pParse 的科研历程

袁作飞

2012-01-03

pParse 是我博士生阶段第一个完整的科研成果，经历了调研、实验、写作、投稿、拒稿、接收的全过程，历时三年多。整个过程走下来确实非常不容易，心态也发生一些变化，从刚开始的不服，到后来的看淡，再到最后的自信。不管是科研能力，还是自信心，都有了质的提高。我把这个过程记录下来，希望能给大家一些启示。

总体历程

2008 年

09.12：贾伟、卢庄来所里讨论母离子质量校准的问题。

09.13-12.19：调研、实验

2009 年

01.17-12.26：实验、写作、投稿 MCP。

2010 年

03.01：MCP 返回审稿意见，拒稿。

03.01-09.16：修改、投稿 JPR。

10.24：JPR 返回审稿意见，拒稿。

10.24-12.31：完成 Proteomics 的初稿。

2011 年

02.10：投稿 Proteomics。

04.04：Proteomics 返回审稿意见，拒稿但可以大修改再投。

08.31：修改、投 Proteomics 第二稿。

10.19：Proteomics 返回审稿意见，小修改。

10.31：修改、投 Proteomics 第三稿。

11.02：Proteomics 返回审稿意见，接收。

详细历程

2008 年

9 月 12 日，北京蛋白质组研究中心的贾伟和卢庄来计算所找付岩师兄讨论岩藻糖的定量实验。他们发现定量数据的鉴定结果中存在母离子单同位素峰判错的情况。我和刘超在做相关的工作，所以付岩师兄安排我们参与这个问题的研究：他负责文献调研，当时搜索到两篇相关文献；我负责通过扩大母离子质量误差的方式估计鉴定的谱图中有多大比例属于母离子判错；刘超协助我的工作。通过扩大母离子质量误差得到在岩藻糖的定量数据上母离子判错的比例为 10%左右，初步判断母离子判错问题对鉴定影响较大，当时查到和这个问题相关的文献很少，所以决定研究这个问题，需要判断母离子的单同位素峰和电荷。

贾伟看谱时在 RAW 文件中只显示一级质谱，然后检查在连续的一级质谱中母离子的单同位素峰是否正确。受他的启发，我在判断母离子的单同位素峰时，除了检查当前一级质谱上的母离子单同位素峰，还需要检查相邻一级质谱上的母离子单同位素峰。最初的方法是在当前一级质谱和前后各三张一级质谱上分别判

断单同位素峰，然后投票产生最终的结果。和卢庄讨论后她建议把一级质谱的范围扩大到整个保留时间范围。在岩藻糖的定量数据上通过判断母离子的单同位素峰和电荷之后，提高了 20%左右。考虑到岩藻糖的定量数据经过富集，富集的谱图数比不富集的谱图数少，海鹏师兄建议到一般的数据上试试，比如 Haas 的酵母数据。

12 月 19 日，我到北京生命科学研究所参加文献讲评，在去的路上想到了利用同位素峰簇的强度随保留时间变化的相似性。过去之后，景志毅所讲的文献 MaxQuant 刚好利用了这个信息。当时的第一感觉是大家所见略同，第二感觉是这个信息确实很重要。回所后开始思考我的方法和 MaxQuant 的不同之处，发现它在判断单同位素峰时使用的是平均氨基酸模型的方法，这点是我可以突破的地方，我采用同位素峰强度比值的约束来判断。另外我的方法考虑了混合谱，这是 MaxQuant 忽略的地方。这期间，我注意到二级谱的相似性也可以提高谱图鉴定率，并做了相应的实验。

2009 年

1 月 17 日完成的第 0.3 版完整中文稿中包含母离子判断和二级谱相似性聚类的工作。

2 月 2 日春节回来后修改中文第 0.3 版，母离子判断和二级谱相似性聚类直接关系不大，为了突出重点，放弃了二级谱相似性聚类的研究，重点研究母离子判断，修改成中文第 1 版，两个星期后完成英文第 0.1 版。中间发现母离子判断的一个重要工作是方法的评价，当时觉得正反库 FDR 的评价方法不够可靠，组长和海鹏师兄提了建议，除了反转库的评价还需要其他验证手段，采用了海鹏师兄写的提取三个特征的 Matlab 程序：匹配离子的连续性、匹配离子占总离子的强度比例、匹配离子占总离子的个数比例。

3 月，做了相似度分布、强度误差分布、母子离子的质量误差分布的统计，并利用 pFind 当时最新支持的非对称误差设置进行搜索；和贺老师详细讨论了需要补充的实验，并完成中文第 2、3 版。

4 月，根据付岩师兄的审阅意见修改第 3 版；判断保留时间的范围；测试母离子判断的代码。

5 月，进行数据分析的流程设计，并和 extract_msn 比较；修改中文第 3 版；调试母离子判断的代码。

6 月，稳定了母离子判断 pParse 的版本；利用非对称窗口重新搜索数据，进行过滤结果的比较；和贺老师讨论工作进展和计划；测试 MaxQuant；完成中文第 4 版。

7 月，冻结 pParse 的版本，发现 pXtract 和 RawXtract 的不同，重新测试 Haas 数据，并开始测试 DDDT 数据。

8 月，重做 Haas 和 DDDT 数据，比较 pParse、MaxQuant、pXtract、RawXtract 四个软件，完成中文第 5.1 版。

9 月，阅读周耀旗关于写作方法的文章，完成中文第 5.2 和 6 版；完成英文第 1、2、3 版；pParse 支持 MS2 格式。

10 月，完成中文第 7.1、7.2、7.3 版和英文第 4.1、4.2 版；搜集了张师姐对中文第 7.2 版的意见和组长、海鹏师兄对英文第 4.2 版的意见；和海鹏师兄一起修改了英文第 4.3 版，自己完成了英文第 5.1 版。

11 月，和组长详细讨论图表的形式、文章的内容以及结构；和董老师修改文

章的引言和方法部分；和贾伟详细讨论文章的内容以及结构；完成英文第 6.2 版；搜集付岩师兄、贺老师的意见，修改句子的表达和文章的连贯性；完成英文第 7、8 版。

12 月，修改 pParse 判断母离子的类别，修改 pParse 的界面，完成英文第 9、10、11 版。12 月 26 日，投稿到期刊 MCP。

2010 年

3 月 1 日，MCP 的审稿意见如下：

主编拒稿的理由：

1. 没有可用的软件、算法的文档描述和代码；
2. 缺乏性能的评价；
3. 缺乏方法不足的讨论。

第一个审稿人的主要意见：

1. 在附件中提供算法的详细步骤和 Matlab 代码；
2. 在结果和讨论部分增加讨论的分量，解释一些现象；
3. 引用一些文献；
4. 对其它数据格式进行调研，是否存在 miss-assignment 的现象；
5. 和 Mascot 的同位素峰选项比较；
6. 解释这个参数的含义；
7. 解释数据的选取原因；
8. 解释 centroid 和 profile 模式的影响；
9. 图的解释不清楚，或者放到附录；
10. 解释数据的比较；
11. 解释为什么不输出所有的母离子；
12. 对定量的展望多一些。

第二个审稿人的主要意见：

1. 文章缺乏新意；
2. 对去同位素峰的概念不清楚；
3. 没有可执行的程序。

9 月 16 日，投稿 JPR 版本的修改如下：

1. 在附件中提供算法的详细步骤和 Matlab 代码；
2. 增加了灵敏度的评价试验；
3. 重新选择了数据集；
4. 需要在讨论部分中解释 centroid 和 profile 两种模式的区别和影响，在选择数据时要考虑数据模式；
5. 比较输出一张二级谱的所有母离子和前两个母离子的区别和影响；
6. 增加 Mascot 的同位素峰选项的比较实验；
7. 增加 pParse 的缺点描述；
8. 对余下的部分进行增加、解释等。

10 月 24 日，JPR 的审稿意见如下：

主编拒稿的理由：

根据两位审稿人的意见以及主编自己的评价，不能接收这篇文章。

第一个审稿人的主要意见：
1. 引用一些文献；
2. 比较去同位素峰的方法时不准确；
3. 注意仪器上更新对本文问题的影响；
4. pParse 依赖第三方的软件；
5. 信噪比的计算方法描述不清楚；
6. 参数 1.00307 是如何得到的。
第二个审稿人的主要意见：
1. 语言上有描述不清楚的地方；
2. 和已有软件的比较不清楚；
3. 强度比值和平均氨基酸模型的区别在哪里；
4. 怎么确定 noise，相似度的阈值怎么确定；
5. 强度加权平均对低丰度谱峰是否有效；
6. 系统误差校准是如何起作用的；
7. 不用数据导出工具的质量误差分布的比较描述不清；
8. pParse 对肽鉴定率有提高，其它软件呢；
9. 混合谱的搜索、过滤和原来的谱有没有不同；
10. 混合谱的发现应放在讨论中，共洗脱离子的相对强度描述不清；
11. 碎裂窗口外母离子的鉴定效果不明显；
12. pParse 和数据库搜索方法的比较描述不清；
13. 中心化的细节描述不清，改进效果也不明显；
14. 表 1、2、3 在文章没有讨论，可放到附录中；
15. 表 7 的自我解释性太差；
16. 仪器的更新对母离子挑错有影响；
17. 分离的方法对混合谱的比例有影响；
18. FDR 用 2%的原因是什么。

12 月 31 日， Proteomics 初稿的修改如下：
1. 把文献综述写完整；
2. 根据实验需要调整了数据集；
3. 增加了 precision 的评价；
4. 调整了实验结果和讨论的内容；
5. 增加蛋白覆盖率的内容；
6. 通过预搜索确定参数的阈值；
7. 增加了对重叠谱峰的判断；
8. 强调了不同实验条件导致不同的结果。

3 月 1 日收到 MCP 拒稿消息的时候，感觉自己受到了不公正的对待，认为有的审稿人不负责任；后来再看这些审稿意见，觉得主要责任在自己，做了很好的工作，但没有展示清楚，文章写作上有很多漏洞。一项好的研究，除了实验效果好，方法有创新，文章也要写的清楚明白。两次拒稿的经历，让我有了好的心态，我相信能把文章写好。

pParse 的程序版本比较稳定，开发了如下四个版本：pParse_20100322、pParse_20100531、pParse_20100910、pParse_20101227。第一个是 MCP 的版本；

第二、三个是 JPR 的版本，第四个是 Proteomics 的版本，和上面的修改内容对应。

2011 年

4 月 4 日，Proteomics 的审稿意见如下：

主编拒稿的理由：

文章在当期被接收的优先级不够高，建议大修改后再投。

第一个审稿人的主要意见：

1. 需要明确指出本文和已有方法的区别是实现方式；
2. 混合谱的鉴定中一谱鉴定出多肽的情况有多少；
3. pParse 和 BioWorks 的结果进行比较；
4. 混合谱单独鉴定的肽段和蛋白的比例在表述上不准确，有两种修改方式；
5. 把图 4 中的 a、b、c 清晰的划到 Tabb data 上，d、e、f 的划到 Haas data 上。

第二个审稿人的主要意见：

1. 本文把四种基本技术放在一起，没有新颖性；
2. 标注集的选择太随意；
3. 参数太多，参数如何调节，是否有过拟合的现象；
4. 标注集只选择了有限的数据，所以 sensitivity 的定义不准确；
5. 评价指标只提了 sensitivity，没提 specificity；
6. pParse 和 BioWorks 的结果进行比较；
7. 肽段同位素峰簇的判断和碎裂是两回事，一级谱去同位素峰和碎裂窗口没关系；
8. 在标注集中的'labeled'一词有歧义。

8 月 31 日，投 Proteomics 第二稿包括：

1. 方法的新意在于“平均氨基酸模型”的利用，即同位素峰簇中最高峰的位置和质量的关系；
2. 标注集的获得以及画 ROC 曲线；
3. 同类软件的比较；
4. 利用 UIS 鉴定共洗脱母离子；
5. 共洗脱母离子对实验的影响。

10 月 19 日，Proteomics 的审稿意见如下：

主编的意见：

小修改，建议找英语国家的人帮忙修改。

第一个审稿人的主要意见：

1. 一些语法错误需要修改；
2. 注意时态。

第二个审稿人的主要意见：

没有进一步的问题了。

11 月 2 日，文章被接收；12 月 2 日，排版后的校正；12 月 20 日，在网上公布。