
中科院计算所
生物信息学研究组
欢迎全国各地
新老朋友出席
首届中国
计算蛋白质组学
研讨会

首届中国计算蛋白质组学研讨会简介

中国科学院计算技术研究所

贺思敏

2010年11月10日

计算蛋白质组学：新方向

- 基因组学 **Genomics**
  - →生物信息学 **Bioinformatics**
  - →蛋白质组学 **Proteomics**
    - →计算蛋白质组学 **Computational Proteomics**

2008-05-05	Google	PubMed	IEEE	ACM
“Genomics”	14,600,000	25,143	375	1,197
“Bioinformatics”	9,910,000	20,389	3,185	2,177
“Proteomics”	4,730,000	16,393	191	241
“Computational Proteomics”	13,100	33	1	9

计算蛋白质组学：国际会议 · 美国

- **2005.01.11-14: MBI Workshop on Computational Proteomics and Mass Spectrometry, Ohio State University**
- **2005.05.16-17: Workshop on Structural and Computational Proteomics of Biological Complexes, Rice Univ.**
- **2006.12.02-03: The first RECOMB Satellite Workshop on Computational Proteomics, UCSD**

计算蛋白质组学：国际会议 · 欧洲

- **2005**.11.20-25: The Dagstuhl Seminar on **Computational Proteomics**, No. 05471
- **2006**.08.24-25: International BCB-Workshop on **Computational Proteomics**
- **2008**.03.02-07: The Dagstuhl Seminar on **Computational Proteomics**, No. 08101

计算蛋白质组学：国际会议 · 其他

- **2005 - ：** International Symposium on Computational Life Science, Topic on **Computational Proteomics**
- **2006 - ：** IEEE International Symposium on Computer-Based Medical Systems (CBMS), Special Track on **Computational Proteomics**
- **2006, 2007：** Pacific Symposium on Biocomputing, Session on **Computational Proteomics**

计算蛋白质组学：国际期刊·综述

- **2001**.07：E.T. Maggio and K. Ramnarayan, Recent developments in computational proteomics, TRENDS in Biotechnology, vol. 19, no. 7, pp. 266-272, July 2001.
- **2002**.03：E. Razvi, Market opportunity in computational proteomics, Biotechniques, Suppl: 54-8, 60-2, March 2002.
- **2007.07：J. Colinge and K.L. Bennett, Introduction to Computational Proteomics, PLoS Comput. Biol. 2007 July; 3(7): e114.**
- **2007.08：R. Matthiesen, Methods, algorithms and tools in computational proteomics: a practical point of view, Proteomics, vol. 7, no. 16, pp. 2815–2832, Aug. 2007.**

计算蛋白质组学：国际期刊·专辑

- **2002.03**：BioTechniques, Supplement on Computational Proteomics, Mar. 2002.

- **2008.01**：**J. Proteome Research, Special Issue on Statistical and Computational Proteomics, Jan. 2008.**

- **2008.03**：**Briefings in Bioinformatics, Special Issue on Computational Proteomics, Mar. 2008.**

计算蛋白质组学：定义

- **Computational Proteomics**
  - **CBMS: Management and Analysis of Proteomics Data**
  - **PSB: High-Throughput Analysis for Systems Biology**

计算蛋白质组学：内涵

系统生物学的高通量分析

蛋白质表达-结构-功能的高通量分析

基于质谱技术的
规模化蛋白质表达分析

计算蛋白质组学：新动向

PNAS PNAS

SPECIAL FEATURE: PERSPECTIVE

Precision proteomics: The case for high resolution and high mass accuracy

Matthias Mann*† and Neil L. Kelleher‡*

*Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany; and ‡Department of Chemistry and the Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801

Edited by Jack Halpern, University of Chicago, Chicago, IL, and approved August 20, 2008 (received for review February 1, 2008)

Proteomics has progressed radically in the last 5 years and is now on par with most genomic technologies in throughput and comprehensiveness. Analyzing peptide mixtures by liquid chromatography coupled to high-resolution mass spectrometry (LC-MS) has emerged as the main technology for in-depth proteome analysis whereas two-dimensional gel electrophoresis, low-resolution MALDI, and protein arrays are playing niche roles. MS-based proteomics is rapidly becoming quantitative through both label-free and stable isotope labeling technologies. The latest generation of mass spectrometers combines extremely high resolving power, mass accuracy, and very high sequencing speed in routine proteomic applications. Peptide fragmentation is mostly performed in low-resolution but very sensitive and fast linear ion traps. However, alternative fragmentation methods and high-resolution fragment analysis are becoming much more practical. Recent advances in computational proteomics are removing the data analysis bottleneck. Thus, in a few specialized laboratories, "precision proteomics" can now identify and quantify almost all fragmented peptide peaks. Huge challenges and opportunities remain in technology development for proteomics; thus, this is not "the beginning of the end" but surely "the end of the beginning."

Analysis of individual proteins by classical methods and by mass spectrometry (MS) has been an indispensable cornerstone of biochemistry for many decades. Large-scale analysis of the whole protein complement of cells, tissues, and body fluids (proteomics) would additionally enable the unbiased comparison of different cellular states in biology and medicine at a "systems-wide" level. However, technological challenges associated with proteomics have long prevented its widespread adoption. Two-dimensional (2D) gel electrophoresis was conceived more than 30 years ago (1). This technology has been useful for low-complexity protein mixtures but never matured into a comprehensive and accurate proteomics technology. The introduction of high-sensitivity protein identification by MS at first seemed to help 2D gel analysis, but in fact it revealed that the thousands of spots seen in the gel maps are actually variants of a few hundred of the most abundant proteins (2). Recently, it has also become clear that quantitation of even these proteins is far from accurate because of spot overlap (3). Accordingly, "biomarkers" found by these technologies tend to be the same regardless of the system under investigation (4).

In principle, protein arrays might be applicable to proteomics in a similar way that gene chips have been to the measurement of RNA. However, the challenge associated with expressing thousands of full-length proteins and immobilizing them in a native state on a chip is daunting (5, 6). In practice, the role of protein arrays has been limited, and the literature contains few examples of their successful use. MS technology with low resolving power, especially in the form of the so-called SELDI method (7), caught the imagination of clinicians a few years ago. This approach involves measuring a MALDI spectrum of proteins from the body fluid of a patient and then employs machine learning to differentiate disease and healthy states. However, from a mass-spectrometric point of view, SELDI boils down to simple MALDI spectra of very complex mixtures and would be expected to only yield a subset of the most abundant low-mass peptides and protein fragments. Such species could still have proven sufficient to classify patient samples. However, as the scientific community demanded identification of the peaks comprising the SELDI patterns, these usually turned out to belong to the same nonspecific proteins unlikely to be directly associated with the disease.

In contrast to the above approaches, which were discussed as promising proteomics technologies as late as a few years ago, MS-based proteomics has taken great strides in development. MS-based protein science has always been extremely useful in studies focused on individual proteins, but large-scale proteomics is increasingly realizing its turn-of-the-millennium promises, too. In particular, technological improvements in the last 5 years have dramatically increased the routine availability of extremely high-performance MS. In many but not all cases, these technologies already existed but could only be applied in specialized situations by expert laboratories and at low throughput. The main purpose of this perspective is to show that MS techniques with high accuracy can and should now be applied routinely in most proteomics contexts, and that there is no penalty for their use. In fact, we argue that precise and comprehensive analysis of complex proteomes is best achieved by using high-resolution proteomics technologies. There are many other important aspects of MS-based proteomics that have been the subjects of recent reviews and that will not serve as focal points here. For example, the remarkable inroads of proteomics strategies into the quantitative analysis of posttranslational modifications (8), the determination of protein interactions (9), and the ongoing integration of MS technology with other powerful tools of molecular biology (10) are not discussed here.

The Importance of Being Highly Resolved

The current mainstream format in large-scale proteomics involves the analysis of very complex peptide mixtures. In this "shotgun approach" (11), tens of thousands of peptides with very large dynamic range (i.e., the concentration difference between the most and least abundant peptides) have to be analyzed in several chromatographic runs. If these mixtures are measured with ion traps or other MS instruments of lower resolving power, coeluting peptides with similar m/z ratios frequently overlap. This precludes accurate mass analysis,

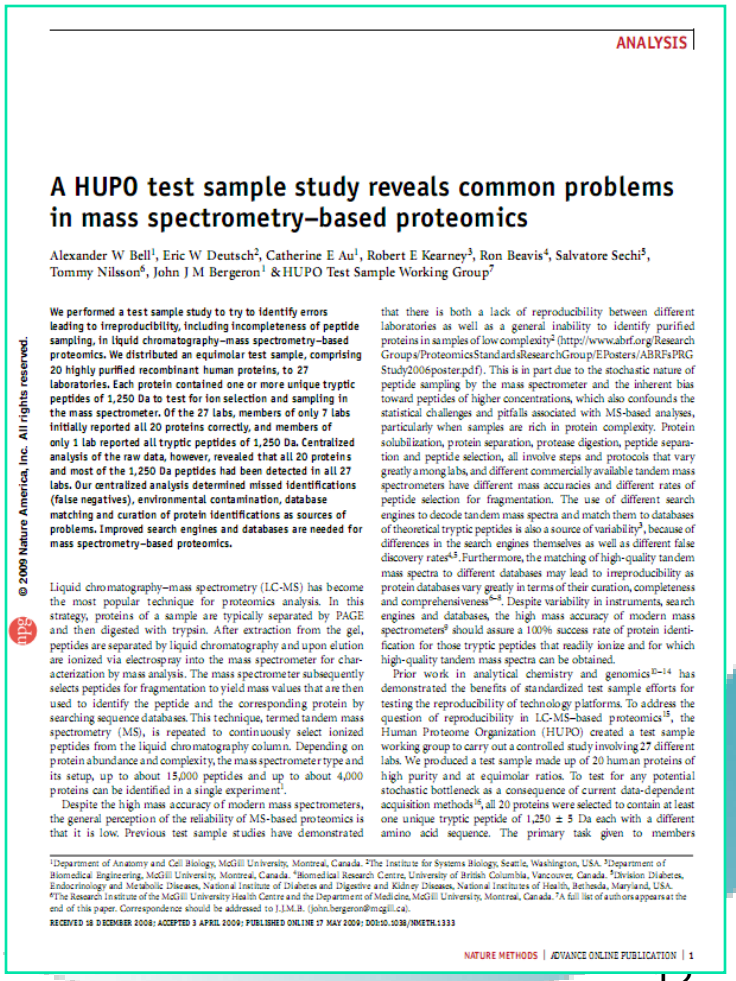
Author contributions: M.M. and N.L.K. wrote the paper.
The authors declare no conflict of interest.
This article is a PNAS Direct Submission.
†To whom correspondence may be addressed. E-mail: mmann@biochem.mpg.de or kelleher@scs.uiuc.edu.

计算蛋白质组学：新动向

HUPO (Human Proteome Organisation):

- We distributed an equimolar test sample, comprising 20 highly purified recombinant human proteins, **to 27 laboratories**.
- Each protein contained one or more unique tryptic peptides of 1,250 Da to test for ion selection and sampling in the mass spectrometer.
- **Only 7 labs** initially reported all 20 proteins correctly, and members of **only 1 lab** reported all tryptic peptides of 1,250 Da.

Bell, A.W. et al. (2009) "A HUPO test sample study reveals common problems in mass spectrometry-based proteomics."
Nature Methods 6(6): 423-30.



ANALYSIS

A HUPO test sample study reveals common problems in mass spectrometry–based proteomics

Alexander W Bell[1], Eric W Deutsch[2], Catherine E Au[1], Robert E Kearney[3], Ron Beavis[4], Salvatore Sechi[5], Tommy Nilsson[6], John J M Bergeron[1] & HUPO Test Sample Working Group[7]

We performed a test sample study to try to identify errors leading to irreproducibility, including incompleteness of peptide sampling, in liquid chromatography–mass spectrometry–based proteomics. We distributed an equimolar test sample, comprising 20 highly purified recombinant human proteins, to 27 laboratories. Each protein contained one or more unique tryptic peptides of 1,250 Da to test for ion selection and sampling in the mass spectrometer. Of the 27 labs, members of only 7 labs initially reported all 20 proteins correctly, and members of only 1 lab reported all tryptic peptides of 1,250 Da. Centralized analysis of the raw data, however, revealed that all 20 proteins and most of the 1,250 Da peptides had been detected in all 27 labs. Our centralized analysis determined missed identifications (false negatives), environmental contamination, database matching and curation of protein identifications as sources of problems. Improved search engines and databases are needed for mass spectrometry–based proteomics.

Liquid chromatography–mass spectrometry (LC-MS) has become the most popular technique for proteomics analysis. In this strategy, proteins of a sample are typically separated by PAGE and then digested with trypsin. After extraction from the gel, peptides are separated by liquid chromatography and upon elution are ionized via electrospray into the mass spectrometer for characterization by mass analysis. The mass spectrometer subsequently selects peptides for fragmentation to yield mass values that are then used to identify the peptide and the corresponding protein by searching sequence databases. This technique, termed tandem mass spectrometry (MS), is repeated to continuously select ionized peptides from the liquid chromatography column. Depending on protein abundance and complexity, the mass spectrometer type and its setup, up to about 15,000 peptides and up to about 4,000 proteins can be identified in a single experiment[1].

Despite the high mass accuracy of modern mass spectrometers, the general perception of the reliability of MS-based proteomics is that it is low. Previous test sample studies have demonstrated that there is both a lack of reproducibility between different laboratories as well as a general inability to identify purified proteins in samples of low complexity[2] (http://www.abrf.org/ResearchGroups/ProteomicsStandardsResearchGroup/EPosters/ABRFsPRGStudy2006poster.pdf). This is in part due to the stochastic nature of peptide sampling by the mass spectrometer and the inherent bias toward peptides of higher concentrations, which also confounds the statistical challenges and pitfalls associated with MS-based analyses, particularly when samples are rich in protein complexity. Protein solubilization, protein separation, protease digestion, peptide separation and peptide selection, all involve steps and protocols that vary greatly among labs, and different commercially available tandem mass spectrometers have different mass accuracies and different rates of peptide selection for fragmentation. The use of different search engines to decode tandem mass spectra and match them to databases of theoretical tryptic peptides is also a source of variability[3], because of differences in the search engines themselves as well as different false discovery rates[4,5]. Furthermore, the matching of high-quality tandem mass spectra to different databases may lead to irreproducibility as protein databases vary greatly in terms of their curation, completeness and comprehensiveness[6–8]. Despite variability in instruments, search engines and databases, the high mass accuracy of modern mass spectrometers[9] should assure a 100% success rate of protein identification for those tryptic peptides that readily ionize and for which high-quality tandem mass spectra can be obtained.

Prior work in analytical chemistry and genomics[10–14] has demonstrated the benefits of standardized test sample efforts for testing the reproducibility of technology platforms. To address the question of reproducibility in LC-MS–based proteomics[15], the Human Proteome Organization (HUPO) created a test sample working group to carry out a controlled study involving 27 different labs. We produced a test sample made up of 20 human proteins of high purity and at equimolar ratios. To test for any potential stochastic bottleneck as a consequence of current data-dependent acquisition methods[16], all 20 proteins were selected to contain at least one unique tryptic peptide of 1,250 ± 5 Da each with a different amino acid sequence. The primary task given to members

[1]Department of Anatomy and Cell Biology, McGill University, Montreal, Canada. [2]The Institute for Systems Biology, Seattle, Washington, USA. [3]Department of Biomedical Engineering, McGill University, Montreal, Canada. [4]Biomedical Research Centre, University of British Columbia, Vancouver, Canada. [5]Division Diabetes, Endocrinology and Metabolic Diseases, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland, USA. [6]The Research Institute of the McGill University Health Centre and the Department of Medicine, McGill University, Montreal, Canada. [7]A full list of authors appears at the end of this paper. Correspondence should be addressed to J.J.M.B. (john.bergeron@mcgill.ca).

RECEIVED 18 DECEMBER 2008; ACCEPTED 3 APRIL 2009; PUBLISHED ONLINE 17 MAY 2009; DOI:10.1038/NMETH.1333

计算蛋白质组学：新动向

RECOMB Satellite Conference on Computational Proteomics 2010

March27-28, 2010 Calit2 Auditorium, Atkinson Hall, UC San Diego

http://proteomics.ucsd.edu/recombcp2010

In collaboration with **Journal of Proteome Research**

RECOMB Satellite Conference on Computational Proteomics 2011

March 11-13, 2011 Calit2 Auditorium, Atkinson Hall, UC San Diego

http://proteomics.ucsd.edu/recombcp2010

In collaboration with **Journal of Proteome Research**

计算蛋白质组学：新动向

2008-05-05	Google	PubMed	IEEE	ACM
“Genomics”	14,600,000	25,143	375	1,197
“Bioinformatics”	9,910,000	20,389	3,185	2,177
“Proteomics”	4,730,000	16,393	191	241
“Computational Proteomics”	13,100	33	1	9

2010-10-27	Google	PubMed	IEEE	ACM
“Genomics”	13,100,000	61,943	2,078	1,997
“Bioinformatics”	11,600,000	77,801	10,455	4,173
“Proteomics”	2,110,000	26,530	615	424
“Computational Proteomics”	12,300	50	11	25

首届中国计算蛋白质组学研讨会

- 北京蛋白质组研究中心贺福初团队
- 华大基因研究院刘斯奇团队
- 中科院大化所邹汉法团队
- 暨南大学何庆瑜团队
- 复旦大学杨芃原团队
- 上海生科院曾嵘团队
- 清华大学张学工团队
- 协和医科大学高友鹤团队
- 上海生物信息中心李亦学团队
- 北京生命科学研究所董梦秋团队

首届中国计算蛋白质组学研讨会

159人注册参会：全国14省区155人，美加4人

- [1] 北京 101
- [2] 上海 10
- [3] 黑龙江 2
- [4] 吉林 4
- [5] 辽宁 10
- [6] 内蒙古1
- [7] 河北 2
- [8] 江苏 1
- [9] 浙江 4
- [10] 湖北 5
- [11] 湖南 3
- [12] 江西1
- [13] 广东9
- [14] 香港2
- [15] 美国2
- [16] 加拿大2

- 研究所84
- 高校49
- 医院12
- 公司11
- 其他2

首届中国计算蛋白质组学研讨会

159人注册参会：全国14省区155人，美加4人

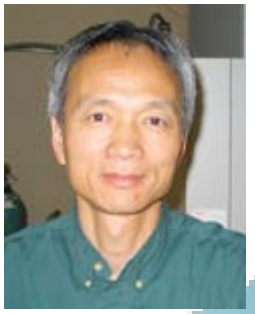
- 研究所84
- 高校49
- 医院12
- 公司11
- 其他2

首届中国计算蛋白质组学研讨会

72人注册培训：全国14省区71人，美1人

- [1] 北京 43
- [2] 上海 4
- [3] 黑龙江 3
- [4] 吉林 4
- [5] 辽宁 3
- [6] 内蒙古 1
- [8] 江苏 1
- [10] 湖北 2
- [11] 湖南 3
- [13] 广东 6
- [14] 香港 1
- [15] 美国 1

- 研究所 32
- 高校 23
- 医院 9
- 公司 8



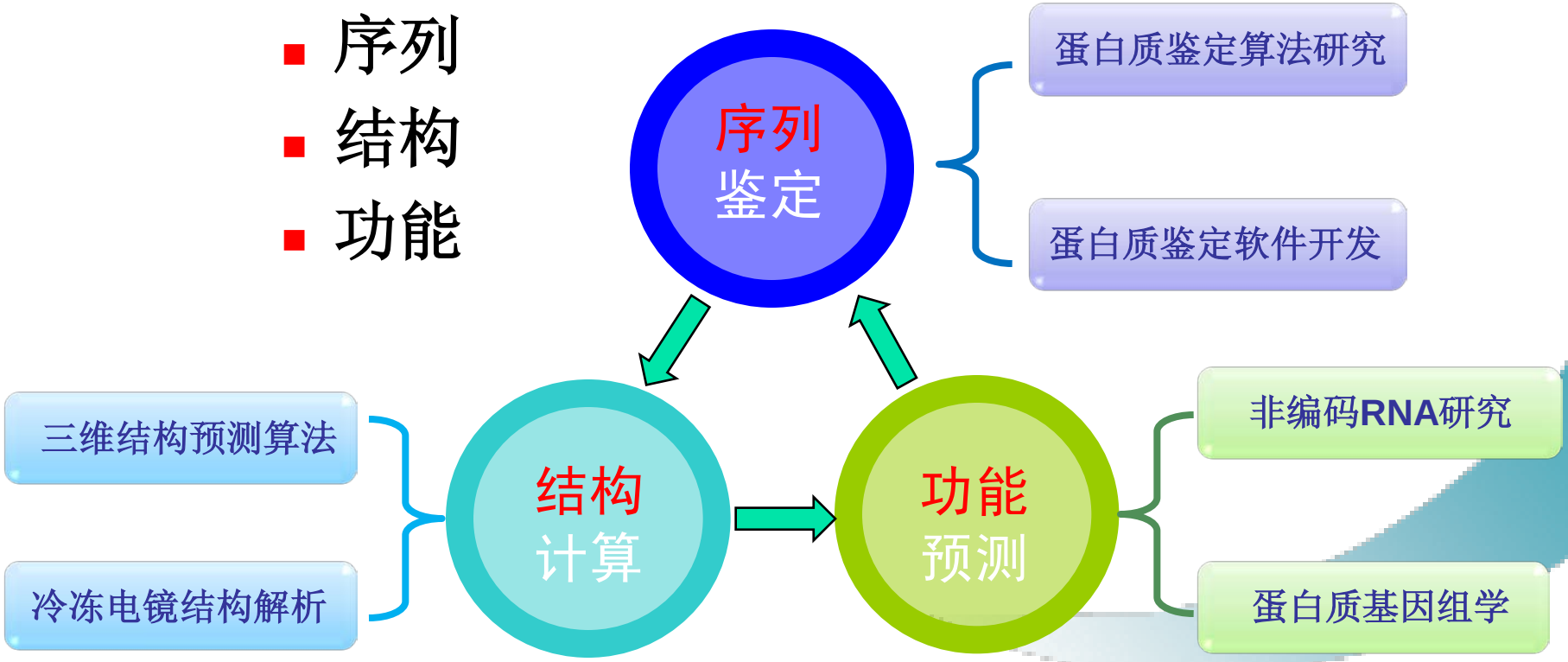
首届中国计算蛋白质组学研讨会

首届中国计算蛋白质组学研讨会

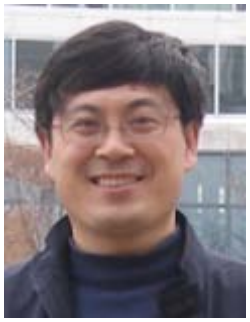
首届中国计算蛋白质组学研讨会

首届中国计算蛋白质组学研讨会

- 计算所的计算蛋白质组学研究
  - 序列
  - 结构
  - 功能

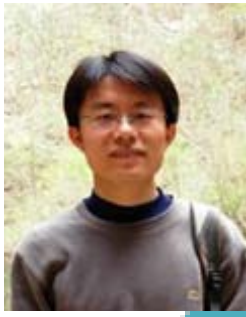


首届中国计算蛋白质组学研讨会



计算所的计算蛋白质组学研究

- 序列
- 结构
- 功能



首届中国计算蛋白质组学研讨会

	高友鹤 中国协和 医科大学	贺福初 北京蛋白质组 研究中心	李亦学 / 曾嵘 上海生命 科学研究院	贺思敏 中科院计算所	其他单位
2010		马洁	李虹	王海鹏	
2009		李涛，孙爱华			于长永 [东北大学]
2008	邵晨	郝运伟，刘伟	王连水，盛泉虎		沈菊文 [上海药物所]
2007		张纪阳，贠栋	李素君	付岩，李德泉	唐凯临 [同济大学]
2006		李栋	俞晓晶	张京芬	张卓 [生物物理所]
2005	孙伟	吴松峰，万平			

首届中国计算蛋白质组学研讨会

	高友鹤 中国协和 医科大学	贺福初 北京蛋白质组 研究中心	李亦学 / 曾嵘 上海生命 科学研究院	贺思敏 中科院计算所	其他单位
2010					
2009					于长永 [东北大学]
2008			王		沈菊文 [上海药物所]
2007		张纪阳，贡栋			唐凯临 [同济大学]
2006		李栋			张卓 [生物物理所]
2005	孙伟	吴松峰，万平			

首届中国计算蛋白质组学研讨会

首届中国计算蛋白质组学研讨会

首届中国计算蛋白质组学研讨会

- 报告人
  - 每人安排30分钟，充分表达但不拖延
  - 建议22分钟报告，8分钟问答
  - 最多25分钟报告，最少5分钟问答
- 参会者
  - 请在会场内关闭手机声音
  - 请在会场外接听电话
  - 请在会场外使用无线网络

谢谢！

让我们静下心来，在喧嚣的北京，
一起欣赏一场纯粹的学术报告会：

首届中国计算蛋白质组学研讨会