的嗅觉受体基因，如果与任意一条已经分配好基因标识符的嗅觉受体序列的同一性大于或等于60%而小于80%，

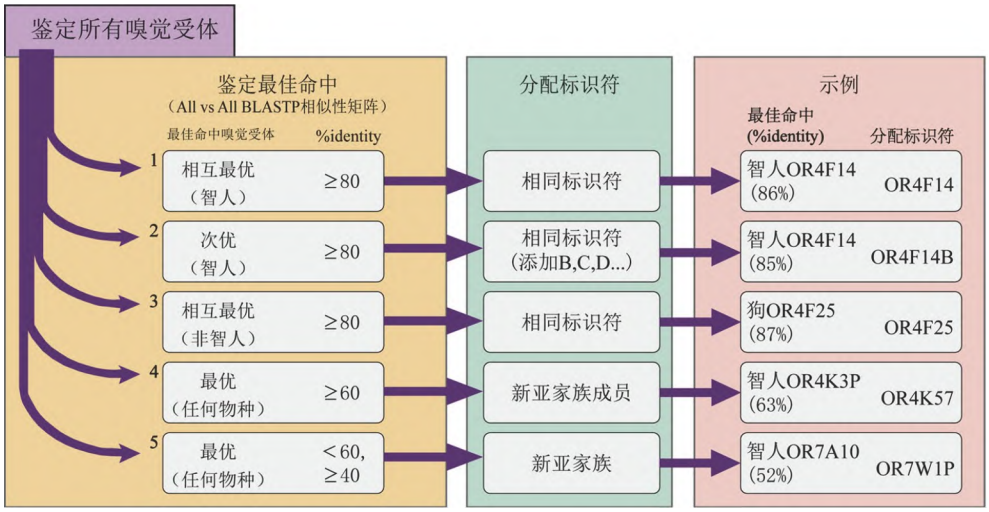
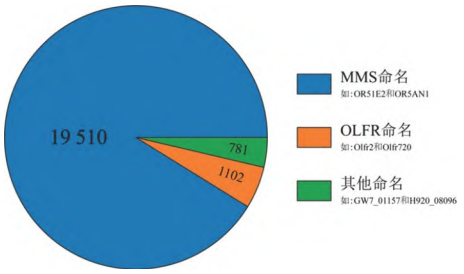


图1　使用相互最大相似度算法进行嗅觉受体基因命名的算法流程(改编自Olender等[26])

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

则将其分配到该亚家族并赋予新的基因标识符；(5)如果与任意一条已经分配好基因标识符的嗅觉受体的同一性大于或等于40%而小于60%。则将其分配到该家族并赋予新的亚家族和基因标识符。简单来说，该算法是一种系统的命名法，用于根据物种间分层成对的序列相似性，以人类的嗅觉受体基因标识符为中心，将嗅觉受体基因标识符分配给任意嗅觉受体基因。

截至目前，Doron等[26]提出的相互最大相似度算法已经成为目前最流行的嗅觉受体基因命名法，已经为大鼠、黑猩猩和斑马鱼等八种脊椎动物的10 249个嗅觉受体基因分配了基因标识符，并将这些标识符收录在HORDE数据库中[26,46]。除了人的嗅觉受体基因命名方案被HUGO基因命名委员会采纳外[46,47]，其他的特定物种基因命名委员会正在考虑采纳Olender等[26]的命名方案，如小鼠的MGNC[49]、大鼠的RGNC[50]、斑马鱼的ZNC[51]，还有为黑猩猩、牛、狗和马等更多脊椎动物基因命名的VGNC[52]等。此外，还有一些物种的嗅觉受体基因标识符是通过Gene Ontology(GO)自动分配获得的[53]，如UniProt数据库中记录了两万多个嗅觉受体基因标识符[54](图2)。



MMS命名指的是相互最大相似度算法分配的嗅觉受体基因标识符，OLFR命名指的是一种全基因组范围编号的嗅觉受体基因命名法

图2 UniProt数据库的嗅觉受体基因标识符的统计

2 主流嗅觉受体基因命名法存在诸多漏洞和限制

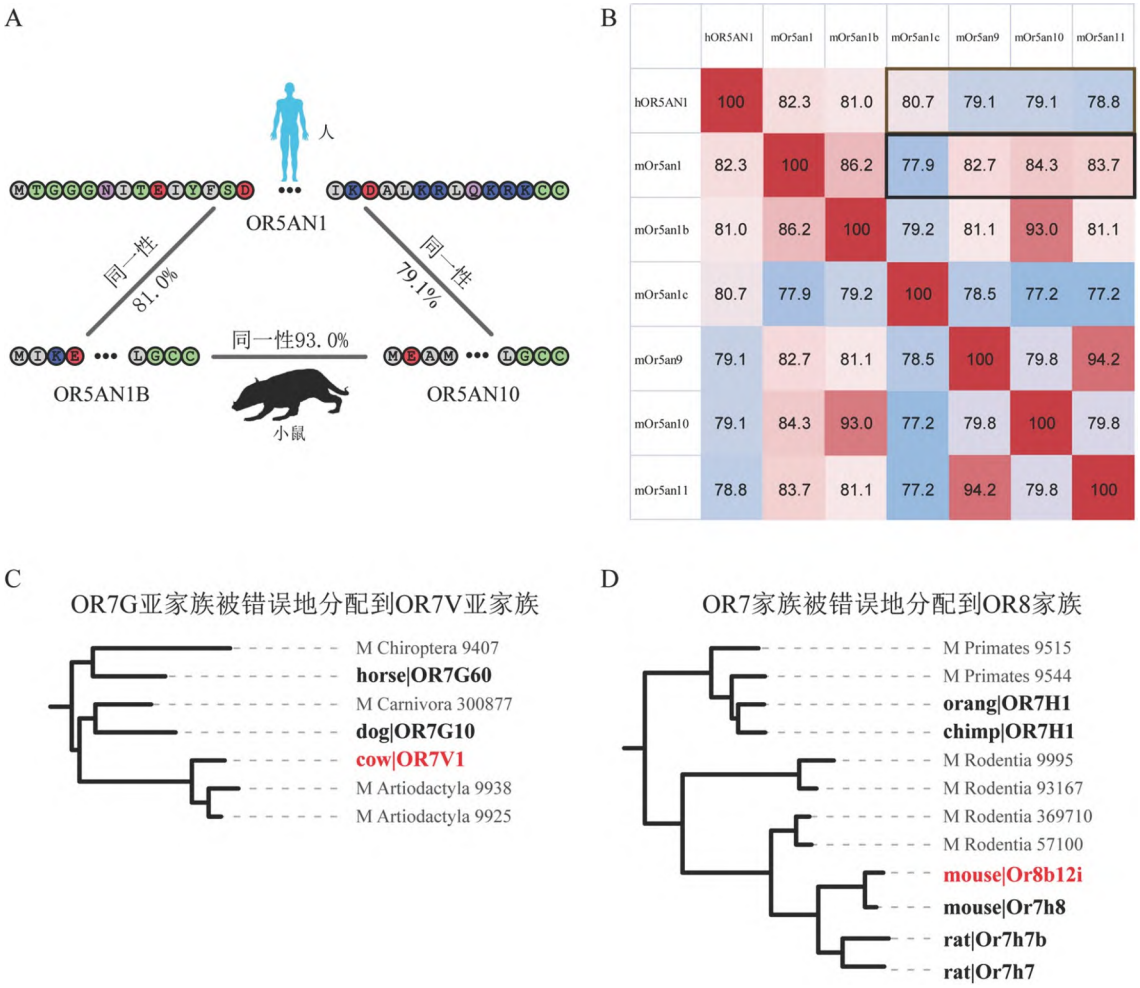
虽然，Olender等[26]提出的新方法已经成为目前主流的嗅觉受体基因命名法，但是笔者发现这种命名法存在以下四个缺陷。第一，对于非人类的物种来说，该算法只关注物种间序列的相似性关系，而忽略物种内部的序列相似性关系导致基因标识符的错配，如小鼠中两条同一性高达93.0%的嗅觉受体序列，其中一条序列与人的OR5AN1相似性未达到80%阈值要求而被分配了不同的标识符OR5AN10，另一条则采用与人类相同的标识符(图3A)。第二，选取不同的嗅觉受体基因作为参照将会得到不同的命名结果，从而导致基因标识符的冲突；以图3B为例，如果以mOr5an1(第二行)为参照而不是hOR5AN1(第一行)，小鼠的mOR5an9、mOr5an10和mOr5an11将会被分配为mOr5an1基因的不同子类型，而mOr5an1c将会分配以其他的基因标识符。第三，该算法采用局部最优的处理方式造成了嗅觉受体基因亚家族甚至是家族的错误分配(如图3C中原本属于OR7G亚家族的嗅觉受体基因被错误地分配了OR7V1标识符，图3D中原本属于OR7家族的嗅觉受体基因被错误地分配了OR8B12I标识符)。第四，该算法拓展性欠佳，很难快速且准确地为大量的嗅觉受体基因分配标识符。事实上，Olender等[26]也仅为包括人在内的九个物种的嗅觉受体基因分配了完整的标识符。截至到2022年，Han等[9]已经从近1 700个脊索动物中注释得到了约80万个嗅觉受体基因，如果以Olender等[26]的方法进行命名，将需要巨大的工作量并且基因标识符之间会随着已命名物种数量的增加而产生越来越多的冲突[17]。总的来说，Olender等[26]的命名新方法最核心的矛盾在于以人类的嗅觉受体基因标识符为中心为其他物种分配标识符，只依据局部最相似的关系进行标识符分配而忽略了全局的相似性关系。该方法造成的后果十分严重，它会对跨物种嗅觉受体基因搜寻和标识算法实际应用等造成较大偏差，也会在基于机器学习的嗅觉受体蛋白序列、结构和功能预测中引入错误的归纳偏置。

3 讨论与展望

3.1 新嗅觉受体基因命名法应有的特征

嗅觉受体具有重要的研究意义。脊索动物嗅觉受体基因的数量巨大、假基因比率高、在物种及染色体上分布差异大等多方面的原因，使嗅觉受体基因的命名等基础问题在最初的十年间都处于混乱的状态。这段时间内研究者对嗅觉受体基因的命名策略尚未达成共识，他们分别尝试以克隆名称、克隆方法或环境、染色体位置、全基因

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net



A：小鼠嗅觉受体基因物种内部不进行序列相似性比较而导致的矛盾；B：小鼠嗅觉受体OR5AN亚家族的6个成员与人类OR5AN1蛋白序列的同一性矩阵；表头中hOR5AN1是human OR5AN1的缩写，mOr5an1是mouse Or5an1的缩写，其他缩写方式类似；C、D：嗅觉受体基因的系统发生树(部分)

图3　相互最大相似度算法分配嗅觉受体基因标识符的几个缺陷实例

组范围的顺序编号和随机编号，甚至是随意指定等多种方式完成对嗅觉受体基因的命名。到20世纪末，随着测序技术的飞速发展，随之而来的是来自不同物种的海量嗅觉受体基因，研究者们认识到他们迫切需要一个具有明确逻辑的命名法，以便更好地描述和讨论嗅觉受体基因。

在前人的基础上，Ziegler等[24]和Glusman等[25]分别提出了具有明确命名逻辑的嗅觉受体基因命名法。Ziegler等[24]提出的方法将物种、染色体编号、受体类型、等位基因、功能基因及假基因等多方面信息整合到基因标识符内，有利于研究者们之间分享信息。Glusman等[25]提出了一种基于分歧演化模型的嗅觉受体基因命名法，通过基因的演化关系划分出家族和亚家族。相较于Ziegler等[24]提出的嗅觉受体基因命名法，Glusman等[25]提出的方法不是简单地将多方面的信息拼接起来，而是基于基因的演化信息，因此在研究嗅觉受体基因的动态演化、物种间嗅觉受体基因的直系同源关系、物种内的旁系同源关系和序列分类等问题上具有优势。经过近二十年的发展，Olender等[26]在之前的经验上进一步提出相互最大相似度算法，旨在构建一个统一的嗅觉受体基因命名系统，为所有脊椎动物的嗅觉受体基因分配标识符。由于以人类嗅觉受体基因标识符为中心为其他物种的嗅觉受体基因分配标识符而忽略了其他物种之间和内部的序列关系，导致该方法存在基因标识符错配、冲突、嗅觉受体基因亚家族和家族的错误分配以及拓展性欠佳等多方面的缺陷。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

当前流行的嗅觉受体基因命名法，在为大量新注释出的嗅觉受体基因合理地命名时变得越来越困难，因此迫切需要一个新的嗅觉受体基因命名法。与之前的命名法相比，新方法应具有以下几方面的特点才能应对当下的挑战。第一，新方法应基于全局序列相似性进行聚类，从而避免因局部最优而导致的嗅觉受体基因名称混乱与冲突等缺陷；第二，新方法应具有良好的分配效率和可拓展性。

3.2 机器学习对嗅觉受体基因命名法的要求

鉴于脊索动物嗅觉受体基因数量多、差异大、假基因比率高等特点，大规模的生物序列建模和蛋白质结构及功能预测非常适合使用机器学习的方法。其中，自动聚类学习算法就可以建立在反映高维度参数特征的嗅觉受体基因名称上。机器学习对嗅觉受体基因命名法的要求包括满足聚类全局序列相似性、命名分配的空间匀称性和命名规则的可拓展性。

聚类算法的本质是无监督机器学习，在语言建模和机器翻译中得到广泛使用，所采用的主流框架为Encoder-Decoder，如近年来引起广泛关注的ChatGPT和它的核心技术Transformer。在Transformer网络框架中，从编码器输入的序列首先会经过一个自注意力层，该层帮助编码器在对每个序列单元编码时关注输入序列的其他单元。自注意力层的输出会传递到前馈神经网络中，每个位置的单元对应的前馈神经网络都一样。解码器中也有编码器的自注意力层和前馈层。除此之外，这两个层之间还有一个注意力层，用来关注输入序列的相关部分。从生物信息学算法的角度看，嗅觉受体的功能发现就是设计一个合乎计算生物学逻辑的Transformer网络框架，输入一个嗅觉受体基因序列，输出一个嗅觉受体基因名称。一个反映高维度参数特征的嗅觉受体基因命名法可以使机器学习程序有效地实现聚类；即从已有的脊索动物嗅觉受体基因序列数据库自动构建相应的蛋白质结构和功能知识库。

3.3 新命名法需要为实验生物学家提供便利

对嗅觉受体家族进行合理命名是为了服务于功能研究。因此，新命名法需要严格区分嗅觉受体基因的各个类别，赋予含有内在命名逻辑的标识符，并为功能验证研究带来如下便利。(1)严格、清晰、明确的家族和亚家族关系将会使嗅觉受体基因的功能注释具有更高的质量。前人提出的命名法存在给相同基因分配不同的标识符、给不同的基因分配相同的标识符等问题，这对于功能注释来说是一场灾难，往往使实验生物学家对嗅觉受体基因间功能上的联系感到困惑。新的命名法需要从整个嗅觉受体家族的全局出发，给相同的嗅觉受体基因分配统一的标识符，这会很大程度上改善功能注释的质量以帮助实验生物学家理解诸多嗅觉受体基因的功能。(2)结合隐马尔可夫模型等机器学习方法将会提高嗅觉受体基因功能注释的速度。新方法需要为具有相同标识符的基因构建隐马尔可夫模型，并将之用于注释新发现的嗅觉受体基因以提高注释速度。因此，新命名法需要明确嗅觉受体基因和标识符之间的对应关系并为每一个嗅觉受体基因类别都构建好隐马尔可夫模型。基于这些构建好的隐马尔可夫模型，可以快速地为新发现的嗅觉受体基因做好家族和亚家族的归属，使生物实验科学家在开始正式研究之前便对该新基因潜在功能有一个整体上的认识。

参 考 文 献

[1] Zhou C, Liu Y, Zheng X, et al. Characterization of olfactory receptor repertoires provides insights into the high-altitude adaptation of the yak based on the chromosome-level genome. Int J Biol Macromolecules, 2022, 209: 220-230

[2] Liu H, Chen C, Lv M, et al. A chromosome-level assembly of blunt snout bream (*Megalobrama amblycephala*) genome reveals an expansion of olfactory receptor genes in freshwater fish. Mol Biol Evol, 2021, 38(10): 4238-4251

[3] Wang K, Tian S, Galindo-González J, et al. Molecular adaptation and convergent evolution of frugivory in Old World and neotropical fruit bats. Mol Ecol, 2020, 29(22): 4366-4381

[4] Cheng J, Yang Z, Ge XY, et al. Autonomous sensing of the insulin peptide by an olfactory G protein-coupled receptor modulates glucose metabolism. Cell Metab, 2022, 34(2): 240-255

[5] Seo J, Choi S, Kim H, et al. Association between olfactory receptors and skin physiology. Ann Dermatol, 2022, 34(2):

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

87-94

[6] Orecchioni M, Kobiyama K, Winkels H, et al. Olfactory receptor 2 in vascular macrophages drives atherosclerosis by NLRP3-dependent IL-1 production. Science, 2022, 375(6577): 214-221

[7] Buck L, Axel R. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell, 1991, 65(1): 175-187

[8] Vadevoo SMP, Gunassekaran GR, Lee CE, et al. The macrophage odorant receptor Olfr78 mediates the lactate-induced M2 phenotype of tumor-associated macrophages. Proc Natl Acad Sci USA, 2021, 118(37): e2102434118

[9] Han W, Wu Y, Zeng L, et al. Building the chordata olfactory receptor database using more than 400,000 receptors annotated by genome2OR. Sci China Life Sci, 2022, 65(12): 2539-2551

[10] Hasin-Brumshtein Y, Lancet D, Olender T. Human olfaction: from genomic variation to phenotypic diversity. Trends Genet, 2009, 25(4): 178-184

[11] Hoover KC. Evolution of olfactory receptors. Methods Mol Biol, 2013, 1003: 241-249

[12] Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet, 2008, 9(12): 951-963

[13] Bear DM, Lassance JM, Hoekstra HE, et al. The evolving neural and genetic architecture of vertebrate olfaction. Curr Biol, 2016, 26(20): R1039-R1049

[14] Zozulya S, Echeverri F, Nguyen T. The human olfactory receptor repertoire. Genome Biol, 2001, 2(6): RESEARCH0018

[15] Olender T, Feldmesser E, Atarot T, et al. The olfactory receptor universe-from whole genome analysis to structure and evolution. Genet Mol Res, 2004, 3(4): 545-553

[16] Go Y, Niimura Y. Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. Mol Biol Evol, 2008, 25(9): 1897-1907

[17] Glusman G, Yanai I, Rubin I, et al. The complete human olfactory subgenome. Genome Res, 2001, 11(5): 685-702

[18] Schurmans S, Muscatelli F, Miot F, et al. The OLFR1 gene encoding the HGMP07E putative olfactory receptor maps to the 17p13p12 region of the human genome and reveals an MspI restriction fragment length polymorphism. Cytogenet Genome Res, 1993, 63(3): 200-204

[19] Feingold EA, Penny LA, Nienhuis AW, et al. An olfactory receptor gene is located in the extended human β-globin gene cluster and is expressed in erythroid cells. Genomics, 1999, 61(1): 15-23

[20] Ben-Arie N, Lancet D, Taylor C, et al. Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. Hum Mol Genet, 1994, 3(2): 229-235

[21] Healy MD, Smith JE, Singer MS, et al. Olfactory receptor database (ORDB): a resource for sharing and analyzing published and unpublished data. Chem Senses, 1997, 22(3): 321-326

[22] Mori T, Sakai M, Matsuoka I, et al. Analysis of promoter activity of 5′-upstream regions of zebrafish olfactory receptor genes. Biol Pharm Bull, 2000, 23(2): 165-173

[23] Mezler M, Konzelmann S, Freitag J, et al. Expression of olfactory receptors during development in *Xenopus laevis*. J Exp Biol, 1999, 202(4): 365-376

[24] Ziegler A, et al. Polymorphic olfactory receptor genes and HLA loci constitute extended haplotypes. In: Kasahara, M. (eds) Major Histocompatibility Complex. Springer, Tokyo, 2000

[25] Glusman G, Bahar A, Sharon D, et al. The olfactory receptor gene superfamily: data mining, classification, and nomenclature. Mamm Genome, 2000, 11(11): 1016-1023

[26] Olender T, Jones TEM, Bruford E, et al. A unified nomenclature for vertebrate olfactory receptors. BMC Evol Biol, 2020, 20(1): 42

[27] Hughes GM, Teeling EC, Higgins DG. Loss of olfactory receptor function in hominin evolution. PLoS One, 2014, 9(1): e84714

[28] Hughes GM, Boston ESM, Finarelli JA, et al. The birth and death of olfactory receptor gene families in mammalian niche adaptation. Mol Biol Evol, 2018, 35(6): 1390-1406

[29] Khan I, Yang Z, Maldonado E, et al. Olfactory receptor subgenomes linked with broad ecological adaptations in sauropsida. Mol Biol Evol, 2015, 32(11): 2832-2843

[30] Yohe LR, Fabbri M, Lee D, et al. Ecological constraints on highly evolvable olfactory receptor genes and morphology in neotropical bats. Evolution, 2022, 76(10): 2347-2360

[31] Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods, 2008, 5(1): 16-18

[32] Mombaerts P. Molecular biology of odorant receptors in vertebrates. Annu Rev Neurosci, 1999, 22(1): 487-509

[33] Alioto TS, Ngai J. The odorant receptor repertoire of teleost fish. BMC Genomics, 2005, 6(1): 173

[34] Niimura Y, Nei M. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc Natl Acad Sci USA, 2005, 102(17): 6039-6044

[35] Dugas JC, Ngai J. Analysis and characterization of an odorant receptor gene cluster in the zebrafish genome. Genomics, 2001, 71(1): 53-65

[36] Kondo R, Kaneko S, Sun H, et al. Diversification of olfactory receptor genes in the Japanese medaka fish, Oryzias latipes. Gene, 2002, 282(1-2): 113-120

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[37] Niimura Y, Matsui A, Touhara K. Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals. Genome Res, 2014, 24(9): 1485-1496

[38] Nebert DW, Nelson DR, Coon MJ, et al. The P450 superfamily: update on new sequences, gene mapping, and recommended nomenclature. DNA Cell Biol, 1991, 10(1): 1-14

[39] Lancet D, Ben-Arie N. Olfactory receptors. Curr Biol, 1993, 3(10): 668-674

[40] Nelson DR, Koymans L, Kamataki T, et al. P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics, 1996, 6(1): 1-42

[41] Burchell B, Nebert DW, Nelson DR, et al. The UDP glucuronosyltransferase gene super family: suggested nomenclature based on evolutionary divergence. DNA Cell Biol, 1991, 10(7): 487-494

[42] White JA, McAlpine PJ, Antonarakis S, et al. Guidelines for human gene nomenclature (1997). HUGO nomenclature committee. Genomics, 1997, 45(2): 468-471

[43] Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. Methods Enzymol, 1983, 91: 524-545

[44] Tweedie S, Braschi B, Gray K, et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res, 2021, 49(D1): D939-D946

[45] Bruford EA, Braschi B, Denny P, et al. Guidelines for human gene nomenclature. Nat Genet, 2020, 52(8): 754-758

[46] Olender T, Nativ N, Lancet D. HORDE: comprehensive resource for olfactory receptor genomics. Methods Mol Biol, 2013, 1003: 23-38

[47] Safran M, Chalifa-Caspi V, Shmueli O, et al. Human gene-centric databases at the weizmann institute of science: geneCards, UDB, croW 21 and HORDE. Nucleic Acids Res, 2003, 31(1): 142-146

[48] Zhang X, Firestein S. The olfactory receptor gene superfamily of the mouse. Nat Neurosci, 2002, 5(2): 124-133

[49] Smith CL, Blake JA, Kadin JA, et al. Mouse genome database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res, 2018, 46(D1): D836-D842

[50] Shimoyama M, De Pons J, Hayman GT, et al. The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic Acids Res, 2015, 43(D1): D743-D750

[51] Howe DG, Bradford YM, Conlin T, et al. ZFIN, the zebrafish model organism database: increased support for mutants and transgenics. Nucleic Acids Res, 2013, 41(D1): D854-D860

[52] Yates B, Braschi B, Gray KA, et al. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res, 2017, 45(D1): D619-D625

[53] Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. Nat Genet, 2000, 25(1): 25-29

[54] UniProt C, Martin MJ, Orchard S, et al. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res, 2023, 51(D1): D523-D531

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net