十年回顾与展望

贺思敏

2020-7-12

2020 年是 CNCP（China Workshop on Computational Proteomics，中国计算蛋白质组学研讨会）召开十周年。原计划在 8 月份举办第六届 CNCP 会议，由于新冠疫情的影响，会议不得不推迟到明年 8 月再举办。作为主办方，我总觉得在 CNCP 十周年之际应该做点什么，于是决定邀请咱们 CNCP 报告人在群内做一次“十年回顾与展望”的分享。作为倡导者，我肯定得带个头，以示诚意。

我想谈三点：CNCP 这十年，我这十年，未来十年。

CNCP 这十年

首届 CNCP 于 2010 年召开，迄今已经办过五届、历经十年。大家先看看下面五届会议的合影，找找当年的自己在哪里。

CNCP 不是什么大事，但是任何一件事情历经十年，都不那么轻巧。我本是一个激情有余、长性不足之人，所以对于 CNCP 能坚持到今天，内心也稍感惊讶。CNCP 这十年主要做了三件事情：一是会议本身，二是技术培训和技术评测，三是 CNCP 论坛。

- **关于会议本身**

五届十年，比较难的是保持水准。一般来讲，第一届的时候，肯定小心翼翼、兢兢业业，第二、第三届往往会松懈许多，第四、第五届即使有些倦怠也不难理解。这是人之常情，在中国尤其如此，大家参观一下开办十年的五星酒店就会深有感触。但是批评别人很容易，轮到自己就很难说了，所以不断提醒自己、提醒同伴，要警惕自身的懈怠。

总体上说，我觉得五届 CNCP 会议保持了“学术优先、其他从简”的风格，办成了“小人物的聚会”。这与当下“精英”“峰会”的办会时尚还是很有些差异，所以也会引起不解。记得有一年我在计算所年终述职，提到当年主办了 CNCP，坐在台下的计算所老所长李国杰院士说他可以帮我邀请贺福初院士参会（他俩都是湖南人），蛋白质组学的会议没有贺院士参加可能水平不够。我说我们这个会议特意不请院士，万一院士讲得兴起不下台，我就不好

控制时间了。老所长一片好意，我却没领情；老所长在所内各种场合也很支持 pFind 组所做的交叉研究，我却一次也没邀请他去 CNCP 讲个话——当然我这也是一片好意。

- **关于技术培训和技术评测**

这是常规会议之外的 10%创新。技术培训是指董梦秋在 CNCP-2014、CNCP-2018 会议前后主办的两届交联质谱技术培训，不仅讲课，而且学员实际操作完整干湿流程。交联质谱技术算是领域新技术，梦秋在国内是先行者，在国际上也是知名学者，她把自己的看家本领与同行分享，这在科研竞争空前激烈的当前殊为不易。派人参加过培训的黄超兰、周虎团队后来利用交联质谱技术分别支持施一公和蒋华良团队在 Science 和 NSMB 发表了文章。下面的 CNCP-2018 培训合影中有主教官梦秋，我还认出有第四届 CNCP 报告人赵群。

技术评测是指 CNCP-2018 同期刘超组织的首届蛋白质组学技术评测。刘超在 pFind 团队读博士期间主攻定量软件 pQuant，曾经在梦秋实验室学习过色谱质谱操作，博士毕业留组后又到田瑞军实验室合作研究过色谱质谱技术，是屈指可数的水陆两栖战士。刘超基于迟浩主持研发的最新版 pFind 设计了评测内容，利用后端的数据分析环节为前端的数据采集环节提供诊断意见。如果说开个会还只是纸上谈兵，那技术评测就是实兵演练了，特别是陆军开始给水军提建议，水陆交叉更上一层楼，我觉得很有意义。今年在周虎和迟浩协助下，刘超将组织第二届技术评测，评测范围从上一次的色谱质谱环节进一步推进到样品制备环节，算是又有了 10%创新。

- **关于 CNCP 论坛**

这是指 cncp-speaker@这个邮件群，目前包括五届 CNCP 报告人共计 114 人。特意没有加入杨芃原教授，因为他虽是首届报告人，但是不属于“小人物”。这个论坛把两年一次的会议交流扩展为随时在线的网上交流。2018 年 2 月 4 日我写过一个邮件“CNCP 论坛 2017 年大事记”，有个总结：

“CNCP 论坛 2015 年开张，已经运行三年，每年增加一点儿变化，我称之为‘10%创新’：2015 年主要是邀请同行评价 CNCPer 作品，2016 年增加了 CNCPer 每月作品荟萃，2017 年增加了 CNCPer 作者回顾。回顾的意义，不仅在于彼此启发、彼此照亮，也定义了‘你的作品’：你可以为之写回顾的文章才是你的作品。

“今年最可喜的事情，就是群体性地突破 Nature 子刊。特别值得感谢的是，刘铭琪/曾文锋、李栋/李杨、叶明亮、肖传乐不仅回顾了研究/写作/投稿历程，而且分享了审稿意见，这使得顶刊审稿过程不再神秘，相信这会帮助更多的 CNCPer 团队及早完成突破 CNS 子刊这一阶段性任务。

“还有同样可喜的事情，那就是论坛开始出现学术批评，比如围绕如何评价 Olsen 文章展开的交锋。特别感谢张弓率先亮剑，激发了后续的讨论。程仲毅与张弓有关科学情怀和产业化之间碰撞与相融的交锋，张弓和陶生策应邀发表的评论，也同样精彩，令人回味无穷。”

稍作补充说明：

2015 年首次发表的同行评论，是周愿评论李灵军的 AC 封面文章，后来我邀请新朋友写作评论总是拿周愿这个评论做范本。

2017 年首次发表的作者回顾，是屠成剑、李婧的作品自述，虽然作者自称是小文章，但是我挺喜欢自述中展现的生动历程，这两天还重新看了一遍。

2017 年 9 月我推荐了 Olsen 团队关于 HeLa 细胞蛋白质组接近完全覆盖的 Cell Rep 新作，近两年 Mann、Cox 文章认为这是人类蛋白质组深度覆盖的代表性成果之一。当时贾辰熙、张弓、杨靖、黄超兰、刘铭琪、叶明亮、田瑞军、谭敏佳先后参加评论，但总体评价不高，令我深感意外。CNCP-2018 会后铭琪和我还有新的讨论，铭琪的态度好像有些变化。

Nature 子刊审稿意见和投稿历程分享，后来还有 2018 年迟浩的 Open-pFind 投稿 NBT 的分享，2019 年我的博士生陈镇霖的新版 pLink 投稿 NC 的分享。黄光明 2019 年初分享的科研历程回顾也十分精彩。

简单做个总结，三点感想：

第一，虽然 CNCP 会议名称为“计算”蛋白质组学，但是 CNCP 会议上偏计算的报告始终不如偏实验的多，说明计算方面的人才培养还是不够快，这肯定制约蛋白质组学的发展。

第二，大陆蛋白质组学成果开始突破 CNS 期刊，相关审稿意见和成文历程的分享有助于破除所谓顶刊的神秘感。期待看到更多的分享，促成更多的 CNCP 朋友跨过这一关，从而实现“对 CNS 病毒的群体免疫”:-)。

第三，Olsen 文章价值之争，是科研价值观和方法论差异的表现。这是 CNCP 这十年我印象最深刻的一件事。

我这十年

这次 CNCP 十周年之际的回顾虽然源于“CNCP”，其实重点在于“十年”。无论对于谁，十年都是沉甸甸的，CNCP 只是沧海一粟。所以我想谈谈我这十年、也是 pFind 这十年。回顾十年的历程实在太不轻松了，所以我重点回顾其中一年。我会选择哪一年呢？

我会选择 2010 年。是因为 2010 年举办了首届 CNCP 吗？那只是其中一个因素。最重要的原因，是那一年我和 pFind 团队发生了很多“首次”：

- **3 月份：首次参加 ABRF 会议、RECOMB-CP 会议**

孙瑞祥去美国加州州府 Sacramento 参加了 ABRF（Association of Biomolecular Resource Facilities）年会。他和迟浩利用 pFind 引擎参加了 ABRF 的蛋白质组信息学研究组 iPRG 组织

的大规模磷酸化质谱图的鉴定评测，鉴定数量和一致性排在前五名，这比 2008 年 pFind 首次参加 iPRG 评测表现好很多，从平均水平提升到第一梯队。

紧接着瑞祥和迟浩去美国 UCSD 参加了 RECOMB-CP 国际会议。RECOMB 是生物信息学两大顶会之一（另一个是 ISMB），而 RECOMB-CP 是计算蛋白质组学专题分会，是计算蛋白质组学最合适的国际会议，2006 第一届，2010 第二届。会上迟浩报告了 pNovo 软件，瑞祥 3 月 29 日 11:29 向组里发来题为“pNovo 轰动 UCSD”的邮件：“刚才 pNovo 的报告彻底轰动了 UCSD，Pevzner said 'I am shocked by the performance of pNovo on HCD...' 问问题是最多的一个报告。会后有很多希望跟我们合作的人。” Pevzner 是计算蛋白质组学的老大。

这是瑞祥、迟浩第一次去美国，见识了国际同行对于中国学者与会的意外和惊奇。瑞祥回国后写了一个总结，题目很煽情：“pFind 在 2010 年春起步迈向国际舞台”。他还写到：“开会期间还偷访了 Pevzner 的办公室，标志 pFind 人到此一游。”下有照片为证，可惜忘了刻字留念了。瑞祥还提到顺访蛋白质组学的老大之一 Yates：“这次见到 Yates，一个地下工作者。”因为 Yates 的办公室和实验室都在地下一层。瑞祥真是个有趣的人。

- **4 月份：首次为中日韩生物信息学培训班授课**

我、付岩、迟浩带领三位硕士生访问上海。付岩在谢鹭组织的第十届中日韩生物信息学培训班利用 pFind 软件为学员讲授“蛋白质组学质谱数据分析”。我们和曾嵘团队交流了科学院课题的进展，和谢鹭、陆豪杰团队交流了 973 课题“蛋白质组海量质谱数据的深度解析”中腾冲嗜热菌实验和数据分析初步结果，当时的认识是：“对于高精度串联质谱数据的深度利用，对于谱图鉴定率低、蛋白覆盖率低等老问题，需要下大功夫攻坚，否则依照目前的常规数据分析手段，难有新意和进展。”可见当时对深度解析这个问题还完全没有思路，更不会想到八年之后迟浩在深度解析方面居然突破 NBT。豪杰还专门带我参观了杨芃原教授的质谱仪研制现场，我很受鼓舞，也很受感动，因为研制硬件比我们研制软件需要更多的投入、更长的时间、更大的勇气。

- **5 月份：首次应邀作报告，首次发布 pFind 软件**

我和瑞祥、袁作飞、王乐珩（pFind 软件架构师）去丽江参加了杨芃原、钱小红教授组织的“蛋白质组数据处理暨全国第三届生物质谱学术交流会”。我和瑞祥首次应邀在国内会议上作报告，而且我在会上首次宣布 pFind 2.4 面向国内外用户开放下载。

我的报告在 5 月 16 日。为了准备首秀，两个月前就开始准备，作飞、刘超作为我的助手。5 月 15 日夜里我在丽江准备报告，而迟浩则在北京带着几位博士生、硕士生与在丽江的乐珩一起为 pFind 软件首次发布鏖战。5 月 16 日凌晨 2:32 我给 pFind 全组发邮件：

“我在丽江，刚刚准备完毕明天的报告。这是连续第二天熬夜了。可喜的是我的报告有很大改进，思想上有许多新认识，明天早晨应当会不辱使命。不是明天早晨，应该是 6 个小时之后的今天早晨。看到刘超的邮件，更加心情激动。

“我知道这几天，多数 pFinder 也一样在高强度战斗，很多 pFinder 也在熬夜。从瑞祥到一年级新生，我不想一一指出他们的名字，只想说：高手之间智力上、精神上互相砥砺，互相激发，是人生值得回忆和纪念的事情。我为有你们这样的同伴感到自豪！

“记住此刻！”

我的报告核心内容是基于 pFind 的 500Da 大窗口搜索来深度解析质谱图，可以发现肽段母离子挑峰错误、多个母离子的混合谱、意外修饰和意外酶切及其组合。报告也重点提及 2009 年 NM 发表的 HUPO（Human Proteome Organization）组织的质谱鉴定国际评测，绝大多数实验室表现不佳，这对八年后我决定在 CNCP-2018 组织首次技术评测影响很大。

我的报告得到了前辈钱小红、同辈谢鹭、晚辈李虹（谢鹭学生）的肯定，谢谢她们当年对我首秀的鼓励。会上认识了徐平、余维川，他俩在我报告现场有提问（猜猜下面举手的那位是谁）。还认识了张丽华、叶明亮。当年 10 月我访问了徐平实验室，12 月我参加成都 973 项目进展会之后去大连访问了丽华实验室。

- **6 月份：首次见到 Mann**

刘超去德国参加了 MaxQuant Summer School，并与 Mann 合影留念。这是刘超第一次出国，当时他感觉自己的 pQuant 已经超越了 Yates 的 Census，但是这一次培训发现还没有超越 Mann 的 MaxQuant，于是又是几年拼搏，直到 2014 年才有足够的信心投稿 AC 并发表。刘超在读博时就见到了本领域的老大之一，当时的刘超幸福地闭上了眼睛。

- **7 月份：首次突破 ISMB，首次参会 ASMS，首次集体参加质谱大会**

付岩指导硕士生叶叮完成了开放式谱库搜索软件 pMatch，被生物信息学顶级国际会议 ISMB 录用，文章同时在 Bioinformatics 发表，很兴奋，因为当时我们发表一篇英文论文已觉不易，中稿顶会就更觉困难（团队第二次中稿 ISMB 已经是九年之后的 2019 年了）。付岩和叶叮参加了 ISMB 会议，这也是我们组首次支持硕士生出国开会。付岩还首次参加了 ASMS。

7 月 28 日到 8 月 4 日，pFind 团队集体去长春参加全国质谱大会和华人质谱大会，瑞祥、付岩作学术报告。我对刘斯奇教授在会上所作的文献综述报告有印象，他对蛋白质组学做了

反思和辩护。会后游览了长春净月潭、长白山天池。由于松花江涨水，我和瑞祥在漂流时充气艇翻了，而且恰好倒扣在身上，会游泳也没用，新配的眼镜也献给龙王了。这是我们组第一次集体远行。那时也是我们组人员最多的时候。

- **11 月份：首次举办 CNCP**

当时最纠结的是要不要办这个会。倒不是怕办这一届，而是一旦办了第一届，那么后面就得坚持办下去。最终还是下决心办！瑞祥费心最多。很感谢老关（慎恒）会前两天长达 10 个小时的质谱基础知识培训，我现在也没这个水平。下面的培训学员合影中有主教官老关，我还认出有首届 CNCP 报告人王全会，还有 CNCP 论坛第一位评论员周愿。

时刻不忘学习的 Thermo 中国公司的方宇以个人身份参加了首届 CNCP，觉得会议挺好，后来在她的积极建议下，Thermo 中国公司赞助了后几届 CNCP，李静、贾伟、张伟、周岳等 Thermo 中国公司的朋友也给予各种支持。

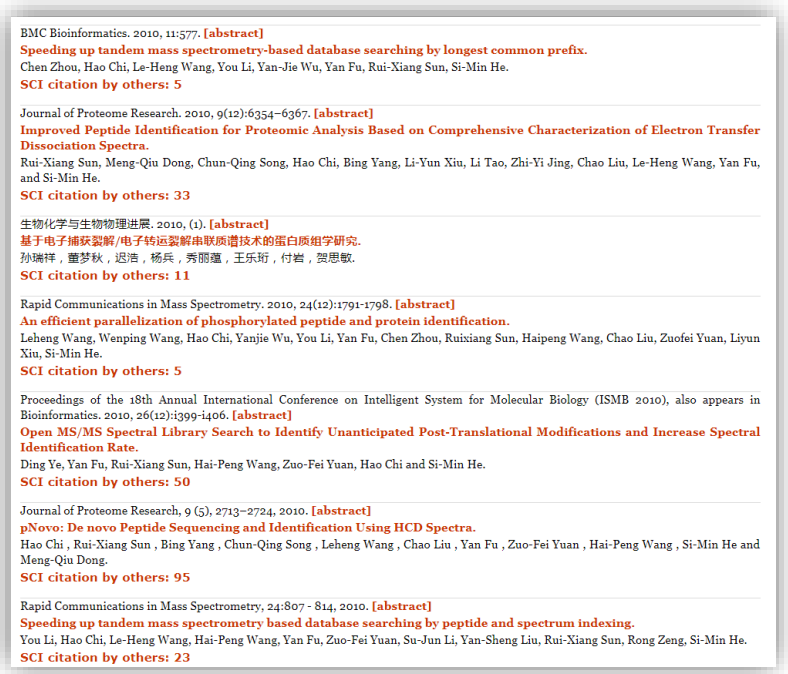
- **全年：首次发表六篇英文期刊文章**

这六篇文章包括 JPR 两篇，RCM 两篇，Bioinformatics 和 BMC Bioinformatics 各一篇，并不高大上。但是每当我回忆 2010 年时，首先想到的却是这六篇文章，为什么呢?

首先是因为数量。本来我们英文写作能力就不强，又刚进入一个新领域，所以写文章一直是我们的难题。这一年发表六篇文章，又都是自己成员主写，对于我们组来讲，不仅是“空前”，而且直到现在还属于“绝后”。时至今日，写作对于我们组，依旧不是轻巧之事。

其次是因为很多“第一次”。比如迟浩的 pNovo 是他主写的第一篇文章，不仅是他的代表作之一，而且也是从头测序领域的经典文献之一。这也是我们团队首次在 JPR 发表文章。瑞祥的代表作 ETD 研究也发表在 JPR。这两篇文章都是与梦秋合作完成。Bioinformatics 那篇是我们组第一次突破 ISMB，前面提到过。BMC Bioinformatics 那篇虽然引用不多，但是用到了一个比较高级的数据结构，我很欣赏，而且 Pevzner 团队成员、MS-GF+作者 Kim 在 2012 年 RECOMB-CP 的培训报告中有一页特别介绍了我们这篇文章，刘晓文也引用过，能得到少数几个小同行认可我也很高兴。

再次是因为这是我们组硕士生毕业表现最好的一年。我们对硕士生有个期望，那就是毕业时有篇文章发表，两个人合写一篇 RCM 就可以。2010 年，两位本科毕业于华中科大计算机系的硕士生在毕业前一年就发表了文章，另外四位硕士生两两合作分别完成一篇 RCM 文章，捍卫了自己的荣誉。此后组内毕业的硕士生少数能完成一次写作投稿，其中两三位能发表，我有些遗憾，部分原因可能是期望比 RCM 高了，一时达不到。



BMC Bioinformatics. 2010, 11:577. [abstract]
Speeding up tandem mass spectrometry-based database searching by longest common prefix.
Chen Zhou, Hao Chi, Le-Heng Wang, You Li, Yan-Jie Wu, Yan Fu, Rui-Xiang Sun, Si-Min He.
SCI citation by others: 5

Journal of Proteome Research. 2010, 9(12):6354–6367. [abstract]
Improved Peptide Identification for Proteomic Analysis Based on Comprehensive Characterization of Electron Transfer Dissociation Spectra.
Rui-Xiang Sun, Meng-Qiu Dong, Chun-Qing Song, Hao Chi, Bing Yang, Li-Yun Xiu, Li Tao, Zhi-Yi Jing, Chao Liu, Le-Heng Wang, Yan Fu, and Si-Min He.
SCI citation by others: 33

生物化学与生物物理进展. 2010, (1). [abstract]
基于电子捕获裂解/电子转运裂解串联质谱技术的蛋白质组学研究.
孙瑞祥，董梦秋，迟浩，杨兵，秀丽蕴，王乐珩，付岩，贺思敏.
SCI citation by others: 11

Rapid Communications in Mass Spectrometry. 2010, 24(12):1791-1798. [abstract]
An efficient parallelization of phosphorylated peptide and protein identification.
Leheng Wang, Wenping Wang, Hao Chi, Yanjie Wu, You Li, Yan Fu, Chen Zhou, Ruixiang Sun, Haipeng Wang, Chao Liu, Zuofei Yuan, Liyun Xiu, Si-Min He.
SCI citation by others: 5

Proceedings of the 18th Annual International Conference on Intelligent System for Molecular Biology (ISMB 2010), also appears in Bioinformatics. 2010, 26(12):i399-i406. [abstract]
Open MS/MS Spectral Library Search to Identify Unanticipated Post-Translational Modifications and Increase Spectral Identification Rate.
Ding Ye, Yan Fu, Rui-Xiang Sun, Hai-Peng Wang, Zuo-Fei Yuan, Hao Chi and Si-Min He.
SCI citation by others: 50

Journal of Proteome Research, 9 (5), 2713–2724, 2010. [abstract]
pNovo: De novo Peptide Sequencing and Identification Using HCD Spectra.
Hao Chi , Rui-Xiang Sun , Bing Yang , Chun-Qing Song , Leheng Wang , Chao Liu , Yan Fu , Zuo-Fei Yuan , Hai-Peng Wang , Si-Min He and Meng-Qiu Dong.
SCI citation by others: 95

Rapid Communications in Mass Spectrometry, 24:807 - 814, 2010. [abstract]
Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing.
You Li, Hao Chi, Le-Heng Wang, Hai-Peng Wang, Yan Fu, Zuo-Fei Yuan, Su-Jun Li, Yan-Sheng Liu, Rui-Xiang Sun, Rong Zeng, Si-Min He.
SCI citation by others: 23

我对 2010 年印象深刻还有一个原因。pFind 组是 2002 年为了申请和完成第一个 973 课题“基于信息技术的蛋白质组研究”而成立的，到 2010 年算是第一个十年接近终点，可以说是“开局的结束”。2010 年又开始第二个 973 课题“蛋白质组海量质谱数据的深度解析”，是新的开始。从对计算蛋白质组学一无所知，到 2010 年的诸多“首次”，可以看得出团队的发展态势：虽然并不知道还能走多远，但是愿意继续走下去。结果一走就走到 2020 年，正好是第二个十年，对比 2010 年，最大的感受有三点：

第一， 成果进步显著。

文章方面，总体上我们已经从 RCM 的水平稳定进入到蛋白质组学主流期刊比如 JPR、Proteomics、AC、MCP 的水平；而且通过与梦秋、杨芃原团队合作或独立研究，2012 年交联质谱鉴定软件 pLink 突破 NM，2017 年完整糖肽鉴定软件 pGlyco 2.0 突破 NC，2018 年开放式搜索鉴定软件 Open-pFind 突破 NBT，2019 年新版 pLink 再破 NC，学术自信初步确立。回想 1997 年我从清华大学毕业时，只有两篇国内期刊文章在手；导师告诉我，整个清华大学信息学院（计算机系、自动化系、无线电系、微电子所）每年能发表两篇 SCI 文章的只有两人！当时能在国际会议发表一篇文章已经非常难了，很多重要年会中国常年缺席，或者每年至多有一两篇，在主流国际期刊上发表文章也屈指可数。记得 1995 年我作为博士生接待过一名在 Science 发表文章的美国学者，我都没带他见导师，因为那时候我完全不明白 Science 是什么。

软件方面，2010 年首次发布 pFind 之后，十年来我们还发布了 pLink、pNovo、pQuant、pAnno、pTop、pGlyco、pMatch、pCluster、pParse、pDeep 等新软件和新版本。2019 年底统计，国内外同行注册下载软件 5700+，利用软件发表文章 336 篇（包括 CNS 正刊 23 篇、CNS 重要子刊 45 篇）。文章与软件并重，pFind 团队的科研特色初步确立。2017 年 2 月我写过一封邮件讨论此事，今年五一节发表在 pFind 网站，名为《pFind 团队的追求与道路》。我在计算所的同事包云岗研究员今年初写过一篇文章《伯克利科研模式的启发》，计算所所长孙凝晖院士称之为“科研重工业模式”，推荐诸位一读。

我时常想起付岩十五年前（2005-6-26）给我的一个老邮件：“今天课题组汇报时，有几位老师对 pFind 的推广和定位很关心，也有质疑。似乎 pFind 已经是成形甚至成熟的软件了。但是我觉得 pFind 现在还有很多欠缺，可能连个‘another’都还不是，如果现在过分宣传，可能会适得其反。”付岩早在 2004 年就在 Bioinformatics 发表了 pFind 组第一篇文章，其中的 KSDP 打分算法后来成为 pFind、pLink 的第一个打分算法，但是他冷静地认识到从算法到软件没那么简单。2010 年对外发布 pFind 2.4 之后，pFind 才逐步成为一个同行可用的软件，但是总体上属于“just another”。直到 2018 年 Open-pFind 在 NBT 发表，才算有了一点“the other”的味道，这距离付岩的邮件已经十三年了。

第二， 人才脱颖而出。

2010 年作为首届 CNCP 报告人的付岩，当时已从 pFind 组博士毕业三年，2011 年底从计算所调到数学所，作为 PI 已经快十个年头，培养的博士生也已经毕业三位，其中一位现在美国贝勒医学院，希望还在咱们领域。

2010 年和我一同去丽江开会的作飞，2012 年博士毕业后去美国宾州大学 Garcia 实验室博士后学习，2018 年成为 CNCP 报告人。此刻的他带着妻子和两个孩子已经开始在 St Jude Children's Research Hospital 的 Center for Proteomics and Metabolomics 做 Senior Bioinformatics Research Scientist。

2010 年刚刚在 JPR 发表第一篇文章 pNovo 的迟浩，2013 年博士毕业，2014 年成为 CNCP 报告人，2018 年在 NBT 发表了 Open-pFind 文章，正在成长为 pFind 团队的新队长。

2010 年见过 Mann 的刘超，2014 年博士毕业，2016 年成为 CNCP 报告人，2018 年组织了首届 CNCP 技术评测。此刻的他正在北航建设自己的实验室，并在组织第二届 CNCP 技术评测。

2010 年刚刚大学毕业、进入 pFind 组开始读研究生的曾文锋，2016 年博士毕业，2017 年与复旦团队合作在 NC 发表了 pGlyco 2.0 文章，同年独立发表 pDeep 软件（领域内第一个也是目前性能最佳的深度学习预测肽段质谱图的软件），2018 年成为 CNCP 报告人。此刻的他带着刚博士毕业的妻子和一岁大的儿子正在前往德国 Mann 实验室深造的路上。

pFind 团队自 2002 年成立以来，博士毕业仅仅 11 人，目前只有上述 5 人继续坚守在计算蛋白质组学领域。他们经过风雨，见过彩虹，比我在他们这个年龄有成绩、有名气多了，我相信他们会以计算蛋白质组学作为一生的学术追求。星星之火，可以燎原。文章软件只能管一时，人才辈出才能管一世。

第三，当年真有活力。

先说出差。四月上海、五月丽江、七月长春、十二月成都和大连，一年五次出差，对我来讲属于空前绝后！

再说作报告。当年首次应邀作报告还比较有勇气，之后十年胆子越来越小，除了课题答辩、年终述职之外，我只应丽华之邀在 2013 年北京分析测试学术报告会 BCEIA 做过一次报告，比 2010 年的丽江会议报告有进步，相关内容后来在丽华/明亮、周虎/敏佳、恒樑/徐平、豪杰/铭琪/张扬、仲毅实验室小范围分享过。

再说交朋友。2010 年见过很多老朋友、结识很多新朋友，后来大都成为 CNCP 报告人。尤其是结识徐平、丽华之后，两位与我和梦秋一同成为 CNCP 组织者。最近几年在 CNCP 之外好像没结识新朋友，即使老朋友一年也都难见一回。

再说写总结。这一次为了写回顾，我特意重新阅读了 2010 年的几个参会、访问总结，其中丽江会议总结 3,000 字，长春会议总结 5,500 字，成都会议总结 13,000 字。现在我每次出差还是要写总结，自己定的规矩不敢轻易破坏，但是没过去那么详实了。

令我欣慰的是，每次 CNCP 群内分享 ASMS 参会见闻，pFinder 的总结总是最长的，可见写总结已经变成 pFinder 的习惯。也正因为十年前的文字记录，我才觉得当年我也年轻过、奋斗过，所以当我今日活力不再的时候，我才没感觉那么惭愧。今日的我做事要靠毅力，比如这个 CNCP 十周年回顾与展望的活动，其实也可以不搞，搞的话不仅自己累，也不一定有很多人响应，但是我还是决定搞，哪怕只有我这一篇文字，哪怕只给十年后的我自己看。

未来十年

我不时回顾过去，但是我很少展望未来。也许是因为安心做个小人物并不需要如此劳神，也许是因为一直觉得未来还很遥远。但是 CNCP 十周年之际，回顾之余总该做点展望，我作为倡导者更是义不容辞。这个展望会有点“浪漫主义”色彩，也就是凭感觉、凭信念、凭心愿，没那么多依据；但是基调是现实主义，着重思考：如何解决好此前十年没解决好的问题？我始终觉得咱们中国有的是人才，但很大一个问题是“信心”。

- **对蛋白质组学的信心**

2010 年长春质谱大会上，刘斯奇教授的报告中提到了当年 NBT 发表了一期回顾，其中有一篇专题文章《Proteomics Retrenches》，十年后的今天我特意看了一遍，也建议诸位一读。文中访谈的几位蛋白质组学的著名人物 Yates、Liebler、Kuster、Carr、Mann、Aebersold、Bergeron、Moritz 等都在反思蛋白质组学，大意是十年时间花了很多钱，却没办成什么事。

十年后的今天，虽然我依旧没深究蛋白质组学做成了什么事，但是我感觉蛋白质组学已经成为主流科学的一部分，不需要再为自己辩护了，而基因组学如日中天的现实就是蛋白质组学的未来——我觉得投身于这样一个极具成长性的研究方向是一件幸运的事情，无论对于我们搞计算的外来户，还是对于传统的生物学研究者！

我很欣赏原神州数码工程院院长谢耘《中国企业技术产品创新中的几个问题分析》中的观点：作为当代技术创新发展的基本特征，技术创新、产品和技术应用的突破主要是通过渐进式的积累而得以实现的，而不是依靠某个天才的灵光一现；只要这个领域的客户需求是一个长期稳定的需求，而且无法一次或经过短暂努力就能比较彻底地满足，那么企业就可以通过持续努力来建立技术壁垒。我觉得这对于蛋白质组研究也同样适用。

- **对计算蛋白质组学的信心**

计算蛋白质组学领域远不如信息领域热闹，所以在吸引、挽留优秀信息学人才方面不算很成功。计算所报考热门方向的研究生比报考 pFind 组的研究生多十倍、二十倍，pFind 组毕业的博士一半多会流失到信息领域大公司。从学术上讲，计算蛋白质组学中具有普遍意义的计算问题不多，所以我也一直苦恼如何在计算所做一个学术报告。

但是现在的我已经比过去更有信心了。既然蛋白质组学不是昙花一现、永远在成长，那么即使计算蛋白质组学只是应用成熟的计算技术来解决蛋白质组学的问题，那也足够有价值。我已经认识到“知不易、行益难”，即学知识不易、用知识更难，尤其是跨领域的知识迁移。而且在这个新方向上可以避免“踩踏效应”。

虽然要继续关注信息领域的技术进步，但是不必留恋信息领域，因为计算蛋白质组学/生物信息学已经成长为一个独立的方向，蛋白质组学/生物学有着独特的问题和广阔的成长空间，更能体现计算的价值。就像计算机科学，曾经是数学的边缘学科，受尽歧视，最终独立，反而比数学发展更快，而且更能体现数学的价值。

- **对质谱技术的信心**

我觉得，即使蛋白质组学不成功，质谱技术也绝不会失败。作为最灵敏的分析仪器，质谱仪太有用了，而充分发挥质谱仪的潜能又有太大的空间。质谱技术的稳定进步，始终是蛋白质组学发展进步的最大驱动力。以计算质谱学为主要内容的计算蛋白质组学一定会有用武之地。质谱大数据和深度学习会带来某种模式甚至范式的改变。我特别建议大家更热烈地拥抱质谱技术。

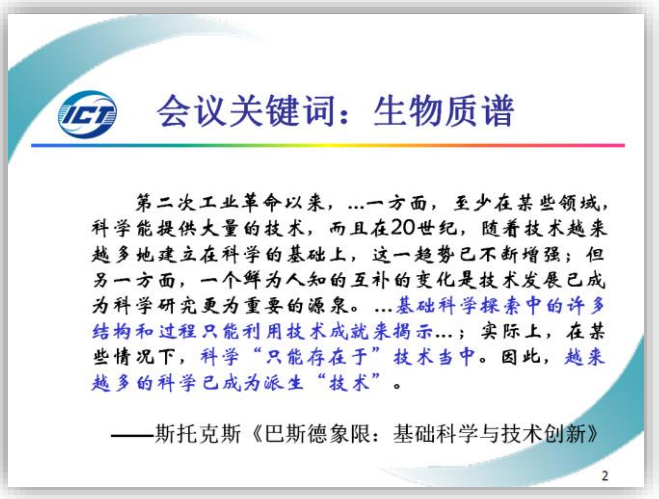
十多年来我一直没有忘记 2007 年 Zubarev 和 Mann 在 MCP 的一篇文章中的观点：

“MS-based proteomics can inherently be extremely precise and accurate. /基于质谱技术的蛋白组学本质上可以做到非常精确。

"This high accuracy is in stark contrast to many other areas of biology, which have large intrinsic errors and gray zones of interpretation. /与这种高精度形成鲜明对照，生物学的众多其他领域存在很大的固有误差和解读的灰色区域。

"It potentially makes MS-based proteomics one of the most 'digital' and error-free of the life sciences at least as concerns peptide identification./基于质谱技术的蛋白质组学，有可能成为生命科学中最为'数字化'和精确无误的分支之一，至少肽鉴定可以做到如此。"

很多科学家对于技术不屑一顾，认为技术只是科学的工具，科学才是根本的追求。我觉得诸多同行对于 Olsen 文章不屑一顾，部分出于这样的原因。我不这样看，我认为科学与技术都既是方法、又是目的，你中有我、我中有你，没有什么高下之分；技术看起来不过是量变，但是技术的量变积累到一定程度会引发科学的质变。2010 年丽江生物质谱会议上我的报告开篇就引述了这样的观点：



就像真理只在大炮的射程之内，科学也只存在于技术可达之处。杨靖实验室主页引用一句名言："Every scientific advance is an advance in method." 遗憾的是，我们既没有创造 DDA、DIA，也没有创造 SRM、PRM；既没有创造 Olsen 的 HeLa 全覆盖纪录，也没有创造 Coon 的酵母一小时全覆盖纪录 (前者可以借口代价太大，后者就不好以此为借口了)。2008 年 Mann 以质谱技术进步突破 Nature，2020 年再次以质谱技术进步登上 Nature，其中我都没看到太多生物故事。未来十年中国蛋白质组学肯定有很多巧妙的应用成果，期待在相关技术方面也有深刻的创新，尤其是质谱技术。

- **对文章突破的信心**

文章是学术界的硬通货。在主流期刊持续发表文章，不时尝试突破 CNS 期刊，是学术生活的常态。对于学生如此，对于老师亦如此；突破 CNS 期刊之前如此，突破之后亦如此。

1995 年的我尚不知 Science 为何物，整个清华都觉得发表一篇英文文章不容易，但是今天的清华大学，写文章已经不再是个问题，Nature 封面都不在话下。即使咱们 CNCP 这个小生境，突破 CNS 子刊已是新常态，突破 CNS 正刊这两年也已经开始。

写文章多了，开始有不满，也有不满足。对于前述包云岗《伯克利科研模式的启发》的观点，计算机学会的同行讨论热烈，有篇评论《科研模式与评估体制》，推荐大家看看。我觉得，突破 CNS 期刊显然不能作为科研的终极目标，但是作为科研的一个阶段目标有其合

理性。作为对比，山东理工大学毕玉遂成功研制聚氨酯发泡技术，转让费 5.2 亿，当然是件利国利民利己的大好事，但是足足花费 15 年，作为终极目标尚可，作为阶段目标就不现实了。

2018 年 CNHUPO 会后，我专程拜访肖传乐，祝贺他突破 NM 和 MC，传乐有个观点说得非常好：工作量要够。的确，创新有没有、大不大，见仁见智，但是工作量大不大，一目了然。当年梦秋指导杨兵和我的学生研制 pLink 时，为了给软件提供足够可信的训练数据，选择 38 条肽段逐一合成、两两交联、手工上样、每组一针、一母三谱，上千次实验形成了 741 组交联肽段的质谱数据集，这是迄今为止合成肽段交联质谱最大的实验数据集，比十年后的 2020 年 NC 最新发表的类似数据集规模还要大。杨兵做这个实验“霸占”了两个月质谱机时，正值春节，而且正值孩子出生，他却必须在实验室坚守，真是不容易。铭琪/文锋 2017 年冲击 NC 时鉴定结果破万，迟浩 2018 冲击 NBT 时对比了八个引擎、六个数据集，也是一下子就把我震了——这么大的工作量，至少说明作者对自己的作品有信心！

冲击 CNS 期刊，需要一点运气。只要不断尝试，运气就会越来越大。但是即使冲顶成功，也不能感觉万事大吉——下一篇同样难写。这两年我阅读文献稍微多一些，发现很多著名团队的文章很有新意，工作量也很大，写作也很清晰，但是并不在 CNS 期刊发表，我猜测是尝试过但是没成功。所以我觉得这些同行的“板凳深度”不可小觑，咱们不要被一点好运气所迷惑。今年 5 月我在 CNCP 论坛转发了我新招的国科大生物系学生李金洋的《CMU 访学总结》，后来我看到一篇文章《美国教育同样残酷，我们也真的不如美国人勤奋》，就此文征询金洋同学的意见，他的基本观点是：“我觉得应该说中国的顶尖学生没有美国的顶尖学生勤奋。”中美竞争是未来十年的重头戏，顶尖人才的比拼可能更具决定性。

- **对软件突破的信心**

钱小红教授在 CNCP-2014 的晚宴上有一句话令我印象深刻：“没有生物信息学，就没有话语权！”而没有软件，就没有生物信息学。

Mann 和 Cox 合作过的文章中，2008 年发表的酵母 SILAC 定量蛋白质组分析的 Nature 文章目前 Web of Science 统计引用为 660+，而同年发表的 MaxQuant 软件的 NBT 文章目前引用 5,900+，是前者的 9 倍！其 2011 年发表的 Andromeda 软件的 JPR 文章居然也高达 2,300！

我非常好奇的是，MaxQuant 软件的设计开发，除了 Cox，到底还有多少人参与？我听说过另外两个软件的相关信息：Mascot 软件，整个公司只有 11 个人；UCSF 的 Protein Prospector 软件，2010 年长春会议的火车上听老关说，其实只有一个人在编写和维护，但是我发现这个软件功能也挺丰富，比如在常规鉴定之外，在交联鉴定领域也占有一席之地。

相比之下，pFind 系列全部软件注册下载量还没有超过 MaxQuant 的引用量；就 pFind 引擎而言，应该比 Andromeda 强很多，但是就定性、定量综合功能看，距离 MaxQuant 差距还不小。迟浩的 NBT 文章发表后，我破例在个人微信朋友圈发布消息，其中葛峰评论道：“蛋白质组学领域有了中国芯，我等俱有荣焉！”朋友殷殷之情令我感动。但是正如我在《pFind 团队的追求与道路》所言，没有 CPU 不行，只有 CPU 也不够，未来十年内 pFind 是否能赶超 MaxQuant，或者至少在国内、在 CNCP 范围内成为首选软件，对迟浩是个很大的挑战。

- **对学术生涯的信心**

常乘 2018-2-16 在 CNCP 论坛发过一贴：

“另外想说一下自己的困惑：作为一名做计算方法研究的‘陆军’，相关生物学知识的欠缺，不仅让我感觉很对不住自己拿的生物学学历，而且愈发觉得举步维艰。回过头来再看薛宇老师当年的博文（http://blog.sciencenet.cn/blog-404304-834869.html），心里难免有些慌。因此，自己在恶补生化知识的同时，也想求教包括贺老师在内的各位前辈、老师，当初是否有我这种感受？如有，怎么度过的呢？”

我这里正式做个回复：

先读一读《师从天才》。比如第 67 页。而这本书最令我印象深刻的是这样一句话：安心做个小人物。在你这个阶段，不用太在意薛宇的博文，与其关注自己的定位而心慌，不如关注具体的研究问题更心安。走自己的路，让薛宇心慌去吧:-)。

如果你身为信息学工作者，觉得恶补生化知识很艰难，那么假设你变身为一个生物学工作者，会觉得恶补计算技术更轻松吗？你觉得相关生物学知识欠缺而恶补，你觉得质谱技术不欠缺、不用恶补吗？在高通量、高灵敏、高特异的质谱技术冲击下，生物学的研究方式正在经历一场革命，有多少传统技术（什么东西南北 Blotting、抗原抗体 ELISA）还能保持生命力？生物学知识，与质谱技术，哪一个更稀缺？哪一个更简单？成为一个全栈科学家当然好，但是必然也只能追求一专多能，而非样样精通，那么依照你自己的背景，你的“一专”选定在哪里？我觉得从 Mann 的作品看，他不像是个生物学家，而是个非常好的质谱学家、蛋白质组学家；由于他在质谱蛋白质组学上的竞争优势，生物学家愿意和他合作解决问题。

心慌是难免的，而且不仅是现在，未来每一个十年都有心慌的时刻。一个学者的成长自然分成几个阶段：(1) 博士。会为能不能发表一两篇主流期刊文章达到毕业要求而心慌。(2) 博士后。会为能不能发表一篇 CNS 期刊文章或者 N 篇主流期刊文章而心慌。(3) PI，比如迟浩、刘超、肖传乐、钟传奇，开始为经费、团队心慌。(4) 资深 PI，比如付岩进入第二个十年，我则进入最后一个十年，开始为一辈子科研到底做出什么亮点而心慌，为除了写文章还会干什么而心慌。学术生涯是个马拉松，固然不能跑得太慢，但是也没必要在每一个阶段都领跑，最关键的是要一直在赛道上奔跑，因为学术马拉松没有终点，只要还在跑就永远有机会。过去十年看到不少在前两个阶段、甚至前三个阶段都非常优秀的青年学者因各种原因退出赛道，令我深为惋惜。学术生涯先苦后甜，虽不能大富大贵，但是疫情之下也不会因担忧失业而心慌。

我猜想，如果你现在有一篇 CNS 期刊文章在手，心慌会缓解很多。但是你博士毕业五年，已经参与写作 32 篇文章，而我博士毕业五年时一篇国际期刊文章都没有，博士毕业十年后才有第一篇国际期刊文章发表。你博士毕业五年，还在继续博士研究大方向，而我博士毕业后完全转向，毕业十年后再次转向，简直如过山车般地心慌，现在不也挺过来了嘛！我有时庆幸当年我的学术生涯那么不顺利，否则的话我会变得很嚣张，后果也许更严重。

段奕宏说：心慌总比安逸好。毛浩然说：心慌使人更进步。我说：拥抱质谱技术。

- **对自己的信心**

我，也是一代人的代表。我和徐平、梦秋作为 CNCP 的主办者，都已经跨过人生 50 的门槛，原则上讲，十年后将退休。不敢说老，却不再年轻，精力、活力、动力大不如前，也有个信心的问题。未来十年，有所不为才能有所为，具体想到三件事。

第一件事，研究一点具体问题。

我计划把已经开展研究的两个一般性问题继续做下去，争取断其一指：

（1）深度解析：2018 年迟浩的 Open-pFind 突破 NBT 之后，常规单肽质谱图解析率已经比较高了，但是 pLink、pGlyco、pTop 等的谱图解析率依然很低，求解也肯定更为困难，值得一试。我努力把 pLink 的深度解析做好。

（2）精准鉴定：常规单肽质谱图的所谓高解析率是相对 TDA-FDR ≤ 1%的判据而言，而这个判据所依据的 1:1 假设并非牛顿定律那么令人放心；即使 1:1 假设、1%错误均可信，那么定位这 1%的错误也是非常必要的。我努力把 pFind、pLink 的精准鉴定做好。

杨靖有个微信评论很有意思：社会公平这事吧，就像蛋白质组搜库结果，个体的悲惨都被“整体上还行”给稀释了。如果未来十年能把 pFind、pLink 的谱图解析率和鉴定准确率问题解决好，改善质谱鉴定中的“社会公平”，我退休也就心安了。

第二件事，把 CNCP 办好。

不求有多大创新，但求保持水准。创新总是激动人心的，保持则是平淡无奇的，但是往往最难做好的就是平淡无奇之事。ABRF 的 sPRG、iPRG 评测活动时断时续，而 RECOMB-CP 则在 2006、2010、2011、2012 举办之后停办，可见保持之难。

也许会适当尝试一点 CNCP 国际化，比如某一届增加一天日程邀请国际同行做一次学术交流，比如技术评测向国际同行也开放。

第三件事，促成《计算蛋白质组学》成书。

pFind 组大概是国际上研制各类蛋白质鉴定引擎最多的团队，写专著有一定条件。此外，现在迫切感到计算蛋白质组学人才培养效率不高，如果有个教材，学生进入领域快、出成果早，兴趣大、信心足，也许更愿意留在领域发展。

结语

2011 年底，我在 pFind 组内分享过黄炎培与毛泽东的窑洞夜话。1945 年黄炎培在延安同毛泽东谈话时讲到：

“我生六十多年，耳闻的不说，所亲眼看到的，真所谓‘其兴也勃焉，其亡也忽焉’。一人，一家，一团体，一地方，乃至一国，不少单位都没有能跳出这周期律的支配力。大凡初时聚精会神，没有一事不用心，没有一人不卖力，也许那时艰难困苦，只有从万死中觅取一生。既而环境渐渐好转了，精神也就渐渐放下了。有的因为历时长久，自然地惰性发作，由少数演为多数，到风气养成，虽有大力，无法扭转，并且无法补救。也有为了区域一步步扩大了，它的扩大，有的出于自然发展，有的为功业欲所驱使，强求发展，到干部人才渐见竭蹶、艰于应付的时候，环境倒越加复杂起来了，控制力不免趋于薄弱了。一部历史，‘政怠宦成’的也有，‘人亡政息’的也有，‘求荣取辱’的也有，总之没有能跳出这周期律。”

黄炎培的话，我总结为一句，未来十年，与诸君共勉：<u>一时之功，不足一世之用</u>。