首届中国计算蛋白质组学研讨会

The First China Workshop on Computational Proteomics (CNCP2010)

程序册

Advanced Program

2010 年 11 月 10 日-11 日

中国科学院计算技术研究所 北京

目录

会议基本信息

会议题目和网址

首届中国计算蛋白质组学研讨会

The First China Workshop on Computational Proteomics (CNCP2010)

http://cncp2010.ict.ac.cn/

会议地点

中国科学院计算技术研究所（简称中科院计算所）

地址：北京海淀区中关村科学院南路 6 号，邮编：100190

网址：www.ict.ac.cn，联系电话：010-62600114

会议语言

中文

会场分布

注册地点：中科院计算所一楼大厅

培训注册：11 月 8 日 8:30-9:00；会议注册：11 月 10 日 8:30-9:00

说明：参加培训的人员在 11 月 8 日注册；只参加会议的人员在 11 月 10 日注册

培训费：学生 500 元（需学生证），其他人员 800 元，只接受现金

培训教室：中科院计算所四楼报告厅

时间：11 月 8 日和 9 日

主会场：中科院计算所一楼多功能报告厅

时间：11 月 10 日和 11 日

宾馆分布

燕山大酒店，地址：北京市海淀区中关村大街甲 38 号

网址：http://www.yanshanhotel.com/，电话：010-62563388

说明：距离会场约 15 分钟车程（10 日和 11 日早上和晚上有专车接送）

乘车地点：燕山大酒店和计算所门口

乘车时间：10 日 8:10 从燕山大酒店发车，11 日 8:30 从燕山大酒店发车

晚餐后通知具体返回时间

天创宾馆，地址：北京市海淀区中关村南一条甲 1 号

网址：http://www.tianchuanghotel.com.cn，电话：010-51192000

说明：到会场步行约 5 分钟

线路及交通

说明：以下时间不包括堵车的时间——地铁除外

首都机场——燕山大酒店

乘坐的士：大约 32 公里，1 小时，100 多元

乘坐地铁：机场快轨（首都机场-三元桥）、地铁 10 号线（三元桥-海淀黄庄）、地铁 4 号线（海淀黄庄-人民大学），从人民大学 A2 出口步行 390 米可到（图 1）

总时间 1.5 小时，机场快轨 25 元，地铁 2 元（地铁换乘不需再购票）

首都机场——天创宾馆或中科院计算所

乘坐的士：大约 30 公里，50 多分钟，90 多元

乘坐地铁：机场快轨（首都机场-三元桥）、地铁 10 号线（三元桥-知春里），从知春里 A 出口步行 780 米到天创宾馆，再步行 380 米到中科院计算所（图 2）

总时间约 1.5 小时，机场快轨 25 元，地铁 2 元

机场大巴：从首都机场坐机场大巴 5 线到中关村终点站中关村四桥（图 3），下车后步行 580 米到中科院计算所，再步行 380 米到天创宾馆（图 4）

机场大巴有 50 多分钟路程，16 元

北京站——燕山大酒店

乘坐的士：大约 16 公里，40 多分钟，30 多元

乘坐地铁：地铁 2 号线（北京站-宣武门）、地铁 4 号线（宣武门-人民大学），从人民大学 A2 出口步行 390 米可到（图 1）

总时间约 44 分钟，地铁 2 元（地铁换乘不需再购票）

北京站——天创宾馆或中科院计算所

乘坐的士：大约 16 公里，40 多分钟，30 多元

乘坐地铁：地铁 2 号线（北京站-雍和宫）、地铁 5 号线（雍和宫-惠新西街南口）、地铁 10 号线（惠新西街南口-知春里），从知春里 A 出口步行 780 米到天创宾馆，再步行 380 米到中科院计算所（图 2）

总时间约 1 小时，地铁 2 元（地铁换乘不需再购票）

北京西站——燕山大酒店

乘坐的士：大约 10 公里，20 多分钟，20 多元

北京西站——天创宾馆或中科院计算所

乘坐的士：大约 12 公里，30 多分钟，30 多元

乘坐公交：从北京西站坐 319 路到海淀交通支队站下，步行 410 米到天创宾馆，再步行 380 米到中科院计算所（图 5）

319 路公交路程大约 14 公里，1 小时，1.5 元

北京南站——燕山大酒店

乘坐的士：大约 18 公里，40 多分钟，40 多元

乘坐地铁：地铁 4 号线（北京南站-人民大学），从人民大学 A2 出口步行 390 米可到（图 1）

总时间约 40 分钟，地铁 2 元

北京南站——天创宾馆或中科院计算所

乘坐的士：大约 20 公里，50 多分钟，50 多元

乘坐地铁：地铁 4 号线（北京南站-海淀黄庄）、地铁 10 号线（海淀黄庄-知春里），从知春里 A 出口步行 780 米到天创宾馆，再步行 380 米到中科院计算所（图 2）

总时间约 50 分钟，地铁 2 元（地铁换乘不需再购票）

图 1：从人民大学 A2 出口步行到燕山大酒店

图 2：从知春里 A 出口步行到天创宾馆或中科院计算所

图 3：从首都机场坐机场大巴 5 线到中关村终点站中关村四桥

图 4：从中关村四桥步行到中科院计算所或天创宾馆

图 5：从海淀交通支队公交站到天创宾馆或中科院计算所

图 6：从中科院计算所步行到融科 C 座

图 7：会议地点、宾馆及周边交通

图 8：北京市地铁线路(部分)

就餐地点和时间

早餐：在宾馆餐厅用餐，用餐时间 7:00-7:30

午餐：融科 C 座地下一层金白领（图 6），用餐时间 12:30-13:00

晚餐：融科 C 座地下一层金白领（图 6），用餐时间 18:00-18:30

说明：**到金白领就餐的人员需凭代表证**

联系方式

会议网站： http://cncp2010.ict.ac.cn

邮件： cncp2010@ict.ac.cn

电话： 010-62601352 任菲 刘玉东

会务组织： 中国科学院计算技术研究所 pFind 研发组

会务组织负责人：贺思敏 139-1076-5853

孙瑞祥 186-0003-8430

培训与会议注册：付 岩 136-8333-5378

燕山大酒店接待：孙瑞祥 186-0003-8430

天创宾馆接待： 迟 浩 138-1065-7903

培训与会议资料：袁作飞 138-1182-5909

培训与会议会场：刘 超 186-1119-0741

会议网站与邮件：张 昆 135-8156-9858

会议就餐： 樊盛博 158-1127-4880

特别提醒

由于大会报告时间紧张，敬请所有报告专家务必提前一日将报告幻灯拷贝至会务组，会务组提供会场放映电脑，不接受个人电脑接入。大会报告时间为 30 分钟（含 5 分钟提问）。非常感谢您的配合！

近期北京早晚温度较低，请您注意加衣保暖。

图 1：从人民大学 A2 出口步行到燕山大酒店

图 2：从知春里 A 出口步行到天创宾馆或中科院计算所

图 3：从首都机场坐机场大巴 5 线到中关村终点站中关村四桥

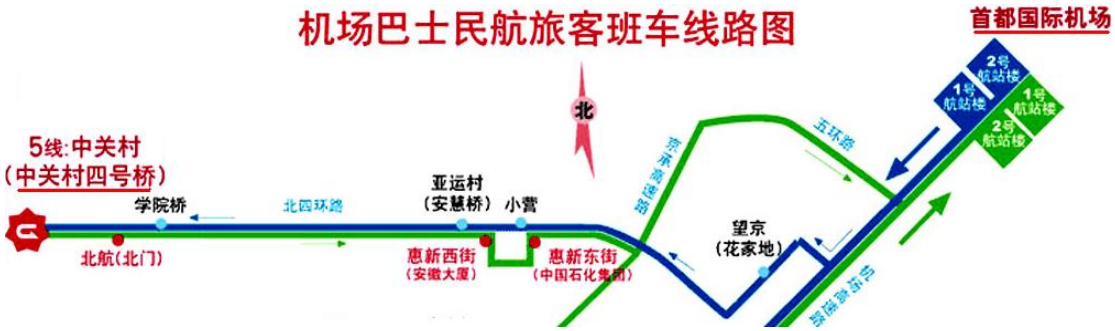


图 4：从中关村四桥步行到中科院计算所或天创宾馆

图 5：从海淀交通支队公交站到天创宾馆或中科院计算所

图 6：从中科院计算所步行到融科 C 座

图 7：会议地点、宾馆及周边交通

图 8：北京市地铁线路(部分)

会议总体日程

日期	时间	内容	地点
2010-11-8 **星期一**	8:30-9:00	注册，缴纳培训费	计算所一楼大厅
	9:00-12:10	质谱培训：质谱技术和蛋白质组学	计算所四楼报告厅
	12:10-13:30	午餐	融科 C 座地下一层
	13:30-18:00	质谱培训：数据库搜索和翻译后修饰	计算所四楼报告厅
	18:00-19:30	晚餐	融科 C 座地下一层
2010-11-9 **星期二**	9:00-12:10	质谱培训：肽从头测序	计算所四楼报告厅
	12:10-13:30	午餐	融科 C 座地下一层
	13:30-18:00	质谱培训：肽/蛋白质定量	计算所四楼报告厅
	18:00-19:30	晚餐	融科 C 座地下一层
2010-11-10 **星期三**	8:30-9:00	签到注册（参加培训的不需再注册）	计算所一楼大厅
	9:00-9:10	首届中国计算蛋白质组学研讨会简介	计算所一楼报告厅
	9:10-9:20	欢迎词	计算所一楼报告厅
	9:20-9:40	合影	计算所一楼大厅
	9:40-11:10	大会报告（第 1 至 3 个）	计算所一楼报告厅
	11:10-11:20	休息	
	11:20-12:20	大会报告（第 4 至 5 个）	计算所一楼报告厅
	12:20-13:30	午餐	融科 C 座地下一层
	13:30-15:30	大会报告（第 6 至 9 个）	计算所一楼报告厅
	15:30-15:50	休息	
	15:50-17:50	大会报告（第 10 至 13 个）	计算所一楼报告厅
	17:50-19:30	专家招待会	
2010-11-11 **星期四**	8:30-9:00	签到注册（未注册的人员）	计算所一楼大厅
	9:00-11:00	大会报告（第 14 至 17 个）	计算所一楼报告厅
	11:00-11:20	休息	
	11:20-12:20	大会报告（第 18 至 19 个）	计算所一楼报告厅
	12:20-13:30	午餐	融科 C 座地下一层
	13:30-15:30	大会报告（第 20 至 23 个）	计算所一楼报告厅
	15:30-15:50	休息	
	15:50-17:20	大会报告（第 24 至 26 个）	计算所一楼报告厅
	17:20-17:30	会议总结	计算所一楼报告厅
	17:30-19:00	晚餐	

质谱培训课程日程

2010 年 11 月 8 日星期一: 质谱技术与蛋白质组学基础培训(一)* Monday, November 8, 2010: Training on the fundamentals of mass spectrometry and proteomics I 地点: 中科院计算所四楼报告厅			
时间 Time	**培训内容 Topic**	**报告人 Tutor**	**备注 Note**
8:30-9:00	注册, 缴纳培训费	(位置在计算所一楼大厅内)	学生 500 元, 其他人员 800 元
9:00-10:30	1.质谱技术基础 Introduction to mass spectrometry	关慎恒 Guan Shenheng	
10:30-10:40	休息 Break		
10:40-12:00	2.蛋白质组学基础 Fundamentals of proteomics	关慎恒 Guan Shenheng	
12:00-12:10	问题与讨论 Questions and discussions		
12:10-1:30	午餐 Lunch		
1:30-3:00	3.蛋白质序列数据库搜索 Protein sequence database searching	关慎恒 Guan Shenheng	
3:00-3:10	休息 Break		
3:10-4:30	4.翻译后修饰 Post-translational modifications	关慎恒 Guan Shenheng	
4:30-6:00	用 pFind 软件实地分析质谱数据 Using pFind studio software	袁作飞 Yuan Zuofei	学员请带上笔记本电脑
6:00-7:30	晚餐 Supper		

2010 年 11 月 9 日星期二: 质谱技术与蛋白质组学基础培训(二)*

Tuesday, November 9, 2010:
Training on the fundamentals of mass spectrometry and proteomics II
地点: 中科院计算所四楼报告厅

时间 Time	培训内容 Topic	报告人 Tutor	备注 Note
9:00-10:30	5. 肽从头测序/自顶向下蛋白质组学 De novo peptide sequencing/Topdown proteomics	关慎恒 Guan Shenheng	
10:30-10:40	休息 Break		
10:40-12:00	6. 肽从头测序:算法与软件 De novo peptide sequencing algorithms and software	迟浩 Chi Hao	学员请带上笔记本电脑
12:00-12:10	问题与讨论 Questions and discussions		
12:10-1:30	午餐 Lunch		
1:30-3:30	7.肽/蛋白质定量 Peptide/protein quantification	关慎恒 Guan Shenheng	
3:30-3:40	休息 Break		
3:40-5:40	8.蛋白质定量:算法与软件 Protein quantification: algorithms and software	刘超 Liu Chao	学员请带上笔记本电脑
5:40-6:00	问题与讨论 Questions and discussions		
6:00-7:30	晚餐 Supper		

*: 注册时只接受现金形式的培训费, 凭代表证入场和就餐.

会议报告日程

2010 年 11 月 10 日星期三上午: 大会邀请报告(一) Wednesday, November 10, 2010: Invited talks 地点: 中科院计算所一楼多功能报告厅 主持人: 王通 应万涛				
时间 Time	**报告题目 Title**	**报告人 Speaker**	**报告人单位 Institution**	**报告摘要页码 Abstract Page**
8:30-9:00	签到注册	参加培训的不需注册	(不收注册费)	
9:00-9:10	首届中国计算蛋白质组学研讨会简介 Brief introduction to CNCP2010	贺思敏	中科院计算所	
9:10-9:20	欢迎词 Opening Ceremony	所领导	中科院计算所	
9:20-9:40	合影 Photo	全体	(计算所一楼大厅)	
9:40-10:10	糖蛋白结构的质谱数据库	杨芃原	复旦大学	19
10:10-10:40	核心岩藻糖化蛋白质特异性发掘的系统解决方案 Establishment of a systematic method coupling consecutive MS^n and software tools for charactering core-fucosylated glycoproteins	应万涛	北京蛋白质组研究中心	20
10:40-11:10	利用串联质谱技术解析多糖结构 Glycan Structure Sequencing with Tandem Mass Spectrometry	张凯中	加拿大西安大略大学	21
11:10-11:20	休息 Break			
11:20-11:50	解码细胞迁移过程中的信号通路网络 Deciphering the Signaling Network in the Leading Edge of the Migrating Cells	汪迎春	中科院遗传与发育生物学研究所	22
11:50-12:20	信号通路分析辅助的功能蛋白质组学研究策略 Pathway analysis-assisted study strategy in functional proteomics	王通	暨南大学	23
12:20-13:30	午餐 Lunch	全体		

2010 年 11 月 10 日星期三下午: 大会邀请报告(二)
Wednesday, November 10, 2010: Invited talks
地点: 中科院计算所一楼多功能报告厅
主持人: 谢鹭 陆豪杰

时间 Time	报告题目 Title	报告人 Speaker	报告人单位 Institution	报告摘要页码 Abstract Page
1:30-2:00	利用稳定同位素代谢标记研究哺乳动物动态蛋白质组的数据处理平台 A data processing platform for mammalian proteome dynamics studies using stable isotope metabolic labeling	关慎恒	美国加州大学旧金山分校	24
2:00-2:30	大规模 SILAC 标记定量蛋白质组学研究中的数据分析 Data analysis in large scale quantitative proteomics study with SILAC approach	徐平	北京蛋白质组研究中心	25
2:30-3:00	体内终端氨基酸标记在定量蛋白质组学中的应用 In vivo termini amino acid labeling for quantitative proteomics	陆豪杰	复旦大学	26
3:00-3:30	利用基于肽段计数的无标记定量技术揭示线粒体蛋白质组的功能特性 Quantitative Analysis of Mitochondrial Proteomes using Normalized Spectral Abundance Factor	邓宁	浙江大学	27
3:30-3:50	休息 Break			
3:50-4:20	尿液蛋白质疾病标志物数据库 The urinary protein biomarker database	邵晨	中国协和医科大学	28
4:20-4:50	基于质谱数据发现小鼠基因组新蛋白质编码区域 The discovery of novel protein-coding features in mouse genome based on mass spectrometry data	谢鹭	上海生物信息中心	29
4:50-5:20	从新一代测序技术的组学到基于质谱仪的蛋白质组学 -- 华大基因的生物信息学 From NGS Genomics to MS-based Proteomics -- BGI's bioinformatics activities	张勇	深圳华大基因研究院	30
5:20-5:50	腾冲嗜热菌的多温度条件下的蛋白质组基因组学研究	赵屹	中科院计算所	31
5:50-7:30	宴会 Banquet	邀请专家		

2010 年 11 月 11 日星期四上午: 大会邀请报告(三) Thursday, November 11, 2010: Invited talks 地点: 中科院计算所一楼多功能报告厅 主持人: 邹汉法 孙瑞祥				
时间 Time	**报告题目 Title**	**报告人 Speaker**	**报告人单位 Institution**	**报告摘要页码 Abstract Page**
8:30-9:00	签到注册	未注册的人员	(不收注册费)	
9:00-9:30	基于 HCD 谱图的肽段从头测序 De novo Sequencing of Peptides Using HCD Spectra	董梦秋	北京生命科学研究所	32
9:30-10:00	从未知基因组到可测定的蛋白质组：通过从头测序来研究依赖于 pH 值的 N10 细菌蛋白质组 From an unknown genome to a measurable proteome: Studying on the pH-dependent proteomes in N10 bacteria by denovo sequencing	王全会	中科院北京基因组研究所	33
10:00-10:30	利用质谱和同源数据库进行全蛋白测序 Complete Protein Sequencing with MS/MS and a Homologous Database	马斌	加拿大滑铁卢大学	34
10:30-11:00	电子转运裂解质谱:特征发现与鉴定应用 Electron Transfer Dissociation: Characterization and Applications in Protein Identification	孙瑞祥	中科院计算所	35
11:00-11:20	休息 Break			
11:20-11:50	基于质谱的蛋白质组学数据处理新方法和平台发展 Development of Methods and Platform for Data Processes in Mass Spectrometry Based Proteome Research	邹汉法	大连化学物理研究所	36
11:50-12:20	基于优化的肽质量指纹谱方法鉴定蛋白质混合物 Optimization-Based Peptide Mass Fingerprinting for Protein Mixture Identification	余维川	香港科技大学	37
12:20-13:30	午餐 Lunch	全体		

2010 年 11 月 11 日星期四下午: 大会邀请报告(四)
Thursday, November 11, 2010: Invited talks
地点: 中科院计算所一楼多功能报告厅
主持人: 张红雨 付岩

时间 Time	报告题目 Title	报告人 Speaker	报告人单位 Institution	报告摘要页码 Abstract Page
1:30-2:00	基于相关谱图对的非限制性修饰检测 Unrestrictive modification detection based on related spectral pairs	付岩	中科院计算所	38
2:00-2:30	评价诱饵库设计，搜索策略，匹配误差和质量控制对鸟枪法蛋白质组学中肽段鉴定精确性的影响 Evaluation of the effect of decoy design, search strategy, mass tolerance and quality control method on the accuracy of peptide identifications in shotgun proteomics	朱云平	北京蛋白质组研究中心	39
2:30-3:00	BuildSummary : 一个基于目标-诱饵策略的蛋白质鉴定整合软件 BuildSummary: A software tool for assembling protein	盛泉虎	上海生命科学研究院	40
3:00-3:30	冷冻电镜中的计算方法：图像数据处理和三维重构 Computational methods in cryo-electron microscopy: image data processing and 3D structure reconstruction	张法	中科院计算所	41
3:30-3:50	休息 Break			
3:50-4:20	DomainRBF: 一种对疾病相关蛋白质结构域进行优先排序的贝叶斯回归方法 DomainRBF: a Bayesian regression approach to the prioritization of associations between protein domains and human complex diseases	江瑞	清华大学	42
4:20-4:50	蛋白质结构“字母表”设计 Designing Succinct Structural Alphabets	卜东波	中科院计算所	43
4:50-5:20	蛋白质作为分子化石 Proteins as molecular fossils	张红雨	华中农业大学	44
5:20-5:30	会议总结	杨芃原	复旦大学	
5:30-7:00	晚餐 Supper			

会议报告摘要

报告题目：糖蛋白结构的质谱数据库

杨芃原

复旦大学

摘要：糖蛋白的结构分析需要借助于糖链结构数据库， 但由于糖链结构的复杂性， 构建一个理论的糖链结构数据库似乎是不太可行的方法。因此，通过各种技术构建专业性强、针对性明显的糖链结构数据库已经引起了关注。 我们的研究基于生物质谱的数据分析， 建立了蛋白质糖基化位点以及糖链结构数据库。通过自己开发的一种对比实验方法，先鉴定去糖链结构的糖基化肽段序列，建成理论糖肽库，为后期在实际样品中的快速鉴定糖肽组成提供数据基础。 开发了一套糖蛋白鉴定和糖链结构确立的理论算法，并将理论算法在我们创建的软件 GRIP: Glycopeptide Reveal & Interpretation Platform 中全部实现。 实际的分析表明， 我们的新方法可以有效的进行通量化的糖蛋白结构质谱分析， 展现了比较好的应用前景。

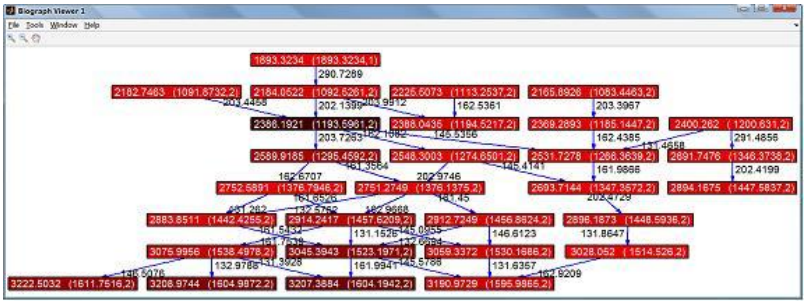


Fig.1 Topo-strucure of GRIP for exp MS tree

报告题目：核心岩藻糖化蛋白质特异性发掘的系统解决方案

Establishment of a systematic method coupling consecutive MS^n and software tools for charactering core-fucosylated glycoproteins

应万涛

北京蛋白质组研究中心

钱小红

北京蛋白质组研究中心

Abstract: Core-fucosylated glycoproteins have been known to involve a variety of biological and pathological processes, especially in hepatocellular carcinoma, TGF-β1 and EGF signaling pathways, pancreatic cancer and so on. Current methods, such as lectin blotting, could map core-fucosylated glycoproteins. However, when dealing with the glycosite information, it is hard to discriminate core-fucosylated ones from general glycosites. A more efficient and targeted approach is needed to screen core-fucosylated glycoproteins. Here, we simplified carbohydrate parts of three core-fucosylated glycopeptides, following with LTQ-FT analysis under MS2 and MS3 mode. Several special fragments signal appear to be particular and helpful to identify core fucosylated N-linked glycopeptides, which may be useful in its large scale identification by mass spectrometer and searching engine directly. The most obvious characteristic in the MS2 spectra of CID model in ion trap was that the intensity of the highest fragment peak was about ten times higher than that of the second one. These peaks were attributed to mass of the peptide parent ion mass and a GlcNAc residue, which resulted from neutral loss of the fucose unit. In a consecutive MS3 analysis, a neutral loss of GlcNAc appears to be the most strong peak (about two times higher than other peaks), and the peaks' intensity in MS3 distribute much more even than that in MS2. Standing on these characteristics summarized from the three glycopeptides' data, we are developing software tools to identify core fucosylated glycopeptides automatically, and design a strategy to determine core-fucosylated glycoproteins from normal and cancer individual. This strategy will be aim to establish a high confident core fucosylated glycopeptides database for biomarker discovery.

报告题目：利用串联质谱技术解析多糖结构

Glycan Structure Sequencing with Tandem Mass Spectrometry

张凯中

加拿大西安大略大学

Abstract: Glycosylation which adds carbohydrate moieties to a protein is one of the important post-translational modifications. The carbohydrates in glycoproteins are commonly referred as glycans which are assembled from simple sugars. Since each sugar can have up to five linkage sites, glycans are assembled in a tree-like structure. The structural variation in glycans is fundamental to their biological activity. The glycan structure sequencing problem is to determine the tree-like structure. We study glycan de novo sequencing with tandem mass spectrometry. We show results concerning complexity and algorithm of the problem. Some experimental results based on a heuristic algorithm will also be presented.

报告题目：解码细胞迁移过程中的信号通路网络

Deciphering the Signaling Network in the Leading Edge of the Migrating Cells

汪迎春

中国科学院遗传与发育生物学研究所

Abstract:Cell polarization to form a dominant pseudopodium in the leading edge is necessary for sustained directional cell migration, a process that plays a central role in many important physiological and pathological pathways including wound healing and cancer metastasis. However, the mechanism underlying polarization of migratory cells is still unclear because it has not been possible to biochemically isolate the front and the back compartments for large-scale protein analysis. Here we use microporous filters to differentially isolate the pseudopodium and cell body compartments for large-scale (> 3000 proteins) quantitative proteomic analysis. We found that proteins involved in the regulation of actin cytoskeleton, focal adhesion dynamics, and MAPK signaling are highly enriched and/or differentially phosphorylated in the pseudopodium, whereas proteins involved in nuclear functions and metabolism are highly enriched and/or differentially phosphorylated in the cell body. Importantly, several hypothetical phosphoproteins were also differentially distributed and phosphorylated in polarized cells, including the novel tyrosine kinase (PEAK1, pseudopodial-enriched atypical kinase one). Functional analysis of PEAK1 revealed that PEAK1 colocalizes with actin cytoskeleton and focal adhesions, and is necessary for cell migration and tumor progression. Our findings provided novel insight into the spatial organization of complex signal transduction pathways that control directional cell migration and cancer cell metastasis and reveal PEAK1 is a new tyrosine kinase that regulates the cell migration and invasion machinery.

王　通

暨南大学

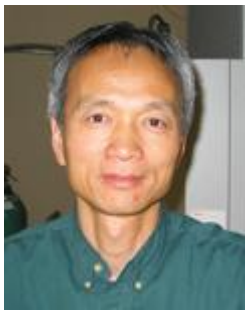
何庆瑜

暨南大学

报告题目：信号通路分析辅助的功能蛋白质组学研究策略

Pathway analysis-assisted study strategy in functional proteomics

Abstract: Major breakthroughs in proteomics, particularly in mass spectrometry and computational proteomics, are continuously increasing the capacity of identifying low abundance proteins as well as acquiring more relative quantification information. As a result, new ways of experimental designs and the amount of valid data are expanding at an extremely high rate. As one of the greatest challenges, functional interpretation of these data is a crucial link between proteomic profiling and biological probing, turning investigations from descriptive into mechanistic. Regarding this tough job, numerous pathway analysis tools, including both public and commercial ones, are available. Among them, Ingenuity Pathway Analysis (IPA) is one of the concentrated and specialized biological function mining tools. In this talk, IPA-assisted functional proteomic investigations performed by our group will be presented as examples. These include a proteomic modeling study of HIV infected astrocyte-microglia crosstalk, addressing the protection effects of astrocytes in containing HIV infection in central nervous system, as well as a lung cancer cell proteomic model, defining new endpoints of epithelial to mesenchymal transition (EMT). These studies employed the strategy of proteomic profiling--pathway analysis--biological validation, successfully identifying target pathways and nodes of regulation by computational pathway analysis, as well as directing downstream biological investigations and validating, in turn, the accuracy of pathway analysis.



报告题目：利用稳定同位素代谢标记研究哺乳动物动态蛋白质组的数据处理平台

A data processing platform for mammalian proteome dynamics studies using stable isotope metabolic labeling

关慎恒

加州大学旧金山分校

Abstract: Mammalian in vivo heavy metabolic labeling using heavy isotopes such as 15N allows for proteome scale investigations of protein dynamics in various tissues. These proteomic dynamics studies can provide a deeper understanding of healthy development and well-being of complex organisms, as well as the possible causes and progression of diseases. An essential and enabling component of these large scale investigations is a robust data processing platform, which is capable of reduction of a large set of LCMSMS raw data files into the desirable protein turnover rate constants. The data processing platform described here is the integration of a variety of software modules into a workflow. This software platform consists of some established software tools such as the database search engine and several novel data processing modules specifically developed for 15N metabolic labeling, such as (1) cross-extraction of 15N-containing ion intensities from raw data files of varying biosynthetic incorporation times, (2) computation of peptide 15N incorporation distributions, (3) aggregation of multiple peptide relative isotope abundance curves into a protein curve. In order to reduce the propagation errors in a long chain of the processing steps, processing parameter optimization and noise reduction procedures are performed in some necessary processing modules.

报告题目：大规模 SILAC 标记定量蛋白质组学研究中的数据分析

Data analysis in large scale quantitative proteomics study with SILAC approach

徐　平

北京蛋白质组研究中心

Abstract: Protein ubiquitination is a highly conserved modification that plays a central regulatory role in eukaryotic cells. Ubiquitin (Ub) is covalently attached to protein substrates by a cascade of enzymatic reactions involving activating enzymes (E1), conjugating enzymes (E2), and ligases (E3). These reactions generally form an isopeptide bond between the carboxyl group of the C terminus of Ub (G76) and the 3-amino group of a lysine residue within the substrates. All seven lysines in Ub contribute to the assembly of polyUb chains, thereby producing a variety of structures with diverse lengths and linkages. On the other hand, ubiquitinated substrates may be sorted into different pathways based on diverse polyUb structures. Canonical K48-linked polyUb chains are believed to be the principal signal for targeting substrates for degradation by the 26S proteasome, whereas K63-linked chains act in a range of other processes, including protein trafficking, DNA repair, and inflammation. However, the prevalence of other polyUb topologies and their roles in biological processes are not well known. Recent proteomics developments in mass spectrometry will facilitate functional study on ubiquitinated proteins. However large scale identification of heterogeneously ubiquitinated protein and quantification for low abundant proteins is still remaining challenge. Large-scale analyses of ubiquitinated proteins are usually performed by combining affinity purification strategies with mass spectrometry. Here, we provide a strategy for differentiating ubiquitinted proteins from co-purified unmodified components by reconstructing virtual Western blots for proteins analyzed by gel electrophoresis and mass spectrometry. Because protein ubiquitination, especially polyubiquitination, causes a dramatic shift of molecular weight, the difference between experimental and expected molecular weight was used to confirm the status of ubiquitination. Experimental molecular weight of putative yeast ubiquitin-conjugates was computed from the value and distribution of spectral counts in the gel using a Gaussian curve fitting approach. Unmodified proteins in yeast cell lysate were also analyzed as a control to assess the accuracy of the method. Multiple thresholds that incorporated the mass of ubiquitin and/or experimental variations were evaluated with respect to sensitivity and specificity. Furthermore, by profiling both the entire yeast proteome and ubiquitinated proteins in wild-type and ubiquitin K11R mutant strains using mass spectrometry, we identified K11 linkage-specific substrates. In order to study K11 chain function, we applied quantitative proteomics analysis through SILAC approach. Both biological and technical replicates have been utilized for evaluating the variation in large scale quantitative proteomics study. By combining all of these strategies, we found that Ubc6 primarily synthesizes K11-linked chains, and K11 linkages function in the ERAD pathway.

报告题目：体内终端氨基酸标记在定量蛋白质组学中的应用

In vivo termini amino acid labeling for quantitative proteomics

陆豪杰

复旦大学

摘要：定量蛋白质组学已经成为当今科学研究的一个热点和挑战，在多种定量分析策略中，细胞稳定同位素标记（SILAC）方法已被广泛而有效的应用于各种重要的生物学过程和医学问题的研究中。在这里，我们建立了一个新的定量策略---体内终端氨基酸标记方法，重精氨酸-6 和重赖氨酸-6 稳定同位素分别培养细胞，蛋白混合后经过 Lys-N 和 Arg-C 酶解，使肽段两端分别含有轻重精赖氨酸，这样在一级质谱中为同一质荷比，但在串级质谱中可以观察到成对的 b, y 碎片离子，通过直接比较对峰的强度可以得到肽段和蛋白的定量信息。相比传统的 SILAC，由于一级质谱中肽段母离子质量相同，这样降低了样品分析的复杂度，而且提高了一级质谱分析的灵敏度，同时克服了 iTRAQ 方法低端报告离子的抑制效应，可以在全质量范围内提供多对碎片离子的定量信息，为准确的肽段和蛋白定量提供了可靠的依据。

报告题目：利用基于肽段计数的无标记定量技术揭示线粒体蛋白质组的功能特性

Quantitative Analysis of Mitochondrial Proteomes using Normalized Spectral Abundance Factor

邓宁

浙江大学

Abstract: As double-membrane organelles in the cell, mitochondria are essential for cell metabolism, transport, biosynthesis, and signaling. Quantitative proteomics has been increasingly recognized as an essential proteomics tool to generate significant insight in biomedical research. In this study, we used the normalized spectral abundance factor (NSAF) to quantitatively analyze the mitochondrial proteome for human heart, mouse heat and mouse liver samples. In this approach, the spectral counts of a protein were divided by its length and normalized to the total sum of spectral counts / length in a given LC-MS/MS analysis.

A total of five human cardiac mitochondrial samples, eight murine cardiac mitochondrial samples and seven murine liver mitochondrial samples have been collected and submitted for the LC-MS/MS analyses to generate MS Spectra. The spectra were then searched against IPI databases using the SEQUEST algorithm and statistically validated by Scaffold software package. In-house software was developed to generate NSAF value for each identified protein for further quantitative analyses and comparisons.

The mitochondrial proteins identified from human heart, mouse heart and mouse liver were functionally clustered into several groups including metabolism, OXPHOS, transport, biosynthesis, signaling, apoptosis etc. As a result, proteins involved in electron transport chain (ETC) show highest abundances in all three mitochondrial proteomes, especially in the heart. On the other hand, metabolism related proteins and urea cycle proteins show more abundant in the liver.

In conclusion, our study provides a comprehensive proteomic way to quantitatively understand the mitochondrial biology and medicine.

报告题目：尿液蛋白质疾病标志物数据库

The urinary protein biomarker database

邵　晨

中国协和医科大学

摘要：尿液是重要的疾病标志物来源。目前的尿蛋白质组疾病标志物研究面临着两个问题：（1）当前技术手段的限制和正常人个体间较大的差异，降低了候选标志物的可信度；（2）由于实验时未对所有可能影响尿蛋白质组的疾病同时比较，很多候选标志物不具有疾病特异性。通过实验手段改善这两个问题必须大幅度增加工作量，而通过对已有研究结果进行数据整合和对比分析，可以以最小的代价获取最多的信息，指导后续的实验工作。本研究通过收集、整理目前已发表的蛋白质组研究文献，建立了尿液中的蛋白质标志物数据库。以互联网平台的形式展现出来，使得相关领域的研究者可以通过这个平台方便、快捷、全面地对自己的数据进行分析，研究泌尿系统的生理功能，或选取最有可能成为疾病标志物的蛋白质分子进行下一步的验证，从而降低实验耗费，提高研究效率。研究还通过对不同实验条件和手段、不同疾病类型的数据进行整合和对比分析，以及对建立的疾病—蛋白质标志物的网络图结构进行分析，对数据库收集的潜在疾病标志物的可信度和疾病特异性进行了评价。

高友鹤

中国协和医科大学

报告题目：基于质谱数据发现小鼠基因组新蛋白质编码区域

The discovery of novel protein-coding features in mouse genome based on mass spectrometry data

谢　鹭

上海生物信息中心

Abstract: Identifying protein-coding genes in eukaryotic genomes remains a challenge in post-genome era due to the complex gene models. Proteogenomics strategy has been employed in many eukaryotes to validate protein-coding genes directly on protein level in large scale and has provided valuable complementary information in genome annotation.We applied a proteogenomics method to detect un-annotated protein-coding regions in mouse genome. Two searchable proteomic datasets were constructed, one with all possible encoded exon junctions (EJCT dataset) for the discovery of novel exon splice events, and the other with all putative encoded exons (ORF dataset) for finding uninterrupted novel protein-coding regions. The two datasets were combined with a public full-length protein dataset (competitive dataset) respectively and queried against 496 high-accuracy tandem mass spectrometry (MS/MS) RAW files from diverse mouse samples. By setting a strict spectra false discovery rate threshold at close to 0, thirty two unique peptides (matching 149 spectra) from EJCT dataset were discovered which straddle novel exon junctions, and 104 unique peptides (matching 450 spectra) from ORF dataset were located in 99 unique protein-coding regions. Aligning backwards from these peptides to the parent genome could classify them into seven distinct new protein coding categories based on their genome loci relative to adjacent previously annotated genes or exons. Some of the novel peptides were under cross validation by RT-PCR. Our work discovered a number of novel mouse protein-coding regions directly at the translational product level, which implied that proteogenomics strategy could be performed to provide substantial evidences for genome annotation in encoded genes and help novel gene identifications.

报告题目： 从新一代测序技术的组学到基于质谱仪的蛋白质组学 -- 华大基因的生物信息学

From NGS Genomics to Ms-based Proteomics -- BGI's bioinformatics activities

张 勇

华大基因

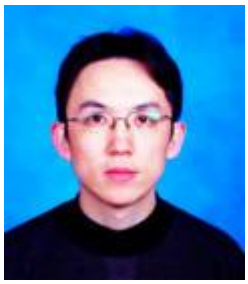
摘要： 1999 年 9 月 9 日，随着“国际人类基因组计划 1%项目”的正式启动，华大基因在北京正式成立。过去 11 年，先后完成了若干具有国际领先地位的大型基因组项目，在《Nature》和《Science》等国际一流的杂志上发表多篇论文。2007 年，华大基因主力南下深圳，基因组领域迎来第二代测序浪潮，新型测序仪使得基因组从物种水平上升到个体水平。一直以来，测序和生物信息学堪称华大基因的两大核心竞争力。对于海量数据的信息分析和挖掘成为华大基因立足世界基因组领域的根本。新一代测序技术被广泛用于 DNA, RNA 测序，针对动植物、人，华大基因与许多科学家合作，一起启动并完成了若干项目。

华大基因的研究思路是通过利用海量数据的信息学分析从而识别关键要素。这种工业化、规模化的科研思路充分发挥了高通量、低成本的仪器特性。除了测序仪，质谱仪无疑成为蛋白质组领域的高通量仪器。华大基因也逐步从 DNA, RNA 水平，向 Protein 水平研究发展，拟通过海量质谱数据，进行相关的生物信息挖掘，彻底从 DNA, RNA, Protein 水平解读生命的奥妙。

DNA 水平，目前主要侧重于短序列比对；基因组组装、注释；突变识别；拷贝数变异识别等。

RNA 水平，目前主要侧重于转录组研究；基因表达；基因选择性剪切；融合基因识别等。

Protein 水平，简单的峰图识别；定量蛋白质组；蛋白质鉴定与翻译后修饰鉴定。



报告题目：腾冲嗜热菌的多温度条件下的蛋白质组基因组学研究

赵　屹

中科院计算技术研究所

摘要：蛋白质组基因组学是近几年诞生的一门用蛋白质组数据信息，也是结合转录组数据来解析基因组的新兴学科。MS/MS质谱实验辅助基因组注释已经在多种物种中成功运用。这里我们整合腾冲嗜热菌的多温度条件下的蛋白质组质谱数据及转录组深度测序数据进行分析。结果显示：(1)转录组数据可以覆盖 70%以上的 2588 核酸库基因；(2) 74%以上的质谱鉴定肽段可以寻找到转录组证据。此外，通过基因的定量表达分析，我们比较了不同温度条件下基因转录表达的差别：初步得出在四种温度条件下都有表达的基因 359 条，而每个温度条件下都有自己特异表达的基因；结合两种独立来源的组学数据互相验证研究，发现其中有 80 多个不属于 2588 范围内的转录区，其中 2 个区域强烈支持质谱鉴定得到的新基因区域，21 个区域属于可能的新非编码 RNA。我们还对于腾冲嗜热菌基因组中预测的 2588 条基因进行了注释修正。这种系统的数据整合分析方法无疑对于生物学的研究，尤其是基因组学，将起到非常大的辅助作用。

报告题目：基于 HCD 谱图的肽段从头测序

De novo Sequencing of Peptides Using HCD Spectra

董梦秋

北京生命科学研究所

Abstract: Previous de novo sequencing efforts could correctly interpret only 30% of high- and medium-quality tandem mass spectra generated by collision-induced dissociation (CID). This compares poorly to database search, and hence, de novo sequencing did not find any practical use in biological research. Our study shows that higher-energy collisional dissociation (HCD) is much more superior to CID for de novo sequencing. This is because HCD produces high mass accuracy tandem mass spectra, the majority of which contain complete ion series. Besides, abundant internal and immonium ions in the HCD spectra can help differentiate between similar sequences. Based on the characteristics of HCD spectra, an algorithm called pNovo was developed for efficient de novo sequencing. Tested with HCD data of tryptic digests of either a simple protein mixture or a highly complex C. elegans whole-worm lysate, pNovo correctly assigned peptide sequences to at least 80% of the HCD spectra that were independently identified by database search. The number of correct full-length peptide sequences generated by pNovo was comparable with that obtained by database search. A distinct advantage of de novo sequencing is that deamidated peptides and peptides with amino acid mutations can be identified efficiently without extra cost in computation. In short, by taking advantage of the HCD characteristics, pNovo makes an excellent tool for de novo peptide sequencing.

报告题目：从未知基因组到可测定的蛋白质组：通过从头测序来研究依赖于 pH 值的 N10 细菌蛋白质组

From an unknown genome to a measurable proteome: Studying on the pH-dependent proteomes in N10 bacteria by denovo sequencing

王全会

中科院北京基因组研究所

刘斯奇

中科院北京基因组研究所

Abstract: It is very important to understand why alkaliphiles could survive from alkaline environment, which is still poorly understood yet. Alkalimonas amylolytica N10 is such a kind of alkaliphilic bacteria found in China, which could survive from pH 7.5 to 11.0. Till now, only limited numbers of proteins were found in alkaliphiles, which are involved in regulation of pH homeostasis and play important roles in the adaptive mechanism. More proteomic surveys are highly expected to explore the pH-dependent proteins in alkaliphilic bacteria. Being absent of genome data of N10, a combined strategy, including derivative with 4-sulfophenyl isothiocyanate (SPITC) and non-derivative de novo peptide sequencing were used to interpret proteome data. The protein identification rate was highly improved from 18.1% to 73.6%, and approximately 40 pH-dependent proteins were identified from the 72 differential spots. This study give us the confidence that mass data could be directly interpreted but not dependent upon genomics.

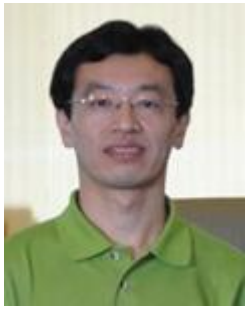
报告题目：利用质谱和同源数据库进行全蛋白测序

Complete Protein Sequencing with MS/MS and a Homologous Database

马 斌

加拿大滑铁卢大学

Abstract: Protein identification with database search and peptide de novo sequencing are routinely carried out with tandem mass spectrometry (MS/MS) today. However it is still a challenging problem to use MS/MS to sequence a complete protein that is not in a sequence database. Such possibility has been demonstrated by biochemists with human interpretation of the data. We present a novel algorithm and automated software, named CHAMPS, for sequencing the complete protein from MS/MS data and a homologous protein database. Experiments with two standard proteins showed that our automated method yields greater than 99% sequence coverage and 100% sequence accuracy on these two proteins.



报告题目：电子转运裂解质谱:特征发现与鉴定应用

Electron Transfer Dissociation: Characterization and Applications in Protein Identification

孙瑞祥

中科院计算技术研究所

Abstract: In recent years, electron transfer dissociation (ETD) has enjoyed widespread applications from sequencing of peptides with or without post-translational modifications to top-down analysis of intact proteins. However, peptide identification rates from ETD spectra compare poorly with those from collision induced dissociation (CID) spectra, especially for doubly charged precursors. This is in part due to an insufficient understanding of the characteristics of ETD and consequently a failure of database search engines to make use of the rich information contained in the ETD spectra. In this study, we statistically characterized ETD fragmentation patterns from a collection of 461,440 spectra, and subsequently implemented our findings into pFind, a database search engine developed earlier for CID data. From ETD spectra of doubly charged precursors, pFind 2.1 identified 63~122% more unique peptides than Mascot 2.2 under the same 1% false discovery rate. For higher charged peptides as well as phosphopeptides, pFind 2.1 also consistently obtained more identifications. Of the features built into pFind 2.1, the following two greatly enhanced its performance: 1) refined automatic detection and removal of high-intensity peaks belonging to the precursor, charge-reduced precursor, or related neutral loss species, whose presence often set spectral matching askew; 2) a thorough consideration of hydrogen-rearranged fragment ions such as z+H and c-H for peptide precursors of different charge states. Our study has revealed that different charge states of precursors result in different hydrogen rearrangement patterns. For a fragment ion, its propensity of gaining or losing a hydrogen depends on (1) the ion type (c or z) and (2) the size of the fragment relative to the precursor, and both dependencies are affected by (3) the charge state of the precursor. In addition, we discovered ETD characteristics that are unique for certain types of amino acids (AAs), such as a prominent neutral loss of SCH2CONH2 (90.0014 Da) from z ions with a carbamidomethylated cysteine at the N-terminus and a neutral loss of histidine side chain C4N2H5 (81.0453 Da) from precursor ions containing histidine. The comprehensive list of ETD characteristics summarized in this paper should be valuable for automated database search, de novo peptide sequencing, and manual spectral validation.

报告题目：基于质谱的蛋白质组学数据处理新方法和平台发展

Development of Methods and Platform for Data Processes in Mass Spectrometry Based Proteome Research

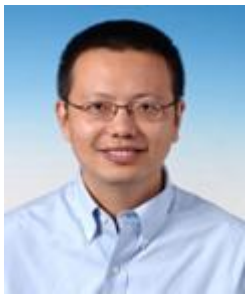
邹汉法

中科院大连化学物理研究所

摘要：在基于质谱的蛋白质组学研究中，高可信度的肽段和蛋白的鉴定信息是蛋白质组学研究的目的和后续生物信息分析挖掘的源头。我们在蛋白质组学数据处理方法和平台方面，分别发展了针对非修饰肽段和磷酸化肽段鉴定的数据筛选方法，并建立了一个磷酸化蛋白质组学数据处理专用平台以促进蛋白质组学研究。对于非修饰肽段的鉴定，发展了一种基于遗传算法的数据筛选门槛优化方法以克服基于模型的鉴定结果筛选方法的高风险性和大部分数据筛选标准仍基于经验筛选标准的问题。通过结合正伪数据库检索，该方法可以安全快速的针对每个数据集得出指定可信度下的最优筛选门槛，在不降低可信度的前提下提高了肽段和蛋白鉴定的数目。对于非修饰肽段的鉴定，还发展了一种基于实例的肽段鉴定概率的计算方法。该方法基于 Shannon 信息熵和 k 近邻算法，利用局部假阳性率实现对肽段鉴定正确概率进行计算。计算得到概率可以很好的与实际概率相匹配。由于该方法不基于假设模型，因而可以安全的应用于各种具有不同特点的数据处理。该方法通过合理的结合多个具有判别能力的参数值，大大提高了肽段和蛋白鉴定的数目，并且计算得到的概率可以直接应用于蛋白鉴定概率的计算。针对磷酸化蛋白质组学中磷酸化肽段鉴定难，假阳性率高，主要依赖于人工验证的现状，发展了一种结合 MS2 和 MS3 图谱以及正伪数据库检索的自动磷酸化肽段鉴定方法。该方法结合了 MS2 和 MS3 的鉴定信息，提高了磷酸化肽段鉴定的灵敏度和可信度，可以自动的对磷酸化肽段进行鉴定而无需进一步的人工验证。进而发展了一种基于分类筛选的磷酸化肽段鉴定方法，该方法结合了 MS2/MS3 方法的高可信度，并且考虑了部分不易发生中性丢失的磷酸化肽段的鉴定，进一步提高了磷酸化肽段鉴定的灵敏度。此外发展了一种利用小库进行正伪检索，实现对单个蛋白质磷酸化的深度分析。建立了一个跨平台，支持多个数据库检索软件的磷酸化蛋白质组学数据处理平台，克服了目前磷酸化蛋白质组学数据处理中的繁琐性，大大加快了磷酸化蛋白质组学数据处理的速度，提高了鉴定结果的灵敏度和可靠性。并且该平台可以同时适用于非修饰和具有其他修饰的肽段和蛋白数据处理。

叶明亮

中科院大连化学物理研究所



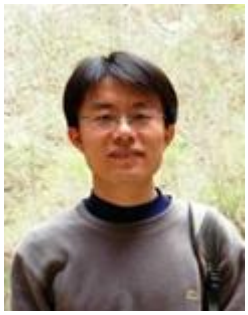
报告题目：基于优化的肽质量指纹谱方法鉴定蛋白质混合物

Optimization-Based Peptide Mass Fingerprinting for Protein Mixture Identification

余维川

香港科技大学

Abstract: In current proteome research, the most widely used method for protein mixture identification is probably peptide sequencing. Peptide sequencing is based on tandem Mass Spectrometry (MS/MS) data. The disadvantage is that MS/MS data only sequences a limited number of peptides and leaves many more peptides uncovered. Peptide Mass Fingerprinting (PMF) has been widely used to identify single purified proteins from single-stage MS data. Unfortunately, this technique is less accurate than the peptide sequencing method and can not handle protein mixtures, which hampers the widespread use of PMF technique. In this talk, we tackle the problem of protein mixture identification from an optimization point of view. We show that some simple heuristics can find good solutions to the optimization problem. Through a comprehensive simulation study, we identify a set of limiting factors that hinder the performance of PMF-based protein mixture identification. We argue that it is feasible to remove these limitations and PMF can be a powerful tool in the analysis of protein mixtures, especially in the identification of low-abundance proteins which are less likely to be sequenced by MS/MS scanning.



付 岩

中科院计算技术研究所

报告题目：基于相关谱图对的非限制性修饰检测

Unrestrictive modification detection based on related spectral pairs

Abstract: Identification of proteins and their modifications via liquid chromatography-tandem mass spectrometry (LC-MS/MS) is an important task for the field of proteomics. However, due to the complexity of tandem mass spectra, the majority of the spectra cannot be identified. The presence of unanticipated protein modifications is among the major reasons for the low spectral identification rate. Efficient and comprehensive detection of protein modifications has become one of the most important and challenging problems in MS/MS-based proteomics. In this report, I will introduce two new computational approaches for unrestrictive detection of modifications. Both of them are based on the fact that the modified and unmodified versions of a peptide are usually present simultaneously in a sample. The first approach takes full advantage of the precursor information and can detect abundant modifications in a very fast speed. The second approach is by open spectral library searching and can detect lower-abundance modifications. On various datasets, our methods successfully detected many anticipated modifications, yielding deep insights into the data. The spectral identification rates were significantly increased, and many modified peptides were identified.

报告题目：评价诱饵库设计，搜索策略，匹配误差和质量控制对鸟枪法蛋白质组学中肽段鉴定精确性的影响

Evaluation of the effect of decoy design, search strategy, mass tolerance and quality control method on the accuracy of peptide identifications in shotgun proteomics

朱云平

北京蛋白质组研究中心

Abstract: In the last decade, a large number of quality control methods have been developed to improve the sensitivity of peptide identifications. False discovery rate (FDR) and q-value have been widely used to estimate/control false positive identifications based on target-decoy strategy. However, the accuracy of q-value estimation and its influence factors need to be carefully examined. Using a widely used standard protein dataset with high mass accuracy, we found that decoy database of reversing tryptic peptides, composite database search generated significant lower deviation from actual q-values across various parent mass tolerance and quality control methods and that parent mass tolerance could significantly affect q-value accuracy when separate search was applied. This analysis offered an important clue as to how to find the optimal decoy design, search strategy, search parameters, and quality control method to achieve sensitive and accurate peptide identifications in shotgun proteomics.

报告题目：BuildSummary ：一个基于目标-诱饵策略的蛋白质鉴定整合软件

BuildSummary: A software tool for assembling protein

盛泉虎

上海生命科学院

Abstract:Target-decoy database search strategy has been widely accepted as a standard method for evaluating the confidence of peptide/protein identifications in high throughput shotgun proteomics. In order to maximize the number of confident proteins above a threshold of false discovery rate (FDR), a post-processing procedure is often used to integrate results from different peptide search engines for the same dataset. Specifically, three issues should be addressed in the post-processing procedure: 1) how to maximize the number of peptide-spectrum matches (PSMs) above a given peptide FDR? 2) how to maximize the number of confident proteins above a given protein FDR? and 3) how to integrate results from different search engines? We show that the PSMs with different charges, different number of missed internal cleavage sites, different modification states, different number of protease termini should be grouped into different categories for filtering PSMs above an FDR threshold. We also developed an iterative procedure that simultaneously assigns FDRs to identified peptides and corresponding proteins in each group of protein. Finally, we build a general framework to integrate results from different peptide search engines (or the same engine by using different parameters) based on the post-processing procedure. The framework allows a user to combine many independent PSM scoring algorithms including de novo sequencing and spectrum library search algorithms, as long as the same peptide FDR is applied to each of them by using target-decoy search approach. We implemented the methods described above in user-friendly software called BuildSummary, which report a list of identified proteins under a user-defined protein FDR as well as a list of identified peptides under a low peptide FDR in these proteins. The software is tested by using several shotgun proteomics datasets acquired by multiple LC/MS instruments from two different biological samples, and shows satisfactory performance. BuildSummary can be downloaded as part of software suite ProteomicsTools freely from http://www.proteomics.ac.cn/software/index.htm.

报告题目：冷冻电镜中的计算方法：图像数据处理和三维重构

Computational methods in cryo-electron microscopy: image data processing and 3D structure reconstruction

张 法

中科院计算技术研究所

Abstract: Cryo-electron microscopy (cryoEM) has emerged as a method of choice for determining the three-dimensional (3D) structures of biological complexes towards atomic resolution. It has the potential of revealing structural details of large macromolecular machines at near atomic scale without requiring crystallization, and unlike other techniques, can examine structural changes occurring during operation of the molecule. To obtain a high resolution structure using cryoEM, large numbers (104~105) of individual images of the structure is combined to form an average three-dimensional electron density map. Particle selection has becoming a significant bottleneck in cryoEM. Moreover, each of individual images requires a complex series of computing to obtain the final result. So the needs for computational power are constantly growing with the increasing complexity of algorithms and the amount of data needed to push the resolution limits. In this report, I will present some research progresses of computational methods of our group in cryoEM, which includes: 1) Picker, an automated particle selection algorithm, 2) ParaEMAN, a high efficient parallel strategy for 3D structure reconstruction of cryoEM and 3) ISAFRS, a new 3D structure reconstruction system in cryoEM.

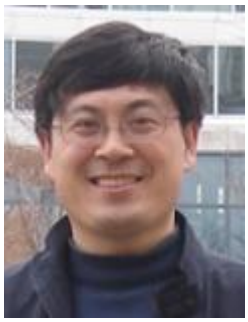
报告题目：DomainRBF: 一种对疾病相关蛋白质结构域进行优先排序的贝叶斯回归方法

DomainRBF: a Bayesian regression approach to the prioritization of associations between protein domains and human complex diseases

江 瑞

清华大学

Abstract: Domains are basic units of proteins, and thus exploring associations between protein domains and human inherited diseases will greatly improve our understanding of the pathogenesis of human complex diseases and further benefit the medical prevention, diagnosis and treatment of these diseases. Based on the assumption that domain proximities in a domain-domain interaction network imply phenotype similarities between human diseases, we propose in this paper a Bayesian regression approach named "domainRBF" (domain Rank with Bayes Factor) to prioritize associations between candidate domains and human diseases. Using a compiled a dataset that contains 1,614 associations between 671 domains and 1,145 phenotypes, we demonstrate the effectiveness of this approach through three large-scale leave-one-out cross-validation experiments (random control, simulated linkage interval, and genome-wide scan) in terms of three criteria (precision, mean rank ratio, and AUC score). Results show that the approach can effectively rank susceptible domains among the top of a few candidates. We further show that the proposed approach is robust to parameters involved and the underlying domain-domain interaction network through a series of permutation tests. With the validity of this approach being accessed, we show the possibility of ab initio inference of domain-disease associations and gene-disease associations, and we illustrate the well agreement of our inference with evidence of genome-wide association studies for four common diseases (type 1 diabetes, type 2 diabetes, Crohn's disease, and breast cancer). Finally, provide a precalculated genome-wide landscape of associations between 5,490 protein domains and 5,080 human diseases and offer free access to this resource.



报告题目：蛋白质结构“字母表”设计

Designing Succinct Structural Alphabets

卜东波

中科院计算技术研究所

Abstract: The three dimensional structure of a protein sequence can be assembled from the substructures corresponding to small segments of this sequence, \cite{Sim97,Levitt2002}. For each small sequence segment, there are only a few more likely substructures. We call them the "structural alphabet" for this segment. Classical approaches such as ROSETTA used sequence profile and secondary structure information, to predict structural fragments. In contrast, we utilize more structural information, such as solvent accessibility and contact capacity, for finding structural fragments. Integer linear programming technique is applied to derive the best combination of these sequence and structural information items. This approach generates significantly more accurate and succinct structural alphabets with more than 50\% improvement over the previous accuracies. With these novel structural alphabets, we are able to construct more accurate protein structures than the state-of-art Ab Initio protein structure prediction programs such as ROSETTA. We are also able to reduce the Kolodny's library size by a factor of 8, at the same accuracy.

报告题目：蛋白质作为分子化石

Proteins as molecular fossils

张红雨

华中农业大学

Abstract: On the stage of evolution, fossils are central players. Fossils not only refer to preserved body remains of ancient organisms, but also include molecules that are durable during the geological history. The latter are also known as molecular fossils. Traditionally, molecular fossils are small-molecule metabolites, including biomarkers (e.g., porphyrins, hopanoids, biphytanes and sterols), and very conserved coenzymes (e.g., ATP, NAD, NADP and FAD). Recently, accumulating evidence indicated that proteins can also serve as molecular fossils, which include protein remains of ancient organisms, such as collagen peptides in dinosaurs, and extremely conserved protein features, such as protein folds. Using a phylogenomic structural census in hundreds of proteomes, we build phylogenies and timelines of domains at fold and fold superfamily levels of structural complexity. These timelines correlate approximately linearly with geological timescales and can be used to date some crucial events in life history, such as planet oxygenation and organism diversification.